HAL open science

Estimation de point terminal dans le domaine d'attraction de Weibull par une méthode des moments d'ordre élevé

Gilles Stupfler, Stéphane Girard, Armelle Guillou

▶ **To cite this version:**

Gilles Stupfler, Stéphane Girard, Armelle Guillou. Estimation de point terminal dans le domaine d'attraction de Weibull par une méthode des moments d'ordre élevé. 44e Journées de Statistique, May 2012, Bruxelles, Belgique. hal-00801381

HAL Id: hal-00801381

https://hal.science/hal-00801381

Submitted on 18 Mar 2013

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 de point terminal dans le domaine d'attraction de Weibull par une méthode des moments d'ordre élevé

Gilles Stupfler [1], Stéphane Girard [2] & Armelle Guillou [3]

[1] *Université de Strasbourg & CNRS, IRMA, UMR 7501, 7, rue René Descartes, 67084 Strasbourg Cedex, France. E-mail : stupfler@math.unistra.fr*
[2] *Team Mistis, INRIA Rhône-Alpes & LJK, Inovallée, 655, avenue de l'Europe, Montbonnot, 38334 Saint-Ismier Cedex, France. E-mail : stephane.girard@inria.fr*
[3] *Université de Strasbourg & CNRS, IRMA, UMR 7501, 7, rue René Descartes, 67084 Strasbourg Cedex, France. E-mail : guillou@math.unistra.fr*

Résumé. On présente une méthode pour estimer le point terminal d'un échantillon de variables aléatoires réelles indépendantes de même loi lorsque leur fonction de répartition commune appartient au domaine d'attraction de Weibull. La méthode repose sur une transformation de la variable d'intérêt et sur l'utilisation de moments d'ordre élevé de la variable positive obtenue de cette façon. On suppose que l'ordre des moments tend vers l'infini. Des conditions sur la vitesse de divergence de l'ordre des moments sont données pour obtenir la consistance de l'estimateur ainsi que sa normalité asymptotique. Les performances de l'estimateur sont illustrées par quelques simulations.

Mots-clés. Estimation de point terminal, moments d'ordre élevé, consistance, normalité asymptotique.

Abstract. We present a method for estimating the endpoint of a unidimensional sample when the distribution function belongs to the Weibull-max domain of attraction. The approach relies on transforming the variable of interest and then using high order moments of the positive variable obtained this way. It is assumed that the order of the moments goes to infinity. We give conditions on the rate of divergence to get the weak and strong consistency as well as the asymptotic normality of the estimator. The good performance of the estimator is illustrated on some finite sample situations.

Keywords. Endpoint estimation, high order moments, consistency, asymptotic normality.

1 Introduction

Soit $(X_1, \ldots, X_n)$ un échantillon de copies indépendantes d'une variable aléatoire X, dont la fonction de répartition possède un point terminal à droite $\theta < \infty$, où θ est supposé inconnu. On s'intéresse à l'estimation de θ. Parmi les travaux récents sur ce

problème, on peut citer une approche par vraisemblance bayésienne (Hall et Wang, 2005), des estimateurs utilisant la vraisemblance censurée (Li *et al.*, 2011a) et une méthode basée sur la vraisemblance empirique (Li *et al.*, 2011b). Dans Girard *et al.* (2012), un nouvel estimateur de θ est présenté ; il utilise des moments d'ordre élevé des X_k. D'un point de vue pratique, l'utilisation des moments d'ordre élevé donne (exponentiellement) plus de poids aux X_k proches de θ. Une étude théorique détaillée de l'estimateur est réalisée en supposant que X est une variable aléatoire positive dont la fonction de répartition appartient au domaine d'attraction de Weibull.

On considère ici le problème d'estimation de θ avec des moments d'ordre élevé, sans hypothèse de positivité sur X. On ne peut pas utiliser de moments de la variable X, puisque $|X|$ pourrait avoir une moyenne infinie. Pour résoudre ce problème, on remarque que la variable aléatoire e^X est à support borné $\left[0,\, e^{\theta}\right]$ et que, si $\mu_{p_n} := \mathbb{E}\left(e^{p_n X}\right)$, $p_n \to \infty$, est le moment d'ordre p_n de e^X, alors pour tout $u \geq 1$, on a $\mu_{p_n}/\mu_{p_n+u} \to e^{-u\theta}$ quand $n \to \infty$. Pour tout $a > 0$, il s'ensuit que

$$\Theta_n := \frac{1}{a}\left\{\log\left[\frac{\mu_{p_n}}{\mu_{p_n+1}}\right] - \log\left[\frac{\mu_{(a+1)p_n}}{\mu_{(a+1)p_n+a+1}}\right]\right\} = \theta(1+\mathrm{o}(1)).$$

On introduit alors un estimateur utilisant des moments d'ordre élevé de la variable aléatoire e^X : en remplaçant le vrai moment μ_{p_n} par le moment empirique correspondant $\widehat{\mu}_{p_n}$ dans l'expression de Θ_n, on obtient un estimateur $\widehat{\theta}_n$ défini par

$$\widehat{\theta}_n := \frac{1}{a}\left\{\log\left[\frac{\widehat{\mu}_{p_n}}{\widehat{\mu}_{p_n+1}}\right] - \log\left[\frac{\widehat{\mu}_{(a+1)p_n}}{\widehat{\mu}_{(a+1)p_n+a+1}}\right]\right\} \quad \text{où} \quad \widehat{\mu}_{p_n} := \frac{1}{n}\sum_{i=1}^{n} e^{p_n X_i}.$$

2 Résultats

Les résultats de consistance s'obtiennent sans aucune hypothèse paramétrique sur X :

Théorème 1 (Consistance faible). Si $\dfrac{n\,\mu_{(a+1)p_n}}{e^{(a+1)p_n\theta}} \to \infty$ quand $n \to \infty$, alors $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ quand $n \to \infty$.

Théorème 2 (Consistance forte). Si $\dfrac{1}{\log n}\,\dfrac{n\,\mu_{(a+1)p_n}}{e^{(a+1)p_n\theta}} \to \infty$ quand $n \to \infty$, alors $\widehat{\theta}_n \xrightarrow{\text{p.s.}} \theta$ quand $n \to \infty$.

Pour obtenir la normalité asymptotique de notre estimateur et en particulier sa vitesse de convergence, on demande que la fonction de survie $\overline{F} = 1 - F$ de X appartienne au domaine d'attraction de Weibull et vérifie une hypothèse de second ordre :

(A_1) $\forall\, x < \theta$, $\overline{F}(x) = (\theta - x)^{\alpha}\, L((\theta - x)^{-1})$ où $\alpha > 0$ et L est une fonction à variation lente, c'est-à-dire telle que $L(ty)/L(y) \to 1$ quand $y \to \infty$, pour tout $t > 0$.

(A_2) $\forall\, x \geq 0$, $L(x) = c \exp\left(\int_1^x \eta(t)\, t^{-1}\, dt\right)$, où $c > 0$ et η est une fonction telle que

- η est bornée, tend vers 0 en $+\infty$ et est non identiquement nulle ;
- η est continûment différentiable sur $(0,\, \infty)$;
- η est monotone à l'infini ;
- $|\eta'|$ est à variation régulière et il existe $\nu \leq 0$ tel que $x\, \eta'(x)/\eta(x) \to \nu$ quand $x \to \infty$.

Sous ces hypothèses, on peut donner un résultat de normalité asymptotique :

Théorème 3 (Normalité asymptotique). On suppose que (A_1) et (A_2) sont vérifiées. Si $n\, p_n^{-\alpha}\, L(p_n) \to \infty$ et $n\, p_n^{-\alpha}\, L(p_n)\, \eta^2(p_n) \to 0$, alors

$$v_n \left(\widehat{\theta}_n - \theta\right) \xrightarrow{d} \mathcal{N}\left(0,\, V(\alpha,\, a)\right) \quad \text{quand } n \to \infty,$$

où $v_n = \sqrt{n\, L(p_n)}\, p_n^{-\alpha/2+1}$ et

$$V(\alpha,\, a) = \frac{\alpha+1}{a^2\, \Gamma(\alpha)} \left[2^{-\alpha-2} - 2\frac{(a+1)^{\alpha+1}}{(a+2)^{\alpha+2}} + 2^{-\alpha-2}(a+1)^{\alpha}\right].$$

3 Etude sur simulations

On examine ici les performances de notre estimateur sur deux modèles différents. Dans le premier modèle noté (M_1), X a pour fonction de survie

$$\forall\, x < 0, \quad \overline{F}(x) = \left[1 + (-x)^{-\tau_1}\right]^{-\tau_2}$$

où $\tau_1,\, \tau_2 > 0$, c'est-à-dire $X = -1/Z$ où Z a une loi de Burr$(1,\, \tau_1,\, \tau_2)$ type XII. Ici $\theta = 0$, $\alpha = \tau_1\, \tau_2$ et $\nu = -\tau_1$.

Dans le second modèle noté (M_2), X a pour fonction de survie

$$\forall\, x < 0, \quad \overline{F}(x) = \int_{\log(1-1/x)}^{\infty} \lambda^2\, t\, e^{-\lambda t}\, dt$$

où $\lambda > 0$, c'est-à-dire $X = -1/(e^Z - 1)$ où Z a une loi Gamma$(2,\, \lambda)$. Ici $\theta = 0$, $\alpha = \lambda$ et $\nu = 0$.

Pour chaque modèle, on teste différentes valeurs des paramètres, voir Table 1. On prend $p_n = n^{1/\alpha}/\log\log n$, et un ensemble $\mathcal{A} = \{0.1,\, 0.2,\, 0.3,\, \ldots,\, 25\}$ de valeurs de a est testé.

Distribution	Maximum	Aarssen & de Haan	$\widehat{\theta}_n$
Modèle (M_1)			
$(\tau_1,\, \tau_2) = (1,\, 1)$	$2.0 \cdot 10^{-3}$	$2.1 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$
$(\tau_1,\, \tau_2) = (5/6,\, 6/5)$	$2.0 \cdot 10^{-3}$	$2.0 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$
$(\tau_1,\, \tau_2) = (2/3,\, 3/2)$	$2.2 \cdot 10^{-3}$	$2.0 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$
$(\tau_1,\, \tau_2) = (1/2,\, 2)$	$2.3 \cdot 10^{-3}$	$2.3 \cdot 10^{-3}$	$2.0 \cdot 10^{-3}$
Modèle (M_2)			
$\lambda = 1$	$2.3 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$	$1.9 \cdot 10^{-4}$
$\lambda = 5/4$	$1.1 \cdot 10^{-3}$	$9.2 \cdot 10^{-4}$	$8.5 \cdot 10^{-4}$
$\lambda = 5/3$	$5.7 \cdot 10^{-3}$	$4.6 \cdot 10^{-3}$	$4.1 \cdot 10^{-3}$
$\lambda = 5/2$	$3.1 \cdot 10^{-2}$	$2.4 \cdot 10^{-2}$	$2.3 \cdot 10^{-2}$

Table 1: Erreur L^1 moyenne associée aux estimateurs dans chaque situation.

Dans chaque situation, on simule $N = 1000$ échantillons de taille $n = 500$ et l'erreur L^1- moyenne

$$E(a) = \frac{1}{N} \sum_{j=1}^{N} \left| \widehat{\theta}^{(j,\, a)} - \theta \right|$$

est calculée, où $\widehat{\theta}^{(j,\, a)}$ est l'estimation calculée sur le $j-$ième échantillon avec $a \in \mathcal{A}$ et $\theta = 0$. La valeur "optimale" de a est conservée : $a^\star = \operatorname{argmin}\{E(a),\, a \in \mathcal{A}\}$. La même procédure est appliquée à l'estimateur de Aarssen et de Haan (1994), qui utilise un seuil $k \in \{2,\, 3,\, \ldots,\, n-1\}$. L'estimateur du maximum est aussi considéré. Les résultats sont résumés dans la Table 1, où $E(a^\star)$ est donné. On remarque que dans toutes les situations, notre estimateur donne de meilleurs résultats que ses deux concurrents.

Sur la Figure 1, on compare les erreurs E associées aux estimateurs dans deux cas. Sur le modèle (M_1), alors que l'erreur associée à l'estimateur de Aarssen et de Haan semble très sensible au choix de k, celle de notre estimateur est stable pour un grand éventail de valeurs de a. Les résultats sont similaires dans les autres cas.

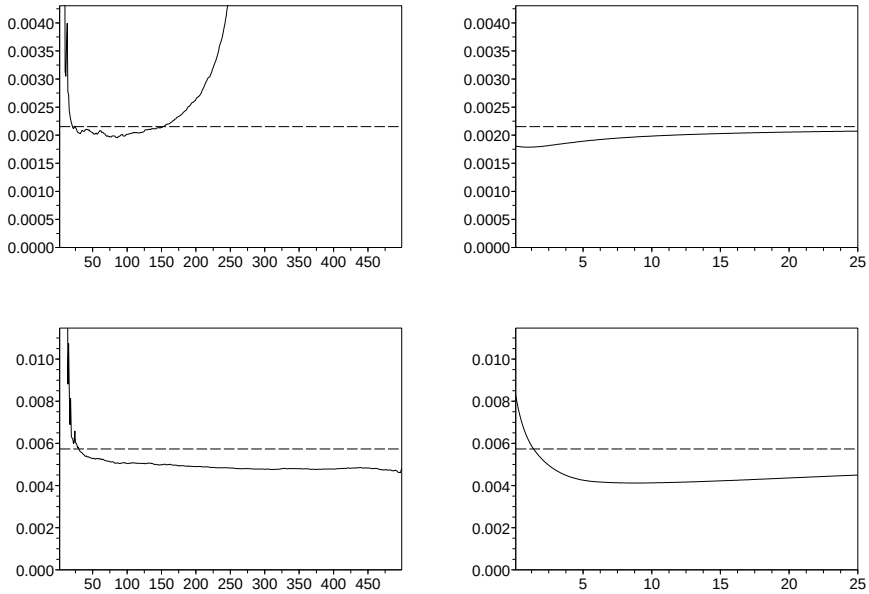

Figure 1: Comparaison des estimateurs sur les modèles (M_1) et (M_2). A gauche, en abscisses : seuil k, en ordonnées : erreur E, en pointillés : maximum, en trait plein : estimateur de Aarssen & de Haan. A droite, en abscisses : paramètre a, en ordonnées : erreur E, en pointillés : maximum, en trait plein : estimateur $\widehat{\theta}_n$. En haut : modèle (M_1), $(\tau_1,\, \tau_2) = (2/3,\, 3/2)$. En bas : modèle (M_2), $\lambda = 5/3$.

Bibliographie

[1] Aarssen, K. et de Haan, L., 1994, On the maximal life span of humans, *Math. Popul. Stud.*, 4(4), 259–281.

[2] Girard, S., Guillou, A. et Stupfler, G., 2012, Estimating an endpoint with high order moments, *Test*, to appear. DOI : 10.1007/s11749-011-0277-8.

[3] Hall, P. et Wang, J.Z., 2005, Bayesian likelihood methods for estimating the end point of a distribution, *J. Roy. Statist. Soc. Ser. B*, 67(5), 717–729.

[4] Li, D., Peng, L. et Xu, X., 2011a, Bias reduction for endpoint estimation, *Extremes*, 14, 393–412.

[5] Li, D., Peng, L. et Qi, Y., 2011b, Empirical likelihood confidence intervals for the endpoint of a distribution function, *Test*, 20, 353–366.