Goodness-of-fit tests for arbitrary multivariate models

Lolian Shtembari[*] and Allen Caldwell

Max Planck Institute for Physics, Munich 80805, Germany

(Received 20 March 2023; accepted 25 October 2023; published 4 December 2023)

Goodness-of-fit tests are often used in data analysis to test the agreement of a distribution to a set of data. These tests can be used to detect an unknown signal against a known background or to set limits on a proposed signal distribution in experiments contaminated by poorly understood backgrounds. Out-of-the-box nonparametric tests that can target any proposed distribution are only available in the univariate case. In this paper, we discuss how to build goodness-of-fit tests for arbitrary multivariate distributions or multivariate data generation models.

DOI: 10.1103/PhysRevD.108.123006

I. INTRODUCTION

Goodness-of-fit tests are often used in data analysis to test the agreement of a distribution to a set of data. These tests can be used to detect an unknown signal against a known background or to set limits on a proposed signal distribution in experiments contaminated by poorly understood backgrounds. Out-of-the-box nonparametric tests that can target any proposed distribution are only available in the univariate case: the Kolmogorov-Smirnov (KS) test [1], the Anderson-Darling test [2] or the recursive product of spacings (RPS) test [3]. In this paper, we discuss how to build goodness-of-fit tests for arbitrary multivariate distributions or multivariate data generation models. The resulting tests perform an unbinned analysis and do not need any trials factor or look-elsewhere correction since the multivariate data can be analyzed all at once. The proposed distribution or generative model is used to transform the data to an uncorrelated space where the tests are developed. Depending on the complexity of the model, it is possible to perform the transformation analytically or numerically with the help of a normalizing flow algorithm.

The flexibility of targeting vastly different univariate distributions is made possible by the probability integral transformation [4,5]. We start by reviewing this transformation in the univariate case and then extend it to the multivariate case. We then discuss different ways of performing a multivariate uniformity test and how to adapt this tool in the case of signal discovery or setting upper limits.

Finally, we consider examples for each application, using either real or artificial data, in order to test the sensitivity of our methods.

A. Univariate probability integral transformation

Given m univariate samples $\{x_i\}$ assumed to be independent and identically distributed (IID) according to a known distribution, $f(x)$, we can perform quantitative tests based on the probability integral transformation. Considering only continuous distributions $f(x)$ with cumulative $F(x)$, we first transform the samples onto the unit interval [0, 1] via $u_i = F(x_i)$. This reduces the task at hand to test transformed samples $\{u_i\}$ being distributed according to the standard uniform distribution $\mathcal{U}(0,1)$. Many tests have been developed for this univariate distribution. The take-away message from the univariate case is that, in order to develop a test statistic, it is easier to do so in a standardized space, such as the uniform interval [0, 1].

II. MULTIVARIATE PROBABILITY INTEGRAL TRANSFORMATION

Much like the univariate case, the goal in multivariate cases (in n dimensions) is to develop uniformity tests in the unit hypercube $[0,1]_n$. In order to target any given multivariate distribution $\mathbf{M}$, we need to transform the probability space described by $\mathbf{M}$ into $[0,1]_n$. This transformation can be easy or difficult depending on the distribution $\mathbf{M}$, specifically depending on the correlation among the dimensions of $\mathbf{M}$. In the following we show how to perform the transformation into the unit hypercube in three main cases: first, distributions comprised of uncorrelated dimensions are considered, moving then to distributions with correlated dimensions or sample generating processes for which a probabilistic model is not available. Finally hierarchical models are discussed.

[*] lolian@mpp.mpg.de

A. Independent dimensions

If the dimensions of the proposed distribution $\boldsymbol{M}$ are all independent of each other, then $\boldsymbol{M}$ is just a composition of n independent univariate distributions:

$$\boldsymbol{M} = [M_1, M_2, \ldots, M_n], \tag{1}$$

where M_j is the distribution of the jth dimension. Much like the univariate case, it is possible to transform the jth component of each sample using the corresponding cumulative distribution function F_{M_j}. Thus, the transformation of sample $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})$ in $[0,1]_n$ is simply

$$\boldsymbol{u}_i = [u_{i,1}, u_{i,2}, \ldots, u_{i,n}] = [F_{M_1}(x_{i,1}), \ldots, F_{M_n}(x_{i,n})]. \tag{2}$$

B. Correlated dimensions and generative models

If the dimensions of the distribution to be compared to the data are not mutually independent, then it might be difficult to write down a transformation to the hypercube. This is still possible when dealing with nicely behaved distributions, such as a multivariate normal distribution whose covariance matrix is not diagonal, but that might not be the case for a more complex distribution, such as a weighted sum of distributions. In such cases, it is possible to learn the transformation to the unit hypercube by using a normalizing flow (NF) which can perform a whitening of the distribution; i.e., transform the distribution so that it becomes a diagonal multivariate normal distribution in the new coordinates. Once the original distribution is transformed in this way, it is then possible to further transform it to the unit hypercube one component at a time as shown earlier.

The normalizing flow (NF) is made up of a neural network which is trained using samples from the proposed distribution $\boldsymbol{M}$. The samples needed for training can be obtained from an associated generative model or by sampling $\boldsymbol{M}$ using a Markov chain Monte Carlo. The use of the generative model is particularly interesting because it allows to train the NF without having a normalized distribution or any model at all. In such cases, the NF is learning the associated distribution and the transformation all at once. References [6,7] offer a nice review of the theory and some of the many applications of normalizing flows. In order to show the feasibility of this approach, a proof of principle example is presented where a normalizing flow is used to whiten data sampled from a sum of three two-dimensional normal distributions. A sampled distribution is depicted in Fig. 1 and the resulting marginal distributions of the whitened samples are shown in Fig. 2. The normalizing flow used for this example was adapted from [7].

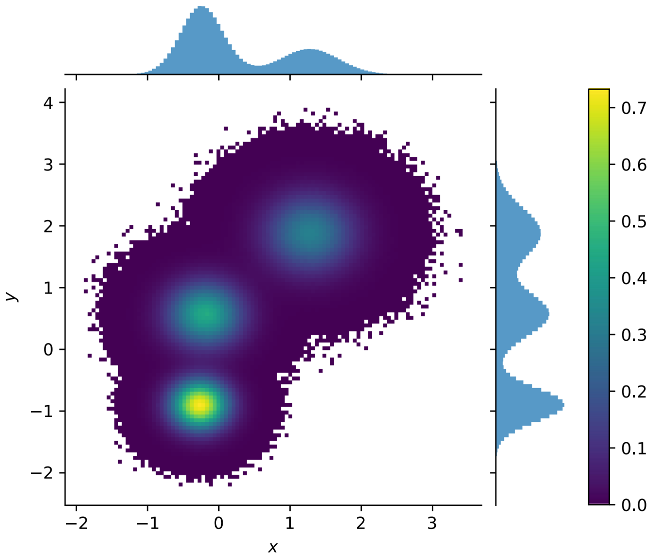

FIG. 1. Sample distribution of the sum of three two-dimensional Gauss distributions.

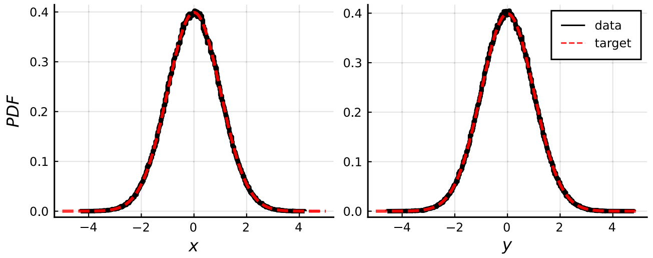

FIG. 2. Whitened marginal distributions after transforming with the normalizing flow.

C. Hierarchical models

Given a hierarchical model, the distribution of some components of the data is dependent of the values of other components, which are referred to as hyperparameters of the model. If the hyperparameters are mutually independent or if a transformation to the unit hypercube is available for their distribution and if the same is true for all the dependent parameters at each layer of depth of the hierarchical model, then it is possible to transform the whole distribution into the unit hypercube in stages.

Consider for example a two-layer hierarchical model producing distributions $\boldsymbol{M} = [\boldsymbol{M}_1, \boldsymbol{M}_2(\boldsymbol{M}_1)]$. $\boldsymbol{M}_1$ models the distribution of the hyperparameters $\boldsymbol{x}^{\mathrm{high}}$ of the model and these components can be transformed to the corresponding uniform unit hyperspace using the associated function T_{M_1}. The distribution of the dependent parameters $\boldsymbol{x}^{\mathrm{low}}$ is affected by the observed value of the hyperparameters $\boldsymbol{x}^{\mathrm{high}}$:

$$\boldsymbol{x}_i^{\mathrm{low}} \sim \boldsymbol{M}_2(\boldsymbol{x}_i^{\mathrm{high}}). \tag{3}$$

For any given sample $\boldsymbol{x}_i$, the value of the hyperparameters $\boldsymbol{x}_i^{\mathrm{high}}$ is fixed, so the distribution $\boldsymbol{M}_2(\boldsymbol{x}_i^{\mathrm{high}})$ is fully defined and it is possible to compute the corresponding transformation to the unit hyperspace. While $T_{\boldsymbol{M}_1}$ is sample independent, $T_{\boldsymbol{M}_2}$ is sample dependent. In case of hierarchical models with more layers, this staged transformation approach is to be repeated for each layer.

III. UNIFORMITY TESTS IN THE UNIT HYPERCUBE

In the following we discuss various methods that allow to perform a multivariate uniformity test by reducing this task to a series of univariate uniformity tests. These tests are sensitive to nonuniformities in a transformed dataset and their application is twofold: (1) detection of clustering of events against a uniform background, in a discovery scenario, and (2) upper limit on the rate of events corresponding to the uniform component of the data, representative of a proposed signal, against unknown backgrounds. For the latter, a desired confidence level is set in advance.

A. Projection—Discovery

Assume we have m samples within a unit hypercube $\{\boldsymbol{u}_i\} \in [0,1]_n$. The n components of each sample are assumed independent of one another after the necessary transformations. The projections of the samples along each axis of the hypercube therefore yield n univariate uniformly distributed sets of data: $\{u_{i,j}\}$ for the jth dimension. For each one of these projected datasets $\{u_{i,j}\}$ it is possible to perform a uniformity test using a test statistic of choice and condense the information for the jth dimension in one scalar p value p_j. Given our assumptions, the expected distribution of each p value p_j is uniform, and moreover, the p values will be independent of one another.

On this resulting dataset, $\{p_j\}$, it is possible to perform a uniformity test using a test statistic of choice in order to check whether there are any significant deviations from uniformity. The result of this last uniformity check results in one last p value p_{final} which is the overall p value of the multivariate goodness-of-fit test.

As pointed out in the discussion above, in order to obtain the intermediate p values, $\{p_j\}$, and then the final one, p_{final}, it is possible to use any test statistic of choice, as long as the chosen statistics preserve the noncorrelation among dimensions (results from tests that have a Poisson dependent factor, for example, will be correlated, since the same number of samples is projected on all dimensions). What is important is that the distribution of the resulting p values is uniform. This implies that the test statistic used for the evaluation of the intermediate p values, $\{p_j\}$, does not have to be the same as the one used to evaluate p_{final}; as a matter of fact, one could also use different tests for different dimensions in the evaluation of $\{p_j\}$, but it might be a more consistent approach to consider all dimensions equally and use the same test for all projections.

In the previous discussion, we considered a dataset of m samples $\{\boldsymbol{u}_i\} \in [0,1]_n$. In such a case, if the number of events m is large, it might be appropriate to use a test such as RPS or KS in order to pick up on a signal in any of the projections. Afterwards, when considering the n p values $\{p_j\}$, it could be better to look for outliers, since already one of a few small p_j could be indicative of the presence of a signal in our data. In this case, especially when dealing with low-dimensionality spaces (n small), instead of using RPS or KS on the set $\{p_j\}$ it might be more informative to look at the smallest p value or rather their product in case we want to improve the sensitivity in the presence of multiple small p values.

1. Minimum p value

As discussed, observing one small p value might already be enough to point to a possible signal in the data. Under the assumption of a uniform distribution of $\{p_j\}$, the distribution of $p_{\mathrm{min}} = \min\{p_j\}$ is simply the first order statistic, and it follows a Beta distribution [8]:

$$p_{\mathrm{min}} = \min_j\{p_j\} \sim \mathrm{Beta}(1,n), \tag{4}$$

where n is the dimensionality of the original data. Thus the final p value is

$$p_{\mathrm{final}} = F_{\mathrm{Beta}}(p_{\mathrm{min}}; 1, n), \tag{5}$$

where $F_{\mathrm{Beta}}(x; a, b)$ is the cumulative distribution function of the Beta distribution with parameters (a, b).

2. Product of p values

Given more than one small p value p_j, looking only at the smallest one might be reductive and we could gain in sensitivity by combining the small p values together. One way of doing so is to consider the product of all p values:

$$p_{\mathrm{prod}} = \prod_{j=1}^{n} p_j. \tag{6}$$

Once again, we expect all $\{p_j\}$ to be uniformly distributed, and the distribution of p_{prod} is known [9]:

$$P(p_{\mathrm{prod}} = x; n) = \frac{(-1)^{n-1}}{(n-1)!}[\ln(x)]^{n-1} \tag{7}$$

thus the final p value p_{final} is

$$p_{\mathrm{final}} = F(p_{\mathrm{prod}}; n) = p_{\mathrm{prod}} \cdot \sum_{j=1}^{n} \frac{(-1)^{j-1}}{(j-1)!}[\ln(p_{\mathrm{prod}})]^{j-1}. \tag{8}$$

B. Projection—Limit setting

Several spacings-based tests have been developed for this task in one dimension, such as the maximum-gap or optimum-interval (OI) methods [10], as well as the sum-of-largest-sorted spacings (SLSS) or the product-of-complementary spacings (PCS) [11].

When setting limits in the univariate case, given a test T with cumulative distribution F_T, its Poisson-averaged p value is calculated as

$$1 - p = F_{T,\mathrm{Pois}}(t_{\mathrm{obs}}|\mu) = \sum_{n=0}^{\infty} F_T(t_{\mathrm{obs}}|n) \cdot \frac{\mu^n e^{-\mu}}{n!}, \quad (9)$$

where t_{obs} is the observed value of the test statistic. Given Eq. (9) it is possible to find the event rate μ_{lim} with a confidence level (C.L.) such that

$$F_{T,\mathrm{Pois}}(t_{\mathrm{obs}}|\mu_{\mathrm{lim}}) = \mathrm{C.L.} \quad (10)$$

For a more complete discussion regarding how to set upper limits, see Ref. [11].

For the multivariate case, as discussed before, given m uniformly distributed samples $\{\boldsymbol{u}_i\} \in [0,1]_n$, we consider the projection of the samples on the n axes, knowing these will be uniformly distributed as well. For each one of these projected datasets $\{u_{i,j}\}$ it is possible to estimate an upper limit μ_j on the event rate with confidence level C_1.

Out of the upper limits $\{\mu_j\}, j = 1, \ldots, n$ obtained from each projection, we can use a best of the bunch approach and select the smallest one as the final limit:

$$\mu_{\mathrm{final}} = \min_j \{\mu_j\}. \quad (11)$$

At this point we must consider the confidence level C_n associated with this estimate. If the projected limits $\{\mu_j\}$ were completely independent of one another, then we might consider that selecting the smallest limit amounts to a resulting confidence level C_n equal to the product of n Bernoulli variables with rate C_1, thus:

$$C_n = (C_1)^n. \quad (12)$$

Under this assumption, we could easily select the confidence level C_1 of the individual projection limit estimations in order to ensure that C_n is equal to the desired value.

This assumption is however not correct. Although the distribution of the projected events on each axis is independent, the number of samples projected on each axis is not: if there are m samples in the multidimensional space then there will be m samples on each projected dataset $\{u_{i,j}\}, j = 1, \ldots, m$. In order to set a limit we consider both the distribution of events and the total number of events, merging a goodness-of-fit test with a Poisson test. Since all projected datasets $\{u_{i,j}\}$ share the same number of events, this introduces a correlation in the Poisson statistic part of each limit-setting estimation, rendering all resulting limits correlated.

Although the projection-independence assumption is not valid if applied after the Poisson averaging, it is possible to calculate the corrections necessary to ensure the desired final confidence level C_n. We assume that C_n is a function of the projection specific confidence level C_1 and that it is dependent on the value of the reconstructed limit μ_{final}, for a given number of dimensions n: $C_n(\mu_{\mathrm{final}}, C_1|n)$. If we seek a specific confidence level (C.L.), then we need to find the value of C_1 that for the resulting best limit μ_{final} yields

$$C_n(\mu_{\mathrm{final}}(C_1), C_1|n) = \mathrm{C.L.} \quad (13)$$

This equation is just a one-dimensional root finding problem in C_1 which can be solved iteratively (for example using a bisection algorithm) by estimating the error at $\mu_{\mathrm{final}}(C_1)$ for a proposed value of C_1. The estimation of the error rate can be done via Monte Carlo simulations, producing data according to a uniform distribution in the n-dimensional hypercube, since Eq. (13) only needs to hold in this nominal case.

Although this procedure might seem complicated, it is easy to devise and can be performed well before any real analysis has to be run, during the method validation phase, allowing for the tabulation, interpolation, and sharing of $C_n(\mu_{\mathrm{final}}, C_1|n)$. We have calculated the exact correction for the SLSS method and an approximate correction for the OI method up to 5 dimensions.

C. Product of complementary spacings—Limit setting

1. Best projection

As discussed above, if one calculates the Poisson-averaged p value on each projected dataset and then chooses the most significant value, a correction needs to be calculated to account for the correlation of these values due to the fixed number of samples on each axis. In order to avoid this problem, if the definition of the test statistic chosen allows it, it is possible to perform the selection of the best p value before averaging with a Poisson distribution. In such a case it would be trivial to calculate the correct confidence level without having to resort to numerical corrections.

The product-of-complementary spacings, PCS, is defined as [11]

$$T(\{u_i\}) = -\sum_{i=1}^{n+1} \log(1 - u_i + u_{i-1}) \quad (14)$$

for a univariate ordered set of data $\{u_i\}$ where $u_0 = 0$ and $u_{n+1} = 1$. For each of the projected datasets, one can compute the corresponding value of the test T_j and its p value [here $p_j = F_T(T_j)$]. The n projected p values p_j form an order statistic with uniform distribution. If we were to select the largest $F_T(T_j)$, its distribution would be simply

$$f\left(\max_j(F_T(T_j))\right) = \mathrm{Beta}(n, 1). \tag{15}$$

Given the test-statistic values T_j for each projection, the Poisson-averaged p value of the largest one, $T_{\max} = \max_j(T_j)$, is

$$F_{T,\mathrm{Pois}}(T_{\max}|\mu) = \sum_{m=1}^{\infty} F_{\mathrm{Beta}}[F_T(T_{\max}|m)|n] \cdot \frac{\mu^m e^{-\mu}}{m!}. \tag{16}$$

It follows that the upper limit μ_{lim}, with a C.L., is such that

$$F_{T,\mathrm{Pois}}(T_{\max}|\mu_{\mathrm{lim}}) = \mathrm{C.L.} \tag{17}$$

2. Sum of projections

Given the PCS test-statistic values T_j on each projection, instead of selecting the largest, we can consider their sum:

$$T_{\mathrm{sum}} = \sum_{j=1}^{n} T_j \tag{18}$$

which can be interpreted as a product of the product of complementary spacings. Assuming we know the distribution of T_{sum} for a fixed number of events m, $F(T_{\mathrm{sum}}|m)$, then we can compute the Poisson-averaged p value of this test for a given event rate μ:

$$F_{\mathrm{Pois}}(T_{\mathrm{sum}}|\mu) = \sum_{m=1}^{\infty} F(T_{\mathrm{sum}}|m) \cdot \frac{\mu^m e^{-\mu}}{m!}. \tag{19}$$

Given this definition, it is possible to invert the formula and find the upper limit on the event rate up to a desired confidence level. For example, the 90% confidence level upper limit μ_{lim} is such that

$$F_{\mathrm{Pois}}(T_{\mathrm{sum}}|\mu_{\mathrm{lim}}) = 0.9. \tag{20}$$

If $F(T_j|m)$ is known, it is rather easy to compute $F(T_{\mathrm{sum}}|m)$. Since T_j are all IID, the distribution of T_{sum} is just $f_{T,m}$ convolved $n-1$ times with itself:

$$f(T_{\mathrm{sum}}|m) = \underbrace{f(T|m) * f(T|m) * .. * f(T|m)}_{n\ \mathrm{times}}. \tag{21}$$

Since $F(T|m)$ has been tabulated in the Julia package SPACINGSTATISTICS.JL [12] and is available as a monotonic cubic spline polynomial function, it is possible to easily obtain its derivative $f(T|m)$, transform it to the Fourier space using an FFT, raise it to the power of n and transform back to the real space to obtain $f(T_{\mathrm{sum}}|m)$:

$$f(T_{\mathrm{sum}}|m) = FFT^{-1}\{[FFT(f(T|m))]^n\}. \tag{22}$$

This procedure is used for the tabulated $F_{\mathrm{PCS},m}$ ($m \leq 10^4$). For values of m larger than 10^4 we use the asymptotic distribution of $F_{\mathrm{PCS},m}$, which is a Gaussian distribution, thus rendering the calculation of the convolution much easier.

These two approaches show how to adapt the PCS test to a multivariate limit-setting scenario, similarly to how the minimum p value and product of p values were used in the multivariate discovery case. Although we discussed the PCS test specifically, these corrections apply in general to any test statistic T calculated where the Poisson averaging can be calculated as a final step.

D. Volume transformation method

Finally, we consider a different dimensionality reduction strategy. Given m samples $\{\boldsymbol{u}_i\} \in [0, 1]_n$, instead of projecting them onto the axes and obtaining n independent sets of univariate data, we can use a dimension-reducing transformation to map them all at once onto a single univariate dataset. To achieve this, we calculate the volume contained in the hyper-rectangle defined by its projections simply by taking the product of its coordinates:

$$v_i = V(\boldsymbol{u}_i) = \prod_{j=1}^{n} u_{i,j}. \tag{23}$$

Calculating the volume in this way for each multivariate sample we obtain a simple univariate dataset: $\{\boldsymbol{u}_i\} \xrightarrow{V} \{v_i\}$. Since the $\{\boldsymbol{u}_i\}$ were IID samples, so are the $\{v_i\}$ (although not uniformly distributed). Since v_i is the product of n independent uniform variables, whose distribution is given by Eq. (7), its probability distribution is known. Using the probability integral transformation, Eq. (8), we can therefore transform $\{v_i\}$ into a set of uniform IID samples $\{z_i\}$. We can use these to then perform a univariate uniformity test using a test statistic of choice; standard discovery and limit-setting tests can then be used in order to analyze the data.

IV. EXAMPLE—nD DISCOVERY

Here we illustrate how the proposed goodness-of-fit tests can be used in a scenario where a possible “new physics” model is searched for but it is not wished to specify how the new physics might populate the data space. It is then to be tested whether the data follows a known distribution, which is a “background” to a possible new signal. After having collected some data, one wants to quantify the goodness of fit of the background-only distribution to the data and a resulting low p value could indicate the presence of events distributed according to an additional, previously unknown, signal distribution.

A. Multivariate Gaussian signal

In this example the background is modeled by a simple uniform distribution in the five-dimensional hypercube $[0,1]_5$ and in order to illustrate how the presence of an actual signal (alternative hypothesis) would affect the outcome, additional events are injected, following a multivariate normal distribution randomly positioned within the hypercube with isotropic variance of either 0.01 or 0.1. The number of events is Poisson fluctuated for both background and signal populations, with expected values of $\langle n_b \rangle = 10^4$ and expected values of $\langle n_s \rangle$ ranging up to 10^3.

The p-value distributions under the assumption of H^0 (i.e., only background is present) are shown in Fig. 3: the results corresponding to the narrow signal ($\Sigma = I_5 \cdot 0.01$) are on the left (first column) and those corresponding to the broad signal ($\Sigma = I_5 \cdot 0.1$) are on the right (second column); the first two rows present p-value distributions calculated using projection methods while the third row presents p-value distributions obtained with the volume transformation method; the fourth row presents the sensitivity of each scenario quantified as the median p value for each distribution. Regarding the results of the projection method, the evaluation of the intermediate p values was performed using the KS test, given the large count rates, while the evaluation of the final p value, since there are only five dimensions, was performed using the two tests previously described, namely the minimum and the product of intermediate p values, corresponding to the first and second row, respectively. Similarly, the KS test statistic was used in the final uniformity test after performing the volume transformation.

Distributions with no signal ($\langle n_s \rangle = 0$) show a flat p-value distribution, as expected, while the distributions of trials with injected signals are trending towards smaller p values, indicating the worsened goodness of fit for the background-only model. The distributions of trials where the signal has smaller variance (left) are much more skewed towards small p value compared to those where a larger variance signal was injected (right). This shows how the sensitivity of the tests varies when targeting clusters of varying width and strength relative to the background.

In this example, since the signal can be spotted in the projection of multiple dimensions, the product of the p-values test (second row) offers the largest rejection probability of the null hypothesis compared to the volume-transformed p value (third row) or the minimum p-value test (first row).

B. Multivariate Gaussian-shell signal

Instead of injecting a clustered signal, we assess the sensitivity of our methods in the case of a Gaussian-shell signal. Our signal is five dimensional and characterized by a radius $r = 0.25$, a radial standard deviation of either $\sigma_r = 0.02$ or $\sigma_r = 0.1$, and the center of the distribution μ chosen at random within the hypercube $[0.25, 0.75]_5$. The results are shown in Fig. 4. In this case, we notice that the sensitivity to either signal thickness, σ_r, is very similar, which shows that all methods are mostly sensitive to the shell-like structure and its radial extension. Of the three tested methods, the product of p values shows the highest sensitivity, followed by the minimum p value and then the volume transformed p value.

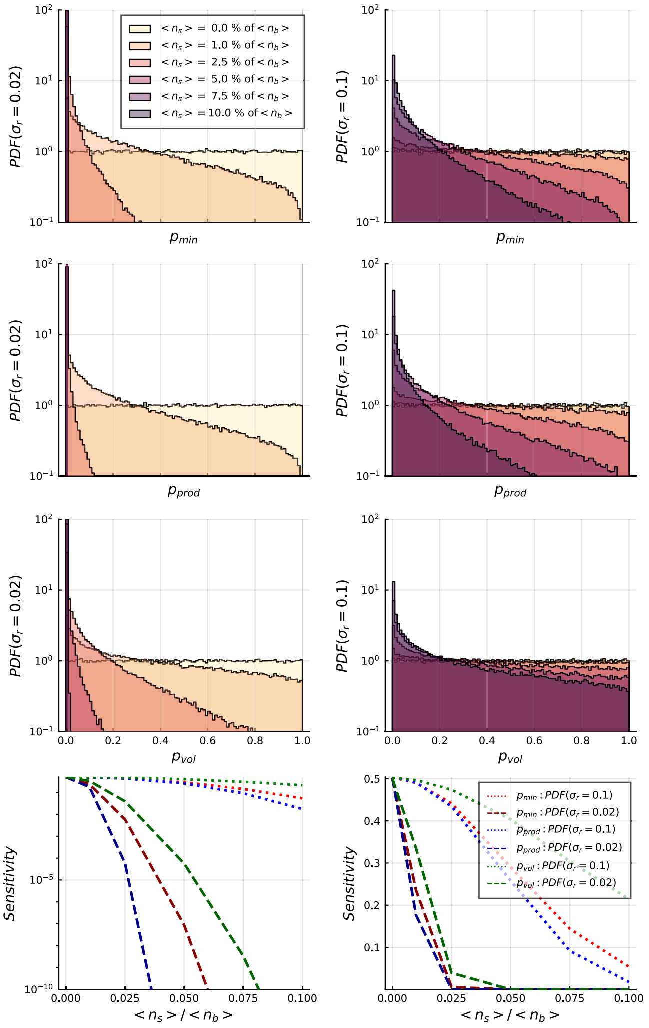

FIG. 3. Distributions of p values for background-only samples ($\langle n_s \rangle = 0$) and background plus randomized signal injections from a 5D Gaussian distribution: "narrow" signal with random $\mu \in [0.2, 0.8]$, $\Sigma = 0.01 \cdot \mathrm{I}_5$ (left) and "wide" signal with random $\mu \in [0.2, 0.8]$, $\Sigma = 0.1 \cdot \mathrm{I}_5$ (right) of varying strength; comparison to the background model for either the minimum p-value statistic (first row), the product of p-values statistic (second row) or the volume-transformed p value (third row); median p value (sensitivity) both in linear and logarithmic scale (fourth row).

Note that the data in the previous examples were analyzed all in one pass for each trial, meaning that the

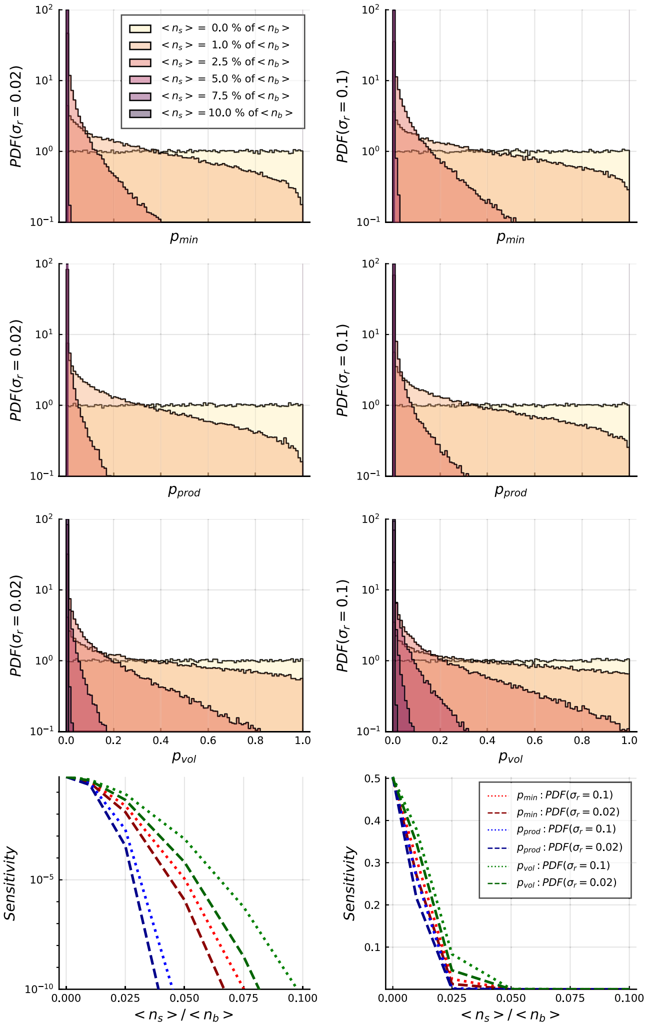

FIG. 4. Distributions of p values for background-only samples ($\langle n_s \rangle = 0$) and background plus randomized signal injections from a 5D Gaussian-shell distribution: narrow signal with random $\mu \in [0.25, 0.75]$, $r = 0.25$, $\sigma_r = 0.02$ (left) and wide signal with random $\mu \in [0.25, 0.75]$, $r = 0.25$, $\sigma_r = 0.1$ (right) of varying strength; comparison to the background model for either the minimum p-value statistic (first row), the product of p-values statistic (second row) or the volume-transformed p value (third row); median p-value (sensitivity) distribution both in linear and logarithmic scale (fourth row).

extracted p values do not need any corrections for a “trials effect” or “look-elsewhere effect.” Of course, if one analyzes many separate sets of data, the resulting p value will need to be corrected as is usually done in the univariate case.

C. Extragalactic cosmic rays

In the previous examples, we demonstrate the sensitivity and flexibility of our methods using synthetic data produced by known signal distributions. Here, we present the analysis of real multivariate data; specifically the arrival coordinates and energies of cosmic rays of extragalactic origin, as measured by the Pierre Auger Observatory [13]. Here, we consider the latest public dataset [14], consisting of 2,635 ultrahigh energy cosmic rays above 32 EeV. The arrival directions of each event are expressed in equatorial coordinates, (α, δ), the right ascension (R.A.) and declination (Dec.) respectively, as described in [15]. The expected distribution of the measured right ascension is uniform, whereas the distribution of the declination depends on the maximum zenith angle of arrival, θ_m, as described in [16]. Finally, the expected distribution of the energy of cosmic rays (E) is estimated in [17].

The measured coordinates and reconstructed energies all suffer from uncertainties, as described in the respective publications. In order to account for these in the predicted distributions, we convolve the predicted spectra with a Gaussian distribution with $\sigma = 1°$ for the angular coordinates and $\sigma = 0.074 \cdot E$ for the energy (meaning that the uncertainty increases for more energetic events). Given these distributions, and given that the two angular coordinates (R.A. and Dec.) and energy are all independent, we can transform the data to the unit hypercube and analyze them using a test statistic of choice and one of the methods described above. Examining the dataset, we notice that the data is reported up to a precision of 0.1° and 0.1 EeV for the angular coordinates and energies, respectively. This implies multiple events with identical coordinates in one of the dimensions, which skews the results of many test statistics. In order to avoid these biases, we produce replicas of the original data with added noise, modeled as a uniform distribution $\mathcal{U}[-\Delta/2, +\Delta/2]$, where Δ is the precision of each dimension. A p value derived from such a data manipulation will be affected by statistical fluctuations. In order to remove this effect, we produce 10^4 replicas of the original dataset and analyze them, deriving a distribution of p values and select the median as representative of the original dataset.

As previously seen, the product of p values offers the highest sensitivity for signal discovery, so we adopt this method in this analysis. The intermediate p values, one per data projection, are obtained with either the Kolmogorov-Smirnov (KS) [1] or best-sum-of-spacings (BSS) [18] test statistics. Through the KS test it is possible to determine the position of the signal by identifying where the value of D_{sup}, the KS statistic, is observed. The description of the BSS test is reported briefly in the Appendix below and in Ref. [18], but differently from the KS test, it allows to identify an interval where the putative signal might be located.

In Fig. 5, we show the original data [14] as a scatter plot in galactic coordinates, scaling the size of the points with the square of the respective energy (purely for pictorial clarity). We also show the median location for the largest value of the KS statistic (in blue), as well as the median region producing the smallest value of the BSS statistic (in red). The median p values obtained from analyzing the data

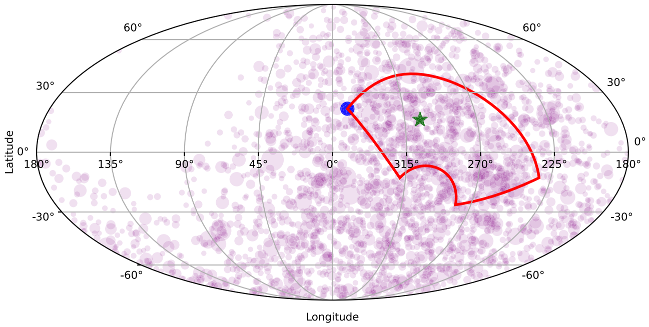

FIG. 5. Scatter plot of the Pierre Auger Observatory data [14], in galactic coordinates, where the size scales with the square of the reconstructed energy; in blue the median location of the largest KS statistic; in red the median region identified by the BSS test; in green the highest local significance estimated by the Auger collaboration [15].

are 0.82×10^{-3} and 0.84×10^{-1} for the KS and BSS tests, respectively. Comparing these results with the search for angular correlation between cosmic rays of extragalactic origin and their sources conducted in [15], we notice that the region highlighted by the BSS test agrees with the region of highest significance identified by the Pierre Auger collaboration. The median energy range corresponding with the smallest values of the BSS statistic is [58, 93] EeV, whereas the median value corresponding to the largest KS statistic is ~54 EeV. The median angular location determined with the KS test sits at a corner of the BSS region, and does not coincide with the point location of highest local significance found by the Auger collaboration. The latter had Galactic coordinates $(l, b) = (305.4°, 16.2°)$ in [15] (reported in green in Fig. 5), and was found above an energy threshold of 41 EeV using a top-hat window analysis with angular range 24° and comparing the number of recorded events with the expected number. We note that the results of our tests rely primarily on the distribution of events and consider the whole dataset at once, such that our result does not need a trial factor correction.

V. EXAMPLE—nD LIMIT SETTING

The performance of our proposed methods for limit setting is explored in a series of simulated experiments for multivariate sample distributions. We consider the case where a background model is not present, and only a distribution of counts according to a signal model is available. In this case, the task is to set a limit on the signal strength of the signal model.

A. Background-free experiment

We start by considering the case in which no background contaminates the experiment, in order to estimate the baseline of the different methods. Figure 6 shows the median of the C.L. = 0.90 upper limits on the event rate normalized to the median limit of the Poisson test. We notice that in this baseline scenario the Poisson test is the best of the bunch, as expected, but it does not drastically outperform the others.

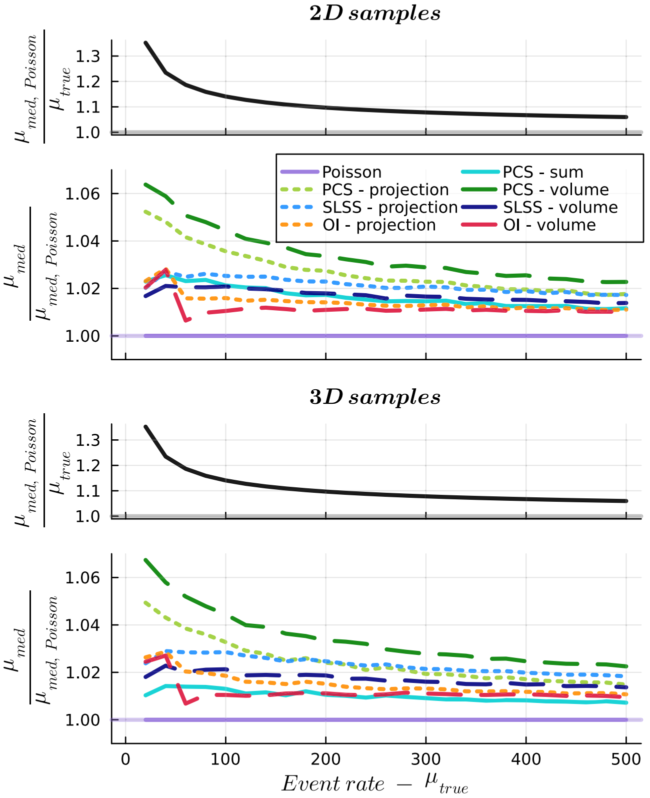

FIG. 6. Median C.L. = 0.90 upper limit for the Poisson test, upper panels, and for tests discussed in the text normalized to the limit from a standard Poisson probability test, lower panels, for 2D (top) and 3D (bottom) uniform signal distributions and no background.

B. Background-only experiment

Next we investigate the case in which a background is present in our simulations and the signal strength is negligible in comparison: this mimics a rare process search in which the signal is absent.

1. *Exponential distribution*

We first consider a background resulting from the product of n independent exponential distributions of rate 0.1 in each dimension.

Figure 7 reports the median C.L. = 0.90 upper limits of the measured event rate normalized to the smallest median result for a specific background event rate μ_{bkg}. Analyzing these results, we notice that the volume transformation method provides the best limits, regardless of the test used. All other projection-based methods perform similarly: in the two-dimensional scenario, the limits are a factor 1.5–2 worse than the volume transformation results, and in

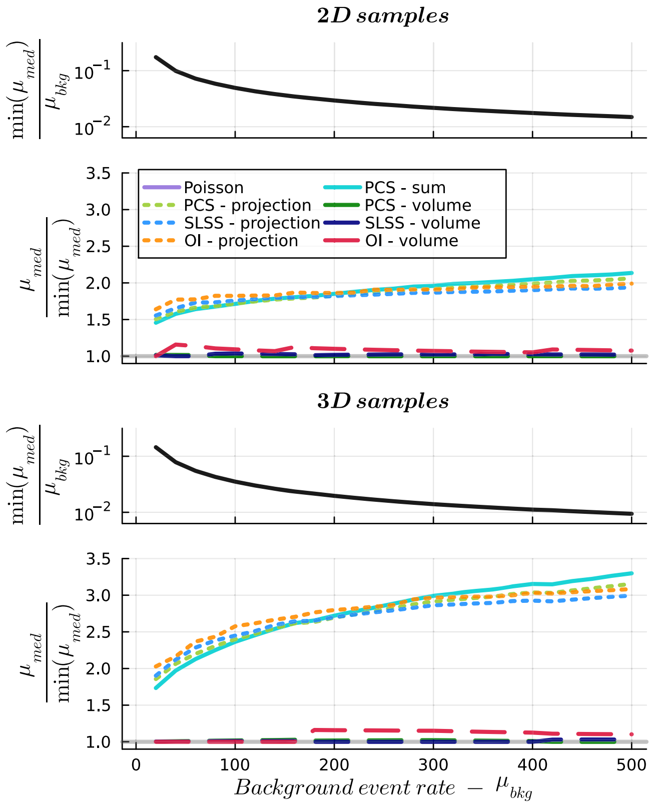

FIG. 7. Median C.L. $= 0.90$ signal upper limit for the best available test, normalized to the background strength, upper panels, and for tests discussed in the text normalized to the limit from the best test, lower panels, for 2D (top) and 3D (bottom) distributions containing only an exponentially distributed background.

the case of a three-dimensional distribution a factor 2–3 worse.

2. Gaussian distribution

Next we consider a background distributed according to a multivariate Gaussian centered at the middle of the hypercube and with covariance matrix $\Sigma = I \cdot 0.01$.

Figure 8 reports the median C.L. $= 0.90$ upper limits of the measured event rate normalized to the smallest median result for a specific background event rate μ_{bkg}. Once again, the volume transformation method provides the best limits, regardless of the test used. Out of these, the SLSS test is the best of the bunch, since it is better suited to analyze datasets that present multiple disconnected low density regions.

The projection-based methods provide weaker limits: the SLSS and PCS version being up to a factor 1.25(1.5) larger in the 2D (3D) case, respectively; the OI test limits are weaker by a factor 1.5(1.75) in the 2D (3D) case, respectively. This is understandable since this test relies only on one low density region to estimate its limit.

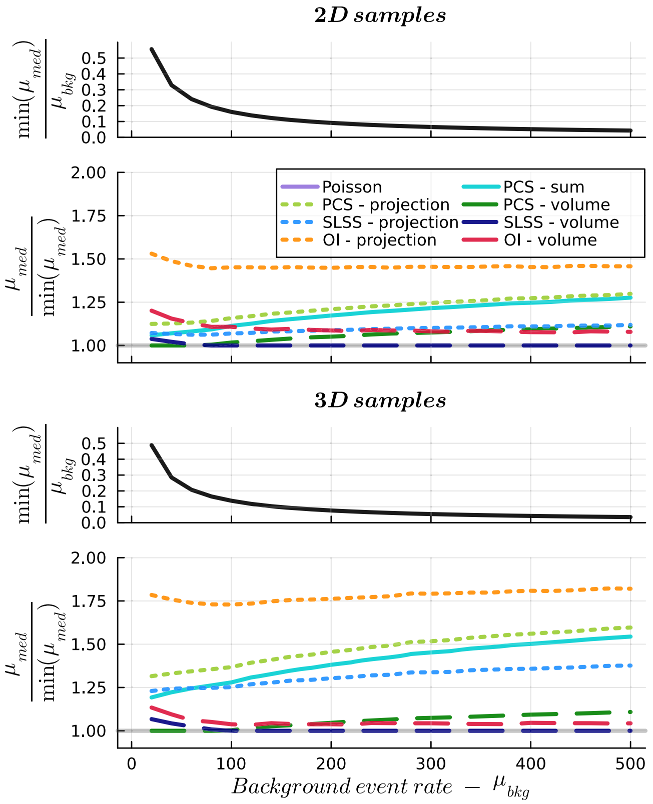

FIG. 8. C.L. $= 0.90$ upper limit normalized to minimum median result respectively for 2D (top) and 3D (bottom) multivariate normal distributions with $\Sigma = I \cdot 0.01$ centered in the middle of the hypercube. The upper panels in each case show the best limit result normalized to the background expectation.

3. Concave distribution

Finally, we consider a bowl shaped background, obtained by reversing the roles of signal and background distribution of the previous example: assuming a uniform background and a Gaussian signal in the real space (truncated to the unit interval [0, 1] with $\mu = 0.5$ and $\sigma = 0.1$), we perform the probability integral transformation with respect to the latter, obtaining a bowl shaped background distribution in the cumulative space.

Figure 9 reports the median C.L. $= 0.90$ upper limits of the measured event rate normalized to the smallest median result for a specific background event rate μ_{bkg}. In this case we show results for four- and five-dimensional distributions of events. We notice that the best results in this case are set by the OI test with volume transformation. This is reasonable since there is only one fully connected region of low event density, namely the basin of the bowl, thus being the best-suited case for the OI test. The next best results are obtained by the SLSS and PCS volume transformations, which yield no more than 25% larger limits. Finally, the projection-based methods yield the most conservative limits, with the OI test being the best of this subset.

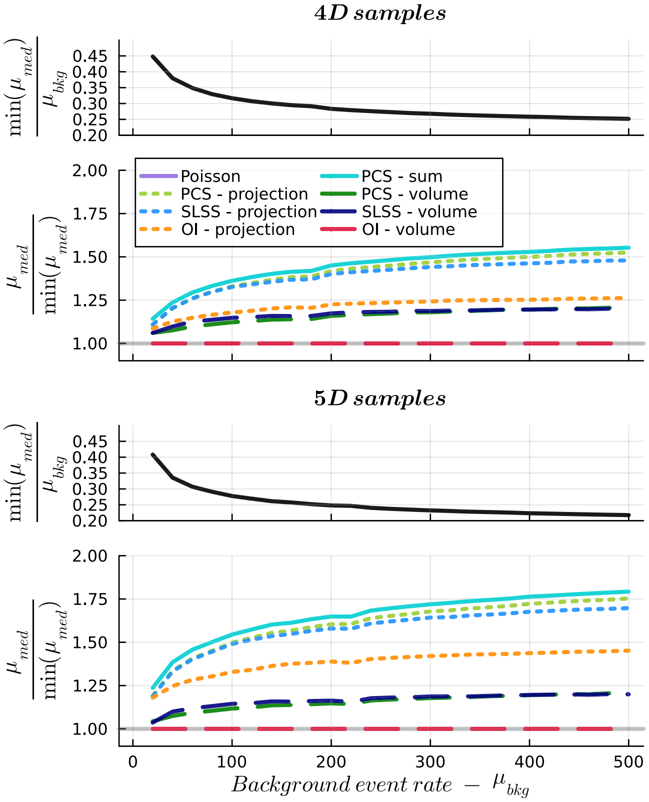

FIG. 9. C.L. $= 0.90$ upper limit normalized to minimum median result, respectively, for 2D (top) and 3D (bottom) multivariate normal distributions for the concave background model. The upper panels in each case show the best limit result normalized to the background expectation.

VI. CONCLUSIONS

We have provided novel nonparametric statistics to perform goodness-of-fit tests targeting any given multivariate distribution or multivariate generative model by means of a transformation to the uniform unit hypercube. Our approaches allow for unbiased tests, either by considering the volumes identified by each sample or by taking into account their projections. The tests developed with these methods perform an unbinned analysis of the data and do not need any trials factor or look-elsewhere correction since the multivariate data is analyzed all at once. These novel methods allow to test for the presence of a signal beyond the known background expectation, or to set a limit on a signal's event rate in cases where the background is not well modeled. The sensitivity of our proposals was tested in the context of mock signal searches as well as using real data released by the Pierre Auger collaboration. We have also compared the limit setting capabilities of our methods in simulated rare process searches under a variety of background behaviors.

The test statistics described in this paper are simple to use and the code is available to interested users.

ACKNOWLEDGMENTS

We thank Dr. Francesca Capel for her help and useful discussions regarding the analysis of the Pierre Auger data. We thank Dr. Oliver Schulz for many helpful discussions and comments on this note and its content. We thank Dr. Vasyl Hafych and Michael Dudkowiak for their help in the implementation of the normalizing flow used in this paper.

APPENDIX: BEST-SUM-OF-SPACINGS

Here, we briefly describe the best-sum-of-spacings statistic, but for a detailed discussion we refer to Chapter 3.6.1 of [18]. Given a set of n univariate data, we consider their distribution after the cumulative probability transformation, i.e., we consider them to be uniformly distributed in the unit interval [0, 1], and we refer to their ith order statistic as $u_{(i)}$. We define the spacing of rank k (containing k intervals) as $S_{i,k} = u_{(i)} - u_{(i-k)}$ for any i. For any rank k, we can identify the smallest such spacing: $S_k^{\min} = \min_i S_{i,k}$. This represents the most likely candidate for a cluster of k events out of n. We can quantify the significance of such a possible cluster of events for any rank k available in the given data by computing the p value for each one, p_k (assuming we know the distribution of $S_k^{\min}$ for the given values of k and n). This leaves us with a list of n p values, one for each rank, and we choose to construct the BSS test statistic out of the smallest p_k: $\mathrm{BSS} = \min_k p_k$. The spacing corresponding to the smallest p_k is the most likely cluster of events deviating from a uniform distribution. The definition of this test statistic is similar to the optimum interval method, which relies on an analogous construction, considering the largest p_k instead of the smallest. The distributions of $S_k^{\min}$ and BSS are tabulated numerically, based on a large number of simulations and interpolations, and are described in detail in [18].

[1] A. N. Kolmogorov, G. Ist. Ital. Attuari **4**, 83 (1933).

[2] T. W. Anderson and D. A. Darling, J. Am. Stat. Assoc. **49**, 765 (1954).

[3] P. Eller and L. Shtembari, J. Instrum. **18**, P03048 (2023).

[4] K. Pearson, Biometrika **1**, 390 (1933).

[5] K. Pearson, Biometrika **25**, 379 (1933).

[6] I. Kobyzev, S. J. Prince, and M. A. Brubaker, IEEE Trans. Pattern Anal. Mach. Intell. **43**, 3964 (2021).

[7] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2019), Vol. 32, https://proceedings.neurips.cc/paper_files/paper/2019/hash/7ac71d433f282034e088473244df8c02-Abstract.html.

[8] H. A. David and H. N. Nagaraja, *Order Statistics* (Wiley, New York, 2003).

[9] M. Springer, *The Algebra of Random Variables* (Wiley, New York, 1979).

[10] S. Yellin, Phys. Rev. D **66**, 032005 (2002).

[11] L. Shtembari and A. Caldwell, preceding paper, Phys. Rev. D **108**, 123005 (2023).

[12] L. Shtembari, SPACINGSTATISTICS.JL, https://github.com/bat/SpacingStatistics.jl/tree/dev.

[13] J. Abraham *et al.*, Nucl. Instrum. Methods Phys. Res., Sect. A **523**, 50 (2004).

[14] T. P. A. Collaboration, Pierre auger observatory open data, 2022, https://zenodo.org/records/6867688.

[15] P. Abreu *et al.*, Astrophys. J. **935**, 170 (2022).

[16] P. Sommers, Astropart. Phys. **14**, 271 (2001).

[17] A. Aab *et al.* (The Pierre Auger Collaboration), Phys. Rev. D **102**, 062005 (2020).

[18] L. Shtembari, On non-parametric tests for discovery and limit setting in one and multiple dimensions, Ph.D. thesis, Technische Universität München, 2023.