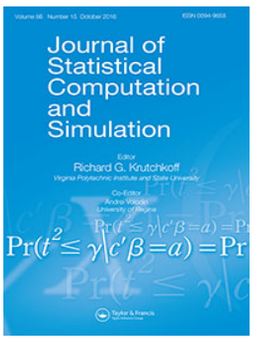

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Estimation of a probability density function using interval aggregated data

Jianhua Z. Huang, Xueying Wang, Ximing Wu & Lan Zhou

To cite this article: Jianhua Z. Huang, Xueying Wang, Ximing Wu & Lan Zhou (2016) Estimation of a probability density function using interval aggregated data, Journal of Statistical Computation and Simulation, 86:15, 3093-3105, DOI: 10.1080/00949655.2016.1150481

To link to this article: https://doi.org/10.1080/00949655.2016.1150481

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2016
VOL. 86, NO. 15, 3093–3105
http://dx.doi.org/10.1080/00949655.2016.1150481

Taylor & Francis
Taylor & Francis Group

Estimation of a probability density function using interval aggregated data

Jianhua Z. Huang[a], Xueying Wang[b], Ximing Wu[c] and Lan Zhou[a]

[a]Department of Statistics, Texas A&M University, College Station, TX, USA; [b]Department of Mathematics and Statistics, Washington State University, Pullman, WA, USA; [c]Department of Agriculture Economics, Texas A&M University, College Station, TX, USA

ABSTRACT

In economics and government statistics, aggregated data instead of individual level data are usually reported for data confidentiality and for simplicity. In this paper we develop a method of flexibly estimating the probability density function of the population using aggregated data obtained as group averages when individual level data are grouped according to quantile limits. The kernel density estimator has been commonly applied to such data without taking into account the data aggregation process and has been shown to perform poorly. Our method models the quantile function as an integral of the exponential of a spline function and deduces the density function from the quantile function. We match the aggregated data to their theoretical counterpart using least squares, and regularize the estimation by using the squared second derivatives of the density function as the penalty function. A computational algorithm is developed to implement the method. Application to simulated data and US household income survey data show that our penalized spline estimator can accurately recover the density function of the underlying population while the common use of kernel density estimation is severely biased. The method is applied to study the dynamic of China's urban income distribution using published interval aggregated data of 1985–2010.

ARTICLE HISTORY
Received 5 September 2015
Accepted 1 February 2016

KEYWORDS
Density estimation; Gini index; growth incidence curve; income disparity; Lorenz curve; penalized splines

1. Introduction

Aggregated data in the form of summary statistics by intervals are prevalent. Statistical agencies from many countries report the average income of different income groups according to the quantiles of income distributions. The underlying individual data are often not available for the reason of data confidentiality (e.g. income or wealth information) or missing (e.g. historical data). For example, the Chinese Statistical Yearbooks report the average incomes of eight income quantiles; The Penn World Tables (maintained by the Center for International Comparisons at the University of Pennsylvania) provide information on production, income and price for more than 150 countries, many of their income data are summarized by intervals. In either case, the underlying individual data are not available to the public. Sometimes aggregated data are the only data available to researchers. Econometricians have been using such data in economic studies; see, for example, [1–8].

Due to the lack of a better terminology, we refer to summary statistics organized as intervals by interval aggregated data. This type of data provides a succinct description of the overall distribution, but they do not lend themselves to more definite analysis that is commonly undertaken on individual

CONTACT Xueying Wang xueying@math.wsu.edu

data. For instance, statisticians routinely calculate measures of dispersion, such as the variance or the inter-quantile range; economists use the Lorenz curve and Gini index to measure the degree of income disparity. We note that although the individual level data are not available, most analysis can be conducted based on an estimation of the underlying probability density function. This is precisely the purpose of the current work: to develop a method of probability density estimation based on interval aggregated data. Our goal is a flexible modeling approach – we do not make strong parametric assumptions on the population distribution as typically used in the literature.[9]

Here is a typical example of interval aggregated data:

Intervals	0–10%	10–20%	20–40%	40–60%	60–80%	80–100%
Averages	512	720	978	1103	1435	1976

This table says that the average of the individual level data below the 10% quantile is 512, between the 10% and 20% quantile is 720, etc. One should not confuse interval aggregated data with a histogram, a widely used concise summary of the data distribution that is also based on grouping of individual level data into intervals. To draw a histogram, we need to know (i) the values of the breaks for creating the intervals and (ii) the frequency of individual level data in each interval. The interval aggregated data we consider do not contain the information for drawing a histogram. For the interval aggregated data, (i) we do not know the values of the breaks, we only know that they are the sample quantiles but the actual values of the quantiles are unknown; (ii) we know the sample average in each interval. Thus the interval aggregated data do not provide a crude but transparent density estimate like the histogram, and so one cannot simply use it to compute the quantities of interest such as the variance, the Lorentz curve, and the Gini index. The kernel density estimation has been applied to interval aggregated data as a flexible alternative to parametric models to obtain poverty and inequality estimates and to describe global or national income distributions.[10–13] However, Minoiu and Reddy [14] recently reports that the kernel density estimator performs poorly and usually gives rise to nontrivial biases. The simulation study reported in this paper also confirms their finding. This is not surprising since the naive application of the kernel density estimation treats the aggregated data as a simple random sample and does not take into account the data aggregation process. So far we are not aware of any statistically rigorous nonparametric density estimation method that treats the aggregated data properly. We will fill in this gap in this paper.

Our approach is based on an explicit expression of the interval aggregated data as a function of the empirical quantile function of the individual level data. We model the theoretical quantile function as an integral of an exponential of a spline function, match the aggregated data with their theoretical counterpart using least squares, and use a roughness penalty to regularize the density estimation. Although the penalized spline method [15,16] has been widely used as a tool for flexible function estimation, its application to density estimation based on interval aggregated data are novel. Moreover, the specific penalty function we propose is non-standard for density estimation and causes some complication in computation. We develop a computational algorithm to implement our method. Numerical results in this paper show that the proposed penalized spline method does not have the reported caveat of the kernel density estimator.

The rest of the paper is organized as follows. Section 2 formulates the statistical estimation problem and Section 3 develops the penalized spline density estimator. Results from a simulation study are reported in Section 4. Section 5 applies the proposed method to the US household income data. Section 6 studies the dynamic of China's income distributions using the proposed method.

2. Problem formulation

Let u_i, $i = 0, \ldots, I$, be a set of distinct probability values (referred to as the quantile levels) that are listed in increasing order, where $u_0 = 0$ and $u_I = 1$. Suppose the individual level data are a random sample from a population. Let $\hat{Q}(u_i)$ denote the $(100 \times u_i)$th sample quantile of the individual level

data. The ith aggregated data point, denoted as $\hat{m}_i$, is the sample average of the individual level data that are in the interval $(\hat{Q}(u_{i-1}), \hat{Q}(u_i)]$ (or referred to as the quantile interval $(u_{i-1}, u_i]$). The available data are the aggregated data points $\hat{m}_i$, while the individual level data and the sample quantiles are not available. Our goal is to estimate the probability density function of the population using only the group averages $\hat{m}_i$'s.

As an example, consider a set of 11 quantile levels: 0, 0.1, 0.2, ..., 0.8, 0.9, 1. These probability values separate the individual level data into 10 equally sized groups by the corresponding sample quantiles. The sample averages of data in these groups are the interval aggregated data available to us for estimating the underlying probability density function. In general, the quantile levels may not be evenly spaced. For example, to capture the heavy right tail of the income distribution, 0.95 could be added to the above set of quantile levels by the data reporting agency to create finer quantile intervals at high values.

Let $F(x)$ denote the cumulative distribution function of the population and $Q(u) = F^{-1}(u)$ be the corresponding quantile function. Suppose the individual level data are $x_1, \ldots, x_n$, and denote the order statistics as $x_{(1)}, \ldots, x_{(n)}$. Let $\hat{F}$ and $\hat{Q}$ denote the empirical cumulative distribution function and quantile function based on the individual level data. Note that an individual level data point (order statistics) $x_{(j)}$ falls in the interval $(\hat{Q}(u_{i-1}), \hat{Q}(u_i)]$ if and only if $u_{i-1} < j \le u_i$. Thus, the sample average of individual level data falling in the quantile interval $(u_{i-1}, u_i]$ is

$$\hat{m}_i = \frac{1}{n(u_i - u_{i-1})} \sum_{nu_{i-1} < j \le nu_i} x_{(j)} = \frac{1}{u_i - u_{i-1}} \int_{\hat{Q}(u_{i-1})}^{\hat{Q}(u_i)} x \, \mathrm{d}\hat{F}(x). \tag{1}$$

By the law of large numbers, $\hat{Q}(u) \to Q(u)$ and $\hat{F}(x) \to F(x)$ almost surely, as $n \to \infty$.[17] It follows that

$$\hat{m}_i \to \frac{1}{u_i - u_{i-1}} \int_{Q(u_{i-1})}^{Q(u_i)} x \, \mathrm{d}F(x) = \frac{1}{u_i - u_{i-1}} \int_{u_{i-1}}^{u_i} Q(u) \, \mathrm{d}u. \tag{2}$$

This suggests that, if the quantile function is suitably parametrized, then we can estimate the parameters by matching the observations $\hat{m}_i$ and their limiting values. In the next section, we will use Equation (2) to construct a penalized spline estimator for flexible modeling of the quantile function.

We obtain the probability density function, denoted as $f(x)$, from the quantile function $Q(u)$ by using the following relationship:

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} = \frac{1}{\mathrm{d}Q(u)/\mathrm{d}u}\bigg|_{u=F(x)} = \exp\{-s(u)\}, \tag{3}$$

where x and u are linked by $u = F(x)$ and $x = Q(u)$. Note that u is in the space of probability values. To obtain the corresponding x, we need to solve the equation $Q(u) = x$ and any root finding algorithm can be applied. For example, the Newton–Raphson algorithm starts from an initial value $u^{(0)}$, and iterates according to

$$u^{(t+1)} = u^{(t)} - \frac{Q(u^{(t)}) - x}{Q'(u^{(t)})}$$

until convergence is reached. For a strictly increasing quantile function, the solution to the equation exists and is unique.

3. Penalized spline estimation

The quantile function $Q(u)$ should be monotonically non-decreasing and we assume it to be strictly increasing. Assume further that $Q(u)$ is differentiable. Thus $Q'(u)$ is positive for all u. We model

$\log Q'(u)$ using a basis expansion

$$\log Q'(u) = \sum_{k=1}^{K} \beta_k \phi_k(u) = \boldsymbol{\phi}^{\mathrm{T}}(u)\boldsymbol{\beta}, \tag{4}$$

where $\boldsymbol{\phi}(u) = (\phi_1(u), \ldots, \phi_K(u))^{\mathrm{T}}$ is a vector of basis functions, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^{\mathrm{T}}$ is the associated vector of the basis coefficients. In reality, Equation (4) is only an approximation. In our implementation, we use cubic B-splines [18] as basis functions and put the knots of the splines at the quantile levels that form the quantile intervals. Note that the number of parameters used in the basis expansion may be larger than the number of aggregated data points. To resolve the overparametrization problem, we later introduce a roughness penalty to regularize the spline fit.

Define $s(u) = \boldsymbol{\phi}^{\mathrm{T}}(u)\boldsymbol{\beta}$. It follows from Equation (4) that $Q'(u)$ is an exponential spline and so

$$Q(u) = \alpha + \int_0^u \exp\{s(v)\}\,\mathrm{d}v = \alpha + \int_0^u \exp\{\boldsymbol{\phi}^{\mathrm{T}}(v)\boldsymbol{\beta}\}\,\mathrm{d}v, \tag{5}$$

where $\alpha = Q(0)$ is a constant. With income data as a motivating example, we focus in this paper on positive random variables, and so we require that $\alpha = 0$. When α is nonzero, it simply adds an extra parameter in the nonlinear penalized least-squares problem to be introduced.

Defining

$$m(\boldsymbol{\beta}; a, b) = \frac{1}{b-a}\int_a^b \int_0^u \exp\{\boldsymbol{\phi}^{\mathrm{T}}(v)\boldsymbol{\beta}\}\,\mathrm{d}v\,\mathrm{d}u. \tag{6}$$

We can rewrite Equation (2) as

$$\hat{m}_i \to m(\boldsymbol{\beta}; u_{i-1}, u_i), \quad i = 1, \ldots, I.$$

This motivates us to estimate the spline coefficients by minimizing the following penalized least-squares criterion:

$$\sum_{i=1}^{I} \{\hat{m}_i - m(\boldsymbol{\beta}; u_{i-1}, u_i)\}^2 + \lambda J(s), \tag{7}$$

where $J(s)$ is a roughness penalty to regularize the spline fit and λ is a penalty parameter that controls the trade-off between the goodness-of-fit of the data and the smoothness of the estimated function.

Following Ramsay,[19] we can use the following penalty:

$$J_1(s) = \int_0^1 \{s'(u)\}^2\,\mathrm{d}u = \int_0^1 \left\{\frac{Q''(u)}{Q'(u)}\right\}^2\,\mathrm{d}u,$$

which penalizes the relative curvature of the quantile function $Q(u)$, since $s'(u) = Q''(u)/Q'(u)$ measures the size of the curvature $Q''(u)$ relative to the slope $Q'(u)$. One advantage of this penalty is that it can be written as a quadratic form of the vector of spline coefficients, which is convenient for solving the penalized least-squares problem. In fact, denoting

$$\dot{\boldsymbol{\phi}}(u) = (\phi_1'(u), \ldots, \phi_K'(u))^{\mathrm{T}}$$

and using the basis expansion (4), we can rewrite $J_1(s)$ as a quadratic form in $\boldsymbol{\beta}$:

$$J_1(s) = \boldsymbol{\beta}^{\mathrm{T}}\left\{\int_0^1 \dot{\boldsymbol{\phi}}(u)\dot{\boldsymbol{\phi}}^{\mathrm{T}}(u)\,\mathrm{d}u\right\}\boldsymbol{\beta}.$$

The above penalty function regularizes the roughness of the function $s(u)$, which is essentially the log density indexed by the quantile levels. However, our simulation studies indicate that this penalty

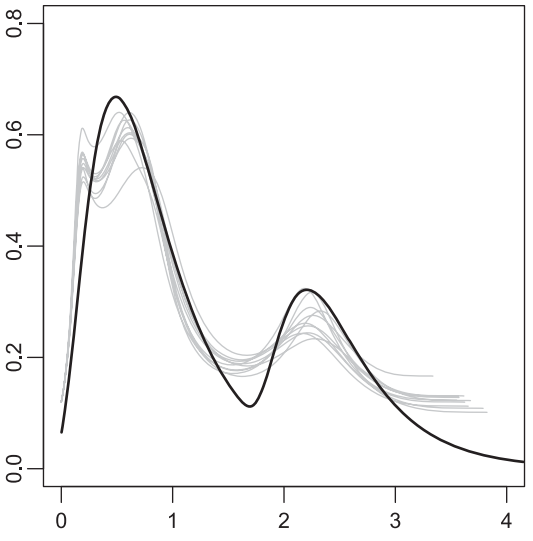

Figure 1. Density estimation using Ramsay's penalty for the same simulation step as the top right panel of Figure 2. The true density function is drawn in black and estimated functions for 10 simulation runs are in gray colour.

function does not work well; it focuses too much on the smoothness at the tail of the distribution but less so in the centre of the distribution. See Figure 1 for an example of using this penalty.

Now we consider a penalty that regularizes the roughness of the density function itself. Using the curvature of the density function to measure the roughness, we define

$$J(s) = \int_0^\infty \{f''(x)\}^2 \, \mathrm{d}x, \quad \text{with } f(x) = \exp\{-s(u)\}, \quad \text{and} \quad u = Q^{-1}(x). \tag{8}$$

By direct calculation and changing of variable $x = Q(u)$, we obtain an explicit expression of this penalty

$$\begin{aligned} J(s) &= \int_0^1 \{\mathrm{e}^{-3s\{u(x)\}}(2[s'\{u(x)\}]^2 - s''\{u(x)\})^2 \, \mathrm{e}^{s(u)} \, \mathrm{d}u, \\ &= \int_0^1 (\mathrm{e}^{-5s(u)/2}[2\{s'(u)\}^2 - s''(u)])^2 \, \mathrm{d}u. \end{aligned} \tag{9}$$

One nice consequence of this change of variable is that the domain of integration becomes the unit interval. In light of the basis expansion $s(u) = \boldsymbol{\phi}^{\mathrm{T}}(u)\boldsymbol{\beta}$, $J(s)$ is a function of $\boldsymbol{\beta}$. To make the dependence of $J(s)$ on $\boldsymbol{\beta}$ more transparent, we write (with a slight abuse of notation)

$$J(\boldsymbol{\beta}) = \int_0^1 \{\mathrm{e}^{-5\boldsymbol{\phi}(u)^{\mathrm{T}}\boldsymbol{\beta}/2}[2\{\dot{\boldsymbol{\phi}}(u)^{\mathrm{T}}\boldsymbol{\beta}\}^2 - \ddot{\boldsymbol{\phi}}(u)^{\mathrm{T}}\boldsymbol{\beta}\}]^2 \, \mathrm{d}u. \tag{10}$$

In the implementation, the integral appeared in this penalty function is calculated using the composite trapezoidal rule.

Then the penalized least-square criterion (7) can be written as

$$\sum_{i=1}^{I} \{\hat{m}_i - m(\boldsymbol{\beta}; u_{i-1}, u_i)\}^2 + \lambda J(\boldsymbol{\beta}). \tag{11}$$

We use the Gauss–Newton method to minimize this criterion over $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}$ denote the minimizer of Equation (7). Then the quantile function is estimated by plugging $\hat{\boldsymbol{\beta}}$ into Equation (5), and the probability density function is estimated by plugging $\hat{\boldsymbol{\beta}}$ into Equation (3).

The penalty parameter λ controls the trade-off between fit to the data and the smoothness of the fit. We choose the value of λ by minimizing the following crossvalidation (CV) criterion:

$$\mathrm{CV}(\lambda) = \sum_{i=2}^{i=I-1} \{\hat{m}_i - m(\hat{\boldsymbol{\beta}}^{(-i)}; u_{i-1}, u_i)\}^2, \tag{12}$$

where $\hat{\boldsymbol{\beta}}^{(-i)}$ is obtained by minimizing (11) using $\hat{m}_1, \ldots, \hat{m}_{i-1}, \hat{m}_{i+1}, \ldots, \hat{m}_I$. Here the crossvalidation is conducted for i from 2 to $I-1$, since we would like to avoid extrapolation at the end points.

4. Simulation

We conducted a simulation study to evaluate the performance of the proposed penalized spline method. We considered the following two scenarios of population distribution.

Scenario 1. Unimodal distribution. It is a truncation of a shifted log-normal distribution. Specifically, it is the conditional distribution of the random variable X provided $X \geq 0$, where

$$X = Y + c, \tag{13}$$

Y is log-normal $(-0.05, 0.5^2)$ distributed, and $c = -0.25$. The probability density function of X is

$$f_1(x; c) = \begin{cases} \dfrac{a}{\sqrt{2\pi}(0.5)(x-c)} \exp\left[-\dfrac{\{\log(x-c) + 0.05\}^2}{2(0.5)^2}\right], & \text{if } x > \max(0, c), \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

where a is a normalizing constant such that $f_1(x; c)$ is a proper density function. We can draw realizations of X through rejection sampling method.

Scenario 2. Bimodal distribution. It is a mixture of two unimodal distributions each of which is the same kind of distribution as in Scenario 1. The probability density function has the form

$$f_2(x; c_1, c_2) = 0.7 f_1(x; c_1) + 0.3 f_1(x; c_2), \tag{15}$$

where $f_1(x; c)$ is defined as in Equation (14), $c_1 = -0.25$, and $c_2 = 1.5$.

For each population distribution, we generated a random sample of size n from it and then created the I interval aggregated data points according to Equation (1), where the quantile limits $u_i, i = 0, \ldots, I$ are equally spaced points on [0,1]. We considered two values of n ($n = 10^3$ and $n = 10^4$) and three values of I ($I = 5$,10, and 20). All possible combinations of n and I give totally six setups for each scenario of simulation.

For each set of interval aggregated data, we applied the proposed penalized spline method and the kernel density estimator. The kernel density estimator of $f(x)$ based on the aggregated data is computed as follows:

$$\hat{f}_{kde}(x) = \frac{1}{Ih} \sum_{i=1}^{I} K\left(\frac{\hat{m}_i - x}{h}\right),$$

where $\hat{m}_i$'s are the interval aggregated data, $K(\cdot)$ is a kernel function, and h is the bandwidth. The density function of the standard Gaussian distribution is a typical choice of kernel function. The

bandwidth can be chosen as

$$h = 0.79 \times \text{IQR} \times I^{-1/5}.$$

See [20] for a comprehensive account of kernel density estimation. Even for positive random variables as considered in our simulation study, the kernel estimated density can be positive on the negative half of the real line. We thus modify the kernel density estimator by forcing its support to be on the positive half of the real line. Specifically, we defined the modified kernel density estimator to be

$$\hat{f}_{kde2}(x) = \begin{cases} C\hat{f}_{kde}(x), & \text{if } x > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where C is a normalizing constant to ensure $\hat{f}_{kde2}(x)$ to be a proper density function.

We ran the simulation 200 times for each setup. We used the integrated absolution error (IAE) and the integrated squared error (ISE) to assess the quality of the density estimation. The summary statistics of the IAE and ISE are given in Table 1 by displaying the simulated mean and standard derivation (SD). From this table, we have the following observations: (i) The proposed penalized spline method significantly outperforms the kernel density estimator and its modified version. (ii) The modified kernel density estimator slightly improves over the standard kernel density estimator. (iii) For all three methods, the IAE and ISE gets smaller when the sample size n of individual level data gets larger and when the number of intervals for data aggregation gets larger. The inferior performance of the kernel density estimators is mainly caused by the bias. As an example, Figure 2 shows estimated density functions in 50 simulation runs for each of the two scenarios with $n = 10^3$ and $I = 10$. It is clear that the kernel density estimators seriously underestimate the peak for the unimodal density and smooth out the two peaks for the bimodal density. The penalized spline method can capture well the features of the true density function. For the bimodal density function, the penalized spline slightly oversmooths the valley – this is the bias caused by data aggregation. When we increase the number of intervals, I, from 10 to 20, the bias at the valley is significantly reduced.

Table 1. Simulated mean (SD) of the IAE and ISE for estimating density functions by three methods: the penalized spline, the kernel density estimator, and the modified kernel density estimator, based on 200 simulation runs.

			Penalized spline		Kernel		Modified kernel	
Scenario	n	I	IAE	ISE	IAE	ISE	IAE	ISE
1	10^3	5	40.7	2.5	261.5	95.6	215.9	83.6
			(11.9)	(1.4)	(13.2)	(7.5)	(10.9)	(8.4)
		10	32.0	1.7	155.3	36.4	136.5	29.6
			(11.5)	(1.1)	(11.4)	(4.8)	(10.6)	(4.5)
		20	28.7	1.7	103.5	18.5	100.8	22.7
			(1.1)	(1.0)	(12.0)	(1.1)	(10.6)	(3.6)
	10^4	5	30.6	1.3	252.2	95.2	182.3	81.9
			(6.7)	(0.3)	(5.2)	(2.9)	(9.3)	(7.7)
		10	21.5	0.7	152.7	37.2	114.7	36.7
			(4.1)	(0.3)	(4.9)	(1.6)	(8.3)	(2.6)
		20	13.1	0.3	99.5	16.4	86.6	11.0
			(4.1)	(0.2)	(4.9)	(1.4)	(5.0)	(1.2)
2	10^3	5	74.5	10.3	178.8	63.1	151.8	41.7
			(8.5)	(1.7)	(12.4)	(7.8)	(11.0)	(7.5)
		10	28.7	1.3	170.2	46.2	130.4	39.4
			(6.9)	(1.2)	(9.4)	(5.3)	(9.3)	(5.7)
		20	27.2	1.3	142.0	37.3	125.4	24.5
			(4.8)	(0.9)	(9.5)	(5.1)	(9.3)	(4.9)
	10^4	5	70.6	9.7	173.4	52.5	150.3	39.3
			(7.7)	(0.4)	(10.7)	(4.8)	(9.5)	(6.0)
		10	20.5	0.6	152.7	37.2	114.7	36.7
			(5.1)	(0.2)	(8.3)	(2.3)	(7.3)	(2.3)
		20	19.1	0.5	112.8	16.4	108.9	23.0
			(2.9)	(0.2)	(8.2)	(2.0)	(6.4)	(1.9)

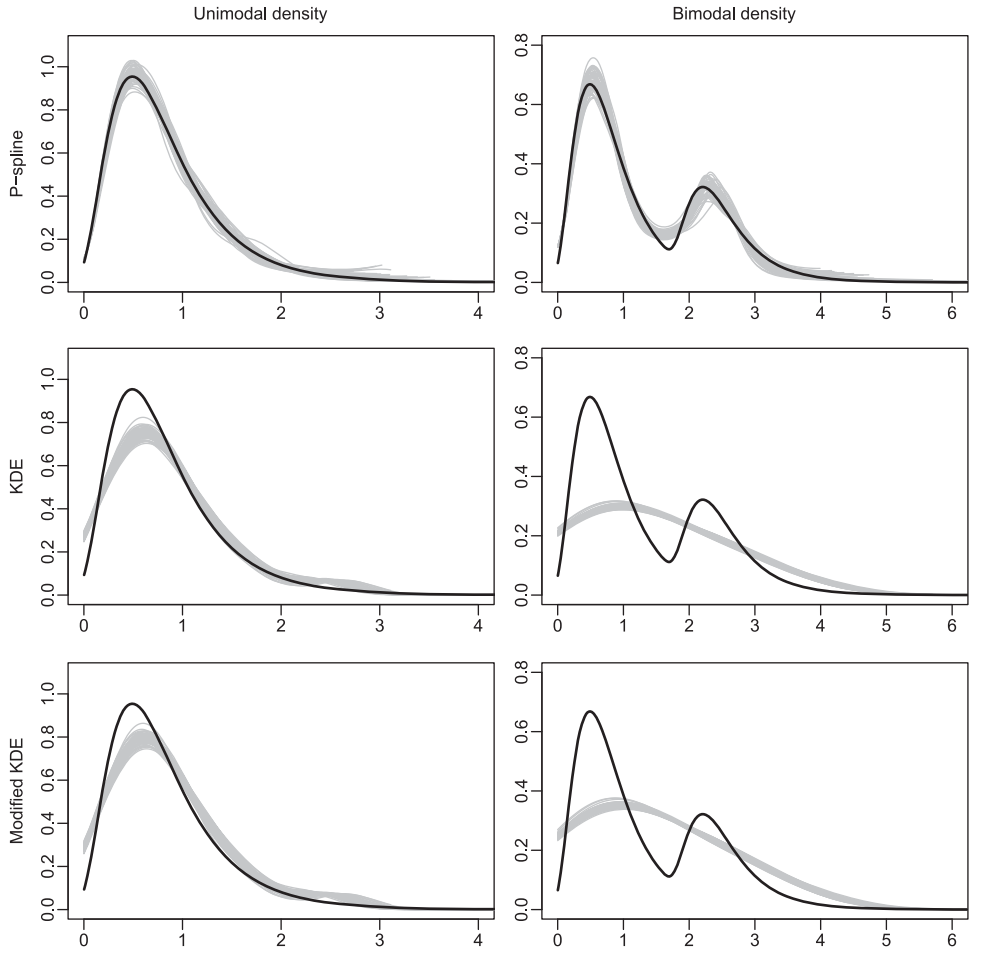

Figure 2. Comparison of the penalized spline and the kernel density estimators in a simulation study with $n = 10^3$ and $I = 10$. The true density functions are shown as solid lines and the estimated density functions from 50 simulation runs are shown as gray lines.

5. US household income data

The Current Population Survey (CPS), sponsored jointly by the US Census Bureau and the US Bureau of Labor Statistics, is the primary source of labor force statistics for the population of the United States. We used the US annual household income individual level data for year 2010, available from CPS, in a simulation study to create artificial aggregated data for illustration of the proposed method. We created the aggregated data from the individual level data and applied the proposed method and the kernel density estimators to the aggregated data. The histogram of the individual data provides a golden standard to evaluate the density estimation methods. One advantage of having the individual level data is that we can test the methods and see how the performance varies at different levels of the data aggregation.

In the first experiment we considered eight quantile intervals with the following probability limits: .05, .1, .2, .4, .6, .8, .9. The aggregated data were obtained as the sample averages in the quantile intervals, as formally defined in Equation (1). The China income data to be reported later in the next section has the same set of probability limits. We also considered 50 quantile intervals whose probability limits are evenly distributed in [0, 1]. Figure 3 shows the estimated density function by the three methods. With only eight aggregated data points, the penalized spline estimator captures very well

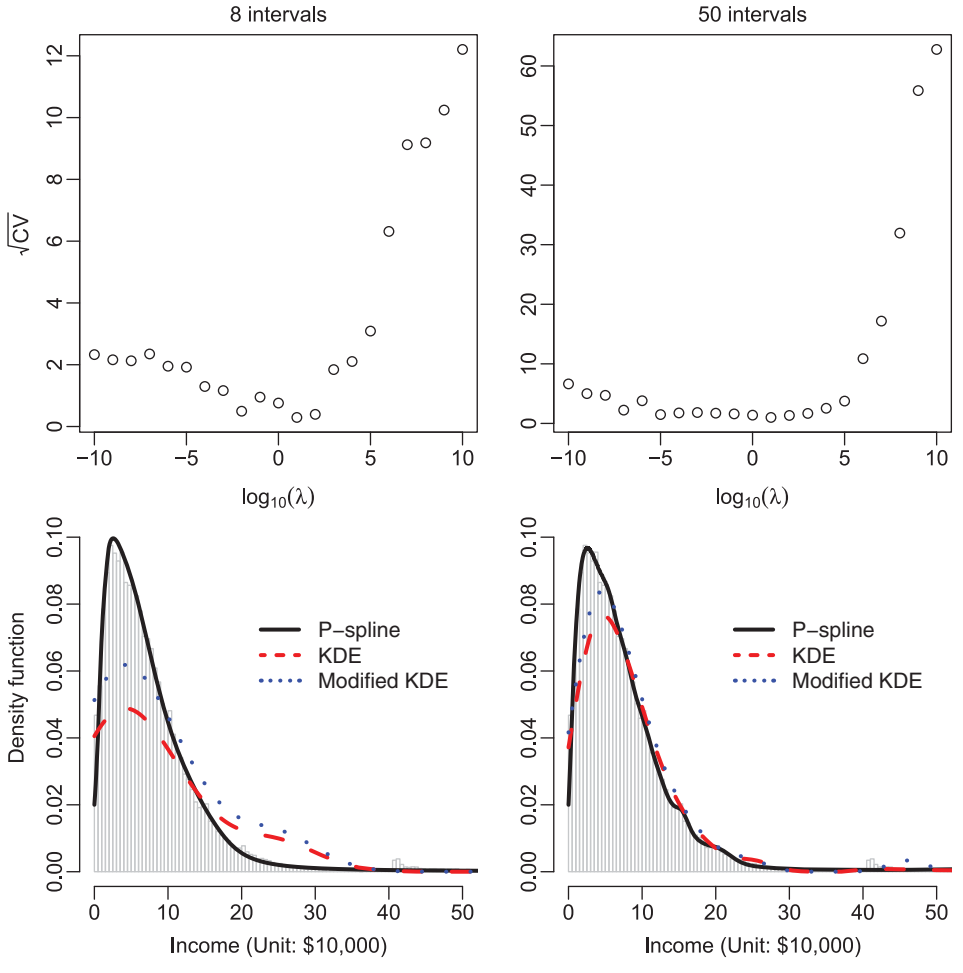

Figure 3. Density estimation of the US income distribution for year 2010 based on interval aggregated data. The first row shows the selection of the penalty parameter by the crossvalidation. The second row shows the estimated density by the penalized spline estimator and the kernel density estimator and its modified version. The histogram of the individual level data is also shown in the second row. The two columns show respectively the results for 8 and 50 quantile intervals.

the shape of the population distribution. When the number of intervals increases to 50, more subtle details of the distribution are revealed by the penalized spline estimator, for example, the small bump at around the point of $150,000. As comparisons, the kernel density estimators are seriously biased, although the modified version improves slightly. The crossvalidation was used to select the penalty parameter, as shown in the first row of the figure.

6. China urban household income data

China has registered unprecedented rapid economic growth in the past three decades, at the same time its economic inequality has increased substantially.[21] The economic disparity in China has attracted a great deal of interest from the academia, policy-makers and the general public. The actual degree of the economic inequality, however, has remained elusive since governmental disclosure of this matter has been sporadic at best. The National Statistics Bureau of China (NSBC) has maintained an expansive nationally representative annual survey of Chinese households but does not release the individual level survey data. Fortunately, the Chinese Statistical Yearbook, the official publication of

the NSBC, reports interval summary statistics of the urban income distributions. Starting in 1985, the Yearbook reports the average income of eight income groups separated by income quantiles at levels of 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9; the quantile limits are not reported. In this section, we apply the proposed penalized spline method to estimate the probability density function of China's income distribution based solely on the interval summary statistics reported by the Yearbooks. We deflate incomes with the Chinese annual Consumer Price Index and all variables are denominated by 1985 Chinese Yuan. In this application, the penalty is defined by the integrated second derivative of the density function, and the penalty parameter λ is selected by the crossvalidation.

There is a dramatically large increase in the income values for the 26-year span we considered. To ease comparison, we report the estimated density function for years 1985–1999 and 2000–2010 separately in Figure 4. Note that the units of the two plots are different: 1000 Yuan for the earlier period and 10,000 for the later. The overall distributions shifted rightwards steadily during the sample periods due to the economic growth. At the same time, the dispersion of distributions increased considerably, implying that the degree of economic disparity has risen substantially.

In addition to visual inspection, the estimated density functions allow us to examine the different facets in the evolution of China's income distributions. Given that we have estimated the quantile functions as a by-product of density estimation, a natural measure of the dispersion is the inter quartile range (IQR). The upper-left panel of Figure 5 shows the IQR together with the mean and median income as a function of time. The unequal incidence of economic growth is apparent in this plot. From 1985 to 1995, the median income tracks the average income closely. Since around 1995, the

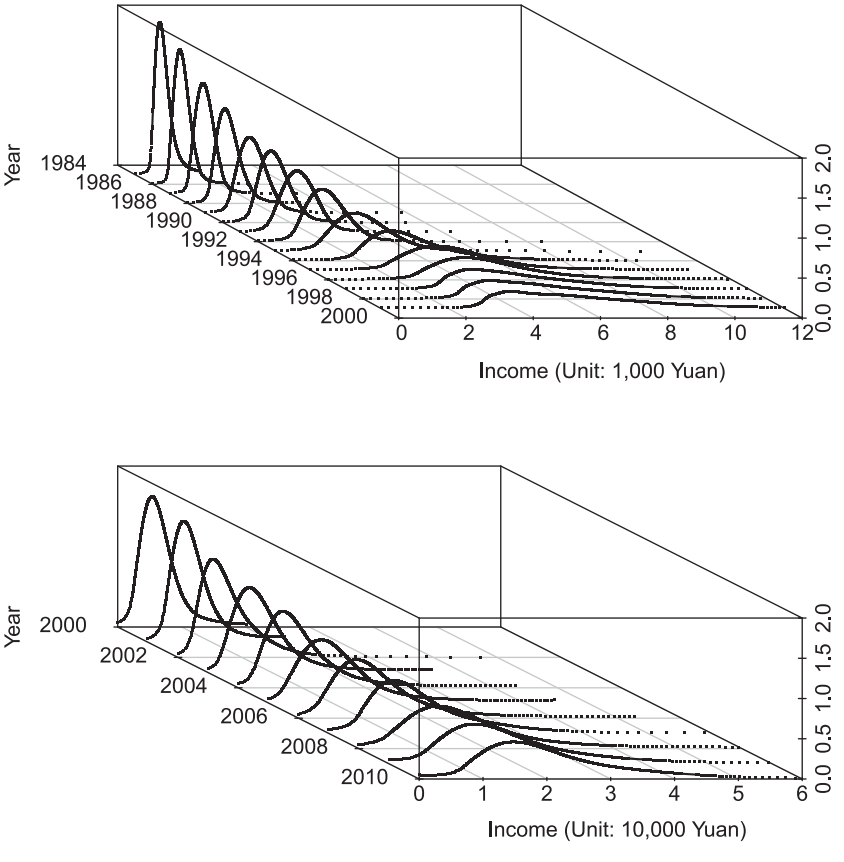

Figure 4. Estimated density function for China urban household income distributions. Left: 1985–1999 with income unit 1000 Yuan; Right: 2000–2010 with income unit 10,000 Yuan.

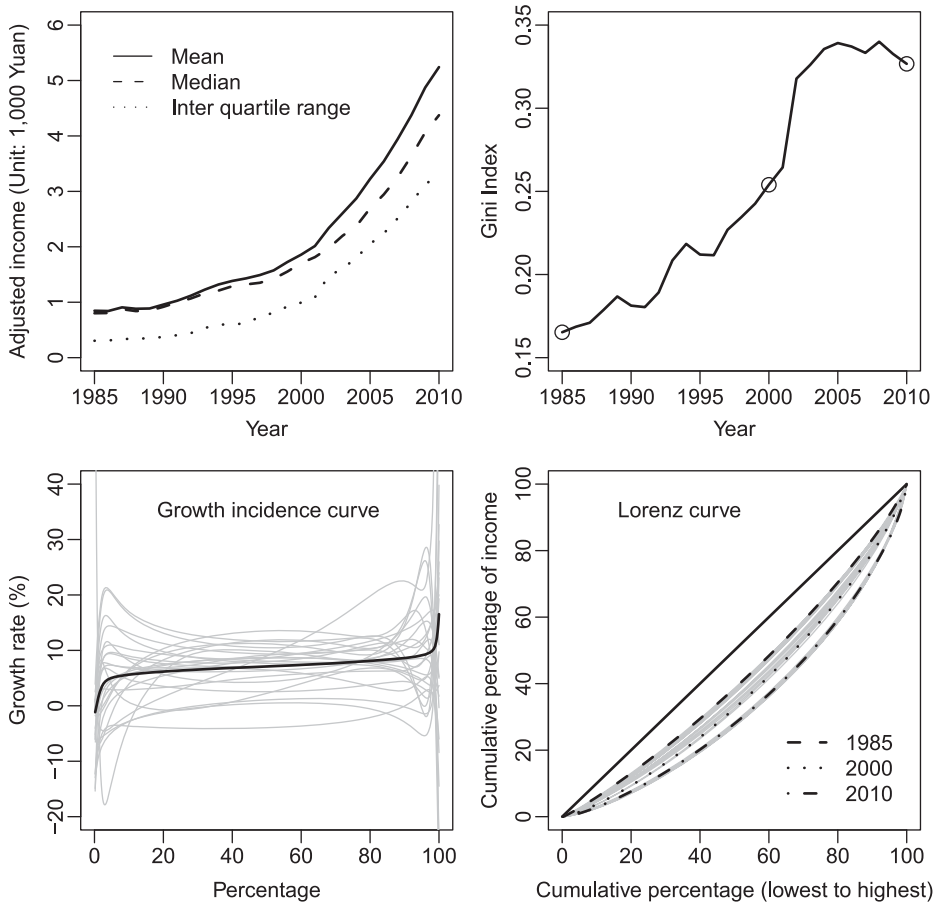

Figure 5. Dynamics of characteristics of China urban household income distributions 1985–2010. Upper-left: mean, median and IQR of the inflation-adjusted income. Upper-right: Gini index. Lower-left: growth incidence curves (yearly curves in gray and the average curve in black). Lower-right: Lorenz curves.

average income begins to overtake the median income, and the gap increases over time, reflecting the faster growth of incomes above the median. More telling is the rapid growth of the IQR during the same period. In 1985, the median, mean income and IQR are, respectively 801, 846 and 305 Yuan. These numbers rise to 4374, 5218 and 3309 Yuan in year 2010, increased by 5.5, 6.2 and 10.8 folds respectively. Clearly, the rising tide does not raise all boats equally.

To see how the growth of income is distributed over the whole population, we examine the growth incidence curve (GIC), which shows the growth rate in income between two points in time at each percentile of the distribution. Let Q_t and Q_{t+s} be the quantile functions at years t and $t+s$. The annualized GIC from year t to year $t+s$ is defined as

$$\mathrm{GIC}(u) = \left\{ \frac{Q_{t+s}(u)}{Q_s(u)} \right\}^{1/s} - 1, \quad u \in [0,1].$$

The GIC illustrates the complete profile of growth along a distribution. When $s = 1$, the GIC measures the growth in one-year period; when $s > 1$, the GIC measures the average growth over s years. If a GIC is everywhere positive, then the distribution at the second period first-order stochastically dominates that in the first period. The lower-left panel of Figure 5 shows the yearly GICs for all years and the average GIC over the 26-year period. The yearly GICs show substantial variability, especially at the

tails of the distribution. We observe that the GICs mostly take positive values, indicating growth of the income. The average GIC indicates that during the 26-year period, the entire population benefit from the rapid economic growth, with a median income growth rate around 7%, and the majority of the population enjoy a growth between 5 % and 9%. The average GIC is strictly increasing, suggesting that the income of the rich grew at a faster rate than that of the poor. The growth rate of those at the bottom of the income distribution is about 1%, while that at the top of the distribution is as high as 12%. However, this result should not be over-interpreted since the tails of the distribution can not be reliably estimated because of sampling bias and measurement errors.

To study the income inequality, we used the Lorenz curve, which shows the proportion of the total income assumed by the bottom u% of the population for $u \in [0, 100]$. Mathematically, the Lorenz curve is defined in terms of the quantile function $Q(u)$ as

$$L(u) = \frac{\int_0^u Q(u)\,\mathrm{d}u}{\int_0^1 Q(u)\,\mathrm{d}u}, \quad u \in [0, 1].$$

For a completely egalitarian distribution where everybody receives the same income, the Lorenz curve coincides with the diagonal of the unit square. As long as a distribution is unequal, the Lorenz curve is strictly below the diagonal. The further away the Lorenz curve is from the diagonal, the higher the level of inequality. The lower-right panel of Figure 5 depicts the estimated Lorenz curves. The Lorenz curves of the latter years do not cross those of the earlier years, suggesting that the steady increase in income disparity occurs throughout the entire distribution. It is also interesting to see that the Lorenz curves for later years (after 2002) are hardly distinguishable in the plot.

The Gini index is a popular measure of income disparity. The Gini index is calculated as twice of the area between the diagonal and the Lorenz curve. The upper-right panel of Figure 5 shows the Gini index of the income as a function of time. The Gini index increases steadily from 0.17 in 1985 to 0.33 in 2010, effectively doubled in 26 years. Although the NSBC does not release the Gini index regularly, the director of NSBC indicated in a press conference during the 2012 Chinese National Congress that the latest Gini index for the urban area is 0.33. Our estimates based on merely eight summary statistics is extremely close to the governmental figure calculated from the full survey of more than 60,000 households. This illustrative application highlights the efficacy of proposed method in facilitating informative and rigorous statistical analyses based on limited amount of interval aggregated data.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Huangs research was partly sponsored by NSF (DMS-0907170, DMS-1007618), and Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST). Xueying's work was partially supported by a grant from the Simons Foundation (#317047 to X. Wang). Zhous research was partially sponsored by NSF (DMS-0907170).

References

[1] Kakwani NC, Podder N. Efficient estimation of the lorenz curve and associated inequality measures from grouped observations. Econometrica. 1976;44(1):137–148.
[2] N Kakwani. On the estimation of Engel elasticities from grouped observations with application to Indonesian data. J Econ. 1977;6(1):1–19.
[3] Beach CM, Davidson R. Distribution-free statistical inference with lorenz curves and income shares. Rev Econ Stud. 1983;50(4):723–735.
[4] Villaseñor JA, Arnold BC. Elliptical Lorenz curves. J Econ. 1989;40(2):327–338.
[5] Chotikapanich D, Griffiths WE, Rao DSP. Estimating and combining national income distributions using limited data. J Bus Econ Stat. 2007;25:97–109.

[6] Chotikapanich D, Valenzuela MR, Rao DSP. Global and regional inequality in the distribution of income: estimation with limited/incomplete data. Empirical Econ. 2007;20:533–546.
[7] Wu X, Perloff JM. GMM estimation of a maximum entropy distribution with interval data. J Econ. 2007;138(2):532–546.
[8] Minoiu C, Reddy SG. Estimating poverty and inequality from grouped datal how well do parametric methods perform? J Income Distrib. 2009;18(2):160–179.
[9] Chen S, Ravallion M. The developing world is poorer than we thought, but no less successful in the fight against poverty. Q J Econ. 2010;125(4):1577–1625.
[10] Sala-i Martin X. The disturbing 'rise' of global income inequality. Technical report. National Bureau of Economic Research; 2002.
[11] Sala-i Martin X. The world distribution of income: falling poverty and . . . convergence, period. Q J Econ. 2006;121(2):351–397.
[12] Zhang Y, Wan G. Globalization and the urban poor in China. Research Paper 2006/42, UNU-WIDER, United Nations University (UNU); 2006.
[13] Ackland R, Dowrick S, Freyens B. Measuring global poverty: why PPP methods matter. Rev Econ Stat. 2013;95:813–824.
[14] Minoiu C, Reddy SG. Kernel density estimation on grouped data: the case of poverty assessment. J Econ Inequality. 2012.
[15] PHC Eilers, Marx BD. Flexible smoothing with B-splines and penalties. Statist Sci. 1996;11:89–102.
[16] Ruppert D, Wand P, Carroll RJ. Semiparametric regression. Cambridge Series in Statistical and Probabilistic Mathematics. London: Cambridge University Press; 2003.
[17] van der Vaart AW. Asymptotic statistics. Cambridge: Cambridge University Press; 1998.
[18] de Boor C. A practical guide to splines. Applied Mathematical Sciences 27. New York: Springer; 2001.
[19] JO Ramsay. Estimating smooth monotone functions. J R Statist Soc: Ser B (Statist Methodol). 1998;60(2):365–375.
[20] Silverman B. Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability. London: Chapman and Hall; 1986.
[21] Wu X, Perloff JM. China's income distribution, 1985–2001. Rev Econ Stat. 2005;87(4):763–775.