How to Retrieve the Lorenz Curve from Sparse Data

Michael Braulke
Fachbereich Wirtschaftswissenschaften
Universität Osnabrück
D-4500 Osnabrück, F.R.G.

It is widely felt among empirically oriented economists that too much of the information on income distribution that is actually collected by statistical offices is lost in what is eventually published. Typically, only a few points of the Lorenz curve are made available whereas in applied work often the entire curve is required. This need to know the entire Lorenz curve arises already when the intention is merely to extract e.g. the Gini concentration ratio with a reasonable degree of accuracy, and it becomes more urgent when the informational content of such uni-dimensional inequality measures is to be supplemented by additional information from the Lorenz curve. For instance, calculation of the minor concentration ratio proposed by Hagerbaumer (1977), which is meant to measure the relative position of the poor, requires sufficiently reliable information on the shape of the Lorenz curve over at least the lower income deciles, whereas the derivation of Lorenz coefficients as suggested by Koo, Quan, and Rasche (1981), which may convey useful information on the contribution of certain income classes to overall inequality[1] presupposes in essence knowledge of the Lorenz curve over all its range. Finally, knowledge of the entire Lorenz curve is obviously indispensable when the extent of overlapping of income distributions of pairs of countries has to be determined as would be necessary, for example, in tests of Linder's hypothesis on intra-industry trade[2].

There are many other applications in applied work where one would prefer having the entire Lorenz curve rather than only a few points on it. The question therefore arises how one may retrieve the unknown Lorenz curve from the sparse data points which are typically available. It is the purpose of this paper to discuss three different approaches to this task and to report briefly the problems encountered in each case.

1 See also Thon (1983) and Koo, Quan, and Rasche (1983).

2 One may in fact wonder whether the poor empirical support obtained so far really speak against the validity of Linder's convincing hypothesis or rather reflect the crude approximations to the extent of income similarity between countries that had to be used due to lack of better information on income distribution. See e.g. Fortune (1979).

Measurement in Economics
ed. by W. Eichhorn
© Physica-Verlag Heidelberg 1988

The approaches considered are (i) fitting some specified functional form to the known data points of the Lorenz curve and taking this estimate as a substitute for the unknown Lorenz curve, (ii) interpolating the known data points by a well-behaved (monotone, convex and differentiable) spline function, and finally (iii) constructing the upper and lower bounds of the band through which the true Lorenz curve must pass. The calculation of such bounds may be of interest on its own, but it will serve here essentially as a yardstick with which to judge the reliability of the other two approaches mentioned.

1. Retrieving the Lorenz Curve by Fitting Specified Functional Forms

Let p and y denote the cumulative proportions of income units and of income received. The best known functional form to represent a Lorenz curve is then presumably that of Kakwani and Podder (1976),

$$\eta = \gamma\pi^{\alpha}(\sqrt{2}-\pi)^{\beta} \quad \text{with} \quad 0<\alpha,\beta\leq 1 \quad \text{and} \quad \gamma\geq 0 \quad , \qquad (1)$$

which expresses the Lorenz curve in terms of the new coordinates $\eta=(p-y)/\sqrt{2}$ and $\pi=(p+y)/\sqrt{2}$. This specification was extensively and obviously also successfully used by Jain (1975) in her compilation of size distributions of income. From a computational point of view this representation combines the advantage of being particularly easy to estimate[3] with the disadvantage of being awkward when it comes to integrating the associated Lorenz curve over some subinterval. But what is more important, the Kakwani and Podder specification (1) has been seriously criticized by Rasche, Gaffney, Koo, and Obst (1980) on theoretical grounds for failing to comply with one of the basic properties of a Lorenz curve, i.e. to stay within the lower triangle of the unit square[4].

---

3 Note that after taking logarithms, (1) is linear in its parameters.

4 Indeed, it is not difficult to check that the Kakwani and Podder representation of the Lorenz curve will leave the unit square at either or both of its endpoints (0,0) and (1,1) in all but two permissible parameter constellations which are $\gamma=0$ and $\alpha=\beta=1$, $\gamma\leq 1/\sqrt{2}$. In all other cases, the derivative $d\eta/d\pi$ approaches plus or minus infinity at the two endpoints $\pi=0$ and $\pi=\sqrt{2}$, respectively, which is tantamount to a slope of -1 of the associated Lorenz curve at its two endpoints (0,0) and (1,1). Kakwani argued (1980) that this rather general defect will be of little practical importance since the Lorenz curve implied by (1) will normally possess over almost its entire range the required monotonicity. This is hardly much consolation because the essential question is how long it will take the estimated Lorenz curve to reenter the lower triangle of the unit square, and on this account, the order of magnitude involved is too often not negligible. This may not be obvious, but e.g. Jain's compilation is full of such cases. For instance, the estimates for Botswana 1971-72 imply $y<0$ for all $p\varepsilon(0,.75)$ and those for Brazil 1960 (first column) imply $p>1$ for all y in the interval (.75,1).

To avoid this flaw, Rasche et al. proposed using the generalized Pareto Lorenz curve instead,

$$y = \{1-(1-p)^{\alpha}\}^{1/\beta} \quad \text{with} \quad 0<\alpha,\beta\leq 1 \quad , \tag{2}$$

which has all the properties (such as monotonicity and convexity) a Lorenz curve must possess. From the point of view of computational effort involved, it is perhaps comparable to Kakwani and Podder's specification (1), for it requires non-linear estimation but is somewhat easier to integrate.

As an alternative to these two specifications we will also use the functional form

$$y = \gamma p^{1/\alpha} + (1-\gamma)\{1-(1-p)^{\beta}\} \quad \text{with} \quad 0<\alpha,\beta\leq 1 \quad \text{and} \quad 0\leq\gamma\leq 1 \quad , \tag{3}$$

which is another generalization of the Pareto Lorenz curve, as it is simply pieced together as a convex combination of the Pareto Lorenz curve proper and its mirror image across the main diagonal of the unit square. Thus, it is clearly monotone and convex. Like the Rasche et al. specification (2), it requires non-linear estimation techniques, but unlike it, it is easily integrated[5].

2. Retrieving the Lorenz Curve by Interpolation

A priori there is little reason to believe that an actual Lorenz curve should obey one or another simple functional form. And since it is hardly any merit for a functional form to have particularly few parameters, one wonders whether or not to try retrieving the Lorenz curve by interpolation rather than through curve fitting. The simplest way of interpolation would, of course, consist of joining the known data points by linear segments. This is in no way an unusual method of interpolation - e.g. Paukert (1973) used it to calculate the Gini coefficients for his 56 country sample - but it implies identical incomes for all units within each given bracket and thus leads, as is well known, to the lowest Gini coefficient that is compatible with the data. Looking then for a smoother representation of the Lorenz curve we will require that it be continuously differentiable in addition to being monotone and convex. One should be aware, however, that differentiability in contrast to monotonicity and convexity is at best justified on intuitive grounds and constitutes no necessary attribute of a Lorenz curve.

---

5 In particular, integration of (3) does not require any inconvenient recourse to the Beta distribution, as do both the specifications (1) and (2). See Kakwani and Podder (1976) and Rasche et al. (1980).

Of the many conceivable interpolation schemes, we will consider only the quadratic spline function since with this representation differentiability, monotonicity and convexity are particularly easy to control. Let p_i $(i=1,\ldots,n)$ be the cumulative proportions of the population for which the associated values of the Lorenz curve, y_i, are known and define the divided differences

$$d_i = (y_{i+1}-y_i)/(p_{i+1}-p_i) \qquad i=1,\ldots,n-1 \quad ,$$

i.e. the slopes of the linear segments joining the known data points. Now, the quadratic spline function is[6]

$$S(p) = (p_{i+1}-p_i)^{-2}\{y_i(p_{i+1}-p)^2+2a_i(p_{i+1}-p)(p-p_i)+y_{i+1}(p-p_i)^2\} \qquad (4)$$

$$\text{for} \quad p_i \leqq p \leqq p_{i+1} \quad \text{and} \quad i=1,\ldots,n-1$$

with

$$s_{i+1} = 2d_i-s_i \qquad (5.1)$$

$$a_i = y_i+s_i(p_{i+1}-p_i)/2 \quad . \qquad (5.2)$$

Note that the s_i are the the first dervatives of the quadratic spline function $S(p)$ at the given points p_i. From (5.1) it is evident that merely s_1, the slope at the point $p_1=0$, has to be fixed in order to make $S(p)$ unique. Passow (1977, Theorem 2) has shown that given the conditions $d_i \geqq d_{i-1}$ and choosing s_1 to satisfy $0 \leqq s_1 \leqq 2d_1$, the spline function $S(p)$ will be unique and, in particular, monotone. Clearly, these conditions can always be met since all the points on a Lorenz curve must satisfy $d_i \geqq d_{i-1} \geqq 0$ by construction. Passow furthermore proved that $S(p)$ will be convex if the inequalities $0 \leqq s_1 \leqq d_1$, $s_1 \geqq 2d_1-d_2$ and $d_i-2d_{i-1}+d_{i-2} \geqq 0$ for $i=3,\ldots,n-1$ hold. However, these conditions are not overly helpful in the search for a particular s_1 that makes $S(p)$ convex. Also, they can be shown to be unnecessarily strong. We therefore propose using the weaker condition[7]

$$\max(0,b_2,b_4,\ldots) \leqq s_1 \leqq \min(b_1,b_3,b_5,\ldots) \qquad (5.3)$$

$$\text{with } b_1=d_1 \quad \text{and} \quad b_i=b_{i-1}-(-1)^i(d_i-d_{i-1}) \quad \text{for} \quad i=2,\ldots,n-1 \quad ,$$

which in view of (5.1) follows directly from the convexity condition

---

6 It is not difficult to check that the following representation of a quadratic spline function together with the specifications (5.1) and (5.2) is indeed an interpolation and continuously differentiable.

7 I owe this condition to Wilhelm Forst.

(implying $s_{i+1} \geq s_i$) in conjunction with the monotonicity condition (requiring $s_1 \geq 0$).

The key question is whether the permissible range of s_1 according to (5.3) is empty or not, for if it is empty, there is no quadratic spline function such as (4) that is compatible with the data and convex. Nonexistence may occur for two reasons. Either there is an error in the data in that they violate the natural conditions $d_i \geq d_{i-1} \geq 0$, and then there exists, of course, no convex function at all which could accomodate these incorrect data. Or the quadratic spline (4) with its one degree of freedom is too inflexible to interpolate the given data without violating the convexity requirement in some range. Both cases occured when we tried to interpolate Paukert's data by the quadratic spline. If the failure is due to inflexibility, one might think of trying higher order splines. However, that would inevitably require a much more complex effort to control for convexity and, more importantly, this approach may not work either[8]. In the few cases where (4) actually failed, we recurred instead to McAllister and Roulier's "point insertion algorithm", a much simpler and more meaningful method that solves the convexity problem by inserting, if necessary, at most one additional (variable) knot between each pair of data points. As an additional advantage, it permits (and requires) specification of the slopes s_i at the n data points[9]. When these slopes are selected, define the variable knots

$$\bar{p}_i = (s_{i+1}p_{i+1} - s_i p_i + y_i - y_{i+1})/(s_{i+1} - s_i) \tag{6.1}$$

$$\bar{y}_i = (a_i^1(p_{i+1} - \bar{p}_i) + a_i^2(\bar{p}_i - p_i))/(p_{i+1} - p_i) \tag{6.2}$$

$$\text{with} \quad a_i^1 = y_i + s_i(\bar{p}_i - p_i)/2, \quad a_i^2 = y_{i+1} + s_{i+1}(\bar{p}_i - p_{i+1})/2 \quad .$$

McAllister and Roulier's shape-preserving quadratic spline is then given by

$$S^1(p) = (\bar{p}_i - p_i)^{-2}\{y_i(\bar{p}_i - p)^2 + 2a_i^1(\bar{p}_i - p)(p - p_i) + \bar{y}_i(p - p_i)^2\} \tag{7.1}$$

---

8 First of all, one may encounter data which are incompatible with a convex cubic spline even though a convex quadratic spline exists. More importantly, one may in fact construct convex data which require the shape-preserving spline to have an arbitrarily high degree. See McAllister and Roulier (1981).

9 Such information is, at times, available. When it is not, as we assume here, the slopes may be chosen more or less aribtrarily as long as they meet the natural conditions $0 \leq s_1 \leq d_1$, $d_1 \leq s_2 \leq d_2, \ldots, d_{n-1} \leq s_n \leq \infty$. However, whenever $d_i = d_{i+1}$, the Lorenz curve is necessarily linear over the corresponding interval (p_i, p_{i+2}), i.e. $S(p) = y_i + d_i(p - p_i)$ for $p_i \leq p \leq p_{i+2}$. In this case, the slopes at the two endpoints of the linear segment must be identical so that $s_i = s_{i+2} = d_i$ has to be chosen.

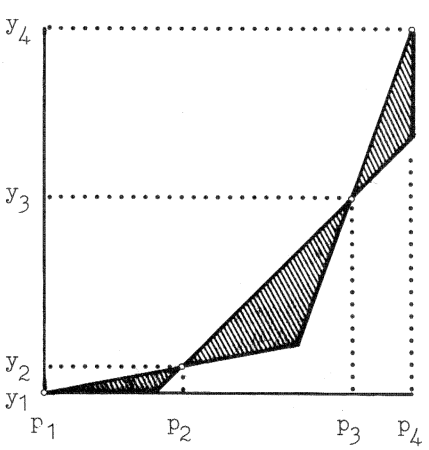
$$S^2(p) = \begin{cases} \text{for } p_i \le p \le \bar{p}_i \quad , \\ (p_{i+1}-\bar{p}_i)^{-2}(\bar{y}_i(p_{i+1}-p)^2+2a_i^2(p_{i+1}-p)(p-\bar{p}_i)+y_{i+1}(p-\bar{p}_i)^2) \\ \text{for } \bar{p}_i \le p \le p_{i+1} \quad . \end{cases} \quad (7.2)$$

It is easy to check that (7) indeed interpolates the data points, observes the specified slopes, and is convex if the data are.

3. Bracketing the Lorenz Curve by Upper and Lower Bounds

Interpolating the known points on the Lorenz curve rather than fitting some simple specification such as (1) through (3) has the obvious advantage that at least the given points will be exactly hit. Yet even the interpolation form, like these simpler specifications, arbitrarily forces some specific curvature (and smoothness) on the Lorenz curve which may or may not come close to the unknown truth. Rather than relying on such a degree of arbitrariness, one may wish to extract from the known data points only that information about the Lorenz curve that can safely be trusted. This may appear to be a modest intention, but as will be demonstrated below, the resulting information is rather precise and very useful when checking the reliability of the other approaches.

Gastwirth (1972) has extensively described how to determine upper and lower bounds for the Lorenz curve, and even though his method is not entirely applicable to our problem, since we do not assume, as he does, to know the slopes of the Lorenz curve in addition to its ordinate values y_i at the n given points p_i, we will merely briefly outline the method. Consider the following figure where n=4 is assumed. Given the four points (y_i,p_i), it is obvious that the true Lorenz curve must lie somewhere inside the shaded band. This construct relies on the assumption



that individual incomes are non-negative and finite, and it only uses the property that units in each interval (p_i, p_{i+1}) earn at least the average income of the previous interval and at most the average income of the following interval. Normally, more than just two inner points of the Lorenz curve will be given and then the associated band will, of course, be substantially narrower and the sharpness of the bounds improve accordingly.

4. The Performance of the Three Approaches

To reach a judgement on the reliability of the three methods outlined in the previous sections, we used Paukert's data for a sample of 56 countries in which, because of an obvious data error, the data for South Africa were replaced by comparable data from Jain (1975).

Consider first the curve fitting method. In view of the fact that Paukert's data encompass just five inner points on each Lorenz curve to which our two- or three-parameter specifications have to be fitted[10], it may come as a surprise that even in terms of goodness of fit all three specifications perform poorly. The Rasche et al. specification (2) with its two parameters clearly gives the worst fit. Calculated on the basis of the entire 56-country sample, it misses the given Lorenz curve points on an average by .0086 and thus almost by a full percentage point. In relative terms, this amounts at times to enormous errors[11]. The Kakwani and Podder specification (1) and the alternative specification proposed in (3), which both have three parameters, give a somewhat better fit in that they miss on an average by only half as much as the Rasche et al. specification. In relative terms this is often still too much. One could now argue that when it comes to extract summary distribution measures these errors tend to cancel out. While this indeed seems to be the case for global measures that rely on the shape and position of the Lorenz curve over its entire range[12], this argument is certainly not

---

10 All three specifications meet the two endpoints (0,0) and (1,1) by construction.

11 In the cases of Morocco and Tunisia, for example, the Rasche et al. specification underestimates the share of the first quintile by 65 and 68 percent, respectively.

12 For instance, all estimates of the Gini concentration ratio based on the specifications (1) and (3) remained within the theoretical bounds suggested by Paukert's data. Only those derived on the basis of the Rasche et al. specification (2) failed twice, namely with respect to Morocco and Tunisia again. It should be noted that the calculation of upper and lower bounds for e.g. the Gini coefficient compatible with the given data is somewhat more involved than the simple construction of the band depicted in the figure. The lower bound is clearly given by the Gini coefficient associated with the linear connection of the given data

valid for supplementary measures such as Lorenz coefficients that may refer to smaller sections of the Lorenz curve. Here the estimates based on the fitted specifications (1) through (3) fell too often outside the range suggested by the theoretical bounds to be of any practical use.

Consider then the method of interpolating the Lorenz curve by a convex quadratic spline function. Goodness of fit is no issue, as the function will meet the given data points by construction. Furthermore, with regard to the resulting Gini concentration ratio or Lorenz coefficients estimates, this approach performs excellently since in all our calculations we did not find a single case where the estimate based on spline interpolation fell outside the range suggested by the associated theoretical upper and lower bounds.

Consider finally the method of calculating upper and lower bounds for either the Lorenz curve or associated distribution measures using parts of or the entire Lorenz curve. Unfortunately, the available data often will not, and in the case of Paukert's data, in fact do not come in a form that permits use of Gastwirth's approach to deriving upper and lower bounds for the Lorenz curve. When only cumulative income shares rather than maximum and minimum incomes for each income group are known, only the weak Lorenz curve property can be used, according to which the incomes within each group cannot be smaller than the previous group's average nor larger than the average of the following group. As a consequence, the resulting bounds are often not very sharp as is indicated by ranges that amounted on the average to about five per cent, however, at times to as much as 30 per cent of the associated lower bound. Nevertheless, these bounds were in numerous cases sharp enough to discredit the estimates on the basis of the curve fitting method as unreliable. At any rate, if they are not very sharp, rather than pretending to be precise, they reveal the lack of precision that goes along with the given data.

5. Summary

We have discussed three different ways ways of retrieving the entire Lorenz curve from a few given points, i.e. the curve fitting method, interpolation by a convex quadratic spline, and calculation of upper and lower bounds, which rests on the intrinsic properties of the Lorenz

---

points and thus easily derived. The upper bound, however, involves finding the piecewise linear and convex Lorenz curve that hits the data points and minimizes the area below subject to the constraint that incomes are non-negative and finite. While the procedure is straightforward, the calculatory details are too tedious to be given here.

curve. In applying these three methods to Paukert's data, it turned out that curve fitting in general and the Rasche et al. specification (2) in particular, are too unreliable to warrant practical application. Both from the computational effort involved and particularly from its impeccable performance in terms of staying within the theoretical bounds, the use of an interpolating convex quadratic spline function proved to be clearly preferable. This approach, however, still suffers from an important defect: the precision of the spline interpolation is more apparent than real since it is inevitably the result of an arbitrary selection of a specific functional form, the choice of which is guided by mathematical simplicity rather than by facts. The most dependable method thus appears to consist of calculating the upper and lower bounds for this alone guarantees that not more information is extracted from the given data than what they actually contain.

In practical application, however, one may often be willing to sacrifice at least some of the reliability for practicability and accordingly one may prefer to work with reasonable point estimates rather than with reliable but unwieldy bounds. The midpoint of the range spanned by the bounds would then constitute a possible candidate for such a point estimate and, indeed, a rather likely one. Suppose this midpoint is actually chosen. It would then make little difference whether one uses these point estimates or the computationally much simpler estimates from the spline interpolation, because in all of the calculations we made, an almost perfect correlation between these two was found.

References

Fortune, J. N., "Income Distribution and Linder's Thesis", Southern Economic Journal, Vol. 46 (1979), 158-167.

Gastwirth, J. L., "The Estimation of the Lorenz Curve and Gini Index", Review of Economics and Statistics, Vol. 54 (1972), 306-316.

Hagerbaumer, J. B., "The Gini Concentration Ratio and the Minor Concentration Ratio: A Two-Parameter Index of Inequality", Review of Economics and Statistics, Vol. 59 (1977), 377-379.

Jain, S., Size Distribution of Income, A Compilation of Data, Washington, D.C.: The World Bank, 1975.

Kakwani, N., "Functional Forms for Estimating the Lorenz Curve: A Reply", Econometrica, Vol. 48 (1980), 1063-1064.

---, and N. Podder, "Efficient Estimation of the Lorenz Curve and Associated Inequality Measures from Grouped Observations", Econometrica, Vol. 44 (1976), 137-148.

Koo, A. Y. C., N. T. Quan, and R. Rasche, "Identification of the Lorenz Curve by Lorenz Coefficient", Weltwirtschaftliches Archiv, Band 117 (1981), 125-135.

---, ---, ---, "Identification of the Lorenz Curve by Lorenz Coefficient: A Reply", Weltwirtschaftliches Archiv, Band 119 (1983), 368-369.

McAllister, D. F., and J. A. Roulier, "An Algorithm for Computing a Shape-Preserving Osculatory Quadratic Spline", ACM Transactions on Mathematical Software, Vol. 7 (1981), 331-347.

Passow, E., "Monotone Quadratic Spline Interpolation", Journal of Approximation Theory, Vol. 19 (1977), 143-147.

Paukert, F., "Income Distribution at Different Levels of Development: A Survey of Evidence", International Labour Review, Vol. 108 (1973), 97-125.

Rasche, R. H., J. Gaffney, A. Y. C. Koo, and N. Obst, "Functional Forms for Estimating the Lorenz Curve", Econometrica, Vol. 48 (1980), 1061-1062.

Thon, D., "Lorenz Curves and Lorenz Coefficients: A Sceptical Note", Weltwirtschaftliches Archiv, Band 119 (1983), 364-367.