J Econ Inequal (2007) 5:21–37
DOI 10.1007/s10888-006-9022-z

Robust stochastic dominance: A semi-parametric approach

Frank A. Cowell · Maria-Pia Victoria-Feser

Received: 23 November 2004 / Accepted: 8 February 2006 /
Published online: 14 September 2006

Abstract Lorenz curves and second-order dominance criteria, the fundamental tools for stochastic dominance, are known to be sensitive to data contamination in the tails of the distribution. We propose two ways of dealing with the problem: (1) Estimate Lorenz curves using parametric models and (2) combine empirical estimation with a parametric (robust) estimation of the upper tail of the distribution using the Pareto model. Approach (2) is preferred because of its flexibility. Using simulations we show the dramatic effect of a few contaminated data on the Lorenz ranking and the performance of the robust semi-parametric approach (2). Since estimation is only a first step for statistical inference and since semi-parametric models are not straightforward to handle, we also derive asymptotic covariance matrices for our semi-parametric estimators.

Key words Lorenz curve · M-estimators · Pareto model · welfare dominance

1. Introduction

The Lorenz curve is central to the analysis of income distributions, embodying fundamental intuition about inequality comparisons. Ranking theorems based on Lorenz dominance and the associated concept of stochastic dominance are fundamental to the theoretical welfare economics of distributions. But formal welfare propositions can only be satisfactorily invoked for empirical constructs if sample data can be taken as a reasonable representation of the underlying income distributions under consideration. In practice income-distribution data may be contaminated by recording errors, measurement errors and the like and, if the data cannot be purged of these, welfare conclusions drawn from the data can be seriously

F. A. Cowell
London School of Economics,
Houghton Street, London, WC2A 2AE, UK

M.-P. Victoria-Feser (✉)
HEC, Université de Geneve,
40, bd du Pont d'Arve, Geneve 4, CH-1211, Switzerland
e-mail: maria-pia.victoriafeser@hec.unige.ch

misleading. The purpose of this paper is to provide a rigorous method for handling some of these potential problems, one that accords well with pragmatic procedures that are sometimes adopted by applied researchers in this field.

The point of departure is recent research which has shown that Lorenz and stochastic dominance results are non-robust [5]. This means that small amounts of data contamination in the wrong place can reverse unambiguous ranking orders: The "wrong place" usually means in the upper tail of the distribution. This is of particular interest in view of a burgeoning recent literature that has focused on empirical issues concerning the upper tail of both income distributions and wealth distributions [1, 19, 22, 23]. So it is important to have an approach that enables one to control for the distortionary effect of upper-tail contamination in a systematic fashion. We need a robust method of estimating Lorenz curves and implementing stochastic dominance criteria.

There are two main ways of avoiding misleading conclusions due to non-robust ranking tools in the presence of contaminated data. One is based on statistics that automatically remove from the sample any data that are potentially troublesome. The other relies on the specification of parametric models for the distribution of the data and uses robust estimators of the parameters. The first approach, based on the concept of trimmed Lorenz curves, raises issues which go beyond the scope of this paper and are handled in Cowell and Victoria-Feser [7]. Here we focus on parametric approaches, which are of particular interest because of their *ad hoc* use in practical treatment of problems associated with the upper tails of distributions. For example a Pareto tail is sometimes fitted to data in cases where data are sparse in order to provide better estimates of upper tail probabilities or higher quantiles.[1] This approach is related to problems in the field of extreme value distributions. Several models for extreme value distributions (which often include the Pareto distribution as a special case) as well as several estimators for the parameters of these models have been proposed. A now classic reference is Embrechts et al. [13]; Dupuis and Field [12] have proposed a methodology for the robust estimation of the parameters of a generalized extreme value distribution. They actually concentrate only on the upper tail of the distribution. However, our approach is quite different in that we not only consider the whole distribution (extreme and not extreme) and that this (robust) estimation part is just a first step in estimating Lorenz curves in a robust fashion. This is necessary for drawing conclusion on stochastic dominance results.

The paper is organized as follows. While section 2 sets the background, in section 3 we discuss two ways of implementing a parametric approach to the estimation of Lorenz curves. In section 4 we analyze both simulated data and a real example, and section 5 concludes. In the Appendix, we provide the necessary tools for inference on robust Lorenz curves.

2. The background

Let $\mathfrak{F}$ be the set of all univariate probability distributions and X be a random variable with probability distribution $F \in \mathfrak{F}$ and support $\mathfrak{X} \subseteq \mathbb{R}$. F can be thought of

[1] An important recent example of this is provided in Atkinson [1].

as a parametric model F_θ. We shall write statistics of any distribution $F \in \mathfrak{F}$ as a functional $T(F)$; in particular we write the mean as $\mu(F) := \int x dF(x)$.

A key distributional concept derived from F is given by

DEFINITION 1. *The q^{th} cumulative functional is the functional* C: $\mathfrak{F} \times [0,1] \mapsto \mathfrak{X}$: *such that:*

$$C(F;q) := \int_{\underline{x}}^{Q(F;q)} x dF(x) = c_q. \tag{1}$$

where $\underline{x} := \inf\ \mathfrak{X}$ *and*

$$Q(F;q) = \inf\ \{x | F(x) \geq q\} = x_q \tag{2}$$

is the quantile functional.

The importance of this concept is considerable in the practical analysis of income distributions: For a given $F \in \mathfrak{F}$, the graph of $C(F,q)$ against q describes the *generalized Lorenz curve* (GLC); normalizing by the mean functional $\mu(F) = C(F,1)$ one has the *Relative Lorenz curve* (RLC) [20]:

$$L(F;q) := \frac{C(F;q)}{\mu(F)} \tag{3}$$

The GLC and RLC are fundamental to a number of theorems drawing welfare-conclusions from income-distribution data and other types of data.

Cumulative functionals can obviously be estimated empirically by replacing F in (1) by the empirical distribution $F^{(n)}$. However, this can lead to misleading conclusions when it comes to comparing distributions in terms of their cumulative functionals when there is data contamination [5].

In order to present an alternative robust approach we will make use of the *influence function* (*IF*).[2] The primary usage of the *IF* is to characterize the sensitivity of a statistic to point contamination in the data [16] but can also be used to derive asymptotic results such as asymptotic covariance matrices of for example cumulative functionals [6, 7]. Let Δ_z be a point mass distribution giving probability 1 to an arbitrary point $z \in \mathfrak{X}$ and define the mixture distribution

$$F_\varepsilon^{(z)} = (1-\varepsilon)F + \varepsilon\Delta_z \tag{4}$$

$F_\varepsilon^{(z)}$ defines a distribution which generates with a large probability $(1-\varepsilon)$ data from the true model F and with a small probability ε arbitrary data z. The *IF* of a statistic $T(F)$ is defined as

$$IF(z;T,F) = \lim_{\varepsilon \downarrow 0} \frac{T(F_\varepsilon^{(z)}) - T(F)}{\varepsilon} \tag{5}$$

which becomes $\frac{\partial}{\partial \varepsilon} T(F_\varepsilon^{(z)})|_{\varepsilon=0}$ if T is differentiable. If the *IF* of a statistic T is unbounded or can take large values, then T is said to be not robust in the infin-

[2] Also called the influence curve and first introduced by Hampel [14, 15].

itesimal sense: an infinitesimal amount of contaminated data at z can change drastically the value of T.

Furthermore, the *IF* can be used as a fundamental tool to compute the asymptotic covariance matrix of T: Under very mild conditions on T one has that $\sqrt{n}\big(T(F^{(n)}) - T(F)\big)$ is asymptotically normal with asymptotic covariance matrix

$$\mathrm{cov}(\sqrt{n}T(F^{(n)}) = \int_{\mathfrak{X}} IF(z;T,F)IF'(z;T,F)dF(z) \tag{6}$$

(see [16] and references given in [6]). This result will be used when computing the asymptotic covariance matrix of semi-parametric functionals – see the Appendix.

3. Robust estimation of Lorenz curves

Before considering the robust fitting of the upper tail of the distribution defining a semi-parametric approach, we will briefly consider a full parametric approach and point out at its limitations.

3.1. A full parametric approach

A parametric approach to modelling the Lorenz curve requires the specification of a functional form for modelling the data. One then estimates the parameters of the model robustly and uses the estimated distributions to compute the (estimated) Lorenz curves. To be more precise, suppose we choose F_θ as model for the data and estimate θ robustly by $\widehat{\theta}$, then robust estimates of the GLC and the RLC are, respectively, given by (1) and (3) in which F is replaced by $F_{\widehat{\theta}}$. The *IF* of the estimators of the Lorenz curves will then depend on the *IF* of the parameter's estimator. Indeed, the Lorenz curves depend on the data only through the estimator $\widehat{\theta}$. If we write the latter as a functional of the contaminated distribution given in (4), i.e., $\widehat{\theta}(F_\varepsilon^{(z)})$, then we have

$$IF(z;C,F_\theta) = \frac{\partial}{\partial\theta} C(F_\theta;q)\cdot IF(z;\widehat{\theta},F_\theta). \tag{7}$$

Note that $\frac{\partial}{\partial\theta} C(F_\theta;q)$ does not depend on z, so that only if the estimator is robust, or in other words if its *IF* is bounded, the Lorenz curve estimated through a parametric model is also robust. Optimal bounded-influence estimators have been developed in the statistical literature for general parametric models [16] and used in particular for income distribution [25, 26]. Other types of robust estimators, for example ones based on robust moment estimators, could also be used.

However, in the present context, a full parametric approach is inappropriate because it forces the data into the mould of a functional form that may not be suitable for comparisons. For example, if one supposes that the data are lognormally distributed, then a "parametric Lorenz" comparison of two distributions based on the lognormal will always yield a strict dominance order! The parametric approach is therefore only appropriate provided that the postulated model is capable of yielding Lorenz curves that can cross: this may require specification of a complicated functional form that is difficult to estimate and to interpret.

3.2. A semi-parametric approach

In light of the above considerations, we suggest using a semi-parametric approach. If the range of X is bounded below – 0 is a typical value – the problems with contaminated data occur in the upper tail of the distribution [5]. A case can therefore be made for using parametric modelling only in the upper tail and estimating the parameter of the upper-tail model robustly. The rest of the distribution is estimated using the empirical distribution function. If no restriction is imposed on the range of the random variable of interest, then the results below can easily be extended accordingly.

Although the approach proposed here is suitable for any parametric model for the upper tail of the distribution, a model that is of special relevance empirically is the Pareto distribution given by

$$F_{\theta,x_0}(x) = 1 - \left[\frac{x}{x_0}\right]^{-\theta}, x > x_0 \tag{8}$$

with density $f(x;\theta) = \theta x^{-(\theta+1)} x_0^{\theta}$. The parameter of interest is θ.[3] A semi-parametric approach will combine a non-parametric RLC for say the $(1-\alpha)\%$ lower incomes and a parametric RLC based on the Pareto distribution for the $\alpha\%$ upper incomes. Therefore we suppose that x_0 is determined by the $1-\alpha$ quantile $Q(F;1-\alpha)$ defined in (2). The full semi-parametric distribution $\widetilde{F}$ of the income variable X is then

$$\widetilde{F}(x) = \begin{cases} F(x) & x \le Q(F;1-\alpha) \\ 1-\alpha\left(\frac{x}{Q(F;1-\alpha)}\right)^{-\theta} & x > Q(F;1-\alpha) \end{cases}. \tag{9}$$

For $x > Q(F;1-\alpha)$, the density $\widetilde{f}$ is

$$\widetilde{f}(x;\theta) = \alpha\theta Q(F;1-\alpha)^{\theta} x^{-\theta-1}.$$

In particular

$$\widetilde{f}(x_{1-\alpha};\theta) = \frac{\alpha\theta}{x_{1-\alpha}}. \tag{10}$$

To estimate the Pareto model for the upper tail of the distribution, one can use the maximum likelihood estimator (MLE). Unfortunately, the MLE for the Pareto model is known to be very sensitive to data contamination [25]. Here we propose the use a bounded *IF M*-estimator [17] with minimal asymptotic covariance matrix known as *optimal B-robust estimators* (OBRE). The expression of *M*-estimators is similar to that of the MLE. Given a sample $\{x_i, i = 1,\ldots n\}$ and a bound c on the *IF*, they are defined implicitly by the solution $\widehat{\theta}(\widetilde{F})$ in

$$\int_{Q(F;1-\alpha)}^{\infty} \psi(x;\widehat{\theta}(\widetilde{F}), Q(F;1-\alpha)) d\widetilde{F}(x) = 0.$$

When ψ is the score function $s(x;\theta, Q(F;1-\alpha)) = \frac{1}{\theta} - \log(x) + \log(Q(F;1-\alpha))$ we get the MLE. We get the OBRE when

$$\psi(x;\theta) = [s(x;\theta) - a(\theta)] W_c(x;\theta)$$

[3] θ is assumed to be greater than 2 for the variance to exist.

with

$$W_c(x;\theta) = \min\left\{1; \frac{c}{\|A(\theta)[s(x;\theta) - a(\theta)]\|}\right\} \tag{11}$$

where $\|\cdot\|$ denotes the Euclidean norm, and the matrix $A(\theta)$ and vector $a(\theta)$ are defined implicitly by

$$E\left[\psi(x;\theta)\psi'(x;\theta)\right] = \left[A(\theta)'A(\theta)\right]^{-1}$$
$$E[\psi(x;\theta)] = 0.$$

The weights (11) are attributed to each observation according to its influence on the estimator. The constant c is a regulator between efficiency and robustness: The lower c the more robust is the OBRE but also the less efficient. A common method for choosing the constant c is to choose an efficiency level (compared to the MLE) and derive the corresponding value for c. Indeed, one can use the asymptotic covariance of $\sqrt{n}\widehat{\theta}$ given by

$$\text{var}(\widehat{\theta}) = \frac{1}{M^2(\theta)} \int \psi^2(x;\theta) dF_\theta(x)$$

with

$$M(\theta) = -\int \frac{\partial}{\partial\theta}\psi(x;\theta) dF_\theta(x)$$
$$= \int \psi(x;\theta)s(x;\theta) dF_\theta(x)$$

(see [16]). For the Pareto model, a value of $c = 2$ leads to an OBRE achieving (approximately) 85% efficiency.

Finally, for the choice of the proportion α of upper incomes to model by means of the Pareto distribution, we propose to use the robust approach developed in Dupuis and Victoria-Feser [11]. The latter develop a robust prediction error criterion (*RC*-criterion) by viewing the Pareto model as a regression model. Indeed, rearranging (8) one gets

$$\log\left(\frac{x}{x_0}\right) = -\frac{1}{\theta}\log(1 - F_\theta(x))$$

showing that there is a linear relationship between the log of the x and the log of the inverse cumulative distribution function. Given a sample of ordered data $X_{[i]}$, the Pareto regression plot of $\log(X_{[i]})$ *versus* $-\log\left(\frac{n+1-i}{n+1}\right)$, $i = 1, \ldots, n$ 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 will use the results of Dupuis and Victoria-Feser [11] when analyzing the dataset in section 4.2.

3.3. First and second order semi-parametric rankings

The quantile functional is obtained using (9) and is given by

$$Q(\widetilde{F}, q) = \begin{cases} Q(F, q) & q \le 1 - \alpha \\ Q(F; 1-\alpha)\left(\dfrac{1-q}{\alpha}\right)^{-1/\theta(\widetilde{F})} & q > 1 - \alpha \end{cases}.$$

Hence the cumulative income functional defining the semi-parametric GLC becomes

$$
\begin{aligned}
C(\widetilde{F};q) &= \int_{\underline{x}}^{Q(\widetilde{F},q)} x d\widetilde{F}(x) \\
&= \begin{cases} \displaystyle\int_{\underline{x}}^{Q(F,q)} x dF(x) & q \le 1-\alpha \\ \displaystyle + \alpha \int_{Q(F,1-\alpha)}^{Q(F;1-\alpha)\left(\frac{1-q}{\alpha}\right)^{-1/\widehat{\theta}(\widetilde{F})}} x dF_{\widehat{\theta}(\widetilde{F}),Q(F;1-\alpha)} & q > 1-\alpha \end{cases} \\
&= \begin{cases} \displaystyle\int_{\underline{x}}^{Q(F,q)} x dF(x) & q \le 1-\alpha \\ \displaystyle + \alpha \frac{\widehat{\theta}(\widetilde{F})}{1-\widehat{\theta}(\widetilde{F})} Q(F;1-\alpha)\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\widehat{\theta}(\widetilde{F})-1}{\widehat{\theta}(\widetilde{F})}} - 1\right] & q > 1-\alpha \end{cases}
\end{aligned}
$$

where $\underline{x} := \inf \mathfrak{X}$. An estimator is given by $\widehat{c}_q = C(F^{(n)};q)$. The mean of the semi-parametric distribution is given by:

$$
\begin{aligned}
C(\widetilde{F};1) &= \int_{\underline{x}}^{Q(F,1-\alpha)} x dF(x) - \alpha Q(F;1-\alpha) \frac{\widehat{\theta}(\widetilde{F})}{1-\widehat{\theta}(\widetilde{F})} \\
&= c_{1-\alpha} - \alpha x_{1-\alpha} \frac{\theta}{1-\theta} \\
&= \mu(\widetilde{F})
\end{aligned}
$$

The semi-parametric RLC is simply

$$
L(\widetilde{F};q) = \frac{C(\widetilde{F};q))}{\mu(\widetilde{F})} \tag{12}
$$

which is estimated by $\widehat{l}_q = L(F^{(n)};q)$.

In the Appendix, we provide the necessary tools for inference with semi-parametric LCs.

4. Data analysis

4.1. Simulated examples

In order to test our semi-parametric RLC we performed the following simulation exercise. Two samples of 10 000 observations were simulated from a Dagum type I distribution given by

$$
f(x;\beta,\lambda,\delta) = (\beta+1)\lambda\delta x^{-(\delta+1)}(1+\lambda x^{-\delta})^{-(\beta+1)} \tag{13}
$$

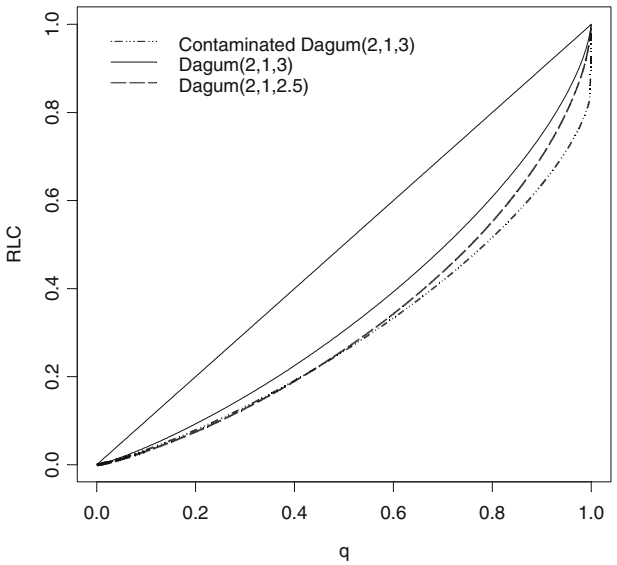

Figure 1. Contaminated Dagum-I distribution

[8].[4] The values of the parameters were chosen in order to get two distributions such that one exactly RLC-dominates the other. They are the Dagum(2,1,3) (i.e., $\beta = 2$, $\lambda = 1$, $\delta = 3$) and the Dagum(2,1,2.5). We then contaminated the Dagum(2,1,3) by multiplying 0.25% of the largest observations by 10. It should be noted that we chose to contaminate only one distribution for simplicity of exposition. The aim is to show that with a semi-parametric robust estimator the RLC is not biased by data contamination, so that if it is the case for one distribution, it is also expected to be the case for the other distribution. Obviously, this relatively extreme example is intended to show the robustness of our RLC semi-parametric robust estimator and not necessarily to represent a situation frequently encountered in practice. A practical example is analyzed in section 4.2. The RLC for the uncontaminated and contaminated Dagum(2,1,3) and the Dagum(2,1,2.5) are given in Figure 1. We can see that the original dominance order no longer holds because the contaminated Dagum(2,1,3) is completely determined by 0.25% extreme observations introduced into the data.

The non-parametric RLC clearly gives a misleading picture. We can avoid this by modelling the upper tail of the Dagum(2,1,3) distribution using the Pareto-tail model as explained above. We used INeQ [18] which computes the MLE and the OBRE for the Pareto model and chose $c = 2$ and $\alpha = 5\%$. The values of $\widehat{\theta}$ (with standard errors) for the non-contaminated sample are, respectively, $\widehat{\theta} = 2.82(0.126)$ for the MLE and $\widehat{\theta} = 2.78(0.134)$ for the OBRE, whereas for the contaminated sample they are, respectively, $\widehat{\theta} = 2.11(0.094)$ for the MLE and $\widehat{\theta} = 2.78(0.134)$ for the OBRE. We can see that the OBRE remains stable whereas the MLE is influenced by data contamination. We then estimated the semi-parametric RLC using (12) in which $\widetilde{F}$ is replaced by $F^{(n)}$ and using either the MLE or the OBRE for θ.

[4] The form (13) has the property that for large values of x, the distribution converges to the Pareto distribution. Note also that this model can be seen as a particular case of the generalized Beta distribution proposed by McDonald and Ransom [21] and is a well-known model for income distributions.

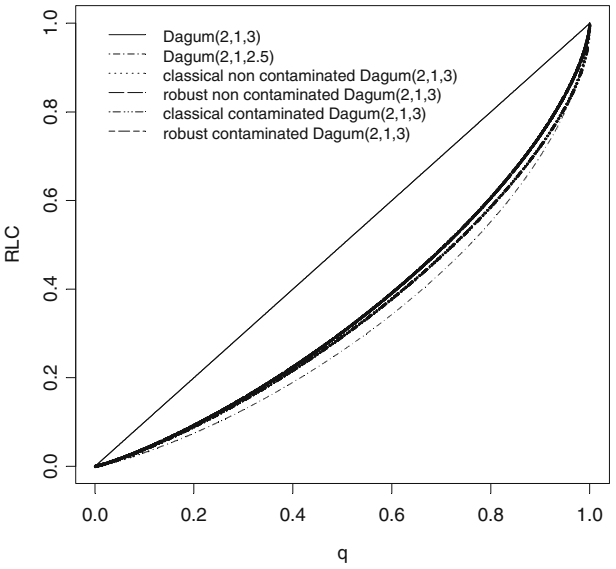

Figure 2. Semi-parametric approach RLCs

The results are compared to the non-parametric RLC using the non-contaminated sample in Figure 2. Figure 3 presents the same picture but zoomed in the upper tail of the distribution. We can see that the semi-parametric RLC on non-contaminated data and/or using a robust estimator are very near to the non-parametric RLC with non-contaminated data. However, when one uses a semi-parametric RLC with a classical estimator on contaminated data, the picture is distorted and the resulting RLC actually crosses the RLC of the Dagum(2,1,2.5) data. It should be noted that it is

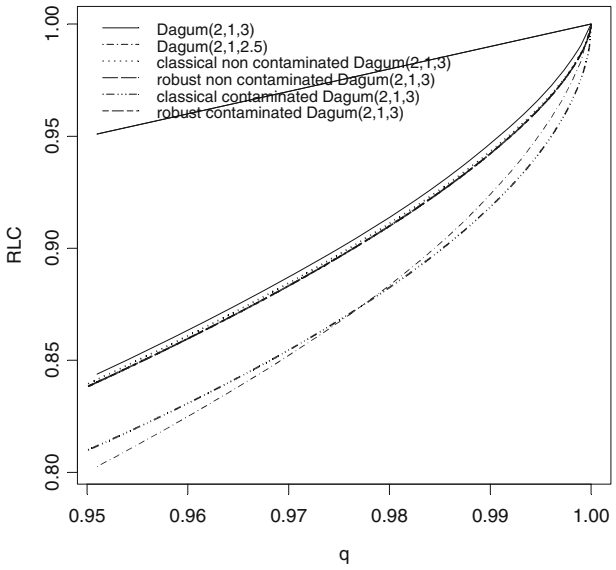

Figure 3. Semi-parametric Lorenz rankings: Classical and robust

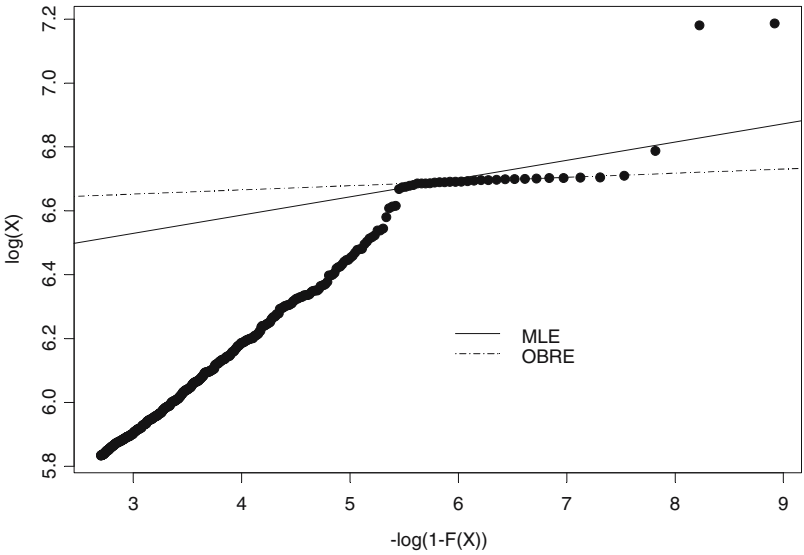

Figure 4. Pareto regression plot with fitted lines (MLE and OBRE with $c = 2$) of the UK income data

not as distorted as with the non-parametric RLC given in Figure 1. Hence, with the robust semi-parametric RLC, the dominance order is preserved with or without contamination, whereas with the classical semi-parametric RLC on contaminated data the curves cross, thus contradicting the original order.

4.2. UK incomes

The data are for household disposable incomes in the UK, 1981 ($n = 7470$).[5] This data set has also been analyzed by Dupuis and Victoria-Feser [11] who found, using a robust approach, that the upper 22 observations should be modelled by means of a Pareto distribution. Using INeQ [18], we found for the MLE and the OBRE ($c = 2$) of the Pareto parameter, respectively, $\widehat{\theta} = 17.5\ (3.73)$ and $\widehat{\theta} = 76.65\,(17.62)$. This large difference in estimates can be properly seen in the Pareto regression plot in Figure 4 on which the estimated regression lines based on $\widehat{\theta}$ have been drawn. While the OBRE properly captures the linear part of the upper tail of the income distribution, the MLE is unduly influenced by the three extreme observations. In Figure 5 are presented the empirical and the two semi-parametric RLC of the UK data (only the 0.5% upper tail). Even if it is small, one can see a difference between the three estimates, in that the MLE follows the empirical RLC up to roughly the 0.1% of the top distribution, while the OBRE leads to an estimated RLC showing less inequality on the entire 0.5% top range.

[5] The data set is Households Below Average Income which, despite its name, actually provides a representative sample of households over the whole income range – see Department of Social Security [10] for details.

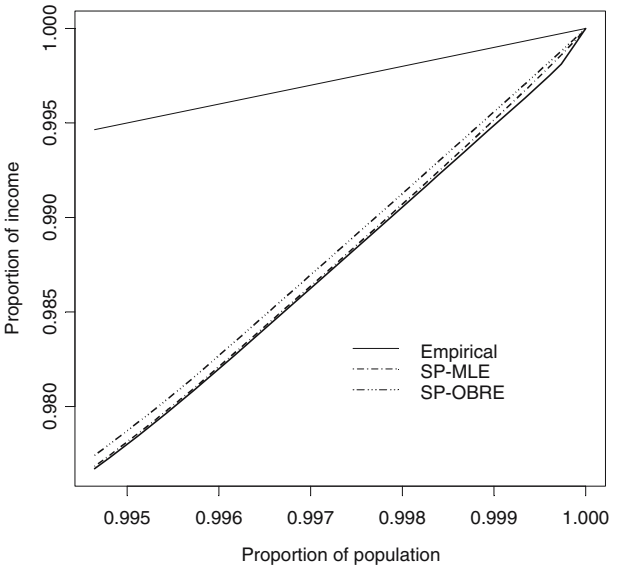

Figure 5. RLC (top 0.5%) estimates (empirical and semi-parametric with MLE and OBRE with $c = 2$) of the UK income data

5. Conclusion

Using ranking criteria to compare distributions is of immense theoretical advantage and practical convenience. In welfare economics they provide a connection between the philosophical basis of welfare judgments and elementary statistical tools for describing distributions. In practical applications they suggest useful ways in which simple computational procedures may be used to draw inferences from collections of empirical distributions. However, since it has been shown that second order rankings are not robust to data contamination, especially in the upper tail of the distribution, it is important to provide the empirical researcher with computational devices which can be used to draw inferences about the properties of distributional comparisons in a robust fashion.

One way forward might be to estimate Lorenz curves through an appropriately specified parametric model and to estimate the model parameters robustly. However, this approach is too restrictive because tractable parametric models are unlikely to be sufficiently flexible to capture some of the essential nuances of Lorenz comparisons. For example, in order for Lorenz curves to be able to cross, a parametric model would usually need to incorporate at least three parameters, which itself may lead to serious estimation complications.

The method proposed here is a semi-parametric approach in that the upper tail of the distribution is robustly fitted using the Pareto model and a semi-parametric Lorenz curve is then built which combines non-parametric cumulative functionals and estimated ones. Simulated examples have proved not only that a few extreme data can reverse the ranking order, but also that the robust parametric Lorenz curve restores the initial ordering. Inference can be made for comparing two distributions even in the semi-parametric setting, by extending the general setting provided in Cowell and Victoria-Feser [6]. For variances too, a robust approach provides reasonable estimates when there is contamination.

Finally note that although we took the Pareto distribution as a suitable parametric model for the upper tail, and although we considered the (most common) case of a range of definition for the variable bounded below, our results can be extended to other models and/or to a two-tail modeling in a straightforward manner.

Acknowledgments We would like to thank Emmanuel Flachaire for useful comments. The second author is partially supported by the Swiss National Fund (grant no 610-057883.99 and PP001-106465).

Appendix: Interference with semi-parametric LCs

As noted in section 2 the *IF* can be used to derive asymptotic covariance matrices. This can be done quite easily using the same approach as in Cowell and Victoria-Feser [6]; see also Beach and Davidson [2], Bishop et al. [3], Bishop et al. [4] and Davidson and Duclos [9].

First we need to compute the *IF* of $\widehat{\theta}(\widetilde{F})$; this is given in the following theorem:

THEOREM 1. *If $\widehat{\theta}(\widetilde{F})$ is a consistent estimator of θ which is implied by (Fisher consistency)*

$$\int_{x_{1-\alpha}}^{\infty} \psi(x;\theta,x_{1-\alpha})dF_{\theta,x_{1-\alpha}}(x) = 0 \tag{14}$$

then we have that the IF of $\widehat{\theta}(\widetilde{F})$ is

$$IF(z;\widehat{\theta},\widetilde{F}) = \frac{1}{\alpha M(\theta)}\psi(z;\theta,x_{1-\alpha})\iota(z > x_{1-\alpha}) \tag{15}$$

Proof. (14) implies $\frac{\partial}{\partial a}\left[\int_a^{\infty}\psi(x;\theta,a)dF_{\theta,x_{1-\alpha}}(x)\right]_{a=x_{1-\alpha}} = -\int_{x_{1-\alpha}}^{\infty}\psi(x;\theta,x_{1-\alpha})\frac{\partial}{\partial x_{1-\alpha}}\log f(x;\theta,x_0)dF_{\theta,x_{1-\alpha}}(x) = -\frac{\theta}{x_{1-\alpha}}\int_{x_{1-\alpha}}^{\infty}\psi(x;\theta,x_{1-\alpha})dF_{\theta,x_{1-\alpha}}(x) = 0$. Applying (5), $IF(z;\widehat{\theta},\widetilde{F}) = \frac{\partial}{\partial\varepsilon}\widehat{\theta}(F_\varepsilon)_{\varepsilon=0}$ is obtained through $\frac{\partial}{\partial\varepsilon}\left[\int_{Q(F_\varepsilon;1-\alpha)}^{\infty}\psi(x;\widehat{\theta}(F_\varepsilon),Q(F_\varepsilon;1-\alpha))dF_\varepsilon(x)\right]_{\varepsilon=0} = 0$ which is

$$\begin{aligned}
&\frac{\partial}{\partial\varepsilon}\left[(1-\varepsilon)\int_{Q(F_\varepsilon;1-\alpha)}^{\infty}\psi(x;\widehat{\theta}(F_\varepsilon),Q(F_\varepsilon;1-\alpha))d\widetilde{F}(x)\right]_{\varepsilon=0}\\
&\quad+\frac{\partial}{\partial\varepsilon}\left[\varepsilon\psi(z;\widehat{\theta}(F_\varepsilon),Q(F_\varepsilon;1-\alpha))\iota(z>Q(F_\varepsilon;1-\alpha))\right]_{\varepsilon=0}\\
&=-\alpha\int_{x_{1-\alpha}}^{\infty}\psi(x;\theta,x_{1-\alpha})dF_{\theta,x_{1-\alpha}}(x)\\
&\quad+\frac{\partial}{\partial\varepsilon}\left[\int_{Q(F_\varepsilon;1-\alpha)}^{\infty}\psi(x;\widehat{\theta}(F_\varepsilon),Q(F_\varepsilon;1-\alpha))d\widetilde{F}(x)\right]_{\varepsilon=0}\\
&\quad+\psi(z;\theta,x_{1-\alpha})[\iota(z>x_{1-\alpha})]\\
&=+\alpha\frac{\partial}{\partial a}\left[\int_a^{\infty}\psi(x;\theta,a)dF_{\theta,x_{1-\alpha}}(x)\right]_{a=x_{1-\alpha}}\frac{\partial}{\partial\varepsilon}Q(F_\varepsilon;1-\alpha)\Big|_{\varepsilon=0}\\
&\quad+\alpha\left[\int_{x_{1-\alpha}}^{\infty}\frac{\partial}{\partial\theta}\psi(x;\theta,x_{1-\alpha})dF_{\theta,x_{1-\alpha}}(x)\right]\frac{\partial}{\partial\varepsilon}\widehat{\theta}(F_\varepsilon)\Big|_{\varepsilon=0}\\
&\quad+\psi(z;\theta,x_{1-\alpha})[\iota(z>x_{1-\alpha})]\\
&=0
\end{aligned}$$

Solving for $\frac{\partial}{\partial\varepsilon}\widehat{\theta}(F_\varepsilon)\Big|_{\varepsilon=0}$ we get (15). □

To derive the asymptotic covariance matrix of RLC ordinates, we then need the *IF* of the cumulative income functionals.

THEOREM 2. *The IF of* $\widehat{c}_q$ *is*

$$IF(z;\widehat{c}_q,\widetilde{F}) = \begin{cases} qx_q - c_q + \iota(x_q \geq z)[z - x_q] & \text{if } q \leq 1-\alpha \\ C(q) + D(q)[\iota(x_{1-\alpha} \geq z)] & \\ +[\iota(x_{1-\alpha} \geq z)][z - x_{1-\alpha}] & \\ +E(q)\frac{1}{M(\theta)}\psi(z;\theta,x_{1-\alpha})[\iota(z > x_{1-\alpha})] & \text{if } q > 1-\alpha \end{cases} \tag{16}$$

where ι *is the indicator function and*

$$C(q) = (1-\alpha)x_{1-\alpha} - c_{1-\alpha} + \frac{(1-\alpha)x_{1-\alpha}}{1-\theta}\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}} - 1\right] \tag{17}$$

$$D(q) = -\frac{x_{1-\alpha}}{1-\theta}\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}} - 1\right] \tag{18}$$

$$E(q) = \frac{x_{1-\alpha}}{\theta(1-\theta)}\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}}\log\left(\frac{1-q}{\alpha}\right) + \frac{\theta}{(1-\theta)}\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}} - 1\right]\right] \tag{19}$$

with

$$C(1) = (1-\alpha)x_{1-\alpha} - c_{1-\alpha} - \frac{(1-\alpha)x_{1-\alpha}}{1-\theta}$$
$$D(1) = \frac{x_{1-\alpha}}{1-\theta}$$
$$E(1) = -\frac{x_{1-\alpha}}{(1-\theta)^2}$$

Proof. For $q \leq 1-\alpha$ see Cowell and Victoria-Feser [5]. For $q > 1-\alpha$, applying (5) we get

$$\frac{\partial}{\partial\varepsilon}\left[\int_{\underline{x}}^{Q(F_\varepsilon,1-\alpha)} x dF_\varepsilon(x) + \alpha\frac{\widehat{\theta}(F_\varepsilon)}{1-\widehat{\theta}(F_\varepsilon)}Q(F_\varepsilon;1-\alpha)\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\widehat{\theta}(F_\varepsilon)-1}{\widehat{\theta}(F_\varepsilon)}} - 1\right]\right]_{\varepsilon=0}$$

$$= (1-\alpha)x_{1-\alpha} - c_{1-\alpha} + [\iota(x_{1-\alpha} \geq z)][z - x_{1-\alpha}]$$
$$+ \alpha\left[x_{1-\alpha}\frac{\partial}{\partial\varepsilon}\left[\frac{\widehat{\theta}(F_\varepsilon)}{1-\widehat{\theta}(F_\varepsilon)}\right]_{\varepsilon=0} + \left(\frac{\theta}{1-\theta}\right)\frac{\partial}{\partial\varepsilon}Q(F_\varepsilon;1-\alpha)\Big|_{\varepsilon=0}\right]\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}} - 1\right]$$
$$+ \alpha\frac{\theta}{1-\theta}x_{1-\alpha}\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}}\log\left(\frac{1-q}{\alpha}\right)\frac{\partial}{\partial\varepsilon}\left(\frac{\widehat{\theta}(F_{\varepsilon)-1}}{\widehat{\theta}(F_\varepsilon)}\right)_{\varepsilon=0}\right]$$

Given that $\frac{\partial}{\partial \varepsilon} Q(F_\varepsilon; 1-\alpha)\big|_{\varepsilon=0} = \frac{q-\iota(x_{1-\alpha}\geq z)}{f(x_{1-\alpha})}$ Staudte and Sheather [24] and using (10) and (15) we get

$$
\begin{aligned}
&(1-\alpha)x_{1-\alpha} - c_{1-\alpha} + \frac{(1-\alpha)x_{1-\alpha}}{1-\theta}\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}} - 1\right] \\
&\quad - \frac{x_{1-\alpha}}{1-\theta}\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}} - 1\right][\iota(x_{1-\alpha}\geq z)] \\
&\quad + \frac{x_{1-\alpha}}{\theta(1-\theta)}\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}}\log\left(\frac{1-q}{\alpha}\right)\right. \\
&\quad \left. + \frac{\theta}{(1-\theta)}\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}} - 1\right]\right]\frac{1}{M(\theta)}\psi(z;\theta,x_{1-\alpha})[\iota(z > x_{1-\alpha})] \\
&\quad + [\iota(x_{1-\alpha}\geq z)][z - x_{1-\alpha}]
\end{aligned}
$$

On rearranging we then get (16). □

We then use (6) to obtain the asymptotic covariances for the semi-parametric income functionals.

THEOREM 3. *For any $q, q', q \leq q'$ the asymptotic covariance of $\sqrt{n}\widehat{c}_q$ and $\sqrt{n}\widehat{c}_{q'}$ is*

$$
\omega_{qq'} = \begin{cases}
s_q + (qx_q - c_q)(x_{q'} - q'x_{q'} + c_{q'}) - x_q c_q & q, q' \leq 1-\alpha \\
s_q + (qx_q - c_q)(c_{1-\alpha} + \alpha x_{1-\alpha} - \alpha D(q')) - c_q x_q & q \leq 1-\alpha < q' \\
\begin{aligned}
&s_{1-\alpha} - 2c_{1-\alpha}x_{1-\alpha} + (1-\alpha)x_{1-\alpha}^2 \\
&+ (C(q) + D(q) + C(q') + D(q'))(c_{1-\alpha} - (1-\alpha)x_{1-\alpha}) \\
&+ C(q)C(q') + (1-\alpha)C(q)D(q') \\
&+ (1-\alpha)C(q')D(q) + (1-\alpha)D(q')D(q) \\
&+ \alpha E(q')E(q)var(\widehat{\theta})
\end{aligned} & 1-\alpha < q, q'
\end{cases} \quad . \tag{20}
$$

Proof. (a) $q, q' \leq 1-\alpha$: see Cowell and Victoria-Feser [5, 6], Theorem 2. b) $q \leq 1-\alpha < q'$: We have to integrate with respect to $\widetilde{F}$ the quantity

$$
\begin{aligned}
&C(q')\big[qx_q - c_q + \iota(x_q \geq z)[z - x_q]\big] \\
&\quad + D(q')[\iota(x_{1-\alpha}\geq z)]\big[qx_q - c_q + \iota(x_q \geq z)[z - x_q]\big] \\
&\quad + E(q')\frac{1}{M(\theta)}\psi(z;\theta,x_{1-\alpha})[\iota(z > x_{1-\alpha})]\big[qx_q - c_q + \iota(x_q \geq z)[z - x_q]\big] \\
&\quad + [\iota(x_{1-\alpha}\geq z)][z - x_{1-\alpha}]\big[qx_q - c_q + \iota(x_q \geq z)[z - x_q]\big]
\end{aligned}
$$

which gives the second line in (20). c) $1-\alpha < q, q'$: We have to integrate with respect to $\widetilde{F}$ the quantity

$$
\begin{aligned}
&C(q)C(q') + C(q)D(q')[\iota(x_{1-\alpha} \geq z)] \\
&+ C(q)E(q')\frac{1}{M(\theta)}\psi(z;\theta,x_{1-\alpha})[\iota(z > x_{1-\alpha})] \\
&+ C(q)[\iota(x_{1-\alpha} \geq z)][z - x_{1-\alpha}] + D(q)C(q')[\iota(x_{1-\alpha} \geq z)] \\
&+ D(q)D(q')[\iota(x_{1-\alpha} \geq z)] \\
&+ D(q)E(q')\frac{1}{M(\theta)}\psi(z;\theta,x_{1-\alpha})[\iota(x_{1-\alpha} \geq z)][\iota(z > x_{1-\alpha})] \\
&+ D(q)[\iota(x_{1-\alpha} \geq z)][z - x_{1-\alpha}] \\
&+ E(q)C(q')\frac{1}{M(\theta)}\psi(z;\theta,x_{1-\alpha})[\iota(z > x_{1-\alpha})] \\
&+ E(q)D(q')\frac{1}{M(\theta)}\psi(z;\theta,x_{1-\alpha})[\iota(z > x_{1-\alpha})][\iota(x_{1-\alpha} \geq z)] \\
&+ E(q)E(q')\frac{1}{M^2(\theta)}\psi^2(z;\theta,x_{1-\alpha})[\iota(z > x_{1-\alpha})] \\
&+ E(q)\frac{1}{M(\theta)}\psi(z;\theta,x_{1-\alpha})[\iota(z > x_{1-\alpha})][\iota(x_{1-\alpha} \geq z)][z - x_{1-\alpha}] \\
&+ C(q')[\iota(x_{1-\alpha} \geq z)][z - x_{1-\alpha}] + D(q')[\iota(x_{1-\alpha} \geq z)][z - x_{1-\alpha}] \\
&+ E(q')\frac{1}{M(\theta)}\psi(z;\theta,x_{1-\alpha})[z - x_{1-\alpha}][\iota(x_{1-\alpha} \geq z)][\iota(z > x_{1-\alpha})] \\
&+ [\iota(x_{1-\alpha} \geq z)][z - x_{1-\alpha}][z - x_{1-\alpha}]
\end{aligned}
$$

which gives the last four lines in (20). □

The estimation of $\omega_{qq'}$ is relatively straightforward. Given a sample $\{x_{[1]}, \ldots, x_{[n]}\}$ of ordered data, letting $n_{1-\alpha} = \text{int}((n-1)(1-\alpha)+1)$ we can obtain $\widehat{\theta}$ and $var(\widehat{\theta})$ from $\{x_{[n_{1-\alpha}]}, \ldots, x_{[n]}\}$. The set of proportions $\{q_i = \frac{i}{n}, i = 1, n\}$ is then defined and $\omega_{qq'}$ is estimated by $\widehat{\omega}_{q_i q_j}$ obtained by replacing in (20), (17), (18) and (12), q by q_i and q' by q_j, x_q by $x_{[i]}$ and $x_{q'}$ by $x_{[j]}$ and $x_{1-\alpha}$ by $x_{[n_{1-\alpha}]}$, c_q by $\frac{1}{n}\sum_{k=1}^{i} x_{[k]}$ and $c_{q'}$ by $\frac{1}{n}\sum_{k=1}^{j} x_{[k]}$ and $c_{1-\alpha}$ by $\frac{1}{n}\sum_{k=1}^{n_{1-\alpha}} x_{[k]}$, s_q by $\frac{1}{n}\sum_{k=1}^{i} x_{[k]}^2$ and $s_{1-\alpha}$ by $\frac{1}{n}\sum_{k=1}^{n_{1-\alpha}} x_{[k]}^2$, and θ by $\widehat{\theta}$.

To extend the results for the cumulative income functional to the Lorenz curve is also straightforward. Indeed, the covariance between $\sqrt{n}\widehat{l}_q$ and $\sqrt{n}\widehat{l}_{q'}$ is obtained using the standard results on limiting distributions of differentiable functions of random variables, and is given by

$$v_{qq'} = \frac{1}{\mu^4}\left[\mu^2\omega_{qq'} - \mu\left(c_{q'}\omega_{q1} + c_q\omega_{q'1}\right) + c_q c_{q'}\omega_{11}\right]$$

where $\mu = \mu(\widetilde{F})$. It is estimated in the same manner as $\omega_{qq'}$.

Empirical comparison of variances

It is interesting to compare asymptotic variances for RLC ordinates when computed on empirical RLC or semi-parametric RLC and with or without contaminated data.

We did this by taking the simulated samples used when we compared the two approaches, i.e., 10 000 data from a Dagum(2,1,3), a contaminated Dagum(2,1,3) and a Dagum(2,1,2.5). We computed the asymptotic variances for the empirical RLC and for the semi-parametric RLC using the MLE and the OBRE ($c = 2$) and their standard errors obtained on the top 5% of the data. The results are presented in Figure 6.

We can draw the following conclusions. First, the semi-parametric approach leads to similar variances in the non-contaminated samples (top two panels). In these cases the semi-parametric approach using the OBRE leads to relatively larger variances when compared to the MLE, which is to be expected since the OBRE is less efficient than the MLE. Second, when there is contamination, variances obtained through the non-parametric approach are excessively large when compared to the uncontaminated case (bottom-left panel). Third, with contaminated data, variances for the semi-parametric RLC are considerably larger with the MLE than with the OBRE (bottom-left panel). Fourth, variances for the semi-parametric RLC with the OBRE in the contaminated case are comparable to the nonparametric and semi-parametric cases in the uncontaminated case (bottom-right panel). So, in cases where there are contaminated data, it is always better to use a semi-parametric approach in which the unknown parameters are estimated robustly.

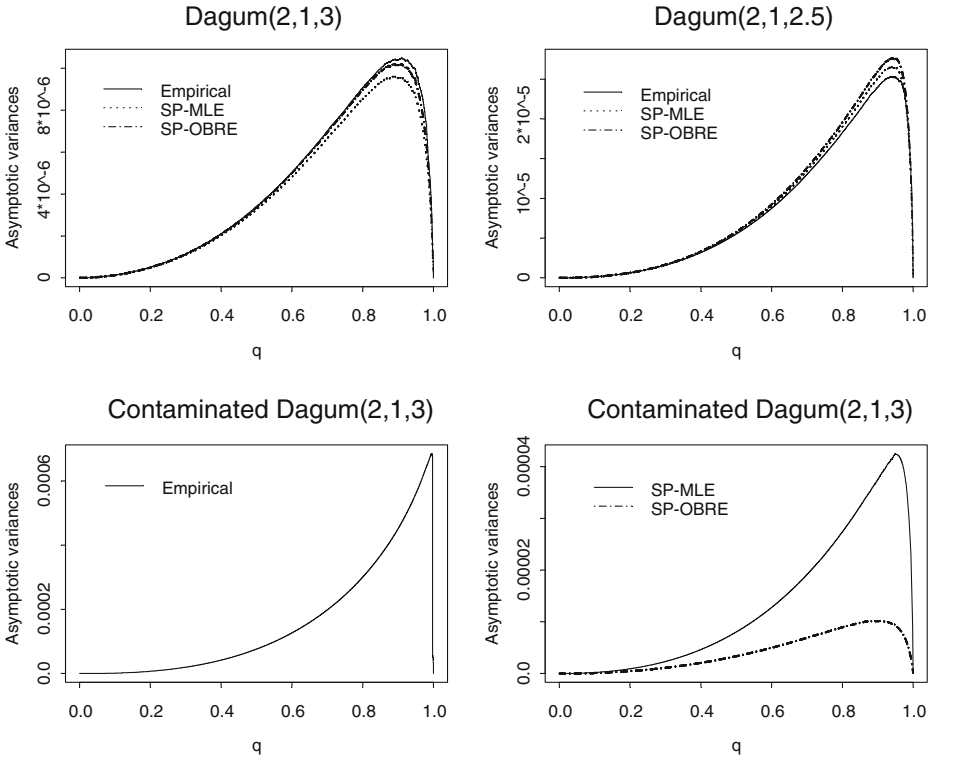

Figure 6. Variance comparisons between empirical and semi-parametric RLC, with and without contamination

References

1. Atkinson, A.B.: Income tax and top incomes over the twentieth century. Hacienda Pública Esp. **168**, 123–141 (2004)
2. Beach, C.M., Davidson, R.: Distribution-free statistical inference with Lorenz curves and income shares. Rev. Econ. Stud. **50**, 723–735 (1983)
3. Bishop, J.A., Chakraborti, S., Thistle, P.D.: Large sample tests for absolute Lorenz dominance. Econ. Lett. **26**, 291–294 (1988)
4. Bishop, J.A., Chakraborti, S., Thistle, P.D.: Asymptotically distribution-free statistical inference for generalized Lorenz curves. Rev. Econ. Stat. **71**(11), 725–727 (1989)
5. Cowell, F.A., Victoria-Feser, M.-P.: Welfare rankings in the presence of contaminated data. Econometrica **70**, 1221–1233 (2002)
6. Cowell, F.A., Victoria-Feser, M.-P.: Distribution-free inference for welfare indices under complete and incomplete information. Journal of Economic Inequality **1**, 191–219 (2003)
7. Cowell, F.A., Victoria-Feser, M.-P.: Distributional dominance with trimmed data. J. Bus. Econ. Stat. (2006). To appear
8. Dagum, C.: A new model of personal income distribution: Specification and estimation. Econ. Appl. **30**, 413–436 (1977)
9. Davidson, R., Duclos, J.-Y.: Statistical inference for stochastic dominance and for the measurement of poverty and inequality. Econometrica **68**, 1435–1464 (2000)
10. Department of Social Security: Households below average income: A statistical analysis, 1979–1988/9. HMSO, London (1992)
11. Dupuis, D., Victoria-Feser, M.-P.: A robust prediction error criterion for Pareto modeling of upper tails. Cahiers du GERAD G-2005-29, Montreal (2005)
12. Dupuis, D.J., Field, C.A.: Robust estimation of extremes. Can. J. Stat. **26**, 199–215 (1998)
13. Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events. Applications of Mathematics: Stochastic modelling and applied probability. Springer, Berlin Heidelberg New York (1997)
14. Hampel, F.R.: A general qualitative definition of robustness. The Annals of Mathematics and Statistics **42**, 1887–1896 (1971)
15. Hampel, F.R.: The influence curve and its role in robust estimation. J. Am. Stat. Assoc. **69**, 383–393 (1974)
16. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: Robust statistics: The approach based on influence functions. Wiley, New York (1986)
17. Huber, P.J.: Robust statistics. Wiley, New York (1981)
18. INeQ: Software for distributional analysis, Distributional Analysis Research Programme, STICERD. London School of Economics, London WC2A 2AE, UK (2001)
19. Kopczuk, W., Saez, E.: Top wealth shares in the United States, 1916–2000: Evidence from estate tax returns. NBER Working Paper 10399, National Bureau for Economic Research, Cambridge, Massachusetts (2004)
20. Lorenz, M.O.: Methods for measuring concentration of wealth. J. Am. Stat. Assoc. **9**, 209–219 (1905)
21. McDonald, J.B., Ransom, M.R.: Functional forms, estimation techniques and the distribution of income. Econometrica **47**, 1513–1525 (1979)
22. Piketty, T.: Les hauts revenus en France au 20eme siécle – Inégalités et redistributions, 1901–1998. Editions Grasset, Paris (2001)
23. Piketty, T., Saez, E.: Income inequality in the United States, 1913–1998. Q. J. Econ. **118**, 1–39 (2003)
24. Staudte, R.G., Sheather, S.J.: Robust estimation and testing. Wiley, New York (1990)
25. Victoria-Feser, M.-P., Ronchetti, E.: Robust methods for personal income distribution models. Can. J. Stat. **22**, 247–258 (1994)
26. Victoria-Feser, M.-P., Ronchetti, E.: Robust estimation for grouped data. J. Am. Stat. Assoc. **92**, 333–340 (1997)