Calibrated estimation of a nonparametric income distribution from a few percentiles

Jing Dai
Universität Kassel, Department of Economics, Nora-Platiel-Str. 5, D-34109 Kassel,

Ignacio Moral-Arce
Institute for Fiscal Studies, Ministry of Finance, Madrid, Spain

Stefan Sperlich*
Université de Genève, Département des sciences économiques and Research Center for Statistics, Bd du Pont d'Arve 40, CH-1204 Genève
stefan.sperlich@unige.ch

Abstract

For different welfare studies, often an estimate of the income or consumption distribution is needed even if only few percentiles (e.g. quantiles and quintiles) are available. In the past, simulation methods, interpolation or smoothing methods with strong oversmoothing properties were applied to obtain an idea of the entire distribution. A method for estimation of a convex function based on spline smoothing is proposed for estimating the Lorenz curve from sparse data points. Compared to the currently available methods, the new estimate does not require constrained optimization nor simulations. The use of the functional form for the Lorenz curve enables us to provide a nonparametric density or cumulative distribution function that is consistent with the given percentiles. Further, we can easily derive important welfare measures such as the Gini coefficient for inequality. In the simulation study and a real application with quintile share data on US income, it can be seen that our estimation method performs very well.

Key Words: Income distribution, Lorenz curve, smoothing splines, nonparametric estimation.

Subject Category: A2, E1, D5

1 Introduction

When Sala-I-Martin (2006) in his seminal paper calculated the income distribution of 138 countries he used nonparametric kernel density estimation. Typically, for such an estimation sufficient large data sets have to be available or to be constructed artificially by different data matching, forecasting, or extrapolation (from neighboring countries) methods. Often, however, quantiles, quintiles or even more percentiles are indeed available and probably more reliable than particular (small) samples or own artificial constructs. Sometimes, it is also that for many small areas we have basically the information of percentiles, would like to estimate the distribution for each as well as the total distribution which again respects given percentiles - one might speak here of a scaling problem. Our target is to construct easy-to-calculate analytical estimators based on only few percentiles which allow to construct the distribution functions as well as derivatives like the Lorenz

curve or inequality or poverty measures. The main idea is to use a constraint spline estimator for the Lorenz curve which provides an analytic - since parametric - specification of all the other functions and parameters of interest. This function can be forced to pass through the given percentiles and is therefore by construction calibrated. As all the other constructs are analytical derivatives, not based on simulation or numerical methods, they automatically are calibrated as well. The minimum of information needed are quantiles while there is no upper limit. However, with deciles we get already excellent approximations of the real underlying distribution, and the use of more than centiles does typically not contribute new information.

In case only a few points of the distribution are available instead of microdata sets, how one may retrieve the unknown Lorenz curve then? Different approaches to this task were reported and summarized by Braulke (1988). Unfortunately, they had a poor performance. The method of interpolation by a well-behaving (monotone, convex, differentiable) quadratic spline (Passow, 1977 and Lam, 1990) performs very well as the fitted curve will definitely go through the given data points, and the estimate is surely lying inside the range suggested by the theoretical bounds. Unlike them, the main purpose of the present work is to use a nonparametric estimate of a regression function under certain shape restrictions such as monotonicity and convexity, and based on sparse data points on the Lorenz curve for estimating the entire Lorenz curve, the density, etc. The convex regression function estimate was described in detail in Birke and Dette (2007). It provides good fits to the Lorenz curve of many income distributions, and allows to easily compute Gini's Index and other inequality measures. Moreover, one can derive a explicit income density function.

2 A new estimator for convex functions

Consider the nonparametric regression model

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i, \quad i = 1, \cdots, n, \tag{2.1}$$

where $X_1, \cdots, X_n$ are independent and identically distributed (i.i.d.) random variables with density f and $\epsilon_1, \cdots, \epsilon_i$ are i.i.d. with mean 0, variance 1 and finite fourth moment. Further, assume that the variance function σ^2 is continuous and the density and regression function are three times continuously differentiable. Note that m is strictly convex if and only if its derivative m' is strictly increasing. Suppose for a set of specified population percentile indexed in increasing order ($p_i < p_{i+1}$) and bounded in $[0, 1]$, one has observed the corresponding share of aggregate income y_i. Then, the method can be described as follows:

- Construct a strictly isotonic estimate of the derivative of the regression function $m(p)$. The penalized least squares smoothing spline is used as an unconstrained estimate of the regression function.

- Construct a density estimate

$$\frac{1}{Nh_d}\sum_{i=1}^{N} K_d\Big(\frac{\hat{m}'(i/N)-u}{h_d}\Big) \tag{2.2}$$

from the estimated values $\hat{m}'(i/N)(i=1,\cdots,N)$ of function m'. K_d denotes a symmetric, twice continuously differentiable kernel with compact support on $[-1,1]$ with finite second moment, and h_d the corresponding bandwidth converging to zero with increasing sample size n.

- Referring to the ideas of Dette et al. (2006) for making a function estimate isotonic

$$\hat{\Upsilon}_{h_d} = \frac{1}{Nh_d}\sum_{i=1}^{N}\int_{-\infty}^{t} K_d\Big(\frac{\hat{m}'(i/N)-u}{h_d}\Big)du \tag{2.3}$$

is a consistent estimate of the function $(m')^{-1}$ at the point t. Note that this function estimate is strictly increasing if h_d is sufficiently small. Consequently, its inverse is a strictly isotonic and smooth estimate of the derivative of the function $m(\cdot)$.

- Since the estimate $\hat{\Upsilon}_{h_d}^{-1}$ is strictly increasing (and continuous), the estimated functional

$$\hat{m}_l(p,u_0) = m(u_0) + \int_{u_0}^{p} \hat{\Upsilon}_{h_d}^{-1}(z)dz \tag{2.4}$$

is strictly convex. $\hat{\Upsilon}_{h_d}^{-1}$ is an arbitrary point at $u_0 \in (0,1)$. In our case, the choice of the initial point u_0 is not crucial for the performance.

Different spline estimates are considered in our study while applying Birke and Dette (2007):

- S1: A smoothing spline estimate that minimizes the criterion:

$$\sum_{i=1}^{n}(y_i - m(p_i))^2 + \lambda\int (m'')^2\ , \tag{2.5}$$

implemented in R in the 'mi' package, see the function $sreg()$.

- S2: A so-called penalized spline estimate

$$\sum_{i=1}^{n}(y_i - m(p_i))^2 + \lambda\int (D^r m)^2, \tag{2.6}$$

where D^r is a linear differential operator and r denotes the order of the derivative to be penalized. This is implemented in R in the 'pspline' package, see $smooth.Pspline()$. Due to the sparse data we have to smooth the data using a second order polynomial spline.

- S3: Basically as S2 but now $m(p)$ is a piecewise polynomial of order $2r-1=4$ by setting $r=2.5$.

In the literature, there are some suggestions on choosing and using smoothing splines under certain shape constraints. However, they all propose methods that can hardly be applied on a problem with very sparse data like we face it in this work.

3 Application on US data

While in the complete original paper an extensive simulation study will be presented to understand well the pros and cons of the method, and to also get an idea of the parameter choices, we limit our presentation here to an application. Specifically, we consider here the US household income quintiles of the year 2000 listed in Table 1, based on a micro data sample with $514'779$ observations from the US Census Bureau. These data have been taken because we also have the (almost, see Table) individual information and can therefore compare the method with a kernel estimate based on the full information.

Table 1: Percent share of aggregate income (in US Dollars) received by each fifth US household in 2000 (514779).

Quintile	1	2	3	4	5	mean income
Share of aggregate income	.037	.093	.152	.234	.484	$57'195$

Source: US Census Bureau: Consumer Income Reports (P60-213)

You can see the average household income of residents, and, ranking the households from the poorest to the wealthiest, the fife resulting income quintiles (1 being poorest and 5 being wealthiest), containing each 20% of the population. As can be seen, the poorest 20 percent of the population had roughly 3.7 percent of total income, the next poorest 20 percent of the population had roughly 9.3 percent of total income, etc.

In Figures 1 and 2 are given nonparametric function estimates for the Lorenz curve and income distribution based on the $514,779$ observations, compared to our different estimates which are only based on the quintile information. As can be seen, the estimators based on Birke and Dette's method perform pretty well describing quite well the shape of the truly underlying functions even though we are provided with only 5 data points. Note that, quite importantly, all our new estimators pass through the given quintiles whereas for example the nonparametric estimates using all data do not. This is why we speak of a *calibrated* method. There are three important remarks that should be added: First, the precision rises dramatically with the number of percentiles. Second, for aggregated income above 575'000 USD we have only one data point to estimate the density and Lorenz curve. Although we still meet the correct quintile, it is clear that the right tail can hardly be met without further information. Third, and most importantly, as all important wealth, income and inequality measures (functions as well as indexes) can be derived from each other, our method guarantees that all these measured are consistent to each other as they are based on the same model.

Let us close with one simple example. The Gini coefficient is one commonly accepted measure for inequality of income or wealth. Based on the Lorenz

Figure 1: Estimated Lorenz curves

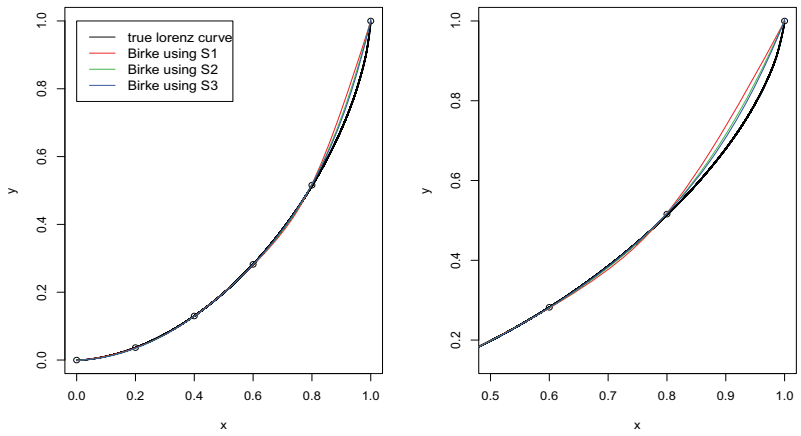
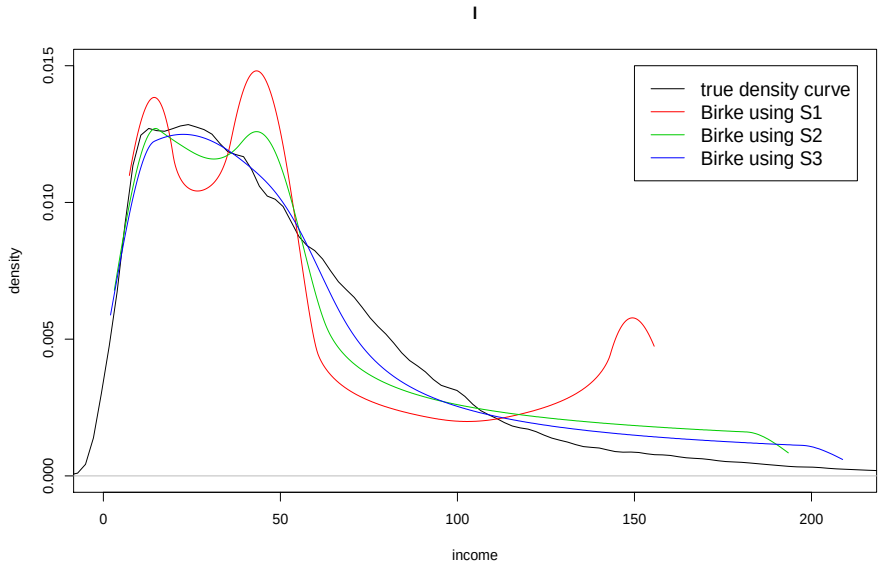


Figure 2: Estimated income distributions. Scale: x-axis 10^3, y-axis: 10^{-3}.



curve the Gini coefficient is defined mathematically as the ratio of the area that lies between the line of equality and the Lorenz curve over the total area under the line of perfect equality, i.e. the 45% line. This ratio can be

Table 2: Gini estimates

census estimate	$S1$	$S2$	$S3$
0.448	0.432	0.437	0.438

determined by $1 - 2\int_0^1 L(X)dX$. Table 2 compares the estimates based on the $514,779$ observations to our three different estimates based on just the quintiles. The resulting Gini estimates confirm that especially the Birke and Dette method using S3 is pretty close to the census estimate.

References

Birke, M. and Dette, H. (2007) "Estimating a convex function in nonparametric regression," *Scandinavian Journal of Statistics*, 34, 384–404.

Braulke, M. (1988) "How to retrieve the Lorenz curve from sparse data," in W. Eichhorn (Ed.), *Measurement in Economics*, Heidelberg, Physica-Verlag, 373-382.

Dette, H., Neumeyer, N. and Pilz, K.F. (2006) "A simple nonparametric estimator of a monotone regression function," *Bernoulli*, 12, 469–490.

Lam, M.H. (1990) "Monotone and Convex Quadratic Spline Interpolation," *Virginia Journal of Science*, 41(1), 3-13.

Passow, E. (1977) "Monotone quadratic spline interpolation," *Journal of Approximation Theory*, 19, 143-147.

Sala-I-Martin, X. (2006) "The world distribution of income: falling poverty and ... convergence, period," *The Quarterly Journal of Economics*, 121(2), 351-397.