

# Estimation du paramètre de second ordre par la méthode des moments pondérés

Julien Worms, Rym Worms

► **To cite this version:**

Julien Worms, Rym Worms. Estimation du paramètre de second ordre par la méthode des moments pondérés. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France, France. 2009. <inria-00386771>

**HAL Id: inria-00386771**

**<https://hal.inria.fr/inria-00386771>**

Submitted on 22 May 2009

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# ESTIMATION DU PARAMÈTRE DE SECOND ORDRE PAR LA MÉTHODE DES MOMENTS PONDÉRÉS

Julien Worms (1) & Rym Worms (2)

(1) *Laboratoire de Mathématiques de Versailles, UMR 8100,  
Université de Versailles-St-Quentin, 45 av. des Etats-Unis, 78035 Versailles cedex*

(2) *Laboratoire d'Analyse et de Mathématiques Appliquées, UMR 8050,  
Université Paris XII Val-de-Marne, 61 av. du G<sup>l</sup> de Gaulle, 94010 Créteil cedex*

*Résumé:* La méthode P.O.T. (Peaks Over Threshold) consiste à utiliser la loi de Pareto Généralisée (GPD) pour approximer la loi des excès au delà d'un seuil élevé. Dans ce travail, la méthode des moments pondérés (PWM) est utilisée pour estimer les paramètres de second ordre qui permettent d'affiner cette approximation : en particulier cela donne lieu à l'introduction d'un nouvel estimateur du paramètre de second ordre  $\rho$ , qui sera comparé à d'autres estimateurs récents par le biais de simulations lors de l'exposé. Des résultats de normalité asymptotique sont établis pour ces estimateurs, et confirment que l'estimation de  $\rho$  doit reposer sur un plus grand nombre d'excès que lorsque l'on désire estimer l'indice  $\gamma$  des valeurs extrêmes.

*Abstract:* The P.O.T. method (Peaks Over Threshold) consists in using the generalized Pareto distribution (GPD) as an approximation for the distribution of excesses over a high threshold. In this work, the method of probability-weighted moments (PWM) is called upon in order to estimate the second order parameters which yield a refinement of this approximation : in particular, this leads us to introduce a new estimator for the second order parameter  $\rho$ , which will be compared to other recent estimators through some simulations during the course of the talk. Asymptotic normality is proved for the estimators and these results confirm that, in order to estimate  $\rho$ , much more excesses must be used than in the case of the estimation of the tail index  $\gamma$ .

*Mots-clés:* Valeurs extrêmes, queue de distribution, loi de Pareto généralisée, paramètre de second ordre, moments pondérés.

## 1. Introduction

On se place dans le cadre classique de la théorie des valeurs extrêmes, en considérant une suite de réalisations indépendantes  $(X_i)_{i \geq 1}$  d'une même variable aléatoire réelle  $X$  de loi de fonction de répartition  $F$  appartenant au domaine d'attraction d'une des lois des valeurs extrêmes  $H_\gamma$ , où  $\gamma \in \mathbb{R}$  est communément appelé indice des valeurs extrêmes associée à  $F$ . L'étude de techniques d'estimation de cet important paramètre  $\gamma$  a fait et fait encore l'objet de nombreux travaux : l'affinage de ces techniques repose souvent sur une bonne connaissance d'un autre paramètre d'intérêt,  $\rho$ , paramètre dit de second ordre, défini plus bas et dont l'estimation est un des objectifs du présent travail.

Les résultats présentés ici s'inscrivent dans le cadre classique des méthodes basées sur l'information apportée par les observations excédant un seuil grand donné. Si  $s_+(F) = \sup\{x; F(x) < 1\}$  désigne le point terminal (à droite) de  $F$ , on considère la fonction de répartition  $F_u$  de la loi des excès au delà d'un seuil  $u$  donné, définie pour  $x \geq 0$  par

$$F_u(x) = \mathbb{P}(X \leq u + x | X > u).$$

Comme cela a été établi dans Balkema & De Haan (1974), l'appartenance de  $F$  au domaine d'attraction de la loi des valeurs extrêmes  $H_\gamma$  équivaut à l'existence d'une fonction d'échelle  $\sigma(\cdot)$  strictement positive telle que la loi des excès au delà d'un seuil  $u$  grand puisse être approximée par une loi de Pareto généralisée (GPD), autrement dit telle que

$$\lim_{u \nearrow s_+(F)} \sup_{0 < x < s_+(F) - u} |F_u(x) - G_{\gamma, \sigma(u)}(x)| = 0.$$

Cette caractérisation est à la base de techniques d'estimation de  $\gamma$  (méthodes de type *Peaks Over Threshold*, ou POT; voir Pickands (1975)) et surtout d'approximation de la queue de  $F$  : pour une suite de seuils  $(u_n)$  telle que  $u_n \rightarrow s_+(F)$  par valeurs inférieures, en se basant sur les observations  $X_{1, N_n} \leq \dots \leq X_{N_n, N_n}$  qui, parmi  $X_1, \dots, X_{N_n}$ , dépassent le seuil  $u_n$  (on notera plus loin  $k_n$  la réalisation du nombre aléatoire d'excès  $N_n$ ), il est possible d'estimer les paramètres  $\gamma$  et  $\sigma_n = \sigma(u_n)$  par la méthode des moments pondérés (abrégée PWM dans la suite). Hosking & Wallis (1987) l'on fait sous l'hypothèse que  $F$  est exactement une loi GPD, arguant du fait que la méthode du maximum de vraisemblance, applicable dans ce cadre et la plus efficace asymptotiquement, présente en pratique de sérieux défauts. Dans Diebolt, Guillou, Worms (2003), les résultats de normalité asymptotique établis dans Hosking & Wallis (1975) ont été étendus au cas considéré ici, c'est-à-dire pour  $F$  seulement supposée appartenir au domaine d'attraction d'une GPD : ils ont également été complétés par un résultat fonctionnel de convergence en loi de l'estimateur correspondant  $G_{\hat{\gamma}_n, \hat{\sigma}_n}$  de  $F_{u_n}$ . Plus précisément, ces résultats ont été prouvés en supposant satisfaites les conditions classiques suivantes :

(C0)  $F$  est deux fois dérivable et inversible d'inverse  $F^{-1}$

(C1)  $\lim_{t \rightarrow +\infty} A(t) = 0$

(C2)  $A$  est de signe constant à l'infini et il existe  $\rho \leq 0$  tel que  $|A| \in RV_\rho$

où  $A$  est définie par  $A(t) = V''(\ln t)/V'(\ln t) - \gamma$ , avec  $V(t) = \bar{F}^{-1}(e^{-t})$ . Les conditions (C1) et (C2) sont communément appelées conditions de premier et de second ordre, la première suffisant à assurer la propriété d'approximation de la loi des excès par une GPD et la seconde permettant de préciser la vitesse avec laquelle cette approximation a lieu, par le biais du paramètre  $\rho$ .

Le point de départ du présent travail était d'essayer d'améliorer l'estimation de la queue de distribution dans ce cadre, en se basant sur le résultat suivant : sous les hypothèses décrites ci-dessus, il a été démontré dans Raoult & Worms (2003) que, quand  $u_n \rightarrow s_+(F)$ , on a pour tout  $y$ ,

$$F_{u_n}(\sigma_n y) - G_\gamma(y) = -a_n D_{\gamma, \rho}(y) + o(a_n)$$

où  $G_\gamma = G_{\gamma, 1}$ ,  $\sigma_n = V'(V^{-1}(u_n))$ ,  $a_n = A(e^{V^{-1}(u_n)})$  (donc  $a_n \rightarrow 0$  en vertu de (C1)) et  $D_{\gamma, \rho}$  est une fonction d'expression explicite que nous ne détaillerons pas ici. Ceci suggère donc que, pour tout  $y$ , la probabilité  $F_{u_n}(x)$  est mieux approximée par  $G_{\gamma, \sigma_n}(x) - a_n D_{\gamma, \rho}(x/\sigma_n)$  que par  $G_{\gamma, \sigma_n}(x)$  seul.

Cette approximation de  $F_{u_n}$  nous a permis d'estimer les paramètres de second ordre  $\rho$  et  $a_n$ , en utilisant la méthode des moments pondérés. L'estimation de  $\rho$  n'est pas un sujet nouveau : depuis une quinzaine d'années plusieurs estimateurs de  $\rho$  ont été proposés, et le nouvel estimateur introduit ici, et pour lequel on démontre la normalité asymptotique, apparaît comme compétitif en particulier dans des cadres où  $\rho$  est proche de 0. Plusieurs simulations seront présentées pendant l'exposé pour attester de ce fait.

## 2. Méthode PWM pour l'estimation des paramètres de second ordre

La méthode PWM est basée sur l'utilisation des moments pondérés de la loi de fonction de répartition  $\tilde{G}_n(y) = G_{\gamma, \sigma_n}(y) - a_n D_{\gamma, \rho}(y/\sigma_n)$  plutôt que ceux de la loi GPD seule  $G_{\gamma, \sigma_n}$ , comme cela était le cas dans Diebolt, Guillou, Worms (2003). On considère les trois premiers moments (note: leur dépendance en  $n$  ne sera pas incluse dans les notations afin de ne pas alourdir celles-ci) définis par

$$\tilde{v}_j = \mathbb{E}(X(1 - \tilde{G}_n)^j(X)), \text{ pour } j = 0, 1, 2$$

et lorsque  $\gamma < 1$ , on a  $\tilde{v}_0 = v_0$  et  $\tilde{v}_j = v_j + o(a_n)$  (pour  $j = 1, 2$ ) où

$$v_j := \frac{\sigma_n}{(j+1)(j+1-\gamma)} + \frac{a_n \sigma_n}{u_j} \quad \text{et} \quad u_j := (j+1)(j+1-\gamma)(j+1-\gamma-\rho).$$

Ce sont ces approximations des 3 premiers moments pondérés de  $B$  qui seront utilisées ici, et dans un premier temps le paramètre  $\gamma$  est considéré connu : en inversant le système liant  $v_0, v_1, v_2$  aux 3 paramètres restants  $\rho, a_n, \sigma_n$  et en remplaçant ces moments pondérés par des estimateurs  $\hat{v}_0, \hat{v}_1, \hat{v}_2$ , nous obtenons des estimateurs notés  $\hat{\rho}_\gamma, \hat{a}_{n, \gamma}, \hat{\sigma}_{n, \gamma}$ . Les estimateurs  $\hat{v}_j$  choisis sont les moments pondérés de la loi empirique de l'échantillon des

excès  $Y_{1,N_n}, \dots, Y_{N_n,N_n}$  au delà du seuil  $u_n$  (i.e.  $Y_{i,N_n} := X_{i,N_n} - u_n$ ) : conditionnellement à  $N_n = k_n$ , cela donne, pour  $j = 0, 1, 2$ ,

$$\hat{v}_j := \frac{1}{j+1} \sum_{i=1}^{k_n} \left( \left(1 - \frac{i-1}{k_n}\right)^{k+1} - \left(1 - \frac{i}{k_n}\right)^{j+1} \right) Y_{i,k_n}.$$

### 3. Résultats de normalité asymptotique

Pour obtenir les résultats de normalité asymptotique des estimateurs des moments pondérés (et donc a fortiori des estimateurs des paramètres), nous avons recours à la condition suivante, dite condition de troisième ordre, introduite dans Fraga Alves, De Haan & Lin (2003) :

(C3) il existe une fonction  $B$  et une constante  $\beta \neq 0$  telles que  $B$  soit de signe constant à l'infini,  $|B| \in RV_\beta$ , et pour tout  $x \in \mathbb{R}$

$$\lim_{t \rightarrow +\infty} \frac{1}{B(e^t)} \left( \frac{\frac{V(t+x)-V(t)}{V'(t)} - \int_0^x e^{\gamma s} ds}{A(e^t)} - I_{\gamma,\rho}(x) \right) = \int_0^x e^{\gamma s} \int_0^s e^{\rho z} \int_0^z e^{\beta y} dy dz ds$$

#### Théorème 1

Sous les conditions (C0)-(C3), si  $-1 < \gamma < 1/2$  et si  $\exists \lambda_1, \lambda_2 \in \mathbb{R}$  telles que

$$\lim_{n \rightarrow \infty} \sqrt{n(1-F(u_n))} a_n b_n = \lambda_1 \quad (1)$$

$$\lim_{n \rightarrow \infty} \sqrt{n(1-F(u_n))} a_n^2 = \lambda_2 \quad (2)$$

$$\lim_{n \rightarrow \infty} \sqrt{n(1-F(u_n))} a_n = \pm \infty \quad (3)$$

alors, pour presque toute suite  $k_n \rightarrow +\infty$ , conditionnellement à  $N_n = k_n$ , le vecteur de coordonnées  $\sqrt{k_n}(\hat{v}_j/\sigma_n - v_j/\sigma_n)$  ( $j = 0, 1, 2$ ) converge en loi vers une loi gaussienne d'espérance  $\lambda_1 C$  et de matrice de covariance  $\Gamma$  ( $C$  et  $\Gamma$  sont définies explicitement,  $C$  étant fonction de  $\gamma, \rho, \beta$  et  $\Gamma$  de  $\gamma$  uniquement).

#### Théorème 2

Sous les conditions (C0)-(C3) et (1)-(3), si l'on suppose  $-1 < \gamma < 1/2$  connu, alors pour presque toute suite  $k_n \rightarrow +\infty$ , conditionnellement à  $N_n = k_n$ , les suites

$$\sqrt{k_n} a_n (\hat{\rho}_\gamma - \rho), \quad \sqrt{k_n} a_n (\hat{a}_{n,\gamma}/a_n - 1), \quad \sqrt{k_n} (\hat{\sigma}_{n,\gamma}/\sigma_n - 1)$$

convergent en loi vers des lois gaussiennes définies explicitement.

Le théorème précédent met en évidence les normalisations nécessaires à l'obtention de la normalité asymptotique des estimateurs de  $\rho$ ,  $a_n$ , et  $\sigma_n$ , mais dans le cas où  $\gamma$

est supposé connu. On constate en particulier que le nombre d'excès  $k_n$  nécessaire pour aboutir à la normalité asymptotique de  $\hat{\rho}_\gamma$  doit être bien plus grand que le nombre d'excès habituellement pris pour l'estimation de  $\gamma$ ; ce constat a déjà été fait dans d'autres cadres et par d'autres auteurs, et se trouve vérifié ici également. Le résultat ci-dessous montre que ces normalisations ne peuvent toutefois être conservées lorsque l'on remplace  $\gamma$  par un estimateur de celui-ci.

On considère que  $(\tilde{u}_n)$  désigne une suite de seuils (convergeant vers  $s_+(F)$ ) et que  $\tilde{N}_n$  est égal au nombre d'excès parmi  $X_1, \dots, X_n$  au delà du seuil  $\tilde{u}_n$ . On pose  $\tilde{a}_n = A(e^{V^{-1}(\tilde{u}_n)})$ .

### **Théorème 3**

*Supposons les conditions (C0)-(C3) et (1)-(3) satisfaites et  $-1 < \gamma < 1/2$ .*

*Soit  $\hat{\gamma} = \hat{\gamma}_n$  un estimateur de  $\gamma$  qui soit fonction des  $\tilde{N}_n$  excès au delà de  $\tilde{u}_n$ , et  $(k_n)$  et  $(\tilde{k}_n)$  des suites tendant vers  $+\infty$  telles que, pour une certaine constante  $\tilde{\lambda}$ ,*

$$\tilde{k}_n = o(k_n) \quad \text{et} \quad \tilde{k}_n^{1/2} \tilde{a}_n \rightarrow \tilde{\lambda} \quad \text{quand } n \rightarrow \infty. \quad (4)$$

*Si conditionnellement à  $\tilde{N}_n = \tilde{k}_n$  on a*

$$\tilde{k}_n^{1/2}(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}(\tilde{\lambda}c, d) \quad \text{quand } n \rightarrow \infty$$

*alors pour presque toutes suites  $(k_n)$  et  $(\tilde{k}_n)$  satisfaisant à (4), on a, conditionnellement à  $N_n = k_n$  et  $\tilde{N}_n = \tilde{k}_n$ , convergence en loi vers des gaussiennes (définies explicitement) des suites*

$$\sqrt{\tilde{k}_n a_n}(\hat{\rho}_{\hat{\gamma}} - \rho), \quad \sqrt{\tilde{k}_n a_n}(\hat{a}_{n, \hat{\gamma}}/a_n - 1) \quad \text{et} \quad \sqrt{\tilde{k}_n}(\hat{\sigma}_{n, \hat{\gamma}}/\sigma_n - 1).$$

### **Bibliographie**

- [1] A. Balkema & L. de Haan (1974), Residual life time at a great age, *Annals of Probability*, vol 2, 792-801.
- [2] J. Diebolt, A. Guillou, R. Worms (2003), Asymptotic behaviour of the probability weighted moments and penultimate approximation, *ESAIM Probability & Statistics*, vol 7, 219-238.
- [3] M.I. Fraga Alves, M.I. Gomes, L. de Haan (2003), A new class of semi-parametric estimators of the second order parameter, *Portugaliae Mathematica*, vol 60, no. 2, 193-213.
- [4] M. I. Fraga Alves, L. de Haan & T. Lin (2003), Estimation of the parameter controlling the speed of convergence in extreme value theory, *Mathematical Methods of Statistics*, vol. 12, no. 2, 155-176.
- [5] J. Pickands III (1975), Statistical inference using extreme order statistics, *Annals of Statistics*, vol 3, 119-131.
- [6] J.P. Raoult, R. Worms (2003), Rate of convergence for the generalized Pareto approximation of the excesses, *Advances in Applied Probability*, vol 35, no. 4, 1007-1027.