Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Data re-arranging techniques leading to proper variable selections in high energy physics

To cite this article: Václav Ks and Petr Bou 2017 *J. Phys.: Conf. Ser.* **936** 012063

View the article online for updates and enhancements.

This content was downloaded from IP address 188.184.3.52 on 29/01/2018 at 13:50

IOP Conf. Series: Journal of Physics: Conf. Series **936** (2017) 012063 doi:10.1088/1742-6596/936/1/012063

Data re-arranging techniques leading to proper variable selections in high energy physics

Václav Kůs, Petr Bouř

Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Trojanova 13, 12000 Praha 2, Czech Republic

E-mail: vaclav.kus@fjfi.cvut.cz, petr.bour@fjfi.cvut.cz

Abstract. We introduce a new data based approach to homogeneity testing and variable selection carried out in high energy physics experiments, where one of the basic tasks is to test the homogeneity of weighted samples, mainly the Monte Carlo simulations (weighted) and real data measurements (unweighted). This technique is called 'data re-arranging' and it enables variable selection performed by means of the classical statistical homogeneity tests such as Kolmogorov-Smirnov, Anderson-Darling, or Pearson's chi-square divergence test. P-values of our variants of homogeneity tests are investigated and the empirical verification through 46 dimensional high energy particle physics data sets is accomplished under newly proposed (equiprobable) quantile binning. Particularly, the procedure of homogeneity testing is applied to re-arranged Monte Carlo samples and real DATA sets measured at the particle accelerator Tevatron in Fermilab at DØ experiment originating from top-antitop quark pair production in two decay channels (electron, muon) with 2, 3, or 4+ jets detected. Finally, the variable selections in the electron and muon channels induced by the re-arranging procedure for homogeneity testing are provided for Tevatron top-antitop quark data sets.

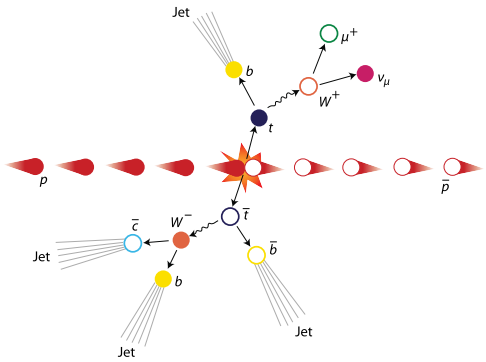
1. Particle physics tasks and data structure

In this paper, we deal with the top-antitop quark pair production after annihilated collisions of accelerated proton and antiproton beams at Tevatron synchrotron within DØ experiment in Fermilab [1] at 2TeV centered energy. The data set (DATA) measured is provided with the Monte Carlo simulation set (MC). The top-antitop quark decay process, together with the levels of contamination of the relevant signal S with respect to different background decays B, are shown in Figure 1. The dimensions of MC and DATA are both limited to 46 selected physical variables of various ranges and supports listed in Table 1.

Table 1. Short names of variables in DØ top quark decay channel.

1	Apla	9	M0nl	17	Centr	25	LepEta	33	DRJetJet2	41	Jteta3
2	Spher	10	M1nl	18	DRminejet	26	Jetm	34	DRJetJet3	42	Jteta4
3	HTL	11	MT0nl	19	DiJetDrmin	27	Metper	35	Jtpt1	43	JtMVA1
4	JetMt	12	Met	20	Ht	28	Metpar	36	Jtpt2	44	JtMVA2
5	HT3	13	Mtt	21	Ht20	29	DphiJetJet1	37	Jtpt3	45	JtMVA3
6	MEvent	14	Mva_max	22	Ktminp	30	DphiJetJet2	38	Jtpt4	46	JtMVA4
7	MT1NL	15	Wmt	23	Lepdphimet	31	DphiJetJet3	39	Jteta1		
8	M01mall	16	Wpt	24	LeppT	32	DRJetJet1	40	Jteta2		

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd



#jets	Electron channel	Muon channel
2	2011176	2001522
3	719484	668072
4+	295932	277214
#jets	S/B ratio (%)	S/B ratio (%)
2	1.13%	0.88%
3	12.24%	10.98%
4+	39.09%	39.34%

Figure 1. Number of generated samples in Top-antitop quark pair production from $p\bar{p}$ annihilation at Tevatron and signal (S) to background (B) ratio in MC simulations.

Prior to any supervised physical task based on MC samples [2], it is vital to guarantee the homogeneity of MC and DATA. However, in most cases, the MC samples are re-weighted according to the actual detection ability of the accelerator. In TMVA ROOT package, the homogeneity testing procedures are simply based on weighted histograms found for equidistant binning in each single dimension (variable). We propose the homogeneity testing approach based on equiprobable quantile binning and re-arranged MC sample in Sections 2 and 3, which is applicable to arbitrary weighted observations. Then we apply the tests to the DØ top quark production at Tevatron. This data set of 46 variables represents a bunch of wide-ranging and significantly complex array of observations, which are sometimes positive, negative, bounded, discrete or continuous, etc. When this homogeneity is not possible to accept, the variable selection is consequently induced, i.e., we are forced to considerable dimensionality reduction in all six top quark decay channels, as it is treated in Section 4.

2. Tests of homogeneity for weighted samples

Let X_1 denote the random variable representing the selected MC physical variable distributed by F and X_2 be the corresponding DATA associated variable distributed by G. Then the homogeneity testing means the statistical test of hypothesis $H_0 : F = G$ at a given significance level $\alpha \in (0, 1)$. Let $\boldsymbol{X}_1 = (X_1^{(1)}, \ldots, X_{n_1}^{(1)})$ denote repeated real valued random variables identically and independently distributed (i.i.d.) from a cumulative distribution function (cdf) F and let $(w_1^{(1)}, \ldots, w_{n_1}^{(1)})$ be their corresponding nonnegative weights. Then we define *weighted empirical distribution function* (wedf) of $\boldsymbol{X}_1$ by

$$F_{n_1}^{W_1}(x) = \frac{1}{W_1} \sum_{i=1}^{n_1} w_i^{(1)} I_{(-\infty,x]}(X_i^{(1)}), \qquad \text{where } W_1 = \sum_{i=1}^{n_1} w_i^{(1)}, \tag{1}$$

and $I_A(\cdot)$ denotes the characteristic function (indicator) of the set A. Assume further that $\boldsymbol{X}_2 = (X_1^{(2)}, \ldots, X_{n_2}^{(2)})$ represents repeated i.i.d. variables under cdf G with the corresponding nonnegative weights $(w_1^{(2)}, \ldots, w_{n_2}^{(2)})$. Then the same definition (1) gives us the wedf $G_{n_2}^{W_2}$ with $W_2 = \sum_{j=1}^{n_2} w_j^{(2)}$. In our HEP representation, $\boldsymbol{X}_2$ corresponds to DATA and so all the weights $w_j^{(2)} = 1$, $W_2 = n_2$, and $G_{n_2}^{W_2} = G_{n_2}$ is the usual edf.

In this paper, we use classical two sample Kolmogorov-Smirnov (K-S) test [3] and two sample Anderson-Darling (A-D) test [4, 5], both applied to wedf's. Further, the χ^2 divergence test of homogeneity [6] is applied to weighted histograms, which does not employ the concept of wedf's. First of all, it is necessary to set up the number of bins k and the subsequent binning $t_0 < \ldots < t_k$. The most frequently used binning is the equidistant one made within a chosen interval $[a, b]$

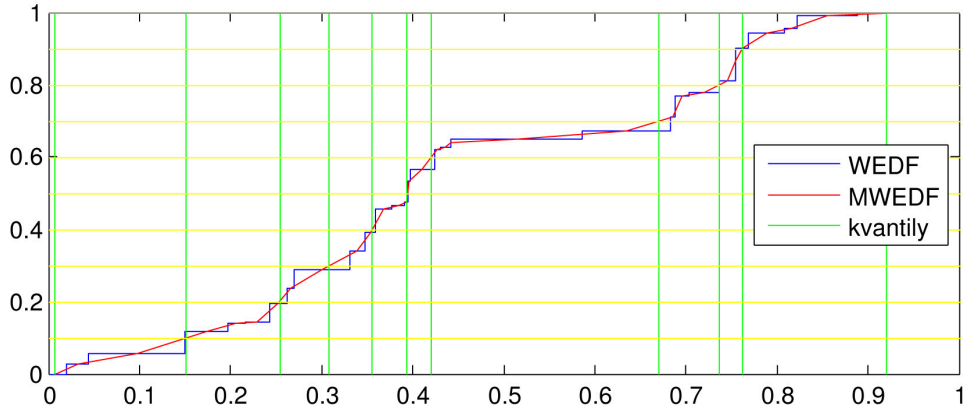


Figure 2. Quantile binning technique for weighted pooled sample $\{\boldsymbol{X}_1, \boldsymbol{X}_2\}$ through mwedf.

spread out over the data sample. However, this approach can be generally insufficient due to poor diversification of data into the equidistant bins, thus the local distribution of data can be lost. Therefore, for a given fixed k, we propose using of the so called equiprobable *quantile binning*, $t_0 < \ldots < t_k$, developed in [7], which assigns approximately the same portion of (weighted) observations to every cell $(t_{j-1}, t_j]$. This method of binning initially consists in finding out the wedf of pooled sample $\{\boldsymbol{X}_1, \boldsymbol{X}_2\}$ with $W = W_1 + W_2$ weighted observations, then creating a modified continuous version of wedf (say mwedf) by linear interpolation of all neighbouring middle points of the constant segments of wedf. Consequently, equi-balanced (j/k)-th quantiles t_j of mwedf are evaluated for all $j = 1, \ldots, k-1$, as illustrated in Figure 2. Thus the quantile binning guarantees the uniformly distributed information contained in weighted data sample since approximately the same sum of weights is concentrated in every cell.

After we have defined the equiprobable pattern of quantile binning $t_0 < \ldots < t_k$ for a fixed k, we make a judicious choice of k. Because of the large number of observations in DATA and large sums of weights in MC samples, the χ^2 divergence test could potentially loose its power with excessively increasing number of bins k, as was shown in [8]. Thus, we decided to choose the following wise choice of the number of bins $k = \lceil 1 + \log_2 W \rceil$ proposed and tested in [9].

Table 2. Sample sizes and sums of weights W_1, W_2 in MC and DATA measurements in top quark decay channel at Tevatron, $W = W_1 + W_2$, and number of bins $k = \lceil 1 + \log_2 W \rceil$.

Top quark decay		MC samples $\boldsymbol{X}_1$		DATA samples $\boldsymbol{X}_2$		Weights	#Bins
	#jets	n_1	W_1	n_2	W_2	W	k
Electron	2	2011176	59118.87	59121	59121	118239.87	17
	3	719484	11904.55	11905	11905	23809.55	15
	4+	295932	3006.98	3007	3007	6013.98	13
Muon	2	2001522	44736.64	44736	44736	89472.64	17
	3	668072	9098.06	9098	9098	18196.06	15
	4+	277214	2325.02	2325	2325	4650.02	13

Now, we apply the three homogeneity tests K-S, A-D, and χ^2 to the set of all 46 variables from top quark pair decay production at Tevatron introduced in Section 1. The factual sample sizes with corresponding sums of weights W and number of bins k are given in Table 2. Figure 3 provides us with the comparison of weighted χ^2 divergence test of homogeneity with K-S and A-D tests based on wedf's. Notice that divergence tests produce generally slightly higher p-values compared with K-S and A-D tests.

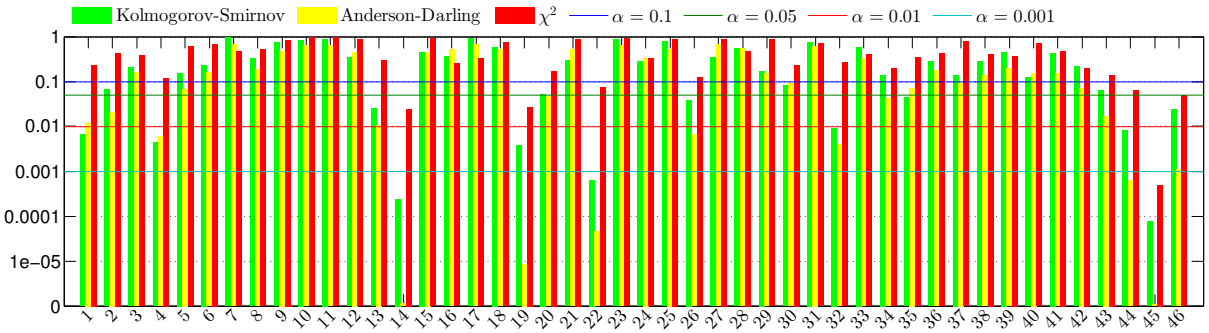


Figure 3. P-values of weighted homogeneity tests K-S, A-D, and χ^2, of MC and DATA distributions for all $m = 46$ variables in Electron 4+ Jets channel. (log-scale)

3. Re-arranging technique for DØ top quark data

The regular asymptotic properties derived for classical unweighted K-S, A-D, and χ^2 tests motivate us to plug into the testing an unweighted data set instead of weighted MC samples. Therefore, we propose a certain transformation of the weighted MC data file into an aggregated unweighted $\mathrm{MC}_\dagger$ data array. We make two requirements for such a re-arrangement. First, we desire preserving or even exploiting the information contained in MC weighting, since the weights linked to certain observations refer the distribution layout in close neighbourhood of these observations. Secondly, we require the sum of weights in MC to correspond to the number of observations in the unweighted re-arranged $\mathrm{MC}_\dagger$.

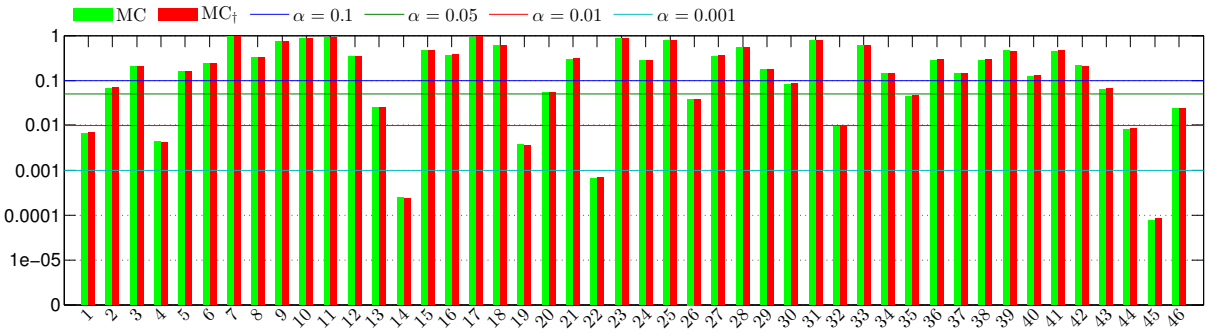


Figure 4. P-values of K-S tests for all $m = 46$ variables in MC and $\mathrm{MC}_\dagger$ data sets of Electron 4+ Jets channel. (log-scale)

For simplicity, let us denote by $\boldsymbol{X} = \left(X_{(1)}, \ldots, X_{(n)}\right)$ the ordered MC sample with the corresponding ordered weights $(w_{(1)}, \ldots, w_{(n)})$ and let $W = \sum_{i=1}^{n} w_i$. Let $\widetilde{n} = \lfloor W \rfloor + 1$ denote the desired minimal number of observations in the new transformed file $\mathrm{MC}_\dagger$, where $\lfloor W \rfloor$ denotes the biggest integer smaller than W. Afterwards, we construct the special weighted averages $\boldsymbol{Y}$ from $\boldsymbol{X}$, presuming that $0 \leq w_{(i)} \leq 1$ for all $i \in 1, \ldots, n$. To define the first weighted average $Y_{(1)}$, we use the smallest possible number of observations $\left(X_{(1)}, \ldots, X_{(k_1)}\right)$ such that $1 \leq \sum_{i=1}^{k_1} w_{(i)} < 2$. Thereby, for all $l < k_1$ it holds that $\sum_{i=1}^{l} w_{(i)} < 1$. The residual portion of weight $w_{(k_1)}$ of the observation $X_{(k_1)}$ equal to $\sum_{i=1}^{k_1} w_{(i)} - 1$ is denoted as r_{k_1}. Thereafter the first $\mathrm{MC}_\dagger$ observation $Y_{(1)}$ can be defined as the following weighted average

$$Y_{(1)} = \frac{\sum_{i=1}^{k_1} w_{(i)} X_{(i)} - r_{k_1} X_{(k_1)}}{\sum_{i=1}^{k_1} w_{(i)} - r_{k_1}} = \sum_{i=1}^{k_1 - 1} w_{(i)} X_{(i)} + (w_{(k_1)} - r_{k_1}) X_{(k_1)}, \tag{2}$$

where we simply substituted for r_{k_1} in (2). The residual portion r_{k_1} of $X_{(k_1)}$ will be included into the next weighted average $Y_{(2)}$. If this re-arranging procedure is applied recursively to the successive observations in $\boldsymbol{X}$, we obtain for all $j = 1, \ldots, \widetilde{n} - 1$,

$$r_{k_j} = \sum_{i=k_{j-1}+1}^{k_j} w_{(i)} - r_{k_j - 1} - 1 \qquad (r_{k_0} \triangleq 0), \tag{3}$$

$$Y_{(j)} = r_{k_{j-1}} X_{(k_{j-1})} + \sum_{i=k_{j-1}+1}^{k_j - 1} w_{(i)} X_{(i)} + (w_{(k_j)} - r_{k_j}) X_{(k_j)}, \tag{4}$$

$$Y_{(\widetilde{n})} = r_{k_{\widetilde{n}}} X_{(n)}, \qquad \text{where } r_{k_{\widetilde{n}}} = \begin{cases} 0, & \text{if } w_{(n)} - r_{k_{\widetilde{n}}-1} < 1/2; \\ w_{(n)} - r_{k_{\widetilde{n}}-1}, & \text{otherwise.} \end{cases} \tag{5}$$

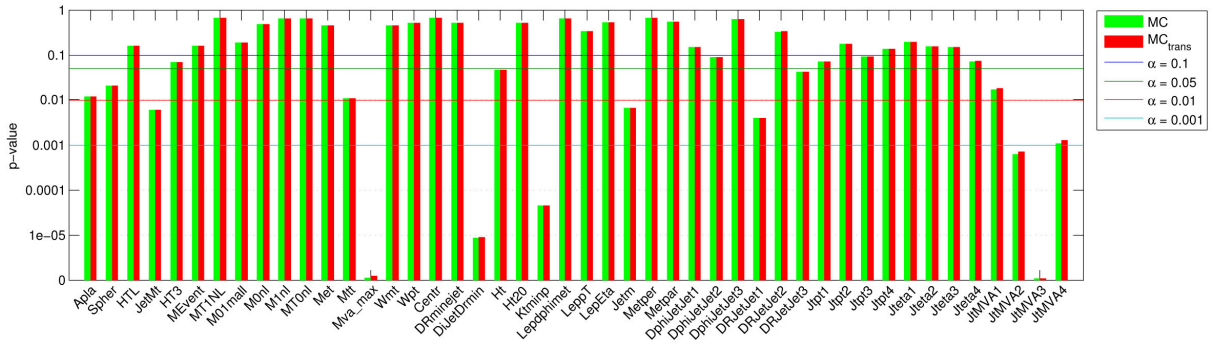


Figure 5. P-values of A-D tests for all $m = 46$ variables in MC and MC$_\dagger$ data sets of Electron 4+ Jets channel. (log-scale)

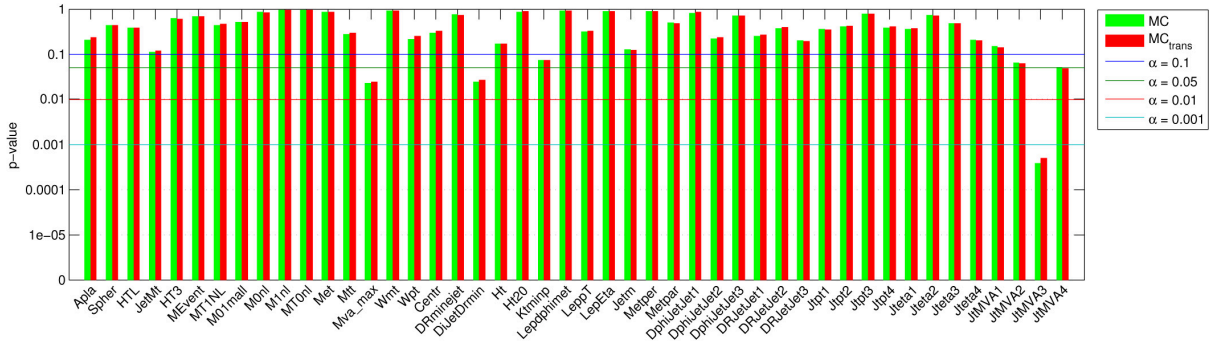


Figure 6. P-values of χ^2 homogeneity tests for all $m = 46$ variables in MC and MC$_\dagger$ data sets of Electron 4+ Jets channel. (log-scale)

Thus we have transformed the original weighted MC array $\boldsymbol{X} = \big(X_{(1)}, \ldots, X_{(n)}\big)$ into the new unweighted MC$_\dagger$ vector $\boldsymbol{Y} = \big(Y_{(1)}, \ldots, Y_{(\widetilde{n})}\big)$ by re-distributing the original weights from MC sample to all the unit weights of vector $\boldsymbol{Y}$. Therefore, we are authorized to apply the standard unweighted theoretical asymptotic properties of K-S, A-D, and χ^2. Indeed, the resulting unweighted p-values (red bars) in Figures 4, 5, 6 from all standard K-S, A-D, χ^2 tests, performed over MC$_\dagger$, remarkably matches with p-values (green bars) obtained by the weighted tests performed over original MC sample. This coincidence holds true even for small magnitudes of p-values.

4. Induced variable selection for DØ top quark data

Finally, we applied all the three homogeneity tests to carry out the variable selection of the complete 46 re-arranged MC and pure DATA samples measured in DØ top quark pair decay channel at Tevatron accelerator. In Table 3 we present the list of variables for which the hypotheses H_0 of homogeneity were rejected at the significance level $\alpha = 0.01$ for at least two out of three applied tests of homogeneity. The results differ for our six decay channels. Notice that 16 different variables were rejected in Muon 3 Jets channel, whilst only 5 different variables were rejected in Muon 4+ Jets decay channel. It means that we are forced to reduce considerably the dimension of Muon 3 Jets data sets to only 30 variables accepted by the homogeneity test procedure. Similar variable selection to only 31 reduced dimensions is induced by the homogeneity rejections in Muon 2 Jets channel. Moreover, the variables Jtpt1-3, Ht, Ht20, HT3, HTL were rejected only in Muon 2 Jets and Muon 3 Jets channels while they work quite well in all Electron channels. Also, Jteta3 and Centr variables are not properly generated only in the case of 3 jets channels (Electron or Muon). On the contrary, the variable DRminejet should be used in a consequent HEP analysis only in 4+ Jets channels (Electron or Muon) and the worst result was achieved for Mva_max variable, which is not allowed to enter any HEP analysis since it was rejected in all top quark channels considered in Table 3.

Table 3. Variables rejected ($\star$) by at least two of three considered K-S, A-D, χ^2 homogeneity tests of MC versus DATA at significance level $\alpha = 0.01$

	Jets	3	4	5	8	10	14	16	17	18	19	20	21	22	23	24	25	27	32	33	35	36	37	41	43	44	45	46
Muon	4+						$\star$																		$\star$	$\star$	$\star$	$\star$
	3	$\star$	$\star$	$\star$			$\star$		$\star$	$\star$		$\star$	$\star$		$\star$				$\star$		$\star$	$\star$	$\star$	$\star$	$\star$		$\star$	
	2	$\star$	$\star$		$\star$		$\star$			$\star$		$\star$	$\star$	$\star$	$\star$		$\star$	$\star$			$\star$	$\star$			$\star$	$\star$		
Elec	4+		$\star$			$\star$	$\star$	$\star$			$\star$			$\star$												$\star$	$\star$	
	3						$\star$		$\star$	$\star$	$\star$									$\star$				$\star$		$\star$	$\star$	
	2						$\star$			$\star$						$\star$	$\star$								$\star$	$\star$		

Acknowledgments

This work was supported by the grants LG15047 (MYES), LM2015068 (MYES), SGS15/214/OHK4/3T/14 (CTU), and GA16-09848S (GACR).

References

[1] Kvita J 2009 *Measurement of Differential Cross-Sections in the ttbar to Lepton+Jets Channel at sqrt(s) = 1.96 TeV with the D0 Experiment at Fermilab* (Prague: Charles University)
[2] Bouř P, Kůs V and Franc J 2016 Statistical classification techniques in high energy physics (SDDT algorithm) *J. Phys.: Conf. Series* **738** 012034
[3] Smirnov N J 1944 Approximate laws of distribution of random variables from empirical data *Usp. Mat. Nauk* **10** 179–206
[4] Pettitt A N 1976 A two-sample Anderson-Darling rank statistic *Biometrika* **1** 161–168
[5] Engmann S and Cousineau D 2011 Comparing distributions: the two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test *Journal of Applied Quantitative Methods* **6** 1–17
[6] Pardo L 2006 *Statistical Inference Based on Divergence Measures* (Chapman & Hall/CRC)
[7] Bouř P 2016 *Statistical nonparametric and divergence methods for data processing in HEP* (Disertation thesis, FNSPE CTU in Prague (in progress))
[8] Mann H B and Wald A 1942 On the Choice of the Number of Class Intervals in the Application of the Chi Square Test *Annals of Mathematical Statistics* **13** 306–317
[9] Kececioglu D B 1993 *Reliability and Life Testing Handbook, 1st ed* (Prentice Hall)