Tuning Parameter-Free Nonparametric Density Estimation from Tabulated Summary Data

Ji Hyung Lee∗ Yuya Sasaki† Alexis Akira Toda‡ Yulong Wang§

May 19, 2023

Abstract

Administrative data are often easier to access as tabulated summaries than in the original format due to confidentiality concerns. Motivated by this practical feature, we propose a novel nonparametric density estimation method from tabulated summary data based on maximum entropy and prove its strong uniform consistency. Unlike existing kernel-based estimators, our estimator is free from tuning parameters and admits a closed-form density that is convenient for post-estimation analysis. We apply the proposed method to the tabulated summary data of the U.S. tax returns to estimate the income distribution.

Keywords: grouped data, income distribution, maximum entropy

JEL codes: C14, D31

∗ Ji Hyung Lee: jihyung@illinois.edu. Department of Economics, University of Illinois, 214 David Kinley Hall, 1407 West Gregory Drive, Urbana, IL 61801, USA
† Yuya Sasaki: yuya.sasaki@vanderbilt.edu. Department of Economics, Vanderbilt University, VU Station B #351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819, USA
‡ Alexis Akira Toda: atoda@ucsd.edu. Department of Economics, University of California San Diego, 9500 Gilman Dr, #0508, La Jolla, CA 92093-0508, USA
§ Yulong Wang: ywang402@syr.edu. Department of Economics, Syracuse University, 110 Eggers Hall, Syracuse, NY 13244-1020, USA
1 Introduction

Researchers can often access a tabulated summary of data more easily than the original data containing confidential information at the individual level. Examples include administrative tax data containing detailed information about individual-level records of income. Tax authorities often release summary statistics of the income distributions in a tabulated format, such as the number of taxpayers and their average income, grouped by bins of income levels. Despite their lack of details, such tabulated summary data are still useful for researchers in analyzing income distributions, especially for historically old times for which micro data are no longer available.

A typical econometric method for estimating the cross-sectional distribution of a continuous random variable, such as the kernel density estimator and the empirical cumulative distribution function, requires individual-level information, and thus is not suitable when researchers only have access to tabulated summaries. In this paper, we propose a novel method for nonparametrically estimating the probability density function of an absolutely continuous distribution from tabulated summary data based on the maximum entropy (ME) principle. This ME density estimator enjoys a number of desirable properties. First, it is piecewise exponential, which allows for analytical post-estimation integration to calculate cumulative distribution estimates such as the mean, variance, top income shares, Lorenz curve, and the Gini coefficient. Second, and more importantly, unlike alternative kernel-based estimators (Blower and Kelsall, 2002; Sun, 2014), the ME density estimator is free from tuning parameters such as the bandwidth. This feature is attractive in practice and provides a complete theoretical justification of the method, unlike the existing kernel-based methods for which the theory fails to formally account for the tuning parameter choice in practice. We establish the strong uniform consistency for the ME density and cumulative distribution estimators under the asymptotic framework in which the resolution of bins in the table becomes finer at certain rates as the underlying sample size increases.

To illustrate our proposed method, we consider two applications. The first is a simulation study of the calculation of top income shares. Compared to the existing methods, our proposed method generates smaller bias and root mean squared error (RMSE). The second is the estimation of the income distribution and top income shares from the tabulated summaries of the U.S. tax returns data. Unlike the popular Pareto interpolation method of Piketty (2003), which relies on parametric assumptions in the upper tail, our method allows for the estimation in
mid-sample as well as in the tails without parametric assumptions.

**Related literature** Our paper is related to a large literature in economics on inequality measures as well as in statistics and econometrics of estimation with grouped data.

Applied researchers working with income inequality measures have long been studying the interpolation problem from grouped data; see for instance Cowell and Mehta (1982) for an early review. The parametric methods for estimating the Lorenz curve (and hence computing the Gini coefficient) of Kakwani and Podder (1976) and Villaseñor and Arnold (1989) have been used by the World Bank. Feenberg and Poterba (1993) and Piketty (2003) interpolate the upper tail of the income distribution by the Pareto distribution with local Pareto exponents estimated from a tabulated summary. See Piketty and Saez (2003) for an application to the U.S. income distribution. More recently, Blanchet, Fournier, and Piketty (2022) apply spline interpolation to what they call the inverted Pareto coefficients. Compared to this applied literature, our approach enjoys several advantages such as that (i) the method is supported by the rigorous econometric theory that guarantees the performance in large samples, (ii) the method is nonparametric across the entire distribution of income, and (iii) the output is a closed-form density that is analytically tractable and convenient for post-estimation analysis.

In terms of estimating the income distribution with tabulated data, our proposed method differs from the existing methods, which can be categorized into parametric and nonparametric ones. For the former, Hajargasht, Griffiths, Brice, Rao, and Chotikapanich (2012) assume a parametric income distribution and develop a generalized method of moment (GMM) estimator for the unknown parameters. See also Chen (2018) and Hajargasht and Griffiths (2020) for other GMM methods relying on parametric assumptions. Despite their good fit in some data sets (e.g., Jorda, Sarabia, and Jäntti, 2021), parametric methods in general suffer from model misspecifications. Specifically, although the estimator of the coefficients in the parametric model might still converge to some pseudo true values, the implied density estimator is in general inconsistent. The theoretical effect of misspecification on further estimations of other features, say moments and top income shares, are unknown. They may exhibit large biases in finite samples, as we show by simulation studies in Section 4. In contrast, our estimator is nonparametric and relatively robust to parametric assumptions.

For nonparametric estimation with binned/grouped data, Scott and Sheather (1985) study the kernel density estimator when the observations are equally spaced.
On the other hand, we do not require data to be equally spaced. Without the equal spacing restriction, Blower and Kelsall (2002) propose an alternative estimator with Gaussian kernel function, which Sun (2014) extends by allowing for other kernel functions. These kernel approaches require a bandwidth as a tuning parameter, whose choice is challenging in practice especially under the nonstandard sampling setup of binned/grouped/tabulated data. In contrast, our proposed ME estimator is free from tuning parameters and therefore more attractive in practice. Furthermore, this tuning-parameter-free feature of our proposed method provides a complete theoretical justification under weak regularity condition. Reyes, Francisco-Fernández, and Cao (2016) derive the orders of magnitude of the bias and the variance of Scott and Sheather’s estimator under very strong conditions: (i) the bin size is of a smaller order than $h^2$, where $h$ denotes the bandwidth of the kernel estimator and (ii) the underlying distribution function is seven-times differentiable with bounded derivatives. The first condition implies that the group structure is negligible and hence the bias and the variance are asymptotically the same as in the case with individual observations. The second condition is strong and rules out some candidate distributions, such as double Pareto, which is used in empirical studies of income distribution. Unlike the existing kernel-based methods, our ME estimator only requires the underlying density to be Lipschitz continuous and is based on minimizing the Kullback-Leibler divergence, which lead to both numerical and theoretical advantages. See Section 3 for details.

Finally, our paper is related to the large literature that applies the maximum entropy (ME) principle. Historically, the ME principle was developed in physics to infer the population distribution (e.g., energy distribution of gas molecules) from macroscopic variables (e.g., temperature); see Jaynes (1957). In economics, applications of the ME method include general equilibrium theory (Foley, 1994; Toda, 2010, 2015), diagnosis of asset pricing models (Stutzer, 1995), derivative pricing (Stutzer, 1996), discretization of probability distributions and stochastic processes (Tanaka and Toda, 2013, 2015; Farmer and Toda, 2017), among others. In this paper, we apply the ME principle as a tool to impose moment restrictions implied by the tabulated summary data. Applications in econometrics include Kitamura and Stutzer (1997) and Wu (2003), among others. To our knowledge, none of the existing ME methods work for tabulated summary data.

**Organization of the paper** Section 2 introduces the general data framework and previews the U.S. tax return data set used in the application. Section 3 introduces the nonparametric density estimator and proves its strong uniform
consistency. Section 4 presents simulation studies. Section 5 applies the proposed method to the U.S. tax return data set.

2 Tabulated summary data

2.1 General data framework

Consider the latent sample \( \{Y_i\}_{i=1}^n \), which is not directly observed by the researcher. Denote the order statistics in descending order by \( \{Y_{(i)}\}_{i=1}^n \) such that

\[
Y_{(1)} \geq Y_{(2)} \geq \cdots \geq Y_{(n)}.
\]

For each \( m \in \{1, \ldots, n\} \), define the partial sum of top \( m \) order statistics

\[
S_m := \sum_{i=1}^{m} Y_{(i)}.
\] (2.1)

Consider a positive number \( K \) of bins, where \( \{t_k\}_{k=0}^K \) denotes the sequence of bin threshold values such that

\[
\infty = t_0 > t_1 > t_2 > \cdots > t_K.
\]

Let \( n_k \) be the number of order statistics included in the top \( k \) bins, that is, \( n_k := \# \{ i : Y_{(i)} \geq t_k \} \). The tabulated data are summarized as \( \{(t_k, n_k, S_{n_k})\}_{k=1}^K \), which is observed by the researcher.

2.2 Example: summary of income data by tax authorities

As a concrete example of the data framework just described, consider the latent sample \( \{Y_i\}_{i=1}^n \) of the values \( Y_i \) of income, where \( i \) indexes potential taxpayers and \( n \) is the sample size. Due to confidentiality concerns, in general there is no public access to administrative data of income. Publicly available data on the income distribution released from tax authorities often take the form of the tabulated summary \( \{(t_k, n_k, S_{n_k})\}_{k=1}^K \).

Table 1 presents an example data set from the 2019 U.S. tax returns.\(^1\) In this case, Table 1 shows partial information from Internal Revenue Service, Statistics of Income (SOI) Individual Income Tax Returns Publication 1304 [https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-returns-complete-report-publication-1304], Table 1.4 under “Basic Tables”. Adjusted gross income (AGI) is AGI less deficit. We omit the row corresponding to negative income.
example, the number of income groups is $K = 18$. Column (1) shows the lower threshold $t_k$ of adjusted gross income (AGI) for each income group $k \in \{1, \ldots, K\}$. Column (2) shows the number of taxpayers within each income group $k$, which corresponds to $n_k - n_{k-1}$ in our notations, where we set $n_0 = 0$ by convention. Column (3) shows the total income (AGI) accruing to taxpayers in each income group $k$ in units of 1,000 U.S. dollars, which corresponds to $(S_{n_k} - S_{n_{k-1}})/1,000$ in our notations, where we set $S_0 = 0$ by convention. We thus observe the tabulated summary data $\{(t_k, n_k, S_{n_k})\}_{k=1}^K$ of AGI, where $K = 18$.

Table 1: Income distribution in the United States, 2019.

| Income group $k$ | Adjusted gross income (AGI) $t_k$ | # returns | Total income |
|-----------------|-----------------------------------|-----------|--------------|
| 18              | $1$                               | $9,866,880$ | $24,439,988$ |
| 17              | $5,000$                           | $9,925,940$ | $74,584,857$ |
| 16              | $10,000$                          | $11,087,737$ | $138,230,399$ |
| 15              | $15,000$                          | $10,039,446$ | $175,255,963$ |
| 14              | $20,000$                          | $9,493,968$  | $213,660,160$ |
| 13              | $25,000$                          | $9,289,939$  | $254,877,708$ |
| 12              | $30,000$                          | $16,090,602$ | $560,073,192$ |
| 11              | $40,000$                          | $12,503,041$ | $560,258,808$ |
| 10              | $50,000$                          | $22,238,948$ | $1,366,892,948$ |
| 9               | $75,000$                          | $14,118,568$ | $1,222,947,425$ |
| 8               | $100,000$                         | $21,997,582$ | $3,004,363,636$ |
| 7               | $200,000$                         | $7,297,883$  | $2,090,808,696$ |
| 6               | $500,000$                         | $1,162,371$  | $781,920,814$  |
| 5               | $1,000,000$                       | $254,197$   | $305,561,848$  |
| 4               | $1,500,000$                       | $103,075$   | $176,961,208$  |
| 3               | $2,000,000$                       | $143,514$   | $425,088,995$  |
| 2               | $5,000,000$                       | $34,738$    | $237,781,553$  |
| 1               | $10,000,000$                      | $20,876$    | $590,230,011$  |

Note: “AGI threshold” is the lower threshold of adjusted gross income (AGI) that defines the income groups. “# returns” is the number of tax returns with income weakly above the lower AGI threshold and strictly below the upper AGI threshold. “Total income” is the total income (in units of 1,000 U.S. dollars) accruing to taxpayers in each income group.

3 Main results

This section investigates a method to characterize a well-behaved density function of the distribution of $Y$ from the tabulated summary data $\{(t_k, n_k, S_{n_k})\}_{k=1}^K$ intro-
duced in Section 2. We first consider the case when the sample size is infinite and there are no sampling errors in bin probabilities and conditional means. We next propose a feasible estimator and study its asymptotic properties as the sample size tends to infinity.

3.1 Maximum entropy density

Let $F$ denote the true cumulative distribution function (CDF) of $Y$, which is assumed to be absolutely continuous with probability density function denoted by $f = F'$. Suppose that the thresholds satisfy

$$\infty = t_0 > t_1 > t_2 > \cdots > t_K =: t,$$

and let $I_k := [t_k, t_{k-1})$ denote the interval for the top $k$-th bin with top fractile denoted by $p_k = P(Y \geq t_k) = 1 - F(t_k)$. For each $k$, the bin probability and conditional mean are defined by

$$q_k := P(Y \in I_k) = \int 1_{I_k}(y) \, dF(y) = p_k - p_{k-1}, \quad (3.1a)$$

$$y_k := E[Y \mid Y \in I_k] = \frac{1}{q_k} \int 1_{I_k}(y) y \, dF(y), \quad (3.1b)$$

where $1_I(\cdot)$ denotes the indicator function indicating that the argument belongs to the set $I$.

Obviously, given only the finite tabulation $\{(p_k, y_k)\}_{k=1}^K$, we do not have sufficient moment restrictions to pin down the true density function $f$. The maximum entropy (ME) method is useful when only certain moment conditions are given. In our context of characterizing the distribution of $Y$ from a tabulation, we can proceed as follows.

Letting $g$ denote a generic density, the given moment conditions consistent with (3.1) are

$$\int 1_{I_k}(y) g(y) \, dy = q_k, \quad (3.2a)$$

$$\int 1_{I_k}(y) yg(y) \, dy = q_k y_k, \quad (3.2b)$$

for each $k \in \{1, \ldots, K\}$. The ME density $f^*$ is defined by the density $g$ on $I := [t, \infty)$ that minimizes the Kullback-Leibler divergence (with respect to the
improper uniform density)
\[ \int_I g(y) \log g(y) \, dy \quad (3.3) \]
subject to the moment restrictions (3.2). Below, we let \( L_1^+(I) \) denote the equivalence class (identified by the \( L^1 \) norm) of nonnegative, measurable, and integrable functions \( g : I \to \mathbb{R} \). The following proposition characterizes the solution to the ME problem.

**Proposition 1.** Let \( \infty = t_0 > t_1 > \cdots > t_K \) be the thresholds and \( y_k \in (t_k, t_{k-1}) \) be the average of \( Y \) in group \( k \). Then the function
\[
J_k(\lambda; t) := \begin{cases} 
  y_k \lambda - \log \left( \frac{e^{\lambda y_k} - e^{\lambda t_k}}{\lambda} \right) & (\lambda \neq 0) \\
  -\log (t_{k-1} - t_k) & (\lambda = 0)
\end{cases} \quad (3.4)
\]
is strictly concave in \( \lambda \in \mathbb{R} \) and achieves a unique maximum \( \lambda_k^* \), which satisfies
\[
\lambda_1^* = -\frac{1}{y_1 - t_1} < 0 \quad \text{and} \quad \lambda_k^* \begin{cases} < 0 & \text{if } y_k < \frac{t_k + t_{k-1}}{2}, \\
= 0 & \text{if } y_k = \frac{t_k + t_{k-1}}{2}, \quad \text{for } k \geq 2, \\
> 0 & \text{if } y_k > \frac{t_k + t_{k-1}}{2},
\end{cases} \quad (3.5)
\]
The ME problem has a unique solution \( f^* \in L_1^+(I) \), which is piecewise exponential and satisfies
\[
f^*(y) = \begin{cases} 
  q_k \lambda_k^* e^{\lambda_k^* y} \quad (\lambda_k^* \neq 0) \\
  q_k \frac{e^{\lambda_k^* t_k} - e^{\lambda_k^* t_{k-1}}}{t_{k-1} - t_k} \quad (\lambda_k^* = 0)
\end{cases} \quad (3.6)
\]
for \( y \in I_k = [t_k, t_{k-1}) \). The minimum value of the Kullback-Leibler divergence (3.3) is given by
\[
J^*(t_1, \ldots, t_K) := \sum_{k=1}^K q_k (J_k(\lambda_k^*; t) + \log q_k). \quad (3.7)
\]

### 3.2 Estimation

This section proposes a feasible analog of the ME density characterized in Section 3.1 for estimation of the true density function \( f \) of \( Y \). We then establish its strong uniform consistency.

We construct a feasible ME density estimator by replacing \( q_k \) and \( y_k \) with their empirical analogs \( \hat{q}_k = (n_k - n_{k-1})/n \) and \( \hat{y}_k = (S_{n_k} - S_{n_{k-1}})/(n_k - n_{k-1}) \), respectively. Letting \( t := \{t_k\}_{k=1}^K \) denote the vector of thresholds, we thus define the
sample-analog ME estimator $\hat{f}$ of $f$ as the solution to the constrained optimization problem of minimizing (3.3) subject to (3.2) with $\hat{Q}_k$ and $\hat{y}_k$ in place of $q_k$ and $y_k$, respectively.

We now establish the almost sure uniform consistency of $\hat{f}$ for the true density function $f$ over any compact subset of the domain of $f$ as $n \to \infty$. To this end, consider the following conditions.

**Assumption 1.** The following conditions hold.

(i) $\{Y_i\}_{i=1}^n$ is IID with density $f$.

(ii) The density $f : \mathbb{R} \to [0, \infty)$ is Lipschitz continuous with constant $L > 0$, that is, for all $y_1, y_2 \in \mathbb{R}$ we have

$$|f(y_1) - f(y_2)| \leq L |y_1 - y_2|.$$  

(iii) There exist some constants $c_1, c_2 > 0$ and $0 < r_2 \leq r_1 < 1/4$ such that

$$c_1 n^{-r_1} \leq \min_{2 \leq k \leq K} \{t_{k-1} - t_k\} \leq \max_{2 \leq k \leq K} \{t_{k-1} - t_k\} \leq c_2 n^{-r_2}. \quad \text{(3.8)}$$

Furthermore, letting $D := \{y \in \mathbb{R} : f(y) > 0\}$ be the domain of $f$, we have

$$\lim \sup_{n \to \infty} t_K \leq \inf D \quad \text{and} \quad \lim \inf_{n \to \infty} t_1 \geq \sup D.$$

Condition (i) assumes a random sample. Condition (ii) implies that $f$ is almost everywhere differentiable with a bounded derivative, which is much weaker than typical assumptions on kernel estimators that require high-order smoothness conditions. Condition (iii) requires that the length of any bin is neither too large nor too small, as well as that the interval $[\lim \sup_{n \to \infty} t_K, \lim \inf_{n \to \infty} t_1]$ covers the domain $D$ of $f$. With these conditions, the following theorem establishes the strong uniform consistency of the maximum entropy density estimator $\hat{f}$ for the true density function $f$ over any compact subset of the domain and also bounds the convergence rate. The proof is non-trivial and deferred to Appendix B.

**Theorem 2.** Suppose that Assumption 1 holds and let $C \subset D$ be compact. Then, as $n \to \infty$, we have

$$\sup_{y \in C} \left| \hat{f}(y) - f(y) \right| \xrightarrow{a.s.} 0,$$  

$$\sup_{y \in C} \left| \hat{f}(y) - f(y) \right| = O(n^{-r_2}) + O_p\left(n^{-\frac{1-4r_1}{2}}\right).$$  

9
A few remarks are in order regarding this result on the convergence rate. First, while the rate is decomposed into the deterministic part and the stochastic part, we do not have control over the trade-off between these two components in the absence of a tuning parameter. This implies a drawback of the tuning parameter-free approach. Second, the rate depends on the parameters $r_1$ and $r_2$ of bin lengths. Slowly vanishing bin lengths (i.e., small $r_1$ and $r_2$) yield small variances at the expense of large biases. Quickly vanishing bin lengths (i.e., large $r_1$ and $r_2$) yield small biases at the expense of large variances. Third, suppose $r_1 = r_2$ for simplicity. Then, $1/6 < r_1 = r_2 < 1/4$ implies that the stochastic part dominates, while $r_1 = r_2 < 1/6$ implies that the deterministic part dominates. In the latter case, the limit distribution has a biased center, and it is difficult to conduct statistical inference in general. This is another limitation of the tuning parameter-free approach.

Reyes et al. (2016) derive the convergence rate of the kernel estimator proposed by Scott and Sheather (1985). Reyes et al. (2016) assume that the bin lengths are $o(h_n^2)$, where $h_n$ denotes the bandwidth satisfying $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. This condition implies that the group/bin structure is asymptotically negligible, and the resulting orders, $O(h_n^2)$ and $O_p(1/\sqrt{nh_n})$, of the deterministic and stochastic parts, respectively, are the same as those in the standard case with individual observations. On the one hand, choosing a certain bandwidth $h_n$ could lead to a smaller bias or variance than our ME estimator. On the other hand, the assumption that the bin lengths are $o(h_n^2)$ is very restrictive and could be violated in empirical studies where the bins are not too small. Furthermore, Reyes et al. (2016) assume that the underlying distribution function is seven-times differentiable with bounded derivatives. In contrast, our ME estimator only requires Lipschitz continuity for $f$, which is another advantage.

Having established the strong uniform consistency of the density, it is straightforward to establish the same for the cumulative distribution function (CDF) and quantiles. Define the estimator $\hat{F}$ of $F$ by

$$\hat{F}(y) = \int_{-\infty}^{y} \hat{f}(y) \, dy.$$ 

The following corollary shows the strong uniform consistency of $\hat{F}$.

**Corollary 3.** Suppose that Assumption 1 holds and let $C \subset D$ be compact. If
$-\infty < \inf D$, then, as $n \to \infty$, we have

$$\sup_{y \in C} \left| \hat{F}(y) - F(y) \right| \xrightarrow{a.s.} 0.$$

Let $Q_\tau = \inf \{ y : \tau \leq F(y) \}$ denote the $\tau$-th quantile of $F$. Given $\hat{F}$, we can estimate $Q_\tau$ by the analog $\hat{Q}_\tau = \inf \left\{ y : \tau \leq \hat{F}(y) \right\}$. The following corollary shows that this quantile estimator is also consistent.

**Corollary 4.** Suppose that Assumption 1 holds and let $C \subset D$ be compact. If $-\infty < \inf D$ and $Q_\tau \in \text{int} C$, then, as $n \to \infty$, we have $\hat{Q}_\tau \xrightarrow{a.s.} Q_\tau$.

### 3.3 Discussion

In an early review of interpolation methods from grouped data of income, Cowell and Mehta (1982) list the following ten desirable properties (with slight rewording) that the hypothetical interpolated distribution should possess.

1. The bin probability and conditional mean ($q_k$ and $y_k$ in our notation in (3.2)) of the estimated density $\hat{f}$ agree with the tabulated summary.

2. $\hat{f} \geq 0$.

3. $\hat{f}$ is continuous within any interval $I_k = [t_k, t_{k-1})$.

4. $\hat{f}$ is continuous.

5. $\hat{f}$ is differentiable.

6. $\lim_{y \to \infty} \hat{f}(y) = 0$.

7. $\lim_{y \to -\infty} \hat{f}(y) = 0$.

8. $\hat{f}$ has “few” turning points within each bin.

9. The range of $\hat{f}$ is “small” on any given interval.

10. $\hat{f}$ admits a closed-form expression to compute inequality measures.

In addition to these properties, we would like to add:

11. $\hat{f}$ converges to the true density $f$ as the sample size tends to infinity.
The methods reviewed in Cowell and Mehta (1982) as well as those proposed thereafter satisfy only a few of these properties. For instance, the recent method of Blanchet et al. (2022) does not satisfy 8 and 9 (because it uses polynomial interpolation), 10 (because it interpolates the inverted Pareto coefficients, not the density), or 11 (they do not provide formal theorems).

In contrast, our ME density estimator \( \hat{f} \) satisfies all properties except 4 and 5. To see this, property 1 holds by construction and 11 is established in Theorem 2. All other properties (except 4 and 5) hold because \( \hat{f} \) is piecewise exponential explicitly given by (3.6) and hence is nonnegative, continuously differentiable, and monotonic on each interval. Regarding properties 4 and 5, they clearly hold except at bin thresholds.

To illustrate property 10 further, we present some integral formulas that are useful when computing the CDF and top income shares when the density is piecewise exponential. Consider the piecewise exponential density (3.6). To simplify the notation, let \( q_k = q \), \( \lambda_k = \lambda \), \( t_k = a \), and \( t_{k-1} = b \). Therefore, for \( y \in [a,b) \), the density is

\[
f(y) = \begin{cases} 
\frac{q e^{\lambda y}}{e^{\lambda a} - e^{\lambda b}}, & (\lambda \neq 0) \\
\frac{q}{b-a}, & (\lambda = 0)
\end{cases}
\]

The counter CDF (tail probability) can be computed using

\[
\int_y^b f(y) \, dy = \begin{cases} 
q \frac{e^{\lambda y} - e^{\lambda b}}{e^{\lambda a} - e^{\lambda b}}, & (\lambda \neq 0) \\
q \frac{b-y}{b-a}, & (\lambda = 0)
\end{cases}
\]

Applying integration by parts, the tail expectation can be computed using

\[
\int_y^b y f(y) \, dy = \begin{cases} 
\frac{b e^{\lambda y} - ye^{\lambda y} - \frac{1}{\lambda} (e^{\lambda b} - e^{\lambda y})}{e^{\lambda a} - e^{\lambda b}}, & (\lambda \neq 0) \\
\frac{q b^2 - y^2}{2 \cdot b-a}, & (\lambda = 0)
\end{cases}
\]

Putting all the pieces together, we obtain the following closed-form expressions for the CDF and tail expectation. (We assume \( \lambda_k \neq 0 \) for simplicity, and we use the notation \( y \vee t = \max \{y, t\} \).)

\[
1 - F(y) = \int_y^\infty f(y) \, dy = \sum_{k=1}^K 1(y < t_{k-1}) q_k \frac{e^{\lambda_k t_{k-1} - e^{\lambda_k (y \vee t_k)}}}{e^{\lambda_k t_{k-1} - e^{\lambda_k t_k}}},
\]

\[
\int_y^\infty y f(y) \, dy = \sum_{k=1}^K 1(y < t_{k-1}) q_k \frac{(t_{k-1} - 1/\lambda_k) e^{\lambda_k t_{k-1} - e^{\lambda_k t_k}} - ((y \vee t_k) - 1/\lambda_k) e^{\lambda_k y - e^{\lambda_k t_k}}}{e^{\lambda_k t_{k-1} - e^{\lambda_k t_k}}}. 
\]
4 Simulation studies

We conduct three simulation studies that examine the performance of our proposed ME method relative to existing methods.

4.1 Density estimation

We first present density estimates from one large simulation draw with the sample size comparable to those in our empirical data. We consider four typical models for income distribution, namely lognormal, gamma, Weibull, and double Pareto. In each case, we choose the scale parameter so that the population mean normalizes to 1. Table 2 summarizes these distributions.

| Name       | Density $f(y)$                                                                 | Normalization | Parameter(s)   |
|------------|-------------------------------------------------------------------------------|---------------|----------------|
| Lognormal  | $\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\log y - \mu)^2}{2\sigma^2}}$         | $\mu = -\sigma^2/2$ | $\sigma = 1.5$ |
| Gamma      | $\frac{b^a}{\Gamma(a)}y^{a-1}e^{-by}$                                         | $b = a$       | $a = 0.5$      |
| Weibull    | $bk^k y^{k-1}e^{-by}$                                                         | $b = \Gamma(1 + 1/k)^k$ | $k = 0.7$      |
| Double Pareto | $\begin{cases} \frac{\alpha \beta}{\alpha + \beta} M^{-\beta} y^{\beta-1} & (y \leq M) \\ \frac{\alpha \beta}{\alpha + \beta} M^\alpha y^{-\alpha-1} & (y \geq M) \end{cases}$ | $M = \frac{(\beta+1)(\alpha-1)}{\alpha \beta}$ | $(\alpha, \beta) = (1.5, 0.5)$ |

The simulation design is as follows. For each model, we generate a random sample $\{Y_i\}_{i=1}^n$ with size $n = 10^7$. Such a large $n$ is coherent with the number of taxpayers in our empirical data set; see Table 1. We set the top fractiles to $(p_k)_{k=1}^K = (0.001, 0.005, 0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99, 0.995, 0.999, 1)$ (so $K = 26$) and define the $k$-th threshold $t_k$ as the top $p_k$-th quantile of $\{Y_i\}_{i=1}^n$. We then estimate the ME density $\hat{f}$ as in Section 3.

Figure 1 shows the population and estimated densities for each model. Because the population densities are skewed, for visibility we plot the density of $\log Y$. In each case, the two densities $f$ and $\hat{f}$ are nearly identical.

4.2 Top income shares

Next, we estimate the top income shares, which correspond to the Lorenz curve flipped along the 45 degree line. There are many existing methods for estimat-
ing the Lorenz curve as discussed in Section 1. We implement those proposed by Kakwani and Podder (1976) (henceforth KP) and Villaseñor and Arnold (1989) (henceforth VA), both of which have been used by the World Bank. In addition, we implement a more recently developed method by Hajargasht, Griffiths, Brice, Rao, and Chotikapanich (2012) (henceforth HGBRC), which has been further extended by Chen (2018) and Hajargasht and Griffiths (2020). The KP and VA methods impose some parametric assumptions on the Lorenz curve and essentially run linear regressions of the group mean ($\hat{y}_k$ in our notation) on some transformation of the proportion of each group ($\hat{q}_k$ in our notation). The HGBRC method imposes some parametric assumptions on the underlying density and constructs a generalized method of moments estimation. Regarding KP, we implement their Method III as described in their Section 4. Regarding VA, we implement their method with $a = 1$ and $d = 0$ as described in their Section 4. Regarding HGBRC, we adopt their assumption of the generalized beta distribution of the second kind (GB2, McDonald, 1984) and the diagonal weighting matrix as proposed by Chotikapanich, Griffiths, and Rao (2007).

Our data generating process is as follows. We suppose that the population
distribution is double Pareto with parameters $\alpha = 2.3$, $\beta = 1.1$, and $M = 1$ (normalization). There are two reasons for using the double Pareto distribution with these parameters. First, this distribution has been shown to fit the income distribution very well; see for instance Toda (2012, Fig. 1(a)). Second, unlike other parametric distributions used in Figure 1, the double Pareto distribution admits a closed-form CDF and Lorenz curve as discussed in Appendix D, which is convenient for numerical evaluation. Appendix F considers other distributions.

We treat the population top $0.1, 1, 5, 10, \ldots, 95, 100$ percentiles as the observed thresholds ($K = 22$) and compute the population top income shares. Next, we generate random samples with sizes $n = 10^4, 10^5, 10^6$ from the population distribution and record the proportions of observations and their average incomes within each group, which we treat as our data.

Implementing our proposed ME, the KP, the VA, and the HGBRC methods, we report their relative bias and relative root mean squared error (RMSE) for the income shares of the top $p_0$ fractile with $p_0 \in \{0.001, 0.01, 0.05, 0.1, 0.2, \ldots, 0.9\}$. More specifically, let $s_0$ denote the true top income share and $\hat{s}_m$ the estimator in the $m$-th simulation draw with $m \in \{1, \ldots, M\}$. We define the relative bias and RMSE by

$$\text{Relative Bias} = \frac{1}{M} \sum_{m=1}^{M} (\hat{s}_m/s_0 - 1),$$

$$\text{Relative RMSE} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (\hat{s}_m/s_0 - 1)^2},$$

respectively. Table 3 presents the results based on $M = 1,000$ simulations.

The findings can be summarized as follows. First, our proposed ME method performs very well in terms of both bias and RMSE. They decrease as $n$ increases and are smaller than those of the other three methods for most of the $(n, p_0)$ combinations, especially the bias. Second, the KP and the VA methods both impose some parametric assumptions on the Lorenz curve and hence implicitly on the underlying density. In particular, the VA method imposes that the Lorenz curve is a part of an ellipse. This assumption implies that the underlying density $f(y)$ (after a location- and scale-transformation) is proportional to $(1 + y^2/2)^{-3/2}$.

---

2Since the logarithm of a double Pareto random variable is Laplace, which is double exponential, we can generate a double Pareto random variable using $\log(Y/M) = \frac{1}{\alpha}X_1 - \frac{1}{\beta}X_2$, where $X_1, X_2$ are independent exponential random variables with parameter 1. Therefore $Y = MU_1^{-1/\alpha}U_2^{1/\beta}$, where $U_1, U_2$ are independent uniform random variables on $[0, 1]$. 

---

15
Table 3: Relative bias for top income shares.

| Relative Bias |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|---------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
|               | 0.001| 0.01 | 0.05 | 0.1  | 0.2  | 0.3  | 0.4  | 0.5  | 0.6  | 0.7  | 0.8  | 0.9  |      |      |
| n             |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 10^4          | 0.440| 0.133| 0.053| 0.033| 0.019| 0.012| 0.009| 0.006| 0.004| 0.001| 0.001| 0.000|      |      |
| 10^5          | 0.136| 0.042| 0.017| 0.010| 0.006| 0.004| 0.003| 0.002| 0.001| 0.001| 0.000| 0.000|      |      |
| 10^6          | 0.046| 0.014| 0.005| 0.003| 0.002| 0.001| 0.001| 0.000| 0.000| 0.000| 0.000| 0.000|      |      |
| ME            |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| n             |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 10^4          | 0.427| 0.370| 0.062| 0.130| 0.102| 0.035| 0.021| 0.037| 0.031| 0.014| 0.002| 0.005|      |      |
| 10^5          | 0.357| 0.346| 0.029| 0.123| 0.099| 0.032| 0.021| 0.037| 0.031| 0.014| 0.002| 0.005|      |      |
| 10^6          | 0.348| 0.343| 0.024| 0.123| 0.099| 0.032| 0.021| 0.037| 0.031| 0.014| 0.002| 0.005|      |      |
| KP            |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| n             |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 10^4          | 0.268| 0.079| 0.051| 0.048| 0.034| 0.022| 0.041| 0.036| 0.020| 0.005| 0.002| 0.002|      |      |
| 10^5          | 0.254| 0.052| 0.023| 0.030| 0.022| 0.013| 0.039| 0.035| 0.019| 0.005| 0.002| 0.002|      |      |
| 10^6          | 0.253| 0.049| 0.014| 0.008| 0.005| 0.004| 0.008| 0.005| 0.003| 0.001| 0.000| 0.000|      |      |
| VA            |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| n             |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 10^4          | 0.233| 0.137| 0.071| 0.044| 0.027| 0.015| 0.011| 0.009| 0.005| 0.002| 0.001| 0.000|      |      |
| 10^5          | 0.093| 0.049| 0.024| 0.019| 0.021| 0.011| 0.006| 0.005| 0.003| 0.001| 0.000| 0.000|      |      |
| 10^6          | 0.054| 0.026| 0.011| 0.015| 0.020| 0.011| 0.005| 0.005| 0.002| 0.001| 0.000| 0.000|      |      |
| HGBRC         |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| n             |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 10^4          | -0.353| -0.348| 0.025| 0.125| 0.101| 0.033| -0.020| -0.036| -0.031| -0.014| 0.002| 0.005|      |      |
| 10^5          | -0.349| -0.343| 0.023| 0.123| 0.099| 0.032| -0.020| -0.037| -0.031| -0.014| 0.002| 0.005|      |      |
| 10^6          | -0.347| -0.342| 0.023| 0.123| 0.099| 0.032| -0.021| -0.037| -0.031| -0.014| 0.002| 0.005|      |      |

Note: Relative Bias and RMSE for the income shares of the top \( p_0 \) fractile. ME denotes our proposed method. KP, VA, and HGBRC denote the methods proposed respectively by Kakwani and Podder (1976), Villaseñor and Arnold (1989), and Hajargasht et al. (2012). See the main text for the data generating process. The results are based on 1,000 simulation draws.

which is the Student \( t \) distribution with two degrees of freedom (Villaseñor and Arnold, 1989, Theorem 4). The KP method introduces a new coordinate system and imposes another parametric form on the Lorenz curve. The implied density is still parametric but does not have a closed-form expression. The HGBRC method assumes the GB2 density, which has Pareto upper and lower tails. Therefore, its performance is substantially better than those of KP and VA. In summary, these
and any other parametric assumptions could lead to large bias and RMSE caused by misspecification, which do not decrease with \( n \).

### 4.3 Comparison to kernel estimator

Finally, we compare our proposed ME estimator with the nonparametric kernel estimator proposed by Blower and Kelsall (2002). Given a bandwidth \( h \), define \( K_h(u) \) as the PDF of the normal distribution with mean zero and variance \( h^2 \), that is,

\[
K_h(u) = \frac{1}{\sqrt{2\pi h}} \exp \left( -\frac{u^2}{2h^2} \right).
\]

We implement Blower and Kelsall (2002, eq.(1.4)) by constructing the density estimator

\[
\hat{f}_{BK}(y) = \sum_{k=2}^{K} \hat{q}_k \int_{I_k} K_h(s - y) \hat{f}_0(s) \frac{ds}{\int_{I_k} f_0(s) ds},
\]

where \( \hat{f}_0(s) \) is the histogram estimator

\[
\hat{f}_0(y) = \sum_{k=2}^{K} \frac{\hat{q}_k}{t_k - t_{k-1}} \mathbb{1}_{I_k}(y).
\]

Using the fact that \( \int_{I_k} \hat{f}_0(s) ds = \hat{q}_k \) and \( K_h \) is the normal density, we can simplify \( \hat{f}_{BK} \) as follows:

\[
\hat{f}_{BK}(y) = \sum_{k=2}^{K} \int_{I_k} K_h(s - y) \hat{f}_0(s) ds = \sum_{k=2}^{K} \frac{\hat{q}_k}{t_{k-1} - t_k} \int_{I_k}^{t_{k-1}} K_h(s - y) ds
\]

\[
= \sum_{k=2}^{K} \frac{\hat{q}_k}{t_{k-1} - t_k} \left( \Phi \left( \frac{t_{k-1} - y}{\hat{h}} \right) - \Phi \left( \frac{t_k - y}{\hat{h}} \right) \right),
\]

where \( \Phi \) denotes the CDF of the standard normal distribution.

Blower and Kelsall (2002) do not derive any asymptotic properties of this estimator nor theoretical requirements on the choice of the bandwidth. We implement a variety of choices of \( h \) to examine its finite sample performance. Specifically, we use the rule-of-thumb choice \( h = c\hat{\sigma}n^{-1/5} \), where \( \hat{\sigma} \) is the sample standard deviation based on individual observations (which is in principle infeasible given the tabulated data) and \( c \in \{0.1, 0.5, 1.0, 1.5\} \) is some constant.

Figure 2 presents the relative RMSE for the density estimators when the data generating process is double Pareto, lognormal, and gamma as in Section 4.2 with sample size \( n = 10^4, 10^5, 10^6 \). Although these figures are not necessarily easy to
read, the RMSEs for the ME (BK) estimator are indicated with solid (dashed) lines. As is clear from this figure, the RMSEs for the ME estimator is generally closest to the horizontal axis uniformly across quantiles, so the performance of our proposed ME method is outstanding. In addition, we also implement the kernel estimator studied by Reyes et al. (2016). Its performance is substantially worse than that proposed by Blower and Kelsall (2002) and hence not reported.

Figure 2: Relative RMSE of density estimators.

Note: The figure presents the relative RMSE for the density estimators when the data generating process is double Pareto, lognormal, and gamma with sample size $n = 10^4, 10^5, 10^6$. ME denotes our proposed maximum entropy method, whereas BK denotes the kernel estimator of Blower and Kelsall (2002) with various bandwidth parameters $c$.

5 Income distribution in the United States

We consider two empirical applications of our method. First, we estimate the distribution of U.S. income distribution for particular years. Second, we estimate the top income shares (including mid-sample) over the past century.
5.1 Income distribution in 1946 and 2019

We estimate the distribution of U.S. adjusted gross income (AGI) in 1946 and 2019. We choose 2019 because it is the most recent year for which data is available. Before World War II, because only a small fraction of the population filed for taxes, the tax returns data is not representative for the population.\(^3\) For this reason, we choose 1946 because it is one of the earliest years for which the tax returns data is representative for the population. Note that unlike in recent years, the tabulated summary data set is almost the only publicly available income data set in early years such as 1946.

To make the results comparable across years, we measure income in 2019 dollars by adjusting with the Consumer Price Index (CPI). The number of income groups is \(K = 48\) for 1946 and \(K = 18\) for 2019. Figure 3a shows the ME density estimates \(\hat{f}(e^x) e^x\) of log income \(x = \log y\) in a semi-log scale.

We can summarize the findings as follows. First, for each year the log income distribution is bell-shaped but slightly asymmetric. Second, observe that although \(\hat{f}\) is piecewise exponential, it is not necessarily continuous at the bin thresholds as can be seen from the spikes in the 1946 density. Third, the 2019 density is more spread-out than 1946, which suggests that income inequality has increased. Finally, Figure 3b shows the tail probability \(1 - \hat{F}(y)\) in a log-log scale, which is continuous. The fact that the 2019 tail probability is higher than 1946 implies that the 2019 (real) income distribution first-order stochastically dominates the 1946 one, possibly due to economic growth. The graphs also show a straight-line pattern for high incomes, which is consistent with a Pareto upper tail documented elsewhere; see for instance de Vries and Toda (2022) and the references therein. Because the slope is steeper for 1946 than in 2019, the income Pareto exponent is smaller (top income inequality is higher) in 2019.

Because the ME density \(\hat{f}\) is piecewise exponential, which is analytically tractable, it is straightforward to compute statistics such as top income shares; see Section 3.3. Figure 4 shows the top income shares, both in original and log-log scales. Figure 4a (original scale) is essentially the Lorenz curve flipped along the 45 degree line. The fact that the 2019 curve is above the 1946 one suggests that income inequality has increased. The straight-line pattern in log-log scale (Figure 4b) is consistent with a Pareto upper tail.

\(^3\)The fraction of tax filers among potential tax units has been stable at around 80–90% postwar but in the range of 1–20% before 1940; see the discussion in Piketty and Saez (2003).
5.2 Top income shares over the century

Finally, we apply the proposed method to estimate the top $p$ fractile income share for various values of $p \in [0, 1]$. To construct the top income shares, we use the following approach. First, we collect the tabulated summaries of income similar to Table 1 for each year from the IRS Statistics of Income.\footnote{See the appendix in Lee, Sasaki, Toda, and Wang (2022) for specific details.} These tables contain information on the number of tax units (an individual or a married couple with dependents if any) and their total income within each income group. As these tables contain only tax filers, we complement them with the total number of potential tax units and total income estimated by Piketty and Saez (2003).\footnote{We obtain the total number of tax units from the spreadsheet \url{https://eml.berkeley.edu/~saez/TabFig2018.xls}, Table A0, Column B, and total income from Column I.} We suppose that non-filers are low income households and thus do not affect the

\textit{Figure 3: U.S. income distributions in 1946 and 2019.}

(a) Probability density of log income. \hspace{2cm} (b) Log-log plot of tail probability.

\textit{Figure 4: U.S. top income shares in 1946 and 2019.}

(a) Original scale. \hspace{2cm} (b) Log-log scale.
calculation of the top $p$ fractile income share if $p$ is small enough. Because the fraction of tax filers among potential tax units exceeds 0.1 (0.8) since 1936 (1945), we construct the top $p$ fractile income share for $p \in \{0.01, 0.05, 0.1\}$ since 1936 and also for for $p \in \{0.2, 0.4, 0.6, 0.8\}$ since 1945. Figure 5 shows the results.

![Figure 5: Top income shares in U.S., 1936–2018.](image)

Note: From bottom to top, the figure plots the top $p$ fractile income share for $p \in \{0.01, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8\}$. Mid-sample fractiles ($p > 0.1$) are omitted pre-1945 because the fraction of tax filers is too small.

The top 1%, 5%, 10% income shares exhibit an inverse U-shaped pattern, which is well known. To the best of our knowledge, the top income shares for mid-sample fractiles (e.g., $p = 0.2, 0.4, 0.6, 0.8$) have not been reported in the previous literature. We find that the mid-sample top income shares exhibit an abrupt upward jump between 1986 and 1987. This could be due to the Tax Reform Act of 1986, which significantly altered the treatment of capital gains income.\(^6\)

We next compare our top income shares to those constructed by Piketty and Saez (2003).\(^7\) Figure 6 shows the top 1%, 5%, 10% income shares constructed in two ways. We find that post-1986, our top income shares are nearly identical to

\(^{6}\)According to Piketty and Saez (2001, p. 40), the fraction of capital gains income included in AGI was 100% until 1933, 70% in 1934–1937, 60% in 1938–1941, 50% in 1942–1978, 40% in 1979–1986, and 100% since 1987. Because high income earners tend to hold more financial asset (and generate more capital gains), the large exclusion of capital gains in 1934–1986 likely causes the top income shares to be biased downwards.

\(^{7}\)We obtain the top income shares from Table A3 in the spreadsheet [https://eml.berkeley.edu/~saez/TabFig2018.xls](https://eml.berkeley.edu/~saez/TabFig2018.xls). Piketty (2001, Appendix B, Section 1.1, pp. 592–599) provides the details of the method for constructing these top income shares. Since it is written in French, we describe the method in Appendix E for the convenience of the readers.
those from Piketty and Saez (2003), even though our method is nonparametric while their method is parametric (assuming a Pareto upper tail). This is likely because the upper tail of the income distribution can be well approximated by the Pareto distribution. However, there are large discrepancies between the two series pre-1986 because we have used the raw AGI without adjusting for the excluded capital gains income discussed in Footnote 6.

![Figure 6: Comparison to Piketty and Saez (2003) series.](image)

6 Concluding remarks

Existing standard nonparametric estimators of density and cumulative distribution functions require individual-level data. Even when individual-level information is difficult to access due to confidentiality concerns, tabulated summaries of such data are often publicly available. Administrative data of income are the leading examples. In this paper, we propose a novel method of maximum entropy density estimation from tabulated summary data and establish the strong uniform consistency of the density and cumulative distribution estimators. This method enjoys many desirable properties. First, the estimator is piecewise exponential, which is analytically tractable. Using its functional form, it is straightforward to compute statistics such as top income shares. Second, and more importantly, our estimator is free from tuning parameters unlike existing kernel-based methods, which is attractive in practice. This feature provides a complete theoretical justification that our proposed estimator works in practice, unlike the existing kernel-based estima-
tors for which the theory does not formally account for the effects of bandwidth choice in practice.

Appendix

Appendix A provides a smoothed version of the proposed ME estimator. Appendix B contains mathematical proofs of the strong uniform consistency. Appendix C solves some numerical issues in constructing the proposed estimator. Appendix D computes the Lorenz curve for the double Pareto distribution. Appendix E describes the details of Piketty (2003)'s Pareto interpolation method, which was originally written in French. Appendix F contains additional simulation studies.

A Smoothed ME estimation

Our estimator described in Section 3.1 in the main text is generally discontinuous at the bin thresholds $t_1, \ldots, t_K$. In practice, an ad hoc yet simple method to smooth the estimator is to replace it with some local polynomial function within small neighborhoods of $t_1, \ldots, t_K$. As a more systematic alternative, we can slightly shift the bin boundaries so that the implied ME estimator becomes continuous. To that end, we can treat $t_1, \ldots, t_K$ as arguments and further minimize the Kullback-Leibler divergence (3.7) over $t_1, \ldots, t_K$ as follows. Because the minimum threshold $t_K$ is unidentified, we set it to an arbitrary value $t_K < y_K$, say, $t_K = 0$. Define the set of admissible thresholds by

$$T = \{ (t_1, \ldots, t_{K-1}) \in \mathbb{R}^{K-1} : (\forall k) t_k < y_k < t_{k-1} \}$$

(A.1)

which is a nonempty open interval in $\mathbb{R}^{K-1}$. The following theorem shows that $J^*$ in (3.7) always achieves a unique minimum on $T$, which is continuous.

**Theorem 5.** Fix $t_K < y_K$ and let $J^* : T \to \mathbb{R}$ be defined by (3.7). Then $J^*$ is strictly convex and achieves a unique minimum $t^* \in T$. Furthermore, the corresponding maximum entropy density $f^*$ in (3.6) with $t = t^*$ is continuous on $[t_K, \infty)$.

The proof is deferred to Appendix B.
B Proofs

Proof of Proposition 1. To simplify notation, we suppress the dependence of \(J_k, \lambda_k^*,\) etc. on \(t\). We use Fenchel duality (Borwein and Lewis, 1991) to solve the maximum entropy problem. The Lagrangian of the maximum entropy problem subject to the moment conditions (3.2) is

\[
L(g, \mu, \lambda) = \int g \log g \, dy + \sum_{k=1}^{K} \left( \mu_k \left( q_k - \int 1_{I_k} g \, dy \right) + \lambda_k \left( q_k y_k - \int 1_{I_k} yg \, dy \right) \right)
\]

where \(\mu_k\) and \(\lambda_k\) are the Lagrange multipliers corresponding to the moment conditions (3.2a) and (3.2b). The dual objective function is the minimum of \(L\) over unconstrained \(g \in L^1(I)\). Taking the first order condition (Gâteaux derivative) pointwise, for \(y \in I_k\) we obtain

\[
0 = \log g + 1 - (\mu_k + \lambda_k y) \iff g(y) = e^{\mu_k + \lambda_k y - 1}. \tag{B.1}
\]

Hence the dual objective function becomes

\[
J(\mu, \lambda) := \min_{g \in L^1_+(I)} L(g, \mu, \lambda) = \sum_{k=1}^{K} q_k (\mu_k + y_k \lambda_k) - \sum_{k=1}^{K} \int_{I_k} e^{\mu_k + \lambda_k y - 1} \, dy.
\]

The dual problem maximizes \(J\) with respect to \(\mu, \lambda\), which is a concave maximization problem. The first order condition with respect to \(\mu_k\) is

\[
0 = \frac{\partial J}{\partial \mu_k} = q_k - \int_{I_k} e^{\mu_k + \lambda_k y - 1} \, dy \iff \mu_k = 1 + \log q_k - \log \left( \int_{I_k} e^{\lambda_k y} \, dy \right). \tag{B.2}
\]

Then (with a slight abuse of notation) the objective function becomes

\[
J(\lambda) := \max_{\mu \in \mathbb{R}^K} J(\mu, \lambda) = \sum_{k=1}^{K} \left( q_k y_k \lambda_k + q_k \log q_k - q_k \log \left( \int_{I_k} e^{\lambda_k y} \, dy \right) \right)
\]

\[
= \sum_{k=1}^{K} q_k \left( y_k \lambda_k + \log q_k - \log \left( \int_{I_k} e^{\lambda_k y} \, dy \right) \right)
\]

\[
= \sum_{k=1}^{K} q_k (J_k(\lambda_k) + \log q_k), \tag{B.3}
\]

24
where \( J_k \) is given by (3.4). Since \( J \) is additively separable, it suffices to maximize \( J_k \) for each \( k \).

Let us now show that \( J_k(\lambda) \) achieves a unique maximum over \( \lambda \in \mathbb{R} \). It is straightforward to show the strict concavity of \( J_k \) by applying Hölder’s inequality to the function \( \lambda \mapsto \log \left( \int_{I_k} e^{\lambda y} \, dy \right) \). As \( \lambda \to \pm \infty \), it follows from (3.4) that

\[
J_k(\lambda) = y_k \lambda - \log \left( \frac{e^{\lambda y_k} - e^{\lambda t_k}}{e^{\lambda y_k} - e^{\lambda t_k}} \right) \to -\infty
\]

because \( t_k < y_k < t_{k-1} \). Since \( J_k \) is continuous and strictly concave on its domain, it achieves a unique maximum \( \lambda_k \in \mathbb{R} \). When \( k = 1 \), we have \( J_1(\lambda) = -\infty \) for \( \lambda \leq 0 \), and analytically maximizing \( J_1 \) for \( \lambda > 0 \), we obtain the expression for \( \lambda_1^* \) in (3.5). Differentiating \( J_k \) under the integral sign, we obtain

\[
J'_k(\lambda) = y_k - \frac{\int_{I_k} y e^{\lambda y} \, dy}{\int_{I_k} e^{\lambda y} \, dy} \implies J'_k(0) = y_k - \frac{\int_{I_k} y \, dy}{\int_{I_k} \, dy} = y_k - \frac{t_k + t_{k-1}}{2},
\]

so \( \lambda_k^* \) can be signed as in (3.5).

If \( \lambda_k^* \neq 0 \), using (B.1) and (B.2) and suppressing the asterisks, we obtain the solution

\[
f^*(y) = g(y) = e^\mu y + \lambda_k y^{-1} = \frac{q_k \lambda_k e^{\lambda_k y}}{e^{\lambda_k t_k} - e^{\lambda_k t_k}},
\]

which is (3.6) and is exponential in \( y \). The derivation for \( \lambda_k^* = 0 \) is similar. The minimum value (3.7) follows from (B.3).

\[\Box\]

**Lemma 6**. Define the function \( \phi : (0, \infty) \to \mathbb{R} \) by

\[
\phi(x) = \frac{\cosh x}{\sinh x} - \frac{1}{x},
\]

where \( \cosh x = \frac{e^x + e^{-x}}{2} \) and \( \sinh x = \frac{e^x - e^{-x}}{2} \). Then \( \phi'(x) > 0, \phi''(x) < 0, \phi(0+) = 0, \phi(\infty) = 1 \), and \( \phi'(0+) = 1/3 \).

**Proof.** To simplify notation, let \( c(x) = \cosh x \) and \( s(x) = \sinh x \). \( \phi(\infty) = 1 \) is trivial since \( c(x) \sim e^x/2 \sim s(x) \) as \( x \to \infty \). Using the definition of \( c \) and \( s \), we have \( c' = s, s' = c, \) and \( c^2 - s^2 = 1 \). Taking the derivative of \( \phi \), we have

\[
\phi'(x) = \frac{c' s - c s'}{s^2} + \frac{1}{x^2} = \frac{s^2 - c^2}{s^2} + \frac{1}{x^2} = -\frac{1}{s^2} + \frac{1}{x^2}.
\]

Noting that \( s(x) = x + x^3/6 + o(x^3) \) as \( x \to 0 \) by Taylor’s theorem, a straightforward
calculation yields $\phi'(0+) = 1/3$.

To show $\phi' > 0$, it suffices to show $s > x$. To this end, define $g(x) = s(x) - x$. Then $g(0) = 0$ and

$$g'(x) = s'(x) - 1 = c(x) - 1 = \frac{e^x + e^{-x}}{2} - 1 = \frac{(e^{x/2} - e^{-x/2})^2}{2} > 0$$

for $x > 0$, so $g(x) > 0$ for $x > 0$. This shows $\phi'(x) > 0$. Taking the derivative once again, we have

$$\phi''(x) = \frac{2s'}{s^3} - \frac{2}{x^3} = 2 \left( \frac{c}{s^3} - \frac{1}{x^3} \right).$$

Therefore to show $\phi'' < 0$, it suffices to show $s^3/c > x^3$. To this end, define $h(x) = s(x)^3/c(x) - x^3$. Then

$$h'(x) = \frac{3s^2sc' - s^3c'}{c^2} - 3x^2 = \frac{3s^2c^2 - s^4}{c^2} - 3x^2$$

$$= 3(c^2 - 1) - \frac{(c^2 - 1)^2}{c^2} - 3x^2 = 2c^2 - 1 - \frac{1}{c^2} - 3x^2,$$

$$h''(x) = 4cc' + \frac{2c'}{c^3} - 6x = 4sc + \frac{2s}{c^3} - 6x,$$

$$h'''(x) = 4s'c + 4sc' + \frac{2s'}{c^3} - \frac{3}{c^4} - \frac{6s^2}{c^2} - 6 = 4c^2 + \frac{2c^2 - 6s^2}{c^4} + 6$$

$$= 8c^2 - 10 - \frac{4}{c^2} + \frac{6}{c^4} = \frac{2}{c^4} (c^2 - 1)^2 (4c^2 + 3) > 0$$

for $x > 0$, where the last inequality follows from $c > 1$ for $x > 0$. Noting that $h'''(0) = h''(0) = h'(0) = h(0) = 0$, it follows that $h(x) > 0$ for $x > 0$, implying $s^3/c > x^3$ and hence $\phi'' < 0$. $\square$

Below, extend the domain of $\phi$ in (B.4) to the entire real line by setting $\phi(0) = 0$ and $\phi(x) = -\phi(-x)$ for $x < 0$. Then $\phi : \mathbb{R} \to (-1, 1)$, $\phi$ is strictly increasing, and $\phi$ is concave (convex) for $x > 0$ ($x < 0$). Figure 7 shows its graph.

We need the following lemma to prove Theorem 2.

Lemma 7. Let $a < y < b$. Then the function

$$J(\lambda) := \begin{cases} y\lambda - \log \left( \frac{e^{y\lambda} - e^{x\lambda}}{\lambda} \right) & (\lambda \neq 0) \\ - \log(b - a) & (\lambda = 0) \end{cases}$$

is strictly concave in $\lambda \in \mathbb{R}$ and achieves a unique maximum $\lambda^*$. Furthermore,
letting \( c = (a + b)/2 \) and \( d = b - a > 0 \), we have

\[
\lambda^* = \frac{2}{d} \phi^{-1} \left( \frac{2(y - c)}{d} \right).
\]

(B.5)

**Proof.** The strict concavity of \( J \) and the existence and uniqueness of \( \lambda^* \) follow from Proposition 1. If \( y = c \), then by (3.5) we have \( \lambda^* = 0 \) and (B.5) is trivial. Suppose \( y \neq c \). Then by (3.5) we have \( \lambda^* \geq 0 \) according as \( y \geq c \).

When \( \lambda \neq 0 \), by definition

\[
J(\lambda) = (y - c)\lambda - \log \left( \frac{e^{\lambda d/2} - e^{-\lambda d/2}}{\lambda} \right).
\]

Since \( J \) is differentiable and \( \lambda^* \) maximizes \( J \), taking the derivative we obtain

\[
0 = J'(\lambda^*) = y - c + \frac{1}{\lambda^*} - \frac{d \cosh(\lambda^* d/2)}{2 \sinh(\lambda^* d/2)} \iff \frac{\cosh x}{\sinh x} - \frac{1}{x} = \frac{2(y - c)}{d},
\]

where \( x = \lambda^* d/2 \). Noting that \( y \in (a, b) \), \( c = (a + b)/2 \), and \( d = b - a \), we have \( 2(y - c)/d \in (-1, 1) \). Therefore (B.5) follows from the definition of \( \phi \) and Lemma 6.

**Proof of Theorem 2.** We divide our proof into four steps.

**Step 1. Reduction to the case when \( C \) is an interval.**

Since \( f \) is Lipschitz, it is continuous, so the domain \( D = \{ y \in \mathbb{R} : f(y) > 0 \} \) is
open. Let \( C \subset D \) be compact. For each \( y \in C \) we can take an open interval \((a_y, b_y)\) with \( y \in (a_y, b_y) \) and \([a_y, b_y] \subset D\). Since \( C \) is compact and \( C \subset \bigcup_{y \in C} (a_y, b_y) \), we can take a finite subcover \( C \subset \bigcup_{j=1}^J (a_{y_j}, b_{y_j}) \). Therefore \( C \subset \bigcup_{j=1}^J [a_{y_j}, b_{y_j}] \subset D \), so the uniform consistency on \( C \) follows if we show the uniform consistency on each compact interval \([a, b] \subset D\).

Below, to simplify the argument, fix \( C = [a, b] \) and \( C' = [a', b'] \) such that \( C \subset C' \subset D \) and \( a' < a < b < b' \). By relabeling the thresholds \( \{t_k\}_{k=1}^K \) if necessary, without loss of generality we may assume \( a' < t_K < a < b < t_1 < b' \) by Condition (iii).

**Step 2.** Let \( r_3 := \frac{1-4r_1}{2} > 0 \). As \( n \to \infty \), we have

\[
\max_{2 \leq k \leq K} \left| \frac{\hat{q}_k - q_k}{(t_{k-1} - t_k)^2} \right| = o_{a.s.}(1) = O_p(n^{-r_3}), \tag{B.6a}
\]

\[
\max_{2 \leq k \leq K} \left| \frac{\hat{y}_k - y_k}{t_{k-1} - t_k} \right| = o_{a.s.}(1) = O_p(n^{-r_3}). \tag{B.6b}
\]

Let \((\Omega, \mathcal{F}, P)\) be the probability space over which the random variables \( \{Y_i\}_{i=1}^\infty \) are defined. For any \( \eta > 0 \), define the event

\[
E_n := \left\{ \omega \in \Omega : \max_{2 \leq k \leq K} \left| \frac{\hat{q}_k - q_k}{(t_{k-1} - t_k)^2} \right| > \eta \right\}.
\]

Let \( I_{t_k} \) denote the indicator function of \( I_k \). Using \( \sum_{i=1}^n I_{t_k}(Y_i) = n_k - n_{k-1} = n\hat{q}_k \), we obtain

\[
P(E_n) \leq \sum_{k=2}^K \mathbb{P} \left( \left| \frac{\hat{q}_k - q_k}{(t_{k-1} - t_k)^2} \right| > \eta \right) = \sum_{k=2}^K \mathbb{P} \left( \frac{1}{n} \sum_{i=1}^n I_{t_k}(Y_i) - q_k \left| \frac{1}{n} \sum_{i=1}^n \frac{I_{t_k}(Y_i) - q_k}{(t_{k-1} - t_k)^2} \right| > \eta \right) \leq \sum_{k=2}^K \mathbb{P} \left( \sum_{i=1}^n (I_{t_k}(Y_i) - q_k) > \eta c_1^2 n^{1-2r_1} \right),
\]

where the last line follows from (3.8). Applying Hoeffding (1963)’s inequality to the last term, we obtain

\[
P(E_n) \leq \sum_{k=2}^K 2 \exp(-2c_1^4 \eta^2 n^{2-4r_1}/n) = 2(K-1) \exp(-2c_1^4 \eta^2 n^{1-4r_1}). \tag{B.7}
\]

Since \( 1 - 4r_1 > 0 \) by Assumption 1(iii), it follows that \( \sum_{n=1}^\infty P(E_n) < \infty \). By
the Borel-Cantelli lemma, we have

$$P(E_n \text{ infinitely often}) = P \left( \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} E_n \right) = 0.$$ 

Since $\eta > 0$ is arbitrary, we obtain $\max_{2 \leq k \leq K} \left| \frac{\tilde{q}_k - q_k}{(t_{k-1} - t_k)^2} \right| \to 0$ almost surely as $n \to \infty$.

To show that this term is $O_p(n^{-r_3})$ for $r_3 = \frac{1 - 4r_1}{2}$, for any $\epsilon > 0$, choose $\delta > 0$ such that $2(K - 1) \exp(-2\epsilon^2 \delta^2) < \epsilon$. Then the definition of $E_n$ and (B.7) imply

$$P \left( n^{-\frac{1 - 4r_1}{2}} \max_{2 \leq k \leq K} \left| \frac{\tilde{q}_k - q_k}{(t_{k-1} - t_k)^2} \right| > \delta \right) < \epsilon$$

for all $n$, so by the definition $r_3$ and that of the order in probability, we obtain (B.6a).

We next estimate the term $\max_{2 \leq k \leq K} \left| \frac{\tilde{q}_k - q_k}{(t_{k-1} - t_k)^2} \right|$. Using

$$\sum_{i=1}^{n} Y_i I_{I_k}(Y_i) = S_{n_k} - S_{n_{k-1}} = (n_k - n_{k-1})\hat{y}_k = n\hat{y}_k \hat{y}_k$$

and noting that $Y_i \in C'$ (compact set) whenever $1_{I_k}(Y_i) > 0$, a similar calculation yields

$$\max_{2 \leq k \leq K} \left| \frac{\tilde{q}_k \hat{y}_k - q_k y_k}{(t_{k-1} - t_k)^2} \right| = O_p(n^{-r_3}) = o_{a.s.}(1). \quad (B.8)$$

Using the triangle inequality and $\hat{y}_k \in I_k \subset [t_k, t_1] \subset [a', b']$, we obtain

$$q_k |\hat{y}_k - y_k| \leq |\hat{q}_k \hat{y}_k - q_k y_k| + |\hat{y}_k| |\hat{q}_k - q_k| \leq |\hat{q}_k \hat{y}_k - q_k y_k| + b' |\hat{q}_k - q_k|. \quad (B.9)$$

Assumption 1(ii) implies that we can take $0 < m \leq M < \infty$ such that $f : C' \to [m, M]$. By the mean value theorem for integrals, for each $k \in \{2, \ldots, K\}$, there exists some $u_k \in [t_k, t_{k-1})$ such that

$$\frac{q_k}{t_{k-1} - t_k} = \frac{f_{t_k}^{t_{k-1}} f(y) \, dy}{t_{k-1} - t_k} = f(u_k) \in [m, M]. \quad (B.10)$$

Therefore dividing (B.9) by $(t_{k-1} - t_k)^2$ and using (B.10), we obtain

$$m \left| \frac{\hat{y}_k - y_k}{t_{k-1} - t_k} \right| \leq \left| \frac{\hat{q}_k \hat{y}_k - q_k y_k}{(t_{k-1} - t_k)^2} \right| + b' \left| \frac{\hat{q}_k - q_k}{(t_{k-1} - t_k)^2} \right|$$

for each $k \in \{2, \ldots, K\}$. Taking the maximum over $k$, dividing by $m > 0$, and
Letting $n \to \infty$, we obtain (B.6b).

**Step 3.** Let $\bar{t}_k = (t_k + t_{k-1})/2$ be the midpoint of $I_k$, $m = \inf_{y \in C'} f(y) > 0$, and $L > 0$ be the Lipschitz constant for $f$. Then

$$
\max_{2 \leq k \leq K} \left| \frac{y_k - \bar{t}_k}{(t_{k-1} - t_k)^2} \right| \leq \frac{L}{6m}. \tag{B.11}
$$

Since by Assumption 1(ii) $f$ is Lipschitz continuous with constant $L > 0$, it is absolutely continuous. Theorem 3.35 of Folland (1999) implies that $f$ is almost everywhere differentiable with $|f'| \leq L$ and we can apply the fundamental theorem of calculus (in particular, integration by parts). Therefore using $q_k = \int_{I_k} f(y) \, dy$ and $q_k y_k = \int_{I_k} y f(y) \, dy$, we obtain

$$
q_k(y_k - \bar{t}_k) = \int_{I_k} (y - \bar{t}_k) f(y) \, dy
\leq \frac{1}{2} (y - \bar{t}_k)^2 f(y) \bigg|_{t_k}^{t_{k-1}} - \int_{I_k} \frac{1}{2} (y - \bar{t}_k)^2 f'(y) \, dy
= \frac{1}{8} (t_{k-1} - t_k)^2 (f(t_{k-1}) - f(t_k)) - \int_{I_k} \frac{1}{2} (y - \bar{t}_k)^2 f'(y) \, dy.
$$

Therefore, for every $k \in \{2, \ldots, K\}$, we have

$$
q_k |y_k - \bar{t}_k| \leq \frac{1}{8} (t_{k-1} - t_k)^2 |f(t_{k-1}) - f(t_k)| + \int_{I_k} \frac{1}{2} (y - \bar{t}_k)^2 L \, dy
\leq \frac{1}{8} (t_{k-1} - t_k)^3 L + \frac{1}{24} (t_{k-1} - t_k)^3 L
= \frac{L}{6} (t_{k-1} - t_k)^3.
$$

Dividing both sides by $q_k (t_{k-1} - t_k)^2$ and using (B.10), we obtain

$$
\left| \frac{y_k - \bar{t}_k}{(t_{k-1} - t_k)^2} \right| \leq \frac{L}{6m}
$$

uniformly over $k \in \{2, \ldots, K\}$, which shows (B.11).

**Step 4.** Strong consistency and the convergence rate of the estimator $\hat{f}(y)$ for $f(y)$ uniformly over $y \in C$.

Letting $I_k = [t_k, t_{k-1})$, by assumption we have $C \subset \bigcup_{k=2}^{K} I_k \subset C'$. Therefore for any $y \in C$, we can take $k$ such that $y \in I_k$. Using the moment conditions and

30
the mean value theorem for the integrals, we can take \( u(y), v(y) \in I_k \) such that

\[
\hat{q}_k = \int_{I_k} \hat{f}(y) \, dy = (t_{k-1} - t_k) \hat{f}(u(y)), \tag{B.12a}
\]

\[
q_k = \int_{I_k} f(y) \, dy = (t_{k-1} - t_k) f(v(y)). \tag{B.12b}
\]

Therefore,

\[
\left| \hat{f}(y) - f(y) \right| \leq \left| \hat{f}(y) - \hat{f}(u(y)) \right| + \left| \hat{f}(u(y)) - f(v(y)) \right| + |f(v(y)) - f(y)| =: A_1(y) + A_2(y) + A_3(y). \tag{B.13}
\]

Using Assumption 1(ii), \( y, v(y) \in I_k \), and Assumption 1(iii), we can uniformly bound \( A_3 \) as

\[
\sup_{y \in C} A_3(y) = \sup_{y \in C} \left| f(v(y)) - f(y) \right| \leq \sup_{y \in C} L \left| v(y) - y \right| \leq L \max_{2 \leq k \leq K} (t_{k-1} - t_k) \leq LC_2 n^{-r_2}. \tag{B.14}
\]

Using (B.12), Step 2, and Assumption 1(iii), we can uniformly bound \( A_2 \) as

\[
\sup_{y \in C} A_2(y) = \max_{2 \leq k \leq K} \left| \hat{q}_k - q_k \right| \leq c_2 n^{-r_2} \max_{2 \leq k \leq K} \left| \frac{\hat{q}_k - q_k}{(t_{k-1} - t_k)^2} \right| = c_2 n^{-r_2} O_p(n^{-r_3}). \tag{B.15}
\]

Noting that \( \hat{f} \) is exponential on \( I_k \) and hence monotonic, it follows from the functional form in (3.6) and the mean value theorem that

\[
\sup_{y \in C} A_1(y) = \sup_{y \in C} \left| \hat{f}(y) - \hat{f}(u(y)) \right| \leq \max_{2 \leq k \leq K} \left| \hat{f}(t_{k-1}) - \hat{f}(t_k) \right| \leq \max_{2 \leq k \leq K} \hat{q}_k |\lambda_k| = \max_{2 \leq k \leq K} \frac{\hat{q}_k}{t_{k-1} - t_k} \max_{2 \leq k \leq K} |\lambda_k| (t_{k-1} - t_k). \tag{B.16}
\]

By Step 2 and (B.10), we can bound the first term in the right-hand side of (B.16) as

\[
\max_{2 \leq k \leq K} \frac{\hat{q}_k}{t_{k-1} - t_k} \leq \max_{2 \leq k \leq K} \frac{q_k}{t_{k-1} - t_k} + \max_{2 \leq k \leq K} \left| \frac{\hat{q}_k - q_k}{t_{k-1} - t_k} \right| \leq M + O_p(n^{-r_3}). \tag{B.17}
\]
To bound the second term, note from Steps 2 and 3 that
\[
\max_{2 \leq k \leq K} \left| \frac{\hat{y}_k - \hat{t}_k}{t_{k-1} - t_k} \right| \leq \max_{2 \leq k \leq K} \left| \frac{\hat{y}_k - y_k}{t_{k-1} - t_k} \right| + \max_{2 \leq k \leq K} \frac{y_k - \hat{t}_k}{(t_{k-1} - t_k)^2} (t_{k-1} - t_k)
\leq O_p(n^{-r_3}) + \frac{L}{6m} c_2 n^{-r_2}.
\]
Therefore it follows from Lemma 7 that
\[
\max_{2 \leq k \leq K} |\lambda_k| (t_{k-1} - t_k) \leq 2\phi^{-1} \left( \max_{2 \leq k \leq K} \frac{2 |\hat{y}_k - \hat{t}_k|}{t_{k-1} - t_k} \right)
\leq 2\phi^{-1} \left( \frac{L}{3m} c_2 n^{-r_2} + O_p(n^{-r_3}) \right).
\]
Combining (B.16), (B.17) and (B.18), we obtain
\[
\sup_{y \in C} A_1(y) \leq 2(M + O_p(n^{-r_3}))\phi^{-1} \left( \frac{L}{3m} c_2 n^{-r_2} + O_p(n^{-r_3}) \right).
\]
(B.19)
The uniform bound (3.9b) follows from (B.13), (B.14), (B.15), (B.19), and Lemma 6. Noting that the $O_p(n^{-r_3})$ term is also $o_{a.s.}(1)$ in this case, we also obtain the uniform strong consistency (3.9a).

**Proof of Corollary 4.** Since $Q_\tau \in \text{int } C$, there exists a $\delta_1$-ball $B_{\delta_1}(Q_\tau)$ around $Q_\tau$ such that $B_{\delta_1}(Q_\tau) \subset C$. Also, since $Q_\tau \in \text{int } C \subset D$, $f(Q_\tau) > 0$ by Assumption 1(ii). Furthermore, $f$ is continuous by Assumption 1(ii). Therefore, $f$ is bounded away from zero in a neighborhood of $Q_\tau$. That is, there exists a $\delta_2$-ball $B_{\delta_2}(Q_\tau)$ and $\bar{f} > 0$ such that $f(y) \geq \bar{f}(y)$ for all $y \in B_{\delta_2}(Q_\tau)$. Let $\delta = \frac{1}{2} (\delta_1 \wedge \delta_2)$.

We first show that
\[
F(Q_\tau - \delta) + \delta \cdot \bar{f} \leq \hat{F}(\hat{Q}_\tau) \leq F(Q_\tau + \delta) - \delta \bar{f}.
\]
(B.20)
By way of contradiction, suppose that $\hat{F}(\hat{Q}_\tau) > F(Q_\tau + \delta) - \delta \bar{f}$. Then, since $F$ and $\hat{F}$ are continuous by construction,
\[
F(Q_\tau) = \tau = \hat{F}(\hat{Q}_\tau) > F(Q_\tau + \delta) - \delta \bar{f}
\]
and hence $\delta \bar{f} \geq F(Q_\tau + \delta) - F(Q_\tau) > \delta \bar{f}$, which is a contradiction. Therefore, we must have $\hat{F}(\hat{Q}_\tau) \leq F(Q_\tau + \delta) - \delta \bar{f}$. Similarly, $F(Q_\tau - \delta) + \delta \bar{f} \leq \hat{F}(\hat{Q}_\tau)$ must hold too. Therefore, (B.20) holds.

Since $F$ is non-decreasing by construction, and it is also strictly increasing on
B_δ(Q_τ) \subset B_{δ_2}(Q_τ), (B.20) implies

\hat{Q}_τ \in B_δ(Q_τ) \subset C \quad \text{provided that} \quad \sup_{y \in C} \left| \hat{F}(y) - F(y) \right| < \delta f. \quad (B.21)

Also, noting that \( \hat{F}(\hat{Q}_τ) = \tau = F(Q_τ) \), we obtain

\hat{F}(\hat{Q}_τ) - F(\hat{Q}_τ) + F(\hat{Q}_τ) - F(Q_τ) = 0. \quad (B.22)

With all these pieces put together, we get

\begin{align*}
P \left( \limsup_{n \to \infty} \left\{ \omega \in \Omega : \left| \hat{Q}_τ - Q_τ \right| = 0 \right\} \right) \\
\overset{(i)}{=} P \left( \limsup_{n \to \infty} \left\{ \omega \in \Omega : \int_{Q_τ} f(y) \, dy = 0 \right\} \right) \\
= P \left( \limsup_{n \to \infty} \left\{ \omega \in \Omega : \left| F(\hat{Q}_τ) - F(Q_τ) \right| = 0 \right\} \right) \\
\overset{(ii)}{=} P \left( \limsup_{n \to \infty} \left\{ \omega \in \Omega : \left| \hat{F}(\hat{Q}_τ) - F(\hat{Q}_τ) \right| = 0 \right\} \right) \\
\geq P \left( \limsup_{n \to \infty} \left\{ \omega \in \Omega : \sup_{y \in C} \left| \hat{F}(y) - F(y) \right| = 0 \right\} \right) \\
\overset{(iii)}{\geq} P \left( \limsup_{n \to \infty} \left\{ \omega \in \Omega : \sup_{y \in C} \left| \hat{F}(y) - F(y) \right| = 0 \right\} \right) \\
\overset{(iv)}{=} 1.
\end{align*}

Here equality (i) follows from \( f(y) > f_\bar{y} \) for all \( y \in B_δ(Q_τ) \), equality (ii) follows from (B.22), inequality (iii) follows from (B.21), and equality (iv) follows from Corollary 3. This completes a proof of the corollary.

**Proof of Theorem 5.** Let \( \text{cl} T = [y_2, y_1] \times [y_3, y_2] \times \cdots \times [y_K, y_{K-1}] \) be the closure of \( T \) in (A.1), which is a nonempty compact interval. Let \( \partial T := \text{cl} T \setminus T \) be the boundary of \( T \). Extend \( J^* \) to \( \text{cl} T \) by defining \( J^* = \infty \) on \( \partial T \).

We divide our proof into three steps.

**Step 1.** \( J^* : \text{cl} T \to \mathbb{R} \cup \{\infty\} \) is lower semicontinuous.
Motivated by (3.4) and (3.7), define $J : \mathbb{R}^K \times \text{cl } T \to (-\infty, \infty]$ by

$$J(\lambda; t) = \begin{cases} \sum_{k=1}^{K} q_k \left( y_k \lambda_k - \log \left( \frac{e^{\lambda_k t_k - t_k}}{\lambda_k} \right) + \log q_k \right), & (t \in T) \\ \infty, & (t \in \partial T) \end{cases}$$

where the case $\lambda_k = 0$ needs to be separately defined in the obvious way using (3.4). Then we can easily verify that $J(\lambda; \cdot) : \text{cl } T \to (-\infty, \infty]$ is continuous. Therefore $J^*(t) = \sup_{\lambda} J(\lambda; t)$ is lower semicontinuous.

**Step 2.** $J^* : T \to \mathbb{R}$ is strictly convex.

Since $J^*(t) = \sup_{\lambda} J(\lambda; t)$ and the supremum of convex functions is convex, it suffices to show that $J(\lambda; t)$ is strictly convex in $t \in T$. To this end it suffices to show that its Hessian $\nabla^2 J$ is positive definite almost everywhere on $T$. A tedious but straightforward calculation yields

$$\frac{\partial J}{\partial t} = \frac{q_k \lambda_k}{e^{\lambda_k t_k} - 1} - \frac{q_{k+1} \lambda_{k+1}}{1 - e^{\lambda_{k+1} t_{k+1} - t_k}},$$

$$\frac{\partial^2 J}{\partial t^2} = \frac{q_k \lambda_k^2 e^{\lambda_k t_k} (t_k - t_{k+1})}{(e^{\lambda_k t_k} - 1)^2} + \frac{q_{k+1} \lambda_{k+1}^2 e^{\lambda_{k+1} t_{k+1} - t_k}}{(1 - e^{\lambda_{k+1} t_{k+1} - t_k})^2},$$

$$\frac{\partial^2 J}{\partial t \partial t_{k+1}} = -\frac{q_{k+1} \lambda_{k+1}^2 e^{\lambda_{k+1} t_{k+1} - t_k}}{(1 - e^{\lambda_{k+1} t_{k+1} - t_k})^2},$$

$$\frac{\partial^2 J}{\partial t \partial t_{k+\ell}} = 0 \text{ for } \ell > 1,$$

which are valid for $k = 1$ by setting $e^{\lambda_1 t_0} = 0$. Collecting the partial derivatives (B.23) into a matrix, we obtain the Hessian

$$\nabla^2 J = H_{K-1} := \begin{bmatrix} c_1 + c_2 & -c_2 & 0 & \cdots & 0 \\ -c_2 & c_2 + c_3 & -c_3 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -c_{K-2} & c_{K-2} + c_{K-1} & -c_{K-1} \\ 0 & \cdots & 0 & -c_{K-1} & c_{K-1} + c_K \end{bmatrix},$$

(B.24)

where

$$c_k := \frac{q_k \lambda_k^2 e^{\lambda_k t_k - t_k}}{(e^{\lambda_k t_k} - 1)^2} \geq 0,$$

with strict inequality if $\lambda_k \neq 0$. Since by (3.5) we have $\lambda_k = 0$ if and only if $y_k = \frac{t_k + t_{k+1}}{2}$, we have $c_k > 0$ for all $k$ almost everywhere on $T$. Below, consider such $t$. 

34
As is well known, a real symmetric matrix is positive definite if and only if all leading principal minors are positive. In general, for a tridiagonal matrix

\[
A_n := \begin{bmatrix}
    d_1 & a_1 & 0 & \cdots & 0 \\
    b_2 & d_2 & a_2 & \ddots & \vdots \\
    0 & \ddots & \ddots & \ddots & 0 \\
    \vdots & \ddots & b_{n-1} & d_{n-1} & a_{n-1} \\
    0 & \cdots & 0 & b_n & d_n
\end{bmatrix},
\]

its determinant \(D_n := \det A_n\) satisfies the three-term recurrence relation

\[
D_n = d_n D_{n-1} - b_n a_{n-1} D_{n-2},
\]

(B.25)

where \(D_0 = 1\) and \(D_1 = d_1\) (El-Mikkawy, 2004). Applying (B.25) to \(H_{K-1}\) in (B.24) with \(n = K - 1\), since \(d_n = c_n + c_{n+1}\) and \(b_n = a_{n-1} = -c_n\), we obtain

\[
D_n = (c_n + c_{n+1})D_{n-1} - c_n^2 D_{n-2}.
\]

(B.26)

Clearly \(D_1 = d_1 = c_1 + c_2 > 0\). Let us show by induction that

\[
D_n = c_{n+1} D_{n-1} + \prod_{k=1}^{n} c_k,
\]

(B.27)

which implies \(D_n > 0\) for all \(n\). If \(n = 1\), then \(D_1 = c_1 + c_2 = c_2 D_0 + c_1\) because \(D_0 = 1\), so (B.27) holds. Suppose (B.27) holds for some \(n\). Then for \(n + 1\), using (B.26) and the induction hypothesis, we obtain

\[
D_{n+1} = (c_{n+1} + c_{n+2}) D_n - c_{n+1}^2 D_{n-1}
\]

\[
= c_{n+2} D_n + c_{n+1} \left( c_{n+1} D_{n-1} + \prod_{k=1}^{n} c_k \right) - c_{n+1}^2 D_{n-1}
\]

\[
= c_{n+2} D_n + \prod_{k=1}^{n+1} c_k,
\]

so (B.27) holds for \(n + 1\). This completes the proof of the strict convexity of \(J(\lambda; t)\) in \(t\), and hence \(J^*(t) = \max_\lambda J(\lambda; t)\) is strictly convex in \(t \in T\).

**Step 3.** There exists a unique \(t^* \in T\) that minimizes \(J^*\). For this \(t^*\), the maximum entropy density \(f^*\) in (3.6) is continuous on \([t_K, \infty)\).

Since \(\text{cl} T\) is nonempty compact, \(J^* : \text{cl} T \to (-\infty, \infty]\) is lower semicontinuous,
and $J^*$ is finite-valued on $T$, it achieves a minimum at some $t^* \in \text{cl} T$. Since $J^* = \infty$ on $\partial T$, it must be $t^* \in T$. Since $J^*$ is strictly convex on $T$, the minimizer $t^*$ is unique.

Let us next show the continuity of $f^*$. Since $f^*$ is piecewise exponential (hence continuous), it suffices to show the continuity at the thresholds $t^*_k$. Since $J^*(\lambda; \cdot) : T \to \mathbb{R}$ is differentiable, it follows from the first-order condition, the envelope theorem, and (B.23a) that

$$0 = \frac{\partial J^*}{\partial t_k}(t^*) = \frac{q_k \lambda_k}{e^{\lambda_k t_k} - e^{\lambda_k t_{k-1}}} = \frac{q_k \lambda_k}{e^{\lambda_k t_k} - e^{\lambda_k t_{k-1}}} - \frac{q_{k+1} \lambda_{k+1}}{1 - e^{\lambda_{k+1} (t_{k+1} - t_k)}}.$$

(B.28)

Comparing (3.6) and (B.28), we obtain $f^*(t^*_k -) = f^*(t^*_k +)$, so $f^*$ is continuous.

C Numerical issues

Although conceptually straightforward, numerically maximizing $J_k$ in (3.4) over $\lambda$ can be unstable because the values of $\{(t_k, y_k, t_{k-1})\}_{k=1}^K$ change by many orders of magnitude in typical data, for instance $t_K = 1$ and $t_1 = 10^7$ in Table 1. For this reason it is useful to choose some scaling factor $s > 0$ and consider $(st_k, sy_k, st_{k-1})$.

Define

$$J_k(\lambda; s) := sy_k \lambda - \log \left( \frac{e^{\lambda t_{k-1}} - e^{\lambda t_k}}{\lambda} \right) = J_k(\lambda s) - \log s.$$

Maximizing both sides over $\lambda \in \mathbb{R}$ and using the definition of $\lambda_k$ in Proposition 1, it follows that

$$\lambda_k = s \arg \max_{\lambda \in \mathbb{R}} J_k(\lambda; s) \quad \text{and} \quad J_k(\lambda_k) = \max_{\lambda \in \mathbb{R}} J_k(\lambda; s) + \log s.$$

Therefore to numerically compute $\lambda_k$ and $J_k(\lambda_k)$, we may maximize $J_k(\lambda, s)$ for some scaling factor $s > 0$ (say $s = 1/t_k-1$ so that $st_k < sy_k < st_{k-1} = 1$), multiply its maximizer by $s$, and add $\log s$ to its maximum value.
D Lorenz curve for double Pareto distribution

Integrating the double Pareto density (with \( M = 1 \)), it is easy to show that the CDF of double Pareto is

\[
F(y) = \begin{cases} 
\frac{\alpha}{\alpha + \beta} y^\beta, & (0 \leq y \leq 1) \\
1 - \frac{\alpha}{\alpha + \beta} y^{-\alpha}. & (y \geq 1)
\end{cases}
\]

Let us compute the Lorenz curve, which is implicitly defined by

\[
L(F(y)) = \frac{\int_0^y tf(t) \, dt}{\mu},
\]

where \( \mu \) is the mean. If \( y \leq 1 \), then

\[
\int_0^y tf(t) \, dt = \int_0^y \frac{\alpha \beta}{\alpha + \beta} t^\beta \, dt = \frac{\alpha \beta}{(\alpha + \beta)(\beta + 1)} y^{\beta + 1}.
\]

If \( y \geq 1 \), then

\[
\int_0^y t f(t) \, dt = \int_0^1 t f(t) \, dt + \int_1^y t f(t) \, dt
\]

\[
= \frac{\alpha \beta}{(\alpha + \beta)(\beta + 1)} + \int_1^y \frac{\alpha \beta}{\alpha + \beta} t^{-\alpha} \, dt
\]

\[
= \frac{\alpha \beta}{(\alpha + \beta)(\alpha - 1)} \left( \frac{\alpha + \beta}{\beta + 1} - y^{-\alpha + 1} \right).
\]

Letting \( y \to \infty \), the mean of double Pareto (assuming \( \alpha > 1 \)) is

\[
\mu = \frac{\alpha \beta}{(\alpha - 1)(\beta + 1)}.
\]

Thus if \( y \leq 1 \) (or equivalently \( F = F(y) \leq \frac{\alpha}{\alpha + \beta} \)), then

\[
L(F) = \frac{\alpha - 1}{\alpha + \beta} y^{\beta + 1} = \frac{\alpha - 1}{\alpha + \beta} \left( \frac{\alpha + \beta}{\alpha} F \right)^{1+1/\beta}
\]

\[
= (\alpha - 1)(\alpha + \beta)^{1/\beta} \alpha^{-1-1/\beta} F^{1+1/\beta}.
\]

If \( y \geq 1 \) (or equivalently \( F = F(y) \geq \frac{\alpha}{\alpha + \beta} \)), then

\[
L(F) = 1 - \frac{\beta + 1}{\alpha + \beta} y^{-\alpha + 1} = 1 - \frac{\beta + 1}{\alpha + \beta} \left( \frac{\alpha + \beta}{\beta} (1 - F) \right)^{1-1/\alpha}
\]
Putting all the pieces together, the Lorenz curve for a double Pareto distribution with $\alpha > 1$ and $\beta > 0$ is

\[
L(x) = \begin{cases} 
(\alpha - 1)(\alpha + \beta)^{1/\beta} x^{1+1/\beta}, & (0 \leq x \leq \frac{\alpha}{\alpha + \beta}) \\
1 - (\beta + 1)(\alpha + \beta)^{-1/\alpha} (1 - x)^{1-1/\alpha}. & (\frac{\alpha}{\alpha + \beta} \leq x \leq 1)
\end{cases}
\]

### E Piketty (2003)’s Pareto interpolation method

Piketty (2003)’s Pareto interpolation method for constructing top income shares is widely used, for example in Piketty and Saez (2003) and the World Inequality Database.\(^8\) Because the detailed description of the method is relegated to Piketty (2001, Appendix B, Section 1.1, pp. 592–599), which is written in French, for completeness we describe the method here.

We use the same notation as in Section 3.1. For simplicity assume that we are interested in the income distribution and the top income shares of the taxpayers, so the sample size is $n = n_K$. Suppose we observe the lower income threshold for income group $k$, which we denote by $t_k := Y_{(n_k)}$. Let $p_k := n_k/n$ be the top fractile corresponding to the $k$-th income threshold $t_k$ and

\[
s_k := \frac{1}{n_k} \sum_{i=1}^{n_k} Y_i = \frac{S_{n_k}}{n_k}
\]

be the average income of taxpayers with income above $t_k$.

Let $b_k = s_k/t_k$ be the ratio between the average income of individuals exceeding $t_k$ and the income threshold $t_k$. If income $Y$ is Pareto distributed (in the upper tail), then for large enough income level $y$, the CDF takes the form $F(y) = 1 - Ay^{-\alpha}$ for some $A > 0$ and Pareto exponent $\alpha > 1$. Therefore for any large enough income threshold $t$, we have

\[
b(t) := \frac{E[Y \mid Y \geq t]}{t} = \frac{\int_{t}^{\infty} yF'(y) \, dy}{t(1 - F(t))} = \frac{\alpha}{\alpha - 1}.
\]

(E.1)

Piketty (2001) refers to $b_k = s_k/t_k$ as the (local) Pareto coefficient. When the income distribution has a Pareto upper tail, the (local) Pareto exponent can be

\(^8\)https://wid.world/
recovered from (E.1) as
\[ \alpha_k = \frac{b_k}{b_k - 1} = \frac{1}{1 - t_k/s_k}. \] (E.2)

Let \( p \in (0, 1] \) and suppose we would like to construct the top \( p \) fractile income share. For example, \( p = 0.01 \) corresponds to the top 1% income share. Piketty (2001) proceeds as follows to construct the top \( p \) fractile income share. First, let \( p_k \) be the closest proportion to \( p \) directly observed in the data, and \( t_k \) and \( \alpha_k \) be the corresponding income threshold and local Pareto exponent. Then one supposes that the income distribution is locally exactly Pareto, and therefore the CDF is
\[ F(y) = 1 - p_k(y/t_k)^{-\alpha_k}. \]

The income threshold corresponding to the top \( p \) fractile can be computed as
\[ 1 - p = 1 - p_k(y/t_k)^{-\alpha_k} \iff y = t(p) := t_k(p_k/p)^{1/\alpha_k}. \]

Noting that the sample size is \( n \), the total income of taxpayers in the top \( p \) fractile can be computed as
\[
S(p) := n \int_{t(p)}^{\infty} y F'(y) \, dy = n \int_{t(p)}^{\infty} \alpha_k p_k (y/t_k)^{-\alpha_k} \, dy \\
= n \frac{\alpha_k}{\alpha_k - 1} p_k t_k^{\alpha_k} t(p)^{1-\alpha_k} = n \frac{\alpha_k}{\alpha_k - 1} p t_k^{1/\alpha_k} p^{1-1/\alpha_k} \\
= n s_k p_k^{1/\alpha_k} p^{1-1/\alpha_k}. \]
(E.3)

The top \( p \) fractile income share can then be computed as \( S(p)/S(1) \). If the researcher is interested in the top income share of all income earners (including those who do not file for taxes), one would adjust the sample size \( n \) and the total income \( S(1) \) from other sources (e.g., population statistics and national accounts); see for example Piketty and Saez (2001, Appendix A).

### F Additional simulations

In the main text, we used the double Pareto distribution to evaluate the bias and RMSE in the top income shares. Here, we repeat the analysis in Table 3 with data generated from lognormal, gamma, and Weibull distributions, which are described in Table 2. For the lognormal distribution, we set \( \sigma = 1.5 \). For the gamma and the
Weibull distributions, we set \( a = 1 \) and \( k = 1 \), respectively, so that they are both identical to the exponential distribution with unit mean (for which it is possible to compute top income shares in closed-form). Other setups and the four estimation methods are the same as those described in Section 4. Tables 4 and 5 present the results.

Table 4: Simulation results of four methods for top shares with lognormal data.

| \( \eta \) | Relative Bias | \( n \) | ME | \( \eta \) | KP | \( \eta \) | VA | \( \eta \) | HGBRC |
|---|---|---|---|---|---|---|---|---|---|
| \( 10^4 \) | 0.001 | 0.016 | -0.003 | -0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.01 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.05 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.2 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.3 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.4 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.5 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.6 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.7 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.8 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.9 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

| \( \eta \) | Relative RMSE | \( n \) | ME | \( \eta \) | KP | \( \eta \) | VA | \( \eta \) | HGBRC |
|---|---|---|---|---|---|---|---|---|---|
| \( 10^4 \) | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.01 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.05 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.2 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.3 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.4 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.5 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.6 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.7 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.8 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 0.9 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

Note: See Table 3 for details.

Our findings here are similar to those presented in Section 4. First, our proposed ME method performs very well in terms of both the bias and RMSE. They
Table 5: Simulation results of four methods for top shares with gamma data.

| Relative Bias | \( p_0 \) | 0.001 | 0.01 | 0.05 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---------------|----------|-------|------|------|----|----|----|----|----|----|----|----|----|
| \( n \)       |          |       |      |      |    |    |    |    |    |    |    |    |    |
| ME            | \( 10^4 \) | -0.020 | 0.001 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
|               | \( 10^5 \) | 0.000  | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
|               | \( 10^6 \) | -0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| KP            | \( 10^4 \) | -0.886 | 0.313 | 0.206 | 0.085 | -0.003 | -0.027 | -0.029 | -0.021 | -0.011 | -0.001 | 0.004 | 0.004 | 0.004 |
|               | \( 10^5 \) | -0.905 | 0.313 | 0.205 | 0.084 | -0.003 | -0.027 | -0.029 | -0.022 | -0.011 | -0.001 | 0.004 | 0.004 | 0.004 |
|               | \( 10^6 \) | -0.906 | 0.313 | 0.205 | 0.084 | -0.003 | -0.027 | -0.029 | -0.022 | -0.011 | -0.001 | 0.004 | 0.004 | 0.004 |
| VA            | \( 10^4 \) | 11.773 | 1.777 | 0.236 | -0.012 | -0.107 | -0.096 | -0.062 | -0.029 | -0.008 | 0.002 | 0.003 | 0.001 | 0.001 |
|               | \( 10^5 \) | 11.763 | 1.775 | 0.235 | -0.013 | -0.108 | -0.096 | -0.062 | -0.029 | -0.008 | 0.002 | 0.003 | 0.001 | 0.001 |
|               | \( 10^6 \) | 11.761 | 1.774 | 0.235 | -0.013 | -0.108 | -0.096 | -0.062 | -0.029 | -0.008 | 0.002 | 0.003 | 0.001 | 0.001 |
| HGBRC         | \( 10^4 \) | 0.285  | 0.113 | 0.046 | 0.027 | 0.014 | 0.008 | 0.006 | 0.004 | 0.002 | 0.001 | 0.001 | 0.000 | 0.000 |
|               | \( 10^5 \) | 0.090  | 0.035 | 0.015 | 0.009 | 0.004 | 0.003 | 0.002 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 |
|               | \( 10^6 \) | 0.027  | 0.013 | 0.007 | 0.004 | 0.002 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

See Table 3 for details.

are smaller than those of the other methods for most of the \((n, p_0)\) combinations, especially the bias. Second, the KP and the VA methods both impose some parametric assumptions on the Lorenz curve and hence implicitly on the underlying density. The misspecification biases could be quite large depending on the true data generating process, especially for the VA method.
References

Thomas Blanchet, Juliette Fournier, and Thomas Piketty. Generalized Pareto curves: Theory and applications. Review of Income and Wealth, 68(1):263–288, March 2022. doi:10.1111/roiw.12510.

Gordon Blower and Julia E. Kelsall. Nonlinear kernel density estimation for binned data: Convergence in entropy. Bernoulli, 8(4):423–449, August 2002.

Jonathan M. Borwein and Adrian S. Lewis. Duality relationships for entropy-like minimization problems. SIAM Journal on Control and Optimization, 29(2):325–338, March 1991. doi:10.1137/0329017.

Yi-Ting Chen. A unified approach to estimating and testing income distributions with grouped data. Journal of Business & Economic Statistics, 36(3):438–455, 2018. doi:10.1080/07350015.2016.1194762.

Duangkamon Chotikapanich, William E. Griffiths, and D. S. Prasada Rao. Estimating and combining national income distributions using limited data. Journal of Business & Economic Statistics, 25(1):97–109, January 2007. doi:10.1198/073500106000000224.

Frank A. Cowell and Fatemeh Mehta. The estimation and interpolation of inequality measures. Review of Economic Studies, 49(2):273–290, April 1982. doi:10.2307/2297275.

Tjeerd de Vries and Alexis Akira Toda. Capital and labor income Pareto exponents across time and space. Review of Income and Wealth, 68(4):1058–1078, December 2022. doi:10.1111/roiw.12556.

Moawwad E. A. El-Mikkawy. On the inverse of a general tridiagonal matrix. Applied Mathematics and Computation, 150(3):669–679, March 2004. doi:10.1016/S0096-3003(03)00298-4.

Leland E. Farmer and Alexis Akira Toda. Discretizing nonlinear, non-Gaussian Markov processes with exact conditional moments. Quantitative Economics, 8(2):651–683, July 2017. doi:10.3982/QE737.

Daniel R. Feenberg and James M. Poterba. Income inequality and the incomes of very high-income taxpayers: Evidence from tax returns. Tax Policy and the Economy, 7:145–177, 1993. doi:10.1086/tpe.7.20060632.
Duncan K. Foley. A statistical equilibrium theory of markets. *Journal of Economic Theory*, 62(2):321–345, April 1994. doi:10.1006/jeth.1994.1018.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, Hoboken, NJ, 2 edition, 1999.

Gholamreza Hajargasht and William E. Griffiths. Minimum distance estimation of parametric Lorenz curves based on grouped data. *Econometric Reviews*, 39(4):344–361, April 2020. doi:10.1080/07474938.2019.1630077.

Gholamreza Hajargasht, William E. Griffiths, Joseph Brice, D. S. Prasada Rao, and Duangkamon Chotikapanich. Inference for income distributions using grouped data. *Journal of Business & Economic Statistics*, 30(4):563–575, October 2012. doi:10.1080/07350015.2012.707590.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963. doi:10.1080/01621459.1963.10500830.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, May 1957. doi:10.1103/PhysRev.106.620.

Vanesa Jorda, José María Sarabia, and Markus Jäntti. Inequality measurement with grouped data: Parametric and non-parametric methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(3):964–984, July 2021. doi:10.1111/rssa.12702.

Nanak C. Kakwani and Nripesh Podder. Efficient estimation of the Lorenz curve and associated inequality measures from grouped observations. *Econometrica*, 44(1):137–148, January 1976. doi:10.2307/1911387.

Yuichi Kitamura and Michael Stutzer. An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(4):861–874, July 1997. doi:10.2307/2171942.

Ji Hyung Lee, Yuya Sasaki, Alexis Akira Toda, and Yulong Wang. Capital and labor income Pareto exponents in the United States, 1916–2019. June 2022. URL https://arxiv.org/abs/2206.04257.

James B. McDonald. Some generalized functions for the size distribution of income. *Econometrica*, 52(3):647–663, May 1984. doi:10.2307/1913469.
Thomas Piketty. *Les Hauts Revenus en France au XXe Siècle: Inégalités et Redistributions, 1901–1998*. Bernard Grasset, Paris, 2001.

Thomas Piketty. Income inequality in France, 1901–1998. *Journal of Political Economy*, 111(5):1004–1042, October 2003. doi:10.1086/376955.

Thomas Piketty and Emmanuel Saez. Income inequality in the United States, 1913–1998. NBER Working Paper 8467, 2001. URL https://www.nber.org/papers/w8467.

Thomas Piketty and Emmanuel Saez. Income inequality in the United States, 1913–1998. *Quarterly Journal of Economics*, 118(1):1–41, February 2003. doi:10.1162/00335530360535135.

Miguel Reyes, Mario Francisco-Fernández, and Ricardo Cao. Nonparametric kernel density estimation for general grouped data. *Journal of Nonparametric Statistics*, 28(2):235–249, March 2016. doi:10.1080/10485252.2016.1163348.

David W. Scott and Simon J. Sheather. Kernel density estimation with binned data. *Communication in Statistics - Theory and Methods*, 14(6):1353–1359, January 1985. doi:10.1080/03610928508828980.

Michael Stutzer. A Bayesian approach to diagnosis of asset pricing models. *Journal of Econometrics*, 68(2):367–397, August 1995. doi:10.1016/0304-4076(94)01656-K.

Michael Stutzer. A simple nonparametric approach to derivative security valuation. *Journal of Finance*, 51(5):1633–1652, December 1996. doi:10.1111/j.1540-6261.1996.tb05220.x.

Xu Sun. Asymmetric kernel density estimation based on grouped data with applications to loss model. *Communications in Statistics - Simulation and Computation*, 43(3):657–672, March 2014. doi:10.1080/03610918.2012.712184.

Ken’ichiro Tanaka and Alexis Akira Toda. Discrete approximations of continuous distributions by maximum entropy. *Economics Letters*, 118(3):445–450, March 2013. doi:10.1016/j.econlet.2012.12.020.

Ken’ichiro Tanaka and Alexis Akira Toda. Discretizing distributions with exact moments: Error estimate and convergence analysis. *SIAM Journal on Numerical Analysis*, 53(5):2158–2177, 2015. doi:10.1137/140971269.
Alexis Akira Toda. Existence of a statistical equilibrium for an economy with endogenous offer sets. *Economic Theory*, 45(3):379–415, December 2010. doi:10.1007/s00199-009-0493-6.

Alexis Akira Toda. The double power law in income distribution: Explanations and evidence. *Journal of Economic Behavior and Organization*, 84(1):364–381, September 2012. doi:10.1016/j.jebo.2012.04.012.

Alexis Akira Toda. Bayesian general equilibrium. *Economic Theory*, 58(2):375–411, February 2015. doi:10.1007/s00199-014-0849-4.

José A. Villaseñor and Barry C. Arnold. Elliptical Lorenz curves. *Journal of Econometrics*, 40(2):327–338, February 1989. doi:10.1016/0304-4076(89)90089-4.

Ximing Wu. Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics*, 115(2):347–354, August 2003. doi:10.1016/S0304-4076(03)00114-3.