K-2 rotated goodness-of-fit for multivariate data

Sara Algeri*

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, USA

(Received 24 September 2021; accepted 7 February 2022; published 24 February 2022)

Consider a set of multivariate distributions, $F_1, \ldots, F_M$, aiming to explain the same phenomenon. For instance, each F_m may correspond to a different candidate background model for calibration data, or to one of many possible signal models we aim to validate on experimental data. In this article, we show that tests for a wide class of apparently different models F_m can be mapped into a single test for a reference distribution Q. As a result, valid inference for each F_m can be obtained by simulating only the distribution of the test statistic under Q. Furthermore, Q can be chosen conveniently simple to substantially reduce the computational time.

DOI: 10.1103/PhysRevD.105.035030

I. INTRODUCTION

Despite the popularity of classical goodness-of-fit tests such as Pearson's X^2 [1], likelihood ratio and Kolmogorov-Smirnov [2,3], their applicability often faces serious challenges in many situations relevant to modern experiments. For instance, when conducting multidimensional searches in a binned data regime, the limited sample size may affect the validity of the χ^2 approximation for X^2. Moreover, if the expected number of events is small, the X^2 statistics may be biased, that is, its power can be smaller than the prescribed significance level [4]. Unfortunately, this may occur even when a reasonable χ^2 approximation for it exists, leaving little hope when aiming to address the problem by means of Monte Carlo simulations. Similarly, the likelihood ratio may suffer from additional biases due to the estimation of the unknown parameters [e.g., [5]]. These problems can often be overcome in the unbinned data regime by means of tests such as Kolmogorov-Smirnov, Cramer-von-Mises, and Anderson-Darling. In this case, the price to pay is the loss of distribution-freeness when the models under study are multivariate and/or involve unknown parameters that need to be estimated. As a result, one needs to derive or simulate the distribution of the test statistic on a case-by-case basis.

In this article, we discuss a simulation-based testing strategy which allows us to overcome all these shortcomings and equips experimentalists with a novel tool to perform goodness-of-fit while reducing substantially the computational costs. The rationale behind the solution is somewhat close in spirit (but different in nature) to that of the well-known Metropolis-Hasting algorithm [6,7]. When aiming to sample data from a complex distribution F, the Metropolis-Hasting algorithm circumvents the difficulties associated with sampling directly from F by considering a much simpler distribution Q. The choice of Q is arbitrary and thus one can often compute integrals in F, or approximate the latter, solely relying on samples from Q. In a similar manner, the tests presented here consist of converting the testing problem for a given distribution F into a test for a *reference-distribution* Q. We show that tests for many different distributions $F_1, \ldots, F_M$ can all be mapped into one single test for Q. Also in this case, Q can be chosen conveniently simple. It follows that one can calculate the prescribed test statistic on the data, for one or more candidate models F_m, and compare its observed value directly with the simulated distribution of the test statistic under Q, avoiding M separate simulations.

From a theoretical standpoint, the key element of the solution is the *Khmaladze-2* (K-2) *transform*,[1] also known as *Khmaladze's rotation*, a novel unitary-transformation for empirical processes introduced in recent years by [9,10]. The test statistics proposed in this article are extensions of the Kolmogorov, Cramer-von-Mises and Anderson-Darling's statistics and adequately constructed to account for the variability associated with the estimation of the parameters. For the specific case of Anderson-Darling, we will see that the reference distribution Q also plays the role of weighting function. That is, it can be used to assign the desired weights to the tails of the distribution. Finally, we evaluate the performance of the tests proposed through a suite of simulation studies.

The remainder of the manuscript is organized as follows. In Sec. II we provide an overview on the classical empirical process, that is, the main object at the core of classical

*salgeri@umn.edu

[1]The Khmaladze-2 transformation has not to be confused with the well-known "Khmaladze transformation," also referred to in literature as Khmaladze-1 (K-1) transform, and originally proposed by the same author in [8].

goodness-of-fit tests. Section III is devoted to extend the classical empirical process to the multivariate parametric setting and introduces the projected empirical process. While the latter is shown to provide remarkable computational advantages, its main relevance for us is that of setting the ground to perform distribution-free goodness-of-fit. Distribution-freeness is the focus of Sec. IV. There, we introduce the K-2 transform and investigate its properties through a suite of simulation studies. Some final remarks are collected in Sec. V. Details on the mathematical derivations are provided in the Appendix.

II. THE CLASSICAL EMPIRICAL PROCESS

Consider a sample $x_1, \ldots, x_n$ for which each measurement x_i is the realization of a random variable X_i. For the moment, we assume that the X_is take values on the interval $[L, U]$, are independent and identically distributed (i.i.d.) with cumulative distribution function (cdf), P, either continuous or discrete. In this setup, the empirical process is

$$v_{P,n}(x) = \sqrt{n}[P_n(x) - P(x)] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\mathbb{1}_{\{x_i \leq x\}} - P(x)] \tag{1}$$

where $P_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}}$ is the empirical cumulative distribution of $x_1, \ldots, x_n$ and which is known to converge to P, when $n \to \infty$. From the first equality in (1), it is clear that, for every point in $[L, U]$, v_n consists of a "magnified" difference between the empirical cumulative distribution of the data and P, where the "magnifying factor" is $\sqrt{n}$. Hence, when replacing P with any $F \not\equiv P$, the differences between P_n and F becomes more and more obvious as $n \to \infty$.

The literature investigating the properties of v_n is vast (see Wellner [11] for a review), and mainly focuses on the case where F is fixed. In practical applications, however, F typically depends on unknown parameters to be estimated. It is therefore important to extend (1) to this setting.

III. THE MULTIVARIATE PARAMETRIC REGIME

Consider a sample of i.i.d. observations over the search region $\boldsymbol{\mathcal{X}} \subseteq \mathbb{R}^d$ and let $P(\boldsymbol{x}) = P(x_1, \ldots, x_d)$ be their true underlying distribution. Despite P is unknown, suppose we are given a simplified candidate model $Q_{\boldsymbol{\theta}}(\boldsymbol{x})$ for the data, with $\boldsymbol{\theta}$ being a set of p unknown parameters, and let $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ be the respective probability density function (pdf) or probability mass function (pmf). We assume that $Q_{\boldsymbol{\theta}}$ is easy to simulate from, to evaluate, and to estimate its parameters. For instance, $Q_{\boldsymbol{\theta}}$ may be the cdf of a d-dimensional normal distribution with independent components, known variance and mean vector depending on $\boldsymbol{\theta}$. Moreover, suppose another model, $F_{\boldsymbol{\beta}}$, is given and let $\boldsymbol{\beta}$ be the set of parameters characterizing it. The distribution $F_{\boldsymbol{\beta}}$ may be arbitrarily complex and, potentially, much harder to simulate from, to estimate, and even to evaluate than $Q_{\boldsymbol{\theta}}$. In this section and those to follow, we will show that we can construct two test statistics, one to test $F_{\boldsymbol{\beta}}$ and one to test $Q_{\boldsymbol{\theta}}$, whose null distribution is the same. In order to achieve this goal we begin by constructing a test for $Q_{\boldsymbol{\theta}}$ based on the so-called *projected empirical process*.

A. The projected empirical process

An extension of (1) to this setup is given by the parametric empirical process

$$v_{Q,n}(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{x}_i) \quad \text{with} \tag{2}$$

$$\psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{x}_i) = [\mathbb{1}_{\{\boldsymbol{x}_i \leq \boldsymbol{x}\}} - Q_{\boldsymbol{\theta}}(\boldsymbol{x})] \tag{3}$$

and $\mathbb{1}_{\{\boldsymbol{x}_i \leq \boldsymbol{x}\}} = \mathbb{1}_{\{x_{1i} \leq x_1, \ldots, x_{pi} \leq x_p\}}$ takes value one for all the data points whose coordinates are smaller or equal than $\boldsymbol{x} = (x_1, \ldots, x_d)$ and zero otherwise.

Denote with $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$, which we assume satisfies the classical regularity conditions [e.g., [12], p. 500] (see also [13] for a high-level review). We denote the score vector of $Q_{\boldsymbol{\theta}}$ with $\boldsymbol{u}_{\boldsymbol{\theta}}$, i.e.,

$$\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}) = [u_{\boldsymbol{\theta}_1}(\boldsymbol{x}), \ldots, u_{\boldsymbol{\theta}_p}(\boldsymbol{x})]^T \tag{4}$$

where each element $u_{\boldsymbol{\theta}_j}(\boldsymbol{x})$ corresponds to

$$u_{\boldsymbol{\theta}_j}(\boldsymbol{x}) = \frac{\partial}{\partial \theta_j} \log q_{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{5}$$

with θ_j, $j = 1, \ldots, p$ being the components of the parameter vector $\boldsymbol{\theta}$. We denote with $\Gamma_{\boldsymbol{\theta}}$ the Fisher-information matrix, i.e., the matrix of elements

$$\Gamma_{\boldsymbol{\theta} jk} = \langle u_{\boldsymbol{\theta}_j}, u_{\boldsymbol{\theta}_k} \rangle_{Q_{\boldsymbol{\theta}}}. \tag{6}$$

The inner product in (6) is defined as

$$\langle g, h \rangle_{Q_{\boldsymbol{\theta}}} = \int_{\boldsymbol{\mathcal{X}}} g(\boldsymbol{t}) h(\boldsymbol{t}) q_{\boldsymbol{\theta}}(\boldsymbol{t}) \mathrm{d}\boldsymbol{t} \quad \text{if } Q_{\boldsymbol{\theta}} \text{ is continuous.} \tag{7}$$

If $Q_{\boldsymbol{\theta}}$ is discrete, the integral in (7) is replaced by a summation over all the points of the search region $\boldsymbol{\mathcal{X}}$. Lastly, we consider the normalized score function

$$\boldsymbol{b}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \Gamma_{\boldsymbol{\theta}}^{-1/2} \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{8}$$

and we denote with $b_{\boldsymbol{\theta}_j}(\boldsymbol{x})$, $j = 1, \ldots, p$, its components. The operation in Eq. (8) consists of normalizing the vector $\boldsymbol{u}_{\boldsymbol{\theta}}$ in (4) by multiplying it by the inverse of the square root

matrix of the Fisher information.[2] The resulting functional vector $\boldsymbol{b}_{\boldsymbol{\theta}}$ in (8) consists of the normalized score functions $b_{\boldsymbol{\theta}_j}$, which have mean zero, unit variance, and are uncorrelated from one another under model $Q_{\boldsymbol{\theta}}$.

It was shown in [15] that, when replacing $\boldsymbol{\theta}$ in (2) with $\hat{\boldsymbol{\theta}}$, the resulting process, namely $v_{Q,n}(\boldsymbol{x},\hat{\boldsymbol{\theta}})$, can be rewritten as a projection of $v_{Q,n}(\boldsymbol{x},\boldsymbol{\theta})$ parallel to the normalized score functions $b_{\boldsymbol{\theta}_j}$. Specifically, a Taylor expansion and suitable algebraic manipulations lead to

$$v_{Q,n}(\boldsymbol{x},\hat{\boldsymbol{\theta}}) \approx v_{Q,n}(\boldsymbol{x},\boldsymbol{\theta}) - \frac{1}{\sqrt{n}}\sum_{j=1}^{p}\sum_{i=1}^{n} b_{\boldsymbol{\theta}_j}(\boldsymbol{x}_i)\langle b_{\boldsymbol{\theta}_j}, \psi_{\boldsymbol{x},\boldsymbol{\theta}}\rangle_{Q_{\boldsymbol{\theta}}}, \quad (9)$$

where the error of the approximation is $o_p(1)$,[3] that is, it quickly converges to zero in probability. The inner product in (9) can be computed as in (7). Details on the derivation of (9) are provided in the Appendix.

It follows that, given the set of functions

$$\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t}) = \psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t}) - \sum_{j=1}^{p} b_{\boldsymbol{\theta}_j}(\boldsymbol{t})\langle b_{\boldsymbol{\theta}_j}, \psi_{\boldsymbol{x},\boldsymbol{\theta}}\rangle_{Q_{\boldsymbol{\theta}}}, \quad (10)$$

we can specify the projected empirical process $\tilde{v}_n(\boldsymbol{x},\boldsymbol{\theta})$ as

$$\tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{x}_i), \quad (11)$$

and it is such that

$$v_{Q,n}(\boldsymbol{x},\hat{\boldsymbol{\theta}}) = \tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\theta}) + o_p(1); \quad (12)$$

hence, $v_{Q,n}(\boldsymbol{x},\hat{\boldsymbol{\theta}})$ and $\tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\theta})$ have the same asymptotic distribution.

B. Testing Q

A notable advantage of working with empirical processes is that they allow us to construct an entire family of goodness-of-fit tests. For instance, to test the hypothesis $H_0: P = Q_{\boldsymbol{\theta}}$, many different test statistics can be constructed by simply taking functionals of $\tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\theta})$. Some of these tests will be more powerful then others with respect to different alternatives, and thus, it is particularly valuable to be able to access a variety of them. Here, we focus on three main statistics which can be seen as a generalization of Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling's statistics, i.e.,

$$\hat{D}_Q = \sup_{\boldsymbol{x}}|\tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\theta})|, \hat{\omega}_Q^2 = \int_{\mathcal{X}} \tilde{v}_{Q,n}^2(\boldsymbol{x},\boldsymbol{\theta}) q_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$
$$\text{and} \quad \hat{A}_Q^2 = \int_{\mathcal{X}} \tilde{v}_{Q,n}^2(\boldsymbol{x},\boldsymbol{\theta}) w_{\boldsymbol{\theta}}(\boldsymbol{x}) q_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \quad (13)$$

with $w_{\boldsymbol{\theta}}(\boldsymbol{x}) = [Q_{\boldsymbol{\theta}}(\boldsymbol{x})(1 - Q_{\boldsymbol{\theta}}(\boldsymbol{x}))]^{-1}$ being the weighting function which allows us to highlight differences between the empirical cumulative distribution and $Q_{\boldsymbol{\theta}}$ in the tails.

It is worth emphasizing that, in principle, one can use as test statistics the equivalent of those in (13) with $\tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\theta})$ replaced by $v_{Q,n}(\boldsymbol{x},\hat{\boldsymbol{\theta}})$. There are, however, two main advantages of working with $\tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\theta})$ instead of $v_{Q,n}(\boldsymbol{x},\hat{\boldsymbol{\theta}})$. First of all, as we will discuss in details in Sec. IV, $\tilde{v}_n(\boldsymbol{x},\boldsymbol{\theta})$ sets the foundations to perform distribution-free tests. Second, $\tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\theta})$ provides substantial gain, compared to $v_{Q,n}(\boldsymbol{x},\hat{\boldsymbol{\theta}})$, from a computational stand point.

Specifically, in both cases, since $\boldsymbol{\theta}$ is unknown, one needs to simulate the distribution of the test statistics by means of the paramteric bootstrap, that is, we compute the MLE of $\boldsymbol{\theta}$ on the data observed, namely $\hat{\boldsymbol{\theta}}_{\text{obs}}$, and, at each replicate, we sample datasets from $Q_{\hat{\boldsymbol{\theta}}_{\text{obs}}}(\boldsymbol{x})$. The bootstrap procedure has been proven to lead to consistent results under very general conditions by Babu and Rao [16]. They have shown that by simulating the distribution of continuous functionals of the parametric empirical process one can recover their true distribution if the parameters are estimated via MLE and the classical regularity conditions [e.g., [12], p. 500] hold.

When working with $v_{Q,n}(\boldsymbol{x},\hat{\boldsymbol{\theta}})$, to account for the variability introduced by the estimation process, one needs to repeat the maximization of the likelihood on each simulated bootstrap sample. Moreover, at each replicate, the cdf $Q_{\boldsymbol{\theta}}$ also needs to be evaluated on each point $\boldsymbol{x} \in \mathcal{X}$ considered, and with $\boldsymbol{\theta}$ replaced by its estimated value on the simulated bootstrap sample. On the other hand, when working with $\tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\theta})$, to account for the uncertainty associated with the estimation of $\boldsymbol{\theta}$, instead of maximizing the likelihood at each iteration, we only need to evaluate the normalized score functions in $\boldsymbol{b}_{\hat{\boldsymbol{\theta}}_{\text{obs}}}(\boldsymbol{x})$ on each simulated samples. Furthermore, despite we still need to evaluate $Q_{\boldsymbol{\theta}}$ at each $\boldsymbol{x} \in \mathcal{X}$ considered, as well as the integrals/summations in $\langle b_{\boldsymbol{\theta}_j}, \psi_{\boldsymbol{x},\boldsymbol{\theta}}\rangle_{Q_{\boldsymbol{\theta}}}$, these only need to be computed once, that is, for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\boldsymbol{obs}}$, reducing substantially the computational time. This approach is particularly advantageous since the error of approximating $v_{Q,n}(\boldsymbol{x},\hat{\boldsymbol{\theta}})$ with $\tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\theta})$ is only $o_p(1)$ [see Eq. (12)], and thus, it is negligible even for samples which are only moderately large.

[2] In the applications to follow, the square root matrix has been computed via the Schur method [e.g., [14], Ch. 6]. Nonetheless, other methods to construct the square root matrix, such as diagonalization, Jordan decomposition, etc, are also viable options.

[3] The notation $o_p(1)$ is an abbreviation used in statistics to indicate that a sequence of random vectors converges to zero in probability. In general, given two random sequences R_n and S_n, we write $R_n = o_p(S_n)$ to indicate that $\frac{R_n}{S_n}$ converges in probability to zero.

To illustrate these aspects with a toy example, let Q be the distribution of a bivariate normal with independent components, truncated over the region $\mathcal{X} = [1, 20] \times [1, 25]$, and with density

$$q_{\boldsymbol{\theta}}(\boldsymbol{x}) \propto e^{-\frac{1}{2\theta_3}[(x_1-\theta_1)^2+(x_2-\theta_2)^2]}, \tag{14}$$

We draw a sample of $n = 100$ observations from (14) with $\boldsymbol{\theta} = (-2, 5, 25)$, and which will be considered our "observed data." We estimate $\boldsymbol{\theta}$ on such sample and we obtain $\hat{\theta}_{\text{obs}} = (-0.77, 6.32, 22.02)$. We proceed by simulating the distribution of the Kolmogorov-Smirnov's statistics, $\sup_{\boldsymbol{x}}|\tilde{v}_{Q,n}(\boldsymbol{x}, \boldsymbol{\theta})|$ and $\sup_{\boldsymbol{x}}|v_{Q,n}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})|$, via the parametric bootstrap. To emphasize the validity of the bootstrap procedure, we also simulate the distribution of $\sup_{\boldsymbol{x}}|v_{Q,n}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})|$ via Monte Carlo; that is, the data are generated from $Q_{\boldsymbol{\theta}}(\boldsymbol{x})$ (instead of $Q_{\hat{\boldsymbol{\theta}}_{\text{obs}}}(\boldsymbol{x})$ as in the parametric bootstrap) and the estimation process is repeated at each replicate. In all the three cases, the supremum is taken over a grid of 2000 equidistant points over $\mathcal{X}$. The results obtained are shown in Figure 1. The three simulated distributions are effectively overlapping, providing evidence that the parametric bootstrap does recover the distribution of $\sup_{\boldsymbol{x}}|v_{Q,n}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})|$. Not surprisingly, this is true even when relying on $\tilde{v}_{Q,n}(\boldsymbol{x}, \boldsymbol{\theta})$ instead of $v_{Q,n}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ due to the small error associated with approximating the latter with the former. Notice that, this is true even if our sample size is limited to 100 observations. Moreover, working with the projected empirical process, $\tilde{v}_{Q,n}(\boldsymbol{x}, \boldsymbol{\theta})$, provides a remarkable computational gain

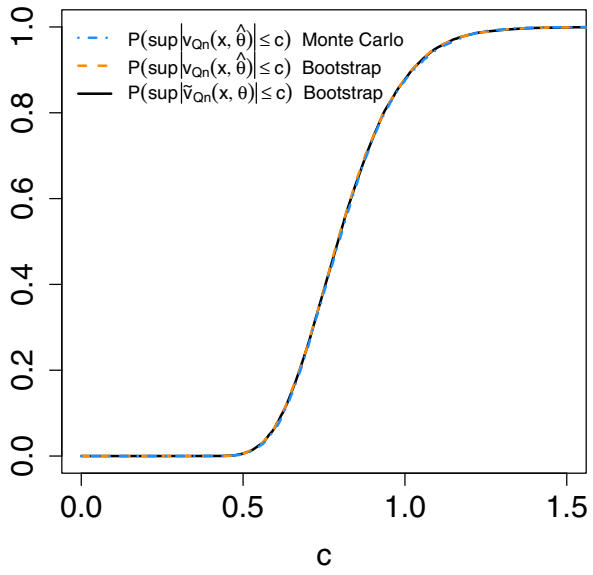

FIG. 1. Comparing the bootstrapped distributions of the Kolmogorov-Smirnov's statistics $\sup_{\boldsymbol{x}}|\tilde{v}_{Q,n}(\boldsymbol{x}, \boldsymbol{\theta})|$ and $\sup_{\boldsymbol{x}}|v_{Q,n}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})|$ and the distribution of $\sup_{\boldsymbol{x}}|v_{Q,n}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})|$ simulated via Monte Carlo. In all three cases the simulation consist of 10,000 replicates and the sample size is $n = 100$.

TABLE I. Overall (system + user) CPU time needed to simulate the distributions of the test statistics $\sup_{\boldsymbol{x}}|\tilde{v}_{Q,n}(\boldsymbol{x}, \boldsymbol{\theta})|$ and $\sup_{\boldsymbol{x}}|v_{Q,n}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})|$ via the parametric bootstrap over 10,000 replicates and $n = 100$ observations.

	$\sup_{\boldsymbol{x}}\lvert\tilde{v}_{Q,n}(\boldsymbol{x}, \boldsymbol{\theta})\rvert$	$\sup_{\boldsymbol{x}}\lvert v_{Q,n}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})\rvert$
CPU time	9.429 mins	12.198 hrs

compared to $v_{Q,n}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$. As shown in Table I, simulating the distribution of $\sup_{\boldsymbol{x}}|v_{Q,n}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})|$ using 10,000 replicates required approximately 12 hours of (user + system) CPU time, whereas simulating the distribution of $\sup_{\boldsymbol{x}}|\tilde{v}_{Q,n}(\boldsymbol{x}, \boldsymbol{\theta})|$ required 9.5 minutes.

IV. CONNECTING TESTS FOR F AND TESTS FOR Q

In principle, we could proceed testing any $F_{\boldsymbol{\beta}} \not\equiv Q_{\boldsymbol{\theta}}$ following exactly the same steps described in Sec. III B. In many practical situations, however, $F_{\boldsymbol{\beta}}$ may be sufficiently complex to make the evaluation of the score functions over several samples impractical. To overcome this limitation, we proceed by constructing a new set of test statistics, namely $\tilde{D}_F$, $\tilde{\omega}^2_F$, and $\tilde{A}^2_F$, whose limiting distributions, under $F_{\boldsymbol{\beta}}$, are the same as those of $\hat{D}_Q$, $\hat{\omega}^2_Q$, and $\hat{A}^2_Q$ in (13), under $Q_{\boldsymbol{\theta}}$. As a result, one can compute $\tilde{D}_F$, $\tilde{\omega}^2_F$, and $\tilde{A}^2_F$ only once on the data observed, and compare their values with the simulated distribution of $\hat{D}_Q$, $\hat{\omega}^2_Q$, and $\hat{A}^2_Q$. This can be done by means of the K-2 transform [9,10] as described below.

Let $\boldsymbol{\beta} \in \mathbb{R}^p$ be the vector of unknown parameters characterizing $F_{\boldsymbol{\beta}}$, let $f_{\boldsymbol{\beta}}(\boldsymbol{x})$ be its density (either pdf or pmf) and denote with $a_{\boldsymbol{\beta}_j}$, $j = 1, \ldots, p$, its normalized score functions. The latter can be constructed as in (8) by replacing $q_{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ with $f_{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, respectively. For what follows, we require that $f_{\boldsymbol{\beta}}(\boldsymbol{x}) = 0$ if and only if $q_{\boldsymbol{\theta}}(\boldsymbol{x}) = 0$, that is, the two densities must share the same support. Moreover, we assume that $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, have the same dimension p.

Equations (10)–(11) imply that the process $\tilde{v}_{Q,n}(\boldsymbol{x}, \boldsymbol{\theta})$ "lives" in the space of functions $\mathcal{L}_\perp(Q_{\boldsymbol{\theta}})$ such that

$$\mathcal{L}_\perp(Q_{\boldsymbol{\theta}}) = \{\tilde{\psi} : \langle\tilde{\psi}, \tilde{\psi}\rangle_{Q_{\boldsymbol{\theta}}} < \infty \tag{15}$$

$$\langle\tilde{\psi}, 1\rangle_{Q_{\boldsymbol{\theta}}} = 0, \quad \text{and} \tag{16}$$

$$\langle\tilde{\psi}, b_{\boldsymbol{\theta}_j}\rangle_{Q_{\boldsymbol{\theta}}} = 0, \quad \text{for all } j = 1, \ldots, p\} \tag{17}$$

That is, each function in $\mathcal{L}_\perp(Q_{\boldsymbol{\theta}})$ is square-integrable with respect to $Q_{\boldsymbol{\theta}}$, has mean zero, and is orthogonal to the normalized score functions $b_{\boldsymbol{\theta}_j}$, $j = 1, \ldots, p$, under $Q_{\boldsymbol{\theta}}$. Moreover, one can show that, under $Q_{\boldsymbol{\theta}}$, the

process $\tilde{v}_n(\boldsymbol{x},\boldsymbol{\theta})$ is asymptotically Gaussian with mean $\langle\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}},1\rangle_{Q_{\boldsymbol{\theta}}}=0$ and covariance $\langle\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}},\tilde{\psi}_{\boldsymbol{s},\boldsymbol{\theta}}\rangle_{Q_{\boldsymbol{\theta}}}<\infty$.

The rationale behind the K-2 transformation is that of constructing a suitable map which allows us to transform functions $\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}\in\mathcal{L}_\perp(Q_{\boldsymbol{\theta}})$ into functions in $\mathcal{L}_\perp(F_{\boldsymbol{\beta}})$, i.e.,

$$\mathcal{L}_\perp(F_{\boldsymbol{\beta}})=\{\tilde{\phi}:\langle\tilde{\phi},\tilde{\phi}\rangle_{F_{\boldsymbol{\beta}}}<\infty, \tag{18}$$

$$\langle\tilde{\phi},1\rangle_{F_{\boldsymbol{\beta}}}=0, \quad \text{and} \tag{19}$$

$$\langle\tilde{\phi},a_j\rangle_{F_{\boldsymbol{\beta}}}=0, \quad \text{for all } j=1,\ldots,p\}, \tag{20}$$

where $\langle\cdot,\cdot\rangle_{F_{\boldsymbol{\beta}}}$ can be defined similarly to $\langle\cdot,\cdot\rangle_{Q_{\boldsymbol{\theta}}}$ in (7). Notice that $\mathcal{L}_\perp(F_{\boldsymbol{\beta}})\subset\mathcal{L}(F_{\boldsymbol{\beta}})\subset L^2(F_{\boldsymbol{\beta}})$, with

$$L^2(F_{\boldsymbol{\beta}})=\{\tilde{\phi}:\langle\tilde{\phi},\tilde{\phi}\rangle_{F_{\boldsymbol{\beta}}}<\infty\}, \quad \text{and}$$
$$\mathcal{L}(F_{\boldsymbol{\beta}})=\{\tilde{\phi}:\langle\tilde{\phi},1\rangle_{F_{\boldsymbol{\beta}}}=0, \quad \langle\tilde{\phi},\tilde{\phi}\rangle_{F_{\boldsymbol{\beta}}}<\infty\}.$$

It follows that, for suitable choices of $\tilde{\phi}$, namely $\tilde{\phi}_{\boldsymbol{x},\boldsymbol{\lambda}}$ (soon to be defined), the process $\tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\theta})$ in (11) and the empirical process

$$\tilde{v}_{F,n}(\boldsymbol{x},\boldsymbol{\lambda})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\phi}_{\boldsymbol{x},\boldsymbol{\lambda}}(\boldsymbol{x}_i), \quad \text{with} \quad \boldsymbol{\lambda}=(\boldsymbol{\theta},\boldsymbol{\beta}), \tag{21}$$

have the same asymptotic distribution (under $Q_{\boldsymbol{\theta}}$ and $F_{\boldsymbol{\beta}}$, respectively). Specifically, in virtue of Gaussianity, we can fully characterize the distribution of $\tilde{v}_{F,n}(\boldsymbol{x},\boldsymbol{\theta})$ and $\tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\lambda})$ considering only their mean and covariance. Therefore, to achieve our purpose, it is sufficient to identify a set of functions $\tilde{\phi}_{\boldsymbol{x},\boldsymbol{\lambda}}\in\mathcal{L}_\perp(F_{\boldsymbol{\beta}})$ such that the mean and covariance functions of $\tilde{v}_{F,n}(\boldsymbol{x},\boldsymbol{\theta})$ and $\tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\lambda})$ are the same, i.e.,

$$\langle\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}},1\rangle_{Q_{\boldsymbol{\theta}}}=\langle\tilde{\phi}_{\boldsymbol{x},\boldsymbol{\lambda}},1\rangle_{F_{\boldsymbol{\beta}}}=0 \quad \text{and}$$
$$\langle\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}},\tilde{\psi}_{\boldsymbol{s},\boldsymbol{\theta}}\rangle_{Q_{\boldsymbol{\theta}}}=\langle\tilde{\phi}_{\boldsymbol{x},\boldsymbol{\lambda}},\tilde{\phi}_{\boldsymbol{s},\boldsymbol{\lambda}}\rangle_{F_{\boldsymbol{\beta}}}$$

The functions $\tilde{\phi}_{\boldsymbol{x},\boldsymbol{\lambda}}$ can be constructed as outlined below.

Step 1—Map the functions $\psi_{\boldsymbol{x},\boldsymbol{\theta}}$ in Eq. (3) and the normalized score functions $b_{\boldsymbol{\theta}_j}$ into $L^2(F_{\boldsymbol{\beta}})$ via the isometry

$$l(\boldsymbol{x})=\sqrt{\frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{f_{\boldsymbol{\beta}}(\boldsymbol{x})}}.$$

Obtain

$$l(\boldsymbol{t})\psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t})\in L^2(F_{\boldsymbol{\beta}}) \quad \text{and} \tag{22}$$

$$l(\boldsymbol{t})b_{\boldsymbol{\theta}_j}(\boldsymbol{t})\in L^2(F_{\boldsymbol{\beta}}). \tag{23}$$

For instance, to see (22), consider the inner product

$$\begin{aligned}\langle l\psi_{\boldsymbol{x},\boldsymbol{\theta}},l\psi_{\boldsymbol{x},\boldsymbol{\theta}}\rangle_{F_{\boldsymbol{\beta}}}&=\int_{\mathcal{X}}l^2(\boldsymbol{t})\psi^2_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t})f_{\boldsymbol{\beta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}\\&=\int_{\mathcal{X}}\frac{q_{\boldsymbol{\theta}}(\boldsymbol{t})}{f_{\boldsymbol{\beta}}(\boldsymbol{t})}\psi^2_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t})f_{\boldsymbol{\beta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}\\&=\int_{\mathcal{X}}\psi^2_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t})q_{\boldsymbol{\theta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}=\langle\psi_{\boldsymbol{x},\boldsymbol{\theta}},\psi_{\boldsymbol{x},\boldsymbol{\theta}}\rangle_Q<\infty.\end{aligned}$$

Equivalent calculations can be used to show (23).

Step 2—Map the functions in (22) and (23) into $\mathcal{L}(F_{\boldsymbol{\beta}})$ by means of the unitary operator,[4] K, and defined as

$$Kh(\boldsymbol{t})=h(\boldsymbol{t})-\frac{1-l(\boldsymbol{t})}{1-\langle l,1\rangle_{F_{\boldsymbol{\beta}}}}\langle 1-l,h\rangle_{F_{\boldsymbol{\beta}}}, \tag{24}$$

where the notation $Kh(\boldsymbol{t})$ is used to indicate that the operator K acts on everything on its right. Obtain

$$Kl(\boldsymbol{t})\psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t})\in\mathcal{L}(F_{\boldsymbol{\beta}}) \tag{25}$$

$$\text{and} \quad c_{\lambda_j}(\boldsymbol{t})=Kl(\boldsymbol{t})b_{\boldsymbol{\theta}_j}(\boldsymbol{t})\in\mathcal{L}(F_{\boldsymbol{\beta}}). \tag{26}$$

To see (25), write 9.5

$$\begin{aligned}Kl(\boldsymbol{t})\psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t})&=l(\boldsymbol{t})\psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t})\\&\quad-\frac{1-l(\boldsymbol{x})}{1-\int_{\mathcal{X}}l(\boldsymbol{t})f_{\boldsymbol{\beta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}}\int_{\mathcal{X}}l(t)\psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t})f_{\boldsymbol{\beta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}.\end{aligned}$$

It follows that 9.5

$$\begin{aligned}\langle Kl\psi_{\boldsymbol{x},\boldsymbol{\theta}},1\rangle_F&=\int_{\mathcal{X}}l(\boldsymbol{x})\psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t})f_{\boldsymbol{\beta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}\\&\quad-\frac{1-\int_{\mathcal{X}}l(\boldsymbol{t})f_{\boldsymbol{\beta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}}{1-\int_{\mathcal{X}}l(\boldsymbol{t})f_{\boldsymbol{\beta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}}\int_{\mathcal{X}}l(\boldsymbol{t})\psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t})f_{\boldsymbol{\beta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}=0.\end{aligned}$$

One can proceed similarly for (26).

Step 3—Map each function c_{λ_j} in (26) with $j>1$ into functions $\tilde{c}_{\lambda_j}$ orthogonal to each $a_{\boldsymbol{\beta}_k}$ with $k<j$. This can be done by means of the unitary operator

$$U_{a_{\beta_j}c_{\lambda_j}}h(\boldsymbol{t})=h(\boldsymbol{t})-\frac{\langle a_{\boldsymbol{\beta}_j}-c_{\lambda_j},\cdot\rangle_{F_{\boldsymbol{\beta}}}}{1-\langle a_{\boldsymbol{\beta}_j},c_{\lambda_j}\rangle_{F_{\boldsymbol{\beta}}}}(a_{\boldsymbol{\beta}_j}(\boldsymbol{t})-c_{\lambda_j}(\boldsymbol{t})). \tag{27}$$

One can easily verify that the operator $U_{a_{\beta_j}c_{\lambda_j}}$ maps the functions $a_{\boldsymbol{\beta}_j}$ into functions c_{λ_j}, and vice-versa, whereas, it leaves functions orthogonal to both $a_{\boldsymbol{\beta}_j}$ and c_{λ_j} unchanged.

We construct $\tilde{c}_2,\ldots,\tilde{c}_p$, by combining operators of the form in (27), i.e.,

[4]A unitary operator is an operator that preserves the inner product. That is, if an operator K is unitary in the Hilbert space $\mathcal{H}$ equipped with the inner product $\langle,\rangle_{\mathcal{H}}$, then $\langle Kh_1,Kh_2\rangle_{\mathcal{H}}=\langle h_1,h_2\rangle_{\mathcal{H}}$, for every $h_1,h_2\in\mathcal{H}$.

$$\tilde{c}_{\lambda_2}(t) = U_{a_{\beta_1} c_{\lambda_1}} c_{\lambda_2}(t)$$
$$\tilde{c}_{\lambda_3}(t) = U_{a_{\beta_2} \tilde{c}_{\lambda_2}} U_{a_{\beta_1} c_{\lambda_1}} c_{\lambda_3}(t)$$
$$\dots$$
$$\tilde{c}_{\lambda_p}(t) = U_{a_{\beta_{(p-1)}} \tilde{c}_{\lambda_{(p-1)}}} \dots U_{a_{\beta_1} c_{\lambda_1}} c_{\lambda_p}(t), \tag{28}$$

where each operator $U_{a_{\beta_j} c_{\lambda_j}}$ acts on everything on its right. As highlighted in what follows, these functions are needed to rotate c_{λ_j}s into a_{β_j}s.

Step 4—Consider the unitary operator

$$\boldsymbol{U}h(t) = U_{a_{\beta_p} \tilde{c}_{\lambda_p}} \dots U_{a_{\beta_2} \tilde{c}_{\lambda_2}} U_{a_{\beta_1} c_{\lambda_1}} h(t) \tag{29}$$

and set

$$\phi_{\boldsymbol{x},\boldsymbol{\lambda}}(t) = \boldsymbol{U}Kl(t)\psi_{\boldsymbol{x},\boldsymbol{\theta}}(t) \tag{30}$$

Map each c_{λ_j} into a_{β_j} via $\boldsymbol{U}$ and apply the latter to $Kl\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}$. Obtain $\tilde{\phi}_{\boldsymbol{x},\boldsymbol{\lambda}} \in \mathcal{L}_\perp(F_{\boldsymbol{\beta}})$ such that

$$\tilde{\phi}_{\boldsymbol{x},\boldsymbol{\lambda}}(t) = \boldsymbol{U}Kl(t)\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}(t) \tag{31}$$

$$= \boldsymbol{U}K\left[l(t)\psi_{\boldsymbol{x},\boldsymbol{\theta}}(t) - \sum_{j=1}^{p} l(t) b_{\boldsymbol{\theta}_j}(t) \langle l b_{\boldsymbol{\theta}_j}, l\psi_{\boldsymbol{x},\boldsymbol{\theta}} \rangle_{F_\beta}\right] \tag{32}$$

$$= \boldsymbol{U}\left[Kl(t)\psi_{\boldsymbol{x},\boldsymbol{\theta}}(t) - \sum_{j=1}^{p} c_{\lambda j} \langle c_{\lambda_j}, Kl\psi_{\boldsymbol{x},\boldsymbol{\theta}} \rangle_{F_\beta}\right] \tag{33}$$

$$= \phi_{\boldsymbol{x},\boldsymbol{\lambda}}(t) - \sum_{j=1} a_{\boldsymbol{\beta}_j}(t) \langle a_{\boldsymbol{\beta}_j}, \phi_{\boldsymbol{x},\boldsymbol{\lambda}} \rangle_{F_\beta}. \tag{34}$$

Where (32) follows from the definition of the functions $\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}$ in (10). Equation (33) follow from (26), from the fact that K is unitary (and thus it preserve the inner product), and because the isometry l is such that $\langle lh, lh \rangle_{F_\beta} = \langle h, h \rangle_{Q_\theta}$. Equation (34) follows from (30) and the properties of the operator $\boldsymbol{U}$ (that is, it is unitary and it maps each $c_{\boldsymbol{\lambda}_j}$ into $a_{\boldsymbol{\beta}_j}$). To see the latter, consider for instance $\boldsymbol{U}c_{\lambda_1}(t)$, i.e.,

$$\boldsymbol{U}c_{\lambda_1}(t) = U_{a_{\beta_p} \tilde{c}_{\lambda_p}} \dots U_{a_{\beta_2} \tilde{c}_{\lambda_2}} U_{a_{\beta_1} c_{\lambda_1}} c_{\lambda_1}(t) \tag{35}$$

$$= U_{a_{\beta_p} \tilde{c}_{\lambda_p}} \dots U_{a_{\beta_2} \tilde{c}_{\lambda_2}} a_{\beta_1}(t) \tag{36}$$

$$= a_{\beta_1}(t) \tag{37}$$

where (36) follows since $U_{a_{\beta_1} c_{\lambda_1}}$ maps c_{λ_1} into a_{β_1}. Whereas, (37) follows from the fact that each a_{β_j} and $\tilde{c}_{\lambda_j}$, with $j \geq 2$, are orthogonal to a_{β_1} and each $U_{a_{\beta_j} \tilde{c}_{\lambda_j}}$ leaves functions orthogonal to a_{β_j} and $\tilde{c}_{\lambda_j}$ unchanged. Moreover, to see that $\tilde{\phi}_{\boldsymbol{x},\boldsymbol{\lambda}} = \boldsymbol{U}Kl\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}} \in \mathcal{L}_\perp(F_{\boldsymbol{\beta}})$, consider

$$\langle \boldsymbol{U}Kl\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}, a_{\boldsymbol{\beta}_j} \rangle_{F_\beta} = \langle \boldsymbol{U}Kl\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}, \boldsymbol{U}c_{\lambda_j} \rangle_{F_\beta} \tag{38}$$

$$= \langle Kl\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}, c_{\lambda_j} \rangle_{F_\beta} \tag{39}$$

$$= \langle l\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}, l b_{\boldsymbol{\theta}_j} \rangle_{F_\beta} \tag{40}$$

$$= \langle \tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}, b_{\boldsymbol{\theta}_j} \rangle_{Q_\theta} = 0, \tag{41}$$

where the equalities in (39)–(40) follow from the properties $\boldsymbol{U}$, K, and l.

Clearly, for $Q_{\boldsymbol{\theta}}$ and $F_{\boldsymbol{\beta}}$ discrete, all the integrals involved in Steps 1–4 need to be replaced by summations over all the points of the search region $\boldsymbol{\mathcal{X}}$. Moreover, it should be noted that, in virtue of the properties of the $\boldsymbol{U}$, K, and l we have

$$\langle b_{\boldsymbol{\theta}_j}, \tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}} \rangle_{Q_\theta} = \langle c_{\lambda_j}, Kl\psi_{\boldsymbol{x},\boldsymbol{\theta}} \rangle_{F_\beta} = \langle a_{\boldsymbol{\beta}_j}, \phi_{\boldsymbol{x},\boldsymbol{\lambda}} \rangle_{F_\beta}. \tag{42}$$

Hence, when evaluating the functions $\tilde{\phi}_{\boldsymbol{x},\boldsymbol{\lambda}}(t)$ in (31), one can avoid computing $\langle a_{\boldsymbol{\beta}_j}, \phi_{\boldsymbol{x},\boldsymbol{\lambda}} \rangle_{F_\beta}$ by replacing it with $\langle b_{\boldsymbol{\theta}_j}, \tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}} \rangle_{Q_\theta}$.

From (31), it is easy to see that K-2 effectively consists of a combination of the unitary operators $\boldsymbol{U}$, K and the isometry l. Intuitively, in Step 1, the isometry l allows us to convert our functions $\psi_{\boldsymbol{x},\boldsymbol{\theta}}$, square-integrable in $Q_{\boldsymbol{\theta}}$, into square integrable functions in $F_{\boldsymbol{\beta}}$. The resulting functions $l\psi_{\boldsymbol{x},\boldsymbol{\theta}}$ and $l_\lambda b_{\theta j}$, however, do not have zero-mean with respect to $F_{\boldsymbol{\beta}}$ (they are not orthogonal to one). Therefore, in Step 2, we apply the unitary operator K. This brings us to the space $\mathcal{L}(F_{\boldsymbol{\beta}})$. If $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ were known, that is, if the two models were fully specified, the isometry l and the operator K would only need to be applied to the functions $\psi_{\boldsymbol{x},\boldsymbol{\theta}}$ (as there would be no score functions) and no further mapping would be needed. Whereas, for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ unknown, two extra steps are necessary. That is because, in this setting, $\mathcal{L}(F_{\boldsymbol{\beta}})$ is not quite yet be in the space we want to be [i.e., $\mathcal{L}_\perp(F_{\boldsymbol{\beta}})$] as we have not yet achieved orthogonality with respect to the score functions $a_{\boldsymbol{\beta}_j}$ s. Hence, in Step 3, we exploit the unitary operator $\boldsymbol{U}$ to map our $c_{\boldsymbol{\lambda}_j} = Klb_{\theta j}$ into $\tilde{c}_{\boldsymbol{\lambda}_j}$ functions which are orthogonal to the $a_{\boldsymbol{\beta}_j}$. Finally, in Step 4, we rotate the $c_{\lambda j}$s into $a_{\boldsymbol{\beta}_j}$s via $\boldsymbol{U}$. The same operator is applied also to the functions $Kl\psi_{\boldsymbol{x},\boldsymbol{\theta}}$ to ensure that the functions $\tilde{\phi}_{\boldsymbol{x}\boldsymbol{\lambda}} = Kl\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}$ in (31) are in $\mathcal{L}_\perp(F_{\boldsymbol{\beta}})$.

To test the hypothesis $H_0 : P = F_{\boldsymbol{\beta}}$, we consider the K-2 rotated equivalent of the test statistics in (13), i.e.,

$$\tilde{D}_F = \sup_{\boldsymbol{x}} |\tilde{v}_{F,n}(\boldsymbol{x},\boldsymbol{\lambda})|, \qquad \tilde{\omega}_F^2 = \int_{\mathcal{X}} \tilde{v}_{F,n}^2(\boldsymbol{x},\boldsymbol{\lambda}) q_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$
$$\text{and} \quad \tilde{A}_F^2 = \int_{\mathcal{X}} \tilde{v}_{F,n}^2(\boldsymbol{x},\boldsymbol{\lambda}) w_{\boldsymbol{\theta}}(\boldsymbol{x}) q_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \tag{43}$$

with $\tilde{v}_{F,n}(\boldsymbol{x},\boldsymbol{\lambda})$ as in (21). Under $F_{\boldsymbol{\beta}}$ and $Q_{\boldsymbol{\theta}}$, respectively, $\tilde{v}_{F,n}(\boldsymbol{x},\boldsymbol{\lambda})$ and $\tilde{v}_{Q,n}(\boldsymbol{x},\boldsymbol{\theta})$ have the same asymptotic

distribution, and the same is true for the statistics in (13) and (43).

Notice that, in practice, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are unknown. Hence, in order to compute steps 1–4, one can proceed by simply plugging-in their MLEs $\hat{\boldsymbol{\beta}}_{\text{obs}}$ and $\hat{\boldsymbol{\theta}}_{\text{obs}}$ obtained on the observed data. In the case where $P \equiv F_{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_{\text{obs}}$ converges, in probability, to the true value of $\boldsymbol{\beta}$, whereas, $\hat{\boldsymbol{\theta}}_{\text{obs}}$ converges to the values of $\boldsymbol{\theta}$ which minimizes the Kullback-Leibler divergence between $F_{\boldsymbol{\beta}}$ and $Q_{\boldsymbol{\theta}}$ [e.g., [17], p. 147]. The integrals can be computed as Darboux sums over a grid of possible $\boldsymbol{x}$ values on the search region $\mathcal{X}$. Finally, it is worth pointing out that all the operators considered are linear, and thus, when p is large, their implementation may be tedious but yet relatively simple; especially since they only need to be computed once in order to evaluate (43) on the data observed.

A. Empirical studies

To assess the performance of the testing procedure described above, we consider a dataset of $n = 100$ observations generated from a bivariate Cauchy distribution, P, truncated over the range $\mathcal{X} = [1, 20] \times [1, 25]$, and density

$$p(\boldsymbol{x}) \propto (2\pi)^{-1}|\boldsymbol{\Sigma}|^{-1/2}[1 + (\boldsymbol{x} - \mu)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \mu)]^{-3/2} \quad (44)$$

where $\mu = (0, 3)^T$, $\boldsymbol{\Sigma}$ is a matrix of diagonal elements $\sigma_{11} = \sigma_{22} = 20$ and off-diagonal elements $\sigma_{12} = \sigma_{21} = 10$. Our goal is to test the validity of three different models for our data. Specifically,

$$\begin{aligned} f_1(\boldsymbol{x};\boldsymbol{\beta}) &\propto x_1^{(\beta_1-1)}x_2^{(\beta_2-1)}\exp\{-\beta_3(x_1 + x_2)\}, \\ f_2(\boldsymbol{x};\boldsymbol{\beta}) &\propto \frac{\beta_3}{2\pi}[(x_1 - \beta_1)^2 + (x_2 - \beta_2)^2 + \beta_3]^{-3/2}, \\ f_3(\boldsymbol{x};\boldsymbol{\beta}) &\propto e^{-\frac{1}{200}[(\frac{x_1}{\beta_1}-1)^2+(\frac{x_2}{\beta_2}-1)^2-\beta_3(\frac{x_1}{\beta_1}-1)(\frac{x_2}{\beta_2}-1)]}, \end{aligned} \quad (45)$$

that is, f_1 is the pdf of a bivariate Gamma with independent components, f_2 is the pdf of a bivariate Cauchy with dependent component [but with dependence structure different from (44)], and f_3 is the pdf of a multivariate normal with dependent components. We denote with F_1, F_2 and F_3 the respective cdfs. Finally, we consider as reference distribution, Q, the bivariate normal with independent components introduced in Sec. III B and with pdf as in Eq. (14). Notice that all the models in (45) are quite different from each other as well as from (14). Moreover, each of these models is characterized by $p = 3$ unknown parameters.

We proceed by simulating the null distributions of the three test statistics in (13) under Q and their counterparts for each of the F_m, $m = 1, 2, 3$, models considered; we denote the latter with $\hat{D}_{F_m}, \hat{\omega}^2_{\text{6F}_m}$ and $\hat{A}^2_{F_m}$. The results are

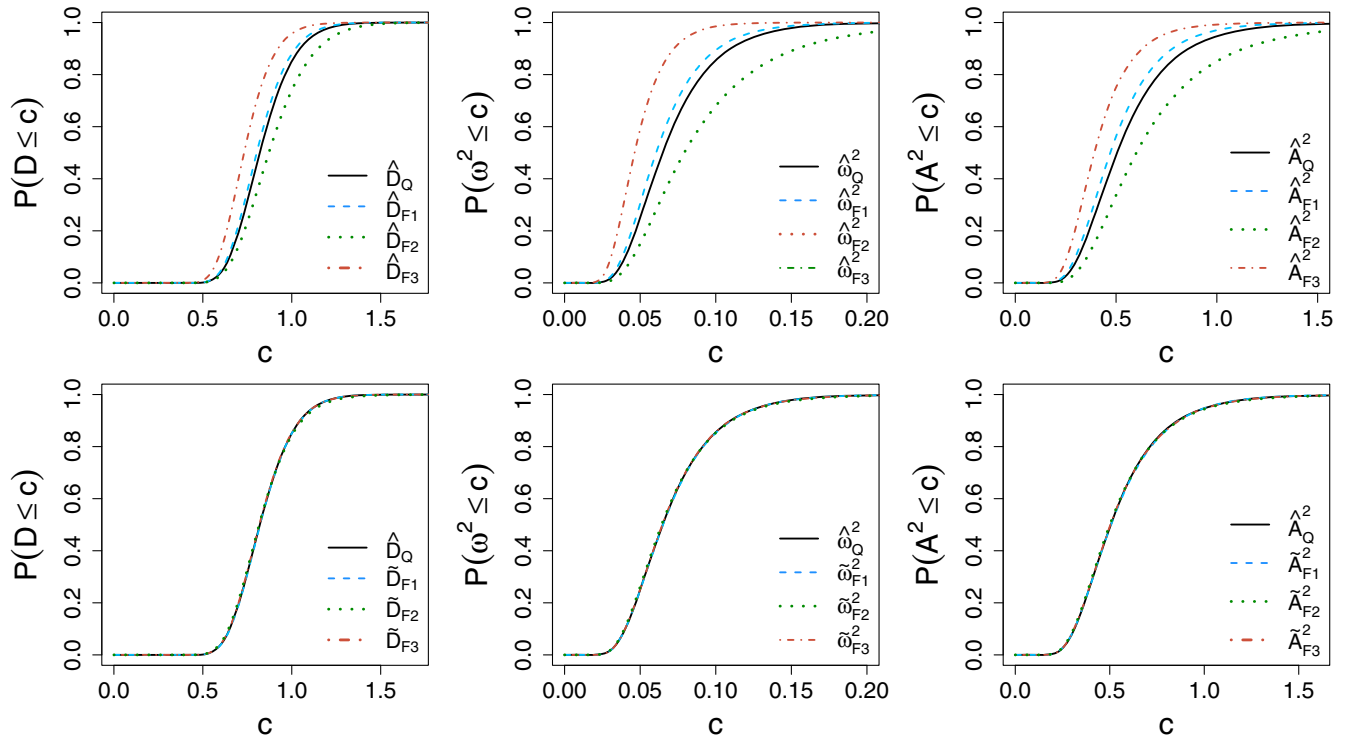

FIG. 2. Upper panels: Comparing the simulated null distributions of the test statistics in (13) for q in (14) and for each candidate model f_m, $m = 1, \ldots, 3$, in (45). Bottom panels: Comparing the simulated null distributions of the test statistics in (13) for q with the K-2 rotated statistics in (43) for each f_m, $m = 1, \ldots, 3$. Each simulation involves 100,000 replicates and 100 observations.

TABLE II. Comparing the power of the test statistics in (13) for each F_m, $m = 1, \ldots, 3$, in (45) with that of the K-2 rotated statistics in (43). The true model from which the data are generated is that in (44). Each simulation involves 100,000 replicates and 100 observations. The significance levels considered are $\alpha = 0.001$ (3.29σ), $\alpha = 0.05$ (1.96σ), and $\alpha = 0.1$ (1.64σ).

	$\alpha = 0.001$						$\alpha = 0.05$						$\alpha = 0.1$					
				$\tilde{D}$	$\tilde{\omega}^2$	$\tilde{A}^2$				$\tilde{D}$	$\tilde{\omega}^2$	$\tilde{A}^2$				$\tilde{D}$	$\tilde{\omega}^2$	$\tilde{A}^2$
Null distribution	$\hat{D}$	$\hat{\omega}^2$	$\hat{A}^2$	(K-2 rotated)			$\hat{D}$	$\hat{\omega}^2$	$\hat{A}^2$	(K-2 rotated)			$\hat{D}$	$\hat{\omega}^2$	$\hat{A}^2$	(K-2 rotated)		
Q	.4773	.7785	.4633	⋯	⋯	⋯	.9331	.9817	.9382	⋯	⋯	⋯	.9679	.9914	.9722	⋯	⋯	⋯
F_1	.3872	.6762	.4815	.1578	1	1	.8623	.9529	.9092	.6971	1	1	.9221	.9748	.9505	.8086	1	1
F_2	.0036	.0025	.0053	.0058	.0226	.0156	.1078	.1019	.1237	.1336	.2422	.2541	.1876	.185	.2127	.2233	.3618	.3770
F_3	.6452	.7947	.0295	.5062	.7975	.6036	.9528	.9820	.6356	.9153	.9746	.9470	.9757	.9915	.7974	.9543	.9874	.9730

shown in the upper panels of Fig. 2. Despite the null distribution of the three statistics under Q and F_1 appear fairly close, as expected, they are substantially different from those of F_2 and F_3. Therefore, in order to achieve distribution-freeness, we consider the test statistics in (43) obtained by implementing Steps 1–4 in Sec. IV. We simulate their null distributions and we compare them with those of $\hat{D}_Q, \hat{\omega}^2_Q$, and A^2_Q, under model Q. The results are shown in bottom panels Fig. 2.

The distributions of the K-2 rotated statistics $\tilde{D}_{F_m}, \tilde{\omega}^2_{F_m}$ and $\tilde{A}^2_{F_m}$, $m = 1, 2, 3$, cannot be distinguished from those of $\hat{D}_Q, \hat{\omega}^2_Q$ and $\hat{A}^2_Q$. Therefore, one can test Q, F_1, F_2, and F_3 by relying solely on the simulated distribution of $\hat{D}_Q, \hat{\omega}^2_Q$ and $\hat{A}^2_Q$, reducing the computational time by a factor of at least three (as we need to perform just one simulation instead of four).

Table II collects the results of a power study. There, we compare the power of the K-2 rotated test statistics in (43) with that of their classical counterparts in (13), and for different significance levels. Interestingly, for model F_2, that is, the closest to the true distribution P among those considered, the power of the K-2 rotated Kolmogorov-Smirnov and Cramer-von Mises statistics is higher compared to that of their nonrotated version. When testing F_1 and F_3, the power decreases for Kolmogorov-Smirnov. The power is comparably high in all the other cases. Notice that the power of the K-2 rotated statistics is not universally higher than their nonrotated counterparts. That is because, the K-2 rotated test statistics are simply new test statistics which may perform better than the classical Kolmogorov-Smirnov, Cramer-von Mises and Anderson Darling in some scenarios, but not in others.

V. FINAL REMARKS

The K-2 transformation is a very powerful tool to achieve distribution-freeness in a simulation-based settings. Researchers can rely on simulations under a simplified model, Q, whose likelihood is easily accessible, and then construct suitable test statistics for one or more complex models F which can be compared with the same simulated distribution.

It is worth emphasizing that the approximation of the null distribution of the statistics in (43) with those of (13) does depend on the sample size. That is because the K-2 transform maps the limiting distribution of the process $\tilde{v}_{F,n}(\boldsymbol{x}, \boldsymbol{\lambda})$ into that of $\tilde{v}_{Q,n}(\boldsymbol{x}, \boldsymbol{\theta})$. In light of this, in order to achieve a good approximation for moderately large samples (e.g., 100 observations), it is recommended to choose Q "sufficiently close to F" so that the entire search region is sampled reasonably often under both Q and F.

To compute the K-2 rotation, one needs to evaluate the score functions of F. In situations where the likelihood is not tractable in closed-form, a possible solution is that of constructing templates for the score, starting from the likelihood templates and applying the definition of derivative. Their evaluation does not need to be repeated on multiple runs, and it is only needed to evaluate the K-2 rotated test statistics on the data observed.

ACKNOWLEDGMENTS

The author thanks an anonymous referee whose feedback has been substantial to improve the overall clarity of the paper.

APPENDIX: DERIVING EQ. (12)

Consider the empirical process

$$v_{Q,n}(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\boldsymbol{x}, \hat{\boldsymbol{\theta}}}(\boldsymbol{x}_i). \tag{A1}$$

and the vectors of derivatives $\dot{\boldsymbol{\psi}}_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t})$ and $\dot{\boldsymbol{q}}_{\boldsymbol{\theta}}(\boldsymbol{t})$ with components

$$\dot{\psi}_{\boldsymbol{x},\boldsymbol{\theta}_j}(\boldsymbol{t}) = \frac{d}{d\theta_j} \psi_x(\boldsymbol{t}) \quad \text{and} \tag{A2}$$

$$\dot{q}_{\boldsymbol{\theta}_j}(\boldsymbol{t}) = \frac{d}{d\theta_j} q_{\boldsymbol{\theta}}(\boldsymbol{t}). \tag{A3}$$

Where,

$$\dot{\psi}_{\boldsymbol{x},\boldsymbol{\theta}_j}(\boldsymbol{t}) = \frac{d}{d\hat{\theta}_j}\psi_{x\boldsymbol{\theta}}(\boldsymbol{t})|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = -\frac{d}{d\theta_j}Q_{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{A4}$$

$$= -\frac{d}{d\theta_j}\int_{-\infty}^{x} q_{\boldsymbol{\theta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t} = -\int_{-\infty}^{x}\dot{q}_{\boldsymbol{\theta}_j}(\boldsymbol{t})\mathrm{d}\boldsymbol{t} \tag{A5}$$

$$= -\int_{-\infty}^{x}\frac{\dot{q}_{\boldsymbol{\theta}_j}(\boldsymbol{t})}{q_{\boldsymbol{\theta}}(\boldsymbol{t})}q_{\boldsymbol{\theta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}. \tag{A6}$$

where the integrals in (A4)–(A6) are all multidimensional. A Taylor expansion of (A1) leads to

$$v_{Q,n}(\boldsymbol{x},\hat{\boldsymbol{\theta}}) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{x}_i) + (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{x}_i) \tag{A7}$$

The asymptotic expansion of $(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ [e.g., [18], p. 53] is

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\Gamma_{\boldsymbol{\theta}}^{-1}\sum_{i=1}^{n}\frac{\dot{\boldsymbol{q}}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{q_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} + o_p(1) \tag{A8}$$

$$= \frac{1}{\sqrt{n}}\Gamma_{\boldsymbol{\theta}}^{-1/2}\sum_{i=1}^{n}\Gamma_{\boldsymbol{\theta}}^{-1/2}\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + o_p(1) \tag{A9}$$

$$= \frac{1}{\sqrt{n}}\Gamma_{\boldsymbol{\theta}}^{-1/2}\sum_{i=1}^{n}\boldsymbol{b}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + o_p(1) \tag{A10}$$

where, as in (8), $\Gamma_{\boldsymbol{\theta}}$ is the Fisher information matrix, and $\boldsymbol{b}_{\boldsymbol{\theta}}(\boldsymbol{x})$ is vector of normalized score functions $b_{\boldsymbol{\theta}_j}(\boldsymbol{x})$. Combining (A6), (A7), (A8), and (A10) we have

$$v_{Q,n}(\boldsymbol{x},\hat{\theta}) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{x}_i) \tag{A11}$$

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{b}_{\boldsymbol{\theta}}^T(\boldsymbol{x}_i)\int_{-\infty}^{x}\boldsymbol{b}_{\boldsymbol{\theta}}(\boldsymbol{t})q_{\boldsymbol{\theta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t} \tag{A12}$$

where the error of the approximation has be shown by Khmaladze [15] to be $o_p(1)$. Moreover, simple algebra can be applied to show that $\langle b_{\boldsymbol{\theta}_j},\psi_{\boldsymbol{x},\boldsymbol{\theta}}\rangle_Q = \int_{-\infty}^{x} b_{\boldsymbol{\theta}_j}(\boldsymbol{t})q_{\boldsymbol{\theta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}$. Specifically,

$$\langle b_{\boldsymbol{\theta}_j},\psi_{\boldsymbol{x},\boldsymbol{\theta}}\rangle_Q = \int_{-\infty}^{\infty} b_{\boldsymbol{\theta}_j}(\boldsymbol{t})\psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{t})q_{\boldsymbol{\theta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t} \tag{A13}$$

$$= \int_{-\infty}^{\infty} b_{\boldsymbol{\theta}_j}(\boldsymbol{t})[\mathbb{1}_{\{\boldsymbol{t}\le\boldsymbol{x}\}} - Q_{\boldsymbol{\theta}}(\boldsymbol{x})]q_{\boldsymbol{\theta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t} \tag{A14}$$

$$= \int_{-\infty}^{x} b_{\boldsymbol{\theta}_j}(\boldsymbol{t})q_{\boldsymbol{\theta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t} - Q_{\boldsymbol{\theta}}(\boldsymbol{x})\int_{-\infty}^{\infty} b_{\boldsymbol{\theta}_j}(\boldsymbol{t})q_{\boldsymbol{\theta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t} \tag{A15}$$

$$= \int_{-\infty}^{x} b_{\boldsymbol{\theta}_j}(\boldsymbol{t})q_{\boldsymbol{\theta}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t} \tag{A16}$$

where (A16) follows from (A15), and the fact that the normalized score vector $\boldsymbol{b}_{\boldsymbol{\theta}}$ has mean zero under $Q_{\boldsymbol{\theta}}$. Finally, combining (A11) and (A13)–(A16), we obtain

$$v_{Q,n}(\boldsymbol{x},\hat{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{x}_i) + o_p(1) \tag{A17}$$

where $\tilde{\psi}_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{x}_i) = \psi_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{x}_i) - \sum_{j=1}^{p} b_{\boldsymbol{\theta}_j}(\boldsymbol{x}_i)\langle b_{\boldsymbol{\theta}_j},\psi_{\boldsymbol{x},\boldsymbol{\theta}}\rangle_{Q_{\boldsymbol{\theta}}}$.

[1] K. Pearson, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, The London, Edinburgh, and Dublin Philos. Mag. J. Sci. **50**, 157 (1900).

[2] A. Kolmogorov, Sulla determinazione empirica di una lgge di distribuzione, Giornale dell'Instituto Italiano degli Attuari **4**, 83 (1933).

[3] N. V. Smirnov, On the estimation of the discrepancy between empirical curves of distribution for two independent samples, Bull. Math. Univ. Moscou **2**, 3 (1939).

[4] S. J. Haberman, A warning on the use of chi-squared statistics with frequency tables with small expected cell counts, J. Am. Stat. Assoc. **83**, 555 (1988).

[5] N. Cressie and T. R. C. Read, Pearson's χ^2 and the log likelihood ratio statistic g^2: A comparative review, Int. Stat. Rev. **57**, 19 (1989).

[6] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. **21**, 1087 (1953).

[7] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika **57**, 97 1970.

[8] E. V. Khmaladze, Martingale approach in the theory of goodness-of-fit tests, Theory Probab. Its Appl. **26**, 240 (1982).

[9] E. Khmaladze, Unitary transformations, empirical processes and distribution free testing, Bernoulli **22**, 563 (2016).

[10] E. Khmaladze, Distribution free testing for conditional distributions given covariates, Stat. Probab. Lett. **129**, 348 (2017).
[11] J. A. Wellner, Empirical processes in action: A review, Int. Stat. Rev. **60**, 247 (1992).
[12] H. Cramér, *Mathematical Methods of Statistics* (Princeton University Press, Princeton, NJ, 1999), Vol. 43.
[13] S. Algeri, J. Aalbers, K. Dundas Morå, and J. Conrad, Searching for new phenomena with profile likelihood ratio tests, Nat. Rev. Phys. **2**, 245 (2020).
[14] N. J. Higham, *Functions of Matrices: Theory and Computation* (SIAM, Philadelphia, 2008).
[15] E. V. Khmaladze, The use of ω^2 tests for testing parametric hypotheses, Theory Probab. Its Appl. **24**, 283 (1980).
[16] G. J. Babu and C. R. Rao, Goodness-of-fit tests when parameters are estimated, Sankhyā: The Indian Journal of Statistics **66**, 63 (2004).
[17] A. C. Davison, *Statistical Models* (Cambridge University Press, Cambridge, England, 2003), Vol. 11.
[18] A. Van der Vaart, *Asymptotic Statistics* (Cambridge University Press, Cambridge, England, 2000), Vol. 3.