MAX-PLANCK-INSTITUT
FÜR PHYSIK

Spacing statistics: what they are and how to use them

Lolian Shtembari, Philipp Eller & Allen Caldwell

PHYSTAT Seminar

09.11.2022

Why spacings ?

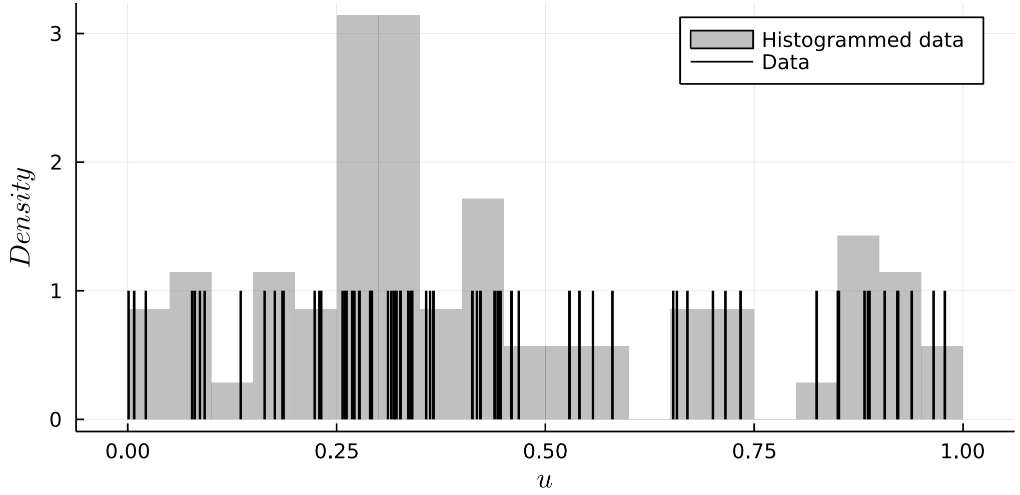
MAX-PLANCK-INSTITUT
FÜR PHYSIK

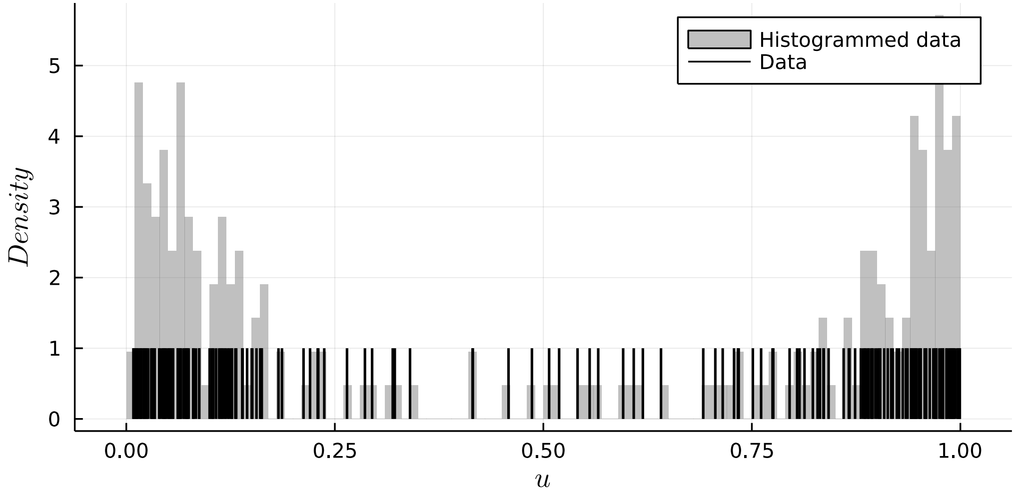
- Do you notice anything strange in these datasets?
- What if you expected a uniform distribution on $[0,1]$?

- Spacings between events correlate with the local event density

- How significant is the cluster?
- Given the previous expectation, can you estimate the event rate?

- *“How well does the model describe the data?”*
    - we need a **Goodness-of-fit test (GOF)**



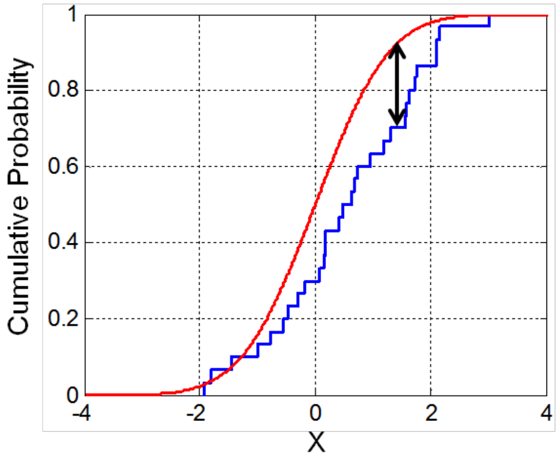


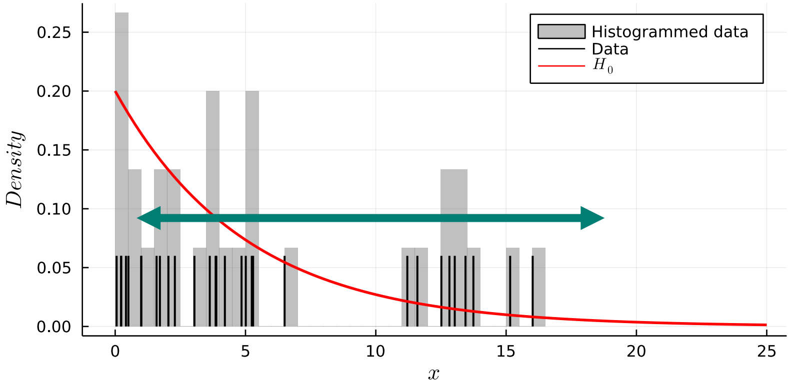
Test statistic

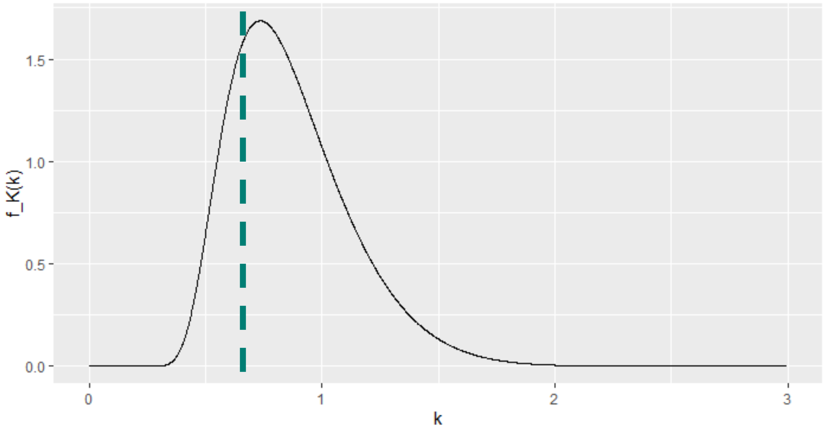
MAX-PLANCK-INSTITUT FÜR PHYSIK

- to perform a Goodness-of-fit test we need a **test statistic** TS
- **condenses the information** available in the data into one value $t = TS(\boldsymbol{x}|\, H_\mathbf{0})$
- each test statistic corresponds to a different question we can ask about the observed data $\boldsymbol{x}$
- we need the distribution $\mathrm{CDF}_T(t|H_0, N)$ in order to assess the **rarity of the observation**

$$\textbf{(p-value)}\ \ p_0 = \Pr(T \leq t \mid H_0)\ \text{ or }\ \Pr(T \geq t \mid H_0)$$



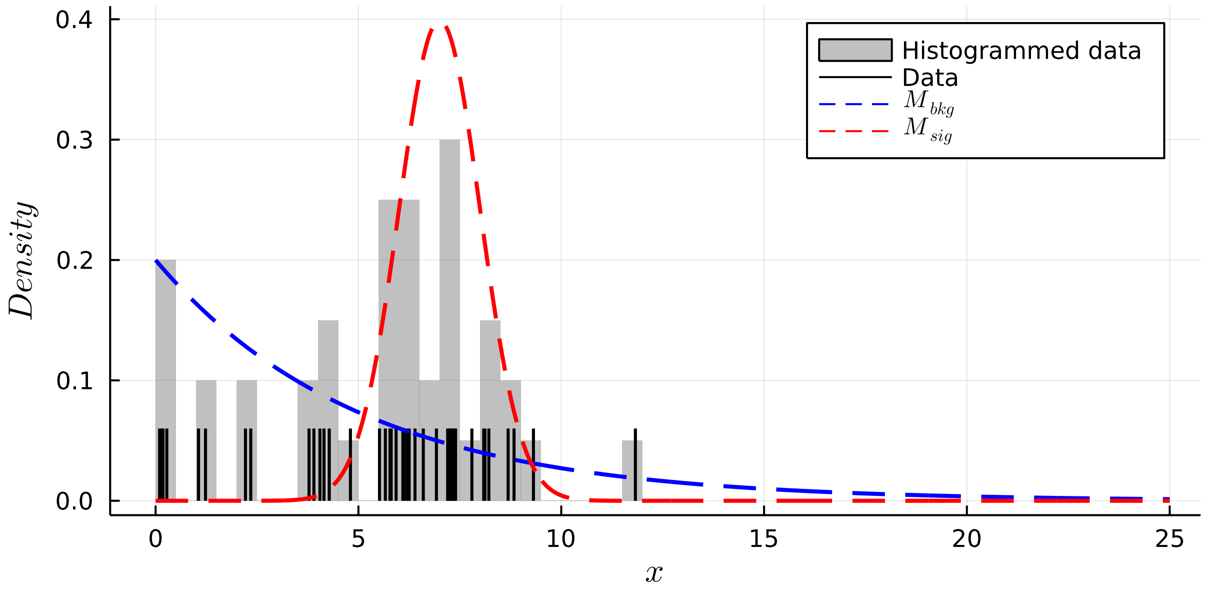




Sources of data

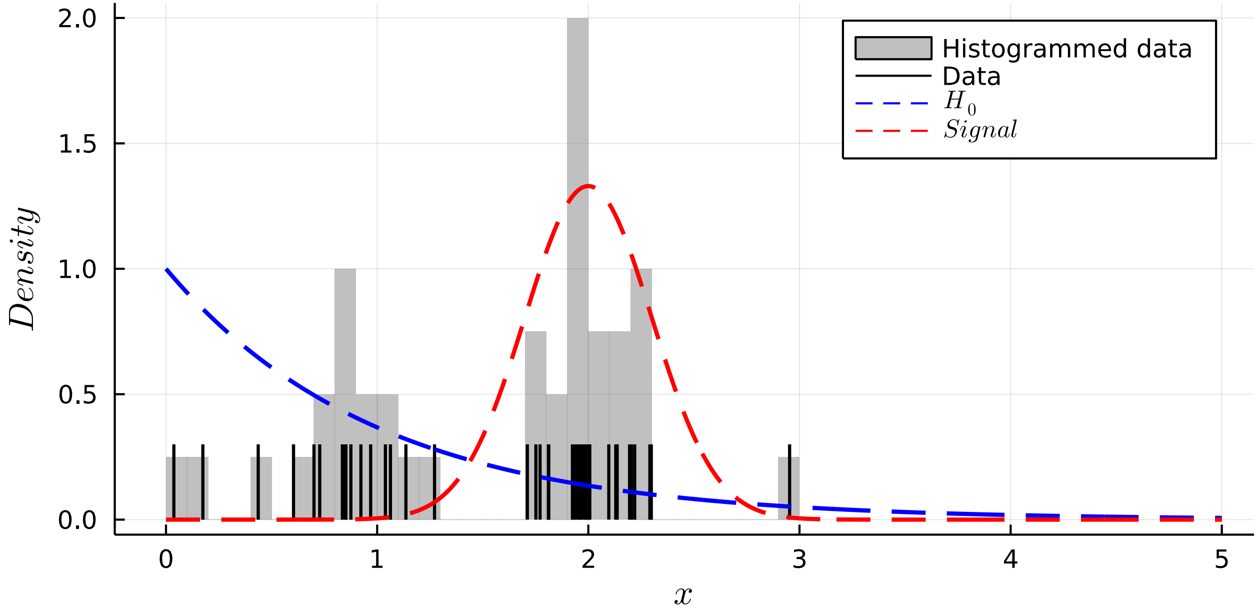
MAX-PLANCK-INSTITUT
FÜR PHYSIK

- The source of data can be split in two families:
    - expected
    - unexpected
- Depending on the scenario, we call them **background** or **signal**
- Depending on goal and knowledge, how to use a GOF?



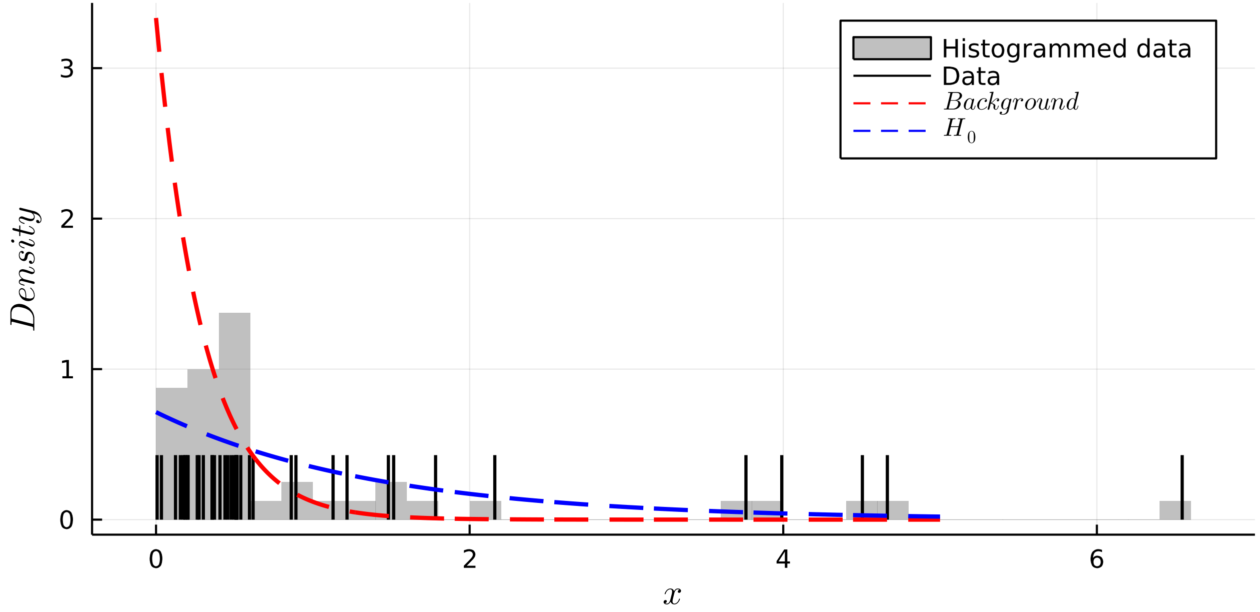
GOF for Discovery

- Use GOF for **discovery**:
  - the **background** is known: H_0
  - possibly an unexpected **signal**
  - reject H_0 if p-value is too small (Confidence Level)
  - **no assumption on signal** (no alternative hypothesis H_1)
  - use model to filter data



GOF for Limit-Setting

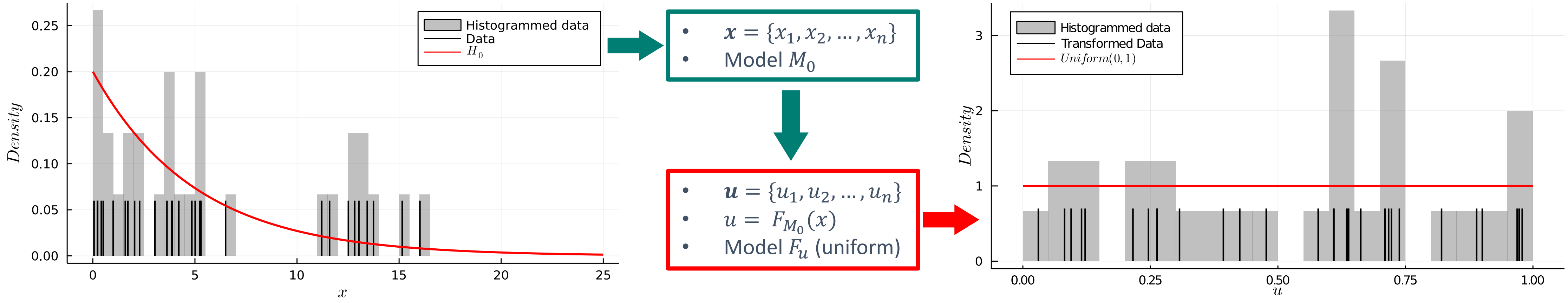
- Use GOF to **set a limit**:
  - the **signal** is know but not the rate: $H(\mu)$
  - possibly an unexpected **background**
  - **no assumption on background shape**
  - select μ to match a target p-value (Confidence Level)
  - use data to filter "models" (μ)



Probability Integral Transformation

MAX-PLANCK-INSTITUT FÜR PHYSIK

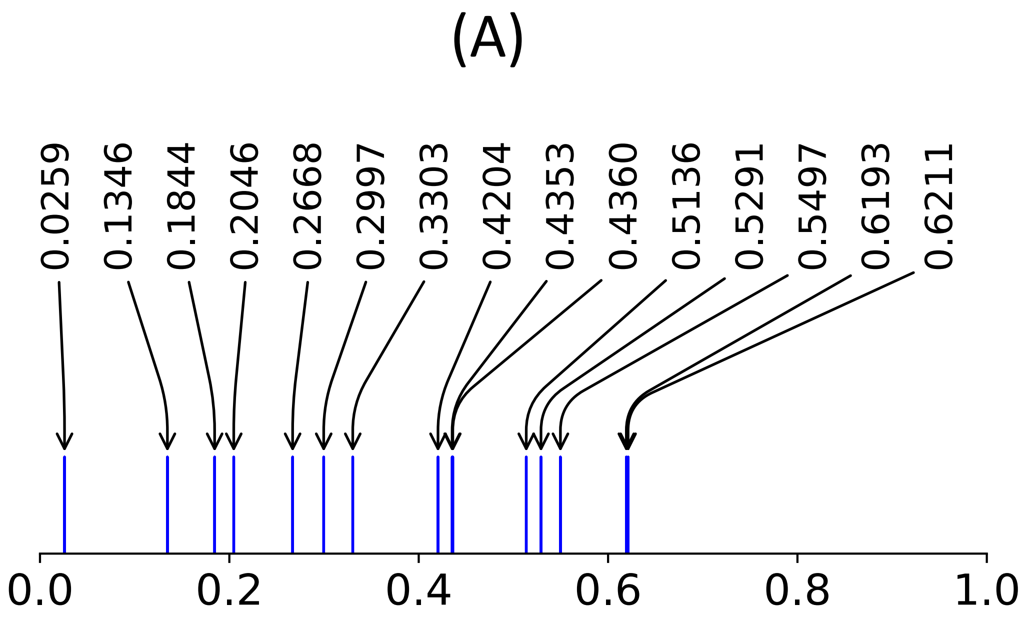
- Often Test Statistics are developed assuming uniform distribution of the null-hypothesis
  - is this enough?
  - What if the null-hypothesis H_0, is not uniform?
- **Probability integral transformation**
  - Transform the data into the cumulative space: $u_i = F_{H_0}(x_i)$
  - We now test if $\boldsymbol{u}$ are distributed according to the standard uniform $\mathcal{U}(0,1)$

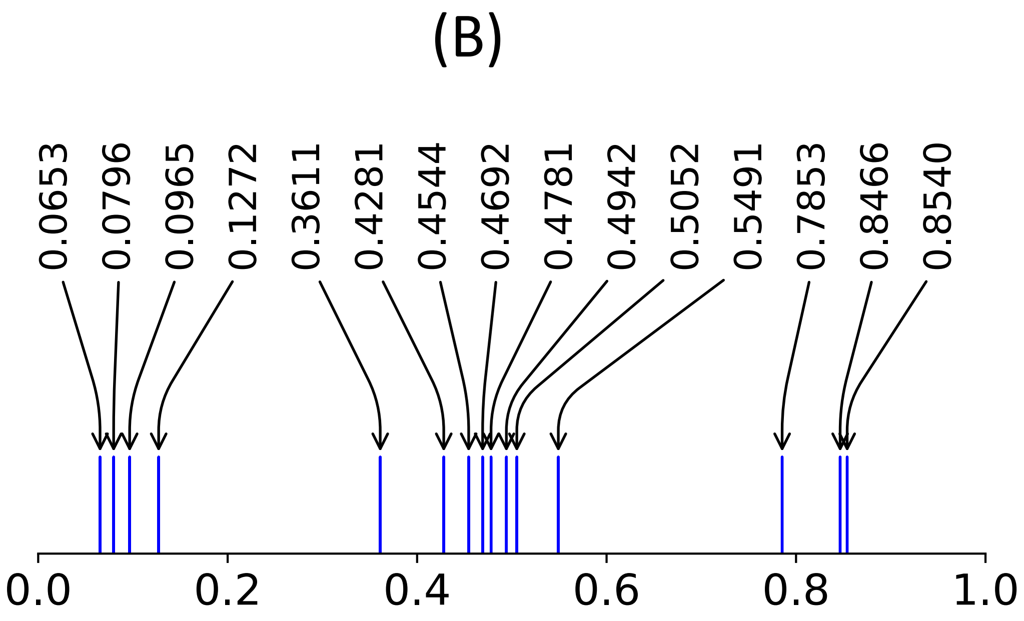


Which one is uniform?

MAX-PLANCK-INSTITUT
FÜR PHYSIK

- Which set of samples is drawn from a uniform distribution?





Analysis with test statistics

MAX-PLANCK-INSTITUT FÜR PHYSIK

- Binned analysis: χ^2 test $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

- Unbinned analysis: compute the likelihood $L = \prod_{i=1}^{n} p(x_i)$
  - but we need the distribution of the likelihood to get a p-value

- EDF tests:
  - Kolmogorov-Smirnov test (KS)

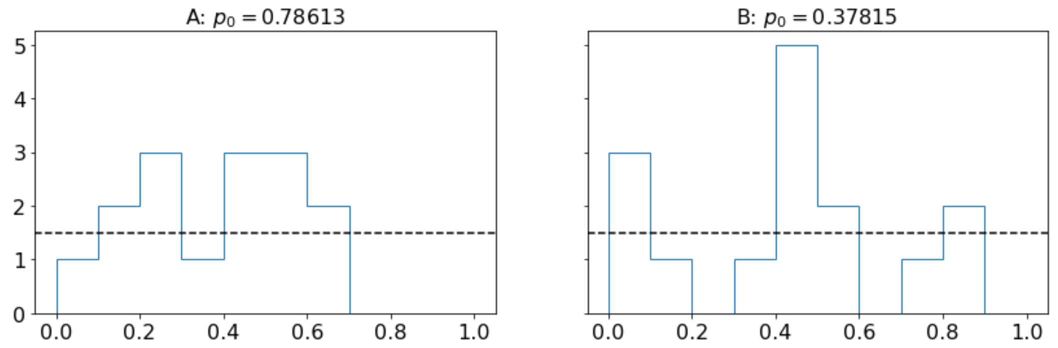
$$D_n = \sup_x |F_n(x) - F(x)|$$

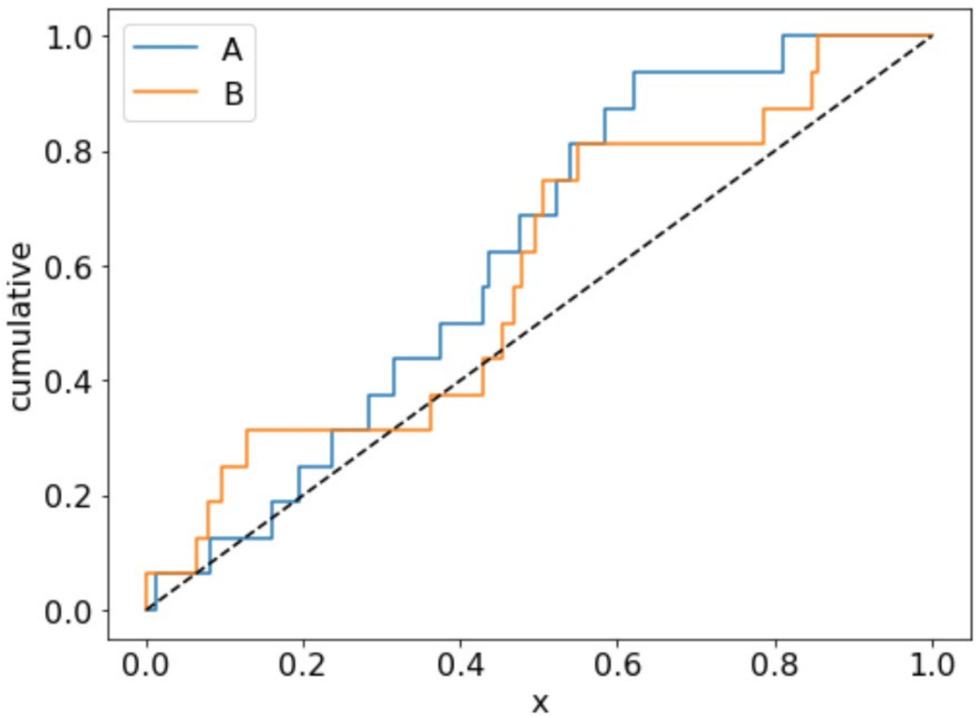
  - Cramér-von Mises test (CvM)

$$T = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \, dF(x)$$

  - Anderson-Darling test (AD)

$$T = n \int_{-\infty}^{+\infty} \frac{(F_n(x) - F(x))^2}{F(x)\,(1 - F(x))} \, dF(x)$$





Ordered samples and spacings

- A feature of our data that we have not yet fully explored
- We can order the data → welcome to field of **Order Statistic**

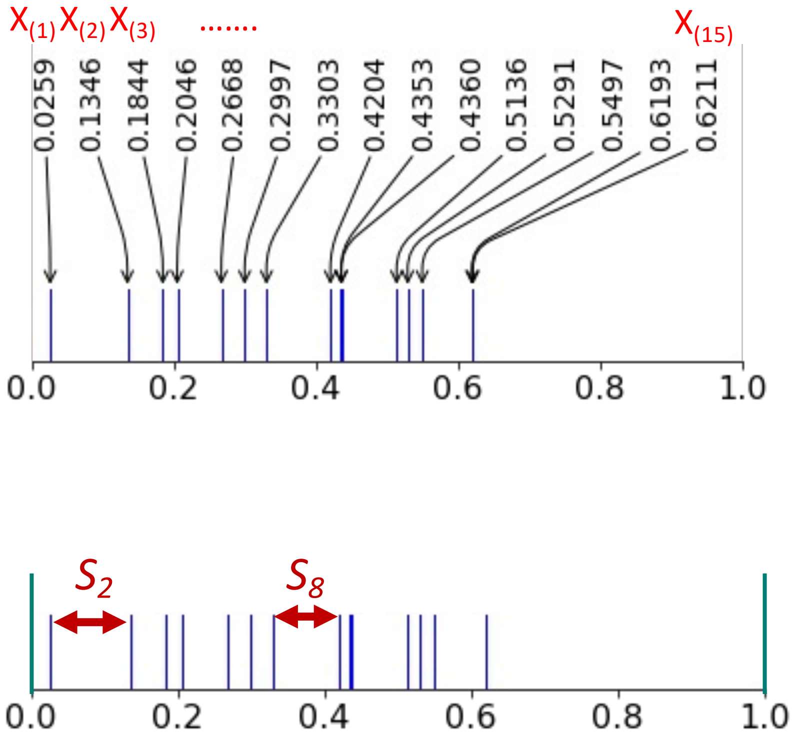
$$\{x_1, x_2, \ldots, x_n\} \Rightarrow \{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}, \text{ where } x_{(i)} < x_{(i+1)} \ \forall i$$

$$x_{(k)} \sim Beta(k, n - k + 1)$$

- Given n samples we can define $n+1$ ordered spacings $\boldsymbol{s}$:
- With left and right edges $x_{(0)} = 0$ and $x_{(n+1)} = 1$

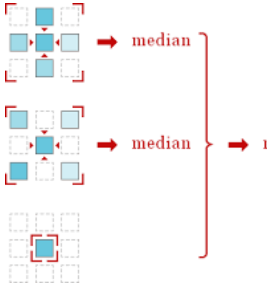
$$s_i = x_{(i)} - x_{(i-1)}$$

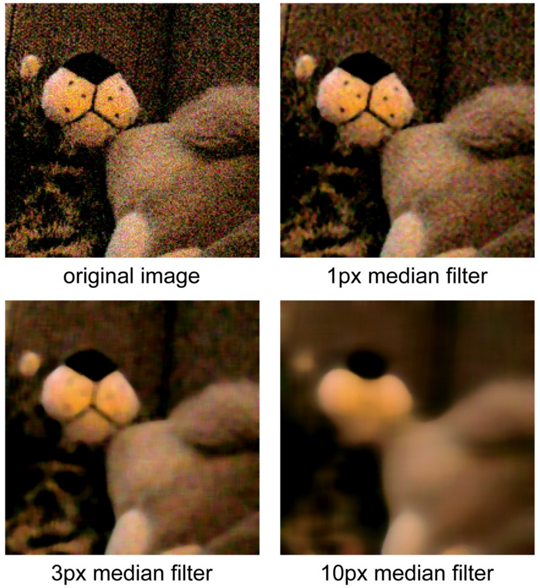
$$x_{(k)} - x_{(j)} \sim Beta(k - j, n - (k - 1) + 1)$$



History of Order Statistics

- Order statistics make their appearance in many areas of statistical theory and practice and by no means is it a new subject…
  - **Extremes** $X_{(1)}$ and $X_{(n)}$:
    - study of floods and droughts
    - problems of breaking strength and fatigue failure
    - auction theory
  - **Median** $X_{(n/2)}$:
    - robust estimator of location
    - used as smoother for time series (median filter) in signal and image processing
  - **Linear functions of order statistics**:
    - can be used to estimate parameters of location and scale of a distribution, especially with “censored” data (no time info on samples)





Tests based on spacings

MAX-PLANCK-INSTITUT
FÜR PHYSIK

- The literature regarding tests based on spacings is very rich…

Tests based on sum: $F_n = \sum_{i=1}^{n} f_n(s_i)$

- Greenwood (1946): $f_n(x) = x^2$
- Kimball (1950): $f_n(x) = x^r$ for $r > 0$
- Irwin (1946): $f_n(x) = \left(\frac{x}{n+1}\right)^2$
- Kendall (1946): $f_n(x) = \left|\frac{x}{n+1}\right|^2$
- Moran, Darling (1953): $f_n(x) = \log(x)$
- Darling (1953): $f_n(x) = \frac{1}{x}$

Tests based on ranked spacings:

$g_i = i$-th smallest spacing

- Fisher (1929): g_1 and g_{n+1} (Darling, Pincus, etc…)
- Kendall (1946): $\frac{g_{n+1}}{g_1}$ and $g_{n+1} - g_1$
- Mauldon (1951): $g_{(n-k+1)} + g_{(n-k+2)} + \ldots + g_{(n+1)}$
- $s_1 + s_2 + \ldots + s_k$

arXiv:2008.02048

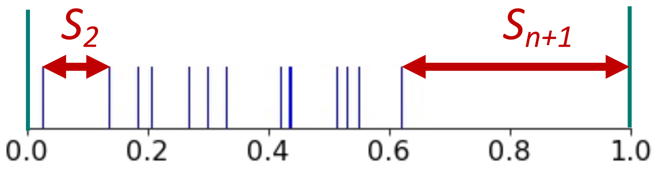
Recursive Product of Spacings

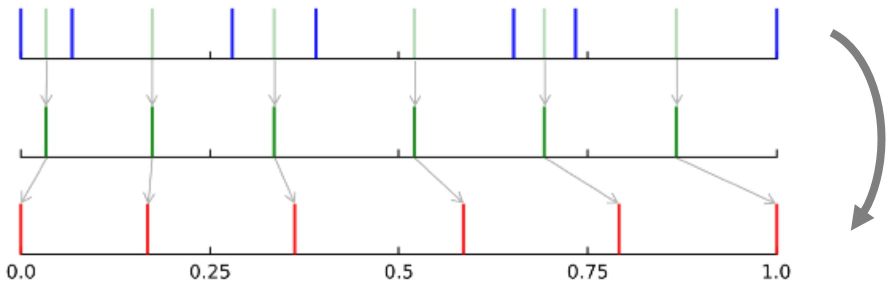
MAX-PLANCK-INSTITUT
FÜR PHYSIK

- Using Moran's test statistic:

$$M^{n+1} = -\sum_{i=1}^{n+1} \log s_i$$

- Reduce levels using mean value and normalize
- Apply Moran to all levels
- Sum contribution from all layers





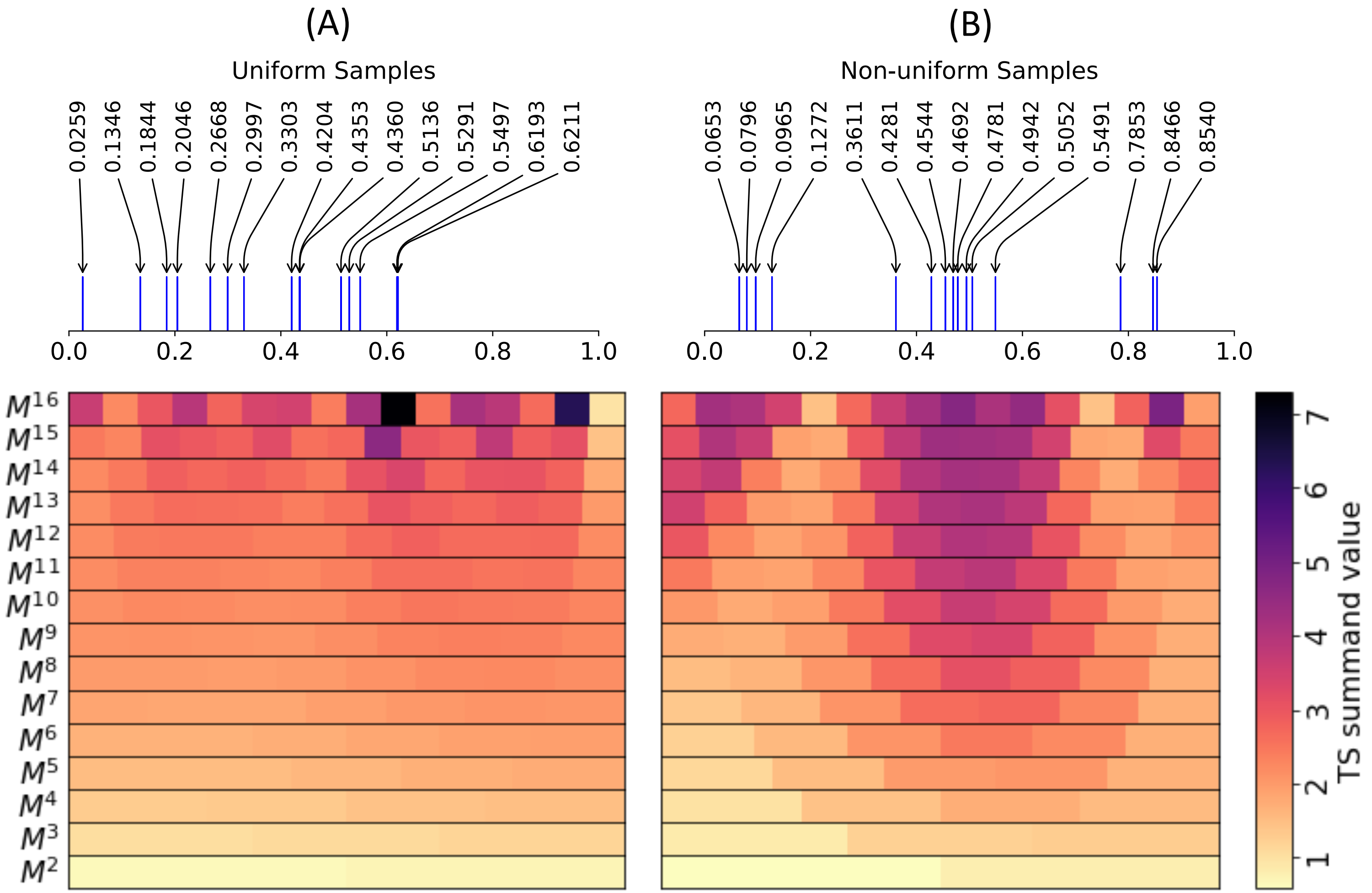
$$M^j = -\sum_{i=1}^{j} \log\left(s_i^j\right)$$

$$RPS(n) = M^{n+1} + M^n + \cdots + M^1$$

$$s_i^j = \frac{s_i^{j+1} + s_{i+1}^{j+1}}{\sum_i s_i^j}$$

arxiv.org/abs/2111.02252

RPS in practice

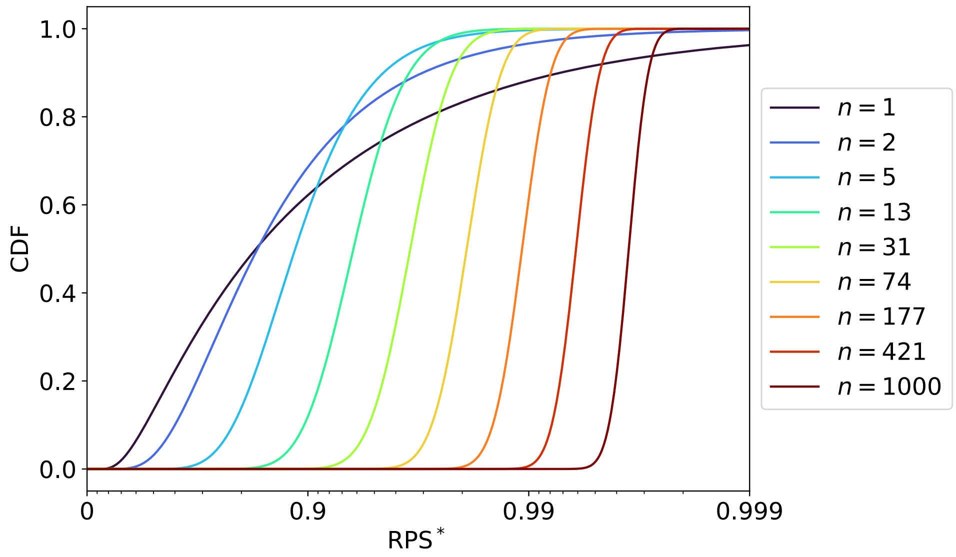


$p_0 = 0.53$

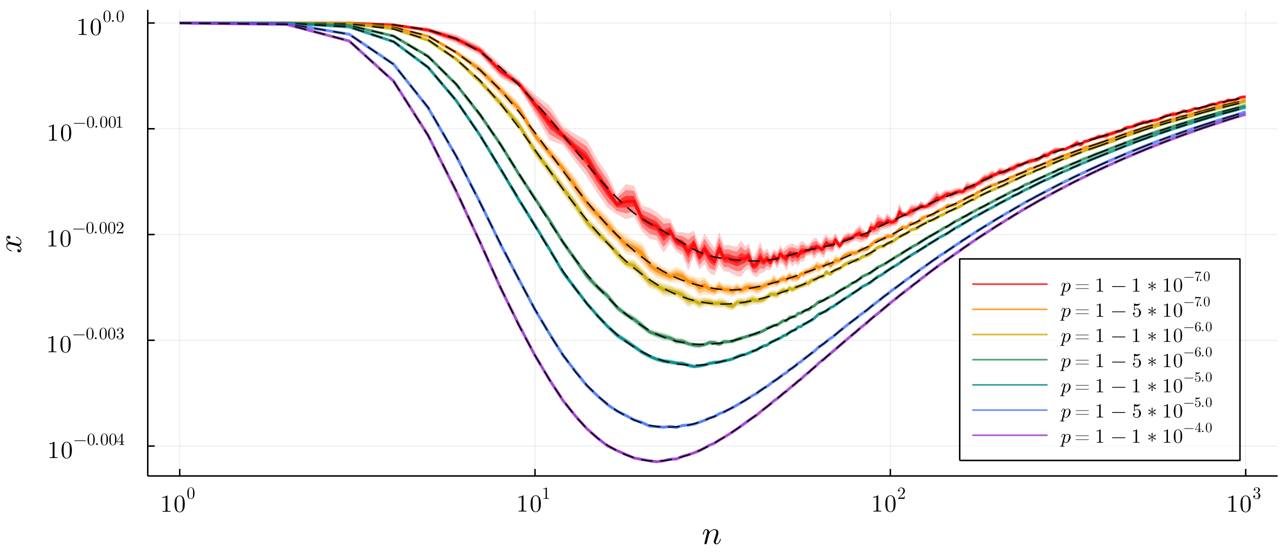
$p_0 = 0.057$

RPS distribution

- We need the cumulative distribution $CDF(RPS \mid n)$
  - deriving the distribution of combinations of spacings is not trivial for $n > 2$
  - Instead of an analytic formula, parametrize the distribution of RPS
  - 2D spline interpolation, based on a large set of simulations for $n \leq 1000$



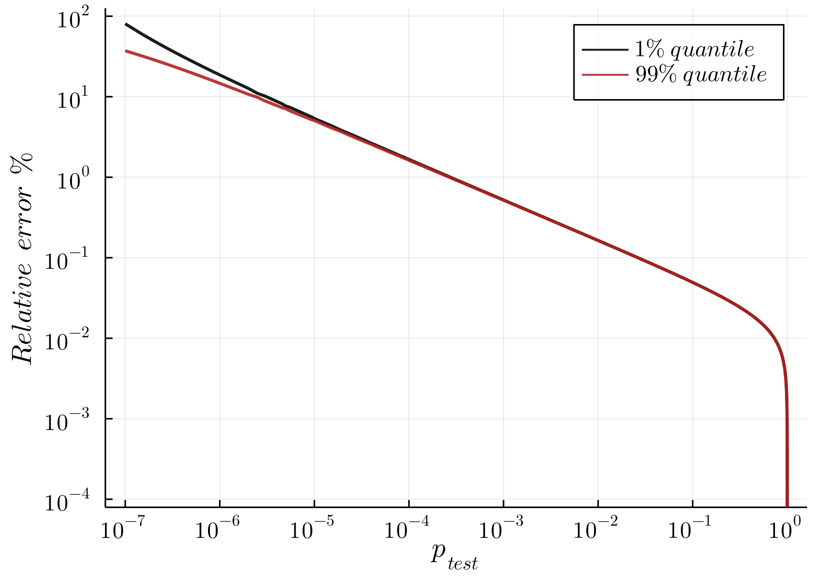
arxiv.org/abs/2111.02252

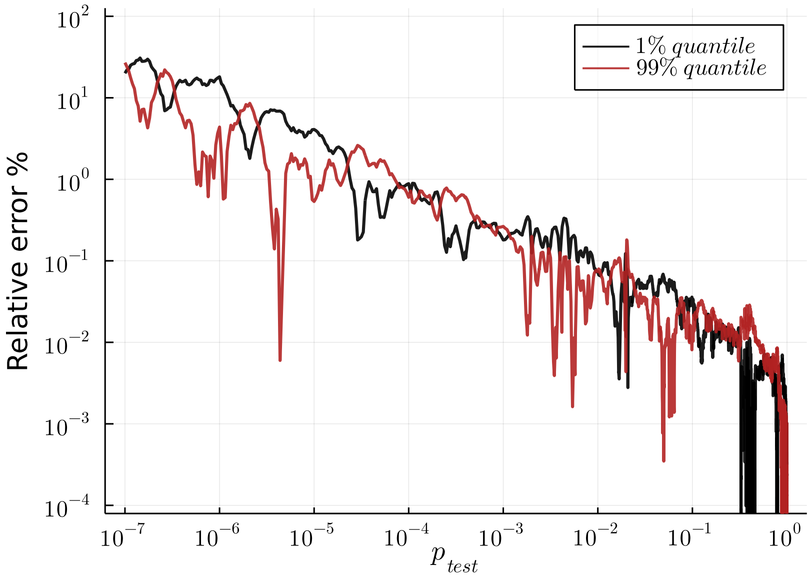


pypi.org/project/spacings/

Error estimation

- It's possible to estimate the relative error of an EDF constructed with n samples
- For any set of i.i.d. variables, the corresponding set of EDF quantiles is a random set of uniform variables
- The EDF quantiles are order statistics (their distribution given n samples is known)
- We can estimate the relative error of an observed quantile against its distribution



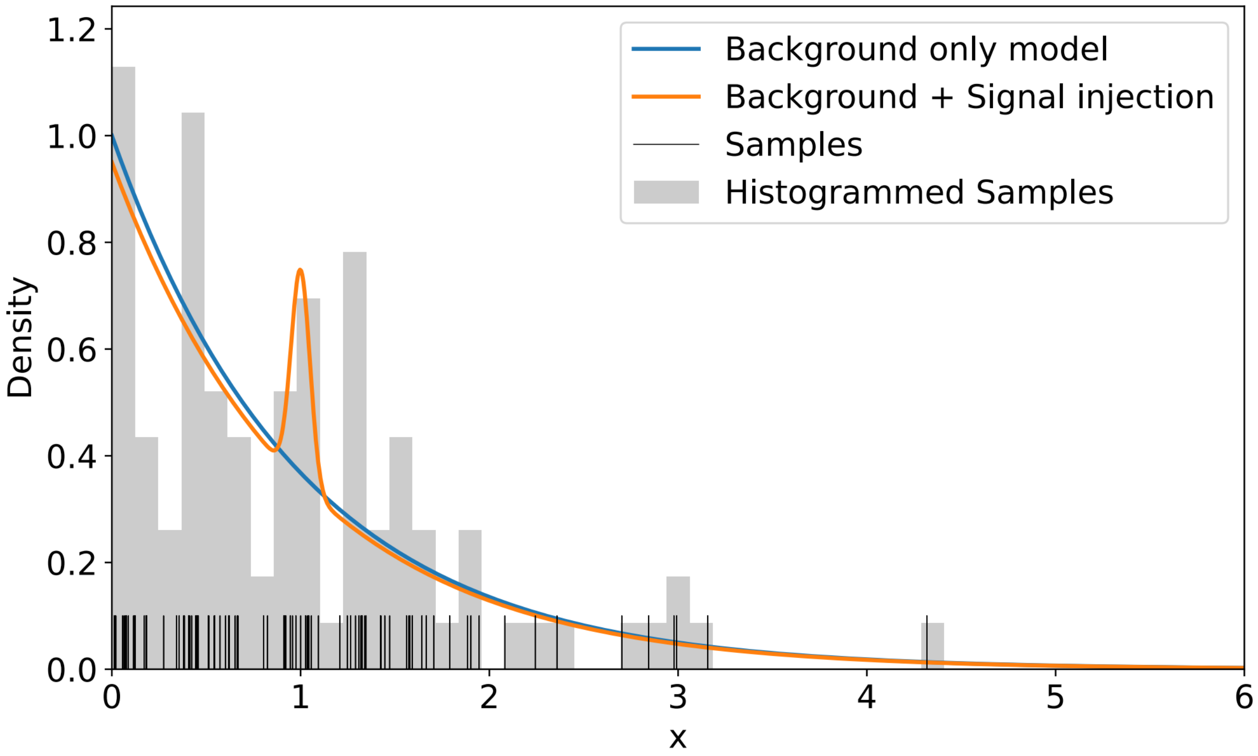


comparison to numerical estimation

Realistic example

MAX-PLANCK-INSTITUT
FÜR PHYSIK

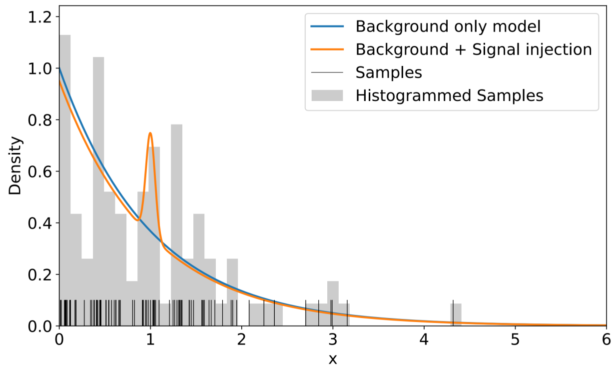
- Quantify how (in)compatible my observations are with a background distribution (here exponential)
- For a test signal, inject some events as a narrow Gaussian $\mathcal{N}(1, 0.05)$

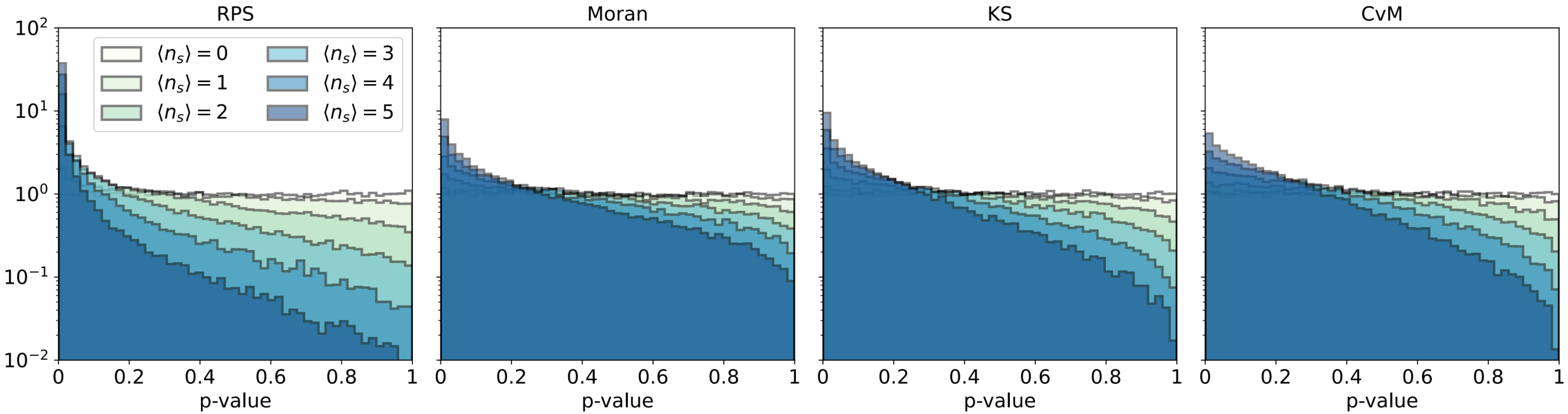


Realistic example

MAX-PLANCK-INSTITUT
FÜR PHYSIK

- Repeated trials of random number of background events: $n_b \sim Poisson(100)$
- Inject signal random number of signal events: $n_s \sim Poisson$





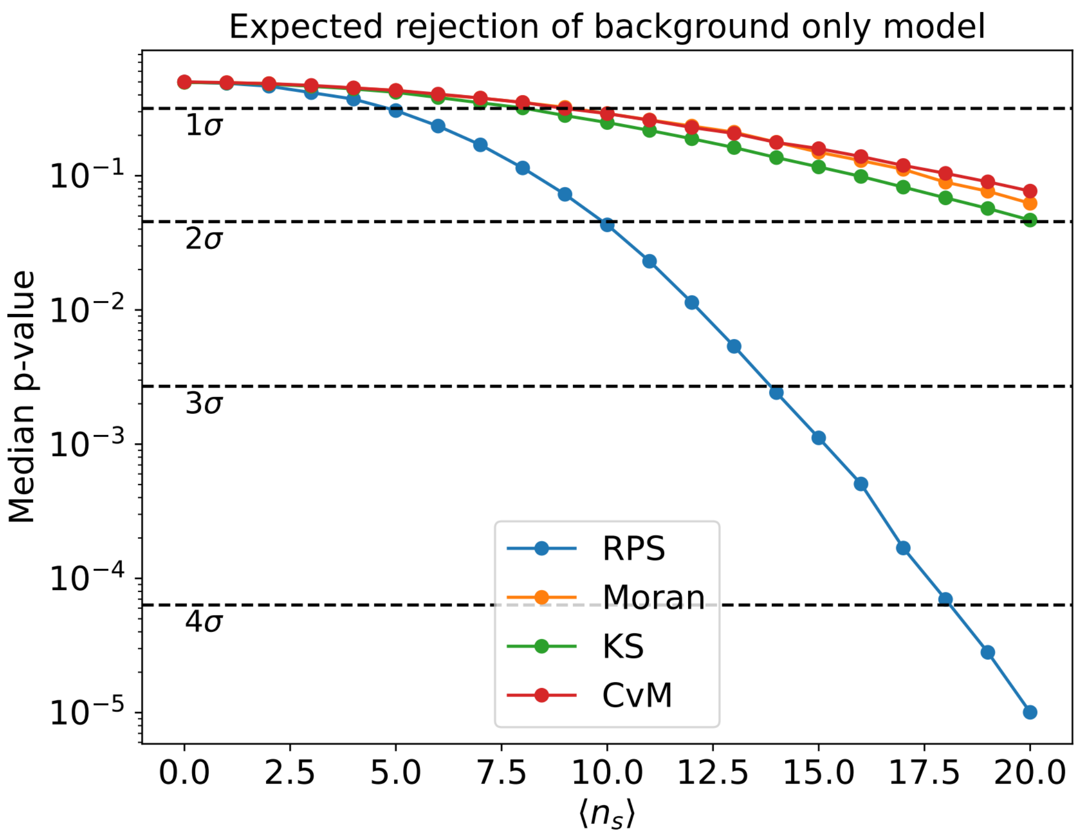
- for $\langle n_s \rangle > 0$ all p-value distributions trend towards smaller p-values → worsened GOF for bkg only model
- RPS test offers the largest rejection probability of the null hypothesis.

Sensitivity

- What p-value (**median**) would one expect as a function of the injected signal?

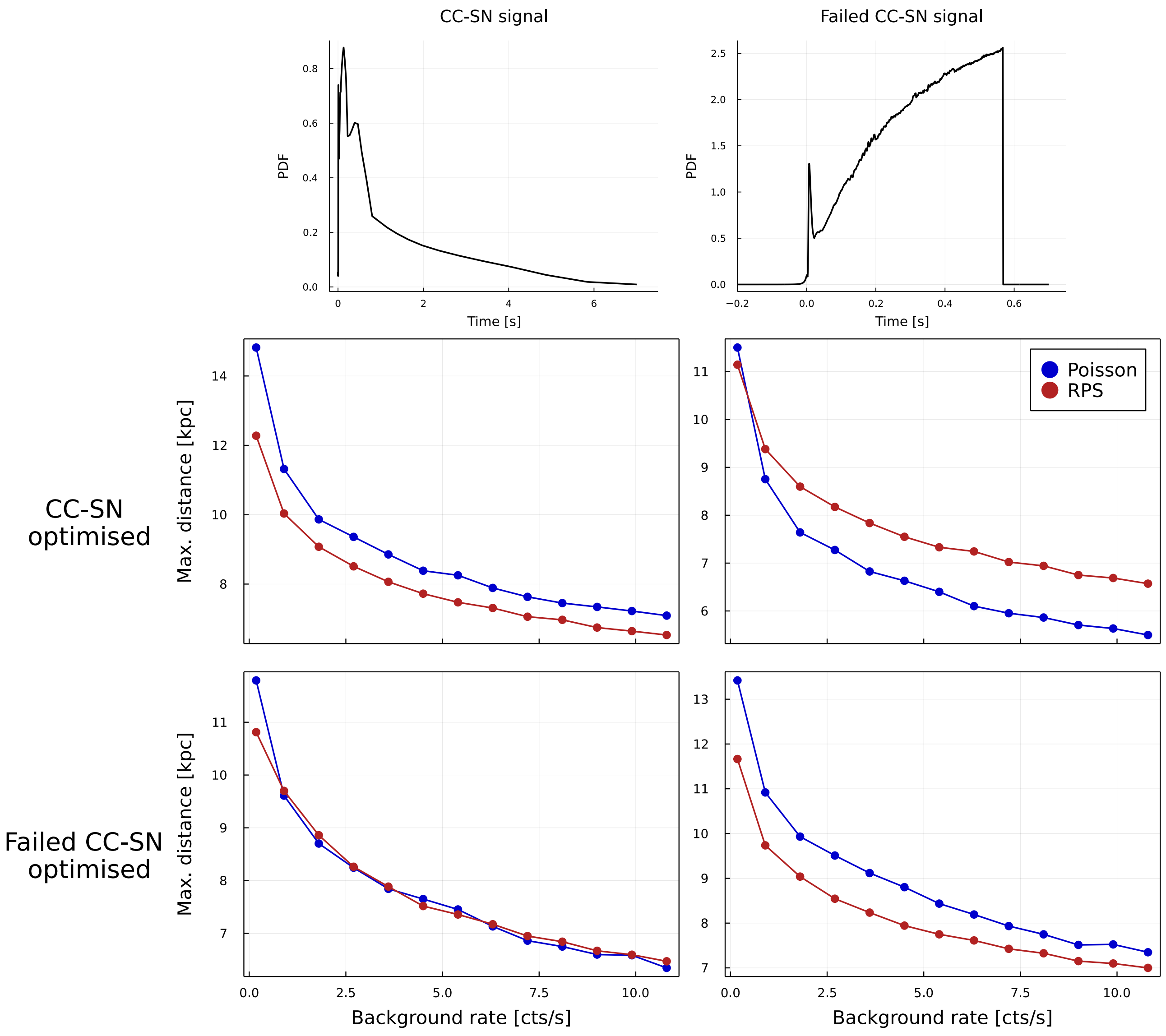
- The TS with highest **sensitivity** can be used to validate model selections for further studies

- TS can be used as a fast filter in large datasets in order to select “interesting” sections of data to be later analyzed more in depth (Bayesian analysis for example)



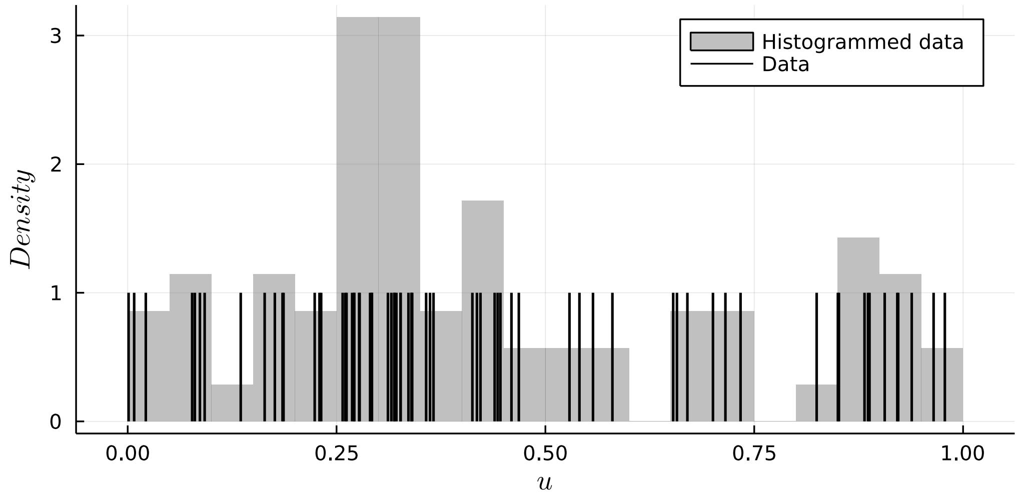
RPS for online trigger of SN Neutrino Bursts

- ❑ Online search of **Neutrino bursts from Supernovae**
- ❑ Successful signal recognition depends on:
  - signal distance → signal rate
  - background rate
- ❑ Optimize analysis parameters to maximize detection horizon at set success rate:
  - **optimization dependent on signal hypothesis**
- ❑ Study how detection horizon of a frozen optimized model changes:
  - **signal change**
  - **background increases**
- ❑ RPS test more robust against signal variation and increased background
- ❑ Case study for RES-NOVA JCAP 10 (2022) 024

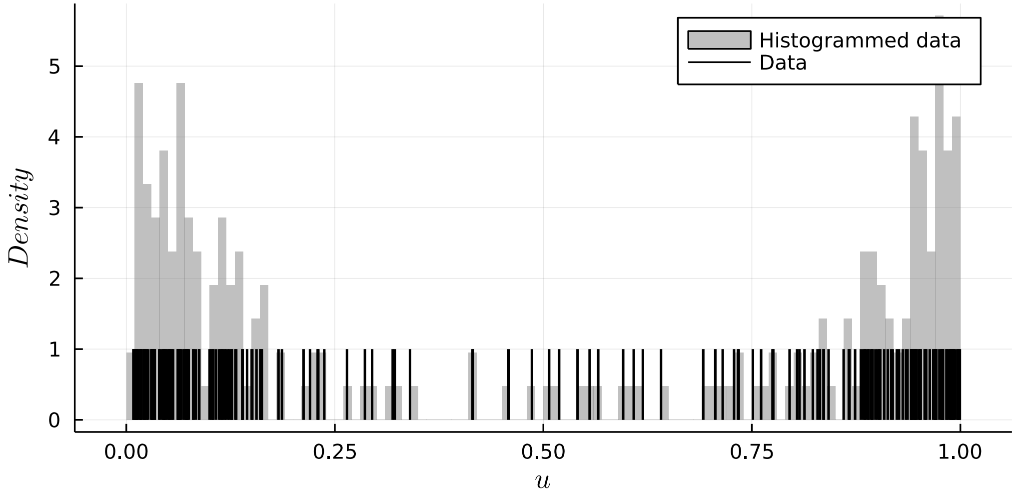


Setting an upper limit

- ❑ **Goodness-of-fit test** for **discovery**



- ❑ **Goodness-of-fit** test for **limit setting**
  - prefer regions with low event density
  - look at large spacings
  - filter out regions with high event density



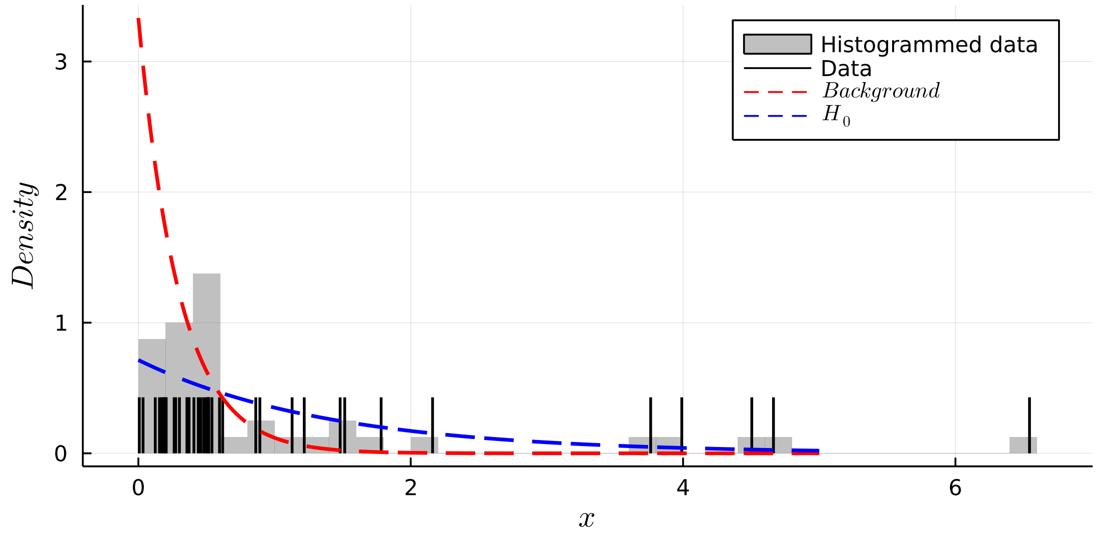
Setting limits with spacings

MAX-PLANCK-INSTITUT
FÜR PHYSIK

- Consider a univariate dataset
- Suppose we know the shape of the signal distribution
- Unknow normalization
- There might also be an unknown background

↓

- Estimate an upper limit on the signal
- Yellin* proposes 2 methods: arXiv:physics/0203002
  - "Maximum Gap" and "Optimum Interval"
  - used in many direct dark matter search experiments (CRESST, CDMS, EDELWEISS,…)



"Refining" test statistics: number of samples

MAX-PLANCK-INSTITUT FÜR PHYSIK

- So far, we only considered test statistics (TS) for a given number of observed events, n
- What if n is an observable (random variable) too?
  - We assign n a distribution and integrate TS over it:

$$F(t|\mu) = \sum_{n=N_{min}}^{\infty} p(n|\mu) \cdot F(t|n)$$

$$n \sim \text{Poisson}(\mu)$$

$$t = TS(\boldsymbol{x})$$

  - N_{min} is the smallest number of samples that produce the desired test statistic

Maximum Gap method

- Spectrum dN/dE for a proposed cross section σ
- Total expected number of events:

$$\mu = \int_{E_{min}}^{E_{max}} \frac{dN}{dE}$$

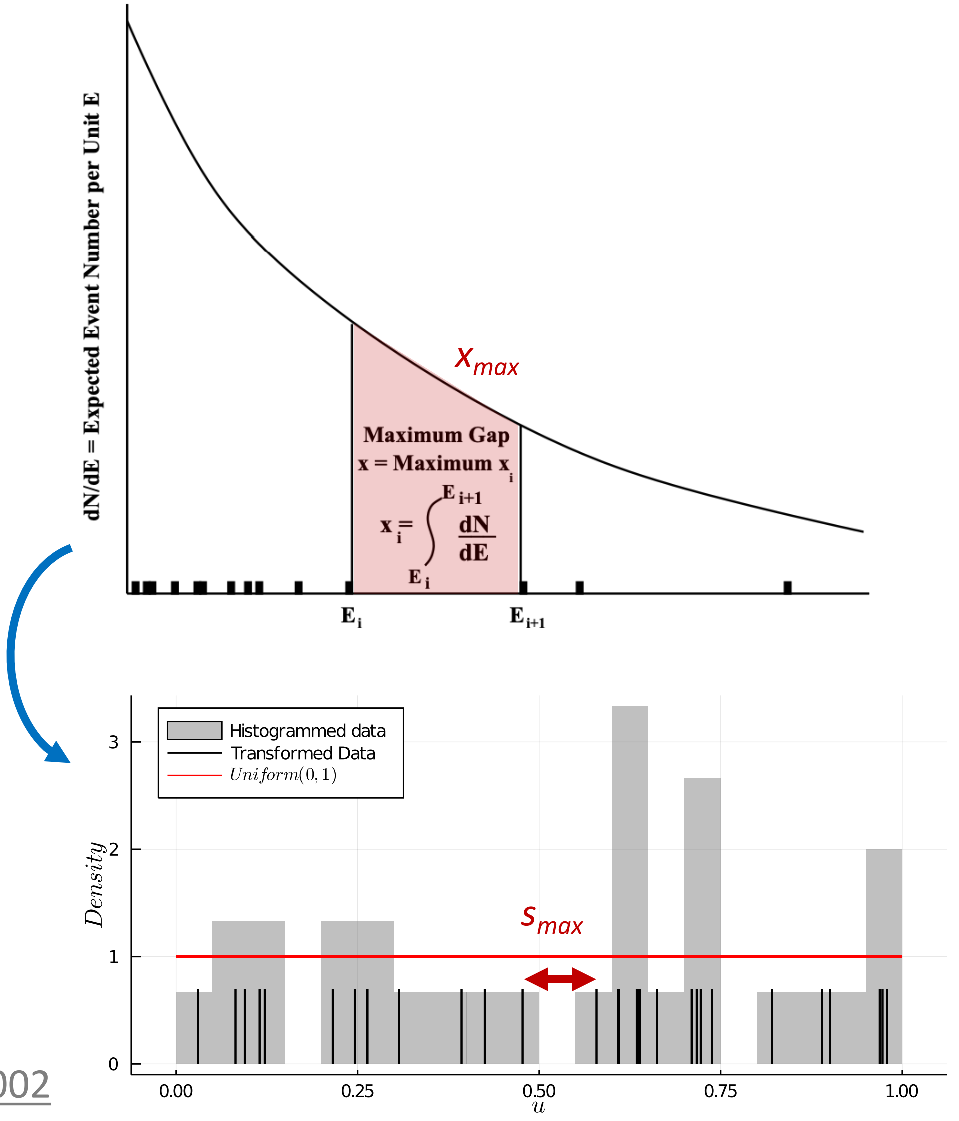
- Evaluate expected number of events in each gap
  - similar to the "probability integral transform"
- Test statistic:

$$TS(\boldsymbol{E}) = \frac{x_{max}}{\mu} = s_{max} = \max_i s_i$$

- p-value:

$$p = \Pr(TS \le s_{max} \mid \mu)$$

- Upper limit at 90% Confidence Level:
  - Find μ such that $p = 0.9$



arXiv:physics/0203002

Optimum Interval method

MAX-PLANCK-INSTITUT FÜR PHYSIK

- Instead of looking only at one gap at a time, look at collections of gaps
  - $s_i = s_{1,i} = u_i - u_{i-1}$ **(Ordered Spacings)**
  - $s_{k,i} = u_i - u_{i-k}$ **(Sum of Ordered Spacings)**
- For each order k, find the largest sum or ordered spacings and its p-value:
  - $s_k^{max} = \max_i s_{k,i}$
  - $p_k = \Pr(s_k^{max} \leq obs. \mid \mu)$
- Test statistic:

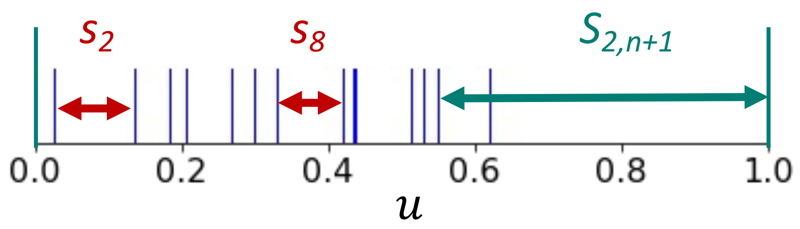
$$TS(\boldsymbol{u}) = p_{max} = \max_k p_k$$

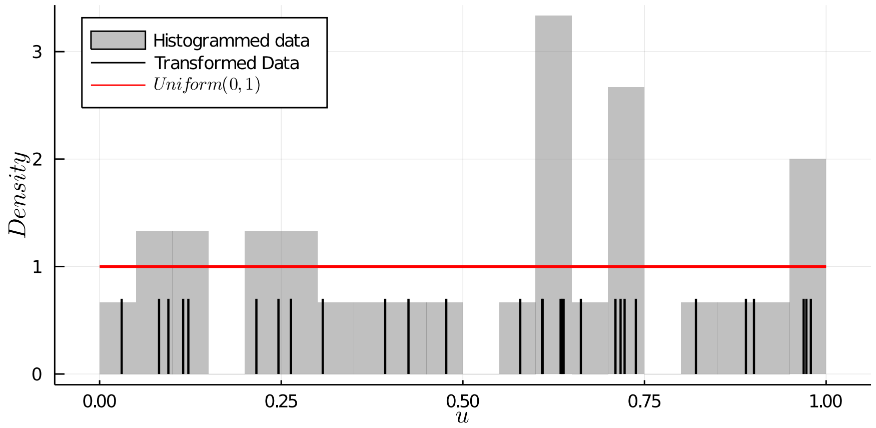
- Final p-value:

$$p_{final} = \Pr(TS \leq p_{max} \mid \mu)$$

- Upper limit at 90% Confidence Level:
  - Find μ such that $p_{final} = 0.9$

arXiv:physics/0203002





Sum of Sorted Spacings

- Sort spacings -> new event list
  - $g_1 = \min_i s_i, \ldots, g_{n+1} = \max_i s_i$ **(Sorted Spacings)**
- Consider **Sum of largest Sorted spacings:**

$$G_k = \sum_{i=n+2-k}^{n+1} g_i$$

- For each order k, get p-value of the sum of largest sorted spacings:

$$p_k = \Pr(G_k \leq obs. \mid \mu)^*$$

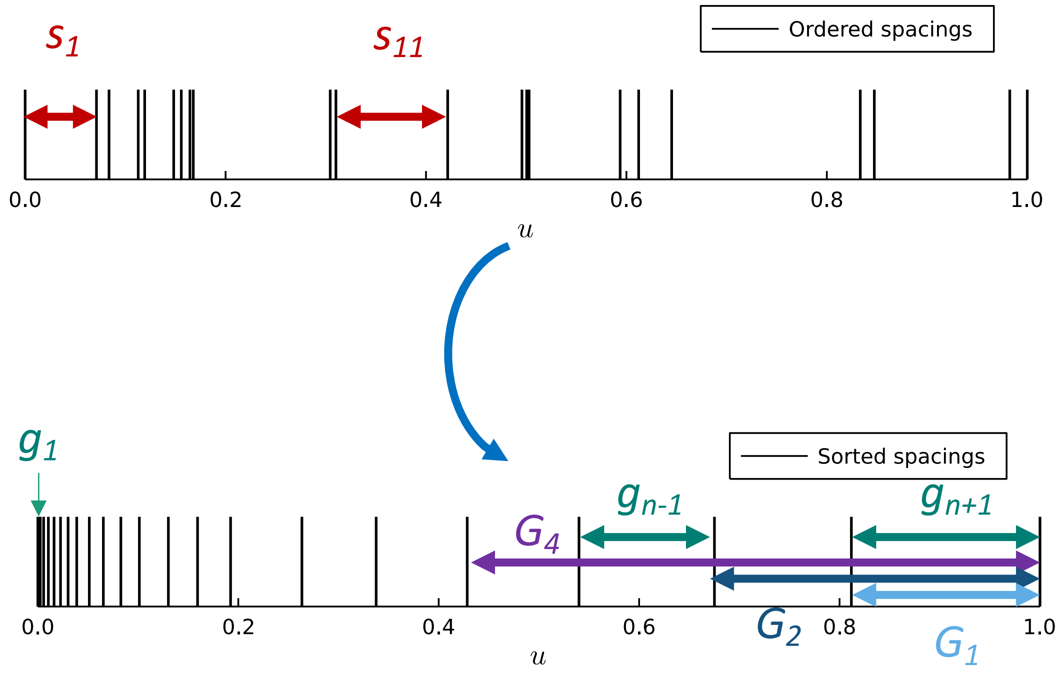
- Test statistic:

$$TS(\boldsymbol{u}) = p_{max} = \max_k p_k$$

- Final p-value:

$$p_{final} = \Pr(TS \leq p_{max} \mid \mu)$$

- Upper limit at 90% Confidence Level:
  - Find μ such that $p_{final} = 0.9$



arXiv:2008.02048

Product of Complementary Spacings

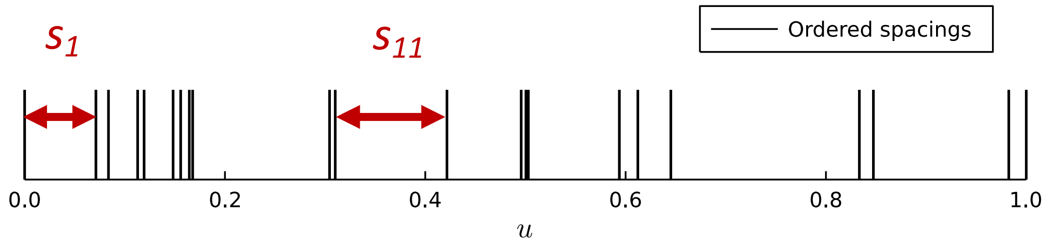
MAX-PLANCK-INSTITUT
FÜR PHYSIK

- Moran's test (sensitive to small spacings):

$$M(\boldsymbol{s}) = -\sum_{i=1}^{n+1} \log s_i$$

- Make this test sensitive to large spacings:
  - **consider complementary of each spacing**

$$PCS(\boldsymbol{s}) = -\sum_{i=1}^{n+1} \log(1 - s_i)$$



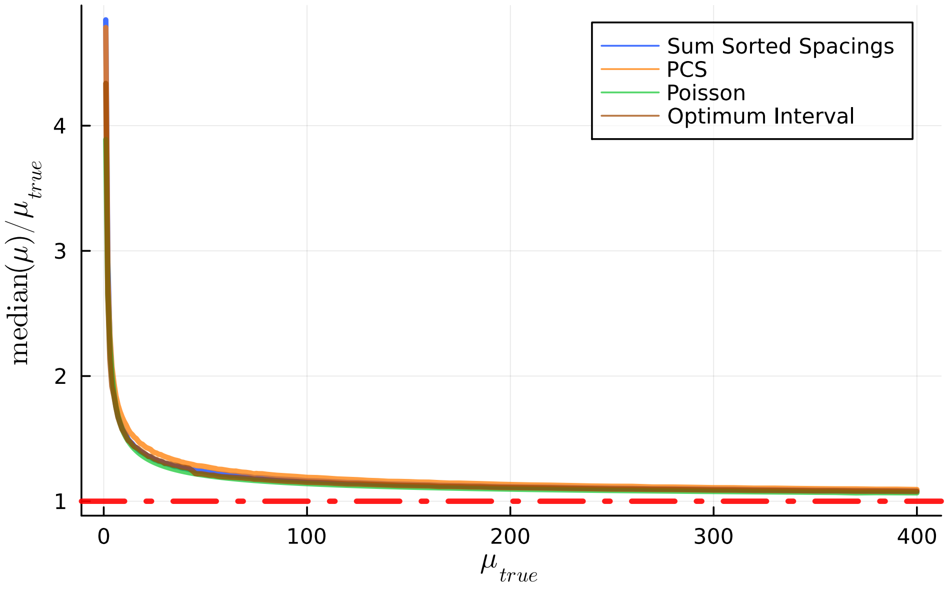
- Final p-value:

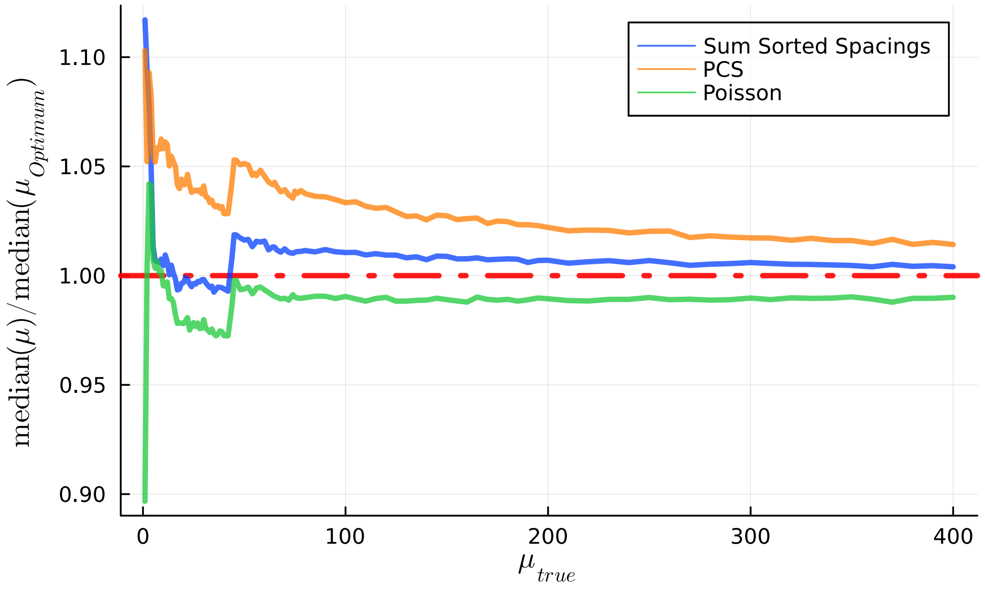
$$p_{final} = \Pr(PCS \leq obs. \,|\, \mu)$$

- Upper limit at 90% Confidence Level:
  - Find μ such that $p_{final} = 0.9$
- Simpler definition -> easier to work with

Limit setting: no Background

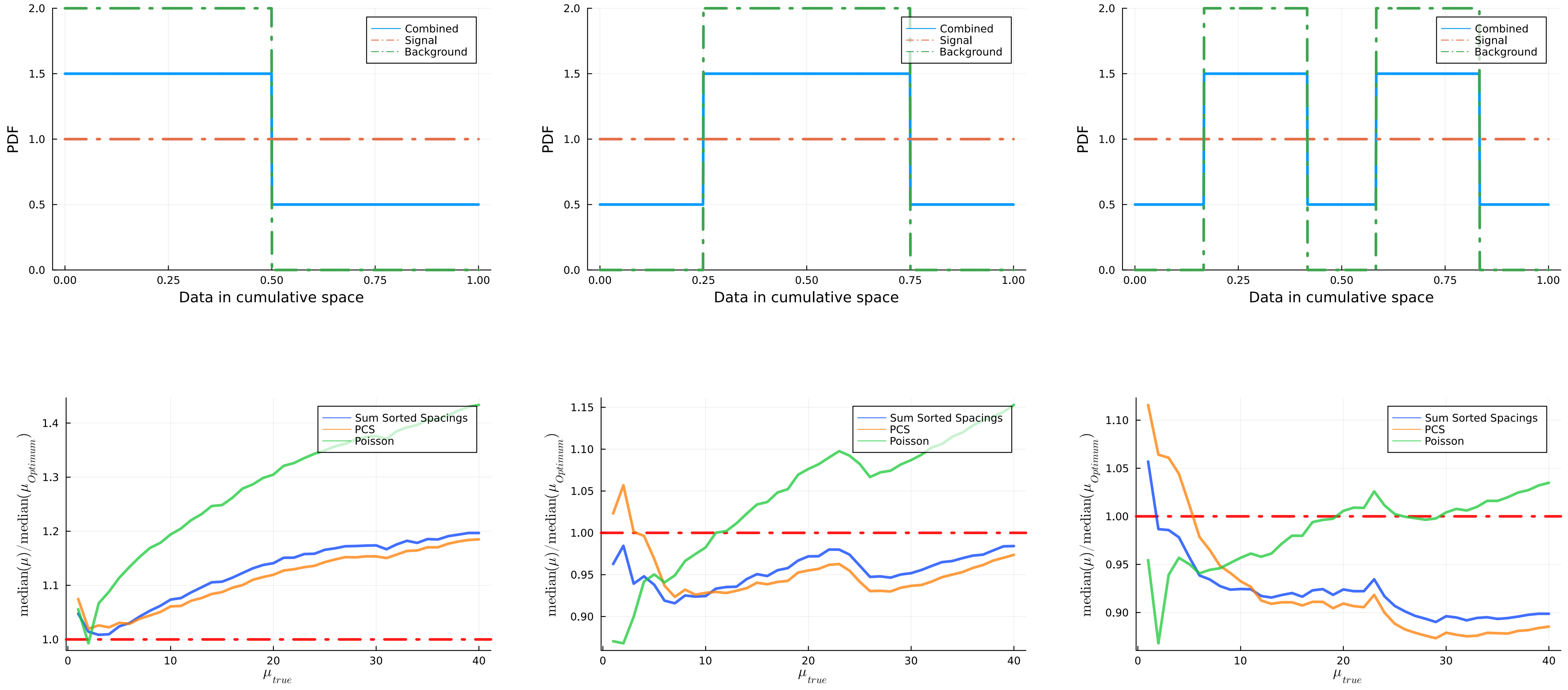
- ❑ Repeated trial experiments to estimate 90% CL limit
- ❑ Consider median of estimated event rates
- ❑ Compare medians to Optimum Interval method





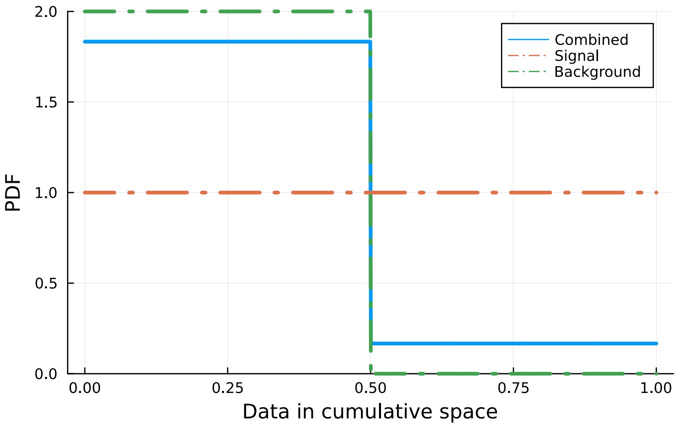
Limit setting: inject Background

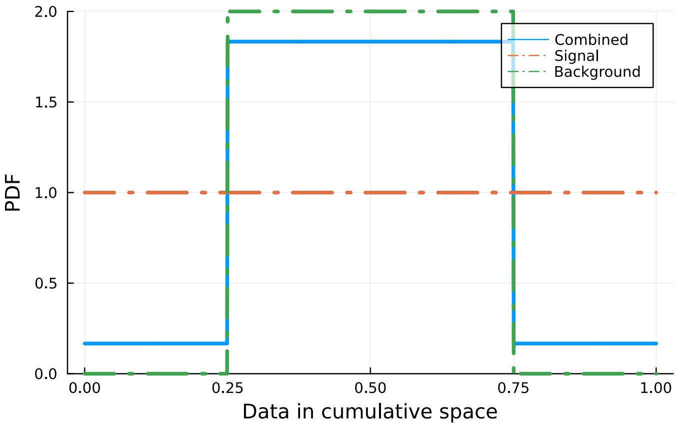
- Background / Signal = 1
- Background width = 0.5

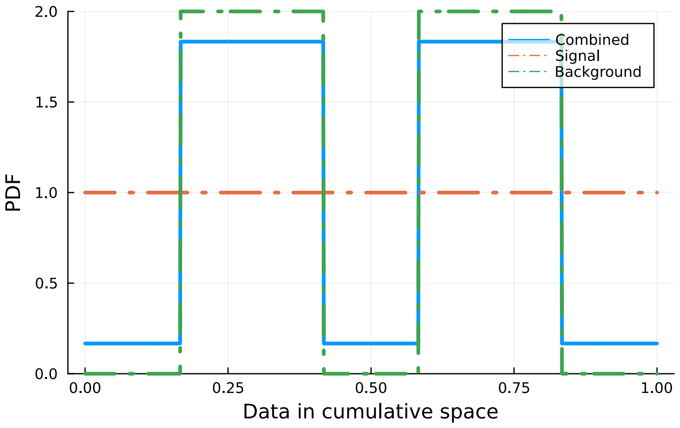


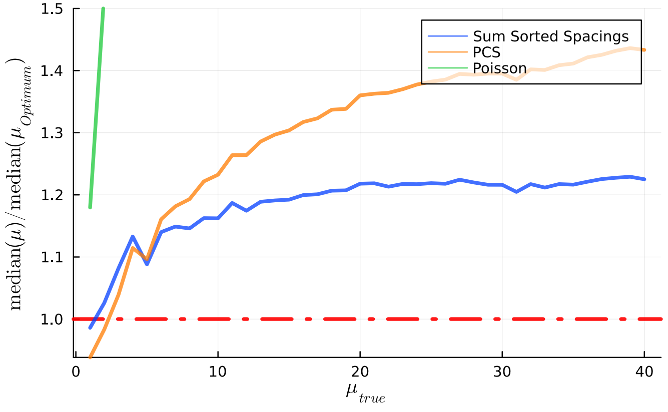
Limit setting: high Background

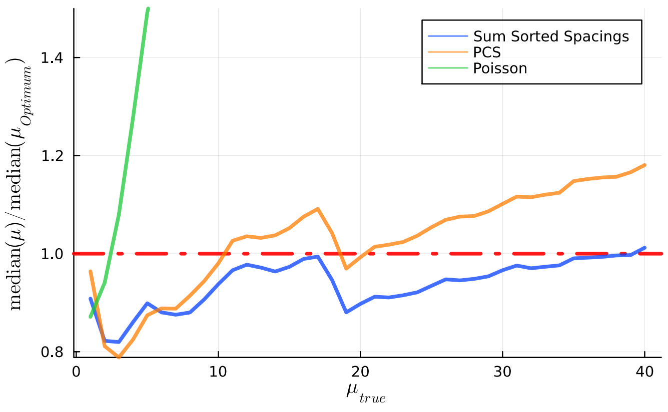
- Background / Signal = 5
- Background width = 0.5

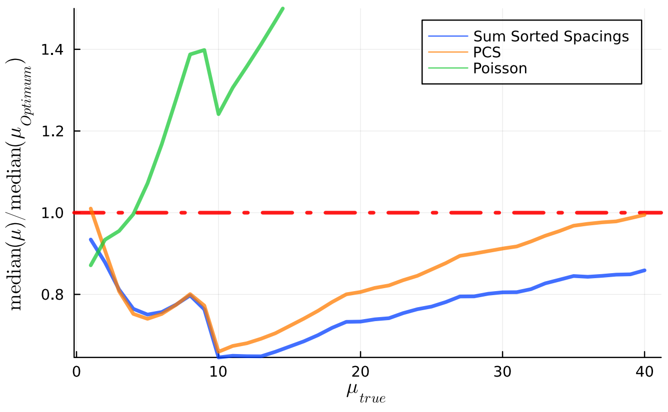








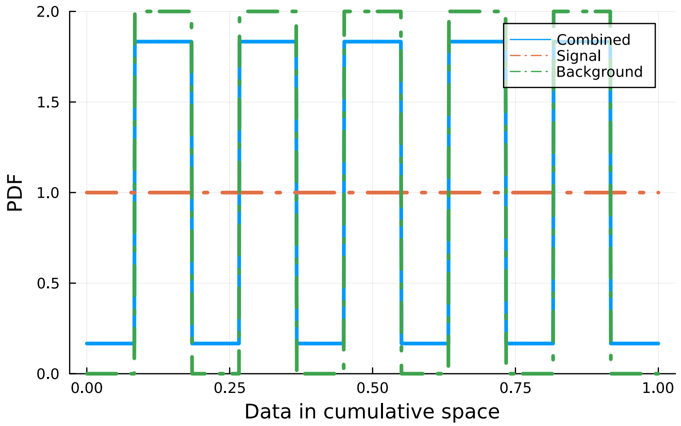


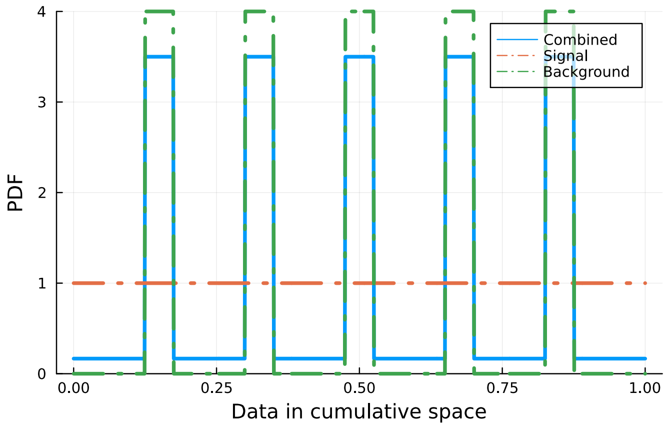


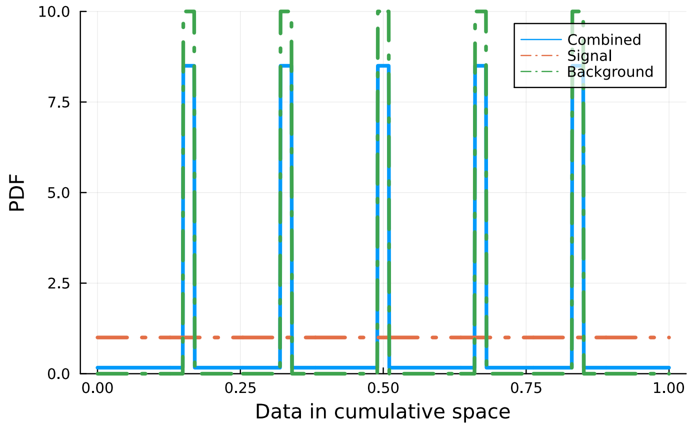
Limit setting: non smooth Background

MAX-PLANCK-INSTITUT FÜR PHYSIK

- Background / Signal = 5



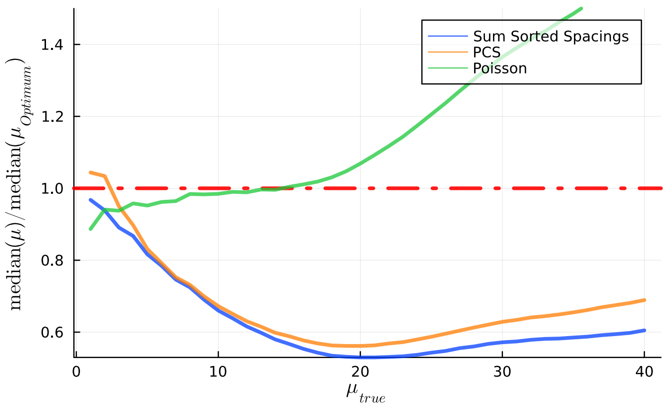


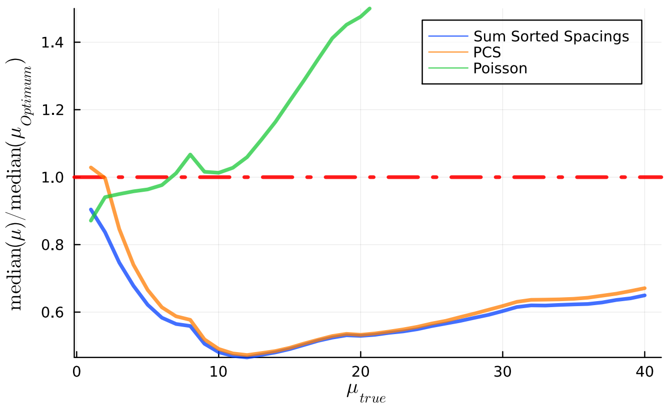


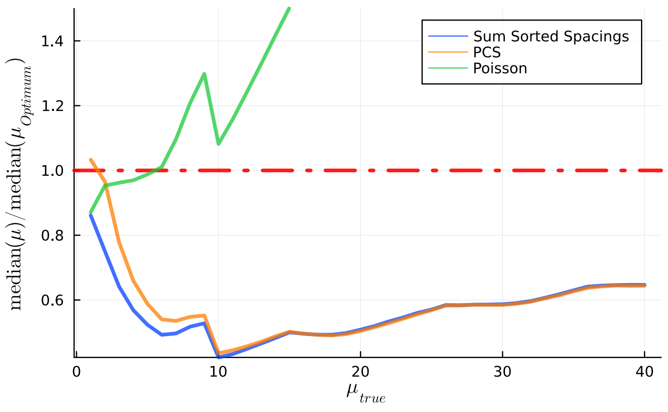
Background width = 0.5

Background width = 0.25

Background width = 0.1

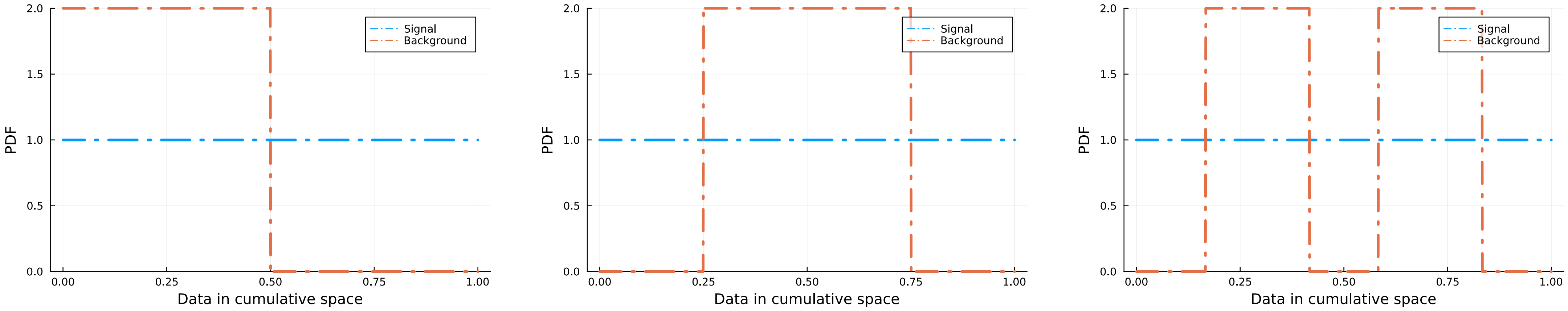


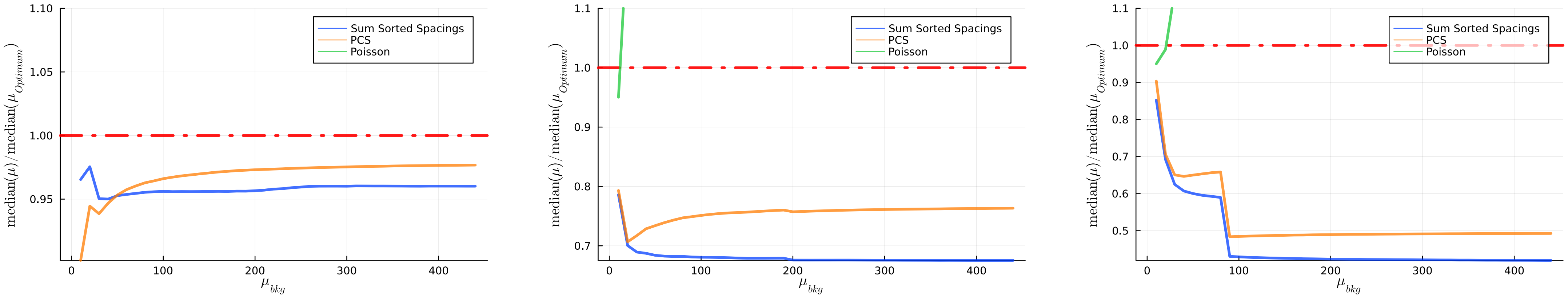




Limit setting: no Signal

- Background width = 0.5



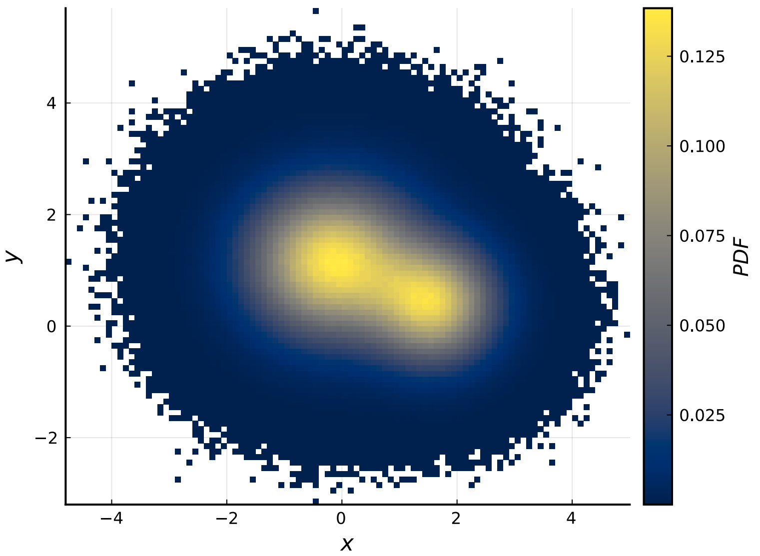


Ongoing work

- Currently working on targeting multivariate distributions

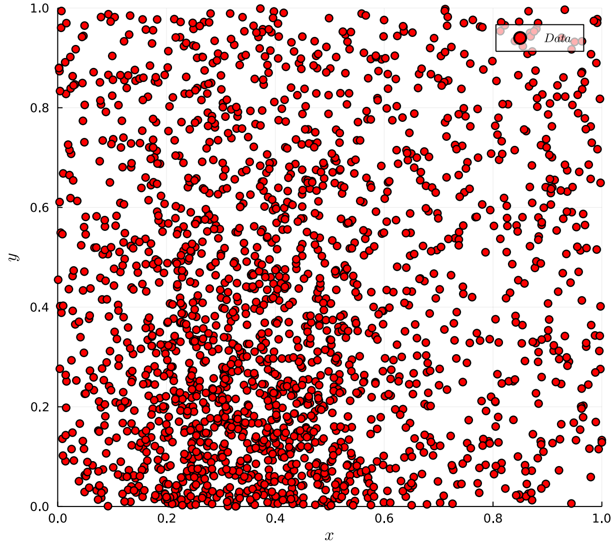
A general method for goodness-of-fit tests for arbitrary multivariate models

arXiv:2211.03478



Limit setting in multiple dimensions

publication in preparation



Conclusions and future work

- Test Statistics and their distributions are very useful
- There is no one right test statistic, it is very problem-dependent (unfortunately)
- Introduced **RPS** for univariate GOF
- Introduced **Sum of Sorted Spacings** and **Product of Complementary Spacings** for univariate limit-setting

- Develop method for **multivariate GOF**
- [WIP] **limit-setting in n dimensions** (2 candidates)

Thank you for your attention !