

UN TEST SIMPLE POUR LE PARAMÈTRE DE FORME D'UNE LOI DE PARETO GÉNÉRALISÉE

Ali Chaouche & Jean-Noël Bacro

ENGREF/GRESE 19, Avenue du Maine 75732 Paris cedex 15

&

*Université Montpellier II , I3M, UMR CNRS 5149, CC51, Place Eugène Bataillon
34095 Montpellier cedex 5*

Résumé

L'estimation par maximum de vraisemblance des paramètres de forme et d'échelle, respectivement ξ et β , d'une distribution de Pareto généralisée conduit à un système d'équations n'admettant pas de solutions explicites. Ce système peut être résumé sous la forme d'une simple équation $h_{\mathbf{x}}(\rho) = 0$ où \mathbf{x} désigne un n -échantillon de réalisations et $\rho = \frac{\xi}{\beta}$. L'étude de la fonction $h_{\mathbf{x}}(\rho)$ au voisinage de 0 permet de mettre en évidence une statistique $S(\mathbf{x})$ indépendante de β . Cette statistique, fondée sur les deux premiers moments empiriques, peut alors être utilisée pour construire un test statistique simple pour ξ . La loi asymptotique de $S(\mathbf{x})$ est caractérisée. Le test obtenu est de même puissance que les approches concurrentes de la littérature, lesquelles sont plus difficiles à mettre en oeuvre en pratique.

Abstract

The maximum likelihood system of equations for the shape and scale parameters ξ and β of a generalized Pareto distribution is summarized through a single equation $h_{\mathbf{x}}(\rho) = 0$ where \mathbf{x} designates the n -sample of realizations and $\rho = \frac{\xi}{\beta}$. The study of the function $h_{\mathbf{x}}(\rho)$ in the neighborhood of 0 puts forward a simple statistic $S(\mathbf{x})$ only based on the first two empirical moments which is independent from β . This statistic allows to build an immediate and easy to use statistical test for ξ . The power of this test appears as quite similar to the one of already known test procedures which are much more difficult to use in practice.

Mots clés : Distribution de Pareto généralisée ; maximum de vraisemblance ; test pour le domaine d'attraction du maximum.

1 Introduction

La distribution de Pareto généralisée (DPG) a été introduite par Pickands (1975) et Balkema et de Haan (1974) comme distribution limite des dépassements normalisés d'un

seuil élevé et est devenue une notion centrale en valeurs extrêmes (Davidson et Smith (1990)). La fonction de répartition d'une DPG, $G_{\xi,\beta}$, est donnée par :

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \frac{\xi}{\beta}x)^{-1/\xi} & \xi \neq 0, \\ 1 - \exp(-\frac{x}{\beta}) & \xi = 0, \end{cases} \quad (1)$$

où $\beta > 0$ et x tel que $0 < x < \infty$ si $\xi \geq 0$, $0 < x < -\beta/\xi$ si $\xi < 0$. Le cas $\xi = 0$ correspond au cas limite $\xi \rightarrow 0$. Les paramètres ξ et β représentent respectivement les paramètres de forme et d'échelle de la DPG.

En pratique, les paramètres ξ et β doivent être estimés sur la base d'un échantillon d'observations. Les approches les plus utilisées sont le maximum de vraisemblance, la méthode des moments ou encore la méthode des moments pondérés. L'approche par maximum de vraisemblance apparaît comme la plus performante uniquement pour de grands échantillons (Hosking et Wallis (1986)). Les équations du maximum de vraisemblance n'admettent pas de solutions explicites et leurs résolutions nécessitent l'utilisation d'algorithmes d'optimisation. En pratique, une utilisation *en aveugle* de ces algorithmes peut être dangereuse : la vraisemblance de la DPG peut en effet présenter plusieurs maxima ou des singularités sources de sérieux problèmes pour les approches numériques automatiques.

Dans la section 2, nous mettons en évidence une fonction particulière de $\frac{\xi}{\beta}$ dont les racines permettent d'atteindre les solutions du système d'équations du maximum de vraisemblance. L'étude de cette fonction permet alors, entre autres, de construire un test sur ξ . Le test est présenté en section 3.

Dans la suite, on note $\mathbf{x} = (x_1, \dots, x_n)$ un n -échantillon de réalisations d'une v.a. X de loi $G_{\xi,\beta}$, $\mu_k = \int_0^{+\infty} X^k dG_{\xi,\beta}(X)$, k entier > 0 , le moment d'ordre k de X et $m_k(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n x_i^k$, k entier > 0 , son moment empirique d'ordre k .

2 La fonction $h_{\mathbf{x}}(\cdot)$

Soit $g_{\xi,\beta}(\cdot)$ la fonction de densité associée à la DPG. La log-vraisemblance fonction sachant \mathbf{x} , $\mathcal{L}(\xi, \beta; \mathbf{x}) \equiv \mathcal{L}$, est alors donnée par :

$$\mathcal{L} = -n \ln \beta + \ln \left(\prod_{i=1}^n \left(\left(1 + \xi \frac{x_i}{\beta} \right)^{-\frac{1}{\xi}-1} \right) \right). \quad (2)$$

Le système des équations du maximum de vraisemblance s'écrit :

$$\begin{cases} \frac{1}{n} \sum_{i=1}^n \ln \left(1 + \xi \frac{x_i}{\beta} \right) = \xi \\ \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + \xi \frac{x_i}{\beta}} = \frac{1}{1 + \xi} \end{cases} \quad (3)$$

ou, de façon équivalente,

$$\begin{cases} \xi = \ln \left(m_G \left(1 + \frac{\xi}{\beta} \mathbf{x} \right) \right) \\ 1 + \ln \left(m_G \left(1 + \frac{\xi}{\beta} \mathbf{x} \right) \right) - m_H \left(1 + \frac{\xi}{\beta} \mathbf{x} \right) = 0 \end{cases} \quad (4)$$

où $m_G(\mathbf{x}) = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}}$ désigne la moyenne géométrique et $m_H(\mathbf{x}) = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$ la moyenne

harmonique.

Soit $\rho = \frac{\xi}{\beta}$; on considère la fonction $h_{\mathbf{x}}(\rho)$ définie par :

$$h_{\mathbf{x}}(\rho) = 1 + \ln(m_G(1 + \rho \mathbf{x})) - m_H(1 + \rho \mathbf{x}) \quad (5)$$

Les estimations de ξ et β peuvent être déduites des racines de $h_{\mathbf{x}}(\rho)$. La valeur 0 est racine évidente. L'approche classique pour déterminer d'éventuelles autres racines est de caractériser les variations de $h_{\mathbf{x}}(\rho)$. Si l'on pose $y_i = \frac{1}{1 + \rho x_i}$, $\mathbf{y} = (y_1, \dots, y_n)$, $var(\mathbf{y}) = m_2(\mathbf{y}) - (m_1(\mathbf{y}))^2$, on obtient

$$\frac{d}{d\rho} h_{\mathbf{x}}(\rho) = \frac{1}{\rho m_1(\mathbf{y})^2} [var(\mathbf{y}) - m_1(\mathbf{y})(1 - m_1(\mathbf{y}))^2] \quad (6)$$

mais la résolution de l'équation $var(\mathbf{y}) - m_1(\mathbf{y})(1 - m_1(\mathbf{y}))^2 = 0$ est très difficile et nous oblige à une approche alternative : le comportement de $h_{\mathbf{x}}(\rho)$ est caractérisé au travers d'une étude combinant résultats analytiques et de simulations. Cette approche permet, au final, de proposer une procédure fiable de recherche des solutions du maximum de vraisemblance. Parmi les retombées de cette étude, un test simple pour ξ peut être mis en évidence, lequel est détaillé dans la section suivante.

3 Test pour ξ

Au voisinage de 0 la fonction $h_{\mathbf{x}}(\rho)$ peut s'écrire :

$$h_{\mathbf{x}}(\rho) = \left(\frac{m_2(\mathbf{x})}{2} - m_1^2(\mathbf{x}) \right) \rho^2 + 2 \left(m_1(\mathbf{x}) m_2(\mathbf{x}) - \frac{m_3(\mathbf{x})}{3} \right) \rho^3 + o(\rho^3)$$

Le coefficient $\frac{m_2(\mathbf{x})}{2} - m_1^2(\mathbf{x})$ est un estimateur de $\left(\frac{\mu_2}{2} - \mu_1^2 \right)$, paramètre nul dans le cas d'une loi exponentielle ($\xi = 0$). Afin de travailler avec une statistique indépendante de β , on considère $S(\mathbf{x}) = \frac{m_2(\mathbf{x})}{2m_1^2(\mathbf{x})} - 1$, estimateur de $\frac{\mu_2}{2\mu_1^2} - 1 \equiv \frac{\xi}{1 - 2\xi}$ pour $\xi < \frac{1}{2}$. La loi asymptotique de S est donnée par le théorème suivant :

Théorème

Pour tout $\xi < \frac{1}{2}$, on a :

$$\sqrt{n} \left(S(\mathbf{x}) - \frac{1 - \xi}{1 - 2\xi} \right) \xrightarrow{\mathcal{D}} \mathcal{N} \left(0, \frac{(1 - \xi)^2(1 - \xi + 6\xi^2)}{(1 - 2\xi)^3(1 - 3\xi)(1 - 4\xi)} \right) \text{ quand } n \rightarrow \infty,$$

où $\mathcal{N}(0, \sigma^2)$ désigne la loi normale centrée de variance σ^2 .

Un test statistique pour ξ peut donc être construit à partir de la statistique S . L'hypothèse nulle peut concerner n'importe quelle valeur de ξ , mais le test de $\xi = 0$ contre $\xi \neq 0$ est le plus utilisé en pratique. La figure 1 présente la courbe de puissance empirique d'un tel test. Les courbes obtenues sont similaires à celles obtenues avec des approches concurrentes (Chaouche et Bacro (2004)) mais le test fondé sur S à l'avantage d'être bien plus simple à mettre en oeuvre en pratique.

Bibliographie

- Balkema, A.A. et de Haan, L. (1974). Residual lifetime at great age. *Ann. Probab.*, 2, 792-804.
- Chaouche, A. et Bacro, J.N. (2004). A statistical test procedure for the shape parameter of a generalized Pareto distribution. *Comp. Stat. and Data Anal.*, 45, 787-803.
- Davison, A.C. et Smith, R.L. (1990). Models for exceedances over high thresholds. *J.R. Statist. Soc. B*, 52-3, 393-442.
- Hosking, J.R.M. et Wallis, J.R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29, 339-349.
- Pickands, J. (1975). Statistical inference using extreme order statistics, *Ann. Stat.*, 3, 119-131.