Published for SISSA by Springer

Received: March 10, 2021
Accepted: April 20, 2021
Published: May 18, 2021

Erratum: Clustering of $\bar{B} \to D^{(*)}\tau^-\bar{\nu}_\tau$ kinematic distributions with ClusterKinG

Jason Aebischer,[a] **Thomas Kuhr**[b,c] **and Kilian Lieret**[b,c]

[a] *Excellence Cluster Universe, Garching, Germany*
[b] *Excellence Cluster Origins, Garching, Germany*
[c] *Ludwig Maximilian University, Munich, Germany*

E-mail: jason.aebischer@tum.de, thomas.kuhr@lmu.de, kilian.lieret@lmu.de

Erratum to: JHEP04(2020)007

ArXiv ePrint: 1909.11088

In our paper we considered two histograms H_k, $k = 1, 2$ with bin contents n_{ki}, $i = 1, \ldots, N$ and associated uncertainties σ_{ki}. Our null hypothesis was that the bin contents of the histograms are drawn from two distributions with identical means. Let us first consider the case of uncorrelated random variables n_{ki} with standard deviations σ_{ki} and normalisations $N_k := \sum_{i=1}^{N} n_{ki}$. For a corresponding two sample test, we considered a distance measure $\tilde{\chi}^2$ with

$$\tilde{\chi}^2(H_1, H_2) = \sum_{i=1}^{N} \frac{(N_1 n_{2i} - N_2 n_{1i})^2}{N_1^2 \sigma_{2i}^2 + N_2^2 \sigma_{2i}^2}. \tag{1}$$

We have added a tilde to avoid confusion with the "true" χ_r^2 distribution of r degrees of freedom. In the following we assume that the n_{ki} are normally distributed. Under the null hypothesis $\tilde{\chi}^2(H_1, H_2)$ will be approximately distributed according to a χ_{N-1}^2 distribution [1], not a χ_N^2 distribution as assumed throughout our paper.

However, this changes the interpretation of our results but slightly, because the exact probability distribution of the metric is only important for the choice and interpretation of the stopping criterium of the clustering. The choice of our algorithm and stopping criterium guaranteed that any two points in the same cluster are "indistinguishable" and that any two distinct clusters contain at least two points that are "distinguishable". We called two points distinguishable if the corresponding histograms have a distance of $\tilde{\chi}^2/N > 1$ and

https://doi.org/10.1007/JHEP05(2021)147

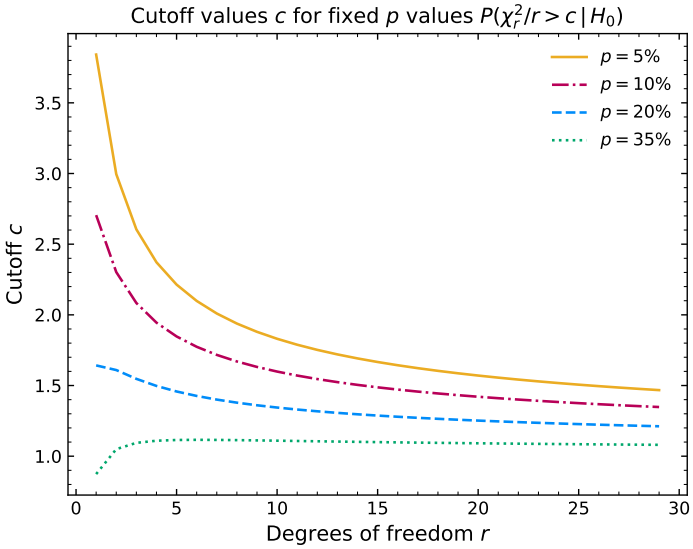

Figure 1. p values and corresponding cut-off values.

indistinguishable if not. For the 9 bins in our examples, $\tilde{\chi}^2/N = 1$ then corresponded to a p value of 34% (not 32% as claimed). In the paper it was also missed to point out that this value depends on the number of bins. The dependency is shown in figure 1.

It should also be highlighted that the approximation of the distribution of $\tilde{\chi}^2$ values to the χ^2_{N-1} distribution can break down if the uncertainties σ_{ki} are very imbalanced, though this does not usually happen if the uncertainties are dominated by Poisson uncertainties.

More importantly, we correct our treatment of the more general case of histograms with correlated bin contents. The approach of our paper to transform this problem into a scenario where (1) again leads to a χ^2_r distribution is incorrect. Instead, we define $\Delta_i = \frac{n_{1i}}{N_1} - \frac{n_{2i}}{N_2}$. With Σ_1 and Σ_2 the covariance matrices of n_{1i} and n_{2i}, we define $\Sigma = \frac{\Sigma_1}{N_1^2} + \frac{\Sigma_2}{N_2^2}$ and

$$\bar{\chi}^2(H_1, H_2) = \sum_{i,j=1}^{N} \Delta_i (\Sigma^{-1})_{ij} \Delta_j. \tag{2}$$

Under the null hypothesis, $\bar{\chi}^2(H_1, H_2)$ approximates a χ^2_{N-1} distribution. In the case of uncorrelated bin contents, $\bar{\chi}^2 = \tilde{\chi}^2$. As the examples presented in our paper did not consider correlation, this does not affect any of the results.

Both mistakes have been corrected in `ClusterKinG` release 1.1.0. The χ^2 metric is now implemented using equation (2) and we define $d(c_1, c_2) = \bar{\chi}^2(H_{c_1}, H_{c_2})/(N-1)$ (such that the expectation value is fixed to unity as originally intended). Additional unit tests automatically run several small toy studies to confirm that our current implementation produces correct approximations of χ^2_r distributions.

We have performed further validation studies for our statistical treatment of the examples shown in the paper: for each point in parameter space we consider the corresponding histogram and its covariance matrix. Toy histograms are generated by drawing random values from the multivariate normal distribution with matching means and covariance matrix.

JHEP05(2021)147

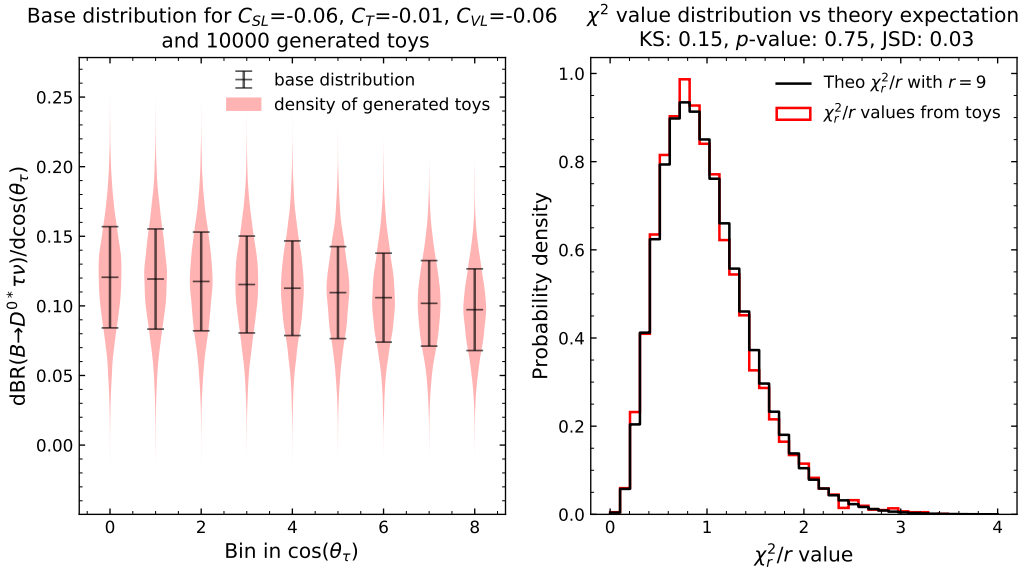

Figure 2. Toy experiments to validate the implementation of the $\bar{\chi}^2$ metric. The right sided figure also reports several values quantifying the similarity of the toy distribution to the theoretical expectation: the Kolmogorov-Smirnov test statistic (KS), its corresponding p value and the Jensen-Shannon Divergence (JSD).

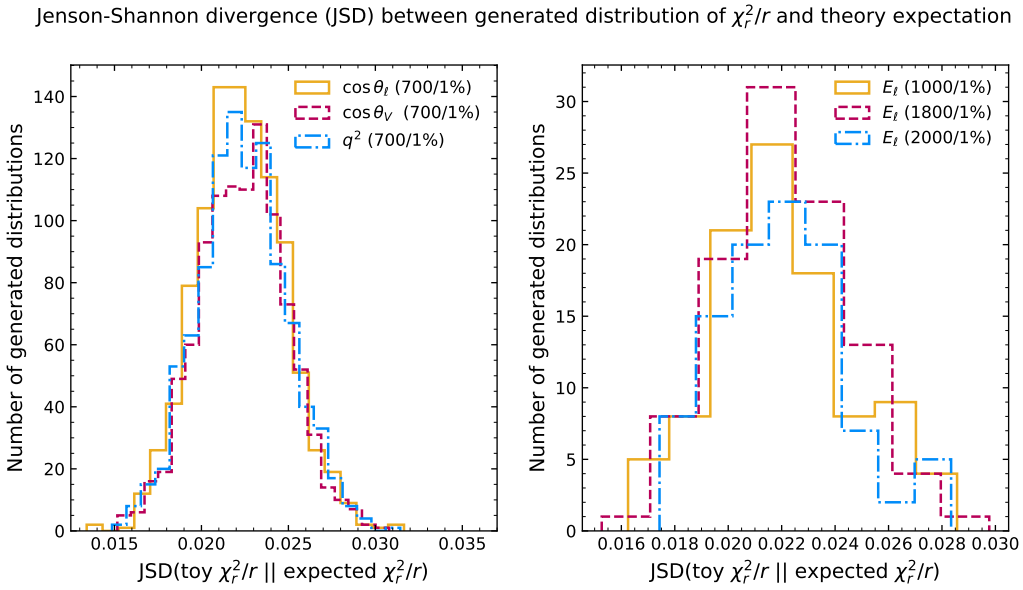

Figure 3. Validating the shape of the $\bar{\chi}^2$ distribution for all points in the parameter space. The numbers in parentheses denote the assumed total yield corresponding to the Poisson uncertainties and the uncorrelated systematic uncertainty.

JHEP05(2021)147

We then calculate the test statistic $\bar{\chi}^2$ between each toy histogram and the original histogram. The distribution of all $\bar{\chi}^2/(N-1)$ values is binned and compared to the calculated expected distribution $\chi^2_{N-1}/(N-1)$ using the Jenson-Shannon Divergence (JSD).

An example for one particular point in parameter space is shown in figure 2. Both histograms agree nicely, resulting in a low JSD value. The result of repeating the same procedure across all points is shown in figure 3, showing satisfactorily low divergence values.

Additional code to reproduce the figures shown here and to validate the statistical treatment has been added to the `ClusterKinG` repository.

References

[1] N.D. Gagunashvili, *Chi-square tests for comparison weighted histograms*, *Nucl. Instrum. Meth. A* **614** (2010) 287 [arXiv:0905.4221] [INSPIRE].

JHEP05(2021)147