UNIVERSITÉ JOSEPH FOURIER - GRENOBLE I
U.F.R. D'INFORMATIQUE
ET DE MATHÉMATIQUES APPLIQUÉES

THÈSE
pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER

Discipline : Mathématiques Appliquées

Présentée et soutenue publiquement par

Myriam GARRIDO

Le 12 juin 2002

Directeur de thèse : Jean DIEBOLT

MODÉLISATION DES ÉVÉNEMENTS RARES ET ESTIMATION DES QUANTILES EXTRÊMES, MÉTHODES DE SÉLECTION DE MODÈLES POUR LES QUEUES DE DISTRIBUTION

Jury

Jean-Louis Soler	Président
Jean-Pierre Raoult	Rapporteur
Armelle Guillou	Rapporteur
Janet Heffernan	Rapporteur
Jean Diebolt	Directeur de thèse
Stéphane Girard	Examinateur
Catherine Bauby	Examinateur

Remerciements

Le moment est venu d'exprimer toute ma reconnaissance à mon directeur de thèse, Jean Diebolt, pour son enthousiasme et son soutien sans faille au cours de ces trois années. Malgré son éloignement, il a toujours été disponible pour me prodiguer ses conseils, et m'inculquer sa grande rigueur.

Je tiens également à remercier Gilles Celeux pour m'avoir accueillie au sein de l'équipe IS2 de l'Inria Rhône-Alpes. Les discussions fructueuses que nous avons eues aux moments clés m'ont beaucoup aidée, et je lui en suis très reconnaissante.

À Jean-Pierre Raoult j'exprime ma profonde gratitude pour avoir accepté, malgré ses nombreuses occupations, d'être rapporteur de ma thèse, pour avoir trouvé le temps de la lire en détail et surtout de me rencontrer pour me faire part de ses remarques. Je lui suis très reconnaissante pour l'attention qu'il a portée à ce travail.

J'adresse mes sincères remerciements à Armelle Guillou et Janet Heffernan pour l'intérêt qu'elles ont manifesté pour ma thèse en acceptant d'en être rapporteur. Je remercie Armelle Guillou pour son enthousiasme qui m'a aidé à contrôler le stress, notamment au moment délicat de la fin de rédaction, et pour ses suggestions qui m'ont été d'un grand intérêt. Je sais gré à Janet Heffernan pour avoir lu ma thèse en Français. Je lui suis également reconnaissante pour ses remarques judicieuses qui m'ont aidée à mieux justifier les méthodes développées, et pour les perpectives qu'elle m'a montrées concernant les méthodes bayésiennes dans le contexte des événements rares.

Je tiens aussi à remercier Jean-Louis Soler qui m'a fait l'honneur de présider mon jury de thèse. Sa gentillesse, sa simplicité ont pleinement contribué à me mettre en confiance dans la dernière ligne droite et au merveilleux souvenir que je garde de ma soutenance.

Je tiens également à remercier chaleureusement Stéphane Girard pour avoir accepté de faire partie de mon jury de thèse, et surtout pour les collaborations fructueuses que nous avons eues, son soutien dans les moments difficiles, et ses nombreux conseils sur la recherche, la thèse, et l'avenir. Je remercie également Catherine Trottier qui m'a accueillie à l'Inria, a accompagné mes premiers pas en recherche et en particulier dans le domaine de l'estimation des événements rares, puis a continué à travailler avec moi après son départ de l'Inria.

J'exprime toute ma reconnaissance aux membres de la division recherche et développement de EDF sans qui rien n'aurait été possible. Je tiens particulièrement à remercier Valérie Durbeck et Benjamin Villain, mes premiers correspondants EDF, ainsi que Catherine Bauby, que je remercie d'avoir accepté de faire partie de mon jury, et Dominique Lagrange, qui les ont remplacés et ont eu la lourde tâche de reprendre un travail déjà bien avancé. Les discussions fructueuses que nous avons eues et leurs descriptions des besoins des ingénieurs m'ont beaucoup aidée. Je tiens aussi à remercier chaleureusement pour leur soutien et leur aide Serge Hugonard-Bruyère et André Lannoy.

Je souhaite aussi remercier Mhamed-Ali El-Aroui et Didier Chauveau pour toute l'aide qu'ils m'ont apportée, notamment en informatique, et pour leur disponibilité.

L'occasion m'est donnée ici de remercier tous les membres d'IS2, notamment mes collègues de bureau : Yann et sa bienveillance de nouveau docteur envers les petits jeunes ; Nathalie et Henri aux conseils avisés de ceux qui viennent de soutenir ; Véro pour nous avoir un peu sorti de nos statistiques ; Jean-Baptiste pour sa patience et son aide en informatique ; Cécile d'un enthousiasme et d'une gaité à toute épreuve ; et Jêrome pour m'avoir supportée, lui qui a eu la malchance d'arriver au pire moment de la rédaction. Grâce à mes amis de IS2, de l'Inria et du LMC, mon séjour à Grenoble a été très agréable et le stress de la thèse moins pesant. Je leur en suis à tous très reconnaissante.

Enfin, j'exprime ma gratitude à ma famille qui m'a toujours soutenue et encouragée dans la voie que je m'étais fixée. Je remercie mes parents qui m'ont stimulée et encouragée pendant mes études. Mes pensées vont à Laurent pour sa patience à toute épreuve et son soutien sans faille tout au long de ces trois années. Je remercie mes grands-parents, ainsi que Christiane et Alain qui sont venus me soutenir lors de ma présentation. Surtout, bonne chance et bon courage à ma petite soeur Edith pour sa future thèse en biologie.

Table des matières

Notations et conventions

Les variables aléatoires sont représentées par des lettres majuscules et les réalisations de ces variables aléatoires par des lettres minuscules.

Soit $\underline{x}_n = (x_1, \dots, x_n)$ un échantillon de taille n. L'échantillon ordonné en ordre croissant est noté $(x_{(1)}, \dots, x_{(n)})$.

On note $f^{\leftarrow}$ l'inverse généralisée de la fonction f, et si f est inversible, f^{-1} l'inverse de f.

Modèles et estimations :

- ▷ modèle $\{F_\theta, \theta \in \Theta\}$: on définit un modèle par l'ensemble des fonctions de répatition F_θ de paramètre θ appartenant au domaine de définition Θ.
- ▷ f_θ la densité de paramètres θ, pour le modèle $\{F_\theta, \theta \in \Theta\}$.
- ▷ F_θ la fonction de répartition de paramètres θ, pour le modèle $\{F_\theta, \theta \in \Theta\}$.
- ▷ $L_n(\theta) = \prod_{i=1}^n f_\theta(x_i)$ la fonction de vraisemblance de θ pour le modèle de densité f_θ et l'échantillon $\underline{x}_n$.
- ▷ $\hat{\theta}_n$ l'estimateur du maximum de vraissemblance (evm) des paramètres θ, pour le modèle $\{F_\theta, \theta \in \Theta\}$ et l'échantillon $\underline{x}_n$.

Abréviations

- ▷ Fdr : fonction de répartition.
- ▷ i.i.d. : indépendant et identiquement distribué (échantillon ou variables aléatoires).
- ▷ POT (méthode) : *Peaks Over Threshold*, méthode des excès.
- ▷ GPD : *Generalised Pareto Distribution*, loi de Pareto généralisée.
- ▷ PWM (estimateur) : *Probability Weighted Moments*, estimateur des moments pondérés de Hosking et Wallis [37] (pour les paramètres de la loi GPD).

Domaines d'attraction des valeurs extrêmes

- ▷ DA(Gumbel) : domaine d'attraction de Gumbel ($\gamma = 0$; la loi des excès tend vers une loi exponentielle ; contient les lois normale, lognormale, exponentielle, gamma et de Weibull).

- ▷ DA(Fréchet) : domaine d'attraction de Fréchet ($\gamma > 0$; le support des fonctions de répartition n'est pas borné supérieurement ; contient les lois de Pareto, de Cauchy et de Student).
- ▷ DA(Weibull) : domaine d'attraction de Weibull ($\gamma < 0$; le support des fonctions de répartition est borné supérieurement ; contient la loi uniforme et les lois beta).

Quantiles ET et GPD

- ▷ $q_{1-p} := F^{\leftarrow}(1-p) = (1-F)^{\leftarrow}(p)$ quantile d'ordre $1-p$ de la loi de fonction de répartition F que l'on souhaite estimer ($F^{\leftarrow}$ étant l'inverse généralisé de F).
- ▷ $q_{ET,n}$ approximation ET (*Exponential Tail*) du quantile q_{1-p} d'ordre $1-p$.
- ▷ $q_{GPD,n}$ approximation GPD du quantile q_{1-p} d'ordre $1-p$.
- ▷ $\hat{q}_{ET,n}$ estimation ET (*Exponential Tail*) du quantile q_{1-p} d'ordre $1-p$.
- ▷ $\hat{q}_{GPD,n}$ estimation GPD du quantile q_{1-p} d'ordre $1-p$.
- ▷ $\delta_n := q_{1-p} - q_{ET,n}$ (resp., $q_{1-p} - q_{GPD,n}$) l'erreur d'approximation entre le quantile q_{1-p} d'ordre $1-p$ et son approximation ET (resp., son approximation GPD).
- ▷ $\hat{q}_{\text{param},n} := F^{-1}_{\hat{\theta}_n}(1-p)$ estimateur paramétrique (pour le modèle de fonctions de répartition F_θ) du quantile q_{1-p} d'ordre $1-p$ (on suppose ici que F_θ est inversible).
- ▷ On note avec une étoile en exposant les estimateurs calculés pour les échantillons bootstrappés : $\hat{q}^*_{\text{param},n}$, $\hat{q}^*_{ET,n}$, $\hat{q}^*_{GPD,n}$ et δ^*_n.

Bayésien

- ▷ $\Pi_\gamma(\theta)$ densité de la loi a priori sur le paramètre θ, d'hyperparamètres γ.
- ▷ $\Pi_\gamma(\theta|\, \underline{x}_n)$ densité de la loi a posteriori de θ au vu de l'échantillon $\underline{x}_n$.
- ▷ $f_\gamma(x|\, \underline{x}_n)$ densité de la loi prédictive de X au vu de l'échantillon $\underline{x}_n$.
- ▷ $q_{\max}$ un des éléments de l'avis d'expert sur la queue de distribution : valeur rarement atteinte selon l'expert, qu'il s'agit d'interpréter avec l'aide de l'expert comme un quantile d'ordre compris entre deux bornes $1-p_1$ et $1-p_2$.

Introduction

L'étude de la fréquence des événements pluviométriques extrêmes et des crues qui peuvent en résulter constitue certainement auprès du grand public l'illustration la plus frappante de l'intérêt du sujet de cette thèse : l'estimation correcte des risques fournit des éléments indispensables pour construire aux endroits critiques des digues d'une hauteur appropriée, déterminer les zones inconstructibles, définir la périodicité des opérations de nettoyage des rivières et des estuaires, afin de protéger efficacement la population et les biens. Beaucoup d'autres domaines sont concernés par l'évaluation de la fréquence d'événements rares : les sciences de l'environnement (par exemple, pour la prévision des pics de pollution), la climatologie (en particulier, l'étude de l'évolution du climat), l'industrie (notamment en fiabilité des structures), les assurances (en particulier en ré-assurance), l'analyse financière (par exemple, pour la prédiction des krachs boursiers ou des crises monétaires), etc.

Cette thèse ayant été initiée par les problèmes concrets de *Électricité de France* (EDF), qui l'a co-financée, nous nous sommes placés dans le contexte industriel de la fiabilité des structures. En effet, dans ce domaine, on cherche souvent à évaluer la probabilité d'occurrence d'événements rares, tels que charges extrêmes, défaillances de dispositifs critiques, valeurs extrêmes de ténacité, etc. Cependant, les méthodes développées peuvent aussi être appliquées plus généralement à tout type d'échantillon dont on souhaite estimer la queue de distribution.

Dans ce contexte de fiabilité des structures, les événements rares peuvent être notamment des défauts importants. Par exemple, les ingénieurs EDF peuvent être amenés à étudier des tailles de fissures importantes qui induisent un risque de rupture élevé pour les matériaux, mais qui sont rarement ou jamais observées. Il faut donc extrapoler à partir des données observées pour estimer la probabilité d'occurrence d'événements rares comme ceux-ci, c'est-à-dire la queue de distribution des défauts étudiés. Les quantiles extrêmes que nous chercherons ensuite à estimer correspondent à des tailles de défauts critiques, dépassées avec de très faibles probabilités, p, proches de 0. Les tailles de défauts supérieures à un quantile extrême étant très rares (de probabilité p), le quantile extrême correspond à une marge de sécurité avec un risque (p) quantifié et très faible.

La théorie des valeurs extrêmes (voir Embrechts *et al.*, [34] chapitre 3, ainsi que Galambos [35], chapitres 1 et 2) a été développée pour l'estimation de la probabilité d'occurrence des événements rares. Elle permet d'extrapoler le comportement de la queue de distribution des données à partir des plus grandes données observées (les données extrêmes de l'échantillon). Le théorème suivant sur la loi des valeurs extrêmes (ou EVD, *Extreme Value Distribution*) est, pour le maximum de n observations, un analogue du théorème central limite pour la moyenne. Il décrit les limites possibles de la loi du maximum de n observations, correctement normalisée à l'aide de deux suites $(\alpha_n)_{n\geq 1}$ et $(\beta_n)_{n\geq 1}$, de n variables aléatoires indépendantes et identiquement distribuées. Si F désigne la fonction de répartition de la loi commune à ces n variables aléatoires, la fonction de répartition de la loi de leur maximum est la puissance n-ième F^n de F.

Théorème 1 *Soit F la fonction de répartition de la loi d'intérêt. Sous certaines conditions de régularité sur F, il existe $\tau \in I\!R$ et deux suites normalisantes réelles $(\alpha_n)_{n\geq 1}$ et $(\beta_n)_{n\geq 1}$ ($\beta_n > 0$) tels que*

$$\forall x \in I\!R, \quad \lim_{n\to\infty} F^n\left(\alpha_n + \beta_n x\right) = H_\tau(x)\,,$$

où H_τ est la fonction de répartition de la loi des valeurs extrêmes (EVD, Extreme Value Distribution*) :*

$$H_\tau(x) := \begin{cases} \exp\left[-\left(1+\tau x\right)^{-1/\tau}\right] & \textit{pour tout } x \textit{ tel que } 1+\tau x > 0 \quad \textit{si } \tau \neq 0 \\ \exp\left(-e^{-x}\right) & \textit{pour tout } x \in I\!R \quad \textit{si } \tau = 0\,. \end{cases}$$

Lorsque $\tau \neq 0$, si $1 + \tau x \leq 0$, alors $H_\tau(x) = 0$.

On dit que la fonction de répartition F (ou la loi correspondante) est dans le domaine d'attraction de Fréchet, de Gumbel ou de Weibull selon que $\tau > 0$, $\tau = 0$ ou $\tau < 0$. On note ces domaines respectivement DA(Fréchet), DA(Gumbel) et DA(Weibull). On appelle *point terminal* d'une fonction de répartition F le réel $\omega(F)$ (fini ou infini) défini par $\omega(F) = \sup\{x : F(x) < 1\}$. On peut remarquer que pour les lois du DA(Fréchet), par exemple les lois de Student, le point terminal est infini ($\omega(F) = +\infty$) ; pour les lois du DA(Weibull), comme par exemple les lois beta, le point terminal est toujours fini ($\omega(F) < +\infty$) ; et dans le DA(Gumbel), le point terminal peut être fini ou infini, par exemple pour les lois normale, lognormale, gamma et de Weibull.

Une deuxième méthode d'estimation des queues de distribution est la **méthode des excès** (encore appelée **POT**, *Peaks Over Threshold*) introduite par de Haan et Rootzen [21]. Soit u un réel "suffisamment grand" et inférieur au point terminal ($u < \omega(F)$), appelé seuil. La méthode des excès s'appuie sur une approximation de la loi des excès au-dessus du seuil u de la variable aléatoire réelle X, c'est-à-dire de la loi conditionnelle de la variable aléatoire réelle positive $X - u$ sachant que $X > u$. La fonction de répartition des excès au-delà du seuil u est définie par

$$F_u(y) = P\left(X - u \leq y | X > u\right) = \begin{cases} \dfrac{F(u+y) - F(u)}{1 - F(u)} & \text{pour } \ y \geq 0 \\ 0 & \text{si } \ y < 0. \end{cases} \tag{1}$$

La loi asymptotique des excès est donnée par le théorème suivant, démontré par Pickands [41], ainsi que par Balkema et de Hann [1, 2, 3].

Théorème 2 (théorème de Pickands) *Si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes (DA(Fréchet), DA(Gumbel) ou DA(Weibull)), alors il existe une fonction $\sigma(u)$ positive, définie à une équivalence près*[1] *quand $u \to \omega(F)$, et un réel γ tels que*

$$\lim_{u \to \omega(F)} \sup_{0<y<\omega(F)-u} \left| F_u(y) - G_{\gamma,\,\sigma(u)}(y) \right| = 0\,, \tag{2}$$

où $G_{\gamma,\sigma}(y)$ est la fonction de répartition de la loi de Pareto généralisée *(ou* **loi GPD***,* Generalised Pareto Distribution*) définie pour $\sigma > 0$ par*

$$G_{\gamma,\sigma}(x) = \begin{cases} 1 - \left(1 + \dfrac{\gamma x}{\sigma}\right)^{-1/\gamma} & si\ \gamma \neq 0, \\ 1 - \exp(-x/\sigma) & si\ \gamma = 0, \end{cases}$$

pour $x \in I\!R_+$ si $\gamma \geq 0$, et $x \in [0,\, -\sigma/\gamma[$ si $\gamma < 0$.

Nous nous appuyons sur cette loi asymptotique des excès (la loi GPD) pour produire un estimateur d'un quantile extrême : l'estimateur ET (*Exponential Tail*) relatif à des lois appartenant au DA(Gumbel) ou plus généralement l'estimateur GPD. Nous utilisons pour cela un seuil aléatoire $\hat{u}_n = X_{(n-m_n)}$, la $(n-m_n)$-ème observation ordonnée. Le nombre d'excès m_n est choisi ; il doit tendre vers l'infini avec la taille n d'échantillon, mais rester petit devant n (notamment pour que le seuil $\hat{u}_n$ soit "suffisamment grand") :

$$\lim_{n\to\infty} m_n = +\infty \qquad \text{et} \qquad \lim_{n\to\infty} \frac{m_n}{n} = 0\,. \tag{3}$$

On peut remarquer que si on considère une suite croissante de seuils non aléatoires u_n tendant vers le point terminal, alors le nombre d'excès N_n devient une variable aléatoire qui suit la loi binomiale de paramètres n et $1 - F(u_n)$, avec cette fois

$$\lim_{n\to\infty} 1 - F(u_n) = 0 \qquad \text{et} \qquad \lim_{n\to\infty} n\,(1 - F(u_n)) = +\infty\,. \tag{4}$$

L'avantage de ces méthodes issues de la théorie des valeurs extrêmes est qu'elles sont non paramétriques. En effet, lorsque l'on utilise la méthode des excès (comme la méthode des maxima découlant du théorème 1), on s'affranchit de toute hypothèse de modèle paramétrique sur la loi des données. Cependant ces méthodes présentent un certain nombre d'inconvénients, en particulier dans un contexte industriel comme le nôtre.

Le principal problème pour l'utilisation des méthodes issues de la théorie des valeurs extrêmes provient du fait qu'en industrie (contrairement à la finance, l'hydrologie, la climatologie...) il arrive qu'on ne dispose que d'échantillons de taille petite (nous verrons des exemples

1. Si $\sigma_1(u)$ est une version de $\sigma(u)$ et $\sigma_2(u)$ est telle que $\lim_{u\to\omega(F)} \sigma_1(u)/\sigma_2(u) = 1$ alors $\sigma_2(u)$ est une autre version de $\sigma(u)$.

d'échantillons de taille de l'ordre de $n = 20$) ou moyenne (par exemple de l'ordre de $n = 100$). Les données contiennent donc peu d'information, et les excès (en nombre m_n petit devant n, en général d'ordre inférieur ou égal à $n/2$) très peu, alors que nous souhaitons utiliser des théorèmes asymptotiques. Nous avons donc eu l'idée d'utiliser l'information contenue dans l'ensemble des données, et non plus seulement dans les excès, en adaptant des modèles paramétriques estimés à partir de tout l'échantillon. Par ailleurs, ce choix possède l'avantage que les modèles paramétriques classiques sont interprétables par les ingénieurs, ou les physiciens. De plus, ces modèles sont disponibles dans les logiciels classiques et facilement réutilisables. Par exemple, lorsque les données étudiées s'intègrent ensuite dans des systèmes physiques complexes, modélisés dans des logiciels existants, il est plus simple d'introduire la modélisation de la queue de distribution des données par des modèles paramétriques classiques (connus et souvent déjà intégrés dans les différents logiciels) que par les méthodes issues de la théorie des valeurs extrêmes.

Enfin, dans certains cas industriels, il nous faut disposer d'un modèle global, en particulier de densité continue, c'est-à-dire un modèle qui produise une bonne modélisation de la probabilité d'occurrence non seulement pour les événements fréquents, mais aussi pour les événements rares. Par exemple, dans le contexte de la prévision des risques de rupture, un événement rare et contraignant peut avoir une grande influence sur la résistance des matériaux et donc sur le risque de rupture, mais de nombreux événements de peu d'influence sur la robustesse peuvent, par addition de leurs effets, avoir aussi une forte incidence sur la probabilité de rupture. Dans de tels cas, il est donc important d'estimer précisément la loi des données pour les valeurs les plus probables aussi bien que les valeurs les plus rares.

Afin de montrer la nécessité d'un modèle global, prenons un exemple plus précis en fiabilité des structures. On suppose que l'on veut étudier des tailles de fissures, qui se propagent au cours du temps dans un matériau, et qui peuvent, à partir d'une taille critique, provoquer la rupture de ce matériau. Au temps t_0, on dispose de données qui nous permettent d'estimer la distribution des tailles de défauts. Puis au cours du temps, les différentes fissures vont se propager, et certaines autres vont apparaître. Il existe des modèles de propagation pour les mécanismes de dégradation qui permettent de déterminer la taille d'un défaut à l'instant t_i à partir de sa taille mesurée à l'instant t_0. Cela permet d'obtenir la densité des tailles de défauts à l'instant t_i à partir de la densité estimée à l'instant t_0. On suppose ici que l'on a une propagation linéaire des fissures au cours du temps. Cela revient à translater, d'une certaine valeur a_i, la distribution des fissures au temps t_0 pour obtenir la distribution des fissures au temps t_i (voir la figure 1). On dispose aussi d'une distribution pour les tailles critiques de fissure, car lors d'essais en laboratoire on a constaté que différentes tailles de fissures pouvaient mener à la rupture. Au temps t_0, les ruptures sont provoquées par de rares fissures de grande taille (ainsi que de rares valeurs faibles de la taille critique). Pour prévoir la probabilité de rupture, il est donc surtout important de bien modéliser la queue de distribution des tailles de fissures. Mais lorsque la distribution des fissures se sera suffisamment translatée, des tailles de fissures de plus en plus fréquentes (mais toujours de grande taille) provoqueront aussi des ruptures (pour des valeurs critiques rares et de petite taille), les tailles de fissures les plus rares (et de tailles les plus grandes) continuant de provoquer des ruptures (mais pour des tailles critiques plus fréquentes et plus grandes), voir la figure 1. Pour prévoir les ruptures, il nous faut alors avoir bien estimé non seulement la queue de distribution des tailles de fissures, mais aussi la probabilité d'occurrence de tailles de fissures plus fréquentes, c'est-à-dire une partie plus centrale (plus proche du mode) de la distribution. Il nous faut donc adapter aux tailles de fissures au temps t_0 un modèle global, qui produise une bonne modélisation de la probabilité d'occurrence d'événements fréquents et d'événements rares. Il est de plus souhaitable que la densité de ces tailles de fissures soit continue, ou même continûment dérivable.

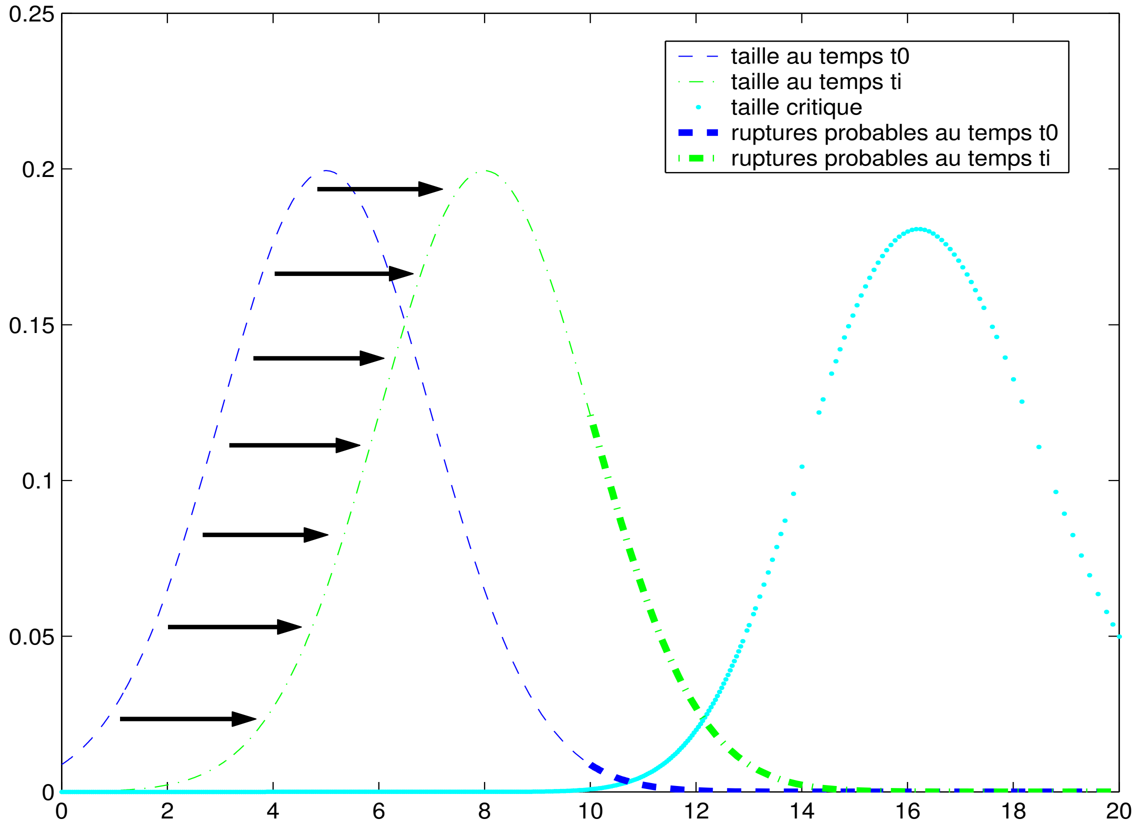

FIG. 1 – *Necessité d'un modèle global – Exemple en fiabilité des structures : Tailles de fissures ou de fractures, avec prise en compte d'une propagation simple au cours du temps – tirets : distribution des tailles au temps t_0 (en gras, la partie de la distribution pouvant provoquer des ruptures), discontinu : distribution des tailles au temps t_i (en gras, la partie de la distribution pouvant provoquer des ruptures), points : distribution des tailles critiques.*

Pour résumer, d'une part nous souhaitons pouvoir travailler sur des échantillons de petite taille pour lesquels un modèle paramétrique classique estimé à partir de tout l'échantillon permet d'utiliser toute la mince information dont on dispose ; alors que travailler, avec la méthode des excès, uniquement sur des excès en nombre trop restreint pourrait nous faire perdre de l'information. D'autre part, nous souhaitons disposer d'un modèle global de densités continues pour décrire nos données. Cela nous conduit à tenter d'utiliser des modèles paramétriques usuels pour estimer, notamment, le comportement des données en queue de distribution. Or, ces modèles sont estimés à l'aide de l'échantillon dans son ensemble, et sont sélectionnés à l'aide de tests d'adéquation classiques, comme les tests de Anderson-Darling ou de Cramér-von Mises décrits par D'Agostino et Stephens [17] (chapitre 4). Ces tests sont donc principalement influencés par les valeurs les plus probables de la variable, et ne permettent pas de juger la qualité de l'extrapolation en queue de distribution. Par conséquent, nous souhaitons introduire un test en queue de distribution pour ces modèles, afin de sélectionner des lois pour lesquelles l'erreur d'estimation des quantiles extrêmes soit réduite.

Le chapitre 1 présente donc une méthode de sélection de modèles paramétriques classiques adaptés à la queue de distribution. Cette méthode, le test ET, repose sur la comparaison en queue de distribution (plus précisément, pour un quantile extrême à déterminer) entre l'estimation donnée par un modèle paramétrique classique et celle induite par la méthode des excès. Le test ET concerne les lois du domaine d'attraction de Gumbel. Il permet en particulier de tester l'adéquation en queue de distribution des modèles normal, lognormal, exponentiel, gamma et de Weibull dont les paramètres ont été estimés par la méthode du maximum de vraisemblance.

Cette procédure laisse cependant en suspens un certain nombre de points traités dans le chapitre 2. Tout d'abord, on cherche parfois à déterminer une loi de probabilité qui s'ajuste globalement aux données, c'est-à-dire une loi qui produise une bonne modélisation à la fois des événements les plus fréquents (en partie centrale de la distribution, c'est-à-dire proche du mode) et des événements plus rares (en queue de distribution). Pour déterminer des modèles adéquats, nous disposons des tests usuels (qui sont focalisés sur les valeurs les plus probables de la variable) et du test ET pour la queue de distribution. Cependant, il se peut qu'aucun des modèles testés ne convienne pour modéliser l'ensemble de la distribution. Lorsqu'aucun modèle n'est accepté à la fois par un test usuel (dit ici "central") et un test extrême (le test ET), nous proposons une procédure de régularisation afin de produire un nouveau modèle plus adapté. Notre procédure de régularisation repose sur un soubassement bayésien dont le point de départ est un modèle accepté par un test central, et s'appuie sur un avis d'expert relatif à la queue de distribution pour modifier l'estimation des événements rares.

Nous avons implémenté le test ET et la procédure de régularisation bayésienne dans une maquette logiciel MATLAB, afin que EDF puisse expérimenter et utiliser facilement les méthodes développées. Dans un deuxième temps, nous détaillons, à travers la présentation de cette maquette, les conditions d'application de ces procédures. Enfin, le test ET n'étant applicable que dans le DA(Gumbel), nous avons souhaité étendre ce type de test à tous les domaines d'attraction des valeurs extrêmes. Nous proposons donc une généralisation du test ET, le test GPD.

Dans le chapitre 3, nous revenons à la méthode des excès pour l'estimation des queues de distribution. Lorsque l'on dispose de peu de données, les échantillons d'excès sont de très petite taille et contiennent très peu d'information. Pour utiliser la méthode des excès, il est alors souhaitable de rajouter de l'information en queue de distribution, par exemple un avis d'expert que l'on peut introduire à l'aide d'une méthode bayésienne. L'estimation des paramètres de la loi GPD étant essentielle à l'utilisation de la méthode des excès, nous proposons une procédure d'estimation bayésienne des paramètres de la loi GPD qui approxime la loi des excès. Nous avons tout d'abord appliqué cette méthode (sans avis d'expert) sur des échantillons simulés de loi GPD, puis sur des échantillons d'excès (sans et surtout avec avis d'expert sur la queue de distribution des données initiales). L'introduction d'un avis d'expert sur la queue de distribution dans le cadre de cette procédure bayésienne permet de réduire le biais des estimations de quantiles extrêmes par la méthode des excès. Cette méthode bayésienne permet en outre de disposer de la loi a posteriori du couple des paramètres, ou même du quantile extrême que l'on souhaite estimer. Ces lois a posteriori permettent de quantifier l'incertitude sur nos estimations (des paramètres ou d'un quantile extrême). Enfin, la méthode bayésienne permet de proposer une loi prédictive a posteriori, qui est la moyenne des lois GPD dont les paramètres suivent la loi a posteriori, pour reconstituer le mieux possible la véritable loi des excès.

Chapitre 1

Le test ET : test d'adéquation d'un modèle central à une queue de distribution

Étant donné un échantillon $(x_1, \ldots, x_n)$, nous voulons vérifier si un modèle paramétrique $\{F_\theta, \theta \in \Theta\}$ permet d'obtenir une bonne approximation de la loi des données, particulièrement en queue de distribution.

Les tests d'adéquation classiques (Cramér-von Mises ou Anderson-Darling, par exemple) mesurent principalement l'adéquation aux données de la partie centrale de la loi, c'est-à-dire sur une partie centrale de l'intervalle où est situé l'échantillon. Or, dans un contexte de fiabilité par exemple, il est souhaitable de pouvoir étudier des événements extrêmes. Notre but est donc de développer un test d'adéquation de la queue de distribution aux données. Lorsque le modèle recherché doit s'adapter aussi bien à l'ensemble de la distribution des données qu'à la queue de distribution (c'est-à-dire être global), on s'intéresse aux modèles acceptés par un test classique (c'est-à-dire les modèles qui induisent une bonne adéquation générale de la fonction de répartition estimée aux données). On veut de plus vérifier si la fonction de répartition correspondant au modèle testé, dont les paramètres sont estimés par maximum de vraisemblance par exemple, est bien adaptée en queue de distribution. Un test de ce type a déjà été envisagé par Diebolt et Girard [29] (chapitre 4). Ce travail s'appuie sur leurs démonstrations et en est donc une continuation.

Après quelques définitions et notations préliminaires (parties 1.1.1 et 1.1.2), nous présentons les différentes versions du test ET (une version asymptotique, partie 1.1.3, et deux versions bootstrap paramétriques, partie 1.1.4). Suivent des résultats asymptotiques sur le niveau de signification des différentes versions du test (parties 1.2.1 et 1.2.2), puis sur leur puissance (parties 1.2.3 et 1.2.4). Enfin, le test ET est appliqué à des données simulées (partie 1.3) et réelles (partie 1.4).

1.1 Présentation du test ET : test d'adéquation d'un modèle central à une queue de distribution

On dispose d'un échantillon $(x_1, \ldots, x_n)$ issu de variables aléatoires réelles, indépendantes et identiquement distribuées (i.i.d.) $X_1, \ldots, X_n$ de fonction de répartition (Fdr) F inconnue.

On considère un modèle $\{F_\theta, \theta \in \Theta\}$ dont l'adéquation générale aux données a éventuellement été acceptée par un test classique. Le test développé ici permet de vérifier l'adéquation aux données de la queue de distribution de la loi de Fdr $F_{\widehat{\theta}_n}$, dont les paramètres sont estimés, par exemple, par maximum de vraisemblance. Les hypothèses du test se ramènent donc à

$$\mathcal{H}_0 : F = F_{\widehat{\theta}_n} \quad \text{contre} \quad \mathcal{H}_1 : F \neq F_{\widehat{\theta}_n}. \tag{1.1}$$

Le test s'appuie sur la comparaison de deux estimateurs (l'un paramétrique, l'autre non paramétrique) d'un quantile extrême de la loi des données. On se limite aux modèles pour lesquels la loi des excès tend vers une loi exponentielle (d'où le nom de "ET" pour "*Exponential Tail*"), c'est-à-dire des lois appartenant au domaine d'attraction de Gumbel (DA(Gumbel)). Cette condition n'est pas trop restrictive, puisque la plupart des lois qui nous intéressent (les modèles usuels exponentiel, normal, lognormal, Weibull et gamma; particulièrement utilisés en fiabilité des structures) appartiennent au DA(Gumbel). De plus, dans le cadre d'échantillons de petite taille (l'un des contextes pour lequel nous développons ce test), nous disposons parfois de si peu d'excès qu'il est hasardeux d'estimer à partir de ceux-ci les deux paramètres d'une loi de Pareto généralisée (GPD), ainsi que de déterminer si l'on est ou non dans le DA(Gumbel). Il est alors pratique (mais invérifiable) de supposer que les excès sont issus d'une loi exponentielle dont on pourra alors mieux estimer le seul paramètre.

On commence par décrire, au paragraphe 1.1.1 (page 11), la méthode des excès pour l'estimation des quantiles extrêmes, méthode non paramétrique que nous utilisons pour construire le test. Puis on définit au paragraphe 1.1.2 (page 13) les classes de fonctions dans le DA(Gumbel) pour lesquelles nous pourrons écrire des conditions de convergence du test.

Trois versions différentes de ce test ont été construites. Tout d'abord, on présente au paragraphe 1.1.3 (page 15) une première version du test ET basée sur la loi asymptotique d'un estimateur non paramétrique d'un quantile extrême. Mais on vérifie en pratique que cette loi n'est approchée de manière satisfaisante que pour de très grandes tailles d'échantillon, alors que nous souhaitons pouvoir travailler avec peu de données. Nous proposons donc au paragraphe 1.1.4 (page 20) d'utiliser la méthode du bootstrap paramétrique pour simuler les fluctuations d'échantillonnage. La deuxième version du test ET, basée sur cette méthode de bootstrap (paragraphe 1.1.4.2 page 21), utilise le bootstrap paramétrique sur les deux estimateurs d'un quantile extrême. Enfin, cette méthode du bootstrap paramétrique pouvant se révéler lourde en calculs (par exemple pour certaines lois de mélange, et en général pour toute loi pour laquelle le calcul des estimateurs du maximum de vraisemblance est long), on propose au paragraphe 1.1.4.3 (page 21) une troisième version simplifiée du test ET qui n'applique plus le bootstrap qu'à l'estimateur non paramétrique d'un quantile extrême.

1.1.1 Méthode des excès pour l'estimation des quantiles extrêmes

Notre propos ici est l'estimation d'un quantile extrême, c'est-à-dire d'un quantile qui sera généralement situé au-delà des observations $x_1, \dots, x_n$. Choisissons un nombre p_n positif et inférieur ou égal à $1/n$, de sorte que le quantile correspondant $q_{1-p_n} := F^{\leftarrow}(1-p_n)$, d'ordre $1-p_n$, soit en général supérieur à l'observation maximale $x_{(n)}$ ($F^{\leftarrow}$ étant l'inverse généralisée de F et $x_{(n)}$ une estimation du quantile d'ordre $1-1/n$). On s'intéresse plus précisément à l'estimation non paramétrique du quantile basée sur un cas particulier de la méthode des excès, la méthode ET.

Soit u un réel suffisamment élevé appelé *seuil*. On définit les excès au-delà du seuil u comme l'ensemble de variables aléatoires $\{Y_j\} = \{X_j - u\,;\ X_j > u\}$. La fonction de répartition des excès au-delà du seuil u est

$$F_u(y) = P(X - u \leq y | X > u) = \frac{F(u+y) - F(u)}{1 - F(u)}. \tag{1.2}$$

On se limite dans ce chapitre aux lois appartenant au domaine d'attraction de Gumbel, l'un des trois domaines d'attraction de la loi des valeurs extrêmes. Ces domaines sont définis par le théorème suivant, démontré dans l'ouvrage de Embrechts *et al.* [34] (théorème 3.2.3, page 121) ainsi que dans celui de Galambos [35] (théorèmes 2.1.1, 2.1.2 et 2.1.3, pages 53 et 54). Ce théorème nous donne, après normalisation, la loi asymptotique du maximum de l'échantillon (de Fdr F^n).

Théorème 3 *Sous certaines conditions de régularité sur F, il existe $\tau \in \mathbb{R}$ et deux suites réelles $(\alpha_n)_{n\geq 1}$ et $(\beta_n)_{n\geq 1}$ ($\beta_n > 0$) tels que $\forall x \in \mathbb{R}$, $\lim_{n\to\infty} F^n(\alpha_n + \beta_n x) = H_\tau(x)$, où H_τ est la fonction de répartition de la loi des valeurs extrêmes (EVD):*

$$H_\tau(x) = \begin{cases} \exp\left[-(1+\tau x)_+^{-1/\tau}\right] & si\ \tau \neq 0, \quad où\ y_+ = \max(0, y). \\ \exp(-\exp(-x)) & si\ \tau = 0. \end{cases}$$

On dit alors que F appartient au domaine d'attraction de H_τ : $F \in DA(H_\tau)$.

Remarque 1.1 *Le domaine d'attraction de Gumbel est alors défini de la façon suivante : $DA(Gumbel) = DA(H_0)$, où $H_0(x) = \exp(-\exp(-x))$.*

Le théorème de Pickands [2, 3, 41], que nous énonçons ci-dessous dans le cas particulier du DA(Gumbel), nous permet de construire une approximation ET du quantile.

Théorème 4 (théorème de Pickands) *Si $F \in DA(Gumbel)$, il existe une fonction σ, définie à une équivalence asymptotique près*[1]*, telle que*

$$\lim_{u \to \omega(F)} \sup_{0<y<\omega(F)-u} \left| F_u(y) - \left(1 - \exp\left(-\frac{y}{\sigma(u)}\right)\right) \right| = 0, \tag{1.3}$$

où $\omega(F) = \sup\{x : F(x) < 1\}$ est le point terminal de la Fdr F.

Remarque 1.2 *On déduit de Diebolt et El-Aroui [24] (voir la démonstration du théorème 1.10, et le théorème 1.6) que l'une des valeurs possibles de $\sigma(u)$ est*

$$\sigma(u) = \frac{1}{1 - F(u)} \int_u^{\omega(F)} [1 - F(t)]\, dt.$$

Remarque 1.3 *Pour les lois du DA(Gumbel), $\omega(F)$ peut être fini ou infini. Les lois usuelles du DA(Gumbel) ont un point terminal infini (par exemple les lois normale, lognormale, de Weibull, exponentielle, gamma). On se limitera donc ici à des Fdr dont le point terminal est infini. Un exemple de loi du DA(Gumbel) ayant un point terminal fini est donné par Embrechts* et al. *[34] (voir l'exemple 3.3.22). La Fdr de cette loi, à comportement exponentiel au voisinage du point terminal fini x_F, s'exprime comme $F(x) = 1 - K \exp(-\alpha/(x_F - x))$ pour $x < x_F$, α, $K > 0$.*

À présent, on suppose que F appartient au domaine d'attraction de Gumbel; on écrit $F \in \mathrm{DA(Gumbel)}$.

Pour appliquer le théorème précédent, on choisit le seuil comme étant le quantile d'ordre $1-m_n/n$ de la loi des données : $u_n = F^{\leftarrow}(1-m_n/n)$, où m_n est un entier tel que $1 < m_n < n$ et $m_n \to \infty$ lorsque $n \to \infty$ (pour que u_n soit un seuil suffisamment grand), et $F^{\leftarrow}$ est l'inverse généralisée de la Fdr F. Pour construire une approximation ET du quantile d'ordre $1 - p_n$, on utilise la décomposition suivante (déduite de l'expression de la Fdr des excès donnée par l'équation (1.2)) :

$$p_n = \mathrm{P}(X > q_{1-p_n}) = (1 - F(u_n))(1 - F_{u_n}(q_{1-p_n} - u_n)).$$

Or par définition de u_n, on a

$$1 - F(u_n) \le \frac{m_n}{n} \le \lim_{u \to u_n,\, u > u_n} (1 - F(u)),$$

les égalités étant vérifiées lorsque $1 - m_n/n$ est un point de continuité de F. On approche donc $1 - F(u_n)$ par m_n/n, les deux quantités étant généralement égales puisque l'ensemble des points de discontinuité de F est au plus dénombrable. En outre, d'après le théorème

1. Si $\sigma_1(u)$ est une version de $\sigma(u)$ et $\sigma_2(u)$ est telle que $\lim_{u \to \omega(F)} \sigma_1(u)/\sigma_2(u) = 1$ alors $\sigma_2(u)$ est une autre version de $\sigma(u)$.

de Pickands et puisque $F \in \mathrm{DA}(\mathrm{Gumbel})$, $1 - F_{u_n}(y)$ est approchée par $\exp(-y/\sigma(u_n))$, où $y = q_{ET,n} - u_n$. Donc l'expression suivante définit $q_{ET,n}$:

$$p_n = \frac{m_n}{n} \exp\left(-\frac{q_{ET,n} - u_n}{\sigma(u_n)}\right).$$

On en déduit l'expression de l'approximation ET du quantile d'ordre $1 - p_n$ de la loi des données (voir Breiman *et al.* [10]) :

$$q_{ET,n} = u_n + \sigma(u_n) \ln\left(\frac{m_n}{np_n}\right). \tag{1.4}$$

Remarque 1.4 *Une valeur possible pour $\sigma(u_n)$ (voir Barbe et Diebolt, [4] paragraphe 3.3) est $\vartheta_n = J(m_n/ne) - J(m_n/n)$ où $J(p) = F^{\leftarrow}(1-p)$, $0 < p < 1$.*

À présent, on doit estimer u_n, le quantile d'ordre m_n/n ($> 1/n$). Il se trouve donc généralement à l'intérieur de l'intervalle défini par l'échantillon. On l'estime simplement par la $(n - m_n)$-ème observation ordonnée $\widehat{u}_n = x_{(n-m_n)}$. Quant à $\sigma(u_n)$, qui est le paramètre de la loi exponentielle qui approche la loi des excès, on l'estime naturellement par la moyenne empirique des excès :

$$\widehat{\sigma}_n = \frac{1}{m_n} \sum_{i=1}^{m_n} Y_i. \tag{1.5}$$

On obtient donc l'estimation suivante du quantile (voir Breiman *et al.* [10]) :

$$\widehat{q}_{ET,n} = X_{(n-m_n)} + \widehat{\sigma}_n \ln\left(\frac{m_n}{np_n}\right). \tag{1.6}$$

1.1.2 Classes de fonctions dans le domaine d'attraction de Gumbel

Afin de pouvoir écrire explicitement des conditions de convergence, notre étude se restreindra à des modèles dont les Fdr appartiennent à des classes définies à partir des fonctions lisses à variations régulières. Nous utiliserons les propriétés de ces fonctions décrites par Bingham *et al.* [8] (voir les parties 1.4, 1.5, 1.8, 2.3 et 2.4) et que nous rappelons dans l'annexe B.1 (page 163). Ces fonctions sont définies au voisinage de $+\infty$, donc sur un intervalle $[\,a, +\infty\,[$, $a \in I\!R$ quelconque.

Définition 1.1 (Fonctions à variations lentes) *Une fonction ℓ, positive, Lebesgue mesurable sur $[\,a, +\infty\,[$, $a \in I\!R$, est à variations lentes en $+\infty$ (noté $\ell \in \mathcal{R}_0$) si*

$$\forall t > 0, \quad \lim_{x \to \infty} \frac{\ell(tx)}{\ell(x)} = 1.$$

Définition 1.2 (Fonctions à variations régulières) *Une fonction f, positive, Lebesgue mesurable sur $[\,a, +\infty\,[$, $a \in I\!R$, est à variations régulières d'indice $\rho \in I\!R$ en $+\infty$ (noté $f \in \mathcal{R}_\rho$) si*

$$\forall t > 0, \quad \lim_{x \to \infty} \frac{f(tx)}{f(x)} = t^\rho.$$

Définition 1.3 (Fonctions lisses à variations régulières) *Une fonction f, positive, Lebesgue mesurable sur $[\,a, +\infty\,[$, $a \in I\!R$, est lisse à variations régulières d'indice $\alpha \in I\!R$ en $+\infty$ (noté $f \in \mathcal{SR}_\alpha$) si la fonction h définie par $h(x) = \ln(f(e^x))$ est $\mathcal{C}^\infty$ et*

$$h'(x) \underset{x\to\infty}{\to} \alpha, \quad h^{(i)}(x) \underset{x\to\infty}{\to} 0 \quad \forall i \geq 2.$$

On considère à présent l'ensemble $\mathcal{D}$ des fonctions de répartition F strictement croissantes, $\mathcal{C}^\infty$, et telles que $\omega(F) = \infty$.

Remarque 1.5

- *Si $F \in \mathcal{D}$, alors F est inversible (car elle est continue et strictement croissante).*
- *Si $F \in \mathcal{D}$ et si on pose $H = -\ln(1-F)$, alors H est définie sur tout $I\!R$ (car $\omega(F) = \infty$), H est strictement croissante (car F l'est), H est $\mathcal{C}^\infty$ (car F est $\mathcal{C}^\infty$) et $H(x) \to +\infty$ quand $x \to +\infty$.*

On considère le sous-ensemble de $\mathcal{D}$ suivant :

$$\mathcal{E} = \{F \in \mathcal{D} \text{ telles que } -\ln(1-F) \in \mathcal{C}^1_\rho \cup \mathcal{C}^2 \cup \mathcal{C}^3_\rho\}, \tag{1.7}$$

où les ensembles $\mathcal{C}^1_\rho$, $\mathcal{C}^2$ et $\mathcal{C}^3_\rho$ sont définis par :

- la classe des fonctions lisses à variations régulières :

$$\begin{cases} \mathcal{C}^1_\rho = \mathcal{SR}_\rho, \text{ si } \rho > 0 \text{ et } \rho \neq 1. \\ \mathcal{C}^1_1 = \{f \in \mathcal{SR}_1 \;:\; |(f^{-1})''| \in \mathcal{SR}_{-1-\tau} \text{ avec } \tau \geq 0\,;\text{ ou } f^{-1} \text{ est une fonction affine}\}. \end{cases}$$

- la classe des fonctions à variations rapides et lisses :

$$\mathcal{C}^2 = \{f,\, f^{-1} \in \mathcal{SR}_0,\, |(f^{-1})'| \in \mathcal{SR}_{-1}\}.$$

- la classe des fonctions à variations lentes et lisses :

$$\mathcal{C}^3_\rho = \{f,\, f^{-1} = \exp g,\, g \in \mathcal{SR}_\rho,\, 0 < \rho < 1\}.$$

Remarque 1.6 *Donnons des exemples de lois appartenant à chacune des classes précédentes.*

- *Appartiennent à la classe $\mathcal{C}^1_\rho$ ($\rho \neq 1$) les fonctions usuelles suivantes : la loi normale ($\rho = 2$) et la loi de Weibull ($\rho = \beta$, le paramètre de forme).*
- *Appartiennent à la classe $\mathcal{C}^1_1$ la loi exponentielle (f^{-1} est une fonction affine) et la loi gamma ($\tau = 1$). On peut donner comme exemples supplémentaires les lois de Fdr $F = 1 - \exp(-H)$ avec $H(x) \sim x/\ln x$ lorsque $x \to \infty$ (donc $H^{-1}(x) \sim x \ln x$ lorsque $x \to \infty$ et $\tau = 0$) ou $H(x) \sim c_1 x(1 + c_2 x^{-\kappa})$ lorsque $x \to \infty$ (introduit par Barbe et Diebolt, [4] page 14) où $\kappa > 0$ ($\tau = \kappa > 0$).*
- *Pour la classe $\mathcal{C}^2$, on peut donner comme exemple la loi double exponentielle telle que pour X de loi double exponentielle, $\exp X$ est de loi exponentielle.*
- *La loi lognormale, quant à elle, appartient à la classe $\mathcal{C}^3_\rho$ ($\rho = 1/2$).*

Des classes similaires ont été définies et étudiées par Diebolt et Girard [31] ainsi que par Ramdani-Worms [43, 49].

Les propriétés des fonctions H appartenant à ces trois classes sont rappelées en annexe B.1 (page 163).

Lemme 1 *$\mathcal{E}$ est inclus dans le domaine d'attraction de Gumbel.*

Ce lemme a été démontré par Diebolt et Girard [28] (cf. la démonstration de la proposition 1).

Lemme 2 *Si $H \in \mathcal{C}^1_\rho$, $H \in \mathcal{C}^2$ ou $H \in \mathcal{C}^3_\rho$, l'une des expressions possibles de $\sigma(u)$ est $\sigma(u) = 1/H'(u)$.*

Ce lemme est prouvé par Diebolt et Girard [28] (voir la démonstration du lemme 6). Dans un cadre plus général, une expression identique de $\sigma(u)$ est utilisée par Ramdani-Worms (voir par exemple [49] équation (12) ou [43]).

1.1.3 Test ET : version basée sur la loi asymptotique de $\widehat{q}_{ET,n}$, l'estimateur ET d'un quantile

Choisissons un nombre p_n positif et inférieur à $1/n$. On peut estimer les quantiles d'ordre $1 - p_n$ de la loi des données de deux façons :

Estimation paramétrique En utilisant le modèle paramétrique dont on veut tester l'adéquation en queue de distribution, on obtient l'estimateur paramétrique

$$\widehat{q}_{\mathrm{param},\,n} = F^{-1}_{\widehat{\theta}_n}(1 - p_n)\,, \tag{1.8}$$

où $\widehat{\theta}_n$ est un estimateur consistant (par exemple l'estimateur du maximum de vraisemblance s'il existe) des paramètres de la fonction de répartition F_θ de ce modèle.

Estimation non paramétrique Il s'agit de l'estimateur ET (défini par l'équation (1.6), page 13) :

$$\widehat{q}_{ET,\,n} = X_{(n-m_n)} + \widehat{\sigma}_n \ln\left(\frac{m_n}{np_n}\right).$$

L'idée est de construire un intervalle de confiance pour le quantile de la loi des données à partir de $\widehat{q}_{ET,n}$ l'estimateur ET, puis de vérifier si $\widehat{q}_{\mathrm{param},\,n}$, l'estimateur paramétrique de ce quantile appartient à cet intervalle.

Le principe de construction de l'estimateur ET du quantile d'ordre $1-p_n$ de la loi des données introduit une erreur d'approximation entre le vrai quantile q_{1-p_n} et son approximation ET, $q_{ET,n}$ (définie par l'équation (1.4), page 13), ainsi qu'une erreur d'estimation entre $q_{ET,n}$

l'approximation ET du quantile et $\widehat{q}_{ET,n}$ l'estimateur ET. On peut donc décomposer l'erreur entre le vrai quantile q_{1-p_n} et l'estimateur ET $\widehat{q}_{ET,n}$ de la façon suivante

$$q_{1-p_n} - \widehat{q}_{ET,n} = \underbrace{q_{1-p_n} - q_{ET,n}}_{\text{biais ou erreur d'approximation } \delta_n} + \underbrace{q_{ET,n} - \widehat{q}_{ET,n}}_{\text{erreur d'estimation}},$$

où d'après l'équation (1.4), page 13, et le lemme 2, page 15 (qui donne une expression de la fonction σ), l'approximation ET du quantile s'exprime comme

$$q_{ET,n} = u_n + \sigma_n \ln\left(\frac{m_n}{np_n}\right) \quad \text{avec} \quad \sigma_n = \frac{1}{H'(u_n)}. \tag{1.9}$$

Remarque 1.7 *On peut montrer (voir Diebolt et Girard [31], ainsi que Ramdani-Worms [43]) que, sous certaines conditions, l'erreur d'approximation relative δ_n/q_{1-p_n} tend vers 0.*

Concernant l'erreur d'estimation, Barbe et Diebolt [4] ont donné la loi asymptotique de $\widehat{q}_{ET,n}$ (voir aussi de Haan et Rootzen [21] et Davis et Resnick [19]), que l'on rappelle dans le théorème suivant :

Théorème 5 *Soit $F \in DA(Gumbel)$ et m_n tel que*

$$m_n \underset{n\to\infty}{\to} \infty \qquad \text{avec} \qquad m_n/n \underset{n\to\infty}{\to} 0. \tag{1.10}$$

Soit $J(u) = F^{-1}(1-u)$, $0 < u < 1$. On suppose que pour tout intervalle compact $K \subset]0,\infty[$

$$\lim_{n\to\infty} \sqrt{m_n} \sup_{x\in K} \left| \frac{J(m_n x/n) - J(m_n/n)}{J(m_n/ne) - J(m_n/n)} - \ln\left(\frac{1}{x}\right) \right| = 0. \tag{1.11}$$

Alors

$$\sqrt{m_n}\, \frac{\widehat{q}_{ET,n} - \tilde{q}_{ET,n}}{\big[J(m_n/ne) - J(m_n/n)\big] \ln(m_n/np_n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \tag{1.12}$$

avec $\tilde{q}_{ET,n} = u_n + \vartheta_n \ln(m_n/np_n)$ et $\vartheta_n = J(m_n/ne) - J(m_n/n)$.

A présent, nous souhaitons écrire des conditions explicites sur m_n pour obtenir les conditions (1.10) et (1.11) pour la convergence (1.12) de l'erreur d'estimation. Dans le cas où F appartient à la classe $(\mathcal{E})$ des fonctions que nous allons maintenant considérer, les conditions sur m_n sont données par la proposition suivante :

Proposition 1 *Soit $F \in \mathcal{E}$ et notons $H = -\ln(1-F)$. Alors les conditions (1.10) et (1.11) sont vérifiées dès que m_n vérifie les conditions suivantes :*

- *si $H \in \mathcal{C}^1_\rho$, $\rho \neq 1$, ou $H \in \mathcal{C}^2$: $m_n = o((\ln n)^2)$.*
- *si $H \in \mathcal{C}^1_1$ et H^{-1} est une fontion affine : $m_n = o(n)$.*
- *si $H \in \mathcal{C}^1_1$ avec $|(H^{-1})''| \in \mathcal{SR}_{-1-\tau}$: $\forall\delta > 0$, $m_n = O((\ln n)^{2(1+\tau)-\delta})$.*
- *si $H \in \mathcal{C}^3_\rho$: $\forall\delta > 0$, $m_n = O((\ln n)^{2(1-\rho)-\delta})$.*

La preuve de cette proposition utilise les deux lemmes suivants dont la démonstration est reportée en annexe B.2 (page 165).

Lemme 3 *On suppose que $H \in \mathcal{C}^1_\rho$ ($\rho > 0$), $H \in \mathcal{C}^2$ ou $H \in \mathcal{C}^3_\rho$ ($0 < \rho < 1$), et que $\ln x / \ln \zeta \to 0$ lorsque $\zeta \to 0$. Alors, pour $\zeta \to 0$,*

$$\frac{(H^{-1})''(\kappa_2 - \ln\zeta - \kappa_3(\kappa_2 + \kappa_1 \ln x))}{(H^{-1})'(\kappa_2 - \ln\zeta)} \sim \frac{(H^{-1})''(-\ln\zeta)}{(H^{-1})'(-\ln\zeta)}, \ \forall \kappa_1, \kappa_2, \kappa_3 \in [0,1].$$

Ce premier lemme sera utile dans plusieurs autres démonstrations.

Lemme 4 *Soient $F \in \mathcal{E}$, $H = -\ln(1-F)$ et $c_n = m_n/n$. On suppose que m_n satisfait l'une des conditions de la proposition 1 page 16. Alors,*

$$\sqrt{m_n} \frac{(H^{-1})''(-\ln c_n)}{(H^{-1})'(-\ln c_n)} \underset{n\to\infty}{\to} 0.$$

Démonstration de la proposition 1 : Pour tout $x \in K$ intervalle compact inclus dans $]0, \infty[$, on considère la quantité

$$R_n(x) = \frac{J(c_n x) - J(c_n)}{J(c_n/e) - J(c_n)} - \ln\left(\frac{1}{x}\right)$$

où $J(u) = H^{-1}(-\ln u)$ et $c_n = m_n/n$. Alors,

$$\begin{aligned}
R_n(x) &= \frac{H^{-1}(-\ln c_n - \ln x) - H^{-1}(-\ln c_n)}{H^{-1}(-\ln c_n + 1) - H^{-1}(-\ln c_n)} - \ln\left(\frac{1}{x}\right) \\
&= -\ln x \frac{(H^{-1})'(-\ln c_n - \kappa_n^{(1)} \ln x)}{(H^{-1})'(-\ln c_n + \kappa_n^{(2)})} - \ln\left(\frac{1}{x}\right) \quad \text{où } \kappa_n^{(1)},\ \kappa_n^{(2)} \in [0,1] \\
&= -\ln x \left(\frac{(H^{-1})'(-\ln c_n - \kappa_n^{(1)} \ln x) - (H^{-1})'(-\ln c_n + \kappa_n^{(2)})}{(H^{-1})'(-\ln c_n + \kappa_n^{(2)})}\right) \\
&= \ln x(\kappa_n^{(1)} \ln x + \kappa_n^{(2)}) \frac{(H^{-1})''(\kappa_n^{(2)} - \ln c_n + \kappa_n^{(3)}(\kappa_n^{(2)} + \kappa_n^{(1)} \ln x))}{(H^{-1})'(\kappa_n^{(2)} - \ln c_n)} \quad \text{où } \kappa_n^{(3)} \in [0,1] \\
&\sim \ln x(\kappa_n^{(1)} \ln x + \kappa_n^{(2)}) \frac{(H^{-1})''(-\ln c_n)}{(H^{-1})'(-\ln c_n)} \quad \text{(d'après le lemme 3 page 17).}
\end{aligned}$$

Soit C_K une constante telle que $\sup_{x\in K} |\ln x(\kappa_n^{(1)} \ln x + \kappa_n^{(2)})| \le C_K$. Il s'ensuit que

$$\sqrt{m_n} \sup_{x \in K} |R_n(x)| \le C_K \sqrt{m_n} \frac{(H^{-1})''(-\ln c_n)}{(H^{-1})'(-\ln c_n)} \underset{n\to\infty}{\to} 0,$$

d'après le lemme 4 page 17.

■

Dans le cas de fonctions de répartition appartenant à la classe $\mathcal{E}$, on peut alors énoncer un résultat de convergence pour l'erreur d'estimation.

Proposition 2 *Soit $F \in \mathcal{E}$. On suppose que m_n satisfait les conditions de la proposition 1 :*
- *si $H \in \mathcal{C}^1_\rho$, $\rho \neq 1$, ou $H \in \mathcal{C}^2$: $m_n = o((\ln n)^2)$.*
- *si $H \in \mathcal{C}^1_1$ et H^{-1} est une fontion affine : $m_n = o(n)$.*
- *si $H \in \mathcal{C}^1_1$ avec $|(H^{-1})''| \in \mathcal{SR}_{-1-\tau}$: $\forall \delta > 0$, $m_n = O((\ln n)^{2(1+\tau)-\delta})$.*
- *si $H \in \mathcal{C}^3_\rho$: $\forall \delta > 0$, $m_n = O((\ln n)^{2(1-\rho)-\delta})$.*

Alors

$$\sqrt{m_n}\,\frac{\widehat{q}_{ET,n} - q_{ET,n}}{\sigma_n \ln(m_n/np_n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{1.13}$$

Démonstration de la proposition 2 : Remarquons que, si on note $c_n = m_n/n$,

$$\begin{aligned}
\frac{\vartheta_n}{\sigma_n} - 1 &= \frac{J(c_n/e) - J(c_n)}{(H^{-1})'(-\ln c_n)} - 1 \\
&= \frac{H^{-1}(-\ln(c_n/e)) - H^{-1}(-\ln c_n)}{(H^{-1})'(-\ln c_n)} - 1 \\
&= \frac{H^{-1}(-\ln c_n + 1) - H^{-1}(-\ln c_n) - (H^{-1})'(-\ln c_n)}{(H^{-1})'(-\ln c_n)} \\
&= \frac{(H^{-1})''(-\ln c_n + \kappa_n)}{2(H^{-1})'(-\ln c_n)} \quad \text{où} \quad \kappa_n \in [0,1] \\
&\sim \frac{(H^{-1})''(-\ln c_n)}{2(H^{-1})'(-\ln c_n)} \quad \text{(d'après le lemme 3 page 17).}
\end{aligned}$$

D'après le lemme 11 (page 164), dans tous les cas, $(H^{-1})''/(H^{-1})' \in \mathcal{SR}_\alpha$ avec $\alpha < 0$, donc tend vers 0. On en déduit que $\vartheta_n/\sigma_n - 1 \to 0$ lorsque $n \to \infty$ c'est-à-dire que $\vartheta_n/\sigma_n \sim 1$. De plus, d'après la proposition 1 (page 16), on montre aussi que

$$\sqrt{m_n}\left(\frac{\vartheta_n}{\sigma_n} - 1\right) \sim \sqrt{m_n}\,\frac{(H^{-1})''(-\ln c_n)}{2(H^{-1})'(-\ln c_n)} \underset{n\to\infty}{\rightarrow} 0. \tag{1.14}$$

A présent, remarquons que

$$\sqrt{m_n}\,\frac{\widehat{q}_{ET,n} - q_{ET,n}}{\sigma_n \ln(m_n/np_n)} = \sqrt{m_n}\,\frac{\widehat{q}_{ET,n} - \tilde{q}_{ET,n}}{\sigma_n \ln(m_n/np_n)} + \sqrt{m_n}\,\frac{\tilde{q}_{ET,n} - q_{ET,n}}{\sigma_n \ln(m_n/np_n)}.$$

Les deux termes de la somme vont être étudiés séparément. Le premier terme,

$$\begin{aligned}
\sqrt{m_n}\,\frac{\widehat{q}_{ET,n} - \tilde{q}_{ET,n}}{\sigma_n \ln(m_n/np_n)} &= \sqrt{m_n}\,\frac{\widehat{q}_{ET,n} - \tilde{q}_{ET,n}}{\vartheta_n \ln(m_n/np_n)}\,\frac{\vartheta_n}{\sigma_n} \\
&\sim \sqrt{m_n}\,\frac{\widehat{q}_{ET,n} - \tilde{q}_{ET,n}}{\vartheta_n \ln(m_n/np_n)},
\end{aligned}$$

converge en loi vers $\mathcal{N}(0{,}1)$, d'après le théorème 5 (page 16). Quant au second terme,

$$\sqrt{m_n}\,\frac{\tilde{q}_{ET,n} - q_{ET,n}}{\sigma_n \ln(m_n/np_n)} = \sqrt{m_n}\,\frac{\vartheta_n - \sigma_n}{\sigma_n},$$

il tend vers 0 quand n tend vers l'infini, grâce à l'équation (1.14).
■

De la loi limite de l'estimateur ET, donnée par l'équation (1.13) de la proposition 2 (page 18), on déduit un intervalle théorique asymptotique de probabilité $1-\alpha$ pour $q_{ET,n}$

$$I_{ET,n} = \left[\widehat{q}_{ET,n} \pm \sigma_n \frac{\ln(m_n/np_n)}{\sqrt{m_n}} z_{1-\alpha/2}\right],$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée et réduite, et la notation $\pm$ pour un intervalle de confiance s'interprète ainsi : $[a \pm b] = [a-b\,;\,a+b]$. Par translation de δ_n, l'erreur d'approximation, de l'intervalle $I_{ET,n}$, on obtient un intervalle théorique asymptotique de probabilité $1-\alpha$ pour le vrai quantile q_{1-p_n} :

$$I_{th,n} = \left[\widehat{q}_{ET,n} + \delta_n \pm \sigma_n \frac{\ln(m_n/np_n)}{\sqrt{m_n}} z_{1-\alpha/2}\right]. \tag{1.15}$$

De façon théorique, on peut construire un test (ET version 0) dans lequel on rejette l'hypothèse $\mathcal{H}_0$ si $\widehat{q}_{\mathrm{param},n} \notin I_{th,n}$. Mais, puisque la vraie loi et donc la Fdr correspondante F sont inconnues, les quantités δ_n et σ_n le sont aussi. Ce test est donc inutilisable en pratique, mais il présente un intérêt théorique : déterminer la méthode permettant de montrer des résultats asymptotiques.

On peut calculer une valeur approchée (au premier ordre) d_n du biais d'approximation δ_n, qui s'exprime en fonction du type de modèle auquel appartient la loi des données (normal, lognormal, exponentiel, gamma, ou de Weibull) et de ses paramètres, du seuil u_n, de la taille de l'échantillon n, du nombre d'excès m_n, de l'ordre du quantile $1-p_n$, ainsi que de la liaison entre n, m_n et p_n (voir l'annexe B.3 page 166). Ici la vraie loi de Fdr F est inconnue, mais sous $\mathcal{H}_0$, on peut utiliser la loi de Fdr $F_{\widehat{\theta}_n}$ qui en est une bonne approximation. D'autre part, au lieu de σ_n, on utilise $\widehat{\sigma}_n$, la moyenne empirique des excès, qui est une estimation naturelle de σ_n, le paramètre d'échelle de la loi exponentielle que suivent approximativement les excès. L'intervalle de confiance pour le vrai quantile q_{1-p_n} que l'on utilise concrètement s'écrit :

$$IC_{re,n} = \left[\widehat{q}_{ET_n} + d_n \pm \widehat{\sigma}_n \frac{\ln(m_n/np_n)}{\sqrt{m_n}} z_{1-\alpha/2}\right], \tag{1.16}$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée et réduite. Le test que l'on utilise en pratique est donc le suivant :

Test ET (version 1) On rejette l'hypothèse $\mathcal{H}_0$ si $\widehat{q}_{\mathrm{param},n} \notin IC_{re,n}$.

1.1.4 Test ET : versions basées sur la méthode du bootstrap paramétrique

Au paragraphe précédent, on a utilisé la loi asymptotique de $\widehat{q}_{ET,n}$ pour construire un intervalle de confiance pour le quantile de la Fdr F. Or, nous avons remarqué sur simulations que cette loi asymptotique n'était bien approchée que pour de très grandes tailles d'échantillon, alors que nous souhaitons pouvoir travailler avec peu de données. En outre, la construction de l'intervalle de confiance $IC_{re,n}$ pour le vrai quantile q_{1-p_n} implique l'approximation des quantités σ_n et δ_n (resp. par $\widehat{\sigma}_n$ et d_n), ce qui rend difficile la maîtrise du niveau α du test ET version 1. Pour un nombre modéré de données, nous proposons d'approcher, à l'aide de la méthode du bootstrap paramétrique, la loi de l'estimateur ET $\widehat{q}_{ET,n}$ du quantile, ainsi que celle de l'estimateur paramérique $\widehat{q}_{\text{param},n}$.

1.1.4.1 Intervalle de confiance bootstrap

Sous $\mathcal{H}_0$, le modèle est $(F_\theta)_{\theta\in\Theta}$. Nous avons étudié les cas des modèles normal, lognormal, exponentiel, gamma et de Weibull, ce catalogue pouvant être aisément étendu à d'autres modèles du DA(Gumbel). Pour les modèles considérés, on estime θ par son estimateur du maximum de vraisemblance $\widehat{\theta}_n$.

On souhaite simuler les fluctuations d'échantillonnage des estimateurs ET $\widehat{q}_{ET,n}$ et paramétrique $\widehat{q}_{\text{param},n}$. Pour cela, on utilise la théorie du bootstrap paramétrique :

- On se place sous l'hypothèse $\mathcal{H}_0$ et on estime la Fdr F de la loi des données par $F_{\widehat{\theta}_n}$.
- On génère N échantillons indépendants entre eux, chaque échantillon étant i.i.d., de loi de Fdr $F_{\widehat{\theta}_n}$ et de taille n, la même que l'échantillon initial. Sous $\mathcal{H}_0$, ces N échantillons seront donc comparables à l'échantillon de départ (même taille et loi proche).
- Pour chacun de ces échantillons, on calcule l'estimateur ET associé $\widehat{q}^*_{ET,n}$, donné par l'équation (1.6) page 13. L'étoile en exposant permet par convention de différencier les quantités calculées à partir des échantillons bootstrap.
- Pour chacun de ces échantillons, on calcule un estimateur $\widehat{\theta}^*_n$ des paramètres du modèle $(F_\theta)_{\theta\in\Theta}$. Ceci nous permet d'évaluer le quantile d'ordre $1-p_n$ de la loi de Fdr $F_{\widehat{\theta}^*_n}$ donné par $\widehat{q}^*_{\text{param},n} = F^{-1}_{\widehat{\theta}^*_n}(1-p_n)$.

On dispose ainsi de N valeurs $\widehat{q}^*_{ET,n}$ de l'estimateur ET et de N valeurs $\widehat{q}^*_{\text{param},n}$ de l'estimateur paramétrique. Cela nous permet de calculer N écarts $\widehat{\delta}^*_n = \widehat{q}^*_{\text{param},n} - \widehat{q}^*_{ET,n}$ entre les deux estimateurs, correspondant à N estimations de l'erreur d'approximation $\delta_n = q_{1-p_n} - q_{ET,n}$.

De cet échantillon de $\widehat{\delta}^*_n$, on peut déduire un intervalle de confiance pour l'erreur d'approximation δ_n. Par exemple, pour un intervalle de confiance à 90%, on ordonne l'échantillon des N valeurs $\widehat{\delta}^*_n$, puis on ôte 5% des valeurs les plus grandes et 5% des plus petites. On appelle la plus petite valeur de l'échantillon restant $\widehat{\delta}^*_{\min,n}$ et la plus grande $\widehat{\delta}^*_{\max,n}$.

L'intervalle de confiance bootstrap pour l'erreur d'approximation $\delta_n = q_{1-p_n} - q_{ET,n}$ est alors

$$IC_{\delta,BP,n} = [\widehat{\delta}^*_{\min,n}, \widehat{\delta}^*_{\max,n}]. \tag{1.17}$$

1.1.4.2 Seconde version du test ET : version bootstrap paramétrique complète

À partir de l'échantillon initial, on calcule $\widehat{q}_{ET,n}$ et $\widehat{q}_{\text{param},n}$, estimateurs de q_{1-p_n}, le quantile d'ordre $1-p_n$ de F.

On construit alors un test d'adéquation en queue de distribution du modèle $(F_\theta)_{\theta\in\Theta}$ en regardant si l'erreur d'approximation estimée sur l'échantillon des données, $\widehat{\delta}_n = \widehat{q}_{\text{param},n} - \widehat{q}_{ET,n}$, appartient à l'intervalle de confiance $IC_{\delta,BP,n}$ que l'on vient de construire par la méthode du bootstrap. Si ceci est vérifié, on peut raisonnablement supposer que le modèle s'ajuste correctement en queue de distribution, tout au moins au voisinage du quantile d'ordre $1-p_n$. La version du test ET basée sur la méthode du bootstrap paramétrique est donc la suivante :

Test ET (version 2, test ET-BP complet) : On rejette l'hypothèse $\mathcal{H}_0$ si $\widehat{\delta}_n \notin IC_{\delta,BP,n}$.

1.1.4.3 Version simplifiée du test ET avec bootstrap paramétrique

Les variations d'échantillonnage de $\widehat{q}_{\text{param},n}$ (en $1/\sqrt{n}$) peuvent être considérées comme négligeables par rapport à celles de $\widehat{q}_{ET,n}$ (en $1/\sqrt{m_n}$), car m_n est petit devant n. On décide donc ici de ne plus bootstrapper l'estimation paramétrique du quantile. Cette simplification présente l'avantage de réduire notablement le temps de calcul lorsque l'estimation paramétrique du quantile est longue à calculer, par exemple pour des modèles de mélanges auxquels on peut souhaiter étendre ce type de test.

Dans ce cas, il n'est plus utile de soustraire $\widehat{q}_{\text{param},n}$ (qui reste maintenant constant) de $q_{ET,n}$ ni de $\widehat{q}_{ET,n}$, car cela revient seulement à translater $q_{ET,n}$ et son intervalle de confiance de la même façon. Il suffit de construire un intervalle de confiance pour $q_{ET,n}$, et de vérifier s'il contient ou non $\widehat{q}_{ET,n}$. Pour cela, on ordonne l'échantillon des N valeurs $\widehat{q}^*_{ET,n}$, puis (pour un intervalle à 90% par exemple) on ôte 5% des données les plus grandes et 5% des plus petites. On appelle la plus petite valeur de l'échantillon restant $\widehat{q}^*_{ET,\min,n}$ et la plus grande $\widehat{q}^*_{ET,\max,n}$. L'intervalle de confiance bootstrap (à 90% par exemple) pour $q_{ET,n}$ est alors

$$IC_{ET,BP,n} = [\widehat{q}^*_{ET,\min,n}, \widehat{q}^*_{ET,\max,n}]. \tag{1.18}$$

Il s'ensuit la dernière version du test ET :

Test ET (version 3, test ET-BP simplifié): On rejette l'hypothèse $\mathcal{H}_0$ si $\widehat{q}_{ET,n} \notin IC_{ET,BP,n}$.

Remarque 1.8 *On fait l'hypothèse que les données sont issues du modèle $(F_\theta)_{\theta\in\Theta}$. Pour le test ET-BP complet (version 2), on l'utilise bien évidemment lors du calcul des estimateurs* *paramétriques* $\widehat{q}_{param,n}$ *(calculé à partir de l'échantillon de départ) et* $\widehat{q}^*_{param,n}$ *(issus du bootstrap). Mais, en particulier pour le test ET-BP simplifié (version 3) qui ne travaille pas sur l'estimateur paramétrique, cette hypothèse intervient aussi dans le calcul des estimateurs ET* $\widehat{q}^*_{ET,n}$ *issus du bootstrap. En effet, on estime chaque* $\widehat{q}^*_{ET,n}$ *à partir d'un échantillon tiré selon la loi de Fdr* $F_{\widehat{\theta}_n}$*. On utilise donc l'hypothèse* $\mathcal{H}_0$ *lors de la simulation de ces échantillons.*

Remarque 1.9 *Supposons que les données de départ ne sont pas issues du modèle* $(F_\theta)_{\theta\in\Theta}$*, c'est-à-dire que notre hypothèse* $\mathcal{H}_0$ *est fausse. Dans ce cas, les échantillons de loi de Fdr* $F_{\widehat{\theta}_n}$ *que l'on simule ont tendance à être différents de l'échantillon de départ. Lorsque la loi de Fdr* $F_{\widehat{\theta}_n}$ *a été acceptée par un test central et qu'elle est donc considérée comme suffisamment satisfaisante pour modéliser le centre de la distribution, cette différence sera probablement peu sensible au niveau des valeurs les plus fréquentes. Par contre, cette différence sera vraisemblablement plus tangible en queue de distribution, c'est-à-dire pour les valeurs élevées de l'échantillon et au-delà, en particulier au niveau des quantiles extrêmes estimés pour le test ET. Ainsi, les* $\widehat{q}^*_{ET,n}$ *auront tendance à être, avec une forte probabilité, assez différents de* $\widehat{q}_{ET,n}$ *pour que celui-ci soit à l'extérieur de l'intervalle de confiance qu'ils permettent de construire. Ceci permet d'espérer que le test ET-BP simplifié aura une puissance raisonnable. Le test ET-BP complet permettant d'introduire plus d'information sur le modèle (puisqu'on utilise aussi la méthode du bootstrap sur l'estimateur paramétrique) que le test ET-BP simplifié, il est logique de supposer qu'il sera certainement plus puissant.*

1.2 Propriétés des différentes versions du test ET

Nous nous proposons dans cette partie d'étudier le niveau et la puissance (asymptotique et/ou à distance finie) des trois versions du test ET.

1.2.1 Niveau de signification du test ET version 1

Nous explorons tout d'abord le niveau de signification du test ET version 1 dans le cas où l'hypothèse nulle est simple, du type $\mathcal{H}_0 : F = F_0$, la Fdr F_0 étant entièrement déterminée et appartenant au domaine d'attraction de Gumbel. On suppose que $F_0 \in \mathcal{E}$ et on note $H_0 = -\ln(1-F_0)$, ainsi que :

- l'approximation ET du quantile sous $\mathcal{H}_0$: $q_{ET,n}(F_0) = u_n^{(0)} + \sigma_n^{(0)} \ln(m_n/np_n)$ avec $u_n^{(0)} = (1-F_0)^{-1}(m_n/n)$ et $\sigma_n^{(0)} = 1/H_0'(u_n^{(0)})$ (voir le lemme 2 page 15).
- la valeur du quantile sous $\mathcal{H}_0$: $q_{0,n} = F_0^{-1}(1-p_n)$.
- l'erreur d'approximation sous $\mathcal{H}_0$: $\delta_{0,n} = q_{0,n} - q_{ET,n}(F_0)$.

Malgré son intérêt uniquement théorique, nous nous proposons tout d'abord d'étudier le niveau asymptotique du test ET version 0 dans le but de déterminer dans un cas simple la méthode que nous utiliserons pour étudier le niveau asymptotique des autres versions du test.

Théorème 6 *On suppose que l'on est sous l'hypothèse $\mathcal{H}_0$ et que m_n vérifie les conditions de la proposition 1 (page 16) :*
- *si $H_0 \in \mathcal{C}^1_\rho$, $\rho \neq 1$, ou $H_0 \in \mathcal{C}^2$: $m_n = o((\ln n)^2)$.*
- *si $H_0 \in \mathcal{C}^1_1$ et H_0^{-1} est une fonction affine : $m_n = o(n)$.*
- *si $H_0 \in \mathcal{C}^1_1$ avec $|(H_0^{-1})''| \in \mathcal{SR}_{-1-\tau}$: $\forall \delta > 0,\ m_n = O((\ln n)^{2(1+\tau)-\delta})$.*
- *si $H_0 \in \mathcal{C}^3_\rho$: $\forall \delta > 0,\ m_n = O((\ln n)^{2(1-\rho)-\delta})$.*

La probabilité que la valeur du quantile sous $\mathcal{H}_0$, $q_{0,n}$, appartienne à l'intervalle

$$I_{th,n} = \left[\widehat{q}_{ET,n} + \delta_{0,n} \pm \sigma_n^{(0)} \frac{\ln(m_n/np_n)}{\sqrt{m_n}} z_{1-\alpha/2}\right],$$

intervalle théorique asymptotique de probabilité $1-\alpha$ pour le vrai quantile, tend vers $1-\alpha$ lorsque n tend vers l'infini.

Démonstration : L'erreur entre la valeur du quantile sous $\mathcal{H}_0$ et son estimation ET s'écrit

$$q_{0,n} - \widehat{q}_{ET,n} = \delta_{0,n} + q_{ET,n}(F_0) - \widehat{q}_{ET,n}.$$

Or, lorsque m_n vérifie les conditions de la proposition 1 (page 16), on a (voir proposition 2 page 18)

$$\sqrt{m_n}\, \frac{\widehat{q}_{ET,n} - q_{ET,n}(F_0)}{\sigma_n^{(0)} \ln(m_n/np_n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

On en déduit que $q_{0,n} - \widehat{q}_{ET,n} = \delta_{0,n} + \sigma_n^{(0)} \ln(m_n/np_n)/\sqrt{m_n}\, \xi_n$ où $\xi_n \xrightarrow{\mathcal{L}} \xi$, $\xi \sim \mathcal{N}(0,1)$. Posons $B_n = \ln(m_n/np_n)/\sqrt{m_n}$. On a

$$\begin{aligned} P &= \mathrm{P}\left(q_{0,n} \in \left[\widehat{q}_{ET,n} + \delta_{0,n} \pm \sigma_n^{(0)} B_n z_{1-\alpha/2}\right] \mid F = F_0\right) \\ &= \mathrm{P}\left(\widehat{q}_{ET,n} + \delta_{0,n} + \sigma_n^{(0)} B_n \xi_n \in \left[\widehat{q}_{ET,n} + \delta_{0,n} \pm \sigma_n^{(0)} B_n z_{1-\alpha/2}\right]\right) \\ &= \mathrm{P}\left(\xi_n \in \left[\pm z_{1-\alpha/2}\right]\right) \to 1-\alpha \quad \text{lorsque} \quad n \to \infty. \end{aligned}$$

■

On suppose à présent que $p_n = n^{-p'}(\ln n)^{-q'}$ et $m_n = n^{1-p}(\ln n)^{-q}$ où

$$p \leq 1 \text{ avec } \begin{cases} q \leq 0 & \text{si } p = 1 \\ q \text{ quelconque} & \text{si } p < 1 \end{cases} \quad \text{et} \quad p' \geq 1 \text{ avec } \begin{cases} q' \geq 0 & \text{si } p' = 1 \\ q' \text{ quelconque} & \text{si } p' > 1 \end{cases}$$

Ces conditions sont bien telles que $p_n \leq 1/n$ et m_n vérifie l'équation (1.10), page 16. De plus, elles permettent de contrôler l'éloignement de p_n par rapport à $1/n$ (c'est-à-dire l'éloignement du quantile au-delà de l'observation maximale $x_{(n)}$) et la croissance de m_n par rapport à n (c'est-à-dire l'éloignement du seuil $\hat{u}_n = x_{(n-m_n)}$ en deçà de l'observation maximale $x_{(n)}$). Des conditions analogues sont définies par Diebolt et Girard [28, 31]. Si on pose $a_n = -\ln(m_n/n)$ et $b_n = -\ln p_n$, ces conditions deviennent

$$a_n = p \ln n + q \ln \ln n \quad \text{et} \quad b_n = p' \ln n + q' \ln \ln n, \tag{1.19}$$

avec les mêmes restrictions sur p, p', q et q'.

Lemme 5 *Soit $F_0 \in \mathcal{E}$, $H_0 = -\ln(1-F_0)$,*

$$D_n^{(0)} = \sqrt{m_n}\,\frac{\delta_{0,n}}{\sigma_n^{(0)}\ln(m_n/np_n)} \quad et \quad C_n^{(0)} = \sqrt{m_n}\,\frac{d_{0,n}-\delta_{0,n}}{\sigma_n^{(0)}\ln(m_n/np_n)}, \tag{1.20}$$

où $d_{0,n}$ est l'approximation au premier ordre de $\delta_{0,n}$. On suppose que a_n et b_n sont donnés par l'équation (1.19). Alors, des conditions suffisantes pour que $\lim_{n\to\infty} D_n^{(0)} = 0$ sont :

- *Si $H_0 \in \mathcal{C}^1_\rho$ $(\rho \neq 1)$ ou $H_0 \in \mathcal{C}^2$: $p = p' = 1$ et $|q| < 2$.*
- *Si $H_0 \in \mathcal{C}^1_1$ et H_0^{-1} est une fonction affine : aucune condition supplémentaire sur p, p', q et q'.*
- *Si $H_0 \in \mathcal{C}^1_1$ avec $|(H_0^{-1})''| \in \mathcal{SR}_{-1-\tau}$: $p = p' = 1$ et $|q| < 2(1+\tau)$.*
- *Si $H_0 \in \mathcal{C}^3_\rho$: $p = p' = 1$ et $|q| < 2(1-\rho)$.*

De plus, sous les mêmes conditions, $\lim_{n\to\infty} C_n^{(0)} = 0$.

Démonstration : On note $\epsilon_n = (q_{0,n} - q_{ET,n}(F_0))/q_{0,n}$ l'erreur d'approximation relative, et e_n son approximation au premier ordre. On a :

$$\delta_{0,n} = q_{0,n}\epsilon_n = q_{0,n}\,e_n\epsilon_n/e_n = q_{0,n}\,e_n\,(1+o(1)) \sim q_{0,n}\,e_n.$$

Si, de plus, on remarque que $\ln(m_n/np_n) = a_n - b_n$ lorsque $n \to \infty$, on a

$$D_n^{(0)} \sim \sqrt{m_n}\,\frac{q_{0,n}\,e_n}{\sigma_n^{(0)}\ln(m_n/np_n)} = \sqrt{m_n}\,\frac{H_0^{-1}(b_n)\,e_n}{(H_0^{-1})'(a_n)(a_n-b_n)}.$$

D'autre part, on peut montrer (cf. [31] lemme 3) que

$$e_n = \frac{1}{2}(a_n-b_n)^2\frac{(H_0^{-1})''(\varrho_n)}{H_0^{-1}(b_n)} \quad \text{où} \quad \varrho_n \in [b_n,a_n].$$

Il s'ensuit que lorsque $n \to \infty$,

$$D_n^{(0)} \sim \frac{\sqrt{m_n}}{2}\,(a_n-b_n)^2\,\frac{(H_0^{-1})''(\varrho_n)}{H_0^{-1}(b_n)}\,\frac{H_0^{-1}(b_n)}{(H_0^{-1})'(a_n)}\,\frac{1}{a_n-b_n} = \frac{\sqrt{m_n}}{2}\,(a_n-b_n)\frac{(H_0^{-1})''(\varrho_n)}{(H_0^{-1})'(a_n)}.$$

Il nous faut à présent distinguer deux cas :

- Si $H_0 \in \mathcal{C}^1_1$ et H_0^{-1} est une fonction affine, $D_n^{(0)} = 0$. Il n'y a donc pas de condition supplémentaire sur p, p', q et q'.

- Sinon, on suppose $p = p' = 1$. On peut écrire $\varrho_n = a_n + \kappa(b_n - a_n)$ avec $\kappa \in [0,1]$. En choisissant $\kappa_2 = 0$, $\kappa_3 = 1$, $\kappa_1 = 1$, $\ln x = \kappa(a_n - b_n)$, et $\ln\zeta = -a_n$, on obtient que $\ln x/\ln\zeta = \kappa(b_n/a_n - 1) \to 0$ lorsque $n \to \infty$. Donc, d'après le lemme 3 (page 17),

$$D_n^{(0)} \sim \frac{\sqrt{m_n}}{2}\,(a_n-b_n)\,\frac{(H_0^{-1})''(a_n)}{(H_0^{-1})'(a_n)} = \frac{\sqrt{m_n}}{2}\,\frac{a_n-b_n}{a_n}\,\frac{a_n(H_0^{-1})''(a_n)}{(H_0^{-1})'(a_n)}.$$

- Si $H_0 \in \mathcal{C}^1_\rho$ ($\rho > 0, \rho \neq 1$) ou $H_0 \in \mathcal{C}^2$, on suppose que $|q| < 2$. D'après le lemme 10 (page 164, propriété 1) et puisque $a_n \to \infty$ quand $n \to \infty$, on a

$$\frac{1}{2}\;\frac{a_n (H_0^{-1})''(a_n)}{(H_0^{-1})'(a_n)} \underset{n\to\infty}{\to} cte\ (\neq 0).$$

On en déduit que

$$D_n^{(0)} \sim \frac{cte}{2}\;\frac{a_n - b_n}{a_n}\sqrt{m_n} \sim \frac{cte}{2}\;(q-q')(\ln\ln n)(\ln n)^{|q|/2-1}.$$

Donc, puisque l'on suppose que $|q| < 2$, $D_n^{(0)} \to 0$ lorsque $n \to \infty$.

- Si $H_0 \in \mathcal{C}^1_1$ avec $|(H_0^{-1})''| \in \mathcal{SR}_{-1-\tau}$, on suppose que $|q| < 2(1+\tau)$. Or on sait d'après le lemme 11 (page 164) que $(H_0^{-1})''/(H_0^{-1})' \in \mathcal{SR}_{-1-\tau}$. On en déduit que

$$\begin{array}{rcl} D_n^{(0)} & \sim & \sqrt{m_n}\ (a_n - b_n)\ a_n^{-1-\tau} L(a_n) \\ & \sim & (q-q')\sqrt{m_n}\ (\ln\ln n)\ a_n^{-1-\tau} L(a_n) \\ & \sim & (q-q')(\ln\ln n)(\ln n)^{|q|/2-1-\tau} L(\ln n), \end{array}$$

où L est une fonction à variations lentes. On en déduit que $D_n^{(0)} \to 0$ lorsque $n \to \infty$.

- Si $H_0 \in \mathcal{C}^3_\rho$, on suppose que $|q| < 2(1-\rho)$. Or d'après le lemme 11 (page 164, propriété 1), on a

$$\frac{x(H_0^{-1})''(x)}{(H_0^{-1})'(x)} \in \mathcal{SR}_\rho.$$

Il s'ensuit que

$$D_n^{(0)} \sim \sqrt{m_n}\frac{a_n - b_n}{a_n} a_n^\rho L(a_n) \sim (q-q')(\ln\ln n)(\ln n)^{|q|/2-1+\rho} L(a_n),$$

où L est une fonction à variations lentes, et $D_n^{(0)} \to 0$ lorsque $n \to \infty$.

D'autre part, on a : $d_{0,n} - \delta_{0,n} = q_{0,n}(e_n - \epsilon_n) = q_{0,n}\, e_n(1 - \epsilon_n/e_n) = q_{0,n}\, e_n\, o(1)$. On en déduit que $C_n^{(0)} = D_n^{(0)} o(1)$, et donc que $C_n^{(0)} \to 0$ lorsque $n \to \infty$.
■

Théorème 7 (niveau du test ET version 1 ($IC_{re,n}$))
Pour le test ET version 1, on utilise l'intervalle de confiance réel de degré de confiance $1-\alpha$ pour le quantile q_{1-p_n} défini par l'équation (1.16) :

$$IC_{re,n} = \left[\widehat{q}_{ET_n} + d_{0,n} \pm \widehat{\sigma}_n \frac{\ln(m_n/np_n)}{\sqrt{m_n}} z_{1-\alpha/2}\right].$$

Sous les conditions du lemme 5 (page 24) :

- *Si $H_0 \in \mathcal{C}^1_\rho$ ($\rho \neq 1$) ou $H_0 \in \mathcal{C}^2$: $p = p' = 1$ et $|q| < 2$,*

- *Si* $H_0 \in \mathcal{C}_1^1$ *et* H_0^{-1} *est une fonction affine : aucune condition supplémentaire,*
- *Si* $H_0 \in \mathcal{C}_1^1$ *avec* $|(H_0^{-1})''| \in \mathcal{SR}_{-1-\tau}$: $p = p' = 1$ *et* $|q| < 2(1+\tau)$,
- *Si* $H_0 \in \mathcal{C}_\rho^3$: $p = p' = 1$ *et* $|q| < 2(1-\rho)$,

le niveau du test ET version 1 tend vers α *quand* n *tend vers l'infini.*

Démonstration : Les conditions de la proposition 1 (page 16) sont impliquées par celles du lemme 5 (page 24). Donc, comme précédemment, puisque les conditions de la remarque 1 (page 16) sont vérifiées, on a la propriété $q_{0,n} - \widehat{q}_{ET,n} = \delta_{0,n} + \sigma_n^{(0)} \ln(m_n/np_n)/\sqrt{m_n}\, \xi_n$ où $\xi_n \xrightarrow{\mathcal{L}} \xi$, $\xi \sim \mathcal{N}(0,1)$.

Posons $B_n = \ln(m_n/np_n)/\sqrt{m_n}$ et $z = z_{1-\alpha/2}$. Par définition, le niveau s'exprime de la façon suivante :

$$\begin{aligned}
\text{Niveau} &= \mathrm{P}(\text{ rejeter } \mathcal{H}_0 \mid F = F_0) \\
&= \mathrm{P}\left(q_{0,n} \notin [\widehat{q}_{ET,n} + d_{0,n} \pm \widehat{\sigma}_n B_n z] \mid F = F_0\right) \\
&= \mathrm{P}\left(\widehat{q}_{ET,n} + \delta_{0,n} + \sigma_n^{(0)} B_n \xi_n \notin [\widehat{q}_{ET,n} + d_{0,n} \pm \widehat{\sigma}_n B_n z]\right) \\
&= \mathrm{P}\left(\xi_n \notin \left[C_n^{(0)} \pm \frac{\widehat{\sigma}_n}{\sigma_n^{(0)}} z\right]\right).
\end{aligned}$$

A présent, posons

$$\mathrm{Int}_n = \left[C_n^{(0)} \pm \frac{\widehat{\sigma}_n}{\sigma_n^{(0)}} z\right], \tag{1.21}$$

et montrons que $\mathrm{P}(\xi_n \in \mathrm{Int}_n) \to 1 - \alpha$ lorsque $n \to \infty$. On sait (voir Davis et Resnick [19]) que lorsque $m_n \to \infty$ et $m_n/n \to 0$ quand $n \to \infty$, on a

$$\frac{\widehat{\sigma}_n}{\sigma_n^{(0)}} \xrightarrow{P} 1.$$

Ceci signifie que

$$\forall \eta > 0,\ \mathrm{P}\left(\left|\frac{\widehat{\sigma}_n}{\sigma_n^{(0)}} - 1\right| > \eta\right) \underset{n\to\infty}{\to} 0.$$

Remarquons que

$$\mathrm{P}\left(\left|\frac{\widehat{\sigma}_n}{\sigma_n^{(0)}} - 1\right| > \eta\right) = \mathrm{P}\left(\frac{\widehat{\sigma}_n}{\sigma_n^{(0)}} - 1 > \eta\right) + \mathrm{P}\left(1 - \frac{\widehat{\sigma}_n}{\sigma_n^{(0)}} > \eta\right).$$

La probabilité du membre de gauche tendant vers zéro, on en déduit que les probabilités à droite de l'égalité tendent vers zéro. On en déduit que

$$\mathrm{P}\left(\frac{\widehat{\sigma}_n}{\sigma_n^{(0)}} > 1 + \eta\right) \underset{n\to\infty}{\to} 0 \quad \text{et} \quad \mathrm{P}\left(\frac{\widehat{\sigma}_n}{\sigma_n^{(0)}} < 1 - \eta\right) \underset{n\to\infty}{\to} 0.$$

Avec la notation $I_{1,n} = [C_n^{(0)} \pm z(1+\eta)]$, on en déduit que $\forall \eta > 0$,

$$\begin{aligned} \mathrm{P}\left(\mathrm{Int}_n \subset I_{1,n}\right) &= \mathrm{P}\left(\left[\pm \frac{\widehat{\sigma}_n}{\sigma_n^{(0)}}\right] \subset [\pm(1+\eta)]\right) = \mathrm{P}\left(\frac{\widehat{\sigma}_n}{\sigma_n^{(0)}} \leq 1+\eta\right) \\ &= 1 - \mathrm{P}\left(\frac{\widehat{\sigma}_n}{\sigma_n^{(0)}} > 1+\eta\right) \underset{n\to\infty}{\to} 1. \end{aligned}$$

D'autre part, puisque sous les conditions du lemme 5 (page 24), $C_n^{(0)} \to 0$ quand $n \to \infty$, on peut écrire :

$$\forall c > \eta,\ \ \exists N_0(c) \in I\!N,\ \ \forall n \geq N_0(c),\ \ I_{1,n} \subset [\pm z(1+c)].$$

Si on prend $c = 2\eta$, on a : $\forall n \geq N_0(2\eta)$,

$$\mathrm{P}\left(\mathrm{Int}_n \subset I_{1,n}\right) \leq \mathrm{P}\left(\mathrm{Int}_n \subset [\pm z(1+2\eta)]\right).$$

Donc $\mathrm{P}(\mathrm{Int}_n \subset [\pm z(1+2\eta)]) \to 1$ lorsque $n \to \infty$, ce qui signifie que

$$\forall \nu_1 > 0,\ \exists N_1(\nu_1,\eta) \in I\!N \ \text{ tel que } \forall n \geq N_1(\nu_1,\eta),\ \ \mathrm{P}\left(\mathrm{Int}_n \not\subset [\pm z(1+2\eta)]\right) < \nu_1.$$

Alors, on a : $\forall n \geq N_1(\nu_1,\eta)$

$$\begin{aligned} \mathrm{P}\left(\xi_n \in \mathrm{Int}_n\right) &= \mathrm{P}\left((\xi_n \in \mathrm{Int}_n) \cap (\mathrm{Int}_n \subset [\pm z(1+2\eta)])\right) \\ &+ \mathrm{P}\left((\xi_n \in \mathrm{Int}_n) \cap (\mathrm{Int}_n \not\subset [\pm z(1+2\eta)])\right) \\ &\leq \mathrm{P}\left(\xi_n \in [\pm z(1+2\eta)]\right) + \mathrm{P}\left(\mathrm{Int}_n \not\subset [\pm z(1+2\eta)]\right) \\ &\leq \mathrm{P}\left(\xi_n \in [\pm z(1+2\eta)]\right) + \nu_1. \end{aligned} \tag{1.22}$$

D'autre part, on sait que $\forall \nu_2 > 0$, il existe $C(\nu_2) > 0$ et $N_2(\nu_2)$ tels que

$$\forall \eta \in]0, C(\nu_2)] \ \text{ et } \ \forall n \geq N_2(\nu_2),\ \ |\mathrm{P}(\xi_n \in [\pm z(1+2\eta)]) - (1-\alpha)| < \nu_2,$$

ce qui implique que

$$\forall \nu_2 > 0,\ \ \forall \eta \in]0, C(\nu_2)],\ \ \limsup_{n\to\infty} \mathrm{P}(\xi_n \in [\pm z(1+2\eta)]) \leq 1 - \alpha + \nu_2.$$

Puisque $\mathrm{P}(\xi_n \in \mathrm{Int}_n) \leq \mathrm{P}(\xi_n \in [\pm z(1+2\eta)]) + \nu_1$ pour $n \geq N_1(\nu_1,\eta)$ (voir l'équation (1.22) page 27), on en déduit que $\forall \nu_1,\ \nu_2 > 0$

$$\limsup_{n\to\infty} \mathrm{P}\left(\xi_n \in \mathrm{Int}_n\right) \leq 1 - \alpha + \nu_1 + \nu_2.$$

Avec le même type de raisonnement, on peut montrer que $\forall \nu_1,\ \nu_2 > 0$,

$$\liminf_{n\to\infty} \mathrm{P}\left(\xi_n \in \mathrm{Int}_n\right) \geq 1 - \alpha - \nu_1 - \nu_2.$$

Il s'ensuit que $\forall \nu_1,\ \nu_2 > 0$,

$$1 - \alpha - \nu_1 - \nu_2 \leq \liminf_{n\to\infty} \mathrm{P}(\xi_n \in \mathrm{Int}_n) \leq \limsup_{n\to\infty} \mathrm{P}(\xi_n \in \mathrm{Int}_n) \leq 1 - \alpha + \nu_1 + \nu_2.$$

Comme ν_1 et ν_2 sont arbitrairement petits, on en déduit que la limite inférieure et la limite supérieure de $\mathrm{P}(\xi_n \in \mathrm{Int}_n)$ sont égales toutes les deux à $1 - \alpha$. Donc la limite existe, et vaut $1 - \alpha$.

■

1.2.2 Niveau de signification des tests ET-BP complet et simplifié

Dans ce paragraphe, nous voulons étudier à distance finie, c'est-à-dire pour n fixé, le niveau de signification des deux tests ET-BP. Malheureusement, il n'existe pas, à notre connaissance, d'éléments permettant de mener cette étude complètement. Pour obtenir malgré tout une idée approximative des résultats auxquels on peut s'attendre, nous proposons de démontrer deux énoncés en nous plaçant dans un cadre simplifié : au lieu de travailler avec des échantillons bootstrap issus de $F_{\widehat{\theta}_n}$, nous allons supposer que ces échantillons bootstrap sont issus de la vraie loi sous $\mathcal{H}_0$. En d'autres termes, nous nous limitons au cas où l'hypothèse nulle est simple, $\mathcal{H}_0 : F = F_0$, la fonction F_0 étant supposée entièrement déterminée, et appartenant au domaine d'attraction de Gumbel. Il s'agit alors de simulation par Monte-Carlo.

Proposition 3 (niveau du test ET-BP complet)
On note H_n la Fdr de la variable aléatoire $\widehat{\delta}_n = q_{1-p_n}(\widehat{\theta}_n) - \widehat{q}_{ET,n}$ (estimation de l'erreur d'approximation δ_n sous $\mathcal{H}_0$), où $q_{1-p_n}(\widehat{\theta}_n)$ est l'estimation paramétriquesous $\mathcal{H}_0$, et $\widehat{q}_{ET,n}$ l'estimation ET du quantile d'ordre $1 - p_n$. On suppose que pour tout n, la fonction H_n est continue et strictement croissante, donc inversible. Alors, quelle que soit la taille n de l'échantillon, le niveau de signification du test ET-BP complet, pour l'hypothèse nulle $\mathcal{H}_0$: $F = F_0$, tend vers α lorsque le nombre N d'échantillons bootstrap tend vers l'infini.

Proposition 4 (niveau du test ET-BB simplifié)
On note G_n la Fdr de la variable $\widehat{q}_{ET,n}$ et on suppose que pour tout n cette fonction est continue et strictement croissante, donc inversible. Alors quelle que soit la taille n de l'échantillon, le niveau de signification du test ET-BP simplifié, pour l'hypothèse nulle $\mathcal{H}_0 : F = F_0$, tend vers α lorsque le nombre N d'échantillons bootstrap tend vers l'infini.

Pour simplifier la présentation, nous commençons par la démonstration la plus simple, celle de la proposition 4.

Démonstration de la proposition 4 : On suppose que l'échantillon initial de variables aléatoires $\mathbf{X} = (X_1, \ldots, X_n)$ est de loi de Fdr F_0. Puis on tire N échantillons indépendants $\mathbf{X}^{*(j)} = (X_1^{*(j)}, \ldots, X_n^{*(j)})$, chacun étant i.i.d de loi de Fdr F_0. Par conséquent, les N variables aléatoires $\widehat{q}_{ET,n}^{*(j)}$, $1 \leq j \leq N$ obtenues à partir de ces N échantillons bootstrap sont i.i.d, de loi commune de Fdr G_n. Il s'ensuit que les quantiles empiriques associés (relatifs aux réplications bootstrap) d'ordre $\alpha/2$ et $1 - \alpha/2$, respectivement, convergent presque sûrement lorsque $N \to \infty$ vers les quantiles correspondants de la Fdr G_n. De plus, on note $\mathbb{X} = (\mathbf{X}^{*(1)}, \ldots, \mathbf{X}^{*(N)})$ l'ensemble des échantillons bootstrap. Par suite, $\forall\ n$ fixé,

$$\begin{aligned}
\text{Niveau} &= \mathrm{P}(\text{ rejeter } \mathcal{H}_0 |\, F = F_0) \\
&= \mathrm{P}(\widehat{q}_{ET,n} \notin [\widehat{q}^{*}_{ET,\min,n}, \widehat{q}^{*}_{ET,\max,n}] |\, F = F_0) \\
&= \mathrm{E}[\mathrm{I}(\widehat{q}_{ET,n} \notin [\widehat{q}^{*}_{ET,\min,n}, \widehat{q}^{*}_{ET,\max,n}]) |\, F = F_0] \\
&= \mathrm{E}\{\mathrm{E}[\mathrm{I}(\widehat{q}_{ET,n} \notin [\widehat{q}^{*}_{ET,\min,n}, \widehat{q}^{*}_{ET,\max,n}]) |\, \mathbb{X}] |\, F = F_0\} \\
&= \mathrm{E}[\mathrm{P}(\widehat{q}_{ET,n} \notin [\widehat{q}^{*}_{ET,\min,n}, \widehat{q}^{*}_{ET,\max,n}] |\, \mathbb{X}) |\, F = F_0] \qquad (1.23) \\
&\to \alpha \quad \text{quand} \quad N \to \infty
\end{aligned}$$

d'après le théorème de convergence dominée, car

$$\mathrm{P}(\widehat{q}_{ET,n} \notin [\widehat{q}^*_{ET,\min,n}, \widehat{q}^*_{ET,\max,n}] | \mathbb{X}) \xrightarrow{p.s.} \alpha \quad \text{lorsque} \quad N \to \infty, \tag{1.24}$$

et puisque ces probabilités conditionnelles restent entre 0 et 1.
■

Démonstration de la proposition 3 : Elle est essentiellement analogue à la démonstration précédente. Mais cette fois-ci, on calcule l'estimateur $\widehat{\theta}_n^{*(j)}$ de θ calculé à partir de $\mathbf{X}^{*(j)}$, $1 \le j \le N$, et on forme

$$\widehat{\delta}_n^{*(j)} = q_{1-p_n}(\widehat{\theta}_n^{*(j)}) - \widehat{q}^{*(j)}_{ET,n}, \;\; 1 \le j \le N.$$

Comme les $\widehat{\delta}_n^{*(j)}$ $1 \le j \le N$, sont indépendants et identiquement distribués, de loi commune de Fdr H_n, on conclut comme ci-dessus.
■

Ces résultats et leurs démonstrations permettent de mieux comprendre pourquoi on peut espérer que le niveau des tests ET-BP soit à distance finie, c'est-à-dire pour une taille d'échantillon n fixée, très proche du niveau cherché α.

La situation est beaucoup plus compliquée dans le cas du vrai bootstrap paramétrique, lorsque les $\mathbf{X}^{*(j)}$ sont simulés selon la loi de Fdr $F_{\widehat{\theta}_n}$ (dont les paramètres sont estimés) et non plus la loi de Fdr F_0 (entièrement connue). L'égalité (1.23), page 28, reste vraie, mais la convergence (1.24), page 29, ne l'est plus. En effet, à n fixé, puisque les $X_i^{*(j)}$ sont issus de la loi de Fdr $F_{\widehat{\theta}_n}$, et non de la loi de Fdr F_0, les quantiles $\widehat{q}^*_{ET,\min,n}$ et $\widehat{q}^*_{ET,\max,n}$ convergent presque sûrement quand $N \to \infty$ vers les quantiles correspondants de la loi de Fdr G_n^* au lieu de G_n, où G_n^* est la Fdr de la loi commune des $\widehat{q}^{*(j)}_{ET,n}$ (c'est-à-dire des estimateurs ET du quantile d'ordre $1-p_n$ calculés à partir d'échantillons de loi de Fdr $F_{\widehat{\theta}_n}$ et non F_0). De même, les quantiles $\widehat{\delta}^*_{\min,n}$ et $\widehat{\delta}^*_{\max,n}$ convergent presque sûrement quand $N \to \infty$ vers les quantiles correspondants de la loi de Fdr H_n^* au lieu de H_n, où H_n^* est la Fdr de la loi commune des $\widehat{\delta}_n^{*(j)}$ (c'est-à-dire des estimateurs de l'erreur d'approximation calculés à partir d'échantillons de loi de Fdr $F_{\widehat{\theta}_n}$ et non F_0).

Pour pouvoir conclure, il faudrait pouvoir évaluer précisément l'écart entre les quantiles extrêmes de G_n (resp. H_n) et ceux de G_n^* (resp. H_n^*), donc connaître précisément l'écart entre G_n (resp. H_n) et G_n^* (resp. H_n^*), en particulier en queue de distribution. Même dans le cas ET-BP simple, on a

$$\widehat{q}^{*(j)}_{ET,n} = X^{*(j)}_{n-m_n} + \left(\frac{1}{m_n} \sum_{k=1}^{m_n} Y_k^{*(j)} \right) \ln \left(\frac{m_n}{np_n} \right),$$

où les $Y_k^{*(j)}$ sont les excès au-dessus du seuil $X^{*(j)}_{(n-m_n)}$ de l'échantillon $\mathbf{X}^{*(j)}$, les variables aléatoires $X_i^{*(j)}$ étant indépendantes et identiquement distribuées de loi de Fdr $F_{\widehat{\theta}_n}$. Par conséquent, chaque $Y_k^{*(j)}$ admet pour fonction de répartition

$$K_n^*(y) = \frac{F_{\widehat{\theta}_n}(X_{n-m_n}^{*(j)} + y) - F_{\widehat{\theta}_n}(X_{n-m_n}^{*(j)})}{1 - F_{\widehat{\theta}_n}(X_{n-m_n}^{*(j)})}$$

au lieu de la fonction de répartition

$$K_n(y) = \frac{F_0(X_{n-m_n} + y) - F_0(X_{n-m_n})}{1 - F_0(X_{n-m_n})}$$

pour les excès initiaux Y_j.

L'étude de la différence $K_n^*(y) - K_n(y)$ paraît insurmontable, et, finalement, ce sont les simulations présentées dans la partie 1.3 (page 38) qui indiquent le mieux la qualité de l'approximation du niveau α à distance finie.

1.2.3 Puissance du test ET version 1

Nous allons à présent explorer la puissance du test ET version 1 dans le cas où l'hypothèse nulle est simple, $\mathcal{H}_0 : F = F_0$. On note F_1 la vraie Fdr. Nous allons montrer que la puissance en $F = F_1$ tend vers 1. Les fonctions F_0 et F_1 sont entièrement déterminées et appartiennent au DA(Gumbel). On suppose que $F_0 \in \mathcal{E}$ et $F_1 \in \mathcal{E}$, et on note $H_0 = -\ln(1 - F_0)$ et $H_1 = -\ln(1 - F_1)$. Pour $i \in \{0,1\}$, on introduit :

- l'approximation ET du quantile sous $\mathcal{H}_i$, soit $q_{ET,n}(F_i) = u_n^{(i)} + \sigma_n^{(i)} \ln(m_n/np_n)$ avec $u_n^{(i)} = (1 - F_i)^{-1}(m_n/n)$ et $\sigma_n^{(i)} = 1/H_i'(u_n^{(i)})$;
- la valeur du quantile sous $\mathcal{H}_i$, soit $q_{i,n} = F_i^{-1}(1 - p_n)$.

Posons

$$\rho_n = \sqrt{m_n}\, \frac{q_{ET,n}(F_1) - q_{ET,n}(F_0)}{\sigma_n^{(1)} \ln(m_n/np_n)} \quad \text{et} \quad \zeta_n = \frac{\sigma_n^{(0)}}{\sigma_n^{(1)}}. \tag{1.25}$$

On définit également

$$K_i(a_n) = \frac{a_n (H_i^{-1})'(a_n)}{H_i^{-1}(a_n)} \quad \text{pour } i \in \{0,1\} \quad \text{et} \quad K_{01}(a_n) = \frac{H_0^{-1}(a_n)}{H_1^{-1}(a_n)}, \tag{1.26}$$

où $a_n = -\ln(m_n/n)$ est défini par l'équation (1.19) page 23, ainsi que $b_n = -\ln p_n$.

Lemme 6 *On a*

$$\rho_n = \sqrt{m_n}\, \frac{K_{01}(a_n)}{K_1(a_n)} \left[\frac{a_n}{b_n - a_n} \left(\frac{1}{K_{01}(a_n)} - 1 \right) + \frac{K_1(a_n)}{K_{01}(a_n)} - K_0(a_n) \right], \tag{1.27}$$

et

$$\zeta_n = \frac{K_{01}(a_n)}{K_1(a_n)} K_0(a_n)\,. \tag{1.28}$$

Démonstration : On remarque que, avec la notation $c_n = m_n/n$,

$$\begin{aligned}
\rho_n &= \frac{\sqrt{m_n}\left[(1-F_1)^{-1}(c_n) - (1-F_0)^{-1}(c_n) + (\sigma_n^{(1)} - \sigma_n^{(0)})\ln(c_n/p_n)\right]}{\sigma_n^{(1)}\ln(c_n/p_n)} \\
&= \sqrt{m_n}\frac{H_1^{-1}(a_n) - H_0^{-1}(a_n) + \left[(H_1^{-1})'(a_n) - (H_0^{-1})'(a_n)\right](b_n - a_n)}{(H_1^{-1})'(a_n)(b_n - a_n)} \\
&= \sqrt{m_n}\left[\frac{H_1^{-1}(a_n) - H_0^{-1}(a_n)}{(H_1^{-1})'(a_n)(b_n - a_n)} + 1 - \frac{(H_0^{-1})'(a_n)}{(H_1^{-1})'(a_n)}\right] \\
&= \sqrt{m_n}\left[\left(\frac{1}{K_1(a_n)} - \frac{K_{01}(a_n)}{K_1(a_n)}\right)\frac{a_n}{b_n - a_n} + 1 - \frac{K_{01}(a_n)K_0(a_n)}{K_1(a_n)}\right].
\end{aligned}$$

Pour ζ_n, on a

$$\zeta_n = \frac{\sigma_n^{(0)}}{\sigma_n^{(1)}} = \frac{(H_0^{-1})'(a_n)}{(H_1^{-1})'(a_n)} = \frac{K_{01}(a_n)}{K_1(a_n)}K_0(a_n).$$

■

Lemme 7 *On considère les cas où les fonctions H_0 et H_1 appartiennent aux classes suivantes:*

- $H_0 \in \mathcal{C}^1_{\beta_0}$ *et* $H_1 \in \mathcal{C}^1_{\beta_1}$ *avec* $0 < \beta_0 < \beta_1$;
- $H_0 \in \mathcal{C}^3_{\rho}$ *et* $H_1 \in \mathcal{C}^1_{\beta}$ *avec* $\beta > 0$ *et* $0 < \rho < 1$;
- $H_0 \in \mathcal{C}^1_{\beta}$ *et* $H_1 \in \mathcal{C}^2$ *avec* $\beta > 0$;
- $H_0 \in \mathcal{C}^3_{\rho}$ *et* $H_1 \in \mathcal{C}^2$ *avec* $0 < \rho < 1$;
- $H_0 \in \mathcal{C}^3_{\rho_0}$ *et* $H_1 \in \mathcal{C}^3_{\rho_1}$ *avec* $0 < \rho_1 < \rho_0 < 1$.

Alors $\rho_n \to -\infty$ *et* $\zeta_n \to +\infty$ *lorsque* $n \to \infty$, *avec* $\zeta_n = o(|\rho_n|)$.

Les conditions du lemme 7 sélectionnent les cas où F_0 est à queue plus lourde que F_1. Les autres cas sont étudiés dans le lemme 8.

Lemme 8 *On considère les cas où les fonctions H_0 et H_1 appartiennent aux classes suivantes :*

- $H_0 \in \mathcal{C}^1_{\beta_0}$ *et* $H_1 \in \mathcal{C}^1_{\beta_1}$ *avec* $0 < \beta_1 < \beta_0$;
- $H_0 \in \mathcal{C}^1_{\beta}$ *et* $H_1 \in \mathcal{C}^3_{\rho}$ *avec* $\beta > 0$ *et* $0 < \rho < 1$;
- $H_0 \in \mathcal{C}^2$ *et* $H_1 \in \mathcal{C}^1_{\beta}$ *avec* $\beta > 0$;
- $H_0 \in \mathcal{C}^2$ *et* $H_1 \in \mathcal{C}^3_{\rho}$ *avec* $0 < \rho < 1$;
- $H_0 \in \mathcal{C}^3_{\rho_0}$ *et* $H_1 \in \mathcal{C}^3_{\rho_1}$ *avec* $0 < \rho_0 < \rho_1 < 1$.

Alors $\rho_n \to +\infty$ *et* $\zeta_n \to 0$ *lorsque* $n \to \infty$.

On démontre ci-dessous le lemme 7. La démonstration du lemme 8 est analogue.

Démonstration du lemme 7.

- Supposons que $H_0 \in \mathcal{C}^1_{\beta_0}$ et $H_1 \in \mathcal{C}^1_{\beta_1}$ avec $0 < \beta_0 < \beta_1$. D'après le lemme 10 (page 164), lorsque $n \to \infty$, $K_0(a_n) \sim 1/\beta_0$ et $K_1(a_n) \sim 1/\beta_1$. D'autre part, $K_{01} \in SR_{1/\beta_0 - 1/\beta_1}$ où $1/\beta_0 - 1/\beta_1 > 0$. On en déduit que $K_{01}(a_n) \to \infty$. Par conséquent, on a
$$\rho_n \sim -\sqrt{m_n}\beta_1 K_{01}(a_n)\left[\frac{a_n}{b_n - a_n} + \frac{1}{\beta_0}\right] \underset{n\to\infty}{\to} -\infty \quad \text{et} \quad \zeta_n \sim \frac{\beta_1}{\beta_0}K_{01}(a_n) \underset{n\to\infty}{\to} +\infty,$$
avec $\zeta_n = o(|\rho_n|)$.

- Supposons que $H_0 \in \mathcal{C}^3_\rho$ et $H_1 \in \mathcal{C}^1_\beta$ avec $\beta > 0$ et $0 < \rho < 1$. D'après le lemme 10 (page 164), lorsque $n \to \infty$, $K_1(a_n) \sim 1/\beta$ et $K_0(x) = xg'(x) \in SR_\rho$, donc $K_0(x) \to +\infty$ quand $x \to \infty$. Enfin, on a $K_{01}(a_n) = \exp(g(a_n))/H_1^{-1}(a_n) \to +\infty$. Il s'ensuit que
$$\rho_n \sim -\beta\sqrt{m_n}K_{01}(a_n)K_0(a_n)\left[\frac{a_n}{b_n - a_n}\frac{1}{K_0(a_n)} + 1\right] \underset{n\to\infty}{\to} -\infty$$
et
$$\zeta_n \sim \beta K_0(a_n)K_{01}(a_n) \underset{n\to\infty}{\to} +\infty \quad \text{avec} \quad \zeta_n = o(|\rho_n|).$$

- Supposons que $H_0 \in \mathcal{C}^1_\beta$ et $H_1 \in \mathcal{C}^2$ avec $\beta > 0$. D'après le lemme 10 (page 164), lorsque $n \to \infty$, $K_0(a_n) \sim 1/\beta$ et $K_1(x) \to 0$ lorsque $x \to \infty$. Enfin, on a $K_{01}(a_n) = H_0^{-1}(a_n)/H_1^{-1}(a_n) \to +\infty$ car K_{01} appartient à $SR_{1/\beta}$. On en déduit que
$$\rho_n \sim -\sqrt{m_n}\frac{K_{01}(a_n)}{K_1(a_n)}\left[\frac{a_n}{b_n - a_n} + \frac{1}{\beta}\right] \underset{n\to\infty}{\to} -\infty \quad \text{et} \quad \zeta_n \sim \frac{K_{01}(a_n)}{\beta\ K_1(a_n)} \underset{n\to\infty}{\to} +\infty,$$
avec $\zeta_n = o(|\rho_n|)$.

- Supposons que $H_0 \in \mathcal{C}^3_\rho$ et $H_1 \in \mathcal{C}^2$ avec $0 < \rho < 1$. On a déjà vu que $H_0 \in \mathcal{C}^3_\rho$ implique que $K_0(a_n) \to +\infty$ et $H_1 \in \mathcal{C}^2$ implique que $K_1(x) \to 0$ lorsque $x \to \infty$. D'autre part, $K_{01}(x) = \exp(g(x))/H_1^{-1}(x)$ où $g \in SR_\rho$ et $H_1^{-1} \in SR_0$. On en déduit que $K_{01}(a_n) \to +\infty$. D'où on a
$$\rho_n \sim -\sqrt{m_n}\frac{K_{01}(a_n)}{K_1(a_n)}\left[\frac{a_n}{b_n - a_n} + K_0(a_n)\right] \underset{n\to\infty}{\to} -\infty$$
et
$$\zeta_n \sim \frac{K_{01}(a_n)K_0(a_n)}{K_1(a_n)} \underset{n\to\infty}{\to} +\infty \quad \text{avec} \quad \zeta_n = o(|\rho_n|).$$

- Supposons que $H_0 \in \mathcal{C}^3_{\rho_0}$ et $H_1 \in \mathcal{C}^3_{\rho_1}$ avec $0 < \rho_1 < \rho_0 < 1$. On a déjà vu que $H_i \in \mathcal{C}^3_{\rho_i}$ signifie que $H_i^{-1}(x) = \exp(g_i(x))$ où $g_i \in SR_{\rho_i}$ $(i \in \{0,1\})$. On en déduit que $K_i(x) = xg_i'(x) \in SR_{\rho_i} \to +\infty$ quand $x \to \infty$. D'autre part, $K_{01}(x) = \exp(g_0(x) - g_1(x))$, où $g_0 - g_1 \in SR_\rho$ avec $\rho = \max(\rho_0,\rho_1) = \rho_0 > 0$, implique que $K_{01}(a_n) \to +\infty$. De plus, $K_0(a_n)/K_1(a_n) \in SR_{\rho_0 - \rho_1} \to +\infty$. D'où $K_{01}(a_n)K_0(a_n)/K_1(a_n) \to +\infty$. Il s'ensuit que
$$\rho_n \sim -\sqrt{m_n}\frac{K_{01}(a_n)K_0(a_n)}{K_1(a_n)}\left[\frac{a_n}{b_n - a_n}\frac{1}{K_0(a_n)} + 1\right] \underset{n\to\infty}{\to} -\infty$$

et
$$\zeta_n = \frac{K_{01}(a_n)K_0(a_n)}{K_1(a_n)} \underset{n\to\infty}{\to} +\infty \qquad \text{avec} \qquad \zeta_n = o(|\rho_n|).$$

■

Bien que la version 0 du test soit inutilisable en pratique, nous nous proposons d'étudier sa puissance asymptotique dans le but de déterminer dans un cadre simple la méthode que nous utiliserons pour étudier les autres versions du test.

Théorème 8 *On suppose que m_n vérifie les conditions de la proposition 1 (page 16) sous l'hypothèse nulle simple $\mathcal{H}_0 : F = F_0$ et que la vraie loi de Fdr $F = F_1$ vérifie les hypothèses des lemmes 7 et 8, page 31 :*

- *si $H_0 \in \mathcal{C}^1_\rho$ ($\rho \neq 1$) et $H_1 \in \mathcal{E} \setminus \mathcal{C}^1_\rho$, ou $H_0 \in \mathcal{C}^2$ et $H_1 \in \mathcal{E} \setminus \mathcal{C}^2$: $m_n = o((\ln n)^2)$.*
- *si $H_0 \in \mathcal{C}^1_1$ avec H_0^{-1} est une fonction affine et $H_1 \in \mathcal{E} \setminus \mathcal{C}^1_1$: $m_n = o(n)$.*
- *si $H_0 \in \mathcal{C}^1_1$ avec $|(H_0^{-1})''| \in \mathcal{SR}_{-1-\tau}$ et $H_1 \in \mathcal{E} \setminus \mathcal{C}^1_1$: $\forall\delta > 0,\ m_n = O((\ln n)^{2(1+\tau)-\delta})$.*
- *si $H_0 \in \mathcal{C}^3_\rho$ et $H_1 \in \mathcal{E} \setminus \mathcal{C}^3_\rho$: $\forall\delta > 0,\ m_n = O((\ln n)^{2(1-\rho)-\delta})$.*

Alors, la probabilité que la valeur du quantile sous $\mathcal{H}_0$, $q_{0,n}$, n'appartienne pas à l'intervalle
$$I_{th,n} = \left[\widehat{q}_{ET,n} + \delta_n \pm \sigma_n \frac{\ln(m_n/np_n)}{\sqrt{m_n}} z_{1-\alpha/2}\right],$$
intervelle théorique asymptotique de probabilité $1-\alpha$ pour le vrai quantile issu de la loi de Fdr $F = F_1$, tend vers 1 lorsque n tend vers l'infini.

Démonstration : L'erreur entre la valeur du quantile sous $\mathcal{H}_0$ et son estimation ET s'écrit
$$q_{0,n} - \widehat{q}_{ET,n} = \delta_{0,n} + q_{ET,n}(F_0) - q_{ET,n}(F_1) + q_{ET,n}(F_1) - \widehat{q}_{ET,n},$$
où $\delta_{0,n} = q_{0,n} - q_{ET,n}(F_0)$ est l'erreur d'approximation sous $\mathcal{H}_0$. On sait (proposition 2 page 18) que lorsque les conditions de la remarque 1 (page 16) sont vérifiées, on a
$$\sqrt{m_n}\,\frac{\widehat{q}_{ET,n} - q_{ET,n}(F_1)}{\sigma_n^{(1)}\ln(m_n/np_n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$
et on en déduit que
$$q_{0,n} = \widehat{q}_{ET,n} + \delta_{0,n} + q_{ET,n}(F_0) - q_{ET,n}(F_1) + \sigma_n^{(1)}\frac{\ln(m_n/np_n)}{\sqrt{m_n}}\xi_n, \tag{1.29}$$
où $\xi_n \xrightarrow{\mathcal{L}} \xi$, $\xi \sim \mathcal{N}(0,1)$. Posons $\Delta_n = q_{ET,n}(F_1) - q_{ET,n}(F_0)$, $z_{1-\alpha/2} = z$ et $B_n = \ln(m_n/np_n)/\sqrt{m_n}$. Par définition, la puissance du test s'exprime de la façon suivante :
$$\begin{aligned}
\text{Puissance} &= 1 - \mathrm{P}(\text{ accepter } \mathcal{H}_0 \mid F = F_1) \\
&= \mathrm{P}\left(q_{0,n} \notin \left[\widehat{q}_{ET,n} + \delta_{0,n} \pm \sigma_n^{(0)}B_n z\right] \mid F = F_1\right) \\
&= \mathrm{P}\left(\widehat{q}_{ET,n} + \delta_{0,n} - \Delta_n + \sigma_n^{(1)}B_n\xi_n \notin [\widehat{q}_{ET,n} + \delta_{0,n} \pm \sigma_n^{(0)}B_n z]\right) \\
&= \mathrm{P}(\xi_n \notin \underbrace{[\rho_n \pm \zeta_n z]}_{J_n}),
\end{aligned}$$

où ρ_n et ζ_n sont définis par l'équation (1.25), page 30.

Les bornes de l'intervalle J_n sont donc : $s_n = \rho_n - \zeta_n z$ et $t_n = \rho_n + \zeta_n z$. On veut montrer que $\mathrm{P}(\xi_n \notin [s_n,\, t_n]) \to 1$, ou, de façon équivalente, que $\mathrm{P}(\xi_n \in [s_n,\, t_n]) \to 0$ lorsque $n \to \infty$. Les lemmes 7 et 8 (page 31) montrent que l'on doit considérer deux cas.

- Dans un premier cas (cf. lemme 7 page 31), on a $\rho_n \to -\infty$ et $\zeta_n \to +\infty$ avec $\zeta_n = o(|\rho_n|)$ lorsque $n \to \infty$. Alors on montre que $t_n \to -\infty$ (puisque $\zeta_n z = o(|\rho_n|)$) et $s_n \to -\infty$ (puisque $s_n < t_n$) lorsque $n \to \infty$. A présent, si on note F_n la Fdr de ξ_n, on a $0 \leq \mathrm{P}(\xi_n \in [s_n,\, t_n]) = F_n(t_n) - F_n(s_n) \leq F_n(t_n)$. Or, si on note Φ la Fdr de la loi normale centrée et réduite, on a

$$\begin{aligned} 0 \leq F_n(t_n) = \Phi(t_n) + F_n(t_n) - \Phi(t_n) &\leq \Phi(t_n) + |F_n(t_n) - \Phi(t_n)| \\ &\leq \Phi(t_n) + \sup_{x \in I\!R} |F_n(x) - \Phi(x)|. \end{aligned}$$

  On sait que $\Phi(t_n) \to 0$ lorsque $n \to \infty$, puisqu'alors $t_n \to -\infty$. De plus, d'après le théorème de Polya [44] (p. 120), puisque $F_n \to \Phi$ lorsque $n \to \infty$, Φ étant une fonction continue, $\sup_{x \in I\!R} |F_n(x) - \Phi(x)| \to 0$ lorsque $n \to \infty$. On en déduit que $F_n(t_n) \to 0$, et donc que $\mathrm{P}(\xi_n \in [s_n,\, t_n]) \to 0$ lorsque $n \to \infty$.

- Dans le second cas (cf. lemme 8 page 31), on a $\rho_n \to +\infty$ et $\zeta_n \to 0$ lorsque $n \to \infty$. Alors on montre que $s_n \to +\infty$ et $t_n \to +\infty$ lorsque $n \to \infty$. Maintenant, on a $0 \leq \mathrm{P}(\xi_n \in [s_n,\, t_n]) = F_n(t_n) - F_n(s_n) \leq 1 - F_n(s_n)$. Or,

$$\begin{aligned} 0 \leq 1 - F_n(s_n) &= 1 - \Phi(s_n) - (F_n(s_n) - \Phi(s_n)) \\ &\leq 1 - \Phi(s_n) + |F_n(s_n) - \Phi(s_n)| \\ &\leq 1 - \Phi(s_n) + \sup_{x \in I\!R} |F_n(x) - \Phi(x)|. \end{aligned}$$

  On sait que $1 - \Phi(s_n) \to 0$ lorsque $n \to \infty$, puisqu'alors $s_n \to +\infty$. Par le même argument que précédemment, on en déduit que $1 - F_n(s_n) \to 0$, et donc que $\mathrm{P}(\xi_n \in [s_n,\, t_n]) \to 0$ lorsque $n \to \infty$.

On a donc montré que dans les deux cas la puissance tend vers 1.
■

Théorème 9 (puissance du test ET version 1 ($IC_{re,n}$))
On utilise pour le test l'intervalle de confiance réel de degré de confiance $1 - \alpha$ pour le quantile q_{1-p_n} défini par l'équation (1.16), page 19. On suppose que m_n vérifie les conditions de la proposition 1 (page 16) sous l'hypothèse nulle simple $\mathcal{H}_0 : F = F_0$ et que la vraie loi de Fdr $F = F_1$ vérifie les hypothèses des lemmes 7 et 8, page 31 :

- *Si $H_0 \in \mathcal{C}^1_\rho$ $(\rho \neq 1)$ et $H_1 \in \mathcal{E} \setminus \mathcal{C}^1_\rho$, ou $H_0 \in \mathcal{C}^2$ et $H_1 \in \mathcal{E} \setminus \mathcal{C}^2$: $p = p' = 1$ et $|q| < 2$.*
- *Si $H_0 \in \mathcal{C}^1_1$ avec H_0^{-1} est une fonction affine et $H_1 \in \mathcal{E} \setminus \mathcal{C}^1_1$: aucune condition supplémentaire sur p, p', q et q'.*

– *Si* $H_0 \in \mathcal{C}_1^1$ *avec* $|(H_0^{-1})''| \in \mathcal{SR}_{-1-\tau}$ *et* $H_1 \in \mathcal{E} \setminus \mathcal{C}_1^1$ *:* $p = p' = 1$ *et* $|q| < 2(1+\tau)$*.*
– *Si* $H_0 \in \mathcal{C}_\rho^3$ *et* $H_1 \in \mathcal{E} \setminus \mathcal{C}_\rho^3$ *:* $p = p' = 1$ *et* $|q| < 2(1-\rho)$*.*

Alors la puissance du test ET version 1, lorsque $F = F_1$*, tend vers 1 quand* $n \to \infty$*.*

Démonstration : Pour alléger les notations, comme dans la démonstration du théorème 8, on pose $\Delta_n = q_{ET,n}(F_1) - q_{ET,n}(F_0)$, $B_n = \ln(m_n/np_n)/\sqrt{m_n}$ et $z = z_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$. L'intervalle de confiance réel pour le quantile q_{1-p_n} de niveau α utilisé pour le test ET version 1 est le suivant :

$$IC_{re,n}(\mathcal{H}_0) = \left[\widehat{q}_{ET,n} + d_{0,n} \pm \widehat{\sigma}_n B_n z\right],$$

où $d_{0,n}$ est l'approximation au premier ordre de $\delta_{0,n}$, l'erreur d'approximation sous $\mathcal{H}_0$.

On suppose à présent que $F = F_1$. Par définition de l'estimateur ET donné par l'équation (1.29) page 33, on peut alors écrire

$$q_{0,n} = \widehat{q}_{ET,n} + \delta_{0,n} - \Delta_n + \sigma_n^{(1)} B_n \xi_n,$$

où $\xi_n \xrightarrow{\mathcal{L}} \xi$, $\xi \sim \mathcal{N}(0,1)$. Par définition, la puissance du test s'exprime de la façon suivante :

$$\begin{aligned}
\text{Puissance} &= 1 - \mathrm{P}(\text{ accepter } \mathcal{H}_0 \mid F = F_1) \\
&= \mathrm{P}\left(q_{0,n} \notin [\widehat{q}_{ET,n} + d_{0,n} \pm \widehat{\sigma}_n B_n z] \mid F = F_1\right) \\
&= \mathrm{P}\left(\widehat{q}_{ET,n} + \delta_{0,n} - \Delta_n + \sigma_n^{(1)} B_n \xi_n \notin [\widehat{q}_{ET,n} + d_{0,n} \pm \widehat{\sigma}_n B_n z]\right) \\
&= \mathrm{P}\left(\xi_n \notin \left[C_n^{(0)} \zeta_n + \rho_n \pm \frac{\widehat{\sigma}_n}{\sigma_n^{(1)}} z\right]\right),
\end{aligned}$$

où ρ_n et ζ_n sont définis par l'équation (1.25) page 30; et $C_n^{(0)}$ est défini, sous l'hypothèse $\mathcal{H}_0$, de façon analogue à C_n qui est donné par l'équation (1.20), page 24. On veut donc montrer que $\mathrm{P}(\xi_n \notin J_n') \to 1$ lorsque $n \to \infty$, où

$$J_n' = \left[C_n^{(0)} \zeta_n + \rho_n \pm \frac{\widehat{\sigma}_n}{\sigma_n^{(1)}} z\right].$$

Sous l'hypothèse $\mathcal{H}_1$, Davis et Resnick [19] ont montré que si $m_n \to \infty$ et $m_n/n \to 0$ quand $n \to \infty$, alors $\widehat{\sigma}_n/\sigma_n^{(1)} \xrightarrow{P} 1$. De même que dans la démonstration du théorème 7 page 25, si on note

$$I_n = \left[C_n^{(0)} \zeta_n + \rho_n \pm z(1+\eta)\right],$$

cela implique que $\forall \eta > 0$, $\mathrm{P}(J_n' \subset I_n) \to 1$, donc que $\mathrm{P}(J_n' \not\subset I_n) \to 0$ lorsque $n \to \infty$.

D'autre part, sous les conditions du lemme 5 page 24, on sait que $C_n^{(0)} \to 0$ lorsque $n \to \infty$. De plus, d'après les lemmes 7 et 8 page 31, on sait que $\zeta_n = o(|\rho_n|)$. Il s'ensuit que $C_n^{(0)} \zeta_n = o(|\rho_n|)$, et donc que

$$I_n = \left[\rho_n(1+\epsilon_n) \pm z(1+\eta)\right],$$

où $\epsilon_n = C_n^{(0)}\zeta_n/\rho_n \to 0$ lorsque $n \to \infty$ (ϵ_n étant de signe quelconque).

Comme dans la preuve du théorème 7 page 25, on en déduit que $\forall \eta > 0$, $\mathrm{P}(\xi_n \notin I_n) \to 1$ lorsque $n \to \infty$, car $\rho_n \to \pm\infty$ quand $n \to \infty$. Si on remarque que

$$\begin{aligned}\mathrm{P}(\xi_n \notin I_n) &= \mathrm{P}\{(\xi_n \notin J'_n) \cap (J'_n \subset I_n)\} + \mathrm{P}\{(\xi_n \notin J'_n) \cap (J'_n \not\subset I_n)\}\\ &\leq \mathrm{P}(\xi_n \notin J'_n) + \mathrm{P}(J'_n \not\subset I_n),\end{aligned}$$

cela implique que $1 \geq \mathrm{P}(\xi_n \notin J'_n) \geq \mathrm{P}(\xi_n \notin I_n) - \mathrm{P}(J'_n \not\subset I_n) \to 1$ lorsque $n \to \infty$. On en déduit donc que $\mathrm{P}(\xi_n \notin J'_n)$, et donc la puissance, tendent vers 1.
■

1.2.4 Puissance des tests ET-BP complet et simplifié

Dans ce paragraphe, nous souhaitons étudier la puissance des deux tests ET-BP. Compte tenu du fait que, pour les mêmes raisons que dans le paragraphe 1.2.2 (page 28), nous ne pouvons pas espérer des résultats complets, nous nous plaçons à nouveau dans un cadre simplifié. Nous allons montrer, dans le cas où l'hypothèse nulle est simple, $\mathcal{H}_0 : F = F_0$, que la puissance en $F = F_1$ tend vers 1.

Proposition 5 (puissance du test ET-BP, version simplifiée)
On suppose que m_n vérifie les conditions de la proposition 1 (page 16) sous l'hypothèse nulle simple $\mathcal{H}_0 : F = F_0$ et que la vraie loi de Fdr $F = F_1$ vérifie les hypothèses des lemmes 7 et 8, page 31 :

- *Si $H_0 \in \mathcal{C}^1_\rho$ ($\rho \neq 1$) et $H_1 \in \mathcal{E} \setminus \mathcal{C}^1_\rho$, ou $H_0 \in \mathcal{C}^2$ et $H_1 \in \mathcal{E} \setminus \mathcal{C}^2$: $p = p' = 1$ et $|q| < 2$.*
- *Si $H_0 \in \mathcal{C}^1_1$ avec H_0^{-1} est une fonction affine et $H_1 \in \mathcal{E} \setminus \mathcal{C}^1_1$: aucune condition supplémentaire sur p, p', q et q'.*
- *Si $H_0 \in \mathcal{C}^1_1$ avec $|(H_0^{-1})''| \in \mathcal{SR}_{-1-\tau}$ et $H_1 \in \mathcal{E} \setminus \mathcal{C}^1_1$: $p = p' = 1$ et $|q| < 2(1+\tau)$.*
- *Si $H_0 \in \mathcal{C}^3_\rho$ et $H_1 \in \mathcal{E} \setminus \mathcal{C}^3_\rho$: $p = p' = 1$ et $|q| < 2(1-\rho)$.*

Alors, la puissance du test ET-BP simplifié, lorsque $F = F_1$ tend vers 1 lorsque N, puis n, tendent vers l'infini.

Démonstration : On suppose que l'échantillon des données $\mathbf{X} = (X_1, \ldots, X_n)$ est de loi de Fdr F_1. On tire N échantillons indépendants, chacun étant i.i.d., $\mathbf{X}^{*(j)} = (X_1^{*(j)}, \ldots, X_n^{*(j)})$ de loi de Fdr F_0, et on note $\mathbb{X} = (\mathbf{X}^{*(1)}, \ldots, \mathbf{X}^{*(N)})$ l'ensemble des échantillons bootstrap. De même que pour le niveau (voir l'équation (1.23) page 28), on montre que

$$\begin{aligned}\text{Puissance} &= \mathrm{P}\,(\,\text{rejeter } \mathcal{H}_0 \,|\, F = F_1\,)\\ &= E\Big[\mathrm{P}\big(\widehat{q}_{ET,n} \notin [\,\widehat{q}^*_{ET,\min,n},\, \widehat{q}^*_{ET,\max,n}\,] \,\big|\, \mathbb{X}\big) \,\big|\, F = F_1\Big]\\ &= E\Big[\mathrm{P}\big(dq_n^{(0)} \notin [dq^{*(0)}_{\min,n},\, dq^{*(0)}_{\max,n}\,] \,\big|\, \mathbb{X}\big) \,\Big|\, F = F_1\Big],\end{aligned}$$

où

$$dq_n^{(0)} = \sqrt{m_n}\frac{\widehat{q}_{ET,n} - q_{ET,n}(F_0)}{\sigma_n^{(0)}\ln(m_n/np_n)},$$

$$dq_{\min,n}^{*(0)} = \sqrt{m_n}\frac{\widehat{q}^*_{ET,\min,n} - q_{ET,n}(F_0)}{\sigma_n^{(0)}\ln(m_n/np_n)} \quad \text{et} \quad dq_{\max,n}^{*(0)} = \sqrt{m_n}\frac{\widehat{q}^*_{ET,\max,n} - q_{ET,n}(F_0)}{\sigma_n^{(0)}\ln(m_n/np_n)}\,.$$

On note $G_n^{(0)}$ la Fdr de la loi commune des différences (on en calcule une pour chaque échantillon bootstrap)

$$dq_{j,n}^{*(0)} = \sqrt{m_n}\frac{\widehat{q}^*_{ET,j,n} - q_{ET,n}(F_0)}{\sigma_n^{(0)}\ln(m_n/np_n)} \quad 1 \leq j \leq N\,.$$

Sous $\mathcal{H}_0$, les quantiles empiriques associés convergent alors presque sûrement lorsque $N \to \infty$ vers les quantiles correspondants de la fonction $G_n^{(0)}$:

$$dq_{\min,n}^{*(0)} \xrightarrow{ps} (G_n^{(0)})^{-1}(\alpha/2) \quad \text{et} \quad dq_{\max,n}^{*(0)} \xrightarrow{ps} (G_n^{(0)})^{-1}(1-\alpha/2) \quad \text{quand } N \to \infty$$

De même, on pose

$$dq_n^{(1)} = \sqrt{m_n}\frac{\widehat{q}_{ET,n} - q_{ET,n}(F_1)}{\sigma_n^{(1)}\ln(m_n/np_n)} \sim G_n^{(1)},$$

et on obtient

$$dq_n^{(0)} = \frac{dq_n^{(1)} + \rho_n}{\zeta_n},$$

où ρ_n et ζ_n sont donnés par l'équation (1.25), page 30. Par conséquent, à n fixé,

$$\begin{aligned}\lim_{N\to\infty} \text{Puissance} = 1 \;&+\; G_n^{(1)}\big[-\rho_n + \zeta_n (G_n^{(0)})^{-1}(1-\alpha/2)\big] \\ &-\; G_n^{(1)}\big[-\rho_n + \zeta_n (G_n^{(0)})^{-1}(\alpha/2)\big]\,. \end{aligned} \tag{1.30}$$

Ceci donne, pour chaque n fixé, une approximation de la puissance quand on choisit N grand. Montrons maintenant que le membre de droite de l'équation (1.30) tend vers 1 lorsque $n \to \infty$. On utilise le fait que les Fdr $G_n^{(0)}$ et $G_n^{(1)}$ convergent uniformément vers Φ, Fdr de la loi normale centrée et réduite, quand $n \to \infty$ d'après la proposition 2, page 18 (sous les hypothèses de la remarque 1, page 16). On obtient donc $(G_n^{(0)})^{-1}(1-\alpha/2) \to z_{1-\alpha/2}$ et $(G_n^{(0)})^{-1}(\alpha/2) \to z_{\alpha/2}$. Par conséquent, ces deux suites sont bornées. Puisque $\zeta_n = o(|\rho_n|)$, on en déduit que $\zeta_n(G_n^{(0)})^{-1}(\alpha/2) = o(|\rho_n|)$ et $\zeta_n(G_n^{(0)})^{-1}(1-\alpha/2) = o(|\rho_n|)$ quand $n \to \infty$. Alors, en raisonnant de la même manière dans la preuve du théorème 8 (page 33), on en déduit que la puissance tend vers 1 quand n et N tendent vers l'infini.
■

Remarque 1.10 (puissance du test ET-BP, version simplifiée)

1. *Cette démonstration montre que la puissance du test ET-BP simplifié étudiée ci-dessus, ainsi que celle du test version 1, est directement reliée à la quantité $|\rho_n|$ (définie par*

l'équation (1.25), page 30) et au fait qu'elle tend vers l'infini quand $n \to \infty$. La puissance est donc principalement liée à l'écart entre les approximations ET des quantiles d'ordre $1 - p_n$ des Fdr F_0 et F_1.

2. *Il ne nous semble pas possible d'établir un résultat analogue pour établir la convergence de la puissance du test pour une hypothèse nulle composite $\mathcal{H}_0 : F \in \{F_\theta : \theta \in \Theta\}$. En effet, dans ce cas, il faut estimer θ sous $\mathcal{H}_0$. Lorsque l'échantillon initial est issu de la loi de Fdr F_1, on ne connaît pas les propriétés de la suite $(\widehat{\theta}_n)_n$. Tout au plus peut-on dire que, puisque dans ρ_n, $q_{ET,n}(F_1) - q_{ET,n}(F_0)$ serait remplacé par $q_{ET,n}(F_1) - q_{ET,n}(F_{\widehat{\theta}_n})$, on peut espérer que la puissance du test tend vers 1 pourvu que*

$$\inf_{\theta \in \Theta} \sqrt{m_n} \frac{|q_{ET,n}(F_1) - q_{ET,n}(F_\theta)|}{\sigma_n^{(1)} \ln(m_n/np_n)} \underset{n \to \infty}{\to} \infty \,. \tag{1.31}$$

Notons aussi que $\sigma_n^{(0)}$ devrait être remplacé par $\sigma_n(F_{\widehat{\theta}_n})$, quantité que l'on ne maîtrise pas. Mais on peut conjecturer, de façon analogue au cas simplifié que nous avons étudié, que le rapport correspondant à ζ_n reste négligeable devant l'équation (1.31) correspondant à $|\rho_n|$.

Remarque 1.11 (puissance du test ET-BP, version complète)
Comme pour le niveau d'une part, et pour le test ET-BP simplifié d'autre part, on rencontre des difficultés insurmontables pour étudier la puissance du test ET-BP complet dans le cas d'une hypothèse composite. Même dans le cas d'une hypothèse simple, la présence dans les bornes bootstrap de l'intervalle de confiance $IC_{\delta,BP}$ des quantités $q_{1-p_n,n}(F_{\widehat{\theta}_n^{*(j)}})$ nous empêche d'utiliser une démonstration analogue au cas simplifié. La simplification qui consiste à remplacer l'estimation du quantile paramétrique $q_{1-p_n,n}(F_{\widehat{\theta}_n^{*(j)}})$ par la vraie valeur $q_{1-p_n,n}(F_0) = q_{0,n}$ pour les échantillons bootstrap nous permet d'achever la démonstration dans ce cas particulier puisque l'on retombe exactement sur le cas du test ET-BP simplifié. Une fois encore, pour les cas plus complexes, nous sommes obligés d'avoir recours à des simulations pour comparer les puissances des différentes versions du test ET.

Dans la partie suivante, nous mettons numériquement en évidence les propriétés des différents tests qu'il n'a pas été possible d'établir théoriquement.

1.3 Données simulées

Pour les différents modèles envisagés, nous cherchons à savoir quelles valeurs de m_n utiliser en fonction de n, p_n et de la version du test ET pour qu'en pratique le niveau et la puissance de ce test restent corrects. On applique donc les trois versions du test ET à des données simulées selon l'un des modèles étudiés, avec différentes tailles d'échantillons, différents nombres d'excès et différents ordres de quantiles. Dans un premier temps, nous présentons l'exemple de la loi gamma pour des échantillons de taille 100. On ne retiendra que les nombre d'excès pour lesquels le niveau expérimental (c'est-à-dire le pourcentage d'échantillons rejetés à tort) atteint ou dépasse le niveau théorique α. De telles valeurs de m_n sont celles que l'on conseille d'utiliser, dans le paragraphe 1.3.3 (page 44), pour appliquer le test ET.

On calcule aussi la puissance expérimentale (pourcentage d'échantillons rejetés à raison) pour différentes alternatives et différentes valeurs de n, m_n et p_n. Un exemple est donné dans une seconde partie pour des données issues d'une loi de Weibull proche de la loi exponentielle et une hypothèse nulle d'exponentialité. Ceci nous permet de contrôler jusqu'à quel point les différentes versions du test discriminent des lois proches comme celles-là.

Enfin, nous avons effectué des simulations intensives sur le niveau et la puissance expérimentaux des différentes versions du test, pour différentes hypothèses et alternatives. Elles nous permettent de déterminer des valeurs de m_n, le nombre d'excès, et p_n ($1-$ l'ordre du quantile extrême pour lequel on effectue le test) pour lesquels le niveau et la puissance expérimentaux atteignent des valeurs convenables. Dans une troisième partie, on présente une synthèse des valeurs de m_n et p_n ainsi déterminées, que l'on conseille alors d'utiliser.

1.3.1 Niveau des versions du test ET

Afin d'explorer le niveau expérimental des différentes versions du test ET, on simule 500 jeux de données selon l'une des lois que nous étudions (normale, lognormale, exponentielle, gamma ou Weibull), et nous calculons le pourcentage d'acceptation du modèle correspondant par le test ET (version 1, 2 ou 3 successivement). On obtient donc le pourcentage d'acceptation à raison du modèle testé. Le pourcentage de rejet à tort du modèle correspond à un niveau expérimental pour le test appliqué. On a calculé le niveau expérimental, ou plus précisément le pourcentage d'acceptation à raison, pour chacune des trois versions du test et pour chacun des modèles normal, lognormal, exponentiel, gamma et Weibull.

On ne présente ici, à titre d'exemple, que le cas du modèle gamma pour les trois versions du test ET. On simule 500 jeux de données de loi $\mathcal{G}amma(3,3)$ et de taille 100 afin de calculer les pourcentages d'acceptation du modèle gamma pour différents quantiles d'ordre $1 - p_n$, différentes valeurs de m_n et différents niveaux de signification α des tests. On conseille de préférence l'utilisation des valeurs de m_n pour lesquelles le pourcentage d'acceptation calculé sur l'application du test à ces 500 échantillons est proche du pourcentage théorique $100(1-\alpha)$, afin que le niveau théorique $1 - \alpha$ annoncé soit approximativement atteint en pratique.

Remarque 1.12

- *On constate expérimentalement que la valeur du pourcentage d'acceptation ne dépend pas des paramètres des lois normale et lognormale, ni des paramètres d'échelle, (resp.) λ, η et η, des lois (resp.) gamma, exponentielle, et de Weibull, ni des paramètres de forme, (resp.) a et β, (resp.) des lois gamma et de Weibull pour autant que ces paramètres de forme soient suffisamment grands ($a >> 1$ et $\beta >> 1$).*

- *En théorie, on démontre (voir l'annexe C page 171) que*
  - *les résultats des trois tests sont indépendants des paramètres d'échelle et de position à distance finie (pour les lois étudiées ici).*
  - *les résultats des trois tests sont approximativement indépendants du paramètre $\beta >> 1$ de la loi de Weibull à distance finie.*

Pour la plupart des lois, ceci permet de limiter le nombre de simulations à effectuer pour évaluer le niveau et la puissance.

À ces jeux de données de taille 100 et de loi $\mathcal{G}amma(3{,}3)$, on applique tout d'abord le test ET version 1, basé sur la loi asymptotique de $\widehat{q}_{ET}$. Les résultats sont présentés dans le tableau 1.1. On constate sur ce tableau que le pourcentage d'acceptation du modèle gamma pour le test ET version 1 n'est que rarement proche du niveau attendu. Pour ne pas se restreindre à trop peu de valeurs de m_n, les valeurs de m_n que nous conseillons d'utiliser sont celles pour lesquelles la fréquence expérimentale d'acceptation présentée dans le tableau 1.1 est supérieure ou égale au niveau théorique du test. Ce niveau théorique étant parfois largement dépassé, cela risque, pour de telles valeurs de m_n, de réduire la puissance du test ET version 1.

	$p_n = 10^{-2}$		$p_n = 10^{-3}$		$p_n = 10^{-4}$		$p_n = 10^{-5}$		$p_n = 10^{-6}$	
α	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
$m_n = 5$	92	87.4	92	90	94.4	90.6	94.6	90.8	93.4	92.4
$m_n = 10$	94.6	92.4	96.6	95	98	95.2	97.4	96.4	97.8	95.2
$m_n = 20$	98.2	95.8	98.8	96.4	99.4	97	99.6	98	99.6	99
$m_n = 30$	98.8	97.8	99.2	97	99.8	96.2	99.4	94	99.4	92.6
$m_n = 40$	99.8	98	100	94.6	99.8	92.6	99	84.4	98.8	86
$m_n = 50$	99.6	95.2	99	92.6	96.8	82.4	96.2	74.8	94	65.8
$m_n = 60$	99.4	93.6	96	78.8	91.4	66.2	85.6	57.8	82	46
$m_n = 70$	98.8	87.8	91.4	60	76.4	40	71.2	10.8	50.6	7.2
$m_n = 80$	93.8	67.8	66.6	32.8	44	14.2	19.6	0.8	9.6	0.4

TAB. 1.1 – *Pourcentage moyen d'acceptation du test ET version 1 pour* 500 *échantillons de taille* $n = 100$ *et de loi gamma.*

On applique ensuite le test ET-BP complet (version 2) à ces mêmes 500 échantillons de taille $n = 100$ et de loi $\mathcal{G}amma(3, 3)$, avec $N = 500$ réplications bootstrap. Le tableau 1.2 présente les résultats obtenus. Cette fois-ci, la fréquence expérimentale d'acceptation est proche du niveau théorique pour un nombre relativement élevé de valeurs de m_n, contrairement au cas du test ET version 1. De plus, ce niveau théorique est peu dépassé. On peut donc supposer que la puissance du test ET-BP complet sera meilleure que celle du test ET version 1.

En dernier lieu, on applique le test ET-BP simplifié (version 3) avec $N = 500$ réplications bootstrap. Les résultats obtenus sont présentés dans le tableau 1.3. Pour cette dernière version du test, on constate que pour de nombreuses valeurs de m_n le niveau théorique est largement dépassé. On conseille donc d'utiliser les valeurs de m_n pour lesquelles la fréquence expérimentale d'acceptation est le plus proche du niveau théorique. On peut supposer que la puissance de cette version du test sera moins bonne que celle du test ET-BP complet.

	$p_n = 10^{-2}$		$p_n = 10^{-3}$		$p_n = 10^{-4}$		$p_n = 10^{-5}$		$p_n = 10^{-6}$	
α	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
$m_n = 5$	99	95	99.6	94.8	99.6	95.6	99.6	96	99.8	96.2
$m_n = 10$	98.4	94.2	98.6	93.6	98.6	93.4	98.8	93.4	98.8	93.2
$m_n = 20$	99.4	94.6	99.4	95.2	99.4	95.4	99.4	95.4	99.4	95.4
$m_n = 30$	99.2	96.8	99.4	96.8	99.2	96.8	99.2	96.8	99.2	96.8
$m_n = 40$	98.6	94.6	98.6	95	98.8	95	98.8	95	98.8	95.2
$m_n = 50$	99.2	95.2	99.4	95.6	99.4	95.4	99.4	95.6	99.4	95.4
$m_n = 60$	98.4	94.8	98.6	95	98.6	95.4	98.6	95.4	98.8	95.6
$m_n = 70$	99.6	96.8	99.6	96.8	99.6	96.8	99.6	97	99.6	97.2
$m_n = 80$	99.4	96	99.6	96.2	99.6	96.6	99.6	96.8	99.4	97
$m_n = 90$	98.6	95.8	98.8	95.8	98.8	96.4	98.8	96.4	98.8	96.6

TAB. 1.2 – *Pourcentage moyen d'acceptation du test ET-BP complet pour* 500 *échantillons de taille* $n = 100$ *et de loi gamma.*

	$p_n = 10^{-2}$		$p_n = 10^{-3}$		$p_n = 10^{-4}$		$p_n = 10^{-5}$		$p_n = 10^{-6}$	
α	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
$m_n = 5$	99.8	98.4	99.4	98.2	99.8	96	99.4	97.2	98.8	94.8
$m_n = 10$	100	99	99.8	98	100	97.4	99.6	97.6	99.6	95.8
$m_n = 20$	100	98.8	100	98.8	99.8	98	100	98.4	99.6	98.4
$m_n = 30$	100	100	100	99.6	99.6	98.8	100	98.8	99.8	99
$m_n = 40$	100	99.8	99.8	100	100	99.6	100	99.8	99.8	99.2
$m_n = 50$	100	100	100	99.6	100	100	100	99.6	100	99.6
$m_n = 60$	100	100	100	100	100	100	100	99.8	100	99.8
$m_n = 70$	100	100	100	100	100	100	100	100	100	100
$m_n = 80$	100	100	100	100	100	100	100	100	100	100

TAB. 1.3 – *Pourcentage moyen d'acceptation du test ET-BP simplifié pour* 500 *échantillons de taille* $n = 100$ *et de loi gamma.*

Les pourcentages calculés dans les tableaux 1.1 à 1.3 sont expérimentaux, calculés sur 500 échantillons. On cherche donc à donner un intervalle de confiance pour la valeur effective. Pour cela, il suffit d'estimer l'écart-type par la formule suivante, liée à la loi binomiale : puisque nous avons simulé 500 jeux de données, si la fréquence est $f \in [0,1]$, alors l'écart-type est approximativement $\sigma_f = \sqrt{f(1-f)/500}$. Par exemple, pour les pourcentages d'acceptation qui nous intéressent, on obtient les écarts-types présentés dans le tableau 1.4.

$100f$	80%	81%	82%	83%	84%	85%	86%	87%	88%	89%
$100\sigma_f$	1.79	1.75	1.72	1.68	1.64	1.6	1.55	1.5	1.45	1.4
$100f$	90%	99%	92%	93%	94%	95%	96%	97%	98%	99%
$100\sigma_f$	1.34	1.28	1.21	1.14	1.06	0.97	0.88	0.76	0.63	0.44

TAB. 1.4 – *Table des écarts-types pour 500 réplications.*

Les fréquences calculées dans le paragraphe suivant sont, elles aussi, obtenues par simulation. On pourra donc à nouveau utiliser le tableau 1.4 des écarts-types pour construire des intervalles de confiance pour les fréquences de rejet.

1.3.2 Puissance des versions du test ET

On souhaite, par des simulations, donner une idée de la puissance des différentes versions du test ET. Des simulations, dont les résultats seront présentés succinctement au paragraphe 1.3.3 (page 44), ont été réalisées pour étudier la puissance des trois versions du test ET pour des hypothèses et des alternatives appartenant aux modèles étudiés : normal, lognormal, gamma et de Weibull (le modèle exponentiel pouvant être considéré comme un cas particulier des modèles gamma ou de Weibull, on ne l'a pas intégré à ces essais). Ces modèles sont dans le plus souvent assez éloignés les uns des autres pour que l'on puisse en général espérer une bonne puissance des trois versions du test quand il s'agit de les discriminer.

Dans ce paragraphe, on étudie un cas plus critique, pour lequel la loi hypothèse et la vraie loi sont proches. Dans ce cas que l'on a voulu extrême, on souhaite ainsi savoir si les différents tests ET sont capables de discriminer les modèles choisis. On se place dans un cas simple : on simule 500 échantillons de taille 100, de loi de Weibull $\mathcal{W}(1, 0.7)$ (proche de la loi exponentielle puisque le paramètre de forme est proche de 1) et on teste l'adéquation de la loi exponentielle à de tels échantillons.

Tout d'abord, on étudie la puissance du test ET version 1, test basé sur la convergence en loi de $\widehat{q}_{ET,n}$. La puissance de ce test est alors très mauvaise : la puissance maximale est de 2% de discrimination, et, dès que $p_n < 10^{-2}$, la puissance est toujours nulle. Les lois réelle et hypothèse étant proches, il est difficile de les différencier. Cette première version du test ET ne parvient donc pas à discriminer deux lois proches. On montre au paragraphe 1.3.3 (page 44) qu'elle se comporte mieux pour deux lois éloignées, bien que sa puissance soit plus faible que celle des deux versions du test basées sur la méthode du bootstrap paramétrique.

On calcule ensuite la puissance expérimentale du test ET-BP complet avec $N = 500$ réplications bootstrap. Le tableau 1.5 présente les résultats obtenus. Au contraire du test ET version 1, la puissance du test ET-BP complet est satisfaisante, étant donné la proximité des lois de simulation et hypothèse. En effet, on peut rejeter (à raison) la loi hypothèse dans 80 à 90% des cas.

	$p_n = 10^{-2}$		$p_n = 10^{-3}$		$p_n = 10^{-4}$		$p_n = 10^{-5}$		$p_n = 10^{-6}$	
α	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
$m_n = 5$	45	67.6	30.4	51.4	27.4	45.6	26.2	43.4	25.8	42
$m_n = 10$	54.2	73.6	47	66.8	43	63.8	41.8	62.4	41	61.6
$m_n = 15$	64.4	79.6	59.6	73.8	57	72.2	56.2	71.6	55.6	71.4
$m_n = 20$	67.8	83.4	63.6	80.6	61.2	78.6	60.2	78	60	77.4
$m_n = 25$	70.4	85.8	67.2	81.2	66.8	81	66	80.4	65.4	80.4
$m_n = 30$	76.2	88.6	73	86.4	71.4	86	70.6	85.6	70	85.6
$m_n = 35$	80	90.4	78.2	89.6	77.6	89.4	77.4	89.4	77	89.4
$m_n = 40$	82.6	93	81.4	92.6	80.6	92.2	80.2	91.6	80	91.6

TAB. 1.5 – *Puissance expérimentale moyenne du test ET-BP version complète pour une hypothèse $\mathcal{H}_0$ exponentielle contre une loi de Weibull $\mathcal{W}(1, 0.7)$, calculée sur* 500 *jeux de données de taille $n = 100$.*

Enfin, le tableau 1.6 présente la puissance expérimentale du test ET-BP simplifié, toujours pour $N = 500$ échantillons bootstrap. On trouve dans ce cas des puissances bien meilleures que pour la première version du test ET, mais un peu moins bonnes que pour le test ET-BP complet (de l'ordre de $65 - 75\%$ de discrimination). Ces résultats semblent aussi assez satisfaisants, étant donné la proximité des lois réelle et hypothèse et donc la difficulté de les discriminer.

	$p_n = 10^{-2}$		$p_n = 10^{-3}$		$p_n = 10^{-4}$		$p_n = 10^{-5}$		$p_n = 10^{-6}$	
α	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
$m_n = 5$	37.8	60	29	50	27.6	46.6	25.6	44.2	25.6	43.2
$m_n = 10$	40.8	64	40	60	38.8	58.4	38.2	57.6	37.8	57.2
$m_n = 15$	44.8	67.8	44.8	66	44.8	65.8	45	65	45	64.4
$m_n = 20$	45.2	68.6	46.8	69.4	48	69.4	48.4	69.4	48.8	69.2
$m_n = 25$	44.6	72	49.8	74.4	50.2	75.2	50.4	75.2	50.6	75.2
$m_n = 30$	42	72.8	48	75.2	50.2	77.2	51.2	77	51.8	77.2
$m_n = 35$	37.2	69.8	43.6	75.6	44.6	77	46	78	47	77.8
$m_n = 40$	29.4	68.2	38.2	73.2	41.8	75.2	43	76.6	43.6	76.8

TAB. 1.6 – *Puissance expérimentale moyenne du test ET-BP version simplifiée pour une hypothèse $\mathcal{H}_0$ exponentielle contre une loi de Weibull $\mathcal{W}(1, 0.7)$, calculée sur* 500 *jeux de données de taille $n = 100$.*

1.3.3 Conseils sur le nombre d'excès à utiliser

Les valeurs de m_n à utiliser lorsque l'on veut appliquer le test ET sont celles pour lesquelles le niveau expérimental de la version du test utilisée est proche du niveau théorique, et pour lesquelles la puissance expérimentale moyenne sera en général bonne.

On s'appuie principalement sur les résultats concernant le niveau des différentes versions du test (voir le paragraphe 1.3.1 page 39 et Garrido [25]), car selon l'alternative considérée, les valeurs de m_n préconisées pour obtenir une bonne puissance expérimentale sont différentes. Il peut même arriver, pour certaines alternatives, que la puissance expérimentale contredise le niveau obtenu par simulation quant aux valeurs de m_n semblant adéquates, c'est-à-dire que les valeurs de m_n pour lesquelles le niveau est expérimental est bon impliquent une faible puissance, alors que la puissance est plus forte pour des valeurs de m_n pour lesquelles le niveau expérimental est trop faible. Il se peut aussi que la discrimination entre l'hypothèse et l'alternative soit si faible que l'on ne puisse conseiller aucune valeur de m_n permettant d'obtenir une bonne puissance.

Les valeurs de m_n que l'on doit utiliser dépendent de la taille n de léchantillon, de la version du test ET que l'on applique, du modèle testé, de l'ordre $1 - p_n$ du quantile utilisé pour construire l'intervalle de confiance du test, et du niveau α du test.

1.3.3.1 Test ET version 1 (basé sur la loi asymptotique de $\widehat{q}_{ET}$)

De tableaux du même type que ceux présentés au paragraphe 1.3.1 (page 39) pour la loi gamma (voir Garrido [25], annexe B), on déduit les valeurs de m_n à utiliser pour chaque type d'hypothèse nulle. Ces valeurs de m_n sont celles pour lesquelles le niveau expérimental égale ou dépasse le niveau théorique du test appliqué. Le tableau 1.7 présente les valeurs de m_n à utiliser pour des hypothèses nulles de type normale, exponentielle, gamma et Weibull. Les simulations sont faites pour la loi normale $\mathcal{N}(6,\ 1)$, la loi de Weibull $\mathcal{W}(3, 3)$, la loi exponentielle $\mathcal{E}xp(3)$, et la loi gamma $\mathcal{G}amma(3, 3)$. D'après la remarque 1.12 (page 39), les résultats obtenus sont valables pour toute loi normale, toute loi exponentielle, ainsi que tout paramètre d'échelle pour les lois de Weibull (η) et gamma (λ), puisque les paramètres intervenant sont des paramètres de position ou d'échelle. On peut aussi utiliser le tableau 1.7 pour des lois gamma et de Weibull dont le paramètre de forme est supérieur à 2 (cas de paramètres de forme grands par rapport à 1).

Attention : Pour les lois gamma et de Weibull, on a constaté expérimentalement que le test n'est applicable que lorsque le paramètre de forme est supérieur à 1, puisque dans le cas contraire le niveau théorique n'est jamais effectivement atteint, ni même suffisamment approché. Le tableau 1.7 n'est donc valable, pour les lois gamma et de Weibull, que lorsque le paramètre de forme est supérieur à 1. On constate le même phénomène pour la loi log-normale, quels que soient ses paramètres. Le point commun de ces lois est une queue de distribution relativement lourde. Le test ET version 1 ne semble donc pas applicable aux lois à queue lourde (cf. figure 1.1), tout au moins pour des tailles d'échantillon raisonnables.

n	p_n	$\mathcal{N}(\mu,\sigma)$	$\mathcal{W}(\eta,\beta)$ ($\beta \geq 2$)	$\mathcal{E}xp(\eta)$	$\mathcal{G}amma(a,\lambda)$ ($a \geq 2$)
20	10^{-2}	$[6,14]$	$[8,16]$	$[8,10]$	$[12,16]$
	10^{-3}	$[6,10]$	$[8,14]$	$[10,12]$	$[10,14]$
	10^{-4}	$[6,10]$	$[8,14]$	$[10,12]$	$[10,14]$
	10^{-5}	$[6,8]$	$[8,14]$	$[10,12]$	$[10,12]$
	10^{-6}	$[6,8]$	$[10,14]$	10	$[10,12]$
50	10^{-2}	$[10,25]$	$[10,30]$	$[15,25]$	$[20,30]$
	10^{-3}	$[10,15]$	$[10,25]$	$[15,25]$	$[20,25]$
	10^{-4}	$[10,15]$	$[10,20]$	$[15,25]$	20
	10^{-5}	10	$[10,20]$	$[20,25]$	20
	10^{-6}	10	$[15,20]$	$[20,25]$	20
100	10^{-2}	$[15,40]$	$[10,50]$	$[25,35]$	$[30,50]$
	10^{-3}	$[10,20]$	$[15,30]$	$[30,40]$	$[20,40]$
	10^{-4}	$[10,20]$	$[20,30]$	$[30,40]$	$[20,30]$
	10^{-5}	10	$[20,30]$	$[30,50]$	$[20,30]$
	10^{-6}	10	$[20,30]$	$[40,50]$	20
500	10^{-3}	$[20,50]$	$[20,70]$	$[30,100]$	$[40,75]$
	10^{-4}	$[20,35]$	$[20,60]$	$[50,80]$	$[35,60]$
	10^{-5}	$[20,30]$	$[25,50]$	$[60,90]$	$[30,50]$
	10^{-6}	20	$[30,50]$	$[60,80]$	$[30,40]$

TAB. 1.7 – *Intervalles de valeurs de m_n à utiliser pour le test ET version 1.*

On souhaite à présent explorer la puissance du test. Le tableau 1.8 présente un encadrement de la puissance expérimentale (pourcentage de rejets justifiés de la loi hypothèse) calculée pour chacun des modèles hypothèses et pour les alternatives classiques (ces mêmes modèles). Ce tableau est déduit de résultats du type de ceux du paragraphe 1.3.2 (page 42).

Description des tableaux sur la puissance des trois versions du test ET Les lois selon lesquelles ont été simulés les échantillons pour lesquels on a exploré numériquement la puissance du test sont indiqués dans la colonne "*simulation*". Elles correspondent à l'alternative F_1. Il s'agit des lois normale $\mathcal{N}(6,1)$, lognormale $\mathcal{LN}(0,1)$, de Weibull $\mathcal{W}(3,3)$ et gamma $\mathcal{G}amma(3,3)$. Pour chacune de ces lois, on a simulé des échantillons de taille $n = 20$, 50, 100 et 500 afin de montrer l'influence de la taille de l'échantillon sur la puissance. Les tailles d'échantillon sont indiquées dans la colonne "n". Les modèles F_θ testés (hypothèse nulle $\mathcal{H}_0 : F = F_\theta$) sont indiquées dans la ligne "*hypothèse*". Il s'agit des modèles normal, lognormal, de Weibull et gamma, sauf pour le test ET version 1 qui n'est pas applicable à la loi lognormale (voir la mise en garde page 44). En dernier lieu, on précise, lorsque c'est utile, les valeurs de m_n et de p_n à utiliser pour obtenir une meilleure puissance (sur les lignes "m_n" et "p_n"). En effet, pour un modèle hypothèse, une alternative et une taille d'échantillon fixés, la puissance de chaque version du test ET peut fortement varier en fonction des valeurs de m_n et p_n utilisées. Par exemple, la puissance du test ET version 1 pour un modèle hypothèse normal, une alternative lognormale et un échantillon de taille $n = 100$, la puissance peut varier entre 0 et 99%. Il est donc important de préciser les valeurs de m_n (petites, moyennes ou grandes parmi les valeurs de m_n conseillées à partir du niveau dans le tableau 1.7 pour le test ET version 1, le tableau 1.9 pour le test ET-BP complet et le tableau 1.11 pour le test ET-BP simplifié) et p_n (petites, moyennes ou grandes parmi les valeurs 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5} et 10^{-6} envisagées) qui permettent d'obtenir la meilleure puissance en fonction des lois hypothèse et alternative ainsi que de la taille d'échantillon.

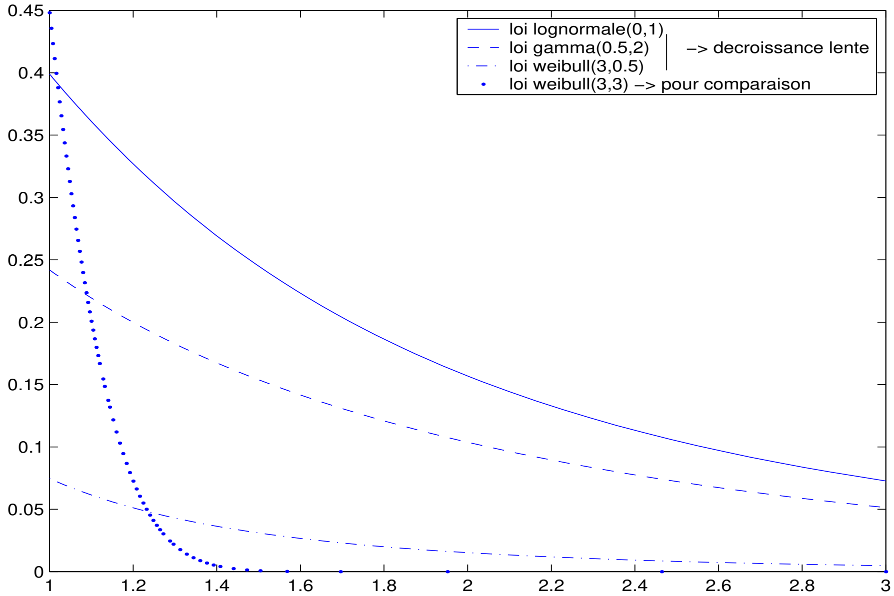

FIG. 1.1 – *Exemples de queues de distributions lourdes, fonctions de survies. – trait plein : loi lognormale* $\mathcal{LN}(0,1)$*, tirets : loi gamma* $\mathcal{G}amma(0.5,2)$*, discontinu : loi de Weibull* $\mathcal{W}(3,0.5)$ *(distributions à queue lourde), pointillés : loi de Weibull* $\mathcal{W}(3,3)$ *(pour comparaison).*

On constate (voir le tableau 1.8) que la loi gamma est très bien discriminée par cette première version du test ET. La loi normale est elle aussi assez bien rejetée lorsqu'il le faut, sauf dans le cas où l'on a simulé selon une loi de Weibull $\mathcal{W}(3,3)$. Mais alors, les lois estimées pour les modèles normal et Weibull sont quasiment confondues, ce qui explique l'acceptation par le test. Enfin, la loi de Weibull est très mal discriminée. La souplesse de la loi de Weibull, due à son paramètre de forme, explique le grand nombre d'acceptations par le test ET. En premier lieu, elle implique souvent une grande proximité entre la vraie loi et le modèle de Weibull estimé. De plus, le moindre changement de ce paramètre de forme donne des quantiles estimés différents. Cela implique que les intervalles bootstrap calculés pour la loi de Weibull sont étendus, et donc que le test a tendance à souvent accepter la loi de Weibull, en particulier à tort. Cela explique la mauvaise discrimination de la loi de Weibull, due à la souplesse qu'implique son paramètre de forme.

hypothèse		normale	de Weibull	gamma
simulation	n			
normale $\mathcal{N}(6,1)$	20	$\times$	$[0, 38.4]$	100
	50	$\times$	$[0, 36.8]$	100
	100	$\times$	$[0, 42.2]$	100
	500	$\times$	$[2.8, 47]$	100
	m_n	$\times$	petit	–
	p_n	$\times$	(1)	–
lognormale $\mathcal{LN}(0,1)$	20	$[0, 45.2]$	$[0, 4.2]$	$[0, 35.4]$
	50	$[0, 84.6]$	0	$[32, 85.8]$
	100	$[0, 99]$	$[0, 0.4]$	$[71, 100]$
	500	$[99.8, 100]$	$[31.4, 99.2]$	100
	m_n	grand	grand	grand
	p_n	(2)	(3)	(4)
de Weibull $\mathcal{W}(3,3)$	20	$[0, 8.6]$	$\times$	$[0.2, 64]$
	50	$[0, 5.8]$	$\times$	$[5.2, 82.6]$
	100	$[0, 5.6]$	$\times$	$[0.6, 96.8]$
	500	$[0, 17]$	$\times$	$[0, 81]$
	m_n	grand	$\times$	grand
	p_n	(5)	$\times$	(6)
gamma $\mathcal{G}amma(3,3)$	20	$[0, 60.6]$	$[0, 1.8]$	$\times$
	50	$[0, 62.2]$	$[0, 1]$	$\times$
	100	$[0, 71.6]$	$[0, 13.8]$	$\times$
	500	$[44.4, 100]$	$[0, 55.6]$	$\times$
	m_n	grand	grand	$\times$
	p_n	(7)	(8)	$\times$

TAB. 1.8 – *Puissance du test ET version 1 pour les lois classiques.*

(1) : p_n petit ($n = 20$ ou 50) ou grand ($n = 100$ ou 500)
(2) : p_n petit ($n = 20$) ou moyen ($n = 50$ ou 100)
(3) : p_n petit ($n = 20$) ou moyen ($n = 500$)
(4) : p_n petit ($n = 20$) ou grand ($n = 50$ ou 100)
(5) : p_n petit ($n = 20$), moyen ($n = 50$ ou 100) ou grand ($n = 500$)
(6) : p_n moyen ($n = 20$) ou grand ($n = 50$, 100 ou 500)
(7) : p_n petit ($n = 20$ ou 50), moyen ($n = 100$) ou grand ($n = 500$)
(8) : p_n petit ($n = 20$, 50 ou 100) ou grand ($n = 500$)

1.3.3.2 Test ET-BP complet (version 2)

On souhaite maintenant donner les valeurs de m_n à utiliser lorsque l'on applique la version 2 du test ET. Ces valeurs, déduites de tableaux du même type que du paragraphe 1.3.1 page 39 (voir Garrido [25], annexe D), sont celles pour lesquelles le niveau expérimental dépasse le niveau théorique du test ET-BP complet appliqué. Ceci étant vérifié pour un grand nombre de valeurs de m_n, on se restreint aux valeurs pour lesquelles le niveau expérimental est le plus proche du niveau théorique. Le tableau 1.9 présente les valeurs de m_n à utiliser pour des hypothèses nulles de type normale, lognormale, exponentielle, gamma et Weibull. Les lois de simulation sont les mêmes que pour le test ET version 1, sauf pour le modèle normal. Dans ce cas, on simule des jeux de données de loi normale $\mathcal{N}(5,1)$. D'après la remarque 1.12 (page 39), les résultats obtenus sont valables pour toute loi normale, toute loi exponentielle, toute loi de Weibull et toute loi gamma, à condition pour les lois gamma et Weibull que leur paramètre de forme soit supérieur à 2.

n	p_n	$\mathcal{N}(\mu,\sigma)$	$\mathcal{W}(\eta,\beta)$ $(\beta \geq 2)$	$\mathcal{E}xp(\eta)$	$\mathcal{LN}(\mu,\sigma)$	$\mathcal{G}amma(a,\lambda)$ $(a \geq 2)$
20	10^{-2}	$[6,12]$	$[2,8]$	$[2,6]$	$\times$	$[6,14]$
	10^{-3}	$[6,12]$	$[2,8]$	$[2,6]$	$[10,14]$	$[6,14]$
	10^{-4}	$[6,10]$	$[2,6]$	$[2,6]$	$[6,8]$	$[6,14]$
	10^{-5}	$[4,6]$	$[2,6]$	$[2,6]$	$[4,6]$	$[6,14]$
	10^{-6}	$[4,6]$	$[2,6]$	$[2,6]$	$[2,4]$	$[6,14]$
50	10^{-2}	$[10,30]$	$[2,16]$	$[2,8]$	$\times$	$[2,16]$
	10^{-3}	$[10,25]$	$[2,16]$	$[2,8]$	$[30,32]$	$[2,16]$
	10^{-4}	$[10,25]$	$[2,12]$	$[2,8]$	$[16,20]$	$[2,16]$
	10^{-5}	$[10,25]$	$[2,12]$	$[2,8]$	$[8,12]$	$[2,16]$
	10^{-6}	$[10,25]$	$[2,12]$	$[2,8]$	$[4,6]$	$[2,16]$
100	10^{-2}	$[5,65]$	$[5,30]$	$[5,15]$	$\times$	$[25,60]$
	10^{-3}	$[5,55]$	$[5,25]$	$[5,20]$	$[70,75]$	$[25,60]$
	10^{-4}	$[5,50]$	$[5,25]$	$[5,20]$	$[35,45]$	$[25,60]$
	10^{-5}	$[5,30]$	$[5,25]$	$[5,25]$	$[20,30]$	$[25,60]$
	10^{-6}	$[5,30]$	$[5,25]$	$[5,25]$	$[10,15]$	$[25,60]$
500	10^{-3}	$[150,350]$	$[10,110]$	$[20,80]$	$[430,440]$	$[40,130]$
	10^{-4}	$[80,250]$	$[30,90]$	$[10,70]$	$[330,350]$	$[40,130]$
	10^{-5}	$[80,220]$	$[40,90]$	$[10,70]$	$[200,230]$	$[40,130]$
	10^{-6}	$[80,220]$	$[40,90]$	$[10,70]$	$[100,140]$	$[40,130]$

TAB. 1.9 – *Intervalles de valeurs de m_n à utiliser pour le test ET-BP complet.*

Le tableau 1.10, déduit de résultats du type de ceux du paragraphe 1.3.2 (page 42), présente un encadrement de la puissance expérimentale du test ET-BP complet, calculée pour chacun des modèles hypothèses et pour les alternatives classiques (ces mêmes modèles). Dans ce tableau, on précise aussi, lorsque c'est utile, les valeurs de m_n et de p_n à utiliser pour obtenir une meilleure puissance.

hypothèse		normale	lognormale	de Weibull	gamma
simulation	n				
normale $\mathcal{N}(5,1)$	20	$\times$	$[92.8, 100]$	$[0, 2.6]$	100
	50	$\times$	$[99.8, 100]$	$[0, 5]$	100
	100	$\times$	100	$[0.2, 4]$	100
	500	$\times$	100	$[0, 11.6]$	100
	m_n	$\times$	–	(1)	–
	p_n	$\times$	–	(2)	–
lognormale $\mathcal{LN}(0,1)$	20	$[0.6, 61]$	$\times$	$[1, 21.8]$	$[42.2, 76]$
	50	$[3.8, 93.6]$	$\times$	$[11.4, 50.6]$	$[31.6, 76.2]$
	100	$[8.8, 100]$	$\times$	$[33.4, 71.6]$	$[64.8, 82.6]$
	500	$[28, 100]$	$\times$	$[89, 100]$	$[95.6, 98.8]$
	m_n	petit	$\times$	(4)	(5)
	p_n	(3)	$\times$	grand	(6)
de Weibull $\mathcal{W}(3,3)$	20	$[0.2, 6.4]$	$[6.4, 76.8]$	$\times$	100
	50	$[0.4, 7.6]$	$[5.6, 87]$	$\times$	100
	100	$[0, 7.6]$	$[10.8, 98.6]$	$\times$	100
	500	$[1.6, 18.8]$	$[73.4, 100]$	$\times$	100
	m_n	(7)	grand	$\times$	–
	p_n	grand	grand	$\times$	–
gamma $\mathcal{G}amma(3,3)$	20	$[0.4, 19]$	$[2.2, 24.8]$	$[0.6, 7.4]$	$\times$
	50	$[1, 46.2]$	$[4, 32.4]$	$[1.8, 12.6]$	$\times$
	100	$[1, 75.8]$	$[2.4, 35.6]$	$[5, 20]$	$\times$
	500	$[1.6, 100]$	$[0, 69.4]$	$[22.8, 59.2]$	$\times$
	m_n	petit	petit	moyen	$\times$
	p_n	(8)	(9)	grand	$\times$

TAB. 1.10 – *Puissance du test ET-BP complet pour les lois classiques.*

(1) : m_n grand ($n = 50$) ou petit ($n = 100$ ou 500)
(2) : p_n moyen ($n = 20$ ou 50) ou grand ($n = 100$ ou 500)
(3) : p_n petit ($n = 500$), moyen ($n = 50$) ou grand ($n = 100$)
(4) : m_n moyen ($n = 20$, 50 ou 100) ou grand ($n = 500$)
(5) : m_n grand ($n = 50$ ou 100) ou petit ($n = 500$)
(6) : p_n grand ($n = 20$, 50 ou 100) ou petit ($n = 500$)
(7) : m_n petit ($n = 20$) ou moyen ($n = 50$, 100 ou 500)
(8) : p_n petit ($n = 20$) ou grand ($n = 50$, 100 ou 500)
(9) : p_n grand ($n = 20$, 50 ou 100) ou moyen ($n = 500$)

La puissance de cette version du test est généralement meilleure que pour la version 1. On constate à nouveau que la loi gamma est très bien discriminée, alors que la loi de Weibull l'est assez mal. Cette fois-ci, on peut appliquer le test pour la loi lognormale, qui est relativement bien discriminée.

1.3.3.3 Test ET-BP simplifié (version 3)

De même que pour les deux précédentes versions du test, en s'appuyant sur des tableaux résultats du type de ceux du paragraphe 1.3.1 (page 42), on donne les valeurs de m_n à utiliser lorsque l'on applique la version 3 du test ET. Ainsi que pour la version 2, parmi les valeurs de m_n pour lesquelles le niveau expérimental égale ou dépasse le niveau théorique du test appliqué, on se restreint aux valeurs pour lesquelles le niveau expérimental est le plus proche du niveau théorique. Le tableau 1.11 présente les valeurs de m_n à utiliser sous les mêmes hypothèses que pour le test ET-BP complet.

n	p_n	$\mathcal{N}(\mu,\sigma)$	$\mathcal{W}(\eta,\beta)$ $(\beta \geq 2)$	$\mathcal{E}xp(\eta)$	$\mathcal{LN}(\mu,\sigma)$	$\mathcal{G}amma(a,\lambda)$ $(a \geq 2)$
20	10^{-2}	$[2,14]$	$[2,6]$	$[2,6]$	$[8,12]$	$[2,6]$
	10^{-3}	$[2,8]$	$[2,6]$	$[2,6]$	10	$[2,6]$
	10^{-4}	$[2,6]$	$[2,6]$	$[2,6]$	10	$[2,6]$
	10^{-5}	$[2,6]$	$[2,6]$	$[2,4]$	10	$[2,6]$
	10^{-6}	$[2,6]$	$[2,6]$	$[2,4]$	10	$[2,6]$
50	10^{-2}	$[5,20]$	$[5,20]$	$[5,20]$	$[5,30]$	$[5,10]$
	10^{-3}	$[5,15]$	$[5,20]$	$[5,20]$	$[5,20]$	$[5,10]$
	10^{-4}	$[5,15]$	$[5,20]$	$[5,10]$	$[5,15]$	$[5,10]$
	10^{-5}	$[10,20]$	$[5,10]$	$[5,10]$	$[10,20]$	$[5,10]$
	10^{-6}	$[10,20]$	$[5,10]$	$[5,10]$	$[10,20]$	$[5,10]$
100	10^{-2}	$[5,50]$	$[5,90]$	$[5,50]$	$[10,50]$	$[5,25]$
	10^{-3}	$[5,30]$	$[5,25]$	$[5,50]$	$[10,50]$	$[5,25]$
	10^{-4}	$[15,30]$	$[5,25]$	$[5,40]$	$[25,50]$	$[5,25]$
	10^{-5}	$[10,40]$	$[5,25]$	$[5,25]$	$[35,50]$	$[5,25]$
	10^{-6}	$[10,20]$	$[5,25]$	$[5,25]$	$[35,50]$	$[5,25]$
500	10^{-3}	$[10,150]$	$[10,100]$	$[10,200]$	$[10,200]$	$[10,100]$
	10^{-4}	$[10,150]$	$[10,100]$	$[10,150]$	$[50,200]$	$[10,100]$
	10^{-5}	$[10,100]$	$[10,100]$	$[50,150]$	$[100,200]$	$[10,50]$
	10^{-6}	$[10,100]$	$[10,100]$	$[50,150]$	$[150,200]$	$[10,50]$

TAB. 1.11 – *Intervalles de valeurs de m_n à utiliser pour le test ET-BP simplifié.*

Le tableau 1.12, déduit de résultats du type de ceux du paragraphe 1.3.2 (page 42), présente un encadrement de la puissance expérimentale du test ET-BP simplifié calculée pour chacun des modèles hypothèses et pour les alternatives classiques (ces mêmes modèles). Dans ce tableau, on précise aussi, lorsque c'est utile, les valeurs de m_n et de p_n à utiliser pour obtenir une meilleure puissance.

hypothèse		normale	lognormale	de Weibull	gamma
simulation	n				
normale $\mathcal{N}(5,1)$	20	$\times$	$[0.2, 3.2]$	$[0, 3]$	100
	50	$\times$	$[0.4, 14.6]$	$[0, 8.4]$	100
	100	$\times$	$[2.8, 27.2]$	$[0.4, 66.6]$	100
	500	$\times$	$[51.6, 95.8]$	$[1.8, 47.8]$	100
	m_n	$\times$	petit	grand	–
	p_n	$\times$	petit	(1)	–
lognormale $\mathcal{LN}(0,1)$	20	$[0.4, 51.4]$	$\times$	$[0, 7.4]$	100
	50	$[2.4, 90]$	$\times$	$[0, 22.2]$	100
	100	$[4, 99.6]$	$\times$	$[0, 41.6]$	100
	500	$[96, 100]$	$\times$	$[77.4, 99.6]$	100
	m_n	petit	$\times$	(3)	–
	p_n	(2)	$\times$	petit	–
de Weibull $\mathcal{W}(3,3)$	20	$[0, 4]$	$[0.4, 8.4]$	$\times$	$[99.2, 100]$
	50	$[0, 3.4]$	$[2.2, 37]$	$\times$	100
	100	$[0, 6]$	$[16.6, 72]$	$\times$	100
	500	$[0, 6]$	$[99.4, 100]$	$\times$	100
	m_n	petit	(4)	$\times$	–
	p_n	petit	petit	$\times$	–
gamma $\mathcal{G}amma(3,3)$	20	$[0, 16]$	$[0, 1.8]$	$[0, 4.8]$	$\times$
	50	$[0.4, 36.2]$	$[0, 9.4]$	$[0, 6.8]$	$\times$
	100	$[1.4, 60.2]$	$[1.2, 20.6]$	$[0, 10]$	$\times$
	500	$[65, 100]$	$[55.8, 95]$	$[7.2, 44.8]$	$\times$
	m_n	(5)	(7)	(9)	$\times$
	p_n	(6)	(8)	petit	$\times$

TAB. 1.12 – *Puissance du test ET-BP simplifié pour les lois classiques.*

(1) : p_n petit ($n = 20$), moyen ($n = 50$) ou grand ($n = 100$ ou 500)
(2) : p_n petit ($n = 20$) ou moyen ($n = 50$)
(3) : m_n petit ($n = 20$, 50 ou 100) ou moyen
(4) : m_n moyen ($n = 50$) ou petit ($n = 100$ ou 500)
(5) : m_n petit ($n = 20$ ou 50) ou grand ($n = 500$)
(6) : p_n petit ($n = 500$)
(7) : m_n petit ($n = 50$ ou 100) ou moyen ($n = 500$)
(8) : p_n petit ($n = 20$ ou 50) ou moyen ($n = 100$)
(9) : m_n petit ($n = 20$, 50 ou 100) ou moyen

La puissance de cette dernière version du test est généralement meilleure que la puissance de la première version, mais moins bonne que la puissance de la version complète du test ET-BP. Cette fois encore, on constate que la loi gamma est très bien discriminée, alors que la loi de Weibull l'est assez mal. On peut aussi appliquer cette version du test pour la loi lognormale qui est relativement bien discriminée, principalement pour de grandes tailles d'échantillons.

1.4 Données réelles

Nous allons à présent appliquer les différentes versions du test ET à deux jeux de données réelles, le premier de taille réduite, l'autre de taille plus importante.

1.4.1 Un premier jeu de données : nombre moyen de transitoires thermiques

Le premier jeu de données étudié ici, de taille $n = 24$, concerne le nombre moyen de transitoires thermiques (variations rapides et importantes de la température de l'eau) auxquels sont soumises certaines tuyauteries qui transportent de l'eau dans les centrales nucléaires d'EDF. Ces transitoires dégradent la résistance des tuyaux et peuvent provoquer des fissures, donc des fuites. La connaissance du nombre de transitoires par cycle (par exemple par an), et en particulier de ses valeurs extrêmes, est donc un élément important pour évaluer l'état des tuyaux.

Le jeu de donnés étant ici de taille $n = 24$, nous nous trouvons dans la situation d'un échantillon de petite taille, pour lequel les excès sont très peu nombreux et contiennent donc peu d'information. L'utilisation de la méthode des excès de semble donc pas judicieuse dans ce cas, et nous préférons adapter un modèle global à cet échantillon, ce qui nous permet d'utiliser toute l'information contenue dans les 24 données. Un histogramme des données, ainsi que les graphes des densités des différentes lois ajustées, sont présentés en figure 1.2.

Le test de Anderson-Darling

Tout d'abord, nous appliquons un test classique (le test de Anderson-Darling) pour plusieurs niveaux de signification.

niveau	Normale	Exponentielle	Weibull	Lognormale	Gamma
0.25	rejetée	rejetée	rejetée	acceptée	acceptée
0.1	acceptée	rejetée	rejetée	acceptée	acceptée
0.05	acceptée	rejetée	acceptée	acceptée	acceptée
0.025	acceptée	rejetée	acceptée	acceptée	acceptée
0.01	acceptée	rejetée	acceptée	acceptée	acceptée

TAB. 1.13 – *Nombre moyen de transitoires thermiques – Test de Anderson-Darling*

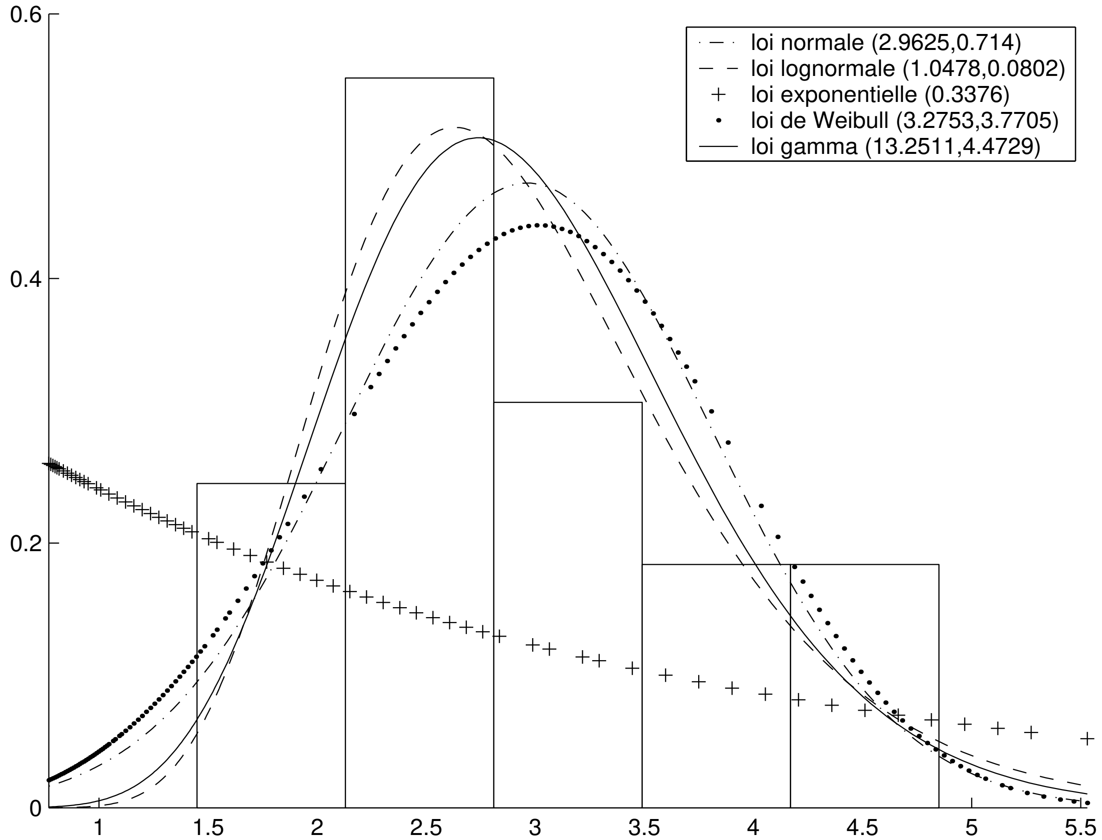

FIG. 1.2 – *Nombre moyen de transitoires thermiques – Histogramme des données et densités des lois estimées – discontinu : loi normale* $\mathcal{N}(2.9625, 0.714)$, *tirets : loi lognormale* $\mathcal{LN}(1.0478, 0.0802)$, *croix : loi exponentielle* $\mathcal{E}xp(0.3376)$, *pointillés : loi de Weibull* $\mathcal{W}(87.6502, 3.7705)$, *trait plein : loi gamma* $\mathcal{G}amma(13.5211, 4.4729)$.

On peut remarquer que le test de Cramér-von Mises donne exactement les mêmes résultats que le test de Anderson-Darling. Les lois qui semblent le plus en adéquation avec les données sont celles qui sont acceptées pour tous les niveaux de signification envisagés : les lois lognormale et gamma. Ensuite, par ordre décroissant d'adéquation aux données, on trouve les lois normale, de Weibull, et enfin la loi exponentielle, qui est toujours rejetée.

Même si l'application du test avec différents niveaux de signification nous pousse à préférer les lois lognormale et gamma, on ne peut pas discriminer ces deux lois. De plus, aux niveaux de signification les plus utilisés (par exemple 5% et 10%), seule la loi exponentielle est rejetée. Le test ET peut ainsi être utile pour discriminer des lois que les tests classiques n'ont pu différencier. Dans un cadre plus général, le test ET nous permet de déterminer si une loi convient pour modéliser des événements extrêmes, que ce soit une loi acceptée par un test usuel (en particulier lorsque l'on souhaite disposer d'un modèle global) ou non (si l'on ne souhaite modéliser que la queue de distribution des données).

Les différentes versions du test ET.

On applique les différentes versions du test pour les valeurs de m_n et de p_n donnant la meilleure puissance d'après les tableaux du paragraphe 1.3.3 (page 44). En particulier, on applique le test ET version 1 pour p_n petit, i.e. $p_n = 10^{-6}$, le test ET-BP complet pour p_n, grand i.e. $p_n = 0.001$, et le test ET-BP simplifié pour p_n petit, i.e. $p_n = 10^{-6}$. Le nombre d'excès que l'on emploie dépend de la version du test appliqué et de la loi hypothèse. On précise donc dans chaque cas la valeur de m_n utilisée. On présente les résultats obtenus dans le tableau 1.14. On se fie davantage au test ET-BP complet qu'au test ET-BP simplifié, pour lequel la puissance est plus faible. On peut donc considérer que les données sont de loi lognormale.

test ET version 1 (basé sur la loi asymptotique de $\widehat{q}_{ET}$), pour $p_n = 10^{-6}$				
loi	m_n	résultat	intervalle de confiance	$\widehat{q}_{\text{param},n}$
normale	8	acceptée	[5.9753 , 20.3393]	6.979
exponentielle	10	acceptée	[17.5043 , 65.1655]	40.9285
Weibull	14	rejetée	[8.0323 , 20.3369]	6.5720
lognormale	test non applicable aux distributions à queue lourde			
gamma	12	acceptée	[7.5985 , 20.3189]	8.5372
test ET version 2 (test ET-BP complet), pour $p_n = 0.001$				
loi	m_n	résultat	intervalle de confiance	$\widehat{\delta}_n$
normale	6	rejetée	[-0.5563 , 2.5142]	3.1898
exponentielle	6	rejetée	[-9.6465 , 13.2769]	-11.7007
Weibull	6	rejetée	[-0.5064 , 2.6076]	3.2951
lognormale	10	acceptée	[-1.0643 , 2.3645]	0.6534
gamma	6	rejetée	[-0.9773 , 2.5801]	2.6413
test ET version 3 (test ET-BP simplifié), pour $p_n = 10^{-6}$				
loi	m_n	résultat	intervalle de confiance	$\widehat{q}_{ET,n}$
normale	6	rejetée	[6.4393 , 14.2435]	15.7864
exponentielle	6	rejetée	[17.436 , 76.1775]	15.7864
Weibull	6	rejetée	[6.0334 , 15.0736]	15.7864
lognormale	10	acceptée	[7.9378 , 18.2286]	12.6403
gamma	6	acceptée	[6.7169 , 16.493347]	15.7864

TAB. 1.14 – *Nombre moyen de transitoires thermiques par an – Résutats des différentes versions du test ET, au niveau* $\alpha = 0.05$

Il semble que le test basé sur la loi asymptotique de $\widehat{q}_{ET}$ ne soit pas fiable sur un aussi petit échantillon. En effet, il accepte des lois dont les quantiles paramétriques sont parfois très différents (la loi exponentielle et la loi normale en particulier). On constate aussi la grande différence de résultats entre les versions complète et simplifiée du test ET-BP sur un petit échantillon comme celui-ci. Le test complet pousse à choisir la loi lognormale, et apporte une information supplémentaire par rapport aux tests centraux, alors que le test simplifié ne discrimine pas non plus les lois gamma et lognormale.

1.4.2 Un second jeu de données : teneurs de brins d'acier en azote

Le second jeu de données consiste en teneurs en azote (N) mesurées sur $n = 118$ brins d'acier. Ces brins d'acier sont des échantillons de l'alliage utilisé pour des composants critiques de centrales nucléaires d'EDF. Les concentrations en azote (aussi bien qu'en autres substances chimiques) correspondent à des contaminations résiduelles qui peuvent détériorer la qualité de l'alliage, et par conséquent altérer les caractéristiques de fiabilité des composants. Cela a déjà été étudié en détails dans [23]. La figure 1.3 représente un histogramme des données ainsi que les densités des lois estimées.

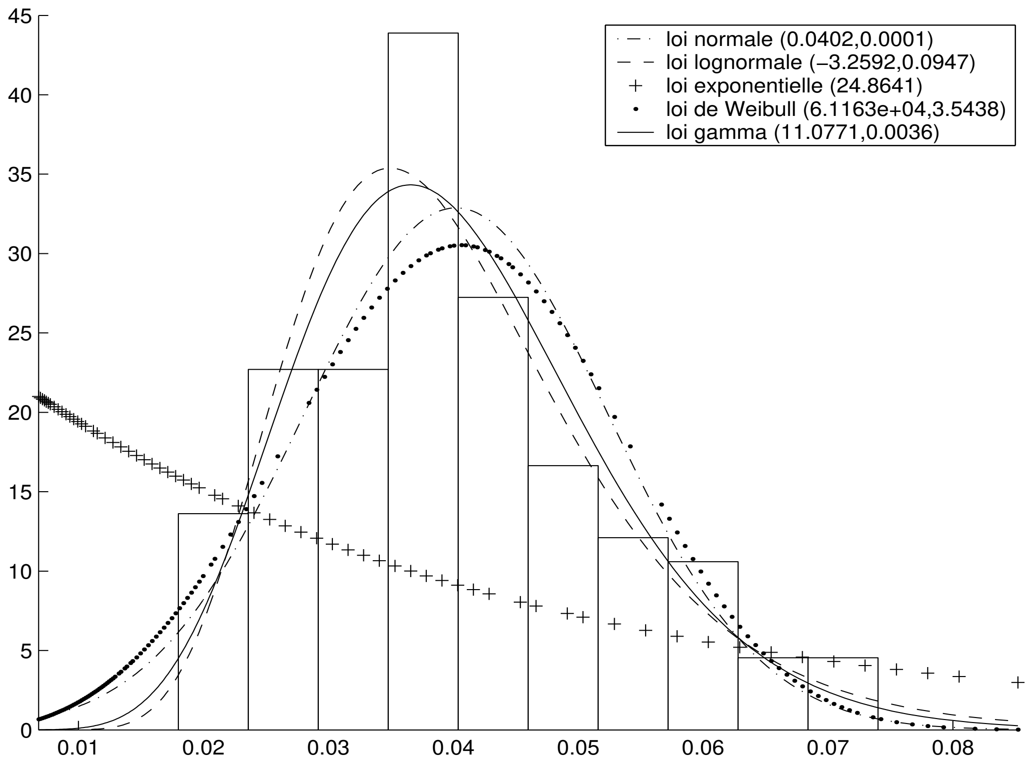

FIG. 1.3 – *Teneurs en azote – Histogramme des données et densités des lois estimées – discontinu : loi normale* $\mathcal{N}(0.0402, 0.0001)$, *tirets : loi lognormale* $\mathcal{LN}(-3.2592, 0.0947)$, *croix : loi exponentielle* $\mathcal{E}xp(24.8641)$, *points : loi de Weibull* $\mathcal{W}(61163, 3.5438)$, *trait plein : loi gamma* $\mathcal{G}amma(11.0771, 0.0036)$.

Cet histogramme possède deux barres consécutives de hauteur comparable en queue de distribution. On peut donc raisonnablement penser que cet échantillon est susceptible de comporter des données aberrantes. Cependant, Diebolt *et al.* [23] n'ont trouvé aucune preuve d'existence de données aberrantes pour ces teneurs en azote. De plus, les tests d'adéquation usuels ne rejettent pas l'exponentialité de la loi des excès pour des valeurs convenables du nombre d'excès m_n. Il est donc justifié d'appliquer un test sur la queue de distribution pour vérifier l'adéquation extrême des lois classiques. Le jeu de données étudié ici est de taille ($n = 118$) un peu plus importante que le précédent ($n = 24$), mais les excès peuvent rester en nombre restreint (nous avons utilisé $m_n = 15$ excès au minimum).

Le test d'Anderson-Darling accepte les modèles lognormal et gamma au niveau 5%. Malgré le nombre raisonnable de données, un test central ne discrimine donc pas ces modèles. Pour le premier jeu de données, le test ET nous avait permis de discriminer entre plusieurs modèles acceptés par un test central. Nous appliquons donc le test ET aux modèles gamma et lognormal, pour plusieurs valeurs de m_n que l'on précise, ainsi que les valeurs de p_n (voir le tableau 1.15).

test ET version 1, (basée sur la loi asymptotique de $\widehat{q}_{ET}$)					
loi	m_n	p_n	résultat	intervalle de confiance	$\widehat{q}_{\text{param},n}$
lognormale	test inapplicable pour les distributions à queue lourde				
gamma	25	0.001	acceptée	[0.0774 , 0.1140]	0.0880
	30	0.001	acceptée	[0.0808 , 0.1182]	0.0880
test ET version 2 (ET-BP complet), pour 500 échantillons bootstrap					
loi	m_n	p_n	résultat	intervalle de confiance	$\widehat{\delta}_n$
lognormale	70	0.001	rejetée	[0.0060 , 0.0238]	0.0052
	75	0.001	rejetée	[0.0073 , 0.0250]	0.0057
gamma	50	0.001	acceptée	[0.0056 , 0.0255]	0.0182
	55	0.001	acceptée	[0.0068 , 0.0269]	0.0195
	60	0.001	acceptée	[0.0092 , 0.0289]	0.0204
test ET version 3 (ET-BP simplifié), pour 500 échantillons bootstrap					
loi	m_n	p_n	résultat	intervalle de confiance	$\widehat{q}_{ET,n}$
lognormale	15	0.0001	acceptée	[0.0933 , 0.1624]	0.1139
	20	0.0001	acceptée	[0.0956 , 0.1555]	0.1125
	25	0.0001	acceptée	[0.0982 , 0.1585]	0.1155
gamma	15	0.0001	acceptée	[0.0864 , 0.1376]	0.1139
	20	0.0001	acceptée	[0.0903 , 0.1398]	0.1125
	25	0.0001	acceptée	[0.0900 , 0.1418]	0.1155

TAB. 1.15 – *Teneurs en chrome – Résultats des trois versions du test ET, au niveau* $\alpha = 0.05$.

Seul le test ET-BP complet discrimine ces deux lois : il rejette la queue de distribution du modèle lognormal, et ne rejette pas la queue de distribution gamma au niveau 5%. Par conséquent, puisqu'on ne suspecte aucune valeur aberrante dans ces données, on peut supposer que le modèle gamma est approprié. L'absence de données aberrantes semble d'ailleurs confirmé par le fait que le test ET-BP complet rejette le modèle lognormal qui présente une queue de distribution lourde, et lui préfère le modèle gamma à queue de distribution plus légère.

1.5 Conclusion.

Les méthodes spécifiques développées pour l'estimation des événements rares (méthode du maximum et méthode des excès) sont appliquées à partir d'une partie restreinte de l'échantillon de départ (maxima locaux ou excès au-delà d'un seuil). Dans le contexte d'échantillons de petite taille, il est donc déconseillé d'appliquer ce type de méthodes puisqu'elles ne pourront incorporer que très peu d'information (à partir d'une partie restreinte de l'échantillon lui-même petit). Nous préférons alors utiliser toute l'information contenue dans les données en adaptant un modèle usuel à l'échantillon, la loi des données étant estimée par des méthodes classiques comme le maximum de vraisemblance.

Cependant, lorsque l'on s'intéresse à l'estimation des événements rares, les tests usuels (par exemple Anderson-Darling ou Cramér-von Mises) ne permettent pas de déceler une mauvaise estimation de la queue de distribution. Nous avons donc proposé le test ET, qui permet, dans le cadre restreint du domaine d'attraction de Gumbel, de tester l'adéquation d'une loi à la queue de distribution. Nous avons décrit trois versions de ce test, la première basée sur une loi asymptotique, les deux autres utilisant la méthode du bootstrap paramétrique pour l'estimation des fluctuations d'échantillonnage.

Au cours des simulations, la première version du test s'est révélée peu puissante et elle ne peut être appliquée aux distributions à queue lourde. Nous conseillons plutôt d'utiliser les versions de type bootstrap paramétrique, qui sont plus puissantes et s'appliquent à toutes les lois. La version la plus puissante est la version bootstrap paramétrique complète. La version simplifiée est par contre plus rapide à appliquer, ce qui présente un intérêt certain pour les lois dont les estimateurs du maximum de vraisemblance sont longs à calculer (par exemple, les lois de mélange). En effet, pour le test ET-BP complet, les estimateurs du maximum de vraisemblance des paramètres sont calculés pour chacun des N échantillons bootstrap, c'est-à dire plusieurs centaines de fois, alors que pour le test ET-BP simplifié, on n'a besoin de les calculer que pour l'échantillon des données.

Les simulations montrent que la loi gamma est particulièrement bien discriminée. Les lois normale et lognormale sont en général assez bien discriminées, sauf dans le cas d'alternatives très proches. La loi de Weibull est par contre assez mal discriminée, très certainement à cause de sa souplesse, qui la fait s'adapter à de nombreuses formes. Le test ET présente cependant plusieurs inconvénients pour lesquels nous présentons des solutions dans le chapitre suivant.

Tout d'abord, supposons que l'on cherche une loi présentant une bonne adéquation en partie centrale (c'est-à-dire pour les valeurs les plus probables de la variable) ainsi qu'en extrême (pour les valeurs les plus rares). Lorsqu'aucune distribution n'est acceptée à la fois par un test usuel et par le test ET, il nous faut construire un nouveau modèle conservant une bonne adéquation en partie centrale de la distribution, mais avec un meilleur comportement en queue. À cet effet, nous proposons une procédure de régularisation bayésienne (voir partie 2.1 page 60). Cette procédure a pour point de départ un modèle accepté par un test central, c'est-à-dire un modèle révélant une bonne adéquation aux valeurs les plus probables. Mais, puisque l'on suppose que ce même modèle a été rejeté par le test ET, il nous faut modifier la queue de distribution afin d'en obtenir une meilleure modélisation. Nous utilisons pour cela une méthode bayésienne qui produit des modifications légères et régulières de la densité et qui permet d'introduire un avis d'expert sur le comportement en queue de distribution.

D'autre part, le test ET ne peut être appliqué que dans le domaine d'attraction de Gumbel. Nous présentons donc (sans l'étudier en détails) une version étendue de ce test, le test GPD (voir partie 2.3 page 91), qui peut être appliqué dans les trois domaines d'attraction des valeurs extrêmes : Fréchet, Gumbel, et Weibull. Puisque l'étude du test ET a montré les limites des lois asymptotiques dans ce cadre, pour construire le test GPD nous utilisons directement la méthode du bootstrap paramétrique pour l'estimation des fluctuations d'échantillonnage.

Chapitre 2

Développements consécutifs au test ET : la procédure de régularisation bayésienne et le test GPD

Nous nous plaçons dans le contexte de l'estimation des queues de distribution et des quantiles extrêmes. D'une part, nous souhaitons pouvoir travailler avec de petits échantillons ce qui n'est pas possible avec les méthodes usuelles d'estimation des événements rares qui n'utilisent qu'une petite partie des données, trop restreinte dans le cas de petits échantillons. D'autre part, dans certains cas, nous souhaitons déterminer des modèles qui s'ajustent globalement aux données, c'est-à-dire qui produisent une bonne estimation de la probabilité d'occurrence à la fois d'événements fréquents et d'événements rares. Que ce soit afin d'utiliser toute l'information contenue dans les données (dans le contexte de petits échantillons), ou pour étudier un modèle produisant une bonne modélisation des événements fréquents (lorsque l'on cherche un modèle global), nous utilisons ici des modèles paramétriques classiques. En particulier, la loi cherchée doit correctement modéliser les plus grandes observations, ainsi que fournir une bonne estimation de la queue de distribution au-delà de l'observation maximale. Nous avons construit un test d'adéquation spécifique, le test ET (voir le chapitre 1), pour vérifier la qualité de l'ajustement en queue. Mais, d'une part, le test ET n'est défini que pour des lois appartenant au domaine d'attraction de Gumbel ; et d'autre part, le test ET ne permet pas de proposer un modèle globalement adapté aux données lorsque la loi testée révèle une bonne adéquation centrale et une mauvaise adéquation extrême.

Nous nous penchons tout d'abord sur le problème de la construction d'un modèle ayant une bonne adéquation globale aux données lorsque les modèles centraux sont rejetés par le test ET. Nous proposons à cet effet une procédure de régularisation (partie 2.1 page 60), c'est-à-dire une procédure qui conserve l'allure générale de la loi initiale (qui produit une bonne modélisation des événements fréquents) et permet un meilleur ajustement de la queue de distribution. Cette procédure met en œuvre des outils bayésiens, et prend en compte l'opinion d'experts. La loi prédictive a posteriori qui en découle est proposée comme nouveau modèle.

Elle est obtenue comme mélange continu de densités du modèle initial par rapport à la loi a posteriori. Elle est donc régulière et aisément simulable. Nous détaillons numériquement cette méthode pour les modèles normal, lognormal, exponentiel, gamma et Weibull (pour ce dernier, nous travaillons soit sur le paramètre d'échelle, soit sur le paramètre de forme). Puis, nous illustrons la méthode sur des données simulées et des données réelles.

Le test ET et la procédure de régularisation bayésienne sont implémentés dans une maquette logiciel Matlab que nous avons conçue et fournie à EDF. Cette maquette permet une utilisation simple et rapide de ces méthodes par les ingénieurs de EDF. Dans la partie 2.2 (page 80) nous détaillons, à travers la présentation des modalités de cette maquette, les conditions d'application de ces procédures.

Nous nous penchons ensuite sur l'inconvénient de la restriction au DA(Gumbel) pour l'application du test ET. Nous proposons alors (partie 2.3 page 91) une version généralisée du test ET, le test GPD. Ce test est applicable dans les trois domaines d'attraction des valeurs extrêmes (Fréchet, Gumbel et Weibull). Nous décrivons le principe du test GPD, puis nous présentons quelques simulations et exemples, mais l'étude systématique des propriétés de ce test reste à faire.

En dernier lieu (partie 2.4 page 96), nous remarquons que la construction des tests ET et GPD ne les limite pas à être appliqués à des modèles acceptés par un test central. Nous explorons succinctement comment construire un modèle en queue de distribution auquel nous pourrions aussi appliquer les tests ET et/ou GPD.

2.1 Une procédure de régularisation bayésienne pour une meilleure adéquation extrême

Nous avons ici pour but de construire des modèles paramétriques réguliers et faciles à utiliser, avec une bonne adéquation globale aux données, c'est-à-dire à la fois pour les domaines centraux et extrêmes de la loi du jeu de données. Nous souhaitons que les lois de ces modèles aient des densités régulières. De plus, ces lois doivent être aisément simulables et permettre des calculs analytiques ou des calculs numériques simples. En particulier, nous devons être capables de calculer les fonctions de répartition et les fonctions quantiles.

En général des familles de modèles paramétriques sont sélectionnées grâce aux tests d'adéquation usuels (voir D'Agostino et Stephens [17] chapitre 4). Cependant, de tels tests ne se focalisent pas sur la queue de distribution des lois testées, puisque c'est un domaine où aucune observation n'a été effectuée. Seul le comportement central des données (c'est-à-dire le comportement des valeurs les plus probables) est réellement pris en compte. Ceci peut se révéler très dangereux lorsque l'on travaille ensuite sur les queues de distribution pour l'estimation de quantiles extrêmes, la prédiction d'événements rares, la simulation de valeurs extrêmes, etc. (voir Ditlevsen [33], Hahn et Meeker [36] ou Diebolt et El-Aroui [23]).

Pour l'adéquation en queue de distribution, nous avons développé un test extrême, le test ET ("Exponential Tail") défini au chapitre 1. Nous pouvons donc appliquer deux types de tests : l'un pour l'adéquation centrale (c'est-à-dire l'adéquation aux valeurs les plus probables de la variable) et l'autre pour l'adéquation en queue de distribution (c'est-à-dire pour l'adéquation aux valeurs rares de la variable). En pratique, lorsque l'on teste l'adéquation aux données de différents modèles, il peut arriver qu'aucun d'entre eux ne soit accepté à la fois en partie centrale et en extrême. En d'autres termes, un modèle paramétrique donnant une bonne description des valeurs les plus probables n'est pas forcément adapté aux observations les plus grandes, comme l'indique Ditlevsen [33]. Dans le but de construire une loi avec une bonne adéquation globale, une procédure de régularisation a été proposée. Le point de départ de cette procédure est un modèle accepté en partie centrale (par un test d'adéquation usuel), et on cherche à obtenir une distorsion de la queue de distribution suffisante pour aboutir à une meilleure adéquation de cette queue.

D'autre part, on souhaite pouvoir prendre en compte dans la procédure des informations extérieures à l'échantillon. En pratique, on suppose que l'on possède un (ou plusieurs) avis d'expert (qui nous indique comment distordre la queue de distribution). Celui-ci peut être subjectif ou relié à de précédents résultats expérimentaux. La procédure de régularisation décrite en partie 2.1.1 (page 60) suit une stratégie bayésienne. D'autres approches bayésiennes pour l'estimation des quantiles extrêmes ont été explorées par Coles et Dixon [14], Coles et Powell [15], Coles et Tawn [16], ainsi que dans le dernier chapitre de cette thèse. Ces approches introduisent les techniques bayésiennes dans l'estimation non paramétrique d'un quantile extrême (basée sur l'approximation de la loi des excès au-delà d'un seuil par une loi de Pareto généralisée). L'approche que nous proposons ici est au contraire basée sur une estimation paramétrique des quantiles extrêmes, et les techniques bayésiennes nous permettent de modifier nos modèles en vue d'une meilleure adéquation extrême. Les outils bayésiens ne permettent pas seulement de prendre en compte un avis d'expert, mais ils produisent aussi une distorsion régulière de la queue de distribution, ce qui est l'un de nos prérequis.

La procédure de régularisation bayésienne a d'abord été introduite par Catherine Trottier [27, 32]. Dans une première partie (paragraphe 2.1.1, page 62), pour des raisons de continuité, on rappelle la procédure qu'elle a décrite. La sortie principale de la procédure bayésienne (décrite au paragraphe 2.1.1.1 page 62) est la loi prédictive a posteriori, qui est alors proposée comme la nouvelle loi. Dans le paragraphe 2.1.1.2 (page 63), nous rappelons les lois prédictive a posteriori pour les différents modèles standards (exponential, gamma, normal, lognormal et Weibull). Enfin, le paragraphe 2.1.1.3 (page 65) définit la forme proposée pour l'avis d'expert, et son utilisation pour déterminer les hyperparamètres.

Pour des questions de simplicité des calculs, la procédure bayésienne n'est appliquée que sur l'un des paramètres de chaque modèle, généralement le paramètre le plus influent pour le comportement de la queue de distribution. Cependant, alors que pour le modèle de Weibull, le paramètre de forme est le plus important en queue de distribution, Catherine Trottier n'a implémenté la procédure bayésienne que pour le paramètre d'échelle de la loi de Weibull. Nous avons donc complété cette procédure en ajoutant le cas du paramètre de forme de la loi de Weibull décrit dans la partie 2.1.2 (page 66). En l'absence de loi a priori conjuguée, le choix de la loi a priori pour le paramètre de forme est plus délicat. De plus, les calculs analytiques étant dès lors impossibles, la procédure est plus calculatoire dans le cas du paramètre de forme de la loi de Weibull.

La partie 2.1.3 (page 70) présente les résultats de cette procédure de régularisation bayésienne sur des jeux de données simulés. Un jeu de données réelles (fourni par EDF) est présenté dans la partie 2.1.4 (page 76).

2.1.1 Une approche de régularisation bayésienne

On suppose à présent que $F_{\widehat{\theta}_n}$ est la fonction de répartition d'une loi acceptée en partie centrale, mais pas en queue de distribution. Notre but est de construire à partir de $F_{\widehat{\theta}_n}$ un modèle avec une bonne adéquation globale. Ce nouveau modèle devrait produire une bonne adéquation aux observations extrêmes et prendre en compte l'information donnée par des experts.

Une première idée est d'agréger $F_{\widehat{\theta}_n}$ tronquée à partir d'un certain seuil à l'approximation exponentielle de F à partir de ce seuil. Malheureusement, ceci conduit à une densité discontinue. De plus, cette méthode ne permet pas d'intégrer une information a priori. Nous avons donc choisi d'appliquer la méthodologie bayésienne.

2.1.1.1 Une procédure bayésienne

Dans ce cadre, les opinions des experts sont transformées en information a priori. On présente alors une procédure bayésienne qui est une méthode de régularisation régulière. Elle consiste à mettre une loi a priori de densité Π_γ, d'hyperparamètre γ, sur le (ou les) paramètre θ. Dans le paragraphe 2.1.1.3 (page 65), nous expliquons comment cette loi a priori prend en compte l'information apportée par des experts. Cela permet d'"attirer" θ vers un ensemble de valeurs plus adaptées à la queue de distribution privilégiée par l'expert.

Notons $\underline{x}_n = (x_1, \ldots, x_n)$ l'échantillon des données. La densité de la loi a posteriori est donnée par

$$\forall \theta \in \Theta, \quad \Pi_\gamma(\theta | \underline{x}_n) \propto \Big(\prod_{i=1}^{n} f_\theta(x_i)\Big)\, \Pi_\gamma(\theta)\,.$$

La densité de la loi prédictive a posteriori est alors obtenue en intégrant la densité $f_\theta(x)$ du modèle par rapport à la densité a posteriori de θ :

$$\forall x \in I\!R, \quad f_\gamma(x|\,\underline{x}_n) = \int_\Theta f_\theta(x)\Pi_\gamma(\theta|\,\underline{x}_n)\ d\theta\,.$$

Nous proposons de prendre la loi *prédictive a posteriori* comme loi régularisée. C'est un mélange continu de la densité du modèle par rapport à la loi a posteriori de θ, qui peut donc être aisément simulé. Il suffit pour cela de simuler des θ_i selon la loi a posteriori de θ, de densité $\Pi_\gamma(\theta|\underline{x}_n)$, puis de simuler selon les lois de densités $f_{\theta_i}(x)$.

2.1.1.2 En pratique

Catherine Trottier [27, 32] s'est intéressée aux cinq familles standard de lois : exponentielle, gamma, normale, lognormale et Weibull. Toutes ces distributions, sauf la loi exponentielle $\mathcal{E}xp(\lambda)$, ont un paramètre θ bi-dimensionnel. Pour faciliter les calculs, on a choisi de mettre une loi a priori uniquement sur l'une des composantes de θ, l'autre composante étant constante, égale à sa valeur estimée. Autant que possible, la loi a priori est mise sur la composante du paramètre qui paraît être la plus influente sur le comportement en queue de distribution.

Dans le cas de la loi normale, on travaille sur $1/\sigma^2$. Pour les lois $\mathcal{G}amma(\alpha, \lambda)$, on prend le paramètre d'échelle λ, puisque l'exponentielle est le terme le plus influent pour la queue de distribution. Pour les lois de Weibull $\mathcal{W}(\lambda, \beta)$, alors que le paramètre de forme β est le plus important en queue de distribution, Catherine Trottier a mis une loi a priori sur le paramètre d'échelle λ pour éviter de lourds calculs numériques. Nous avons ajouté le cas de la loi a priori sur le paramètre de forme β, qui est présenté au paragraphe 2.1.2 (page 66). Le traitement des cas lognormal (travail sur $1/\sigma^2$) et Weibull (avec β fixé, $\beta = \widehat{\beta}_n$: travail sur le paramètre déchelle λ) se réduit respectivement aux cas normal et exponentiel, par changement de variable.

On utilise autant que possible des lois a priori conjuguées, toujours pour des raisons de simplicité des calculs. Dans chaque cas étudié par Catherine Trottier et présenté dans ce paragraphe (exponentiel, normal avec moyenne fixée, gamma avec paramètre de forme fixé) la loi a priori conjuguée est une loi $\mathcal{G}amma(a, b)$. Cela permet d'écrire analytiquement la densité de la loi prédictive a posteriori. Le tableau 2.1 indique les densités des différentes lois prédictives a posteriori. On peut constater que l'on obtient :

- pour le cas exponentiel, une loi de Pareto généralisée,
- pour le cas gamma, une loi betaII avec changement d'échelle,
- pour le cas normal, une densité qui n'est pas usuelle mais peut être considérée comme une version étendue de la densité d'une loi de Student. Dans ce cas, une intégration numérique est nécessaire pour calculer la fonction de répartition et donc les quantiles de la loi prédictive a posteriori. Au contraire, la simulation selon la loi prédictive a posteriori est simple. Puisque, comme d'ailleurs dans le cas général, il s'agit d'un

modèle $f_\theta(x)$	loi a priori $\Pi_\gamma(\theta)$	loi a posteriori $\Pi_\gamma(\theta\vert\,\underline{x}_n)$	loi prédictive a posteriori $f_\gamma(x\vert\,\underline{x}_n)$
$\mathcal{E}xp(\theta)$ $(\theta=\lambda)$	$\mathcal{G}amma(a,b)$	$\mathcal{G}amma(a',b')$ avec $\begin{array}{lcl} a' & = & a+n \\ b' & = & b+\sum_{i=1}^{n} x_i \end{array}$	$GPD(\gamma,\sigma)$ avec $\begin{array}{lcl} \gamma & = & (a+n)^{-1} \\ \sigma & = & \dfrac{n\overline{x}+b}{n+a} \end{array}$
$\mathcal{G}amma(\alpha,\theta)$ $(\theta=\lambda)$	$\mathcal{G}amma(a,b)$	$\mathcal{G}amma(a',b')$ avec $\begin{array}{lcl} a' & = & a+n\alpha \\ b' & = & b+\sum_{i=1}^{n} x_i \end{array}$	$\dfrac{\Gamma(\alpha+a')}{\Gamma(\alpha)\Gamma(a')b'}\left(\dfrac{x}{b'}\right)^{\alpha-1}\left(1+\dfrac{x}{b'}\right)^{-(\alpha+a')}$
$\mathcal{N}\left(\mu,\dfrac{1}{\theta}\right)$ $\left(\theta=\dfrac{1}{\sigma^2}\right)$	$\mathcal{G}amma(a,b)$	$\mathcal{G}amma(a',b')$ avec $\begin{array}{lcl} a' & = & a+n/2 \\ b' & = & b+\dfrac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 \end{array}$	$\dfrac{\Gamma(a'+1/2)}{\sqrt{(2\pi b')}\Gamma(a')}\left(1+\dfrac{(x-\mu)^2}{2b'}\right)^{-(a'+1/2)}$
$X\sim\mathcal{LN}\left(\mu,\dfrac{1}{\theta}\right)$ $Y=\ln(X)\sim\mathcal{N}\left(\mu,\dfrac{1}{\theta}\right)$ $\left(\theta=\dfrac{1}{\sigma^2}\right)$	Utiliser le cas normal pour Y et avec l'échantillon $\underline{y}_n=\ln(\underline{x}_n)$		
$X\sim\mathcal{W}(\lambda,\beta)$ $Y=X^\beta\sim\mathcal{E}xp(\theta)$ $\left(\theta=\dfrac{1}{\lambda^\beta}\right)$	Utiliser le cas exponentiel pour Y et avec l'échantillon $\underline{y}_n=(\underline{x}_n)^\beta$ où $\theta=1/\lambda^\beta$		

Tab. 2.1 – *Distributions : du modèle, a priori, a posteriori et prédictive a posteriori.*

mélange de lois normales $\mathcal{N}(\mu, 1/\theta)$ (la loi du modèle hypothèse) avec comme mesure de mélange $\mathcal{G}amma(a', b')$ (loi a posteriori de θ, où a' et b' sont donnés dans le tableau 2.1), on simule simplement $\theta \sim \mathcal{G}amma(a', b')$ et $X \sim \mathcal{N}(\mu, 1/\theta)$.

2.1.1.3 De l'avis d'expert aux lois a priori

Traduire l'information a priori sur le modèle, c'est-à-dire déterminer les hyperparamètres de la loi a priori gamma, est un point important de la procédure. La fonction régularisée dépend des hyperparamètres inconnus $\gamma = (a, b)$. Puisqu'une approche de type bayésien hiérarchique n'est pas envisagée ici, ces hyperparamètres doivent être spécifiés. Pour cela, on définit tout d'abord à partir de l'avis d'expert un intervalle de variation $[\theta_1, \theta_2]$ pour θ, auquel on accorde un degré de confiance de $1 - \varepsilon$ pour un petit $\varepsilon > 0$. En approchant la loi a priori gamma par une loi normale centrée en $(\theta_1 + \theta_2)/2$, γ peut être explicitement calculé (la loi normale met une masse de $\varepsilon/2$ sur les intervalles $]-\infty, \theta_1]$ et $[\theta_2, +\infty[$).

Pour calculer θ_1 et θ_2, nous allons utiliser l'avis d'expert. Le type d'informations sur la queue de distribution que l'on souhaite obtenir d'un expert (ou d'un groupe d'experts, en agrégeant leurs opinions) concerne les valeurs extrêmes (comme pour Coles et Tawn [16]). On peut demander à l'expert une valeur $q_{\max}$ de la quantité d'intérêt dont il pense qu'elle est rarement atteinte. Il est ensuite nécessaire, avec son aide, de quantifier cette rareté. Ceci est réalisé en déterminant une borne inférieure p_1 et une borne supérieure p_2 pour le risque p associé à $q_{\max}$, c'est-à-dire la probabilité d'obtenir une valeur de la variable supérieure à $q_{\max}$. Les bornes p_1 et p_2 peuvent être d'ordres différents. Typiquement, pour un échantillon de taille 50 à 100, on peut prendre $p_1 = 10^{-2}$, et $p_2 = 10^{-4}$ ou 10^{-5}. Les effets de tels choix sont montrés dans les paragraphes 2.1.3 et 2.1.4 (pages 70 et 76) qui présentent des résultats numériques de cette méthode bayésienne de régularisation sur données simulées et réelles respectivement. Finalement, pour calculer les bornes θ_1 et θ_2, on interprète $q_{\max}$ comme un quantile d'ordre $1 - p_1$ et $1 - p_2$, respectivement :

$$F_{\theta_i}^{-1}(q_{\max}) = 1 - p_i \quad i = 1{,}2.$$

Notons que ces valeurs θ_1 et θ_2 peuvent être calculées analytiquement (cas exponentiel et Weibull associé) ou numériquement (cas normal, gamma et lognormal).

L'intervalle $[\theta_1, \theta_2]$ que l'on vient de calculer correspond à un intervalle de masse $1 - \varepsilon$ pour la loi a priori sur θ (la loi $\mathcal{G}amma(a, b)$). A présent, on approxime la loi a priori $\mathcal{G}amma(a, b)$ par une loi normale de mêmes moyenne et variance $\mathcal{N}(a/b, a/b^2)$, centrée sur $(\theta_1 + \theta_2)/2$. Les propriétés des intervalles de confiance de la loi normale $\mathcal{N}(\mu, \sigma^2)$ ($\mu = (\theta_1 + \theta_2)/2$ et $z_{1-\varepsilon/2}\sigma = (\theta_2 - \theta_1)/2$) nous permettent de déduire les approximations suivantes pour a et b :

$$a = \left(\frac{\theta_1 + \theta_2}{\theta_1 - \theta_2}\right)^2 z_{1-\varepsilon/2}^2 \quad \text{et} \quad b = 2\frac{\theta_1 + \theta_2}{(\theta_1 - \theta_2)^2}\, z_{1-\varepsilon/2}^2,$$

où $z_{1-\varepsilon/2} = \Phi^{-1}(1 - \varepsilon/2)$ est le quantile d'ordre $1 - \varepsilon/2$ d'une loi normale centrée et réduite.

2.1.2 Le paramètre de forme de la loi de Weibull

Dans ce paragraphe, nous nous focalisons sur des lois a priori pour le paramètre de forme β de la loi de Weibull, le paramètre d'échelle $\lambda = \widehat{\lambda}_n$ étant constant, égal à son estimateur du maximum de vraisemblance.

2.1.2.1 Choix de la loi a priori

Malheureusement, il n'existe pas de loi a priori conjuguée dans ce cas. Afin de choisir une famille convenable de lois a priori, remarquons tout d'abord que les formes des densités des lois de Weibull sont très différentes selon que $0 < \beta < 1$ ou $\beta > 1$ (voir la figure 2.1).

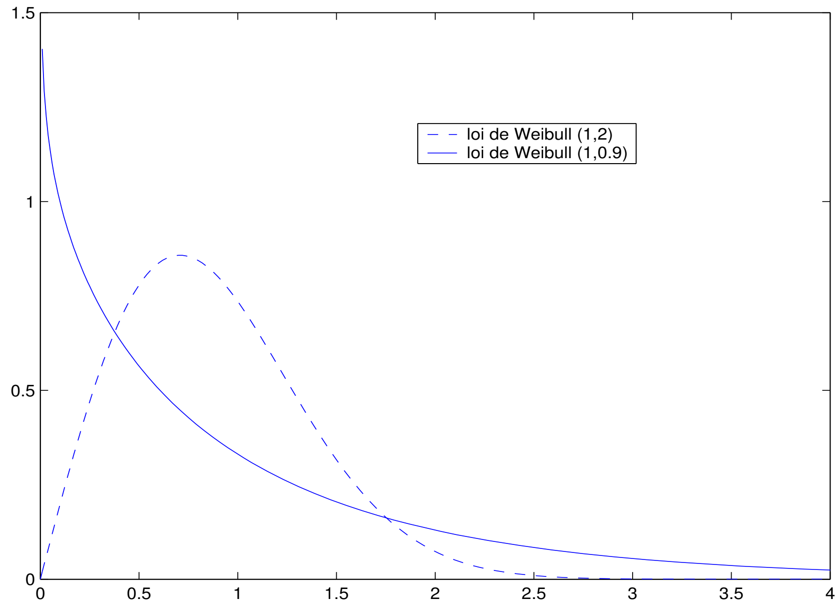

FIG. 2.1 – *Exemples de densités de lois de Weibull de paramètre de forme* $\beta > 1$ *ou* $0 < \beta < 1$ – *trait plein : loi de Weibull* $\mathcal{W}(1, 0.9)$*, tirets : loi de Weibull* $\mathcal{W}(1, 2)$.

En conséquence, une loi prédictive a posteriori dont la densité serait un mélange de densités de lois de Weibull avec $0 < \beta \leq 1$ pour certaines d'entre elles et $\beta > 1$ pour les autres pourrait ne pas s'adapter à l'échantillon puisque la loi de Weibull estimée, $\mathcal{W}(\widehat{\lambda}_n, \widehat{\beta}_n)$, est censée avoir été acceptée en central, avec soit $0 < \widehat{\beta}_n \leq 1$ soit $\widehat{\beta}_n > 1$ (mais pas les deux). En effet, le mélange de lois de Weibull avec $0 < \beta \leq 1$ pour certaines et $\beta > 1$ pour les autres peut produire une densité ayant une asymptote verticale en zéro, ainsi qu'un mode, alors que la forme de la vraie loi doit plutôt correspondre à celle de la Weibull acceptée en central qui possède soit une asymptote verticale en zéro (si $0 < \widehat{\beta}_n \leq 1$), soit un mode (si $\widehat{\beta}_n > 1$), mais pas les deux.

Par exemple, considérons la densité du mélange continu de densités de lois de Weibull, avec $0 < \beta \leq 1$ pour certaines et $\beta > 1$ pour les autres, selon une loi de mélange uniforme. Bien sûr, la loi a priori beta n'étant pas une loi conjuguée, rien n'indique que la loi a posteriori pourrait appartenir à cette même famille (comme la loi uniforme), mais nous avons choisi ce mélange simple pour l'exemple. Dans les trois cas d'une loi de mélange uniforme sur les intervalles $[0.2, 5]$, $[0.4, 3]$ ou $[0.5, 2]$, on obtient une asymptote verticale en zéro, ainsi qu'un mode, de moins en moins marqué (voir la figure 2.2). Si les données étaient issues de lois avec de telles densités, une loi de Weibull n'aurait pas pu être acceptée par un test central, les formes possibles de sa densité étant trop différentes (voir la figure 2.1). À l'inverse, si les données sont issues d'une loi de Weibull, ce type de densités est clairement inadapté. Pour des données issues d'une loi possédant soit un mode, soit une asymptote verticale en zéro, on pourrait certes supposer qu'en général l'influence de l'échantillon (à travers le calcul de la loi a posteriori) permet de privilégier des densités prédictives ayant elles-aussi soit un mode, soit une asymptote, et tend à repousser les densités possédant les deux comme celles de la figure 2.2. Mais rien ne permet de montrer cette conjecture. Puisque lorsqu'on mélange des lois de Weibull avec $0 < \beta \leq 1$ pour certaines et $\beta > 1$ pour les autres, on ne peut pas assurer que la loi prédictive n'aura pas un mode et une asymptote verticale en zéro, nous préférons exclure ce cas.

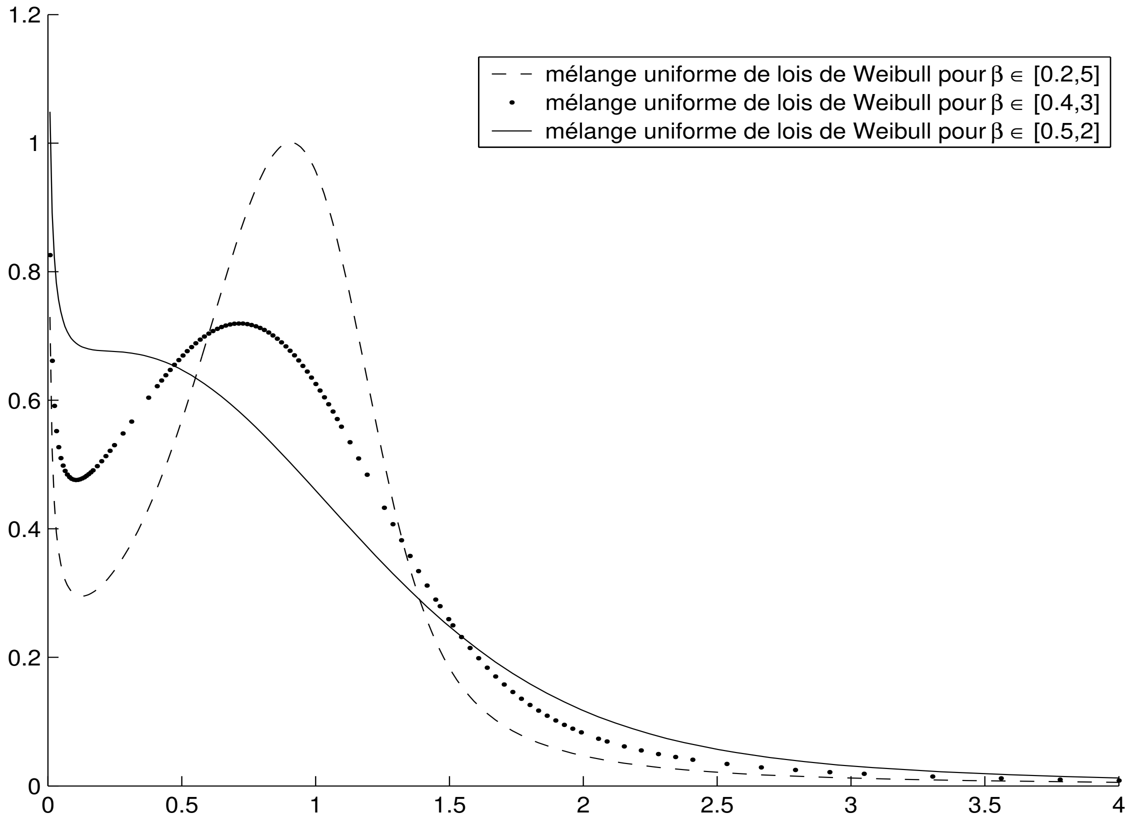

FIG. 2.2 – *Exemples de mélanges continus de densités de lois de Weibull, avec* $0 < \beta \leq 1$ *pour certaines et* $\beta > 1$ *pour les autres, selon une loi de mélange uniforme – mélange uniforme de lois de Weibull avec* $\beta \in [0.2, 5]$ *(tirets),* $\beta \in [0.4, 3]$ *(points), et* $\beta \in [0.5, 2]$ *(trait plein).*

En conséquence, il est tout d'abord nécessaire que le support des lois a priori puisse être compris soit dans $[0,1]$ soit dans $[1,\infty[$. Nous avons donc choisi une famille de lois a priori à support compact $[\ell_m, \ell_M]$ inclus soit dans $[0,1]$ soit dans $[1,\infty[$. Cet intervalle $[\ell_m, \ell_M]$ doit être choisi par l'utilisateur au début de la procédure. La valeur de $\widehat{\beta}_n$ peut contribuer à localiser ℓ_m et ℓ_M, puisqu'elle correspond à un modèle adéquat tout au moins en partie centrale. D'autre part, on peut noter qu'en fiabilité β plus grand que 10 n'a pas de signification physique réaliste, ce qui peut aider à choisir ℓ_M.

Puisque l'intervalle $[\theta_1, \theta_2]$ déterminé à partir de l'avis d'expert (voir le paragraphe 2.1.1.3 page 65) doit être contenu dans l'intervalle $[\ell_m, \ell_M]$, lui-même inclus soit dans $[0,1]$ soit dans $[1,\infty[$, nous ne pouvons pas accepter que l'on ait $\theta_1 < 1 < \theta_2$. Dans un tel cas, il apparaît une incertitude autour d'une queue de distribution exponentielle. Si l'estimateur $\widehat{\beta}_n$ du paramètre de forme est lui aussi proche de 1, il est conseillé d'appliquer la procédure pour la loi exponentielle. Dans tous les cas la procédure bayésienne sur le paramètre de forme de la loi de Weibull doit alors être interrompue.

D'autre part, une contradiction apparaît lorsque l'intervalle $[\theta_1, \theta_2]$ est d'un côté de 1 et $\widehat{\theta}_n$ de l'autre côté. Dans ce cas, il y a incompatibilité entre le modèle central (correspondant par exemple à une queue de distribution à décroissance rapide si $\widehat{\theta}_n > 1$) et l'information a priori sur la queue de distribution (correspondant alors à une queue de distribution lourde c'est-à-dire à décroissance lente). À nouveau, la procédure bayésienne sur le paramètre de forme de la loi de Weibull doit être interrompue pour remettre en question soit le modèle, soit l'avis d'expert.

Puisque les lois a priori doivent être à support compact, un choix naturel est la famille des lois beta de paramètres $\gamma = (a, b)$ (voir l'annexe A page 161). Avec une loi a priori de ce type, la densité de la loi a posteriori $\Pi_\gamma(\theta | \underline{x}_n)$ n'a pas de forme analytique puisque son cœfficient de normalisation n'en a pas. En conséquence, la densité de la loi prédictive a posteriori n'a pas non plus de forme analytique. Étant donné l'échantillon $\underline{x}_n$, pour chaque valeur de ℓ_m, ℓ_M, a et b, des méthodes numériques sont nécessaires pour calculer la densité de la loi prédictive a posteriori, sa fonction de répartition et ses quantiles extrêmes.

2.1.2.2 Détermination des hyperparamètres

Afin de spécifier les hyperparamètres a et b, nous devons tout d'abord calculer θ_1 and θ_2 de la même manière qu'au paragraphe 2.1.1.3 (page 65). L'expert doit déterminer une valeur extrême $q_{\max}$ et un encadrement $[p_1, p_2]$ du risque p associé à cette valeur. On en déduit les valeurs de θ_1 et θ_2 en interprétant $q_{\max}$ comme un quantile d'ordre $1-p_1$ et $1-p_2$ respectivement : $F_{\theta_i}^{-1}(q_{\max}) = 1 - p_i$. Il nous treste maintenant à déterminer les hyperparamètres a et b, en fonction de θ_1, θ_2, et de la confiance que l'on octroie à l'avis d'expert. Nous considérons alors trois cas dépendant de la confiance que nous accordons à l'intervalle $[\theta_1, \theta_2]$ (et à travers lui à l'expert) :

1. Si notre *confiance* en $[\theta_1, \theta_2]$ est *élevée*, alors nous approchons la loi beta par une loi normale centrée sur $[\theta_1, \theta_2]$, de mêmes moyenne et variance que la loi beta. Puis, nous interprétons $[\theta_1, \theta_2]$ comme un $(1-\varepsilon)$ intervalle de confiance, où $\varepsilon > 0$ est petit et quantifie plus précisément notre confiance en l'avis d'expert. Comme on peut le constater sur la partie gauche de la figure 2.3, dans ce cas la loi met une forte probabilité sur $[\theta_1, \theta_2]$.

2. Si l'on accorde une *confiance moyenne* à $[\theta_1, \theta_2]$, on souhaite que la densité de la loi beta soit significativement non nulle sur un intervalle plus grand que $[\theta_1, \theta_2]$. Des expériences numériques préliminaires nous ont montré que lorsque les deux paramètres a et b sont proches de 5, on obtient une densité en forme de cloche qui met un poids significatif sur la majeure partie du domaine $[\ell_m, \ell_M]$. Nous proposons donc de prendre l'un des paramètres égal à 5 et de fixer le mode de la densité au milieu de l'intervalle $[\theta_1, \theta_2]$. Si le mode $(= (\theta_1 + \theta_2)/2)$ est plus grand que le milieu de $[\ell_m, \ell_M]$, nous recommandons de choisir $a = 5$, ou dans le cas contraire $b = 5$. Cette situation est illustrée par la partie centrale de la figure 2.3.

3. Si l'on accorde une *faible confiance* à $[\theta_1, \theta_2]$, on souhaite que la loi beta ait à peu près la forme d'une loi uniforme (obtenue pour $a = b = 1$). De même, des expériences préliminaires nous amènent à fixer l'un des paramètres à 1.2 et à localiser le mode au milieu de $[\theta_1, \theta_2]$ pour obtenir une distribution en forme d'arche (voir la partie droite de la figure 2.3). À nouveau, nous recommandons de prendre $a = 1.2$ lorsque le mode est supérieur au milieu de $[\ell_m, \ell_M]$, et $b = 1.2$ dans le cas contraire.

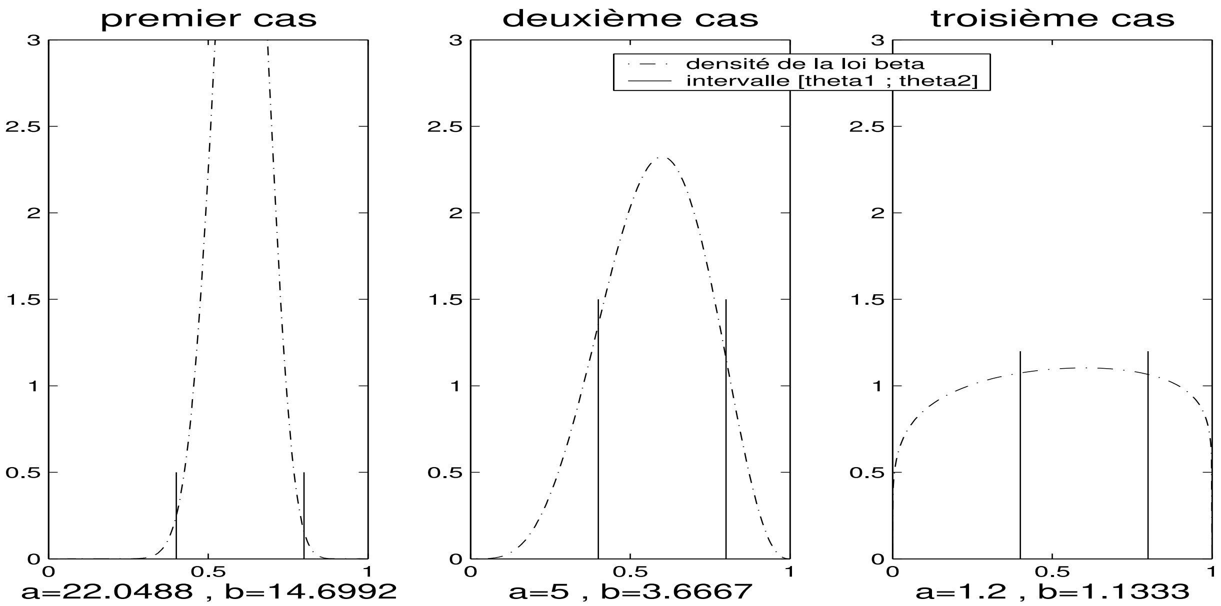

FIG. 2.3 – *Densité de la loi beta pour* $\theta_1 = 0.4$, $\theta_2 = 0.8$ *et* $[l_m, l_M] = [0, 1]$ *dans les trois cas de confiance en l'expert*

2.1.2.3 Simulation selon la loi prédictive a posteriori

Ainsi que mentionné précédemment, nous souhaitons pouvoir simuler selon la loi prédictive a posteriori. De même que dans le paragraphe 2.1.1.1 (page 62), cela se réduit à simuler des θ_i selon la densité a posteriori de θ, $\Pi_\gamma(\theta|\,\underline{x}_n)$, puis simuler selon les lois de densités $f_{\theta_i}(x)$. Il s'agit donc de savoir simuler selon la loi a posteriori de θ. Soit $L_n(\theta) = \prod_{i=1}^n f_\theta(x_i)$ la fonction de vraisemblance et $L_n(\widehat{\theta}_n)$ sa valeur maximale. Nous avons implémenté l'algorithme d'acceptation/rejet suivant pour la simulation selon $\Pi_\gamma(\theta|\,\underline{x}_n)$:

1. `Simuler` $Y \sim \mathcal{B}eta(a, b, [\ell_m, \ell_M])$;
2. `Simuler indépendamment` $U \sim \mathcal{U}[0, 1]$;

   `-- si` $U \leq \dfrac{L_n(Y)}{L_n(\widehat{\theta}_n)}\Pi_\gamma(Y)$ `alors prendre` $\theta = Y$;

   `-- sinon rejeter` Y.

Le nombre moyen de rejets que l'on obtient avant qu'une acceptation n'ait lieu est le rapport $L_n(\widehat{\theta}_n)/\int_{\ell_m}^{\ell_M} L_n(\theta)\,\Pi_\gamma(\theta)\,d\theta$. En général, ce rapport est de l'ordre de 1 à 5, voire plus petit, ce qui permet une simulation rapide. Cependant, pour certains de nos essais, nous avons trouvé un rapport de l'ordre de 100.

Mettre une loi a priori sur le paramètre de forme β d'une loi de Weibull est beaucoup plus complexe que de mettre une loi a priori sur le paramètre d'échelle λ. Cependant, cela peut aider à détecter des situations dans lesquelles les comportements en partie centrale et en queue de distribution sont incompatibles. Les performances de cette approche, ainsi que de celles présentées au paragraphe 2.1.1 (page 62), sont étudiées et comparées dans le paragraphe suivant.

2.1.3 Résultats sur simulations

On souhaite travailler avec des échantillons qui nous placent dans la situation pour laquelle a été développée la procédure de régularisation bayésienne : modèle central existant, mais rejeté en queue de distribution, alors que l'on cherche un modèle global. Pour un modèle donné dans notre ensemble de lois (normale, lognormale, exponentielle, gamma et Weibull), le point de départ de ces simulations est donc d'exhiber un échantillon accepté en partie centrale par le test de Cramér-von Mises mais rejeté en queue de distribution par le test ET-BP complet. Pour cela, on choisit une loi de simulation ressemblant en partie centrale au modèle donné, mais avec une queue différente. Avec un tel échantillon, nous estimons le modèle puis appliquons la procédure de régularisation. Les sorties sont :

- Différents quantiles extrêmes (pour différents risques p donnés par l'utilisateur) calculés pour la vraie loi, la loi estimée pour le vrai modèle, la loi estimée pour le modèle hypothèse, la loi prédictive a posteriori et la méthode ET. Cela permet de donner une idée quantitative des changements de comportement de la queue de distribution.

- La distance de Cramér-von Mises (distance CVM) et l'intervalle de confiance bootstrap du test ET-BP simplifié pour les différents modèles (hypothèse et prédictifs). Ceci permet de noter les changements apparus sur ces quantités lors de l'application de la procédure de régularisation bayésienne. Remarquons que le test de Cramér-von Mises ne peut pas être directement appliqué aux lois prédictives a posteriori, puisque les valeurs critiques correspondantes ne sont pas documentées (même si l'on peut parfois se ramener à un modèle est usuel, l'estimation des paramètres n'est pas classique). Nous pouvons seulement mesurer l'évolution de la distance de Cramér-von Mises par rapport au modèle hypothèse, et vérifier qu'elle ne se dégrade pas trop (ce qui ne constitue qu'un critère subjectif). D'autre part, afin de vérifier l'adéquation extrême, en dépit du fait que les lois prédictives n'appartiennent pas au DA(Gumbel) nous appliquons le test ET pour la loi prédictive (voir la remarque 2.1). Nous utilisons la version simplifiée du test ET-BP, même pour le modèle initial, afin de pouvoir comparer avec les lois prédictives a posteriori (pour lesquelles seul ce test simplifié peut être appliqué, à nouveau pour des raisons d'estimation des paramètres).
- Un tracé des différentes fonctions de survie, à partir du seuil. Ceci permet de constater visuellement les changements induits sur la queue de distribution par la procédure de régularisation bayésienne.

Remarque 2.1 *Les différentes lois prédictives a posteriori auxquelles nous aboutissons n'appartiennent pas au DA(Gumbel), mais au DA(Fréchet). Cependant, remarquons que nous ne disposons pour l'instant que du test ET et que l'exponentialité des excès a été acceptée pour un nombre m_n d'excès. De plus, la loi prédictive a posteriori reste relativement proche du modèle (en partie centrale et surtout pour la partie de la queue de distribution proche de la valeur maximale de l'échantillon à laquelle on applique généralement le test ET) qui, lui, appartient au DA(Gumbel). Nous supposons donc (en attendant, lorsque l'on dispose d'assez de données et donc d'excès pour estimer les deux paramètres de la loi GPD, de pouvoir appliquer le test GPD plus général défini en partie 2.3 page 91) qu'en première approximation nous pouvons appliquer le test ET pour vérifier l'adéquation extrême de la loi prédictive a posteriori.*

Pour l'information a priori, nous utilisons en tant que $q_{\max}$ la valeur du vrai quantile d'ordre $1-10^{-3}$ de la loi de simulation. En ce qui concerne le test ET, le nombre d'excès m_n que nous considérons en pratique est suggéré par les expériences numériques faites sur le test ET (voir la partie 1.3 page 38).

Le cas exponentiel

La loi de simulation utilisée dans ce cas est une loi de Weibull $\mathcal{W}(3, 1.3)$. Nous avons généré un échantillon de taille 100. Le modèle exponentiel estimé est $\mathcal{E}xp(0.3710)$ avec une distance CVM de 0.1789 pour une valeur critique à 5 % de 0.2216. Le test ET-BP simplifié appliqué à l'ordre $1-10^{-3}$ et au niveau 5 % produit l'intervalle $IC_{ET,BP,n} = [12.2673\,, 27.0092]$, alors que l'estimateur ET pour cet échantillon, si l'on considère $m_n = 14$ excès, est $\widehat{q}_{ET,n} = 12.2601$.

Donc dans ce cas le modèle exponentiel est accepté en partie centrale mais rejeté en queue de distribution. Le tableau 2.2 présente les résultats de la procédure de régularisation.

	loi prédictive a posteriori pour l'avis d'expert		loi du modèle estimé	vraie loi estimée	vraie loi de simulation	estimateur ET
	$p_1 = 10^{-2}$ $p_2 = 10^{-4}$	$p_1 = 10^{-3}$ $p_2 = 10^{-5}$				
$q_{0.99}$	11.2397	9.5427	12.4124	9.9728	9.7120	9.0307
$q_{0.999}$	16.9822	14.3947	18.6185	13.8661	13.2668	12.2601
$q_{0.9999}$	22.8082	19.3011	24.8247	17.5191	16.5528	15.4895
θ_1	0.3471	0.5207	$\times$			
θ_2	0.6942	0.8678	$\times$			
d_{CVM}	0.4785	1.2683	0.1789	(valeur critique : 0.2216)		
test ET	acceptée	acceptée	rejetée			
$IC_{ET,BP,n}$	11.5512	9.2169	12.2673			
	24.0674	19.5390	27.0092			

TAB. 2.2 – *Cas exponentiel* – $n = 100$ – *Résultats de la régularisation*

La première partie de ce tableau met en évidence la pertinence de cette procédure, puisque le modèle exponentiel accepté en central produit des quantiles surestimés (d'autant plus surestimés que l'ordre du quantile est grand), alors que pour les deux interprétations de l'avis d'expert utilisés, le quantile estimé par la loi prédictive a posteriori tend à être attiré vers la vraie valeur. Ceci peut aussi être constaté sur le graphique des fonctions de survie (tracées à partir du seuil) de la figure 2.4, où la queue de distribution de la loi régularisée est située entre les queues de distribution de la vraie loi et de la loi estimée pour modèle.

Comme prévu, la correction produite par la procédure est très légère. Il n'y a pas de grandes modifications ni dans l'estimation des quantiles, ni dans l'intervalle $IC_{ET,BP,n}$. Pour l'interprétation $\{p_1 = 10^{-2}, p_2 = 10^{-4}\}$ de l'avis d'expert, le paramètre θ du modèle, initialement estimé à la valeur 0.3710, est forcé de varier avec probabilité de 0.95 à l'intérieur de l'intervalle $[0.3471, 0.6942]$, ce qui donne des quantiles estimés plus faibles c'est-à-dire une estimation sensiblement meilleure de la queue de distribution. Avec une contrainte plus forte sur la queue $\{p_1 = 10^{-3}, p_2 = 10^{-5}\}$, la loi régularisée a comportement encore meilleur en queue de distribution (voir les quantiles estimés et le déplacement de l'intervalle $IC_{ET,BP,n}$ autour de la valeur $\widehat{q}_{ET,n} = 12.2601$) mais sa partie centrale est moins adéquate. En effet, la distance de Cramér-von Mises obtenue alors est de 1.2683, ce qui semble beaucoup comparé à la distance de 0.1789 calculée pour le modèle exponentiel et la valeur critique de 0.2216 correspondant à ce modèle exponentiel. Pour ce jeu de données simulé, nous avons donc réussi, dans le cas le plus adapté $\{p_1 = 10^{-2}, p_2 = 10^{-4}\}$, à construire une loi régularisée acceptée en queue de distribution, sans déterioration importante de la distance de Cramér-von Mises.

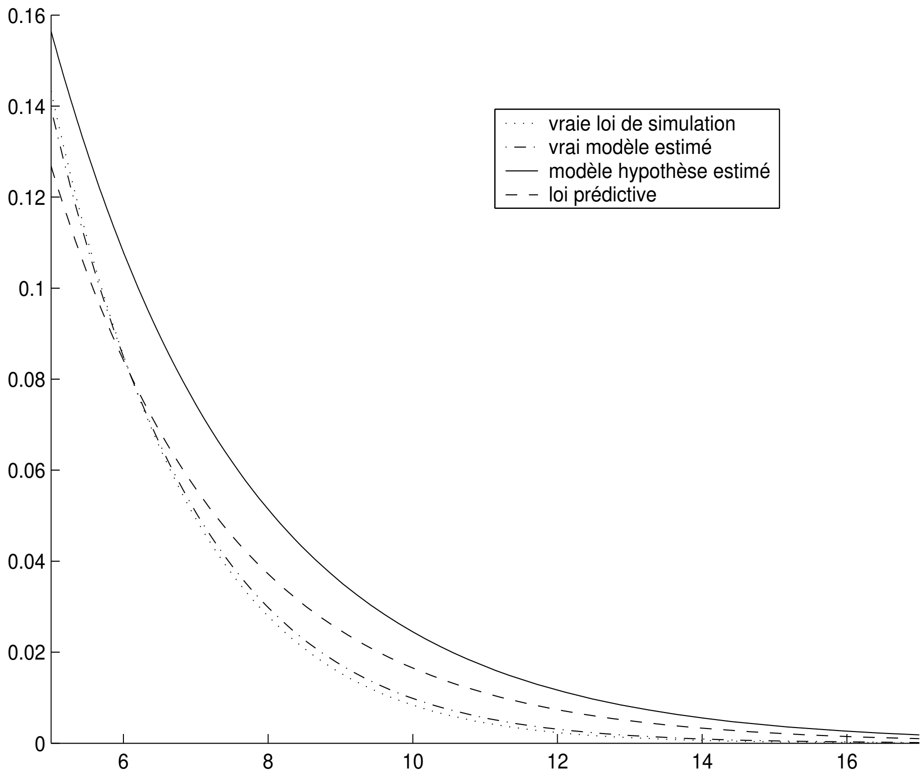

FIG. 2.4 – *Cas exponentiel – fonctions de survie à partir du seuil – pointillés : loi de simulation, discontinu : vrai modèle de Weibull estimé, trait plein : modèle hypothèse exponentiel estimé, tirets : loi prédictive a posteriori obtenue pour l'avis d'expert* $\{p_1 = 10^{-2}, p_2 = 10^{-4}\}$

Le cas normal

Nous avons utilisé une loi de Student à 4 degrés de liberté pour simuler un échantillon de taille 50. Le modèle estimé est $\mathcal{N}(-0.0239, 1.5160)$ avec une distance CVM de 0.1005 pour une valeur critique à 5 % de 0.1248. L'intervalle de confiance du test ET-BP simplifié, calculé à l'ordre $1 - 10^{-3}$ et pour $m_n = 8$ excès, est $IC_{ET,BP,n} = [2.7378, 7.0153]$ et l'estimateur ET est $\widehat{q}_{ET,n} = 8.4589$. Le modèle est donc accepté en central et rejeté en extrêmes. Le tableau 2.3 montre les résultats de la procédure de régularisation. Contrairement au cas précédent, ici la vraie loi a une queue plus lourde que la loi du modèle hypothèse (voir les quantiles) et l'estimateur ET calculé sur l'échantillon simulé est plus grand que la borne supérieure de l'intervalle de confiance bootstrap du test ET-BP simplifié. À nouveau, la procédure de régularisation réussit à attirer la queue de distribution vers de meilleures valeurs. Comme dans le cas exponentiel, la correction est petite et plus les contraintes sur la queue de distribution sont fortes, moins le modèle prédictif est adéquat en partie centrale, mais ici la dégradation est moins sensible. Il ne semble y avoir aucun effet particulier dû à la faible taille de l'échantillon.

	loi prédictive a posteriori avec avis d'expert $p_1 = 10^{-2}$ $p_2 = 10^{-4}$	$p_1 = 10^{-3}$ $p_2 = 10^{-5}$	loi du modèle estimé	vraie loi de simulation	estimateur ET
$q_{0.98}$	4.1666	3.7555	3.0896	2.9985	4.0484
$q_{0.999}$	6.3629	5.7182	4.6609	7.1732	8.4589
$q_{0.9999}$	7.8229	7.0920	5.6141	13.0337	11.8489
θ_1	0.1045	0.1844	$\times$		
θ_2	0.2670	0.3512	$\times$		
d_{CVM}	0.3249	0.2246	0.1005	(valeur critique : 0.1248)	
test ET	acceptée	acceptée	rejetée		
$IC_{ET,BP,n}$	4.5788 11.3167	4.2197 10.0413	2.7378 7.0153		

TAB. 2.3 – *Cas normal* – $n = 50$ – *Résultats de la régularisation*

Le cas Weibull

On utilise la loi $\mathcal{G}amma(3, 3)$ pour les simulations. Le modèle estimé est $\mathcal{W}(1.1325, 1.5246)$ avec une distance CVM de 0.0596 pour une valeur critique à 5 % de 0.1216. En appliquant les deux versions du test ET-BP à l'ordre $1 - 10^{-3}$ et pour $m_n = 20$ excès, on obtient les intervalles $IC_{ET,BP,n} = [3.3442, 5.6401]$ pour l'estimateur ET $\widehat{q}_{ET,n} = 5.3416$, et $IC_{\delta,BP,n} = [-0.3852, 1.2964]$ pour l'erreur d'approximation $\widehat{\delta}_n = 1.3184$. Le modèle de Weibull est donc accepté en central, mais rejeté par le test ET-BP complet, bien qu'accepté par le test ET-BP simplifié. Le tableau 2.4 présente les résultats détaillés de la procédure de régularisation.

	loi prédictive a posteriori pour $p_1 = 10^{-2}$ et $p_2 = 10^{-4}$ — paramètre de forme — cas 1	cas 2	cas 3	param. d'échelle	loi du modèle estimé	vraie loi estimée	vraie loi de simulation	estimateur ET
$q_{0.99}$	3.0700	3.1110	3.1230	3.0330	3.0837	3.3055	2.8020	3.6181
$q_{0.999}$	4.0400	4.1430	4.1710	3.9759	4.0232	4.5742	3.7430	5.3416
$q_{0.9999}$	4.9660	5.1770	5.2360	4.8245	4.8587	5.8052	4.6427	7.0650
θ_1	1.2775	1.2775	1.2775	0.6156	$\times$			
θ_2	1.8573	1.8573	1.8573	1.2312	$\times$			
d_{CVM}	0.0556	0.0585	0.0596	0.0536	0.0596	(valeur critique : 0.1216)		
test ET	acc.	acc.	acc.	acc.	acc.			
$IC_{ET,BP,n}$	3.4471 5.6906	3.4743 5.8036	3.3876 5.8373	3.3709 5.5288	3.3442 5.6401			

TAB. 2.4 – *Cas Weibull* – $n = 100$ – *Résultats de la régularisation*

À nouveau, la correction est légère. Il n'apparait aucun gros changement, ni dans l'estimation des quantiles ni dans la distance de Cramér-von Mises. La correction en queue de distribution semble plus importante et plus adéquate lorsque l'on travaille sur le paramètre de forme de la loi de Weibull. Ce fait semble logique puisque le paramètre de forme est le plus influent sur le comportement en queue de distribution de la loi de Weibull, et qu'ici on ne constate aucune incompatibilité entre la forme des lois de Weibull, le comportement central induit par l'échantillon, et le comportement en queue de distribution induit par l'avis d'expert.

Notons que la régularisation n'implique pas toujours une détérioration de l'adéquation à la partie centrale de la distribution. Dans ce cas par exemple, la distance CVM décroît pour deux des lois prédictives a posteriori calculées. Mais il faut aussi noter ici que la courbe de la fonction de survie de la loi de Weibull estimée (modèle hypothèse) est située entre la courbe de la vraie fonction de survie (loi gamma $\mathcal{G}amma(3,3)$) et la courbe de la fonction de survie estimée par maximum de vraisemblance pour le vrai modèle (gamma) (voir figure 2.5).

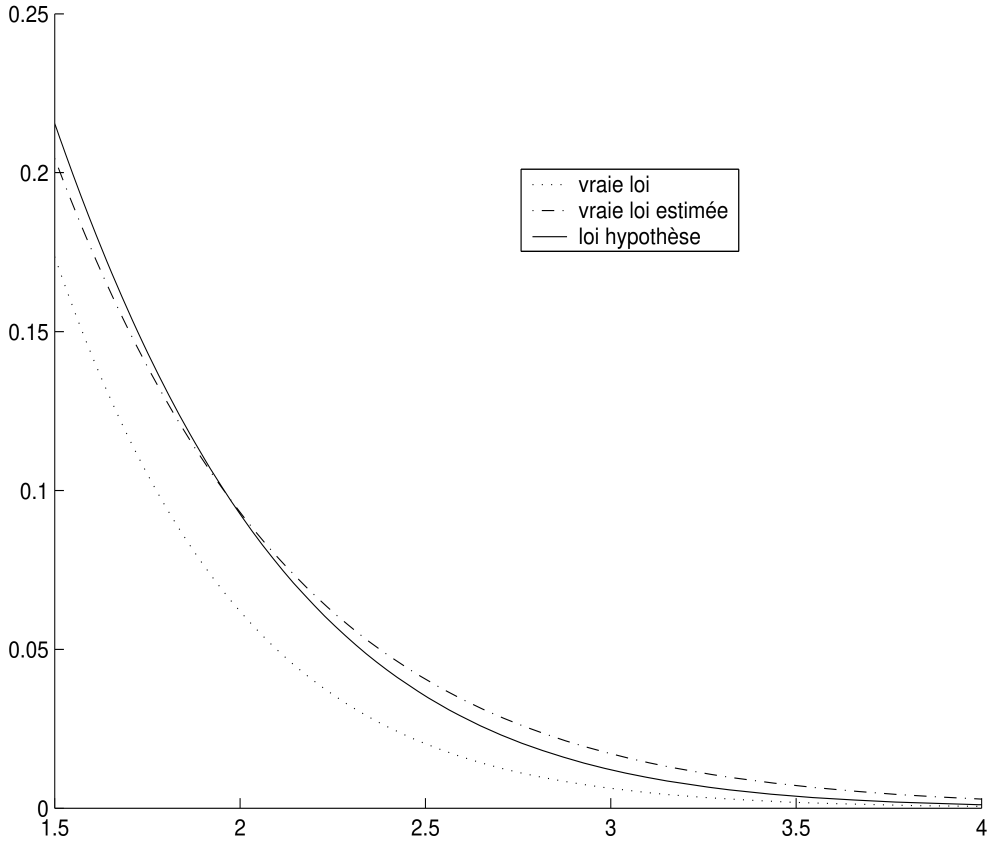

FIG. 2.5 – *Cas Weibull – fonctions de survie à partir du seuil – pointillés : loi de simulation, discontinu : vrai modèle gamma estimé, trait plein : modèle hypothèse de Weibull estimé*

2.1.4 Données réelles

Les données étudiées, de taille $n = 11$, sont des hauteurs de défauts de soudure fournis par EDF :

$$X = [1.80, 2.20, 2.50, 2.60, 2.20, 1.50, 1.70, 2.30, 2.20, 2.50, 1.30].$$

Il s'agit ici d'un jeu de données de taille très petite, ce qui est courant en industrie. Cela représente déjà un challenge pour une étude classique, et encore plus lorsque l'on souhaite étudier la queue de distribution des données, ou contruire un modèle global. Dans ce cas, l'information apportée par l'avis d'expert prend toute son importance. Que l'on souhaite disposer d'un modèle global, ou que l'on s'intéresse uniquement à la queue de distribution, l'avis d'expert fournit des indications sur le comportement extrême des données (que l'échantillon indique d'autant moins qu'il est de petite taille).

Lorsque l'on applique le test de Anderson-Darling ou Cramér-von Mises à ce jeu de données, seule la loi exponentielle est rejetée aux niveaux de signification usuels ($\alpha = 0.01$, 0.05 ou 0.1). La version complète du test ET-BP rejette la même distribution avec $m_n = 4$ excès, $p_n = 1/n$, 0.05 ou 0.01, et pour un niveau de signification $\alpha = 0.05$. L'acceptation de la plupart des lois n'est pas étonnant vu le petit nombre de données dont nous disposons. Le rejet de la loi exponentielle est dû au décalage des données et à l'absence de paramètre de position du modèle exponentiel considéré.

Pour la régularisation bayésienne, l'expert fournit la valeur $q_{\max} = 3.2$ mm et les interprétations suivantes en termes de quantiles : soit $p_1 = 10^{-2}$ et $p_2 = 10^{-3}$, soit $p_1 = 10^{-2}$ et $p_2 = 10^{-4}$. Afin de voir les indications sur la queue de distribution apportée par cet avis d'expert, nous appliquons la procédure de régularisation à différents modèles. Tout d'abord, l'application de la procédure bayésienne au modèle normal donne les résultats présentés dans le tableau 2.5.

		loi prédictive a posteriori		loi du modèle	ET
		$p_1 = 10^{-2}$ $p_2 = 10^{-3}$	$p_1 = 10^{-2}$ $p_2 = 10^{-4}$	$\hat{\mu}_n = 2.0727$ $\hat{\sigma}_n^2 = 0.1882$	
quantiles	0.01	3.0401	2.9527	3.0819	3.1882
	0.001	3.3646	3.2568	3.4133	3.8214
	0.0001	3.6326	3.5154	3.6860	4.4547
intervalle	θ_1	4.2588	4.2588	$\times$	
	θ_2	7.5149	10.8842	$\times$	(valeur
distance CVM		0.0870	0.1097	0.0795	critique :
test ET	réponse	acceptée	acceptée	acceptée	0.1205)
	b_{inf}	2.7286	2.7742	2.7947	
	b_{sup}	5.5590	5.2760	5.6696	

TAB. 2.5 – *Données réelles – Régularisation bayésienne pour le modèle normal*

On constate que les quantiles des deux lois prédictives a posteriori restent très proches de ceux de la loi normale. De plus, la distance de Cramér-von Mises est approximativement la même pour le modèle et les lois prédictives a posteriori. La procédure de régularisation fournit donc des lois prédictives a posteriori proches du modèle. En conséquence, le modèle normal semble être en harmonie avec les données et avec l'avis d'expert.

Nous appliquons ensuite la procédure de régularisation au modèle lognormal. Les résultats sont présentés dans le tableau 2.6. Les quantiles obtenus pour les deux lois prédictives a posteriori sont éloignés de ceux de la loi lognormale; par contre, ils se rapprochent distinctement de ceux du modèle normal. Cela confirme l'accord entre le modèle normal et l'avis d'expert. La figure 2.6 montre les densités des loi normale, et lognormale estimées, ainsi que de la loi prédictive a posteriori issue du modèle lognormal, avec $p_1 = 10^{-2}$ et $p_2 = 10^{-4}$.

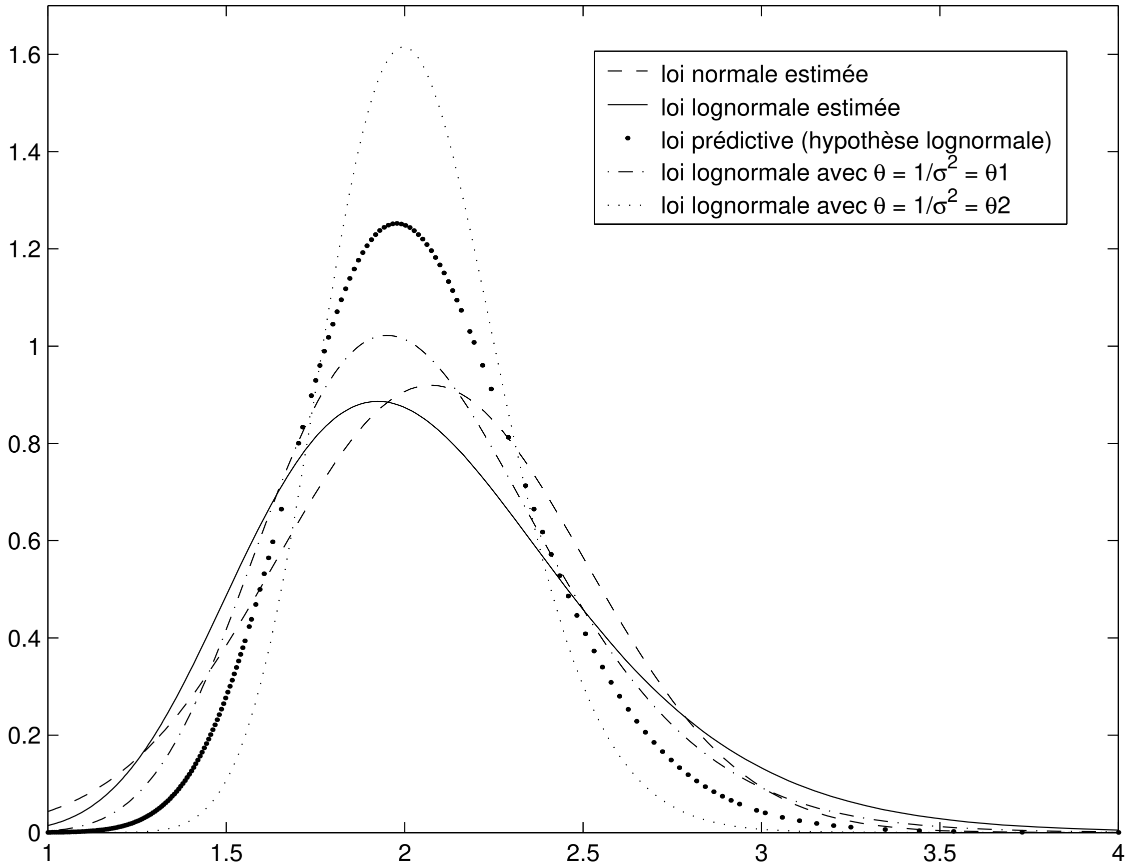

FIG. 2.6 – *Données réelles – comparaison des densités des lois normale, lognormale et prédictive a posteriori – tirets : loi normale estimée, trait plein : loi lognormale estimée, points : loi prédictive a posteriori pour le modèle lognormal avec* $\{p_1 = 10^{-2}, p_2 = 10^{-4}\}$*, discontinu : loi lognormale avec* $\theta = 1/\sigma^2 = \theta_1$*, pointillés : loi lognormale avec* $\theta = 1/\sigma^2 = \theta_2$ *(*θ_1 *et* θ_2 *étant l'interprétation de l'avis d'expert pour le paramètre* $\theta = 1/\sigma^2$ *de la loi lognormale)*

		loi prédictive a posteriori $p_1 = 10^{-2}$, $p_2 = 10^{-3}$	loi prédictive a posteriori $p_1 = 10^{-2}$, $p_2 = 10^{-4}$	loi du modèle $\hat{\mu}_n = 0.7066$, $\hat{\sigma}_n^2 = 0.0519$	ET
quantiles	0.01	3.0233	2.9539	3.4439	3.1882
	0.001	3.4578	3.3648	4.0986	3.8214
	0.0001	3.8693	3.7618	4.7299	4.4547
intervalle	θ_1	25.9642	25.9642	$\times$	
	θ_2	45.8149	66.3561	$\times$	(valeur
distance CVM		0.1715	0.1996	0.0992	critique :
test ET	réponse	acceptée	acceptée	acceptée	0.1205)
	b_{inf}	2.6315	2.5449	3.1024	
	b_{sup}	5.4675	5.2372	6.8167	

TAB. 2.6 – *Données réelles – Régularisation bayésienne pour le modèle lognormal*

Enfin, nous présentons la procédure de régularisation appliquée au paramètre d'échelle du modèle de Weibull. Le tableau 2.7 présente les résultats obtenus avec cette procédure. À nouveau, les quantiles obtenus pour les deux lois prédictives a posteriori s'éloignent de ceux du modèle régularisé (Weibull) et sont proches de ceux de la loi normale. Que ce soit dans le cas d'une surestimation (modèle lognormal) ou d'une sous-estimation (modèle de Weibull) de cette queue normale, l'avis d'expert produit des quantiles de la loi prédictive a posteriori se rapprochant de ceux de la loi normale. Il apparaît donc clairement que selon l'avis d'expert le modèle normal est celui qui s'adapte le mieux à ce jeu de données.

		loi prédictive a posteriori $p_1 = 10^{-2}$, $p_2 = 10^{-3}$	loi prédictive a posteriori $p_1 = 10^{-2}$, $p_2 = 10^{-4}$	loi du modèle $\hat{\lambda}_n = 2.2382$, $\hat{\beta}_n = 6.2877$	ET
quantiles	0.01	3.0827	2.9953	2.8535	3.1882
	0.001	3.2915	3.2032	3.0436	3.8214
	0.0001	3.4492	3.3620	3.1860	4.4547
intervalle	θ_1	0.0031	0.0031	$\times$	
	θ_2	0.0046	0.0061	$\times$	(valeur
distance CVM		0.1696	0.0861	0.0761	critique :
test ET	réponse	accepté	accepté	accepté	0.1169)
	b_{inf}	2.9226	2.8702	2.7753	
	b_{sup}	5.3545	5.4482	5.0847	

TAB. 2.7 – *Données réelles – Régularisation bayésienne pour le cas Weibull paramètre d'échelle*

2.1.5 Conclusion

Dans cette partie, nous avons présenté une procédure de régularisation bayésienne conçue pour améliorer l'adéquation en queue de distribution, selon un avis d'expert. Nos simulations numériques montrent que cela est possible avec seulement une faible dégradation centrale. Cependant, comme prévu, avec cette approche bayésienne la correction en queue de distribution est légère; et pour certains modèles même une forte régularisation ne pourrait suffisamment améliorer la queue. Considérons par exemple un modèle de Weibull dont le comportement en partie centrale (donné par $\widehat{\beta}$) implique un $\beta > 1$ (donc un mode, et une queue de distribution à décroissance rapide) et l'avis d'expert une queue lourde ($\beta < 1$). Selon l'avis d'expert, il faudrait donc alourdir fortement la queue de distribution. Mais les valeurs de β ainsi obtenues impliquent une asymptote verticale, donc un mauvais comportement central (puisqu'on a en fait un mode). À l'inverse, si l'on veut conserver le comportement central ($\beta > 1$), on ne peut obtenir une queue de distribution lourde.

En pratique, cette procédure de régularisation peut être utile pour construire un modèle global dans différents cas :

- Lorsqu'un modèle accepté en partie centrale est rejeté en queue, la procédure propose une meilleure modélisation de la queue de distribution prenant en compte un avis d'expert. Ceci était notre motivation initiale;
- Lorsque la distribution est acceptée à la fois par un test usuel (sur la partie centrale) et un test pour la queue de distribution, la régularisation est une approche bayésienne pour construire un meilleur modèle (principalement pour la queue) en prenant en compte un avis d'expert sur la queue de distribution;
- Lorsque plusieurs distributions sont acceptées à la fois par un test usuel et un test pour la queue de distribution, cette procédure fournit des indications pour aider à la sélection du meilleur modèle selon l'avis d'expert sur les queues de distribution. C'est le cas pour notre jeu de données réelles.

On peut aussi utiliser la procédure de régularisation bayésienne lorsque l'on s'intéresse uniquement à la queue de distribution. Il s'agit alors d'utiliser l'information apportée par l'avis d'expert pour obtenir une meilleure modélisation de la queue de distribution, sans se préocuper du comportement central des modèles estimés.

Notre procédure a été étudiée pour un ensemble de familles usuelles de lois. Cet ensemble peut être élargi. Les lois du domaine d'attraction de Fréchet pourraient être particulièrement d'intérêt. Comme dans le cas d'une loi a priori sur le paramètre de forme de la loi de Weibull, cela conduira probablement à des procédures de calcul très lourdes.

La procédure de régularisation bayésienne, ainsi que le test ET, sont programmés dans un logiciel présenté dans la partie suivante. À travers la présentation de cette maquette, nous détaillons la procédure d'utilisation des méthodes implémentées.

2.2 La maquette logiciel EXTREMES

La maquette logiciel EXTREMES (co-propriété d'EDF et de l'INRIA) a pour but de construire une distribution qui produise une bonne modèlisation des données, en particulier en queue de distribution. Dans certains cas, on s'intéresse seulement à la queue de distribution des données, mais la petite taille des échantillons dont on dispose nous empêche d'utiliser les techniques usuelles d'estimation des événements rares, comme la méthode des excès. Dans d'autres cas, on cherche une distribution qui s'adapte correctement aux données, que ce soit au niveau de l'échantillon, ou au-delà, c'est-à-dire pour des valeurs extrêmes. Il s'agit alors de modéliser la loi des données dans son ensemble.

Une approche naturelle consiste à adapter un certain nombre de lois usuelles aux données, puis tester leur adéquation, éventuellement par des tests classiques (pour l'adéquation centrale), et surtout par des tests en la queue de distribution (pour l'adéquation extrême). L'utilisation de modèle paramétriques classiques peut s'appliquer qu'il s'agisse d'utiliser toute l'information contenue dans les données (dans le cas d'échantillons de petite taille), ou pour déterminer des modèles ayant une bonne adéquation centrale (lorsque l'on recherche un modèle global). Pour l'instant, nous nous sommes limités aux lois usuelles suivantes : normale, exponentielle, gamma, lognormale et de Weibull (appartenant au DA(Gumbel)).

Nous commençons, lorsque cela semble utile, par appliquer un test classique (voir la description des fonctionnalités de la procédure au paragraphe 2.2.2 page 82). Nous avons programmé les tests de Anderson-Darling et Cramér-von Mises. Ces tests usuels nous indiquent si les lois testées modélisent correctement la partie centrale de la loi de l'échantillon, c'est-à-dire les valeurs les plus probables de la variable. Mais ils ne donnent aucune indication quant à la qualité de ces lois pour des événements situés au-delà des observations. Pour vérifier l'adéquation extrême, nous avons implémenté le test ET (défini au chapitre 1), test d'adéquation de ces lois au niveau de la queue de distribution (voir la description des fonctionnalités du module au paragraphe 2.2.4 page 86).

Lorsque l'on recherche un modèle global, dans les cas où certaines lois sont acceptées par un test classique, et où d'autres lois sont acceptées par le test ET, nous utilisons une méthode de régularisation bayésienne (voir le principe dans la partie 2.1 page 60, et la procédure implémentée au paragraphe 2.2.5 page 88). Le but de cette méthode est de modifier légèrement la loi (ou l'une des lois) acceptée par un test classique, de façon à mieux l'adapter au comportement des événements extrêmes. Un avis d'expert est alors utilisé pour contrôler la modification de la queue de distribution.

Il faudra ultérieurement compléter cette maquette en étendant ces méthodes à un catalogue de lois plus important (notamment des lois appartenant au DA(Fréchet)), et en introduisant le test GPD (une extension du test ET) présenté en partie 2.3 (page 91).

2.2.1 Lancer la maquette logiciel EXTREMES

La maquette logiciel EXTREMES est actuellement développée en MATLAB. Il est envisagé de réécrire cette maquette en C, un langage de programmation compilé, donc bien plus rapide que MATLAB, qui est un langage interprété. Plusieurs interfaces seront alors proposées pour une utilisation du module C sous différents environnements, par exemple sous MATLAB, SCILAB, S+ ou EXCEL.

Pour lancer la maquette logiciel EXTREMES, il suffit de lui fournir l'échantillon. Tous les autres paramètres sont demandés au cours de l'exécution des différents modules. Lors du lancement de EXTREMES, on obtient une fenêtre (voir la figure 2.7) qui propose de choisir parmi les différentes procédures de traitement des données implémentées dans la maquette.

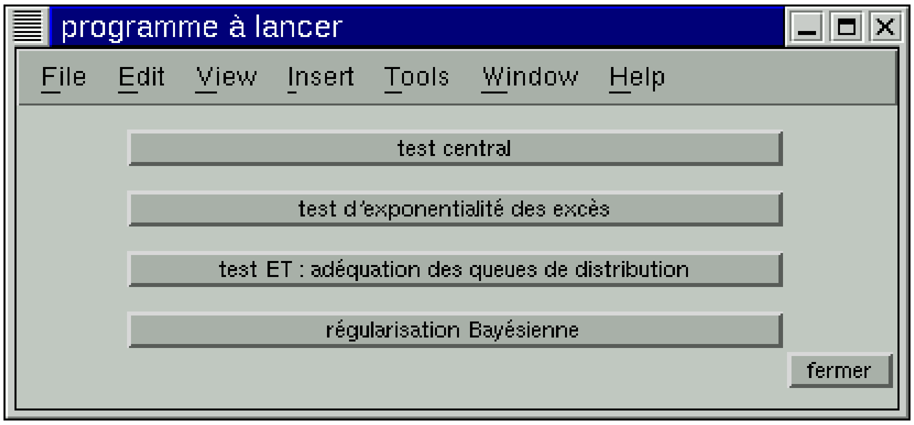

FIG. 2.7 – *Fenêtre de lancement de* EXTREMES *proposant les différents modules au choix*

Lorsque notre but est de déterminer un modèle permettant de représenter au mieux la loi des données dans son ensemble (que ce soit en partie centrale ou en extrême) il est conseillé d'appliquer (si cela est possible et/ou nécessaire) les quatre modules proposés, dans l'ordre de présentation. Si l'on s'intéresse uniquement à la queue de distribution, on peut s'abstenir d'appliquer la première procédure, le test central. Il est alors conseillé d'appliquer dans l'ordre les trois modules suivants. Chacun des quatre boutons de la fenêtre de lancement de EXTREMES donne accès à une nouvelle fenêtre correspondant à la procédure choisie.

Dans le cas où nous souhaitons obtenir un modèle global, il faut en particulier que le modèle soit adapté en partie centrale. Le module *test central* (détaillé au paragraphe 2.2.2 page 82, cliquer sur le premier bouton), permet de contrôler le comportement de la partie centrale d'une loi estimée en appliquant l'un des deux tests usuels programmés : Anderson-Darling ou Cramér-von Mises.

Le test ET d'adéquation à la queue de distribution n'étant défini que dans le DA(Gumbel), il nous faut vérifier que la loi du jeu de données appartient bien à ce domaine avant de pouvoir appliquer le test ET. Le module *test d'exponentialité des excès* (décrit au paragraphe 2.2.3 page 84, cliquer sur le second bouton), nous permet de nous faire une idée sur l'appartenance au DA(Gumbel). Plus précisément, on vérifie la caractéristique du DA(Gumbel) que nous utilisons pour construire le test ET, c'est-à-dire que pour un nombre d'excès m_n adéquat, la loi des excès au-delà du seuil $\hat{u}_n = X_{(n-m_n)}$, la $(n - m_n)$-ème observation ordonnée, est approximativement une loi exponentielle. De plus, ce test d'exponentialité des excès nous permet de choisir un nombre d'excès m_n approprié pour appliquer le test ET.

Lorsqu'on peut supposer que l'on est dans le DA(Gumbel), on peut appliquer la prodédure du *test ET* (décrite au paragraphe 2.2.4 page 86, cliquer sur le troisième bouton), test d'adéquation d'un modèle à la queue de distribution des données. Si l'on recherche un modèle global, le test central nous a déjà permis de vérifier l'adéquation en partie centrale de nos modèles. Le test ET permet alors de la compléter en testant l'adéquation extrême.

Enfin, notamment lorsque l'on cherche un modèle global et qu'aucun des modèles n'est accepté à la fois par un test central et par le test ET, on peut appliquer la procédure de *régularisation bayésienne* (détaillée au paragraphe 2.2.5 page 88, cliquer sur le dernier bouton). Dans ce cas, partant d'un modèle conduisant à une bonne estimation de la partie centrale de la loi, la régularisation bayésienne modifie la queue de distribution, en fonction d'un avis d'expert, pour aboutir à un meilleur modèle global. Lorsque l'on s'intéresse uniquement aux événements rares, on peut souhaiter appliquer la procédure de régularisation bayésienne afin d'obtenir une queue de distribution plus conforme à l'avis d'expert.

2.2.2 Les tests usuels : Anderson-Darling et Cramér-von Mises

On souhaite tout d'abord pouvoir vérifier l'adéquation de nos modèles à la partie centrale de la loi des données. La fenêtre correspondant à cette procédure de *test central* est présentée en figure 2.8.

Tout d'abord, sont proposés les deux tests implémentés : le test de Anderson-Darling et le test de Cramér-von Mises. Si l'on souhaite dès à présent prendre un peu en compte les valeurs les plus extrêmes de l'échantillon (qui donnent une première idée de l'ajustement en queue), on utilise de préférence le test de Anderson-Darling (choisi par défaut). Au contraire, on peut considérer que le comportement de queue sera exploré lors de l'application du test ET. Dans ce cas, on préfère se focaliser pour le test central sur les valeurs les plus probables de la variable et on utilise plutôt le test de Cramér-von Mises.

D'autre part, il faut choisir le niveau α du test appliqué. Les niveaux proposés sont ceux tabulés dans le D'Agostino et Stephens [17] (chapitre 4) pour toutes les lois étudiées ici (normale, lognormale, exponentielle, gamma et de Weibull). La valeur $\alpha = 0.05$, l'une des plus courantes pour le niveau de signification d'un test, est la valeur par défaut. Sont aussi proposées les valeurs $\alpha = 0.25$, 0.1, 0.025 et 0.01.

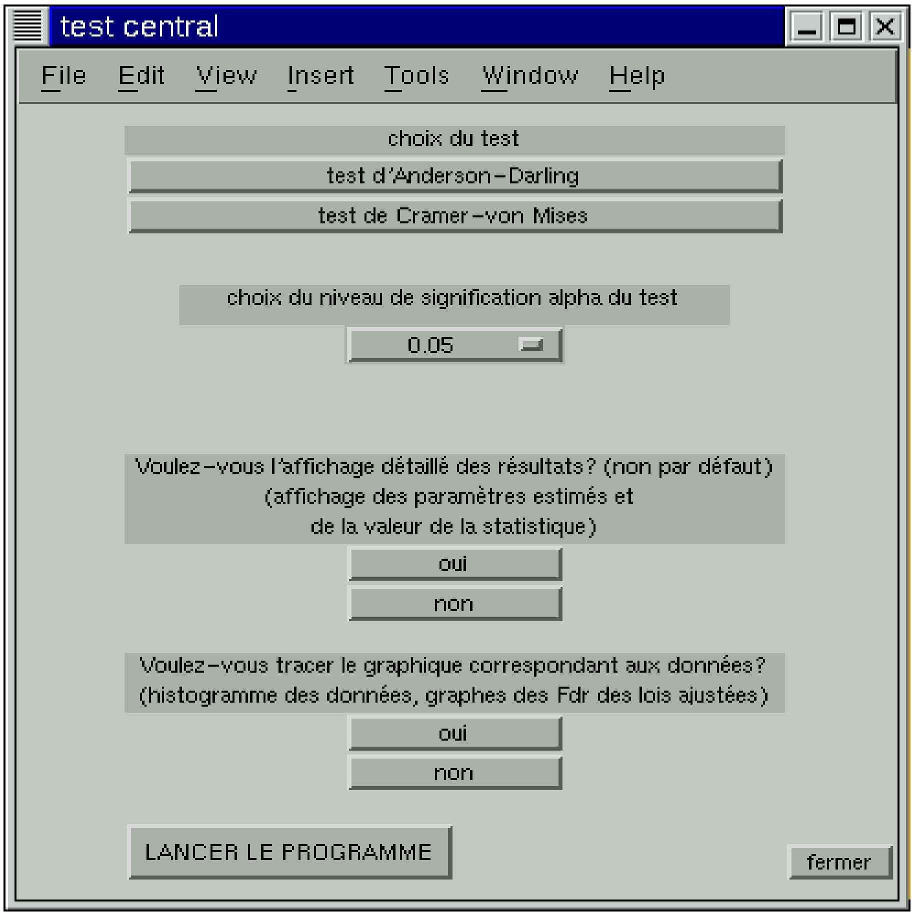

FIG. 2.8 – *Fenêtre pour la procédure* test central

Lorsque l'affichage n'est pas détaillé, la maquette logiciel indique seulement pour chacune des lois testées si elle est acceptée ou rejetée. Dans le cas contraire, il est rappelé lequel des deux tests a été appliqué et pour quel niveau de signification α. En outre, il est précisé la valeur des estimateurs du maximum de vraisemblance des paramètres du modèle, la valeur de la statistique de test, ainsi que la valeur critique correspondante.

Enfin, le programme permet de tracer des graphiques correspondant aux données et aux différents modèles estimés (voir la figure 2.9). Sur un premier graphique, on trace les densités des modèles estimés, ainsi que l'histogramme des données, normalisé pour que sa surface soit égale à 1. Cette normalisation permet d'obtenir un histogramme à la même échelle que les densités, comme on peut le constater sur la première partie de la figure 2.9, et donc de pouvoir comparer leurs formes. Le deuxième graphique fournit le tracé de la fonction de répartition empirique, ainsi que celui des fonctions de répartition estimées pour les modèles testés (voir la deuxième partie de la figure 2.9) que l'on peut aussi comparer.

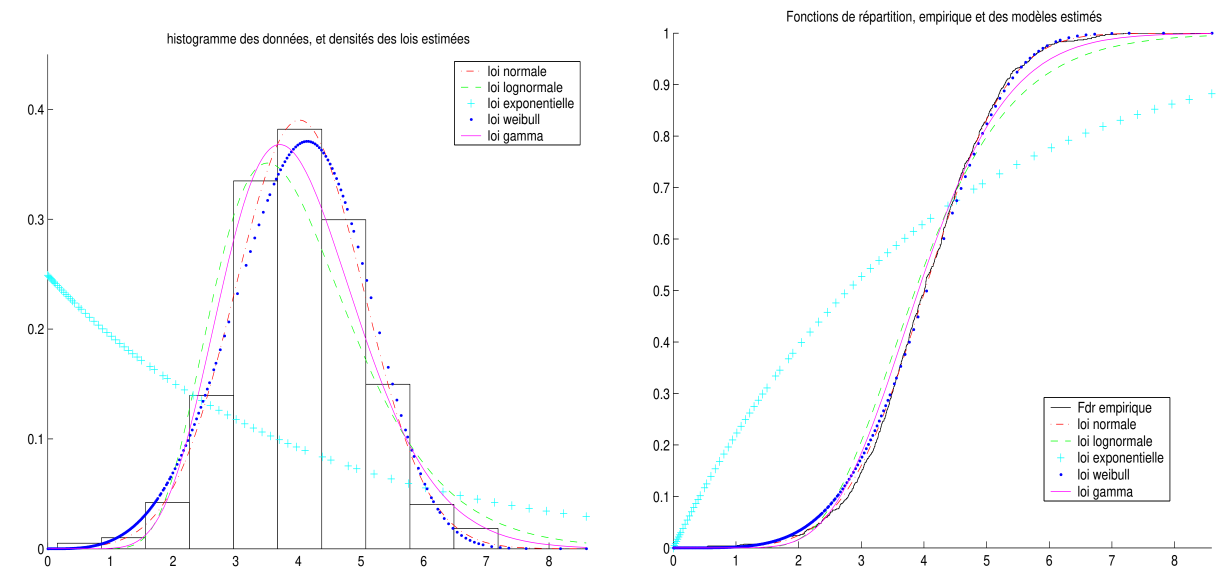

FIG. 2.9 – *Exemple de sortie graphique de la procédure* test central*: à gauche, tracé des densités estimées et de l'histogramme normalisé; à droite, tracé des fonctions de répartition estimées et empirique*

Remarque 2.2 *Les différentes procédures numériques permettant de calculer les estimateurs du maximum de vraisemblance des paramètres de la loi de Weibull peuvent toutes donner (dans quelques cas particuliers, différents selon les méthodes) des estimateurs erronés. On peut contrôler graphiquement (par exemple sur le graphe des densités et de l'histogramme normalisé) si ces paramètres semblent correctement estimés. Lorsqu'ils sont visiblement mal estimés, il faut changer de procédure numérique d'estimation.*

2.2.3 Appartenance au Domaine d'attraction de Gumbel : le test d'exponentialité pour les excès

Cette maquette a principalement été conçue pour la prévision des événements rares. Le principal problème est donc de vérifier l'adéquation extrême. Le test ET a été créé à cet effet. Mais il ne peut être appliqué que dans le DA(Gumbel). Il faut donc s'assurer que l'on est dans ce domaine.

Lorsqu'on a construit le test ET (voir le chapitre 1), on a utilisé la propriété suivante du DA(Gumbel) : pour tout échantillon issu d'une loi appartenant au DA(Gumbel), et pour un seuil convenablement choisi, la loi des excès au-dessus de ce seuil tend vers une loi exponentielle, lorsque la taille de l'échantillon, et donc le seuil, tendent vers l'infini. On travaille ici à distance finie : la taille de l'échantillon est fixée, ainsi que le seuil qui, pour m_n excès, correspond à la $(n-m_n)$-ème observation ordonnée. Mais on peut vérifier si, pour un nombre d'excès donné (c'est-à-dire un seuil donné), la loi des excès est approximativement une loi exponentielle (on peut expérimenter plusieurs valeurs de m_n, c'est-à-dire plusieurs valeurs du seuil). Il suffit pour cela d'appliquer un test classique (par exemple le test de Anderson-Darling) pour vérifier l'adéquation de la loi exponentielle à l'échantillon des excès. Pour appliquer le *test d'exponentialité des excès*, on doit préciser le niveau de signification α du test, ainsi que le nombre d'excès m_n, qui doit être petit devant n. Pour déterminer m_n (qui est ensuite utilisé pour appliquer le test ET), il est conseillé de s'aider des tableaux 1.7 à 1.12 de la partie 1.3.3 (page 44) qui présentent les valeurs de m_n pour lesquelles le niveau et la puissance de chaque version du test ET sont convenables.

Si l'on rejette l'exponentialité des excès, cela ne signifie pas forcément que l'on n'est pas dans le DA(Gumbel). On peut simplement, par exemple, avoir choisi un nombre d'excès inadéquat. Avant de rejeter l'appartenance au DA(Gumbel), il faut d'abord appliquer le test pour différents nombres d'excès (petits devant n, voir les valeurs de m_n conseillées pour appliquer le test ET en partie 1.3.3 page 44). Lorsque l'hypothèse d'exponentialité des excès est rejetée pour toutes les valeurs de m_n testées, on peut supposer que l'on n'est pas dans le DA(Gumbel). Dans ce cas, puisque les lois testées appartiennent au DA(Gumbel), on sait, sans même appliquer le test ET, qu'elles ne peuvent pas être adaptées. On peut alors, soit appliquer la procédure de régularisation bayésienne (en particulier si l'on dispose d'un avis d'expert sur la queue de distribution), soit utiliser d'autres modèles n'appartenant pas au DA(Gumbel) (par exemple la loi uniforme ou la loi beta appartenant au domaine d'attraction de Weibull, ou, dans le domaine d'attraction de Fréchet, les lois de Fréchet, Student ou Pareto). Dans les deux cas, le nouveau modèle (comme les données puisque l'exponentialité des excès a été rejetée) n'appartient pas au DA(Gumbel)), et pour vérifier l'adéquation extrême, il faut utiliser de préférence le test GPD, plus général, décrit dans la partie 2.3 (page 91), à condition que l'on dispose d'assez de données et donc d'excès pour estimer les deux paramètres de la loi GPD.

Remarque 2.3 *D'autres méthodes ont été proposées pour, d'une part vérifier l'appartenance à l'un des domaines d'attraction des valeurs extrêmes, et d'autre part aider au choix d'une valeur adéquate pour le nombre d'excès m_n (en particulier dans le cadre de l'estimation des quantiles extrêmes). Signalons deux méthodes graphiques :*

- *La méthode du* Quantile-Quantile plot *(*QQplot*) (décrite par Embrechts* et al. *[34] paragraphe 6.2.1, page 290) indique l'appartenance probable à l'un des trois domaines d'attraction des valeurs extrêmes. Il s'agit de tracer les quantiles de la loi exponentielle standard $(-\ln(i/m_n))_{i=1,\ldots,m_n}$ versus les excès ordonnés donnés par $(X_{(n-m_n+i)} - X_{(n-m_n)})_{i=1,\ldots,m_n}$. On suppose que le nombre d'excès utilisé est adéquat. Sous cette condition, si les données appartiennent au DA(Gumbel), alors les excès sont approximativement de loi exponentielle. Les points du graphe du QQplot sont donc approximativement alignés et on peut calculer les paramètres (d'échelle et de position) de la loi exponentielle correspondante par régression linéaire. Lorsque les données n'appartiennent pas au DA(Gumbel), on observe une incurvation du QQplot, vers le haut si* *les données sont dans le domaine d'attraction de Fréchet (apparition d'une convexité), vers la droite si les données sont dans le domaine d'attraction de Weibull (apparition d'une concavité). Une illustration est donnée par la figure 2.10.*

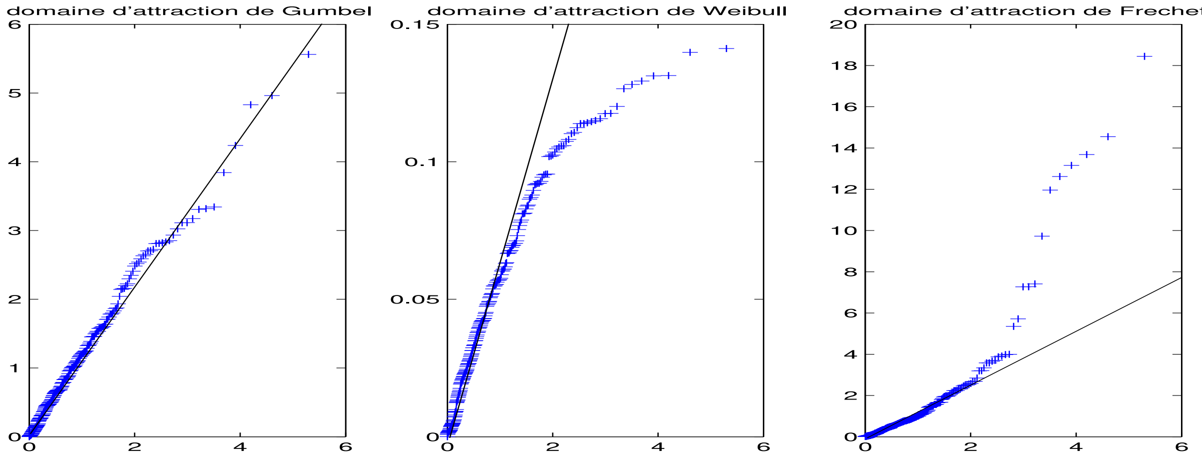

FIG. 2.10 – *Exemples de qqplot dans le domaine d'attraction de Gumbel (à gauche, pour un échantillon de loi normale), dans le domaine d'attraction de Weibull (au centre, pour un échantillon de loi beta) et dans le domaine d'attraction de Fréchet (à droite, pour un échantillon de loi de Student) pour des échantillons de taille* $n = 1000$ *et* $m_n = 200$ *excès*

- *Pour choisir graphiquement un nombre d'excès adéquat pour l'estimation des quantiles extrêmes, on peut utiliser la méthode du* Pareto plot *(ou* "horror plot"*) décrite par Embrechts* et al. *[34] (pages 357 à 365). Il s'agit de tracer l'estimateur ET (ou GPD) d'un quantile extrême en fonction du nombre d'excès. Il est alors conseillé de choisir le nombre d'excès utilisé ensuite parmi les valeurs de* m_n *pour lesquelles on observe une stabilisation de l'estimateur du quantile.*

2.2.4 Le test ET

Supposons que le test d'exponentialité des excès nous ait permis de déterminer des valeurs du nombre d'excès pouvant être utilisées pour appliquer le test ET. On souhaite à présent déterminer si l'un des modèles envisagés permet de correctement modéliser la queue de distribution. Il faut donc vérifier l'adéquation extrême de ces modèles, c'est-à-dire leur appliquer le test ET. La fenêtre correspondant à cette procédure est présentée en figure 2.11.

Il existe trois versions du test ET, l'une basée sur la loi asymptotique de l'estimateur ET $\hat{q}_{ET}$, les deux autres basées sur la méthode du bootstrap paramétrique pour la simulation des fluctuations d'échantillonnage. Il nous faut donc d'abord choisir entre ces deux méthodes de construction du test. La version du test basée sur la loi asymptotique s'étant révélée très peu puissante lors des essais, on préfère utiliser les versions basées sur la méthode du bootstrap (proposées par défaut). Dans ce cas, il faut aussi choisir entre le test ET-BP complet et le test ET-BP simplifié. La version complète étant la plus puissante, nous conseillons de préférence son utilisation, et elle est proposée par défaut. La version simplifiée est plus rapide et présente un intérêt principalement lorsque les estimateurs du maximum de vraisemblance des paramètres de la loi testée sont longs à calculer, par exemple pour les lois de mélange.

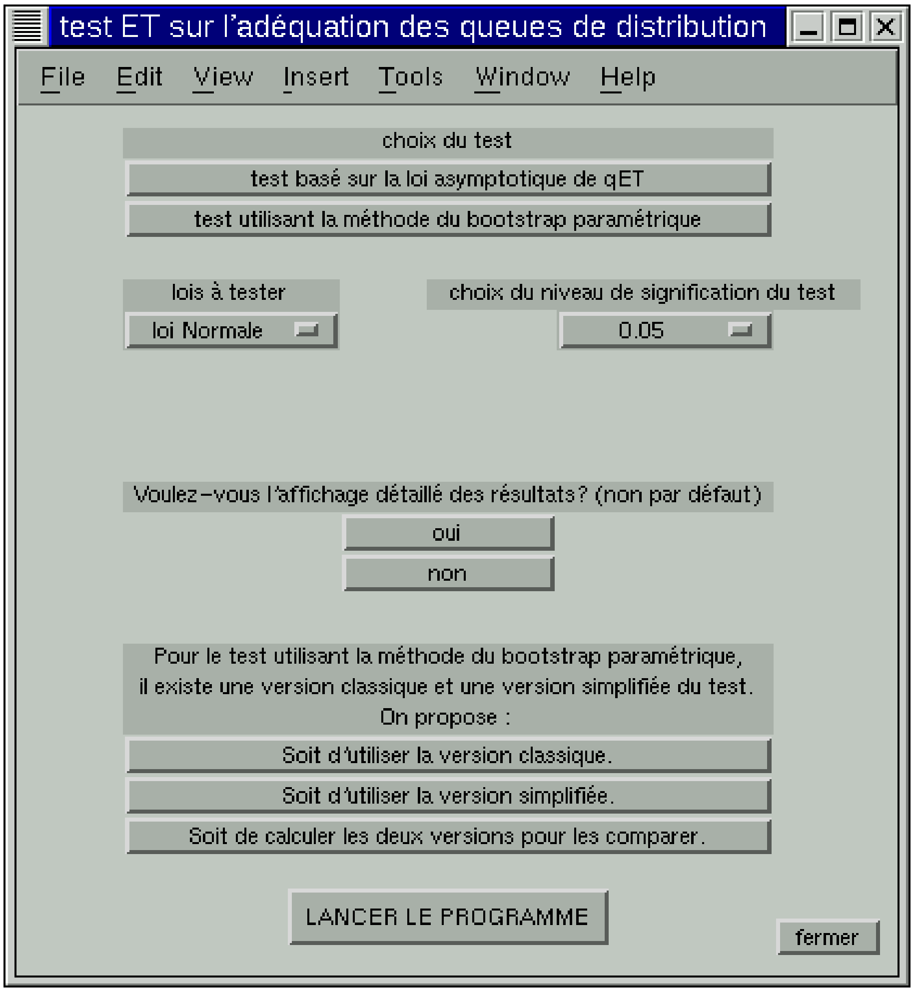

FIG. 2.11 – *Fenêtre pour la procédure* test ET

Il nous faut ensuite choisir le modèle auquel on souhaite appliquer le test (normal, lognormal, exponentiel, gamma, Weibull, ou les cinq modèles simultanément) ainsi que le niveau de signification du test ($\alpha = 0.25$, 0.1, 0.05, 0.025 ou 0.01, comme pour le test central). On applique généralement le test ET séparément pour chacun des modèles envisagés, puisque le nombre d'excès (précisé ultérieurement) pour lequel il est conseillé de l'employer varie souvent en fonction du modèle testé. De plus, le résultat du test ET-BP est obtenu nettement plus lentement dans les cas gamma et Weibull, surtout pour la version complète du test ET-BP. C'est une conséquence de la longueur de l'évaluation de leurs estimateurs du maximum de vraisemblance et de leurs quantiles. La lenteur de calcul du test ET-BP complet pour les modèles gamma et de Weibull permet de comprendre l'importance de disposer du test ET-BP simplifié (beaucoup plus rapide) dans des cas plus calculatoires, comme le cas des modèles de mélange.

Le *test ET* ne peut être appliqué que lorsque les excès au-dessus d'un seuil peuvent être considérés comme étant approximativement de loi exponentielle. On applique donc le test ET avec un nombre d'excès m_n déduit des tableaux 1.7 à 1.12, pages 45 à 51, (qui permettent de choisir des valeurs de m_n pour lesquelles le niveau et la puissances du test ET sont convenables) et pour lequel l'exponentialité des excès a été acceptée par le test précédent. Les tableaux 1.7 à 1.12 permettent aussi de choisir l'ordre du quantile pour lequel appliquer la version du test ET choisie de sorte que sa puissance soit maximisée.

Enfin, pour les versions du test ET basées sur la méthode du bootstrap paramétrique, il faut préciser le nombre N d'échantillons bootstrap que l'on utilise pour construire l'intervalle de confiance du test. Il est évident que les intervalles de confiance sont plus stables lorsqu'on augmente N. En effet, plus on effectue de simulations, plus on diminue l'amplitude des fluctuations des quantités calculées à partir de ces simulations. Cependant, pour des raisons de temps de calcul, il n'est pas rentable de trop augmenter le nombre N de simulations bootstrap. Remarquons que la construction d'un intervalle de confiance à partir de valeurs bootstrappées nécessite d'ôter $[N\alpha/2]$ valeurs de chaque côté de l'échantillon ordonné, où $[a]$ désigne la partie entière de a. Afin que l'intervalle de confiance exclue les cas les plus atypiques, on souhaite effectivement ôter les valeurs les plus extrêmes de l'échantillon simulé. On conseille donc de choisir N de telle sorte que la partie entière de $N\alpha/2$ soit strictement positive ($[N\alpha/2] > 0$), d'ordinaire au moins de l'ordre de 2 ou 3. Les choix $N = 500$ ou 1000 semblent en général raisonnables.

L'affichage simplifié rappelle le nombre d'excès choisi, la version du test ET utilisée et le modèle testé, puis donne le résultat de ce test. Lors d'un affichage plus détaillé, on précise les valeurs des estimateurs ET $\hat{q}_{ET}$ et paramétrique $\hat{q}_{\text{param}}$, ainsi que leur différence $\hat{\delta}_n$ (estimation de l'erreur d'approximation δ). On indique aussi l'intervalle de confiance utilisé pour le test, en précisant la quantité sur laquelle il porte.

2.2.5 La régularisation bayésienne

Lorsqu'à des fins d'exploration, nous souhaitons appliquer la régularisation bayésienne à des échantillons simulés, la première difficulté consiste à simuler un échantillon pour lequel un modèle déterminé (le modèle hypothèse pour la régularisation) est accepté par un test central et rejeté par un test extrême (pour un nombre d'excès et un quantile déterminés). Pour cela, il est nécessaire que la fonction de répartition de la loi de simulation de l'échantillon soit proche, en partie centrale, de celle d'une loi de la famille hypothèse. On tire ensuite des échantillons selon cette loi de simulation jusqu'à en obtenir un tel que le modèle hypothèse soit accepté par un test central mais rejeté en extrêmes par le test ET. C'est ainsi qu'ont été simulé les échantillons du paragraphe 2.1.3 (page 70). D'autre part, puisque l'on connait la vraie loi (la loi de simulation), on peut simuler un avis d'expert, en calculant le quantile d'ordre $1-p$ de cette vraie loi (avec $p \leq 1/n$ pour que ce soit un quantile extrême), et en l'interprétant comme un quantile respectivement d'ordre $1-p_1$ et $1-p_2$. En général, on les choisit tels que p soit compris entre p_1 et p_2.

On peut à présent appliquer la procédure de régularisation bayésienne à l'un des échantillons simulés selon la méthode précédente, ou à un jeu de données réelles. La fenêtre correspondante est présentée en figure 2.12. Pour appliquer la procédure de régularisation bayésienne (cliquer sur le bouton du même nom), il faut tout d'abord indiquer le modèle hypothèse, celui qui va être régularisé, parmi les modèles normal, lognormal, exponentiel, gamma et de Weibull. Lors de la procédure de régularisation bayésienne, on met une loi a priori sur l'un des paramètres (θ) du modèle étudié, qui est de ce fait considéré comme aléatoire. Pour la loi de Weibull, il faut préciser si le paramètre aléatoire θ sur lequel on travaille est le paramètre de forme ou le paramètre d'échelle, les deux cas ayant été envisagés.

FIG. 2.12 – *Fenêtre pour la procédure de* régularisation bayésienne

Il faut ensuite intégrer l'avis d'expert, c'est-à-dire entrer une valeur extrême $q_{\max}$ (qui a un faible risque p d'être dépassée) ainsi qu'un encadrement de la probabilité p de dépasser cette valeur (p_1 et p_2, valeurs inférieures ou égales à $1/n$). De plus, on précise la confiance que l'on a en l'avis d'expert. Pour les cas normal, lognormal, exponentiel, gamma et Weibull paramètre d'échelle, on introduit pour cela une valeur ε (proche de 0) correspondant, pour la loi a priori, à la proportion de valeurs du paramètre aléatoire θ se situant à l'extérieur de l'intervalle $[\theta_1, \theta_2]$ déterminé grâce à l'avis d'expert. Pour le cas Weibull paramètre de forme, nous avons envisagé trois cas : le cas d'une confiance élevée en l'avis d'expert (on utilise à nouveau ε proche de 0), le cas d'une confiance moyenne, et le cas d'une confiance faible en l'avis d'expert.

Nous souhaitons à présent pouvoir comparer le modèle, la (les différentes) loi(s) prédictive(s), éventuellement la loi de simulation et le modèle de simulation estimé. Pour cela, on peut calculer différents quantiles extrêmes pour des ordres $1 - p_i$ tels que $p_i \leq 1/n$ pour tout i. Cela permet de mettre en évidence les modifications induites par l'avis d'expert sur la queue de distribution, ainsi que, dans le cas de données simulées, l'efficacité de la correction en queue de distribution. D'autre part, afin de visualiser ces différents changements, on peut tracer les densités, ou les fonctions de répartition (ou de survie) du modèle estimé, de la (les) loi(s) prédictive(s), éventuellement de la loi de simulation et du modèle de simulation estimé.

À présent, nous souhaitons donner une idée quantitative de l'adéquation centrale (elle est en général détériorée, mais cela n'est pas toujours le cas). Nous souhaiterions donc appliquer l'un des test usuels. Cependant, les modèles prédictifs ne sont généralement pas des modèles classiques, et leurs paramètres ne sont pas estimés par une méthode usuelle. Nous ne disposons donc pas de valeurs de rejet tabulées pour les différentes lois prédictives. Nous pouvons seulement calculer la statistique d'un test classique (par exemple Anderson-Darling ou Cramér-von Mises) pour la loi prédictive et la comparer à la statistique de test calculée pour le modèle hypothèse (et éventuellement à la valeur de rejet correspondante pour le modèle, en précisant un niveau α). On suppose que lorsque la statistique de test n'augmente pas trop entre le modèle hypothèse et la loi prédictive, le comportement central de la loi prédictive reste correct. Cette évaluation est cependant subjective et laissée à l'appréciation des utilisateurs.

D'autre part, nous souhaitons vérifier l'adéquation extrême (elle devrait être améliorée) de la loi prédictive. Les différentes lois prédictives auxquelles nous aboutissons n'appartiennent pas au DA(Gumbel). En toute rigueur, le test ET n'est donc pas applicable à ces lois. Cependant, le test GPD défini à l'extérieur du DA(Gumbel) (voir la partie 2.3 page 91), n'est pas encore implémenté dans la maquette logiciel car il est encore en cours d'étude. En conséquence (voir la remarque 2.1 page 71), puisque la loi prédictive reste relativement proche du modèle initial, qui appartient au DA(Gumbel), et que l'exponentialité des excès a été acceptée pour l'échantillon étudié, nous appliquons, en première approximation, le test ET pour vérifier l'adéquation extrême de la loi prédictive. Plus précisément, nous appliquons le test ET-BP simplifié pour le même nombre d'excès m_n (puique l'exponentialité des excès a été acceptée) et le même quantile d'ordre $1 - p$ que pour le modèle hypothèse, et pour un nombre N d'échantillons bootstrap et un niveau α à préciser.

Enfin, nous avons intégré la possibilité de simuler un échantillon de taille n_{pred} (à choisir par l'utilisateur) selon la loi prédictive. Étant donné que la loi prédictive s'exprime sous forme de mélange continu du modèle selon la loi a posteriori, lorsque la loi a posteriori est une loi classique (cas normal, lognormal, exponentiel, gamma et Weibull paramètre d'échelle), la simulation est immédiate. Par contre, dans le cas Weibull paramètre de forme, la loi a posteriori n'a pas de forme analytique et doit être simulée par la méthode d'acceptation-rejet (voir paragraphe 2.1.2.3 page 70). La simulation selon la loi a posteriori, et en conséquence selon la loi prédictive, est donc bien plus longue dans ce cas.

Nous avons constaté, notamment lorsque l'on veut vérifier l'adéquation extrême de la loi prédictive issue de la procédure de régularisation bayésienne, les limites du test ET qui n'est défini que dans le DA(Gumbel). Nous proposons donc dans la partie suivante une version étendue de ce test, le test GPD, définie dans les trois domaines d'attraction des valeurs extrêmes.

2.3 Le test GPD, premier résultats

Nous avons vu au paragraphe 1.5 (page 57) que le test ET présente principalement deux inconvénients. Premièrement, lorsqu'aucune loi n'est acceptée à la fois par un test central et un test ET, comment trouver un modèle global c'est-à-dire adapté à la partie centrale et à la queue de la distribution des données ? Une solution à ce problème, la procédure de régularisation bayésienne, a été proposée au paragraphe 2.1 (page 60). L'autre problème que nous nous proposons de résoudre dans ce paragraphe est le fait que le test ET ne peut s'appliquer qu'à des lois appartenant au DA(Gumbel). Nous souhaitons donc construire une version étendue (le test GPD) de ce test d'adéquation de queue de distribution, qui soit défini dans les trois domaines d'attraction des valeurs extrêmes : Fréchet, Gumbel et Weibull. De plus, les résultats sur l'approximation pénultième de Rym Worms-Ramdani [43, 49] indiquent que les lois des excès d'une importante classe de lois du DA(Gumbel) sont mieux approchées par des lois GPD avec $\gamma \neq 0$ que par des lois exponentielles. On peut donc penser que même dans le DA(Gumbel), le test GPD (basé sur une approximation GPD de la loi des excès) peut être plus approprié que le test ET (basé sur une approximation exponentielle de cette même loi).

On dispose d'un échantillon de données $(x_1, \dots, x_n)$ issu de variables aléatoires réelles i.i.d. $X_1, \dots, X_n$ de fonction de répartition (Fdr) F inconnue. Nous voulons vérifier si un modèle paramétrique $\{F_\theta, \theta \in \Theta\}$ permet d'obtenir une bonne approximation de la loi des données, particulièrement en queue de distribution. Plus précisément, nous vérifions l'adéquation aux données de la queue de distribution de la loi de Fdr $F_{\widehat{\theta}_n}$, dont les paramètres sont estimés (par exemple) par la méthode du maximum de vraisemblance. Les hypothèses du test se ramènent donc à

$$\mathcal{H}_0 : F = F_{\widehat{\theta}_n} \quad \text{contre} \quad \mathcal{H}_1 : F \neq F_{\widehat{\theta}_n}.$$

Le test GPD est défini de la même façon que le test ET, et s'appuie donc lui aussi sur la comparaison de deux estimateurs (l'un paramétrique, l'autre non paramétrique) d'un quantile extrême de la loi des données. Au paragraphe 2.3.1 (page 92), nous commençons par compléter la présentation de la méthode des excès pour l'estimation des quantiles extrêmes décrite au paragraphe 1.1.1 (page 11) : nous présentons la méthode GPD d'estimation non paramétrique, une extension de la méthode ET à tous les domaines d'attraction.

Nous avons défini trois versions différentes du test ET, mais la version asymptotique du test s'étant révélée peu puissante, dans le cadre plus général de ce paragraphe, nous utilisons directement la méthode du bootstrap paramétrique pour définir le test GPD (voir paragraphe 2.3.2 page 94). De même que pour le test ET, nous proposons une version bootstrap complète, ainsi qu'une version bootstrap simplifiée du test GPD.

2.3.1 Méthode des excès pour l'estimation des quantiles extrêmes, compléments

Nous souhaitons ici estimer un quantile extrême, c'est-à-dire un quantile généralement situé au-delà des observations. Choisissons un nombre p_n positif et inférieur à $1/n$, de sorte que le quantile q_{1-p_n}, d'odre $1-p_n$, correspondant soit en général supérieur à $x_{(n)}$ l'observation maximale, estimation du quantile d'ordre $1/n$. On s'intéresse plus précisément à $\widehat{q}_{GPD,n}$, l'estimateur non paramétrique du quantile basée sur la méthode des excès.

Soit u un réel suffisamment élevé appelé *seuil*. On définit les excès au-delà du seuil u comme l'ensemble des variables aléatoires $\{Y_j\} = \{X_j - u\,;\ X_j > u\}$. La fonction de répartition des excès au-delà du seuil u est

$$F_u(y) = P(X - u \leq y | X > u) = \frac{F(u+y) - F(u)}{1 - F(u)}. \tag{2.1}$$

Le théorème de Pickands [2, 3, 41] (décrit dans le cas particulier du DA(Gumbel) dans le théorème 4 page 11) nous permet de construire une approximation GPD du quantile.

Théorème 10 (théorème de Pickands) *Soit F appartenant à l'un des trois domaines d'attraction des valeurs extrêmes, DA(Fréchet), DA(Gumbel) ou DA(Weibull). Alors il existe une fonction $\sigma(u)$ positive, définie à une équivalence près*[1] *quand $u \to \omega(F)$, où $\omega(F) = \sup\{x : F(x) < 1\}$ est le point terminal de la Fdr F; et un réel γ tels que*

$$\lim_{u \to \omega(F)} \sup_{0<y<\omega(F)-u} \left| F_u(y) - G_{\gamma,\sigma(u)}(y) \right| = 0, \tag{2.2}$$

où $G_{\gamma,\sigma}(x)$ est la fonction de répartition de la loi de Pareto généralisée (Generalised Pareto Distribution : GPD), définie pour $\sigma > 0$ par

$$G_{\gamma,\sigma}(x) = \begin{cases} 1 - \left(1 + \dfrac{\gamma x}{\sigma}\right)^{-1/\gamma} & si \quad \gamma \neq 0 \\ 1 - \exp(-x/\sigma) & si \quad \gamma = 0 \end{cases}$$

pour $x \in \mathbb{R}_+$ si $\gamma \geq 0$ et $x \in [0, -\sigma/\gamma]$ si $\gamma < 0$.

1. Si $\sigma_1(u)$ est une version de $\sigma(u)$ et $\sigma_2(u)$ est telle que $\lim_{u\to\omega(F)} \sigma_1(u)/\sigma_2(u) = 1$ alors $\sigma_2(u)$ est une autre version de $\sigma(u)$.

Pour appliquer ce théorème, on choisit le seuil comme étant le quantile d'ordre $1 - m_n/n$ de la loi des données : $u_n = F^{\leftarrow}(1 - m_n/n)$, où $F^{\leftarrow}$ est l'inverse généralisée de la Fdr F. Le nombre d'excès m_n est un entier fixé tel que $1 < m_n < n$, tendant vers l'infini avec la taille n d'échantillon, mais restant petit devant n :

$$\lim_{n\to\infty} m_n = +\infty \qquad \text{et} \qquad \lim_{n\to\infty} \frac{m_n}{n} = 0\,. \tag{2.3}$$

Pour construire une approximation GPD du quantile d'ordre $1 - p_n$, on utilise la décomposition suivante (déduite de l'expression de la Fdr des excès donnée par l'équation (2.2)) :

$$p_n = \mathrm{P}(X > q_{1-p_n}) = (1 - F(u_n))(1 - F_{u_n}(q_{1-p_n} - u_n)).$$

Or par définition de u_n, on a

$$1 - F(u_n) \leq \frac{m_n}{n} \leq \lim_{u\to u_n,\, u>u_n} (1 - F(u)),$$

les égalités étant vérifiées lorsque $1 - m_n/n$ est un point de continuité de F. On approche donc $1 - F(u_n)$ par m_n/n, les deux quantités étant généralement égales puisque l'ensemble des points de discontinuité de F est au plus dénombrable. De plus, d'après le théorème de Pickands, $1 - F_{u_n}(y)$ est approchée par $1 - G_{\gamma_n,\sigma_n}(y)$, pour un couple (γ_n, σ_n) convenablement choisi, avec $y = q_{GPD,n} - u_n$. Donc, lorsque $\gamma_n \neq 0$, l'expression suivante définit $q_{GPD,n}$:

$$p_n = \frac{m_n}{n}\left(1 + (q_{GPD,n} - u_n)\frac{\gamma_n}{\sigma_n}\right)^{-1/\gamma_n}.$$

On en déduit l'expression de l'approximation GPD du quantile d'ordre $1 - p_n$ de la loi des données (voir Breiman *et al.* [10]) :

$$q_{GPD,n} = u_n + \frac{\sigma_n}{\gamma_n}\left[\left(\frac{np_n}{m_n}\right)^{-\gamma_n} - 1\right]. \tag{2.4}$$

À présent, on doit estimer u_n, le quantile d'ordre m_n/n ($> 1/n$). Il se trouve donc généralement à l'intérieur de l'intervalle défini par l'échantillon. On l'estime simplement par la $(n - m_n)$-ème observation ordonnée $\widehat{u}_n = x_{(n-m_n)}$. Quant aux paramètres σ_n et γ_n de la loi GPD (qui approche la loi des excès), on les estime, par exemple, par la méthode de Hill généralisée décrite par Dekkers *et al.* [22] :

$$\widehat{\sigma}_n = \frac{\widehat{u}_n M_1}{\rho(\widehat{\gamma}_n)} \quad , \quad \widehat{\gamma}_n = M_1 + 1 - \frac{1}{2}\left(1 - \frac{(M_1)^2}{M_2}\right)^{-1}, \tag{2.5}$$

$$\text{où} \quad M_1 = \frac{1}{m_n}\sum_{i=1}^{m_n} \ln(Y_{(n-m_n+i)}) - \ln(\widehat{u}_n)\ , \ M_2 = \frac{1}{m_n}\sum_{i=1}^{m_n} \left(\ln(Y_{(n-m_n+i)}) - \ln(\widehat{u}_n)\right)^2,$$

$$\text{et} \quad \rho(\gamma) = \begin{cases} 1 & \text{si } \gamma \geq 0, \\ \dfrac{1}{1-\gamma} & \text{si } \gamma < 0. \end{cases}$$

On obtient donc l'estimation suivante du quantile (voir Breiman *et al.* [10]) :

$$\widehat{q}_{GPD,n} = \widehat{u}_n + \frac{\widehat{\sigma}_n}{\widehat{\gamma}_n} \left[\left(\frac{np_n}{m_n} \right)^{-\widehat{\gamma}_n} - 1 \right]. \tag{2.6}$$

2.3.2 Présentation du test GPD : test d'adéquation d'un modèle central à une queue de distribution

Lors de l'étude du test ET, nous avons constaté que la loi asymptotique de l'estimateur ET, $\widehat{q}_{ET,n}$ n'était bien approchée que pour de très grandes tailles d'échantillon (voir le paragraphe 1.1.4 page 20). Cela implique notamment (voir les paragraphes 1.3.2 et 1.3.3 pages 42 et 44) que la version 1 du test ET, basée sur cette loi asymptotique, est peu puissante, tout au moins pour des échantillons de taille raisonnable. Or, le test ET a justement été créé dans le but de pouvoir, dans le cadre d'échantillons de petite taille (fréquents en industrie), estimer les queues de distribution à l'aide de modèle paramétriques classiques qui permettent de prendre en compte toute la maigre information contenue dans les données, lorsque les excès sont en nombre trop restreint pour pouvoir espérer utiliser avec succès les méthodes classiques d'estimation des queues de distribution.

Puisqu'il est très probable que de tels problèmes de convergence lente apparaissent lorsque l'on utilise l'estimateur GPD, $\widehat{q}_{GPD,n}$, au lieu de $\widehat{q}_{ET,n}$, et que nous souhaitons aussi pouvoir travailler avec de petits échantillons, nous n'avons pas défini de version du test GPD basée sur la loi asymptotique de $\widehat{q}_{GPD,n}$. Nous utilisons directement la méthode du bootstrap paramétrique pour simuler les fluctuations d'échantillonnage de l'estimateur GPD $\widehat{q}_{GPD,n}$:

- On se place sous l'hypothèse $\mathcal{H}_0$ et on estime la Fdr F de la loi des données par $F_{\widehat{\theta}_n}$.

- On génère N échantillons indépendants entre eux, chaque échantillon étant i.i.d., de loi de Fdr $F_{\widehat{\theta}_n}$ et de taille n, la même que l'échantillon initial. Sous $\mathcal{H}_0$, ces N échantillons seront donc comparables à l'échantillon de départ (même taille et loi proche).

- Pour chacun de ces échantillons, on calcule l'estimateur GPD associé $\widehat{q}^*_{GPD,n}$ donné par l'équation (2.6). L'étoile en exposant permet par convention de différencier les quantités calculées à partir des échantillons bootstrap.

- Pour chacun de ces échantillons, on calcule un estimateur $\widehat{\theta}^*_n$ des paramètres du modèle $(F_\theta)_{\theta\in\Theta}$. Ceci nous permet d'évaluer le quantile d'ordre $1-p_n$ de la loi de Fdr $F_{\widehat{\theta}^*_n}$ donné par $\widehat{q}^*_{\text{param},n}$.

2.3.2.1 Première version du test GPD : version bootstrap paramétrique complète

La méthode du bootstrap paramétrique nous fournit N valeurs $\widehat{q}^*_{GPD,n}$ de l'estimateur GPD et N valeurs $\widehat{q}^*_{\text{param},n} = F^{-1}_{\widehat{\theta}^*_n}(1-p_n)$ de l'estimateur paramétrique, ce qui nous permet de calculer N écarts $\widehat{\delta}^*_n = \widehat{q}^*_{\text{param},n} - \widehat{q}^*_{GPD,n}$ entre les deux estimateurs, ce qui correspond à N estimations de l'erreur d'approximation $\delta_n = q_{1-p_n} - q_{GPD,n}$.

De cet échantillon de $\widehat{\delta}^*_n$, on déduit un intervalle de confiance pour l'erreur d'approximation δ_n. Par exemple, pour un intervalle de confiance à 90%, on ordonne l'échantillon des N valeurs $\widehat{\delta}^*_n$, puis on ôte 5% des valeurs les plus grandes et 5% des plus petites. On appelle la plus petite valeur de l'échantillon restant $\widehat{\delta}^*_{\min,n}$ et la plus grande $\widehat{\delta}^*_{\max,n}$. L'intervalle de confiance bootstrap (à 90% par exemple) pour l'erreur d'approximation $\delta_n = q_{1-p_n} - q_{GPD,n}$ est alors

$$IC_{\delta,BP,n} = [\widehat{\delta}^*_{\min,n},\, \widehat{\delta}^*_{\max,n}]. \tag{2.7}$$

A partir de l'échantillon initial, on calcule $\widehat{q}_{GPD,n}$ et $\widehat{q}_{\text{param},n}$, estimateurs de q_{1-p_n}, le quantile d'ordre $1-p_n$ de F. Si l'écart $\widehat{\delta}_n = \widehat{q}_{\text{param},n} - \widehat{q}_{GPD,n}$ entre les deux estimateurs appartient à l'intervalle $IC_{\delta,BP,n}$ que l'on vient de construire, on peut raisonnablement supposer que le modèle accepté en partie centrale s 'ajuste aussi assez bien en queue de distribution. La version complète du test GPD (basée sur la méthode du bootstrap paramétrique) est donc la suivante :

Test GPD (version complète) : On rejette l'hypothèse $\mathcal{H}_0$ si $\widehat{\delta}_n \notin IC_{\delta,BP,n}$.

2.3.2.2 Seconde version du test GPD : version bootstrap paramétrique simplifiée

Les variations d'échantillonnage de $\widehat{q}_{\text{param},n}$ (en $1/\sqrt{n}$) peuvent être considérées comme petites par rapport à celles de $\widehat{q}_{GPD,n}$ (en $1/\sqrt{m_n}$), puisque m_n est petit devant n. On décide donc ici de ne plus bootstrapper l'estimation paramétrique du quantile. Alors, il n'est plus utile de soustraire $\widehat{q}_{\text{param},n}$ (qui reste maintenant constant) de $q_{GPD,n}$ ni de $\widehat{q}_{GPD,n}$, car cela revient seulement à translater $q_{GPD,n}$ et son intervalle de confiance de la même façon. Il suffit de construire un intervalle de confiance pour $q_{GPD,n}$, et de vérifier s'il contient ou non $\widehat{q}_{GPD,n}$.

Pour cela, on ordonne l'échantillon des N valeurs $\widehat{q}^*_{GPD,n}$, puis (pour un intervalle à 90% par exemple) on ôte 5% des données les plus grandes et 5% des plus petites. On appelle la plus petite valeur de l'échantillon restant $\widehat{q}^*_{GPD,\min,n}$ et la plus grande $\widehat{q}^*_{GPD,\max,n}$. L'intervalle de confiance bootstrap (à 90% par exemple) pour $q_{GPD,n}$ est alors

$$IC_{GPD,BP,n} = [\widehat{q}^*_{GPD,\min,n},\, \widehat{q}^*_{GPD,\max,n}]. \tag{2.8}$$

Il s'ensuit la version simplifiée du test GPD (basée sur la méthode du bootstrap paramétrique) :

Test GPD (version simplifiée): On rejette l'hypothèse $\mathcal{H}_0$ si $\widehat{q}_{GPD,n} \notin IC_{GPD,BP,n}$.

2.4 Relaxation de l'hypothèse d'un modèle central

Que ce soit lors de la description du test ET, au chapitre 1, ou bien lors de celle du test GPD, au paragraphe 2.3 (page 91), nous avons supposé que nous disposions d'un modèle classique, dont nous estimons les paramètres par des méthodes usuelles, qui prennent en compte tout l'échantillon, et sont donc principalement influencées par les valeurs les plus probables de la variable. Ceci correspond tout d'abord à notre premier objectif d'utiliser toute l'information contenue dans les données dans le cadre de petits échantillons, ou à notre second but de déterminer des modèles globaux (c'est-à-dire adaptés à la fois en parties centrale et extrême), mais aussi à notre habitude de manipuler de tels modèles (en particulier pour l'estimation de leurs paramètres et le choix de modèles).

Or, lors de la définition des tests ET et GPD, nous ne nous appuyons pas sur le fait que les modèles testés jusqu'à présent sont des modèles usuels, dont les paramètres (estimés par maximum de vraisemblance) dépendent principalement des valeurs les plus probables de la variable. Dans leur forme actuelle, les tests ET et GPD pourraient donc être appliqués à d'autres types de modèles. Nous pensons en particulier à des modèles de queues de distribution, qui auraient été estimés en queue de distribution. Le problème consiste alors à construire de tels modèles, estimer leurs paramètres et choisir parmi ceux-ci.

Des modèles spécifiques pour la queue de distribution pourraient être construits à partir de la théorie des valeurs extrêmes. Les tests ET et GPD étant définis à partir de la même théorie, nous ne pourrions donc pas les appliquer à de tels modèles. Le problème du choix de modèle resterait donc posé. De plus, nous avons vu au paragraphe 1.1.3 (page 15) que les erreurs dues à l'approximation de la loi des excès par une loi exponentielle, ou plus généralement par une loi GPD, peuvent impliquer un biais pour l'estimation des quantiles. Nous avons donc abandonné une telle approche.

Nous proposons alors d'utiliser des modèles connus (normal, lognormal, Student, etc.), en estimant leurs paramètres de façon à estimer au mieux la queue de distribution. Plus précisément, considérons un modèle de fonction de répartition F et de densité f. Puis choisissons un seuil u de façon à nous placer en queue de distribution (de ce modèle et des données), en considérant la fonction de répartition de la loi des excès au-delà du seuil u

$$F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)}.$$

On en déduit la densité de la loi des excès au-delà du seuil u :

$$f_u(y) = \frac{f(u+y)}{1 - F(u)}.$$

Fixons $u = \hat{u}_n = X_{(n-m_n)}$, c'est-à-dire que l'on a m_n excès. Il doit alors être possible (notamment lorsque F est connue analytiquement), avec plus ou moins de difficultés numériques, de déterminer des estimateurs des paramètres de la loi de densité f à partir de l'échantillon des excès et de la loi des excès. On peut, par exemple, envisager de calculer la vraisemblance, puis la logvraisemblance, ainsi que ses dérivées, ce qui permet de calculer les estimateurs du maximum de vraisemblance (puisque, sachant $\hat{u}_n = X_{(n-m_n)}$, les excès sont i.i.d.). Cependant les estimateurs du maximum de vraisemblance sont réputés être de mauvais estimateurs pour de petits échantillons, comme le sont en général les échantillons d'excès que nous avons rencontrés (taille d'environ un dixième de la taille de l'échantillon de départ). On peut alors envisager d'autres types d'estimateurs, comme les estimateurs des moments ou des moments pondérés. Mais, la loi des excès n'étant pas une loi classique, il faut tout d'abord pouvoir calculer ses deux premiers moments. Ceci doit pouvoir être le cas, au moins lorsqu'on connaît la forme analytique de F.

Pour le choix du modèle à adapter à la queue de distribution, on ne peut plus, comme pour un modèle classique, se laisser d'abord guider par la forme de l'histogramme des données. Cependant, la méthode graphique du *QQ-plot* (voir la remarque 2.3 page 85) permet de déterminer grossièrement la vitesse de décroissance de la queue de distribution à partir des données. Nous sommes alors poussés à choisir des modèles ayant le même type de décroissance en queue. Ces modèles peuvent ensuite être départagés à l'aide des tests ET et/ou GPD.

2.5 Conclusion

Nous avons remarqué au paragraphe 1.5 (page 57) que le test ET, d'adéquation d'un modèle central à la queue de distribution, présente principalement deux inconvénients. Premièrement, le test ET ne permet de proposer aucun modèle global pour les données lorsque les lois testées révèlent une bonne adéquation centrale, et une mauvaise adéquation extrême. Nous avons tout d'abord proposé une solution à ce problème (voir le paragraphe 2.1 page 60) avec la méthode de régularisation bayésienne. Cette méthode concerne un modèle central dont la queue de distribution (et par conséquent la loi dans son ensemble, mais dans une moindre mesure) est déformée en fonction d'un avis d'expert. Cette correction sur la queue est faible. On peut donc raisonnablement espérer que le nouveau modèle conserve sa bonne adéquation centrale. En contrepartie, cette correction peut être trop faible pour aboutir à une modélisation correcte de la queue de distribution.

Le second inconvénient du test ET est qu'il n'est défini que pour des lois appartenant au DA(Gumbel). Nous avons donc souhaité l'étendre à des lois appartenant aux trois domaines d'attraction des valeurs extrêmes. Nous proposons donc au paragraphe 2.3 (page 91) un test plus général d'adéquation d'un modèle en queue de distribution : le test GPD. Cependant, comme pour le test ET, on peut se poser la question de construire un modèle global lorsqu'aucun des modèles testés n'est accepté à la fois par un test central et le test GPD. On peut alors envisager d'étendre la procédure de régularisation bayésienne à une plus grande plage de modèles, notamment des modèles n'appartenant pas au DA(Gumbel) comme les lois de Fréchet, Student, Pareto du DA(Fréchet), ou les lois beta du DA(Weibull).

Chapitre 3

Estimation bayésienne de la loi GPD, loi asymptotique des excès au-delà d'un seuil

Nous nous plaçons à nouveau dans le cadre de l'estimation des queues de distribution. Il s'agit en particulier d'estimer, à partir d'un échantillon de taille n, les quantiles extrêmes $q_{1-p} = F^{-1}(1-p)$ d'une loi de fonction de répartion F inconnue, pour p de l'ordre de ou plus petit que $1/n$ (quantile généralement situé au-delà de l'observation maximale).

Nous souhaitons à présent utiliser une méthode non paramétrique pour l'estimation de quantiles extrêmes. Nous nous intéressons ici à la méthode des excès (ou POT : *Peaks Over Threshold*, voir de Haan et Rootzen [21]) pour l'estimation des queues de distribution, dont un cas particulier, la méthode ET, a dejà été utilisé dans les chapitres précédents (notamment pour le test ET au chapitre 1). Nous utilisons alors le fait que la loi des excès au-delà d'un seuil, lorsque ce seuil tend vers l'infini, tend vers une loi de Pareto généralisée (GPD : *Generalized Pareto Distribution*). Il est donc essentiel de bien estimer les paramètres de cette loi GPD.

Nous souhaitons notamment pouvoir travailler sur des échantillons de petite taille, pour lesquels l'information sur la queue de distribution (apportée par les excès qui sont en nombre restreint) est très réduite. Or, en particulier pour de petits échantillons, de ce type de méthodes non paramétriques découle un biais d'estimation, notamment dû à l'approximation de la loi des excès par une loi GPD. Nous souhaitons donc mieux estimer les paramètres de la loi GPD, en particulier en vue de réduire le biais d'estimation des quantiles extrêmes. Notre problème provenant notamment du fait que nous disposons de peu d'information sur la queue de distribution, nous souhaitons pouvoir introduire des indications supplémentaires en extrême. Nous proposons alors d'utiliser une méthode bayésienne qui permet d'incorporer de l'information sur la queue de distribution à travers un avis d'expert. Nous avons développé une méthode bayésienne d'estimation des paramètres de la loi GPD dans le but d'améliorer l'estimation de quantiles extrêmes par la méthode des excès, en particulier lorsque l'on dispose d'un avis d'expert. Nous commençons ici par présenter une méthode bayésienne peu informative (c'est-à-dire sans avis d'expert), puis nous introduisons un avis d'expert dans la procédure.

3.1 Présentation de la procédure bayésienne

3.1.1 Représentation des lois GPD sous forme de mélange continu

Notre approche bayésienne s'appuie sur une représentation des lois GDP sous forme d'un mélange continu de lois exponentielles avec une loi de mélange gamma. On considère le paramétrage GPD(α, β) (avec soit $\alpha > 0$ et $\beta > 0$, soit $\alpha < 0$ et $\beta < 0$) dont la densité s'exprime comme :

$$f_{\alpha,\beta}(x) = f(x|\,\alpha, \beta) = \frac{\alpha}{\beta}\left(1 + \frac{x}{\beta}\right)^{-\alpha-1} \qquad \text{pour} \quad \begin{cases} x \in I\!R_+ \text{ si } \alpha > 0. \\ x \in [0, -\beta[\text{ si } \alpha < 0. \end{cases} \tag{3.1}$$

On se restreint dans la suite au cas $\boldsymbol{\alpha > 0}$ et $\boldsymbol{\beta > 0}$, qui est celui pour lequel la densité de la loi GPD donnée par l'équation (3.1) peut se représenter sous la forme d'un mélange de lois exponentielles.

Remarque 3.1 *Le paramétrage usuel est GPD(σ, γ) (avec $\gamma \in I\!R$ et $\sigma > 0$), avec la densité*

$$f_{\sigma,\gamma}(x) = \frac{1}{\sigma}\left(1 + \frac{x\gamma}{\sigma}\right)^{-1/\gamma-1} \qquad \textit{pour} \quad \begin{cases} x \in I\!R_+ \textit{ si } \gamma > 0. \\ x \in [0, -\sigma/\gamma[\textit{ si } \gamma < 0, \end{cases}$$

avec $\alpha = 1/\gamma$ et $\beta = \sigma/\gamma$; et

$$f_{\sigma,0}(x) = \frac{1}{\sigma}\exp\left(-\frac{x}{\sigma}\right) \quad \textit{pour} \;\; x \in I\!R_+.$$

La restriction au cas $\alpha > 0$ et $\beta > 0$ correspond, pour le paramétrage usuel, à $\boldsymbol{\gamma > 0}$ (avec $\sigma > 0$ par définition). Nous nous plaçons donc dans le cadre du domaine d'attraction de Fréchet (DA(Fréchet)), domaine contenant des lois au point terminal infini et à queues lourdes, comme les lois de Cauchy, de Student, de Fisher, de Pareto, loggamma et de Burr. Le domaine d'attraction de Gumbel (DA(Gumbel)) correspondant au cas $\gamma = 0$ est exclu de ce cadre, mais on peut s'en approcher en faisant tendre α et β tous les deux vers l'infini de manière à ce que le rapport β/α tende vers une constante σ.

Avec le paramétrage GPD(α, β), $\boldsymbol{\alpha > 0}$ et $\boldsymbol{\beta > 0}$, on montre que la densité $f(x|\,\alpha, \beta)$ s'exprime sous la forme du mélange

$$f(x|\,\alpha, \beta) = \int_0^\infty z\, e^{-xz}\, g(z|\,\alpha, \beta)\, dz\,, \tag{3.2}$$

où la loi de mélange est une loi gamma de paramètres $a = \alpha$ et $\lambda = \beta$ (α et β doivent donc être positifs puisque les paramètres d'une loi gamma le sont), de densité

$$g(z|\,\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, z^{\alpha-1}\, e^{-\beta z}\, I\!I_{\{z>0\}}\,,$$

où $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\, e^{-t}\, dt$ est le coefficient de normalisation de la loi gamma (cf. annexe D.2.1 page 180).

3.1.2 Lois conjuguées pour une loi gamma

Nous allons à présent utiliser la représentation sous forme de mélange des lois GPD ainsi que la classe des lois a priori conjuguées pour la loi de mélange gamma, pour produire un schéma bayésien quasi conjugué pour les lois GPD, dans le cas où $\alpha > 0$ et $\beta > 0$.

Les densités a priori conjuguées pour les lois gamma ont été déterminées par Damsleth [18]. Nous avons choisi de les utiliser ici. Damsleth utilise une loi dite gamcon de type II (ou gamconII) dont la densité s'exprime sous la forme

$$\xi_{c,\,d}(x) = I_{c,\,d}^{-1}\,\Gamma(dx+1)\,(\Gamma(x))^{-d}\,(cd)^{-dx}\, I\!I_{\{x>0\}}\,, \tag{3.3}$$

avec $c > 1$ et $d > 0$, où $I_{c,\,d} = \int_0^\infty \Gamma(dx+1)\,(\Gamma(x))^{-d}\,(cd)^{-dx}\,dx$ désigne la constante de normalisation dont on ne connaît pas de forme explicite simple à calculer. Remarquons que pour $d = 0$ la loi gamconII devient la loi impropre égale à la mesure de Lebesgue sur $I\!R_+$, et pour $d = 1$ on retrouve une loi $\mathcal{G}amma(2, \ln c)$.

Les lois a priori conjuguées étudiées par Damsleth sont les suivantes :

- $\pi(\beta|\,\alpha)$, densité conditionnelle de la loi a priori de β sachant α : loi $\mathcal{G}amma(\delta\alpha+1, \delta\eta)$.
- $\pi(\alpha)$, densité de la loi a priori de α : loi gamconII$(\eta/\mu, \delta)$.

Ici, les hyperparamètres δ, η et μ vérifient $\delta > 0$ et $\eta > \mu > 0$ (puisque $d = \delta$ et $c = \eta/\mu$). On obtient alors comme lois a posteriori, étant donné un échantillon $\underline{z}_n = (z_1, \ldots, z_n)$ issu d'une loi gamma :

- $\pi(\alpha|\,\underline{z}_n)$, densité de la loi a posteriori de α sachant l'échantillon $\underline{z}_n$ (de loi gamma) : loi gamconII$(\eta'/\mu', \delta')$ avec

$$\delta' = \delta + n\,, \quad \eta' = \frac{\delta\eta + \sum_{i=1}^n z_i}{\delta + n} \quad \text{et} \quad \mu' = \mu^{\delta/(\delta+n)} \left(\prod_{i=1}^n z_i\right)^{1/(\delta+n)}.$$

- $\pi(\beta|\,\alpha, \underline{z}_n)$, densité conditionnelle de la loi a posteriori de β sachant α et l'échantillon $\underline{z}_n$: loi $\mathcal{G}amma(\delta'\alpha + 1, \delta'\eta')$.

Remarque 3.2 *Les hyperparamètres η et μ agissent sur les statistiques exhaustives $\sum_{i=1}^n z_i$ et $\sum_{i=1}^n \ln z_i$, tandis que δ mesure l'importance à accorder à ces modifications.*

Remarque 3.3 *Pour $\delta = 0$, η' est la moyenne arithmétique et μ' la moyenne géométrique des z_i.*

Remarque 3.4 *Pour $\delta = 1$, l'introduction de ces lois a priori revient à ajouter artificiellement une observation $z_0 = \eta$ dans le calcul de la moyenne arithmétique, et une observation différente $z_0' = \mu$ dans le calcul de la moyenne géométrique, avec $z_0/z_0' = \eta/\mu = c > 1$.*

3.1.3 Algorithme de Gibbs pour l'estimation bayésienne des lois GPD

Il est possible de transférer la loi a posteriori du couple $\theta = (\alpha, \beta)$ de paramètres d'une loi gamma, sachant un échantillon $\underline{z}_n$ de loi gamma, en une loi a posteriori pour le couple $\theta = (\alpha, \beta)$ de paramètres d'une loi GPD, sachant un échantillon $\underline{x}_n$ de loi GPD. On note

$$\pi(\theta|\,\underline{x}_n) = \frac{\prod_{i=1}^n f(x_i|\,\theta)\,\pi(\theta)}{f_\pi(\underline{x}_n)} \qquad \text{où} \qquad f_\pi(\underline{x}_n) = \int_{I\!R_+^2} \prod_{i=1}^n f(x_i|\,\theta')\pi(\theta')\,d\theta',$$

la densité de la loi a posteriori associée à la loi a priori de densité $\pi(\theta)$ au vu des observations $\underline{x}_n$ (issues d'une loi GPD),

$$\pi(\theta|\,\underline{z}_n) = \frac{\prod_{i=1}^n g(z_i|\,\theta)\,\pi(\theta)}{g_\pi(\underline{z}_n)} \qquad \text{où} \qquad g_\pi(\underline{z}_n) = \int_{I\!R_+^2} \prod_{i=1}^n g(z_i|\,\theta')\pi(\theta')\,d\theta',$$

la densité de la loi a posteriori associée à la loi a priori de densité $\pi(\theta)$ au vu des observations $\underline{z}_n$ (issues d'une loi gamma), et

$$p(\underline{x}_n|\,\underline{z}_n) = \Big(\prod_{i=1}^n z_i\Big)\exp\Big(-\sum_{i=1}^n x_i z_i\Big).$$

On obtient (cf. annexe D.1 page 179) une loi a posteriori pour θ sachant $\underline{x}_n$ du type

$$\pi(\theta|\,\underline{x}_n) = \int_{I\!R_+^n} q_\pi(\underline{z}_n|\,\underline{x}_n)\,\pi(\theta|\,\underline{z}_n)\,d\underline{z}_n\ , \quad \text{où} \quad q_\pi(\underline{z}_n|\,\underline{x}_n) = \frac{p(\underline{x}_n|\,\underline{z}_n)g_\pi(\underline{z}_n)}{\int_{I\!R_+^n} p(\underline{x}_n|\,\underline{z}_n')g_\pi(\underline{z}_n')\,d\underline{z}_n'}\ .$$

La loi a posteriori de θ sachant $\underline{x}_n$, l'échantillon de loi GPD, est donc un mélange continu avec pour loi de mélange la loi a posteriori connue de θ sachant $\underline{z}_n$, échantillon de loi gamma.

La densité de la loi a posteriori de θ sachant $\underline{x}_n$ est complexe. Remarquons que l'on peut aisément calculer la densité de la loi conditionnelle de θ sachant $\underline{z}_n$ (et sachant $\underline{x}_n$), $\pi(\theta|\,\underline{z}_n)$, ainsi que celle de $\underline{z}_n$ sachant θ (et sachant $\underline{x}_n$), $q_\pi(\underline{z}_n|\,\underline{x}_n, \theta)$, qui est le produit des

$$q_\pi(z_i|\,x_i, \alpha, \beta) = \frac{z_i\ \exp(-x_i z_i)\,g(z_i|\,\alpha, \beta)}{f_{\alpha,\beta}(x_i)} = \frac{(\beta + x_i)^{\alpha+1}}{\Gamma(\alpha+1)}\,z_i^\alpha\ \exp(-(\beta + x_i)z_i)\,,$$

pour $i = 1,\ldots,n$, qui sont les densités de lois $\mathcal{G}amma(\alpha+1, \beta+x_i)$. On en déduit l'algorithme de Gibbs suivant :

1. $\forall i = 1,\ldots,n$, simuler indépendamment $z_i^{(m)}$ selon la loi a posteriori de z sachant x_i, α et β, de densité $q_\pi(z|\, x_i, \alpha^{(m)}, \beta^{(m)})$, qui est une loi $\mathcal{G}amma(\alpha^{(m)} + 1, \beta^{(m)} + x_i)$.
2. Simuler $\alpha^{(m+1)}$ selon la loi a posteriori de α sachant $\underline{z}_n^{(m)}$, de densité $\pi(\alpha|\, \underline{z}_n^{(m)})$, qui est une loi gamconII$(\eta'/\mu', \delta')$ avec
$$\delta' = \delta + n\,, \quad \eta' = \frac{\delta\eta + \sum_{i=1}^n z_i^{(m)}}{\delta + n} \quad \text{et} \quad \mu' = \mu^{\delta/(\delta+n)} \left(\prod_{i=1}^n z_i^{(m)}\right)^{1/(\delta+n)}.$$
3. Simuler $\beta^{(m+1)}$ selon la loi a posteriori de β sachant $\alpha^{(m+1)}$ et l'échantillon $\underline{z}_n^{(m)}$, de densité $\pi(\beta|\, \alpha^{(m+1)}, \underline{z}_n^{(m)})$, qui est une loi $\mathcal{G}amma(\delta'\alpha^{(m+1)} + 1, \delta'\eta')$.

Pour appliquer cet algorithme, il nous faut maintenant savoir simuler selon la loi gamconII (voir le paragraphe 3.1.4 ci-dessous), et déterminer les hyperparamètres δ, η et μ. Nous avons étudié plusieurs cadres où les solutions proposées pour déterminer les hyperparamètres sont différentes. Le paragraphe 3.2 (page 107) présente le cas d'échantillons de loi GPD, tandis que dans le paragraphe 3.3 (page 125) on travaille sur des échantillons d'excès. Dans les deux cas précédents, on se limite à une analyse bayésienne empirique puisque l'on n'introduit pour le moment aucun avis d'expert. Le paragraphe 3.4 (page 135) présente une première tentative (partielle) pour introduire un avis d'expert.

Remarque 3.5 *On peut aussi envisager une structure hiérarchique en considérant les hyperparamètres δ, η et μ comme aléatoires (voir par exemple le paragraphe 3.6 page 152). Ce type méthode, qui permettrait entre autres d'étudier conjointement plusieurs échantillons (notamment de petite taille) ou d'introduire une dépendance spatiale entre plusieurs sites de mesure, est actuellement à l'étude.*

3.1.4 Simulation selon la loi gamcon de type II

Cette loi n'est pas une loi usuelle, et il n'existe pas d'algorithme de simulation reconnu et programmé. Nous avons tout d'abord pensé utiliser un algorithme d'acceptation-rejet, mais ce procédé (basé sur une loi gamma dont la densité apparaît assez naturellement dans la formule de la loi gamconII) s'est révélé extrêment long (nombre moyen de rejets avant une acceptation très élevé : au moins de l'ordre de 10^5).

3.1.4.1 Approximation normale de Laplace

Dans un premier temps, nous avons décidé d'utiliser une approximation normale de Laplace de la loi gamconII. Cette approche est applicable lorsque le paramètre d de la loi gamconII est suffisamment grand. Puisqu'ici il s'agit de simuler selon la loi a posteriori, et que dans ce cas $d = \delta' = \delta + n$, où n est la taille de l'échantillon, on peut supposer que pour des tailles d'échantillon raisonnables d sera suffisamment grand pour que l'approximation normale soit applicable. Il s'agit alors de déterminer le mode de la loi gamconII, puis d'approcher cette loi par une loi normale de même mode (égal à la moyenne pour une loi normale) et dont la variance est donnée par un développement de Taylor au voisinage de ce mode de l'exposant de la densité de la loi gamconII.

Calcul du mode d'une loi gamconII Le mode d'une loi est le point (que l'on suppose unique) où la densité atteint son maximum. Puisque la fonction logarithme est croissante, le mode y^* d'une loi gamconII est

$$\begin{aligned} y^* &= \operatorname{argmax}_y \left[d \ln y + \ln \Gamma(yd+1) - d \ln \Gamma(y+1) - dy \ln d - dy \ln c \right] \\ &= \operatorname{argmax}_y \left[\ln y + \frac{1}{d} \ln \Gamma(yd+1) - \ln \Gamma(y+1) - y \ln d - y \ln c \right]. \end{aligned}$$

Notons

$$A_{c,d}(y) = \ln y + \frac{1}{d} \ln \Gamma(yd+1) - \ln \Gamma(y+1) - y \ln d - y \ln c\,. \tag{3.4}$$

Tout point y^* où $A_{c,d}(y)$ atteint son maximum est un point où $A'_{c,d}(y)$, sa dérivée par rapport à y, est nulle. Ce point est unique si la dérivée seconde de $A''_{c,d}(y)$ reste toujours du même signe, ce qui est bien le cas ici (cf. annexe D.2.2 page 181). On cherche donc l'unique y^* tel que

$$A'_{c,d}(y^*) = \psi(y^* d + 1) - \psi(y^* + 1) + \frac{1}{y^*} - \ln d - \ln c = 0\,,$$

où ψ est la fonction digamma, c'est-à-dire la dérivée du logarithme de la fonction gamma (certaines propriétés de la fonction ψ sont données en annexe D.2.1 page 180). Notons que en raison d'une relation simple vérifiée par la fonction ψ (donnée en annexe D.2.1 page 180), on a

$$A'_{c,d}(y^*) = \psi(y^* d + 1) - \psi(y^*) - \ln d - \ln c\,.$$

On calcule y^* par une simple dichotomie sur la fonction $A'_{c,d}(y)$, avec les bornes de départ $1/(4 \ln c)$ et $2/ \ln c$ (voir l'annexe D.2.3 page 182). On a choisi la dichotomie plutôt que la méthode de Newton parce que cette dernière peut produire des points négatifs en cours d'algorithme, alors que la dichotomie permet de rester entre les deux bornes positives de départ de l'algorithme.

Calcul de la variance de la loi normale approximante On utilise la fonction $A_{c,d}(y)$ définie ci-dessus pour réexprimer la densité de la loi gamconII :

$$\xi_{c,d}(y) = I_{c,d}^{-1} \exp(d\, A_{c,d}(y))\,.$$

Puis on effectue un développement de Taylor de la fonction $A_{c,d}(y)$ autour de son mode y^* (égal à celui de $\xi_{c,d}(y)$) :

$$A_{c,d}(y^* + h) = A_{c,d}(y^*) + h\, A'_{c,d}(y^*) + \frac{h^2}{2} A''_{c,d}(y^*) + \frac{h^3}{6} A'''_{c,d}(y_h)\,,$$

où y_h est situé entre y^* et $y^* + h$ (h pouvant être positif ou négatif),

$$A''_{c,d}(y) = d\,\psi'(yd+1) - \psi'(y+1) - \frac{1}{y^2} = d\,\psi'(yd+1) - \psi'(y)\ ,$$

puisque $\psi'(y+1) = \psi'(y) - 1/y^2$ (voir l'annexe D.2.1 page 180), et

$$A'''_{c,d}(y) = d^2\,\psi''(yd+1) - \psi''(y)\,.$$

Puisque par définition de y^* on a $A'_{c,d}(y^*) = 0$, en remplaçant $A_{c,d}(y)$ par son développement de Taylor dans l'expression de $\xi_{c,d}(y)$, pour $y = y^* + h$, on obtient

$$\begin{aligned}
\xi_{c,d}(y) &= \xi_{c,d}(y^*+h) \\
&= I_{c,d}^{-1} \exp\Big(dA_{c,d}(y^*) + \frac{h^2}{2}\,dA''_{c,d}(y^*) + \frac{h^3}{6}\,dA'''_{c,\,d}(y_h)\Big), \ \text{ pour } y_h \text{ entre } y^* \text{ et } y^*+h, \\
&\approx \mathrm{cte}_{c,d}\ \exp\Big(-\ \frac{(y-y^*)^2}{2{\sigma^*_d}^2}\Big), \quad \text{ pour } y \text{ proche de } y^* \text{ c'est-à-dire } |h| \text{ assez petit,}
\end{aligned}$$

où on a négligé le reste en h^3 du développement de Taylor (voir la justification en annexe D.2.5 page 185). La notation $\mathrm{cte}_{c,d}$ désigne la constante de normalisation de la loi normale approximante, qui correspond approximativement à la quantité $I_{c,d}^{-1}\exp(d\,A_{c,d}(y^*))$ (constante de l'expression de $\xi_{c,d}$). On a noté $h = y - y^*$, et ${\sigma^*_d}^2 = 1/(-d\,A''_{c,d}(y^*))$ désigne la variance de la loi normale approximante.

Remarque 3.6 *On peut montrer que la variance ${\sigma^*_d}^2$ de la loi normale approximant la loi gamconII est équivalente à $1/d(\ln d)^2$ lorsque $d \to \infty$ (voir l'annexe D.2.4 page 184), et que l'approximation normale n'est acceptable que pour $h = O(1/\sqrt{d}\ln d)$ (voir l'annexe D.2.5 page 185).*

Simulation selon une loi gamconII On simule selon la loi normale approximante $\mathcal{N}(y^*, \sigma^*_d)$ où le mode y^* est l'unique racine de l'équation

$$\psi(yd+1) - \psi(y) - \ln d - \ln c = 0\,, \tag{3.5}$$

et l'écart-type est

$$\sigma^*_d = \frac{1}{\sqrt{d\psi'(y^*) - d^2\psi'(y^*d+1)}}\,. \tag{3.6}$$

Une illustration graphique de cette méthode est donnée en figure 3.1 où l'on compare la loi gamconII et son approximation normale au travers de leurs densités, pour différents valeurs des paramètres c et d de la loi gamconII.

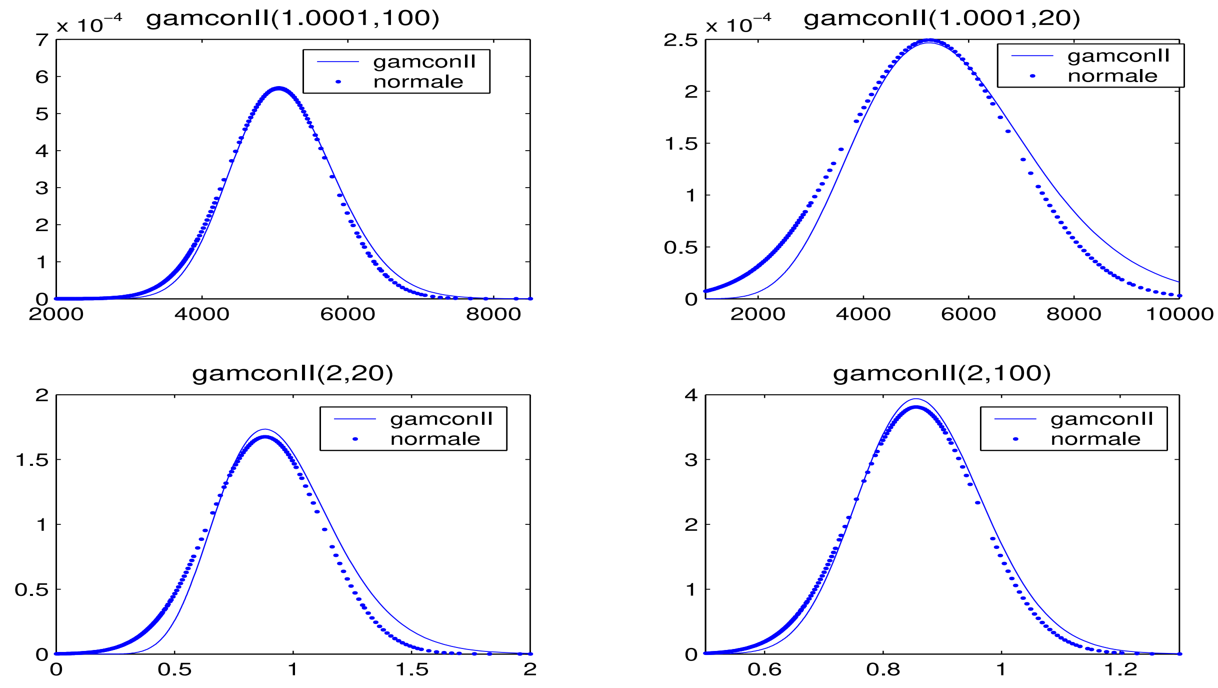

FIG. 3.1 – *Approximation normale pour les lois gamconII de paramètres (1.0001, 100), (1.0001, 20), (2, 100) et (2, 20) – trait plein : gamconII; points : normale approximante.*

3.1.4.2 Étape de Hastings-Metropolis

Afin de simuler réellement selon la loi gamconII, nous avons ensuite utilisé une procédure de Hastings-Metropolis [13] (décrite en annexe D.3.2 page 188). Pour limiter le nombre de rejets lorsque l'on est en queue de distribution (on veut alors avoir une forte probabilité d'accepter le point suivant), on choisit comme loi instrumentale une loi à queue lourde. On utilise ici une loi de Cauchy de même mode et dont la densité prend la même valeur au mode que la loi normale approximante. L'étape de Hastings-Metropolis est la suivante :

- simuler Y selon une loi de Cauchy de même mode y^* que la loi normale approximante et telle que $f(y^*) = 1/\sqrt{2\pi}\sigma_d^*$, où f est la densité de la loi de Cauchy instrumentale :

$$Y = y^* + \sigma_d^* \sqrt{\frac{2}{\pi}} \tan(\pi(U - 0.5)),$$

  où U est de loi uniforme sur [0,1], y^* est le mode de la loi normale approximante, et σ_d^* son écart-type.

- former le rapport

$$\rho = \min\left(1, \frac{f(\alpha^{(m)})g(Y)}{g(\alpha^{(m)})f(Y)}\right),$$

où

$$f(x) = \left(\sigma_d^* \sqrt{2\pi}\left(1 + \frac{\pi(x - y^*)^2}{2\sigma_d^{*2}}\right)\right)^{-1}$$

est la densité de la loi de Cauchy instrumentale, et g est la densité de la loi gamconII. Le fait que la loi de Cauchy ait une queue lourde implique que lorsque $\alpha^{(m)}$ est grand et Y moyen (ce qui sera souvent le cas), ρ sera plutôt proche de 1. Il s'ensuit que lorsque $\alpha^{(m)}$ est en queue de distribution on a une probabilité importante d'accepter Y.

- prendre $\alpha^{(m+1)} = Y$ avec probabilité ρ, et $\alpha^{(m+1)} = \alpha^{(m)}$ sinon.

L'introduction de cette étape ralentit le déroulement du programme car elle implique un test logique (précédé d'un calcul numérique et d'une simulation, tous deux rapides en temps de calcul) assez coûteux en temps de calcul, en particulier sous un langage interprété comme Matlab. D'après nos simulations, à nombre d'itérations constant, le temps de calcul ne semble cependant pas trop ralenti pour un nombre d'itérations modéré. De plus, cette étape de Hastings-Metropolis ne semble pas trop influer sur la rapidité de convergence de l'algorithme de Gibbs. S'il fallait de nombreuses itérations supplémentaires pour obtenir approximativement la convergence de l'algorithme de Gibbs vers son régime stationnaire, l'augmentation du temps de calcul dû aux itérations supplémentaires pourrait nous pousser à préférer l'approximation normale. Mais nos expérimentations numériques montrent que ce n'est généralement pas le cas. Nous utilisons donc dans la suite des simulations une étape de Hastings-Metropolis pour la simulation selon la loi gamconII.

3.2 Application de la procédure bayésienne à des données de loi GPD

Notre but est, à terme, d'utiliser cette procédure pour des échantillon d'excès en vue d'utiliser la méthode POT d'estimation des queues de distribution. Nous commençons cependant par appliquer cette procédure à des échantillons de loi GPD, afin d'en contrôler le déroulement et la qualité.

3.2.1 Détermination des hyperparamètres dans le cadre bayésien empirique

Bien que notre but soit d'utiliser à terme un avis d'expert pour améliorer l'estimation de la loi GPD et des quantiles extrêmes, nous commençons par nous placer dans le cadre bayésien empirique, plus simple, pour lequel l'information a priori provient de l'échantillon lui-même, et non d'un expert.

La première intuition naturelle est de déduire de l'information de la moyenne et de la variance empiriques. Ceci équivaut en fait à utiliser les estimateurs des moments des paramètres α

et β de la loi GPD. Nous avons ensuite introduit, au lieu des estimateurs des moments, les estimateurs des moments pondérés de Hosking et Wallis (PWM : *Probability Weighted Moments*) [37] qui sont meilleurs. On peut envisager d'utiliser encore d'autres estimateurs de α et β (comme les estimateurs du maximum de vraisemblance). Mais, les résultats obtenus étant très satisfaisants avec les estimateurs des moments pondérés (qui sont réputés être de très bons estimateurs paramétriques pour une grande région de l'ensemble des couples (α, β) que nous considérons, et pour des tailles d'échantillon modérées), nous n'avons pas utilisé d'autres estimateurs.

À présent, on dispose donc d'estimations $\tilde{\alpha}$ et $\tilde{\beta}$ (des moments ou des moments pondérés) des paramètres α et β de la loi GPD. Ces valeurs des paramètres sont en général assez satisfaisantes. Par exemple, les estimateurs des moments pondérés sont d'assez bons estimateurs lorsque $-0.4 \leq \gamma \leq 0.4$ approximativement, c'est-à-dire pour nous lorsque $\alpha \geq 2.5$. On va donc choisir des lois a priori pour α et pour β sachant α, telles que ces valeurs $\tilde{\alpha}$ et $\tilde{\beta}$ soient parmi les plus probables. En particulier, on peut choisir de déterminer les hyperparamètres de telle façon que $\tilde{\alpha}$ et $\tilde{\beta}$ soient les moyennes ou les modes des lois a priori (marginale pour α, conditionnelle sachant α pour β) correspondantes.

Remarque 3.7 *Puisque $\delta' = \delta + n$ avec $\delta > 0$, le choix de $\delta = 1$ semble naturel et introduit une modification raisonnable par rapport aux données, puisqu'il revient approximativement à ajouter artificiellement une observation (de loi gamma, voir la remarque 3.4 page 102). D'autre part, la loi a priori sur α est une loi gamconII de paramètres $c = \eta/\mu$ et $d = \delta$ qui est en général difficile à manipuler. Mais, lorsque $d = \delta = 1$, cette loi gamconII se transforme en une loi $\mathcal{G}amma(2, \ln c)$, bien plus commode.*

Pour simplifier les calculs, on a donc choisi de fixer $\boldsymbol{\delta = 1}$.

3.2.1.1 Cas de la moyenne a priori

La loi a priori de β sachant α est une loi $\mathcal{G}amma(\delta\alpha + 1, \delta\eta)$. La moyenne de cette loi est $(\delta\alpha + 1)/\delta\eta = (\alpha + 1)/\eta$ puisqu'on a choisi $\delta = 1$. On suppose que cette valeur moyenne est égale à $\tilde{\beta}$, l'estimation de β (par la méthode des moments ou des moments pondérés), et on remplace α par sa valeur moyenne $\tilde{\alpha}$. On obtient alors

$$\eta = \frac{\tilde{\alpha} + 1}{\tilde{\beta}} \,. \tag{3.7}$$

La loi a priori sur α est une loi gamconII de paramètres $c = \eta/\mu$ et $d = \delta = 1$ qui se simplifie en une loi $\mathcal{G}amma(2, \ln c)$. La moyenne de cette loi gamma est $2/\ln c$, que l'on suppose égale à $\tilde{\alpha}$, l'estimation de α (par la méthode des moments ou des moments pondérés). On obtient que $\eta/\mu = c = \exp(2/\tilde{\alpha})$, et donc que

$$\mu = \frac{\eta}{c} = \exp\left(-\frac{2}{\tilde{\alpha}}\right) \frac{\tilde{\alpha} + 1}{\tilde{\beta}} \,. \tag{3.8}$$

Remarque 3.8 *Si $\delta \neq 1$, on ne peut pas conclure. En effet, on ne connaît pas d'expression analytique simple de la moyenne de la loi gamconII, ce qui nous empêche de trouver la valeur de c et donc celle de μ.*

3.2.1.2 Cas du mode a priori

La loi a priori sur β sachant α est une loi $\mathcal{G}amma(\delta\alpha + 1, \delta\eta)$. Le mode de cette loi est $\delta\alpha/\delta\eta = \alpha/\eta$. On suppose donc que cette valeur modale est égale à $\tilde{\beta}$, l'estimation de β, et on remplace α par sa valeur moyenne $\tilde{\alpha}$. On obtient donc maintenant

$$\eta = \frac{\tilde{\alpha}}{\tilde{\beta}}. \tag{3.9}$$

La loi a priori sur α est une loi gamconII de paramètres $c = \eta/\mu$ et $d = \delta = 1$ qui se simplifie en une loi $\mathcal{G}amma(2, \ln c)$. Le mode de cette loi gamma est $1/\ln c$, que l'on suppose égal à $\tilde{\alpha}$, l'estimation de α. On obtient $\eta/\mu = c = \exp(1/\tilde{\alpha})$, et donc

$$\mu = \frac{\eta}{c} = \exp\left(-\frac{1}{\tilde{\alpha}}\right)\frac{\tilde{\alpha}}{\tilde{\beta}}. \tag{3.10}$$

Remarque 3.9 *Si $\delta \neq 1$, cette fois-ci on peut conclure (mais alors le problème devient : comment choisir δ ?). En effet, le mode de la loi gamconII n'a pas non plus d'expression analytique, mais on sait qu'il vérifie l'équation (3.5) page 105. Puisque pour déterminer les hyperparamètres, on suppose que la loi a priori (la loi gamconII) de α est de mode égal à $\tilde{\alpha}$, on remplace dans l'équation (3.5) y par $\tilde{\alpha}$ et $d = \delta$ par sa valeur. Puis on résout l'équation en c, et on en déduit que*

$$\mu = \frac{\eta}{c} = \frac{\tilde{\alpha}}{\tilde{\beta}} \exp(\ln\delta + \psi(\tilde{\alpha}) - \psi(\tilde{\alpha}\delta + 1)).$$

3.2.2 Estimation

Dans le cadre de l'estimation bayésienne, deux possibilités nous sont offertes pour estimer la densité d'un modèle de loi. Tout d'abord, nous pouvons, comme dans le cadre d'une estimation paramétrique, estimer les paramètres du modèle par la méthode bayésienne, et en déduire une loi estimée appartenant au modèle étudié. Cette possibilités est explorée au paragraphe 3.2.2.1. Mais il est aussi possible d'estimer directement la densité de la loi dont est issu l'échantillon par un mélange continu des densités du modèle avec pour loi de mélange la loi a posteriori des paramètres du modèle. Nous traitons ce cas dans le paragraphe 3.2.2.2. Dans ce cas, la loi obtenue (dite loi prédictive a posteriori, ou plus simplement loi prédictive) n'appartient plus au modèle.

3.2.2.1 Estimation bayésienne de α et β

Dans ce premier cas, on veut estimer les paramètres de la loi GPD par leurs estimateurs bayésiens. Il existe différents estimateurs bayésiens des paramètres d'une loi, dépendant de la fonction de coût choisie. Nous avons choisi dans un premier temps d'estimer les paramètres du modèle GPD par la moyenne a posteriori (estimateur de $\theta =$ moyenne de la loi a posteriori) qui correspond au classique mais contesté coût quadratique. Puis on a utilisé comme estimateur le mode a posteriori (estimateur de $\theta =$ mode de la loi a posteriori) correspondant au coût 0-1, qui intuitivement peut donner de meilleures estimations pour des lois asymétriques comme le sont les lois a posteriori que nous obtenons. Enfin, la moyenne étant sensible aux valeurs aberrantes, nous avons aussi utilisé comme estimateur la médiane de la loi a posteriori, qui est peu influencée par les valeurs extrêmes.

Le problème est qu'il n'existe pas d'expression analytique de la loi a posteriori de θ sachant $\underline{x}_n$. Par contre, l'algorithme de Gibbs nous donne, lorsqu'il a approximativement atteint son régime stationnaire (faire une vérification graphique, voir la figure 3.2 page 112, ou numérique, voir l'annexe D.3.4 page 189), des réalisations (de taille s) $\underline{\alpha}_s$ de la loi a posteriori de α sachant $\underline{x}_n$ et $\underline{\beta}_s$ de la loi a posteriori de β sachant α et $\underline{x}_n$. On peut donc approximer l'estimateur bayésien de θ grâce à ces échantillons $\underline{\alpha}_s$ et $\underline{\beta}_s$ (attention à éliminer les premières valeurs de l'algorithme de façon à avoir approximativement atteint la loi stationnaire).

Cas de la moyenne a posteriori Il suffit d'estimer les moyennes des lois a posteriori marginales de α et β par les moyennes empiriques des échantillons $\underline{\alpha}_s$ et $\underline{\beta}_s$ obtenus par l'algorithme de Gibbs (après atteinte approximative de la loi stationnaire). On a donc les estimations

$$\hat{\alpha} = \overline{\underline{\alpha}_s} \quad \text{et} \quad \hat{\beta} = \overline{\underline{\beta}_s}\,.$$

Cas de la médiane a posteriori Cette fois aussi, il nous suffit d'estimer les médianes des lois a posteriori marginales de α et β par les médianes empiriques des échantillons $\underline{\alpha}_s$ et $\underline{\beta}_s$ de taille s obtenus par l'algorithme de Gibbs (après atteinte approximative de la loi stationnaire). On a donc les estimations

$$\hat{\alpha} = \alpha_{([s/2])} \quad \text{et} \quad \hat{\beta} = \beta_{([s/2])}\,,$$

où $[x]$ désigne la partie entière de x, et $x_{(i)}$ la i-ème valeur de l'échantillon ordonné.

Cas du mode a posteriori Dans ce cas, il est plus difficile d'approximer l'estimateur bayésien. Il faut tout d'abord construire un estimateur lissé de la densité a posteriori de α (respectivement β). Nous proposons d'utiliser la méthode des estimateurs à noyau pour estimer la densité de l'échantillon $\underline{\alpha}_s$ (resp. $\underline{\beta}_s$) (pour plus de précisions, voir l'annexe D.3.5 page 190). On obtient donc une estimation de la densité de la loi a posteriori de α (resp. β). L'estimateur bayésien, qui est ici le mode a posteriori, est alors estimé par le mode de l'estimateur à noyau de la densité de la loi a posteriori.

3.2.2.2 Estimation de la loi de l'échantillon

La première possibilité pour estimer la loi dont sont issues les données $\underline{x}_n$ est d'estimer les paramètres de la loi GPD, dans notre cas par la méthode bayésienne comme on l'a fait au pragraphe précédent. Mais dans le cadre bayésien, on peut aussi estimer directement la densité par la loi prédictive, qui est un mélange continu des densités des lois GPD avec comme loi de mélange la loi a posteriori de $\theta = (\alpha, \beta)$ sachant $\underline{x}_n$.

À nouveau, cette loi a posteriori n'est pas connue analytiquement. On a donc recours à une approximation de Monte-Carlo de cette intégrale en utilisant l'échantillon $\underline{\theta}_s$ de valeurs de $\theta = (\alpha, \beta)$ obtenu par l'algorithme de Gibbs (encore une fois, l'échantillon n'est créé que lorsque la loi stationnaire est approximativement atteinte ; on élimine donc les premières valeurs obtenues). Dans ce cas, on estime la densité (resp. la fonction de répartition) de la loi prédictive a posteriori par la moyenne empirique des densités $f_\theta(x)$ (resp. des fonctions de répartition $F_\theta(x)$) des lois GPD correspondant aux valeurs de $\theta = (\alpha, \beta)$ obtenues par l'algorithme de Gibbs.

Soit s le nombre de valeurs de $\theta = (\alpha, \beta)$ obtenues par l'algorithme de Gibbs (après que l'on ait approximativement atteint la loi stationnaire). On note $\theta^{(i)} = (\alpha^{(i)}, \beta^{(i)})$ la i-ème valeur du couple. La densité de la loi prédictive est alors estimée par

$$g_{\text{pred}}(x) = \frac{1}{s} \sum_{i=1}^{s} f(x|\, \alpha^{(i)}, \beta^{(i)}) \ , \tag{3.11}$$

où $f(x|\, \alpha, \beta)$ est la densité de la loi GPD de paramètres (α,β). De même, la fonction de répartition de la loi prédictive est estimée par

$$G_{\text{pred}}(x) = \frac{1}{s} \sum_{i=1}^{s} F(x|\, \alpha^{(i)}, \beta^{(i)}) \ , \tag{3.12}$$

où $F(x|\, \alpha, \beta)$ est la fonction de répartition de la loi GPD de paramètres (α, β).

3.2.3 Exploration numérique de cette méthode bayésienne pour des échantillons de loi GPD

Nous illustrons tout d'abord les resultats obtenus sur un exemple. Puis nous présentons des simulations intensives qui permettent de juger la qualité de notre méthode et de comparer les différentes versions de cette méthode :

- différents estimateurs bayésiens des paramètres de la loi GPD : par le mode, la moyenne, la médiane a posteriori.
- différentes façons de déterminer les hyperparamètres (par le mode ou la moyenne a priori).

3.2.3.1 Exemple sur un échantillon de taille $n = 100$ simulé selon la loi GPD(3,3)

En premier lieu, nous voudrions vérifier que la loi stationnaire de notre algorithme de Gibbs est approximativement atteinte. A terme, il faudrait introduire une procédure automatique de vérification permettant un choix automatique du nombre d'itérations. Pour l'instant, nous effectuons par défaut $k = 1000$ itérations, et nous retenons les $r = 500$ dernières en supposant qu'alors la stationnarité est approximativement atteinte. Pour cet exemple, nous avons vérifié cette supposition à l'aide de méthodes graphiques. Nous avons tracé, pour les paramètres α et β, les différentes valeurs obtenues au cours de l'algorithme de Gibbs, ainsi que l'évolution des moyennes calculées sur les 500 (ou moins en début d'algorithme) itérations précédant la i-ème (cf. les quatre graphiques de la figure 3.2).

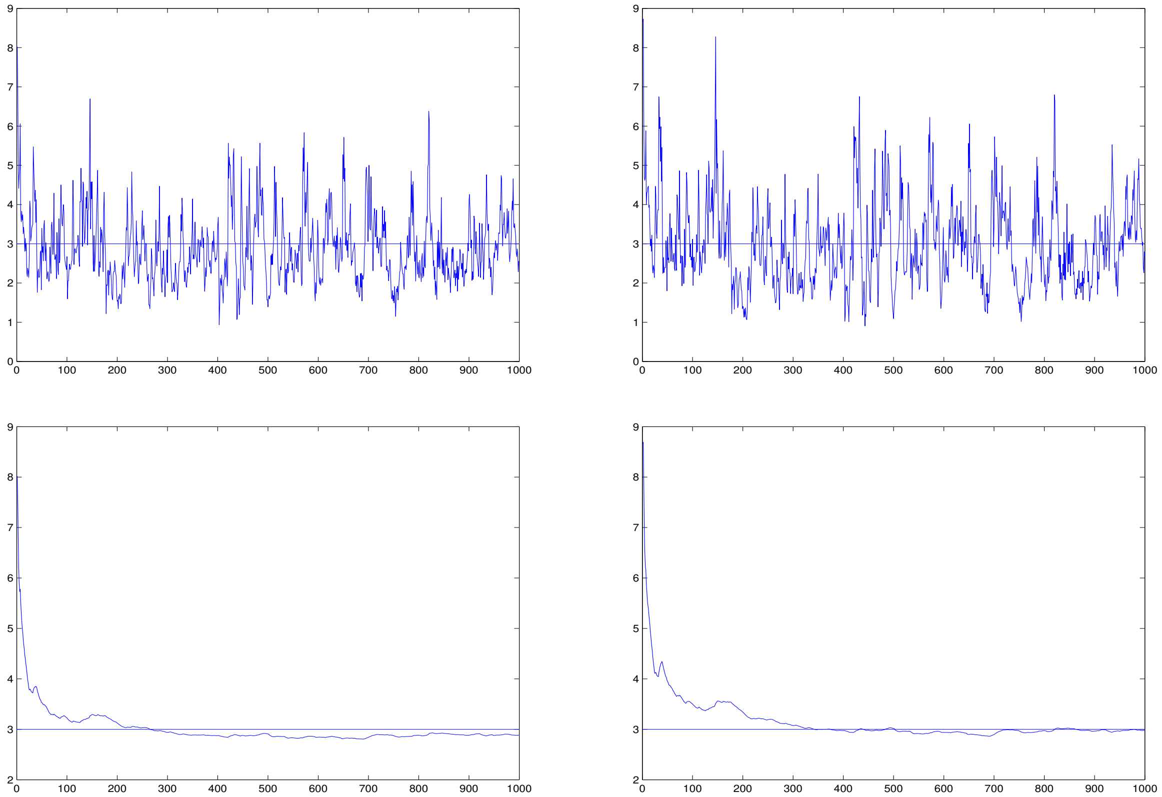

FIG. 3.2 – *Évolution des valeurs de α (en haut à gauche) et β (en haut à droite) obtenues au cours de l'algorithme de Gibbs et de leurs moyennes (en bas, à gauche pour α et à droite pour β), pour 1000 itérations, dans le cas de la moyenne a priori.*

Ici, le point de départ de l'agorithme de Gibbs ($\alpha_0 = 20$ et $\beta_0 = 20$) est volontairement choisi assez éloigné de la vraie valeur ($\alpha = 3$ et $\beta = 3$). On voit que même dans ce cas les valeurs de α et β obtenues se rapprochent très vite des vraies valeurs (seules les deux ou trois premières valeurs sont autour de 8 ou 9) et qu'ensuite elles varient aléatoirement dans un intervalle raisonnable autour de cette vraie valeur. D'autre part, on peut voir que la valeur moyenne (des 500 dernières valeurs, ou moins) calculée se stabilise très vite à une valeur proche de la vraie valeur que nous souhaitons estimer (pour α comme pour β). Ces deux indications nous laissent donc supposer que la loi stationnaire est approximativement atteinte.

Nous souhaitons maintenant avoir une idée de la forme des lois a posteriori marginales de α et β. À partir des valeurs obtenues par l'algorithme de Gibbs (après atteinte approximative de la stationnarité, on utilise donc les $r = 500$ dernières valeurs) on trace à cet effet les histogrammes correspondants (cf. les deux graphes de la figure 3.3).

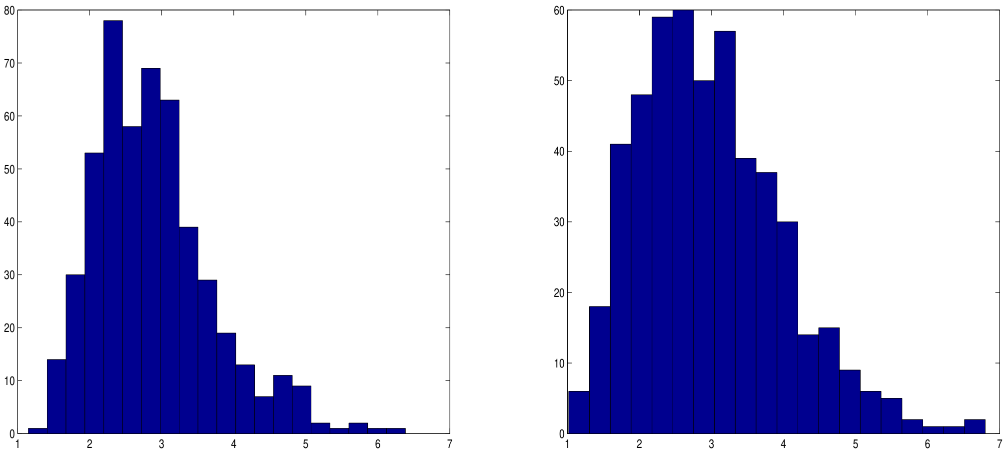

FIG. 3.3 – *Histogramme des 500 dernières valeurs obtenues par l'algorithme de Gibbs pour α (à gauche) et β (à droite), dans le cas de la moyenne a priori*

On peut voir que les lois a posteriori marginales de α ainsi que de β sont des lois asymétriques. C'est l'une des raisons pour lesquelles nous avons eu l'idée d'utiliser le mode de ces lois marginales pour estimer les paramètres.

Dans cet exemple, les différentes estimations des paramètres α et β (estimations bayésiennes par le mode, la moyenne, la médiane des 500 dernières valeurs de l'échantillonneur de Gibbs, estimateurs des moments pondérés de Hosking et Wallis [37]) sont les suivantes.

estimations par	Hosking et Wallis	mode	médiane	moyenne
		de l'algorithme de Gibbs		
$\hat{\alpha}$	2.3858	2.3304	**2.7699**	**2.8843**
$\hat{\beta}$	2.3541	2.7150	**2.8547**	**2.9817**

Remarque 3.10 *Nous comparons nos estimateurs aux estimateurs des moments pondérés [37] de préférence aux estimateurs du maximum de vraisemblance communément utilisés pour la plupart des lois en raison de leurs mauvaises performances dans le cadre des lois GPD mises en évidence par de Hosking et Wallis [37], en particulier dans le cas de petits*

échantillons. De plus, puisque nous introduisons les estimateurs des moments pondérés dans la procédure bayésienne peu informative que nous utilisons, cela nous permet de constater les performances de l'estimation bayésienne lorsque les estimateurs des moments pondérés qui lui apportent de l'information sont mauvais.

Les vraies valeurs des paramètres étant $\alpha = 3$ et $\beta = 3$, les estimations les plus proches des vraies valeurs sont données par la moyenne et la médiane des 500 dernières valeurs de l'algorithme de Gibbs (pour 1000 itérations). D'autre part, afin de comparer nos résultats aux autres travaux, nous allons dans la suite souvent nous ramener à l'estimation des paramètres γ et σ du paramétrage usuellement utilisé. Dans notre cas, les vraies valeurs sont $\gamma = 1/\alpha = 1/3$ et $\sigma = \alpha/\beta = 1$, et les estimateurs sont les suivants

estimations par	Hosking et Wallis	mode de l'algorithme de Gibbs	médiane de l'algorithme de Gibbs	moyenne de l'algorithme de Gibbs
$\hat{\gamma}$	0.4191	0.4291	**0.3610**	**0.3467**
$\hat{\sigma}$	**1.0135**	0.8583	**0.9703**	**0.9673**

Les plus proches sont les estimateurs bayésiens de la médiane et de la moyenne, que ce soit pour le couple (α, β) ou pour le couple (γ, σ).

Dans notre contexte, nous nous intéressons principalement au calcul de quantiles. Nous souhaitons donc appréhender l'influence de nos erreurs d'estimation sur le calcul des quantiles, et donc sur la fonction de répartition (Fdr). Nous traçons donc à présent les graphes des différentes Fdr estimées, que l'on compare à la Fdr de la vraie loi de simulation des données. Ces fonctions estimées sont les Fdr de :

- la loi GPD de paramètres les estimateurs des moments pondérés [37],
- la loi GPD de paramètres les estimateurs bayésiens pour le mode a posteriori,
- la loi GPD de paramètres les estimateurs bayésiens pour la médiane a posteriori,
- la loi GPD de paramètres les estimateurs bayésiens pour la moyenne a posteriori,
- la loi prédictive (mélange continu de lois GPD avec pour loi de mélange la loi a posteriori des paramètres).

Nous avons réalisé deux graphiques séparés (voir la figure 3.4) montrant les comportements des différentes Fdr en partie centrale (sur les valeurs les plus probables de la distribution), et en queue de distribution (pour les événements plus rares et les quantiles extrêmes que nous souhaitons estimer). Si en partie centrale les différentes Fdr estimées semblent assez proches de la vraie Fdr, en queue de distribution, on voit nettement que la Fdr correspondant aux estimateurs des moments pondérés ainsi que celle correspondant aux estimateurs bayésiens du mode donneront de mauvaises estimations des quantiles extrêmes (tracer un trait horizontal au niveau du $1-p$ choisi, puis un trait vertical à l'endroit où l'on coupe la courbe de la Fdr choisie. L'estimateur du quantile d'ordre $1-p$ se lit alors sur l'axe des abscisses).

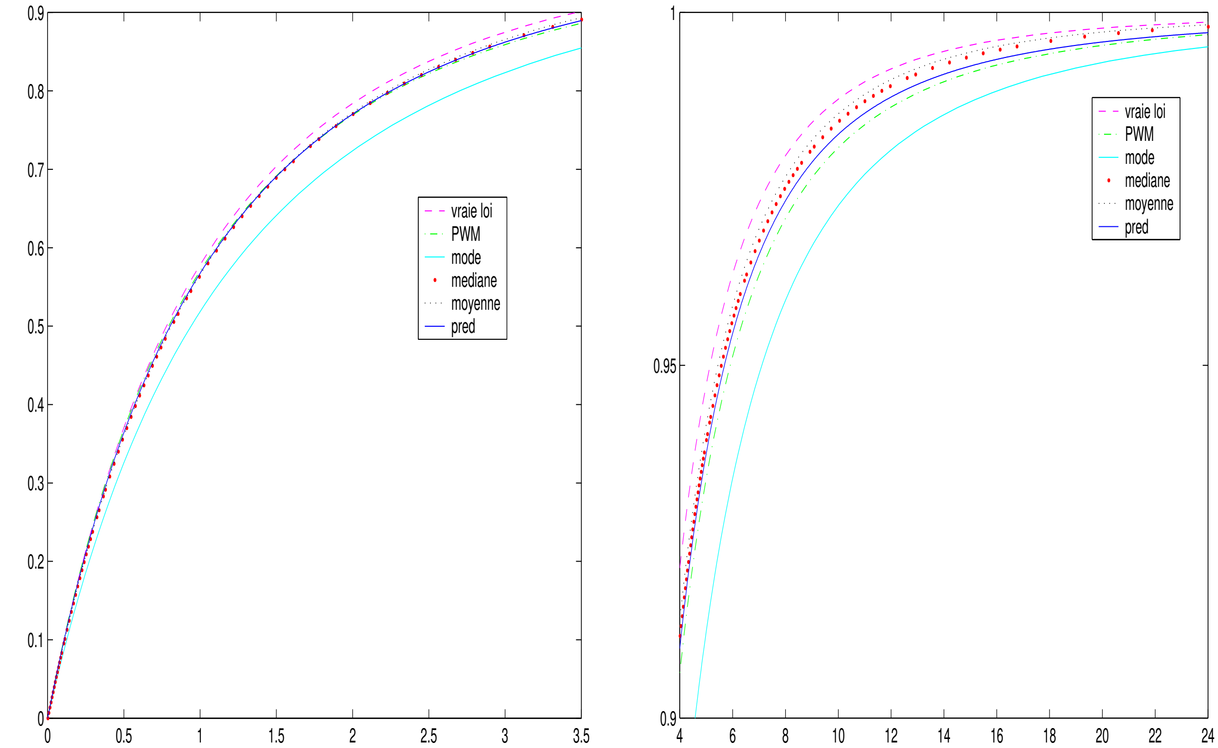

FIG. 3.4 – *Graphe des différentes Fdr pour la partie la plus probable de la distribution (à gauche) et la queue de distribution (à droite) – tirets : vraie loi GPD, discontinu : loi GPD estimée par la méthode de Hosking et Wallis, trait plein clair : loi GPD estimée par la méthode bayésienne du mode a posteriori, points : loi GPD estimée par la méthode bayésienne de la médiane a posteriori, pointillés : loi GPD estimée par la méthode bayésienne de la moyenne a posteriori, trait plein foncé : loi prédictive obtenue avec la méthode bayésienne*

3.2.3.2 Résultats de simulations intensives dans le cas de la moyenne a priori

Nous allons maintenant présenter des tableaux résumant des simulations intensives sur les calculs des estimateurs bayésiens de la GPD dans le cas de la moyenne a priori (pour déterminer les hyperparamètres). On simule $N = 100$ échantillons de loi GPD(α, β), avec $\alpha = \beta$. On choisit des paramètres de simulation α et β égauxcar dans le paramétrage classique, cela correspond à $\gamma = 1/\alpha$ (le paramètre de forme) et $\sigma = 1$ (le paramètre d'échelle). En effet, notre but principal est de bien estimer le paramètre γ qui indique la rapidité de décroissance de la queue de distribution.

Sur les $N = 100$ échantillons simulés, on calcule les estimateurs des moments pondérés de Hosking et Wallis ainsi que les estimateurs bayésiens (par le mode, la moyenne, la médiane a posteriori) pour différentes valeurs d'initialisation de l'algorithme de Gibbs. Puis on calcule des mesures de l'erreur quadratique entre les vraies valeurs des paramètres ((α, β) ou (γ, σ)) et leurs différentes estimations (bayésiennes ou des moments pondérés). Enfin, nous souhaitons aussi évaluer l'impact de l'estimation des paramètres sur les fonctions de survie (Fds) (et à travers celles-ci sur l'estimation des quantiles), pour la partie centrale, ainsi que pour la queue de distribution.

Dans les tableaux suivants sont présentées les moyennes de ces différentes quantités (estimations, erreurs sur les estimations, erreurs sur les Fds) sur les 100 échantillons. Ces tableaux ont été produits dans le cas où la loi gamconII est simulée (au cours de l'algorithme de Gibbs) à l'aide d'une étape de Hastings-Metropolis. On rappelle que les estimateurs bayésiens calculés sont le mode, la moyenne et la médiane des lois a posteriori marginales pour α et β que l'on approche à l'aide des échantillons de valeurs issus de ces lois obtenues par l'algorithme de Gibbs (après atteinte approximative de la loi stationnaire).

On commence par simuler des échantillons de variables aléatoires X_i issues de la loi GPD$(0.5, 0.5)$, puis de GPD$(1, 1)$ et enfin de GPD$(2, 2)$. Dans tous ces cas, $\alpha < 2.5$ (donc $\gamma > 0.4$), et par suite les estimateurs des moments pondérés ne donnent pas de bons résultats. Pourtant, les estimateurs bayésiens obtenus (en introduisant comme information ces estimateurs des moments pondérés, qui sont erronés, voire nettement erronés dans le premier cas) sont proches des vraies valeurs des paramètres (voir les tableaux 3.1 à 3.9 pages 116 à 120). On peut tout d'abord remarquer que les résultats (moyenne et erreurs) des différents estimateurs bayésiens restent très proches lorsque l'on change le point initial de l'algorithme de Gibbs. Ceci confirme que l'on atteint approximativement la loi stationnaire de l'algorithme de Gibbs, puisque l'on semble ainsi "oublier" le point de départ de l'algorithme.

		estimateurs de		erreurs sur		
		α	β	(α, β)	α	β
moments pondérés		1.0500	18.7035	2522.3611	0.3049	2522.0562
point initial du Gibbs : (1,1)						
estimateurs	mode	0.5161	0.5727	0.0399	0.0045	0.0354
bayésiens	médiane	0.5297	0.6218	0.0538	0.0053	0.0485
par	moyenne	0.5364	0.6537	0.0665	0.0058	0.0606
point initial du Gibbs : (100,100)						
estimateurs	mode	0.5160	0.5600	0.0393	0.0050	0.0343
bayésiens	médiane	0.5291	0.6196	0.0545	0.0054	0.0491
par	moyenne	0.5365	0.6509	0.0671	0.0060	0.0612
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs	mode	0.5170	0.5698	0.0458	0.0059	0.0398
bayésiens	médiane	0.5299	0.6222	0.0547	0.0054	0.0493
par	moyenne	0.5370	0.6538	0.0683	0.0060	0.0623

TAB. 3.1 – *Estimation des paramètres (α, β) pour des échantillons de loi GPD$(0.5, 0.5)$*

		estimateurs de γ	estimateurs de σ	erreurs sur (γ, σ)	erreurs sur γ	erreurs sur σ
moments pondérés		0.9543	18.2665	2430.7974	1.0951	2429.7022
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	1.9681	1.1057	0.1569	0.0614	0.0955
	médiane	1.9165	1.1686	0.1729	0.0606	0.1122
	moyenne	1.8924	1.2129	0.1975	0.0631	0.1344
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	1.9717	1.0832	0.1640	0.0661	0.0980
	médiane	1.9197	1.1657	0.1765	0.0619	0.1147
	moyenne	1.8926	1.2068	0.1974	0.0640	0.1334
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	1.9710	1.0968	0.1660	0.0676	0.0985
	médiane	1.9160	1.1683	0.1701	0.0596	0.1105
	moyenne	1.8905	1.2110	0.1973	0.0631	0.1342

TAB. 3.2 – *Estimation des paramètres* (γ, σ) *pour des échantillons de loi GPD*$(0.5, 0.5)$

		erreurs sur le centre	erreurs en queue
moments pondérés		239.9471	2712.7202
point initial du Gibbs : (1,1)			
estimateurs bayésiens par	mode	4.1422	76.6402
	médiane	3.5963	90.8402
	moyenne	3.5963	90.8402
loi prédictive		3.5255	52.4450
point initial du Gibbs : (100,100)			
estimateurs bayésiens par	mode	4.6582	88.1355
	médiane	3.7762	87.6236
	moyenne	3.6183	93.2278
loi prédictive		3.5360	52.3713
point initial du Gibbs : estimateurs des moments pondérés			
estimateurs bayésiens par	mode	4.4727	101.0367
	médiane	3.5504	86.5047
	moyenne	3.5415	92.8373
loi prédictive		3.4633	51.3285

TAB. 3.3 – *Estimation de la fonction de survie pour des échantillons de loi GPD*$(0.5, 0.5)$

		estimateurs de α	estimateurs de β	erreurs sur (α, β)	erreurs sur α	erreurs sur β
moments pondérés		1.3725	1.8194	1.1082	0.2092	0.8990
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	1.0144	1.0260	0.1131	0.0326	0.0805
	médiane	1.0573	1.1307	0.1638	0.0394	0.1244
	moyenne	1.0833	1.1936	0.2113	0.0480	0.1634
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	1.0216	1.0497	0.1577	0.0425	0.1152
	médiane	1.0655	1.1465	0.1815	0.0433	0.1382
	moyenne	1.0910	1.2100	0.2298	0.0507	0.1791
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	1.0161	1.0243	0.1322	0.0406	0.0917
	médiane	1.0606	1.1365	0.1747	0.0423	0.1324
	moyenne	1.0881	1.2012	0.2311	0.0524	0.1787

TAB. 3.4 – *Estimation des paramètres* (α, β) *pour des échantillons de loi GPD*$(1,1)$

		estimateurs de γ	estimateurs de σ	erreurs sur (γ, σ)	erreurs sur γ	erreurs sur σ
moments pondérés		0.7515	1.3289	0.2559	0.0769	0.1790
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	1.0133	1.0060	0.0647	0.0263	0.0384
	médiane	0.9724	1.0589	0.0671	0.0243	0.0428
	moyenne	0.9508	1.0897	0.0759	0.0262	0.0496
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	1.0115	1.0199	0.0811	0.0297	0.0514
	médiane	0.9654	1.0646	0.0698	0.0244	0.0454
	moyenne	0.9440	1.0962	0.0794	0.0263	0.0531
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	1.0164	1.0038	0.0730	0.0297	0.0433
	médiane	0.9711	1.0596	0.0679	0.0258	0.0421
	moyenne	0.9487	1.0902	0.0772	0.0279	0.0493

TAB. 3.5 – *Estimation des paramètres* (γ, σ) *pour des échantillons de loi GPD*$(1,1)$

		erreurs sur le centre	erreurs en queue
moments pondérés		5.2593	488.6246
point initial du Gibbs : (1,1)			
estimateurs bayésiens par	mode	3.3466	104.7564
	médiane	2.7519	110.3050
	moyenne	2.7652	128.0334
loi prédictive		2.6996	61.2702
point initial du Gibbs : (100,100)			
estimateurs bayésiens par	mode	4.0968	132.9339
	médiane	2.7578	118.5945
	moyenne	2.7434	132.8808
loi prédictive		2.6668	62.3744
point initial du Gibbs : estimateurs des moments pondérés			
estimateurs bayésiens par	mode	4.1655	129.6272
	médiane	2.6846	116.9767
	moyenne	2.7189	137.9132
loi prédictive		2.6500	63.7559

TAB. 3.6 – *Estimation de la fonction de survie pour des échantillons de loi GPD*$(1,1)$

		estimateurs de		erreurs sur		
		α	β	(α,β)	α	β
moments pondérés		2.4466	2.5666	3.3443	1.1277	2.2166
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	1.9625	1.8892	0.9392	0.3342	0.6049
	médiane	2.1930	2.2381	1.7333	0.5614	1.1719
	moyenne	2.3383	2.4492	2.5265	0.8270	1.6996
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	1.9433	1.9017	1.0656	0.3522	0.7135
	médiane	2.1732	2.2037	1.8431	0.6099	1.2332
	moyenne	2.3110	2.4158	2.5499	0.8290	1.7209
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	1.9495	1.8964	1.0572	0.3215	0.7358
	médiane	2.1794	2.2201	1.7270	0.5624	1.1646
	moyenne	2.3210	2.4397	2.6636	0.8261	1.8376

TAB. 3.7 – *Estimation des paramètres* (α,β) *pour des échantillons de loi GPD*$(2,2)$

		estimateurs de γ	estimateurs de σ	erreurs sur (γ, σ)	erreurs sur γ	erreurs sur σ
moments pondérés		0.4565	1.0161	0.0483	0.0197	0.0286
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	0.5437	0.9493	0.0444	0.0178	0.0266
	médiane	0.4930	0.9967	0.0420	0.0156	0.0264
	moyenne	0.4674	1.0206	0.0438	0.0162	0.0276
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	0.5461	0.9621	0.0463	0.0164	0.0299
	médiane	0.4979	0.9898	0.0396	0.0151	0.0245
	moyenne	0.4727	1.0175	0.0423	0.0157	0.0266
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	0.5480	0.9503	0.0531	0.0196	0.0335
	médiane	0.4966	0.9935	0.0414	0.0158	0.0255
	moyenne	0.4725	1.0196	0.0454	0.0170	0.0284

TAB. 3.8 – *Estimation des paramètres (γ, σ) pour des échantillons de loi GPD$(2, 2)$*

		erreurs sur le centre	erreurs en queue
moments pondérés		2.6393	292.2566
point initial du Gibbs : (1,1)			
estimateurs bayésiens par	mode	3.0927	130.8545
	médiane	2.6466	163.1511
	moyenne	2.5691	216.5239
loi prédictive		2.4138	80.5171
point initial du Gibbs : (100,100)			
estimateurs bayésiens par	mode	3.2220	113.0376
	médiane	2.4680	165.0744
	moyenne	2.4492	206.8278
loi prédictive		2.3464	80.5602
point initial du Gibbs : estimateurs des moments pondérés			
estimateurs bayésiens par	mode	3.1519	113.8710
	médiane	2.4449	160.0570
	moyenne	2.4437	206.7846
loi prédictive		2.3530	82.4594

TAB. 3.9 – *Estimation de la fonction de survie pour des échantillons de loi GPD$(2, 2)$*

Les résultats des trois types d'estimateurs bayésiens des paramètres $((\alpha, \beta)$ ou $(\gamma, \sigma))$ sont assez semblables du point de vue du biais (voir la moyenne des estimations) ou du point de vue de la variance (voir les erreurs sur les estimateurs qui prennent en compte le biais et la variance). Il semble souhaitable, pour de petites valeurs de α et β, d'utiliser l'estimateur bayésien par le mode a posteriori, en particulier pour estimer le couple (α, β), ou l'estimateur bayésien par la médiane a posteriori, surtout pour estimer le couple (γ, σ). Par contre, lorsque l'on s'intéresse aux erreurs sur les Fds (qui impliquent des erreurs sur les quantiles), les différences entre les trois types d'estimateurs bayésiens sont plus marquées. Dans ce cadre, on est conduit à utiliser de préférence la loi prédictive ou l'estimateur bayésien par la médiane a posteriori.

On se place maintenant dans le cadre $\alpha > 2{,}5$ (c'est-à-dire $\gamma < 0.4$) où les estimateurs des moments pondérés donnent de bons résultats. On simule successivement selon une loi GPD$(5, 5)$, puis selon une loi GPD$(10, 10)$. Les résultats obtenus sont présentés dans les tableaux 3.10 à 3.15. Comme précédemment, on peut remarquer que les résultats des différents estimateurs bayésiens semblent à peu près indépendants du point initial de l'algorithme de Gibbs. Cela confirme que l'on atteint approximativement la loi stationnaire de l'algorithme de Gibbs. Pour les paramètres (α, β), les estimations obtenues par la méthode des moments pondérés restent éloignées de celles obtenues par les différentes variantes de la méthode bayésienne (par le mode, la moyenne ou la médiane a posteriori). Par contre, pour les paramètres (γ, σ) ces différentes estimations sont plus proches, les estimateurs bayésiens semblant sensiblement meilleurs, notamment en terme d'erreurs (voir les tableaux 3.11 et 3.14).

		estimateurs de		erreurs sur		
		α	β	(α, β)	α	β
moments pondérés		7.8733	8.6586	233.5897	89.5299	144.0598
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	4.4971	4.6702	11.3971	4.6297	6.7674
	médiane	5.6439	5.9556	22.3121	8.9284	13.3837
	moyenne	6.8283	7.3702	59.2083	23.5178	35.6906
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	4.6397	4.9304	28.7583	10.6048	18.1534
	médiane	5.6943	5.9832	21.5289	8.6213	12.9077
	moyenne	6.9198	7.4377	70.9667	28.5977	42.3690
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	4.6375	5.0129	39.3812	13.7764	25.6048
	médiane	5.7983	6.2047	28.1779	10.2691	17.9088
	moyenne	6.9276	7.5900	74.6785	27.1524	47.5260

TAB. 3.10 – *Estimation des paramètres (α, β) pour des échantillons de loi GPD$(5, 5)$*

		estimateurs de		erreurs sur		
		γ	σ	(γ, σ)	γ	σ
moments pondérés		0.1981	1.0304	0.0320	0.0094	0.0226
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	0.2567	1.0247	0.0644	0.0108	0.0537
	médiane	0.2111	1.0275	0.0244	0.0062	0.0183
	moyenne	0.1872	1.0456	0.0261	0.0060	0.0200
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	0.2593	1.0389	0.1460	0.0111	0.1349
	médiane	0.2112	1.0216	0.0243	0.0067	0.0176
	moyenne	0.1885	1.0387	0.0259	0.0063	0.0196
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	0.2606	1.0389	0.1177	0.0109	0.1068
	médiane	0.2064	1.0331	0.0265	0.0060	0.0205
	moyenne	0.1833	1.0510	0.0285	0.0058	0.0227

TAB. 3.11 – *Estimation des paramètres (γ, σ) pour des échantillons de loi GPD$(5, 5)$*

		erreurs sur le centre	erreurs en queue
moments pondérés		1.8882	133.1590
point initial du Gibbs : (1,1)			
estimateurs bayésiens par	mode	6.9496	131.0031
	médiane	1.9577	70.6222
	moyenne	2.0017	97.0599
loi prédictive		1.9686	63.3389
point initial du Gibbs : (100,100)			
estimateurs bayésiens par	mode	13.2566	173.9650
	médiane	1.8749	74.4207
	moyenne	1.9599	102.0981
loi prédictive		1.9203	67.9633
point initial du Gibbs : estimateurs des moments pondérés			
estimateurs bayésiens par	mode	9.8888	106.7868
	médiane	2.0095	63.4261
	moyenne	2.0894	83.4895
loi prédictive		2.0423	60.7353

TAB. 3.12 – *Estimation de la fonction de survie pour des échantillons de loi GPD$(5, 5)$*

		estimateurs de		erreurs sur		
		α	β	(α,β)	α	β
moments pondérés		25.9249	26.4127	6938.3136	3424.9513	3513.3623
point initial du Gibbs : (1,1)						
estimateurs	mode	11.2670	10.6567	798.79670	371.4641	427.3329
bayésiens	médiane	12.0640	12.2667	395.2135	176.6953	218.5182
par	moyenne	16.8815	17.3774	1443.7068	660.2799	783.4269
point initial du Gibbs : (100,100)						
estimateurs	mode	10.6260	10.8302	777.7266	357.6751	420.0515
bayésiens	médiane	12.3520	12.2527	423.4184	212.8070	210.6113
par	moyenne	18.3437	18.6276	1920.1966	912.1873	1008.0093
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs	mode	10.2952	9.5916	466.3494	251.3449	215.0044
bayésiens	médiane	11.6659	11.5303	262.7852	132.8726	129.9125
par	moyenne	17.6724	17.7249	1415.9671	724.4668	691.5003

TAB. 3.13 – *Estimation des paramètres* (α,β) *pour des échantillons de loi GPD*$(10,10)$

		estimateurs de		erreurs sur		
		γ	σ	(γ,σ)	γ	σ
moments pondérés		0.1260	0.9747	0.0247	0.0071	0.0176
point initial du Gibbs : (1,1)						
estimateurs	mode	0.1688	0.9328	0.0518	0.0113	0.0405
bayésiens	médiane	0.1332	0.9787	0.0226	0.0054	0.0172
par	moyenne	0.1153	0.9928	0.0202	0.0040	0.0162
point initial du Gibbs : (100,100)						
estimateurs	mode	0.1754	0.9843	0.0745	0.0121	0.0625
bayésiens	médiane	0.1337	0.9720	0.0221	0.0054	0.0168
par	moyenne	0.1146	0.9846	0.0196	0.0040	0.0156
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs	mode	0.1688	0.9433	0.0816	0.0113	0.0703
bayésiens	médiane	0.1331	0.9708	0.0223	0.0055	0.0168
par	moyenne	0.1129	0.9852	0.0210	0.0042	0.0168

TAB. 3.14 – *Estimation des paramètres* (γ,σ) *pour des échantillons de loi GPD*$(10,10)$

		erreurs sur le centre	erreurs en queue
moments pondérés		2.2044	98.8211
point initial du Gibbs : (1,1)			
estimateurs	mode	6.0297	414.3598
bayésiens	médiane	2.2497	67.4802
par	moyenne	2.1877	69.2097
loi prédictive		2.1159	76.8751
point initial du Gibbs : (100,100)			
estimateurs	mode	8.8507	143.1016
bayésiens	médiane	2.2564	73.3691
par	moyenne	2.1877	77.7858
loi prédictive		2.115	82.2928
point initial du Gibbs : estimateurs des moments pondérés			
estimateurs	mode	8.9561	408.4708
bayésiens	médiane	2.2929	73.8463
par	moyenne	2.3532	79.6812
loi prédictive		2.2858	76.4421

TAB. 3.15 – *Estimation de la fonction de survie pour des échantillons de loi GPD*$(10, 10)$

Pour de grandes valeurs de α et β, il semble souhaitable d'utiliser l'estimateur bayésien par la médiane en général (voir le biais et les erreurs sur (α, β) ainsi que sur (γ, σ)) ou l'estimateur bayésien par la moyenne pour estimer le couple (γ, σ) (voir le biais et les erreurs sur (γ, σ)). Lorsque l'on s'intéresse aux erreurs sur les Fds (qui impliquent des erreurs sur les quantiles), on est encore conduit à utiliser de préférence la loi prédictive (qui donne les erreurs les plus faibles) ou l'estimateur bayésien par la médiane (généralement les erreurs les plus faibles pour les paramètres estimés), voire parfois l'estimateur bayésien par la moyenne.

En conclusion, on conseillera d'utiliser l'estimateur bayésien par la médiane a posteriori qui donne de bonnes estimation des couples (α, β) et (γ, σ), ainsi que des Fds, quelles que soient les vraies valeurs des paramètres. Pour l'estimation des Fds (et donc pour l'estimation des quantiles), on recommande aussi d'utiliser la loi prédictive.

3.2.3.3 Simulations intensives dans les autres cas

Nous avons ensuite exploré le cas alternatif du mode a priori pour la détermination des hyperparamètres. L'information est à nouveau apportée par l'échantillon à travers les estimateurs des moments pondérés. Mais pour déterminer les hyperparamètres, on suppose que ces estimateurs sont les modes des lois a priori, marginale sur α et conditionnelle sachant

α sur β (voir le paragraphe 3.2.1.2 page 109). Les résultats obtenus sont présentés dans les tableaux de l'annexe E.1 (page 193). On constate que la stationnarité semble à nouveau approximativement atteinte puisque, les résultats sont semblables quel que soit ce point de départ. Mais, par contre, les résultats sont en général moins bons que ceux obtenus dans le cas de la moyenne a priori, et même parfois que les estimateurs des moments pondérés : les valeurs moyennes obtenues pour les estimateurs bayésiens dans le cas du mode a priori sont plus éloignées des vraies valeurs des paramètres, et les erreurs sont plus importantes. On préfèrera donc utiliser le cas de la moyenne a priori pour déterminer les hyperparamètres.

Dans le but de comparer les résultats obtenus et les temps de calculs pour les deux méthodes de simulation proposées pour la loi gamconII, nous avons reproduit ces simulations lorsque l'on simule, au cours de l'algorithme de Gibbs, la loi gamconII par son approximation normale (voir les tableaux présentés par Garrido [26]). Nous avons tout d'abord constaté que la loi stationnaire semble aussi atteinte au bout de 1000 itérations lorsque l'on simule selon la loi normale approximant la loi gamconII. Pour l'estimation des paramètres ((α,β) ou (γ,σ)) il est difficile de préférer l'une des deux méthodes, chacune d'entre elles donnant dans certains cas la meilleure estimation. Par contre, pour l'estimation de la fonction de survie, les différences constatées sont relativement importantes en faveur de l'étape de Hastings-Metropolis qui donne les plus petites erreurs, sans que le temps de calcul en soit sévèrement augmenté. Nous conseillerons donc en général l'utilisation de l'étape de Hastings-Metropolis pour la simulation selon une loi gamconII. Dans la suite de ce travail, la simulation selon la loi gamconII a toujours été effectuée par une étape de Hastings-Metropolis.

3.3 Application de la procédure bayésienne à des échantillons d'excès

Nous souhaitons à présent utiliser notre procédure bayésienne pour des échantillons d'excès qui, d'après le thérorème de Pickands [2, 3, 41] (voir aussi le théorème 2 page 3), sont approximativement de loi GPD, quand on a convenablement choisi le nombre d'excès.

3.3.1 Calcul des hyperparamètres

Bien que notre but soit à terme d'utiliser un avis d'expert dans le but d'améliorer l'estimation de la loi GPD et des quantiles extrêmes, nous commençons par nous placer dans le cadre bayésien empirique, plus simple, dans lequel l'information a priori provient de l'échantillon et non d'un expert.

Remarquons ensuite que notre procédure n'est valable que pour $\alpha > 0$ et $\beta > 0$. On ne peut donc introduire en tant qu'avis d'expert des estimateurs pouvant produire des valeurs négatives de α ou β. L'estimateur de Hill, bien qu'étant un médiocre estimateur des paramètre de la loi GPD, satisfait cette condition de positivité et a une forme particulièrement simple. Nous introduisons dans notre procédure, à la place d'un avis d'expert, les estimateurs de Hill [34]

$$\tilde{\alpha} = m_n \left(\sum_{i=1}^{m_n} \ln x_{(n-m_n+i)} - \ln x_{(n-m_n)} \right)^{-1} \qquad \text{et} \qquad \tilde{\beta} = x_{(n-m_n)}, \tag{3.13}$$

où m_n est le nombre d'excès, et $x_{(i)}$ désigne la i-ème observation ordonnée. Comme dans le cas d'échantillons de loi GPD (paragraphe 3.2.1 page 107), on choisit les lois a priori pour α et pour β sachant α de telle façon que ces valeurs $\tilde{\alpha}$ et $\tilde{\beta}$ soient les plus probables. Par exemple, nous avons choisi de déterminer les hyperparamètres δ, η et μ de sorte que $\tilde{\alpha}$ et $\tilde{\beta}$ soient les moyennes ou les modes des lois a priori (marginale pour α et conditionnelle sachant α pour β) correspondantes.

À nouveau, afin de simplifier les calculs, on choisit de fixer $\boldsymbol{\delta = 1}$ (voir la remarque 3.7 page 108). Il reste alors à déterminer η et μ. Comme dans le cas d'échantillons de loi GPD (paragraphe 3.2.1 page 107), on considère au choix l'un des deux cas suivants :

Cas de la moyenne a priori Les hyperparamètres sont (cf. paragraphe 3.2.1.1 page 108)

$$\delta = 1\,, \quad \eta = \frac{\tilde{\alpha}+1}{\tilde{\beta}} \qquad \text{et} \qquad \mu = \exp\left(-\frac{2}{\tilde{\alpha}}\right)\frac{\tilde{\alpha}+1}{\tilde{\beta}}\,,$$

où $\tilde{\alpha}$ et $\tilde{\beta}$ sont les estimateurs de Hill de α et β, respectivement. Dans ce cas, on ne sait pas conclure si $\delta \neq 1$ (voir la remarque 3.8 page 109).

Cas du mode a priori Les hyperparamètres sont alors (cf. paragraphe 3.2.1.2 page 109)

$$\delta = 1\,, \quad \eta = \frac{\tilde{\alpha}}{\tilde{\beta}} \qquad \text{et} \qquad \mu = \exp\left(-\frac{1}{\tilde{\alpha}}\right)\frac{\tilde{\alpha}}{\tilde{\beta}}\,,$$

où, respectivement, $\tilde{\alpha}$ et $\tilde{\beta}$ sont les estimateurs de Hill de α et β. Dans ce cas, si $\delta \neq 1$, on peut encore conclure (voir la remarque 3.9 page 109).

3.3.2 Estimation

De même que dans le cas d'échantillons de loi GPD (paragraphe 3.2.2 page 109), la méthode bayésienne nous permet soit d'estimer les paramètres de la loi GPD (on obtient donc une loi GPD estimée) que suivent approximativement les excès, soit d'estimer directement la loi des excès comme un mélange de lois GPD avec pour loi de mélange la loi a posteriori des paramètres (on parle de loi prédictive). On note respectivement $\underline{\alpha}_s$ et $\underline{\beta}_s$ les échantillons, de taille s, des valeurs de α et β simulées par l'algorithme de Gibbs lorsqu'il a approximativement atteint sa loi stationnaire.

Estimation bayésienne de α et β (voir le paragraphe 3.2.2.1 page 110)

- Dans le cas de la moyenne a posteriori, on estime α et β par les moyennes empiriques des échantillons $\underline{\alpha}_s$ et $\underline{\beta}_s$ respectivement (nous estimons ainsi les moyennes des lois a posteriori de α et de β sachant α) : $\hat{\alpha} = \overline{\underline{\alpha}_s}$ et $\hat{\beta} = \overline{\underline{\beta}_s}$.
- Dans le cas de la médiane a posteriori, on estime α et β par les valeurs médianes des échantillons $\underline{\alpha}_s$ et $\underline{\beta}_s$ respectivement (nous estimons ainsi les médianes des lois a posteriori de α et de β sachant α) : $\hat{\alpha} = \alpha_{([s/2])}$ et $\hat{\beta} = \beta_{([s/2])}$, où $[x]$ dénote la partie entière de x, et $x_{(i)}$ la i-ème observation ordonnée.
- Dans le cas du mode a posteriori, on utilise la méthode des estimateurs à noyau (voir l'annexe D.3.5 page 190) pour estimer les densités des échantillons $\underline{\alpha}_s$ et $\underline{\beta}_s$ respectivement, puis on en déduit une estimation de leurs modes respectifs.

Estimation de la loi de l'échantillon (voir le paragraphe 3.2.2.2 page 111) On estime la densité, resp. la Fdr., de la loi prédictive (qui est l'intégrale des densités, resp. des Fdr., des lois GPD avec pour loi de mélange la loi a posteriori de (α, β) sachant les excès $\underline{y}_{m_n}$) par la moyenne empirique :

$$g_{\text{pred}}(x) = \frac{1}{s} \sum_{i=1}^{s} f(x|\,\alpha^{(i)}, \beta^{(i)}) \ , \quad \text{resp.} \quad G_{\text{pred}}(x) = \frac{1}{s} \sum_{i=1}^{s} F(x|\,\alpha^{(i)}, \beta^{(i)}) \ ,$$

où $f(x|\,\alpha, \beta)$ est la densité et $F(x|\,\alpha, \beta)$ la fonction de répartition de la loi GPD de paramètres (α, β), avec $\underline{\alpha}_s = (\alpha^{(1)}, \dots, \alpha^{(s)})$ et $\underline{\beta}_s = (\beta^{(1)}, \dots, \beta^{(s)})$.

On peut être intéressé par l'estimation de l'indice de Pareto, c'est-à-dire le paramètre $\gamma = 1/\alpha$, par exemple dans le cadre des applications financières ou actuarielles. Cependant, dans la plupart des cas, on souhaite surtout estimer des quantiles de la loi des données dont sont issus nos excès. Pour cela, on utilise la méthode des excès (ou POT : Picks Over Threshold, introduite par de Haan et Rootzen [21]) qui nous donne l'estimation suivante pour un quantile extrême d'ordre $1 - p_n$ (avec $p_n \leq 1/n$) :

$$\hat{q}_{GPD,n}(p_n) = \hat{u}_n + \hat{\beta} \left[\left(\frac{np_n}{m_n} \right)^{-1/\hat{\alpha}} - 1 \right] , \tag{3.14}$$

où m_n est le nombre d'excès au-delà du seuil $\hat{u}_n = x_{(n-m_n)}$ (la $(n - m_n)$-ème observation ordonnée), $\hat{\alpha}$ et $\hat{\beta}$ sont des estimateurs (bayésiens ou autres) des paramètres de la loi GPD que suivent approximativement les excès. Dans ce cas, on utilise la loi GPD de paramètres estimés (notamment par notre méthode bayésienne) $(\hat{\alpha}, \hat{\beta})$. Cependant, avec une méthode bayésienne, on dispose aussi d'une loi prédictive (estimation directe de la loi de l'échantillon). On peut donc estimer le quantile en inversant cette loi prédictive :

$$\hat{q}_{\text{pred},\,n}(p_n) = \hat{u}_n + G_{\text{pred}}^{\leftarrow} \left(\frac{np_n}{m_n} \right) ,$$

où $F^{\leftarrow}$ note l'inverse généralisée de la fonction F. Les calculs d'inversion de la fonction de répartition de la loi prédictive sont complexes (contrairement à la loi GPD dont l'inverse admet une forme explicite) et numériquement coûteux. Nous n'avons donc pas encore exploré cette possibilité.

Enfin, on dispose (à travers un échantillon de réalisations $\underline{\alpha}_s$ et $\underline{\beta}_s$) d'une loi a posteriori sur l'espace des paramètres. Cette loi peut être aisément transformée en une loi a posteriori sur l'espace des quantiles extrêmes d'ordre $1 - p_n$, tout simplement par la constitution d'un échantillon de réalisations $\underline{\hat{q}}_s = (q_1, \ldots, q_s)$ tel que

$$q_i = \hat{q}^{(i)}_{GPD,n}(p_n) = \hat{u}_n + \beta^{(i)} \left[\left(\frac{np_n}{m_n} \right)^{-1/\alpha^{(i)}} - 1 \right] ,$$

avec $\underline{\alpha}_s = (\alpha^{(1)}, \ldots, \alpha^{(s)})$ et $\underline{\beta}_s = (\beta^{(1)}, \ldots, \beta^{(s)})$. On déduit de cette loi a posteriori sur le quantile d'ordre $1 - p_n$, une estimation de ce quantile, par la moyenne empirique a posteriori ou la médiane a posteriori :

$$\hat{q}_{\text{post, moy}, n, s}(p_n) = \frac{1}{s} \sum_{i=1}^{s} \hat{q}^{(i)}_{GPD,n}(p_n) \quad \text{ou} \quad \hat{q}_{\text{post, med}, n, s}(p_n) = q_{([s/2])} ,$$

où $[x]$ note la partie entière de x. Cette possibilité non plus n'est pas encore explorée.

3.3.3 Exploration numérique de cette méthode bayésienne pour des échantillons d'excès

Nous appliquons notre méthode bayésienne pour des échantillons d'excès, issus de données simulées de taille $n = 500$. Nous comparons alors les estimations du paramètre γ et du quantile d'ordre $1 - 1/10n = 1 - 1/5000$ de la loi des données originelles, pour différentes valeurs du nombre d'excès m_n (variant de 5 à 495). Les différentes méthodes d'estimation que nous comparons sont : la méthode de Hill, la méthode de Hill généralisée (encore appelée méthode des moments de Dekkers, Einmahl et de Haan [22]), la méthode généralisée des moindre carrés non contraints appelée méthode du Zipf de Beirlant *et al.* [6] (ces trois premières méthodes s'appliquant pour γ réel), notre méthode bayésienne pour la moyenne a priori et la moyenne a posteriori, ainsi que notre méthode bayésienne pour la moyenne a priori et la médiane a posteriori (ces méthodes s'appliquant pour $\gamma > 0$). Nous souhaiterions aussi comparer nos résultats à ceux de la nouvelle méthode de régression de Beirlant *et al.* [5] (méthode pour γ réel). Cependant, pour des raisons de temps d'implémentation et de calcul, nous n'avons pas pu, pour l'instant, l'inclure dans nos essais. Dans chaque cas, nous simulons 100 échantillons de même loi, ce qui nous permet de calculer une moyenne et un intervalle de confiance empiriques pour les différents estimateurs.

Simulons tout d'abord des échantillons de loi de Fréchet de paramètre 1, notée $\mathcal{F}rechet(1)$ (voir l'annexe A page 161). Dans ce cas, la vraie valeur du paramètre γ est $\gamma_0 = 1$, et la vraie valeur du quantile d'ordre $1-1/5000$ est $q_{1-1/5000} = 4999.5$. La figure 3.5 présente les valeurs moyennes obtenues pour les différents estimateurs de γ et du quantile $q_{1-1/5000}$, en fonction du nombre d'excès m_n.

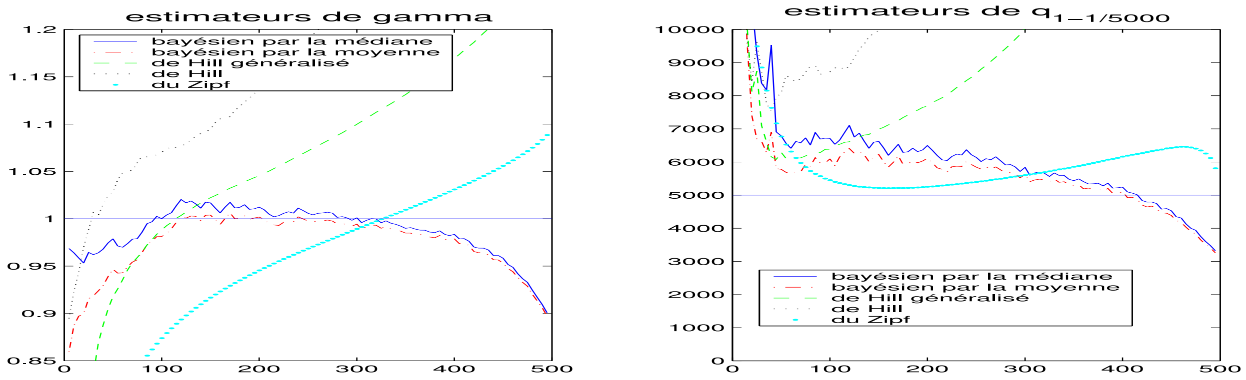

FIG. 3.5 – *Échantillons simulés de loi $\mathcal{F}rechet(1)$ – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) en fonction de m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

On constate tout d'abord une meilleure estimation en moyenne de la valeur de γ et du quantile $q_{1-1/5000}$ avec nos estimateurs bayésiens, et ce pour la plupart des valeurs m_n du nombre d'excès. Seul l'estimateur Zipf du quantile $q_{1-1/5000}$ donne des résultats moyens comparables en terme de biais à ceux des estimateurs bayésiens. Les estimateurs bayésiens sont en outre très stables par rapport au nombre d'excès m_n : leur graphe en fonction de m_n présente une plage horizontale étendue. Seul l'estimateur Zipf du quantile $q_{1-1/5000}$ présente aussi une stablilité horizontale. Ceci permet d'obtenir des estimateurs très peu dépendants du choix du nombre d'excès m_n. De plus, l'existence d'un plateau sur le graphe de l'estimateur en fonction de m_n permet d'utiliser la méthode du *Pareto plot* (ou *"horror plot"*) [34] pour le choix du nombre d'excès. On choisit alors le nombre d'excès à utiliser parmi les valeurs de m_n pour lesquelles on observe une stabilisation horizontale de l'estimateur du quantile.

On souhaite à présent donner une idée de la variance des différents estimateurs calculés. Outre la valeur moyenne, on trace un intervalle de confiance à 90% (voir l'annexe E.2 page 201), en figure E.1 pour les estimateurs de γ, et en figure E.2 pour les estimateurs du quantile $q_{1-1/5000}$. On constate alors que si la variance des différents estimateurs de γ est sensiblement la même, à part pour les grandes valeurs de m_n, la variance des estimateurs du quantile $q_{1-1/5000}$ est par contre visiblement réduite lorsque l'on utilise nos estimateurs bayésiens, en particulier par rapport aux estimateurs Zipf qui présentent un biais moyen équivalent.

En second lieu, nous avons simulé des échantillons de loi de Burr de paramètres $(1,1,1)$, notée $\mathcal{B}urr(1,1,1)$ (voir l'annexe A page 161). La vraie valeur du paramètre γ est encore

$\gamma_0 = 1$, et la vraie valeur du quantile d'ordre $1 - 1/5000$ est $q_{1-1/5000} = 4999$. La figure 3.6 présente les valeurs moyennes obtenues pour les différents estimateurs de γ et du quantile d'ordre $1 - 1/5000$, en fonction du nombre d'excès m_n.

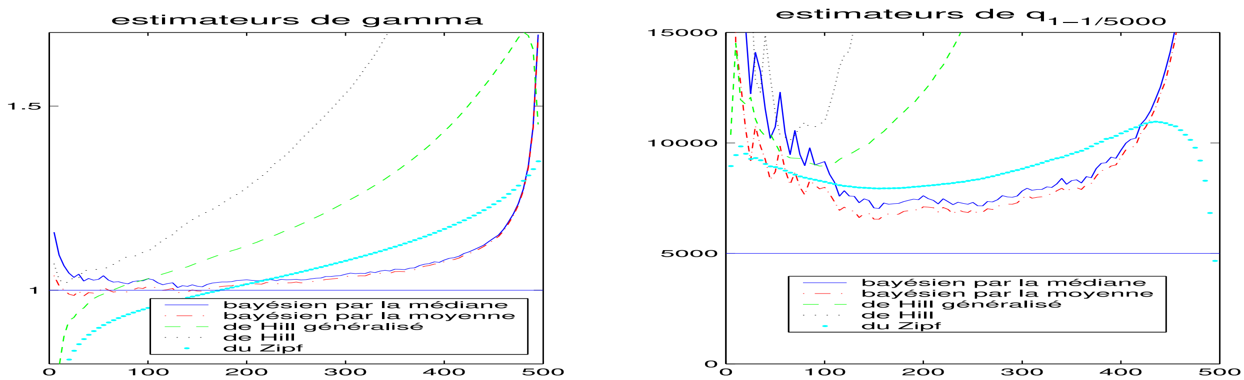

FIG. 3.6 – *Échantillons simulés de loi* $\mathcal{B}urr(1, 1, 1)$ – *Valeurs moyennes des différents estimateurs de* γ *(à gauche) et de* $q_{1-1/5000}$ *(à droite) pour différents* m_n – *continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

On constate une nouvelle amélioration du biais de l'estimation de γ et du quantile d'ordre $1-1/5000$ lorsque l'on utilise nos estimateurs bayésiens, cette fois-ci même par rapport à l'estimateur Zipf du quantile $q_{1-1/5000}$. On retrouve aussi la stabilité des estimateurs bayésiens en fonction du nombre d'excès, c'est-à-dire l'existence d'une plage horizontale étendue dans le graphe des estimateurs bayésiens en fonction de m_n. De plus, la figure E.3 (page 204) montre que la variance des estimateurs bayésiens de γ est plus réduite que celle des autres estimateurs pour les grandes valeurs de m_n. Quant à la variance des estimateurs bayésiens de $q_{1-1/5000}$, elle est bien plus faible que la variance des autres estimateurs, et ce pour la plupart des valeurs de m_n (voir la figure E.4 page 205).

Puis, nous simulons des échantillons de loi de $\mathcal{B}urr(1, 0.5, 2)$ (voir l'annexe A page 161). La vraie valeur du paramètre γ est encore $\gamma_0 = 1$, et la vraie valeur du quantile d'ordre $1 - 1/5000$ est $q_{1-1/5000} = 4859.6$. La figure 3.7 présente les valeurs moyennes obtenues pour les différents estimateurs de γ et du quantile d'ordre $1 - 1/5000$, en fonction du nombre d'excès m_n. Dans ce cas, les estimateurs bayésiens de γ (comme ceux du quantile $q_{1-1/5000}$) sont un peu moins biaisés et légèrement plus stables en fonction du nombre d'excès m_n que les estimateurs de Hill et de Hill généralisé (ils croissent moins vite). Mais, nos estimateurs bayésiens sont un peu plus biaisés que les estimateurs Zipf. Par contre, la variance des estimateurs bayésiens du quantile $q_{1-1/5000}$ est plus faible que celle des estimateurs de Hill, de Hill généralisé, et surtout celle des estimateurs Zipf (voir la figure E.6 page 207), contrairement à la variance des estimateurs de γ qui est sensiblement identique pour tous les estimateurs (voir la figure E.5 page 206). L'introduction d'un avis d'expert sur la queue de distribution permettrait certainement de mieux calibrer les lois a priori et a posteriori, et donc de mieux estimer les quantiles (voir le paragraphe 3.4.3 page 138).

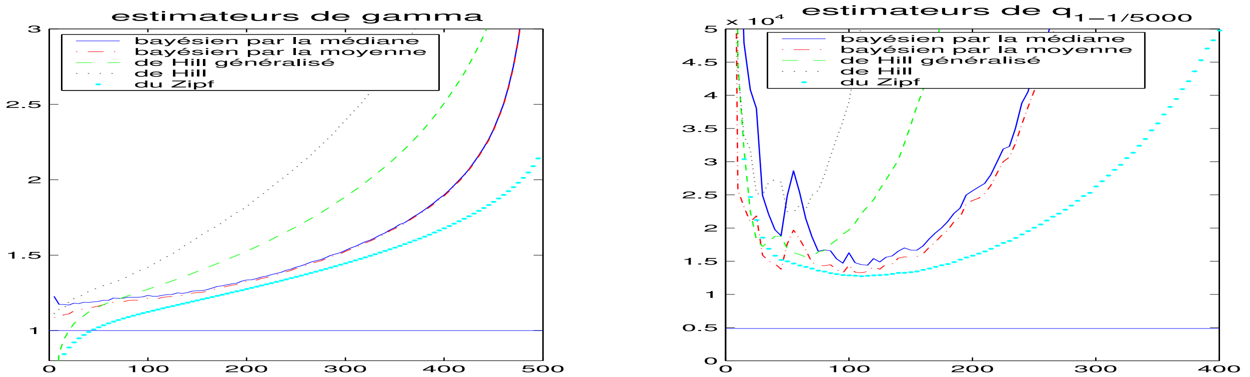

FIG. 3.7 – *Échantillons simulés de loi* $\mathcal{B}urr(1, 0.5, 2)$ – *Valeurs moyennes des différents estimateurs de* γ *(à gauche) et de* $q_{1-1/5000}$ *(à droite) pour différents* m_n – *continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

Par défaut, nous avons fait $k = 1000$ itérations de l'algorithme de Gibbs, puis nous n'avons conservé que les $r = 500$ dernières pour calculer les estimateurs bayésiens. Le fait de faire plus d'itérations au cours de l'algorithme de Gibbs ($k = 5000$ ou 10000), et donc peut-être d'atteindre une meilleure approximation de la loi stationnaire, pourrait permettre d'obtenir des estimateurs plus stables et moins biaisés, notamment pour les quantiles. Nous avons donc appliqué notre procédure bayésienne pour $k = 5000$ itérations de l'algorithme de Gibbs, en n'en conservant que les $r = 500$ dernières pour calculer les estimateurs bayésiens. Les résultats ont été identiques, ce qui montre que notre procédure converge rapidement.

À présent, nous simulons des échantillons qui sont la valeur absolue d'échantillons de lois de Student à ν degrés de liberté. On parle alors de loi de Student absolue, $t_{Abs}(\nu)$ (voir l'annexe A page 161). Tout d'abord simulons des échantillons de loi $t_{Abs}(1)$. La vraie valeur du paramètre γ est encore $\gamma_0 = 1$, et la vraie valeur du quantile d'ordre $1 - 1/5000$ est $q_{1-1/5000} = 3186$. La figure 3.8 présente les valeurs moyennes obtenues pour les différents estimateurs de γ et du quantile d'ordre $1 - 1/5000$, en fonction du nombre d'excès m_n.

Cette fois, la réduction du biais d'estimation de γ lorsque l'on utilise nos estimateurs bayésiens est moins nette pour les grandes valeurs de m_n, et le biais reste alors assez important. Cependant, pour m_n inférieur à 250 on conserve l'avantage d'une relative stabilité et d'un biais réduit pour les estimateurs bayésiens. Ceci n'est pas trop restrictif puisque l'on conseille généralement d'utiliser les valeurs de m_n petites devant n (= 500 ici). Concernant le quantile $q_{1-1/5000}$, les estimateurs bayésiens permettent une réduction du biais pour les plus petites valeurs de m_n (< 200), notamment avec la méthode de la moyenne a posteriori. La stabilité des estimateurs bayésiens a tendance à disparaître lorsque m_n croît, et leur biais augmente. La variance des estimateurs de γ est légèrement réduite lorsque l'on utilise les estimateurs bayésiens (voir la figure E.7 page 208). Quant à l'estimation de $q_{1-1/5000}$ (voir la figure E.8 page 209), la variance des estimateurs bayésiens semble légèrement plus importante pour les petites et moyennes valeurs de m_n, alors qu'elle est très réduite pour les grandes valeurs de m_n, pour lesquelles cependant le biais est plus important.

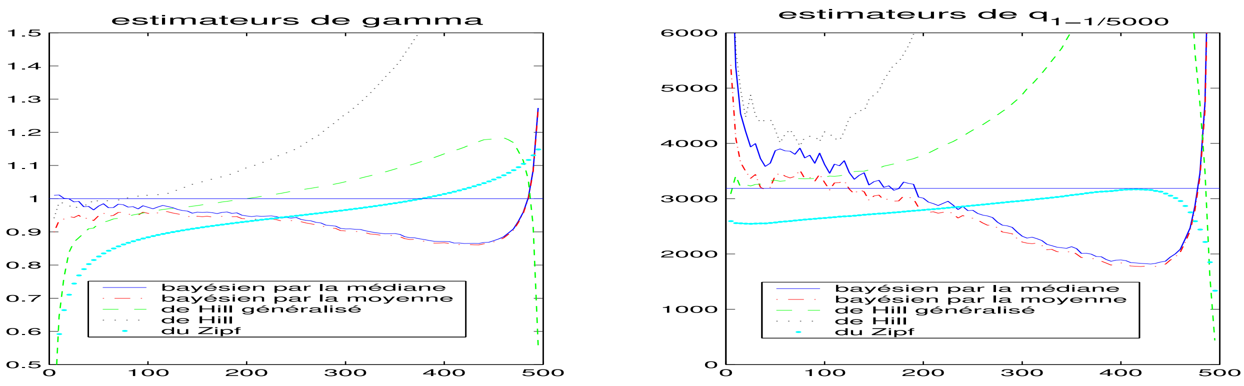

FIG. 3.8 – *Échantillons simulés de loi $t_{Abs}(1)$ – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) pour différents m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

Nous simulons maintenant des échantillons de loi $t_{Abs}(2)$. La vraie valeur du paramètre γ est cette fois $\gamma_0 = 1/2 = 0.5$, et la vraie valeur du quantile d'ordre $1 - 1/5000$ est $q_{1-1/5000} = 70.7001$. La figure 3.9 présente les valeurs moyennes obtenues pour les différents estimateurs de γ et du quantile d'ordre $1 - 1/5000$, en fonction du nombre d'excès m_n. On peut remarquer que les estimateurs bayésiens du quantile sont très satisfaisants. En effet, ils sont peu biaisés et leur variance est faible (voir la figure E.10 page 211), lors du plateau horizontal du graphe de ces estimateurs en fonction de m_n. L'estimateur de Hill généralisé de $q_{1-1/5000}$ présente un biais semblable, mais une variance bien plus élevée. Pour l'estimateur Zipf du quantile, le plateau est bien plus large que pour les estimateurs bayésiens, et le biais très faible, mais la variance est à nouveau nettement plus importante. Au contraire, les estimateurs bayésiens de γ présentent un biais plus important que l'estimateur de Hill généralisé, notamment au niveau du plateau du graphe, mais là aussi une variance plus faible (voir la figure E.9 page 210).

Nous savons que le vrai paramètre γ concerne la loi GPD, c'est-à-dire la loi asymptotique des excès. L'estimation que nous en faisons est basée sur un échantillon d'excès, et donc sur une loi des excès qui n'est pas tout à fait une loi GPD. Si l'on souhaite estimer précisément γ, on cherche à approcher sa vraie valeur théorique γ_0. Mais lorsque l'on souhaite appliquer la méthode POT pour estimer un quantile extrême, une bonne estimation de la vraie valeur théorique γ_0 peut entraîner un biais dans l'estimation du quantile, puisque l'on utilise une loi asymptotique qui n'est pas forcément atteinte en pratique. Il se peut qu'une estimation biaisée de γ produise une loi GPD plus proche de la vraie loi des excès, et donc une meilleure estimation des quantiles, comme cela semble être le cas ici. L'introduction d'un avis d'expert, notamment sur les queues de distribution comme nous le proposons au paragraphe suivant, devrait permettre d'indiquer à l'estimateur bayésien sur quelle quantité se focaliser.

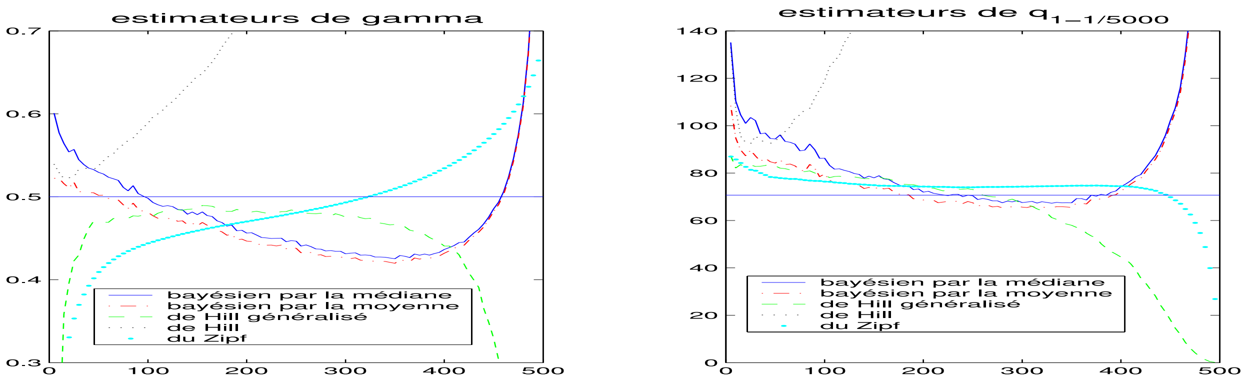

FIG. 3.9 – *Échantillons simulés de loi $t_{Abs}(2)$ – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) pour différents m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

À l'inverse, pour des échantillons simulés de loi $t_{Abs}(4)$ (les vraies valeurs sont $\gamma_0 = 1/4 = 0.25$ et $q_{1-1/5000} = 13.0337$), l'estimation bayésienne produit un biais presque nul pour l'estimation de γ (et une variance très faible), alors que les estimateurs du quantile $q_{1-1/5000}$ sont plus biaisées (mais de variance relativement petite), voir la figure 3.10. Dans les deux cas on observe une plage de stabilité des estimateurs en fonction de m_n, importante pour γ (pour presque toutes les valeurs de m_n), plus réduite pour le quantile (lorsque $m_n \in [5, 250]$ environ). Les figures E.11 et E.12 pages 212 et 213 montrent que notre méthode bayésienne produit des estimateurs de γ et du quantile $q_{1-1/5000}$ de variance plus réduite que celles des autres estimateurs.

On simule maintenant des échantillons de loi $t_{Abs}(8)$. La vraie valeur du paramètre γ est alors $\gamma_0 = 1/8 = 0.125$, et la vraie valeur du quantile d'ordre $1 - 1/5000$ est $q_{1-1/5000} = 6.442$. La difficulté dans ce cas, notamment pour l'estimation bayésienne de γ vient du fait que la vraie valeur γ_0 est très proche de 0. Puisque nos estimateurs bayésiens ne peuvent être négatifs, on peut craindre qu'ils ne surestiment γ, et même que ce biais "artificiel" pour γ ait des répercussions sur l'estimation du quantile $q_{1-1/5000}$. En effet, la figure 3.11 nous montre que les estimateurs bayésiens de γ présentent un biais positif (plus réduit cependant que le biais, négatif, de l'estimateur de Hill généralisé), mais aussi une grande stabilité en fonction de m_n, et une variance très réduite (voir la figure E.13 page 214). Les estimations bayésiennes du quantile $q_{1-1/5000}$, quant à elles, sont très biaisées, croissantes en fonction de m_n, mais ont une variance plutôt réduite (voir la figure E.14 page 215).

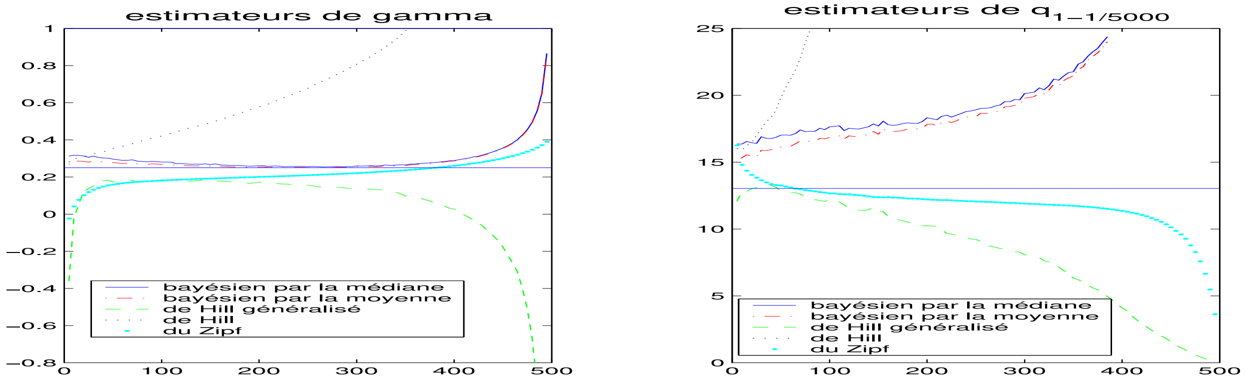

FIG. 3.10 – *Échantillons simulés de loi $t_{Abs}(4)$ – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) pour différents m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

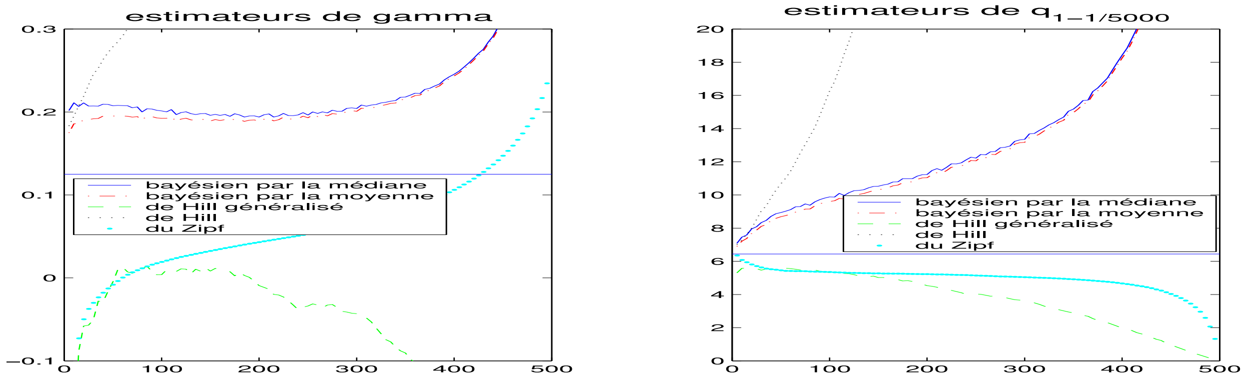

FIG. 3.11 – *Échantillons simulés de loi $t_{Abs}(8)$ – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) pour différents m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

On simule à présent des échantillons de loi loggamma de paramètre 2, notée $\mathcal{L}\Gamma(2)$, c'est-à-dire des échantillons issus d'une variable aléatoire dont l'exponentielle suit une loi gamma : $\mathcal{G}amma(2,1)$. La vraie valeur du paramètre γ est à nouveau $\gamma_0 = 1$, et la vraie valeur du quantile d'ordre $1 - 1/5000$ est $q_{1-1/5000} = 60011$. La figure 3.12 présente les valeurs moyennes obtenues pour les différents estimateurs de γ et du quantile d'ordre $1-1/5000$, en fonction du nombre d'excès m_n. On constate que les estimateurs bayésiens de γ ainsi que du quantile $q_{1-1/5000}$ sont particulièrement stables en fonction de m_n mais biaisés. La variance des estimateurs bayésiens de γ est sensiblement identique à celle des autres estimateurs (voir la figure E.15 page 216). De même, la variance des estimateurs bayésiens du quantile $q_{1-1/5000}$ est comparable à celle des autres estimateurs pour les petites et moyennes valeurs de m_n (voir la figure E.16 page 217), mais plus réduite pour les grandes valeurs de m_n (alors que le biais reste à peu près constant en fonction de m_n).

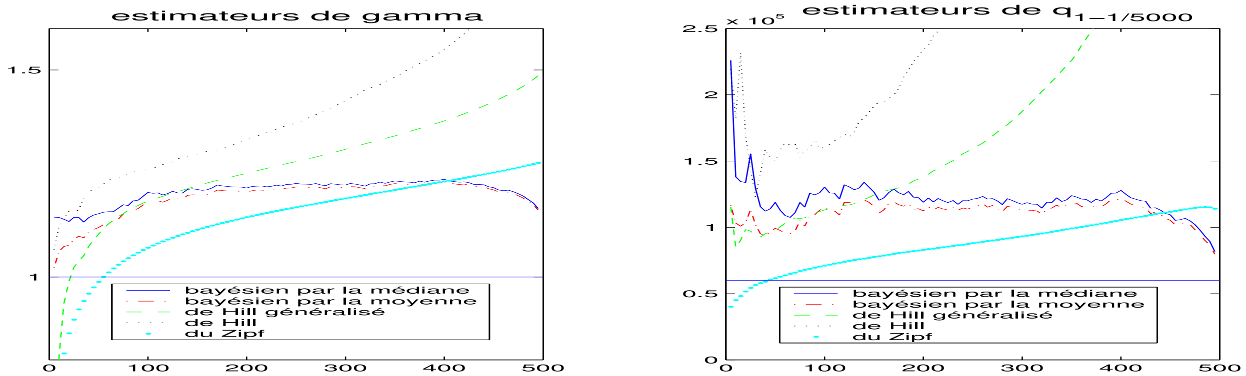

FIG. 3.12 – *Échantillons simulés de loi $\mathcal{L}\Gamma(2)$ – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) pour différents m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

Comme on l'a déjà remarqué, l'introduction d'un avis d'expert pourrait améliorer le biais d'estimation des quantiles. Cette solution est explorée dans le paragraphe suivant.

3.4 Première introduction (partielle) d'un avis d'expert

Nous supposons à présent que nous disposons d'un avis d'expert sur la queue de la loi des données dont sont issus les excès. Le type d'information sur la queue de distribution que l'on peut attendre d'un expert concerne les valeurs extrêmes (voir par exemple Coles et Tawn [16]). Par exemple, l'expert peut déterminer une valeur $q_{\max}$ de la quantité d'intérêt dont il pense qu'elle est rarement atteinte. Pour quantifier cette rareté, il faut, avec l'aide de l'expert, déterminer un intervalle $[p_1, p_2]$ encadrant le risque p associé à $q_{\max}$, c'est-à-dire la probabilité p de dépasser $q_{\max}$.

À présent que nous disposons d'un avis d'expert, il nous faut l'utiliser pour déterminer les hyperparamètres δ, η et μ. Pour cela, on souhaite tout d'abord traduire l'encadrement $[p_1, p_2]$ du risque p associé à $q_{\max}$ par un encadrement sur l'un des paramètres α ou β (la traduction de cette contrainte sur le couple (α, β) étant impossible), ceci étant réalisé en fixant comme précédemment l'autre paramètre à son estimateur de Hill.

En pratique, nous proposons de traduire le fait que le risque associé à $q_{\max}$ est $p \in [p_1, p_2]$ par l'encadrement suivant pour $q_{\max}$, utilisant l'approximation GPD d'un quantile extrême déduite du théorème de Pickands (voir le théorème 10 et la formule (2.4) au paragraphe 2.3.1

page 92) :

$$q_{GPD,n}(p_2) \leq q_{\max} \leq q_{GPD,n}(p_1)\,, \quad \text{où} \quad q_{GPD,n}(p) = u_n + \beta\left[\left(\frac{np}{m_n}\right)^{-1/\alpha} - 1\right].$$

En remplaçant le seuil u_n par $\hat{u}_n = x_{(n-m_n)}$, on en déduit que

$$\beta\left[\left(\frac{np_2}{m_n}\right)^{-1/\alpha} - 1\right] \leq q_{\max} - x_{(n-m_n)} \leq \beta\left[\left(\frac{np_1}{m_n}\right)^{-1/\alpha} - 1\right]. \tag{3.15}$$

Il s'agit ensuite de modifier cet encadrement, soit à β fixé (égal à son estimateur de Hill) lorsque l'on souhaite obtenir un encadrement de α (pour déterminer à partir de l'avis d'expert la loi a priori sur α), soit à α fixé (égal à son estimateur de Hill) lorsque l'on souhaite obtenir un encadrement de β (pour déterminer à partir de l'avis d'expert la loi a priori sur β sachant α).

3.4.1 Cas β fixé : l'avis d'expert agit sur α

Dans ce cas on suppose que $\beta = \hat{\beta}$, son estimateur de Hill. De l'équation (3.15), on déduit l'encadrement suivant pour α :

$$\alpha_1 \leq \alpha \leq \alpha_2 \quad \text{où} \quad \alpha_1 = \frac{\ln(m_n/n) - \ln p_2}{\ln\left(\dfrac{q_{\max} - x_{(n-m_n)}}{\hat{\beta}} + 1\right)} \quad \text{et} \quad \alpha_2 = \frac{\ln(m_n/n) - \ln p_1}{\ln\left(\dfrac{q_{\max} - x_{(n-m_n)}}{\hat{\beta}} + 1\right)}.$$

La loi a priori de β sachant α est une loi $\mathcal{G}amma(\delta\alpha+1, \delta\eta)$. Comme dans le cas bayésien empirique, pour déterminer les hyperparamètres, on suppose que la moyenne (ou éventuellement le mode) de cette loi a priori est égale à $\hat{\beta}$, l'estimateur de Hill de β. On obtient donc $\hat{\beta} = (\delta\alpha + 1)/\delta\eta$ (dans le cas de la moyenne a priori, ou $\hat{\beta} = \alpha/\eta$ dans le cas du mode a priori). Le problème est que la valeur de α est inconnue. On considère alors que α est fixé égal à une “valeur moyenne” déterminée à partir de l'avis d'expert :

$$\alpha_{\text{expert}} = \frac{\alpha_1 + \alpha_2}{2},$$

le centre de l'intervalle $[\alpha_1, \alpha_2]$ déduit de l'avis d'expert pour α. On obtient alors $\hat{\beta} = (\delta\alpha_{\text{expert}} + 2)/(2\delta\eta)$ (dans le cas de la moyenne a priori, ou $\hat{\beta} = \alpha_{\text{expert}}/2\eta$ dans le cas du mode a priori). On en déduit l'expression suivante pour η en fonction de δ :

$$\eta(\delta) = \frac{\delta\alpha_{\text{expert}} + 2}{2\delta\hat{\beta}} \tag{3.16}$$

(dans le cas de la moyenne a priori, ou $\eta(\delta) = \alpha_{\text{expert}}/2\hat{\beta}$ dans le cas du mode a priori).

La loi a priori de α est une loi gamconII$(\delta, \eta/\mu)$. Pour déterminer les hyperparamètres, on suppose que le mode de cette loi gamconII est situé au milieu de l'intervalle $[\alpha_1, \alpha_2]$ déduit de l'avis d'expert pour α, c'est-à-dire que $y^* = (\alpha_1 + \alpha_2)/2 = \alpha_{\text{expert}}$. Puique y^* doit vérifier l'équation (3.5) page 105, on en déduit l'expression suivante de μ en fonction de δ :

$$\mu(\delta) = \frac{\delta\alpha_{\text{expert}} + 2}{2\delta\hat{\beta}} \exp\left(\ln\delta - \psi\left(\alpha_{\text{expert}}\delta + 1\right) + \psi\left(\alpha_{\text{expert}}\right)\right) , \tag{3.17}$$

où ψ est la fonction digamma (voir l'annexe D.2.1 page 180).

Il reste maintenant à déterminer δ. Il est bien sûr possible de fixer $\delta = 1$ de même que dans le cas bayésien empirique, mais on peut à présent exploiter la traduction pour α de l'avis d'expert pour déterminer δ. Pour cela, on utilise l'approximation normale de la loi gamconII, et on suppose que l'intervalle $[\alpha_1, \alpha_2]$ déduit de l'avis d'expert pour α a une masse de $1 - \varepsilon$ (ε proche de 0, mesurant notre défiance vis-à-vis de l'expert) pour cette loi normale approximante. En définitive, on se sert du fait que pour une loi normale on a $z_{1-\varepsilon/2}\, \sigma_d^* = (\alpha_2 - \alpha_1)/2$ où $z_{1-\varepsilon/2}$ est le quantile d'ordre $1 - \varepsilon/2$ de la loi $\mathcal{N}(0, 1)$. Puique σ_d^* doit vérifier l'équation (3.6) page 105, on en déduit que δ vérifie l'équation :

$$\delta\psi'\left(\alpha_{\text{expert}}\right) - \delta^2\psi'\left(\alpha_{\text{expert}}\delta + 1\right) - \frac{4z_{1-\varepsilon/2}^2}{(\alpha_1 - \alpha_2)^2} = 0\,, \tag{3.18}$$

où $z_{1-\varepsilon/2}$ est le quantile d'ordre $1 - \varepsilon/2$ de la loi normale centrée et réduite. Pour l'instant, nous résoudrons cette équation par une méthode de dichotomie, sur un intervalle $[\delta_{\min}, \delta_{\max}]$ suffisamment grand.

3.4.2 Cas α fixé : l'avis d'expert agit sur β

Dans ce cas on suppose que $\alpha = \hat{\alpha}$, son estimateur de Hill. De l'équation (3.15) page 136, on déduit l'encadrement suivant pour β :

$$\beta_1 \leq \beta \leq \beta_2 \quad \text{où} \quad \beta_1 = \frac{q_{\max} - x_{(n-m_n)}}{(np_1/m_n)^{-1/\hat{\alpha}} - 1} \quad \text{et} \quad \beta_2 = \frac{q_{\max} - x_{(n-m_n)}}{(np_2/m_n)^{-1/\hat{\alpha}} - 1}\,.$$

La loi a priori de β sachant $\alpha = \hat{\alpha}$ (l'estimateur de Hill de α) est une loi $\mathcal{G}amma(\delta\hat{\alpha} + 1, \delta\eta)$. On suppose que l'intervalle $[\beta_1, \beta_2]$ déduit de l'avis d'expert pour β a une masse de $1 - \varepsilon$ (ε proche de 0, mesurant notre défiance vis-à-vis de l'expert) pour la loi normale approximant la loi gamma. On doit donc résoudre le système suivant de deux équations à deux inconnues :

$$\begin{cases} \dfrac{\beta_1 + \beta_2}{2} = \dfrac{\delta\hat{\alpha} + 1}{\delta\eta} & \text{(expression de la moyenne)} \\[2ex] \dfrac{\beta_2 - \beta_1}{2} = z_{1-\varepsilon/2}\dfrac{\sqrt{\delta\hat{\alpha} + 1}}{\delta\eta} & \text{(expression de l'écart-type)} \end{cases}$$

où $z_{1-\varepsilon/2}$ est le quantile d'ordre $1-\varepsilon/2$ de la loi $\mathcal{N}(0,1)$. On en déduit que

$$\delta = \frac{1}{\hat{\alpha}}\left[z_{1-\varepsilon/2}^2\left(\frac{\beta_1+\beta_2}{\beta_2-\beta_1}\right)^2-1\right]. \tag{3.19}$$

$$\eta = 2\hat{\alpha}z_{1-\varepsilon/2}\,\frac{\beta_1+\beta_2}{(\beta_2-\beta_1)^2}\left[z_{1-\varepsilon/2}^2\left(\frac{\beta_1+\beta_2}{\beta_2-\beta_1}\right)^2-1\right]^{-1}. \tag{3.20}$$

Il reste à présent à déterminer μ. La loi a priori de α est une loi gamconII$(\delta,\eta/\mu)$. Pour déterminer les hyperparamètres, on suppose que le mode de cette loi a priori est égal à $\hat{\alpha}$, l'estimateur de Hill de α, c'est-à-dire que $y^*=\hat{\alpha}$. Puique y^* doit vérifier l'équation (3.5), on en déduit l'expression suivante pour μ :

$$\mu = \eta\exp\left[\ln\delta+\psi(\hat{\alpha})-\psi\left(z_{1-\varepsilon/2}^2\left(\frac{\beta_1+\beta_2}{\beta_2-\beta_1}\right)^2\right)\right], \tag{3.21}$$

où ψ est la fonction digamma et $z_{1-\varepsilon/2}$ est le quantile d'ordre $1-\varepsilon/2$ de la loi normale centrée et réduite.

3.4.3 Exemple d'application numérique de cette méthode bayésienne avec avis d'expert (partiel) pour des excès

Nous appliquons notre méthode bayésienne pour les mêmes données simulées de taille $n=500$ que pour les essais de la procédure bayésienne empirique sur des échantillons d'excès (voir le paragraphe 3.3.3 page 128). De même, nous explorons les estimations du paramètre γ et du quantile d'ordre $1-1/10n=1-1/5000$ de la loi des données de départ, pour différentes valeurs du nombre d'excès m_n (variant de 5 à 495). Les différentes méthodes d'estimation que nous comparons sont à nouveau : la méthode de Hill, la méthode de Hill généralisée, la méthode du Zipf, notre méthode bayésienne pour la moyenne a priori et la moyenne a posteriori, ainsi que notre méthode bayésienne pour la moyenne a priori et la médiane a posteriori. Dans chaque cas, nous simulons 100 échantillons de même loi, ce qui nous permet de calculer une moyenne et un écart-type empiriques pour les différents estimateurs à comparer.

Nous avons décidé que l'avis d'expert sur la queue de distribution que nous allons introduire dans la procédure baýesienne porterait sur le quantile d'ordre $1-1/10n$ que nous allons ensuite estimer. Ceci permet d'explorer l'influence de l'avis d'expert dans le cas particulièrement favorable où il porte directement sur la quantité que l'on souhaite estimer. On choisit donc comme avis d'expert le vrai quantile de la loi des données originelles ($q_{\max}=q_{1-1/10n}$). Il faut ensuite l'interpréter en terme de risques de dépassement de cette quantité $q_{\max}$. On choisit encore une fois de se placer dans un cadre favorable, où l'encadrement $[p_1,p_2]$ du risque p de dépasser $q_{\max}$ est un intervalle réduit autour de la vraie valeur de $p=1/10n=1/5000$, que l'on connait par construction de $q_{\max}$. On prend donc $p_1=1/1000$ et $p_2=1/10000$.

Nous commençons par explorer les cas où notre méthode bayésienne ne produit pas de bons estimateurs dans le cadre du bayésien empirique, notamment pour les estimateurs du quantile d'ordre $1 - 1/10n$. En effet, puisque l'avis d'expert porte précisément sur ce quantile, nous espérons dans ce cas une amélioration importante de son estimation par la méthode bayésienne. Nous présentons donc tout d'abord la méthode bayésienne avec avis d'expert (partiel) sur des données de loi loggamma de paramètre 2 ($\mathcal{LG}(2)$, voir l'annexe A page 161). Nous avons utilisé les mêmes échantillons que dans le cas du bayésien empirique pour comparer les résultats des deux procédures. Dans le cas sans avis d'expert, les estimateurs bayésiens du paramètre γ et du quantile $q_{1-1/5000}$ sont stables en fonction du nombre d'excès m_n, mais fortement biaisés (voir la figure 3.12 page 135).

Nous introduisons tout d'abord l'avis d'expert sur le paramètre β, c'est-à-dire pour α fixé (voir le paragraphe 3.4.2 page 137). Les résultats obtenus (voir la figure 3.13) ne sont pas satisfaisants : le biais des estimateurs bayésiens (du paramètre γ comme du quantile $q_{1-1/5000}$) n'est pas réduit par cette méthode d'introduction d'un avis d'expert, et leur stabilité en fonction de m_n est dégradée (ils croissent vers l'infini pour de grandes valeurs de m_n). Quant à la variance de ces estimateurs bayésiens (pour γ comparer les figures E.15 page 216 et E.17 page 219 ; pour $q_{1-1/5000}$ comparer les figures E.16 page 217 et E.18 page 219), elle ne semble pas réduite par l'introduction de l'avis d'expert sur β.

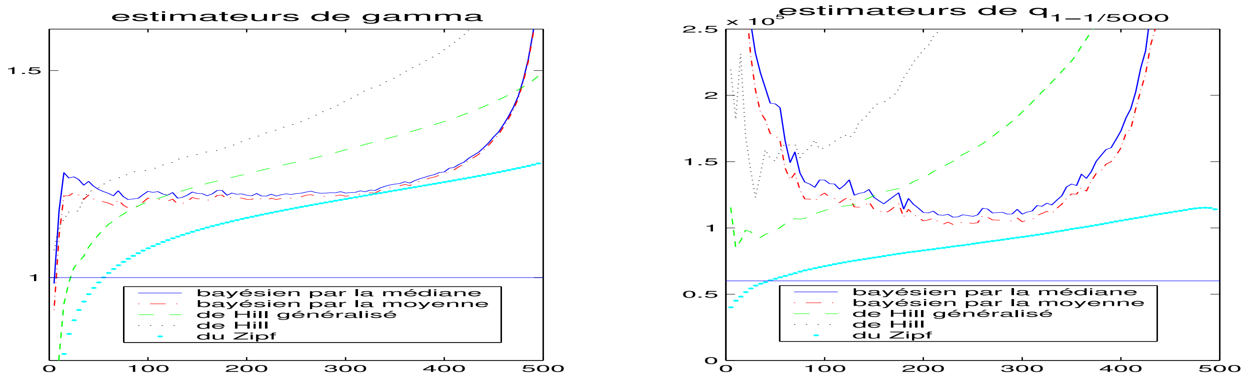

FIG. 3.13 – *Échantillons simulés de loi $\mathcal{LG}(2)$ – Procédure bayésienne avec avis d'expert pour α fixé – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) pour différents m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

Nous introduisons ensuite l'avis d'expert sur le paramètre α, c'est-à-dire pour β fixé (voir le paragraphe 3.4.1 page 136). Les résultats obtenus sont présentés dans la figure 3.14. Dans ce cas, le biais des estimateurs bayésiens du quantile $q_{1-1/5000}$ est visiblement réduit par rapport au cas sans avis d'expert (voir la figure 3.12 page 135). L'estimateur bayésien de $q_{1-1/5000}$ est aussi très stable en fonction du nombre d'excès m_n. Pour l'estimateur bayésien de γ, on constate une légère baisse du biais pour de petites valeurs de m_n. L'avis d'expert portant sur le quantile, nous n'attendions pas forcément d'amélioration pour les estimateurs de γ. De plus, la variance de ces estimateurs bayésiens semble légèrement réduite pour γ (surtout lorsque m_n est petit, voir la figure E.19 page 219), et plus sensiblement réduite pour $q_{1-1/5000}$ (en particulier pour m_n petit, voir la figure E.20 page 220) lorsque l'on introduit un avis d'expert pour β fixé.

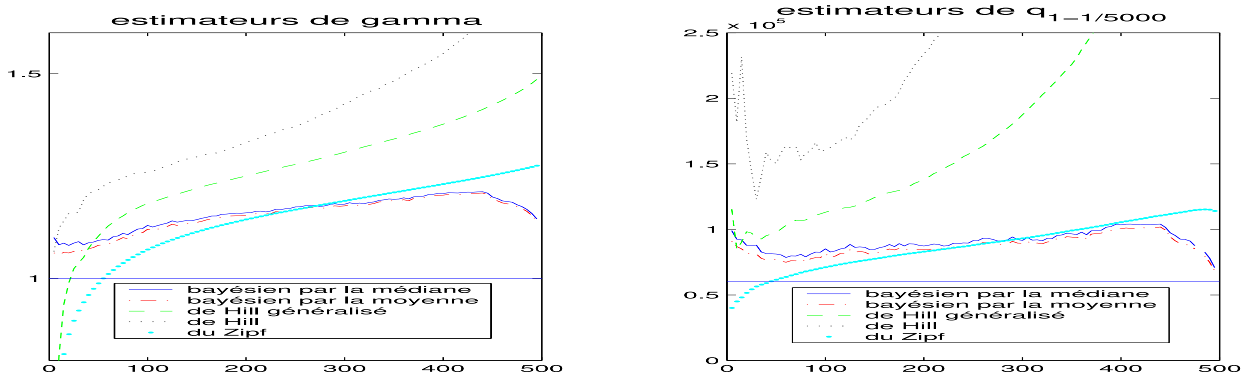

FIG. 3.14 – *Échantillons simulés de loi $\mathcal{LG}(2)$ – Procédure bayésienne avec avis d'expert pour β fixé – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) pour différents m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

Nous appliquons ensuite la méthode bayésienne avec avis d'expert (partiel) sur des données de loi de Student absolue de paramètre 8 ($t_{Abs}(8)$, voir l'annexe A page 161), toujours en utilisant le même échantillon que dans le cas du bayésien empirique pour comparaison. Dans le cas sans avis d'expert, les estimateurs bayésiens du paramètre γ sont stables en fonction du nombre d'excès m_n, maisfortement biaisés, alors que les estimateurs du quantile $q_{1-1/5000}$ sont instables en fonction de m_n et biaisés (voir la figure 3.11 page 134).

Nous introduisons tout d'abord l'avis d'expert sur le paramètre β, c'est-à-dire pour α fixé (voir le paragraphe 3.4.2 page 137). À nouveau, les résultats obtenus (voir la figure 3.15) ne sont pas satisfaisants : le biais des estimateurs bayésiens n'est pas réduit par cette méthode d'introduction d'un avis d'expert, et leur stabilité est dégradée (ils croissent plus vite vers l'infini lorsque le nombre d'excès m_n croît). Quant à la variance de ces estimateurs bayésiens (pour γ comparer les figures E.13 page 214 et E.21 page 220 ; pour $q_{1-1/5000}$ comparer les figures E.14 page 215 et E.22 page 220), elle ne semble que légèrement réduite par l'introduction de l'avis d'expert sur β.

Sur ces premiers essais, traduire l'avis d'expert sur le paramètre β, en fixant le paramètre α n'a pas semblé être la méthode la plus adaptée. Dans la suite des essais, nous avons donc abandonné cette méthode, et nous n'explorons plus que le cas où l'avis d'expert influe sur le paramètre α, c'est-à-dire lorsque β est fixé.

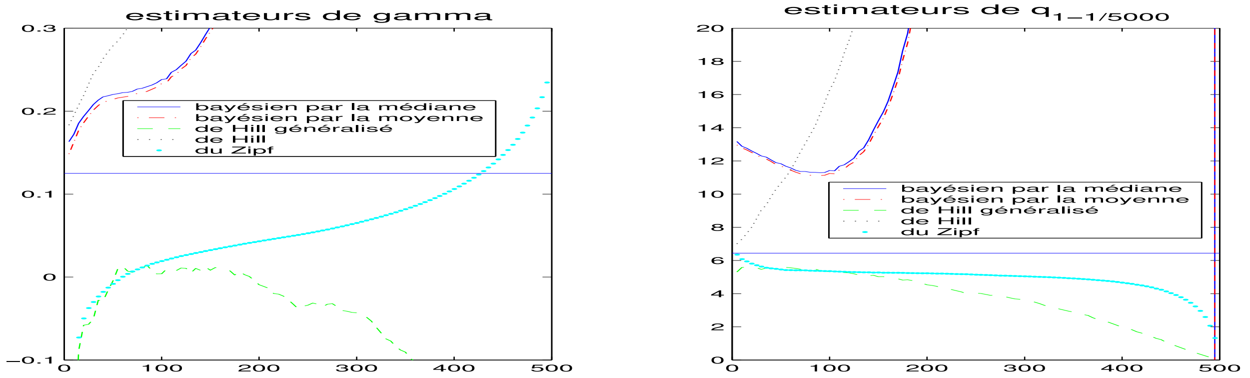

FIG. 3.15 – *Échantillons simulés de loi $t_{Abs}(8)$ – Procédure bayésienne avec avis d'expert pour α fixé – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) pour différents m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

Pour ces échantillons de loi $t_{Abs}(8)$, nous introduisons à présent l'avis d'expert sur le paramètre α, c'est-à-dire pour β fixé (voir le paragraphe 3.4.1 page 136). Les résultats obtenus sont présentés dans la figure 3.16.

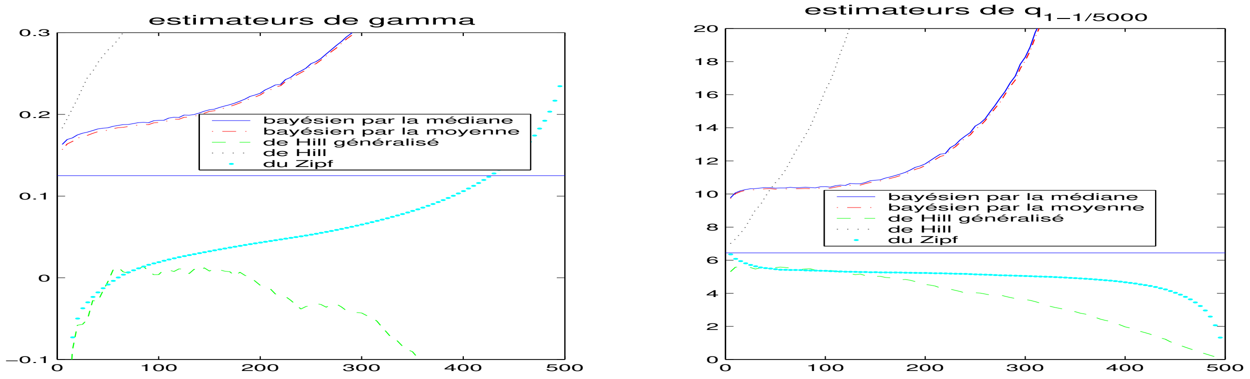

FIG. 3.16 – *Échantillons simulés de loi $t_{Abs}(8)$ – Procédure bayésienne avec avis d'expert pour β fixé – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) pour différents m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

Dans ce cas, pour de petites valeurs du nombre d'excès m_n, le biais des estimateurs bayésiens de γ est visiblement réduit par rapport au cas sans avis d'expert (voir la figure 3.11 page 134). Par contre, la plage de stabilité des estimateurs bayésiens de γ, qui est importante lorsque l'on travaille sans avis d'expert (elle existe pour $m_n < 400$), se réduit fortement lorsque l'on introduit l'avis d'expert sur α (elle n'existe plus que pour $m_n < 200$). Ceci n'est cependant pas un problème car on conseille généralement d'utiliser un nombre d'excès m_n modéré, c'est-à-dire au maximum de l'ordre de $n/2$ (= 250 ici). Cette croissance vers l'infini lorsque m_n devient grand est aussi constatée pour les estimateurs bayésiens du quantile $q_{1-1/5000}$. Cette tendance existe déjà pour toutes les valeurs de m_n lorsque l'on travaille sans avis d'expert, mais en utilisant l'avis d'expert sur α, elle se cantonne aux grandes valeurs de m_n. On obtient donc une petite plage de stabilité des estimateurs bayésiens de $q_{1-1/5000}$ en fonction de m_n (pour $m_n < 150$), dans laquelle cependant ces estimateurs sont encore biaisés. Dans ce cas, la variance des estimateurs bayésiens n'est pas visiblement réduite par l'introduction de l'avis d'expert sur le paramètre α (voir la figure E.23 page 221 pour γ, et la figure E.24 page 221 pour le quantile $q_{1-1/5000}$).

Dans le cas d'échantillons de loi de Student absolue de paramètre 4 ($t_{Abs}(4)$, voir l'annexe A page 161), la procédure bayésienne empirique produit des estimateurs du paramètre γ stables en fonction de m_n et sans biais, mais des estimateurs du quantile $q_{1-1/5000}$ plutôt instables en fonction de m_n et biaisés (voir la figure 3.10 page 134). Nous appliquons maintenant la procédure avec avis d'expert, pour β fixé, sur les mêmes échantillons. Les résultats sont présentés dans la figure 3.17.

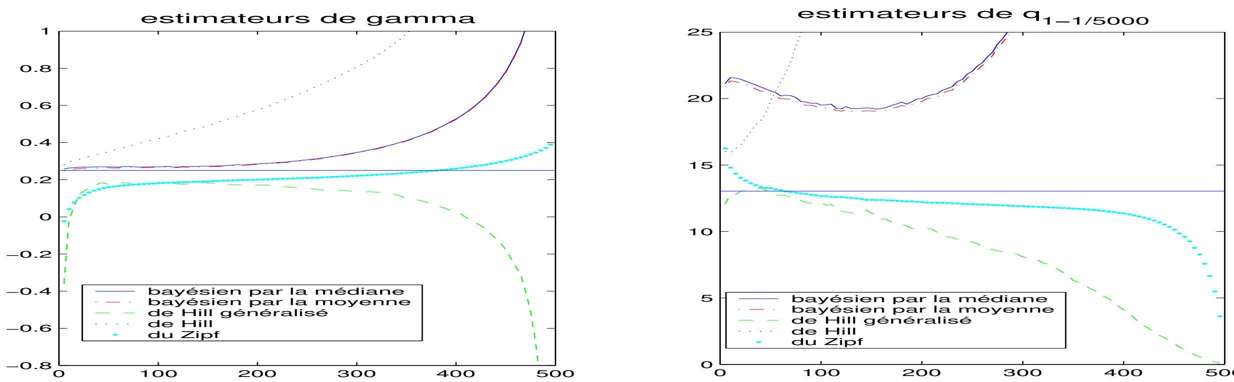

FIG. 3.17 – *Échantillons simulés de loi $t_{Abs}(4)$ – Procédure bayésienne avec avis d'expert pour β fixé – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) pour différents m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

Concernant l'estimation bayésienne du paramètre γ, l'introduction d'un avis d'expert pour β fixé réduit très légèrement le biais pour les petites valeurs de m_n. Cependant la plage de stabilité est alors réduite : de la région $m_n < 400$ sans avis d'expert, on passe à la région $m_n < 250$ avec un avis d'expert pour β fixé. Concernant l'estimation bayésienne du quantile, l'introduction de l'avis d'expert pour β fixé dégrade le biais. On obtient certes une petite plage de stabilité des estimateurs en fonction de m_n (entre 100 et 200), mais elle est plus réduite que dans le cas sans avis d'expert (pour $m_n < 200$). De plus, le biais minimal obtenu avec l'avis d'expert est de l'ordre de 4 ou 5, alors que sans avis d'expert il peut descendre jusqu'à 2. On peut aussi remarquer que l'utilisation de la procédure avec avis d'expert réduit légèrement la variance des estimateurs bayésiens du paramètre γ comme du quantile $q_{1-1/5000}$ (comparer les figures E.11 et E.12 pages 212 et 213 pour le cas bayésien empirique, aux figures E.25 et E.26 pages 221 et 222 pour le cas avec avis d'expert). Dans ce cas, l'introduction de l'avis d'expert n'améliore pas significativement les résultats.

Nous explorons maintenant le cas d'échantillons de loi de Burr de paramètres $(1, 0.5, 2)$ ($\mathcal{B}urr(1, 0.5, 2)$, voir l'annexe A page 161). Dans ce cas, la procédure bayésienne empirique produit des estimateurs du paramètre γ assez stables en fonction de m_n (notamment pour $m_n < n/2 = 250$) et peu biaisés (voir la figure 3.6 page 130). Les estimateurs bayésiens du quantile $q_{1-1/5000}$ sont plutôt instables lorsque $m_n < n/10 = 100$, puis stables lorsque m_n est compris entre 100 et 150. Cependant ces estimateurs de $q_{1-1/5000}$ sont biaisés. Les résultats de la procédure bayésienne avec avis d'expert (pour β fixé) sur les mêmes échantillons sont présentés dans la figure 3.18.

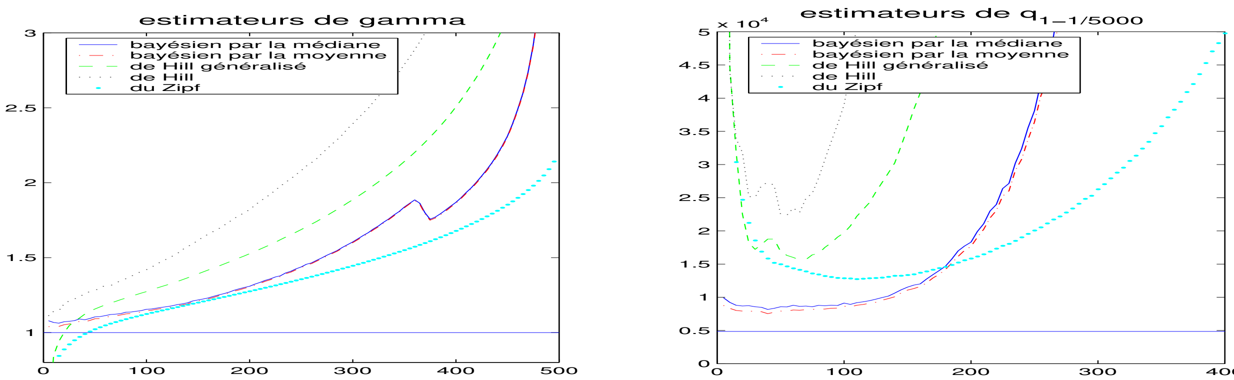

FIG. 3.18 – *Échantillons simulés de loi $\mathcal{B}urr(1, 0.5, 2)$ – Procédure bayésienne avec avis d'expert pour β fixé – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) pour différents m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

Concernant l'estimation bayésienne du paramètre γ, l'introduction d'un avis d'expert pour β fixé réduit légèrement le biais pour les petites valeurs de m_n. Cependant la plage de stabilité qui existe dans la région $m_n < 150$ lorsque l'on travaille sans avis d'expert, disparait lorsque l'on travaille avec un avis d'expert pour β fixé. Concernant l'estimation bayésienne du quantile, par contre, l'introduction de l'avis d'expert pour β fixé réduit fortement le biais des estimateurs bayésiens pour de petites valeurs de m_n. On obtient alors une petite plage de stabilité des estimateurs en fonction de m_n (pour $m_n < 150$), pour laquelle le biais de ces estimateurs est très faible. De plus, la variance des estimateurs bayésiens de γ est légèrement réduite pour les petites valeurs de m_n lorsque l'on introduit un avis d'expert pour β fixé. Quant à la variance des estimateurs bayésiens du quantile $q_{1-1/5000}$, elle est généralement réduite par l'avis d'expert (pour β fixé).

Nous explorons ensuite le cas d'échantillons de loi de $\mathcal{B}urr(1, 1, 1)$ (voir l'annexe A page 161). Dans ce cas, la procédure bayésienne empirique produit des estimateurs du paramètre γ assez stables en fonction de m_n (notamment pour $m_n < n/2 = 250$) et sans biais (voir la figure 3.6 page 130). Les estimateurs bayésiens du quantile $q_{1-1/5000}$ sont plutôt instables lorsque $m_n < n/10 = 100$, puis stables lorsque m_n est compris entre 150 et 300. Cependant ces estimateurs du quantile $q_{1-1/5000}$ sont biaisés. Les résultats de la procédure bayésienne avec avis d'expert (pour β fixé) sur les mêmes échantillons sont présentés dans la figure 3.19.

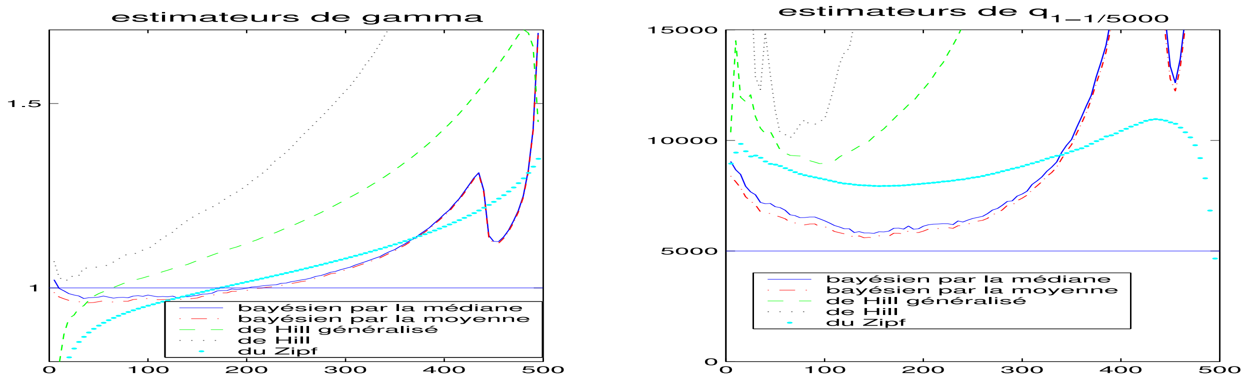

FIG. 3.19 – *Échantillons simulés de loi* $\mathcal{B}urr(1, 1, 1)$ *– Procédure bayésienne avec avis d'expert pour* β *fixé – Valeurs moyennes des différents estimateurs de* γ *(à gauche) et de* $q_{1-1/5000}$ *(à droite) pour différents* m_n *– continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

On constate tout d'abord une perte au niveau de la stabilité des estimateurs qui (surtout pour les estimateurs du quantile) sont stables pour des valeurs de m_n inférieures à 300, mais qui au-delà partent vers l'infini. Ceci n'est cependant pas un problème car on conseille généralement d'utiliser un nombre d'excès m_n modéré, c'est-à-dire au maximum de l'ordre de $n/2$ (= 250 ici). Ensuite, on peut remarquer que les estimateurs bayésiens de γ avec avis d'expert ont, dans la plage de stabilité, des valeurs moyennes légèrement plus faibles que les estimateurs obtenus sans avis d'expert, le biais restant réduit. Enfin, les estimateurs bayésiens du quantile $q_{1-1/5000}$ sont très peu biaisés, dans la plage de stabilité. L'avis d'expert sur la queue de distribution a donc effectivement permis d'améliorer l'estimation du quantile, sans dégrader l'estimation de γ dans de cas. On peut aussi remarquer que l'utilisation de la procédure avec avis d'expert réduit la variance des estimateurs bayésiens du paramètre γ comme du quantile $q_{1-1/5000}$ (comparer les figures E.3 et E.4 pages 204 et 205 pour le cas bayésien empirique, aux figures E.29 et E.30 page 223 pour le cas avec avis d'expert). La réduction de la variance est plus sensible pour les plus petites valeurs du nombre d'excès m_n.

Dans le cas d'échantillons de loi de Fréchet de paramètre 1 ($\mathcal{F}rechet(1)$, voir l'annexe A page 161), la procédure bayésienne empirique produit des estimateurs du paramètre γ stables (tout au moins pour m_n entre 150 et 300) et sans biais (sur la même plage de valeurs, voir la figure 3.5 page 129). Les estimateurs du quantile $q_{1-1/5000}$ sont aussi très stables (pour m_n entre 100 et 400) mais légèrement biaisés. La procédure avec avis d'expert (pour β fixé) appliquée sur les mêmes échantillons produit des estimateurs encore plus stables (lorsque m_n est compris entre 100 et 450 pour γ ou entre 50 et 450 pour $q_{1-1/5000}$), voir la figure 3.20. Enfin, on constate que la variance des estimateurs bayésiens avec avis d'expert est légèrement réduite par rapport à celle des estimateurs de la procédure bayésienne empirique (comparer les figures E.1 et E.2 pages 202 et 203 pour le cas bayésien empirique, aux figures E.31 et E.32 pages 223 et 224 pour le cas avec avis d'expert). La réduction de la variance est plus sensible pour les petites valeurs du nombre d'excès m_n.

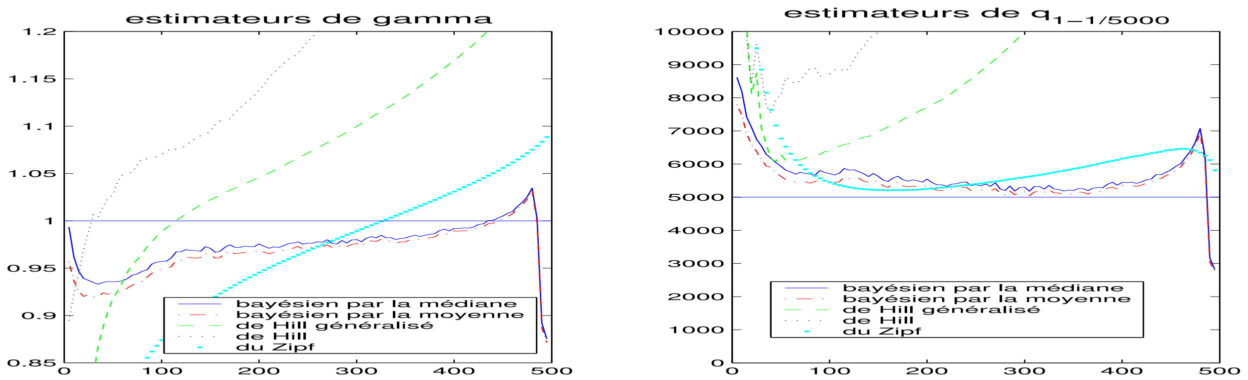

FIG. 3.20 – *Échantillons simulés de loi $\mathcal{F}rechet(1)$ – Procédure bayésienne avec avis d'expert pour β fixé – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) en fonction de m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

Dans le cas d'échantillons de loi de Student absolue de paramètre 1 ($t_{Abs}(1)$, voir l'annexe A page 161), la procédure bayésienne empirique produit des estimateurs du paramètre γ relativement stables (surtout lorsque $m_n < n/2 = 250$) et peu biaisés (sur la même plage de valeurs, voir la figure 3.8 page 132). Les estimateurs du quantile $q_{1-1/5000}$ sont moins stables, mais une légère plage de stabilité semble apparaître pour m_n compris entre 100 et 200, et le biais est faible sur cet intervalle réduit. La procédure avec avis d'expert (pour β fixé) appliquée sur les mêmes échantillons produit des estimateurs de γ très stables (pour $m_n < 400$, voir la figure 3.21. Pour l'estimation du quantile $q_{1-1/5000}$, l'introduction d'un avis d'expert sur la queue de distribution produit des estimateurs bayésiens plus stables pour les petites valeurs du nombre d'excès m_n. Pour $m_n < 100$, les fortes variations des estimateurs bayésiens calculés sans avis d'expert sont gommées par l 'introduction de l'avis d'expert. La plage de stabilité et de biais réduit apparaît maintenant pour m_n compris entre 50 et 200. Mais ensuite continue d'exister un biais négatif, avant que l'estimateur ne parte vers l'infini. De plus, on constate encore une fois la réduction, plus sensible pour les petites valeurs du nombre d'excès m_n, de la variance des estimateurs bayésiens lorsque l'on introduit un avis d'expert (comparer les figures E.7 et E.8 pages 208 et 209 pour le cas bayésien empirique, aux figures E.33 et E.34 page 224 pour le cas avec avis d'expert).

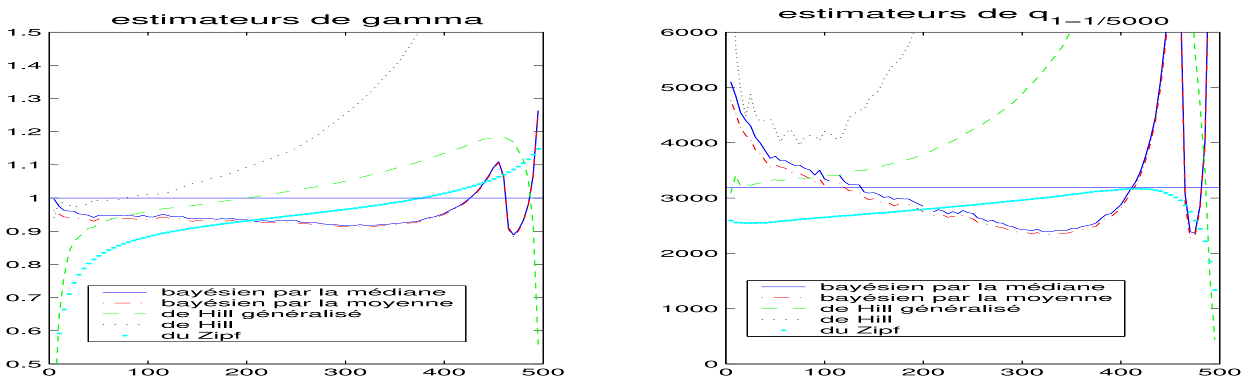

FIG. 3.21 – *Échantillons simulés de loi $t_{Abs}(1)$ – Procédure bayésienne avec avis d'expert pour β fixé – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) pour différents m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

Nous explorons enfin le cas d'échantillons de loi $t_{Abs}(2)$ (voir l'annexe A page 161). Dans ce cas, la procédure bayésienne empirique produit des estimateurs du paramètre γ instables en fonction de m_n et biaisés (voir la figure 3.9 page 133). Les estimateurs bayésiens du quantile $q_{1-1/5000}$ sont plus stables et peu biaisés, notamment lorsque m_n est grand (compris entre 150 et 400). Les résultats de la procédure bayésienne avec avis d'expert (pour β fixé) sur les mêmes échantillons sont présentés dans la figure 3.22.

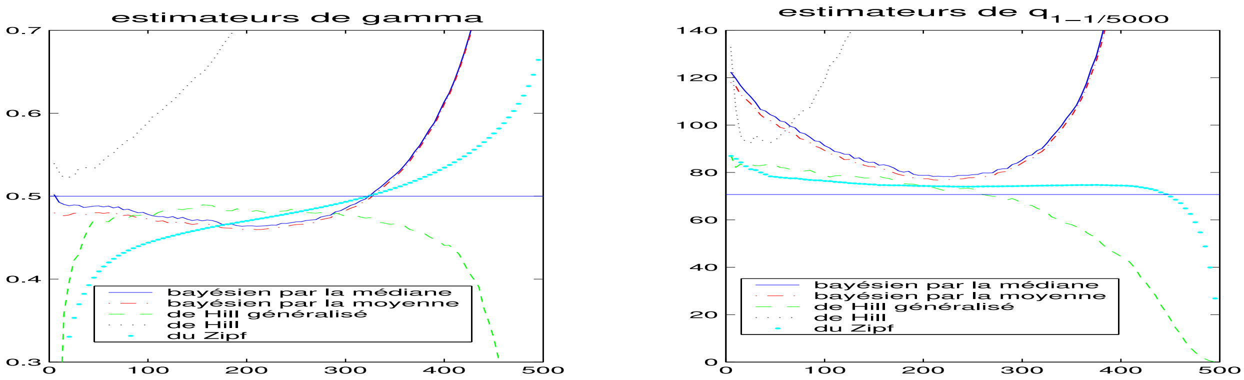

FIG. 3.22 – *Échantillons simulés de loi $t_{Abs}(2)$ – Procédure bayésienne avec avis d'expert pour β fixé – Valeurs moyennes des différents estimateurs de γ (à gauche) et de $q_{1-1/5000}$ (à droite) pour différents m_n – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

On constate tout d'abord une plus grande stabilité des estimateurs bayésiens de γ, principalement pour des valeurs de m_n inférieures à 300. Mais ces estimateurs restent biaisés. Pour les estimateurs du quantile, la plage de stabilité est réduite par l'introduction de l'avis d'expert. En effet, le biais n'étant pas très important sans avis d'expert et pour des valeurs réduites de m_n, il n'est pas amélioré par l'utilisation d'un avis d'expert. Par contre, comme on l'a déjà constaté avec les essais des lois précédentes, l'avis d'expert induit une croissance rapide des estimateurs vers l'infini lorsque m_n devient grand (ici pour $m_n > 300$). Or dans le cas de la loi $t_{Abs}(2)$, les estimateurs obtenus sans avis d'expert donnent de bons résultats pour de grandes valeurs de m_n (entre 200 et 400). Dans ce cas, l'estimation du quantile $q_{1-1/5000}$ étant déjà satisfaisante sans avis d'expert, en particulier pour de grandes valeurs de m_n, l'introduction d'un avis d'expert ne permet pas d'amélioration significative des estimations. On constate même une réduction de la plage de stabilité et de faible biais des estimateurs en fonction de m_n. On peut enfin remarquer que l'utilisation de la procédure avec avis d'expert réduit, ici encore, la variance des estimateurs bayésiens du paramètre γ comme du quantile $q_{1-1/5000}$(comparer les figures E.9 et E.10 pages 210 et 211 pour le cas bayésien empirique, aux figures E.35 et E.36 page 225 pour le cas avec avis d'expert).

3.5 Données réelles

Nous nous proposons à présent d'appliquer notre méthode bayésienne à un jeu de données réelles sur les hauteurs de crue de la rivière Nidd (Yorkshire, Angleterre). Ce jeu de données a été souvent utilisé comme exemple pour les études sur des valeurs extrêmes, par exemple par Hosking et Wallis [37]. Ces données consistent en $n = 154$ excès au-delà du niveau 65 m^3/s de la rivière Nidd durant la période 1934–1969, c'est-à-dire pendant 35 ans. Les Hydrologues ont besoin d'estimer des quantiles extrêmes dans le but de prédire les niveaux de crues records sur de longues périodes (par exemple, tous les 100 ans pour les crues centenales, ou tous les 250 ou même 500 ans). Nous nous intéressons ici par exemple au niveau record pour 250 ans. D'après Davison et Smith [20], cela revient à estimer le quantile d'ordre $1-p$ pour $p = 9.10^{-4}$.

Nous cherchons donc à estimer ce quantile extrême par la méthode POT. Le seuil de 65 m^3/s n'étant pas forcément adapté pour cette méthode, nous ne considérons pas l'échantillon des données comme un échantillon d'excès. On lui applique donc la méthode POT pour $m_n = 1$ à 153 excès. Cependant, pour estimer le niveau de crue record pour 250 ans, il faut garder à l'esprit le fait que les données initiales sont en fait des excès, et donc ajouter 65 (la valeur du seuil initial) au quantile obtenu par la méthode POT appliquée à notre échantillon (qui est déjà un échantillon d'excès). Nous comparons à nouveau les méthodes de Hill, de Hill généralisée, du Zipf, bayésienne pour la moyenne a priori et la moyenne a posteriori, ainsi que bayésienne pour la moyenne a priori et la médiane a posteriori. Nous appliquons pour l'instant la procédure bayésienne sans avis d'expert. La figure 3.23 présente les différentes estimations du paramètre γ de la loi GPD, que suivent approximativement les excès, en fonction des valeurs du nombre d'excès m_n.

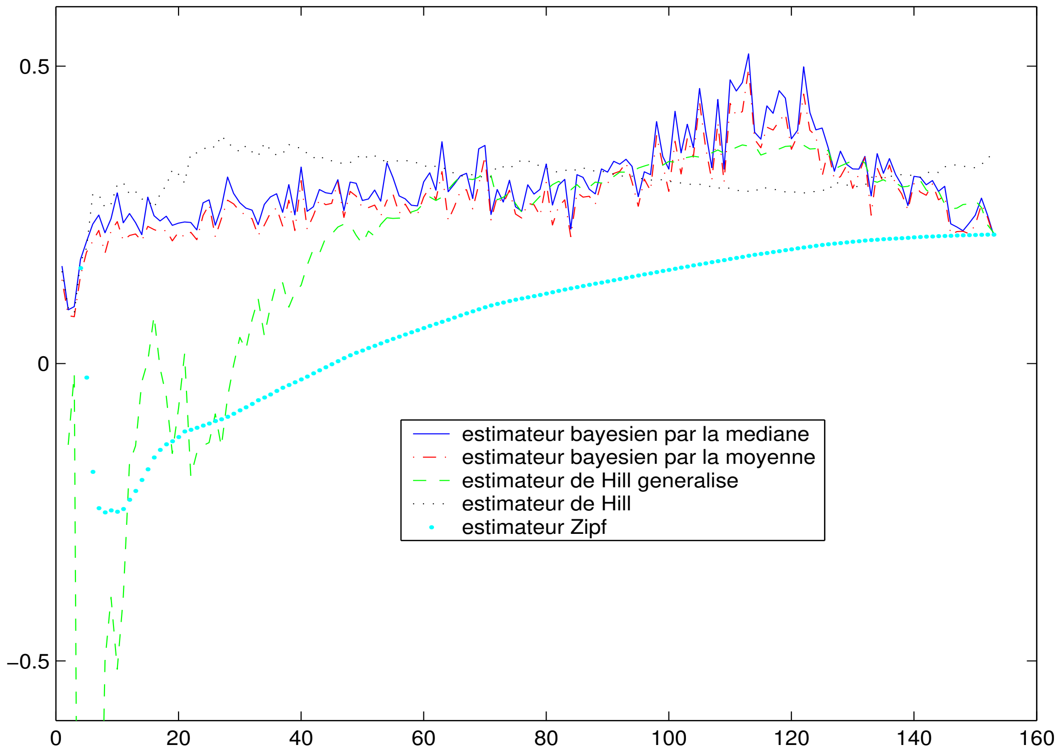

FIG. 3.23 – *Données réelles sur la rivière Nidd – estimation du paramètre γ – continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

On constate qu'il existe une importante plage horizontale de relative stabilité en fonction du nombre d'excès m_n pour les estimateurs bayésiens de γ, et, ce qui est plus étonnant, pour l'estimateur de Hill. L'estimateur de Hill généralisé par contre est très instable pour les petites valeurs de m_n, puis se stabilise un peu, autour des mêmes valeurs que les estimateurs bayésiens, lorsque m_n croît. L'estimateur Zipf croît en fonction de m_n, prend des valeurs généralement plus faibles que celles des autres estimateurs, et ne présente pas de stabilité horizontale. On peut enfin remarquer que les estimateurs strictement positifs que sont l'estimateur de Hill et nos estimateurs bayésiens sont proches de 0 (de l'ordre de 0.2), alors que l'estimateur de Hill généralisé et l'estimateur Zipf produisent des valeurs négatives (au minimum de l'ordre de -2.5 pour Hill généralisé et de -0.2 pour Zipf) pour les petites valeurs de m_n, puis positives lorsque m_n croît. On peut donc se demander si la vraie valeur de γ ne pourrait pas être négative, et donc les méthodes produisant des estimateurs positifs inadaptées.

Étudions maintenant les différentes estimations du quantile d'ordre $1-9.10^{-4}$. La figure 3.24 présente les différentes estimations du quantile d'ordre $1-p$, avec $p = 9.10^{-4}$, en fonction des valeurs du nombre d'excès m_n.

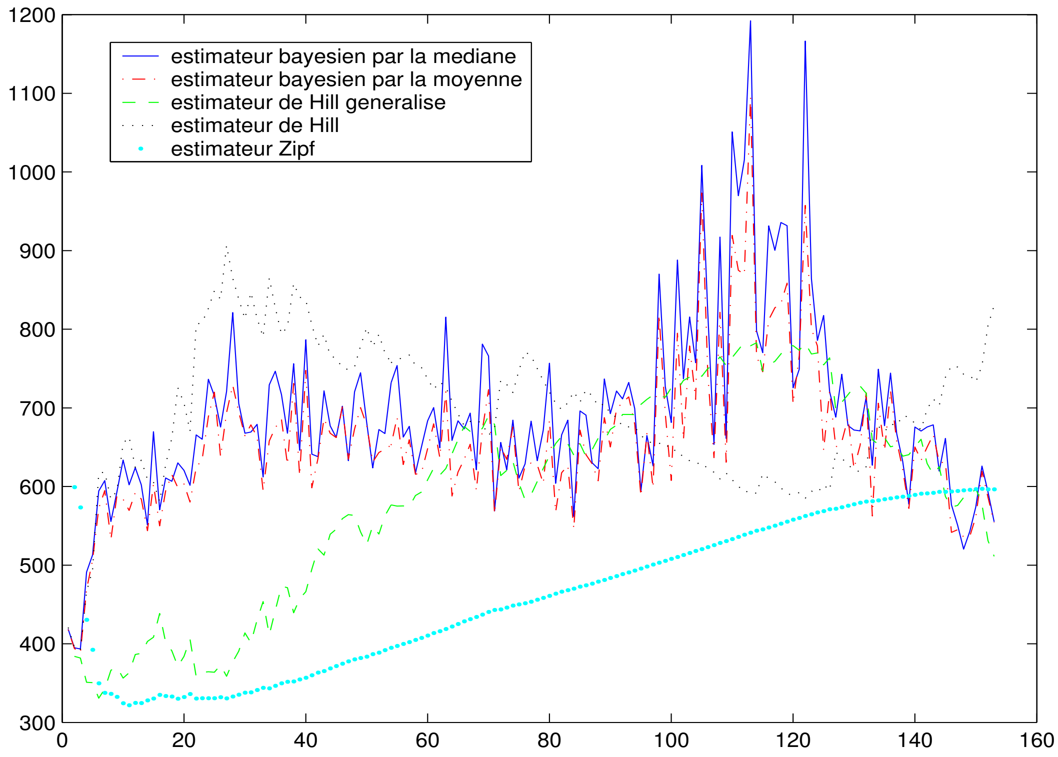

FIG. 3.24 – *Données réelles sur la rivière Nidd – estimation du quantile d'ordre* $1-p$, *avec* $p = 9.10^{-4}$ *– continu : estimateur bayésien par la médiane a posteriori ; discontinu : estimateur bayésien par la moyenne a posteriori ; tirets : estimateur de Hill généralisé ; pointillés : estimateur de Hill ; points : estimateur Zipf.*

On constate un plateau horizontal des estimateurs bayésiens du quantile en fonction du nombre d'excès m_n, bien qu'ils fluctuent fortement autour d'une valeur moyenne horizontale stable de l'ordre de 650. Pour l'estimation du quantile aussi, l'estimateur de Hill généralisé est plus instable que les autres surtout pour les petites valeurs de m_n. Lorsque m_n croît, il se stabilise un peu, autour des mêmes valeurs que les estimateurs bayésiens. Remarquons à présent que la valeur maximale du jeu de données est l'excès 305.75 au-delà de 65, qui correspond donc à une valeur de 370.75 comme estimation du quantile d'ordre $1-p$ pour $p = 1/n = 1/154 = 0.0065$. Or, au minimum, l'estimateur de Hill généralisé propose des valeurs de l'ordre de 350 comme estimation du quantile d'ordre $1-p$ pour $p = 9.10^{-4}$, ce qui semble contradictoire. La même remarque s'applique à l'estimateur Zipf qui là aussi croît en fonction de m_n, prend des valeurs généralement plus faibles que celles des autres estimateurs, et ne présente pas de stabilité horizontale. On aura donc beaucoup plus confiance dans les estimations de ce quantile de l'ordre de 650 obtenues par nos estimateurs bayésiens pour m_n compris entre 20 et 100, et par l'estimateur de Hill généralisé pour m_n compris entre 60 et 100, un nombre d'excès de l'ordre de $m_n = 80$ ayant d'ailleurs déjà été conseillé dans la littérature. Il semble donc, puisque les estimations correspondantes du quantile paraissent erronnées, que les estimations négatives (ou trop proches de 0) de γ données par les estimateurs Zipf (pour la plupart des valeurs de m_n) et de Hill généralisé (pour les petites valeurs de m_n) soient inadaptées. On préfèrera utiliser les estimateurs bayésiens qui semblent donner des estimations adaptées pour une plage étendue de valeurs de m_n.

Les graphes suivants nous permettent de visualiser les lois a posteriori obtenues par la procédure bayésienne, pour $m_n = 82$ excès. Ce nombre d'excès nous a semblé représentatif puisque la plupart des estimateurs du quantile calculés pour ce nombre d'excès sont proches, de l'ordre de 650. De plus, un nombre d'excès de l'ordre de $m_n = 80$ est conseillé dans la littérature. Tout d'abord, on trace donc l'histogramme des 500 dernières valeurs de α et β obtenues au cours de l'algorithme de Gibbs (voir la figure 3.25). Cela nous donne une représentation de la loi a posteriori sur α sachant l'échantillon des excès, ainsi que de la loi a posteriori de β sachant α et l'échantillon des excès.

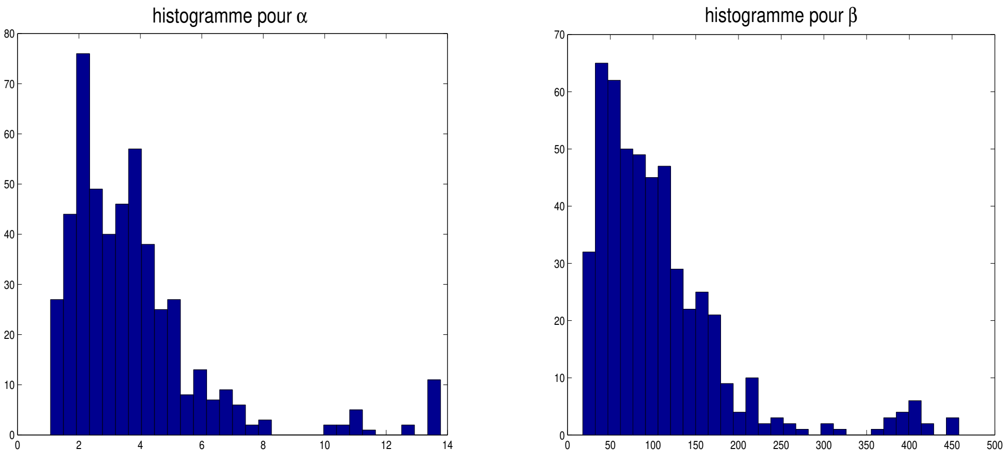

FIG. 3.25 – *Données réelles sur la rivière Nidd – Histogramme des 500 dernières valeurs obtenues par l'algorithme de Gibbs pour α (à gauche) et β (à droite), pour $m_n = 82$ excès*

On constate que ces distributions sont dissymétriques et semblent posséder une queue de distribution lourde. On peut même soupçonner des valeurs aberrantes ou une distribution de mélange multimodale. Ceci peut être dû au fait que le nombre d'itérations de l'algorithme de Gibbs n'a pas été assez important. On a ensuite calculé à partir de ces échantillons de α et β de taille 500 des échantillons de valeurs de γ et σ tels que pour tout $i \in \{1, \ldots, 500\}$, $\gamma_i = 1/\alpha_i$ et $\sigma_i = \beta_i/\alpha_i$. On dispose ainsi de réalisations issues (approximativement) des lois a posteriori que notre schéma bayésien induit sur les paramètres γ et σ. Nous traçons alors l'histogramme de ces valeurs de γ et σ obtenues par transformation des 500 dernières valeurs de α et β de l'algorithme de Gibbs (voir la figure 3.26), ce qui correspond à une représentation des lois a posteriori induites sur γ et σ. Ces lois semblent aussi dissymétriques, mais, dans ce cas, il n'apparaît pas de mode local en queue de distribution.

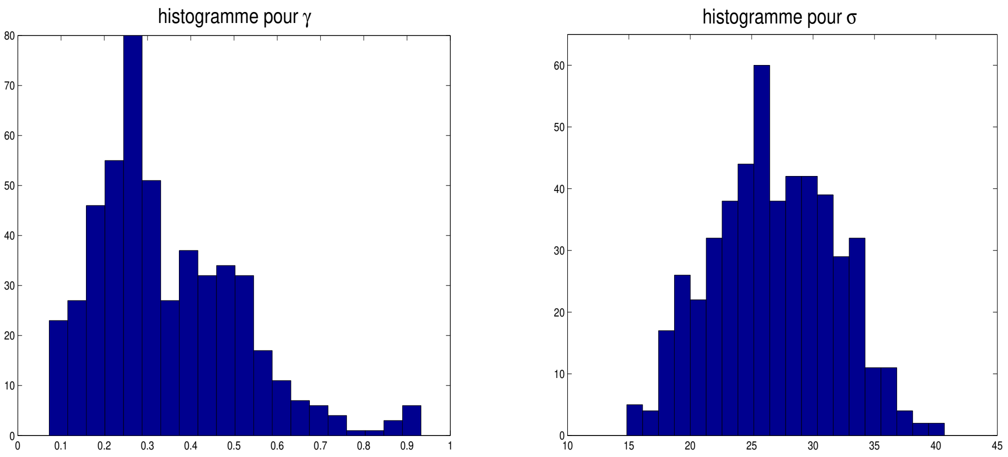

FIG. 3.26 – *Données réelles sur la rivière Nidd – Histogramme des valeurs de γ (à gauche) et σ (à droite) obtenues par transformation des 500 dernières valeurs de α et β de l'algorithme de Gibbs, pour $m_n = 82$ excès*

Enfin, à partir des échantillons de α et β donnés par l'algorithme de Gibbs, on peut calculer un échantillon d'estimations du quantile $q_{1-9.10^{-4}}$ telles que pour tout $i \in \{1, \ldots ,500\}$, $q_i = \hat{u}_n + \beta_i((np/m_n)^{-1/\alpha_i} - 1)$ où $p = 9.10^{-4}$. On dispose ainsi de réalisations issues (approximativement) de la loi a posteriori que notre schéma bayésien induit sur le quantile estimé, loi que l'on représente par l'histogramme de ces valeurs de $q_{1-9.10^{-4}}$ obtenues par transformation des 500 dernières valeurs de α et β de l'algorithme de Gibbs (voir la figure 3.27).

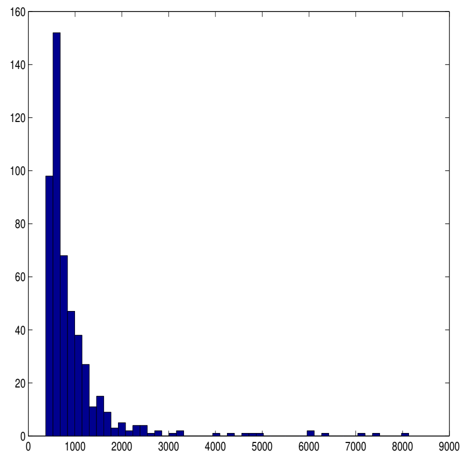

FIG. 3.27 – *Données réelles sur la rivière Nidd – Histogramme des valeurs du quantile $q_{1-9.10^{-4}}$ obtenues par transformation des 500 dernières valeurs de α et β de l'algorithme de Gibbs, pour $m_n = 82$ excès*

3.6 Conclusion/perspectives

La méthode bayésienne proposée ici pour estimer les paramètres d'une loi GPD est encore incomplète puisque lorsque l'on introduit un avis d'expert il n'influence qu'un seul des paramètres. Cependant, déjà dans le cas bayésien empirique (c'est-à-dire sans avis d'expert), les resultats de nos simulations se sont révélés très satisfaisants, notamment lorsque l'information est apportée par les estimateurs des moments pondérés via la moyenne a priori, et que les estimateurs bayésiens sont calculés comme les médianes des lois a posteriori (marginale pour α et conditionnelle sachant α pour β). De plus, l'introduction (même partielle) de l'avis d'expert nous permet dans la plupart des cas une amélioration des estimations, tant au niveau du biais que de la variance.

D'une part, on constate une grande stabilité des estimateurs bayésiens en fonction du nombre d'excès m_n, en particulier pour les estimateurs de γ et ceux des quantiles extrêmes. Dans le cas sans avis d'expert, la plage de stabilité des estimateurs va en général de $m_n = n/5$ à $m_n = n/2$, mais peut être plus large (presque de 0 à n !). Nos estimateurs sont donc peu sensibles au nombre d'excès m_n que l'on choisit. D'autre part, outre un biais généralement réduit, les estimateurs (des paramètres ou des quantiles) issus de l'estimation bayésienne des paramètres de la loi GPD ont une variance faible et surtout bien plus réduite que celle des estimateurs classiques.

Lorsque l'on introduit un avis d'expert pour β fixé (donc partiel) sur les quantiles extrêmes comme nous l'avons proposé au paragraphe 3.4 (page 135), l'information que l'on apporte sur les queues de distribution permet de réduire le biais d'estimation des quantiles extrêmes. Mais ceci se fait au détriment de la stabilité des estimateurs en fonction du nombre d'excès m_n. Les estimateurs issus de cette procédure avec avis d'expert ont des plages de stabilité en fonction de m_n plus restreintes. Cependant, la perte de stabilité s'observe le plus souvent pour m_n grand ($> n/2$), c'est-à-dire dans une région de valeurs de m_n que l'on déconseille généralement d'utiliser. On peut d'ailleurs remarquer qu'un gain de stabilité peut être observé pour les petites valeurs de m_n, celles que l'on souhaite de préférence utiliser.

Dans le cadre d'échantillons d'excès, il faudra plus particulièrement explorer l'utilisation de la loi prédictive. Cette loi étant un mélange de lois GPD, elle peut en effet être plus souple pour trouver un compromis entre les informations apportées par l'échantillon et l'expert, ou s'approcher de la vraie loi des excès (qui est proche d'une loi GPD, mais en général n'en est pas une). Cependant, son utilisation dans le cadre de la méthode POT est plus complexe, puisque, pour l'estimation des quantiles extrêmes, il nous faudrait inverser la fonction de répartition (Fdr) de la loi prédictive, loi qui n'est pas une loi classique et dont la Fdr n'a pas d'expression analytique.

Une alternative plus simple au niveau des calculs serait d'utiliser la loi a posteriori que l'on peut obtenir sur les quantiles extrêmes que l'on souhaite estimer (à partir d'un échantillon d'estimations du quantile calculées par la méthode POT à partir des valeurs de α et β données par l'algorithme de Gibbs). Dans ce cas, comme pour l'estimation des paramètres à travers un échantillon simulé selon leur loi a posteriori, il faudrait explorer les performances des différents estimateurs : par la moyenne a posteriori, ou la médiane a posteriori, voire le mode a posteriori des valeurs du quantile extrême obtenues par la méthode POT avec les valeurs de α et β simulées au cours de l'algorithme de Gibbs. De même, on peut déduire des échantillons de α et β donnés par l'algorithme de Gibbs, des échantillons de valeurs de γ et σ issues de la loi a posteriori sur ces paramètres induite par notre procédure bayésienne.

On pourrait en outre utiliser les différentes lois a posteriori (celles sur α et β, mais aussi celles que l'on peut en déduire sur un quantile extrême ou γ et σ) pour produire des intervalles de crédibilité. Ces intervalles permettent de mesurer l'incertitude des estimateurs bayésiens déduits de ces lois a posteriori, et ceci sans avoir besoin de recourir à des méthodes de type Monte-Carlo.

Dans la prochaine étape, nous souhaiterions exploiter l'avis d'expert sur les deux paramètres α et β. Or, comme on l'a vu au paragraphe 3.4, on ne peut utiliser l'avis d'expert pour l'un des paramètres (par exemple α) que si l'autre paramètre (β pour l'exemple) est fixé.

La solution que nous envisageons consiste alors à introduire l'étape de détermination des hyperparamètres dans l'algorithme de Gibbs, puisqu'à chaque itération (t) de l'algorithme les paramètres sont fixés ($\alpha = \alpha^{(t)}$ ou $\alpha^{(t+1)}$ et $\beta = \beta^{(t)}$ ou $\beta^{(t+1)}$). On obtient donc pour l'algorithme de Gibbs le schéma adaptatif suivant.

1. De la même façon que dans le paragraphe 3.4.2 page 137, calculer à partir de $\alpha^{(t)}$ les hyperparamètres $\delta^{(t+1)}$ et $\mu^{(t+1)}$:
$$\begin{aligned}\delta^{(t+1)} &= \frac{1}{\alpha^{(t)}}\left[z^2_{1-\varepsilon/2}\left(\frac{\beta_1+\beta_2}{\beta_2-\beta_1}\right)^2-1\right],\\ \eta^{(t+1)} &= 2\alpha^{(t)}z_{1-\varepsilon/2}\,\frac{\beta_1+\beta_2}{(\beta_2-\beta_1)^2}\left[z^2_{1-\varepsilon/2}\left(\frac{\beta_1+\beta_2}{\beta_2-\beta_1}\right)^2-1\right]^{-1},\end{aligned}$$
où $\beta_i = (q_{\max}-u_n)/((np_i/m_n)^{-1/\alpha^{(t)}}-1)$, $i = 1{,}2$.
2. Calculer les paramètres de la loi a posteriori de β sachant $\alpha^{(t)}$ et $\underline{z}^{(t)}_{m_n}$:
$$\delta'^{(t+1)} = \delta^{(t+1)} + n \quad \text{et} \quad \eta'^{(t+1)} = \frac{\delta^{(t+1)}\eta^{(t+1)} + \sum_{i=1}^{m_n} z_i^{(t)}}{\delta^{(t+1)} + m_n}.$$
3. Simuler $\beta^{(t+1)}$ selon la loi a posteriori de β sachant $\alpha^{(t)}$ et $\underline{z}^{(t)}_{m_n}$, $\pi(\beta|\,\alpha^{(t)},\underline{z}^{(t)}_{m_n}) = \mathcal{G}amma(\delta'^{(t+1)}\,\alpha^{(t)}+1\,,\delta'^{(t+1)}\eta'^{(t+1)})$.
4. $\forall i = 1,\ldots,m_n$, simuler indépendamment $z_i^{(t+1)}$ selon la loi a posteriori de z sachant x_i, $\alpha^{(t)}$ et $\beta^{(t+1)}$, $q_\pi(z|x_i,\alpha^{(t)},\beta^{(t+1)}) = \mathcal{G}amma(\alpha^{(t)}+1\,,\beta^{(t+1)}+x_i)$.

5. Calculer le dernier hyperparamètre qui, comme dans le paragraphe 3.4.1 page 136, vérifie l'équation (3.5) page 105 :

$$\mu^{(t+1)} = \eta^{(t+1)} \exp\left(\ln \delta^{(t+1)} - \psi(\alpha^{(t)}\delta^{(t+1)} + 1) + \psi(\alpha^{(t)})\right) .$$

6. Calculer le paramètre $\mu'^{(t+1)}$ de la loi a posteriori de α sachant $\beta^{(t+1)}$ et $\underline{z}_{m_n}^{(t+1)}$:

$$\mu'^{(t+1)} = \left(\mu^{(t+1)}\right)^{\delta^{(t+1)}/(\delta^{(t+1)}+m_n)} \left(\prod_{i=1}^{m_n} z_i^{(t+1)}\right)^{1/(\delta^{(t+1)}+m_n)} .$$

7. Simuler $\alpha^{(t+1)}$ selon la loi a posteriori de α sachant $\beta^{(t+1)}$ et $\underline{z}_{m_n}^{(t+1)}$, $\pi(\alpha|\,\beta^{(t+1)}, \underline{z}_{m_n}^{(t+1)})$ $= \mathrm{gamconII}(\eta'^{(t+1)}/\mu'^{(t+1)}, \delta'^{(t+1)})$.

Il faudrait tout d'abord montrer que ce schéma converge vers un régime stationnaire, puis le tester sur des données simulées et réelles.

En dernier lieu, remarquons que pour estimer les deux paramètres α et β, les lois a priori conjuguées conduisent à utiliser trois hyperparamètres δ, η et μ, alors que nous souhaiterions disposer de quatre hyperparamètres, deux pour α et deux pour β. Les deux hyperparamètres relatifs à α (resp. β) doivent nous permettre de coder la valeur médiane ou centrale (position), et la dispersion (échelle) de la loi a priori de α (resp. la loi a priori de β).

L'utilisation d'une procédure hiérarchique permet d'aboutir à ce résultat. Nous choisissons de prendre $\delta = 1$, et nous utilisons la reparamétrisation suivante :

$$\eta = \xi_1 > 0 \quad \text{et} \quad \mu = \xi_1 e^{-\xi_2}, \quad \text{avec} \quad \xi_1 > 0, \xi_2 > 0\,,$$

de sorte que $\eta/\mu = \exp(\xi_2) > 1$. Nous obtenons ainsi

- la densité de la loi a priori de α : $\pi(\alpha|\,\xi_1, \xi_2) = \pi(\alpha|\,\xi_2) = \mathcal{G}amma(2, \xi_2)$.
- $\forall \alpha > 0$, densité conditionnelle de la loi a priori de β sachant α : $\pi(\beta|\,\alpha, \xi_1, \xi_2) = \pi(\beta|\,\alpha, \xi_1) = \mathcal{G}amma(\alpha + 1, \xi_1)$.

On peut remarquer que $1/\xi_2$, resp. $1/\xi_1$, règlent la position centrale et la dispersion a priori de α, resp. β, ce qui n'est pas satisfaisant. Nous avons donc placé sur les hyperparamètres ξ_1 et ξ_2 de niveau un les lois a posteriori suivantes :

- loi a priori de ξ_1 : $\pi_1(\xi_1|\,a_1, b_1) =$ la loi $\mathcal{G}amma(a_1, b_1)$ et
- loi a priori de ξ_2 : $\pi_2(\xi_2|\,a_2, b_2) =$ la loi $\mathcal{G}amma(a_2, b_2)$.

Il s'ensuit les lois a posteriori suivantes :

- sur ξ_1 : $\pi_1(\xi_1|\,\alpha, \beta, a_1, b_1) =$ la loi $\mathcal{G}amma(a_1 + \alpha + 1, b_1 + \beta)$.
- sur ξ_2 : $\pi_2(\xi_2|\,\alpha, a_2, b_2) =$ la loi $\mathcal{G}amma(a_2 + 2, b_2 + \alpha)$.
- sur α : $\pi(\alpha|\,a_2, b_2)$ telle que α/b_2 suit une loi $\mathrm{betaII}(2, a_2)$.
- sur β : $\pi(\beta|\,\alpha, a_2, b_2)$telle que β/b_1 suit une loi $\mathrm{betaII}(\alpha, a_1)$.

Les deux dernières lois a posteriori (pour α et pour β) sont obtenues en intégrant en ξ_2 et ξ_1 respectivement. Il faut choisir $a_1 > 2$ et $a_2 > 2$ pour que ces deux lois admettent des moments d'ordre 2. Des choix appropriés de a_1, b_1, a_2, b_2 (en fonction de l'avis d'expert dont la forme est encore à déterminer) permettront de choisir le centre et la dispersion de chacune de ces deux lois.

Enfin, on pourrait utiliser d'autres structures hiérarchiques pour introduire des covariables, prendre en compte une dépendance spatiale, ou regrouper de petits échantillons. Plus généralement, à partir d'une structure hiérarchique, on envisage d'analyser statistiquement des modèles complexes en exploitant la richesse et la puissance des méthodes de Monte-Carlo par chaînes de Markov.

Conclusion et perspectives

Dans cette thèse, nous avons travaillé sur l'estimation des risques d'occurrence d'événements rares, ainsi que sur l'estimation de quantiles extrêmes. Les exemples applicatifs que nous avons présentés concernent principalement la fiabilité des structures, qui intéressait particulièrement le groupe Fiabilité des Composants et des Structures de la division Recherche et Développement d'EDF, qui a co-financé cette thèse. Nous avons essayé d'apporter des solutions aux problèmes pratiques des ingénieurs de ce groupe, et plus généralement aux problèmes analogues de tous ceux qui travaillent dans le domaine des risques extrêmes liés à des événements rares. En particulier, nous avons souhaité proposer des solutions aux problèmes liés aux données de faible taille.

Nous avons tout d'abord proposé d'utiliser des modèles paramétriques classiques tirant partie de l'information contenue dans tout l'échantillon lorsque la petite partie des données utilisée dans les méthodes spécifiques pour l'estimation des événements rares est trop réduite et ne permet plus l'estimation. Cependant, les modèles paramétriques classiques sont principalement influencés par les valeurs les plus probables de la variable, et il faut pouvoir contrôler leurs propriétés d'extrapolation en queue de distribution. Nous avons donc proposé des méthodes de sélection de modèles paramétriques classiques du point de vue de la qualité de leur ajustement en queue de distribution : il s'agit des tests ET et GPD d'adéquation à la queue de distribution.

Le test ET nous permet d'obtenir des résultats satisfaisants sur données simulées et sur données réelles, et ceci même dans le cas d'échantillons de petite taille. Par exemple, nous obtenons des indications utilisables pour un échantillon d'observations réelles de taille $n = 24$. Le test GPD est encore en cours d'étude, mais nos premiers résultats sont encourageants. Les valeurs du nombre d'excès m_n et de l'ordre $1 - p$ du quantile extrême utilisés pour construire le test doivent être choisis par l'utilisateur. Dans le cas du test GPD, les valeurs à choisir pour retrouver le niveau attendu et obtenir une puissance satisfaisante sont encore à déterminer pour un ensemble représentatif de modèles. D'autre part, les tests ET et GPD ne sont appliqués qu'en un point déterminé de la queue de distribution, le quantile d'ordre $1 - p$. Dans le but d'explorer une plus grande partie de la queue de distribution, on peut envisager d'appliquer ces tests en plusieurs points relativement éloignés les uns des autres, c'est-à-dire pour plusieurs valeurs de p. Il faudrait alors explorer le niveau et la puissance des tests ET et GPD dans cette perspective.

Nous avons ensuite considéré le cas où l'utilisateur n'était pas seulement intéressé par la modélisation de lois d'événements rares, mais aussi par la partie centrale de la loi. Nous cherchons alors un modèle qui s'ajuste correctement à l'ensemble de la loi. Lorsque les modèles testés sont rejetés à la fois par un test classique (qui éprouve l'adéquation en partie centrale) et par un test extrême (comme le test ET ou le test GPD), nous proposons une procédure de régularisation bayésienne. Le point de départ de cette procédure étant un modèle adapté aux valeurs les plus probables de la variable, on cherche à améliorer l'estimation de la loi des événements rares à l'aide d'un avis d'expert portant sur les queues de distribution. Cette procédure bayésienne a été implémentée et étudiée pour les lois normale, lognormale, exponentielle, gamma et de Weibull, c'est-à-dire dans le cadre du domaine d'attraction de Gumbel dans lequel a été développé le test ET. Il serait à présent judicieux d'étendre l'ensemble des modèles étudiés, notamment à des lois appartenant aux autres domaines d'attraction (Fréchet et Weibull) de la loi des valeurs extrêmes, c'est-à-dire des modèles auxquels on applique le test GPD.

Le test ET et la procédure de régularisation bayésienne ont été implémentés au cours de ma thèse dans une maquette logiciel MATLAB fournie à EDF. Ces programmes permettent aux ingénieurs d'explorer et d'utiliser facilement ainsi que de manière conviviale certaines des méthodes que nous avons développées. Il est maintenant nécessaire d'implémenter ces méthodes dans un langage compilé plus rapide, et d'élargir l'ensemble des méthodes proposées, d'une part en ajoutant les autres méthodes étudiées dans cette thèse, et d'autre part en intégrant d'autres méthodes classiques d'inférence pour les valeurs extrêmes. Cette maquette est donc en cours de transformation par un ingénieur en un logiciel industriel qui devrait permettre une meilleure diffusion de nos méthodes. Ce logiciel devrait être disponible sur le site du projet IS2 en 2003.

Le dernier chapitre de cette thèse reste le plus ouvert. Abandonnant les modèles paramétriques classiques, nous souhaitons utiliser la méthode des excès (POT, *Peaks Over Threshold*) pour l'estimation des quantiles extrêmes, laquelle peut produire des estimations biaisées en particulier pour des échantillons de petite taille. Nous cherchons donc à réduire ce biais d'estimation des quantiles extrêmes, en particulier en introduisant de l'information supplémentaire sur la queue de distribution à travers un avis d'expert. Ce biais provient surtout de l'estimation les deux paramètres des lois de Pareto généralisées (GPD, Generalised Pareto Distribution) à partir d'échantillons d'excès au-delà d'un seuil, lesquels ne suivent gńéralement pas une loi GPD (mais une loi proche). En effet, les lois GPD sont seulement les lois asymptotiques des excès au-delà d'un seuil lorsque ce seuil tend vers le point terminal de la loi d'origine. Nous proposons une estimation bayésienne des paramètres de la loi GPD approximante, en nous restreignant au domaine d'attraction de Fréchet (cas où le paramètre de forme γ de la loi GPD est positif). Cette méthode d'estimation bayésienne des paramètres de la loi GPD se révèle satisfaisante dans le cas d'échantillons simulés selon une loi GPD, et aussi dans le cas d'échantillons d'excès, et ceci même lorsqu'on utilise une procédure bayésienne empirique, c'est-à-dire sans incorporer d'avis d'expert. Nous avons ensuite proposé, avec des résultats satisfaisants, une première manière d'introduire un avis d'expert. Il s'agit d'une procédure hybride qui fait porter l'avis d'expert sur l'un des paramètres de la loi GPD, et utilise une information empirique déduite de l'échantillon pour l'autre paramètre. Il serait à présent intéressant d'étudier un schéma bayésien permettant d'introduire un avis d'expert sur les deux paramètres de la loi GPD, du type de celui présenté dans la conclusion du chapitre 3. Une autre perspective de l'approche bayésienne est de pouvoir introduire une dépendance spatiale entre différents sites de mesure, ou de regrouper différents échantillons de mesures, etc., en utilisant une méthode bayésienne hiérarchique. Enfin, cette étude étant restreinte au domaine d'attraction de Fréchet, il serait intéressant de trouver une autre structure de mélange permettant de l'étendre au domaine d'attraction de Weibull.

Annexe A

Paramétrage des lois utilisées

On note f les densités et F les fonctions de répartition des différents modèles.

Loi normale $\mathcal{N}(\mu, \sigma)$ ($\mu \in I\!R$ et $\sigma > 0$) : $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$ pour $x \in I\!R$.

Loi exponentielle

- paramétrage $\mathcal{E}xp(\eta)$ ($\eta > 0$) : $1 - F(x) = e^{-x/\eta}$ pour $x \geq 0$.
- paramétrage $\mathcal{E}xp(\lambda)$ ($\lambda > 0$) : $1 - F(x) = e^{-\lambda x}$ pour $x \geq 0$.

Loi de Weibull

- paramétrage $\mathcal{W}(\eta, \beta)$ ($\eta > 0$ et $\beta > 0$) : $1 - F(x) = \exp(-x^{\beta}/\eta)$ pour $x > 0$.
- paramétrage $\mathcal{W}(\lambda, \beta)$ ($\lambda > 0$ et $\beta > 0$) : $1 - F(x) = \exp\left(-(x/\lambda)^{\beta}\right)$ pour $x > 0$.

Loi lognormale $\mathcal{LN}(\mu, \sigma)$ ($\mu \in I\!R$ et $\sigma > 0$) : $f(x) = \dfrac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\dfrac{(\ln x-\mu)^2}{2\sigma^2}\right)$ pour $x > 0$.

Loi gamma $\mathcal{G}amma(a, \lambda)$ ($a > 0$ et $\lambda > 0$) : $f(x) = \dfrac{\lambda^a}{\Gamma(a)} e^{-\lambda x} x^{a-1}$ pour $x > 0$, où $\Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x} dx$.

Loi double exponentielle $\mathcal{DE}xp(\theta)$ ($\theta > 0$) : $1 - F(x) = \exp(-e^x/\theta)$ pour $x \in I\!R$.

Loi beta $\mathcal{B}eta(a, b, [\ell_m, \ell_M])$ ($a > 0$, $b > 0$, $l_m \in I\!R$ et $l_M \in I\!R$) :
$f(x) = \dfrac{\Gamma(a+b)(x-\ell_m)^{a-1}(\ell_M - x)^{b-1}}{\Gamma(a)\Gamma(b)(\ell_M - \ell_m)^{a+b-1}}$ pour $x \in [\ell_m, \ell_M]$), où $\Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x} dx$.

Loi betaII $\mathcal{B}eta(II)(a,b)$ ($a>0$ et $b>0$) : $f(x) = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}(1+x)^{-(a+b)}$ pour $x>0$, où $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$.
On peut remarquer que si X est de loi $\mathcal{B}eta(II)(a,b)$, alors $Y = X/(X+1)$ est de loi $\mathcal{B}eta(a,b,[0,1])$.

Loi de Pareto généralisée

- paramétrage $GPD(\gamma,\sigma)$ ($\gamma \in I\!R$ et $\sigma > 0$) :
  - Si $\gamma \neq 0$, $f(x) = \dfrac{1}{\sigma}\left(1+\dfrac{\gamma x}{\sigma}\right)^{-1/\gamma-1}$, pour $\begin{cases} x \in I\!R_+^* \text{ si } \gamma > 0, \\ x \in [0,\, -\sigma/\gamma[\text{ si } \gamma < 0. \end{cases}$
  - Si $\gamma = 0$, $f(x) = \dfrac{1}{\sigma}\exp\left(-\dfrac{x}{\sigma}\right)$, pour $x \in I\!R_+^*$.
- paramétrage $GPD(\alpha,\beta)$ ($\alpha > 0$ et $\beta > 0$ ou $\alpha < 0$ et $\beta < 0$) :

$$f_{\alpha,\beta}(x) = \frac{\alpha}{\beta}\left(1+\frac{x}{\beta}\right)^{-\alpha-1} \quad \text{pour} \quad \begin{cases} x \in I\!R_+ \text{ si } \alpha > 0. \\ x \in [0,\, -\beta[\text{ si } \alpha < 0. \end{cases}$$

Loi de Cauchy $\mathcal{C}auchy(\mu,\sigma)$ ($\mu \in I\!R$ et $\sigma > 0$) : $f(x) = \left(\pi\sigma\left(1+\left(\dfrac{x-\mu}{\sigma}\right)^2\right)\right)^{-1}$, pour $x \in I\!R$.

Loi de gamcon de type II gamconII(c,d), ($c>1$ et $d>0$) :
$f(x) = (I_{c,d})^{-1}\Gamma(dx+1)\,(\Gamma(x))^{-d}(cd)^{-dx}$, pour $x \in I\!R_+^*$.

Loi de Fréchet $\mathcal{F}rechet(\beta)$ ($\beta > 0$) : $F(x) = \exp\left(-x^{-1/\beta}\right)$ pour $x > 0$.

Loi de Burr $\mathcal{B}urr(\beta,\tau,\lambda)$ ($\beta>0$, $\tau>0$ et $\lambda>0$) : $F(x) = 1-\left(\dfrac{\beta}{\beta+x^\tau}\right)^\lambda$ pour $x>0$.

Loi de Student à ν degrés de liberté $t(\nu)$ ($\nu > 0$) :
$f(x) = \dfrac{1}{\sqrt{\nu\pi}}\,\dfrac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\left(1+\dfrac{x^2}{\nu}\right)^{-(\nu+1)/2}$ pour $x \in I\!R$.

Loi de Student absolue $t_{Abs}(\nu)$ ($\nu > 0$) : $f(x) = \dfrac{2}{\sqrt{\nu\pi}}\,\dfrac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\left(1+\dfrac{x^2}{\nu}\right)^{-(\nu+1)/2}$ pour $x > 0$.

Loi loggamma $\mathcal{L}\Gamma(a)$ ($a>0$) : $f(x) = \dfrac{1}{\Gamma(a)}\, x^{-2}(\ln x)^{a-1}$ pour $x \in I\!R$, où $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$.

Annexe B

Annexes sur la description du test ET

Nous allons tout d'abord rappeler les propriétés des fonctions lisses à variations régulières. Puis nous donnerons et démontrerons les propriétés des fonctions appartenant aux classes $\mathcal{C}^1_\rho$, $\mathcal{C}^2$ et $\mathcal{C}^3_\rho$. Enfin, nous nous proposons de montrer les lemmes 3 et 4 (page 17) qui n'ont pas été prouvés.

B.1 Propriétés des fonctions lisses à variations régulières et des classes $\mathcal{C}^1_\rho$, $\mathcal{C}^2$ et $\mathcal{C}^3_\rho$

Les propriétés des fonctions lisses à variations régulières que nous utiliserons sont résumées dans le lemme suivant.

Lemme 9 (Propriétés des fonctions lisses à variations régulières)

1. *$\mathcal{SR}_\alpha \subset \mathcal{R}_\alpha$. Donc, si $f \in \mathcal{SR}_\alpha$, alors $\forall t > 0$, $f(tx)/f(x) \to t^\alpha$ lorsque $x \to \infty$ c'est-à-dire $f(tx) \sim t^\alpha f(x)$ lorsque $x \to \infty$.*
2. *Si $f \in \mathcal{SR}_\alpha$, alors $f(x) \underset{x\to\infty}{\to} \begin{cases} +\infty & si \quad \alpha > 0 \\ 0 & si \quad \alpha < 0 \end{cases}$*
3. *Si $f \in \mathcal{SR}_\alpha$, alors $xf'(x)/f(x) \to \alpha$ quand $x \to \infty$, et la fonction $xf'(x)/f(x)$ est continue.*
4. *Si $f \in \mathcal{SR}_\alpha$, $\alpha \neq 0$, alors $|f'| \in \mathcal{SR}_{\alpha-1}$, $1/f \in \mathcal{SR}_{-\alpha}$ et, lorsque f^{-1} est bien définie, on a $f^{-1} \in \mathcal{SR}_{1/\alpha}$.*
5. *Si $f \in \mathcal{SR}_\alpha$, alors $\forall s$ $f^s \in \mathcal{SR}_{s\alpha}$.*
6. *Si $f \in \mathcal{SR}_\alpha$ et $g \in \mathcal{SR}_\beta$, alors $f + g \in \mathcal{SR}_\rho$ où $\rho = \max(\alpha, \beta)$ lorsque $\alpha \neq \beta$, $fg \in \mathcal{SR}_{\alpha+\beta}$ et, si $g(x) \to +\infty$ lorsque $x \to +\infty$, $f \circ g \in \mathcal{SR}_{\alpha\beta}$.*
7. *Si $f \in \mathcal{SR}_\alpha$, alors il existe une fonction $\ell \in \mathcal{R}_0$ telle que $f(x) = x^\alpha \ell(x)$.*
8. *Si $f \in \mathcal{SR}_\alpha$ et $x_n \sim y_n$ alors $f(x_n) \sim f(y_n)$.*

Ces propriétés sont décrites et démontrées dans [8].

Les fonctions H appartenant aux trois classes $\mathcal{C}^1_\rho$, $\mathcal{C}^2$ et $\mathcal{C}^3_\rho$ possèdent les propriétés suivantes.

Lemme 10

1. Si $H \in \mathcal{C}^1_\rho$, $\dfrac{x(H^{-1})'(x)}{H^{-1}(x)} \underset{x\to\infty}{\to} \dfrac{1}{\rho}$ *et* $\dfrac{x(H^{-1})''(x)}{(H^{-1})'(x)} \underset{x\to\infty}{\to} \dfrac{1}{\rho} - 1$.

2. Si $H \in \mathcal{C}^2$, $\dfrac{x(H^{-1})'(x)}{H^{-1}(x)} \underset{x\to\infty}{\to} 0$ *et* $\dfrac{x(H^{-1})''(x)}{(H^{-1})'(x)} \underset{x\to\infty}{\to} -1$.

3. Si $H \in \mathcal{C}^3_\rho$, $\dfrac{x(H^{-1})'(x)}{H^{-1}(x)} \underset{x\to\infty}{\to} \infty$ *et* $\dfrac{x(H^{-1})''(x)}{(H^{-1})'(x)} \underset{x\to\infty}{\to} \infty$.

Démonstration : Elle s'appuie sur les propriétés des fonctions lisses (cf. lemme 9, page 163, en particulier la propriété 3),

- Si $H \in \mathcal{SR}_\rho$, $H^{-1} \in \mathcal{SR}_{1/\rho}$ et $(H^{-1})' \in \mathcal{SR}_{1/\rho-1}$.
- Si $H \in \mathcal{C}^2$, $H^{-1} \in \mathcal{SR}_0$ et $(H^{-1})' \in \mathcal{SR}_{-1}$.
- Si $H \in \mathcal{C}^3_\rho$, $H^{-1} = \exp g$ où $g \in \mathcal{SR}_\rho$. On en déduit que $(H^{-1})' = g' \exp g$ avec $g' \in \mathcal{SR}_{\rho-1}$, et que $(H^{-1})'' = ((g')^2 + g'')\exp g$ avec $(g')^2 + g'' \in \mathcal{SR}_{2(\rho-1)}$.

■

Lemme 11

$$\frac{(H^{-1})'}{H^{-1}} \in \begin{cases} \mathcal{SR}_{-1} & si \quad H \in \mathcal{C}^1_\rho \quad ou\ \mathcal{C}^2 \\ \mathcal{SR}_{\rho-1} & si \quad H \in \mathcal{C}^3_\rho \end{cases} \tag{B.1}$$

et de même, si on note $\mathcal{O}$ *la fonction nulle (telle que pour tout* x *réel* $\mathcal{O}(x) = 0$*), on a*

$$\left|\frac{(H^{-1})''}{(H^{-1})'}\right| \in \begin{cases} \mathcal{SR}_{-1} & si \quad H \in \mathcal{C}^1_\rho,\ \rho \neq 1,\ ou\ \mathcal{C}^2 \\ \mathcal{SR}_{-1-\tau} & si \quad H \in \mathcal{C}^1_1\ avec\ |(H^{-1})''| \in SR_{-1-\tau} \\ \{\mathcal{O}\} & si \quad H \in \mathcal{C}^1_1\ et\ H^{-1}\ est\ une\ fonction\ affine \\ \mathcal{SR}_{\rho-1} & si \quad H \in \mathcal{C}^3_\rho \end{cases} \tag{B.2}$$

Démonstration : On utilise les mêmes propriétés que précédemment pour les fonctions H^{-1} et $(H^{-1})'$, plus le fait que

- Si $H \in \mathcal{C}^1_\rho = \mathcal{SR}_\rho$, pour $\rho \neq 1$, $(H^{-1})' \in \mathcal{SR}_{1/\rho-1}$ et $|(H^{-1})''| \in \mathcal{SR}_{1/\rho-2}$.
- Si $H \in \mathcal{C}^1_1$, $(H^{-1})' \in \mathcal{SR}_0$ et soit $|(H^{-1})''| \in SR_{-1-\tau}$, soit $(H^{-1})'' = \mathcal{O}$.
- Si $H \in \mathcal{C}^2$, $(H^{-1})' \in \mathcal{SR}_{-1}$ et $|(H^{-1})''| \in \mathcal{SR}_{-2}$.
- Si $H \in \mathcal{C}^3_\rho$, $(H^{-1})' = g' \exp g$ où $g' \in \mathcal{SR}_{\rho-1}$ et $(H^{-1})'' = ((g')^2 + g'')\exp g$, où $(g')^2 + g'' \in \mathcal{SR}_{2(\rho-1)}$.

■

B.2 Preuve des lemmes 3 et 4

Démonstration du lemme 3 : Notons $-\ln\zeta' = \kappa_2 - \ln\zeta$ et $-\ln\zeta'' = \kappa_2 - \ln\zeta - \kappa_3(\kappa_2 + \kappa_1 \ln x)$. On peut remarquer que, puisque $\zeta \to 0$, que κ_1, κ_2, et κ_3 restent bornés, et que $\ln x$ est infiniment petit devant $\ln\zeta$, on a $\ln\zeta' \sim \ln\zeta$ et $\ln\zeta'' \sim \ln\zeta$. A présent, considérons deux cas :

- Si H appartient à $\mathcal{C}^1_\rho$ ($\rho > 0$), ou à $H \in \mathcal{C}^2$, alors $(H^{-1})'$ et $(H^{-1})''$ sont des fonctions à variations lisses et le résultat est la conséquence du lemme 9 (page 163, propriété 8) : pour $H \in \mathcal{C}^1_1$, lorsque H^{-1} est une fonction affine, $(H^{-1})''$ n'est pas à variations lisses, mais $(H^{-1})''/(H^{-1})' = 0$ et dans ce cas on a même l'égalité des deux termes de l'équivalence.

- Supposons que $H \in \mathcal{C}^3_\rho$ ($0 < \rho < 1$). On a donc $H = \exp g$ où $g \in \mathcal{SR}_\rho$. On veut montrer que
$$\frac{(H^{-1})''(-\ln\zeta'')}{(H^{-1})'(-\ln\zeta')}\,\frac{(H^{-1})'(-\ln\zeta)}{(H^{-1})''(-\ln\zeta)} \sim 1.$$
Les preuves de $(H^{-1})''(-\ln\zeta'')/(H^{-1})''(-\ln\zeta) \sim 1$ et $(H^{-1})'(-\ln\zeta)/(H^{-1})'(-\ln\zeta')$ ~ 1 sont similaires. Considérons par exemple le premier des deux termes :
$$\begin{aligned}\frac{(H^{-1})''(-\ln\zeta'')}{(H^{-1})''(-\ln\zeta)} &= \frac{(g''+g'^2)(-\ln\zeta'')}{(g''+g'^2)(-\ln\zeta)}\exp(g(-\ln\zeta'') - g(-\ln\zeta)) \\ &\sim \exp(g(-\ln\zeta'') - g(-\ln\zeta)),\end{aligned}$$
puisque $g'' + g'^2 \in SR_{2\rho-2}$ et que $-\ln\zeta'' \sim -\ln\zeta$ (cf. lemme 9 page 163, propriété 8). Un développement de Taylor de g montre qu'il existe $\kappa \in]0,1[$ tel que
$$\frac{(H^{-1})''(-\ln\zeta'')}{(H^{-1})''(-\ln\zeta)} \sim \exp(\ln(\zeta/\zeta'')g'(-\ln\zeta'' - \kappa\ln(\zeta/\zeta''))).$$
A présent, remarquons que $\ln(\zeta/\zeta'') = \kappa_2 - \kappa_3(\kappa_2 + \kappa_1 \ln x)$ est indépendant de ζ, et que $-\ln\zeta'' \to \infty$ lorsque $\zeta \to 0$. On en déduit que $g'(-\ln\zeta'' - \kappa\ln(\zeta/\zeta'')) \to 0$ quand $\zeta \to 0$. puisque g' est une fonction à variations lisses avec un indice négatif (cf. lemme 9 page 163, propriété 2).

■

Démonstration du lemme 4 :

– D'après le lemme 10 (page 164), on sait que si $H \in \mathcal{C}^1_\rho$, $x(H^{-1})''(x)/(H^{-1})'(x) \to 1/\rho - 1$ lorsque $x \to \infty$. Donc, si $\rho \neq 1$, on a $(H^{-1})''(x)/(H^{-1})'(x) \sim (1/\rho - 1)/x$. Alors,
$$\sqrt{m_n}\frac{(H^{-1})''(-\ln c_n)}{(H^{-1})'(-\ln c_n)} \sim \sqrt{m_n}\left(\frac{1}{\rho} - 1\right)\frac{1}{-\ln c_n} \sim \sqrt{m_n}\left(\frac{1}{\rho} - 1\right)\frac{1}{\ln n} \underset{n\to\infty}{\to} 0.$$

- Si $H \in \mathcal{C}_1^1$, d'après le lemme 11 (page 164), on sait que soit $(H^{-1})''(x)/(H^{-1})'(x) = \mathcal{O}$, ce qui implique que
$$\sqrt{m_n}\frac{(H^{-1})''(-\ln c_n)}{(H^{-1})'(-\ln c_n)} = 0,$$
soit $(H^{-1})''(x)/(H^{-1})'(x) \in \mathcal{SR}_{-1-\tau}$. Dans ce dernier cas, on a, pour une fonction L à variations lentes,
$$\begin{aligned}\sqrt{m_n}\frac{(H^{-1})''(-\ln c_n)}{(H^{-1})'(-\ln c_n)} &\sim \sqrt{m_n}(-\ln c_n)^{-1-\tau}L(-\ln c_n)\\ &\sim \sqrt{m_n}(\ln n)^{-1-\tau}L(\ln n) \to 0 \quad \text{lorsque} \quad n \to \infty.\end{aligned}$$
- De même, si $H \in \mathcal{C}^2$, on montre que $(H^{-1})''(x)/(H^{-1})'(x) \sim -1/x$. On en déduit que
$$\sqrt{m_n}\frac{(H^{-1})''(-\ln c_n)}{(H^{-1})'(-\ln c_n)} \sim \frac{-\sqrt{m_n}}{-\ln c_n} \sim -\frac{\sqrt{m_n}}{\ln n} \underset{n\to\infty}{\to} 0.$$
- Si $H \in \mathcal{C}_\rho^3$, on sait que $(H^{-1})''(x)/(H^{-1})'(x) \in \mathcal{SR}_{\rho-1}$ (cf. lemme 11 page 164). Il en résulte que $(H^{-1})''(x)/(H^{-1})'(x) \sim x^{\rho-1}L(x)$ où L est une fonction à variations lentes. On en conclut que
$$\begin{aligned}\sqrt{m_n}\frac{(H^{-1})''(-\ln c_n)}{(H^{-1})'(-\ln c_n)} &\sim \sqrt{m_n}(-\ln c_n)^{\rho-1}L(-\ln c_n)\\ &\sim \sqrt{m_n}(\ln n)^{\rho-1}L(\ln n) \to 0 \quad \text{lorsque} \quad n \to \infty.\end{aligned}$$

■

B.3 Calcul de d_n approximation au premier ordre de l'erreur d'approximation δ_n

On suppose que les valeurs de m_n et p_n que l'on considère à présent sont de la forme
$$p_n = n^{-p'}(\ln n)^{-q'} \quad \text{et} \quad m_n = n^{1-p}(\ln n)^{-q}, \tag{B.3}$$
avec
- $p \leq 1$ et $\begin{cases} q \leq 0 & \text{si } p = 1 \\ q \text{ quelconque} & \text{si } p < 1 \end{cases}$
- $p' \geq 1$ et $\begin{cases} q' \geq 0 & \text{si } p' = 1 \\ q' \text{ quelconque} & \text{si } p' > 1 \end{cases}$

Ces conditions sont bien telles que $p_n < 1/n$ et m_n vérifie l'équation (1.10), page 16. De plus, elles permettent de contrôler l'éloignement de p_n par rapport à $1/n$ et la croissance de m_n par rapport à n. Des conditions analogues sont définies par Diebolt et Girard [28, 31].

Dans ce cadre Girard et Diebolt (voir [30] théorème 1 page 5) ont obtenu, pour chaque modèle considéré, une approximation au premier ordre, notée $\varepsilon^{(1)}_{\text{app},n}$ de l'erreur d'approximation relative $\varepsilon_{\text{app},n} = \delta_n/q_{1-p_n}$. Il suffit alors de multiplier $\varepsilon^{(1)}_{\text{app},n}$ par une approximation au premier ordre du quantile q_{1-p_n} pour en déduire une expression de d_n. Les valeurs obtenues sont énumérées dans le tableau B.1.

modèle	d_n	$\varepsilon^{(1)}_{\text{app},n}$	propriété
$\mathcal{E}xp(\lambda)$	0	0	
$\mathcal{N}(\mu,\sigma^2)$	$-\dfrac{\sigma\sqrt{\ln n}}{4\sqrt{2}}\left(\dfrac{(q'-q)\ln\ln n}{\ln n}\right)^2$	$-\dfrac{1}{8}\left(\dfrac{(q'-q)\ln\ln n}{\ln n}\right)^2$	$p=p'=1$
$\mathcal{LN}(\mu,\sigma^2)$	$-\dfrac{\exp(\mu+\sigma\sqrt{2\ln n})}{4}\left(\dfrac{(q-q')\ln\ln n}{\ln n}\right)^2$	$-\dfrac{1}{4}\left(\dfrac{(q-q')\ln\ln n}{\ln n}\right)^2$	$p=p'=1$
$\mathcal{W}(\eta,\beta)$	$\dfrac{\eta^{1/\beta}(1-\beta)}{2\beta^2}\;\dfrac{(q'-q)^2(\ln\ln n)^2}{(\ln n)^{2-1/\beta}}$	$\dfrac{(1-\beta)}{2\beta^2}\left(\dfrac{(q'-q)\ln\ln n}{\ln n}\right)^2$	$p=p'=1$
$\mathcal{G}amma(a,\lambda)$	$\dfrac{a-1}{\lambda}\;\dfrac{p'}{p}\ln\left(\dfrac{p'}{p}\right)$	$\dfrac{(1-a)\ln(p'/p)}{p\ln n}$	$p\neq p'$

TAB. B.1 – *Valeurs de d_n et $\varepsilon^{(1)}_{app,n}$ sous la condition (B.3).*

On peut remarquer que dans le cadre plus général d'une fonction de répartition $F \in \mathcal{E}$, Girard et Diebolt (voir [31], lemme 3) montrent que

$$\varepsilon^{(1)}_{\text{app},n} = \frac{1}{2}\left(\ln\left(\frac{m_n}{np_n}\right)\right)^2 \frac{(H^{-1})''(r_n)}{H^{-1}(-\ln(p_n))},$$

où $H = -\ln(1-F)$ et $r_n \in [-\ln(m_n/n),\ -\ln(p_n)]$. En multipliant cette expression par $q_{1-p_n} = F^{-1}(1-p_n) = H^{-1}(-\ln(p_n))$, on obtient que

$$d_n = \frac{1}{2}\left(\ln\left(\frac{m_n}{np_n}\right)\right)^2 (H^{-1})''(r_n). \tag{B.4}$$

Les valeurs de d_n du tableau B.1 sont obtenues en multipliant $\varepsilon^{(1)}_{\text{app},n}$ par une approximation au premier ordre de q_{1-p_n}:

- Pour l'exponentielle, l'erreur d'approximation δ_n, l'erreur relative $\varepsilon_{\text{app},n}$ et donc leurs approximations au premier ordre d_n et $\varepsilon^{(1)}_{\text{app},n}$ sont nulles puisqu'alors l'échantillon initial ainsi que l'échantillon des excès sont tous deux issus d'une loi exponentielle, de même paramètre qui plus est. Par conséquent, on a $q_{1-p_n} = -\ln p_n$, $u_n = -\ln(m_n/n)$ et $q_{ET} = -\ln(m_n/n) + \ln(m_n/np_n) = -\ln p_n = q_{1-p_n}$.

- Pour la loi normale, en première approximation (faire un développement limité) on a

$$1 - F(x) \underset{x\to\infty}{\sim} \frac{\sigma^2 f(x)}{x-\mu} \underset{x\to\infty}{\sim} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Il en découle que

$$q_{1-p_n} = (1-F)^{-1}(p_n) \underset{n\to\infty}{\sim} \mu + \sigma\sqrt{-2\ln p_n} \underset{n\to\infty}{\sim} \mu + \sigma\sqrt{2\ln n} \underset{n\to\infty}{\sim} \sigma\sqrt{2\ln n},$$

et donc que

$$d_n \underset{n\to\infty}{\sim} -\frac{\sigma\sqrt{\ln n}}{4\sqrt{2}}\left(\frac{(q'-q)\ln\ln n}{\ln n}\right)^2.$$

- Pour la loi lognormale, on approxime le quantile d'ordre $1-p_n$ par le logarithme de l'approximation au premier ordre du quantile d'ordre $1-p_n$ de la loi normale :

$$q_{1-p_n} \underset{n\to\infty}{\sim} \exp(\mu + \sigma\sqrt{-2\ln p_n}) \underset{n\to\infty}{\sim} \exp(\mu + \sigma\sqrt{2\ln n}).$$

On en déduit que

$$d_n \underset{n\to\infty}{\sim} -\frac{\exp(\mu+\sigma\sqrt{2\ln n})}{4}\left(\frac{(q-q')\ln\ln n}{\ln n}\right)^2.$$

- Pour la loi de Weibull, on connait une expression analytique du quantile d'ordre $1-p_n$:

$$q_{1-p_n} = (1-F)^{-1}(p_n) = \eta^{1/\beta}(-\ln p_n)^{1/\beta} \underset{n\to\infty}{\sim} \eta^{1/\beta}(\ln n)^{1/\beta}.$$

On en conclut que

$$d_n \underset{n\to\infty}{\sim} \frac{\eta^{1/\beta}(1-\beta)}{2\beta^2}\frac{(q'-q)^2(\ln\ln n)^2}{(\ln n)^{2-1/\beta}}.$$

- Enfin, pour la loi gamma, en première approximation (faire un développement limité), on a :

$$1-F(x) \underset{x\to\infty}{\sim} \frac{f(x)}{\lambda} \underset{x\to\infty}{\sim} e^{-\lambda x}.$$

Il s'ensuit que

$$q_{1-p_n} = (1-F)^{-1}(p_n) \underset{n\to\infty}{\sim} -\frac{\ln p_n}{\lambda} \underset{n\to\infty}{\sim} \frac{p'\ln n}{\lambda}.$$

Par conséquent, on a

$$d_n \underset{n\to\infty}{\sim} \frac{a-1}{\lambda}\,\frac{p'}{p}\ln\left(\frac{p'}{p}\right).$$

À présent, remarquons que, d'après l'équation (B.3), lorsque $n \to \infty$, on a les équivalents suivants :

- $-\ln(m_n/n) \sim p \ln n$,
- $-\ln p_n \sim p' \ln n$,
- $\ln(m_n/n) - \ln p_n \sim (p'-p) \ln\ln n$ si $p \neq p'$ et
- $\ln(m_n/n) - \ln p_n \sim (q'-q) \ln\ln n$ si $p = p' = 1$.

On en déduit alors des valeurs plus générales de d_n et $\varepsilon^{(1)}_{\text{app},n}$ présentées dans le tableau B.2.

loi	d_n	$\varepsilon^{(1)}_{\text{app},n}$
$\mathcal{E}xp(\lambda)$	0	0
$\mathcal{N}(\mu,\sigma^2)$	$-\frac{\sigma\sqrt{-\ln p_n}}{4\sqrt{2}}\left(1-\frac{\ln p_n}{\ln(\frac{m_n}{n})}\right)^2$	$-\frac{1}{8}\left(1-\frac{\ln p_n}{\ln(\frac{m_n}{n})}\right)^2$
$\mathcal{LN}(\mu,\sigma^2)$	$-\frac{\exp(\mu+\sigma\sqrt{-2\ln p_n})}{4}\left(1-\frac{\ln p_n}{\ln(\frac{m_n}{n})}\right)^2$	$-\frac{1}{4}\left(1-\frac{\ln p_n}{\ln(\frac{m_n}{n})}\right)^2$
$\mathcal{W}(\eta,\beta)$	$\frac{\eta^{1/\beta}(1-\beta)}{2\beta^2}\ \frac{(\ln(\frac{m_n}{n})-\ln p_n)^2}{(-\ln(\frac{m_n}{n}))^{2-1/\beta}}$	$\frac{1-\beta}{2\beta^2}\left(\frac{\ln(\frac{m_n}{n})}{\ln p_n}\right)^{1/\beta}\left(\frac{\ln(\frac{m_n}{n})-\ln p_n}{\ln(\frac{m_n}{n})}\right)^2$
$\mathcal{G}amma(a,\lambda)$	$\frac{a-1}{\lambda}\ \frac{\ln p_n}{\ln(\frac{m_n}{n})}\ln\left(\frac{\ln p_n}{\ln(\frac{m_n}{n})}\right)$	$\frac{a-1}{\ln p_n}\left(\frac{\ln p_n}{\ln(\frac{m_n}{n})}\right)$

Tab. B.2 – *Valeurs de d_n et $\varepsilon^{(1)}_{app,n}$ dans le cadre général.*

Annexe C

Démonstrations annexes sur les résultats du test ET

On se propose de montrer que les caractéristiques des différentes versions du test ET sont indépendantes de la valeur de certains des paramètres de la loi dont sont issues les données. En particulier, on montre que la probabilité que $\widehat{q}_{\text{param},\,n}$, l'estimateur paramétrique du quantile d'ordre $1 - p_n$, soit à l'intérieur de son intervalle de confiance, est indépendante de la valeur de ces paramètres. Pour montrer cela, il suffit de prouver que les signes des différences entre $\widehat{q}_{\text{param},\,n}$ et les bornes de l'intervalle de confiance sont indépendants des paramètres considérés.

C.1 Version 1 du test ET.

On montre tout d'abord l'indépendance des caractéristiques de la première version du test par rapport à des paramètres de position et d'échelle (ou par rapport au paramètre d'échelle si la loi ne possède pas de paramètre de position).

Soit $F_{v,\,w}$ la fonction de répartition d'une loi admettant un paramètre de position v et un paramètre d'échelle w, i.e. il existe deux réels v et w et une fonction de répartition $F_{0,\,1}$ tels que $F_{v,\,w}(x) = F_{0,\,1}((x - v)/w)$.

Scholie 1 *On veut tester l'adéquation de la famille de lois de fonctions de répartition $F_{v,\,w}$ à laquelle appartient la vraie loi des données. Alors, le résultat du test ET version 1 est indépendant des paramètres de position (le paramètre v) et d'échelle (le paramètre w) (ou simplement du paramètre d'échelle lorsque l'on n'a pas de paramètre de position) pour les lois normale, lognormale, exponentielle, gamma et Weibull.*

Remarque C.1 *Pour m_n fixé, le niveau du test ET version 1 est le même quelle que soit la valeur des paramètres de position et d'échelle. Les valeurs de m_n à utiliser pour obtenir un niveau adéquat de la version 1 du test ET sont donc indépendantes de ces paramètres de position et d'échelle. En effet, ces valeurs de m_n sont conditionnées par la proportion de* *fois où $\widehat{q}_{param,n}$ se trouve à l'intérieur de son intervalle de confiance pour ET. Pour que l'on ait toujours les mêmes valeurs optimales de m_n, il suffit que $\widehat{q}_{param,n}$ ait la même probabilité d'appartenir à son intervalle de confiance, quelles que soient les valeurs de v et w. Cette propriété découle d'une justification analogue à celle qui suit.*

Justification : On s'intéresse à l'intervalle de confiance de la version 1 du test. On veut montrer que, quelles que soient les valeurs des paramètres v et w, la différence entre la borne inférieure (resp.supérieure) de l'intervalle $IC_{re,n}$ et $\widehat{q}_{\text{param},n}$ est toujours du même signe. Cette différence s'exprime de la façon suivante :

$$\Delta_n = b_{inf} - \widehat{q}_{\text{param},n} = \widehat{q}_{ET} - \widehat{q}_{\text{param},n} + d_n - \widehat{\sigma}_n B_n z,$$

où z, le quantile approprié d'une loi $\mathcal{N}(0,1)$, et $B_n = \ln(m_n/np_n)/\sqrt{m_n}$ sont indépendants des paramètres de la loi testée.

On va donc réexprimer en fonction de v et w les quantités qui permettent de calculer Δ_n. Pour cela, on pose $X_i^0 = (X_i - v)/w$ (s'il n'y a qu'un paramètre d'échelle, on prend $v = 0$). Il s'ensuit que :

- $\widehat{\sigma}_n = (1/m_n)\sum_{j=1}^{m_n} X_{(n-m_n+j)} - X_{(n-m_n)} = w\widehat{\sigma}_n^0$, où $\widehat{\sigma}_n^0 = (1/m_n)\sum_{j=1}^{m_n} X_{(n-m_n+j)}^0 - X_{(n-m_n)}^0$;
- $\widehat{q}_{ET,n} = \widehat{u}_n + \widehat{\sigma}_n \ln(m_n/np_n) = v + w\,\widehat{q}_{ET,n}^0$, où $\widehat{q}_{ET,n}^0 = \widehat{u}_n^0 + \widehat{\sigma}_n^0 \ln(m_n/np_n)$, avec $\widehat{u}_n^0 = X_{(n-m_n)}^0$.

D'autre part, on suppose que $F_{v,w}(x) = F_{0,1}((x-v)/w)$, que $1 - F_{v,w}(x) = \exp(-H_{v,w}(x))$ et que $1 - F_{0,1}(x) = \exp(-H_{0,1}(x))$. On en déduit que $F_{v,w}^{-1}(x) = v + w\,F_{0,1}^{-1}(x)$, $H_{v,w}(x) = H_{0,1}((x-v)/w)$, $H_{v,w}^{-1}(x) = v + w\,H_{0,1}^{-1}(x)$, $(H_{v,w}^{-1})'(x) = w\,(H_{0,1}^{-1})'(x)$ et $(H_{v,w}^{-1})''(x) = w\,(H_{0,1}^{-1})''(x)$. Si on note $\widehat{v}_n$ et $\widehat{w}_n$ les estimateurs de v et w resp., il en découle que

- $\widehat{q}_{\text{param},n} = F_{\widehat{v}_n,\widehat{w}_n}^{-1}(1-p_n) = \widehat{v}_n + \widehat{w}_n q_{1-p_n}^0$, où $q_{1-p_n}^0 = F_{0,1}^{-1}(1-p_n)$;
- $d_n = (1/2)\ln^2(m_n/np_n)(H_{v,w}^{-1})''(r_n) = wd_n^0$, où $r_n \in [-\ln(m_n/n), -\ln(p_n)]$ et $d_n^0 = (1/2)\ln^2(m_n/np_n)(H_{0,1}^{-1})''(r_n)$, sous les conditions de l'équation (B.3) page 166, d'après l'équation (B.4) page 167.

Ceci permet de montrer que

$$\Delta_n = w\left[\frac{v - \widehat{v}_n}{w} - \frac{\widehat{w}_n}{w} q_{1-p_n}^0 + (\widehat{q}_{ET,n}^0 + d_n^0 - \sigma_n^0 B_n z)\right],$$

où $\widehat{q}_{ET,n}^0$, $q_{1-p_n}^0$, d_n^0 et σ_n^0 sont indépendants de v et w.

Pour la loi normale (et donc pour la loi lognormale car les calculs sont alors faits sur $\ln X$ de loi normale), remarquons que

$$\widehat{v}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \widehat{w}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{v}_n)^2}.$$

Il s'ensuit que

$$\frac{\widehat{v}_n - v}{w} = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i - v}{w} \quad \text{et} \quad \frac{\widehat{w}_n}{w} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - v}{w} - \frac{\widehat{v}_n - v}{w}\right)^2},$$

où $(X_i - v)/w = X_i^0 \sim F_{0,1}$. On en déduit que ces quantités sont elles aussi indépendantes de v et w. Ceci montre que Δ_n est indépendant de v, et que le signe de Δ_n est indépendant de la valeur de w, qui est toujours positive.

Pour la loi exponentielle, on a (avec par convention $v = 0$ et $\widehat{v}_n = 0$ puisque l'on considère une loi exponentielle sans paramètre de décentrage) $w = \eta$ et $\widehat{w}_n = \overline{X}$, d'où $\widehat{w}_n/w = \overline{X^0}$. De même, on en déduit que Δ_n est indépendant de la valeur du paramètre d'échelle w.

Pour la loi gamma, on a (avec $v = 0$ et $\widehat{v}_n = 0$ puisque sans paramètre de décentrage) $\widehat{w}_n = 1/\widehat{\lambda}_n = \overline{X}/\widehat{a}_n$ (cf. D'Agostino et Stephens [17]). Or, $\widehat{a}_n$ est défini par l'équation

$$\frac{1}{n}\sum_{i=1}^{n} \ln X_i - \ln \overline{X} = \frac{1}{n}\sum_{i=1}^{n} \ln X_i^0 - \ln \overline{X^0} = \psi(\widehat{a}_n) - \ln \widehat{a}_n\,,$$

où ψ est la fonction digamma. L'estimateur $\widehat{a}_n$ est donc indépendant de w. On en déduit que $\widehat{w}_n/w = \overline{X^0}/\widehat{a}_n$ est aussi indépendant de w. Il s'ensuit donc que Δ_n est indépendant de la valeur de w.

Pour la loi de Weibull, on a (avec $v = 0$ et $\widehat{v}_n = 0$ puisque sans paramètre de décentrage) d'après D'Agostino et Stephens [17],

$$\ln \widehat{w}_n = \frac{\ln \widehat{\eta}_n}{\widehat{\beta}_n} = -\frac{1}{\widehat{\beta}_n} \ln \left[\frac{1}{n}\sum_{i=1}^{n} \exp\left(-T_i\,\widehat{\beta}_n\right)\right]\,,$$

où $T_i = -\ln X_i = -\ln X_i^0 - \ln w$. On en déduit que

$$\ln(\widehat{w}_n/w) = -\frac{1}{\widehat{\beta}_n} \ln \left[\frac{1}{n}\sum_{i=1}^{n} \exp(-\widehat{\beta}_n \ln X_i^0)\right]\,.$$

D'autre part, on sait que $\widehat{\beta}_n$ est défini par l'équation

$$\frac{1}{\widehat{\beta}_n} = \overline{T} - \frac{\sum_{i=1}^{n} T_i \exp(-T_i\widehat{\beta}_n)}{\sum_{i=1}^{n} \exp(-T_i\widehat{\beta}_n)} = -\frac{1}{n}\sum_{i=1}^{n} \ln X_i^0 + \frac{\sum_{i=1}^{n} \ln X_i^0 \exp(\widehat{\beta}_n \ln X_i^0)}{\sum_{i=1}^{n} \exp(\widehat{\beta}_n \ln X_i^0)}.$$

Ceci montre que $\widehat{\beta}_n$ est indépendant de w, et donc $\widehat{w}_n/w$ aussi. Il s'ensuit que Δ_n est indépendant de la valeur de w.

Scholie 2 *On veut tester l'adéquation de la famille des lois de Weibull $\mathcal{W}(\eta, \beta)$ à laquelle appartient la vraie loi des données. Pour β grand devant 1 ($\beta \gg 1$), le résultat du test ET version 1 est approximativement indépendant de la valeur de β.*

Remarque C.2 *Comme précédemment, pour m_n fixé, le niveau du test ET version 1 est approximativement le même quelle que soit la valeur du paramètre de forme de la loi de Weibull, à condition que celui-ci soit grand devant 1. Les valeurs de m_n à utiliser pour obtenir un niveau adéquat de la version 1 du test ET sont donc approximativement indépendantes du paramètre de forme de la loi de Weibull, à condition que celui-ci soit grand devant 1.*

Justification : Cette justification est heuristique : on utilise des approximations au premier ordre.

Pour simplifier cette justification, on suppose que l'échantillon des X_i est de loi de Weibull de paramètre d'échelle $\eta = 1$ et de paramètre de forme β, et qu'on n'estime pas le paramètre d'échelle. On s'intéresse à l'intervalle de confiance de la version 1 du test. On veut montrer que, quelles que soient les valeurs de β, la différence entre la borne inférieure (resp. supérieure) de l'intervalle $IC_{re,n}$ et $\widehat{q}_{\text{param},n}$ est toujours du même signe. Cette différence s'exprime de la façon suivante :

$$\Delta_n = b_{inf} - \widehat{q}_{\text{param},n} = \widehat{q}_{ET} - \widehat{q}_{\text{param},n} + d_n - \widehat{\sigma}_n B_n z,$$

où z, le quantile approprié d'une loi $\mathcal{N}(0,1)$, ainsi que $B_n = \ln(m_n/np_n)/\sqrt{m_n}$, sont indépendants des paramètres de la loi testée.

On va donc réexprimer en fonction de β les quantités qui permettent de calculer Δ_n. Pour cela, on pose $X_i^0 = X_i^\beta \sim \mathcal{E}(1)$. Pour x fixé et pour $\beta \gg \ln x$, on utilise l'approximation $x^{1/\beta} \simeq 1 + \ln x/\beta$. Cette approximation est aussi valable pour $\widehat{\beta}_n$, l'estimateur du maximum de vraisemblance (evm) de β, puisque $\widehat{\beta}_n \xrightarrow{\mathcal{L}} \beta$ lorsque $n \to \infty$. Il s'ensuit que :

- $\widehat{q}_{\text{param},n} = (-\ln p_n)^{1/\widehat{\beta}_n} \simeq 1 + \ln q^0_{p_n}/\widehat{\beta}_n$, où $q^0_{p_n} = -\ln p_n$;
- $\widehat{\sigma}_n = (1/m_n)\sum_{j=1}^{m_n} X_{(n-m_n+j)} - X_{(n-m_n)} \simeq \widehat{\sigma}^0_n/\beta$, où $\widehat{\sigma}^0_n = (1/m_n)\sum_{j=1}^{m_n} \ln(X^0_{(n-m_n+j)}) - \ln(X^0_{(n-m_n)})$;
- $\widehat{q}_{ET,n} = \widehat{u}_n + \widehat{\sigma}_n \ln(m_n/np_n) \simeq 1 + \widehat{q}^0_{ET,n}/\beta$, où $\widehat{q}^0_{ET,n} = \ln(\widehat{u}^0_n) + \widehat{\sigma}^0_n \ln(m_n/np_n)$, avec $\widehat{u}^0_n = X^0_{(n-m_n)}$;
- $d_n = (-\ln(m_n/n))^{1/\widehat{\beta}_n}\ (1 - \ln p_n/\ln(m_n/n))^2\ (1-\widehat{\beta}_n)/2\widehat{\beta}_n^2$
  $\simeq -(2\widehat{\beta}_n)^{-1}\ (1 - \ln p_n/\ln(m_n/n))^2$, l'expression de d_n étant déduite du tableau B.2.

Ceci signifie que

$$\Delta_n \simeq \frac{1}{\beta}\left[\widehat{q}^0_{ET,n}\frac{\beta}{\widehat{\beta}_n} - \ln(q^0_{1-p_n}) - \frac{1}{2}\frac{\beta}{\widehat{\beta}_n}(\eta_n - \eta'_n)^2 + \sigma^0_n\right],$$

où $\widehat{q}^0_{ET,n}$, $q^0_{1-p_n}$ et σ^0_n sont indépendants de β.

On sait que $\widehat{\beta}_n$ satisfait l'équation du maximum de vraisemblance qui, lorsque $\alpha = 1$, s'exprime (cf. [17])

$$\frac{1}{\widehat{\beta}_n} = \frac{1}{n}\sum_{i=1}^{n} T_i - \frac{\sum_{i=1}^{n} T_i \exp(-T_i\widehat{\beta}_n)}{\sum_{i=1}^{n}\exp(-T_i\widehat{\beta}_n)},$$

avec $T_i = -\ln X_i = -(1/\beta)\ln X_i^0$. Si on pose $S_i = \ln X_i^0$, cette équation devient

$$\frac{1}{\widehat{\beta}_n} = -\frac{1}{n\beta}\sum_{i=1}^{n} S_i + \frac{1}{\beta}\frac{\sum_{i=1}^{n} S_i \exp(-S_i\widehat{\beta}_n/\beta)}{\sum_{i=1}^{n}\exp(-S_i\widehat{\beta}_n/\beta)}.$$

Soit $\widehat{\gamma} = \widehat{\beta}_n/\beta$. Alors, $\widehat{\gamma}$ est solution de l'équation

$$\widehat{\gamma} = -\frac{1}{n}\sum_{i=1}^{n} S_i + \frac{\sum_{i=1}^{n} S_i \exp(-S_i\widehat{\gamma})}{\sum_{i=1}^{n}\exp(-S_i\widehat{\gamma})},$$

et est donc indépendant de β. Ceci montre que pour β assez grand, on peut considérer que le signe de Δ_n est indépendant de la valeur de β, qui est toujours positive.
■

C.2 Les deux versions du test ET-BP basées sur le bootstrap paramétrique

Scholie 3 *On teste l'adéquation de la famille de lois de fonctions de répartition $F_{v,w}$ à laquelle appartient la vraie loi des données. Le résultat du test ET version 3 (bootstrap paramétrique complet), ainsi que celui du test ET version 2 (bootstrap paramétrique simplifié), est approximativement indépendant du paramètre de position (le paramètre v) et du paramètre d'échelle (le paramètre w), pour les lois normale, lognormale, exponentielle, gamma et Weibull.*

Remarque C.3 *Comme précédemment, le niveau des deux version du test ET-BP étant, pour m_n fixé, le même quelle que soit la valeur des paramètres de position et d'échelle, les valeurs de m_n à utiliser pour obtenir un niveau adéquat des deux versions du test ET-BP sont donc indépendantes de ces paramètres de position et d'échelle.*

Justification : On s'intéresse tout d'abord à l'intervalle de confiance de la version 3 du test, $IC_{\delta,BP} = [\widehat{\delta}^*_{\min,n}, \widehat{\delta}^*_{\max,n}]$. On veut montrer que, quelles que soient les valeurs des paramètres v et w, la différence entre la borne inférieure (resp. supérieure) de l'intervalle $IC_{\delta,BP}$ et $\widehat{\delta}_n$ est toujours du même signe. Cette différence s'exprime comme

$$\Delta_1 = \widehat{\delta}^*_{\min,n} - \widehat{\delta}_n = \widehat{q}^{*(j)}_{ET,n} - \widehat{q}^{*(j)}_{\text{param},n} - (\widehat{q}_{ET,n} - \widehat{q}_{\text{param},n}).$$

On va donc réexprimer en fonction de v et w les quantités qui permettent de calculer cette différence. Pour cela, on pose $X_i^0 = (X_i - v)/w$ et $X_i^{*(j)0} = (X_i^{*(j)} - \widehat{v}_n)/\widehat{w}_n$, où $\widehat{v}_n$ et $\widehat{w}_n$ sont les evm de v et w (resp.) calculés à partir de l'échantillon initial. De même que pour la version 1 du test,

- $\widehat{\sigma}_n = w\widehat{\sigma}^0_n$, où $\widehat{\sigma}^0_n = (1/m_n)\sum_{j=1}^{m_n} X^0_{(n-m_n+j)} - X^0_{(n-m_n)}$;
- $\widehat{\sigma}^{*(j)}_n = \widehat{w}_n\widehat{\sigma}^{*(j)0}_n$, où $\widehat{\sigma}^{*(j)0}_n = (1/m_n)\sum_{j=1}^{m_n} X^{*(j)0}_{(n-m_n+j)} - X^{*(j)0}_{(n-m_n)}$;
- $\widehat{q}_{ET,n} = v + w\widehat{q}^0_{ET,n}$, où $\widehat{q}^0_{ET,n}$ est indépendant de v et w ;
- $\widehat{q}^{*(j)}_{ET,n} = \widehat{v}_n + \widehat{w}_n\widehat{q}^{*(j)0}_{ET,n}$, où $\widehat{q}^{*(j)0}_{ET,n} = \widehat{u}^{*(j)0}_n + \widehat{\sigma}^{*(j)0}_n \ln(m_n/np_n)$, avec $\widehat{u}^{*(j)0}_n = X^{*(j)0}_{(n-m_n)}$.

D'autre part, comme précédemment, on a $F^{-1}_{v,w}(x) = v + wF^{-1}_{0,1}(x)$, $(H^{-1}_{v,w})'(x) = w(H^{-1}_{0,1})'(x)$ et $(H^{-1}_{v,w})''(x) = w(H^{-1}_{0,1})''(x)$. On note $\widehat{v}^{*(j)}_n$ et $\widehat{w}^{*(j)}_n$ les evm de $\widehat{v}_n$ et $\widehat{w}_n$ (resp.) calculé à partir du j-ème échantillon bootstrap simulé selon la loi de Fdr $F_{\widehat{v}_n,\widehat{w}_n}$. Il s'ensuit que

- $\widehat{q}_{\text{param},n} = F^{-1}_{\widehat{v}_n,\widehat{w}_n}(1-p_n) = \widehat{v}_n + \widehat{w}_n q^0_{1-p_n}$, où $q^0_{1-p_n} = F^{-1}_{0,1}(1-p_n)$;
- $\widehat{q}^{*(j)}_{\text{param},n} = F^{-1}_{\widehat{v}^{*(j)}_n,\widehat{w}^{*(j)}_n}(1-p_n) = \widehat{v}^{*(j)}_n + \widehat{w}^{*(j)}_n q^0_{1-p_n}$.

Ceci permet de montrer que

$$\Delta_1 = w\left(\frac{\widehat{v}_n - v}{w} + \frac{\widehat{w}_n}{w}\widehat{q}^{*(j)0}_{ET,n} - \widehat{q}^0_{ET,n} - \frac{\widehat{v}^{*(j)}_n - \widehat{v}_n}{\widehat{w}_n}\,\frac{\widehat{w}_n}{w} - \frac{\widehat{w}^{*(j)}_n - \widehat{w}_n}{\widehat{w}_n}\,\frac{\widehat{w}_n}{w}q^0_{1-p_n}\right),$$

où $\widehat{q}^{*(j)0}_{ET,n}$, $\widehat{q}^0_{ET,n}$ et $q^0_{1-p_n}$ sont indépendants de v et w.

Comme précédemment, on remarque, par exemple pour la loi normale que

$$\frac{\widehat{v}_n - v}{w} = \frac{1}{n}\sum_{i=1}^{n} X^0_i\ ,\quad \frac{\widehat{w}_n}{w} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X^0_i - \frac{\widehat{v}_n - v}{w}\right)^2}\ ,$$

$$\frac{\widehat{v}^{*(j)}_n - \widehat{v}_n}{\widehat{w}_n} = \frac{1}{n}\sum_{i=1}^{n} X^{*(j)0}_i \quad \text{et} \quad \frac{\widehat{w}^{*(j)}_n}{\widehat{w}_n} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X^{*(j)0}_i - \frac{\widehat{v}^{*(j)}_n - \widehat{v}_n}{\widehat{w}_n}\right)^2}.$$

Puisque $X_i^0 \sim F_{0,1}$ et que $X_i^{*(j)0} \sim F_{0,1}$, ces quantités sont elles aussi indépendantes de v et w. Ceci montre que Δ_1 est indépendant de v, et que le signe de Δ_1 est indépendant de la valeur de w, qui est toujours positive pour le cas de la loi normale (et la loi lognormale qui s'y ramène). De même, on démontre l'indépendance de Δ_1 par rapport à la valeur de v pour les lois exponentielle, gamma et Weibull. On en déduit que la proposition 3 (page 175) est vérifiée pour le test ET-BP complet (version 3).

On s'intéresse maintenant à l'intervalle de confiance de la version 2 du test $IC_{ET,BP,n} = \left[\widehat{q}^*_{ET,\min,n}, \widehat{q}^*_{ET,\max,n}\right]$. On veut montrer que, quelles que soient les valeurs des paramètres a et b, la différence entre la borne inférieure (resp. supérieure) de l'intervalle $IC_{ET,BP}$ et $\widehat{q}_{ET,n}$ est toujours du même signe. Cette différence s'exprime comme $\Delta_2 = \widehat{q}^{*(j)}_{ET,n} - \widehat{q}_{ET,n}$. Or, on a déjà montré que $\widehat{q}^{*(j)}_{ET,n} = \widehat{v}_n + \widehat{w}_n \widehat{q}^{*(j)0}_{ET,n}$ et $\widehat{q}_{ET,n} = v + w\widehat{q}^0_{ET,n}$. En utilisant les expressions de l'estimateur du maximum de vraisemblance de w pour les différentes lois envisagées, on en déduit que le signe de $\Delta_2 = w\left[(\widehat{w}_n/b)\widehat{q}^{*(j)0}_{ET,n} - \widehat{q}^0_{ET,n}\right]$ est indépendant de la valeur de w qui est toujours positif pour les lois normale, lognormale, exponentielle, gamma et Weibull.
■

Scholie 4 *On suppose que l'on teste l'adéquation de la famille des lois de Weibull $\mathcal{W}(\eta, \beta)$ à laquelle appartient la vraie loi des données. Alors, pour β grand devant 1 ($\beta \gg 1$), le résultat du test ET version 3 (bootstrap paramétrique complet), ainsi que celui du test ET version 2 (bootstrap paramétrique simplifié), peuvent être considérés comme indépendants du paramètre de forme β.*

Remarque C.4 *Comme précédemment, pour m_n fixé, le niveau des deux versions du test ET-BP est approximativement le même quelle que soit la valeur du paramètre de forme de la loi de Weibull, à condition que celui-ci soit grand devant 1. Les valeurs de m_n à utiliser pour obtenir un niveau adéquat des deux versions du test ET-BP sont donc approximativement indépendantes du paramètre de forme de la loi de Weibull, à condition que celui-ci soit grand devant 1.*

Justification : Pour simplifier, on suppose que l'échantillon des X_i est de loi de Weibull de paramètre d'échelle $\eta = 1$ et de paramètre de forme β.

On s'intéresse tout d'abord à l'intervalle de confiance de la version 3 du test. La différence entre la borne inférieure (resp. supérieure) de l'intervalle $IC_{\delta,BP}$ et $\widehat{\delta}_n$ est

$$\Delta_1 = \widehat{\delta}^*_{\min,n} - \widehat{\delta}_n = \widehat{q}^{*(j)}_{ET,n} - \widehat{q}^{*(j)}_{\text{param},n} - (\widehat{q}_{ET,n} - \widehat{q}_{\text{param},n}).$$

Posons $X_i^0 = X_i^\beta \sim \mathcal{E}(1)$. On note $\widehat{\beta}_n$ l'evm de β calculé à partir de l'échantillon initial, et $\widehat{\beta}_n^{*(j)}$ l'evm de $\widehat{\beta}_n$ calculé à partir du j-ème échantillon bootstrap simulé selon la loi $\mathcal{W}(1, \widehat{\beta}_n)$. On a, comme précédemment :

- $\widehat{\sigma}_n \simeq \widehat{\sigma}_n^0/\beta$, où $\widehat{\sigma}_n^0 = (1/m_n)\sum_{j=1}^{m_n} \ln(X^0_{(n-m_n+j)}) - \ln(X^0_{(n-m_n)})$;

- $\widehat{\sigma}_n^{*(j)} = (1/m_n)\sum_{j=1}^{m_n} X^{*(j)}_{(n-m_n+j)} - X^{*(j)}_{(n-m_n)} \simeq \widehat{\sigma}_n^{*(j)0}/\widehat{\beta}_n$, où
  $\widehat{\sigma}_n^{*(j)0} = (1/m_n)\sum_{j=1}^{m_n} \ln(X^{*(j)0}_{(n-m_n+j)}) - \ln(X^{*(j)0}_{(n-m_n)})$;
- $\widehat{q}_{ET,n} \simeq 1 + \widehat{q}^{0}_{ET,n}/\beta$, où $\widehat{q}^{0}_{ET,n} = \ln(\widehat{u}_n^0) + \widehat{\sigma}_n^0 \ln(m_n/(np_n))$;
- $\widehat{q}^{*(j)}_{ET,n} = \widehat{u}_n^{*(j)} + \widehat{\sigma}_n^{*(j)} \ln(m_n/(np_n)) \simeq 1 + \widehat{q}^{*(j)0}_{ET,n}/\widehat{\beta}_n$,où
  $\widehat{q}^{*(j)0}_{ET,n} = \ln(\widehat{u}_n^{*(j)0}) + \widehat{\sigma}_n^{*(j)0} \ln(m_n/(np_n))$, avec $\widehat{u}_n^{*(j)0} = X^{*(j)0}_{(n-m_n)}$;
- $\widehat{q}_{\text{param},\, n} = (-\ln p)^{1/\widehat{\beta}_n} \simeq 1 + \ln q_p^0/\widehat{\beta}_n$, où $q_p^0 = -\ln p$;
- $\widehat{q}^{*(j)}_{\text{param},\, n} = (-\ln p)^{1/\widehat{\beta}_n^{*(j)}} \simeq 1 + \ln q_p^0/\widehat{\beta}_n^{*(j)}$, où $q_p^0 = -\ln p$.

D'où
$$\Delta_1 \simeq \frac{1}{\beta}\left[\widehat{q}^{*(j)0}_{ET,n} + \ln q_p^0 \frac{\beta}{\widehat{\beta}_n} - \widehat{q}^{0}_{ET,n} - \ln q_p^0 \frac{\beta}{\widehat{\beta}_n}\frac{\widehat{\beta}_n}{\widehat{\beta}_n^{*(j)}}\right],$$
où $\widehat{q}^{*(j)0}_{ET,n}$, q_p^0 et $\widehat{q}^{0}_{ET,n}$ sont indépendants de β. On a montré plus haut que si $\widehat{\beta}_n$ est l'estimateur du maximum de vraisemblance de β, alors le rapport $\beta/\widehat{\beta}_n$ est indépendant de β. Dans notre cas, il en résulte que $\widehat{\beta}_n/\widehat{\beta}_n^{*(j)}$ est indépendant de $\widehat{\beta}_n$. Ceci montre que pour β assez grand, le signe de Δ_1 est approximativement indépendant de la valeur de β. On en déduit que la proposition 4 (page 177) est vérifiée pour le test ET version 3. Le cas du test ET version 2 est analogue.

■

Annexe D

Démonstrations annexes sur l'estimation bayésienne de la loi GPD

D.1 Densité de la loi a posteriori

On sait que la densité de la loi a posteriori s'exprime comme

$$\pi(\theta|\,\underline{x}) = \frac{\prod_{i=1}^{n} f(x_i|\,\theta)\,\pi(\theta)}{f_\pi(\underline{x})}.$$

On remplace alors chacun des $f(x_i|\,\theta)$ par sa forme intégrale (formule 3.2), puis on échange les intégrales et le produit d'après le lemme de Fubini :

$$\pi(\theta|\,\underline{x}) = \frac{\int \ldots \int \left(\prod_{i=1}^{n} z_i\right) \exp\left(-\sum_{i=1}^{n} x_i z_i\right) \left(\prod_{i=1}^{n} g(z_i|\,\theta)\right)\, d\underline{z}\,\pi(\theta)}{\int \prod_{i=1}^{n} f(x_i|\,\theta')\,\pi(\theta')\,d\theta'},$$

où $d\underline{z} = \prod_{i=1}^{n} dz_i$. On reconnaît l'expression de $p(\underline{x}|\,\underline{z})$ que l'on remplace, et on effectue la même transformation au dénominateur :

$$\begin{aligned}
\pi(\theta|\,\underline{x}) &= \frac{\int p(\underline{x}|\,\underline{z}) \prod_{i=1}^{n} g(z_i|\,\theta)\, d\underline{z}\,\pi(\theta)}{\int p(\underline{x}|\,\underline{z}') \prod_{i=1}^{n} g(z_i'|\,\theta')\,\pi(\theta')\,d\theta'\,d\underline{z}'} \\
&= \frac{\int p(\underline{x}|\,\underline{z}) \prod_{i=1}^{n} g(z_i|\,\theta)\, d\underline{z}}{\int p(\underline{x}|\,\underline{z}')\, g_\pi(\underline{z}')\, d\underline{z}'}\,\pi(\theta) \\
&= \int \frac{p(\underline{x}|\,\underline{z})\, g_\pi(\underline{z})}{\int p(\underline{x}|\,\underline{z}')\, g_\pi(\underline{z}')\, d\underline{z}'}\ \frac{\prod_{i=1}^{n} g(z_i|\,\theta)}{g_\pi(\underline{z})}\ d\underline{z}\,\pi(\theta) \\
&= \int \frac{q(\underline{x}|\,\underline{z}) \prod_{i=1}^{n} g(z_i|\,\theta)\,\pi(\theta)}{g_\pi(\underline{z})}\, d\underline{z} \\
&= \int q(\underline{x}|\,\underline{z})\pi(\theta|\,\underline{z})\, d\underline{z}
\end{aligned}$$

D.2 Fonctions gamma, digamma et $A_{c,d}$

Nous souhaitons ici démontrer les propriétés de l'approximation normale de la loi gamconII. Ces propriétés dépendent de la fonction $A_{c,d}$, définie par l'équation (3.4) page 104, et de ses dérivées, elles-mêmes définies à partir des fonctions gamma, et digamma. Nous commençons donc cette annexe (paragraphe D.2.1) par un rappel des propriétés des fonctions gamma, notée Γ, et digamma, notée ψ. Au paragraphe D.2.2, nous montrons l'existance et l'unicité du mode de la densité de la loi gamconII. Puis au paragraphe D.2.3, nous produisons un encadrement de ce mode, qui est aussi le mode de la loi normale approximante. Cela permet le calcul numérique de ce mode par la méthode de dichotomie. Dans le paragraphe D.2.4, nous donnons aussi un encadrement de la variance ${\sigma_d^*}^2$ de la loi normale approximant la loi gamconII. Enfin, le paragraphe D.2.5 sur le contrôle du reste du développement de Taylor de la fonction $A_{c,d}$ nous permet de justifier l'approximation normale, et de donner une plage de validité de cette approximation.

D.2.1 Fonctions gamma et digamma

La fonction gamma $\Gamma(a)$ est le facteur de normalisation de la densité de la loi $\mathcal{G}amma(a, 1)$. Elle est définie pour $t > 0$ par l'intégrale

$$\Gamma(t) = \int_0^\infty e^{-x} x^{t-1}\, dx\,.$$

On démontre par intégration par parties que pour tout $t > 0$

$$\Gamma(t+1) = t\,\Gamma(t)\,. \tag{D.1}$$

On peut montrer que la fonction Γ est strictement convexe, tend vers l'infini quand $t \to 0$ par valeurs positives et $t \to +\infty$, qu'elle atteint son minimum entre 1 et 2, que $\Gamma(1) = \Gamma(2) = 1$ et que $\Gamma(0.5) = \sqrt{\pi}$. Selon la formule de Stirling, on a

$$\Gamma(t+1) \sim \left(\frac{t}{e}\right)^t \sqrt{2\pi\, t} \quad \text{quand} \;\; t \to +\infty\,.$$

La fonction digamma est la dérivée du logarithme de la fonction Γ :

$$\psi(t) = \frac{d}{dt} \ln \Gamma(t) = \frac{\Gamma'(t)}{\Gamma(t)}\,.$$

Comme la fonction gamma, la fonction digamma n'a pas de formule analytique. Des algorithmes de calcul sont cependant accessibles (voir par exemple [38, 42, 47]).

De la propriété (D.1), on déduit pour le logarithme de la fonction Γ la relation fonctionnelle $\ln \Gamma(t+1) = \ln \Gamma(t) + \ln t$ pour tout $t > 0$, donc, pour la fonction digamma,

$$\psi(t+1) = \psi(t) + \frac{1}{t} \quad \text{pour tout} \;\; t > 0\,,$$

et enfin, pour sa dérivée,

$$\psi'(t+1) = \psi'(t) - \frac{1}{t^2} \quad \text{pour tout } \ t > 0\,.$$

D.2.2 Existence et unicité du mode de la densité de la loi gamconII

On veut montrer que la fonction $A_{c,\,d}$ définie par l'équation (3.4) page 104 possède un unique maximum sur son domaine de définition $]0,+\infty[$. Remarquons tout d'abord que $\lim_{y\to 0} A_{c,\,d} = -\infty$ (puisque le terme dominant est alors $\ln y$) et que $\lim_{y\to+\infty} A_{c,\,d} = -\infty$ (puisque $A_{c,\,d}$ est, à une constante près, le logarithme de la densité de la loi gamconII, qui tend vers zéro à l'infini). On en déduit que $A_{c,\,d}$ possède au moins un mode, et on veut montrer que ce mode est unique. Il suffit pour cela de montrer que sa dérivée seconde est toujours strictement négative, c'est-à-dire que $A_{c,\,d}$ est strictement concave. On considère la fonction

$$A''_{c,\,d}(y) = d\psi'(yd+1) - \psi'(y)\ .$$

On utilise la représentation suivante de $\psi'(x)$, déduite des relations fonctionnelles du paragraphe précédent :

$$\psi'(x) = \sum_{i=0}^{\infty} \frac{1}{(x+i)^2}\,. \tag{D.2}$$

Comparons tout d'abord cette série avec l'intégrale

$$\int_0^\infty \frac{dt}{(x+t)^2} = \frac{1}{x}\,.$$

Puisque pour tout $x > 0$ et tout entier $i \geq 0$

$$\frac{1}{(x+i)^2} > \frac{1}{(x+t)^2} \ \text{pour tout } i < t \leq i+1\,,$$

il s'ensuit que

$$\psi'(x) > \int_0^\infty \frac{dt}{(x+t)^2} = \frac{1}{x}\,. \tag{D.3}$$

De même, on montre que pour tout $n > 0$,

$$n\psi'(nx+1) = n\sum_{i=1}^{\infty} \frac{1}{(nx+i)^2} < \int_0^\infty \frac{n\,dt}{(nx+t)^2} = \frac{1}{x}\,,$$

puisque pour tout $x > 0$ et tout entier $i \geq 1$

$$\frac{1}{(x+i)^2} < \frac{1}{(x+t)^2} \ \text{pour tout } i-1 \leq t < i\ . \tag{D.4}$$

On a donc montré que pour tout $x > 0$,

$$n\psi'(nx+1) < \frac{1}{x} < \psi'(x)\,.$$

Il en découle que pour tout $x > 0$, $A''_{c,d}(x)$ est strictement négative, donc que $A_{c,d}$ est concave et possède un seul maximum, c'est-à-dire que $A'_{c,d}$ s'annule une seule fois.

D.2.3 Encadrement du mode de la densité de la loi gamconII

Nous cherchons ici à déterminer un encadrement du mode y^* de la densité de la loi gamconII, notamment en vue d'initialiser la procédure de dichotomie qui permet de calculer y^*. Remarquons tout d'abord que pour tout $y > 0$ et pour tout $d \geq 1$,

$$\psi(yd+1) - \psi(y) = \int_y^{yd+1} \psi'(t)\,dt. \tag{D.5}$$

D'autre part, pour ψ', d'après l'équation (D.3) page 181, on a la minoration $\psi'(x) > 1/x$, $\forall x > 0$. Pour majorer ψ', remarquons à présent que, d'après les équations (D.2) et (D.4) page 181, on a la majoration

$$\psi'(x) = \frac{1}{x^2} + \sum_{i=1}^{\infty} \frac{1}{(x+i)^2} < \frac{1}{x^2} + \int_0^{\infty} \frac{dt}{(x+t)^2} = \frac{1}{x} + \frac{1}{x^2}\,.$$

Par conséquent, on a l'encadrement

$$\frac{1}{x} < \psi'(x) < \frac{1}{x} + \frac{1}{x^2}\,. \tag{D.6}$$

On utilise cet encadrement dans l'équation (D.5), cela donne, pour tout $y > 0$ et pour tout $d \geq 1$,

$$\ln\left(1 + \frac{1}{yd}\right) < \psi(yd+1) - \psi(y) - \ln d < \ln\left(1 + \frac{1}{yd}\right) + \frac{1}{y} - \frac{1}{yd+1}\,.$$

Puisque pour tout $u \geq 0$, $\ln(1+u) \leq u$, la borne supérieure conduit, pour le mode y^*, à

$$0 = \psi(y^*d+1) - \psi(y^*) - \ln d - \ln c < \frac{1}{y^*d} + \frac{1}{y^*} - \frac{1}{y^*d+1} - \ln c < \left(\frac{1}{d} + 1\right)\frac{1}{y^*} - \ln c\,.$$

Il s'ensuit que

$$y^* < \frac{1 + 1/d}{\ln c} \leq \frac{2}{\ln c}\,. \tag{D.7}$$

Pour trouver une borne inférieure de y^*, il nous faut à présent trouver un minorant plus précis pour $\psi'(x)$:

$$\psi'(x) = \frac{1}{x^2} + \sum_{i=1}^{\infty} \frac{1}{(x+i)^2} > \frac{1}{x^2} + \int_1^{\infty} \frac{dt}{(x+t)^2} = \frac{1}{x+1} + \frac{1}{x^2}\,, \tag{D.8}$$

pour tout $x > 0$. On en déduit que pour tout $y > 0$ et tout $d \geq 1$,

$$\psi(yd+1) - \psi(y) > \frac{1}{y} - \frac{1}{yd+1} + \ln\left(\frac{yd+2}{y+1}\right) ,$$

et donc que pour tout $y > 0$, tout $c > 1$ et tout $d \geq 1$,

$$0 = \psi(y^*d+1) - \psi(y^*) - \ln d - \ln c > \frac{1}{y^*} - \frac{1}{y^*d+1} + \ln\left(\frac{y^*d+2}{dy^*+d}\right) - \ln c\,.$$

Il en découle que pour tout $y > 0$, tout $c > 1$ et tout $d \geq 1$,

$$\ln c > \frac{1}{y^*}\,\frac{(d-1)y^*+1}{y^*d+1} + \ln\left(\frac{y^*d+2}{dy^*+d}\right) .$$

Étudions tout d'abord la fonction $h_1(y) = [(d-1)y+1]/[yd+1]$, pour tout $y > 0$. On montre aisément que sa dérivée est négative : $h_1'(y) = -1/(yd+1)$. La fonction $h_1(y)$ est donc décroissante de $h_1(0) = 1$ à $\lim_{y\to\infty} h_1(y) = (d-1)/d = 1 - 1/d$. On en déduit que pour tout $y > 0$, tout $c > 1$ et tout $d \geq 1$,

$$\ln c > \frac{1}{y^*}\left(1 - \frac{1}{d}\right) + \ln\left(\frac{y^*d+2}{dy^*+d}\right) .$$

À présent, étudions la fonction $h_2(y) = (dy+2)/(yd+d)$, pour tout $y > 0$. La dérivée de cette fonction, $h_2'(y) = (d-2)/d(y+1)^2$, est positive si $d > 2$ et négative sinon. Les limites de cette fonction sont $h_2(0) = 2/d$ (c'est la borne inférieure si $d > 2$) et $\lim_{y\to\infty} h_2(y) = 1$ (c'est la borne inférieure si $1 \leq d \leq 2$). On en déduit que

$$\ln c > \begin{cases} \left(1 - \frac{1}{d}\right)\frac{1}{y^*} & \text{si } 1 \leq d \leq 2. \\ \left(1 - \frac{1}{d}\right)\frac{1}{y^*} - \ln\frac{d}{2} & \text{sinon,} \end{cases}$$

Pour tout $y > 0$, tout $c > 1$ et tout $d \geq 2$, on obtient donc l'encadrement

$$\frac{(1-1/d)}{\ln c + \ln(d/2)} < y^* < \frac{2}{\ln c}\,.$$

Nous souhaitons en particulier utiliser cet encadrement pour initialiser la dichotomie qui, au paragraphe 3.1.4 page 103, permet de calculer le mode y^* de la loi gamconII. Il nous a semblé que puique la borne inférieure dépend de d, l'intervalle permettant de calculer y^* par dichotomie pouvait être assez grand pour ralentir les calculs. Après des expérimentations numériques préliminaires, nous avons utilisé la borne inférieure $1/(4\ln c)$, pour des valeurs modérés de n et donc de $d = \delta' = \delta + n$.

D.2.4 Encadrement de la variance ${\sigma_d^*}^2$ de la loi normale approximant la loi gamconII

On souhaite préciser le comportement asymptotique de ${\sigma_d^*}^2 = 1/(-dA''_{c,d}(y^*))$. Pour cela, il nous faut étudier le comportement asymptotique de la fonction $A''_{c,d}$. On utilise l'encadrement de la fonction ψ' donné par l'équation (D.6) page 182 pour produire une borne inférieure de $A''_{c,d}$, ainsi que la borne inférieure pour ψ' donnée par l'équation (D.8) page 182 pour produire une borne supérieure de $A''_{c,d}$:

$$\frac{d}{dy+1} - \frac{1}{y} - \frac{1}{y^2} \leq A''_{c,d}(y) \leq \frac{d}{dy+1} + \frac{d}{(dy+1)^2} - \frac{1}{y+1} - \frac{1}{y^2}. \tag{D.9}$$

L'encadrement de y^* donné par l'équation (D.7) page 182 implique l'encadrement suivant : $A_{2,\inf}(d) \leq A''_{c,d}(y^*) \leq A_{2,\sup}(d)$ où

$$A_{2,\inf}(d) = \frac{d\ln c}{2d+\ln c} - \frac{\ln c + \ln(d/2)}{1-1/d} - \frac{(\ln c + \ln(d/2))^2}{(1-1/d)^2},$$

et

$$A_{2,\sup}(d) = \frac{d\ln c}{2d+\ln c} + \frac{d(\ln c)^2}{(2d+\ln c)^2} - \frac{\ln c + \ln(d/2)}{1-1/d+\ln c+\ln(d/2)} - \frac{(\ln c + \ln(d/2))^2}{(1-1/d)^2}.$$

Or lorsque $d \to \infty$, on a $A_{2,\inf}(d) \sim -(\ln d)^2$ et $A_{2,\sup}(d) \sim -(\ln d)^2$. Ceci montre que $A''_{c,d}(y^*) \sim -(\ln d)^2$, c'est-à-dire que $\sigma_d^{*2} \sim d^{-1}(\ln d)^{-2}$.

Remarque D.1 *L'approximation normale est utilisée au cours de l'algorithme de Gibbs pour simuler selon la loi a posteriori de $\alpha^{(m+1)}$ qui est une loi gamconII(η'/μ',δ'). Dans ce cas, $d = \delta' = \delta + n \to \infty$ lorsque $n \to \infty$, d'où les équivalents pour n donc d tendant vers l'infini. D'autre part, à l'étape m, le paramètre c correspond à $c'_n = \eta'/\mu'$ qui dépend des $z_i^{(m)}$ et de n puisque*

$$\eta' = \frac{\delta\eta + \sum_{i=1}^n z_i^{(m)}}{\delta+n} \quad \textit{et} \quad \mu' = \mu^{\delta/(\delta+n)} \left(\prod_{i=1}^n z_i^{(m)}\right)^{1/(\delta+n)}.$$

Toujours pour m fixé, lorsque $n \to \infty$, on a

$$\eta' = \left(\frac{\delta\eta}{n} + \frac{1}{n}\sum_{i=1}^n z_i^{(m)}\right)\frac{n}{\delta+n} \underset{n\to\infty}{\sim} \frac{1}{n}\sum_{i=1}^n z_i^{(m)} \underset{n\to\infty}{\rightarrow} \frac{\alpha^{(m)}+1}{\beta^{(m)}+x_i},$$

la moyenne de la loi des $z_i^{(m)}$, qui est une constante, à m fixé. De même, on montre que lorsque $n \to \infty$, μ' est équivalent à la moyenne géométrique des $z_i^{(m)}$ qui tend aussi vers une quantité ne dépendant que de m lorsque $n \to \infty$. Il s'ensuit que, à m fixé c'est-à-dire pour chaque étape de l'algorithme de Gibbs, le paramètre $c = c'_n$ tend vers une constante lorsque $n \to \infty$. Ceci justifie que l'on néglige les variations de ce paramètre devant celles de $d = \delta' = \delta + n$.

D.2.5 Contrôle du reste du développement de Taylor de la fonction $A_{c,d}$

On souhaite à présent montrer que le reste du développement de Taylor de la fonction $A_{c,d}$ est négligeable par rapport au dernier terme de la partie principale de ce développement. Étudions tout d'abord le comportement asymptotique, lorsque $d \to \infty$ du dernier terme de la partie principale du développement de Taylor de la fonction $A_{c,d}$:

$$A_{\mathrm{dtpp}}(h,d) = \frac{h^2}{2} A''_{c,d}(y^*) \underset{d\to\infty}{\sim} -(h \ln d)^2 \,, \tag{D.10}$$

d'après l'étude de la fonction $A''_{c,d}(y^*)$ par rapport à d menée dans l'annexe D.2.4 page 184.

Il nous faut ensuite explorer le comportement asymptotique, lorsque $d \to \infty$, du reste du développement de Taylor de $A_{c,d}$:

$$A_{\mathrm{r}}(h,d) = \frac{h^3}{6} A'''_{c,d}(y_h) \,.$$

On souhaite d'abord préciser le comportement asymptotique de la fonction

$$A'''_{c,d}(y) = d^2 \psi''(yd+1) - \psi''(y) \,.$$

Afin d'encadrer $A'''_{c,d}$, on cherche tout d'abord à encadrer ψ''. Par dérivation de l'équation (D.2) page 181, on obtient que

$$\psi''(x) = -2 \sum_{i=0}^{\infty} \frac{1}{(x+i)^3} \,.$$

Puisque pour tout $x > 0$ et tout entier $i \geq 0$,

$$\frac{1}{(x+i)^3} > \frac{1}{(x+t)^3} \quad \text{pour tout } t \text{ tel que } i \leq t \leq i+1 \,,$$

on a la borne supérieure pour ψ''

$$\psi''(x) < -2 \int_0^{\infty} \frac{dt}{(x+t)^3} = -\frac{1}{x^2} \,. \tag{D.11}$$

D'autre part,

$$d^2 \psi''(yd+1) = -2d^2 \sum_{i=1}^{\infty} \frac{1}{(yd+i)^3} > \int_0^{\infty} \frac{-2d^2\,dt}{(yd+t)^3} = -\frac{1}{y^2} \,.$$

On a donc

$$\psi''(y) < -\frac{1}{y^2} < d^2 \psi''(yd+1) \,,$$

ce qui implique que $A'''_{c,d}(y) = d^2\psi''(yd+1) - \psi''(y) > 0$ pour tout $y > 0$.

On souhaite à présent trouver une borne supérieure pour $A'''_{c,d}$. Pour cela, on construit un encadrement de ψ''. Pour minorer ψ'', remarquons que pour tout $x > 0$,

$$\sum_{i=0}^{\infty} \frac{1}{(x+i)^3} = \frac{1}{x^3} + \sum_{i=1}^{\infty} \frac{1}{(x+i)^3} < \frac{1}{x^3} + \int_0^{\infty} \frac{dt}{(x+t)^3} = \frac{1}{x^3} + \frac{1}{2x^2}\,.$$

La majoration de ψ'' étant donnée par l'équation (D.11), on en déduit l'encadrement suivant de ψ''

$$-\frac{1}{x^2} - \frac{2}{x^3} < \psi''(x) < -\frac{1}{x^2}\,.$$

D'où l'encadrement suivant pour $A'''_{c,d}$, $\forall d > 0$

$$0 < A'''_{c,d}(y) < \frac{-d^2}{(yd+1)^2} + \frac{1}{y^2} + \frac{2}{y^3}\,.$$

Dans l'expression du reste de Taylor $A_{\mathrm{r}}(h,d)$ apparaît la quantité $A'''_{c,d}(y_h)$ que l'on souhaite donc encadrer :

$$0 < A'''_{c,d}(y_h) < \frac{-d^2}{(dy_h+1)^2} + \frac{1}{y_h^2} + \frac{2}{y_h^3}\,, \tag{D.12}$$

où y_h est compris entre y^* et $y^* + h$.

Remarquons que puisque $A_{c,d}(y)$ et la loi gamconII ne sont définies que sur $[0,\infty[$, et que la loi normale est définie sur tout $I\!R$, l'approximation normale de la loi gamconII ne peut être adéquate que pour des intervalles bornés. On considère des intervalles du type $[y^* - C\sigma_d^*, y^* + C\sigma_d^*]$, où C est une constante positive, puisque pour de tels intervalles la masse correspondante de la loi normale $\mathcal{N}(y^*, {\sigma_d^*}^2)$ reste constante, égale à $2\Phi(C) - 1$ (où Φ est la Fdr. de la loi $\mathcal{N}(0,1)$), qui est arbitrairement proche de 1 si on prend $C > 0$ assez grand. Par exemple l'intervalle $[y^* - 1.96\sigma_d^*, y^* + 1.96\sigma_d^*]$ contient environ 95% de la loi normale $\mathcal{N}(y^*, {\sigma_d^*}^2)$ approximant la loi gamconII. De tels intervalles contiennent aussi la plus grande partie de la masse de la loi gamconII approximée par la loi normale $\mathcal{N}(y^*, {\sigma_d^*}^2)$, ce qu'on se propose de démontrer dans ce paragraphe.

On choisit h la forme $h(u) = C\sigma_d^* u$ où $u \in [-1,1]$, c'est-à-dire que $h(u) \in [-C\sigma_d^*, C\sigma_d^*]$ et $y_{h(u)} \in [y^* - C\sigma_d^*, y^* + C\sigma_d^*]$. L'encadrement de $A'''_{c,d}(y_h)$ donné par l'équation (D.12) devient alors : $0 < A'''_{c,d}(y_{h(u)}) < A_{3,\sup}(d)$ où

$$\begin{aligned} A_{3,\sup}(d) &= \frac{-d^2}{[d(y^* + C\sigma_d^*) + 1]^2} + \frac{1}{(y^* - C\sigma_d^*)^2} + \frac{2}{(y^* - C\sigma_d^*)^3} \\ &= \frac{-1}{\left(\dfrac{2}{\ln c} + C\sigma_d^* + \dfrac{1}{d}\right)^2} + \frac{1}{\left(\dfrac{1-1/d}{\ln c + \ln(\frac{d}{2})} - C\sigma_d^*\right)^2} + \frac{2}{\left(\dfrac{1-1/d}{\ln c + \ln(\frac{d}{2})} - C\sigma_d^*\right)^3}, \end{aligned}$$

d'après l'encadrement de y^* donné par l'équation (D.7) page 182. Comme lorsque $d \to \infty$, $\sigma_d^* \sim d^{-1/2}(\ln d)^{-1} \to 0$, on a l'équivalent asymptotique suivant pour $A_{3,\,\mathrm{sup}}(d)$:

$$A_{3,\,\mathrm{sup}}(d) \underset{d\to\infty}{\sim} 2(\ln d)^3\,.$$

Pour le reste de Taylor de $A_{c,\,d}(y)$, on en déduit l'encadrement

$$0 < |A_{\mathrm{r}}(h(u), d)| < \frac{C^3\sigma_d^{*3}}{3}(\ln d)^3 \underset{d\to\infty}{\sim} \frac{C^3}{3d^{3/2}}\,,$$

car $\sigma_d^* \sim d^{-1/2}(\ln d)^{-1}$ lorsque $d \to \infty$. On en déduit que $A_{\mathrm{r}}(h(u), d) = O(d^{-3/2})$, alors que, puisque $h = O(d^{-1/2}(\ln d)^{-1})$ et d'après l'équation (D.10) page 185, $A_{\mathrm{dtpp}}(h(u), d) \sim d^{-1}$ lorsque $d \to \infty$, ce qui montre bien que le reste du développement de Taylor est négligeable.

D.3 Description des méthodes classiques utilisées

Nous utilisons principalement, pour la simulation de la loi a posteriori, des méthodes de Monte Carlo par chaînes de Markov (ou méthodes MCMC) [45, 46]. En particulier, nous avons implémenté un algorithme de Gibbs avec une étape de Hastings-Metropolis (ou algorithme de Gibbs hybride) dont nous rappelons le principe. D'autre part, pour l'estimation bayésienne des paramètres de la loi GPD dans le cas du mode a posteriori, nous souhaitons estimer le mode de la densité des échantillons de valeurs de α et β obtenues au cours de l'algorithme de Gibbs. Nous détaillons donc la technique utilisée.

D.3.1 Principe de l'algorithme de Gibbs

Le but de cet algorithme (voir le chapitre 5 de [45]) est de simuler selon la loi du couple (X, Y) connaissant seulement les lois conditionnelles de X sachant Y et de Y sachant X. On travaille avec les densités conditionnelles $f_1(x|y)$ et $f_2(y|x)$. En pratique, on forme la chaîne de Markov homogène $(x^{(m)}, y^{(m)})$:

$$\begin{array}{rcl} x^{(m+1)} & \sim & f_1(x|\,y^{(m)}) \\ y^{(m+1)} & \sim & f_2(y|\,x^{(m+1)}) \end{array}$$

On peut alors montrer (voir le théorème 5.1 de [45]) que la loi de densité jointe $f(x, y)$ est la loi invariante de la chaîne de Markov homogène $(x^{(m)}, y^{(m)})$. Donc lorsque la loi stationnaire est atteinte, les couples (x, y) obtenus sont des réalisations de la loi de densité $f(x, y)$.

Dans notre cas, on a $Y = z$ et $X = \theta = (\alpha, \beta)$, qui est lui aussi simulé grâce à la loi marginale de α (de densité $f_1(\alpha) = \pi(\alpha|\,z)$) et la loi conditionnelle de β sachant α (de densité $f_2(\beta|\,\alpha) = \pi(\beta|\,\alpha, z)$).

Dans un cadre plus général, on simule selon la loi de $(Y_1, \ldots, Y_k)$ en connaissant seulement les lois conditionnelles des Y_i sachant les Y_j $(j \neq i)$, en formant la chaîne de Markov homogène $(y_1^{(m)}, \ldots, y_k^{(m)})$:

`étape i de l'itération` $m+1$: $y_i^{(m+1)} \sim f_i(y_i |\, y_1^{(m+1)}, \ldots, y_{i-1}^{(m+1)}, y_{i+1}^{(m)}, \ldots, y_k^{(m)})$.

D.3.2 Méthode de Hastings-Metropolis pour la simulation

Le but de cet algorithme (voir le chapitre 4 de [45]) est de simuler selon la loi de densité f, connue à une constante de normalisation près. L'algorithme de Hastings-Metropolis repose sur l'utilisation d'une densité conditionnelle $q(y|\,x)$ qui doit être symétrique et aisément simulable. En pratique, on forme la chaîne de Markov homogène $x^{(m)}$:

1. générer $y \sim q(y|\, x^{(m)})$.
2. prendre

$$x^{(m+1)} = \begin{cases} y & \text{avec probabilité } \rho(x^{(m)}, y), \\ x^{(m)} & \text{avec probabilité } 1 - \rho(x^{(m)}, y), \end{cases} \quad \text{où} \quad \rho(x,y) = \min\Big(\frac{f(y)}{f(x)}\,\frac{q(x|\,y)}{q(y|\,x)}, 1\Big).$$

La loi q est appelée loi instrumentale. On peut montrer (voir le théorème 4.1 de [45]) que la loi de densité f est la loi invariante de la chaîne de Markov homogène $x^{(m)}$. Donc lorsque la loi stationnaire est atteinte, les X obtenus sont des réalisations de la loi de densité f.

Nous avons utilisé le cas particulier de l'algorithme de Hastings-Metropolis indépendant où la loi instrumentale $q(y)$ est indépendante de la valeur $x^{(m)}$ de la chaîne de Markov à l'intant m. Nous avons tout d'abord utilisé comme loi instrumentale $q(y)$ la loi normale approximant la vraie loi gamcomII (voir le paragraphe 3.1.4). Cependant, les queues de la loi normale sont à décroissance plutôt rapide, alors que l'on obtient des vitesses de convergence rapides lorsque la loi instrumentale q est à queue lourde. Nous avons donc choisi d'utiliser une distribution à queue lourde, une loi de Cauchy, en conservant le même mode (même position et même valeur de la fonction de répartition) que la vraie loi gamconII (et la loi normale approximante).

D.3.3 Algorithme de Gibbs avec une étape de Hastings-Metropolis

Les algorithmes MCMC hybrides (voir le chapitre 5.4 de [45]) apparaissent notamment lorsque certaines lois conditionnelles de l'algorithme de Gibbs ne sont pas simulables directement. Le compromis suggéré par Müller (voir le chapitre 5.4 de [45], ou pour plus de précisions [39, 40]) est de remplacer chaque étape `i` où une simulation directe selon la loi conditionnelle est impossible par une simulation suivant une loi instrumentale q_i. On introduit donc une étape de Hastings-Metropolis à l'intérieur de l'algorithme de Gibbs. L'étape i de l'itération $m+1$ devient alors :

`i.1` simuler $\tilde{y}_i \sim q_i(y|\, y_1^{(m+1)}, \ldots, y_i^{(m+1)}, y_{i+1}^{(m)}, \ldots, y_p^{(m)})$.

i.2 prendre

$$y_i^{(m+1)} = \begin{cases} y_i^{(m)} & \text{avec probabilité } \rho(\tilde{y}_i, y_i^{(m)}), \\ \tilde{y}_i & \text{avec probabilité } 1 - \rho(\tilde{y}_i, y_i^{(m)}), \end{cases}$$

où

$$\rho(x,y) = \min\left(\frac{f_i(x|\,Y_{-i}^{(m)})}{f_i(y|\,Y_{-i}^{(m)})}\,\frac{q_i(y|\,Y_i^{(m)})}{q_i(x|Y_i^{(m)})}, 1\right),$$

avec

$$\begin{cases} Y_{-i}^{(m)} = (y_1^{(m+1)}, \ldots, y_{i-1}^{(m+1)}, y_{i+1}^{(m)}, \ldots, y_k^{(m)}) \\ Y_i^{(m)} = (y_1^{(m+1)}, \ldots, y_i^{(m+1)}, y_{i+1}^{(m)}, \ldots, y_k^{(m)}) \end{cases}$$

Un point important de cette substitution (de la simulation directe, lorsqu'elle est impossible, par une étape de Hastings-Metropolis) est que l'étape de Hastings-Metropolis n'est utilisée qu'une seule fois lors d'une itération m de l'algorithme de Gibbs. On ne génère donc qu'une seule réalisation de $\tilde{y}_i$ au lieu de chercher à approcher la loi conditionnelle de y_i sachant y_j $(j \neq i)$ en répétant plusieurs fois les simulations selon q_i.

D.3.4 Contrôle de l'atteinte approximative de la loi stationnaire au cours de l'algorithme de Gibbs

Nous utilisons au cours de notre procédure bayésienne d'estimation des paramètres de la loi GPD un algorithme de Gibbs afin de simuler selon la loi a posteriori des paramètres. Cette loi a posteriori est la loi stationnaire de l'algorithme de Gibbs que nous utilisons. On peut considérer que les données simulées au cours de cet algorithme sont issues de la loi a posteriori lorsque la loi stationnaire est approximativement atteinte. Il nous faut donc pouvoir contrôler l'atteinte approximative de la loi stationnaire au cours de l'algorithme de Gibbs.

À notre connaissance, il n'existe pas de méthode numérique simple et rapide pour contrôler l'atteinte approximative de la stationnarité. Nous n'avons donc pas inclu de type de vérification au cours de nos simulations intensives pour ne pas trop augmenter le temps de calcul. Mais nous avons essayé de vérifier l'atteinte de la stationnarité dans quelques cas particuliers. Nous avons notamment utilisé des méthodes graphiques (voir la figure 3.2 page 112) simples et rapides, mais assez imprécises et liées à l'appréciation de l'observateur.

Dans un cas particulier, nous avons appliqué la méthode numérique de contrôle fondée sur le théorème central limite pour les chaînes de Markov développée par Dielbolt et Chauveau [11, 12]. La normalité est en effet une conséquence vérifiable du fait que la chaîne a suffisamment visité le support de la loi cible. Il s'agit alors de tester la normalité d'un échantillon de sommes de fonctions de la chaîne construit à partir de chaînes parallèles initialisées selon une loi suffisamment dispersée.

Nous avons choisi d'utiliser 50 chaînes parallèles dont les points initiaux sont tirés selon une loi uniforme sur un pavé borné et suffisamment étendu de $]0, +\infty[\times]0, +\infty[$. Nous traitons séparément les chaînes du paramètre α et celles du paramètre β puisqu'il est plus simple de contrôler les marginales, que de faire du test de normalité multidimentionnel. Evidemment, le contrôle des marginales n'implique pas que la chaîne de Markov multidimentionnelle (celle concernant le couple (α, β)) est arrivée à la stationnarité, mais cela nous donne une indication raisonnable. On considère la fonction $h(\theta) = \theta$, pour $\theta = \alpha$ et $\theta = \beta$, sur des pavés du support considéré pour α et β.

Pour le paramètre α, le pavé le plus rapide à atteindre la normalité l'atteint en 200 itérations. Pour le pavé le plus lent, il faut attendre 6400 itérations pour atteindre la normalité. Il s'agit de la queue de distribution de droite de la loi de α qui semble donc moins bien explorée par l'algorithme de Gibbs pour un faible nombre d'itérations. Dès $N = 500$ itérations du Gibbs 79.62% de la masse de la loi de α est controlée et approximativement gaussienne. Pour $N = 1000$ itérations 97.29% de la masse de cette loi est controlée.

Concernant le paramètre β, on atteint la normalité pour le pavé le plus rapide en 50 itérations, et pour le pavé le plus lent en 10000 itérations. À nouveau, c'est la queue de distribution de droite de la loi de β qui est la plus lente à atteindre la normalité, donc qui semble moins bien explorée par l'algorithme de Gibbs en un faible nombre d'itérations. Dès $N = 500$ itérations du Gibbs 89.47% de la masse de la loi de β est controlée et approximativement gaussienne. Pour $N = 1000$ itérations 93.04% de la masse de cette loi est controlée.

Ceci nous a semblé suffisant pour nos simulations intensives, le gain en précision lorsque l'on rajoute des itérations n'étant pas assez important pour l'augmentation du temps de calcul que cela occasionnerait. Cependant, lorsque l'on applique la méthode bayésienne sur des données réelles, il est préférable d'utiliser plus d'itérations pour s'assurer de bien atteindre la loi stationnaire, le temps de calcul sur un seul jeu de données restant alors raisonnable.

Lorsque nos simulations intensives ne donnent pas de très bons résultats (par exemple pour la loi loggamma(2)), on a supposé que cela pouvait être dû au fait que l'algorithme de Gibbs n'aie pas atteint sa loi stationnaire. Nous avons donc augmenté le nombre d'itérations de cet algorithme. Mais que l'on effectue $N = 1000$ ou $N = 5000$ itérations, les résultats (basés dans les deux cas sur les 500 desnières itérations) obtenus dans les deux cas sont similaires. Cela semble donc indiquer que l'algorithme de Gibbs a bien approximativement atteint sa loi stationnaire au bout d'environ 500 itérations.

D.3.5 Estimateur d'une densité et de son mode

L'un des estimateurs bayésiens des paramètres de la loi GPD que nous proposons est le mode de la densité a posteriori, de α et de β sachant α, obtenues lors de notre procédure bayésienne. Or nous ne connaissons pas de forme analytique pour ces densités a posteriori.

Nous ne disposons que d'échantillons de valeurs de α et β issues de ces densités a posteriori, de α et de β sachant α. Il nous faut donc, à partir de ces échantillons de valeurs, déterminer les modes des densités correspondantes (voir Vieu [48]).

Pour estimer le(s) mode(s) de la densité d'un échantillon de données, il nous faut tout d'abord disposer d'un estimateur lisse de cette densité. On estimera alors le(s) mode(s) par le(s) mode(s) de l'estimateur lisse de la densité. Nous avons choisi ici d'utiliser l'estimateur à noyau de la densité (voir le chapitre 4 de [9]) qui s'exprime, pour un échantillon $(x_1, \ldots, x_n)$, comme

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

où le noyau K est une densité de probabilité symétrique (par exemple la densité de la loi normale), et h est le paramètre de lissage.

Le problème consiste à présent à déterminer le paramètre de lissage optimal h_{opt}. On choisit pour cela d'utiliser la méthode du double noyau [7] pour laquelle le paramètre de lissage optimal est celui qui minimise la distance L^2 entre les estimateurs obtenus pour deux noyaux K et L. Nous avons choisi d'utiliser le noyau normal et le noyau d'Epanechnikov de densité $L(u) = (3/4)(1 - u^2)$ pour $u \in [-1,1]$. De plus, nous approchons la distance L^2 entre les deux estimateurs à noyau correspondants par la somme des erreurs quadratiques obtenues en chacun des points de l'échantillon dont nous souhaitons estimer la densité (et son mode).

Une fois que le paramètre de lissage optimal h_{opt} est déterminé, on obtient un estimateur $\hat{f}_n(x)$ de la densité calculé avec ce h_{opt}. Il ne reste plus qu'à estimer le mode de cet estimateur $\hat{f}_n(x)$ de la densité, à partir d'une grille de valeurs possibles. Nous utilisons comme grille de test l'ensemble des valeurs de l'échantillon dont nous souhaitons estimer le mode de la densité.

D.4 Méthode de calcul des erreurs

En partie centrale (de 0 au quantile d'ordre 0.9 de la vraie loi) on calcule pour chaque point y_i d'une grille régulière l'erreur relative entre la fonction de survie estimée et la vraie fonction de survie :

$$E_{1,i} = \left| \frac{F(y_i|\, \tilde{\alpha}, \tilde{\beta}) - F(y_i|\, \alpha_{\text{vrai}}, \beta_{\text{vrai}})}{1 - F(y_i|\, \alpha_{\text{vrai}}, \beta_{\text{vrai}})} \right|,$$

où $F(y|\, \alpha, \beta)$ est la fonction de répartition de la loi GPD de paramètres (α, β) ; $\tilde{\alpha}$ et $\tilde{\beta}$ sont les estimations de α et β obtenues par les estimateurs des moments pondérés, ou bien les estimateurs bayésiens par le mode, la moyenne ou la médiane ; et α_{vrai} et β_{vrai} sont les vraies valeurs des paramètres, celles à partir desquelles les différents échantillons ont été simulés. Pour la loi prédictive, on remplace $F(y_i|\, \tilde{\alpha}, \tilde{\beta})$ par la fonction de répartition de la loi prédictive évaluée au point y_i. Puis, on calcule la somme des carrés de ces quantités sur la grille de points :

$$E_1^2 = \sum_{i=1}^{M} E_{1,i}^2 \,,$$

où M est le nombre de points de la grille.

En queue de distribution, nous avons choisi une autre mesure d'erreur, parce que celle utilisée en partie centrale ne permettait pas de discriminer certains mauvais comportements en queue de distribution. En effet, l'erreur relative au point x s'exprime comme

$$E_1(x) = \left| \frac{F(x|\,\tilde{\alpha},\tilde{\beta}) - F(x|\,\alpha_{\text{vrai}},\beta_{\text{vrai}})}{1 - F(x|\,\alpha_{\text{vrai}},\beta_{\text{vrai}})} \right| = \left| \frac{(1+x/\tilde{\beta})^{-\tilde{\alpha}}}{(1+x/\beta_{\text{vrai}})^{-\alpha_{\text{vrai}}}} - 1 \right| .$$

Il s'ensuit que si $\tilde{\alpha} > \alpha_{\text{vrai}}$ (> 0 par hypothèse) alors le numérateur est prédominant, et donc l'erreur relative $E_1(x) \to 1$ lorsque $x \to \infty$. Par contre, lorsque $\alpha_{\text{vrai}} > \tilde{\alpha}$ (> 0 par hypothèse) alors le dénominateur est prédominant, et donc l'erreur relative $E_1(x) \to \infty$ lorsque $x \to \infty$. Ce type d'erreur a donc tendance à ne pas détecter une surestimation de α_{vrai}, même importante, tout en détectant une sous-estimation, même minime, ce qui ne nous satisfait pas. On propose plutôt d'utiliser une erreur du type

$$E_2(x) = \left| \ln \left(\frac{1 - F(x|\,\tilde{\alpha},\tilde{\beta})}{1 - F(x|\,\alpha_{\text{vrai}},\beta_{\text{vrai}})} \right) \right| = \left| \ln \left(\frac{(1+x/\tilde{\beta})^{-\tilde{\alpha}}}{(1+x/\beta_{\text{vrai}})^{-\alpha_{\text{vrai}}}} \right) \right| ,$$

telle que $E_2(x) \to \infty$ lorsque $x \to \infty$, que α_{vrai} soit surestimé ou sous-estimé.

Pour la queue de distribution (du quantile d'ordre 0.9 au quantile d'ordre $1-1/20n$ de la vraie loi), on calcule pour chaque point y_i d'une grille régulière la valeur absolue du logarithme du rapport de la fonction de survie estimée à la vraie fonction de survie :

$$E_{2,i} = \left| \ln \left(\frac{1 - F(y_i|\,\tilde{\alpha},\tilde{\beta})}{1 - F(y_i|\,\alpha_{\text{vrai}},\beta_{\text{vrai}})} \right) \right| ,$$

où $\tilde{\alpha}$ et $\tilde{\beta}$ sont les estimations de α et β obtenues par les estimateurs des moments pondérés, ou bien les estimateurs bayésiens par le mode, la moyenne ou la médiane) ; et α_{vrai} et β_{vrai} sont les vraies valeurs des paramètres, celles selon lesquelles les différents échantillons ont été simulés. Pour la loi prédictive, on remplace $F(y_i|\,\tilde{\alpha},\tilde{\beta})$ par la fonction de répartition de la loi prédictive évaluée au point y_i. Puis, on calcule

$$E_2^2 = \sum_{i=1}^{M} E_{2,i}^2 \,,$$

où M est le nombre de points de la grille.

Annexe E

Simulations complémentaires sur l'estimation bayésienne de la loi GPD

E.1 Simulations dans le cas du mode a priori pour des échantillons de loi GPD

Nous explorons maintenant le cas alternatif (pour la détermination des hyperparamètres) du mode a priori. L'information est ici encore apportée par l'échantillon à travers les estimateurs des moments pondérés. Mais pour déterminer les hyperparamètres, on suppose ici que ces estimateurs sont les modes des lois a priori (marginale sur α et conditionnelle sachant α sur β). Les résultats obtenus sont présentés dans les tableaux suivants.

		estimateurs de		erreurs sur		
		α	β	(α, β)	α	β
moments pondérés		1.0500	18.7035	2522.3611	0.3049	2522.0562
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	0.5216	0.5775	0.0442	0.0058	0.0384
	médiane	0.5328	0.6296	0.0590	0.0057	0.0532
	moyenne	0.5392	0.6617	0.0727	0.0062	0.0665
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	0.5217	0.5840	0.0432	0.0057	0.0375
	médiane	0.5334	0.6345	0.0622	0.0060	0.0562
	moyenne	0.5412	0.6668	0.0774	0.0067	0.0707
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	0.5224	0.5720	0.0393	0.0058	0.0335
	médiane	0.5346	0.6333	0.0599	0.0060	0.0539
	moyenne	0.5413	0.6643	0.0731	0.0066	0.0666

TAB. E.1 – *Estimation des paramètres (α, β) pour des échantillons de loi GPD*$(0.5, 0.5)$

		estimateurs de γ	estimateurs de σ	erreurs sur (γ, σ)	erreurs sur γ	erreurs sur σ
moments pondérés		0.9543	18.2665	2430.7974	1.0951	2429.7022
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	1.9529	1.1046	0.1683	0.0691	0.0992
	médiane	1.9059	1.1757	0.1821	0.0625	0.1196
	moyenne	1.8827	1.2205	0.2083	0.0648	0.1435
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	1.9515	1.1142	0.1585	0.0684	0.0900
	médiane	1.9050	1.1833	0.1902	0.0644	0.1258
	moyenne	1.8777	1.2252	0.2184	0.0686	0.1497
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	1.9497	1.0938	0.1527	0.0681	0.0846
	médiane	1.9003	1.1785	0.1823	0.0646	0.1176
	moyenne	1.8767	1.2206	0.2072	0.0680	0.1392

TAB. E.2 – *Estimation des paramètres* (γ, σ) *pour des échantillons de loi GPD*$(0.5, 0.5)$

		erreurs sur le centre	erreurs en queue
moments pondérés		239.9471	2712.7202
point initial du Gibbs : (1,1)			
estimateurs bayésiens par	mode	4.7490	99.3773
	médiane	3.6751	91.0143
	moyenne	3.5802	95.1603
loi prédictive		3.4960	49.4238
point initial du Gibbs : (100,100)			
estimateurs bayésiens par	mode	4.1901	94.8760
	médiane	3.7800	94.5934
	moyenne	3.7392	103.4580
loi prédictive		3.6373	53.2043
point initial du Gibbs : estimateurs des moments pondérés			
estimateurs bayésiens par	mode	4.6021	101.8578
	médiane	3.6776	94.5775
	moyenne	3.6233	101.1484
loi prédictive		3.5392	54.4785

TAB. E.3 – *Estimation de la fonction de survie pour des échantillons de loi GPD*$(0.5, 0.5)$

		estimateurs de		erreurs sur		
		α	β	(α, β)	α	β
moments pondérés		1.3725	1.8194	1.1082	0.2092	0.8990
point initial du Gibbs : (1,1)						
estimateurs	mode	1.0669	1.1630	0.2286	0.0529	0.1757
bayésiens	médiane	1.1254	1.2674	0.3033	0.0683	0.2350
par	moyenne	1.1601	1.3475	0.4255	0.0929	0.3325
point initial du Gibbs : (100,100)						
estimateurs	mode	1.0603	1.1193	0.1880	0.0474	0.1407
bayésiens	médiane	1.1189	1.2511	0.2945	0.0706	0.2239
par	moyenne	1.1556	1.3423	0.4286	0.0953	0.3333
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs	mode	1.0672	1.1171	0.1669	0.0492	0.1177
bayésiens	médiane	1.1224	1.2570	0.2873	0.0666	0.2207
par	moyenne	1.1551	1.3368	0.3840	0.0848	0.2992

TAB. E.4 – *Estimation des paramètres* (α, β) *pour des échantillons de loi GPD*$(1, 1)$

		estimateurs de		erreurs sur		
		γ	σ	(γ, σ)	γ	σ
moments pondérés		0.7515	1.3289	0.2559	0.0769	0.1790
point initial du Gibbs : (1,1)						
estimateurs	mode	0.9730	1.0766	0.0908	0.0338	0.0570
bayésiens	médiane	0.9204	1.1114	0.0909	0.0333	0.0576
par	moyenne	0.8968	1.1441	0.1063	0.0382	0.0681
point initial du Gibbs : (100,100)						
estimateurs	mode	0.9745	1.0474	0.0822	0.0284	0.0538
bayésiens	médiane	0.9265	1.1036	0.0849	0.0318	0.0530
par	moyenne	0.9012	1.1434	0.1044	0.0371	0.0673
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs	mode	0.9705	1.0449	0.0848	0.0318	0.0531
bayésiens	médiane	0.9225	1.1067	0.0870	0.0328	0.0542
par	moyenne	0.8985	1.1424	0.1017	0.0366	0.0651

TAB. E.5 – *Estimation des paramètres* (γ, σ) *pour des échantillons de loi GPD*$(1, 1)$

		erreurs sur le centre	erreurs en queue
moments pondérés		5.2593	488.6246
point initial du Gibbs : (1,1)			
estimateurs bayésiens par	mode	3.9466	147.2105
	médiane	2.9382	175.4153
	moyenne	2.9428	227.0381
loi prédictive		2.8078	70.4081
point initial du Gibbs : (100,100)			
estimateurs bayésiens par	mode	3.9015	144.7287
	médiane	2.8098	182.0931
	moyenne	2.8720	231.8119
loi prédictive		2.7159	68.2194
point initial du Gibbs : estimateurs des moments pondérés			
estimateurs bayésiens par	mode	4.7728	157.1052
	médiane	2.9750	173.8408
	moyenne	2.9323	210.9890
loi prédictive		2.7788	67.2321

TAB. E.6 – *Estimation de la fonction de survie pour des échantillons de loi GPD*$(1,1)$

		estimateurs de		erreurs sur		
		α	β	(α, β)	α	β
moments pondérés		2.4466	2.5666	3.3443	1.1277	2.2166
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	2.1868	2.2289	1.6611	0.5857	1.0754
	médiane	2.4936	2.6775	3.1920	1.0089	2.1831
	moyenne	2.7453	3.0735	5.8611	1.7934	4.0677
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	2.1937	2.2751	1.5612	0.5252	1.0360
	médiane	2.5364	2.7297	3.7337	1.2314	2.5023
	moyenne	2.7830	3.1131	6.4787	2.0758	4.4028
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	2.1954	2.2323	1.7866	0.4235	1.3631
	médiane	2.5505	2.7497	4.1857	1.3907	2.7951
	moyenne	2.7955	3.1216	7.0454	2.3269	4.7186

TAB. E.7 – *Estimation des paramètres* (α, β) *pour des échantillons de loi GPD*$(2,2)$

		estimateurs de		erreurs sur		
		γ	σ	(γ, σ)	γ	σ
moments pondérés		0.4566	1.0161	0.0483	0.0197	0.0286
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	0.4941	1.0080	0.0535	0.0158	0.0376
	médiane	0.4398	1.0470	0.0489	0.0189	0.0299
	moyenne	0.4095	1.0841	0.0622	0.0243	0.0379
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	0.4924	1.0210	0.0502	0.0159	0.0343
	médiane	0.4339	1.05120	0.0462	0.0184	0.0278
	moyenne	0.4042	1.0883	0.0597	0.0235	0.0362
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	0.4876	0.9948	0.0713	0.0153	0.0560
	médiane	0.4353	1.0499	0.0464	0.0191	0.0273
	moyenne	0.4073	1.0835	0.0596	0.0238	0.0358

TAB. E.8 – *Estimation des paramètres* (γ, σ) *pour des échantillons de loi GPD*$(2, 2)$

		erreurs sur le centre	erreurs en queue
moments pondérés		2.6393	292.2566
point initial du Gibbs : (1,1)			
estimateurs bayésiens par	mode	4.4526	206.5644
	médiane	2.7026	258.2393
	moyenne	2.7321	389.8836
loi prédictive		2.5430	86.2258
point initial du Gibbs : (100,100)			
estimateurs bayésiens par	mode	3.8870	171.2713
	médiane	2.6019	292.9711
	moyenne	2.7868	430.3352
loi prédictive		2.5798	95.1684
point initial du Gibbs : estimateurs des moments pondérés			
estimateurs bayésiens par	mode	5.2037	153.9018
	médiane	2.5644	472.1446
	moyenne	2.7982	472.1446
loi prédictive		2.5743	88.8945

TAB. E.9 – *Estimation de la fonction de survie pour des échantillons de loi GPD*$(2, 2)$

		estimateurs de		erreurs sur		
		α	β	(α, β)	α	β
moments pondérés		7.8733	8.6586	233.5897	89.5299	144.0598
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	6.9685	7.2135	314.3170	123.5669	190.7501
	médiane	8.2858	9.1860	87.4650	35.0793	52.3857
	moyenne	11.6004	13.3776	644.9071	238.6254	406.2818
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	6.5113	7.6964	159.002	58.2004	100.8020
	médiane	8.2407	9.2011	112.9547	42.9502	70.0045
	moyenne	11.0788	12.5905	392.3424	155.6673	236.6751
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	6.5285	7.5898	152.6602	51.8876	100.7726
	médiane	8.4843	9.4675	99.5629	38.4197	61.1432
	moyenne	11.4965	13.1953	452.1513	172.3471	279.8042

TAB. E.10 – *Estimation des paramètres* (α, β) *pour des échantillons de loi GPD*$(5,5)$

		estimateurs de		erreurs sur		
		γ	σ	(γ, σ)	γ	σ
moments pondérés		0.1981	1.0304	0.0320	0.0094	0.0226
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	0.2122	1.0402	0.1430	0.0078	0.1352
	médiane	0.1540	1.0821	0.0330	0.0090	0.0260
	moyenne	0.1296	1.1052	0.0399	0.0092	0.0307
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	0.2075	1.1453	0.2806	0.0071	0.2735
	médiane	0.1571	1.0799	0.0348	0.0069	0.0279
	moyenne	0.1318	1.1018	0.0404	0.0090	0.0314
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	0.2140	1.1342	0.3030	0.0092	0.2938
	médiane	0.1520	1.0839	0.0353	0.0074	0.0279
	moyenne	0.1275	1.1086	0.0421	0.0096	0.0325

TAB. E.11 – *Estimation des paramètres* (γ, σ) *pour des échantillons de loi GPD*$(5,5)$

On constate sur les tableaux E.1 à E.15 que la stationnarité semble approximativement atteinte. Mais d'autre part, les résultats sont en général plus mauvais (plus éloignés des vraies valeurs des paramètres, et erreurs plus grandes) que ceux obtenus dans le cas de la moyenne a priori, et même parfois que les estimateurs des moments pondérés. On préfèrera donc utiliser le cas de la moyenne a priori, pour déterminer les hyperparamètres.

		erreurs sur le centre	erreurs en queue
moments pondérés		1.8882	133.1590
point initial du Gibbs : (1,1)			
estimateurs bayésiens par	mode	12.4488	604.0195
	médiane	2.1969	124.7312
	moyenne	2.2677	186.4450
loi prédictive		2.1523	53.2388
point initial du Gibbs : (100,100)			
estimateurs bayésiens par	mode	17.8258	129.2941
	médiane	2.3528	114.4664
	moyenne	2.4449	195.7755
loi prédictive		2.2444	48.8923
point initial du Gibbs : estimateurs des moments pondérés			
estimateurs bayésiens par	mode	18.6518	220.5821
	médiane	2.3413	130.3663
	moyenne	2.3929	196.8266
loi prédictive		2.2662	51.3663

TAB. E.12 – *Estimation de la fonction de survie pour des échantillons de loi GPD*$(5,5)$

		estimateurs de		erreurs sur		
		α	β	(α,β)	α	β
moments pondérés		25.9249	26.4127	6938.3136	3424.9513	3513.3623
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	16.2347	16.3821	1335.4006	582.9008	752.4998
	médiane	17.1944	17.4204	697.6268	353.5623	344.0645
	moyenne	28.6527	29.5941	4007.5540	1965.4146	2042.1395
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	20.1659	20.9085	3042.9068	1513.6326	1529.2742
	médiane	19.5200	19.8913	1513.9963	772.6969	741.2994
	moyenne	32.5560	33.5170	7514.5322	3747.8231	3766.7091
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	16.5336	16.9972	1489.3098	733.4778	755.8320
	médiane	17.8852	18.3881	831.7708	404.2863	427.4846
	moyenne	26.3984	27.4068	3432.0013	1667.9057	1764.0956

TAB. E.13 – *Estimation des paramètres* (α,β) *pour des échantillons de loi GPD*$(10,10)$

		estimateurs de		erreurs sur		
		γ	σ	(γ, σ)	γ	σ
moments pondérés		0.1260	0.9747	0.0247	0.0071	0.0176
point initial du Gibbs : (1,1)						
estimateurs bayésiens par	mode	0.1311	1.0444	0.2666	0.0067	0.2599
	médiane	0.0933	1.0103	0.0195	0.0027	0.0167
	moyenne	0.0754	1.0286	0.0197	0.0027	0.0171
point initial du Gibbs : (100,100)						
estimateurs bayésiens par	mode	0.1227	1.0488	0.1462	0.0056	0.1406
	médiane	0.0879	1.0186	0.0199	0.0024	0.0175
	moyenne	0.0706	1.0325	0.0193	0.0026	0.0168
point initial du Gibbs : estimateurs des moments pondérés						
estimateurs bayésiens par	mode	0.1230	1.0292	0.2932	0.0055	0.2877
	médiane	0.0898	1.0175	0.0207	0.0028	0.0180
	moyenne	0.0734	1.0295	0.0205	0.0027	0.0178

TAB. E.14 – *Estimation des paramètres* (γ, σ) *pour des échantillons de loi GPD*$(10, 10)$

		erreurs sur le centre	erreurs en queue
moments pondérés		2.2044	98.8211
point initial du Gibbs : (1,1)			
estimateurs bayésiens par	mode	20.8066	∞ (dépasse la précision de matlab)
	médiane	2.5580	82.1437
	moyenne	2.4659	92.5189
loi prédictive		2.4343	66.3918
point initial du Gibbs : (100,100)			
estimateurs bayésiens par	mode	13.7070	209.6594
	médiane	2.5490	75.6219
	moyenne	2.3898	91.3748
loi prédictive		2.3693	58.5546
point initial du Gibbs : estimateurs des moments pondérés			
estimateurs bayésiens par	mode	14.9368	160.2920
	médiane	2.5790	73.3017
	moyenne	2.5260	88.0394
loi prédictive		2.3808	59.0175

TAB. E.15 – *Estimation de la fonction de survie pour des échantillons de loi GPD*$(10, 10)$

On peut cependant remarquer que, dans le cas du mode a priori, pour l'estimation des paramètres ((α, β) ou (γ, σ)), dans le cas de paramètres α et β petits (c'est-à-dire ≤ 2) on choisira de préférence d'utiliser l'estimateur bayésien par le mode ; et pour l'estimation de paramètres α et β grands (≥ 5) on préfèrera utiliser l'estimateur bayésien par le mode ou par la médiane (pour leurs bons compromis entre le biais et la variance : leurs erreurs sont faibles). En revanche, pour l'estimation de la Fds, on est toujours conduit à utiliser de préférence la loi prédictive ou l'estimateur bayésien par la médiane.

E.2 Échantillons d'excès – Intervalles de confiance empiriques pour les estimateurs de γ et de $q_{1-1/5000}$

Nous avons appliqué notre méthode bayésienne pour des échantillons d'excès (de taille m_n variant de 5 à 495), issus de 100 jeux de données simulées de taille $n = 500$. Nous avons comparé, au paragraphe 3.3.3 page 128, les moyennes des estimations du paramètre γ et du quantile d'ordre $1 - 1/10n = 1 - 1/5000$ de la loi des données originelles.

Dans ce paragraphe, on souhaite donner une idée de la variance des différents estimateurs calculés. Outre la valeur moyenne, on trace à présent un intervalle de confiance à 90%, pour les estimateurs de γ, et pour les estimateurs du quantile $q_{1-1/5000}$. Pour construire cet intervalle, pour chaque valeur du nombre d'excès m_n et chaque type d'estimateur, on considère l'échantillon des 100 valeurs de l'estimateur obtenues chacune pour l'un des 100 échantillons simulés. On ordonne ensuite cet échantillon, puis on ôte les deux plus petites et les deux plus grandes valeurs de cet échantillon ordonné. La borne inférieure de l'intervalle est alors la plus petite valeur restante de l'intervalle, et la borne inférieure la plus grande. On dispose ainsi, pour chaque valeur du nombre d'excès m_n et chaque type d'estimateur, d'une borne inférieure et d'une borne supérieure d'un intervalle de confiance empirique à 90%. On peut donc tracer, pour chaque type d'estimateur, des bornes de confiance en fonction de m_n.

Les différentes méthodes d'estimation que nous comparons sont : la méthode de Hill, la méthode de Hill généralisée, la méthode du Zipf, notre méthode bayésienne pour la moyenne a priori et la moyenne a posteriori, notre méthode bayésienne pour la moyenne a priori et la médiane a posteriori. Nous les appliquons pour des échantillons de loi de Fréchet de paramètre 1, de loi de Burr de paramètres $(1, 1, 1)$ et $(1, 0.5, 2)$, de loi de Student absolue de paramètre 1, 2, 4, et 8, et de loi loggamma de paramètre 2.

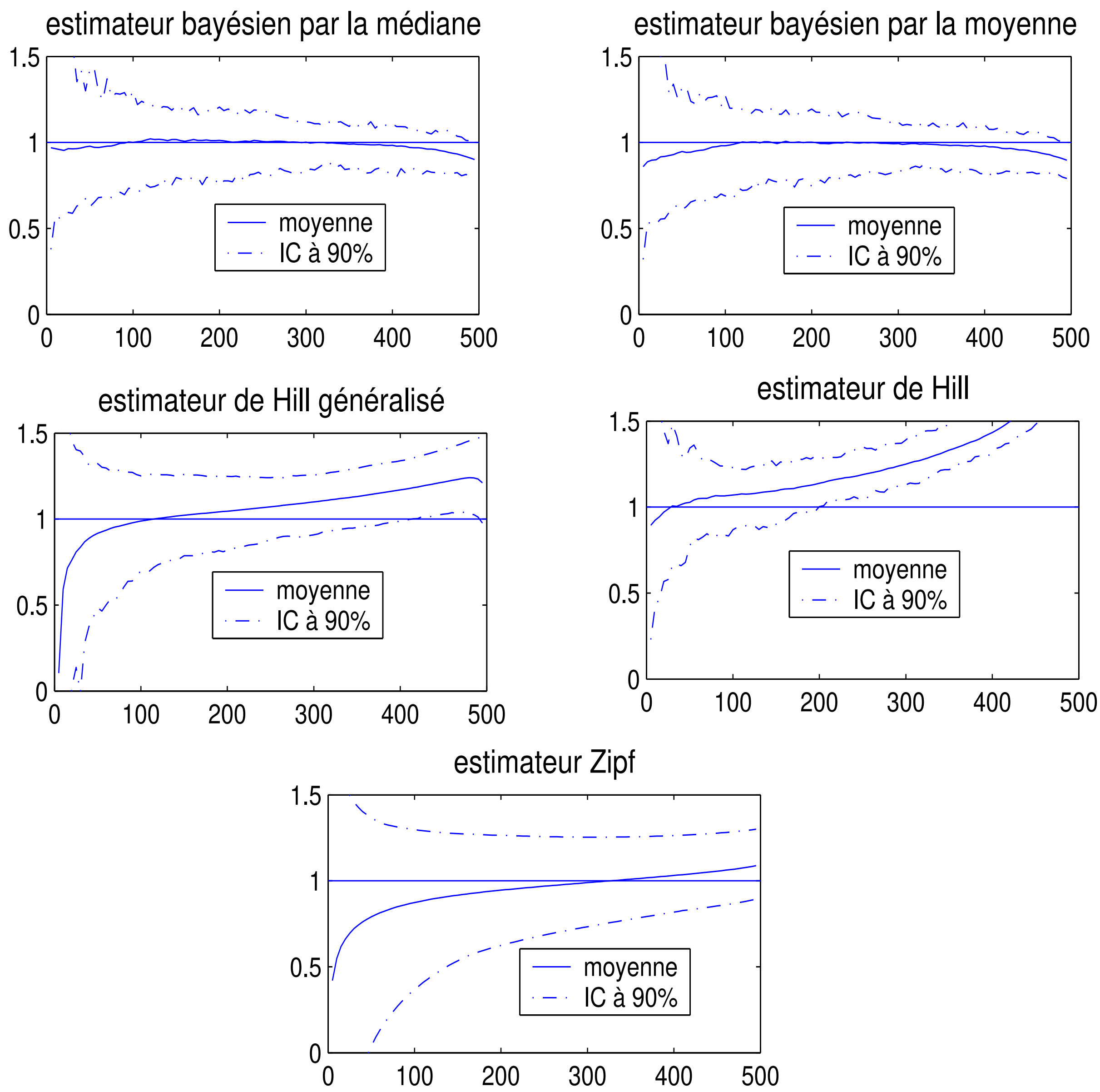

FIG. E.1 – *Échantillons de loi* $\mathcal{F}rechet(1)$ *– Moyennes et intervalles à 90% des différents estimateurs de* γ *en fonction de* m_n *: estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

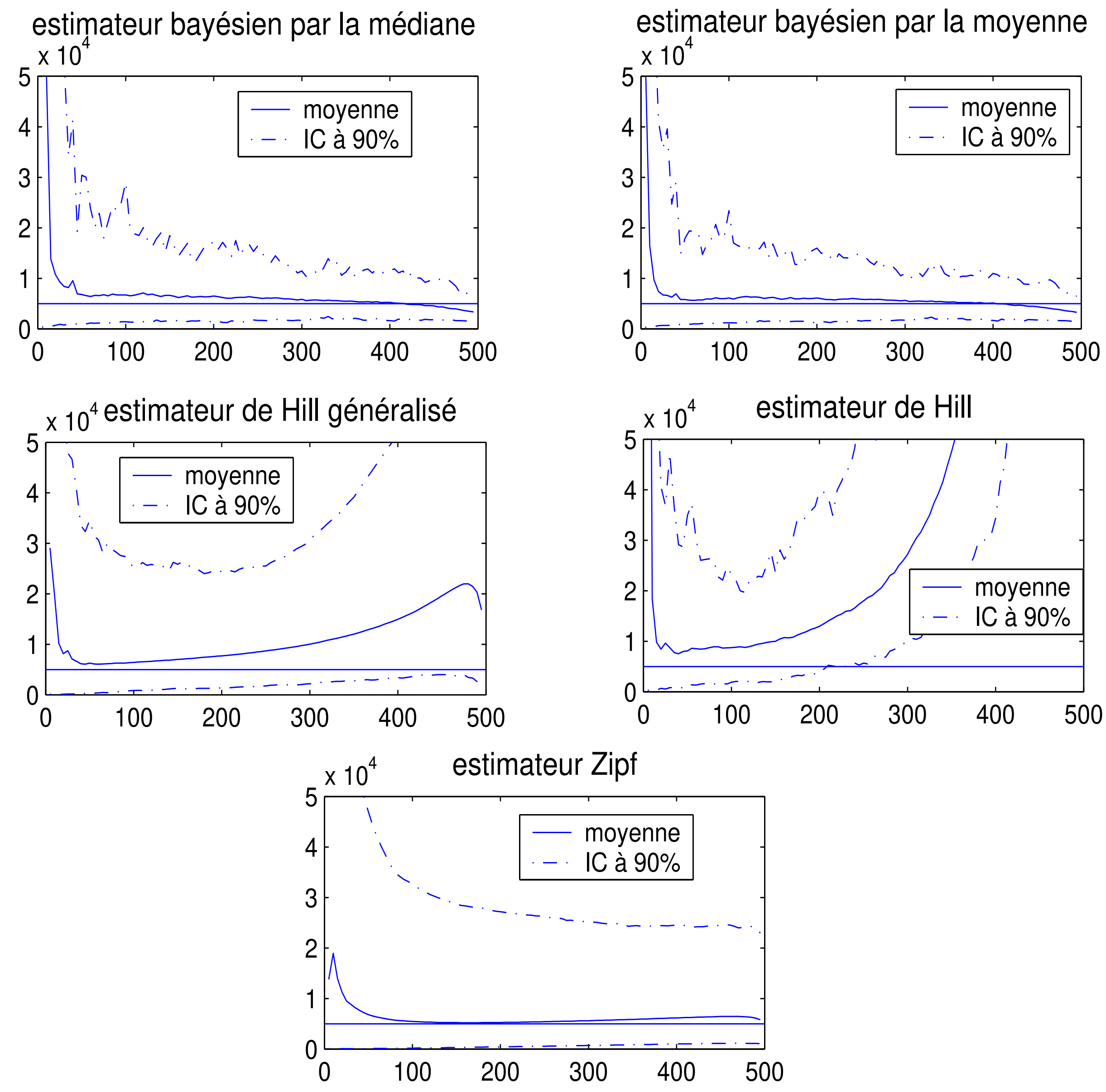

FIG. E.2 – *Échantillons de loi $\mathcal{F}rechet(1)$ – Moyennes et intervalles à 90% des différents estimateurs de $q_{1-1/5000}$ en fonction de m_n : estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

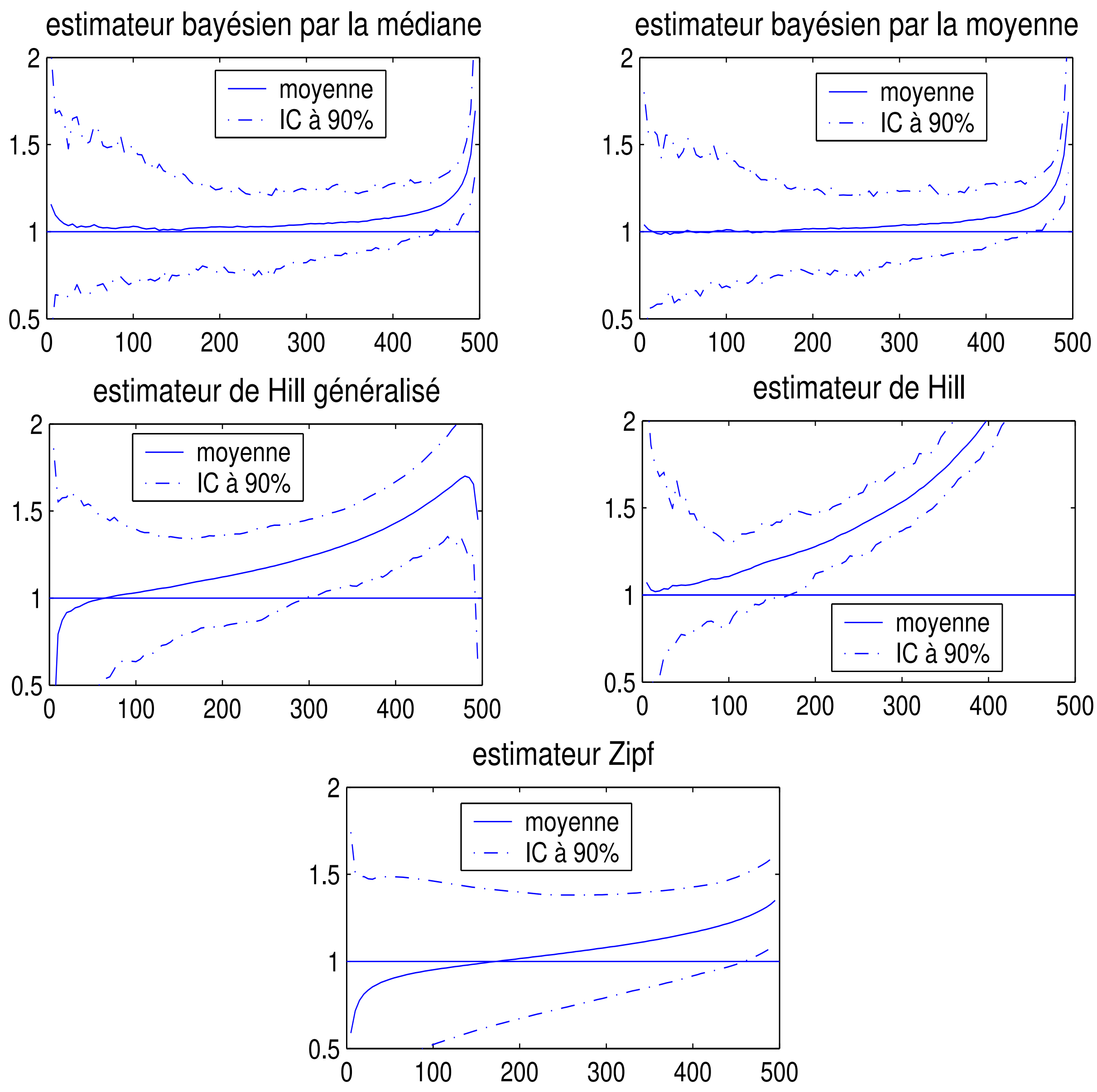

FIG. E.3 – *Échantillons simulés de loi* $\mathcal{B}urr(1,1,1)$ – *Moyennes et intervalles à 90% des différents estimateurs de* γ *en fonction de* m_n *: estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

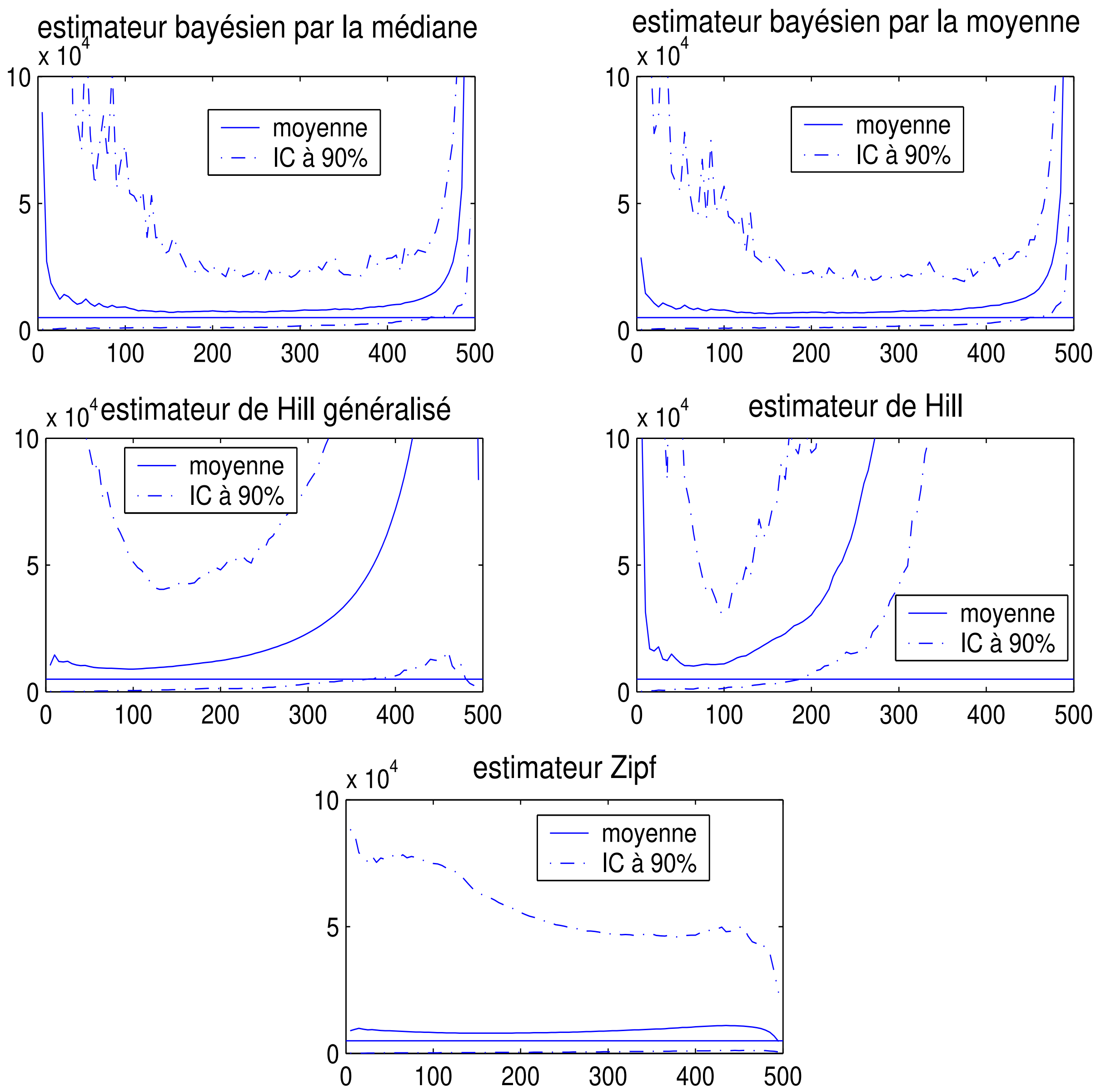

FIG. E.4 – *Échantillons simulés de loi* $\mathcal{B}urr(1,1,1)$ – *Moyennes et intervalles à* 90% *des différents estimateurs de* $q_{1-1/5000}$ *en fonction de* m_n *: estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

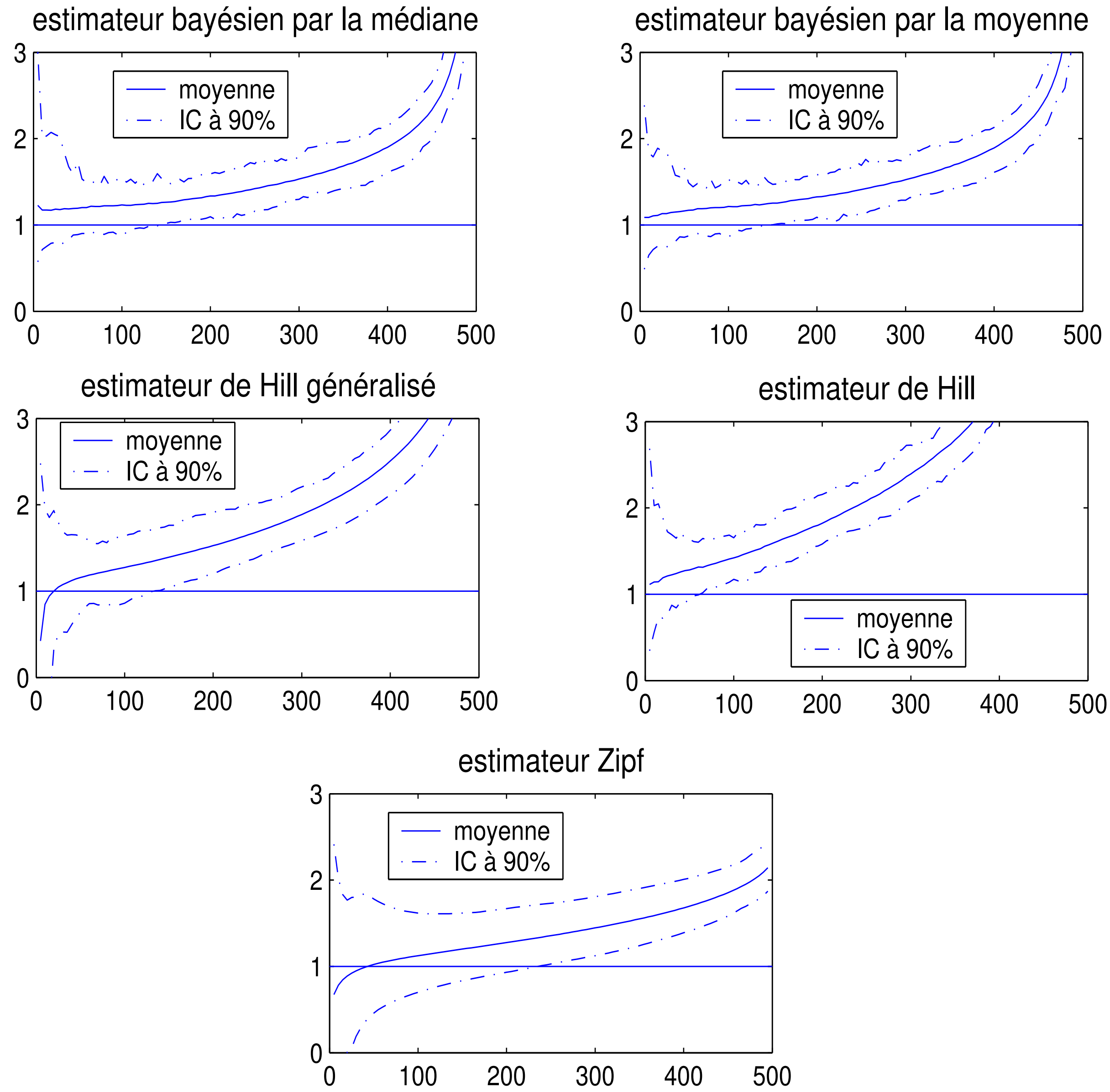

FIG. E.5 – *Échantillons simulés de loi* $\mathcal{B}urr(1, 0.5, 2)$ – *Moyennes et intervalles à* 90% *des différents estimateurs de* γ *en fonction de* m_n *: estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

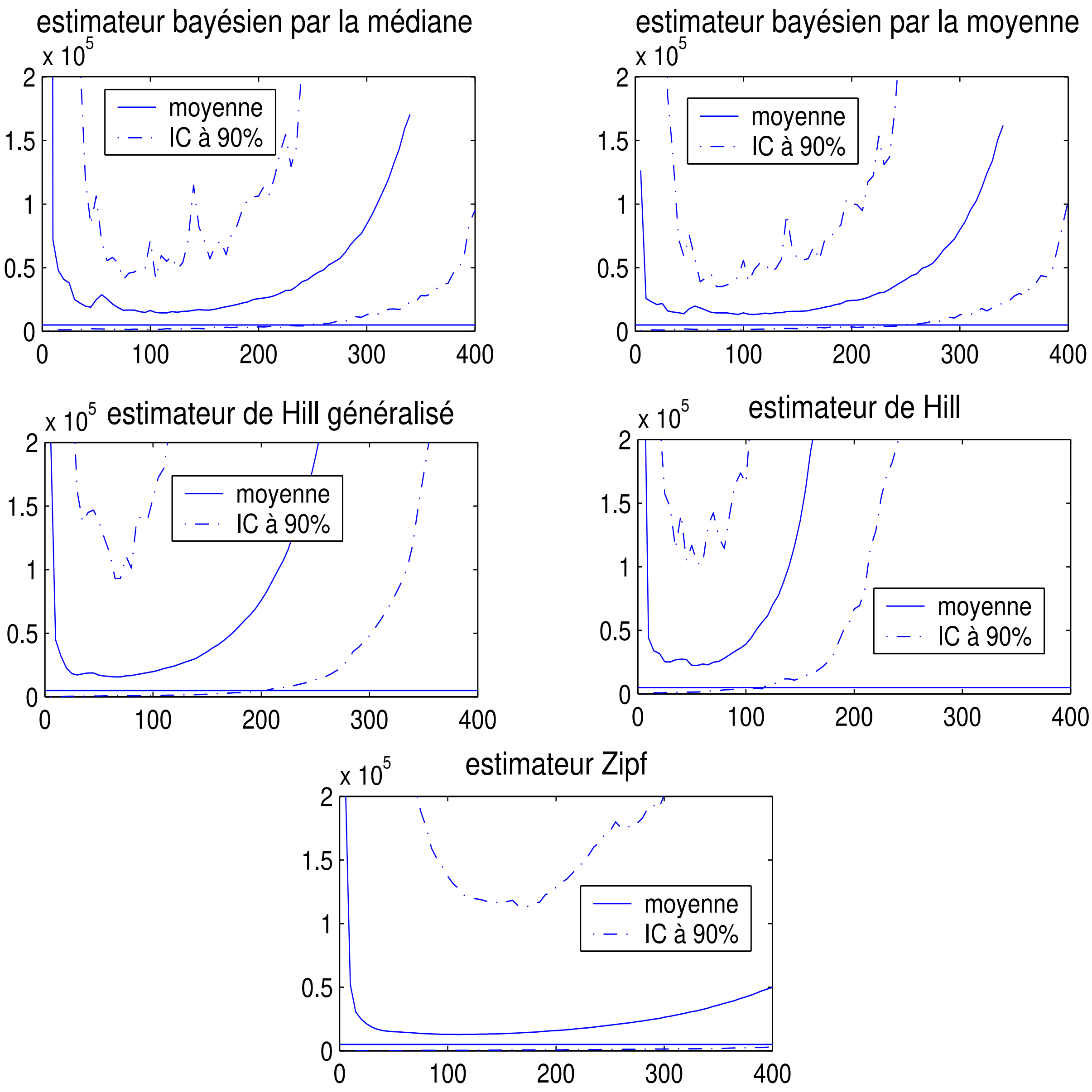

FIG. E.6 – *Échantillons simulés de loi* $\mathcal{B}urr(1, 0.5, 2)$ – *Moyennes et intervalles à* 90% *des différents estimateurs de* $q_{1-1/5000}$ *en fonction de* m_n *: estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

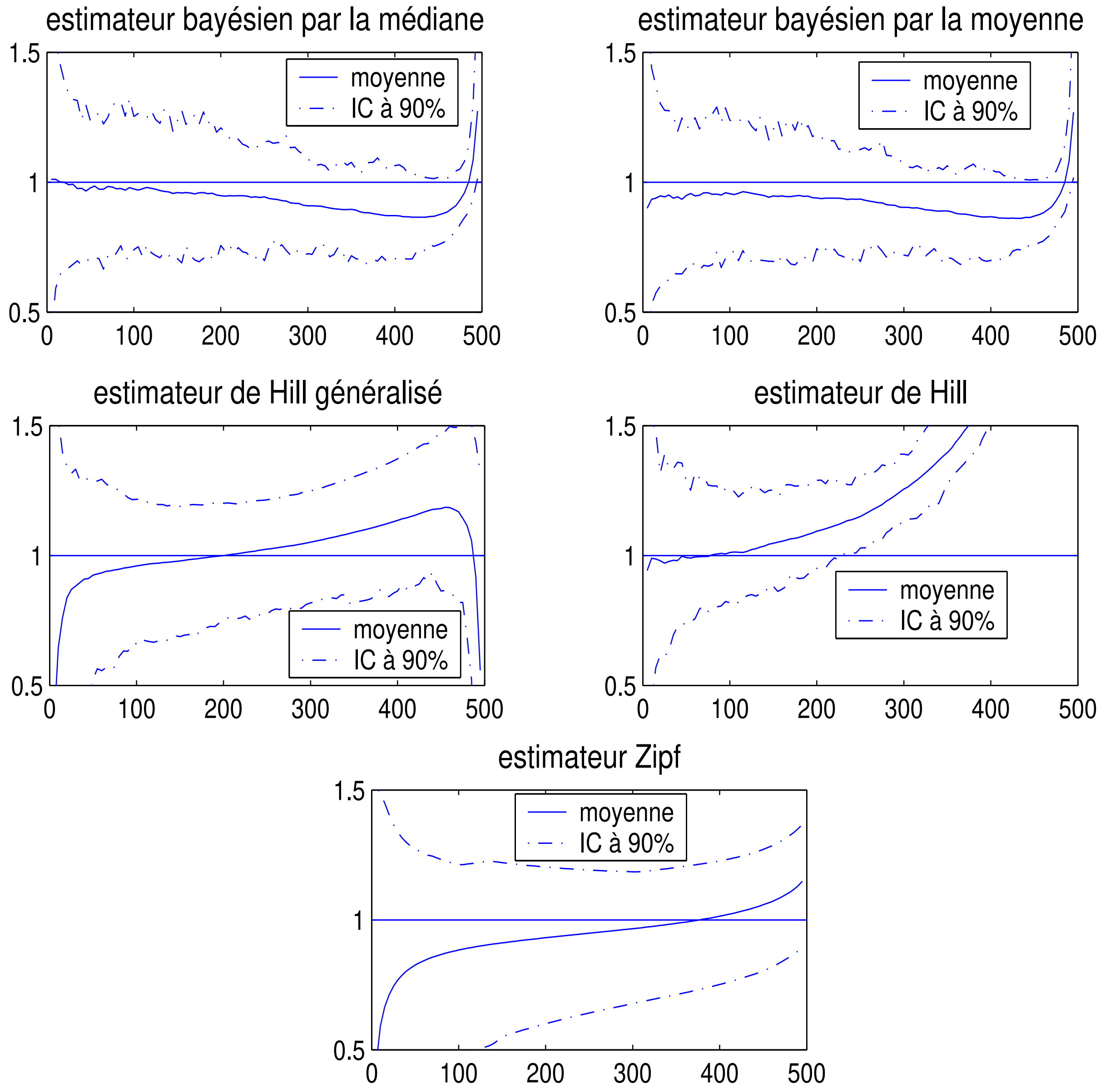

FIG. E.7 – *Échantillons simulés de loi* $t_{Abs}(1)$ – *Moyennes et intervalles à* 90% *des différents estimateurs de* γ *en fonction de* m_n *: estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

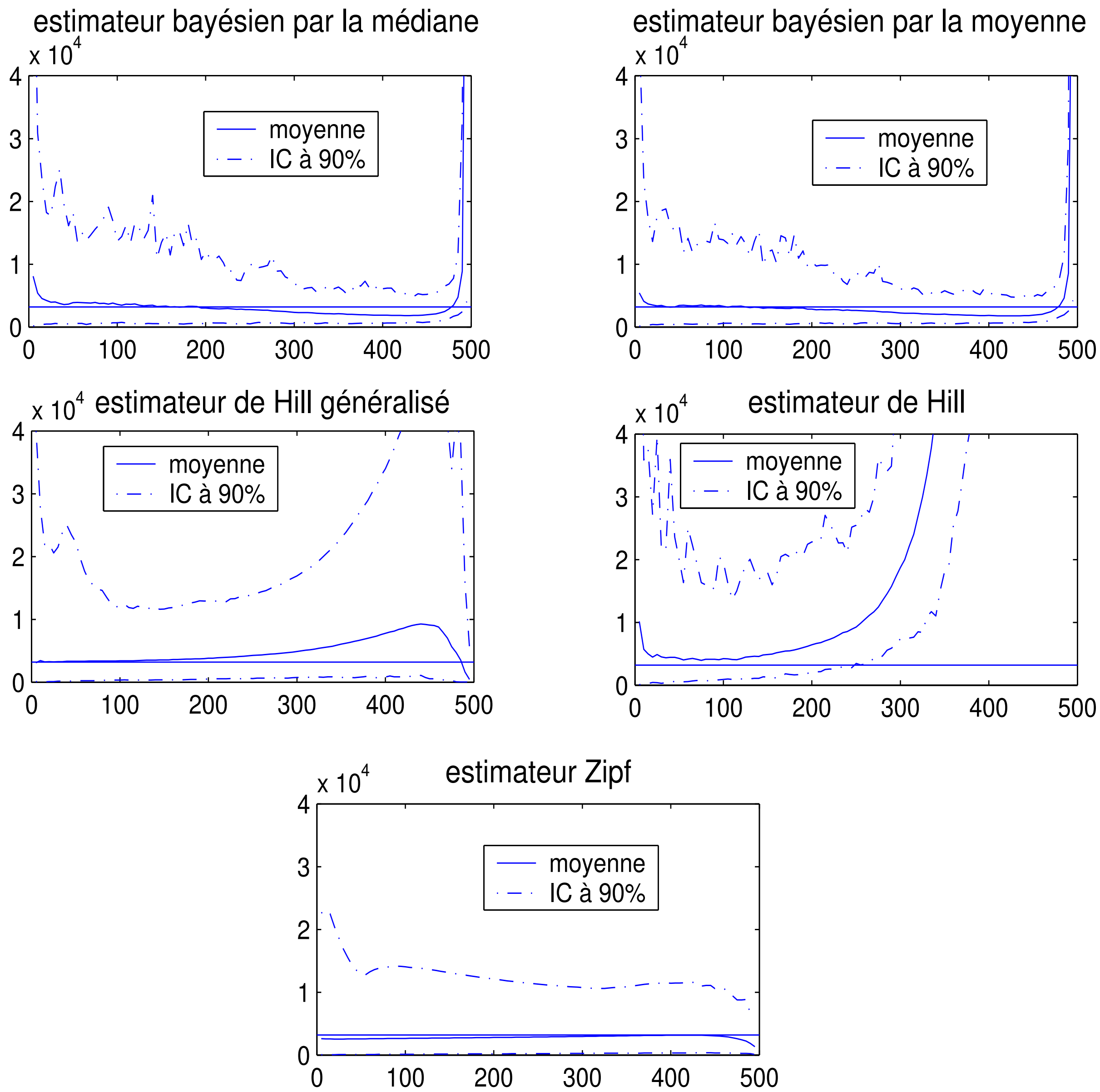

FIG. E.8 – *Échantillons simulés de loi* $t_{Abs}(1)$ – *Moyennes et intervalles à* 90% *des différents estimateurs de* $q_{1-1/5000}$ *en fonction de* m_n *: estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

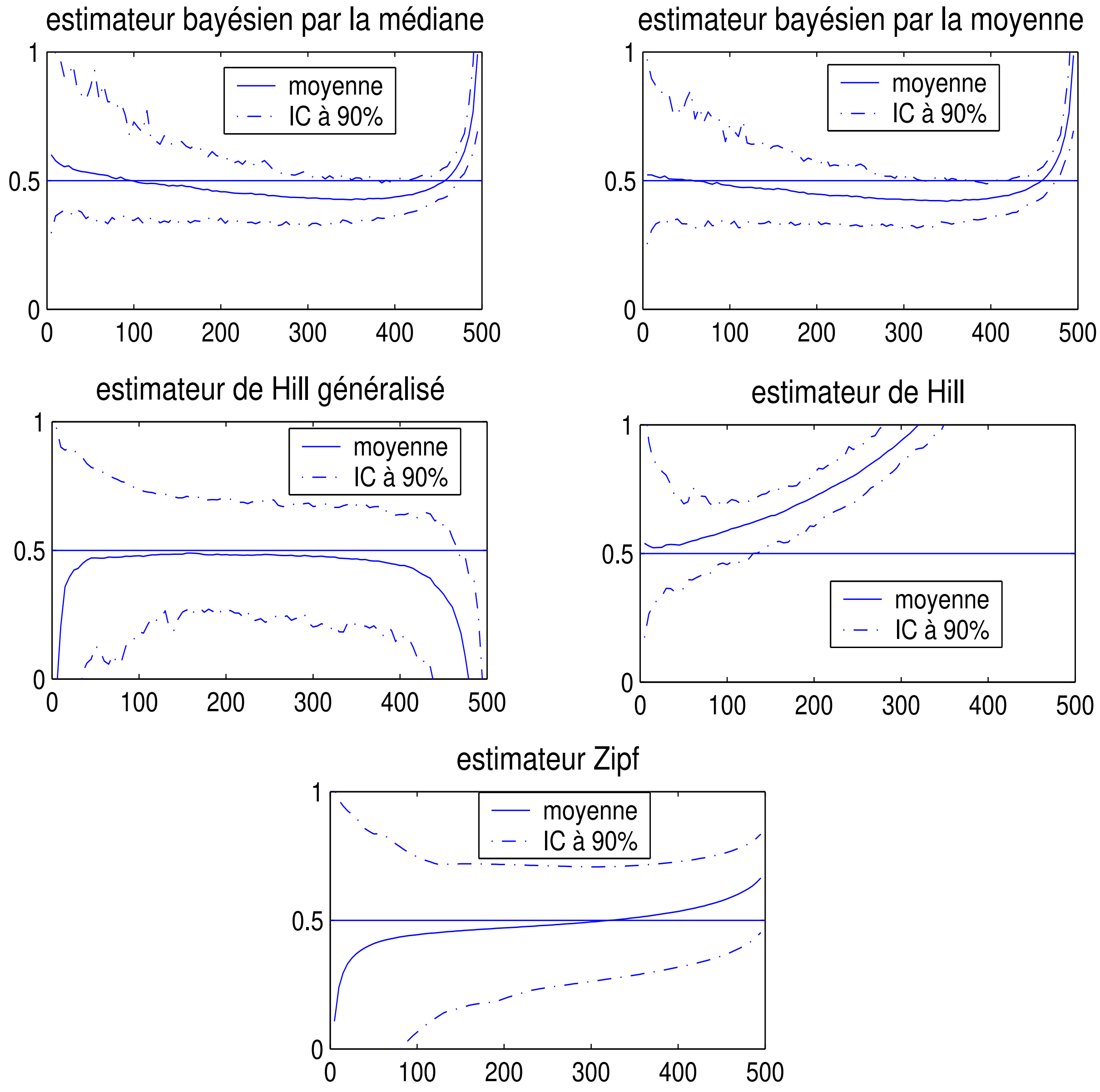

FIG. E.9 – *Échantillons simulés de loi* $t_{Abs}(2)$ – *Moyennes et intervalles à* 90% *des différents estimateurs de* γ *en fonction de* m_n *: estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

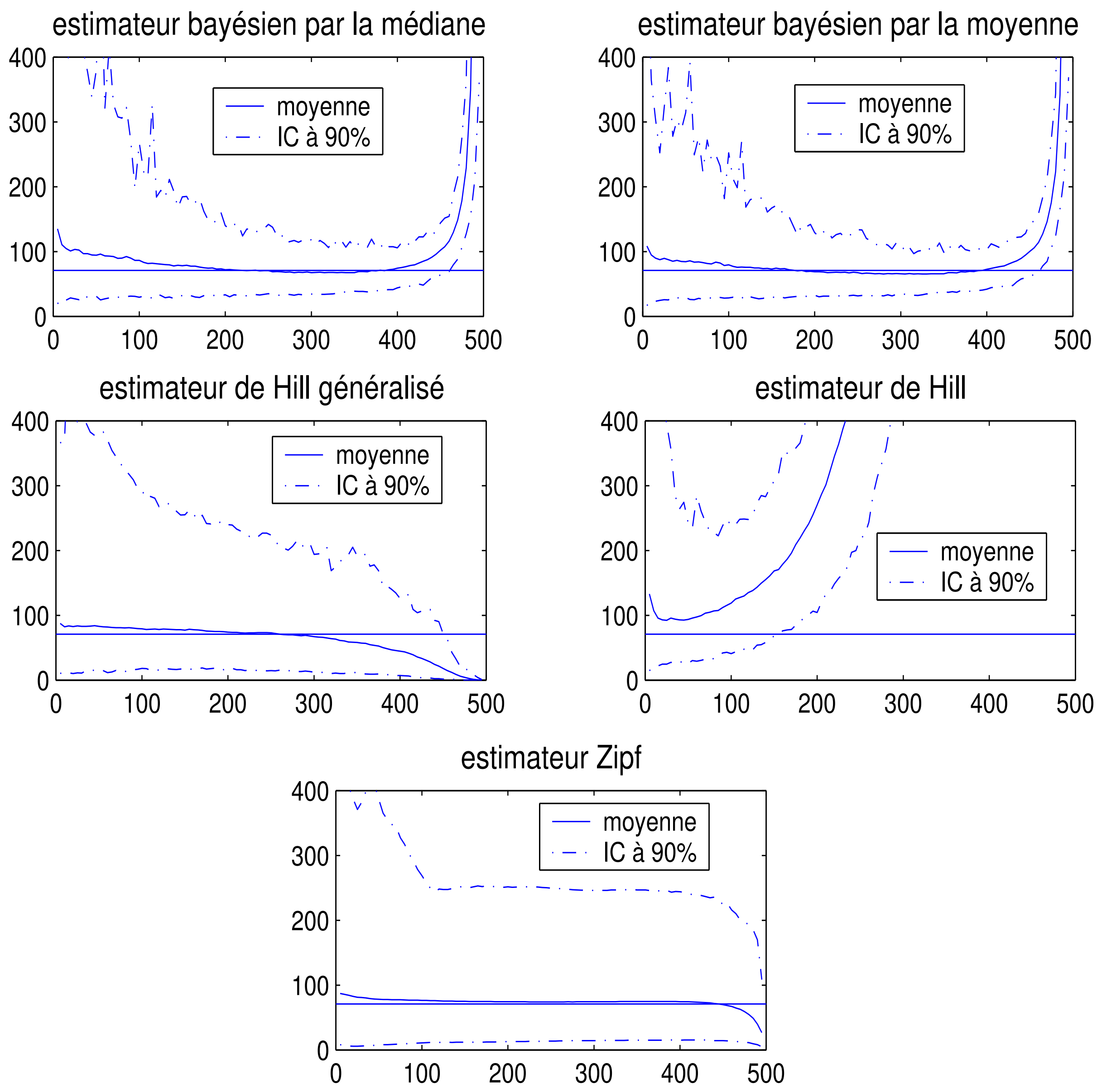

FIG. E.10 – *Échantillons simulés de loi* $t_{Abs}(2)$ – *Moyennes et intervalles à* 90% *des différents estimateurs de* $q_{1-1/5000}$ *en fonction de* m_n *: estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

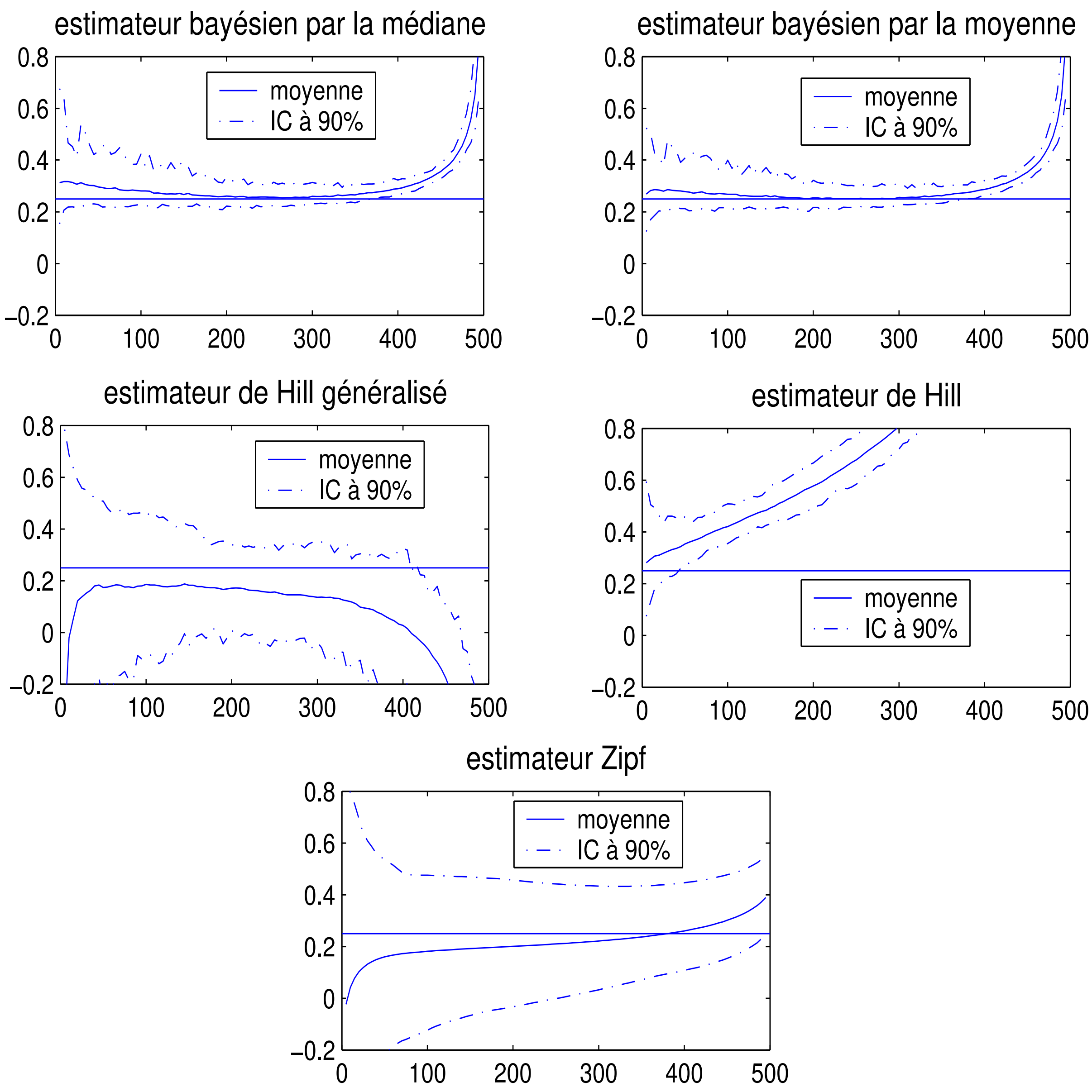

FIG. E.11 – *Échantillons simulés de loi* $t_{Abs}(4)$ – *Moyennes et intervalles à 90% des différents estimateurs de* γ *en fonction de* m_n *: estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

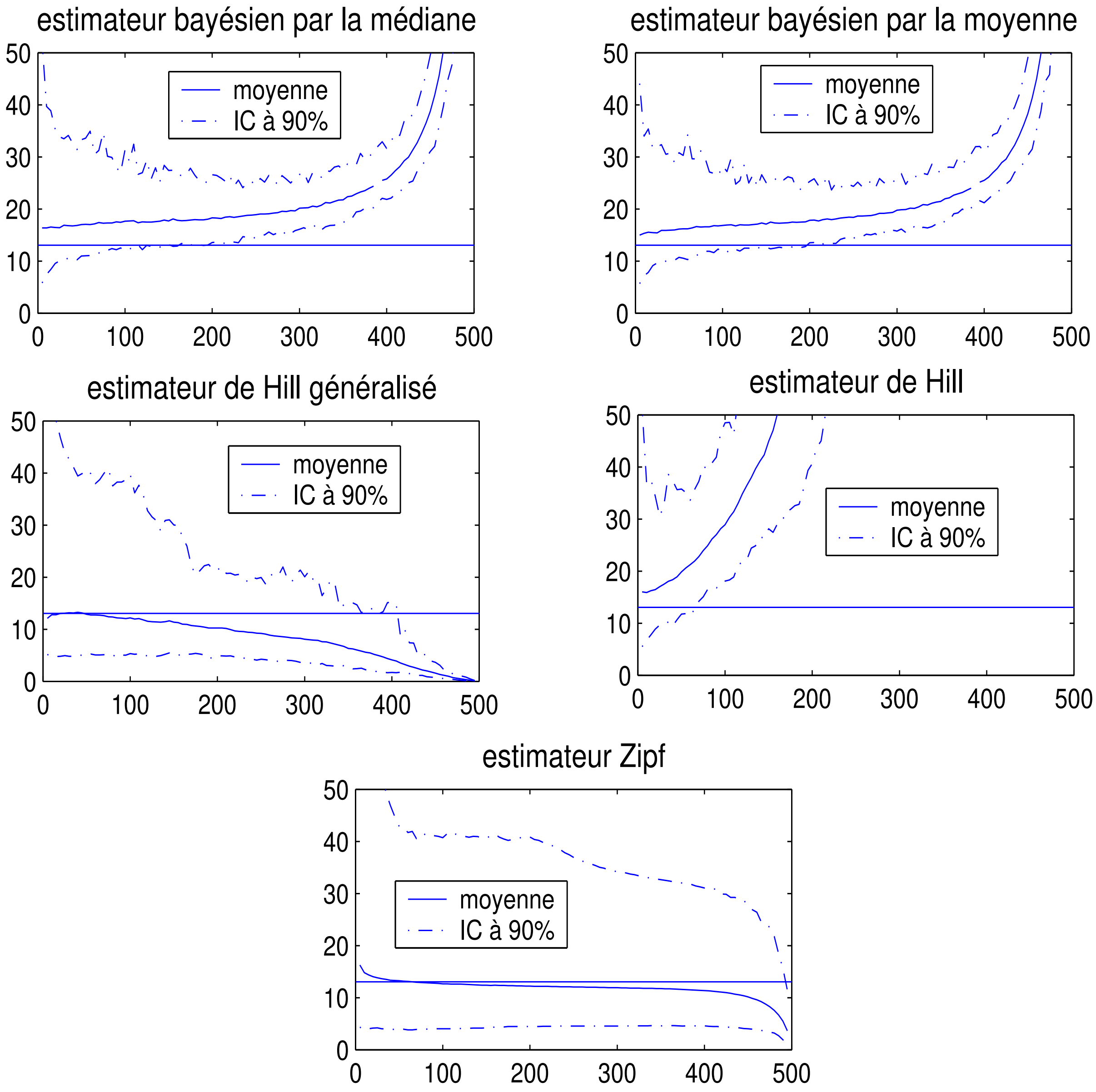

FIG. E.12 – *Échantillons simulés de loi* $t_{Abs}(4)$ *– Moyennes et intervalles à* 90% *des différents estimateurs de* $q_{1-1/5000}$ *en fonction de* m_n *: estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

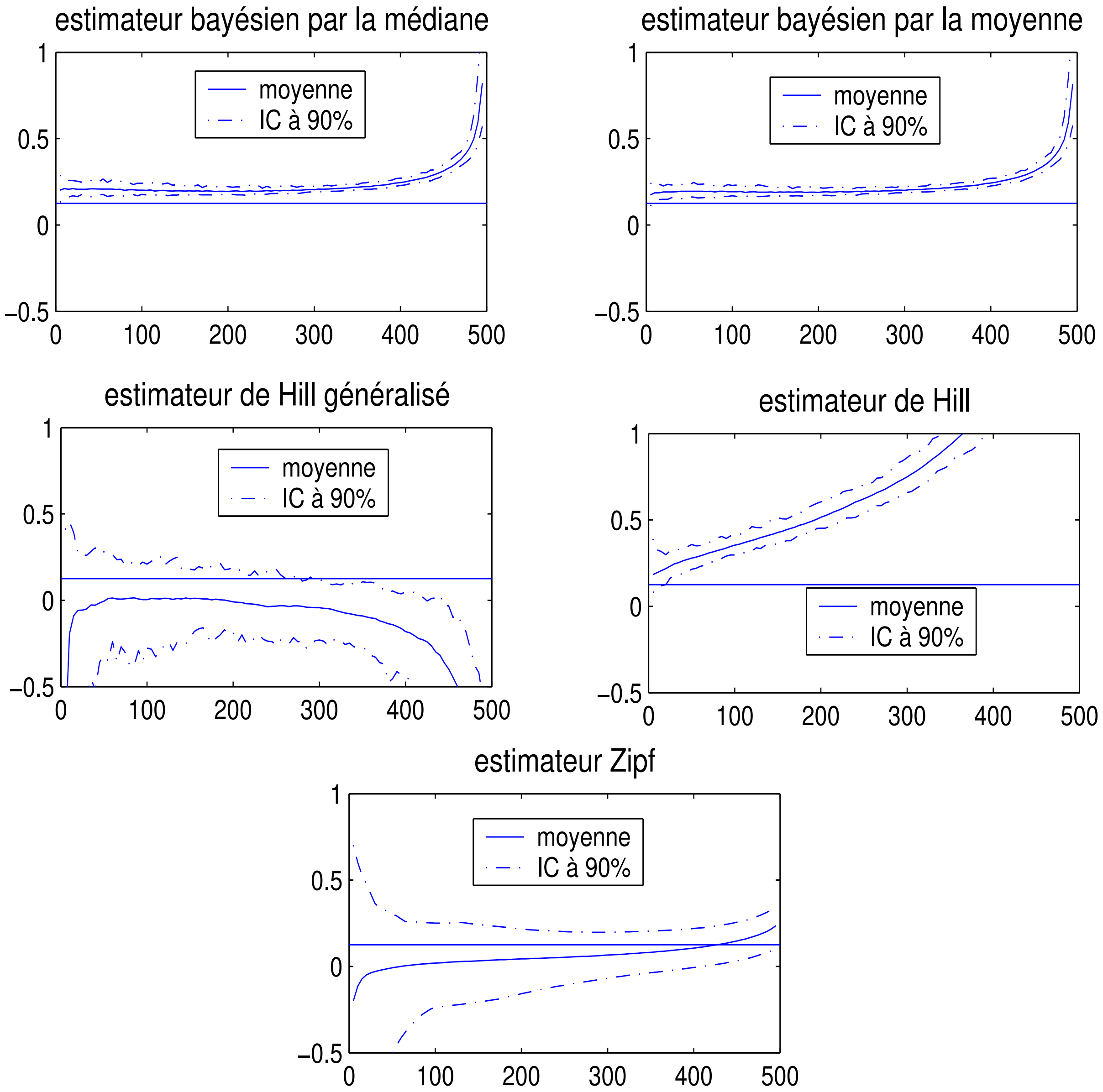

FIG. E.13 – *Échantillons simulés de loi* $t_{Abs}(8)$ – *Moyennes et intervalles à 90% des différents estimateurs de* γ *en fonction de* m_n *: estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

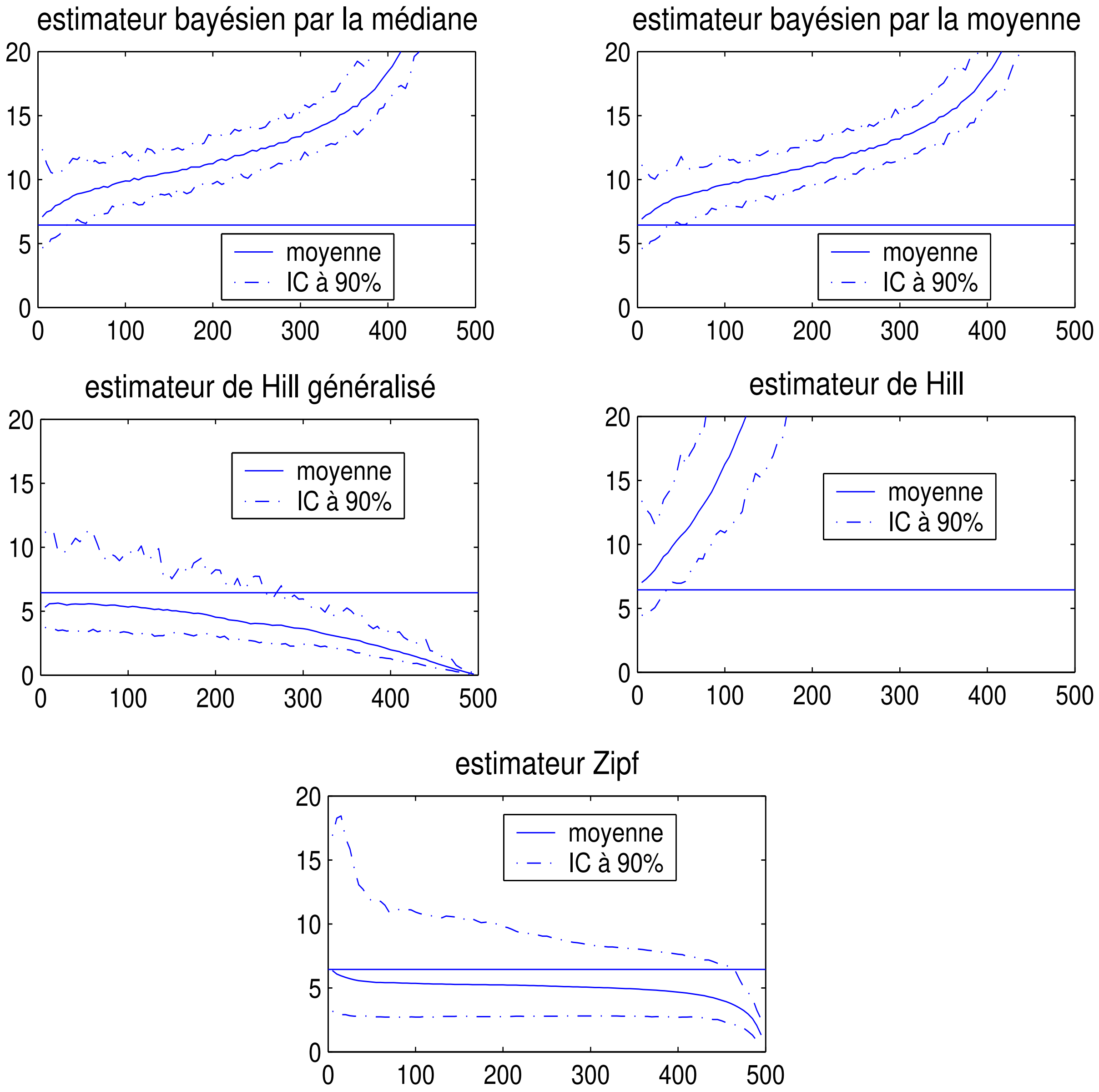

Fig. E.14 – *Échantillons simulés de loi* $t_{Abs}(8)$ – *Moyennes et intervalles à 90% des différents estimateurs de* $q_{1-1/5000}$ *en fonction de* m_n *: estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

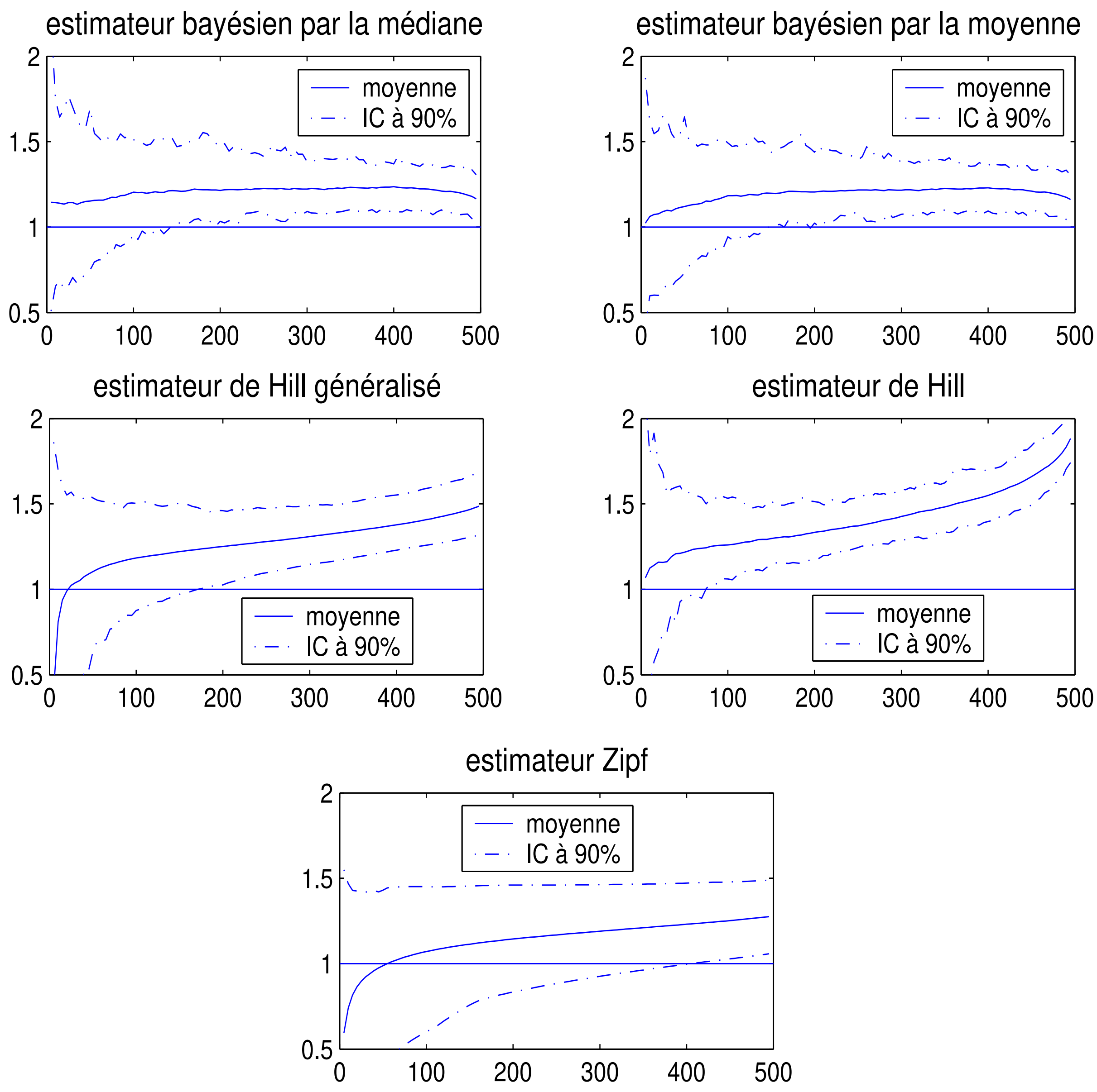

FIG. E.15 – *Échantillons simulés de loi $\mathcal{L}\Gamma(2)$ – Moyennes et intervalles à 90% des différents estimateurs de γ en fonction de m_n : estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

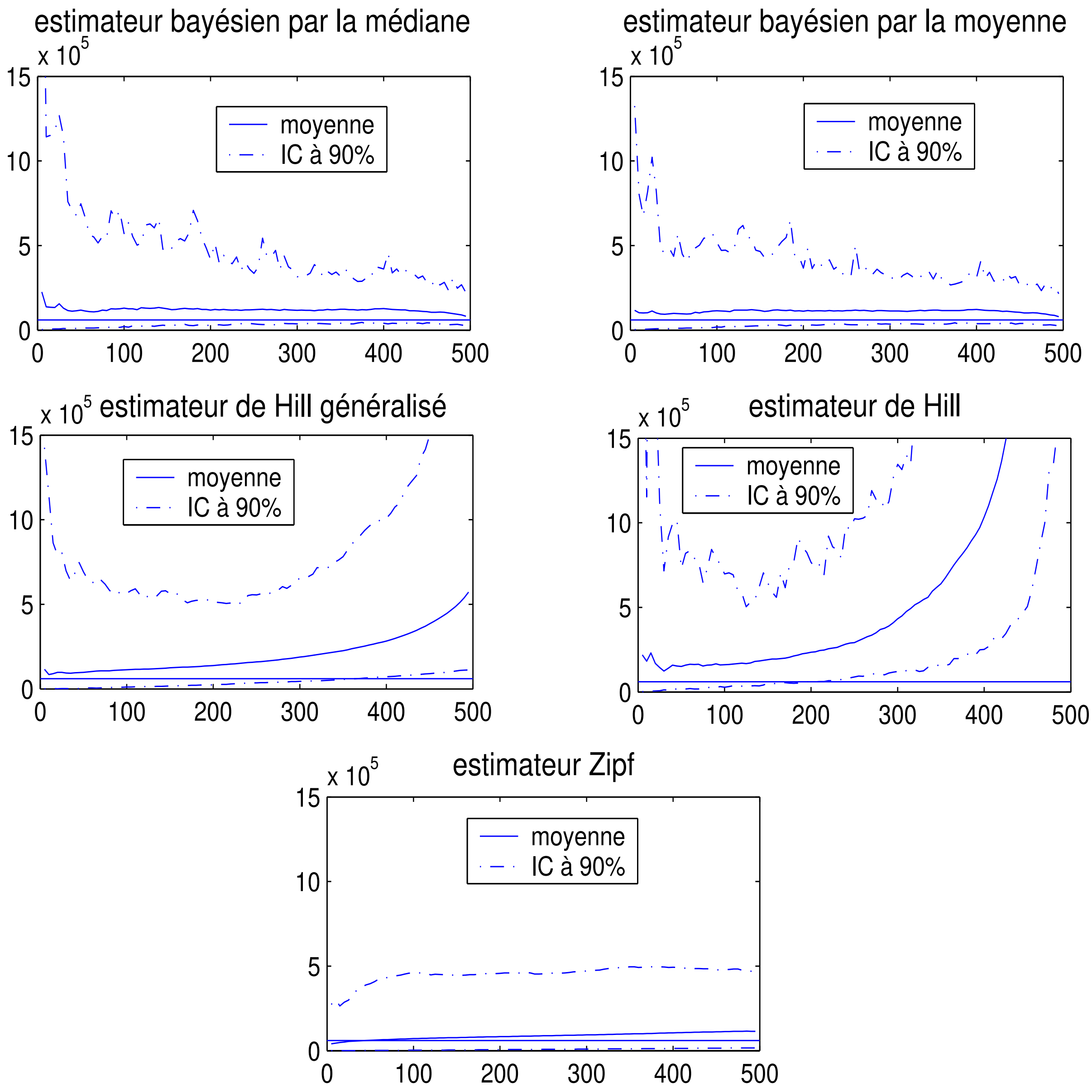

Fig. E.16 – *Échantillons simulés de loi $\mathcal{L}\Gamma(2)$ – Moyennes et intervalles à 90% des différents estimateurs de $q_{1-1/5000}$ en fonction de m_n : estimateurs bayésiens par la médiane a posteriori (en haut à gauche), estimateurs bayésiens par la moyenne a posteriori (en haut à droite), estimateurs de Hill généralisé (au centre à gauche), estimateurs de Hill (au centre à droite), et estimateurs Zipf (en bas).*

E.3 Échantillons d'excès avec avis d'expert – Intervalles de confiance empiriques pour les estimateurs de γ et de $q_{1-1/5000}$

Nous avons appliqué notre méthode bayésienne pour des échantillons d'excès (de taille m_n variant de 5 à 495), issus de 100 jeux de données simulées de taille $n = 500$, en introduisant un avis d'expert sur la queue de distribution concernant le quantile d'ordre $1 - 1/10n = 1 - 1/5000$ que nous souhaitons estimer. Nous avons comparé, au paragraphe 3.4.3 page 138, les moyennes des estimations du paramètre γ et du quantile d'ordre $1 - 1/10n = 1 - 1/5000$ de la loi des données originelles.

Dans ce paragraphe, on souhaite donner une idée de la variance des différents estimateurs calculés. Outre la valeur moyenne, on trace à présent un intervalle de confiance à 90%, pour les estimateurs de γ, et pour les estimateurs du quantile $q_{1-1/5000}$. Pour construire cet intervalle, pour chaque valeur du nombre d'excès m_n et chaque type d'estimateur, on considère l'échantillon des 100 valeurs de l'estimateur obtenues chacune pour l'un des 100 échantillons simulés. On ordonne ensuite cet échantillon, puis on ôte les deux plus petites et les deux plus grandes valeurs de cet échantillon ordonné. La borne inférieure de l'intervalle est alors la plus petite valeur restante de l'intervalle, et la borne inférieure la plus grande. On dispose ainsi, pour chaque valeur du nombre d'excès m_n et chaque type d'estimateur, d'une borne inférieure et d'une borne supérieure d'un intervalle de confiance empirique à 90%. On peut donc tracer, pour chaque type d'estimateur, des bornes de confiance en fonction de m_n.

Les différentes méthodes d'estimation que nous comparons sont : la méthode de Hill, la méthode de Hill généralisée, la méthode du Zipf, notre méthode bayésienne pour la moyenne a priori et la moyenne a posteriori, notre méthode bayésienne (avec avis d'expert cette fois-ci) pour la moyenne a priori et la médiane a posteriori. Nous les appliquons pour des échantillons de loi de Fréchet de paramètre 1, de loi de Burr de paramètres $(1, 1, 1)$ et $(1, 0.5, 2)$, de loi de Student absolue de paramètre 1, 2, 4, et 8, et de loi loggamma de paramètre 2. Ces échantillons sont les mêmes que ceux que l'on a utilisé pour la procédure bayésienne sans avis d'expert au paragraphe 3.3.3 page 128. Les valeurs moyennes, ainsi que les bornes de confiance sont donc inchangées pour les estimateurs Zipf, de Hill et de Hill généralisé. On ne reproduit donc pas ici les graphiques de bornes de confiance en fonction de m_n pour ces estimateurs, car ils sont déjà présentés dans l'annexe E.2 page 201. On ne présente dans cette annexe que les graphiques des bornes de confiance en fonction de m_n pour nos estimateurs bayésiens avec avis d'expert de la moyenne et la médiane a posteriori.

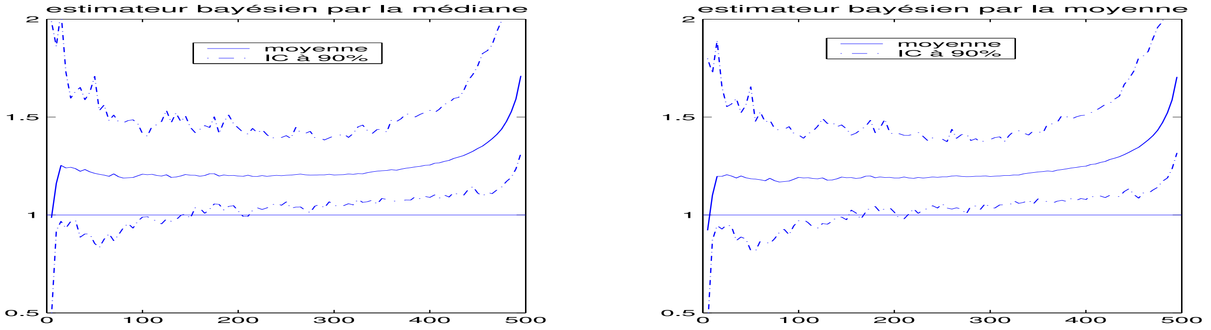

FIG. E.17 – *Échantillons simulés de loi* $\mathcal{LG}(2)$ – *Procédure bayésienne avec avis d'expert pour* α *fixé – intervalle de confiance à* 90% *pour les estimateurs bayésiens de* γ*, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

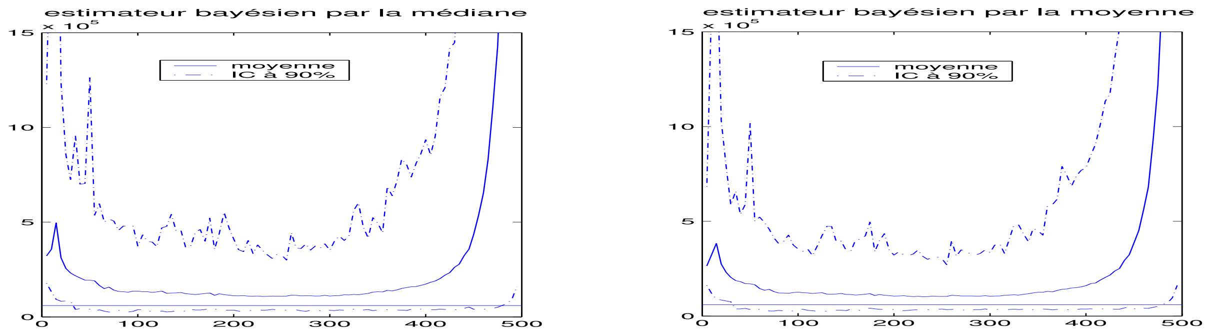

FIG. E.18 – *Échantillons simulés de loi* $\mathcal{LG}(2)$ – *Procédure bayésienne avec avis d'expert pour* α *fixé – intervalle de confiance à* 90% *pour les estimateurs bayésiens de* $q_{1-1/5000}$*, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

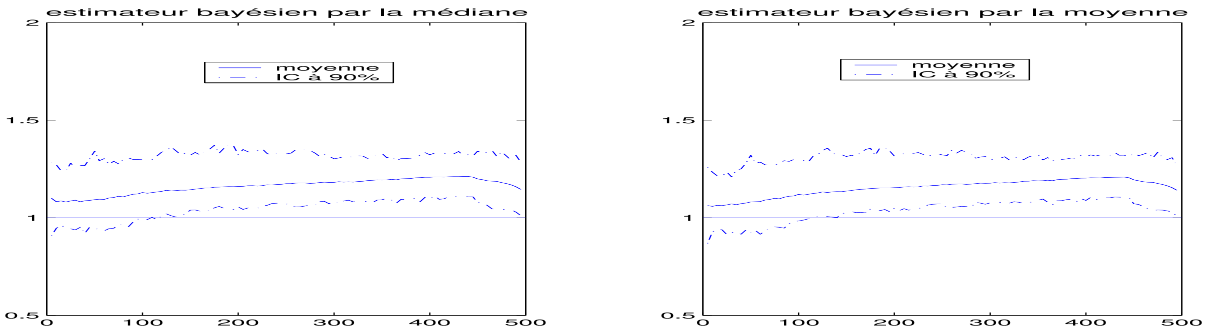

FIG. E.19 – *Échantillons simulés de loi* $\mathcal{LG}(2)$ – *Procédure bayésienne avec avis d'expert pour* β *fixé – intervalle de confiance à* 90% *pour les estimateurs bayésiens de* γ*, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

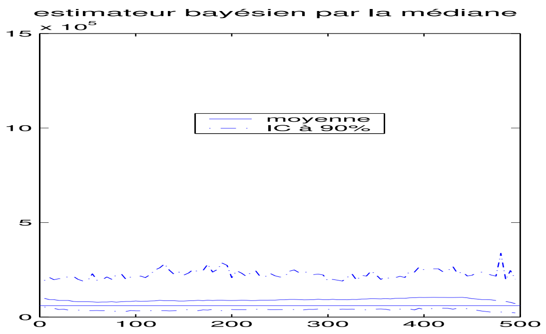

FIG. E.20 – *Échantillons simulés de loi* $\mathcal{LG}(2)$ *– Procédure bayésienne avec avis d'expert pour* β *fixé – intervalle de confiance à* 90% *pour les estimateurs bayésiens de* $q_{1-1/5000}$*, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

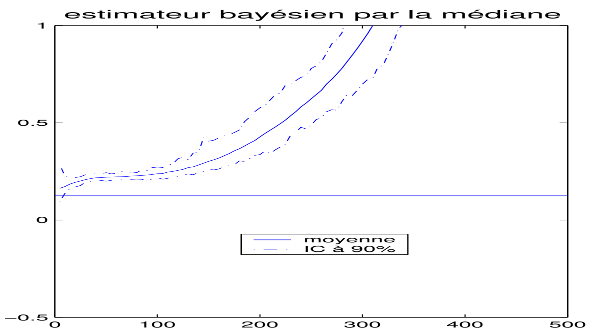

FIG. E.21 – *Échantillons simulés de loi* $t_{Abs}(8)$ *– Procédure bayésienne avec avis d'expert pour* α *fixé – intervalle de confiance à* 90% *pour les estimateurs bayésiens de* γ*, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

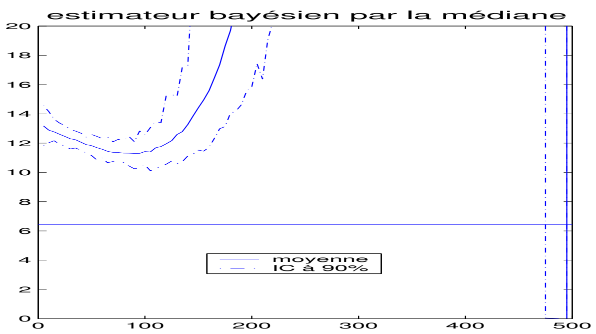

FIG. E.22 – *Échantillons simulés de loi* $t_{Abs}(8)$ *– Procédure bayésienne avec avis d'expert pour* α *fixé – intervalle de confiance à* 90% *pour les estimateurs bayésiens de* $q_{1-1/5000}$*, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

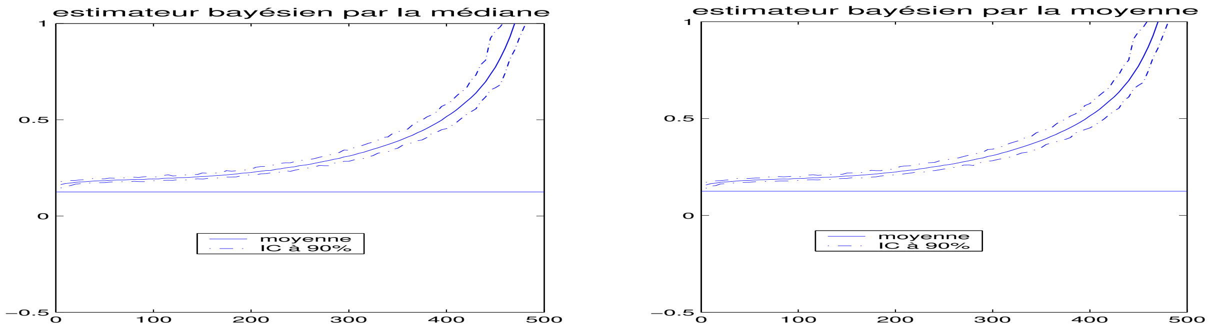

FIG. E.23 – *Échantillons simulés de loi $t_{Abs}(8)$ – Procédure bayésienne avec avis d'expert pour β fixé – intervalle de confiance à 90% pour les estimateurs bayésiens de γ, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

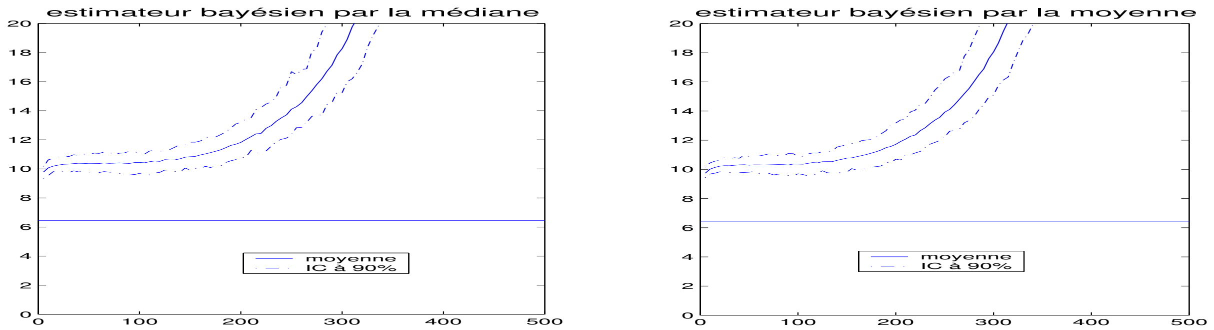

FIG. E.24 – *Échantillons simulés de loi $t_{Abs}(8)$ – Procédure bayésienne avec avis d'expert pour β fixé – intervalle de confiance à 90% pour les estimateurs bayésiens de $q_{1-1/5000}$, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

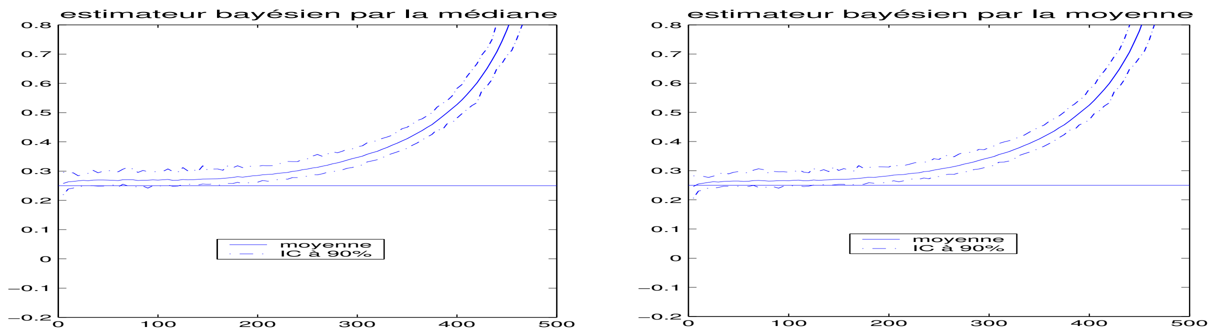

FIG. E.25 – *Échantillons simulés de loi $t_{Abs}(4)$ – Procédure bayésienne avec avis d'expert pour β fixé – intervalle de confiance à 90% pour les estimateurs bayésiens de γ, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

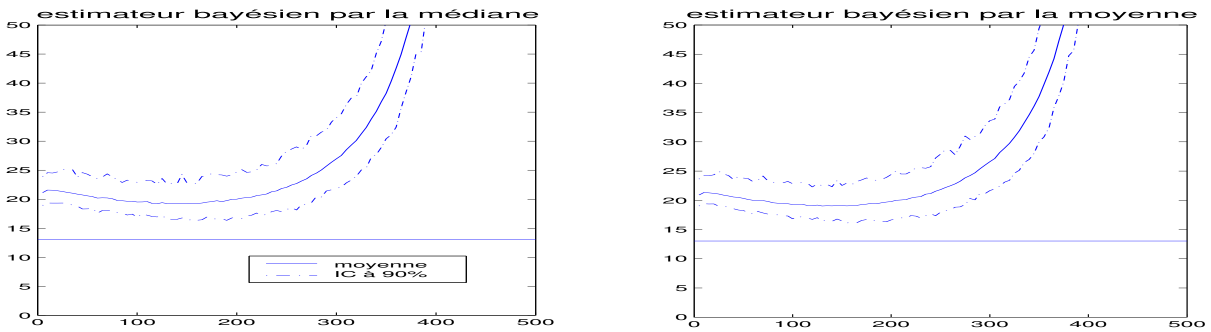

FIG. E.26 – *Échantillons simulés de loi* $t_{Abs}(4)$ – *Procédure bayésienne avec avis d'expert pour* β *fixé – intervalle de confiance à* 90% *pour les estimateurs bayésiens de* $q_{1-1/5000}$, *par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

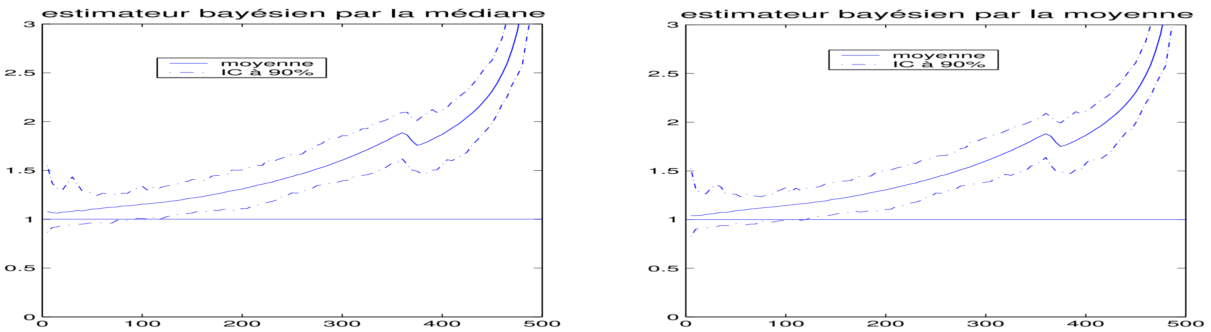

FIG. E.27 – *Échantillons simulés de loi* $\mathcal{B}urr(1, 0.5, 2)$ – *Procédure bayésienne avec avis d'expert pour* β *fixé – intervalle de confiance à* 90% *pour les estimateurs bayésiens de* γ, *par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

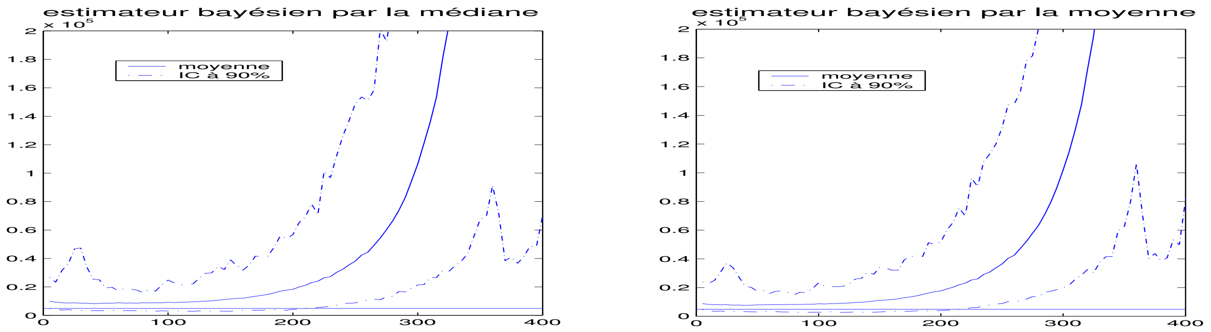

FIG. E.28 – *Échantillons simulés de loi* $\mathcal{B}urr(1, 0.5, 2)$ – *Procédure bayésienne avec avis d'expert pour* β *fixé – intervalle de confiance à* 90% *pour les estimateurs bayésiens de* $q_{1-1/5000}$, *par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

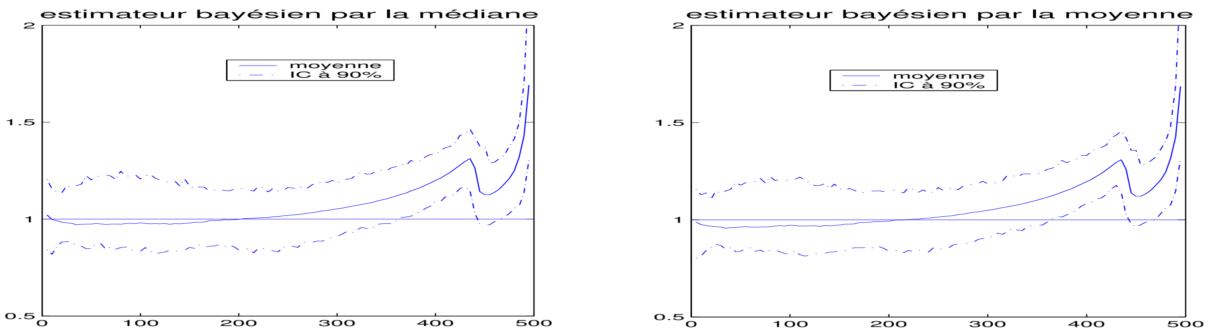

FIG. E.29 – *Échantillons simulés de loi* $\mathcal{B}urr(1,1,1)$ *– Procédure bayésienne avec avis d'expert pour* β *fixé – intervalle de confiance à 90% pour les estimateurs bayésiens de* γ*, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

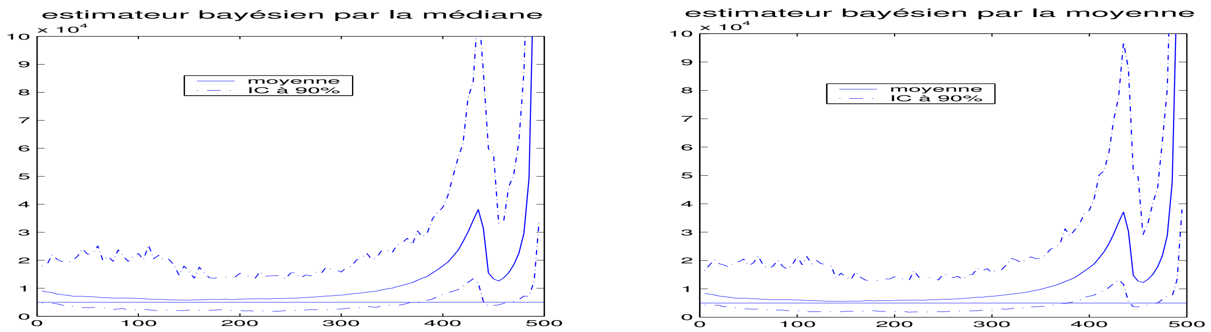

FIG. E.30 – *Échantillons simulés de loi* $\mathcal{B}urr(1,1,1)$ *– Procédure bayésienne avec avis d'expert pour* β *fixé – intervalle de confiance à 90% pour les estimateurs bayésiens de* $q_{1-1/5000}$*, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

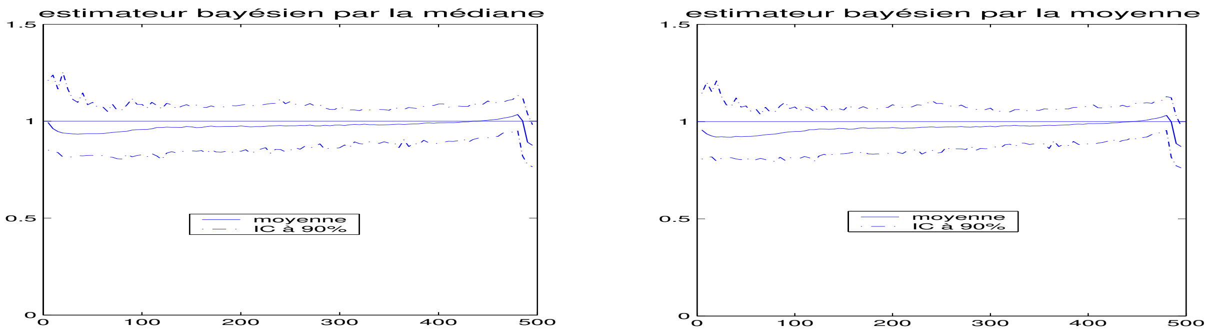

FIG. E.31 – *Échantillons simulés de loi* $\mathcal{F}rechet(1)$ *– Procédure bayésienne avec avis d'expert pour* β *fixé – intervalle de confiance à 90% pour les estimateurs bayésiens de* γ*, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

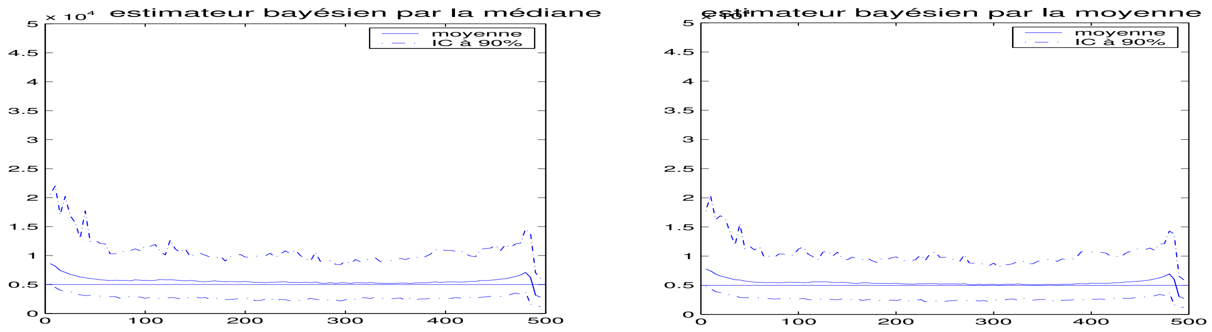

FIG. E.32 – *Échantillons simulés de loi $\mathcal{F}rechet(1)$ – Procédure bayésienne avec avis d'expert pour β fixé – intervalle de confiance à 90% pour les estimateurs bayésiens de $q_{1-1/5000}$, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

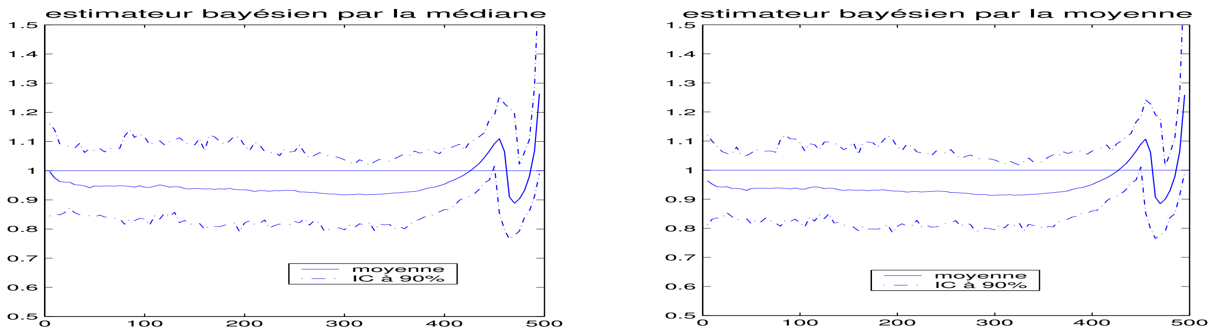

FIG. E.33 – *Échantillons simulés de loi $t_{Abs}(1)$ – Procédure bayésienne avec avis d'expert pour β fixé – intervalle de confiance à 90% pour les estimateurs bayésiens de γ, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

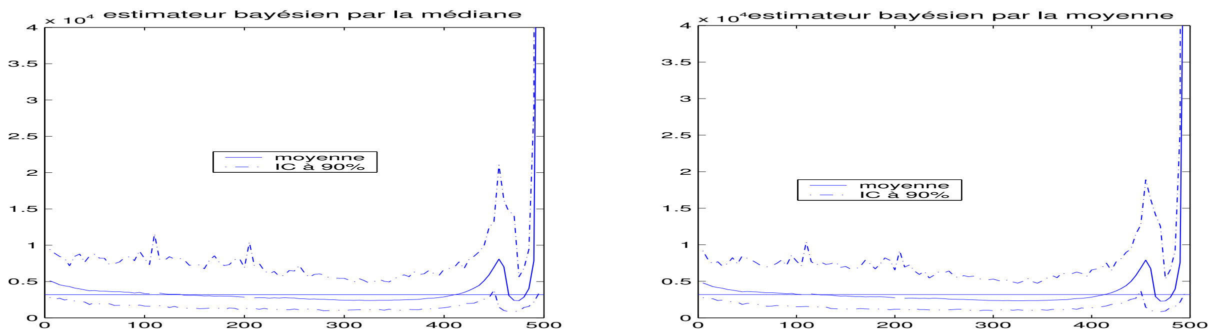

FIG. E.34 – *Échantillons simulés de loi $t_{Abs}(1)$ – Procédure bayésienne avec avis d'expert pour β fixé – intervalle de confiance à 90% pour les estimateurs bayésiens de $q_{1-1/5000}$, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

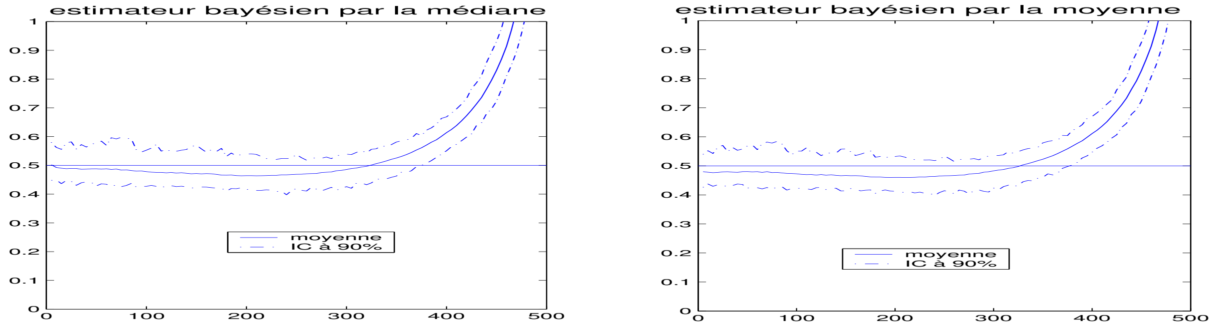

FIG. E.35 – *Échantillons simulés de loi* $t_{Abs}(2)$ – *Procédure bayésienne avec avis d'expert pour* β *fixé – intervalle de confiance à* 90% *pour les estimateurs bayésiens de* γ*, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

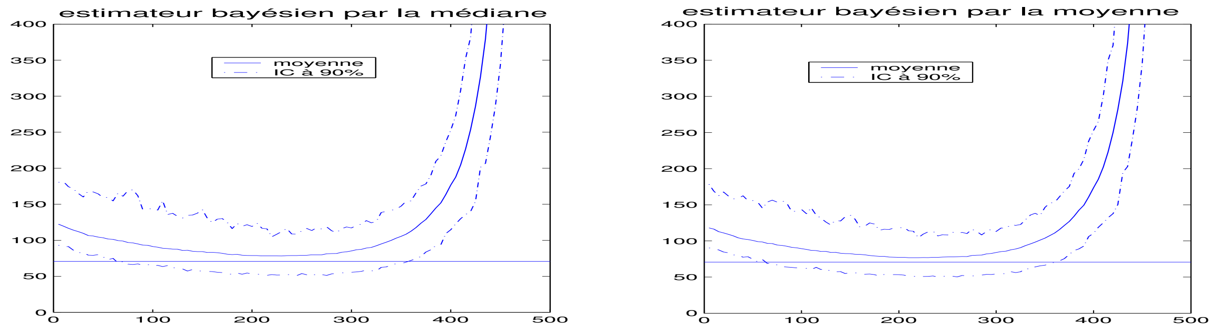

FIG. E.36 – *Échantillons simulés de loi* $t_{Abs}(2)$ – *Procédure bayésienne avec avis d'expert pour* β *fixé – intervalle de confiance à* 90% *pour les estimateurs bayésiens de* $q_{1-1/5000}$*, par la médiane a posteriori (à gauche) et la moyenne a posteriori (à droite).*

Publications

Rapports de recherche

- J. Diebolt, M. Garrido, S. Girard. Le test ET : test d'adéquation d'un modèle central à une queue de distribution, Rapport de Recherche Inria, n^{0}4170, avril 2001. http://www.inria.fr/rrrt/rr-4170.html
- J. Diebolt, M. Garrido, C. Trottier. A Bayesian Regularization Procedure for a Better Extremal Fit, Rapport de Recherche Inria n^{0}4211, juin 2001. http://www.inria.fr/rrrt/rr-4211.html

Conférences internationales

- M. Garrido, J. Diebolt. The ET test, a Goodness-of-Fit Test for the Distribution Tail, *MMR'2000 Methodology, Practice and Inference, Second International Conference on Mathematical Methods in Reliability*, pages 427–430, Bordeaux (France), 4–7 juillet 2000.
- M. Garrido, J. Diebolt, S. Girard. The ET test, a Goodness-of-Fit Test for the Distribution Tail, *Extremes 2001, International Symposium on Extreme Value Analysis: Theory and Practice*, Leuven (Belgique), 5–10 août 2001.
- M. Garrido, J. Diebolt, C. Trottier. A Bayesian Regularization Procedure for a Better Extremal Fit, *Extremes 2001, International Symposium on Extreme Value Analysis: Theory and Practice*, Leuven (Belgique), 5–10 août 2001.
- J. Diebolt, M. Garrido. A New Bayesian Approach to GPD's, *Extremes 2001, International Symposium on Extreme Value Analysis: Theory and Practice*, Leuven (Belgique), 5–10 août 2001.
- C. Bauby, D. Lagrange, J. Diebolt, M. Garrido. Goodness-of-Fit Test for Distribution Tails and Bayesian Regularization Procedure to Estimate Low Probabilities for System Decision, *lambdamu 13, ESREL 2002, Aide à la décision et maîtrise des risques*, pages 258–261, Lyon (France), 18–21 mars 2002.

- ▷ N. Devictor, M. Marques, S. Boulègue, P. Lamagnère, M-P. Valeta, M. Eid, M. Garrido. Analyse statistique de la ténacité : utilisation de méthodes de Monte-Carlo pour estimer l'incertitude sur la distribution de la ténacité, *lambdamu 13, ESREL 2002, Aide à la décision et maîtrise des risques*, pages 480–483, Lyon (France), 18–21 mars 2002.

Conférences nationales

- ▷ M. Garrido, J. Diebolt. Le test ET : test d'adéquation d'un modèle central à une queue de distribution, *XXXIIIèmes Journées de Statistique*, pages 363–366, Nantes (France), 14–18 mai 2001.
- ▷ M. Garrido, J. Diebolt, S. Girard. Une nouvelle approche bayésienne pour l'estimation des paramètres d'une loi GPD, *XXXIVèmes Journées de Statistique*, Bruxelles-Louvain (Belgique), 13–17 mai 2002.

Séminaires - Groupes de travail

- ▷ Les différentes versions du test ET, un test d'adéquation pour la queue de distribution, *Groupe de travail Fiabilité*, Université de Marne-la-Vallée, 28 janvier 2000.
- ▷ Un test d'adéquation d'une loi de probabilité aux plus grandes observations, *Journée sur l'estimation des quantiles extrêmes*, INRIA Rhône-Alpes, 23 mars 2000.
- ▷ Prédiction des événements rares : introduction d'un test d'adéquation et d'une procédure de régularisation bayésienne, *Demi-journée "innovation" SDM*, Électricité de France (EDF), 10 octobre 2000.
- ▷ Le test ET : test d'adéquation à la queue de distribution pour un modèle central, *Séminaire SMS (Statistique et Modélisation Stochastique)*, Université Joseph-Fourier (IMAG–LMC) & Institut National Polytechnique de Grenoble & Université Pierre Mendès-France (LABSAD) & INRIA Rhône-Alpes (projet IS2), 4 janvier 2001.
- ▷ Estimation bayésienne des paramètres d'une loi de Pareto généralisée, *Groupe de travail Modèles Aléatoires pour la Fiabilité et la Maintenance des Systèmes*, IMAG (Laboratoire de Modélisation et Calcul) & INRIA Rhône-Alpes (projet IS2), 4 octobre 2001.
- ▷ Journée d'initiation à la maquette logiciel EXTREME : présentation des méthodes implémentées (test ET et régularisation bayésienne); présentation de la maquette; expérimentation de la maquette par les auditeurs, *Journée d'initiation à la maquette logiciel* EXTREME, EDF, 28 mai 2002.

Bibliographie

[1] Balkema, A. et de Haan, L. – Residul life time at great age. *the Annals fo Probability*, vol. 2, n° 5, 1974, pp. 792–804.

[2] Balkema, A. et de Haan, L. – Limit distributions for order statistics. i. *SIAM Theory of Probability and its Applications*, vol. 23, n° 1, 1978, pp. 77–92.

[3] Balkema, A. et de Haan, L. – Limit distributions for order statistics. ii. *SIAM Theory of Probability and its Applications*, vol. 23, n° 2, 1978, pp. 341–358.

[4] Barbe, P. et Diebolt, J. – *Empirical process of excesses above random thresholds.* – Rapport de Recherche n° 08-2000, Université de Marne-la-Vallée, 2000.

[5] Beirlant, J., Dierckx, G., Goegebeur, Y. et Matthys, G. – Tail index estimation and an exponential regression model. *Extremes*, vol. 2, n° 2, 1999, pp. 177–200.

[6] Beirlant, J., Dierckx, G. et Guillou, A. – Estimation of the extreme value index and regression on generalized quantile plots. 2002. – soummis.

[7] Belkacem, A. – L_2 version of the double kernel method. *Statistics - A Journal of Theoretical and Applied Statistics*, vol. 32, n° 3, 1999, pp. 249–266.

[8] Bingham, N.H., Goldie, C.M. et Teugels, J.L. – *Regular Variation.* – Cambridge University Press, 1987, *Encyclopedia of Mathematics and its application*, volume 27.

[9] Bosq, D. et Lecoutre, J.P. – *Théorie de l'estimation fonctionnelle.* – Paris, Economica, 1987, *Economie et Statistiques avancées.*

[10] Breiman, L., Stone, C.J. et Kooperberg, C. – Robust confidence bounds for extreme upper quantiles. *Journal of Statistical Computation and Simulation*, vol. 37, 1990, pp. 127–149.

[11] Chauveau, D. et Diebolt, J. – *An automated stopping rule for MCMC convergence assessment.* – Rapport de Recherche n° RR-3566, Inria, 1998.

[12] Chauveau, D. et Diebolt, J. – An automated stopping rule for mcmc convergence assessment. *Computationnal Statistics*, vol. 14, n° 3, 1999, pp. 419–442.

[13] Chib, S. et Greenberg, E. – Understanding the Metropolis-Hastings algorithm. *The American Statistician*, vol. 49, n° 4, novembre 1995, pp. 327–335.

[14] Coles, S. G. et Dixon, M. J. – Likelihood-based inference for extreme value models. *Extremes*, vol. 2, 1999, pp. 5–23.

[15] Coles, S. G. et Powell, E. A. – Bayesian methods in extreme value modelling. *International Statistical Review*, vol. 64, 1996, pp. 119–136.

[16] Coles, S. G. et Tawn, J. A. – A bayesian analysis of extreme data rainfalls. *Applied Statistics*, vol. 45, 1999, pp. 463–478.

[17] D'Agostino, R.B. et Stephens, M.A. – *Goodness-of-fit Techniques.* – New York and Basel, Marcel Dekker, 1986, *Statistics textbooks and monographs*, volume 68.

[18] Damsleth, E. – Conjugate classes for gamma distributions. *Scandinavian Journal of Statistics, Theory and Applications*, vol. 2, 1975, pp. 80–84.

[19] Davis, R. et Resnick, S. – Tail estimates motivated by extreme value theory. *The Annals of Statistics*, vol. 12, n° 4, 1984, pp. 1467–1487.

[20] Davison, A. et Smith, R. – Models for exceedances over high thresholds. *Journal of Royal Statistical Society, B*, vol. 52, n° 3, 1990, pp. 393–442.

[21] de Haan, L. et Rootzen, H. – On the estimation of high quantiles. *Journal of Statistical Planning and Inference*, vol. 35, n° 1, 1993, pp. 1–13.

[22] Dekkers, A.L.M., Einmahl, J.H.J. et de Haan, L. – A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, vol. 17, n° 4, 1989, pp. 1833–1855.

[23] Diebolt, J., Durbec, V., El Aroui, M. A. et Villain, B. – Estimation of extreme quantiles: empirical tools for method assessment and comparison. *International Journal of Reliability, Quality and Safety Engineering*, vol. 7, n° 1, 2000, pp. 75–94.

[24] Diebolt, J. et El Aroui, M.A. – *Modélisation des queues de distributions et estimation de quantiles extrêmes (1).* – Rapport de contrat, INRIA–EDF, 1997.

[25] Diebolt, J. et Garrido, M. – *Les différentes versions du test ET, test d'adéquation à la queue de distribution.* – Rapport de contrat, INRIA–EDF, octobre 1999.

[26] Diebolt, J. et Garrido, M. – *Estimation bayésienne de la loi GPD, loi asymptotique des excès au delà d'un seuil.* – Rapport de contrat, INRIA–EDF, octobre 2001.

[27] Diebolt, J., Garrido, M. et Trottier, C. – *A bayesian regularization procedure for a better extremal fit.* – Rapport de Recherche n° RR-4211, Inria, 2001. http://www.inria.fr/rrrt/rr-4211.html.

[28] Diebolt, J. et Girard, S. – *Consistency of the ET method and smooth variations.* – Rapport de Recherche n° 98-08, Université Montpellier 2, 1998.

[29] Diebolt, J. et Girard, S. – *Modélisation des queues de distributions et estimation de quantiles extrêmes (2).* – Rapport de contrat, INRIA–EDF, 1998.

[30] Diebolt, J. et Girard, S. – *On the Convergence of the ET Method for Extreme Upper Quantile Estimation.* – Rapport de Recherche n° RR-3389, Inria, 1998.

[31] Diebolt, J. et Girard, S. – Consistency of the ET method and smooth variations. *Comptes Rendus de l'Académie des Sciences de Paris, Série I, Mathematique*, vol. 329, 1999, pp. 821–826.

[32] Diebolt, J. et Trottier, C. – *Régularisation de distribution pour une meilleure adéquation extrême.* – Rapport de contrat, INRIA–EDF, 1999.

[33] Ditlevsen, O. – Distribution Arbitrariness in Structural Reliability. *In: Structural Safety and Reliability*, éd. par Schuller, Shinozuka et Yao, pp. 1241–1247. – Rotterdam, Balkema, 1994.

[34] Embrechts, P., Klüppelberg, C. et Mikosh, T. – *Modelling Extremal Events.* – Springer-Verlag, 1997, *Applications of Mathematics*, volume 33.

[35] Galambos, J. – *The Asymptotic Theory of Extreme Order Statistics.* – R.E. Krieger publishing compagny, 1987.

[36] Hahn, G. et Meeker, W. – Pitfalls and practical considerations in product life analysis, part 1: Basic concepts and dangers of extrapolation. *Journal of Quality Technology*, vol. 14, 1982.

[37] Hosking, J. et Wallis, J. – Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, vol. 29, n° 3, 1987, pp. 339–349.

[38] Luke, Y. L. – *Algorithms for the Computation of Mathematical Functions.* – 1977, *Computer Science and Applied Mathematics.*

[39] Müller, P. – *A generic approach to posterior integration and Gibbs sampling.* – Tech. Report n° 91-09, West Lafayette, Indiana, Purdue Univeristy, 1991.

[40] Müller, P. – *Alternatives to the Gibbs sampling scheme.* – Tech. report, Intitute of Statistics and Decision Science, Duke University, 1992.

[41] Pickands, J. – Statistical inference using extreme order statistics. *The Annals of Statistics*, vol. 3, 1975, pp. 119–131.

[42] Press, W., Flannery, B., Teukolsky, S. et Vetterling, W. – *Numerical Recipes : the Art of Scientific Computing.* – New York, Cambridge University Press, 1986.

[43] Ramdani-Worms, R. – *Vitesse de convergence pour l'approximation des queues de distributions.* – Thèse de doctorat, PhD, Université de Marne-la-Vallée, décembre 2000.

[44] Rao, C.R. – *Linear Statistical Inference and its Applications.* – J. Wiley and sons, 1973, *Wiley series in probability and mathematical statistics.*

[45] Robert, C. – *Méthodes de Monte Carlo par chaînes de Markov.* – Paris, Economica, 1996, *Statistique mathématique et probabilité.*

[46] Robert, C. – *Discretization and MCMC Convergence Assessment.* – Spinger, 1998, *Lecture Notes in Statistics.*

[47] Spouge, J. L. – Computation of the gamma, digamma, and trigamma functions. *SIAM Journal on Numerical Analysis*, vol. 31, n° 3, 1994, p. 931.

[48] Vieu, P. – A note on density mode estimation. *Statistics and Probability Letters*, vol. 26, 1996, pp. 297–307.

[49] Worms, R. – Penultimate approximation of the distribution of the excesses. *European Series in Applied and Industrial Mathematics : Probability and Statistics*, 2002. – accepté, en révision.

TITLE:

Modelling of rare events and estimation of extreme quantiles, model selecting methods for distribution tails

---

ABSTRACT:

This PhD. work deals with the modelling of rare event and the estimation of extreme quantiles through different models, and with the selection of these models. Extreme value theory allows for nonparametric estimation of distribution tails. Since these estimations are biased, we use usual parametric models. However, these models are estimated and selected by procedures involving the whole sample. The outputs are thus driven by the most likely values of the variables. Therefore, we propose two goodness-of-fit tests for distribution tail, the ET (Exponential Tail) test and the GPD (Generalised Pareto Distribution) test, to select models providing good estimations of the tail (comparing to the POT method). When we wish to have a good picture of the distribution both for the most likely values and for the extreme values, we can first apply a usual goodness-of-fit test to a set of models and then an adequacy test for the distribution tail. When no distribution is accepted by both tests, we propose a Bayesian regularisation procedure which improves the tail fit of a central model (adapted to the most likely values), thanks to an expert opinion concerning the distribution tail. Finally, when we wish to use the POT method, we have to reduce its estimation bias, especially in extreme quantile estimation. Since POT is based on the approximation of the distribution of excesses over a threshold by a GPD distribution, we try to produce a better estimation of this distribution through a Bayesian procedure. The Bayesian estimation of the GPD parameters leads to a reduced estimation bias for extreme quantiles computed from POT, in particular when we introduce an expert opinion concerning the distribution tail.

KEY WORDS:

extreme value theory, upper quantile, bias reduction, Peaks over Threshold method, exponential tail, goodness-of-fit test, Bayesian Statistics, conjugate distribution, mixture distribution.

RESUME :
Cette thèse étudie la modélisation d'événements rares et l'estimation de quantiles extrêmes, à travers différents types de modèles et le choix de ces modèles. La théorie des valeurs extrêmes, et en particulier la méthode des excès (POT, *Peaks Over Threshold*), permettent une estimation non paramétrique, mais biaisée, des queues de distribution. Nous souhaitons donc utiliser des modèles paramétriques classiques. Cependant, ces modèles étant estimés et sélectionnés par des tests usuels à partir de l'échantillon complet, les résultats sont surtout influencés par les valeurs les plus probables de la variable. Nous proposons deux tests d'adéquation pour la queue de distribution, le test ET (*Exponential Tail*) et le test GPD (*Generalised Pareto Distribution*), pour sélectionner, par comparaison avec la méthode POT, les modèles produisant de bonnes estimations de la queue de distribution. Lorsqu'on souhaite reconstituer la loi dont sont issues les observations aussi bien dans la région centrale que dans la région extrême, on applique d'abord à un ensemble de modèles un test usuel (d'adéquation aux valeurs les plus probables), puis un test d'adéquation de la queue de distribution. Si aucune loi n'est acceptée par les deux types de tests, nous proposons une procédure de régularisation bayésienne qui, à partir d'un modèle adapté aux valeurs les plus probables, permet d'améliorer l'adéquation extrême grâce à un avis d'expert sur la queue de distribution. Enfin, si on revient à la méthode POT, il faut en réduire le biais d'estimation, notamment pour l'estimation des quantiles extrêmes. Cette méthode étant fondée sur l'approximation de la loi des excès au-delà d'un seuil par une loi GPD, nous cherchons à mieux en estimer les paramètres. L'inférence bayésienne sur les paramètres de la loi GPD permet de réduire le biais d'estimation des quantiles extrêmes par la méthode POT, en particulier quand on introduit un avis d'expert sur la queue de distribution.

MOTS-CLES :
théorie des valeurs extrêmes, réduction de biais, méthode des excès, queue exponentielle, test d'adéquation, statistique bayésienne, loi a priori conjuguée, loi de mélange.

DISCIPLINE : Mathématiques Appliquées

Thèse réalisée conjointement dans le laboratoire LMC de l'université Joseph Fourier et au sein du projet IS2 de l'INRIA Rhône-Alpes ; co-financée par l'INRIA et EDF.