

# Application of statistical methods for the comparison of data distributions

S. Guatelli, B. Mascialino, A. Pfeiffer, M.G. Pia, A. Ribon, P. Viarengo

**Abstract**—Data analysis is an essential section of all physics experiments; in spite of this only a few analysis standard toolkits are available. Concerning the comparison between distributions, almost all these toolkits are limited to the Chi-squared test. Statistics provides a whole chapter of Goodness-of-Fit tests, from the Chi-squared to tests based on maximum distance (Kolmogorov-Smirnov, Kuiper, Goodman), to tests based on quadratic distance (Fisz-Cramer-von Mises, Anderson-Darling, Tiku). All of these Goodness-of-Fit tests have been collected in a new open-source Statistical Toolkit. This Toolkit matches a sophisticated statistical data treatment with the most advanced computing techniques, such as object-oriented technology with the use of design patterns and generic programming. None of the Goodness-of-Fit tests included in the system is optimum for every case. Unfortunately, statistics does not provide a universal recipe for specific distributions and furthermore the only rare available guidelines refer to the comparison between smooth theoretical distributions. With the aim of helping the user in the algorithm choice, we present the results of an intrinsic statistical comparison among many of the Goodness-of-Fit tests contained in the Statistical Toolkit in terms of relative efficiency.

**Index Terms**—Data Analysis, Goodness-of-Fit testing, Goodness-of-Fit Statistical Toolkit, Power Comparison.

## I. INTRODUCTION

STATISTICAL comparison of distributions is an essential section of data analysis in any field and in particular in high energy physics experiments. In spite of this, only a few basic tools for statistical analysis were available in the public domain FORTRAN libraries for high energy physics, such as CERN Libraries (CERNLIB) [1] and PAW [2]. Nowadays the situation is unchanged even among the libraries for data analysis of the new generation, such as for instance Anaphe [3], JAS [4], Open Scientist [5] and Root [6]. In any case, the user is offered to compare two data distributions by means of the Chi-squared test and little more, ignoring the variety of tests that statistics offers. In fact, it is well known that non-parametric statistics provides a whole chapter of Goodness-of-Fit (GoF) tests, from the Chi-squared to tests based on maximum distance (Kolmogorov-Smirnov, Kuiper, Goodman), to tests based on

quadratic distance (Fisz-Cramer-von Mises, Anderson-Darling, Tiku). All of these GoF tests have been recently collected in a new open-source Statistical Toolkit [7], that is downloadable from the web [8], offering the user the possibility of comparing two samples.

Unfortunately, statistics literature does not provide clear guidelines concerning performance comparisons among tests. Some power comparisons among non-parametric tests are available [12], and predominantly concern tests aimed at detecting the location differences between two distributions. More in detail, the comparison studies available in literature are very interesting, but, being berthed to the very specific alternatives set, do not provide a general guideline under more non-parametric conditions.

For these reasons, the purpose of this paper is to compare the power of several GoF two-sample tests in many different situations. More in detail, we focused our attention on a selection of the GoF tests contained in the GoF Statistical Toolkit.

## II. HYPOTHESIS TESTING: THE TWO-SAMPLE PROBLEM

A statistical hypothesis is an assertion about the probability function of one or more random variables or a statement about the populations from which one or more samples are drawn. The inference can concern its form, shape or the parameter values. Statistical inference is often concerned with the comparison between observed data and theoretical functions. It must be noticed that many applications deal with the comparison of two samples, in physics as well as in statistics or in other fields. If the null hypothesis  $H_0$  states that the underlying distributions of the two samples are identical in shape and differ only in location or scale (the *location-scale* problem), many statistical tests are available, parametric and non-parametric. The distributions considered in these studies cover almost all the range of possibilities: symmetric, asymmetric and among these last ones, short, medium and long tailed distributions [10], [11], [?]. In the opposite case, in which the underlying distributions of the two samples differ also in shape (the *general shape alternative* problem), among the ones available in literature Kolmogorov-Smirnov and Cramér-von Mises tests are the most known and, therefore, used [9]. Almost all the studies available in literature concentrate on location-scale alternatives.

More in detail, let us consider two real valued random variables  $X$  and  $Y$ , and let  $(X_1, \dots, X_n)$   $(Y_1, \dots, Y_m)$  be two samples of independent and identically distributed observations with empirical distribution functions (*edf*)  $F_n$ , and

Manuscript received November 1, 2004.

S. Guatelli, B. Mascialino and M.G. Pia are with the Italian Institute for Nuclear Research, via Dodecaneso 33, 16146 Genova - Italy (telephone: 0039 010 353 6420 - e-mail: Susanna.Guatelli@ge.infn.it, Barbara.Mascialino@ge.infn.it, MariaGrazia.Pia@ge.infn.it)

A. Pfeiffer and A. Ribon are with CERN, CH 1122, Geneva 23 - Switzerland (telephone: 0041 22 767 5029 - e-mail: Andreas.Pfeiffer@cern.ch, Alberto.Ribon@cern.ch).

P. Viarengo is with the National Cancer Research Institute (telephone: 0039 010 5600 796 - e-mail: paolo.viarengo@istge.it).

Hypothesis	General	Location-scale
$H_0$	$F(z) = G(z)$	$F(z) = G(z)$
$H_1$	$F(z) \neq G(z)$	$F(z) = G(\frac{z-\theta}{\tau})$

TABLE I

HYPOTHESIS UNDER STUDY IN THE GENERAL ALTERNATIVE AND IN THE LOCATION-SCALE PROBLEM.

$G_m$ . Let us consider  $(Z_1, \dots, Z_N)$  as the combined sample  $(X_1, \dots, X_n, Y_1, \dots, Y_m)$ , with  $N = n + m$ .

According to these definitions, the two hypothesis under test in the general alternative and location-scale two-sample problem are summarised in Table 1. In both cases, the null hypothesis must be valid for all the values of  $z$ . In the general alternative, the null hypothesis will be rejected if  $F$  and  $G$  will differ for at least one  $z$ . In the location-scale problem, the alternative hypothesis depends on two parameters:  $\theta$  (the location parameter), and  $\tau$  (the scale parameter). These two parameters are subject to the following constraints:  $-\infty < \theta < \infty$ ,  $\theta \neq 0$  and/or  $\tau > 0$ ,  $\tau \neq 1$ .

It must be observed that:

- in the specific case  $\tau = 1$ , the alternative hypothesis concerns a pure location problem,
- if  $\theta = 0$ , the two-samples test will reduce to a pure scale alternative.

On the basis of the location-scale alternative hypothesis, it is stated that the two distributions  $F$  and  $G$  are differently located or scaled, but it is assumed that the shape of the two distributions is the same.

In the general hypothesis, the two distributions  $F$  and  $G$  are assumed different, not simply in location and/or scale, but also in shape. As an example, in this specific case  $F$  could be the gaussian distribution and  $G$  the exponential one.

### III. GOF TWO-SAMPLE TESTS

GoF tests measure the compatibility of a random sample with a theoretical probability distribution function or between the empirical distributions of two different populations coming from the same theoretical distribution. From a general point of view, the aim may consist also in testing whether the distributions of two random variables are identical against the alternative that they differ in some way.

#### A. Chi-squared Test

With the purpose of quantifying the measure of the deviation between two distributions, many software toolkits for physics data analysis solve the problem by means of the well known and wide-spread Chi-squared test. This test was introduced to describe discrete distributions, but it can be useful also in case of unbinned distributions. In this case the researcher must group data into classes, sacrificing in this way a good deal of the information conveyed by the distribution itself. In spite of the fact that this test has a general applicability, it must be noticed

that the Chi-squared asymptotic distribution is *not* valid if the theoretical frequencies involved in the computation are lower than 5. For these reasons, a powerful and up-dated statistical Toolkit for physics data analysis should supplement the Chi-squared test with other statistical tests, involving individual sample values.

#### B. The Kolmogorov-Smirnov Family

A common alternative to the Chi-squared tests includes tests based on Kolmogorovs *edf* definition [13]. These tests are: **Kolmogorov-Smirnov** (KS) [13], [15], [16], [18], **Goodman** [19] and **Kuiper** [17]. In any case, the test statistics is a linear function of the maximum vertical distance between the *edfs* of the two distributions. Tests belonging to the Kolmogorov-Smirnov family can be applied only to *continuous* distributions. Some other limitations are related with the fact that these tests tend to be more sensitive near the center of the distribution with respect to the tails.

#### C. The Anderson-Darling Family

The Anderson-Darling family of tests measures the integrated quadratic deviation of the two *edfs* suitably weighted by a weighting function  $\psi(F(x))$ . Different mathematical formulations of the weighting function  $\psi$  define the **Anderson-Darling** (AD) [20], [24], [28], **Fisz-Cramer-von Mises** (CVM) [18], [22], [25]–[27] and **Tiku** [23] test statistics. These tests can be performed on both binned or unbinned data and they are satisfactory for symmetric and right-skewed distributions [21]. It must be pointed out the fact that these tests give more weight to the tails than the test belonging to the Kolmogorov-Smirnov type.

### IV. THE GOODNESS-OF-FIT STATISTICAL TOOLKIT

The Goodness-of-Fit Statistical Toolkit [7] is an easy to use, up-dated and versatile tool for data comparison in physics analysis. This Toolkit provides some GoF algorithms for the comparison of data distributions. The GoF Toolkit gathers together some of the most important GoF tests available in literature, such as the ones described in the previous Section. Its aim is to provide a wide set of algorithms in order to test the compatibility of the distributions of two variables (two-sample problem).

The GoF Toolkit matches a sophisticated data treatment with the most advanced computing techniques, such as object oriented technology with the use of design patterns and generic programming. The adoption of the Abstract Interfaces for Data Analysis [29] for its user layer decouples the usage of the GoF Toolkit from any concrete analysis system. The code is open-source and can be downloaded from the web together with user and software process documentation [8].

## V. TESTS PERFORMANCE ASSESSMENT: POWER COMPARISON

From its definition, the power of a test represents the probability  $P$  that the test statistic  $T$  will lead to the rejection of the null hypothesis  $H_0$ :

$$\text{Power} = P(T \in R)$$

It is directly related with the probability of a correct decision, which implies  $H_0$  false. In this case, the power value will depend on:

- the degree of falseness of  $H_0$ ,
- the confidence level  $\alpha$ , *i.e.* the subset of real numbers chosen for  $R$ ,
- the sample sizes.

A test is said to be more powerful for a specific alternative hypothesis if no other test of the same size has greater power against the same alternative.

Power comparisons in non-parametric testing are generally performed in the following way:

- 1) selecting the specific hypothesis tests one is interested in,
- 2) designing for every test the same hypothesis testing situation.

## VI. POWER STUDY

We define a *pseudoexperiment* a random drawing of two samples;  $S_1$  of size  $n$  and  $S_2$  of size  $m$ , from a parent distribution. Given these two samples, we can calculate the distance,  $d$ , between them according to the GoF statistical test described in detail in Section III, and then from the distance we can calculate the corresponding p-value. For each test, the p-value computed by the GoF Toolkit derives from the analytical calculation of the *asymptotic* distribution, often depending on the samples sizes.

For each pseudoexperiment we obtain:

$$S_1^{(j)}, S_2^{(j)} \rightarrow d^{(j)}, p^{(j)}$$

where  $j = 1, \dots, k$ , with  $k$  representing the number of pseudoexperiments performed.

With the aim of avoiding bias, the two samples object of the comparison are drawn from the probability distributions, which were investigated in a recent paper [9]. These distributions differ in shape, covering at the same time a wide range of short-tailed up to very long tailed distributions as well as symmetric ones. They include also two different functions deriving from the contamination of normal distributions.

With the aim of quantifying these parameters, two indexes are introduced, a *skewness* index and a *tailweight* index [9]:

$$S = \frac{x_{0.975} - x_{0.5}}{x_{0.5} - x_{0.025}},$$

$$T = \frac{x_{0.975} - x_{0.025}}{x_{0.875} - x_{0.125}}.$$

Distributions are classified in terms of  $S$  and  $T$  as follows:

- $S = 1$ : the distribution is a symmetric one,

	Probability distributions	Skewness $S$	Tailweight $T$
$f_1$	uniform	1	1.267
$f_2$	$N(0, 1)$	1	1.704
$f_3$	$\frac{1}{2}e^{- x }$	1	2.161
$f_4$	$\frac{1}{\pi} \frac{1}{1+x^2}$	1	5.263
$f_5$	$e^{-x}$	4.486	1.883
$f_6$	$0.9N(0, 1) + 0.1N(0, 9)$	1	1.991
$f_7$	$0.5N(1, 4) + 0.5N(-1, 1)$	1.769	1.693

TABLE II

DETAILS OF THE SEVEN PARENT PROBABILITY DISTRIBUTIONS FROM WHICH THE TWO SAMPLES WERE DRAWN. COLUMNS THREE AND FOUR INDICATE THE MEASURES OF SKEWNESS AND TAILWEIGHT OF THE DISTRIBUTIONS.

- $S < 1$ : the distribution is skewed to the left,
- $S > 1$ : the distribution is skewed to the right,
- $T$  is always greater than one, the longer the tail, the greater the value of  $T$ .

The measures of  $S$  and  $T$  are location and scale invariant. The seven parent distributions are listed in Table 2, together with their  $S$  and  $T$  indexes.

Power is evaluated as the ratio of the number of correctly rejected pseudoexperiments with respect to the number  $k$  of pseudoexperiments performed. To evaluate the power of the specific tests we generated  $k = 1000$  pairs of equal-sized samples ( $n = m = 25, 40, 60$  in this study). The confidence level was set to 5%.

## VII. RESULTS

Because of the limited number of pages, we are going to summarise the results obtained providing in any of the following subsections only a few examples.

### A. Location and scale alternatives

As the p-value derives from the analytical calculation of the asymptotic distribution, there is an evident dependence on the sample size. Power increases proportionally to the samples size. When the asymptotic limit is reached, the GoF tests power approaches the confidence level set ( $\alpha = 0.05$ ).

### B. Different shape alternatives

1) *Symmetric versus symmetric*: Let us fix one of the two parent distributions: the double exponential. Let us compare it with samples drawn from the other symmetric distributions considered (presenting different tailweights values). We find that:

- for short- and medium-tailed distributions the different tests have almost the same power (as an example,  $KS \sim CVM \sim AD$ ),
- for long-tailed distributions the AD test turns out to have a power that is double with respect to the other tests (as an example,  $KS \sim CVM < AD$ ).

Quantitative results concerning this issue are presented in Table 3.

Tailweight	Power KS	Power CVM	Power AD
1.704	5.1	5.7	5.4
1.991	7.1	6.8	6.5
5.263	9.7	9.6	18.5

TABLE III

GENERAL SHAPE ALTERNATIVE PROBLEM: SYMMETRIC PARENT DISTRIBUTION VERSUS SYMMETRIC PARENT DISTRIBUTION. FIXED ONE OF THE TWO PARENT DISTRIBUTIONS ( $f_3$ ), WE SEE THAT THE POWER (EXPRESSED IN %) OF THE TESTS INCREASES WITH THE TAILWEIGHT OF THE SECOND PARENT DISTRIBUTION. FOR A SHORT- AND/OR MEDIUM-TAILED DISTRIBUTION THE THREE TESTS SELECTED SEEM TO BE EQUIVALENT, WHILE IN THE CASE OF A LONG-TAILED ONE THE AD TEST HAS A POWER DOUBLE THAN THE OTHER ONES.

2) *Skewed versus symmetric*: Let us consider the practical example of the skewed distribution  $f_7$  ( $S = 1.769$ ) and let us compare it respectively with three symmetric ones ( $S = 1$ ) having different tailweights  $T$ :

- $f_7 \leftrightarrow f_2$ ,
- $f_7 \leftrightarrow f_6$ ,
- $f_7 \leftrightarrow f_3$ .

We obtain that in any case the AD test turns out to be the most powerful test, with respect to the KS and CVM ones. More in detail, the KS power decreases as the tailweight of the second parent distribution considered increases, while the CVM power shows the opposite behavior.

On the basis of the results obtained in this study and very briefly illustrated in the previous Sections, we provide a very schematic guideline. To select one test in practice:

- 1) first classify the type of the distributions in terms of skewness  $S$  and tailweight  $T$ ,
- 2) choose the most appropriate test for the classified type of distribution on the basis of the following criteria:

	Short-tailed ( $1 < T < 1.3$ )	Medium-tailed ( $1.3 < T < 2$ )	Long-tailed ( $T > 2$ )
$S \sim 1$	KS	KS-CVM	CVM-AD
$S > 1.5$	KS-AD	AD	CVM-AD

Of course this is only a schematic classification of the results we obtained on the basis of the features of the parent distributions we considered. The thresholds between short-, medium- and long-tailed distributions are to be considered only indicative and not as absolute values. The same is for the threshold for skewness. Further studies are in progress aimed at investigating much in detail this issue.

### VIII. REAL DATA EXAMPLES

On the basis of the results achieved, let us consider two practical real data examples:

- Example 1: an X-ray fluorescence spectrum (see Fig. 2a),
- Example 2: a dosimetric distribution from a medical LINAC (see Fig. 2b).

In both cases, we would like to compare two data-distributions  $F$  and  $G$ , one coming from experimental data ( $X$ -variable) and the other from Geant4 simulations ( $Y$ -variable). The aim is to measure the compatibility of the two distributions.

What are the most suitable GoF tests for these data? In order to take a decision, let us calculate for both  $X$ - and  $Y$ -variables the values of  $S$  and  $T$ :

Example 1	$S_1$	$T_1$	Example 2	$S_2$	$T_2$
$X$ -variable	4.00	1.43	$X$ -variable	1.53	1.34
$Y$ -variable	4.00	1.50	$Y$ -variable	1.27	1.36

- Example 1: both the  $X$ - and  $Y$ -variables have a medium tail, but they are extremely skewed. On the basis of the results scheme we select the Anderson-Darling test.
- Example 2: both the  $X$ - and  $Y$ -variables have a short-medium tail, with a moderate skewness. On the basis of the results scheme we select the Kolmogorov-Smirnov test.

The following Table summarises the statistical results obtained in these two practical two-sample problems.

Test selected	Test statistics	Test result
Anderson-Darling	$A^2 = 0.085$	$p > 0.05$
Kolmogorov-Smirnov	$D = 0.27$	$p > 0.05$

### IX. CONCLUSIONS

This paper summarises a general and wide study concerning the comparison of several non-parametric test (the two-sample problem). Both the location-scale problem and the general shape one have been considered with several different-shaped distributions (from symmetric to skewed, including short-, medium- and long-tailed ones).

As a general consideration, it turns out that there is no clear *winner* among all the GoF tests considered for all the parent distributions selected. Probably this is related to the fact that the different tests behave better/worse depending on the specific distributions we are considering (in terms of their skewness  $S$  and tailweight  $T$ ).

As a general guideline, following [9], [30], [31], we suggest to classify the type of underlying distribution functions  $F$  and  $G$  with respect to some measures like skewness  $S$  and tailweight  $T$  and then to select the most appropriate GoF test.

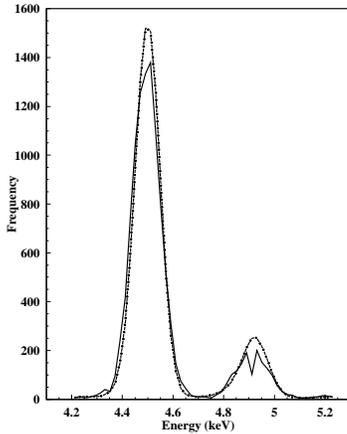
Deeper studies are now in progress aimed at investigating this crucial issue, that is almost neglected even in the statistics domain, further.

### ACKNOWLEDGMENT

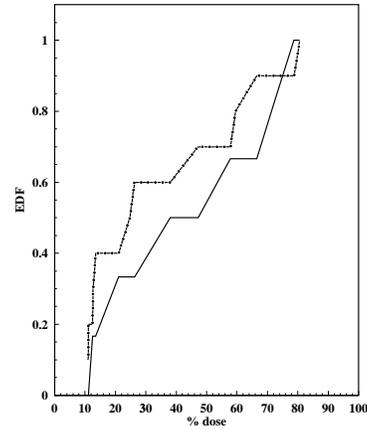
The authors would like to thank Professor Subhabrata Chakraborti for the fruitful collaboration concerning the main issues of this paper.

### REFERENCES

- [1] J. Shiers, "CERNLIB - CERN Program Library Short Writeups", CERN Geneva, 1996.
- [2] R. Bock, R. Brun, O. Couet, J.C. Marin, R. Nierhaus, L. Pape, et al., "PAW-Towards a Physics Analysis Workstation", *Computer Physics Communications*, vol. 45, pp. 181-190, 1987.



(a) X-ray fluorescence spectrum example: original distributions of the X- and Y-variables.



(b) Dosimetric distribution from a medical LINAC example: the two *edfs* of the X- and Y-variables.

- [3] O. Couet, B. Ferrero-Merlino, Z. Molnár, J.T.Moscicki, A.Pfeiffer, M.Sang, "ANAPHE-OO Libraries and Tools for Data Analysis", Proceedings of the CHEP 2001 Conference, Beijing, China, Spetember 2001.
- [4] T. Johnson, V. Serbo, M. Turri, M. Donszelmann, J. Perl, "JAS3-A general purpose Data Analysis Framework for High Energy Physics and beyond", Proceedings of the IEEE Nuclear Science Symposium, Portland, Oregon, October 2003.
- [5] <http://www.lal.in2p3.fr/OpenScientist>
- [6] R. Brun, F. Rademakers "ROOT, An object-oriented Data Analysis Framework", *NIM - Section A*, vol.389, pp.81-86, 1997.
- [7] G.A.P. Cirrone, S. Donadio, S. Guatelli, A. Mantero, B. Mascialino, S. Parlati M.G. Pia, A. Pfeiffer, A. Ribon, P. Viarengo "A Goodness-of-Fit Statistical Toolkit" Transactions on Nuclear Science, vol. 51 (5), pp. 2056-2063, 2004.
- [8] <http://www.ge.infn.it/geant4/analysis/HEPstatistics/index.html>
- [9] H. Buning, S. Chakraborti *Power comparison of several two-sample tests for general alternatives* Allgemeines Statistisches Archiv, 1999 83: 190-210.
- [10] H. Buning, "Robust und adaptive tests", Ed. Berlin: De Gruyter, 1991.
- [11] B.S. Duran, "A survey of nonparametric tests for scale", *Communications in statistics - Theory and Methods*, 1976 5: 1287-1312.
- [12] M.A. Stephens, "Introduction to Kolmogorov (1933) on the empirical determination of a distribution", S. Kotz, N.L. Johnson, "Breakthrough in Statistics", vol. 2, Ed. New York: Springer Verlag, 1992.
- [13] A.N. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione", *Giornale dell'istituto italiano degli attuari*, vol.4, pp. 83-91, 1933.
- [14] W.J. Conover, "Practical Nonparametric Statistics", Ed. New York: John Wiley & Sons, 1991
- [15] N.V. Smirnov, "Sur les écarts de la courbe de distribution empirique (Russian/French summary)", *Matematicheski Sbornik N.S.*, vol.6, pp. 3-26, 1939.
- [16] N.V. Smirnov, "Table for estimating the Goodness-of-Fit of empirical distributions", *Annals of Mathematical Statistics*, vol.19, pp. 279-281, 1948.
- [17] N.H. Kuiper, Proc. Koninkl. Neder. Akad. van. Wetenschappen A 63 (1960) 38.
- [18] D.A. Darling, "The Kolmogorov-Smirnov, Cramer-von Mises tests", *Annals of Mathematical Statistics*, vol. 28, pp. 823-838, 1957.
- [19] L.A. Goodman, "Kolmogorov-Smirnov tests for psychological research", *Psychological Bulletin*, vol. 51, pp. 160-168, 1954.
- [20] T.W. Anderson, D.A. Darling, "A test of Goodness of Fit", *Journal of the American Statistical Association*, vol.49, pp. 765-769, 1954.
- [21] T.W. Aksenov, M.A. Savageau, "Mathematica and C programs for minimum distance estimation of the S distribution and for calculation of Goodness-of-Fit by bootstrap", 2002, in press.
- [22] T.W. Anderson, "On the distribution of the two-sample Cramer-von Mises criterion", *Annals of Mathematical Statistics*, vol. 33, pp. 1148-1159, 1962.
- [23] M.L. Tiku, *Biometrika* 52 (1965) 630.
- [24] A. Heckert (2003, April 4). NIST, Available: <http://www.itl.nist.gov/div898/software/dataplot/refman1/>
- [25] H. Cramer, "On the composition of elementary errors", *Second paper: statistical applications*, *Skand. Aktuarietidskrift*, vol.11, pp. 13-74, pp. 141-180, 1928.
- [26] R. von Mises, "Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik", *Leipzig: F.Duticke*, 1931.
- [27] M. Fisz, "On a result by M.Rosenblatt concerning the von Mises-Smirnov test", *Annals of Mathematical Statistics*, vol. 31, pp. 427-429, 1960.
- [28] T.W. Anderson, D.A. Darling, "Asymptotic theory of certain Goodness of fit criteria based on stochastic processes", *Annals of Mathematical Statistics*, vol. 23, pp. 193-212, 1952.
- [29] G. Barrand, P. Binko, M. Donszelmann, A. Johnson, A. Pfeiffer, "Abstract Interfaces for Data Analysis - Component Architecture for Data Analysis Tools", proceedings of CHEP, Beijing, September 2001, Science Press, pp. 215-218.
- [30] R.V. Hogg *Adaptive robust procedures. A partial review and some suggestions for future applications and theory.* Journal of the American Statistical Association, 1974 69: 909-927.
- [31] R.V. Hogg, D.M. Fisher, R.H. Randles *A two sample adaptive distribution-free test* Journal of the American Statistical Association, 1974 70: 656-661.