Electronic Journal of Statistics
Vol. 15 (2021) 5570–5597
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1926

Informative goodness-of-fit for multivariate distributions

Sara Algeri

School of Statistics, University of Minnesota,
0461 Church St SE, Minneapolis, MN 55455, USA.
e-mail: salgeri@umn.edu

Abstract: This article discusses an informative goodness-of-fit (iGOF) approach to study multivariate distributions. When the null model is rejected, iGOF allows us to identify the underlying sources of mismodeling and naturally equips practitioners with additional insights on the nature of the deviations from the true distribution. The informative character of the procedure is achieved by exploiting smooth tests and random field theory to facilitate the analysis of multivariate data. Simulation studies show that iGOF enjoys high power for different types of alternatives. The methods presented here directly address the problem of background mismodeling arising in physics and astronomy. It is in these areas that the motivation of this work is rooted.

MSC2020 subject classifications: Primary 62H15, 62P35, 62M40.
Keywords and phrases: Multivariate goodness-of-fit, smooth tests, background mismodeling.

Received April 2021.

Contents

1. Introduction

Scientific motivations. When searching for the signals of new particles, or when aiming to detect new astronomical objects, a common difficulty arising in the analysis of the data collected by the detectors is the impossibility of correctly specifying the background distribution. In physics and astronomy, we typically refer with "background" or "noise" to the signal of all the astrophysical sources which are not those we aim to discover. Unfortunately, since many sources contribute to the background, its distribution is particularly difficult to model [e.g., 4, 14, 30].

Moreover, if the model postulated by the scientists is rejected, it is often difficult to identify the invalidating causes. For instance, instrumental errors may lead to unexpected perturbations in the data distribution, or there may be unpredicted cosmic sources with non-negligible contributions. Moreover, given the complexity of the models investigated through physics experiments, it is often convenient to consider simplified versions of them (typically Gaussian approximations, [e.g., 10]). Hence, it is particularly important to assess the reliability of the simplified models for the data available and, if needed, provide adequate adjustments for them.

Statistical formulation of the problem. In statistical terms, these difficulties translate into two main questions arising in the statistical analysis of multivariate data. Specifically, given a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$, we may wonder:

Q1. *is the distribution of* $\boldsymbol{X}$ *correctly specified and, if not, in what way does the true data distribution diverge from that hypothesized under the null hypothesis?*

Q2. *How can we improve our postulated model? Or in other words, can we provide a data-driven correction for it?*

As noted by Pearson [29], smooth tests, originally introduced by Neyman [26], naturally allow us to capture and model the departure of f from g and thus, they offer the framework to address Q1 and Q2 directly.

In order to provide a high level overview on smooth tests, let f be the true (unknown) probability density function (pdf) of a random variable $X \in \mathbb{R}$, g is the hypothesized density and G the respective cumulative distribution function (cdf). For example, in the above-mentioned problem of background mismodeling, f represents the true background distribution and g is the background model postulated by the scientists. A smooth model for the true probability law f can be specified as

$$f(x) = g(x)d(x) = g(x)\left\{1 + \sum_{j\geq 1} \theta_j T_j\big[G(x)\big]\right\}, \tag{1.1}$$

where $d(x) = \frac{f(x)}{g(x)}$ is the likelihood ratio and the term in the curly brackets is an orthonormal expansion for it. A smooth test [e.g., 26, 11, 19] consists of testing if any of the coefficients θ_j in (1.1) is different from zero. Finally, by estimating

$d(x)$ and constructing adequate confidence bands, it is possible to visualize the nature of the departure of f from g.

Despite their usefulness, smooth tests are mainly limited to the univariate setting. In light of this, the main methodological task of this work is to extend this framework to allow for the analysis of multivariate data.

Main results and organization. The theoretical framework is presented in Section 2. There, we define a suitable expansion of the likelihood ratio through orthonormal functions on the unit cube. As shown in Sections 3 and 4, such representation substantially simplifies the subsequent stages of estimation, model selection and (post-selection) inference. In Section 5, we discuss a simple ANOVA-like testing strategy to identify possible sources of mismodeling. Power studies are conducted via simulations in both Sections 4 and 5. As noted above, this work finds its main motivations in the context of astrophysical searches. Therefore, in Section 6 we illustrate how iGOF can be used to address the problem of mismodeling of the cosmic background considering a realistic simulation from the Fermi Large Area Telescope [8]. While this article mainly focuses on the analysis of continuous data, extensions to the discrete setting are discussed in Section 7. Section 8 collects a summary of the results and a discussion of the limitations of iGOF. A summary of the main notation used throughout the paper is available in Appendix A. Technical proofs are provided in Appendix B. Codes are provided in the Supplementary Material.

2. Theoretical framework

2.1. An orthonormal expansion for the likelihood ratio

Suppose F is the true distribution function of a p-dimensional random vector, $\mathbf{X}$, taking values in $\mathcal{X} \subseteq \mathbb{R}^p$, and denote with G its hypothesized distribution. F and G are assumed to be continuous with densities f and g. Furthermore, assume that $f(\boldsymbol{x}) = 0$ whenever $g(\boldsymbol{x}) = 0$. For every $\boldsymbol{x} = (x_1, \dots, x_p) \in \mathcal{X}$, the hypothesized density g is such that

$$g(\boldsymbol{x}) = \prod_{d=1}^{p} g_d(x_d|\boldsymbol{x}_{<d}),$$

where $\boldsymbol{x}_{<d} = (x_1, \dots, x_{d-1})$ and $g_1, \dots, g_p$ are suitable densities with associated cdfs and quantile functions G_d and G_d^{-1}, for all $d = 1, \dots, p$. The likelihood ratio between F and G can be specified as

$$d(\boldsymbol{u}) = \frac{f\big(\boldsymbol{G}_R^{-1}(\boldsymbol{u})\big)}{g\big(\boldsymbol{G}_R^{-1}(\boldsymbol{u})\big)}, \quad \boldsymbol{u} \in [0,1]^p \tag{2.1}$$

where $\boldsymbol{u} = (u_1, \dots, u_p) = \big(G_1(x_1), \dots, G_p(x_p|\boldsymbol{x}_{<p})\big) = \boldsymbol{G}_R(\boldsymbol{x})$ is the Rosenblatt transformation [33][1], and

$$\boldsymbol{x} = (x_1, \dots, x_d) = \big(G_1^{-1}(u_1), \dots, G_p^{-1}(u_p|\boldsymbol{x}_{<p})\big) = \boldsymbol{G}_R^{-1}(\boldsymbol{u}).$$

[1]Notice that, in general, $\boldsymbol{G}_R(\boldsymbol{x}) \not\equiv G(\boldsymbol{x})$ as the Rosenblatt's transform $\boldsymbol{G}_R(\boldsymbol{x}) \in [0,1]^d$ whereas the cdf $G(\boldsymbol{x}) \in [0,1]$.

In the bivariate setting, for instance, let $G_1 \equiv G_{X_1}$ and $G_2 \equiv G_{X_2|X_1}$, i.e., the hypothesized marginal cdf of X_1 and the hypothesized conditional cdf of $X_2|X_1$, respectively. Hence, (2.1) specifies as

$$d(u_1, u_2) = \frac{f_{X_1X_2}\big(G_1^{-1}(u_1), G_2^{-1}(u_2|x_1)\big)}{g_{X_1X_2}\big(G_1^{-1}(u_1), G_2^{-1}(u_2|x_1)\big)}.$$

Remark 2.1. *As a plausible alternative to Rosenblatt's transform, one could choose each $G_d \equiv G_{X_d}$, which corresponds to assuming independence among the components of $\boldsymbol{X}$. In this setting, if the marginal distributions are correctly specified, (2.1) is the copula density [e.g., 25] of $\boldsymbol{X}$ under G. While this choice could simplify substantially the computations, it would not allow us to test models G which assume a specific dependence structure and the interest is in assessing if the joint distribution G is misspecified. Moreover, it is worth pointing out that there are situations where such transformation cannot be specified (e.g., Section 6).*

To provide a sufficiently detailed representation of the substructures characterizing the distribution of $\boldsymbol{X}$ (see Q1 in Section 1), a natural approach is that of expressing (2.1) by means of a suitable orthonormal basis in $L^2[0,1]^p$. For instance, let $T_{j_d}(u_d)$ be the j_d-th normalized shifted Legendre polynomial evaluated at $u_d = G_d(x_d|\boldsymbol{x}_{<d})$, with $T_0(u_d) = 1$, $T_1(u_d) = \sqrt{12}(u_d - 0.5)$, etc (e.g., Appendix C). Each $\{T_{j_d}(u_d)\}_{j_d \geq 0}$ forms a basis in $L^2[0,1]$. Hence, we can exploit a well known result in Hilbert space theory [e.g., Proposition 2 in 32, p.50] which asserts that given two orthonormal bases $\{\psi_j\}$, $\{\phi_k\}$ for the Hilbert spaces $\mathcal{H}_1$, $\mathcal{H}_2$, then $\{\psi_j \otimes \phi_k\}$ is an orthonormal basis for $\mathcal{H}_1 \otimes \mathcal{H}_2$. It follows that the tensor product basis $\{T_{j_1,\ldots,j_p}(\boldsymbol{u})\}_{j_1\ldots j_p \geq 0}$ of functions

$$T_{j_1\ldots j_p}(\boldsymbol{u}) = \prod_{d=1}^{p} T_{j_d}(u_d) \tag{2.2}$$

forms an orthonormal basis on $L^2[0,1]^p$, the Hilbert space of square integrable function over the p-dimensional unit cube.

Notice that while any orthonormal basis in $[0,1]$ could be used to construct a tensor product basis in $[0,1]^p$, here we focus on the normalized shifted Legendre polynomials. This choice is justified by the fact that the latter are special cases of the so called LP-score functions [e.g., 22]. As discussed in Section 7, these basis functions allow for extensions to the discrete setting.

Finally, under the assumption that $d(\boldsymbol{u}) \in L^2[0,1]^p$, we can write

$$d(\boldsymbol{u}) = \sum_{j_1 \geq 0, \ldots, j_p \geq 0} \theta_{j_1\ldots j_p} T_{j_1\ldots j_p}(\boldsymbol{u}), \qquad \boldsymbol{u} \in [0,1]^p \tag{2.3}$$

with $\theta_{j_1\ldots j_p} = \int_{[0,1]^p} T_{j_1\ldots j_p}(\boldsymbol{u}) d(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}$. The expansion in (2.3) follows from Theorem II.6 in Reed and Simon [32] and it is equivalent to say that the sum on the right-hand side converges to $d(\boldsymbol{u})$ in $L^2[0,1]^p$.

As noted by an anonymous referee, the likelihood ratio can also be expanded on the original domain $\mathcal{X} \subseteq \mathbb{R}^p$ by means of any set of bounded functions which are orthogonal with respect to G. In our context, this is achieved by combining the Legendre polynomials and Rosemblatt's transform. The latter provides the additional advantage of allowing us to work on the compact compact domain $[0,1]^p$. As it will become clear in Section 4, this is particularly useful as one can exploit results from random field theory to construct simultaneous confidence bands even when $\mathcal{X} \equiv \mathbb{R}^p$. Moreover, for visualization purposes, it may be particularly advantageous to work in the quantile domain when testing long tailed distributions. In this setting, we may expect that only a few observations have been detected over large regions of the $\mathcal{X}$ domain and thus the quantile representation allows us to magnify the differences observed over the the most "data-abundant" regions. An more detailed discussion of this aspect, and adequate graphical comparisons can be found in Algeri [3]. Finally, it is worth pointing out that, the quantile functions G_d^{-1} in $\boldsymbol{G}_R^{-1}(\boldsymbol{u})$ are used in (2.1) with the only purpose of highlighting the dependence of d on $\boldsymbol{u}$, when working in the quantile domain. In practice, however, estimation and inference focus entirely on (2.3) (see Sections 3 and 4) and thus one needs not to compute $\boldsymbol{G}_R^{-1}(\boldsymbol{u})$.

3. Estimation

The summations in (2.3) are taken up to infinity. However, to make the expansion operational, it is necessary to truncate the series in (2.3) at integers values $m_1, \ldots, m_p$. That is because, effectively, the coefficients $\theta_{j_1 \ldots j_p}$ need to be estimated and, consequently, the more terms are included in (2.3), the larger the variance of the resulting estimator of $d(\boldsymbol{u})$ (see Section 4.2 for a more detailed discussion on model selection).

For the sake of simplifying the notation in this section and those to follow, denote with $\mathcal{K}$ the set

$$\mathcal{K} := \left\{ \{j_1 \ldots j_p\}, \text{ with } j_d = 0, \ldots, m_d, \text{ for all } d = 1, \ldots, p, \text{ and } \sum_{d=1}^{p} j_d \neq 0 \right\} \tag{3.1}$$

of cardinality $|\mathcal{K}| = M = \prod_{d=1}^{p}(m_d + 1) - 1$. That is, $\mathcal{K}$ contains all the p-tuples $\{j_1 \ldots j_p\}$ of indexes $j_d = 0, \ldots, m_d$, $d = 1, \ldots, p$ apart from the p-tuple $\{0 \ldots 0\}$, since $\theta_{0 \ldots 0} = 1$ (see (B.11) in Appendix B). Let $\boldsymbol{\theta}$ be the $M \times 1$ vector of components θ_k, with $k \in \mathcal{K}$. Similarly, denote with $\boldsymbol{T}(\boldsymbol{u})$ the $M \times 1$ vector of elements $T_k(\boldsymbol{u})$, $k \in \mathcal{K}$.

Consider $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, a sample of n i.i.d. observations from $\boldsymbol{X}$, and let $\boldsymbol{U} = \boldsymbol{G}_R(\boldsymbol{X})$ be the respective Rosenblatt transformation. Denote with $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ the sample of elements $\boldsymbol{u}_i = \boldsymbol{G}_R(\boldsymbol{x}_i)$. The parameter $\boldsymbol{\theta}$ can be estimated by means of the vector $\widehat{\boldsymbol{\theta}}$ of components

$$\widehat{\theta}_k = \frac{1}{n} \sum_{i=1}^{n} T_k\big(\boldsymbol{G}_R(\boldsymbol{x}_i)\big) = \frac{1}{n} \sum_{i=1}^{n} T_k(\boldsymbol{u}_i) \quad \text{for all } k \in \mathcal{K}\,, \tag{3.2}$$

The mean and covariance matrix of $\widehat{\boldsymbol{\theta}}$ and an estimator of $d(\boldsymbol{u})$ are given in Proposition 3.1.

Proposition 3.1. *The likelihood ratio* $d(\boldsymbol{u})$ *is the density of the random vector* $\boldsymbol{U}$ *and*

$$E[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta} \quad \text{and} \quad Cov(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\Sigma} \tag{3.3}$$

where $\boldsymbol{\Sigma}$ *has diagonal elements* $\frac{\sigma_k^2}{n} = \frac{1}{n}V\big[T_k(\boldsymbol{U})\big]$ *and non-diagonal elements* $\frac{\sigma_{k,h}}{n} = \frac{1}{n}Cov\big[T_k(\boldsymbol{U}), T_h(\boldsymbol{U})\big]$, *with* $k, h \in \mathcal{K}$. *Furthermore, if* $F \equiv G$, *the equalities in* (3.3) *reduce to*

$$E[\widehat{\boldsymbol{\theta}}] = \mathbf{0} \quad \text{and} \quad Cov(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}\boldsymbol{I}_M, \tag{3.4}$$

where $\mathbf{0}$ *is the* $M \times 1$ *zero vector and* $\boldsymbol{I}_M$ *is the* $M \times M$ *identity matrix.*

Finally, an estimator of $d(\boldsymbol{u})$ *is*

$$\widehat{d}(\boldsymbol{u}) = 1 + \widehat{\boldsymbol{\theta}}'\boldsymbol{T}(\boldsymbol{u}), \tag{3.5}$$

and has variance $V\Big[\widehat{d}(\boldsymbol{u})\Big] = \boldsymbol{T}(\boldsymbol{u})'\boldsymbol{\Sigma}\boldsymbol{T}(\boldsymbol{u})$.

Combining (1.1), (2.1) and (3.5) an estimate of f is

$$\widehat{f}(\boldsymbol{x}) = g(\boldsymbol{x})\widehat{d}(\boldsymbol{x}) = g(\boldsymbol{x})\big[1 + \widehat{\boldsymbol{\theta}}'\boldsymbol{T}\big(\boldsymbol{G}_R(\boldsymbol{x})\big)\big] \tag{3.6}$$

Notice that the estimator $\widehat{f}$ incorporates the information carried by the hypothesized model g; whereas, the estimator in the square brackets provides a data-driven correction for it.

Furthermore, define the integrated squared bias (ISB) of $\widehat{d}(\boldsymbol{u})$ as

$$ISB = \int_{[0,1]^p} \Big(E[\widehat{d}(\boldsymbol{u})] - d(\boldsymbol{u})\Big)^2 \mathrm{d}\boldsymbol{u}. \tag{3.7}$$

From Proposition 3.2 it follows that the closer g is to f in terms of squared normalized distance the lower the ISB of $\widehat{d}(\boldsymbol{u})$.

Proposition 3.2. *The integrated squared bias of the estimator in* (3.5) *is*

$$\begin{aligned} ISB &= \int_{[0,1]^p} \left(\frac{f\big(\boldsymbol{G}_R^{-1}(\boldsymbol{u})\big) - g\big(\boldsymbol{G}_R^{-1}(\boldsymbol{u})\big)}{g\big(\boldsymbol{G}_R^{-1}(\boldsymbol{u})\big)}\right)^2 d\boldsymbol{u} - \boldsymbol{\theta}'\mathbf{1} \\ &= \int_{\mathcal{X}} \left(\frac{f(\boldsymbol{x}) - g(\boldsymbol{x})}{g(\boldsymbol{x})}\right)^2 g(\boldsymbol{x})\, d\boldsymbol{x} - \boldsymbol{\theta}'\mathbf{1} \end{aligned} \tag{3.8}$$

where $\mathbf{1}$ *is the* $M \times 1$ *unit vector.*

The estimate in (3.6) is essentially that of a smooth model [e.g., 31], that is, a smoothed version of the true underlying probability function. Similarly to the smooth model proposed by Barton [11] in the univariate setting, the estimator

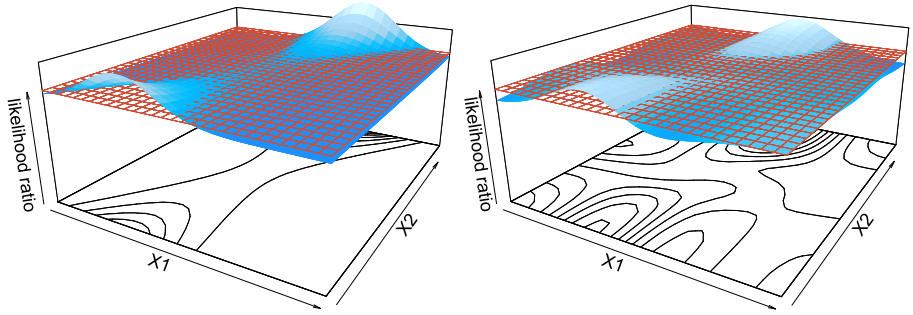

FIG 1. *True (left panel) and estimated (right panel) likelihood ratio for Example I. The estimate on the right has been obtained via (3.5) with $m_1 = 4$ and $m_2 = 3$. The components of the $\widehat{\boldsymbol{\theta}}$ vector have been selected via the AIC criterion in (4.7).*

in (3.6) may lead to estimate that are not *bona-fide*, i.e, they may be negative and/or they may not integrate/sum up to one. In this manuscript we focus on (3.6) mostly for the sake of mathematical convenience when constructing simultaneous confidence bands in Section 4. Nonetheless, bona-fide estimators can be constructed similarly to the univariate case as described in Algeri and Zhang [6].

Example I. In direct searches for dark matter, the dominant background sources are neutron recoils which may produce signals mimicking those expected from dark matter candidates [e.g., 40]. As a toy example, suppose we are interested in assessing the validity of a given distribution for the nuclear recoil background specified over the energy region $\mathcal{X} = [5, 20]KeVnr \times [0, 17]KeVnr$. Each observations in $\mathcal{X}$ corresponds to the scintillation of photons (X_1) and ionization electrons (X_2) [e.g., 7]. The hypothesized background distribution, $G_{X_1X_1}$, is that of a truncated bivariate normal with mean vector $(12, 8)$, variances 8 and 12 and covariance 2. Moreover, suppose that one additional background source is present. The latter is also a bivariate normal with the same mean vector, variances 4 and 20 and covariance 5. Thus, the true model, $F_{X_1X_2}$, involves a mixture of two, overlapping truncated bivariate Gaussians with mixture parameter 0.15. In order to estimate the likelihood ratio, set $G_1 = G_{X_1}(x_1)$ and $G_2 = G_{X_2|X_1}$. The estimated likelihood ratio, obtained over a sample of $n = 5,000$, is shown in the right panel of Figure 1, whereas the left panel shows the true likelihood ratio. A closed form expression for the estimate shown on the right panel is given in equation (C.1) in Appendix C. The estimate obtained recovers the main departures from uniformity. Specifically, the second mixture component contributing to $F_{X_1X_2}$ leads the an inflation of the variance of X_2; as a result, $F_{X_1X_2}$ exhibits higher tails than $G_{X_1X_2}$ in the direction of the first eigenvector. Such departure becomes more and more prominent when moving from the center of the distribution towards the truncation points $X_2 = 0$ and $X_2 = 17$. Finally, the contours of $\widehat{d}(\boldsymbol{u})$ highlight that the estimator is rather noisy. Therefore, it is important to investigate the properties of (3.5) to assess the significance of the deviations observed.

4. Inference and model selection

4.1. Pre-selection inference

A smooth test for $H_0 : G \equiv F$ versus $H_1 : G \not\equiv F$ consists in reformulating the problem as a test for uniformity of $\boldsymbol{U}$. Specifically, (2.1) implies that $F \equiv G$ whenever $d(\boldsymbol{u}) = 1$, and thus

$$H_0 : d(\boldsymbol{u}) = 1 \quad \text{for all } \boldsymbol{u} \in [0,1]^p \quad \text{vs} \quad H_1 : \exists \boldsymbol{u} \in [0,1]^p \text{ such that } d(\boldsymbol{u}) \neq 1. \tag{4.1}$$

It is easy to see that $d(\boldsymbol{u}) = 1$ for all $\boldsymbol{u} \in [0,1]^p$, when all θ_k, $k \in \mathcal{K}$, are identically equal to zero. Hence, in practice, we test

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \qquad \text{vs} \qquad H_1 : \boldsymbol{\theta} \neq \mathbf{0}. \tag{4.2}$$

Notice that H_0 in (4.1) implies H_0 in (4.2), but the opposite is not true in general. Whereas, H_1 in (4.2) does imply H_1 in (4.1). With a little abuse of nomenclature, in this section and those to follow, we will refer to G as the "null model". Furthermore, we will refer to H_0 in (4.1) when generically saying "under H_0". However, most of the results presented here, only require validity of the milder H_0 in (4.2).

To conduct inference, we rely on the so-called *deviance* test statistics, i.e.,

$$D = n\widehat{\boldsymbol{\theta}}'\widehat{\boldsymbol{\theta}}. \tag{4.3}$$

Its asymptotic null distribution is given in Theorem 4.1.

Theorem 4.1. *If H_0 is true, then*

$$\sqrt{n}\widehat{\boldsymbol{\theta}} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}), \quad \text{as } n \to \infty \tag{4.4}$$

where $N(\mathbf{0}, \boldsymbol{I})$ denotes a standard multivariate normal distribution. Furthermore,

$$D \xrightarrow{d} \chi^2_M, \quad \text{as } n \to \infty, \tag{4.5}$$

where M is the size of $\widehat{\boldsymbol{\theta}}$.

The asymptotic distribution of the random field $\widehat{d}(\boldsymbol{u})$ is derived in Theorem 4.2 below. This result is particularly useful for us to construct simultaneous confidence bands as described in Section 4.2.

Theorem 4.2. *Denote with $\{\widehat{d}(\boldsymbol{u})\}$ the random field indexed by $\boldsymbol{u} \in [0,1]^p$ with components as in (3.5). Moreover, assume that $\widehat{\theta}_k = o(n^{-1/2})$ for all $k \notin \mathcal{K}$. If H_0 is true,*

$$\left\{ \frac{\widehat{d}(\boldsymbol{u}) - 1}{\sqrt{\frac{1}{n}\boldsymbol{T}(\boldsymbol{u})'\boldsymbol{T}(\boldsymbol{u})}} \right\} \xrightarrow{d} \boldsymbol{Z}(\boldsymbol{u}), \quad \text{as } n \to \infty, \tag{4.6}$$

where $\boldsymbol{Z}(\boldsymbol{u})$ denotes a Gaussian random field with mean zero, unit variance and covariance function $Cov\Big(\boldsymbol{Z}(\boldsymbol{u}), \boldsymbol{Z}(\boldsymbol{u}^\dagger)\Big) = \frac{\boldsymbol{T}(\boldsymbol{u})'\boldsymbol{T}(\boldsymbol{u}^\dagger)}{\sqrt{\boldsymbol{T}(\boldsymbol{u})'\boldsymbol{T}(\boldsymbol{u})\boldsymbol{T}(\boldsymbol{u}^\dagger)'\boldsymbol{T}(\boldsymbol{u}^\dagger)}}$.

At this stage, constructing inference on the basis of Theorems 4.1 and 4.2 would be tempting. However, to guarantee the validity of our results we must take into account that, when estimating the likelihood ratio in (3.5), a model selection procedure is likely to be implemented. Unfortunately, when a model is selected from a pool of possibilities, such process introduces an additional source of variability, and thus, the resulting inference is automatically affected [e.g., 12]. Section 4.2 addresses this aspect directly.

4.2. Post-selection inference

The estimate of the likelihood ratio considered so far involves up to M functions $T_k(\boldsymbol{u})$. Nonetheless, it is possible that not all of these M terms are needed to capture the departures of G from F and indeed, it is often convenient to remove some of them to avoid unnecessary sources of noise. Various criteria have been proposed in literature for density estimation and smooth models [e.g., 21, 3] and they can be easily extended to the multivariate setting. Here, we focus on the approach of Mukhopadhyay [21] and which specifies as follows.

Let $\widehat{\theta}_{(k)}$ be the k-th largest $\widehat{\theta}_k$ estimate in order of magnitude, for $k \in \mathcal{K}$, i.e., $\widehat{\theta}^2_{(1)} \geq \widehat{\theta}^2_{(2)} \geq \cdots \geq \widehat{\theta}^2_{(M)}$. Select the K largest coefficients which maximize either

$$\text{BIC}(K) = \sum_{(k)=1}^{K} \widehat{\theta}^2_{(k)} - \frac{K \log n}{n} \quad \text{or} \quad \text{AIC}(K) = \sum_{(k)=1}^{K} \widehat{\theta}^2_{(k)} - \frac{2K}{n}. \tag{4.7}$$

Notice that, as defined in (3.1), each $k \in \mathcal{K}$ is a p-tuple of indexes $j_1 \dots j_p$, whereas (k) is the integer value corresponding to the order of magnitude of the respective coefficient $\widehat{\theta}_k$. Hence, the summations in (4.7) and those to follow are taken over $(k) = 1, \dots, K$, that is, the K p-tuples of indexes $j_1 \dots j_p$ with the K-th largest estimates $\widehat{\theta}_k$.

An estimate of $d(\boldsymbol{u})$, is then selected via (4.7) from the family of estimators

$$\widehat{d}_{(K)}(\boldsymbol{u}) = 1 + \sum_{(k)=1}^{K} \widehat{\theta}_{(k)} T_{(k)}(\boldsymbol{u}), \quad K = 1, \dots, M \tag{4.8}$$

where the subscript (K) on the left-hand-side is used to emphasize that the estimator in (4.8) includes only the K-th largest $\widehat{\theta}_k$ estimated coefficients. Clearly, the choice of BIC or AIC is arbitrary and, from a practical standpoint, when $n > 7 (\approx e^2)$, the BIC assigns a heavier penalty than AIC for increasing values of K and, consequently, it often leads to smoother estimators than AIC.

The selection rules in (4.7) compare M possible models assuming that each m_d, for $d = 1, \dots, D$ was fixed before the researcher looked at the data. Valid post-selection inference can then be constructed as in Corollaries 4.1 and 4.2. The respective proofs are provided in Appendix B.

Corollary 4.1. *Denote with* $\widehat{d}_{(K*)}$ *the estimator of* $d(\boldsymbol{u})$ *selected via* (4.7)*, and let* $D_{(K^*)} = \sum_{(k)=1}^{K^*} \widehat{\theta}^2_{(k)}$ *be the respective deviance statistics. As* $n \to \infty$*, a valid post-selection bound for the p-value to test* (4.1) *is*

$$p\text{-}value_{adj} = P(\chi^2_M > D_{obs}), \tag{4.9}$$

where D_{obs} *is the value of* $D_{(K^*)}$ *observed.*

The bound in (4.9) follows from the fact that the estimators in (4.8) are nested, for all $K = 1, \ldots, M-1$. As a result, each $D_{(K)} = \sum_{(k)=1}^{K} \widehat{\theta}^2_{(k)}$ is stochastically lower or equal than $D_{(M)} = \sum_{(k)=1}^{M} \widehat{\theta}^2_{(k)}$, that is, for all $K = 1, \ldots, M-1$, $P(D_{(K)} > D_{\text{obs}})$ is smaller than $P(D_{(M)} > D_{\text{obs}})$.

In order to grasp further insights on the deviations of G from F, it is worth constructing adequate confidence bands. This can be done, while accounting for post-selection adjustments, as in Corollary 4.2.

Corollary 4.2. *Denote with* $\widehat{d}_{(K^*)}$ *the estimator of* $d(\boldsymbol{u})$ *selected via* (4.7)*, and let* $SE_0\big[\widehat{d}_{(K^*)}(\boldsymbol{u})\big]$ *be its standard error under* H_0*. Moreover, assume that,*

$$\left[\sup_{\boldsymbol{u}}\left\{\frac{\widehat{d}_{(K)}(\boldsymbol{u})-1}{SE_0\big(\widehat{d}_{(K)}(\boldsymbol{u})\big)}\right\}\middle| K^*=K\right] \preceq \left[\sup_{\boldsymbol{u}}\left\{\frac{\widehat{d}_{(M)}(\boldsymbol{u})-1}{SE_0\big(\widehat{d}_{(M)}(\boldsymbol{u})\big)}\right\}\middle| K^*=K\right] \tag{4.10}$$

for all $K = 1, \ldots, M-1$*. Valid (post-selection adjusted) simultaneous* $(1-\alpha)\%$ *confidence regions, under* H_0 *in* (4.1)*, are*

$$\left[1 - c_{\alpha/2} SE_0\big[\widehat{d}_{(K^*)}(\boldsymbol{u})\big], 1 + c_{\alpha/2} SE_0\big[\widehat{d}_{(K^*)}(\boldsymbol{u})\big]\right] \qquad \text{for all } \boldsymbol{u} \in [0,1]^p \tag{4.11}$$

with $c_{\alpha/2}$ *such that*

$$P\left(\sup_{\boldsymbol{u}}\left\{\frac{\widehat{d}_{(M)}(\boldsymbol{u})-1}{SE_0\big[\widehat{d}_{(M)}(\boldsymbol{u})\big]}\right\} > c_{\alpha/2}\middle| H_0\right) = \frac{\alpha}{2}. \tag{4.12}$$

In (4.10), "$\preceq$" indicates that the left-hand-side is stochastically lower or equal than the right-hand-side. Intuitively, the validity of (4.10) in practical settings follows from the fact that $\widehat{d}_{(M)}$ is the least smooth among all the estimators considered; thus, we expect that the random field resulting from $\widehat{d}_{(M)}$ has the largest probability of crossing the fixed level $c_{\alpha/2}$.

The confidence bands in (4.11) are constructed around 1, not around $\widehat{d}_{(K^*)}(\boldsymbol{u})$. That is because $\widehat{d}_{(K^*)}(\boldsymbol{u})$ only accounts for the K^* largest terms in (2.3), therefore, it is a biased estimator of $d(\boldsymbol{u})$ (see Proposition 3.2). It follows that, when the bias is large, confidence bands constructed around $\widehat{d}_{(K^*)}(\boldsymbol{u})$ would be shifted away from the true density $d(\boldsymbol{u})$. However, under H_0, both the bias at a point $\boldsymbol{u}$ and the integrated bias are equal to zero. Hence, (4.11) are reliable confidence bands under H_0.

From a theoretical perspective, a highly non-trivial aspect in the construction of (4.11) is the estimation of the quantile $c_{\alpha/2}$. Probabilities such (4.12) are known in literature as *excursion probabilities* [e.g., 1] and cannot be expressed in closed form. A possible solution for constructing the confidence bands in (4.11), is to proceed by estimating $SE_0\big[\widehat{d}_{(K^*)}(\boldsymbol{u})\big]$ and $c_{\alpha/2}$ via Monte Carlo simulations [see 6, Algorithm 1]. Unfortunately, in the most crucial (astro)physical searches the level of significance required to claim a new discovery is typically in the order of $O(10^{-7})$ [e.g., 20], and thus Monte Carlo simulations may be computationally prohibitive. This is further aggravated when dealing with complex models for which even a single Monte Carlo replicate can be highly expensive in terms of both computational and time resources.

As a valid alternative, for continuous F and G, accurate approximations for (4.12) exist under mild smoothness conditions [e.g., 36]. In our setting, smoothness follows from the fact that the random field in (4.6) and the respective limit can be written as a linear combination of the functions $\frac{T_k(\boldsymbol{u})}{\sqrt{\sum_{k\in\mathcal{K}} T_k^2(\boldsymbol{u})}}$ (see proof of Theorem 4.2 in Appendix B) which are composition of Legendre polynomials, and thus, admit infinite partial derivatives.

An approximation for the left-hand side of (4.12) is

$$\big(1-\Phi(c_{\alpha/2})\big)+\mathcal{L}_1\frac{e^{-\frac{c^2_{\alpha/2}}{2}}}{\pi}+\mathcal{L}_2\frac{e^{-\frac{c^2_{\alpha/2}}{2}}}{\sqrt{2}\pi^{3/2}}+O\left(\exp\left(-\frac{\gamma c^2_{\alpha/2}}{2}\right)\right),\quad \text{as } n\to\infty, \tag{4.13}$$

for some $\gamma > 1$ [37]. In (4.13), $\mathcal{L}_1$ and $\mathcal{L}_2$ are constant known as Lipischitz-Killing curvatures and are typically estimated numerically [e.g., 5]. Notice that the error rate in (4.13) decreases exponentially fast, as $\alpha\to\infty$. Therefore, this solution is particularly amenable to overcome the issues arising when dealing with stringent significance requirements.

As one may expect, the simplicity of the post-selection adjustments in (4.9) and (4.11) comes with a price. Specifically, they can be rather conservative for increasing values of M. However, as shown below for Example I and in the sections to follow, (4.9) still leads to high power even if the sample size is only moderately large. Similarly, (4.11) can be quite accurate and match closely the confidence regions obtained by simulating directly the distribution of (4.6), and repeating the selection process at each replicate.

Example I (continued). The estimate of the likelihood ratio in the right panel of Figure 1 has been obtained by setting $m_1 = 4$ and $m_2 = 3$ and selecting the terms of the respective tensor basis via the AIC rule in (4.7).

The AIC procedure selects 9 terms out of $M = 19$. The post-selection adjusted p-value and 95% confidence regions are shown in the right panel of Figure 2. The confidence contours are constructed by setting equal to one all the values of $\widehat{d}$ contained within the bands in (4.11). Whereas the quantile c_α has been calculated by solving

$$\big(1-\Phi(c_{\alpha/2})\big)+\mathcal{L}_1\frac{e^{-\frac{c^2_{\alpha/2}}{2}}}{\pi}+\mathcal{L}_2\frac{e^{-\frac{c^2_{\alpha/2}}{2}}}{\sqrt{2}\pi^{3/2}}-\frac{\alpha}{2}=0 \tag{4.14}$$

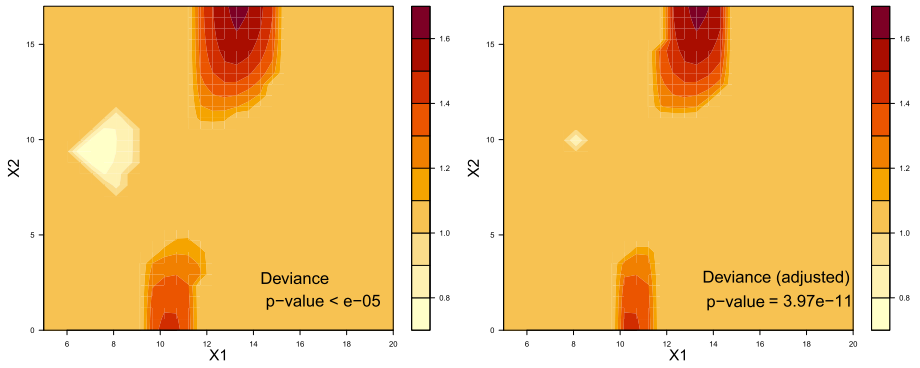

FIG 2. *Simulated and approximated confidence regions for Example I. The left panel corresponds to the (post-selection) confidence regions and deviance p-value obtained via a simulation of size* 10,000. *The right panel shows to the (post-selection adjusted) confidence regions and deviance p-value computed as in* (4.9) *and* (4.11). *Darker shades correspond to significant deviations of the estimated likelihood ratio above one. Lighter shades correspond to significant deviations below one.*

TABLE 1
Simulated probability of type I error and power for Example I considering different sample sizes. The nominal level is chosen to be $\alpha = 0.05$. *Each simulation involves* 10,000 *replicates.*

	$n=500$	$n=1000$	$n=2000$	$n=5000$	$n=7000$	$n=10,000$
Type I error ($\pm$ SE)	.0540 ($\pm$ .0023)	.0500 ($\pm$ .0022)	.0499 ($\pm$ .0022)	.0482 ($\pm$ .0021)	.0508 ($\pm$ .0022)	.04930 ($\pm$ .0022)
Power ($\pm$ SE)	.2157 ($\pm$ .0041)	.4456 ($\pm$ .0050)	.8063 ($\pm$ .0040)	.9995 ($\pm$ .0002)	1 ($\pm$ .0000)	1 ($\pm$ .0000)

and estimating $\mathcal{L}_1$ and $\mathcal{L}_2$ by means of the `R` package `TOHM` [2] as described in Algeri and van Dyk [5]. This approach led to $c_\alpha = 3.5568$. The confidence contours suggest that the most prominent deviations occur in correspondence of the regions $[10,12] \times [0,5]$ and $[12,15] \times [12,17]$. Here, the estimator of $d(\boldsymbol{u})$ shows significant deviations above one and thus we conclude that the postulated model underestimates the truth over these areas. The presence of significant departures of $G_{X_1X_2}$ from $F_{X_1X_2}$ are confirmed by the deviance test (adjusted p-value $\sim 3.97 \cdot 10^{-11}$). The left panel of Figure 2 shows the confidence regions and deviance p-value obtained by means of a Monte Carlo simulation involving 10,000 replicates. The selection procedure has been implemented at each replicate. While more conservative, the confidence regions computed via (4.11) and (4.14), approximate reasonably well those obtained via simulation.

Finally, we investigate the probability of type I error and the power of the deviance test based on (4.9). Table 1 reports the results obtained considering a suite of five simulations, each of size $B = 10,000$, conducted using five different sample sizes. For all n considered, the probability of type I error observed is

approximately the same than the nominal level $\alpha = 0.05$. Whereas, the power increases rapidly with n. For the smallest samples sizes considered, i.e., $n = 500$ and $n = 1000$, the power is rather low ($\sim$ 22% and $\sim$ 45%, respectively). However, it has to be noted that, in our example, the mixture parameter is 0.15; therefore the deviations from the postulated model effectively account for only ~ 75 and ~ 150 data points when $n = 500$ and $n = 1000$, respectively.

5. iGOF-diagnostic analysis

The constructs introduced so far allow us to assess the validity of the postulated model, obtain an estimate of the likelihood ratio test to visualize where and how departures of g from f occur, and construct a data driven correction for the initial model g (equation (3.6)). Unfortunately, however, a visual inspection is only possible when $p \leq 3$. Nevertheless, when $p > 3$, more insights on the sources of mismodeling affecting G can be obtained by conducting an ANOVA-like analysis where random sub-vectors of $\boldsymbol{X}$ are tested individually, from the largest to the smallest.

Without loss of generality, let $\boldsymbol{X}_q = (X_1, \ldots, X_q)$ be the random collecting the first $q < p$ components of $\boldsymbol{X}$. Denote with F_q the true cdf of $\boldsymbol{X}_q$ and let G_q be its postulated cdf. Moreover, assume that the density of G_q can be specified as

$$g_q(\boldsymbol{x}_q) = \prod_{d=1}^{q} g_d(x_d|\boldsymbol{x}_{<d}) \quad \text{for all } d = 1, \ldots, q. \tag{5.1}$$

As in (2.1) and (2.3), we can then express the likelihood ratio of $\boldsymbol{X}_q$ on the q dimensional unit cube via

$$d(\boldsymbol{u}_q) = \sum_{j_1 \geq 0, \ldots, j_q \geq 0} \theta_{j_1 \ldots j_q} T_{j_1 \ldots j_q}(\boldsymbol{u}_q), \quad \boldsymbol{u_q} \in [0,1]^q \tag{5.2}$$

where $\boldsymbol{u}_q = \big(G_1(x_1), \ldots, G_q(x_q|\boldsymbol{x}_{<q})\big)$, and thus $\boldsymbol{u}_q$ is a sub-vector of $\boldsymbol{u} = \boldsymbol{G}_R(\boldsymbol{x})$. Whereas, similarly to (2.2), one can write the tensor basis functions $T_{j_1 \ldots j_q}$ as

$$T_{j_1 \ldots j_q}(\boldsymbol{u}_q) = \prod_{d=1}^{q} T_{j_d}(u_d) = \prod_{d=1}^{p} T_{j_d}(u_d) \quad \text{with } j_d = 0, \text{ for all } d = q+1. \tag{5.3}$$

The last equality follows from the fact that $T_0\big(G(x_d|\boldsymbol{x}_{<d})\big) = 1$ for all $d = 1, \ldots, p$, and thus each $T_{j_1 \ldots j_q}(\boldsymbol{u}_q) = T_{j_1 \ldots j_q 0 \ldots 0}(\boldsymbol{u})$. Consequently, the $\theta_{j_1 \ldots j_q}$ coefficients are equal to $\theta_{j_1 \ldots j_p}$ whenever $j_d = 0$, for all $d = q+1$. As a result, we can easily perform inference for $\boldsymbol{X}_q$ by means of the estimators $\widehat{\theta}_k$ in (3.2), without the need of repeating the estimation procedure.

Specifically, denote with $\mathcal{K}_q$ and $\mathcal{K}^*$ the subsets of $\mathcal{K}$ in (3.1)

$$\mathcal{K}_q := \Big\{ k = \{j_1 \ldots j_p\} \in \mathcal{K} \text{ with } j_d = 0, \text{ for all } d = q+1, \ldots, p \Big\} \tag{5.4}$$

$$\mathcal{K}^* := \left\{ k = \{j_1 \dots j_p\} \in \mathcal{K} \text{ with } (k) \leq K^* \right\} \tag{5.5}$$

of cardinality $|\mathcal{K}_q| = M_q = \prod_{d=1}^{q}(m_d + 1) - 1$ and $|\mathcal{K}^*| = K^*$. Recall that K^* is the value minimizing either the AIC or BIC in (4.7), and thus, $\mathcal{K}^*$ collects all the p-tuple of indexes in $\mathcal{K}$ which have been ultimately selected when constructing the estimator $\widehat{d}_{(K^*)}$ and the deviance statistics $D_{(K^*)}$ in Corollaries 4.1 and 4.2. To test

$$H_0 : G_q = F_q \quad \text{versus} \quad H_1 : G_q \neq F_q \tag{5.6}$$

we consider the test statistics $D_q = n \sum_{k \in \mathcal{K}_q} \widehat{\theta}_k^2$ and proceed as in Theorem 4.1. Whereas, valid post-selection inference can be obtained as in Theorem 5.1.

Theorem 5.1. *As $n \to \infty$, a valid post-selection bound for the p-value to test (5.6) is*

$$p\text{-}value_{q,adj} = P(\chi^2_{M_q} > D_{obs}), \tag{5.7}$$

where D_{obs} being the value of the test statistics

$$D_q^* = n \sum_{k \in \mathcal{K}_q \cap \mathcal{K}^*} \widehat{\theta}_k^2 \tag{5.8}$$

observed, $\mathcal{K}_q$ and $\mathcal{K}^*$ as in (5.4) and (5.5) and $\widehat{\theta}_k$ as in (3.2).

Theorem 5.1 follows directly from (5.2) and (5.3), orthogonality of the T_k functions, and from condition (5.1).

Because of condition (5.1), Theorem 5.1 holds only for random sub-vectors of $\boldsymbol{X}$ whose Rosenblatt transform u_q includes all the conditioning, from the higher to the lower, necessary to recover $g_q(\boldsymbol{x}_q)$. To some extent, this condition can be seen as the iGOF counterpart of the marginality principle advocated by Nelder [24] in the context of ANOVA, and which consists of taking into account the hierarchy of the main effects and interactions in a given model.

Similarly to the ANOVA, Theorem 5.1 allows us to construct an iGOF-diagnostic table to identify the source of mismodeling for a given random vector $\boldsymbol{X}$ and its components. Below we show how this can be done in practice for the case of a 7-dimensional random vector.

Example II. We consider a sample of $n = 5000$ observations from a random vector $\boldsymbol{X} = (X_1, \dots, X_7)$ with components distributed as summarized in the second column of Table 2. Table 3 collects the results obtained by applying Theorem 5.1 to test the validity of the models specified for different sub-vectors of $\boldsymbol{X}$. The overall deviance test is reported in the first row and correctly reject the null model. Similarly, the test in the second row, rejects the hypotheses that the vector (X_1, X_2, X_5, X_6) is modelled correctly, and fails to rejects the model for (X_1, X_2, X_5).

This aspect is particularly important as it highlights that the mismodeling occurs only with respect to the conditional distribution of $X_6|X_1, X_2, X_5$. The tests in the fourth and fifth row show that the vector (X_3, X_4) has been mismodeled and one source of mismodeling is the marginal of X_3. Ultimately, the

TABLE 2
True and postulated model for Example II. The last column highlights where mismodeling occurs.

Variables	True (F)	Hypothesized (G)	Correct
$X_6 \mid X_1, X_2, X_5$	$\text{Laplace}\left[e^{.03x_1+.02x_2+.01x_2^2+.02x_5}, 1\right]$	$\text{Laplace}\left[e^{.03x_1+.02x_2+.02x_5}, 1\right]$	No
X_1, X_2, X_5	$N\left[\begin{pmatrix}10\\15\\11\end{pmatrix}, \begin{pmatrix}4 & .5 & 0\\.5 & 3 & 1\\0 & 1 & 5\end{pmatrix}\right]$	$N\left[\begin{pmatrix}10\\15\\11\end{pmatrix}, \begin{pmatrix}4 & .5 & 0\\.5 & 3 & 1\\0 & 1 & 5\end{pmatrix}\right]$	Yes
$X_4 \mid X_3$	$\text{Exponential}\left(\frac{1}{x_3}\right)$	$\text{Exponential}\left(\frac{1}{x_3}\right)$	Yes
X_3	$\text{Exponential}(1)$	$\text{Exponential}(.9)$	No
X_7	T_3	$\text{Cauchy}(0, 1)$	No

TABLE 3
iGOF-diagnostic table. The third column reports the (post-selection adjusted) deviance p-values in (5.7) with $\boldsymbol{X}_q$ specified as in the first column. The second column corresponds to the degrees of freedom used in the calculation of the p-value, namely, M_q.

$\boldsymbol{X}_q$	df	(Adjusted) p-value
$\boldsymbol{X}$	16383	$< 10^{-130}$
(X_1, X_2, X_5, X_6)	256	$< 10^{-130}$
(X_1, X_2, X_5)	63	1
(X_3, X_4)	15	$8.467 \cdot 10^{-08}$
X_3	3	$1.525 \cdot 10^{-13}$
X_7	3	$2.599 \cdot 10^{-122}$

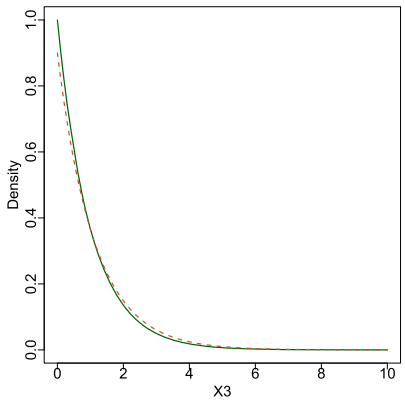

FIG 3. *Comparing the postulated (red dashed line) and the true model (green solid line) of X_3.*

test for X_7 also correctly rejects the null hypothesis of Cauchy distribution. Table 4 collects the results of a simulation obtained by repeating the diagnostic analysis in Table 3 through a simulation of $B = 10,000$ replicates, while considering different sample sizes. Even when the sample size considered is only 500, the most prominent deviations are captured with probability one, whereas, the model for (X_1, X_2, X_5) is never rejected. More issues arise in diagnosing mis-

TABLE 4

Performance of the iGOF-diagnostic analysis for different sample sizes. For values different from zero and one the Monte Carlo errors ($\pm SE$) are also reported. The significance level considered is $\alpha = 0.05$.

$\boldsymbol{X}_q$	**Sample size (n)**					
	500	1000	2000	3000	5000	10,000
$\boldsymbol{X}$	1	1	1	1	1	1
(X_1, X_2, X_5, X_6)	1	1	1	1	1	1
(X_1, X_2, X_5)	0	0	0	0	0	0
(X_3, X_4)	.0069 (±.0008)	.0342 (±.0018)	.2384 (±.0049)	.5939 (±.0049)	.9615 (± .0019)	1
X_3	.2360 (±.0043)	.5560 (±.0050)	.9153 (±.0028)	.9868 (±.0011)	.9999 (± .0001)	1
X_7	1	1	1	1	1	1

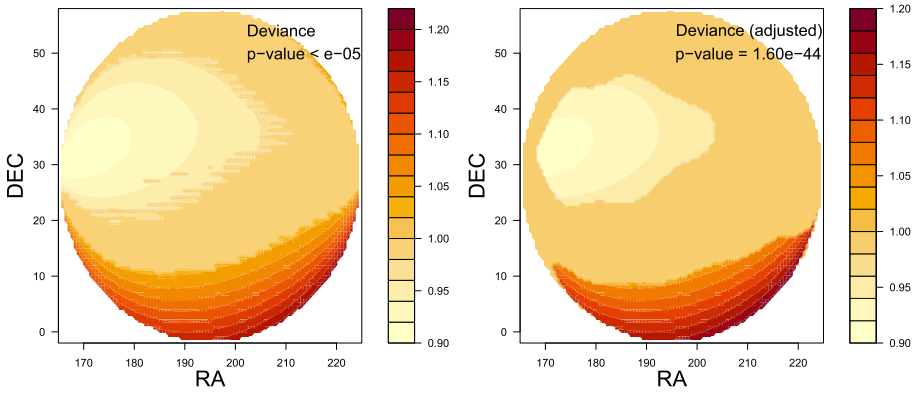

FIG 4. *Simulated and approximated confidence regions for the Fermi LAT simulation. The left panel corresponds to the (post-selection) confidence regions and deviance p-value obtained via a simulation of size* 10,000. *The right panel shows to the (post-selection adjusted) confidence regions and deviance p-value computed as in* (4.9) *and* (4.11). *Darker shades correspond to significant deviations of the estimated likelihood ratio above one. Lighter shades correspond to significant deviations below one.*

modeling of X_3 and, consequently, (X_3, X_4) for smaller samples. For instance, even when $n = 1000$ the power of the procedure in detecting departures of G_{X_3} from F_{X_3} is only $\sim 56\%$ and $\sim 3\%$ for (X_3, X_4). It has to be noted, however, that detecting mismodeling of X_3 is a particularly challenging task. As shown in Figure 3, the postulated and the true pdf of X_3 are very close one-another; this minor differences are further "diluted" when considering the joint distribution of (X_3, X_4), since $X_4|X_3$ is correctly specified. Nevertheless, such minor deviations are detected with high power for larger sample sizes.

6. A diagnosis of background mismodeling

When conducting searches for new phenomena, mismodeling of the background distribution can dramatically compromise the sensitivity of the experiment.

Specifically, overestimating the background can increase the chances of false negatives. Whereas, underestimating the background may lead to claiming false discoveries. To illustrate how iGOF can be used to understand if and how the postulated background model have been misspecified, we consider a simulated observation by the Fermi Large Area Telescope (LAT) [8] obtained with the *gtobssim* package[2] and previously published in Algeri and van Dyk [5]. The simulation includes a realistic representations of the instrumental noise of the detector and present backgrounds.

The region of interest corresponds to a disc in the sky of 30° radius and centered at (195 RA, 28 DEC), where RA and DEC are the coordinates in the sky. Here we assume that, while the cosmic background is known to follow a uniform distribution over the search area, but it is unclear if the instrumental error is effectively negligible, or if it has a prominent effect on the underlying distribution. Therefore, we set $G_{X_1X_2}$ to be the cdf of a uniform distribution with support $\mathcal{X}_1 \times \mathcal{X}_2 = [165, 195] \times [28 - \sqrt{30^2 - (x - 195)^2}, 28 + \sqrt{30^2 - (x - 195)^2}]$ and we proceed by estimating the likelihood ratio via (3.5) over a sample of $n = 68658$ observations. Specifically, we set $m_1 = m_2 = 4$ and we select the components of $\widehat{\boldsymbol{\theta}}$ via the BIC criterion in (4.7). The resulting estimate is

$$\widehat{d}(x_1, x_2) = 1 + 0.022T_1\big(G_1(x_1)\big) - 0.043T_1\big(G_2(x_2|x_1)\big) + 0.041T_2\big(G_2(x_2|x_1)\big). \tag{6.1}$$

In order to assess the significance of the deviations captured by (6.1), we compute simultaneous confidence regions and deviance p-values via (4.11) and (4.9). The results are reported in the right panel of Figure 4, whereas the left panel shows the confidence regions and deviance p-value obtained via simulation. Similarly to what we have observed for Example I (see Figure 2), despite the approximate confidence bands are more conservative, they still allow to capture the main departures from uniformity. Indeed, in both cases, we can see that the prominent deviations of the true underlying model from the postulated uniform distribution occur in proximity of low values of X_2. Whereas, at the center-left of the search area, the uniform model significantly underestimates the model inclusive of the instrumental error. Finally, it follows from (3.6) that an updated model background distribution which accounts for these deviations can be constructed as in equation (3.6) by simply multiplying the uniform pdf by the estimated likelihood ratio in (6.1).

7. Extensions to the discrete case

The methods discussed so far focus on the case where F and G are continuous. However, extensions to the discrete setting can be derived by rewriting the expansion in (2.3) through an orthonormal set of functions suitable to model discrete data. This can be done, for instance, by means of the so-called "LP[3] score

[2] http://fermi.gsfc.nasa.gov/ssc/data/analysis/software

[3] In the LP acronym, the letter L typically denotes nonparametric methods based on quantiles, whereas P stands for polynomials [22, Supp S1].

functions", recently introduced in literature [e.g., 22] and which can be seen as a generalization of the Legendre polynomials valid in both the continuous and discrete setting.

Specifically, when $p = 1$, a complete orthonormal basis of LP score functions in $L^2(G)$ can be specified by letting the first component to be $T_0\big[G(x)\big] = 1$. Subsequent components $\{T_j\big[G(x)\big]\}_{j>0}$ are obtained by Gram-Schimidt orthonormalization of powers of

$$T_1\big[G(x)\big] = \frac{G_{\text{mid}}(x) - E[G_{\text{mid}}(x)]}{\sqrt{V(G_{\text{mid}}(x))}} = \frac{G(x) - 0.5p_G(x) - 0.5}{\sqrt{[1 - \sum_{x\in\mathcal{X}} p_G^3(x)]/12}}, \tag{7.1}$$

where $G_{\text{mid}}(x) = G(x) - 0.5p_G(x)$ is the *mid-distribution function*, which it has been shown in Parzen [28] to have mean 0.5 and variance $[1 - \sum_{x\in\mathcal{X}} p_G^3(x)]/12$, with $\mathcal{X}$ being the set of distinct points in the support of X and $p_G(x) = P(X = x)$ if $X \sim G$. Therefore, $T_1\big[G(x)\big]$ is the standardized mid-distribution and orthonormality of the $T_j\big[G(x)\big]$ functions in $L^2(G)$ follows by the first equality in (7.1) and by Gram-Schmidt process. Notice that, for continuous X, $G_{\text{mid}}(x) = G(x)$ and $\sum_{x\in\mathcal{X}} p_G^3(x) = 0$, consequently, the LP score functions reduce to normalized shifted Legendre polynomials. The latter are effectively the result of a Gram-Schmidt orthonormalization applied to powers of $G(x)$. Whereas, the LP score functions are obtained by orthonormalizing powers of the standardized mid-distribution function with respect to the measure G.

Recall that, in our context, the cdfs G_d, $d = 1, \ldots, p$ are the conditional and marginal distribution functions specified in the Rosenblatt's transform $\boldsymbol{G}_R(\boldsymbol{x})$. Hence, an orthonormal basis in $L^2\big(G_d\big)$ is given by the set of functions $T_{j_d}\big[G_d(x_d|\boldsymbol{x}_{<d})\big]$, $j_d \geq 0$ with $T_0\big[G_d(x_d|\boldsymbol{x}_{<d})\big] = 1$ and subsequent components

$$T_{j_d}\big[G_d(x_d|\boldsymbol{x}_{<d})\big] = \frac{\mathring{\mathrm{T}}_{j_d}\big[G_d(x_d|\boldsymbol{x}_{<d})\big]}{\big|\big|\mathring{\mathrm{T}}_{j_d}\big[G_d(x_d|\boldsymbol{x}_{<d})\big]\big|\big|_{G_d}}, \quad \text{for all } j_p \geq 1, \text{ where} \tag{7.2}$$

$$\mathring{\mathrm{T}}_{j_d}\big[G_d(x_d|\boldsymbol{x}_{<d})\big] = T_1^{j_d}\big[G_d(x_d|\boldsymbol{x}_{<d})\big] \tag{7.3}$$

$$- \sum_{k=1}^{j_p-1} \big\langle T_1^{j_d}\big[G_d(x_d|\boldsymbol{x}_{<d})\big], T_k\big[G_d(x_d|\boldsymbol{x}_{<d})\big]\big\rangle_{G_d} T_k\big[G_d(x_d|\boldsymbol{x}_{<d})\big], \tag{7.4}$$

$$\text{with } \big\langle T_1^{j_d}\big[G_d(x_d|\boldsymbol{x}_{<d})\big], T_{k_d}\big[G_d(x_d|\boldsymbol{x}_{<d})\big]\big\rangle_{G_d} = \tag{7.5}$$

$$\int T_1^{j_d}\big[G_d(x_d|\boldsymbol{x}_{<d})\big] T_{k_d}\big[G_d(x_d|\boldsymbol{x}_{<d})\big] \mathrm{d}G_d(x_d|\boldsymbol{x}_{<d}) \tag{7.6}$$

and $||\cdot||_{G_d} = \sqrt{<\cdot,\cdot>_{G_d}}$.

When $p > 1$, a suitable tensor basis in $L^2(G)$ can then be constructed as in (2.2). $\{T_{j_1,\ldots,j_p}(\boldsymbol{u})\}_{j_1\ldots j_p\geq 0}$. Orthonormality of the $T_{j_1,\ldots,j_p}(\boldsymbol{u})$ score functions can be verified directly as shown in Appendix D.

8. Discussion

This work proposes an informative approach to goodness-of-fit which connects exploratory and confirmatory data analysis to study multivariate distributions. By transforming the likelihood ratio on the unit cube, confidence regions can be constructed as in Corollary 4.2 to identify regions of the support where significant deviations occurs. While this approach is practical only for problems in at most three dimensions, in more dimensions, a detailed diagnosis of mismodeling can be achieved by means of the iGOF-diagnostic analysis proposed in Section 5. These tools can be used to directly address Q1 in Section 1. For instance, given the panacea of theories available on the nature of dark matter, experimentalists aiming to detect it often face the dilemma of selecting which of the tens of theoretical models (mainly non-nested) available should be tested [e.g., 34]. If one was to test it using the procedure discussed in this paper, even when a given model is rejected, it is possible to gain further insight on the shape of the departure of the true data distribution and the null model and ultimately use such information to "rule out" other models which would be inconsistent with such deviation.

Moreover, as we aimed for when formulating Q2 in Section 1, the true probability function of the data can be estimated semi-parametrically via (3.6), while assessing the validity of the model postulated by the scientists. Interestingly, the resulting estimate incorporates the knowledge carried by the hypothesized model and thus, it provides a data-driven update for it in the direction of the true distribution of the data.

Despite the usefulness of the methods presented here in applied settings, and in the physical sciences in particular (e.g., Section 6), they are not exempt from limitations. For instance, several problems in physics and astronomy, often involve no more than 8 or 10 dimensions and/or can be reduced to 2D planes [e.g., 7]. In this context, choosing m_d equal to 3 or 4 for all $d = 1, \dots, p$, is often sufficient to avoid overfitting and, eventually, lack of power by implementing adequate model selection strategies and for sufficiently large samples (see Sections 4 and 5). In more dimensions, however, the method suffers from the curse of dimensionality [e.g. 17], as the size of the LP tensor basis increases exponentially fast with p. In this context, a regularized solution could be particularly valuable [see for instance 35]. Alternatively, if the interest is merely in detecting signals without assuming a specific background model, a data-driven solution for high-dimensional data has been recently proposed by Chakravarti et al. [13].

Furthermore, the unitary representation of the likelihood ratio in (2.1) relies on the Rosenblatt transform and which can lead to different configurations of $\boldsymbol{U}$ and, potentially, different estimators. While this aspect would require adequate treatment on its own, it is worth noting that this problem is essentially the same arising in the context of vine copulas [e.g., 23] and for which adequate model selection procedures exists [e.g., 15, 27].

Finally, the inferential procedures presented here extend classical smooth tests to the multivariate setting and allow us to visualize graphically the departure of G from F and study their substructures. Despite this article focuses on

simple null hypothesis, that is, the postulated model is assumed to be fully specified, classical results on smooth tests [e.g., 38, Sec 4.2.2.3 and 5.2.2.3] can be used to derive asymptotic tests in the parametric setting. Unfortunately, however, the asymptotic approximations are known to be rather slow in this case. Therefore, in practical applications, when G depends on unknown parameters it is recommended to perform inference by means of the parametric bootstrap and which has been shown by Babu and Rao [9] to be consistent also in the multivariate setting.

Appendix A: Notation

TABLE 5
A summary of the main notation used throughout the paper.

Symbol	Description
$\boldsymbol{X} = (X_1, \dots, X_p)$	Random vector of components $X_d, d = 1, \dots, p$
$\mathcal{X}$	Support of $\boldsymbol{X}$
F, f	True cdf and density of $\boldsymbol{X}$
G, g	Postulated cdf and density of $\boldsymbol{X}$
G_d, G_d^{-1}, g_d	Conditional cdf, quantile function and density of X_d
$\boldsymbol{U} = \boldsymbol{G}_R(\boldsymbol{X})$	Rosenblatt's transform of $\boldsymbol{X}$
$\boldsymbol{G}_R^{-1}(\boldsymbol{U}) = \boldsymbol{X}$	Inverse of the Rosenblatt's transform
$\boldsymbol{x} = (x_1, \dots, x_p) = \boldsymbol{G}_R^{-1}(\boldsymbol{u})$	Realization of $\boldsymbol{X}$ with components $x_d = G_d^{-1}(u_d)$
$\boldsymbol{u} = (u_1, \dots, u_p) = \boldsymbol{G}_R(\boldsymbol{x})$	Realization of $\boldsymbol{U}$, with components $u_d = G_d(x_d \vert \boldsymbol{x}_{<d})$
$\boldsymbol{x}_{<d} = (x_1, \dots, x_{d-1})$	First $d-1$ components of $\boldsymbol{x}$
$\boldsymbol{u}_{<d} = (u_1, \dots, u_{d-1})$	First $d-1$ components of $\boldsymbol{u}$
$d(\boldsymbol{u}), \widehat{d}(\boldsymbol{u})$	Likelihood ratio and its estimate
$T_{j_d}\big[G_d(x_d \vert \boldsymbol{x}_{<d})\big] = T_{j_d}(u_d)$	j_d-th normalized shifted Legendre polynomial in $L^2(G_d)$ and $[0,1]$
$T_k\big[\boldsymbol{G}_R(\boldsymbol{x})\big] = T_k(\boldsymbol{u})$	Tensor product of T_{j_d} functions with $k \in \mathcal{K}$, in $L^2(G)$ and $[0,1]^p$
$\mathcal{K}$	sets of p-tuple $\{j_1 \dots j_p\}$, $\sum_d j_d \neq 0$, $j_d = 0, \dots, m_d$
$\vert\mathcal{K}\vert = M$	Cardinality of $\mathcal{K}$
$\boldsymbol{T}(\boldsymbol{u}) = \boldsymbol{T}\big[\boldsymbol{G}_R(\boldsymbol{x})\big]$	$M \times 1$ vector of components $T_k(\boldsymbol{u}) = T_k\big[\boldsymbol{G}_R(\boldsymbol{x})\big]$
$\boldsymbol{\theta}$	$M \times 1$ vector collecting the coefficients θ_k
$\widehat{\boldsymbol{\theta}}$	$M \times 1$ vector collecting the estimates $\widehat{\theta}_k$
D	Deviance statistics

Appendix B: Proofs

Proof of Proposition 3.1. First, to show that $d\big(\boldsymbol{u}\big)$ is the density of the random vector $\boldsymbol{U} = (U_1, \dots, U_p)$, consider

$$P(\boldsymbol{U} \leq \boldsymbol{u}) = P(U_1 \leq u_1, \dots, U_p \leq u_p) \tag{B.1}$$

$$= P\big(G_1(X_1) \leq u_1, \ldots, G_p(X_p|\boldsymbol{x}_{<p}) \leq u_p\big) \tag{B.2}$$

$$= P\big(X_1 \leq G_1^{-1}(u_1), \ldots, X_p \leq G_p^{-1}(u_p|\boldsymbol{x}_{<p})\big) \tag{B.3}$$

$$= F\big(G_1^{-1}(u_1), \ldots, G_p^{-1}(u_p|\boldsymbol{x}_{<p})\big). \tag{B.4}$$

By applying the chain rule to (B.4) we obtain,

$$\frac{\partial^p P(\boldsymbol{U} \leq \boldsymbol{u})}{\partial u_1 \ldots \partial u_p} = \frac{\partial^p F\big(G_1^{-1}(u_1), \ldots, \partial G_p^{-1}(u_p|\boldsymbol{x}_{<p})\big)}{\partial G_1^{-1}(u_1) \ldots \partial G_p^{-1}(u_p|\boldsymbol{x}_{<p})} \frac{\partial G_1^{-1}(u_1), \ldots, G_p^{-1}(u_p|\boldsymbol{x}_{<p})}{\partial u_1 \ldots \partial u_p} \tag{B.5}$$

$$= \frac{g\big(G_1^{-1}(u_1), \ldots, G_p^{-1}(u_p|\boldsymbol{x}_{<p})\big)}{f\big(G_1^{-1}(u_1), \ldots, G_p^{-1}(u_p|\boldsymbol{x}_{<p})\big)} = \frac{f\big(\boldsymbol{G}_R^{-1}(\boldsymbol{u})\big)}{g\big(\boldsymbol{G}_R^{-1}(\boldsymbol{u})\big)} = d\big(\boldsymbol{u}\big). \tag{B.6}$$

Let's now derive the mean and variance of our $\widehat{\theta}_k$ estimators. Let $\boldsymbol{U}_i$, $i = 1, \ldots, n$, be the random vectors of which the $\boldsymbol{u}_i$ are realizations,

$$E[\widehat{\theta}_k] = E\Big[\frac{1}{n}\sum_{i=1}^{n} T_k(\boldsymbol{U}_i)\Big] = E\Big[T_k(\boldsymbol{U})\Big] = \int_{[0,1]^p} T_k(\boldsymbol{u}) d\big(\boldsymbol{u}\big) \mathrm{d}\boldsymbol{u} = \theta_k, \quad \text{for all } k \in \mathcal{K}, \tag{B.7}$$

where the third equality follows from the fact that $d(\boldsymbol{u})$ is the pdf of $\boldsymbol{U}$. Moreover, if $F \equiv G$, then $d\big(\boldsymbol{u}\big) = 1$ and thus, by orthogonality of each $T_k(\boldsymbol{u})$, $k \in \mathcal{K}$ with respect to one, we have that $\theta_k = 0$ for all $k \in \mathcal{K}$. Similarly,

$$V[\widehat{\theta}_k] = V\Big(\frac{1}{n}\sum_{i=1}^{n} T_k(\boldsymbol{U}_i)\Big) = \frac{1}{n}\int_{[0,1]^p} \Big(T_k(\boldsymbol{u}) - \theta_k\Big)^2 d\big(\boldsymbol{u}\big) \mathrm{d}\boldsymbol{u} = \frac{\sigma_k^2}{n}, \quad \text{for all } k \in \mathcal{K}. \tag{B.8}$$

Furthermore, if $F \equiv G$, then for all $k \in \mathcal{K}$, and by orthonormality of the $T_k(\boldsymbol{u})$ we have $\sigma_k^2 = 1$.

The covariance between $\widehat{\theta}_k$ and $\widehat{\theta}_h$ is

$$\mathrm{Cov}\big(\widehat{\theta}_k, \widehat{\theta}_h\big) = \mathrm{Cov}\Big(\frac{1}{n}\sum_{i=1}^{n} T_k(\boldsymbol{U}_i), \frac{1}{n}\sum_{i=1}^{n} T_h(\boldsymbol{U}_i)\Big) \tag{B.9}$$

$$= \frac{1}{n}\int_{[0,1]^p} \big(T_k(\boldsymbol{u}) - \theta_k\big)\big(T_h(\boldsymbol{u}) - \theta_h\big) d\big(\boldsymbol{u}\big) \mathrm{d}\boldsymbol{u} = \frac{\sigma_{k,h}}{n}, \quad \text{for all } k, h \in \mathcal{K}. \tag{B.10}$$

In this case, if $F \equiv G$, then for all $k, h \in \mathcal{K}$, and by orthogonality of the $T_k(\boldsymbol{u})$ we have $\sigma_{k,h} = 0$.

Finally, the estimator in (3.5) is obtained by truncating the series in (2.3) at $m_1, \ldots, m_d$, plugging-in the estimates in (3.2) and noticing that, since $T_0(u_d) = 1$ for all $d = 1, \ldots, p$, we have that $T_{0\ldots0}(\boldsymbol{u}) = \prod_{d=1}^{p} T_0(u_d) = 1$ for all $\boldsymbol{u} \in [0,1]^p$. Consequently,

$$\theta_{0\ldots0} = \int_{[0,1]^p} T_{0\ldots0}(\boldsymbol{u}) d(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} = \int_{[0,1]^p} d(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} = 1. \tag{B.11}$$

□

Proof of Proposition 3.2. Denote with $\bar{\mathcal{K}}$ the complement of $\mathcal{K} \cup \{0, \dots, 0\}$, i.e.,

$$\bar{\mathcal{K}} := \left\{ \{j_1, \dots, j_p\}, j_d = m_d, \dots, \infty, \text{ for all } d = 1, \dots, p, \sum_{d=1}^{p} j_d \neq 0 \right\}, \quad \text{(B.12)}$$

and let $\Omega = \{0, \dots, 0\} \cup \mathcal{K} \cup \bar{\mathcal{K}}$. Hence, we can rewrite the expansion in (2.3) as

$$d(\boldsymbol{u}) = \sum_{k \in \Omega} \theta_k T_k(\boldsymbol{u}), \qquad \boldsymbol{u} \in [0,1]^p. \quad \text{(B.13)}$$

For each $\boldsymbol{u} \in [0,1]^p$, let the bias of (3.5) at $\boldsymbol{u}$ be

$$\text{Bias}\big(\widehat{d}(\boldsymbol{u})\big) = E\big[\widehat{d}(\boldsymbol{u})\big] - d(\boldsymbol{u}) = 1 + \sum_{k \in \mathcal{K}} \theta_k T_k(\boldsymbol{u}) - d(\boldsymbol{u}) \quad \text{(B.14)}$$

$$= 1 + \sum_{k \in \mathcal{K}} \theta_k T_k(\boldsymbol{u}) - \sum_{k \in \mathcal{K} \cup \{0 \dots 0\}} \theta_k T_k(\boldsymbol{u}) - \sum_{k \in \bar{\mathcal{K}}} \theta_k T_k(\boldsymbol{u}) = \sum_{k \in \bar{\mathcal{K}}} \theta_k T_k(\boldsymbol{u}), \quad \text{(B.15)}$$

where the second equality follows in (B.14) from (3.5) and (3.3), the first equality in (B.15) follows from (B.13) and the second equality from (B.11).

Therefore, by pluggin-in (B.15) in (3.7) we have

$$ISB = \int_{[0,1]^p} \Big(\sum_{k \in \bar{\mathcal{K}}} \theta_k T_k(\boldsymbol{u}) \Big)^2 \mathrm{d}\boldsymbol{u} = \sum_{k,h \in \bar{\mathcal{K}}} \theta_k \theta_h \int_{[0,1]^p} T_k(\boldsymbol{u}) T_h(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} \quad \text{(B.16)}$$

$$= \sum_{k \in \bar{\mathcal{K}}} \theta_k^2 = \sum_{k \in \Omega} \theta_k^2 - \sum_{k \in \mathcal{K}} \theta_k^2 - 1 = \int_{[0,1]^p} \big(d(\boldsymbol{u}) - 1 \big)^2 \mathrm{d}\boldsymbol{u} - \sum_{k \in \mathcal{K}} \theta_k^2 \quad \text{(B.17)}$$

$$= \int_{[0,1]^p} \left(\frac{f\big(\boldsymbol{G}_R^{-1}(\boldsymbol{u})\big) - g\big(\boldsymbol{G}_R^{-1}(\boldsymbol{u})\big)}{g\big(\boldsymbol{G}_R^{-1}(\boldsymbol{u})\big)} \right)^2 \mathrm{d}\boldsymbol{u} - \sum_{k \in \mathcal{K}} \theta_k^2, \quad \text{(B.18)}$$

where the first equality in (B.17) follows from orthonormality of the functions $T_k(\boldsymbol{u})$. The third equivalence in (B.17) follows from Parseval's identity. Finally, (B.18) follows from (2.1). □

Proof of Theorem 4.1. By the multivariate central limit theorem, we have that, under H_0, $\sqrt{n}\widehat{\boldsymbol{\theta}}$ converges to a standard multivariate normal, as $n \to \infty$. Whereas, (4.5) follows from (4.3). □

Proof of Corollary 4.2. Let $\psi_k(\boldsymbol{u}) = \frac{T_k(\boldsymbol{u})}{\sqrt{\sum_{k \in \mathcal{K}} T_k^2(\boldsymbol{u})}}$, with $\sum_{k \in \mathcal{K}} \psi_k^2(\boldsymbol{u}) = 1$ for all $\boldsymbol{u} \in [0,1]^2$, and write

$$\frac{\widehat{d}(\boldsymbol{u}) - 1}{\sqrt{\frac{1}{n}\boldsymbol{T}(\boldsymbol{u})'\boldsymbol{T}(\boldsymbol{u})}} = \sqrt{n} \sum_{k \in \mathcal{K}} \widehat{\theta}_k \psi_k(\boldsymbol{u}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{k \in \mathcal{K}} T_k(U_i) \psi_k(\boldsymbol{u}). \quad \text{(B.19)}$$

Under H_0, $E[T_k(\boldsymbol{U}_i)] = 0$ and $V[T_k(\boldsymbol{U}_i)] = 1$, for all $i = 1, \ldots, n$. Moreover, each $\psi_k(\boldsymbol{u})$ is differentiable on the unit cube and therefor Lipschitz continuous, i.e., for each $\psi_k(\boldsymbol{u})$ there exist a constant L_k such that

$$|\psi_k(\boldsymbol{u}) - \psi_k(\boldsymbol{v})| \leq L_k ||\boldsymbol{u} - \boldsymbol{v}||. \tag{B.20}$$

Consequently,

$$\left| \sum_{k \in \mathcal{K}} T_k(\boldsymbol{U}_i)\psi_k(\boldsymbol{u}) - \sum_{k \in \mathcal{K}} T_k(\boldsymbol{U}_i)\psi_k(\boldsymbol{v}) \right| \leq \sum_{k \in \mathcal{K}} |T_k(\boldsymbol{U}_i)||\psi_k(\boldsymbol{u}) - \psi_k(\boldsymbol{v})| \leq M||\boldsymbol{u} - \boldsymbol{v}|| \tag{B.21}$$

where $M = \sum_{k \in \mathcal{K}} |T_k(\boldsymbol{U}_i)|$. Moreover, by compactness of $[0,1]^p$, the ϵ-entropy integral is finite (see Dudley [16, Sec. 7]). Hence, the result in (4.6) follows from Jain and Marcus [18, Theorem 1]. □

Proof of Corollary 4.1. To prove Corollary 4.1, we need to show that there exists a value D_α for which

$$P\Big(\bigcup_{K=1}^{M} \Big[\{D_{(K)} > D_\alpha\} \cap \{K^* = K\} \Big] \Big| H_0 \Big) \leq \alpha, \tag{B.22}$$

with $D_{(K)} = \sum_{(k)=1}^{K} \widehat{\theta}^2_{(k)}$ and the event $\{K^* = K\}$ indicates that K is the value selected by the AIC or BIC procedure in (4.7). The left-hand-side of (B.22) is the probability that at least one model leads to incorrectly reject H_0 when selected (while accounting for the selection probability). Hence,

$$P\Big(\bigcup_{K=1}^{M} \Big[\{D_{(K)} > D_\alpha\} \cap \{K^* = K\} \Big] \Big| H_0 \Big) \tag{B.23}$$

$$= \sum_{K=1}^{M} P\big(D_{(K)} > D_\alpha \big| K^* = K, H_0\big) P\big(K^* = K \big| H_0\big) \tag{B.24}$$

$$\leq \sum_{K=1}^{M} P\big(D_{(M)} > D_\alpha \big| K^* = K, H_0\big) P\big(K^* = K \big| H_0\big) \tag{B.25}$$

$$= P\big(D_{(M)} > D_\alpha \big| H_0\big) \xrightarrow[n \to \infty]{} P\big(\chi^2_M > D_\alpha\big) = \alpha \tag{B.26}$$

with $D_\alpha = \chi^2_{M,1-\alpha}$. Equation (B.24) follows from the fact that, for each $K \neq K'$,

$$P\big(\{K^* = K\} \cap \{K^* = K'\} \big| H_0\big) = 0 \tag{B.27}$$

since one and only one model is selected and thus all the intersection events of the type

$$\Big[\{D_{(K)} > D_\alpha\} \cap \{K^* = K\}\Big] \cap \Big[\{D_{(K')} > D_\alpha\} \cap \{K^* = K'\}\Big]$$

have zero probability for all $K' \neq K$. The inequality in (B.25) follows from the fact that $D_{(K)} \leq D_{(M)}$, for all K. Finally, the equality in (B.26) follows from the law of total probabilities, whereas, the limit in (B.26) follows from Theorem 4.1, with $\chi^2_{M,1-\alpha}$ denoting the quantile of order $1-\alpha$ of a χ^2_M distributed random variable. □

Proof of Corollary 4.2. Similarly to the proof of Corollary 4.1, we need to show that there exists a value $c_{\alpha/2}$ for which

$$P\left(\bigcup_{K=1}^{M}\left[\left\{\sup_{\boldsymbol{u}}\left\{\frac{\widehat{d}_{(K)}(\boldsymbol{u})-1}{SE_0\big(\widehat{d}_{(K)}(\boldsymbol{u})\big)}\right\}\leq c_{\alpha/2}\right\}\cap\big\{K^*=K\big\}\right]\Big|H_0\right)\geq 1-\frac{\alpha}{2}, \quad \text{(B.28)}$$

where $SE_0\big(\widehat{d}_{(K)}(\boldsymbol{u})\big)$ denotes the standard error of (4.8) under H_0. We have

$$P\left(\bigcup_{K=1}^{M}\left[\left\{\sup_{\boldsymbol{u}}\left\{\frac{\widehat{d}_{(K)}(\boldsymbol{u})-1}{SE_0\big(\widehat{d}_{(K)}(\boldsymbol{u})\big)}\right\}\leq c_{\alpha/2}\right\}\cap\big\{K^*=K\big\}\right]\Big|H_0\right) \quad \text{(B.29)}$$

$$=\sum_{K=1}^{M}P\left(\sup_{\boldsymbol{u}}\left\{\frac{\widehat{d}_{(K)}(\boldsymbol{u})-1}{SE_0\big(\widehat{d}_{(K)}(\boldsymbol{u})\big)}\right\}\leq c_{\alpha/2}\Big|K^*=K,H_0\right)P\big(K^*=K\big|H_0\big) \quad \text{(B.30)}$$

$$\geq\sum_{K=1}^{M}P\left(\sup_{\boldsymbol{u}}\left\{\frac{\widehat{d}_{(M)}(\boldsymbol{u})-1}{SE_0\big(\widehat{d}_{(M)}(\boldsymbol{u})\big)}\right\}\leq c_{\alpha/2}\Big|K^*=K,H_0\right)P\big(K^*=K\big|H_0\big) \quad \text{(B.31)}$$

$$=P\left(\sup_{\boldsymbol{u}}\left\{\frac{\widehat{d}_{(M)}(\boldsymbol{u})-1}{SE_0\big(\widehat{d}_{(M)}(\boldsymbol{u})\big)}\right\}\leq c_{\alpha/2}\Big|H_0\right) \quad \text{(B.32)}$$

$$=1-\frac{\alpha}{2} \quad \text{(B.33)}$$

where (B.30) follows from (B.27), (B.31) follows from (4.10). Whereas (B.32) follows from the law of total probabilities and (B.33) from (4.12). □

Appendix C: Likelihood Ratio Estimate for Example I

The first few normalized shifted Legendre polynomials over $[0,1]$ are

$$\begin{aligned}
T_0(u) &= 1\\
T_1(u) &= \sqrt{12}(u-0.5)\\
T_2(u) &= \sqrt{5}(6u^2-6u+1)\\
T_3(u) &= \sqrt{7}(20u^3-30u^2+12u-1)\\
T_4(u) &= 3(70u^4-140u^3+90u^2-20u+1), \quad \text{etc.}
\end{aligned}$$

The likelihood ratio estimate for Example I (Figure 1, right panel) can then be specified as

$$\begin{aligned}\widehat{d}(x_1,x_2) =& 1 - 0.067T_2\big[G_1(x_1)\big] + 0.067T_2\big[G_2(x_2|x_1)\big] + 0.0264_3\big[G_2(x_2|x_1)\big] \\ &+ 0.055T_{11}\big[G_1(x_1), G_2(x_2|x_1)\big] + 0.022T_{13}\big[G_1(x_1), G_2(x_2|x_1)\big] \\ &- 0.029T_{21}\big[G_1(x_1), G_2(x_2|x_1)\big] - 0.035T_{22}\big[G_1(x_1), G_2(x_2|x_1)\big] \\ &- 0.033T_{31}\big[G_1(x_1), G_2(x_2|x_1)\big] - 0.037T_{33}\big[G_1(x_1), G_2(x_2|x_1)\big]\end{aligned} \tag{C.1}$$

In (C.1), $T_{j_1}\big[G_1(x_1)\big]$ and $T_{j_2}\big[G_2(x_2|x_1)\big]$ denote the j_1^{th} and the j_2^{th}, normalized shifted Legendre polynomial evaluated at $G_1(x_1)$ and $G_2(x_2|x_1)$, respectively. Whereas, $T_{j_1 j_2}\big[G_1(x_1), G_2(x_2|x_1)\big] = T_{j_1}\big[G_1(x_1)\big]T_{j_2}\big[G_2(x_2|x_1)\big]$.

For instance,

$$\begin{aligned}T_{13}\big[G_1(x_1), G_2(x_2|x_1)\big] =& \sqrt{12}\big(G_1(x_1) - 0.5\big)\sqrt{7}\big(20G_2(x_2|x_1)^3 - 30G_2(x_2|x_1)^2 \\ &+ 12G_2(x_2|x_1) - 1\big).\end{aligned}$$

Appendix D: Properties of the LP tensor basis

For $p = 1$, the set $\big\{T_1^j\big[G(x)\big]\big\}_{j\geq 0}$, constructed as in Section 7 of the manuscript, spans $L^2(G)$. While this result is obvious in the continuous case, for G discrete we can write the least squares problem as

$$\sum_{x\in\mathcal{X}}\Big\{h\big[G(X)\big] - \sum_{j=0}^{J}\beta_j T_1^j\big[G(X)\big]\Big\}^2 p_G(x) = \tag{D.1}$$

$$\sum_{x\in\mathcal{X}}\Big\{h\big[G(X)\big] - \sum_{j=0}^{J}\beta_j\big(G(x) - 0.5p_G(x) - 0.5\big)^j\Big\}^2 p_G(x). \tag{D.2}$$

and has solution $\beta_j = \sum_{x\in\mathcal{X}}\big(G(x) - 0.5p_G(x) - 0.5\big)^j h\big[G(X)\big]p_G(x)$ for all $j = 0,\dots,J$. Thus, letting $|\mathcal{X}|$ be the number of distinct points in the support of X, one can make (D.1) arbitrary small by choosing J sufficiently close to $|\mathcal{X}|-1$. The orthonormal set $\{T_j\big[G(x)\big]\}_{j\geq 0}$ is then obtained via orthonormalization of the functions $T_1^j\big[G(x)\big]$. And since none of the resulting $T_j\big[G(x)\big]$ is identically equal to zero, they are all linearly independent. Hence, $\{T_j\big[G(x)\big]\}_{j\geq 0}$ is a basis in $L^2(G)$.

For $p > 1$, we consider the well known fact that, given two orthonormal bases $\{\psi_j\}$, $\{\phi_k\}$ for the Hilbert spaces $\mathcal{H}_1$, $\mathcal{H}_2$, then $\{\psi_j \otimes \phi_k\}$ is an orthonormal basis for $\mathcal{H}_1 \otimes \mathcal{H}_2$ [Proposition 2, 32, p.50]. By applying this result in our setting, we have that, $\{T_{j_1}[G_1(x_1)] \otimes T_{j_2}[G_2(x_2|x_1)]\}$ forms an orthonormal basis for $L^2\big(G_1 \times G_2\big)$. Choosing for instance $G_1 = G_{X_1}$ and $G_2 = G_{X_2|X_1}$, it is clear that $G_{X_1} \times G_{X_2|X_1} = G_{X_1X_2}$. One can then proceed analogously for $d > 2$ to show that $\{\bigotimes_{d=1}^{p} T_{j_d}\}$ is an orthonormal basis for $L^2\big(G_1 \times \cdots \times G_d\big) \equiv L^2\big(G\big)$.

Orthonormality of the score functions $T_{j_1\dots j_p}\big(\boldsymbol{G}_R(\boldsymbol{x})\big)$ can be verified directly. For $p=2$,

$$\int T_{j_1 j_2}\big(\boldsymbol{G}_R(\boldsymbol{x})\big)T_{k_1 k_2}\big(\boldsymbol{G}_R(\boldsymbol{x})\big)\mathrm{d}G(\boldsymbol{x}) \tag{D.3}$$

$$=\iint T_{j_1}\big[G_1(x_1)\big]T_{k_1}\big[G_1(x_1)\big]T_{j_2}\big[G_2(x_2|x_1)\big]T_{k_2}\big[G_2(x_2|x_1)\big]\mathrm{d}G_2(x_2|x_1)\mathrm{d}G_1(x_1) \tag{D.4}$$

$$=\int T_{j_1}\big[G_1(x_1)\big]T_{k_1}\big[G_1(x_1)\big]\left[\int T_{j_2}\big[G_2(x_2|x_1)T_{k_2}\big[G_2(x_2|x_1)\big]\mathrm{d}G_2(x_2|x_1)\right]\mathrm{d}G_1(x_1) \tag{D.5}$$

$$=\delta_{j_2k_2}\int T_{j_1}\big[G_1(x_1)\big]T_{k_1}\big[G_1(x_1)\big]\mathrm{d}G_1(x_1)=\delta_{j_2k_2}\delta_{j_1k_1} \tag{D.6}$$

with $\delta_{j_dk_d}=0$ if $j_d=k_d$ and $\delta_{j_dk_d}=1$ otherwise, $d=1,2$. The equivalence in (D.5) follows from Fubini-Tonelli's theorem and the last two equalities follow from orthonormality of the T_{j_d} score functions in $L^2(G_d)$, $d=1,2$. Similarly, orthonormality of $T_{j_1j_2}\big(\boldsymbol{u}\big)$ in $L^2[0,1]^2$ follows by substituting $u_1=G_1(x_1)$ and $u_2=G_2(x_2|x_1)$ in (D.3)-(D.6). One can proceed in a similar manner to show orthonormality for any $p>2$.

Acknowledgments

The author thanks sincerely G. Jogesh Babu and two anonymous referees for the useful suggestions and comments. Their valuable feedback has led to a substantial improvement of the quality and clarity of the manuscript.

Supplementary Material

Codes and data

(doi: 10.1214/21-EJS1926SUPP; .zip). The folder ***Codes_and_data*** collects the data used in Section 6 and the codes used for the analyses in Figures 1, 2, and 4. The interested reader is directed to the supplementary file ***Codes_description.pdf*** for a more detailed description of all the codes and files available.

References

[1] Adler, R. J. (2000). On excursion sets, tube formulas and maxima of random fields. *Annals of Applied Probability* 1–74.
[2] Algeri, S. (2019). TOHM: Testing One Hypothesis Multiple Times R package version 1.3.
[3] Algeri, S. (2020). Detecting new signals under background mismodeling. *Phys. Rev. D* **101** 015003.
[4] Algeri, S. et al. (2018). Statistical challenges in the search for dark matter. *arXiv:1807.09273*.

[5] Algeri, S. and van Dyk, D. A. (2020). Testing one hypothesis multiple times: the multidimensional case. *Journal of Computational and Graphical Statistics* **29** 358–371.

[6] Algeri, S. and Zhang, X. (2020). Exhaustive goodness-of-fit via smoothed inference and graphics. *arXiv preprint arXiv:2005.13011*.

[7] Aprile, E., Aalbers, J., Agostini, F., Alfonsi, M., Amaro, F., Anthony, M., Arneodo, F., Barrow, P., Baudis, L., Bauermeister, B. et al. (2017). First dark matter search results from the XENON1T experiment. *Physical review letters* **119** 181301.

[8] Atwood et al., W. B. (2009). The Large Area Telescope on the Fermi Gamma-Ray Space Telescope Mission. *The Astrophysical Journal* **697** 1071.

[9] Babu, G. J. and Rao, C. R. (2004). Goodness-of-fit tests when parameters are estimated. *Sankhya* **66** 63–74.

[10] Balázs, C. et al. (2017). ColliderBit: a GAMBIT module for the calculation of high-energy collider observables and likelihoods. *The European Physical Journal C* **77** 795.

[11] Barton, D. E. (1953). On Neyman's smooth test of goodness of fit and its power with respect to a particular system of alternatives. *Scandinavian Actuarial Journal* **1953** 24–63.

[12] Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L. et al. (2013). Valid post-selection inference. *The Annals of Statistics* **41** 802–837.

[13] Chakravarti, P., Kuusela, M., Lei, J. and Wasserman, L. (2021). Model-Independent Detection of New Physics Signals Using Interpretable Semi-Supervised Classifier Tests. *arXiv preprint arXiv:2102.07679*.

[14] Dauncey, P. D., Kenzie, M., Wardle, N. and Davies, G. J. (2015). Handling uncertainties in background shapes: the discrete profiling method. *Journal of Instrumentation* **10** P04015.

[15] Dissmann, J., Brechmann, E. C., Czado, C. and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis* **59** 52–69.

[16] Dudley, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis* **1** 290–330.

[17] Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The elements of statistical learning* **1**. Springer series in statistics New York.

[18] Jain, N. C. and Marcus, M. B. (1975). Central limit theorems for C (S)-valued random variables. *Journal of Functional Analysis* **19** 216–231.

[19] Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association* **89** 1000–1005.

[20] Lyons, L. (2013). Discovering the Significance of 5 sigma. *arXiv:1310.1284*.

[21] Mukhopadhyay, S. (2017). Large-scale mode identification and data-driven sciences. *Electronic Journal of Statistics* **11** 215–240.

[22] Mukhopadhyay, S. and Wang, K. (2020). Nonparametric High-dimensional K-sample Comparison. *Biometrika (to appear)*.

[23] Nagler, T., Schellhase, C. and Czado, C. (2017). Nonparametric estimation of simplified vine copula models: comparison of methods. *Dependence Modeling* **5** 99–120.
[24] Nelder, J. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society: Series A (General)* **140** 48–63.
[25] Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.
[26] Neyman, J. (1937). Smooth test for goodness of fit. *Scandinavian Actuarial Journal* **1937** 149–199.
[27] Panagiotelis, A., Czado, C., Joe, H. and Stöber, J. (2017). Model selection for discrete regular vine copulas. *Computational Statistics & Data Analysis* **106** 138–152.
[28] Parzen, E. (2004). Quantile probability and statistical data modeling. *Statistical Science* **19** 652–662.
[29] Pearson, E. S. (1938). The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika* **30** 134–148.
[30] Priel, N., Rauch, L., Landsman, H., Manfredini, A. and Budnik, R. (2017). A model independent safeguard against background mismodeling for statistical inference. *Journal of Cosmology and Astroparticle Physics* **2017** 013.
[31] Rayner, J. C. W. and Best, D. J. (1990). Smooth tests of goodness of fit: an overview. *International Statistical Review/Revue Internationale de Statistique* 9–17.
[32] Reed, M. and Simon, B. (1980). Methods of modern mathematical physics I: functional Analysis.
[33] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics* **23** 470–472.
[34] Scott, P. (2018). Dark matter model comparison. *BIRS Workshop on Dark matter model comparison, DM-Stat: Statistical Challenges in the Search for Dark Matter*.
[35] Signoretto, M., Dinh, Q. T., De Lathauwer, L. and Suykens, J. A. (2014). Learning with tensors: a framework based on convex optimization and spectral regularization. *Machine Learning* **94** 303–351.
[36] Taylor, J. E. and Worsley, K. J. (2008). Random fields of multivariate test statistics, with applications to shape analysis. *Ann. Statist.* **36** 1–27.
[37] Taylor, J., Takemura, A., Adler, R. J. et al. (2005). Validity of the expected Euler characteristic heuristic. *The Annals of Probability* **33** 1362–1396.
[38] Thas, O. (2010). *Comparing distributions*. Springer.
[39] Tyson, J. A. (2002). Large synoptic survey telescope: overview. In *Survey and Other Telescope Technologies and Discoveries* **4836** 10–20. International Society for Optics and Photonics.
[40] Westerdale, S. S. (2016). A study of nuclear recoil backgrounds in dark matter detectors, PhD thesis, Princeton University.