A study of Weighted Kolmogorov-Smirnov homogeneity test's properties

Jakub Trusina

FNSPE CTU in Prague
Department of Mathematics

21 June 2018

Homogeneity tests in HEP

- Data vs. MC
- Signal vs. background
- Weighted samples
- Problems of `TH1::chi2test` and `TH1::KolmogorovTest`
- What is the substance of weights?
  - Deterministic constants
  - Random Variables

Unweighted Kolmogorov-Smirnov test

- $(X_i)_{i=1}^n$ $i.i.d. \sim F_X := F$, $\quad (Y_i)_{i=1}^m$ $i.i.d. \sim F_Y := G$
- $H_0 : F = G \quad \text{vs.} \quad H_1 : F \neq G$
- Empirical distribution function: $\quad F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leq x)$

Glivenko-Cantelli theorem

Let $(X_i)_{i=1}^n$ $i.i.d. \sim F$. Then $\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0$.

Test statistic and its distribution

$$T_{nm}(F_n, G_m) = \sqrt{\frac{nm}{n+m}} \sup_{x \in \mathbb{R}} |F_n(x) - G_m(x)|$$

$$\lim_{n,m \to +\infty} P(T_{nm}(F_n, G_m) < \lambda) = K(\lambda) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2\lambda^2}$$

Weighted Kolmogorov-Smirnov test

- $(X_i, W_i)_{i=1}^n$ $i.i.d. \sim (F_X, F_W)$, $(Y_i, V_i)_{i=1}^m$ $i.i.d. \sim (F_Y, F_V) := (G_Y, G_V)$
- Weighted empirical distribution function: $F_n^{\boldsymbol{W}}(x) = \frac{1}{W_\cdot} \sum_{i=1}^n W_i \mathbf{1}(X_i \leq x)$

Generalized Glivenko-Cantelli theorem

Let $(X_i, W_i)_{i=1}^n$ $i.i.d. \sim (F_X, F_W)$ and $(X_1, ..., X_n, W_1, ..., W_n)$ are independent r. v. Let $W_1 \in \mathbb{R}_0^+, \mathrm{E}[W_1] \in \mathbb{R}^+$, and $\mathrm{Var}[W_1] \in \mathbb{R}_0^+$. Then

$$D(F_n^{\boldsymbol{W}}, F_X) = \sup_{x \in \mathbb{R}} |F_n^{\boldsymbol{W}}(x) - F_X(x)| \xrightarrow{\text{a.s.}} 0.$$

Test statistic

$$T_{n_e m_e}(F_n^{\boldsymbol{W}}, G_m^{\boldsymbol{V}}) = \sqrt{\frac{n_e m_e}{n_e + m_e}} D(F_n^{\boldsymbol{W}}, G_m^{\boldsymbol{V}}),\ n_e = \frac{\left(\sum_{i=1}^n W_i\right)^2}{\sum_{i=1}^n W_i^2} \text{ a } m_e = \frac{\left(\sum_{i=1}^n V_i\right)^2}{\sum_{i=1}^n V_i^2}$$

Probability of type-I error

- Type-I error: we reject H_0 even though it is true
- Generally: $P(\text{Type-I error}) \leq \alpha$
- p-value $\doteq 1 - K(T_{n_e m_e}(F_n^{\boldsymbol{W}}, G_m^{\boldsymbol{V}})) \sim \mathrm{U}(0,1)$
- Our case: $P(\text{Type-I error}) = \alpha$
- Ratio of rejected tests r is estimation of $P(\text{Type-I error})$
- If $n_e, m_e \to +\infty$, r is distributed by

$$\frac{\mathrm{Bi}(\alpha n_{\text{tests}}, \alpha(1-\alpha) n_{\text{tests}})}{n_{\text{tests}}}$$

- Using Moivre-Laplace CLT we obtain that asymptotically:

$$r \sim N\left(\alpha, \frac{\alpha(1-\alpha)}{n_{\text{tests}}}\right)$$

Experiment's description

- $H_0 : \widetilde{F} = G \quad \text{vs.} \quad H_1 : \widetilde{F} \neq G!$

1. $(X_i)_{i=1}^n$ $i.i.d. \sim \mathrm{N}(0.2, 1.5^2)$, $\qquad (Y_i)_{i=1}^m$ $i.i.d. \sim \mathrm{N}(0,1)$
   $n = 50m$, kde $m \in \{50, 100, 200, 500, 1000, 2000\}$
2. Computation of weights
3. Application of tests
4. 10 000 repetions of experiment, computation of rejection's ratio at three significance levels $\alpha = \{0.05, 0.01, 0.001\}$

Weights	Using histograms	Using PDF	Using CDF
W_i	$\sum_{j=1}^{k} \mathbf{1}(X_i \in \mathrm{I}_j) \frac{\sum_{l=1}^{m} \mathbf{1}(Y_l \in \mathrm{I}_j)}{\sum_{l=1}^{n} \mathbf{1}(X_l \in \mathrm{I}_j)}$	$\frac{m f_Y(X_i)}{n f_X(X_i)}$	$m(F_Y(X_i) - F_Y(X_{i-1}))$
V_i	1	1	1

Ratio of rejected tests

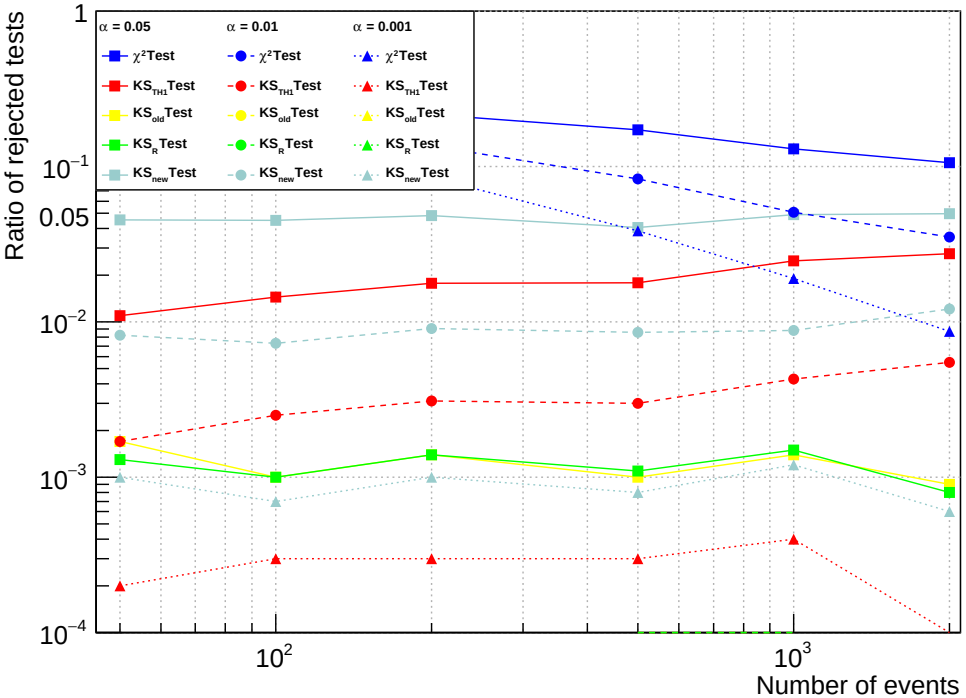

Figure: Weights computed by using PDF

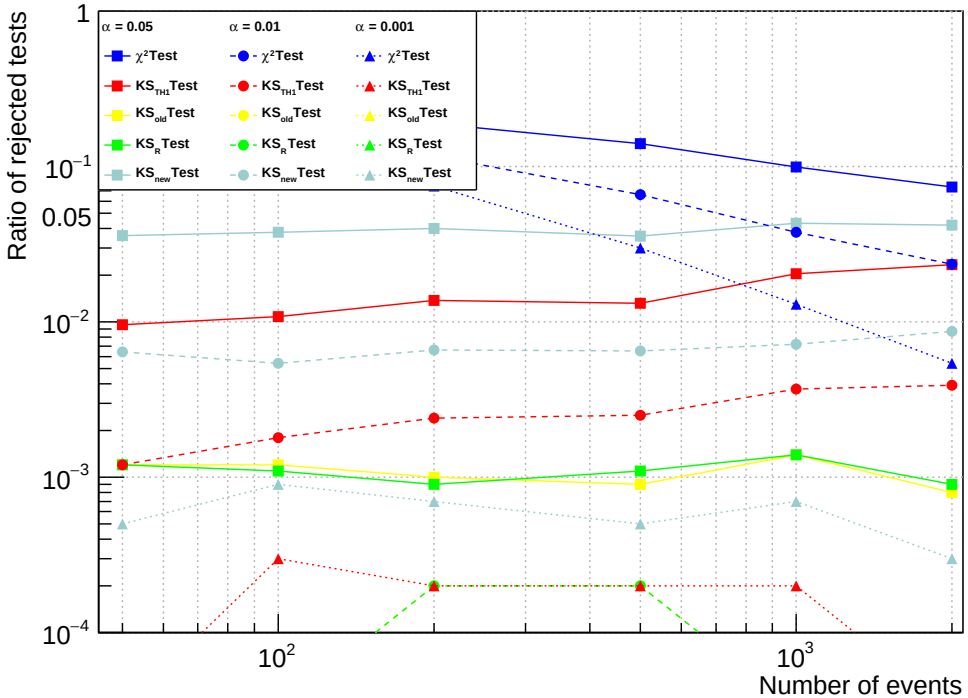

Figure: Weights computed by using CDF

KS_{old} Test a KS_{new} Test

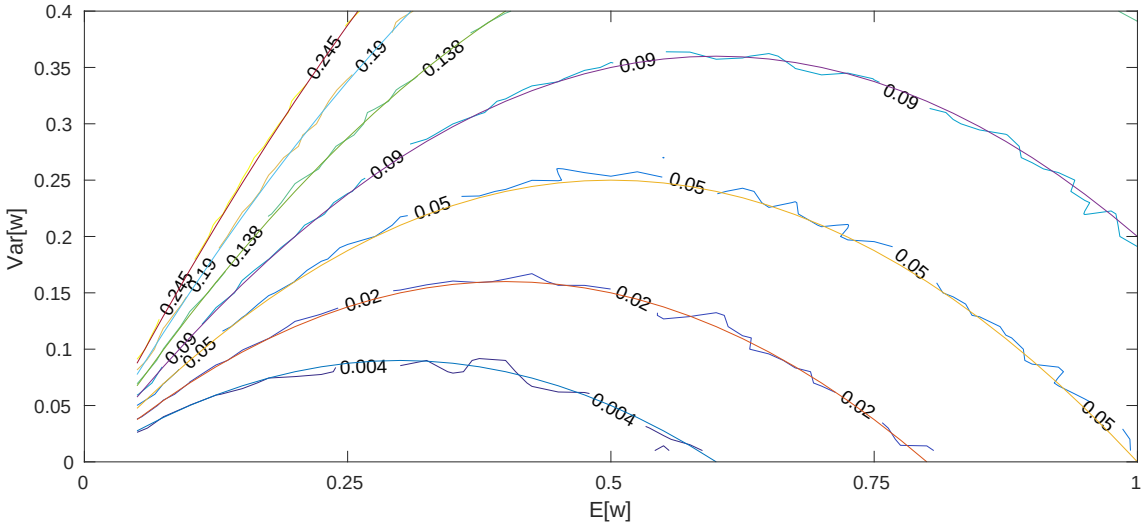

Figure: KS_{old} Test, weights produced from uniform distribution

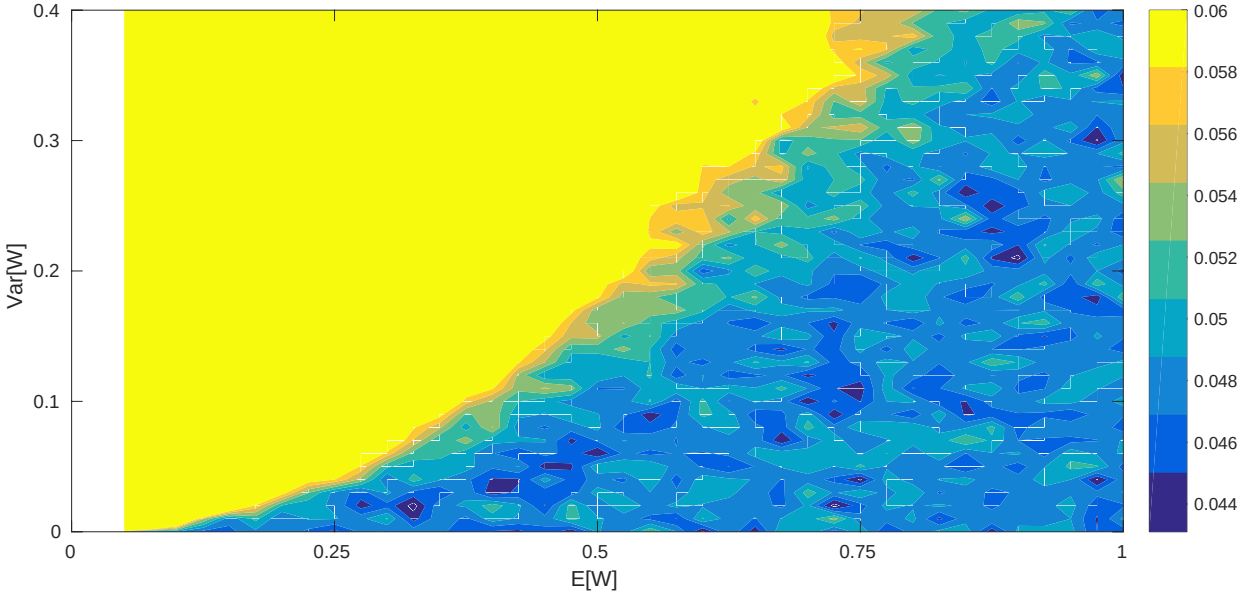

Figure: KS_{new} Test, weights produced from uniform distribution

KS_{new} Test a KS_R Test

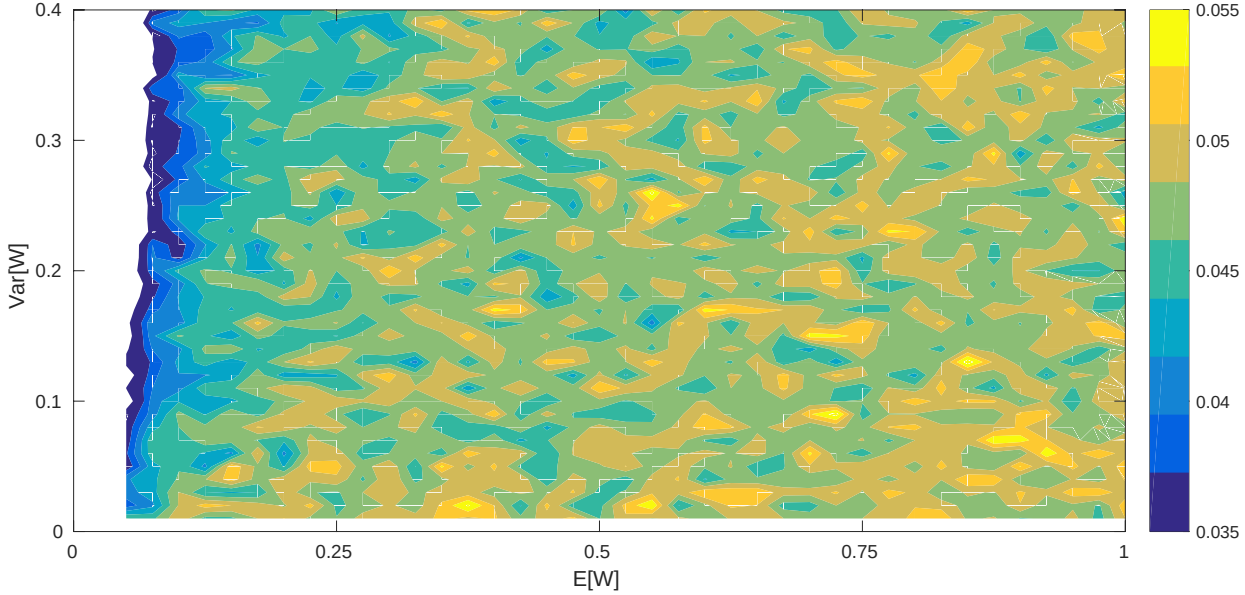

Figure: KS_{new} Test, weights produced from gamma distribution

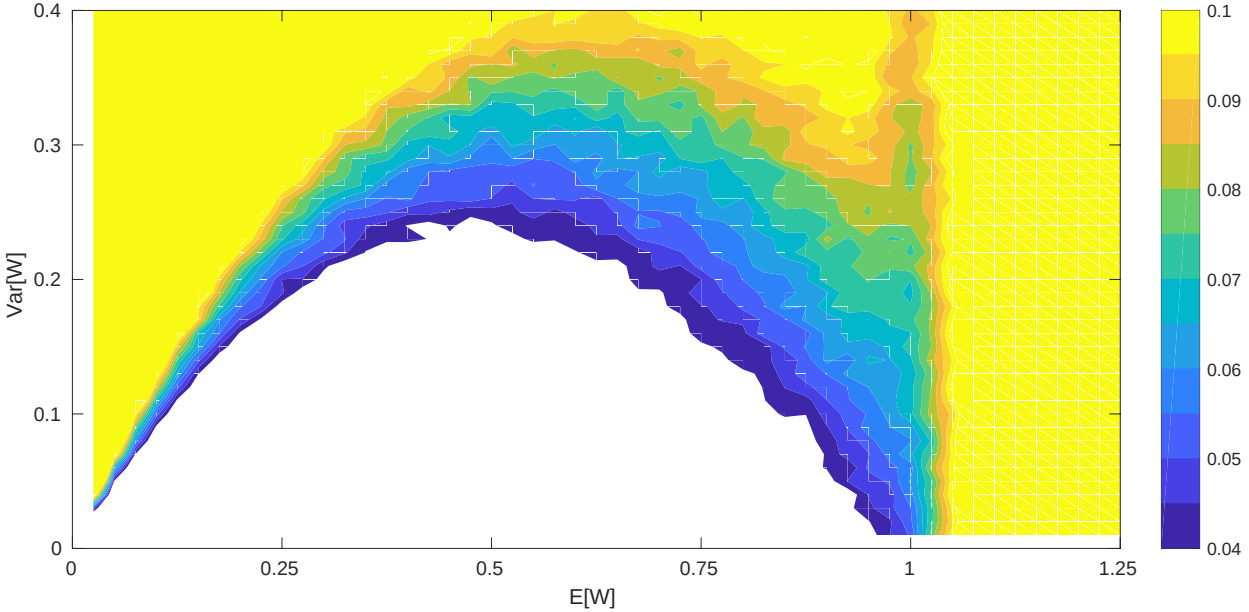

Figure: KS_R Test, weights produced from gamma distribution

KS_{new} Test a KS_{old} Test

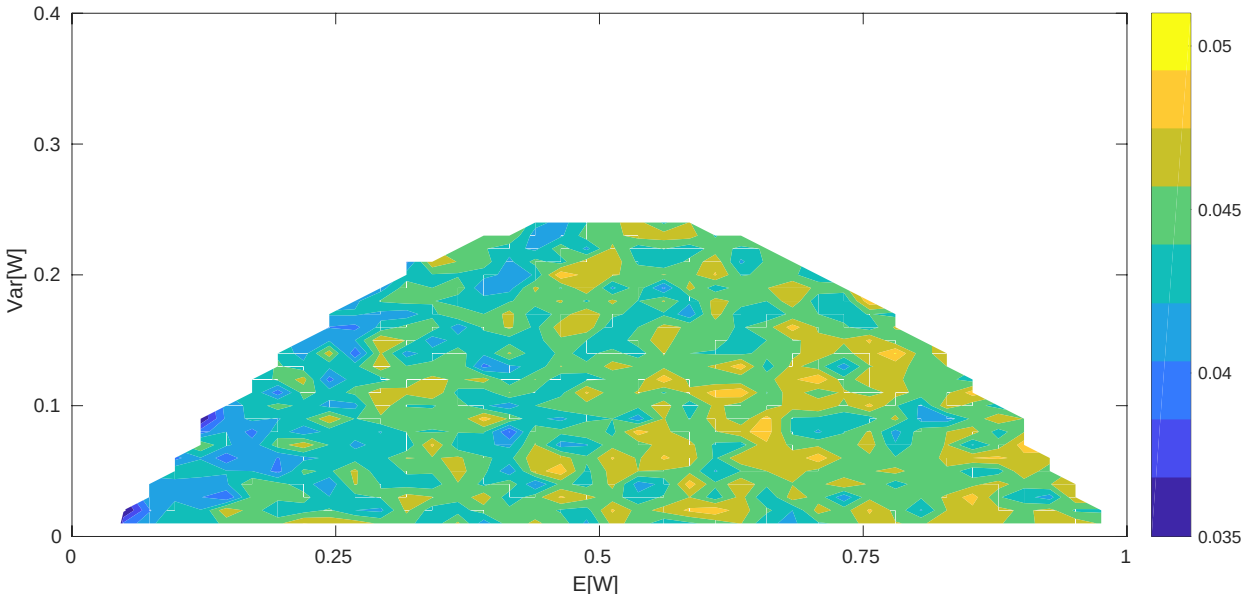

Figure: KS_{new} Test, weights produced from beta distribution

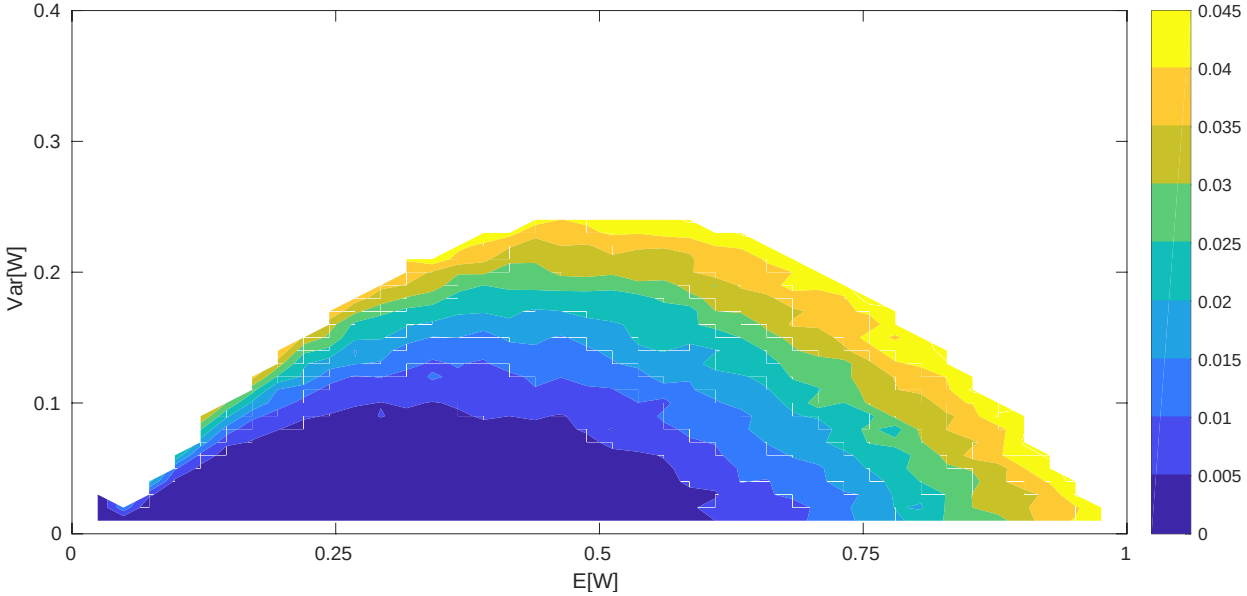

Figure: KS_{old} Test, weights produced from beta distribution

KS_{TH1} Test a χ^2 Test

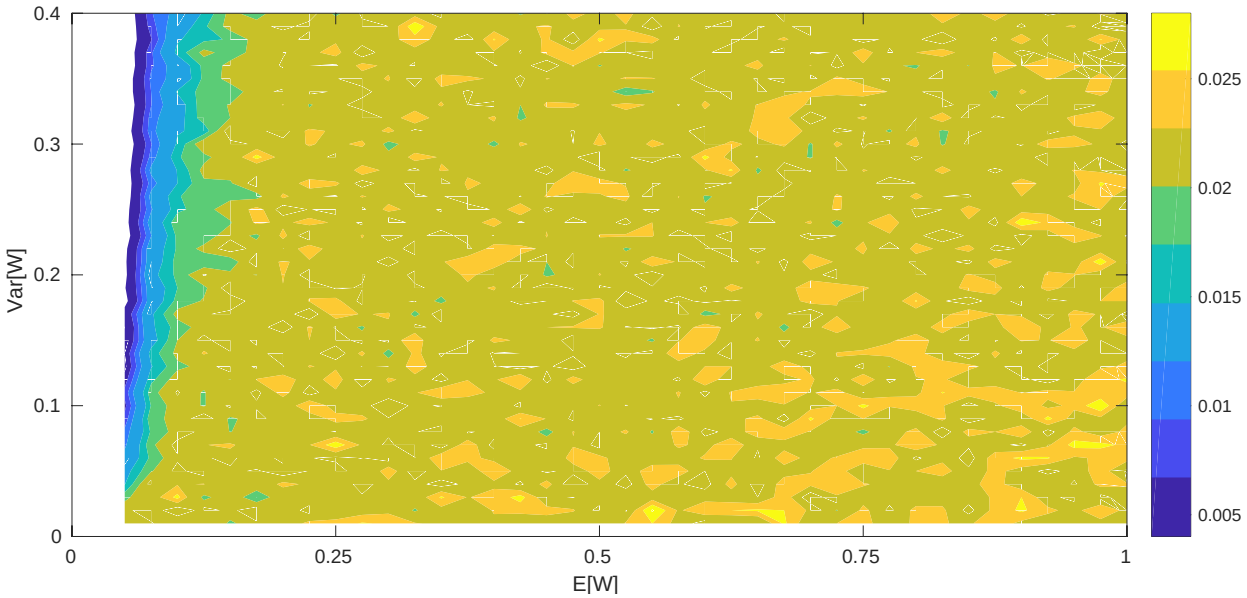

Figure: KS_{new} Test, weights produced from gamma distribution

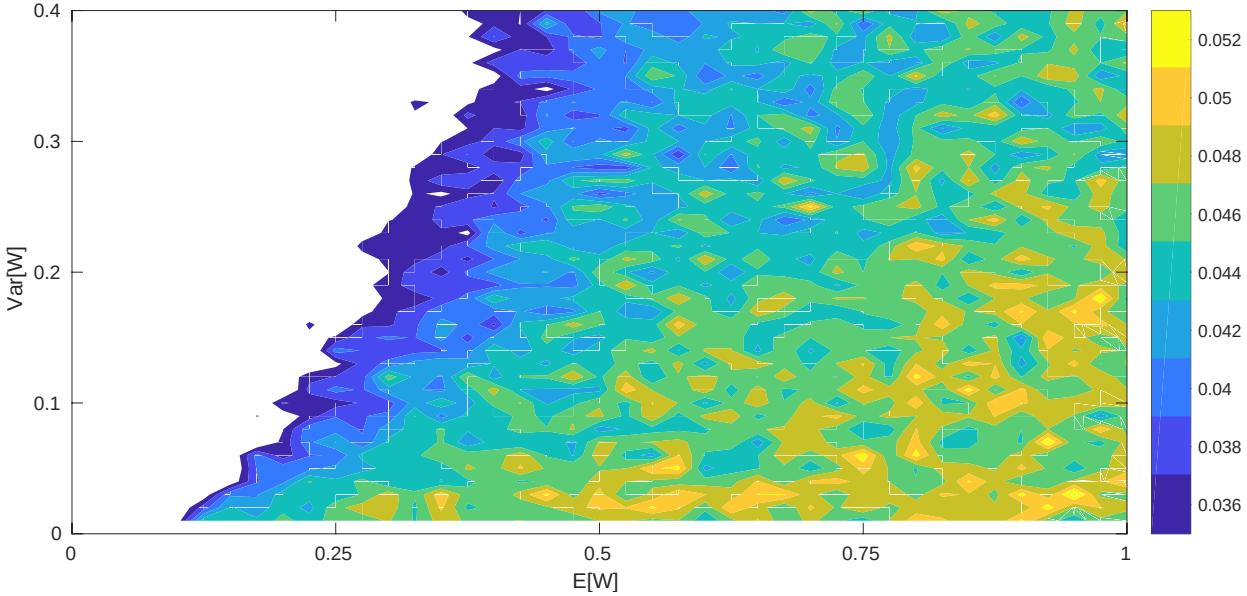

Figure: χ^2 Test, weights produced from gamma distribution

Data vs. MC and signal vs. background

- How does the ratio of rejected tests change as the ratio of sample sizes k increases?

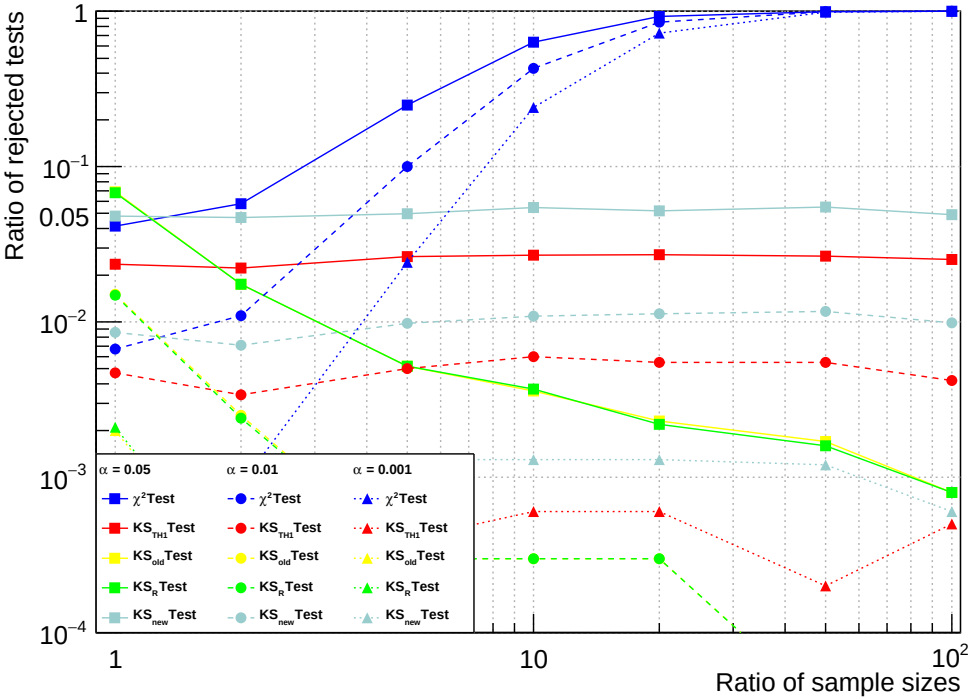

Figure: Data vs. MC

W	V
1	Gamma$(5,5)/k$

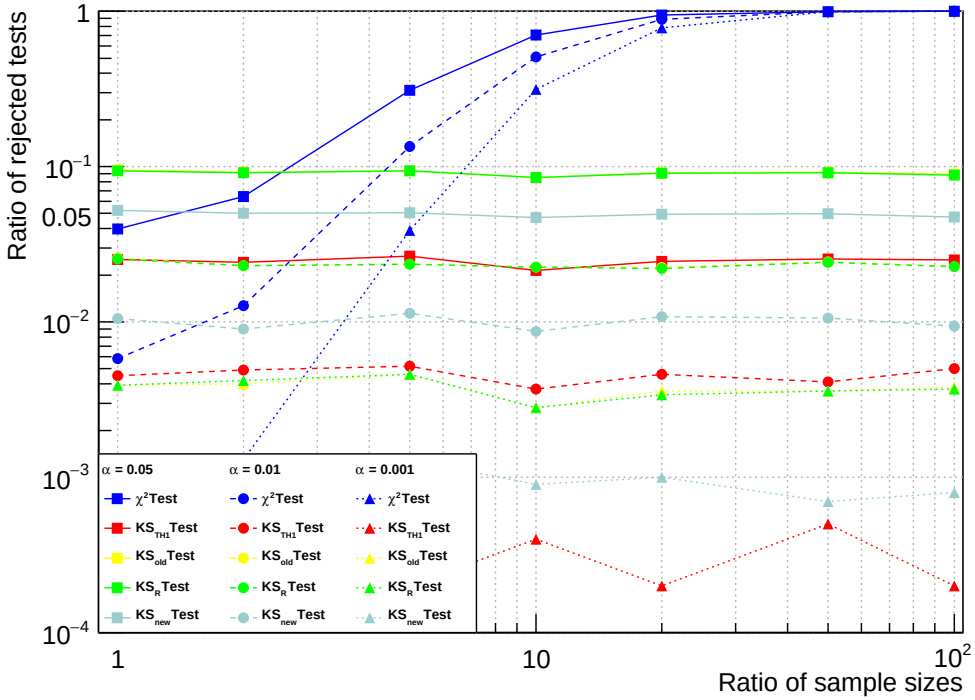

Figure: Signal vs. background

W	V
Gamma$(5,5)$	Gamma$(5,5)$

Summary

- Formulation and proof of generalized Glivenko-Cantelli theorem
  - Is it applicable in practice?
- Three experiments: KS_{new}Test has ratios of rejected tests close to significance level
- For future:
  - Contribute KS_{new}Test to ROOT project
  - Usage of gained knowledge for generalizations of other homogeneity tests
    - Adam Novotný will cover this topic in next presentation
  - Theoretical proof of asymptotic distribution of KS_{new}Test's test statistic