Spring 4-29-2022

On Misuses of the Kolmogorov–Smirnov Test for One-Sample Goodness-of-Fit

Anthony Zeimbekakis
anthony.zeimbekakis@uconn.edu

On Misuses of the Kolmogorov–Smirnov Test for One-Sample Goodness-of-Fit

Anthony Zeimbekakis
anthony.zeimbekakis@uconn.edu
Department of Statistics, University of Connecticut

Thesis Advisor: Jun Yan
Honors Advisor: Elizabeth Schifano

Abstract

The Kolmogorov–Smirnov (KS) test is one of the most popular goodness-of-fit tests for comparing a sample with a hypothesized parametric distribution. Nevertheless, it has often been misused. The standard one-sample KS test applies to independent, continuous data with a hypothesized distribution that is completely specified. It is not uncommon, however, to see in the literature that it was applied to dependent, discrete, or rounded data, with hypothesized distributions containing estimated parameters. For example, it has been "discovered" multiple times that the test is too conservative when the parameters are 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.

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 from some continuous distribution. The null hypothesis H_0 is that the population distribution is F. Let $F_n(\cdot)$ be the empirical cumulative distribution. The KS statistic is

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

The distribution of D under H_0 turns out to be independent of the distribution F and a table of critical values for D has been constructed (Massey, 1951) for various sample sizes n and significance levels α. If the value of D exceeds the test's corresponding critical value the null hypothesis is rejected. The KS test is available in popular statistical software packages, such as function `ks.test` in R (R Core Team, 2020; 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.

Noether (1963) provides a brief demonstration of the conservative character of the KS test when using discrete functions. There are remedies that allow the KS test to be used in discrete cases. Conover (1972) describes a method for finding the exact critical level for discrete cases and derives its power. Gleser (1985) provides methods for adapting existing algorithms to provide results when $F(x)$ is discontinuous. The discrete KS test is available in various statistical software packages. The package dgof implements the methods from Conover (1972) for one-sided tests and the methods from Gleser (1985) for exact two-sided p-values (Arnold and Emerson, 2011). The KSgeneral package provides a fast and accurate method for when $F(x)$ is arbitrary, discontinuous, or continuous that uses an alternative fast Fourier transform based method (Dimitrova et al., 2020). This issue with discrete or rounded is not the focus of this paper, though is a common misuse that should be noted.

The standard KS test is not applicable when the hypothesized distribution contains fitted parameters. Steinskog et al. (2007) "discovers" the change in power when using estimated parameters and stresses caution in using the KS test in such ways. Lilliefors (1967) shows for the normal distribution that using the standard table when values of the mean and standard deviation are estimated obtains extremely conservative results. This is supported by Capasso et al. (2009), who concludes that failing to re-estimate the parameters may lead to wrong, overly-conservative approximations to the distributions of goodness-of-fit test statistics based on the empirical distribution function. Capasso et al. (2009) also notes that the impact of this mistake may turn out to be dramatic and does not vanish as the sample size increases. Remedies are provided by Babu and Rao (2004) and Genest and Rémillard (2008) in the form of bootstrap. Babu and Rao (2004) details the bootstrap procedure for goodness-of-fit tests and notes that both parametric and non-parametric procedures lead to correct asymptotic levels, however there is a correction required for the non-parametric case. Genest and Rémillard (2008) provides validity for using parametric bootstrap with various goodness-of-fit tests.

In the case of serially dependent data, Durilleul and Legendre (1992) demonstrates that the KS statistic is too liberal for medium-to-high positive autocorrelation values. Durilleul and Legendre (1992) also shows that for negative autocorrelation values, the behavior is asymmetrical with respect to positive values. For remedies, Weiss (1978) provides a procedure that is applicable specifically for data modeled by the second-order auto-regressive (AR) process where the parameters are known. Lanzante (2021) tests various strategies for dealing with temporal dependence and concludes that a test based on Monte-Carlo simulations performed the best. We propose a parametric bootstrap procedure involving copulas to account for dependence.

The contribution of this paper is a demonstration of misuses of the one-sample KS test in three scenarios and their remedies in practice. The scenarios 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. In order to set up the demonstrations, simulated data is used throughout. The remedies are also performed on various families of distributions.

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 is serially dependent. A bootstrap procedure employing copulas is a working solution. Section 4 explores the case where a combination of the first two scenarios occurs. The copula procedure can be adjusted for fitted parameters to be available as a fix. Section 5 concludes with a discussion.

2 Unspecified Parameters

The null distribution of the KS statistic changes when the hypothesized distribution contains fitted parameters. In this scenario, the null hypothesis is H_0: the random sample $X_1, \ldots, X_n$ comes from a continuous distribution F_θ with unspecified parameter vector θ. Let $\hat{\theta}_n$ be an estimator of θ, which could be, for example, maximum likelihood estimator of moment estimator. The test statistic is

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

Since $F_{\hat{\theta}_n}$ is not the same as the true data generating F_θ, the null distribution of D obtained in existing implementations in software packages which assumes completely known F_θ no longer applies (Steinskog et al., 2007).

To demonstrate the consequences of this problem, a simulation is performed. A random sample $X_1, \ldots, X_n$ is generated from the standard normal distribution with sample size $n = 100$. The p-values of 1000 replicate tests are displayed in the Naive plot of Figure 1. Each

KS test was performed using fitted parameters, i.e. the hypothesized distribution is $N(\bar{X}, s^2)$ where $\bar{X}$ is the sample mean and s^2 is the sample variance. Since the data is generated from a standard normal distribution with seemingly all assumptions met, a uniform distribution of $U(0,1)$ is expected for the p-values. However this is under the assumption that the KS test assumptions hold, which no longer do due to using fitted parameters. Therefore, there is notable deviation from the uniform distribution.

To fix the problem, parametric bootstrap can be used to approximate the null distribution of the testing statistic.

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

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

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

Nonparametric bootstrap can also be used to approximate the null distribution of the testing statistic. The procedure is similar, however the resampling is performed with the empirical distribution instead of the fitted parametric distribution and there is a correction for bias that is required (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 the sample and obtain estimated $\hat{\theta}_n^*$
3. Obtain the empirical distribution function of the random sample F_n^*

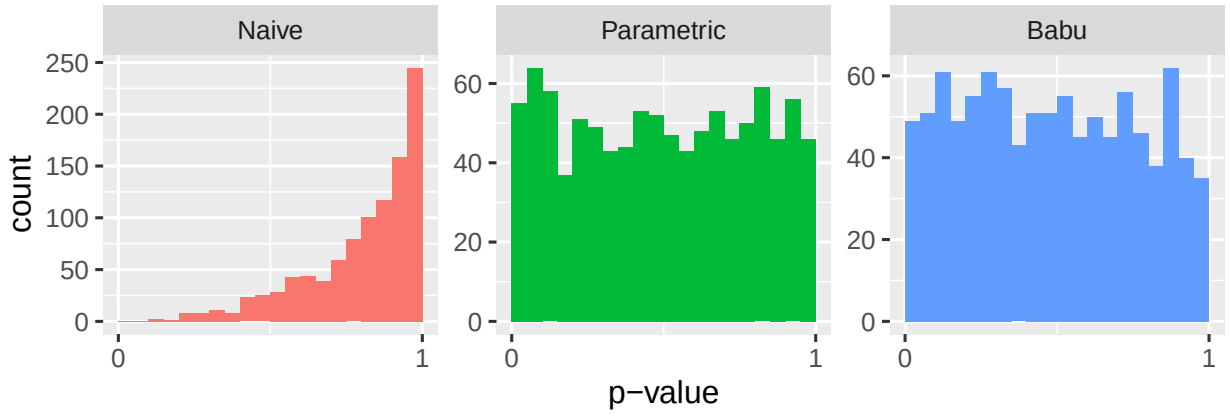

Figure 1: The Naive plot (left) is the histogram of p-values from the standard KS test with fitted parameters. The Parametric plot (middle) is the histogram of p-values from implementing parametric bootstrap. The Babu plot (right) is the histogram of p-values from implementing nonparametric bootstrap with a correction for bias. In each case, 1000 replicate tests were performed on the standard normal distribution with sample size $n = 100$. $B = 1000$ bootstrap samples are obtained for each test.

4. Calculate bootstrap KS statistic

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

where $B_n(x) = \sqrt{n}(F_n(x) - F_{\hat{\theta}_n}(x))$ is the known bias term (Babu and Rao, 2004)

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

The p-value can be calculated by counting the number of bootstrap KS statistics greater than or equal to the observed KS statistic, and then dividing by the number of bootstrap samples. Figure 1 displays the results of from our simulations. We would expect the distribution of p-values to be uniform in the case where the KS test holds its size. It is clear from the figure that both parametric and nonparametric bootstrap processes correct for the problem of fitted parameters. The plots for the bootstrapped p-values appear to be $U(0, 1)$, unlike the naive p-values.

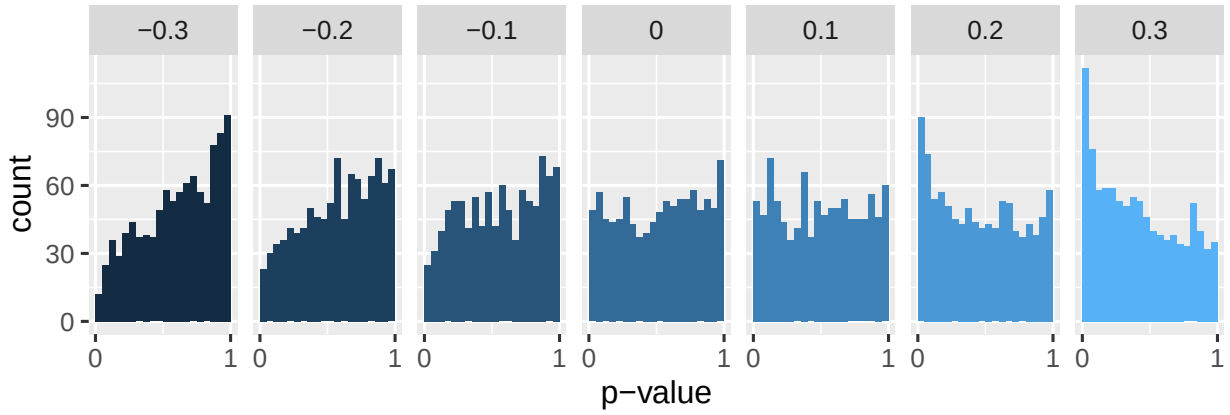

Figure 2: Each histogram is a plot of the p-values of standard KS tests performed on dependent data simulated from an AR(1) process of the standard normal distribution. The titles represent the various levels of the AR coefficient ψ tested. The sample size is $n = 100$ and 1000 replicate tests were performed for each AR coefficient.

3 Serially Dependent Data

The KS test also displays issues in the case of dependent data. As mentioned, an assumption of the test is that the data is independent. Unfortunately, real data is often temporally or spatially dependent and the results of a goodness-of-fit test would be valuable. When the KS test is performed on dependent data it performs poorly. In this situation, the null hypothesis is H_0: the random sample $X_1, \ldots, X_n$ come from a continuous distribution F where F has completely specified parameters. The test statistic is the same as Equation (1). However, since the data is dependent, the null hypothesis is wrong and must be corrected. This is demonstrated with a simulation in Figure 2. Data is generated from a first-order autoregressive model (AR(1)) with a standard normal distribution. P-values are gathered from 1000 replicate tests for different levels of the AR coefficient ψ, varying from $(-3, 3)$. In the presence of serially dependent data, the distribution of p-values no longer follows $U(0, 1)$ as would be expected of a valid test. The results echo those of Durilleul and Legendre (1992) that the KS statistic is too liberal for positive autocorrelation values, and that the behavior is asymmetrical for negative values.

In order to correct this, we can employ a parametric bootstrap procedure which assumes

a working serial dependence structure through copulas. A copula is a multivariate distribution with standard uniform marginal distributions, which completely characterizes the dependence structure of a multivariate distribution (Hofert et al., 2020, 2018). For simplicity, we assume a normal copula with an AR(1) structure to characterize the serial dependence of the observations. The AR(1) parameter r of the normal copula is set to match the be the sample serial Spearman's rho of the observed sample. The procedure is as follows.

1. Generate $Z_1, \ldots, Z_n$ from an AR(1) process with autocorrelation coefficient r such that the Z_i's are $N(0,1)$ variables.

2. Form a bootstrap sample $X_i^* = F^{-1}[\Phi(Z_i)]$, where Φ is the distribution function of $N(0,1)$, $i = 1, \ldots, n$, whose first-order sample Spearman's rho matches that of the observed data.

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

4. Calculate bootstrap KS statistic

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

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.

Throughout the simulation we assume a working AR(1) dependence structure regardless of the true dependence. If the true dependence is indeed a normal copula with an AR(1) structure, this method is exact. When the true dependence is not AR(1), it may still give a reasonable approximation that can be useful for practical purposes. This is demonstrated with different dependence structures in Figure 4.

Figure 3 displays the results of the copula remedy for dependent data. The data is generated from the standard normal distribution with an AR(1) dependence structure where

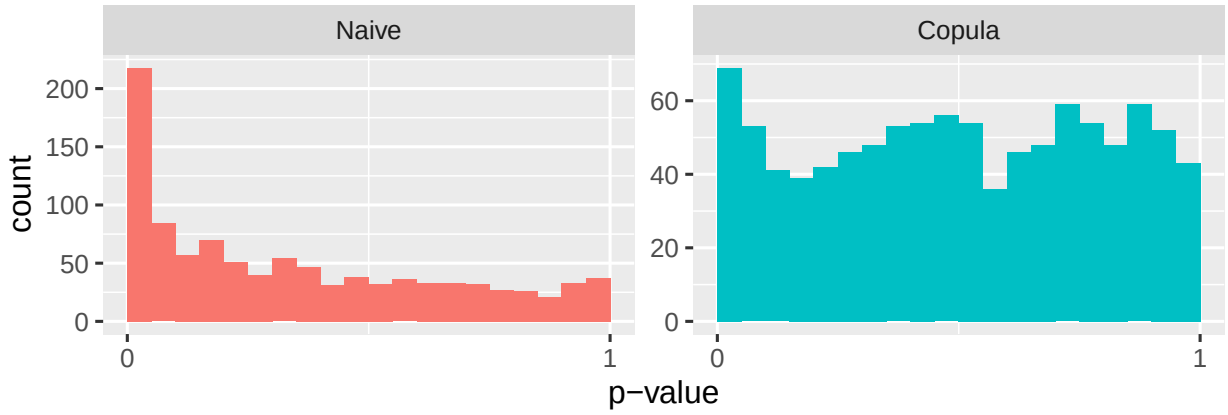

Figure 3: The Naive plot (left) is the histogram of p-values from the standard KS test. The Copula plot (right) is the histogram of p-values from implementing parametric bootstrap with copulas for dependence. In each case, 1000 replicate tests were performed with the data generated from an AR(1) process of the standard normal distribution with $\psi = .5$. The sample size is $n = 100$ and $B = 1000$ bootstrap samples are obtained for each test.

$\psi = .5$. The copula used is to model dependence is the normal copula. The distribution of p-values is again expected to be $U(0, 1)$ for a valid test. The Naive plot is clearly not uniform and reinforces the results shown in Figure 2. The Copula plot, which implements the aforementioned procedure to correct for dependence in the data, appears to be uniform and restores the size of the KS test.

The procedure detailed in this section can also provide a reasonable approximation in cases where the true dependence structure is not far from AR(1). Figure 4 shows the distribution of p-values for dependence structures MA(1), ARMA(1, 1), and AR(2). As expected, naively performing the KS test without correcting for dependence provides plots of p-values that deviate from the uniform distribution. In the case of MA(1) and ARMA(1, 1), the true dependence structure is close enough to our assumption of AR(1) that the bootstrap procedure provides a reasonable approximation. However, the AR(2) copula plot shows the limitation of this technique as no AR(1) process can approximate an AR(2) process unless the second-order coefficient is close to zero.

To further demonstrate the effectiveness of the procedure, we perform similar simulations on other families of distributions. Figure 5 displays the results of tests using dependent data

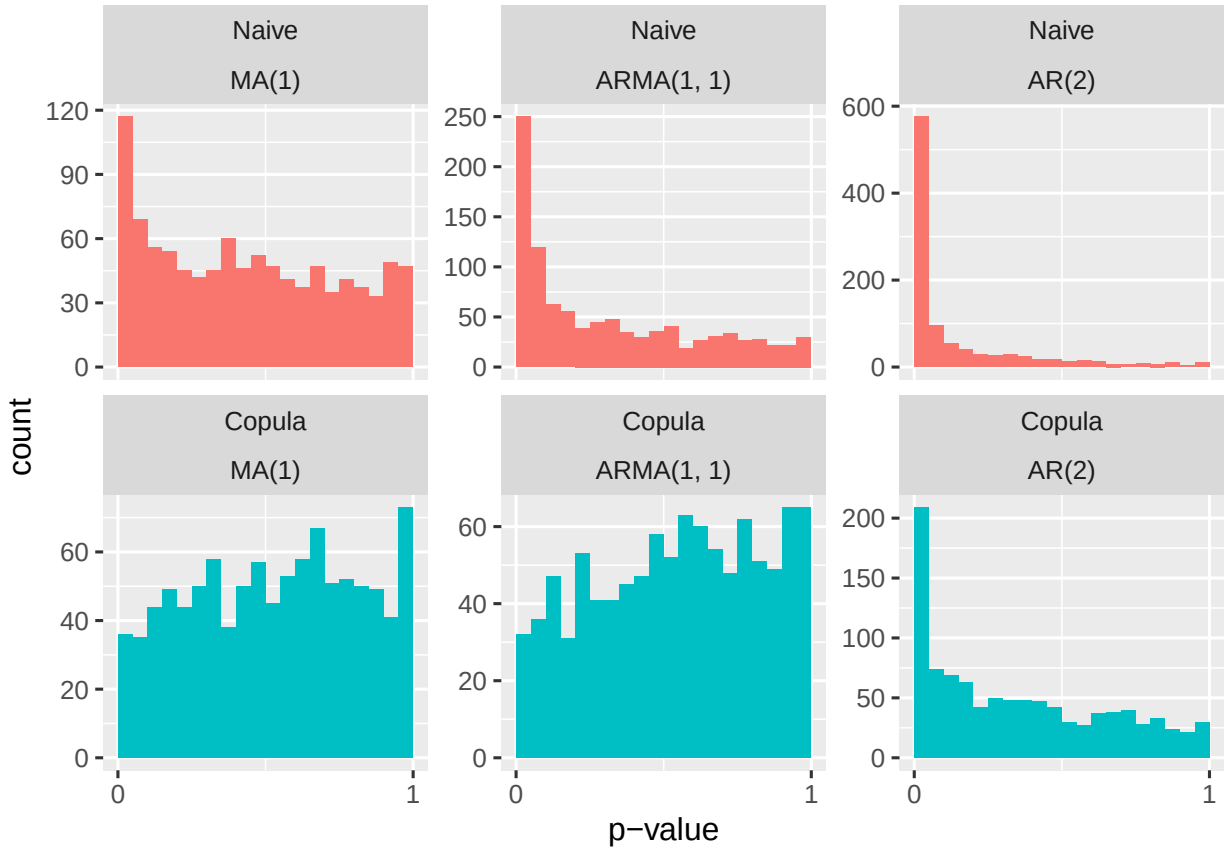

Figure 4: The Naive plots are the histograms of p-values from the standard KS test. The Copula plots are the histograms of p-values from parametric bootstrap with copulas for dependence. The data is generated from an MA(1) process (left) with $\theta = .5$, ARMA(1, 1) process (middle) with $\psi = .5$ and $\theta = .3$, and AR(2) process (right) with $\psi = (.5, .3)$ of the standard normal distribution. In each case, 1000 replicate tests were performed with sample size $n = 100$ and $B = 1000$ bootstrap samples.

generated from the gamma and the generalized extreme value distributions. The procedure corrects the distribution of p-values for both distributions and is shown to be applicable irregardless of the family of distribution tested.

4 Unspecified Parameters and Serially Dependent Data

The procedure demonstrated in Section 3 works when the data is dependent and the hypothesized distribution is completely specified. However, this is not practical. In practice we do not know the parameters of the hypothesized distribution. Therefore, it is valuable to

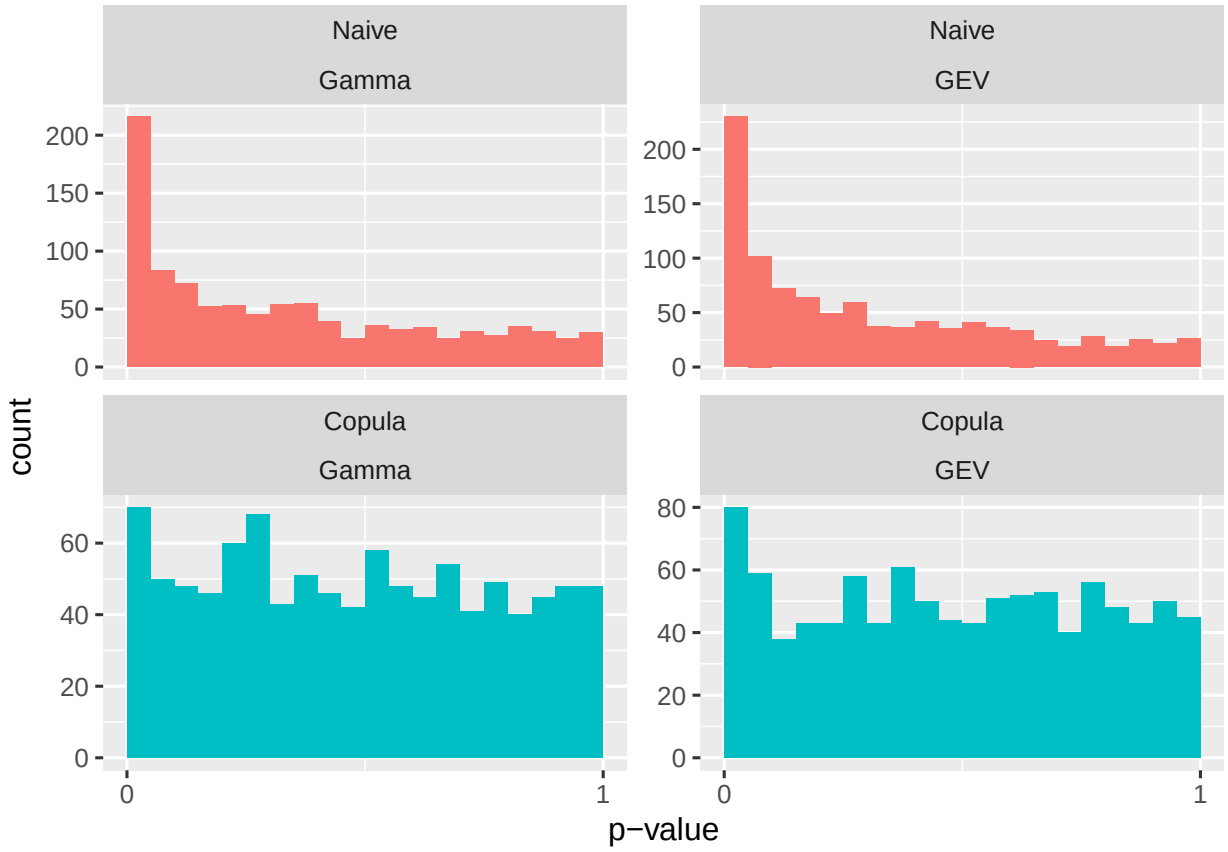

Figure 5: The Naive plots are the histograms of p-values from the standard KS test. The Copula plots are the histograms of p-values from parametric bootstrap with copulas for dependence. The data is generated from $Gamma(3, 1)$ (left) and $GEV(0, .2, 1)$ (right) with a correlation coefficient of $r = 0.5$. In each case, 1000 replicate tests were performed with sample size $n = 100$ and $B = 1000$ bootstrap samples.

have a procedure that corrects for both fitted parameters and serially dependent data. The remedy using copulas can be modified to be effective in the case where both assumptions must be violated. The null hypothesis is equivalent to Section 2 and the test statistic is equal to Equation (2). The bootstrap procedure in Section 2 is no longer valid because the serial dependence is not accounted for. Let r be the AR(1) coefficient of the normal copula that matches the sample first-order Spearman's rho of the observed data as obtained by the `iRho` function in the copula package (Hofert et al., 2020). We propose the following bootstrap procedure to assess the significance of the observed KS statistic.

1. Generate $Z_1, \ldots, Z_n$ from an AR(1) process with autocorrelation coefficient r such that

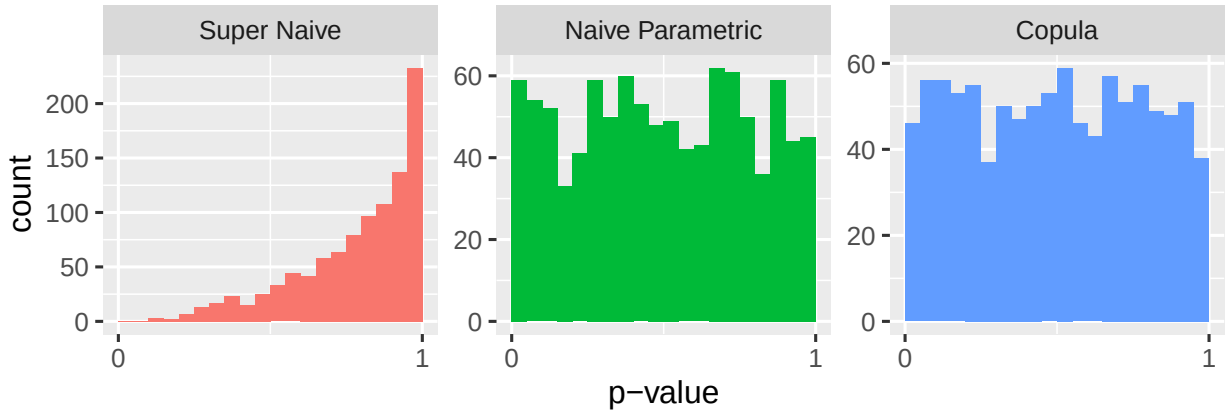

Figure 6: The Super Naive plot (left) is the histogram of p-values from the standard KS test with fitted parameters. The Naive Parametric plot (middle) is the histogram of p-values from implementing parametric bootstrap. The Copula plot is the histogram of p-values from implementing parametric bootstrap with copulas for dependence and correcting for fitted parameters. In each case, 1000 replicate tests were performed with the data generated from an AR(1) process of the standard normal distribution with $\psi = .5$. The sample size is $n = 100$ and $B = 1000$ bootstrap samples are obtained for each test.

the Z_i's are $N(0, 1)$ variables.

2. Form a bootstrap sample $X_i^* = F_{\hat{\theta}_n}^{-1}[\Phi(Z_i)]$, $i = 1, \ldots, n$, whose first-order sample Spearman's rho matches that of the observed data.

3. Fit F_θ to the sample $X_1^*, \ldots, X_n^*$ and obtain estimator $\hat{\theta}_n^*$

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

5. Calculate bootstrap KS statistic

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

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.

Figure 6 shows the results of our simulation on data simulated from an AR(1) process. As should be expected, naively fitting parameters while providing no adjustment for dependent

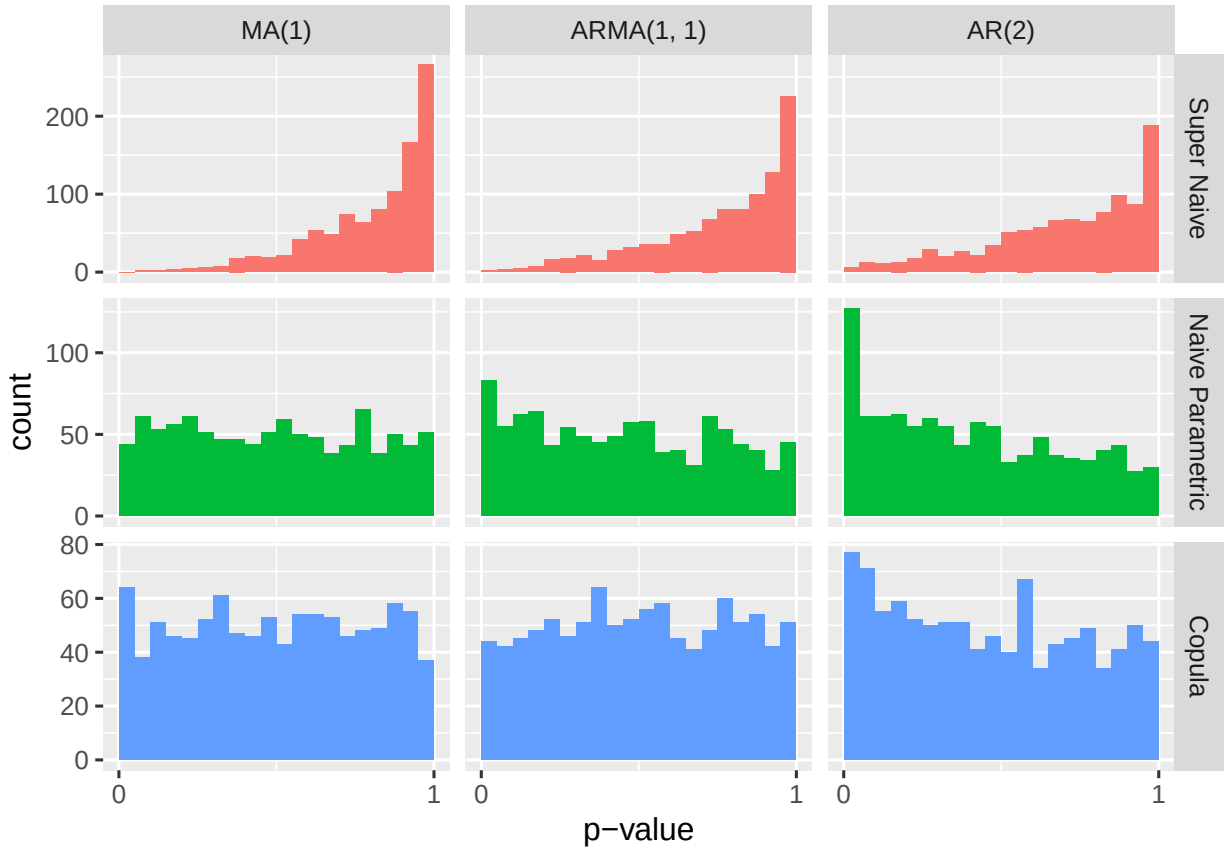

Figure 7: The Super Naive plots (left) are the histograms of p-values from the standard KS test with fitted parameters. The Naive Parametric plots (middle) are the histograms of p-values from implementing parametric bootstrap. The Copula plots (bottom) are the histograms of p-values from implementing parametric bootstrap with copulas for dependence and correcting for fitted parameters. The data is generated from an MA(1) process (left) with $\theta = .5$, ARMA(1, 1) process (middle) with $\psi = .5$ and $\theta = .3$, and AR(2) process (right) with $\psi = (.5, .3)$ of the standard normal distribution. In each case, 1000 replicate tests were performed with sample size $n = 100$ and $B = 1000$ bootstrap samples.

data invalidates the KS test. As well, only using the parametric bootstrap remedy for fitted parameters from Section 2 seems to provide weaker results than the copula remedy presented above. The results show that copula remedy generates uniform p-values and restores the size of the KS test.

Similar to Section 3, the copula fix is not a complete solution. Regardless of the true dependence in the data we assume an AR(1) dependence structure by taking the lag-1 sample auto-spearman rho. However, we can show that as long as the AR(1) assumption is a close

approximation, the correction still provides a reasonable approximation that can be useful for practical purposes. Figure 7 shows the results of the procedure performed on data generated with dependence structures of MA(1), ARMA(1, 1), and AR(2). Naively fitting parameters and not adjusting for dependence clearly deviates from a uniform distribution of p-values. Parametric bootstrap provides some correction but does not account for dependence, so therefore the results of the copula remedy are more accurate and favorable. In the case of MA(1) and ARMA(1, 1), the true dependence appears close enough to our assumption of AR(1) that the results are reasonable. AR(2) however appears to be just far enough from our assumption, showing a limitation of the procedure.

It is also possible to apply our procedure to other families of distributions. The data in Figure 8 is generated from the gamma and generalized extreme value distributions. The results are favorable and show that the procedure outlined as a fix for fitted parameters and spatially dependent can be adapted for various families of distributions.

5 Conclusion

The KS test has base assumptions that the hypothesized distribution is completely specified and the data is independent. When these assumptions are violated, the test is no longer accurate and remedies must be performed. In the case of fitted parameters, parametric and non-parametric bootstrap can restore the size of the test. A bias correction is required if the non-parametric form is used (Babu and Rao, 2004). In the case of dependent data, a procedure using bootstrap with copulas to model dependency shows positive results. When both assumptions are violated, that is the case where the data has dependence and parameters must be fitted, an adjusted copula procedure also shows positive results. The tests appear effective for a variety of families of distributions. The copula remedy is not a complete solution and has limitations. Regardless of the true dependence, we assume an AR(1) dependence structure. Therefore, if the AR(1) dependence structure is a close approximation

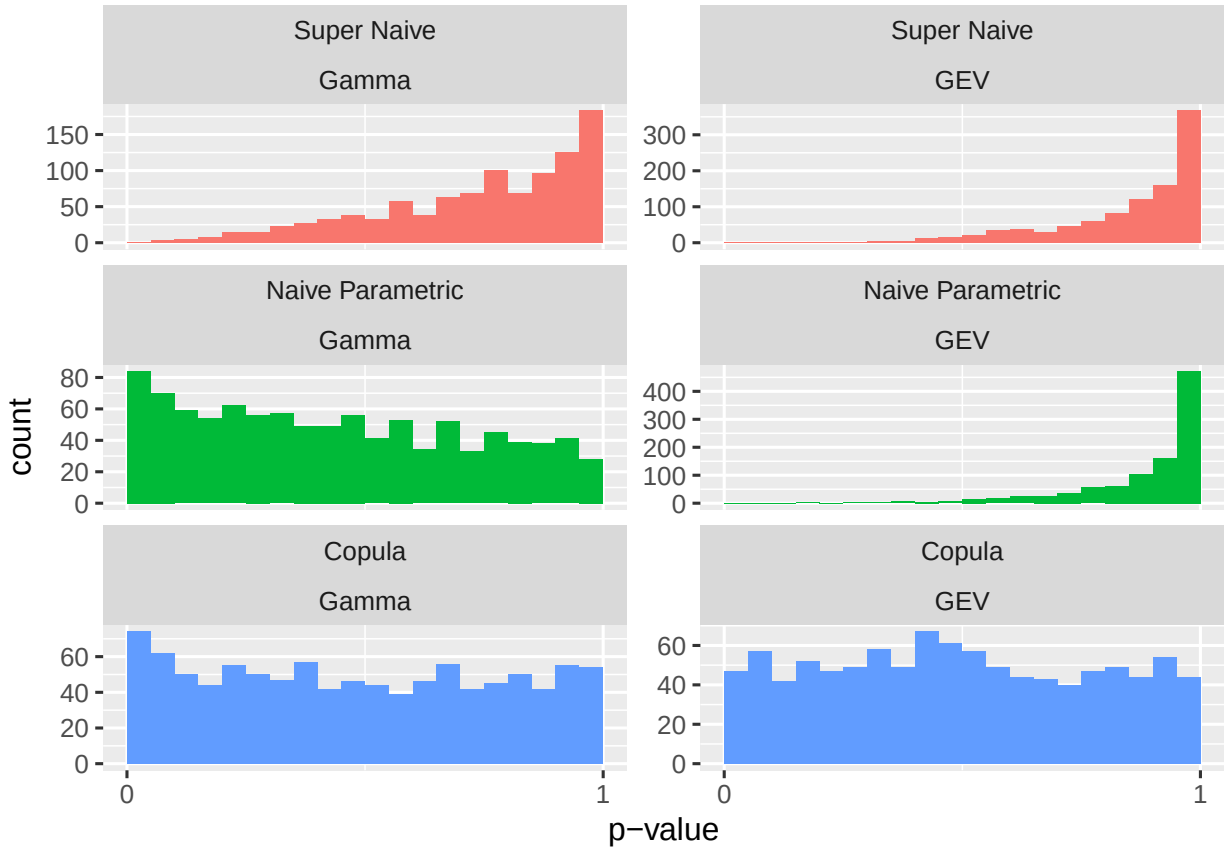

Figure 8: The Super Naive plots (left) are the histograms of p-values from the standard KS test with fitted parameters. The Naive Parametric plots (middle) are the histograms of p-values from implementing parametric bootstrap. The Copula plots are the histograms of p-values from implementing parametric bootstrap with copulas for dependence and correcting for fitted parameters. The data is generated from $Gamma(3, 1)$ (left) and $GEV(0, .2, 1)$ (right) with a correlation coefficient of $r = 0.5$. In each case, 1000 replicate tests were performed with sample size $n = 100$ and $B = 1000$ bootstrap samples.

of the truth, the fix can work. However, in cases such as AR(2), if the approximation is too far from the true dependence structure the fix does not completely solve the issue. As well, tests were only performed with the normal copula. It is possible that other copulas could provide stronger results based on the true dependence of the data.

References

Arnold, T. A. and J. W. Emerson (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal 3*(2), 34–39.

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

Capasso, M., L. Alessi, M. Barigozzi, and G. Fagiolo (2009). On approximating the distributions of goodness-of-fit tets statistics based on the empirical distribution function: The case of unknown parameters. *Advances in Complex Systems 12*, 157–167.

Conover, W. J. (1972). A Kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association 67*(339), 591–596.

Dimitrova, D. S., V. K. Kaishev, and S. Tan (2020). Computing the Kolmogorov-Smirnov distribution when the underlying cdf is purely discrete, mixed, or continuous. *Journal of Statistical Software 95*, 1–42.

Durilleul, P. and P. Legendre (1992). Lack of robustness in two tests of normality against autocorrelation in sample data. *Journal of Statistical Computation and Simulation 42*(1-2), 79–91.

Genest, C. and B. Rémillard (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 44*(6), 1096–1127.

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

Hofert, M., I. Kojadinovic, M. Maechler, and J. Yan (2018). *Elements of Copula Modeling with R*. Springer Use R! Series.

Hofert, M., I. Kojadinovic, M. Maechler, and J. Yan (2020). *copula: Multivariate Dependence with Copulas*. R package version 1.0-1.

Lanzante, J. (2021, 05). 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*(318), 399–402.

Marsaglia, G., W. W. Tsang, and J. Wang (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*(253), 68–78.

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

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Steinskog, D. J., D. B. Tjøstheim, and N. G. Kvamstø (2007). A cautionary note on the use of the Kolmogorov-Smirnov test for normality. *Monthly Weather Review 135*(3), 1151–1157.

Weiss, M. S. (1978). Modification of the Kolmogorov-Smirnov statistic for use with correlated data. *Journal of the American Statistical Association 73*(364), 872–875.