ISBN 1352-2469 Distributional Analysis Research Programme

Robust Lorenz Curves: A Semi-Parametric Approach

by

Frank A. Cowell
and
Maria-Pia Victoria-Feser

London School of Economics and Université de Genève

Discussion Paper
No. DARP 50
May 2001

Distributional Analysis Research Programme
The Toyota Centre
Suntory and Toyota International
Centres for Economics and
Related Disciplines
London School of Economics
Houghton Street
London WC2A 2AE

Distributional Analysis Research Programme

The Distributional Analysis Research Programme was established in 1993 with funding from the Economic and Social Research Council. It is located within the Suntory and Toyota International Centres for Economics and Related Disciplines (STICERD) at the London School of Economics and Political Science. The programme is directed by Frank Cowell. The Discussion Paper series is available free of charge and most papers are downloadable from the website. To subscribe to the DARP paper series, or for further information on the work of the Programme, please contact our Research Secretary, Sue Coles on:

Telephone:	UK+20 7955 6678
Fax:	UK+20 7955 6951
Email:	s.coles@lse.ac.uk
Web site:	http://sticerd.lse.ac.uk/DARP

Abstract

Lorenz curves and second-order dominance criteria are known to be sensitive to data contamination in the right tail of the distribution. We propose two ways of dealing with the problem: (1) Estimate Lorenz curves using parametric models for income distributions, 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 approach (2). Statistical inference tools are also provided.

Keywords: Welfare dominance; Lorenz curve; Pareto model, M-estimators.

JEL Classification: C13, D63

Correspondence to: Professor M.-P. Victoria-Feser, Department of Econometrics, University of Geneva, 1211 Geneva, Switzerland. (maria-pia.victoriafeser@pse.unige.ch)

We would like to thank Emmanuel Flachaire for useful comments. The second author is partially supported by the “Fond National Suisse pour la Recherche Scientifque”.

Abstract

Lorenz curves and second-order dominance criteria are known to be sensitive to data contamination in the right tail of the distribution. We propose two ways of dealing with the problem: (1) Estimate Lorenz curves using parametric models for income distributions, 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 approach (2). Statistical inference tools are also provided.

Keywords: Welfare dominance; Lorenz curve; Pareto model, M-estimators.

JEL Classification: C13,D63

Correspondence to: Professor M.-P. Victoria-Feser, Department of Econometrics, University of Geneva, 1211 Geneva, Switzerland. (maria-pia.victoriafeser@pse.unige.ch)

We would like to thank Emmanuel Flachaire for useful comments. The second author is partially supported by the "Fond National Suisse pour la Recherche Scientifique".

1 Introduction

The Lorenz curve is central to the analysis of income distributions. It embodies some fundamental intuition about inequality comparisons. Ranking theorems based on the associated concepts 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 misleading. The purpose of this paper is to provide a rigorous method for handling 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 (Cowell and Victoria-Feser 1996, 2002), This means that small amounts of data contamination in the wrong place can reverse unambiguous welfare conclusions: the "wrong place" usually means in the upper tail of the distribution. 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 (Cowell and Victoria-Feser 2001), raises issues which go beyond the scope of this paper. 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 income and wealth distributions.[1]

The paper is organised as follows. In sections 3 and 4 we discuss two ways of implementing a parametric approach to the estimation of Lorenz curves.

[1] For example a Pareto tail is sometimes fitted to data in cases where data are sparse in order to provide better estimates of inequality measures.

In section 5 we provide the necessary tools for inference on robust Lorenz curves. Section 6 concludes.

2 The Background

Let $\mathfrak{F}$ be the set of all univariate probability distributions and X be a random variable which may be thought of as "income", with probability distribution $F \in \mathfrak{F}$ and support $\mathfrak{X} \subseteq \mathbb{R}$. F can be thought of 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 income *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 in practical analysis of income distributions is considerable: note, for example, that the mean functional emerges as one particular case – $\mu(F) = C(F,1)$ – and the income share of the bottom q of the population is given by $C(F,q)/C(F,1)$. Moreover, for a given $F \in \mathfrak{F}$, the graph of $C(F,q)$ against q describes the *generalised Lorenz curve* (GLC). From the fundamental concept of the cumulative income functional one obtains two other important analytical tools for drawing welfare-conclusions from income data, namely the *Relative Lorenz curve* (RLC)(Lorenz 1905):

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

and the *Absolute Lorenz Curve* (ALC) (Moyes 1987):

$$A(F;q) := C(F;q) - q\mu(F) \tag{4}$$

Cumulative income functionals can obviously be estimated empirically by replacing F in (1) by the empirical distribution. However, this can lead to misleading conclusions when it comes to comparing distributions in terms of their cumulative income functionals when there is data contamination (Cowell and Victoria-Feser 2002).

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 characterise the sensitivity of a statistic to point contamination in the data (see e.g. Hampel et al. 1986) but can also be used to derive asymptotic results such as asymptotic covariance matrices of for example cumulative income functionals (Cowell and Victoria-Feser 1999, 2001). Let Δ_z be a point mass distribution giving probability 1 to an arbitrary point $z \in \mathfrak{X}$ and define the mixture distribution

$$F_{\varepsilon}^{(z)} = (1-\varepsilon)F + \varepsilon\Delta_z \tag{5}$$

$F_{\varepsilon}^{(z)}$ 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

$$IF(z;T,F) = \lim_{\varepsilon \downarrow 0} \frac{T(F_{\varepsilon}^{(z)}) - T(F)}{\varepsilon} \tag{6}$$

which becomes $\frac{\partial}{\partial \varepsilon} T(F_{\varepsilon}^{(z)})\Big|_{\varepsilon=0}$ if T is differentiable. If the *IF* of a statistic T is unbounded or can take large values, then T is said not robust in the infinitesimal sense in that an infinitesimal amount of contaminated data at z can change drastically the value of T. Using the *IF*, one can also compute the asymptotic covariance matrix of T which is obtained by means of

$$\text{cov}(\sqrt{n}T(F^{(n)}) = \int_{\mathfrak{X}} IF(z;T,F)IF'(z;T,F)dF(z) \tag{7}$$

(see Hampel et al. 1986). This result will be used when computing the asymptotic covariance matrix of semi-parametric income functionals.

3 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 robustly

[2] Also called the influence curve and first introduced by Hampel (1971, 1974).

the parameters of the model 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 say $\hat{\theta}$, then robust estimates of the GLC, RLC and ALC are given by respectively

$$C(\hat{\theta}; q) = \int_{\underline{x}}^{Q(F_{\hat{\theta}};q)} x dF_{\hat{\theta}}(x), \tag{8}$$

$$L(\hat{\theta}; q) = \frac{C(F_{\hat{\theta}}; q)}{\mu(F_{\hat{\theta}})}, \tag{9}$$

$$A(\hat{\theta}; q) = C(F_{\hat{\theta}}; q) - \mu(F_{\hat{\theta}}) \cdot q, \tag{10}$$

where $\mu(F_{\hat{\theta}}) = \int x dF_{\hat{\theta}}(x)$. 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 $\hat{\theta}$. If we write the latter as a functional of the contaminated distribution given in (5), i.e. $\hat{\theta}(F_\varepsilon^{(z)})$, then we have

$$IF(z; C, F_\theta) = \frac{\partial}{\partial \theta} C(F_\theta; q) \cdot IF(z; \hat{\theta}, F_\theta). \tag{11}$$

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 (Hampel et al. 1986) and for income distribution (Victoria-Feser and Ronchetti, 1994, 1997). 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 welfare comparisons. For example, if one supposes that the income 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.

4 A semi-parametric approach

In light of the above considerations, we suggest using a semi-parametric approach. If the income range is bounded below – 0 is a typical value – the problems with contaminated data occur in the upper tail of the distribution (Cowell and Victoria-Feser 2002). 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.

4.1 Robust estimation

A suitable model for the upper tail is the Pareto distribution given by

$$F_{\theta,x_0}(x) = 1 - \left[\frac{x}{x_0}\right]^{-\theta} \tag{12}$$

with density $f(x;\theta) = \theta x^{-(\theta+1)} x_0^{\theta}$. The parameter of interest is θ and is assumed to be greater than 2 for the variance to exist. 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

$$\widetilde{F}(x) = \begin{cases} F(x) & x \leq Q(F;1-\alpha) \\ 1 - \alpha\left(\frac{x}{Q(F;1-\alpha)}\right)^{-\theta} & x > Q(F;1-\alpha) \end{cases} . \tag{13}$$

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

$$\widetilde{f}(x_{1-\alpha};\theta) = \frac{\alpha\theta}{x_{1-\alpha}} . \tag{14}$$

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

(Victoria-Feser 1993). Robust estimators for general parametric models have been developed by Hampel et al. (1986) and now implemented in INeQ (2001) for the Pareto model. These are actually bounded *IF* M-estimators (Huber 1981) with minimal asymptotic covariance matrix and are called *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 $\hat{\theta}(\widetilde{F})$ in

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

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 ML estimator. We get the OBRE when

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

with

$$W_c(x; \theta) = \min \left\{ 1; \frac{c}{\|A(\theta)[s(x; \theta) - a(\theta)]\|} \right\} \tag{15}$$

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

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

The weights (15) 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. Finally, the asymptotic covariance of $\sqrt{n}\hat{\theta}$ is given by

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

with

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

(see Hampel et al. 1986).

4.2 First and second order semi-parametric rankings

The quantile functional is obtained using (13) and is given by

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

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

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

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

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

The semi-parametric RLC is simply

$$L(\widetilde{F};q)=\frac{C(\widetilde{F};q))}{\mu(\widetilde{F})} \tag{16}$$

which is estimated by $\widehat{l_q}=L(F^{(n)};q)$. A question arises here about the choice of the proportion α of data to model. We propose a simple rationale

based on prior knowledge of the quality of the data. There are to points to be stressed. First, α should be as small as possible to avoid putting too much of the parametric approach into the ranking exercise, for the reason we mentioned above. Second, α should be large enough so that a majority of data points in the upper tail subsample are uncontaminated data. Let ε be the (suspected) proportion of contaminated data in the whole sample, which should be relatively small. Suppose that the data analyst has a fairly good idea of that quantity which in general depends on the data source. We propose the adoption of a minimax approach in that we assume that the contamination will result in the worst scenario, i.e. in extremely large incomes. To prevent the ranking exercise being completely determined by this proportion ε of contaminated data, one then should get the information on the upper tail through the estimation of the parameter of the Pareto distribution. Then α, the proportion of data to model, should be such that $\frac{\varepsilon}{\alpha}$ equals that proportion of contaminated data that the chosen robust estimator can withstand before it breaks down (see Hampel et al. 1986). In our experience, the OBRE for the Pareto model can withstand up to 10% of contaminated data, so that $\alpha = \frac{1}{0.1}\varepsilon$.

4.3 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

$$f(x;\beta,\lambda,\delta) = (\beta+1)\lambda\delta x^{-(\delta+1)}(1+\lambda x^{-\delta})^{-(\beta+1)} \tag{17}$$

(Dagum 1977).[3] 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. 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 does not anymore hold because the contaminated Dagum(2,1,3) is completely determined by 0.25% extreme observations introduced into the data.

[3]The form (17) 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 (1979).

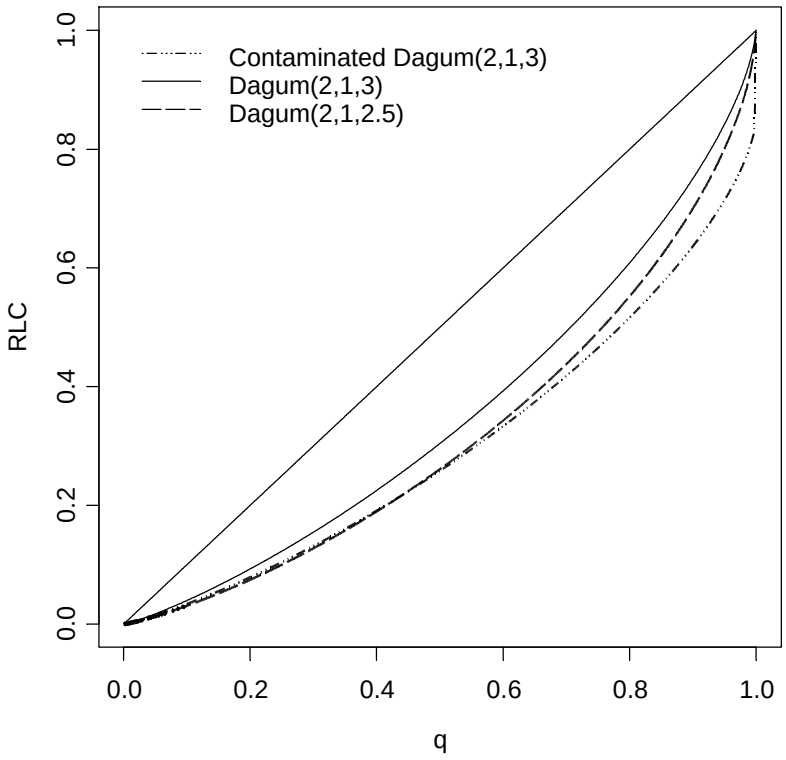

Figure 1: Contaminated Dagum-I distribution

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 (2001) which computes the MLE and the OBRE for the Pareto model and chose $c = 2$ and $\alpha = 5\%$. The values of $\hat{\theta}$ (with standard errors) for the non-contaminated sample are respectively $\hat{\theta} = 2.82(0.126)$ for the MLE estimator and $\hat{\theta} = 2.78(0.134)$ for the OBRE, whereas for the contaminated sample they are respectively $\hat{\theta} = 2.11(0.094)$ for the MLE estimator and $\hat{\theta} = 2.78(0.134)$ for the OBRE. We can see that the OBRE remains very stable whereas the MLE seems to

be quite influenced by data contamination. We then estimated the semi-parametric RLC using (16) in which $\widetilde{F}$ is replaced by $F^{(n)}$ and using either the MLE or the OBRE for θ. 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 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.

5 Inference with semi-parametric LCs

As noted in section 2 the *IF* can be used to derive asymptotic covariance matrices. Although this has been done for a wide variety of data settings and welfare statistics the semi-parametric case has not yet been tackled; nevertheless it can be developed quite easily using the same approach as in Cowell and Victoria-Feser (1999).

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

Theorem 1 *If $\hat{\theta}(\widetilde{F})$ is a consistent estimator of θ which implies that (Fisher consistency)*

$$\int_{x_{1-\alpha}}^{\infty} \psi(x;\theta,x_{1-\alpha})dF_{\theta,x_{1-\alpha}}(x) = 0 \tag{18}$$

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

$$IF(z;\hat{\theta},\widetilde{F}) = \frac{1}{\alpha M(\theta)}\psi(z;\theta,x_{1-\alpha})\iota(z > x_{1-\alpha}) \tag{19}$$

Proof. See Appendix A.1.

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

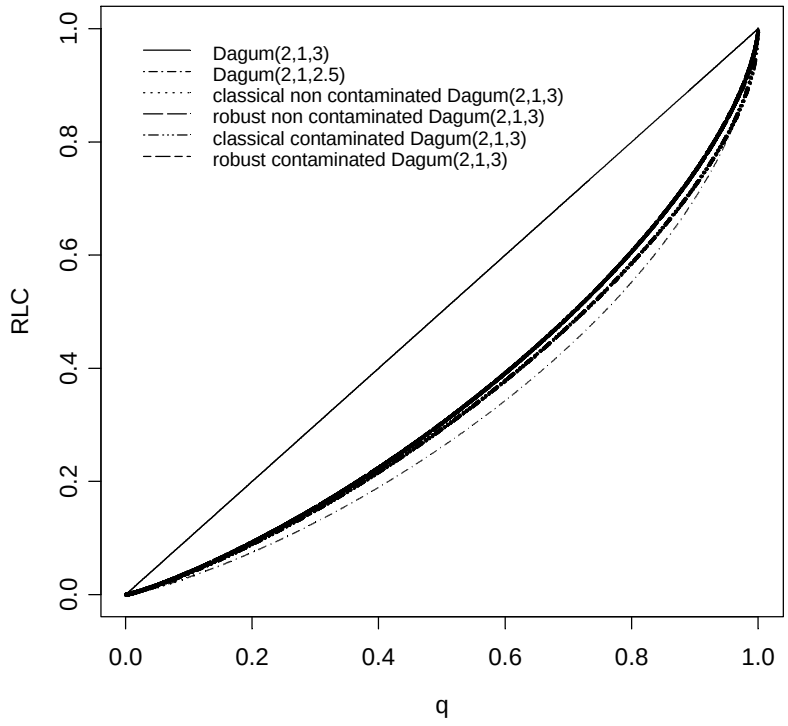

Figure 2: Semi-parametric approach RLCs

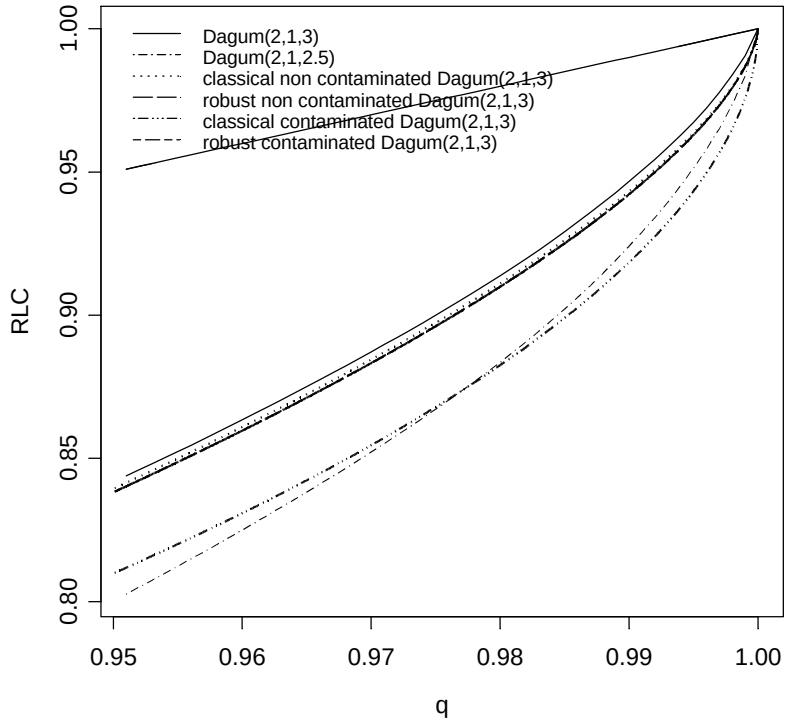

Figure 3: Semi-parametric Lorenz rankings: classical and robust

Theorem 2 *The IF of* $\hat{c}_q$ *is*

$$IF(z;\hat{c}_q,\widetilde{F}) = \begin{cases} qx_q - c_q + \iota(x_q \geq z)[z - x_q] & \text{if } q \leq 1-\alpha \\ C(q) + D(q)\left[\iota(x_{1-\alpha} \geq z)\right] & \\ +\left[\iota(x_{1-\alpha} \geq z)\right]\left[z - x_{1-\alpha}\right] & \\ +E(q)\frac{1}{M(\theta)}\psi(z;\theta,x_{1-\alpha})\left[\iota(z > x_{1-\alpha})\right] & \text{if } q > 1-\alpha \end{cases} \tag{20}$$

where ι *is the indicator function and*

$$\begin{aligned} C(q) &= (1-\alpha)x_{1-\alpha} - c_{1-\alpha} \\ &\quad + \frac{(1-\alpha)x_{1-\alpha}}{1-\theta}\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}} - 1\right] \end{aligned} \tag{21}$$

$$D(q) = -\frac{x_{1-\alpha}}{1-\theta}\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}} - 1\right] \tag{22}$$

$$\begin{aligned} E(q) &= \frac{x_{1-\alpha}}{\theta(1-\theta)}\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}} \log\left(\frac{1-q}{\alpha}\right)\right. \\ &\quad \left. + \frac{\theta}{(1-\theta)}\left[\left(\frac{1-q}{\alpha}\right)^{\frac{\theta-1}{\theta}} - 1\right]\right] \end{aligned} \tag{23}$$

with

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

Proof. See Appendix A.2.

We then use (7) to obtain the asymptotic covariances for the semi-parametric income functionals.

Theorem 3 *For any* q, q', $q \leq q'$ *the asymptotic covariance of* $\sqrt{n}\hat{c}_q$ *and*

$\sqrt{n}\hat{c}_{q'}$ *is*

$$\omega_{qq'} = \begin{cases} s_q + (qx_q - c_q)(x_{q'} - q'x_{q'} + c_{q'}) - x_q c_q & q, q' \leq 1-\alpha \\ \\ s_q + (qx_q - c_q)(c_{1-\alpha} + \alpha x_{1-\alpha} - \alpha D(q')) - c_q x_q & q \leq 1-\alpha < q' \\ \\ \begin{array}{l} s_{1-\alpha} - 2c_{1-\alpha}x_{1-\alpha} + (1-\alpha)x_{1-\alpha}^2 + \\ (C(q) + D(q) + C(q') + D(q'))\,(c_{1-\alpha} - (1-\alpha)x_{1-\alpha}) \\ +C(q)C(q') + (1-\alpha)C(q)D(q') \\ +(1-\alpha)C(q')D(q) + (1-\alpha)D(q')D(q) \\ +\alpha E(q')E(q)var(\hat{\theta}) \end{array} & 1-\alpha < q, q' \end{cases} \tag{24}$$

Proof. See Appendix A.3.

The estimation of $\omega_{qq'}$ is relatively straightforward. Given a sample $\left\{x_{[1]}, \ldots, x_{[n]}\right\}$ of ordered data, letting $n_{1-\alpha} = \text{int}\left((n-1)(1-\alpha)+1\right)$ we can obtain $\hat{\theta}$ and $\text{var}(\hat{\theta})$ from $\left\{x_{[n_{1-\alpha}]}, \ldots, x_{[n]}\right\}$. The set of proportions $\left\{q_i = \frac{i}{n}, i = 1, n\right\}$ is then defined and $\omega_{qq'}$ is estimated by $\hat{\omega}_{q_i q_j}$ obtained by replacing in (24), (21), (22) and (16), 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 $\hat{\theta}$.

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

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

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

5.1 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

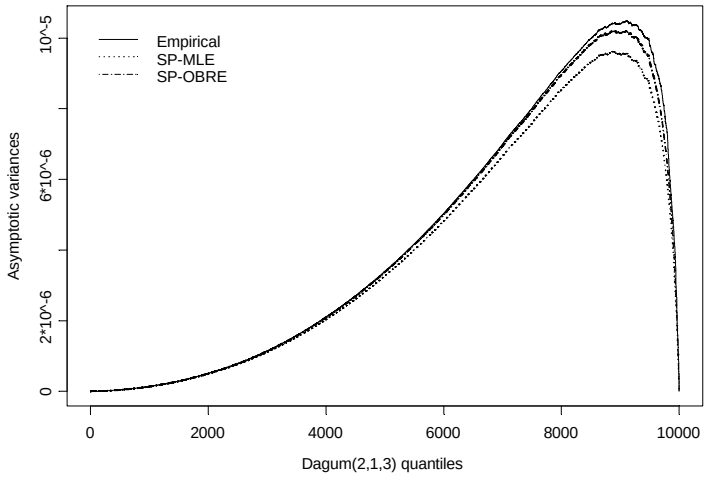

Figure 4: Variance comparisons between empirical and semi-parametric RLC (without contamination)

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 Figures 4 to 7 where in each case the horizontal axis plots 10 $000q$ for $0 < q \leq 1$.

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 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.

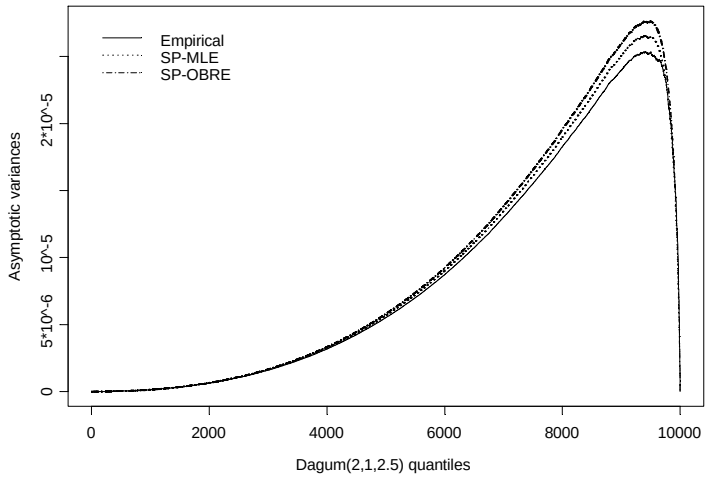

Figure 5: Variance comparisons between empirical and semi-parametric RLC (without contamination)

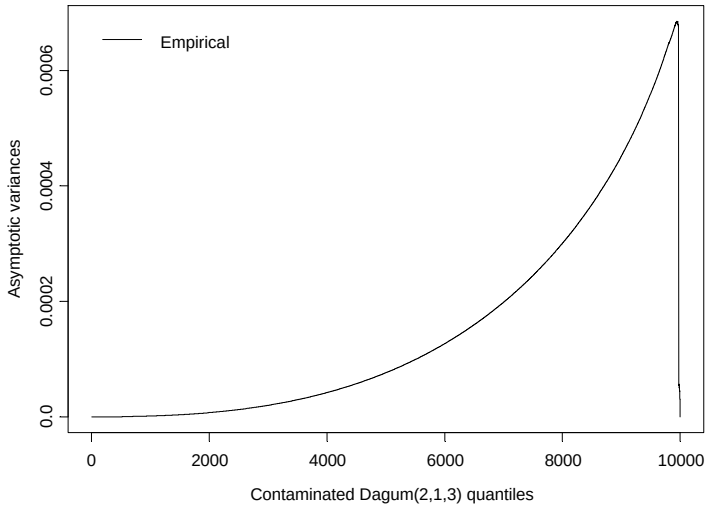

Figure 6: Variance of the empirical RLC (with contamination)

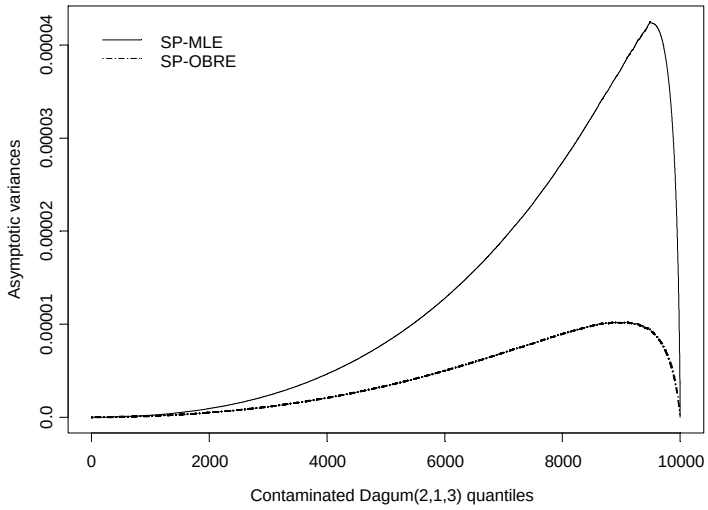

Figure 7: Variance of the semi-parametric RLC (with contamination)

6 Conclusion

Using ranking criteria to make welfare inferences about income distributions is of immense theoretical advantage and practical convenience. As abstract theoretical constructs 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 income 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 restricted welfare inferences about the properties of distributional comparisons in a robust fashion.

One way is to estimate Lorenz curves through the specification of a parametric model and the robust estimation of its parameters. However, this approach is too restrictive in that in order to have the possibility of having crossing Lorenz curves, the models should be very flexible and incorporate at

least three parameters, which might lead serious estimation complications.

The way we propose 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 build which combines non parametric cumulated incomes and estimated ones. Simulated examples have proven not only that a few extreme incomes 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 (1999). For variances too, a robust approach provides reasonable estimates when there is contamination.

Finally, it should be stressed that, although our main discussion is couched in the language of income distribution, all of our analysis is immediately applicable to cognate fields such as the comparison of probability distributions in finance.

References

Cowell, F. A. and M.-P. Victoria-Feser (1996). Welfare judgements in the presence of contaminated data. Distributional Analysis Discussion Paper 13, STICERD, London School of Economics, London WC2A 2AE.

Cowell, F. A. and M.-P. Victoria-Feser (1999). Statistical inference for welfare indices under complete and incomplete information. Distributional Analysis Discussion Paper 47, STICERD, London School of Economics, London WC2A 2AE.

Cowell, F. A. and M.-P. Victoria-Feser (2001). Distributional dominance with dirty data. Distributional Analysis Discussion Paper 51, STICERD, London School of Economics, London WC2A 2AE.

Cowell, F. A. and M.-P. Victoria-Feser (2002). Welfare rankings in the presence of contaminated data. *Econometrica (forthcoming)*.

Dagum, C. (1977). A new model of personal income distribution: Specification and estimation. *Economie Appliquée 30*, 413–436.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematics and Statistics 42*, 1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association 69*, 383–393.

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

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

INeQ (2001). Software for distributional analysis, Distributional Analysis Research Programme, STICERD, London School of Economics, London WC2A 2AE, UK.

Lorenz, M. O. (1905). Methods for measuring concentration of wealth. *Journal of the American Statistical Association 9*, 209–219.

McDonald, J. B. and M. R. Ransom (1979). Functional forms, estimation techniques and the distribution of income. *Econometrica 47*, 1513–1525.

Moyes, P. (1987). A new concept of Lorenz domination. *Economics Letters 23*, 203–207.

Victoria-Feser, M.-P. (1993). *Robust Methods for Personal Income Distribution Models*. Ph. D. thesis, University of Geneva, Switzerland. Thesis no 384.

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

Victoria-Feser, M.-P. and E. Ronchetti (1997). Robust estimation for grouped data. *Journal of the American Statistical Association 92*, 333–340.

A Proofs

A.1 Theorem 1

(18) 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 (6),
$IF(z;\hat{\theta},\widetilde{F}) = \frac{\partial}{\partial\varepsilon}\hat{\theta}(F_\varepsilon)_{\varepsilon=0}$ is obtained through
$\frac{\partial}{\partial\varepsilon}\left[\int_{Q(F_\varepsilon;1-\alpha)}^\infty \psi(x;\hat{\theta}(F_\varepsilon),Q(F_\varepsilon;1-\alpha))dF_\varepsilon(x)\right]_{\varepsilon=0} = 0$ which is

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

Solving for $\frac{\partial}{\partial\varepsilon}\hat{\theta}(F_\varepsilon)\Big|_{\varepsilon=0}$ we get (19). ■

A.2 Theorem 2

For $q \leq 1-\alpha$ see Cowell and Victoria-Feser 2002. For $q > 1-\alpha$, applying (6) we get

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

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

Given that $\frac{\partial}{\partial \varepsilon} Q(F_\varepsilon; 1-\alpha)\big|_{\varepsilon=0} = \frac{q - \iota(x_{1-\alpha} \geq z)}{f(x_{1-\alpha})}$ (Cowell and Victoria-Feser 1999) and using (14) and (19) we get

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

On rearranging we then get (20). ■

A.3 Theorem 3

a) $q, q' \leq 1-\alpha$: see Cowell and Victoria-Feser 2002, 1999, Theorem 2.

b) $q \leq 1-\alpha < q'$: we have to integrate with respect to $\widetilde{F}$ the quantity

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

which gives the second line in (24).

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 (24). ■