Multivariate Goodness-of-Fit Tests Based on Wasserstein Distance

Marc Hallin
ECARES, Université libre de Bruxelles

Gilles Mordant
LIDAM, ISBA, UCLouvain

Johan Segers,
LIDAM, ISBA, UCLouvain

March 2020

ECARES working paper 2020-06

ECARES
ULB - CP 114/04
50, F.D. Roosevelt Ave., B-1050 Brussels BELGIUM
www.ecares.org

Multivariate Goodness-of-Fit Tests Based on Wasserstein Distance

Marc Hallin

ECARES, Université libre de Bruxelles
Avenue F.D. Roosevelt 42, 1050 Brussels, Belgium
e-mail: mhallin@ulb.ac.be

Gilles Mordant and Johan Segers*

LIDAM/ISBA, UCLouvain
Voie du Roman Pays 20/L1.04.01, B-1348 Louvain-la-Neuve, Belgium
e-mail: gilles.mordant@uclouvain.be; johan.segers@uclouvain.be

Abstract: Goodness-of-fit tests based on the empirical Wasserstein distance are proposed for simple and composite null hypotheses involving general multivariate distributions. This includes the important problem of testing for multivariate multivariate normality with unspecified location and covariance and, more generally, testing for elliptical symmetry with given standard radial density, unspecified location and scatter parameters. The calculation of test statistics boils down to solving the well-studied semi-discrete optimal transport problem. Exact critical values can be computed for some important particular cases, such as null hypotheses of ellipticity with given standard radial density and unspecified location and scatter; else, approximate critical values are obtained via parametric bootstrap. Consistency is established, based on a result on the convergence to zero, uniformly over certain families of distributions, of the empirical Wasserstein distance—a novel result of independent interest. A simulation study establishes the practical feasibility and excellent performance of the proposed tests.

Keywords and phrases: Copula, Elliptical distribution, Goodness-of-fit, Multivariate normality, Optimal transport, Semi-discrete problem, Skew-t distribution, Wasserstein distance.

1. Introduction

Wasserstein distances are metrics on spaces of probability measures with certain finite moments. They measure the distance between two such distributions by the minimal cost needed to move probability mass in order to transform one distribution into the other one. Wasserstein distances have a long history and continue to attract interest from diverse fields in statistics, machine learning and computer science, in particular image analysis; see for instance the monographs and reviews by Santambrogio (2015), Peyré and Cuturi (2019), and Panaretos and Zemel (2019).

A natural application of any meaningful distance between distributions is to the goodness-of-fit (GoF) problem—namely, the problem of testing the null

*J. Segers gratefully acknowledges funding from FNRS-F.R.S. grant CDR J.0146.19.

hypothesis that a sample comes from a population with fully specified distribution P_0 or with unspecified distribution within some postulated parametric model $\mathcal{M}$. GoF problems certainly are among the most fundamental and classical ones in statistical inference. Typically, GoF tests are based on some distance between the empirical distribution $\widehat{\mathrm{P}}_n$ and the null distribution P_0 or an estimated distribution in the model $\mathcal{M}$. The most popular ones are the Cramér–von Mises (Cramér, 1928; von Mises, 1928) and Kolmogorov–Smirnov (Kolmogorov, 1933; Smirnov, 1939) tests, involving distances between the cumulative distribution function of P_0 and the empirical one. Originally defined for univariate distributions only, they have been extended to the multivariate case, for instance in Khmaladze (2016), who proposes a test that has nearly all properties one could wish for, including asymptotic distribution-freeness, but whose implementation is computationally quite heavy and quickly gets intractable.

Many other distances have been considered in this context, though. Among them, distances between densities (after kernel smoothing) have attracted much interest, starting with Bickel and Rosenblatt (1973) in the univariate case. Bakshaev and Rudzkis (2015) recently proposed a multivariate extension; the choice of a bandwidth matrix, however, dramatically affects the outcome of the resulting testing procedure. Fan (1997) considers a distance between characteristic functions, which accommodates arbitrary dimensions; the idea is appealing but the estimation of the integrals involved in the distance seems tricky. McAssey (2013) proposes a heuristic test that relies on a comparison of the empirical Mahalanobis distance with a simulated one under the null. Still in a multivariate setting, Ebner, Henze and Yukich (2018) define a distance based on sums of powers of weighted volumes of kth nearest neighbour spheres.

The use of the Wasserstein distance for GoF testing has been considered mostly for univariate distributions (Munk and Czado, 1998; del Barrio et al., 1999; del Barrio et al., 2000; del Barrio, Giné and Utzet, 2005). For the multivariate case, available methods are restricted to discrete distributions (Sommerfeld and Munk, 2018) and Gaussian ones (Rippl, Munk and Sturm, 2016). Indeed, serious difficulties, both computational and theoretical, hinder the development of Wasserstein GoF tests for general multivariate continuous distributions, particularly in the case of composite null hypotheses. Composite null hypotheses are generally more realistic than simple ones. Of particular practical importance is the case of location–scale families: tests of multivariate Gaussianity, tests of elliptical symmetry with given standard radial density, etc., belong to that type. Although the asymptotic null distribution of empirical processes with estimated parameters is well known (van der Vaart, 1998, Theorem 19.23), the actual exploitation of that theory in GoF testing remains problematic because of the difficulty of computing critical values.

The aim of this paper is to explore the potential of the Wasserstein distance for GoF tests of simple (consisting of one fully specified distribution) and composite (consisting of a parametric family of distributions) null hypotheses involving continuous multivariate distributions. The tests we are proposing are based on the Wasserstein distance between $\widehat{\mathrm{P}}_n$ and the distribution P_0 in the case of a simple null hypothesis and on the Wasserstein distance between $\widehat{\mathrm{P}}_n$ and a

model-based distribution estimate in the case of a composite null hypothesis. They are computationally feasible, have the correct size, and enjoy good power properties in comparison with other tests available in the literature.

We concentrate on the continuous case, i.e., the distributions under the null hypothesis are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. The test statistic involves the Wasserstein distance between $\widehat{\mathrm{P}}_n$, which is discrete, and a distribution from the null hypothesis to be tested, which is continuous. Calculating such a distance requires solving the so-called semi-discrete transportation problem, an active area of research in computer science.

In case of a simple null hypothesis, the null distribution of the test statistic does not depend on unknown parameters. Exact critical values can be calculated with arbitrary precision via a Monte Carlo procedure, by simulating from the null distribution and computing empirical quantiles. Exact critical values can also be computed for Wasserstein tests for the GoF of a location–scatter family of elliptical distributions with known radial distribution. We handle the presence of unknown nuisance parameters by using empirically standardized data. An important and well-studied special case is that of testing for multivariate normality. Out of the many available tests in the literature, we select the ones of Royston (1982), Henze and Zirkler (1990) and Rizzo and Székely (2016) as benchmark for our Wasserstein test.

For general parametric models, we rely on the bootstrap to calculate critical values. The question whether the method has the correct size under the null hypothesis remains open. A proof of that property would require knowledge of non-degenerate limit distributions of the empirical Wasserstein distance—a hard and long-standing open problem, which we briefly review in Section 1.2. Monte Carlo experiments, however, suggest that our tests have the correct asymptotic size.

In all cases, we show that our Wasserstein GoF tests are consistent against fixed alternatives, that is, the null hypothesis under such alternatives is rejected with probability tending to one. For the general parametric case, this property relies on the uniform consistency in probability of the empirical distribution with respect to the Wasserstein distance, uniformly over adequate classes of probability measures. To the best of our knowledge, this result, which is of independent interest, is new in the literature.

Measure transportation has attracted much interest in the recent statistical literature. Carlier et al. (2016), Chernozhukov et al. (2017) and del Barrio et al. (2018) propose measure transportation-based concepts of multivariate ranks, signs, and quantiles. These notions have been successfully applied by Shi, Drton and Han (2019), Deb and Sen (2019), and Ghosal and Sen (2019) in the construction of distribution-free tests in a multivariate context, by Hallin, La Vecchia and Liu (2019) for R-estimation of VARMA models with unspecified innovation densities.

The outline of the paper is as follows. In the remainder of this introduction, we introduce the Wasserstein distance (Section 1.1), review the asymptotic theory of empirical Wasserstein distance (Section 1.2), and provide some information on the computational methods for the semi-discrete problem underlying the

implementation of the Wasserstein GoF tests (Section 1.3). In Section 2, we give a formal description of the GoF test procedure for simple null hypotheses. Section 3 addresses the composite null hypothesis that the unknown distribution belongs to an elliptical family with unknown location and scatter (covariance) parameters and known radial distribution; the multivariate normal family is an important special case. Composite null hypotheses covering general parametric models are treated in Section 4. In Section 5, we conduct a simulation study to assess the finite-sample performance of the Wasserstein tests in comparison to other GoF tests, both for simple and composite null hypotheses. In Appendix A, the convergence of the empirical Wasserstein distance uniformly over certain classes of underlying distributions is stated and proved. In Appendix B, the algorithms we are using in the computation of critical values are listed and explained.

1.1. Wasserstein distance

Let $\mathcal{P}(\mathbb{R}^d)$ be the set of Borel probability measures on $\mathbb{R}^d$ and let $\mathcal{P}_p(\mathbb{R}^d)$ be the subset of such measures with a finite moment of order $p \in [1, \infty)$. For $\mathrm{P}, \mathrm{Q} \in \mathcal{P}(\mathbb{R}^d)$, let $\Gamma(\mathrm{P}, \mathrm{Q})$ be the set of probability measures γ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals P and Q, i.e., such that $\gamma(B \times \mathbb{R}^d) = \mathrm{P}(B)$ and $\gamma(\mathbb{R}^d \times B) = \mathrm{Q}(B)$ for Borel sets $B \subseteq \mathbb{R}^d$. The p-Wasserstein distance $W_p(\mathrm{P}, \mathrm{Q})$ between $\mathrm{P}, \mathrm{Q} \in \mathcal{P}_p(\mathbb{R}^d)$ is

$$W_p(\mathrm{P}, \mathrm{Q}) := \left(\inf_{\gamma \in \Gamma(\mathrm{P}, \mathrm{Q})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, \mathrm{d}\gamma(x, y) \right)^{1/p},$$

with $\|\cdot\|$ the Euclidean norm. In terms of random variables $X \sim \mathrm{P}$ and $Y \sim \mathrm{Q}$, the p-Wasserstein distance is the smallest value of $\{\mathbb{E}(\|X - Y\|^p)\}^{1/p}$ over all possible couplings $(X, Y) \sim \gamma$.

The p-Wasserstein distance W_p defines a metric on $\mathcal{P}_p(\mathbb{R}^d)$ which, when endowed with the Wasserstein distance W_p, is a complete separable metric space (Villani, 2009, Theorem 6.18 and the bibliographical notes). Convergence in the W_p metric is equivalent to weak convergence plus convergence of moments of order p; see for instance Bickel and Freedman (1981, Lemmas 8.1 and 8.3) and Villani (2009, Theorem 6.9). It is thus quite natural to consider W_p in the construction of GoF tests for multivariate distributions.

For univariate distributions P and Q with distribution functions F and G, the p-Wasserstein distance boils down to the L^p-distance

$$W_p(\mathrm{P}, \mathrm{Q}) = \left(\int_0^1 \left| F^{-1}(u) - G^{-1}(u) \right|^p \mathrm{d}u \right)^{1/p}$$

between the respective quantile functions F^{-1} and G^{-1}. This representation considerably facilitates both the computation of the distance and the asymptotic theory of its empirical versions. Also, the optimal transport plan mapping $X \sim \mathrm{P}$ to $Y \sim \mathrm{Q}$ is immediate: if F has no atoms, then $Y := G^{-1} \circ F(X) \sim \mathrm{Q}$, while monotonicity of $G^{-1} \circ F$ implies the optimality of the coupling (X, Y).

1.2. Asymptotic theory

Let $\mathbf{X}_n = (X_1, \ldots, X_n)$ be an independent random sample from $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d)$. Its distribution as a random vector in $(\mathbb{R}^d)^n$ is the n-fold product P^n of P with itself. Let $L_n : (\mathbb{R}^d)^n \to \mathcal{P}(\mathbb{R}^d)$ map $\mathbf{x}_n = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ to the discrete probability measure $L_n(\mathbf{x}_n) := n^{-1} \sum_{i=1}^n \delta_{x_i}$, with δ_x the Dirac measure at x. The empirical distribution of the sample is $\widehat{\mathrm{P}}_n := L_n(\mathbf{X}_n) = n^{-1} \sum_{i=1}^n \delta_{X_i}$. We study its distribution as a random element in $\mathcal{P}(\mathbb{R}^d)$.

The Wasserstein distance between the empirical distribution $L_n(\mathbf{x}_n)$ and a probability measure $\mathrm{P} \in \mathcal{P}_p(\mathbb{R}^d)$ is the value at $\mathbf{x}_n \in (\mathbb{R}^d)^n$ of the map

$$W_p(L_n, \mathrm{P}) : \mathbf{x}_n \in (\mathbb{R}^d)^n \mapsto W_p(L_n(\mathbf{x}_n), \mathrm{P}) \in [0, \infty).$$

Consider the distribution of this map under P^n, i.e., for an independent random sample of size n from P. In perhaps more familiar notation, the random variable of interest is the empirical Wasserstein distance $W_p(\widehat{\mathrm{P}}_n, \mathrm{P})$.

According to Bickel and Freedman (1981, Lemma 8.4), if $\mathrm{P} \in \mathcal{P}_p(\mathbb{R}^d)$, the empirical distribution is strongly consistent in the Wasserstein distance: for an i.i.d. sequence $X_1, X_2, \ldots$ with common distribution P, we have $W_p(\widehat{\mathrm{P}}_n, \mathrm{P}) \to 0$ almost surely as $n \to \infty$. The corresponding consistency rates have been studied intensively; see Panaretos and Zemel (2019, Section 3.3) for a review. If P is non-degenerate, then $\mathbb{E}[W_p(\widehat{\mathrm{P}}_n, \mathrm{P})]$ is at least of the order $n^{-1/2}$, and if P is absolutely continuous, which is the case of interest here, the convergence rate cannot be faster than $n^{-1/d}$. Actually, the rate can be arbitrarily slow (Bobkov and Ledoux, 2019, Theorem 3.3). Precise rates under additional moment assumptions are given for instance in Fournier and Guillin (2015).

Asymptotic distribution results for the empirical Wasserstein distance in dimension $d \geq 2$ are, however, surprisingly scarce. The one-dimensional case is well-studied thanks to the link to empirical quantile processes, see for instance del Barrio, Giné and Utzet (2005). Also for discrete distributions, non-degenerate limit distributions are known (Sommerfeld and Munk, 2018; Tameling, Sommerfeld and Munk, 2019). For multivariate Gaussian distributions, a central limit theorem for the empirical Wasserstein between the true normal distribution and the one with estimated parameters is given in Rippl, Munk and Sturm (2016). Although interesting and useful for GoF testing (see Section 5.1 below), this result does not cover the case of the empirical distribution $\widehat{\mathrm{P}}_n$.

Important steps have been taken recently by del Barrio and Loubes (2019) who, quite remarkably, manage to obtain some asymptotic results under alternatives. For general $\mathrm{P}, \mathrm{Q} \in \mathcal{P}_{4+\delta}(\mathbb{R}^d)$ for some $\delta > 0$, they establish a central limit theorem for

$$n^{1/2} \Big[W_2(\widehat{\mathrm{P}}_n, \mathrm{Q}) - \mathbb{E}\{W_2(\widehat{\mathrm{P}}_n, \mathrm{Q})\} \Big].$$

Unfortunately, if $\mathrm{Q} = \mathrm{P}$, which is the case of interest here, the asymptotic variance is zero, meaning that the random fluctuations of $W_2(\widehat{\mathrm{P}}_n, \mathrm{P})$ around its mean are of order less than $n^{-1/2}$. Moreover, as mentioned above, $\mathbb{E}\{W_p(\widehat{\mathrm{P}}_n, \mathrm{P})\}$ may converge to zero at a slower rate than $n^{-1/2}$. The crucial problem of the

limiting distribution of the empirical Wasserstein distance under the null so far remains an important and difficult open problem, which apparently precludes the implementation of multivariate analogues of the existing one-dimensional procedures. The most recent progress perhaps has been booked in Goldfeld and Kato (2020), who obtain a central limit theorem for the empirical W_1-distance after smoothing the empirical and true distributions with a Gaussian kernel.

To construct critical values of Wasserstein GoF tests of general parametric models, we will propose in Section 4 the use of the parametric bootstrap. In general, proving consistency of the parametric bootstrap typically requires having non-degenerate limit distributions under contiguous alternatives of the statistic of interest (Beran, 1997; Capanu, 2019). As the above review shows, such results are still beyond the horizon.

1.3. Computational issues

Important numerical developments have taken place recently in the area of measure transportation and, more particularly, in the computation of the 2-Wasserstein distance between a discrete and a continuous distribution, the so-called semi-discrete optimal transportation problem; see for instance Leclaire and Rabin (2019) and the references therein. The efficiency and high accuracy of the algorithms developed by Mérigot (2011), Lévy (2015), or Kitagawa, Mérigot and Thibert (2017) make it possible to simulate from the exact null distribution of empirical Wasserstein distances. Moreover, Kitagawa, Mérigot and Thibert (2017) establish, under certain assumptions, the convergence of their algorithm. This opens the door for the implementation, based on simulated critical values, of the Wasserstein distance-based GoF tests in dimension $d \geq 1$ for which asymptotic critical values remain unavailable.

Most algorithms to date rely on the dual formulation of the problem, assuming that the source continuous probability measure P admits a density f w.r.t. the Lebesgue measure on $\mathbb{R}^d$. We follow Santambrogio (2015, Section 6.4.2) for a brief exposition. In line with the set-up relying on the empirical measure, we work with the quadratic cost function ($p = 2$) and a discrete measure $\mathrm{P}_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$ over n distinct atoms $x_1, \ldots, x_n \in \mathbb{R}^d$, each of mass $1/n$.

The semi-discrete problem requires constructing a power diagram or Laguerre–Voronoi diagram, partitioning $\mathbb{R}^d$ into power cells

$$V_\psi(i) := \left\{ x \in \mathbb{R}^d \ : \ \tfrac{1}{2}\|x - x_i\|^2 - \psi_i \leq \tfrac{1}{2}\|x - x_j\|^2 - \psi_j, \ \forall j = 1, \ldots, n \right\}$$

for $i = 1, \ldots, n$, defined in terms of a vector $\psi = (\psi_1, \ldots, \psi_n) \in \mathbb{R}^n$. Each power cell $V_\psi(i)$ corresponds to a set of linear constraints and, therefore, is a convex polyhedron. The dual to the problem of minimizing the expected transportation cost $\int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{1}{2}\|x - y\|^2 \, \mathrm{d}\gamma(x, y)$ over the couplings $\gamma \in \Gamma(\mathrm{P}_n, \mathrm{P})$ is then the maximisation, with respect to the vector ψ, of the objective function

$$F(\psi) := \frac{1}{n} \sum_{i=1}^n \psi_i + \sum_{i=1}^n \int_{V_\psi(i)} \left(\tfrac{1}{2}\|x - x_i\|^2 - \psi_i \right) f(x) \, \mathrm{d}x.$$

The function $\psi \mapsto F(\psi)$ is differentiable. Setting its partial derivatives to zero yields the equations

$$\int_{V_\psi(i)} f(x)\,\mathrm{d}x = \frac{1}{n}, \qquad i = 1, \ldots, n,$$

specifying that each power cell $V_\psi(i)$ should receive mass $\mathrm{P}_n(\{x_i\}) = n^{-1}$ under P. The optimal transport plan from P to P_n then consists in mapping all points in the interior of $V_\psi(i)$ to x_i.

The dual formulation above is the basis for the multi-scale algorithm developed in Mérigot (2011) based on the method for solving constrained least-squares assignment problems in Aurenhammer, Hoffmann and Aronov (1998). For the Monte Carlo simulation experiments, we use the implementation of that algorithm in the function semidiscrete provided by the R package transport (Schuhmacher et al., 2019).

Further improvements of the multi-scale algorithm are introduced in Lévy (2015) and Kitagawa, Mérigot and Thibert (2017). Recently, stochastic algorithms in Genevay et al. (2016) and Leclaire and Rabin (2019) are claimed to perform even better. To the best of our knowledge, implementations of these algorithms are not yet available in R (R Core Team, 2018), the language in which we programmed the simulation experiments. Our aim in this paper is to demonstrate the feasability of goodness-of-fit tests for multivariate distributions based on the Wasserstein distance. Advances in computational methods and software implementations can only strengthen that case.

2. Wasserstein GoF tests for simple null hypotheses

Let $\mathbf{X}_n = (X_1, \ldots, X_n)$ be an independent random sample from some unknown distribution $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d)$. For some given fixed $\mathrm{P}_0 \in \mathcal{P}_p(\mathbb{R}^d)$, consider testing the simple null hypothesis

$$\mathcal{H}_0^n : \mathrm{P} = \mathrm{P}_0 \qquad \text{against} \qquad \mathcal{H}_1^n : \mathrm{P} \neq \mathrm{P}_0$$

based on the observations $\mathbf{X}_n$. Note that P, under the alternative, is not required to have finite moments of order p.

Consider the test statistic $T_n := W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P}_0)$, the pth power of the p-Wasserstein distance between the empirical distribution $\widehat{\mathrm{P}}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ and the distribution P_0 specified by the null hypothesis. Having bounded support, $\widehat{\mathrm{P}}_n$ trivially belongs to $\mathcal{P}_p(\mathbb{R}^d)$, so that T_n is well-defined.

Actual computation of T_n amounts to solving the semi-discrete optimal transport problem, as reviewed in Section 1.3 for $p = 2$. In the simulations of Section 5, we therefore limit ourselves to $p = 2$; the theory developed in this section, however, is developed for general $p \geq 1$.

For $0 < \alpha < 1$, the test $\phi_{\mathrm{P}_0}^n$ we are proposing has the form

$$\phi_{\mathrm{P}_0}^n = \begin{cases} 1 & \text{if } T_n > c(\alpha, n, \mathrm{P}_0), \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

with critical value

$$c(\alpha, n, \mathrm{P}_0) := \inf\left\{c > 0 : \mathrm{P}_0^n[T_n > c] \leq \alpha\right\} \tag{2}$$

where P_0^n stands for n-fold product measure of P_0 on $(\mathbb{R}^d)^n$, that is, the distribution under $\mathcal{H}_0^n$ of the observation $\mathbf{X}_n$. By construction, the exact size of the GoF test in (1) is

$$\mathrm{P}_0^n\left[T_n > c(\alpha, n, \mathrm{P}_0)\right] \leq \alpha,$$

with equality if the law of T_n under P_0^n is continuous. The risk of a false rejection is thus bounded by the nominal size α, and often equal to it.

Although the critical level $c(\alpha, n, \mathrm{P}_0)$ cannot be calculated analytically, its value can be approximated to any desired degree of precision via Monte Carlo simulation. To this end, draw a large number N, say, of independent random samples of size n from P_0 and compute the test statistic for each such sample. The empirical $(1-\alpha)$ quantile of the N simulated test statistics thus obtained is then a consistent and asymptotically normal estimator of $c(\alpha, n, \mathrm{P}_0)$ as $N \to \infty$ provided that the distribution of T_n has a continuous and positive density at $c(\alpha, n, \mathrm{P}_0)$. The approximation error is of the order $N^{-1/2}$ and can be made arbitrarily small by choosing N sufficiently large. The null distribution of T_n depends on P_0, so that $c(\alpha, n, \mathrm{P}_0)$ needs to be calculated for each P_0 separately.

Under the alternative hypothesis, the following proposition establishes that the test is rejecting the null with probability tending to one, i.e., is consistent against any fixed alternative $\mathrm{P} \neq \mathrm{P}_0$.

Proposition 1 (Consistency)**.** *For every* $\mathrm{P}_0 \in \mathcal{P}_p(\mathbb{R}^d)$*, the test* $\phi_{\mathrm{P}_0}^n$ *is consistent against any* $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d)$ *with* $\mathrm{P} \neq \mathrm{P}_0$*:*

$$\lim_{n\to\infty} \mathrm{P}^n[\phi_{\mathrm{P}_0}^n = 1] = 1 \qquad \textit{for any } \alpha > 0.$$

Proof. Fix $\mathrm{P}_0 \in \mathrm{P}_p(\mathbb{R}^d)$. For any $\alpha > 0$, the critical value $c(\alpha, n, \mathrm{P}_0)$ tends to zero as $n \to \infty$. Indeed, by Bickel and Freedman (1981, Lemma 8.4), we have $T_n \to 0$ in P_0^n-probability and thus $\lim_{n\to\infty} \mathrm{P}_0^n[T_n > \varepsilon] = 0$ for any $\varepsilon > 0$. It follows that, for every $\alpha > 0$ and every $\varepsilon > 0$, we have $c(\alpha, n, \mathrm{P}_0) \leq \varepsilon$ for all sufficiently large n.

Let $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d)$ with $\mathrm{P} \neq \mathrm{P}_0$. We consider two cases according as P has finite moments of order p or not.

First, suppose that $\mathrm{P} \in \mathcal{P}_p(\mathbb{R}^d)$. Still by Bickel and Freedman (1981, Lemma 8.4), we have $W_p(\widehat{\mathrm{P}}_n, \mathrm{P}) \to 0$ in P^n-probability as $n \to \infty$. The triangle inequality for the metric W_p yields

$$\left|W_p(\widehat{\mathrm{P}}_n, \mathrm{P}_0) - W_p(\mathrm{P}, \mathrm{P}_0)\right| \leq W_p(\widehat{\mathrm{P}}_n, \mathrm{P}) \to 0, \qquad n \to \infty$$

in P^n-probability. Hence $T_n = W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P}_0) \to W_p^p(\mathrm{P}, \mathrm{P}_0)$ in P^n-probability as $n \to \infty$. But $W_p^p(\mathrm{P}, \mathrm{P}_0) > 0$ since $\mathrm{P}, \mathrm{P}_0 \in \mathcal{P}_p(\mathbb{R}^d)$ and $\mathrm{P} \neq \mathrm{P}_0$ by assumption. It follows that $\lim_{n\to\infty} \mathrm{P}^n[T_n > c(\alpha, n, \mathrm{P}_0)] = 1$, as required.

Second, suppose that $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d) \setminus \mathcal{P}_p(\mathbb{R}^d)$. Let δ_0 denote the Dirac measure at $0 \in \mathbb{R}^d$. Since W_p is a metric, the triangle inequality implies

$$W_p(\widehat{\mathrm{P}}_n, \mathrm{P}_0) \geq W_p(\widehat{\mathrm{P}}_n, \delta_0) - W_p(\mathrm{P}_0, \delta_0).$$

Now, $W_p(\mathrm{P}_0, \delta_0)$ is a constant and $W_p^p(\widehat{\mathrm{P}}_n, \delta_0) = n^{-1} \sum_{i=1}^n \|X_i\|^p$. As the expectation of $\|X_1\|^2$ under P is infinite, the law of large numbers implies that $W_p^p(\widehat{\mathrm{P}}_n, \delta_0) \to \infty$ in P^n-probability as $n \to \infty$. But the same then is true for T_n and thus

$$\lim_{n\to\infty} \mathrm{P}^n[T_n > c(\alpha, n, \mathrm{P}_0)] = 1,$$

as required. □

3. Wasserstein GoF tests for elliptical families

The distribution $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d)$ of a d-dimensional random vector Z with density f is called *spherical with radial density* f_{rad} if $f(z)$ is of the form $f_{\mathrm{rad}}(\|z\|)$ for $z \in \mathbb{R}^d$ where $\int_0^\infty f_{\mathrm{rad}}(r)\,\mathrm{d}r = 1$. The radial density f_{rad} is called *standard* if $\int_0^\infty r^2 f_{\mathrm{rad}}(r)\,\mathrm{d}r = d$. The distribution P is then in $\mathcal{P}_2(\mathbb{R}^d)$—denote it by $\mathrm{P}_{f_{\mathrm{rad}}}$—and Z has mean zero and covariance matrix I_d.

The distribution $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d)$ of a d-dimensional random vector X is called *elliptical with standard radial density* f_{rad} if there exist $\mu \in \mathbb{R}^d$ and a full-rank $d \times d$ matrix A such that the distribution of $A^{-1}(X - \mu)$ is spherical with radial density f_{rad} satisfying $\int r^2 f_{\mathrm{rad}}(r)\mathrm{d}r = d$; the distribution P then is in $\mathcal{P}_2(\mathbb{R}^d)$ and X has mean μ and covariance matrix $\Sigma = AA'$. We refer to Cambanis, Huang and Simons (1981) or Fang, Kotz and Ng (1990) for details.

Let $\mathcal{E}(f_{\mathrm{rad}})$ denote the family of d-variate elliptical distributions with standard radial density f_{rad}. Such families are indexed by a location vector $\mu \in \mathbb{R}^d$ and a positive definite $d \times d$ covariance matrix Σ; the choices $\mu = 0$ and $\Sigma = I_d$ yield the spherical $\mathrm{P}_{f_{\mathrm{rad}}}$. Common examples of elliptical families are the multivariate normal family, with f_{rad} the density of the root of a χ^2_d variable, and the multivariate Student t distribution with $\nu > 2$ degrees of freedom, where f_{rad} is the density of the root of a rescaled Fisher $F(d, \nu)$ variable. In general, elliptical distributions are not subject to moment constraints ($\Sigma := AA'$ is then a scatter rather than a covariance matrix), but here we intend to use the Wasserstein distance of order $p = 2$ and therefore restrict to elliptical families with finite second-order moments.

Given an i.i.d. sample $X_1, \ldots, X_n$ from some unspecified $\mathrm{P} \in \mathcal{P}_2(\mathbb{R}^d)$, we wish to test the null hypothesis that P is elliptical with specified standard radial density f_{rad}, namely,

$$\mathcal{H}_0^n : \mathrm{P} \in \mathcal{E}(f_{\mathrm{rad}}) \quad \text{against} \quad \mathcal{H}_1^n : \mathrm{P} \notin \mathcal{E}(f_{\mathrm{rad}}). \tag{3}$$

The location vector μ and the covariance matrix Σ of P are unknown nuisance parameters. In contrast to Section 2, the null hypothesis is thus a composite one.

Our testing strategy is to compute residuals of the form

$$\hat{Z}_{n,i} := \hat{A}_n^{-1}(X_i - \hat{\mu}_n), \qquad i = 1, \ldots, n, \tag{4}$$

yielding an empirical distribution $\widehat{\mathrm{P}}_n^{\hat{Z}} := n^{-1} \sum_{i=1}^n \delta_{\hat{Z}_{n,i}}$. The test statistic we propose is

$$T_{\mathcal{E}(f_{\mathrm{rad}}),n} := W_2^2(\widehat{\mathrm{P}}_n^{\hat{Z}}, \mathrm{P}_{f_{\mathrm{rad}}}).$$

If the null distribution of $(\hat{Z}_{n,1}, \ldots, \hat{Z}_{n,n})$ does not depend on the unkown μ and Σ, then we can define critical values for $T_{\mathcal{E}(f_{\mathrm{rad}}),n}$ as if $\mu = 0$ and $\Sigma = I_d$. As in Section 2, such critical values can then be approximated with any desired accuracy via Monte Carlo random sampling from $\mathrm{P}_{f_{\mathrm{rad}}}$.

In the sequel, we let $\hat{\mu}_n = n^{-1} \sum_{i=1}^n X_i = \overline{X}_n$ and choose for $\hat{A}_n$ the Cholesky triangle $L_{n,X} \in \mathbb{R}^{d \times d}$ of the empirical covariance matrix

$$S_{n,X} := \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)'.$$

Recall that for every symmetric positive definite matrix $S \in \mathbb{R}^{d \times d}$, there exists a unique lower triangular matrix $L \in \mathbb{R}^{d \times d}$ with positive diagonal elements, called Cholesky triangle, producing the Cholesky decomposition $S = LL'$ (Golub and Van Loan, 1996, Theorem 4.2.5). If Σ is invertible, then $S_{n,X}$ is invertible with probability tending to one; even more, for an i.i.d. sequence $X_1, X_2, \ldots$ from $\mathrm{P} \in \mathcal{E}(f_{\mathrm{rad}})$, with probability one, the matrix $S_{n,X}$ is invertible for all n large enough depending on the sample. On the event that $S_{n,X}$ is invertible, the residuals (4) are thus

$$\hat{Z}_{n,i} = L_{n,X}^{-1}(X_i - \overline{X}_n), \qquad i = 1, \ldots, n. \tag{5}$$

For completeness, on the event that $S_{n,X}$ is not invertible, we set $\hat{Z}_{n,i} = 0$ for $i = 1, \ldots, n$, although a precise definition is immaterial for the results to follow.

Let us show that the joint distribution of the vector of residuals computed in this way does not depend on the unknown μ or Σ. The key is the following elementary property.

Lemma 1. *Let $x_1, \ldots, x_n \in (\mathbb{R}^d)^n$ have mean vector $\overline{x}_n = n^{-1} \sum_{i=1}^n x_i$ and covariance matrix $S_{n,x} = (n-1)^{-1} \sum_{i=1}^n (x_i - \overline{x}_n)(x_i - \overline{x}_n)' \in \mathbb{R}^{d \times d}$. Let $\mu \in \mathbb{R}^d$ and let $L \in \mathbb{R}^{d \times d}$ be lower triangular with positive diagonal elements. Put*

$$z_i := L^{-1}(x_i - \mu), \qquad i = 1, \ldots, n;$$

with obvious notation, similarly define $\overline{z}_n$ and $S_{n,z}$. Then, $S_{n,x}$ is invertible if and only $S_{n,z}$ is, in which case their Cholesky factors $L_{n,x}$ and $L_{n,z}$ are related by $L_{n,x} = LL_{n,z}$, which implies

$$L_{n,x}^{-1}(x_i - \overline{x}_n) = L_{n,z}^{-1}(z_i - \overline{z}_n), \qquad i = 1, \ldots, n.$$

Proof. We have $x_i = \mu + Lz_i$ for all $i = 1, \dots, n$, whence

$$\overline{x}_n = \mu + L\overline{z}_n \quad \text{and} \quad S_{n,x} = LS_{n,z}L'.$$

Since L is invertible, $S_{n,x}$ is invertible if and only if $S_{n,z}$ is. Suppose they both are, and let $L_{n,x}$ and $L_{n,z}$ denote their Cholesky factors. The matrix $LL_{n,z}$ is lower triangular, has positive diagonal elements, and satisfies

$$LL_{n,z}(LL_{n,z})' = LS_{n,z}L' = S_{n,x}.$$

By the uniqueness of the Cholesky decomposition, $L_{n,x} = LL_{n,z}$. Finally,

$$L_{n,x}^{-1}(x_i - \overline{x}_n) = (LL_{n,z})^{-1}\{(\mu + Lz_i) - (\mu + L\overline{z}_n)\} = L_{n,z}^{-1}(z_i - \overline{z}_n), \quad i = 1, \dots, n.$$

□

Proposition 2. *Let $X_1, \dots, X_n$ be an i.i.d. sample from $\mathrm{P} \in \mathcal{E}(f_{\text{rad}})$ with mean μ and full-rank covariance Σ. The joint distribution of $\hat{Z}_{n,i}$ in (5) for $i = 1, \dots, n$ does not depend on μ nor Σ.*

Proof. Let L be the Cholesky factor of Σ. In view of Lemma 1, it is sufficient to show that $Z_i = L^{-1}(X_i - \mu)$, for $i = 1, \dots, n$, is an independent random sample of the spherical distribution $\mathrm{P}_{f_{\text{rad}}}$ with mean zero, covariance identity, and standard radial density f_{rad}. By the assumption on P, there exists an invertible $A \in \mathbb{R}^{d \times d}$ with $AA' = \Sigma$ such that $X_i = \mu + A\zeta_i$ for $i = 1, \dots, n$, where $\zeta_1, \dots, \zeta_n$ is an independent random sample from $\mathrm{P}_{f_{\text{rad}}}$. Then, $Z_i = L^{-1}A\zeta_i$ for all $i = 1, \dots, n$, where the matrix $L^{-1}A$ is orthogonal: indeed,

$$(L^{-1}A)(L^{-1}A)' = L^{-1}AA'(L')^{-1} = L^{-1}\Sigma(L')^{-1} = L^{-1}LL'(L')^{-1} = I_d.$$

It thus follows from sphericity that the common distribution of $Z_1, \dots, Z_n$ is the same as that of $\zeta_1, \dots, \zeta_n$, that is, $\mathrm{P}_{f_{\text{rad}}}$. □

For the hypothesis testing problem (3), we propose the test

$$\phi^n_{\mathcal{E}(f_{\text{rad}})} := \begin{cases} 1 & \text{if } T_{\mathcal{E}(f_{\text{rad}}),n} > c_{\mathcal{E}}(\alpha, n, f_{\text{rad}}), \\ 0 & \text{otherwise,} \end{cases}$$

at level $\alpha \in (0, 1)$ and with critical value

$$c_{\mathcal{E}}(\alpha, n, f_{\text{rad}}) = \inf\left\{c > 0 : \mathrm{P}^n_{f_{\text{rad}}}[T_{\mathcal{E}(f_{\text{rad}}),n} > c] \leq \alpha\right\}. \tag{6}$$

The probability in (6) is calculated under the spherical distribution with radial density f_{rad}, which is free of nuisances. By Proposition 2, the size of the test is at most α: for all $\mathrm{P} \in \mathcal{E}(f_{\text{rad}})$,

$$\mathrm{P}^n[T_{\mathcal{E}(f_{\text{rad}}),n} > c_{\mathcal{E}}(\alpha, n, f_{\text{rad}})] = \mathrm{P}^n_{f_{\text{rad}}}[T_{\mathcal{E}(f_{\text{rad}}),n} > c_{\mathcal{E}}(\alpha, n, f_{\text{rad}})] \leq \alpha.$$

The size of the test is equal to α if the null distribution of $T_{\mathcal{E}(f_{\text{rad}}),n}$ is continuous at the critical value. In practice, calculation of the critical value is implemented by Monte Carlo simulation, see Algorithm 2 in Appendix B.

The consistency of the test follows from a large of law numbers in 2-Wasserstein distance for the empirical distribution of the residuals defined in (5).

Proposition 3. *Let* $\mathrm{P} \in \mathcal{P}_2(\mathbb{R}^d)$ *have mean vector* $\mu \in \mathbb{R}^d$ *and invertible covariance matrix* $\Sigma \in \mathbb{R}^{d\times d}$ *with Cholesky triangle* $L \in \mathbb{R}^{d\times d}$. *Let* $X_1, X_2, \ldots$ *be a sequence of i.i.d. random vectors with common distribution* P. *For* $\hat{Z}_{n,i}$ *as in* (5), *we have*

$$W_2^2(\widehat{\mathrm{P}}_n^{\hat{Z}}, \mathrm{Q}_0) \to 0 \qquad \text{almost surely as } n \to \infty,$$

where $\widehat{\mathrm{P}}_n^{\hat{Z}} := n^{-1}\sum_{i=1}^n \delta_{\hat{Z}_{n,i}}$ *and* $\mathrm{Q}_0 \in \mathcal{P}_2(\mathbb{R}^d)$ *is the distribution of* $L^{-1}(X_1 - \mu)$.

Proof. The random vectors $Z_i = L^{-1}(X_i - \mu)$ for $i = 1, 2, \ldots$ form an i.i.d. sequence with common distribution Q_0. By the strong law of large numbers,

$$\overline{X}_n \to \mu \quad \text{and} \quad S_{n,X} \to \Sigma \text{ a.s. as } n \to \infty. \tag{7}$$

With probability one, $S_{n,X}$ is invertible for n large enough (depending on the sample) and admits a unique Cholesky factor $L_{n,X}$. The map that sends a positive definite symmetric matrix to its Cholesky triangle is differentiable (Smith, 1995) and thus continuous. It follows that

$$L_{n,X} \to L \quad \text{and} \quad L_{n,X}^{-1} \to L^{-1} \text{ a.s. as } n \to \infty. \tag{8}$$

Let $\widehat{\mathrm{P}}_n^{Z} := n^{-1}\sum_{i=1}^n \delta_{Z_i}$. By the triangle inequality for the Wasserstein distance,

$$W_2(\widehat{\mathrm{P}}_n^{\hat{Z}}, \mathrm{Q}_0) \le W_2(\widehat{\mathrm{P}}_n^{\hat{Z}}, \widehat{\mathrm{P}}_n^{Z}) + W_2(\widehat{\mathrm{P}}_n^{Z}, \mathrm{Q}_0).$$

We already know that $W_2(\widehat{\mathrm{P}}_n^{Z}, \mathrm{Q}_0) \to 0$ almost surely as $n \to \infty$ (Bickel and Freedman, 1981, Lemma 8.4). It remains to show that $W_2(\widehat{\mathrm{P}}_n^{\hat{Z}}, \widehat{\mathrm{P}}_n^{Z}) \to 0$ in probability as $n \to \infty$.

Consider the coupling of $\widehat{\mathrm{P}}_n^{\hat{Z}}$ and $\widehat{\mathrm{P}}_n^{Z}$ via the discrete uniform distribution on the pairs $(\hat{Z}_{n,i}, Z_i)$ for $i = 1, \ldots, n$. From the definition of the Wasserstein distance, we have

$$W_2^2(\widehat{\mathrm{P}}_n^{\hat{Z}}, \widehat{\mathrm{P}}_n^{Z}) \le \frac{1}{n}\sum_{i=1}^n \|\hat{Z}_{n,i} - Z_i\|^2.$$

For each $i = 1, \ldots, n$, the identity $X_i = \mu + LZ_i$ yields

$$\begin{aligned}\|\hat{Z}_{n,i} - Z_i\| &= \|L_{n,X}^{-1}(\mu + LZ_i - \overline{X}_n) - Z_i\| \\ &\le \|L_{n,X}^{-1}\|\,\|\mu - \overline{X}_n\| + \|L_{n,X}^{-1}L - I_d\|\,\|Z_i\|\end{aligned}$$

featuring the matrix norm on $\mathbb{R}^{d\times d}$ associated with the Euclidean norm on $\mathbb{R}^d$. It follows that

$$W_2^2(\widehat{\mathrm{P}}_n^{\hat{Z}}, \widehat{\mathrm{P}}_n^{Z}) \le 2\|L_{n,X}^{-1}\|^2\,\|\mu - \overline{X}_n\|^2 + 2\|L_{n,X}^{-1}L - I_d\|^2\,\frac{1}{n}\sum_{i=1}^n \|Z_i\|^2.$$

The right-hand side converges to zero almost surely as $n \to \infty$ in view of the law of large numbers for $n^{-1}\sum_{i=1}^n \|Z_i\|^2$, (7), and (8). The result follows. □

Proposition 4 (Consistency)**.** *The sequence of tests $\phi^n_{\mathcal{E}(f_{rad})}$ is consistent against any $\mathrm{P} \in \mathcal{P}_2(\mathbb{R}^d) \setminus \mathcal{E}(f_{rad})$ with positive definite covariance matrix:*

$$\lim_{n\to\infty} \mathrm{P}^n\big[\phi^n_{\mathcal{E}(f_{rad})} = 1\big] = 1 \quad \text{for every } \alpha > 0.$$

Proof. Let $\alpha > 0$ and $\varepsilon > 0$. By Proposition 3, $\lim_{n\to\infty} \mathrm{P}^n_{f_{\mathrm{rad}}}[T_{\mathcal{E}(f_{\mathrm{rad}}),n} > \varepsilon] = 0$ and thus $c_{\mathcal{E}}(\alpha, n, f_{\mathrm{rad}}) \le \varepsilon$ for all sufficiently large n (depending on α and ε). It follows that $\lim_{n\to\infty} c_{\mathcal{E}}(\alpha, n, f_{\mathrm{rad}}) = 0$.

Let P be as in the statement. It is sufficient to show that there exists $\varepsilon > 0$ such that $\lim_{n\to\infty} \mathrm{P}^n[T_{\mathcal{E}(f_{\mathrm{rad}}),n} > \varepsilon] = 1$.

By Proposition 3, we have

$$W_2(\widehat{\mathrm{P}}^{\hat{Z}}_n, \mathrm{Q}_0) \to 0 \quad \text{in } \mathrm{P}^n\text{-probability}, \qquad n \to \infty,$$

with Q_0 as in the statement of that proposition. By the triangle inequality,

$$W_2(\widehat{\mathrm{P}}^{\hat{Z}}_n, \mathrm{P}_{f_{\mathrm{rad}}}) \ge W_2(\mathrm{P}_{f_{\mathrm{rad}}}, \mathrm{Q}_0) - W_2(\widehat{\mathrm{P}}^{\hat{Z}}_n, \mathrm{Q}_0).$$

By assumption, $\mathrm{Q}_0 \neq \mathrm{P}_{f_{\mathrm{rad}}}$ and thus $W_2(\mathrm{P}_{f_{\mathrm{rad}}}, \mathrm{Q}_0) > 0$ since otherwise $\mathrm{P} \in \mathcal{E}(\mathrm{P}_0)$. For $\varepsilon > 0$ less than $W_2^2(\mathrm{P}_{f_{\mathrm{rad}}}, \mathrm{Q}_0)$, we obtain $\lim_{n\to\infty} \mathrm{P}^n[T_{\mathcal{E}(f_{\mathrm{rad}}),n} > \varepsilon] = 1$, as required. □

4. Wasserstein GoF tests for general parametric families

Extending the scope of Section 3, consider the problem of testing whether the unknown common distribution P of a sample of observations belongs to some parametric family $\mathcal{M} := \{\mathrm{P}_\theta : \theta \in \Theta\}$ of distributions on $\mathbb{R}^d$ where the parameter space Θ is some metric space and the map $\theta \mapsto \mathrm{P}_\theta$ is assumed to be one-to-one and continuous in a sense to be specified. Given an independent random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$ from some unknown $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d)$, the goodness-of-fit problem is about testing

$$\mathcal{H}^n_0 : \mathrm{P} \in \mathcal{M} \quad \text{against} \quad \mathcal{H}^n_1 : \mathrm{P} \notin \mathcal{M}. \tag{9}$$

Assume that every $\mathrm{P}_\theta \in \mathcal{M}$ has a finite moment of order $p \in [1, \infty)$, that is, $\mathcal{M} \subseteq \mathcal{P}_p(\mathbb{R}^d)$. The test statistic we propose is

$$T_{\mathcal{M},n} := W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P}_{\hat{\theta}_n})$$

where $\hat{\theta}_n = \theta_n(\mathbf{X}_n)$ is some consistent (under $\mathcal{H}^n_0$) estimator sequence of the true parameter θ. The distribution of $\mathbf{X}_n$ under $\mathcal{H}^n_0$ in (9) being P^n_θ for some $\theta \in \Theta$, we would like to take

$$c_{\mathcal{M}}(\alpha, n, \theta) = \inf\{c > 0 : \mathrm{P}^n_\theta[T_{\mathcal{M},n} > c] \le \alpha\} \tag{10}$$

as the critical value of a test with nominal size $\alpha \in (0,1)$. This choice is infeasible, however, since the true parameter θ is unknown. Therefore, we propose to replace it by the bootstrapped quantity $c_{\mathcal{M}}(\alpha, n, \hat{\theta}_n)$, yielding the test

$$\phi^n_{\mathcal{M}} := \begin{cases} 1 & \text{if } T_{\mathcal{M},n} > c_{\mathcal{M}}(\alpha, n, \hat{\theta}_n), \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

rejecting $\mathcal{H}^n_0$ whenever $T_{\mathcal{M},n}$ exceeds $c_{\mathcal{M}}(\alpha, n, \hat{\theta}_n)$.

Given the parameter estimate $\hat{\theta}_n$, the proposed critical value can be approximated by resampling from the estimated distribution $\mathrm{P}_{\hat{\theta}_n}$. The idea is as follows and is given in more detail in Appendix B, in particular Algorithm 3:

1. generate a large number B of samples $\mathbf{X}^*_{n,b} = (X^*_{1,b}, \ldots, X^*_{n,b}) \in (\mathbb{R}^d)^n$, say, for $b = 1, \ldots, B$, of size n from $\mathrm{P}_{\hat{\theta}_n}$;
2. letting $\widehat{\mathrm{P}}^*_{n,b} = n^{-1} \sum_{i=1}^n \delta_{X^*_{i,b}}$ denote the empirical distribution of the bootstrap sample number b, compute
   (a) the parameter estimate $\hat{\theta}^*_{n,b} = \theta_n(\mathbf{X}^*_{n,b})$, and
   (b) the test statistic $T^*_{\mathcal{M},n,b} = W_2^2(\widehat{\mathrm{P}}^*_{n,b}, \mathrm{P}_{\hat{\theta}^*_{n,b}})$;
3. compute the empirical quantile

$$c_{\mathcal{M},B}(\alpha, n, \hat{\theta}_n) = \inf\left\{ c > 0 : B^{-1} \textstyle\sum_{b=1}^B I(T^*_{\mathcal{M},n,b} > c) \le \alpha \right\}.$$

As $B \to \infty$ and since, conditionally on the data, the Monte Carlo approximation $c_{\mathcal{M},B}(\alpha, n, \hat{\theta}_n)$ converges to the true quantile $c_{\mathcal{M}}(\alpha, n, \hat{\theta}_n)$ of the distribution of $T_{\mathcal{M},n}$ under $\mathrm{P}^n_{\hat{\theta}_n}$, provided the latter distribution has a positive and continuous density at the stated limit point. The rate of convergence in probability is $\mathrm{O}(1/\sqrt{B})$. In what follows, we assume we can compute $c_{\mathcal{M}}(\alpha, n, \hat{\theta}_n)$ to any desired degree of accuracy.

By construction, we have

$$\forall \theta \in \Theta, \qquad \mathrm{P}^n_\theta \left[T_{\mathcal{M},n} > c_{\mathcal{M}}(\alpha, n, \theta) \right] \le \alpha,$$

that is, if we could use the critical value at the true parameter θ, the risk of a false rejection of the null hypothesis would be bounded by α; it would be even equal to α if the distribution of $T_{\mathcal{M},n}$ is continuous at $c_{\mathcal{M}}(\alpha, n, \theta)$. But as the true parameter θ is unknown, we use the estimated one $\hat{\theta}_n$ instead, so that the risk of a false rejection is

$$\mathrm{P}^n_\theta \left[T_{\mathcal{M},n} > c_{\mathcal{M}}(\alpha, n, \hat{\theta}_n) \right].$$

The question remains open whether under the null hypothesis the actual size of the test indeed converges to α. To prove this would require non-degenerate limit distribution theory for $W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P}_\theta)$, not only for fixed $\theta \in \Theta$, but even for sequences θ_n converging to θ at certain rates. As discussed in Section 1.2, such limit results are still beyond the horizon. In the simulation study, however,

we check that the proposed bootstrap method indeed produces a test with approximately the right size.

Nevertheless, against a fixed alternative, the consistency of the test (11) based on the parametric bootstrap can be established theoretically. The key is the uniform convergence in probability of the empirical Wasserstein distance treated in Appendix A. For the parameter estimator $\hat{\theta}_n$, we need to assume weak consistency locally uniformly in θ: if ρ denotes the metric on Θ and if $\mathcal{K}(\Theta)$ denotes the collection of compact subsets of Θ, we will require that

$$\forall \varepsilon > 0,\ \forall K \in \mathcal{K}(\Theta), \qquad \lim_{n\to\infty} \sup_{\theta\in K} \mathrm{P}_\theta^n\big[\rho(\hat{\theta}_n, \theta) > \varepsilon\big] = 0. \tag{12}$$

As illustrated in Remark 1 below, this condition is satisfied, for instance, for moment estimators of a Euclidean parameter under a uniform integrability condition.

Proposition 5 (Consistency)**.** *Let $\mathcal{M} = \{\mathrm{P}_\theta : \theta \in \Theta\} \subseteq \mathcal{P}_p(\mathbb{R}^d)$, $p \in [1, \infty)$, be a model indexed by a metric space (Θ, ρ). Assume that the following conditions are satisfied:*

(a) the map $\Theta \to \mathcal{P}_p(\mathbb{R}^d) : \theta \mapsto \mathrm{P}_\theta$ is one-to-one and W_p-continuous;
(b) $\hat{\theta}_n$ is weakly consistent locally uniformly in $\theta \in \Theta$, i.e., (12) holds.

Then, the following properties hold:

(i) $T_{\mathcal{M},n} \to 0$ in P_θ^n-probability locally uniformly in $\theta \in \Theta$, i.e.,

$$\forall \varepsilon > 0,\ \forall K \in \mathcal{K}(\Theta), \qquad \lim_{n\to\infty} \sup_{\theta\in K} \mathrm{P}_\theta^n\big[T_{\mathcal{M},n} > \varepsilon\big] = 0;$$

(ii) the critical values $c_{\mathcal{M}}(\alpha, n, \theta)$ tend to zero uniformly in θ, i.e.,

$$\lim_{n\to\infty} \sup_{\theta\in K} c_{\mathcal{M}}(\alpha, n, \theta) = 0 \qquad \forall \alpha > 0,\ \forall K \in \mathcal{K}(\Theta);$$

(iii) for every $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d) \setminus \mathcal{M}$ such that there exists $K \in \mathcal{K}(\Theta)$ with

$$\mathrm{P}^n\big[\hat{\theta}_n \in K\big] \to 1 \qquad \text{as } n \to \infty,$$

we have $\lim_{n\to\infty} \mathrm{P}^n\big[\phi_{\mathcal{M}}^n = 1\big] = 1$.

Proof. (i) By the triangle inequality, it follows that

$$T_{\mathcal{M},n}^{1/p} = W_p(\widehat{\mathrm{P}}_n, \mathrm{P}_{\hat{\theta}_n}) \le W_p(\widehat{\mathrm{P}}_n, \mathrm{P}_\theta) + W_p(\mathrm{P}_\theta, \mathrm{P}_{\hat{\theta}_n}) \tag{13}$$

for all $\theta \in \Theta$. For compact $K \subseteq \Theta$, it is then sufficient to show that each of the W_p-distances on the right-hand side of (13) converges to 0 in P_θ^n-probability uniformly in $\theta \in K$.

First, since K is compact and $\theta \mapsto \mathrm{P}_\theta$ is W_p-continuous, the set

$$\mathcal{M}_K := \{\mathrm{P}_\theta : \theta \in K\}$$

is compact in $\mathcal{P}_p(\mathbb{R}^d)$ equipped with the W_p-distance. By Bickel and Freedman (1981, Lemma 8.3(b)) or Villani (2009, Definition 6.8(b) and Theorem 6.9) and a subsequence argument, it follows that $x \mapsto \|x\|^p$ is uniformly integrable with respect to $\mathcal{M}_K$, i.e.,

$$\lim_{r\to\infty} \sup_{\theta\in K} \int_{\|x\|>r} \|x\|^p \, \mathrm{dP}_\theta(x) = 0.$$

Corollary 1 then implies that $W_p(\widehat{\mathrm{P}}_n, \mathrm{P}_\theta) \to 0$ in P_θ^n-probability as $n \to \infty$, uniformly in $\theta \in K$.

Second, as K is compact and $\theta \to \mathrm{P}_\theta$ is W_p-continuous, there exists, for every scalar $\varepsilon > 0$, a scalar $\delta = \delta(\varepsilon) > 0$ such that[1]

$$\forall \theta \in K,\ \forall \theta' \in \Theta, \qquad \rho(\theta, \theta') \le \delta \implies W_p(\mathrm{P}_\theta, \mathrm{P}_{\theta'}) \le \varepsilon.$$

It follows that

$$\forall \theta \in K, \qquad \mathrm{P}_\theta^n \big[W_p(\mathrm{P}_\theta, \mathrm{P}_{\hat\theta_n}) > \varepsilon \big] \le \mathrm{P}_\theta^n \big[\rho(\theta, \hat\theta_n) > \delta \big].$$

By condition (b), the latter probability converges to 0 as $n \to \infty$ uniformly in $\theta \in K$.

(ii) Fix $\alpha > 0$, $\varepsilon > 0$, and $K \in \mathcal{K}(\Theta)$. By (i), there exists an integer $n(\varepsilon) \ge 1$ such that

$$\forall n \ge n(\varepsilon),\ \forall \theta \in K, \qquad \mathrm{P}_\theta^n \big[T_{\mathcal{M},n} > \varepsilon \big] \le \alpha.$$

By definition of the critical values, also $c_{\mathcal{M}}(\alpha, n, \theta) \le \varepsilon$ for all $n \ge n(\varepsilon)$ and $\theta \in K$.

(iii) Let P and K be as in the statement. Put $c_n = \sup_{\theta\in K} c_{\mathcal{M},n}(\alpha, n, \theta)$. We have

$$\begin{aligned} \mathrm{P}^n \big[\phi_{\mathcal{M}}^n = 1 \big] &\ge \mathrm{P} \big[T_{\mathcal{M},n} > c_{\mathcal{M}}(\alpha, n, \hat\theta_n),\ \hat\theta_n \in K \big] \\ &\ge \mathrm{P} \big[T_{\mathcal{M},n} > c_n,\ \hat\theta_n \in K \big]. \end{aligned}$$

In view of *(ii)*, we have $c_n \to 0$ as $n \to \infty$, so that it is sufficient to show that there exists $\varepsilon > 0$, depending on P and $\mathcal{M}$, such that $\lim_{n\to\infty} \mathrm{P}^n \big[T_{\mathcal{M},n} > \varepsilon \big] = 1$. Consider two cases, $\mathrm{P} \in \mathcal{P}_p(\mathbb{R}^d) \setminus \mathcal{M}$ and $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d) \setminus \mathcal{P}_p(\mathbb{R}^d)$, according as P has a finite moment of order p or not.

First, suppose that $\mathrm{P} \in \mathcal{P}_p(\mathbb{R}^d) \setminus \mathcal{M}$. We have $W_p(\mathrm{P}, \mathrm{P}_\theta) > 0$ for every $\theta \in \Theta$ while the map $\theta \mapsto W_p(\mathrm{P}, \mathrm{P}_\theta)$ is continuous. As K is compact, $\eta :=$

[1] This is a slight generalization of the well-known property that a continuous function on a compact set is uniformly compact. As a proof, fix $\varepsilon > 0$ and consider for each $\theta \in K$ a scalar $\delta(\theta) > 0$ such that for all $\theta' \in \Theta$ with $\rho(\theta, \theta') \le \delta(\theta)$ we have $W_p(\mathrm{P}_\theta, \mathrm{P}_{\theta'}) \le \varepsilon/2$. Cover K by open balls with centers $\theta \in K$ and radii $\delta(\theta)/2$. By compactness, extract a finite cover with centers $\theta_1, \ldots, \theta_m \in K$. Put $\delta = \min_j \delta(\theta_j)/2$. For every $\theta \in K$ and $\theta' \in \Theta$ with $\rho(\theta, \theta') \le \delta$, there exists $j = 1, \ldots, m$ such that $\rho(\theta, \theta_j) < \delta(\theta_j)/2$ and then also $\rho(\theta', \theta_j) < \delta(\theta_j)$. By the triangle inequality, $W_p(\mathrm{P}_\theta, \mathrm{P}_{\theta'}) \le W_p(\mathrm{P}_{\theta_j}, \mathrm{P}_\theta) + W_p(\mathrm{P}_{\theta_j}, \mathrm{P}_{\theta'}) \le \varepsilon$.

$\inf\{W_p(\mathrm{P}, \mathrm{P}_\theta) : \theta \in K\} > 0$. On the event $\{\hat{\theta}_n \in K\}$, the triangle inequality implies

$$\begin{aligned} T_{\mathcal{M},n}^{1/p} = W_p(\widehat{\mathrm{P}}_n, \mathrm{P}_{\hat{\theta}_n}) &\geq W_p(\mathrm{P}, \mathrm{P}_{\hat{\theta}_n}) - W_p(\widehat{\mathrm{P}}_n, \mathrm{P}) \\ &\geq \eta - W_p(\widehat{\mathrm{P}}_n, \mathrm{P}). \end{aligned}$$

We obtain that

$$\begin{aligned} \mathrm{P}^n\big[\phi_{\mathcal{M}}^n = 1\big] &\geq \mathrm{P}\big[T_{\mathcal{M},n}^{1/p} > c_n^{1/p},\, \hat{\theta}_n \in K\big] \\ &\geq \mathrm{P}\big[W_p(\widehat{\mathrm{P}}_n, \mathrm{P}) < \eta - c_n^{1/p},\, \hat{\theta}_n \in K\big]. \end{aligned}$$

As $\eta > 0$ and $\lim_{n\to\infty} c_n = 0$, the latter probability converges to one by the assumption made on K and the fact that $W_p(\widehat{\mathrm{P}}_n, \mathrm{P}) \to 0$ in P^n-probability as $n \to \infty$.

Second, suppose that $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d) \setminus \mathcal{P}_p(\mathbb{R}^d)$. Let δ_0 be the Dirac measure at $0 \in \mathbb{R}^d$. Since $\theta \mapsto W_p(\mathrm{P}_\theta, \delta_0)$ is continuous, $s = \sup_{\theta\in K} W_p(\mathrm{P}_\theta, \delta_0)$ is finite. By the triangle inequality, on the event $\{\hat{\theta}_n \in K\}$,

$$\begin{aligned} T_{\mathcal{M},n}^{1/p} = W_p(\widehat{\mathrm{P}}_n, \mathrm{P}_{\hat{\theta}_n}) &\geq W_p(\widehat{\mathrm{P}}_n, \delta_0) - W_p(\mathrm{P}_{\hat{\theta}_n}, \delta_0) \\ &\geq W_p(\widehat{\mathrm{P}}_n, \delta_0) - s. \end{aligned}$$

Moreover, $W_p^p(\widehat{\mathrm{P}}_n, \delta_0) = n^{-1}\sum_{i=1}^n \|X_i\|^p$ diverges to $\int \|x\|^p \,\mathrm{dP}(x) = \infty$ in P^n-probability by the weak law of large numbers. It follows that

$$\lim_{n\to\infty} \mathrm{P}^n\big[T_{\mathcal{M},n} > c_n,\, \hat{\theta}_n \in K\big] = 1. \qquad \square$$

Remark 1 (Uniform consistency). Under a mild moment condition, the uniform consistency condition (b) in Proposition 5 is satisfied for *method of moment estimators*—call them *moment estimators*—of a Euclidean parameter $\theta \in \Theta \subseteq \mathbb{R}^k$. In the method of moments, an estimator $\hat{\theta}_n$ of θ is obtained by solving (with respect to θ) the equations

$$\frac{1}{n}\sum_{i=1}^n f_j(X_i) = \mathrm{E}_\theta[f_j(X)], \qquad j = 1, \ldots, k,$$

for some given k-tuple $f := (f_1, \ldots, f_k)$ of functions such that $m : \theta \mapsto \mathrm{E}_\theta[f(X)]$ is a homeomorphism between Θ and $m(\Theta)$; see, for instance, van der Vaart (1998, Chapter 4). The consistency of $\hat{\theta}_n = m^{-1}(n^{-1}\sum_{i=1}^n f(X_i))$ uniformly in $\theta \in K$ for any compact $K \subseteq \Theta$ then follows from the uniform consistency over K of $n^{-1}\sum_{i=1}^n f(X_i)$ as an estimator of $\mathrm{E}_\theta[f(X)]$ for such θ. By van der Vaart and Wellner (1996, Proposition A.5.1), a sufficient condition for the latter is that the functions f_j are P_θ-uniformly integrable for $\theta \in K$, i.e.,

$$\lim_{M\to\infty} \sup_{\theta\in K} \mathrm{E}_\theta\big[|f_j(X)|\, I\{|f_j(X)| > M\}\big] = 0, \qquad j = 1, \ldots, k.$$

Since $I\{|f_j(X)| > M\} \leq |f_j(X)|^\eta / M^\eta$ for $\eta > 0$, a further sufficient condition is that there exists $\eta > 0$ such that $\sup_{\theta\in K} \mathrm{E}_\theta[|f_j(X)|^{1+\eta}] < \infty$ for $j = 1, \ldots, k$.

Remark 2 (Parameter estimate under the alternative). In Proposition 5(iii), the condition that there exists a compact $K \subseteq \Theta$ such that $\lim_{n\to\infty} \mathrm{P}^n[\hat{\theta}_n \in K] = 1$ holds, for instance, when Θ is locally compact and $\hat{\theta}_n$ is consistent for a pseudo-parameter $\theta(\mathrm{P}) \in \Theta$. This is the case for the moment estimators of Remark 1 when $\Theta \subseteq \mathbb{R}^k$ is open and f is P-integrable with $\int f(x)\,\mathrm{dP}(x) \in m(\Theta)$.

Remark 3 (Location–scale parameters). Let $p = 2$ and consider a parametric model

$$\mathcal{M} = \{\mathrm{Q}_\psi : \psi \in \Psi\} \subseteq \mathcal{P}_2(\mathbb{R}^d)$$
$$\text{where } \psi = (\mu, \sigma, \theta) \in \mathbb{R}^d \times (0,\infty)^d \times \Theta,$$

such that, for $X_i = (X_{11}, \ldots, X_{id}) \sim \mathrm{Q}_\psi \in \mathcal{M}$, we have

$$\mu_j = \mathrm{E}[X_{ij}] \quad \text{and } \sigma_j = \sqrt{\mathrm{var}(X_{ij})} \quad \text{for all } j = 1, \ldots, d.$$

The range Θ of θ is supposed not to depend on the location–scale parameter vectors μ and σ. For instance, θ could be a vector of shape parameters for the marginal distributions and/or determine the copula of Q_ψ.

Then, we can simplify the procedure by employing estimated residuals of the form $\hat{Z}_{n,i} = (\hat{Z}_{n,i1}, \ldots, \hat{Z}_{n,id})$ with

$$\hat{Z}_{n,ij} = (X_{ij} - \overline{X}_{n,j})/s_{n,j,X}, \qquad i = 1, \ldots, n,\ j = 1, \ldots, d, \tag{14}$$

where $\overline{X}_{n,j}$ and $s_{n,j,X}$ are the empirical means and standard deviations, respectively, of $X_{1j}, \ldots, X_{nj}$. The joint distribution of these estimated residuals depends only on θ but not on (μ, σ). Indeed, we have $X_{ij} = \mu_j + \sigma_j Z_{ij}$ where the distribution of $Z_i = (Z_{i1}, \ldots, Z_{id})$ is P_θ, which is defined as Q_ψ with $\psi = ((0,\ldots,0),(1,\ldots,1),\theta)$, that is, $\mu_j = 0$ and $\sigma_j = 1$ for all $j = 1, \ldots, d$. In obvious notation, we have $\hat{Z}_{n,ij} = (Z_{ij} - \overline{Z}_{n,j})/s_{n,j,Z}$.

Let $\hat{\theta}_n$ denote a strongly consistent estimator of θ that depends on the data only through $\hat{Z}_{n,1}, \ldots, \hat{Z}_{n,n}$. Consider the empirical distributions

$$\widehat{\mathrm{P}}_n^{\hat{Z}} := n^{-1} \sum_{i=1}^n \delta_{\hat{Z}_{n,i}} \quad \text{and} \quad \widehat{\mathrm{P}}_n^{Z} := n^{-1} \sum_{i=1}^n \delta_{Z_i}.$$

To test the hypothesis $\mathcal{H}_0^n : \mathrm{P} \in \mathcal{M}$, we propose the location-scale adjusted statistic

$$T_{\mathcal{M},n}^{\mathrm{ls}} := W_2^2\big(\widehat{\mathrm{P}}_n^{\hat{Z}}, \mathrm{P}_{\hat{\theta}_n}\big). \tag{15}$$

Its distribution still depends on θ but no longer on μ or σ. Critical values can thus be computed as if $\mu_j = 0$ and $\sigma_j = 1$ for all $j = 1, \ldots, d$. For a test of size $\alpha \in (0,1)$, we reject the null hypothesis as soon as the test statistic exceeds the critical value $c_{\mathcal{M}}^{\mathrm{ls}}(\alpha, n, \hat{\theta}_n)$ where

$$c_{\mathcal{M}}^{\mathrm{ls}}(\alpha, n, \theta) := \inf\big\{c \geq 0 : \mathrm{P}_\theta^n\big[T_{\mathcal{M},n}^{\mathrm{ls}} > c\big] \leq \alpha\big\}, \qquad \theta \in \Theta. \tag{16}$$

In practice, critical values are calculated by a parametric bootstrap procedure as before. The advantage of the estimated residuals (14) is that the critical values are a function of θ only rather than a function of $\psi = (\mu, \sigma, \theta)$, which greatly simplifies their computation.

Under the null hypothesis, we have $T^{\mathrm{ls}}_{\mathcal{M},n} \to 0$ almost surely as $n \to \infty$ since it is bounded by a multiple of

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}(\hat{Z}_{n,ij} - Z_{ij})^2 + W_2^2\big(\widehat{\mathrm{P}}_n^Z, \mathrm{P}_\theta\big) + W_2^2\big(\mathrm{P}_\theta, \mathrm{P}_{\hat{\theta}_n}\big)$$

where each of the three terms converges to zero almost surely. Under an alternative $\mathrm{P} \in \mathcal{P}_2(\mathbb{R}^d) \setminus \mathcal{M}$ such that $\hat{\theta}_n$ remains in a compact set with probability tending to one, $T^{\mathrm{ls}}_{\mathcal{M},n}$ remains bounded away from zero and the test is consistent by an argument similar to the proof of Proposition 5.

5. Finite-sample performance of GoF tests

This section is devoted to a numerical assessment of the finite-sample performance of the Wasserstein-based GoF tests introduced in the previous sections and we compare them, whenever possible, with other tests. The case of a simple null hypothesis (Section 2) is treated in Section 5.1. The performances of various tests for multivariate normality, which is a special case of the hypothesis of elliptical symmetry considered in Section 3, are compared in Section 5.2, along with an illustration involving a Student t distribution with known degrees of freedom. Section 5.3 considers, in line with Remark 3, the more general composite null hypothesis of a parametric family indexed by marginal location and scale along with a copula parameter θ. Numerical results support the validity of the bootstrap-based calculation of critical values. To the best of our knowledge, no GoF test is available in the literature for such cases except for the method described by Khmaladze (2016), the numerical implementation of which, however, remains unsettled.

Throughout, we consider the Wasserstein distance of order $p = 2$. The level α of the tests is set to 5%, the sample size is $n = 200$, and the number of replicates considered in the estimation of power curves is 1000. We rely on the R package transport (Schuhmacher et al., 2019), which is why we restrict ourselves to dimension $d = 2$. As explained in Section 1.3, stochastic algorithms have recently been proposed to solve the semi-discrete problem in higher dimensions, but these are not yet implemented in R.

5.1. Simple null hypotheses

The setting is as in Section 2: given an independent random sample $X_1, \ldots, X_n$ from some unknown $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d)$, we consider testing the simple null hypothesis $\mathcal{H}_0^n : \mathrm{P} = \mathrm{P}_0$, where $\mathrm{P}_0 \in \mathcal{P}_2(\mathbb{R}^d)$ is fully specified.

5.1.1. Other GoF tests

Two other goodness-of-fit tests will be used as benchmarks.

Rippl, Munk and Sturm (2016) consider the fully specified Gaussian null hypothesis $\mathcal{H}_0^n : \mathrm{P} = \mathcal{N}_d(\mu_0, \Sigma_0)$ with given mean and covariance. Recall that the squared 2-Wasserstein distance between two d-variate Gaussian distributions is

$$W_2^2\big(\mathcal{N}_d(\mu_1,\Sigma_1),\mathcal{N}_d(\mu_2,\Sigma_2)\big) = \|\mu_1-\mu_2\|^2 + \operatorname{tr}\big\{\Sigma_1+\Sigma_2-2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\big\}.$$

The Rippl–Munk–Sturm test statistic is $W_2^2(\mathcal{N}_d(\overline{X}_n, S_{n,X}), \mathcal{N}_d(\mu_0,\Sigma_0))$, with $\overline{X}_n$ and $S_{n,X}$ the sample mean and sample covariance matrix, respectively. This test is sensitive to changes in the parameters of the Gaussian distribution but not to other types of alternatives. Calculation of the test statistic is straightforward. To compute critical values, we relied on a Monte Carlo approximation, drawing many samples from the Gaussian null distribution and taking the empirical quantile of the resulting test statistics.

Khmaladze (2016) constructs empirical processes in such a way that they are asymptotically distribution-free, which facilitates their use for hypothesis testing. A special case of the construction is as follows. Let the d-variate cumulative distribution function (cdf) F be absolutely continuous with joint density f, marginal densities $f_1,\dots,f_d$, and copula density c. Define

$$l(x) = \{c(F_1(x_1),\dots,F_d(x_d))\}^{1/2}, \qquad x \in \mathbb{R}^d,$$

with $F_1,\dots,F_d$ the marginal cdfs of F. The d-variate cdf $G(x) = \prod_{j=1}^d F_j(x_j)$ has the same margins as F, but coupled via the independence copula. Letting

$$\kappa(x) = \int_{(-\infty,x]} l(y)\, f(y)\,\mathrm{d}y \quad \text{and} \quad \kappa = \int l(y)\, f(y)\,\mathrm{d}y,$$

it follows from Corollary 4 in Khmaladze (2016) that the empirical process

$$\tilde{v}_{F,n}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \big\{l(X_i)I(X_i\le x)-\kappa(x)\big\} - \frac{G(x)-\kappa(x)}{1-\kappa}\frac{1}{\sqrt{n}}\sum_{i=1}^n\big\{l(X_i)-\kappa\big\}$$

based on an independent random sample $X_1,\dots,X_n$ from F converges weakly to a G-Brownian bridge, i.e., has the same weak limit as the ordinary empirical process

$$v_{G,n}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\big\{I(Y_i\le x)-G(x)\big\}$$

based on an independent random sample $Y_1,\dots,Y_n$ from G. The asymptotic distribution of a test statistic based on $\tilde{v}_{F,n}$ which is invariant with respect to coordinate-wise continuous monotone increasing transformations is thus the same as if F (or G) were the uniform distribution on $[0,1]^d$. This includes the Kolmogorov–Smirnov type statistic $\sup_{x\in\mathbb{R}^d}\big|\tilde{v}_{F,n}(x)\big|$, which (with F the cdf of P_0) we are considering below for comparison with our Wasserstein-based test.

In case F has independent margins, F and G coincide and the procedure reduces to a classical Kolmogorov–Smirnov test. To ensure that the test has the right size at finite sample size, we calculate critical values by Monte Carlo approximation rather than relying on the asymptotic theory.

5.1.2. Results

In Figure 1, we assess the performance of the GoF tests of $\mathcal{H}_0^n : \mathrm{P} = \mathrm{P}_0$ where $\mathrm{P}_0 = \mathcal{N}_2(0, I_2)$ is a centered bivariate Gaussian with identity covariance matrix. The alternatives P in panels (a)–(f) are as follows:

(a) $\mathrm{P} = \mathcal{N}_2\big(\binom{\mu}{\mu}, I_2\big)$ with location shift μ along the main diagonal (rejection frequencies plotted against $\mu \in \mathbb{R}$);
(b) $\mathrm{P} = \mathcal{N}_2\big(0, \left(\begin{smallmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{smallmatrix}\right)\big)$ (rejection frequencies plotted against $\sigma^2 > 0$);
(c) $\mathrm{P} = \mathcal{N}_2\big(0, \left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right)\big)$ with correlation ρ (rejection frequencies plotted against $\rho \in (-1, 1)$);
(d) P has standard normal margins but Gumbel copula with parameter θ (rejection frequencies plotted against $\theta \in [1, \infty)$);
(e) P has standard Gaussian margins but a bivariate Student t copula with $\nu = 4$ degrees of freedom and correlation parameter ρ (rejection frequencies plotted against $\rho \in (-1, 1)$);[2]
(f) P is the "banana-shaped" Gaussian mixture described in Appendix C (rejection frequencies plotted against the mixing weight $p \in (-1, 1)$).[3]

The Gumbel and Student t copula simulations in (d) and (e) were implemented from the R package copula (Hofert et al., 2018).

Inspection of Figure 1 indicates that the Khmaladze test, as a rule, is uniformly outperformed by the Rippl–Munk–Sturm and Wasserstein tests. The Rippl–Munk–Sturm test, of course, does relatively well under the Gaussian alternatives of panels (a)–(c) where, however, the Wasserstein test is almost as powerful (while its validity, contrary to that of the Rippl–Munk–Sturm test, extends largely beyond the Gaussian null hypothesis). Against the non-Gaussian alternatives in panels (d)–(f), the Wasserstein test has higher power than the Rippl–Munk–Sturm and Khmaladze tests, with the exception of the Gumbel copula alternative in panel (d), where the Rippl–Munk–Sturm and Wasserstein tests perform equally well. For the "banana mixture" of panel (f), the Rippl–Munk–Sturm test fails to capture the change in distribution.

Figures 2 and 3 are dealing with non-Gaussian simple null distributions P_0, so that the Rippl–Munk–Sturm test no longer applies. In Figure 2, the null distribution is the mixture of Gaussians $\mathrm{P}_0 = 0.5\,\mathcal{N}_2(0, I_2) + 0.5\,\mathcal{N}_2\left(\binom{3}{0}, I_2\right)$. The alternatives in both panels are

(a) $\mathrm{P} = 0.5\,\mathcal{N}_2(0, I_2) + 0.5\,\mathcal{N}_2\left(\binom{3+\delta}{0}, I\right)$ (rejection frequencies plotted against the location shift $\delta \in \mathbb{R}$);

[2] Note that P is not Gaussian, even for $\rho = 0$.
[3] The mixture is constructed so that the first and second moments of P remain close to those of P_0.

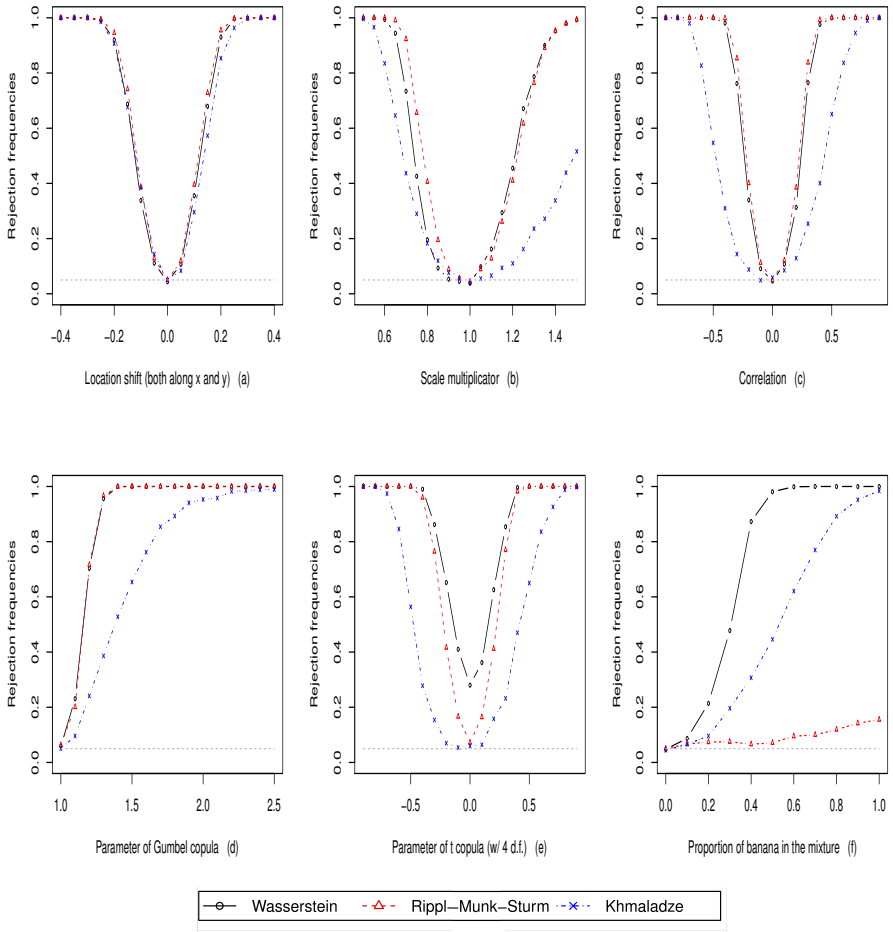

FIG 1. *Empirical powers of various GoF tests for the simple Gaussian null hypothesis* $\mathcal{H}_0^n : \mathrm{P} = \mathcal{N}_2(0, I_2)$. *Three tests are considered: the Wasserstein-2 distance (Section 2), the Rippl–Munk–Sturm test (Rippl, Munk and Sturm, 2016), and the Khmaladze Kolmogorov–Smirnov type test (Khmaladze, 2016), see Section 5.1.1. The alternatives* P *in panels (a)–(f) are described in Section 5.1.2 (note that in (e),* P *is not Gaussian even when* $\rho = 0$*).*

(b) $\mathrm{P}_0 = \lambda \mathcal{N}_2(0, I_2) + (1-\lambda)\mathcal{N}_2\left(\binom{3}{0}, I_2\right)$ (rejection frequencies plotted against the mixing weight $\lambda \in [0, 1]$).

The Wasserstein test uniformly outperforms the Khmaladze one.

In Figure 3, P_0 has standard Gaussian margins and a Gumbel copula with parameter $\theta = 1.7$. The alternative P is of the same form but with another

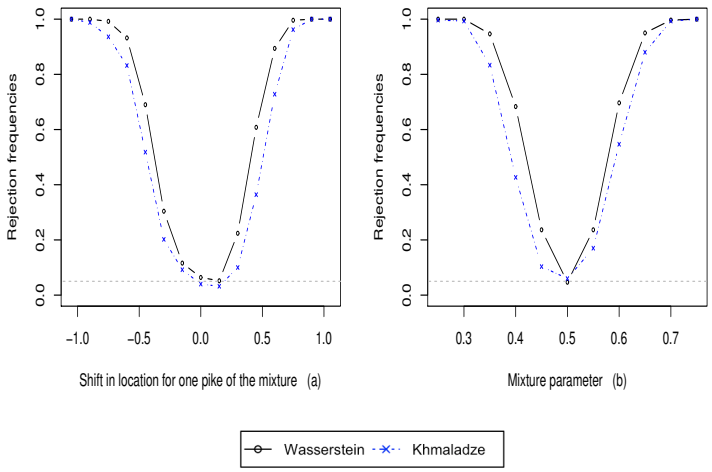

FIG 2. *Empirical powers of the Wasserstein and Khmaladze (2016) tests for the simple null hypothesis* $\mathcal{H}_0^n : \mathrm{P} = \mathrm{P}_0$ *with* P_0 *an equal-weights mixture of* $\mathcal{N}_2(0, I_2)$ *and* $\mathcal{N}_2\left(\binom{3}{0}, I_2\right)$. *In panel (a), the alternative* P *is an equal-weights mixture of* $\mathcal{N}_2(0, I_2)$ *and* $\mathcal{N}_2\left(\binom{3+\delta}{0}, I_2\right)$; *rejection frequencies are plotted against* $\delta \in [-1, 1]$. *In panel (b), the alternative* P *is a mixture of the same two components, but with weights* $\lambda \in (0, 1)$ *and* $(1 - \lambda)$; *rejection frequencies are plotted against* $\lambda \in [0.25, 0.75]$.

value $\theta \neq 1.7$ of the copula parameter $\theta \in [1, \infty)$. Again, the Wasserstein test yields uniformly higher empirical power.

5.2. *Elliptical families*

If the radial density f_{rad} is the density of the root of a chi-square random variable with d degrees of freedom, the elliptical family $\mathcal{E}(f_{\mathrm{rad}})$ corresponds to the Gaussian family. The null hypothesis in (3) then is that P is multivariate Gaussian with unknown mean vector and positive definite covariance matrix.

Testing multivariate normality is a well-studied problem for which many tests have been put forward. As benchmarks, we will consider here the tests proposed in Royston (1982), Henze and Zirkler (1990), and Rizzo and Székely (2016). Royston's test is a generalisation of the well-known Shapiro–Wilks test. The Henze–Zirkler test statistic is an integrated weighted squared distance between the characteristic function under the null and its empirical counterpart. Interestingly, Ramdas, García Trillos and Cuturi (2017) showed that the Wasserstein distance and the energy distance of Rizzo and Székely (2016) are connected, as the so-called entropy-penalized Wasserstein distance interpolates between them two. We borrowed the implementation of these tests from the R package MVN (Korkmaz, Goksuluk and Zararsiz, 2014). The test by Rippl, Munk and Sturm

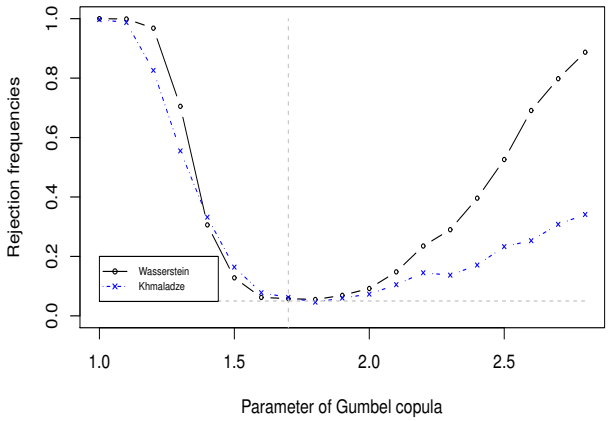

FIG 3. *Empirical powers of the Wasserstein and Khmaladze (2016) tests (see Section 5.1.1) for the simple null hypothesis* $\mathcal{H}_0^n : \mathrm{P} = \mathrm{P}_0$ *with* P_0 *a bivariate distribution with standard Gaussian margins and Gumbel copula with parameter* $\theta = 1.7$*; rejection frequencies are plotted against the copula parameter* θ*.*

(2016) considered in Section 5.1 does not apply here, since it only can handle fully specified Gaussian distributions.

The alternatives in the two panels of Figure 4 are

(a) P with standard normal margins and a Gumbel copula with parameter θ ranging over $[1, \infty)$;
(b) P with independent margins, one of which is standard normal while the other one is Student t with $\nu > 0$ degrees of freedom.

Inspection of Figure 4 reveals that the Wasserstein test has the highest power against the copula alternative in panel (a), while Royston's test has no power at all. For the Student t alternative in panel (b), Royston's test comes out as most sensitive, but the Wasserstein and energy tests (Rizzo and Székely, 2016) performe quite well too.

In Figure 5, we consider the bivariate Student ($\nu = 12$ degrees of freedom) elliptical family, with radial density f_{rad} the density of the root of a rescaled Fisher $F(d, 12)$ variable. Figure 5 provides a plot of rejection frequencies under bivariate skew-t alternatives (Azzalini, 2014) with marginal skewness parameters α_1 and α_2. Simulations were based on the function rmst from the R package sn (Azzalini, 2020). In principle, the empirical process approach in Khmaladze (2016) leads to test statistics that are asymptotically distribution-free, but their numerical implementation involves a number of multiple integrals, the computation of which remains problematic.

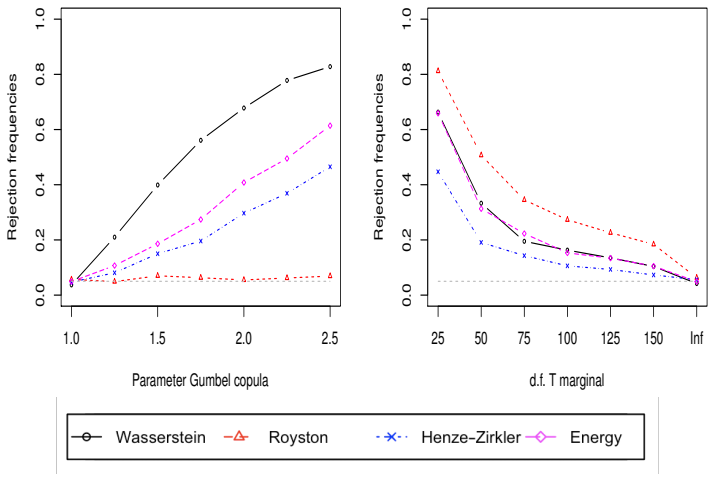

FIG 4. *Empirical power curves of various tests of the hypothesis that* P *is bivariate Gaussian with unknown mean vector and covariance matrix. The Wasserstein test in Section 3 is compared to three other multivariate normality tests mentioned in Section 5.2. In (a), the alternative* P *has a Gumbel copula with parameter θ; rejection frequencies are plotted against $\theta \in [1, \infty)$. In (b), one of the marginals of* P *is a Student* t *distribution with ν degrees of freedom; rejection frequencies are plotted against $\nu > 0$.*

5.3. *General parametric families*

We now turn to the more general example of a non-elliptical parametric model $\mathcal{M}$ where the parametric bootstrap procedure described in Section 4 nevertheless applies. In the notation of Remark 3, let $\mathcal{M} = \{Q_\psi : \psi \in \Psi\}$ consist of the bivariate distributions with Gaussian marginals and an Ali–Mikhail–Haq (AMH) copula, yielding a five-dimensional parameter vector $\psi = (\mu_1, \sigma_1, \mu_2, \sigma_2, \theta)$ where $\mu_1, \mu_2 \in \mathbb{R}$ and $\sigma_1, \sigma_2 \in (0, \infty)$ are marginal location and scale parameters, and $\theta \in \Theta = [-1, 1]$ is the AMH copula parameter. We applied the method involving the location-scale reduction described in Remark 3. Following Genest, Ghoudi and Rivest (1995), the copula parameter θ was estimated via a rank-based maximum pseudo-likelihood estimator. Obviously, the componentwise ranks of the data and those of the residuals in (14) coincide, so that $\hat{\theta}_n$, as required, depends on the data only through the residuals.

We first checked the validity of the parametric bootstrap procedure of Section 4. To do so, we simulated 1 000 independent random samples of size $n = 200$ from $P \in \mathcal{M}$ with $\theta = 0.7$. For each sample, we calculated the test statistic $T^{\mathrm{ls}}_{\mathcal{M},n}$ in (15) and checked whether or not it exceeds the bootstrapped critical value $c^{\mathrm{ls}}_{\mathcal{M}}(\alpha, n, \hat{\theta}_n)$ for α equal to multiples of 5%. The critical value function $\theta \mapsto c^{\mathrm{ls}}_{\mathcal{M}}(\alpha, n, \theta)$ in (16) was pre-computed by Algorithm 3, or rather a

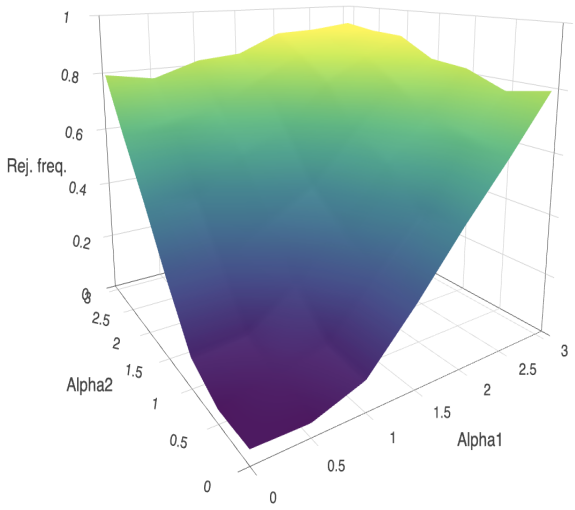

FIG 5. *Empirical power of the Wasserstein test in Section 3 for the hypothesis that* P *is bivariate Student* t *with* $\nu = 12$ *degrees and unknown mean vector and covariance matrix. The alternatives* P *are bivariate skew-*t *with skewness parameters* α_1 *and* α_2.

variation thereof taking into account the estimated residuals in Eq. (14). The points in Figure 6(a) show the empirical type I errors as a function of α. The diagonal line fits the points well, lending support to the validity of the parametric bootstrap method (if not proving it).

Figure 6(b) similarly displays the rejection frequencies of the Wasserstein test under an alternative P whose copula belongs to the Frank family with parameter η. If $\eta = 0$, the Frank copula reduces to the independence copula, which is a member of the AMH family too. Again, the approach in Khmaladze (2016) in principle also applies, but its actual implementation is intricate and remains unsettled.

References

AURENHAMMER, F., HOFFMANN, F. and ARONOV, B. (1998). Minkowski-type theorems and least-squares clustering. *Algorithmica* **20** 61–76.

AZZALINI, A. (2014). *The Skew-Normal and Related Families. Institute of Mathematical Statistics (IMS) Monographs* **3**. Cambridge University Press, Cambridge With the collaboration of Antonella Capitanio. MR3468021

AZZALINI, A. (2020). The R package `sn`: The Skew-Normal and Related Distributions such as the Skew-*t*., Università di Padova, Italia.

BAKSHAEV, A. and RUDZKIS, R. (2015). Multivariate goodness-of-fit tests based on kernel density estimators. *Nonlinear Analysis. Modelling and Control* **20** 585–602.

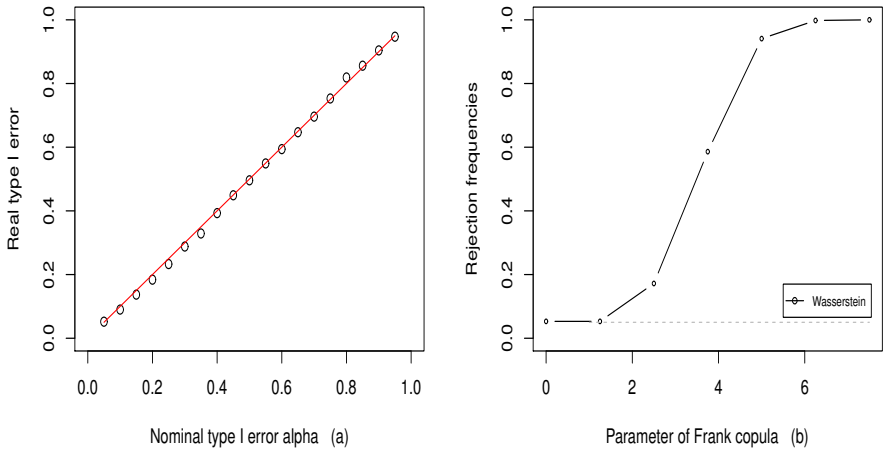

FIG 6. *Wasserstein test for $\mathcal{H}_0^n : \mathrm{P} \in \mathcal{M}$ with $\mathcal{M}$ the family of bivariate distributions with Gaussian margins with unknown location–scale parameters and Ali–Mikhail–Haq (AMH) copula with unknown parameter $\theta \in [-1, 1]$ (Section 5.3). Test statistic and critical values computed based on estimated residuals and parametric bootstrap as in Remark 3. Panel (a) shows real versus nominal type I errors α based on 1 000 samples of size $n = 200$ drawn from $\mathrm{P} \in \mathcal{M}$ with $\theta = 0.8$. Panel (b) shows the power against alternatives* P *with Gaussian marginals and Frank copula with parameter $\eta \geq 0$; if $\eta = 0$, the Frank copula is the independence one, which is part of the AMH family too.*

BERAN, R. (1997). Diagnosing bootstrap success. *Annals of the Institute of Statistical Mathematics* **49** 1–24.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics* **9** 1196–1217.

BICKEL, P. J. and ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics* **1** 1071–1095.

BOBKOV, S. and LEDOUX, M. (2019). One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Mem. Amer. Math. Soc.* **261** v+126. MR4028181

CAMBANIS, S., HUANG, S. and SIMONS, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis* **11** 368–385.

CAPANU, M. (2019). A unified approach to proving parametric bootstrap consistency for some goodness-of-fit tests. *Statistics* **53** 58–80.

CARLIER, G., CHERNOZHUKOV, V., GALICHON, A. et al. (2016). Vector quantile regression: an optimal transport approach. *The Annals of Statistics* **44** 1165–1192.

CHERNOZHUKOV, V., GALICHON, A., HALLIN, M., HENRY, M. et al. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*

45 223–256.

Cramér, A. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal* **1** 13-74.

Deb, N. and Sen, B. (2019). Multivariate Rank-based Distribution-free Nonparametric Testing using Measure Transportation. *arXiv preprint arXiv:1909.08733*.

del Barrio, E., Giné, E. and Utzet, F. (2005). Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli* **11** 131–189.

del Barrio, E. and Loubes, J. M. (2019). Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability* **47** 926–951.

del Barrio, E., Cuesta-Albertos, J. A., Matrán, C. and Rodríguez-Rodríguez, J. M. (1999). Tests of goodness of fit based on the L_2-Wasserstein distance. *The Annals of Statistics* **27** 1230–1239.

del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., Csörgö, S., Cuadras, C. M., de Wet, T., Giné, E., Lockhart, R., Munk, A. and Stute, W. (2000). Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *Test* **9** 1–96.

del Barrio, E., Cuesta-Albertos, J. ., Hallin, M. and Matrán, C. (2018). Center-Outward Distribution Functions, Quantiles, Ranks, and Signs in $\mathbb{R}^d$. *arXiv preprint arXiv:1806.01238*.

Ebner, B., Henze, N. and Yukich, J. E. (2018). Multivariate goodness-of-fit on flat and curved spaces via nearest neighbor distances. *Journal of Multivariate Analysis* **165** 231–242.

Fan, Y. (1997). Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function. *Journal of Multivariate Analysis* **62** 36–63.

Fang, K.-T., Kotz, S. and Ng, K.-W. (1990). *Symmetric multivariate and related distributions*. Chapman & Hall, London.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields* **162** 707–738.

Genest, C., Ghoudi, K. and Rivest, L. P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82** 543–552.

Genevay, A., Cuturi, M., Peyré, G. and Bach, F. (2016). Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems* 3440–3448.

Ghosal, P. and Sen, B. (2019). Multivariate ranks and quantiles using optimal transportation and applications to goodness-of-fit testing. *arXiv preprint arXiv:1905.05340*.

Goldfeld, Z. and Kato, K. (2020). Limit Distribution Theory for Smooth Wasserstein Distance with Applications to Generative Modeling. *arXiv preprint arXiv:2002.01012*.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*, third ed. The Johns Hopkins University Press, Baltimore and London.

Hallin, M., La Vecchia, D. and Liu, H. (2019). Center-Outward R-Estimation for Semiparametric VARMA Models. *arXiv preprint arXiv:1910.08442*.

Henze, N. and Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics. Theory and Methods* **19** 3595–3617.

Hofert, M., Kojadinovic, I., Maechler, M. and Yan, J. (2018). copula: Multivariate Dependence with Copulas R package version 0.999-19.1.

Horowitz, J. and Karandikar, R. L. (1994). Mean rates of convergence of empirical measures in the Wasserstein distance. *Journal of Computational and Applied Mathematics* **55** 261–273.

Khmaladze, E. V. (2016). Unitary transformations, empirical processes and distribution free testing. *Bernoulli* **22** 563–588.

Kitagawa, J., Mérigot, Q. and Thibert, B. (2017). Convergence of a Newton algorithm for semi-discrete optimal transport. *arXiv preprint arXiv:1603.05579v2*.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari* **4** 83-91.

Korkmaz, S., Goksuluk, D. and Zararsiz, G. (2014). MVN: An R Package for Assessing Multivariate Normality. *The R Journal* **6** 151–162.

Leclaire, A. and Rabin, J. (2019). A Fast Multi-layer Approximation to Semi-discrete Optimal Transport. In *Scale Space and Variational Methods in Computer Vision* (J. Lellmann, M. Burger and J. Modersitzki, eds.) 341–353. Springer International Publishing, Cham.

Lévy, B. (2015). A numerical algorithm for L2 semi-discrete optimal transport in 3D. *ESAIM: Mathematical Modelling and Numerical Analysis* **49** 1693–1715.

McAssey, M. P. (2013). An empirical goodness-of-fit test for multivariate distributions. *Journal of Applied Statistics* **40** 1120–1131.

Mérigot, Q. (2011). A multiscale approach to optimal transport. In *Computer Graphics Forum* **30** 1583–1592. Wiley Online Library.

Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60** 223–241.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual review of statistics and its application* **6** 405–431.

Peyré, G. and Cuturi, M. (2019). Computational Optimal Transport. *Foundations and Trends® in Machine Learning* **11** 355–607.

Ramdas, A., García Trillos, N. and Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy* **19** Paper No. 47, 15.

Rippl, T., Munk, A. and Sturm, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis* **151** 90–109.

Rizzo, M. L. and Székely, G. J. (2016). Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics* **8** 27–38.

Royston, J. P. (1982). An extension of Shapiro and Wilk's W test for normality

to large samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **31** 115–124.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians. Progress in Nonlinear Differential Equations and their Applications* **87**. Birkhäuser/Springer, Cham.

Schuhmacher, D., Bähre, B., Gottschlich, C., Hartmann, V., Heinemann, F. and Schmitzer, B. (2019). transport: Computation of Optimal Transport Plans and Wasserstein Distances R package version 0.12-1.

Shi, H., Drton, M. and Han, F. (2019). Distribution-free consistent independence tests via Hallin's multivariate rank. *arXiv preprint arXiv:1909.10024*.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou* **2** 3-14.

Smith, S. P. (1995). Differentiation of the Cholesky Algorithm. *Journal of Computational and Graphical Statistics* **4** 134–147.

Sommerfeld, M. and Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 219–238.

Tameling, C., Sommerfeld, M. and Munk, A. (2019). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability* **29** 2744–2781.

R Core Team (2018). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer Sciences+Business Media, New York.

Villani, C. (2009). *Optimal Transport: Old and New*. Springer-Verlag, Berlin.

von Mises, R. E. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit*. Julius Springer.

Appendix A: Uniform convergence of the empirical Wasserstein distance

The aim of this appendix is to establish the convergence to zero in probability, uniformly in the underlying distribution $\mathrm{P} \in \mathcal{M}$, of the empirical Wasserstein distance $W_p(\widehat{\mathrm{P}}_n, \mathrm{P})$ when $\mathcal{M} \subseteq \mathcal{P}_p(\mathbb{R}^d)$ has a compact W_p-closure. Actually, Theorem 1 establishes the stronger result that the convergence to zero holds uniformly in the p-th mean. The Markov inequality then implies (Corollary 1) the desired uniform convergence in probability. The notation is that of Section 1.2, with $\mathbb{E}_{\mathrm{P}}$ standing for expectation under an independent random sample from P.

Theorem 1. *Let $\mathcal{M} \subseteq \mathcal{P}_p(\mathbb{R}^d)$ be such that*

$$\lim_{r\to\infty} \sup_{\mathrm{P}\in\mathcal{M}} \int_{\|x\|>r} \|x\|^p \,\mathrm{dP}(x) = 0.$$

Then,

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathcal{M}} \mathbb{E}_{\mathrm{P}} \left\{ W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P}) \right\} = 0.$$

The condition on $\mathcal{M}$ is equivalent to the one that the closure of $\mathcal{M}$ in the metric space $(\mathcal{P}_p(\mathbb{R}^d), W_p)$ is compact. This follows from Prohorov's theorem and the characterization of W_p-convergence in Bickel and Freedman (1981, Lemma 8.3) or Villani (2009, Theorem 6.9). The convergence rate of $\mathbb{E}_{\mathrm{P}}\{W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P})\}$ has been studied intensively; see, for instance, Fournier and Guillin (2015, Theorem 1). However, those rates require the existence of moments of order q higher than p.

Proof of Theorem 1. The following smoothing argument is inspired by the proof of Theorem 1.1 in Horowitz and Karandikar (1994). Let U_σ denote the Lebesgue-uniform distribution on the ball $\{x \in \mathbb{R}^d : \|x\| \le \sigma\}$ in $\mathbb{R}^d$ with radius $\sigma \in (0, \infty)$ and centered at the origin. Denoting by $*$ the convolution of probability measures, we have, for any $\mathrm{Q} \in \mathcal{P}_p(\mathbb{R}^d)$,

$$W_p(\mathrm{Q} * \mathrm{U}_\sigma, \mathrm{Q}) \le \sigma.$$

Indeed, if X and Y are independent random vectors with distributions Q and U_σ, respectively, then $(X + Y, X)$ is a coupling of $\mathrm{Q} * \mathrm{U}_\sigma$ and Q, so that

$$W_p^p(\mathrm{Q} * \mathrm{U}_\sigma, \mathrm{Q}) \le \mathbb{E}[\|Y\|^p] \le \sigma^p.$$

By the triangle inequality, it follows that

$$W_p(\widehat{\mathrm{P}}_n, \mathrm{P}) \le 2\sigma + W_p(\widehat{\mathrm{P}}_n * \mathrm{U}_\sigma, \mathrm{P} * \mathrm{U}_\sigma).$$

Taking expectations and using the elementary inequality

$$(a+b)^p \le 2^{p-1}(a^p + b^p) \quad \text{for } p \ge 1,\ a \ge 0, \quad \text{and } b \ge 0,$$

we obtain

$$\mathbb{E}_{\mathrm{P}} \left\{ W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P}) \right\} \le 2^{p-1} \left[2^p \sigma^p + \mathbb{E} \left\{ W_p^p(\widehat{\mathrm{P}}_n * \mathrm{U}_\sigma, \mathrm{P} * \mathrm{U}_\sigma) \right\} \right].$$

If we can show that

$$\forall \sigma > 0, \qquad \lim_{n\to\infty} \sup_{\mathrm{P}\in\mathcal{M}} \mathbb{E} \left\{ W_p^p(\widehat{\mathrm{P}}_n * \mathrm{U}_\sigma, \mathrm{P} * \mathrm{U}_\sigma) \right\} = 0, \tag{17}$$

then it will follow that

$$\forall \sigma > 0, \qquad \limsup_{n\to\infty} \sup_{\mathrm{P}\in\mathcal{M}} \mathbb{E}_{\mathrm{P}} \left\{ W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P}) \right\} \le 2^{2p-1}\sigma^p.$$

But then, this latter lim sup is actually a lim and is equal to zero, as required.

Let us proceed to show (17). Fix $\sigma > 0$ for the remainder of the proof. Let f_σ denote the density function of U_σ. The measures $\widehat{\mathrm{P}}_n * \mathrm{U}_\sigma$ and $\mathrm{P} * \mathrm{U}_\sigma$ are absolutely continuous too and have density functions $x \mapsto n^{-1} \sum_{i=1}^n f_\sigma(x - X_i)$ and $x \mapsto \int_{\mathbb{R}^d} f_\sigma(x-y)\mathrm{dP}(y)$, respectively. The Wasserstein distance can be controlled by weighted total variation (Villani, 2009, Theorem 6.15):

$$\begin{aligned} W_p^p(\widehat{\mathrm{P}}_n * \mathrm{U}_\sigma, \mathrm{P} * \mathrm{U}_\sigma) &\le 2^{p-1} \int_{\mathbb{R}^d} \|x\|^p \, \mathrm{d}|\widehat{\mathrm{P}}_n * \mathrm{U}_\sigma - \mathrm{P} * \mathrm{U}_\sigma|(x) \\ &= 2^{p-1} \int_{\mathbb{R}^d} \|x\|^p \left| \frac{1}{n} \sum_{i=1}^n f_\sigma(x - X_i) - \int_{\mathbb{R}^d} f_\sigma(x-y) \, \mathrm{dP}(y) \right| \mathrm{d}x. \end{aligned}$$

Take expectations and apply Fubini's theorem to see that

$$\mathbb{E}_\mathrm{P} \left\{ W_p^p(\widehat{\mathrm{P}}_n * \mathrm{U}_\sigma, \mathrm{P} * \mathrm{U}_\sigma) \right\} \le 2^{p-1} \int_{\mathbb{R}^d} \|x\|^p g_n(x; \mathrm{P}) \, \mathrm{d}x \tag{18}$$

where

$$g_n(x; \mathrm{P}) = \mathbb{E}_\mathrm{P} \left[\left| \frac{1}{n} \sum_{i=1}^n f_\sigma(x - X_i) - \int_{\mathbb{R}^d} f_\sigma(x-y) \, \mathrm{dP}(y) \right| \right].$$

Let $r > \sigma$ and split the integral in (18) according to whether $\|x\| > r$ or $\|x\| \le r$. Note that $f_\sigma(u) = f_\sigma(0)$ if $\|y\| \le \sigma$ and $f_\sigma(u) = 0$ otherwise. For any $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d)$ and any $x \in \mathbb{R}^d$, we have, by the Cauchy–Schwarz inequality,

$$g_n(x; \mathrm{P}) \le n^{-1/2} f_\sigma(0).$$

It follows that

$$\lim_{n \to \infty} \sup_{\mathrm{P} \in \mathcal{P}(\mathbb{R}^d)} \int_{\|x\| \le r} \|x\|^p g_{n;\mathrm{P}}(x) \, \mathrm{d}x = 0.$$

But then, in view of (18), we have

$$\limsup_{n \to \infty} \sup_{\mathrm{P} \in \mathcal{M}} \mathbb{E}_\mathrm{P} \left\{ W_p^p(\widehat{\mathrm{P}}_n * \mathrm{U}_\sigma, \mathrm{P} * \mathrm{U}_\sigma) \right\} \le \limsup_{n \to \infty} \sup_{\mathrm{P} \in \mathcal{M}} 2^{p-1} \int_{\|x\| > r} \|x\|^p g_n(x; \mathrm{P}) \, \mathrm{d}x.$$

By the triangle inequality, we also have, for all n,

$$g_n(x; \mathrm{P}) \le 2 \int_{\mathbb{R}^d} f_\sigma(x-y) \mathrm{dP}(y).$$

Applying Fubini's theorem once more, we find that

$$\begin{aligned} \int_{\|x\| > r} \|x\|^p g_n(x; \mathrm{P}) \, \mathrm{d}x &\le 2 \int_{\|x\| > r} \|x\|^p \int_{y \in \mathbb{R}^d} f_\sigma(x-y) \, \mathrm{dP}(y) \, \mathrm{d}x \\ &= 2 \int_{y \in \mathbb{R}^d} \int_{\|x\| > r} \|x\|^p f_\sigma(x-y) \, \mathrm{d}x \, \mathrm{dP}(y) \\ &= 2 \int_{y \in \mathbb{R}^d} \int_{\|u+y\| > r} \|u+y\|^p f_\sigma(u) \, \mathrm{d}u \, \mathrm{dP}(y). \end{aligned}$$

Since $f_\sigma(u) = 0$ whenever $\|u\| > \sigma$ and since $r > \sigma$, we have

$$\int_{\|u+y\|>r} \|u+y\|^p f_\sigma(u)\,\mathrm{d}u \leq \begin{cases} 2^{p-1}(\sigma^p + \|y\|^p) & \text{if } \|y\| > r-\sigma, \\ 0 & \text{otherwise.} \end{cases}$$

Choosing $r > 2\sigma$, we get that $\|y\| > \sigma$ for all y in the non-zero branch above, and thus, for all n,

$$\int_{\|x\|>r} \|x\|^p g_n(x;\mathrm{P})\,\mathrm{d}x \leq 2^{p+1} \int_{\|y\|>r-\sigma} \|y\|^p\,\mathrm{dP}(y).$$

It follows that, for every $r > \sigma$,

$$\limsup_{n\to\infty} \sup_{\mathrm{P}\in\mathcal{M}} \mathbb{E}_\mathrm{P}\left\{W_p^p(\widehat{\mathrm{P}}_n * \mathrm{U}_\sigma, \mathrm{P} * \mathrm{U}_\sigma)\right\} \leq 2^{2p} \sup_{\mathrm{P}\in\mathcal{M}} \int_{\|y\|>r-\sigma} \|y\|^p\,\mathrm{dP}(y).$$

The left-hand side does not depend on r. The condition on $\mathcal{M}$ implies that the right-hand side converges to zero as $r \to \infty$. It follows that the left-hand side must be equal to zero. But this is exactly (17), as required. The proof is complete. □

Corollary 1. *For $\mathcal{M}$ as in Theorem 1, we have*

$$\forall \varepsilon > 0, \qquad \lim_{n\to\infty} \sup_{\mathrm{P}\in\mathcal{M}} \mathrm{P}^n\left[W_p^p(L_n, \mathrm{P}) > \varepsilon\right] = 0,$$

i.e., $W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P}) \to 0$ in probability as $n \to \infty$, uniformly in $\mathrm{P} \in \mathcal{M}$.

Proof. By Markov's inequality, for every $\varepsilon > 0$ and every $\mathrm{P} \in \mathcal{P}_p(\mathbb{R}^d)$, we have

$$\mathrm{P}^n\left[W_p(L_n, \mathrm{P}) > \varepsilon\right] \leq \varepsilon^{-p} \int_{(\mathbb{R}^d)^n} W_p^p(L_n, \mathrm{P})\,\mathrm{dP}^n.$$

In view of Theorem 1, the integral converges to zero uniformly in $\mathrm{P} \in \mathcal{M}$. □

Appendix B: Algorithms for the computation of critical values

Our test statistics involve the Wasserstein distance between an empirical measure and a continuous one. Their calculation requires solving a semi-discrete optimal transport problem (Section 1.3), for which we relied on the function semidiscrete in the R package transport (Schuhmacher et al., 2019), which implements the method of Mérigot (2011). The method starts from a discretization of the source density. The quality of approximation can be set by choosing a sufficiently fine mesh and selecting the tolerance parameter to a low value. The meshes considered here consisted of approximately 10^5 cells.

Below we provide pseudo-code algorithms to sketch the main steps in the actual computation of the critical values. We start with the case of a simple null hypothesis (Algorithm 1), then turn to elliptical families with a given

generator (Algorithm 2) and finally propose the bootstrap procedure for a general parametric family (Algorithm 3). The empirical distribution associated with a sample $\mathbf{X} = (X_1, \ldots, X_n) \in (\mathbb{R}^d)^n$ is denoted by $\widehat{\mathrm{P}}_n(\mathbf{X})$. The largest integer not larger than a scalar $x \in \mathbb{R}$ is denoted by $\lfloor x \rfloor$.

In Algorithm 3, we first compute $c_{\mathcal{M}}(\alpha, n, \theta)$ for θ in a finite mesh $\Theta_1 \subseteq \Theta$. From these values, we reconstruct the function $\theta \mapsto c_{\mathcal{M}}(\alpha, n, \theta)$ by smoothing. It is into the resulting function that we plug in the actual estimate $\hat{\theta}_n$. Further, we restrict the bootstrap parameter estimates $\hat{\theta}^*_{n,b}$ to be in another finite mesh $\Theta_2 \subset \Theta$, because calculation of the bootstrapped test statistics $T^*_{\mathcal{M},n,b}$ requires a preliminary discretization of the density associated to $\hat{\theta}^*_{n,b}$ in order to solve the corresponding semi-discrete optimal transport problem. The first loop in Algorithm 3 is discretizing the densities of P_θ for $\theta \in \Theta_2$. The second loop is calculating $c_{\mathcal{M}}(\alpha, n, \theta)$ for $\theta \in \Theta_1$ by drawing B samples of size n from P_θ. The final step of the algorithm consists of reconstructing the function $\theta \mapsto c_{\mathcal{M}}(\alpha, n, \theta)$ by smoothing. This smoothing step is illustrated in Figure 7 for the five-parameter bivariate Gaussian–AMH model in Section 5.3, applying the location–scale reduction in Remark 3.

The quality of the approximate critical thresholds is ensured by choosing a large enough number of Monte Carlo replications N (Algorithms 1 and 2) or bootstrap replicates B (Algorithm 3). In the simulation experiments, we chose N between 3 000 and 10 000 depending on the time required, while $B = 1\,000$.

Algorithm 1: Computation of $c(\alpha, n, \mathrm{P}_0)$ in Eq. (2)

Input:

- A mesh that supports the source density f associated to P_0
- A number of replications N
- A sample size n
- A level α

Output: An approximation of $c(\alpha, n, \mathrm{P}_0)$

$T \leftarrow [0, \ldots, 0] \in \mathbb{R}^N$ `// Initialization`
for $i = 1$ *to* N **do**
  $\mathbf{X} \leftarrow \mathrm{rand}(n, \mathrm{P}_0)$ `// Generation of sample of size` n `from` P_0
  $T[i] \leftarrow W_2^2(\widehat{\mathrm{P}}_n(\mathbf{X}), \mathrm{P}_0)$
$\mathrm{sort}(T)$
$c(\alpha, n, \mathrm{P}_0) \leftarrow T[\lfloor (1-\alpha)N \rfloor]$ `// Empirical quantile (order statistic)`

Algorithm 2: Computation of $c_{\mathcal{E}}(\alpha, n, f_{\text{rad}})$ in Eq. (6)

Input:

- A mesh that supports the source density f associated to P_0.
- A number of replications N
- A sample size n
- A level α

Output: An approximation of $c_{\mathcal{E}}(\alpha, n, f_{\text{rad}})$

$T \leftarrow [0, \dots, 0] \in \mathbb{R}^N$ `// Initialization`
for $i = 1$ *to* N **do**
  $\mathbf{X} \leftarrow \text{rand}(n, \mathrm{P}_{f_{\text{rad}}})$ `// Generation of sample of size` n `from` $\mathrm{P}_{f_{\text{rad}}}$
  $\hat{\mathbf{Z}} \leftarrow \text{standardize}(\mathbf{X}, \text{method} = \text{“Cholesky”})$ `// Residuals as in (5)`
  $T[i] \leftarrow W_2^2(\widehat{\mathrm{P}}_n(\hat{\mathbf{Z}}), \mathrm{P}_{f_{\text{rad}}})$
sort(T)
$c_{\mathcal{E}}(\alpha, n, f_{\text{rad}}) \leftarrow T[\lfloor (1-\alpha)N \rfloor]$

Algorithm 3: Computation of $\theta \mapsto c_{\mathcal{M}}(\alpha, n, \theta)$ in (10)

Input:

- A finite set $\Theta_1 \subseteq \Theta$ of values of θ at which to calculate $c_{\mathcal{M}}(\alpha, n, \theta)$ initially
- A larger finite set $\Theta_2 \subseteq \Theta$ into which to force the bootstrapped estimates $\hat{\theta}^*_{n,b}$
- A number of bootstrap replications B
- A sample size n
- A level α

Output: An approximation of $\theta \mapsto c_{\mathcal{M}}(\alpha, n, \theta)$

$T \leftarrow 0 \in \mathbb{R}^{N \times |\Theta_1|}$ `// Initialization`
for θ *in* Θ_2 **do**
  $\text{Grid}[\theta] = \text{discretization}(\mathrm{P}_\theta)$ `// Discretization of density of` P_θ
for θ *in* Θ_1 **do**
  for $b = 1$ *to* B **do**
    $\mathbf{X} \leftarrow \text{rand}(n, \mathrm{P}_\theta)$
    $\hat{\theta}^*_n \leftarrow \theta_n(\mathbf{X})$ `// While ensuring` $\hat{\theta}^*_n \in \Theta_2$
    $T[b, \theta] \leftarrow W_2^2(\widehat{\mathrm{P}}_n(X), \mathrm{P}_{\hat{\theta}^*_n})$ `// Requires` $\text{Grid}[\hat{\theta}^*_n]$
ColumnSort(T) `// For each` $\theta \in \Theta_1$`, sort` $T[\,\cdot\,, \theta]$
for θ *in* Θ_1 **do**
  $c_{\mathcal{M}}(\alpha, n, \theta) \leftarrow T[\lfloor (1-\alpha)N \rfloor, \theta]$
$c_{\mathcal{M}}(\alpha, n, \cdot\,) \leftarrow \text{Smooth}\left((\theta, c_{\mathcal{M}}(\alpha, n, \theta)) : \theta \in \Theta_1\right)$ `// function` $\theta \mapsto c_{\mathcal{M}}(\alpha, n, \theta)$

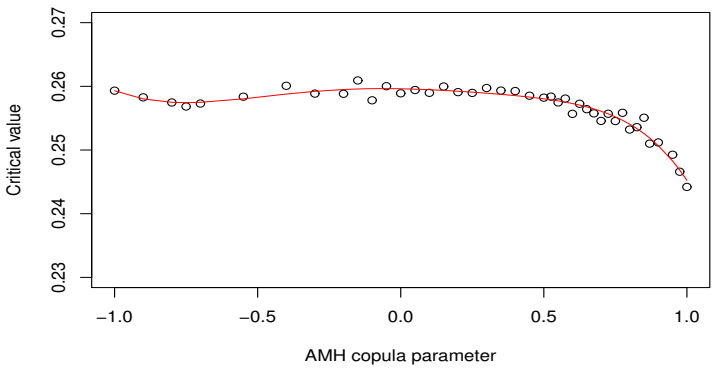

FIG 7. *Illustration of the last step in Algorithm 3 for the bivariate five-parameter Gaussian–AMH model in Section 5.3 using the location–scale reduction in Remark 3. The function* $\theta \mapsto c^{\mathrm{ls}}_{\mathcal{M}}(\alpha, n, \theta)$ *(in red) is constructed by smoothing Monte Carlo estimates (circles) of* $c^{\mathrm{ls}}_{\mathcal{M}}(\alpha, n, \theta)$ *for* $\theta \in \Theta_1 \subseteq \Theta = [-1, 1]$*, with* $\alpha = 0.05$*,* $n = 200$ *and* $B = 1\,000$ *samples per point. The smoother is a 6th-degree polynomial fitted by ordinary least squares.*

Appendix C: A banana-shaped distribution

The "banana-shaped" distribution in Section 5.1 and Figure 1(f) is a mixture

$$(1-2p)\,\mathcal{N}_2\left(\begin{pmatrix}0\\-0.7\end{pmatrix}, \begin{pmatrix}0.35^2 & 0\\0 & 0.35^2\end{pmatrix}\right) + p\,\mathcal{N}_2\left(\begin{pmatrix}-0.9\\0.3\end{pmatrix}, \begin{pmatrix}0.358 & -0.55\\-0.55 & 1.02\end{pmatrix}\right) + p\,\mathcal{N}_2\left(\begin{pmatrix}0.9\\0.3\end{pmatrix}, \begin{pmatrix}0.358 & 0.55\\0.55 & 1.02\end{pmatrix}\right). \tag{19}$$

of three Gaussian components. Figure 8 shows a scatterplot for $p = 0.35$ of a random sample of size $n = 500$ from this distribution.

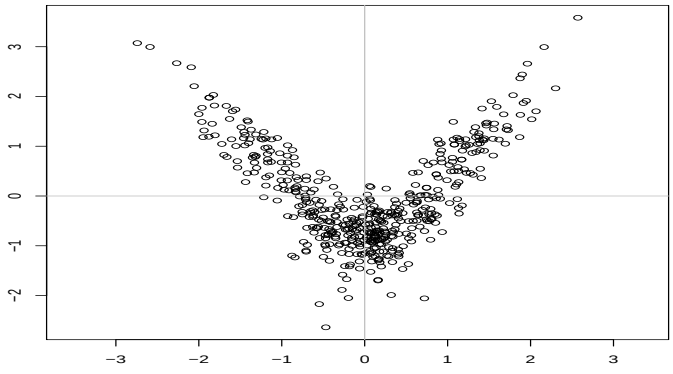

FIG 8. *Scatterplot of a sample of size* 500 *from the "banana-shaped" mixture* (19).