Electronic Journal of Statistics
Vol. 15 (2021) 1328–1371
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1816

Multivariate goodness-of-fit tests based on Wasserstein distance[*]

Marc Hallin

ECARES and Département de Mathématique, Université libre de Bruxelles
Avenue F.D. Roosevelt 50, 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. For group families, the procedure is to be implemented after preliminary reduction of the data via invariance. This property allows for calculation of exact critical values and p-values at finite sample sizes. Applications include testing for location–scale families and testing for families arising from affine transformations, such as elliptical distributions with given standard radial density and unspecified location vector and scatter matrix. A novel test for multivariate normality with unspecified mean vector and covariance matrix arises as a special case. For more general parametric families, we propose a parametric bootstrap procedure to calculate critical values. The lack of asymptotic distribution theory for the empirical Wasserstein distance means that the validity of the parametric bootstrap under the null hypothesis remains a conjecture. Nevertheless, we show that the test is consistent against fixed alternatives. To this end, we prove a uniform law of large numbers for the empirical distribution in Wasserstein distance, where the uniformity is over any class of underlying distributions satisfying a uniform integrability condition but no additional moment assumptions. The calculation of test statistics boils down to solving the well-studied semi-discrete optimal transport problem. Extensive numerical experiments demonstrate the practical feasibility and the excellent performance of the proposed tests for the Wasserstein distance of order $p = 1$ and $p = 2$ and for dimensions at least up to $d = 5$. The simulations also lend support to the conjecture of the asymptotic validity of the parametric bootstrap.

Keywords and phrases: Copula, elliptical distribution, goodness-of-fit, group families, multivariate normality, optimal transport, semi-discrete problem, skew-t distribution, Wasserstein distance.

Received October 2020.

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

Contents

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 of moving probability mass in order to transform one distribution into the other. 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 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 null 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 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. Such hypotheses are generally more realistic than simple ones. Of particular practical importance is the case of location–scale and location–scatter 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 the empirical distribution of the data or estimated residuals on the one hand and a model-based estimate thereof on the other hand. We concentrate on the continuous case, i.e., the distributions under the null hypothesis are absolutely continuous with respect to the d-dimensional

Lebesgue measure. The test statistic involves the Wasserstein distance between a discrete empirical distribution and a continuous distribution specified by the null hypothesis. Calculating such a distance requires solving the 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 group family, that is, a model that arises by applying a transformation group to some specified distribution (Lehmann and Casella, 1998, pp. 16–23). If the parameter estimate is equivariant, the data can be reduced in such a way that their distribution no longer depends on the unknown parameter. The Wasserstein distance between this parameter-free distribution and the empirical distribution of the reduced data then provides a test statistic whose null distribution does not depend on the unknown parameter either. Important special cases include elliptical distributions with known radial distribution and unknown location vector and scatter matrix. In particular, our approach yields a novel test for multivariate normality with unknown mean vector and covariance matrix.

For general parametric models, the test statistic measures the Wasserstein distance between the empirical distribution and the model-based one with estimated parameter. A reduction via invariance is no longer possible and we rely on the parametric bootstrap to calculate critical values. Still, some parameters, such as location-scale parameters, can be factored out, again by relying on transformation groups. The question whether the parametric bootstrap has the correct size under the null hypothesis remains open. A proof of that property would require asymptotic distribution theory for the empirical Wasserstein distance—a hard and long-standing open problem, which we briefly review in Section 1.2, the solution of which is beyond the scope of this paper. Monte Carlo experiments, however, support our conjecture that the parametric bootstrap has the correct size at least asymptotically.

In all cases, even in the general parametric case, 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 as the sample size tends to infinity. For the general parametric case, the proof relies on the uniform consistency in probability of the empirical distribution with respect to the Wasserstein distance, uniformly over families of distributions that satisfy a uniform integrability condition. To the best of our knowledge, this result is new.

We conduct an extensive simulation study to assess the finite-sample performance of the Wasserstein tests of order $p \in \{1, 2\}$ in comparison to other GoF tests. The set-up involves both simple and composite null hypotheses as well as a wide variety of alternatives. The experiments lend support to the conjecture that the parametric bootstrap is valid asymptotically. In comparison to other GoF tests available in the literature, the Wasserstein test demonstrates good power. This is especially true for the test of multivariate normality, where, out of the many available tests in the literature, we select the ones of Royston

(1983), Henze and Zirkler (1990) and Rizzo and Székely (2016) as benchmarks.

In a recent strand of literature, measure transportation serves to link a multivariate probability measure to a standard reference distribution, yielding novel concepts of multivariate ranks, signs, and quantiles (Carlier et al., 2016; Chernozhukov et al., 2017; Hallin et al., 2020). Here we do not make this step, as the Wasserstein distances we are considering are between distributions defined on the sample space.

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 transportation 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 some group family. Composite null hypotheses covering general parametric models are treated in Section 4. In Section 4.1 we mention a hybrid approach, where some components of the parameter vector are factored out by relying on a transformation group. In Section 5, finally, we report on the results of our numerical experiments. In Appendix A, the convergence of the empirical Wasserstein distance uniformly over certain classes of underlying distributions is stated and proved. Appendix B is related to the consistency of the parametric bootstrap. Appendices C–E contain further details on the simulation study.

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 P and 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 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 and Y with laws P and Q, respectively, the p-Wasserstein distance is the smallest value of $\{\mathbb{E}(\|X - Y\|^p)\}^{1/p}$ over all possible joint distributions $\gamma \in \Gamma(\mathrm{P}, \mathrm{Q})$ of (X, Y).

The p-Wasserstein distance W_p defines a metric on $\mathcal{P}_p(\mathbb{R}^d)$, which thereby becomes 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).

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

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

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), see for instance Panaretos and Zemel (2019, Section 1.2.3).

1.2. Asymptotic theory: results and an open problem

To construct critical values for 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 requires having, under contiguous alternatives, non-degenerate limit distributions of the statistic of interest (Beran, 1997; Capanu, 2019). For Wasserstein distances involving empirical distributions, such results are still far beyond the horizon, as the following short survey will show.

Let $X_1, \ldots, X_n$ be an i.i.d. (independent and identically distributed) sample from $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d)$. The empirical distribution of the sample is $\widehat{\mathrm{P}}_n := n^{-1} \sum_{i=1}^n \delta_{X_i}$, with δ_x the Dirac measure at x. Assuming that P has a finite moment of order $p \in [1, \infty)$, we are interested in the empirical Wasserstein distance $W_p(\widehat{\mathrm{P}}_n, \mathrm{P})$.

According to Bickel and Freedman (1981, Lemma 8.4), 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 $n \to \infty$. Bounds and rates for the expectation of the empirical Wasserstein distance 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, even in the one-dimensional case (Bobkov and Ledoux, 2019, Theorem 3.3). Precise rates under additional moment assumptions are given, for instance, in Fournier and Guillin (2015). In Appendix A, we will show that the convergence in pth mean takes place uniformly over families $\mathcal{M} \subset \mathcal{P}_p(\mathbb{R}^d)$ of probability measures satisfying a uniform integrability condition. For distributions on compact metric spaces, Weed and Bach (2019) provide sharp rates for $\mathbb{E}[W_p(\widehat{\mathrm{P}}_n, \mathrm{P})]$ in terms of what they coin the *Wasserstein dimension* of P. For Lebesgue-absolutely continuous measures on $\mathbb{R}^d$, this dimension is just d. Moreover, they exploit McDiarmid's bounded difference inequality to derive a concentration inequality of $W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P})$ around its expectation.

Asymptotic results on the distribution of the empirical Wasserstein distance in dimension $d \geq 2$ are, however, surprisingly scarce. The question is whether there exist sequences $a_n > 0$ and $b_n \geq 0$ such that $a_n\{W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P}) - b_n\}$ converges in distribution to a non-degenerate limit. Although this problem has already attracted a lot of attention, a general answer remains elusive.

The one-dimensional case is well-studied thanks to the link (1) to empirical quantile processes (del Barrio, Giné and Utzet, 2005; Bobkov and Ledoux, 2019). For discrete distributions, large-sample theory for the empirical Wasserstein distance is available too (Sommerfeld and Munk, 2018; Tameling, Sommerfeld and Munk, 2019). For multivariate Gaussian distributions, a central limit theorem for the empirical Wasserstein of order $p = 2$ between the true distribution and the one with estimated parameters is given in Rippl, Munk and Sturm (2016). Although interesting and useful for GoF testing (see Section 5.1.1), this result does not cover the empirical distribution $\widehat{\mathrm{P}}_n$.

Ambrosio, Stra and Trevisan (2018) exploit the possibility to linearize the 2-Wasserstein distance in dimension $d = 2$ in case the optimal transport plan is close to the identity. The technique requires balancing the errors due to the dual Sobolev norm approximation and a smoothing step. Mena and Niles-Weed (2019) derive a limit theorem for the empirical entropic optimal transport cost. We refer to the latter for an introduction to optimal transport with entropic regularization. Recent progress has been booked in Goldfeld and Kato (2020), who obtain a central limit theorem for the empirical 1-Wasserstein distance after smoothing the empirical and the true distributions with a Gaussian kernel.

Important advances on the limit distribution have been made by del Barrio and Loubes (2019) who obtained results under fixed 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

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

The result is proved using the Efron–Stein inequality combined with stability of optimal transport plans. Unfortunately, if $\mathrm{Q} = \mathrm{P}$, 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 smaller than $n^{-1/2}$. The authors conclude that their proof technique is of little use for the case we are interested in. The crucial problem of the limiting distribution of the empirical Wasserstein distance thus remains an important and difficult open problem.

1.3. Computational issues

In the last decade, important numerical developments have taken place in the area of measure transportation. The problem to be faced here is the computation of the Wasserstein distance between a discrete and a continuous distribution, the so-called semi-discrete optimal transportation problem. 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 $\mathbb{R}^d$; see, e.g., Santambrogio (2015, Section 6.4.2) for a didactic exposition. This formulation is the basis for the multi-scale algorithm for the squared Euclidean distance ($p = 2$) developed in Mérigot (2011), with further improvements in Lévy (2015) and Kitagawa, Mérigot and Thibert (2017). It requires constructing a power diagram or Laguerre–Voronoi diagram, partitioning $\mathbb{R}^d$ into convex polyhedra called power cells. With the Euclidean distance as cost function ($p = 1$) the edges of the cells involved in the tessellation are no longer linear, making the computation more demanding (Hartmann and Schuhmacher, 2020). Genevay et al. (2016) show that a semi-discrete reformulation of the dual program can be tackled by the stochastic averaged gradient (SAG) method (Schmidt, Le Roux and Bach, 2017).

In our numerical experiments in Section 5, we assess the finite-sample performance of the test statistic based on the p-Wasserstein distance for $p \in \{1, 2\}$ and for d-variate distributions for $d \in \{2, 5\}$. To the best of our knowledge, an implementation of the SAG method is not yet available in R (R Core Team, 2018). After preliminary tests and running time assessment, we made the following choices of algorithms and implementations:

- In case $p = 2$ and $d = 2$, we relied on the R package transport (Schuhmacher et al., 2019), which implements the multi-scale algorithm in Mérigot (2011).
- In all other cases ($p = 1$ or $d = 5$), we relied on our own C implementation of the SAG method as employed in Genevay et al. (2016).

A first version of the package making our implementation available is to be found on `https://github.com/gmordant/WassersteinGoF`.

2. Wasserstein GoF tests for simple null hypotheses

Let $\mathbf{X}_n = (X_1, \dots, X_n)$ 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.

Let $\widehat{\mathrm{P}}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ denote the empirical distribution and consider the test statistic

$$T_n := W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P}_0), \tag{2}$$

the pth power of the p-Wasserstein distance between $\widehat{\mathrm{P}}_n$ 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. In the numerical experiments of Section 5, we provide results for $p \in \{1, 2\}$. The theory, however, is developed for general $p \geq 1$.

Let $F_n(t) = \mathrm{P}_0^n[T_n \leq t]$ for $t \in [0, \infty)$ denote the distribution function of the test statistic under $\mathcal{H}_0^n$. Here, P_0^n stands for the distribution under $\mathcal{H}_0^n$ of the observation $\mathbf{X}_n$, the n-fold product measure of P_0 on $(\mathbb{R}^d)^n$. The p-value of the test statistic is $1 - F_n(T_n)$, while the critical value for a test of size $\alpha \in (0, 1)$ is

$$c_n(\alpha, \mathrm{P}_0) := \inf \left\{ t > 0 : F_n(t) \geq 1 - \alpha \right\} \tag{3}$$

The test we propose is then

$$\phi^n_{\mathrm{P}_0} = \begin{cases} 1 & \text{if } 1 - F_n(T_n) \leq \alpha \text{ or, equivalently, } T_n \geq c_n(\alpha, \mathrm{P}_0), \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The exact size of the GoF test in (4) is $1 - F_n(c_n(\alpha, \mathrm{P}_0)) \leq \alpha$, with equality if and only if F_n is continuous at $c_n(\alpha, \mathrm{P}_0)$. The type I error is thus bounded by the nominal size α, and often equal to it. The null distribution of T_n depends on P_0, so that $c_n(\alpha, \mathrm{P}_0)$ needs to be calculated for each P_0 separately.

Although p-values and critical values usually cannot be calculated analytically, they can be approximated with any desired degree of precision via the following simple Monte Carlo algorithm. Draw a large number of independent random samples of size n from P_0, compute the test statistic for each such sample, and approximate F_n by the empirical distribution function of the simulated test statistics. Critical values and p-values then can be calculated from the approximated F_n. By the Donsker theorem, any desired accuracy can be achieved by drawing sufficiently many samples.

Under the alternative hypothesis, the test rejects the null hypothesis 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* $\phi^n_{\mathrm{P}_0}$ *is consistent against any* $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d)$ *with* $\mathrm{P} \neq \mathrm{P}_0$*:*

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

Proof. Fix $\mathrm{P}_0 \in \mathcal{P}_p(\mathbb{R}^d)$. 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^n_0-probability and thus $\lim_{n\to\infty} \mathrm{P}^n_0[T_n > \varepsilon] = 0$ for any $\varepsilon > 0$. It follows that, for every $\alpha > 0$ and every $\varepsilon > 0$, we have $c_n(\alpha, \mathrm{P}_0) \leq \varepsilon$ 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 $\lim_{n\to\infty} \mathrm{P}^n[T_n > c_n(\alpha, \mathrm{P}_0)] = 1$, 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 expected value of $\|X_1\|^p$ 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 $n \to \infty$. The same then holds for T_n and thus

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

3. Wasserstein GoF tests for group families

Let $\mathrm{Q}_0 \in \mathcal{P}(\mathbb{R}^d)$ and let G be a group of measurable transformations g of $\mathbb{R}^d$. That is, G is closed under composition ($g_1, g_2 \in G$ implies $g_1 \circ g_2 \in G$) and inversion ($g \in G$ implies $g^{-1} \in G$). If the random variable Z has distribution Q_0, the random variable $g(Z)$ has distribution $g_\# \mathrm{Q}_0 := \mathrm{Q}_0 \circ g^{-1}$, where the subscript $\#$ denotes the push-forward of a measure by a measurable function. Let $\mathcal{M} = \{g_\# \mathrm{Q}_0 : g \in G\}$ be the group family generated by G and Q_0. We assume further that the transformation g is identifiable, that is, the map $g \mapsto g_\# \mathrm{Q}_0$ is one-to-one, so that $g_1 \neq g_2$ implies that $g_1(Z)$ and $g_2(Z)$ have different distributions, with again $Z \sim \mathrm{Q}_0$. Note that for any element P of $\mathcal{M}$ we have $\mathcal{M} = \{g_\# \mathrm{P} : g \in G\}$, so that the choice of Q_0 in $\mathcal{M}$ is in some sense arbitrary.

Group families form one of the two principal classes of models covered in Lehmann and Casella (1998). Here are some prominent examples of transformation groups G on $\mathbb{R}^d$ and some models $\mathcal{M}$ that they generate.

Example 1 (Location–scale families). For $(a, b) \in \mathbb{R}^d \times (0, \infty)^d$, consider the transformations $g_{a,b} : \mathbb{R}^d \to \mathbb{R}^d$ defined by $g_{a,b}(x) = (a_j + b_j x)_{j=1}^d$ for $x \in \mathbb{R}^d$. The model $\mathcal{M}$ is the location-scale family generated by P_0. We can also consider the location family generated by the subgroup $x \mapsto g_{a,1}(x) = (x_j + a_j)_{j=1}^d$ and the scale family generated by the subgroup $x \mapsto g_{0,b}(x) = (b_j x_j)_{j=1}^d$. In dimension $d = 1$, we can generate in this way the normal and exponential families, for instance.

Example 2 (Affine transformations and elliptical families). For $a \in \mathbb{R}^d$ and nonsingular $B \in \mathbb{R}^d \times \mathbb{R}^d$, define $g_{a,B} : \mathbb{R}^d \to \mathbb{R}^d$ by $g_{a,B}(x) = a + Bx$ for $x \in \mathbb{R}^d$. If Q_0 is the d-variate standard normal distribution $\mathcal{N}_d(0, I_d)$, then $\mathcal{M}$ is the family of all d-variate Gaussian distributions with positive definite covariance matrix. More generally, if Q_0 is spherically symmetric around the origin, then $\mathcal{M}$ is the family of elliptical distributions with a given characteristic generator and positive definite scatter matrix (Cambanis, Huang and Simons, 1981; Fang, Kotz and Ng, 1990). Besides the Gaussian family, another common example is the multivariate Student t distribution with a fixed number of degrees of freedom. For elliptical families, the matrix B is not identifiable from the model but only the matrix BB' is. Identifiability can be restored by restricting B to the set of lower triangular matrices with positive elements on the diagonal.[1] Note that the case of elliptical distributions with possibly degenerate scatter matrices is not covered here, as the corresponding affine transformation is not invertible.

[1] 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).

In the examples above, the group G is parametrized by a Euclidean parameter $\theta \in \Theta$ with $\Theta \subseteq \mathbb{R}^k$ for some dimension k, so that $G = \{g_\theta : \theta \in \Theta\}$. We will assume this to be the case in general and write $\mathrm{P}_\theta = (g_\theta)_\# \mathrm{Q}_0$. The model then takes the form $\mathcal{M} = \{\mathrm{P}_\theta : \theta \in \Theta\}$. The mappings $\theta \mapsto g_\theta$ and $g \mapsto g_\# \mathrm{Q}_0$ are assumed to be one-to-one. The parametrization $\theta \mapsto \mathrm{P}_\theta$ then is also one-to-one, i.e., the model parameter θ is identifiable. Models generated by infinite-dimensional transformation groups exist as well, but the theory here is intended for the finite-dimensional situation, as the conditions to come seem too restrictive otherwise.

Let $\mathrm{Q}_0 \in \mathcal{P}_p(\mathbb{R}^d)$ for some $p \in [1, \infty)$. Assume that, for all $g \in G$, there exists $c_g > 0$ such that

$$\forall x \in \mathbb{R}^d, \qquad \|g(x)\| \leq c_g(1 + \|x\|). \tag{5}$$

Then it is easy to verify that $g_\# \mathrm{Q}_0$ belongs to $\mathcal{P}_p(\mathbb{R}^d)$ for every $g \in G$ too and, therefore, $\mathcal{M} \subset \mathcal{P}_p(\mathbb{R}^d)$. This condition on g is fulfilled for the transformations in Examples 1 and 2.

Let $\mathcal{M} = \{\mathrm{P}_\theta = (g_\theta)_\# \mathrm{Q}_0 : \theta \in \Theta\} \subset \mathcal{P}_p(\mathbb{R}^d)$ be a group family as just described. Given an i.i.d. sample $\mathbf{X}_n = (X_1, \dots, X_n)$ from some unspecified $\mathrm{P} \in \mathcal{P}_p(\mathbb{R}^d)$, we wish to test the hypothesis

$$\mathcal{H}_0^n : \mathrm{P} \in \mathcal{M} \quad \text{against} \quad \mathcal{H}_1^n : \mathrm{P} \notin \mathcal{M}. \tag{6}$$

The parameter θ of the transformation g_θ is an unknown nuisance. In contrast to Section 2, the null hypothesis is thus a composite one. An important special case is when Q_0 is the d-variate standard normal distribution and G is the affine group in Example 2: the testing problem (6) then concerns the hypothesis of multivariate normality with unspecified positive definite covariance matrix.

Our testing strategy is to choose some estimator $\hat{\theta}_n$ for θ and compute "residuals" of the form

$$\hat{Z}_{n,i} := g_{\hat{\theta}_n}^{-1}(X_i), \qquad i = 1, \dots, n, \tag{7}$$

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

$$T_{\mathcal{M},n} := W_p^p\big(\widehat{\mathrm{P}}_n^{\hat{Z}}, \mathrm{Q}_0\big). \tag{8}$$

If the null distribution of $(\hat{Z}_{n,1}, \dots, \hat{Z}_{n,n})$ does not depend on the unkown parameter θ, then we can compute critical values and p-values for $T_{\mathcal{M},n}$ as if the true distribution is Q_0. As in Section 2, the null distribution of $T_{\mathcal{M},n}$ then can be computed up to any desired accuracy via Monte Carlo random sampling from Q_0, and this prior to having observed the sample.

For any $g \in G$, let $\bar{g} : \Theta \to \Theta$ denote the mapping $\theta \mapsto \bar{g}(\theta)$ characterized by $g \circ g_\theta = g_{\bar{g}(\theta)}$, so that $g_\# \mathrm{P}_\theta = \mathrm{P}_{\bar{g}(\theta)}$. The estimator $\hat{\theta}_n = \theta_n(\mathbf{X}_n)$ is said to be *equivariant* (Lehmann and Casella, 1998, Definition 2.5) if for every $g \in G$ and for every $(x_1, \dots, x_n) \in (\mathbb{R}^d)^n$, we have

$$\theta_n\big(g(x_1), \dots, g(x_n)\big) = \bar{g}\big(\theta_n(x_1, \dots, x_n)\big). \tag{9}$$

Equivariance is a natural symmetry requirement and is satisfied for many common estimators. For a location parameter, it is satisfied by the mean and the median, or in fact any weighted average of the order statistics. For a scale parameter, it is satisfied by the standard deviation and by the mean or median absolute deviation. For the affine group in Example 2 with B restricted to be lower triangular and with positive diagonal elements, it is satisfied by the lower Cholesky triangle of the empirical covariance matrix. The proof of the latter is elementary and follows from the uniqueness of the Cholesky decomposition and the fact that the set of lower triangular matrices with positive diagonal elements forms a multiplicative group. Equivariance is also satisfied by maximum likelihood estimators provided the transformations g are diffeormorphisms: by the change-of-variables formula, $\hat{\theta}_n$ maximizes the likelihood given the sample $x_1, \ldots, x_n$ if and only if $\bar{g}(\hat{\theta}_n)$ maximizes the likelihood given the sample $g(x_1), \ldots, g(x_n)$.

In the group model $\mathcal{M} = \{\mathrm{P}_\theta = (g_\theta)_\# \mathrm{Q}_0 : \theta \in \Theta\}$, if the estimator $\hat{\theta}_n$ is equivariant, then for an i.i.d. sample $X_1, \ldots, X_n$ from $\mathrm{P}_\theta \in \mathcal{M}$, the joint distribution of $(\hat{Z}_{n,i})_{i=1}^n$ in (7) does not depend on $\theta \in \Theta$ and is the same as if $X_1, \ldots, X_n$ were an i.i.d. sample from Q_0.

As a consequence, the distribution of $T_{\mathcal{M},n}$ in (8) under any $\mathrm{P}_\theta \in \mathcal{M}$ is the same as under Q_0. Let $F_{\mathcal{M},n}$ denote its cumulative distribution function. The p-value of the observed test statistic is

$$1 - F_{\mathcal{M},n}(T_{\mathcal{M},n})$$

while the critical value at level $\alpha \in (0,1)$ is

$$c_{\mathcal{M},n}(\alpha) = \inf\{c > 0 : F_{\mathcal{M},n}(c) \geq 1 - \alpha\}.$$

For the testing problem (6), we propose the test

$$\phi^n_{\mathcal{M}} := \begin{cases} 1 & \text{if } 1 - F_{\mathcal{M},n}(T_{\mathcal{M},n}) \leq \alpha, \text{ or equivalently, } T_{\mathcal{M},n} \geq c_{\mathcal{M},n}(\alpha), \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The actual size of the test is $1 - F_{\mathcal{M},n}(c_{\mathcal{M},n}(\alpha)) \leq \alpha$, with equality if and only if $F_{\mathcal{M},n}$ is continuous in $c_{\mathcal{M},n}(\alpha)$. Formally, the case of a single null hypothesis in Section 2 can be seen as a special case by letting $\mathrm{P}_0 = \mathrm{Q}_0$ and G the trivial group containing only the identity mapping.

Since the null distribution $F_{\mathcal{M},n}$ does not depend on any unknown parameter, critical values and p-values values can be computed with arbitrary precision by a Monte Carlo algorithm as we did in Section 2. The difference is now that we generate samples from Q_0. Note that the critical values can be computed prior to having seen the data.

To show that the test is consistent, we need an extra assumption on G: for every $\theta \in \Theta$ there exists $m_\theta > 0$ such that for all $\theta' \in \Theta$ in some neighbourhood of θ, we have

$$\sup_{x \in \mathbb{R}^d} \frac{\|g_{\theta'}^{-1} \circ g_\theta(x) - x\|}{1 + \|x\|} \leq m_\theta \|\theta' - \theta\|. \tag{11}$$

The condition is fulfilled for the transformation groups and parametrizations in Examples 1 and 2. For the affine group in Example 2, the property (11) follows from continuity of matrix inversion with respect to the matrix norm induced by the Euclidean norm. We will also need weak consistency of the estimator: for every $\theta \in \Theta$, we have $\hat{\theta}_n \to \theta$ as $n \to \infty$ in P_θ^n-probability. To prove this for the affine group in Example 2, it is helpful to know that the map that sends a positive definite symmetric matrix to its Cholesky triangle is differentiable (Smith, 1995) and thus continuous.

Proposition 2 (Consistency against fixed alternatives)**.** *Let the group family $\mathcal{M} = \{\mathrm{P}_\theta = (g_\theta)_\#\mathrm{Q}_0 : \theta \in \Theta\} \subset \mathcal{P}_p(\mathbb{R}^d)$ be such that θ is identifiable as above and such that (5) and (11) are satisfied. Let $\hat{\theta}_n$ be an equivariant and weakly consistent estimator sequence of $\theta \in \Theta$.*

(i) We have $T_{\mathcal{M},n} \to 0$ as $n \to \infty$ in P_θ^n-probability for any $\theta \in \Theta$.

(ii) Let $\mathrm{P} \in \mathcal{P}_p(\mathbb{R}^d) \setminus \mathcal{M}$. If $\hat{\theta}_n$ converges weakly to some $\theta \in \Theta$ under P^n, then $\mathrm{P}^n[\phi_{\mathcal{M}}^n = 1] \to 1$ as $n \to \infty$.

The pseudo-parameter θ in Proposition 2(ii) depends on the estimator: for instance, in a location-scale model and if $p \geq 2$, if we estimate the location and scale parameters by the empirical mean and standard deviation, respectively, then θ denotes the vector of population means and standard deviations.

Proof. (i) The sample $(X_1, \ldots, X_n)$ is equal in distribution to $(g_\theta(Z_i))_{i=1}^n$ for some $\theta \in \Theta$, where $(Z_i)_{i=1}^n$ is an i.i.d. sample from Q_0. Since we are interested in convergence in probability, we can then in fact suppose that $X_i = g_\theta(Z_i)$ for all $i = 1, \ldots, n$ and compute probabilities under Q_0^n.

The empirical distribution of $(Z_i)_{i=1}^n$ is denoted by $\widehat{\mathrm{P}}_n^Z$. By the triangle inequality for the Wasserstein distance,

$$T_{\mathcal{M},n}^{1/p} = W_p\big(\widehat{\mathrm{P}}_n^{\hat{Z}}, \mathrm{Q}_0\big) \leq W_p\big(\widehat{\mathrm{P}}_n^{\hat{Z}}, \widehat{\mathrm{P}}_n^Z\big) + W_p\big(\widehat{\mathrm{P}}_n^Z, \mathrm{Q}_0\big). \tag{12}$$

Since Q_0 has a finite moment of order p, the second term on the right-hand side converges to zero in probability by Bickel and Freedman (1981, Lemma 8.4).

To bound the first term on the right-hand side of the previous equation, 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 $i = 1, \ldots, n$. It follows that

$$W_p^p\big(\widehat{\mathrm{P}}_n^Z, \mathrm{Q}_0\big) \leq \frac{1}{n}\sum_{i=1}^n \|\hat{Z}_i - Z_i\|^p = \frac{1}{n}\sum_{i=1}^n \|g_{\hat{\theta}_{n,Z}}^{-1}(Z_i) - Z_i\|^p,$$

where $\hat{\theta}_{n,Z} = \theta_n(Z_1, \ldots, Z_n)$ is the estimated parameter from $(Z_i)_{i=1}^n$. Let $\theta_e \in \Theta$ denote the parameter that corresponds to the identity transformation: $g_{\theta_e}(x) = x$ for all $x \in \mathbb{R}^d$. Then $\mathrm{P}_{\theta_e} = \mathrm{Q}_0$ and, by assumption, $\hat{\theta}_{n,Z} \to \theta_e$ as $n \to \infty$ in probability. Let $\varepsilon > 0$ be small enough so that (11) holds for all $\theta' \in \Theta$ with $\|\theta' - \theta_e\| \leq \varepsilon$. Then, on the event that $\|\hat{\theta}_{n,Z} - \theta_e\| \leq \varepsilon$, we have

$$\frac{1}{n}\sum_{i=1}^n \|g_{\hat{\theta}_{n,Z}}^{-1}(Z_i) - Z_i\|^p \leq m_{\theta_e}^p \|\hat{\theta}_{n,Z} - \theta_e\|^p \cdot \frac{1}{n}\sum_{i=1}^n (1 + \|Z\|_i)^p\,.$$

As $\mathrm{Q}_0 \in \mathcal{P}_p(\mathbb{R}^d)$, the weak consistency of $\hat{\theta}_{n,Z}$ and the law of large numbers imply that, in probability,

$$\frac{1}{n}\sum_{i=1}^{n}\|g^{-1}_{\hat{\theta}_{n,Z}}(Z_i) - Z_i\|^p \to 0, \qquad n \to \infty.$$

We conclude that both terms in the bound (12) for $T^{1/p}_{\mathcal{M},n}$ converge to zero in probability. Hence the same is true for $T_{\mathcal{M},n}$.

(ii) By (i), it follows that $\lim_{n\to\infty} F_{\mathcal{M},n}(\varepsilon) = 1$ for every $\varepsilon > 0$. It is thus sufficient to show that, under the alternative hypothesis, there exists $\varepsilon > 0$ such that $\lim_{n\to\infty} \mathrm{P}^n[T_{\mathcal{M},n} > \varepsilon] = 1$.

Let $\mathrm{Q}_1 = (g^{-1}_\theta)_\# \mathrm{P}$, that is, Q_1 is the law of $Y = g^{-1}_\theta(X)$, where $g^{-1}_\theta \in G$ is the inverse transformation of g_θ and where X has law P. By assumption, $\mathrm{Q}_1 \neq \mathrm{Q}_0$, for otherwise $\mathrm{P} \in \mathcal{M}$. Also, $\mathrm{Q}_1 \in \mathcal{P}_p(\mathbb{R}^d)$, since $\mathrm{P} \in \mathcal{P}_p(\mathbb{R}^d)$ and since each g in G satisfies (5).

Put $Y_i = g^{-1}_\theta(X_i)$ for $i = 1, \ldots, n$, an i.i.d. sample from Q_1 and denote by $\widehat{\mathrm{P}}^Y_n = n^{-1}\sum_{i=1}^n \delta_{Y_i}$ its empirical distribution. The estimated residuals are $\hat{Z}_i = g^{-1}_{\hat{\theta}_n}(X_i) = g^{-1}_{\hat{\theta}_n} \circ g_\theta(Y_i)$. By the same argument as in (i), we have

$$W_p\big(\widehat{\mathrm{P}}^{\hat{Z}}_n, \mathrm{Q}_1\big) \le \left[\frac{1}{n}\sum_{i=1}^{n}\|g^{-1}_{\hat{\theta}_n} \circ g_\theta(Y_i) - Y_i\|^p\right]^{1/p} + W_p\big(\widehat{\mathrm{P}}^{Y}_n, \mathrm{Q}_1\big) \to 0,$$

as $n \to \infty$ in probability. By the continuous mapping theorem, it follows that $T_{\mathcal{M},n} \to W^p_p(\mathrm{Q}_1, \mathrm{Q}_0) > 0$ in probability as $n \to \infty$. But this implies that $1 - F_{\mathcal{M},n}(T_{\mathcal{M},n}) \to 0$ as $n \to \infty$ in probability: the null hypothesis, thus, is rejected with probability tending to one. □

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 $\mathbb{R}^d$. 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 distribution $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d)$, the goodness-of-fit problem consists of testing

$$\mathcal{H}^n_0 : \mathrm{P} \in \mathcal{M} \quad \text{against} \quad \mathcal{H}^n_1 : \mathrm{P} \notin \mathcal{M}. \tag{13}$$

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

$$T_{\mathcal{M},n} := W^p_p(\widehat{\mathrm{P}}_n, \mathrm{P}_{\hat{\theta}_n}) \tag{14}$$

where $\hat{\theta}_n = \theta_n(\mathbf{X}_n)$ is some consistent (under $\mathcal{H}_0^n$) estimator sequence of the true parameter θ. The distribution of $\mathbf{X}_n$ under $\mathcal{H}_0^n$ in (13) being P_θ^n for some $\theta \in \Theta$, let $F_{n,\theta}(t) = \mathrm{P}_\theta^n[T_{\mathcal{M},n} \leq t]$ for $t \in \mathbb{R}$ denote the null distribution function of the test statistic. As p-value and critical value, we would like to take

$$1 - F_{n,\theta}(T_{\mathcal{M},n}) \text{ and } c_{n,\theta}(\alpha) = \inf\{t \geq 0 : F_{n,\theta}(t) \geq 1 - \alpha\}, \tag{15}$$

respectively, for some $\alpha \in (0,1)$. This choice is infeasible, however, since the true parameter θ is unknown. Therefore, we propose to replace $c_{n,\theta}(\alpha)$ by the bootstrapped quantity $c_{n,\hat{\theta}_n}(\alpha)$, yielding the test

$$\phi_{\mathcal{M}}^n := \begin{cases} 1 & \text{if } 1 - F_{n,\hat{\theta}_n}(T_{\mathcal{M},n}) \leq \alpha \text{ or, equivalently, } T_{\mathcal{M},n} \geq c_{n,\hat{\theta}_n}(\alpha), \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

We reject $\mathcal{H}_0^n$ as soon as $T_{\mathcal{M},n}$ exceeds the critical value at the estimated parameter. The substitution of θ by $\hat{\theta}_n$ qualifies as a parametric bootstrap.

To compute the critical value $c_{n,\hat{\theta}_n}(\alpha)$ in practice, we rely, as before, on a Monte Carlo approximation: resample from $\mathrm{P}_{\hat{\theta}_n}$, compute the test statistic, and approximate $F_{n,\hat{\theta}_n}$ by the empirical distribution function of the resampled test statistics. By the Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990), the difference between $F_{n,\hat{\theta}_n}(t)$ and its Monte Carlo approximation can be controlled explicitly and uniformly in $t \geq 0$ and in the unknown parameter, and this in terms of the Monte Carlo sample size only. To speed up the calculations in case of a low-dimensional parameter space, we pre-compute an approximation of the critical value function $\theta \mapsto c_{n,\theta}(\alpha)$ in this way for θ in a finite grid $\Theta' \subset \Theta$ and then compute $c_{n,\hat{\theta}_n}(\alpha)$ by interpolation and/or smoothing.

Under the null hypothesis and if the true parameter is θ, the size of the test is now the random quantity

$$1 - F_{n,\theta}\big(c_{n,\hat{\theta}_n}(\alpha)\big).$$

In contrast to Sections 2 and 3, it is no longer guaranteed that this risk is bounded by α. The question remains open whether under the null hypothesis the actual size of the test indeed converges to α. To prove this conjecture 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 which depend on the model $\mathcal{M}$ under study (Beran, 1997; Capanu, 2019). As discussed in Section 1.2, such asymptotic distribution theory is still far beyond the horizon. Our numerical experiments in Section 5, however, support the conjecture that the parametric bootstrap produces a test with the right asymptotic size. For any $\theta \in \Theta$, any $\varepsilon > 0$, and a sufficiently regular parametric model $\mathcal{M}$ and estimator sequence $\hat{\theta}_n$, we conjecture that $\mathrm{P}_\theta^n[1 - F_{n,\theta}(c_{n,\hat{\theta}_n}(\alpha)) \leq \alpha + \varepsilon]$ converges to one as $n \to \infty$. In Appendix B, we provide a theoretical justification of the consistency of the parametric bootstrap in the univariate case, for which the asymptotic distribution theory of the empirical Wasserstein distance is well developed.

Nevertheless, against a fixed alternative, the consistency of the test (16) based on the parametric bootstrap can be established theoretically. The key is a law of large numbers for the empirical distribution in Wasserstein distance uniformly over classes of distributions that satisfy a uniform integrability condition, see Appendix A. For the parameter estimator $\hat{\theta}_n$, we 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

$$\forall \varepsilon > 0,\, \forall K \in \mathcal{K}(\Theta), \qquad \lim_{n\to\infty} \sup_{\theta \in K} \mathrm{P}_\theta^n \big[\rho(\hat{\theta}_n, \theta) > \varepsilon\big] = 0. \tag{17}$$

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

(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., (17) 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 $c_{n,\theta}(\alpha)$ tend to zero uniformly in θ, i.e.,

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

(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

$$T_{\mathcal{M},n}^{1/p} = W_p(\widehat{\mathrm{P}}_n, \mathrm{P}_{\hat{\theta}_n}) \le W_p(\widehat{\mathrm{P}}_n, \mathrm{P}_\theta) + W_p(\mathrm{P}_\theta, \mathrm{P}_{\hat{\theta}_n}) \tag{18}$$

for all $\theta \in \Theta$. It is then sufficient to show that, for any compact $K \subseteq \Theta$, each of the W_p-distances on the right-hand side of (18) 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^n_θ-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[2]

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

It follows that

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

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

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

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

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

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

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

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

[2] This is a slight generalization of the well-known property that a continuous function on a compact set is uniformly continuous. 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 $\rho(\theta, \theta_j) < \delta(\theta_j)/2$, hence $\rho(\theta', \theta_j) < \delta(\theta_j)$. By the triangle inequality, $W_p(\mathrm{P}_\theta, \mathrm{P}_{\theta'}) \le W_p(\mathrm{P}_{\theta_j}, \mathrm{P}_\theta) + W_p(\mathrm{P}_{\theta_j}, \mathrm{P}_{\theta'}) \le \varepsilon$.

We obtain that

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

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 $\mathrm{P} \in \mathcal{P}(\mathbb{R}^d) \setminus \mathcal{P}_p(\mathbb{R}^d)$. Since $\theta \mapsto W_p(\mathrm{P}_\theta, \delta_0)$ is continuous, $\sup_{\theta\in K} W_p(\mathrm{P}_\theta, \delta_0)$ is finite, with K as in (iii) and δ_0 the Dirac measure at $0 \in \mathbb{R}^d$. By an argument similar to the second part of the proof of Proposition 1, it follows that $\mathrm{P}^n[T_{\mathcal{M},n} > c_n,\, \hat{\theta}_n \in K] \to 1$ as $n \to \infty$. □

Remark 1 (Uniform consistency). Under a mild moment condition, the uniform consistency condition (b) in Proposition 3 is satisfied for *method of moment estimators*—call them *moment estimators*—of $\theta \in \Theta \subseteq \mathbb{R}^k$. 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 $m\colon \theta \mapsto \mathrm{E}_\theta[f(X)]$ 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 3(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 (Non locally compact parameter spaces). Proposition 3 allows for infinite-dimensional parameter spaces Θ. An example would be the space of all copulas of given dimension equipped with a metric that metrizes weak convergence, a space that is still compact thanks in view of Prohorov's theorem. If Θ is not locally compact, however, then condition (iii) is too severe and the compact set K should be replaced by its enlargement

$$K^\delta = \{\theta \in \Theta : \exists\, \theta' \in K, \rho(\theta, \theta') < \delta\} \text{ for some sufficiently small } \delta > 0$$

(van der Vaart and Wellner, 1996, Definition 1.3.7). The conditions on $\theta \mapsto \mathrm{P}_\theta$ and the estimator $\hat{\theta}_n$ then should be modified accordingly. We are grateful to an anonymous Referee for pointing this out.

4.1. Parametric models with group subfamilies

Consider again the testing problem (13). Sometimes the unknown parameter can be decomposed as $\theta = (\psi, \eta) \in \Psi \times H = \Theta$, where, for fixed ψ, the subfamily $\mathcal{M}_\psi = \{\mathrm{P}_{\psi,\eta} : \eta \in H\}$ is a group family as in Section 3, generated by a group $G = \{g_\eta : \eta \in H\}$ of transformations $g_\eta : \mathbb{R}^d \to \mathbb{R}^d$ independent of ψ. Think for instance of the case where ψ is a vector of shape parameters and η a vector of location–scale parameters, with G the group of Example 1.

Suppose further that a weakly consistent estimator $\hat{\theta}_n = (\hat{\psi}_n, \hat{\eta}_n)$ exists with the following two properties:

(i) $\hat{\psi}_n$ is invariant under G: writing $\hat{\psi}_n = \psi_n(X_1, \ldots, X_n)$, we have

$$\psi_n(x_1, \ldots, x_n) = \psi_n\big(g_\eta(x_1), \ldots, g_\eta(x_n)\big) \tag{19}$$

for all $\eta \in H$ and all possible samples $(x_i)_{i=1}^n$.

(ii) $\hat{\eta}_n$ is equivariant as in (9).

Then we propose a hybrid approach: compute the estimated residuals

$$\hat{Z}_{i,n} = g_{\hat{\eta}_n}^{-1}(X_i), \qquad i = 1, \ldots, n, \tag{20}$$

and form the test statistic

$$T_{\mathcal{M},n} = W_p^p\big(n^{-1}\textstyle\sum_{i=1}^n \delta_{\hat{Z}_{i,n}}, \mathrm{P}_{\hat{\psi}_n,\eta_e}\big), \tag{21}$$

with $\eta_e \in H$ the parameter yielding the identity transformation $g_{\eta_e}(x) = x$. For $\theta = (\psi, \eta)$, the distribution function $t \mapsto F_{n,\theta}(t) = \mathrm{P}_\theta^n[T_{\mathcal{M},n} \le t]$ of the test statistic depends on ψ but not on η. It can thus be computed as if $\eta = \eta_e$, that is, $F_{n,\psi,\eta} = F_{n,\psi,\eta_e}$. The proof of this invariance property relies on (i)–(ii) above.

The actual p-value of $T_{\mathcal{M},n}$ under $\mathcal{H}_n^0$ is

$$1 - F_{n,\psi,\eta_e}(T_{\mathcal{M},n})$$

while the critical value for a test of size $\alpha \in (0, 1)$ is now

$$c_{n,\psi}(\alpha) = \inf\{t \ge 0 : F_{n,\psi,\eta_e}(t) \ge 1 - \alpha\}.$$

Both are infeasible, however, since the null distribution of $T_{\mathcal{M},n}$ depends on the unknown ψ. We therefore compute p-values and critical values under $(\hat{\psi}_n, \eta_e)$. More precisely, we apply the parametric bootstrap. The test thus takes the form

$$\phi_{\mathcal{M}}^n = \begin{cases} 1 & \text{if } 1 - F_{n,\hat{\psi}_n,\eta_e}(T_{\mathcal{M},n}) \le \alpha \text{ or, equivalently, } T_{\mathcal{M},n} \ge c_{n,\hat{\psi}_n}(\alpha), \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

In practice, the distribution $F_{n,\hat{\psi}_n,\eta_e}$ and the associated critical values $c_{n,\hat{\psi}_n}(\alpha)$ are computed by Monte Carlo approximation, as described in the paragraph following (16). If the dimension of ψ is sufficiently low, we can pre-compute the critical values $c_{n,\psi}(\alpha)$ for ψ on a finite grid $\Psi' \subset \Psi$ and then reconstruct the critical value function by interpolation or smoothing. The reduction from θ to ψ thus brings a clear computational benefit.

We conjecture that, asymptotically, the test has the correct size under the null hypothesis. The obstacle for the proof is the same as before: required is the asymptotic distribution of the empirical Wasserstein distance, which is a very hard, long-standing open problem. In Section 5.3, we provide numerical support for the conjecture by an application to a five-dimensional distribution with separate location-scale parameters for each margin (ten parameters in total) and a single copula parameter. By exploiting invariance, the computation of the critical value is facilitated as the copula parameter remains as single argument.

Since the distribution of $T_{\mathcal{M},n}$ under $\theta = (\psi, \eta)$ is the same as under $\theta_e = (\psi, \eta_e)$, consistency of the hybrid test can be established by combining ideas from Propositions 2 and 3.

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. 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 a group model in Section 3, are compared in Section 5.2.1, along with an illustration involving a Student t distribution with known degrees of freedom in Section 5.2.2. Section 5.3 considers, in line with Section 4.1, the more general composite null hypothesis of a parametric family indexed by marginal location and scale parameters along with a copula parameter. Numerical results support the conjecture of the (asymptotic) validity of the parametric bootstrap for calculating 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 distances of order $p \in \{1, 2\}$. 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 1 000. As mentioned in Section 1.3 we relied on the R package transport (Schuhmacher et al., 2019) in case $p = 2$ and $d = 2$ and on our own C implementation of the algorithm proposed by Genevay et al. (2016) in all other cases. See also Appendix D for some details on the calculation of the critical values.

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 $\mathrm{P}_0 \in \mathcal{P}_p(\mathbb{R}^d)$ is fully specified.

Two other goodness-of-fit tests will be used as benchmarks: the test by Rippl, Munk and Sturm (2016), which is based on the 2-Wasserstein distance and is specific for multivariate Gaussian distributions, and the adaptation of the Kolmogorov–Smirnov test by Khmaladze (2016), which is based on empirical process theory. Both tests are described in some detail in Appendix C.

5.1.1. Bivariate Gaussian distribution

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) $\mathrm{P} = \mathcal{N}_2(0, \sigma^2 I_2)$ (rejection frequencies plotted against $\sigma^2 > 0$);
(c) $\mathrm{P} = \mathcal{N}_2\big(0, \big(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\big)\big)$ 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 $\rho \in (-1, 1)$);[3]
(f) P is the "boomerang-shaped" Gaussian mixture described in Appendix E (rejection frequencies plotted against the mixing weight $p \in (-1, 1)$).[4]

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 "boomerang mixture" of panel (f), the Rippl–Munk–Sturm test fails to capture the change in distribution. There is little difference between the Wasserstein tests with $p = 1$ and $p = 2$, except for the t-copula, where $p = 2$ yields a more powerful test than $p = 1$.

5.1.2. Mixture of bivariate Gaussian distributions

Figures 2 to 4 concern non-Gaussian null distributions P_0, so that the Rippl–Munk–Sturm test no longer applies. In Figure 2, the null distribution is the

[3] Note that P is not Gaussian, even for $\rho = 0$.

[4] The mixture is constructed so that the first and second moments of P remain close to those of P_0.

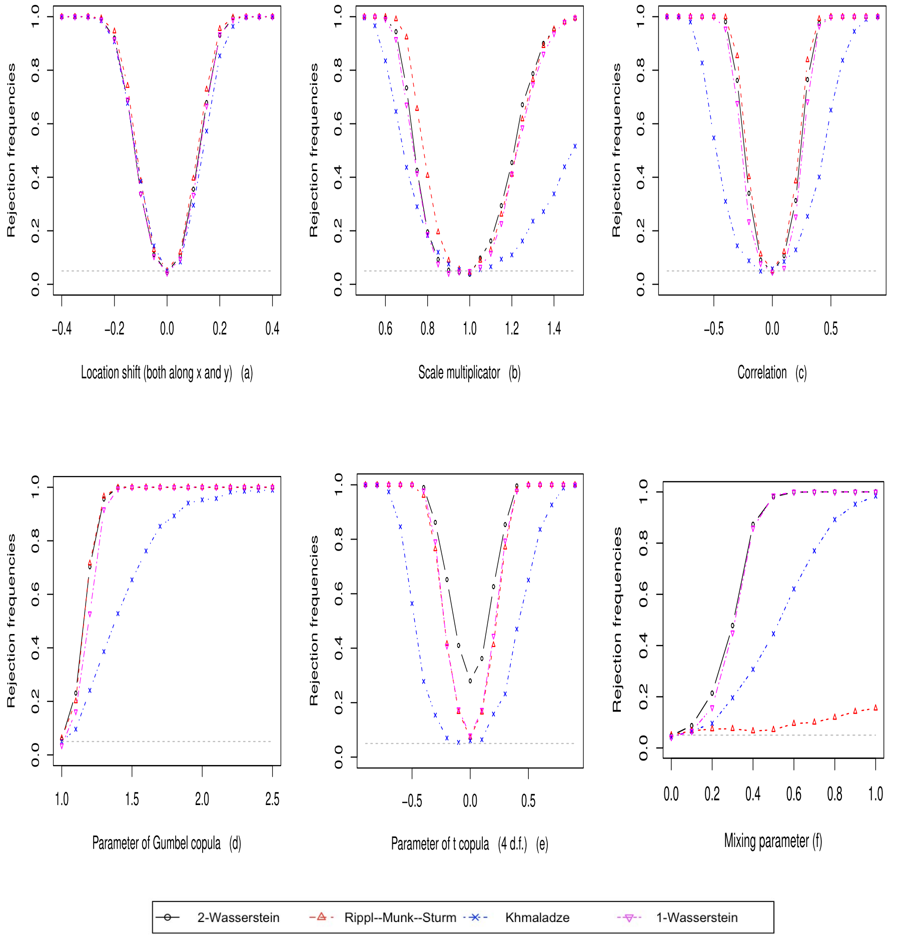

FIG 1. *Empirical powers of various GoF tests for the simple Gaussian null hypothesis* $\mathcal{H}_0^n$: $\mathrm{P} = \mathcal{N}_2(0, I_2)$. *Four tests are considered: the 2- and 1-Wasserstein distances (Section 2), the Rippl–Munk–Sturm test (Rippl, Munk and Sturm, 2016), and the Khmaladze Kolmogorov–Smirnov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

Gaussian mixture $\mathrm{P}_0 = 0.5\,\mathcal{N}_2(0, I_2) + 0.5\,\mathcal{N}_2\left(\binom{3}{0}, I_2\right)$. The alternatives in both panels are as follows:

(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 $\delta \in \mathbb{R}$);

(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 $\lambda \in [0, 1]$).

Both Wasserstein tests have higher power than the Khmaladze (2016) test.

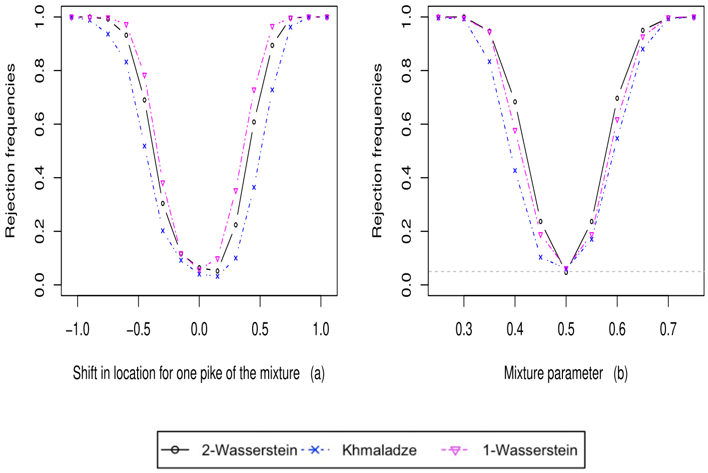

Fig 2. *Empirical powers of the Wasserstein and Khmaladze (2016) tests for the simple null hypothesis* $\mathcal{H}_0^n : \mathrm{P} = \mathrm{P}_0$ *with* P_0 *an equal-weights mixture of* $\mathcal{N}_2(0, I_2)$ *and* $\mathcal{N}_2\left(\binom{3}{0}, I_2\right)$. *The alternatives in (a)–(b) are described in Section 5.1.2.*

5.1.3. Gumbel copula and Gaussian marginals

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 value $\theta \neq 1.7$ of the copula parameter $\theta \in [1, \infty)$. Again, the Wasserstein tests at $p \in \{1, 2\}$ have quite comparable performance and yield higher empirical powers in most cases.

5.1.4. A five-dimensional Student t distribution

Let us turn now to a higher-dimensional case. In Figure 4, we test for the null hypothesis $\mathcal{H}_0^n : \mathrm{P} = \mathrm{P}_0$ with $\mathrm{P}_0 = \otimes_{i=1}^5 t_{25}$ a five-dimensional distribution with independence copula and Student t_{25} margins. The following alternatives are considered:

(a) A distribution with independence copula and Student t_ν margins. The rejection frequencies are plotted against ν.
(b) A distribution with independence copula and margins equal to the Student t_{25} distribution shifted by μ. The rejection frequencies are plotted against μ.
(c) A distribution $t_{25} \otimes t_{25} \otimes t_{25} \otimes t_{25,\delta}$, where $t_{\nu,\delta}$ is the bivariate Student t distribution with ν degrees of freedom and dependence parameter δ. Note that $\delta = 0$ does not correspond to the null hypothesis. The rejection frequencies are plotted against δ.

The Khmaladze Kolmogorov–Smirnov test is most sensitive to the change in location (b), although the two Wasserstein tests perform quite well too. For the other two alternatives, the Wasserstein tests have much higher power than the

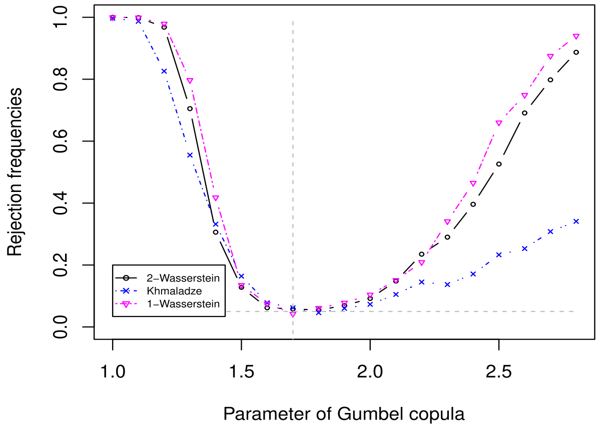

Fig 3. *Empirical powers of the Wasserstein and Khmaladze (2016) tests for the simple null hypothesis $\mathcal{H}_0^n$: $\mathrm{P} = \mathrm{P}_0$ with P_0 a bivariate distribution with standard Gaussian margins and Gumbel copula with parameter $\theta = 1.7$. Rejection frequencies are plotted against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

Khmaladze test. For the Wasserstein test, there is little difference between $p = 1$ and $p = 2$, except for case (c), in which the choice of $p = 2$ yields higher power.

5.2. Elliptical families as group models

Elliptical distributions arise as group models for the group of affine transformations, see Example 2 in Section 3. Two notable examples are the multivariate normal family and the multivariate Student t distribution with a fixed number of degrees of freedom. We assess the finite-sample performance of the Wasserstein test in (10) for $p \in \{1, 2\}$ with residuals computed by

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

with $\hat{\mu}_n$ the sample mean vector and $\hat{L}_n$ the lower Cholesky triangle of the empirical covariance matrix of $X_1, \ldots, X_n$.

5.2.1. Testing for multivariate normality

Testing for 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 (1983), Henze and Zirkler (1990), and Rizzo and Székely (2016). Royston's test is a generalisation of the well-known Shapiro–Wilks test. It only tests whether the margins are Gaussian and ignores the dependence structure. The Henze–Zirkler test statistic is an integrated weighted squared distance between the characteristic function of the multivariate standard normal distribution and the empirical characteristic function of the empirically standardized data. Interestingly, Ramdas, García Trillos and Cuturi (2017) showed that the

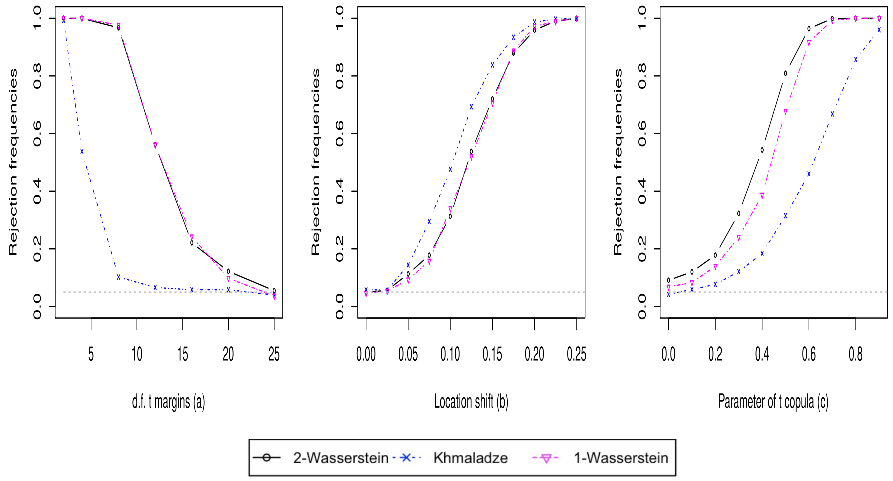

Fig 4. *Empirical powers of various GoF tests for the simple hypothesis $\mathcal{H}_0^n$ that* $\mathrm{P} \in \mathcal{P}_2(\mathbb{R}^5)$ *is the five-fold product of the Student* t_{25} *distribution with itself. Shown are rejection frequencies obtained via the Wasserstein test of order* $p \in \{1, 2\}$ *and the Khmaladze Kolmogorov–Smirnov one. Alternatives (a)–(c) are described in Section 5.1.4; in (c), no setting is corresponding to the null hypothesis.*

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 (2016) considered in Section 5.1 does not apply, since it only can handle fully specified Gaussian distributions, while here, the mean vector and covariance matrix are unknown.

In Figure 5, we consider dimensions $d = 2$ [top row, panels (a) and (b)] and $d = 5$ [bottom row, panels (c) and (d)]. Here are the alternative distributions P:

(a) A bivariate distribution with standard normal margins and a bivariate Gumbel copula with parameter $\psi \in [1, \infty)$. Rejection frequencies are plotted against $\psi \in [1, \infty)$.

(b) A bivariate distribution with independent margins, one of which is standard normal while the other one is Student t with $\nu > 0$ degrees of freedom. Rejection frequencies are plotted against $\nu > 0$.

(c) A five-variate distribution given by $\mathcal{N}_3(0, I_3) \otimes \mathcal{D}$, where $\mathcal{D}$ is a bivariate distribution with Gumbel copula indexed by a parameter $\psi \in [1, \infty)$ and with standard normal margins. Rejection frequencies are plotted against ψ.

(d) A five-variate distribution with independent margins, all of which are Student t_ν. Rejection frequencies are plotted against the common parameter ν.

The Wasserstein tests have the highest power against the copula alternatives in (a) and (c), while Royston's test has no power at all, as expected. For

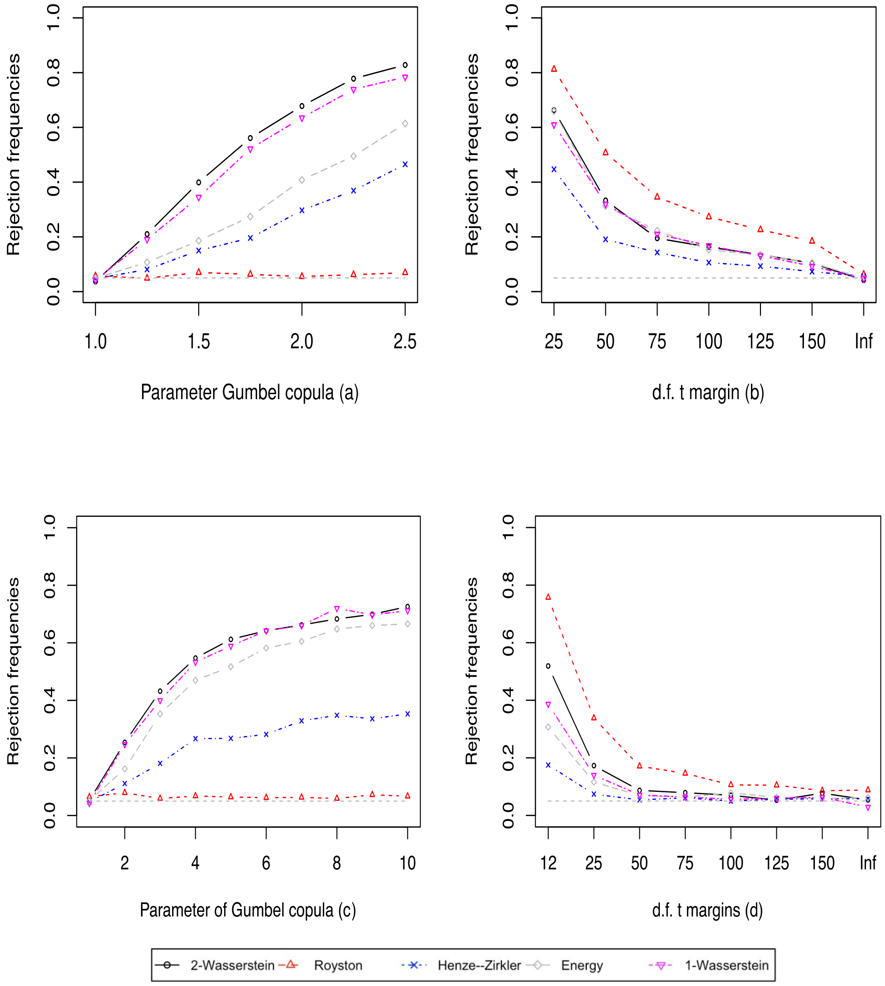

FIG 5. *Empirical power curves of various tests that* P *is d-variate Gaussian with unknown mean vector and covariance matrix. Top:* $d = 2$*. Bottom:* $d = 5$*. The Wasserstein test in Section 3 is compared to three other multivariate normality tests mentioned in Section 5.2.1, where the alternatives (a)–(d) are described as well.*

the Student t alternatives in panels (b) and (d), Royston's test comes out as most powerful, but the Wasserstein and energy tests (Rizzo and Székely, 2016) perform quite well too. It is also worth noticing that in panel (b), Royston's test had a type I error of 6.3% and in panel (d) this rose to 8.8%.

5.2.2. *Bivariate elliptical Student t with unknown location and scatter*

For fixed scalar $\nu > 0$, the d-variate elliptical Student t family with ν degrees of freedom and unknown location and scatter is generated by the affine trans-

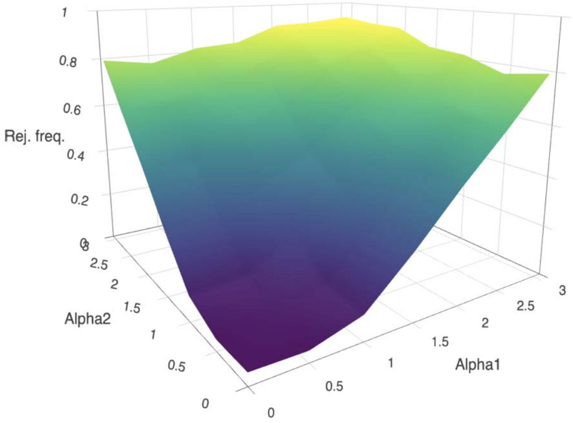

FIG 6. *Empirical power of the Wasserstein test in* (10) *for the hypothesis that* P *is bivariate Student t with $\nu = 12$ degrees and unknown mean vector and covariance matrix. The alternatives* P *are bivariate skew-t with skewness parameters α_1 and α_2.*

formation group in Example 2 applied to a spherical distribution whose radial density is that of the square root of a rescaled Fisher $F(d, \nu)$ variable. In dimension $d = 2$, we consider the hypothesis that P is of this form with $\nu = 12$.

Figure 6 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 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.

5.3. *General parametric families*

Turning to more general parametric families $\mathcal{M} = \{\mathrm{P}_\theta : \theta \in \Theta\}$, we investigate the finite-sample performance of the Wasserstein-based tests in Section 4. We consider models indexed by location, scale, and shape parameters. As in Section 4.1, the location-scale parameters are treated as stemming from the transformation group in Example 1 of Section 3, while for the shape parameters, we apply the parametric bootstrap. The test is thus the one we define in (22). We numerically investigate our conjecture that the test has the correct size asymptotically. In theory, the Khmaladze (2016) approach also applies, but its implementation is intricate and remains unsettled, especially when there are multiple parameters.

5.3.1. *Gaussian margins and AMH copula*

Let $\mathcal{M}$ consist of the bivariate distributions with Gaussian marginals and an Ali–Mikhail–Haq (AMH) copula, yielding a vector $\theta = (\psi, \mu_1, \sigma_1, \mu_2, \sigma_2)$ of five parameters: the AMH copula parameter $\psi \in \Theta = [-1, 1]$, the means $\mu_1, \mu_2 \in \mathbb{R}$, and the standard deviations $\sigma_1, \sigma_2 \in (0, \infty)$. The means and standard deviations are estimated by their empirical counterparts, so that the residuals (20) are $\hat{Z}_{i,n} = (\hat{Z}_{i,1,n}, \hat{Z}_{i,2,n})$ with

$$\hat{Z}_{i,j,n} = (X_{i,j} - \hat{\mu}_{j,n})/\hat{\sigma}_{j,n} \tag{23}$$

for $i = 1, \ldots, n$ and $j = 1, 2$. Following Genest, Ghoudi and Rivest (1995), the copula parameter ψ is estimated via a rank-based maximum pseudo-likelihood estimator. Obviously, the component-wise ranks of the data and those of the residuals in (23) coincide, so that $\hat{\psi}_n$, as required, depends on the data only through the residuals. The test statisti $T_{\mathcal{M},n}$ in (21) is the Wasserstein distance between the empirical distribution of the residuals and the bivariate distribution with standard Gaussian margins and AMH copula with the estimated parameter.

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 $\mathrm{P} \in \mathcal{M}$ with $\psi = 0.7$. For each sample, we calculated the test statistic $T_{\mathcal{M},n}$ in (21) and checked whether or not it exceeds the bootstrapped critical value $c_{\mathcal{M}}(\alpha, n, \hat{\psi}_n)$ for α equal to multiples of 5%. The critical values were computed as described below (22). The points in Figure 7(a) show the empirical type I errors as a function of α. The diagonal line fits the points well, supporting the conjecture that the parametric bootstrap is asymptotically valid.

Figure 7(b) similarly displays the rejection frequencies of the Wasserstein test for $p = 2$ under an alternative P whose copula belongs to the Frank family. If the Frank copula parameter is equal to zero, the Frank copula reduces to the independence copula, which is a member of the AMH family too.

5.3.2. *A multivariate Gumbel max-stable family*

Next, let $\mathcal{M} = \{\mathrm{Q}_\theta : \theta \in \Theta\}$ be the family of d-variate distributions with Gumbel margins with unknown location and scale parameters $(l_j, s_j) \in \mathbb{R} \times (0, \infty)$ for $j = 1, \ldots, d$ and a Gumbel copula with unknown shape parameter $\psi \in [1, \infty)$. Each $\mathrm{P}_\theta \in \mathcal{M}$ is thus a d-variate max-stable distribution, that is, a possible large-sample limit of the vector of affinely normalized component-wise maxima of an i.i.d. sample from a common distribution (Beirlant et al., 2004, Chapter 9).

There are $2d+1$ parameters in total. We treat the $2d$ location-scale parameters as indexing the transformation group in Example 1 of Section 3. The parameters are estimated in two stages:

1. The $2d$ location-scale parameters are estimated separately for each margin $j = 1, \ldots, d$ by maximum likelihood, producing $\hat{l}_j$ and $\hat{s}_j$.
2. The copula parameter is estimated by maximum likelihood on the basis of the estimated residuals $\hat{Z}_{i,n} = (\hat{Z}_{i,j,n})_{j=1}^d$ with

$$\hat{Z}_{i,j,n} = (X_{i,j} - \hat{l}_j)/\hat{s}_j, \qquad i = 1, \ldots, n.$$

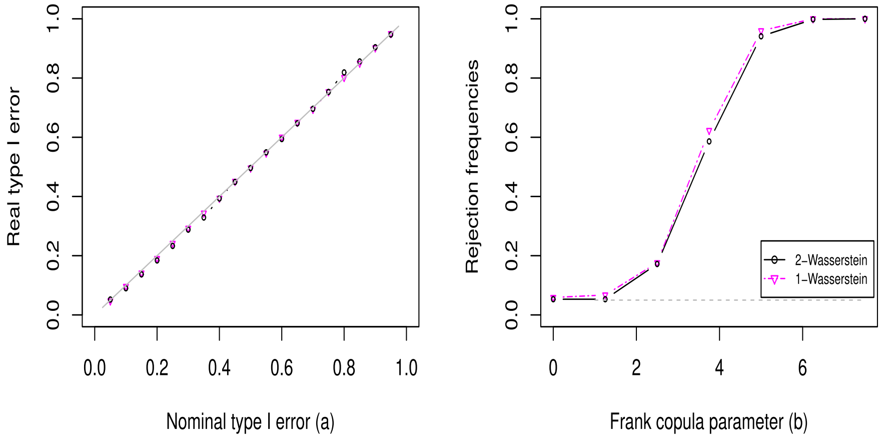

Fig 7. *Wasserstein tests for* $\mathcal{H}_0^n : \mathrm{P} \in \mathcal{M}$ *with* $\mathcal{M}$ *the parametric family of bivariate distributions with Gaussian margins and AMH copula (Section 5.3). Test statistics and critical values are computed from estimated residuals via a parametric bootstrap as in Section 4.1. Panel (a): real versus nominal type I errors* α *based on* $1\,000$ *samples of size* $n = 200$ *drawn from* $\mathrm{P} \in \mathcal{M}$ *with* $\psi = 0.7$. *Panel (b): powers against alternatives* P *with Gaussian marginals and Frank copula with varying parameter; if the latter is zero, the Frank copula is the independence one, which belongs to the AMH family too.*

This two-stage maximum pseudo-likelihood estimation procedure usually enjoys a high relative efficiency with respect to the full maximum likelihood estimator and is computationally much simpler (Joe, 2005). The location-scale estimators are equivariant under location-scale transformations. The residuals and the copula parameter estimator are thus invariant under such transformations.

We then proceed as in Section 4.1. The goodness-of-fit statistic $T_{\mathcal{M},n}$ in (21) measures the Wasserstein distance between the empirical distribution of the estimated residuals and the d-variate max-stable distribution with standard Gumbel margins and Gumbel copula with the estimated parameter. The goodness-of-fit test is carried out as in (22).

Figure 8 shows simulation results in dimensions $d = 2$ and $d = 5$ on top and bottom rows, respectively.

- On the left, the evaluation of the bootstrap accuracy is carried out as in Figure 7(a). Samples are generated from a distribution in the model with Gumbel copula parameter $\psi = 5/3$. The results support the conjecture that the Wasserstein-based tests with critical values calculated by the parametric bootstrap have the correct size, at least asymptotically.
- On the right, the power is calculated against alternative distributions whose margins are Generalized Extreme-Value (GEV) distributions with common shape parameter ξ indicated on the horizontal axis. Note that $\xi = 0$ corresponds to the null hypothesis. For $\xi > 0$, the distribution has finite moments up to order $p < 1/\xi$ only. This explains perhaps why the power of the Wasserstein test for $p = 2$ is less than for $p = 1$.

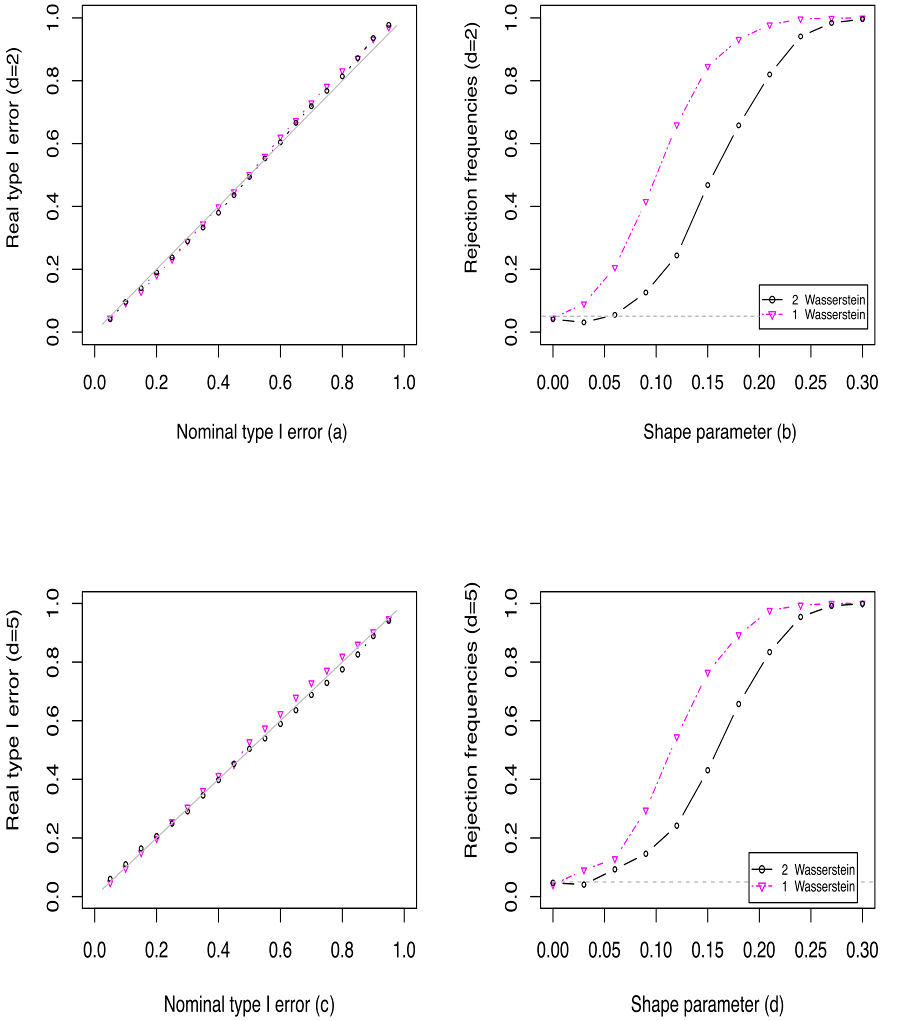

FIG 8. *Wasserstein test for* $\mathcal{H}_0^n : \mathrm{P} \in \mathcal{M}$ *with* $\mathcal{M}$ *the family of d-variate distributions with Gumbel margins with unknown location–scale parameters and Gumbel copula with unknown shape parameter* $\psi \in [1, \infty)$ *(Section 5.3.2). Test statistic and critical values based on estimated residuals and parametric bootstrap as in Section 4.1. Top:* $d = 2$. *Bottom:* $d = 5$. *Left: real versus nominal type I errors* α *based on* $1\,000$ *samples of size* $n = 200$ *drawn from* $\mathrm{P} \in \mathcal{M}$ *with Gumbel copula shape parameter* $\psi = 5/3$. *Right: power against alternatives* P *with Gumbel copula and GEV marginals (shape parameter on the horizontal axis).*

Acknowledgments

The authors gratefully acknowledge the remarks and comments made by the reviewers that greatly helped improve the paper.

References

Ambrosio, L., Stra, F. and Trevisan, D. (2018). A PDE approach to a 2-dimensional matching problem. *Probability Theory and Related Fields* 1–45. MR3916111

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. MR3468021

Bakshaev, A. and Rudzkis, R. (2015). Multivariate goodness-of-fit tests based on kernel density estimators. *Nonlinear Analysis. Modelling and Control* **20** 585–602. MR3403352

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J. L., De Waal, D. and Ferro, C. (2004). *Statistics of Extremes: Theory and Applications*. Wiley. MR2108013

Beran, R. (1997). Diagnosing bootstrap success. *Annals of the Institute of Statistical Mathematics* **49** 1–24. MR1450689

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics* **9** 1196–1217. MR0630103

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. MR0348906

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models. Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins University Press, Baltimore, MD. MR1245941

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. MR0629795

Capanu, M. (2019). A unified approach to proving parametric bootstrap consistency for some goodness-of-fit tests. *Statistics* **53** 58–80. MR3900080

Carlier, G., Chernozhukov, V., Galichon, A. et al. (2016). Vector quantile regression: an optimal transport approach. *The Annals of Statistics* **44** 1165–1192. MR3485957

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. MR3611491

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

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. MR2121458 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. MR3916938

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. MR1740113

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. MR1790430

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. MR3768763

Fan, Y. (1997). Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function. *Journal of Multivariate Analysis* **62** 36–63. MR1467872

Fang, K.-T., Kotz, S. and Ng, K.-W. (1990). *Symmetric multivariate and related distributions*. Chapman & Hall, London. MR1071174

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. MR3383341

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. MR1366280

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.

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

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*, third ed. The Johns Hopkins University Press, Baltimore and London. MR1417720

Hallin, M., del Barrio, E., Cuesta Albertos, J. and Matrán, C. (2020). Distribution and quantile functions, ranks, and signs in dimension d: a measure transportation approach. *The Annals of Statistics* (to appear).

Hartmann, V. and Schuhmacher, D. (2020). Semi-discrete optimal transport: a solution procedure for the unsquared Euclidean distance case. *Mathematical Methods of Operations Research* **92** 133–163. MR4152920

Henze, N. and Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics. Theory and Methods* **19** 3595–3617. MR1089501

Hofert, M., Kojadinovic, I., Maechler, M. and Yan, J. (2018). copula: Multivariate Dependence with Copulas R package version 0.999-19.1. MR3887600

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. MR1329874

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis* **94** 401–419. MR2167922

Khmaladze, E. V. (2016). Unitary transformations, empirical processes and distribution free testing. *Bernoulli* **22** 563–588. MR3449793

Kitagawa, J., Mérigot, Q. and Thibert, B. (2017). Convergence of a Newton algorithm for semi-discrete optimal transport. *arXiv preprint 1603.05579v2*. MR3985609

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.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer Science+Business Media, New York. MR1639875

Lévy, B. (2015). A numerical algorithm for L2 semi-discrete optimal transport in 3D. *ESAIM: Mathematical Modelling and Numerical Analysis* **49** 1693–1715. MR3423272

Massart, P. (1990). The tight constant in the Dvoretsky–Kiefer–Wolfowitz inequality. *The Annals of Probability* **18** 1269–1283. MR1062069

McAssey, M. P. (2013). An empirical goodness-of-fit test for multivariate distributions. *Journal of Applied Statistics* **40** 1120–1131. MR3286299

Mena, G. and Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems* 4541–4551.

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. MR1625620

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and its Applications* **6** 405–431. MR3939527

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. MR3608466

Rippl, T., Munk, A. and Sturm, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis* **151** 90–109. MR3545279

Rizzo, M. L. and Székely, G. J. (2016). Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics* **8** 27–38. MR3457239

Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro–Wilk W. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **32** 121–133.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians. Progress in Nonlinear Differential Equations and their Applications* **87**. Birkhäuser/Springer, Cham. MR3409718

Schmidt, M., Le Roux, N. and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming* **162**. MR3612933

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.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscow* **2** 3-14. MR0002062

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. MR3744719

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. MR4019874

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. MR1652247

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer-Verlag, New York. MR1385671

Villani, C. (2009). *Optimal Transport: Old and New*. Springer-Verlag, Berlin. MR2459454

von Mises, R. E. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit*. Julius Springer, Berlin. MR0041364

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli* **25** 2620–2648. MR4003560

Appendix A: Uniform convergence of the empirical Wasserstein distance

We establish here 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. The result is thus a law of large numbers for the empirical distribution in Wasserstein distance uniformly in the underlying distribution akin to Chung's uniform law of large numbers (van der Vaart and Wellner, 1996, Proposition A.5.1).

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*

$$\lim_{r\to\infty} \sup_{\mathrm{P}\in\mathcal{M}} \int_{\|x\|>r} \|x\|^p \,\mathrm{dP}(x) = 0. \tag{24}$$

Then we have

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

The condition on $\mathcal{M}$ is equivalent to assuming 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. In Fournier and Guillin (2015, Theorem 1), for instance, the expectation is bounded by an explicit expression involving n, p, d, and the moment of order q of P for some $q > p$. Bounds on such moments for all $\mathrm{P} \in \mathcal{M}$ then imply a uniform rate of convergence in $\mathrm{P} \in \mathcal{M}$. In contrast, we do not impose the existence of moments of order q higher than p, but only the uniform integrability of the p-th order moment.

The challenge in the proof of Theorem 1 is to obtain a sufficiently sharp and explicit bound on $\mathbb{E}_{\mathrm{P}}\{W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P})\}$. Such a bound is known for absolutely continuous measures in terms of a weighted total variation distance (Villani, 2009, Theorem 6.15). To apply that bound, an additional smoothing step is needed, and the whole procedure needs to work uniformly in the underlying probability measure, relying only on the uniform integrability condition (24).

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 $\{x \in \mathbb{R}^d : \|x\| \leq \sigma\}$ 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 $\mathrm{Q} \in \mathcal{P}_p(\mathbb{R}^d)$,

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

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

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

By the triangle inequality, it follows that

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

Taking expectations and using the elementary inequality

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

we obtain

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

If we can show that

$$\forall \sigma > 0, \qquad \lim_{n\to\infty} \sup_{\mathrm{P}\in\mathcal{M}} \mathbb{E}\left\{W_p^p(\widehat{\mathrm{P}}_n * \mathrm{U}_\sigma, \mathrm{P} * \mathrm{U}_\sigma)\right\} = 0, \tag{25}$$

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 the lim sup is actually a limit and is equal to zero, as required.

Let us proceed to show (25). 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

$$\mathbb{E}_{\mathrm{P}}\left\{W_p^p(\widehat{\mathrm{P}}_n * \mathrm{U}_\sigma, \mathrm{P} * \mathrm{U}_\sigma)\right\} \le 2^{p-1}\int_{\mathbb{R}^d} \|x\|^p g_n(x;\mathrm{P})\,\mathrm{d}x \tag{26}$$

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 (26) 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

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

But then, in view of (26), 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 obtain

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

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

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

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

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

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

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

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 (25), as required. The proof is complete. □

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

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

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

$$\mathrm{P}^n\big[W_p(\widehat{\mathrm{P}}_n, \mathrm{P}) > \varepsilon\big] \le \varepsilon^{-p}\,\mathbb{E}_\mathrm{P}\left\{W_p^p(\widehat{\mathrm{P}}_n, \mathrm{P})\right\}.$$

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

For a single $\mathrm{P} \in \mathcal{P}_p(\mathbb{R}^d)$, Lemma 8.4 in Bickel and Freedman (1981) says that $W_p(\widehat{\mathrm{P}}_n, \mathrm{P}) \to 0$ almost surely as $n \to \infty$. Whether Corollary 1 can be strengthened to almost sure convergence uniformly in P, i.e., whether

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

remains an open problem.

Appendix B: Consistency of the parametric bootstrap in the univariate case

In Section 4, we left open the conjecture of the consistency of the parametric bootstrap procedure for the Wasserstein GoF test in general parametric families. Proving it requires asymptotic distribution theory for the empirical Wasserstein distance, which, in general, is a difficult and long-standing open problem (Section 1.2). In the univariate case, however, the large-sample distribution of the empirical Wasserstein distance is known, enabling a theoretical analysis.

Consider the same notation as in Section 4 and assume $d = 1$. In the univariate case, the Wasserstein distance can be expressed as the L_p distance between quantile functions, see Panaretos and Zemel (2019, Section 1.2.3) and the references therein. Let $\hat{F}_n^{-1}$ denote the empirical quantile function of the sample $X_1, \ldots, X_n$ and let F_θ^{-1} denote the quantile function of P_θ, defined as the (generalized) inverse of the cumulative distribution function F_θ of P_θ. The normalized Wasserstein GoF test statistic in Section 4 takes the form

$$R_n := n^{p/2} T_{\mathcal{M},n} = n^{p/2}\, W_p^p\big(\widehat{\mathrm{P}}_n, \mathrm{P}_{\hat{\theta}_n}\big) = \int_0^1 |\zeta_n(u)|^p \,\mathrm{d}u, \tag{27}$$

where ζ_n is the empirical quantile process at the estimated parameter:

$$\zeta_n(u) = \sqrt{n}\big\{\hat{F}_n^{-1}(u) - F_{\hat{\theta}_n}^{-1}(u)\big\}, \qquad u \in (0,1). \tag{28}$$

We follow the notation and logic of Beran (1997). Let $H_n(\theta)$ denote the sampling distribution of R_n under P_θ^n. We would like to use $H_n(\theta)$ to draw inference based on the observed value of R_n, for instance by comparing the latter to a critical value computed under $H_n(\theta)$. Since we do not know θ, we do not know $H_n(\theta)$ either. The parametric bootstrap consists of estimating the unknown sampling distribution of R_n by the random probability measure $H_n(\hat{\theta}_n)$, the sampling distribution of the statistic R_n under the estimated parameter. In practice, we calculate relevant quantities related to $H_n(\hat{\theta}_n)$ such as critical values or p-values by Monte Carlo simulation, drawing many bootstrap samples $X_1^*, \ldots, X_n^*$ from $\mathrm{P}_{\hat{\theta}_n}$ and calculating the test statistic R_n^* from those.

The question is whether inference drawn from R_n based on $H_n(\hat{\theta}_n)$ rather than on $H_n(\theta)$ is still valid, at least asymptotically. Let $H(\theta)$ denote the limit distribution of R_n under P_θ^n as $n \to \infty$, assuming it exists. If $H_n(\theta_n)$ converges weakly to the same limit $H(\theta)$ for any sequence $\theta_n \in \Theta$ such that

$$\sqrt{n}(\theta_n - \theta) = \mathrm{O}(1), \qquad n \to \infty, \tag{29}$$

then it follows that in P_θ^n-probability, the estimated sampling distribution $H_n(\hat{\theta}_n)$ converges weakly to $H(\theta)$ for all estimator sequences $\hat{\theta}_n$ such that $\sqrt{n}(\hat{\theta}_n - \theta)$ is $\mathrm{O}_{\mathrm{P}_\theta^n}(1)$ as $n \to \infty$. [The estimator $\hat{\theta}_n$ in the definition of R_n need not even be the same as the one under which we calculate the sampling distribution $H_n(\hat{\theta}_n)$, but for simplicity, we assume it is.]

The normalized test statistic R_n in Eq. (27) is a functional of the empirical quantile process ζ_n in Eq. (28). In Proposition 4 below, we will show that the weak limits as $n \to \infty$ of the finite-dimensional distributions of ζ_n under $\mathrm{P}_{\theta_n}^n$ for $\theta_n \to \theta \in \Theta$ do not depend on the particular sequence θ_n as long as Eq. (29) and certain assumptions on the model and the estimator sequence are fulfilled.

Assumption 1. $\mathcal{M} = \{\mathrm{P}_\theta : \theta \in \Theta\} \subset \mathcal{P}(\mathbb{R})$ is a parametric model and $\hat{\theta}_n$ is an estimator sequence satisfying the following properties:

(A1) Θ is an open subset of $\mathbb{R}^k$;
(A2) P_θ has density f_θ with respect to one-dimensional Lebesgue measure for every $\theta \in \Theta$;
(A3) $\mathcal{M}$ is differentiable in quadratic mean at any $\theta \in \Theta$ with score function $\dot{\ell}_\theta = \nabla_\theta \log f_\theta(x) : \mathbb{R} \to \mathbb{R}^k$ and non-singular $k \times k$ Fisher information matrix $\mathcal{I}_\theta = \mathbb{E}_\theta[\dot{\ell}_\theta(X)\dot{\ell}_\theta(X)^\top]$;
(A4) the estimator sequence $\hat{\theta}_n$ is regular and asymptotically linear at $\theta \in \Theta$ with influence function ψ_θ.

Asymptotic linearity in Assumption (A4) means that

$$n^{1/2}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_\theta(X_i) + \mathrm{o}_{\mathrm{P}_\theta^n}(1), \qquad n \to \infty, \tag{30}$$

with P_θ-square integrable influence function $\psi_\theta : \mathbb{R} \to \mathbb{R}^k$ satisfying

$$\mathbb{E}_\theta[\psi_\theta(X)] = 0.$$

Regularity in (A4) means that the influence function ψ_θ satisfies

$$\psi_\theta - \tilde{\ell}_\theta \perp_\theta \dot{\ell}_\theta$$

where $\tilde{\ell}_\theta = \mathcal{I}_\theta^{-1}\dot{\ell}_\theta$ is the efficient influence function for estimating θ and $\perp_\theta$ means orthogonality in $L_2(P_\theta)$; an equivalent criterion is that

$$\mathbb{E}_\theta[\psi_\theta(X)\, \dot{\ell}_\theta(X)^\top] = I_k, \tag{31}$$

the $k \times k$ identity matrix. We refer to van der Vaart (1998, Chapters 7–8) and Bickel et al. (1993, Chapter 2) for more background on these assumptions, which are standard.

Proposition 4. *Let $\mathcal{M} = \{\mathrm{P}_\theta : \theta \in \Theta\} \subset \mathcal{P}(\mathbb{R})$ be a parametric model and $\hat{\theta}_n$ an estimator sequence such that Assumption 1 is satisfied. Suppose that f_θ is continuous and strictly positive on the interior of the support of P_θ and that, for every fixed $u \in (0,1)$, the quantile function $F_\theta^{-1}(u)$ is continuously differentiable as a function of θ. Then, for all $\theta, \theta_n \in \Theta$ satisfying Eq. (29) and for every vector $(u_1, \ldots, u_m) \in (0,1)^m$, the quantile process ζ_n in Eq. (28) satisfies*

$$\big(\zeta_n(u_j)\big)_{j=1}^m \overset{\theta_n}{\rightsquigarrow} \mathcal{N}_m(0, \Gamma_\theta), \qquad n \to \infty, \tag{32}$$

where Γ_θ is a certain $m \times m$ covariance matrix given in Eq. (37) below and where the arrow means weak convergence of the law under $\mathrm{P}_{\theta_n}^n$ of the random vector on the left-hand side to the law of the random vector on the right-hand side.

The proof of Proposition 4 is given below. Passing from the asymptotics of the finite-dimensional distributions of ζ_n to those of R_n in Eq. (27) requires two things: asymptotic tightness of ζ_n as well as regularity conditions on P_θ controlling the tails of the quantile functions to be integrated. The former is a classical topic in empirical process theory, see for instance Chapters 19 and 21 in van der Vaart (1998). Regarding the latter, see del Barrio, Giné and Utzet (2005) for the case $p = 2$ and Bobkov and Ledoux (2019) for general $p \geq 1$.

The important thing in Proposition 4 is that the limit (32) does not depend on the sequence $(\theta_n)_n$. If integration and passage to the limit in Eq. (27) is permitted, the asymptotic equivariance in law in Eq. (32) for ζ_n continues to hold for the normalized test statistic R_n. The consistency of the parametric bootstrap for the Wasserstein GoF test then follows as explained in the lines below Eq. (29).

Proof of Proposition 4. By a subsequence argument, we can and will assume that $h_n = \sqrt{n}(\theta_n - \theta) \to h \in \mathbb{R}^k$ as $n \to \infty$. The proof proceeds by Le Cam's third lemma following the strategy in van der Vaart (1998, Section 7.5).

By Theorem 7.2 in the same reference, Assumption 1 implies that the log-likelihood ratio of $\mathrm{P}_{\theta_n}^n$ with respect to P_θ^n admits the expansion

$$\log \prod_{i=1}^n \frac{f_{\theta_n}}{f_\theta}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h^\top \dot{\ell}_\theta(X_i) - \frac{1}{2} h^\top \mathcal{I}_\theta h + \mathrm{o}_{\mathrm{P}_\theta^n}(1), \qquad n \to \infty, \tag{33}$$

with $n^{-1/2} \sum_{i=1}^n \dot{\ell}_\theta(X_i)$ asymptotically $\mathcal{N}_k(0, \mathcal{I}_\theta)$ under P_θ^n as $n \to \infty$. This means that the sequence of statistical experiments $\{\mathrm{P}_\theta^n : \theta \in \Theta\}$ is locally asymptotically normal.

To show Eq. (32), we need to find the joint limit distribution under θ of the finite-dimensional distributions of ζ_n together with the log-likelihood ratio in Eq. (33). If the joint limit is Gaussian and if the asymptotic cross-covariance with the term corresponding to the log-likelihood ratio is zero, then by Le Cam's third lemma (van der Vaart, 1998, Example 6.7), the asymptotic distribution of ζ_n under θ_n is the same as under θ, as required.

Fix $0 < u < 1$. We derive an asymptotically linear expansion of $\zeta_n(u)$. By the delta method (van der Vaart, 1998, Theorem 3.1) and the asymptotic linearity of $\hat{\theta}_n$ in Eq. (30), we have

$$\begin{aligned}\sqrt{n}\{F_{\hat{\theta}_n}^{-1}(u) - F_\theta^{-1}(u)\} &= \sqrt{n}(\hat{\theta}_n - \theta)^\top \nabla_\theta F_\theta^{-1}(u) + \mathrm{o}_{\mathrm{P}_\theta^n}(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_\theta(X_i)^\top \nabla_\theta F_\theta^{-1}(u) + \mathrm{o}_{\mathrm{P}_\theta^n}(1), \qquad n \to \infty.\end{aligned}$$

Further, by the functional delta method, the empirical quantile function satisfies

$$\sqrt{n}\{\hat{F}_n^{-1}(u) - F_\theta^{-1}(u)\} = -\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\mathbb{1}\{X_i \le F_\theta^{-1}(u)\} - u}{f_\theta(F_\theta^{-1}(u))} + \mathrm{o}_{\mathrm{P}_\theta^n}(1), \quad n \to \infty,$$

see van der Vaart (1998, Corollary 21.5). Subtract both expansions to get

$$\zeta_n(u) = -\frac{1}{\sqrt{n}} \sum_{i=1}^n g_\theta(X_i, u) + \mathrm{o}_{\mathrm{P}_\theta^n}(1), \qquad n \to \infty, \tag{34}$$

where

$$g_\theta(x, u) = \frac{\mathbb{1}\{x \le F_\theta^{-1}(u)\} - u}{f_\theta\big(F_\theta^{-1}(u)\big)} + \psi_\theta(x)^\top \nabla_\theta F_\theta^{-1}(u).$$

Differentiating the identity $u = \int_{-\infty}^{F_\theta^{-1}(u)} f_\theta(x)\,\mathrm{d}x$ with respect to θ using Leibniz' integral rule yields, after some calculations, the identity

$$\nabla_\theta F_\theta^{-1}(u) = -\frac{1}{f_\theta\big(F_\theta^{-1}(u)\big)} \mathbb{E}_\theta[\dot{\ell}_\theta(X)\, \mathbb{1}\{X \le F_\theta^{-1}(u)\}]. \tag{35}$$

The regularity property (31) of the influence function ψ_θ, the centering property $\mathbb{E}_\theta[\dot{\ell}_\theta(X)] = 0$ of the score function in Assumption 1 and the identity (35) for the gradient $\nabla_\theta F_\theta^{-1}(u)$ combine to imply that

$$\mathbb{E}_\theta\left[\dot{\ell}_\theta(X)\, g_\theta(X, u)\right] = 0. \tag{36}$$

Let $u_j \in (0, 1)$ for $j = 1, \ldots, m$. By the multivariate central limit theorem, the expansions (33) and (34) combine with Slutsky's lemma to yield the convergence in distribution of the sequence of $(m + 1)$-dimensional random vectors

$$\left(\zeta_n(u_1), \ldots, \zeta_n(u_m), \log \prod_{i=1}^n \frac{f_{\theta_n}}{f_\theta}(X_i)\right)$$

to a certain $(m + 1)$-variate normal variable. Each of the first m components of the limiting normal random vector is centred and, by Eq. (36), uncorrelated with the $(m+1)$th one. By Le Cam's third lemma (van der Vaart, 1998, Example 6.7), the limit distribution of $(\zeta_n(u_j))_{j=1}^m$ under θ_n is then the same as under θ: an m-variate centred normal distribution with covariance matrix Γ_θ having elements

$$\Gamma_\theta(j_1, j_2) = \mathbb{E}_\theta\left[g_\theta(X, u_{j_1})\, g_\theta(X, u_{j_2})\right], \qquad j_1, j_2 \in \{1, \ldots, m\}. \tag{37}$$

□

The above argument for the consistency of the parametric bootstrap in case $d = 1$ was made possible by the representation in Eqs (27)–(28) of the normalized test statistic R_n in terms of the empirical quantile process ζ_n. This representation made it possible to find the invariant limit distribution of R_n under contiguous alternatives $\theta_n = \theta + \mathrm{O}(1/\sqrt{n})$. In case $d \geq 2$, however, no sufficiently explicit representations of the empirical Wasserstein distance are hitherto known to enable a similar analysis.

Appendix C: Some other GoF tests

We provide details about the tests appearing in the comparisons in Section 5.1.

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

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

The Rippl–Munk–Sturm test statistic is $W_2^2\bigl(\mathcal{N}_d(\overline{X}_n, S_{n,X}),\, \mathcal{N}_d(\mu_0, \Sigma_0)\bigr)$, 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 quantiles 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 $f_1, \ldots, f_d$, and copula density c. Define

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

with $F_1, \ldots, F_d$ 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

$$\tilde{v}_{F,n}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \bigl\{l(X_i)\, \mathbb{1}(X_i \leq x) - \kappa(x)\bigr\} - \frac{G(x) - \kappa(x)}{1 - \kappa} \frac{1}{\sqrt{n}} \sum_{i=1}^n \bigl\{l(X_i) - \kappa\bigr\}$$

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

$$v_{G,n}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \bigl\{\mathbb{1}(Y_i \leq x) - G(x)\bigr\}$$

of 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 coordinatewise 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 $\sup_{x \in \mathbb{R}^d} |\tilde{v}_{F,n}(x)|$, which (with F the cdf of P_0) we consider in Section 5.1 for comparison with the 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 finite-sample size, we calculate critical values by Monte Carlo approximation rather than by relying on asymptotic theory.

Appendix D: Algorithms for the computation of critical values

Our test statistics involve the Wasserstein distance between an empirical measure and a continuous one. Calculating such a distance requires solving a semi-discrete optimal transport problem (Section 1.3).

In dimension $d = 2$ and for the Wasserstein distance of order $p = 2$, 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 are providing approximately 10^5 cells.

For the simulations involving the Wasserstein distance of order $p = 1$ or in dimension d larger than two, we resort to our own implementation of the stochastic average gradient algorithm as employed in Genevay et al. (2016). The number of random points chosen for the reference measure was 2×10^5 which corresponds to a thousand times the sample sizes considered in the various simulation settings. The C parameter appearing in their algorithm was set to 1.

The test statistics in (2) and (8) involve a fixed continuous measure P_0 or Q_0, respectively, but those in (14) and (21) concern a continuous measure with estimated parameter $\hat{\theta}_n$ or $\hat{\psi}_n$. Moreover, to calculate critical values with the parametric bootstrap, even a single execution of the test requires a large number of evaluations of the test statistic at random parameter values. To speed up the calculations, we perform the following two steps prior to observing the data:

1. We compute the discretizations of the target density mentioned above for each value of the unknown parameter in a large but finite subset of the parameter space. We then force the Monte Carlo replications of the parameter estimates to take values in that subset. In this way, we do not need to recompute the discretization of the target density each time.
2. We compute the critical values at a finite subset of the parameter space, by drawing random samples of the test statistic for each value of the parameter in that finite subset and applying the reduction of Step 1. Then we learn the critical value as a function of the (continuous) parameter by smoothing. See Figure 9 for an illustration.

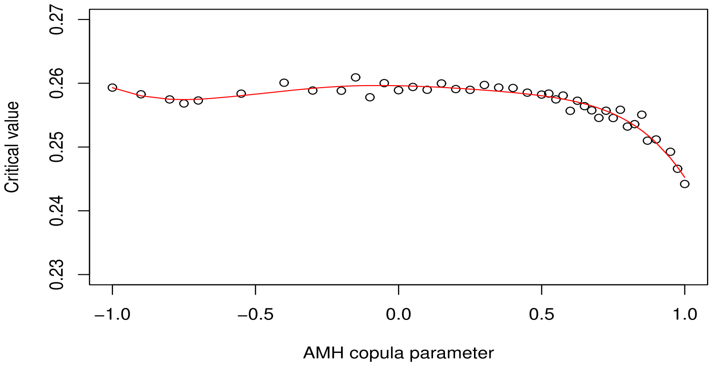

Fig 9. *Illustration of Step 2 in Appendix D for learning the critical value function of the* 2*-Wasserstein GoF test for the bivariate five-parameter Gaussian–AMH model in Section 5.3.1 using the location–scale reduction in Section 4.1. The function* $\psi \mapsto c_{\mathcal{M}}(\alpha, n, \psi)$ *(in red) is constructed by smoothing Monte Carlo estimates (circles) of* $c_{\mathcal{M}}(\alpha, n, \psi)$ *for* $\psi \in \Psi' \subseteq \Psi = [-1, 1]$*, 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.*

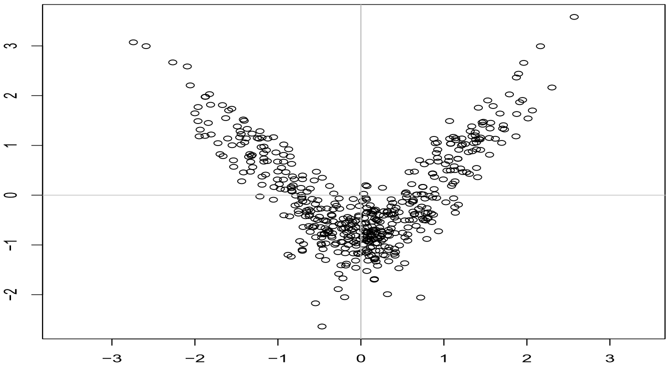

Fig 10. *Scatterplot of a sample of size* 500 *from the "boomerang-shaped" mixture* (38).

Appendix E: A boomerang-shaped distribution

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

$$(1-2p)\,\mathcal{N}_2\left(\begin{pmatrix} 0 \\ -0.7 \end{pmatrix}, \begin{pmatrix} 0.35^2 & 0 \\ 0 & 0.35^2 \end{pmatrix}\right) + p\,\mathcal{N}_2\left(\begin{pmatrix} -0.9 \\ 0.3 \end{pmatrix}, \begin{pmatrix} 0.358 & -0.55 \\ -0.55 & 1.02 \end{pmatrix}\right) + p\,\mathcal{N}_2\left(\begin{pmatrix} 0.9 \\ 0.3 \end{pmatrix}, \begin{pmatrix} 0.358 & 0.55 \\ 0.55 & 1.02 \end{pmatrix}\right). \tag{38}$$

of three Gaussian components. Figure 10 shows a scatterplot for $p = 0.35$ of a random sample of size $n = 500$ from this distribution.