Analysis of Statistical Algorithms for the Comparison of Data Distributions in Physics Experiments

Anton Lechner, *Student Member, IEEE,* Andreas Pfeiffer, Maria Grazia Pia and Alberto Ribon

***Abstract*—The relative power of statistical algorithms for the comparison of data distributions was examined for selected physics use cases. Various Goodness-of-Fit tests, applicable to binned and unbinned data sets, were evaluated. The results provide guidance about the usage of a particular test in certain scenarios.**

***Index Terms*—Goodness-of-Fit tests, data analysis, statistics**

I. Introduction

THE statistical comparison of data distributions in physics analysis often relies on the χ^2 test. However, in various situations other algorithms, based on different test statistics, might be preferable, since they can be demonstrated to be more powerful and exhibit a higher level of sensitivity to critical features of the data distributions under study.

The presented study compares the behaviour of several Goodness-of-Fit (GoF) tests for a selected set of use cases derived from experimental practice. The relative power of the tests was estimated in pseudoexperiments. The considered scenarios involve typical characteristics appearing in various domains of physics data analysis. Therefore the obtained results might provide guidance about which tests are likely to be more powerful for similar applications.

II. Goodness-of-Fit tests

Goodness-of-Fit tests provide a measure if two random samples derive from the same parent distribution. Relatively little exists in literature as systematic comparison of GoF tests; moreover what is available usually applies to the comparison of distributions derived from mathematical functions and may not be directly applicable to physics use cases.

A. Statistical Toolkit

The study of the power of Goodness-of-Fit tests was performed by employing the Statistical Toolkit [1], [2], an open source software package written in C++. Its functionality enables the statistical comparison of both binned and unbinned distributions; interfaces to AIDA-compliant [3] data analysis systems and ROOT [4] are provided.

Manuscript received November 23, 2007.

A. Lechner is with the Atomic Institute of the Austrian Universities, Vienna University of Technology, Vienna, Austria and CERN, Geneva, Switzerland.

M. G. Pia is with INFN Sezione di Genova, Via Dodecaneso 33, I-16146 Genova, Italy (phone: +39 010 3536420, fax: +39 010 313358, e-mail: MariaGrazia.Pia@ge.infn.it).

A. Pfeiffer and A. Ribon are with CERN, Geneva, Switzerland.

TABLE I
Overview of Goodness-of-Fit Tests included in the Statistical Toolkit (Version 2.3)

Tests for binned data sets	**Tests for unbinned data sets**
Anderson-Darling	Anderson-Darling
Anderson-Darling approximated	Anderson-Darling approximated
χ^2	-
Fisz-Cramer von Mises	Fisz-Cramer von Mises
-	Girone
-	Goodman
-	Kolmogorov-Smirnov
Tiku	Tiku
-	Watson
-	Weighted Cramer von Mises
-	Weighted Kolmogorov-Smirnov *(AD weighting function)*
-	Weighted Kolmogorov-Smirnov *(Buning weighting function)*

A summary of all tests, which are available in the Statistical Toolkit (version 2.3), is given in Table I. A noteworthy feature of this package is the rich collection of Goodness-of-Fit tests included; the availability of different algorithms in the same software system facilitates a study of their relative performance.

B. Relative power of GoF tests

Given two samples deriving from different parent distributions, a Goodness-of-Fit test can be considered powerful, if the probability for not rejecting the hypothesis that the samples have the same parent distribution is low [5]. Comparison of the behaviour of GoF tests in different applications allows an assessment of their relative power.

In physics use cases, scenarios might occur where data sets derive in principle from the same parent distribution, but are subject to perturbations, like systematic errors or fluctuations, which cause discrepancies between them. Therefore it is interesting to evaluate the relative ability of GoF tests of detecting such perturbations as a function of the extent of the perturbation itself.

C. Pseudoexperiments

The relative power of different GoF tests was studied by comparing perturbed data distributions against their unperturbed reference. For each physical use case considered,

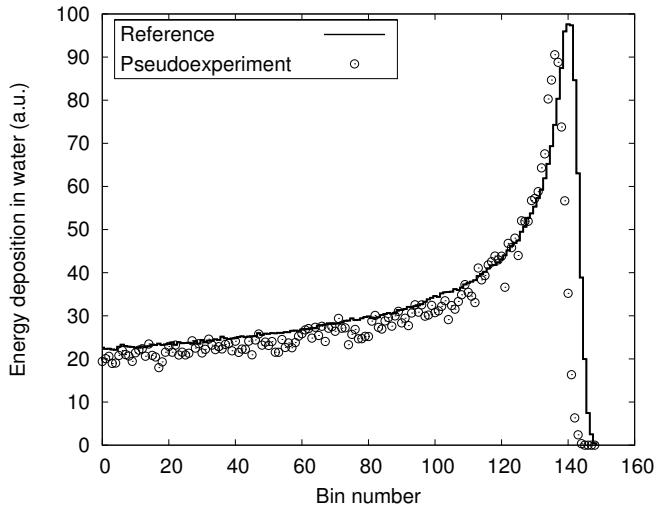


Fig. 1. Proton Bragg peak in water: Experimental curve (reference) and a data sample deriving from a pseudoexperiment which simulates perturbing effects like shifts and scaling.

a large number of pseudoexperiments was performed and the behaviour of the p-value distributions obtained with the different statistical algorithms was studied.

A pseudoexperiment consisted of a stochastic procedure where a distribution was randomly sampled from an analytical function or from an experimental distribution according to the relative error imposed by the measurement. A statistical comparison against a reference distribution was performed for each pseudoexperiment and the resulting p-values were recorded.

A convenient way to investigate the power of a GoF test is to consider the fraction of p-values <(1-CL), $0 \leq \mathrm{CL} < 1$, where CL refers to a specified confidence level. For the current investigations a CL of 0.9 was imposed on all scenarios under consideration to treat the compared distributions as compatible.

III. Application of GoF-Test in Physics scenarios

Three physical use cases as encountered in experimental data analysis were considered as playground for the examination of GoF tests. For the comparison of binned data sets the Anderson-Darling (AD), Cramer-Von Mises (CvM), Tiku and χ^2 algorithms are considered; algorithms applied to unbinned distributions are the Anderson-Darling, Kolmogorov-Smirnov, Tiku, Watson and Cramer-Von Mises.

A. Scale-location problem

The scale-location problem is a well-known scenario in statistics. It generally deals with the situation of compatible distributions, which are subject to shift or relative scaling.

Displaced signals and peaks or scaling effects commonly appear in physics applications. In this study the scale-location problem is addressed in a use case concerning energy vs depth distributions, representing the proton Bragg peak in water (see Fig. 1).

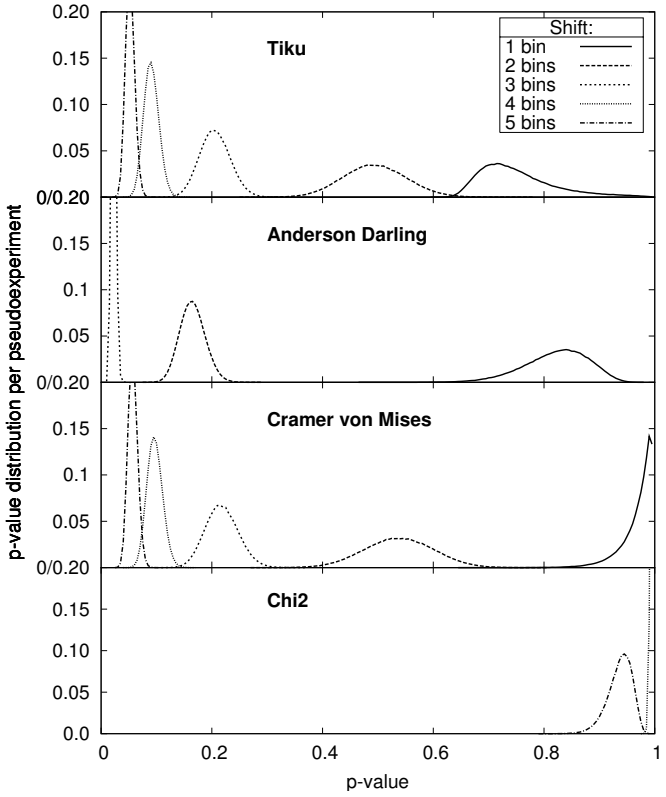


Fig. 2. p-value distributions per pseudoexperiment deriving from the statistical comparison between the reference proton Bragg peak and shifted curves, that were randomly sampled. Shifts are specified in terms of bins of the energy vs depth histograms. Results are presented for the Tiku, Anderson-Darling, Cramer von Mises and χ^2 GoF tests (binned versions).

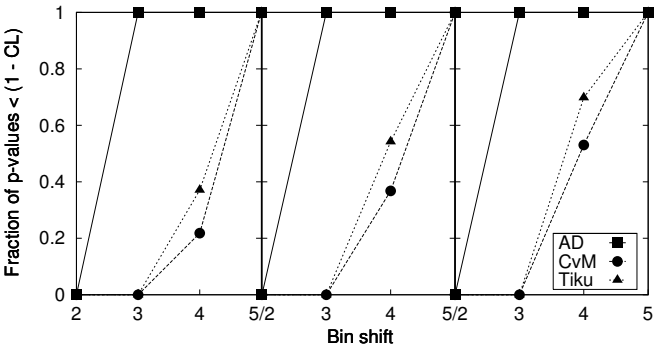
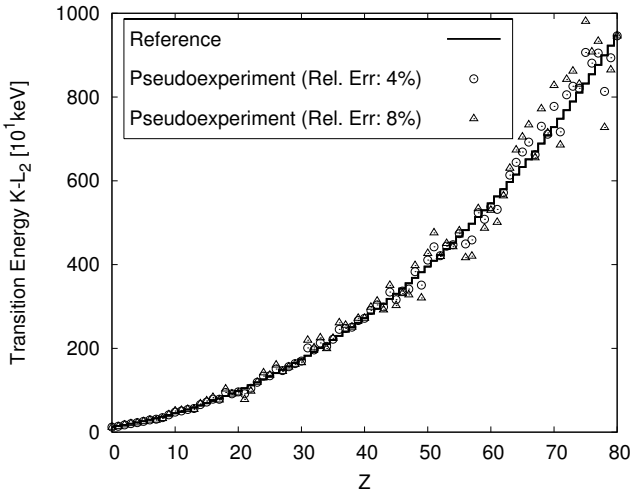


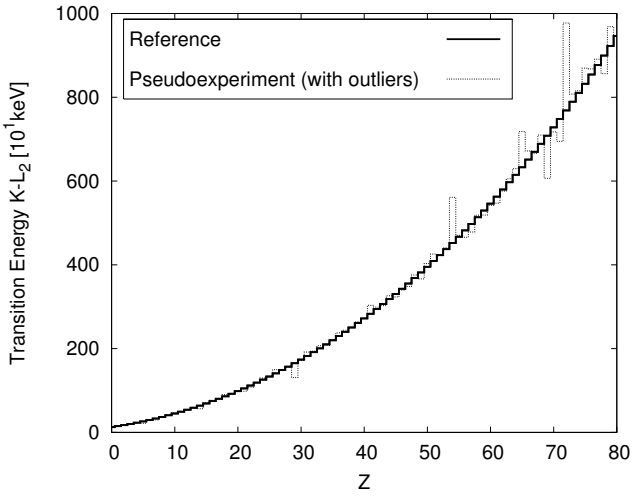
Fig. 3. Scale-location problem: Fraction of p-values <(1-CL) as a function of bin shift for ensembles of pseudoexperiments involving scaling factors of 0.90 (left), 0.95 (center) and 1.00 (right).

As a reference distribution, an experimental depth-dose profile, measured at the CATANA hadrontherapy facility at INFN LNS [6], was utilized. Scaling and shifting factors were applied to produce the results associated with pseudoexperiments.

As a first step the sole influence of displacement with respect to the reference curve is investigated, neglecting any rescaling of the depth-dose profiles. Figure 2 shows the p-value distributions corresponding to shifts up to 5 bins, where the bin size is equal to 200 μm. The full width at half maximum (FWHM) of the Bragg peak is 21 bins, equivalent to 4.2 mm.



(a)



(b)

Fig. 4. Fluorescence spectrum for the K-L_2 transition: Reference curves and distributions deriving from pseudoexperiments that simulate (a) fluctuations and (b) outliers.

The profiles expose that the Anderson-Darling test is most sensitive to shifts compared to the other algorithms, while the χ^2 test only shows a response for shifts larger than 3 bins; hence the χ^2 test reacts slower as a function of the perturbation to recognize incompatibilities in the distributions due to the introduced shift. The similar behaviour of the Tiku and Cramer-Von Mises tests for shifts larger than 1 bin is not surprising, since the Tiku test statistics is based on the Cramer-Von Mises one.

Comparable results are obtained if both perturbing effects are combined. Figure 3 shows the fractions of p-values $<$ (1-CL) as a function of the applied shift for different scaling factors. The Anderson-Darling test rejects the assumption of compatibility for shifts larger than 2 bins, independently of the considered rescaling, and thus shows the highest sensitivity. The behaviour of the Tiku test is slightly different from the response of the Cramer-Von Mises test. For the χ^2 test all p-value distributions were $>$ (1-CL), i.e. all data samples were considered as compatible with the original Bragg peak profile.

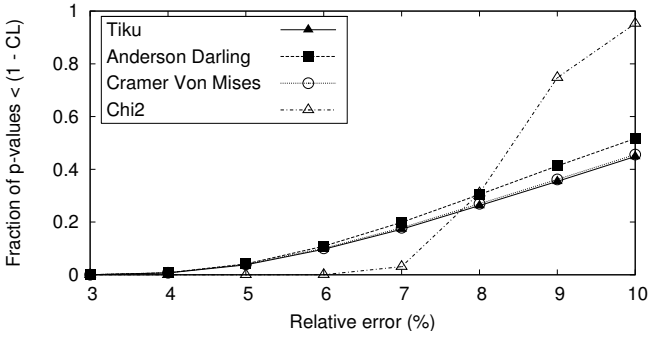


Fig. 5. Fluctuations in the fluorescence spectrum: Fraction of p-values $<$(1-CL) as a function of the relative error in the data sample.

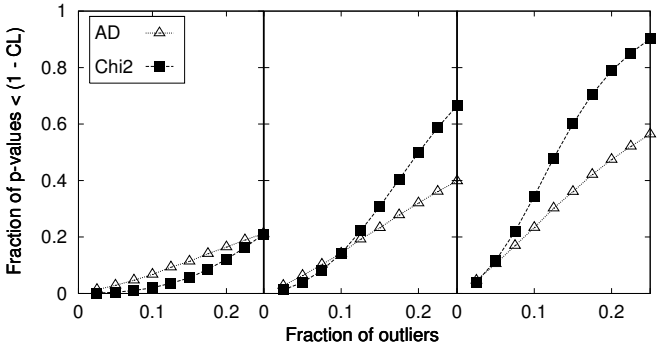


Fig. 6. Outliers in the fluorescence spectrum: Fraction of p-values $<$(1-CL) as a function of the fraction of the data sample, which are outliers. Results are shown for outliers corresponding to relative errors in the spectrum of 15% (left), 20% (center) and 25% (right).

B. Fluctuations and outliers

Experimental data often contain fluctuations and outliers. In this study the response of the GoF tests to such perturbations is examined. The reference distribution of such use case is taken from the distribution of X-ray fluorescence energies of K-shell transitions for different elements of the periodic system. Experimental distributions are generated for each element as values displaced from their reference according to Gaussian experimental errors (see Fig. 5). Such process can generate outliers, i.e. points largely displaced from the original reference.

In the case of fluctuations the relative errors are assumed to apply to the entire data set; errors from 3% to 10% are considered. For outliers, relative errors are introduced only for a fraction of data points, but are taken to be considerably larger than the simulated fluctuations (considered values range from 15% to 25%).

The results reveal that all examined unbinned GoF tests are not capable of identifying either the fluctuations or the outliers in the spectrum, i.e. their p-values were between 0.7 and 1.0 in all cases; hence the focus is on the evaluation of binned tests.

Fig. 5 presents the fraction of p-values $<$(1-CL) in dependency on the relative error in the data sample. The Tiku, Cramer-Von Mises and Anderson-Darling algorithms show a slowly increasing response to the fluctuations for errors larger than 4%; the latter algorithm is slightly more sensitve than

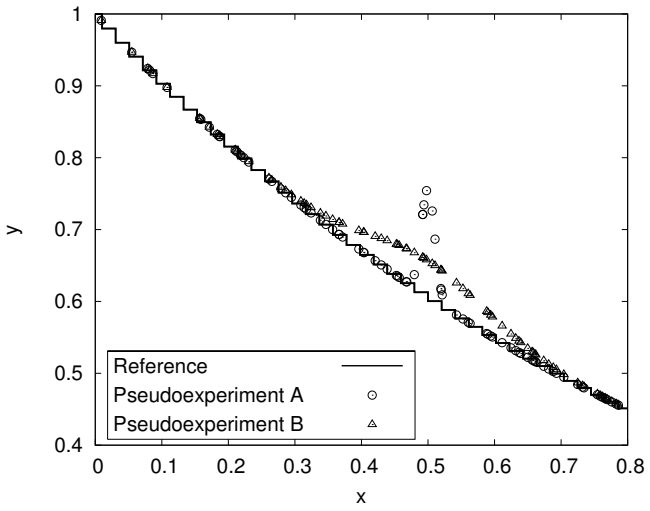


Fig. 7. Exponential background spectrum and data samples deriving from pseudoexperiments which superimposes a Gaussian signal.

the first two tests. In contrast, the χ^2 test responds only for higher fluctuations; but its sensitivity grows considerably faster than for the other algorithms and rapidly exceeds their level of reaction to the perturbations.

In the case of outliers, the behaviour of the Tiku, Cramer-Von Mises and Anderson-Darling tests is nearly identical; hence only the Anderson-Darling and χ^2 algorithms are compared. The curves in Fig. 6 presents the response with respect to the fraction of outliers in the data sample. Results are shown for outliers of different magnitude: for the smallest considered outliers the Anderson-Darling test is more sensitive than the χ^2, while the behaviour swaps if the magnitude increases. This is consistent with the findings obtained for fluctuations.

C. Signal over background

The third physics use case investigates if the GoF tests are capable of identifying a signal over a background. As illustrated in Fig. 7, sample distributions deriving from the superposition of a Gaussian and an exponential curve are compared against a plain exponential background spectrum. Only unbinned tests were considered.

Fig. 8 presents the fraction of p-values <(1-CL) as a function of sigma. The various algorithms recognize the signal, but respond with different discriminating capabilities when sigma is increased; the Kolmogorov-Smirnov and Watson tests are most sensitive to broader signals, while the Anderson-Darling test shows the lowest level of response. The Cramer-Von Mises and Tiku algorithms are again nearly identical.

IV. Conclusions

An evaluation of the relative power of statistical algorithms for the comparison of data distribution was performed for different physics use cases. It was demonstrated, that certain GoF test show a higher sensitivity to perturbations, and hence might be preferable for analysis procedures in similar applications.

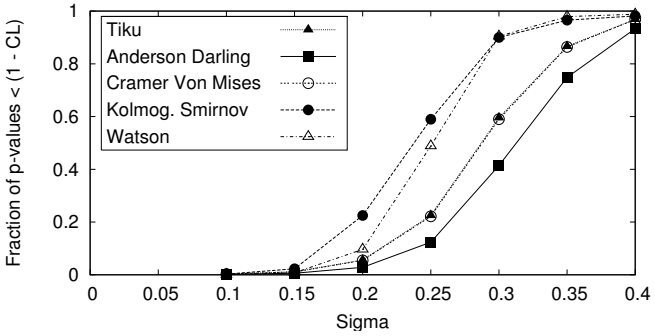


Fig. 8. Gaussian signal over background: p-values <(1-CL) as a function of sigma.

Acknowledgment

The authors would like to thank the CATANA group (INFN LNS, Catania/Italy) for providing the experimental proton Bragg peak data used in this study.

References

[1] G.A.P. Cirrone et al., "A Goodness-of-Fit Statistical Toolkit", *IEEE TNS*, Vol 51, Issue 5, pp. 1056-63, 2004.

[2] B. Mascialino et al. "New Developments of the Goodness-of-Fit Statistical Toolkit" *IEEE TNS*, Vol 53, Issue 6, pp. 3834-41, 2006.

[3] G. Barrand, P. Binko, M. Donszelmann, A. Johnson, and A. Pfeiffer, "Abstract interfaces for data analysis: component architecture for data analysis tools", in *Proc. of CHEP 2001 Int. Conf. on Computing in High Energy and Nuclear Physics*, pp. 215-218, Science Press, Beijing, 2001.

[4] R. Brun and F. Rademakers, "ROOT, An Object-oriented Data Analysis Framework", *NIM - Section A*, vol. 389, pp. 81-86, 1997.

[5] B. Aslan, G. Zech, "Comparison of different goodness-of-fit tests" in *Proc. of Conf. on Advanced Statistical Techniques in Particle Physics*, pp. 166-175, Durham, 2002.

[6] G.A.P. Cirrone et al., "A 62 MeV proton beam for the treatment of ocular melanoma at Laboratori Nazionali del Sud-INFN (CATANIA)", in *IEEE Trans. Nucl. Sci.*, vol. 51, no. 3, pp. 860-865, 2004.