Article

A Robust Prediction Error Criterion for Pareto Modeling of Upper Tails

DUPUIS, Debbie, VICTORIA-FESER, Maria-Pia

Abstract

Estimation of the Pareto tail index from extreme order statistics is an important problem in many settings. The upper tail of the distribution, where data are sparse, is typically fitted with a model, such as the Pareto model, from which quantities such as probabilities associated with extreme events are deduced. The success of this procedure relies heavily not only on the choice of the estimator for the Pareto tail index but also on the procedure used to determine the number k of extreme order statistics that are used for the estimation. The authors develop a robust prediction error criterion to choose k and estimate the Pareto index. A simulation study shows the good performance of the new estimator and the analysis of real data sets shows that a robust procedure for selection, and not just for estimation, is needed.

The Canadian Journal of Statistics
Vol. 34, No. 4, 2006, Pages ???-???
La revue canadienne de statistique

A robust prediction error criterion for Pareto modelling of upper tails

Debbie J. DUPUIS and Maria-Pia VICTORIA-FESER

Key words and phrases: Extreme values; tail index; income distribution; value at risk; M-estimators; regression.
MSC 2000: Primary 62F35; secondary 62P05.

Abstract: Estimation of the Pareto tail index from extreme order statistics is an important problem in many settings. The upper tail of the distribution, where data are sparse, is typically fitted with a model, such as the Pareto model, from which quantities such as probabilities associated with extreme events are deduced. The success of this procedure relies heavily not only on the choice of the estimator for the Pareto tail index but also on the procedure used to determine the number k of extreme order statistics that are used for the estimation. The authors develop a robust prediction error criterion to choose k and estimate the Pareto index. A simulation study shows the good performance of the new estimator and the analysis of real data sets shows that a robust procedure for selection, and not just for estimation, is needed.

Un critère robuste de prévision pour la modélisation Pareto des ailes supérieures

Résumé : L'estimation de l'indice de Pareto à partir des statistiques d'ordre extrêmes est un problème important dans plusieurs disciplines. L'aile supérieure de la distribution, où les données sont rares, est souvent ajustée par un modèle tel que le modèle de Pareto. Les probabilités des événements extrêmes sont ensuite calculées à partir de ce dernier. Le succès de cette procédure dépend beaucoup non seulement du choix de l'estimateur de l'indice, mais aussi de la procédure utilisée pour déterminer le nombre de statistiques d'ordre extrêmes inclus dans l'estimation. Les auteurs développent un critère robuste de prévision pour la sélection de k et l'estimation de l'indice de Pareto. Une étude de simulation démontre la bonne performance du nouvel estimateur et une analyse de données réelles démontre qu'une procédure robuste de sélection, et non seulement d'estimation, est requise.

1. INTRODUCTION

Whether it is in economics, finance, insurance, engineering, or environmental issues, there is much interest in the upper tails of distributions. In economics, data in the upper tails of income distributions can be sparse and therefore tails are fitted with parametric models (typically the Pareto distribution) in order to properly estimate inequality measures or Lorenz curves (see e.g. Cowell & Victoria-Feser, 2006). In finance, determination of the value at risk or expected shortfall from the lower tail of the returns' distribution is central to portfolio management. The lower tail is modelled for better estimation of these risk measures, and much work has been done on foreign exchange rates and stock returns: Koedijk, Schafgans & de Vries (1990), Jansen & de Vries (1991), Hols & de Vries (1991), Phillips, McFarland & McMahon (1996), Danielsson & de Vries (1997), and Dacorogna, Müller, Pictet & de Vries (2001). Finally, in insurance, a similar argument holds for

the estimation of probabilities associated with given levels of losses; see Embrechts, Klüppelberg & Mikosch (1997) for a full treatment.

Let $X_1, X_2, \ldots, X_n$ be a sequence of positive independent and identically distributed random variables, each with distribution function F. We are interested here in the upper tail of the sample, namely $X_{[n-k+1]}, \ldots, X_{[n]}$ with $X_{[i]}$ denoting the ith order statistic. Our interest lies in Pareto type tails and we suppose that for sufficiently large quantiles x, F is such that there exists a positive constant θ for which $1 - F(x) = x^{-\theta} l(x)$, where $l(x)$ is a so-called slowly varying function at infinity (see e.g. Beirlant, Vynckier & Teugels, 1996). In this paper we choose the function $l(x)$ corresponding to the Pareto model, but our results can be extended to other models (see below). Moreover, as our simulation study will show, our method works well with data from other thick-tailed distributions. We wish to estimate the Pareto index $\theta > 0$ (or $1/\theta$) and simultaneously determine k, the number of observations in the upper-tail to be included in the estimation. Here, we focus on estimation of the upper-tail, but results apply to the lower-tail after proper relabelling.

The problem of estimating the parameter of interest, i.e. the Pareto index $\theta > 0$, raises two important challenges. The first is the simultaneous determination of k, or equivalently the threshold x_0 above which observations in the upper-tail are included in the estimation, and the second is the choice of the estimator. For the threshold, a compromise should be sought between bias and variance: choosing a threshold too close to the central data will cause bias and selecting too extreme a threshold will yield large variances for the resulting estimator.

For the choice of the Pareto index estimator, we consider here the problem of robust estimation. At first sight, a robust analysis seems to contradict an extreme value analysis. It is well known that in general, robust estimators "downweight extreme data", and in extreme value analysis, one is actually interested in extreme data. However, "extreme" does not have the same meaning in robust statistics as in extreme value analysis. Indeed, extreme data for a robust procedure, i.e. outliers, means data that are in some sense "far from the model" that has supposedly generated the majority of the data ("far" may be quantified in more than one way; we say more about this in Section 2.3). If a model is postulated for extreme values, it is by definition adequate for the extremes, and a robust estimator does not downweight these values which constitute the core of the model. On the other hand, one can reasonably expect that not all the data in the upper tail of the sample follow exactly the postulated model.

Formally, let F_θ be the postulated (parametric) model for the upper tail of the sample and let x_0 be the quantile above which F_θ is the correct model. The cumulative distribution on the whole range of x (not only the upper tail) is then given by

$$F(x) = \begin{cases} G(x), & x \leq x_0, \\ G(x_0) + (1 - G(x_0))F_\theta(x), & x \geq x_0, \end{cases} \tag{1}$$

where G is an unknown distribution function defined on the real line. We consider the following contamination model. We suppose that the data are generated from F_ε, where F_ε is as F in (1) but with $F_\theta(x)$ replaced by $(1-\varepsilon)F_\theta(x) + \varepsilon H(x)$, where ε is small, and where $H(x)$ is a distribution function defined on (x_0, ∞). For example, $H(x)$ can be a point mass distribution at an arbitrary point z or it can be $F_{\theta'}$ for $\theta' \neq \theta$. This type of model contamination is not as general as the mixture $(1-\varepsilon)F(x) + \varepsilon H(x)$, but since the deviations will be measured only on the upper tail of the distribution (i.e. where the parametric model is assumed), there is no loss of generality in assuming the neighbourhood defined by F_ε.

The effect of data generated by F_ε in the estimation procedure can be important in that not only can the Pareto index estimators be biased, but also any estimator of k that uses these index estimators can be biased. In other words, because the Pareto index and k are simultaneously estimated, gross errors in the upper tail of the sample can lead to biased estimates of both. For example, a biased estimate of k could lead to some observations from the lower tail of the sample entering the upper-tail sample used for the estimation of the Pareto index. The estimate of the latter would then also be biased.

In this paper, we therefore propose an approach robust to *both* challenges: (1) the estimation of the Pareto index θ, and (2) the simultaneous selection of the suitable number k of order statistics in the upper tail to use in that estimation. To achieve this, we will view the Pareto model as a regression model, e.g. see Beirlant, Vynckier & Teugels (1996).

Consider the conditional distribution $P\left(X/x_0 < x | X > x_0\right)$ of relative excesses over high thresholds x_0. This conditional distribution is known to converge to $1 - x^{-\theta}$ for all $x > 1$, leading to the Pareto (1896) model

$$F_\theta(x) = 1 - \left(\frac{x}{x_0}\right)^{-\theta}, \quad x \geq x_0, \tag{2}$$

with density

$$f(x;\theta) = \theta x^{-(\theta+1)} x_0^\theta, \quad x \geq x_0, \tag{3}$$

and with $F_\theta^{-1}(q) = x_0 \left(1-q\right)^{-1/\theta}$. Rearranging (2) one gets

$$\log\left(\frac{x}{x_0}\right) = -\frac{1}{\theta}\log\left(1 - F_\theta(x)\right), \quad x > x_0, \tag{4}$$

showing that there is a linear relationship between the log of the $x > x_0$ and the log of the survival function. Let $Q(F;q) = \inf\{x | F(x) \geq q\}$ and let $X^*_{[i]}$, $i = 1, \ldots, k$, be the ordered largest k observations, so that $X^*_{[i]} = Q(F_{(n)}; i/(k+1))$, with $F_{(n)}$ the empirical distribution of $X^*_{[i]}$. The empirical counterpart of (4) is the Pareto quantile plot

$$\log\left(\frac{Q(F_{(n)}; i/(k+1))}{x_0}\right) = -\frac{1}{\theta}\log\left(\frac{k+1-i}{k+1}\right), \quad i = 1, \ldots, k. \tag{5}$$

The plot of $\log\left(X_{[i]}\right)$ versus $-\log\left((n+1-i)/(n+1)\right)$, $i = 1, \ldots, n$ is often used to detect graphically the quantile $X_{[i]}$ above which the Pareto relationship is valid, i.e. the point above which the plot yields a straight line. We note that there is a clear relationship between x_0 and k in that $k = \sum_{i=1}^n I(X_{[i]} \geq x_0)$, I being the indicator function.

A general approach in determining k is the minimization of an estimate of the asymptotic mean squared error (AMSE) of the estimator of θ. The classical estimator of θ is the MLE

$$\widehat{\theta} = \left[\frac{1}{k}\sum_{i=1}^{k} \log X_{[n-i+1]} - \log X_{[n-k]}\right]^{-1}. \tag{6}$$

The latter is derived in Hill (1975) and is known as the Hill estimator. In this paper we use another criterion to determine k, namely a prediction error criterion that is estimated robustly. Proceeding as in Ronchetti & Staudte (1994), we develop a robust prediction error criterion based on the Pareto quantile plot estimate. We call this criterion the *RC*-criterion and minimize it in order to find x_0 and thus indirectly k. The *RC*-criterion depends on the choice of a robust estimator for the Pareto index, an estimator which depends in turn on a value for x_0. We will consider suitable estimators in the class of weighted maximum likelihood estimators (WMLE) of Dupuis & Morgenthaler (2002) which downweight observations that are "far" from the Pareto model in terms of either probabilities associated to the Pareto model (2) or the size of the residuals with respect to the Pareto regression model.

An anonymous referee has suggested that we introduce measurement error in the model to capture some or all of the model contamination. In our approach, this means replacing the Pareto model with a more flexible model, which could also be another thick-tailed distribution. Our *RC*-criterion could be derived for these models. However, the calculations can become more complex and lead to increased computational cost. Moreover, independently of the complexity of the postulated models, there is no guarantee that real data will follow exactly these models, or in

other words that measurement error models will capture all of the model contamination, so that a robust approach is also needed.

It should be stressed that some authors have already shown concern about the non-robustness of statistics used in risk theory for insurance or finance, e.g. excess of loss premiums and probability of ruin as studied by Marceau & Rioux (2001) and Brazauskas (2003). Both authors showed that robust parametric fitting of the extreme value distribution is the suitable approach for the robust calculation of these statistics. Robust estimation of extreme value distributions has been studied by Victoria-Feser & Ronchetti (1994) and Vandewalle, Beirlant & Hubert (2004) for the Pareto model, by Peng & Welsh (2001) for the Generalized Pareto model, and by Dupuis & Field (2004) for these and other distributions. All these proposals assume that the threshold x_0 is known. On the other hand, Dupuis (1999) considered using robust methods for threshold selection only.

The remainder of the paper is organized as follows. In Section 2, a general formulation for the robust prediction error is presented and then developed more precisely for our Pareto approach. This leads to our *RC*-criterion for determining k. The criterion requires a robust estimator for θ and these are also discussed in Section 2. In Section 3, we draw comparisons between our criteria and presently available options. In Section 4, a simulation study is presented to complete our comparisons. Applications to finance and to economic data are given in Section 5.

2. ROBUST PREDICTION ERROR CRITERION

2.1 General formulation

Let $Y = (Y_1, \ldots, Y_n)^T$ be a random sample of observations from a distribution F. Consider the most general case of Y having covariance Σ. Given a model, let $\hat{Y} = \left(\hat{Y}_1, \ldots, \hat{Y}_n\right)^T$ be the predicted values for Y. One can define a prediction error criterion as

$$\frac{1}{n}\mathrm{tr}\left\{\mathrm{E}\left[\Sigma^{-1}\left(\hat{Y} - \mathrm{E}[Y]\right)\left(\hat{Y} - \mathrm{E}[Y]\right)^T\right]\right\}. \tag{7}$$

In the case of independent observations, (7) reduces to

$$\Gamma = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\left(\frac{\hat{Y}_i - \mathrm{E}[Y_i]}{\sigma_i}\right)^2\right], \tag{8}$$

where $\sigma_i^2 = \mathrm{var}(Y_i)$. For an arbitrary $n \times n$ covariance matrix (7) may be numerically intractable. Therefore, even when observations are dependent, one might resort to (8) as a prediction error criterion.

However, prediction error criterion (8) gives equal weight to all observations and if $\hat{Y}_i$ are obtained from non-robust estimators of model parameters, estimators of (8) will be sensitive to outliers and other departures from model assumptions. Following Ronchetti and Staudte (1994), we define a rescaled mean squared weighted prediction error

$$\Gamma_R = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\hat{w}_i^2\left(\frac{\hat{Y}_i - \mathrm{E}[Y_i]}{\sigma_i}\right)^2\right], \tag{9}$$

where $\hat{w}_i \in [0,1]$ is the fitted weight of the i^{th} observation under a robust fit of the model, to yield $\hat{Y}_i$.

LEMMA 1. *The rescaled mean squared weighted prediction error can be written as*

$$\Gamma_R = \frac{1}{n}E\left[\sum_{i=1}^{n}\hat{w}_i^2\left(\frac{Y_i - \hat{Y}_i}{\sigma_i}\right)^2\right] + \frac{2}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}cov\left[\hat{w}_i Y_i, \hat{w}_i\hat{Y}_i\right] - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}var\left[\hat{w}_i Y_i\right]. \tag{10}$$

The proof uses standard techniques and the derivations can be found in Dupuis & Victoria-Feser (2005). We seek an unbiased estimator of Γ_R and choose

$$C_R = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i^2 \left(\frac{Y_i - \hat{Y}_i}{\sigma_i}\right)^2 + \frac{2}{n}\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\text{cov}\left[\hat{w}_i Y_i, \hat{w}_i \hat{Y}_i\right] - \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\text{var}\left[\hat{w}_i Y_i\right], \tag{11}$$

with suitable estimators for σ_i^2, $\text{cov}\left[\hat{w}_i Y_i, \hat{w}_i \hat{Y}_i\right]$, and $\text{var}\left[\hat{w}_i Y_i\right]$.

2.2 Prediction error criteria for the Pareto model

For the Pareto model, Γ_R is applied to the upper tail of the sample, i.e. where the Pareto model is supposed to have generated the data. Then, given a value for x_0 (and hence for k), in order to apply (5) we take $Y_i = \log\left(X^*_{[i]}/x_0\right), i = 1, \ldots, k$ and $\hat{Y}_i = -1/\hat{\theta}\log\left[(k+1-i)/(k+1)\right], i = 1, \ldots, k$, where $\hat{\theta}$ is an estimator of θ. There are a few options for the latter estimator and their merits are discussed in the next section. To find suitable estimators for the terms in (11) we make use of the following results.

LEMMA 2. *Given a value for x_0, the ordered k quantiles $X^*_{[i]} \geq x_0$ have density*

$$f_{i:k}(x) \quad = (k-(i-1))\tbinom{k}{i-1}\left\{1 - \left(\tfrac{x}{x_0}\right)^{-\theta}\right\}^{i-1}\left\{\left(\tfrac{x}{x_0}\right)^{-\theta}\right\}^{k-(i-1)}\theta x^{-1} \quad \text{for } x \geq x_0.$$

LEMMA 3. *The variance of Y_i is*

$$\sigma_i^2 = \sum_{j=1}^{i}\frac{1}{\theta^2(k-i+j)^2} = \frac{1}{\theta^2}\left[\frac{1}{k^2} + \frac{1}{(k-1)^2} + \ldots + \frac{1}{(k+1-i)^2}\right]. \tag{12}$$

The proof of Lemma 2 is straightforward and that of Lemma 3 is in the Appendix.

In a first instance, we consider a very special case. If θ is replaced by the MLE $\hat{\theta}$ and we set $\hat{w}_i = 1$ for all i, we have the following result.

PROPOSITION 1. *Let*

$$Y_i = \log\left(\frac{X^*_{[i]}}{x_0}\right) = \log\left(\frac{Q(F_{(n)}; i/(k+1))}{x_0}\right), \tag{13}$$

and for the predicted values

$$\hat{Y}_i = -\frac{1}{\hat{\theta}}\log\left(\frac{k+1-i}{k+1}\right), \tag{14}$$

where $\hat{\theta}$ is the Hill estimator $\hat{\theta} = \left[\frac{1}{k}\sum_{i=1}^{k}\log X^*_{[i]} - \log x_0\right]^{-1}$, an estimator of Γ_R is (up to $O(1/k)$) given by

$$\begin{aligned} C(x_0) = \quad & \frac{\hat{\theta}^2}{k}\sum_{i=1}^{k}\left[\frac{1}{k^2} + \frac{1}{(k-1)^2} + \ldots + \frac{1}{(k+1-i)^2}\right]^{-1}\left[\log\left(\frac{X^*_{[i]}}{x_0}\right) + \frac{1}{\hat{\theta}}\log\left(\frac{k+1-i}{k+1}\right)\right]^2 \\ & + \frac{2}{k^2}\sum_{i=1}^{k}\left[\frac{1}{k^2} + \frac{1}{(k-1)^2} + \ldots + \frac{1}{(k+1-i)^2}\right]^{-1}\log\left(\frac{k+1-i}{k+1}\right)^2 - 1\,. \end{aligned} \tag{15}$$

The proof is in the Appendix. When we choose x_0 to minimize (15), we refer to this as the C-criterion. Recall that choosing x_0 simultaneously establishes k since $k = \sum_{i=1}^{n} I(X_{[i]} \geq x_0)$. The performance of this new classical (i.e. non-robust) criterion will be assessed in Section 4. We can however evaluate (11) more generally to yield a robust criterion. First, note that an estimator of $\text{var}\left[\hat{w}_i Y_i\right]$ is $\text{E}\left[\hat{w}_i^2 Y_i^2\right] - \text{E}\left[\hat{w}_i Y_i\right]^2$ where

$$\text{E}\left[\hat{w}_i^j Y_i^j\right] = \int_{x_0}^{\infty} \hat{w}(x)^j \log(x/x_0)^j f_{i:k}(x) dx, \tag{16}$$

and where both $\hat{w}(x)$ and $f_{i:k}$ depend on θ which is replaced by $\hat{\theta}$. Similarly, an estimator of $\text{cov}\left[\hat{w}_i Y_i, \hat{w}_i \hat{Y}_i\right]$ is then $\text{E}\left[\hat{w}_i^2 Y_i \hat{Y}_i\right] - \text{E}\left[\hat{w}_i Y_i\right] \text{E}\left[\hat{w}_i \hat{Y}_i\right]$ where

$$\text{E}\left[\hat{w}_i \hat{Y}_i\right] = -\int_{x_0}^{\infty} \frac{1}{\hat{\theta}} \log\left(\frac{k+1-i}{k+1}\right) \hat{w}(x) f_{i:k}(x) dx. \tag{17}$$

Also similarly,

$$\text{E}\left[\hat{w}_i^2 Y_i \hat{Y}_i\right] = -\int_{x_0}^{\infty} \frac{1}{\hat{\theta}} \log\left(\frac{k+1-i}{k+1}\right) \hat{w}(x) log(x/x_0) f_{i:k}(x) dx. \tag{18}$$

Integrals (16)-(18) can be evaluated either analytically or numerically for the MLE of θ (see Proposition 1) . Any robust estimator of θ will be the solution of an implicit equation and we choose to use Monte Carlo simulations. Sufficient accuracy was obtained with 1000 simulations.

Substituting (12) and Monte Carlo estimates in (11), replacing θ by the robust estimators described below (which depend on the data and on x_0), and replacing the weights $\hat{w}$ by the weights used in the robust estimation, one obtains an *estimated* robust prediction error $C_R(x_0)$. We propose to choose x_0 so as to minimize the latter and refer to this as the RC-criterion. While choosing x_0, we simultaneously find $k = \sum_{i=1}^{n} I(X_{[i]} \geq x_0)$ and $\widehat{\theta}$ in a robust fashion.

A minimal number of order statistics are required for the criterion to be properly evaluated according to our numerical approach as otherwise Monte Carlo error in numerical approximations to integrals (16)-(18) overwhelm estimates of (11). A simulation study (results not shown) established $k = 20$ to be sufficient.

2.3 Possible robust estimators

In this section we propose two robust estimators for the parameter θ of model (2) for a given value of x_0. They are computed on the $k = \sum_{i=1}^{n} I(X_{[i]} \geq x_0)$ largest observations. We propose to implement the weighted maximum likelihood approach of Dupuis & Morgenthaler (2002). This call for the computation of an M-estimator, defined as the solution $\hat{\theta}$ in θ of

$$\sum_{i=1}^{k} \psi(X_{[i]}^*; \theta) = 0, \tag{19}$$

with suitable (mild) conditions on ψ (see Huber, 1981). The weighted MLE (WMLE) has $\psi(x;\theta) = w(x;\theta)\frac{\partial}{\partial\theta}\log(x;\theta)$ where $w(x;\theta)$ is a weight function with values in $[0,1]$. Depending on the model and the choice of the weight function, the resulting WMLE can be biased so that Dupuis & Morgenthaler (2002) propose a bias-corrected WMLE as

$$\widetilde{\theta} = \widehat{\theta} - B(\widehat{\theta}), \tag{20}$$

where

$$B(\hat{\theta}) = -\frac{\int \left(w(x;\theta)\frac{\partial}{\partial\theta}\log f(x;\theta)\right)\big|_{\hat{\theta}}\, dF_{\hat{\theta}}(x)}{\int \left(\frac{\partial}{\partial\theta}w(x;\theta)\frac{\partial}{\partial\theta}\log f(x;\theta) + w(x;\theta)\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right)\big|_{\hat{\theta}}\, dF_{\hat{\theta}}(x)}. \tag{21}$$

The final weight attributed to each data point x is computed using the bias-corrected WMLE $\widetilde{\theta}$, i.e. it is $w(x;\widetilde{\theta})$. Different weight functions $w(x;\theta)$ will lead to different robust estimators. We note that Victoria-Feser & Ronchetti (1994) use optimal bias robust estimators (OBRE) to fit the Pareto model robustly, but since it is rather complex computationally we prefer to use a WMLE.

One possibility for the weighting function is probability based weighting (Field & Smith, 1994) where

$$w(x;\theta) = \begin{cases} F_\theta(x)/p_1, & \text{if } F_\theta(x) < p_1, \\ 1, & \text{if } p_1 < F_\theta(x) < 1-p_2, \\ \{1-F_\theta(x)\}/p_2, & \text{if } F_\theta(x) > 1-p_2, \end{cases} \tag{22}$$

with $F_\theta(x)$ is as in (2) and p_1 and p_2 being constants regulating the amount of robustness. This type of downweighting was used by Dupuis & Morgenthaler (2002) in the context of bivariate extreme values. Any points which do not lie in the central p_1 to $1-p_2$ part of the distribution, determined by the value of θ under consideration, will be smoothly downweighted. With the relatively simple form of the Pareto distribution and weighting scheme (22), it is possible to obtain an analytical expression for the bias correction term (21).

LEMMA 4. *A WMLE of θ, where $F_\theta(x)$ is as in (2) and $w(x;\theta)$ is as in (22), has bias correction term $B(\theta)$ equal to*

$$\left(\frac{\theta}{2}\right) \frac{2(1-p_1)^2\log(1-p_1) + p_1(1-p_1) + p_1(1-p_2) + 2p_1p_2\log p_2}{[(1-p_1)\log(1-p_1)]^2 - p_1(1-p_1) - p_1(1-p_2) + p_1p_2(\log p_2)^2}. \tag{23}$$

The proof is in the Appendix.

The question of a suitable choice for the robustness tuning constants p_1 and p_2 is in general made on the basis of efficiency arguments. The latter is measured as the ratio between the variances of the MLE and the robust estimator of the Pareto index. The former can be shown to be θ^2/k, while the latter is $-\frac{1}{k}\left[\int_{x_0}^{\infty}\frac{\partial}{\partial x}\psi(x;\theta)dF_\theta(x)\right]^{-2}\int_{x_0}^{\infty}\psi(x;\theta)^2dF_\theta(x)$. The larger the tuning constants, the more robust the resulting estimator, but also the less efficient. Simulations show that $\approx 95\%$ efficiency is achieved at the Pareto model for $p_1 = p_2 = 0.005$, for any value of θ. Larger values of p_1 and p_2 lead to more robust, but less efficient, estimators. While suitable in many applications, this form of downweighting seems somewhat unnatural in the case of the Pareto distribution and estimation of its tail index θ since the most upper-tail is systematically downweighted.

We propose the following more appealing downweighting option. Recall that if the Pareto relationship is valid, a plot of $\log\left(X^*_{[i]}\right)$ versus $-\log\left[(k+1-i)/(k+1)\right]$, $i = 1,\ldots,k$ will yield a straight line. While the complicated dependence structure of the response variables does not allow for regression to be used for estimation of θ (and any regression which ignores these dependences yields very poor results), we can make use of the *standardized residual* $r_i = (Y_i - \hat{Y}_i)/\sigma_i$ where Y_i is as in (13), $\hat{Y}_i$ is as in (14), σ_i^2 is as in (12), and $\hat{\theta}$ is the WMLE. To establish the weight, we take a Huber type approach and set

$$w(X^*_{[i]};\theta) = \begin{cases} 1, & \text{if } |r_i| < c, \\ c/|r_i|, & \text{if } |r_i| > c, \end{cases} \tag{24}$$

where c is a constant regulating the amount of robustness. Points lying far away from the Pareto regression line, after accounting for non-constant variance, are not well fit by the Pareto model and downweighted. For this weighting scheme, we can use the following result:

LEMMA 5. *A WMLE of θ where $F_\theta(x)$ is as in (2) and $w(x;\theta)$ is as in (24) has approximate bias correction term $\widehat{B(\theta)}$ equal to*

$$-\frac{\sum_{i=1}^{k}\left(w(X^*_{[i]};\theta)\frac{\partial}{\partial\theta}\log f(X^*_{[i]};\theta)\right)\Big|_{\hat{\theta}}\left(F_{\hat{\theta}}(X^*_{[i]}) - F_{\hat{\theta}}(X^*_{[i-1]})\right)}{\sum_{i=1}^{k}\left(\frac{\partial}{\partial\theta}w(X^*_{[i]};\theta)\frac{\partial}{\partial\theta}\log f(X^*_{[i]};\theta) + w(X^*_{[i]};\theta)\frac{\partial^2}{\partial\theta^2}\log f(X^*_{[i]};\theta)\right)\Big|_{\hat{\theta}}\left(F_{\hat{\theta}}(X^*_{[i]}) - F_{\hat{\theta}}(X^*_{[i-1]}\right)},$$

Table 1: Efficiencies of the WMLE with regression weighting estimation of θ and with k known, based on 50,000 simulations.

		c					
		2.5	2.25	2.0	1.75	1.50	1.25
	10	0.73	0.73	0.72	0.70	0.65	0.56
k	50	0.74	0.73	0.72	0.72	0.70	0.65
	100	0.81	0.81	0.80	0.80	0.75	0.60
	500	0.93	0.92	0.86	0.66	0.40	0.22

where $X^*_{[0]}$ is set to x_0.

Proof of Lemma 5. *The last expression is simply (21) with the integrals discretized over the sample of order statistics.*

As before, the value of the tuning constant c can be chosen based on efficiency. Although 95% efficiency seems a reasonable amount *a priori*, it is however difficult to simultaneously maintain 95% efficiency and guard against the contamination that can appear in real data sets. The contamination can simply be too large and we have to lose efficiency to gain robustness. How does one proceed in practice? We suggest using both a slightly robust ($\approx$ 95% efficient) and very robust ($\approx$ 60% efficient) criterion. In the absence of contamination, the results will be almost the same and we will proceed with confidence. When contamination is present, the more robust criterion will identify the problematic points and give us the required robustness and further insights into our data. Table 1 gives both small-sample and large-sample efficiencies for different values of the tuning parameter c as obtained through simulations with 50,000 replications.

3. OTHER CRITERIA

Our approach is unique since others have developed methods based on estimates of the AMSE of the Hill estimator to determine k. Here, we briefly explain some of these methods and draw comparisons with our C-criterion where appropriate. Hall & Welsh (1985) and Beirlant, Dierckx, Goegebeur & Matthys (1999) compare some of the AMSE-based estimators.

Beirlant, Vynckier & Teugels (1996) obtain an estimate for the optimal k by minimizing a nonparametric estimate of the AMSE of the Hill estimator for $1/\theta$. The weighted MSE expression minimized is (in our notation)

$$\mathrm{MSE}_{\mathrm{opt}}(k) = \frac{1}{k}\sum_{i=1}^{k} w_{i,k}^{\mathrm{opt}} \left(\log \frac{X^*_{[i]}}{x_0} + \frac{1}{\hat{\theta}} \log\left(\frac{k+1-i}{k+1} \right) \right)^2,$$

for some sequence of weights $w_{i,k}^{\mathrm{opt}}$ which depends on ρ, a non-positive index that is assumed to characterize the slowly varying function and which must also be estimated. The index ρ is also estimated nonparametrically and thus estimation of the AMSE, and the optimal k, is an iterative procedure. Note that the factor $1/\hat{\theta}$ is the estimate of $1/\theta$ obtained by using the Hill estimator with k observations. Beirlant, Vynckier & Teugels (1996) rely on probabilistic deductions to establish optimal weights $w_{i,k}^{\mathrm{opt}}$. Essentially, we have

$$w_{i,k}^{\mathrm{opt}} = \delta_{1,k} w_{i,k}^{(1)} + \delta_{2,k} w_{i,k}^{(2)}, \tag{25}$$

where $\delta_{1,k}$ and $\delta_{2,k}$ are scaling constants that depend on k and ρ, and $w_{i,k}^{(1)}$ and $w_{i,k}^{(2)}$ are chosen weight functions. They used $w_{i,k}^{(1)} = 1$ and $w_{i,k}^{(2)} = (k+1-i)/(k+1)$. The values of $\delta_{1,k}$ and $\delta_{2,k}$

must be obtained numerically as even for simple choices of $w_{i,k}$, and assuming a fixed value of ρ, we cannot get an analytical expression. However, $\delta_{1,k}$ and $\delta_{2,k}$ will be functions of k, and we keep this notation in an attempt to compare $\mathrm{MSE}_{\mathrm{opt}}(k)$ with the $C(x_0)$ in (15). We thus have

$$\begin{aligned}\mathrm{MSE}_{\mathrm{opt}}(k) = & \frac{1}{k}\sum_{i=1}^{k}\left(\delta_{1,k}+\delta_{2,k}\left(\frac{k+1-i}{k+1}\right)\right)\left[\log\frac{X^*_{[i]}}{x_0}+\frac{1}{\hat\theta}\log\left(\frac{k+1-i}{k+1}\right)\right]^2 \\ = & \frac{\delta_{2,k}}{k(k+1)}\sum_{i=1}^{k}(k+1-i)\left(\log\frac{X^*_{[i]}}{x_0}+\frac{1}{\hat\theta}\log\left(\frac{k+1-i}{k+1}\right)\right)^2+ \\ & \frac{\delta_{1,k}}{k}\sum_{i=1}^{k}\left(\log\frac{X^*_{[i]}}{x_0}+\frac{1}{\hat\theta}\log\left(\frac{k+1-i}{k+1}\right)\right)^2,\end{aligned} \tag{26}$$

and the following may be noted about the first term in $\mathrm{MSE}_{\mathrm{opt}}(k)$: 1) the sum is scaled by $\delta_{2,k}/k(k+1)$, an implicit function of the estimated $\mathrm{AMSE}(1/\hat\theta)$, compared to $\hat\theta^2/k$ for $C(x_0)$; and 2) relative downweighting is a linear function in i in the Beirlant case while nonlinear in the C-criterion, the C-criterion putting more relative weight on points closer to x_0, see Figure 1.

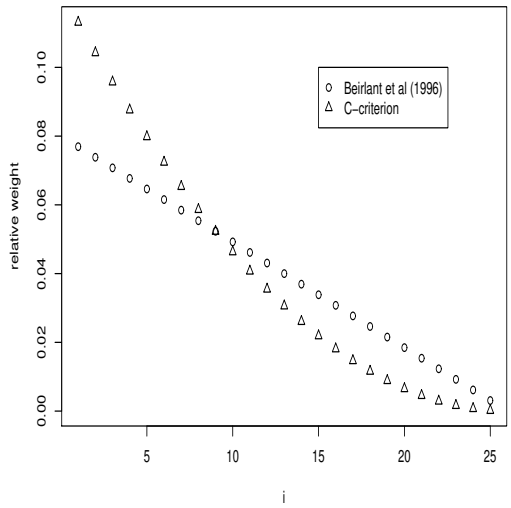


Figure 1: Relative weights assigned to ith term in first sum under C-criterion (15) and $\mathrm{MSE}_{\mathrm{opt}}(\mathrm{k})$ (26), respectively, when $k = 25$.

An expansion of the second term in $\mathrm{MSE}_{\mathrm{opt}}(k)$ leads to

$$\frac{\delta_{1,k}}{k}\frac{1}{\hat\theta^2}\sum_{i=1}^{k}\log\left(\frac{k+1-i}{k+1}\right)^2+\frac{\delta_{1,k}}{k}\sum_{i=1}^{k}\log\left(\frac{X^*_{[i]}}{x_0}\right)^2+\frac{2\delta_{1,k}}{k\hat\theta}\sum_{i=1}^{k}\log\left(\frac{X^*_{[i]}}{x_0}\right)\log\left(\frac{k+1-i}{k+1}\right). \tag{27}$$

The following may be noted about the first term in (27): 1) the sum is scaled by $\delta_{1,k}/k\hat\theta^2$ versus the simple $2/k^2$ in the second term in $C(x_0)$; and 2) the squared log terms all have weight 1 versus the variable weights $(1/k^2+\ldots+1/(k+1-i)^2)^{-1}$ in $C(x_0)$. Note that the variable weights in

$C(x_0)$ are such that weights decrease with increased size in the contribution from the squared log term. The third term in $C(x_0)$ is the constant -1. From (27) it is easily seen that $\mathrm{MSE}_{\mathrm{opt}}(k)$ is somewhat more complex with the presence of its remaining terms

$$\frac{\delta_{1,k}}{k}\sum_{i=1}^{k}\log\left(\frac{X^*_{[i]}}{x_0}\right)^2+\frac{2\delta_{1,k}}{k\hat{\theta}}\sum_{i=1}^{k}\log\left(\frac{X^*_{[i]}}{x_0}\right)\log\left(\frac{k+1-i}{k+1}\right).$$

Note that minimizing the latter is minimizing

$$\frac{\delta_{1,k}}{k}\left(\sum_{i=1}^{k}Y_i^2-2\sum_{i=1}^{k}Y_i\hat{Y}_i\right)$$

in our regression setting. The direct impact of the latter is unclear.

A theoretical comparison of our C-criterion to other recently suggested methods for selecting k is not so easily derived. We describe these other methods briefly here and include them in a simulation study in Section 4. Beirlant, Dierckx, Guillou & Stărică (2002) find an asymptotic representation of $k_{n,opt}$ and derive an estimator for $k_{n,opt}$ based on that representation. The automatic method for selection is somewhat *ad hoc* as it is really the median of $\hat{k}_{n,k_0}$ for $k_0 = 3,\ldots,n/2$ that is the recommended threshold. The choice is said to be *practical*, but is not justified mathematically. The estimator requires consistent estimators for ρ, $1/\theta$, and $g((n+1)/(k_0+1))$, where g is a rate function satisfying $g(x)\to 0$ as $x\to\infty$, characterizing the slowly-varying function l. Least squares estimators based on regression models with additive noise are used. Both Hall (1990) and Danielsson, de Haan, Peng & de Vries (2001) use subsample bootstrapping to estimate the MSE of the Hill estimator. Drees & Kaufmann (1998) present a sequential procedure, based on 'stopping times' for the sequence $H_{k,n}$ of Hill estimators that are asymptotically equivalent to a deterministic sequence, to select the optimal $k_{n,opt}$. Guillou & Hall (2001) propose to choose $H_{\hat{k},n}$ where $\hat{k}$ is the smallest value of $k\in[n^a,n^b]$ for which a scaled least squares estimate of $g((n+1)/(k+1))$ is larger than a critical value.

The implicit definition of the robust weights in our RC-criterion do not allow for an analytical comparison to be carried out with these previously mentioned complex estimators either and a simulation study is our only available tool for comparison. Results are shown in the following section.

4. SIMULATION STUDY

In this section we present the results of simulation studies showing the good properties, in terms of MSE, of the classical C- and robust RC- criteria when compared to competing methods for data from a distribution with Pareto tail behaviour. We also assess the performance of the C- and RC-criteria with and without data contamination.

When k is known, comparing our approach with others results in comparing the performance (bias and variance) of our robust estimators with the MLE in the presence of model contamination. In Dupuis & Victoria-Feser (2005), it is shown in a simulation study involving the MLE, the WMLE with probabilistic weighting and the WMLE with regression weighting, that all methods perform well at the model and that the MLE clearly fails under model contamination as small as 2.5%. The WMLE with probabilistic weighting has however a larger bias than the WMLE with regression weighting. Therefore, in what follows we consider only the WMLE with regression weighting.

4.1 Comparing the C- and RC-criteria to other criteria without model contamination

In order to evaluate the usefulness of the C- and RC-criteria, we carry out a simulation study and compare results with those of others. Beirlant *et al.* (2002) carried out an extensive simulation study

Table 2: RMSE for $\hat{\theta}^{-1}$ with the Burr distribution ($\theta = 1$).

		n		
ρ		500	1000	1500
-0.5	$C-$criterion	0.295	0.269	0.247
	$RC-$criterion	0.312	0.287	0.268
	Method 1	0.334	0.238	0.220
	Method 2	0.305	n/a	n/a
	Method 3	0.382	n/a	n/a
	Method 4	0.352	n/a	n/a
	Method 5	0.381	n/a	n/a
-1.0	C criterion	0.151	0.121	0.104
	$RC-$criterion	0.161	0.128	0.109
	Method 1	0.264	0.148	0.132
-1.5	C criterion	0.107	0.082	0.071
	$RC-$criterion	0.113	0.084	0.072
	Method 1	0.150	0.100	0.085

and we will refer to some of their results, along with those of Beirlant, Vynckier & Teugels (1996). More specifically, methods compared to the C- and RC-criteria are:

- Method 1 - Beirlant, Vynckier & Teugels (1996)
- Method 2 - Beirlant *et al.* (2002)
- Method 3 - Danielsson *et al.* (2001), based on Hall (1990).
- Method 4 - Drees & Kaufmann (1998)
- Method 5 - Guillou & Hall

The root mean squared error (RMSE) of $\widehat{\theta}^{-1}$ is reported. Note that the C-criterion should further outperform these asymptotic-based criteria in smaller sample sizes, but the smallest sample size previously considered and available for comparison is $n = 500$. We consider a Burr distribution parametrized as $F(x) = 1 - (1 + x^{-\rho})^{1/\rho}$ for some parameter $\rho < 0$. Beirlant, Vynckier & Teugels (1996) consider $\rho = -0.5, -1$, and -1.5 (all leading to $\theta = 1$) and sample sizes of $n = 500$, 1000, and 1500. Their simulation results are based on 200 replications. Results based on 100 replications were reported by Beirlant *et al.* (2002) for Methods 2-5 and a sample size of $n = 500$. All these results, along with those for the C-criterion and the RC-criterion with the residual based WMLE ($c = 2.5$) are listed in Table 2. Both the $C-$ and $RC-$ criteria do quite well, especially for $n = 500$ and/or $\rho = -1$ and -1.5. Other distributions are considered in Dupuis & Victoria-Feser (2003).

4.2 Comparing the C- to the RC-criteria with model contamination

In order to have examples with a relatively clear distinction between the Pareto upper-tail and the rest of the distribution, samples of size $n = 500$ from a triangular/Pareto mixture for the upper-tail were also considered. More precisely, we generated $\lfloor \alpha n \rfloor$ data (where $\lfloor x \rfloor$ denotes the integer part of x) from a Pareto distribution as in (2) and $n - \lfloor \alpha n \rfloor$ data from a triangular distribution with density $f(x) = 2(x - 1.5)$ for $1.5 \leq x \leq 2.5$. Parameter settings were $\alpha = 0.1$, $x_0 = 2.5$, and $\theta = 1$. The Pareto regression plot for this sample is *hockey stick* shaped and the line bends quite abruptly around $x_0 = 2.5$. Ideally, a k-selection criterion would properly capture the break. Boxplots of the estimates for k and θ, for different efficiency levels of the WMLE (i.e. different values of c) are shown in Figure 2. Results are as expected with both approaches selecting a threshold slightly

above x_0 and the robust approach leading to slightly more variable estimates of θ. Moreover, we see that there is very little loss of efficiency in the estimation of x_0 and θ when reducing the value of c.

We repeat the study, contaminating the 2% largest observations in each sample by multiplying them by 1000. The type of contamination should in principle mimic the type of situation one could encounter in practice, however a robust procedure should be stable across all types of contamination, even those to which we have not yet been exposed. Our point here is merely to show that, given a type of contamination, the classical approach fails while the robust approach remains stable. Considering other types of contamination would merely show that the C-criterion may be affected to a lesser degree and the performance of the RC-criterion would remain. Boxplots of the estimates for these contaminated samples are shown in Figure 3. The contamination destroys the MLE. Robust estimates of θ are quite good, showing only modest bias and no more variability than in the non-contaminated case. The increased robustness for smaller values of c is also clear.

All these results clearly demonstrate the necessity of our new weighting approach.

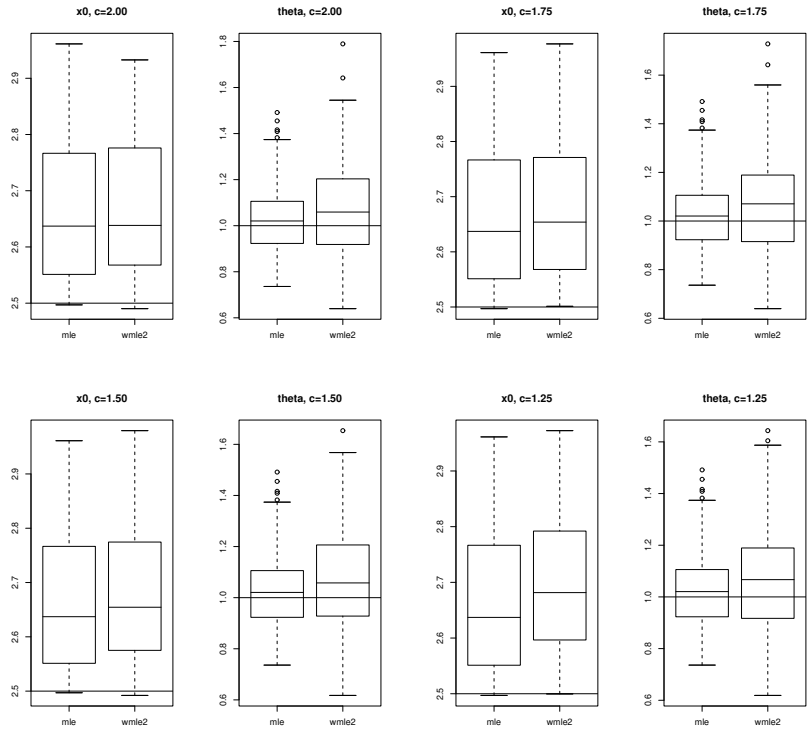


Figure 2: Estimates for x_0 and θ under triangular/Pareto mixture and no contamination. `wmle2` is WMLE with regression weighting. The value of the robustness constant c is shown for each panel.

5. FINANCE AND INCOME APPLICATIONS

In this section we examine the classical C-criterion and its robust counterpart RC-criterion for two published data sets.

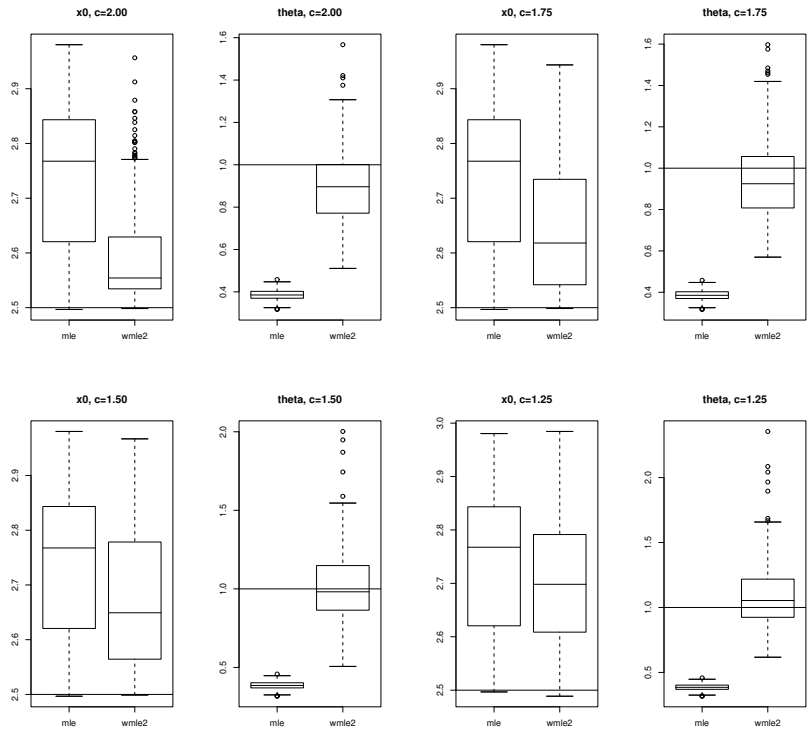
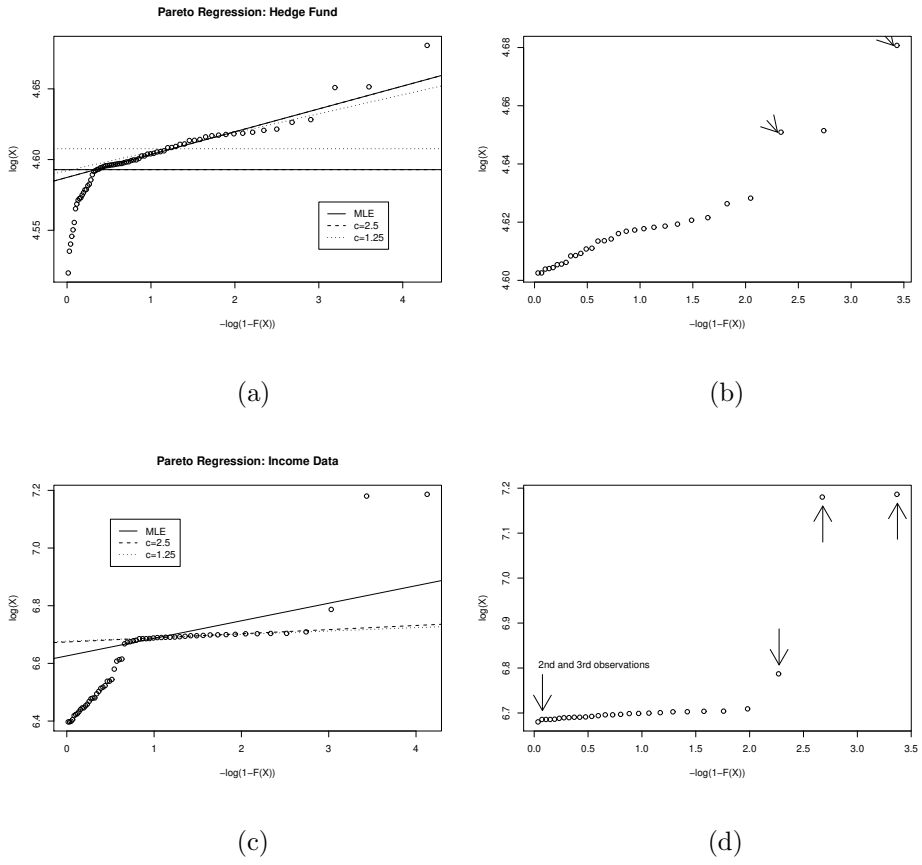


Figure 3: Estimates for x_0 and θ under triangular/Pareto mixture and 2% contamination in upper tail. `wmle2` is WMLE with regression weighting. The value of the robustness constant c is shown for each panel.

5.1 Finance data

In finance theory, one important issue is the ability to establish the value at risk (VaR) of an investment (an asset, a portfolio of assets, etc.). The latter can be defined as the level of loss (i.e. a quantile usually in the lower tail of the distribution of returns) on a portfolio which is expected to be equaled or exceeded with a given (usually small) probability (see e.g. Jorion, 1997). Since the returns in the tails of the distribution are sparse, it is therefore important to be able to model them. Empirical studies on the tails of log-returns have indicated that a Pareto-type model is usually suitable. As an example, we consider here a series of log-returns in 100% on alternative investments on a monthly basis between January 1997 and December 2002 (i.e. $n = 72$ observations). See Credit Suisse First Boston (CSFB) / Tremont hedge fund index at `www.hedgeindex.com` and Perret-Gentil & Victoria-Feser (2003) for a more detailed description of the data. The data do not show significant autocorrelations (results not shown here). We actually take 100 minus the log-return for the evaluation of the downside risk. The C-criterion yields $x_0 = 98.77$, corresponding to $k = 51$, and $\theta = 61.90$. The RC-criterion yields $x_0 = 98.77$ ($k = 51$) and $\theta = 61.92$, and $x_0 = 100.24$ ($k = 22$) and $\theta = 74.45$ for $c = 2.5$ and $c = 1.25$, respectively. All three fits are shown in Figure 4 (a) and downweighted observations ($c = 1.25$) are identified in Figure 4 (b). It is interesting to note that weighting scheme (24) is indeed performing as intended, downweighting observations away from the Pareto regression line, and not necessarily only the largest observations. If one was to erroneously remove the three largest observations, thinking them to be outliers, and proceed,



(a) (b)

(c) (d)

Figure 4: (a) Pareto regression plot for Hedge fund data. Fitted regression line based on classical $C-$ and robust $RC-$ criteria added. Horizontal lines indicate the corresponding chosen thresholds. Solid and dashed lines are superimposed; (b) Pareto regression plot of Hedge data above robustly chosen threshold. Downweighted observations following $RC-$ criterion, $c = 1.25$, are identified; (c) Pareto regression plot for income data. Fitted regression line based on classical $C-$ and robust $RC-$ criteria added. Only incomes above 600 are shown for clarity; (d) Pareto regression plot of income data above robustly chosen threshold. Downweighted observations following $RC-$ criterion, $c = 1.25$, are identified.

Table 3: Estimated VaR (ES) for the Finance data

	probabilities			
	0.05	0.01	0.005	0.001
C-criterion	103.09 (1297)	105.81 (6756)	107.00 (13579)	109.82 (68163)
idem with 3 largest removed	102.13 (1214)	104.13 (6325)	105.00 (12713)	107.06 (63818)
RC-criterion, $c = 2.5$	103.09 (1297)	105.81 (6756)	107.00 (13579)	109.82 (68163)
RC-criterion, $c = 1.25$	102.71 (590)	104.95 (3074)	105.93 (6179)	108.25 (31015)

the $C-$criterion would yield $x_0 = 98.96$ ($k = 48$) and $\theta = 82.96$ (to be compared to $\theta = 74.45$ provided by the RC-criterion with $c = 1.25$ on the whole sample).

To have a better sense of the differences implied by the classical and robust approaches, we compute the VaR given the different estimates of the Pareto tail index. The latter appear in Table 3 along with another financial risk measure, the expected shortfall (ES), defined as the mean return in the lower tail of the distribution or in the upper tail after proper relabeling, i.e. $ES(F;q) = \frac{1}{q}\int_{Q(F;q)} xdF(x)$ where $F(x)$ is given in (1). The expected shortfall is then

$$ES(F;q) = \frac{\alpha\theta x_0^\theta}{q}\int_{Q(F;q)} x^{-\theta}dx = \alpha\frac{\theta}{\theta-1}\frac{x_0}{q}(1-q)^{\frac{\theta-1}{\theta}}$$

with $\alpha = 1 - G(x_0)$ and which is estimated using estimates of θ and x_0.

The C- and RC- criteria all lead to very similar VaR when all observations are considered but the values are quite different when the three largest observations are removed. The story is different with the ES in that the C-criterion without the three largest observations provides estimated ES that lie between the two RC-criterion estimates, while the RC-criterion provides ES that are nearer to the ones provided by means of the C-criterion. The conclusion here is that the pure *ad hoc* procedure of removing suspect observations can lead to different results from a proper robust approach, especially in cases where, as it is here, the largest observations are not all "extreme" at least as measured by means of standardized residuals.

5.2 Income Data

The data are incomes ($n = 7469$) in the UK in 1981 (see Department of Social Security (1992) and Cowell & Victoria-Feser (1996) for a more detailed description). With income data, the determination of the value x_0 is an important issue since traditionally data in the upper tail are sparse and therefore are fitted with Pareto-type distributions for the estimation of inequality measures. The value of x_0 can also be used as a cutting point for a semi-parametric approach to stochastic dominance comparisons (see Cowell & Victoria-Feser, 2006). The corresponding fitted Pareto-regression lines for the C- and $RC-$ criteria are given in Figure 4 (c). The classical and high efficiency robust curves are different. The selected thresholds are similar: the $C-$criterion leading to $x_0 = 783.9$ ($k = 32$) and high efficiency $RC-$criterion leading to $x_0 = 802.9$ ($k = 22$), however estimates of θ are quite different, 16.4 and 69.7, respectively. Using the very robust (less efficient) $RC-$criterion, we find $x_0 = 803.3$ ($k = 22$) and $\theta = 85.5$. Figure 4 (d) shows observations above the robustly selected threshold $x_0 = 803.3$ and arrows indicate the downweighted observations. Variance in the lower-tail of the Pareto regression plot is quite small and an observation does not have to fall too far out of line to be considered outlying. The opposite is true in the upper-tail. The three largest observations here were considerably off the mark and were flagged.

APPENDIX

Proof of Lemma 3. Since X is a Pareto random variable, $Y_i = log\,(X_i/x_0)\,,\ i = 1,\ldots,k$ are exponential random variables with mean $1/\theta$. Let $Y_{[i]}$ be the corresponding order statistics. From Bickel & Doksum (2001) (problem 14, p. 528), we see that $Z_i = \theta(k-i+1)\left(Y_{[i]} - Y_{[i-1]}\right)$ (with $Y_{[0]} = 0$) are i.i.d. exponentially distributed random variables with mean 1. We can write

$$Y_{[i]} = \sum_{j=1}^{i} \frac{1}{\theta(k-i+j)} Z_{i-j+1},$$

so that

$$\mathrm{var}\left(Y_{[i]}\right) = \sum_{j=1}^{i} \frac{1}{\theta^2(k-i+j)^2}.$$

Proof of Lemma 4. Substituting the Pareto model (2), the partial derivative w.r.t. θ of the log of its corresponding density (3), and the weight function (22) into the numerator of (21), we find that this numerator becomes

$$\frac{1}{p_1}\int_{x_0}^{x_0(1-p_1)^{-1/\theta}} [1-(x/x_0)^{-\theta}]\left[\frac{1}{\theta}+\log\frac{x_0}{x}\right]\theta x^{-(\theta+1)}x_0^{\theta}dx + \int_{x_0(1-p_1)^{-1/\theta}}^{x_0 p_2^{-1/\theta}} \left[\frac{1}{\theta}+\log\frac{x_0}{x}\right]\theta x^{-(\theta+1)}x_0^{\theta}dx \ + \frac{1}{p_2}\int_{x_0 p_2^{-1/\theta}}^{\infty}(x/x_0)^{-\theta}\left[\frac{1}{\theta}+\log\frac{x_0}{x}\right]\theta x^{-(\theta+1)}x_0^{\theta}dx\ .$$

Standard calculus and algebra yields

$$\frac{-1}{4p_1\theta}\left[2(1-p_1)^2\log(1-p_1)+p_1(1-p_1)+p_1(1-p_2)+2p_1p_2\log p_2\right].$$

The denominator of (21) can be evaluated similarly, and minus the ratio yields (23).

Proof of Proposition 1. Upon substituting the values for Y_i and $\hat{Y}_i$ in (11), and provided that all $\hat{w}_i = 1$, we obtain

$$C_R = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{\sigma_i^2}\left(\log\left[\frac{Q(F_{(n)};i/(k+1))}{x_0}\right]+\frac{1}{\hat{\theta}}\log\left[\frac{k+1-i}{k+1}\right]\right)^2+\frac{2}{k}\sum_{i=1}^{k}\frac{1}{\sigma_i^2}\mathrm{cov}\left(Y_i,\hat{Y}_i\right)-1\ . \quad (28)$$

To compute $\mathrm{cov}(Y_i,\hat{Y}_i)$ we use the approach based on the influence function (IF) which is equivalent to the delta method or the infinitesimal jackknife; see Efron (1982). The IF is defined for a statistic $\hat{\alpha}$ as

$$IF(z;\hat{\alpha},F_\theta) = \left.\frac{\partial}{\partial\varepsilon}\hat{\alpha}\left(F_\varepsilon\right)\right|_{\varepsilon=0},$$

where $F_\varepsilon = (1-\varepsilon)F_\theta + \varepsilon\Delta_z$ and Δ_z is the probability measure that puts mass 1 on the point z. The IF can be used to compute variances and covariances between two statistics $\hat{\alpha}$ and $\hat{\beta}$. We get, up to $O(1/k)$,

$$\mathrm{cov}\left(\hat{\alpha},\hat{\beta}\right) \simeq \int IF(z;\hat{\alpha},F_\theta)IF(z;\hat{\beta},F_\theta)dF_\theta(z)\ ;$$

see e.g. Hampel *et al.* (1986).
The IF of Y_i is

$$\begin{aligned} IF(z;Y_i,F_\theta) &= \tfrac{\partial}{\partial\varepsilon}\left[\log\left(\tfrac{Q(F_\varepsilon;i/(k+1))}{x_0}\right)\right]_{\varepsilon=0} \\ &= \frac{1}{Q(F_\theta;i/(k+1))}\left[\frac{\partial}{\partial\varepsilon}Q(F_\varepsilon;i/(k+1))\right]_{\varepsilon=0} \\ &= \frac{i/(k+1)-I(Q(F_\theta;i/(k+1))\geq z)}{Q(F_\theta;i/(k+1))f(Q(F_\theta;i/(k+1)),\theta)}, \end{aligned}$$

where we have used the result $IF(z;Q(\cdot,q),F_\theta) = (q - I(Q(F_\theta;q) \geq z))/f(Q(F_\theta;q),\theta)$; see e.g. Staudte and Sheather (1990). For the IF of $\hat{Y}_i$ we have

$$\begin{aligned} IF(z;\hat{Y}_i,F_\theta) = & \frac{\partial}{\partial\varepsilon}\left[-\frac{1}{\hat{\theta}(F_\varepsilon)}\log\left(\frac{k+1-i}{k+1}\right)\right]_{\varepsilon=0} \\ = & \log\left(\frac{k+1-i}{k+1}\right)\frac{1}{\theta^2}IF(z;\hat{\theta},F_\theta), \end{aligned}$$

for whose evaluation we will also need $IF(z;\hat{\theta},F_\theta)$. We know that $IF(z;\hat{\theta},F_\theta) = M(\theta)^{-1}s(z,\theta)$, where

$$\begin{aligned} s(z,\theta) = & \frac{\partial}{\partial\theta}\log\left(f(z;\theta)\right) \\ = & \frac{1}{\theta} + \log x_0 - \log z, \end{aligned}$$

and

$$\begin{aligned} M(\theta) = & \int_{x_0} s(z,\theta)^2 dF_\theta(z) \\ = & \int_{x_0}^{\infty}\left(\tfrac{1}{\theta} + \log x_0 - \log z\right)^2 \theta x_0^\theta z^{-\theta-1}dz \\ = & \left(\frac{1}{\theta} + \log x_0\right)^2 \theta x_0^\theta \int_{x_0}^{\infty} z^{-\theta-1}dz \\ & -2\left(\frac{1}{\theta} + \log x_0\right)\theta x_0^\theta \int_{x_0}^{\infty}\log(z)\, z^{-\theta-1}dz \\ & +\theta x_0^\theta \int_{x_0}^{\infty}(\log z)^2 z^{-\theta-1}dz. \end{aligned}$$

The integrals are straightforward and subsequent simplifications yield $M(\theta) = 1/\theta^2$, so that $IF(z;\hat{\theta},F_\theta) = \theta^2(1/\theta + \log x_0 - \log z)$ and $IF(z;\hat{Y}_i,F_\theta) = \log\left[(k+1-i)/(k+1)\right](1/\theta + \log x_0 - \log z)$. Thus, we now have

$$\begin{aligned} \operatorname{cov}\left(Y_i,\hat{Y}_i\right) = & \tfrac{1}{k}\int_{x_0}^{\infty}\left(\frac{i/(k+1) - I(Q(F_\theta;i/(k+1)) \geq z)}{Q(F_\theta;i/(k+1))f(Q(F_\theta;i/(k+1)),\theta)}\right) \\ & \log\left(\tfrac{k+1-i}{k+1}\right)\left(\tfrac{1}{\theta} + \log x_0 - \log z\right)\theta x_0^\theta z^{-\theta-1}dz \\ = & \frac{\log\left(\frac{k+1-i}{k+1}\right)}{k\theta\left(\frac{k+1-i}{k+1}\right)}\left[i/(k+1)\frac{1}{\theta} + i/(k+1)\log x_0\right. \\ & -i/(k+1)\int_{x_0}^{\infty}\log(z)\,\theta x_0^\theta z^{-\theta-1}dz - \frac{1}{\theta}\int_{x_0}^{Q(F_\theta;i/(k+1))}\theta x_0^\theta z^{-\theta-1}dz \\ & \left.-\log x_0\int_{x_0}^{Q(F_\theta;i/(k+1))}\theta x_0^\theta z^{-\theta-1}dz + \int_{x_0}^{Q(F_\theta;i/(k+1))}\log(z)\,\theta x_0^\theta z^{-\theta-1}dz\right]. \end{aligned}$$

But it can easily be shown that

$$\int_{x_0}^{Q(F_\theta;i/(k+1))}\log(z)\,\theta x_0^\theta z^{-\theta-1}dz = \frac{i}{k+1}\log x_0 + \frac{1}{\theta}\left(1 - \frac{i}{k+1}\right)\log\left(1 - \frac{i}{k+1}\right) + \frac{1}{\theta}\frac{i}{k+1},$$

and straightforward, although tedious, simplifications lead to

$$\operatorname{cov}\left(Y_i,\hat{Y}_i\right) = \frac{1}{k}\frac{1}{\theta^2}\log\left(\frac{k+1-i}{k+1}\right)^2. \tag{29}$$

Substituting (12) and (29) into (28) and replacing θ by the Hill estimator in (6), one obtains the estimated prediction error (15).

We note that an approach based on the IF can also be used to get an $O(1/k)$ approximation to the exact σ_i^2 in (12). Since we obtained essentially the same performance of our criteria with

the $O(1/k)$ form (rather than the exact form) of σ_i^2 (results not shown), we are confident that our $O(1/k)$ approximation to $\text{cov}(Y_i, \hat{Y}_i)$, for which the exact form cannot be obtained, is good.

ACKNOWLEDGEMENTS

The first author acknowledges ongoing grant support from the Natural Sciences and Engineering Research Council of Canada. The second author acknowledges the financial support of the Swiss National Fund and the Swiss NCCR FinRisk. Both authors wish to thank the Editor, Associate Editor, and two referees for valuable comments that improved the presentation.

REFERENCES

J. Beirlant, G. Dierckx, Y. Goegebeur & G. Matthys (1999). Tail index estimation and an exponential regression model. *Extremes*, 2, 177–200.

J. Beirlant, G. Dierckx, A. Guillou & C. Stărică (2002). On exponential representations of log-spacings of extreme order statistics. *Extremes*, 5, 157–180.

J. Beirlant, P. Vynckier & J. L. Teugels (1996). Tail index estimation, Pareto quantile plots, and regression diagnostics. *Journal of the American Statistical Association*, 91, 1659–1667.

P. J. Bickel & K. A. Doksum (2001). *Mathematical Statistics: Basic Ideas and Selected topics*, Volume I, 2nd Edition, Prentice-Hall.

V. Brazauskas (2003). Influence functions of empirical nonparametric estimators of net insurance premiums. *Insurance: Mathematics and Economics*, 32, 115–133.

F. A. Cowell & M.-P. Victoria-Feser (1996). Robustness properties of inequality measures. *Econometrica*, 64, 77–101.

F. A. Cowell & M.-P. Victoria-Feser (2006). Robust Stochastic Dominance: A Semi-Parametric Approach. *Journal of Economic Inequality*, to appear.

M. M. Dacorogna, U. A. Müller, O. V. Pictet & C. G. de Vries (2001). Extremal Forex Returns in Extremely Large Data Sets. *Extremes*, 4, 105–127.

J. Danielsson, L. de Haan, L. Peng & C. de Vries (2001). Using a bootstrap method to choose sample fraction in tail index estimation. *Journal of Multivariate Analysis*, 76, 226–248.

J. Danielsson & C. de Vries (1997). Tail index and quantile estimation with very high frequency data. *Journal of Empirical Finance*, 4, 241–257.

Department of Social Security (1992). *Households below Average Income, 1979-1988/89*, London: HMSO.

H. Drees & E. Kaufmann (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications*, 75, 149–172.

D. J. Dupuis (1999). Exceedances over High Thresholds: A Guide to Threshold Selection. *Extremes*, 1, 251–261.

D. J. Dupuis & C. Field (2004). Large Wind Speeds: Modeling and Outlier Detection. *Journal of Agricultural, Biological and Environmental Sciences*, 9, 105–121.

D. J. Dupuis & S. Morgenthaler (2002). Robust weighted likelihood estimators with an application to bivariate extreme value problems. *Canadian Journal of Statistics*, 30, 17–36.

D. J. Dupuis & M.-P. Victoria-Feser (2003). A Prediction Error Criterion for Choosing the Lower Quantile in Pareto Index Estimation. *Cahiers de Recherche HEC no 2003.10*, University of Geneva, CH-1211 Geneva.

D. J. Dupuis & M.-P. Victoria-Feser (2005). A Robust Prediction Error Criterion for Pareto Modeling of Upper Tails. *Cahiers du GERAD #G-2005-29*, March 2005, 30 pages.

B. Efron (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial an Applied Mathematics, 38, Philadelphia.

P. Embrechts, C. Klüppelberg & T. Mikosch (1997). *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin.

C. Field & B. Smith (1994). Robust estimation - a weighted maximum likelihood approach. *International Statistical Review*, 62, 405–424.

A. Guillou & P. Hall (2001). A diagnostic for selecting the threshold in extreme-value analysis. *Journal of the Royal Statistical Society, Series B*, 63, 293–305.

P. Hall (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate Analysis*, 32, 177–203.

P. Hall & A. H. Welsh (1985). Adaptive Estimates of Parameters of Regular Variation. *The Annals of Statistics*, 13, 331–341.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw & W. A. Stahel (1986). *Robust Statistics: The Approach based on Influence Functions*, John Wiley, New York.

B. M. Hill (1975). A simple approach to inference about the tail of a distribution. *The Annals of Statistics*, 3, 1163–1174.

M. C. A. B. Hols & C. G. de Vries (1991). The limiting distribution of extremal exchange rate returns. *Journal of Applied Econometrics*, 6, 287–302.

P. J. Huber (1981). *Robust Statistics*, John Wiley, New York.

D. W. Jansen & C. G. de Vries (1991). On the frequency of large stock returns: Putting booms and busts into perspective. *Review of Economics and Statistics*, 73, 18–24.

P. Jorion (1997). *Value at Risk*, McGraw Hill, New York.

K. G. Koedijk, M. M. A. Schafgans & C. G. de Vries (1990). The tail index of exchange rate returns. *Journal of International Economics*, 29, 93–108.

E. Marceau & J. Rioux (2001). On robustness in risk theory. *Insurance: Mathematics and Economics*, 29, 167–185.

V. Pareto (1896). Ecrits sur la courbe de la répartition de la richesse. *Oeuvres complètes de Vilfredo Pareto*, Giovanni Busino, Librairie Droz, Genève, 1965.

L. Peng & A. H. Welsh (2001). Robust Estimation of the Generalized Pareto Distribution. *Extremes*, 4, 53–65.

C. Perret-Gentil & M.-P. Victoria-Feser (2003). Robust mean-variance portfolio selection. *Cahiers du Département d'Econométrie no 2003.2*, University of Geneva, CH-1211 Geneva.

P. C. B. Phillips, J. W. McFarland & P. C. McMahon (1996). Robust tests of forward exchange market efficiency with empirical evidence from the 1920s. *Journal of Applied Econometrics*, 11, 1–22.

E. Ronchetti & R. Staudte (1994). A robust version of Mallows' C_p. *Journal of the American Statistical Association*, 94, 550–559.

R. Staudte & S. J. Sheather (1990). *Robust Estimation and Testing*. John Wiley and Sons, New York.

B. Vandewalle, J. Beirlant & M. Hubert (2004). A Robust Estimator of the Tail Index based on Exponential Regression Model. *Theory and Applications of Recent Robust Methods*, Statistics for Industry and Technology Series, Birkhauser, Basel, Editors M. Hubert, G. Pison, A. Struyf & S. Van Aelst.

M.-P. Victoria-Feser & E. Ronchetti (1994). Robust methods for personal income distribution models. *Canadian Journal of Statistics*, 22, 247–258.

Received 8 September 2005
Accepted 23 July 2006

Debbie J. DUPUIS: `debbie.dupuis@hec.ca`
Department of Management Sciences
HEC Montréal
Montréal (Québec), Canada, H3T 2A7

Maria-Pia VICTORIA-FESER: `Maria-Pia.VictoriaFeser@hec.unige.ch`
Faculty of Economics and Social Sciences (HEC)
University of Geneva
Switzerland, CH-1211 Geneva 4