# An asymptotically unbiased minimum density power divergence estimator for the Pareto-tail index

Goedele Dierckx, Yuri Goegebeur, Armelle Guillou

# An asymptotically unbiased minimum density power divergence estimator for the Pareto-tail index

Goedele Dierckx[(1)], Yuri Goegebeur[(2)] & Armelle Guillou[(3)]

[(1)] Faculty Economics and Management, Hogeschool-Universiteit Brussel, Warmoesberg 26, 1000 Brussel, Belgium
[(2)] Department of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark
[(3)] Institut Recherche Mathématique Avancée, UMR 7501, Université de Strasbourg et CNRS, 7 rue René Descartes, 67084 Strasbourg cedex, France

**Abstract.** We introduce a robust and asymptotically unbiased estimator for the tail index of Pareto-type distributions. The estimator is obtained by fitting the extended Pareto distribution to the relative excesses over a high threshold with the minimum density power divergence criterion. Consistency and asymptotic normality of the estimator is established under a second order condition on the distribution underlying the data, and for intermediate sequences of upper order statistics. The finite sample properties of the proposed estimator and some alternatives from the extreme value literature are evaluated by a small simulation experiment involving both uncontaminated and contaminated samples.

**AMS Subject Classifications:** 62G05, 62G20, 62G32, 62G35.

**Keywords:** Pareto-type distribution, tail index, bias-correction, density power divergence.

## 1 Introduction

Basu *et al.* (1998) introduced the idea of the density power divergence for the purpose of developing a robust estimation criterion. In particular, the density power divergence between density functions $f$ and $g$ is given by

$$\Delta_\alpha(f,g) := \begin{cases} \int_{\mathbb{R}} \left[g^{1+\alpha}(y) - \left(1+\frac{1}{\alpha}\right) g^\alpha(y) f(y) + \frac{1}{\alpha} f^{1+\alpha}(y)\right] dy, & \alpha > 0, \\ \int_{\mathbb{R}} \log \frac{f(y)}{g(y)} f(y) dy, & \alpha = 0. \end{cases}$$

Assume that the density function $g$ depends on a parameter vector $\theta$, and let $Y_1, \ldots, Y_n$ be a sample of independent and identically distributed (i.i.d.) random variables according to density function $f$. The minimum density power divergence estimator (MDPDE) is the point $\hat{\theta}$ minimizing the empirical density power divergence

$$\widehat{\Delta}_\alpha(\theta) := \int_{\mathbb{R}} g^{1+\alpha}(y) dy - \left(1+\frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^n g^\alpha(Y_i),$$

for $\alpha > 0$, and

$$\widehat{\Delta}_0(\theta) := -\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i),$$

for $\alpha = 0$. Note that in case $\alpha = 0$, the empirical density power divergence corresponds with minus the log-likelihood function. The parameter $\alpha$ controls the trade-off between efficiency and robustness of the MDPDE: the estimator becomes more efficient but less robust against outliers as $\alpha$ gets closer to zero, whereas for increasing $\alpha$ the robustness increases and the efficiency decreases. In the present paper we will use the density power divergence criterion with the objective to obtain a robust and asymptotically unbiased estimator for the tail parameter of a Pareto-type distribution.

A distribution is said to be of Pareto-type if for some $\gamma > 0$ its survival function is of the form:

$$\bar{F}(x) := 1 - F(x) = x^{-1/\gamma}\ell_F(x), \qquad x > 0, \tag{1}$$

where $\ell_F$ denotes a slowly varying function at infinity, i.e.

$$\frac{\ell_F(\lambda x)}{\ell_F(x)} \to 1 \text{ as } x \to \infty \text{ for all } \lambda > 0. \tag{2}$$

The parameter $\gamma$ is the extreme value index, and clearly governs the tail behavior of $F$, with larger values indicating heavier tails. This parameter assumes a central position in extreme value statistics, and consequently considerable efforts have been dedicated to its estimation. We refer to Beirlant *et al.* (2004) and de Haan and Ferreira (2006) for recent accounts on the estimation of $\gamma$. Consistency of such estimators can usually be achieved under (1), whereas asymptotic normality requires that one imposes some more structure on the tail of $F$. In the present paper we will work with the second order condition of Beirlant *et al.* (2009). Let $RV_\beta$ denote the class of the regularly varying functions at infinity with index $\beta$, i.e. Lebesgue measurable ultimately positive functions $z$ satisfying $\lim_{t\to\infty} z(tx)/z(t) = x^\beta$ for all $x > 0$.

**Condition** ($\mathcal{R}$). *Let $\gamma > 0$ and $\tau < 0$ be constants. The distribution function $F$ is such that $x^{1/\gamma}\bar{F}(x) \to C \in (0,\infty)$ as $x \to \infty$ and the function $\delta$ defined via*

$$\bar{F}(x) = Cx^{-1/\gamma}(1 + \gamma^{-1}\delta(x)),$$

*is ultimately nonzero, of constant sign and $|\delta| \in RV_\tau$.*

For practical purposes, condition ($\mathcal{R}$) is not very restrictive and therefore well accepted in extreme value theory. It is satisfied by e.g. the Hall class of Pareto-type distributions (Hall, 1982), which is obtained if $\delta(x) \sim Dx^\tau$, $x \to \infty$, for some $D \neq 0$. This class contains many commonly used distributions like the Fréchet, Burr, absolute Student $t$, $F$ and generalized Pareto, to name but a few. In case there is doubt about the appropriateness of this model for a given set of data, one can use a formal goodness-of-fit test for Pareto-type behavior with a second order condition, see e.g. Beirlant *et al.* (2006), Goegebeur *et al.* (2008), Koning and Peng (2008). We also refer to Neves and Fraga Alves (2008) for a recent overview of test procedures concerning extreme value conditions.

Clearly, condition $(\mathcal{R})$ implies that the tail quantile function $U$, defined as $U(y) = \inf\{x : F(x) \geq 1 - 1/y\}$, $y > 1$, satisfies $y^{-\gamma}U(y) \to C^{\gamma}$ as $y \to \infty$ and the function $a$ implicitly defined by

$$U(y) = C^{\gamma}y^{\gamma}(1 + a(y)) \tag{3}$$

satisfies $a(y) = \delta(C^{\gamma}y^{\gamma})(1 + o(1))$ as $y \to \infty$, so $|a| \in RV_{\rho}$, with $\rho = \gamma\tau$.

The second order condition $(\mathcal{R})$ can be used to derive the so-called extended Pareto distribution (Beirlant *et al.*, 2004, Beirlant *et al.*, 2009), with distribution function given by

$$G(y) = \begin{cases} 1 - [y(1 + \delta - \delta y^{\tau})]^{-1/\gamma}, & y > 1, \\ 0, & y \leq 1, \end{cases} \tag{4}$$

where $\gamma > 0$, $\tau < 0$, and $\delta > \max\{-1, 1/\tau\}$. In Beirlant *et al.* (2004) and Vandewalle *et al.* (2007) a slightly different form of this model is presented, which consists of a mixture of two Pareto distributions. In fact, as shown in Proposition 2.3 of Beirlant *et al.* (2009), for distribution functions satisfying $(\mathcal{R})$, the distribution function of the relative excess $Y := X/u$ given that $X > u$ can be approximated by (4) with $\delta = \delta(u)$ up to an error that is uniformly $o(\delta(u))$ for $u \to \infty$. Beirlant *et al.* (2009) obtained an asymptotically unbiased estimator for $\gamma$ by applying a maximum likelihood procedure to (4). In the present paper we will introduce a robust asymptotically unbiased estimator for $\gamma$, which is obtained by applying the minimum density power divergence criterion to (4). The maximum likelihood estimator of Beirlant *et al.* (2009) is then a special case of our approach and corresponds with $\alpha = 0$.

Kim and Lee (2008) obtained a robust estimator for $\gamma > 0$ by fitting the strict Pareto distribution to the largest observations in a given dataset with the minimum density power divergence criterion. Their method allows for strong mixing processes, but, since the estimator can be seen as a robust version of the Hill estimator (Hill, 1975), it is not asymptotically unbiased. Juárez and Schucany (2004) obtained the MDPDE for the parameters of the generalized Pareto distribution (GPD). Although the method works in principle also for some $\gamma < 0$, in that $\gamma > -(1 + \alpha)/(2 + \alpha)$, they assumed that the data follow exactly a GPD, and hence the bias which appears when using the GPD as approximation for the true, but in practice unknown, distribution of the excesses over a high threshold, is ignored. Vandewalle *et al.* (2007) fitted a so-called partial density component involving a mixture of two Pareto distributions by minimizing a $L_2$ distance, a method which is closely related to the MDPDE approach presented in this paper with $\alpha = 1$. Although the method seems to be bias-correcting and robust, the authors did not investigate the asymptotic properties of their estimators. Other contributions to the robust estimation of the extreme value index can be found in e.g. Peng and Welsh (2001), Dupuis and Victoria-Feser (2006) and Hubert *et al.* (2012).

The remainder of this paper is organized as follows. In the next section we introduce a robust and asymptotically unbiased estimator for the positive extreme value index, obtained by fitting the extended Pareto distribution to the relative excesses over a high threshold by means of the density power divergence criterion, and also examine the asymptotic properties of the estimator. The finite sample performance of the proposed procedure is evaluated by a small simulation experiment in Section 3. The proofs of all results are given in the appendix.

## 2 Estimation procedure and asymptotic properties

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a distribution function satisfying $(\mathcal{R})$, and denote the associated ascending order statistics by $X_{1,n} \leq \ldots \leq X_{n,n}$. We estimate the parameters $\gamma$ and $\delta$ of the extended Pareto distribution with the minimum density power divergence criterion applied to the relative excesses over the random threshold $u = X_{n-k,n}$, namely $Y_j := X_{n-k+j,n}/X_{n-k,n}$, $j = 1, \ldots, k$.

The density function of the extended Pareto distribution is given by

$$g(y) = \frac{1}{\gamma}\, y^{-1/\gamma-1}[1+\delta(1-y^\tau)]^{-1/\gamma-1}[1+\delta(1-(1+\tau)y^\tau)], \qquad y > 1,$$

where $\gamma > 0$, $\tau < 0$, and $\delta > \max\{-1, 1/\tau\}$. Remember that the parameter $\delta$ reflects in fact the function $\delta(u)$, where $\delta(u) \to 0$ as $u \to \infty$, in condition $(\mathcal{R})$, but we do not make this dependence on the threshold explicit in the notation.

The MDPDE for $(\gamma, \delta)$ satisfies the system of equations

$$0 = \int_1^\infty g^\alpha(y)\frac{\partial g(y)}{\partial \gamma}dy - \frac{1}{k}\sum_{j=1}^k g^{\alpha-1}(Y_j)\frac{\partial g(Y_j)}{\partial \gamma}, \tag{5}$$

$$0 = \int_1^\infty g^\alpha(y)\frac{\partial g(y)}{\partial \delta}dy - \frac{1}{k}\sum_{j=1}^k g^{\alpha-1}(Y_j)\frac{\partial g(Y_j)}{\partial \delta}. \tag{6}$$

These estimating equations depend also on the unknown parameter $\tau$. For this parameter, we use the parameterization $\tau = \rho/\gamma$.

We examine the asymptotic properties of the solutions to the estimating equations (5) and (6), assuming that the underlying distribution of the data satisfies $(\mathcal{R})$. First we consider the case where the true value of $\rho$ is known, and establish existence and consistency of a sequence of solutions to the estimating equations (5) and (6). From now on we denote the true value of $\gamma$ and $\rho$ by $\gamma_0$ and $\rho_0$, respectively.

**Theorem 1** *Let $X_1, \ldots, X_n$ be a sample of i.i.d. random variables from a distribution function satisfying $(\mathcal{R})$. Then, if $k, n \to \infty$ with $k/n \to 0$ we have that with probability tending to 1, there exist solutions $(\hat{\gamma}_n, \hat{\delta}_n)$ of the estimating equations (5) and (6), with $\rho$ fixed at $\rho_0$, such that $(\hat{\gamma}_n, \hat{\delta}_n) \xrightarrow{\mathbb{P}} (\gamma_0, 0)$.*

Next, we state the asymptotic normality of the sequence of consistent solutions of equations (5) and (6). Let $\delta_n := \delta(X_{n-k,n})$.

**Theorem 2** *Let $X_1, \ldots, X_n$ be a sample of i.i.d. random variables from a distribution function satisfying ($\mathcal{R}$), and assume that $(\hat{\gamma}_n, \hat{\delta}_n)$ is a consistent sequence of estimators for $(\gamma_0, 0)$, satisfying (5) and (6), with $\rho$ fixed at $\rho_0$. Then if $k, n \to \infty$ with $k/n \to 0$ and $\sqrt{k}a(n/k) \to \lambda \in \mathbb{R}$, we have that*

$$\sqrt{k}\begin{bmatrix} \hat{\gamma}_n - \gamma_0 \\ \hat{\delta}_n - \delta_n \end{bmatrix} \rightsquigarrow N_2\left(\mathbf{0}, \mathbb{C}^{-1}(\rho_0)\mathbb{B}(\rho_0)\mathbf{\Sigma}(\rho_0)\mathbb{B}'(\rho_0)\mathbb{C}^{-1}(\rho_0)\right),$$

*where the elements of the matrix $\mathbf{\Sigma}(\rho_0)$ are given by (21)–(26), and the matrices $\mathbb{B}(\rho_0)$ and $\mathbb{C}(\rho_0)$ are defined in (27) and (28), respectively.*

Note that the condition $\sqrt{k}a(n/k) \to \lambda$ in Theorem 2 implies that $\sqrt{k}\delta_n = \lambda + o_{\mathbb{P}}(1)$. The estimators are asymptotically unbiased in the sense that whatever the value of $\lambda$, the mean of the normal limiting distribution is equal to zero. In Figure 1 we show the asymptotic standard deviation of $\hat{\gamma}_n$ as a function of $\rho_0$ for some values of $\alpha$ in case $\gamma_0 = 0.5$. As expected, the smallest asymptotic variance is obtained for $\alpha = 0$, which corresponds to maximum likelihood estimation, and is given by $\gamma_0^2(1-\rho_0)^2/\rho_0^2$. Increasing $\alpha$ leads to more robust estimates, but decreases the asymptotic efficiency of the estimation procedure. Table 1 displays some values of the asymptotic relative efficiency of the estimator $\hat{\gamma}_n$ with $\alpha = 0.1, 0.5$ and 1, relative to the estimator $\hat{\gamma}_n$ with $\alpha = 0$. For $\alpha = 1$, the asymptotic relative efficiency has dropped already considerably. The efficiency also tends to decrease for larger values of $\gamma_0$ and $\rho_0$. However, the decreasing asymptotic relative efficiency for increasing values of $\alpha$ is not specific for our estimation procedure. For instance, Juárez and Schucany (2004) studied the estimation of the parameters of the generalized Pareto distribution by means of the density power divergence criterion, and from Table 1 in their paper it is clear that also for their method the asymptotic relative efficiency (compared to maximum likelihood estimation) drops quickly as a function of $\alpha$. In fact our figures are in line with the numbers reported in their Table 1. Also Kim and Lee (2008) illustrate the loss of efficiency of their density power divergence estimator relative to the Hill estimator as a function of $\alpha$. The loss of efficiency is the price to pay to obtain an estimator with better robustness properties. We illustrate the increased robustness by calculating the influence function of our estimator as a function of the point of contamination in case of an underlying Fréchet model with $\gamma_0 = 0.5$ and a threshold for estimation fixed at quantile 0.75, and this for several values of $\alpha$ ; see Figure 2. The calculation of the influence function is based on a straightforward generalization of the idea presented in Vandewalle *et al.* (2007) to the MDPDE framework, and therefore we omit the details. As is clear from Figure 2, if the contamination is reasonably far in the tail of the distribution then the influence function is decreasing in $\alpha$, illustrating the improved robustness. In particular, observe that unlike the bias-corrected maximum likelihood estimator, our bias-corrected robust estimator has a bounded influence function for the larger values of $\alpha$, and therefore it can be considered to be B-robust (Hampel *et al.*, 1986).

The following proposition deals with the behavior of the estimator when the parameter $\rho$ is fixed at some value $\tilde{\rho}$, possibly misspecified.

**Proposition 1** *Let $X_1, \ldots, X_n$ be a sample of i.i.d. random variables from a distribution function satisfying ($\mathcal{R}$) and assume the parameter $\rho$ is fixed at $\tilde{\rho}$ in (5) and (6). Then, if $k, n \to \infty$*

Table 1: Asymptotic relative efficiency of $\hat{\gamma}_n$ with $\alpha = 0.1, 0.5$ and 1 relative to $\hat{\gamma}_n$ with $\alpha = 0$.

| | | $\alpha = 0.1$ | $\alpha = 0.5$ | $\alpha = 1$ |
|---|---|---|---|---|
| $\gamma_0 = 0.5$ | $\rho_0 = -0.5$ | 0.90 | 0.33 | 0.16 |
| | $\rho_0 = -1$ | 0.91 | 0.39 | 0.21 |
| | $\rho_0 = -2$ | 0.93 | 0.47 | 0.28 |
| $\gamma_0 = 1$ | $\rho_0 = -0.5$ | 0.84 | 0.24 | 0.11 |
| | $\rho_0 = -1$ | 0.86 | 0.30 | 0.15 |
| | $\rho_0 = -2$ | 0.88 | 0.38 | 0.22 |
| $\gamma_0 = 2$ | $\rho_0 = -0.5$ | 0.71 | 0.16 | 0.06 |
| | $\rho_0 = -1$ | 0.75 | 0.21 | 0.09 |
| | $\rho_0 = -2$ | 0.78 | 0.28 | 0.14 |

*with $k/n \to 0$ we have that with probability tending to 1, there exist solutions $(\hat{\gamma}_n, \hat{\delta}_n)$ of the estimating equations (5) and (6) such that $(\hat{\gamma}_n, \hat{\delta}_n) \xrightarrow{\mathbb{P}} (\gamma_0, 0)$. If additionally $\sqrt{k}a(n/k) \to \lambda \in \mathbb{R}$, we also have that*

$$\sqrt{k} \begin{bmatrix} \hat{\gamma}_n - \gamma_0 \\ \hat{\delta}_n \end{bmatrix} \rightsquigarrow N_2 \left( -\lambda \mathbb{C}^{-1}(\tilde{\rho}) \mathbb{B}(\tilde{\rho}) \tilde{\boldsymbol{\mu}},\ \mathbb{C}^{-1}(\tilde{\rho}) \mathbb{B}(\tilde{\rho}) \boldsymbol{\Sigma}(\tilde{\rho}) \mathbb{B}'(\tilde{\rho}) \mathbb{C}^{-1}(\tilde{\rho}) \right),$$

*where the elements of the vector $\tilde{\boldsymbol{\mu}}$ are defined in (19), (29) and (20).*

Note that, as expected, by a misspecification of $\rho$ at some value $\tilde{\rho}$, one loses the bias-correcting effect of taking the second order structure of $F$ into account in the estimation. However, the variance expression is the same as in Theorem 2, but with $\rho_0$ replaced by $\tilde{\rho}$.

Finally, we examine the asymptotic behavior of $(\hat{\gamma}_n, \hat{\delta}_n)$ in the case where $\rho$ is replaced by an external consistent estimator $\hat{\rho}_n$ in (5) and (6). In the recent extreme value literature, several estimators for $\rho$ have been proposed that work quite well in practice. We refer to Fraga Alves *et al.* (2003), Ciuperca and Mercadier (2010) and Goegebeur *et al.* (2010). As mentioned in these papers, consistency of estimators for $\rho$ can be obtained under a second order framework, though establishing asymptotic normality would require a third order framework. For our purposes, consistency of the $\rho$ estimator is enough, so there is no need for a third order condition. Estimating the second order parameter $\rho$ externally is a common approach in extreme value statistics, and it leads to bias-corrected estimators for $\gamma$ with a smaller asymptotic variance compared to a joint estimation of $(\gamma, \delta, \rho)$ (see e.g. Gomes *et al.*, 2008, for a detailed discussion about the external estimation of second order scale and rate parameters). As shown in the next theorem, replacing $\rho$ by an external consistent estimator leads to the same limiting distribution as in the case where $\rho_0$ is known.

**Theorem 3** *Let $X_1, \ldots, X_n$ be a sample of i.i.d. random variables from a distribution function satisfying $(\mathcal{R})$. The result of Theorem 1 and 2 continues to hold if $\rho$ is replaced by a consistent estimator $\hat{\rho}_n$ in (5) and (6).*

## 3 Simulation study

To study the finite sample behavior of the MDPDE $\hat{\gamma}_n$, a small simulation study has been performed. In order to make the dependence on the tuning parameter $\alpha$ explicit, we use from now on the notation $\hat{\gamma}_{n,\alpha}$ for the MDPDE of the extended Pareto distribution. The estimator $\hat{\gamma}_{n,\alpha}$ is compared with the well-known Hill (1975) estimator $\hat{\gamma}_H$, the maximum likelihood estimator for the extended Pareto Distribution (Beirlant *et al.*, 2009), which corresponds in fact to $\hat{\gamma}_{n,0}$, and the MDPDE estimator $\hat{\gamma}_{KL,\alpha}$ with $\alpha = 0.5$ by Kim and Lee (2008). The choice $\alpha = 0.5$ for the Kim and Lee (2008) estimator can be motivated by the theoretical and simulation results reported in their paper. Indeed, on uncontaminated data sets this choice still gives a reasonable efficiency compared to the Hill estimator (approximately 70%), whereas on contaminated data taking $\alpha = 0.5$ seems to make a good trade of between robustness and efficiency.

We illustrate the finite sample behavior of the above mentioned estimators in Figures 3 to 7 based on 100 simulated datasets of size $n = 200$ for each of the following cases.

- Figure 3: Uncontaminated Fréchet distribution given by $1-F(x) = 1-\exp(-x^{-\beta})$, $x > 0$, $\beta > 0$, denoted Fréchet($\beta$). We have chosen $\beta = 2$, such that the extreme value index $\gamma_0 = 0.5$. This model satisfies the second order condition ($\mathcal{R}$) with $\rho_0 = -1$.

- Figure 4: Uncontaminated Burr distribution $1 - F(x) = \left(\frac{\eta}{\eta+x^{\tau}}\right)^{\lambda}$, $x > 0$, $\eta$, $\tau$, $\lambda > 0$, denoted Burr($\eta, \tau, \lambda$). This model is of Pareto-type with $\gamma_0 = 1/(\lambda\tau)$, and also satisfies the second order condition with $\rho_0 = -1/\lambda$. We have chosen $\eta = 1$, $\tau = 1$, $\lambda = 2$, such that $\gamma_0 = 0.5$ and $\rho_0 = -0.5$.

- Figure 5: Uncontaminated log-gamma distribution for which

$$1 - F(x) = \int_x^{\infty} \frac{\lambda^{\beta}}{\Gamma(\beta)} w^{-\lambda-1} (\log w)^{\beta-1} dw,$$

  $x > 1$, $\beta, \lambda > 0$. This model is of Pareto-type with $\gamma_0 = 1/\lambda$, but it does not satisfy condition ($\mathcal{R}$). It is included in the simulation experiment to illustrate the robustness of the developed methodology with respect to deviations from the assumed model. We set $\beta = 2$ and $\lambda = 1$ so that $\gamma_0 = 1$.

- Figure 6: A contaminated Fréchet distribution with $F_{\epsilon}(x) = (1-\epsilon)F(x) + \epsilon\tilde{F}(x)$ where the uncontaminated distribution $F$ is a Fréchet(2) and $\tilde{F}(x) = 1 - \left(\frac{x}{x_c}\right)^{-\beta}, x > x_c$. We have chosen $\beta = 0.5$, $\epsilon = 0.01$, $x_c$= 2 times the 99.99% quantile of the uncontaminated Fréchet(2) distribution.

- Figure 7: A contaminated Burr distribution with $F_{\epsilon}(x) = (1-\epsilon)F(x) + \epsilon\tilde{F}(x)$ where the uncontaminated distribution $F$ is a Burr(1,1,2) and $\tilde{F}(x) = 1 - \left(\frac{x}{x_c}\right)^{-\beta}, x > x_c$. We have chosen $\beta = 0.5$, $\epsilon = 0.1$, $x_c$= 1.2 times the 99.99% quantile of the uncontaminated Burr(1,1,2) distribution.

Other choices for $\epsilon$ ranging from 0.01 up to 0.1, and $x_c$ were considered and resulted in similar findings.

In the panels (a) and (b) of Figures 3 to 7, we illustrate the behavior of $\hat{\gamma}_{n,\alpha}$ when different levels of robustness are considered, namely $\alpha = 0.1$ (dotted), $\alpha = 0.5$ (solid) and $\alpha = 1$ (dashed). Concerning the simulation from the uncontaminated distributions we see from Figure 3 and 4 that in terms of bias the behavior of the estimators does not depend strongly on the value of $\alpha$. In terms of MSE we have, as expected from the theory, that the performance deteriorates with increasing values of $\alpha$, though the differences are distribution specific. In Figures 6 and 7 we show the behavior of $\hat{\gamma}_{n,\alpha}$ when contamination is present. The estimator $\hat{\gamma}_{n,0.5}$ seems to be less biased compared to the estimator $\hat{\gamma}_{n,0.1}$, whereas the optimal MSE of $\hat{\gamma}_{n,0.5}$ is at least as good. The estimators $\hat{\gamma}_{n,0.5}$ and $\hat{\gamma}_{n,1}$ are comparable with a slightly better behavior of $\hat{\gamma}_{n,0.5}$. We conclude that for the contamination we considered, the value $\alpha = 0.5$ seems to be most appropriate, and will therefore be used in the subsequent comparison with alternative robust estimation procedures from the extreme value literature. These observations seems to be in line with the earlier theoretical considerations. Indeed, for $\alpha = 0.5$ the efficiency is better than for $\alpha = 1$, while the robustness, as examined by the influence functions, seems to be similar.

In the panels (c) and (d) of Figures 3 to 7, $\hat{\gamma}_{n,0.5}$ (solid) is compared with $\hat{\gamma}_H$ (dashed), $\hat{\gamma}_{n,0}$ (dashed-dotted) and $\hat{\gamma}_{KL,0.5}$ (dotted). For the uncontaminated cases $\hat{\gamma}_{n,0.5}$ is quite competitive in terms of bias and MSE. For the contamination we considered, the estimator $\hat{\gamma}_{n,0.5}$ clearly outperforms the non-robust estimators $\hat{\gamma}_H$ and $\hat{\gamma}_{n,0}$ both in terms of bias and MSE. The minimum value of the MSE of $\hat{\gamma}_{n,0.5}$ and that of $\hat{\gamma}_{KL,0.5}$ are comparable, but usually $\hat{\gamma}_{n,0.5}$ shows a more stable sample path. This was to be expected, since the estimator $\hat{\gamma}_{n,\alpha}$ is designed to have better second order properties. Similar stability results were reported in other studies of bias-corrected estimators like Beirlant *et al.* (1999), Feuerverger and Hall (1999), Gomes *et al.* (2008), and Beirlant *et al.* (2009).

We also evaluated the performance of all estimators on the log-gamma distribution, which is of Pareto-type, but does not satisfy $(\mathcal{R})$; see Figure 5. Clearly, the bias-corrected estimators continue to work better with respect to bias than the uncorrected estimators.

In Figures 3 to 7, the parameter $\rho$ was estimated using the estimator introduced by Fraga Alves *et al.* (2003). In Figure 8 it is illustrated how $\hat{\gamma}_{n,0.5}$ changes when $\rho$ is misspecified to -1. For the smaller $k$-values, the estimation is quite insensitive to this misspecification, whereas for the larger $k$-values, the estimator $\hat{\gamma}_{n,0.5}$ with a misspecification of $\rho$ approximates the true value $\gamma_0 = 0.5$ even better.

From the simulations we can conclude that $\hat{\gamma}_n$ is in general a good alternative to estimate $\gamma$. In uncontaminated cases its behavior is, as expected, similar to that of the maximum likelihood estimator for the extended Pareto distribution when $\alpha$ is kept small, say in the range 0.1 to 0.5. When contamination is present, $\hat{\gamma}_n$ outperforms non-robust estimators while by taking the second order structure of $F$ explicitly into account, the estimator has more stable sample paths than robust estimators using only the first order structure of $F$. These more stable sample paths alleviate the typical issue about the choice of $k$, the number of extremes to be used in the estimation, to some extent. If there is need to determine $k$ in an adaptive way then one could use the method recently introduced by Gomes *et al.* (2012), where a versatile bootstrap algorithm was proposed to select $k$ for bias-corrected estimators.

# Appendix

## An auxiliary lemma

First we state a lemma giving the limiting distribution of the random terms that appear in the derivations of Theorems 1 and 2. For convenience we state the result in terms of convergence of empirical processes. For $s \leq 0$, let

$$
\begin{aligned}
A_{k,n}^{(1)}(s) &:= \frac{1}{k}\sum_{j=1}^{k} Y_j^s, \\
A_{k,n}^{(2)}(s) &:= \frac{1}{k}\sum_{j=1}^{k} Y_j^s \log Y_j, \\
A_{k,n}^{(3)}(s) &:= \frac{1}{k}\sum_{j=1}^{k} Y_j^s \log^2 Y_j,
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{A}_{k,n}^{(1)}(s) &:= \sqrt{k}\left(A_{k,n}^{(1)}(s) - \frac{1}{1-s\gamma_0}\right), \\
\mathbb{A}_{k,n}^{(2)}(s) &:= \sqrt{k}\left(A_{k,n}^{(2)}(s) - \frac{\gamma_0}{(1-s\gamma_0)^2}\right), \\
\mathbb{A}_{k,n}^{(3)}(s) &:= \sqrt{k}\left(A_{k,n}^{(3)}(s) - \frac{2\gamma_0^2}{(1-s\gamma_0)^3}\right).
\end{aligned}
$$

**Lemma 1** *Let $X_1, \ldots, X_n$ be a sample of i.i.d. random variables from a distribution function satisfying $(\mathcal{R})$. Then if $k, n \to \infty$ with $k/n \to 0$ and $\sqrt{k}a(n/k) \to \lambda \in \mathbb{R}$, for every $s_0 < 0$, in $\mathcal{C}^3[s_0, 0]$, we have that*

$$
(\mathbb{A}_{k,n}^{(1)}, \mathbb{A}_{k,n}^{(2)}, \mathbb{A}_{k,n}^{(3)}) \rightsquigarrow (\mathbb{A}^{(1)}, \mathbb{A}^{(2)}, \mathbb{A}^{(3)}),
$$

*a Gaussian process, with, for $s, s_1, s_2 \in [s_0, 0]$, expected values*

$$
\begin{aligned}
\mathbb{E}(\mathbb{A}^{(1)}(s)) &= \lambda\, \frac{s\rho_0}{(1-s\gamma_0)(1-\rho_0-s\gamma_0)}, \\
\mathbb{E}(\mathbb{A}^{(2)}(s)) &= \lambda\left\{\frac{\rho_0}{(1-s\gamma_0)(1-\rho_0-s\gamma_0)} + s\gamma_0\left[\frac{1}{(1-\rho_0-s\gamma_0)^2} - \frac{1}{(1-s\gamma_0)^2}\right]\right\}, \\
\mathbb{E}(\mathbb{A}^{(3)}(s)) &= \lambda 2\gamma_0\left\{\frac{1}{(1-\rho_0-s\gamma_0)^2} - \frac{1}{(1-s\gamma_0)^2} + \gamma_0 s\left[\frac{1}{(1-\rho_0-s\gamma_0)^3} - \frac{1}{(1-s\gamma_0)^3}\right]\right\},
\end{aligned}
$$

*and covariances*

$$\mathbb{C}ov(\mathbb{A}^{(1)}(s_1), \mathbb{A}^{(1)}(s_2)) = \frac{\gamma_0^2 s_1 s_2}{(1-s_1\gamma_0)(1-s_2\gamma_0)[1-(s_1+s_2)\gamma_0]}, \tag{7}$$

$$\mathbb{C}ov(\mathbb{A}^{(2)}(s_1), \mathbb{A}^{(2)}(s_2)) = \gamma_0^2 \left\{ \frac{2}{[1-(s_1+s_2)\gamma_0]^3} - \frac{1}{(1-s_1\gamma_0)^2(1-s_2\gamma_0)^2} \right\}, \tag{8}$$

$$\mathbb{C}ov(\mathbb{A}^{(3)}(s_1), \mathbb{A}^{(3)}(s_2)) = \gamma_0^4 \left\{ \frac{4!}{[1-(s_1+s_2)\gamma_0]^5} - \frac{4}{(1-s_1\gamma_0)^3(1-s_2\gamma_0)^3} \right\}, \tag{9}$$

$$\mathbb{C}ov(\mathbb{A}^{(1)}(s_1), \mathbb{A}^{(2)}(s_2)) = \gamma_0 \left\{ \frac{1}{[1-(s_1+s_2)\gamma_0]^2} - \frac{1}{(1-s_1\gamma_0)(1-s_2\gamma_0)^2} \right\}, \tag{10}$$

$$\mathbb{C}ov(\mathbb{A}^{(1)}(s_1), \mathbb{A}^{(3)}(s_2)) = 2\gamma_0^2 \left\{ \frac{1}{[1-(s_1+s_2)\gamma_0]^3} - \frac{1}{(1-s_1\gamma_0)(1-s_2\gamma_0)^3} \right\}, \tag{11}$$

$$\mathbb{C}ov(\mathbb{A}^{(2)}(s_1), \mathbb{A}^{(3)}(s_2)) = 2\gamma_0^3 \left\{ \frac{3}{[1-(s_1+s_2)\gamma_0]^4} - \frac{1}{(1-s_1\gamma_0)^2(1-s_2\gamma_0)^3} \right\}. \tag{12}$$

**Proof of Lemma 1**

The term $\mathbb{A}_{k,n}^{(1)}(s)$ was already treated in Beirlant *et al.* (2009). The terms $\mathbb{A}_{k,n}^{(2)}(s)$ and $\mathbb{A}_{k,n}^{(3)}(s)$ can be dealt with analogously, and therefore we only give the big lines of the proof.

Let $\breve{Y}_1, \ldots, \breve{Y}_n$ be independent unit Pareto random variables, with order statistics $\breve{Y}_{1,n} \leq \ldots \leq \breve{Y}_{n,n}$. From the inverse probability integral transform we have that

$$(X_{j,n}; j = 1, \ldots, n) \stackrel{\mathcal{D}}{=} (U(\breve{Y}_{j,n}); j = 1, \ldots, n).$$

For unit Pareto order statistics it is well-known that

$$\left( \frac{\breve{Y}_{n-k+j,n}}{\breve{Y}_{n-k,n}}; j = 1, \ldots, k \right) \stackrel{\mathcal{D}}{=} (\tilde{Y}_{j,k}; j = 1, \ldots, k),$$

where $\tilde{Y}_{1,k} \leq \ldots \leq \tilde{Y}_{k,k}$ are the order statistics of a random sample of size $k$ from the unit Pareto distribution, and $\tilde{Y}_{1,k}, \ldots, \tilde{Y}_{k,k}$ are independent of $\breve{Y}_{n-k,n}$. Combining the above two displays we have the following distributional representation

$$(X_{n-k+j,n}; j = 1, \ldots, k) \stackrel{\mathcal{D}}{=} (U(\tilde{Y}_{j,k}\breve{Y}_{n-k,n}); j = 1, \ldots, k). \tag{13}$$

Let $\eta(y) := \log(1 + a(y))$. From (3) and (13), and introducing $\varepsilon_{j,n} := \eta(\tilde{Y}_j \breve{Y}_{n-k,n})/\eta(\breve{Y}_{n-k,n}) - \tilde{Y}_j^{\rho_0}$, $j = 1, \ldots, k$, we then have

$$A_{k,n}^{(1)}(s) \stackrel{\mathcal{D}}{=} \frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{s\gamma_0} e^{s\eta(\breve{Y}_{n-k,n})(\tilde{Y}_j^{\rho_0}-1+\varepsilon_{j,n})},$$

$$A_{k,n}^{(2)}(s) \stackrel{\mathcal{D}}{=} \frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{s\gamma_0} e^{s\eta(\breve{Y}_{n-k,n})(\tilde{Y}_j^{\rho_0}-1+\varepsilon_{j,n})} [\gamma_0 \log \tilde{Y}_j + \eta(\breve{Y}_{n-k,n})(\tilde{Y}_j^{\rho_0} - 1 + \varepsilon_{j,n})],$$

$$A_{k,n}^{(3)}(s) \stackrel{\mathcal{D}}{=} \frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{s\gamma_0} e^{s\eta(\breve{Y}_{n-k,n})(\tilde{Y}_j^{\rho_0}-1+\varepsilon_{j,n})} [\gamma_0 \log \tilde{Y}_j + \eta(\breve{Y}_{n-k,n})(\tilde{Y}_j^{\rho_0} - 1 + \varepsilon_{j,n})]^2.$$

Since $|a| \in RV_{\rho_0}$, $\rho_0 < 0$, we have that $a(y) \to 0$ for $y \to \infty$, and hence $\eta(y) = a(y)(1 + o(1))$. Further, as shown in Corollary 2.2.2 in de Haan and Ferreira (2006), if $k, n \to \infty$ with $k/n \to 0$ we have that $\sqrt{k}(k/n\breve{Y}_{n-k,n} - 1) \rightsquigarrow N(0,1)$ and thus $\breve{Y}_{n-k,n} = n/k(1 + o_{\mathbb{P}}(1))$. The above combined with Proposition B.1.10 in de Haan and Ferreira (2006) (see also Drees, 1998), give that for any $\varepsilon > 0$ and $0 < \xi < -\rho_0$ there exists an $n_0$ such that for $n \geq n_0$, with arbitrary large probability,

$$|\varepsilon_{j,n}| \leq \varepsilon \tilde{Y}_j^{\rho_0+\xi} \leq \varepsilon.$$

Hence $\max_{j=1,\ldots,k} |\varepsilon_{j,n}| = o_{\mathbb{P}}(1)$.

Now use the inequality $|e^z - 1 - z| \leq z^2/2 \max(e^z, 1)$, $z \in \mathbb{R}$, to obtain

$$\sup_{s\in[s_0,0]} \left| A_{k,n}^{(1)}(s) - \frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{s\gamma_0} - s\eta(\breve{Y}_{n-k,n})\frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{s\gamma_0}(\tilde{Y}_j^{\rho_0} - 1) \right| = o_{\mathbb{P}}(k^{-1/2}),$$

$$\begin{aligned} \sup_{s\in[s_0,0]} \Bigg| A_{k,n}^{(2)}(s) - \gamma_0 \frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{s\gamma_0} \log \tilde{Y}_j - \eta(\breve{Y}_{n-k,n})\frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{s\gamma_0}(\tilde{Y}_j^{\rho_0} - 1) \\ - s\gamma_0\eta(\breve{Y}_{n-k,n})\frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{s\gamma_0} \log \tilde{Y}_j(\tilde{Y}_j^{\rho_0} - 1) \Bigg| = o_{\mathbb{P}}(k^{-1/2}), \end{aligned}$$

$$\begin{aligned} \sup_{s\in[s_0,0]} \Bigg| A_{k,n}^{(3)}(s) - \gamma_0^2 \frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{s\gamma_0} \log^2 \tilde{Y}_j - 2\gamma_0\eta(\breve{Y}_{n-k,n})\frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{s\gamma_0} \log \tilde{Y}_j(\tilde{Y}_j^{\rho_0} - 1) \\ - s\gamma_0^2\eta(\breve{Y}_{n-k,n})\frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{s\gamma_0} \log^2 \tilde{Y}_j(\tilde{Y}_j^{\rho_0} - 1) \Bigg| = o_{\mathbb{P}}(k^{-1/2}). \end{aligned}$$

Consider the functions $h_{1,\theta}(y) = y^\theta$, $h_{2,\theta}(y) = y^\theta \log y$ and $h_{3,\theta}(y) = y^\theta \log^2 y$, where $y \geq 1$ and $\theta < 0$, and introduce the classes $\mathcal{H}_i := \{h_{i,\theta} : \theta \in [\theta_0, 0]\}$, $\theta_0 < 0$, $i = 1, 2, 3$. By the mean value theorem we have that

$$\begin{aligned} |h_{1,\theta_1}(y) - h_{1,\theta_2}(y)| &\leq |\theta_1 - \theta_2| \log y, \\ |h_{2,\theta_1}(y) - h_{2,\theta_2}(y)| &\leq |\theta_1 - \theta_2| \log^2 y, \\ |h_{3,\theta_1}(y) - h_{3,\theta_2}(y)| &\leq |\theta_1 - \theta_2| \log^3 y. \end{aligned}$$

Also, $\mathbb{E}(\log^a \tilde{Y}_1) = \Gamma(a + 1)$, $a > 0$, and hence, using the result from Example 19.7 in van der Vaart (2007), the function classes $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$ are Glivenko-Cantelli for the unit Pareto-distribution. Thus, if $k \to \infty$

$$\begin{aligned}
\sup_{\theta\in[\theta_0,0]} \left| \frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{\theta} - \frac{1}{1-\theta} \right| &\stackrel{\mathbb{P}}{\to} 0, \\
\sup_{\theta\in[\theta_0,0]} \left| \frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{\theta} \log \tilde{Y}_j - \frac{1}{(1-\theta)^2} \right| &\stackrel{\mathbb{P}}{\to} 0, \\
\sup_{\theta\in[\theta_0,0]} \left| \frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{\theta} \log^2 \tilde{Y}_j - \frac{2}{(1-\theta)^3} \right| &\stackrel{\mathbb{P}}{\to} 0,
\end{aligned}$$

and therefore

$$\mathbb{A}_{k,n}^{(1)}(s) \stackrel{\mathcal{D}}{=} \sqrt{k}\left(\frac{1}{k}\sum_{j=1}^{k}\tilde{Y}_j^{s\gamma_0} - \frac{1}{1-s\gamma_0}\right) + \lambda\,\frac{s\rho_0}{(1-s\gamma_0)(1-\rho_0-s\gamma_0)} + \varepsilon_n^{(1)}(s), \tag{14}$$

$$\begin{aligned}
\mathbb{A}_{k,n}^{(2)}(s) \stackrel{\mathcal{D}}{=}\ & \gamma_0\sqrt{k}\left(\frac{1}{k}\sum_{j=1}^{k}\tilde{Y}_j^{s\gamma_0}\log\tilde{Y}_j - \frac{1}{(1-s\gamma_0)^2}\right) + \lambda\left\{\frac{\rho_0}{(1-s\gamma_0)(1-\rho_0-s\gamma_0)}\right. \\
& \left. + s\gamma_0\left[\frac{1}{(1-\rho_0-s\gamma_0)^2} - \frac{1}{(1-s\gamma_0)^2}\right]\right\} + \varepsilon_n^{(2)}(s),
\end{aligned} \tag{15}$$

$$\begin{aligned}
\mathbb{A}_{k,n}^{(3)}(s) \stackrel{\mathcal{D}}{=}\ & \gamma_0^2\sqrt{k}\left(\frac{1}{k}\sum_{j=1}^{k}\tilde{Y}_j^{s\gamma_0}\log^2\tilde{Y}_j - \frac{2}{(1-s\gamma_0)^3}\right) + \lambda 2\gamma_0\left\{\frac{1}{(1-\rho_0-s\gamma_0)^2} - \frac{1}{(1-s\gamma_0)^2}\right. \\
& \left. + s\gamma_0\left[\frac{1}{(1-\rho_0-s\gamma_0)^3} - \frac{1}{(1-s\gamma_0)^3}\right]\right\} + \varepsilon_n^{(3)}(s),
\end{aligned} \tag{16}$$

where $\sup_{s\in[s_0,0]}|\varepsilon_n^{(i)}(s)| = o_{\mathbb{P}}(1),\ i=1,2,3.$

Finally we comment on the joint convergence of the processes $\mathbb{A}_{k,n}^{(1)}$, $\mathbb{A}_{k,n}^{(2)}$ and $\mathbb{A}_{k,n}^{(3)}$. To this aim introduce

$$\begin{aligned}
\tilde{\mathbb{A}}_{k,n}^{(1)}(s) &:= \sqrt{k}\left(\frac{1}{k}\sum_{j=1}^{k}\tilde{Y}_j^{s\gamma_0} - \frac{1}{1-s\gamma_0}\right), \\
\tilde{\mathbb{A}}_{k,n}^{(2)}(s) &:= \gamma_0\sqrt{k}\left(\frac{1}{k}\sum_{j=1}^{k}\tilde{Y}_j^{s\gamma_0}\log\tilde{Y}_j - \frac{1}{(1-s\gamma_0)^2}\right), \\
\tilde{\mathbb{A}}_{k,n}^{(3)}(s) &:= \gamma_0^2\sqrt{k}\left(\frac{1}{k}\sum_{j=1}^{k}\tilde{Y}_j^{s\gamma_0}\log^2\tilde{Y}_j - \frac{2}{(1-s\gamma_0)^3}\right).
\end{aligned}$$

Again by the result for parametric classes of functions from Example 19.7 in van der Vaart (2007) we have that the classes $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$ are Donsker for the unit Pareto distribution.

Following van der Vaart (2007) p. 270 this is equivalent to the class of vector valued functions $h = (h_{1,\theta_1}, h_{2,\theta_2}, h_{3,\theta_3})$ to be Donsker and hence, in $\mathcal{C}^3([s_0, 0])$,

$$(\tilde{\mathbb{A}}^{(1)}_{k,n}, \tilde{\mathbb{A}}^{(2)}_{k,n}, \tilde{\mathbb{A}}^{(3)}_{k,n}) \rightsquigarrow (\tilde{\mathbb{A}}^{(1)}, \tilde{\mathbb{A}}^{(2)}, \tilde{\mathbb{A}}^{(3)}),$$

a zero centered Gaussian process with covariance functions as given in (7) till (12). Combining this with the result in (14), (15) and (16), establishes the result of the lemma.

□

## Proof of Theorem 1

To prove the existence and consistency of $(\hat{\gamma}_n, \hat{\delta}_n)$ we adapt the proof of Theorem 5.1 in Chapter 6 of Lehmann and Casella (1998), where existence and consistency of solutions of the likelihood equations is established, to the MDPDE framework. Let $Q_r$ denote the sphere centered at $(\gamma_0, 0)$ and radius $r$, and let $\widehat{\Delta}_\alpha(\gamma, \delta; \rho)$ denote the density power divergence objective function. Note that $r$ should be such that $Q_r$ is a subset of the parameter space. First we show that for any $r$ sufficiently small

$$\mathbb{P}_{(\gamma_0,0)}(\widehat{\Delta}_\alpha(\gamma_0, 0; \rho_0) < \widehat{\Delta}_\alpha(\gamma, \delta; \rho_0) \text{ for all } (\gamma, \delta) \text{ on the surface of } Q_r) \to 1.$$

Let $f_s(\gamma, \delta; \rho_0)$, $s = 1, 2$, denote the derivatives of $\widehat{\Delta}_\alpha(\gamma, \delta; \rho_0)$ with respect to $\gamma$ and $\delta$, respectively, without the common scale factor $1 + \alpha$. Similarly, $f_{st}$ and $f_{stu}$, $s, t, u = 1, 2$, denote the second and third order derivatives, respectively (again apart from the common scaling by $1+\alpha$).

By Taylor's theorem

$$\begin{aligned}
\widehat{\Delta}_\alpha(\gamma, \delta; \rho_0) - \widehat{\Delta}_\alpha(\gamma_0, 0; \rho_0) \quad = \quad & (1 + \alpha) \{ f_1(\gamma_0, 0; \rho_0)(\gamma - \gamma_0) + f_2(\gamma_0, 0; \rho_0)\delta \\
& + \frac{1}{2} \left[ f_{11}(\gamma_0, 0; \rho_0)(\gamma - \gamma_0)^2 + f_{22}(\gamma_0, 0; \rho_0)\delta^2 + 2f_{12}(\gamma_0, 0; \rho_0)(\gamma - \gamma_0)\delta \right] \\
& + \frac{1}{6} \Big[ f_{111}(\tilde{\gamma}, \tilde{\delta}; \rho_0)(\gamma - \gamma_0)^3 + f_{222}(\tilde{\gamma}, \tilde{\delta}; \rho_0)\delta^3 + 3f_{112}(\tilde{\gamma}, \tilde{\delta}; \rho_0)(\gamma - \gamma_0)^2\delta \\
& + 3f_{122}(\tilde{\gamma}, \tilde{\delta}; \rho_0)(\gamma - \gamma_0)\delta^2 \Big] \Big\} \qquad (17) \\
=: \quad & (1 + \alpha)\{S_1 + S_2 + S_3\},
\end{aligned}$$

where $(\tilde{\gamma}, \tilde{\delta})$ is a point on the line segment connecting $(\gamma, \delta)$ and $(\gamma_0, 0)$. After some tedious, but straightforward derivations one obtains

$$
\begin{aligned}
f_1(\gamma_0,0;\rho_0) &= \gamma_0^{-\alpha-2}\left[-\frac{\alpha\gamma_0(1+\gamma_0)}{[1+\alpha(1+\gamma_0)]^2}+\gamma_0 A_{k,n}^{(1)}(-\alpha(1+\gamma_0)/\gamma_0)-A_{k,n}^{(2)}(-\alpha(1+\gamma_0)/\gamma_0)\right],\\
f_{11}(\gamma_0,0;\rho_0) &= \gamma_0^{-\alpha-2}\left[\frac{\alpha+2}{1+\alpha(1+\gamma_0)}-\frac{2\alpha+4}{[1+\alpha(1+\gamma_0)]^2}+\frac{2\alpha+2}{[1+\alpha(1+\gamma_0)]^3}\right.\\
&\quad -(\alpha+1)A_{k,n}^{(1)}(-\alpha(1+\gamma_0)/\gamma_0)+\frac{2\alpha+2}{\gamma_0}A_{k,n}^{(2)}(-\alpha(1+\gamma_0)/\gamma_0)\\
&\quad \left.-\frac{\alpha}{\gamma_0^2}A_{k,n}^{(3)}(-\alpha(1+\gamma_0)/\gamma_0)\right],\\
f_{12}(\gamma_0,0;\rho_0) &= \gamma_0^{-\alpha-2}\left[\frac{1+\alpha(2+\alpha)(1+\gamma_0)}{[1+\alpha(1+\gamma_0)]^2}\right.\\
&\quad -\frac{(1-\rho_0)^2-\alpha[\rho_0(1-\rho_0)-2(1+\gamma_0)(1-\rho_0)]+\alpha^2(1+\gamma_0)(1-\rho_0)}{[1-\rho_0+\alpha(1+\gamma_0)]^2}\\
&\quad -(1+\alpha)A_{k,n}^{(1)}(-\alpha(1+\gamma_0)/\gamma_0)+(\alpha+1)(1-\rho_0)A_{k,n}^{(1)}(-(\alpha(1+\gamma_0)-\rho_0)/\gamma_0)\\
&\quad \left.+\frac{\alpha}{\gamma_0}A_{k,n}^{(2)}(-\alpha(1+\gamma_0)/\gamma_0)-\frac{(\alpha-\rho_0)(1-\rho_0)}{\gamma_0}A_{k,n}^{(2)}(-(\alpha(1+\gamma_0)-\rho_0)/\gamma_0)\right],\\
f_2(\gamma_0,0;\rho_0) &= \gamma_0^{-\alpha-1}\left[-\frac{\alpha\rho_0(1+\gamma_0)}{[1+\alpha(1+\gamma_0)][1-\rho_0+\alpha(1+\gamma_0)]}+A_{k,n}^{(1)}(-\alpha(1+\gamma_0)/\gamma_0)\right.\\
&\quad \left.-(1-\rho_0)A_{k,n}^{(1)}(-(\alpha(1+\gamma_0)-\rho_0)/\gamma_0)\right],\\
f_{22}(\gamma_0,0;\rho_0) &= \gamma_0^{-\alpha-2}\left[\frac{1+\alpha+\gamma_0}{1+\alpha(1+\gamma_0)}-\frac{2(1-\rho_0)(1+\gamma_0+\alpha)}{1-\rho_0+\alpha(1+\gamma_0)}+\frac{(1+\gamma_0)(1-2\rho_0)+\alpha(1-\rho_0)^2}{1-2\rho_0+\alpha(1+\gamma_0)}\right.\\
&\quad -(\alpha+\gamma_0)A_{k,n}^{(1)}(-\alpha(1+\gamma_0)/\gamma_0)+2(1-\rho_0)(\alpha+\gamma_0)A_{k,n}^{(1)}(-(\alpha(1+\gamma_0)-\rho_0)/\gamma_0)\\
&\quad \left.-[(1+\gamma_0)(1-2\rho_0)+(\alpha-1)(1-\rho_0)^2]A_{k,n}^{(1)}(-(\alpha(1+\gamma_0)-2\rho_0)/\gamma_0)\right].
\end{aligned}
$$

Following the lines of proof of Lemma 1 we have that $f_1(\gamma_0,0;\rho_0)\xrightarrow{\mathbb{P}}0$ and $f_2(\gamma_0,0;\rho_0)\xrightarrow{\mathbb{P}}0$, so, for any given $r>0$ we have that $|f_1(\gamma_0,0;\rho_0)|<r^2$ and $|f_2(\gamma_0,0;\rho_0)|<r^2$ with probability tending to 1, and hence, on $Q_r$, $|S_1|<2r^3$ with probability tending to 1.

Concerning $S_2$, let $f_{st}^*(\gamma_0,0;\rho_0)$ denote the limits of the random terms $f_{st}(\gamma_0,0;\rho_0)$, $s,t=1,2$. These can be obtained from the result of Lemma 1, and are given by

$$
\begin{aligned}
f_{11}^*(\gamma_0,0;\rho_0) &= \gamma_0^{-\alpha-2}\,\frac{1+\alpha^2(1+\gamma_0)^2}{[1+\alpha(1+\gamma_0)]^3},\\
f_{12}^*(\gamma_0,0;\rho_0) &= \gamma_0^{-\alpha-2}\,\frac{\rho_0(1-\rho_0)[1+\alpha(1+\gamma_0)+\alpha^2(1+\gamma_0)^2]+\alpha^3\rho_0(1+\gamma_0)^3}{[1+\alpha(1+\gamma_0)]^2[1-\rho_0+\alpha(1+\gamma_0)]^2},\\
f_{22}^*(\gamma_0,0;\rho_0) &= \gamma_0^{-\alpha-2}\,\frac{(1-\rho_0)\rho_0^2+\alpha\rho_0^2(1+\gamma_0)[\alpha(1+\gamma_0)-\rho_0]}{[1+\alpha(1+\gamma_0)][1-\rho_0+\alpha(1+\gamma_0)][1-2\rho_0+\alpha(1+\gamma_0)]}.
\end{aligned}
$$

Now, write

$$\begin{aligned} 2S_2 &= f^*_{11}(\gamma_0,0;\rho_0)(\gamma-\gamma_0)^2 + f^*_{22}(\gamma_0,0;\rho_0)\delta^2 + 2f^*_{12}(\gamma_0,0;\rho_0)(\gamma-\gamma_0)\delta \\ &\quad +[f_{11}(\gamma_0,0;\rho_0) - f^*_{11}(\gamma_0,0;\rho_0)](\gamma-\gamma_0)^2 + [f_{22}(\gamma_0,0;\rho_0) - f^*_{22}(\gamma_0,0;\rho_0)]\delta^2 \\ &\quad +2[f_{12}(\gamma_0,0;\rho_0) - f^*_{12}(\gamma_0,0;\rho_0)](\gamma-\gamma_0)\delta. \end{aligned}$$

Note that the first three terms are in fact a nonrandom positive definite quadratic form in $(\gamma-\gamma_0)$ and $\delta$. By the spectral decomposition this quadratic form can be rewritten as $\lambda_1\xi_1^2+\lambda_2\xi_2^2$, where $0<\lambda_1\leq\lambda_2$ are the eigenvalues and $\xi_1$ and $\xi_2$ are orthogonal transformations of $(\gamma-\gamma_0)$ and $\delta$. Note that in this new coordinate system $Q_r$ becomes $\xi_1^2+\xi_2^2=r^2$. Thus, for the quadratic form we have that $\lambda_1\xi_1^2+\lambda_2\xi_2^2\geq\lambda_1(\xi_1^2+\xi_2^2)=\lambda_1 r^2$. For the random part of $S_2$ we know from Lemma 1 that $f_{st}(\gamma_0,0;\rho_0)\xrightarrow{\mathbb{P}} f^*_{st}(\gamma_0,0;\rho_0)$, $s,t=1,2$, and thus in absolute value the random part is less than $4r^3$ with probability tending to 1. Overall, we have that there exists $c>0$ and $r_0>0$ such that for $r<r_0$

$$S_2 > cr^2$$

with probability tending to 1.

For the term $S_3$, one can show that $|f_{stu}(\gamma,\delta;\rho_0)|\leq M_{stu}(\boldsymbol{Y})$, where $\boldsymbol{Y}'=(Y_1,\ldots,Y_k)$, for $(\gamma,\delta)\in Q_r$, with $M_{stu}(\boldsymbol{Y})\xrightarrow{\mathbb{P}} m_{stu}$, $s,t,u=1,2$, which is bounded. The derivations are straightforward but tedious, and are for brevity omitted from the paper. Thus, with probability tending to 1, $|f_{stu}(\tilde\gamma,\tilde\delta;\rho_0)|<2m_{stu}$, and hence $|S_3|<br^3$ on $Q_r$, where

$$b=\frac{1}{3}\sum_{s=1}^{2}\sum_{t=1}^{2}\sum_{u=1}^{2} m_{stu}.$$

Combining the above we find that with probability tending to 1,

$$\min(S_1+S_2+S_3) > cr^2-(2+b)r^3,$$

where the minimum is over $(\gamma,\delta)$ on the surface of $Q_r$. Clearly, the right-hand side of the above inequality is positive if $r<c/(2+b)$.

To complete the proof of the existence and consistency we adjust the line of argumentation of Theorem 3.7 in Chapter 6 of Lehmann and Casella (1998). For $r>0$, small enough that $Q_r$ is a subset of the parameter space, consider

$$S_n(r) := \{\boldsymbol{y} : \widehat{\Delta}_\alpha(\gamma_0,0;\rho_0) < \widehat{\Delta}_\alpha(\gamma,\delta;\rho_0) \text{ for all } (\gamma,\delta) \text{ on the surface of } Q_r\}.$$

From the above we have that $\mathbb{P}_{(\gamma_0,0)}(S_n(r))\to 1$ for any such $r$, and hence there exists a sequence $r_n\downarrow 0$ such that $\mathbb{P}_{(\gamma_0,0)}(S_n(r_n))\to 1$ as $n\to\infty$. By the differentiability of $\widehat{\Delta}_\alpha(\gamma,\delta;\rho_0)$ we have that $\boldsymbol{y}\in S_n(r_n)$ implies that there exists a point $(\hat\gamma_n(r_n),\hat\delta_n(r_n))\in Q_{r_n}$ for which $\widehat{\Delta}_\alpha(\gamma,\delta;\rho_0)$ attains a local minimum, and thus $f_s(\hat\gamma_n(r_n),\hat\delta_n(r_n);\rho_0)=0$, $s=1,2$. Now let $(\hat\gamma_n^*,\hat\delta_n^*):=(\hat\gamma_n(r_n),\hat\delta_n(r_n))$ for $\boldsymbol{y}\in S_n(r_n)$ and arbitrary otherwise. Clearly

$$\mathbb{P}_{(\gamma_0,0)}(f_1(\hat\gamma_n^*,\hat\delta_n^*;\rho_0)=0, f_2(\hat\gamma_n^*,\hat\delta_n^*;\rho_0)=0)\geq\mathbb{P}_{(\gamma_0,0)}(S_n(r_n))\to 1,$$

as $n \to \infty$. Thus with probability tending to 1 there exists a sequence of solutions to the estimating equations (5) and (6). Let $d(v,w)$ denote the Euclidean distance between the points $v$ and $w$. Then, for any fixed $r>0$ and $n$ sufficiently large

$$\mathbb{P}_{(\gamma_0,0)}(d((\hat{\gamma}_n^*,\hat{\delta}_n^*),(\gamma_0,0))<r) \geq \mathbb{P}_{(\gamma_0,0)}(d((\hat{\gamma}_n^*,\hat{\delta}_n^*),(\gamma_0,0))<r_n) \geq \mathbb{P}_{(\gamma_0,0)}(S_n(r_n)) \to 1,$$

which establishes the consistency of the sequence $(\hat{\gamma}_n^*,\hat{\delta}_n^*)$.

□

## Proof of Theorem 2

First, apply a Taylor series expansion of the estimating equations $f_1(\hat{\gamma}_n,\hat{\delta}_n;\rho_0)=0$ and $f_2(\hat{\gamma}_n,\hat{\delta}_n;\rho_0)=0$ around $(\gamma_0,0)$. This gives

$$\begin{aligned}
0 &= f_1(\gamma_0,0;\rho_0)+f_{11}(\gamma_0,0;\rho_0)(\hat{\gamma}_n-\gamma_0)+f_{12}(\gamma_0,0;\rho_0)\hat{\delta}_n \\
&\quad +\frac{1}{2}\left\{f_{111}(\breve{\gamma}_n,\breve{\delta}_n;\rho_0)(\hat{\gamma}_n-\gamma_0)^2+f_{122}(\breve{\gamma}_n,\breve{\delta}_n;\rho_0)\hat{\delta}_n^2+2f_{112}(\breve{\gamma}_n,\breve{\delta}_n;\rho_0)(\hat{\gamma}_n-\gamma_0)\hat{\delta}_n\right\}, \\
0 &= f_2(\gamma_0,0;\rho_0)+f_{21}(\gamma_0,0;\rho_0)(\hat{\gamma}_n-\gamma_0)+f_{22}(\gamma_0,0;\rho_0)\hat{\delta}_n \\
&\quad +\frac{1}{2}\left\{f_{211}(\breve{\gamma}_n,\breve{\delta}_n;\rho_0)(\hat{\gamma}_n-\gamma_0)^2+f_{222}(\breve{\gamma}_n,\breve{\delta}_n;\rho_0)\hat{\delta}_n^2+2f_{122}(\breve{\gamma}_n,\breve{\delta}_n;\rho_0)(\hat{\gamma}_n-\gamma_0)\hat{\delta}_n\right\},
\end{aligned}$$

where $(\breve{\gamma}_n,\breve{\delta}_n)$ is a point on the line segment connecting $(\hat{\gamma}_n,\hat{\delta}_n)$ and $(\gamma_0,0)$. A straightforward rearrangement gives a set of random equations where interest is in $\sqrt{k}(\hat{\gamma}_n-\gamma_0)$ and $\sqrt{k}\hat{\delta}_n$:

$$-\sqrt{k}\begin{bmatrix} f_1(\gamma_0,0;\rho_0) \\ f_2(\gamma_0,0;\rho_0)\end{bmatrix} = \begin{bmatrix} \tilde{f}_{11}(\gamma_0,0;\rho_0) & \tilde{f}_{12}(\gamma_0,0;\rho_0) \\ \tilde{f}_{12}(\gamma_0,0;\rho_0) & \tilde{f}_{22}(\gamma_0,0;\rho_0)\end{bmatrix}\begin{bmatrix}\sqrt{k}(\hat{\gamma}_n-\gamma_0) \\ \sqrt{k}\hat{\delta}_n\end{bmatrix}, \tag{18}$$

where

$$\begin{aligned}
\tilde{f}_{11}(\gamma_0,0;\rho_0) &:= f_{11}(\gamma_0,0;\rho_0)+\frac{1}{2}\left[f_{111}(\breve{\gamma}_n,\breve{\delta}_n;\rho_0)(\hat{\gamma}_n-\gamma_0)+f_{112}(\breve{\gamma}_n,\breve{\delta}_n;\rho_0)\hat{\delta}_n\right], \\
\tilde{f}_{12}(\gamma_0,0;\rho_0) &:= f_{12}(\gamma_0,0;\rho_0)+\frac{1}{2}\left[f_{122}(\breve{\gamma}_n,\breve{\delta}_n;\rho_0)\hat{\delta}_n+f_{112}(\breve{\gamma}_n,\breve{\delta}_n;\rho_0)(\hat{\gamma}_n-\gamma_0)\right], \\
\tilde{f}_{22}(\gamma_0,0;\rho_0) &:= f_{22}(\gamma_0,0;\rho_0)+\frac{1}{2}\left[f_{222}(\breve{\gamma}_n,\breve{\delta}_n;\rho_0)\hat{\delta}_n+f_{122}(\breve{\gamma}_n,\breve{\delta}_n;\rho_0)(\hat{\gamma}_n-\gamma_0)\right].
\end{aligned}$$

Let

$$\overline{\mathbb{A}}_{k,n}(\rho_0) := \begin{bmatrix} \mathbb{A}_{k,n}^{(1)}(-\alpha(1+\gamma_0)/\gamma_0) \\ \mathbb{A}_{k,n}^{(1)}(-(\alpha(1+\gamma_0)-\rho_0)/\gamma_0) \\ \mathbb{A}_{k,n}^{(2)}(-\alpha(1+\gamma_0)/\gamma_0)\end{bmatrix},$$

$\boldsymbol{\mu}$ a vector with elements

$$\begin{aligned}
\mu_1 &:= -\frac{\alpha\rho_0(1+\gamma_0)}{\gamma_0[1+\alpha(1+\gamma_0)][1-\rho_0+\alpha(1+\gamma_0)]}, &(19)\\
\mu_2 &:= -\frac{[\alpha(1+\gamma_0)-\rho_0]\rho_0}{\gamma_0[1-\rho_0+\alpha(1+\gamma_0)][1-2\rho_0+\alpha(1+\gamma_0)]}, \\
\mu_3 &:= \frac{\rho_0(1-\rho_0)-\alpha^2\rho_0(1+\gamma_0)^2}{[1+\alpha(1+\gamma_0)]^2[1-\rho_0+\alpha(1+\gamma_0)]^2}, &(20)
\end{aligned}$$

and $\mathbf{\Sigma}(\rho_0)$ a symmetric $(3\times 3)$ matrix with elements

$$\begin{aligned}
\sigma_{11}(\rho_0) &:= \frac{\alpha^2(1+\gamma_0)^2}{[1+\alpha(1+\gamma_0)]^2[1+2\alpha(1+\gamma_0)]}, && (21)\\
\sigma_{21}(\rho_0) &:= \frac{\alpha(1+\gamma_0)[\alpha(1+\gamma_0)-\rho_0]}{[1+\alpha(1+\gamma_0)][1-\rho_0+\alpha(1+\gamma_0)][1-\rho_0+2\alpha(1+\gamma_0)]}, && (22)\\
\sigma_{22}(\rho_0) &:= \frac{[\alpha(1+\gamma_0)-\rho_0]^2}{[1-\rho_0+\alpha(1+\gamma_0)]^2[1-2\rho_0+2\alpha(1+\gamma_0)]}, && (23)\\
\sigma_{31}(\rho_0) &:= \gamma_0\left(\frac{1}{[1+2\alpha(1+\gamma_0)]^2}-\frac{1}{[1+\alpha(1+\gamma_0)]^3}\right), && (24)\\
\sigma_{32}(\rho_0) &:= \gamma_0\left(\frac{1}{[1-\rho_0+2\alpha(1+\gamma_0)]^2}-\frac{1}{[1+\alpha(1+\gamma_0)]^2[1-\rho_0+\alpha(1+\gamma_0)]}\right), && (25)\\
\sigma_{33}(\rho_0) &:= \gamma_0^2\left(\frac{2}{[1+2\alpha(1+\gamma_0)]^3}-\frac{1}{[1+\alpha(1+\gamma_0)]^4}\right). && (26)
\end{aligned}$$

Note that, from Lemma 1

$$\overline{\mathbb{A}}_{k,n}(\rho_0) \rightsquigarrow N_3(\lambda\boldsymbol{\mu}, \mathbf{\Sigma}(\rho_0)).$$

Now, introduce

$$\mathbb{B}(\rho_0) := \gamma_0^{-\alpha-2}\begin{bmatrix} \gamma_0 & 0 & -1 \\ \gamma_0 & -\gamma_0(1-\rho_0) & 0 \end{bmatrix}, \tag{27}$$

so that

$$\sqrt{k}\begin{bmatrix} f_1(\gamma_0,0;\rho_0) \\ f_2(\gamma_0,0;\rho_0) \end{bmatrix} = \mathbb{B}(\rho_0)\overline{\mathbb{A}}_{k,n}(\rho_0),$$

leading to the weak convergence

$$\sqrt{k}\begin{bmatrix} f_1(\gamma_0,0;\rho_0) \\ f_2(\gamma_0,0;\rho_0) \end{bmatrix} \rightsquigarrow N_2(\lambda\mathbb{B}(\rho_0)\boldsymbol{\mu}, \mathbb{B}(\rho_0)\mathbf{\Sigma}(\rho_0)\mathbb{B}'(\rho_0)).$$

Concerning the terms $\tilde{f}_{st}(\gamma_0,0;\rho_0)$, $s,t=1,2$, we have by Lemma 1, the consistency of $(\hat{\gamma}_n,\hat{\delta}_n)$ and because $|f_{stu}(\gamma,\delta;\rho_0)| \le M_{stu}(\boldsymbol{Y})$, in some open neighborhood of $(\gamma_0,0)$, with $M_{stu}(\boldsymbol{Y}) = O_{\mathbb{P}}(1)$, $s,t,u=1,2$, that $\tilde{f}_{st}(\gamma_0,0;\rho_0) \xrightarrow{\mathbb{P}} f^*_{st}(\gamma_0,0;\rho_0)$, $s,t=1,2$. Let

$$\mathbb{C}(\rho_0) := \begin{bmatrix} f^*_{11}(\gamma_0,0;\rho_0) & f^*_{12}(\gamma_0,0;\rho_0) \\ f^*_{12}(\gamma_0,0;\rho_0) & f^*_{22}(\gamma_0,0;\rho_0) \end{bmatrix}. \tag{28}$$

From the proof of the consistency, we know that $\mathbb{C}(\rho_0)$ is a positive definite matrix, and thus invertible. Then, according to Lemma 5.2 in Chapter 6 of Lehmann and Casella (1998), for the solution of the system of equations (18), we have the following convergence

$$\sqrt{k}\begin{bmatrix} \hat{\gamma}_n-\gamma_0 \\ \hat{\delta}_n \end{bmatrix} \rightsquigarrow N_2(-\lambda\mathbb{C}^{-1}(\rho_0)\mathbb{B}(\rho_0)\boldsymbol{\mu}, \mathbb{C}^{-1}(\rho_0)\mathbb{B}(\rho_0)\mathbf{\Sigma}(\rho_0)\mathbb{B}'(\rho_0)\mathbb{C}^{-1}(\rho_0)).$$

After tedious calculations one can show that $-\mathbb{C}^{-1}(\rho_0)\mathbb{B}(\rho_0)\boldsymbol{\mu} = [0,1]'$. Taking into account that $\sqrt{k}\delta_n \xrightarrow{\mathbb{P}} \lambda$, the theorem follows.

□

## Proof of Proposition 1

The arguments needed to prove the consistency and asymptotic normality are the same as those used in the proofs of Theorem 1 and 2, and therefore we limit ourselves to giving some comments to the main ideas. Concerning the consistency one works with $\widehat{\Delta}_\alpha(\gamma, \delta; \tilde{\rho})$ and its derivatives. Again by Lemma 1 we have that $f_s(\gamma_0, 0; \tilde{\rho}) \xrightarrow{\mathbb{P}} 0$, $s = 1, 2$, and that $f_{st}(\gamma_0, 0; \tilde{\rho}) \xrightarrow{\mathbb{P}} f^*_{st}(\gamma_0, 0; \tilde{\rho})$, $s, t = 1, 2$, leading to the results for $S_1$ and $S_2$. Also for the third order derivatives we can use the same arguments. This establishes the existence and the consistency. To prove the asymptotic normality one uses the same line of argumentation as in Theorem 2, with $\rho_0$ replaced by $\tilde{\rho}$ in $\overline{\mathbb{A}}_{k,n}(\rho_0)$, $\boldsymbol{\Sigma}(\rho_0)$, $\mathbb{B}(\rho_0)$ and $\mathbb{C}(\rho_0)$, and replacing the vector $\boldsymbol{\mu}$ by $\tilde{\boldsymbol{\mu}}$, having as elements $\tilde{\mu}_1 := \mu_1$, $\tilde{\mu}_3 := \mu_3$ and

$$\tilde{\mu}_2 := -\frac{[\alpha(1+\gamma_0) - \tilde{\rho}]\rho_0}{\gamma_0[1 - \tilde{\rho} + \alpha(1+\gamma_0)][1 - \rho_0 - \tilde{\rho} + \alpha(1+\gamma_0)]}. \tag{29}$$

$\square$

## Proof of Theorem 3

To prove the existence and consistency we will condition on the event that $\hat{\rho}_n \in (\rho_0 - \varepsilon, \rho_0 + \varepsilon)$, for some $\varepsilon > 0$, with $\rho_0 + \varepsilon < 0$. We then have that

$$\begin{aligned}
&\mathbb{P}_{(\gamma_0,0)}(\widehat{\Delta}_\alpha(\gamma_0, 0; \hat{\rho}_n) < \widehat{\Delta}_\alpha(\gamma, \delta; \hat{\rho}_n) \text{ for all } (\gamma, \delta) \text{ on the surface of } Q_r) \\
&\quad \geq \mathbb{P}_{(\gamma_0,0)}(\widehat{\Delta}_\alpha(\gamma_0, 0; \hat{\rho}_n) < \widehat{\Delta}_\alpha(\gamma, \delta; \hat{\rho}_n) \text{ for all } (\gamma, \delta) \text{ on the surface of } Q_r | \hat{\rho}_n \in (\rho_0 - \varepsilon, \rho_0 + \varepsilon)) \\
&\quad \times \mathbb{P}(\hat{\rho}_n \in (\rho_0 - \varepsilon, \rho_0 + \varepsilon)).
\end{aligned}$$

By the consistency of $\hat{\rho}_n$ we have that $\mathbb{P}(\hat{\rho}_n \in (\rho_0 - \varepsilon, \rho_0 + \varepsilon)) \to 1$, so it remains to show that

$$\mathbb{P}_{(\gamma_0,0)}(\widehat{\Delta}_\alpha(\gamma_0, 0; \hat{\rho}_n) < \widehat{\Delta}_\alpha(\gamma, \delta; \hat{\rho}_n) \text{ for all } (\gamma, \delta) \text{ on the surface of } Q_r | \hat{\rho}_n \in (\rho_0 - \varepsilon, \rho_0 + \varepsilon)) \to 1.$$

The arguments are again similar to those used in the proof of Theorem 1 and therefore we only give an outline. First make a Taylor series expansion as in (17), though now with $\rho_0$ replaced by $\hat{\rho}_n$.

As for $S_1$, we have that $f_1(\gamma_0, 0; \hat{\rho}_n)$ does not depend on $\hat{\rho}_n$ and therefore $f_1(\gamma_0, 0; \hat{\rho}_n) \xrightarrow{\mathbb{P}} 0$, whereas $f_2(\gamma_0, 0; \hat{\rho}_n)$ can be analyzed by using Lemma 1. In particular, from the proof of Lemma 1

$$A^{(1)}_{k,n}(s) \stackrel{\mathcal{D}}{=} \frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{s\gamma_0} + s\eta(\breve{Y}_{n-k,n})\frac{1}{k}\sum_{j=1}^{k} \tilde{Y}_j^{s\gamma_0}(\tilde{Y}_j^{\rho_0} - 1) + o_{\mathbb{P}}(k^{-1/2}),$$

uniformly in $s \in [s_0, 0]$. Assume that $(-(\alpha(1+\gamma_0) - (\rho_0 - \varepsilon))/\gamma_0, -(\alpha(1+\gamma_0) - (\rho_0 + \varepsilon))/\gamma_0) \subset$

$[s_0, 0]$. Setting $s = -(\alpha(1+\gamma_0) - \hat{\rho}_n)/\gamma_0$, we have that

$$
\begin{aligned}
A_{k,n}^{(1)}(-(\alpha(1+\gamma_0) - \hat{\rho}_n)/\gamma_0) \quad &\stackrel{\mathcal{D}}{=} \quad \frac{1}{1-\rho_0+\alpha(1+\gamma_0)} + \left( \frac{1}{k} \sum_{j=1}^{k} \tilde{Y}_j^{-(\alpha(1+\gamma_0)-\hat{\rho}_n)} - \frac{1}{1-\hat{\rho}_n+\alpha(1+\gamma_0)} \right) \\
&\quad - \frac{\alpha(1+\gamma_0) - \hat{\rho}_n}{\gamma_0}\, \eta(\breve{Y}_{n-k,n}) \frac{1}{k} \sum_{j=1}^{k} \tilde{Y}_j^{-(\alpha(1+\gamma_0)-\hat{\rho}_n)} (\tilde{Y}_j^{\rho_0} - 1) + o_{\mathbb{P}}(1) \\
&=: \quad \frac{1}{1-\rho_0+\alpha(1+\gamma_0)} + T_1 - \frac{\alpha(1+\gamma_0) - \hat{\rho}_n}{\gamma_0}\, \eta(\breve{Y}_{n-k,n}) T_2 + o_{\mathbb{P}}(1).
\end{aligned}
$$

Again from the proof of Lemma 1

$$
|T_1| \le \sup_{\rho \in (\rho_0 - \varepsilon, \rho_0 + \varepsilon)} \left| \frac{1}{k} \sum_{j=1}^{k} \tilde{Y}_j^{-(\alpha(1+\gamma_0)-\rho)} - \frac{1}{1-\rho+\alpha(1+\gamma_0)} \right| \xrightarrow{\mathbb{P}} 0,
$$

and

$$
|T_2| \le \frac{1}{k} \sum_{j=1}^{k} |\tilde{Y}_j^{\rho_0} - 1| \le 2.
$$

Thus

$$
A_{k,n}^{(1)}(-(\alpha(1+\gamma_0) - \hat{\rho}_n)/\gamma_0) \xrightarrow{\mathbb{P}} \frac{1}{1-\rho_0+\alpha(1+\gamma_0)},
$$

and therefore $f_2(\gamma_0, 0; \hat{\rho}_n) \xrightarrow{\mathbb{P}} 0$.

For $S_2$, write

$$
\begin{aligned}
2S_2 \quad = \quad & f_{11}^*(\gamma_0, 0; \rho_0)(\gamma - \gamma_0)^2 + f_{22}^*(\gamma_0, 0; \rho_0)\delta^2 + 2f_{12}^*(\gamma_0, 0; \rho_0)(\gamma - \gamma_0)\delta \\
& + [f_{11}(\gamma_0, 0; \hat{\rho}_n) - f_{11}^*(\gamma_0, 0; \rho_0)](\gamma - \gamma_0)^2 + [f_{22}(\gamma_0, 0; \hat{\rho}_n) - f_{22}^*(\gamma_0, 0; \rho_0)]\delta^2 \\
& + 2[f_{12}(\gamma_0, 0; \hat{\rho}_n) - f_{12}^*(\gamma_0, 0; \rho_0)](\gamma - \gamma_0)\delta.
\end{aligned}
$$

By arguments similar to those used when dealing with $f_2(\gamma_0, 0; \hat{\rho}_n)$, we have that $f_{st}(\gamma_0, 0; \hat{\rho}_n) \xrightarrow{\mathbb{P}} f_{st}^*(\gamma_0, 0; \rho_0)$, $s, t = 1, 2$, and hence we can proceed as in the proof of Theorem 1. Finally, conditionally on $\hat{\rho}_n \in (\rho_0 - \varepsilon, \rho_0 + \varepsilon)$, also the argument for the third order derivatives holds and the proof for the existence and consistency can be completed in the same way as in the proof of Theorem 1.

For what concerns asymptotic normality, we make as before a Taylor series expansion of the estimating equations, leading to (18), though with $\rho_0$ replaced by $\hat{\rho}_n$. Since $\mathbb{P}(\hat{\rho}_n \in (\rho_0 - \varepsilon, \rho_0 + \varepsilon)) \to 1$, we have that (by an appropriate choice of $s_0$ in Lemma 1)

$$
\overline{\mathbb{A}}_{k,n}(\hat{\rho}_n) \rightsquigarrow N_3(\lambda \boldsymbol{\mu}, \boldsymbol{\Sigma}(\rho_0)),
$$

and hence

$$\sqrt{k}\left[\begin{array}{c} f_1(\gamma_0,0;\hat{\rho}_n) \\ f_2(\gamma_0,0;\hat{\rho}_n) \end{array}\right] = \mathbb{B}(\hat{\rho}_n)\overline{\mathbb{A}}_{k,n}(\hat{\rho}_n) \rightsquigarrow N_2(\lambda\mathbb{B}(\rho_0)\boldsymbol{\mu}, \mathbb{B}(\rho_0)\boldsymbol{\Sigma}(\rho_0)\mathbb{B}'(\rho_0)).$$

The rest of the proof is identical to that of Theorem 2.

$\square$


## Acknowledgement

The authors are very grateful to the anonymous referee for the very constructive comments on the paper. The suggestions have definitely improved the presentation of the material.


## References


[1] Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C., 1998. Robust and efficient estimation by minimizing a density power divergence. Biometrika, 85, 549–559.

[2] Beirlant, J., de Wet, T., Goegebeur, Y., 2006. A goodness-of-fit statistic for Pareto-type behavior. Journal of Computational and Applied Mathematics, 186, 99-116.

[3] Beirlant, J., Dierckx, G., Goegebeur, Y., Matthys, G., 1999. Tail index estimation and an exponential regression model. Extremes, 2, 177–200.

[4] Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., 2004. *Statistics of Extremes – Theory and Applications*. Wiley Series in Probability and Statistics.

[5] Beirlant, J., Joossens, E., Segers, J., 2009. Second-order refined peaks-over-threshold modelling for heavy-tailed distributions. Journal of Statistical Planning and Inference, 139, 2800–2815.

[6] Ciuperca, G., Mercadier, C., 2010. Semi-parametric estimation for heavy tailed distributions. Extremes, 13, 55–87.

[7] de Haan, L., Ferreira, A., 2006. *Extreme Value Theory: An Introduction*. Springer.

[8] Drees, H. (1998). On smooth statistical tail functionals. Scandinavian Journal of Statistics, 25, 187–210.

[9] Dupuis, D.J., Victoria-Feser, M.P., 2006. A robust prediction error criterion for Pareto modelling of upper tails. Canadian Journal of Statistics, 34, 639–658.

[10] Feuerverger, A., Hall, P., 1999. Estimating a tail exponent by modelling departure from a Pareto distribution. The Annals of Statistics, 27, 760–781.

[11] Fraga Alves, M.I., Gomes, M.I., de Haan, L., 2003. A new class of semi-parametric estimators of the second order parameter. Portugaliae Mathematica, 60, 193–213.

[12] Goegebeur, Y., Beirlant, J., de Wet, T., 2008. Linking Pareto-tail kernel goodness-of-fit statistics with tail index at optimal threshold and second order estimation. REVSTAT, 6, 51–69.

[13] Goegebeur, Y., Beirlant, J., de Wet, T., 2010. Kernel estimators for the second order parameter in extreme value statistics. Journal of Statistical Planning and Inference, 140, 2632–2652.

[14] Gomes, M.I., de Haan, L., Rodrigues, L.H., 2008. Tail index estimation for heavy-tailed models: accommodation of bias in weighted log-excesses. Journal of the Royal Statistical Society Series B, 70, 31–52.

[15] Gomes, M.I., Figueiredo, F., Neves, M.M., 2012. Adaptive estimation of heavy right tails: resampling-based methods in action. Extremes, 15, 463–489.

[16] Hall, P., 1982. On some simple estimates of an exponent of regular variation. Journal of the Royal Statistical Society Series B, 44, 37–42.

[17] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. *Robust Statistics - The Approach Based on Influence Functions*. Wiley Series in Probability and Mathematical Statistics.

[18] Hill, B., 1975. A simple general approach to inference about the tail of a distribution. The Annals of Statistics, 3, 1163–1174.

[19] Hubert, M., Dierckx, G., Vanpaemel, D., 2012. Detecting influential data points for the Hill estimator in Pareto-type distributions. Computational Statistics & Data Analysis, doi:10.1016/j.csda.2012.07.011.

[20] Juárez, S.F., Schucany, W.R., 2004. Robust and efficient estimation for the generalized Pareto distribution. Extremes, 7, 237–251.

[21] Kim, M., Lee, S., 2008. Estimation of a tail index based on minimum density power divergence. Journal of Multivariate Analysis, 99, 2453–2471.

[22] Koning, A.J., Peng, L., 2008. Goodness-of-fit tests for a heavy tailed distribution. Journal of Statistical Planning and Inference, 138, 3960–3981.

[23] Lehmann, E.L., Casella, G., 1998. *Theory of point estimation*. Springer.

[24] Neves, C., Fraga Alves, M.I., 2008. Testing extreme value conditions – an overview and recent approaches. REVSTAT, 6, 83–100

[25] Peng, L., Welsh, A.H., 2001. Robust estimation of the generalized Pareto distribution. Extremes, 4, 53–65.

[26] van der Vaart, A.W., 2007. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathemetics.

[27] Vandewalle, B., Beirlant, J., Christmann, A., Hubert, M., 2007. A robust estimator for the tail index of Pareto-type distributions. Computational Statistics & Data Analysis, 51, 6252–6268.

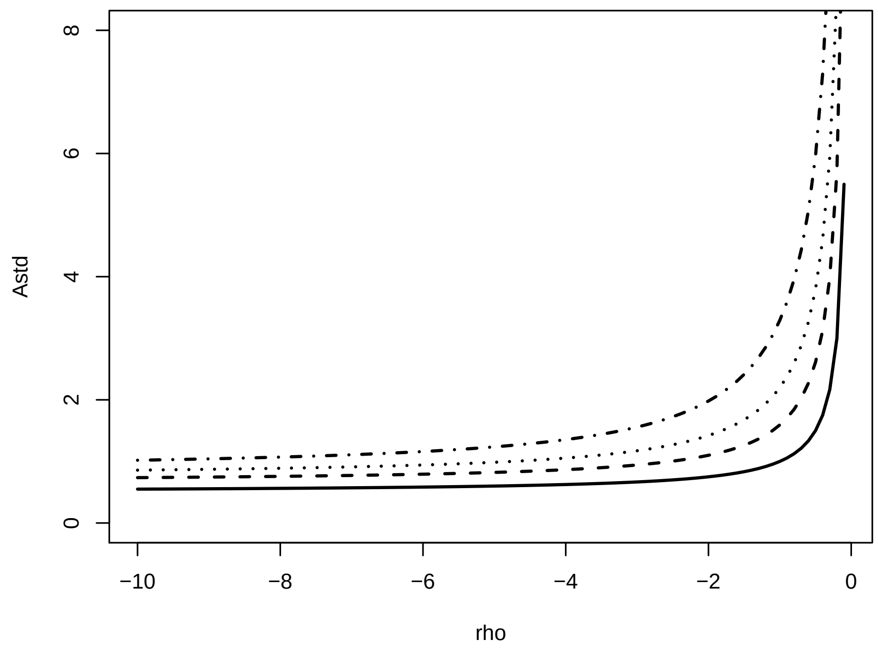


Figure 1: Asymptotic standard deviation of $\hat{\gamma}_n$ as a function of $\rho_0$ in case $\gamma_0 = 0.5:$ $\alpha = 0$ (solid), $\alpha = 0.5$ (dashed), $\alpha = 1$ (dotted), and $\alpha = 2$ (dashed-dotted).

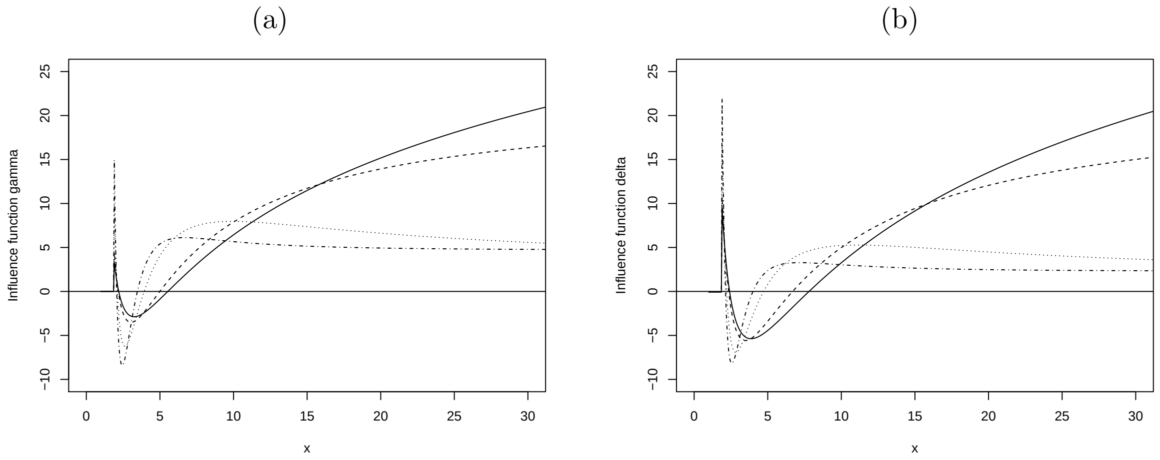


Figure 2: Influence function for (a) the parameter $\gamma$ and (b) the parameter $\delta$, with $\alpha = 0$ (solid line), 0.1 (dashed line), 0.5 (dotted line) and 1 (dashed-dotted line), in case the distribution is Fréchet with $\gamma = 1/2$ and the threshold for estimation is set at quantile 0.75.

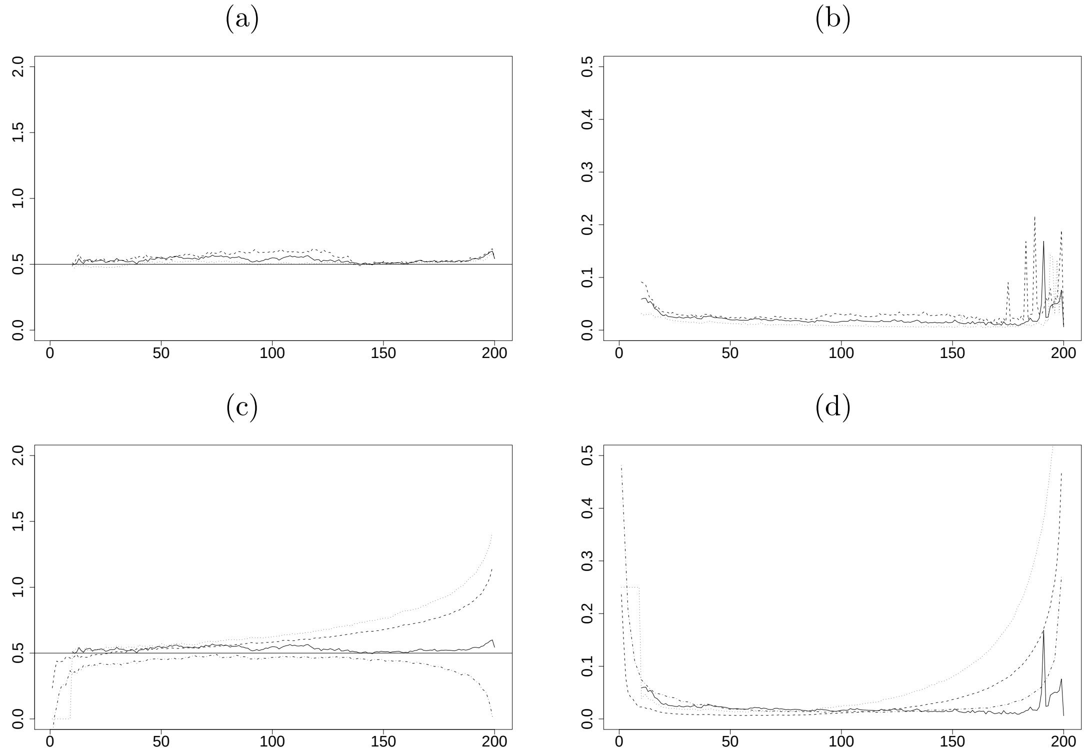


Figure 3: Estimators for $\gamma$ based on 100 samples of size $n = 200$ from an uncontaminated Fréchet(2) distribution. (a) Median and (b) MSE of $\hat{\gamma}_{n,\alpha}$ as a function of $k$ with $\alpha = 0.1$ (dotted), $\alpha = 0.5$ (solid), $\alpha = 1$ (dashed); (c) Median and (d) MSE as a function of $k$ for $\hat{\gamma}_H$ (dashed), $\hat{\gamma}_{KL,0.5}$ (dotted), $\hat{\gamma}_{n,0}$ (dashed-dotted) and $\hat{\gamma}_{n,0.5}$ (solid). The true value $\gamma_0$ is added by a horizontal line.

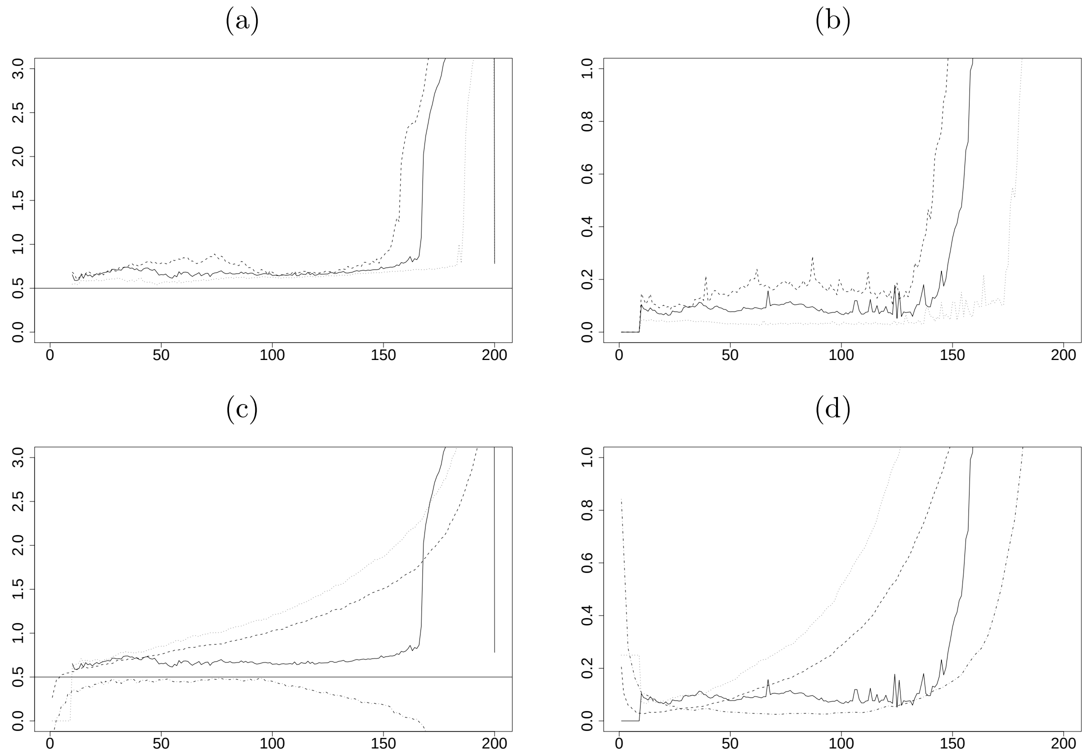


Figure 4: Estimators for $\gamma$ based on 100 samples of size $n = 200$ from an uncontaminated Burr(1,1,2) distribution. (a) Median and (b) MSE of $\hat{\gamma}_{n,\alpha}$ as a function of $k$ with $\alpha = 0.1$ (dotted), $\alpha = 0.5$ (solid), $\alpha = 1$ (dashed); (c) Median and (d) MSE as a function of $k$ for $\hat{\gamma}_H$ (dashed), $\hat{\gamma}_{KL,0.5}$ (dotted), $\hat{\gamma}_{n,0}$ (dashed-dotted) and $\hat{\gamma}_{n,0.5}$ (solid). The true value $\gamma_0$ is added by a horizontal line.

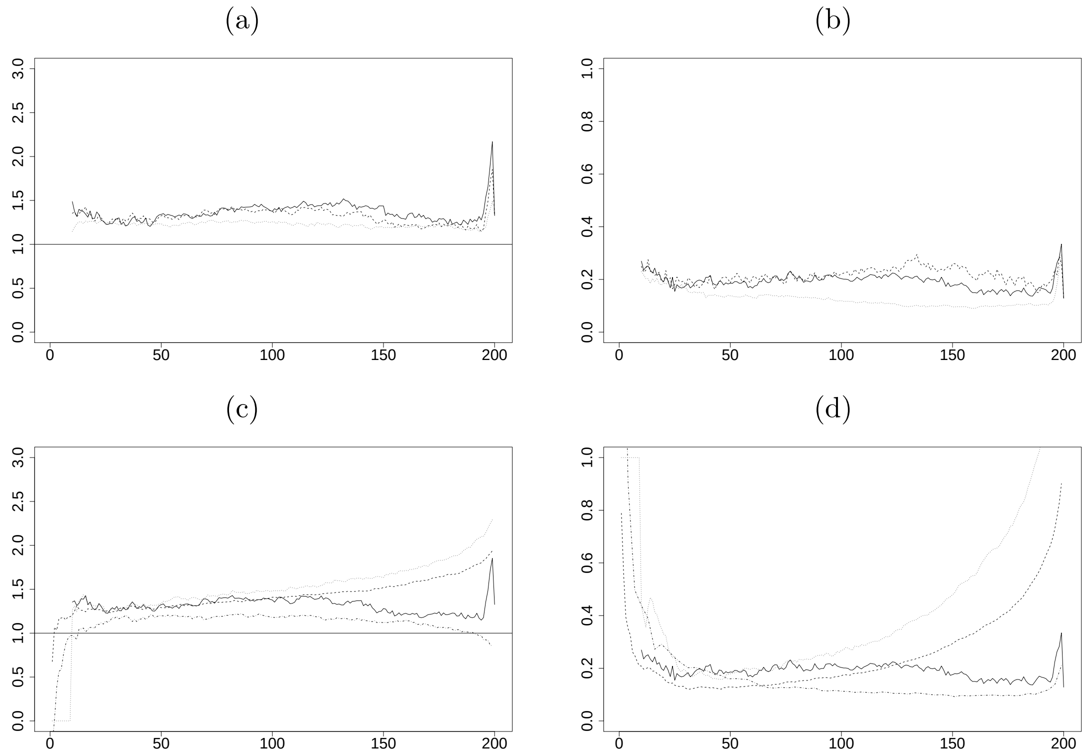


Figure 5: Estimators for $\gamma$ based on 100 samples of size $n = 200$ from an uncontaminated log-gamma distribution. (a) Median and (b) MSE of $\hat{\gamma}_{n,\alpha}$ as a function of $k$ with $\alpha = 0.1$ (dotted), $\alpha = 0.5$ (solid), $\alpha = 1$ (dashed); (c) Median and (d) MSE as a function of $k$ for $\hat{\gamma}_H$ (dashed), $\hat{\gamma}_{KL,0.5}$ (dotted), $\hat{\gamma}_{n,0}$ (dashed-dotted) and $\hat{\gamma}_{n,0.5}$ (solid). The true value $\gamma_0$ is added by a horizontal line.

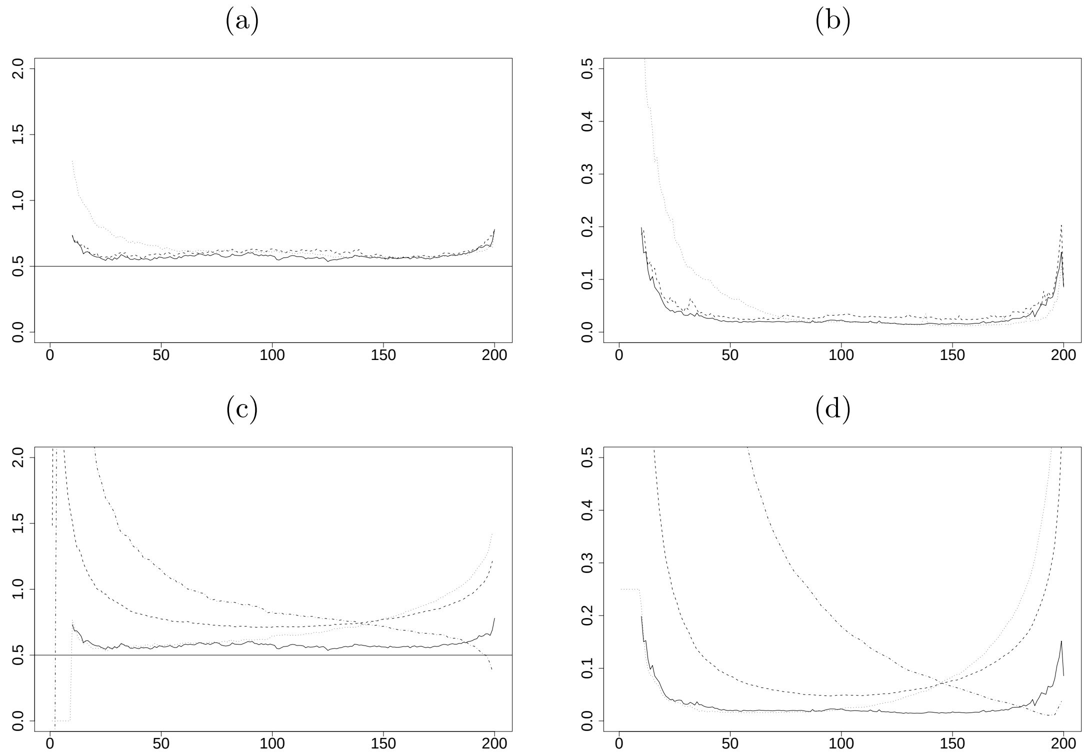


Figure 6: Estimators for $\gamma$ based on 100 samples of size $n = 200$ from a contaminated Fréchet(2) distribution. (a) Median and (b) MSE of $\hat{\gamma}_{n,\alpha}$ as a function of $k$ with $\alpha = 0.1$ (dotted), $\alpha = 0.5$ (solid), $\alpha = 1$ (dashed); (c) Median and (d) MSE as a function of $k$ for $\hat{\gamma}_H$ (dashed), $\hat{\gamma}_{KL,0.5}$ (dotted), $\hat{\gamma}_{n,0}$ (dashed-dotted) and $\hat{\gamma}_{n,0.5}$ (solid). The true value $\gamma_0$ is added by a horizontal line.

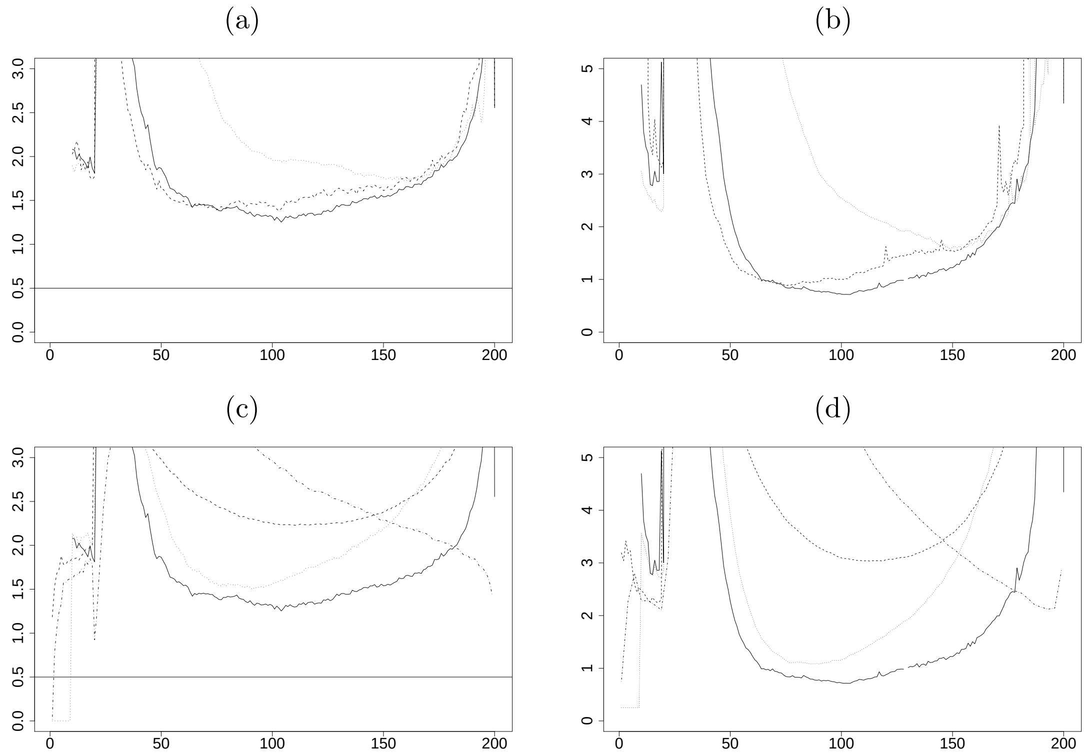


Figure 7: Estimators for $\gamma$ based on 100 samples of size $n = 200$ from a contaminated Burr(1,1,2) distribution. (a) Median and (b) MSE of $\hat{\gamma}_{n,\alpha}$ as function of $k$ with $\alpha = 0.1$ (dotted), $\alpha = 0.5$ (solid), $\alpha = 1$ (dashed); (c) Median and (d) MSE as a function of $k$ for $\hat{\gamma}_H$ (dashed), $\hat{\gamma}_{KL,0.5}$ (dotted), $\hat{\gamma}_{n,0}$ (dashed-dotted) and $\hat{\gamma}_{n,0.5}$ (solid). The true value $\gamma_0$ is added by a horizontal line.

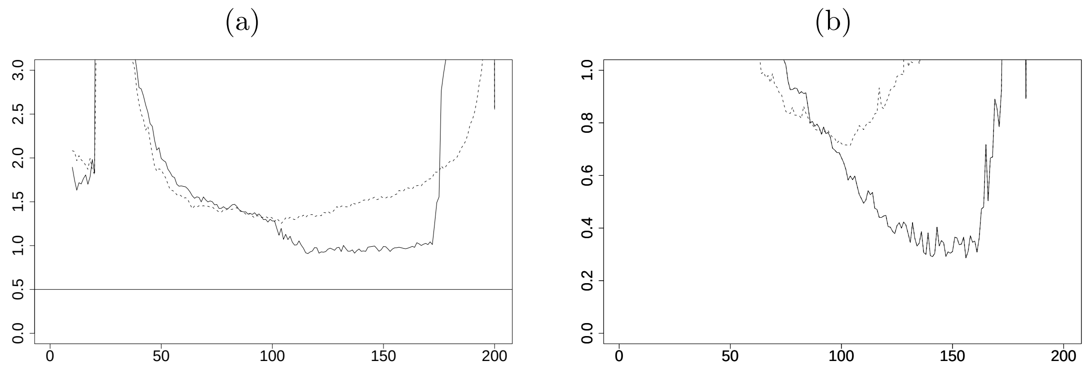


Figure 8: Estimators for $\gamma$ based on 100 samples of size $n = 200$ from a contaminated Burr(1,1,2) distribution. (a) Median and (b) MSE of $\hat{\gamma}_{n,0.5}$ as a function of $k$ with $\rho$ fixed at -1 (solid) and $\rho$ estimated (dashed). The true value $\gamma_0$ is added by a horizontal line.