CIRANO
Knowledge into action

REGULARIZED GOODNESS-OF-FIT STATISTICS AND EXACT NONPARAMETRIC CONFIDENCE BANDS FOR DISTRIBUTIONS WITH APPLICATION TO HOUSEHOLD CONSUMPTION

MAME ASTOU DIOUF
JEAN-MARIE DUFOUR

CS
2026s-15
WORKING PAPER

CIRANO
Knowledge into action

The purpose of the **Working Papers** is to disseminate the results of research conducted by CIRANO research members in order to solicit exchanges and comments. These reports are written in the style of scientific publications. The ideas and opinions expressed in these documents are solely those of the authors.

Les cahiers de la série scientifique visent à rendre accessibles les résultats des recherches effectuées par des chercheurs membres du CIRANO afin de susciter échanges et commentaires. Ces cahiers sont rédigés dans le style des publications scientifiques et n'engagent que leurs auteurs.

CIRANO is a private non-profit organization incorporated under the Quebec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the government of Quebec, and grants and research mandates obtained by its research teams.

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du gouvernement du Québec, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO Partners – *Les partenaires du CIRANO*

Corporate Partners – Partenaires Corporatifs	**Governmental partners - Partenaires gouvernementaux**	**University Partners – Partenaires universitaires**
Autorité des marchés financiers	Ministère des Finances du Québec	École de technologie supérieure
Banque de développement du Canada	Ministère de l'Économie, de l'Innovation et de l'Énergie	École nationale d'administration publique de Montréal
Banque du Canada	Innovation, Sciences et Développement Économique Canada	HEC Montreal
Banque Nationale du Canada	Ville de Montréal	Institut national de la recherche scientifique
Bell Canada		Polytechnique Montréal
BMO Groupe financier		Université Concordia
Caisse de dépôt et placement du Québec		Université de Montréal
Énergir		Université de Sherbrooke
Hydro-Québec		Université du Québec
Intact Corporation Financière		Université du Québec à Montréal
Mouvement Desjardins		Université Laval
Power Corporation du Canada		Université McGill
Pratt & Whitney Canada		
VIA Rail Canada		

CIRANO collaborates with many *centres* and university research chairs; list available on its website. *Le CIRANO collabore avec de nombreux centres et chaires de recherche universitaires dont on peut consulter la liste sur son site web.*

ISSN 2292-0838 (online version)

Regularized goodness-of-fit statistics and exact nonparametric confidence bands for distributions with application to household consumption[*]

Mame Astou Diouf [†], *Jean-Marie Dufour* [‡]

Abstract

We study from a finite-sample viewpoint the problem of building tests and simultaneous confidence bands for cumulative distribution functions (CDFs), continuous or discrete. We emphasize procedures based on reweighted empirical distribution function (EDF) with shrinking bandwidths in the tails of the distribution. Since weighted statistics may have a problematic behavior when scaling factors are small (or zero), we propose to use regularized statistics. We consider a wide class set of modified EDF-type statistics, and give general characterizations of their distributions in the case of i.i.d. observations, so that the relevant distributions can be simulated in finite samples. We show that test criteria in the class studied can be implemented through the technique of Monte Carlo tests (MCT), so that the level is fully controlled in finite samples, irrespective of whether the tested distribution is continuous or discrete, without the need to establish an asymptotic distribution. Standard criteria such as the Kolmogorov-Smirnov (KS), Anderson-Darling (AD), Eicker (E), and Berk–Jones (BJ) statistics are covered as special cases. Confidence bands are built by inverting sup-type goodness-of-fit test statistics. We show that the bands based on regularized AD-type and E-type statistics have closed forms which are especially easy to compute, without nonlinear optimization.
For continuous variables, the null distributions of the statistics do not depend on the CDF tested. For noncontinuous distributions, we show that the MCT approach transparently controls test levels irrespective of the distribution tested. We also establish monotonicity properties (based on nesting image sets) and a general dominance result, so continuous critical values are valid (conservative) critical points. We show in Monte Carlo simulations that the proposed regularized goodness-of-fit tests and confidence bands are numerically tractable, reliable and yield power and precision improvements over standard procedures. The proposed methods are applied to the distribution of households' consumption in Kenya.

Résumé

Nous étudions, dans une perspective d'échantillons finis, le problème de la construction de tests et de bandes de confiance simultanées pour les fonctions de distribution cumulées (FDC),

[*] This work was supported by the William Dow Chair of Political Economy (McGill University), CIREQ, CIRANO, the Natural Sciences and Engineering Research Council of Canada, and the Social Sciences and Humanities Research Council of Canada.

[†] International Monetary Fund.
[‡] CIRANO Researcher and Fellow, William Dow Professor of Economics, McGill University, CIREQ

qu'elles soient continues ou discrètes. Nous mettons l'accent sur les procédures basées sur la fonction de distribution empirique (FDE) repondérée, avec des largeurs de bande décroissantes dans les queues de la distribution. Étant donné que les statistiques pondérées peuvent présenter un comportement problématique lorsque les facteurs d'échelle sont faibles (ou nuls), nous proposons d'utiliser des statistiques régularisées. Nous considérons une vaste classe de statistiques de type EDF modifiées et donnons des caractérisations générales de leurs distributions dans le cas d'observations i.i.d., de sorte que les distributions concernées puissent être simulées dans des échantillons finis. Nous montrons que les critères de test de la classe étudiée peuvent être mis en œuvre à l'aide de la technique des tests de Monte Carlo (MCT), de sorte que le niveau de confiance soit entièrement contrôlé dans des échantillons finis, que la distribution testée soit continue ou discrète, sans qu'il soit nécessaire d'établir une distribution asymptotique. Les critères standard tels que les statistiques de Kolmogorov-Smirnov (KS), d'Anderson-Darling (AD), d'Eicker (E) et de Berk–Jones (BJ) sont traités comme des cas particuliers. Les intervalles de confiance sont construits en inversant les statistiques de test d'adéquation de type sup. Nous montrons que les bandes basées sur les statistiques régularisées de type AD et de type E ont des formes fermées particulièrement faciles à calculer, sans optimisation non linéaire. Pour les variables continues, les distributions nulles des statistiques ne dépendent pas de la fonction de distribution cumulative (FDC) testée. Pour les distributions non continues, nous montrons que l'approche MCT permet de contrôler de manière transparente le niveau de signification, quelle que soit la distribution testée. Nous établissons également des propriétés de monotonie (basées sur des ensembles d'images imbriquées) et un résultat général de dominance, de sorte que les valeurs critiques continues constituent des points critiques valides (conservateurs). Nous montrons, à l'aide de simulations de Monte Carlo, que les tests de bon ajustement régularisés et les intervalles de confiance proposés sont numériquement gérables, fiables et permettent d'améliorer la puissance et la précision par rapport aux procédures standard. Les méthodes proposées sont appliquées à la distribution de la consommation des ménages au Kenya.

Keywords/Mots-clés: Nonparametric inference; empirical distribution function; confidence band; regularization; weighted statistics; Kolmogorov-Smirnov statistic; Anderson-Darling statistic; Eicker statistic; Berk-Jones statistic / Inférence non paramétrique ; fonction de distribution empirique ; intervalle de confiance ; régularisation ; statistiques pondérées ; statistique de Kolmogorov-Smirnov ; statistique d'Anderson-Darling ; statistique d'Eicker ; statistique de Berk-Jones

JEL Codes/Codes JEL: C3; C12; C33; C15; G1; G12; G14.

Pour citer ce document / To quote this document
Diouf, M. A. and Dufour, J-M. (2026). Regularized goodness-of-fit statistics and exact nonparametric confidence bands for distributions with application to household consumption (2026s-15, CIRANO).
https://doi.org/10.54932/JWIP9167

Contents

List of Tables

List of Figures

List of Definitions, Assumptions, Propositions and Theorems

1. Introduction

The problem of determining the distribution function of a random sample is one of the most basic problems in statistics. Related subproblems include goodness-of-fit tests (GF) of specific parametric distributions, and the building of confidence bands for distribution functions in nonparametric setups. While parametric models leave open the basic uncertainty of the parametric family considered, alternative "nonparametric methods" (such as methods based on density estimation) typically rely on potentially fragile asymptotic approximations. In this paper, we take the view that nonparametric inference is indeed a desirable objective, but wish to emphasize an approach which preserves finite-sample validity. In particular, we wish to focus on methods which can lead to finite-sample confidence bands on distribution functions, allowing for the possibility of both continuous and discrete distributions (or mixtures), as well as desirable features such as a shrinking width in distribution tails (where the distribution probabilities are either very small or very large). It is easy to see that methods based on continuous density estimation are not appropriate for discrete distributions.

Following the seminal contributions of Kolmogorov (1933, 1941) and Smirnov (1939, 1944), the only class of methods which can fill the above requirements are those based on the empirical distribution function (EDF). For reviews, see D'Agostino and Stephens (1986), Shorack and Wellner (1986, Chapter 1, Proposition 3), Reiss (1989), Gibbons and Chakraborti (1992, Theorem 3.1). Interestingly, recent work in this area has been scarce. In this paper, we reconsider this approach by proposing several improvements, including: (1) a general finite-sample theory which covers highly modified EDF-type test statistics, including statistics focusing on specific regions of the distribution; (2) a systematic treatment of the construction of finite-sample confidence sets based on a wide class of EDF-type statistics, allowing for *non-uniform* confidence bands which shrink in the tails of the distribution; (4) *regularization* methods which allow one to use potentially irregular standardizations [*e.g.*, scaling factors which can be occasionally zero or very small] in the context of general EDF-type statistics; (5) Monte Carlo test methods (MCT) which yield provably valid procedures with continuous and discrete distributions (even when mass points are not explicitly taken into account). MCT methods play a central role in allowing one to implement general EDF-type statistics in a *transparent* way, with both continuous and discrete distributions.

Several papers proposed EDF-based GF tests using sup-type or integral-type statistics. Cramér (1928) and von Mises (1931) proposed a statistic that is the integral over the real line of the square of the difference between the null distribution function F_0 and the EDF. The seminal papers of Kolmogorov (1933, 1941) and Smirnov (1939, 1944) proposed a test statistic, subsequently called the Kolmogorov-Smirnov (KS) statistic, which has become a popular statistic for nonparametric GF test. The KS statistic is the supremum over all observations of the difference between the null distribution function F_0 and the EDF. The test rejects the distribution function assumed under the null hypothesis if it is too far from the empirical distribution function, the threshold being defined by the critical point of the KS statistic. The KS test owes its popularity to its desirable properties, which have been widely studied [D'Agostino and Stephens (1986), Shorack and Wellner (1986, Chapter 1, Proposition 3), Reiss (1989), Gibbons and Chakraborti (1992, Theorem 3.1)]. Indeed, it is distribution-free when the distribution under the null is continuous, which means that the critical points of the statistic do not depend on the assumed distribution and can be used to test any

continuous distribution function. In addition, the test is consistent against all alternative continuous distributions.

A few papers proposed weighted version of the KS statistic [Anderson and Darling (1952), Darling (1957), Eicker (1979), Jaeschke (1979)] and other modifications that yield more powerful tests, including using a sum instead of the supremum [Finkelstein and Schafer (1971)] or adding a finite sample adjustment factor O'Neill and Stern (2012). Using results from Chernoff (1952), Berk and Jones (1979) proposed a statistic defined as the sup of the likelihood ratio between the EDF and F_0, which he claimed is superior to the KS statistic in the Bahadur sense. Zhang (2002) proposed a parametrization that allows to construct a class of EDF-based statistics, which are the sup or the integral of the weighted Pearson chi-square or likelihood ratio statistics. He showed that the class of statistics imbeds as special cases the statistics of KS, Anderson-Darling, Cramér-Wold-vonMises, and Berk-Jones, as well as two weighted Berk-Jones statistics. Zhang (2002) proved that Berk-Jones type statistics, which are based on the likelihood ratio principle, are more powerful than the KS-type statistics, which are based on the Pearson Chi-square statistic. Other statistics were also proposed, including EDF-based [Henze (1996), Jager and Wellner (2007), Stepanova and Pavlenko (2018), Dümbgen and Wellner (2023)] and non-EDF based [Cochran and Snedecor (1989), Massey (1951), and Noether (1963)]. Shao and Hahn (1996) proposed a normalized log spacing statistic (NLSS) based on the sum of the ratios of consecutive order statistics logarithms, which are distribution-free and consistent. However, all the above tests are only valid in continuous cases, which is a serious drawback that limits their use.

While the above tests cannot be applied for discrete distribution functions, goodness of fit to a discrete distribution is very important as well. Data of discrete nature arise from a discrete distribution (e.g., Binomial or Poisson) or from grouped data generated from continuous or discrete distribution (e.g., measurements rounded to the nearest unit such as millimeters, degrees). Only a couple of GF tests were proposed for discrete distribution functions. Klar (1999) proposed to use the supremum between the integrated distribution function (the integral of the survival function) and its empirical estimation, which is expected to perform well against heavy-tailed distributions. Kocherlakota and Kocherlakota (1986) proposed a similar statistic using probability generating functions, the power series representation of the probability mass function of the random variable.

Discrete versions of popular continuous statistics have also been proposed. Pettitt and Stephens (1977) used a variant of the KS statistic that is applied on grouped data. This statistic does not require having a minimum number of data by cells, as does the chi-square test, but it depends on the cell ordering. Choulakian, Lockhart and Stephens (1994) and Lockhart, Spinelli and Stephens (2007) proposed analogues of the three popular Cramér-vonMises statistics in the discrete case based on the grouped empirical distribution function and modified Cramér-von Mises statistics omitting the weights to improve accuracy in long tails or using weights that allow the modified statistics to remain unaltered if the cells' order is reversed (respectively). The latter (modified) statistics bear similar properties than the original statistics, while being more powerful than the Chi square test when testing fit of a uniform law against alternatives where the probabilities of falling in the cells are in a steadily increasing or decreasing pattern. Studying the properties of the Chi-square statistic and the discrete KS and Cramér-von Mises statistics under the null uniform, Steele and Chaseling (2006) found that the performance of statistics depends on the type of alternative and sample size,

and in particular none of these statistics have good power in small samples of twenty observations. However, the statistics based on EDF for ordinal data (KS, Cramér-von Mises, Anderson-Darling) behave better for trend alternative distributions (*i.e.* with increasing or decreasing cell probabilities). Other statistics for discrete distributions were also proposed [see Arnold and Emerson (2011)].

Several papers examined various related topics, including asymptotic distributions of the one-sided and two-sided KS statistics for discrete distributions – which is expressed as the maximum of a Wiener process over the set of discontinuity points [Wood and Altavela (1978)], approximation of the KS statistic in finite samples [Conover (1972)], asymptotic properties for GF test statistics under the assumption that the null distribution is continuous or in the discrete case [Gleser (1985)] or without such assumptions. KS statistic sensitive to tail alternatives [Mason and Schuenemeyer (1983)], conditional KS test [Andrews (1997)]. Other analyses proposed other types of distribution-free confidence bands for distribution functions [Frey (2008), Dümbgen and Wellner (2023)], unbiased GF test [Frey (2009)] and studied the power of discrete GF tests under an uniform null [Steele and Chaseling (2006)]. Guilbaud (1986) studied properties of statistics that could be rewritten as the supremum of a real valued function of the EDF and a distribution function $F(\cdot)$, with the KS statistic and the Anderson-Darling statistics as special cases. He showed that the statistics corresponding to two distribution functions with embedded ranges are stochastically ordered, which implies monotonicity properties for the corresponding confidence regions for distribution functions.

Our paper aims to construct confidence bands (CBs, henceforth) for distribution functions which are convenient to use and apply to all distribution functions without any restrictions. We invert many EDF-based tests, using the pivotality properties of KS-based statistics. We propose improvements to several limitations of KS-based statistics. First, the KS statistic is quite uninformative in the tails of the distribution. It has little power to discriminate between distributions that differ mainly through their tails, which reduces the performance of the KS test and CB. In particular, the KS confidence band is uniform and, given that its width is constant for all observations, the KS bounds do not converge towards 0 and 1 in the lower and the upper tails of the distribution, as distribution functions do. To correct this drawback, we propose weighted KS statistics based on the three common principles in econometrics: the Wald, Lagrange multiplier, and likelihood-ratio principles. These statistics have been proposed by Anderson and Darling (1952), Eicker (1979), and Berk and Jones (1979). The Anderson-Darling and the Eicker statistics are standardized versions of the KS statistic where the difference between the theoretical and the empirical distributions is divided by a modified standard deviation. These statistics allow discriminating between distributions that differ mostly through their tails, and provide tools to build finite-sample nonparametric CBs whose widths decrease with observations further from the center of the distribution.

Second, the Anderson-Darling and the Eicker statistics have their own drawbacks. The power of the GF tests they yield is lower than the power of the standard KS test when testing distributions with low dispersion that differ more in the center of the distribution than in the tails. Moreover, the weights in the denominators of those statistics become very close to zero for observations in the tails, leading to erratic behavior of the statistics. We propose improved weighted KS statistics to correct these. These statistics are obtained by adding a regularization term in the denominator of the Anderson-Darling and the Eicker statistics. They retain the advantages of the weighted KS statistics but do not suffer from instability, improving the performance of the inference. By inversion of the

regularized statistics, we build improved exact CBs for distribution functions.

The Berk-Jones statistics uses the supremum, over all observations, of the log-likelihood ratio of the empirical distribution function and the theoretical distribution function as a distance between these two functions. This statistic has been proved to dominate any weighted KS statistic, in the sense of Bahadur and is thus a challenging referral for our inference methods. It has been used by Owen (1995) to propose a CB for distribution functions.

In the continuous case, we show that the distributions of the empirical distribution-based statistics are pivotal and that their critical points do not depend on the distribution function being tested under the null hypothesis. Hence, the corresponding CBs depend on the distribution only through the sample; they are built using the same critical points for all continuous distribution functions, which make them easy to compute. For noncontinuous distribution functions, we derive monotonicity properties which exploit the nesting of the image sets of different distributions to narrow the CBs without altering their reliability.

We compare the relative performance of the nonparametric and parametric inference methods. Monte Carlo simulations are performed to study the power of the goodness-of-fit tests under various hypotheses. In both studies, we study carefully the choice of the regularization parameter. The results show that regularized statistics yield more powerful goodness-of-fit tests than the existing ones when applied to distributions with more discrepancy in the tails.

The paper is organized as follows. Section 2 defines the class of statistics studied, along with some standard statistics which are covered as special cases (Kolmogorov-Smirnov, Anderson-Darling, Eicker). In Section 3, we introduce the regularized statistics and we derive explicit expressions to compute them and build CBs for distribution functions. In Section 4, we discuss the finite-sample distributional theory of the general class of EDF-type statistics considered in the paper, In Section 5, we give general dominance results relevant to non-continuous distributions. Section 6 discusses order statistics representations which can be useful for implementation. Section 7 discusses how the technique of Monte Carlo tests can be applied to obtain provably valid p-values. Section 8 provides formulas for building confidence bands based on the proposed statistics. Section 9 presents Monte Carlo simulation results. Section 10 applies the proposed methods to household consumption in Kenya. The proofs, details on the construction of confidence bands, and some additional simulation results are presented in the Appendix.

2. Framework

In this section, we describe the class of test statistics and problems studied in this paper. Our objective consists in building computationally simple finite-sample distribution-free confidence bands for the distribution function of a random variable, without regularity assumptions on the form of the distribution, such as continuity. Consequently, methods based on density estimation are not appropriate here, since a density function (with respect to the Lebesgue measure) may not exist; for further discussion of density estimation, see Rosenblatt (1956), Parzen (1962), Hall, Racine and Li (2004), Guilbaud (1986). So we focus on methods based on empirical distribution functions (EDFs). Further, the fact that we wish to allow for the possibility of an unknown discrete distribution will entail restrictions on the class of EDF-based statistics we shall consider.

We suppose we have a sample of n i.i.d. observations $X_1, \ldots, X_n$ on a real random variable X with $\mathbb{P}[X \leq x] = F(x)$ for $x \in \mathbb{R}$. We also set $F(x-) = \mathbb{P}[X < x] = \lim_{y \uparrow x} F(x)$, $p_F(x) = \mathbb{P}[X = x] = F(x) - F(x-)$, $F(-\infty) = 0$ and $F(\infty) = 1$. To underscore the fact that the probability measure $\mathbb{P}$ can be represented by F, we also write $\mathbb{P}_F$ instead of $\mathbb{P}$, *i.e.* $\mathbb{P}_F[X \leq x] = F_0(x)$ if and only if $\mathbb{P}[X \leq x] = F_0(x) = F(x)$ for all $x \in \mathbb{R}$. We wish to build confidence bands for $F(\cdot)$. The distribution F is not restricted, so F may be continuous or discrete. We denote $\mathscr{F}$ the set of all distribution functions F, $\tilde{\mathscr{F}}$ the set of continuous distribution functions, $\mathscr{F}_S$ the set of distribution functions with support $S \subseteq \mathbb{R}$, and $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$. The empirical distribution of our sample is:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[X_i \leq x] \tag{2.1}$$

where $\mathbf{1}[E] = 1$ if E is true, and $\mathbf{1}[E] = 0$ otherwise.

We consider the problem of testing

$$\mathscr{H}_0(F_0) : X_1, \ldots, X_n \text{ are i.i.d. each one with distribution function } F_0(x)\,. \tag{2.2}$$

The best known statistic for testing $\mathscr{H}_0(F_0)$ is the Kolmogorov-Smirnov (KS) statistic:

$$KS_n(F_0) = \sup_{-\infty < x < +\infty} \sqrt{n}\, \left|\hat{F}_n(x) - F_0(x)\right|\,. \tag{2.3}$$

When $F_0(x)$ is continuous, it is well known that the null distribution of KS does not depend on F_0, so critical values applicable to test any continuous distribution F_0 can be computed and have been tabulated; see Smirnov (1948), D'Agostino and Stephens (1986), Shorack and Wellner (1986, Chapter 1, Proposition 3), Reiss (1989), Gibbons and Chakraborti (1992, Theorem 3.1).
When F_0 is discrete, the distribution of KS depends on F_0, so the problem is more complex. However, the distribution of KS under any non-continuous distribution F_0 is dominated by the corresponding (unique) distribution for the continuous case; see Guilbaud (1986). This entails that using critical values obtained for continuous distributions to test a non-continuous distribution F_0 still allows one to control the level of the test, *i.e.* the probability of rejecting $\mathscr{H}_0(F_0)$ is not larger than the nominal level of the test, but could be lower (leading to a conservative procedure).

The latter feature is especially convenient in order to build a confidence set for an unknown distribution function $F(\cdot)$. If

$$\mathbb{P}_{F_0}\left[KS_n(F_0) > c_{KS}(\alpha; n)\right] = \alpha \tag{2.4}$$

for F_0 a continuous distribution function and $0 < \alpha < 1$, we have $\mathbb{P}_{F_0}\left[KS_n(F_0) > c_{KS}(\alpha; n)\right] \leq \alpha$ irrespective of the form of F_0, for any $0 \leq \alpha \leq 1$. Given the sup form of the KS statistic, we have:

$$\mathbb{P}_{F_0}\left[\left|\hat{F}_n(x) - F_0(x)\right| \leq c_{KS}(\alpha; n)/\sqrt{n}\,, \forall x\right] \geq 1 - \alpha\,, \tag{2.5}$$

which yields the following confidence set with level $1 - \alpha$ for the distribution function of X:

$$C_{KS}(\alpha) = \left\{F_0 \in \mathscr{F} : \hat{F}_n(x) - \frac{c_{KS}(\alpha; n)}{\sqrt{n}} \leq F_0(x) \leq \hat{F}_n(x) + \frac{c_{KS}(\alpha; n)}{\sqrt{n}},\ \forall x\right\} \cap [0, 1] \tag{2.6}$$

where the band entailed by (2.5) is truncated to remain within the admissible interval [0, 1]. The confidence set $C_{KS}(\alpha; n)$ is both general and simple to compute. By contrast, these features do not appear to be available for goodness-of-fit statistics based on weighted integrals (or sums) of the differences $\hat{F}_n(x) - F_0(x)$, such as the Cramér-von -Mises statistic [Cramér (1928), von Mises (1931)], the W^2 statistic proposed by Anderson and Darling (1952), or Watson's $\mathscr{U}^2$ statistic [Watson (1961)]. To the best of our knowledge, easily usable closed-form confidence bands based on such tests have not been proposed, nor dominance results applicable to discrete distributions.

The *KS* statistic, however, is quite uninformative in the tails of the distribution. This is manifest from observing that the critical value $c_{KS}(\alpha; n)/\sqrt{n}$ does not depend on x, so the width of the band in $C_{KS}(\alpha; n)$ is never smaller than $c_{KS}(\alpha; n)/\sqrt{n}$. This property is quite unsatisfactory in view of the fact that the variance of $\hat{F}_n(x) - F_0(x)$ is $F_0(x)[1 - F_0(x)]/n$ which goes to zero as x goes to $\pm\infty$. We will now consider a general class of EDF-based statistics which allows us to improve the precision of the confidence sets for F while keeping the attractive features of KS-type procedures.

We first consider the problem of testing $\mathscr{H}_0(F_0)$ with general statistics of the form

$$D_n(\hat{F}_n, F_0; C_0) = \sup_{x \in C_0} \bar{D}_n[\hat{F}_n(x), F_0(x)] \tag{2.7}$$

where $\bar{D}_n\left[\hat{F}_n(x), F_0(x)\right]$ is a function of $\hat{F}_n(x)$ and $F_0(x)$, which typically represents the difference between $\hat{F}_n(x)$ and $F_0(x)$, and $C_0 = \bar{C}[F_0, X^{(n)}]$ is a subset of $\mathbb{R}$ which may depend on the function $F_0(\cdot)$ and the data $X^{(n)}$. We call C_0 the *optimization domain* of $D_n(\hat{F}_n, F_0; C_0)$, and statistics of the form $D_n(\hat{F}_n, F_0; C_0)$ sup-type EDF-based ($\sup - EDF$) GF statistics. Of course, we can take $C_0 = \mathbb{R}$, but restricting the set C_0 may be required with functions $\bar{D}_n$ which are not well defined over the whole domain $\mathbb{R}$ (or go to infinity at some point). Further, different choices of optimization domain yield different test statistics, which may possess different power features. Note $D_n(\hat{F}_n, F_0; C_0)$ can also be rewritten as

$$D_n(\hat{F}_n, F_0; C_0) = \sup_{-\infty < x < +\infty} \mathbf{1}[x \in C_0]\, \bar{D}_n[\hat{F}_n(x), F_0(x)]\,. \tag{2.8}$$

On setting $C_0 = \mathbb{R}$ and $\Delta[\hat{F}_n(x), F_0(x)] = \sqrt{n}\,[\hat{F}_n(x) - F_0(x)]$, the usual KS statistics (for one-sided and two-sided tests) are then:

$$KS_n^+ = \sup_{-\infty < x < +\infty} \Delta[\hat{F}_n(x), F_0(x)], \quad KS_n^- = \sup_{-\infty < x < +\infty} \Delta[F_0(x), \hat{F}_n(x)], \tag{2.9}$$

$$KS_n = \max\{KS_n^+, KS_n^-\} = \sup_{-\infty < x < +\infty} |\Delta[\hat{F}_n(x), F_0(x)]|. \tag{2.10}$$

KS tests involve clearly taking the supremum of tests comparing different "elementary" tests based on the differences $\sqrt{n}\,[\hat{F}_n(x) - F_0(x)]$ at all possible values of x ($C_0 = \mathbb{R}$), where $\hat{F}_n(x) \sim \text{Bin}[n, F_0(x)]$ follows a binomial distribution under the null hypothesis, so that

$$\sigma_n(x) = \text{Var}(\sqrt{n}\,[\hat{F}_n(x) - F_0(x)]) = F_0(x)[1 - F_0(x)] \to 0 \quad \text{as } x \to \pm\infty. \tag{2.11}$$

Even though this leads to a statistic whose distribution can be established both in finite samples

and asymptotically, the differences follow very different distributions under $\mathscr{H}_0(F_0)$, and in the tails of distribution. This suggests considering other "elementary" tests of the hypotheses $\mathscr{H}_0(x, F_0)$: $F(x) = F_0(x)$, at different values of x.

Natural choices here consist in considering Wald-type, score-type, and LR-type statistics.

1. In Wald-type statistics, $\sigma_n(x)$ is estimated using $\hat{F}_n(x)$, so $\bar{D}_n$ takes one of the following forms:

$$\bar{D}_n^{W+}[\hat{F}_n(x), F_0(x)] := \Delta_n^W[\hat{F}_n(x), F_0(x)], \quad \bar{D}_n^{W-}[\hat{F}_n(x), F_0(x)] := \Delta_n^W[F_0(x), \hat{F}_n(x)], \quad (2.12)$$

$$\bar{D}_n^W[\hat{F}_n(x), F_0(x)] := |\Delta_n^W[\hat{F}_n(x), F_0(x)]|, \ \Delta_n^W[\hat{F}_n(x), F_0(x)] := \frac{\sqrt{n}[\hat{F}_n(x) - F_0(x)]}{\{\hat{F}_n(x)[1-\hat{F}_n(x)]\}^{1/2}}. \quad (2.13)$$

2. In score-type statistics, $\sqrt{n}\,[\hat{F}_n(x) - F_0(x)]$ is scaled by its standard error $\sigma_n(x)$ under $\mathscr{H}_0(x, F_0)$, so $\bar{D}_n$ has one of the forms:

$$\bar{D}_n^{S+}[\hat{F}_n(x), F_0(x)] := \Delta_n^S[\hat{F}_n(x), F_0(x)], \quad \bar{D}_n^{S-}[\hat{F}_n(x), F_0(x)] := \Delta_n^S[F_0(x), \hat{F}_n(x)], \quad (2.14)$$

$$\bar{D}_n^S[\hat{F}_n(x), F_0(x)] := \left|\Delta_n^S[F_0(x), \hat{F}_n(x)]\right|, \ \Delta_n^S[\hat{F}_n(x), F_0(x)] := \frac{\sqrt{n}[\hat{F}_n(x) - F_0(x)]}{\{F_0(x)[1-F_0(x)]\}^{1/2}}. \quad (2.15)$$

3. The log-LR-type statistic uses $\bar{D}_n = \bar{D}_n^{LR}$ where

$$\bar{D}_n^{LR}[\hat{F}_n(x), F_0(x)] := K[\hat{F}_n(x), F_0(x)], \quad (2.16)$$

$$\begin{array}{rll} K(t,x) := & t\,\log\left(\frac{t}{x}\right) + (1-t)\,\log\left(\frac{1-t}{1-x}\right) & \text{if } 0 < x < t < 1, \\ & := 0 & \text{if } 0 \le t \le x \le 1, \\ & := \infty & \text{otherwise.} \end{array} \quad (2.17)$$

The Wald-type and score-type statistics represent natural choices by standardization and studentization of the KS statistic. Standard sup-type EDF-based test statistics include the following ones. Tests of this type can be viewed as applications of the *union-intersection* approach. On taking $\bar{D}_n = \bar{D}_n^{W+}$ and $C_0 = \left(X_{(1)}, X_{(n)}\right)$, where $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ are the order statistics of $X_1, \ldots, X_n$, we get the statistic studied by Eicker (1979):

$$\widehat{V}_n = D_n^E[\hat{F}_n, F_0] = \sup_{X_{(1)} < x < X_{(n)}} \frac{\sqrt{n}[\hat{F}_n(x) - F_0(x)]}{\{F_0(x)[1-F_0(x)]\}^{1/2}} = \sup_{X_{(1)} < x < X_{(n)}} \Delta_n^W\left[\hat{F}_n(x), F_0(x)\right]. \quad (2.18)$$

With $\bar{D}_n = \bar{D}_n^S$ and $C_0 = \mathbb{R}$, we get a weighted KS statistic proposed by Anderson and Darling (1952):

$$D_n^{AD}[\hat{F}_n, F_0] = \sup_{-\infty < x < +\infty} \frac{\sqrt{n}[\hat{F}_n(x) - F_0(x)]}{(F_0(x)[1-F_0(x)])^{1/2}}. \quad (2.19)$$

For $\bar{D}_n = \bar{D}_n^{LR}$ and $C_0 = \{x : 0 \le x \le 1;\, \hat{F}_n(x) > x\}$, we get the LR-type statistic proposed by Berk and Jones (1979):

$$R_n = \sup_{0 \le x \le 1} \{K[\hat{F}_n(x), x]\}. \quad (2.20)$$

The general class of statistics in (2.7) also includes as special cases alternative statistics studied by Donoho and Jin (2004), Donoho and Jin (2008), Cayon, Jin and Treaster (2005), Hall and Jin (2008), Stepanova and Pavlenko (2018), and Dümbgen and Wellner (2023). In particular, Stepanova and Pavlenko (2018) focus on the case where $\bar{D}_n[\hat{F}_n(x), F_0(x)] = \sqrt{n}|\hat{F}_n(x) - F_0(x)|/q(F_0(x))$ with: $q(u) = \sqrt{u(1-u)}$, while Dümbgen and Wellner (2023) proposed a refined Berk-Jones statistic which performs better in the center of the distribution. These authors focus on asymptotic distributional theory.

3. Regularized goodness-of-fit statistics

The statistics discussed in Section 2 raise numerical and statistical difficulties. Because $\hat{F}_n(x) = 0$ in the left tail of the distribution and $\hat{F}_n(x) = 1$ in the right tail, the denominator in $\bar{D}_n^W$ is zero at those points. Very small value of the denominator may also have a strong influence on the distribution of the test statistic. Similarly, if $F_0(x)$ has bounded support – a possibility we allow here – we have $F_0(x) = 0$ or $F_0(x) = 1$ for some values of x, so score-type statistics ($\bar{D}_n^S$) are not (finite) real numbers. On noting that $\lim_{x\to 0^+} x\log(x) = 0$, the same problem shows up for $\bar{D}_n^{LR}$ when $F_0(x) = 0$ or $F_0(x) = 1$. In other words, the corresponding test statistics may not be well defined.

To avoid this difficulty, we consider two approaches: (1) restricted optimization domain, and (2) regularization. The first one restricts the domain C_0 over which $\bar{D}_n[\hat{F}_n(x), F_0(x)]$ is maximized. The second one consists in "regularizing" $\bar{D}_n[\hat{F}_n(x), F_0(x)]$ so it is always finite irrespective of the value of $x \in \mathbb{R}$. In the first method, we may take $C_0 = [x_{\min}, x_{\max}]$ where $x_{\min}$ and $x_{\max}$ are chosen so that $0 < \hat{F}_n(x) < 1$ (for $\bar{D}_n^W$) or $0 < F_0(x) < 1$ (for $\bar{D}_n^S$ and $\bar{D}_n^{LR}$). For the second method, let

$$\delta_n = \delta_n[\hat{F}_n(x), F_0(x)] \tag{3.1}$$

a random variable such that $0 < \delta_n[\hat{F}_n(x), F_0(x)]$ when $\hat{F}_n(x)[1-\hat{F}_n(x)] = 0$ or $F_0(x)[1-F_0(x)] = 0$. In particular, δ_n could be a positive constant, which depends on n. This suggests considering the following modified Wald-type and score-type statistics, and obtained through replacing $\Delta_n^W[\hat{F}_n(x), F_0(x)]$ and $\Delta_n^S[\hat{F}_n(x), F_0(x)]$ by:

$$\Delta_n^W[\hat{F}_n, F_0; \delta_n] = \frac{\sqrt{n}[\hat{F}_n(x) - F_0(x)]}{\{\hat{F}_n(x)[1-\hat{F}_n(x)] + \delta_n\}^{1/2}}, \tag{3.2}$$

$$\Delta_n^S[\hat{F}_n, F_0; \delta_n] = \frac{\sqrt{n}[\hat{F}_n(x) - F_0(x)]}{\{F_0(x)[1-F_0(x)] + \delta_n\}^{1/2}}, \tag{3.3}$$

respectively. Clearly, non-regularized statistics correspond to $\delta_n = 0$. The corresponding sup-type statistics will be denoted as in (2.12) - (2.15), upon replacing $(\hat{F}_n, F_0; C_0)$ by $(\hat{F}_n, F_0; C_{0,}, \delta_n)$.

For the LR-type statistic, an appropriate regularization is less straightforward. One possibility consists in replacing $\log(p)$ and $\log(1-p)$ in $K[\hat{p}, p]$ in (2.17) by

$$\begin{array}{ll} l(p, \delta_n) & = \max\{\log(p), -\delta_n\} \quad \text{if } \delta_n > 0 \\ & = \log(p) \qquad\qquad\quad \text{if } \delta_n = 0 \end{array} \tag{3.4}$$

where δ_n is defined as in (3.1). When $\delta_n > 0$, $l(p)$ is never below $-\delta_n$, so it can go to $-\infty$. When $\delta_n = 0$, no regularization is used (and C_0 should be appropriately restricted). This leads one to replace $K(\hat{p}, p)$ in (2.17) by

$$\begin{aligned} K[\hat{p}, p; \delta_n] &= [\hat{p}\log(\hat{p}) + (1-\hat{p})\log(1-\hat{p})] - [\hat{p}\, l(p, \delta_n) + (1-\hat{p}) l(1-p, \delta_n)] \\ &= \hat{p}\log\left(\frac{\hat{p}}{\bar{p}}\right) + (1-\hat{p})\log\left(\frac{1-\hat{p}}{1-\bar{p}}\right) \end{aligned} \tag{3.5}$$

where $\bar{p} = \exp\{l(p, \delta_{Ln})\}$, and $\bar{D}_n^{LR}\left[\hat{F}_n(x), F_0(x)\right]$ by

$$\bar{D}_n^{LR}\left[\hat{F}_n(x), F_0(x); \delta_n\right] = K[\hat{F}_n(x), F_0(x); \delta_n]\,. \tag{3.6}$$

The two approaches described can easily be combined, yielding the following statistics:

$$D_n^W(\hat{F}_n, F_0; C_0, \delta_n) = \sup_{x \in C_0} \bar{D}_n^W[\hat{F}_n(x), F_0(x); \delta_n]\,, \tag{3.7}$$

$$D_n^S(\hat{F}_n, F_0; C_0, \delta_n) = \sup_{x \in C_0} \bar{D}_n^S[\hat{F}_n(x), F_0(x); \delta_n]\,, \tag{3.8}$$

$$D_n^{LR}(\hat{F}_n, F_0; C_0, \delta_n) = \sup_{x \in C_0} K[\hat{F}_n(x), F_0(x); \delta_n]\,. \tag{3.9}$$

If $\delta_n = 0$ and C_0 is chosen appropriately, a pure domain restriction approach is used. If $\delta_n > 0$ and $C_0 = \mathbb{R}$, only regularization is employed. But, as indicated by (3.7) - (3.9), both approaches can be simultaneously employed.

4. Distributional theory

In this section, we describe a number of general finite-sample distributional properties of sup-type goodness-of-fit statistics based on empirical distribution functions, which will play a central role in the methods proposed in this paper. We first give general lemmas which provide useful representations of the test statistics $D_n(\hat{F}_n, F_0; C_0)$ in terms of the distribution set by the null hypothesis F_0 and the distributions of the $X_1, \ldots, X_n$.

Lemma 4.1 REPRESENTATIONS OF SUP-EDF STATISTICS IN TERMS OF F_0 AND F_1. *Let F_0 and F_1 be two distribution functions, $X^{(n)} = (X_1, \ldots, X_n)'$ a vector of n real random variables, and $D_n(\hat{F}_n, F_0; C_0)$ a statistic of the form given by (2.7). If $X_i = F_1^{-1}(U_i)$ where U_i is a random variable such that $0 < U_i < 1$, for $i = 1, \ldots, n$, then*

$$\hat{F}_n(x) = H_n[F_1(x)]\,,\ \forall x \in \mathbb{R}\,, \tag{4.1}$$

$$D_n(\hat{F}_n, F_0; C_0) = \sup_{x \in C_0} \bar{D}_n\left[H_n\big(F_1(x)\big), F_0(x)\right] \tag{4.2}$$

where

$$H_n(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[U_i \leq u]\,,\ 0 \leq u \leq 1. \tag{4.3}$$

Further,

$$\begin{aligned} D_n(\hat{F}_n, F_0; C_0) &= \sup_{u \in F_0(C_0)} \bar{D}_n\left[H_n(u), u\right] && \text{if } F_1 = F_0\,, \\ &= \sup_{u \in F_0(C_0)} \bar{D}_n\left[H_n\big(F_1[F_0^{-1}(u)]\big), u\right] && \text{if } F_0 \text{ is strictly increasing,} \\ &= \sup_{x \in F_1(C_0)} \bar{D}_n\big[H_n(u), F_0[F_1^{-1}(u)]\big] && \text{if } F_1 \text{ is strictly increasing.} \end{aligned} \tag{4.4}$$

The above lemma is purely analytical in the sense the identities hold everywhere on the sample space. However, it holds under the special assumption $X_i = F_1^{-1}(U_i)$ where U_i is a random variable such that $0 < U_i < 1$, for $i = 1, \ldots, n$. When $U_i \sim U(0,1)$ [*i.e.*, U_i follows a uniform distribution on $(0,1)$], it is well known that $\mathbb{P}[F_1^{-1}(U_i) \leq x] = F_1(x)$. This raises the question whether the assumption $X_i = F_1^{-1}(U_i)$ can be replaced by the more intuitive assumption $\mathbb{P}[X_i \leq x] = F_1(x)$.

If F_1 is continuous, $\mathbb{P}[X_i \leq x] = F_1(x)$ entails $F_1(X_i) \sim U(0,1)$, so an appropriate variable U_i can be built through the transformation $U_i = F_1(X_i)$. However, when F_1 is not continuous, $F_1(X_i)$ is not anymore uniform. In the following lemma, we show that the identities given by Lemma 4.1 remain valid *almost surely* with $U_i = F_1(X_i)$, irrespective of the form of F_1 or whether the null hypothesis $\mathscr{H}_0(F_0)$ holds.

Lemma 4.2 ALMOST SURE REPRESENTATIONS OF SUP-EDF STATISTICS IN TERMS OF F_0 AND F_1. *Let F_0 and F_1 be two distribution functions, $X^{(n)} = (X_1, \ldots, X_n)'$ a vector of n real random variables, and $D_n(\hat{F}_n, F_0; C_0)$ a statistic of the form given by* (2.7). *If $\mathbb{P}[X_i \leq x] = F_1(x)$, $\forall x \in \mathbb{R}$, the identities* (4.1), (4.2) *and* (4.4) *hold* almost surely *with $U_i = F_1(X_i)$, $i = 1, \ldots, n$.*

Lemma 4.2 characterizes the distribution of $D_n(\hat{F}_n, F_0; C_0)$ irrespective of the forms of F_1 and F_0, whether continuous or discrete. Dependence between the variables $X_1, \ldots, X_n$ is also allowed. In the standard case where the observations are i.i.d., we can now give a general representation of the distribution of $D_n(\hat{F}_n, F_0; C_0)$ under both the null and the alternative hypotheses.

Proposition 4.3 DISTRIBUTIONS OF SUP-EDF STATISTICS. *Let F_0 and F be two distribution functions, $X^{(n)} = (X_1, \ldots, X_n)'$ a vector of i.i.d. real random variables such that $\mathbb{P}[X_i \leq x] = F(x)$ for $i = 1, \ldots, n$, and $D_n(\hat{F}_n, F_0; C_0)$ a statistic of the form given by* (2.7). *Then $D_n(\hat{F}_n, F_0; C_0)$ is distributed like the random variable*

$$\tilde{D}_n(\hat{F}_n, F_0; C_0) := \sup_{x \in C_0} \bar{D}_n\left[H_n\big(F(x)\big), F_0(x)\right] \tag{4.5}$$

where $H_n(u)$ is defined by (4.3) *and U_i, $i = 1, \ldots, n$, are i.i.d. random variables with $U(0,1)$*

distribution. Further,

$$\begin{aligned}\tilde{D}_n(\hat{F}_n, F_0; C_0) &= \sup_{u\in F_0(C_0)} \bar{D}_n[H_n(u), u] && \text{if } F = F_0,\\ &= \sup_{u\in F_0(C_0)} \bar{D}_n\left[H_n\left(F[F_0^{-1}(u)]\right), u\right] && \text{if } F_0 \text{ is strictly increasing,}\\ &= \sup_{x\in F(C_0)} \bar{D}_n\left[H_n(u), F_0[F^{-1}(u)]\right] && \text{if } F \text{ is strictly increasing.}\end{aligned} \tag{4.6}$$

The distributions given by Proposition 4.3 all depend on F_0 and F, even under the null hypothesis where $F = F_0$. In the latter case, dependence upon F_0 comes through the optimization domain $F_0(C_0)$, even when F_0 is continuous. When $C_0 = \mathbb{R}$, $F = F_0$ and F_0 is continuous, we have $(0,1) \subseteq F_0(C_0)$ and the distribution of $D_n(\hat{F}_n, F_0; C_0)$ does not depend on F_0. However, this is not the case when F_0 is discrete or $C_0 \neq \mathbb{R}$ (even for continuous distributions). Restricting the optimization domain does not generally yield a distribution which does not depend on F_0.

We will now give a general condition under which the distribution of the test statistic $D_n(\hat{F}_n, F_0; C_0)$ does not depend on F_0 under the null hypothesis.

Assumption 4.1 OPTIMIZATION DOMAIN DISTRIBUTIONAL EQUIVARIANCE. *For all $F_0 \in \mathscr{F}_0$, $C_0 = \bar{C}[F_0, X^{(n)}]$ and $\bar{C}[F_0, X^{(n)}]$ satisfies the following property*

$$F_0\left(\bar{C}[F_0, X^{(n)}]\right) = \tilde{C}[F_0\left(X^{(n)}\right)] \tag{4.7}$$

where $F_0\left(X^{(n)}\right) = \left(F_0(X_1), \ldots, F_0(X_n)\right)'$ and $\tilde{C}[F_0\left(X^{(n)}\right)]$ depends on the function F_0 only through $F_0\left(X^{(n)}\right)$.

Proposition 4.4 DISTRIBUTIONS OF SUP-EDF STATISTICS: I.I.D. CONTINUOUS VARIABLES. *Let $X(n) = (X_1, \ldots, X_n)'$ a vector of n i.i.d. real random variables such that $\mathbb{P}[X_i \leq x] = F_0(x)$ for $i = 1, \ldots, n$, where $F_0 \in \mathscr{F}_0$, and $D_n(\hat{F}_n, F_0; C_0)$ a statistic of the form given by (2.7). If and Assumption 4.1 holds and $F_0 \in \mathscr{F}_0$, $D_n(\hat{F}_n, F_0; C_0)$ is distributed like*

$$\tilde{D}_{0n} = \sup_{u\in\tilde{C}[F_0(X(n))]} \bar{D}_n[H_n(u), u] \tag{4.8}$$

where $H_n(u)$ is defined as in (4.3) and $U_i,\ i = 1, \ldots,\ n$, are i.i.d. random variables with uniform distributions on $(0,1)$. If furthermore the distributions in $\mathscr{F}_0$ are continuous, $D_n(\hat{F}_n, F_0; C_0)$ is distributed like

$$\tilde{D}_{1n} = \sup_{u\in\tilde{C}[U(n)]} \bar{D}_n[H_n(u), u] \tag{4.9}$$

where $U(n) = (U_1, \ldots, U_n)$, and the distribution of $D_n(\hat{F}_n, F_0; C_0)$ is the same for all $F_0 \in \mathscr{F}_0$.

Proposition 4.4 entails that the distribution of $D_n(\hat{F}_n, F_0; C_0)$ is independent of $F(x)$ whenever $F(x)$ is continuous. $D_n(\hat{F}_n, F_0; C_0)$ can be rewritten using only uniform random variables on $(0,1)$. All statistics of the form $D_n(\hat{F}_n, F_0; C_0)$ are pivotal for continuous distribution functions. Hence, the critical points associated to those statistics are also independent of $F(x)$. This property considerably

simplifies the implementation of the tests and CBs associated with such statistics. A unique set of critical points is needed to compute these for all continuous distribution functions.

5. Dominance properties

When $F(x)$ is not continuous, the distribution of $D_n(\hat{F}_n, F_0; C_0)$ is different for each $F(x)$ tested. The associated critical points are also modified by the distribution of the sample. Hence, a new set of critical values need be computed to implement the tests for each distribution, making inference relatively difficult, especially if we wish to build confidence sets for an unknown distribution function. To simplify the inference in such cases, we shall exploit the following general dominance property.

Proposition 5.1 EXTREMAL PROPERTY OF CONTINUOUS DISTRIBUTIONS FOR SUP-EDF STATISTICS. *Let $X_1, \ldots, X_n$ be i.i.d. observations on X and $\hat{F}_n(x)$ the corresponding empirical distribution function. Let $F(x) \in \tilde{\mathscr{F}}$ be a continuous distribution function and $G(x) \in \mathscr{F}$ a non-continuous one. For any level $\alpha,\ 0 \leq \alpha \leq 1,$ the critical value associated with $D_n(\hat{F}_n, F_0; C_0)$ for testing the null hypothesis $\mathscr{H}_0(F)$ as defined by equation (2.2) is larger than or equal to the critical value associated with $D_n(\hat{F}_n, G; C_0)$ for testing the null hypothesis $\mathscr{H}_0(G)$:*

$$\mathbb{P}_G\left[D_n(\hat{F}_n, G; C_0) \geq x\right] \leq \mathbb{P}_F\left[D_n(\hat{F}_n, F_0; C_0) \geq x\right],\ \forall x \tag{5.10}$$

or equivalently

$$\mathbb{P}_F\left[D_n(\hat{F}_n, F_0; C_0) \geq D_{n\alpha}\right] \leq \alpha \Rightarrow \mathbb{P}_G\left[D_n(\hat{F}_n, G; C_0) \geq D_{n\alpha}\right] \leq \alpha,\ \forall D_{n\alpha}. \tag{5.11}$$

Proposition 5.2 FINITE-SAMPLE CONFIDENCE BANDS FOR GENERAL DISTRIBUTION FUNCTIONS. *Let $X_1, \ldots, X_n$ be n i.i.d. observations on X and $\hat{F}_n(x)$ the corresponding empirical distribution function. Let $F(x) \in \tilde{\mathscr{F}}$ be a continuous distribution function and $G(x) \in \mathscr{F}$ be a non-continuous one. For any level $\alpha,\ 0 \leq \alpha \leq 1,$ the confidence band obtained by inverting the test of the null hypothesis $\mathscr{H}_0(F)$ as defined by equation (2.2) using appropriate critical points with level α for $D_n(\hat{F}_n, F_0; C_0)$ yields a confidence band for $G(x)$ with level larger than or equal to $1-\alpha$. Equivalently, if C_n is defined as follows:*

$$C_n(\alpha) = \{H \in \mathscr{F} : D_n(\hat{F}_n, H_1; C_0) \leq c(\alpha, n)\} \tag{5.12}$$

where $c_\alpha = \inf\{D_{n\alpha},\ \mathbb{P}_F\left[D_n(\hat{F}_n, F_0; C_0) \geq D_{n\alpha}\right] \leq \alpha\}$, then

$$\mathbb{P}_G\left[G \in C_n(\alpha)\right] \geq 1-\alpha. \tag{5.13}$$

Propositions 5.1 and 5.2. highlight some interesting properties of the empirical distribution function-based statistics and CBs which simplify their implementation when applied to noncontinuous distributions. Proposition 5.1. states that the critical values of $D_n(\hat{F}_n, G; C_0)$ for continuous distribution functions $F(x)$ are conservative for noncontinuous functions $G(y)$. Using the appropriate critical values of level α for $F(x)$ provide a test of level less than or equal to α for $G(y)$.

Therefore, rejection of the null hypothesis with such test leads to rejection for the nominal level α. In other words, the result of the test based on $D_n(\hat{F}_n, G; C_0)$ remains valid and one can use the conservative critical values to compute tests and CBs for noncontinuous distribution functions. Let's remember that those critical points–that applies to continuous distributions–are independent of the function being tested and are thus, identical for all continuous distributions. Note that these propositions hold for any continuous distribution function $F(x)$.

Likewise, the CBs for $G(y)$ built using appropriate critical points for continuous distribution functions will be of level larger than or equal to $1-\alpha$. Using these properties, critical points from continuous distribution functions can be applied to any sample from a general distribution function. The resulting CBs will be of level at least equal to $1-\alpha$.

Even though the conservative critical points provide valid inference for noncontinuous distribution functions, using them alters the performance of the inference. The question is how far the quality of the performed inference is affected? We assess this question using the properties of tests and CBs. Concerning the tests, when the null hypothesis is accepted with level α, the conclusion of the test remains valid for levels less than or equal to α. Conversely, when the null hypothesis is rejected with level α, it will be still rejected for levels larger than α but might be accepted for lower levels. Concerning CBs, the impact of the using conservative critical points can be studied using the level of confidence (accuracy) and the width (precision) of the CBs. Given that the CBs using conservative critical points have a higher level than the targeted one, they will be wider than the CBs with effective level $1-\alpha$. To reduce the width, exact critical values corresponding to the distribution function under interest can be computed. However, by doing this, the CBs will loose one of their major advantages. To avoid this shortcoming, we derive monotonicity properties that can be used to narrow CBs without altering their reliability. These results are based on information about the set of discontinuities of the distribution function.

Proposition 5.3 RANGE MONOTONICITY OF CRITICAL VALUES. *Let $X_1, \ldots, X_n$ be n i.i.d. observations on X and $\hat{F}_n(x)$ the corresponding empirical distribution function. Let $F(x)$ and $G(y)$ be two distribution functions such that $G(\bar{\mathbb{R}}) \subseteq F(\bar{\mathbb{R}})$. For any level $\alpha,\ 0 \leq \alpha \leq 1,$ the critical value associated with $D_n(\hat{F}_n, F_0; C_0)$ for testing the null hypothesis $\mathscr{H}_0(F)$ as defined by (2.2) is larger than or equal to the critical value associated with $D_n(\hat{F}_n, G; C_0)$ for testing the null hypothesis $\mathscr{H}_0(G)$:*

$$\mathbb{P}_F\left[D_n(\hat{F}_n, F_0; C_0) \geq D_{n\alpha}\right] \leq \alpha \Rightarrow \mathbb{P}_G\left[D_n(\hat{F}_n, G; C_0) \geq D_{n\alpha}\right] \leq \alpha,\ \forall D_{n\alpha}. \tag{5.14}$$

Proposition 5.4 CONFIDENCE BAND RANGE MONOTONICITY. *Let $X_1, \ldots, X_n$ be n i.i.d. observations on X and $\hat{F}_n(x)$ the corresponding empirical distribution function. Let $F(x)$ and $G(y)$ be two distribution functions such that $G(\bar{\mathbb{R}}) \subseteq F(\bar{\mathbb{R}})$. For any level $\alpha,\ 0 \leq \alpha \leq 1,$ the confidence band obtained by inverting the test of the null hypothesis $\mathscr{H}_0(F)$ as defined by equation (2.2) using appropriate critical points with level α for $D_n(\hat{F}_n, G; C_0)$ yields a confidence band for $G(x)$ with level larger than or equal to $1-\alpha$. Equivalently, if C_n is defined as follows:*

$$C_n(\alpha) = \{G : D_n(\hat{F}_n, G; C_0) \leq c_\alpha\}, \tag{5.15}$$

where $c_\alpha = \inf\{D_{n\alpha},\ \mathbb{P}_F\left[D_n(\hat{F}_n, F_0; C_0) \geq D_{n\alpha}\right] \leq \alpha\}$, *then*

$$\mathbb{P}_G\left[G \in C_n(\alpha)\right] \geq 1 - \alpha. \tag{5.16}$$

Propositions 5.3. and 5.4. generalize Propositions 5.1. and 5.2. to all distribution functions. It suggests that CBs can be made narrower by exploiting the nesting of the image sets of different distributions. When studying a discontinuous distribution $G(y)$, we know that $G(y)$ takes its values in a set V^G which is included in $[0,1]$. Thus, the conservative CB for a continuous distribution provides a CB for $G(y)$ with level $1-\delta_1$ greater than or equal to $1-\alpha$. If additional information about the image set of $G(y)$ is available—in particular, if we know there exists a distribution function with image V^F such that $V^G \subseteq V^F$—then the critical points for testing $F(x)$ can be used to derive a CB for $G(y)$ with level $1-\delta_2$ such that $1-\alpha \leq 1-\delta_2 \leq 1-\delta_1$. The CB with level $1-\delta_2$ is narrower than the CB with level $1-\delta_1$ while being reliable. Thus, using information about the nature of the discontinuity of the random variable can be useful for providing shorter CBs for $G(y)$. The more is known about the set of discontinuity points of the distribution, the better the inference will be. However, the improvement can be achieved without knowing all discontinuity points of $G(y)$ and their probability masses. Hence, the main advantage of the KS confidence band, that is its independence of the assumed distribution function, is somehow preserved.

Let's consider a special case of embedded image sets. Let X be a random variable with distribution $H(x) \in \mathscr{F}_{[0,1]}$ such that X is a mixture between a continuous variable bounded on $(0,1]$ and a probability mass of $H(0) \equiv p$ at 0. $H(x)$ is continuous on $(0,1]$ with $H(0) = p$ and $H(1) = 1$.

Corollary 5.5 RANGE MONOTONICITY WITH A MASS AT THE LOWER BOUNDARY. *Let* $X_{21}, \ldots, X_{2n}$ *be n i.i.d. observations on* X_2 *and* $\hat{F}_n(x)$ *the corresponding empirical distribution function. Let* $F_1(x)$ *and* $F_2(x)$ *be two distribution functions continuous on* $(a,b]$ *such that* $p_1 = F_1(a) \leq F_2(a) = p_2$. *For any level* $\alpha,\ 0 \leq \alpha \leq 1$, *the confidence band obtained by inverting the test of the null hypothesis* $\mathscr{H}_0(F_1)$ *as defined by equation* (2.2) *using appropriate critical points with level* α *for* $D_n(\hat{F}_n, F_1; C_0)$ *yields a confidence band for* $F_2(x)$ *with level larger than or equal to* $1-\alpha$. *Equivalently, if* C_n *is defined as follows:*

$$C_n(\alpha) = \{H \in \mathscr{F} : D_n(\hat{F}_n, H_1; C_0) \leq c_\alpha\} \tag{5.17}$$

where $c_\alpha = \inf\{D_{n\alpha},\ \mathbb{P}_{F_1}\left[D_n(\hat{F}_n, F_1; C_0) \geq D_{n\alpha}\right] \leq \alpha\}$, *then*

$$\mathbb{P}_{F_2}\left[F_2 \in C_n(\alpha)\right] \geq 1 - \alpha. \tag{5.18}$$

Corollary 5.5 describes the special case where the distribution function being studied is continuous everywhere except at the lower bound of its support. This case is very interesting because such distribution functions are quite frequent in financial studies, and poverty and inequality analysis. Moreover, this case can be easily extended to those where the discontinuity point is at the upper bound of the support or those where both the lower and the upper bounds are discontinuity points.

6. Order-statistic representations

In this section, we derive expressions for sup-EDF GF statistics which are more convenient for computation than the above definitions. The explicit expressions are based on order statistics.

Let the order statistics of the data $X^{(n)}$ be $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. Let's define $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$, so that $X_{(0)} \leq X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)} \leq X_{(n+1)}$. The empirical distribution function of $X_1, \ldots, X_n$ may then be rewritten as follows:

$$\hat{F}_n(x) = \frac{k}{n} \text{ for } X_{(k)} \leq x < X_{(k+1)}\,, \quad k = 0, 1, \ldots, n; \tag{6.1}$$

see Gibbons and Chakraborti (1992, Theorem 3.1).

Lemma 6.1 PROPOSITION. *Order statistic representation of sup-EDF GF statistics Let X be a random variable with distribution function $F(x) \in \mathscr{F}$, on which n i.i.d. observations $X_1, \ldots, X_n$ are available. Then the statistic $D_n(\hat{F}_n, F_0; C_0)$ defined in* (2.7) *can be rewritten as*

$$D_n(\hat{F}_n, F_0; C_0) = \max_{0 \leq i \leq n} \sup\{\bar{D}_n\left[\hat{F}_n(x), F_0(x)\right] : x \in C_0 \cap \left[X_{(i)}, X_{(i+1)}\right)\}. \tag{6.2}$$

We call (6.2) the *order-statistic representation* of $D_n(\hat{F}_n, F_0; C_0)$. This formula allows one to compute the sup-EDF statistic, through simple optimizations over the intervals determined by the order statistics. In the special case where C_0 is a (semi-open) interval between two order statistics, *i.e.* $C_0 = \left[X_{(j)}, X_{(k+1)}\right)$ for $0 \leq j \leq k \leq n$, we have:

$$D_n(\hat{F}_n, F_0; C_0) = \max_{j \leq i \leq k} \sup\{\bar{D}_n\left[\hat{F}_n(x), F_0(x)\right] : X_{(i)} \leq x < X_{(i+1)}\}\,. \tag{6.3}$$

Further, when the function $\bar{D}_n(\widehat{F}_n, F_0)$ can be expressed as in terms of monotonic functions on intervals $\left[X_{(j)}, X_{(k+1)}\right)$ for $0 \leq j \leq k \leq n$ in the sense defined by equations (6.4) and (6.5) below, then it is possible to derive computable explicit expressions using order statistics. The monotonicity of $\bar{D}_n(\widehat{F}_n, F_0)$ is defined as follows: $\bar{D}_n(\widehat{p}, p)$ is non-increasing in p iff

$$p_1 \leq p_2 <=> \bar{D}_n^+(\widehat{p}, p_2) \leq \bar{D}_n^+(\widehat{p}, p_1) \tag{6.4}$$

and $\bar{D}_n(\widehat{p}, p)$ is non-decreasing in p iff

$$p_1 \leq p_2 <=> \bar{D}_n^+(\widehat{p}, p_1) \leq \bar{D}_n^+(\widehat{p}, p_2)\,. \tag{6.5}$$

Proposition 6.2 ORDER STATISTIC REPRESENTATION OF SUP-EDF GF STATISTICS: MONOTONIC DISCREPANCIES. *Let X be a random variable with distribution function $F(x) \in \mathscr{F}$, on which n i.i.d. observations $X_1, \ldots, X_n$ are available. Suppose $D_n(\hat{F}_n, F_0; C_0)$ defined in* (2.7) *has the form*

$$D_n(\hat{F}_n, F_0; C_0) = \sup\left\{D_n^+, D_n^-\right\} \tag{6.6}$$

where

$$D_n^+(\hat{F}_n, F_0; C_0) = \sup_{x \in C_0} \bar{D}_n^+ \left[\hat{F}_n(x), F_0(x)\right], \quad D_n^- = \sup_{x \in C_0} \bar{D}_n^- \left[\hat{F}_n(x), F_0(x)\right], \tag{6.7}$$

$\bar{D}_n^+ \left[\hat{F}_n(x), F_0(x)\right]$ and $\bar{D}_n^- \left[\hat{F}_n(x), F_0(x)\right]$ are special forms of $\bar{D}_n \left[\hat{F}_n(x), F_0(x)\right]$ such that $\bar{D}_n^+(\widehat{p}, p)$ is non-increasing in p, and $\bar{D}_n^-(\widehat{p}, p)$ is nondecreasing in p. The statistic $D_n(\hat{F}_n, F_0; C_0)$ can be rewritten as follows:

$$D_n(\hat{F}_n, F_0; C_0) = \max_{0 \le i \le n} \sup\{\bar{D}_n^+ \left[i/n, F_0(X_{(i)})\right], \bar{D}_n^- \left[i/n, F_0(X_{(i+1)}-)\right] : x \in C_0 \cap \left[X_{(i)}, X_{(i+1)}\right)\} \tag{6.8}$$

where $F_0(x-) := \lim_{y \uparrow x} F(y) = F_0(x) - p_{F_0}(x)$ and $p_{F_0}(x) := \mathbb{P}_{F_0}(X = x)$.

We study below special cases where the expressions of the statistics further simplify. Let us assume that $C_0 = (-\infty, +\infty)$, we can derive an easily computable order-statistic representation in the case where $D_n(\hat{F}_n, F_0; C_0)$ has the form (6.6):

$$D_n(\hat{F}_n, F_0; C_0) = \max_{0 \le i \le n} \sup\{\bar{D}_n^+ \left[i/n, F_0(X_{(i)})\right], \bar{D}_n^- \left[i/n, F_0(X_{(i+1)}-)\right]\} \tag{6.9}$$

or

$$D_n(\hat{F}_n, F) = \max_{0 \le i \le n} \sup\{\bar{D}_n^+ \left[i/n, F_0(X_{(i)})\right], \bar{D}_n^- \left[i/n, F_0(X_{(i+1)}) - p_F(X_{(i+1)})\right]\}. \tag{6.10}$$

On applying equation (6.9), we get closed-form expressions for sup-EDF GF statistics proposed in Section 3.

Corollary 6.3 ORDER-STATISTIC REPRESENTATIONS OF WALD AND SCORE-TYPE STATISTICS. *The order-statistic representations of the regularized Wald-type and score-type statistics defined in (3.2) and (3.3) are:*

$$D_n^W(\hat{F}_n, F_0; C_0, \delta_n) = \max_{0 \le i \le n} \sup \left\{ \frac{\sqrt{n}\left[(i/n) - F_0(X_{(i)})\right]}{\sqrt{\tilde{\sigma}^2\left[i/n\right] + \delta_n}}, \frac{\sqrt{n}\left[F_0(X_{(i+1)}-) - (i/n)\right]}{\sqrt{\tilde{\sigma}^2\left[i/n\right] + \delta_n}} \right\}, \tag{6.11}$$

$$D_n^S(\hat{F}_n, F_0; C_0, \delta_n) = \max_{0 \le i \le n} \sup \left\{ \frac{\sqrt{n}\left[(i/n) - F_0(X_{(i)})\right]}{\sqrt{\tilde{\sigma}^2\left[F_0(X_{(i)})\right] + \delta_n}}, \frac{\sqrt{n}\left[F_0(X_{(i+1)}-) - (i/n)\right]}{\sqrt{\tilde{\sigma}^2\left[F_0(X_{(i+1)}-)\right] + \delta_n}} \right\}, \tag{6.12}$$

where $C_0 = (-\infty, +\infty)$, $\tilde{\sigma}^2[p] := p(1-p)$ and $F_0(X_{(i+1)}-) = F_0(X_{(i+1)}) - p_F(X_{(i+1)})$.

Setting $\delta_n = 0$ and $C_0 = \left(X_{(1)}, X_{(n)}\right)$ in (6.11) yields the order-statistic representation of the Eicker (1979) statistic defined in (2.18). Similarly, setting $\delta_n = 0$ in (6.12) yields the order-statistic representation of the Anderson and Darling (1952) statistic defined in (2.19). As far as we know, order-statistics representations of the Eicker (1979) and the Anderson and Darling (1952) statistics have not been proposed in the past. Only special cases were incidentally derived, such as for example the expression of the one-sided Eicker (1979) statistic derived by Moscovich, Nadler and

Spiegelman (2016). The latter can be derived as a special case of equation (6.11). Owen (1995) proposed an expression for the Kolmogorov-Smirnov statistic defined in (2.3). However, this expression is valid only for a continuous $F_0(x)$. In the general case where $F_0(x)$ may be noncontinuous, applying equation (6.9) provides the following explicit expression:

$$KS_n = \sqrt{n} \max_{0\le i\le n} \sup \left\{ (i/n) - F_0(X_{(i)}),\, F_0(X_{(i+1)}-) - (i/n) \right\}. \tag{6.13}$$

Equation (6.13) generalizes the expression that is usually given in the literature, which corresponds to the continuous case where $F_0(x-) = F_0(x)$.

Corollary 6.4 ORDER-STATISTIC REPRESENTATION OF LR-TYPE STATISTIC. *The order-statistic representations of the sup-EDF LR-type statistic defined in (*2.16*) is:*

$$D_n^{LR}(\hat{F}_n, F_0; C_0) = \max \left\{ D_n^{LR+}, D_n^{LR-} \right\}, \tag{6.14}$$

$$D_n^{LR+}(\hat{F}_n, F_0; C_0) := \max \left\{ \sup_{0\le i\le n} K\left[i/n, F_0(X_{(i)}) \right], \hat{F}_n(x) \ge F_0(x) \right\}, \tag{6.15}$$

$$D_n^{LR-}(\hat{F}_n, F_0; C_0) := \max \left\{ \sup_{0\le i\le n} K\left[i/n, F_0(X_{(i+1)}-) \right], \hat{F}_n(x) \le F_0(x) \right\}, \tag{6.16}$$

*where $K[\hat{F}_n(x), F_0(x)]$ is defined in (*2.17*) and $F_0(X_{(i+1)}-) = F_0(X_{(i+1)}) - p_F(X_{(i+1)})$.*

Special cases of the expression were given in the literature. (Berk and Jones (1979)) proposed an explicit expression for the one-sided statistic $D_n^{-LR}(\hat{F}_n, F_0; C_0)$ where $C_0 = (-\infty, +\infty)$. Owen (1995) proposed the following numerically less efficient expression for the Berk and Jones (1979) statistic R_n, which we remind is a special case of the LR sup-EDF statistic

$$R_n = \max_{1\le i\le n} \max\{ K[(i-1)/n, F(X_{(i)})], K[i/n, F(X_{(i)})] \}. \tag{6.17}$$

7. Exact Monte Carlo EDF-based tests

In the preceding section, we studied interesting properties for statistics of the form $D_n(\hat{F}_n, F_0; C_0)$. In this section, we show another important advantage of these statistics which make them even more attractive. We show that the test of $\mathscr{H}_0(F)$ as defined by equation (2.2) based on these statistics can be implemented using exact randomized test procedures such as Monte Carlo tests [see Dwass (1957), Dufour (2006), Dufour and Kiviet (1998), and Dufour and Khalaf (2001)]. Given that $D_n(\hat{F}_n, F_0; C_0)$ is pivotal under $\mathscr{H}_0(F)$, the Monte Carlo test based on pivotal statistics can be applied.

Given that the distribution of $D_n(\hat{F}_n, F_0; C_0)$ is noncontinuous, the standard Monte Carlo test procedure cannot be applied. In this case, a randomized tie-breaking procedure [Dufour (2006)] can be applied to test $\mathscr{H}_0(F)$. This procedure is a modified Monte Carlo test adapted for discrete distributions. It can be implemented as follows.

Let D_n denote the test statistic computed from data and D_{n0} the observed value of D_n based on specific realized data. The critical region of the test is $D_{n0} \geq D_{n\alpha}$ where α is the level of the test and $G(D_{n0})$ and $G(x) = \mathbb{P}_F[D_n \geq x]$ is the realized p-value of the test statistic D_{n0}. Suppose we can generate N i.i.d. replications $D_{nj},\ j = 1, \ldots, N$ of $D_n(\hat{F}_n, F_0; C_0)$ under $\mathscr{H}_0(F)$. The following steps apply:

1. Draw $N+1$ i.i.d. variates $W_0, W_1, \ldots, W_N$ independently of D_{nj}.

2. Order the pairs (D_{nj}, W_j) following the lexicographic criterion:

$$(D_{ni}, W_i) \geq (D_{nj}, W_j) \Leftrightarrow \{D_{ni} > D_{nj} \text{ or } (D_{ni} = D_{nj} \text{ and } W_i \geq W_j)\}. \tag{7.1}$$

3. Compute an empirical p-value function:

$$\widetilde{p}_N(x) = \frac{N\widetilde{G}_N(x) + 1}{N+1}, \tag{7.2}$$

$$\widetilde{G}_N(x) := 1 - \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}_{[0,\infty)}(x - D_{nj}) + \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}_{[0]}(D_{nj} - x)\, \mathbf{1}_{[0,\infty)}(W_j - W_0). \tag{7.3}$$

   $\widetilde{p}_N(x)$ is the empirical probability that a value as extreme or more extreme than x is realized if $\mathscr{H}_0(F)$ is true. Note that $N\widetilde{G}_N(x)$ is the number of simulated statistics which are greater or equal to x.

4. Compute the associated Monte Carlo (randomized) critical region:

$$\widetilde{p}_N(D_{n0}) \leq \alpha,\ 0 \leq \alpha \leq 1. \tag{7.4}$$

   where $\widetilde{p}_N(D_{n0})$ may be interpreted as an estimate of $G(D_{n0})$. $\widetilde{p}_N(D_{n0})$ can be interpreted as a realized Monte Carlo p-value associated with D_{n0}. Thus if N is chosen such that $\alpha(N+1)$ is an integer, the critical region (7.4) has the same size as the critical region $G(D_{n0}) \leq \alpha$. Moreover,

$$\mathbb{P}[\widetilde{p}_N(D_{n0}) \leq \alpha] = \frac{I[\alpha(N+1)]}{N+1},\ 0 \leq \alpha \leq 1 \tag{7.5}$$

   where $I[x]$ is the integer part of x.

Thanks to the properties of the Monte Carlo test, the implementation of the studied goodness-of-fit tests along this procedure is convenient. First, if $F(x)$ is free of nuisance parameters and $\alpha(N+1)$ is an integer then the properties of the critical region are irrespective of the number of replications used. Second, the Monte Carlo test does not need to compute exact critical points for each sample size and each distribution function under test. Besides, the number of replications needed to compute the test is not constraining as its the number of replications needed to simulate valid critical points or to get accurate bootstrap results. Third, confidence intervals can be built from such tests using inversion procedures [see Fieller (1940, 1954), for example].

8. Confidence bands for distribution functions

In this section, we derive exact confidence bands for a distribution $F_0(x)$ by inverting the sup-EDF based GF tests proposed in this paper. Given the tests' properties derived in section 4, we obtain a set of exact confidence bands for a distribution with several desirable properties. We suppose $C_0 = (-\infty, +\infty)$ for the rest of the section.

Let $c_W(\alpha; \delta_n)$ be the critical point of the regularized Wald-type statistic such that $\mathbb{P}[D_n^W(\hat{F}_n, F_0; \delta_n) \leq c_W(\alpha; \delta_n)] \geq 1 - \alpha$. Given the sup form of $D_n^W(\hat{F}_n, F_0; \delta_n)$, it is straightforward that:

$$\mathbb{P}_{F_0}\left[\left|\hat{F}_n(x) - F_0(x)\right| \leq \frac{c_W(\alpha; \delta_n)}{\sqrt{n}}(\tilde{\sigma}^2\left[\hat{F}_n(x)\right] + \delta_n)^{1/2}, \forall x\right] \geq 1 - \alpha \tag{8.1}$$

where $\tilde{\sigma}^2[p] := p(1-p)$. Hence, the set $C_W(\alpha; \delta_n)$ defined as follows is a confidence band with level $1 - \alpha$ for $F_0(x)$:

$$C_W(\alpha; \delta_n) = \{F_0 \in \mathscr{F} : C_W^L(\alpha; \delta_n) \leq F_0(x) \leq C_W^U(\alpha; \delta_n),\ \forall x\} \cap [0, 1] \tag{8.2}$$

where

$$C_W^L(\alpha; \delta_n) = \hat{F}_n(x) - \frac{c_W(\alpha; \delta_n)}{\sqrt{n}}\left[\tilde{\sigma}^2\left[\hat{F}_n(x)\right] + \delta_n\right]^{1/2}, \tag{8.3}$$

$$C_W^U(\alpha; \delta_n) = \hat{F}_n(x) + \frac{c_W(\alpha; \delta_n)}{\sqrt{n}}\left[\tilde{\sigma}^2\left[\hat{F}_n(x)\right] + \delta_n\right]^{1/2}. \tag{8.4}$$

Similarly, the confidence band associated to the regularized EDF-based Score-type statistic is:

$$C_S(\alpha; \delta_n) = \left\{F_0 \in \mathscr{F} : C_S^L(x; \delta_n) \leq F_0(x) \leq C_S^U(x; \delta_n),\ \forall x\right\} \cap [0, 1] \tag{8.5}$$

$$C_S^L(x; \delta_n) := \frac{2\hat{F}_n(x) + \frac{c_S^2(\alpha; \delta_n)}{n} - \sqrt{\Delta}}{2\left(1 + \frac{c_S^2(\alpha; \delta_n)}{n}\right)},\ C_S^U(x; \delta_n) := \frac{2\hat{F}_n(x) + \frac{c_S^2(\alpha; \delta_n)}{n} + \sqrt{\Delta}}{2\left(1 + \frac{c_S^2(\alpha; \delta_n)}{n}\right)}, \tag{8.6}$$

$$\Delta := \left[2\hat{F}_n(x) + \frac{c_S^2(\alpha; \delta_n)}{n}\right]^2 - 4\left[1 + \frac{c_S^2(\alpha; \delta_n)}{n}\right]\left(\hat{F}_n^2(x) - \frac{\delta_n c_S^2(\alpha; \delta_n)}{n}\right), \tag{8.7}$$

where $c_S(\alpha; \delta_n)$ is the critical point of the regularized Score-type statistic such that $\mathbb{P}[D_n^S(\hat{F}_n, F_0; \delta_n) \leq c_S(\alpha; \delta_n)] \geq 1 - \alpha$. Setting $\delta_n = 0$ in equation (8.2) and equation (8.5) yields the confidence bands based on the Eicker(1979) and Anderson-Darling(1952) tests respectively.

The regularized Wald-type and Score-type confidence bands present desirable properties inherited from the underlying GF tests. First, owing to the structure of the weights used by the Wald-type and the Score-type statistics, the widths of the corresponding CBs decrease with observations further from the center of the distribution, which improves their performance in the tails of distributions. This is a desirable feature compared to the KS confidence band, which width is uniform. Second, owing to the pivotality of the underlying statistics, the confidence bands can be computed using the same critical points for all continuous distribution functions. This property simplifies greatly their

computation. To end, these confidence bands benefit from similar monotonicity properties than the sup-EDF based GF tests, when applied to distribution functions with embedded image sets.

Regarding the LR-type GF test, the related confidence band with level $1-\alpha$ can be built using the following procedure [Owen (1995)]:

$$C_F^{LR}(\alpha) = \{F_0 \in \mathscr{F} : G_n^L(x) \leq F_0(x) \leq G_n^U(x),\ \forall x\}, \tag{8.8}$$

$$G_n^L(x) = \min\{p : K[\hat{F}_n(x),\, p] \leq c_{LR}(\alpha;n)\},\ G_n^U(x) = \max\{p : K[\hat{F}_n(x),\, p] \leq c_{LR}(\alpha;n)\}, \tag{8.9}$$

$$G_n^U(x) = \max\{p : K[\hat{F}_n(x),\, p] \leq c_{LR}(\alpha;n)\}, \tag{8.10}$$

$$K(\widehat{p}\,,\, p) = \widehat{p}\log\,(\frac{\widehat{p}}{p}) + (1-\widehat{p})\log\,(\frac{1-\widehat{p}}{1-p}), \tag{8.11}$$

where $c_{LR}(\alpha;n)$ satisfies $\mathbb{P}[D_n^{LR}(\hat{F}_n, F_0; 0, C_0) \leq c_{LR}(\alpha;n)] \geq 1-\alpha$. Although useful, the LR-type EDF-based confidence band suffers from important drawbacks compared with the Score-type and Wald-type statistics. Indeed, the LR-type confidence band is computed by performing a likelihood ratio test on the distribution of $\hat{F}_n(x)$, and retaining only candidates $F(x)$ with sufficiently large likelihood for each observation x. Consequently, it carries an important computational cost: building the confidence band requires to perform n optimizations and the performance of the method depends greatly on the performance of the used optimization procedure. Moreover, the large number of optimizations may be time-costly when n is large (large samples).

Other EDF-based statistics for GF tests can also be used to build exact CBs for continuous distribution functions. However, even though the GF tests performed using these statistics may not be difficult to compute, the corresponding CBs generally do not have explicit expressions and must be computed numerically. This is the case of the statistics of (the integral version of) Anderson and Darling (1952), Cramér (1928), and von Mises (1931).

9. Simulations

We illustrate how to operate the EDF-based tests and confidence bands using simulations. The critical points are independent of the observations of the test samples, thanks to their pivotality, which allows us to simulate them using the empirical distribution function of the sample. The critical points are however contingent on the sample size n. Moreover, the critical points of the regularized EDF-based statistics depend on the regularization parameter δ_n.We discuss below how to choose δ_n such that the methods deliver quality inference. For the non-regularized statistics, their critical points are tabulated and the literature proposed approximation formula for some of them.

9.1. Choice of δ_n

Adding a regularization to the statistics improves test performance by controlling the test behavior in the tails of the distribution. However, the magnitude of performance improvement depends on δ_n, which form and value should be fully specified to allow building the tests and confidence bands.

The form and value of the regularization parameter should be carefully chosen for several

reasons. First, it is important that the form of δ_n does not alter the sup form of the statistics $D_n(\hat{F}_n, F_0; \delta_n, C_0)$, which is the foundation of their convenient properties discussed above. Second, the value of δ_n should be estimated independently from the sample $X_1, \ldots, X_n$. The distribution of δ_n, when it is not independent of $X_1, \ldots, X_n$, may modify the distribution of D_n, thus complicating the test implementation. Notably, if the critical points are no longer pivotal and conservative, computing the tests would require computing critical points for each distribution function, which adds a time-consuming extra step to the computation process. Third, given that the value of δ_n influences the test performance, it is preferable to estimate it using a sample with similar properties than $X_1, \ldots, X_n$.

There are several ways of estimating the regularization parameter, while satisfying the constraints discussed above. We propose and illustrate one methodology below. As a starting point, let's assume without any loss of generality that δ_n is constant. We propose to use a split sample procedure (see Dufour and Jasiak, 2001) to estimate the optimal value of δ_n, bearing in mind the second and third constraints discussed above.

The procedure is as follows. Step 1: randomly draw a subsample from the original one without replacement. This sample, called auxiliary sample, will be used to estimate the optimal value of δ_n out of sample. The rest of the sample will be used to build the tests and confidence bands based on D_n,conditionally to the value of δ_n estimated on the auxiliary sample. There is no clear recommendations about the optimal size of the auxiliary sample. Common practice is to use a small part, up to 10 percent, of the initial sample (Dufour, 2001). However, given that the performance of the regularized tests depends greatly on the proper estimation of δ_n, we suggest to use 20 percent of the original sample, provided that the initial sample is large enough. With this out-of-sample approach, the auxiliary and estimation samples are independent, and the statistics' distributions are unchanged by the estimation of δ_n, which guarantees the validity of the tests.

Step 2: Use the auxiliary sample, to assess the tests performance for several values of δ_n and choose the optimal value. An appropriate criterion should be picked to estimate δ_n on the auxiliary sample, depending on the objectives of the inference. For example, if F, the distribution assumed under the null hypothesis $\mathscr{H}_0(F)$, has a shape similar to a uniform distribution, the criterion could be to minimize the simple average of the confidence band widths at each observation of the auxiliary sample. By contrast, if F has a heavy tail, the criterion should assign more weights to observations in the tails of the auxiliary sample than those in the center. If no information is available about the distribution, we suggest choosing the minimum value of δ_n that increases sufficiently the power of the test, compared to the case $\delta_n = 0$. This makes sense as the smaller the value δ_n, the smaller the widths of the confidence bands in the tails of the sample. Note that using exact critical points ensures that the test levels are controlled. The value of δ_n that maximizes the power of the goodness of fit tests also controls the widths of the corresponding confidence bands (Pratt (1963)).

9.2. EDF-based tests with heavy-tail distributions

This section illustrates the relative performance of the sup-EDF goodness-of-fit tests, captured by their level and power, the impact of the regularization, and the performance of EDF-based confidence bands, using Monte Carlo simulations. The test levels are controlled using exact critical points simulated with $N_1 = 3000000$ replications for each value of δ_n. The study focuses thus mainly on

comparing the power of the tests.

Simulations illustrate the performance of tests on a distribution with heavy tails, illustrated using the Singh-Maddala distribution (or Burr Type XII distribution) proposed by Singh and Maddala (1976) and Burr (1942). These distributions mimics closely income and consumption distributions, which is are the type of data we are interested in studying. An analysis of the tests' behavior on a Normal distribution, which is a more standard distribution, yields similar results (see Appendix). The Singh-Maddala cumulative distribution function Burr(a,c,k) is

$$F(x;c,k,a) = 1-\left[1+(x/a)^c\right]^{-k}. \tag{9.1}$$

where $a>0$ is the scale parameter, $c>0$ is the first shape parameter, and $k>0$ is the second shape parameter.

As a first simulation, let's consider the test of $\mathscr{H}_0 : X \sim \text{Burr}(a_0 = 1, c_0 = 2, k_0 = 5)$ against the alternative $\mathscr{H}_1 : X \sim \text{Burr}(a_1, c_1, k_1)$. The level and power of the tests are simulated using a relatively small sample size $n = 500$ to demonstrate finite sample properties and $N_2 = 15000$ Monte Carlo replications, chosen larger than recommended by the literature (Mundform, Schaffer, Kim, Shaw and Thongteeraparp (2011)) to ensure strong confidence in our results. We analyze the behavior of the tests using six alternatives. For each one of the three Singh-Maddala parameters, we pick two alternatives: one alternative with a value of the parameter above its value under the null hypothesis value and a second alternative with a parameter value below its value under the null. The simulations uses exact critical points tabulated in Table A–1 (see Appendix).

The power of the regularized Score- and Wald-type tests improves greatly with increasing values of δ_{500} (see Tables 1 and 2 below). The power of the regularized-Score test is relatively high, even for small values of $\delta_{500}(S)$ starting at $\delta_{500}(S) = 0.001$, as the latter prevents the denominator of the statistics from converging to zero in the tails of the sample. The power continues to improve with values of $\delta_{500}(S)$, culminating between 81 to 95 percent for $\delta_{S,500}$ ranging from 0.01 to 0.25, depending on the alternative hypothesis. Similarly, the power of the regularized Wald-type test increases with $\delta_{500}(W)$. The power reaches relatively high levels, even for small values of $\delta_{500}(W)$ starting at $\delta_{500}(W) = 0.005$. It continues to improve with values of $\delta_{500}(W)$, culminating in the range of 81 to 97 percent for $\delta_{500}(W) = 0.01$ to 0.25 depending on the alternative hypothesis.

We assume for the remainder of the study that $\delta_{500}(S)$ and $\delta_{500}(W)$ could range from 0.01 to 0.25 to deliver tests with reasonably high power.

As a second simulation, we compare the performance of the regularized-EDF based tests to the performance of the other EDF-based tests. Table 3 shows that amongst all the tests, the likelihood-ratio based test achieves the best test, followed by the regularized Score-type test, the Kolmogorov-Smirnov test, and the regularized Wald-type test. The non-regularized Score and Wald tests have much less power than the regularized tests, demonstrating once more the usefulness of our proposed methods.

9.3. EDF-based confidence bands

As a third simulation, we analyze the relative performance of the EDF-based confidence bands for a Burr$(a = 1, c = 2, k = 5)$ distribution with sample size $n = 500$. Based on the test power, we could

Table 1: Power of the regularized-Score (Anderson-Darling) tests for different values of δ_n (in percent)

Test of $H_0 : X \sim \text{Burr}(a_0 = 1, c_0 = 2, k_0 = 5)$ **versus** $H_1 : X \sim \text{Burr}(a_1, c_1, k_1)$

	$c_1 = 2, k_1 = 5$		$a_1 = 1, k_1 = 5$		$a_1 = 1, c_1 = 2$	
δ_{500}	$a_1 = 0.9$	$a_1 = 1.1$	$c_1 = 1.8$	$c_1 = 2.2$	$k_1 = 4.1$	$k_1 = 5.9$
0.0001	16.48	5.08	1.46	18.18	4.88	23.28
0.0005	20.07	5.18	1.42	20.85	5.15	26.94
0.001	42.88	10.20	7.42	34.52	15.62	44.74
0.005	89.26	68.52	72.47	79.28	84.51	84.88
0.01	92.50	77.31	81.32	84.45	90.57	87.91
0.05	94.11	84.96	88.26	86.78	94.09	89.06
0.07	94.45	85.47	89.02	86.55	94.68	88.43
0.1	94.48	85.88	89.75	86.42	94.32	88.34
0.15	93.85	86.27	89.78	86.02	94.39	87.47
0.2	93.98	86.51	89.93	85.70	94.55	87.25
0.25	94.01	85.95	89.96	86.05	94.51	86.50
0.3	93.52	85.92	89.96	85.78	94.37	86.27
0.35	93.66	86.47	90.12	85.45	94.46	86.63
0.4	93.57	86.38	90.13	84.84	94.51	86.01
0.7	93.58	86.19	89.92	85.01	94.40	85.33
1	93.46	86.66	90.02	84.69	94.36	85.15
5	93.20	86.63	89.68	83.98	94.22	84.40
10	93.04	85.95	90.09	84.06	94.16	84.30

Sample size: $n = 500$; level: $\alpha = 0.05$; number of Monte Carlo simulations: $N_2 = 15000$; the critical value $c_S(\alpha; \delta_n)$ for D_n^S [in (3.8)] is obtained by simulation ($N_1 = 3000000$ replications).

use values of $\delta_{500}(S)$ and $\delta_{500}(W)$ ranging from 0.01 to 0.25. However, higher values of $\delta_{500}(S)$ and $\delta_{500}(W)$ yield confidence bands of larger widths in the tails of the distribution. Hence, lower values of regularization parameters are more desirable to ensure strong performance in the tails of the distribution and those values of regularization parameters may not yield the highest test power. There is a trade-off between optimizing the performance of the regularized confidence bands and maximizing the power of the regularized tests. We simulate the regularized confidence bands using values of parameters $\delta_{500}(S) = 0.05$ and $\delta_{500}(W) = 0.01$ (Figures 1 and 2).

The figures show that truncating the Kolmogorov-Smirnov (KS) and Eicker (Wald) confidence bands ensures that they stay within $[0,1]$, improving their performance. Thanks to the truncation, the width of the truncated Eicker (Wald) band converges to zero in the tails of the sample while the original Eicker confidence band goes beyond the $[0,1]$ interval. The Kolmogorov-Smirnov confidence band also goes beyond $[0,1]$ and is improved by the truncation. On the contrary, the Anderson-Darling (Score) confidence band is always continuous and stays within the $[0,1]$ boundary.

For observations in the center of the distribution, the regularized-Wald confidence band has the smallest width, followed by the Kolmogorov-Smirnov band, the regularized-Score band, and the

Table 2: Power of the regularized Wald-type (Eicker) test for different values of δ_n (in percent)

Test of $H_0 : X \sim \text{Burr}(a_0 = 1, c_0 = 2, k_0 = 5)$ **versus** $H_1 : X \sim \text{Burr}(a_1, c_1, k_1)$

	$c_1 = 2, k_1 = 5$		$a_1 = 1, k_1 = 5$		$a_1 = 1, c_1 = 2$	
δ_{500}	$a_1 = 0.9$	$a_1 = 1.1$	$c_1 = 1.8$	$c_1 = 2.2$	$k_1 = 4.1$	$k_1 = 5.9$
0.0001	52.888	60.096	71.82	32.48	87.288	37.188
0.0005	79.992	80.08	86.944	61.876	95.18	67.252
0.001	84.364	84.544	89.384	68.984	96.268	73.988
0.005	90.328	87.848	91.832	77.732	97.124	81.916
0.01	91.472	88.44	91.976	80.46	97.14	83.448
0.05	92.708	88.848	91.96	83.54	96.492	85.704
0.07	93.236	88.564	91.92	83.828	96.412	85.68
0.1	93.536	88.436	92.084	84.052	95.94	85.732
0.15	93.032	88.208	91.556	84.192	95.728	85.548
0.2	93.344	88.076	91.292	84.06	95.456	85.496
0.25	93.492	87.336	91.124	84.716	95.416	85.136
0.3	92.988	87.112	91.06	84.716	95.18	85.092
0.35	93.096	87.472	91.024	84.308	95.1	85.424
0.4	93.088	87.252	90.96	83.996	95.152	85.044
0.7	93.344	86.756	90.412	84.356	94.836	84.556
1	93.22	86.952	90.388	84.148	94.664	84.688
5	93.176	86.68	89.788	83.896	94.256	84.288
10	93.024	85.98	90.14	83.992	94.204	84.272

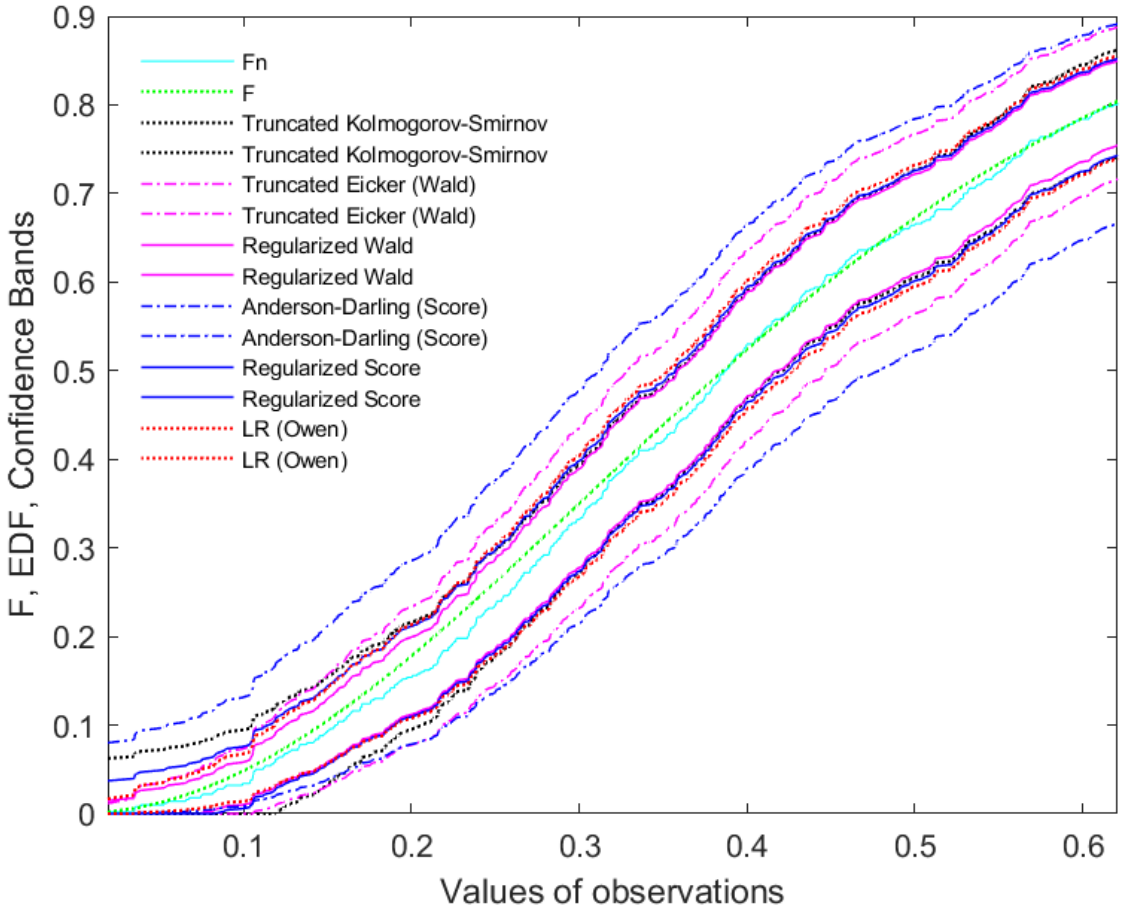
Sample size: $n = 500$; level: $\alpha = 0.05$; number of Monte Carlo simulations: $N_2 = 15000$; the critical value $c_W(\alpha; \delta_n)$ for D_n^W [in (3.7)] is obtained by simulation ($N_1 = 3000000$ replications).

LR band. The truncated Eicker (Wald) and Anderson-Darling (Score) confidence bands have the largest widths, much larger than the widths of the regularized confidence bands.

The relative performances of the inference methods change substantially in the tails of the sample. The LR and the regularized-Wald confidence bands perform generally the best, followed by the regularized-Score band. It follows the truncated Eicker confidence band, the truncated KS band, and the truncated Anderson-Darling; the latter has the largest width amongst the studied confidence bands.

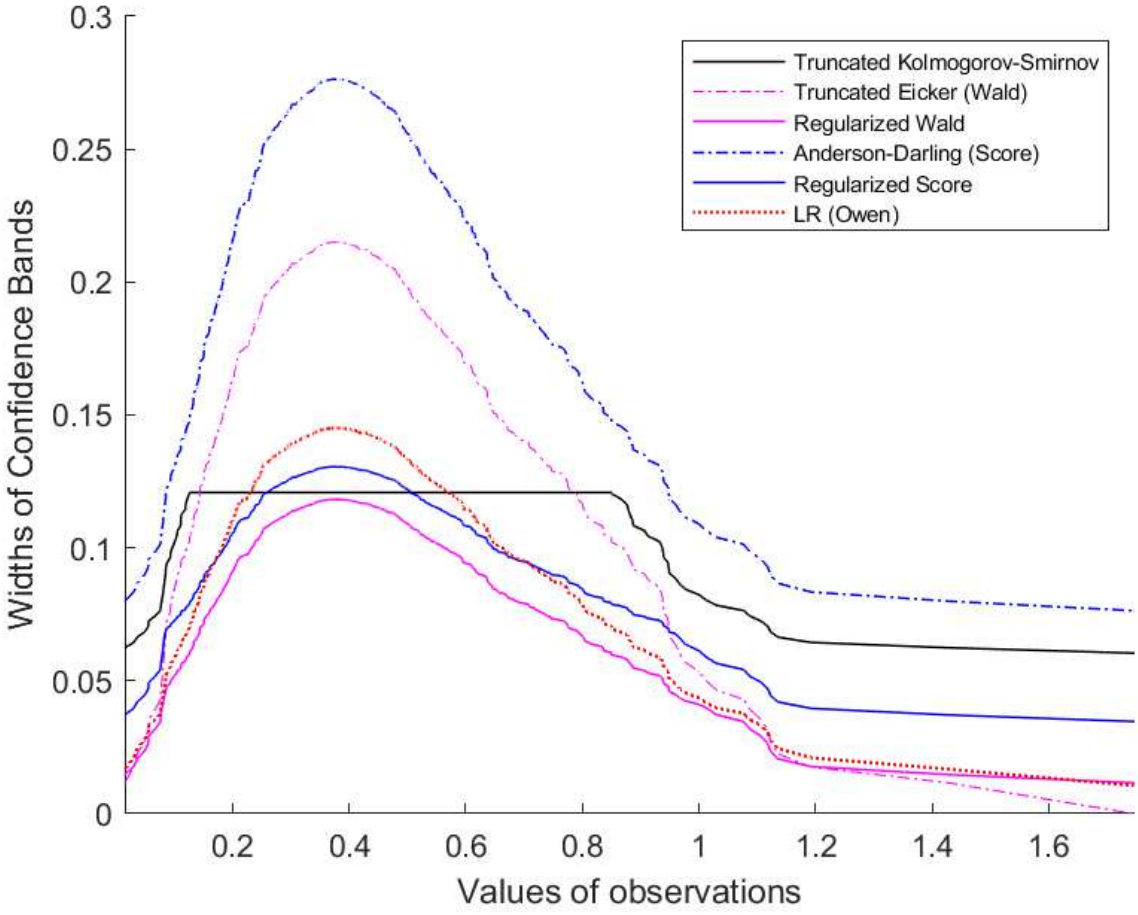
A couple of additional general conclusions emerging from the simulations. First, the regularized-Wald and -Score confidence bands have smaller widths than the non–regularized ones throughout the sample, hence a better inference quality. Regularizing the Eicker (Wald) statistic corrects the discontinuity of the corresponding confidence band. Although for δ_n different from zero, the width of the regularized-Wald band does not converge to zero anymore, the band performs well, as shown by the simulations. Likewise, regularizing the Score statistic improves a lot the performance of the corresponding band both throughout the sample.

Figure 1. EDF-based confidence bands for the Singh-Maddala distribution function Burr$(a=1,c=2,k=5)$ **for** $\delta_{500}(S)=0.05$ **and** $\delta_{500}(W)=0.01$



Sample size: $n=500$, level: $1-\alpha=0.95$, $\delta_{500}(S)=0.05$ for the regularized Score-type test, $\delta_{500}(W)=0.01$ for the regularized Wald-type test, and number of replications to simulate critical points: $N_1=1000000$.

Figure 2. Widths of EDF-based confidence bands for the Singh-Maddala distribution function Burr$(a=1,c=2,k=5)$ **for** $\delta_{500}(S)=0.05$ **and** $\delta_{500}(W)=0.01$



Sample size: $n=500$, level: $1-\alpha=0.95$, $\delta_{500}(S)=0.05$ for the regularized Score-type test, $\delta_{500}(W)=0.01$ for the regularized Wald-type test, and number of replications to simulate critical points: $N_1=1000000$.

Table 3: **Level and power of EDF-based tests,** $n = 500$

Test of

$H_0 : X \sim \text{Burr}(a_0 = 1, c_0 = 2, k_0 = 5)$ **versus** $H_1 : X \sim \text{Burr}(a_1 = 1, c_1 = 1.8, k_1 = 5)$

Tests	$\mathbf{c}(\alpha)$	**Level (percent)**	**Power (percent)**
Kolmogorov-Smirnov	0.060415	4.886667	89.686667
Wald	4.808473	4.813333	19.466667
Regularized Wald	3.230072	4.793333	81.200000
Score	6.447926	4.906667	17.773333
Regularized Score	2.683674	4.706667	91.986667
Berk-Jones	0.011090	5.106667	94.766667

$H_0 : X \sim \text{Burr}(a_0 = 1, c_0 = 2, k_0 = 5)$ **versus** $H_1 : X \sim \text{Burr}(a_1 = 1, c_1 = 2, k_1 = 4.1)$

Tests	$\mathbf{c}(\alpha)$	**Level (percent)**	**Power (percent)**
Kolmogorov-Smirnov	0.060415	5.0466667	94.060000
Wald	4.808473	5.36	32.260000
Regularized Wald	3.230072	5.073333	90.093333
Score	6.447926	5.3	30.540000
Regularized Score	2.683674	5.106667	96.293333
Berk-Jones	0.011090	5.866667	98.946667

Sample size: $n = 500$; level: $1 - \alpha = 0.95$; $\delta_{500}(S) = 0.05$ for the regularized Score-type test; $\delta_{500}(W) = 0.01$ for the regularized Wald-type test; number of replications to simulate critical points: $N_1 = 3000000$; number of Monte Carlo simulations: $N_2 = 15000$.

Second, comparing the regularized Wald- and Score-based confidence bands to the LR-based band (Owen), it follows that the LR confidence band performs worse than the regularized bands in the center of the distribution, but close to the regularized-Eicker bands in the tails of the distribution. However, the Owen CB has a major drawback: it is very computationally demanding. Building the Owen CB requires to perform as many optimizations as there are observations, which is resource-intensive and takes a long time compared to the regularized bands (Table 4), more than 2000 times the time needed to compute the regularized-Wald band.

Third, comparing the surface between the lower and the upper bounds of the confidence bands, the sum of widths of confidence bands over all observations, show that the regularized-Wald method is the best performer under this metric, followed by the regularized-Score method (Table 5) and the LR confidence band. Despite improvements from the truncation, the truncated Kolmogorov-Smirnov and Eicker confidence bands yield larger surface than the three above methods. The Anderson-Darling confidence band has the largest surface.

Fourth and last, the value of the regularization plays an important role and, when choosing its value, there is a trade-off between the performance of the tests and that of the confidence bands. Simulations using values of parameters $\delta_{500}(S) = 0.001$ and $\delta_{500}(W) = 0.07$ (see Figures A–1 and A–2 in Appendix) show that in this case, the Kolmogorov-Smirnov band has the smallest width

Table 4: **Computation time of EDF-based confidence bands for the Singh-Maddala distribution function** Burr$(a=1,c=2,k=5)$**,** $n=500$

Confidence bands	Time (in percent of the regularized-Wald time)
Kolmogorov-Smirnov	0.4
Eicker (Wald)	0.9
Anderson-Darling (Score)	1.8
Regularized-Wald	1.0
Regularized-Score	2.3
LR (Owen)	2883.2

Sample size: $n=500$; level: $1-\alpha=0.95$; $\delta_{500}(S)=0.05$ for the regularized Score-type test; $\delta_{500}(W)=0.01$ for the regularized Wald-type test, and number of replications to simulate critical points: $N_2=1000000$

for observations in the center of the distribution, followed by the regularized-Wald band, the LR band, and the regularized-Score band. With this simulation [$\delta_{500}(S)=0.001$ and $\delta_{500}(W)=0.07$] the widths of the regularized bands are much larger than for $\delta_{500}(S)=0.05$ and $\delta_{500}(W)=0.01$, although the latter values do not maximize the power of the regularized tests. This illustrates the trade-offs between performances of the tests and performances of confidence bands in choosing the values of the regularization parameters, which users have to balance, depending on the objectives of the study.

Table 5: Surface of EDF-based confidence bands for the Singh-Maddala distribution function Burr$(a=1,c=2,k=5)$**,** $n=500$

Confidence bands	Surface
Kolmogorov-Smirnov	58.8369
Truncated Eicker	84.0887
Anderson-Darling	112.181
Regularized-Wald	47.1162
Regularized-Score	54.495
LR (Owen)	57.1345

Sample size: $n=500$, level: $1-\alpha=0.95$, $\delta_{500}(S)=0.05$ for the regularized Score-type test, $\delta_{500}(W)=0.01$ for the regularized Wald-type test, and number of replications to simulate critical points: $N_2=1000000$

10. Application to household consumption in Kenya

This section illustrates how to use the EDF-based inference methods on a data sample. We utilize household consumption data from Kenya's 2015-2016 Household Survey (source: Kenya National

Table 6: Kenya: Households Consumption, 2015-16
Summary statistics (KES)

Average	Median	Minimum	Maximum	Standard deviation
6,625.1	4,954.7	9.4	552,741.8	7372.8

Sources: Kenya National Bureau of Statistics and authors' calculations.

Bureau of Statistics) and use monthly per adult equivalent total consumption expenditure (deflated). The total size of the sample is $n_T = 21773$. For the purpose of this illustration, we assume $\delta_{n_T}(W) = 0.05$ for the regularized Wald-type test and $\delta_{n_T}(S) = 0.001$ for the regularized Score-type test, which are values of the parameters that deliver high enough power of the regularized Wald and the regularized Score tests as per the above simulations.

Kenya's households consumption (expressed in Kenyan Shillings or KES) displays substantial dispersion and a heavy right tail (Table 6). Average monthly consumption in Kenyan households is $KES6625$, while half of Kenyan households consume less than $KES4955$ (median). The minimum value is $KES9$ and the maximum value is $KES552742$, pointing to heterogeneity. In particular the upper tail is sizeable: about 15 percent of households consume more than $KES10,644$ and the top one percent consume more than $KES28326$. The heavy tails of the distribution make the application especially relevant for the methodologies proposed in the paper and, in fine, for poverty and inequality analyses.

Using the formula of the confidence bands derived above and the values of regularization parameters $\delta_{n_T}(S) = 0.001$ and $\delta_{n_T}(W) = 0.05$, we build EDF-based confidence bands for the underlying theoretical distribution of households' consumption using simulated critical points for each EDF statistic and $n = n_T$ (Table A–6, Appendix). Figure A–10 (Appendix) shows the empirical distribution of the households' consumption and the estimated confidence bands for the whole sample and Figures 3, 4, and 5 zoom into the left tail, the center, and the right tail of the distribution (respectively), with smaller number of observations to allow better comparing the EDF-based confidence bands. Figure A–11 (Appendix) compares the widths of the confidence bands for each observation and Figure 6 plots the widths of confidence bands at smaller observations. Several conclusions emerge.

Truncating the Kolmogorov-Smirnov and Eicker confidence bands to the interval $[0,1]$ improves their performance in the tails of distributions. The illustration shows that their lower bands go below 0 while their upper band can be higher than 1, allowing for values of distribution function that are not permissible by definition (see Figures A–6, A–7, A–8, and A–7 in Appendix). We propose to truncate the confidence bands to $[0,1]$, which improves the performance of the confidence bands at a non-negligible number of observations (200 observations in each tail for the Kolmogorov-Smirnov confidence band and 23 observations in each tail for the Eicker statistic). We use the truncated Kolmogorov-Smirnov and Eicker confidence bands in the rest of the illustration. On the contrary, the Anderson-Darling confidence band remains within the boundaries $[0,1]$ and does not require improvement.

The regularized-Score and -Wald confidence bands perform generally better than the non-

regularized ones, exhibiting smaller widths. The regularized-Score confidence band delivers a much smaller width than the Anderson-Darling (Score) confidence band throughout. The regularized-Wald confidence band performs better than the truncated Eicker (Wald) confidence band, except for the smallest observations (493 first observations). The relative performances are unchanged when using the non-truncated Eicker band.

Comparing all confidence bands, the regularized-Score confidence band performs best amongst all the bands on most of the observations, followed by the Owen confidence band, although both bands perform consistently well throughout the sample, whereas the Kolmogorov-Smirnov and the regularized Wald confidence band perform best in the center of the distribution. The widths of the regularized Score and Owen confidence bands are the smallest ones amongst the confidence bands, except in the center of the distribution where they are outperformed by the Kolmogorov-Smirnov band followed by the regularized-Wald band. The regularized Score band exhibit a smaller width than that of the Owen band throughout the sample except for the first 253 observations. In the left tail (for smaller observations), the truncated Eicker band has the third smallest width, followed by the Anderson-Darling band and the regularized-Wald confidence band. The truncated Kolmogorov-Smirnov CB delivers the largest width. As the values of the observations increase, the regularized-Wald confidence band improves, delivering the smallest width, followed by the regularized Score confidence band, except towards the center of the distribution, where the truncated Kolmogorov-Smirnov confidence band provides the smallest width–though only for $3/7$ of the sample (42.9 percent of the sample or $9,333$ observations, from observation number $6,220$ to $15,553$). In the right tail (for larger values of observations), the regularized-Wald confidence band delivers the third smallest width followed by the truncated Wald confidence band, the Anderson-Darling band, and the truncated Kolmogorov-Smirnov confidence band. The latter has the widest width amongst all the confidence bands in both tails.

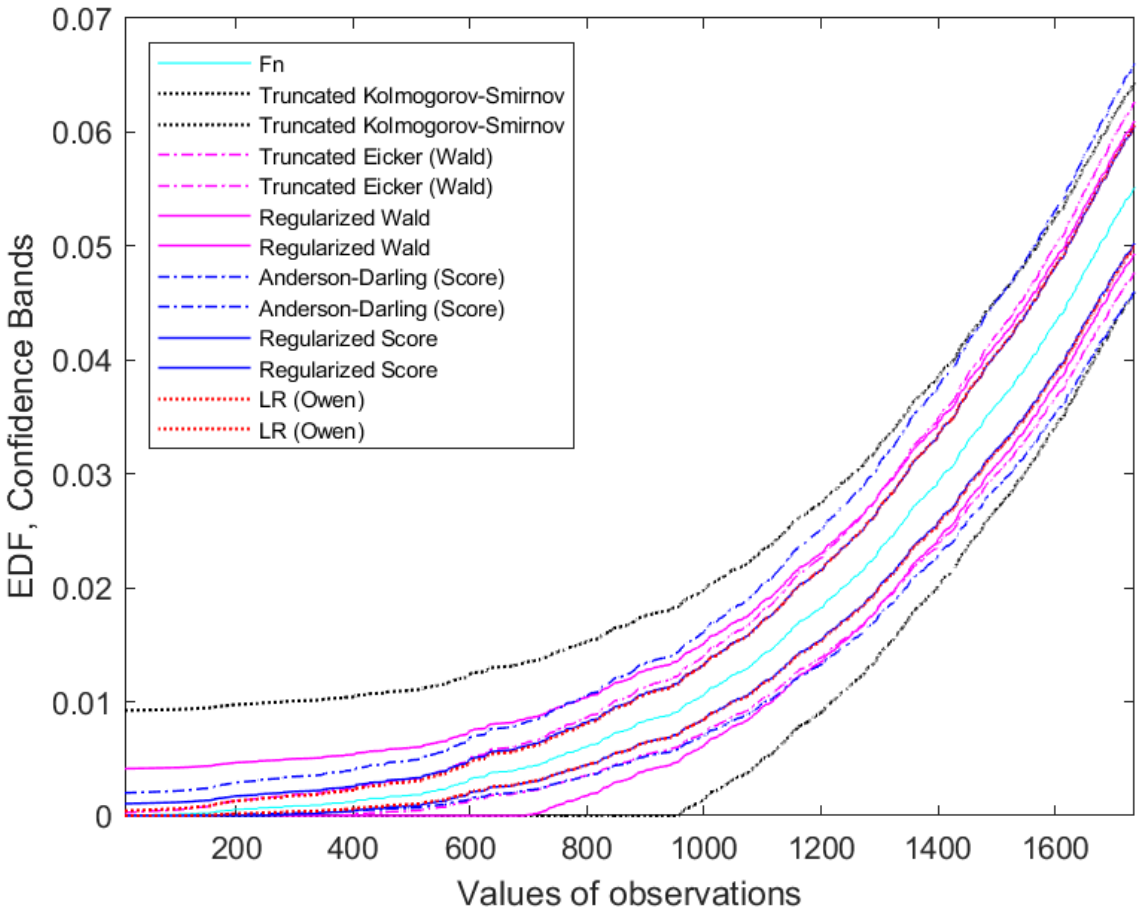
Comparing the surface between the lower and upper bands, the sum of widths of confidence bands over all observations, show that the regularized-Wald method is the best performer under this metric, followed by the regularized-Score method (Table 7). The LR confidence band delivers the third best performance and the truncated Kolmogorov-Smirnov band the fourth best. The Anderson-Darling confidence band has the largest surface.

Note that it is possible to improve the performance of the regularized-EDF based confidence bands by using an out-of-sample procedure to determine the values of $\delta_{n_T}(S)$ and $\delta_{n_T}(W)$. We used for this illustration $\delta_{n_T}(S) = 0.001$ and $\delta_{n_T}(W) = 0.05$ by convenience, as we proved in the previous section that these values improve greatly the performance of the bands. An alternative would be to use an out-of-sample procedure to determine on a subsample the regularization parameters that minimize the widths of the corresponding confidence bands, as discussed in the previous section.

11. Conclusion

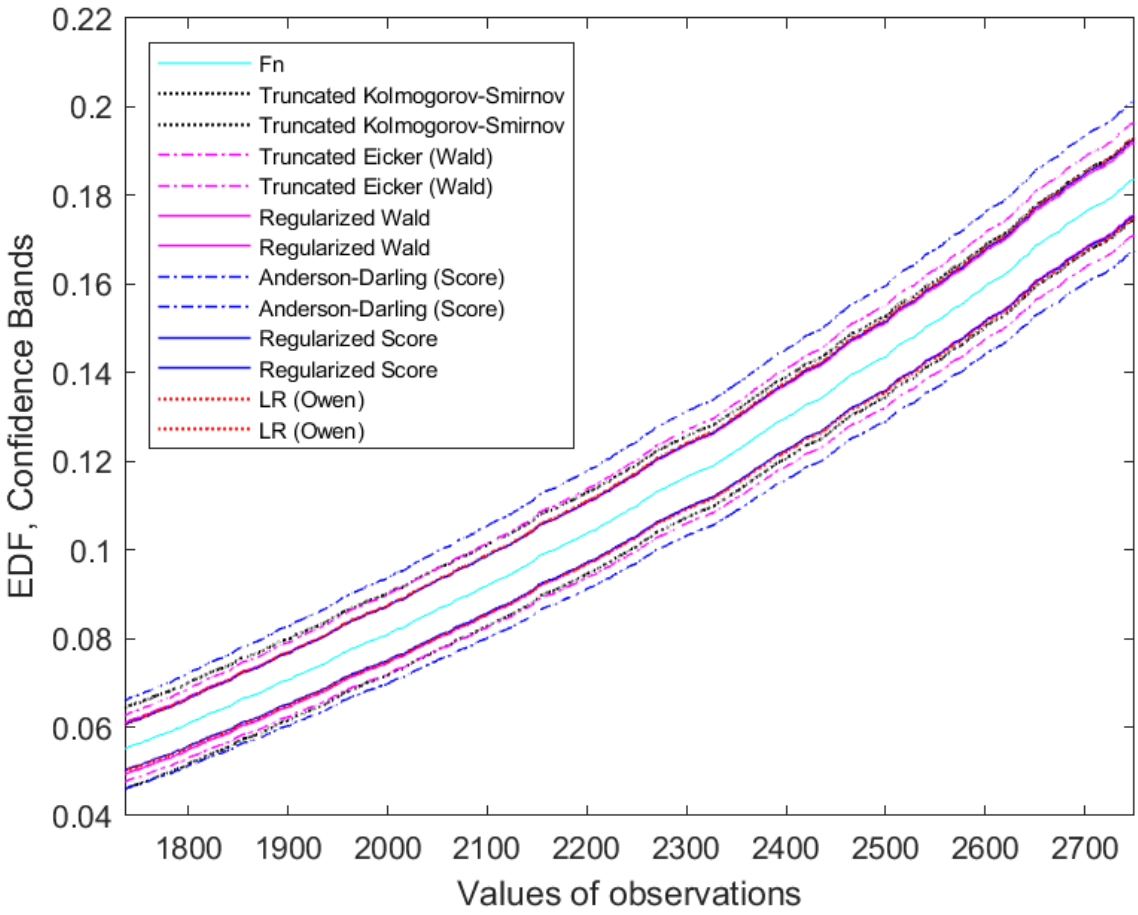
The Kolmogorov-Smirnov test is one of the most popular nonparametric goodness-of-fit tests. It allows to test the null hypothesis that a sample follows a given distribution against the alternative that the sample does not follow this distribution, without any hypothesis on the law of the studied sample. However, the Kolmogorov-Smirnov test does not allow to discriminate a lot between distributions

Figure 3. EDF-based Confidence Bands for the Kenyan Households' Consumption (small observations)



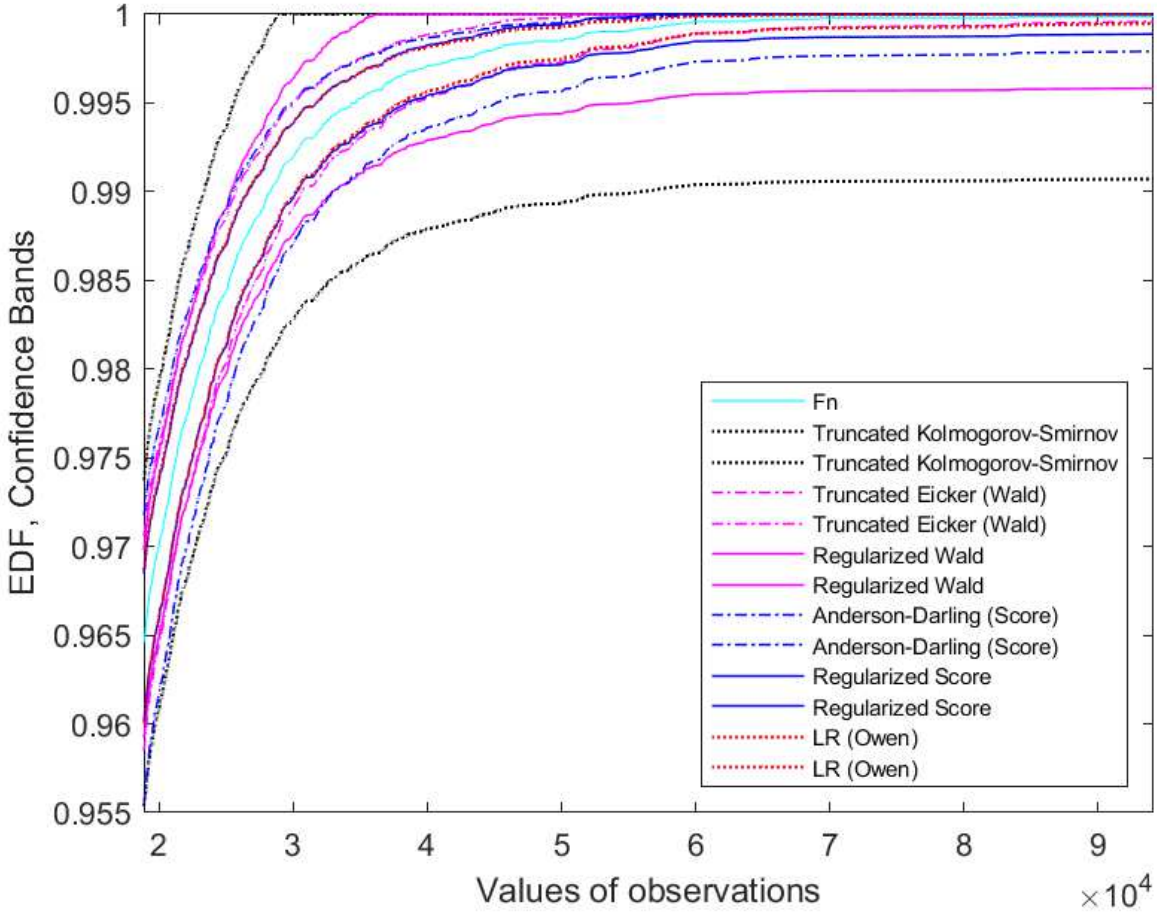
Sample: $n = 1, \ldots, 1,200$; level: $1 - \alpha = 0.95$; $\delta_{n_T}(S) = 0.001$ and $\delta_{n_T}(W) = 0.05$; and number of replications to simulate critical points: $N_2 = 1000000$.

Figure 4. EDF-based Confidence Bands for the Kenyan Households' Adult Consumption (lower center of distribution)



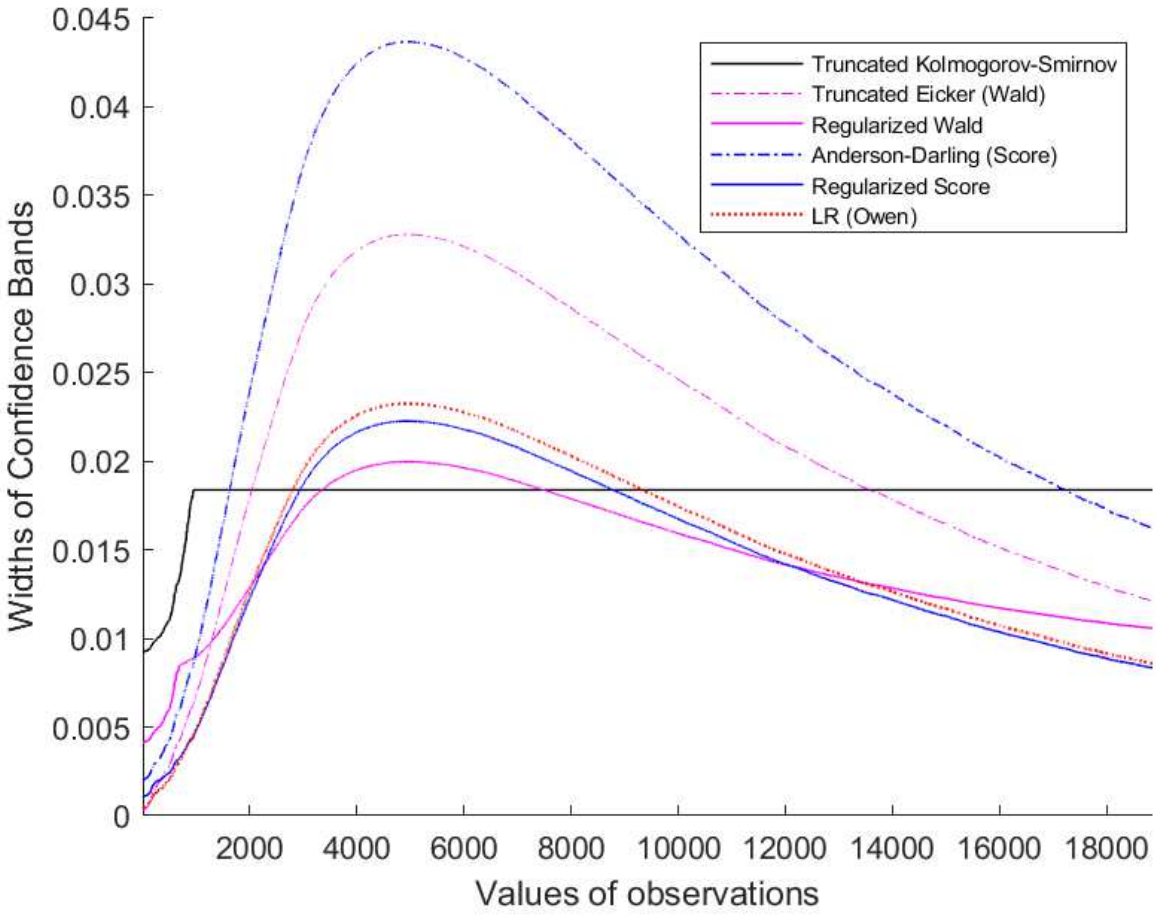
Sample: $n = 1,201,...,4,000$; level: $1-\alpha = 0.95$; $\delta_{n_T}(S) = 0.001$ and $\delta_{n_T}(W) = 0.05$; and number of replications to simulate critical points: $N_2 = 1000000$.

Figure 5. EDF-based Confidence Bands for the Kenyan Households' Consumption (right tail)



Sample: $n = 21,001,...,n_T - 2$; level: $1-\alpha = 0.95$; $\delta_{n_T}(S) = 0.001$ and $\delta_{n_T}(W) = 0.05$; and number of replications to simulate critical points: $N_2 = 1000000$.

Figure 6. Widths of EDF-based Confidence Bands for the Kenyan Households' Consumption (small observations)



Sample: $n = 1, ..., 21000$; level: $1 - \alpha = 0.95$; $\delta_{n_T}(S) = 0.001$ and $\delta_{n_T}(W) = 0.05$; and number of replications to simulate critical points: $N_2 = 1000000$.

Table 7: Surface of EDF-based confidence bands for the Kenyan Households' Consumption

Tests	Surface
Kolmogorov-Smirnov	400.2700048629125
Truncated Kolmogorov-Smirnov	398.4303807424418
Eicker (Wald)	560.9198217199387
Truncated Eicker (Wald)	560.9115139116321
Regularized-Wald	362.4414005675357
Anderson-Darling (Score)	747.5280315956876
Regularized-Score	381.5927706299357
Owen (LR)	397.6378750957000

Sample: $n = n_T = 21773$; level: $1-\alpha = 0.95$; $\delta_{n_T}(S) = 0.001$ and $\delta_{n_T}(W) = 0.05$; number of replications to simulate critical points: $N_2 = 1000000$.

that differ mostly through their tails. The CB it allows to build is uniform: its width is constant for all observations and do not converge to 0 and 1 as do the distribution functions it brackets. To correct this drawback, we study weighted KS statistics based on the three common principles in econometrics: the Wald, likelihood-ratio, and Lagrange multiplier principles.

Weighted Kolmogorov-Smirnov statistics have been proposed by Anderson and Darling (1952), and Eicker (1979). However, these statistics suffer from important drawbacks too. For observations in the tails of distributions, the weights in the denominators of those statistics become very close to zero, leading to erratic behavior of the statistic.

We propose regularized weighted statistics to correct these limits. These statistics are obtained by adding a regularization term in the denominator of the Anderson-Darling and the Eicker statistics. They retain the advantages of the weighted KS statistics but do not suffer from instability, improving the performance of the inference.

We show that in the continuous case, the distributions of the empirical distribution-based statistics of the form of the Kolmogorov-Smirnov and the weighted Kolmogorov-Smirnov statistics are pivotal and that their critical points do not depend on the distribution function being tested under the null hypothesis. Inverting these statistics, we propose exact CBs for distribution functions, which inherit the properties of the statistics they are based on. We show that for all continuous distribution functions, these CBs depend on the distribution only through the sample. A unique set of critical points is needed to build CBs for all continuous distributions, which make them easy to compute. For noncontinuous distribution functions, we derive monotonicity properties that exploit the nesting of the image sets of different distributions to narrow the CBs without altering their reliability. We study a statistic based on the likelihood-ratio principle: the Berk-Jones statistic and the Owen CB, which is derived from it. We show that these statistics and CBs follow the same properties as the Kolmogorov-Smirnov based ones.

We study the performance of these inference methods using Monte Carlo simulations. The results show that the regularized statistics deliver the best performance among the studied inference methods. The regularization increases the power of the tests and the corresponding CBs are thinner

than the other ones. Compared to the weighted statistics, the Owen's band yields generally better results. Nevertheless, it suffers from an important computational shortcoming. In fact, its computation requires to perform as many optimization as the number of observations of the sample we use. This leads to an important loss of time whereas the computation of the Kolmogorov-Smirnov based bands is almost time free.

References

Anderson, T. W. and Darling, D. A. (1952), 'Asymptotic theory of certain goodness-of-fit criteria based on stochastic process', *Annals of Mathematical Statistics* **23**, 193–212.

Andrews, D. W. K. (1997), 'A conditional Kolmogorov test', *Econometrica* **65**(5), 1097–1128.

Arnold, T. B. and Emerson, J. W. (2011), 'Nonparametric goodness-of-fit tests for discrete null distributions', *The R Journal* **3/2**.

Berk, R. H. and Jones, D. H. (1979), 'Goodness-of-fit test statistics that dominate the Kolmogorov statistics', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **47**, 47–59.

Burr, I. W. (1942), 'Cumulative frequency functions', *Annals of Mathematical Statistics* **13**(2), 215–232.

Cayon, L., Jin, J. and Treaster, A. (2005), 'Higher criticism statistic: Detecting and identifying non-gaussianity in the wmap first year data', *Monthly Notices of the Royal Astronomical Society* **362**, 826–832.

Chernoff, H. (1952), 'A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations', *The Annals of Mathematical Statistics* **23**(4), 493–507.

Choulakian, V., Lockhart, R. A. and Stephens, M. A. (1994), Cramér-von Mises statistics for discrete distributions.

Cochran, W. G. and Snedecor, G. W. (1989), *Statistical Methods*, eighth edn, Iowa State University Press, Ames, Iowa.

Conover, W. J. (1972), 'A Kolmogorov goodness-of-fit test for discontinuous distributions', *Journal of the American Statistical Association* **67**, 591–596.

Cramér, H. (1928), 'On the composition of elementary errors', *Skandinavisk Actuarietidskrift* **11**(1), 141–180.

D'Agostino, R. B. and Stephens, M. A., eds (1986), *Goodness-of-Fit Techniques*, Marcel Dekker, New York.

Darling, D. A. (1957), 'The Kolmogorov-Smirnov, Cramér-von Mises tests', *Annals of Mathematical Statistics* **28**, 823–838.

Donoho, D. and Jin, J. (2004), 'Higher criticism for detecting sparse heteregeneous mixtures', *Annals of Statistics* **32**, 962–994.

Donoho, D. and Jin, J. (2008), 'Higher criticism thresholding: Optimal feature selection when useful features are rare and weak', *Proc. Natl Acad. Sci. USA* **105**(39), 14790–14795.

Dufour, J.-M. (2006), 'Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics in econometrics', *Journal of Econometrics* **133**(2), 443–477.

Dufour, J.-M. and Khalaf, L. (2001), Monte Carlo test methods in econometrics, *in* B. Baltagi, ed., 'Companion to Theoretical Econometrics', Blackwell Companions to Contemporary Economics, Basil Blackwell, Oxford, U.K., chapter 23, pp. 494–519.

Dufour, J.-M. and Kiviet, J. F. (1998), 'Exact inference methods for first-order autoregressive distributed lag models', *Econometrica* **66**, 79–104.

Dümbgen, L. and Wellner, J. A. (2023), 'A new approach to tests and confidence bands for distribution functions', *Annals of Statistics* **51**(1), 260–289.

Dwass, M. (1957), 'Modified randomization tests for nonparametric hypotheses', *Annals of Mathematical Statistics* **28**, 181–187.

Eicker, F. (1979), 'The asymptotic distribution of the suprema of the standardized empirical process', *The Annals of Statistics* **7**(1), 116–138.

Fieller, E. C. (1940), 'The biological standardization of insulin', *Journal of the Royal Statistical Society (Supplement)* **7**, 1–64.

Fieller, E. C. (1954), 'Some problems in interval estimation', *Journal of the Royal Statistical Society Series B* **16**, 175–185.

Finkelstein, J. M. and Schafer, R. E. (1971), 'Improved goodness-of-fit tests', *Biometrika* **58**(3), 641–645.

Frey, J. (2008), 'Optimal distribution-free confidence bands for a distribution function', *Journal of Statistical Planning and Inference* **138**, 3086–3098.

Frey, J. (2009), 'Unbiased goodness-of-fit tests', *Journal of Statistical Planning and Inference* **139**, 3690–3697.

Gibbons, J. D. and Chakraborti, S. (1992), *Nonparametric Statistical Inference, Third Edition, Revised and Expanded*, Marcel Dekker, New York.

Gleser, L. J. (1985), 'Exact power of goodness-of-fit tests of Komogorov type for discontinuous distributions', *Journal of the American Statistical Association* **80**(392), 954–958.

Guilbaud, O. (1986), 'Stochastic inequalities for Kolmogorov and similar statistics, with confidence region applications', *Scandinavian Journal of Statistics* **13**(4), 301–305.

Hall, P. and Jin, J. (2008), 'Properties of higher criticism under strong dependence', *The Annals of Statistics* **36**(1), 381–402.

Hall, P., Racine, J. and Li, Q. (2004), 'Cross-validation and the estimation of conditional probability densities', *Journal of the American Statistical Association* **99**(468), 1015–1026.

Henze, N. (1996), 'Empirical-distribution-function goodness-of-fit tests for discrete models', *The Canadian Journal of Statistics* **24**(1), 81–93.

Jaeschke, D. (1979), 'The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals', *The Annals of Statistics* **7**(1), 108–115.

Jager, L. and Wellner, J. A. (2004), A new goodness of fit test: the reversed Berk-Jones statistic, Technical Report 443, Department of Statistics, University of Washington, Seattle, WA.

Jager, L. and Wellner, J. A. (2007), 'Goodness-of-fit tests via phi-divergences', *Annals of Statistics* **35**(5), 2018–2053.

Klar, B. (1999), 'Goodness-of-fit tests for discrete models based on the integrated distribution function', *Metrika* **49**, 53–69.

Kocherlakota, S. and Kocherlakota, K. (1986), 'Goodness of fit tests for discrete distributions', *Communications in Statistics - Theory and Methods* **15**(3), 815–829.

Kolmogorov, A. N. (1933), 'Sulla determinazione empirica di une legge di distribuzione', *Giornale dell'Istituto Italiano degli Attuari* **4**, 83–91.

Kolmogorov, A. N. (1941), 'Confidence limits for an unknown distribution function', *Annals of Mathematical Statistics* **12**, 461–463.

Lockhart, R. A., Spinelli, J. J. and Stephens, M. A. (2007), Cramér-von Mises statistics for discrete distributions with unknown parameters.

Mason, D. M. and Schuenemeyer, J. H. (1983), 'A modified Kolmogorov-Smirnov test sensitive to tail alternatives', *The Annals of Statistics* **11**(3), 933–946.

Massey, F. J. J. (1951), 'The Kolmogorov-Smirnov test for goodness of fit', *Journal of the American Statistical Association* **46**(253), 68–78.

Moscovich, A., Nadler, B. and Spiegelman, C. (2016), 'On the exact berk-jones statistics and their p-value calculation', *Electronic Journal of Statistics* **10**(2), 2329–2354.
*https://doi.org/10.1214/16-EJS1172

Mundform, D., Schaffer, J., Kim, M., Shaw, D. and Thongteeraparp, A. (2011), 'Number of replications required in Monte Carlo simulation studies: A synthesis of four studies', *Journal of Modern Applied Statistical Methods* **10**, 19–28.

Noether, G. E. (1963), 'Note on Kolmogorov statistic in the discrete case', *Metrika* **7**, 115–116.

O'Neill, T. J. and Stern, S. E. (2012), 'Finite population corrections for the Kolmogorov-Smirnov tests', *Journal of Nonparametric Statistics* **24**(2), 497–504.

Owen, A. B. (1995), 'Nonparametric likelihood confidence bands for a distribution function', *American Statistical Association* **90**, 516–521.

Parzen, E. (1962), 'On estimation of a probability density function and mode', *Ann. Math. Statist.* **33**(3), 1065–1076.
*https://doi.org/10.1214/aoms/1177704472

Pettitt, A. N. and Stephens, M. A. (1977), 'The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data', *Technometrics* **19**(2), 205–210.

Pratt, J. W. (1963), 'Shorter confidence intervals for the mean of a normal distribution with known variance', *Annals of Mathematical Statistics* **34**(2), 574–586.
*https://doi.org/10.1214/aoms/1177704170

Reiss, H. D. (1989), *Approximate Distributions of Order Statistics with Applications to Nonparametric Statistics*, Springer Series in Statistics, Springer-Verlag, New York.

Rosenblatt, M. (1956), 'Remarks on some nonparametric estimates of a density function', *Ann. Math. Statist.* **27**(3), 832–837.
*https://doi.org/10.1214/aoms/1177728190

Shao, Y. and Hahn, M. G. (1996), 'On a distribution-free test of fit for continuous distribution functions', *Scandinavian Journal of Statistics* **23**(1), 63–73.

Shorack, G. R. and Wellner, J. A. (1986), *Empirical Processes with Applications to Statistics*, John Wiley & Sons, New York.

Singh, S. K. and Maddala, G. S. (1976), 'A function for size distribution of incomes', *Econometrica* **44**(5), 963–970.

Smirnov, N. V. (1939), 'On the estimation of the discrepancy between empirical curves of distribution for two independent samples', *Moscow University Mathematics Bulletin* **2**(2), 3–14.

Smirnov, N. V. (1944), 'Approximate laws of distribution of random variables from empirical data', *Uspehi Matematiceskih Nauk* **10**, 179–206.

Smirnov, N. V. (1948), 'Table for estimating the goodness of fit of empirical distributions', *Ann. Math. Statist.* **19**(2), 279–281.
*https://doi.org/10.1214/aoms/1177730256

Steele, M. and Chaseling, J. (2006), Powers of discrete goodness-of-fit test statistics for a uniform null against a selection of alternative distributions, Vol. 35, pp. 1067–1075.

Stepanova, N. and Pavlenko, T. (2018), 'Goodness-of-fit tests based on sup-functionals of weighted empirical processes', *Theory of Probability and Its Applications* **63**(2), 292–317.

von Mises, R. (1931), *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und Theoretischen Physik*, F. Deuticke, Leipzig and Wien.

Watson, G. S. (1961), 'Goodness-of-fit tests on a circle', *Biometrika* **48**(1 and 2), 109–114.

Wood, C. L. and Altavela, M. M. (1978), 'Large-sample results for Kolmogorov-Smirnov statistics for discrete distributions', *Biometrika* **65**(1), 235–239.

Zhang, J. (2002), 'Powerful goodness-of-fit tests based on the likelihood ratio', *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **64**(2), 281–294.

Regularized goodness-of-fit statistics and exact nonparametric confidence bands for distributions with application to household consumption

Technical appendix

by
Mame Astou Diouf and Jean-Marie Dufour
International Monetary Fund and McGill University
August 30, 2026

A. Proofs

PROOF OF LEMMA 4.1 By the definition of quantile functions, we have

$$x \geq F_1^{-1}(q) \Leftrightarrow F_1(x) \geq q \text{ , for all } x \in \mathbb{R} \text{ and } q \in (0,1) \tag{A.1}$$

see Shorack and Wellner (1986, Chapter 1, p. 5). Since $X_i = F_1^{-1}(U_i)$ and $0 < U_i < 1$, for $i = 1, \ldots, n$, this entails that

$$X_i \leq x \Leftrightarrow F_1^{-1}(U_i) \leq x \Leftrightarrow U_i \leq F_1(x) \text{ , for all } x \in \mathbb{R}, \tag{A.2}$$

and

$$\mathbf{1}[X_i \leq x] = \mathbf{1}[U_i \leq F_1(x)] \text{ , for all } x \in \mathbb{R}, \tag{A.3}$$

for $i = 1, \ldots, n$. We can thus write

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[X_i \leq x] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[U_i \leq F_1(x)] = H_n\big(F_1(x)\big) \text{ , for all } x \in \mathbb{R}, \tag{A.4}$$

where $H_n\big(\,\cdot\,\big)$ is defined by (4.3), This establishes (4.1), and (4.2) then follows on substituting (4.1) in $D_n(\hat{F}_n, F_0; C_0)$. If $F_1 = F_0$, $F_0(x)$ and $F_1(x)$ in $D_n(\hat{F}_n, F_0; C_0)$ only take values in the set $F_0(C_0)$, hence

$$D_n(\hat{F}_n, F_0; C_0) = \sup_{x \in C_0} \bar{D}_n\left[H_n\big(F_1(x)\big), F_0(x)\right] = \sup_{u \in F_0(C_0)} \bar{D}_n\left[H_n(u), u\right]. \tag{A.5}$$

If F_0 is strictly increasing (with possibly $F_1 \neq F_0$), F_0 is invertible, so that $F_0(C_0) = (0,1)$ and $F_0(x) = u \Leftrightarrow F_0^{-1}(u) = x$, for all $x \in \mathbb{R}$ and $u \in (0,1)$, hence

$$D_n(\hat{F}_n, F_0; C_0) = \sup_{x \in C_0} \bar{D}_n\left[H_n\big(F_1(x)\big), F_0(x)\right] = \sup_{u \in F_0(C_0)} \bar{D}_n\left[H_n\big(F_1(F_0^{-1}(u))\big), u\right]. \tag{A.6}$$

Similarly, if F_1 is strictly increasing, F_1 is invertible, so that $F_1(C_0) = (0,1)$ and $F_1(x) = u \Leftrightarrow F_1^{-1}(u) = x$, for all $x \in \mathbb{R}$ and $u \in (0,1)$, hence

$$D_n(\hat{F}_n, F_0; C_0) = \sup_{x \in C_0} \bar{D}_n \left[H_n\big(F_1(x)\big), F_0(x) \right] = \sup_{u \in F_1(C_0)} \bar{D}_n \left[H_n(u), F_0[F_1^{-1}(u)] \right]. \tag{A.7}$$

This establishes (4.4) and completes the proof. □

PROOF OF LEMMA 4.2 From the definition of $U_i = F_1(X_i)$, we have

$$\mathbb{P}[X_i = F_1^{-1}(U_i)] = \mathbb{P}[X_i = F_1^{-1}(F_1(X_i)] = 1\,,\ i = 1, \ldots, n\,; \tag{A.8}$$

see Shorack and Wellner (1986, Chapter 1, Proposition 3). The condition of Lemma 4.1 [$X_i = F_1^{-1}(U_i)$ with $0 < U_i < 1$, for $i = 1, \ldots, n$] thus holds almost surely, and the identities (4.1), (4.2) and (4.4) also hold almost surely. This completes the proof of the lemma. □

PROOF OF PROPOSITION 4.3 Let $U_1, \ldots, U_n$ be i.i.d. random *random variables with uniform distributions on* $(0,1)$. Then, the variables $\tilde{X}_i = F^{-1}(U_i)$, $i = 1, \ldots, n$, are i.i.d. with distribution function F_1 [this follows from the equivalence (A.1)], and the vectors $\tilde{X}^{(n)} = (\tilde{X}_1, \ldots, \tilde{X}_n)'$ and $X^{(n)} = (X_1, \ldots, X_n)'$ have the same joint distribution. Setting

$$\tilde{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[\tilde{X}_i \le x]\,, \tag{A.9}$$

the random variables $D_n(\tilde{F}_n, F_0)$ and $D_n(\hat{F}_n, F_0; C_0)$ follow the same distribution. We can then apply Lemma 4.1 with $F_1 = F$, and we see that

$$D_n(\tilde{F}_n, F_0) = \sup_{x \in C_0} \bar{D}_n \left[H_n\big(F(x)\big), F_0(x) \right] := \tilde{D}_n(\hat{F}_n, F_0; C_0) \quad \text{a.s.} \tag{A.10}$$

This establishes (4.5). Finally, (4.6) follows in a similar way from (4.4). □

PROOF OF PROPOSITION 4.4 The characterization (4.8) follows directly from Proposition 4.3 with $F = F_0$ along with Assumption 4.1. When furthermore F_0 is continuous,

$$F_0(X^{(n)}) = \big((F_0(X_1), \ldots, F_0(X_n)\big) := (U_1, \ldots, U_n) \tag{A.11}$$

where $U_1, \ldots, U_n$ are i.i.d. according a $U(0,1)$ distribution. Thus the distribution of $D_n(\hat{F}_n, F_0; C_0)$ is the same for all $F_0 \in \mathscr{F}_0$. □

PROOF OF PROPOSITION 5.1 On applying proposition 4.4, it follows that for continuous distribution functions, the statistic $D_n[\hat{F}_n(x), F(x)]$

$$D_n[\hat{F}_n(x), F(x)] = \sup_{-\infty < x < +\infty} \bar{D}_n[\hat{F}_n(x), F(x)]. \tag{A.12}$$

is pivotal and can be reexpressed as follows:

$$D_n[\hat{F}_n(x), F(x)] = \sup_{0\le u\le 1} \bar{D}_n\left[H(U_1,\ldots,U_n,u),u\right]. \tag{A.13}$$

If $G(y)$ is a noncontinuous distribution function,

$$\begin{aligned} D_n\left(G_n(x),G(x)\right) &= \sup_{-\infty<x<+\infty} \bar{D}_n\left(H\left[G(X_1),\ldots,G(X_n),G(x)\right],G(x)\right) \\ &= \sup_{\upsilon\in G(\bar{\mathbb{R}})} \bar{D}_n\left(H\left[G(X_1),\ldots,G(X_n),\upsilon\right],\upsilon\right) \le D_n[\hat{F}_n(x),F(x)] \end{aligned} \tag{A.14}$$

because $G(\bar{\mathbb{R}}) \subseteq F(\bar{\mathbb{R}}) = [0,1]$. The critical points associated with $D_n(\hat{F}_n, F_0; C_0)$ are larger than or equal to the corresponding ones for $D_n(\hat{F}_n, G, C_0)$. Critical values associated with continuous distribution functions are conservative for noncontinuous distributions. □

PROOF OF PROPOSITION 5.2 On applying proposition 5.1, it follows that $D_n\left[\hat{F}_n(x),F(x)\right]$ is greater than or equal to $D_n\left[G_n(x),G(x)\right]$. Hence, for a given level α, the critical point associated to $D_n\left[\hat{F}_n(x),F(x)\right]$ is larger than those associated to $D_n\left[G_n(x),G(x)\right]$ or equivalently, the critical point with level α associated to $D_n\left[\hat{F}_n(x),F(x)\right]$ represents a critical point for $D_n\left[G_n(x),G(x)\right]$ with $\beta \le \alpha$. Therefore, the CB for $G(y)$ using the appropriate critical point for $D_n\left[\hat{F}_n(x),F(x)\right]$ will be of level $1-\beta \ge 1-\alpha$. □

PROOF OF PROPOSITION 5.3 (4.7) In the proof of Proposition 4.4, we showed that

$$\begin{aligned} D_n\left[\hat{F}_n(x),F(x)\right] &= \sup_{-\infty<x<+\infty} \bar{D}_n\left[H(U_1,\ldots,U_n,F(x)),F(x)\right] \\ &= \sup_{u\in F(\bar{\mathbb{R}})} \bar{D}_n\left[H(U_1,\ldots,U_n,u),u\right]. \end{aligned} \tag{A.15}$$

Given that $G(\bar{\mathbb{R}}) \subseteq F(\bar{\mathbb{R}})$, when taking the supremum over $G(\bar{\mathbb{R}})$, the result will be smaller than when taking the supremum over $F(\bar{\mathbb{R}})$. Hence,

$$D_n\left[\hat{F}_n(x),F(x)\right] \ge \sup_{v\in G(\bar{\mathbb{R}})} \bar{D}_n\left[H(U_1,\ldots,U_n,u),u\right] = D_n\left[G_n(x),G(x)\right]. \tag{A.16}$$

□

PROOF OF PROPOSITION 5.4 On applying proposition 5.3, it follows that for a given level α, the critical point associated to $D_n\left[\hat{F}_n(x),F(x)\right]$ is larger than those associated to $D_n\left[G_n(x),G(x)\right]$. In other words, the critical point with level α associated to $D_n\left[\hat{F}_n(x),F(x)\right]$ represents a critical point for $D_n\left[G_n(x),G(x)\right]$ with $\beta \le \alpha$. Hence, using the same reasoning as in the proof of Proposition 5.2., it follows that the CB for $G(y)$ using the appropriate critical point for $F(x)$ will be of level $1-\beta \ge 1-\alpha$. □

PROOF OF COROLLARY 5.5 This proposition is a special case of Proposition 2 5. In this case, $p_1 = F_1(a) \leq F_2(a) = p_2$. This implies that $F_2(\bar{\mathbb{R}}) = [p_2, 1] \subseteq [p_1, 1] = F_1(\bar{\mathbb{R}})$ and applying Proposition 5.4. yields the result.

PROOF OF PROPOSITION 6.1 The expression of the statistic $D_n(\hat{F}_n, F_0)$ proposed in (6.2) follows by decomposing the optimization domain C_0 into intervals $\left[X_{(j)}, X_{(j+1)}\right)$, $j = 0, ..., n+1$. It follows that:

$$\begin{aligned} D_n(\hat{F}_n, F_0; C_0) &= \sup_{x \in C_0} \bar{D}_n \left[\hat{F}_n(x), F_0(x)\right] \\ &= \max_{0 \leq i \leq n} \left\{ \sup_{X_{(i)} \leq x < X_{(i+1)}} \bar{D}_n \left[\hat{F}_n(x), F_0(x)\right] : x \in C_0 \cap \left[X_{(i)}, X_{(i+1)}\right) \right\}. \quad \text{(A.17)} \end{aligned}$$

□

PROOF OF PROPOSITION 6.2 Under the assumptions of the proposition

$$D_n(\hat{F}_n, F_0; C_0) = \sup \left\{ D_n^+, D_n^- \right\} \quad \text{(A.18)}$$

where

$$D_n^+(\hat{F}_n, F_0; C_0) = \sup_{x \in C_0} \bar{D}_n^+ \left[\hat{F}_n(x), F_0(x)\right], \quad D_n^-(\hat{F}_n, F_0; C_0) = \sup_{x \in C_0} \bar{D}_n^- \left[\hat{F}_n(x), F_0(x)\right]. \quad \text{(A.19)}$$

Assuming that $C_0 = \left[X_{(j)}, X_{(k+1)}\right)$ for $0 \leq j \leq k \leq n$ and applying the order-statistic representation formula in equation (6.3) yields

$$\begin{aligned} D_n^+(\hat{F}_n, F_0; C_0) &= \max_{j \leq i \leq k} \sup \left\{ \bar{D}_n^+ \left[\hat{F}_n(x), F_0(x)\right] : X_{(i)} \leq x < X_{(i+1)} \right\} \\ &= \max_{j \leq i \leq k} \sup \left\{ \bar{D}_n^+ \left[\frac{i}{n}, F_0(X_{(i)})\right] : X_{(i)} \leq x < X_{(i+1)} \right\} \\ &= \max_{0 \leq i \leq n} \sup \left\{ \bar{D}_n^+ \left[\frac{i}{n}, F_0(X_{(i)})\right] : x \in C_0 \cap \left[X_{(i)}, X_{(i+1)}\right) \right\} \quad \text{(A.20)} \end{aligned}$$

and

$$\begin{aligned} D_n^-(\hat{F}_n, F_0; C_0) &= \max_{j \leq i \leq k} \sup \left\{ \bar{D}_n^- \left[\hat{F}_n(x), F_0(x)\right] : X_{(i)} \leq x < X_{(i+1)} \right\} \\ &= \max_{j \leq i \leq k} \sup \left\{ \bar{D}_n^- \left[\frac{i}{n}, F_0(X_{(i+1)}-)\right] : X_{(i)} \leq x < X_{(i+1)} \right\} \\ &= \max_{0 \leq i \leq n} \sup \left\{ \bar{D}_n^- \left[\frac{i}{n}, F_0(X_{(i+1)}-)\right] : x \in C_0 \cap \left[X_{(i)}, X_{(i+1)}\right) \right\} \quad \text{(A.21)} \end{aligned}$$

where $F_0(x-) := \lim_{y \uparrow x} F(y) = F_0(x) - p_{F_0}(x)$ and $p_{F_0}(x) := \mathbb{P}_{F_0}(X = x)$. It follows that

$$D_n(\hat{F}_n, F_0; C_0) = \max_{0 \le i \le n} \sup \left\{ \bar{D}_n^+ \left[\frac{i}{n}, F_0(X_{(i)}) \right], \bar{D}_n^- \left[\frac{i}{n}, F_0(X_{(i+1)}-) \right] : x \in C_0 \cap \left[X_{(i)}, X_{(i+1)} \right) \right\}. \tag{A.22}$$

□

PROOF OF COROLLARY 6.3 For $C_0 = \mathbb{R}$, the regularized Wald-type statistic defined in (3.2) is

$$\begin{aligned} D_n^W(\hat{F}_n, F_0; \delta_n) &= \sup_{x \in \mathbb{R}} \sqrt{n} \frac{\left| \hat{F}_n(x) - F_0(x) \right|}{\{\hat{F}_n(x)[1 - \hat{F}_n(x)] + \delta_n\}^{1/2}} \\ &= \sqrt{n} \max \left\{ \sup_{x \in \mathbb{R}} \frac{\hat{F}_n(x) - F_0(x)}{\{\tilde{\sigma}^2 \left[\hat{F}_n(x) \right] + \delta_n\}^{1/2}}, \sup_{x \in \mathbb{R}} \frac{F_0(x) - \hat{F}_n(x)}{\{\tilde{\sigma}^2 \left[\hat{F}_n(x) \right] + \delta_n\}^{1/2}} \right\} \end{aligned} \tag{A.23}$$

where $\tilde{\sigma}^2[p] := p(1-p)$. It follows that D_n^W can be rewritten as

$$D_n^W(\hat{F}_n, F_0; \delta_n) = \sup \left\{ D_n^+, D_n^- \right\} \tag{A.24}$$

where

$$D_n^+ = \sup_{-\infty < x < +\infty} \overline{D}_n^+, \text{ with } \overline{D}_n^+ = \sqrt{n} \frac{\hat{F}_n(x) - F_0(x)}{\sqrt{\tilde{\sigma}^2 \left[\hat{F}_n(x) \right] + \delta_n}}, \tag{A.25}$$

$$D_n^- = \sup_{-\infty < x < +\infty} \overline{D}_n^-, \text{ with } \overline{D}_n^- = \sqrt{n} \frac{F_0(x) - \hat{F}_n(x)}{\sqrt{\tilde{\sigma}^2 \left[\hat{F}_n(x) \right] + \delta_n}}. \tag{A.26}$$

Applying the order-statistic representation with monotonic discrepancies in equation (6.9), we derive:

$$\begin{aligned} D_n^W(\hat{F}_n, F_0; \delta_n) &= \sqrt{n} \max_{0 \le i \le n} \sup \left\{ \frac{\frac{i}{n} - F_0(X_{(i)})}{\sqrt{\frac{i}{n}[1 - \frac{i}{n}] + \delta_n}}, \frac{F_0(X_{(i+1)}-) - \frac{i}{n}}{\sqrt{\frac{i}{n}(1 - \frac{i}{n}) + \delta_n}} \right\} \\ &= \sqrt{n} \max_{0 \le i \le n} \sup \left\{ \frac{\frac{i}{n} - F_0(X_{(i)})}{\sqrt{\tilde{\sigma}^2 \left[\frac{i}{n} \right] + \delta_n}}, \frac{F_0(X_{(i+1)}-) - \frac{i}{n}}{\sqrt{\tilde{\sigma}^2 \left[\frac{i}{n} \right] + \delta_n}} \right\}. \end{aligned} \tag{A.27}$$

Note that for $i = 0$: $\overline{D}_n^+ \left[\hat{F}_n(X_{(i)}), F_0(X_{(i)}) \right] = \overline{D}_n^+ [0,0] = 0$, and for $i = n+1$: $\overline{D}_n^- \left[\hat{F}_n(X_{(i)}), F_0(X_{(i+1)-}) \right] = \overline{D}_n^- [1,1] = 0$. The regularized Score-type statistic defined in (3.3) is:

$$D_n^S(\hat{F}_n, F_0; \delta_n) = \sup_{x \in \mathbb{R}} \sqrt{n} \frac{\left| \hat{F}_n(x) - F_0(x) \right|}{\{F_0(x)[1 - F_0(x)] + \delta_n\}^{1/2}}$$

$$= \sqrt{n}\ \max\left\{\sup_{x\in\mathbb{R}}\frac{\hat{F}_n(x)-F_0(x)}{\{\tilde{\sigma}^2\left[F_0(x)\right]+\delta_n\}^{1/2}},\sup_{x\in\mathbb{R}}\frac{F_0(x)-\hat{F}_n(x)}{\{\tilde{\sigma}^2\left[F_0(x)\right]+\delta_n\}^{1/2}}\right\}$$
$$= \sup\left\{D_n^+,D_n^-\right\} \tag{A.28}$$

where

$$D_n^+ = \sup_{-\infty<x<+\infty}\bar{D}_n^+(\hat{F}_n,F_0;C_0)\text{ for }\bar{D}_n^+(\hat{F}_n,F_0;C_0)=\sqrt{n}\frac{\hat{F}_n(x)-F_0(x)}{\sqrt{\tilde{\sigma}^2\left[F_0(x)\right]+\delta_n}}, \tag{A.29}$$

$$D_n^- = \sup_{-\infty<x<+\infty}\bar{D}_n^-(\hat{F}_n,F_0;C_0)\text{ for }\bar{D}_n^-[\hat{F}_n,F_0]=\sqrt{n}\frac{F_0(x)-\hat{F}_n(x)}{\sqrt{\tilde{\sigma}^2\left[F_0(x)\right]+\delta_n}}. \tag{A.30}$$

$\bar{D}_n^+(\widehat{p},p)$ is non-increasing in p, as its first derivative in p is always negative for $\delta_n\geq 0$ and is equal to zero only at a value

$$p^*=\frac{\frac{i}{2}\frac{i}{n}+\delta_n}{\frac{i}{n}-\frac{i}{2}} \tag{A.31}$$

which does not belong to $[0,1]$. For the same reason $\bar{D}_n^-(\widehat{p},p)$ is non-decreasing in p. Applying the order-statistic representation with monotonic discrepancies in equation (6.9), it follows that

$$D_n^S(\hat{F}_n,F_0;\delta_n)=\sqrt{n}\max_{0\leq i\leq n}\sup\left\{\frac{\left[\frac{i}{n}-F_0(X_{(i)})\right]}{\sqrt{\tilde{\sigma}^2\left[F_0(X_{(i)})\right]+\delta_n}},\frac{\left[F_0(X_{(i+1)}-)-\frac{i}{n}\right]}{\sqrt{\tilde{\sigma}^2\left[F_0(X_{(i+1)}-)\right]+\delta_n}}\right\}. \tag{A.32}$$

Note that for $i=0$: $\overline{D}_n^+\left[\hat{F}_n(X_{(i)}),F_0(X_{(i)})\right]=\overline{D}_n^+[0,0]=0$, and for $i=n$: $\overline{D}_n^-\left[\hat{F}_n(X_{(i)}),F_0(X_{(i+1)-})\right]=\overline{D}_n^-[1,1]=0$. □

PROOF OF COROLLARY 6.4 For $C_0=\mathbb{R}$, the sup-EDF LR-type statistic defined in (2.16) is:

$$D_n^{LR}(\hat{F}_n,F_0;C_0)=\sup_{x\in\mathbb{R}}K[\hat{F}_n(x)\ ,F_0(x)]\,, \tag{A.33}$$

$$K[\hat{p},p]=\hat{p}\log\left(\frac{\hat{p}}{p}\right)+(1-\hat{p})\log\left(\frac{1-\hat{p}}{1-p}\right). \tag{A.34}$$

For $p\not\geq 0$, $K(\widehat{p},p)$ is non-increasing in p for $p\leq\widehat{p}$ and is non-decreasing in p for $p\not\geq\widehat{p}$. It follows that D_n^{LR} can be rewritten as

$$D_n^{LR}(\hat{F}_n,F_0;C_0)=\max\left\{D_n^{+LR},D_n^{-LR}\right\}$$

where

$$D_n^{+LR}(\hat{F}_n,F_0;C_0)=\sup_{x\in\mathbb{R}}\left\{K\left[\hat{F}_n(x)\ ,F_0(x)\right],\hat{F}_n(x)\geq F_0(x)\right\}, \tag{A.35}$$

$$D_n^{-LR}(\hat{F}_n,F_0;C_0)=\sup_{x\in\mathbb{R}}\left\{K\left[\hat{F}_n(x)\ ,F_0(x)\right],\hat{F}_n(x)\leq F_0(x)\ \right\}. \tag{A.36}$$

Applying the order-statistic representation with monotonic discrepancies in equation (6.9) yields:

$$D_n^{LR}(\hat{F}_n, F_0; C_0) = \max\left\{D_n^{+LR}, D_n^{-LR}\right\}, \tag{A.37}$$

$$D_n^{+LR}(\hat{F}_n, F_0; C_0) = \sup\left\{\sup_{0\le i\le n} K\left[\frac{i}{n}, F_0(X_{(i)})\right], \hat{F}_n(x) \ge F_0(x)\right\}, \tag{A.38}$$

$$D_n^{-LR}(\hat{F}_n, F_0; C_0) = \sup\left\{\sup_{0\le i\le n} K\left[\frac{i}{n}, F_0(X_{(i+1)}-)\right], \hat{F}_n(x) \le F_0(x)\right\}, \tag{A.39}$$

where $F_0(X_{(i+1)}-) = F_0(X_{(i+1)}) - p_F(X_{(i+1)})$. Note that for $i=0$: $K\left[\hat{F}_n(X_{(i)}), F_0(X_{(i)})\right] = K[0,0] = 0$ and for $i = n+1$: $K\left[\hat{F}_n(X_{(i)}), F_0(X_{(i+1)}-)\right] = K[1,1] = 0$. □ □

B. Construction of confidence bands

B.1. Score-type (Anderson-Darling) confidence bands

This section derives an explicit expression for the confidence bands obtained by inverting the EDF-based GF tests regularized using a restricted optimization domain. For the Score-weighted EDF-based GF test:

$$D_n^S[\hat{F}_n, F_0] = \sup_{x\in C_0} \bar{D}_n^S\left[\hat{F}_n(x), F_0(x)\right]. \tag{B.1}$$

where $F_0(x) \notin \{0,1\}, \forall x \in C_0$, and

$$\bar{D}_n^S\left[\hat{F}_n(x), F_0(x)\right] = \frac{\sqrt{n}[\hat{F}_n(x) - F_0(x)]}{(F_0(x)[1-F_0(x)])^{1/2}}. \tag{B.2}$$

Let $c_S(\alpha; n)$ *be defined* by

$$\mathbb{P}\left[D_n^S[\hat{F}_n, F_0] \le c_S(\alpha;n)\right] = \mathbb{P}\left[\sup_{x\in C_0} \bar{D}_n^S\left[\hat{F}_n(x), F_0(x)\right] \le c_S(\alpha;n)\right] \ge 1-\alpha. \tag{B.3}$$

Since $\bar{D}_n^S\left[\hat{F}_n(x), F_0(x)\right] \ge 0\ \forall x \in C_0$, we have:

$$\mathbb{P}\left[\frac{\left[\hat{F}_n(x) - F_0(x)\right]^2}{F_0(x)[1-F_0(x)]} \le \frac{1}{n} c_S^2(\alpha;n),\ \forall x \in C_0\right] \ge 1-\alpha, \tag{B.4}$$

$$\mathbb{P}\left[\hat{F}_n^2(x) - 2\hat{F}_n(x)F_0(x) + F_0^2(x) \le \frac{1}{n} c_S^2(\alpha;n)\left(F_0(x) - F_0^2(x)\right),\ \forall x \in C_0\right] \ge 1-\alpha, \tag{B.5}$$

$$\mathbb{P}\left[\left(1 + \frac{1}{n}c_S^2(\alpha;n)\right)F_0^2(x) - \left(2\hat{F}_n(x) + \frac{1}{n}c_S^2(\alpha;n)\right)F_0(x) + \hat{F}_n^2(x) \le 0,\ \forall x \in C_0\right] \ge 1-\alpha. \tag{B.6}$$

This is equivalent to

$$\mathbb{P}\left[\hat{F}_n^L(x) \le F(x) \le \hat{F}_n^U(x) \le 0,\ \forall x \in C_0\right] \ge 1-\alpha\,, \tag{B.7}$$

$$\hat{F}_n^L(x) = \frac{2\hat{F}_n(x) + \frac{1}{n}c_S^2(\alpha;n) - \sqrt{\Delta_n(x)}}{2\left(1+\frac{1}{n}c_S^2(\alpha;n)\right)}\,, \quad \hat{F}_n^U(x) = \frac{2\hat{F}_n(x) + \frac{1}{n}c_S^2(\alpha;n) + \sqrt{\Delta_n(x)}}{2\left(1+\frac{1}{n}c_S^2(\alpha;n)\right)}\,, \tag{B.8}$$

$$\Delta_n(x) = \left[2\hat{F}_n(x) + \frac{1}{n}c_S^2(\alpha;n)\right]^2 - 4F_n^2(x)\left[1+\frac{1}{n}c_S^2(\alpha;n)\right]\,. \tag{B.9}$$

B.2. Wald-type (Eicker) confidence bands

For the Wald-weighted EDF-based GF test:

$$D_n^W[\hat{F}_n, F_0] = \sup_{x\in C_0} \bar{D}_n^W\left[\hat{F}_n(x), F_0(x)\right] \tag{B.10}$$

where $\hat{F}_n(x) \notin \{0,1\}, \forall x \in C_0$,

$$\bar{D}_n^W\left[\hat{F}_n(x), F_0(x)\right] = \frac{\sqrt{n}[\hat{F}_n(x) - F_0(x)]}{(\hat{F}_n(x)[1-\hat{F}_n(x)])^{1/2}}\,. \tag{B.11}$$

Let $c_W(\alpha;n)$ be the critical point such that

$$\mathbb{P}[D_n^W[\hat{F}_n, F_0] \le c_W(\alpha;n)] = \mathbb{P}[\sup_{x\in C_0} \bar{D}_n^W\left[\hat{F}_n(x), F_0(x)\right] \le c_W(\alpha;n)] \ge 1-\alpha\,. \tag{B.12}$$

Given that $\bar{D}_n^W\left[\hat{F}_n(x), F_0(x)\right] \ge 0,\ \forall x \in C_0$,then

$$\mathbb{P}\left[-c_W(\alpha;n) \le \frac{\sqrt{n}[\hat{F}_n(x) - F_0(x)]}{\left(\hat{F}_n(x)[1-\hat{F}_n(x)]\right)^{1/2}} \le c_W(\alpha;n),\ \forall x \in C_0\right] \ge 1-\alpha\,, \tag{B.13}$$

$$\mathbb{P}\left[\hat{F}_n(x) - \frac{1}{\sqrt{n}}c_W(\alpha;n)\left(\hat{F}_n(x)[1-\hat{F}_n(x)]\right)^{1/2} \le F_0(x) \le \hat{F}_n(x) + \frac{1}{\sqrt{n}}c_W(\alpha;n)\left(\hat{F}_n(x)[1-\hat{F}_n(x)]\right)^{1/2},\ \forall x \in C_0\right] \ge 1-\alpha\,. \tag{B.14}$$

B.3. Regularized Score-type confidence bands

Let $c_S(\alpha;\delta_n)$ such that $\mathbb{P}[D_n^S(\hat{F}_n, F_0;\,\delta_n) \le c_S(\alpha;\delta_n)] \ge 1-\alpha$ *i.e.*,

$$\mathbb{P}\left[\sup_{-\infty<x<+\infty} \left|\frac{\sqrt{n}[\hat{F}_n(x) - F_0(x)]}{\sqrt{F(x)[1-F(x)]+\delta_n}}\right| \le c_S(\alpha;\delta_n)\right] \ge 1-\alpha\,. \tag{B.15}$$

Then, with probability greater than or equal to $1-\alpha$

$$\frac{\left[\hat{F}_n(x)-F_0(x)\right]^2}{F_0(x)[1-F_0(x)]+\delta_n} \leq \frac{c_S^2(\alpha;\delta_n)}{n} \qquad \forall\, x \tag{B.16}$$

i.e., $\forall\, x$,

$$\big(1+\frac{c_S^2(\alpha;\delta_n)}{n}\big)F_0^2(x)-\big(2\hat{F}_n(x)+\frac{c_S^2(\alpha;\delta_n)}{n}\big)F_0(x)+\hat{F}_n^2(x)-\frac{\zeta c_S^2(\alpha;\delta_n)}{n} \leq 0 \tag{B.17}$$

This condition is satisfied if and only if $\forall\, x$

$$F_0^L(x) \leq F_0(x) \leq F_0^U(x) \tag{B.18}$$

where

$$F_0^L(x)=\frac{2\hat{F}_n(x)+\frac{c_S^2(\alpha;\delta_n)}{n}-\sqrt{\Delta}}{2\big(1+\frac{c_S^2(\alpha;\delta_n)}{n}\big)}\ ,\ \ F_0^U(x)=\frac{2\hat{F}_n(x)+\frac{c_S^2(\alpha;\delta_n)}{n}+\sqrt{\Delta}}{2\big(1+\frac{c_S^2(\alpha;\delta_n)}{n}\big)}\,, \tag{B.19}$$

$$\Delta=\big[2\hat{F}_n(x)+\frac{c_S^2(\alpha;\delta_n)}{n}\big]^2-4\big[1+\frac{c_S^2(\alpha;\delta_n)}{n}\big]\big[\hat{F}_n^2(x)-\frac{\delta c_S^2(\alpha;\delta_n)}{n}\big]\,. \tag{B.20}$$

B.4. Regularized Wald-type confidence bands

Suppose $\forall x \in C_0=\mathbb{R}$. Let's define the critical point $c_W(\alpha;\delta_n)$ of the regularized Wald-type statistic defined in (3.2) such that $\mathbb{P}[D_n^W(\hat{F}_n,F_0;\delta)\leq c_W(\alpha;\delta_n)]\geq 1-\alpha,\ \textit{i.e.},$

$$\mathbb{P}\left[\sup_{-\infty<x<+\infty}\left|\frac{\sqrt{n}[\hat{F}_n(x)-F(x)]}{\sqrt{\hat{F}_n(x)[1-\hat{F}_n(x)]+\delta_n}}\right|\leq c_W(\alpha;\delta_n)\right]\geq 1-\alpha\ . \tag{B.21}$$

Then, with probability greater than or equal to $1-\alpha$

$$-c_W(\alpha;\delta_n)\leq\frac{\sqrt{n}[\hat{F}_n(x)-F(x)]}{\sqrt{\hat{F}_n(x)[1-\hat{F}_n(x)]+\delta_n}}\leq c_W(\alpha;\delta_n) \qquad \forall\, x \tag{B.22}$$

or equivalently,

$$\hat{F}_n(x)-\frac{c_W(\alpha;\delta_n)}{\sqrt{n}}\big(\hat{F}_n(x)[1-\hat{F}_n(x)]+\delta_n\big)^{1/2}\leq F(x)\leq\hat{F}_n(x)+\frac{c_W(\alpha;\delta_n)}{\sqrt{n}}\big(\hat{F}_n(x)[1-\hat{F}_n(x)]+\delta_n\big)^{1/2}. \tag{B.23}$$

B.5. LR-type confidence bands

Owen (1995) suggests the following computational procedure:

1. Compute (for $c_{BJ}(\alpha;n) > 0$)

$$H_i = \begin{cases} 1 - e^{-c_{BJ}(\alpha)} \text{ if } i = 0 \\ 1 \text{ if } i = n \\ \max\limits_{\frac{i}{n} \leq p \leq 1} \{p : K[F_n(x), p] \leq c_{BJ}(\alpha;n)\} \text{ if } 2 \leq i \leq n-1 \end{cases} \tag{B.24}$$

2. Deduce $L_i = 1 - H_{n-i}$ for $0 \leq i \leq n$ (by symmetry)

3. Compute the values of the confidence band for each observation $X_{(i)}$ using

$$F_n^L(X_{(i)}) \leq F(X_{(i)}) \leq F_n^U(X_{(i)}), \tag{B.25}$$

$$F_n^L(X_{(i)}) = \min(L_{i-1}, L_i) = L_{i-1} \quad F_n^U(X_{(i)}) = \max(H_{i-1}, H_i) = H_i. \tag{B.26}$$

The following polynomial approximation can be used for $c_{BJ}(\alpha)$(Jager and Wellner (2004)):

$$c_{LR}(0.05;n) = \begin{cases} \frac{1}{n}\left[3.6792 + 0.5720 \log n - 0.0567 \log^2(n) - 0.0027 \log^3(n)\right] & \text{for } 1 < n \leq 100, \\ \frac{1}{n}\left[3.7752 + 0.5062 \log n - 0.0417 \log^2(n) + 0.0016 \log^3(n)\right] & \text{for } 100 < n \leq 1000, \end{cases} \tag{B.27}$$

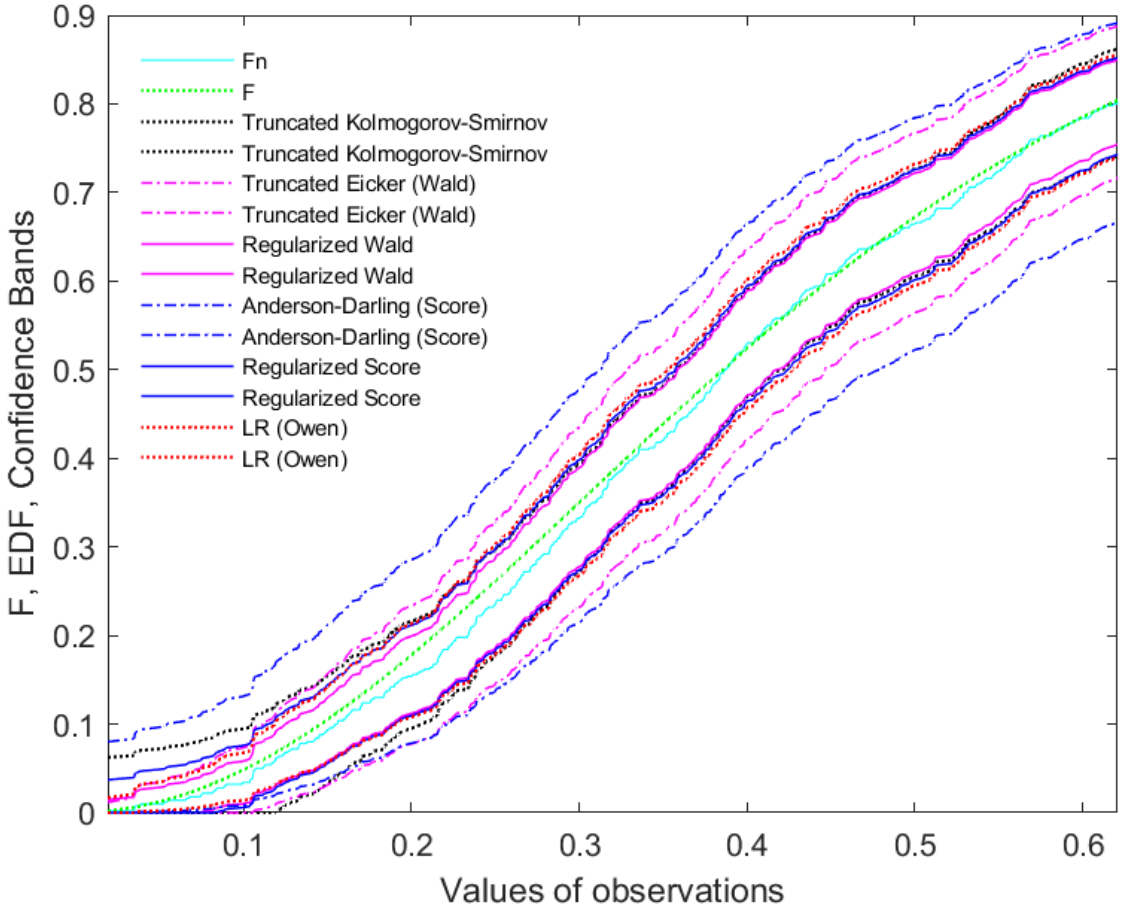
$$c_{LR}(0.01;n) = \begin{cases} \frac{1}{n}\left[5.3318 + 0.5539 \log n - 0.0370 \log^2(n)\right] & \text{for } 1 < n \leq 100, \\ \frac{1}{n}\left[5.6392 + .04018 \log n - 0.0183 \log^2(n)\right] & \text{for } 100 < n \leq 1000. \end{cases} \tag{B.28}$$

C. Performance of tests: Singh-Maddala distribution

C.1. Choice of δ_n: trade-off between the performances of the tests and confidence bands

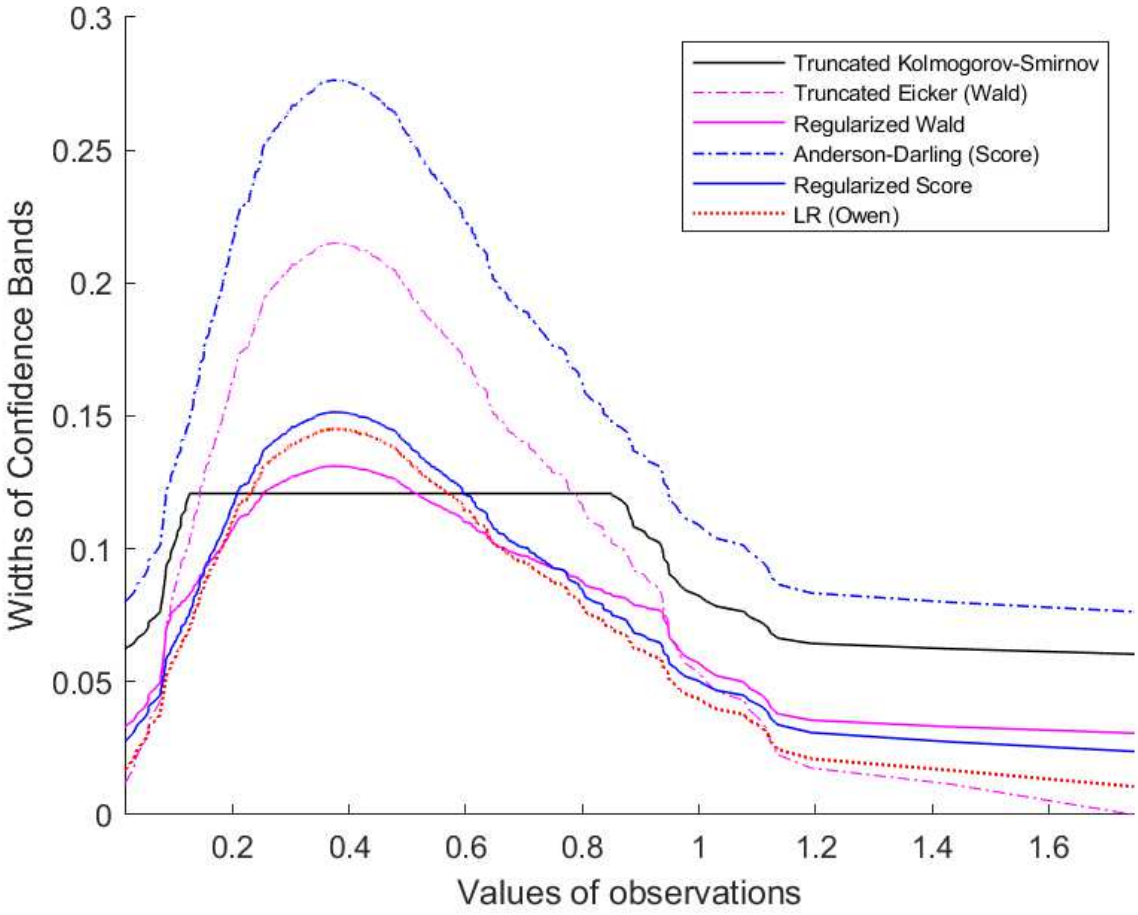
To illustrate the trade-off between the performance of the tests and that of the confidence bands, we simulate the regularized confidence bands using values of parameters $\delta_{500}(S) = 0.001$ and $\delta_{500}(W) = 0.07$. Figures A–1 and A–2 below show that for observations in the center of the distribution, the Kolmogorov-Smirnov band has the smallest width, followed by the regularized-Wald band, the LR band, and the regularized-Score band. The truncated Eicker (Wald) and Anderson-Darling (Score) confidence bands have the largest widths, larger than the widths of the regularized bands. The relative performances of the inference methods change substantially in the tails of the sample. The LR and the regularized-Score bands perform generally the best, followed by the regularized-Wald band. It follows the truncated Eicker confidence band, the truncated KS band, and the truncated Anderson-Darling, which has the largest width amongst the studied confidence bands. With this simulation ($\delta_{500}(S) = 0.001$ and $\delta_{500}(W) = 0.07$), the widths of the regularized bands are much larger than for $\delta_{500}(S) = 0.05$ and $\delta_{500}(W) = 0.01$ (See Figures 1 and 2 in the main text), although the latter values of δ_n do not maximize the power of the regularized tests. This illustrates the trade-offs between performances of the tests and performances of confidence bands in choosing the values of the regularization parameters.

Figure A–1. EDF-based confidence bands for the Singh-Maddala distribution function
Burr$(a = 1, c = 2, k = 5)$



Source: Authors' calculations
Sample size: $n = 500$, level: $1 - \alpha = 0.95$, $\delta_{500}(S) = 0.001$ for the regularized Score-type test, $\delta_{500}(W) = 0.07$ for the regularized Wald-type test, and number of replications to simulate critical points: $N_1 = 1000000$.

Figure A–2. Widths of EDF-based confidence bands for the Singh-Maddala distribution function Burr$(a = 1, c = 2, k = 5)$



Source: Authors' calculations

Sample size: $n = 500$, level: $1 - \alpha = 0.95$, $\delta_{500}(S) = 0.001$ for the regularized Score-type test, $\delta_{500}(W) = 0.07$ for the regularized Wald-type test, and number of replications to simulate critical points: $N_2 = 1000000$.

Table A–1: Critical Values of the regularized-Wald and Score tests for different values of δ_n, $n = 500$

δ_{500}	$c_W(\alpha;\delta_n)$	$c_S(\alpha;\delta_n)$
0.0001	16.37528088649020	4.17824446566989
0.0005	7.36154998914373	3.56522368315369
0.001	5.36547819989887	3.41662932946573
0.005	3.51147983248528	3.16081268057037
0.01	3.22831769744001	3.05280703340687
0.05	2.71936799333338	2.68345849655426
0.07	2.59119208512667	2.56645322481850
0.1	2.44199663292903	2.42481199115093
0.15	2.25246882600000	2.24137497200000
0.2	2.10251561300000	2.09483575100000
0.25	1.98181689200000	1.97607161800000
0.3	1.88147707600000	1.87704608200000
0.35	1.79394441000000	1.79004253300000
0.4	1.71950513900000	1.71629372100000
0.7	1.40757174200000	1.40600396200000
1	1.22161581700000	1.22083933500000
5	0.59082202900000	0.59071731000000
10	0.42224449300000	0.42221946900000

Sample size: $n = 500$; level: $\alpha = 0.05$; $c_W(\alpha;\delta_n)$: critical value of the regularized Wald-type test; $c_S(\alpha;\delta_n)$: critical value of the regularized Score-type test; number of replications to simulate critical points: $N_1 = 3000000$.

D. Performance of tests: normal distribution

This section illustrates the behavior of the sup-EDF goodness-of-fit tests, captured by their level and power, the impact of the regularization using Monte Carlo simulations on these, and the performance of EDF-based confidence bands. The test levels are controlled thanks to exact critical points simulated with $N_1 = 3000000$ replications. The study focuses on comparing the power of the tests. Simulations illustrate the performance of tests on a standard distribution with thin tails, illustrated using the Normal distribution. Simulations using a distribution with a heavier tail, illustrated using the Singh-Maddala distribution, which mimics more closely income or consumption distributions are provided in the paper.

D.1. Performance of the tests

As a first simulation, let's test the null hypothesis $\mathscr{H}_0 : X \sim N(0,1)$ against the alternative $\mathscr{H}_1 : X \sim N(0,1.2)$. The level and power of the tests are simulated using a sample of size $n = 500$ and $N_2 = 15\,000$ Monte Carlo replications, chosen larger than recommended by the literature (Mundform

Table A–2: Level and power of the regularized-Score (AD) test and values of δ_n, $n = 500$

Test of $H_0 : X \sim N(0,1)$ **versus** $H_1 : X \sim \mathrm{N}(0, 1.2)$

δ_{500}	Critical value ($c_{S_\delta}(\alpha)$)	Level (percent)	Power (percent)
0.0001	4.178244466	4.820	97.92666667
0.0005	3.565223683	4.8	99.40666667
0.001	3.416629329	5.213333333	99.54666667
0.005	3.160812681	5.18	99.24
0.05	2.683458497	4.806666667	96.15333333
0.07	2.566453225	4.726666667	95.08666667
0.1	2.424811991	4.76	93.20666667

Sample size: $n = 500$; level: $\alpha = 0.05$; $c_S(\alpha; \delta_n)$: critical value of the regularized Score-type test; number of replications to simulate critical points: $N_1 = 3000\,000$; number of Monte Carlo simulations: $N_2 = 15000$.

et al. (2011)) to give strong confidence in our results.

Table A–2 and A–3 (see below) show the performance of the regularized Score- and Wald-type tests (respectively) for increasing values of δ_{500}. The power of the regularized-Score test is relatively high even for small values of δ_{500}, as the latter prevents the denominator of the statistics from converging to zero in the tails of the sample, and continues improving with values of δ_{500} (Table A–2). In these simulations, the power of the regularized-Score test culminates at 99.55 percent for $\delta_{500} = 0.001$. Similarly, the power of the regularized Wald-type test increases with the value of the regularization parameter (Table A–3), however less quickly than for regularized-Score test. The power culminates at 86.13 percent for $\delta_{500} = 0.05$. We assume for the remainder of the study that $\delta_{S,500} = 0.001$ for the regularized-Score test and $\delta_{W,500} = 0.05$ for the regularized-Wald test.

The second simulation compares the relative performance of all the EDF-based tests. Table A–4 shows that amongst all the tests, the regularized Score-type test achieves the highest power at 99.46 percent. The likelihood-ratio based test achieves the second best power at 98.63 percent, followed by the regularized Wald-type test (power of 85.09 percent). The non-regularized Score and Wald tests have much less power than the regularized tests, demonstrating once more the usefulness of our proposed method. The Kolmogorov-Smirnov statistic, which is unweighted, delivers the least power.

As a last step, we analyze the behavior of the tests with distribution of increasingly heavy tails, using Monte Carlo simulations under the same assumptions as the previous simulations. We test the null hypothesis $\mathscr{H}_0 : X \sim \mathrm{N}(0, 1)$ against the alternative $\mathscr{H}_1 : X \sim \mathrm{N}(0, \sigma)$ for σ ranging from 0.5 to 1.5 in increments of 0.05. Graph 1.1 compares the power of the different tests as σ increases. The results show that the tests perform differently when σ is larger than 1 than when σ is smaller than 1. For $\sigma > 1$ – in other words, when the studied sample actually comes from a distribution with heavier tails than the distribution assumed under $\mathscr{H}_0$, the regularized Score test delivers the best power amongst the tests. The likelihood-ratio test achieves the second best power followed by the regularized Wald-type test. The Kolmogorov-Smirnov and the Score tests perform equivalently, with a

Table A–3: Level and power of the regularized Wald-type (Eicker) test and values of δ_n, $n = 500$

Test of $H_0 : X \sim N(0,1)$ **versus** $H_1 : X \sim \mathrm{N}(0,1.2)$

δ_{500}	Critical value ($c_{S_\delta}(\alpha)$)	Level (percent)	Power (percent)
0.0001	16.37528089	5.186666667	0.006666667
0.0005	7.361549989	5	0.006666667
0.001	5.3654782	5.3	0.766666667
0.005	3.511479832	5.2	70.58
0.05	2.719367993	4.733333333	86.13333333
0.07	2.591192085	4.84	85.50666667
0.1	2.441996633	4.906666667	84.04

Sample size: $n = 500$; level: $\alpha = 0.05$; $c_W(\alpha;\delta_n)$: critical value of the regularized Wald-type test; number of replications to simulate critical points: $N_1 = 3000\,000$; number of Monte Carlo simulations: $N_2 = 15\,000$.

slightly better power than the Wald test. The latter performs the worst amongst the inference methods. Conversely, for $\sigma < 1$ – when the studied sample comes from a distribution with thinner tails than the one assumed under $\mathscr{H}_0$, the Wald-type tests perform better than the other tests. The Wald test achieves the best power. The regularized Wald test achieves the second best power, followed by the likelihood-ratio test. The regularized Score test performs better than the Kolmogorov-Smirnov test, except when the distributions under the null and the alternative assumptions are very close.

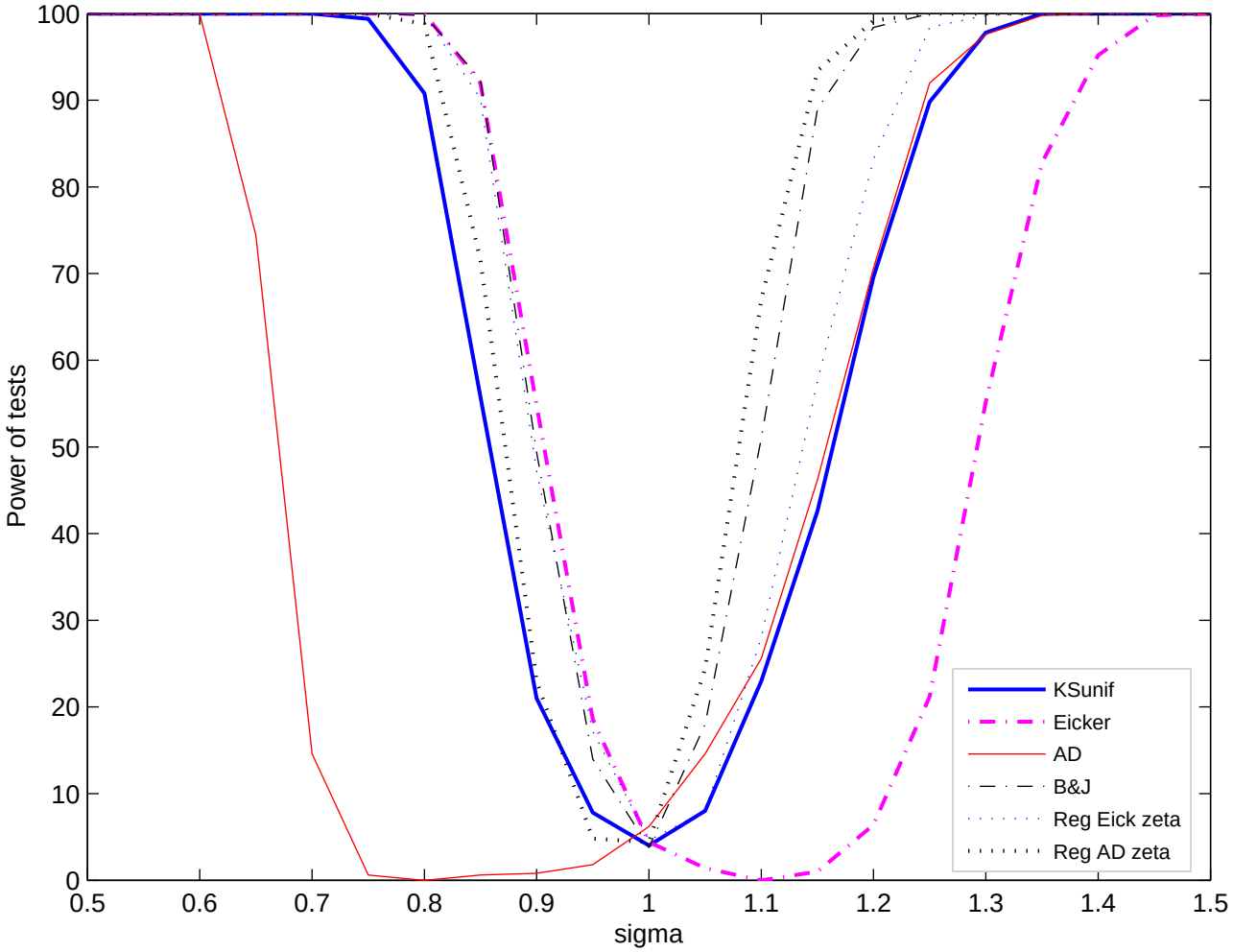
More generally, the results demonstrate that the Wald-type tests allows more discrimination than the Score-type tests when the distribution under $\mathscr{H}_1$ has heavier tails than the one under $\mathscr{H}_0$. In addition, as expected, when the distribution under $\mathscr{H}_1$ is very close to the distribution under $\mathscr{H}_0$, the Kolmogorov-Smirnov test performs better than the weighted statistics. To end, the results show that the non regularized Score and Wald-type are biased. The power of these tests reaches their minimum values when σ is slightly different from 1, rather than when $\sigma = 1$ ($\mathscr{H}_0$ is equivalent to $\mathscr{H}_1$).

Let's highlight another important advantage of the simulation procedures. Usually, regularized statistics are biased due to the distortion the regularization term introduces to the distribution of the original statistics. Simulating the critical values controls the level of the tests and the CBs, which offsets the bias. Likewise, regularizing the statistics also offsets the bias of the non-regularized Score- and Wald-type tests.

D.2. Performance of EDF-based confidence bands

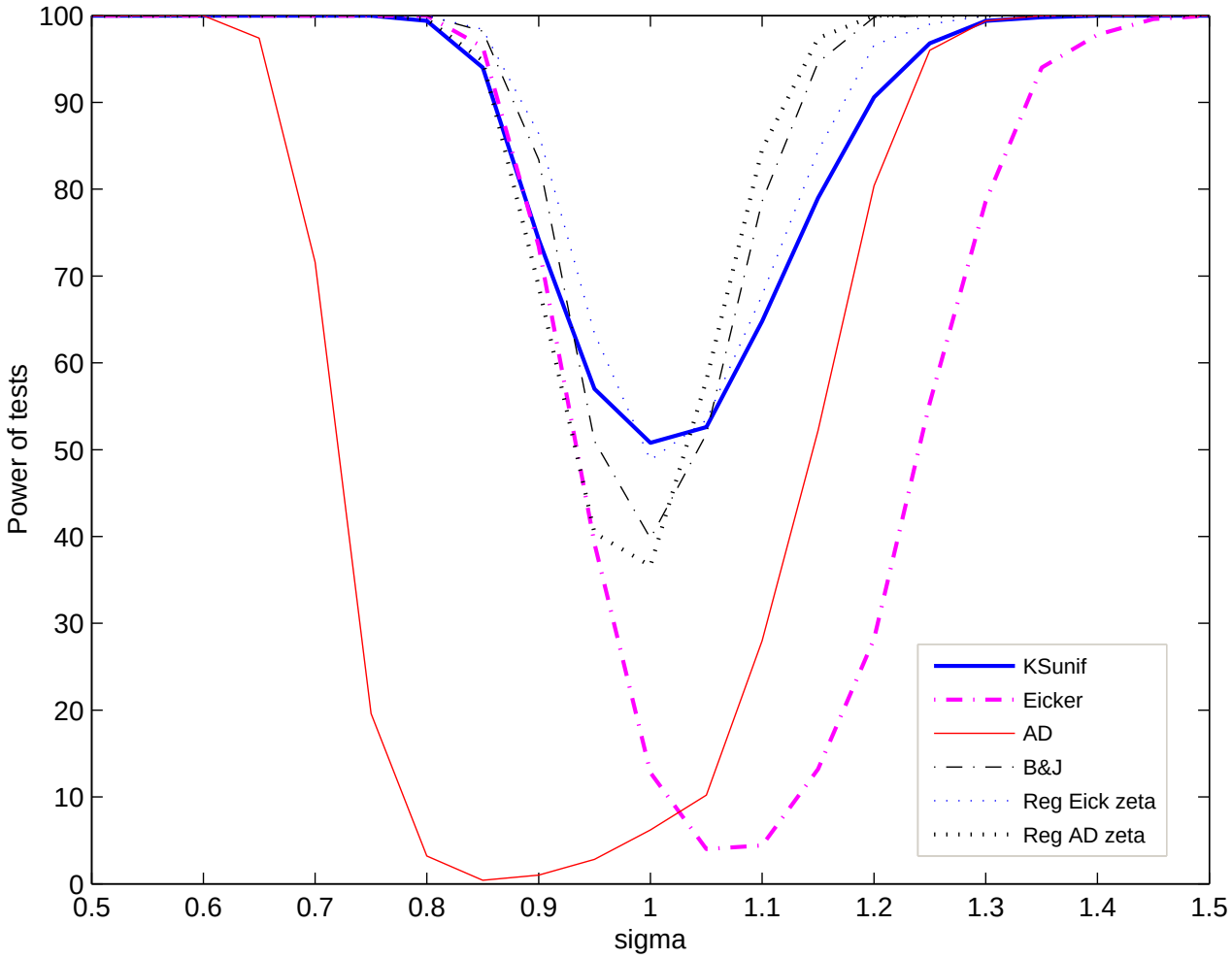
This subsection illustrates how to construct the EDF-based confidence bands in practice and assess their relative performance. Let's suppose a sample of n i.i.d. observations $X_1, \ldots, X_n$. Let's also assume that critical points of the respective test statistics, and sample size n are available, either simulated or taken from the literature, where available. We assume the same values of regularization parameters δ_n as previously.

Figure A–3. Power of EDF-based tests for Normal distributions with the same means
Test of $H_0 : X \sim N(0,1)$ **versus** $H_1 : X \sim \mathrm{N}(0,\sigma)$, **for** $\sigma = 0.5, 0.55, 0.6, \ldots, 1.5$



Sample size: $n = 500$; level: $\alpha = 0.05$; $\delta_{500}(S) = 0.001$ for the regularized Score-type test; $\delta_{500}(W) = 0.07$ for the regularized Wald-type test; number of Monte Carlo simulations: $N_1 = 3000000$; number of replications to simulate critical points: $N_2 = 15000$.

Figure A–4. Power of EDF-based tests for Normal distributions with different means
Test of $H_0 : X \sim N(0,1)$ **versus** $H_1 : X \sim N(0.1, \sigma)$, **for** $\sigma = 0.5, 0.55, 0.6, \ldots, 1.5$



Sample size: $n = 500$; level: $\alpha = 0.05$; $\delta_{500}(S) = 0.001$ for the regularized Score-type test; $\delta_{500}(W) = 0.07$ for the regularized Wald-type test; number of replications to simulate critical points: $N_1 = 3000\,000$; number of Monte Carlo simulations: $N_2 = 15\,000$.

Table A–4: **Level and power of EDF-based tests, sample of size** $n = 500$
Test of $H_0 : X \sim N(0,1)$ **versus** $H_1 : X \sim \mathrm{N}(0, 1.2)$

Tests	$c(\alpha)$	Level (percent)	Power (percent)
Kolmogorov-Smirnov	0.06039953611469	4.60	68.44
Wald	4.79617625286106	5.35	6.25
Regularized Wald	2.59103603317055	4.64	85.09
Score	6.45272451410767	4.83	71.81
Regularized Score	3.42243384382137	4.83	99.46
Berk-Jones	0.01138040175450	4.57	98.63

Sample size: $n = 500$; level: $\alpha = 0.05$; $\delta_{500}(S) = 0.001$ for the regularized Score-type test; $\delta_{500}(W) = 0.07$ for the regularized Wald-type test; number of replications to simulate critical points: $N_1 = 3000000$; number of Monte Carlo simulations: $N_2 = 15000$.

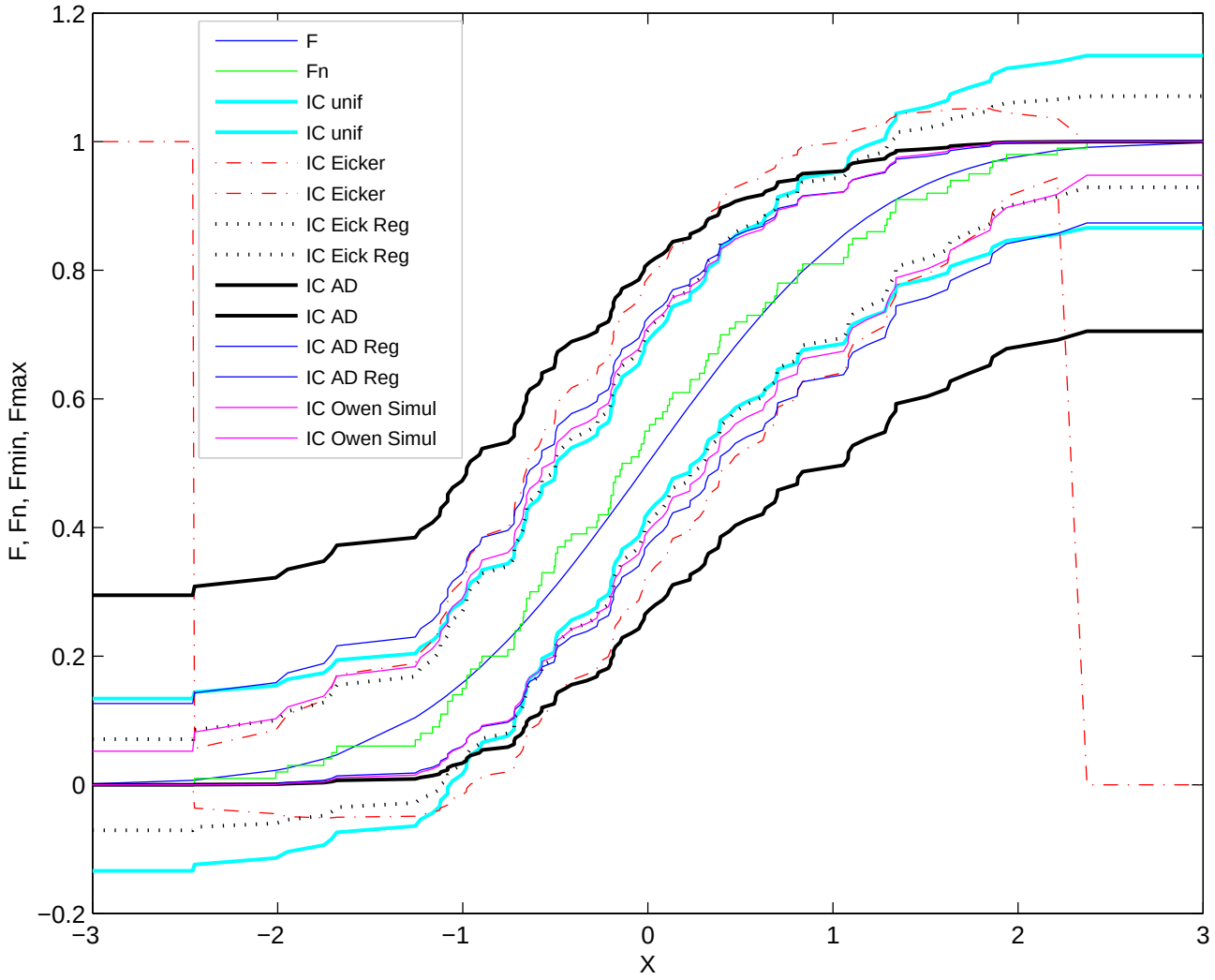
As a first illustration, we build confidence bands for a Normal distribution $\mathrm{N}(0,1)$ and sample of size $n = 100$.The small number of observations allows having a clear graph that helps better compare the confidence bands.

Figure A–5 shows that for observations in the center of the distribution, the Kolmogorov-Smirnov CB has the smallest width among the EDF-based CBs, closely followed by the regularized Wald-type CB. The LR-type and regularized Eicker-type CBs achieve the best following performance. The widths of the confidence bands based on non-regularized Wald and Score tests are the largest amongst the studied inference methods in the center of the distribution.

In the tails of the distribution, the ranking of the inference methods changes a lot. The width of the KS-based CB is larger than that of other CBs. The regularized Score-type CB performs the best followed by the regularized Wald-type CB. The Wald statistic provides a CB whose width tends to zero as observations go further from the center of the distribution. However, owing to its discontinuity at the first and last observations, for all values of x lower than the first observation $(x < X_{(1)})$ or greater or equal to the last observation $(x \geq X_{(n)})$, the CB becomes the non informative $[0,1]$ interval. At the opposite, the Score CB is always continuous. Its width equals to $c_S^2(\alpha;n)/(n + c_S^2(\alpha;n))$ in the tail of the sample which is to compare to $2\ c_{KS}(\alpha;n)$, the constant width of the Kolmogorov-Smirnov CB. Adding the regularization term to the Wald statistic corrects the discontinuity problem of the Wald CB. Nevertheless, for δ_n different from zero, the width of the corresponding CB does not reach zero anymore but performs well, with a width—equal to $2c_W^2(\alpha;n)\zeta^{1/2}/n^{1/2}-$ around half the width of the uniform CB for $n = 100$. Likewise, adding the regularization term to the Score statistic improves a lot the performance of the CB, in particular in the tails. We simulate the critical values of the statistics for sample sizes from 50 to 1000 using $N = 1000000$ replications. The results (see Table A–5) shows that for n smaller than 150, the KS confidence band performs better than the Anderson-Darling CB in the very end of the tails of continuous distributions but the Score CB becomes better in the tails for n greater than 150. The statistics using regularization parameters yield CBs with smaller widths than those without regularization and than the uniform CB.

Overall, as expected, weighted KS statistics achieve better performance than the unweighted one

Figure A–5. EDF-based confidence bands for the distribution function $N(0,1)$



Sample size: $n = 100$; level: $1 - \alpha = 0.95$; $\delta_{100}(S) = 0.001$ for the regularized Score-type test; $\delta_{100}(W) = 0.07$ for the regularized Wald-type test; number of replications to simulate critical points: $N_1 = 3000000$; number of Monte Carlo simulations: $N_2 = 15000$.

at some observations of the sample and for large and moderately large samples, but do not clearly dominate the latter. Conversely, the CBs based on the regularized statistics deliver smaller widths than the other statistics in the tails of distributions, for all sample sizes. Moreover, the regularized Score-type CB performs the best CB amongst the EDF-based CBs–the widths of the regularized Score-type CB in the center of the distribution is very close to the width of the KS CB while it achieves the best width in the tails.

When comparing the KS based CBs to the Owen one—using simulated critical points which allow to control the level of the test and CBs, it appears that in the center of the distribution, the Owen CB perform worse than the uniform CB and the regularized Anderson Darling-type CB but better than the regularized Eicker-type CB. Conversely, in the tails of the distribution, the Owen CB overcomes all the other CBs. However, the Owen critical has a major drawback: it is very computationally demanding. Building the Owen CB requires to perform as many optimizations as there are observations, which is time demanding and condition the performance of the inference method to that of the optimization method used.

Table A–5: Simulated critical points of EDF-based statistics and width of the corresponding confidence bands in the tails of distributions
$n = 50, 100, \ldots, 1000$ **using** $N = 1000000$ **replications**

A. Critical Points

	n		
	50	100	150
KS	0.188335597	0.134056205	0.109638527
E	4.522280174	4.666516217	4.70725121
$E_{\delta n}$	2.793607145	2.673939389	2.635861357
AD	6.43881938	6.457275098	6.474963046
$AD_{\delta n}$	4.14789771	3.780490269	3.681454542
BJ	0.104133743	0.053727502	0.036421672

	n			
	200	250	500	1000
KS	0.095178969	0.085207603	0.060368076	0.042771814
E	4.739536779	4.763956841	4.798134574	4.823671809
$E_{\delta n}$	2.619270234	2.610421227	2.587466879	2.580100277
AD	6.458800396	6.45095347	6.454634534	6.444683756
$AD_{\delta n}$	3.600576974	3.540729297	3.424467947	3.352876861
BJ	0.027593217	0.022293228	0.011369705	0.005804824

B. Width of CBs in the tails of distributions

n	KS	E	$E_{\delta n}$	AD	$AD_{\delta n}$	BJ
50	0.38	1.00	0.21	0.45	0.37	0.10
100	0.27	1.00	0.14	0.29	0.23	0.05
150	0.22	1.00	0.11	0.22	0.17	0.04
200	0.19	1.00	0.10	0.17	0.14	0.03
250	0.17	1.00	0.09	0.14	0.13	0.02
500	0.12	1.00	0.06	0.08	0.08	0.01
1000	0.09	1.00	0.04	0.04	0.06	0.01

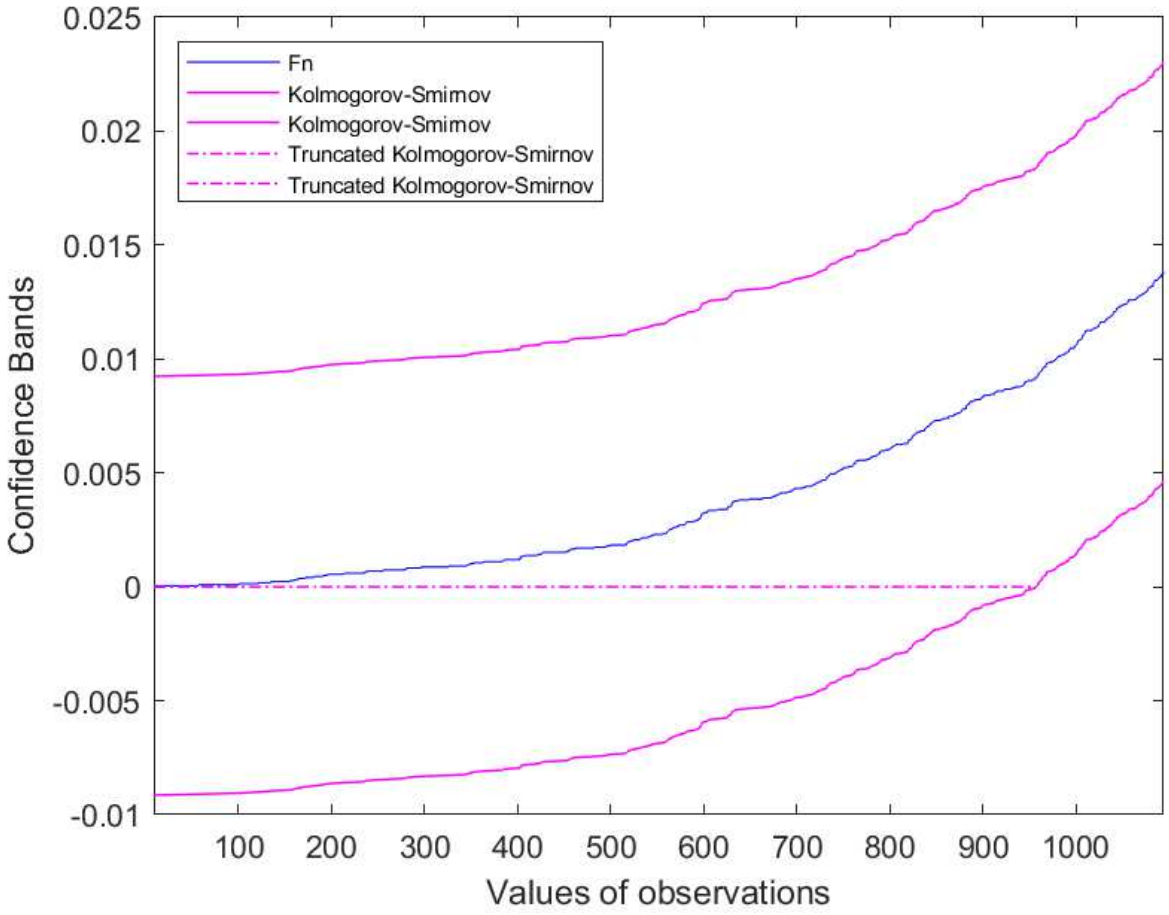
E. Kenyan consumption: additional results

Table A–6: Simulated critical points of EDF-based statistics

Test statistic	Critical Points
Kolmogorov-Smirnov	0.009191889148555
Eicker (Wald)	4.840073009947004
Regularized-Wald	2.690892495171526
Anderson-Darling (Score)	6.450506803795759
Regularized-Score	3.280057922215571
Owen (LR)	0.00028060465474

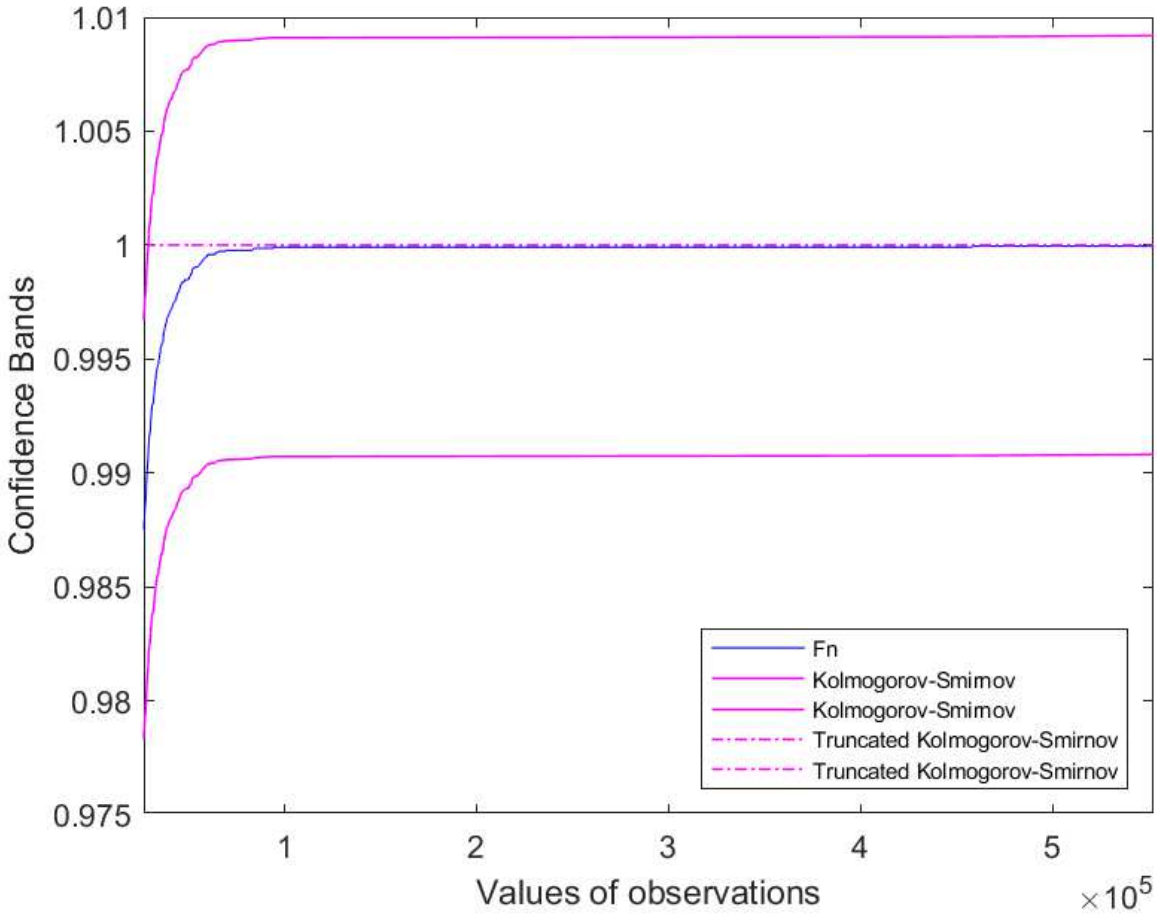
Sample: $n = n_T = 21773$, level: $1-\alpha = 0.95$, $\delta_{n_T}(S) = 0.001$ and $\delta_{n_T}(W) = 0.05$, and number of replications to simulate critical points: $N_2 = 1000000$.

Figure A–6. Kolmogorov-Smirnov and Truncated Kolmogorov-Smirnov Confidence Bands for the Kenyan Households' Adult Consumption (small observations)



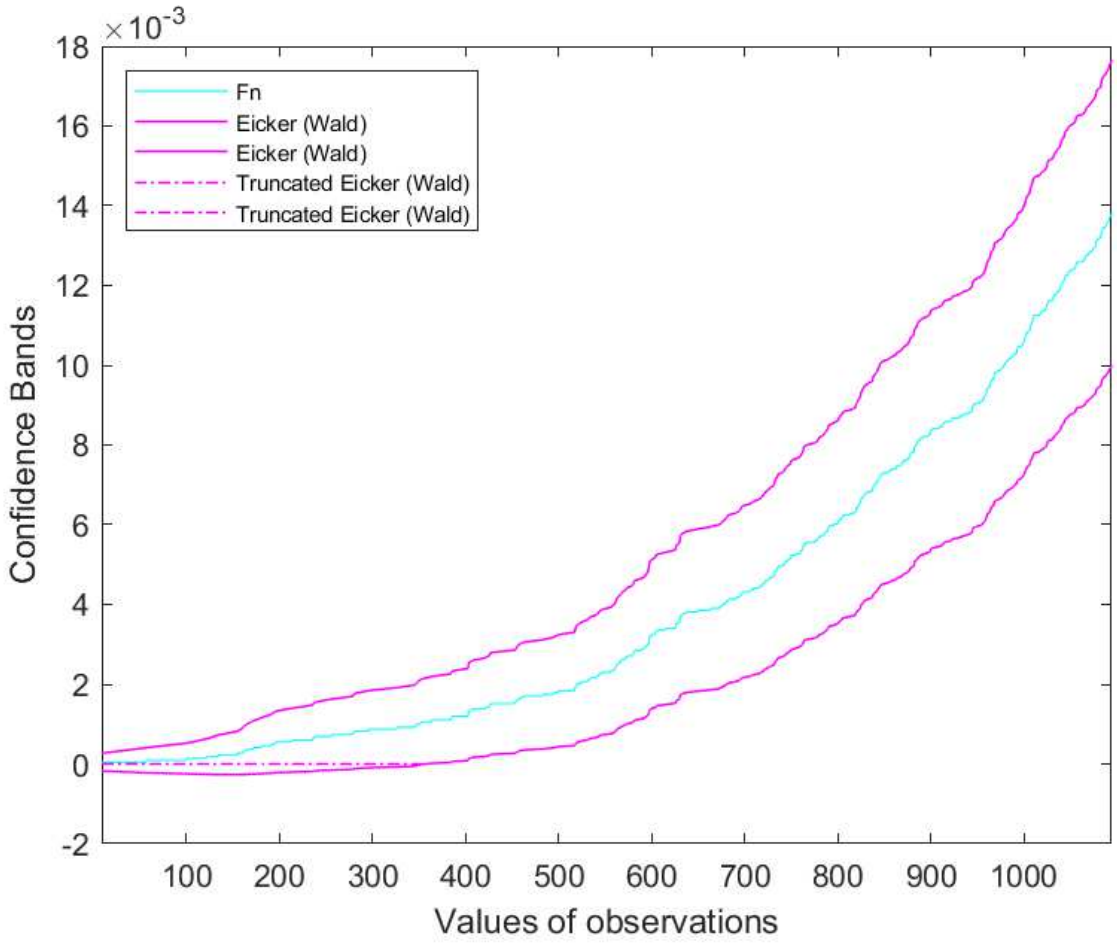
Sample: $n = 1, ..., 300$, level: $1 - \alpha = 0.95$, and number of replications to simulate critical points: $N_2 = 1000000$.

Figure A–7. Kolmogorov-Smirnov and Truncated Kolmogorov-Smirnov Confidence Bands for the Kenyan Households' Adult Consumption (largest observations)



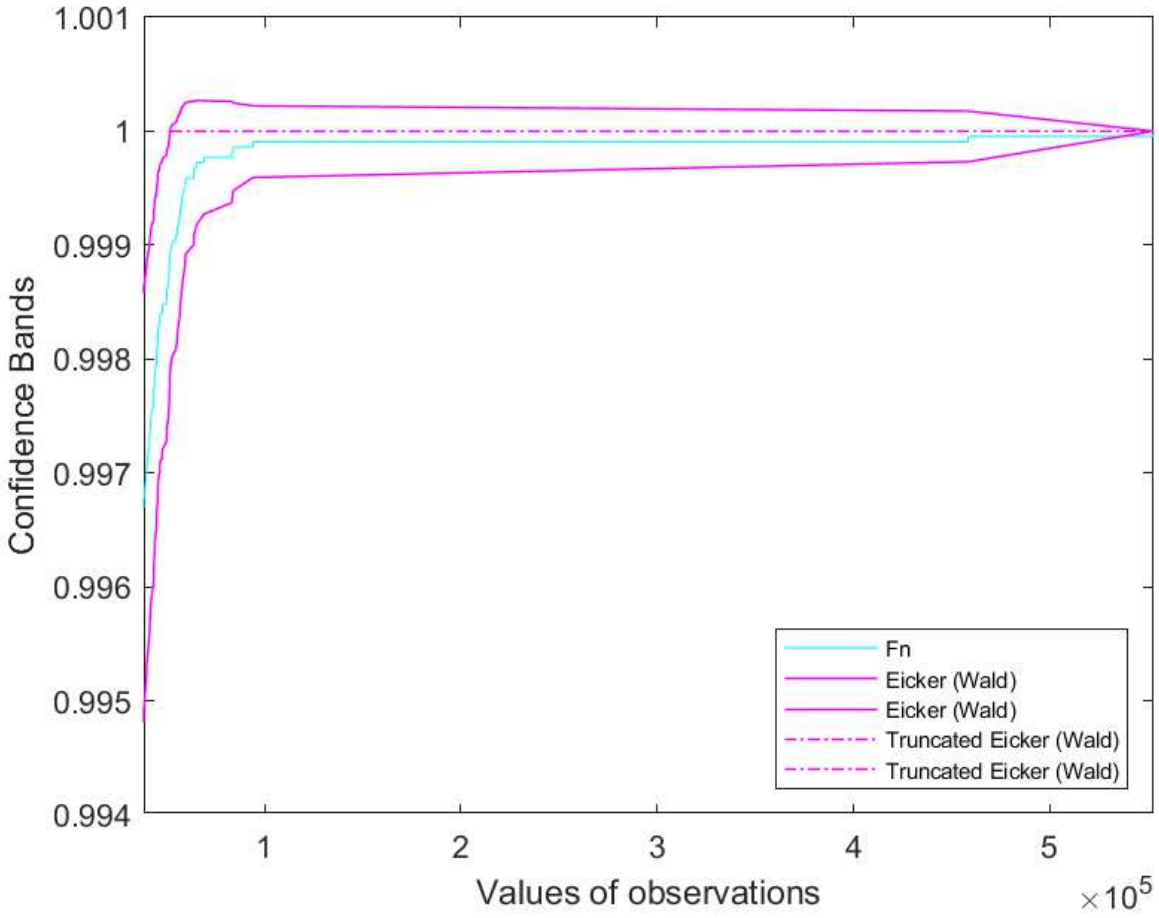
Sample: $n = 21,500,...,n_T$, level: $1-\alpha = 0.95$, and number of replications to simulate critical points: $N_2 = 1000000$.

Figure A–8. Eicker and Truncated Eicker Confidence Bands for the Kenyan Households' Adult Consumption (small observations)



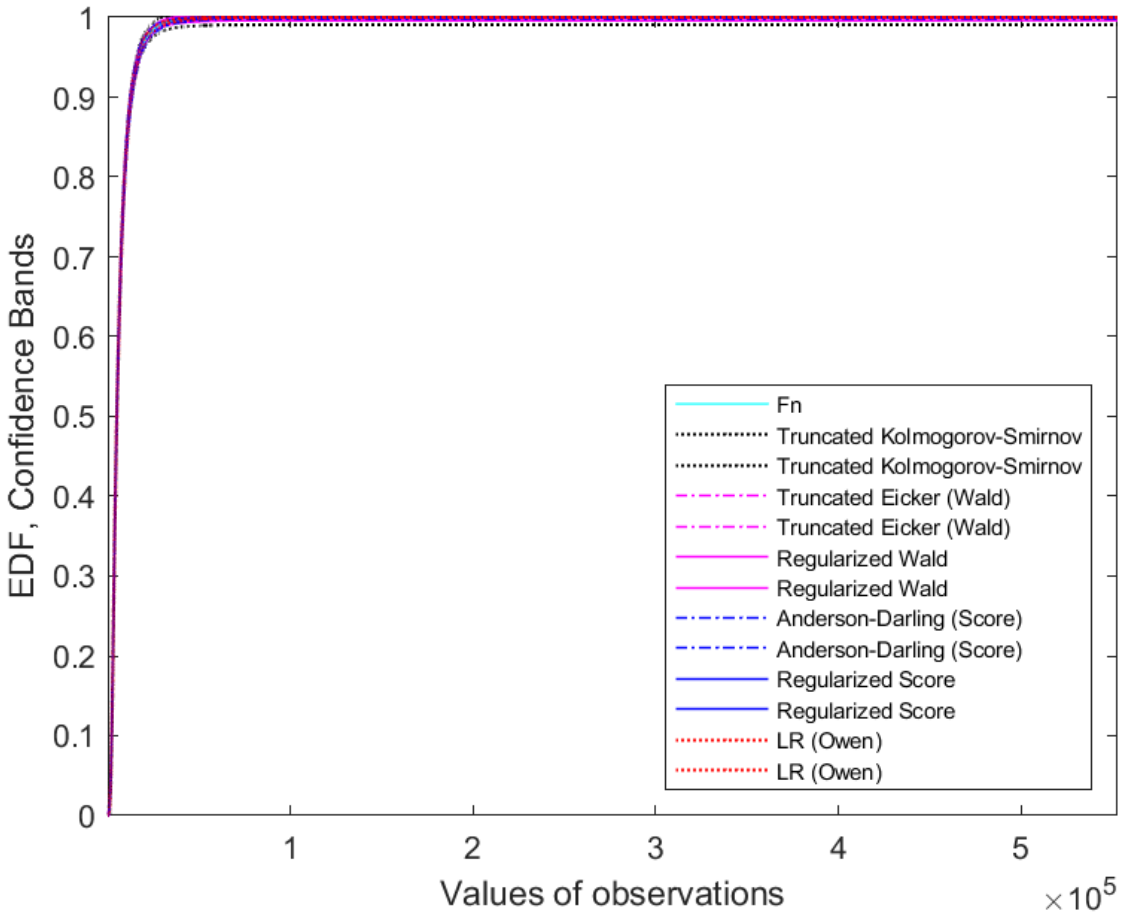
Sample: $n = 1, ..., 300$, level: $1 - \alpha = 0.95$, and number of replications to simulate critical points: $N_2 = 1000000$.

Figure A–9. Eicker and Truncated Eicker Confidence Bands for the Kenyan Households' Adult Consumption (largest observations)



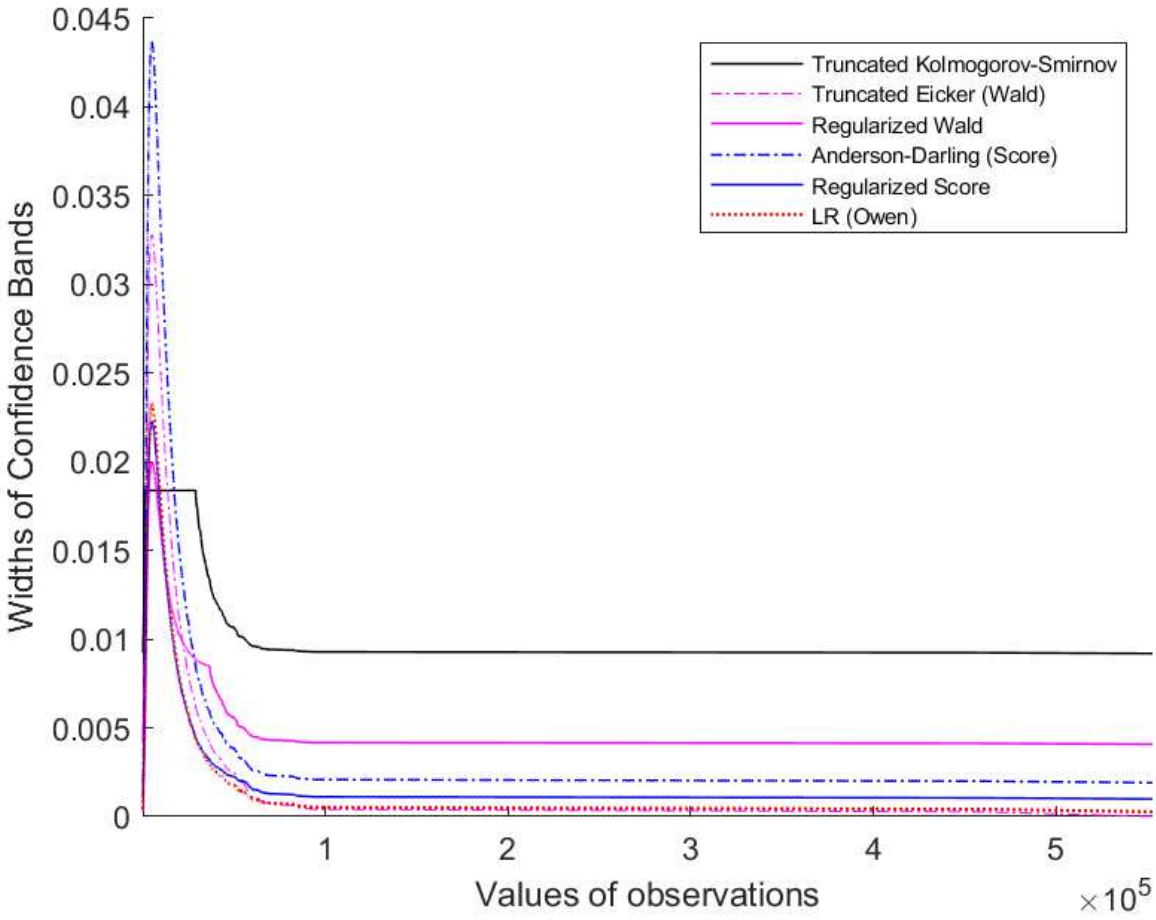
Sample: $n = 21,500,...,n_T$, level: $1-\alpha = 0.95$, and number of replications to simulate critical points: $N_2 = 1000000$.

Figure A–10. EDF-based Confidence Bands for the Kenyan Households' Adult Consumption



Sample: $n = 1, \ldots, n_T$, level: $1 - \alpha = 0.95$, $\delta_{n_T}(S) = 0.001$ and $\delta_{n_T}(W) = 0.05$, and number of replications to simulate critical points: $N_2 = 1000000$.

Figure A–11. Widths of EDF-based Confidence Bands for the Kenyan Households' Adult Consumption



Sample: $n = 1,...,n_T$, level: $1-\alpha = 0.95$, $\delta_{n_T}(S) = 0.001$ and $\delta_{n_T}(W) = 0.05$, and number of replications to simulate critical points: $N_2 = 1000000$.