MINIMUM DISTANCE TESTING AND TOP INCOME SHARES IN KOREA

By

Jin Seo Cho, Myung-Ho Park, and Peter C. B. Phillips

June 2015

COWLES FOUNDATION DISCUSSION PAPER NO. 2007

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Minimum Distance Testing and Top Income Shares in Korea[*]

JIN SEO CHO

School of Economics

Yonsei University

50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Korea

MYUNG-HO PARK

Center for Long-Term Fiscal Projections

Korea Institute of Public Finance

1924 Hannuri-daero, Sejong 339-007, Korea

PETER C.B. PHILLIPS

Yale University, University of Auckland

Singapore Management University &

University of Southampton

First version: October 2014 This version: April 2015

Abstract

We study Kolmogorov-Smirnov goodness of fit tests for evaluating distributional hypotheses where unknown parameters need to be fitted. Following work of Pollard (1979), our approach uses a Cramér-von Mises minimum distance estimator for parameter estimation. The asymptotic null distribution of the resulting test statistic is represented by invariance principle arguments as a functional of a Brownian bridge in a simple regression format for which asymptotic critical values are readily delivered by simulations. Asymptotic power is examined under fixed and local alternatives and finite sample performance of the test is evaluated in simulations. The test is applied to measure top income shares using Korean income tax return data over 2007 to 2012. When the data relate to the upper 0.1% or higher tail of the income distribution, the conventional assumption of a Pareto tail distribution cannot be rejected. But the Pareto tail hypothesis is rejected for the top 1.0% or 0.5% incomes at the 5% significance level.

Key Words: Brownian bridge, Cramér-von Mises statistic, Distribution-free asymptotics, null distribution; minimum distance estimator; empirical distribution; goodness-of-fit test; Crámer-von Mises distance; top income shares; Pareto interpolation.

JEL Subject Classifications: C12, C13, D31, E01, O15.

[*] The authors acknowledge helpful discussions with Hoon Hong, Jinook Jeong, Sun-Bin Kim, Tae-Hwan Kim, Rami Tabri, Byungsam Yoo, and participants of NZESG in Brisbane, February, 2015. Pyung Gang Kim provided excellent research assistance. The National Tax Service of Korea provided the data. Phillips thanks the NSF for research support under grant No. SES 12-58258.

1 Introduction

Distributional hypotheses can play a significant role in the nature and quality of inference in many different areas of econometric work. In the quantification of inequality, for instance, interpolation methods based on the Pareto distribution are widely used for measuring top income shares. In recent influential work involving such exercises, Piketty and Saez (2003) follow an approach that is now quite typical by assuming that top incomes are well modeled by a Pareto distribution, which is then used to estimate the top income shares. As another example, the probability integral transform (PIT) is frequently used in density forecasting exercises (see Diebold, Gunther and Tay, 1998) so that PIT transformed quantities follow a standard uniform distribution, which is extremely useful in statistical testing and forecast evaluation, if the assumed distribution is correct. In such applications, it is of considerable interest to assess whether the distributional hypotheses are supported by the data.

Many test procedures are available to make such assessments. The most commonly used methods for testing distributional hypotheses in practical work involve goodness-of-fit (GOF) test statistics based on the Kolmogorov and Smirnov (KS) and Crámer-von Mises (CM) test statistics. When a particular distribution is hypothesized, these GOF test statistics are known to converge weakly to certain functionals of a Brownian bridge process under the null. Early fundamental work on the use of weak convergence methods for the development of such limit theory was done by Durbin (1973).

The practical efficacy of GOF test statistics is often limited by the presence of unknown parameters in the parent distribution. As Durbin (1973) and Henze (1996) pointed out, the KS test statistic is not distribution free. In consequence, when unknown parameters of the hypothesized distribution are estimated, the estimation error typically affects the asymptotic null distribution, so that different models will generate different asymptotic critical values for the test. This limitation applies also to other GOF test statistics.

The main goal of the current paper is to introduce a methodology for improving the practical efficacy of GOF testing. We concentrate our attention on the KS test in view of its popularity in applied work and, for this statistic, we follow the work of Pollard (1980) and examine the use of the minimum Crámer-von Mises distance (MCMD) estimator in dealing with parameter estimation. Bolthausen (1977) and Pollard (1980), among others, studied the asymptotic behavior of minimum distance (MD) estimators and provided asymptotic results for generalized GOF tests. For the purposes of the current study, we exploit the fact that the MD estimator can be analyzed in the context of regression when the CM distance is used for MD estimation. The MCMD estimator in turn simplifies the asymptotic null distribution of the KS test statistic: just as for the KS test statistic when there are no unknown parameters, the limit theory is again a functional of a Brownian bridge process, although in the case where parameter estimation is employed, the limit theory is not given by the same functional. The important practical implication is that asymptotic critical values can be obtained by applying invariance principle arguments in the same way as for the original KS test statistic. The current paper provides the form of the functional and demonstrates its implementation in a practical application.

In prior work on this topic, the practical inefficacy of the KS test statistic has been tackled by methodologies that are numerically intensive. For example, Henze (1996) and Klar (1999) recommend applying the parametric bootstrap to GOF tests, and Khmaladze (1981, 1993, 2013) modified the GOF tests by a transformation so that the asymptotic null distribution is unaffected by parameter estimation errors under the null. The procedure given here to obtain asymptotic critical values is not as numerically intensive as the parametric bootstrap or the martingale transformation. Simulations show that the new methodology has performance characteristics similar to those of the parametric bootstrap.

The KS test statistic has null and local alternative asymptotic distributions that depend upon data types. The practical import of this feature of the test is that grouped (or discretely distributed) data and continuously distributed data have different asymptotic distributions. We first examine grouped data and consider the implications for the KS test of using MCMD parameter estimation in the construction of the test By focusing on grouped data, the regression nature of the MCMD estimator is manifest and it becomes clear how to simulate to obtain asymptotic critical values of the test. We next extend the discussion to include continuously distributed data. A large group size limit distribution of the KS test statistic is derived from the large sample size limit distribution by increasing the group size and keeping the the data range fixed. By this process, the asymptotic null and local alternative distributions of the KS test of Bolthausen (1977) and Pollard (1980) can be obtained in a different way. This process also enables us to identify the Gaussian process associated with the asymptotic null distribution as another functional of the Brownian bridge, so that the associated Gaussian process can be easily simulated.

As an empirical application of the KS test statistic presented here, we revisit the problem of estimating top income shares of the income distribution by means of Pareto interpolation. Since Kuznets (1953, 1955) first examined the top income shares in US income data, these quantities have been commonly used in empirical work to assist in addressing drawbacks in Gini coefficient measures that focus more on the central tendencies of income data. Piketty and Saez (2003), Piketty (2003), Atkinson and Leigh (2007, 2008), Moriguchi and Saez (2010), and Kim and Kim (2014), among many others, assume a Pareto distribution for grouped income data in the US, France, Australia, New Zealand, Japan, and Korea (respectively) and estimate top income shares by Pareto interpolation.[1] Using our methodology and Korean income tax return data from 2007 to 2012, we test the underlying hypothesis of a Pareto distribution and conclude that the Pareto distributional hypothesis does not hold for top 1.0% and higher incomes, although the hypothesis is not rejected further in the tail of the distribution for the top 0.10% and higher incomes. The income data we use here are of very high quality and were provided by the National Tax Service of Korea. Although they are grouped, the group intervals are narrow: out of the 6 years of data we used, the smallest and largest group sizes were 2,760 and 4,241, respectively. With this degree of detail, we may expect to be able to test the Pareto distributional hypothesis with some precision using the asymptotic theory.

The plan of this paper is as follows. Section 2 develops the limit theory for the MCMD estimator and associated KS test statistic for grouped data. The asymptotic null distribution, power, and local power of the test statistic are also derived. Section 3 examines the large sample size limit distribution for continuously distributed data. In Section 4, we conduct Monte Carlo experiments to evaluate the adequacy of the asymptotic theory. Section 5 applies the KS test statistic to the Korean income tax return data from 2007 to 2012. Conclusions are given in Section 6. Proofs, related technical material, and data explanations are provided in the Appendix.

A brief word on notation. A function mapping $f : \mathcal{X} \mapsto \mathcal{Y}$ is denoted by $f(\cdot)$, evaluated derivatives such as $f'(x)|_{x=x_*}$ are written simply as $f'(x_*)$, the vector derivative $\nabla_{\boldsymbol{\theta}} F(x, \boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\theta})F(x, \boldsymbol{\theta})$, and for $i = 0, 1$, $\partial_j^i F(x, \boldsymbol{\theta}) := (\partial^i/\partial\theta_j)F(x, \boldsymbol{\theta})$ so that, for $i = 0$, $\partial_j^i F(x, \boldsymbol{\theta}) \equiv F(x, \boldsymbol{\theta})$.

2 Testing Distributional Hypotheses for Grouped Data

Suppose that data is available in a group frequency format whereby some variable of interest, X_t, may be unobserved but the numbers of times X_t lies within certain specified groups are observed. Typical income data have this format. In such cases, if X_t is annual income earned by individual or household t, then income data is provided in the following pairs

$$\{(x_i, x_{i+1}], \#\{X_t \in (x_i, x_{i+1}]\} : i = 0, 1, \ldots, k; t = 1, 2, \ldots, n\} \tag{1}$$

along with total group incomes. According to this scheme, n is the sample size of the number of individuals or households, k is the number of income groups, x_i and x_{i+1} are the lower and upper bounds of each interval

[1] Measuring top income shares is now a popular research topic in applied economics. Atkinson, Piketty, and Saez (2011) provide a well-organized summary of results for many countries, including Argentina, Canada, China, Finland, German, India, Indonesia, Ireland, Italy, the Netherlands, Norway, Portugal, Singapore, Spain, Sweden, and Switzerland.

i, and $\#\{A\}$ is the frequency A is observed. The end points x_0 and x_k are fixed at b and u, respectively.

Using data that fit the above format, Piketty and Saez (2003), Piketty (2003), Atkinson (2005), Atkinson and Leigh (2007, 2008), Moriguchi and Saez (2010), and Kim and Kim (2014) among others have estimated the top income shares in the US, UK, France, Australia, New Zealand, Japan, and Korea. The data in these studies may all be understood as continuously distributed grouped data or as collections of discretely distributed observations.

Suppose that an applied investigator is interested in testing some distributional assumption regarding the generating mechanism, or probability measure $\mathbb{P}$ with cumulative distribution function (cdf) F, of the latent variable X_t underlying the observed grouped data. The following hypotheses are considered:

$$\mathcal{H}_0 : \text{for all } x_i, \text{ there is a parameter value } \boldsymbol{\theta}_* \text{ and } cdf\ F \text{ such that } \mathbb{P}(X_t \leq x_i \mid b \leq X_t \leq u) = F(x_i, \boldsymbol{\theta}_*);$$

$$\mathcal{H}_1 : \text{ there is no } \boldsymbol{\theta}_* \text{ or } cdf\ F \text{ such that for all } x_i, \mathbb{P}(X_t \leq x_i \mid b \leq X_t \leq u) = F(x_i, \boldsymbol{\theta}_*).$$

Under these hypotheses, the parameter value $\boldsymbol{\theta}_*$ is properly defined only under the null $\mathcal{H}_0$. For practical reasons in what follows we consider distributions bounded below and above by $b = x_0$ and $u = x_k$, respectively. Empirical income data generally do not follow a Pareto law for low or medium income levels, so it is natural to focus on income levels that are bounded below in investigating the suitability of a Pareto law. It is also convenient to bound income levels by some (possibly very large) upper bound, which avoids extreme observations adversely affecting inference.

The null hypothesis $\mathcal{H}_0$ is motivated by commonly used estimation procedures. For example, Piketty and Saez (2003), Piketty (2003), Atkinson and Leigh (2007, 2008), Moriguchi and Saez (2010), and Kim and Kim (2014) each assume an underlying Pareto law for income data and estimate top income levels by a Pareto interpolation method. The validity of this method relies on the validity of the Pareto law, and violations of the law lead to biased and inconsistent estimation, thereby motivating tests of the distributional hypothesis.

The literature provides a variety of distributional test methodologies. Primary among these are Goodness-of-fit (GOF) tests and of these the most popular in empirical work is the Kolmogorov and Smirnov (KS) statistic. The limit theory of the traditional KS test statistic is a simple functional of a Brownian bridge process under the null, and Smirnov (1948) provides critical values for the test when the available data are continuously distributed. For grouped or discretely distributed data, Wood and Altavela (1978), Pettitt and Stephens (1977) and Choulakian, Lockhart, and Stephens (1994), among others, give the asymptotic null distribution of the KS test statistic in terms of another functional of the same Brownian bridge limit process.

When parameters are estimated, the limit distributions are affected. In his original treatment, Durbin (1973) pointed out that the asymptotic null distribution of the KS test statistic is not invariant to parameter estimation, so the test is not distribution free. Henze (1996) observed the same property for discretely distributed data. This limitation affects practical implementation of the KS test and has led to various numerically intensive procedures. One such procedure is the parametric bootstrap, which provides effective size control of the KS test asymptotically. A second approach, due to Khmaladze (1981, 1993), modifies the test statistic by a martingale transformation to eliminate the effect of parameter estimation asymptotically for continuously distributed observations. Khmaladze (2013) and Lee (2014) have given alternate transformations that may be used for discretely distributed observations.

The approach pursued in the present work differs from the prior literature. Instead of leaving the parametric estimator undefined in the test statistic, we suggest a particular estimator that ensures the null limit distribution of the KS test is pivotal and readily implementable for practical work. The estimator that achieves this purpose is the minimum Crámer-von Mises distance (MCMD) estimator. As it turns out, the MCMD estimator can be analyzed in a simple regression context which enables us to represent the asymptotic null distribution of the resulting KS test as a new functional of the same Brownian bridge process that appears in the original KS limit theory where there are no unknown parameters. The regression characteristic of the estimator is more apparent in the context of grouped data. Importantly, while the modified KS test statistic has a limit functional form that differs from the KS test with no parameter estimation error, the statistic is still asymptotically pivotal and depends only on the same Brownian bridge process, which is easily simulated to obtain critical values for the test.

Before examining the MCMD estimator, we provide the following conditions to fix ideas.

Assumption A. *(i) $\{X_t \in \mathbb{R}\}$ is independently and identically distributed (IID) with a continuous distribution and cdf $F(x_i, \boldsymbol{\theta})$;*

(ii) For every $i = 1, 2, \ldots, k$, $F(x_i, \cdot) : \boldsymbol{\Theta} \mapsto [0, 1]$ is in $\mathcal{C}^{(1)}(\boldsymbol{\Theta})$ where $\boldsymbol{\Theta} \subset \mathbb{R}^d$ is a compact and convex set with $k > d$ and such that $-\infty < b = x_0 < x_1 < \ldots < x_k = u < \infty$;

(iii) $\boldsymbol{\theta}_o := \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q(\boldsymbol{\theta}) \in \mathrm{int}(\boldsymbol{\Theta})$ and is unique in $\boldsymbol{\Theta}$, where for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ $Q(\boldsymbol{\theta}) := \sum_{i=1}^{k}\{F(x_i, \boldsymbol{\theta}) - p(x_i)\}^2$ and for every $i = 1, 2, \ldots, k$, $p(x_i) := \mathbb{P}(X_t \leq x_i \mid b \leq X_t \leq u)$; and

(iv) $\mathbf{z}'\mathbf{z}$ is positive definite, where $\mathbf{z} := [\nabla_{\boldsymbol{\theta}} F(x_1, \boldsymbol{\theta}_o), \ldots, \nabla_{\boldsymbol{\theta}} F(x_i, \boldsymbol{\theta}_o), \ldots, \nabla_{\boldsymbol{\theta}} F(x_k, \boldsymbol{\theta}_o)]'$. □

Both $Q(\cdot)$ and $\mathbf{z}$ depend on the number of groups k, so it would be more appropriate to indicate this dependence with the notation $Q(\cdot; k)$ and $\mathbf{z}_k$, but this additional notational complexity is suppressed for notational simplicity and will be implicit in what follows.

2.1 Model Estimation and Limit Theory

To estimate the unknown parameter $\boldsymbol{\theta}_*$ we first employ the empirical distribution function

$$\widehat{p}_n(x) := n^{-1}\#\{X_t \leq x\},$$

where x is generic notation for x_1, x_2, . . ., x_k. Since $p(x) = \mathbb{P}(X_t \leq x \mid b \leq X_t \leq u)$ is the conditional mean of $\widehat{p}_n(x)$, we have by standard limit theory $\sqrt{n}\{\widehat{p}_n(x)-p(x)\} \overset{\text{A}}{\sim} N[0, p(x)(1-p(x))]$ and $\sqrt{n}(\widehat{p}_n(\cdot)-p(\cdot)) \Rightarrow \mathcal{B}^o(\cdot)$, where the limit process $\mathcal{B}^o$ is a Brownian bridge. To estimate the unknown parameter $\boldsymbol{\theta}_*$ in the posited model with *cdf* $F(\cdot, \boldsymbol{\theta})$ we then perform a least squares minimum distance estimation between the nonparametric estimate $\widehat{p}(\cdot)$ and the parametric model $F(\cdot, \boldsymbol{\theta})$ over the end points of the intervals of observation, giving the MCMD estimator $\widehat{\boldsymbol{\theta}}_n := \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q_n(\boldsymbol{\theta})$, where

$$Q_n(\boldsymbol{\theta}) := \sum_{i=1}^{k} \{\widehat{p}_n(x_i) - F(x_i, \boldsymbol{\theta})\}^2 .$$

The structure of the MCMD estimator promotes analysis in terms of a nonlinear regression of $\widehat{p}_n(\cdot)$, which is a data-based uniformly consistent estimate of $p(\cdot)$, on the nonstochastic mean regressor function $F(\cdot, \boldsymbol{\theta})$ that equals $p(\cdot)$ under the null. It is convenient to maintain this regression interpretation in what follows. Note that

$$\begin{aligned} Q_n(\boldsymbol{\theta}) &= \sum_{i=1}^{k} \left[\{\widehat{p}_n(x_i) - p(x_i)\}^2 + 2\{\widehat{p}_n(x_i) - p(x_i)\}\{p(x_i) - F(x_i, \boldsymbol{\theta})\} + \{p(x_i) - F(x_i, \boldsymbol{\theta})\}^2\right] \\ &= Q(\boldsymbol{\theta}) + o_{\text{a.s.}}(1), \end{aligned}$$

where the last equality holds uniformly in $\boldsymbol{\theta}$ because (i) $\widehat{p}_n(\cdot)$ is uniformly consistent for $p(\cdot)$, and (ii) $|p(\cdot) - F(\cdot, \boldsymbol{\theta})|$ is bounded between 0 and 2 uniformly in $\boldsymbol{\theta}$. Therefore, $\arg\min_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}) + o_{\text{a.s.}}(1)$ and $\widehat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}_o = \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})$.

Theorem 1. *Given Assumption A,* $\widehat{\boldsymbol{\theta}}_n \overset{\text{a.s.}}{\to} \boldsymbol{\theta}_o$. □

When the null hypothesis is correct, we have $\boldsymbol{\theta}_o = \boldsymbol{\theta}_*$ because $F(\cdot, \boldsymbol{\theta}_*) = p(\cdot)$ under the null, whereas $\boldsymbol{\theta}_*$ is undefined under the alternative. Correct specification under the null therefore suffices to ensure that $\boldsymbol{\theta}_* = \boldsymbol{\theta}_o$. The convergence rate of $\widehat{\boldsymbol{\theta}}_n$ is determined by that of $\widehat{p}_n(\cdot)$ and is $O(\sqrt{n})$ because the empirical distribution function whose convergence rate is $O\left(\sqrt{n}\right)$ is the only data-dependent component involved in the objective function $Q_n(\cdot)$.

To find the limit distribution of the MCMD estimator we consider the usual linear approximation based on the expansion $F(x_i, \widehat{\boldsymbol{\theta}}_n) = F(x_i, \boldsymbol{\theta}_o) + \nabla'_{\boldsymbol{\theta}} F(x_i, \boldsymbol{\theta}_o)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) + o_{\mathbb{P}}(n^{-1/2})$, which implies in view of theorem 1 that

$$Q_n(\widehat{\boldsymbol{\theta}}_n) = \sum_{i=1}^{k} \{\widehat{p}_n(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n)\}^2 = \sum_{i=1}^{k} \{\underbrace{[\widehat{p}_n(x_i) - F(x_i, \boldsymbol{\theta}_o)]}_{=Y_i} - \underbrace{\nabla'_{\boldsymbol{\theta}} F(x_i, \boldsymbol{\theta}_o)}_{=\mathbf{z}'_i}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o)\}^2 + o_{\mathbb{P}}(1).$$

We define (as suggested in the brace underbars of the above equation) the variables $Y_i = \widehat{p}_n(x_i) - F(x_i, \boldsymbol{\theta}_o)$ and $\mathbf{z}_i = \nabla_{\boldsymbol{\theta}} F(x_i, \boldsymbol{\theta}_o)$, so that the leading term on the right side is a sum of the squared residuals in a regression of Y_i on $\mathbf{z}_i$ with regression coefficient $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o$. Thus, the MCMD estimator is asymptotically equivalent to the least squares regression on a linear pseudo-model involving Y_i and $\mathbf{z}_i$, viz.,

$$(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{Y} + O_{\mathbb{P}}(n^{-1}), \tag{2}$$

where $\mathbf{Y} := [Y_1, \ldots, Y_i, \ldots, Y_k]'$ and $\mathbf{z} := [\mathbf{z}_1, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_k]'$.

Since $\sqrt{n}\{\widehat{p}(\cdot) - p(\cdot)\} \Rightarrow \mathcal{B}^o(\cdot)$, we also have

$$\sqrt{n}\{Y_{(\cdot)} - y_{(\cdot)}\} = \sqrt{n}\{\widehat{p}_n(\cdot) - p(\cdot)\} \Rightarrow \mathcal{B}^o(\cdot), \tag{3}$$

where for each $i = 1, 2, \ldots, k$, we define $y_i := p(x_i) - F(x_i, \boldsymbol{\theta}_o)$. Note that for $i = 1, 2, \ldots, k$, $y_i = 0$ under the null, whereas for some i, $y_i \neq 0$ under the alternative. This difference ensures the consistency of the KS test statistic that we introduce in the next subsection. The regression coefficient $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o$ now satisfies the following property:

$$\sqrt{n}\{\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o - \underbrace{(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}}_{=\mathbf{0}}\} = \sqrt{n}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'(\mathbf{Y} - \mathbf{y}) + O_{\mathbb{P}}(n^{-1}) \Rightarrow (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{B}^o, \tag{4}$$

where $\mathbf{y} := [y_1, \ldots, y_k]'$ and $\mathbf{B}^o := [\mathcal{B}^o(p(x_1)), \ldots, \mathcal{B}^o(p(x_i)), \ldots, \mathcal{B}^o(p(x_k))]'$. The limit theory (4) appears to indicate that the MCMD estimator has an asymptotic bias involving $(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}$. However, this bias is zero, as implied by the brace underbar in the first member of (4). Specifically, note that the first-order condition for the optimum $\boldsymbol{\theta}_o$ of $Q\left(\theta\right)$ in $\boldsymbol{\Theta}$ implies that

$$\sum_{i=1}^{k}\{F(x_i,\boldsymbol{\theta}_o)-p(x_i)\}\nabla_{\boldsymbol{\theta}}F(x_i,\boldsymbol{\theta}_o)=\mathbf{0},\text{ or }\mathbf{z}'\mathbf{y}=\mathbf{0}, \tag{5}$$

which orthogonality ensures that $(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}=\mathbf{0}$. It follows from (4) that $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_o)\Rightarrow(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{B}^o$, which leads directly to the limit distribution of the MCMD estimator.

Theorem 2. *Given Assumption A,* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_o)\overset{\text{A}}{\sim}N\left[\mathbf{0},(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\boldsymbol{\Sigma}^o\mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\right]$, *where*

$$\boldsymbol{\Sigma}^o:=\begin{bmatrix} p(x_1)(1-p(x_1)) & p(x_1)(1-p(x_2)) & \cdots & p(x_1)(1-p(x_k)) \\ p(x_1)(1-p(x_2)) & p(x_2)(1-p(x_2)) & \cdots & p(x_2)(1-p(x_k)) \\ \vdots & \vdots & \ddots & \vdots \\ p(x_1)(1-p(x_k)) & p(x_2)(1-p(x_k)) & \cdots & p(x_k)(1-p(x_k)) \end{bmatrix}. \qquad \square$$

Some remarks on this result are in order. First, note that the core of this limit theory is the simple linear functional $(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{B}^o$ of $\mathbf{B}^o$, implying that the asymptotic distribution of the MCMD estimator $\widehat{\boldsymbol{\theta}}_n$ has the form of a functional of the Brownian bridge process that appears in the asymptotic null distribution of the KS test statistic with no unknown parameters. Correspondingly, the KS test statistic that depends on the use of the MCMD estimator of these unknown parameters has an asymptotic null distribution that is also a functional of the same Brownian bridge process. This result is typically quite different from the outcome of using other estimators of $\boldsymbol{\theta}$ in the KS test.

Second, the asymptotic distribution of the MCMD estimator is related to the asymptotic results in Bolthausen (1977) and Pollard (1980). In particular, Pollard (1980) derived the asymptotic distribution of the MD estimator using a general functional that extends the L_2-norm of Bolthausen (1977). The asymptotic distribution in Theorem 2 can also be derived by letting the objective function $Q_n(\boldsymbol{\theta})$ in our formulation be a special case of the general functional used in Pollard (1980) and applying Pollard's theorem 5.6 to deliver the asymptotic distribution of this general functional. The regression framework for the MCMD estimator used here enables asymptotic critical values of the limit distribution theory to be obtained by a simple simulation calculation.

2.2 Testing the Hypothesis

We now examine the KS test statistic

$$\widehat{T}_n := \sup_{i\leq k} |\sqrt{n}\{\widehat{p}_n(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n)\}|, \tag{6}$$

which has the same form as the usual KS statistic given in the literature (e.g., Durbin, 1973; Henze, 1996), the sole difference being the use of the MCMD estimator $\widehat{\boldsymbol{\theta}}_n$ in (6). We distinguish $\widehat{T}_n$ from the usual statistic with no parameter estimation error which we define as $T_n := \sup_{i\leq k} |\sqrt{n}\{\widehat{p}_n(x_i) - F(x_i, \boldsymbol{\theta}_*)\}|$.

2.2.1 The Null Limit Distribution

We first develop asymptotic theory under the null where $\boldsymbol{\theta}_o = \boldsymbol{\theta}_*$. Hence, for each x_i we have $\widehat{p}_n(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n) = \widehat{p}_n(x_i) - F(x_i, \boldsymbol{\theta}_*) - \nabla'_{\boldsymbol{\theta}} F(x_i, \boldsymbol{\theta}_*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) + o_{\mathbb{P}}(1)$, which implies that

$$\sup_{i\leq k} \sqrt{n}|\widehat{p}_n(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n)| \Rightarrow \sup_{i\leq k} |\mathcal{B}^o(p(x_i)) - \mathbf{z}'_i(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{B}^o| \tag{7}$$

by continuous mapping, where $\mathbf{B}^o$ and $\mathbf{z}$ are as in (4) and (2). The null limit distribution is therefore bounded in probability as a functional of the Gaussian process $\mathcal{B}^o$ and the component $\mathcal{B}^o(p(x_i)) - \mathbf{z}'_i(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{B}^o$ in (7) is the i-th row element of $\mathbf{m}\mathbf{B}^o$ where $\mathbf{m} := \mathbf{I} - \mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'$. Therefore,

$$\widehat{\mathbf{S}}_n := \sqrt{n}[\widehat{p}_n(x_1) - F(x_1, \widehat{\boldsymbol{\theta}}_n), ..., \widehat{p}_n(x_k) - F(x_k, \widehat{\boldsymbol{\theta}}_n)]' \Rightarrow \mathbf{m}\mathbf{B}^o \sim N(\mathbf{0}, \mathbf{m}\boldsymbol{\Sigma}^o\mathbf{m}). \tag{8}$$

The matrix $\mathbf{m}$ projects onto the orthogonal complement of the range of the $k \times d$ matrix $\mathbf{z}$, so the rank of $\mathbf{m}\boldsymbol{\Sigma}^o\mathbf{m}$ is $k - d$. For notational simplicity, let $\mathbf{G}$ and G_i denote $\mathbf{m}\mathbf{B}^o$ and the i-th row element of $\mathbf{G}$, respectively. Then, $\widehat{T}_n \Rightarrow \mathcal{Z}_k := \max[|G_1|, \ldots, |G_k|]$ under the null.

The asymptotic null distribution given by (7) differs from that of the KS test statistic with no unknown parameters because in that case

$$\mathbf{S}_n := \sqrt{n}[\widehat{p}_n(x_1) - F(x_1, \boldsymbol{\theta}_*), ..., \widehat{p}_n(x_k) - F(x_k, \boldsymbol{\theta}_*)]' \Rightarrow \mathbf{B}^o \sim N(\mathbf{0}, \boldsymbol{\Sigma}^o), \tag{9}$$

so that the KS test statistic weakly converges to $\max[|\mathcal{B}^o(p(x_1))|, |\mathcal{B}^o(p(x_2))|, \ldots, |\mathcal{B}^o(p(x_k))|]$ under the null. Hence, in the limit $\widehat{T}_n$ and T_n are constructed from different functionals of the same Brownian bridge process $\mathcal{B}^o$.

In both cases, the only stochastic component determining the null limit distribution is the Brownian bridge. The deterministic component $\mathbf{m}$ is a constant matrix that depends on the border values of the data groups, which are known, and the parameter value $\boldsymbol{\theta}_o = \boldsymbol{\theta}_*$ under the null, which may be consistently estimated. Thus, $\widehat{p}_n(\cdot)$ is the only stochastic source that determines the asymptotic null distribution of $\widehat{T}_n$, because the MCMD estimator of $\boldsymbol{\theta}_*$ can be represented asymptotically as a linear functional of the same Brownian bridge. If another estimator is used, the limit distribution typicaly involves a functional of another Gaussian process with covariance kernel different from that of the Brownian bridge. The transform device of Khmaladze's (1981, 1993, 2013) works to remove such components (by a non-orthogonal projection) that modifies the statistic so that the asymptotic null distribution is identical to that of T_n. Since the Brownian bridge component $\mathscr{B}^o$ is the only stochastic part of our limiting KS test statistic, we do not have to eliminate the parameter estimation error part from our test basis in order to construct an easily implemented test, as we now discuss.

The asymptotic null distribution of the KS test statistic can be approximated simply by estimating the covariance matrix $\mathbf{m}\boldsymbol{\Sigma}^o\mathbf{m}$. Both $\mathbf{m}$ and $\boldsymbol{\Sigma}^o$ involve $\boldsymbol{\theta}_*$ with $\mathbf{z}_i = \nabla_{\boldsymbol{\theta}} F(x_i, \boldsymbol{\theta}_*)$ and $p(x_i) = F(x_i, \boldsymbol{\theta}_*)$, so that replacing $\boldsymbol{\theta}_*$ with $\widehat{\boldsymbol{\theta}}_n$ we have the consistent estimates $\widehat{\mathbf{z}}_i := \nabla_{\boldsymbol{\theta}} F(x_i, \widehat{\boldsymbol{\theta}}_n) \overset{\text{a.s.}}{\to} \mathbf{z}_i$ and $F(x_i, \widehat{\boldsymbol{\theta}}_n) \overset{\text{a.s.}}{\to} F(x_i, \boldsymbol{\theta}_*)$ since $\widehat{\boldsymbol{\theta}}_n \overset{\text{a.s.}}{\to} \boldsymbol{\theta}_*$. Then $\widehat{\mathbf{z}} := [\widehat{\mathbf{z}}_1, \ldots, \widehat{\mathbf{z}}_k]' \overset{\text{a.s.}}{\to} \mathbf{z}'$, $\widehat{\mathbf{m}} := \mathbf{I} - \widehat{\mathbf{z}}(\widehat{\mathbf{z}}'\widehat{\mathbf{z}})^{-1}\widehat{\mathbf{z}}' \overset{\text{a.s.}}{\to} \mathbf{m}$, and $\widehat{\boldsymbol{\Sigma}}_n^o \overset{\text{a.s.}}{\to} \boldsymbol{\Sigma}^o$, where

$$\widehat{\boldsymbol{\Sigma}}_n^o := \begin{bmatrix} F(x_1, \widehat{\boldsymbol{\theta}}_n)(1 - F(x_1, \widehat{\boldsymbol{\theta}}_n)) & F(x_1, \widehat{\boldsymbol{\theta}}_n)(1 - F(x_2, \widehat{\boldsymbol{\theta}}_n)) & \cdots & F(x_1, \widehat{\boldsymbol{\theta}}_n)(1 - F(x_k, \widehat{\boldsymbol{\theta}}_n)) \\ F(x_1, \widehat{\boldsymbol{\theta}}_n)(1 - F(x_2, \widehat{\boldsymbol{\theta}}_n)) & F(x_2, \widehat{\boldsymbol{\theta}}_n)(1 - F(x_2, \widehat{\boldsymbol{\theta}}_n)) & \cdots & F(x_2, \widehat{\boldsymbol{\theta}}_n)(1 - F(x_k, \widehat{\boldsymbol{\theta}}_n)) \\ \vdots & \vdots & \ddots & \vdots \\ F(x_1, \widehat{\boldsymbol{\theta}}_n)(1 - F(x_k, \widehat{\boldsymbol{\theta}}_n)) & F(x_2, \widehat{\boldsymbol{\theta}}_n)(1 - F(x_k, \widehat{\boldsymbol{\theta}}_n)) & \cdots & F(x_k, \widehat{\boldsymbol{\theta}}_n)(1 - F(x_k, \widehat{\boldsymbol{\theta}}_n)) \end{bmatrix}.$$

The distribution of $\mathscr{Z}_k$ can be approximated by that of $\widehat{\mathscr{Z}}_k := \max[|\widehat{G}_1|, |\widehat{G}_2|, \ldots, |\widehat{G}_k|]$, where for each i, $\widehat{G}_i$ is the i-th row element of $\widehat{\mathbf{G}}$, and $\widehat{\mathbf{G}} \sim_d N[\mathbf{0}, \widehat{\mathbf{m}}\widehat{\boldsymbol{\Sigma}}^o\widehat{\mathbf{m}}]$.

2.2.2 Limit Behavior under Fixed Alternatives

For a fixed alternative, we necessarily have $y_i = p(x_i) - F(x_i, \boldsymbol{\theta}_o) \neq 0$ for some i. It then follows that under such an alternative,

$$\begin{aligned}
& \sqrt{n}\{[\widehat{p}_n(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n)] = \sqrt{n}[\widehat{p}_n(x_i) - p(x_i))] + \sqrt{n}\left[p(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n)\right] \\
= \; & \sqrt{n}[\widehat{p}_n(x_i) - p(x_i))] + \sqrt{n}\left[p(x_i) - F(x_i, \boldsymbol{\theta}_o) + \left\{F(x_i, \boldsymbol{\theta}_o) - F(x_i, \widehat{\boldsymbol{\theta}}_n)\right\}\right] \\
= \; & \sqrt{n}[\widehat{p}_n(x_i) - p(x_i))] + \sqrt{n}\left[y_i - \left\{\mathbf{z}_i'(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) + o_{\mathbb{P}}(n^{-1/2})\right\}\right] \\
= \; & \sqrt{n}[\widehat{p}_n(x_i) - p(x_i))] + \sqrt{n}\left[y_i - \left\{\mathbf{z}_i'(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'(\mathbf{Y} - \mathbf{y}) + o_{\mathbb{P}}(n^{-1/2})\right\}\right] \qquad (10) \\
= \; & O_{\mathbb{P}}(n^{1/2}) \text{ for some } i \qquad (11)
\end{aligned}$$

since $\sqrt{n}[\widehat{p}_n(x_i) - p(x_i))]$, $\sqrt{n}(\mathbf{Y} - \mathbf{y}) = O_{\mathbb{P}}(1)$ whereas $y_i - \mathbf{z}_i'(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}$ is the i-th element of $\mathbf{my}$, and at least one of the elements in $\mathbf{my}$ is different from zero under the alternative. This follows because the first-order condition (5) for $\boldsymbol{\theta}_o$ implies that $\mathbf{z}'\mathbf{y} = \mathbf{0}$, so that $\mathbf{my} = \mathbf{y}$. Therefore, the ith element of $\mathbf{my}$ is necessarily $y_i = p(x_i) - F(x_i, \boldsymbol{\theta}_o) \neq 0$ under the alternative. Then, $\widehat{T}_n = \sqrt{n}\max[|y_1|, \ldots, |y_k|] + O_{\mathbb{P}}(1) = O_{\mathbb{P}}(n^{1/2})$ as indicated in (11), and the KS test is consistent under any fixed alternative for which $y_i = p(x_i) - F(x_i, \boldsymbol{\theta}_o) \neq 0$ for some i. Thus, the KS test statistic that relies on the MCMD estimator has unit power and is consistent under fixed alternatives.

Observe that, upon recentering $\sqrt{n}\{[\widehat{p}_n(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n)]$ and using $\mathbf{z}'\mathbf{y} = \mathbf{0}$, we have from (10)

$$\begin{aligned}
& \sqrt{n}\{[\widehat{p}_n(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n)] - [y_i - \mathbf{z}_i'(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}]\} = \sqrt{n}\{[\widehat{p}_n(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n)] - y_i\} \\
= \; & \sqrt{n}[\{\widehat{p}_n(x_i) - p(x_i)\} - \left\{\mathbf{z}_i'(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'(\mathbf{Y} - \mathbf{y}) + o_{\mathbb{P}}(n^{-1/2})\right\}] \Rightarrow G_i, \qquad (12)
\end{aligned}$$

which gives the limit distribution under fixed alternatives.

2.2.3 Limit Theory for Local Alternatives

We next consider the limit behavior of the KS test under the following local alternative

$$\mathcal{H}_\ell : \mathbb{P}(X_t \leq x_i \mid x_0 \leq X_t \leq x_k) = F(x_i, \boldsymbol{\theta}_o) + h(x_i)/\sqrt{n}, \qquad (13)$$

for some uniformly bounded function $h(\cdot)$. Under this local alternative, we have $y_i = p(x_i) - F(x_i, \boldsymbol{\theta}_o) = h(x_i)/\sqrt{n}$, and $\mathbf{y} = n^{-1/2}\mathbf{h}$ where $\mathbf{h} := [h(x_1), \ldots, h(x_i), \ldots, h(x_k)]'$, so that

$$\sqrt{n}\{\widehat{p}_n(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n)\} \Rightarrow h_i - \mathbf{z}_i'(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{h} + G_i \tag{14}$$

under the local alternative by (12), where h_i is the ith element of $\mathbf{h}$ and G_i is the ith element of $\mathbf{m}\mathbf{B}^o$ as above. Evidently the component $h_i - \mathbf{z}_i'(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{h}$ in (14) is the ith element of $\mathbf{mh}$, so that in place of (8) we have

$$\widehat{\mathbf{S}}_n = \sqrt{n}[\widehat{p}_n(x_1) - F(x_1, \widehat{\boldsymbol{\theta}}_n), ..., \widehat{p}_n(x_k) - F(x_k, \widehat{\boldsymbol{\theta}}_n)]' \Rightarrow \mathbf{m}(\mathbf{h} + \mathbf{B}^o) \sim N(\mathbf{mh}, \mathbf{m}\boldsymbol{\Sigma}^o\mathbf{m}), \tag{15}$$

which reduces to the null limit theory when $\mathbf{mh} = \mathbf{0}$. Similar to the null case, the limit theory when $\boldsymbol{\theta}_o$ is not estimated is given by

$$\mathbf{S}_n := \sqrt{n}[\widehat{p}_n(x_1) - F(x_1, \boldsymbol{\theta}_o), ..., \widehat{p}_n(x_k) - F(x_k, \boldsymbol{\theta}_o)]' \Rightarrow \mathbf{h} + \mathbf{B}^o \sim N(\mathbf{h}, \boldsymbol{\Sigma}^o), \tag{16}$$

in place of (9).

Defining $\xi_i = h(x_i) - \mathbf{z}_i'(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{h}$ and using the fact noted above that $\mathbf{y} = n^{-1/2}\mathbf{h}$, it is apparent that $\xi_i = h(x_i)$ because $\mathbf{z}'\mathbf{y} = n^{-1/2}\mathbf{z}'\mathbf{h} = \mathbf{0}$ from first-order conditions for $\boldsymbol{\theta}_o$. It also trivially follows that $\mathbf{mh}$ is necessarily different from $\mathbf{0}$ whenever $h_i \neq 0$ for some i, thereby ensuring that the local alternative differs from the null. Further, the KS test statistic has the following limit

$$\widehat{T}_n \Rightarrow \mathcal{Z}_k^a := \max[|\xi_1 + G_1|, \ldots, |\xi_k + G_k|]. \tag{17}$$

Thus, tests based on $\widehat{T}_n$ have non-negligible power under the local alternative $\mathcal{H}_\ell$ in (13). We further note that the weak limit of T_n under the local alternative is a functional of the same form but one that involves $(\mathbf{h} + \mathbf{B}^o)$ rather than $\mathbf{m}(\mathbf{h} + \mathbf{B}^o)$, as is apparent by comparing (15) and (16). It follows that estimation of the parameter $\boldsymbol{\theta}$ using MCMD modifies the limit distribution of T_n by scaling the components entering (17) with the projector $\mathbf{m} = \mathbf{I} - \mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'$.

We summarize the key claims of this section in the following theorem.

Theorem 3. *Given Assumption A,*

(i) $\widehat{T}_n \Rightarrow \mathcal{Z}_k := \max[|G_1|, \ldots, |G_k|]$ *under* $\mathcal{H}_0$*;*

(ii) $\widehat{T}_n = \sqrt{n}\max[|y_1|, \ldots, |y_k|] + O_{\mathbb{P}}(1)$ *under* $\mathcal{H}_1$*; and*

(iii) if $|h(\cdot)|$ is uniformly bounded on $\{x_1,\ldots,x_k\}$, $\widehat{T}_n \Rightarrow \mathcal{Z}_k^a := \max[|h(x_1)+G_1|,\ldots,|h(x_k)+G_k|]$ under $\mathcal{H}_\ell$. □

Before moving to the next section, we discuss some aspects of this test and its implementation. Two methods are available to compute critical values of the test. The first method is to estimate the idempotent matrix $\mathbf{m}$ and $\mathbf{\Sigma}^o$ by a plug in method using $\widehat{\mathbf{m}} := \mathbf{I}-\widehat{\mathbf{z}}(\widehat{\mathbf{z}}'\widehat{\mathbf{z}})^{-1}\widehat{\mathbf{z}}'$ and $\widehat{\mathbf{\Sigma}}^o$, as mentioned above. This method produces valid critical values asymptotically by virtue of the invariance principle (Donsker, 1951), consistency of $\widehat{\boldsymbol{\theta}}_n$ under the null, and continuous mapping. In practice, the process $\mathcal{B}^o(\cdot)$ can be evaluated on the unit interval at the points $F(x_i;\widehat{\boldsymbol{\theta}}_n)$ $(i=1,2,\ldots,k)$ and the functional $\widehat{\mathbf{m}}[\mathcal{B}^o(F(x_1;\widehat{\boldsymbol{\theta}}_n)),\ldots,\mathcal{B}^o(F(x_k;\widehat{\boldsymbol{\theta}}_n))]'$ can be used to approximate the weak limit. An alternative method is to apply a parametric bootstrap by generating data with n number of observations from $F(\cdot,\widehat{\boldsymbol{\theta}_n})$ and computing the null distribution by iteratively replicating the test many times. We examine these two approaches in Section 4 and compare their performance with the KS test statistic that is based on the use of the (Q)ML estimator.

Pollard (1980) provided a general theory on the asymptotic distribution of the MD estimator and the GOF test statistic for both of which the same norm is assumed. The results given in Theorem 3 are closely related but differ in that the CM distance is used for parameter estimation of $\boldsymbol{\theta}$, whereas the KS distance is used for testing goodness-of-fit. This approach offers the advantage of a regression formulation of the KS test and convenient simulation-based calculation of asymptotic critical values for the test.

3 Hypothesis Testing using Grouped Data with Large Group Size

This section extends the analysis to the KS test statistic formed with continuously distributed data. We exploit the large sample size weak limit theory of the KS test statistic given in Section 2 by using sequential asymptotics in which large sample size asymptotics with $n \to \infty$ are followed by infill asymptotics in which the data range $u-b$ is fixed but the group interval is reduced. Then the sequential weak limit of the KS test can be linked to the large sample size limit of the KS test for continuously distributed data.

For convenience of notation, we distinguish symptotics in which the group size k tends to infinity from those in which the sample size n tends to infinity by affixing 'k' and 'n' to the relevant weak convergence symbols. Thus, '$\overset{n}{\Rightarrow}$' and '$\overset{k}{\Rightarrow}$' denote weak convergence in which $n \to \infty$ and $k \to \infty$, respectively.

3.1 Estimation Limit Theory with Large Group Size

We develop a large group size limit theory for the MCMD estimator with the data range fixed, and proceed by examining the corresponding limits of the components in Theorem 1. First, for technical convenience, we suppose that the interval distance c_k is the same for each group, so that $c_k \cdot k = u - b$. Next, let $\bar{F}(\cdot) := F((1-(\cdot))b + (\cdot)u; \boldsymbol{\theta}_o)$, which is defined on the unit interval, and for $\ell = 1, 2, \ldots, k$, $j = 1, 2, \ldots, d$, and $i = 0, 1$, define

$$\partial_j^i \bar{F}_k(x) := \begin{cases} \partial_j^i \bar{F}(\ell/k), & \text{if } x \in [(\ell-1)/k, \ell/k); \\ \partial_j^i \bar{F}(1), & \text{if } x = 1. \end{cases}$$

Note that $\partial_j^i \bar{F}_k(\cdot)$ is continuous from the right with limits from the left. As k tends to infinity with the distance between x_0 and x_k being constant, $\partial_j^i \bar{F}_k(\cdot)$ converges uniformly to $\partial_j^i \bar{F}(\cdot)$, provided that $\partial_j^i \bar{F}(\cdot)$ is continuous on $[0,1]$. Therefore, the large group size limit of the i-row and j-column element of $k^{-1}\mathbf{z}'\mathbf{z}$ is obtained as follows. Observe that

$$\frac{1}{k}\sum_{\ell=1}^{k} \partial_i F(x_\ell; \boldsymbol{\theta}_o)\partial_j F(x_\ell; \boldsymbol{\theta}_o) = \int_0^1 \partial_i \bar{F}_k(x)\partial_j \bar{F}_k(x)dx \xrightarrow{k} \int_0^1 \partial_i \bar{F}(x)\partial_j \bar{F}(x)dx, \tag{18}$$

which holds by monotone convergence. As the group size $k \to \infty$, the group interval $c_k \to 0$ when the range $u - b$ is held constant. Therefore, as k increases, the group interval decreases. If we further let $\nabla_{\boldsymbol{\theta}} \bar{F}(\cdot) := [\partial_1 \bar{F}(\cdot), \ldots, \partial_d \bar{F}(\cdot)]'$ and $\nabla_{\boldsymbol{\theta}} \bar{F}_k(\cdot) := [\partial_1 \bar{F}_k(\cdot), \ldots, \partial_d \bar{F}_k(\cdot)]'$, we obtain as in (18)

$$\mathbf{A}_k := \frac{1}{k}\mathbf{z}'\mathbf{z} = \int_0^1 \nabla_{\boldsymbol{\theta}} \bar{F}_k(x)\nabla'_{\boldsymbol{\theta}} \bar{F}_k(x)dx \xrightarrow{k} \mathbf{A}_o := \int_0^1 \nabla_{\boldsymbol{\theta}} \bar{F}(x)\nabla'_{\boldsymbol{\theta}} \bar{F}(x)dx.$$

Next, we examine the large group size limit of $k^{-1}\mathbf{z}'\mathbf{B}^o$. Let $\bar{p}(\cdot) := p((1-(\cdot))b + (\cdot)u)$ and for $\ell = 1, 2, \ldots, k$, set

$$\bar{p}_k(x) := \begin{cases} \bar{p}(\ell/k), & \text{if } x \in [(\ell-1)/k, \ell/k); \\ 1, & \text{if } x = 1. \end{cases}$$

As $k \to \infty$, $\bar{p}_k(\cdot)$ converges uniformly to $\bar{p}(\cdot)$, provided that $\bar{p}(\cdot)$ is continuous on $[0,1]$. We also let $\bar{\mathcal{B}}^o(\cdot) := \mathcal{B}^o(\bar{p}(\cdot))$ and for $\ell = 1, 2, \ldots, k$, set

$$\bar{\mathcal{B}}_k^o(x) := \begin{cases} \bar{\mathcal{B}}^o(\ell/k), & \text{if } x \in [(\ell-1)/k, \ell/k); \\ 0, & \text{if } x = 1. \end{cases}$$

Since $\bar{\mathcal{B}}^o(\cdot)$ is a continuous process on $[0,1]$ almost surely, $\bar{\mathcal{B}}_k^o(\cdot)$ is uniformly bounded and also uniformly converges to $\bar{\mathcal{B}}^o(\cdot)$ with probability 1. Therefore, we find that

$$\mathbf{Z}_k := \frac{1}{k}\mathbf{z}'\mathbf{B}^o = \int_0^1 \nabla_{\boldsymbol{\theta}} \bar{F}_k(x)\bar{\mathcal{B}}_k^o(x)dx \xrightarrow{k} \mathbf{Z} := \int_0^1 \nabla_{\boldsymbol{\theta}} \bar{F}(x)\bar{\mathcal{B}}^o(x)dx,$$

with probability 1, which implies that $\mathbf{Z}_k \overset{k}{\Rightarrow} \mathbf{Z}$. The large group size weak limit is therefore a normally distributed random variable $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{B}_o)$, with covariance matrix

$$\mathbf{B}_o := \int_0^1 \int_0^{x'} \bar{p}(x)(1-\bar{p}(x'))\nabla_{\boldsymbol{\theta}}\bar{F}(x)\nabla'_{\boldsymbol{\theta}}\bar{F}(x')dxdx' + \int_0^1 \int_{x'}^1 \bar{p}(x')(1-\bar{p}(x))\nabla_{\boldsymbol{\theta}}\bar{F}(x)\nabla'_{\boldsymbol{\theta}}\bar{F}(x')dxdx'.$$

The following additional conditions are imposed to deliver the asymptotic behavior of the KS test.

Assumption B. *(i) The interval size of each group is identical and equal to $c_k = (u-b)/k$, and $u-b$ is constant;*

(ii) For each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $j = 1, 2, \ldots, d$, $F(\cdot; \boldsymbol{\theta})$ and $\partial_j F(\cdot; \boldsymbol{\theta})$ are continuous on $[b, u]$;

(iii) $p(\cdot)$ is continuous on $[b, u]$;

(iv) $\int_0^1 \nabla_{\boldsymbol{\theta}}\bar{F}(x)\nabla'_{\boldsymbol{\theta}}\bar{F}(x)dx$ is finite and positive definite; and

(v) $\mathbf{B}_o$ is finite and positive definite. □

The equal group interval size Assumption B(*i*) is technically convenient but might well be weakened and is by no means necessary to derive the large group size limit results of the KS test statistic. Assumptions B(*ii* and *iii*) are useful because as continuous functions defined on a bounded space they are integrable. These conditions ensure that the stated limits $\mathbf{A}_o$, $\mathbf{B}_o$, and $\mathbf{Z}$ are all well defined as the group size k tends to infinity. Assumption B(*iii*) is redundant under the null, because $p(\cdot) = F(\cdot; \boldsymbol{\theta}_*)$ and $\boldsymbol{\theta}_* = \boldsymbol{\theta}_0$, so that (*ii*) implies (*iii*). Assumptions B(*iv)* and (*v*) are standard conditions that ensure the sequential limit distribution of the KS test statistic behaves regularly.

The stated results are collected in the following lemma.

Lemma 1. *Given the Assumptions A and B,*

(i) $k^{-1}\mathbf{z}'\mathbf{z} \overset{k}{\to} \mathbf{A}_o := \int_0^1 \nabla_{\boldsymbol{\theta}}\bar{F}(x)\nabla'_{\boldsymbol{\theta}}\bar{F}(x)dx$; and

(ii) $k^{-1}\mathbf{z}'\mathbf{B}^o \overset{k}{\Rightarrow} \mathbf{Z} := \int_0^1 \nabla_{\boldsymbol{\theta}}\bar{F}(x)\bar{\mathcal{B}}^o(x)dx \sim N(\mathbf{0}, \mathbf{B}_o)$. □

Some remarks are warranted. First, from the first-order condition for $\boldsymbol{\theta}_o$, $\mathbf{z}'\mathbf{y} = \mathbf{0}$ holds uniformly in k. Furthermore, $k^{-1}\mathbf{z}'\mathbf{y}$ has the following large group size limit as $k \to \infty$

$$\frac{1}{k}\mathbf{z}'\mathbf{y} = \int_0^1 \nabla_{\boldsymbol{\theta}}\bar{F}_k(x)[\bar{p}_k(x) - \bar{F}_k(x)]dx \overset{k}{\to} \int_0^1 \nabla_{\boldsymbol{\theta}}\bar{F}(x)[\bar{p}(x) - \bar{F}(x)]dx, \tag{19}$$

therefore implying that $\int_0^1 \nabla_{\boldsymbol{\theta}}\bar{F}(x)\bar{p}(x)dx = \int_0^1 \nabla_{\boldsymbol{\theta}}\bar{F}(x)\bar{F}(x)dx$. Second, Lemma 1 implies a straightforward large group size weak limit for the MCMD estimator. The following theorem trivially holds by joint convergence.

Theorem 4. *Given Assumptions A and B,* $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \overset{n}{\Rightarrow} \mathbf{W}_k := (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{B}^o \overset{k}{\Rightarrow} \mathbf{W} := \mathbf{A}_o^{-1}\mathbf{Z} \sim N(\mathbf{0}, \mathbf{C}_o)$ *with* $\mathbf{C}_o := \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}$. □

This result corresponds to theorem 5.1 of Bolthausen (1977), in which the asymptotic distribution of the MD estimator obtained using CM distance is derived. Theorem 4 shows that theorem 5.1 of Bolthausen (1977) can also be obtained in a regression context. As we shown in the next subsection, the same mode of analysis suggests a way to obtain symptotic critical values for the KS test.

The asymptotic results in Theorem 4 are obtained by assuming that the data range $[u, b]$ is fixed. Data and models with unbounded range may be similarly analyzed by transforming group border values using the probability integral transform, so that the standard uniform distribution is set as the null distribution.

3.2 Testing Hypotheses with Large Group Size

Using the large group size weak limit result for the MCMD estimator given in Theorem 4, we examine the large group size limit distributions of the KS test statistic under the null and local alternatives. Note that $\widehat{T}_n$ is not bounded in probability under the alternative as shown in Section 2.

3.2.1 Null Limit Theory

We first examine the large group size null limit distribution of the KS test statistic. We start the discussion from (7). Defining

$$\bar{\mathcal{G}}_k^o(\cdot) := \bar{\mathcal{B}}_k^o(\cdot) - \nabla'_{\boldsymbol{\theta}}\bar{F}_k(\cdot)\mathbf{A}_k^{-1}\mathbf{Z}_k \quad \text{and} \quad \bar{\mathcal{G}}^o(\cdot) := \bar{\mathcal{B}}^o(\cdot) - \nabla'_{\boldsymbol{\theta}}\bar{F}(\cdot)\mathbf{A}_o^{-1}\mathbf{Z},$$

then $\bar{\mathcal{G}}_k^o(\cdot) \overset{k}{\to} \bar{\mathcal{G}}^o(\cdot)$ uniformly with probability 1, because $\bar{F}_k(\cdot) \overset{k}{\to} \bar{F}(\cdot)$ and $\bar{\mathcal{B}}_k^o(\cdot) \overset{k}{\to} \bar{\mathcal{B}}^o(\cdot)$ uniformly on $[0, 1]$ with probability 1. Furthermore, for each i, $\mathcal{B}^o(p(x_i)) - \mathbf{z}_i'(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{B}^o = \bar{\mathcal{G}}_k^o(i/k)$, so that $\mathcal{Z}_k = \sup_{i \le k} |\bar{\mathcal{G}}_k^o(i/k)| \overset{k}{\to} \mathcal{Z} := \sup_{z \in [0,1]} |\bar{\mathcal{G}}^o(z)|$ with probability 1, viz., the large group size null weak limit of $\mathcal{Z}_k$ is obtained as a functional of $\bar{\mathcal{G}}^o(\cdot)$ such that for each x, $\mathbb{E}[\bar{\mathcal{G}}^o(x)] = 0$, and if $x \le \tilde{x}$,

$$\mathbb{E}[\bar{\mathcal{G}}^o(x)\bar{\mathcal{G}}^o(\tilde{x})] = \bar{p}(x)(1 - \bar{p}(\tilde{x})) - \nabla'_{\boldsymbol{\theta}}\bar{F}(x)\mathbf{A}_o^{-1}\mathbf{D}(\tilde{x}) - \nabla'_{\boldsymbol{\theta}}\bar{F}(\tilde{x})\mathbf{A}_o^{-1}\mathbf{D}(x) + \nabla'_{\boldsymbol{\theta}}\bar{F}(x)\mathbf{C}_o\nabla_{\boldsymbol{\theta}}\bar{F}(\tilde{x}),$$

where for each $x \in [0, 1]$, $\mathbf{D}(x) := (1 - \bar{p}(x)) \int_0^x \bar{p}(z)\nabla_{\boldsymbol{\theta}}\bar{F}(z)dz + \bar{p}(x) \int_x^1 (1 - \bar{p}(z))\nabla_{\boldsymbol{\theta}}\bar{F}(z)dz$. This covariance kernel corresponds to that of theorem 1 of Durbin (1973). The difference is that Durbin's Gaussian process is derived as a variation of the Brownian bridge affected by the limit distribution of the ML estimator, whereas our result arises from MCMD estimation.

The gain from using MCMD estimation is that the distribution of $\bar{\mathcal{G}}^o(\cdot)$ is easily simulated: the distribution of $\bar{\mathcal{G}}^o(\cdot)$ can still be approximated by applying Donsker's (1951) invariance principle. If we let

$$B^o_\ell(\cdot) := \frac{1}{\sqrt{\ell}} \sum_{j=1}^{\lceil \ell(\cdot) \rceil} U_t - \frac{(\cdot)}{\sqrt{\ell}} \sum_{j=1}^{\ell} U_t, \tag{20}$$

$B^o_\ell(\cdot) \Rightarrow \mathcal{B}^o(\cdot)$ as $\ell \to \infty$, where the ceiling function $\lceil \cdot \rceil$ gives the smallest integer greater than or equal to its argument, and $U_j \sim$ IID $N(0,1)$. Therefore, we may approximate $\bar{\mathcal{G}}^o(\cdot)$ by

$$\widehat{\mathcal{G}}^o_\ell(\cdot) := \widehat{B}^o_\ell(\cdot) - \nabla'_{\boldsymbol{\theta}} \widehat{F}_\ell(\cdot) \widehat{\mathbf{A}}^{-1}_{o,\ell} \widehat{\mathbf{Z}}_\ell,$$

where for each $i = 1, 2, \ldots, \ell$,

$$\widehat{B}^o_\ell(i/\ell) := B^o_\ell(F((1-i/\ell)b + (i/\ell)u, \widehat{\boldsymbol{\theta}}_n)); \quad \nabla_{\boldsymbol{\theta}} \widehat{F}_\ell(i/\ell) := \nabla_{\boldsymbol{\theta}} F((1-i/\ell)b + (i/\ell)u, \widehat{\boldsymbol{\theta}}_n);$$

$$\widehat{\mathbf{A}}_{o,\ell} := \ell^{-1} \sum_{i=1}^{\ell} \nabla_{\boldsymbol{\theta}} \widehat{F}_\ell(i/\ell) \nabla'_{\boldsymbol{\theta}} \widehat{F}_\ell(i/\ell); \quad \text{and} \quad \widehat{\mathbf{Z}}_\ell := \ell^{-1} \sum_{i=1}^{\ell} \nabla_{\boldsymbol{\theta}} \widehat{F}_\ell(i/\ell) \widehat{B}^o_\ell(i/\ell).$$

By iteratively generating $\sup_{i \leq \ell} |\widehat{\mathcal{G}}^o_\ell(i/\ell)|$ many times, an approximate distribution for $\widehat{T}_n$ can be derived, with improvements in the approximation occurring as ℓ becomes large. This simulation method is effective in practice because the parameter estimation error is linked to the same Brownian bridge as that obtained for the empirical process. For other estimators, this type of linkage in the limit theory is not obvious and so cannot be relied upon in simulations.

3.2.2 Local Alternative Limit Theory

We next examine the large group size local alternative limit distribution of $\widehat{T}_n$. For this purpose, suppose $h(\cdot)$ is a continuous function on $[b, u]$, let $\bar{h}(\cdot) := h((1-(\cdot))b + (\cdot)u)$, and for $\ell = 1, 2, \ldots, k$, define

$$\bar{h}_k(x) := \begin{cases} \bar{h}(\ell/k), & \text{if } x \in [(\ell-1)/k, \ell/k); \\ \bar{h}(1), & \text{if } x = 1. \end{cases}$$

As $k \to \infty$, $\bar{h}_k(\cdot)$ converges uniformly to $\bar{h}(\cdot)$ and is uniformly bounded on $[0,1]$, because $\bar{h}(\cdot)$ is a continuous function on a compact interval. Hence,

$$\mathbf{Q}_k := \frac{1}{k}\mathbf{z}'\mathbf{h} = \frac{1}{k} \int_0^1 \nabla_{\boldsymbol{\theta}} \bar{F}_k(x) \bar{h}_k(x) \xrightarrow{k} \mathbf{Q} := \int_0^1 \nabla_{\boldsymbol{\theta}} \bar{F}(x) \bar{h}(x) dx.$$

Note that (19) implies that the right side is **0** from the local alternative that $\bar{p}(\cdot) - \bar{F}(\cdot) = h(\cdot)/\sqrt{n}$. Therefore, if we let

$$\bar{\xi}_k(\cdot) := \bar{h}_k(\cdot) - \nabla'_{\boldsymbol{\theta}} \bar{F}_k(\cdot)\mathbf{A}_k^{-1}\mathbf{Q}_k,$$

for each $i = 1, 2, \ldots, k$, $\xi_i = \bar{h}_k(i/k) - \nabla'_{\boldsymbol{\theta}} \bar{F}_k(i/k)\mathbf{A}_k^{-1}\mathbf{Q}_k$ and $\bar{\xi}_k(\cdot) \overset{k}{\to} \bar{\xi}(\cdot) := \bar{h}(\cdot) - \nabla'_{\boldsymbol{\theta}} \bar{F}(\cdot)\mathbf{A}^{-1}\mathbf{Q} \equiv \bar{h}(\cdot)$ uniformly on $[0, 1]$, so that $\bar{\xi}_k(\cdot) + \bar{\mathcal{G}}_k^o(\cdot) \overset{k}{\to} \bar{\xi}(\cdot) + \bar{\mathcal{G}}^o(\cdot) = \bar{h}(\cdot) + \bar{\mathcal{G}}^o(\cdot)$ uniformly on $[0, 1]$ with probability 1. Hence, letting $\mathcal{Z}_k^a := \sup_{i \leq k} |\bar{\xi}_k(i/k) + \bar{\mathcal{G}}_k^o(i/k)|$ and $\mathcal{Z}^a := \sup_{z \in [0,1]}, |\bar{\xi}(z) + \bar{\mathcal{G}}^o(z)|$, the sequential weak limit of $\widehat{T}_n$ is obtained as $\mathcal{Z}_k^a \overset{k}{\to} \mathcal{Z}^a$ with probability 1, which implies that $\mathcal{Z}_k^a \overset{k}{\Rightarrow} \mathcal{Z}^a$. Thus the localizing parameter of the limit Gaussian process shifts from zero under the null to $\bar{\xi}(\cdot)$ under the local alternative, which is identical to $\bar{h}(\cdot)$. It is therefore apparent that, if $h(\cdot) \neq 0$, the KS test statistic has non-negligible local power asymptotically.

Thes results are summarized in the following theorem.

Theorem 5. *Given Assumptions A and B,*

(i) $\widehat{T}_n \overset{n}{\Rightarrow} \mathcal{Z}_k := \sup_{i \leq k} |\bar{\mathcal{G}}_k^o(i/k)| \overset{k}{\Rightarrow} \mathcal{Z} := \sup_{z \in [0,1]} |\bar{\mathcal{G}}^o(z)|$ *under* $\mathcal{H}_0$*; and*

(ii) if $h(\cdot)$ *is continuous on* $[b, u]$, $\widehat{T}_n \overset{n}{\Rightarrow} \mathcal{Z}_k^a := \sup_{i \leq k} |\bar{h}_k(i/k) + \bar{\mathcal{G}}_k^o(i/k)| \overset{k}{\Rightarrow} \mathcal{Z}^a := \sup_{z \in [0,1]} |\bar{h}(z) + \bar{\mathcal{G}}^o(z)|$ *under* $\mathcal{H}_\ell$*.* □

4 Simulations

This section reports simulations conducted to assess the relevance of the asymptotic theory in finite samples. For this purpose, we suppose that a Pareto distribution is hypothesized for positively valued grouped data. Specifically, the hypothetical data distribution for X_t is given by

$$\mathbb{P}(X_t \leq x) = 1 - (x_{\min}/x)^{\theta},$$

where $x_{\min}$ is the minimal value of X_t, and θ is the shape parameter.

Given this model assumption, we further suppose that data are grouped in the format of (1) such that b is greater than or equal to $x_{\min}$, and u is finite. This framework implies that the unconditional distribution is modified to the conditional distribution

$$\mathbb{P}(X_t \leq x_i \mid b \leq X_t \leq u) = 1 - \frac{(u/x_i)^{\theta} - 1}{(u/b)^{\theta} - 1}.$$

We denote this distribution as Pareto(θ) and let the right side of the equation be the model for the grouped data. That is, for $i = 1, 2, \ldots, k$,

$$F(x_i, \theta) = 1 - \frac{(u/x_i)^\theta - 1}{(u/b)^\theta - 1}. \tag{21}$$

In our simulations, we use the following parameter settings: the bounds are $b = 1.0$, and $u = 10.0$; for every $i = 1, 2, \ldots, k$, the interval length is $x_i - x_{i-1} = c_k$ and we consider four cases for $c_k \in [0.1, 0.2, 0.5, 1.0]$. For data generated according to this schematic, we examine the finite sample properties of the KS test statistic under the null, alternative, and local alternative hypotheses.

4.1 Testing under the Null Hypothesis

We implement the following procedures for examining the KS test statistic under the null. First, we let $\theta_* = 2.0$ and generate n observations with conditional distribution as in (21). Six sample sizes are considered: 100, 200, 400, 600, 800, and 1,000. Second, we consider three approaches to assess the adequacy of the limit theory in Theorem 3(*i*), and we call these methods A, B, and C, respectively. Method A first generates the asymptotic null distribution in Theorem 3(*i*) by the simulation method in Section 3. Specifically, $\mathbf{m} = \mathbf{I}_k - \mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}$, and $\mathbf{z}$ is a vector with ith element

$$z_i = \left(\frac{u}{b}\right)^{\theta_*} \log\left(\frac{u}{b}\right) \left[\left(\frac{u}{x_i}\right)^{\theta_*} - 1\right] - \left(\frac{u}{x_i}\right)^{\theta_*} \log\left(\frac{u}{x_i}\right) \left[\left(\frac{u}{b}\right)^{\theta_*} - 1\right].$$

Next, the Brownian bridge $\mathcal{B}^o(\cdot)$ is generated by the invariance principle (Donsker, 1951) with $\ell = 10,000$ of (20), and we compute $\mathbf{m}\mathfrak{B}^o$, where $\mathfrak{B}^o$ is a $k \times 1$ vector with ith element $B^o_\ell(F(x_i, \theta_*))$ $(i = 1, 2, \ldots, k)$. Finally, we iterate this process 200 times and compare the distribution obtained by this process with the test statistic value. The whole process is iterated 5,000 times, and we compute the average rejection rate. Note that method A generates the asymptotic null distribution by assuming that θ_* is known, so that it cannot be used in practical work with empirical data. Nonetheless, method A is useful in corroborating Theorem 3(*i*). Method B estimates $\mathbf{m}$ and $\mathbf{B}^o$ using the MCMD estimator. Instead of θ_*, we iterate the same process using $\widehat{\theta}_n$. Specifically, for each $i = 1, 2, \ldots, k$, we let

$$\widehat{z}_i = \left(\frac{u}{b}\right)^{\widehat{\theta}_n} \log\left(\frac{u}{b}\right) \left[\left(\frac{u}{x_i}\right)^{\widehat{\theta}_n} - 1\right] - \left(\frac{u}{x_i}\right)^{\widehat{\theta}_n} \log\left(\frac{u}{x_i}\right) \left[\left(\frac{u}{b}\right)^{\widehat{\theta}_n} - 1\right]$$

be the ith element of $\widehat{\mathbf{z}}$ and estimate $\mathbf{m}$ by $\widehat{\mathbf{m}} := \mathbf{I}_k - \widehat{\mathbf{z}}(\widehat{\mathbf{z}}'\widehat{\mathbf{z}})^{-1}\widehat{\mathbf{z}}'$. We also estimate $\mathfrak{B}^o$ by $\widehat{\mathfrak{B}}^o := [\ldots, B^o_\ell(F(x_i, \widehat{\theta}_n)), \ldots]'$, and compare $\widehat{T}_n$ with the asymptotic critical values implied by $\widehat{\mathbf{m}}\widehat{\mathfrak{B}}^o$. The null distribution is obtained by independently generating $\widehat{\mathbf{m}}\widehat{\mathfrak{B}}^o$ 200 times, and we compute the empirical rejection rate of the KS test by iterating the whole process 5,000 times. Method C implements the parametric bootstrap. The bootstrap iteration number is 200, and the entire number of replications is again 5,000.

In addition to our KS test statistic, we also apply the (Q)ML estimator to the same data and compare its performance with our KS test statistic. For the (Q)ML procedure the following KS test statistic is computed:

$$\widetilde{T}_n := \sup_{i \leq k} |\sqrt{n}\{\widehat{p}_n(x_i) - F(x_i, \widetilde{\theta}_n)\}|,$$

where $\widetilde{\theta}_n$ denotes the (Q)ML estimator. For method A, we suppose that parameter estimation error is absent and obtain the empirical rejection rates by iteratively simulating

$$\max[|B^o_\ell(p(x_1))|, |B^o_\ell(p(x_2))|, \ldots, |B^o_\ell(p(x_k))|]$$

200 times. This method is conducted to explore how use of the ML estimator affects the asymptotic null distribution. Next, we also apply method B to the KS test statistic when it is computed using the (Q)ML estimator. In this case, we iteratively simulate

$$\max[|B^o_\ell(F(x_1, \widetilde{\theta}_n))|, |B^o_\ell(F(x_2, \widetilde{\theta}_n))|, \ldots, |B^o_\ell(F(x_k, \widetilde{\theta}_n))|]$$

200 times and obtain the corresponding critical values. Third, we apply the parametric bootstrap. Finally, Khmaladze's (2013) distribution-free test is applied. Specifically, θ_* is estimated by (Q)ML, and the following KS test statistic is computed based upon the transformation:

$$\ddot{T}_n := \max_{s \leq k} \left| \sum_{j=1}^{s} \widetilde{Z}_{n,j} \right|,$$

where $\widetilde{Z}_{n,j}$ is the jth-row element of $\widetilde{\mathbf{Z}}_n := \widetilde{\mathbf{Y}}_n - \widetilde{\mathbf{Y}}'_n \mathbf{a}_3(\mathbf{a}_3 - \mathbf{b}_3) - \widetilde{\mathbf{Y}}'_n \mathbf{a}_4(\mathbf{a}_4 - \mathbf{b}_4)$ with $\widetilde{\mathbf{Y}}_n := [\widetilde{Y}_{n,1}, \widetilde{Y}_{n,2}, \ldots, \widetilde{Y}_{n,k}]'$,

$$\widetilde{Y}_{n,j} := \frac{\#\{X_t \in (x_i, x_{i+1}]\} - n\widetilde{c}_j}{\sqrt{n\widetilde{c}_j}},$$

and $\widetilde{c}_j := F(x_i, \widetilde{\theta}_n) - F(x_{i-1}, \widetilde{\theta}_n)$; $\mathbf{a}_3 := \widetilde{\mathbf{a}}_3/(\widetilde{\mathbf{a}}'_3\widetilde{\mathbf{a}}_3)^{1/2}$ with $\widetilde{\mathbf{a}}_3 := \mathbf{r} - (\mathbf{r}'\widetilde{\mathbf{q}})\widetilde{\mathbf{q}} - (\mathbf{r}'\widehat{\mathbf{q}})\widehat{\mathbf{q}}$, $\mathbf{r} := [1, 0, \ldots, 0]'$, $\widetilde{\mathbf{q}} := [\sqrt{\widetilde{c}_1}, \sqrt{\widetilde{c}_2}, \ldots, \sqrt{\widetilde{c}_k}]'$, and $\widehat{\mathbf{q}} := \ddot{\mathbf{q}}/(\ddot{\mathbf{q}}'\ddot{\mathbf{q}})^{1/2}$ with $\ddot{\mathbf{q}} := [\widetilde{d}_1/\sqrt{\widetilde{c}_1}, \widetilde{d}_2/\sqrt{\widetilde{c}_2}, \ldots, \widetilde{d}_k/\sqrt{\widetilde{c}_k}]'$ and $\widetilde{d}_i := (\partial/\partial\theta)F(x_i, \widetilde{\theta}_n) - (\partial/\partial\theta)F(x_{i-1}, \widetilde{\theta}_n)$; $\mathbf{a}_4 := \widetilde{\mathbf{a}}_4/(\widetilde{\mathbf{a}}'_4\widetilde{\mathbf{a}}_4)^{1/2}$ with $\widetilde{\mathbf{a}}_4 := \widehat{\mathbf{r}} - (\widehat{\mathbf{r}}'\widetilde{\mathbf{q}})\widetilde{\mathbf{q}} - (\widehat{\mathbf{r}}'\widehat{\mathbf{q}})\widehat{\mathbf{q}} - (\widehat{\mathbf{r}}'\mathbf{a}_3)\mathbf{a}_3$ and $\widehat{\mathbf{r}} := [0, 1, 0, \ldots, 0]'$; $\mathbf{b}_3 := \widetilde{\mathbf{b}}_3/(\widetilde{\mathbf{b}}'_3\widetilde{\mathbf{b}}_3)^{1/2}$ with $\widetilde{\mathbf{b}}_3 := \widetilde{\mathbf{q}} - (\widetilde{\mathbf{q}}'\mathbf{r})\mathbf{r} - (\widetilde{\mathbf{q}}'\widehat{\mathbf{r}})\widehat{\mathbf{r}}$; and $\mathbf{b}_4 := \widetilde{\mathbf{b}}_4/(\widetilde{\mathbf{b}}'_4\widetilde{\mathbf{b}}_4)^{1/2}$ with $\widetilde{\mathbf{b}}_4 := \widehat{\mathbf{q}} - (\widehat{\mathbf{q}}'\mathbf{r})\mathbf{r} - (\widehat{\mathbf{q}}'\widehat{\mathbf{r}})\widehat{\mathbf{r}} - (\widehat{\mathbf{q}}\mathbf{b}_3)\mathbf{b}_3$. Then, $\ddot{T}_n$ weakly converges to $\widetilde{Z} := \max_{s \leq k} |\sum_{j=3}^{s} Z_j|$ under the null by Khmaladze's (2013) corollary 4, where $Z_j \sim$ IID $N(0,1)$. The asymptotic critical values are obtained by simulating the limit random variable 1 million times. We call this approach method D.

Tables 1 and 2 contain the empirical rejection rates of $\widehat{T}_n$ and $\widetilde{T}_n$, respectively. The simulation results can be summarized as follows.

1. The simulation results in Table 1 generally well support the theory given in Theorem 3(i). The nominal rejection rates in Table 1 are consistently well estimated by the empirical rejection rates, and more precise empirical rejection rates are obtained as the sample size increases. In particular, the simulation results using methods A and B reveal that the test statistic computed via first computing the MCMD estimator is almost identical to that computed with no parameter estimation error.

2. Table 2 shows results that are very different. As pointed out by Durbin (1973), Henze (1996), and Khmaladze (2013), the KS test statistic with a plug in ML estimator has significant level distortions that persist even when the sample size is large. These distortions occur mainly because $\widetilde{T}_n$ has an asymptotic distribution that is affected by the ML estimator. Methods A and B therefore yield substantial level distortions in this case. These distortions are relieved by using the parametric bootstrap method C, which accommodates the parameter estimation error and has the same asymptotic null distribution as that of $\widetilde{T}_n$. Khmaladze's (2013) transformation method D removes the parameter estimation error from the test basis, and $\ddot{T}_n$ becomes distribution free.

3. Looking at Table 1 again, we note that there is a tendency for the empirical rejection rates to be closer to the nominal levels when c_k is small.

4. Comparing methods B and C in Table 1 we find that applying the asymptotic null distribution directly to the test yields more precise empirical rejection rates than applying the parametric bootstrap and Khmaladze's (2013) transformation. These results indicate that the $\widehat{T}_n$ test performs best under the null when it is constructed by data observations grouped into small intervals and compared with the asymptotic null distribution.

4.2 Testing under the Alternative

We now examine test power. For this purpose, we change the distribution of X_t from Pareto to the following exponential distribution as the generating mechanism:

$$\mathbb{P}(X_t \leq x | b \leq X_t \leq u) = \frac{1 - \exp(-\lambda_*(x-b))}{1 - \exp(-\lambda_*(u-b))}.$$

We denote this distribution $\text{Exp}(\lambda_*)$. We group the observations from $\text{Exp}(1.2)$ in the same way as in Section 4.1 and test the Pareto distributional assumption using methods B and C for $\widehat{T}_n$ and $\widetilde{T}_n$. In this case, the parameter θ_* is not defined under the alternative hypothesis, so that method A is infeasible.

The empirical rejection rates of $\widehat{T}_n$ and $(\widetilde{T}_n, \ddot{T}_n)$ are contained in Tables 3 and 4, respectively. The results can be summarized as follows.

1. First, $\widehat{T}_n$, $\widetilde{T}_n$, and $\ddot{T}_n$ are consistent. As the sample size increases, the rejection rates approach unity for methods B, C, and D.

2. The empirical rejection rates of $\widetilde{T}_n$ using method B are uniformly dominated by $\widehat{T}_n$ using methods B and C. This is mainly because the asymptotic critical values of $\widetilde{T}_n$ implemented by method B are too large, as evidenced in the substantial level distortions under the null seen in Table 2.

3. The overall power of $\widetilde{T}_n$ when the test is implemented by method C is similar to that of $\widehat{T}_n$ implemented by methods B or C and always dominate that of $\ddot{T}_n$ implemented by method D.

4. The empirical rejection rates of $\widehat{T}_n$ implemented by method B are close to those of method C. Even when the sample size is as small as 100, the empirical rejection rates are similar. So, the asymptotic null distribution based critical values yield performances similar to those based upon the parametric bootstrap.

5. When the sample size is small, the power of $\widetilde{T}_n$ implemented by method C is slightly higher than that of $\widehat{T}_n$ implemented by methods B or C, but the differences are very small.

4.3 Testing under the Local Alternative

To examine the local power of the test statistic we construct a mixed distribution of the null and alternative distributions using draws from both. Specifically, when $Z_t \sim \text{Exp}(1.2)$ and $W_t \sim \text{Pareto}(2.0)$, we let

$$X_t = \frac{5}{\sqrt{n}} Z_t + \left(1 - \frac{5}{\sqrt{n}}\right) W_t,$$

so that X_t is a mixture of Pareto and exponential random variables for which the mixture distribution of X_t converges to the Pareto distribution at an $n^{-1/2}$ convergence rate. For this generating mechanism, we test the Pareto distributional assumption using methods B, C, and D.

The simulation results of $\widehat{T}_n$ and $(\widetilde{T}_n, \ddot{T}_n)$ are contained in Tables 5 and 6, respectively. We summarize the results as follows.

1. As the sample size increases, the empirical rejection rates converge to levels that exceed nominal size except for the test $\widetilde{T}_n$ implemented by method B for which power is less than size. Hence, the test $\widehat{T}_n$ (resp. $\ddot{T}_n$) has nontrivial power under local alternatives when method B or C (resp. method D) is applied, but $\widetilde{T}_n$ has nontrivial powers only when method C is applied.

2. Local power of $\widetilde{T}_n$ is not given for method B in many cases because the critical values of $\widetilde{T}_n$ exceed the upper bound and test size is zero as noted in the discussion of Table 2.

3. Methods B and C have similar power patterns for the test $\widehat{T}_n$. Although method C yields more precise results than method B in the sense that the empirical rejection rates from the smaller sample sizes are closer to those for the large sample size, the differences are slight. We deduce from these results that the performance of methods B and C are similar under local alternatives.

4. The overall empirical rejection rates of $\widehat{T}_n$ are similar to those of $\widetilde{T}_n$ when that test is implemented by method C, implying that we can expect similar local power from $\widehat{T}_n$ and $\widetilde{T}_n$ when using parametric bootstrap methods. Furthermore, the local power of $\ddot{T}_n$ implemented by method D is uniformly dominated by that of $\widehat{T}_n$ implemented by method B.

5 Empirical Applications

We now proceed to apply these distributional tests in measuring top income shares. Estimating top income shares has been a longstanding topic of interest in the inequality literature since Kuznets (1953,1955), who calculated upper income shares for the US over the period 1913 to 1948. The widely used Gini coefficient is an alternative inequality measure but has been found to be insensitive to variations in upper income levels. In view of this limitation of the Gini coefficient, upper $x\%$ income shares have become commonly used as an additional, easily interpreted measure of income inequality.

The conventional approach to measuring upper income levels is to continuously interpolate the top $x\%$ income levels by relying on estimates from a Pareto distribution. Most income data are available in a group frequency format, making interpolation necessary for implementing this approach.

In spite of its popularity, the Pareto distribution for income data is restrictive and may be a misleading representation for top incomes in some cases. Feenberg and Poterba (1993) test the validity of the top income share estimates obtained by the Pareto interpolation method with those obtained by using micro-data. For the top 0.50% US income data from 1979 to 1989, they found that the results from these two different methodologies yielded almost identical results. This outcome is suggestive, indicating that the

Pareto distribution condition may be a reasonable assumption for these US data. On the other hand, Atkinson (2005) introduced a nonparametric method called the mean-split histogram method that estimates the top income shares under certain underlying conditions on the income distributions. Thus, both parametric and nonparametric methods have been used in past work on inequality measurement, and empirical tests have been used to assess the adequacy of the parametric assumptions in upper income share estimation.

With the same motivation as Feenberg and Poterba (1993), we apply our KS test statistic to Korean income tax return data from 2007 to 2012. Our empirical goal is to calculate estimates of upper income shares for Korea using our new methodology and compare findings with those available in the prior literature.

5.1 Korean Income Data from 2007 to 2012

Top income shares are estimated by comparing income tax return data of Korea with population data. The source and nature of the data are briefly discussed in what follows in this subsection. More detailed explanations on data constructions are given in the Appendix.

The Statistical Yearbook of National Tax published by the National Tax Service (NTS) contains annual Korean income tax statistics for each year, and the data therein were used for measuring the top income shares by Kim and Kim (2014). The number of income groups in *The Statistical Yearbook of National Tax* differ from year to year, and there are at most around 10 income groups. Although the NTS provides income tabulations for a long period[2], tests of the Pareto distributional assumption are better suited to the methodology when the group size is much bigger.

We, therefore, use another set of income tax return data that are also provided by the NTS for the years from 2007 to 2012. These data have a different format from those in *The Statistical Yearbook of National Tax*. Table 7 provides summary statistics of the income tax return data used herein. Several features stand out. The most noticeable feature of the data for our purposes is group size. For example, our 2010 data have 3988 groups, whereas the conventional data in *the Statistical Yearbook of National Tax* have only 10 groups for the same year. This large group size is obtained by making the group interval much smaller than those in the conventional income data. The first and the last group intervals for the year 2010 are $(0.0, \text{KRW50 mil.}]$ and $(\text{KRW39}, 910\text{ mil.}, \infty]$. For the other groups, the data are provided in the same format with each group interval width being KRW10 mil. For example, the second smallest income group is $(\text{KRW 50 mil.}, \text{KRW 60 mil.}]$. By contrast the conventional income tax data have irregular group patterns. Out of the 10 income groups, the first and last groups are $(0.0, \text{KRW21 mil.}]$ and $(\text{KRW536 mil.}, \infty]$, respec-

[2] Kim and Kim (2014) measure top income shares only from 1979 using the information in *the Statistical Yearbook of National Tax*.

tively. For the other 8 groups, the smallest income group interval has width KRW12 mil., and the largest group interval width is KRW211 mil. A second important feature of the data is that there is no double counting from the same income source, a phenomenon that arises with some data, such as the Japanese data examined by Moriguchi and Saez (2010). A third feature of interest is the time period covered by our data. The time span includes the global financial crisis, which opens up the possibility of studying the impact of the global financial crisis on the distribution of income in Korea with these data.

We also obtain total income for each year to compute upper $x\%$ income shares. For this calculation, we follow the approach in Pikety and Saez (2003) and Moriguchi and Saez (2010), where total income is derived from the national accounts for personal income by adjusting non-taxable income. This adjustment is a commonly used process in the literature for obtaining total income, as detailed in the Appendix.

Finally, we obtain population data in Korea. Various population data have been used in the prior literature. For example, Picketty and Saez (2003) and Atkinson (2005) employ US family data and UK individual unit data, respectively, accordingly to the country tax units available. For Korea, the tax unit is the individual unit, and a significant number of men serve mandatory military service in their 20s. So we calculate population in terms of the working-age population of age 20 and above by excluding conscripted personnel such as soldiers and call this measure *employment*. In addition to this definition of population, we construct another measure to assist in making comparisons of top income shares with other studies. This measure includes the working-age population aged 15 and over, so that conscripted individuals are included in the population, and we call this measure the *labor force*. These two populations measures for Korea aged 15 and above and aged 20 and above correspond with population measures used in studies of other countries such as the UK and Japan in Atkinson (2005) and Moriguchi and Saez (2008). The population data are reported in Table 7.

5.2 Empirical Analysis

Using the income tax return data described above, we estimate the top income shares in Korea from 2007 to 2012. The specific procedures are as follows:

1. We first identify the income group for the top $x\%$ income level to ensure inclusion. The size of top $x\%$ income population is computed using the population data, and we let $(x_{\sharp-1}, x_{\sharp}]$ denote this group. Note that $x_{\sharp} - x_{\sharp-1}$ is KRW 10 million for our data sets.

2. We test the Pareto distributional assumption for the grouped data. We choose b and u so that $b \leq x_{\sharp-1}$ and $u \geq x_{\sharp}$ and estimate θ_* by the MCMD estimator to test the Pareto distributional hypothesis. The asymptotic critical values are estimated and applied. Readers are referred to our discussion below on

how b and u are determined.

3. We estimate the top $x\%$ income level and denote this level $\widehat{x}_n$. This procedure involves first estimating the preliminary top $x\%$ income level by choosing it as $x_\dagger := F^{-1}(q; \widehat{\theta}_n)$, where

$$q := \frac{\text{top } x\% \text{ income population size} - \text{population size with incomes greater than } u}{\text{population size with incomes} \in (b, u]}.$$

   If $x_\dagger \in (x_{\sharp-1}, x_\sharp]$, we let $\widehat{x}_n$ be $x_\dagger$; if $x_\dagger > x_\sharp$, let $\widehat{x}_n$ be the upper bound $x_\sharp$ of the interval; otherwise, let $\widehat{x}_n$ be $x_{\sharp-1}$. This additional restriction is imposed because $\widehat{x}_n$ must lie between $x_{\sharp-1}$ and $x_\sharp$ by virtue of the first-step requirement.

4. We finally compute the top $x\%$ share of incomes. We first estimate the total income greater than $\widehat{x}_n$ by

$$\widehat{m}_n := \left(\frac{F(x_\sharp; \widehat{\theta}_n) - F(\widehat{x}_n; \widehat{\theta}_n)}{F(x_\sharp; \widehat{\theta}_n) - F(x_{\sharp-1}; \widehat{\theta}_n)} \right) \times I_\sharp + \sum_{i=\sharp+1}^{k} I_i,$$

   where I_i denotes the total income in the group of $(x_{i-1}, x_i]$, and k is the group size as before. The top x-% share of income is computed by dividing $\widehat{m}_n$ with total income from the national account.

Several remarks on this process are in order. First, the Pareto distribution condition is tested in Step 2. Even if the null hypothesis is rejected, we proceed to Step 3 by assuming that the Pareto distribution is a good approximation to the top income distribution and then examine how the Pareto assumption affects the estimation of the top income shares. Below we compare the top $x\%$ income shares estimated by the Pareto interpolation method with those obtained by the mean-split histogram method.[3] Second, when implementing Step 2, the bottom and top border values (b and u) have to be selected in such a way that the interval $(x_{\sharp-1}, x_\sharp]$ is a subgroup of the grouped data. In principle, this selection may affect inference - that is, when the initial bottom and top border values are modified, test results from using $\widehat{T}_n$ may also be modified. However, for our data, if the top $x\%$ income level is high enough, the test results turn out to be insensitive to the selection of b and u.

The top $x\%$ income levels are estimated and contained in Tables 8 and 9. Table 8 contains the findings from 2007 to 2009, and Table 8 contains results from 2010 to 2012. We summarize the key properties of our estimates as follows.

[3] Atkinson's (2005) mean-split histogram method estimates top income shares by a piecewise linear interpolation method that is constructed by upper and lower bounds for income density function under the assumption that income density is not an increasing function around the region of interest. Atkinson, Piketty, and Saez (2011) survey that top income shares are estimated by this method for many countries such as Australia, Finland, Netherlands, New Zealand, Norway, Singapore, and UK.

1. When the top 1.0% income level is estimated, the Pareto assumption does not hold for every year from 2007 to 2012. For example, for 2007, the p-value of $\widehat{T}_n$ is zero regardless of the population data. As mentioned above, the value of $\widehat{T}_n$ is dependent on the selection of $(b, u]$. In fact, we tried many selections of $(b, u]$ and had to reject the null hypothesis for every selection. The reported interval is one of these trials. This test outcome shows that the Pareto distribution assumption is hard to accept as holding for the 1.0% and higher incomes.

2. Although the results are not reported in the tables, even for the top 0.5% of incomes, the Pareto distribution assumption does not hold for every year in the sample data.

3. When the top 0.10%, 0.05%, or 0.01% and higher incomes are estimated, we could not reject the Pareto distribution assumption. More precisely, for every year, we could find intervals $(b, u]$ such that the null hypothesis cannot be rejected. Finding such an interval was not difficult. When an interval was arbitrarily selected, the Pareto hypothesis could not be rejected at the first stage for most cases. If the null hypothesis was rejected at the first trial, we searched for bottom and top values of the interval (b and u) until the Pareto hypothesis could not be rejected. Sequential testing in this way is justified asymptotically, thereby avoiding the data snooping problem that arises when hypotheses are tested iteratively. These findings imply that for the Pareto distribution assumption to hold, at least the top 0.10% and higher incomes need to be considered.

4. The estimated x% top income levels ($\widehat{x}_n$) are between $(x_{\sharp-1}, x_\sharp]$ for most cases. Sometimes, the preliminary estimates of the top income levels ($x_\dagger$) are greater than the presumed border value $x_\sharp$. For such cases, we let $\widehat{x}_n$ be $x_\sharp$ as required in Step 3. For example, when the top 1.00% income level is estimated using 2007 data and the population involves ages 20 or older, $\widehat{x}_n = 0.9000$ for $x_{\sharp-1} = 0.8900$ and $x_\sharp = 0.9000$. We added the superscript '$\sharp$' to the figures to indicate such an occurence. The preliminary estimates of the top income levels ($x_\dagger$) are not substantially different from the boundary values ($x_\sharp$) for every case, and this happens because the data set has a narrow interval width of KRW 10 million.

Using the estimated top income levels, we next implement Step 4 and estimate the top x% income shares. For each population data set and each year, we compute the shares and provide the estimates in Table 10. We summarize the findings given in these Tables as follows.

1. Table 10 compares our top income shares with those obtained by Atkinson's (2005) mean-split histogram method which does not impose a particular parametric distributional assumption for income.

Under the condition that the population density function is not an increasing function of income around the region of interest, the mean-split histogram method estimates top income shares by estimating tight lower and upper bounds for the income density function. Instead of obtaining top income shares by linearly interpolating histogram points, the method interpolates points by using a piecewise linear function that is derived by using lower and upper bounds for the income distribution under the condition that the income density is a non-increasing function around the region to which the top x% income level belongs. Figures in round parentheses in Table 10 are the income shares estimated by the mean-split histogram method that were computed by Park and Jeon (2014). The estimated top income shares from this method are generally very close to our own estimates, but show greater differences at the 1.00% top income level, the level for which the Pareto hypothesis is rejected and non parametric estimates may be preferred.

2. We also compare our findings with those of Kim and Kim (2014; KK) who estimated the top income shares using the income tax table from 1933 to 2010. These authors used population data for adults aged 20 or older and income data from *the Statistical Yearbook of National Tax.* Both data sets differ from those used here and have certain limitations, as discussed earlier. In spite of these differences, the KK estimates are similar to our own, with the greatest difference being 0.69%, which occurs for the top 1.00% income shares in year 2010. For higher income shares, the differences are small. We therefore conclude that our findings concerning upper income shares in Korea corroborate those obtained by KK over the period 2007 to 2010.

3. The top income shares have a general tendency to rise over time. In year 2009 the income shares went down, most probably due to the global financial crisis, but began to rise again and maintain a rising tendency thereafter, concomitant with the slow recovery in the global economy from the financial crisis. These results indicate that the top income shares can usefully supplement the Gini coefficient, because income inequality as measured by the Gini coefficient has declined since 2009 according to official Korean statistics. The results also match earlier findings in the literature. For instance, Piketty and Saez (2003), Atkinson (2005), Piketty (2003), Atkinson and Leigh (2007, 2008) Moriguchi and Saez (2010), and Kim and Kim (2014), among others, observe that the top income shares of the US, UK, and France, Australia, New Zealand, Japan, and Korea all increased over time between 2000 to 2010, respectively.

4. Despite the general rising tendency of the top income shares over 2007 to 2012, the patterns are not monotonic and have a noticeable blip around 2010 and 2011. We note that jumps are observed from

top $x\%$ income levels over 2010 and 2011. For example, the growth rates of the top 1% income levels in 2010 and 2011 are about 11.79% and 10.34%, whereas the growth rates of 2008, 2009, and 2012 are 2.71, 2.39, and 3.44%, respectively. On the other hand, total income derived from the national accounts does not exhibit such big jumps in 2010 and 2011, although it does jump to 8.25% in 2012 from 5.94%, which partly explains the noticeable blips in income shares in 2010 and 2011. In terms of international comparisons, the top income shares of most other countries do not show definitely rising tendencies since 2007, based on available estimates,[4] although they do show an overall increasing pattern from 2000. On the other hand, some countries such as Germany, US, and Korea maintain a rising tendency over the same period. The upper income earners of these countries have apparently overcome the effects of the global financial crisis more rapidly than other countries that manifest declining top income shares.

6 Conclusion

Issues of income inequality now attract considerable attention at both national and international levels. Of growing interest in the assessment of income inequality is the share of upper incomes within the income distribution and whether and by how much such shares may be growing over time. Analysis of such issues requires quantification of suitable inequality measures and is frequently conducted empirically using explicit distributional assumptions, such as the Pareto, to characterize upper tail shape, as in the research of Piketty and Saez (2003). The tests given in the present work enable applied researchers to evaluate the adequacy of such distributional assumptions in practical empirical studies where, as is most frequently the case, unknown parameters need to be estimated. Our test criteria integrate the Kolmogorov and Smirnov (KS) test criteria with a minimum distance parameter estimation procedure that leads to a convenient limit theory for the test statistic under the null. The test is easily implemented and is shown to perform well under both null and local alternative hypotheses.

Our application of this KS test statistic to Korean income data over 2007 to 2012 shows that the Pareto distribution is supported only for very high income levels. The Pareto tail shape is rejected for the top 1.0% or 0.5% and higher incomes for every year in the data; but for tail observations lying in the top 0.10%, 0.05%, or 0.01% and higher incomes, the Pareto shape is much harder to reject. These empirical findings suggest the use of care in applying Pareto interpolation techniques for measuring top 0.50% or lower income shares. Our results also generally support the observation that upper income shares have been increasing

[4] The world top income database reports the top income shares of 27 countries that are reported in the literature. For example, countries such as Canada, Netherlands, and UK show declining patterns, and this is believed to be due to the global financial crisis.

over time in Korea, in line with more global observations on income shares.

7 Appendix: Proofs and Data Description

7.1 Proofs

Proof of Theorem 1: For each i we have $\widehat{p}_n(x_i) \overset{\text{a.s.}}{\to} p(x_i)$, so that the uniform strong law $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} | \sum_{i=1}^{k}\{F(x_i,\boldsymbol{\theta}) - \widehat{p}(x_i)\}^2 - \sum_{i=1}^{k}\{F(x_i,\boldsymbol{\theta}) - p(x_i)\}^2| \overset{\text{a.s.}}{\to} 0$ holds, which implies that $\widehat{\boldsymbol{\theta}}_n \overset{\text{a.s.}}{\to} \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{k}\{F(x_i,\boldsymbol{\theta}) - p(x_i)\}^2$. By Assumption A(*iii*). $\boldsymbol{\theta}_o$ is unique in $\boldsymbol{\Theta}$. Therefore, $\widehat{\boldsymbol{\theta}}_n \overset{\text{a.s.}}{\to} \boldsymbol{\theta}_o$ as desired. ■

Proof of Theorem 2: By Taylor expansion: $F(x_i,\boldsymbol{\theta}) = F(x_i,\boldsymbol{\theta}_o)+\nabla'_{\boldsymbol{\theta}}F(x_i,\boldsymbol{\theta}_o)(\boldsymbol{\theta}-\boldsymbol{\theta}_o)+O((\boldsymbol{\theta}-\boldsymbol{\theta}_o)^2)$, so that $Q_n(\boldsymbol{\theta}) = \sum_{i=1}^{k}\{Y_i-\mathbf{z}'_i(\boldsymbol{\theta}-\boldsymbol{\theta}_o)[1+O(\boldsymbol{\theta}-\boldsymbol{\theta}_o)]\}^2$. Therefore, $(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_o)[1+O(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_o)] = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{Y}$ using least squares, which in turn implies

$$(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o)[1 + O_{\mathbb{P}}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o)] - (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y} = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'(\mathbf{Y} - \mathbf{y}).$$

Note that $\sqrt{n}(\mathbf{Y} - \mathbf{y}) \Rightarrow \mathbf{B}^o$ and $\mathbf{z}'\mathbf{y} = \mathbf{0}$ from the first-order condition for $\boldsymbol{\theta}_o$, implying that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) = \sqrt{n}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'(\mathbf{Y} - \mathbf{y}) + O_{\mathbb{P}}(n^{-1/2}) \Rightarrow (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{B}^o, \tag{22}$$

completing the proof. ■

Proof of Theorem 3: (*i*) Since, for each $i = 1,\dots,k$, $F(x_i,\widehat{\boldsymbol{\theta}}_n) = F(x_i,\boldsymbol{\theta}_*) + \nabla'_{\boldsymbol{\theta}}F(x_i,\boldsymbol{\theta}_*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) + O_{\mathbb{P}}(n^{-1})$, it follows that

$$\begin{aligned}\sqrt{n}(\widehat{p}_n(x_i) - F(x_i,\widehat{\boldsymbol{\theta}}_n)) &= \sqrt{n}(\widehat{p}_n(x_i) - F(x_i,\boldsymbol{\theta}_*)) - \nabla'_{\boldsymbol{\theta}}F(x_i,\boldsymbol{\theta}_*)\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) + O_{\mathbb{P}}(n^{-1/2})\\ &= \sqrt{n}(\widehat{p}_n(x_i) - F(x_i,\boldsymbol{\theta}_*)) - \mathbf{z}'_i(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\sqrt{n}(\mathbf{Y} - \mathbf{y}) + O_{\mathbb{P}}(n^{-1/2})\\ &\Rightarrow \mathcal{B}^o(p(x_i)) - \mathbf{z}'_i(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{B}^o,\end{aligned}$$

where the last equality and the weak convergence follows from (22). The result holds for every i and jointly, so it follows that $\widehat{\mathbf{S}}_n \Rightarrow \mathbf{B}^o - \mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{B}^o = \mathbf{m}\mathbf{B}^o = \mathbf{G}$. Therefore, continuous mapping delivers the limit result for the test statistic $\widehat{T}_n \Rightarrow \mathcal{Z}_k := \max[|G_1|,\dots,|G_k|]$.

(*ii*) Since for each $i = 1,\dots,k$, $F(x_i,\widehat{\boldsymbol{\theta}}_n) = F(x_i,\boldsymbol{\theta}_o) + \nabla'_{\boldsymbol{\theta}}F(x_i,\boldsymbol{\theta}_o)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) + O_{\mathbb{P}}(n^{-1})$, it follows that

$$\widehat{p}_n(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n) - (p(x_i) - F(x_i, \boldsymbol{\theta}_o)) = \widehat{p}_n(x_i) - p(x_i) - \nabla'_{\boldsymbol{\theta}} F(x_i, \boldsymbol{\theta}_o)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) + O_{\mathbb{P}}(n^{-1}). \quad (23)$$

Given the definition of $y_i := p(x_i) - F(x_i, \boldsymbol{\theta}_o)$, it follows that

$$\begin{aligned}\sqrt{n}\{\widehat{p}_n(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n) - y_i\} &= \sqrt{n}(\widehat{p}_n(x_i) - p(x_i)) - \nabla'_{\boldsymbol{\theta}} F(x_i, \boldsymbol{\theta}_o)\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) + O_{\mathbb{P}}(n^{-1/2}) \\ &= \sqrt{n}(\widehat{p}_n(x_i) - p(x_i)) - \mathbf{z}'_i(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\sqrt{n}(\mathbf{Y} - \mathbf{y}) + O_{\mathbb{P}}(n^{-1/2}) \\ &\Rightarrow \mathcal{B}^o(p(x_i)) - \mathbf{z}'_i(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{B}^o,\end{aligned}$$

where the last equality and the weak convergence hold by (22). The result holds for all i and jointly, so that $\sqrt{n}(\widehat{\mathbf{S}}_n - \mathbf{y}) \Rightarrow \mathbf{G}$, and the desired result follows.

(*iii*) As in the proof of Theorem 1, we note that, for each i, $F(x_i, \widehat{\boldsymbol{\theta}}_n) = F(x_i, \boldsymbol{\theta}_o) + \mathbf{z}'_i(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) + o_{\mathbb{P}}(n^{-1/2})$, implying that

$$\begin{aligned}\widehat{p}_n(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n) &= \widehat{p}_n(x_i) - p(x_i) + p(x_i) - F(x_i, \boldsymbol{\theta}_o) - \mathbf{z}'_i(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) + o_{\mathbb{P}}(n^{-1/2}) \\ &= \widehat{p}_n(x_i) - p(x_i) + y_i - \mathbf{z}'_i(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y} - \mathbf{z}'_i(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'(\mathbf{Y} - \mathbf{y}) + o_{\mathbb{P}}(n^{-1/2}) \\ &= \widehat{p}_n(x_i) - p(x_i) + n^{-1/2}\xi_i - \mathbf{z}'_i(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'(\mathbf{Y} - \mathbf{y}) + o_{\mathbb{P}}(n^{-1/2}),\end{aligned}$$

where the second equality holds by the definition of $y_i := p(x_i) - F(x_i, \boldsymbol{\theta}_o)$ and the fact that $(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y} + (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'(\mathbf{Y} - \mathbf{y}) + O_{\mathbb{P}}(n^{-1})$, and the third equality holds by virture of the local alternative $\mathcal{H}_\ell$ and the definition of $\xi_i := y_i - \mathbf{z}'_i(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}$. Therefore, $\sqrt{n}\{\widehat{p}_n(x_i) - F(x_i, \widehat{\boldsymbol{\theta}}_n)\} \Rightarrow \xi_i + G_i$, which holds for every i and jointly. Therefore, $\widehat{\mathbf{S}}_n \Rightarrow \mathbf{m}(\mathbf{h} + \mathbf{B}^o) \sim N(\mathbf{m}\mathbf{h}, \mathbf{m}\boldsymbol{\Sigma}^o\mathbf{m})$, so that the test statistic $\widehat{T}_n \Rightarrow \mathcal{Z}^a_k := \max[|\xi_1 + G_1|, \ldots, |\xi_k + G_k|]$, as desired. ■

Proof of Lemma 1: (*i*) We note that $\partial_i F_k(\cdot, \boldsymbol{\theta}_o)\partial_j$ and $F(\cdot, \boldsymbol{\theta}_o)$ are piecewise continuous on $[x_0, x_k]$. Therefore, for each i and j, $\int_0^1 \partial_i \bar{F}_k(x)\partial_j \bar{F}_k(x)dx$ is well defined. Similarly, for each i and j, $\partial_i \bar{F}(\cdot)$ and $\partial_j \bar{F}(\cdot)$ are continuous on $[0, 1]$ by Assumption B(*i*), implying that $\partial_i \bar{F}(\cdot)\partial_j \bar{F}(\cdot)$ is also continuous on $[0, 1]$ and that $\int_0^1 \partial_i \bar{F}(x)\partial_j \bar{F}(x)dx$ is also well defined. Furthermore, $\partial_i \bar{F}_k(\cdot)\partial_j \bar{F}_k(\cdot)$ converges uniformly to $\partial_i \bar{F}(\cdot)\partial_j \bar{F}(\cdot)$, and $|\partial_i \bar{F}_k(\cdot)\partial_j \bar{F}_k(\cdot)|$ is uniformly bounded by $\sup_{x\in[0,1]} |\partial_i \bar{F}(x)\partial_j \bar{F}(x)|$ by Theorem B(*ii*). Thus, dominated convergence implies that

$$\int_0^1 \partial_i \bar{F}_k(x)\partial_j \bar{F}_k(x)dx \xrightarrow{k} \int_0^1 \partial_i \bar{F}(x)\partial_j \bar{F}(x)dx.$$

As this limit holds for every combination of i and j, the desired result follows.

(*ii*) By definition $\bar{\mathcal{B}}^o_k(\cdot)$ is piecewise continuous on $[0,1]$, so that for every j, $\int_0^1 \partial_j \bar{F}_k(x)\bar{\mathcal{B}}^o_k(x)dx$ is well defined as before. Furthermore, $\bar{\mathcal{B}}^o(\cdot)$ is continuous with probability 1, implying that $\partial_j \bar{F}(\cdot)\bar{\mathcal{B}}^o(\cdot)$ is also continuous on $[0,1]$ with probability 1. Therefore, for each j, $\int_0^1 \partial_j \bar{F}(x)\bar{\mathcal{B}}^o(x)dx$ is also well defined with probability 1. Furthermore, note that $|\partial_j \bar{F}_k(\cdot)\bar{\mathcal{B}}^o_k(\cdot)|$ is uniformly bounded by $\sup_{x\in[0,1]} |\partial_j \bar{F}(x)\bar{\mathcal{B}}^o(x)|$ with probability 1 by Assumption B(*ii*). Therefore, dominated convergence implies that for each j,

$$\int_0^1 \partial_j \bar{F}_k(x)\bar{\mathcal{B}}^o_k(x)dx \overset{k}{\to} \int_0^1 \partial_j \bar{F}(x)\bar{\mathcal{B}}^o(x)dx,$$

with probability 1. As this result holds for every j and jointly, $\mathbf{Z}_k \overset{k}{\Rightarrow} \mathbf{Z}$. We also note that Assumptions B(*iv* and *v*) imply that $\mathbf{A}_o^{-1}$ exists and $\mathbb{E}[\int_0^1 \int_0^1 \nabla_{\boldsymbol{\theta}} \bar{F}(x)\nabla'_{\boldsymbol{\theta}} \bar{F}(x')\bar{\mathcal{B}}^o(x)\bar{\mathcal{B}}^o(x')dxdx'] = \mathbf{B}_o$ by Fubini and Tonelli, respectively. This completes the proof. ■

Proof of Theorem 4: The desired result simply follows by combining results in Lemma 1(*i* and *ii*) using joint convergence. ■

Proof of Theorem 5: (*i*) Weak convergence as $n \to \infty$ is already provided in the proof of Theorem 3(*i*) and so we need only show weak convergence with respect to k. Note that $\bar{\mathcal{B}}^o_k(\cdot) \overset{k}{\to} \bar{\mathcal{B}}^o(\cdot)$ uniformly on $[0,1]$ with probability 1, and $\nabla_{\boldsymbol{\theta}} \bar{F}_k(\cdot) \overset{k}{\to} \nabla_{\boldsymbol{\theta}} \bar{F}(\cdot)$ by Assumption B(*ii*). Finally, $\mathbf{A}_k^{-1}\mathbf{Z}_k \overset{k}{\Rightarrow} \mathbf{A}_o^{-1}\mathbf{Z}$ by Lemma 1. It follows that $\bar{\mathcal{B}}^o_k(\cdot) - \nabla'_{\boldsymbol{\theta}} \bar{F}_k(\cdot)\mathbf{A}_k^{-1}\mathbf{Z}_k \overset{k}{\Rightarrow} \bar{\mathcal{B}}^o(\cdot) - \nabla'_{\boldsymbol{\theta}} \bar{F}(\cdot)\mathbf{A}_o^{-1}\mathbf{Z}$ by joint convergence. Now simply apply the definitions of $\bar{\mathcal{G}}^o_k(\cdot)$ and $\bar{\mathcal{G}}^o(\cdot)$ to the left and right side of this large group size weak convergence result.

(*ii*) Weak convergence as $n \to \infty$ is again provided in the proof of Theorem 3(*ii*). The proof of Theorem 5(*i*) shows that $\bar{\mathcal{G}}^o_k(\cdot) \overset{k}{\Rightarrow} \bar{\mathcal{G}}^o(\cdot)$, and $\bar{\xi}_k(\cdot) = \bar{h}_k(\cdot) \overset{k}{\to} \bar{h}(\cdot)$ uniformly on $[0,1]$ by virtue of the structure of $\bar{h}_k(\cdot)$. The desired result follows directly. ■

7.2 Data Description: Korean Income Data from 2007 to 2012

We provide more detailed source and nature of the data in this subsection.

7.2.1 Grouped Income Data

The income tax return data are formed from several different income sources. The following items are included in the data: business income, interests and dividends (above KRW40 million), pension benefits, wage income, and other income. All these are taxable income sources. Interests and dividends (below KRW40 million), retirement income, and capital gains are not included in our data. As our main focus of here is in the distribution of high income groups, the exclusion of interest and dividend income below KRW40 million is unlikely to affect inferences. High income groups with more than KRW40 million in interest and dividend income are included in our data. Consistent with other countries, the high income groups in our data are mostly determined by wage income, business income, and interest and dividend income.

On the other hand, our data do not include all types of non-taxable income. Most non-taxable income data are not easy to verify and some income is not voluntarily reported especially that relating to financial income for high income groups. This aspect of income data produces multiple sources of measurement error and difficulties in correcting income data to include all types of non-taxable incomes. Therefore, we delimit attention to taxable income, as is commonly done for other countries, in order to maintain a consistent income definition throughout the time period studied and to minimize the effects of measurement error in the data.

7.2.2 Total Income Calculation

The way total income is derived from the national accounts for personal income differs depending on which of the two income tax systems is in use: negative and positive income tax systems. The negative income tax system includes almost all types of income earned by personnel and uses these to construct income tax statistics. The positive income tax system additionally includes non-taxable personal income, so that the income total corresponds to household income data information retrieved from the national accounts.

The Korean income tax system is built upon the positive income tax system, so that we estimate total income using personal income information in the national accounts. There are three types of personal income in the accounts that need to be adjusted: imputed employers' social contributions that are the part of the compensation for employees, imputed rents on taxpayers' owned houses that are part of operating surpluses, and indirectly measured financial intermediary services (IMFIS) that are part of property incomes. These three sources of income are not included in tax base although they are attributed to households in the national accounts.

For the exclusion, we follow these procedures. First, employers' social contributions are simply subtracted from employee compensation. The national accounts separately report compensation under three headings: wages, benefits, and social contributions made by employers. The last item is again divided into actual and imputed contributions. The final item is excluded. Second, imputed rents on taxpayers' owner occupied houses are estimated in two steps: we compute the ratio of houses owned by taxpayers using the yearly national census data and multiply the ratio to housing service operating surpluses that are given by the input-output tables each year. The amount of imputed rents on taxpayers' owner occupied houses is estimated by this product. Finally, instead of estimating the amount of IMFIS, we use the information in *the Statistical Yearbook of National Tax* each year that reports household interest and dividend income earned from financial institutions. Using this information for our household property income, there is no need to estimate the IMFIS.

7.2.3 Population Calculation

Some further remarks on the Korean population data are in order. First, some of the data for those aged 15 or 20 and above involve projections. Statistics Korea conduct a national census every five years which is used to project population over the next five-year period and to correct the prior five-year projections. Currently, the Korean census population data aged 15 or 20 and above for the years 2011 and 2012 are not yet available. We therefore use data projections for these years. Second, the employment and labor force data are based on population bases that are collected monthly by Statistics Korea. The population bases are constructed to include individuals who are capable of working and who are not soldiers, individuals who are required to work in social services (including the police force), and individuals who are incarcerated and serving fixed sentences. Statistics Korea announces the population bases every month and provides detailed statistics segregated by gender, age, and other characteristics. The labor force and employment are estimated by adding to these population bases as required by the definitions of these populations.

References

ALVAREDO, F., ATKINSON, T., PIKETTY, T., AND SAEZ, E. (2015): *The World Top Incomes Databse*, (accessed April 15, 2015).

ATKINSON, A. (2005): "Top Incomes in the UK over the 20th Centry," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 168, 325–343.

ATKINSON, A. (2007): "Measuring Top Incomes: Methodological Issues." In: Atkinson, Anthony B. and Piketty, Thomas (Eds.), *Top Incomes over the Twentieth Century : a Contrast between Continental European and English-Speaking Countries*. Oxford, UK: , Oxford University Press.

ATKINSON, A. AND LEIGH, A. (2007): "The Distribution of Top Incomes in Australia," *Economic Record*, 83, 247–261.

ATKINSON, A. AND LEIGH, A. (2008): "Top Incomes in New zealand 1921–2005: Understanding the Effects of Marginal Tax Rates, Migration Threat, and the Macroeconomy," *Review of Income and Wealth*, 54, 149–165.

ATKINSON, A., PIKETTY, T., AND SAEZ, E. (2011): "Top Incomes in the Long Run of History," *Journal of Economic Literature*, 49, 3–71.

BANK OF KOREA (2007–2012): *Economic Statistics System*. (accessed August 1, 2014).

BOLTHAUSEN, E. (1997): "Convergence in Distribution of Minimum-Distance Estimators ," *Metrika*, 24, 215–227.

CHOULAKIAN, V., LOCKHART, A., AND STEPHENS, M. (1994): "Cramér-von Mises Statistics for Discrete Distributions," *Canadian Journal of Staistics*, 22, 125–137.

DIEBOLD, F., T. A. GUNTHER AND A. S. TAY, (1998): Evaluating Density Forecasts with Application to Financial Risk Management," International Economic Review, 39, 863-883.

DONALD, P., STEWARR, W., AND KOPECKY, K. (1979): "Distribution-Free Regression Analysis of Grouped Survival Data," *Biometric*, 35, 785–793.

DONSKER, M. (1951): "An Invariance Principle for Certain Probability Limit Theorems," *Memoirs of the American Mathematical Society*, 6, 1–10

DURBIN, (1973): "Weak Convergence of the Sample Distribution Function when Parameters are Estimated," *Annals of Statistics*, 1, 279–290.

FEENBERG, D. AND POTERBA, J. (1993): "Income Inequality and the Incomes of Very High-Income Taxpayers: Evidence from Tax Returns." In: Poterba, J. (Ed.), *Tax Policy and the Economy*. Vol. 7. Cambridge, MA: MIT Press.

FEENBERG, D. AND POTERBA, J. (2000): "The Income and Tax Share of Very High-Income Households, 1960–1995" *American Economic Review, Papers and Proceedings*, 90, 264–270.

HENZE, N. (1996): "Empirical-Distribution-Function Goodness-of-Fit Tests for Discrete Models," *Canadian Journal of Statistics*, 39, 2795–3443.

KHMALADZE, E. (1981): "Martingale Approach in the Theory of Goodness-of-fit Tests," *Theoretical Probability and Its Applications*, 26, 240–257.

KHMALADZE, E. (1993): "Goodness of Fit Problems and Scanning Innovation Martingales," *Annals of Statistics*, 21, 798–829.

KHMALADZE, E. (2013): "Note on Distribution Free Testing for Discrete Distributions," *Annals of Statistics*, 41, 2979–2993.

KIM, N. AND KIM, J. (2014): "Income Inequality in Korea, 1933-2010: Evidence from Income Tax Statistics," Working Paper, Naksungdae Institute of Economic Research.

KLAR, B. (1999): "Goodness-of-Fit Tests for Discrete Models Based on the Integrated Distribution Function," *Metrika*, 49, 53–69.

KUZNETS, S. (1953): *Shares of Upper Income Groups in Income and Savings*. New York: National Bureau of Economic Research.

KUZNETS, S. (1953): "Economic Growth and Economic Inequality," *American Economic Review*, 45, 1–28.

LEE, S. (2014): "Goodness of Fit Test for Discrete Random Variables," *Computational Statistics and Data Analysis*, 69, 92–100.

MORIGUCHI, C. AND SAEZ, E. (2008): "The Evolution of Income Concentration in Japan, 1886-2005: Evidence from Income Tax Statistics," *Review of Economics and Statistics*, 90, 713–734.

MORIGUCHI, C. AND SAEZ, E. (2010): "The Evolution of Income Concentration in Japan, 1886-2005: Evidence from Income Tax Statistics." In: Atkinson, Anthony B. and Piketty, Thomas (Eds.), *Top Incomes: A Global Perspective*. Oxford, UK: Oxford University Press. Series updated by Facundo Alvaredo, Chiaki Moriguchi and Emmanuel Saez (2012, Methodological Notes)

NATIONAL TAX SERVICE (2007–2012): *Annals of National Tax Statistics*. Seoul: Korea Government.

THE STATISTICS OF KOREA (2007–2012): "Future Population Projections," *National Statistics Portal (KOSIS)*, (accessed September 29, 2014).

THE STATISTICS OF KOREA (2005, 2010): "Surveys on Economically Active Population," *National Statistics Portal (KOSIS)*, (accessed September 29, 2014).

PIKETTY, T. (2001): "Income Inequality in France, 1901–1998." Discussion Paper no. 2876. London: Centre for Economic Policy Research.

PIKETTY, T. (2003): "Income Inequality in France, 1901–1998," *Journal of Political Economy*, 111, 1004–1042.

PIKETTY, T. AND SAEZ, E. (2003): "Income Inequality in the United States, 1913–1998," *Quarterly Journal of Economics*, 118, 1–39.

PARK, M. AND JEON, B. (2014): "Changes in Income Allocations and Policy Suggestions," Research Report, Korea Institute of Public Finance (in Korean).

PETTITT, A. AND STEPHENS, M. (1977): "The Kolmogorov-Smirnov Goodness-of-Fit Statistic with Discrete and Grouped Data," *Technometrics*, 19, 205–210.

POLLARD, D. (1980): "The Minimum Distance Method of Testing," *Metrika*, 27, 43–70.

SMIRNOV, N. (1948): "Table for Estimating the Goodness of Fit of Empirical Distributions," *Annals of Mathematical Statistics*, 19, 279–281.

WOOD, L. AND ALTAVELA, M. (1978): "TrustLarge-Sample Results for Kolmogorov-Smirnov Statistics for Discrete Distributions," *Biometrika*, 65, 235–239.

b	Methods	Levels \ n	100	200	400	600	800	1,000
0.10	A	1%	0.58	0.72	0.74	0.92	0.92	1.10
		5%	4.24	4.62	4.52	4.40	4.74	4.78
		10%	9.24	9.56	9.78	9.50	10.14	10.06
	B	1%	1.46	1.18	1.76	1.66	1.44	1.36
		5%	5.56	5.38	6.40	5.70	5.24	5.56
		10%	10.64	10.12	11.56	10.66	10.06	10.64
	C	1%	1.52	1.34	1.72	1.48	1.40	1.44
		5%	5.44	5.12	6.28	5.86	5.44	5.96
		10%	10.42	9.90	11.36	10.84	10.54	10.52
0.20	A	1%	0.70	0.72	0.74	0.90	0.86	0.82
		5%	5.36	5.42	5.22	5.00	4.54	4.80
		10%	10.24	10.26	9.86	9.94	9.20	9.86
	B	1%	1.64	1.48	1.40	1.74	1.80	1.58
		5%	5.62	5.00	5.72	5.54	5.30	5.90
		10%	10.60	10.06	10.52	10.88	9.84	10.72
	C	1%	1.72	1.48	1.44	1.44	1.68	1.66
		5%	5.64	4.98	5.46	5.72	5.52	5.84
		10%	10.64	9.88	10.56	10.46	10.00	10.60
0.50	A	1%	1.26	1.00	0.86	0.90	0.84	0.86
		5%	5.44	4.90	4.76	4.68	4.98	4.78
		10%	10.72	9.80	9.84	9.68	9.64	9.96
	B	1%	1.58	1.48	1.14	1.50	1.60	1.26
		5%	5.54	5.78	4.76	5.64	5.82	5.16
		10%	10.92	10.84	10.16	10.74	10.34	10.32
	C	1%	1.34	1.50	1.30	1.62	1.66	1.18
		5%	5.68	5.78	4.92	5.64	5.68	5.12
		10%	10.64	10.98	10.66	10.34	10.78	10.12
1.00	A	1%	1.00	1.04	1.16	0.96	1.02	0.96
		5%	4.90	5.02	4.76	4.94	4.74	4.26
		10%	9.42	9.30	9.46	9.28	9.40	9.08
	B	1%	1.14	1.16	1.14	1.56	1.46	1.30
		5%	5.56	4.90	5.10	5.22	5.84	5.66
		10%	10.06	9.32	10.00	10.08	11.18	10.48
	C	1%	1.18	1.08	1.06	1.32	1.56	1.80
		5%	5.04	4.76	4.98	5.00	5.68	5.44
		10%	9.98	9.38	9.76	10.02	10.96	10.56

Table 1: Empirical Levels of the Test Statistic Using the MCMD Estimator. Repetitions: 5,000. Bootstrap and Null Distribution Repetitions: 200. DGP: $X_t \sim$ Pareto(θ_*); (θ_*) $= (2.0)$; Bottom Value of Data Range (b): 1.00; Top Value of Data Range (u): 10.00; n observations are grouped into $(u-b)/d$ number of intervals such that for each $i = 1, \ldots, k$, $x_i - x_{i-1} = d$. Model: for each $i = 1, 2, \ldots, k$, $F(x_i, \theta) = 1 - [(u/x_i)^\theta - 1]/[(u/b)^\theta - 1]$.

b	Methods	Levels \ n	100	200	400	600	800	1,000
0.10	A	1%	0.06	0.06	0.06	0.04	0.06	0.06
		5%	0.32	0.28	0.40	0.54	0.28	0.42
		10%	1.34	1.06	1.24	1.24	1.20	1.04
	B	1%	0.02	0.04	0.04	0.02	0.02	0.06
		5%	0.18	0.30	0.48	0.32	0.28	0.32
		10%	0.78	1.00	1.06	0.98	1.16	1.08
	C	1%	1.14	1.42	1.50	1.34	1.58	1.52
		5%	5.62	5.20	5.50	5.24	5.94	5.58
		10%	10.54	10.80	10.96	10.10	10.92	10.80
	D	1%	1.56	1.34	1.18	1.42	1.54	1.34
		5%	6.14	5.60	5.80	6.36	6.44	6.24
		10%	11.00	10.76	10.60	11.76	11.96	11.64
0.20	A	1%	0.06	0.04	0.02	0.02	0.04	0.06
		5%	0.32	0.36	0.30	0.30	0.26	0.34
		10%	0.98	0.90	0.80	1.06	0.88	0.94
	B	1%	0.04	0.04	0.04	0.08	0.04	0.06
		5%	0.30	0.32	0.22	0.34	0.36	0.28
		10%	0.76	0.98	0.98	1.26	1.04	1.20
	C	1%	1.34	1.60	1.34	1.58	1.38	1.62
		5%	5.64	5.72	5.52	5.86	5.18	5.78
		10%	10.36	10.84	10.96	10.96	10.50	10.92
	D	1%	1.20	1.20	1.24	1.34	1.26	1.30
		5%	5.04	5.62	5.78	5.70	5.70	5.62
		10%	10.70	10.42	11.00	10.98	11.12	11.08
0.50	A	1%	0.00	0.00	0.00	0.00	0.00	0.00
		5%	0.04	0.10	0.12	0.10	0.00	0.04
		10%	0.36	0.42	0.48	0.26	0.22	0.36
	B	1%	0.02	0.00	0.02	0.00	0.00	0.00
		5%	0.10	0.08	0.08	0.06	0.08	0.08
		10%	0.38	0.40	0.40	0.34	0.28	0.42
	C	1%	1.74	1.54	1.52	1.62	1.28	1.78
		5%	5.54	5.58	6.00	5.54	5.34	5.68
		10%	10.06	10.02	10.84	9.98	10.30	10.94
	D	1%	1.10	1.24	1.32	1.24	1.00	1.10
		5%	5.04	5.82	5.60	5.44	5.30	5.78
		10%	9.84	10.72	10.82	10.64	10.22	11.02
1.00	A	1%	0.00	0.00	0.00	0.00	0.00	0.00
		5%	0.00	0.00	0.00	0.00	0.00	0.00
		10%	0.00	0.00	0.00	0.00	0.02	0.00
	B	1%	0.00	0.00	0.00	0.00	0.00	0.00
		5%	0.00	0.02	0.00	0.00	0.00	0.00
		10%	0.00	0.02	0.02	0.00	0.00	0.00
	C	1%	1.46	1.66	1.48	1.42	1.52	1.48
		5%	5.24	5.42	5.26	5.38	5.70	5.30
		10%	10.00	10.80	10.04	10.20	10.22	10.30
	D	1%	0.86	1.00	1.18	0.92	1.06	1.02
		5%	5.06	5.52	5.24	5.42	4.86	5.08
		10%	10.18	10.86	10.30	11.04	9.56	10.42

Table 2: EMPIRICAL LEVELS OF THE TEST STATISTIC USING THE ML ESTIMATOR. Repetitions: 5,000. Bootstrap and Null Distribution Repetitions: 200. DGP: $X_t \sim$ Pareto(θ_*); $(\theta_*) = (2.0)$; Bottom Value of Data Range (b): 1.00; Top Value of Data Range (u): 10.00; n observations are grouped into $(u-b)/d$ number of intervals such that for each $i = 1, \ldots, k$, $x_i - x_{i-1} = d$. Model: for each $i = 1, 2, \ldots, k$, $F(x_i, \theta) = 1 - [(u/x_i)^\theta - 1]/[(u/b)^\theta - 1]$.

b	Methods	Levels \ n	100	200	400	600	800	1,000
0.10	B	1%	43.56	77.84	98.88	99.90	100.0	100.0
		5%	68.24	93.18	99.84	100.0	100.0	100.0
		10%	81.24	97.50	99.94	100.0	100.0	100.0
	C	1%	44.80	78.32	98.94	99.82	100.0	100.0
		5%	69.74	93.66	99.80	100.0	100.0	100.0
		10%	81.28	97.46	99.94	100.0	100.0	100.0
0.20	B	1%	41.54	79.08	98.64	99.94	100.0	100.0
		5%	67.26	93.26	99.72	100.0	100.0	100.0
		10%	78.90	97.34	99.96	100.0	100.0	100.0
	C	1%	41.40	79.70	98.74	99.94	100.0	100.0
		5%	66.04	93.40	99.78	100.0	100.0	100.0
		10%	79.70	97.42	99.96	100.0	100.0	100.0
0.50	B	1%	47.60	83.18	99.02	99.94	100.0	100.0
		5%	69.36	94.80	99.82	100.0	100.0	100.0
		10%	80.32	97.60	99.96	100.0	100.0	100.0
	C	1%	45.78	82.58	98.76	99.90	100.0	100.0
		5%	70.20	94.70	99.86	100.0	100.0	100.0
		10%	80.76	97.76	99.90	100.0	100.0	100.0
1.00	B	1%	40.06	75.76	98.46	99.94	100.0	100.0
		5%	61.06	90.10	99.86	100.0	100.0	100.0
		10%	71.10	95.12	99.96	100.0	100.0	100.0
	C	1%	37.10	74.46	98.34	99.94	100.0	100.0
		5%	59.04	90.34	99.82	100.0	100.0	100.0
		10%	71.56	95.02	99.96	100.0	100.0	100.0

Table 3: Empirical Powers of the Test Statistic Using the MCMD Estimator. Repetitions: 5,000. Bootstrap and Null Distribution Repetitions: 200. DGP: $X_t \sim$ Exp(λ_*); $\lambda_* = 1.2$; Bottom Value of Data Range (b): 1.00; Top Value of Data Range (u): 10.00; n observations are grouped into $(u - b)/d$ number of intervals such that for each $i = 1, \ldots, k$, $x_i - x_{i-1} = d$. Model: for each $i = 1, 2, \ldots, k$, $F(x_i, \theta) = 1 - [(u/x_i)^\theta - 1]/[(u/b)^\theta - 1]$.

b	Methods	Levels \ n	100	20 0	400	600	800	1,000
0.10	B	1%	6.00	27.02	78.38	99.96	99.72	100.0
		5%	20.22	57.46	95.70	99.90	100.0	100.0
		10%	35.26	75.18	98.68	100.0	100.0	100.0
	C	1%	41.56	79.02	99.08	100.0	100.0	100.0
		5%	66.96	93.32	99.92	100.0	100.0	100.0
		10%	78.30	97.24	99.98	100.0	100.0	100.0
	D	1%	2.26	5.68	13.16	22.80	35.10	47.30
		5%	8.00	14.60	29.46	45.40	59.04	70.82
		10%	13.62	22.72	41.08	58.10	71.60	82.94
0.20	B	1%	5.62	26.04	78.48	97.14	99.68	100.0
		5%	19.46	57.42	95.60	99.88	100.0	100.0
		10%	34.30	75.44	98.84	99.98	100.0	100.0
	C	1%	42.14	79.80	99.26	99.98	100.0	100.0
		5%	67.20	93.62	99.94	100.0	100.0	100.0
		10%	79.42	97.28	100.0	100.0	100.0	100.0
	D	1%	4.08	7.14	17.26	30.70	45.84	58.70
		5%	11.48	19.26	38.22	58.64	75.60	89.70
		10%	18.54	29.76	53.10	76.12	91.36	98.28
0.50	B	1%	5.68	26.72	78.74	96.82	99.72	99.96
		5%	19.40	56.46	95.04	99.72	99.96	100.0
		10%	34.12	73.00	98.22	99.92	100.0	100.0
	C	1%	51.04	84.20	99.42	99.98	100.0	100.0
		5%	73.40	95.58	99.98	100.0	100.0	100.0
		10%	83.26	97.96	99.98	100.0	100.0	100.0
	D	1%	15.60	34.70	75.42	93.40	98.96	99.94
		5%	34.40	62.90	93.08	99.14	99.92	100.0
		10%	47.22	76.30	96.94	99.80	100.0	100.0
1.00	B	1%	0.04	1.36	16.78	51.26	82.06	95.40
		5%	1.58	11.48	59.42	90.18	99.12	99.78
		10%	5.82	28.12	82.24	97.66	99.94	99.94
	C	1%	38.90	78.32	98.82	99.94	100.0	100.0
		5%	62.22	92.16	99.86	100.0	100.0	100.0
		10%	73.46	95.96	100.0	100.0	100.0	100.0
	D	1%	28.62	57.62	92.28	99.04	99.94	100.0
		5%	50.30	80.20	98.58	99.96	100.0	100.0
		10%	63.20	88.82	99.46	100.0	100.0	100.0

Table 4: EMPIRICAL POWERS OF THE TEST STATISTIC USING THE ML ESTIMATOR. Repetitions: 5,000. Bootstrap and Null Distribution Repetitions: 200. DGP: $X_t \sim$ Exp(λ_*); $\lambda_* = 1.2$; Bottom Value of Data Range (b): 1.00; Top Value of Data Range (u): 10.00; n observations are grouped into $(u - b)/d$ number of intervals such that for each $i = 1, \ldots, k$, $x_i - x_{i-1} = d$. Model: for each $i = 1, 2, \ldots, k$, $F(x_i, \theta) = 1 - [(u/x_i)^\theta - 1]/[(u/b)^\theta - 1]$.

b	Methods	Levels $\setminus$ n	100	200	400	600	800	1,000
0.10	B	1%	10.88	11.32	11.54	11.32	11.36	11.94
		5%	24.16	26.00	25.94	25.68	25.12	26.78
		10%	35.26	36.82	37.68	37.06	35.82	37.56
	C	1%	11.18	11.46	11.52	11.34	11.88	12.38
		5%	24.14	26.42	25.98	25.76	25.60	26.70
		10%	35.48	37.14	38.08	36.68	36.18	38.26
0.20	B	1%	10.96	11.18	10.86	11.72	11.58	11.44
		5%	25.20	24.90	25.28	25.88	26.46	26.18
		10%	36.54	35.94	36.80	36.38	37.36	37.02
	C	1%	10.86	11.18	10.72	11.56	11.20	11.48
		5%	25.28	24.90	25.38	25.66	26.08	26.30
		10%	36.78	36.18	37.02	36.56	37.68	37.14
0.50	B	1%	11.52	11.46	11.90	10.72	11.78	10.94
		5%	25.14	24.70	24.12	23.64	24.36	24.20
		10%	35.56	34.78	34.60	34.10	34.16	34.30
	C	1%	11.62	11.48	11.58	11.00	11.34	10.86
		5%	24.78	24.26	24.82	23.86	24.70	23.72
		10%	35.54	34.74	34.42	34.02	34.36	34.56
1.00	B	1%	6.80	7.06	5.94	6.20	5.84	5.72
		5%	16.24	16.54	14.84	15.06	14.82	14.32
		10%	25.46	25.14	23.00	23.20	23.32	21.98
	C	1%	6.34	6.56	5.68	6.34	5.72	5.46
		5%	15.60	16.30	14.88	14.92	14.38	14.50
		10%	25.18	24.70	23.06	22.52	23.44	22.24

Table 5: Empirical Local Powers of the Test Statistic Using the MCMD Estimator. Repetitions: 5,000. Bootstrap and Null Distribution Repetitions: 200. DGP: $X_t \sim (1 - 5/\sqrt{n})$Pareto$(\theta_*)$ + $(5/\sqrt{n})$Exp(λ_*); $(\theta_*, \lambda_*) = (2.0, 1.2)$; Bottom Value of Data Range (b): 1.00; Top Value of Data Range (u): 10.00; n observations are grouped into $(u - b)/d$ number of intervals such that for each $i = 1, \ldots, k$, $x_i - x_{i-1} = d$. Model: for each $i = 1, 2, \ldots, k$, $F(x_i, \theta) = 1 - [(u/x_i)^\theta - 1]/[(u/b)^\theta - 1]$.

b	Methods	Levels \ n	100	200	400	600	800	1,000
0.10	B	1%	0.46	0.72	1.08	0.92	0.90	0.72
		5%	3.42	4.44	4.20	4.34	4.76	4.58
		10%	7.66	9.46	10.08	9.66	10.00	9.80
	C	1%	10.14	11.22	11.32	11.28	11.86	11.20
		5%	23.70	25.22	25.42	25.58	26.42	25.98
		10%	34.18	36.82	35.78	36.56	37.26	37.60
	D	1%	1.26	1.70	1.80	1.90	2.04	1.96
		5%	5.14	6.30	6.96	7.06	7.46	7.48
		10%	8.90	11.24	12.00	12.70	13.14	13.30
0.20	B	1%	0.64	0.96	0.74	0.98	0.74	0.88
		5%	3.86	4.40	4.14	4.58	4.00	4.24
		10%	8.68	9.02	8.80	9.72	9.06	9.46
	C	1%	11.66	11.72	11.22	11.82	11.28	11.52
		5%	25.80	26.52	25.26	25.86	25.20	26.52
		10%	36.60	36.80	36.00	36.92	35.86	37.16
	D	1%	2.10	2.24	2.14	2.44	2.72	2.68
		5%	6.96	7.88	7.80	8.54	8.92	8.56
		10%	12.04	13.62	14.40	14.68	15.88	15.46
0.50	B	1%	0.34	0.28	0.38	0.26	0.32	0.30
		5%	2.28	2.22	2.66	2.06	2.42	2.20
		10%	5.44	5.74	5.96	5.60	5.76	5.22
	C	1%	11.94	12.00	12.18	12.32	12.78	12.24
		5%	26.12	25.48	25.42	24.60	26.48	25.10
		10%	35.74	35.76	35.52	34.88	36.72	34.60
	D	1%	4.58	5.28	4.72	4.74	4.96	4.70
		5%	14.30	14.18	14.06	13.62	13.78	14.36
		10%	21.08	21.82	22.34	21.70	22.02	23.82
1.00	B	1%	0.00	0.00	0.00	0.02	0.00	0.00
		5%	0.00	0.00	0.04	0.06	0.06	0.02
		10%	0.20	0.18	0.30	0.30	0.24	0.20
	C	1%	6.12	6.28	6.76	6.02	6.08	5.68
		5%	15.62	15.94	16.48	15.84	15.20	14.92
		10%	25.24	23.94	24.42	23.86	24.00	23.84
	D	1%	5.56	5.26	4.50	4.52	3.94	3.90
		5%	15.44	14.38	12.80	13.16	13.18	13.14
		10%	22.82	22.10	21.00	20.64	21.62	21.60

Table 6: EMPIRICAL LOCAL POWERS OF THE TEST STATISTIC USING THE ML ESTIMATOR. Repetitions: 5,000. Bootstrap and Null Distribution Repetitions: 200. DGP: $X_t \sim (1 - 5/\sqrt{n})$Pareto$(\theta_*) + (5/\sqrt{n})Exp(\lambda_*)$; $(\theta_*, \lambda_*) = (2.0, 1.2)$; Bottom Value of Data Range (b): 1.00; Top Value of Data Range (u): 10.00; n observations are grouped into $(u - b)/d$ number of intervals such that for each $i = 1, \ldots, k$, $x_i - x_{i-1} = d$. Model: for each $i = 1, 2, \ldots, k$, $F(x_i, \theta) = 1 - [(u/x_i)^\theta - 1]/[(u/b)^\theta - 1]$.

Statistics\ Years	2007	2008	2009	2010	2011	2012
Sample Size	10,464,206	11,066,599	11,590178	12,448,203	13,265,840	14,104,742
Total Group Size	2,761	3,352	3,418	3,988	3,553	4,241
First Group	(0.0, 0.5]	(0.0, 0.5]	(0.0, 0.5]	(0.0, 0.5]	(0.0, 0.6]	(0.0, 0.6]
Last Group	(276.40 ∞]	(335.50, ∞]	(342.10, ∞]	(399.10, ∞]	(355.70, ∞]	(424.50, ∞]
Sample Average$_1$	0.3325	0.3361	0.3331	0.3456	0.3582	0.3640
Sample Average$_2$	0.8434	0.8443	0.8443	0.8723	1.0428	1.0334
Sample Variance	2.2668	2.2194	2.1780	2.5191	3.6052	3.4154
Sample Skewness	59.823	67.084	70.287	73.061	59.930	68.368
Sample Kurtosis	6369.6	8075.6	8937.7	9378.1	6498.4	8807.1
Sample Median	138.45	168.00	171.30	199.80	178.10	212.50
Sample Mode Group	(0.5, 0.6]	(0.5, 0.6]	(0.5, 0.6]	(0.5, 0.6]	(0.6, 0.7]	(0.6, 0.7]
Populations ≥ 15	39,873,045	40,459,969	40,949,973	41,434,992	42,008,528	42,445,378
Populations ≥ 20	36,640,987	37,133,082	37,536,274	37,967,813	38,540,049	39,021,687
Labor Forces	39,170,000	39,598,000	40,092,000	40,590,000	41,052,000	41,582,000
Employments	23,433,000	23,577,000	23,506,000	23,829,000	24,244,000	24,681,000

Table 7: Descriptive Statistics of Income Tabulations and Populations of Korea (2007–2012) All other groups from the second to the second to last have the same group interval size: KRW 10 mil. All income figures are measured in KRW 100 mil. Sample Average$_1$ is the average of sample incomes computed by the National Tax Service of Korea using all individual observations. Sample average$_2$, sample variance, sample skewness, sample kurtosis, sample mode group, and sample median group are the statistics obtained by using the group median values and their frequencies from the second to the second to last groups.

Years	Top x-%	Statistics \ Populations	$\geq$ 15 year old	$\geq$ 20 year old	measured by Labor Forces	measured by Employments
2007	1.00%	b	0.5000	0.5000	0.5000	0.5000
		u	2.5000	2.8000	2.5000	2.5000
		$\widehat{x}_n$	0.8802	$0.9000^{\sharp}$	0.8860	1.0735
		p-value of $\widehat{T}_n$	0.0000	0.0000	0.0000	0.0000
	0.10%	b	2.1000	2.1000	2.1000	2.1000
		u	4.1000	4.1000	4.1000	4.1000
		$\widehat{x}_n$	2.3370	2.4484	2.3599	3.1541
		p-value of $\widehat{T}_n$	45.500	45.500	45.500	45.000
	0.05%	b	2.1000	2.1000	2.1000	3.5000
		u	4.1000	4.1000	4.1000	6.1000
		$\widehat{x}_n$	3.4698	3.6512	3.5070	4.8222
		p-value of $\widehat{T}_n$	45.500	45.500	45.500	94.000
	0.01%	b	8.5000	8.5000	8.5000	12.000
		u	10.500	11.500	10.500	14.000
		$\widehat{x}_n$	9.5534	10.072	9.6618	13.206
		p-value of $\widehat{T}_n$	97.500	99.500	97.500	78.500
2008	1.00%	b	0.5000	0.5000	0.5000	0.5000
		u	2.5000	2.5000	2.5000	2.5000
		$\widehat{x}_n$	0.9286	0.9593	0.9362	$1.1000^{\sharp}$
		p-value of $\widehat{T}_n$	0.0000	0.0000	0.0000	0.0000
	0.10%	b	2.4000	2.4000	2.4000	2.4000
		u	4.4000	4.4000	4.4000	4.4000
		$\widehat{x}_n$	2.4325	2.5497	2.4614	3.2908
		p-value of $\widehat{T}_n$	100.00	100.00	100.00	100.00
	0.05%	b	2.4000	2.4000	2.4000	4.3000
		u	4.4000	4.4000	4.4000	6.5000
		$\widehat{x}_n$	3.5975	3.7852	3.6435	4.9975
		p-value of $\widehat{T}_n$	100.00	100.00	100.00	37.000
	0.01%	b	9.0000	9.0000	9.0000	12.200
		u	11.000	11.000	11.000	14.200
		$\widehat{x}_n$	9.5740	10.117	9.7095	13.291
		p-value of $\widehat{T}_n$	72.500	72.500	72.500	52.500
2009	1.00%	b	0.5000	0.5000	0.5000	0.5000
		u	2.5000	2.5000	2.5000	2.5000
		$\widehat{x}_n$	0.9382	0.9692	0.9458	1.1609
		p-value of $\widehat{T}_n$	0.0000	0.0000	0.0000	0.0000
	0.10%	b	2.4000	2.5000	2.5000	2.5000
		u	4.4000	5.8000	5.8000	4.8000
		$\widehat{x}_n$	2.4943	2.6136	2.5191	3.3987
		p-value of $\widehat{T}_n$	100.00	14.000	14.000	100.00
	0.05%	b	2.5000	2.5000	2.5000	2.5000
		u	5.8000	5.8000	5.8000	5.8000
		$\widehat{x}_n$	3.6794	3.8672	3.7241	5.0761
		p-value of $\widehat{T}_n$	14.000	14.000	14.000	14.000
	0.01%	b	8.4000	8.4000	8.4000	12.000
		u	11.000	11.000	11.000	14.200
		$\widehat{x}_n$	9.4237	9.9125	9.5404	12.946
		p-value of $\widehat{T}_n$	100.00	100.00	100.00	19.000

Table 8: EMPIRICAL TESTING AND TOP INCOME ESTIMATION OF KOREA (2007–2009). Notes: b and u are the lower and upper border values of the grouped data; $\widehat{x}_n$ is the estimated top x-% income out of the given population; superscript $\sharp$ indicates that the estimated top x-% income is identical to $x_{\sharp}$. The units of b, u, and $\widehat{x}_n$ are KRW 100 mil., and p-values are in %.

Years	Top x-%	Statistics \ Populations	$\geq$ 15 year old	$\geq$ 20 year old	measured by Labor Forces	measured by Employments
2010	1.00%	b	0.5000	0.5000	0.5000	0.5000
		u	2.5000	2.5000	2.5000	2.5000
		$\widehat{x}_n$	1.0000$^\sharp$	1.0686	1.0400	1.2923
		p-value of $\widehat{T}_n$	0.0000	0.0000	0.0000	0.0000
	0.10%	b	2.2000	2.2000	2.2000	2.2000
		u	6.1000	6.1000	6.1000	6.1000
		$\widehat{x}_n$	2.7646	2.8973	2.7953	3.7364
		p-value of $\widehat{T}_n$	22.000	22.000	22.000	22.000
	0.05%	b	2.2000	2.2000	2.2000	2.2000
		u	6.1000	6.1000	6.1000	6.1000
		$\widehat{x}_n$	4.0410	4.2463	4.0883	5.5887
		p-value of $\widehat{T}_n$	22.000	22.000	22.000	22.000
	0.01%	b	8.0000	10.000	9.0000	10.000
		u	12.000	15.000	12.000	15.000
		$\widehat{x}_n$	10.392	10.981	10.500$^\sharp$	14.626
		p-value of $\widehat{T}_n$	35.500	58.500	36.000	58.500
2011	1.00%	b	0.6000	0.6000	0.6000	0.6000
		u	2.6000	2.6000	2.6000	2.6000
		$\widehat{x}_n$	1.0739	1.1000$^\sharp$	1.0836	1.3000$^\sharp$
		p-value of $\widehat{T}_n$	0.0000	0.0000	0.0000	0.0000
	0.10%	b	2.8000	2.8000	2.8000	2.8000
		u	6.9000	6.9000	6.9000	6.9000
		$\widehat{x}_n$	3.0855	3.2428	3.1268	4.2407
		p-value of $\widehat{T}_n$	29.000	29.000	29.000	29.000
	0.05%	b	2.8000	2.8000	2.8000	2.8000
		u	6.9000	6.9000	6.9000	6.9000
		$\widehat{x}_n$	4.6096	4.8471	4.6719	6.3624
		p-value of $\widehat{T}_n$	29.000	29.000	29.000	29.000
	0.01%	b	10.000	11.000	10.000	15.000
		u	13.000	13.100	13.000	17.000
		$\widehat{x}_n$	11.825	12.396	11.980	16.248
		p-value of $\widehat{T}_n$	17.500	100.00	17.500	57.000
2012	1.00%	b	0.6000	0.6000	0.6000	0.6000
		u	2.6000	2.6000	2.6000	2.6000
		$\widehat{x}_n$	1.1000$^\sharp$	1.1521	1.1239	1.3817
		p-value of $\widehat{T}_n$	0.0000	0.0000	0.0000	0.0000
	0.10%	b	2.0000	2.0000	2.0000	2.0000
		u	7.0000	7.0000	7.0000	7.0000
		$\widehat{x}_n$	3.1511	3.2980	3.1863	4.2427
		p-value of $\widehat{T}_n$	100.00	100.00	100.00	100.00
	0.05%	b	2.0000	2.0000	2.0000	2.0000
		u	7.0000	7.0000	7.0000	7.0000
		$\widehat{x}_n$	4.6182	4.8442	4.6723	6.3367
		p-value of $\widehat{T}_n$	100.00	100.00	100.00	100.00
	0.01%	b	9.7000	10.500	11.000	15.700
		u	12.000	13.500	13.000	17.900
		$\widehat{x}_n$	11.789	12.381	11.931	16.161
		p-value of $\widehat{T}_n$	66.500	34.000	78.500	100.00

Table 9: EMPIRICAL TESTING AND TOP INCOME ESTIMATION OF KOREA (2010–2012). Notes: b and u are the lower and upper border values of the grouped data; $\widehat{x}_n$ is the estimated top x-% income out of the given population; superscript $\sharp$ indicates that the estimated top x-% income is identical to $x_\sharp$. The units of b, u, and $\widehat{x}_n$ are KRW 100 mil., and p-values are in %.

Top x-%	Years \ Populations	$\geq$ 15 year old	$\geq$ 20 year old	measured by Labor Forces	measured by Employments
1.00%	2007	11.45 (11.70)	11.06 (11.19)	11.33 (11.59)	8.63 (8.93)
	2008	11.37 (11.79)	10.80 (11.26)	11.23 (11.65)	8.79 (8.94)
	2009	11.21 (11.69)	10.64 (11.17)	11.07 (11.56)	8.24 (8.82)
	2010	12.25 (12.38)	11.09 (11.83)	11.56 (12.24)	8.57 (9.38)
	2011	12.55 (12.89)	12.07 (12.33)	12.37 (12.74)	9.53 (9.84)
	2012	12.22 (12.29)	11.37 (11.77)	11.82 (12.16)	8.86 (9.40)
0.10%	2007	4.10 (4.10)	3.96 (3.96)	4.07 (4.07)	3.32 (3.31)
	2008	4.10 (4.11)	3.96 (3.96)	4.07 (4.02)	3.32 (3.30)
	2009	4.05 (4.05)	3.90 (3.91)	4.01 (4.02)	3.22 (3.22)
	2010	4.34 (4.34)	4.19 (4.19)	4.30 (4.31)	3.46 (3.46)
	2011	4.50 (4.61)	4.33 (4.45)	4.45 (4.57)	3.56 (3.67)
	2012	4.31 (4.32)	4.16 (4.17)	4.27 (4.28)	3.44 (3.44)
0.05%	2007	3.11 (3.11)	3.01 (3.00)	3.09 (3.08)	2.51 (2.51)
	2008	3.10 (3.10)	3.00 (3.00)	3.07 (3.07)	2.49 (2.49)
	2009	3.04 (3.04)	2.93 (2.93)	3.01 (3.01)	2.42 (2.42)
	2010	3.28 (3.27)	3.17 (3.16)	3.25 (3.25)	2.62 (2.62)
	2011	3.35 (3.45)	3.23 (3.33)	3.32 (3.42)	2.64 (2.75)
	2012	3.24 (3.23)	3.13 (3.12)	3.21 (3.20)	2.58 (2.57)
0.01%	2007	1.61 (1.62)	1.56 (1.56)	1.60 (1.60)	1.28 (1.29)
	2008	1.61 (1.61)	1.56 (1.56)	1.60 (1.60)	1.29 (1.29)
	2009	1.57 (1.57)	1.52 (1.52)	1.56 (1.56)	1.26 (1.26)
	2010	1.72 (1.72)	1.66 (1.66)	1.71 (1.71)	1.38 (1.38)
	2011	1.65 (1.76)	1.59 (1.70)	1.63 (1.74)	1.29 (1.40)
	2012	1.64 (1.64)	1.58 (1.59)	1.63 (1.63)	1.31 (1.31)

Table 10: TOP INCOME SHARES OF KOREA(2007–2012, IN %). The figures show the share of the top x-% income out of total income of each year. The figures are the same shares measured by Atkinson's (2005) mean-split histogram method computed by Park and Jeon (2014).