A Simple Consistent Nonparametric Estimator of the Lorenz Curve

Yu Yvette Zhang^{*} Ximing Wu[†] Qi Li[‡]

July 29, 2015

Abstract

We propose a nonparametric estimator of the Lorenz curve that satisfies its theoretical properties, including monotonicity and convexity. We adopt a transformation approach that transforms a constrained estimation problem into an unconstrained one, which is estimated nonparametrically. We utilize the splines to facilitate the numerical implementation of our estimator and to provide a parametric representation of the constructed Lorenz curve. We conduct Monte Carlo simulations to demonstrate the superior performance of the proposed estimator. We apply our method to estimate the Lorenz curve of the U.S. household income distribution and calculate the Gini index based on the estimated Lorenz curve.

Keywords: Lorenz Curve; Spline Estimation; Monotonicity; Convexity; Gini Index

---

^{*}Department of Agricultural Economics, Texas A&M University, College Station, TX 77843; email: yzhang@tamu.edu

[†]Department of Agricultural Economics, Texas A&M University, College Station, TX 77843; email: xwu@tamu.edu

[‡]The corresponding author: Department of Economics, Texas A&M University, College Station, TX 77843; email: qi@econmail.tamu.edu

1 Introduction

The Lorenz curve is an important tool of economic analysis. For instance, the Lorenz curve of an income distribution describes the proportion of total income assumed by the bottom u proportion of the population, where $u \in [0, 1]$. The Gini index, a popular measure of economic inequality, is closely related to the Lorenz curve: The value of Gini index equals twice the area between a Lorenz curve and the 45-degree line of a unit square, which corresponds to a completely egalitarian distribution. Lorenz curve is also useful for the testing of stochastic dominance. If Lorenz curve A is everywhere above Lorenz curve B , then the corresponding distribution A stochastically dominates distribution B . For general overviews of the Lorenz curve and its applications in economics, see, e.g., Chotikapanich (2008) and Silber (1999)[Chapters 10 and 11].

Consider a positive random variable Y with a distribution function $F : R^+ \rightarrow [0, 1]$. Throughout this study, we assume that F is strictly monotone with a quantile function $Q(u) = F^{-1}(u)$ and continuously differentiable with a density function $f(y) = F'(y)$. The Lorenz curve can be constructed as follows:

$$L(u) = \frac{\int_0^u Q(t)dt}{\int_0^1 Q(t)dt} = \frac{\int_0^u Q(t)dt}{\mu}, \quad u \in [0, 1],$$

where $\mu = \int y dF(y) < \infty$ is the population mean.

The Lorenz curve has the following properties:

1. $0 \leq L(u) \leq u$ for all $u \in [0, 1]$.
2. $L(0) = 0$ and $L(1) = 1$.
3. $L'(u) > 0$ for all $u \in [0, 1]$.
4. $L''(u) > 0$ for all $u \in [0, 1]$.

In other words, the Lorenz curve resides below the 45-degree of the unit square and is monotonically increasing and strictly convex (except for a completely egalitarian distribution). Throughout the text for simplicity, we shall use convexity in the place of strict convexity. For a detailed treatment of the properties of the Lorenz curve, see e.g. Gastwirth (1972).

Analogous to the empirical distribution function, the empirical Lorenz curve can be constructed from data in a straightforward manner. Consider $\{Y_i\}_{i=1}^n$ an *i.i.d* sample from distribution F and $\{Y_{(i)}\}_{i=1}^n$ its corresponding order statistics arranged in the ascending

order. Define

$$s_i \equiv s(u_i) = \frac{\sum_{j=1}^i Y_{(j)}}{\sum_{j=1}^n Y_{(j)}} \quad \text{and} \quad u_i = \frac{i}{n}, \quad i = 1, \dots, n. \quad (1)$$

Similar to the empirical distribution function, $s(u) \xrightarrow{P} L(u)$ as $n \rightarrow \infty$ uniformly for $u \in [0, 1]$ under very mild regularity conditions. Although the empirical Lorenz curve can be easily constructed, it suffers from the drawbacks that it is not a smooth function and does not allow easy portability (like the empirical distribution function, its parametrization requires the entire sample of data). Luo (2013) presents a kernel-based estimator that yields a smoothed empirical Lorenz curve. However this kernel estimator, like the empirical Lorenz curve, lacks portability.

This study concerns the estimation of the Lorenz curve. There exists in the literature two general approaches of estimation. The parametric approach assumes some parametric forms, often out of mathematical convenience. Parametric Lorenz curves usually satisfy its theoretical properties and converge at root- N rates. Their limitations include the lack of flexibility and possible misspecification errors. Sarabia (2008) reviews parametric estimation of Lorenz curves. In contrast, the nonparametric approach employs smoothing techniques to encourage flexibility (see e.g. Ryu and Slottje (1996) for series-based estimators, Luo (2013) for kernel-based estimators, and Cowell and Victoria-Fester (2008) for semiparametric estimations of the Lorenz curves and robustness issues). This advantage, however, is gained at the price of slower convergence rates and oftentimes violations to the theoretical properties of the Lorenz curve. There is also a small literature that utilizes the Dirichlet distribution in the estimation of Lorenz curves (Chotikapanich and Griffiths (2002) and Hasegawa and Kozumi (2003)).

The goal of this study is to propose a flexible nonparametric estimator of the Lorenz curve that satisfies all its theoretical properties. We adopt a transformation approach, in the spirit of Ramsay (1998), to incorporate the theoretical constraints. This approach transforms a constrained estimation problem into an unconstrained one. We model the unconstrained component using nonparametric methods to gain flexibility. In particular, we use the spline method in our nonparametric modeling. This approach facilitates the numerical implementations of the proposed estimator. In addition, it provides the convenience of portability such that the resultant Lorenz curve, albeit nonparametrically constructed, is parametrized by a small number of parameters.

The rest of the text is organized as follows. Section 2 provides a brief review of the relevant literature. Section 3 presents our alternative estimator and Section 4 proposes a nonparametric estimator of the proposed method. Sections 5 and 6 report Monte Carlo

simulations and an empirical application. The last section concludes the paper. Some technical details are provided in the Appendix.

2 Estimation of Lorenz Curves: A Brief Review

There exists a sizable literature on the estimation of the Lorenz curve. Chotikapanich (2008) and Silber (1999)[Chapters 10 and 11] offer excellent overviews of this literature. In this section, we provide a brief review of some commonly used estimators of the Lorenz curve.

In terms of modeling strategy, there are generally two possibilities: the direct and the indirect approaches. The direct approach models the Lorenz curve, usually via statistical fitting of the empirical Lorenz curves. In contrast, the indirect approach first models some related subject such as the density, distribution or quantile function and then constructs its corresponding Lorenz curve.

Regardless of the modeling strategy, either the parametric or nonparametric method can be used to estimate the Lorenz curve. Kakwani and Podder (1973) propose a parsimonious Lorenz curve as follows:

$$L(u) = u \exp(-h(1-u) + \mu), \quad (2)$$

which provides reasonable fits to some commonly used income distributions. Basmann et al. (1990) generalize model (2) to provide enhanced flexibility. Their model takes the form

$$L(u) = u^{au+b} \exp(-g(1-u^2) - h(1-u) + \mu). \quad (3)$$

Both Models (2) and (3), via a logarithm transformation, can be estimated by OLS based on the empirical Lorenz coordinates (1).

Ryu and Slottje (1996) propose two nonparametric estimators of the Lorenz curve. Their first estimator takes the indirect approach. A two step procedure is suggested. First, the quantile function is approximated by a J -th degree exponential polynomial series:

$$Q(u) \approx \exp\left(\sum_{j=0}^J a_j u^j\right). \quad (4)$$

This approach is equivalent to approximating the logarithm of a quantile function by polynomials and can be easily implemented by OLS. Denote the estimated coefficients by $\hat{a} =$

$(\hat{a}_0, \dots, \hat{a}_J)$. The Lorenz curve is then constructed as

$$\hat{L}(u_i) = \frac{1}{n\hat{\mu}} \sum_{k=1}^i \exp \left(\sum_{j=0}^J \hat{a}_j u_k^j \right), \quad i = 1, \dots, n, \quad (5)$$

where $\hat{\mu}$ is the sample average.

Ryu and Slottje (1996) also propose a direct estimator of the Lorenz curve based on the Bernstein polynomials. This estimator is given by

$$\hat{L}(u) = \sum_{j=0}^J L(j/J) \binom{J}{j} u^j (1-u)^{J-j}. \quad (6)$$

In practice, the unknown Lorenz curve $L(j/J)$ is replaced by its empirical counterpart given in (1).

By construction, the direct estimators (2) and (3) satisfy all theoretical properties of the Lorenz curve. They are, however, vulnerable to misspecification errors because of their parametric nature. In contrast, the two nonparametric estimators converge to the true Lorenz curve asymptotically under some mild regularity conditions, but they are not without their limitations. The exponential polynomial estimator (5) does not guarantee convexity because the exponential polynomial approximation to a quantile function, given in (4), is not necessarily monotonically increasing. Furthermore, it is based on the modeling of the quantile function; an optimal approximation to the quantile function may not be optimal for the Lorenz curve. Lastly, the Bernstein estimator (6) enjoys favorable theoretical properties: it preserves the properties of $L(\cdot)$ such as monotonicity and convexity (see Ryu and Slottje (1996) for details). We note that, however, this estimator essentially interpolates data and suffers from a rather slow convergence rate. Our experiments indicate that a considerably large number of parameters are required to obtain performance comparable to that of other estimators with a more economical parametrization.

3 An Alternative Estimator

Aiming to avoid the limitations of the existing methods discussed above, in this section we propose an alternative estimator that: (i) estimates the Lorenz curve directly; (ii) allows flexible functional forms; and (iii) complies with all theoretical properties of the Lorenz curve. The key innovation of our estimator is to employ integral transformations to transform a constrained problem into an unconstrained one. Subsequently, nonparametric method is

applied to the unconstrained component to obtain flexibility.

Our estimator is inspired by the monotone spline estimator of Ramsay (1998). Suppose that one is interested in modeling a strictly monotonic relationship $y = g(x)$ with $g'(x) > 0$ for $x \in R^+$. Ramsay (1998) suggests the following parametrization

$$g(x) = \int_0^x \exp(r(t))dt.$$

It follows that $g'(x) = \exp(r(x)) > 0$. One advantage of Ramsay (1998)'s approach is that the resultant estimates satisfy the monotonicity constrain globally. This approach has been used to model monotonic relationships implied by economic theories; see, e.g., Zhang et al. (2011) and Liu et al. (2015) on the modeling of monotone bidding processes of auctions.

To further impose convexity, we propose the following model:

$$g(x) = \int_0^x \exp\left(\int_0^s r(t)dt\right)ds.$$

We then have

$$g'(x) = \exp\left(\int_0^x r(t)dt\right) > 0, \quad g''(x) = g'(x)r(x).$$

Thus for convexity, we can choose $r(\cdot)$ such that $r(x) > 0$ for all x .

We therefore advocate the following parameterization of a Lorenz curve:

$$L(u) = \frac{\int_0^u \exp\left(\int_0^s r(t)dt\right)ds}{\int_0^1 \exp\left(\int_0^s r(t)dt\right)ds}, \quad u \in [0, 1], \quad (7)$$

with

$$r(\cdot) : [0, 1] \rightarrow R^+,$$

where the denominator ensures that $L(1) = 1$. It can be verified that all four properties of the Lorenz curve listed in Section 1 are satisfied.

We next suggest an economic interpretation of $r(\cdot)$ related to income mobility, which is inspired by Basmann et al. (1991). Let $E(u)$ designate the logarithmic derivative of $L(u)$, i.e.,

$$E(u) = uL'(u)/L(u).$$

Basmann et al. (1991) note that $E(u)$ satisfies the Bernoulli differential equation

$$uE'(u) - \{I(u) + 1\}E(u) + E^2(u) = 0,$$

where $I(u)$ is the logarithmic derivative of $L'(u)$. It follows that

$$I(u) = uE'(u)/E(u) + \{E(u) - 1\}.$$

Basmann et al. (1991) suggest that $I(u)$ reflects the relative difficulty in moving up the income graduation (as captured by the Lorenz curve) and therefore can be viewed as a ‘mobility function’. For the proposed estimator, note that

$$r(u) = L''(u)/L'(u). \quad (8)$$

Thus $r(\cdot)$ is the relative curvature of the Lorenz curve. In the spirit of Basmann et al. (1991), r can be interpreted as a mobility function, or more precisely an ‘immobility function’ of the underlying distribution. By construction, it is strictly positive on $u \in [0, 1]$. Therefore, our estimation strategy has an interesting economic interpretation in that we transform the modeling of Lorenz curve to that of its associated mobility function.

Below we plot the density and mobility function (8) of the standard exponential and log-normal distributions. The exponential density is infinite at zero and downward sloping on its entire range. Its mobility function is also infinite at origin and descends quickly. Its slope, however, eventually becomes positive at the right end of the distribution. Intuitively, it is most difficult to escape extreme poverty under an exponential distribution that peaks at origin. The difficulty of moving up the distribution then decreases along the distribution until one gets to the right tail of the distribution, where it again becomes difficult to further move up. In contrast, the log-normal density is finite at origin; it increases until it peaks at $\exp(\mu - \sigma^2)$ and then decreases. Interestingly, its mobility function has a symmetric U shape, which tends to infinity at both ends of the distribution. The symmetry of its mobility function is consistent with the fact that the log-normal distribution is Lorenz symmetric in the sense that it is symmetric about the anti-diagonal of the unit square.

4 Nonparametric Estimation

We now proceed to the statistical estimation of the proposed model. To ensure convexity, we further write $r(t) = h(m(t))$, where $h(\cdot) : R \rightarrow R^+$ and $m(\cdot) : R \rightarrow R$. The parametrization (7) thus transforms a constrained problem (modeling of $g(\cdot)$) into an unconstrained one (modeling of $m(\cdot)$), enforcing the convexity of the thus constructed Lorenz curve.

To obtain flexibility, we model $m(\cdot)$ nonparametrically. Kernel smoothing and spline approximation are two commonly used nonparametric methods. In this study we adopt the

Figure 1: Density and mobility functions of the standard exponential distribution (top) and the standard log-normal distribution (bottom)

spline estimator out of the consideration of portability: despite its nonparametric nature, the spline estimator admits a parametric representation which facilitates subsequent model manipulation and analysis, especially when sample size is large. In addition, our estimator (7) entails numerical evaluation of a nested integration. It is computationally expensive to implement this estimator if $m(\cdot)$ is modeled by a kernel smoother.

The spline estimation is essentially a series estimation with piecewise polynomial basis functions. For instance the truncated power series, a commonly used spline basis, are given by

$$1, t, \dots, t^k, (t - z_1)_+^k, \dots, (t - z_J)_+^k,$$

where $t_+ = \max(t, 0)$ and $z = \{z_1 < \dots < z_J\} \in \mathcal{T}$, the support of t , are the so-called *knots* of spline functions. A spline approximation to a smooth function, $m(t) : \mathcal{T} \rightarrow R$, consists of a linear combination of spline basis:

$$m_k(t; z) = \beta_0 + \beta_1 t + \dots + \beta_k t^k + \sum_{j=1}^J b_j (t - z_j)_+^k. \quad (9)$$

Generally the k -th order global polynomials determine the overall shape of the curve, while the piecewise polynomials characterize local deviations from the ‘baseline’ curve and thus facilitate flexibility. One important advantage of the spline basis relative to the power series is its robustness against potentially untamed oscillations associated with high order power series. Spline estimators with a small k and a relatively large J are customarily used to achieve both robustness and flexibility.

Given a spline approximation to $m(\cdot)$, next we need to specify $h(\cdot)$ such that $h(m(t)) > 0$ to ensure the convexity of the resultant Lorenz curve. In principle, any specification satisfying $h(t) > 0$ for all t suffices, such as $h(t) = \exp(t)$, $h(t) = t^2$ or $h(t) = |t|$. For the proposed estimator (7), we advocate the quadratic specification $h(t) = t^2$. Since the square of (piecewise) power series remains power series, it follows that the inner integration of (7) admits an analytical solution and thus avoids nested integration by numerical methods, which is difficult and numerically rather expensive. We provide in the Appendix an algorithm to numerically evaluate

$$s(u; \theta) \equiv \int_0^u \exp \left(\int_0^s m_k^2(t; z) dt \right) ds, \quad (10)$$

where $\theta = (\beta_0, \dots, \beta_k, b_1, \dots, b_J)$ and $m_k(\cdot; z)$ is given by (9). We stress that the choice of quadratic specification is based on computation consideration. We have experimented with the exponential specification $h(t) = \exp(t)$. The results are essentially identical to those from quadratic $h(t)$. However the computation time increases substantially because it now

involves expensive nested numerical integration in iterations.

Our estimator of Lorenz curve is then defined by the following minimization problem:

$$\min_{\theta} \sum_{i=1}^n (s_i - s(u_i; \theta))^2,$$

where s_i , given in (1), is the empirical Lorenz curve evaluated at $u_i, i = 1, \dots, n$. It turns out that the parametrization (10) considerably simplifies the implementation of the proposed estimator, which can be solved by the method of nonlinear least squares. As is suggested by a referee, one can use alternative objective function such as the minimum absolute deviation, which is more robust to possible outliers. This alternative, however, significantly complicates the estimation and inference. Nonetheless, we report in the next section simulation results in terms of both sum of residual squares and sum of absolute residuals. The proposed estimator dominates its competitors according to either criterion.

The practical implementation of spline estimations entails the selection of its polynomial order k and a vector of knots z . For a general treatment of spline specification and estimation, see, e.g., Ruppert et al. (2003). Low ordered spline basis, e.g. $k = 1, 2$ or 3 , is often used. It is known that a k -th order spline series is $(k - 1)$ times differentiable. Our estimator is at least twice differentiable regardless of the smoothness of $r(\cdot)$ due to the presence of double integration in its construction. For simplicity, in our simulations and empirical application below we use the linear splines.

Equally-spaced knots or equal-percentile knots are two commonly used principles of knot placement. In our case, the Lorenz curve depicts the relationship between cumulative income share, s , against population share, u , which is uniform on $[0, 1]$. Therefore the two principles are equivalent. The number of knots can be determined by data driven methods that balance the trade-off between the goodness of fit and model parsimony. For instance, one can use the information criterion (such as the Akaike Information Criterion or Bayesian Information Criterion) or the method of cross validation.

We conclude this section with a discussion of the covariance matrix of Lorenz curve. Similar to the empirical CDF's, the empirical Lorenz curves are not independently distributed. Beach and Davidson (1983) derive the asymptotic covariance matrix of a vector of Lorenz curve coordinates. Let $s = (s_1, \dots, s_n)'$ be a vector of Lorenz curve associated with accumulated income shares $u = (u_1 < u_2 < \dots < u_n)' \in (0, 1)$. Denote the corresponding quantiles by $\xi_i = \min\{x : u_i \leq F(x)\}, i = 1, \dots, n$. Define the truncated mean $\mu_i = E[x|x \leq \xi_i]$, the corresponding Lorenz curves can be written as $s_i = u_i \mu_i / \mu, i = 1, \dots, n$.

Next let $u_{n+1} = 1$, it follows that $\mu_{n+1} = \mu$. Define

$$\nu = (u_1\mu_1 \dots, u_{n+1}\mu_{n+1})'.$$

The covariance matrix of ν , denoted by Ω , is then given by

$$\Omega_{ij} = u_i \left\{ \sigma_i^2 + (1 - u_j)(\xi_i - \mu_i)(\xi_j - \mu_j) + (\xi_i - \mu_i)(\mu_j - \mu_i) \right\}, 1 \leq i \leq j \leq n + 1,$$

where $\sigma_i^2 = E[(x - \mu_i)^2 | x \leq \xi_i], i = 1, \dots, n + 1$. Define $G_1 = \text{diag}(1/\mu)$, a n -dimensional diagonal matrix and $G_2 = (-u_1\mu_1/\mu^2, \dots, -u_n\mu_n/\mu^2)'$. Denote by $G = [G_1 | G_2]$ the $n \times (n+1)$ dimensional gradient of s with respect to ν . We then have the covariance matrix of s given by

$$V = G\Omega G'.$$

In practice, we replace V with its sample analog. In particular, we set, for $i = 1, \dots, n$

$$u_i = i/n, \hat{\xi}_i = Y_{(i)},$$

$$\hat{\mu}_i = \frac{1}{i} \sum_{j=1}^i Y_{(j)}, \hat{\sigma}_i^2 = \frac{1}{i} \sum_{j=1}^i (Y_{(j)} - \hat{\mu}_i)^2,$$

and $\hat{V}$ is calculated accordingly.

Our estimator is then given by the following weighted quadratic objective function

$$\min_{\theta} (s - s(u; \theta))' \hat{V}^{-1} (s - s(u; \theta)).$$

Lastly we advocate regulating $\hat{V}$ to avoid numerical instability due to sampling variation. For a general treatment of regulated or penalized spline estimations, see e.g. Ruppert et al. (2003). In the simulations below, we use instead $\tilde{V} = \hat{V} + I_n/\sqrt{n}$ as the weighting matrix, which is found to provide satisfactory performance. Here I_n is an identity matrix of dimension n .

5 Simulations

We use Monte Carlo experiments to explore the numerical performance of the proposed shape-restricted spline estimator (S-Spline hereafter). Following Ryu and Slottje (1996), we consider three distributions that are commonly used to model income distributions: the Log-normal distribution, the Gamma distribution and the Generalized Beta distribution of

type II (GB2), whose density function is given by

$$f(x) = \frac{a(x/b)^{ap-1}}{bB(p, q)(1 + (x/b))^{p+q}}, x > 0.$$

In particular, we consider the standard Log-normal distribution, a Gamma distribution with the shape parameter being 3, and a GB2 distribution with $a = 5.0573, b = 13.5815, p = 0.2961, q = 0.6708$ (as given by McDonald (1984)).

For comparison, we also consider the four commonly used estimators of the Lorenz curve described in Section 2. The estimator by Basmann et al. (1990) (BHJS hereafter) is parametrized by five parameters. To facilitate comparison, we restrict the number of parameters for all nonparametric estimators to be the same. In particular, we set the order of the exponential polynomial estimator (Exp-Poly hereafter) and the Bernstein estimator (Bern hereafter) to be five, and use a linear spline basis with knots $z = (1/4, 1/2, 3/4)$ for S-Spline. The estimator by Kakwani and Podder (1973) (KP hereafter) is a special case of BHJS with two parameters and is included for completeness.

We consider three sample sizes: $n = 50, 100$ and 200 ; each experiment is repeated 1,000 times. We gauge the estimation performance by the Mean Square Error (MSE) evaluated at a 99-point equally-spaced grid $(0.01, 0.02, \dots, 0.99)$:

$$\text{MSE} = \frac{1}{99} \sum_{i=1}^{99} (s_0(i/100) - \hat{s}(i/100))^2,$$

where $s_0(\cdot)$ is the true Lorenz curve calculated from the underlying distribution and $\hat{s}(\cdot)$ is its estimate. Table 1 reports the average and median MSE across all repetitions for each experiment. In all experiments, the proposed estimator outperforms the others, often by considerable margins. Some remarks are in order.

- KP, a restricted case of BHJS, is dominated by BHJS. BHJS provides the second best overall performance, plausibly due to its flexible parametrization and theoretical consistency. Nevertheless, it is strictly dominated by S-Spline in all experiments.
- Exp-Poly provides satisfactory results for the log-normal distribution, but is severely biased for the Gamma and GB2 distributions. In all three cases, the performance does not seem to improve with sample size. We conjecture its subpar performance might be associated with the following reasons: (i) this estimator approximates the underlying quantile function rather than the Lorenz curve; (ii) the approximation is based on (exponential) power series, which suffers from possible oscillations and can be sensitive

to extreme values; (iii) the resultant Lorenz curve estimate is not necessarily convex.

- Despite its appealing theoretical properties, Bern suffers from a rather slow convergence rate, as is evident in our simulations. In contrast, S-Spline not only provides better results, but also converges faster. For instance, under the log-normal distributions, the average MSE of Bern improves from 7.01 with $n = 50$ to 5.99 with $n = 200$, by 15%; while that of S-Spline improves from 1.54 to 0.41 by 73%.

Table 1: Simulation results (KP: Kakwani and Podder (1973) estimator; BHJS: Basmann et al. (1990) estimator; Exp-Poly: exponential polynomial estimator of Ryu and Slottje (1996); Bern: Bernstein estimator of Ryu and Slottje (1996); S-Spline: shape-restricted spline estimator of the current study)

		KP	BHJS	Exp-Poly	Bern	S-Spline
Log-normal						
$n = 50$	Mean MSE	6.35	2.55	1.64	7.01	1.54
	Median MSE	5.80	1.87	1.62	6.80	0.81
$n = 100$	Mean MSE	6.10	2.21	1.64	6.37	0.84
	Median MSE	5.77	1.90	1.65	6.15	0.46
$n = 200$	Mean MSE	6.04	1.99	1.66	5.99	0.41
	Median MSE	5.91	1.81	1.66	5.97	0.22
Gamma						
$n = 50$	Mean MSE	1.08	0.50	6.95	1.66	0.33
	Median MSE	0.64	0.30	6.85	1.52	0.19
$n = 100$	Mean MSE	0.88	0.34	6.92	1.47	0.16
	Median MSE	0.61	0.21	6.87	1.39	0.09
$n = 200$	Mean MSE	0.77	0.26	6.98	1.40	0.09
	Median MSE	0.62	0.17	6.99	1.36	0.06
GB2						
$n = 50$	Mean MSE	2.74	1.14	5.73	2.43	0.73
	Median MSE	1.71	0.69	5.59	2.23	0.44
$n = 100$	Mean MSE	2.34	0.89	5.70	2.13	0.41
	Median MSE	1.61	0.55	5.58	2.01	0.25
$n = 200$	Mean MSE	2.15	0.78	5.64	1.99	0.23
	Median MSE	1.76	0.52	5.52	1.92	0.14

Note: All figures have been multiplied by 1,000 to improve readability.

Since Exp-Poly, Bern and S-Spline are nonparametric estimators, we next explore their performance under data-driven specification of tuning parameters. In particular, we use the Bayesian Information Criterion (BIC) to select the order of polynomials for Exp-Poly and Bern and the number of knots for S-Spline. The simulation results are reported in Table

2. In addition to the MSE's, we also report the median number of parameters selected by each estimator. The results suggest that all three estimators perform better under data-driven selection of tuning parameters than under fixed specification as is reported in Table 1. Again, S-Spline is seen to provide the best overall performance and clearly dominates the other two estimators under small or modest sample size. Only modest improvement is observed for Exp-Poly, suggesting its rigidity. In contrast, Bern improves substantially. But this improvement comes at the price of a considerably larger number of parameters; for instance, the median number of parameters is 12 for the experiments on log-normal distributions when $n = 50$. As is discussed above, this is because Bern effectively interpolates the data.

6 Empirical Example

In this section we apply the proposed estimator of Lorenz curve to a real income dataset. Our data consist of observations of U.S. household income, extracted from the Current Population Survey March Supplement of Year 2013. We include only households with positive income, resulting in a sample of 200,452 households. We construct the empirical Lorenz curve of this sample and use 99 points of this curve, evaluated at $(0.01, 0.02, \dots, 0.99)$ of the population share, in our estimation. We also consider the four other estimators and use the same specifications as in the previous section. Since KP is clearly dominated by BHJS, we do not report its results in this application. The estimated Lorenz curves are plotted in Figure 2, along with the underlying empirical Lorenz curve. The graphs suggest that S-Spline provides a superior fit relative to its competitors.

We next report in the top panel of Table 3 some summary statistics of the estimation error $e_i = s_0(i/100) - \hat{s}(i/100)$, $i = 1, \dots, 99$, with s_0 and $\hat{s}$ being the empirical Lorenz curve and its estimate. The results confirm the superior performance of S-Spline demonstrated in Figure 2. Lastly, we compare the Gini index calculated from the full sample of more than 200,000 observations to those obtained from the estimated Lorenz curves. The results are reported in the last row of Table 3. The sample Gini is 0.4419 and that based on the Lorenz curve with S-Spline is 0.4440. In contrast, all other estimates show considerable biases.

7 Concluding Remarks

In this study we have proposed an estimator of the Lorenz curve that is both flexible and consistent with all theoretical properties of the Lorenz curve. The proposed estimator is easy

Figure 2: Estimated and Empirical Lorenz Curve

Table 2: Simulation results (Exp-Poly: exponential polynomial estimator of Ryu and Slottje (1996); Bern: Bernstein estimator of Ryu and Slottje (1996); S-Spline: shape-restricted spline estimator of the current study; K : number of parameters)

		Exp-Poly	Bern	S-Spline
Log-normal				
$n = 50$	Mean MSE	1.23	2.90	1.47
	Median MSE	1.13	1.61	0.80
	Median K	3	12	3
$n = 100$	Mean MSE	1.02	1.22	0.77
	Median MSE	0.99	0.62	0.43
	Median K	3	19	3
$n = 200$	Mean MSE	0.86	0.50	0.38
	Median MSE	0.86	0.19	0.22
	Median K	3	17	4
Gamma				
$n = 50$	Mean MSE	6.52	0.63	0.31
	Median MSE	6.32	0.32	0.19
	Median K	3	12	3
$n = 100$	Mean MSE	6.23	0.29	0.19
	Median MSE	6.25	0.09	0.09
	Median K	3	20	4
$n = 200$	Mean MSE	6.20	0.13	0.09
	Median MSE	6.12	0.03	0.05
	Median K	3	15	4
GB2				
$n = 50$	Mean MSE	4.70	0.98	0.62
	Median MSE	4.52	0.33	0.34
	Median K	3	12	3
$n = 100$	Mean MSE	4.59	0.45	0.33
	Median MSE	4.59	0.19	0.18
	Median K	3	17	3
$n = 200$	Mean MSE	4.54	0.18	0.28
	Median MSE	4.52	0.08	0.14
	Median K	3	15	4

Note: MSE's multiplied by 1,000 to improve readability.

to implement and satisfies the theoretical constraints globally. Our Monte Carlo simulations and empirical example demonstrate excellent performance of the estimator. We expect the new estimator to be a useful addition to the toolbox of researchers in the areas of income distribution, inequality and welfare analysis. We conjecture that statistical inference

Table 3: Estimation results

	sample	BHJS	Exp-Poly	Bern	S-Spline
Mean of e_i^2		.0018	.0025	.0034	.0012
Mean of $ e_i $		.0307	.0450	.0505	.0071
Max. of $ e_i $		.0895	.0736	.0996	.0438
Gini	.4419	.3710	.5028	.3744	.4440

Note: $e_i = s_0(i/100) - \hat{s}(i/100), i = 1, \dots, 99$.

can be conducted along the line of nonlinear least squares or nonlinear penalized spline estimation in conjunction with the results by Beach and Davidson (1983). Another possibility is the likelihood approach adopted by Chotikapanich and Griffiths (2002) and Hasegawa and Kozumi (2003). Formal inference with the Lorenz curves using the proposed estimator will be a topic of interest for future research.

References

- Basmann, R., Hayes, K., Johnson, J., and Slottje, D. (1990), “A General Functional Form for Approximating the Lorenz Curve,” *Journal of Econometrics*, 43, 77–90.
- Basmann, R., Hayes, K., and Slottje, D. (1991), “The Lorenz Curve and the Mobility Function,” *Economics Letters*, 35, 105–111.
- Beach, C. and Davidson, R. (1983), “Distribution-Free Statistical Inference with Lorenz Curves and Income Shares,” *Review of Economic Studies*, 50, 723–735.
- Chotikapanich, D. (2008), *Modeling Income Distributions and Lorenz Curves*, New York: Springer.
- Chotikapanich, D. and Griffiths, W. (2002), “Estimating Lorenz Curves Using a Dirichlet Distribution,” *Journal of Business and Economic Statistics*, 20, 290–295.
- Cowell, F. and Victoria-Fester, M. (2008), “Modelling Lorenz Curves: Robust and Semiparametric Issues,” in *Modeling Income Distributions and Lorenz Curves*, ed. Chotikapanich, D., New York: Springer.
- Gastwirth, J. (1972), “The Estimation of the Lorenz Curve and Gini Index,” *The Review of Economics and Statistics*, 54, 306–316.

- Hasegawa, H. and Kozumi, H. (2003), “Estimation of Lorenz Curves: a Bayesian Nonparametric Approach,” *Journal of Econometrics*, 115, 277–291.
- Kakwani, N. and Podder, N. (1973), “On the Estimation of Lorenz Curve from Grouped Observations,” *International Economic Review*, 14, 278–292.
- Liu, W., Zhang, Y., and Li, Q. (2015), “A Semiparametric Varying-Coefficient Model of Monotone Auction Bidding Processes,” *Empirical Economics*, 48, 313–335.
- Luo, S. (2013), “New Non-Parametric Methods for Income Distributions,” Dissertation, Georgia State University.
- McDonald, D. (1984), “Some Generalized Functions for the Size Distribution of Income,” *Econometrica*, 52, 647–665.
- Ramsay, J. (1998), “Estimating Smooth Monotone Functions,” *Journal of the Royal Statistical Society: Series B*, 60, 365–375.
- Ruppert, D., Wand, W., and Carroll, R. (2003), *Semiparametric Regression*, Cambridge University Press.
- Ryu, H. and Slottje, D. (1996), “Two Flexible Functional Form Approaches for Approximating the Lorenz Curve,” *Journal of Econometrics*, 72, 251–274.
- Sarabia, J. (2008), “Parametric Lorenz Curve: Models and Applications,” in *Modeling Income Distributions and Lorenz Curves*, ed. Chotikapanich, D., New York: Springer.
- Silber, J. (1999), *Handbook of Income Inequality Measurement*, New York: Springer.
- Zhang, Y., Gu, J., and Li, Q. (2011), “Nonparametric Panel Estimation of Online Auction Process,” *Empirical Economics*, 40, 51–68.

Appendix

In this Appendix we present an algorithm to numerically evaluate the integration (10). An R-code implementing the estimation procedure is available from the authors upon request. For simplicity, we consider the case of linear splines with

$$m_1(t; z) = \beta_0 + \beta_1 t + \sum_{j=1}^J b_j (t - z_j)_+. \quad (11)$$

Generalizations to higher order splines are straightforward.

Define $(\alpha_0 = \beta_0, a_0 = \beta_1)$ and $(\alpha_j = \beta_0 - \sum_{i=1}^j b_i z_i, a_j = \beta_1 + \sum_{i=1}^j b_i)$ for $j = 1, \dots, J$. It follows that

$$m_1(t; z) = \begin{cases} \alpha_0 + a_0 t, & \text{if } 0 \leq t \leq z_1 \\ \alpha_1 + a_1 t, & \text{if } z_1 < t \leq z_2 \\ \vdots & \\ \alpha_J + a_J t, & \text{if } z_J < t \leq 1. \end{cases}$$

Define $h(\alpha, a; s_1, s_2) = \int_{s_1}^{s_2} (\alpha + at)^2 dt = (\alpha^2 t + \alpha a t^2 + \frac{a^2}{3} t^3)|_{s_1}^{s_2}$. Let $z_0=0$ and $h_0 = h(\alpha_0, a_0; z_0, z_1)$, $h_1 = h(\alpha_1, a_1; z_1, z_2), \dots, h_J = h(\alpha_J, a_J; z_J, 1)$. Next define $J(s) = \max\{j : z_j \leq s, j = 1, \dots, J\}$. We then have

$$\int_0^s m_1^2(t; z) dt = \sum_{j=0}^{J(s)-1} h_j + h(\alpha_{J(s)}, a_{J(s)}; z_{J(s)}, s) \equiv H_{J(s)-1} + h(\alpha_{J(s)}, a_{J(s)}; z_{J(s)}, s).$$

Finally for an arbitrary $u \in [0, 1]$, we have

$$\begin{aligned} & \int_0^u \left\{ \exp \left(\int_0^s m_1^2(t; z) dt \right) \right\} ds \\ &= \sum_{j=0}^{J(u)-1} \int_{z_j}^{z_{j+1}} \exp \left(\sum_{i=0}^j H_i \right) ds + \int_{z_{J(u)}}^u \exp \left(\sum_{i=0}^{J(u)} H_i + h(\alpha_{J(u)}, a_{J(u)}; z_{J(u)}, s) \right) ds \\ &= \sum_{j=0}^{J(u)-1} (z_{j+1} - z_j) \exp \left(\sum_{i=0}^j H_i \right) + \exp \left(\sum_{i=0}^{J(u)} H_i \right) \int_{z_{J(u)}}^u \exp (h(\alpha_{J(u)}, a_{J(u)}; z_{J(u)}, s)) ds. \end{aligned}$$

Although direct evaluation of *nested* integration, as required in (7), is difficult and computationally expensive, we have shown that our procedure only requires numerical evaluation of the following *single layer* integration

$$\int_{z^*}^u \exp(h(s)) ds,$$

where z^* is the largest element of the knots that is smaller than u and $h(\cdot)$ is a simple power series. This quantity can be easily and accurately evaluated by a standard numerical integration routine.