Homogeneity Testing of Weighted Datasets in High Energy Physics

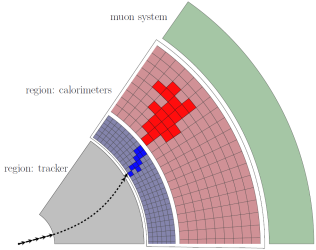
Kristina Jarůšková
FNSPE CTU in Prague

The 12th International Conference SPMS 2021

June 2021

Simulations in HEP

- Simulations of elementary particle interactions - essential tool, representation of theory
- Usage - algorithm training (regression, classification), tuning of data processing steps
- *Detector simulations*
  - Simulations of detector response (electric signal)
  - Followed by reconstruction steps - calculation of different quantities (energy, momentum, ...)
- Standard approach - Monte Carlo-based algorithms
- Agreement between simulations and real data ?



Homogeneity testing of reconstructed quantites

- Two datasets (eg. MC simulations and real measurements)
  - Do they come from the same distribution?
- Unknown parametric family $\longrightarrow$ two-sample nonparametric test of homogeneity
- MC simulations - often weighted samples (sample $x_j \longrightarrow$ weight w_j)
- *Problem:* Standard homogeneity tests are not built to handle weighted samples.

- *In general:* Two i.i.d. weighted datasets:
  - Observations $X_1, \ldots, X_n \sim F$ with weights $W_1, \ldots, W_n \sim F_W$
  - Observations $Y_1, \ldots, Y_m \sim G$ with weights $V_1, \ldots, V_m \sim G_V$

Homogeneity testing - example

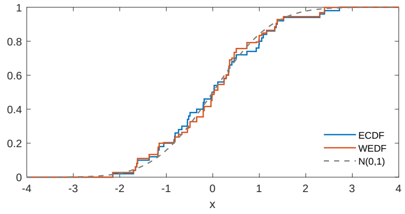
Kolmogorov-Smirnov test

- Kolmogorov distance: $K(F,G) = \sup_{x\in\mathbb{R}} |F(x) - G(x)|$
- Empirical distribution function (EDF): $F_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{I}_{(-\infty,x]}(X_j)$
- Test statistics: $K_{n,m} = \sup_{x\in\mathbb{R}} |F_n(x) - G_m(x)|$
- H_0 rejected $\Leftrightarrow \sqrt{\frac{nm}{n+m}} K_{n,m} \geq h_{1-\alpha}$
  where $H(\lambda) = 1 - 2\sum_{k=1}^{+\infty}(-1)^{k-1}e^{-2k^2\lambda^2}$, $\lambda > 0$

Homogeneity testing of weighted datasets

Possible approaches:

1. Modify test statistic to account for weighted data
   - Empirical distribution function → weighted EDF
     $F_n^W(x) = \frac{1}{W} \sum_{j=1}^{n} W_j \, \mathbf{I}_{(-\infty,x]}(X_j), \quad \forall x \in \mathbb{R}$
   - n number of observations → effective sample size
     $n_e = \frac{\left(\sum_{j=1}^{n} W_j\right)^2}{\sum_{j=1}^{n} W_j^2} \approx n \frac{(\mathrm{E}\, W)^2}{\mathrm{E}\, W^2}$
   - Asymptotic distribution of modified test statistic - unknown



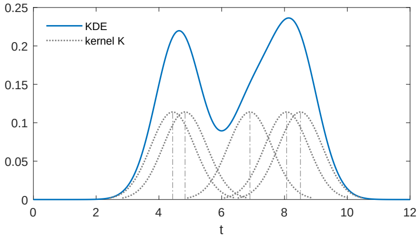
Homogeneity testing of weighted datasets

Possible approaches:

2. Estimate distrib. on weighted data and generate unweighted dataset
   - Weighted kernel density estimates (KDE)
     $\hat{f}(t) = \frac{1}{h\sum_{j=1}^{n} W_j} \sum_{j=1}^{n} W_j \, K\left(\frac{t-X_j}{h}\right), \quad \forall t \in \mathbb{R}$
   - $K : \mathbb{R}_0 \rightarrow \mathbb{R}_0^+$ kernel function

Draw samples from KDE:

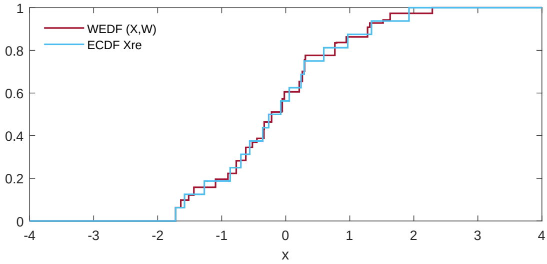
- Randomly select X_I
- Generate $\varepsilon \sim K$
- Unweighted obs.
  $X_I + h\varepsilon$



Homogeneity testing of weighted datasets

Possible approaches:

3. Re-arranging
   - Transformation of weighted data to unweighted
   - Based on weighted averages



Homogeneity testing - numerical simulations

- Verify functioning of modifications - **numerical simulations**
  - H_0: both datasets drawn from the same distribution
  - Get KDE $\rightarrow$ generate unweighted
  - Get AKDE (adaptive KDE) $\rightarrow$ generate unweighted
  - Re-arranging (data transformation)
  - KS statistic modification to weighted data
- Portion (%) of H_0 rejections (estimate of type I error)
  - Good functioning - % of H_0 rejections $\approx$ signif. level α.
- Distribution of p-values, power of test

Functioning of weighted homogeneity testing

- Two weighted datasets, weighted dataset vs. unweighted
- No. of observations $s \in \{500, 1000, 1500, \ldots, 3500\}$
- Verification for selected families of distributions (observations, weights)

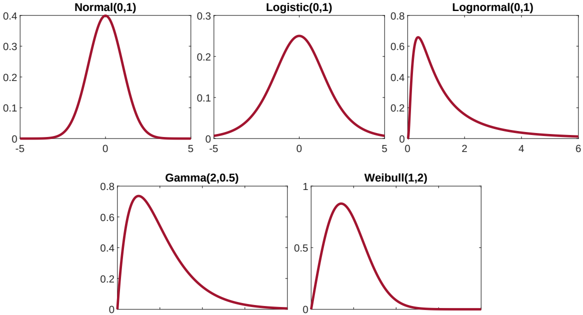


Figure: Distribution of X

Functioning of weighted homogeneity testing

- Two weighted datasets, weighted dataset vs. unweighted
- No. of observations $s \in \{500, 1000, 1500, \ldots, 3500\}$
- Verification for selected families of distributions (observations, weights)

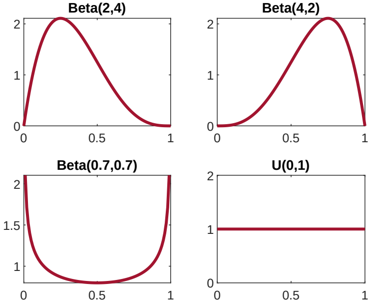
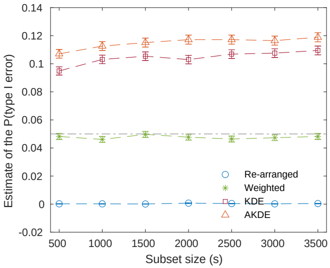


Figure: Distribution of W

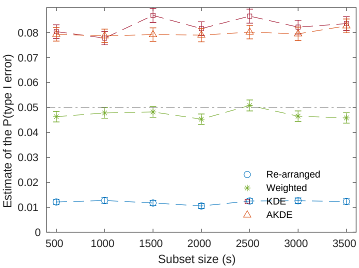
Results of simulations

Estimate of type I. error, observations $N(0,1)$, weights $Beta(2,4)$, $\alpha = 0.05$

- Similar results for other distributions
- KDE-based test - type I error $\gg \alpha$
- Re-arranging - type I error $\ll \alpha$



Weighted vs. weighted

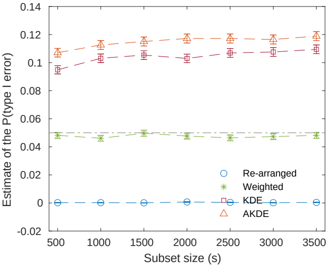


Weighted vs. unweighted

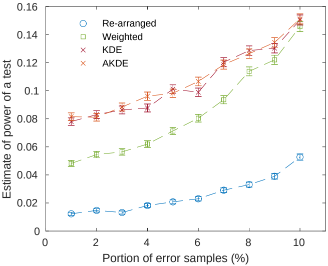
Results of simulations

Estimate of type I. error, observations $N(0,1)$, weights $Beta(2,4)$, $\alpha = 0.05$

- Similar results for other distributions
- KDE-based test - type I error $\gg \alpha$
- Re-arranging - type I error $\ll \alpha \rightarrow$ low power of test



Weighted vs. weighted

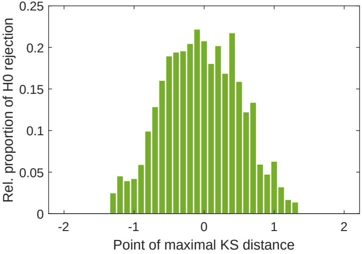


Power of a test

KDE-based tests

Why does KDE-based approach not work properly?

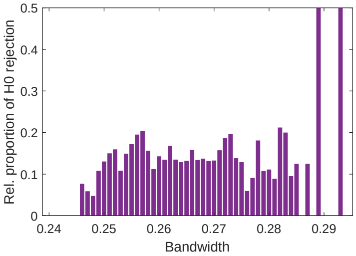
- Problem in tail estimation - false
- Problem in parameter h - false
- Problem in data generation - false
  - Comparison with different method
- Assume knowledge of parametric family $\rightarrow$ estimate parameters $\rightarrow$ generate unweighted data



KDE-based tests

Why does KDE-based approach not work properly?

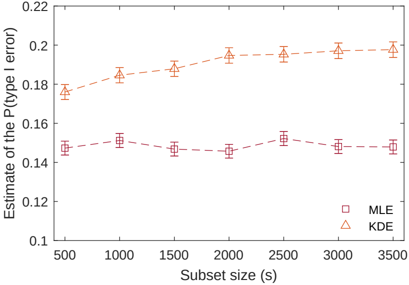
- Problem in tails estimation - false
- Problem in parameter h - false
- Problem in data generation - false
    - Comparison with different method
- Assume knowledge of parametric family $\rightarrow$ estimate parameters $\rightarrow$ generate unweighted data



KDE-based tests

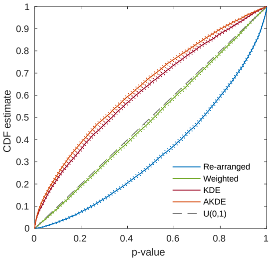
Why does KDE-based approach not work properly?

- Problem in tails estimation - false
- Problem in parameter *h* - false
- Problem in data generation - false
  - Comparison with different method
- Assume knowledge of parametric family $\rightarrow$ estimate parameters $\rightarrow$ generate unweighted data



Summary

- Test with modified statistics
  - Type I error around signif. level α
- Test with re-arranging
  - Type I error below $\alpha \rightarrow$ low power of a test
- Test with KDE/AKDE
  - Accumulation of inaccuracies $\rightarrow$ large type I error
  - Similar results for different distributions $\rightarrow$ determine critical values for H_0 rejection from numerical simulations



Thank you.