Robust test statistics for data sets with missing correlation information

Lukas Koch*

Particle Physics Department, University of Oxford, DWB, Keble Road, OX1 3RH Oxford, United Kingdom

(Received 12 February 2021; accepted 25 May 2021; published 21 June 2021)

Not all experiments publish their results with a description of the correlations between the data points. This makes it difficult to do hypothesis tests or model fits with that data, since just assuming no correlation can lead to an overestimation or underestimation of the resulting uncertainties. This work presents robust test statistics that can be used with datasets with missing correlation information. They are exact in the case of no correlation and either guaranteed to be conservative—i.e., the uncertainty is never underestimated—in the presence of correlations, or they are also exact in the degenerate case of perfect correlation between the data points.

DOI: 10.1103/PhysRevD.103.113008

I. INTRODUCTION

Some datasets are published without a full covariance matrix, describing the correlations between the data points of the result. The implied assumption in these datasets is that the correlation in the uncertainties is 0, i.e., the data is uncorrelated. This is not always the case though,[1] and users of the data are put in the unenviable situation of having to use correlated results, without knowing what the correlations actually are. Usually one would use the fully correlated Mahalanobis distance [1,2] or its square, $D^2 = \boldsymbol{\Delta}^T S^{-1} \boldsymbol{\Delta}$, to judge how well a certain model fits the data.[2] Here $\boldsymbol{\Delta}$ is the difference between the data and the model prediction, and S is the covariance matrix describing the uncertainty of the result. Just ignoring the correlations and applying a "naive," uncorrelated Mahalanobis distance to compare a model to the data can lead to plainly wrong results.

Consider multivariate normal distributed data with ten dimensions. In fact, unless otherwise stated, let us assume that the examples in this paper are all multivariate normal distributed and we know the correct diagonal elements of the respective covariance matrices. The only problem we will address here is the missing of information regarding the correlations between the variables. The naive test statistic would consist of just summing up the squared z-scores, i.e., the residuals normalized by the uncertainty:

$$\text{naive}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{s}) = \sum_i \frac{(x_i - \mu_i)^2}{S_{ii}}, \tag{1}$$

where $\boldsymbol{x}$ is the data result, $\boldsymbol{\mu}$ is a prediction of the expectation value from some model, and $S_{ii} = s_i^2$ is the variance of the data points. This test statistic will be chi-square distributed if there are no correlations present in the data (see any introductory statistics text book, e.g., Chapter 6 of [3]), but using it in the presence of correlations leads to underestimation or overestimation of uncertainty, depending on the correlation and the actual value of the statistic. Figure 1 shows this for ten-dimensional toy datasets thrown with different levels of correlation. The diagonal terms of the covariance matrix are kept constant at 1, while the off diagonals are set to the values 0, 0.5, 0.9, and 0.99, in the different sets.

The left plot shows the cumulative probability density functions (CDFs) of the expected distribution of the test statistic in the absence of correlations, as well as the actual CDFs of the different toy datasets. The different distributions affect which significance level (or p-value) a certain value of the test statistic corresponds to. The right plot shows how an assumed significance level (as calculated with the expected CDF) translates to the actual significance level (as calculated from the actual CDFs). To put it another way: The x axis shows how often one would like to make a type-I error (rejecting a true hypothesis), while the y axis shows how often one actually makes a type-I error, given the different levels of correlation in the data. If the actual significance level is larger than assumed, one rejects a true hypothesis more often than intended. In terms of error bars or confidence regions, this means that the size of the uncertainties is effectively underestimated.

This behavior is clearly undesirable. If the data is (suspected to be) strongly correlated, it would be better

*lukas.koch@physics.ox.ac.uk

[1]If, e.g., the data points vary smoothly within their error bands or when they are supposed to describe a "shape-only" uncertainty, it is clear that there is a correlation there.

[2]In the particle physics community, this is often simply called "the chi-square." To avoid confusion with other chi-square distributed test statistics, it is useful to use its proper name, though.

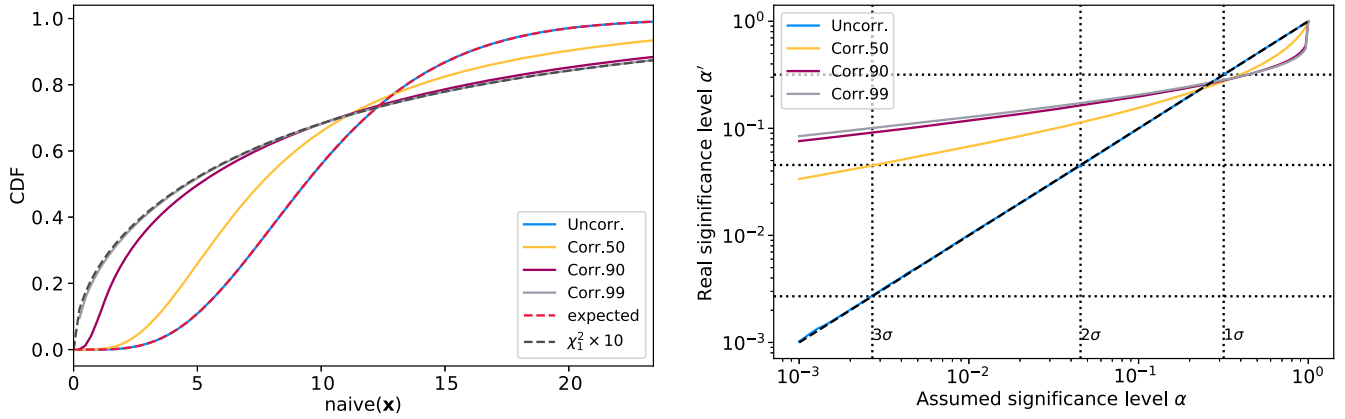


FIG. 1. CDFs (left) for the naive test statistic for different levels of correlations in the data. When using the uncorrelated CDF to calculate the assumed significance level (or p-value) of a value of the statistic, the actual level will differ from the assumption depending on the correlations (right). As the correlation increases, the distribution of the naive test statistic approaches that of a χ^2_1 distributed variable which is multiplied by the number of bins (in this case ten).

to use a different test statistic that is able to perform consistently under different levels of correlation in the data. For such a test statistic, the following properties would be desirable:

1. Exact in the case of no correlations.
2. Guaranteed to be conservative when not exact.
3. Low deviations from exactness when not exact.
4. Exact in the case of 100% correlation.
5. Exact at every possible level of correlation.

Some of these properties are contained in one another. The naive use of the uncorrelated Mahalanobis distance has property 1 but none of the others. The following sections will describe some test statistics that have more of these properties. After that, we will compare them in Sec. IV, and apply them to some real experimental results from neutrino scattering experiments in Sec. V.

II. FITTING THE COVARIANCE TO THE DATA

The problem of estimating both the mean and covariance of multivariate normal distributed data has been extensively discussed in statistical literature (see e.g., [4] and references therein). This includes work on estimators for the covariance when the number of observations is smaller than the number of dimensions of the data space [5]. Unfortunately, the problem addressed here is somewhat unique, since we only have access to a single observation from the distribution we would like to estimate. This observation is the published result. We cannot assume that the models we try to test are drawn from the same distribution, since this is exactly the hypothesis we want to test. Another difference to the widely discussed case—this time in our favor—is that we can assume to know the diagonal elements of the covariance matrix as well as the mean values of the distribution. Also, we are not interested in the actual values of the full covariance matrix, as long as we can construct a test statistic that performs well without knowing these values.

The first considered test statistic thus arises from treating the off-diagonal elements of the covariance matrix as nuisance parameters of the statistical model. For any given predicted mean value in the N-dimensional data space $\boldsymbol{\mu}$ and a given sample (i.e., the data) $\boldsymbol{x}$, it is possible to choose the off-diagonal elements of S in a way to minimize the resulting squared Mahalanobis distance $D^2 = (\boldsymbol{x}-\boldsymbol{\mu})^T S^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$. This is different from maximizing the likelihood of the data, since the probability density of a multivariate normal distribution also depends on the determinant of the covariance matrix:

$$L = (2\pi)^{-\frac{N}{2}} \det(S)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}} S^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \tag{2}$$

A minimal Mahalanobis distance is only equivalent to a maximal likelihood if the determinant of the covariance matrix is constant. This is not the case here. Furthermore, for $N \geq 3$ the supremum of the likelihood in a maximization over the covariance elements is always $+\infty$, rendering it useless as a test statistic. This will be shown below.

Since the Mahalanobis distance is invariant under a linear transformation of the variables [2,6], we can simplify the minimization by transforming the variable space to make the last variable x_N independent of the others:

$$\boldsymbol{y} = \begin{pmatrix} \mathbf{1} & -\frac{S_N}{S_{NN}} \\ 0 & 1 \end{pmatrix} \boldsymbol{x} \tag{3}$$

$$= \begin{pmatrix} 1 & 0 & \dots & 0 & -\frac{S_{1N}}{S_{NN}} \\ 0 & 1 & \dots & 0 & -\frac{S_{2N}}{S_{NN}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -\frac{S_{(N-1)N}}{S_{NN}} \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} \boldsymbol{x}. \tag{4}$$

Here S_{NN} is the variance of the Nth variable, and $\boldsymbol{S}_N$ is the vector of the $N-1$ covariances between the Nth and the other variables:

$$\boldsymbol{S}_N = \begin{pmatrix} S_{1N} \\ \vdots \\ S_{(N-1)N} \end{pmatrix}. \tag{5}$$

The covariance and expectation values for $\boldsymbol{y}$ are then

$$S^y = (\boldsymbol{\nabla y}^T)^T S(\boldsymbol{\nabla y}^T) \tag{6}$$

$$= \begin{pmatrix} \mathbf{1} & -\frac{\boldsymbol{S}_N}{S_{NN}} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} S^{/N} & \boldsymbol{S}_N \\ \boldsymbol{S}_N^T & S_{NN} \end{pmatrix} \begin{pmatrix} \mathbf{1} & 0 \\ -\frac{\boldsymbol{S}_N^T}{S_{NN}} & 1 \end{pmatrix} \tag{7}$$

$$= \begin{pmatrix} S^{/N} - \frac{\boldsymbol{S}_N \boldsymbol{S}_N^T}{S_{NN}} & 0 \\ \boldsymbol{S}_N^T & S_{NN} \end{pmatrix} \begin{pmatrix} \mathbf{1} & 0 \\ -\frac{\boldsymbol{S}_N^T}{S_{NN}} & 1 \end{pmatrix} \tag{8}$$

$$= \begin{pmatrix} S^{/N} - \frac{\boldsymbol{S}_N \boldsymbol{S}_N^T}{S_{NN}} & 0 \\ 0 & S_{NN} \end{pmatrix}, \tag{9}$$

$$\boldsymbol{\mu}^y = \begin{pmatrix} \mathbf{1} & -\frac{\boldsymbol{S}_N}{S_{NN}} \\ 0 & 1 \end{pmatrix} \boldsymbol{\mu}, \tag{10}$$

where $S^{/N}$ is the original covariance matrix for the remaining $N-1$ variables.

The contribution of $y_N = x_N$ to the total Mahalanobis distance $D^2 = (\boldsymbol{y} - \boldsymbol{\mu}^y)^T (S^y)^{-1} (\boldsymbol{y} - \boldsymbol{\mu}^y)$ is fixed, since it only depends on S_{NN}, which has a given constant value. The contribution of the remaining variables $\boldsymbol{y}^{/N}$ could be minimized to 0 by choosing the off-diagonal elements of the covariance such that their expectation value $\boldsymbol{\mu}^{y/N}$ is equal to the actual value:

$$\boldsymbol{\mu}^{y/N} \stackrel{!}{=} \boldsymbol{y}^{/N} \tag{11}$$

$$\boldsymbol{\mu}^{/N} - \frac{\boldsymbol{S}_N}{S_{NN}} \mu_N = \boldsymbol{x}^{/N} - \frac{\boldsymbol{S}_N}{S_{NN}} x_N \tag{12}$$

$$\boldsymbol{S}_N = S_{NN} \frac{\boldsymbol{x}^{/N} - \boldsymbol{\mu}^{/N}}{x_N - \mu_N} \tag{13}$$

$$= \begin{pmatrix} \frac{\Delta_i}{\sqrt{S_{ii}}} \frac{\sqrt{S_{NN}}}{\Delta_N} \sqrt{S_{ii} S_{NN}} \\ \vdots \\ \frac{\Delta_{N-1}}{\sqrt{S_{(N-1)(N-1)}}} \frac{\sqrt{S_{NN}}}{\Delta_N} \sqrt{S_{(N-1)(N-1)} S_{NN}} \end{pmatrix}. \tag{14}$$

Here $\Delta_i/\sqrt{S_{ii}} = (x_i - \mu_i)/\sqrt{S_{ii}}$ is the (positive or negative) z-score of the ith variable, and $\sqrt{S_{ii} S_{NN}}$ is the maximum allowed absolute value of the covariance between the ith and Nth variable. The latter arises from the fact that the correlation coefficients $S_{ij}/\sqrt{S_{ii} S_{NN}}$ must be within $[-1, 1]$. The vector of covariances $\boldsymbol{S}_N$ is thus the ratio of the $N-1$ z-scores over the Nth z-score, multiplied by the maximum allowed value for each covariance. This is a valid choice of covariances when the Nth variable has the largest absolute z-score, meaning that all z-score ratios are within $[-1, 1]$.

We can always reorder the variables such that the Nth is the one with the largest absolute z-score. Thus, by eliminating the contribution of the other variables as shown above, the minimal achievable Mahalanobis distance under variation of the off-diagonal elements of the covariance matrix is equal to the largest absolute z-score of the single variables. This behavior is illustrated in Fig. 2 for two dimensions. Note that we only need $N-1$ of the $N(N-1)/2$ covariance parameters to ensure the value of the Mahalanobis distance. The remaining elements can be chosen freely as long as they result in a valid covariance matrix. In fact, they could be chosen in a way to make the determinant of S arbitrarily

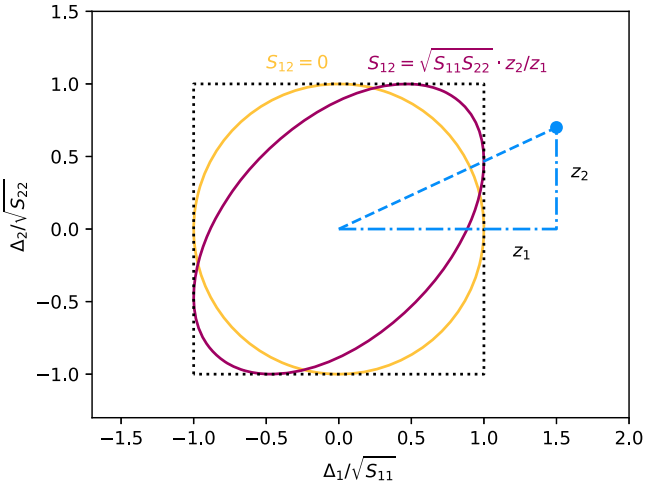


FIG. 2. Minimum achievable Mahalanobis distance for two dimensions. The minimum achievable Mahalanobis distance when varying the off-diagonal covariance element is equal to the largest absolute z-score of the single variables. The surface where the Mahalanobis distance is equal to 1 is an ellipse contained within the square with its edges at $\Delta_i/\sqrt{S_{ii}} = \pm 1$. Varying the off-diagonal element of the covariance matrix does not rotate the principal axes of the ellipse, but it changes where it touches the edges of the square. When chosen correctly, the ellipse touches the edge of the box at the point where the data is projected onto it. Because of the linearity of the Mahalanobis distance, this means that the total distance of the data point is then simply the largest z-score. This two-dimensional minimization can be done for all marginal projections of pairs of variables in N-dimensional problems. To achieve the minimal total Mahalanobis distance, only the pairs involving the overall largest absolute z-score need to be specified like this, but applying the scheme to all pairs ensures a valid covariance matrix.

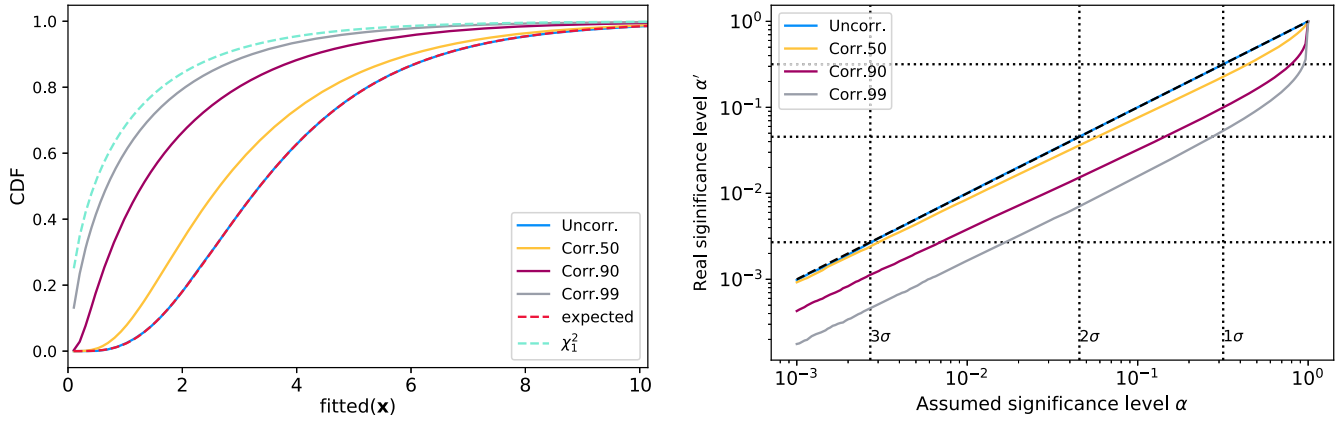


FIG. 3. CDFs (left) for the fitted test statistic for different levels of correlations in the data. When using the uncorrelated CDF to calculate the assumed significance level (or p-value) of a value of the statistic, the actual level will differ from the assumption depending on the correlations (right). In the presence of correlations, the real significance is consistently higher (the significance level is lower) than the assumption. This means the uncertainties are overestimated and the statistic behaves conservatively. As the correlations increase, the distribution of the fitted test statistic approaches the χ^2_1 distribution.

small, leading to the infinite supremum of the likelihood maximization over the covariance elements for $N \geq 3$.

Since the minimum achievable Mahalanobis distance is always equal to the maximum absolute z-score among the variables, no actual fitting or optimization needs to be done for this test statistic. Let us call b the largest absolute z-score, and we can define the "fitted" test statistic as

$$\text{fitted}(\boldsymbol{\Delta}|\boldsymbol{s}) = b^2 = \max_i \left(\frac{\Delta_i^2}{S_{ii}}\right). \tag{15}$$

It is straightforward to derive the expected distribution of this test statistic in the case of no correlations. The CDF of b, $F_b(b')$, is just the probability of the absolute values of *all* z-scores being smaller than or equal to b':

$$\begin{aligned} F_b(b') &= P(b \leq b') \\ &= \int_{-b'}^{+b'} \cdots \int_{-b'}^{+b'} f(z_1, \ldots, z_N) \mathrm{d}z_1 \ldots \mathrm{d}z_N, \end{aligned} \tag{16}$$

where $f(\boldsymbol{z})$ denotes the probability density function (PDF) of the potentially negative z-scores. With uncorrelated, standard normal distributed z-scores this evaluates to

$$F_b(b') = \text{erf}^N\left(\frac{b'}{\sqrt{2}}\right). \tag{17}$$

With this we can write down the CDF of b^2 as

$$F_{b^2}(y) = F_b(\sqrt{y}), \tag{18}$$

which defines the distribution of the fitted test statistic. We will call the distribution the "Bee-square" distribution (as a nod to the chi-square distribution) and we have

$$\text{fitted}(\boldsymbol{\Delta}|\boldsymbol{s}) \sim \text{Bee}^2_N, \tag{19}$$

for uncorrelated normal distributed Δ_i. A Python implementation of the distribution can be found in Listing 1 in the Appendix.

Figure 3 shows how this test statistic fares for different levels of correlation in the toy data. For no correlations, the distribution follows the expectation. With increasing correlations, the distribution deviates more and more, approaching a chi-square distribution with 1 degree of freedom.[3] Compared to the naive approach, we can see that the deviation from an exact statistic has been decreased for a wide range of significance levels. But more importantly, the fitted test statistic is conservative for all significance levels and all correlation strengths. The real significance level of a result is always equal to or lower than the assumed significance that was evaluated using the expected Bee-square distribution.[4] A proof of this for the two-dimensional case can be found in Appendix.

III. ASYMPTOTICALLY INVARIANT TEST STATISTICS

The fitted test statistic described above is "safe" to use in the sense that it is always conservative. Unfortunately it gets more and more conservative with increasing correlations in the data. It would be advantageous if the test statistic was exact at all levels of correlations, or at least at both no correlations, and (in the limit of) 100%

[3]As the data gets more and more correlated, the z-scores will approach being equal in all cases, and the maximum z-score will be distributed like a single standard normal distributed variable.

[4]That is, the probability of a result at least as extreme as the observed is actually lower than what the assumed distribution suggests; the uncertainty is overestimated.

correlations. To achieve the latter, it is useful to view the problem in the "CDF space" of the data points.

Instead of the distribution of $\mathbf{\Delta}$, let us consider the CDFs of the squares of the *single* variables:

$$y_i = F_{\chi_1^2}(\Delta_i^2/S_{ii}), \tag{20}$$

where $F_{\chi_1^2}$ is the CDF of Δ_i^2/S_{ii}, since we assume Δ_i to be normal distributed with a variance of S_{ii}. Since y is a function of a random variable, it is itself a random variable. Also, by definition, y_i is uniformly distributed between 0 and 1:

$$y_i \sim \mathrm{U}(0,1). \tag{21}$$

This is true irrespective of the possible correlations between the data points, as long as the marginal distribution of each single data point is known. In fact, the single data points do not have to be normal distributed. If they follow a different (but known) distribution, its CDF can be substituted in Eq. (20).

If and only if the different variables are independently distributed, the combined probability density of all y_i, $f_y(\boldsymbol{y})$ will also be uniform within the N-cube defined by the N unit vectors:

$$f_y(\boldsymbol{y}) = \begin{cases} 1 & \text{if } 0 \le y_i \le 1 \ \ \forall\, i \\ 0 & \text{else.} \end{cases} \tag{22}$$

This follows from simply multiplying the PDFs of the single variables y_i.

If, on the other hand, the variables are perfectly correlated, the values of all y_i will be identical in each random sampling. This means the *combined* PDF must be zero wherever the y_i are not identical. In this case, the combined probability density will be a delta function that concentrates all probability on the main diagonal of the hypercube:

$$f'_y(\boldsymbol{y}) = \begin{cases} \prod_{i=1}^{N-1} \delta(y_i - y_{i+1}) & \text{if } 0 \le y_1 \le 1 \\ 0 & \text{else.} \end{cases} \tag{23}$$

Note that this does not affect the marginal distributions of the single variables. Marginalizing out all but one variable leads again to a uniform distribution in that variable. If we can define a function $z(\boldsymbol{y})$ that is identically distributed under both assumptions, we can use it to define a test statistic that is exact in both cases.

Let us demand that a low value of z indicates a good agreement between data and model, while high values indicate tension between the two. Within the N-dimensional hypercube, this means that z should be low towards the corner at $\boldsymbol{y} = \mathbf{0}$ and increase towards the corner at $\boldsymbol{y} = \mathbf{1}$. For z to be identically distributed with no correlations and with 100% correlations, the surface defined by the implicit function $z(\boldsymbol{y}) = z' = \text{const}$ must enclose the same amount of probability in both cases, as this defines the CDF of z:

$$F_z(z') = \int_{z(\boldsymbol{y}) \le z'} f_y^{(\prime)}(\boldsymbol{y}) \mathrm{d}^N \boldsymbol{y}. \tag{24}$$

In the case of no correlation, this is the volume A of the part of the N-cube enclosed by the implicit function. In the case of perfect correlation, it is equal to the single (identical) y_i coordinates where the surface intersects with the diagonal. Figure 4(a) illustrates this in the case of two dimensions.

Let us call this coordinate x and let us also demand that the function $z(\boldsymbol{y}) = x$ at that point. We then get the following condition:

$$A = \int_{z(\boldsymbol{y}) \le x} \mathrm{d}^N \boldsymbol{y} \overset{!}{=} x, \tag{25}$$

where the integral is understood to be confined to the inside of the N-cube. The challenge is now to find functions $z(\boldsymbol{y})$ which fulfill this condition for any number of dimensions.

A. Invariant 1

The simplest way to fulfill Eq. (25) in two dimensions is to draw straight lines from the points on the diagonal to the "off-diagonal" corners of the square, as shown in Fig. 4(b). We can easily calculate the value of $z(\boldsymbol{y})$ for any given point, as each point can be seen as lying on a straight diagonal line starting at the lower or left edge of the square ($\boldsymbol{y}_s$) and ending on the right or top edge ($\boldsymbol{y}_e$). The fractional distance along this line is the desired z and evaluates to

$$z(\boldsymbol{y}) = \frac{y_{\min}}{y_{\min} + (1 - y_{\max})}, \tag{26}$$

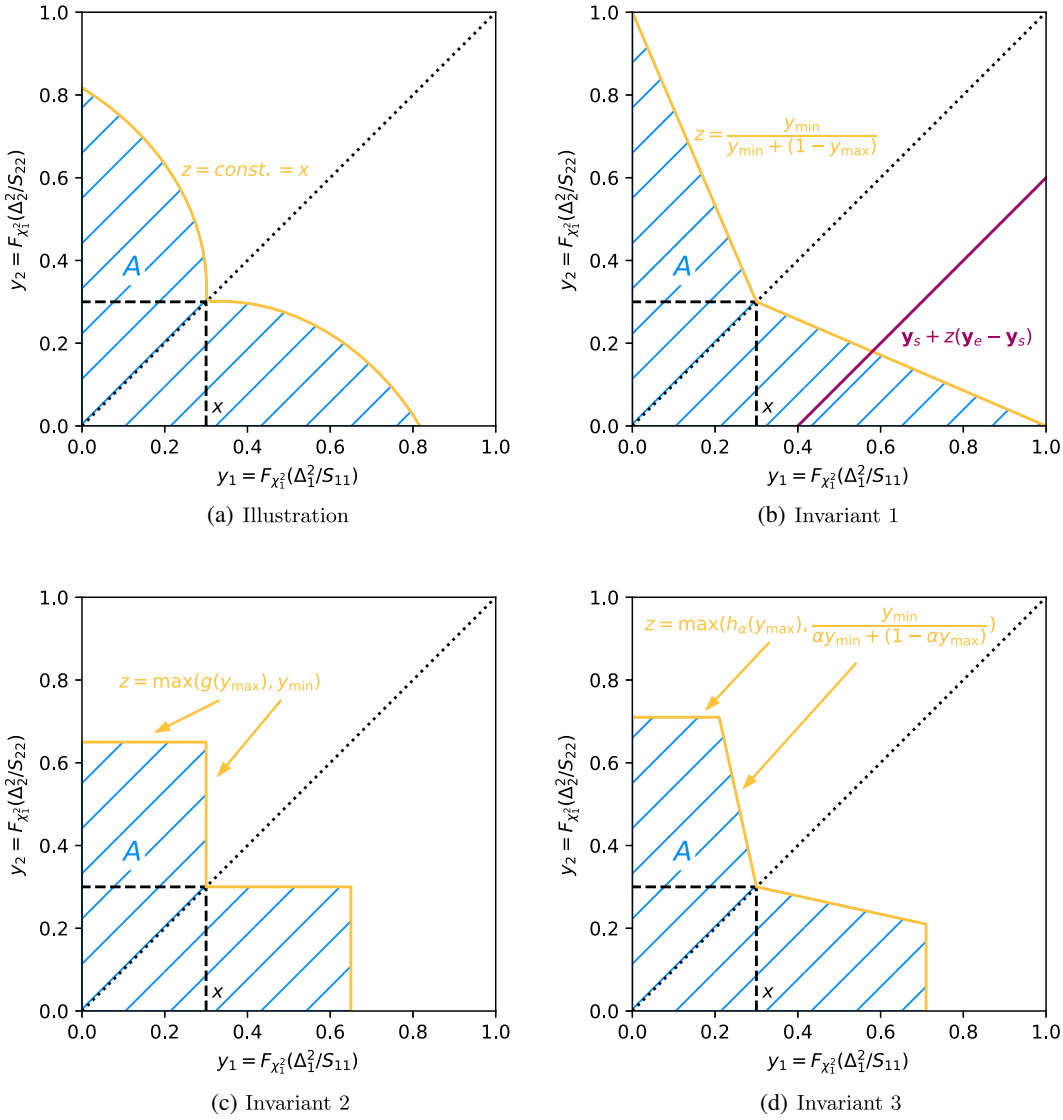
where $y_{\min\,/\,\max}$ are the minimum and maximum of the elements of $\boldsymbol{y}$ respectively. This also directly applies to the N-dimensional case without change.

Now, it would be possible to use z as the test statistic directly when used on its own. When the data is intended to be used in conjunction with other datasets though, e.g., in a global fit, it is useful to use a test statistic that is (approximately) chi-square distributed. To this end, we can simply apply another function to z which is chosen so that the result is chi-square distributed if z is uniformly distributed. This function is just the inverse of the CDF of the chi-square distribution $\overset{-1}{F}_{\chi_1^2}$. Finally we get the first of the "invariant" test statistics:

$$\text{invariant}_1(\mathbf{\Delta}|\boldsymbol{s}) = \overset{-1}{F}_{\chi_1^2}(z(F_{\chi_1^2}(\Delta_i/\sqrt{S_{ii}}), \ldots)), \tag{27}$$

with z as defined in Eq. (26).

Note that we could have chosen a different number of degrees of freedom for the transformation back to



(a) Illustration (b) Invariant 1

(c) Invariant 2 (d) Invariant 3

FIG. 4. Illustration of the test statistics in CDF space. For z to be identically distributed at both no correlations (PDF of $\boldsymbol{y}$ is uniform over square) and at 100% correlations (PDF of $\boldsymbol{y}$ is uniform along diagonal), the area A enclosed by the implicit function $z = \text{const}$ must be equal to the position x where the function crosses the diagonal.

a chi-square distribution. This makes no difference when using the test statistic on a single dataset alone, but it changes the relative weight a dataset has in a global fit with other data when computing the total chi-square. If one is confident that the correlations in the dataset are weak, it might be better to use the actual number of data points. In the presence of medium to strong correlations it could be argued though that there is actually less information in the data than the number of data points suggests, or rather, we are losing "degrees of freedom" by having to make up for the missing information of the covariance parameters.

Figure 5 shows the performance of the test statistic. It does deviate from being exact in the presence of correlations, but the deviation peaks at a certain level and from then on it get more exact again when the correlations are further increased. Unfortunately it is not conservative for low significance levels.

B. Invariant 2

Another simple solution in two dimensions is to add identical rectangles to the two inside-facing sides of the x^2 square, as shown in Fig. 4(c). Let l be the length of the added rectangles:

$$l = \frac{1 - x}{2}. \tag{28}$$

Extended into N dimensions, the shape of A becomes a hypercube H_1 with an edge length of $x + l$, minus another hypercube H_2 at the inside diagonal corner with an edge length of l:

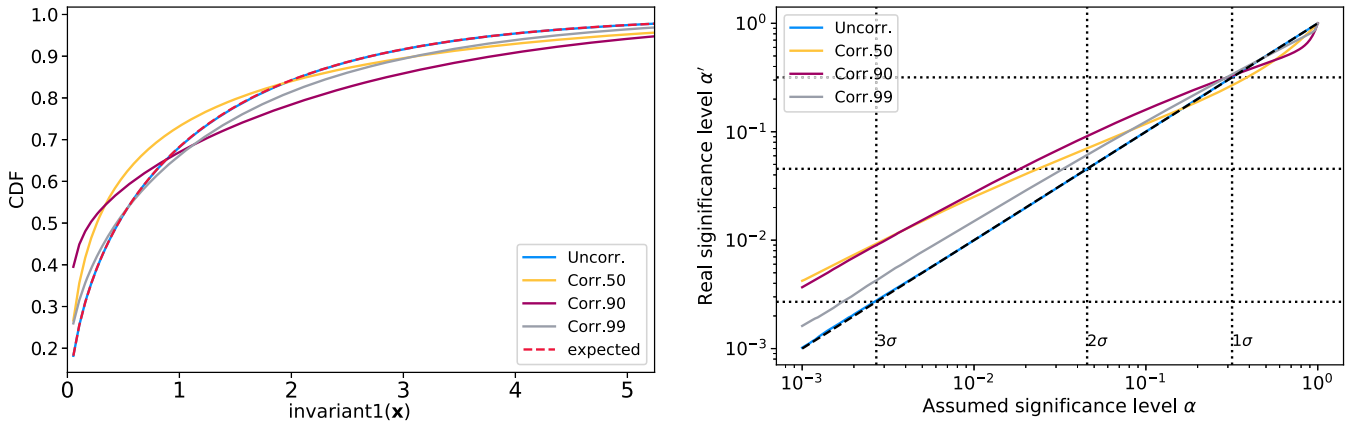


FIG. 5. CDFs (left) for the "invariant 1" test statistic for different levels of correlations in the data. When using the uncorrelated CDF to calculate the assumed significance level (or p-value) of a value of the statistic, the actual level will differ from the assumption depending on the correlations (right). The effect of the correlations is weaker than for the naive test statistic, but it is not consistently conservative as the fitted one. As the correlations increase, the distribution of the test statistic approaches the uncorrelated expectation again.

$$A = (x+l)^N - l^N \stackrel{!}{=} x. \tag{29}$$

Expressed in the total width of the larger cube $d = x + l$ we get

$$d^N - l^N \stackrel{!}{=} d - l. \tag{30}$$

This equation will always have a solution of $l \in (0, d)$ for any $d \in (N^{-1/(N-1)}, 1)$, and can be solved numerically. Let $l(d)$ be that solution, so we can define the function $g\colon [0,1] \to [0,1]$:

$$g(d) = \begin{cases} 0 & \text{if } d \leq N^{-1/(N-1)} \\ d - l(d) & \text{if } N^{-1/(N-1)} < d < 1 \\ 1 & \text{if } d = 1, \end{cases} \tag{31}$$

which can calculate x from a given (possible) edge length d of H_1.

To calculate the z of any given point, we can use that every point on the surface of $(H_1 \backslash H_2)$ is either on one of the "outer" faces of H_1 (where all coordinates are >0), or one of the "inner" faces of H_2 (where at least one coordinate is x). In the former case, the edge length of H_1 is given by the maximum of the y-coordinates, while in the latter case the position of the inner diagonal corner of H_2 is given by the minimum of the coordinates. Since that inner corner is at $y_i = x \ \ \forall \ \ i$ by construction, we can write z as

$$z(\boldsymbol{y}) = \max(g(y_{\max}), y_{\min}). \tag{32}$$

With this z we can then define the $\text{invariant}_2(\boldsymbol{\Delta}|\boldsymbol{s})$ test statistic just like in Eq. (27).

Its performance is shown in Fig. 6. It is conservative for all strengths of correlation and significance levels, and shows the expected limit of exactness at no and very strong correlations.

C. Invariant 3

Finally, Fig. 4(d) shows an intermediate shape between invariant 1 and invariant 2. It can be interpreted as the shape of invariant 1, but instead of connecting the diagonal to the off-diagonal corners of the square, it is connected to the respective corners of a larger square with edge length $1/\alpha$, with the shape parameter $\alpha \in (0, 1)$. To ensure that the resulting area A is equal to x, it is cut off at the edges of a square with edge length d.

In N dimensions, the volume of such a body is

$$A = d^N - \frac{(d-x)^N}{(1-\alpha x)^{N-1}} \stackrel{!}{=} x. \tag{33}$$

This equation has one solution of $x \in (0, d)$ if $(N - \alpha dN + \alpha d) > d^{1-N}$. Let $x(d)$ be that solution and, like before, we can define a function $h_\alpha\colon (0,1) \to (0,1)$:

$$h_\alpha(d) = \begin{cases} 0 & \text{if } (N - \alpha dN + \alpha d) \leq d^{1-N} \\ x(d) & \text{if } (N - \alpha dN + \alpha d) > d^{1-N} > 1 \\ 1 & \text{if } d = 1. \end{cases} \tag{34}$$

This determines the value of z for points on the surface of the hypercube.

The value of z for points on the "cut-off" corner can be calculated just like in the case of invariant 1. Only the scaling of the containing cube needs to be taken into account. The total z function is then again the maximum of the two values:

$$z(\boldsymbol{y}) = \max\left(h_\alpha(y_{\max}), \frac{y_{\min}}{\alpha y_{\min} + (1 - \alpha y_{\max})}\right). \tag{35}$$

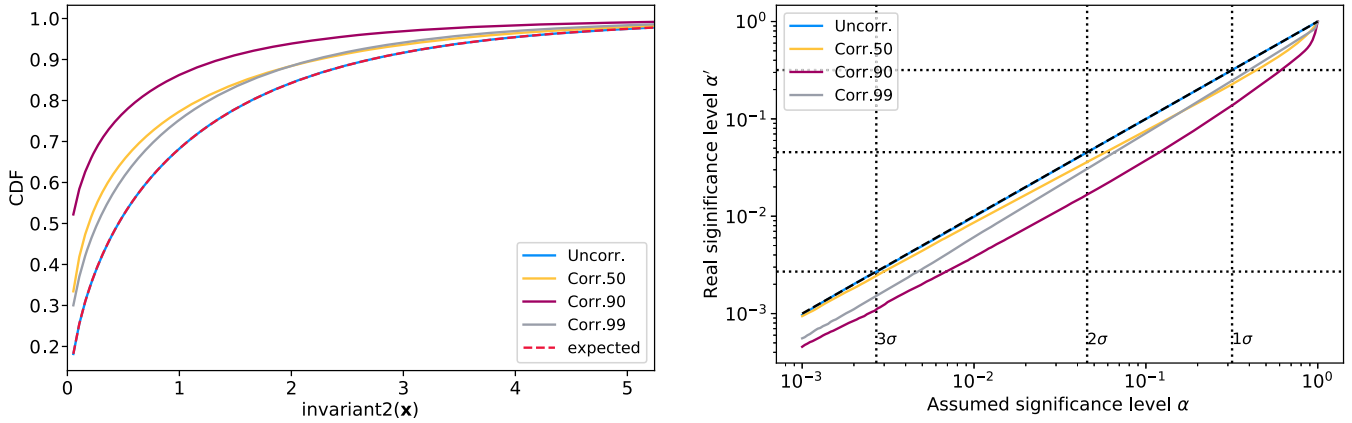


FIG. 6. CDFs (left) for the "invariant 2" test statistic for different levels of correlations in the data. When using the uncorrelated CDF to calculate the assumed significance level (or p-value) of a value of the statistic, the actual level will differ from the assumption depending on the correlations (right). Like the fitted test statistic, this one is consistently conservative. As the correlations increase, the distribution of the test statistic approaches the uncorrelated expectation again.

The final test statistic "invariant 3" is then built according to Eq. (27). A Python implementation of this test statistic is provided in Listing 2 in the Appendix.

This test statistic has one free shape parameter α. It determines the opening angle of the iso-z surface where it meets the main diagonal of the hypercube. In the 2D case, this angle is constant at 90° for $\alpha \to 0$, making it identical to the invariant 2 case. For $\alpha > 0$, the angle starts at 90° at $x = 0$ and then opens up with increasing x. The value of α determines where the angle reaches 180°:

$$x_{180°} = \frac{1}{2\alpha}. \tag{36}$$

For example, for $\alpha = 1$, corresponding to the invariant 1 case, the opening angle is 180° at $x = 0.5$.

In the presence of "medium" correlations, the opening angle can give an indication of whether the test statistic is conservative for the corresponding significance level. An opening angle <180° means that the surface "protrudes" into parts of the CDF space that should "belong" to a higher value of x, suggesting a conservative statistic. Conversely, an opening angle >180° means that the CDF space perpendicular to the diagonal is only covered with higher x, suggesting a coverage that is actually lower than the expectation. At $\alpha = 0.5$, the opening angle is <180° for all x (as x is always <1). This makes that value a conservative choice.

There is room for a more aggressive choice of α though. Figure 7 shows the performance of the invariant 3 test statistic with $\alpha = 2/3$. Out of all considered test statistics, it shows the smallest deviations from exactness, and it is

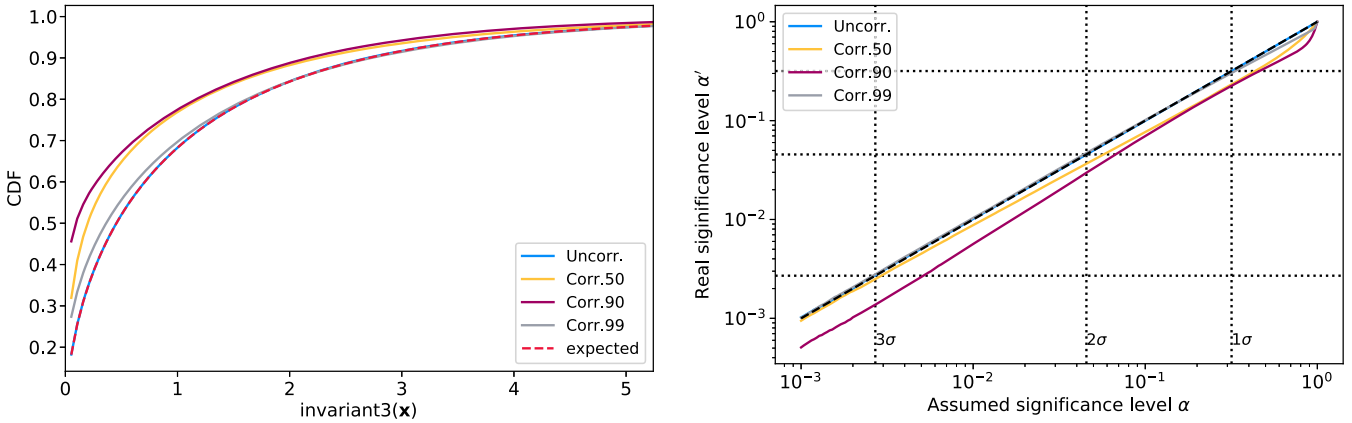


FIG. 7. CDFs (left) for the invariant 3 test statistic for different levels of correlations in the data. When using the uncorrelated CDF to calculate the assumed significance level (or p-value) of a value of the statistic, the actual level will differ from the assumption depending on the correlations (right). The behavior of this statistic depends on the parameter α. For $\alpha = 1$ it is identical to the invariant 1 statistic, while for $\alpha \to 0$ it approaches invariant. Shown here is $\alpha = 2/3$.

conservative for the considered significance levels and correlations.

IV. COMPARISON

It is quite clear that among the considered test statistics, invariant 3 performs the most consistent under many different levels of correlations. It suffers from a kind of arbitrariness though when trying to combine it in larger fits with other datasets. The transformation to a chi-square distribution with 1 degree of freedom is a conservative choice that allows it to be combined in least-squares or likelihood fits. Aside from the case of perfect correlations, it will underestimate the amount of information compared to the other datasets though.

The fitted statistic is quite a bit more conservative and it gets more conservative the stronger the correlations are. It has the advantage though that it corresponds to an "actual Mahalanobis distance" when considering the correlations in the data as nuisance parameters. It should thus easily be included in least-square fits in combination with other datasets. Wilks' theorem does *not* hold for it though, except for very strong correlations. Only then is it distributed like a chi-square with 1 degree of freedom. With no correlations present, it follows the "Bee-square" distribution.

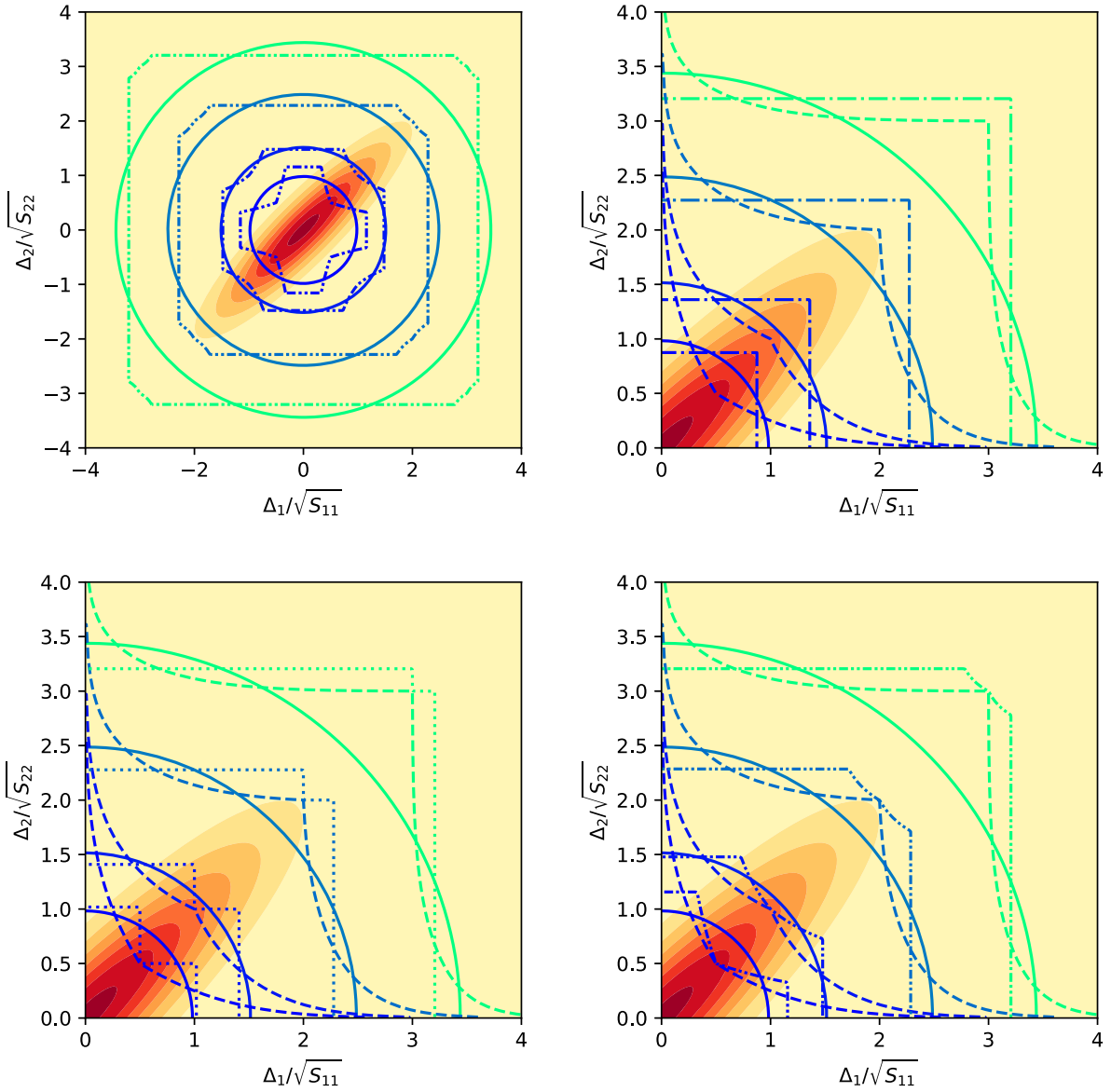


FIG. 8. Shape of confidence regions of the different statistics in two dimensions. The confidence levels correspond to a 0.5σ, 1σ, 2σ, and 3σ deviation of a one-dimensional normal distributed variable. The shown test statistics are: top left, naive (solid) and invariant 3 (dash-dot-dot); top right, naive (solid), fitted (dash-dot), invariant 1 (dashed); bottom left, naive (solid), invariant 1 (dashed), invariant 2 (dotted); bottom right, naive (solid), invariant 1 (dashed), invariant 3 ($\alpha = 2/3$) (dash-dot-dot). Also shown is the distribution of some multivariate normal data with a variance of 1.0 and a correlation of 0.9.

Figure 8 shows the shape of the resulting confidence regions for the different test statistics in two dimensions. The invariant statistics all have a cross shape at low confidence levels and then get progressively more square. The confidence regions of the invariant 1 statistic extend all the way to $\pm\infty$ along the variable axes. This is due to the fact that all lines of the construction in the CDF space go to the edges of the hypercube. For a normal distributed variable a CDF of 1 means a variable value of $\pm\infty$. The coverage is still correct in the case of no correlations, but in principle this means that one would have to include a point in the 1σ region that is arbitrarily far away on one axis, as long as it is close enough to 0 in the other axis. This is clearly counterintuitive and could very well lead to strange behavior in fits. Combined with the fact that it is not conservative for all confidence levels, the invariant 1 statistic should probably not be used.

It is worth stressing that the actual coverage behavior of the test statistics will depend on the actual correlations present in the datasets. The examples shown here were very simple, with constant correlation coefficients between all data points. Figure 9 shows the performance of the invariant 3 test statistic for a dataset with a more complicated correlation structure. The data consists of $N = 100$ data points with a covariance matrix of the form

$$S = \begin{pmatrix} 1 & c_0 & c_1 & \cdots & c_{N-2} \\ c_0 & 1 & c_0 & \cdots & c_{N-3} \\ c_1 & c_0 & 1 & \cdots & c_{N-4} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{N-2} & c_{N-3} & c_{N-4} & \cdots & 1 \end{pmatrix}. \quad (37)$$

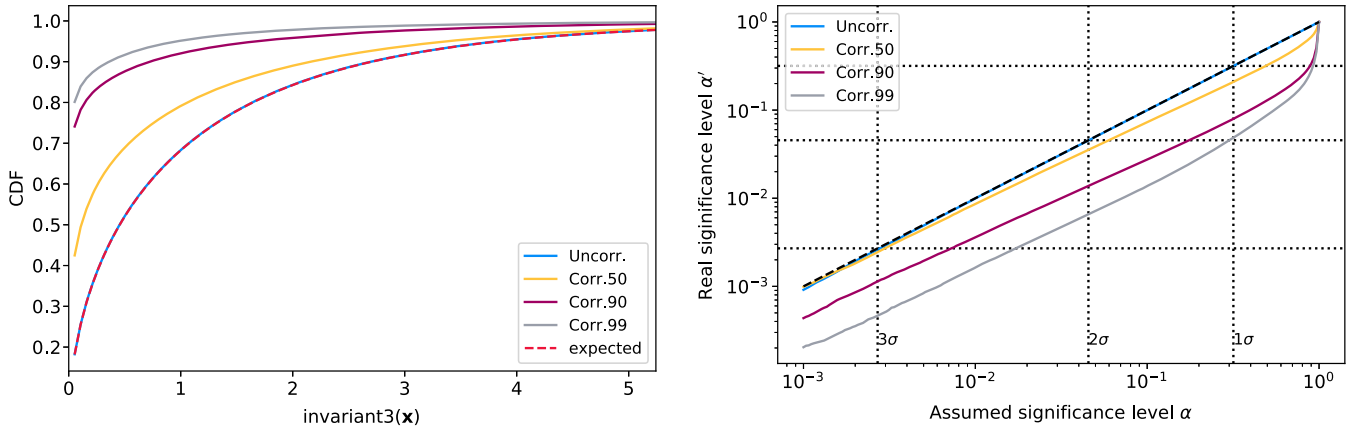


FIG. 9. CDFs (left) for the invariant 3 test statistic for different levels of correlations in data with a more complicated correlation structure (see text). When using the uncorrelated CDF to calculate the significance level (or p-value) of a value of the statistic, the actual level depends on the correlations (right).

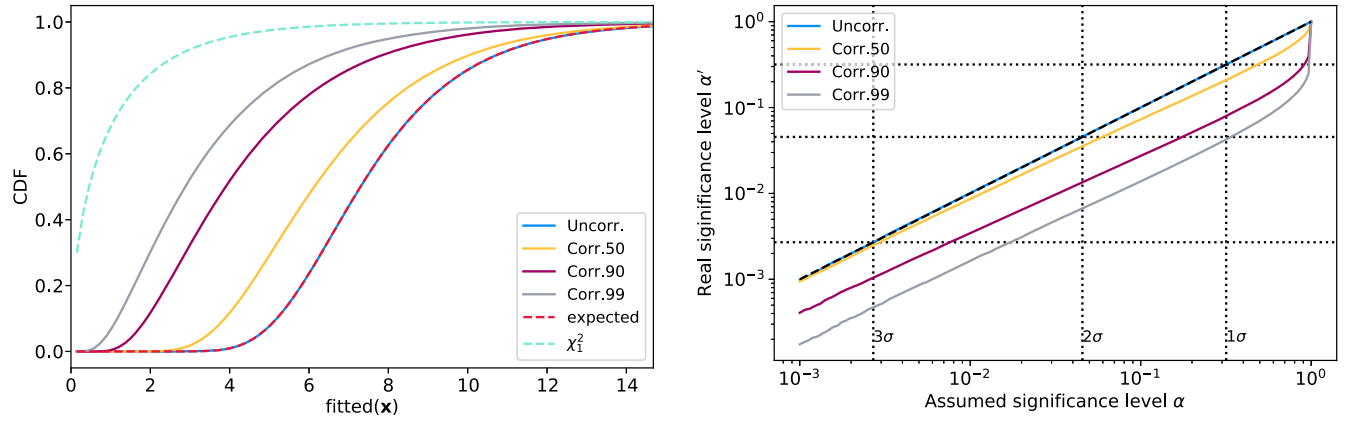


FIG. 10. CDFs (left) for the fitted test statistic for different levels of correlations in data with a more complicated correlation structure (see text). When using the uncorrelated CDF to calculate the significance level (or p-value) of a value of the statistic, the actual level depends on the correlations (right).

Here c_0 is the maximum correlation as specified in the plot labels, and $c_{N-2} = -c_0/2$. All intermediate c_i are linearly equidistant, so $c_i - c_{i+1} = \text{const}$. The correlations are thus stronger for "neighboring" data points and there is some negative correlation as well. Since the total covariance never reaches the limit of "perfect correlation," the limiting exactness is not as efficient in this dataset and the performance is closer to that of the fitted test statistic as shown in Fig. 10. Since the fitted test statistic is considerably easier to calculate, it might be worth choosing it over an invariant one, depending on the reasonably expected correlations and computation requirements.

V. APPLICATION TO REAL NEUTRINO CROSS-SECTION DATA

As a test, let us apply the invariant 3 test statistic to some real data and compare its results with the naive approach. We will use the double-differential, charged-current quasielastic cross section measurement of (anti)muon neutrinos by MiniBooNE [7,8]. This data is presented in the form of a set of differential cross sections with a "shape error" for each bin, plus a relative "normalization error" common to all bins. The publications provide no information about the correlations between the bins, or between the shape and the normalization error.

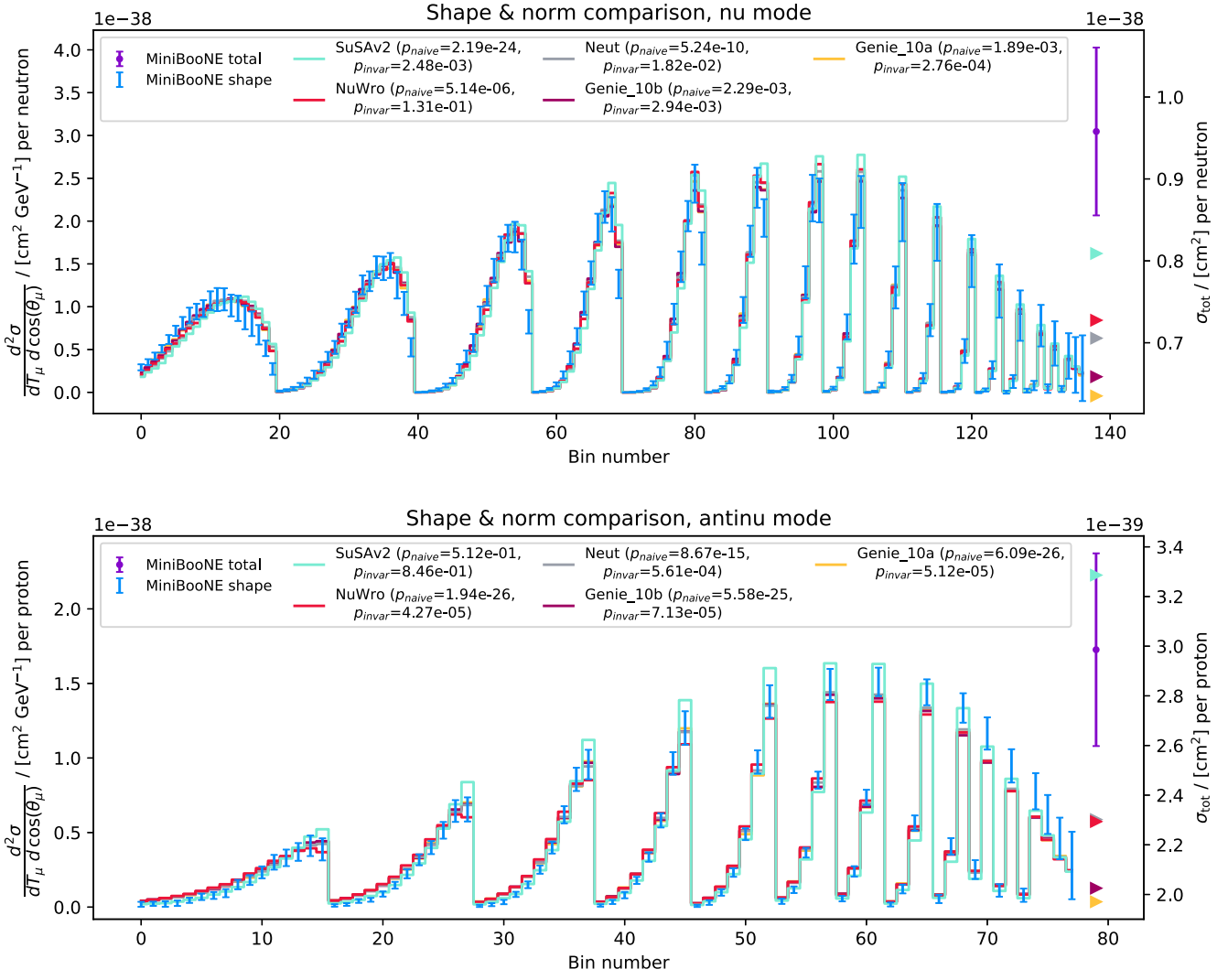


FIG. 11. Comparison of shape and normalization between MiniBooNE data and several model predictions. The cosine of the muon angle increases bin by bin, the muon's kinetic energy increases block by block. The error bars show the shape error of the data. The data point on the right shows the common normalization error. The model predictions have been scaled to the same total cross section as the data for the shape comparison, i.e., the sum over all bins is identical for all models and the data. The points on the right show the actual total cross section of the model predictions. The p-values were calculated using all bins and the normalization. See Table I for shape-only p-values.

TABLE I. P-values from the comparison of models and MiniBooNE data.

			GENIE 10a	GENIE 10b	NEUT	NuWro	SuSAv2
ν	Shape and normalization	Naive	1.89×10^{-03}	2.29×10^{-03}	5.24×10^{-10}	5.14×10^{-06}	2.19×10^{-24}
		Invariant	2.76×10^{-04}	2.94×10^{-03}	1.82×10^{-02}	1.31×10^{-01}	2.48×10^{-03}
	Shape only	Naive	3.50×10^{-02}	2.37×10^{-02}	5.66×10^{-09}	2.27×10^{-05}	3.40×10^{-24}
		Invariant	6.89×10^{-02}	1.40×10^{-01}	1.81×10^{-02}	1.30×10^{-01}	2.46×10^{-03}
$\bar{\nu}$	Shape and normalization	Naive	6.09×10^{-26}	5.58×10^{-25}	8.67×10^{-15}	1.94×10^{-26}	5.12×10^{-01}
		Invariant	5.12×10^{-05}	7.13×10^{-05}	5.61×10^{-04}	4.27×10^{-05}	8.46×10^{-01}
	Shape only	Naive	9.92×10^{-24}	3.71×10^{-23}	2.82×10^{-14}	7.32×10^{-26}	4.96×10^{-01}
		Invariant	5.05×10^{-05}	7.05×10^{-05}	5.54×10^{-04}	4.22×10^{-05}	8.42×10^{-01}

Model predictions for the measurements were generated with NUISANCE [9]. The generators and models considered in these studies are:

1. GENIE [10] v3.00.06 tune G18_10a_02_11a
2. GENIE v3.00.06 tune G18_10b_00_000
3. NEUT [11] v5.4.1
4. NuWro [12] v19.02.2
5. GENIE v3.00.06 with SuSAv2 [13].

To make them comparable to the data, they are split into a shape and a normalization part as described in [8].

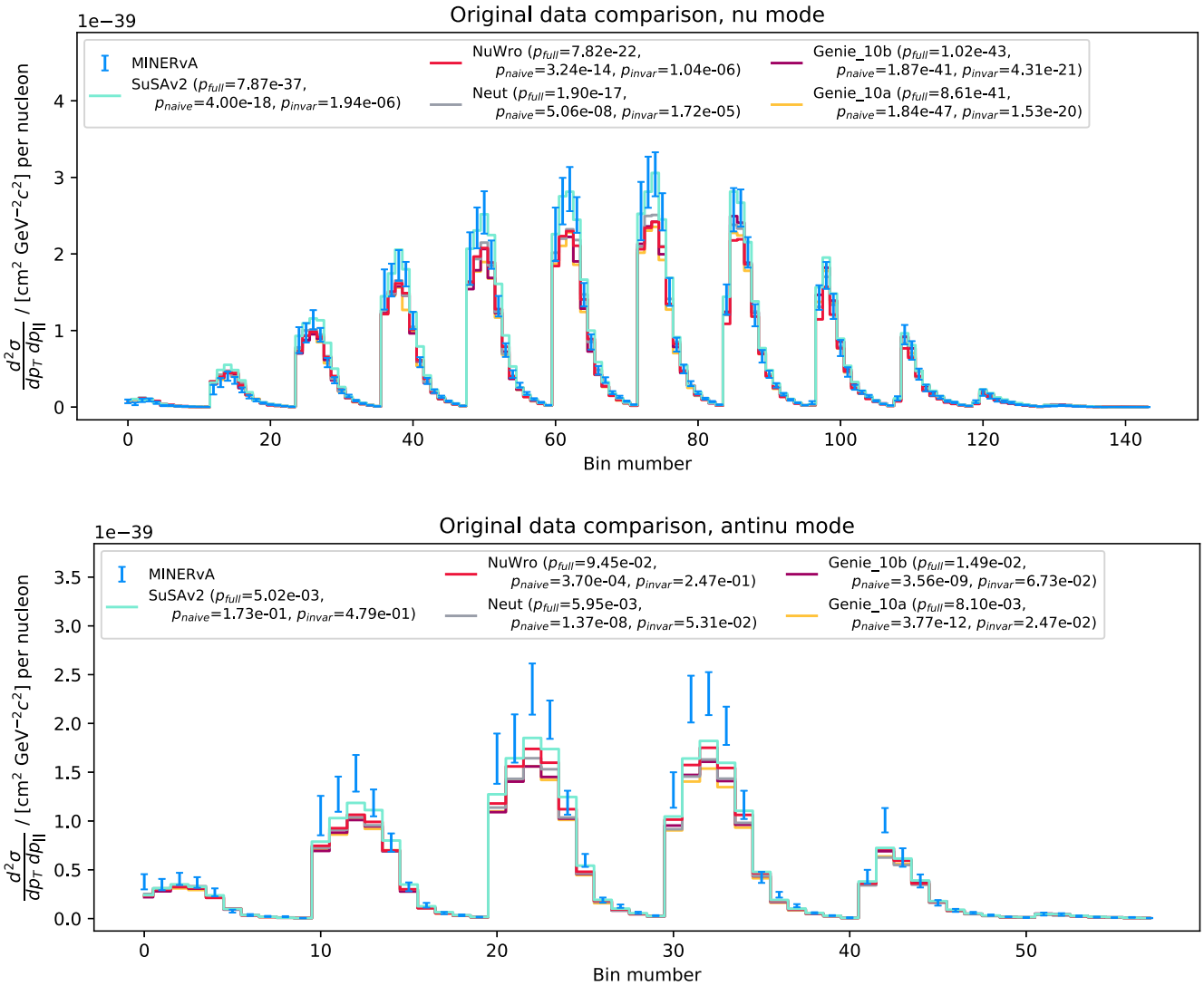


FIG. 12. Comparison between original MINERvA data and several model predictions. The muon's parallel momentum $p_{||}$ increases bin by bin, its transverse momentum p_T block by block. The error bars show the diagonal elements of the full covariance of the data.

The shape part is scaled to the total cross section of the data, so it can be compared directly to the published data points:

$$x^{\text{data/MC,norm}} = \sum_i x_i^{\text{data/MC}} w_i, \tag{38}$$

$$x_i^{\text{MC,shape}} = \frac{x_i^{\text{MC}}}{x^{\text{MC,norm}}} x^{\text{data,norm}}, \tag{39}$$

with the 2D bin area w_i, which is constant for all bins in this case. Figure 11 shows the comparison of the data and the model predictions. Despite the lack of covariance information and the implied claim that the uncertainties are independent, it is clear that the shape errors must be correlated in some way. Not only does the cross section vary too smoothly from bin to bin, but any variation of only the shape must yield a set of points with a constant sum, meaning the uncertainties of the points cannot be uncorrelated. Furthermore, it is reasonable to assume that there could be correlations between the normalization and the shape. Such correlations would easily arise from scaling uncertainties that affect some bins more than others.

Given the data and the model predictions, it is easy to calculate p-values with the "naive" and the invariant 3 test statistics:

$$p_{\text{naive}} = 1 - F_{\chi_N^2}(\text{naive}(\boldsymbol{\Delta}|\boldsymbol{s}^{\text{data}})), \tag{40}$$

$$p_{\text{invar}} = 1 - F_{\chi_1^2}(\text{invariant}_3(\boldsymbol{\Delta}|\boldsymbol{s}^{\text{data}}, \alpha = 2/3)), \tag{41}$$

with $\boldsymbol{\Delta} = \boldsymbol{x}^{\text{data}} - \boldsymbol{x}^{\text{MC}}$, and the number of data points N. Both the vectors $\boldsymbol{x}$ and N explicitly include the added "normalization bin" when comparing both shape and normalization.

Table I shows the p-values resulting from the comparisons using the two test statistics, with and without the normalization bin. In most cases, the naive chi-square statistic suggests a much stronger disagreement between the data and the model than the invariant statistic. This is consistent with the behavior we have seen in the toy studies. Apparent

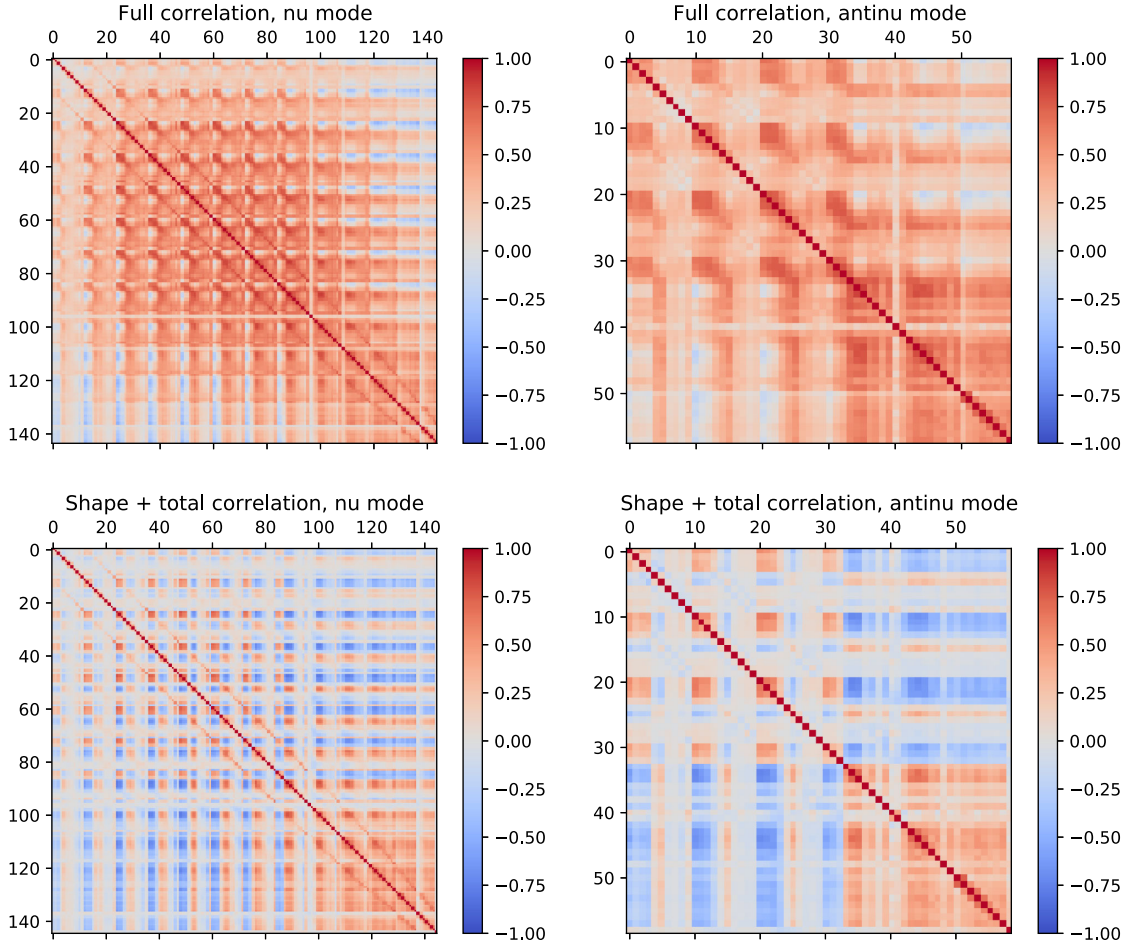


FIG. 13. Correlation matrix of the MINERvA experiment as reported (top) and after decomposition into shape and normalization bin (bottom) for both neutrino (left) and antineutrino (right) modes. The decomposition reduces the strong positive correlations in the data, but considerable positive and negative correlations remain.

strong statistical significance of the naive test statistic actually corresponds to a much weaker significance in the presence of correlations (see Fig. 1), while the invariant test statistic tends to be conservative (see Figs. 7 and 9).

Since the MiniBooNE publications do not provide a full covariance matrix, it is impossible to judge how the "diagonal-only" statistics compare to the "correct" results using the information about correlations. For this we can look at a comparable dataset from the MINERvA experiment. In [14,15] they report double-differential, quasielastic-like cross sections in variables of muon momentum. The number of bins is comparable with the MiniBooNE measurements and they use a similar data unfolding strategy.

Unlike MiniBooNE, MINERvA reports the cross sections with a full covariance matrix, and they do not decompose the uncertainties into a shape and a normalization part. This means we can directly compare the p-values we obtain when using the full Mahalanobis distance, with the ones obtained using the test statistics ignoring the off diagonals of the covariance matrix (see Fig. 12). None of the considered models describe all of the data particularly well. The naive test statistic sometimes yields a better and sometimes a worse fit between model and data compared to the full Mahalanobis distance. The invariant 3 test statistic is consistently conservative. It is worth noting that while the naive statistic tends to be closer to the correct answer for the models with very poor fits, it consistently overestimates the tension between data and model for the better fitting ones, e.g., the NuWro prediction of the antineutrino data.

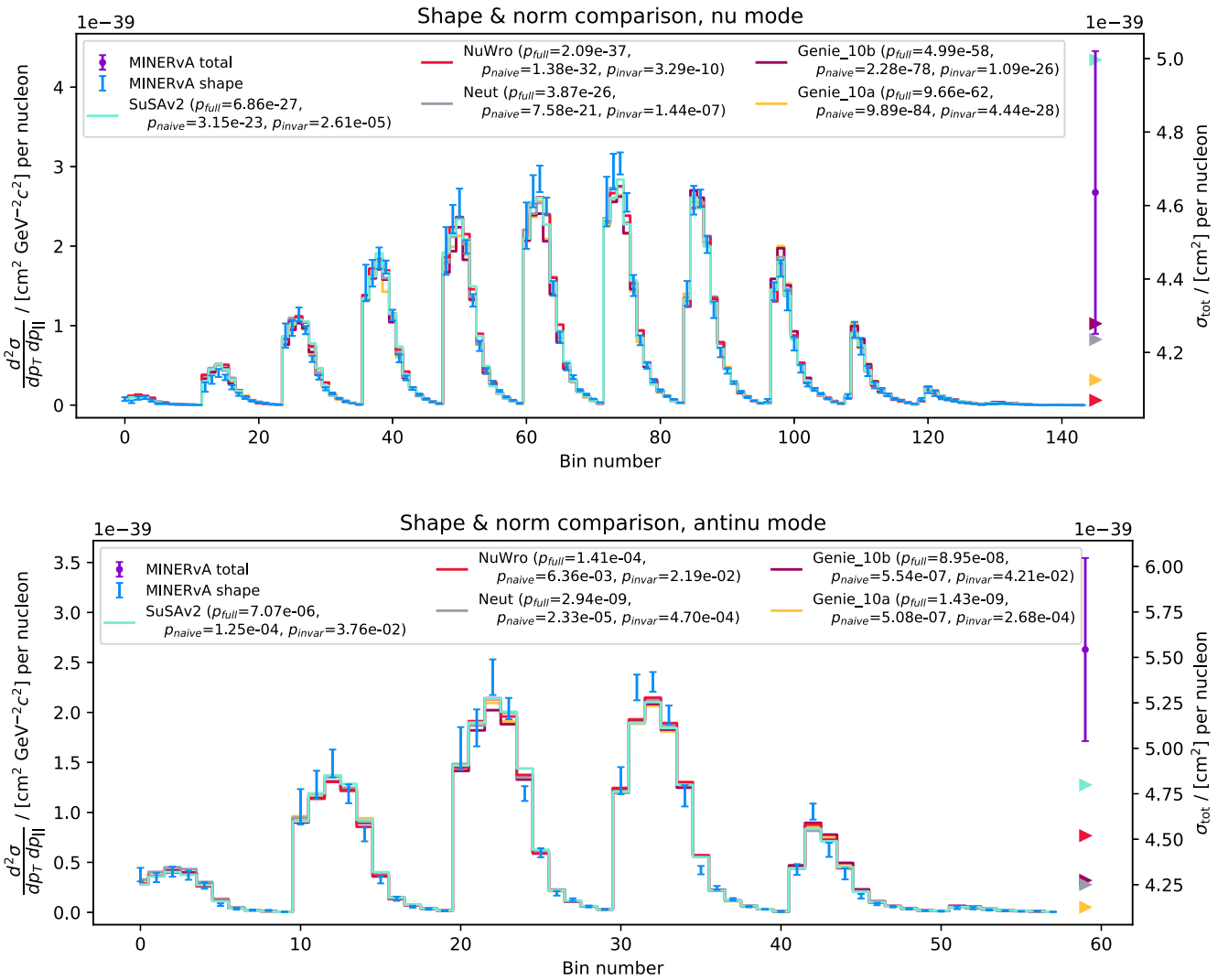


FIG. 14. Comparison between decomposed MINERvA data and several model predictions. The muon's parallel momentum $p_{||}$ increases bin by bin, its transverse momentum p_T block by block. The error bars show the shape error of the data. The data point on the right shows the common normalization error. The model predictions have been scaled to the same total cross section as the data for the shape comparison, i.e., the sum over all bins weighted by the 2D bin widths is identical for all models and the data. The points on the right show the actual total cross section of the model predictions. The p-values were calculated using all shape bins and the normalization. See Table II for shape-only p-values.

We can make the MINERvA data look even more like the MiniBooNE one by decomposing the uncertainty into a shape and a normalization part following Eq. (38). Since we have the covariance of the original cross section, we can also calculate the covariance of the shape and norm:

$$Q_{ij} = \sum_{k,l} J_{ki} M_{ij} J_{lj}. \tag{42}$$

Here M is the original covariance matrix and J is the Jacobian matrix of the combined shape and norm vector:

$$J_{ij} = \begin{cases} \delta_{ij} x_{\text{norm}}^{-1} - x_i w_j x_{\text{norm}}^{-2} & \text{if } 0 \leq i < N \\ w_j & \text{if } i = N, \end{cases} \tag{43}$$

with the number of original cross-section bins N. The correlation matrices before and after the decomposition are shown in Fig. 13. As intended, the decomposition reduces the amount of positive correlation in the data, but considerable positive and negative correlations remain.

Figure 14 shows the model comparisons with different test statistics in this decomposed data. Following MiniBooNE's approach, the shape errors are scaled up by the total cross section in the data for plotting purposes. Again, the invariant 3 test statistic is consistently conservative compared to the full Mahalanobis distance, while the naive test statistic is not. Note that for the calculation of the Mahalanobis distance, the pseudoinverse of the covariance matrix was used. Since shape and norm together have one additional dimension compared to the original cross section, their covariance matrix is not positive definite in general.

Table II summarizes the p-values from the different model comparisons to the MINERvA data. The "full" p-values using the correlated Mahalanobis distance in the shape and norm case are different from the original case, despite the fact that they should contain the same information. This is caused by the different parametrizations of the problem space. Since the decomposition in shape and norm is not a linear transformation, the likelihood surfaces described by the covariance matrices cannot be identical. It is to be expected that this will lead to differing p-values, especially in the low-probability tails. But even for the comparatively well-fitting models with p-values in the order of a few percent, the difference is surprisingly strong. The NuWro prediction of the antineutrino data has a p-value of ∼13% in the original parametrization, but only 0.03% in the decomposed view. This would make a huge difference in the interpretation of the data with respect to this model. Note also that the invariant test statistic in the decomposed case is much closer to the original full p-values in those cases.

Since we have the covariance matrices, we can create toy data that is distributed according to them and check for the coverage properties of the different test statistics, just like we did in the previous sections. The results of these studies is shown in Fig. 15. The top plot shows the performance of the full, naive, and invariant statistics when applied to the data distributed according to MINERvA's original covariance matrix. The middle plots show the same for data distributed according to the decomposed covariance. As in the previous studies, we see that the invariant test statistic shows much more accurate coverage properties than the naive one. The full Mahalanobis distance performs best, as expected.

In the bottom plots, however, we see the performance of the three statistics when generating datasets according to the original covariance, and then applying the decomposition. Because of the nonlinear transformation of the decomposition, the covariance of the shape and

TABLE II. P-values from the comparison of models and MINERvA data.

			GENIE 10a	GENIE 10b	NEUT	NuWro	SuSAv2
ν	Original	Full	8.61×10^{-41}	1.02×10^{-43}	1.90×10^{-17}	7.82×10^{-22}	7.87×10^{-37}
		Naive	1.84×10^{-47}	1.87×10^{-41}	5.06×10^{-08}	3.24×10^{-14}	4.00×10^{-18}
		Invariant	1.53×10^{-20}	4.31×10^{-21}	1.72×10^{-05}	1.04×10^{-06}	1.94×10^{-06}
	Shape and normalization	Full	9.66×10^{-62}	4.99×10^{-58}	3.87×10^{-26}	2.09×10^{-37}	6.86×10^{-27}
		Naive	9.89×10^{-84}	2.28×10^{-78}	7.58×10^{-21}	1.38×10^{-32}	3.15×10^{-23}
		Invariant	4.44×10^{-28}	1.09×10^{-26}	1.44×10^{-07}	3.29×10^{-10}	2.61×10^{-05}
	Shape only	Full	9.37×10^{-62}	4.32×10^{-58}	3.11×10^{-26}	1.96×10^{-37}	7.36×10^{-27}
		Naive	8.78×10^{-84}	1.43×10^{-78}	6.60×10^{-21}	1.63×10^{-32}	2.54×10^{-23}
		Invariant	4.41×10^{-28}	1.08×10^{-26}	1.43×10^{-07}	3.27×10^{-10}	2.60×10^{-05}
$\bar{\nu}$	Original	Full	8.10×10^{-03}	1.49×10^{-02}	5.95×10^{-03}	9.45×10^{-02}	5.02×10^{-03}
		Naive	3.77×10^{-12}	3.56×10^{-09}	1.37×10^{-08}	3.70×10^{-04}	1.73×10^{-01}
		Invariant	2.47×10^{-02}	6.73×10^{-02}	5.31×10^{-02}	2.47×10^{-01}	4.79×10^{-01}
	Shape and normalization	Full	1.43×10^{-09}	8.95×10^{-08}	2.94×10^{-09}	1.41×10^{-04}	7.07×10^{-06}
		Naive	5.08×10^{-07}	5.54×10^{-07}	2.33×10^{-05}	6.36×10^{-03}	1.25×10^{-04}
		Invariant	2.68×10^{-04}	4.21×10^{-02}	4.70×10^{-04}	2.19×10^{-02}	3.76×10^{-02}
	Shape only	Full	6.04×10^{-08}	2.71×10^{-06}	1.28×10^{-08}	2.78×10^{-04}	9.13×10^{-06}
		Naive	3.10×10^{-06}	2.19×10^{-06}	8.74×10^{-05}	1.12×10^{-02}	1.54×10^{-04}
		Invariant	2.64×10^{-04}	4.14×10^{-02}	4.62×10^{-04}	2.15×10^{-02}	3.69×10^{-02}

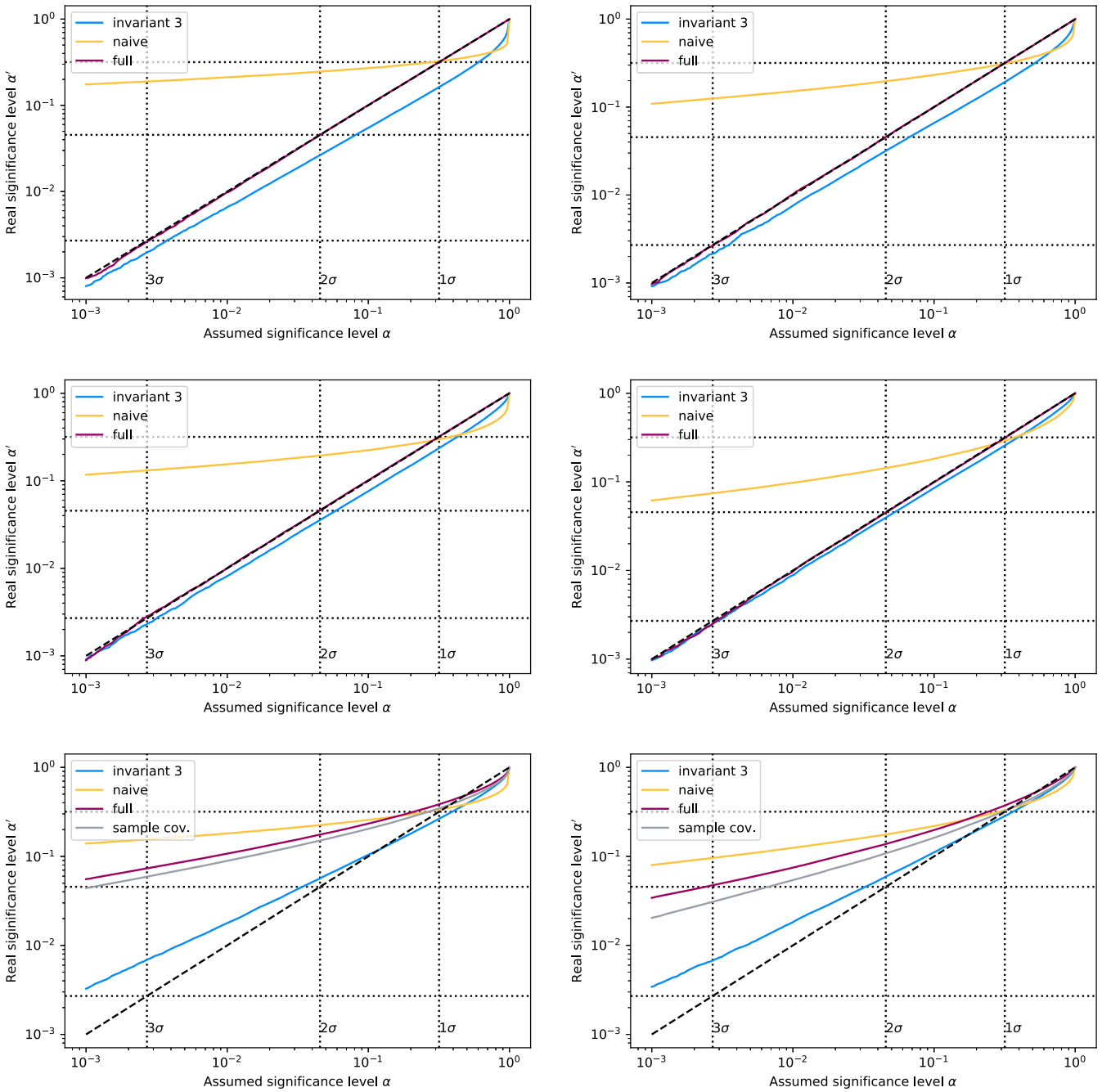


FIG. 15. Assumed vs actual significances with different test statistics in toy data distributed according to the provided MINERvA covariance (top) and the covariance of the decomposition into shape and total cross section (middle) for both the neutrino (left) and antineutrino data (right). The bottom plots show the calculated and true significance levels when generating toy data according to the original covariance and applying the decomposition afterwards. The nonlinear transformation of the decomposition causes a significant distortion in the distribution of the Mahalanobis distance, even when using the full covariance matrix. This improves only slightly when using the sample covariance instead of the linear error propagation. The invariant test statistic seems to be less sensitive to these distortions than the full Mahalanobis distance.

norm parametrization does no longer reflect the actual distribution of the data. This deteriorates the coverage properties of the full Mahalanobis distance test statistic to the point where it performs as bad as, or worse than, the naive test statistic. Interestingly, the invariant test statistic is not as affected by the nonlinear transformation as the other two. It is no longer conservative, but it shows better coverage properties than even the full statistic. This is probably due to the fact that the full statistic is trying to make full use of the assumed shape of the data distribution, i.e., the correlation between the data points, which is distorted due to a nonlinear transformation. The invariant

test statistic on the other hand is designed to work well with a wide range of correlations, in the data. Note that the performance of the full Mahalanobis distance improves slightly when calculating the covariance in the decomposed space directly from the sample, rather than doing linear error propagation. It still performs rather poorly, because of the non-Gaussian shape of the underlying data, though. This underlines the importance of checking that reported uncertainties are actually sufficiently normal distributed in the chosen parametrization, when reporting them as a covariance matrix.

VI. CONCLUSIONS

We have shown that the use of the naive uncorrelated Mahalanobis distance in the presence of unknown correlations leads to wrong coverage properties. The presented alternative test statistics perform more robust under varying degrees of correlation in the data.

The fitted test statistic is motivated by treating the correlations as nuisance parameters. Its distribution in the absence of correlations is known (the "Bee-square distribution") and it is conservative in their presence. It could be used when it is important to have an "actual" Mahalanobis distance[5] for the combination with other datasets, and the overestimation of the error in the presence of correlations is acceptable. It also has the advantage of being incredibly easy to calculate, as it is just the maximum squared z-score among the variables. Depending on the actual correlations in the data, its performance is actually comparable to the invariant test statistics.

The invariant 3 test statistic performs the best across varying levels of correlations. Its level of conservativeness can be tuned with the shape parameter α. A value of 0.5 seems to be a safe choice, while the best performance in the toy datasets in the presented studies was achieved at $\alpha = 2/3$. As it goes to 0, the test statistic becomes equivalent to the invariant 2 statistic, which can be seen as the most conservative version of invariant 3. The invariant 3 statistic could be used when it is important to not overestimate the uncertainties by too much. If necessary, the relative weight of the statistic can be tuned by transforming it to a chi-square distribution with $N > 1$ degrees of freedom. The choice of N is somewhat arbitrary though.

The application to real data from MiniBooNE and MINERvA shows that the invariant 3 test statistic performs as expected. It is consistently conservative, while the naive uncorrelated Mahalanobis distance overestimates the strength of the discrepancy between data and model in multiple instances. When the MINERvA data is decomposed into a shape and a normalization part, the invariant statistic even has better coverage properties than a full Mahalanobis distance that uses a linearly propagated covariance matrix. This is most likely due to the nonlinear nature of the decomposition into shape and norm, which distorts the shape of the distribution. If it was multivariate Gaussian originally, the covariance in the shape and norm space can only ever be an approximation. Since the invariant statistic does not use information about the correlations between the data points, it is not affected as much by this as the fully correlated Mahalanobis distance.

[5]The usefulness of this is probably limited though, as it will not be chi-square distributed in general.

ACKNOWLEDGMENTS

I would like to thank Callum Wilkinson and Stephen Dolan for generating the NUISANCE files that were used in the MiniBooNE and MINERvA studies, as well as providing a space for discussions, feedback, and venting. Thanks also go to Louis Lyons for providing feedback and clarifying discussions about the concepts discussed in this paper. This work was supported by a grant from the Science and Technology Facilities Council.

APPENDIX: PROOF OF CONSERVATIVENESS OF FITTED TEST STATISTIC IN TWO DIMENSIONS

The conservativeness of the fitted test statistic can be proven by showing that the CDF of any given maximum

Listing 1: Python implementation of the Bee-square distribution.

```
import numpy as np
from scipy.stats import rv_continuous
from scipy.special import erf, erfinv

class Bee(rv_continuous):
    def _cdf(self, x, df):
        return erf(x/np.sqrt(2))**df

    def _pdf(self, x, df):
        ret = df*(erf(x/np.sqrt(2)))**(df–1)
        return ret * np.sqrt(2/np.pi)*np.exp(-x**2/2)

    def ppf(self, x, df):
        return erfinv((x)**(1/df)) * np.sqrt(2)

# Instance of the distribution, support starts at 0
bee = Bee(a = 0)

class Bee2(rv_continuous):
    def _cdf(self, x, df):
        b = np.sqrt(x)
        ret = bee.cdf(b, df)
        return ret
    def _pdf(self, x, df):
        ret = df*(erf(np.sqrt(x/2)))**(df-1)
        return ret / np.sqrt(2*np.pi*x) * np.exp(-x/2)

    def_ppf(self, x, df):
        b = bee.ppf(x, df)
        return b**2

# Instance of the distribution, support starts at 0
bee2 = Bee2(a = 0)
```

absolute value b among the N standard normal distributed variables is minimal, when the variables are uncorrelated. In two dimensions this can be done by proving that

$$|S|^{-\frac{1}{2}} \int_{-b}^{b} \int_{-b}^{b} \exp\left(-\frac{1}{2} z^T S^{-1} z\right) \mathrm{d}z_1 \mathrm{d}z_2, \tag{A1}$$

with $S = \left(\begin{smallmatrix} 1 & a \\ a & 1 \end{smallmatrix}\right)$ and $a \in (-1, 1)$, is minimal for $a = 0$.

Differentiating term (A1) with respect to a yields

$$\frac{2(1 - e^{\frac{2ab^2}{a^2-1}}) e^{\frac{b^2(1-a)}{a^2-1}}}{\sqrt{1-a^2}}, \tag{A2}$$

as can be shown e.g., with computer algebra systems. This term is positive when $a > 0$, 0 when $a = 0$, and negative when $a < 0$. Thus term (A1) is minimal at $a = 0$. ∎

Listing 2: Python implementation of the invariant 3 test statistic. Data must be provided in z-scores, i.e., normalized by the variance. This uses caching for improved performance and works with survival functions instead of CDFs for improved numerical accuracy.

```
@np.vectorize
@lru_cache(10000)
def _yfrommax(b, df = 2, alpha = 0.5):
    """(1—diagonal coordinate) from (1—max) of accepted region"""
    # A=(1-b)**df—((y-b)**df)/((1-alpha+alpha*y)**(df-1))=1—y
    beta=1—alpha
    q=(1.0—b) ** df—1
    dfm=df—1

    def f(y):
        return q—((y—b) ** df) / ((beta + alpha * y) ** (dfm)) + y

    if b <=0:
        return 0.0
    if b >= 1:
        return 1.0
    else:
        return root_scalar(f, x0=b, x1=b * 1.001).root

def yfrommax(b, df = 2, alpha = 0.5):
    """Buffer and interpolate values to speed things up."""
    step = 0.0001
    b_=np.floor(b / step, dtype=float) * step
    b__=b_+step
    delta = (b—b_) / step
    x_=_yfrommax(b_, df = df, alpha = alpha)
    x__=_yfrommax(b__, df = df, alpha = alpha)
    return x_+(x__—x_) * delta

def invariant3(x, alpha = 0.5, fast = False):
    """Return test statistic given vector of normalized values."""
    if fast:
        sf=1—chi2.cdf(x ** 2, df = 1) # Faster, but less accurate
    else:
        sf=chi2.sf(x ** 2, df = 1)

    # Get possible diagonal coordinate from maximum CDF value (= minimum SF)
    a=np.min(sf, axis=-1)
    b=np.max(sf, axis=-1)
    yfm = yfrommax(a, df=x.shape[-1], alpha = alpha)

    # Get possible diagonal coordinate from center surface
    yfc=(alpha * (a—b) + b) / (1.0 + alpha * (a—b))

    y= np.minimum(yfc, yfm)
    y = np.maximum(y, 0) # Cap in case of rounding or root finding errors
    return chi2.isf(y, df = 1)
```

This calculation will remain true when integrating over additional statistically independent variables in term (A1). Those will add constant factors to the derivative, but not change the general dependence on a. Starting from a covariance with no correlations, the CDF is thus globally minimal vs variations of any single of the off-diagonal elements.

[1] P. C. Mahalanobis, On the generalized distance in statistics, in *Proceedings National Institute of Science, India* (1936), Vol. 2, pp. 49–55, https://insa.nic.in/writereaddata/UpLoadedFiles/PINSA/Vol02_1936_1_Art05.pdf.

[2] D. S. Wilks, The multivariate normal distribution, in *Statistical Methods in the Atmospheric Sciences* (Elsevier, New York, 2019), pp. 587–615.

[3] J. A. Rice, *Mathematical Statistics and Data Analysis* (Duxbury Press, Pacific Grove, CA, 2006).

[4] H. Tsukuma and T. Kubokawa, *Shrinkage Estimation for Mean and Covariance Matrices* (Springer, Singapore, 2020).

[5] H. Tsukuma, Estimation of a high-dimensional covariance matrix with the stein loss, J. Multivariate Anal. **148**, 1 (2016).

[6] P. K. Bhattacharya and P. Burman, Multivariate analysis, in *Theory and Methods of Statistics* (Elsevier, New York, 2016), pp. 383–429.

[7] A. A. Aguilar-Arevalo *et al.*, First measurement of the muon neutrino charged current quasielastic double differential cross section, Phys. Rev. D **81**, 092005 (2010).

[8] A. A. Aguilar-Arevalo *et al.*, First measurement of the muon antineutrino double-differential charged-current quasielastic cross section, Phys. Rev. D **88**, 032001 (2013).

[9] P. Stowell *et al.*, NUISANCE: A neutrino cross-section generator tuning and comparison framework, J. Instrum. **12**, P01016 (2017).

[10] C. Andreopoulos *et al.*, The GENIE neutrino monte carlo generator, Nucl. Instrum. Methods Phys. Res., Sect. A **614**, 87 (2010).

[11] Y. Hayato, A neutrino interaction simulation program library NEUT, Acta Phys. Pol. B **40**, 2477 (2009), https://inspirehep.net/literature/844435.

[12] T. Golan, J. Sobczyk, and J. Żmuda, NuWro: The Wrocław Monte Carlo generator of neutrino interactions, Nucl. Phys. B, Proc. Suppl. **229–232**, 499 (2012).

[13] R. González-Jiménez, G. D. Megias, M. B. Barbaro, J. A. Caballero, and T. W. Donnelly, Extensions of superscaling from relativistic mean field theory: The SuSAv2 model, Phys. Rev. C **90**, 035501 (2014).

[14] D. Ruterbories *et al.*, Measurement of quasielastic-like neutrino scattering at $\langle E_\nu \rangle \sim 3.5$ GeV on a hydrocarbon target, Phys. Rev. D **99**, 012004 (2019).

[15] C. E. Patrick *et al.*, Measurement of the muon antineutrino double-differential cross section for quasielastic scattering on hydrocarbon at $e_\nu \sim 3.5$ GeV, Phys. Rev. D **97**, 052002 (2018).