Density model checks via the lack-of-fitness*

Valentin Patilea[1*] and François Portier[1]

[1*]CREST - UMR CNRS 9194, Univ Rennes, Ensai, Rennes, F-35000, France.

*Corresponding author(s). E-mail(s): valentin.patilea@ensai.fr;
Contributing authors: francois.portier@ensai.fr;

Abstract

Parametric multivariate density estimators, such as the maximum likelihood, can be generalized by mixing them with a kernel estimator. The mixture weights can be chosen to optimize a measure of the goodness-of-fit. The optimal weight of the kernel estimator, which we call the lack-of-fitness coefficient, then provides a simple check of the parametric model. The test statistic is defined as the appropriately normalized lack-of-fitness coefficient. When the parametric density model is correct, the statistic converges in distribution to the positive part of a standard Gaussian variable, regardless of the dimension of the observations. In addition, the test has good power against alternative hypotheses approaching the density model.

Keywords: Central Limit Theorem, Concavity, Leave-one-out density estimator, Pivotalness

MSC Classification: 62G07; 62G10

1 Introduction

Let $X_1, \ldots, X_n$ be independent and identically distributed $\mathbb{R}^d$-valued random vectors with density f_0. Our goal is to test a composite null hypothesis for this density, that is f_0 belongs to a specified parametric model of densities, such as the multivariate Gaussian model. There are a number of tests available for this problem, derived from three main types of approaches to testing the goodness-of-fit testing for density models: the smooth tests, the tests based on the distance between a nonparametric estimator and a model-based density estimator, and the tests based on the so-called weighted L^2-statistics, including the energy statistics.

One way to measure the discrepancy between the true density of the data and a target density in a given parametric model, is to consider their log-likelihood ratio. A class of tests, called *smooth tests*, is then obtained by considering a suitable orthonormal expansion of the log-likelihood ratio, and testing the nullity of the coefficients in the expansion. Neyman [1] introduces this idea for testing uniformity, and uses Rao's score statistic, with a given number of coefficients in the log-likelihood ratio expansion. Several data-driven versions of Neyman's test, where the number of coefficients is chosen automatically, have been proposed. See for example Ledwina [2] and Fan [3]. Claeskens and Hjort [4] consider a related idea, but they propose to test the nullity of the coefficients in the expansion using a likelihood ratio test (LRT) instead of the score test. While there is an asymptotic equivalence between Rao's score and the LRT statistic when the number of coefficients to be tested is fixed, Claeskens and Hjort [4] provide evidence that their approach performs better. They also propose a data-driven selection of the number of coefficients to be tested using AIC and BIC.

Another natural way to check the adequacy of a density model is to consider a norm between a non-parametric estimator, typically obtained by kernel smoothing, and the density estimator obtained within the model. Cao and Lugosi [5] investigated this approach with the L^1-norm. Their approach allows for general density models, not necessarily parametric. The tests based on the L^2-distance, first considered by Bickel and Rosenblatt [6], are somewhat more convenient since, in this case, the difficult part of the problem is reduced to finding an estimate of the squared L^2-norm of the true density. Fromont and Laurent [7] construct a test

*This version: June 19, 2024

based on this idea when the model, which is either a given univariate density or the location/scale family obtained from it. Their test is related to the data-driven smooth tests, but Fromont and Laurent [7] propose to construct adaptive estimates of the squared L^2-norm of the true density, rather than the density itself. For some alternative ways of using the L^2-distance between the parametric and nonparametric estimators, see also Fan [8], Wen and Wu [9], Tenreiro [10], and the references therein.

The fact that the distribution of a random vector belongs to a given family of probability distributions can be characterized by an infinite set of moment equations. The most elementary example is provided by the distribution function, and this leads to popular tools such as Cramér-von Mises and Kolmogorov-Smirnov tests. See also Khmaladze [11] for extensions to multivariate observations. Alternatively, one can consider the moment generating function or the characteristic function, *etc.* The empirical version of the moments leads to an empirical process that can be used to construct functionals, typically weighted L^2- functionals, to test the goodness-of-fit of a given model. The Cramér-von Mises test, obtained with a univariate distribution function, and the BHEP (Baringhaus-Henze-Epps-Pulley; [12, 13]) tests of multivariate normality, using the characteristic functions, are some such examples. Tests based on the energy distance, which is a weighted L^2-distance between the characteristic functions, are another, more recent example. See Székely and Rizzo [14], Székely and Rizzo [15]. Although the model checking approach based on moment equations does not require a dominated model, most of the attention has been given to parametric models with density, in particular to the Gaussian model. It is worth noting that it is possible to connect some weighted L^2-tests to the L^2-distance kernel based tests by imposing the weights to localize around a point. See Ebner and Henze [16] for an illuminating review.

Our test follows the idea of comparing the density in the model with that of a more general model. However, instead of considering the norm between a nonparametric estimator and the density estimator obtained within the model, we follow an idea introduced by Olkin and Spiegelman [17]. They considered two-component mixtures of densities, where one component is the density estimated in the parametric model, and the other one is a nonparametric density estimator. The mixture weight of the parametric density, which we call the *fitness coefficient*, is chosen by maximum likelihood. The null hypothesis, *i.e.,* the parametric model is correct, is then characterized by the fact that the fitness coefficient is equal to 1. We use the fitness coefficient to construct a test statistic with a the limit in distribution equal to the positive part of a standard normal distribution. Our pivotal statistic is simple, requires no bias correction, and has the same limit distribution under the null hypothesis for any fixed dimension d and general parametric density models. A similar test has proposed by [18] in the context of regression models, where the mixture weight is selected by least squares and thus has an explicit expression.

The paper is organized as follows. Section 2 is devoted to the formal presentation of our method. In Section 2.2 we define the lack-of-fitness coefficient from which the lack-of-fitness test statistic is derived, and provide an insight into our construction. The asymptotic properties of the lack-of-fitness statistic under the composite null hypothesis and alternative hypotheses are derived in Section 3. Our test is consistent against a wide family of alternative hypotheses. Section 4 discusses some implementation aspects, including a simple and effective data-driven bandwidth rule for the lack-of-fitness test based on the likelihood maximization, and a parametric bootstrap method for finite sample corrections. The results of a simulation study are also presented in Section 4. There, the new test is compared with two tests based on a L^2-distance in the spirit of Bickel and Rosenblatt [6]. Their critical values are adjusted by the same simple parametric bootstrap. The parametric models investigated are the two and three dimensional Gaussian families, and the alternative hypotheses are defined as mixtures of Gaussian distributions. Our test shows good performance for sample sizes as small as 50 or 100. We conclude our presentation with a discussion in Section 5. There we note that the lack-of-fit principle is a general one, and can be applied to a wide range of problems, such as testing non-nested parametric models against nonparametric alternatives, testing semiparametric models, testing conditional independence. The Appendix contains the proof of our main results, and additional technical proofs are presented in a Supplementary Material.

2 The method

The aim in this paper is to test a given parametric model of probability densities using an independent sample $X_1, \ldots, X_n$ from $X \in \mathbb{R}^d$ which admits the density f_0. Let $\mathcal{P} = \{f_\theta : \theta \in \Theta\}$ denote the model, where Θ is the set of parameters. The null hypothesis is then

$$\mathcal{H}_0 : \exists \theta_0 \in \Theta \text{ such that } f_0 = f_{\theta_0} \in \mathcal{P}. \tag{1}$$

For simplicity, in the following we assume that θ_0 satisfying (1) is unique. The method proposed below, allows to test the goodness-of-fit of $\mathcal{P}$ against nonparametric alternative hypotheses approaching the model. Let us denote them by $\mathcal{H}_{1,n}$, in which case $f_0 \notin \mathcal{P}$. Under $\mathcal{H}_{1,n}$, the density f_0 may depend on n, but for simplicity we omit this dependence in the notation. In particular, $\mathcal{H}_{1,n}$, $n \geq 1$, can be what is usually called a sequence of local (or directional) alternatives, defined as follows : for some fixed $f_* \notin \mathcal{P}$ and $0 < \delta_n \leq 1$, the sequence of local alternative hypotheses is given by

$$f_0 = (1 - \delta_n) f_{\theta_0} + \delta_n f_*, \quad n \geq 1. \tag{2}$$

With $\delta_n = 0$, $\mathcal{H}_{1,n}$ and $\mathcal{H}_0$ would coincide. It is worth noting, however, that our theory is built without reference to any particular form of f_0, such as that in (2).

To define our method, we need to be more specific on the way θ_0 is estimated, and to introduce the model free estimates.

2.1 Parametric versus nonparametric fit

The null hypothesis (1) is a composite null hypothesis and, for testing it, we need an estimate $\widehat{\theta}_n$ of θ_0. Many common parametric estimation method can be considered to obtain $\widehat{\theta}_n$. Let us consider the class of $M-$estimators [see, for example, 19, Chapter 5]. With independent data, the underlying ideas of this common method for building parametric estimator is to maximize a criterion function of the type

$$\theta \mapsto M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i) \quad \theta \in \Theta.$$

Let

$$\widehat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, M_n(\theta),$$

be the $M-$estimator. This framework includes the maximum likelihood estimator (MLE), for which $m_\theta(X_i) = \log(f_\theta(X_i))$, but also methods based on moments, or robust estimators.

The $M-$estimation approach usually requires that the map $\theta \mapsto \mathbb{E}[M_n(\theta)]$ reaches its maximum at a unique point. Under the null hypothesis (1), the unique maximum of $\mathbb{E}[M_n(\theta)]$ must be θ_0. Under the alternative hypotheses $\mathcal{H}_{1,n}$, we will consider $\overline{\theta}$ a *pseudo-true value* of the model. This is typically defined as

$$\overline{\theta} = \arg\max_{\theta \in \Theta} \mathbb{E}[M_n(\theta)], \tag{3}$$

and, for simplicity, we assume $\overline{\theta}$ to be the uniquely defined one. In general, under $\mathcal{H}_{1,n}$ with $\delta_n > 0$ the pseudo-true value $\overline{\theta}$ depends on n, and $\overline{\theta} \neq \theta_0$. For simplicity, we omit the dependence on n in the notation for the pseudo-true value $\overline{\theta}$. The $\overline{\theta}$ is the element of Θ which corresponds to the least misspecified density in the model $\mathcal{P}$, according to the criterion $M_n(\theta)$. When $\delta_n = 0$ (*i.e.*, under $\mathcal{H}_0$) we have $\overline{\theta} = \theta_0$ for all n.

Remark 1. *Whenever $f_0 \notin \mathcal{P}$ but f_0 approaches the model $\mathcal{P}$ is some sense, for example $\delta_n \downarrow 0$ in (2), it is expected that $\overline{\theta}$ converges to θ_0. It is worth noting that in some cases, $\overline{\theta} = \theta_0$ even when $f_0 \notin \mathcal{P}$. For example when $\mathcal{P}$ is the Gaussian vector model with given variance, $\overline{\theta} - \theta_0$ is exactly equal to the difference between the expectations of the observations under f_0 and f_{θ_0}, respectively. On the one hand, this means that whenever the expectation under f_0 is equal to θ_0, we have $\overline{\theta} = \theta_0$. On the other hand, it is generally expected that the rate of convergence for $\|\overline{\theta} - \theta_0\|$ will be determined by some distance between f_0 and the model $\mathcal{P}$. For example, if f_0 is a sequence of local alternatives as in (2), it is expected that $\|\overline{\theta} - \theta_0\| = O(\delta_n)$.*

Our construction requires also a model free, nonparametric estimate of the true density f_0 of the independent observations at the sample points $X_1, \ldots, X_n \in \mathbb{R}^d$. We consider the *leave-one-out* (LOO) density kernel estimator

$$\widehat{f}_{n,i}^{\text{LOO}} = \frac{1}{(n-1)h_n^d} \sum_{j \neq i} K\left(\frac{X_j - X_i}{h_n}\right), \qquad 1 \leq i \leq n,$$

where $K : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is a symmetric probability density function, and h_n is the bandwidth.

2.2 Lack-of-fitness coefficient

Olkin and Spiegelman [17] considered mixtures of densities $\alpha f_{\widehat{\theta}_n} + (1-\alpha)\widehat{f}_{n,i}$ and proposed a data-driven choice of the mixture weight by maximizing a likelihood criterion :

$$\widehat{\alpha}_n^{\mathrm{OS}} = \arg\max_{\alpha\in[0,1]} \sum_{i=1}^{n} \log\left(\alpha f_{\widehat{\theta}_n}(X_i) + (1-\alpha)\widehat{f}_{n,i}\right). \tag{4}$$

Here, $\widehat{f}_{n,i}$ is the Parzen-Rosenblatt density estimator computed at X_i, which, using our notation, can be written as $n\widehat{f}_{n,i} = (n-1)\widehat{f}_{n,i}^{\mathrm{LOO}} + h_n^{-d}K(0)$. The mixture defined by $\widehat{\alpha}_n^{\mathrm{OS}}$ permits to get a density estimator robust to misspecification while retaining a performance comparable to parametric estimators when the true density is close to the model. Olkin and Spiegelman [17] state that $\widehat{\alpha}_n^{\mathrm{OS}}$ converges to 1 in probability when the model is well specified, and to 0 otherwise. They also claim the rate of convergence of $1-\widehat{\alpha}_n^{\mathrm{OS}}$ when the parametric model is correctly specified. [18] considered the idea of Olkin and Spiegelman in the context of least-squares for regression models, allowing for α on the whole real line, and proved the asymptotic distribution of optimal mixing weight. The appealing feature of least-squares regression estimation is the explicit form of the mixing weight. [20] reconsidered the mixture weight selection (4), with $\widehat{f}_{n,i}$ replaced by a leave-and-repair estimator, which is a modification of $\widehat{f}_{n,i}^{\mathrm{LOO}}$. They named the solution of their maximization problem the *fitness* coefficient. Leave-one-out kernel estimators overcome the undesirable effects caused by this bias term of the Parzen-Rosenblatt when computed at the sample points. They are thus preferable when selecting the mixing weight α.

In this paper, we investigate the asymptotic distribution of a mixture weight selected as in (4). In order to avoid the control of small values of kernel density estimator, we restrict the domain of interest to a compact set $S \subset \mathbb{R}^d$ on which the true density stays away from zero. In particular, this allows to use the more user friendly LOO kernel estimator, and avoid the leave-and-repair estimator considered by [20]. More precisely, let

$$\widehat{\alpha}_n \in \arg\max_{\alpha\in[0,1]} \left\{ \sum_{i=1}^{n} \mathbb{I}_S(X_i) \log\left(\alpha f_{\widehat{\theta}_n}(X_i) + (1-\alpha)\widehat{f}_{n,i}^{\mathrm{LOO}}\right) - (1-\alpha)n\widehat{\beta}_n - \alpha n\widehat{\gamma}_n \right\}, \tag{5}$$

with $\widehat{\beta}_n = \int_S \widehat{f}_n d\lambda$ and $\widehat{\gamma}_n = \int_S f_{\widehat{\theta}_n} d\lambda$. Here, λ denotes the Lebesgue measure on $\mathbb{R}^d$.

Definition 1. *For any bounded set $S \subset \mathbb{R}^d$ such that $\inf_S f_0 > 0$, the random variable $1-\widehat{\alpha}_n$ obtained by (5) is the lack-of-fitness coefficient (on S).*

Let us explain the rationale behind the definition (5). The study of the likelihood-based estimators crucially relies on the property of the Kullback-Leibler (KL) divergence to be non-negative. Restricting the sum in (5) to sample points in S breaks this property. We then need to consider a *generalized Kullback-Leibler (KL) divergence* which extends the KL divergence to non-negative functions which do not necessarily integrate to 1. To define this extension, let us note that for any g_1, g_2 non-negative, measurable functions defined on $\mathbb{R}^d$, such that $\int(g_1+g_2)d\lambda < \infty$, we have :

$$-\int \log(g_1/g_2)g_2 d\lambda + \int (g_1 - g_2)d\lambda \geq \int (\sqrt{g_1} - \sqrt{g_2})^2 d\lambda. \tag{6}$$

In view of this inequality, we consider the following definition for the generalized KL divergence : for any f and f_0 density functions on $\mathbb{R}^d$, and any set $S \subset \mathbb{R}^d$, let

$$KL_S(f_0\|f) = -\int_S \log(f/f_0)f_0 d\lambda + \int_S (f - f_0)d\lambda.$$

As a direct consequence of (6), we have the following result.

Lemma 1. *We have*

$$KL_S(f_0\|f) \geq \int_S (\sqrt{f} - \sqrt{f_0})^2 d\lambda.$$

Consequently, $KL_S(f_0\|f) = 0$ if and only if $f = f_0$ a.e. on S.

Up to terms without influence on the optimization, the criterion optimized in (5) is an approximation of the KL_S divergence, and this explains the role of $\widehat{\beta}_n$ and $\widehat{\gamma}_n$. Let us point out that even in the case where

the densities in the model are supported on the compact S, such that $\widehat{\gamma}_n = 1$, it is still likely that $\widehat{\beta}_n < 1$, and thus $\widehat{\beta}_n$ should be considered in (5).

3 Convergence results for the lack-of-fitness statistics

3.1 The representation of the minimiser of a convex process

A key result used for our theoretical results is an extended version of the "Basic Corollary" given in [21], which we present below. The proof is presented in the Supplementary Material.

Lemma 2. *Let $0 \leq a_n$, $n \geq 1$, be sequence of numbers such that $a_n \to \infty$. Let $A_n : [0, a_n] \to \mathbb{R}$, $n \geq 1$, be a sequence of random convex function such that, for all $g \geq 0$,*

$$A_n(g) = \{g^2 V/2 - g Z_n\} + o_{\mathbb{P}}(1), \tag{7}$$

where $V > 0$ is some constant and $(Z_n)_{n\geq 1}$ is a stochastically bounded sequence of random variables. Then

$$g_n = V^{-1}(Z_n \vee 0) + o_{\mathbb{P}}(1), \qquad \text{where} \quad g_n \in \arg\min_{g\in[0,a_n]} A_n(g).$$

We will use this result with g_n equal to a suitable normalization of the lack-of-fitness coefficient $1-\widehat{\alpha}_n$. Let us point out that the lack-of-fitness coefficient estimator is a $M-$estimator under non-standard conditions, because the limit of the estimator is expected to be on the boundary of the parameter set. Lemma 2 is a powerful theoretical tool allowing to handle general situations, including our non-standard setup.

3.2 Assumptions

For any vector $\mathbf{a}$, $\|\mathbf{a}\|$ denotes its Euclidean norm. Let us recall that f_0 is the true density of X from which the independent sample $X_1, \ldots, X_n$ is drawn. It can depend on n under the alternative hypotheses $\mathcal{H}_{1,n}$.

Assumption A. *The set $S \subset \mathbb{R}^d$ from Definition 1 is compact with nonvoid interior. A constant b exists such that*

$$0 < b \leq \inf_{x\in S} f_0(x) \leq \sup_{x\in S} f_0(x) \leq b^{-1} < \infty.$$

Assumption B. *The density f_0 is twice differentiable on $\mathbb{R}^d$, with squared integrable second order partial derivatives that are uniformly bounded on S, and uniformly with respect to n.*

Assumption C. *The parameter set Θ of the model $\mathcal{P}$, is a subset of some $d_\Theta-$dimensional Euclidean space. Let $\overline{\theta}$ be a pseudo-true value of the model, for example defined as in (3).*

(a) A sequence $\{\delta_n, n \geq 1\} \subset [0,1]$ and a constant $\underline{C} > 0$ exist such that

$$Q_n^* := \int_S \left\{f_{\overline{\theta}} - f_0\right\}^2 d\lambda \geq \underline{C}\delta_n^2. \tag{8}$$

Moreover, we have $\sup_{x\in S} \left|f_{\overline{\theta}}(x) - f_0(x)\right| = O(\delta_n)$.

(b) The estimator $\widehat{\theta}_n$ satisfies $\left\|\widehat{\theta}_n - \overline{\theta}\right\| = O_{\mathbb{P}}(n^{-1/2})$, under the null hypothesis (where $\overline{\theta} = \theta_0$) and under the alternatives $\mathcal{H}_{1,n}$.

(c) For any $x \in S$, $\theta \mapsto f_\theta(x)$ is continuously differentiable. The gradient $\nabla_\theta f_\theta(x) \in \mathbb{R}^{d_\Theta}$ is bounded on $\Theta \times S$ and a constant C_r exists such that

$$\forall \theta, \theta' \in \Theta \text{ and } \forall x \in S, \quad \|\nabla_\theta f_\theta(x) - \nabla_\theta f_{\theta'}(x)\| \leq C_r \|\theta - \theta'\|.$$

Assumption D. *The kernel function $K : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ integrates to 1 and takes one of the two following forms,*

$$(a) \quad K(x) \propto K^{(0)}(\|x\|) \qquad or \qquad (b) \quad K(x) \propto \prod_{k=1}^{d} K^{(0)}(|x_k|),$$

where $K^{(0)} : [0,1] \to \mathbb{R}_{\geq 0}$ *is of bounded variation. The bandwidth sequence satisfies*

$$nh_n^d / \log(n) \to \infty \quad \textit{(variance condition)}, \qquad nh_n^{d+4} \to 0 \quad \textit{(bias condition)}.$$

Assumption A imposes constraints on the choice of S in order to avoid the parts of the support with very low or very high probability. These parts would require a more complex control of quantities related to the nonparametric estimator of the density. In the empirical study section, we use a simple definition of S using some fixed extreme quantiles. Note that the choice of S may also be guided by the practitioner's purposes, who may be interested in focusing on specific parts of the support of X. Assumption B imposes standard regularity conditions on the density model and the density defining the deviation from the model. Condition (a) in Assumption C introduces a distance between the model and the true density of the data. Under the null hypothesis $\mathcal{H}_0$ we necessarily have $Q_n^* = \delta_n = 0$, $\forall n \geq 1$, while $\delta_n > 0$ is expected under the alternative hypotheses $\mathcal{H}_{1,n}$. Combining Assumptions A and B, we get

$$\underline{C} b^2 \delta_n^2 \leq \mathbb{E}\left[\left\{\frac{f_{\theta_0}(X) - f_*(X)}{f_0(X)}\right\}^2 \mathbb{I}_S(X)\right] \leq \lambda(S) b^{-2} \sup_{x \in S} \left|f_{\bar{\theta}}(x) - f_0(x)\right|^2 = O(\delta_n^2),$$

an this double inequality, determined by f_0 but also by the set S, the model $\mathcal{P}$ and the $M-$estimation method, will be used to examine the power of our test. Conditions (b) and (c) in Assumption C introduce mild conditions on the model $\mathcal{P}$ that are often encountered in $M-$estimation; see, for example, [19] and [22]. Finally, the conditions on the kernel function and the bandwidth range in Asssumption D are mild. In order to obtain the technical results involving the density kernel estimators presented in the Supplementary Material, it is very convenient to consider a symmetric, compactly supported kernel K. However, in our implementation, we utilize the Gaussian kernel, supported on $\mathbb{R}^d$. Note that, in practice, it can be considered as compactly supported because its integral over, say, $[-4,4]$ is very close to 1. The conditions imposed on the bandwidth guarantee the uniform convergence of the kernel density estimator [see, e.g., 23]. A range of bandwidths comparable to that we consider is required for the tests based on L^2-distance, see [8]. Our bias condition forces the bias of the kernel estimator to be negligible compared to the variance, which is required to derive a pivotal test statistics. Similar conditions are required for the Bickel and Rosenblatt [6] estimator, see also [24].

3.3 Main results

Let us introduce some more notation. When f is a function defined on $\mathbb{R}^d$, we write f_i instead of $f(X_i)$, such for instance $f_{0,i}$ and $\mathbb{I}_{S,i}$ instead of $f_0(X_i)$ and $\mathbb{I}_S(X_i)$, respectively.

Our main results are obtained by applying Lemma 2 to a suitable objective function $A_n(g)$ derived from the function of α maximized in (5). To build the function $A_n(g)$, we use Taylor expansion, with respect to α, of the function in (5), next suitably rescale α, and finally add some centering terms that do not depend on α. The details are given at the beginning of the proof of Theorem 1, in the Appendix. The following preliminary results, investigates the Taylor expansion terms related to Z_n and V in (7). The first result, which is valid under the null and the alternative hypotheses $\mathcal{H}_{1,n}$, concerns the linear and the quadratic terms related to the nonparametric density estimator in the Taylor expansion of the objective function, that are

$$M_n^{(np)} = \frac{1}{n}\sum_{i=1}^n \left(\frac{\widehat{f}_{n,i}^{\text{LOO}} - f_{0,i}}{f_{0,i}} \mathbb{I}_{S,i} - \widetilde{\beta}_n\right) \quad \text{and} \quad Q_n^{(np)} = \frac{1}{n}\sum_{i=1}^n \left(\frac{\widehat{f}_{n,i}^{\text{LOO}} - f_{0,i}}{f_{0,i}}\right)^2 \mathbb{I}_{S,i},$$

respectively. Here, $\widetilde{\beta}_n = \int_S (\widehat{f}_n - f_0) d\lambda$ is a centering term.

Lemma 3. *Suppose that Assumptions A, B and D are fulfilled. Then, under* $\mathcal{H}_0$ *and* $\mathcal{H}_{1,n}$,

$$h_n^{d/2} n M_n^{(np)} \rightsquigarrow \mathcal{N}(0, 2 v_K \lambda(S)),$$

and

$$h_n^d n Q_n^{(np)} \to v_K \lambda(S), \quad \textit{in probability},$$

where $v_K = \int K^2(u) du$, *and* $\rightsquigarrow$ *denotes convergence in distribution.*

The second preliminary result concerns terms involving the model based estimator :

$$M_n^{(p)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{f_{\widehat{\theta}_n,i} - f_{0,i}}{f_{0,i}}\mathbb{I}_{S,i} - \widetilde{\gamma}_n\right), \quad Q_n^{(p)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{f_{\widehat{\theta}_n,i} - f_{0,i}}{f_{0,i}}\right)^2 \mathbb{I}_{S,i},$$

and

$$C_n^{(p,np)} = \frac{1}{n}\sum_{i=1}^{n}\frac{f_{\widehat{\theta}_n,i} - f_{0,i}}{f_{0,i}}\frac{\widehat{f}_{n,i}^{\mathrm{LOO}} - f_{0,i}}{f_{0,i}}\mathbb{I}_{S,i}. \tag{9}$$

Here, $\widetilde{\gamma}_n = \int_S (f_{\widehat{\theta}_n} - f_0)d\lambda$ is a centering term. These quantities have a different behavior under $\mathcal{H}_0$ and $\mathcal{H}_{1,n}$.

Lemma 4. *Suppose that Assumptions A to C hold true, and δ_n from Condition (a) in Assumption C is such that either $\delta_n \equiv 0$ or $n\delta_n^2 \to \infty$. Then*

$$nM_n^{(p)} = O_{\mathbb{P}}(n^{1/2}\delta_n) + O_{\mathbb{P}}(1).$$

Moreover, a non-random, positive and bounded away from zero sequence $\{u_n, n \geq 1\}$ exists such that

$$nQ_n^{(p)} = n\delta_n^2 \times \{u_n + o_{\mathbb{P}}(1)\} + O_{\mathbb{P}}(1).$$

The third preliminary result concerns a term involving both the model based and model free estimators.

Lemma 5. *Suppose that Assumptions A to D hold true, and δ_n from Condition (a) in Assumption C is such that either $\delta_n \equiv 0$ or $n\delta_n^2 \to \infty$. Then*

$$nC_n^{(p,np)} = O_{\mathbb{P}}(\sqrt{n}h_n^2) + O_{\mathbb{P}}(1) + O_{\mathbb{P}}(nh_n^2\delta_n) + O_{\mathbb{P}}(\sqrt{n}\delta_n).$$

As a consequence of the preliminary lemmas, we deduce that whenever $f_{\overline{\theta}} = f_0$, which is the case under $\mathcal{H}_0$, the linear terms $M_n^{(p)}$, $Q_n^{(p)}$ and the cross-products term $C_n^{(p,np)}$ are negligible compared to $M_n^{(np)}$ as soon as $nh_n^{d+4} \to 0$. As a consequence, the term $h_n^{d/2}nM_n^{(np)}$ will determine the weak convergence of the sequence of lack-of-fitness coefficients $(1 - \widehat{\alpha}_n)$ under $\mathcal{H}_0$.

We now derive the asymptotic behavior of the lack-of-fitness coefficient. We first study the behavior under the null composite hypothesis (1).

Theorem 1. *Suppose that Assumptions A to D are fulfilled. Under $\mathcal{H}_0$,*

$$\left(\frac{v_K\lambda(S)}{2h_n^d}\right)^{1/2}(1 - \widehat{\alpha}_n) \rightsquigarrow \mathcal{N}(0,1) \vee 0,$$

where $\rightsquigarrow$ denotes convergence in distribution, $v_K = \int K^2(u)du$.

Corollary 1. *Let $a \in (0, 1/2)$ and z_a be the $a-$th quantile of the standard normal law. Under the assumptions of Theorem 1, the test defined by*

$$\mathbb{I}\left(\left\{v_K\lambda(S)h_n^{-d}/2\right\}^{1/2}(1 - \widehat{\alpha}_n) \geq z_{1-a}\right), \tag{10}$$

has asymptotic level a.

3.4 Power study

Under the alternative hypotheses where $f_{\overline{\theta}} \neq f_0$, the power of our test is expected to be driven by the quadratic term $Q_n^{(p)}$. Let us first discuss the case of alternatives hypotheses 'slowly' converging towards $\mathcal{H}_0$, that is the case where δ_n from Condition (8) satisfies

$$nh_n^d\delta_n^2 \to \mathfrak{C}, \qquad 0 < \mathfrak{C} \leq \infty. \tag{11}$$

This includes the alternative hypotheses where f_0 does not approach the model $\mathcal{P}$. Under the bias condition $nh_n^{d+4} \to 0$ imposed in Assumption D, condition (11) means that the squared convergence rate of the LOO

kernel estimator, given by $Q_n^{(np)}$, is negligible compared to δ_n^2. By Lemma 4, it also negligible compared to $Q_n^{(p)}$. We show in equation (19) in the Appendix that in this case, a constant $C_{\mathfrak{C}}$ exists, such that $\widehat{\alpha}_n \in [0, C_{\mathfrak{C}}] \subset [0,1)$ with probability tending to 1. In particular, $\widehat{\alpha}_n = o_{\mathbb{P}}(1)$ when $\mathfrak{C} = \infty$. This means that $h_n^{-d/2}(1-\widehat{\alpha}_n) \to \infty$ in probability, and our test is consistent against $\mathcal{H}_{1,n}$ if δ_n satisfies (11). Let us note that $nh_n^d\delta_n^2 = nh_n^{d+4} \times (h_n^{-2}\delta_n)^2$, and given our bias condition $nh_n^{d+4} \to 0$, the condition (11) necessarily requires $h_n^{-2}\delta_n \to \infty$.

By a more refined analysis, we will show that our test is also consistent against alternatives $\mathcal{H}_{1,n}$ defined by sequences δ_n, $n \geq 1$, such that $nh_n^d\delta_n^2 \to 0$ as long as

$$nh_n^{d/2}\delta_n^2 \to \infty \qquad \text{and} \qquad h_n^{-2}\delta_n \to \infty. \tag{12}$$

This case requires a specific treatment because when $nh_n^d\delta_n^2 \to 0$ we get $\widehat{\alpha}_n \to 1$ in probability. However, the second part of condition (12) guarantees that the rate of $Q_n^{(p)}$ dominates those of $C_n^{(p,np)}$ and $M_n^{(np)}$. Then, $h_n^{-d/2}(1-\widehat{\alpha}_n)$ still converges to infinity in probability and our test is able to detect $\mathcal{H}_{1,n}$.

Theorem 2. *Suppose that assumptions of Theorem 1 are fulfilled. Then, under the alternative hypotheses $\mathcal{H}_{1,n}$ satisfying either (11) or (12), for any $C > 0$,*

$$\mathbb{P}\left(h_n^{-d/2}(1-\widehat{\alpha}_n) > C\right) \to 1,$$

and thus the test defined by (10) is consistent.

The condition in (12) can be rewritten under the form

$$\delta_n \gg h_n^2 + \{nh_n^{d/2}\}^{-1/2}. \tag{13}$$

The right-hand side is minimized by

$$h_n \sim n^{-\frac{1}{4+d/2}}. \tag{14}$$

This means that the fastest decreasing rate δ_n allowed by (12) is $\delta_n \gg n^{-2/(4+d/2)}$. In the case of $d = 1$ we get $\delta_n \gg n^{-4/9}$. In particular, assuming in (13) that $h_n^2 \ll \{nh_n^{d/2}\}^{-1/2}$, *i.e.*, $nh_n^{4+d/2} \to 0$, the deviation δ_n can be chosen as $\delta_n \gg \{nh_n^{d/2}\}^{-1/2}$. The same detection rate is established for the Bickel and Rosenblatt [6] test [see 8, Theorem 3.5, case (c2)].

Remark 2. *A careful reading of the proof of Theorem 2 reveals that consistency may be achieved even when the condition $h_n^{-2}\delta_n \to \infty$ is not satisfied. Indeed, the proof is based on the fact that*

$$nh_n^{d/2}\left\{Q_n^{(p)} - C_n^{(p,np)} - M_n^{(p)}\right\} \geq nh_n^{d/2}\delta_n^2\{u_n + o_{\mathbb{P}}(1)\} \to \infty, \qquad \textit{in probability,} \tag{15}$$

with the non-random sequence $\{u_n, n \geq 1\}$ remaining above 0. On the one hand, this is a consequence of Lemma 4 from which we have $M_n^{(p)} = o_{\mathbb{P}}(Q_n^{(p)})$ as soon as $n\delta_n^2 \to \infty$. On the other hand, by Lemma 4 we have $Q_n^{(p)} = \delta_n^2\{u_n + o_{\mathbb{P}}(1)\}$. Consequently, $nh_n^{d/2}Q_n^{(p)} \to \infty$ in probability, as soon as $nh_n^{d/2}\delta_n^2 \to \infty$, as imposed by the first part of the condition (12). Finally, the proof of Lemma 5, given in the Supplement, indicates that if $n\delta_n^2 \to \infty$, $C_n^{(p,np)} = Ch_n^2 \times \{G_n(S) + o_{\mathbb{P}}(1)\}$, where

$$G_n(S) = \int_S \frac{f_{\overline{\theta}} - f_0}{f_0}\,\mathrm{tr}\{\nabla^2 f_0\}d\lambda,$$

and C is a positive constant depending on the kernel K. Here, for any multivariate function f, $\nabla^2 f$ denotes the Hessian matrix of f. It now becomes clear the role of the second part of the condition (12) : if $h_n^{-2}\delta_n \to \infty$, then $C_n^{(p,np)} = o_{\mathbb{P}}(\delta_n^2)$ and, regardless the sign of $G_n(S)$, the term $C_n^{(p,np)}$ is negligible compared to $Q_n^{(p)}$. The consistency of the test is then guaranteed by (15), obtained using $h_n^{-2}\delta_n \to \infty$. However, we deduce from above that the consistency may be achieved even when the condition $h_n^{-2}\delta_n \to \infty$ is not satisfied, depending on the sign of $G_n(S)$, determined by S, f_0 and the $M-$estimation approach used with the model $\mathcal{P}$. The question of how to exploit the behavior of $C_n^{(p,np)}$ to obtain a more powerful test remains open.

4 Simulation study

In this section, we investigate the finite sample behavior of our lack-of-fitness test when testing the Gaussian distribution assumption with i.i.d. data. The lack-of-fitness test is compared with L_2-distance lack-of-fit testing approaches in the spirit of [6] and revisited by [8, 24]. We also consider the BHEP test [12, 13, 16], which will here serve as a benchmark. On contrary to our test or the L_2-distance type test, BHEP is specifically designed to detect departures from the normal distributions. Our simulation study is carried out using the `R` software.

4.1 Implementation aspects

Recall that $f_0 : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ denotes the true density of X from which the independent vectors $X_1, \ldots, X_n$ are drawn, and the hypothesis of interest here is

$$\mathcal{H}_0 \,:\, f_0 \in \{f_\theta \,:\, \theta \in \Theta \subset \mathbb{R}^{d_\Theta}\},$$

where f_θ is the multivariate Gaussian density function with unknown mean μ and covariance Σ. In this case, the components of the parameter θ are given by μ and the entries of the lower triangular part of Σ. The dimension of the parametric model is therefore $d_\Theta = d + d(d+1)/2$. The components of MLE are then obtained from the empirical mean and covariance computed with the sample $X_1, \ldots, X_n$.

Lack-of-fitness calculation.

The lack-of-fitness coefficient $\widehat{\alpha}_n \in [0,1]$ is computed using (5) where the parametric estimate is the MLE $f_{\widehat{\theta}_n}$ and the LOO kernel estimates $\widehat{f}_{n,i}^{\mathrm{LOO}}$ are constructed using the Gaussian kernel K, which in practice can be considered being compactly supported. Strictly speaking, the choice of the Gaussian kernel violates Assumption D, but, as explained below, it facilitates a straightforward implementation of L_2-distance based tests. The set S is defined as the rectangle $[\underline{q}_1, \overline{q}_1] \times \ldots \times [\underline{q}_d, \overline{q}_d]$ where, for any $1 \leq k \leq d$, $\underline{q}_k$ (resp. $\overline{q}_k$) is computed as the $\eta-$th (resp. $(1-\eta)-$th) empirical quantile of the 1-dimensional sample of the k-th coordinates of the observations. We set $\eta = (1 - .95^{1/d})/2$ which, in our setup where the X_i have independent components (see below), means that on average we drop 5% of the data. Note that in our setup, both the integrals $\widehat{\beta}_n$ and $\widehat{\gamma}_n$ have an explicit expression depending on the standard normal distribution function. Given the bandwidth h_n, we use the function `optimize` in `R` to compute $\widehat{\alpha}_n$ in (5). Finally, the lack-of-fitness (LoF) test statistic we use is $h_n^{-d/2}\{1 - \widehat{\alpha}_n\}$.

Alternative approaches.

Some of the competing lack-of-fit approaches we consider are inspired by the test introduced in [6] which follows from the L_2-distance between the nonparametric and the model. The test statistics is defined as

$$\widehat{d}_{1n}^2 = \frac{h_n^{-d/2}}{\sqrt{2\sigma^2}} \left\{ (nh_n^d) \int \left\{ f_{\widehat{\theta}_n}(x) - \widehat{f}_n(x) \right\}^2 dx - v_K \right\}.$$

with $\sigma^2 = \int (\int K(u)K(u+v)du)^2 dv \int \widehat{f}_n(x)^2 dx$. Following [8, Section 5], some explicit formulas are given when a Gaussian kernel is used and when the considered parametric family is Gaussian. We use the same expression in our implementation of the test. A bias corrected variant, proposed in [8] and further studied in [24], is based on the following "smoothed" test statistics

$$\widehat{d}_{2n}^2 = \frac{h_n^{-d/2}}{\sqrt{2\sigma^2}} \left\{ (nh_n^d) \int \left\{ (f_{\widehat{\theta}_n} \star K_{h_n})(x) - \widehat{f}_n(x) \right\}^2 dx - v_K \right\}.$$

In case the kernel is Gaussian and the parametric family is Gaussian, explicit formula similar to $\widehat{d}_{1n}^2$ are available based on properties on the convolution product between Gaussian densities. Let BR1 and BR2 be the tests based on the tests statistics $\widehat{d}_{1n}^2$ and $\widehat{d}_{2n}^2$, respectively.

For the BHEP test we use the implementation proposed in [25] and available in the `mnt` `R` package.

Bandwidth choices.

Both the LoF and the two BR test statistics depend on a bandwidth parameter which plays a crucial role as it calibrates the nonparametric estimate and thus affects the performance of the tests. There is no reason to choose the same bandwidth for the different test statistics, and the asymptotic theory provides little information on how to choose the bandwidths with finite samples in order to ensure accurate level and power.

We consider a set of bandwidths defined as the grid

$$H_n = \left\{ c\widehat{\sigma}_n n^{-1/(4+d/2)} : \ c = 0.5, 0.6, \dots 2 \right\}, \tag{16}$$

where $\widehat{\sigma}_n$ is a scaling parameter. The rate of decrease of the bandwidth is set to be consistent with the discussion in Section 3.4 and faster than the rate in (14). To simplify the computations, we used the fact that the generated X_i have the same variance (see below) and computed the scale factor $\widehat{\sigma}_n$ as the empirical standard deviation using all the nd vector components of the data. Alternatively, the practitioner can pre-process the data by standardizing each component separately, and define H_n in (16) without the $\widehat{\sigma}_n$ factor.

For the LoF, we introduce the following data-driven choice of h on a given grid. Given the maximum likelihood approach to estimate the parameter $\widehat{\theta}_n$ and a value of α, we here follow a similar idea for choosing h by solving

$$\widehat{h}_n \in \arg\max_{h \in H_n} \left\{ \sum_{i=1}^{n} \mathbb{I}_S(X_i) \log\left(\widehat{f}_{n,i}^{\mathrm{LOO}} \right) - n\widehat{\beta}_n \right\}.$$

In this way, both $\widehat{\theta}_n$ and $\widehat{h}_n$ results from optimizing the likelihood. When selecting $\widehat{\alpha}_n$ next, this is meant to put the nonparametric estimate at its advantage, in the context of our likelihood-based approach, and thereby increasing the power of the test. The grid H_n is scaled and thus depends on the sample, but H_n does not changed when calibrating the level using the parametric bootstrap described below. Let $\mathrm{LoF}_{\mathrm{lik}}$ be the test obtained with the statistic the statistic $\widehat{h}_n^{-d/2}\{1-\widehat{\alpha}_n\}$ and the corresponding bootstrap critical values.

To the best of our knowledge, there is no effective data-driven selection of the bandwidth for the BR type tests. Therefore, for comparison purposes, in our simulation study we consider an infeasible bandwidth choice that favours any smoothing-based test and apply it to the two BR tests and also to the LoF test. More precisely, for each of the BR1, BR2 and LoF, we select the bandwidth that gives the best performance among all bandwidths in H_n and leads to a similar or better level (after the same parametric bootstrap calibration) than that obtained with the $\mathrm{LoF}_{\mathrm{lik}}$ approach described above. Similar or better level means a level closest to the nominal one, otherwise closest to the level of $\mathrm{LoF}_{\mathrm{lik}}$. The best performance is measured using the average of the rejection frequencies over the eight deviations from the null hypothesis. Let $\mathrm{LoF}_{\mathrm{best}}$, $\mathrm{BR1}_{\mathrm{best}}$ and $\mathrm{BR2}_{\mathrm{best}}$ be the tests obtained with this infeasible bandwidth rule using the statistics $h_n^{-d/2}\{1-\widehat{\alpha}_n\}$, $\widehat{d}_{1,n}^2$ $\widehat{d}_{2,n}^2$, respectively. Note that the infeasible bandwidth rule we propose gives a fixed bandwidth that does not change across different generated samples, while the data-driven bandwidth used for $\mathrm{LoF}_{\mathrm{lik}}$ depends on the sample.

Parametric bootstrap quantile calibration.

In the context of parametric density models, a simple idea for calibrating the level of the test is to use the parametric bootstrap. Given a data set $X_1, \dots, X_n$, the B bootstrap samples of size n are independently generated from $f_{\widehat{\theta}_n}$, and the bootstrap test statistic values are simply the B values obtained by applying the test statistic to the bootstrap samples. For a level $a \in (0, 1/2)$, the bootstrap critical value is given by the $(1-a)-$th empirical quantile of the $B+1$ test statistic values obtained with the data set $X_1, \dots, X_n$ and the B bootstrap samples. The parametric bootstrap was used for all the tests we consider in the comparisons, that are the LoF, BR1, BR2, and the BHEP tests.

Data generation process.

We now describe the distribution of X under the alternative hypotheses $\mathcal{H}_{1,n}$. The deviations from the Gaussian model are constructed as component-wise mixture models. More precisely, with $X = (X^{(1)}, \dots, X^{(d)})$ we generate each component $X^{(k)}$, $1 \le k \le d$ independently as follows. Let $X_0^{(k)} \sim \mathcal{N}(\mu_0^{(k)}, 1)$, $B^{(k)} \sim \mathcal{B}(\delta)$ and $X_*^{(k)} \sim f_*^{(k)}$ be independent variables, and let

$$X^{(k)} = (1 - B^{(k)})X_0^{(k)} + B^{(k)}X_*^{(k)}.$$

Here, $\mu_0 = (\mu_0^{(1)}, \ldots, \mu_0^{(d)})$ is the true mean vector under the null hypothesis, while true covariance is $\Sigma_0 = I_d$. The multivariate density $f_* = f_*^{(1)} \cdots f_*^{(d)}$ defines the direction of the deviation from the null hypothesis, and δ determines the magnitude of this deviation. When $\delta = 0$, the Bernoulli variables $B^{(k)}$ are all degenerate and equal to 0, and this corresponds to the null hypothesis. When $\delta > 0$, the true distribution of X is a mixture of up to 2^d components. For example, with $d = 2$, $\theta_0 = (\mu_0^{(1)}, \mu_0^{(1)}, 1, 0, 1)$, and

$$f_0(x_1, x_2) = (1-\delta)^2 f_{\theta_0}(x_1, x_2) + \delta(1-\delta) f_*^{(1)}(x_1) f_{\mu_0^{(2)}}(x_2) \\ + \delta(1-\delta) f_{\mu_0^{(1)}}(x_1) f_*^{(2)}(x_2) + \delta^2 f_*^{(1)}(x_1) f_*^{(2)}(x_2),$$

where $f_{\mu_0^{(1)}}$ and $f_{\mu_0^{(2)}}$ are the densities of normal distributions with variance 1 and mean equal to $\mu_0^{(1)}$ and $\mu_0^{(2)}$, respectively. When $\mu_0^{(1)} = \mu_0^{(2)}$ and $f_*^{(1)} = f_*^{(2)}$, the true density of X under the alternative hypotheses is a mixture of three Gaussian densities.

In our simulation setups we consider

$$d \in \{2, 3\}, \quad \delta \in \{0, 0.05, 0.1, \ldots, 0.4\}, \quad B = 1999.$$

Moreover, $n \in \{50, 100, 150, 250\}$, and the level of the test was set to $a = 0.05$. Finally, we consider four types of deviations from $\mathcal{H}_0$. The first two are Gaussian type perturbations given by

$$f_*^{(k)} \sim N(2, 1) \quad \textit{(Model I)} \qquad \text{and} \qquad f_*^{(k)} \sim N(2, 0.25) \quad \textit{(Model II)}, \quad 1 \le k \le d,$$

On the alternative hypotheses, with $d = 2$ and $d = 3$ the true distribution of X is a mixture of three and four multivariate Gaussian densities, respectively. We also consider non Gaussian deviations given by

$$f_*^{(k)} \sim \chi_1^2 \quad \textit{(Model III)} \qquad \text{and} \qquad f_*^{(k)} \sim \chi_1^2 - 1 \quad \textit{(Model IV)}, \quad 1 \le k \le d,$$

respectively.

4.2 Bandwidth sensitivity analysis

A first goal here is to investigate the sensitivity to the bandwidth choice for the smoothing-based tests LoF, BR1 and BR2. We do not yet consider any particular rule for choosing the bandwidth, but rather give a summary of the results obtained considering all the 16 bandwidths in H_n defined in (16). The boxplots of the rejection frequencies are depicted in Figure 1. The rejection frequencies are computed using 1000 independent samples. We consider each method LoF, BR1 and BR2 under the null hypothesis and the deviations like in *Models I* to *IV*, when $d = 2$, $n = 250$ and $h_n \in H_n$. Each boxplot corresponds to a type of test, a type of deviation from the null hypothesis, an amplitude of the deviation δ, and is constructed from 16×1000 points. Thus, the sizes of the boxplots reveal the variability of the rejection frequency with respect to the choice of h_n.

Looking first at the effective level that is achieved by the different test methods, it is clear that the LoF and BR2 are the closest to the nominal level of 0.05. They both show small variability with respect the bandwidth in this situation. The BR1's level is lower than the nominal one, across all the setups. Moreover, while the deviation from the null is getting higher, the LoF is the first to detect the alternative. This occurs in all the setups we consider. Moreover, in other experiments not reported here, the LoF also shows the best power when compared to BR1 and BR2 with other sample sizes ($n = 50, 100, 150$) and dimension ($d = 3$).

To summarize, Figure 1 provides evidence that the LoF approach is much less sensitive to the bandwidth choice compared to the competitors, and is expected to have good level and power for a larger range of bandwidths than the BR type tests.

4.3 Results for selected bandwidths

We now present a comparison between the test LoF_{lik} and the tests LoF_{lik}, as well as the infeasible rule used for LoF_{best}, BR1_{best}, BR2_{best}, introduced in Section 4.1. The aim is to show that the practical LoF_{lik} procedure achieves good performance and compares favorably to all the other three 'optimal bandwidth' approaches which cannot be used in practice. As already mentioned, in the comparison we also include results from the BHEP test which will serve as benchmark. The results with the type of deviation from the null hypothesis corresponding to *Model I*, *Model II*, *Model III* and *Model IV* in the cases $d = 2$ and $d = 3$, are

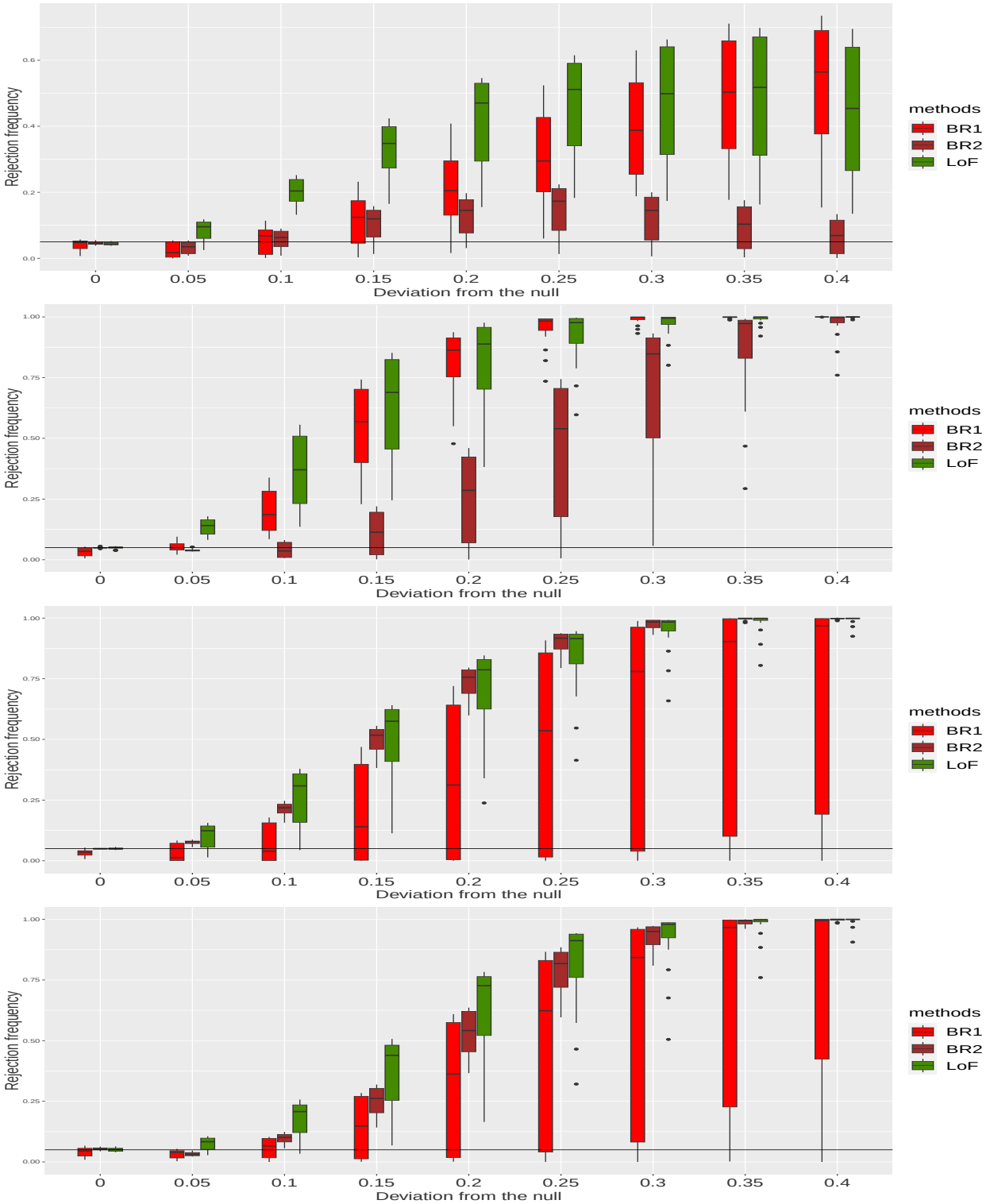

Fig. 1 Boxplot of rejection frequency for LoF, BR1 and BR2 statistics for *Model I* (top) to *Model IV* (bottom) when $n = 250$ and $d = 2$. Each boxplot represents the distribution of the rejection frequency obtained from 1000 independent Monte-Carlo experiments when varying the bandwidth within the range H_n in (16). On the $x-$axis is represented the different deviations δ from the null hypothesis.

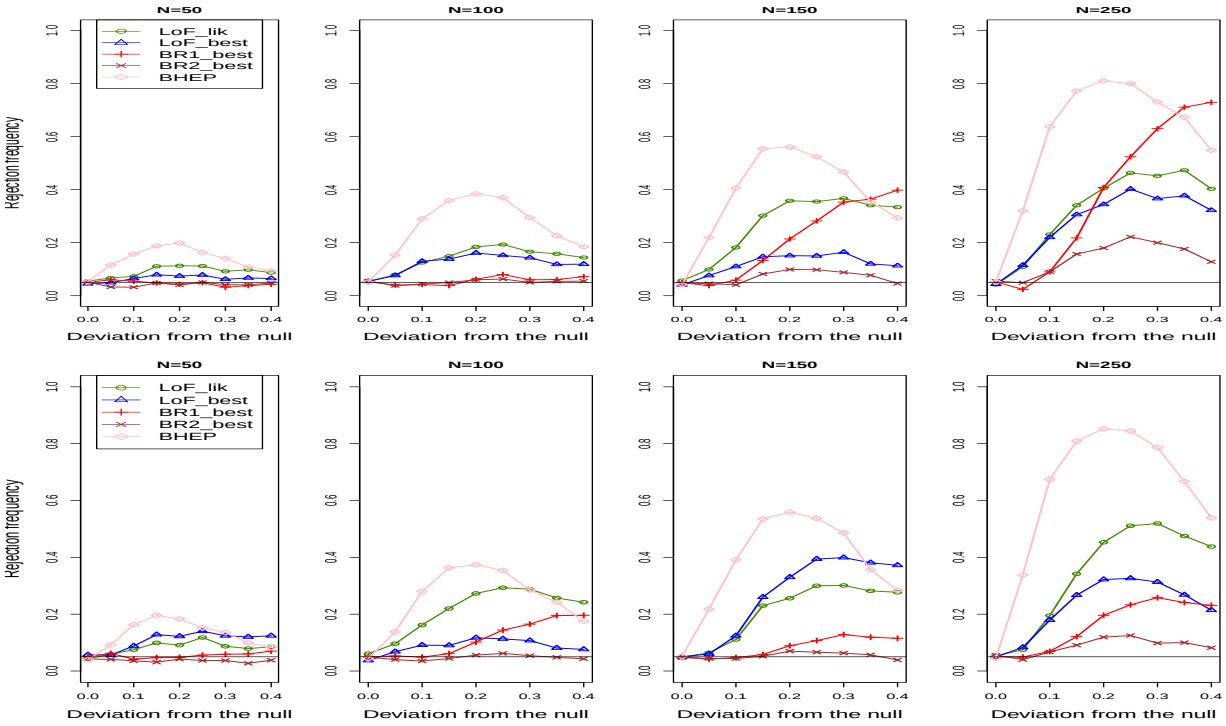

Fig. 2 Rejection frequency obtained from 1000 Monte-Carlo experiments with $\mathrm{LoF}_{\mathrm{lik}}$, $\mathrm{LoF}_{\mathrm{best}}$, $\mathrm{BR1}_{\mathrm{best}}$, $\mathrm{BR2}_{\mathrm{best}}$ and BHEP statistics for $\mathcal{H}_0$ and $\mathcal{H}_{1n}$ obtained with *Model I* (Gaussian mixture with different means and same variance), dimension $d = 2$ (top) and $d = 3$ (bottom), sample sizes $n \in \{50, 100, 150, 250\}$, the bandwidth range H_n in (16). Critical values computed by parametric bootstrap with $B = 1999$.

shown in Figure 2, Figure 3, Figure 4, Figure 5, respectively. First of all, all the approaches achieve good and comparable results under the null as the levels are close to the nominal level $a = 0.05$. Comparing the power of the tests, the $\mathrm{LoF}_{\mathrm{best}}$ test clearly outperforms the $\mathrm{BR1}_{\mathrm{best}}$ and $\mathrm{BR2}_{\mathrm{best}}$test, for almost all n, dimensions d and practically all the types of deviation from the null hypothesis. The only situation where the $\mathrm{BR1}_{\mathrm{best}}$ has a slight advantage is when the alternative is a Gaussian mixture with different mean and variance (i.e., in the *Model II* case) and $d = 3$. In particular, the difference between the two BR and two LoF tests is highly significant with a small sample size. In a consistent way across the models, the BHEP alternative achieves the best results among most of the considered setups, though in some situations $\mathrm{LoF}_{\mathrm{lik}}$ shows better power.

Note that in case of *Model I*, when δ is sufficiently large, the different Gaussian distributions involved tend to overlap, making their mixture indistinguishable from a normal distribution with large variance. This explains the phenomenon observed in Figure 2 where : for δ in the middle of the range of δ, the different tests are able to detect the departure from the null hypothesis, but when δ further increases their power decreases.

5 Discussions and conclusion

As noted by [17], the idea of combining parametric and nonparametric estimators in a two-component mixture is a general principle that, to the best of our knowledge, has not been much explored. In this paper we use this idea to introduce a new goodness-of-fit test for parametric density models. The test statistic does not require a bias correction and, under the null hypothesis, has a simple limit in distribution that does not depend on either the model or the true density. We call our model checking approach lack-of-fitness. [18] propose a related procedure but in a regression context where the estimator of the mixture weight has a closed form. Our statistic requires a bandwidth choice for the nonparametric estimator. We here propose to choose the bandwidth which maximizes the test statistic on a finite grid. The critical values of this adaptive version of our test can be easily calibrated by parametric bootstrapping.

Let us discuss some possible extensions of the lack-of-fitness principle. We focus on how the parametric model $\mathcal{P}$ can be made more general. First, for simplicity, we have assumed that the parameter θ_0 is identifiable. However, a careful inspection of the proofs reveals that we have used this condition to simply control the difference between the $f_{\overline{\theta}}$ and the parametric estimated density $f_{\widehat{\theta}_n}$. This control was used to derive the rates

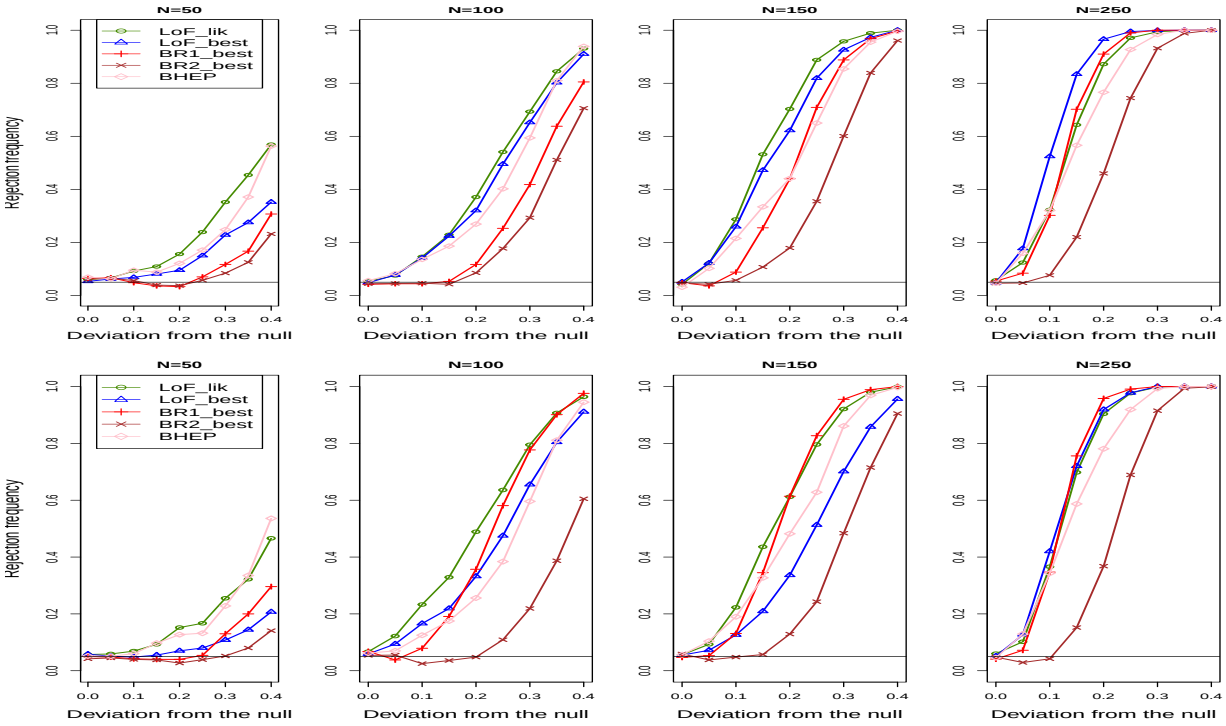

Fig. 3 Rejection frequency obtained from 1000 Monte-Carlo experiments with $\mathrm{LoF}_{\mathrm{lik}}$, $\mathrm{LoF}_{\mathrm{best}}$, $\mathrm{BR1}_{\mathrm{best}}$, $\mathrm{BR2}_{\mathrm{best}}$ and BHEP statistics for $\mathcal{H}_0$ and $\mathcal{H}_{1n}$ obtained with *Model II* (Gaussian mixture with different means and variance), dimension $d = 2$ (top) and $d = 3$ (bottom), sample sizes $n \in \{50, 100, 150, 250\}$, the bandwidth range H_n in (16). Critical values computed by parametric bootstrap with $B = 1999$.

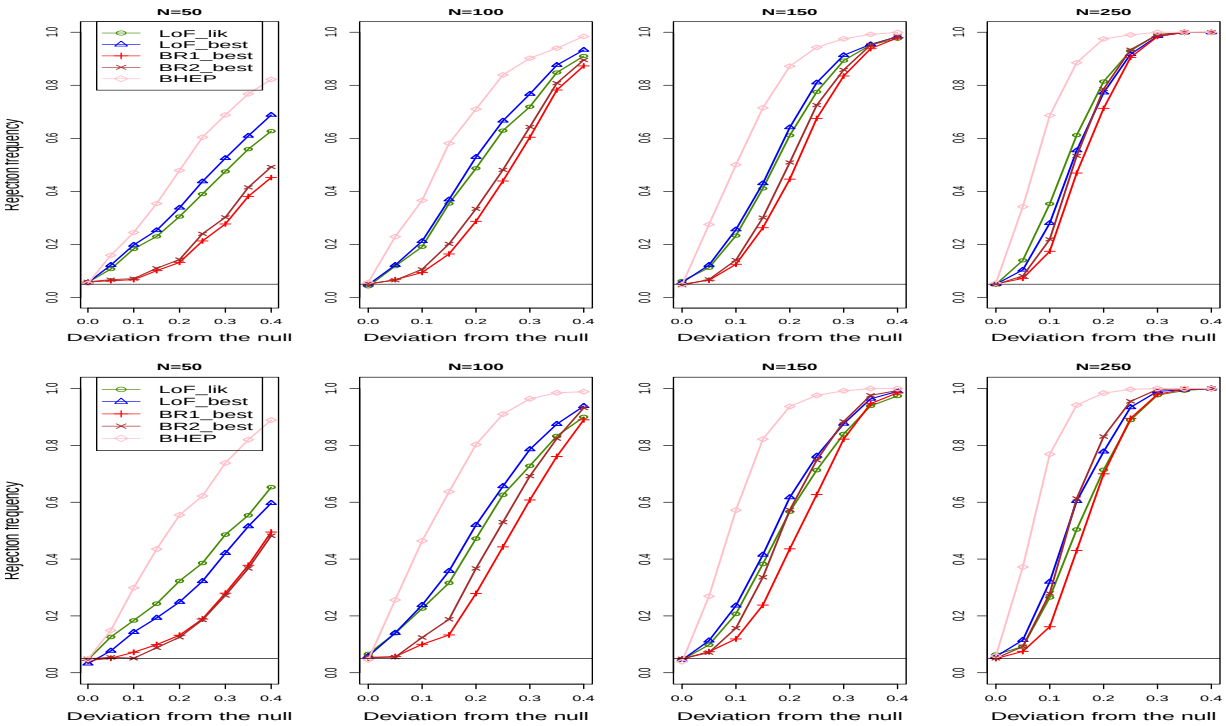

Fig. 4 Rejection frequency obtained from 1000 Monte-Carlo experiments with $\mathrm{LoF}_{\mathrm{lik}}$, $\mathrm{LoF}_{\mathrm{best}}$, $\mathrm{BR1}_{\mathrm{best}}$, $\mathrm{BR2}_{\mathrm{best}}$ and BHEP statistics for $\mathcal{H}_0$ and $\mathcal{H}_{1n}$ obtained with *Model III* (Chi-square deviation), dimension $d = 2$ (top) and $d = 3$ (bottom), sample sizes $n \in \{50, 100, 150, 250\}$, the bandwidth range H_n in (16). Critical values computed by parametric bootstrap with $B = 1999$.

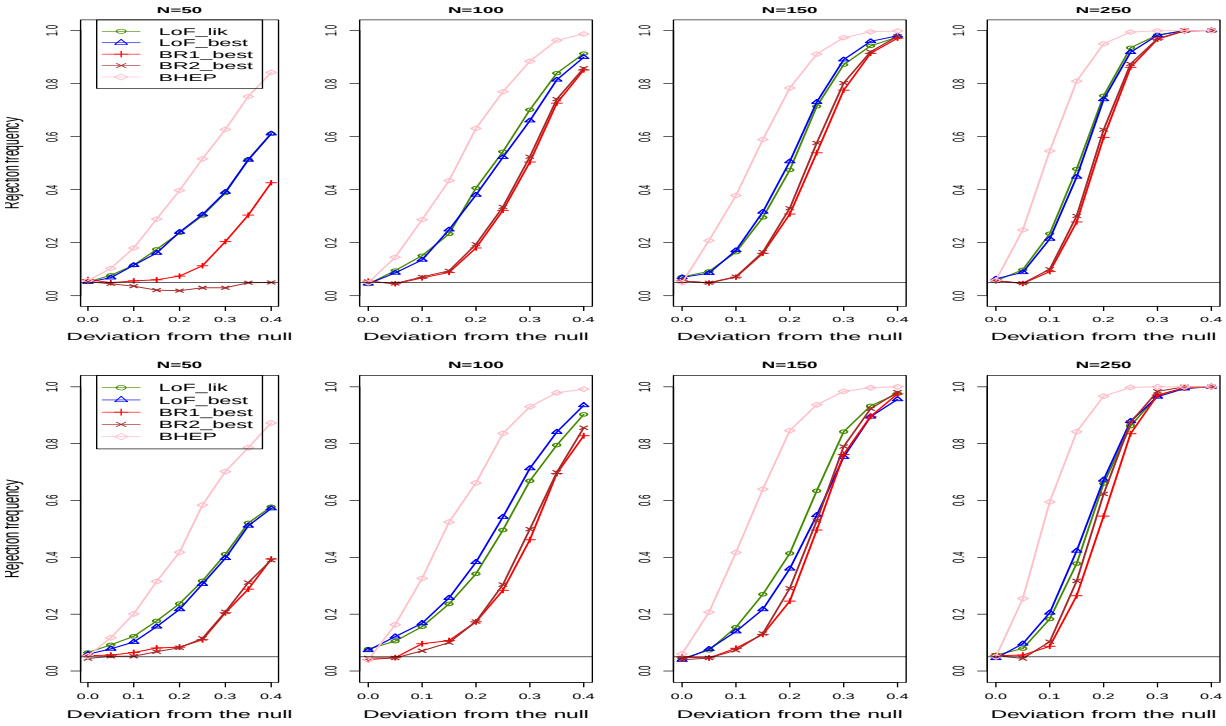

Fig. 5 Rejection frequency obtained from 1000 Monte-Carlo experiments with $\mathrm{LoF}_{\mathrm{lik}}$, $\mathrm{LoF}_{\mathrm{best}}$, $\mathrm{BR1}_{\mathrm{best}}$, $\mathrm{BR2}_{\mathrm{best}}$ and BHEP statistics for $\mathcal{H}_0$ and $\mathcal{H}_{1n}$ obtained with *Model IV* (centered Chi-square deviation), dimension $d = 2$ (top) and $d = 3$ (bottom), sample sizes $n \in \{50, 100, 150, 250\}$, the bandwidth range H_n in (16). Critical values computed by parametric bootstrap with $B = 1999$.

of the different terms in the log-likelihood decomposition with respect to α. The same rates can be derived by using, for example, the uniform convergence rates of empirical processes and $U-$processes indexed by Vapnik-Chervonenkis classes of functions. We thus claim that the assumption of an identifiable parameter θ_0 can be relaxed.

Second, we have considered the problem of testing a model $\mathcal{P}$ against nonparametric alternatives. However, the lack-of-fitness principle also applies to multiple non-overlapping (or separated) parametric models $\mathcal{P}_1, \ldots, \mathcal{P}_K$. See [26–28] for some references on this challenging problem. In this case, we can first construct the K density estimators separately in each model. Next, we can use these K estimators and define $f_{\widehat{\theta}_n}$ as their mixture, with the mixture weights fitted by maximum likelihood. If one of the models $\mathcal{P}_1, \ldots, \mathcal{P}_K$ is correct, then the mixture weight corresponding to the estimators from that model is expected to converge to 1, and all the other $K - 1$ mixture weights are expected to converge to 0. Finally, we apply our lack-of-fitness approach with this $f_{\widehat{\theta}_n}$ and a model-free LOO density kernel estimator. The test statistic is the one studied in this paper which, under mild conditions, has the limit distribution given in Theorem 1 if one of the parametric models is correct. Parametric bootstrap remains a simple way to calibrate the critical values with finite samples.

Before discussing two further extensions, it should be noted that the lack-of-fitness principle consists in mixing an estimator from a 'restricted model' (e.g. parametric) with a 'model free' one (e.g. nonparametric). The theory supporting the lack-of-fitness principle is based on two key aspects. First, a convergence in distribution result as in Theorem 1, which is a consequence of the properties of the 'model free' density estimator. Second, the rate of convergence for the estimator derived in the 'restricted model' that is faster than that of the nonparametric density estimator. This suggests that the lack-of-fitness principle can be applied to more complex situations. For example, let us consider a semiparametric model of elliptical densities $\mathcal{P}_e = \{f(x) \propto g((x-\mu)^\top \Sigma^{-1}(x-\mu)) : \mu \in \mathbb{R}^d, \Sigma \gg 0\}$, where g is some unknown univariate function such that $\int_{\mathbb{R}^d} g(x^\top x)dx = 1$. The mean vector μ and the positive definite matrix Σ can be estimated using the empirical mean and covariance, respectively. Given the estimates $\widehat{\mu}$, $\widehat{\Sigma}$, to construct a semiparametric estimator of the density f_0 in $\mathcal{P}_e$ it suffices to consider the $1-$dimensional nonparametric kernel density estimator of the transformed data $(X_i - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(X_i - \widehat{\mu})$. See [29]. The semiparametric estimator of f_0 can play two roles. On the one hand, it can play the role of a 'model free' estimator if the function g is assumed

to be known in $\mathcal{P}_e$. This would provide a test for a parametric family of elliptical distributions (Gaussian, multivariate $t-$distribution...) against semiparametric alternatives. On the other hand, the semiparametric estimate of f_0 can play the role of a 'restricted model' estimator and be compared with the fully nonparametric kernel density estimator using the lack-of-fitness approach, leading to a nonparametric test of the ellipticity assumption.

Finally, we discuss the problem of testing conditional independence using densities, or equivalently, the problem of nonparametric significance testing for conditional densities. Suppose that $X = (U, V, W) \in \mathbb{R}^d$ with U, V and W random vectors of dimensions d_U, d_V and d_W, respectively. The null hypothesis is the conditional independence $U \perp V \mid W$. Let $f_{U|W}$ (resp. $f_{V|W}$) (resp. f_W) be the conditional density of U given W (resp. the conditional density of V given W) (resp. the density of W). Under the condition $U \perp V \mid W$, the density f_0 of X can be factorized as $f_0 = f_{U|W} f_{V|W} f_W$. Then, the $f_{\widehat{\theta}_n}$ above in the paper can be replaced by the product of the there kernel estimators of the conditional densities $f_{U|W}$, $f_{V|W}$ and the density f_W. This involves smoothing in the dimensions $d_U + d_W$, $d_V + d_W$ and d_W, respectively. On the other hand, f_0 can be estimated by the LOO density estimator in the dimension $d = d_U + d_V + d_W$. Since $d > \max\{d_U + d_W, d_V + d_W\}$, using appropriate bandwidths for the kernel estimators, our lack-of-fitness approach can be used, and the behavior of the lack-of-fitness coefficient under the conditional independence is that given in Theorem 1.

In conclusion, we would like to endorse the opinion of [17] that the lack-of-fitness principle is a resourceful idea that deserves more extensive investigation in the future.

Appendix

Proof of Theorem 1. Since $f_0 \in \mathcal{P}$, we have $\inf_i f_{\hat{\theta}_n, i} > 0$, which guarantees the existence of $\widehat{\alpha}_n$ in (5). Moreover, by standard results on the uniform convergence of the kernel density estimator [see, e.g., 23], with probability tending to 1,

$$\inf_{\alpha \in [0,1]} \min_{i=1,\ldots,n} \{\alpha f_{\widehat{\theta}_n}(X_i) + (1-\alpha)\widehat{f}^{\mathrm{LOO}}_{n,i}\} \geq b/2.$$

Next, we slightly modify the objective function in (5) by centering it using quantities that are not dependent on α. More precisely, let

$$L_n(\alpha) = \sum_{i=1}^{n} \left\{ \log\left(\frac{\alpha f_{\widehat{\theta}_n}(X_i) + (1-\alpha)\widehat{f}^{\mathrm{LOO}}_{n,i}}{f_{0,i}} \right) \mathbb{I}_{S,i} \right\} - (1-\alpha) n \widetilde{\beta}_n - \alpha n \widetilde{\gamma}_n,$$

with $\widetilde{\beta}_n = \int_S (\widehat{f}_n - f_0) d\lambda$ $\widetilde{\gamma}_n = \int_S (f_{\widehat{\theta}_n} - f_0) d\lambda$, and note that $\widehat{\alpha}_n \in \operatorname{argmax}_{[0,1]} L_n(\alpha)$. For any $g \geq 0$, let

$$\alpha_{n,g} = 1 - h_n^{d/2} g \qquad \text{and} \qquad \widehat{g}_n = h_n^{-d/2}(1 - \widehat{\alpha}_n).$$

Since $\alpha_{n,\widehat{g}_n} = \widehat{\alpha}_n$, the value $\widehat{g}_n$ is a point of maximum of the rescaled version of L_n :

$$\widehat{g}_n \in \operatorname{argmax}_{g \in [0, h_n^{-d/2}]} L_n(\alpha_{n,g}).$$

To apply Lemma 2, is actually more convenient to define

$$A_n(g) = -L_n(\alpha_{n,g}) + n M_n^{(p)} - \frac{1}{2} n Q_n^{(p)},$$

which is a convex random function of g, and to notice that

$$\widehat{g}_n \in \operatorname{argmin}_{g \in [0, h_n^{-d/2}]} A_n(g).$$

By second order Taylor expansion of the function, we have

$$\log(1 + (1-\alpha)u + \alpha v) \approx (1-\alpha)u + \alpha v - \frac{1}{2}(1-\alpha)^2 u^2 - \frac{1}{2}\alpha^2 v^2 - \alpha(1-\alpha)uv,$$

for small values of $|u|, |v|$. We can write

$$A_n(g) = -(1 - \alpha_{n,g}) n M_n^{(np)} + (1 - \alpha_{n,g}) n M_n^{(p)}$$

$$+\frac{1}{2}(1-\alpha_{n,g})^2 nQ_n^{(np)} + \frac{1}{2}(\alpha_{n,g}^2-1)nQ_n^{(p)} + (1-\alpha_{n,g})\alpha_{n,g} nC_n^{(p,np)} + R_n(g) =: -gZ_n + g^2 V/2 + \mathfrak{D}_n(g), \quad (17)$$

with $Z_n = h_n^{d/2} nM_n^{(np)}$, $V = v_K \lambda(S)$, $C_n^{(p,np)}$ defined in (9) and $R_n(g)$ is a Taylor expansion remainder. Using that $(\alpha^2-1) = 2(\alpha-1) + (\alpha-1)^2$ and $(1-\alpha)\alpha = (1-\alpha) - (1-\alpha)^2$, we rewrite

$$\mathfrak{D}_n(g) = -g \times \left\{ h_n^{d/2} nQ_n^{(p)} - h_n^{d/2} nC_n^{(p,np)} - h_n^{d/2} nM_n^{(p)} \right\} + g^2 \times \left\{ \frac{1}{2}\left[h_n^d nQ_n^{(np)} - V \right] + h_n^{d/2} \times \left[\frac{1}{2} h_n^{d/2} nQ_n^{(p)} - h_n^{d/2} nC_n^{(p,np)} \right] \right\} + R_n(g). \quad (18)$$

The reminder $R_n(g)$ is clearly negligible. By the Lemmas 4 and 5, under $\mathcal{H}_0$,

$$\forall g \in [0, h_n^{-d/2}], \qquad \mathfrak{D}_n(g) = o_{\mathbb{P}}(1).$$

The application of Lemma 2 then implies

$$\widehat{g}_n = h_n^{-d/2}(1-\widehat{\alpha}_n) = V^{-1}\left(h_n^{d/2} nM_n^{(np)} \vee 0 \right) + o_{\mathbb{P}}(1).$$

Combining Slutsky's Lemma and Lemma 3 leads to the conclusion, under $\mathcal{H}_0$. □

Proof of Theorem 2. Consider first departures from the null hypothesis with δ_n satisfying (11). We show in the following that in this case, $\widehat{\alpha}_n \le c$ with probability tending to 1, where c is some constant in $[0,1)$. Indeed, using the Taylor expansion used in the proof of Theorem 1, but without adding any longer $nM_n^{(p)} - nQ_n^{(p)}/2$ to $L_n(\alpha)$, we get

$$L_n(\alpha) \approx \alpha nM_n^{(p)} + (1-\alpha)nM_n^{(np)} - (1/2)(1-\alpha)^2 nQ_n^{(np)} - (1/2)\alpha^2 nQ_n^{(p)} - \alpha(1-\alpha)nC_n^{(p,np)} + \text{reminder},$$

with a reminder term that can be shown to be sufficiently small, just as in Theorem 1. Lemma 4 implies that $nQ_n^{(p)}/(n\delta_n^2)$ stays larger than $\underline{C}/2$ with probability tending to 1. Moreover, by Lemmas 3, 4 and 5, we get that $nM_n^{(np)}/(n\delta_n^2)$, $nM_n^{(p)}/(n\delta_n^2)$ and $nC_n^{(p,np)}/(n\delta_n^2)$ converge to zero in probability, provided (11) holds true. Moreover, by Lemma 3, we have

$$\frac{nQ_n^{(np)}}{n\delta_n^2} = \frac{nh_n^d Q_n^{(np)}}{nh_n^d \delta_n^2} \to \mathfrak{C}^{-1} \times v_K \lambda(S).$$

We then deduce

$$L_n(\alpha)/(n\delta_n^2) = -\left\{ \mathfrak{C}^{-1} \times v_K \lambda(S) \right\} (1-\alpha)^2/2 - \frac{nQ_n^{(p)}}{n\delta_n^2}\alpha^2/2 + o_{\mathbb{P}}(1).$$

By a simple modification of Lemma 2, we get that, with probability tending to 1,

$$0 \le \arg\min_{\alpha\in[0,1]} \{-L_n(\alpha)/(n\delta_n^2)\} = \arg\max_{\alpha\in[0,1]} L_n(\alpha) = \widehat{\alpha}_n \le C_{\mathfrak{C}} =: \frac{\mathfrak{C}^{-1} \times v_K \lambda(S)}{\underline{C}/2 + \mathfrak{C}^{-1} \times v_K \lambda(S)} \in [0,1). \quad (19)$$

In the regime for δ_n defined by (12) and the condition $nh_n^d \delta_n^2 \to 0$, we have $\widehat{\alpha}_n \to 1$ in probability, and we need a different justification. By Lemmas 4 and 5 and the condition $h_n^{-2}\delta_n \to \infty$, we deduce

$$\mathcal{R}_n := h_n^{d/2} nQ_n^{(p)} - h_n^{d/2} nC_n^{(p,np)} - h_n^{d/2} nM_n^{(p)} = nh_n^{d/2}\delta_n^2 \{u_n + o_{\mathbb{P}}(1)\}.$$

Under the alternative hypotheses, the decomposition (17)-(18) becomes

$$A_n(g) = -gZ_n + g^2V/2 - g\mathcal{R}_n + \text{negligible terms.}$$

Let $C > 0$ and define the event

$$\mathcal{E}_n = \{\mathcal{R}_n > CV - Z_n\},$$

with $Z_n = h_n^{d/2} n M_n^{(np)}$. Under the hypotheses $\mathcal{H}_{1,n}$, we have $\mathbb{P}(\mathcal{E}_n) \to 1$. Let

$$\widehat{g}_n = \arg \min_{0 \le g \le h_n^{-d/2}} A_n(g) \qquad \text{and} \qquad \widehat{g}_{C,n} = \arg \min_{0 \le g \le h_n^{-d/2}} A_{C,n}(g),$$

where $A_{C,n}(g) = -gCV + g^2V/2 +$ negligible terms, with the same negligible terms as in the expression of $A_n(g)$. On the event $\mathcal{E}_n$, we have $\widehat{g}_n \ge \widehat{g}_{C,n}$. On the other hand, the function $A_{C,n}(g) - A_n(g)$ is a linear function of g, so $A_{C,n}(g)$ is also convex. The Lemma 2 implies $\widehat{g}_{C,n} = C + o_{\mathbb{P}}(1)$, and thus $\mathbb{P}(\{\widehat{g}_{C,n} \ge C/2\}) \to 1$. We deduce that $\mathbb{P}(\{\widehat{g}_n \ge C/2\} \cap \mathcal{E}_n) \le \mathbb{P}(\{\widehat{g}_n \ge C/2\}) \to 1$, and the result follows. □

Supplementary information. The proofs of the Lemmas 2 to 5 above, and some additional technical results are provided in a Supplementary Material.

References

[1] Neyman, J.: 'Smooth' test for goodness of fit. Skand Aktuar **1937**(20), 150–199 (1937)

[2] Ledwina, T.: Data-driven version of Neyman's smooth test of fit. J. Amer. Statist. Assoc. **89**(427), 1000–1005 (1994)

[3] Fan, J.: Test of significance based on wavelet thresholding and Neyman's truncation. J. Amer. Statist. Assoc. **91**(434), 674–688 (1996) https://doi.org/10.2307/2291663

[4] Claeskens, G., Hjort, N.L.: Goodness of fit via non-parametric likelihood ratios. Scand. J. Statist. **31**(4), 487–513 (2004)

[5] Cao, R., Lugosi, G.: Goodness-of-fit tests based on the kernel density estimator. Scand. J. Statist. **32**(4), 599–616 (2005) https://doi.org/10.1111/j.1467-9469.2005.00471.x

[6] Bickel, P.J., Rosenblatt, M.: On Some Global Measures of the Deviations of Density Function Estimates. Ann. Statist. **1**(6), 1071–1095 (1973)

[7] Fromont, M., Laurent, B.: Adaptive goodness-of-fit tests in a density model. Ann. Statist. **34**(2), 680–720 (2006)

[8] Fan, Y.: Testing the goodness of fit of a parametric density function by kernel method. Economet. Theor. **10**(2), 316–356 (1994) https://doi.org/10.1017/S0266466600008434

[9] Wen, K., Wu, X.: A guided nonparametric goodness-of-fit test with application to income distributions. Economet. J. **22**(3), 207–222 (2019) https://doi.org/10.1093/ectj/utz007 https://academic.oup.com/ectj/article-pdf/22/3/207/38334816/utz007.pdf

[10] Tenreiro, C.: On automatic kernel density estimate-based tests for goodness-of-fit. TEST **31**(3), 717–748 (2022)

[11] Khmaladze, E.: Unitary transformations, empirical processes and distribution free testing. Bernoulli **22**(1), 563–588 (2016) https://doi.org/10.3150/14-BEJ668

[12] Baringhaus, L., Henze, N.: Limit distributions for mardia's measure of multivariate skewness. The Annals of Statistics, 1889–1902 (1992)

[13] Epps, T.W., Pulley, L.B.: A test for normality based on the empirical characteristic function. Biometrika **70**(3), 723–726 (1983)

[14] Székely, G.J., Rizzo, M.L.: Energy statistics: a class of statistics based on distances. J. Statist. Plann. Inference **143**(8), 1249–1272 (2013) https://doi.org/10.1016/j.jspi.2013.03.018

[15] Székely, G.J., Rizzo, M.L.: The energy of data. Annu. Rev. Stat. Appl. **4**(1), 447–479 (2017)

[16] Ebner, B., Henze, N.: Tests for multivariate normality—a critical review with emphasis on weighted L^2-statistics. TEST **29**(4), 845–892 (2020) https://doi.org/10.1007/s11749-020-00740-0

[17] Olkin, I., Spiegelman, C.H.: A semiparametric approach to density estimation. J. Amer. Statist. Assoc. **82**(399), 858–865 (1987)

[18] Fan, Y., Ullah, A.: Asymptotic normality of a combined regression estimator. J. Multivariate Anal. **71**(2), 191–240 (1999) https://doi.org/10.1006/jmva.1999.1838

[19] Vaart, A.W.: Asymptotic Statistics, p. 443. Cambridge University Press, Cambridge (1998)

[20] Mazo, G., Portier, F.: Parametric versus nonparametric: The fitness coefficient. Scand. J. Statist. **48**(4), 1344–1383 (2021)

[21] Hjort, N.L., Pollard, D.: Asymptotics for minimisers of convex processes. arXiv:1107.3806 (2011)

[22] Newey, W.K., McFadden, D.: Large sample estimation and hypothesis testing. Handbook of Econometrics, vol. 4, pp. 2111–2245. Elsevier (1994)

[23] Giné, E., Guillou, A.: Rates of strong uniform consistency for multivariate kernel density estimators. Ann. Inst. H. Poincaré Probab. Statist. **38**(6), 907–921 (2002)

[24] Fan, Y.: Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters. Economet. Theor. **14**(5), 604–621 (1998)

[25] Henze, N., Wagner, T.: A new approach to the bhep tests for multivariate normality. Journal of Multivariate Analysis **62**(1), 1–23 (1997)

[26] Cox, D.R.: A Return to an Old Paper: Tests of Separate Families of Hypotheses. J. R. Stat. Soc. B **75**(2), 207–215 (2013) https://doi.org/10.1111/rssb.12003 https://academic.oup.com/jrsssb/article-pdf/75/2/207/49510344/jrsssb_75_2_207.pdf

[27] Loh, W.-Y.: A new method for testing separate families of hypotheses. J. Amer. Statist. Assoc. **80**(390), 362–368 (1985)

[28] Li, X., Liu, J., Ying, Z.: Chernoff index for Cox test of separate parametric families. Ann. Statist. **46**(1), 1–29 (2018) https://doi.org/10.1214/16-AOS1532

[29] Liebscher, E.: A semiparametric density estimator based on elliptical distributions. J. Multivariate Anal. **92**(1), 205–225 (2005) https://doi.org/10.1016/j.jmva.2003.09.007