Analysis of Statistical Algorithms for the Comparison of Data Distributions in Physics Experiments

Anton Lechner[1], Andreas Pfeiffer[1], Maria Grazia Pia[2] and Alberto Ribon[1]

[1]CERN, Geneva, Switzerland

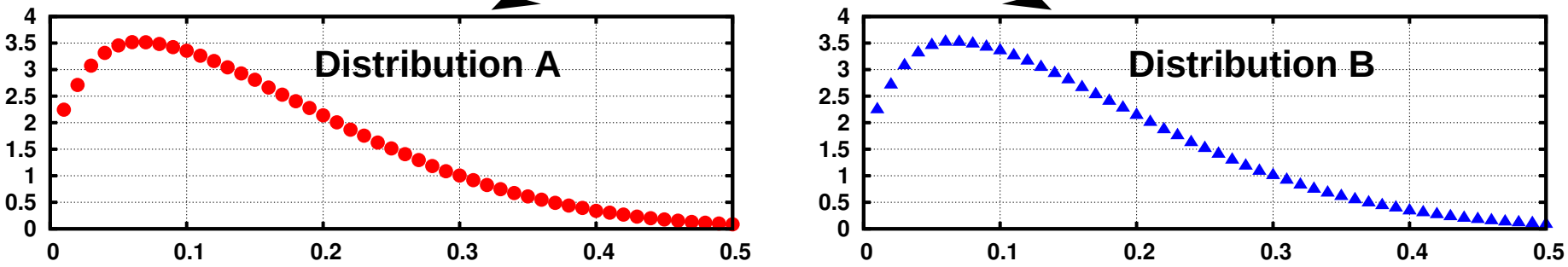
[2]INFN Genova, Genova, Italy

1st November 2007

Statistical comparison of data

Simulation data vs calorimetric results, measured spectra vs theoretical functions,...
⇒ Physics use cases involving the comparison of data sets are diversified

How to compare?



⇒ Goodness of Fit tests can be applied

Goodness of Fit (GoF) tests

Provide a measure for the compatibility of

1. a data sample with a theoretical distribution (**one-sample problem**)
2. two different data samples deriving from the same theoretical distribution (**two-sample problem**)

A variety of GoF-tests exists: χ^2, Anderson Darling, Tiku,...

- But, is a particular test
  - applicable to the considered physics use case?
  - capable of identifying certain characteristics?
- **Relative power of several GoF tests was examined**
  - **Focus on specific issues in physics use cases**
  - **First results available**

Aims of the study

- Evaluating the relative perfomance of tests for certain physics scenarios
  - Fluctuations, outliers, background spectrum,... considered
- Providing guidlines for practical applications
  - Theoretical results presented at NSS 2006, but not close to physics use cases
- Filling partly the gap of information in this domain
  - No extensive hints for physics use cases available yet in literature
  - Novel approach

Power Tests: Tools

Analysis of statistical algorithms performed by employing the "Statistical Toolkit"

HEP Statistics Project: Statistical Toolkit

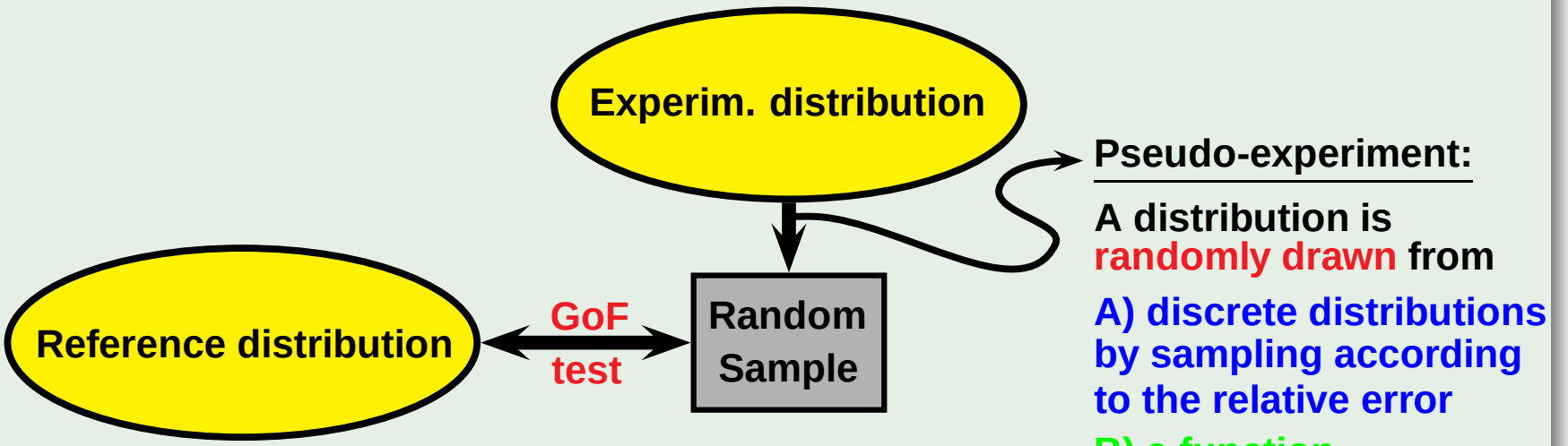
- Open source software toolkit (C++) for statistical data analysis
- Comparison of binned/unbinned data sets possible
- Various Goodness-of-Fit (GoF) tests included
- User layers for AIDA-compliant analysis systems and ROOT
- Reference publication
  - G.A.P. Cirrone *et al*, A Goodness-of-Fit Statistical Toolkit, IEEE TNS, Vol 51, Issue 5, p 1056-63 (2004)
  - B. Mascialino *et al*, New Developments of the Goodness-of-Fit Statistical Toolkit, IEEE TNS, Vol 53, Issue 6, p 3834-41 (2006)

GoF-Tests in the Statistical Toolkit

- Comprehensive collection of GoF tests (see below)
- Hardly any other tool offers a comparable spectrum of tests
- Statistical Toolkit enables an extensive power study

Tests for binned data sets	**Tests for unbinned data sets**
Anderson-Darling (AD)	Anderson-Darling (AD)
Anderson-Darling approximated	Anderson-Darling approximated
χ^2	-
χ^2 *(Incomplete Gamma function)*	-
χ^2 *(Gamma function)*	-
Fisz-Cramer von Mises (CvM)	Fisz-Cramer von Mises (CvM)
-	Girone
-	Goodman
-	Kolmogorov-Smirnov (KS)
Tiku	Tiku
-	Watson
-	Weighted Cramer von Mises
-	Weighted Kolmogorov-Smirnov *(AD or Buning weighting function)*

Power estimation of GoF-tests: Method



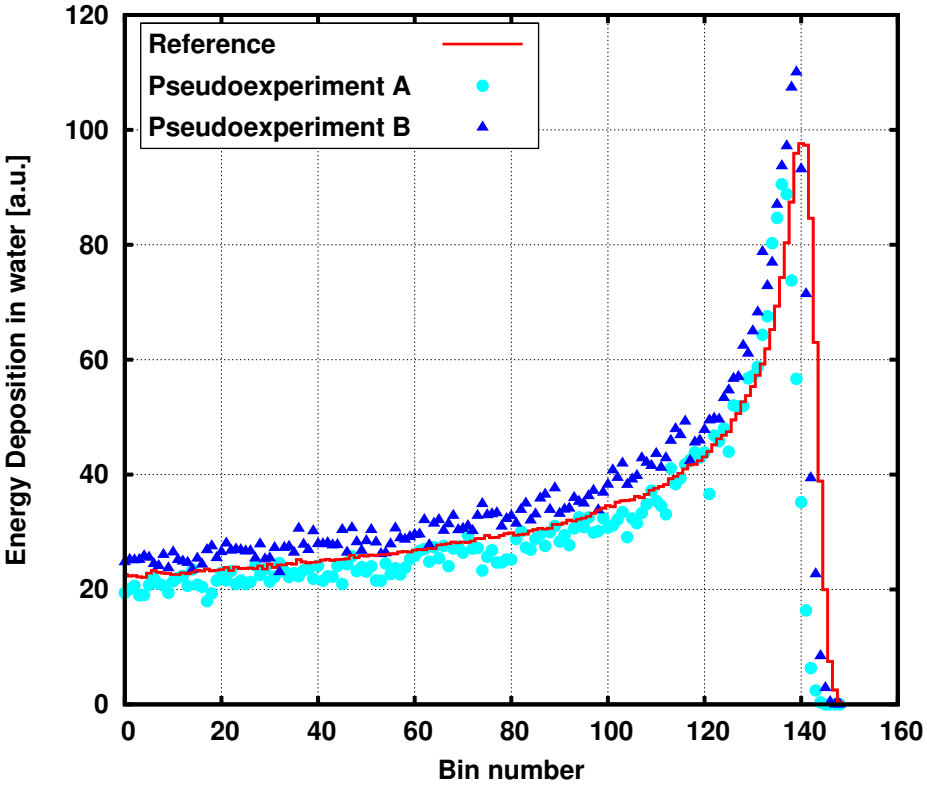
Ensembles of Pseudo-experiments

Large number of pseudo-experiments $\Rightarrow$
GoF-tests characterised by their distribution of p-values

$$\text{Power} = \frac{\#\text{ Pseudo-exp. with p-value} < (1 - \text{CL})}{\#\text{ Pseudo-exp.}}$$

CL = Confidence Level
(here: CL = 0.9)

Proton Bragg peak: Scale-location problem



Proton Bragg peak: Scale-location problem

Are the GoF-Tests sensitive to shifts of the peak position?

Are the GoF-Tests sensitive to variations in the peak height?

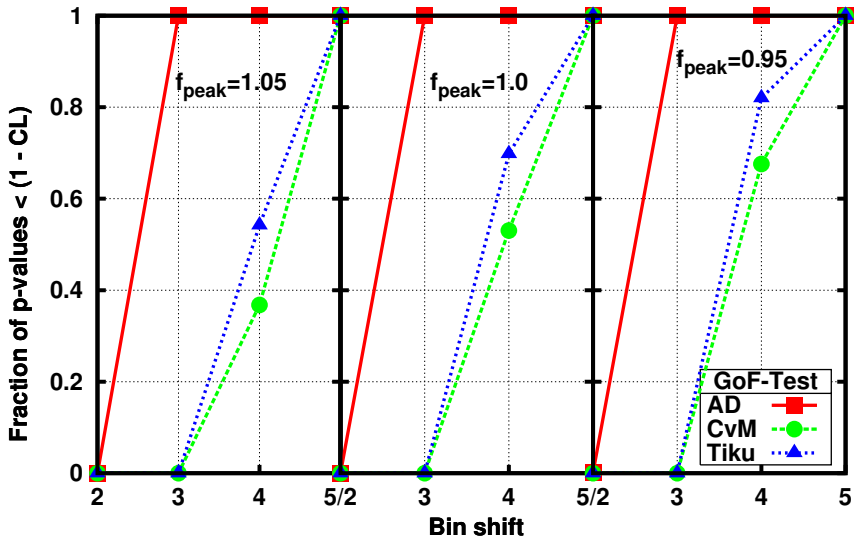
(only binned tests presented)

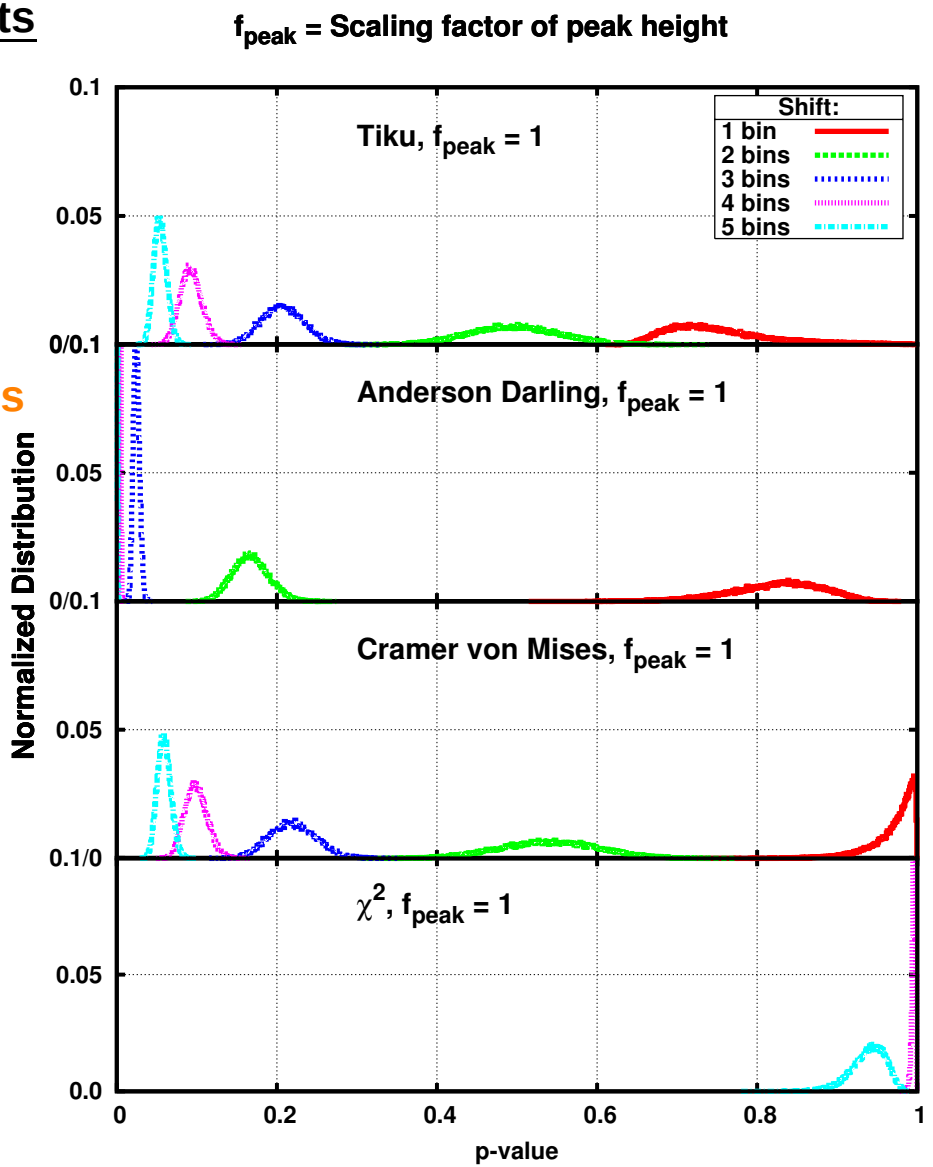
Performance of tests for small relative shifts and deviations in the peak height:

AD, CvM and Tiku: fast rejection of hypothesis that both distributions derive from the same parent distr.

AD shows the most sensitive response

χ^2 reacts much slowlier than the other tests





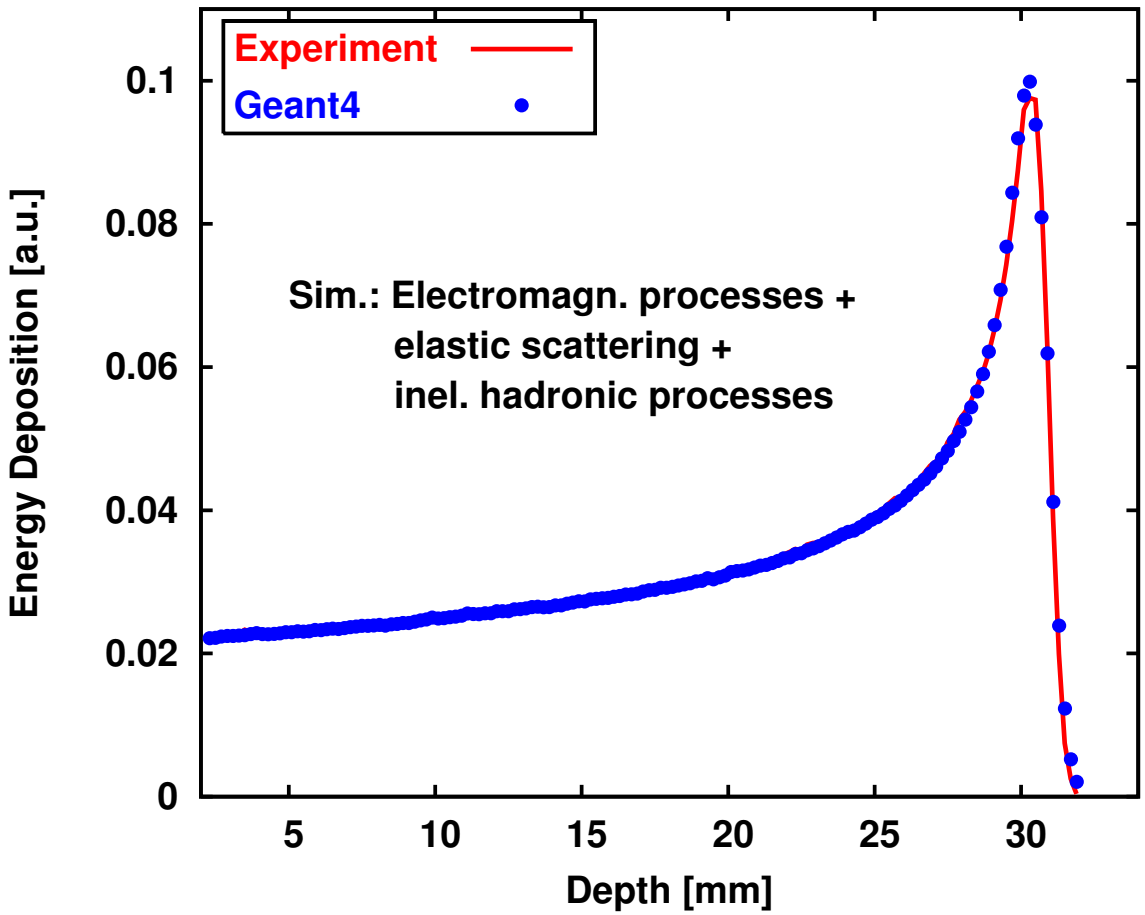
Real physics application

Geant4 validation process

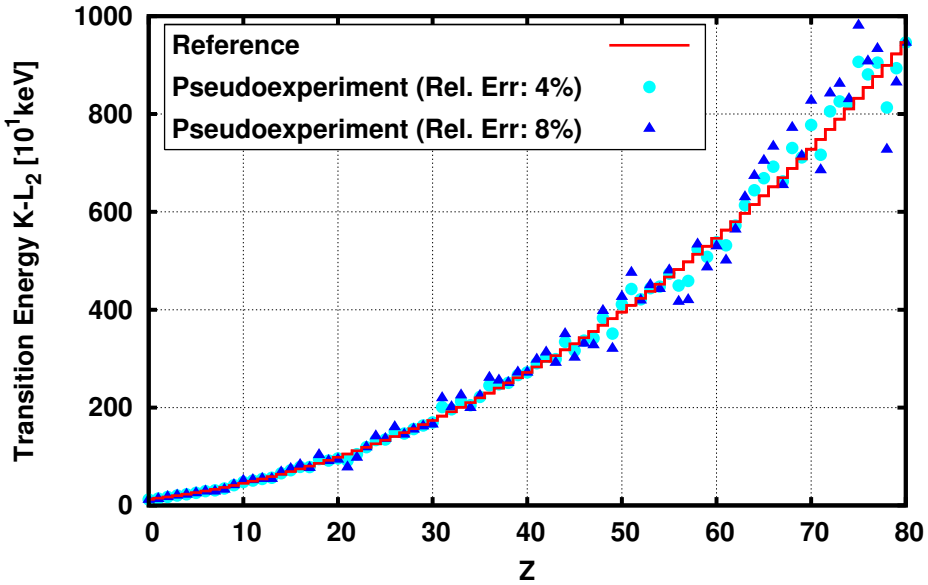
- Simulation of proton energy deposition in water
- Various physics models investigated
- Results compared to experimental data

Performance of physic models evaluated w.r.t. the agreement with the experiment
$\Rightarrow$ **GoF tests applied**

Findings of power study are of great importance how to interpret the results



Fluorescence data: Fluctuations and outliers



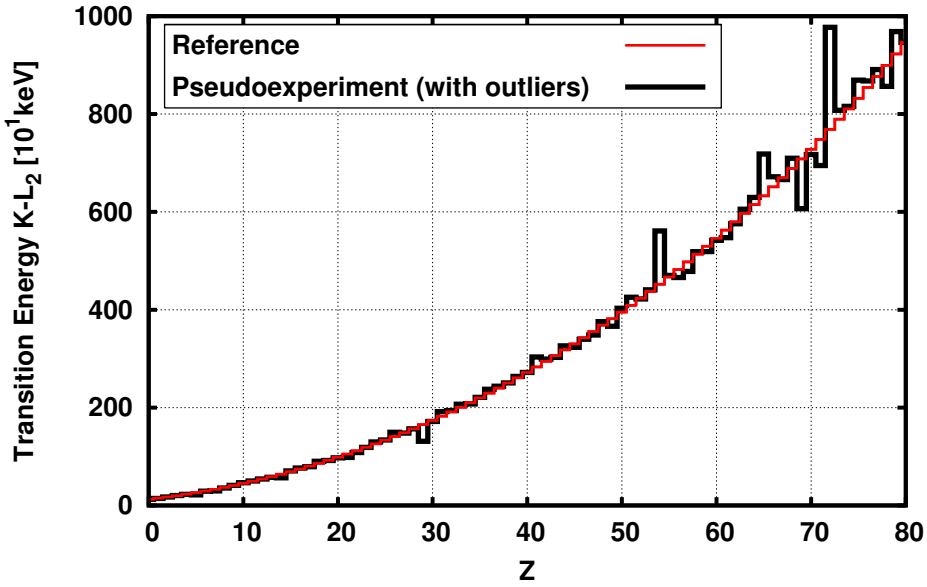
Case 1: Fluctuations

Are the GoF-Tests sensitive to fluctuations?

Are the curves considered as being from the same parent distribution in case of large fluctuations?

Case 2: Outliers

Assuming small fluctuations in the data sample, are outliers recognized?

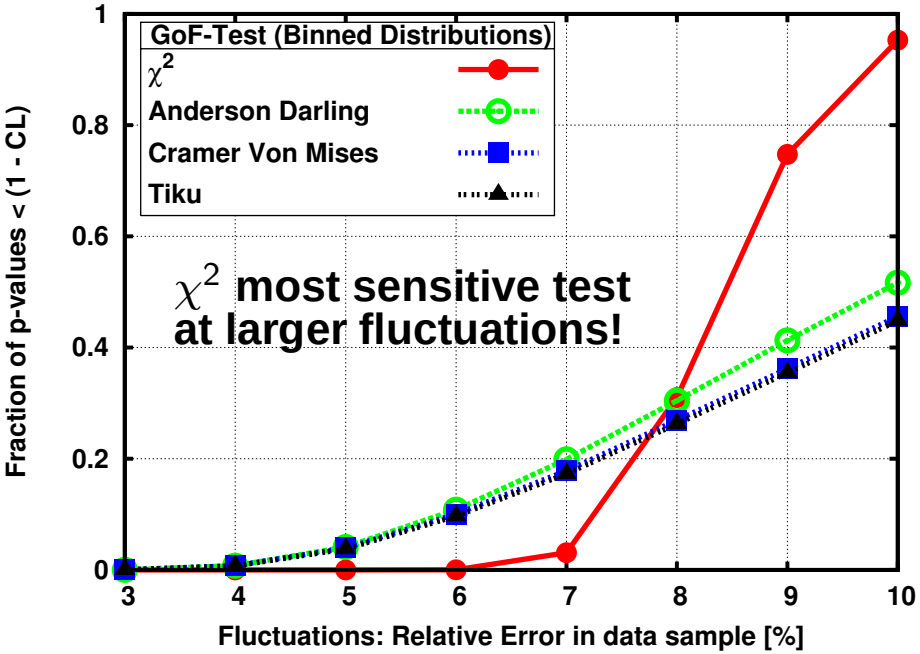


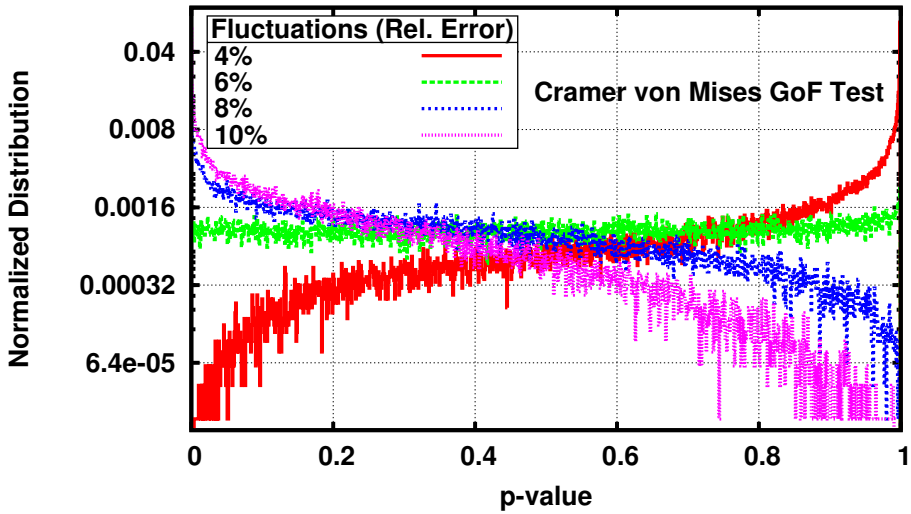
Case 1: Fluctuations

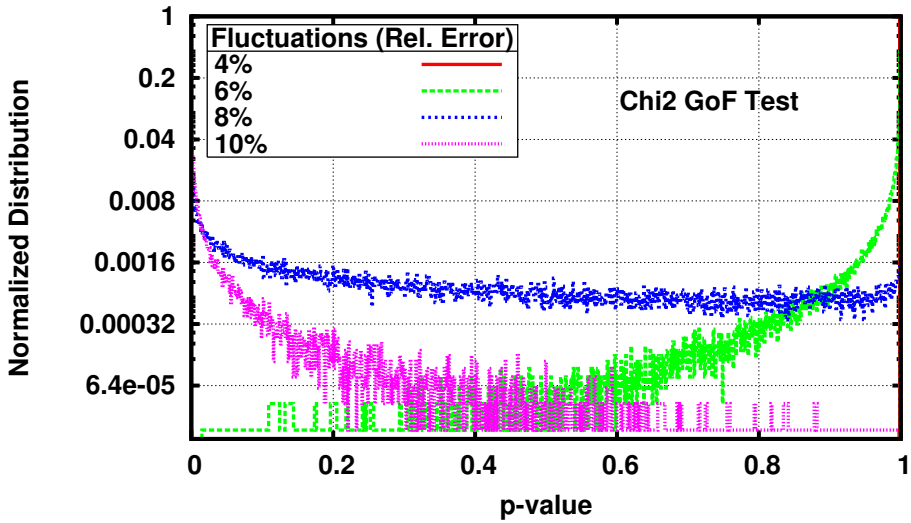
Performance of tests for relative errors from 3% to 10% in data sample:

GoF-Tests for unbinned distributions NOT sensitive to fluctuations: p-values mostly close to 1 (not shown)

Binned comparison: Similar behaviour of AD, CvM and Tiku (see plots)





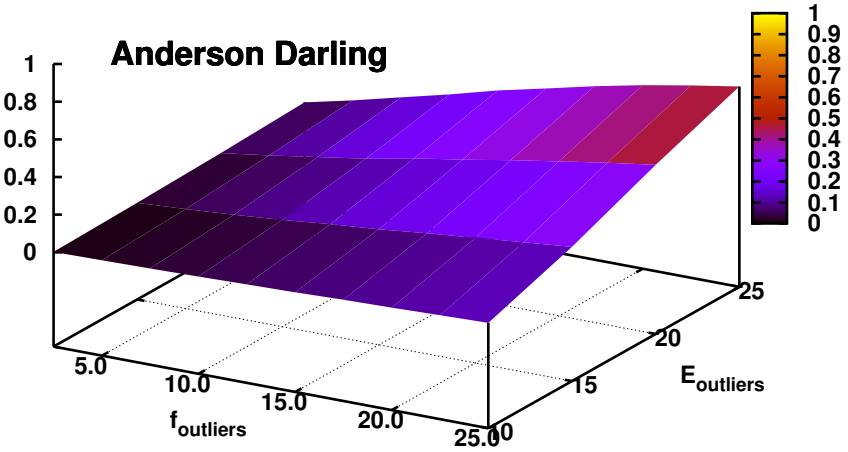


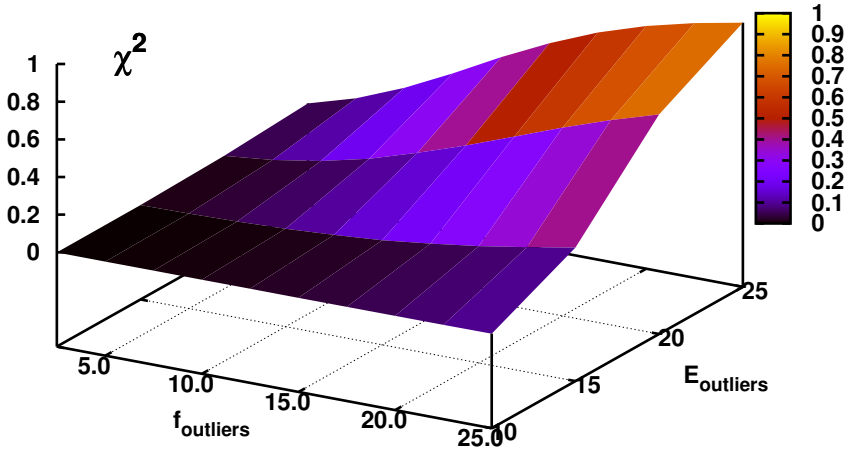
Case 2: Outliers

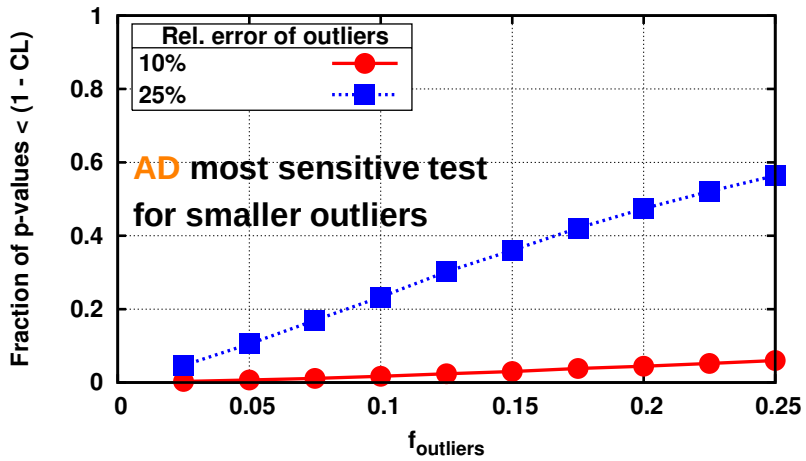
Performance of tests for increasing numbers and errors of outliers:

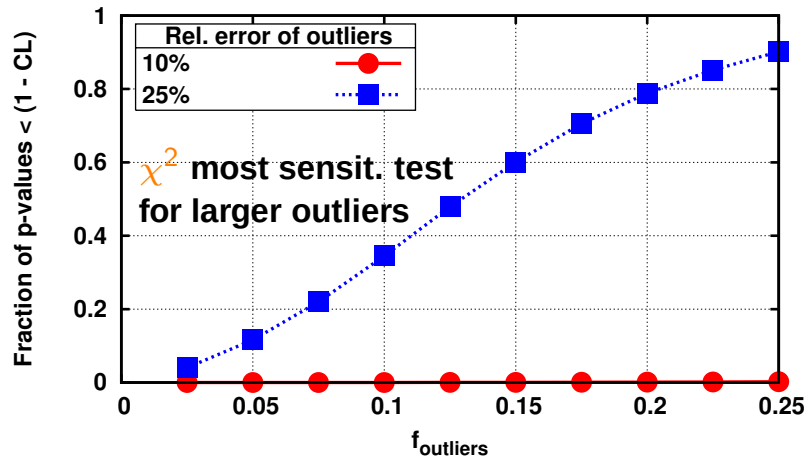
Unbinned comparison: Again no sensitivity

Binned tests: AD, CvM, Tiku similar (only AD shown)





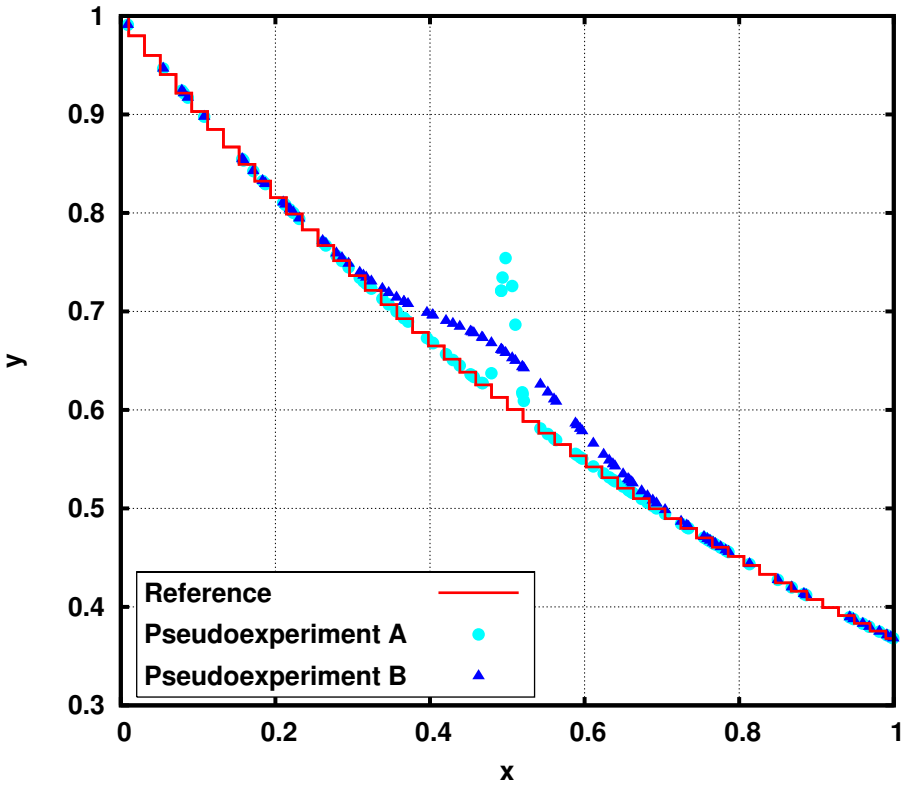




$f_{outliers}$ = Fraction of sample points which are outliers

$E_{outliers}$ = Relative error of outliers

Signal over exponential background

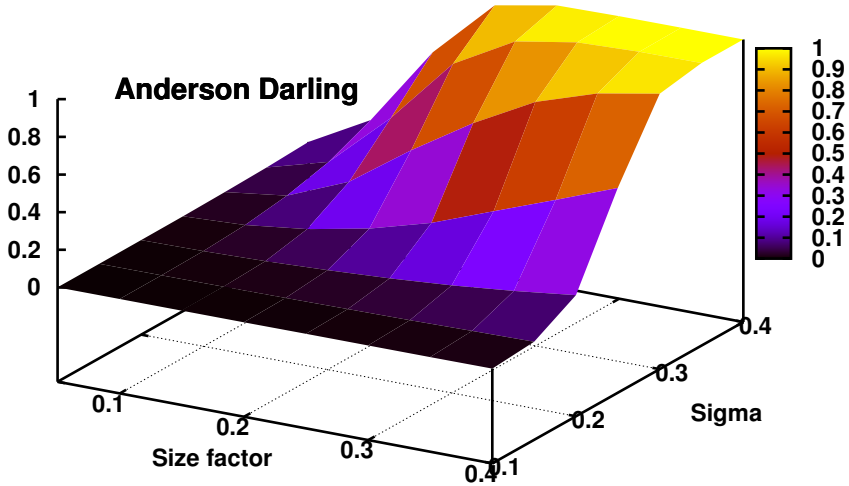


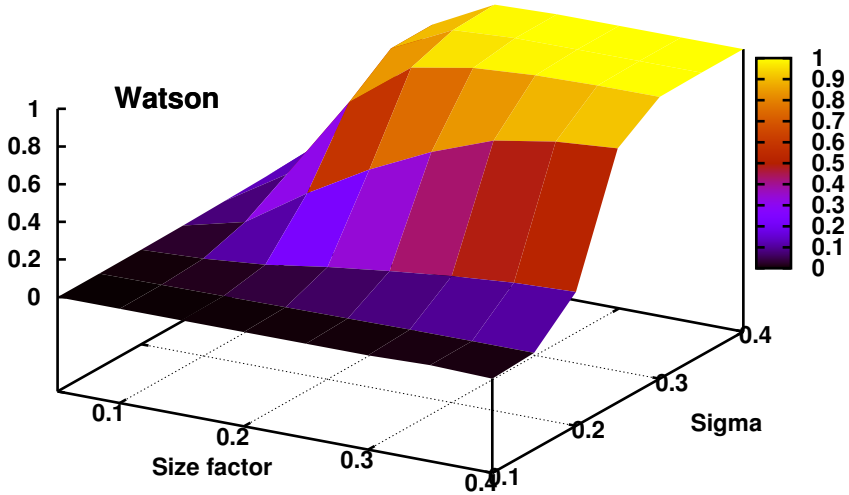
Gaussian signal on top of an exponentially decreasing background spectrum

Are the GoF-tests capable of identifying the differences in the parent distributions?

How is the sensitivity regarding varying sizes and sigmas of the signal?

(only unbinned tests presented)



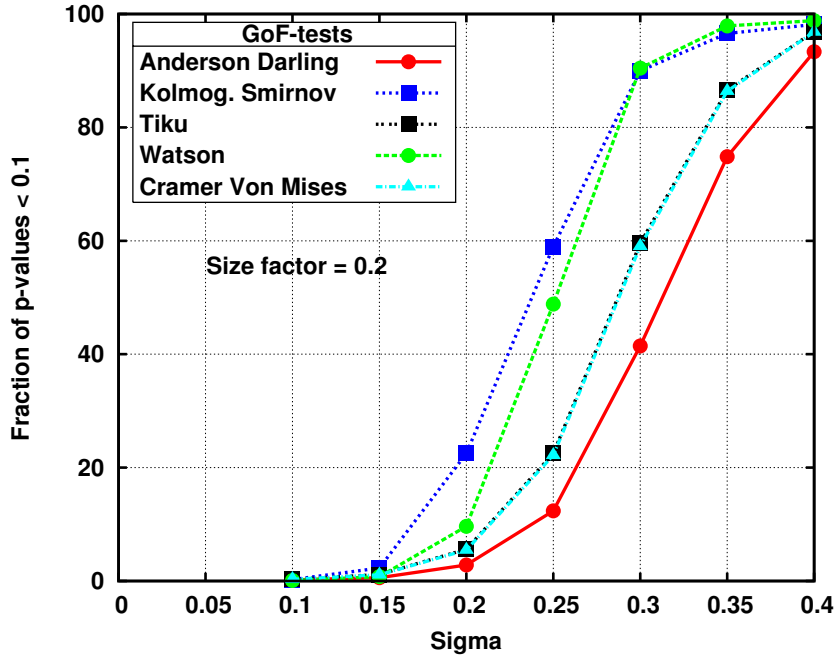


Performance of tests for increasing size and sigma of the Gaussian signal:

All considered tests recognize the signal, but with varying sensitivity

KS and Watson show the most sensitive behaviour
(Watson hardly used in physics analysis!)

AD responses slowlier than all other tests



Summary

- **A power study of GoF tests was performed**
  - Examines a range of GoF tests (Statistical Toolkit)
  - Concentrates on practical aspects
  - Reveals varying sensitivity of GoF tests for different scenarios
  - Gives hints for usage in physics analysis
- **Due to lack of time, only few results shown $\Rightarrow$ Publication with more extensive analysis in preparation**
  - Not much available yet in literature