Paper about Something

Mathew Chandy

Department of Statistics, University of Connecticut

Abstract

The Kolmogorov–Smirnov (KS) test is widely employed to assess the goodness-of-fit of a hypothesized continuous distribution to a sample. Despite its popularity, the test is frequently misused in the literature and practice. While originally intended for independent, continuous data with precisely specified hypothesized distributions, it is erroneously applied to scenarios with dependent, discrete, or rounded data, with hypothesized distributions requiring estimated parameters. For example, it has been "discovered" multiple times that the test is too conservative when the hypothesized distribution has parameters that need to be estimated. We demonstrate misuses of the one-sample KS test in three scenarios through simulation studies: 1) the hypothesized distribution has unspecified parameters; 2) the data are serially dependent; and 3) a combination of the first two scenarios. For each scenario, we provide remedies for practical applications using appropriate bootstrap approaches. The whole demonstration can be used as hands-on education materials on both goodness-of-fit tests and bootstrap.

KEYWORDS: nonparametric bootstrap; parametric bootstrap; working dependence.

1 Introduction

The Kolmogorov-Smirnov (KS) test is one of the most popular goodness-of-fit tests for comparing a sample with a hypothesized parametric distribution. Let $X_1, ..., X_n$ be a random sample of size n from a continuous distribution. The null hypothesis H_0 is that X_i's follow

distribution F. Let $F_n(t) = \sum_{i=1}^{n} I(X_i \leq t)/n$ be the empirical cumulative distribution function of the sample, where $I(\cdot)$ is the indicator function. The KS test statistic is

$$D_n = \sqrt{n} \sup_x |F_n(x) - F(x)|. \tag{1}$$

The asymptotic distribution of D_n under H_0 is independent of the distribution F. As $n \to \infty$, D_n converges in distribution to the supremum of standard Brownian bridge (Kolmogorov, 1933). For large samples, the tests can be performed with a table (Smirnov, 1948). Critical values for small samples ($n \leq 35$) have also been given (Massey, 1951). The KS test is available in popular statistical software packages, such as function `ks.test()` in R package stats (R Core Team, 2022; Marsaglia et al., 2003).

The standard one-sample KS test applies to independent data with a continuous hypothesized distribution that is completely specified. In practice, however, it has often been applied without realizing that one or more of these assumptions do not hold. For example, Noether (1963) showed the conservativeness of the KS test when applied to discontinuous distributions. The null distribution of the KS statistic is no longer distribution-free and depends on the hypothesized distribution F. Computing the exact and asymptotic distribution of D_n is challenging. Fortunately, the null distribution of the KS statistic for discontinuous distributions has been efficiently addressed by Dimitrova et al. (2020) with a companion R package KSgeneral. Although a common misuse of the KS test, the issue with discontinuous data is not our focus.

When the hypothesized distribution F contains unspecified parameters, as is the case in most goodness-of-fit test settings, the standard KS test is not applicable. Steinskog et al. (2007) "discovers" the change in power when using fitted parameters and stresses caution in using the KS test in such ways. In fact, using fitted parameters in place of the true parameters in the KS test has been long known to yield extremely conservative results (e.g., Lilliefors, 1967). This problem can be solved by parametric bootstrap (Efron, 1985; Hall and

Wilson, 1991), where bootstrap samples of the test statistics are constructed from samples generated from the fitted hypothesized distribution. A nonparametric bootstrap solution is not trivial because a nonparametric bootstrap sample of the observed data has ties, which would not happen for continuous distributions. Babu and Rao (2004) derived the bias of the standard nonparametric bootstrap and showed how to correct it. They further noted that both parametric and nonparametric procedures lead to correct asymptotic levels.

The standard KS test does not apply to stationary yet dependent data either. The distribution of the KS statistic would have a higher variance for positively dependent data than that derived when the data are independent because of a smaller effective sample size. For example, for testing normality, Durilleul and Legendre (1992) demonstrate that a naive application of the KS statistic is too liberal for medium-to-high positive serial dependence, and that for negative dependence, the behavior is asymmetrical. For remedies, Weiss (1978) provides a procedure that is applicable specifically for data modeled by the second-order auto-regressive (AR) process where the AR parameters are known. Serial dependence affects the validity of the two-sample KS test too. In a recent paper on online diagnosis of performance variation in high-performance computing systems, for example, the authors provided no details about whether they accounted for serial dependence when applying the two-sample KS test (Tuncer et al., 2019). After comparing various strategies for dealing with serial dependence, Lanzante (2021) concludes that a test based on Monte-Carlo simulations performed the best. When additionally the hypothesized distribution contains unknown parameters, the standard KS test becomes even further inapplicable.

The contribution of this paper is a demonstration of misuses of the one-sample KS test in three scenarios and their remedies in practice. Contrary to the assumptions within the statistics community that anyone familiar with the relevant literature would not be "misusing" the KS test, there is still a considerable journey ahead in educating students and practitioners from diverse fields on the proper utilization of the KS test and similar tests. The scenarios that we consider are where: 1) the hypothesized distribution has unspecified

parameters; 2) the data are serially dependent; and 3) a combination of the first two scenarios. In each scenario, the misuse is performed and the impacts are shown. Then, a remedy is detailed and performed alongside the misuse to show its positive effects. Specifically, unspecified parameters are handled by parametric bootstrap; serial dependence is handled by introducing a working autoregressive moving average (ARMA) model that preserves the serial dependence for a wide range of dependence structures. In order to set up the demonstrations, simulated data are used throughout. The remedies are also performed on various families of distributions. An R implementation of the proposed methods will be available after the paper is published.

The rest of the paper is organized as follows. Section 2 investigates the scenario where the hypothesized distribution has unspecified parameters. Both parametric and nonparametric bootstrap are available to fix the issue. Section 3 investigates the scenario where the data of the empirical distribution are serially dependent. A bootstrap procedure employing a working ARMA model to account for dependence is proposed as a working solution. Section 4 explores the case where a combination of the first two scenarios occurs. An adjusted bootstrap procedure is proposed as a working solution in this case. Section 5 concludes with a discussion.

2 Unspecified Parameters

The null hypothesis of a goodness-of-fit test is often a composite hypothesis instead of a single hypothesis. That is, the hypothesized distribution is a family of distributions with unspecified parameters instead of a specific member in this family. Let F_θ be a family of distributions indexed by parameter vector θ. The null hypothesis is

$$H_0 : \text{the random sample } X_1, \ldots, X_n \text{ comes from a distribution } F_\theta \text{ for some } \theta.$$

Since θ is unknown, one would naturally estimate θ by $\hat{\theta}_n$ from, for example, maximum likelihood or methods of moments. The KS statistic would then be computed as

$$D_n = \sqrt{n} \sup_x |F_n(x) - F_{\hat{\theta}_n}(x)|. \tag{2}$$

An overly large D_n still indicates evidence to reject H_0. To get the p-value of the observed D_n, however, note that the null distribution is not the same as that in the standard case (1). If the same null distribution were used, one would be testing a different H_0 that the random sample comes from the specific member distribution $F_{\hat{\theta}_n}$ instead of the family F_θ.

The consequence of using the wrong null distribution for the KS statistic D_n in (2) can be illustrated through a simple simulation study. A random sample $X_1, \ldots, X_n$ was generated from a normal distribution with both mean and variance parameters set equal to 8 ($N(8, 8)$), and sample size $n = 200$. The maximum likelihood estimates of the mean and variance were used as $\hat{\theta}_n$. The p-value of the test statistic D_n in Equation (2) was obtained by naively calling the `ks.test()` function in R with the fitted normal distribution as the hypothesized distribution. That is, the hypothesized distribution was $N(\bar{X}, s^2)$ where $\bar{X}$ is the sample mean and s^2 is the sample variance (with n in the denominator). This experiment was repeated 1000 times and the probability-probability plot (PP-plot) of the 1000 p-values are displayed in the top left plot of Figure 1. If the test were valid, the p-values would be uniformly distributed in the unit interval and the points in the PP-plot would fall along the diagonal line. Clearly, the distribution of the 1000 p-values is very different from what one would expect to see from 1000 draws from the standard uniform distribution.

We similarly considered random samples $X_1, \ldots, X_n$ generated from a gamma distribution with shape and scale parameters 8 and 1, respectively ($\Gamma(8, 1)$), and sample size $n = 200$. These parameter values were selected so that the mean and variance of the normal and gamma distributions in the simulations are the same. Likewise, the p-value of the test statistic D_n in (2) was obtained by naively calling the `ks.test()` function in R with the fit-

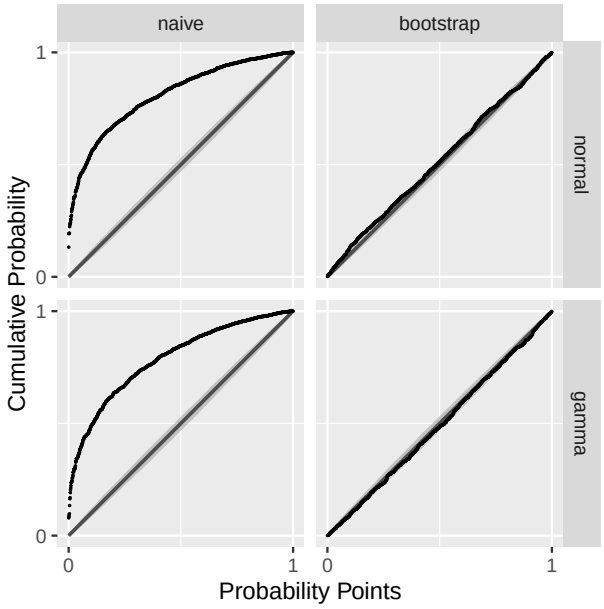

Figure 1: PP-plots for the p-values obtained from KS tests with unspecified parameters with sample size $n = 200$ based on 1000 replicates. The data generating models were $N(8, 8)$ and $\Gamma(8, 1)$.

ted gamma distribution as the hypothesized distribution using the shape and scale maximum likelihood estimates. The PP-plot of the p-values from 1000 replications of this experiment are displayed in the bottom left plot of Figure 1. The p-values are again non-uniformly distributed due to the use of the estimated parameters when specifying the null distribution.

To fix the problem, parametric bootstrap can be used to approximate the null distribution of the test statistic D_n:

1. Draw a random sample $X_1^*, ..., X_n^*$ from the fitted distribution $F_{\hat{\theta}_n}$
2. Fit F_θ to $X_1^*, ..., X_n^*$ and obtain estimate $\hat{\theta}_n^*$ of θ.
3. Obtain the empirical distribution function F_n^* of $X_1^*, \ldots, X_n^*$.
4. Calculate bootstrap KS statistic

$$D_n^* = \sqrt{n} \sup_x |F_n^*(x) - F_{\hat{\theta}_n^*}(x)|.$$

5. Repeat the previous steps a large number B times and use the empirical distribution of D_n^* to approximate the null distribution of the observed statistic.

The p-value of D_n is approximated by the proportion of the D_n^* statistics that are greater than or equal to D_n.

In the same simulation study, p-values were obtained for the 1000 replicates from the parametric bootstrap procedures. The top and bottom right plots of Figure 1 display the PP-plots of the 1000 p-values from the normal and gamma data generation scenarios, respectively. This time, all points fall along the diagonal lines implying that the p-values are coming from standard uniform distributions.

Nonparametric bootstrap can also be used to approximate the null distribution of the test statistic in (2) except that a bias correction is needed (Babu and Rao, 2004); see details of the procedure in the Appendix. The results from nonparametric bootstrap are similar to those from the parametric bootstrap and, hence, are omitted.

3 Serially Dependent Data

Here we consider the situation where the observed data $X_1, \ldots, X_n$ are not independent but serially dependent, and the hypothesized distribution F contains no unknown parameters. That is,

$$H_0 : X_i\text{'s have marginal distribution } F.$$

The KS statistic D_n in Equation (1) is still a good measure of deviation from the null hypothesis. Nonetheless, the null distribution changes when there is serial dependence. When the serial dependence is positive, the effective sample size is less than n. The test statistic would be stochastically greater than that obtained from independent data, resulting in a test that is too liberal (i.e., the null hypothesis is rejected more often than it should). The reasoning is similar when the serial dependence is negative, in which case, the test is conservative.

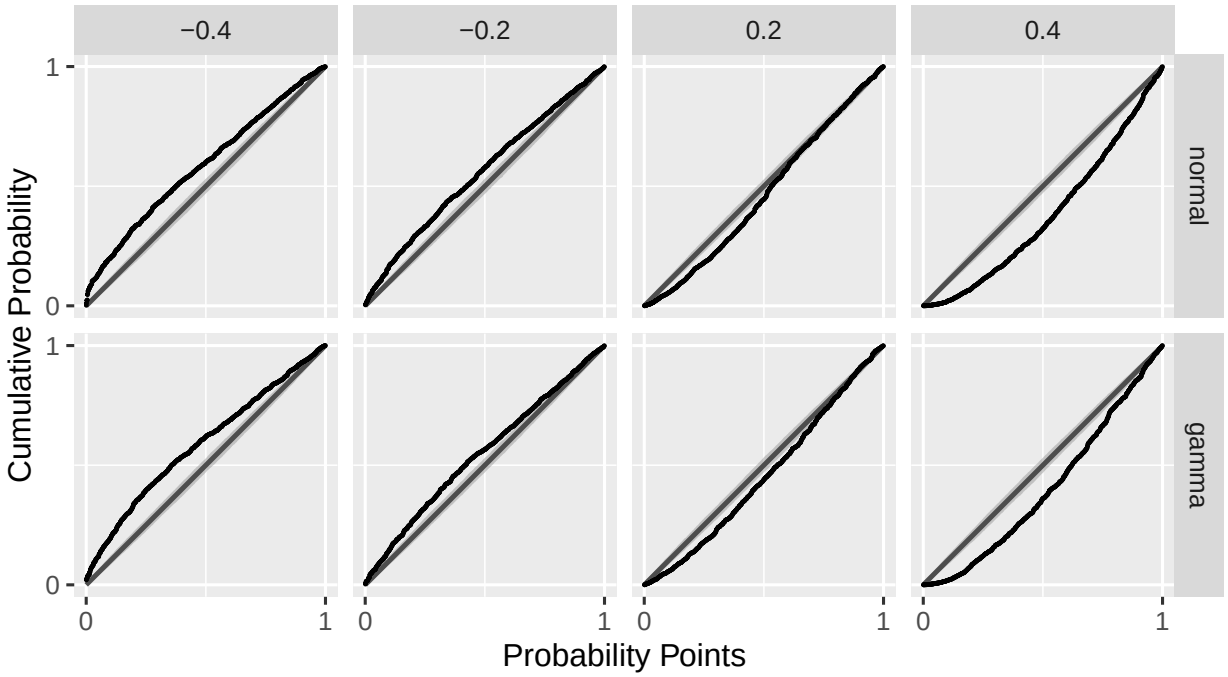

Figure 2: PP-plots for the p-values obtained from KS test with serial dependence ignored and sample size $n = 200$ based on 1000 replicates. The data have marginal distributions $N(8, 8)$ and $\Gamma(8, 1)$, with an AR(1) dependence structure on the standard normal scale. The AR coefficients are $\psi \in \{-0.4, -0.2, 0.2, 0.4\}$.

The invalidity of the standard statistic for serially dependent data can be illustrated by a simple simulation study. Consider a first-order autoregressive or AR(1) model with marginal normal (or gamma) distribution. For each AR coefficient $\psi \in \{-0.4, -0.2, 0.2, 0.4\}$, we generated 1000 series of length 200. For each series, we applied the KS statistic in (1) to test H_0 that the X_i's follow $N(8, 8)$ (or $\Gamma(8, 1)$) marginally. The PP-plots of the p-values from 1000 replicates for each ψ value are displayed in Figure 2. As ψ diverges from 0, the p-values look less likely to be generated from a standard uniform distributions. Positive ψ values led to p-values with a distribution function higher than the standard uniform distribution function as there were more smaller p-values than there should be, and hence, liberal tests. Negative ψ values led to p-values with a distribution function lower than the standard uniform distribution as there were more larger p-values than there should be, and hence, conservative tests.

A complete remedy for KS tests with serially dependent data is challenging. The null distribution depends on the structure of the serial dependence, which can be arbitrary in practice. A parametric bootstrap procedure would need to specify a model for the dependence, whereas the primary interest is to test the marginal distribution of the stationary series. A block nonparametric bootstrap for stationary data (Kunsch, 1989) is tempting, but the counterpart of a bias correction as in Babu and Rao (2004) is not available yet. Is there any approach that does not require full specification of the dependence of the stationary process, but still gives reasonably satisfactory correction to the size of the KS test in certain applications?

We propose a semiparametric bootstrap procedure that assumes a working serial dependence structure which does not need to be correctly specified. The working serial dependence is introduced through a working ARMA process with the hope to approximate the true serial dependence as allowed by the working model. Essentially, this working model covers the most commonly seen dependence structure characterized by a normal copula (Hofert et al., 2018), which separates the dependence structure of a multivariate distribution from its marginal distributions. The parameters of the working normal copula are estimated from fitting an ARMA model to the observed data transformed to the standard normal scale.

In particular, let $Z_i = \Phi^{-1}\{F(X_i)\}$, $i = 1, \ldots, n$, where Φ is the distribution function of the standard normal. Then we fit an ARMA(p, q) model to $Z_1, \ldots, Z_n$ with AR order p and MA order q selected by the Akaike Information Criterion (AIC). This can be done with, for example, function `auto.arima()` from R package `forecast` (Hyndman and Khandakar, 2008). Since it is known that the mean of Z_i's is zero, it is necessary to set the intercept of the ARMA model as zero in the fitting process. This restriction turned out to be critical from our investigation; having the intercept estimated does not lead to desired p-value distributions in the following bootstrap process. The selected ARMA(p, q) model with fitted parameters will be used as the working model to generate bootstrap samples with serial dependence mimicking that in the observed data. The unconditional variance σ^2 of the ARMA model

with unit innovation variance can be obtained with function `tacvfARMA()` from R package `ltsa` (McLeod et al., 2007).

The semiparametric bootstrap procedure is as follows.

1. Generate $Z_1^*, \ldots, Z_n^*$ from the fitted ARMA(p, q) process with innovation variance $1/\sigma^2$ such that the Z_i's are marginally standard normal variables.

2. Form a bootstrap sample $X_i^* = F^{-1}[\Phi(Z_i^*)]$, $i = 1, \ldots, n$, where F^{-1} is the quantile function of F.

3. Obtain the empirical distribution function F_n^* of the bootstrap sample $X_1^*, \ldots, X_n^*$.

4. Calculate bootstrap KS statistic

$$D_n^* = \sup_x |F_n^*(x) - F(x)|.$$

5. Repeat the previous steps a large number B times and use the empirical distribution of the B test statistics to approximate the null distribution of the observed statistic.

The p-value of D_n is, again, approximated by the proportion of the D_n^* statistics that are greater than or equal to D_n.

This process is semiparametric because the dependence structure is specified by a normal copula determined by an ARMA process with normally distributed errors. The closer the approximation is to the truth, the better performance of the size of the KS test. The normal copula of the working ARMA process provides a wide class of dependence structures. If the true dependence indeed has an ARMA structure, this method is exact. When the true dependence is not covered by the ARMA model, the working model may still give a reasonable approximation that can be useful for practical purposes.

The semiparametric bootstrap procedure was evaluated in a simulation study. We considered two marginal distributions: $N(8, 8)$ and $\Gamma(8, 1)$. Three dependence structures were used to link variables generated from each marginal distribution; on the standard normal

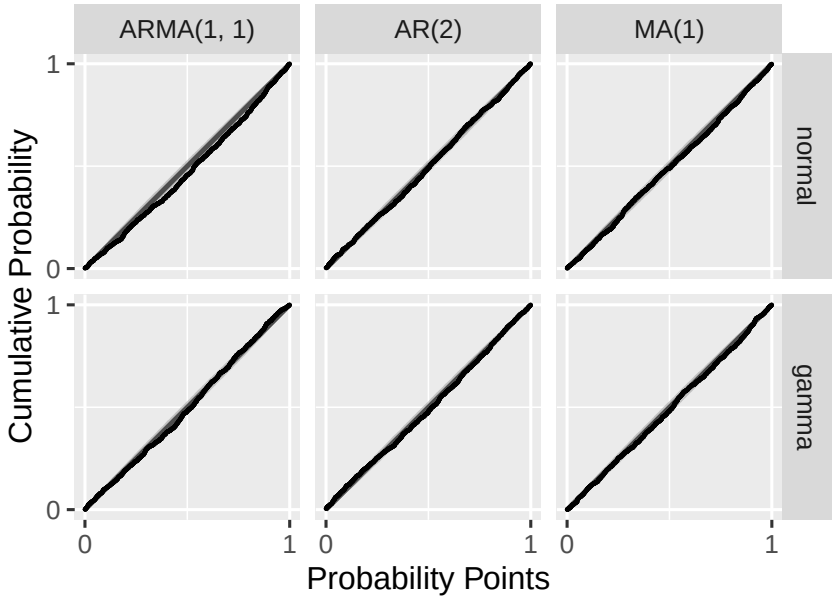

Figure 3: PP-plots for the p-values obtained from KS test with serial dependence accounted for through a working ARMA model and $n = 200$ based on 1000 replicates. The data have marginal distributions $N(8, 8)$ and $\Gamma(8, 1)$. The dependence structure on the standard normal scale was characterized by ARMA(1, 1) model, AR(2) model, and MA(1) model.

scale, they are ARMA(1, 1) with AR coefficient 0.5 and MA coefficient 0.3; AR(2) with coefficients $(0.6, 0.2)$; and MA(1) with coefficient 0.8. For each dependence structure and marginal distribution combination, we generated 1000 series of length 200. For each series, we applied the semiparametric bootstrap procedure with serial dependence accounted for through a working ARMA model to test H_0 that the X_i's marginally follow their true marginal distribution ($N(8, 8)$ or $\Gamma(8, 1)$). Figure 3 displays the PP-plots of the 1000 p-values for each dependence structure and marginal distribution combination. All plots within the figure show the points falling along the diagonal lines, implying that the p-values are coming from standard uniform distributions and that the semiparametric bootstrap procedure is an effective remedy in all cases considered.

4 Unspecified Parameters and Serially Dependent Data

When the observed sample $X_1, \ldots, X_n$ is no longer a random sample but a serially dependent stationary series, we may still be interested in testing whether marginally each X_i follows hypothesized distribution with unknown parameters. That is, the null hypothesis is

$$H_0 : X_i\text{'s have marginal distribution } F_\theta \text{ for some } \theta.$$

The testing statistic D_n in Equation (2) still measures the deviation from H_0, but its null distribution depends on the serial dependence. The bootstrap procedure in the last section can be adapted to handle this situation.

Specifically, let $\hat{\theta}_n$ be the marginally fitted parameters, and $Z_i = \Phi^{-1}\{F_{\hat{\theta}_n}(X_i)\}$, $i = 1, \ldots, n$ be a transformation of the observed data onto the standard normal scale using the fitted parameters $\hat{\theta}_n$. We then fit an ARMA model with orders selected by the AIC using function `auto.arima()` from R package `forecast`. This ARMA model will be used as the working model to introduce serial dependence in the bootstrap samples. Again, we obtain the unconditional variance σ^2 of the ARMA process with unit innovation variance. Our semiparametric bootstrap procedure is as follows.

1. Generate $Z_1^*, \ldots, Z_n^*$ from the working ARMA(p, q) model with innovation variance $1/\sigma^2$ such that the Z_i's are marginally standard normal variables.

2. Form a bootstrap sample $X_i^* = F_{\hat{\theta}_n}^{-1}[\Phi(Z_i^*)]$, $i = 1, \ldots, n$, where $F_{\hat{\theta}_n}^{-1}$ is the quantile function of $F_{\hat{\theta}_n}$.

3. Fit F_θ to $X_1^*, \ldots, X_n^*$ to obtain estimate $\hat{\theta}_n^*$ of θ.

4. Obtain the empirical distribution function F_n^* of the bootstrap sample $X_1^*, \ldots, X_n^*$.

Table 1: Power of rejecting the null hypothesis of normal and gamma distribution at significance level 0.05 when the true distribution was $\Gamma(8,1)$ and $N(8,8)$, respectively.

		H_0 : normal		H_0: gamma	
Dependence	Method	$n = 100$	$n = 200$	$n = 100$	$n = 200$
ARMA	naive	0.404	0.656	0.515	0.779
	copula	0.330	0.556	0.442	0.722
AR	naive	0.411	0.655	0.455	0.705
	copula	0.261	0.448	0.313	0.544
MA	naive	0.374	0.728	0.514	0.817
	copula	0.343	0.693	0.487	0.797

5. Calculate bootstrap KS statistic

$$D_n^* = \sup_x |F_n^*(x) - F_{\hat{\theta}_n^*}(x)|.$$

6. Repeat the previous steps a large number B times and use the empirical distribution of the B test statistics to approximate the null distribution of the observed statistic.

The p-value of D_n is approximated by the proportion of the D_n^* statistics that are greater than or equal to D_n. Compared to the algorithm in the last section, here we use $F_{\hat{\theta}_n^*}$ in place of the known F.

Figure 4 shows the PP-plots resulting p-values of the procedure performed on data generated from same simulation settings as in Section 3. For comparison, we also displayed the PP-plot of a naive bootstrap procedure that adjusts for the unspecified parameters as in Section 2 but not for the serial dependence. The super naive approach that does not adjust for fitted parameters nor serial dependence would have double sin, so the results are not reported. The p-values from naively fitting parameters but not adjusting for dependence clearly deviates from the standard uniform distribution except in the MA(1) case, where the autocorrelation is nonzero with a short memory of only lag 1. Similar to Section 3, the working normal copula approach gives p-values with the desired standard uniform distribution regardless of the true dependence.

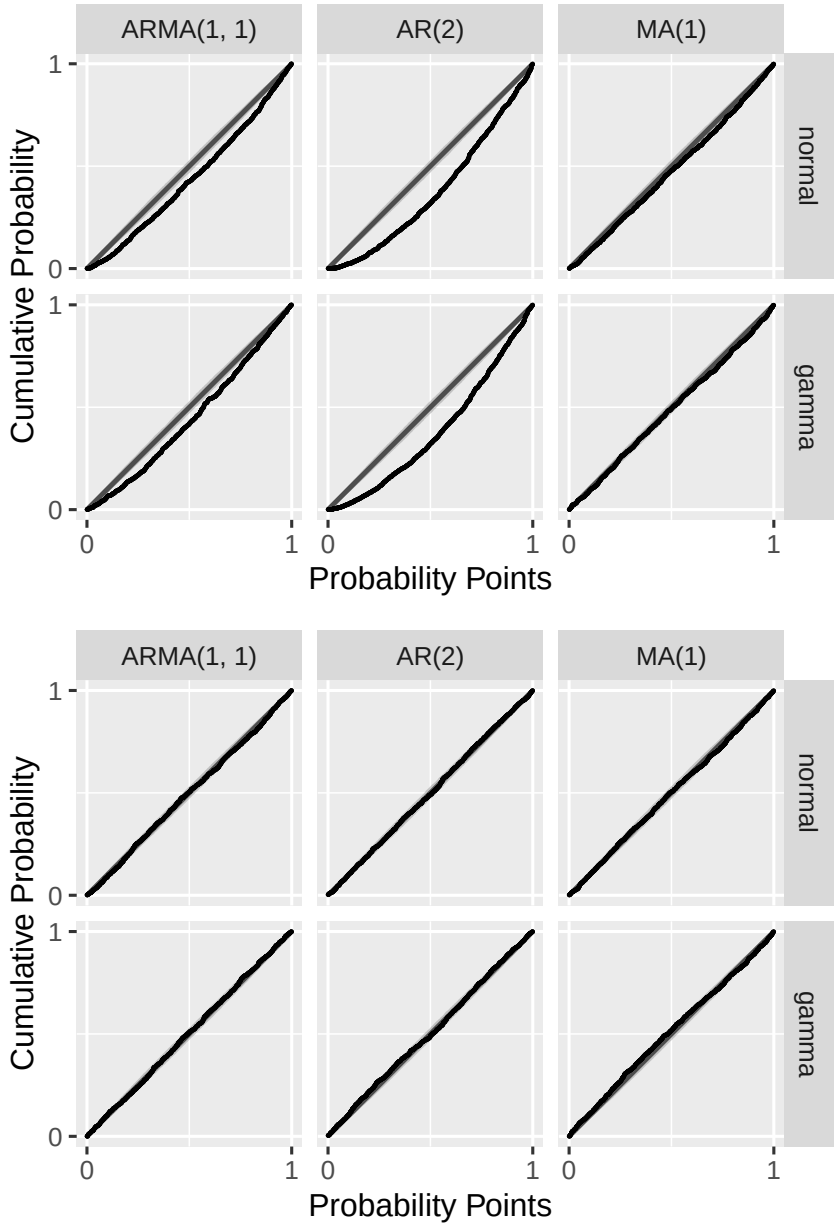

Figure 4: PP-plots for the p-values obtained from KS test without (top) and with (bottom) serial dependence accounted for and unspecified parameters; sample size $n = 200$ based on 1000 replicates. The data have marginal distributions $N(8, 8)$ and $\Gamma(8, 1)$. The dependence structure on the standard normal scale was characterized by ARMA(1, 1) model, AR(2) model, and MA(1) model.

Finally, we report a small study of the power of the one-sample KS test with unspecified parameters and serially dependent data. Using the three dependence models characterized by ARMA(1, 1), AR(2), and MA(1) in the simulation settings in the last section, we generated

serially dependent data with marginal distributions $N(8,8)$ left truncated by zero and $\Gamma(8,1)$. Then, for data from $\Gamma(8,1)$, we tested H_0 that the data has marginally a normal distribution; for data from $N(8,8)$ left truncated by zero, we tested H_0 that the data has marginally a gamma distribution. In both cases, no parameters were specified. Two sample sizes were considered, $n \in \{100, 200\}$. Table 4 summarizes the power of rejecting the null hypothesis at significance level 0.05 from 1000 replicates. The tests based on a working copula have substantial power in rejecting the null hypothesis and the power increases as sample size increases. The power from the naive tests where the serial dependence is ignored, although the unspecified parameters are accounted for (i.e., procedure from Section 2), appear to have higher power than the proposed tests, but these are not reliable since they do not hold their sizes as demonstrated in Figure 4.

5 Concluding Remarks

The three kinds of misuses of the KS test for one-sample goodness-of-fit that we demonstrated here are commonly seen in practice. With the aid of computing tools such as R, they can be easily demonstrated in classroom teaching. Practitioners need to be aware of the assumptions of the standard one-sample KS test that the hypothesized distribution is completely specified and that the data are independent. When these assumptions are not met, naive applications of the test are no longer accurate and remedies must be performed. In the case of fitted parameters, parametric and nonparametric bootstrap can restore the size of the test. The nonparametric version has long been developed too but is lesser known (Babu and Rao, 2004). In the case of serially dependent data, our semiparametric bootstrap procedure with a working dependence structure controlled by an automatically selected ARMA model restores the size of the test for a wide range of dependence structures. When both assumptions are violated, i.e., where the data has serial dependence and the parameters must be fitted, a combination of the two individual remedies shows good results. The misuses of the KS test

also apply to other goodness-of-fit tests based on the empirical distribution function, such as the Anderson–Darling test or the Cramér-—von Mise test, which are known to be superior to KS in certain respects (Stephens, 2017). Similar bootstrap remedies can be developed for these tests.

The semiparametric bootstrap procedure with a working ARMA dependence structure is not a complete solution and has limitations. When the true serial dependence can be well approximated by an ARMA model, the model selection process could be replaced with autoregressive sieve, i.e., a sequence of autoregressive models the order of which increases simultaneously with the sample size although at a suitably slower rate (e.g., Psaradakis and Vávra, 2017, 2020). When the true serial dependence structure cannot be well approximated by an ARMA model with normally distributed error terms, the remedy may not restore the size of the test. The working dependence structure is essentially the normal copula uniquely determined by the ARMA model with normal errors. Some copulas are quite different from the normal copula. For example, extreme value copulas have tail dependence, which is not provided by normal copulas. Using a normal copula to approximate an extreme value copula may not capture the true serial dependence structure that is needed to restore the size of the test. A block version of the nonparametric bootstrap (Babu and Rao, 2004), which avoids specification of a working dependence structure, would be more desirable.

Statisticians living in an ivory tower tend to doubt that anyone would be misusing the KS test as we illustrated given that such situations are very well-known in the statistics literature. It is our experience, however, that much effort is needed in propagating many knowns in the statistics community to the real wild world through innovative and impactful education. Such effort may be more valuable than developing more advanced methods that few people would use in practice.

Data Availability Statement

An R implementation of the proposed methods will be made publicly available upon publication.

A Appendix: Nonparametric Bootstrap for KS Test with Fitted Parameters

Using the same notations in the text, the nonparametric bootstrap procedure is summarized as follows (Babu and Rao, 2004).

1. Draw a random sample $X_1^*, ..., X_n^*$ from the empirical distribution F_n with replacement.
2. Fit F_θ to $X_1^*, ..., X_n^*$ and obtain estimate $\hat{\theta}_n^*$ of θ.
3. Obtain the empirical distribution function F_n^* of $X_1^*, \ldots, X_n^*$.
4. Calculate bootstrap KS statistic

$$D_n^* = \sup_x |\sqrt{n}\left(F_n^*(x) - F_{\hat{\theta}_n^*}(x)\right) - B_n(x)|.$$

   where $B_n(x) = \sqrt{n}(F_n(x) - F_{\hat{\theta}_n}(x))$ is the known bias term.

5. Repeat the previous steps a large number B times and use the empirical distribution of D_n^* to approximate the null distribution of the observed statistic.

The p-value of D_n is, again, approximated by the proportion of the D_n^* statistics that are greater than or equal to D_n. Note that the step 1 and step 4 are different from the parametric bootstrap version.

References

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

Dimitrova, D. S., Kaishev, V. K., and Tan, S. (2020), "Computing the Kolmogorov-Smirnov Distribution When the Underlying CDF is Purely Discrete, Mixed, or Continuous," *Journal of Statistical Software*, 95(10), 1–42.

Durilleul, P. and Legendre, P. (1992), "Lack of robustness in two tests of normality against autocorrelation in sample data," *Journal of Statistical Computation and Simulation*, 42, 79–91.

Efron, B. (1985), "Bootstrap confidence intervals for a class of parametric problems," *Biometrika*, 72, 45–58.

Hall, P. and Wilson, S. R. (1991), "Two guidelines for bootstrap hypothesis testing," *Biometrics*, 47, 757–762.

Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2018), *Elements of Copula Modeling with R*, New York: Springer.

Hyndman, R. J. and Khandakar, Y. (2008), "Automatic Time Series Forecasting: The forecast Package for R," *Journal of Statistical Software*, 26(3), 1–22.

Kolmogorov, A. (1933), "Sulla determinazione empirica di una lgge di distribuzione," *Giornale dell'Istituto Italiano degli Attua*, 4, 83–91.

Kunsch, H. R. (1989), "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics*, 17, 1217–1241.

Lanzante, J. R. (2021), "Testing For Differences Between Two Distributions in The Presence Of Serial Correlation Using the Kolmogorov–Smirnov and Kuiper's Tests," *International Journal of Climatology*, 41, 6314—6323.

Lilliefors, H. W. (1967), "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown," *Journal of the American Statistical Association*, 62, 399–402.

Marsaglia, G., Tsang, W. W., and Wang, J. (2003), "Evaluating Kolmogorov's Distribution," *Journal of Statistical Software*, 8(18), 1–4.

Massey, F. J. (1951), "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, 46, 68–78.

McLeod, A. I., Yu, H., and Krougly, Z. (2007), "Algorithms for Linear Time Series Analysis: With R Package," *Journal of Statistical Software*, 23(5), 1–26.

Noether, G. E. (1963), "Note on the Kolmogorov Statistic in the Discrete Case," *Metrika*, 7, 115–116.

Psaradakis, Z. and Vávra, M. (2017), "A Distance Test of Normality for a Wide Class of Stationary Processes," *Econometrics and Statistics*, 2, 50–60.

— (2020), "Normality Tests for Dependent Data: Large-Sample and Bootstrap Approaches," *Communications in Statistics—Simulation and Computation*, 49, 283–304.

R Core Team (2022), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

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

Steinskog, D. J., Tjøstheim, D. B., and Kvamstø, N. G. (2007), "A Cautionary Note on the Use of the Kolmogorov-Smirnov Test for Normality," *Monthly Weather Review*, 135, 1151–1157.

Stephens, M. A. (2017), "Tests Based on EDF Statistics," in *Goodness-of-fit-techniques*, eds. D'Agostino, R. B. and Stephens, M. A., Routledge, pp. 97–194.

Tuncer, O., Ates, E., Zhang, Y., Turk, A., Brandt, J., Leung, V. J., Egele, M., and Coskun, A. K. (2019), "Online Diagnosis of Performance Variation in HPC Systems Using Machine Learning," *IEEE Transactions on Parallel and Distributed Systems*, 30, 883–896.

Weiss, M. S. (1978), "Modification of the Kolmogorov–Smirnov Statistic for Use with Correlated Data," *Journal of the American Statistical Association*, 73, 872–875.