Advances in Complex Systems, Vol. 12, No. 2 (2009) 157–167

ON APPROXIMATING THE DISTRIBUTIONS OF GOODNESS-OF-FIT TEST STATISTICS BASED ON THE EMPIRICAL DISTRIBUTION FUNCTION: THE CASE OF UNKNOWN PARAMETERS

MARCO CAPASSO

Tjalling C. Koopmans Research Institute, and
Urban and Regional Research Centre Utrecht
Utrecht, The Netherlands

LUCIA ALESSI

European Central Bank, Frankfurt, Germany, and
Laboratory of Economics and Management
Sant'Anna School of Advanced Studies
Pisa, Italy

MATTEO BARIGOZZI

Max Planck Institute of Economics, Jena, Germany, and
Laboratory of Economics and Management
Sant'Anna School of Advanced Studies
Pisa, Italy

GIORGIO FAGIOLO*

Laboratory of Economics and Management
Sant'Anna School of Advanced Studies
Pisa, Italy
giorgio.fagiolo@sssup.it

Received 25 April 2008
Revised 20 October 2008

This paper discusses some problems possibly arising when approximating via Monte-Carlo simulations the distributions of goodness-of-fit test statistics based on the empirical distribution function. We argue that failing to re-estimate unknown parameters on each simulated Monte-Carlo sample — and thus avoiding to employ this information to build the test statistic — may lead to wrong, overly-conservative testing.

*Corresponding author. Address: Sant'Anna School of Advanced Studies, Laboratory of Economics and Management, Piazza Martiri della Libertà 33, I-56127 Pisa, Italy. Tel.: +39-050-883359, Fax: +39-050883344.

Furthermore, we present some simple examples suggesting that the impact of this possible mistake may turn out to be dramatic and does not vanish as the sample size increases.

Keywords: Goodness-of-fit tests; critical values; Anderson–Darling statistic; Kolmogorov–Smirnov statistic; Kuiper statistic; Cramér–Von Mises statistic; empirical distribution function; Monte-Carlo simulations.

PACS Classification: 02.50.Ng, 02.70.Uu, 05.10.Ln

1. Introduction

This paper discusses some problems possibly arising when approximating — via Monte-Carlo simulations — the distributions and critical values of the most commonly employed goodness-of-fit (GoF) tests based on empirical distribution function (EDF) statistics [3, 21].

This situation arises very frequently — in many areas of statistical physics or econophysics — when the researcher aims at fitting some experimental or empirical (univariate) sample with a parametric (univariate) probability distribution whose parameters are unknown. In such cases, the goodness of fit may be ex-post evaluated by employing standard statistical tests based on the EDF. However, critical-value tables are only available for testing particular distributions (e.g. normal, exponential, etc.). If, as it typically happens, critical-value tables are not available, one has to resort to Monte-Carlo methods to derive the approximated distribution of the test statistics under analysis.

Some issues related to this problem have been discussed from a theoretical perspective in Refs. 20 and 2. Here we take a more applied approach and we show that, when testing with unknown parameters, the Monte-Carlo procedure employed to approximate the critical values must involve the maximum-likelihood (ML) re-estimation of unknown parameters on each simulated sample. The common procedure of estimating the unknown parameters once and for all at the beginning of the Monte-Carlo procedure leads to wrong, overly-conservative hypothesis tests.

The rest of this paper is organized as follows. Section 2 formalizes the general GoF test under study and discusses the main problems associated with the approximation of EDF-based GoF test statistic distributions from a theoretical perspective. Section 3 presents an application to the case of normality with unknown parameters. Finally, Sec. 4 discusses the main findings of the paper and concludes with a few remarks.

2. Approximating EDF-Based GoF Test Statistic Distributions

In many applied contexts, the researcher faces the problem of assessing whether an empirical univariate sample $\underline{x}_N = (x_1, \ldots, x_N)$ comes from a (continuous) distribution $F(x; \underline{\theta})$, where $\underline{\theta}$ is a vector of *unknown* parameters. EDF-based GoF tests [3, 21] employ statistics that are non-decreasing functions of some distance

between the theoretical distribution under the null hypothesis $H_0 : \underline{x}_N \sim F(x; \underline{\theta})$ and the empirical distribution function constructed from $\underline{x}_N$, provided that some estimate of the unknown parameters is given.

In what follows, we will begin by focusing on the simplest case where $F(x; \underline{\theta})$ has only location and scale unknown parameters. Furthermore, we will limit the analysis to four out of the most used EDF test statistics, namely Kolmogorov–Smirnov [15, 16], Kuiper [13], Cramér–Von Mises [17] and Quadratic Anderson–Darling [1], with small-sample modifications usually considered in the literature.[a]

It is well-known that if one replaces $\underline{\theta}$ with its maximum likelihood (ML) empirical-sample estimate $\hat{\underline{\theta}}(\underline{x}_N)$, the distribution of the EDF test statistic under study can be shown to be independent of the unknown true parameter values [5]. However, test statistic distributions are hard to derive analytically. They must therefore be simulated via Monte Carlo and critical values must be computed accordingly. To do so, let us consider a first possible procedure:

Procedure A

Step A1 *Generate a sufficiently large number (say, $M \gg 0$) of independently-drawn N-sized samples $\underline{z}_N^j = (z_1^j, \ldots, z_N^j)$, $j = 1, \ldots, M$, where each z_i^j is an i.i.d. observation from a $F(x; \hat{\underline{\theta}}(\underline{x}_N))$, i.e. from the distribution under H_0 where unknown parameters are replaced by their empirical-sample estimates.*

Step A2 *For each N-sized sample $\underline{z}_N^j$, compute the relevant EDF test statistic by comparing the EDF built from $\underline{z}_N^j$ with the theoretical distribution $F(\underline{z}_N^j, \hat{\underline{\theta}}(\underline{x}_N))$ where the unknown parameters are always replaced with the corresponding estimates obtained once and for all from the empirical sample. Repeat for all the M samples.*

Step A3 *Compute the empirical distribution function T of the test statistic.*

Step A4 *Compute (upper-tailed) critical values, for any given significance level α, by employing the empirical distribution function T of the EDF test statistic as obtained in Step A3.*

Procedure A is not correct, in the sense that it generates a completely wrong approximation to the "true" distribution of the test statistic under the null hypothesis. The reason why Procedure A is not correct lies in Step A2. More precisely, when we compare the EDF constructed from $\underline{z}_N^j$ with the theoretical distribution $F(\underline{z}_N^j, \hat{\underline{\theta}}(\underline{x}_N))$, we are assuming that our estimate for $\underline{\theta}$ does not depend on the actual sample $\underline{z}_N^j$ under analysis. This is the same as presuming that the hypothesis test is performed for *known parameters*. On the contrary, sticking to the null hypothesis implies that the theoretical distribution which should be compared to the

[a]For more formal definitions, see Ref. 3, Chap. 4, Table 4.2. Small-sample modifications have been applied to benchmark our results to those presented in the literature. However, our main findings remain qualitatively unaltered if one studies test statistic distributions without small sample modifications.

EDF of $\underline{z}_N^j$ must have parameter estimates that depend on the actual Monte-Carlo sample $\underline{z}_N^j$. In other words, scale and location parameters $\underline{\theta}$ must be re-estimated (via, e.g. ML) *each time we draw the Monte-Carlo sample.* Let $\hat{\underline{\theta}}(\underline{z}_N^j)$ be such an estimate for sample j. This means that the theoretical distribution to be used to compute the test statistic would be $F(\underline{z}_N^j, \hat{\underline{\theta}}(\underline{z}_N^j))$ and not $F(\underline{z}_N^j, \hat{\underline{\theta}}(\underline{x}_N))$. The correct procedure therefore reads:

Procedure B

Step B1 *Same as A1.*

Step B2 *For each N-sized sample $\underline{z}_N^j$, compute the relevant EDF test statistic by comparing the EDF built from $\underline{z}_N^j$ with the theoretical distribution $F(\underline{z}_N^j, \hat{\underline{\theta}}(\underline{z}_N^j))$, where the unknown parameters are replaced with estimates $\hat{\underline{\theta}}(\underline{z}_N^j)$ obtained from the j-th Monte-Carlo sample. Repeat for all the M samples.*

Step B3 *Same as A3.*

Step B4 *Same as A4.*

In the next section, we show that the shifts in the Monte-Carlo approximation to the distribution of the test statistics under study are significant when we compare Procedures A and B.

3. Application: Testing for Normality with Unknown Parameters

Let us consider the null hypothesis that the empirical sample comes from a normal distribution $N(\mu, \sigma)$ with unknown mean (μ) and standard deviation (σ). In such a case, parameters may be replaced by their ML estimates $(\hat{m}(\underline{x}_N), \hat{s}(\underline{x}_N))$, i.e. sample mean and standard deviation. For the normal distribution, critical values for the four test statistics under study are already available. Our goal is therefore to compare Monte-Carlo approximations to the distributions of the four test statistics obtained under Procedures A and B.

We thus have two setups. In the first one (Procedure A), we do not re-estimate the parameters and we always employ $(\hat{m}(\underline{x}_N), \hat{s}(\underline{x}_N))$ to build the theoretical distribution. In the second one (Procedure B), we re-estimate via ML mean and standard deviation on each simulated sample by computing $(\hat{m}(\underline{z}_N^j), \hat{s}(\underline{z}_N^j))$ and then use them to approximate the theoretical distribution of the test statistic.

Our simulation strategy is very simple. Since the argument put forth above does not depend on the observed sample's mean and standard deviation, we can suppose that $(\hat{m}(\underline{x}_N), \hat{s}(\underline{x}_N)) = (0, 1)$ without loss of generality.[b] For each of the four test statistics considered, we run Monte-Carlo simulations to proxy its distribution

[b]Alternatively, one can standardize the observed sample and generate Monte-Carlo sample replications from a $N(0, 1)$ without loss of generality.

under the two setups above.[c] In both setups, we then compute the critical values of the test statistics associated to any significance level (or p-value).

To begin with, Table 1 shows critical values for all 4 tests at $\alpha = 0.05$ significance level, and for different combinations of N (sample size) and M (Monte-Carlo replications). If we employ Procedure B, we obtain the same critical values published in the relevant literature for the case of normality with *unknown* parameters (compare, e.g. our Table 1 with Table 1A-1.3 on p. 732 in Ref. 18). On the contrary, if we employ procedure A, critical values dramatically increase. The effect is of course more evident in the case of so-called "quadratic statistics" (Cramér–Von Mises and Quadratic Anderson–Darling), but is equally relevant also in the case of "supremum statistics" (Kolmogorov–Smirnov and Kuiper). What is more, Procedure A allows us to obtain critical-value figures which are very similar to those found in the literature for the case of normality with *completely specified, known parameters*.

The figures in Table 1 imply also that if we wrongly employ Procedure A, we end up with test statistics that are dramatically more conservative (at $\alpha = 0.05$) than if we correctly employ Procedure B. This is true irrespective of the significance level. As Fig. 1 shows, the A versus B gap between critical values remains relevant

Table 1. The normal distribution. Critical values at significance level $\alpha = 0.05$ for the four EDF tests considered. Procedure A: Always using empirical-sample estimates. Procedure B: Parameters are re-estimated each time on Monte-Carlo sample. KS = Kolmogorov–Smirnov; KUI = Kuiper; CVM = Cramér–Von Mises; AD2 = Quadratic Anderson–Darling.

		KS		KUI		CVM		AD2	
N	M	Proc B	Proc A	Proc B	Proc A	Proc B	Proc A	Proc B	Proc A
10	100	0.8220	1.2252	1.4235	1.6138	0.0690	0.2807	0.7194	2.2727
	1000	0.8564	1.3035	1.4485	1.7270	0.0706	0.3707	0.6996	2.6139
	10000	0.8648	1.3482	1.4565	1.7314	0.0757	0.3800	0.7442	2.7667
50	100	0.8159	1.4965	1.4376	1.7671	0.0985	0.5941	0.6863	3.1936
	1000	0.8928	1.3813	1.4783	1.7715	0.1111	0.4512	0.7424	2.4106
	10000	0.8931	1.3623	1.4867	1.7489	0.1146	0.4446	0.7553	2.5414
100	100	0.9380	1.4601	1.5831	1.7954	0.1308	0.4902	0.7664	2.7079
	1000	0.8839	1.3525	1.5083	1.7162	0.1129	0.4634	0.7172	2.6332
	10000	0.8969	1.3587	1.4933	1.7407	0.1199	0.4478	0.7425	2.4740
500	100	0.9177	1.2389	1.531	1.7857	0.1146	0.3455	0.7231	2.0889
	1000	0.9125	1.3316	1.5139	1.7250	0.1273	0.4333	0.7769	2.3941
	10000	0.9108	1.3576	1.4998	1.7544	0.1261	0.4628	0.7543	2.5423
1000	100	0.9443	1.3723	1.4753	1.7879	0.1312	0.4116	0.7998	2.3795
	1000	0.9041	1.3858	1.5103	1.7814	0.1246	0.4674	0.7700	2.5817
	10000	0.9121	1.3606	1.5108	1.7575	0.1267	0.4578	0.7581	2.5076
5000	100	0.9689	1.2911	1.5656	1.8264	0.1405	0.4107	0.8099	2.2523
	1000	0.8944	1.3807	1.4801	1.7478	0.1204	0.4513	0.7180	2.4596
	10000	0.9116	1.3606	1.5120	1.7424	0.1274	0.4617	0.7613	2.5073

[c]All simulations are performed using MATLAB®, version 7.4.0.287 (R2007a).

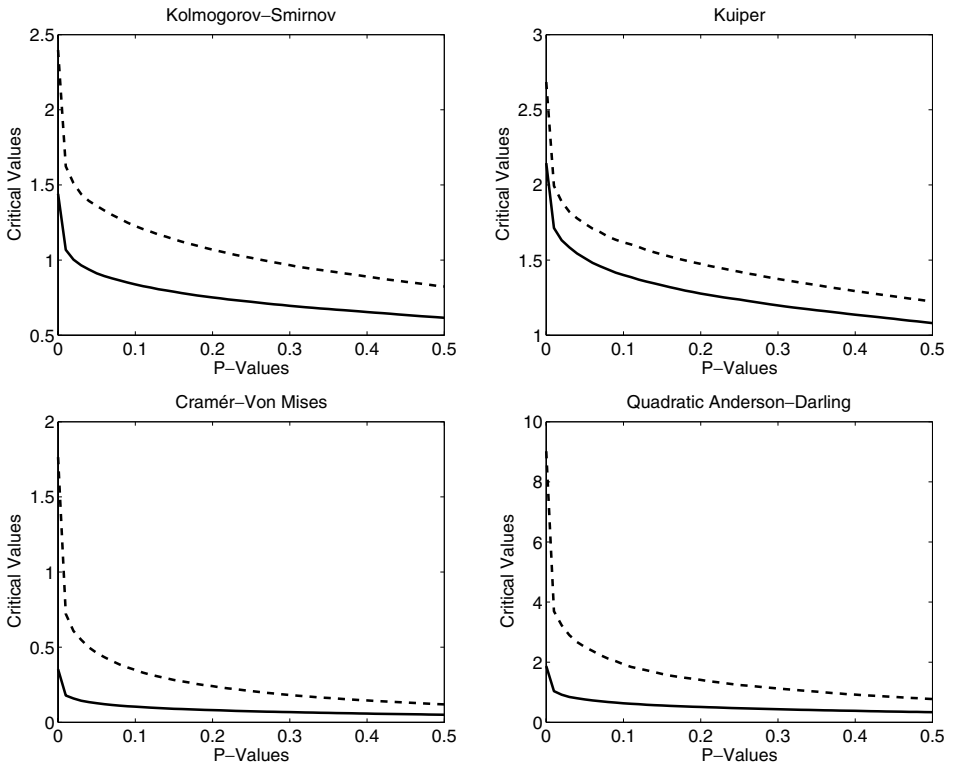

Fig. 1. The normal distribution. Critical values versus p-values for the four test statistics under study. Empirical sample size: $N = 5000$. Number of Monte-Carlo replications: $M = 10000$. Solid line: Procedure B (parameters are re-estimated each time on Monte-Carlo sample). Dashed line: Procedure A (always using empirical-sample estimates).

for all (reasonable) p-value levels. In other words, the wrong choice of employing Procedure A induces a shift to the right of (and reshapes) the entire test statistic distribution. To see this, in Fig. 2 we plot the estimated cumulative distribution of all 4 test statistics under the two setups.

It is worth noting that the above results do not depend on the empirical sample size. In fact, one might argue that the mismatch between the two procedures may be relevant only for small N's but should vanish as N gets large. This is not true: the gap remains there as N increases within an empirically-reasonable range and for any sufficiently large number of Monte-Carlo replications (M) — see Fig. 3 for the case $M = 10000$.

4. Discussion and Concluding Remarks

In this paper, we have argued that failing to re-estimate unknown parameters on each simulated Monte-Carlo sample (and not employing this information to compute the theoretical distribution to be compared with the sample EDF) may

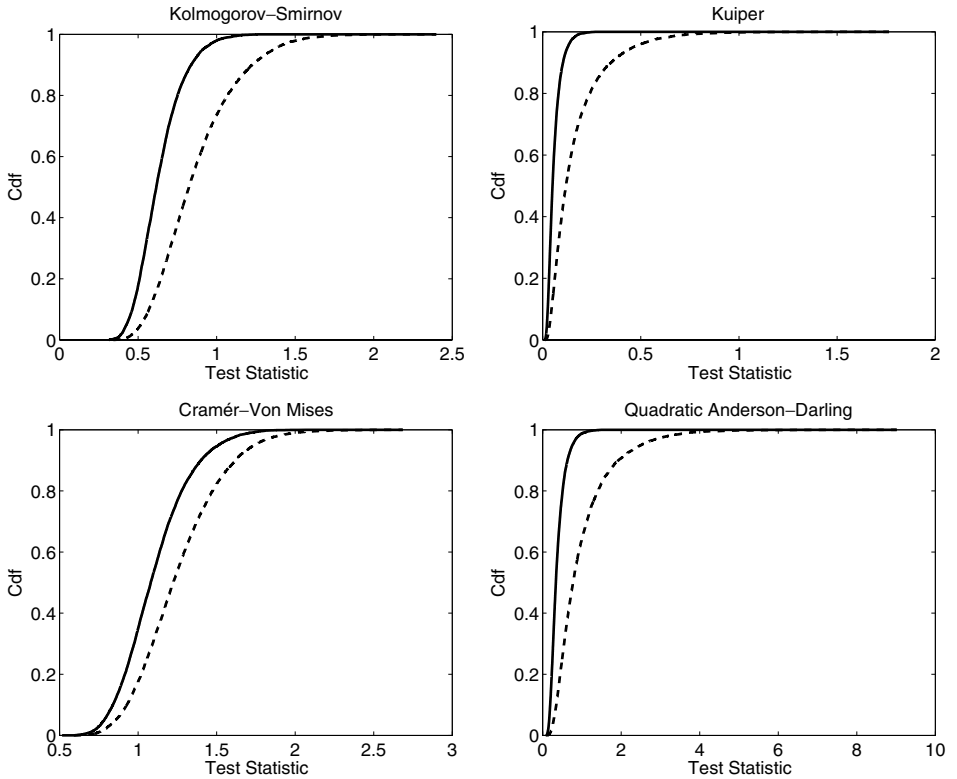

Fig. 2. The normal distribution. Estimates of cumulative distribution function (Cdf) for the four test statistics under study. Empirical sample size: $N = 5000$. Number of Monte-Carlo replications: $M = 10000$. Solid line: Procedure B (parameters are re-estimated each time on Monte-Carlo sample). Dashed line: Procedure A (always using empirical-sample estimates).

lead to wrong, overly-conservative approximations to the distributions of GoF test statistics based on the EDF. Furthermore, as our simple application shows, the impact of this possible mistake may turn out to be dramatic and does not vanish as the sample size increases.

Notice that similar issues have already been discussed in the relevant literature [9, 12, 14, 18]. More specifically, Ref. 19 shows that the mean of the Anderson–Darling statistic shifts towards the left when the parameters of the population distribution are unknown. Furthermore, Ref. 20 and Ref. 2 discuss the problem of approximating EDF test statistics from a rather theoretical perspective. Yet, despite the success of EDF-based GoF tests, no clear indications were given — to the best of our knowledge — about the practical correct Monte-Carlo procedure to be followed in order to approximate test statistic distributions in the case of unknown parameters. This paper aims at shedding more light on the risks resulting from a wrong specification of the Monte-Carlo simulation strategy, in all cases where critical-value tables are not already available. Given the lack of contributions

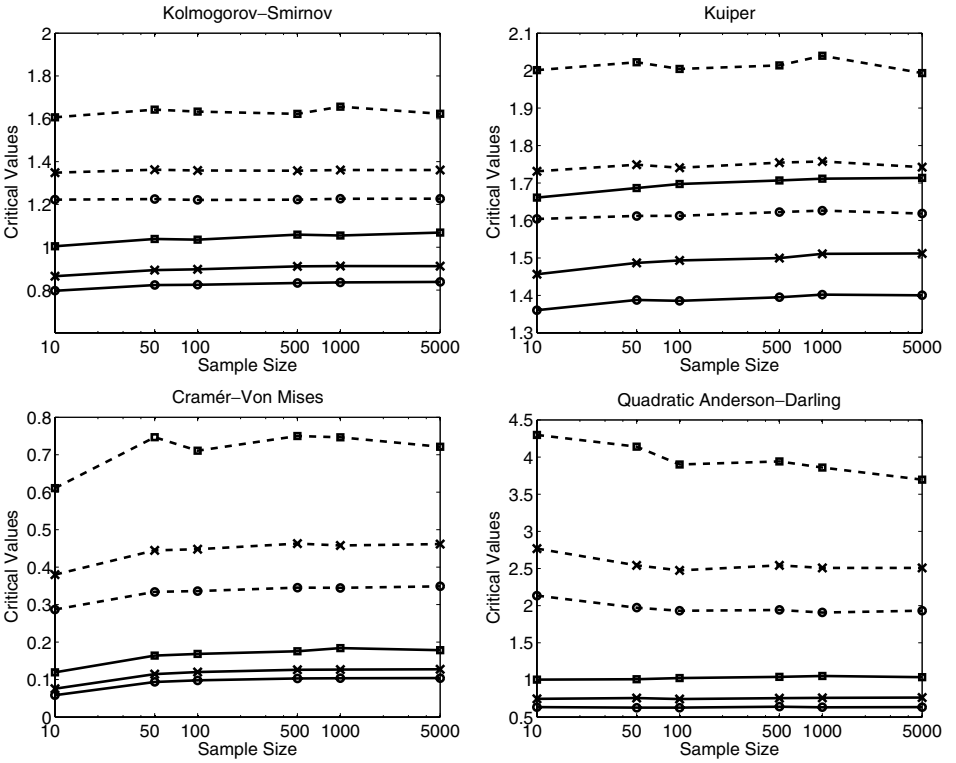

Fig. 3. The normal distribution. Critical values versus empirical sample size N (in log scale) for the four test statistics under study. Number of Monte-Carlo replications: $M = 10000$. Solid line: Procedure B (parameters are re-estimated each time on Monte-Carlo sample). Dashed line: Procedure A (always using empirical-sample estimates). Symbols stand for significance levels: $\circ = 0.10$, $\times = 0.05$, $\square = 0.01$.

addressing this topic, and the subtle nature of the choice between Procedure A and B, our feeling is that mistakes may be more likely than it may seem.

A final important remark is in order. In our discussion we deliberately focused on the case where parameters to be estimated are location and scale. In such an "ideal" situation, as we noted, the distributions of the four EDF-based test statistics that we have considered do not depend on the true unknown parameters. Therefore, in principle, to approximate their distributions one may generate, in Step B1, a sufficiently large number of independently-drawn N-sized samples from a $F(x; \underline{\theta}^*)$, where $\underline{\theta}^*$ is any given value of the unknown parameters, and not necessarily their empirical-sample estimates $\underline{\hat{\theta}}(\underline{x}_N)$. Since the distribution of the test is location- and scale-invariant, we just need to make sure to apply Step B2 (i.e. re-estimation of $\underline{\theta}$ using $\underline{z}_N^j$) in order to avoid the implicit assumption that parameters are known.

What happens if instead parameters are not location and scale but are still unknown? In such a case, test statistic distributions do depend on the true unknown parameter values [4,8]. Therefore, Step B1 may be considered as a first (good) guess

towards the approximation of test statistic distributions. In fact, when parameters are not location and scale, one cannot employ any given θ^* to generate Monte-Carlo samples. Since the "true" test statistic distribution depends on the "true" unknown parameter values, one would like to approximate it with a sufficiently similar (although not exactly equal) distribution, which can be easily obtained — provided that Procedure B is carried out — by employing the empirical sample estimates $\hat{\theta}(\underline{x}_N)$.

To shed more light on this situation, we consider here two additional distributions, namely Beta [11] and Generalized Extreme Value (GEV) [10], for which some shape parameters must be estimated. We repeat the exercise done in the previous sections by using samples drawn, respectively, from a Beta distribution having both (shape) parameters equal to 2, and from a GEV distribution having parameters equal to 0.2 (shape), 2 (scale) and 1 (location). The cumulative distributions of the four test statistics for both setups are reported in Figs. 4 (Beta distribution) and 5 (GEV distribution).

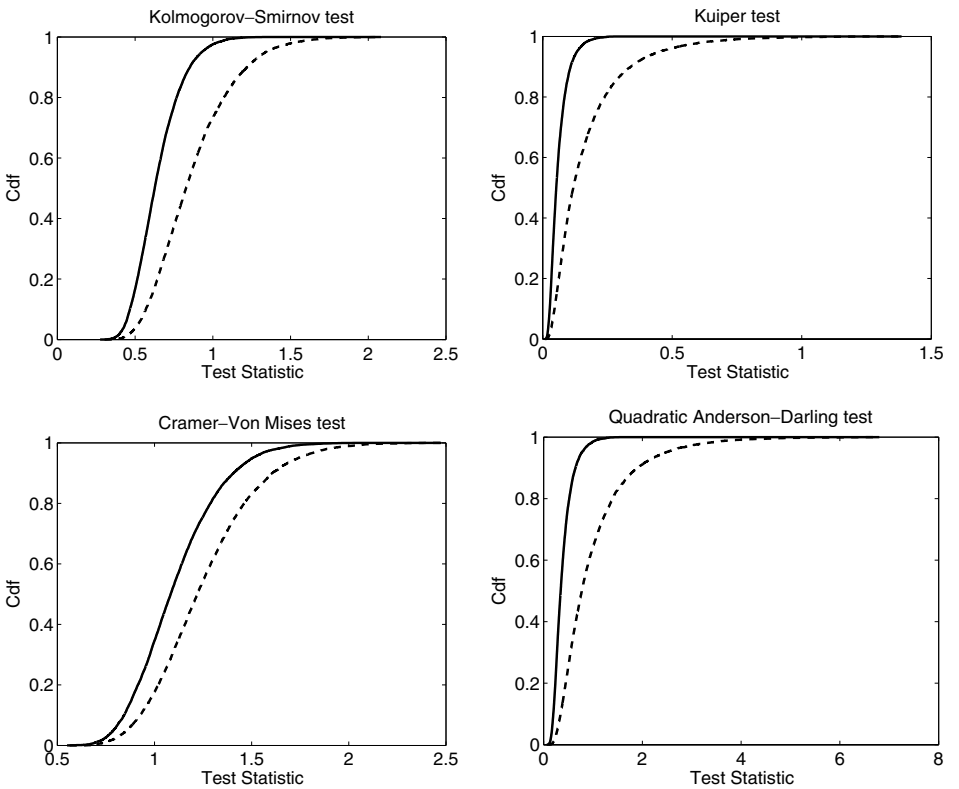

Fig. 4. The Beta distribution. Estimates of cumulative distribution function (Cdf) for the four test statistics under study. Empirical sample size: $N = 5000$. Number of Monte-Carlo replications: $M = 10000$. Solid line: Procedure B (parameters are re-estimated each time on Monte-Carlo sample). Dashed line: Procedure A (always using empirical-sample estimates).

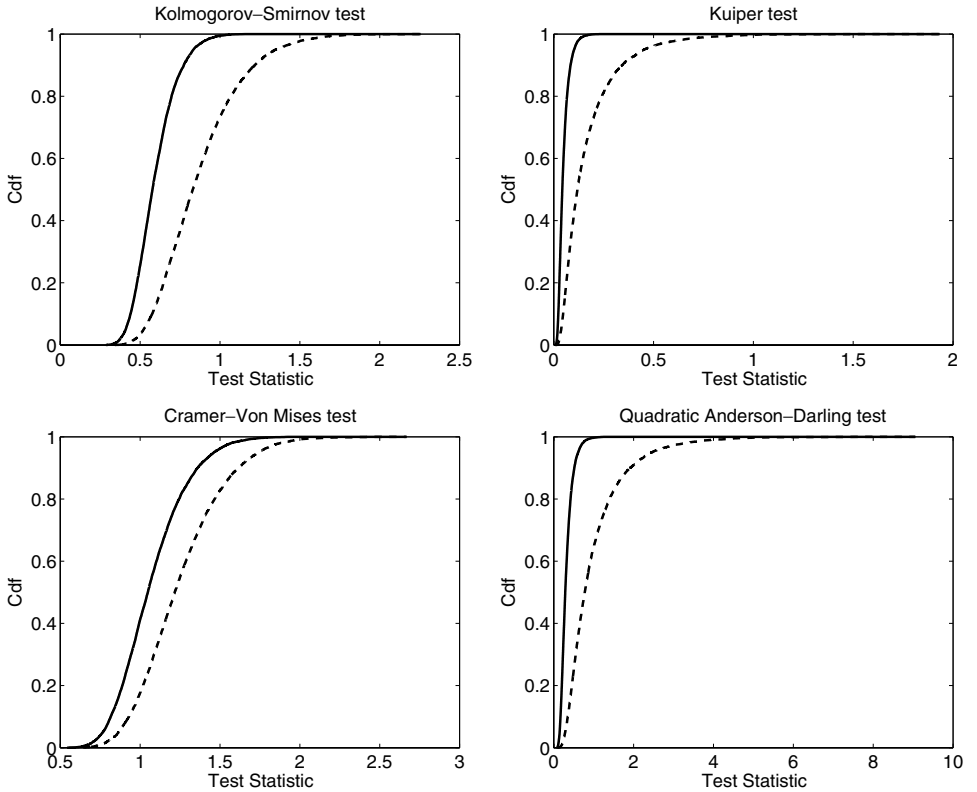

Fig. 5. The Generalized Extreme Value distribution. Estimates of cumulative distribution function (Cdf) for the four test statistics under study. Empirical sample size: $N = 5000$. Number of Monte-Carlo replications: $M = 10000$. Solid line: Procedure B (parameters are re-estimated each time on Monte-Carlo sample). Dashed line: Procedure A (always using empirical-sample estimates).

As the two figures show, Procedures A and B still produce different critical values for the four GoF tests considered. Therefore, the main insight of the paper is confirmed also when estimation does not concern location or scale, but focuses on the more general and frequent case of unknown shape parameters. In such situations, critical-value tables are not typically available, because they would depend on the empirical sample to be tested. Monte-Carlo simulations are therefore required and choosing the correct Procedure (B) instead of the wrong one (A) might become even more crucial than in the location/scale case.

Acknowledgments

This article enormously benefitted from comments and suggestions made by Richard Lockhart, who patiently went through the main points with his illuminating explanations, and by two anonymous referees. Thanks also to Michael Stephens and

Fabio Clementi for their very useful remarks and suggestions. Any remaining errors are the sole responsibility of the authors.

References

[1] Anderson, T. W. and Darling, D. A., A test for goodness-of-fit, *Journal of the American Statistical Association* **49** (1954) 765–769.

[2] Babu, G. J. and Rao, C. R., Goodness-of-fit tests when parameters are estimated, *Sankhya: The Indian Journal of Statistics* **66** (2004) 63–74.

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

[4] Darling, D. A., The Cramér-Smirnov test in the parametric case, *The Annals of Mathematical Statistics* **26** (1955) 1–20.

[5] David, F. N. and Johnson, N. L., The probability integral transformation when parameters are estimated from the sample, *Biometrika* **35** (1948) 182–192.

[6] Davis, C. S. and Stephens, M. A., Algorithm AS 248: Empirical distribution function goodness-of-fit tests, *Applied Statistics* **38** (1989) 535–543.

[7] Devroye, L., *Non-Uniform Random Variate Generation* (Springer-Verlag, New York, 1986).

[8] Durbin, J., Kolmogorov–Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings, *Biometrika* **62** (1975) 5–22.

[9] Dyer, A. R., Comparisons of tests for normality with a cautionary note, *Biometrika* **61** (1974) 185–189.

[10] Embrechts, P., Klüppelberg, C. and Mikosch, T., *Modelling Extremal Events for Insurance and Finance* (Springer-Verlag, Berlin, 1997).

[11] Evans, M., Hastings, N. and Peacock, B., *Statistical Distributions* (Wiley, New York, 2000).

[12] Green, J. R. and Hegazy, Y. A. S., Powerful modified-EDF goodness-of-fit tests, *Journal of the American Statistical Association* **71** (1976) 204–209.

[13] Kuiper, N. H., Tests concerning random points on a circle, *Proc. of the Koninklijke Nederlandse Akademie van Wetenschappen B* **63** (1962) 38–47.

[14] Lilliefors, H. W., On the Kolmogorov–Smirnov tests for normality with mean and variance unknown, *Journal of the American Statistical Association* **62** (1967) 399–402.

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

[16] Owen, D. B., *A Handbook of Statistical Tables* (Addison-Wesley, Reading, 1962).

[17] Pearson, E. S. and Stephens, M. A., The goodness-of-fit tests based on Wk and Uk, *Biometrika* **49** (1962) 397–402.

[18] Stephens, M. A., EDF statistics for goodness-of-fit and some comparisons, *Journal of the American Statistical Association* **69** (1974) 730–737.

[19] Stephens, M. A., Asymptotic results for goodness-of-fit statistics with unknown parameters, *The Annals of Statistics* **4** (1976) 357–369.

[20] Stute, W., Gonzalez-Manteiga, W. and Presedo-Quinmidil, M., Bootstrap based goodness-of-fit tests, *Metrika* **40** (1993) 243–256.

[21] Thode, H. C., Jr., *Testing for Normality* (Marcel Dekker, New York, 2002).