Generalization of Homogeneity Tests Used in HEP Experiments

Jakub Trusina, Jiří Franc, and Adam Novotný

Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering,
Czech Technical University in Prague

jakub.trusina@fjfi.cvut.cz

Goals:

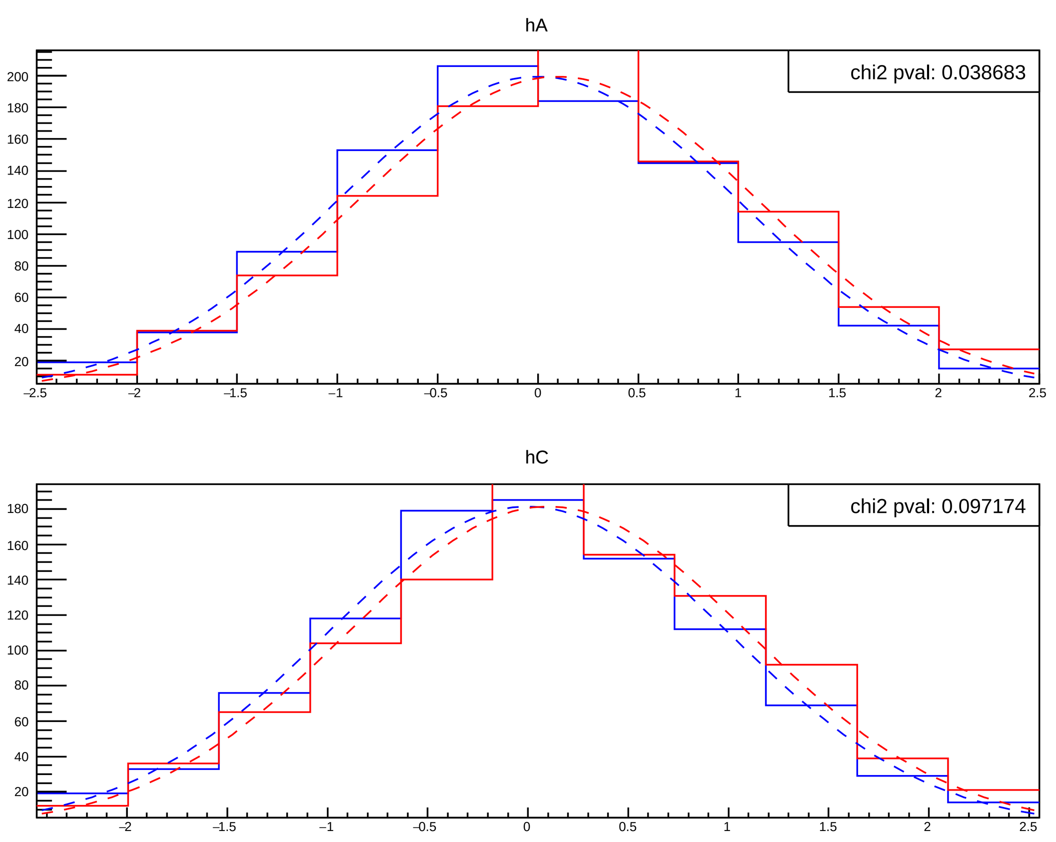
- To develop homogeneity (two sample) tests which can be applied to weighted unbinned data samples in ROOT
- To verify that suggested generalized tests statistics have their presumed distribution
- To compare power of tests for χ^2 test, Kolmogorov-Smirnov test, Anderson-Darling test, and Cramér-von Mises test

I. Introduction

Homogeneity tests currently available in ROOT

- **TH1::Chi2Test**
  - it allows testing weighted data
  - it is unreliable when sample sizes are significantly different
  - it can be applied only to binned data; however, various binning can lead to different test's conclusion
- **TH1::KolmogorovTest** is a modification of the Kolmogorov-Smirnov (KS) test that can be applied to binned weighted data; however, returned p-value is higher than the true one
- **TH1::AndersonDarlingTest** is a modification of the Anderson-Darling (AD) test, it can be applied to binned unweighted data only
- **TMath::KolmogorovTest** is the classical KS test which can be applied only to unweighted and unbinned data
- **ROOT::Math::GoFTest**
  - this class contains implementations of KS and AD test
  - both tests are applicable to unweighted and unbinned samples
  - AD test can be applied also to binned data

Problems with binned data



An example of various binning of the same sample and its effect. Samples were produced from N(0,1) and N(0.1,1)

- **Two different binning configuration**

nbins = 10, min = -2.5, max = 2.5, pval = 0.0387

nbins = 11, min = -2.45, max = 2.55, pval = 0.0972

- **Different results**
  - χ^2 test (or any other test of binned data) can lead to different decision if user adjusts the binning configuration
  - higher number of bins usually leads to higher p-value
- **Advantages of tests based on EDF**
  - while histogram loses information of sample's distribution within each bin, EDF keeps complete information
  - every difference inside bin can be counted (such as Cramér-von Mises (CvM) or AD test)
  - binned KS test does not find the true maximum distance between EDFs but maximum distance between cumulative histograms which is most likely lower

II. Generalized homogeneity tests

We suggest modifications of KS, CvM and AD homogeneity tests statistics. Let $(\boldsymbol{X},\boldsymbol{W}) = ((X_1,...,X_n)',(W_1,...,W_n)')$ be first sample with its weights and $(\boldsymbol{Y},\boldsymbol{V}) = ((Y_1,...,Y_m)',(V_1,...,V_m)')$ the second one. Let $W_{\bullet} = \sum_{i=1}^{n} W_i$ and $K_{\frac{1}{4}}(\cdot)$ be Bessel function of the third kind.

- **Fundamental definitions**

Weighted EDF	Effective sample size	Mixed sample's WEDF
$F_n^{\boldsymbol{W}}(x) = \frac{1}{W_{\bullet}}\sum\limits_{i=1}^{n} W_i \mathbf{1}_{(-\infty,X_i]}(x)$	$n_e = \frac{\left(\sum\limits_{i=1}^{n} W_i\right)^2}{\sum\limits_{i=1}^{n} W_i^2}$	$H_{n_e,m_e}^{\boldsymbol{W},\boldsymbol{V}}(x) = \frac{n_e F_n^{\boldsymbol{W}}(x) + m_e F_m^{\boldsymbol{V}}(x)}{n_e+m_e}$

- **Kolmogorov-Smirnov test**

Test statistic
$$T_{n,m}^{\boldsymbol{W},\boldsymbol{V}} = \sqrt{\frac{n_e m_e}{n_e+m_e}} \sup_{x\in\mathbb{R}} \left| F_n^{\boldsymbol{W}}(x) - F_m^{\boldsymbol{V}}(x) \right|$$

Presumed asymptotic distribution
$$K(\lambda) = 1 - 2\sum_{k=1}^{+\infty} (-1)^{k+1} e^{-2k^2\lambda^2}$$

- **Cramér-von Mises test**

Test statistic
$$T_{n,m}^{\boldsymbol{W},\boldsymbol{V}} = \frac{n_e m_e}{n_e+m_e} \int_{\mathbb{R}} \left(F_n^{\boldsymbol{W}}(x) - F_m^{\boldsymbol{V}}(x) \right)^2 dH_{n_e,m_e}^{\boldsymbol{W},\boldsymbol{V}}$$

Presumed as. dist.
$$L_{\mathrm{CvM}}(z) = \frac{1}{\pi\sqrt{z}} \sum_{k=0}^{+\infty} (-1)^k \binom{-\frac{1}{2}}{k} \sqrt{1+4k} \exp\left(-\frac{(1+4k)^2}{16z}\right) K_{\frac{1}{4}}\left(\frac{(1+4k)^2}{16z}\right)$$

- **Anderson-Darling test**

Test statistic
$$T_{n,m}^{\boldsymbol{W},\boldsymbol{V}} = \frac{n_e m_e}{n_e+m_e} \int\limits_{0<H_{n_e,m_e}^{\boldsymbol{W},\boldsymbol{V}}(x)<1} \frac{\left(F_n^{\boldsymbol{W}}(x) - F_m^{\boldsymbol{V}}(x)\right)^2}{H_{n_e,m_e}^{\boldsymbol{W},\boldsymbol{V}}(x)\left(1-H_{n_e,m_e}^{\boldsymbol{W},\boldsymbol{V}}(x)\right)} dH_{n_e,m_e}^{\boldsymbol{W},\boldsymbol{V}}$$

Presumed asympt. distribution
$$L_{\mathrm{AD}}(z) = \frac{\sqrt{2\pi}}{z} \sum_{k=0}^{+\infty} \binom{-\frac{1}{2}}{k} (1+4k) \exp\left(-\frac{(1+4k)^2\pi^2}{8z}\right) \int_0^{+\infty} \exp\left(\frac{z}{8(w^2+1)} - \frac{(1+4k)^2\pi^2 w^2}{8z}\right) dw$$

III. Numerical verification of presumed distributions

Since no theoretical proof of asymptotic properties has been done yet, we can demonstrate them numerically. If we consider data as random variables, distribution of test statistic is a continuous function, and the null hypothesis is true then

$$\text{p-value} \doteq 1 - F_T\left(T_{n,m}^{\boldsymbol{W},\boldsymbol{V}}\right) \sim \mathrm{U}(0,1).$$

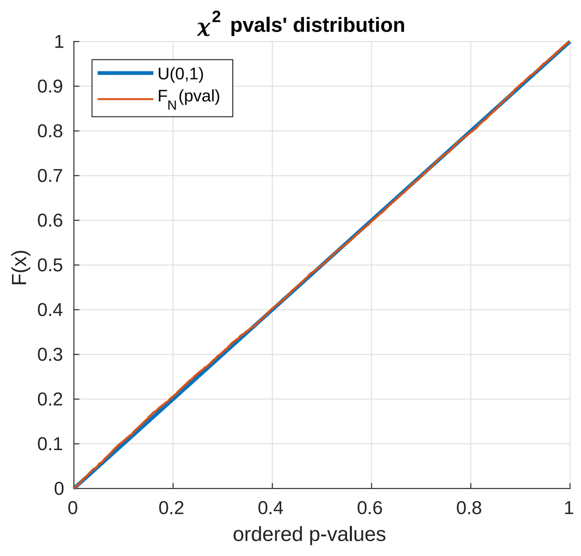
We carried out an experiment in which we produced two samples from different distributions and assigned them weights in such a way that their WEDFs converge to the same distribution. Afterward, we applied homogeneity tests.

Experiment's description

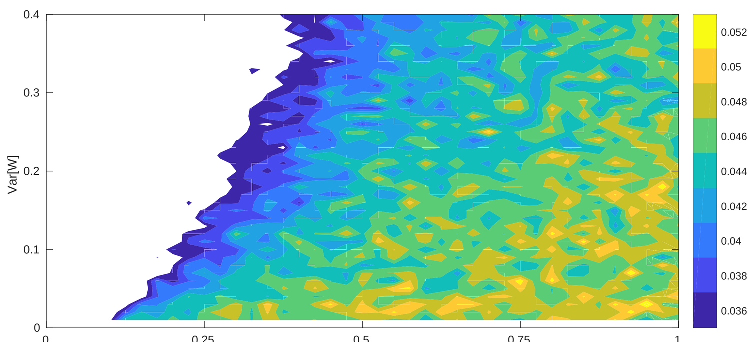
We repeated the whole procedure 10 000 times. Then we plotted EDF of each test's p-values and compared it to CDF of U(0,1).

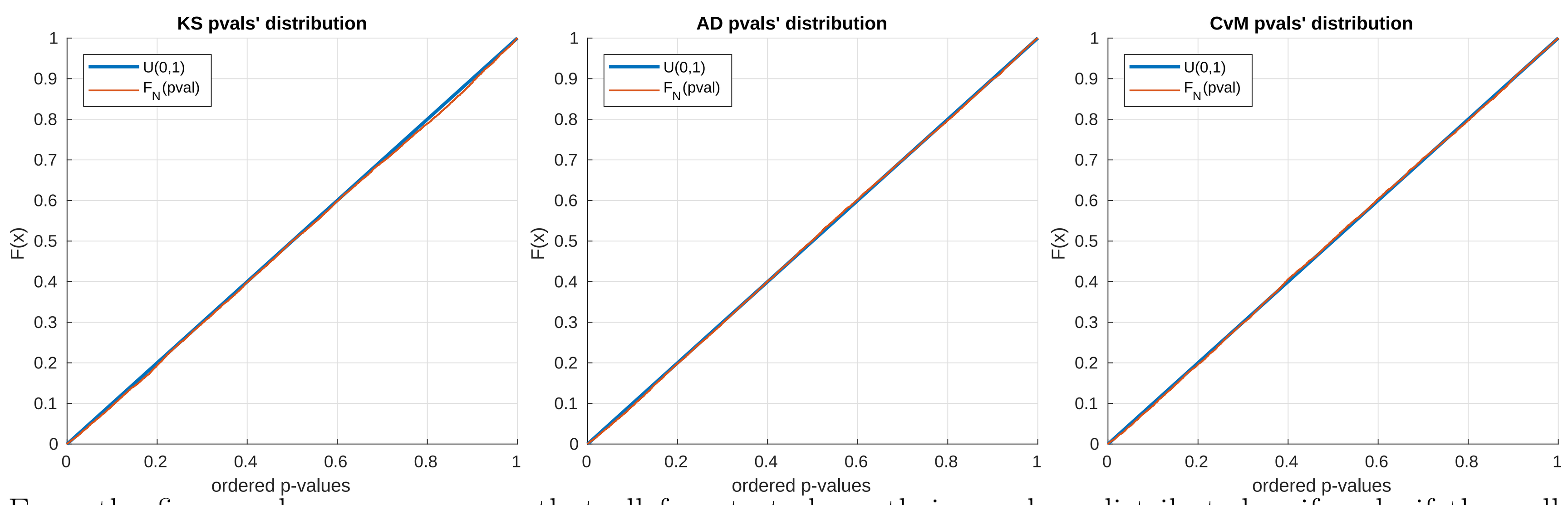
- **First sample**
  - distribution: $X \sim$ N(0,1)
  - sample size: $n = 1000$
  - weights: $W_i = 1$
- **Second sample**
  - distribution: $Y \sim$ N(0.5,1.5[2])
  - sample size: $m = 100\,000$
  - weights: $V_i = \frac{1.5\,\varphi(Y_i)}{100\,\varphi\left(\frac{Y_i-0.5}{1.5}\right)}$ where φ is the standard normal distribution

Results



- Even though p-values of χ^2 test are correct in this experiment, we can show an counterexample. Let now $X, Y \sim$ N(0,1) and $W_i, V_i \sim$ Gamma(k, θ).
- The ratios of rejection (on significance level 0.05) for this another experiment differs significantly for various combinations of k and θ.





From the figures above, we can see that all four tests have their p-values distributed uniformly if the null hypothesis is true. In this case, we consider the null hypothesis not as $F_X = F_Y$ but as $F_n^{\boldsymbol{W}}(x) \to F(x)$ a.s. and $F_m^{\boldsymbol{V}}(x) \to F(x)$ a.s. for every $x \in \mathbb{R}$. For KS, AD and CvM tests we also verified p-values distribution in case of $F_X = F_Y$ and both samples have weights produced independently from some random nonnegative distribution (as in the counterexample).

IV. Power of test comparison

Power of test differs for various experiments' setting. We carried out another experiment in which we observed the effect of six parameters on the power of test.

Experiment's description

We produced two samples from N(0,1) and N(μ_s, $(1+\sigma_s)^2$). All weights of the first sample are equal to 1 while weights of the second sample were independently generated from Gamma(k,θ). Parameters k and θ will be represented by mean (μ_w) and variance (σ_w) of weights. The first sample's size is equal to n while the other sample's is equal to $k{\cdot}n$. For every setting of (μ_s, σ_s, μ_w, μ_w, σ_w, n, k) we repeated procedure 1000 times and calculated ratio of rejected tests (r) on significance level $\alpha = 0.05$ which is power of test's estimate.

Results

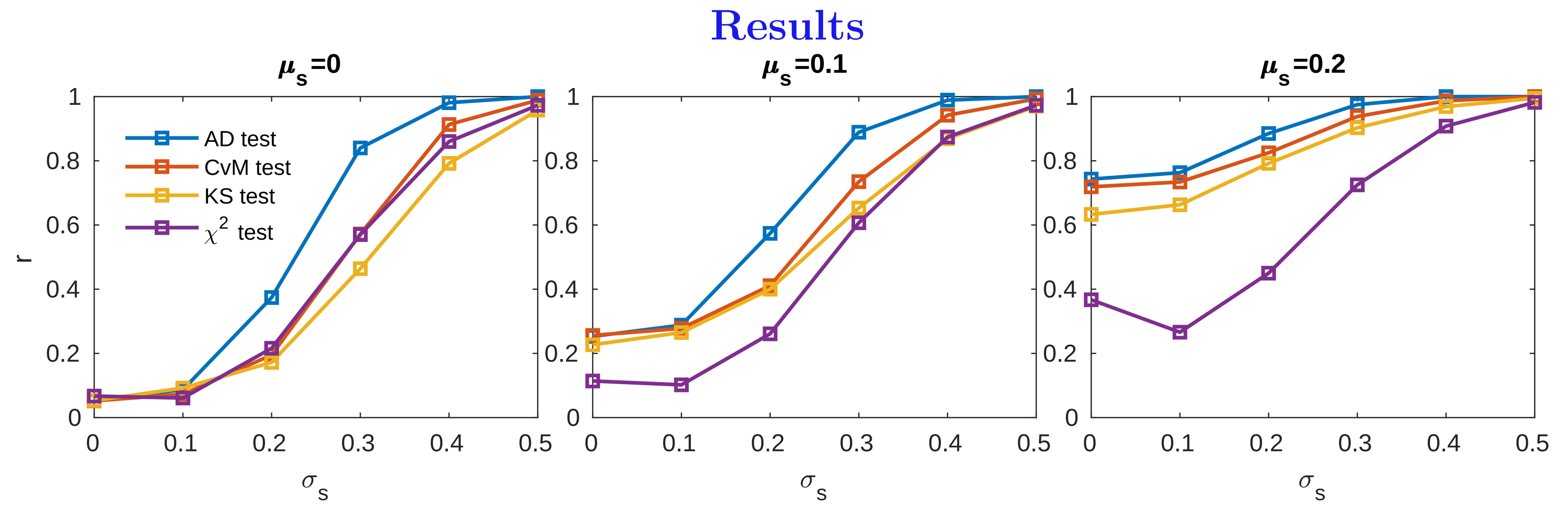


Figure 1: Parameter setting: $(\mu_w, \sigma_w, n, k) = (0.3, 0.1, 200, 10)$. AD test has the highest ratio of rejected tests for both changing parameter μ_s and σ_s. This is also true for $\mu_s = 0.3, 0.4, 0.5$.

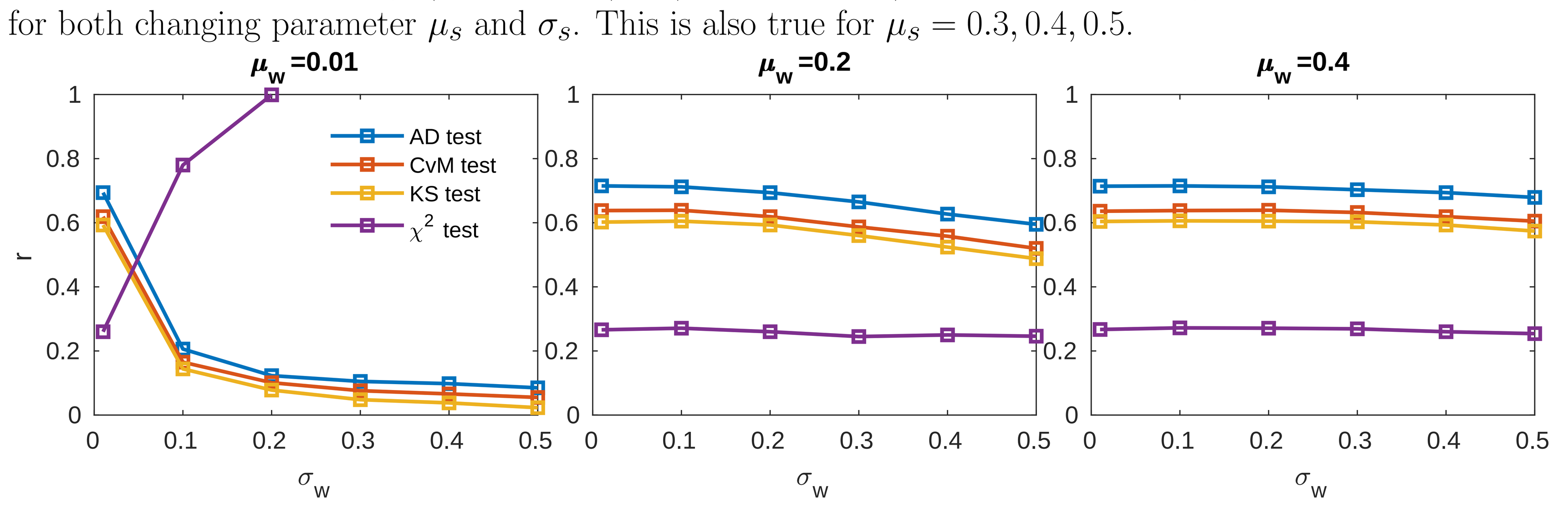


Figure 2: Parameter setting: $(\mu_s, \sigma_s, n, k) = (0.1, 0.1, 500, 20)$. AD test has the highest r for both changing parameter μ_w and σ_w. However, it is interesting discovery that rising σ_w lowers power of test. χ^2 test is unstable for small number of events (when $\mu_w = 0.01$).

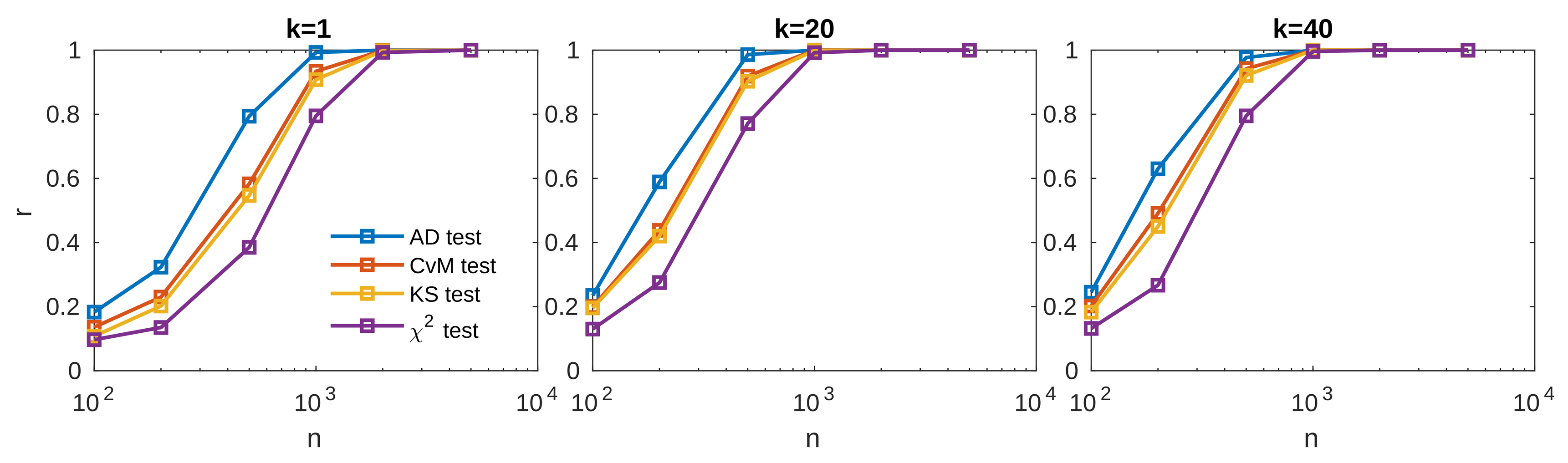


Figure 3: Parameter setting: $(\mu_s, \sigma_s, \mu_w, \sigma_w) = (0.1, 0.2, 0.4, 0.01)$. AD test has again the highest ratio of rejected tests for both changing parameter k and n.

Acknowledgment We acknowledge support from the Czech CTU grant SGS18/188/OHK4/3T/14 and MEYS grants LM2015068 and LTT18001.