# Practical Kolmogorov-Smirnov Testing by Minimum Distance Applied to Measure 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: May 2016

## Abstract

We study Kolmogorov-Smirnov goodness of fit tests for evaluating distributional hypotheses where unknown parameters need to be fitted. Following work of Pollard (1980), 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 estimating the upper 0.1% or higher income shares, the conventional assumption of a Pareto tail distribution cannot be rejected. But the Pareto tail hypothesis is rejected for estimating the top 1.0% or 0.5% income shares at the 5% significance level. A Supplement containing proofs and data descriptions is available online.
*Key Words:* Distribution-free asymptotics, null distribution, minimum distance estimator, Crámer-von Mises distance, top income shares, Pareto interpolation.

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

---

*Corresponding author: jinseocho@yonsei.ac.kr +82 2 2123 5448

# 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 tests.

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 from using the MCMD estimator for the KS distance is that asymptotic critical values can be obtained by applying invariance principle arguments in the same way as for the original KS test statistic without unknown parameter. The current paper further provides the form of the functional linked to the Brownian bridge and demonstrates its implementation in a practical application by a convenient simulation-based calculation of asymptotic critical values. By this, an efficient test procedure is delivered that overcomes the inefficacy of the KS test pointed out in Durbin (1973) and Henze (1996).

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) recommends 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. Furthermore, simulations show that the new methodology has performance characteristics similar to those of the parametric bootstrap, whereas it saves the computation times more substantially than the parametric bootstrap and exhibits better power than the test obtained by Khmaladze's (2013) transformation.

The KS test statistic has null and local alternative asymptotic distributions that depend upon data types. The practical importance 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 limit distribution of the KS test statistic is derived from the large sample limit distribution by increasing the number of groups and keeping 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 also 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 simulated.

As an empirical application of the KS test statistic presented here, we revisit the problem of esti-

mating 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 (2015), 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. Atkinson, Piketty, and Saez (2011) also 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. 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 when estimating the top 1.0% income, although the hypothesis is not rejected further in the tail of the distribution when estimating 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 numbers of groups were 2,760 and 4,241, respectively. Based on this degree of detail for our data, we also compare our top income shares with those estimated by Kim and Kim (2015) using Korean tax income tax data with at most 10 income groups. From this, we find that our estimates are very close to those in Kim and Kim (2015) and affirm the usefulness of their methodology.

The plan of this paper is as follows. Section 2 develops the limit theory for the MCMD estimator and associated KS test for grouped data. The asymptotic null distribution, power, and local power of the test are also derived. Section 3 examines the large sample 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 to the Korean income tax return data from 2007 to 2012. Conclusions are given in Section 6. Proofs, related material, and data explanations are provided in the Supplement.

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 $\nabla_\theta F(x, \theta) = (\partial/\partial\theta)F(x, \theta)$, and for $i = 0, 1$, $\partial_j^i F(x, \theta) := (\partial^i/\partial\theta_j)F(x, \theta)$ so that, for $i = 0$, $\partial_j^i F(x, \theta) \equiv F(x, \theta)$. Finally, $\|X\|_{\max} := \max_{i,j} |x_{i,j}|$, where $x_{i,j}$ is the $i$-th row and $j$-th column element of $X$.

# 2  Testing Distributional Hypotheses for Grouped Data

Suppose that data is available in a group frequency format whereby some variable of interest, $X_i$, may be unobserved but the numbers of times $X_i$ lies within certain specified groups are observed. Typical income data have this format. In such cases, if $X_i$ is annual income earned by individual $i$, then income data is provided in pairs: $\{[c_j, c_{j+1}), \#\{X_i \in [c_j, c_{j+1})\} : j = 0, \ldots, k-1; i = 1, \ldots, n\}$ along with total group incomes $I_{j+1} := \sum_{i=1}^{n} X_i \times \{X_i \in [c_j, c_{j+1})\}$. According to this scheme, $n$ is the sample size of the number of individuals, $k$ is the number of income groups, $c_j$ and $c_{j+1}$ are the lower and upper bounds of $(j+1)$-th interval, $\#\{A\}$ is the frequency $A$ is observed, and $\{\cdot\}$ is the indicator function. The end points $c_0$ and $c_k$ 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 (2015) among others have estimated the top income shares in the US, UK, France, Australia, New Zealand, Japan, and Korea. Their 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 $X_i$ underlying the observed grouped data. For practical reasons in what follows we consider distributions bounded below and above by $b = c_0$ and $u = c_k$, respectively. For example, 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 following hypotheses are considered:

$\mathcal{H}_0$: for all $c_j$, there is a parameter value $\theta_*$ for *cdf* $F$ such that $\mathbb{P}(X_i \leq c_j \mid b \leq X_i \leq u) = F(c_j, \theta_*)$;

$\mathcal{H}_1$: there is no $\theta_*$ for *cdf* $F$ such that for all $c_j$, $\mathbb{P}(X_i \leq c_j \mid b \leq X_i \leq u) = F(c_j, \theta_*)$.

Under these hypotheses, the parameter value $\theta_*$ is properly defined only under the null $\mathcal{H}_0$.

The null hypothesis 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 (2015) 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 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. For grouped or discretely distributed data, Wood and Altavela (1978) 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) has 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 that approximates its finite sample distribution well and is 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 still has a null distribution that is well approximated by its asymptotic distribution 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)* $\{X_i \in \mathbb{R}\}$ *is independently and identically distributed (IID) with a continuous cdf* $p(\cdot) := \mathbb{P}(X_i \leq (\cdot) \mid b \leq X_i \leq u)$; *(ii) For every* $j = 1, 2, \ldots, k$, $F(c_j, \cdot) : \Theta \mapsto [0, 1]$ *is in* $\mathcal{C}^{(1)}(\Theta)$, *the space of continuously differentiable functions, where* $\Theta \subset \mathbb{R}^d$ *is a compact and convex set with* $k > d$

*and such that* $-\infty < b = c_0 < \ldots < c_k = u < \infty$; *(iii)* $\theta_o := \arg\min_{\theta \in \Theta} Q(\theta) \in \text{int}(\Theta)$ *and is unique in* $\Theta$, *where for each* $\theta \in \Theta$, $Q(\theta) := \sum_{j=1}^{k}\{F(c_j, \theta) - p(c_j)\}^2$; *and (iv)* $Z'Z$ *is positive definite, where* $Z := [\nabla_\theta F(c_1, \theta_o), \ldots, \nabla_\theta F(c_k, \theta_o)]'$. $\qquad\qquad\square$

Some remarks are warranted. First, note that $Z'Z = \sum_{j=1}^{k-1} \nabla_\theta F(c_j, \theta_o) \nabla'_\theta F(c_j, \theta_o)$ due to the fact that $\nabla_\theta F(c_k, \theta_o) \equiv 0$. We include $\nabla_\theta F(c_k, \theta_o)$ in $Z$ so that the MCMD estimator can be viewed analogously to the least squares estimator, below. Second, the practical application of the current approach may depend on $\dim(\theta)$. If $\dim(\theta)$ is so large that the number of bins $k$ is less than $d$, $Z'Z$ becomes singular, and it is hard to satisfy Assumption A(iv). As $(Z'Z)^{-1}$ is a component constituting the asymptotic distribution of the MCMD estimator as we detail below, we require that $Z'Z$ be positive definite in addition to $k > d$. Finally, both $Q(\cdot)$ and $Z$ depend on the number of groups $k$, so it would be more appropriate to indicate this dependence with the notation $Q(\cdot; k)$ and $Z_k$, 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 $\theta_*$ we first employ the empirical distribution function $\widehat{p}_n(c) := n^{-1}\#\{X_i \leq c\}$, where $c$ is generic notation for $c_1, \ldots, c_k$. Since $p(c) = \mathbb{P}(X_i \leq c \mid b \leq X_i \leq u)$ is the conditional mean of $\widehat{p}_n(c)$, we have by standard limit theory $\sqrt{n}\{\widehat{p}_n(c) - p(c)\} \overset{A}{\sim} N[0, p(c)(1 - p(c))]$ and $\sqrt{n}(\widehat{p}_n(\cdot) - p(\cdot)) \Rightarrow \mathcal{B}^o(\cdot)$, where the limit process $\mathcal{B}^o(\cdot)$ is a Brownian bridge. To estimate the unknown parameter $\theta_*$ in the posited model with *cdf* $F(\cdot, \theta)$ we then perform a least squares minimum distance estimation between the non-parametric estimate $\widehat{p}_n(\cdot)$ and the parametric model $F(\cdot, \theta)$ over the end points of the intervals of observation, giving the MCMD estimator $\widehat{\theta}_n := \arg\min_{\theta \in \Theta} Q_n(\theta)$, where $Q_n(\theta) := \sum_{j=1}^{k}\{\widehat{p}_{n,j} - F_j(\theta)\}^2$ with $\widehat{p}_{n,j} := \widehat{p}_n(c_j)$ and $F_j(\cdot) := F(c_j, \cdot)$ for $j = 1, \ldots, k$. Similarly, we also let $p_j := p(c_j)$.

The MCMD structure promotes analysis in terms of a regression of $\widehat{p}_n(\cdot)$, which is a data-based uniformly consistent estimate of $p(\cdot)$, on the nonstochastic mean regressor function $F(\cdot, \theta)$ that equals $p(\cdot)$ under the null. It is convenient to maintain this regression interpretation in what follows. Note that (i) $\widehat{p}_n(\cdot)$ is uniformly consistent for $p(\cdot)$, and (ii) $|p(\cdot) - F(\cdot, \theta)|$ is bounded between 0 and 2 uniformly in $\theta$. Therefore, $\arg\min_\theta Q_n(\theta) = \arg\min_\theta Q(\theta) + o_{\text{a.s.}}(1)$ and $\widehat{\theta}_n$ is consistent for $\theta_o := \arg\min_\theta Q(\theta)$:

**Theorem 1.** *Given Assumption A,* $\widehat{\theta}_n \overset{\text{a.s.}}{\to} \theta_o$. $\qquad\qquad\square$

When the null hypothesis is correct, we have $\theta_o = \theta_*$ because $F(\cdot, \theta_*) = p(\cdot)$ under the null, whereas $\theta_*$ is undefined under the alternative. The convergence rate of $\widehat{\theta}_n$ 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 $F_j(\widehat{\theta}_n) = F_j(\theta_o) + \nabla'_\theta F_j(\theta_o)(\widehat{\theta}_n - \theta_o) + o_{\mathbb{P}}(n^{-1/2})$, which implies in view of Theorem 1 that $Q_n(\widehat{\theta}_n) = \sum_{j=1}^{k}[\widehat{p}_{n,j} - F_j(\widehat{\theta}_n)]^2 = \sum_{j=1}^{k}\{[\widehat{p}_{n,j} - F_j(\theta_o)] - \nabla'_\theta F_j(\theta_o)(\widehat{\theta}_n - \theta_o)\}^2 + o_{\mathbb{P}}(1)$. We let $\widehat{y}_j := \widehat{p}_{n,j} - F_j(\theta_o)$ and $Z_j := \nabla_\theta F_j(\theta_o)$, so that the leading term on the right side is a sum of the squared residuals in a regression of $\widehat{y}_j$ on $Z_j$ with regression coefficient $\widehat{\theta}_n - \theta_o$. Thus, the MCMD estimator is asymptotically equivalent to the least squares regression on a linear pseudo-model involving $\widehat{y}_j$ and $Z_j$, viz., $(\widehat{\theta}_n - \theta_o) = (Z'Z)^{-1}Z'\widehat{Y} + O_{\mathbb{P}}(n^{-1})$, where $\widehat{Y} := [\widehat{y}_1, \dots, \widehat{y}_k]'$ and $Z := [Z_1, \dots, Z_k]'$.

Since $\sqrt{n}\{\widehat{p}_n(\cdot) - p(\cdot)\} \Rightarrow \mathcal{B}^o(\cdot)$, we also have $\sqrt{n}\{\widehat{y}_{(\cdot)} - y_{(\cdot)}\} = \sqrt{n}\{\widehat{p}_n(\cdot) - p(\cdot)\} \Rightarrow \mathcal{B}^o(\cdot)$, where for each $j$, we define $y_j := p_j - F_j(\theta_o)$. Note that for each $j$, $y_j = 0$ under $\mathcal{H}_0$, whereas for some $j$, $y_j \neq 0$ under $\mathcal{H}_1$. This difference ensures the consistency of the KS test statistic that we introduce in the next subsection. The regression coefficient $\widehat{\theta}_n - \theta_o$ now satisfies the following property: $\sqrt{n}\{\widehat{\theta}_n - \theta_o - (Z'Z)^{-1}Z'Y\} = \sqrt{n}(Z'Z)^{-1}Z'(\widehat{Y} - Y) + O_{\mathbb{P}}(n^{-1/2}) \Rightarrow (Z'Z)^{-1}Z'W$, where $Y := [y_1, \dots, y_k]'$ and $W := [\mathcal{B}^o(p_1), \dots, \mathcal{B}^o(p_k)]'$. Here, the asymptotic bias of the MCMD estimator $(Z'Z)^{-1}Z'Y$ is zero from that the first-order condition for the optimum $\theta_o$ of $Q(\theta)$ in $\Theta$ implies that $\sum_{j=1}^{k}\{F_j(\theta_o) - p_j\}\nabla_\theta F_j(\theta_o) = 0$, or $Z'Y = 0$, which orthogonality ensures that $(Z'Z)^{-1}Z'Y = 0$. Therefore, $\sqrt{n}(\widehat{\theta}_n - \theta_o) \Rightarrow (Z'Z)^{-1}Z'W$, which leads directly to the limit distribution of $\widehat{\theta}_n$.

**Theorem 2.** *Given Assumption A, $\sqrt{n}(\widehat{\theta}_n - \theta_o) \overset{A}{\sim} N[0, (Z'Z)^{-1}Z'\Sigma Z(Z'Z)^{-1}]$, where $\Sigma$ is a $k \times k$ matrix whose $i$-th row and $j$-th column element is $\min[p_i, p_j](1 - \max[p_i, p_j])$.* $\square$

Some remarks on this result are in order. First, the core of this limit theory is the simple linear functional $(Z'Z)^{-1}Z'W$ of $W$, implying that the asymptotic distribution of $\widehat{\theta}_n$ has the form of a functional of the Brownian bridge process that appears in the asymptotic null distribution of the KS test with no unknown parameters. Correspondingly, the KS test that depends on the use of $\widehat{\theta}_n$ 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 $\theta$ in the KS test.

Second, the asymptotic distribution of $\widehat{\theta}_n$ 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 $Q_n(\cdot)$ 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 $\widehat{\theta}_n$ 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 $\widehat{T}_n := \sup_{j \leq k} |\sqrt{n}\{\widehat{p}_{n,j} - F_j(\widehat{\theta}_n)\}|$, 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 $\widehat{\theta}_n$ in $\widehat{T}_n$. We distinguish $\widehat{T}_n$ from the usual statistic with no parameter estimation error which we define as $T_n := \sup_{j \leq k} |\sqrt{n}\{\widehat{p}_{n,j} - F_j(\theta_*)\}|$.

We first develop asymptotic theory under $\mathcal{H}_0$, where $\theta_o = \theta_*$. Hence, for each $c_j$ we have $\widehat{p}_{n,j} - F_j(\widehat{\theta}_n) = \widehat{p}_{n,j} - F_j(\theta_*) - \nabla'_\theta F_j(\theta_*)(\widehat{\theta}_n - \theta_*) + o_{\mathbb{P}}(1)$, which implies that $\sup_{j \leq k} \sqrt{n}|\widehat{p}_{n,j} - F_j(\widehat{\theta}_n)| \Rightarrow \sup_{j \leq k} |\mathcal{B}^o(p_j) - Z'_j(Z'Z)^{-1}Z'W|$ by continuous mapping. The null limit distribution is therefore bounded in probability as a functional of $\mathcal{B}^o(\cdot)$, and the component $\mathcal{B}^o(p_j) - Z'_j(Z'Z)^{-1}Z'W$ is the $j$-th row element of $MW$, where $M := I - Z(Z'Z)^{-1}Z'$. Therefore, $\widehat{S}_n := \sqrt{n}[\widehat{p}_{n,1} - F_1(\widehat{\theta}_n), ..., \widehat{p}_{n,k} - F_k(\widehat{\theta}_n)]' \Rightarrow MW \sim N(0, M\Sigma M)$. The matrix $M$ projects onto the orthogonal complement of the range of the $k \times d$ matrix $Z$, so the rank of $M\Sigma M$ is $k - d$. For notational simplicity, let $G$ denote $MW$. Then, $\widehat{T}_n \Rightarrow \max \|G\|_{\max}$. Note that the asymptotic null distribution of $\widehat{S}_n$ differs from that of the KS test statistic with no unknown parameters because in that case $S_n := \sqrt{n}[\widehat{p}_{n,1} - F_1(\theta_*), ..., \widehat{p}_{n,k} - F_k(\theta_*)]' \Rightarrow W \sim N(0, \Sigma)$, so that in the limit $\widehat{T}_n$ and $T_n$ are constructed from different functionals of the same Brownian bridge process $\mathcal{B}^o(\cdot)$.

In both cases, the only stochastic component determining the null limit distribution is the Brownian bridge. The deterministic component $M$ is a constant matrix that depends on the border values of the data groups, which are known, and the parameter value $\theta_o = \theta_*$ 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 $\widehat{\theta}_n$ can be represented asymptotically as a linear functional of the same Brownian bridge. If another estimator is used, the limit distribution typically 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 $\mathcal{B}^o(\cdot)$ is the only stochastic part of our limiting KS test, 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 $M\Sigma M$. Both $M$ and $\Sigma$ involve $\theta_*$ with $Z_j := \nabla_\theta F_j(\theta_*)$ and $p_j := F_j(\theta_*)$, so that replacing $\theta_*$ with $\widehat{\theta}_n$ we have the consistent estimates $\widehat{Z}_j := \nabla_\theta F_j(\widehat{\theta}_n) \overset{a.s.}{\to} Z_j$ and $F_j(\widehat{\theta}_n) \overset{a.s.}{\to} F_j(\theta_*)$ since $\widehat{\theta}_n \overset{a.s.}{\to} \theta_*$. Then $\widehat{Z} := [\widehat{Z}_1, \ldots, \widehat{Z}_k]' \overset{a.s.}{\to} Z$, $\widehat{M} := I - \widehat{Z}(\widehat{Z}'\widehat{Z})^{-1}\widehat{Z}' \overset{a.s.}{\to} M$, and $\widehat{\Sigma}_n \overset{a.s.}{\to} \Sigma$, where $\widehat{\Sigma}_n$ is a $k \times k$ matrix whose $i$-th row and $j$-th column element is $F_i(\widehat{\theta}_n)(1 - F_j(\widehat{\theta}_n))$ when $F_i(\widehat{\theta}_n) \le F_j(\widehat{\theta}_n)$. The distribution of $\max \|G\|_{\max}$ can be approximated by that of $\|\widehat{G}\|_{\max}$, where $\widehat{G} \sim_d N[0, \widehat{M}\widehat{\Sigma}\widehat{M}]$.

For a fixed alternative, we necessarily have $y_j := p_j - F(c_j, \theta_o) \ne 0$ for some $j$. We also note that $\sqrt{n}\{[\widehat{p}_{n,j} - F_j(\widehat{\theta}_n)] = \sqrt{n}[\widehat{p}_{n,j} - p_j] + \sqrt{n}[p_j - F_j(\theta_o) + \{F_j(\theta_o) - F_j(\widehat{\theta}_n)\}]$, and $\{F_j(\theta_o) - F_j(\widehat{\theta}_n)\} = -Z_j'(\widehat{\theta}_n - \theta_o) + o_{\mathbb{P}}(n^{-1/2}) = -Z_j'(Z'Z)^{-1}Z'(\widehat{Y} - Y) + o_{\mathbb{P}}(n^{-1/2})$. It then follows that $\sqrt{n}\{[\widehat{p}_{n,j} - F_j(\widehat{\theta}_n)] = \sqrt{n}[\widehat{p}_{n,j} - p_j] + \sqrt{n}[y_j - \{Z_j'(Z'Z)^{-1}Z'(\widehat{Y} - Y) + o_{\mathbb{P}}(n^{-1/2})\}] = O_{\mathbb{P}}(n^{1/2})$, since $\sqrt{n}[\widehat{p}_{n,j} - p_j] = O_{\mathbb{P}}(1)$ and $\sqrt{n}(\widehat{Y} - Y) = O_{\mathbb{P}}(1)$, whereas $y_j - Z_j'(Z'Z)^{-1}Z'Y$ is the $j$-th element of $MY$, and at least one of the elements in $MY$ is different from zero. This follows because the first-order condition for $\theta_o$ implies that $Z'Y = 0$, so that $MY = Y$. Therefore, the $j$-th element of $MY$ is necessarily $y_j = p_j - F_j(\theta_o) \ne 0$. Then, $\widehat{T}_n = \sqrt{n}\|Y\|_{\max} + O_{\mathbb{P}}(1) = O_{\mathbb{P}}(n^{1/2})$, and the KS test is consistent under any fixed alternative for which $y_j = p_j - F_j(\theta_o) \ne 0$ for some $j$.

This also implies that the limit distribution can be obtained by removing $\sqrt{n}y_j$ from $\sqrt{n}[\widehat{p}_{n,j} - F_j(\widehat{\theta}_n)]$ for every $j$. In more detail, observe that $Z'Y = 0$ from the first-order condition for $\theta_o$, and upon recentering on $y_j$, we have $\sqrt{n}\{[\widehat{p}_{n,j} - F_j(\widehat{\theta}_n)] - y_j\} = \sqrt{n}[\{\widehat{p}_{n,j} - p_j\} - \{Z_j'(Z'Z)^{-1}Z'\widehat{Y} + o_{\mathbb{P}}(n^{-1/2})\}] \Rightarrow g_j$, where $g_j$ is the $j$-th element of $G := MW$, and this gives the limit distribution under fixed alternatives.

We next consider the limit behavior of the KS test under the following local alternative $\mathcal{H}_\ell$: $\mathbb{P}(X_i \le c_j \mid c_0 \le X_i \le c_k) = F_j(\theta_o) + h(c_j)/\sqrt{n}$ for some uniformly bounded function $h(\cdot)$. Under this local alternative, we have $y_j = p_j - F_j(\theta_o) = h(c_j)/\sqrt{n}$, and $Y = n^{-1/2}H$, where $H := [h(c_1), \ldots, h(c_k)]'$, so that $\sqrt{n}\{\widehat{p}_{n,j} - F_j(\widehat{\theta}_n)\} \Rightarrow h_j - Z_j'(Z'Z)^{-1}Z'H + g_j$ under the local alternative by the fact that $\sqrt{n}\{[\widehat{p}_{n,j} - F_j(\widehat{\theta}_n)] - y_j\} \Rightarrow g_j$, where $h_j$ is the $j$-th element of $H$. Evidently the component $h_j - Z_j'(Z'Z)^{-1}Z'H$ is the $j$-th element of $MH$, so that we have $\widehat{S}_n \Rightarrow M(H + W) \sim N(MH, M\Sigma M)$, which reduces to the null limit theory when $MH = 0$. Similar to the null case, the limit theory when $\theta_o$ is not estimated is given by $S_n \Rightarrow H + W \sim N(H, \Sigma)$ under $\mathcal{H}_\ell$.

Defining $\xi_j := h(c_j) - Z_j'(Z'Z)^{-1}Z'H$ and using the fact noted above that $Y = n^{-1/2}H$, it is apparent that $\xi_j = h(c_j)$ because $Z'Y = n^{-1/2}Z'H = 0$ from first-order conditions for $\theta_o$. It also trivially follows that $MH$ is necessarily different from $0$ whenever $h_j \ne 0$ for some $j$, thereby ensuring that the local alternative differs from the null. Further, the KS test has the following limit $\widehat{T}_n \Rightarrow \|H + G\|_{\max}$. Thus,

tests based on $\widehat{T}_n$ have non-negligible power under $\mathcal{H}_\ell$. 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 $(H + W)$ rather than $M(H + W)$. It follows that estimation of the parameter $\theta$ using MCMD modifies the limit distribution of $T_n$ by scaling the components entering the weak limit of $\widehat{T}_n$ with the projector $M = I - Z(Z'Z)^{-1}Z'$.

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

**Theorem 3.** *Given Assumption A, (i) $\widehat{T}_n \Rightarrow \|G\|_{\max}$ under $\mathcal{H}_0$; (ii) $\widehat{T}_n = \sqrt{n}\|Y\|_{\max} + O_{\mathbb{P}}(1)$ under $\mathcal{H}_1$; and (iii) if $\sup_{j\leq k} |h(c_j)| < \infty$, $\widehat{T}_n \Rightarrow \|H + G\|_{\max}$ under $\mathcal{H}_\ell$.* □

Two methods are available to compute critical values of the test. The first method is to estimate the idempotent matrix $M$ and $\Sigma$ by a plug-in method using $\widehat{M} := I - \widehat{Z}(\widehat{Z}'\widehat{Z})^{-1}\widehat{Z}'$ and $\widehat{\Sigma}$, as mentioned above. This method produces valid critical values asymptotically by virtue of the invariance principle, consistency of $\widehat{\theta}_n$ under the null, and continuous mapping. In practice, the process $\mathcal{B}^o(\cdot)$ can be evaluated on the unit interval at the points $F_j(\widehat{\theta}_n)$ and the functional $\widehat{M}[\mathcal{B}^o(F_1(\widehat{\theta}_n)), \ldots, \mathcal{B}^o(F_k(\widehat{\theta}_n))]'$ 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 $F(\cdot, \widehat{\theta}_n)$ and computing the null distribution by iteratively replicating the test many times.

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, 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   Testing using Continuous Data

This section extends the analysis to the KS test formed with continuously distributed data. We exploit the large sample weak limit theory of the KS test given in Section 2 by using sequential asymptotics in which large sample 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 limit of the KS test for continuously distributed data.

For convenience of notation, we distinguish asymptotics in which $k$ tends to infinity from those in which $n$ 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.

10

## 3.1 Estimation Limit Theory with Continuous Data

We develop a large group 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, we note that group interval distances can be possibly different for different groups, and it may not be easy to handle this structure for the goal. So, we first reorganize the structure for technical convenience by letting $\omega$ be a new distance of every group such that for each $j \in \{1, 2, \ldots, k\}$, $n_j \cdot \omega = \omega_j$ for some $n_j \in \mathbb{N}$. We also let $m_k := \sum_{j=1}^{k} n_j$, so that $\omega \cdot m_k \equiv u - b$, and further let $d_j$ be the upper bound of the new $j$-th group, where $j = 1, 2, \ldots, m_k$ such that $d_0 = b$ and $d_{m_k} = u$. As $k$ is finite, it is not hard to select $\omega$ to satisfy the given condition. This new structure is considered only for technical convenience and could be irrelevant to the original data. Even so, the limit behaviors of the KS test can be easily obtained from this, as $k$ tends to infinity. Note that if $k$ tends to infinity, $m_k$ also tends to infinity but $\omega$ tends to zero by construction. Next, let $\bar{F}(\cdot, \theta_o) := F((1 - (\cdot))b + (\cdot)u, \theta_o)$, which is defined on the unit interval, and for $j = 1, \ldots, d$, and $i = 0, 1$, define $\partial_j^i \bar{F}_k(x, \theta_o) := \partial_j^i F(c(x), \theta_o)$, where $c(x) := \max\{c_j : (1 - x)b + xu \leq c_j, j = 0, 1, \ldots, k\}$. Note that $\partial_j^i \bar{F}_k(\cdot, \theta_o)$ is càdlàg. As $k$ tends to infinity with the distance between $c_0$ and $c_k$ being constant, $\partial_j^i \bar{F}_k(\cdot, \theta_o)$ converges uniformly to $\partial_j^i \bar{F}(\cdot, \theta_o)$, provided that $\partial_j^i \bar{F}(\cdot, \theta_o)$ is continuous on $[0, 1]$. Therefore, the large group limit of the $i$-row and $j$-column element of $A_k := m_k^{-1} Z'Z$ is obtained as $m_k^{-1} \sum_{\ell=1}^{k} \partial_i F(c_\ell, \theta_o) \partial_j F(c_\ell, \theta_o) = m_k^{-1} \sum_{\ell=1}^{m_k} \partial_i F(c(d_\ell), \theta_o) \partial_j F(c(d_\ell), \theta_o) = \int_0^1 \partial_i \bar{F}_k(x, \theta_o) \partial_j \bar{F}_k(x, \theta_o) dx \overset{k}{\to} \int_0^1 \partial_i \bar{F}(x, \theta_o) \partial_j \bar{F}(x, \theta_o) dx$, which holds by monotone convergence. If we further let $\nabla_\theta \bar{F}(\cdot, \theta_o) := [\partial_1 \bar{F}(\cdot, \theta_o), \ldots, \partial_d \bar{F}(\cdot, \theta_o)]'$ and $\nabla_\theta \bar{F}_k(\cdot, \theta_o) := [\partial_1 \bar{F}_k(\cdot, \theta_o), \ldots, \partial_d \bar{F}_k(\cdot, \theta_o)]'$, we obtain $A_k = \int_0^1 \nabla_\theta \bar{F}_k(x, \theta_o) \nabla_\theta' \bar{F}_k(x, \theta_o) dx \overset{k}{\to} A_o := \int_0^1 \nabla_\theta \bar{F}(x, \theta_o) \nabla_\theta' \bar{F}(x, \theta_o) dx$. Next, we examine the large group limit of $m_k^{-1} Z'W$. Let $\bar{p}(\cdot) := p((1 - (\cdot))b + (\cdot)u)$ and set $\bar{p}_k(x) := p(c(x))$. 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 set $\bar{\mathcal{B}}_k^o(\cdot) := \mathcal{B}^o(\bar{p}_k(\cdot))$. 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 $U_k := m_k^{-1} Z'W = \int_0^1 \nabla_\theta \bar{F}_k(x, \theta_o) \bar{\mathcal{B}}_k^o(x) dx \overset{k}{\to} U := \int_0^1 \nabla_\theta \bar{F}(x, \theta_o) \bar{\mathcal{B}}^o(x) dx$ with probability 1, which implies that $U_k \overset{k}{\Rightarrow} U$. The large group weak limit is therefore a normally distributed random variable $U \sim N(0, B_o)$, with covariance matrix $B_o := \int_0^1 \int_0^{x'} \bar{p}(x)(1 - \bar{p}(x')) \nabla_\theta \bar{F}(x, \theta_o) \nabla_\theta' \bar{F}(x', \theta_o) dx dx' + \int_0^1 \int_{x'}^1 \bar{p}(x')(1 - \bar{p}(x)) \nabla_\theta \bar{F}(x, \theta_o) \nabla_\theta' \bar{F}(x', \theta_o) dx dx'$.

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

**Assumption B.** *(i) $u - b$ is constant; (ii) For each $\theta \in \Theta$ and $j$, $F(\cdot, \theta)$ and $\partial_j F(\cdot, \theta)$ are continuous on*

$[b, u]$; (iii) $p(\cdot)$ *is continuous on* $[b, u]$; (iv) $\int_0^1 \nabla_\theta \bar{F}(x, \theta_o) \nabla'_\theta \bar{F}(x, \theta_o) dx$ *is finite and positive definite; and* (v) $B_o$ *is finite and positive definite.*  □

Although the interval distances are not identical, the desired integrals are well defined as long as the maximum interval distance tends to zero. 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 $A_o$, $B_o$, and $U$ are all well defined as $k$ tends to infinity. Assumption B(*iii*) is redundant under the null, because $p(\cdot) = F(\cdot, \theta_*)$ and $\theta_* = \theta_0$, so that (*ii*) implies (*iii*). Assumptions B(*iv*) and (*v*) are standard conditions that ensure the sequential limit distribution of the KS test behaves regularly.

The stated results are collected in the following lemma.

**Lemma 1.** *Given the Assumptions A and B, (i)* $m_k^{-1} Z'Z \overset{k}{\to} A_o := \int_0^1 \nabla_\theta \bar{F}(x, \theta_o) \nabla'_\theta \bar{F}(x, \theta_o) dx$*; and (ii)* $m_k^{-1} Z'W \overset{k}{\Rightarrow} U := \int_0^1 \nabla_\theta \bar{F}(x, \theta_o) \bar{\mathcal{B}}^o(x) dx \sim N(0, B_o)$.  □

Some remarks are warranted. First, from the first-order condition for $\theta_o$, $Z'Y = 0$ holds uniformly in $k$. Furthermore, $m_k^{-1} Z'Y = \int_0^1 \nabla_\theta \bar{F}_k(x, \theta_o)[\bar{p}_k(x) - \bar{F}_k(x, \theta_o)] dx \overset{k}{\to} \int_0^1 \nabla_\theta \bar{F}(x, \theta_o)[\bar{p}(x) - \bar{F}(x, \theta_o)] dx$, therefore implying that $\int_0^1 \nabla_\theta \bar{F}(x, \theta_o) \bar{p}(x) dx = \int_0^1 \nabla_\theta \bar{F}(x, \theta_o) \bar{F}(x, \theta_o) dx$. Second, Lemma 1 implies a straightforward large group weak limit for $\widehat{\theta}_n$. The following theorem trivially holds by joint convergence.

**Theorem 4.** *Given Assumptions A and B,* $\sqrt{n}(\widehat{\theta} - \theta_o) \overset{n}{\Rightarrow} (Z'Z)^{-1} Z'W \overset{k}{\Rightarrow} A_o^{-1} U$.  □

Note that $A_o^{-1} U \sim N(0, A_o^{-1} B_o A_o^{-1})$ that 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 shown in the next subsection, the same mode of analysis suggests a way to obtain asymptotic 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 Continuous Data

Using the large group weak limit result for the MCMD estimator given in Theorem 4, we examine the large group 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.

We first examine the large group null limit distribution of the KS test statistic. We start the discussion from the asymptotic null distribution. Defining $\bar{\mathcal{G}}_k^o(\cdot) := \bar{\mathcal{B}}_k^o(\cdot) - \nabla_\theta' \bar{F}_k(\cdot, \theta_o) A_k^{-1} U_k$ and $\bar{\mathcal{G}}^o(\cdot) := \bar{\mathcal{B}}^o(\cdot) - \nabla_\theta' \bar{F}(\cdot, \theta_o) A_o^{-1} U$, then $\bar{\mathcal{G}}_k^o(\cdot) \overset{k}{\to} \bar{\mathcal{G}}^o(\cdot)$ uniformly with probability 1, because $\bar{F}_k(\cdot, \theta_o) \overset{k}{\to} \bar{F}(\cdot, \theta_o)$ 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 $i$ (resp. $j$), there is $j$ (resp. $i$) such that $\mathcal{B}^o(p_i) - Z_i'(Z'Z)^{-1} Z'W = \bar{\mathcal{G}}_k^o(j/m_k)$ by the construction of $\bar{\mathcal{G}}_k^o(\cdot)$, so that $\sup_{j \le m_k} |\bar{\mathcal{G}}_k^o(j/m_k)| \overset{k}{\to} \sup_{z \in [0,1]} |\bar{\mathcal{G}}^o(z)|$ with probability 1, viz., the large group null weak limit is obtained as a functional of $\bar{\mathcal{G}}^o(\cdot)$ such that for each $z$, $\mathbb{E}[\bar{\mathcal{G}}^o(z)] = 0$, and if $z \le \tilde{z}$, $\mathbb{E}[\bar{\mathcal{G}}^o(z)\bar{\mathcal{G}}^o(\tilde{z})] = \bar{p}(z)(1 - \bar{p}(\tilde{z})) - \nabla_\theta' \bar{F}(z, \theta_o) A_o^{-1} D(\tilde{z}) - \nabla_\theta' \bar{F}(\tilde{z}, \theta_o) A_o^{-1} D(z) + \nabla_\theta' \bar{F}(z, \theta_o) A_o^{-1} B_o A_o^{-1} \nabla_\theta \bar{F}(\tilde{z}, \theta_o)$, where for each $z \in [0,1]$, $D(z) := (1 - \bar{p}(z)) \int_0^z \bar{p}(x) \nabla_\theta \bar{F}(x, \theta_o) dx + \bar{p}(z) \int_z^1 (1 - \bar{p}(x)) \nabla_\theta \bar{F}(x, \theta_o) dx$. 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.

We next examine the large group 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 define $\bar{h}_k(x) := h(c(x))$. 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, $Q_k := m_k^{-1} Z'H = \int_0^1 \nabla_\theta \bar{F}_k(x, \theta_o) \bar{h}_k(x) dx \overset{k}{\to} Q := \int_0^1 \nabla_\theta \bar{F}(x, \theta_o) \bar{h}(x) dx$. Note that $\bar{p}(\cdot) - \bar{F}(\cdot, \theta_o) = \bar{h}(\cdot)/\sqrt{n}$, so that $Q$ is proportional to $\int_0^1 \nabla_\theta \bar{F}(x, \theta_o) (\bar{p}(x) - \bar{F}(x, \theta_o)) dx$ that is the asymptotic limit of $m_k^{-1} Z'Y$ as $k \to \infty$. Furthermore, $Z'Y = 0$. Therefore, $Q = 0$, and if we let $\bar{\xi}_k(\cdot) := \bar{h}_k(\cdot) - \nabla_\theta' \bar{F}_k(\cdot, \theta_o) A_k^{-1} Q_k$, for each $j$, $\xi_j = \bar{h}_k(j/m_k) - \nabla_\theta' \bar{F}_k(j/m_k, \theta_o) A_k^{-1} Q_k$ and $\bar{\xi}_k(\cdot) \overset{k}{\to} \bar{\xi}(\cdot) := \bar{h}(\cdot) - \nabla_\theta' \bar{F}(\cdot, \theta_o) A^{-1} Q \equiv \bar{h}(\cdot)$ 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, $\sup_{j \le m_k} |\bar{\xi}_k(j/m_k) + \bar{\mathcal{G}}_k^o(j/m_k)| \overset{k}{\to} \sup_{z \in [0,1]}, |\bar{\xi}(z) + \bar{\mathcal{G}}^o(z)|$ with probability 1, which implies that $\sup_{j \le k} |\bar{\xi}_k(j/m_k) + \bar{\mathcal{G}}_k^o(j/m_k)| \overset{k}{\Rightarrow} \sup_{z \in [0,1]} |\bar{\xi}(z) + \bar{\mathcal{G}}^o(z)|$. Thus the localizing parameter of the limit Gaussian process shifts from zero under the null to $\bar{\xi}(\cdot)$ under $\mathcal{H}_\ell$, which is identical to $\bar{h}(\cdot)$. It is therefore apparent that, if $h(\cdot) \ne 0$, the KS test statistic has non-negligible local power asymptotically.

These results are summarized in the following theorem.

**Theorem 5.** *Given Assumptions A and B, (i)* $\widehat{T}_n \overset{n}{\Rightarrow} \sup_{j \le m_k} |\bar{\mathcal{G}}_k^o(\frac{j}{m_k})| \overset{k}{\Rightarrow} \sup_{z \in [0,1]} |\bar{\mathcal{G}}^o(z)|$ *under* $\mathcal{H}_0$ ; *and (ii) if* $h(\cdot) \in \mathcal{C}([b, u])$, $\widehat{T}_n \overset{n}{\Rightarrow} \sup_{j \le m_k} |\bar{h}_k(\frac{j}{m_k}) + \bar{\mathcal{G}}_k^o(\frac{j}{m_k})| \overset{k}{\Rightarrow} \sup_{z \in [0,1]} |\bar{h}(z) + \bar{\mathcal{G}}^o(z)|$ *under* $\mathcal{H}_\ell$.  $\square$

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 the invariance principle. If we let $B_\ell^o(\cdot) :=$

$\frac{1}{\sqrt{\ell}} \sum_{j=1}^{\lceil \ell(\cdot) \rceil} u_j - \frac{(\cdot)}{\sqrt{\ell}} \sum_{j=1}^{\ell} u_j$, $B_\ell^o(\cdot) \Rightarrow \mathcal{B}^o(\cdot)$ 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 $N(0,1)$. Thus, we approximate $\bar{\mathcal{G}}^o(\cdot)$ by $\widehat{\mathcal{G}}_m^o(\cdot) := B_\ell^o(F(\cdot, \widehat{\theta}_n)) - \nabla_\theta F(\cdot, \widehat{\theta}_n) \widehat{A}_{o,m}^{-1} \widehat{U}_m$, where $\widehat{A}_{o,m} := m^{-1} \sum_{j=1}^{m} \nabla_\theta F(c_j, \widehat{\theta}_n) \nabla'_\theta F(c_j, \widehat{\theta}_n)$; and $\widehat{U}_m := m^{-1} \sum_{j=1}^{m} \nabla_\theta F(c_j, \widehat{\theta}_n) B_\ell^o(F(c_j, \widehat{\theta}_n))$. Here, $m$ is a sufficiently large number set by the researcher for the number of groups such that each group has an equal group interval, so that $c_j = b + (u-b) \times j/m$. If $m$ tends to increase, the group interval distance given by $(u-b)/m$ tends to zero, so that for sufficiently large $m$, $\widehat{A}_{o,m}$ and $\widehat{U}_m$ well approximate their corresponding limits. This is also the case for $B_\ell^o(\cdot)$ when $\ell$ is sufficiently large, so that $\widehat{\mathcal{G}}_m^o(\cdot)$ approximates well $\bar{\mathcal{G}}^o(\cdot)$ if $m$ and $\ell$ are sufficiently large. Also, note that 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.

# 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 $X_i$ is given by $\mathbb{P}(X_i \leq c) = 1 - (c_{\min}/c)^\theta$, where $c_{\min}$ is the minimal value of $X_i$, and $\theta$ is the shape parameter.

We further suppose that data are grouped such that $b$ is greater than or equal to $c_{\min}$, and $u$ is finite. This framework implies that the unconditional distribution is modified to the conditional distribution $\mathbb{P}(X_i \leq c_j \mid b \leq X_i \leq u) = 1 - \{(u/c_j)^\theta - 1\}/\{(u/b)^\theta - 1\}$. We denote this distribution as Pareto$(\theta)$ and let the right side of the equation be the model for the grouped data. That is, for each $j$, $F_j(\theta) = 1 - \{(u/c_j)^\theta - 1\}/\{(u/b)^\theta - 1\}$. In our simulations, we use the following parameter settings: the bounds are $b = 1.0$, and $u = 10.0$; for every $j$, the interval length is $c_j - c_{j-1} = \omega$ and we consider two cases for $\omega \in \{0.1, 1.0\}$. 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 under the null. First, we let $\theta_* = 2.0$ and generate $n$ observations with conditional distribution $F_j(\theta_*)$. Six sample sizes are considered: 100,

200, 400, 600, 800, and 1,000. Second, we consider two approaches to assess the adequacy of the limit theory in Theorem 3($i$), and we call these methods A and B, respectively.

Method A first estimates $M$ and $W$ using the MCMD estimator $\widehat{\theta}_n$. Specifically, for each $j$, we let $\widehat{z}_j = (u/b)^{\widehat{\theta}_n} \log(u/b)[(u/c_j)^{\widehat{\theta}_n} - 1] - (u/c_j)^{\widehat{\theta}_n} \log(u/c_j)[(u/b)^{\widehat{\theta}_n} - 1]$ be the $j$-th element of $\widehat{Z}$ and estimate $M$ by $\widehat{M} := I_k - \widehat{Z}(\widehat{Z}'\widehat{Z})^{-1}\widehat{Z}'$. Next, $\mathcal{B}^o(\cdot)$ is approximated by $B_\ell^o(\cdot)$ with $\ell = 10,000$, and we let $\widehat{\mathfrak{B}}^o := [\ldots, B_\ell^o(F_j(\widehat{\theta}_n)), \ldots]'$, and compare $\widehat{T}_n$ with the asymptotic critical values implied by $\widehat{M}\widehat{\mathfrak{B}}^o$. The null distribution is obtained by independently generating $\widehat{M}\widehat{\mathfrak{B}}^o$ 200 times, and we iterate the whole process 5,000 times. Method B implements the parametric bootstrap in the same way.

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: $\widetilde{T}_n := \sup_{j \leq k} |\sqrt{n}\{\widehat{p}_{n,j} - F_j(\widetilde{\theta}_n)\}|$, where $\widetilde{\theta}_n$ denotes the (Q)ML estimator. For method A, we iteratively simulate $\max[\ldots, |B_\ell^o(F_j(\widetilde{\theta}_n))|, \ldots]$ 200 times and obtain the corresponding critical values. Third, we apply the parametric bootstrap. Finally, Khmaladze's (2013) distribution-free test is applied. Specifically, $\theta_*$ is estimated by (Q)ML, and the following KS test statistic is computed based upon $\ddot{T}_n := \max_{s \leq k} |\sum_{j=1}^s \widetilde{z}_{n,j}|$, where $\widetilde{z}_{n,j}$ is the $j$-th-row element of $\widetilde{Z}_n := \widetilde{\widehat{Y}}_n - \widetilde{\widehat{Y}}_n' A_3(A_3 - B_3) - \widetilde{\widehat{Y}}_n' A_4(A_4 - B_4)$ with $\widetilde{\widehat{Y}}_n := [\widetilde{y}_{n,1}, \ldots, \widetilde{y}_{n,k}]'$, $\widetilde{y}_{n,j} := \{\#\{X_i \in [c_j, c_{j+1})\} - n\widetilde{c}_j\}/\sqrt{n\widetilde{c}_j}$, and $\widetilde{c}_j := F_j(\widetilde{\theta}_n) - F_{j-1}(\widetilde{\theta}_n)$; $A_3 := \widetilde{A}_3/(\widetilde{A}_3'\widetilde{A}_3)^{1/2}$ with $\widetilde{A}_3 := R - (R'\widetilde{Q})\widetilde{Q} - (R'\widehat{Q})\widehat{Q}$, $R := [1, 0, \ldots, 0]'$, $\widetilde{Q} := [\sqrt{\widetilde{c}_1}, \ldots, \sqrt{\widetilde{c}_k}]'$, and $\widehat{Q} := \ddot{Q}/(\ddot{Q}'\ddot{Q})^{1/2}$ with $\ddot{Q} := [\widetilde{d}_1/\sqrt{\widetilde{c}_1}, \ldots, \widetilde{d}_k/\sqrt{\widetilde{c}_k}]'$ and $\widetilde{d}_j := (\partial/\partial\theta)F(c_j, \widetilde{\theta}_n) - (\partial/\partial\theta)F(c_{j-1}, \widetilde{\theta}_n)$; $A_4 := \widetilde{A}_4/(\widetilde{A}_4'\widetilde{A}_4)^{1/2}$ with $\widetilde{A}_4 := \widehat{R} - (\widehat{R}'\widetilde{Q})\widetilde{Q} - (\widehat{R}'\widehat{Q})\widehat{Q} - (\widehat{R}'A_3)A_3$ and $\widehat{R} := [0, 1, 0, \ldots, 0]'$; $B_3 := \widetilde{B}_3/(\widetilde{B}_3'\widetilde{B}_3)^{1/2}$ with $\widetilde{B}_3 := \widetilde{Q} - (\widetilde{Q}'R)R - (\widetilde{Q}'\widehat{R})\widehat{R}$; and $B_4 := \widetilde{B}_4/(\widetilde{B}_4'\widetilde{B}_4)^{1/2}$ with $\widetilde{B}_4 := \widehat{Q} - (\widehat{Q}'R)R - (\widehat{Q}'\widehat{R})\widehat{R} - (\widehat{Q}B_3)B_3$. 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 C.

Tables 1 contains the empirical rejection rates of $\widehat{T}_n$ and $\widetilde{T}_n$. The simulation results can be summarized as follows: (i) The simulation results for $\widehat{T}_n$ generally well support the theory given in Theorem 3($i$). The nominal rejection rates of $\widehat{T}_n$ are consistently well estimated by the empirical rejection rates, and more precise empirical rejection rates are obtained as $n$ increases. (ii) $\widetilde{T}_n$ shows results that are very different. As pointed out by Durbin (1973), the KS test statistic with a plug-in ML estimator has significant level distortions that persist even when $n$ is large. These distortions occur mainly because $\widetilde{T}_n$ has an asymptotic distribution that is affected by the ML estimator. Method A therefore yields substantial level distortions in this case, and they are relieved by using method B, which accommodates the parameter estimation error

and has the same asymptotic null distribution as that of $\widetilde{T}_n$. Method C removes the parameter estimation error from the test basis, and $\ddot{T}_n$ becomes distribution free. (iii) There is a tendency for the empirical rejection rates of $\widehat{T}_n$ to be closer to the nominal levels when $\omega$ is small. (iv) Applying the asymptotic null distribution directly to the test yields more precise empirical rejection rates than applying methods B and C. These results indicate that $\widehat{T}_n$ performs best under the null when it is constructed by data observations grouped into small intervals and compared with the asymptotic null distribution.

Some additional remarks are in order. First, the asymptotic null limit distribution is different for different $\omega$. Figure 1(a) shows the null limit distributions for various $\omega$'s. They are obtained using method A with $\widehat{\theta}_n$ being replaced by $\theta_*$. Note that the null limit distribution converges as $\omega$ tends to zero. Second, we examine the methodology given below Theorem 5 to test the distributional hypothesis of continuously distributed $X_i$. By letting $\ell = 10,000$ and $m = 2n$, we draw the percentile-percentile (PP) plots between the level of significance and the $p$-value for $n = 100, 200, 400, 600,$ and $800$. The simulation environments are identical to those for Table 1. Note that the resulting PP-plots shown in Figure 1(b) are close to the 45-degree line even when the sample size is as small as 100, affirming the claims in Theorem 5.

## 4.2 Testing under the Alternative

We now examine test power. For this purpose, we change the distribution of $X_i$ from Pareto to the following exponential distribution as the generating mechanism: $\mathbb{P}(X_i \leq x | b \leq X_i \leq u) = \{1 - \exp(-\lambda_*(x - b))\}/\{1 - \exp(-\lambda_*(u - b))\}$. We denote this distribution $\mathrm{Exp}(\lambda_*)$. We group the observations from $\mathrm{Exp}(1.2)$ in the same way as in Section 4.1 and test the Pareto distributional assumption as before.

The empirical rejection rates of $\widehat{T}_n$ and $(\widetilde{T}_n, \ddot{T}_n)$ are contained in Table 2. The results can be summarized as follows: (i) 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 A, B, and C. (ii) The empirical rejection rates of $\widetilde{T}_n$ using method A are uniformly dominated by $\widehat{T}_n$ using methods A and B. This is mainly because the asymptotic critical values of $\widetilde{T}_n$ implemented by method A are too large, as evidenced in the substantial level distortions under the null seen in Table 1. (iii) The overall power of $\widetilde{T}_n$ when the test is implemented by method B is similar to that of $\widehat{T}_n$ implemented by methods A or B and always dominate that of $\ddot{T}_n$ implemented by method C. (iv) The empirical rejection rates of $\widehat{T}_n$ implemented by method A are close to those of method B. 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. (v) When the sample size is small, the power of $\widetilde{T}_n$ implemented by method B is slightly higher than that of $\widehat{T}_n$ implemented by methods A or B, 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 $Z_i \sim \mathrm{Exp}(\lambda_*)$ and $W_i \sim \mathrm{Pareto}(2.0)$, we let $X_i = \frac{5}{\sqrt{n}} Z_i + (1 - \frac{5}{\sqrt{n}}) W_i$, so that $X_i$ is a mixture of Pareto and exponential random variables for which the mixture distribution of $X_i$ converges to the Pareto distribution at an $n^{-1/2}$ convergence rate. For $\lambda_* = 0.8$, $1.0$, $1.2$, $1.4$, and $1.6$, we test the Pareto distributional assumption using methods A, B, and C.

The simulation results of $\widehat{T}_n$ and $(\widetilde{T}_n, \ddot{T}_n)$ are contained in Table 3. We let $n = 500$ and summarize the results as follows: (i) For every $\lambda_*$, the empirical rejection rates exceed nominal size except for the test $\widetilde{T}_n$ implemented by method A 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 A or B (resp. method C) is applied, but $\widetilde{T}_n$ has nontrivial powers only when method B is applied. (ii) Local power of $\widetilde{T}_n$ is not given for method A in many cases because the critical values of $\widetilde{T}_n$ exceed the upper bound and test size is zero as seen in Table 1. (iii) Methods A and B have similar power patterns for the test $\widehat{T}_n$. We deduce from these results that the performance of methods A and B are similar under local alternatives. (iv) The overall empirical rejection rates of $\widehat{T}_n$ are similar to those of $\widetilde{T}_n$ when that test is implemented by method B, implying that we can expect similar local powers from $\widehat{T}_n$ and $\widetilde{T}_n$ when using parametric bootstrap methods. (v) The local power of $\ddot{T}_n$ implemented by method C is almost uniformly dominated by that of $\widehat{T}_n$ implemented by method A, although the local power difference decreases as $\lambda_*$ increases.

## 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

17

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 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 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 Supplement.

*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 (2015). 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, tests of the Pareto distributional assumption are better suited to the methodology when the number of groups 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 1 of the Supplement to this paper provides summary statistics of the income tax return data used herein. Several features stand out. The most noticeable feature of the data is the number of groups. 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 number of groups 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 $[0.0, \text{KRW}50 \text{ mil.})$ and $[\text{KRW}39,910 \text{ mil.}, \infty)$. For the other groups, the data are provided in the same format with each group interval width being KRW10 mil. By contrast the conventional income tax data have irregular group patterns. A second important feature 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 Piketty 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 Supplement.

Finally, we obtain population data in Korea. Various population data have been used in the prior literature. For example, Piketty 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, and we call this measure the *labor force*. These two populations measures correspond with population measures used in studies of other countries such as the UK and Japan in Atkinson (2005) and Moriguchi and Saez (2008).

## 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: (i) 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 $[c_{\sharp-1}, c_{\sharp})$ denote this group. Note that $c_{\sharp} - c_{\sharp-1}$ is KRW10 million for our data sets. (ii) We test the Pareto distributional assumption for the grouped data. We choose $b$ and $u$ so that $b \leq c_{\sharp-1}$ and $u \geq c_{\sharp}$ and estimate $\theta_*$ 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. (iii) 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 $c_{\dagger} := F^{-1}(q, \widehat{\theta}_n)$, 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 $c_{\dagger} \in [c_{\sharp-1}, c_{\sharp})$, we let $\widehat{x}_n$ be $c_{\dagger}$; if $c_{\dagger} > c_{\sharp}$, let $\widehat{x}_n$ be the upper bound $c_{\sharp}$ of the interval; otherwise, let $\widehat{x}_n$ be $c_{\sharp-1}$. This additional restriction is imposed because $\widehat{x}_n$ must lie between $c_{\sharp-1}$ and $c_{\sharp}$ by virtue of the first-step requirement. (iv) We finally compute the top $x\%$ share of incomes. We first estimate the total income greater than $\widehat{x}_n$ by $\widehat{m}_n := (\{F(c_{\sharp}, \widehat{\theta}_n) - F(\widehat{x}_n, \widehat{\theta}_n)\}/\{F(c_{\sharp}, \widehat{\theta}_n) - F(c_{\sharp-1}, \widehat{\theta}_n)\}) \times I_{\sharp} + \sum_{j=\sharp+1}^{k} I_j$, where $I_j$ denotes the total income in the group of $[c_{j-1}, c_j)$, and $k$ is the number of groups 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 condition is tested in Step (ii). Even if the null is rejected, we proceed to Step (iii) 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 Atkinson's (2005) mean-split histogram method, which 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. According to Atkinson, Piketty, and Saez (2011), top income shares are estimated by this method for many countries such as Australia, New Zealand, Norway, and UK. Second, when implementing Step (ii), $b$ and $u$ have to be selected in such a way that the interval $[c_{\sharp-1}, c_{\sharp})$ 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 results turn out to be insensitive to the selection of $b$ and $u$.

The top $x\%$ income levels are estimated and contained in Tables 4. We summarize the key properties of our estimates as follows: (i) 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 in Table 1 of the Supplement is one of these trials. This shows that the Pareto assumption is hard to accept as holding for estimation the top 1.0% income. (ii) Although the results are not reported in the tables, even for estimating the top 0.5% income, the Pareto assumption does not hold for every year in the sample data. (iii) When the top 0.10%, 0.05%, or 0.01% and higher incomes are estimated, we could not reject the Pareto hypothesis. 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 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 assumption to be properly exploited, at least the top 0.10% and higher income shares need to be estimated. (iv) The estimated $x$% top income levels ($\widehat{x}_n$) are between $[c_{\sharp-1}, c_\sharp)$ for most cases. Sometimes, the preliminary estimates of the top income levels ($c_\dagger$) are greater than the presumed border value $c_\sharp$. For such cases, we let $\widehat{x}_n$ be $c_\sharp$ as required in Step (iii). We added the superscript '$\sharp$' to the figures in Table 4 to indicate such an occurrence. The preliminary estimates of the top income levels ($c_\dagger$) are not substantially different from the boundary values ($c_\sharp$) for every case.

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 Figure 2, whose numerical values are provided in the Supplement. We summarize the findings as follows: (i) Figure 2 can be compared with those obtained by Atkinson's (2005) mean-split histogram method that are provided by Park and Jeon (2014) for the same data. They 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. (ii) We also compare our findings with those of Kim and Kim (2015; 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 the differences, the KK estimates are similar to our own, with the greatest difference being 0.69% points, 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. (iii) 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 (2015), 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. (iv) 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, although they do show an overall increasing pattern from 2000, the top income shares of most other countries do not show definitely rising tendencies since 2007, based on available estimates. The world top income database provides the top income shares of 27 countries that are reported in the literature, as reviewed in Atkinson, Piketty, and Saez (2011). For example, countries such as Canada, Netherlands, and UK show declining patterns, and this is believed to be due to the global financial crisis. 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 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 when estimating the top 1.0% or 0.5% income shares for every year in the data; but for tail observations lying in the top 0.10%, 0.05%, or 0.01% and higher income shares, 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 over time in Korea, in line with more global observations on income shares.

# Acknowledgements

# Supplementary Materials

**Title: Supplement to "Practical Kolmogorov-Smirnov Testing by Minimum Distance Applied to Measure Top Income Shares in Korea."** This supplement provides proofs of the results stated in the paper and data descriptions for the empirical work reported here. (PDF type)

# References

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

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.

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

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.

Durbin, J. (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 *Tax Policy and the Economy*, Vol. 7. ed. Poterba, J., Cambridge, MA: MIT Press.

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. (2015), "Income Inequality in Korea, 1933-2010: Evidence from Income Tax Statistics," *Hitotsubashi Journal of Economics*, 56, 1–19.

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

Kuznets, S. (1955), "Economic Growth and Economic Inequality," *American Economic Review*, 45, 1–28.

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 *Top Incomes: A Global Perspective*, eds. Atkinson, Anthony B. and Piketty, Thomas, Oxford, UK: Oxford University Press. Series updated by Facundo Alvaredo, Chiaki Moriguchi and Emmanuel Saez (2012, Methodological Notes).

Park, M., and Jeon, B. (2014), "Changes in Income Allocations and Policy Suggestions," Research Report 14-02, Sejong, Korea: Korea Institute of Public Finance (in Korean).

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.

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

Wood, L., and Altavela, M. (1978), "Large-Sample Results for Kolmogorov-Smirnov Statistics for Discrete Distributions," *Biometrika*, 65, 235–239.
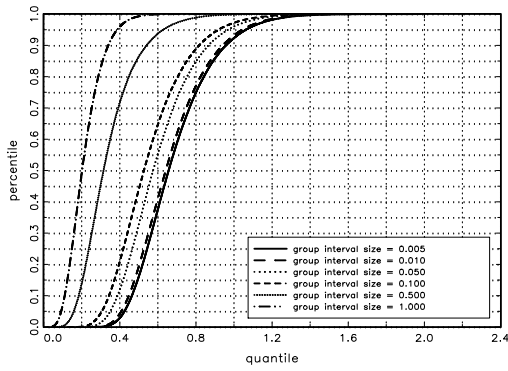
(a) Null Limit Distributions
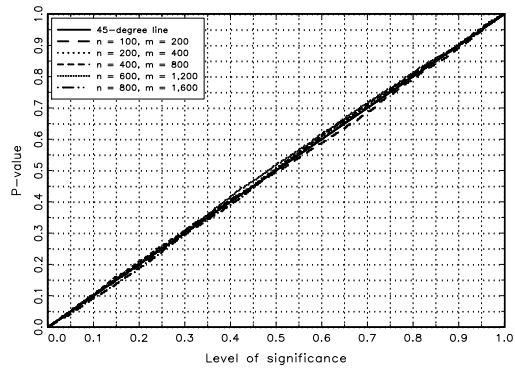


(b) PP-plots

Figure 1: Null Limit Distributions and PP-plots. Figure1(a) shows the null limit distributions for $\omega = 1.00, 0.50, 0.10, 0.05, 0.01$, and $0.005$. They are obtained by Method using the unknown $\theta_*$, and the experiment repetition is 100,000. Figure 1(b) shows the PP-plots between the level of significance and the $p$-value for continuous data that are implemented by the method below Theorem 5. Refer to Table 1 for simulation environments.



Top 0.01% Income Shares
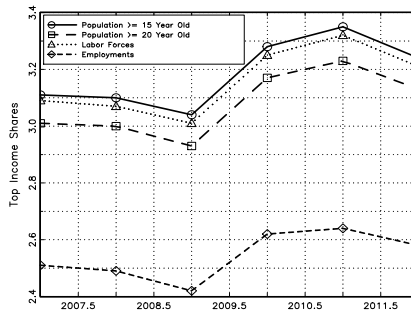


Top 0.05% Income Shares
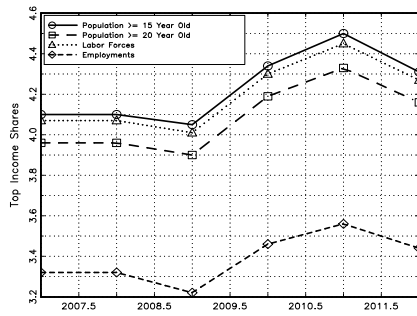


Top 0.10% Income Shares



Top 1.00% Income Shares

Figure 2: Top Income Shares over Time. The figures show the estimated top income shares for 0.01, 0.05, 0.10, and 1.00%. The four different populations are used to estimate the shares.

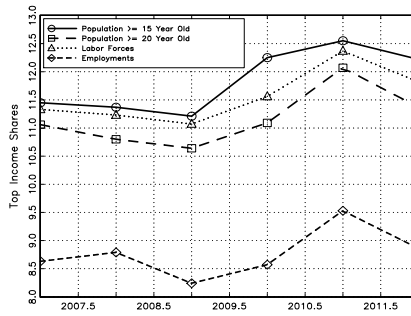| Estimators | $\omega$ | Methods | Levels $\setminus n$ | 100 | 200 | 400 | 600 | 800 | 1,000 |
|---|---|---|---|---|---|---|---|---|---|
| MCMD | 0.10 | A | 1% | 1.46 | 1.18 | 1.76 | 1.66 | 1.44 | 1.36 |
| | | | 5% | 5.56 | 5.38 | 6.40 | 5.70 | 5.24 | 5.56 |
| | | | 10% | 10.64 | 10.12 | 11.56 | 10.66 | 10.06 | 10.64 |
| | | B | 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 |
| | 1.00 | A | 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 |
| | | B | 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 |
| ML | 0.10 | A | 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 |
| | | B | 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 |
| | | C | 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 |
| | 1.00 | A | 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 |
| | | B | 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 |
| | | C | 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 1: Empirical Levels of the Test Statistic Using the MCMD and ML Estimators. Repetitions: 5,000. Bootstrap and Null Distribution Repetitions: 200. DGP: $X_i \sim \text{Pareto}(\theta_*)$; $(\theta_*) = (2.0)$; Bottom Value of Data Range ($b$): 1.00; Top Value of Data Range ($u$): 10.00; $n$ observations are grouped into $(u - b)/\omega$ number of intervals such that for each $j = 1, \ldots, k$, $c_j - c_{j-1} = \omega$. Model: for each $j = 1, 2, \ldots, k$, $F_j(\theta) = 1 - [(u/c_j)^\theta - 1]/[(u/b)^\theta - 1]$.

| Estimators | $\omega$ | Methods | Levels $\backslash$ $n$ | 100 | 200 | 400 | 600 | 800 | 1,000 |
|---|---|---|---|---|---|---|---|---|---|
| MCMD | 0.10 | A | 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 |
| | | B | 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 |
| | 1.00 | A | 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 |
| | | B | 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 |
| ML | 0.10 | A | 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 |
| | | B | 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 |
| | | C | 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 |
| | 1.00 | A | 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 |
| | | B | 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 |
| | | C | 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 2: Empirical Powers of the Test Statistic Using the MCMD and ML Estimators. Repetitions: 5,000. Bootstrap and Null Distribution Repetitions: 200. DGP: $X_i \sim \text{Exp}(\lambda_*)$; $\lambda_* = 1.2$. Refer to Table 1 for other simulation environments.

| Estimators | $\omega$ | Methods | Levels \ $\lambda_*$ | 0.80 | 1.00 | 1.20 | 1.40 | 1.60 |
|---|---|---|---|---|---|---|---|---|
| MCMD | 0.10 | A | 1% | 28.88 | 18.38 | 11.52 | 8.14 | 6.14 |
| | | | 5% | 51.74 | 35.90 | 26.36 | 19.44 | 16.16 |
| | | | 10% | 64.04 | 47.74 | 37.70 | 28.88 | 25.06 |
| | | B | 1% | 29.14 | 18.28 | 11.84 | 8.24 | 6.00 |
| | | | 5% | 50.84 | 35.58 | 26.30 | 19.76 | 15.92 |
| | | | 10% | 63.78 | 47.86 | 37.42 | 29.34 | 25.04 |
| | 1.00 | A | 1% | 15.04 | 8.40 | 6.30 | 4.18 | 2.82 |
| | | | 5% | 30.32 | 20.22 | 15.04 | 11.54 | 9.40 |
| | | | 10% | 41.16 | 30.20 | 23.40 | 19.26 | 16.32 |
| | | B | 1% | 15.00 | 8.98 | 5.90 | 4.02 | 2.60 |
| | | | 5% | 30.24 | 20.26 | 14.56 | 11.86 | 8.98 |
| | | | 10% | 40.92 | 29.42 | 23.62 | 19.24 | 16.58 |
| ML | 0.10 | A | 1% | 2.94 | 1.68 | 0.90 | 0.36 | 0.38 |
| | | | 5% | 11.96 | 7.16 | 4.34 | 2.86 | 2.16 |
| | | | 10% | 22.12 | 14.12 | 9.40 | 6.68 | 5.12 |
| | | B | 1% | 27.10 | 17.26 | 11.86 | 8.46 | 6.24 |
| | | | 5% | 47.98 | 35.58 | 26.22 | 19.88 | 15.42 |
| | | | 10% | 60.80 | 47.70 | 37.92 | 30.22 | 24.50 |
| | | C | 1% | 3.54 | 2.82 | 1.92 | 1.52 | 1.42 |
| | | | 5% | 10.24 | 9.30 | 6.82 | 6.30 | 5.60 |
| | | | 10% | 17.22 | 14.32 | 12.00 | 11.90 | 10.94 |
| | 1.00 | A | 1% | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | | | 5% | 0.28 | 0.12 | 0.00 | 0.02 | 0.00 |
| | | | 10% | 1.34 | 0.58 | 0.20 | 0.06 | 0.10 |
| | | B | 1% | 15.16 | 9.62 | 6.80 | 4.32 | 3.73 |
| | | | 5% | 31.06 | 21.84 | 15.08 | 11.86 | 10.44 |
| | | | 10% | 41.76 | 31.16 | 23.16 | 19.64 | 17.10 |
| | | C | 1% | 11.22 | 6.40 | 4.86 | 3.40 | 2.36 |
| | | | 5% | 26.72 | 17.26 | 13.60 | 10.64 | 9.06 |
| | | | 10% | 37.40 | 27.52 | 21.74 | 17.94 | 15.60 |

Table 3: Empirical Local Powers of the Test Statistic Using the MCMD and ML Estimators. Repetitions: 5,000. Bootstrap and Null Distribution Repetitions: 200. Sample Size: 500. DGP: $X_i \sim (1 - 5/\sqrt{n})\text{Pareto}(\theta_*) + (5/\sqrt{n})\text{Exp}(\lambda_*)$; $\theta_* = 2.0$. Refer to Table 1 for other simulation environments.

| Years | Top $x$% | Statistics \ Populations | $\geq$ 15 year old | $\geq$ 20 year old | measured by Labor Forces | measured by Employments |
|---|---|---|---|---|---|---|
| 2007 | 1.00% | $\widehat{x}_n$ | 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% | $\widehat{x}_n$ | 2.3370 | 2.4484 | 2.3599 | 3.1541 |
| | | $p$-value of $\widehat{T}_n$ | 45.500 | 45.500 | 45.500 | 45.000 |
| | 0.05% | $\widehat{x}_n$ | 3.4698 | 3.6512 | 3.5070 | 4.8222 |
| | | $p$-value of $\widehat{T}_n$ | 45.500 | 45.500 | 45.500 | 94.000 |
| | 0.01% | $\widehat{x}_n$ | 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% | $\widehat{x}_n$ | 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% | $\widehat{x}_n$ | 2.4325 | 2.5497 | 2.4614 | 3.2908 |
| | | $p$-value of $\widehat{T}_n$ | 100.00 | 100.00 | 100.00 | 100.00 |
| | 0.05% | $\widehat{x}_n$ | 3.5975 | 3.7852 | 3.6435 | 4.9975 |
| | | $p$-value of $\widehat{T}_n$ | 100.00 | 100.00 | 100.00 | 37.000 |
| | 0.01% | $\widehat{x}_n$ | 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% | $\widehat{x}_n$ | 0.9382 | 0.9692 | 0.9458 | 1.1609 |
| | | $p$-value of $\widehat{T}_n$ | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | 0.10% | $\widehat{x}_n$ | 2.4943 | 2.6136 | 2.5191 | 3.3987 |
| | | $p$-value of $\widehat{T}_n$ | 100.00 | 14.000 | 14.000 | 100.00 |
| | 0.05% | $\widehat{x}_n$ | 3.6794 | 3.8672 | 3.7241 | 5.0761 |
| | | $p$-value of $\widehat{T}_n$ | 14.000 | 14.000 | 14.000 | 14.000 |
| | 0.01% | $\widehat{x}_n$ | 9.4237 | 9.9125 | 9.5404 | 12.946 |
| | | $p$-value of $\widehat{T}_n$ | 100.00 | 100.00 | 100.00 | 19.000 |
| 2010 | 1.00% | $\widehat{x}_n$ | 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% | $\widehat{x}_n$ | 2.7646 | 2.8973 | 2.7953 | 3.7364 |
| | | $p$-value of $\widehat{T}_n$ | 22.000 | 22.000 | 22.000 | 22.000 |
| | 0.05% | $\widehat{x}_n$ | 4.0410 | 4.2463 | 4.0883 | 5.5887 |
| | | $p$-value of $\widehat{T}_n$ | 22.000 | 22.000 | 22.000 | 22.000 |
| | 0.01% | $\widehat{x}_n$ | 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% | $\widehat{x}_n$ | 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% | $\widehat{x}_n$ | 3.0855 | 3.2428 | 3.1268 | 4.2407 |
| | | $p$-value of $\widehat{T}_n$ | 29.000 | 29.000 | 29.000 | 29.000 |
| | 0.05% | $\widehat{x}_n$ | 4.6096 | 4.8471 | 4.6719 | 6.3624 |
| | | $p$-value of $\widehat{T}_n$ | 29.000 | 29.000 | 29.000 | 29.000 |
| | 0.01% | $\widehat{x}_n$ | 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% | $\widehat{x}_n$ | 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% | $\widehat{x}_n$ | 3.1511 | 3.2980 | 3.1863 | 4.2427 |
| | | $p$-value of $\widehat{T}_n$ | 100.00 | 100.00 | 100.00 | 100.00 |
| | 0.05% | $\widehat{x}_n$ | 4.6182 | 4.8442 | 4.6723 | 6.3367 |
| | | $p$-value of $\widehat{T}_n$ | 100.00 | 100.00 | 100.00 | 100.00 |
| | 0.01% | $\widehat{x}_n$ | 11.789 | 12.381 | 11.931 | 16.161 |
| | | $p$-value of $\widehat{T}_n$ | 66.500 | 34.000 | 78.500 | 100.00 |

Table 4: Empirical Testing and Top Income Estimation of Korea (2007–2012). Notes: $\widehat{x}_n$ is the estimated top $x$% income out of the given population; superscript $\sharp$ indicates that the estimated top $x$% income is identical to $c_\sharp$. The unit of $\widehat{x}_n$ is KRW100 mil., and $p$-values are in %.