

# Data-driven stabilizations of goodness-of-fit tests

Alberto Fernández-de-Marcos<sup>1,2</sup> and Eduardo García-Portugués<sup>1</sup>

## Abstract

Exact null distributions of goodness-of-fit test statistics are generally challenging to obtain in tractable forms. Practitioners are therefore usually obliged to rely on asymptotic null distributions or Monte Carlo methods, either in the form of a lookup table or carried out on demand, to apply a goodness-of-fit test. There exist simple and useful transformations of several classic goodness-of-fit test statistics that stabilize their exact- $n$  critical values for varying sample sizes  $n$ . However, detail on the accuracy of these and subsequent transformations in yielding exact  $p$ -values, or even deep understanding on the derivation of several transformations, is still scarce nowadays. The latter stabilization approach is explained and automated to (i) expand its scope of applicability and (ii) yield upper-tail exact  $p$ -values, as opposed to exact critical values for fixed significance levels. Improvements on the stabilization accuracy of the exact null distributions of the Kolmogorov–Smirnov, Cramér–von Mises, Anderson–Darling, Kuiper, and Watson test statistics are shown. In addition, a parameter-dependent exact- $n$  stabilization for several novel statistics for testing uniformity on the hypersphere of arbitrary dimension is provided. A data application in astronomy illustrates the benefits of the advocated stabilization for quickly analyzing small-to-moderate sequentially-measured samples.

**Keywords:** Exact distribution; Goodness-of-fit;  $p$ -value; Stabilization; Uniformity.

## 1 Introduction

The classical one-sample goodness-of-fit problem is concerned with testing the null hypothesis in which the cumulative distribution function (cdf)  $F$  of an independent and identically distributed (iid) random sample  $X_1, \dots, X_n$  equals a certain prescribed cdf  $F_0$ . The most popular class of goodness-of-fit statistics for testing  $\mathcal{H}_0 : F = F_0$  is arguably that based on  $F_n$ , the empirical cumulative distribution function (ecdf) of  $X_1, \dots, X_n$ . Ecdf-based test statistics confront  $F_n$  against  $F_0$ , their best-known representatives being the Kolmogorov–Smirnov ( $D_n$ ), Cramér–von Mises ( $W_n^2$ ), and Anderson–Darling ( $A_n^2$ ) statistics, all of them generating omnibus tests of  $\mathcal{H}_0$  against  $\mathcal{H}_1 : F \neq F_0$ . When  $F_0$  is continuous, testing  $\mathcal{H}_0$  reduces to testing whether the iid sample  $U_1, \dots, U_n$ ,  $U_i := F_0(X_i)$ ,  $i = 1, \dots, n$ , is distributed as  $\text{Unif}(0, 1)$ , the continuous uniform distribution on  $(0, 1)$ . Hence, tests of uniformity, despite their a priori limited applicability, provide powerful approaches to most of the goodness-of-fit problems concerned with fully-specified null hypotheses. In particular, the above ecdf-based statistics have the attractive property of being distribution-free, i.e., their exact null distributions do not depend on  $F_0$ .

Both ecdf-based tests and uniformity tests have been exported to deal with data naturally arising in supports different from  $\mathbb{R}$  or subsets thereof. This is the case of directional data, that is, data supported on the unit hypersphere  $\mathbb{S}^{p-1} := \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\| = 1\}$ ,  $p \geq 2$ , which commonly occurs in the form of circular ( $p = 2$ ) or spherical ( $p = 3$ ) data. The analysis of directional data faces specific challenges due to the non-Euclidean nature of the support; see Mardia and Jupp (1999) for a book-length treatment of tailored statistical methods and Pewsey and García-Portugués (2021) for a review of recent advances. In particular, tests of uniformity on  $\mathbb{S}^{p-1}$  must be invariant under arbitrary rotations of the data coordinates, as these do not alter the uniform/non-uniform nature of the data.

<sup>1</sup>Department of Statistics, Carlos III University of Madrid (Spain).

<sup>2</sup>Corresponding author. e-mail: albertfe@est-econ.uc3m.es.

While a sizable number of tests of uniformity on  $\mathbb{S}^{p-1}$  exist (see a review in García-Portugués and Verdebout (2018)), perhaps the two most known omnibus tests are those of Kuiper (1960) and Watson (1961) on  $\mathbb{S}^1$ : their statistics,  $V_n$  and  $U_n^2$ , can be regarded as the rotation-invariant versions of the Kolmogorov–Smirnov and Cramér–von Mises tests of uniformity, respectively. Moving beyond  $\mathbb{S}^1$  has proven a challenging task for ecdf-based tests up to relatively recent years, with Cuesta-Albertos et al. (2009) using a Kolmogorov–Smirnov test on random projections data and García-Portugués et al. (2020) proposing a class of projected-ecdf statistics that extends Watson (1961)’s test to  $\mathbb{S}^{p-1}$  (see Section 3.1). As in the classical setting, tests of uniformity on  $\mathbb{S}^{p-1}$  allow for testing the goodness-of-fit of more general distributions: in  $\mathbb{S}^1$ , this is a straightforward application of the probability integral transform in the angles space  $[-\pi, \pi)$ ; the case  $\mathbb{S}^{p-1}$ ,  $p \geq 3$ , is remarkably more complex and has been recently put forward in Jupp and Kume (2020).

| Statistic | Exact distribution approximations                                                                                                                                                                                            |
|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $D_n$     | Massey (1950, 1951) <sup>*,†</sup> , Birnbaum (1952) <sup>‡</sup> , Maag and Dicaire (1971) <sup>§</sup> , Marsaglia et al. (2003) <sup>‡</sup> , Brown and Harvey (2007) <sup>†,††</sup> , Facchinetti (2009) <sup>††</sup> |
| $W_n^2$   | Marshall (1958) <sup>††</sup> , Pearson and Stephens (1962) <sup>¶,‡‡</sup> , Tiku (1965) <sup>‡</sup> , Stephens and Maag (1968) <sup>‡,¶,‡‡</sup> , Knott (1974) <sup>**</sup> , Csörgö and Faraway (1996) <sup>‡</sup>    |
| $V_n$     | Stephens (1965) <sup>*</sup> , Maag and Dicaire (1971) <sup>§</sup> , Durbin (1973); Arsham (1988) <sup>††</sup>                                                                                                             |
| $U_n^2$   | Pearson and Stephens (1962) <sup>¶,‡‡</sup> , Tiku (1965) <sup>‡</sup> , Quesenberry and Miller Jr (1977) <sup>‡‡</sup>                                                                                                      |
| $A_n^2$   | Lewis (1961) <sup>‡‡</sup> , Marsaglia and Marsaglia (2004) <sup>  </sup>                                                                                                                                                    |

Table 1: Summary of existing specific approaches for approximating exact distributions of several goodness-of-fit test statistics. The approximations rely of the following main techniques: difference equations<sup>\*</sup>, recursive formulae<sup>†</sup>, truncated approximations<sup>‡</sup>, asymptotic expansions<sup>§</sup>, approximation of distribution moments<sup>¶</sup>, correction factors<sup>||</sup>, characteristic function approximation<sup>\*\*</sup>, direct formulae<sup>††</sup>, and Monte Carlo simulations<sup>‡‡</sup>.

Historically, applications of goodness-of-fit tests were somehow hampered due to the absence of exact distribution theory for finite sample sizes. Statisticians focused on giving extensive tables of critical values for each statistic’s exact distribution and, alternatively, approximating exact distributions of remarkable statistics. Table 1 lists the approximations available for the exact distributions of  $D_n$ ,  $W_n^2$ ,  $V_n$ ,  $U_n^2$ , and  $A_n^2$ , as well as the main techniques behind them. Although these specific approximations are highly accurate, the complexity of their expressions, and the lack of straightforward applicability to other statistics beyond the ones they were designed for, have not displaced the customary use of Monte Carlo simulations, asymptotic distributions, or even lookup tables when emitting general test decisions. In order to reduce the size of lookup tables, Stephens (1970) transformed several statistics  $T_n$  (among others,  $D_n$ ,  $V_n$ ,  $W_n^2$ , and  $U_n^2$ ) into  $T_n^*$  in such a way that the upper tails of  $T_n^*$  remain roughly constant on  $n$ . Comparing  $T_n^*$  (and not  $T_n$ ) with certain fixed asymptotic critical values for  $T_n$  gives a more accurate test calibration for small-to-moderate  $n$ ’s. This approach also allowed finding finite-sample approximations in a wider set of goodness-of-fit problems: Stephens (1974, 1977, 1979) and D’Agostino and Stephens (1986) derived analogous transformations for  $D_n$ ,  $V_n$ ,  $W_n^2$ ,  $U_n^2$ , and  $A_n^2$  when testing the goodness-of-fit of normal, exponential, logistic, and extreme value distributions. Other authors, such as Dufour and Maag (1978), found modifications for  $D_n$  to use with truncated or censored samples, and Crown (2000) applied this method to an  $A_n^2$ -related statistic for testing normal and exponential distributions. Hegazy and Green (1975) found transformations for new test statistics by fitting a functional relationship between the critical values and the sample size, introducing the first explicit use of a regression view to stabilize test statistics and offering insight into Stephens’ original work. Pettitt (1977) also applied this regression approach to  $A_n^2$  for normality tests. Johannes and Rasche (1980) proposed an improved modification for Durbin (1969)’s  $C$  statistic, finding a specific transformation for each significance level; these approximations give more accurate results for a wider set of significance levels,

yet at the expense of tabulating a higher number of transformations. More recently, using several regressions for different significance levels too, Marks (1998, 2007) found transformations for  $D_n$  to test for Erlang distributions, while Heo et al. (2013) did the same for  $A_n^2$  with several extreme value distributions. As Table 2 shows, Stephens’ transformations are present in nowadays’ R software for goodness-of-fit testing, which also implements some of the statistic-specific approaches from Table 1.

| Methodology          | R package             | Statistics and references                                                   |
|----------------------|-----------------------|-----------------------------------------------------------------------------|
| Exact distributions  | <code>goftest</code>  | $W_n^2$ (Csörgö and Faraway, 1996), $A_n^2$ (Marsaglia and Marsaglia, 2004) |
|                      | <code>stats</code>    | $D_n$ (Marsaglia et al., 2003)                                              |
| Transformation-based | <code>circular</code> | $V_n, U_n^2$ (Stephens, 1970)                                               |
|                      | <code>sphunif</code>  | $D_n, W_n^2, V_n, U_n^2$ (Stephens, 1970)                                   |
|                      | <code>EnvStats</code> | $D_n, W_n^2, A_n^2$ (D’Agostino and Stephens, 1986)                         |

Table 2: R packages implementing different approximation methods to compute exact  $p$ -values of goodness-of-fit tests: `circular` (Agostinelli and Lund, 2017), `sphunif` (García-Portugués and Verdebout, 2021), `EnvStats` (Millard, 2013), `goftest` (Faraway et al., 2019), and `stats` (R Core Team, 2021).

In this paper we build on Stephens’ transformations to expand and automate them. First, we present a data-driven procedure to achieve a better stabilization, with respect to the sample size  $n$ , of the exact null distribution of a generic test statistic  $T_n$  of interest, for a wider range of significance levels  $\alpha$  (i.e., upper  $\alpha$ -quantiles of  $T_n$ ). Specifically, new modifications for the (one-sample) Kolmogorov–Smirnov, Cramér–von Mises, Kuiper, and Watson test statistics are derived and shown to extend the scope of applicability of previous approaches. To the best of our knowledge, we also provide the first instance of such a stabilization for the Anderson–Darling test statistic. Second, we provide a method to approximate upper-tail exact  $p$ -values for the tests constructed from stabilized statistics. Through an extensive simulation study, we show a significant improvement in the precision of the stabilization of the exact critical values of  $T_n$  for several sample sizes, as well as a competitive computational cost when compared with statistic-specific methods for evaluating exact null distributions. We also show large improvements, both in precision and computational efficiency, over the use of Monte Carlo simulation, arguably the most popular test calibration approach nowadays. Third, we develop an extension of our stabilization procedure to deal with several recent test statistics for assessing uniformity on  $\mathbb{S}^{p-1}$ ,  $p \geq 2$ , and which hence have dimension-dependent distributions. In particular, we stabilize the exact null distribution of a novel Anderson–Darling test statistic for circular data. Finally, the introduced stabilization methodology allows us to perform tests in batches of small-to-moderate samples in an accurate and fast manner that does not require Monte Carlo simulation. This is illustrated in an astronomical dataset comprised of the longitudes at which sunspots appear, which exhibits a suspected temporal mix of uniform and non-uniform patterns.

The rest of the paper is organized as follows. Section 2 introduces Stephens’ approach (Section 2.1) and our proposed extension (Section 2.2), together with simulation studies and a comparison between several modifications (Section 2.3). Section 3 briefly introduces the projected-ecdf statistics for testing uniformity on the hypersphere (Section 3.1), develops the parameter-dependent transformations to achieve their stabilization (Section 3.2), and analyzes the empirical performance of these transformations (Section 3.3). Section 4 gives an application of the modified statistics in astronomy. A final discussion of the obtained results concludes the paper in Section 5. Further analyses and empirical results are included in the Supplementary Material (SM). All the code and data are available at <https://github.com/afernandezdemarcos/approxstats>.

## 2 Stabilization of ecdf statistics

### 2.1 On Stephens' stabilization

Stephens (1970) stabilization aims to transform a statistic  $T_n$  into  $T_n^*$  through a function of  $n$ , so that the upper  $\alpha$ -quantiles of  $T_n^*$  are well approximated by the upper  $\alpha$ -quantiles of  $T_\infty$ , the random variable distributed as the asymptotic null distribution of  $T_n$ , for small-to-moderate sample sizes. The transformation can be interpreted as a two-step stabilization. First, in the *quantile ratios stabilization*,  $T_n$  is modified to the statistic  $T_n^{\alpha_0-s}$  so that the ratios of  $T_n^{\alpha_0-s}$ 's upper  $\alpha$ -quantiles with respect to a certain reference upper  $\alpha_0$ -quantile are roughly constant as a function of  $n$ . Second, in the *asymptotic stabilization*,  $T_n^{\alpha_0-s}$  is transformed into  $T_n^*$  so that the upper  $\alpha$ -quantiles of  $T_n^*$  are approximately equal to the asymptotic upper  $\alpha$ -quantiles for small-to-asymptotic sample sizes. For the sake of brevity, and since we are concerned only with upper-tail tests, henceforth we will use “ $\alpha$ -quantile” as a replacement for “upper  $\alpha$ -quantile”.

The ratios involved in the first step are  $T_{n;\alpha}/T_{n;\alpha_0}$ , where  $T_{n;\alpha}$  is the  $\alpha$ -quantile of the distribution for sample size  $n$ , i.e.,  $\mathbb{P}[T_n \geq T_{n;\alpha}] = \alpha$ . Obviously, these ratios do not have to be constant for all  $n$ , as Figure 1 shows for  $W_n^2$ . The *quantile ratios stabilization* step searches for a transformed statistic,  $T_n^{\alpha_0-s}$ , whose quantile ratios  $T_{n;\alpha}^{\alpha_0-s}/T_{n;\alpha_0}^{\alpha_0-s}$  do not depend on  $n$ . In other words, the desideratum is that these quantile ratios, for any sample size  $n$ , equal the asymptotic quantile ratios  $T_{\infty;\alpha}/T_{\infty;\alpha_0}$ , where  $T_{\infty;\alpha}$  is the asymptotic  $\alpha$ -quantile. One way to find such transformation is by setting  $T_n^{\alpha_0-s} := T_n - p(n)$  for a certain function  $p : \mathbb{N} \rightarrow \mathbb{R}$  such that it verifies  $\lim_{n \rightarrow \infty} p(n) = 0$  and the second equality below, for all  $n$  and  $\alpha$ :

$$\frac{T_{n;\alpha}^{\alpha_0-s}}{T_{n;\alpha_0}^{\alpha_0-s}} = \frac{T_{n;\alpha} - p(n)}{T_{n;\alpha_0} - p(n)} = \lim_{n \rightarrow \infty} \frac{T_{n;\alpha} - p(n)}{T_{n;\alpha_0} - p(n)} = \frac{T_{\infty;\alpha}}{T_{\infty;\alpha_0}} =: k_{\infty;\alpha}. \quad (1)$$

Hence,  $p$  is such that

$$p(n) = \frac{T_{n;\alpha} - k_{\infty;\alpha} \cdot T_{n;\alpha_0}}{1 - k_{\infty;\alpha}},$$

which clearly depends on  $\alpha$ . Stephens fitted  $p$  (see the end of this section) for a specific value of  $\alpha$ , at the expense of accuracy for other significance levels. Upon this step, the quantile ratios of  $T_n^{\alpha_0-s}$  are roughly constant for all  $n$ , as Figure 1 shows for  $W_n^2$ . This first step can be omitted for statistics with quantile ratios that are already roughly stable, as it is remarkably the case of  $D_n$  and  $V_n$  (Stephens, 1970, Section 5). In this case,  $p \approx 0$ .

The *asymptotic stabilization* step aims to transform the already modified statistic,  $T_n^{\alpha_0-s}$ , into  $T_n^*$  so that the  $\alpha$ -quantiles of this latter statistic are well approximated by the asymptotic  $\alpha$ -quantiles of the original statistic  $T_n$ . For that goal,  $g : \mathbb{N} \rightarrow \mathbb{R}$  is defined as  $g(n) := T_{\infty;\alpha}/T_{n;\alpha}^{\alpha_0-s}$ . Owing to (1), in principle this function does not depend on the significance level  $\alpha$ , only on  $\alpha_0$ :

$$\frac{T_{\infty;\alpha}}{T_{n;\alpha}^{\alpha_0-s}} = \frac{T_{\infty;\alpha_0}}{T_{n;\alpha_0}^{\alpha_0-s}}, \quad (2)$$

which holds for any value of  $\alpha$ . However, when  $p$  and  $g$  are fitted in practice, (2) will approximately hold for a certain set of  $\alpha$  values, as those shown in Figure 1. The function  $g$  is estimated from the ratio between  $T_{\infty;\alpha}/T_{n;\alpha}^{\alpha_0-s}$  for a particular value  $\alpha_1$  (possibly different from  $\alpha_0$ ), which could result in a loss of accuracy for other quantiles.

The final modified form of  $T_n$  is

$$T_n^* = T_n^{\alpha_0-s} \cdot g(n) = (T_n - p(n)) \cdot g(n), \quad (3)$$

where we highlight that in practice the functions  $p$  and  $g$  have to be estimated beforehand. Once these fits are readily available, the main benefit of (3) is the simplicity of its use, which only

requires evaluating a simple  $n$ -dependent transformation of  $T_n$ . The fits of  $p$  and  $g$  were originally handcrafted on a case-by-case basis (even “found by trial”, Stephens, 1970, Section 5), or were heavily influenced by Stephens’ functional forms, which pose significant limitations in terms of automation and flexibility. Moreover, the approximation error to the exact quantiles of  $T_n$  that is obtained is, first, dependent on  $\alpha_1$  and, second, significant for  $\alpha$ -quantiles different from  $\alpha_1$ . An additional downside of (3) is the initial stabilization step, which increases the complexity and tuning required (selection of  $\alpha_0$ ), and is a source of uncertainty for the final approximation. In order to overcome these problems, we present in the next section an enhanced stabilization approach that improves the accuracy of exact  $\alpha$ -quantiles while retaining the simplicity of the transformation.

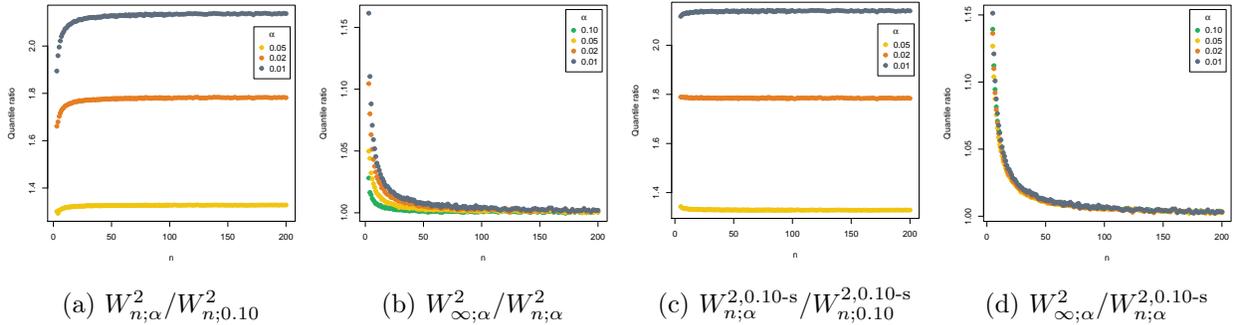


Figure 1: Quantile ratios of the Cramér–von Mises statistic  $W_n^2$  (leftmost two figures) and its ratio-stabilized statistic  $W_n^{2,0.10-s}$  (rightmost two figures).

## 2.2 $(n, \alpha)$ -stabilization

Our stabilization consists of a single-step transformation of the original statistic  $T_n$  into  $T_n^*(\alpha)$  by a function that depends on the sample size  $n$  and the significance level  $\alpha$  at which the test is to be performed, so that the exact  $\alpha$ -quantile of  $T_n^*(\alpha)$  is closely approximated by the  $\alpha$ -quantile of  $T_\infty$ . Additionally to its improved accuracy and simplicity, an advantage of our modification is that it compresses extensive lookup tables: critical values do not need to be available for different significance levels because  $\alpha$  is already included within the transformation.

The ratios  $T_{\infty;\alpha}/T_{n;\alpha}$ , shown in Figure 1 for  $W_n^2$ , can be directly modeled as a function  $g : \mathbb{N} \times (0, 1) \rightarrow \mathbb{R}$  of  $(n, \alpha)$ , hence condensing the two steps from Section 2.1 into one. To that aim, we define  $g$  as the function satisfying

$$\alpha = \mathbb{P}[T_n \geq T_{n;\alpha}] = \mathbb{P}[T_n \geq T_{\infty;\alpha}/g(n, \alpha)], \quad (4)$$

for all  $(n, \alpha)$ , and our transformed statistic (for the  $\alpha$  significance level) as

$$T_n^*(\alpha) := T_n \cdot g(n, \alpha).$$

It is very convenient to reexpress  $g$ , as defined in (4), as

$$\frac{T_{\infty;\alpha}}{T_{n;\alpha}} = g(n, \alpha) + \varepsilon, \quad (5)$$

where  $\varepsilon = 0$  if (4) is perfectly satisfied for all  $(n, \alpha)$ . Indeed, Equation (5) casts the stabilization problem as an error-free fixed-design regression problem with predictors  $(n, \alpha)$ , response  $Y := T_{\infty;\alpha}/T_{n;\alpha}$ , and unknown regression function  $g$ . Casting Stephens’ stabilizations as a regression problem was early introduced in Hegazy and Green (1975), Pettitt (1977), and Johannes and Rasche (1980). Yet these works focus on using the sample size as the unique predictor, for isolated  $\alpha$ -quantiles, an approach that has been later applied in Marks (1998, 2007) and Heo et al. (2013).

We introduce now a sufficiently flexible parametric specification for  $g$  in (5) that allows its effective estimation in practice. We resort to a linear model featuring negative powers of the sample size  $n$  and significance level  $\alpha$  as predictors, as well as the corresponding interaction effects between them. Precisely, we consider the following saturated model:

$$g(n, \alpha) = 1 + \frac{\beta_1}{\sqrt{n}} + \frac{\beta_2}{n} + \frac{\beta_3}{\sqrt{n\alpha}} + \frac{\beta_4}{\sqrt{n\alpha}} + \frac{\beta_5}{n\sqrt{\alpha}} + \frac{\beta_6}{n\alpha}. \quad (6)$$

The fixed intercept and negative powers of  $n$  were included to guarantee that  $\lim_{n \rightarrow \infty} g(n, \alpha) = 1$ , thus in accordance with  $\lim_{n \rightarrow \infty} T_{\infty; \alpha} / T_{n; \alpha} = 1$ . Powers of  $n^{-1/2}$  resemble the sample size factors in the terms of an Edgeworth series. The powers of  $\alpha^{-l/2}$ ,  $l = 1, 2$ , were experimentally found to be an appropriate specification for capturing the interactions with  $n$ . The appropriateness of the model specification (6) is exhaustively investigated in A in the SM. Upon available samples of the form  $\{(n_j, \alpha_j, Y_j)\}_{j=1}^J$ ,  $Y_j := T_{\infty; \alpha_j} / T_{n_j; \alpha_j}$ , model (6) is estimated through weighted least squares, using the weight  $w_j := n_j^{-1/2} 1_{\{0 < \alpha_j \leq 0.25\}}$  for the  $j$ -th observation to give heavier weight to the approximation error on lower sample sizes. The indicator in  $w_j$  reflects our interest in only stabilizing the upper tail of the test statistic  $T_n$ , hence disregarding those quantiles associated with non-rejections of the test based on  $T_n$ . B in the SM provides more detail on the selection of the weight function among several alternatives.

The data required for fitting (6) is to be produced under the (fairly realistic nowadays) assumption that it is feasible to simulate a large number of statistics  $T_n$  under the null hypothesis and for varying sample sizes. Specifically, we have carried out the following simulation for the test statistics  $D_n$ ,  $W_n^2$ ,  $V_n$ ,  $U_n^2$ , and  $A_n^2$ . We produced  $M = 10^7$  Monte Carlo random samples for  $T_n$ , for each of the sample sizes  $n$  in the set  $\mathcal{N} := \{5, \dots, 100, 102, \dots, 200, 204, \dots, 300, 308, \dots, 404, 420, \dots, 500\}$ . We then condensed these statistics as the quantiles  $\{T_{n_j; \alpha_j} : n_j \in \mathcal{N}, \alpha_j \in \mathcal{A}\}$ , for  $\mathcal{A} := \{a/A : a = 1, \dots, A\}$ ,  $A = 10^3$ . The asymptotic  $\alpha$ -quantiles  $\{T_{\infty; \alpha_j} : \alpha_j \in \mathcal{A}\}$  were computed from the statistics' asymptotic null distributions, as those were readily available in the literature. The generated sample is therefore  $\{(n_j, \alpha_j, Y_j)\}_{j=1}^J$ ,  $J = \#\mathcal{N} \times A$ . Clearly, this is a computationally-intensive process, although it only needs to be done once per kind of test statistic. The procedure is analogous for other one-sample test statistics that are feasible to simulate under the simple null hypothesis at hand. If the limiting distribution is not available or tractable, a sufficiently large sample size  $n$  could be used to approximate  $T_{\infty; \alpha}$  by  $T_{n; \alpha}$  by Monte Carlo.

Using the sample  $\{(n_j, \alpha_j, Y_j)\}_{j=1}^J$ , we advocate the use of stepwise regression for performing model selection within (6). Specifically, we performed a forward-backward search for minimizing the Bayesian Information Criterion (BIC) on the space of models contained in (6). The search was initiated with the model featuring only the predictors used in Stephens' modifications, i.e.,  $n^{-1/2}$  and  $n^{-1}$ . To attain simpler models than the BIC-optimal one, a final step was implemented to iteratively drop one-by-one the predictors that contributed the least to the adjusted  $R^2$  of the resulting model. The threshold was established to keep only three final terms (for simplicity), the predictors removed decreasing less than 0.15% the  $R_{\text{adj}}^2$  which, averaged across the five statistics, was larger than 0.96.

The resulting modified forms for  $D_n$ ,  $W_n^2$ ,  $V_n$ ,  $U_n^2$ , and  $A_n^2$  are collected in Table 3. All of the transformations have three correcting terms, one dependent on  $n$  and the other two related to  $n$  and  $\alpha$ ,  $(n\sqrt{\alpha})^{-1}$  being common to the five statistics. Interestingly, the same correction terms are present within the groups of supremum- and quadratic-norm statistics, as well as in the pairs of linear and circular variants. These forms are valid for  $n \geq 5$ , which anecdotally gives a minor improvement over Stephens' forms, valid for  $n \geq 8$ . The steps to use them with the upper-tail test for  $\mathcal{H}_0$  that is based on  $T_n$  and that is carried out at the significance level  $\alpha$  are as follows:

- (i) Compute the test statistic  $T_n$  using its original form.
- (ii) Calculate the corresponding modified test statistic,  $T_n^*(\alpha)$ , in Table 3.
- (iii) Retrieve an asymptotic critical value  $T_{\infty; \alpha}$  in Table 3. If  $T_n^*(\alpha) > T_{\infty; \alpha}$ , reject  $\mathcal{H}_0$  at significance level  $\alpha$ .

| $T_n$   | $T_n^*(\alpha)$                                                                                           | $T_{\infty;0.15}$ | $T_{\infty;0.1}$ | $T_{\infty;0.05}$ | $T_{\infty;0.025}$ | $T_{\infty;0.01}$ |
|---------|-----------------------------------------------------------------------------------------------------------|-------------------|------------------|-------------------|--------------------|-------------------|
| $D_n$   | $D_n \left( 1 + \frac{0.1575}{\sqrt{n}} + \frac{0.0192}{n\sqrt{\alpha}} - \frac{0.0051}{n\alpha} \right)$ | 1.1380            | 1.2239           | 1.3581            | 1.4803             | 1.6277            |
| $W_n^2$ | $W_n^2 \left( 1 - \frac{0.1651}{n} + \frac{0.0749}{n\sqrt{\alpha}} - \frac{0.0014}{n\alpha} \right)$      | 0.2841            | 0.3473           | 0.4613            | 0.5806             | 0.7435            |
| $V_n$   | $V_n \left( 1 + \frac{0.2330}{\sqrt{n}} + \frac{0.0276}{n\sqrt{\alpha}} - \frac{0.0068}{n\alpha} \right)$ | 1.5370            | 1.6196           | 1.7473            | 1.8625             | 2.0010            |
| $U_n^2$ | $U_n^2 \left( 1 - \frac{0.1505}{n} + \frac{0.0917}{n\sqrt{\alpha}} - \frac{0.0018}{n\alpha} \right)$      | 0.1313            | 0.1518           | 0.1869            | 0.2220             | 0.2685            |
| $A_n^2$ | $A_n^2 \left( 1 + \frac{0.0360}{n} - \frac{0.0234}{n\sqrt{\alpha}} + \frac{0.0006}{n\alpha} \right)$      | 1.6212            | 1.9331           | 2.4922            | 3.0775             | 3.8784            |

Table 3: Modified statistics for sample size  $n$  and significance level  $\alpha$ . Modified forms are valid for  $n \geq 5$  and  $0 < \alpha \leq 0.25$ .  $\mathcal{H}_0$  is rejected at significance level  $\alpha$  if  $T_n^*(\alpha) > T_{\infty;\alpha}$ .

The transformed statistics can also be used to obtain approximations to exact  $p$ -values, provided the asymptotic quantiles  $\mathcal{T}_{\infty} := \{T_{\infty;\alpha_j} : \alpha_j \in \mathcal{A}\}$  have been precomputed. This is done in two steps. First,  $p$ -value bounds  $[\alpha_1, \alpha_2]$  are obtained from the grid  $\mathcal{A}$  such that  $T_n^*(\alpha_1) \leq T_{\infty;\alpha_1}$  and  $T_n^*(\alpha_2) > T_{\infty;\alpha_2}$ . Once these discrete bounds for  $p$ -value are available, a linear interpolation is applied to define  $t_{\infty}(\alpha) := T_{\infty;\alpha_1} + (T_{\infty;\alpha_2} - T_{\infty;\alpha_1})(\alpha - \alpha_1)/(\alpha_2 - \alpha_1)$  for  $\alpha \in [\alpha_1, \alpha_2]$  and then the root  $\alpha^* \in [\alpha_1, \alpha_2]$  of

$$T_n^*(\alpha^*) = t_{\infty}(\alpha^*) \quad (7)$$

is obtained by Newton–Raphson (NR). The approximate  $p$ -value is then set to  $\alpha^*$ . If  $\alpha_1 \geq \alpha_{\max}$ ,  $\alpha_{\max} = 0.25$  being the maximum element in  $\mathcal{A}$  for which the transformation has been estimated,  $p$ -value =  $\alpha_{\max}$  is returned. Algorithm 1 summarizes this process.

---

**Algorithm 1**  $p$ -value approximation using the  $(n, \alpha)$ -modification

---

```

1: function PVALUE_APPROX( $T_n, n, \mathcal{T}_{\infty}, \mathcal{A}$ )
2:   for  $j$  from 1 to  $\#\mathcal{A}$  do
3:      $T_{\text{mod},\alpha} \leftarrow T_n^*(T_n, n, \mathcal{A}[j])$ 
4:     if  $T_{\text{mod},\alpha} > \mathcal{T}_{\infty}[j]$  then
5:       if  $j = 1$  then
6:          $(\alpha_1, \alpha_2) \leftarrow (\mathcal{A}[j], \mathcal{A}[j+1])$ 
7:          $(T_{\infty;\alpha_2}, T_{\infty;\alpha_1}) \leftarrow (\mathcal{T}_{\infty}[j], \mathcal{T}_{\infty}[j+1])$ 
8:       else
9:          $(\alpha_1, \alpha_2) \leftarrow (\mathcal{A}[j-1], \mathcal{A}[j])$ 
10:         $(T_{\infty;\alpha_2}, T_{\infty;\alpha_1}) \leftarrow (\mathcal{T}_{\infty}[j-1], \mathcal{T}_{\infty}[j])$ 
11:         $\alpha^* \leftarrow \text{NR}(T_n^*(T_n, n, \alpha) - t_{\infty}(\alpha, \mathcal{T}_{\infty}, \alpha_1, \alpha_2))$ 
12:        return  $\alpha^*$ 
13:   return 0.25

```

---

When there is no  $\alpha_1$  in  $\mathcal{A}$  such that  $T_n^*(\alpha_1) \leq T_{\infty;\alpha_1}$ , the  $p$ -value is set as the nonnegative extrapolation of the root in (7), with  $\alpha_1$  and  $\alpha_2$  being the two lowest elements in  $\mathcal{A}$ .

### 2.3 Simulation study

For the test statistics  $D_n$ ,  $V_n$ ,  $W_n^2$ ,  $U_n^2$ , and  $A_n^2$ , we evaluate next the divergence of the exact- $n$  critical values under  $\mathcal{H}_0$  from their corresponding approximations given by: (a) Stephens’ modified forms; (b) the particular approximation methods from Table 2; (c) Monte Carlo approximation with  $10^4$  trials; and (d) our proposed transformations. Figure 2 displays the relative errors for the rejection proportions generated by approximated critical values based on methods (a)–(d). These relative

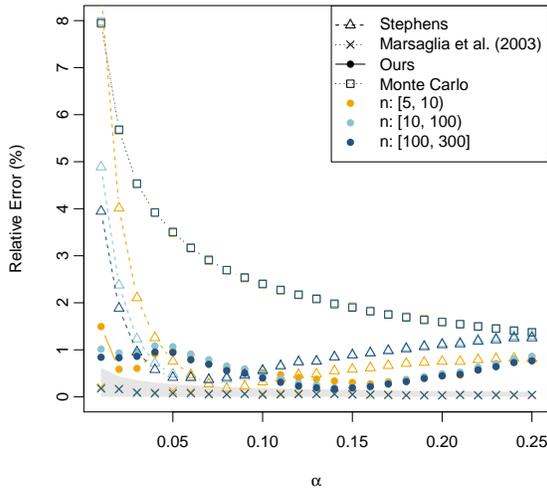
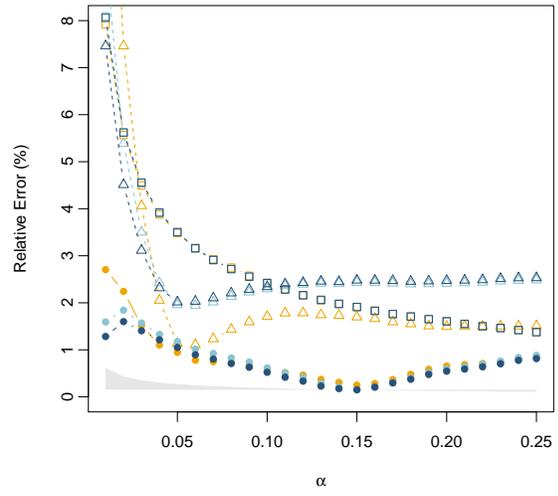
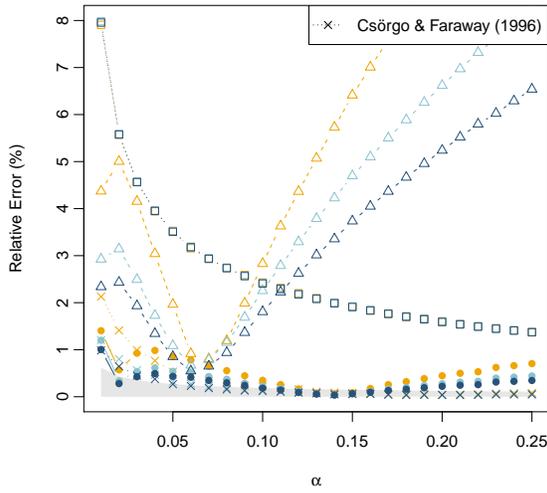
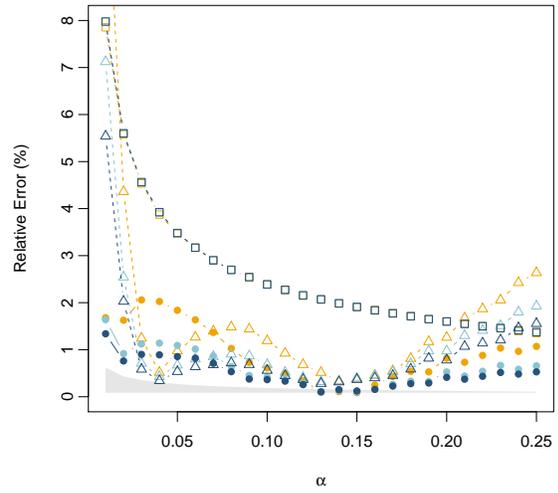
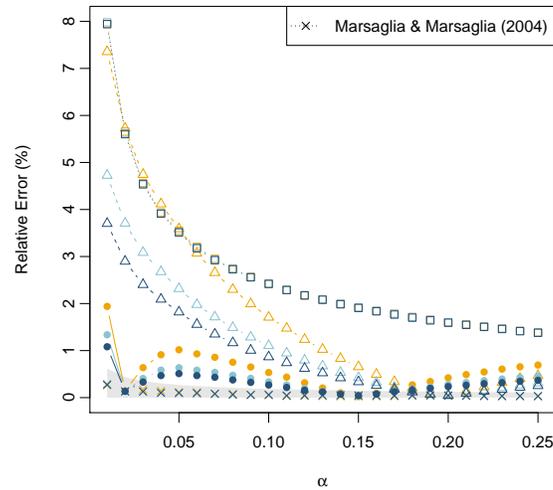
(a)  $D_n$ (b)  $V_n$ (c)  $W_n^2$ (d)  $U_n^2$ (e)  $A_n^2$ 

Figure 2: Relative error (in %)  $|\alpha - \tilde{\alpha}|/\alpha$  between the significance level  $\alpha$  and  $\tilde{\alpha}$ , the empirical rejection rate using an approximated exact- $n$  critical value, averaged across different sample sizes  $n$ . The legend in Figure 2a details the approximation methods considered and applies to the rest of the panels, with different specific methods in Figures 2c and 2e. The gray shaded area corresponds to the 95% confidence interval of the relative error when  $\tilde{\alpha}$  is produced by the exact- $n$  critical value estimated by  $M = 10^7$  Monte Carlo samples.

errors are defined as  $|\alpha - \tilde{\alpha}|/\alpha$ , where  $\alpha$  is the significance level and  $\tilde{\alpha}$  is the empirical rejection rate obtained with  $M = 10^7$  Monte Carlo samples when using an  $\alpha$ -critical value computed by each approximation method. The  $M = 10^7$  Monte Carlo samples under  $\mathcal{H}_0$  were drawn for each of the sample sizes  $n$  in  $\mathcal{N}_{\text{test}} := \{5, \dots, 10, 20, \dots, 50, 100, 200, 300\}$ . The sample quantiles for the significance levels in  $\mathcal{A}_{\text{test}} := \{a/100 : a = 1, \dots, 25\}$  were computed for each sample size and statistic. For the critical value approximations (a) and (d), critical values were computed by applying the corresponding inverse transformation from Table 8 to the asymptotic  $\alpha$ -critical value  $T_{\infty;\alpha}$ . Obtaining the critical values in (b) is straightforward using the functions `stats::C_pKolmogorov2x` (R Core Team, 2021) for  $D_n$ , and `gofest::pCvM` and `gofest::pAD` (Faraway et al., 2019) for  $W_n^2$  and  $A_n^2$ , respectively. For the critical value approximation based on (c), the (random) relative error for each critical value was averaged over  $10^3$  simulations to give an estimate of the average Monte Carlo relative error. Each panel in Figure 2 shows the relative error along  $\mathcal{A}_{\text{test}}$  averaged for three sets of sample sizes:  $5 \leq n < 10$ ,  $10 \leq n < 100$ , and  $n \geq 100$ .

Along  $\mathcal{A}_{\text{test}}$ , the average relative errors of our stabilizations are 0.5%, 0.3%, 0.5%, 0.3%, and 0.7% for  $D_n$ ,  $W_n^2$ ,  $U_n^2$ ,  $A_n^2$ , and  $V_n$ , respectively. The relative errors remain fairly stable for every significance value in  $\mathcal{A}_{\text{test}}$  without significant differences between the sets of sample sizes analyzed. Compared to Stephens' stabilizations, our relative error is lower by a factor of  $\times 2$ ,  $\times 12$ ,  $\times 2$ ,  $\times 3$ , and  $\times 4$  on average, respectively. The largest improvements are achieved for  $\alpha \neq 0.05$ , since Stephens' stabilizations were tuned for  $\alpha = 0.05$ , and for sample sizes  $n \leq 100$ . This behavior is more obvious in  $W_n^2$  and  $U_n^2$ , which are the statistics that, in Stephens' approach, use an additional prior step for stabilizing the quantile ratios. When compared to the Monte Carlo approximation with  $10^4$  samples, our relative error is lower for every significance level and sample size tested, and improves by  $\times 5$ ,  $\times 10$ ,  $\times 5$ ,  $\times 9$ , and  $\times 4$  on average, respectively. As expected, the approximation methods that are specifically designed for each test statistic achieve the lowest relative errors.

Table 4 presents a comparison of the running times between our  $p$ -value approximation (Algorithm 1) and the already implemented  $p$ -value approximation methods for  $D_n$ ,  $W_n^2$ , and  $A_n^2$  described in Table 2. Our method is shown to be  $\times 3.8$ ,  $\times 5.4$ , and  $\times 4.8$  faster than Marsaglia et al. (2003), Csörgö and Faraway (1996), and Marsaglia and Marsaglia (2004), respectively. These methods are already implemented in C++, except for Csörgö and Faraway (1996) which is in R. Hence, C++ and R versions implementing Algorithm 1 were developed for each statistic to allow a fair comparison. In addition, Table 5 compares the running times between the  $p$ -value approximation based on Algorithm 1 and a Monte Carlo  $p$ -value approximation based on  $10^4$  trials, which shows that our method is  $\times 75 \cdot 10^4$ ,  $\times 58 \cdot 10^4$ , and  $\times 93 \cdot 10^4$  faster. Monte Carlo approximation was implemented in R code with calls to C++-coded statistics (the most time-consuming part), and the C++ version of Algorithm 1 was used. All comparisons were carried out using `microbenchmark` package (Mersmann, 2019). In order to compute the median running time of each function for a given sample size  $n$  and significance level  $\alpha$ ,  $10^3$  evaluations of the compiled functions were run after 10 warm-up runs using the same machine, a regular desktop computer with a 3.6GHz processor. In all cases, the computation of the original statistic  $T_n$  was excluded from the timings. R and C++ integration was done with the `Rcpp` package (Eddelbuettel and François, 2011).

The empirical results show that our stabilized statistics give more accurate results than those by Stephens, while still retaining the simplicity of the latter. When it comes to the Monte Carlo approximation (with  $10^4$  trials), relative errors on the empirical rejection rates are lowered by a factor that varies from  $\times 4$  to  $\times 10$ , depending on the statistic. In addition, Table 5 shows how our stabilization algorithm outperforms Monte Carlo execution times. Part of these improvements could be attributed to the R-C++ mix, as opposed to pure C++. Yet, given the massive difference in timing orders, we regard this effect as second-order. Arguably, for  $D_n$ ,  $W_n^2$ , and  $A_n^2$ , the tailored approximation methods are to be preferred due to their better accuracy. Even in these highly-competitive settings, our stabilizations still offer comparative advantages, as Figure 2 shows that their average relative error is  $< 0.7\%$ , sufficing for most practical applications, while Table 4 shows an improvement of  $\times 5$  in running times with respect to specific methods.

| $\alpha$                                                 | $n$   |       |       |       |       |       |       |       |       |       |
|----------------------------------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|                                                          | 5     | 6     | 7     | 8     | 9     | 10    | 20    | 30    | 40    | 50    |
| $D_n$ : Marsaglia et al. (2003) vs. Algorithm 1          |       |       |       |       |       |       |       |       |       |       |
| 0.01                                                     | 2.48  | 2.56  | 3.04  | 2.96  | 3.00  | 3.23  | 7.09  | 10.74 | 17.08 | 23.38 |
| 0.02                                                     | 2.28  | 2.28  | 2.40  | 2.75  | 2.80  | 2.85  | 4.80  | 9.62  | 12.04 | 17.39 |
| 0.05                                                     | 1.61  | 1.97  | 1.90  | 1.87  | 1.90  | 2.29  | 3.06  | 5.94  | 7.29  | 10.50 |
| 0.10                                                     | 1.22  | 1.21  | 1.44  | 1.43  | 1.48  | 1.49  | 2.24  | 3.33  | 4.17  | 6.05  |
| 0.15                                                     | 1.02  | 0.96  | 0.98  | 1.13  | 1.15  | 1.19  | 1.42  | 2.74  | 3.45  | 3.67  |
| 0.25                                                     | 0.68  | 0.71  | 0.70  | 0.71  | 0.70  | 0.81  | 1.04  | 1.48  | 1.82  | 2.64  |
| $W_n^2$ : Csörgö and Faraway (1996) vs. Algorithm 1      |       |       |       |       |       |       |       |       |       |       |
| 0.01                                                     | 10.43 | 10.40 | 10.17 | 10.12 | 10.03 | 10.00 | 10.60 | 10.68 | 10.66 | 11.82 |
| 0.02                                                     | 8.69  | 8.47  | 8.42  | 8.47  | 8.73  | 8.75  | 8.92  | 9.06  | 8.85  | 8.99  |
| 0.05                                                     | 5.54  | 5.53  | 5.61  | 5.57  | 5.56  | 5.58  | 5.67  | 5.68  | 5.64  | 5.68  |
| 0.10                                                     | 3.46  | 3.50  | 3.48  | 3.46  | 3.45  | 3.48  | 3.50  | 3.48  | 3.48  | 3.49  |
| 0.15                                                     | 2.50  | 2.48  | 2.49  | 2.54  | 2.48  | 2.55  | 2.57  | 2.50  | 2.51  | 2.52  |
| 0.25                                                     | 1.62  | 1.62  | 1.63  | 1.59  | 1.59  | 1.64  | 1.61  | 1.61  | 1.65  | 1.64  |
| $A_n^2$ : Marsaglia and Marsaglia (2004) vs. Algorithm 1 |       |       |       |       |       |       |       |       |       |       |
| 0.01                                                     | 6.66  | 6.28  | 6.23  | 6.14  | 6.20  | 6.42  | 6.29  | 6.18  | 6.29  | 6.43  |
| 0.02                                                     | 6.00  | 6.52  | 5.91  | 6.18  | 6.13  | 6.22  | 5.91  | 6.26  | 6.14  | 6.60  |
| 0.05                                                     | 5.12  | 5.24  | 5.72  | 5.74  | 5.36  | 6.04  | 5.24  | 5.23  | 5.44  | 5.44  |
| 0.10                                                     | 4.26  | 4.39  | 4.35  | 4.26  | 4.26  | 4.81  | 4.35  | 4.52  | 4.36  | 4.32  |
| 0.15                                                     | 3.70  | 3.62  | 3.64  | 3.78  | 3.64  | 3.65  | 3.78  | 3.75  | 3.72  | 3.74  |
| 0.25                                                     | 2.87  | 3.19  | 2.79  | 2.85  | 3.10  | 3.02  | 2.83  | 2.85  | 2.98  | 2.87  |

Table 4: Running time ratios between specific  $p$ -value approximation methods and our  $p$ -value approximation method (Algorithm 1). Ratios are computed for the median running times of  $10^3$  evaluations, for each pair  $(n, \alpha)$ . The averages of the median running times of Algorithm 1 are  $3.65\mu s$ ,  $225\mu s$  (for R version,  $4.5\mu s$  for C++ version), and  $3\mu s$  for  $D_n$ ,  $W_n^2$ , and  $A_n^2$ , respectively.

| $\alpha$                              | $n$ |    |    |    |    |    |    |     |     |     |
|---------------------------------------|-----|----|----|----|----|----|----|-----|-----|-----|
|                                       | 5   | 6  | 7  | 8  | 9  | 10 | 20 | 30  | 40  | 50  |
| $D_n$ : Monte Carlo vs. Algorithm 1   |     |    |    |    |    |    |    |     |     |     |
| 0.05                                  | 14  | 16 | 19 | 23 | 16 | 28 | 69 | 118 | 182 | 261 |
| $W_n^2$ : Monte Carlo vs. Algorithm 1 |     |    |    |    |    |    |    |     |     |     |
| 0.05                                  | 10  | 12 | 14 | 13 | 17 | 21 | 51 | 94  | 146 | 203 |
| $A_n^2$ : Monte Carlo vs. Algorithm 1 |     |    |    |    |    |    |    |     |     |     |
| 0.05                                  | 15  | 19 | 22 | 26 | 32 | 33 | 80 | 150 | 227 | 325 |

Table 5: Running time ratios, in scale  $\times 10^4$ , between a  $p$ -value Monte Carlo approximation based on  $10^4$  trials and our  $p$ -value approximation method (Algorithm 1). Ratios are computed for the median running times of  $10^3$  evaluations, for each pair  $(n, \alpha)$ . The averages of the median running times for the Monte Carlo approximation are  $2.34s$ ,  $2.35s$ , and  $2.35s$  for  $D_n$ ,  $W_n^2$ , and  $A_n^2$ , respectively.

### 3 Stabilization of parameter-dependent statistics

This section gives an extension of the  $(n, \alpha)$ -transformations introduced in Section 2.2 that is designed to stabilize the exact distributions of statistics that depend on a (known) parameter. Instances of the transformation are given for testing uniformity on  $\mathbb{S}^{p-1}$ ,  $p \geq 2$  being the statistic parameter.

### 3.1 Projected-ecdf test statistics

García-Portugués et al. (2020) proposed a class of test statistics to evaluate the null hypothesis of uniformity of an iid sample  $\mathbf{X}_1, \dots, \mathbf{X}_n$  on  $\mathbb{S}^{p-1}$ . Projected-ecdf statistics compute the weighted quadratic discrepancy between  $F_{n,\gamma}$ , the ecdf of  $\gamma'\mathbf{X}_1, \dots, \gamma'\mathbf{X}_n$  for  $\gamma \in \mathbb{S}^{p-1}$ , and  $F_p$ , the cdf of the random variable  $\gamma'\mathbf{X}$  when  $\mathbf{X} \sim \text{Unif}(\mathbb{S}^{p-1})$ . The weighted quadratic discrepancies are integrated over all possible directions  $\gamma \in \mathbb{S}^{p-1}$ , a convenient specification of the projected-ecdf statistics being

$$P_{n,p}^w := n \int_{\mathbb{S}^{p-1}} \left[ \int_{-1}^1 (F_{n,\gamma}(x) - F_p(x))^2 w(F_p(x)) dF_p(x) \right] d\gamma,$$

where  $w : [0, 1] \rightarrow \mathbb{R}$  is a certain weight function and the cdf  $F_p$  is that of the random variable  $T$ , with  $T^2 \sim \text{Beta}(1/2, (p-1)/2)$ .

The weights  $w \equiv 1$  and  $w(u) = 1/(u(1-u))$  result in the Projected Cramér–von Mises statistic,  $P_{n,p}^{\text{CvM}}$ , and the Projected Anderson–Darling statistic,  $P_{n,p}^{\text{AD}}$ , respectively. The test based on  $P_{n,p}^{\text{CvM}}$  happens to be an extension of the Watson test to  $\mathbb{S}^{p-1}$ ,  $p \geq 2$ , since  $P_{n,2}^{\text{CvM}} = U_n^2/2$ . Moreover, the test based on  $P_{n,3}^{\text{CvM}}$  is equivalent to the chordal-based test on  $\mathbb{S}^2$  by Bakshaev (2010), whose statistic for  $p \geq 2$  is

$$N_{n,p} := n\mathbb{E}_{\mathcal{H}_0} [\|\mathbf{X}_1 - \mathbf{X}_2\|] - \frac{1}{n} \sum_{i,j=1}^n \|\mathbf{X}_i - \mathbf{X}_j\|.$$

The statistic  $P_{n,p}^{\text{AD}}$  represents the first instance of the Anderson–Darling statistic in the context of directional data. Particularly,  $P_{n,2}^{\text{AD}}$  can be regarded as the circular variant of  $A_n^2$ , just as  $U_n^2$  is the circular variant of  $W_n^2$ . Asymptotic distributions and computational formulae for  $P_{n,p}^{\text{CvM}}$  and  $P_{n,p}^{\text{AD}}$  are provided in García-Portugués et al. (2020), while the `sphunif` R package (García-Portugués and Verdebout, 2021) provides implementations for  $P_{n,p}^{\text{CvM}}$ ,  $P_{n,p}^{\text{AD}}$ , and  $N_{n,p}$ , for all  $p \geq 2$ .

### 3.2 Stabilization of projected-ecdf statistics

Let  $T_{n,p}$  be a statistic depending on  $p \in \mathbb{N}$ . From expression (4), the ratios  $T_{\infty,p;\alpha}/T_{n,p;\alpha}$  can be modeled as a function  $g : \mathbb{N} \times \mathbb{N} \times (0, 1) \rightarrow \mathbb{R}$  of  $(n, p, \alpha)$ . Hence, the modified version of the statistic  $T_{n,p}$  is defined as

$$T_{n,p}^*(\alpha) := T_{n,p} \cdot g(n, p, \alpha).$$

As in expression (5), the stabilization of  $T_{n,p}$  can be approached as a regression problem, now with predictors  $(n, p, \alpha)$ , response  $Y := T_{\infty,p;\alpha}/T_{n,p;\alpha}$ , and unknown regression function  $g$ .

The connection between  $P_{n,2}^{\text{CvM}}$  and  $U_n^2$  implies the stabilized form of  $P_{n,2}^{\text{CvM}}$  to have the same set of predictors based on  $(n, \alpha)$  as the Watson statistic already presented in Table 3:  $\mathcal{R} := \{1/n, 1/(n\sqrt{\alpha}), 1/(n\alpha)\}$ . An additional reflection suggests the adequacy of choosing  $\mathcal{R}$  for stabilizing  $P_{n,p}^{\text{CvM}}$ , also when  $p \geq 3$ , due to its appearance in all the transformations for quadratic-ecdf statistics in Table 3 and the quadratic nature of  $P_{n,p}^{\text{CvM}}$ . For different particular values of  $p \geq 2$ , it was noted that, if regression models were fitted to the ratios  $P_{\infty,p;\alpha}^{\text{CvM}}/P_{n,p;\alpha}^{\text{CvM}}$ , the coefficients fitted for each predictor  $r \in \mathcal{R}$  could be modeled as a smooth function of  $p$  denoted as  $q_r : \mathbb{N} \rightarrow \mathbb{R}$ . Unsurprisingly, given its similarity to  $P_{n,p}^{\text{CvM}}$ , the same considerations also hold for  $P_{n,p}^{\text{AD}}$ . Moreover, the statistic  $N_{n,p}$  can also be stabilized through  $\mathcal{R}$  and  $q_r$ , a fact explained by the closeness between  $P_{n,p}^{\text{CvM}}$  and  $N_{n,p}$  when  $p \neq 3$  and its equivalence when  $p = 3$ . Empirical investigations suggested the following saturated model for  $q_r$ , for each  $r \in \mathcal{R}$ :

$$q_r(p) = \frac{\beta_{r,1}}{\sqrt{p}} + \frac{\beta_{r,2}}{p}.$$

Thus, the resulting saturated model for  $g$  is set as

$$g(n, p, \alpha) = 1 + q_{1/n}(p) \cdot \frac{1}{n} + q_{1/(n\alpha)}(p) \cdot \frac{1}{n\alpha} + q_{1/(n\sqrt{\alpha})}(p) \cdot \frac{1}{n\sqrt{\alpha}}. \quad (8)$$

Once training samples of the form  $\{(n_j, \alpha_j, p_j, Y_j)\}_{j=1}^J$ ,  $Y_j := T_{\infty, p_j; \alpha_j} / T_{n_j, p_j; \alpha_j}$ , are available, model (8) is estimated following the same methodology described in Section 2.2. For each of the three test statistics  $P_{n,p}^{\text{CvM}}$ ,  $P_{n,p}^{\text{AD}}$ , and  $N_{n,p}$ , we obtained  $M = 10^7$  Monte Carlo random samples for each sample size  $n$  in  $\mathcal{N} := \{5, \dots, 100, 102, \dots, 200, 204, \dots, 300, 308, \dots, 404, 420, \dots, 500\}$  and for each dimension  $p$  in  $\mathcal{P} := \{2, \dots, 11, 21, 31, 41, 51, 61, 71, 81, 91, 101, 151, 201, 251, 301\}$ . We then summarized these statistics as the quantiles  $\{T_{n_j, p_j; \alpha_j} : n_j \in \mathcal{N}, p_j \in \mathcal{P}, \alpha_j \in \mathcal{A}\}$  for  $\mathcal{A} := \{a/A : a = 1, \dots, A\}$ ,  $A = 10^3$ . The asymptotic  $\alpha$ -quantiles  $T_{\infty, p; \alpha}$  were approximated through  $T_{500, p; \alpha}$  due to the accuracy limitations on inverting the asymptotic cdfs of the three statistics for large dimensions. Table 6 lists the approximated  $T_{\infty, p; \alpha}$  for the first ten dimensions.

| Critical value                       | $p$      |        |        |        |        |        |        |        |        |        |        |
|--------------------------------------|----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
|                                      | $\alpha$ | 2      | 3      | 4      | 5      | 6      | 7      | 8      | 9      | 10     | 11     |
| $P_{\infty, p; \alpha}^{\text{CvM}}$ | 0.10     | 0.3035 | 0.2768 | 0.2606 | 0.2500 | 0.2421 | 0.2361 | 0.2312 | 0.2272 | 0.2239 | 0.2210 |
|                                      | 0.05     | 0.3735 | 0.3288 | 0.3027 | 0.2858 | 0.2735 | 0.2641 | 0.2568 | 0.2508 | 0.2458 | 0.2416 |
|                                      | 0.01     | 0.5358 | 0.4461 | 0.3960 | 0.3638 | 0.3413 | 0.3244 | 0.3115 | 0.3008 | 0.2922 | 0.2849 |
| $P_{\infty, p; \alpha}^{\text{AD}}$  | 0.10     | 1.6871 | 1.5604 | 1.4816 | 1.4279 | 1.3883 | 1.3576 | 1.3327 | 1.3124 | 1.2957 | 1.2809 |
|                                      | 0.05     | 2.0293 | 1.8214 | 1.6951 | 1.6106 | 1.5494 | 1.5023 | 1.4651 | 1.4347 | 1.4092 | 1.3875 |
|                                      | 0.01     | 2.8197 | 2.4096 | 2.1679 | 2.0090 | 1.8969 | 1.8126 | 1.7471 | 1.6931 | 1.6493 | 1.6121 |
| $N_{\infty, p; \alpha}$              | 0.10     | 2.4034 | 2.2141 | 2.1003 | 2.0231 | 1.9673 | 1.9238 | 1.8887 | 1.8601 | 1.8367 | 1.8158 |
|                                      | 0.05     | 2.9906 | 2.6305 | 2.4320 | 2.3034 | 2.2119 | 2.1423 | 2.0879 | 2.0437 | 2.0067 | 1.9752 |
|                                      | 0.01     | 4.3495 | 3.5687 | 3.1669 | 2.9136 | 2.7402 | 2.6112 | 2.5124 | 2.4314 | 2.3661 | 2.3108 |

Table 6: Asymptotic critical values for modified uniformity statistics with dimension  $p$ , sample size  $n$ , and significance level  $\alpha$ .

| $T_{n,p}$              | $T_{n,p} \left( 1 + q_{1/n} \cdot \frac{1}{n} + q_{1/(n\alpha)} \cdot \frac{1}{n\alpha} + q_{1/(n\sqrt{\alpha})} \cdot \frac{1}{n\sqrt{\alpha}} \right)$ |                            |                                              |  |
|------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------|----------------------------------------------|--|
|                        | $q_{1/n}$                                                                                                                                                | $q_{1/(n\alpha)}$          | $q_{1/(n\sqrt{\alpha})}$                     |  |
| $P_{n,p}^{\text{CvM}}$ | $\frac{0.1130}{\sqrt{p}} - \frac{0.5415}{p}$                                                                                                             | $-\frac{0.0031}{\sqrt{p}}$ | $\frac{0.1438}{\sqrt{p}}$                    |  |
| $P_{n,p}^{\text{AD}}$  | $\frac{0.0978}{\sqrt{p}} - \frac{0.3596}{p}$                                                                                                             | $-\frac{0.0025}{\sqrt{p}}$ | $\frac{0.1126}{\sqrt{p}}$                    |  |
| $N_{n,p}$              | $\frac{0.1189}{\sqrt{p}} - \frac{0.5838}{p}$                                                                                                             | $-\frac{0.0030}{\sqrt{p}}$ | $\frac{0.1210}{\sqrt{p}} + \frac{0.0385}{p}$ |  |

Table 7: Modified uniformity statistics for dimension  $p$ , sample size  $n$ , and significance level  $\alpha$ . Modified forms are valid for  $2 \leq p \leq 300$ ,  $n \geq 5$ , and  $\alpha \leq 0.25$ .  $\mathcal{H}_0$  is rejected at significance level  $\alpha$  if  $T_{n,p}^*(\alpha) > T_{\infty, p; \alpha}$ , where  $T_{\infty, p; \alpha}$  is given in Table 6 for  $p = 2, \dots, 11$ .

The resulting modified forms for  $P_{n,p}^{\text{CvM}}$ ,  $P_{n,p}^{\text{AD}}$ , and  $N_{n,p}$  are presented in Table 7, where each fitted  $q_r$  is shown for each predictor  $r \in \mathcal{R}$ . An algorithm similar to Algorithm 1 for computing an approximated  $p$ -value has been implemented for these statistics, with the only difference being that the modified statistic function in lines 3 and 11 is the corresponding dimension-dependent version which also includes the parameter  $p$  as an input.

### 3.3 Simulation study

In the same manner as in Section 2.3, the empirical stabilization of the modified forms of the projected-ecdf statistics is investigated (Figure 3) in terms of the relative error between the significance level and the empirical rejection rate of the  $T_{n,p}^*(\alpha)$ -test for sample sizes  $n \in \mathcal{N}_{\text{test}}$  and dimensions  $p \in \mathcal{P}_{\text{test}}$ , where  $\mathcal{N}_{\text{test}}$  was defined in Section 2.3 and  $\mathcal{P}_{\text{test}} := \{2, \dots, 11, 21, 51, 101, 151, 201, 301\}$ . As for most non-heavily studied test statistics, Monte Carlo is the only method readily available to approximate the exact- $n$   $p$ -values of  $P_{n,p}^{\text{CvM}}$ ,  $P_{n,p}^{\text{AD}}$ , and  $N_{n,p}$ . Figure 3 shows an average improvement of our stabilizations' accuracy over Monte Carlo approximations (using  $10^4$  trials) of  $\times 3$ ,  $\times 4$ , and  $\times 4$ , for each of the three statistics, respectively. We point out the steadiness of our relative errors regardless of the significance level and the dimension  $p$  (except for  $\alpha = 0.01$ , which increases for large  $p$ 's), which on average are 1.3%, 0.9%, and 1% respectively. In almost all circumstances, our relative errors are largely below those obtained by Monte Carlo (except for  $\alpha = 0.25$  when  $p > 10$  in  $P_{n,p}^{\text{CvM}}$  and  $N_{n,p}$ ).

We conclude this section by summarizing in Table 8 a comparison of the modified forms found by Stephens (1970) and our results, for each of the classical ecdf-based statistics, and their corresponding versions for circular data, along with the circular particularizations of the projected-ecdf statistics. We emphasize the simplicity of the formulae in both approaches, with the Mean Relative Error (MRE) being reduced for the second by  $\times 2$  for  $D_n$  and  $U_n^2$ , by  $\times 9$  for  $W_n^2$ , and by  $\times 4$  for  $A_n^2$  and  $V_n$ . The stabilizations for the projected-ecdf statistics are such MRE  $< 0.9\%$  for the circular case, which aligns with the results specifically attained for  $U_n^2$  and  $P_{n,2}^{\text{AD}}$ , and supports the convenience of the extension proposed in this section for the  $(n, \alpha)$ -stabilization.

| $T_n$                               | Stephens' $T_n^*$                                                          | MRE   | $T_n^*(\alpha)$                                                                                                   | MRE   |
|-------------------------------------|----------------------------------------------------------------------------|-------|-------------------------------------------------------------------------------------------------------------------|-------|
| $D_n$                               | $D_n \left(1 + \frac{0.12}{\sqrt{n}} + \frac{0.11}{n}\right)$              | 1.44% | $D_n \left(1 + \frac{0.1575}{\sqrt{n}} + \frac{0.0192}{n\sqrt{\alpha}} - \frac{0.0051}{\sqrt{n\alpha}}\right)$    | 0.63% |
| $W_n^2$                             | $(W_n^2 - \frac{0.4}{n} + \frac{0.6}{n^2}) \left(1 + \frac{1}{n}\right)$   | 3.28% | $W_n^2 \left(1 - \frac{0.1651}{n} + \frac{0.0749}{n\sqrt{\alpha}} - \frac{0.0014}{n\alpha}\right)$                | 0.36% |
| $A_n^2$                             | $A_n^2$ (*)                                                                | 1.42% | $A_n^2 \left(1 + \frac{0.0360}{n} - \frac{0.0234}{n\sqrt{\alpha}} + \frac{0.0006}{n\alpha}\right)$                | 0.38% |
| $V_n$                               | $V_n \left(1 + \frac{0.155}{\sqrt{n}} + \frac{0.24}{n}\right)$             | 3.40% | $V_n \left(1 + \frac{0.2330}{\sqrt{n}} + \frac{0.0276}{n\sqrt{\alpha}} - \frac{0.0068}{\sqrt{n\alpha}}\right)$    | 0.85% |
| $U_n^2 \equiv P_{n,2}^{\text{CvM}}$ | $(U_n^2 - \frac{0.1}{n} + \frac{0.1}{n^2}) \left(1 + \frac{0.8}{n}\right)$ | 1.62% | $U_n^2 \left(1 - \frac{0.1505}{n} + \frac{0.0917}{n\sqrt{\alpha}} - \frac{0.0018}{n\alpha}\right)$                | 0.63% |
|                                     | –                                                                          | –     | $P_{n,2}^{\text{CvM}} \left(1 - \frac{0.1908}{n} + \frac{0.1017}{n\sqrt{\alpha}} - \frac{0.0022}{n\alpha}\right)$ | 0.88% |
| $P_{n,2}^{\text{AD}}$ (†)           | –                                                                          | –     | $P_{n,2}^{\text{AD}} \left(1 - \frac{0.0751}{n} + \frac{0.0692}{n\sqrt{\alpha}} - \frac{0.0014}{n\alpha}\right)$  | 0.74% |
|                                     | –                                                                          | –     | $P_{n,2}^{\text{AD}} \left(1 - \frac{0.1106}{n} + \frac{0.0796}{n\sqrt{\alpha}} - \frac{0.0018}{n\alpha}\right)$  | 0.83% |

Table 8: Modified forms of ecdf-based statistics for sample size  $n$  and significance level  $\alpha$ . MRE refers to the Mean Relative Error between the expected rejection proportion and the approximated proportion for each pair of  $(n, \alpha)$  with  $n \in \mathcal{N}_{\text{test}}$  and  $\alpha \in \{0.25, 0.2, 0.15, 0.1, 0.05, 0.02, 0.01\}$ . The  $T_n^*(\alpha)$  forms are valid for  $n \geq 5$  and  $\alpha \leq 0.25$ . (\*) Stephens (1974) states the best modification for Anderson–Darling statistic for  $n \geq 5$  is its asymptotic distribution. (†) Both the modified form estimated for  $p = 2$  (top row) and the  $(n, p, \alpha)$ -modification particularized for  $p = 2$  (bottom row) are given for  $P_{n,2}^{\text{CvM}}$  and  $P_{n,2}^{\text{AD}}$ .

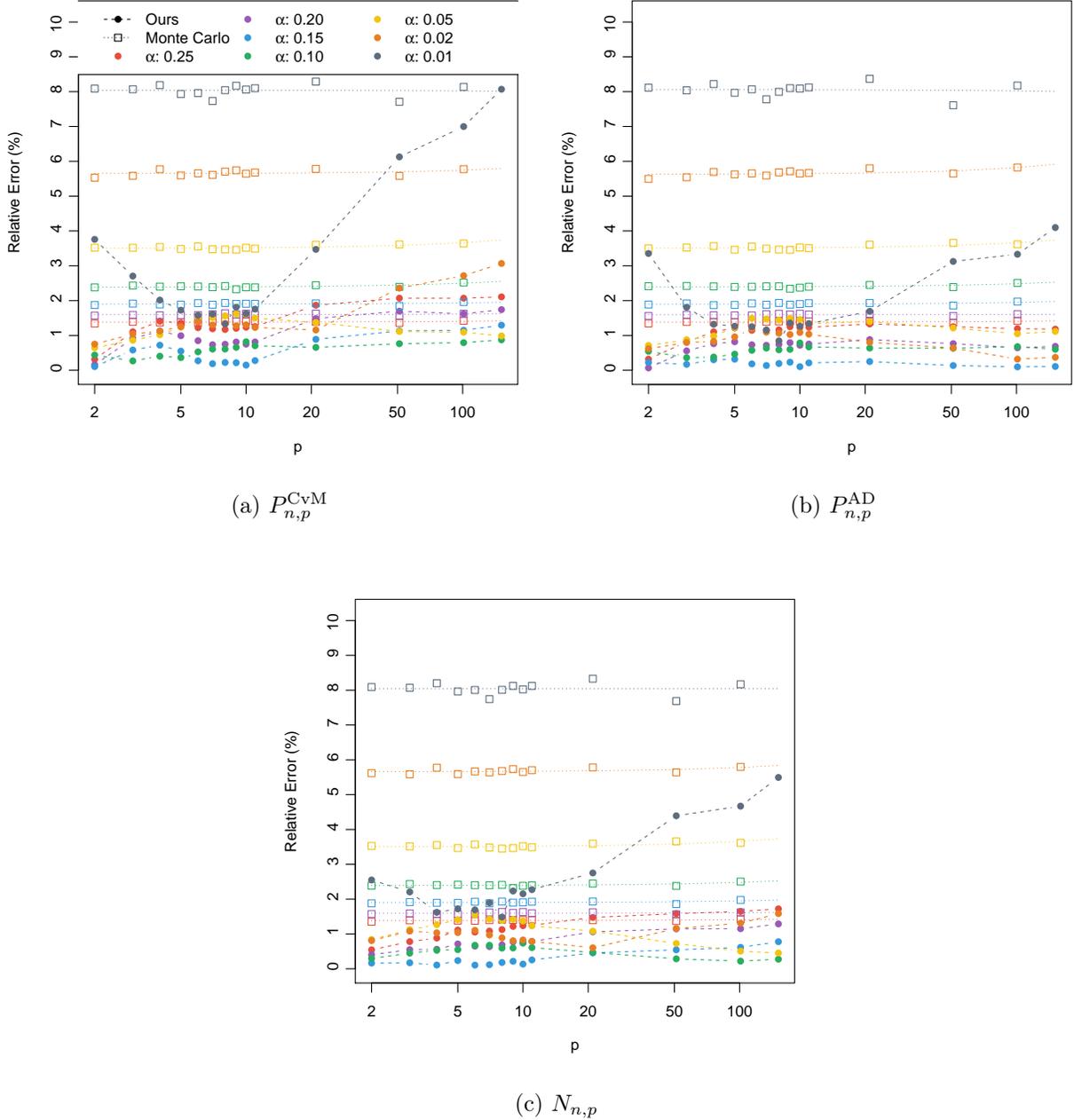


Figure 3: Relative error (in %)  $|\alpha - \tilde{\alpha}|/\alpha$  between the significance level  $\alpha$  and  $\tilde{\alpha}$ , the empirical rejection rate using an approximated exact- $n$  critical value, averaged over  $5 \leq n \leq 300$ . For the Monte Carlo approximation method, a regression fit is shown for each significance level  $\alpha$  to show no trend on the error with respect to  $p$ . The legend in Figure 3a details the approximation methods considered and significance levels, and applies to the rest of the panels.

## 4 Detecting temporal longitudinal non-uniformity in sunspots

The Sun’s magnetic field presents periodic behavioral patterns of about 11 years. During this period, the magnetic field is pulled around the Sun’s surface as the solar plasma rotates. Near the equator this pull is stronger due to the Sun’s differential rotation (equatorial latitudes rotate faster than poles), causing the field to wrap in a spiral-like shape until its polarity is eventually reversed and the wrapping restarts, indicating the beginning of a new solar cycle (see, e.g., Babcock, 1961). Sunspots

are created by high-intensity magnetic loops emerging from the Sun’s interior convection zone to the surface, producing darker, cooler regions on the Sun’s photosphere. Through their lifespans, which can last from hours to days, they experience continuous changes in shape, area, and location. The total number of active sunspots varies throughout the cycle, showing the maximum activity during the middle years (see Figure 4). Sunspots appear in a marked rotationally symmetric fashion: they are mainly distributed in latitudinal belts that are initially situated at  $\pm 30^\circ$  and that decay to the equator as the solar cycle advances (a phenomenon known as the *Spörer’s law*).

Sunspots also appear to cluster in *active longitudes*. Non-uniform patterns may appear by *preferred zones of occurrence* where sunspots had originated previously, as early described by Babcock (1961, pages 574 and 581). The existence of active longitudes was also suggested in Bogart (1982) upon inspection of the significant autocorrelation of daily sunspot numbers. Since daily sunspot numbers have no positional information, such analysis shows either there is one active longitude band or there are two active longitude bands separated by  $180^\circ$ , as Berdyugina and Usoskin (2003) concluded in both hemispheres, observing the alternation of major solar activity between both longitudes in about 1.5 to 3 years. This is known as the *flip-flop* phenomenon (Eltner and Korhonen, 2005).

Analyzing the presence of solar active longitudes requires knowledge from the Carrington period (or solar rotation period). It corresponds to the mean synodic rotation period of sunspots, which is about 27.275 days. Differential rotation causes the migration of active longitudes in the Carrington reference frame, causing a lag of 2.5 Carrington rotations per solar cycle that blurs the longitudinal pattern if more than one solar cycle is analyzed at once (Berdyugina and Usoskin, 2003). In order to ensure the adequate detection of active longitudes, a sequential analysis of data limited to a certain number of Carrington rotations, from 3–7 (Bogart, 1982; de Toma et al., 2000) to 10–15 (Pelt et al., 2010), is preferable.

The data we analyze is based on the Debrecen Photoheliographic Data (DPD) sunspot catalog (Baranyi et al., 2016; Györi et al., 2016). It contains observations of sunspots locations since 1974 and is a continuation of the Greenwich Photoheliographic Results (GPR) catalog, which spanned 1872–1976. The dataset `sunspots_births`, available in the R package `rotasym` (García-Portugués et al., 2021), accounts just for the first-ever observation (referred to as “birth” henceforth) of a group of sunspots.

In our analysis, summarized in Figure 4, we first applied the  $P_{n,2}^{\text{AD}}$ -based uniformity test sequentially to the longitudes of sunspot births —which include a total of 6195, 4551, and 5373 observations for the 21st, 22nd, and 23rd cycle, respectively— within a rolling window whose size is 10-Carrington rotations (approximately, nine months). The corresponding  $p$ -values were computed using Algorithm 1 for northern (blue), southern (red), and both (black) hemispheres. In addition, the  $p$ -value was also computed by Monte Carlo with  $5 \times 10^3$  samples, in order to compare the running times between the two methods. Our method runs in an average of 1.6 s per solar cycle, while Monte Carlo completes it in 1600 s per solar cycle. In order to account for dependency between sequential tests, Benjamini and Yekutieli (2001)’s FDR correction was applied to the  $p$ -values obtained with the test based on  $P_{n,2}^{\text{AD}}$ . These corrected  $p$ -values are shown in the top row of Figure 4. Second, circular-linear kernel density estimation (García-Portugués et al., 2013) of sunspot births for northern (middle-top figure) and southern (middle-bottom) hemispheres allowed us to compute several level sets, represented as contour lines labeled as “100 $p$ %”. Each of these sets is the smallest set containing  $1 - p$  of the probability of the estimated density function. Hence, darker sets represent higher-density zones of sunspot births, both through time and longitude. Third, a scatter plot of sunspot births is shown in the bottom figures, along with the circular Nadaraya–Watson (Di Marzio et al., 2012) regression for northern (blue), southern (red), and both (black) hemispheres. The Nadaraya–Watson regression gives a moving circular mean of the longitudes of sunspot births through time. Both density and regression kernel estimates use “rule-of-thumb” bandwidths for normal (Silverman, 1986) and von Mises–Fisher (García-Portugués, 2013) distributions, given the similarity of marginal distributions with these respective distributions and the marked undersmoothing that resulted from

cross-validation bandwidths.

We draw the following conclusions from the analysis:

- (i) In general, the two hemispheres seem to have different behavioral patterns, both in terms of longitudinal non-uniformities and sunspots activity level, along solar cycles. During the 21st cycle, the northern hemisphere presents 33% of the tests rejected at significance level  $\alpha = 0.05$ . In cycles 22 and 23, the southern hemisphere presents more non-uniform periods (9% and 10% of the tests are rejected for  $\alpha = 0.05$ , respectively) than the northern hemisphere (5% and 3% are rejected, respectively).
- (ii) Non-uniformity periods are intermittent during the lifetime of the solar cycle, without a clear association with the intensity of the sunspots appearance. The length and quantity of non-uniformity periods differ between solar cycles.
- (iii) Sunspots seem to appear in preferred zones of occurrence. Highest density sets, together with Nadaraya–Watson regressions, show consistent patterns of activity within certain longitudinal zones. In particular, periods in which uniformity is rejected at significance level  $\alpha = 0.05$ , the northern sunspot births seem to cluster around  $0^\circ$  (1982, 1990, 2000),  $135^\circ$  (1983–1984), and  $180^\circ$  (1977–1978, 1979–1980), while the southern hemisphere sunspot births cluster around  $-135^\circ$  (1991, 2004, 2008). However, non-uniformity periods are too few to claim the existence of active longitudes.
- (iv) The flip-flop phenomenon between  $180^\circ$  active longitudes is not obvious throughout all the cycles. Although longitudes  $0$  and  $180^\circ$  seem to accumulate more sunspots in the northern hemisphere, the alternation between supplementary longitudes is not a clear, fixed-duration pattern.

## 5 Discussion

We have presented a general, automated approach to construct simple yet effective approximations for the upper tail of the exact- $n$  null distribution of numerous goodness-of-fit statistics. The simulation results demonstrate that these approximations are accurate enough for practical applications of several upper-tail tests, even when these depend on a varying (yet known) parameter.

Although state-of-the-art statistic-specific algorithms like those of Marsaglia et al. (2003), Csörgö and Faraway (1996), and Marsaglia and Marsaglia (2004) provide arbitrarily accurate upper-tail  $p$ -values for the  $D_n$ ,  $W_n^2$ , and  $A_n^2$  statistics, respectively, our  $p$ -value approximation method offers significant computational improvements, has a reasonable precision (mean relative errors below 1%), and, most importantly, can be applied to a wide range of statistics. Compared to the general and omnipresent  $p$ -value approximation by  $M$  Monte Carlo trials, our method presents two key advantages: (i) more accurate results (at least, when  $M = 10^4$ ); and (ii) faster running times by several orders of magnitude. This computational expediency makes the stabilized statistic especially convenient for sequential tests, as illustrated in the data application.

The  $(n, \alpha)$ -stabilization significantly extends the scope of applicability of stabilizations like those of Stephens (1970). The stabilization focuses only on the upper tail of  $T_n$ , as this is usually the most useful in practice. However, stabilizations for the lower tail can be analogously derived. Obtaining modifications that stabilize the whole distribution, while still retaining simplicity, would offer the advantage of having approximated  $p$ -values that are roughly uniformly distributed under the null hypothesis. This task is left for future research.

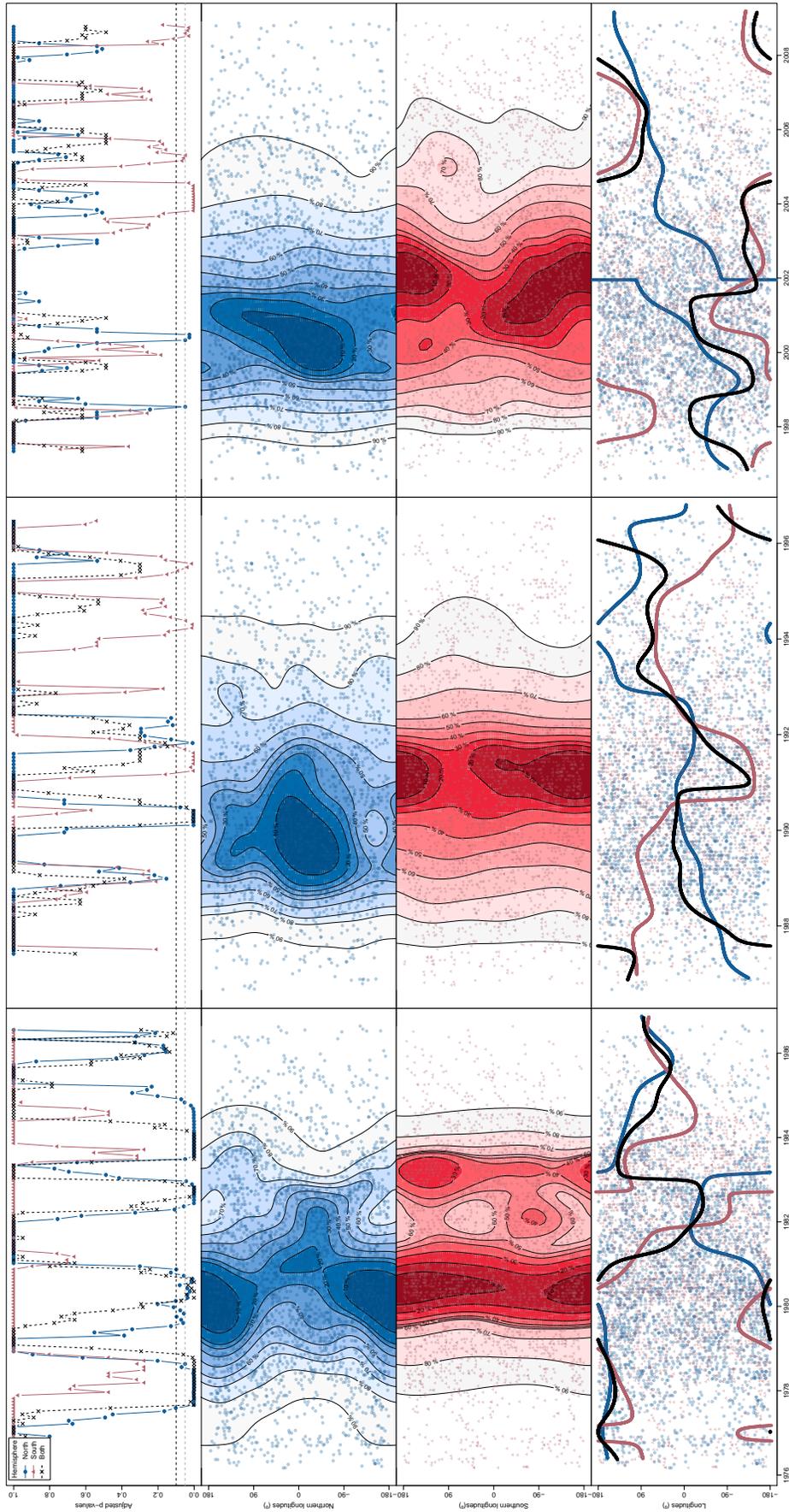


Figure 4: Longitudinal non-uniformity patterns of sunspot births. Each column represents the analysis for each of the 21st, 22nd, and 23rd solar cycles. Northern (blue), southern (red), and both (black) hemispheres were separately analyzed. Top figures:  $P_{n,2}^{\text{AD}}$ -based uniformity test of sunspot births longitudes. The  $p$ -values shown are corrected by Benjamini and Yekutieli (2001)'s FDR. Middle figures: Circular-linear kernel density level sets of sunspot births through time and longitude. Bottom figures: Sunspot births (points) along with the corresponding Nadaraya–Watson regression (lines).

## Acknowledgments

The second author acknowledges support by grants PGC2018-097284-B-100 and IJCI-2017-32005 by Spain's Ministry of Science, Innovation and Universities. The two grants were co-funded with ERDF funds. The computational resources of the Supercomputing Center of Galicia (CESGA) are greatly appreciated. The authors greatly acknowledge the comments of three referees.

## References

- Agostinelli, C. and Lund, U. (2017). *R package circular: Circular Statistics*. R package version 0.4-93.
- Arsham, H. (1988). Kuiper's P-value as a measuring tool and decision procedure for the goodness-of-fit test. *J. Appl. Stat.*, 15(2):131–135.
- Babcock, H. W. (1961). The topology of the Sun's magnetic field and the 22-year cycle. *Astrophys. J.*, 133(2):572–587.
- Bakshaev, A. (2010).  $N$ -distance tests of uniformity on the hypersphere. *Nonlinear Anal. Model. Control.*, 15(1):15–8.
- Baranyi, T., Györi, L., and Ludmány, A. (2016). On-line tools for solar data compiled at the Debrecen observatory and their extensions with the Greenwich sunspot data. *Sol. Phys.*, 291(9):3081–3102.
- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, 29(4):1165–1188.
- Berdyugina, S. V. and Usoskin, I. G. (2003). Active longitudes in sunspot activity: Century scale persistence. *Astron. Astrophys.*, 405(3):1121–1128.
- Birnbaum, Z. W. (1952). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. *J. Am. Stat. Assoc.*, 47(259):425–441.
- Bogart, R. S. (1982). Recurrence of solar activity: Evidence for active longitudes. *Solar Phys.*, 76(1):155–165.
- Brown, J. R. and Harvey, M. E. (2007). Rational arithmetic mathematica functions to evaluate the one-sided one sample K–S cumulative sampling distribution. *J. Stat. Softw.*, 19(6):1–32.
- Crown, J. S. (2000). Percentage points for directional Anderson–Darling goodness-of-fit tests. *Commun. Stat. Simul. Comput.*, 29(2):523–532.
- Csörgö, S. and Faraway, J. J. (1996). The exact and asymptotic distributions of Cramér-von Mises statistics. *J. R. Stat. Soc. Ser. B Methodol.*, 58(1):221–234.
- Cuesta-Albertos, J. A., Cuevas, A., and Fraiman, R. (2009). On projection-based tests for directional and compositional data. *Stat. Comput.*, 19(4):367–380.
- D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-Fit Techniques*, volume 68 of *Statistics: Textbooks and Monographs*. Marcel Dekker, New York.
- Di Marzio, M., Panzera, A., and Taylor, C. C. (2012). Smooth estimation of circular cumulative distribution functions and quantiles. *J. Nonparametr. Stat.*, 24(4):935–949.
- Dufour, R. and Maag, U. R. (1978). Distribution results for modified Kolmogorov–Smirnov statistics for truncated or censored. *Technometrics*, 20(1):29–32.

- Durbin, J. (1969). Tests for serial correlation in regression analysis based on the periodogram of least-squares residuals. *Biometrika*, 56(1):1–15.
- Durbin, J. (1973). *Distribution Theory for Tests Based on the Sample Distribution Function*, volume 9 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia.
- Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *J. Stat. Softw.*, 40(8):1–18.
- Elstner, D. and Korhonen, H. (2005). Flip-flop phenomenon: observations and theory. *Astron. Nachr.*, 326(3-4):278–282.
- Facchinetti, S. (2009). A procedure to find exact critical values of Kolmogorov–Smirnov test. *Stat. App.*, 21(3–4):337–359.
- Faraway, J., Marsaglia, G., Marsaglia, J., and Baddeley, A. (2019). *goftest: Classical Goodness-of-Fit Tests for Univariate Distributions*. R package version 1.2-2.
- García-Portugués, E. (2013). Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. *Electron. J. Stat.*, 7:1655–1685.
- García-Portugués, E., Crujeiras, R. M., and González-Manteiga, W. (2013). Kernel density estimation for directional-linear data. *J. Multivar. Anal.*, 121:152–175.
- García-Portugués, E., Navarro-Esteban, P., and Cuesta-Albertos, J. A. (2020). On a projection-based class of uniformity tests on the hypersphere. *arXiv:2008.09897*.
- García-Portugués, E., Paindaveine, D., and Verdebout, T. (2021). *rotasym: Tests for Rotational Symmetry on the Hypersphere*. R package version 1.1.3.
- García-Portugués, E. and Verdebout, T. (2018). A review of uniformity tests on the hypersphere. *arXiv:1804.00286*.
- García-Portugués, E. and Verdebout, T. (2021). *sphunif: Uniformity Tests on the Circle, Sphere, and Hypersphere*. R package version 1.0.1.
- Györi, L., Baranyi, T., and Ludámny, A. (2016). Comparative analysis of Debrecen sunspot catalogues. *Mon. Not. R. Astron. Soc.*, 465(2):1259–1273.
- Hegazy, Y. A. S. and Green, J. R. (1975). Some new goodness-of-fit tests using order statistics. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 24(3):299–308.
- Heo, J.-H., Shin, H., Nam, W., Om, J., and Jeong, C. (2013). Approximation of modified Anderson–Darling test statistics for extreme value distributions with unknown shape parameter. *J. Hydrol.*, 499:41–49.
- Johannes, J. M. and Rasche, R. H. (1980). Additional information on significance values for Durbin’s  $c^+$ ,  $c^-$  and  $c$  statistics. *Biometrika*, 67(2):511–514.
- Jupp, P. E. and Kume, A. (2020). Measures of goodness of fit obtained by almost-canonical transformations on Riemannian manifolds. *J. Multivar. Anal.*, 176:104579.
- Knott, M. (1974). The distribution of the Cramér–von Mises statistic for small sample sizes. *J. R. Stat. Soc. Ser. B Methodol.*, 36(3):430–438.
- Kuiper, N. H. (1960). Tests concerning random points on the circle. *K. Ned. Akad. Van Wet. A*, 63:38–47.

- Lewis, P. A. W. (1961). Distribution of the Anderson–Darling statistic. *Ann. Math. Stat.*, 32(4):1118–1124.
- Maag, U. R. and Dicaire, G. (1971). On Kolmogorov–Smirnov type one-sample statistics. *Biometrika*, 58(3):653–656.
- Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley, Chichester.
- Marks, N. B. (1998). Modification of the Kolmogorov–Smirnov test for the Erlang-2 distribution. *Commun. Stat. Simul. Comput.*, 27(1):39–49.
- Marks, N. B. (2007). Kolmogorov–Smirnov test statistic and critical values for the Erlang-3 and Erlang-4 distributions. *J. Appl. Stat.*, 34(8):899–906.
- Marsaglia, G. and Marsaglia, J. (2004). Evaluating the Anderson–Darling distribution. *J. Stat. Softw.*, 9(2):1–5.
- Marsaglia, G., Tsang, W. W., and Wang, J. (2003). Evaluating Kolmogorov’s distribution. *J. Stat. Softw.*, 8(18):1–4.
- Marshall, A. W. (1958). The small sample distribution of  $n\omega_n^2$ . *Ann. Math. Stat.*, 29(1):307–309.
- Massey, F. J. (1950). A note on the estimation of a distribution function by confidence limits. *Ann. Stat.*, 21(1):116–119.
- Massey, F. J. (1951). The Kolmogorov–Smirnov test for goodness of fit. *J. Am. Stat. Assoc.*, 46(253):68–78.
- Mersmann, O. (2019). *microbenchmark: Accurate Timing Functions*. R package version 1.4-7.
- Millard, S. P. (2013). *EnvStats*. Springer, New York.
- Pearson, E. S. and Stephens, M. A. (1962). The goodness-of-fit tests based on  $W_N^2$  and  $U_N^2$ . *Biometrika*, 49(3/4):397–402.
- Pelt, J., Korpi, M. J., and Tuominen, I. (2010). Solar active regions: A nonparametric statistical analysis. *Astron. Astrophys.*, 513:A48.
- Pettitt, A. N. (1977). Testing the normality of several independent samples using the Anderson–Darling statistic. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 26(2):156–161.
- Pewsey, A. and García-Portugués, E. (2021). Recent advances in directional statistics. *Test*, 30(1):1–58.
- Quesenberry, C. P. and Miller Jr, F. L. (1977). Power studies of some tests for uniformity. *J. Stat. Comput. Simul.*, 5(3):169–191.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
- Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. Wiley, New York.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.
- Stephens, M. A. (1965). The goodness-of-fit statistic  $V_n$ : distribution and significance points. *Biometrika*, 52(3/4):309–321.

- Stephens, M. A. (1970). Use of the Kolmogorov–Smirnov, Cramér-von Mises and related statistics without extensive tables. *J. R. Stat. Soc. Ser. B Methodol.*, 32(1):115–122.
- Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *J. Am. Stat. Assoc.*, 69(347):730–737.
- Stephens, M. A. (1977). Goodness of fit for the extreme value distribution. *Biometrika*, 64(3):583–588.
- Stephens, M. A. (1979). Tests of fit for the logistic distribution based on the empirical distribution function. *Biometrika*, 66(3):591–595.
- Stephens, M. A. and Maag, U. R. (1968). Further percentage points for  $W_N^2$ . *Biometrika*, 55(2):428–430.
- Tiku, M. L. (1965). Chi-square approximations for the distributions of goodness-of-fit statistics  $U_n^2$  and  $W_n^2$ . *Biometrika*, 52(3/4):630–633.
- de Toma, G., White, O. R., and Harvey, K. L. (2000). A picture of solar minimum and the onset of solar cycle 23. I. Global magnetic field evolution. *Astrophys. J.*, 529(2):1101.
- Watson, G. S. (1961). Goodness-of-fit tests on a circle. *Biometrika*, 48(1/2):109–114.

# Supplementary material for “Data-driven stabilizations of goodness-of-fit tests”

Alberto Fernández-de-Marcos<sup>1,2</sup> and Eduardo García-Portugués<sup>1</sup>

## Abstract

This supplementary material is divided into two sections. Appendix A contains an analysis of regression metrics for different polynomial forms of (6) in Section 2.2. Appendix B presents an analysis of different weight functions for the regression in Section 2.2.

## A Selection of $(n, \alpha)$ -model form

In order to choose and justify the final form of model (6), an analysis of different models is presented in Tables 9 and 10. The explored model specifications are

$$g_{\lambda,\mu}(n, \alpha) := 1 + \{\beta_{l,m} n^{-l/2} \alpha^{-m/2}\}_{l=1, m=0}^{\lambda,\mu}, \quad (9)$$

where  $\lambda, \mu \in \mathbb{N}$  are to be tuned. With this notation, the model considered in (6) in Section 2.2 is  $g_{2,2}$ .

Table 9 shows a comparative study of the performance of  $g_{\lambda,2}(n, \alpha)$  for  $\lambda = 2, \dots, 5$ , in order to determine the optimal power of  $n$  predictors, while keeping the  $\alpha$ -predictors unchanged. Conversely, Table 10 shows the performance of models  $g_{2,\mu}(n, \alpha)$  where  $\mu = 2, \dots, 5$  to analyze the effect of  $\alpha$ , while not varying the  $n$ -predictors. In both tables the same main model-selection procedure applied in Section 2.2 is carried out: forward-backward model selection is run from Stephens’ set of predictors, using weighted least squares with weights  $w_j = n_j^{-1/2} 1_{\{0 < \alpha_j \leq 0.25\}}$ , and with extended scope given by (9) (interactions are included). The last dropping step is excluded from the analysis.

Table 9 shows that, when increasing  $\lambda$  the BIC decreases marginally. The standard deviation of the residuals  $\hat{\sigma}$  only decreases in the sixth decimal, while the  $R_{\text{adj}}^2$  remains almost constant for  $D_n$ , and  $W_n^2$ . Moreover, the multicollinearity present in the model increases by an order of magnitude per unit increment in  $\lambda$ , with high values of MVIF. In the case of  $A_n^2$ , the BIC-optimal model equals that with  $\lambda = 2$  — including more powers of  $n$  does not improve it.

Regarding the influence of the powers of  $\alpha$ , Table 10 shows similar patterns to those in Table 9 for the three statistics  $D_n$ ,  $W_n^2$ , and  $A_n^2$ . For  $D_n$  and  $W_n^2$ , when including more powers of  $\alpha$ , the BIC and  $\hat{\sigma}$  decrease marginally, and  $R_{\text{adj}}^2$  increases marginally. In exchange, the multicollinearity increases by an order of magnitude per unit increment in  $\mu$ , also attaining high values of MVIF. For  $A_n^2$  the situation is somehow different: from  $\mu = 2$  to  $\mu = 4$  there is a significant increase in  $R_{\text{adj}}^2$ , yet at expenses of a  $\times 100$  increase in MVIF and a more convoluted model.

From the analyzed test statistics, the final form of the saturated model is chosen to be  $g_{2,2}(n, \alpha)$  due to the general small increase in the goodness-of-fit metrics for more complex models and the increment in multicollinearity when increasing  $\lambda$  and  $\mu$ . Importantly, the choice of  $g_{2,2}(n, \alpha)$  allows having a single generic approach for any statistic and provides parsimonious stabilizing transformations.

---

<sup>1</sup>Department of Statistics, Carlos III University of Madrid (Spain).

<sup>2</sup>Corresponding author. e-mail: albertfe@est-econ.uc3m.es.

| $T_n$   | $\lambda$ | $\hat{\sigma}$       | BIC     | $R_{\text{adj}}^2$ | MVIF           | $\text{pred}_{\text{MVIF}}$ |
|---------|-----------|----------------------|---------|--------------------|----------------|-----------------------------|
| $D_n$   | 2         | $7.57 \cdot 10^{-4}$ | -397236 | 0.9993             | $1 \cdot 10^2$ | $n^{-1/2}\alpha^{-1/2}$     |
|         | 3         | $7.50 \cdot 10^{-4}$ | -398020 | 0.9993             | $4 \cdot 10^3$ | $n^{-1}\alpha^{-1/2}$       |
|         | 4         | $7.48 \cdot 10^{-4}$ | -398184 | 0.9993             | $1 \cdot 10^5$ | $n^{-3/2}\alpha^{-1/2}$     |
|         | 5         | $7.48 \cdot 10^{-4}$ | -398212 | 0.9993             | $3 \cdot 10^5$ | $n^{-2}$                    |
| $W_n^2$ | 2         | $9.06 \cdot 10^{-4}$ | -369813 | 0.9835             | $1 \cdot 10^2$ | $n^{-1/2}\alpha^{-1/2}$     |
|         | 3         | $8.84 \cdot 10^{-4}$ | -371092 | 0.9841             | $8 \cdot 10^2$ | $n^{-1}\alpha^{-1/2}$       |
|         | 4         | $8.83 \cdot 10^{-4}$ | -371112 | 0.9841             | $7 \cdot 10^3$ | $n^{-3/2}$                  |
|         | 5         | $8.83 \cdot 10^{-4}$ | -371119 | 0.9841             | $2 \cdot 10^5$ | $n^{-2}$                    |
| $A_n^2$ | 2         | $8.11 \cdot 10^{-4}$ | -376072 | 0.8722             | $2 \cdot 10^1$ | $n^{-1}\alpha^{-1/2}$       |
|         | 3         | $8.11 \cdot 10^{-4}$ | -376072 | 0.8722             | $2 \cdot 10^1$ | $n^{-1}\alpha^{-1/2}$       |
|         | 4         | $8.11 \cdot 10^{-4}$ | -376072 | 0.8722             | $2 \cdot 10^1$ | $n^{-1}\alpha^{-1/2}$       |
|         | 5         | $8.11 \cdot 10^{-4}$ | -376072 | 0.8722             | $2 \cdot 10^1$ | $n^{-1}\alpha^{-1/2}$       |

Table 9: BIC-optimal  $g_{\lambda,2}(n, \alpha)$  for statistics  $D_n$ ,  $W_n^2$ , and  $A_n^2$ , obtained from weighted least squares and forward-backward model selection with scope (9) and  $\lambda = 2, \dots, 5$ . The following goodness-of-fit measures are presented for each selected model: standard deviation  $\hat{\sigma}$  of the residuals of upper-tail observations (i.e.,  $\{\hat{\varepsilon}_j \mid \alpha_j \leq 0.25\}_{j=1}^J$ ), BIC, and  $R_{\text{adj}}^2$ . In addition, the MVIF and  $\text{pred}_{\text{MVIF}}$ , the predictor that takes the maximum variance inflation factor, inform on the multicollinearity of the selected model.

| $T_n$   | $\mu$ | $\hat{\sigma}$       | BIC     | $R_{\text{adj}}^2$ | MVIF           | $\text{pred}_{\text{MVIF}}$ |
|---------|-------|----------------------|---------|--------------------|----------------|-----------------------------|
| $D_n$   | 2     | $7.57 \cdot 10^{-4}$ | -397236 | 0.9993             | $1 \cdot 10^2$ | $n^{-1/2}\alpha^{-1/2}$     |
|         | 3     | $5.22 \cdot 10^{-4}$ | -421343 | 0.9996             | $2 \cdot 10^3$ | $n^{-1/2}\alpha^{-1}$       |
|         | 4     | $3.68 \cdot 10^{-4}$ | -444816 | 0.9998             | $5 \cdot 10^4$ | $n^{-1/2}\alpha^{-3/2}$     |
|         | 5     | $2.83 \cdot 10^{-4}$ | -462192 | 0.9999             | $2 \cdot 10^6$ | $n^{-1}\alpha^{-2}$         |
| $W_n^2$ | 2     | $9.06 \cdot 10^{-4}$ | -369813 | 0.9835             | $1 \cdot 10^2$ | $n^{-1/2}\alpha^{-1/2}$     |
|         | 3     | $5.85 \cdot 10^{-4}$ | -411594 | 0.9949             | $2 \cdot 10^3$ | $n^{-1/2}\alpha^{-1}$       |
|         | 4     | $5.13 \cdot 10^{-4}$ | -427884 | 0.9968             | $5 \cdot 10^4$ | $n^{-1}\alpha^{-3/2}$       |
|         | 5     | $4.94 \cdot 10^{-4}$ | -433359 | 0.9972             | $4 \cdot 10^5$ | $n^{-1}\alpha^{-2}$         |
| $A_n^2$ | 2     | $8.11 \cdot 10^{-4}$ | -376072 | 0.8722             | $2 \cdot 10^1$ | $n^{-1}\alpha^{-1/2}$       |
|         | 3     | $5.99 \cdot 10^{-4}$ | -403690 | 0.9414             | $2 \cdot 10^3$ | $n^{-1/2}\alpha^{-1}$       |
|         | 4     | $4.67 \cdot 10^{-4}$ | -430707 | 0.9726             | $7 \cdot 10^3$ | $n^{-1}\alpha^{-3/2}$       |
|         | 5     | $4.15 \cdot 10^{-4}$ | -447192 | 0.9828             | $3 \cdot 10^5$ | $n^{-1}\alpha^{-2}$         |

Table 10: BIC-optimal  $g_{2,\mu}(n, \alpha)$  for statistics  $D_n$ ,  $W_n^2$ , and  $A_n^2$ , obtained from weighted least squares and forward-backward model selection with scope (9) and  $\mu = 2, \dots, 5$ . The table entries are analogous to those of Table 9.

## B Selection of the weight function

Model (6) is estimated through weighted least squares using the samples  $\{(n_j, \alpha_j, Y_j)\}_{j=1}^J$ ,  $Y_j := T_{\infty; \alpha_j} / T_{n_j; \alpha_j}$ . The weights considered in Section 2.2 are  $w_j := n_j^{-1/2} 1_{\{0 < \alpha_j \leq 0.25\}}$ . The term  $n_j^{-1/2}$  gives heavier weight to the approximation error on lower sample sizes, while the indicator  $1_{\{0 < \alpha_j \leq 0.25\}}$  reflects the interest in prioritizing the stabilization of the upper tail of the test statistic  $T_n$ , where accuracy on the determination of exact- $n$  quantiles is crucial for a precise test decision in the standard significance levels.

The effect of adding other factors to the weights is investigated in this section. First, instead of considering a hard thresholding by  $1_{\{0 < \alpha_j \leq 0.25\}}$ , the factor  $\alpha_j^{-1/2}$  can be introduced to place more weight on the upper quantiles while still incorporating the remaining quantiles. Second, in order to

mitigate the heteroscedasticity of the observations, the asymptotic variance of  $Y_j$ ,

$$\text{AVar}[Y_j] = \frac{T_{\infty; \alpha_j}^2 \alpha_j (1 - \alpha_j)}{M \cdot T_{n_j; \alpha_j}^4 \cdot (f_n(T_{n_j; \alpha_j}))^2}, \quad (10)$$

where  $M$  is the number of Monte Carlo samples and  $f_n$  is the density of  $T_n$ , can be incorporated to give more weight to ratios with smaller variances. Expression (10) is obtained with the delta method from the standard errors of sample quantiles (see, e.g., Serfling, 1980, Section 2.3.3). The evaluation of  $f_n$  can be done by differentiating a monotone spline interpolation of its cdf based on the saved quantiles  $\{T_{n_j; \alpha_j} : n_j \in \mathcal{N}, \alpha_j \in \mathcal{A}\}$  (see Section 2.2).

Different combinations of the previous factors shape the following candidates for weight functions:

- $w_{1,j}(\alpha_j, Y_j) := 1_{\{0 < \alpha_j \leq 0.25\}}$ .
- $w_{2,j}(n_j, \alpha_j) := n_j^{-1/2} 1_{\{0 < \alpha_j \leq 0.25\}}$ .
- $w_{3,j}(\alpha_j, Y_j) := \text{AVar}^{-1/2}[Y_j] 1_{\{0 < \alpha_j \leq 0.25\}}$ .
- $w_{4,j}(n_j, \alpha_j, Y_j) := n_j^{-1/2} \text{AVar}^{-1/2}[Y_j] 1_{\{0 < \alpha_j \leq 0.25\}}$ .
- $w_{5,j}(\alpha_j, Y_j) := \text{AVar}^{-1/2}[Y_j]$ .
- $w_{6,j}(n_j, \alpha_j, Y_j) := n_j^{-1/2} \text{AVar}^{-1/2}[Y_j]$ .
- $w_{7,j}(n_j, \alpha_j) := (n_j \cdot \alpha_j)^{-1/2}$ .

The following analysis compares the results of the weighted least squares regression on model (6) with weights computed by  $w_{k,j}$ ,  $k = 1, \dots, 7$ , according to the methodology described in Section 2.2, except for the last dropping step. The standard deviation of the residuals for the statistics  $D_n$ ,  $W_n^2$ , and  $A_n^2$  is presented from two perspectives: Table 11 shows the residuals based on the  $\alpha_j$  value, while in Table 12 the residuals are divided according to the sample size  $n_j$ . In addition, Table 13 shows the residuals depending on  $n_j$  only for upper-tail observations.

First, we analyze weights  $w_1$ ,  $w_2$ ,  $w_3$ , and  $w_4$ , all of which have the factor  $1_{\{0 < \alpha_j \leq 0.25\}}$  in common. As expected, they present the lowest errors for the upper tail  $\alpha \leq 0.25$ . Despite presenting higher residual deviation for  $\alpha > 0.25$ , the errors differ only by about  $\times 2$  with respect to  $w_5$  and  $w_6$  in average (Table 11). More importantly,  $w_1$  and  $w_2$  produce the smallest residuals for all sample sizes in the upper tail,  $w_2$  exhibiting slightly smaller values for small sample sizes,  $n \in [5, 10]$  (Table 13).

Second,  $w_5$  and  $w_6$  factor in the asymptotic variance, showing both similar results. In particular, they attain the lowest errors for  $\alpha > 0.25$  (Table 11) and small-to-moderate sample sizes (Table 12). However, the standard deviation of upper-tail residuals is one order of magnitude higher than  $w_1$ ,  $w_2$ ,  $w_3$ , and  $w_4$  (Table 11).

Finally,  $w_7$  weights observations by  $\alpha_j^{-1/2}$ . Its behavior is similar to  $w_5$  and  $w_6$  for  $\alpha > 0.25$  and all sample sizes. Yet, errors in the upper tail are lower, but still about  $\times 3$  higher than  $w_1$  and  $w_2$ .

From the previous observations, the final weight function chosen to fit model (6) is  $w_2$  due to two main reasons: (i) the significant difference of errors for  $\alpha \leq 0.25$  against  $w_5$ ,  $w_6$ , and  $w_7$ ; and (ii) the lower residual deviation in the upper tail for small sample sizes when compared to  $w_1$ ,  $w_3$ , and  $w_4$ .

## References

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. New York: Wiley.

| $w_k$ | $T_n$   | Standard deviation                     |                                        |                                        |                                        |
|-------|---------|----------------------------------------|----------------------------------------|----------------------------------------|----------------------------------------|
|       |         | $\alpha \in (0, 0.25]$                 | $\alpha \in (0.25, 0.5]$               | $\alpha \in (0.5, 0.75]$               | $\alpha \in (0.75, 1)$                 |
| $w_1$ | $D_n$   | <b><math>7.55 \cdot 10^{-4}</math></b> | $1.66 \cdot 10^{-3}$                   | $3.11 \cdot 10^{-3}$                   | $7.24 \cdot 10^{-3}$                   |
|       | $W_n^2$ | <b><math>9.03 \cdot 10^{-4}</math></b> | $1.90 \cdot 10^{-3}$                   | $4.08 \cdot 10^{-3}$                   | $1.01 \cdot 10^{-2}$                   |
|       | $A_n^2$ | <b><math>8.11 \cdot 10^{-4}</math></b> | $1.50 \cdot 10^{-3}$                   | $2.13 \cdot 10^{-3}$                   | $2.76 \cdot 10^{-3}$                   |
| $w_2$ | $D_n$   | $7.57 \cdot 10^{-4}$                   | $1.67 \cdot 10^{-3}$                   | $3.13 \cdot 10^{-3}$                   | $7.25 \cdot 10^{-3}$                   |
|       | $W_n^2$ | $9.06 \cdot 10^{-4}$                   | $1.90 \cdot 10^{-3}$                   | $4.07 \cdot 10^{-3}$                   | $1.01 \cdot 10^{-2}$                   |
|       | $A_n^2$ | <b><math>8.11 \cdot 10^{-4}</math></b> | $1.50 \cdot 10^{-3}$                   | $2.13 \cdot 10^{-3}$                   | $2.76 \cdot 10^{-3}$                   |
| $w_3$ | $D_n$   | $8.03 \cdot 10^{-4}$                   | $1.60 \cdot 10^{-3}$                   | $3.05 \cdot 10^{-3}$                   | $7.18 \cdot 10^{-3}$                   |
|       | $W_n^2$ | $9.43 \cdot 10^{-4}$                   | $1.71 \cdot 10^{-3}$                   | $3.83 \cdot 10^{-3}$                   | $9.84 \cdot 10^{-3}$                   |
|       | $A_n^2$ | $8.53 \cdot 10^{-4}$                   | $1.30 \cdot 10^{-3}$                   | $1.88 \cdot 10^{-3}$                   | $2.69 \cdot 10^{-3}$                   |
| $w_4$ | $D_n$   | $8.10 \cdot 10^{-4}$                   | $1.61 \cdot 10^{-3}$                   | $3.06 \cdot 10^{-3}$                   | $7.19 \cdot 10^{-3}$                   |
|       | $W_n^2$ | $9.46 \cdot 10^{-4}$                   | $1.71 \cdot 10^{-3}$                   | $3.83 \cdot 10^{-3}$                   | $9.83 \cdot 10^{-3}$                   |
|       | $A_n^2$ | $8.54 \cdot 10^{-4}$                   | $1.30 \cdot 10^{-3}$                   | $1.87 \cdot 10^{-3}$                   | $2.69 \cdot 10^{-3}$                   |
| $w_5$ | $D_n$   | $6.26 \cdot 10^{-3}$                   | <b><math>1.27 \cdot 10^{-3}</math></b> | <b><math>1.24 \cdot 10^{-3}</math></b> | <b><math>5.07 \cdot 10^{-3}</math></b> |
|       | $W_n^2$ | $4.45 \cdot 10^{-3}$                   | $2.07 \cdot 10^{-3}$                   | $1.17 \cdot 10^{-3}$                   | <b><math>5.79 \cdot 10^{-3}</math></b> |
|       | $A_n^2$ | $1.47 \cdot 10^{-3}$                   | <b><math>3.73 \cdot 10^{-4}</math></b> | <b><math>6.42 \cdot 10^{-4}</math></b> | $2.73 \cdot 10^{-3}$                   |
| $w_6$ | $D_n$   | $6.27 \cdot 10^{-3}$                   | <b><math>1.27 \cdot 10^{-3}</math></b> | <b><math>1.24 \cdot 10^{-3}</math></b> | $5.08 \cdot 10^{-3}$                   |
|       | $W_n^2$ | $4.45 \cdot 10^{-3}$                   | $2.05 \cdot 10^{-3}$                   | <b><math>1.16 \cdot 10^{-3}</math></b> | $5.80 \cdot 10^{-3}$                   |
|       | $A_n^2$ | $1.48 \cdot 10^{-3}$                   | <b><math>3.73 \cdot 10^{-4}</math></b> | $6.46 \cdot 10^{-4}$                   | $2.73 \cdot 10^{-3}$                   |
| $w_7$ | $D_n$   | $2.52 \cdot 10^{-3}$                   | $1.28 \cdot 10^{-3}$                   | $1.55 \cdot 10^{-3}$                   | $5.66 \cdot 10^{-3}$                   |
|       | $W_n^2$ | $2.42 \cdot 10^{-3}$                   | <b><math>1.39 \cdot 10^{-3}</math></b> | $1.24 \cdot 10^{-3}$                   | $7.21 \cdot 10^{-3}$                   |
|       | $A_n^2$ | $1.15 \cdot 10^{-3}$                   | $6.36 \cdot 10^{-4}$                   | $1.14 \cdot 10^{-3}$                   | <b><math>2.61 \cdot 10^{-3}</math></b> |

Table 11: Standard deviation of residuals  $\{\hat{\varepsilon}_j\}_{j=1}^J$  of the weighted least squares regression of (6) with weights  $\{w_{k,j}\}_{j=1}^J$ ,  $k = 1, \dots, 7$ , for statistics  $D_n$ ,  $W_n^2$ , and  $A_n^2$ . Residuals are presented in four blocks, each one considering the residuals of the observations whose  $\alpha$  value lies within the interval in the column header. Bold highlights the best-performing weight per statistic and  $\alpha$ -block.

| $w_k$ | $T_n$   | Standard deviation                     |                                        |                                        |
|-------|---------|----------------------------------------|----------------------------------------|----------------------------------------|
|       |         | $n \in [5, 10)$                        | $n \in [10, 100)$                      | $n \in [100, 300)$                     |
| $w_1$ | $D_n$   | $1.28 \cdot 10^{-2}$                   | $7.25 \cdot 10^{-3}$                   | $4.11 \cdot 10^{-3}$                   |
|       | $W_n^2$ | $1.96 \cdot 10^{-2}$                   | $4.54 \cdot 10^{-3}$                   | $8.21 \cdot 10^{-4}$                   |
|       | $A_n^2$ | $7.40 \cdot 10^{-3}$                   | $1.43 \cdot 10^{-3}$                   | $4.19 \cdot 10^{-4}$                   |
| $w_2$ | $D_n$   | $1.28 \cdot 10^{-2}$                   | $7.24 \cdot 10^{-3}$                   | $4.10 \cdot 10^{-3}$                   |
|       | $W_n^2$ | $1.96 \cdot 10^{-2}$                   | $4.55 \cdot 10^{-3}$                   | $8.37 \cdot 10^{-4}$                   |
|       | $A_n^2$ | $7.40 \cdot 10^{-3}$                   | $1.43 \cdot 10^{-3}$                   | $4.19 \cdot 10^{-4}$                   |
| $w_3$ | $D_n$   | $1.26 \cdot 10^{-2}$                   | $7.13 \cdot 10^{-3}$                   | $4.03 \cdot 10^{-3}$                   |
|       | $W_n^2$ | $1.92 \cdot 10^{-2}$                   | $4.41 \cdot 10^{-3}$                   | $8.03 \cdot 10^{-4}$                   |
|       | $A_n^2$ | $7.19 \cdot 10^{-3}$                   | $1.32 \cdot 10^{-3}$                   | $4.07 \cdot 10^{-4}$                   |
| $w_4$ | $D_n$   | $1.27 \cdot 10^{-2}$                   | $7.12 \cdot 10^{-3}$                   | $4.02 \cdot 10^{-3}$                   |
|       | $W_n^2$ | $1.92 \cdot 10^{-2}$                   | $4.41 \cdot 10^{-3}$                   | $8.17 \cdot 10^{-4}$                   |
|       | $A_n^2$ | $7.19 \cdot 10^{-3}$                   | $1.32 \cdot 10^{-3}$                   | $4.07 \cdot 10^{-4}$                   |
| $w_5$ | $D_n$   | <b><math>1.07 \cdot 10^{-2}</math></b> | <b><math>5.84 \cdot 10^{-3}</math></b> | <b><math>3.44 \cdot 10^{-3}</math></b> |
|       | $W_n^2$ | <b><math>1.66 \cdot 10^{-2}</math></b> | <b><math>3.34 \cdot 10^{-3}</math></b> | <b><math>7.35 \cdot 10^{-4}</math></b> |
|       | $A_n^2$ | $7.08 \cdot 10^{-3}$                   | <b><math>1.12 \cdot 10^{-3}</math></b> | <b><math>3.91 \cdot 10^{-4}</math></b> |
| $w_6$ | $D_n$   | <b><math>1.07 \cdot 10^{-2}</math></b> | <b><math>5.84 \cdot 10^{-3}</math></b> | <b><math>3.44 \cdot 10^{-3}</math></b> |
|       | $W_n^2$ | <b><math>1.66 \cdot 10^{-2}</math></b> | $3.35 \cdot 10^{-3}$                   | $7.39 \cdot 10^{-4}$                   |
|       | $A_n^2$ | <b><math>7.07 \cdot 10^{-3}</math></b> | <b><math>1.12 \cdot 10^{-3}</math></b> | $3.95 \cdot 10^{-4}$                   |
| $w_7$ | $D_n$   | $1.13 \cdot 10^{-2}$                   | $6.23 \cdot 10^{-3}$                   | $3.64 \cdot 10^{-3}$                   |
|       | $W_n^2$ | $1.76 \cdot 10^{-2}$                   | $3.59 \cdot 10^{-3}$                   | $7.75 \cdot 10^{-4}$                   |
|       | $A_n^2$ | $7.13 \cdot 10^{-3}$                   | $1.16 \cdot 10^{-3}$                   | <b><math>3.91 \cdot 10^{-4}</math></b> |

Table 12: Standard deviation of residuals  $\{\hat{\varepsilon}_j\}_{j=1}^J$  of the weighted least squares regression of (6) with weights  $\{w_{k,j}\}_{j=1}^J$ ,  $k = 1, \dots, 7$ , for statistics  $D_n$ ,  $W_n^2$ , and  $A_n^2$ . The results are divided into three blocks, each one considering the residuals of the observations whose  $n$  value lies within the interval in the column header. Bold highlights the best-performing weight per statistic and  $n$ -block.

| $w_k$ | $T_n$   | Standard deviation                     |                                        |                                        |
|-------|---------|----------------------------------------|----------------------------------------|----------------------------------------|
|       |         | $n \in [5, 10)$                        | $n \in [10, 100)$                      | $n \in [100, 300)$                     |
| $w_1$ | $D_n$   | $1.21 \cdot 10^{-3}$                   | <b><math>8.21 \cdot 10^{-4}</math></b> | <b><math>5.23 \cdot 10^{-4}</math></b> |
|       | $W_n^2$ | $3.03 \cdot 10^{-3}$                   | <b><math>7.93 \cdot 10^{-4}</math></b> | <b><math>5.09 \cdot 10^{-4}</math></b> |
|       | $A_n^2$ | <b><math>2.84 \cdot 10^{-3}</math></b> | <b><math>6.96 \cdot 10^{-4}</math></b> | <b><math>4.41 \cdot 10^{-4}</math></b> |
| $w_2$ | $D_n$   | <b><math>1.19 \cdot 10^{-3}</math></b> | $8.22 \cdot 10^{-4}$                   | $5.29 \cdot 10^{-4}$                   |
|       | $W_n^2$ | <b><math>3.01 \cdot 10^{-3}</math></b> | $7.98 \cdot 10^{-4}$                   | $5.25 \cdot 10^{-4}$                   |
|       | $A_n^2$ | <b><math>2.84 \cdot 10^{-3}</math></b> | <b><math>6.96 \cdot 10^{-4}</math></b> | <b><math>4.41 \cdot 10^{-4}</math></b> |
| $w_3$ | $D_n$   | $1.26 \cdot 10^{-3}$                   | $8.71 \cdot 10^{-4}$                   | $5.67 \cdot 10^{-4}$                   |
|       | $W_n^2$ | $3.19 \cdot 10^{-3}$                   | $8.32 \cdot 10^{-4}$                   | $5.11 \cdot 10^{-4}$                   |
|       | $A_n^2$ | $3.02 \cdot 10^{-3}$                   | $7.33 \cdot 10^{-4}$                   | $4.42 \cdot 10^{-4}$                   |
| $w_4$ | $D_n$   | $1.23 \cdot 10^{-3}$                   | $8.80 \cdot 10^{-4}$                   | $5.80 \cdot 10^{-4}$                   |
|       | $W_n^2$ | $3.17 \cdot 10^{-3}$                   | $8.35 \cdot 10^{-4}$                   | $5.26 \cdot 10^{-4}$                   |
|       | $A_n^2$ | $3.02 \cdot 10^{-3}$                   | $7.33 \cdot 10^{-4}$                   | $4.43 \cdot 10^{-4}$                   |
| $w_5$ | $D_n$   | $1.18 \cdot 10^{-2}$                   | $6.74 \cdot 10^{-3}$                   | $4.05 \cdot 10^{-3}$                   |
|       | $W_n^2$ | $1.78 \cdot 10^{-2}$                   | $3.63 \cdot 10^{-3}$                   | $7.71 \cdot 10^{-4}$                   |
|       | $A_n^2$ | $5.24 \cdot 10^{-3}$                   | $1.26 \cdot 10^{-3}$                   | $5.15 \cdot 10^{-4}$                   |
| $w_6$ | $D_n$   | $1.17 \cdot 10^{-2}$                   | $6.76 \cdot 10^{-3}$                   | $4.07 \cdot 10^{-3}$                   |
|       | $W_n^2$ | $1.79 \cdot 10^{-2}$                   | $3.63 \cdot 10^{-3}$                   | $7.77 \cdot 10^{-4}$                   |
|       | $A_n^2$ | $5.19 \cdot 10^{-3}$                   | $1.27 \cdot 10^{-3}$                   | $5.27 \cdot 10^{-4}$                   |
| $w_7$ | $D_n$   | $4.34 \cdot 10^{-3}$                   | $2.54 \cdot 10^{-3}$                   | $1.47 \cdot 10^{-3}$                   |
|       | $W_n^2$ | $6.30 \cdot 10^{-3}$                   | $1.76 \cdot 10^{-3}$                   | $5.32 \cdot 10^{-4}$                   |
|       | $A_n^2$ | $3.18 \cdot 10^{-3}$                   | $8.83 \cdot 10^{-4}$                   | <b><math>4.41 \cdot 10^{-4}</math></b> |

Table 13: Standard deviation of upper-tail residuals  $\{\hat{\varepsilon}_j \mid \alpha_j \leq 0.25\}_{j=1}^J$  of the weighted least squares regression of (6) with weights  $\{w_{k,j}\}_{j=1}^J$ ,  $k = 1, \dots, 7$ , for statistics  $D_n$ ,  $W_n^2$ , and  $A_n^2$ . Residuals are presented in three blocks, each one considering the residuals of the observations whose  $n$  value lies within the interval in the column header. Bold highlights the best-performing weight per statistic and  $n$ -block.