LE LOGICIEL EXTREMES, UN OUTIL POUR L'ETUDE DES QUEUES DE DISTRIBUTION

Jean Diebolt[1], Jérôme Ecarnot[2], Myriam Garrido[3], Stéphane Girard[2,4], Dominique Lagrange[5]

[1] *CNRS, Université de Marne la Vallée, 5 bd. Descartes, 77454 Marne la Vallée Cedex 2 (jean.dieb@wanadoo.fr)*
[2] *IS2, INRIA Rhône-Alpes, 655 av. de l'Europe, 38330 Montbonnot Saint Martin (jerome.ecarnot@inrialpes.fr)*
[3] *Labsad, Université Grenoble 2, BP 47, 38040 Grenoble Cedex 9 (myriam.garrido@upmf-grenoble.fr)*
[4] *LMC-IMAG, Université Grenoble 1, BP 53, 38041 Grenoble Cedex 9 (stephane.girard@imag.fr)*
[5] *EDF R&D/MRI, 6 quai Watier, BP 49, 78401 Chatou Cedex (dominique.lagrange@edf.fr)*

Résumé *: Le logiciel EXTREMES regroupe différents outils pour la modélisation des queues de distribution et l'estimation des quantiles extrêmes. On y trouve en particulier les procédures classiques d'estimation des paramètres des lois décrivant le comportement des valeurs extrêmes, mais aussi des procédures plus complexes de test d'adéquation pour la queue de distribution. Des fonctions d'inférence statistique classique sont aussi implémentées, permettant ainsi la comparaison de modèles paramétriques centraux avec des modèles semi paramétriques extrêmes. Le logiciel est écrit en C++, comporte une interface graphique développée en Matlab et une documentation technique. Le tout est disponible auprès des auteurs.*

Mots-clés *: Queue de distribution, quantile extrême, tests d'adéquation, statistique bayésienne.*

1. Introduction

Le logiciel ***EXTREMES*** regroupe différents outils dédiés à l'étude des valeurs extrêmes tels que des procédures d'estimation des quantiles extrêmes et de sélection de modèles pour les queues de distribution. Il est le fruit d'une collaboration entre l'équipe IS2 de l'INRIA Rhône-Alpes et la division Recherche et Développement d'EDF, et l'aboutissement des travaux de thèse de Myriam Garrido [GAR02]. Il ne s'adresse pas uniquement aux spécialistes des valeurs extrêmes, même s'il offre de nouveaux outils pour l'étude des queues de distribution.

Dans le paragraphe 2, nous décrivons le contexte mathématique permettant l'étude des évènements rares, dans le paragraphe 3 sont exposées les fonctionnalités du logiciel proprement dites et les aspects informatiques sont abordés dans le paragraphe 4.

2. Fondements théoriques

La théorie des valeurs extrêmes [EMB97] a été développée pour l'estimation de probabilités d'occurrences d'évènements rares. Elle permet d'extrapoler le comportement de la queue de distribution à partir des plus grandes données observées. Le résultat suivant sur la loi des valeurs extrêmes 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 variables aléatoires indépendantes et identiquement distribuées correctement normalisé à l'aide de deux suites α_n et β_n.

Soit F la fonction de répartition de la loi d'intérêt. Sous certaines conditions de régularité sur F, il existe $\tau \in R$ et deux suites normalisantes α_n et β_n tels que :

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

où H_τ est la fonction de répartition de la loi des valeurs extrêmes :

$$si\ \tau \neq 0,\ H_\tau(x) = \exp\left(-(1+\tau x)_+^{-1/\tau}\right)$$

$$si\ \tau = 0,\ H_0(x) = \exp(-\exp(-x))$$

et où la notation y_+ désigne $\max(y,0)$.

On dit alors que le fonction de répartition F est dans le domaine d'attraction de Fréchet, de Gumbel ou de Weibull selon que $\tau > 0, \tau = 0$ ou $\tau < 0$.

Une 2ème méthode d'estimation de queues de distribution est la méthode des excès ou POT (Peaks over threshold), introduite dans [PIC75]. Soit u un réel suffisamment grand appelé seuil. La méthode des excès s'appuie sur l'approximation de la loi des excès au-dessus du seuil u de la variable aléatoire X, c'est-à-dire de la loi conditionnelle de la variable aléatoire $X-u$ sachant que $X > u$. La fonction de répartition des excès est définie par :

$$F_u(y) = P(X - u < y \,/\, X > u).$$

D'après le théorème de Pickands, si F appartient à l'un des 3 domaines d'attraction de la loi des valeurs extrêmes, la fonction de répartition F_u peut être approchée par une loi de Pareto généralisée (GPD) définie pour $\sigma > 0$ par :

$$si\gamma \neq 0 \ , \ G_{\gamma,\sigma}(x) = 1 - \left(1 + \frac{\gamma x}{\sigma}\right)_+^{-1/\gamma}$$
$$si\gamma = 0 \ , \ G_{0,\sigma}(x) = 1 - \exp(-x/\sigma).$$

Sur la base de ces résultats, il est possible d'estimer des quantiles extrêmes. Un quantile extrême q_n d'ordre $(1 - p_n)$ est défini par l'équation $F(q_n) = 1 - p_n$ avec $p_n \leq 1/n$, n désignant la taille de l'échantillon. Un tel quantile étant généralement situé au-delà de l'observation maximale, des techniques spécifiques d'estimation sont nécessaires. La méthode POT s'appuie sur le théorème de Pickands pour estimer q_n par :

$$\hat{q}_n = u_n + \frac{\hat{\sigma}_n}{\hat{\gamma}_n}\left(\left(\frac{np_n}{k_n}\right)^{-\hat{\gamma}_n} - 1\right) \qquad [1]$$

où k_n désigne le nombre d'excès au-delà du seuil u_n et $\hat{\sigma}_n$ et $\hat{\gamma}_n$ sont des estimateurs des paramètres de la loi GPD. Pour ces derniers, de nombreuses propositions existent, voir par exemple [EMB97].

3. Fonctionnalités

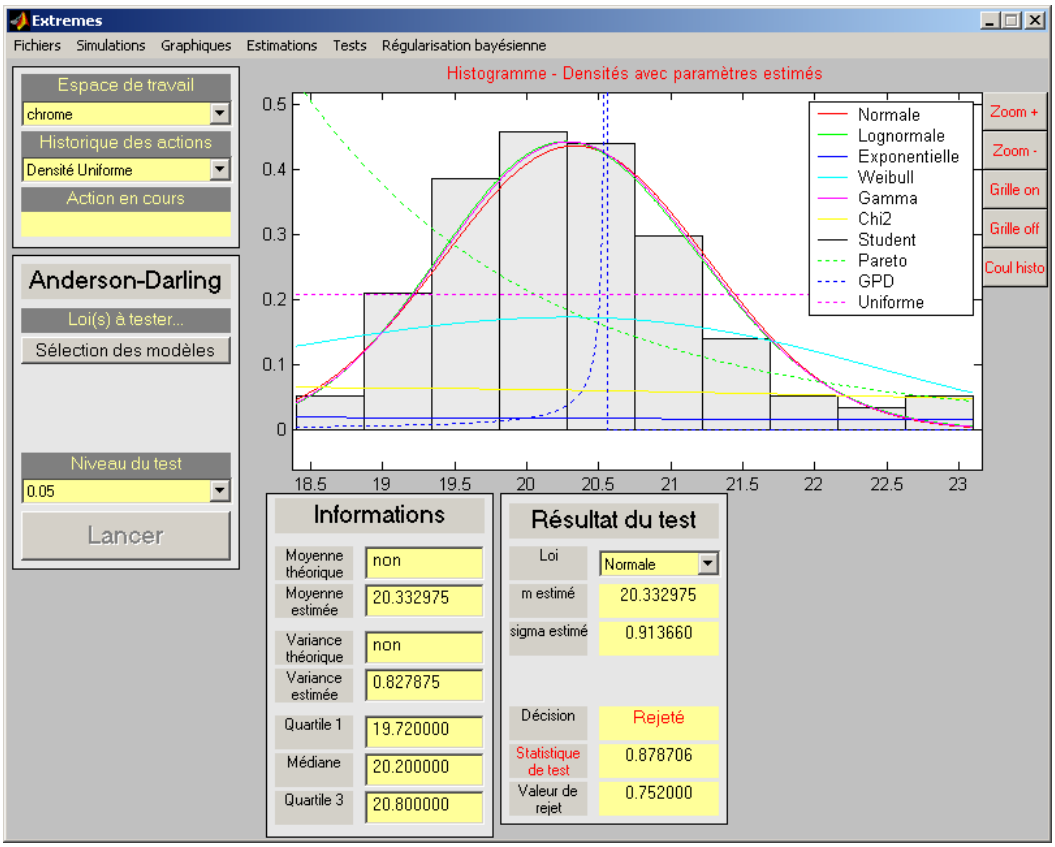
Les fonctions disponibles sont regroupées en 3 catégories.

3.1. Fonctions statistiques classiques

Les fonctions ci-dessous sont d'intérêt général au sens où elles ne sont pas dédiées à l'étude des valeurs extrêmes :

- Simulation de variables aléatoires de lois Normale, Lognormale, Exponentielle, Gamma, Weibull, Chi2, Student, Pareto, Beta, Uniforme et Pareto généralisée.
- Graphique des densités, fonctions de répartition, fonctions de survie, fonctions quantiles des lois précitées.
- Estimation des paramètres des lois précitées.
- Estimation non paramétrique de la densité (méthode de noyau, histogramme).
- Estimation paramétrique des quantiles.
- Test d'Anderson-Darling et Cramer-Von Mises.

Exemple : Illustration du fonctionnement du test d'Anderson-Darling sur le fichier de données réelles *cr.txt* (fourni avec le logiciel, voir paragraphe 4). L'adéquation des lois Normale, Lognormale, Exponentielle, Weibull, Gamma, Chi2, Student, Pareto, GPD et Uniforme a été testée au niveau 5%. Dans la fenêtre *Résultat du test*, le logiciel indique les valeurs des paramètres estimés, la statistique du test, la valeur de rejet et la décision. Les densités correspondantes sont superposées sur l'histogramme des données.



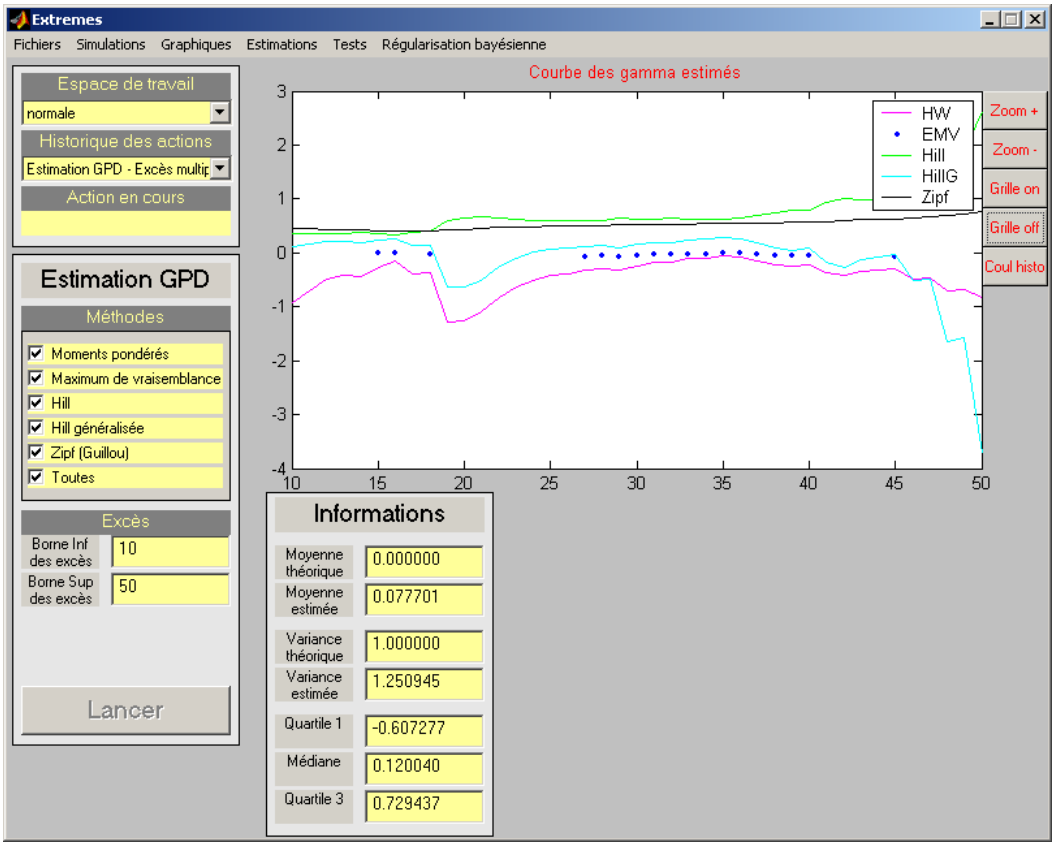
3.2. Fonctions extrêmes classiques

Nous regroupons ici les fonctions d'estimation et de test bien connues dans le domaine de la statistique des valeurs extrêmes.

- Vérification de l'exponentialité des excès : il s'agit de s'assurer que la fonction de répartition des données étudiées est dans le domaine d'attraction de Gumbel, et que le nombre d'excès k_n est convenablement choisi. L'ajustement de la loi Exponentielle aux excès est contrôlé graphiquement en traçant un QQ-plot. Un test d'exponentialité des excès est également proposé.

- Estimation des paramètres de la loi GPD. Sont regroupées ici plusieurs méthodes classiquement utilisées pour estimer le couple (γ, σ), notamment les méthodes de Hill, Hill généralisé, Moments pondérés d'Hosking et Wallis, Maximum de vraisemblance et Zipf. On pourra se reporter à [EMB97] et aux références indiquées dans l'ouvrage pour plus de détails.
- Estimation des quantiles extrêmes. Cette estimation s'appuie sur l'équation [1] et l'estimation des paramètres précédents.

Exemple : Estimation du paramètre γ (indice de valeurs extrêmes) sur un échantillon issu d'une loi normale centrée réduite de taille 100. Pour chaque valeur du nombre d'excès k_n comprise entre 10 et 50 en abscisse, on place en ordonnée la valeur de l'estimateur obtenu. Sur cet exemple, les 5 estimateurs disponibles ont été utilisés. Remarquons que dans cet exemple, la valeur théorique de γ est 0 (bien sûr indépendante du nombre d'excès), la loi Normale étant dans le domaine d'attraction de Gumbel.



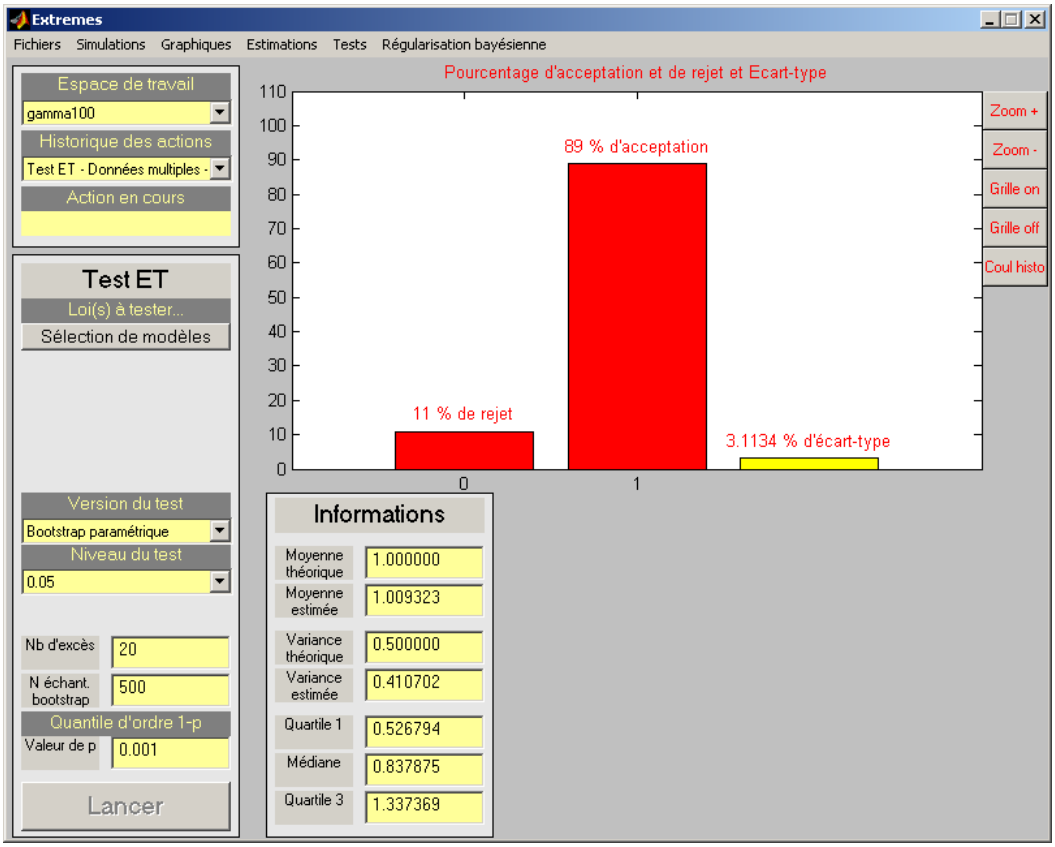
3.3. Procédures introduites dans [GAR02]

Il s'agit de la partie la plus innovante du logiciel. Les fonctions rassemblées ici ont été intégralement développées dans le cadre d'une thèse co-financée par INRIA Rhône-Alpes et EDF :

- Test ET,
- Test GPD,
- Régularisation bayésienne.

Le test ET et le test GPD sont 2 tests d'adéquation pour la queue de distribution. Ils sélectionnent par comparaison avec la méthode POT les modèles centraux produisant de bonnes estimations de la queue de distribution. Lorsqu'on souhaite reconstituer la loi des observations aussi bien dans la région centrale qu'extrême, on applique d'abord à un ensemble de modèles un test usuel (Anderson-Darling ou Cramer-Von Mises) puis un test d'adéquation de la queue de distribution (ET ou GPD). Si aucune loi n'est acceptée par les deux types de tests, la procédure de régularisation bayésienne permet, à partir d'un modèle adapté aux valeurs les plus probables, d'améliorer l'adéquation extrême grâce à un avis d'expert sur la queue de distribution.

Exemple : Estimation de la puissance du test ET sur données simulées. Dans cet exemple, 100 échantillons de taille 100 sont simulés à partir d'une loi Gamma. Sur chacun des jeux de données, la version bootstrap paramétrique du test ET est utilisée pour tester l'adéquation de la queue de distribution à une loi de Weibull. Le test accepte à tort cette hypothèse dans 89% des cas. Cette faible puissance s'explique par la similitude entre les queues des lois Gamma et Weibull dans le cas d'un paramètre de forme proche de 1.



4. Utiliser EXTREMES

Les sources du logiciel EXTREMES sont écrites en langage C++ et une interface graphique a été développée sous Matlab de façon à allier rapidité d'exécution et convivialité. Le logiciel accompagné de quelques jeux de données est disponible (sous forme compilée) à l'adresse suivante :

http://www.inrialpes.fr/is2/pub/software/EXTREMES/accueil.html

Une documentation aux formats Word, Postscript et HTML est fournie à la même adresse.

REFERENCES BIBLIOGRAPHIQUES

[EMB97] Embrechts P., Klüppelberg C., Mikosh T., Modelling extremal events – Springer Verlag, Applications of mathematics, 1997.

[GAR02] Garrido M., Modélisation des évènements rares et estimation des quantiles extrêmes, Méthodes de sélection de modèles pour les queues de distribution, Thèse de doctorat, Université Grenoble 1, 2002.

[PIC75] Pickands J., « Statistical inference using extreme order statistics », *The Annals of statistics, vol. 3,* 1975, p. 119-131.