TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM SCHOOL OF NATURAL SCIENCES


Ph.D. Thesis

# On non-parametric tests for discovery and limit setting in one and multiple dimensions

**Lolian Shtembari**

**2023**

TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM School of Natural Sciences

# On non-parametric tests for discovery and limit setting in one and multiple dimensions

Lolian Shtembari

Vollständiger Abdruck der von der TUM School of Natural Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Sherry Suyu

Prüfer*innen der Dissertation:

1. Hon.-Prof. Dr. Allen C. Caldwell
2. Prof. Dr. Lukas Heinrich

Die Dissertation wurde am 03.05.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Natural Sciences am 29.06.2023 angenommen.

# Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Dr. Allen Caldwell, for his guidance and mentorship throughout my research journey. I am truly grateful for the freedom he granted me in pursuing the problems that interested me most, for sharing his expertise and for his support and constant encouragement, which helped me overcome challenges and complete the work presented here.

I would like to thank Dr. Oliver Schulz for all the help he has given me during my PhD regarding programming and data-science related problems, but most importantly, for having introduced me to the Institute, where I had the pleasure of working in contact with exceptional scientists.

I am thankful for the opportunity to have collaborated with Dr. Philipp Eller, Dr. Nahuel Ferreiro Iachellini, Dr. Luca Pattavina and Dr. Heerak Banerjee. Their knowledge, experience, and technical skills have enriched my understanding of experimental and statistical techniques. The fruitful discussions I had with them and our joint research efforts have played a significant role in the successful realization of this work.

I would also like to extend my thanks to my colleagues from BAT, GERDA, MADMAX and ODSL teams. I relished the time spent working and having fun with you all.

I would like to thank Prof. Maisola for her cherished friendship and my teacher, Maestra Anna, for her affection and the enthusiastic care she has always had for me.

Lastly, I would like to express my profound gratitude to my family for their unwavering love, encouragement, and support throughout my academic journey.

Mami, Babi, Kristi, Nëna dhe Nono, faleminderit për gjithë ndihmën dhe mbështetjen tuaj.

# Abstract

The use of spacings between ordered real-valued numbers is very useful in many areas of science. In particular, either unnaturally small or large spacings can be a signal of an interesting effect. As particle physicists, we are interested in the appearance of the unexpected clustering of values, indicating the presence of a new process, or large gaps between the ordered values, allowing us to set upper limits on the normalization of a distribution. By analyzing the distribution of spacings between consecutive ordered data points, I develop sensitive test statistics, allowing for a quantitative measure of agreement between a model and observed data. The statistics developed in this thesis are used to perform unbinned non-parametric goodness of fit tests, without the need of trials factor or look-elsewhere correction, that can be used to detect an unknown signal against a known background or to set limits on a proposed signal distribution in experiments contaminated by poorly understood backgrounds. This thesis aims to provide a comprehensive understanding of the use of Order Statistics in physics, while also addressing the challenge of extending goodness-of-fit tests to multivariate samples, since out-of-the-box non-parametric tests that can target any proposed distribution are only available in the univariate case. My approach relies on a multivariate probability integral transformation of the data, that can be carried out analytically for simple models or numerically using a Normalizing Flow in the case of arbitrary complex multivariate distributions or multivariate data generation models. Once transformed, the problem is reduced to a multivariate uniformity test, which is simplified by either considering the independent marginal distributions of the data or by considering the volumes identified by each sample. These methods effectively reduce the complexity of one multivariate goodness-of-fit test to a single or multiple univariate ones.

# Kurzfassung

Die Verwendung von Abständen zwischen geordneten reellwertigen Zahlen ist in vielen Bereichen der Wissenschaft sehr nützlich. Hier können entweder unnatürlich kleine oder große Abstände ein Signal für einen interessanten Effekt sein. Als Teilchenphysiker interessieren wir uns für das Auftreten unerwarteter Wertecluster, die auf das Vorhandensein eines neuen Prozesses hindeuten, oder große Lücken zwischen den geordneten Werten, die es uns ermöglichen, Obergrenzen für die Normalisierung einer Verteilung festzulegen. Durch die Analyse der Verteilung von Abständen zwischen aufeinanderfolgenden geordneten Datenpunkten entwickle ich empfindliche Teststatistiken, die eine quantitative Maßnahme für die Übereinstimmung zwischen einem Modell und beobachteten Daten ermöglichen. Die in dieser Arbeit entwickelten Statistiken werden verwendet, um ungebinnte nicht-parametrische Goodness-of-Fit-Tests ohne die Notwendigkeit von Trial-Faktoren oder Look-Elsewhere-Korrekturen durchzuführen. Diese Tests können verwendet werden, um ein unbekanntes Signal gegen einen bekannten Hintergrund zu erkennen oder Grenzen für eine vorgeschlagene Signalverteilung in Experimenten zu setzen, die durch schlecht verstandene Hintergründe kontaminiert sind. Diese Arbeit zielt darauf ab, ein umfassendes Verständnis der Verwendung von Ordnungsstatistiken in der Physik zu vermitteln und gleichzeitig die Herausforderung der Erweiterung von Goodness-of-Fit-Tests auf multivariate Stichproben anzugehen, da Out-of-the-Box nicht-parametrische Tests, die auf jede vorgeschlagene Verteilung abzielen können, nur im univariaten Fall verfügbar sind. Mein Ansatz basiert auf einer multivariaten Wahrscheinlichkeitsintegraltransformation der Daten, die für einfache Modelle analytisch durchgeführt oder numerisch mit einem Normalizing Flow für beliebig komplexe multivariate Verteilungen oder multivariate Datengenerierungsmodelle durchgeführt werden kann. Nach erfolgter Transformation wird das Problem auf einen multivariaten Uniformitätstest reduziert, der entweder durch Betrachtung der unabhängigen marginalen Verteilungen der Daten oder durch Betrachtung der Volumina, die von jeder Stichprobe identifiziert werden, vereinfacht wird. Diese Methoden reduzieren die Komplexität eines multivariaten Goodness-of-Fit-Tests effektiv auf einen oder mehrere univariate Tests.

# Contents

# 1. Motivation and Overview

Physics, at its core, is the relentless study of the universe, a quest towards understanding its fundamental nature and the intricate web of interactions that govern the behaviour of matter and energy. As our knowledge of physics has advanced, so too have the experimental techniques and theoretical frameworks used to probe the world surrounding us. In this pursuit, the analysis of experimental data plays a pivotal role, providing a critical link between the predictions of theoretical models and the observations made in the laboratory.

One essential aspect of data analysis in physics is the ability to test the goodness of fit between a theoretical model and experimental data. The outcome of such tests can be helpful in determining whether a particular model is capable of explaining the observed phenomena, if there is a need to refine the model or if it is necessary to search for alternative explanations. Additionally, goodness of fit tests are often fast, allowing them to be employed to filter a large number of datasets and select only the most promising candidates for further analysis, such as a more detailed but also slower and more costly Bayesian analysis of the data.

In this thesis, I explore unbinned non-parametric goodness of fit tests for discovery and limit setting. The use of non-parametric tests can offer several advantages over that of parametric counterparts. Unlike parametric tests, which rely on parameters that are tuned to the observations, such as a choice of binning or a specific kernel function, and whose value can affect the outcome of the test, non-parametric tests need no such tunable parameters but simply a null-hypothesis, i.e. the assumed distribution of the observed data. This allows for more robust and versatile tests, particularly in cases where the distribution of the data is unknown or poorly understood. Moreover, the tests developed here do not need any trials factor or look-elsewhere correction since the data can be analyzed all at once.

The non-parametric tests considered in this thesis are based on Order Statistics, particularly spacings or gaps. Order Statistics refers to the arrangement of a set of data points in ascending order and by considering the spacings or gaps between consecutive ordered data, it is possible to construct sensitive test statistics that provide a quantitative measure of the agreement between a given model and the observed data. In particular, either unnaturally small or large spacings can be a signal of an interesting effect. These test statistics can be used to perform goodness of fit tests either for signal discovery or to set limits on signals affected by poorly understood backgrounds, thereby facilitating the identification of new physical phenomena and the validation of theoretical predictions.

Order Statistics have been studied in great depth in the statistics community, but unfortunately, the work is poorly known in the physics community, which has led to the rediscovery of results long-known to statisticians. An example of the use of spacings between values in the particle physics community is presented by Yellin [1], where the author proposes a method to set a limit on the interaction rate of putative dark matter particles using the size of gaps in the observed energy spectrum of recorded interactions. In this context, a large gap in the energy spectrum implies an upper limit on the interaction strength. Conversely, in a discovery scenario, an abundance of small spacings located close to one another can point to the presence of a signal on top of the assumed background.

The development and application of such tests to the analysis of univariate datasets is often aided by the probability integral transformation [2, 3], which allows transforming the data at hand into a standardized space, the unit interval, $[0, 1]$, using the cumulative distribution of the provided model. Given this, the test statistics are developed in the unit interval and at their core are uniformity tests.

While the univariate case is of substantial importance, the extension of goodness of fit tests to multivariate samples is an open challenge. Multivariate datasets are increasingly common in modern physics experiments, as advances in detector technology and data acquisition systems enable the simultaneous measurement of multiple observables. The ability to perform goodness of fit tests on multivariate samples would boost the effectiveness of the analysis and interpretation of these data. To address this challenge, I introduce two methods for performing unbinned goodness of fit tests for multivariate samples. The applicability of these methods relies on a multivariate probability integral transformation of the data.

The difficulty of performing this transformation depends on the complexity of the proposed model. While it may be easy to perform the transformation in the case of multivariate distributions with uncorrelated dimensions, it might not be easy to find an exact transformation if correlations are present. When dealing with more complex models, or when only a generative model is available, but no proper definition of a multivariate probability distribution, it might be possible to numerically perform the multivariate probability integral transformation by employing a Normalizing Flow.

Normalizing Flows are a powerful class of machine learning techniques that enable the transformation of complex, high-dimensional probability distributions into simpler, more tractable forms. By leveraging these techniques, it is possible to extend the applicability of univariate goodness of fit tests to a broader range of multivariate problems, including those involving correlated dimensions and intricate statistical structures.

In this thesis, Chapter 2 provides a general introduction to the topic of Order Statistics, focusing mainly on the distribution theory of spacings in the case of an underlying Standard Uniform ($\mathcal{U}(0, 1)$) distribution of samples. Chapter 3 focuses mainly on the description of goodness of fit tests based on spacings aimed at discovery applications: here I introduce three new test statistics and then perform a general comparison of the

performance of these tests using simulated experiments. Chapter 4 concerns test statistics aimed at setting limits on signals affected by poorly understood background: here I discuss the state-of-the-art test statistics present in the literature and introduce two new tests, comparing the performance of all available tests against various background distributions. Chapter 5 describes the extension to multivariate goodness of fit tests, introducing the "*projection*" and "*volume transformation*" methods and presenting examples of their utilization to both discovery and limit setting applications.

The comparisons between tests presented in Chapters 3 to 5 not only serve to validate the effectiveness of the methods in a practical context but also provide valuable insights into their relative performance in different scenarios, which can be referenced as guidelines in the choice of the most sensitive tests for different applications.

Finally, to demonstrate the power and versatility of the newly introduced test statistics, in Chapter 6 I show applications to physics analyses inspired by a proposed experiment (RES-NOVA) to observe Supernovae by detecting neutrino flares and by reanalysing the CRESST dark matter search data using the newly developed tests, hoping these examples encourage their adoption in the broader physics community and stimulate further research in this area. A mock "bump-hunting" example, such as the search for an exotic particle decay with unknown mass, is also considered and a comparison of the sensitivity of the newly introduced tests against a likelihood ratio approach in this scenario is presented. Lastly, Chapter 7 contains a brief summary of my contributions and a discussion of the future research outlooks building upon the results presented here.

# 2. Order Statistics

## 2.1. Introduction

Given a sequence of $n$ random variables $\{X_1, X_2, ..., X_n\}$, we construct their order statistics by sorting them in ascending order:

$$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)} \tag{2.1}$$

where $X_{(i)}$ refers to the $i$-th order statistic, i.e. the $i$-th smallest sample.

In this thesis, we will assume the unordered samples $X_i$ to be statistically independent and identically distributed (i.i.d.). Still, even with this assumption, we recognize that the ordered samples $X_{(i)}$ are necessarily dependent due to the inequality relations between them.

The field of Order statistics explores the properties and applications of these ordered random variables, as well as functions involving them.

Order statistics are used in characterizations and ***goodness-of-fit tests***, which have long-ranging applications in various scientific fields: model validation, signal detection, background-rejection and limit setting. Additionally, most goodness-of-fit tests for arbitrary parent distributions, implicitly involve order statistics, since they often focus on deviations between the empirical quantile function and the hypothesized one. We will focus on these in later chapters, reviewing existing tests and discussing my newly proposed ones, adding to the large body of literature that has been devoted to the study of order statistics.

For a more comprehensive and detailed review of the results that have been achieved in the field of order statistics, I recommend the following books [4, 5] which at times I closely follow in delivering this brief introduction.

## 2.2. Distribution of Order Statistics

In the following, I review the main results of the distribution theory of order statistics. I begin investigating single order statistics and then derive the joint distribution of a set of them. Finally, I review a general result on the conditional distribution of order statistics which highlights how they can be interpreted as a Markov chain.

### 2.2.1. Basic distribution theory

Assuming a sequence $\{X_1, X_2, ...X_n\}$ of $n$ i.i.d. variables with cumulative density function (cdf) $F(x) = \Pr\{X \leq x\}$, it is possible to derive the distribution of any order statistic. Starting with the largest order statistic, $X_{(n)}$, its cumulative distribution is simply:

$$\begin{aligned} F_{(n)}(x) &= \Pr\{X_{(n)} \leq x\} \\ &= \Pr\{\text{all } X_i \leq x\} \\ &= [F(x)]^n \end{aligned} \tag{2.2}$$

Similarly, the cumulative distribution of the smallest order statistics, $X_{(1)}$, is:

$$\begin{aligned} F_{(1)}(x) &= \Pr\{X_{(1)} \leq x\} \\ &= 1 - \Pr\{X_{(1)} > x\} \\ &= 1 - \Pr\{\text{all } X_i > x\} \\ &= 1 - [1 - F(x)]^n \end{aligned} \tag{2.3}$$

Generally, the cumulative distribution of the $k$-th order statistic, $X_{(k)}$, is:

$$\begin{aligned} F_{(k)}(x) &= \Pr\{X_{(k)} \leq x\} \\ &= \Pr\{\text{at least } k \text{ of } X_i \text{ are at most equal to } x\} \\ &= \sum_{j=k}^{n} \Pr\{\text{exactly } j \text{ of } X_i \text{ are at most equal to } x\} \\ &= \sum_{j=k}^{n} \binom{n}{j} [F(x)]^j \, [1 - F(x)]^{(n-j)} \end{aligned} \tag{2.4}$$

Differentiating the cumulative distributions above yields the probability density function (pdf) $f_{(k)}$ of any $X_{(k)}$:

$$f_{(1)}(x) = n \cdot f(x) \cdot [1 - F(x)]^{(n-1)} \tag{2.5}$$

$$f_{(n)}(x) = n \cdot f(x) \cdot [F(x)]^{(n-1)} \tag{2.6}$$

$$f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!} \cdot f(x) \cdot [F(x)]^{(k-1)} \, [1 - F(x)]^{(n-k)} \, . \tag{2.7}$$

Given these results, it is now possible to consider the joint distribution of a set of order statistics. If we consider two order statistics $X_{(j)}$ and $X_{(k)}$, with $1 \leq j < k \leq n$, then

their joint distribution $f_{(j)(k)}(x,y)$ can be derived by considering that $j-1$ observations are at most equal to $x$, one is exactly $x$, another $k-j-1$ are at least $x$ and at most $y$, one is exactly $y$ and the remaining $n-k$ observations are at least equal to $y$. This configuration is depicted in Fig. 2.1, and using the distributions derived above, we can translate our interpretation of this configuration into the formula of $f_{(j)(k)}(x,y)$:

$$f_{(j)(k)}(x,y) = \frac{n!\,[F(x)]^{(j-1)}\,f(x)\,[F(y)-F(x)]^{(k-j-1)}\,f(y)\,[1-F(y)]^{(n-k)}}{(j-1)!(k-j-1)!(n-k)!} \tag{2.8}$$

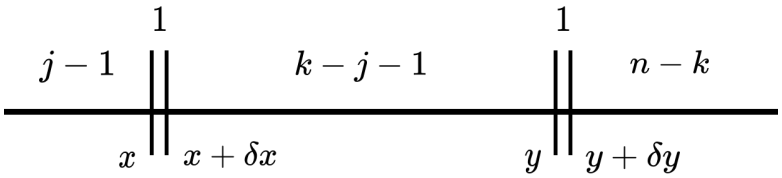



Figure 2.1.: Illustration of the position of order statistics relative to $X_{(j)}$ and $X_{(k)}$.

Based on the above-mentioned interpretation of $f_{(j)(k)}(x,y)$, one can easily generalize and obtain the joint distributions of any combination of order statistics $\{X_{(n_1)},...,X_{(n_k)}\}$ $(1 \le n_1 < ... < n_k \le n)$:

$$\begin{aligned} f_{(n_1)...(n_k)}(x_1,...,x_k) =& \frac{n!}{(n_1-1)!(n_2-n_1-1)!...(n-n_k)!}\cdot \\ & \cdot [F(x_1)]^{(n_1-1)}\,f(x_1)\,[F(x_2)-F(x_1)]^{(n_2-n_1-1)}\,\cdot \\ & \cdot f(x_2)...\,[1-F(x_k)]^{(n-n_k)} \\ =& n!\left[\prod_{j=1}^{k} f(x_j)\right]\cdot\left[\prod_{j=0}^{k}\frac{[F(x_{j+1})-F(x_j)]^{(n_{j+1}-n_j-1)}}{(n_{j+1}-n_j-1)!}\right] \end{aligned} \tag{2.9}$$

where $x_0 = -\infty$, $x_{k+1} = +\infty$, $n_0 = 0$ and $n_{k+1} = n+1$.

If one considers all order statistics at once, their joint distribution is simply the number of equally likely orderings of $\{X_1,...,X_n\}$, the $n!$ permutations of samples, times the likelihood of a given realization:

$$f_{(1)...(n)}(x_1,...,x_n) = n!\prod_{j=1}^{n} f(x_j). \tag{2.10}$$

Integrating Eq. 2.8 yields the cumulative distributions, which can also be obtained directly with similar considerations to Eq. 2.4:

$$
\begin{aligned}
F_{(j)(k)}(x,y) &= \Pr\{\text{at least } j \ X_i \leq x \text{ and at least } k \ X_i \leq y\} \\
&= \sum_{t=k}^{n} \sum_{r=j}^{t} \Pr\{\text{exactly } r \ X_i \leq x \text{ and exactly } t \ X_i \leq y\} \\
&= \sum_{t=k}^{n} \sum_{r=j}^{t} \frac{n! \left[F(x)\right]^r \left[F(y) - F(x)\right]^{(t-r)} \left[1 - F(y)\right]^{(n-t)}}{r!(t-r)!(n-t)!}
\end{aligned} \tag{2.11}
$$

Given the joint distribution of $k$ order statistics, it is possible to derive the distribution of any function of order statistics, using variable transformation methods. For example, considering the gap between any two given order statistics, $W_{jk} = X_{(k)} - X_{(j)}$, we can express its value as $w_{jk} = y - x$ and substitute $(x, y) \rightarrow (x, w_{jk})$ into Eq. 2.8 and marginalize over the variable $x$, in order to obtain the distribution of $w_{jk}$:

$$
\begin{aligned}
f_{W_{jk}}(w_{jk}) = &\frac{n!}{(j-1)!(k-j-1)!(n-k)!} \cdot \int_{-\infty}^{+\infty} F^{j-1}(x) f(x) \cdot \\
&\cdot \left[F(x + w_{jk}) - F(x)\right]^{k-j-1} f(x + w_{jk}) \left[1 - F(x + w_{jk})\right]^{n-k} dx
\end{aligned} \tag{2.12}
$$

Such an approach, in principle, is possible for any function of order statistics $h(X_{(i)})$, but depending on the complexity of $h$ and of the original probability distribution $f$, the resulting integral might not be solvable in closed form. Exact results have been derived for well-known distributions of events, such as the Uniform and Exponential distributions and for "well-behaved" combinations of order statistics, such as the "gaps" mentioned above, which are extremely important in the work of this thesis.

### 2.2.2. Order Statistics as a Markov Chain

Given the joint probability distribution of a combination of order statistics, Eq. 2.9, the joint conditional distribution of $\{X_{(j+1)}, ..., X_{(k-1)}\}$ given $X_{(i)}$ for $i \leq j \vee i \geq k$ is:

$$
\begin{aligned}
&f_{X_{(j+1)} \cdots X_{(k-1)} | X_{(i)} = x_i, i \leq j \vee i \geq k}(x_{j+1}, ..., x_{k-1}) = \\
&\qquad = (k-j-1)! \prod_{i=j+1}^{k-1} \frac{f(x_i)}{F(x_k) - F(x_j)}
\end{aligned} \tag{2.13}
$$

for $x_1 < ... < x_n$. This result means that the conditional distribution of $\{X_{(j+1)}, ..., X_{(k-1)}\}$ is just the distribution of all order statistics in a sample of $k - j - 1$ drawn from the distribution $\frac{f(x)}{[F(x_k) - F(x_j)]}$, i.e. $f(x)$ truncated to the interval $[x_j, x_k]$. Additionally, we notice that this conditional distribution is free of $x_i$ for $i < j \vee i > k$, meaning that

$\{X_{(j+1)}, ..., X_{(k-1)}\}$ is independent of $\{X_{(1)}, ..., X_{(j-1)}, X_{(k-1)}, ..., X_{(n)}\}$ when $X_{(j)}$ and $X_{(k)}$ are given. Conditioning on the lower order statistic Eq. 2.13 leads to:

$$\begin{aligned} &f_{X_{(j+1)}...X_{(n)}|X_{(1)}=x_1,...,X_{(j)}=x_j}(x_{j+1}, ..., x_n) = \\ &= f_{X_{(j+1)}...X_{(n)}|X_{(j)}=x_j}(x_{j+1}, ..., x_n) \end{aligned} \tag{2.14}$$

which shows that the order statistics in a sample from a continuous density $f(x)$ form a Markov chain. The transition density is given by:

$$f_{X_{(j+1)}|X_{(j)}=x}(y) = (n-j)\frac{f(y)}{1-Fx}\left[\frac{1-F(y)}{1-F(x)}\right]^{n-j-1}, \; y > x. \tag{2.15}$$

## 2.3. Uniform Order Statistics

So far, we have considered order statistics pertaining to arbitrary continuous distributions $f(x)$ with cumulative $F(x)$. The case of standard uniformly distributed samples is particularly important, since it allows a simple derivation of many important properties of order statistics.

### 2.3.1. Probability Integral Transformation

Given an arbitrary cumulative distribution $F(x)$, i.e. a non-decreasing and right continuous function with $F(-\infty) = 0$ and $F(\infty) = 1$, its associated *inverse distributions function*, or *quantile function*, is defined by:

$$F^{-1}(u) = \sup\{x : F(x) \leq u\} \tag{2.16}$$

If $U$ is a standard Uniform random variable, $U \sim \mathcal{U}(0, 1)$, then $F^{-1}(U)$ has distribution function $F$:

$$\Pr\{F^{-1}(U) \leq x\} = \Pr\{U \leq F(x)\} = F(x) \tag{2.17}$$

If we consider $n$ i.i.d. standard Uniform random variables $U_i$ and $n$ i.i.d. random variables $X_i$ with distribution $F(x)$, then:

$$\left(X_{(1)}, ..., X_{(n)}\right) \stackrel{d}{=} \left(F^{-1}(U_{(1)}), ..., F^{-1}(U_{(n)})\right) \tag{2.18}$$

$$\left(U_{(1)}, ..., U_{(n)}\right) \stackrel{d}{=} \left(F(X_{(1)}), ..., F(X_{(n)})\right) \tag{2.19}$$

where $\stackrel{d}{=}$ means they are identically distributed.

### 2.3.2. Distribution of Uniform order statistics

In the absolutely continuous case, the relationship highlighted by Eq. 2.18 points out that it is possible to derive the joint density of Eq 2.10 from the much simpler joint probability density of uniform order statistics:

$$f_{(1)\ldots(n)}(u_1, ..., u_n) = n! \tag{2.20}$$

Working with uniform order statistics allows simple derivations of moments and other distributional features of order statistics, which can be translated into the original space by a variable transformation. Additionally, this allows the study of order statistics in a standardized space, defined by the unit interval $[0, 1]$, which allows to formulate constraints that are extremely useful in deriving distributions and asymptotic results. One such constraint is the sum of all consecutive spacings arising from a random sample of uniform order statistics always equals 1. In the following chapter, we will refer to this constraint in our derivations. Furthermore, transforming random variables into uniform ones by means of their cumulative distribution, $U = F(X)$, is referred to as the *probability integral transformation* [2, 3] and is extremely important in hypothesis testing, as will be discussed in further chapters.

Considering $n$ i.i.d. Uniform random variables $U_i$, Eq. 2.5 becomes:

$$f_{U_{(k)}} = \frac{n!}{(k-1)!(n-k)!} u^{k-1}(1-u)^{n-k} \tag{2.21}$$

meaning that the $k$-th order statistic of the uniform distribution is distributed according to a Beta distribution:

$$U_{(k)} \sim \text{Beta}(k, n+1-k). \tag{2.22}$$

The joint distributions of $(U_{(j)}, U_{(k)})$, for $j < k$, becomes:

$$f_{U_{(j)},U_{(k)}}(u, v) = n! \frac{u^{j-1}(v-u)^{k-j-1}(1-v)^{n-k}}{(j-1)!(k-j-1)!(n-k)!} \tag{2.23}$$

and using this to find the distribution of the distance between two uniform order statistics, one finds that it follows a Beta distribution as well:

$$U_{(k)} - U_{(j)} \sim \text{Beta}(k-j, n+1-(k-j)) \tag{2.24}$$

which depends only on the difference $(k-j)$, but not on $k$ or $j$ individually.

## 2.4. Spacings

### 2.4.1. Uniform Spacings

Given a set of $n$ i.i.d. standard Uniform random variables, $\{U_1, U_2, ..., U_n\}$, I am interested in the distance between the corresponding order statistics, $U_{(i)}$, which are referred to as ***spacings***. Begin by considering an extended set of ordered values, namely the boundaries of the range of $U_{(i)}$ themselves: defining $U_{(0)} = 0$ and $U_{(n+1)} = 1$.

The general spacing of rank $k$, $S_{i,k}$, are defined as:

$$S_{i,k} = U_{(i)} - U_{(i-k)} \tag{2.25}$$

for $k \leq i \leq n+1$. These spacings are Beta random variables, since their distribution is given by Eq. 2.24, where the difference between uniform order statistics depends only on the difference of their indices:

$$S_{i,k} \sim \text{Beta}(k, n+1-k). \tag{2.26}$$

In this thesis, I refer to the spacing between consecutive order statistics as *simple spacings*, i.e. first rank spacings $S_{i,1}$, and for convenience, I simplify their notation: $S_i \coloneqq S_{i,1}$. Given $n$ samples, one can define $n+1$ (uniform) simple spacings.

One important constraint when it comes to simple spacings is their sum:

$$\sum_{i=1}^{n+1} S_i = 1 \tag{2.27}$$

The cumulative distribution of a simple spacing is (from Eq. 2.26):

$$F_{S_i}(x) = F_{S_1}(x) = F_{U_{(1)}}(x) = 1 - (1-x)^n \tag{2.28}$$

while the joint probability of any two simple spacings is:

$$\begin{aligned} F_{S_i,S_j}(x,y) &= F_{S_1,S_2}(x,y) = \Pr\{S_1 \leq x, S_2 \leq y\} \\ &= \Pr\{U_{(1)} \leq x, U_{(2)} - u_{(1)} \leq y\} \\ &= n \int_0^x \left[1 - \left(1 - \frac{y}{1-u}\right)^{n-1}\right] (1-u)^{n-1} du \\ &= 1 - \left[(1-x)^n + (1-y)^n - (1-x-y)^n\right]. \end{aligned} \tag{2.29}$$

Given Eq.s 2.28 and 2.29, one can deduce that the joint cumulative distribution of $k$ simple spacings is:

$$F_{S_1,...,S_k}(x_1,...,x_k) = 1 - \left[\left(\sum_{i=1}^{k}(1-x_i)^n\right) - \left(1-\sum_{i=1}^{k}x_i\right)^n\right] \tag{2.30}$$

and their joint probability density is:

$$f_{S_1,...,S_k}(x_1,...,x_k) = \frac{n!}{(n-k)!}\left(1-\sum_{i=1}^{k}x_i\right)^{n-k} \tag{2.31}$$

which can be proven by induction. For general spacings, for samples $X_i$ with arbitrary distribution $F(x)$, these distributions can be obtained either by using directly the joint distribution of the first $k$ order statistics, Eq. 2.9, or by considering that the order statistics can be expressed as a Markov chain, as discussed in Sec. 2.2.2. An example of such a derivation can be found in Pyke's review of tests based on spacings [6].

### 2.4.2. Exponential Spacings

Considering $Z_{(1)} \leq ... \leq Z_{(n)}$, denoting the order statistics of $n$ samples from an exponential distribution with rate $\lambda$ , $f(z) = \lambda e^{-\lambda z}$, then their joint distribution is:

$$f_{Z_{(1)},...,Z_{(n)}}(z_{(1)},...,z_{(n)}) = n!\lambda^n \exp\left(-\lambda\sum_{i=1}^{n}z_{(i)}\right) \tag{2.32}$$

which can be rewritten as:

$$f_{Z_{(1)},...,Z_{(n)}}(z_{(1)},...,z_{(n)}) = n!\lambda^n \exp\left(-\lambda\sum_{i=1}^{n}(n+1-i)(z_{(i)}-z_{(i-1)})\right) \tag{2.33}$$

where $z_{(0)} = 0$. Defining the exponential (simple) spacings as:

$$D_i = Z_{(i)} - Z_{(i-1)} \tag{2.34}$$

for $1 \leq i \leq n$, we can substitue them in Eq. 2.33 and obtain the joint distribution of all $D_i$ [7]:

$$\begin{aligned} f_{D_1,...,D_n}(d_1,...,d_n) &= n!\lambda^n \exp\left(-\lambda\sum_{i=1}^{n}(n+1-i)d_i\right) \\ &= \prod_{i=1}^{n}\lambda(n+1-i)\exp\left[-\lambda(n+1-i)d_i\right] \end{aligned} \tag{2.35}$$

which shows that the joint density function of the exponential spacings is the product of $n$ marginal exponential densities. This means we can interpret $D_i$ as an independent Exponential random variable with rate $\lambda(n+1-i)$. Considering the *normalized spacings*:

$$Y_i = \lambda(n+1-i)D_i \tag{2.36}$$

we notice that the samples $Y_i$ are statistically independent variates, distributed according to a standard Exponential distribution.

As noted in [6], due to the independence of exponential spacings, the exact distribution theory for functions of $D_i$ is relatively simple, and classical limit theorems for independent random variables may be applied to obtain the limiting distribution of functions of exponential spacings.

Finally, the transformation 2.36 allows $Z_{(j)}$ to be expressed as a linear function of exponential i.i.d.:

$$Z_{(j)} = \sum_{i=1}^{j} D_i = \sum_{i=1}^{j} \frac{Y_i}{\lambda(n+1-i)}. \tag{2.37}$$

From this follows that $\{Z_{(1)}, ..., Z_{(n)}\}$ form an additive Markov chain [8].

## 2.5. Construction of spacings

In the following, I list several ways of constructing Uniform spacings, closely following the corresponding chapter from Pyke's review [6].

### 2.5.1. Uniform spacings as Exponential r.v.'s

Given $n+1$ independent standard Exponential random variables $\{Y_1, ..., Y_{n+1}\}$, let $T$ be their sum:

$$T = \sum_{i=1}^{n+1} Y_i \tag{2.38}$$

and $D_i = Y_i/T$, then:

$$f_{(Y_1,...,Y_n,T)}(y_1, ..., y_n, t) = e^{-t} \tag{2.39}$$

from which:

$$f_{(D_1,...,D_n,T)}(d_1, ..., d_n, t) = t^n e^{-t} \tag{2.40}$$

for $d_i \geq 0$, $0 \leq \sum_{i=1}^{n} d_1 \leq 1$ and $t > 0$. Marginalizing over the sum $T$, we get:

$$f_{(D_1,...,D_n)}(d_1, ..., d_n) = n! \tag{2.41}$$

which is the distribution of the first $n$ simple Uniform spacings (Eq. 2.31), thus, $\{D_1, ..., D_{n+1}\}$ are distributed as the set of $n+1$ spacings determined by $n$ independent standard Uniform random variables:

$$\left(D_{(1)}, ..., D_{(n+1)}\right) \stackrel{d}{=} (S_1, ..., S_{n+1}) \tag{2.42}$$

An alternative route would have been to consider the conditional distribution:

$$f_{(Y_1,...,Y_n)|T}(y_1, ..., y_n|t) = n!t^{-n} \tag{2.43}$$

which given $T = 1$ would yield the same result.

### 2.5.2. Uniform spacings as Inter-event times in a Poisson process

Let $N(t)$ for $t \geq 0$ be a Poisson process with parameter $\lambda = E[N(1)]$, and let $T_j$, with $T_i \leq T_j$ for $i < j$, denote the successive times of events in the process. For a given $t \geq 0$, set:

$$Y_i = T_1 - T_{i-1} \tag{2.44}$$

for $1 \leq i \leq N(t)$, with $T_0 = 0$, and set $Y_{N(t)+1} = t - T_{N(t)}$.

Given $N(t) = n$, then the distribution of $\{Y_1/t, ..., Y_{n+1}/t\}$ is the same as $n+1$ Uniform spacings:

$$\left(\frac{Y_1}{t}, ..., \frac{Y_{n+1}}{t}\right) \stackrel{d}{=} (S_1, ..., S_{n+1})\,. \tag{2.45}$$

This is possibly the oldest construction of Uniform spacings and one which represents the most natural relationship between the Poisson process ("random" points on a line) and the Uniform distribution ("random" points on an interval).

### 2.5.3. Uniform spacings as Beta r.v.'s

Given $n$ i.i.d. Uniform random variables, we can produce a set of $n+1$ Uniform spacings $0 \leq S_i \leq 1$ subject to the constraint:

$$\sum_{i=1}^{n+1} S_i = 1.$$

Given this constraint, it is possible to interpret the set of Uniform spacings as a realization of a Dirichlet random variable with all concentration parameters equal to 1: $\alpha_i = 1$ for $1 \leq i \leq n+1$.

Generally, given an ensemble of $n+1$ Gamma variates $Y_i$:

$$Y_i \sim \Gamma(\alpha_i, \beta_i) \tag{2.46}$$

is has been shown that upon normalization they follow a Dirichlet distribution [9]:

$$\left(\frac{Y_i}{\sum_{j=1}^{n+1} Y_j}\right) = (D_i) \sim \mathrm{Dir}(\alpha_1, ..., \alpha_{n+1}) = \frac{\Gamma\left(\sum_{i=1}^{n+1} \alpha_i\right)}{\prod_{i=1}^{n+1} \Gamma(\alpha_i)} \prod_{i=1}^{n+1} d_i^{\alpha_1 - 1} \tag{2.47}$$

This result is not particularly surprising, because in order to construct a set of Uniform spacings, one needs to fix $\alpha_i = 1$, as stated above, in which case the Gamma distribution is equal to the Exponential one, $\Gamma(1, \lambda) = \mathrm{Exp}(\lambda)$, and $Y_i$ are nothing more than exponential random variables, which are distributed as Uniform spacings upon normalization, as shown previously.

Focusing on the description of Uniform spacings as a Dirichlet sample, I show how to relate it to a set of independent Beta variables. This transformation was introduced by Betancourt [10] when dealing with the problem of sampling directly from a Dirichlet distribution in the context of Markov Chain Monte Carlo (MCMC) techniques such as Hamiltonian Monte Carlo.

Given a set of $n+1$ Uniform spacings $\{S_i\}$, consider the transformation:

$$Y_i = \sqrt{S_i} \tag{2.48}$$

with support and distribution given by:

$$\begin{gathered} 0 \leq Y_i \leq 1 \\ \sum_{i=1}^{n+1} Y_i^2 = 1 \\ f_{\{Y_i\}}(\{y_i\}) = 2^{n+1} n! \prod_{i=1}^{n+1} y_i \end{gathered} \tag{2.49}$$

Given the quadratic constraint, instead of dealing with samples on a hyperplane (i.e. the simplex where Dirichlet variables reside), the samples $Y_i$ reside on the surface of a $n+1$ dimensional hypersphere, which can be parametrised by transforming to hyperspherical coordinates [11]:

$$Y_i = r\left(\prod_{k=1}^{i-1}\sin\Theta_k\right)\cdot\begin{cases}\cos\Theta_i, & i<n+1\\ 1, & i=n+1\end{cases}. \tag{2.50}$$

The support and distribution of this transformation is:

$$0\le\theta\le\frac{\pi}{2}$$

$$r^2=1$$

$$f_{(r,\{\Theta_i\})}(r,\{\theta_i\}) = 2^{n+1}n!r^{2n+1}\prod_{i=1}^{n}\cos\theta_i\left(\sin\theta_i\right)^{2(n+1-i)-1} \tag{2.51}$$

which upon marginalization over the radial component becomes:

$$f_{\{\Theta_i\}}(\{\theta_i\}) = 2^n n!\prod_{i=1}^{n}\cos\theta_i\left(\sin\theta_i\right)^{2(n+1-i)-1}. \tag{2.52}$$

Finally, consider the last transformation:

$$B_i = \sin^2\Theta_i \tag{2.53}$$

with $0\le B_i\le 1$, whose joint distribution is:

$$f_{\{B_i\}}(\{b_i\}) = \prod_{i=1}^{n}(n+1-i)b_i^{n-i} = \prod_{i=1}^{n}\mathrm{Beta}(n+1-i,1) \tag{2.54}$$

which can be factored as the product of $n$ independent Beta random variables.

The mapping $S_i\to Y_i\to\Theta\to B_i$ reduces the original Uniform spacings, distributed as a Dirichlet random variable, to a simple product of independent Beta distributions. The inverse transformation to construct Uniform spacings starting from Beta variates $B_i$ is:

$$S_i = \left(\prod_{k=1}^{i-1}B_k\right)\cdot\begin{cases}1-B_i, & i<n+1\\ 1, & i=n+1\end{cases}. \tag{2.55}$$

## 2.6. Basic Asymptotic Theory

### 2.6.1. Order Statistics

So far we focused on the exact distribution theory for order statistics, where we saw that the exact cumulative distributions is often computationally messy except for some very special cases. For large sample sizes, i.e. $n\gg 1$, then it might be beneficial to investigate

the asymptotic behaviour in hope of finding easier approximations to the distributions we seek.

Representing order statistics as $X_{(i)} \stackrel{d}{=} F^{-1}(U_{(i)})$ proves to be useful also in deriving their asymptotic distributions. Given $X_{(\lceil np \rceil)}$, where $p \in [0,1]$ and $n \gg 1$, consider the corresponding uniform order statistic $U_{(\lceil np \rceil)}$ whose distributions is that of a Beta random variable, Eq. 2.22:

$$U_{(\lceil pn \rceil)} \sim \text{Beta}\left(np, n(1-p)\right). \tag{2.56}$$

Such a Beta variable can also be represented as the combination of independent Gamma variates [5], i.e. has the same distribution as:

$$\sum_{i=1}^{n} \frac{V_i}{\sum_{i=1}^{n} V_i \sum_{i=1}^{n} W_i} \tag{2.57}$$

where the variables $V_i$ are i.i.d. from $\Gamma(p,1)$ and the variables $W_i$ are also i.i.d. but from $\Gamma(1-p,1)$ and independent from $V_i$. By using the multivariate central limit theorem and delta method, one can prove that the asymptotic distribution of $U_{(\lceil np \rceil)}$ is Normal [12]:

$$U_{(\lceil np \rceil)} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right). \tag{2.58}$$

From this result, one can derive the asymptotic distribution of the original $X_{(\lceil np \rceil)}$:

$$X_{(\lceil np \rceil)} \sim \mathcal{N}\left(F^{-1}(p), \frac{p(1-p)}{n\left[f(F^{-1}(p))\right]^2}\right). \tag{2.59}$$

For a given value of $p$, $X_{(\lceil np \rceil)}$ does not represent an extreme order statistic, since the ratio $np/n$ is constant as $n$ increases. Extreme order statistics such as $X_{(1)}$ or $X_{(n)}$ have different, non-normal, limiting distributions, as first guessed by Tippett [13] and later proven and derived by Fisher, Tippet [14] and Gnedenko [15]. Considering $X_{(n)}$, its distributions will depend only on the upper tail of $F(x)$. For example, if $\bar{F}(x) = 1 - F(x) \sim cx^{-\alpha}$ for $x \to \infty$, then:

$$\begin{aligned}
\Pr\left\{b_n X_{(n)} \leq x\right\} &= \Pr\left\{X_{(n)} \leq \frac{x}{b_n}\right\} \\
&= \left[F\left(\frac{x}{b_n}\right)\right]^n \\
&= \left[1 - \bar{F}\left(\frac{x}{b_n}\right)\right]^n \\
&= \left[1 - \left(\frac{x}{b_n}\right)^{-\alpha}\right]^n
\end{aligned} \tag{2.60}$$

and if one chooses $b_n = n^{-1/\alpha}$, then $\Pr\left\{b_n X_{(n)} \leq x\right\} = [1 - x^{-\alpha}/n]^n$, which converges to $e^{-x^{-\alpha}}$, often called the extreme value distributions of the Frechét type.

This follows from a more general result, the Fisher–Tippett–Gnedenko theorem, also known as the Fisher–Tippett theorem or the extreme value theorem. This is a general result in extreme value theory concerning the asymptotic distribution of extreme order statistics. The maximum of a sample of i.i.d. random variables can only converge in distribution to one of 3 possible distributions: the Gumbel distribution, the Fréchet distribution, or the Weibull distribution.

### 2.6.2. Spacings

When it comes to spacings, we start considering Exponential spacings $D_i$, which themselves are independent Exponential random variables with rate $\lambda(n+1-i)$, Eq. 2.35. As the number of events increases, $(i, n) \to \infty$ with $i/n \to u$, $0 < u < 1$, then the variable $nD_i$ has distribution:

$$nD_i \sim \text{Exp}\left(\frac{\lambda(n+1-i)}{n}\right) \longrightarrow \text{Exp}(\lambda(1-u)). \tag{2.61}$$

Considering more spacings, letting $j/n \to v$, then we can write the joint cumulative distributions as:

$$\lim_{n\to\infty} F_{(nD_i, nD_j)}(x, y) = \left[1 - e^{-\lambda(1-u)x}\right]\left[1 - e^{-\lambda(1-v)y}\right]. \tag{2.62}$$

Although this is an obvious result for Exponential spacings, it is possible to generalize to spacings derived from arbitrary distributions with cdf $F(x)$ and pdf $f(x)$. For $0 < u, v < 1$, suppose $s = F^{-1}(u)$ and $t = F^{-1}(v)$ are uniquely defined, then if $i/n \to u$ and $j/n \to v$:

$$\lim_{n\to\infty} F_{(nD_i^*, nD_j^*)}(x, y) = \left[1 - e^{-f(s)x}\right]\left[1 - e^{-f(t)y}\right] \tag{2.63}$$

where $D_i^*$ refers to the general spacings obtained from $f(x)$. An outline of the proof of this result can be found in [6]. Additionally, any finite set of spacings retaians the asymptotic independence and Exponential distribution [16].

### 2.6.3. Functions of Uniform Spacings

The study of Order statistics and Spacings is important because using these components we are able to build statistical tools to investigate the fit between a set of data and a proposed model. I will discuss these tests in the next chapter, but for now it will suffice to say that they are built assuming a Uniform distribution of samples. This is due to the Probability Integral Transformation, which can be operated on any set of data provided a

candidate distribution, allowing to develop statisticla tools in a standardized environment where we can always transform into.

Some of the tests I discuss and use in my work can be expressed as a combination of functions of Uniform Spacings. For example, given a set of $n+1$ Uniform Spacings $\{S_i\}$, consider the test statistic:

$$\Omega_n = \sum_{i=1}^{n+1} g(nS_i) \tag{2.64}$$

where $g$ is an arbitrary real valued funciton.

Even when working with Uniform spacings, very rarely it is possible to derive the exact distribution of statistics $\Omega_n$, and even then, it might be difficult to write down suitable explicit expressions for the exact distribution.The first to provide a unified approach to the problem of finding the limiting distribution of a statistic such as $\Omega_n$ (Eq. 2.64) was Darling [17]. In his work, Darling targets the more general class of functions $g_i(nS_i)$, deriving a simple formula for the characteristic function of $\Omega_n$, from which it is possible to obtain general properties of spacings and exact moments of $\Omega_n$.

Subsequently, in 1958, LeCam [18] derived an easier general approach to finding the asymptotic distribution of $\Omega_n$. Consider a set of standard exponential random variables $Y_i$ and set:

$$K = \frac{1}{\sqrt{n}} \sum_{i=1}^{n+1} (Y_i - 1) \tag{2.65}$$

$$J = \sum_{i=1}^{n+1} g(Y_i). \tag{2.66}$$

From the construction of Uniform spacings as Exponential random variables, it follows that:

$$\Pr\{\Omega_n \leq x\} = \Pr\{J_n \leq x | K_n = 0\} \tag{2.67}$$

so instead of directly deriving the asymptotic distribution of $\Omega_n$, LeCam seeks the asymptotic joint distribution $(J_n, K_n)$ and derives the desired conditional distribution of $J$ given $K = 0$. Since $J_n$ and $K_n$ are sums of i.i.d. random variables, their joint asymptotic distribution may be obtained using classical limit theorems:

$$\lim_{n\to\infty} K_n = K \sim \mathcal{N}(0,1) \tag{2.68}$$

$$\lim_{n\to\infty} J_n = J = J_N + J_P \tag{2.69}$$

where $J$ is split between its Normal part, $J_N$, and its non-Normal (Poisson) part, $J_P$. LeCam's theorem states that if $(J_n, K_n) \to (J_N + J_P, K)$ then $\Omega_n \to J - cK$ where $c = \mathrm{E}[J_N K]$.

Specifically, if $(J_n, K_n)$ converges asymptotically to a two-dimensional Normal distribution $(J, K)$, then:

$$\lim_{n\to\infty} \Omega_n \sim \mathcal{N}(0, \sigma_J^2 - \sigma_{JK}^2). \tag{2.70}$$

This does not exhaust the results regarding the asymptotic distribution of combinations of functions of Spacings, regardless whether Uniform or not. For a broader and more detailed overview of limiting distributions of Spacings see [6, 4, 5].

# 3. Goodness-of-fit tests using Spacings

## 3.1. Introduction

Assessing the Goodness-of-Fit (GoF) of a distribution given a number of random samples is an often-encountered problem in data analysis. A GoF test consists in deciding whether a set of i.i.d. samples $\{X_1, ..., X_n\}$ of a univariate random variable $X$ was obtained from a population that can be described by a cumulative density function $F(x)$.

Such statistical hypothesis tests find applications in many fields, ranging from the natural and social sciences over to engineering and quality control.

For most "non-parametric" tests, the goodness-of-fit to a specific distribution $F(x)$, may be reduced to a uniformity test, i.e. testing whether or not the given observations have come from a Uniform population. This is possible due to the probability integral transformation [2, 3]: given the samples $X_i$ and the null-hypothesis distribution $F(x)$, we transform using $U_i = F(X_i)$; if the samples $X_i$ are distributed according to $F(x)$, then the samples $U_i \in [0, 1]$ are distributed according to the standard Uniform distribution $\mathcal{U}(0, 1)$. Such a transformation allows to develop and derive the distribution of a test statistic assuming the underlying distribution is only the standard Uniform one, and use these results by transforming into a space where this assumption holds true under the correct choice of the null-hypotheis $F(x)$. Therefore, in the rest of this work, without loss of generality, I only consider samples $U_i$ assuming a standard Uniform distribution as the null-hypothesis.

Several non-parametric tests exist, some of which have become standard tools, such as the Kolmogorov-Smirnov (KS) test [19, 20], the Cramér–von Mises (CvM) test [21] or the Anderson-Darling (AD) one [22]. Apart from these tests, which are based on the empirical cumulative distribution (ECDF Statistics), there is a rich literature regarding goodness-of-fit methods based on Order Statistics and Spacings. In the following, I briefly list and describe existing test statistics, borrowing from excellent reviews on this topic [23, 24, 6]. I then discuss two new proposed tests, the "Recursive Product of Spacings" (RPS), developed in collaboration with Dr. P. Eller [25], and the "Best Sum of Ordered Spacings". Finally, I present a detailed performance comparison between the tests presented here.

Parts of the text presented in this chapter closely follow [25].

## 3.2. ECDF Statistics

This class of test statistics compares the empirical cumulative distribution function (ECDF) $F_n(u)$ to the cumulative distribution function (CDF) $F(u)$, (here $F(u) = u$).

Clustering of points under the null hypothesis of a uniform distribution would induce a steeper ECDF compared to the expected CDF, leading to a large deviation between the two, as shown in Fig. 3.1.

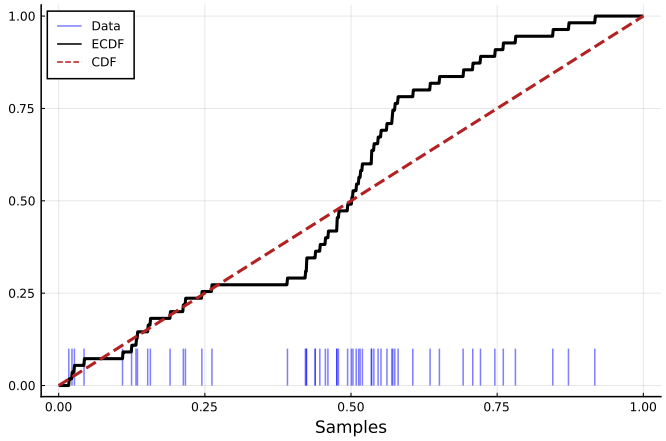



Figure 3.1.: Example of the the empirical distribution function (ECDF) of a set of samples $u_i \in [0, 1]$ against the null-hypothesis CDF $F(u) = u$.

In particular, the following tests are widely used in order to detect such deviations.

**Kolmogorov-Smirnov** (KS) test:

$$D_n = \sup_u |F_n(u) - F(u)| = \max(D_n^+, D_n^-) \tag{3.1}$$

where

$$D_n^+ = \max_i \left\{ \frac{i}{n} - U_{(i)} \right\} \quad \text{and} \quad D_n^- = \max_i \left\{ U_{(i)} - \frac{i-1}{n} \right\} \tag{3.2}$$

which are the largest vertical differences between $F_n(x)$ and $F(x)$. Kolmogorov [19] showed that the distribution of $D_n$, if $F(x)$ is the underlying distribution, is independent of $F(x)$, and derived the asymptotic distribution of $D_n$ as $n \to \infty$, as well as recursion formulae to calculate the pdf of $D_n$ for finite $n$. Later, Smirnov [20] provided a tabulation of the asymptotic distribution. Tabulations of the distribution for finite $n$ have also been provided [26, 27, 28].

**Quadratic** family of tests:

$$Q_n = n \int_{-\infty}^{+\infty} [F_n(u) - F(u)]^2 \, \Psi(u) dF(u) \tag{3.3}$$

where $\Psi(u)$ is a weighting function. When $\Psi(u) = 1$ we have the Cramér–von Mises (CvM) statistic $W_n^2$ [21] and when $\Psi(u) = [F(u)(1 - F(u))]^{-1}$ we deal with the Anderson-Darling (AD) statistic $A_n^2$ [22]. These tests can be expressed in terms of Order Statistics:

$$W_n^2 = \sum_{i=1}^{n} \left[ U_{(i)} - \frac{2i-1}{2n} \right]^2 + \frac{1}{12n} \tag{3.4}$$

$$A_n^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1) \left[ \log(U_{(i)}) \left( 1 - \log(U_{(n+1-i)}) \right) \right] \tag{3.5}$$

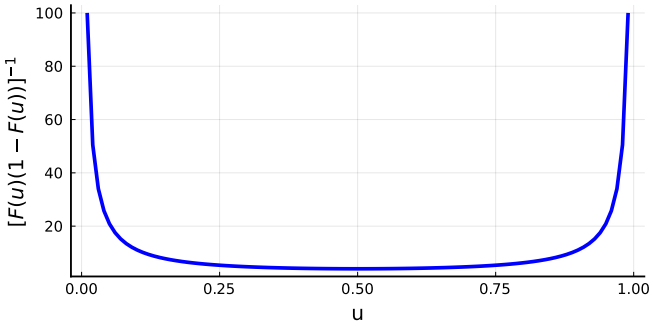



Figure 3.2.: Anderson-Darling weighting function $\Psi(u) = [F(u)(1 - F(u))]^{-1}$ asuming $F(u) = u$.

Regarding the AD test, we notice that the choice of the weighting function $\Psi(u) = [F(u)(1 - F(u))]^{-1}$, shown in Fig. 3.2, has the effect of assigning more importance to discrepancies located at the edges of the unit interval, making it better suited to detect clusters of events located at the edges of the analysis window. The asymptotic distribution of the AD test was derived in the original publication, while the $q$-values of the lower tail of the asymptotic distribution have been tabulated by Lewis [29].

## 3.3. Tests based on Order Statistics

These tests are based on the deviation ($\delta_i$) of each order statistic $U_{(i)}$ from its expected value:

$$\delta_i = U_{(i)} - \frac{i}{n+1}. \tag{3.6}$$

Some tests based on the deviations $\delta_i$ are:

- $C_n^+ = \max_i \delta_i$
- $C_n^- = -\min_i \delta_i$

- $C_n = \max(C_n^+, C_n^-)$
- $K_n = C_n^+ + C_n^-$

where $C_n$ is referred to as Pyke's modified KS test [30] (whose critical values are tabulated in [31]) and $K_n$ is Brunk's modified KS test [32].

## 3.4. Tests based on Spacings

Since the goal of a goodness-of-fit test is detecting discrepancies in the distribution of samples, this can be applied in a discovery context, whereby one searches for a signal by trying to spot an unusual cluster of events. Clusters of points lead to an increased number of unusually small spacings compared to the expectations, thus it is possible to construct tests sensitive to small spacings. Several such test statistics built from spacings $S_i$ are considered in the literature. Typically, they are of two main types:

- sum of a function of the spacings $S_i$, as considered in Eq. 2.64:

$$\Omega_n = \sum_{i=1}^{n+1} g(S_i) \tag{3.7}$$

- function of the ordered spacings $S_{(i)}$, which just like the order statistic $U_{(i)}$ are the ordered set of samples $U_i$, $S_{(i)}$ are the ordered (sorted) set of corresponding spacings $S_i$ ($S_{(i)} < S_{(j)}$ for $i < j$).

Borrowing from Pyke [6], examples of the former are:

- *Greenwood statistic*, proposed in [33]:

$$\Omega_n = \sum_{i=1}^{n+1} S_i^2 \tag{3.8}$$

  whose limiting density function was derived by Moran [34, 35, 36]

- *Kimball statistic* proposed in [37]:

$$\Omega_n = \sum_{i=1}^{n+1} S_i^r \text{ for } r > 0 \tag{3.9}$$

- *Irwin-Kimball statistic*, proposed in [33, 38]:

$$\Omega_n = \sum_{i=1}^{n+1} \left[ S_i - \frac{1}{n+1} \right]^2 \tag{3.10}$$

- *Kendall statistic*, proposed in [33]:

$$\Omega_n = \sum_{i=1}^{n+1} \left| S_i - \frac{1}{n+1} \right| \tag{3.11}$$

  whose limiting density function was derived by Sherman [39]

- *Moran statistic* proposed in [35] and studied by Darling as well [17]:

$$\Omega_n = -\sum_{i=1}^{n+1} \log(S_i) \tag{3.12}$$

- *Darling statistic* proposed in [17], where its limiting distribution was derived:

$$\Omega_n = \sum_{i=1}^{n+1} \frac{1}{S_i} \tag{3.13}$$

The exact distribution of the tests listed here has not been found for a finite value of $n$, apart from a few easy cases (such as $n \leq 3$), which shows that although one might deal with a simple combination of spacings this might already prove too difficult to solve for or to obtain a closed form solution. On the other hand, their asymptotic distributions have been studied in detail, and most of them can be derived using LeCam's theorem [18], as discussed in Sec. 2.6.3.

When it comes to ordered spacings, $S_{(i)}$, there have been fewer proposals; here are some:

- extreme ordered spacings $S_{(1)}$ or $S_{(n+1)}$ (references in [40])
- ratio $S_{(n+1)}/S_{(1)}$ or difference $S_{(n+1)} - S_{(1)}$ of extreme spacings: proposed by Kendall in [33] and whose limiting distributon was derived by Lévy [41] and Darling [17]
- sum of $k$ largest spacings $\sum_{i=n+2-k}^{n+1} S_{(i)}$, whose exact distribution was derived by Mauldon [42].

These tests are very sensitive to an overall mismatch between the empirical distribution of original samples and the null-hypothesis $F(u) = u$, but when it comes to detecting a cluster of events, they tend to be less powerful than the previous proposals since they are agnostic to location information: sorting the spacings has the effect of shuffling the ordered samples $U_{(i)}$, as shown in Fig. 3.3, partially delocalizing the clustered events.

Finally, it is also possible to construct tests based on higher rank spacings, $S_{i,k} = U_{(i)} - U_{(i-k)}$. Cressie considers statistics based on overlapping spacings of rank $k$ and defines generalizations of the Moran and Greenwood statistics:

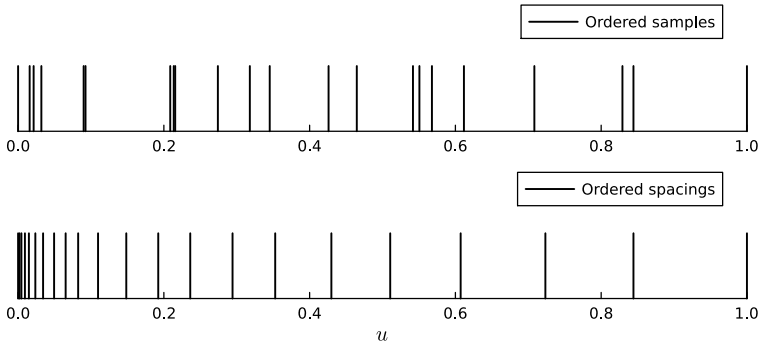



Figure 3.3.: Examples of ordered samples $U_{(i)}$ (top) and corresponding ordered spacings $S_{(i)}$ (bottom).

- Logarithms of higher rank spacings [43]:

$$L_{n,k} = -\sum_{i=1}^{n+2-k} \log S_{i,k} \tag{3.14}$$

- Squares of higher rank spacings [44]:

$$S_{n,k} = \sum_{i=1}^{n+2-k} S_{i,k}^2. \tag{3.15}$$

This concludes the overview of the tests present in literature, but is not an exhaustive list of proposed statistics.

## 3.5. Recursive Product of Spacings

Here I present the first of the newly proposed tests, the "Recursive Product of Spacings". This work was done in collaboration with Dr. P. Eller and presented in [25], which I closely follow for the coming discussion.

### 3.5.1. Definition

The goal is to construct a new test statistic, that has better sensitivity to narrow features or clusters in an otherwise Uniform distribution of samples. The tell-tale sign we are looking for is a localized group of uncommonly small spacings of the ordered data. For this purpose, I propose a new class of test statistics that includes higher rank spacings in a recursive way.

The Recursive Product of Spacings (RPS) can be thought of as an extension of the Moran statistic and is defined as:

$$RPS(n) = \sum_{j=1}^{n} M_j = M_1 + M_2 + \cdots + M_n, \tag{3.16}$$

where the term $M_1$ is the *simple* sum of negative log spacings equivalent to the Moran statistic (Eq. 3.12):

$$M_1 = -\sum_{i=1}^{n+1} \log\left(S_{i,1}\right) \tag{3.17}$$

The sum over all $\log(S_i)$ is the same as the logarithm of the product over all spacings $S_i$, thus the name *product* for the test. Additionally, working with logarithms is numerically more stable than products. All terms in Eq. 3.5.1 are computed in the same way as Moran's test:

$$M_j = -\sum_{i=1}^{n+2-j} \log\left(S^*_{i,j}\right), \tag{3.18}$$

but with modified spacings $S^*_{i,j}$, defined for $1 < j \leq n$ as:

$$S^+_{i,j} = \frac{S^*_{i,j-1} + S^*_{i-1,j-1}}{2} \tag{3.19}$$

$$S^*_{i,j} = \frac{S^+_{i,j}}{\sum_i S^+_{i,j}} \tag{3.20}$$

which there are $n + 2 - j$ of, and that depend on the spacings $S^*_{i,j-1}$ used to compute the previous term $M_{j-1}$ (hence the *recursiveness*). Obviously $S^*_{i,1} = S_{i,1} = S_i$.

In order to better understand Eq. 3.20, turn to Fig. 3.4, where it is shown how to transition from layer $j-1$ (top) to layer $j$ (bottom): in the top plot, it is shown a list of events (blue), where we also the boundaries 0 and 1 are highlighted, since they contribute to defining spacings; in the middle plot the middle points of the top row spacings are shown, forming a reduced set of "events", which is then transformed in order to ensure that the spacings of the new set sum up to 1, as shown in the bottom plot; the number of spacings going from the top plot to the bottom one is reduced by one, showing how there is a finite number of reduction steps in the definition of the $RPS$.

Regarding Eq. 3.18, I would like to point out its sequence reversal invariance: if the events were to be flipped ($\{x_i\} \to \{1 - x_i\}$), then one would obtain the same list of spacings in reversed order at all layers. The time reversal invariance in the formulae follows directly from the commutativity of sums and products.

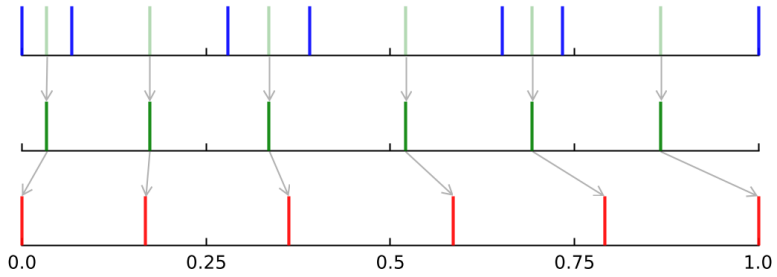



Figure 3.4.: Example of the reduction step included in the $RPS$ calculation. Given an initial set of events (top; blue), the middle points are calculated (top and middle; green) following Eq. 3.19, which are then scaled in order to fill the $[0, 1]$ interval, forming a new set of data (bottom; red), following Eq. 3.20. The evolution of sample positions on the $[0, 1]$ interval are annotated via the arrows.

We can see that term $M_2$ is identical to $L_{n,2}$ (Eq. 3.14) up to a normalization factor $1/\sum_i S_i$. If we considered the most regular and uniform case — a completely equidistant distribution of data, yielding all equal spacings $(1/(n+1))$ — then it would be helpful if the value of the test statistic for such a configuration were an extremum of its support. This is achieved by including a normalization at each layer of RPS. Doing so ensures that the equidistant samples remain equidistant in each layer, thus summing over the minimal contributions to the Moran test, which then yields the smallest possible RPS value. This minimum value of $RPS(n)$, given by the configuration of equidistant samples, can be expressed easily, as each spacing $S^*_{i,j}$ is equal to $\frac{1}{n+2-j}$, and thus:

$$RPS_{min}(n) = -\sum_{j=1}^{n}\sum_{i=1}^{n+2-j} \log\left(\frac{1}{n+2-j}\right) = \sum_{j=1}^{n}(j+1)\cdot\log(j+1). \tag{3.21}$$

At the other extreme, very small spacings will yield a large contribution to the sum of Eq. 3.18, thus $RPS_{max}(n) = \infty$ for any given number of samples $n$. These extrema show that RPS measures the irregularity in sample positions. The RPS statistic increases the more samples aggregate into local clusters.

The RPS quantity calculated so far has an infinite support $[RPS_{min}(n), +\infty)$. In order to bound the support of RPS, consider a new quantity $RPS^*$, with support $[0, 1]$, defined as:

$$RPS^*(n) = \frac{RPS_{min}(n)}{RPS(n)} \tag{3.22}$$

since the bounded interval makes extending the approximating function to the extrema of the test's support easier. This is the definition that should be considered when using the RPS test and for the remainder of this thesis. An interesting property of the construction

of $RPS$ is that spacings in the middle (order-wise, not w.r.t. the analysis window) will have a larger impact on the overall value of the statistic compared to spacings towards the edges: this means that the test is more sensitive to centrally located non-uniformities. Such a behaviour is not uncommon, in fact both the KS and AD tests do not possess uniform sensitivity over the analysis window: KS is more sensitive towards the middle while AD is more sensitive towards the edges.

Similarly to the definition of the RPS test, it is also possible to define an extension to Greenwood $G(n)$ statistic, that instead of summing over logarithms of spacings, sums over the squares of spacings. This means substituting Eq. 3.18 with $G_j = \sum_{i=1}^{n+2-j} \left(S^*_{i,j}\right)^2$, while keeping the definition of $S^*_{i,j}$ from Eq. 3.20. We call this recursive form the "Recursive Sum of Spacings" (RSS) test statistic.

### 3.5.2. Illustration

To better illustrate how the RPS statistic works and to highlight differences to other tests, two sets of samples, one drawn from a Uniform distribution (null hypothesis $H_0$) and one from a non-uniform distribution, shown in Fig. 3.5. The example given is a particularly challenging one and is used to illustrate the workings of different tests and highlight their difference, but it is not meant as a performance comparison between them. Actual performance comparisons using a large number of random replications are given in the following sections.

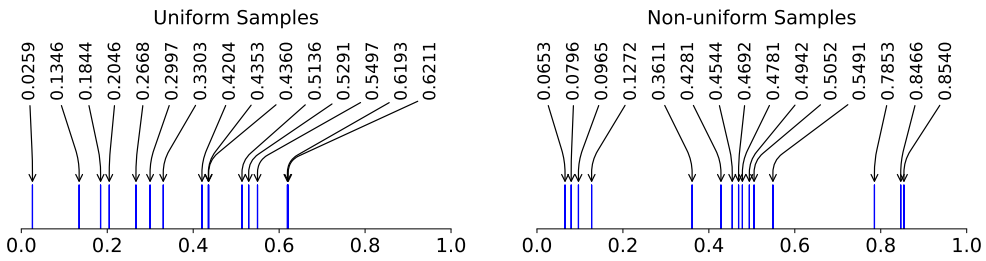



Figure 3.5.: Example of 15 standard uniformly distributed samples (left) and 10 standard uniformly + 5 normally ($\mu = 0.5, \sigma = 0.1$) distributed samples (right). The sample positions on the $[0, 1]$ interval are annotated via the arrows + text.

The Moran test is based on the spacings between samples, and the smallest and largest spacings in the specific example are present in the uniform case. This leads to a more extreme test statistic value and hence p-value $p = F(T \geq t|H_0 = \mathcal{U}(0, 1))$ of 0.117 for the uniform case, while it evaluates to $p = 0.335$ in the non-uniform case.

The KS test can detect such clustering via the CDF, however in this example, it is challenged by the fact that samples trend towards the left in the uniform case, while they are more balanced in the non-uniform case. This leads to p-values of 0.048 for uniform and 0.356 for non-uniform cases respectively.

The RPS test, however, taking into account also higher rank spacings, finds a p-value of 0.532 for the uniform case, and a much lower p-value of 0.057 for the non-uniform samples. The behaviour of RPS is further illustrated in Fig. 3.6, which shows the individual contribution of spacings of all recursion levels that build up the test statistic value. The Moran statistic corresponds to the sum over the first row ($M_1$), while all subsequent levels are added for RPS. By construction, Moran's test does not preserve information about the position of spacings, meaning that the value of the test is unchanged under reordering of spacings (the test's definition is invariant due to the commutative property of sums and products): clusters of samples, as in the non-uniform case, do not affect Moran's test. Including the recursive layers allows to preserve the information about the relative position of small spacings. This can be noticed by the stronger contributions to the RPS test value coming from different layers in the presence of a cluster of events (darker color on the right panel of Fig. 3.6) as opposed to the small contributions coming from layers beyond the first one in the case of uniform events (left).

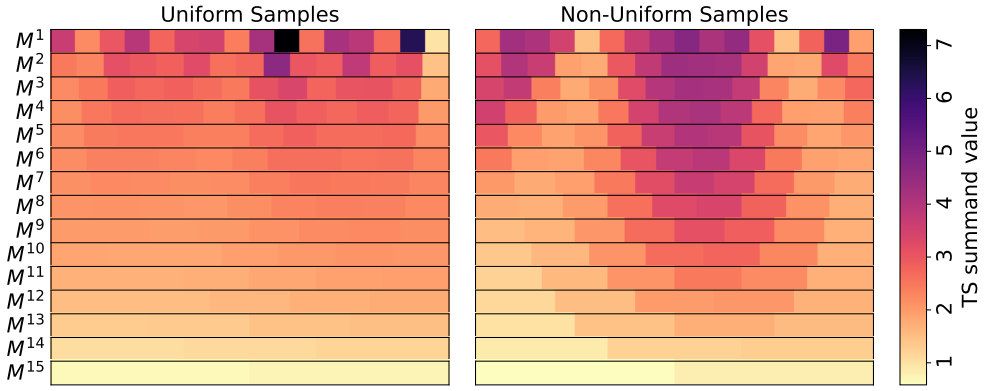



Figure 3.6.: Illustration of the test statistic contributions from all recursion levels for the uniformly distributed samples (left) and the non-uniform samples (right). The sum over the first level only ($M_1$) is equivalent to the Moran statistic.

### 3.5.3. Cumulative Distribution

In order to use RPS as a statistical test yielding p-values, we need its cumulative distribution $F_{RPS^*}$. In the case of $n = 1$, where only two spacings are present—the simplest non-trivial case we can encounter—the distribution of the only event present is the standard Uniform. So it is possible to write the formula of the test as a function of the sample value and find its distribution $RPS^*(1)$ as a simple transformation of random variables, which is:

$$F_{RPS^*}(x; n = 1) = 1 - \sqrt{1 - 4^{\frac{x-1}{x}}} \tag{3.23}$$

For $n \geq 2$, however, it is not simple to derive this distribution, therefore, it is necessary to to numerically approximate the distribution of $RPS^*$.

I have built an approximation for the cumulative distribution $F_{RPS^*}(x; n)$ precise enough to compute meaningful p-values up to relatively extreme values of up to $10^{-7}$, and large sample sizes $n$ of up to 1000. Fig. 3.7 shows some examples of $RPS^*$ distributions for a few values of $n$.

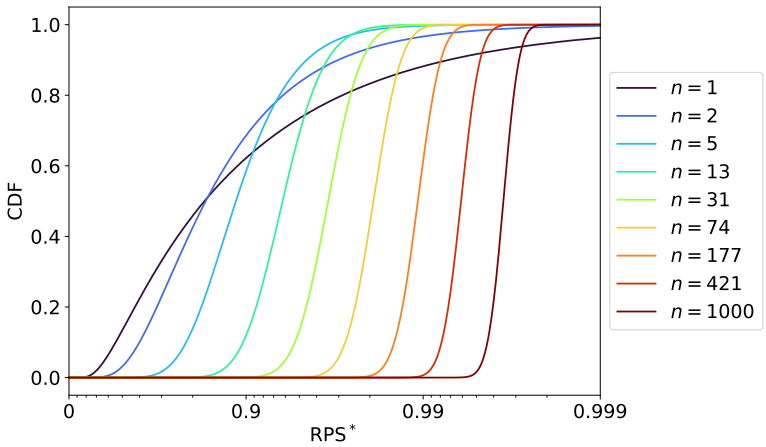



Figure 3.7.: Example of CDFs of the $RPS^*$ distribution for a few different values of $n$. N.B.: the x-axis is displayed in *inverted* logarithm.

The approximate distributions are based on simulations, drawing events with Uniform distribution in the range $[0, 1]$ for a given $n$. The detailed approximation and fitting procedure are described in Appendix C.

## 3.6. Sum of Spacings

Since we want to detect clusters of points by being sensitive to small spacings, instead of considering a combination of functions of spacings, we could rely on the spacings themselves and look for unusually small ones (compared to the expectation).

### 3.6.1. Best Sum of Spacings

Suppose we knew that a number of $k$ events out of the collected data, $n$ in total, are generated from an unknown narrow signal distribution. In such a scenario, we would be looking for clusters of at least $k$ events. One possible strategy to detect such a cluster, would be that of considering all collections of $k$ consecutive samples and select the set with the smallest width as the best signal-cluster candidate.

The metric used to single out this candidate is the distance between the first and $k$-th sample, i.e. the smallest spacing $S_{i,k}$ or rank $k$:

$$S_k^{min}(n) = \min_i S_{i,k} \tag{3.24}$$

In order to quantify how improbable it is that the cluster is due to statistical fluctuations of Uniform samples, we need to calculate the p-value of this observation using the cumulative distribution of $S_k^{min}(n)$. The CDF derivation is discussed in Appendix B, where I present an integral solution. Ultimately, such a solution proves to be impractical to evaluate for large values of $n$, so I rely on a numerical approximation, whose details are presented in Appendix C.

For now, assume we know the distribution, and hence we can calculate the p-value. The assumption we made at the beginning, approximately knowing the number of samples contained in a signal cluster, cannot be expected in a realistic scenario. In a real experiment, we might not know anything about the expected count rate of a possible signal and would like to be sensitive to narrow clusters of any number of samples $k$.

In such a case, we might think of testing for all possible orders $k$, thus obtaining the p-value corresponding to $S_{i,k}$ for of $1 \leq k \leq n$:

$$p_k = \Pr\{S_k^{min} \leq x_{obs,k}\}. \tag{3.25}$$

Out of these $n$ p-values we could reference the smallest one as the one indicative of the largest deviation from the expected behaviour of samples. I refer to this value as the smallest Best Sum of Spacings ($\mathrm{BSS_{min}}$):

$$\mathrm{BSS_{min}} = \min_k p_k \tag{3.26}$$

Since we choose the smallest value of all $p_k$, $\mathrm{BSS_{min}}$ is not a valid p-value anymore. In order to calculate a true p-value, we need to know the distribution of $\mathrm{BSS_{min}}$. Since the distribution of $S_k^{min}(n)$ is approximated numerically, so will be the distribution of $\mathrm{BSS_{min}}$, similarly to the distribution of the RPS test. The approximate distributions are based on simulations, and details about the fitting procedure are described in Appendix C.

### 3.6.2. Best Sum of Ordered Spacings

In the pursuit of discovering sample clustering against a uniform expectation, we could consider the ordered (sorted) spacings $S_{(i)}$ instead of $S_i$. Since clusters of points produce a local abundance of small spacings, it might be possible to compare the observed sorted list of spacings against the expected one. For a given rank $k$, one might consider the statistic defined by the Sum of the $k$ smallest Ordered Spacings, $S_{(k)}^{min}$:

$$S_{(k)}^{min}(n) = \sum_{i=1}^{k} S_{(i)}. \tag{3.27}$$

The exact distribution of $S^{min}_{(k)}(n)$ for a given rank $k$ and a given number of samples $n$ is known:

$$\Pr\left\{S^{min}_{(k)}(n) \leq x\right\} = \\ = \frac{A(k,N)}{N}\sum_{i=1}^{k}\frac{a(i,k)(k+1-i)}{(N+2-i)}\left(1-\left[1-\left(\frac{N+2-i}{k+1-i}\right)x\right]^{N}H\left(x,0,\frac{k+1-i}{N+2-i}\right)\right) \tag{3.28}$$

where $H(x,a,b) = 1$ if $a \leq x \leq b$ and 0 otherwise, while the coefficients $A(k,N)$ and $a(i,k)$ are given by:

$$A(k,N) = \frac{N(N+1)!}{(N+1-k)^{k-1}(N+1-k)!} \tag{3.29}$$

$$a(i,k) = \frac{(-1)^{i-1}(k+1-i)^{k-2}}{(k-i)!(i-1)!}. \tag{3.30}$$

I derived this result using a proof by induction reported in Appendix A. However, I need to mention that after more careful literature research, I found out that this result is not entirely novel. Mauldon [42] derived the exact distribution, using different methods, of the sum of the largest $k$ ordered spacings, $S^{max}_{(k)}(n)$:

$$S^{max}_{(k)}(n) = \sum_{i=n+2-k}^{n+1} S_{(i)} \tag{3.31}$$

and since the sum of all spacings is constrained to 1, then:

$$S^{min}_{(k)}(n) = 1 - S^{max}_{(k)}(n) \tag{3.32}$$

thus, knowing the cumulative distribution of $S^{max}_{(k)}$ for any $k$ and $n$:

$$\Pr\{S^{min}_{(k)}(n) \leq x\} = \Pr\{S^{max}_{(n+1-k)}(n) \geq 1-x\}. \tag{3.33}$$

Much like with the $\mathrm{BSS}_{\mathrm{min}}$ test, we would like to test all available ranks of $S^{min}_{(k)}$ for a given set of data, and compute the respective p-values:

$$p_k = \Pr\{S^{min}_{(k)}(n) \leq x_{obs,k}\} \text{ for } 1 \leq k \leq n+1 \tag{3.34}$$

Out of these p-values, we construct the test statistic based on the smallest one:

$$\mathrm{BSOS}_{\min}(n) = \min_k p_k = \min_k F[S^{min}_{(k)}(n)]. \tag{3.35}$$

The distribution of $\mathrm{BSOS}_{\min}$ is not simple to derive, and although it could be expressed in a nested integral form, like the distribution of the $\mathrm{BSS}_{\min}$ statistic, these are not easily solvable. Therefore, I resort to approximating the distribution of the $\mathrm{BSOS}_{\min}$ statistics with simulations. I have tabulated the distribution of $\mathrm{BSOS}_{\min}(n)$ for relatively few values of $n$ and use these to interpolate the approximate distribution across all values of $n \leq 10^3$, similarly to the RPS test. Details on the interpolation and error estimation of the accuracy of the approximate distribution are reported in Appendix C.

## 3.7. Spacings as time-series

Before showing examples of the applications of the test I proposed in previous sections, I would like to introduce an additional class of tests, derived from spacings.

During the review of the basic results of Order Statistics, I showed in Sec. 2.5 how to transform spacings in a set of independent random variables, either Exponential or Beta distributed. Given these transformations, it is possible to translate the observed spacings $S_i$ in any collection of independent random variables, targeting any arbitrary univariate distribution, using a probability integral transformation. For example, given that we are able to transform the spacing $S_i$ into an independent variate $B_i$ with distribution $\mathrm{Beta}(n+1-i, 1)$, we can then transform $B_i$ into a Standard Normal random variable $Y_i$:

$$Y_i = F^{-1}_{\mathrm{Normal}}\left(F_{\mathrm{Beta}}(B_i)\right). \tag{3.36}$$

Given that we can transform the $n+1$ spacings $\{S_i\}$ into $n$ standard Normal variables $\{Y_i\}$, these can represent a time-series of normally distributed deviations from a baseline. When faced with such a dataset, we are suddenly exposed to a multitude of tests developed in the signal-processing community in order to detect sizable deviations from the expectation in a stream of data.

### 3.7.1. Success runs statistic for Spacings

Given $n$ independent, normally distributed variables $Y_i \sim \mathcal{N}(\mu_i, \sigma_i)$, an observation is considered a *success* if the observed value exceeds the expected value ($Y_1 > \mu_i$), while it is considered a failure is it doesn't ($Y_i < \mu_i$). Based on this definition we focus on success runs: uninterrupted sequences of successes in a set of data. In order to analyse these runs, we resort to the "Run Statistic" introduced by Beaujean & Caldwell [45]:

- split the data $\{Y_i\}$ in runs, keeping the success runs and ignoring failure runs, denoting by $A_j = \{Y_{j_1}, Y_{j_2}, ...\}$ the set of observations in the $j$-th success run

- each run is associated with a weight, $w(A_j)$, which indicates the discrepancy between model and observation:

$$w(A_j) = \chi^2_j = \sum_{i \in \{j_1, j_2, \ldots\}} \frac{(Y_i - \mu_i)^2}{\sigma_i^2} \tag{3.37}$$

- choose the largest weight as the value of the test statistic $T_{RUN}$:

$$T_{RUN} = \max_j \chi^2_j \tag{3.38}$$

- the p-value is:

$$p_{RUN} = \Pr\{T_{RUN} \geq T_{obs}|n\} = 1 - \Pr\{T_{RUN} \leq T_{obs}|n\} \tag{3.39}$$

The exact derivation of the RUN statistic $T_{RUN}$ is reported in [45]. The use of the exact distribution becomes computationally costly for $n \gtrsim 100$ and a high-precision extrapolation from a few dozen up to millions of data points is reported in [46].

We can use this test statistic to analyse events $\{X_i\}$ $(i \leq n)$ against the null-hypothesis $F(x)$ by using the RUN statsitic on a set of Standard Normal random variables $\{Y_i\}$ $(i \leq n)$ obtained using the transformation chain:

$$X_{(i)} \to U_{(i)} \to S_i \to B_i \to Y_i.$$

To be sure that this transformation covers the support of the Standard Normal distribution correctly, we can compare the distribution of p-values obtained using repeated random events produced according to a Standard Uniform distribution and then transformed into Gaussian variates against the p-value distribution obtained by generating directly Gaussian variates with the correct distribution ($\mathcal{N}(0,1)$). The results of this consistency check are shown in Fig. 3.8, where we notice that both distributions are flat.

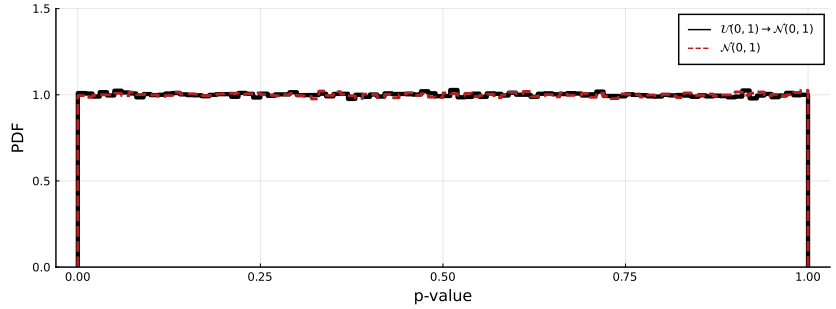



Figure 3.8.: RUN statistic p-value distribution for transformed samples (black) and original samples (red) for $n = 20$.

Although the p-value distribution is flat, there are a few peculiarities regarding the transformation of spacings into Gaussian variates that need to be mentioned. The transformation of random variables from one distribution to another shown in Eq. 3.36 conserves all quantiles, meaning that if $B_i$ corresponds to the median of $F_{\text{Beta}}$, then $Y_i$ will be equal to the median of $F_{\text{Normal}}$ as well. The mean on the other hand is not conserved in random variable tranformations, unless both the starting and target distributions happen to be symmetric. In this case, the starting Beta distribution is not symmetric, thus:

$$\mathrm{E}[Y|f_{\text{Normal}}] = 0 \neq F^{-1}_{\text{Normal}}\left(F_{\text{Beta}}(\mathrm{E}[B|f_{\text{Beta}}])\right). \tag{3.40}$$

Given a set of equidistant events, which represent the expectation of a random collection of $n$ standard Uniform variables and translates into $n+1$ equal spacings $S_i = 1/(n+1)$, after transforming into standard Normal variates we do not have a set of $n$ variables all equal to 0, but instead, the Gaussian samples plotted in Fig. 3.9.

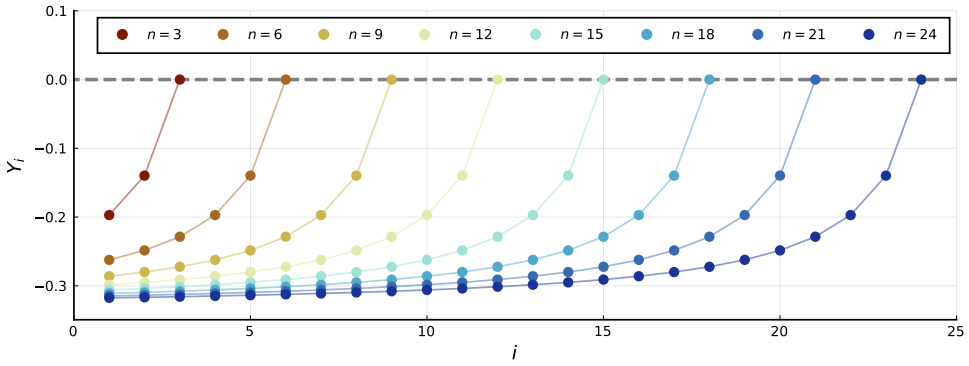



Figure 3.9.: Strandard Normal random variables transformed from equal spacings for different values of $n$.

Additionally, the transformation of spacings is not invariant under sequence reversal: mirroring all samples $U_i$ with respect to 0.5 (i.e. $U_i \to 1 - U_i$), the spacings $S_i$ would be in reverse order, but since the Beta distribution applied to transform them depends on the index $i$, the transformed samples $Y_i$ would not be the same, meaning we transform to a different time-series.

In this thesis, I consider the transformation of spacings without any further modification, since I show that it yields a valid p-value, but in order to render the test invariant under sequence reversal, one suggestion would be to transform both the original and reversed samples, obtaining two time-series, calculate the p-values for both and consider only the smallest one, $p_{min}$. The distribution of $p_{min}$ is not uniform, as shown in Fig. 3.10, so one could parametrize this distribution across the values of $n$, using the same interpolation and approximation schemes describe in Appendix C in order to estimate a valid p-value.

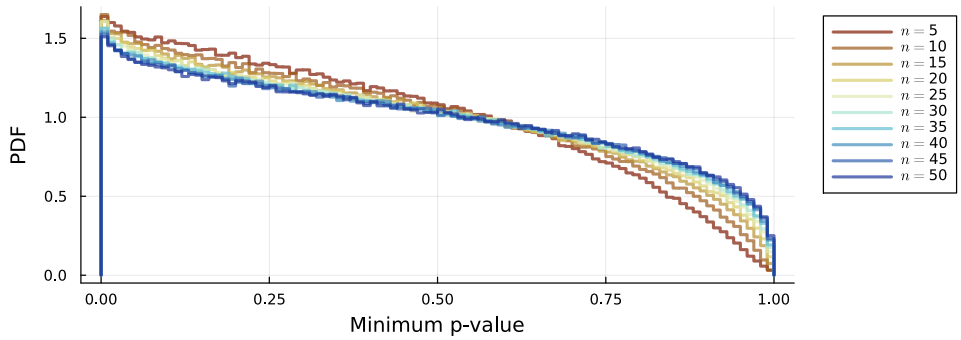



Figure 3.10.: Distribution of smallest p-value of the RUN statistic between the transformation of the spacings and reversed spacings, for different values of $n$.

## 3.8. General Performance Comparison

This section presents an in-depth performance comparison of the proposed tests (RPS, $\text{BSS}_{\text{min}}$, $\text{BSOS}_{\text{min}}$ and RUN statistics) to several other tests referenced in the introduction (KS, AD, CvM and Moran — all those that allow computing p-values).

The goal is to detect small changes in an otherwise uniform distribution, therefore consider the following generic benchmark scenario: for one simulation of a specific test case $H^K(n, s, w)$ generate $(1 - s) \cdot n$ random variates[1] from a standard Uniform distribution $\mathcal{U}(0, 1)$, where $s$ is a *signal* fraction. In addition, include $s \cdot n$ samples distributed according to $\Delta + \mathcal{N}(0, w/2)$ with the offset $\Delta = \mathcal{U}(0, 1 - w)$, i.e. a truncated (in the interval $[0, 1]$) Normal distribution with $\sigma = w/2$ located randomly within the interval $[w/2, 1 - w/2]$. In these tests, I vary all three parameters of $H^K(n, s, w)$: the number of samples $n$, logarithmically distributed between 10 and 1000; the fraction of signal events $s$, which goes from 0 to 30% of the total number of samples $n$; the "width" of the signal distribution $w$, which ranges from 0.01 to 0.35, in order to test narrow and wide signals. A sensitive test should be able to detect the presence of the added, narrower signal samples by reporting a low p-value. For each choice of $H^K(n, s, w)$, I produce a distribution of p-values obtained by $10^5$ trials.

In order to quantify the sensitivity of a test statistic, I compute the median of its corresponding p-value distribution for each $H^K(n, s, w)$. This quantity can be interpreted as the median significance at which a test is expected to be able to reject the null hypothesis. When comparing tests with one another, I interpret a lower reported median p-value as a more powerful test (higher sensitivity).

The performance (sensitivity) of the various tests as a function of $(n, s, w)$ is shown in Fig. 3.11. The outer axes at the top and left of the table of plots indicate a specific choice of number of samples and signal width respectively. Given $(n, w)$, the sensitivity of each test statistic (its median, row-specific log-scale can be seen on the right of Fig. 3.11) is plotted as a function of the signal fraction $s$ (shared horizontal axis of each subplot).

[1] Numbers of samples are rounded to the closest integer

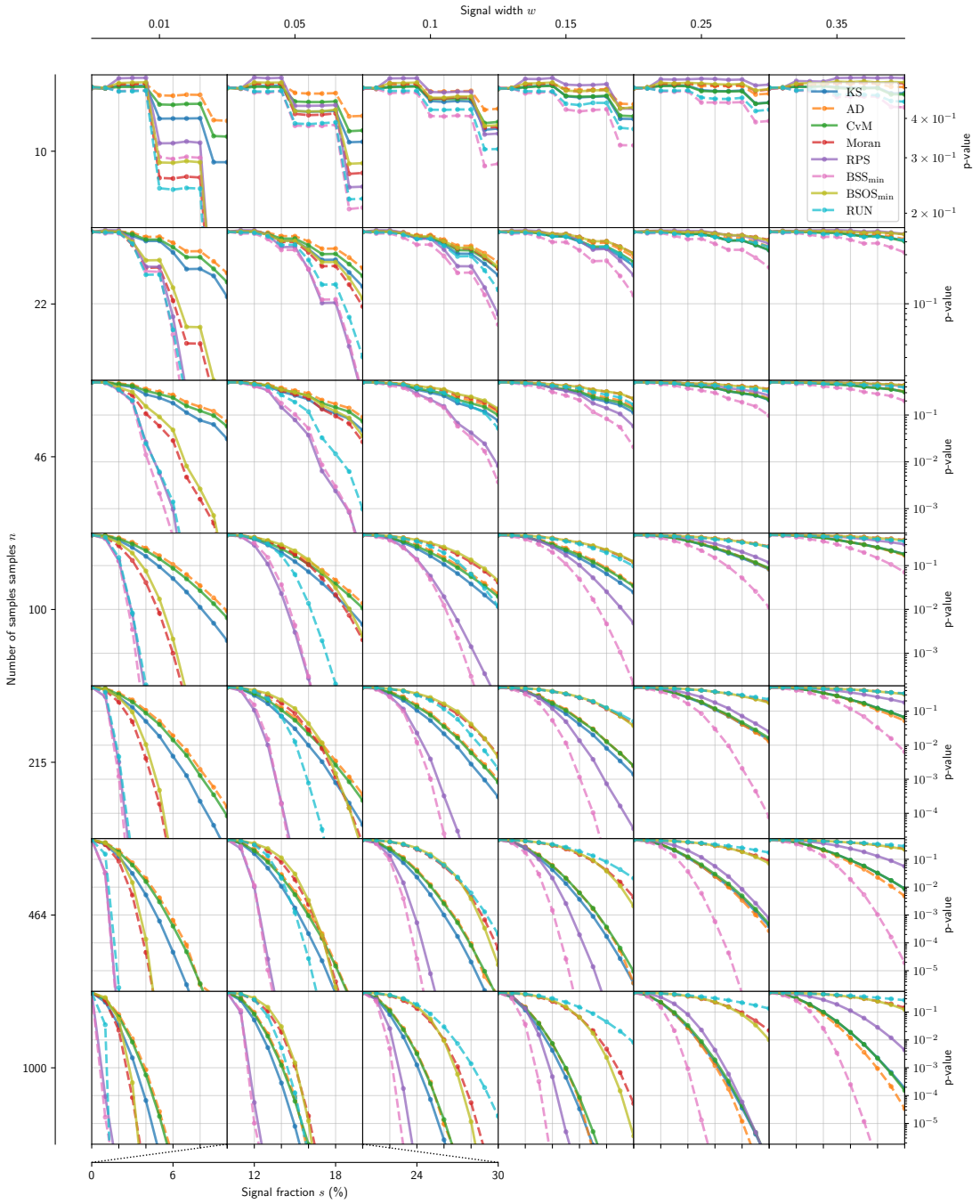



Figure 3.11.: Comparison of the performance (median p-value of repeated trials, individual panel's y-axis) as a function of the number of total samples (large y-axis), the width of the signal (large x-axis) samples, and the fraction of signal samples (individual panel's x-axis). The number of signal samples is rounded to the closest integer, hence the "step"-like features visible mostly in the first few rows.

Examining Fig. 3.11, we notice that for $n < 50$ and $w > 0.1$, the performance of all test statistics is very similar, with the $\mathrm{BSS}_{\mathrm{min}}$ one being slightly more sensitive than the rest, although the overall sensitivity is quite low given the small size and large width of the signal distribution, which cannot be easily distinguished from the null-hypothesis. As the width of the signal decreases, and the number of total events increases, all tests become more sensitive and are able to reach sensitivities of the order of $3 - 5\sigma$.

Focusing on the ECDF statistics (KS, CvM and AD), we notice that they perform very similarly, with the KS test being slightly more sensitive for narrow signals and the AD test overtaking it by a small margin when dealing with large signals ($w > 0.2$) and large signal fractions ($s > 15\%$). Given this grouping, I refer to the sensitivity of the ECDF statistics as a baseline against which I compare the performance of the Spacings based statistics.

Looking at the $\mathrm{BSOS}_{\mathrm{min}}$ and Moran statistics, we also notice that their sensitivities are very close, showing a comparable performance across the board. These statistics prove to be more sensitive than the ECDF ones only when dealing with very narrow signals ($w < 0.05$), but as the width of the signal increases they become less powerful. The reason for this behaviour can be explained by considering that these statistics are based only on the first-rank spacings ($S_i$), and in the case of $\mathrm{BSOS}_{\mathrm{min}}$ the local structure of the signal event-cluster is broken down by the reordering of spacings. The sole presence of small first-rank spacings can be sufficient only if they are extremely small (in the case of a narrow signal with $w < 0.05$). As soon as this condition is lifted, by widening the signal, then the sensitivity of the tests suffers.

The RPS statistic, which relies on higher-ranked spacings through its recursive construction, shows a much better sensitivity. When dealing with signals with width $w < 0.2$, the RPS test's sensitivity is orders of magnitude better than the ECDF ones, being able to reach discovery thresholds ($\sim 5\sigma$) far sooner than the KS test. For $w \lesssim 0.1$ the RPS is practically on par with the best statistics, while it becomes the second best for $w \sim 0.2$. For $w \sim 0.25$ its sensitivity is similar to the ECDF statistics and for wider signals it deteriorates further.

Considering the RUN statistic, we notice that it is on par with the best statistics only for very narrow signals ($w \sim 0.01$); as the width of the signal increases it quickly becomes the second-best for $w \sim 0.05$ and then comparable or less sensitive than the ECDF statistics for $w > 0.05$. The reason for this behaviour is due to the fact that the RUN statistic is based only on the first-rank spacings, much like Moran and $\mathrm{BSOS}_{\mathrm{min}}$. Moreover, if one of the spacings produced by the clustering of signal events is transformed to a Gaussian variate less than 0, then this would break the "run", yielding a larger p-value, thus a lower sensitivity: for example, given 10 consecutive positive small spacings, if they were transformed in 10 consecutive positive Gaussian variates, that would make a "run" of 10, while if one of them is non-positive, this leads to two "runs", of say 3 and 6 samples, which individually are more likely to yield a larger p-value. The probability of having a "run-breaking" spacing increases with wider signals, causing the decline in sensitivity seen in Fig. 3.11.

Finally, looking at the $\mathrm{BSS}_{\mathrm{min}}$ statistic, we notice that it is always the most sensitive

test, regardless of the width or strength of the signal or the number of events. For narrow signals, its sensitivity is matched by the RPS and briefly by the RUN statistics (only for $w \sim 0.01$) but for wide signals it stands out as the best of the bunch, being several orders of magnitude better than the ECDF statistics even for large sample sizes.

Apart from the median, I also investigated other metrics to judge the tests' performance, such as the area under the receiver operating characteristics (ROC) curve between the p-value distribution of signal and null hypothesis trials, but the overall picture does not change substantially.

Given these results, I am able to recommend the use of the $\mathrm{BSS}_{\mathrm{min}}$ statistic as a goodness-of-fit test for any discovery application since it appears to be one of the most sensitive non-parametric tests available.

In case more information is known about the possible signal, for example, if there are upper limits to its width or strength, then the results presented in Fig. 3.11 could guide the selection of the most sensitive tests that can be employed. All of the Spacing-based tests presented here are available using the SpacingStatistics.jl [47] package for Julia, so instead of relying on the general guidelines presented in Fig. 3.11, it is possible to try out these statistics and characterize their sensitivity on a customized model of background and signal, as would be usually done during validation studies before the unblinding and analysis of the data of an experiment.

# 4. Limit setting using spacings

## 4.1. Introduction

Many experiments tasked with the discovery of theorised rare processes might find themselves in a situation where the collected data is insufficient to claim a positive detection. In such cases, the data is used to set an upper limit on the number of events resulting from the rare process under consideration, which in turn can be used to set an upper limit on physical quantities of the proposed model. An example would be the determinations of upper limits on the cross-section of Weakly Interacting Massive Particles (WIMPs) recoiling off atoms in a detector, such as for the CRESST [48] or the CDMS [49] experiments. The experiments in question are often contaminated by a poorly understood background, in which case the signal strength limit must be set from properties of the observed event distribution without any background subtraction. Spacing statistics are one method to go beyond pure event counting in setting signal strength limits.

Since the expected shape of the event distribution produced by the targeted process is known, it is possible to estimate the number of events it accounts for, up to a desired confidence level, leveraging the difference between the observed event distribution and the expected one. Such an analysis is carried out using goodness-of-fit tests allowing for the assumption that the observed number of events collected in the analysis window is just a realization of a random variable following a Poisson distribution with unknown rate $\mu$. For a selected goodness-of-fit test, the goal is to determine the event rate $\mu$ coinciding with the desired confidence level.

In the following, I briefly review how to use goodness-of-fit tests to set upper limits, accounting for the random number of observed events in the definition of the p-value.

I then discuss various spacings-based tests to provide upper limits: I begin with a quick review of the Maximum Gap and Optimum Interval methods [1] and then introduce two new tests based respectively on the sorted list of spacings and on the product of spacings.

The content of this chapter closely follows [50], where these results were first presented.

## 4.2. Setting upper limits with test statistics

Several non-parametric goodness-of-fit tests exist to test a univariate distribution. Targeting vastly different univariate distributions is made possible by the probability integral transformation [2, 3], which basically reduces the goodness-of-fit to a simple uniformity test, as we have previously seen.

As a reminder, given $n$ samples $\{x_i\}$, if we want to quantitatively test the hypothesis of these samples being random variates of a known continuous cumulative distributions $F(x)$, independent and identically distributed (i.i.d.) according to $f(x)$, then we transform the samples onto the unit interval $[0, 1]$ via $u_i = F(x_i)$. This reduces the task at hand to test transformed samples $\{u_i\}$ being distributed according to the standard Uniform distribution $\mathcal{U}(0, 1)$.

In the rest of this chapter, I always consider samples $u_i \in [0, 1]$ distributed according to a standard Uniform null hypothesis unless otherwise stated.

### 4.2.1. Poisson distribution and p-value

In a standard goodness-of-fit test scenario, given a dataset consisting of $n$ uniformly distributed samples $\{u_i\}$, we consider a test statistic $T$, based on a scalar function of the data $t = g(\{u_i\})$. From this, we can extract a p-value, which can be calculated directly from $F_T(t|n)$, where $F_T$ is the cumulative distribution function of the test $T$ for exactly $n$ events. In this case, the p-value treats the number of events in the analysis window as a fixed parameter. If we assume that the observed number of events is not fixed, but rather a random variable with an associated Poisson distribution, then it is possible to correct the definition of the p-value by averaging over all possible numbers of events. Considering a Poisson distribution with rate $\mu$, and an observed test statistic value $t_{obs} = g(\{u_i\})$, the Poisson-averaged p-value is calculated as:

$$p = F_{T,Pois}(t_{obs}|\mu) = \sum_{n=1}^{\infty} F_T(t_{obs}|n) \cdot \frac{\mu^n e^{-\mu}}{n!} \tag{4.1}$$

where the sum starts at $n = 1$ since the test statistic often is not defined for $n = 0$. In case of no observed events, $n = 0$, no improvement upon the simple Poisson statistic is possible, which becomes the extension of this approach in the limit of empty datasets.

### 4.2.2. Setting upper limits

Given a test statistic $T$ and its Poisson-averaged cumulative distribution $F_{T,Pois}$, I showed above how to calculate the p-value of a given dataset $\{x_i\}$ comprised of $n$ events, for a selected value of the event rate $\mu$. Instead of performing a goodness of fit test, assessing how well the event distribution fits that of a uniform distribution for a given $\mu$, we could determine which is the rate $\mu$ that yields a desired p-value, determining the event rate representative of the uniformly distributed subset of events up to a desired confidence level ($CL$).

As an example, for a given dataset $\{u_i\}$ the upper limit on the event rate, $\mu$, at a confidence level $CL$ is such that:

$$p = F_{T,Pois}(t_{obs}|\mu) = CL \tag{4.2}$$

## 4.3. Spacing statistics

Given a uniformly distributed and ordered set of $n$ events $\{U_{(i)}\}$ in the interval $[0, 1]$, we can consider the spacings $S_{i,k}$ between the samples, with $U_{(0)} = 0$ and $U_{(n+1)} = 1$.

Based on these spacings it is possible to construct test statistics capable of setting much more competitive upper limits on the event rate than the simple counting (Poisson) test, since they not only consider the total number of data contained in the analysis window, but also their distribution, taking advantage of regions of relatively low event density in order to estimate the underlying uniform component of the event distribution.

As an intuitive example of the power that spacings can have in estimating upper limits, consider the dataset shown in Fig. 4.1. We notice that the distribution of events is clearly not uniform, since the density of events closer to the edges of the analysis window is evidently higher than the density observed in the middle. Looking at this dataset, we are led to believe that there is some unknown background that produces events closer to the edges 0 and 1, while it does not affect as drastically the distribution of events closer to the middle of the range. Thus we assume that the regions with the lowest density of events are those least affected by additional backgrounds, or in the worst case, only affected by backgrounds that are indistinguishable from the signal under study. Regions with low density of events, are also regions that present larger spacings between samples. In a symmetric situation to the one considered in the previous chapter, here we would like to be sensitive to the presence of large spacings in our data, filtering out collections of small spacings, thus trying to filter out any clusters and only look at the underlying uniform distribution.

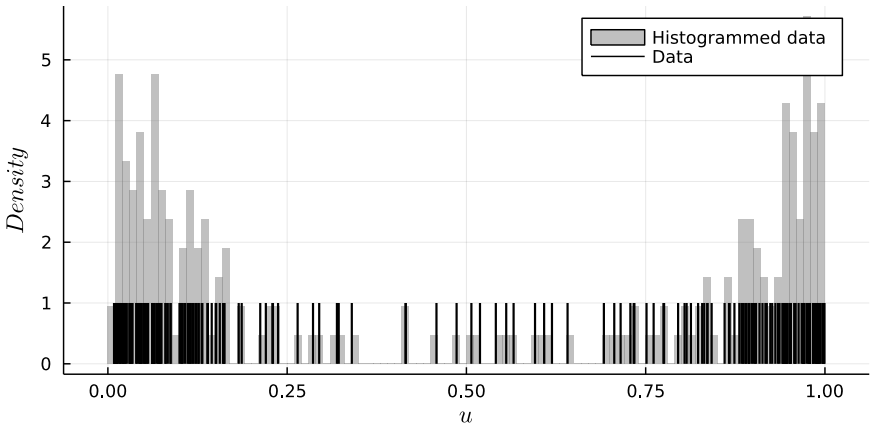



Figure 4.1.: Example of dataset presenting a possible background contamination at the edges of the analysis window.

### 4.3.1. Maximum Gap

The Maximum Gap test has been proposed in the physics literature [1] and earlier, as a test statistic, in the statistics literature [51]. It consists of the largest spacing present in order to determine the upper limit on the event rate. The test statistic is defined as:

$$T_{MG}(\{u_i\}) = S_1^{max} = \max_i(S_{i,1}) \tag{4.3}$$

The distribution of $S_1^{max}$, for a given number of events $n$, is known [51, 1]:

$$\Pr\{S_1^{max} \leq x|L\} = \sum_{t=0}^{m}(-1)^t\binom{n+1}{t}\left(1-\frac{tx}{L}\right)^n \tag{4.4}$$

where $L$ is the total length of the range of events ($u_i \in [0, L]$ and $\sum_i S_i = L$ instead of 1; this is a simple rescaling) and $m$ is the greatest integer such that $m \leq L/x$. Under the assumption of a uniform distribution of events, a large proposed event rate, $\mu$, will lead to a small probability to observe large values of $S_1^{max}$. If the observed value of $S_1^{max}$ is indeed large relative to our expectations, this can be used to exclude values of $\mu$ at a specified confidence level.

This definition of the test is particularly helpful since it is little affected by possible clustering of events in the unit interval, whose distribution deviates from the standard uniform one.

The Poisson averaged cumulative distribution, $F_{MG,Pois}$, can be easily computed analytically and has been given by Yellin [1]:

$$\begin{aligned} F_{MG,Pois}(x|\mu) &= \sum_{t=0}^{m}\sum_{j=0}^{\infty} e^{-\mu}\frac{\mu^j}{j!}(-1)^t\binom{j+1}{t}\left(1-\frac{tx}{\mu}\right)^j \\ &= \sum_{t=0}^{m}\frac{(tx-\mu)^t e^{-tx}}{t!}\left(1+\frac{t}{\mu-tx}\right) \end{aligned} \tag{4.5}$$

where the author considers $L = \mu$ to derive the simplified formula consisting of a finite sum. The $CL$ upper limit on the event rate, $\mu_{MG}$, is such that:

$$F_{MG,Pois}(S_1^{max}|\mu_{MG}) = CL \tag{4.6}$$

### 4.3.2. Optimum Interval

In addition to the Maximum gap method, Yellin proposes also the Optimum Interval (OI) method [1] which instead of looking at the largest spacing, considers sums of spacings: i.e., spacings of higher rank.

The Maximum Gap method compares the size of the largest first rank spacing against the expectation of it containing no events for a given event rate $\mu$. Similarly, given higher rank spacings, for example $k = 2$, we might find the largest second rank spacing, $S_2^{max}$, and compare its size to the expectation of it containing only one event given a proposed event rate $\mu$. Such an investigation can be performed for any rank of spacings allowed by the data ($k \leq n$, where $n$ is the number of events) and would result in $n$ different Poisson-averaged p-values, one for each rank of spacing, for a given event rate $\mu$:

$$S_k^{max} = \max_i(S_{i,k}) \tag{4.7}$$

$$p_k = F_{max,k,Pois}(S_k^{max}|\mu) \tag{4.8}$$

where $F_{max,k}$ is the cumulative distribution of $S_k^{max}$ for a given number of events. The analytic formula of $F_{max,k}$ ($k > 1$) for $n$ events is not known, but Yellin calculated numerical approximations using large Monte Carlo campaigns, similarly to how I derived previous approximate distributions. Additionally, Yellin produced an approximate asymptotic distribution for large values of $n$ [52], leveraging the asymptotically normal behaviour of the spacings for large $n$.

In order to exclude the proposed event rate $\mu$, one might look at the largest available p-value, as the one that most strongly rejects the hypothesis of the rate being $\mu$:

$$p_{max} = \max_k(p_k) \tag{4.9}$$

So defined, this test is similar to the Best Sum of Spacings ($\text{BSS}_{\text{min}}$), with the only difference that we consider the largest p-value instead of the smallest and we account for a random number of events via the Poisson-averaging of the p-value calculation.

Since $p_{max}$ does not have a uniform distribution, it can't be interpreted as a valid p-value, thus one needs to know its cumulative distribution for a given event rate $\mu$, $F_{\text{OI}}$. Knowing this, the final p-value is:

$$p_{\text{OI}} = F_{\text{OI}}(p_{max}|\mu) \tag{4.10}$$

The analytic formula of $F_{\text{OI}}$ is not known, and a numerical approximation is derived using Monte Carlo simulations. Fig. 4.2 shows the value of the 90% quantile of $p_{max}$ as a function of $\mu$.

Finally, the upper limit $\mu_{OI}$ on the event rate up to a given $CL$ is such that:

$$F_{\text{OI}}(p_{max}|\mu_{OI}) = CL \tag{4.11}$$

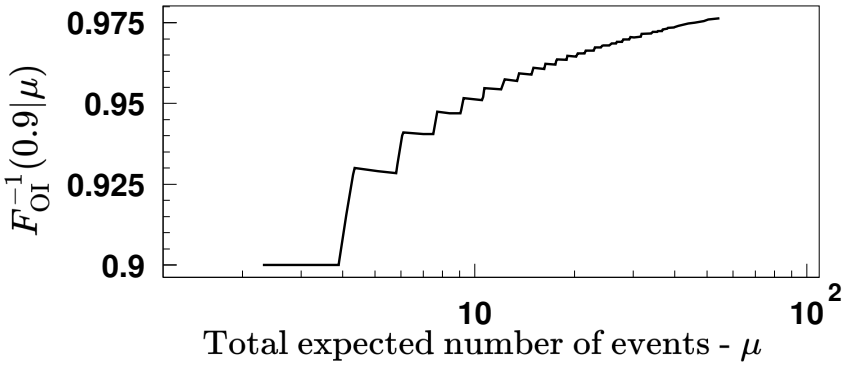



Figure 4.2.: Plot of the 90% quantile of $p_{max}$ as function of $\mu$ for the Optimum Interval method. Image taken from [1].

### 4.3.3. Best Sum of Ordered Spacings

The tests discussed so far were developed in order to be sensitive to the presence of any abnormally large gaps between the events. For relatively few events under analysis, there might be only one such abnormally large gap, in which case the Maximum Gap method might already provide the most stringent upper limit. If more than one such abnormally large gaps are present, and if these gaps happen to be located near one another, then they can be integrated into one higher rank spacing, and the Optimum Interval method provides more competitive limits.

However, if the abnormally large gaps are interspaced by many small gaps, then the Optimum Interval method would lose sensitivity and not offer significant improvements compared to the Maximum Gap method: in these cases, the Best Sum of Spacings approach, i.e. the Optimum Interval, saturates and is dominated by individual low-rank spacings.

Ideally, we would like to combine the abnormally large gaps individually, without the (smaller) gaps interspaced between them. This approach would increase the sensitivity of the test and potentially allow setting more competitive upper limits. We now describe such an approach.

Given a set of $n$ events $\{u_i\}$ in the unit interval $[0, 1]$, consider the ordered set of spacings , $S_{(i)}$. Given $\{S_{(i)}\}$, it is possible to consider higher rank spacings by summing over its elements. The $k$-th Sum of Ordered Spacings, $S_{(k)}^{max}$, is just the sum of the $k$ largest order spacings:

$$S_{(k)}^{max} = \sum_{i=1}^{k} S_{(n+2-i)} = \sum_{i=n+2-i}^{n+1} S_{(i)} \tag{4.12}$$

In total there are up to $n$ non-trivial $S_{(k)}^{max}$ given $n$ events ($1 \leq k \leq n$), since the sum over all $n + 1$ ordered spacings is constrained to be equal to 1.

It is now possible to evaluate the p-value of each rank of sum of ordered spacings for a given event rate $\mu$:

$$p_k = F_{max,(k),Pois}\left(S^{max}_{(k)}\middle|\,\mu\right) \tag{4.13}$$

where $F_{max,(k),Pois}$ is the cumulative distribution of $S^{max}_{(k)}$ for a given number of events. The analytic formula of $F_{max,(k),Pois}$ for $n$ events is known: it was first derived by Mauldon [42], and I have independently re-derived it using an alternative approach, reported in Appendix A.

As with the Optimum Interval method, it is possible to use the largest p-value as a test statistic in order to exclude a value of $\mu$ that is too large. I call this test value the largest Best Sum of Ordered Spacings ($\mathrm{BSOS}_{\max}$):

$$\mathrm{BSOS}_{\max} = \max_k(p_k) \tag{4.14}$$

Since $\mathrm{BSOS}_{\max}$ is not a valid p-value any more, one needs to know its cumulative distribution for a given event rate $\mu$, $F_{\mathrm{BSOS},max}$, and the final p-value is calculated in this case as:

$$p_{\mathrm{BSOS},max} = F_{\mathrm{BSOS},max}\left(\mathrm{BSOS}_{\max}|\mu\right) \tag{4.15}$$

The analytic formula of $F_{\mathrm{BSOS},max}$ is not known, and a numerical approximation is derived from Monte Carlo simulations. All the details on the approximation and fitting of the cumulative distribution are reported in Appendix C. To set a limit, one needs to find the $CL$ upper limit $\mu_{\mathrm{BSOS}}$ such that:

$$F_{\mathrm{BSOS},max}(\mathrm{BSOS}_{\max}|\mu_{\mathrm{BSOS}}) = CL \tag{4.16}$$

Although the analytic formula of $F_{max,(k)}$ for a fixed $n$ is available, it is not well-behaved, since it relies on the iterative difference of extremely large numbers (as $n$ increases), making it susceptible to catastrophic numerical cancellation when used on a computer. In order to avoid these problems, I computed the values of the function with high numerical precision on a suitable grid in order to construct a reliable monotonic cubic-spline interpolation [53] that can be used with default 64-bit floating-point arithmetic. Currently, these interpolations have been tabulated up to $n = 1000$ and allow the estimation of event rates up to $\mu \lesssim 800$, correspondingly limiting the number of events it is possible to analyse. The speed-up tables and the code used to compute the test statistic and the limits are available through the SpacingStatistics.jl [47] package for Julia.

### 4.3.4. Product of Complementary Spacings

Finally, I propose another test statistic that combines spacings between events regardless of their relative location. The stepping stone of this proposal is a test first proposed by Moran [35] which consists of the product of all the spacings between consecutive events Eq. 3.12:

$$M(n) = -\sum_{i=1}^{n+1} \log(S_i).$$

This statistic was proposed as a goodness-of-fit test sensitive to clusters of data against the null-hypothesis of a Uniform distribution: the presence of small spacings will drive the whole product of spacings towards more extreme values.

In order to make this test sensitive to the presence of large spacings, we consider the complements of each spacing and take their product:

$$C(n) = -\sum_{i=1}^{n+1} \log(1 - S_i) \tag{4.17}$$

The distribution of this quantity for a fixed number of events $n$, $F_C(C|n)$, is not known analytically, but I derived a numerical approximation based on Monte Carlo simulations and tabulated them for $n \leq 1000$. Additionally, since the definition of the test is of the form $\sum g(S_i)$, it is possible to derive its asymptotic distribution, as described in Sec. 2.6.3, using Darling's [17] or LeCam's [18] theorems. The asymptotic distribution of $C(n)$ as $n \to \infty$ I derived is:

$$f_C(C|n \to \infty) = \mathcal{N}(n \cdot \mu_\infty,\, n \cdot \sigma_\infty) \tag{4.18}$$

where the parameters are given by:

$$\mu_\infty(x) = e^{-x} \left[\mathrm{E}_1(-x) - 2i\pi\right] \tag{4.19}$$

$$\sigma^2_\infty(x) = e^{-x} \left[2A(x) - 4i\pi B(x) - 2xe^{-x}\mathrm{E}_1(-x)\right] - 1 - (x^2+1)e^{-2x}C(x) \tag{4.20}$$

where:

$$A(x) = xF_{HG}\left(\frac{[1,1,1]}{[2,2,2]}; x\right) + \frac{\ln^2(-x)}{2} + \gamma\ln(-x) + \frac{\pi^2}{12} + \frac{\gamma^2}{2} \tag{4.21}$$

$$B(x) = \gamma + \ln(x) + x \tag{4.22}$$

$$C(x) = \left[e^{-x}\mathrm{E}_1(-x)\right]^2 + 4i\pi e^{-x}\mathrm{E}_1(-x) - 4\pi \tag{4.23}$$

with $\mathrm{E}_1$ being the scaled exponential integral function, $F_{HG}$ the hyper-geometric function and $\gamma$ the Euler–Mascheroni constant.

Given this result, we can use it to estimate the test statistic for any large value of $n$ and effectively extend the applicability of this test and its limit calculation to large numbers of events. The Poisson-averaged p-value of this test for a given event rate $\mu$ is simply:

$$F_{C,Pois}(C|\mu) = \sum_{n=1}^{\infty} F_C(C|n) \cdot \frac{\mu^n e^{-\mu}}{n!} \tag{4.24}$$

Using this formula, one finds the $CL$ upper limit $\mu_C$ such that:

$$F_{C,Pois}(C|\mu_C) = CL \tag{4.25}$$

The tabulated distributions as well as the code used to compute the test-statistic and the limits are available through the SpacingStatistics.jl [47] package for Julia.

## 4.4. Performance comparison

The performance of the proposed methods was extensively studied using simulated examples, where backgrounds of varying shapes and strengths were introduced.

In the following, I compare the new methods described above methods against the standard Poisson test and the Optimum Interval method, which is considered the state of the art for setting limits in experiments affected by unknown backgrounds.

### 4.4.1. Background-free experiment

To begin with, consider the case in which no added background contaminates the experiment, in order to estimate the baseline of the different methods. For simplicity, I chose a uniform distribution for the generation of the events, which coincides with the null-hypothesis of all tests, and I varied the event rate used in the data generation. Fig 4.3 shows the median of the $CL = 0.90$ upper limits on the event rate set using different methods. In order to better discern differences between the efficiency of each method, we can look at the results normalized to the Poisson limit, as shown in Fig. 4.4. We notice that in this baseline scenario, the Poisson test is the best of the bunch, setting the lowest upper limits, as expected. Nevertheless, the results of the Poisson test do not drastically outperform any other test, showing that they all consistently estimate the true rate of events.

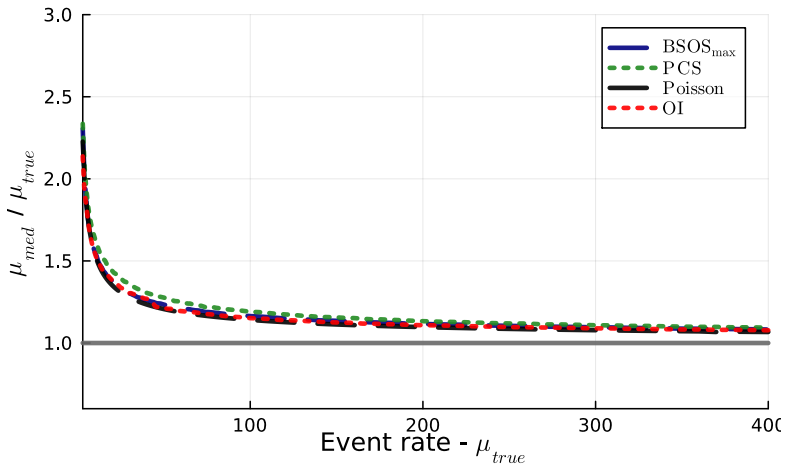



Figure 4.3.: Median $CL = 0.90$ upper limit normalized to the event rate used in the simulations; data generated according to a uniform distribution (background-free).

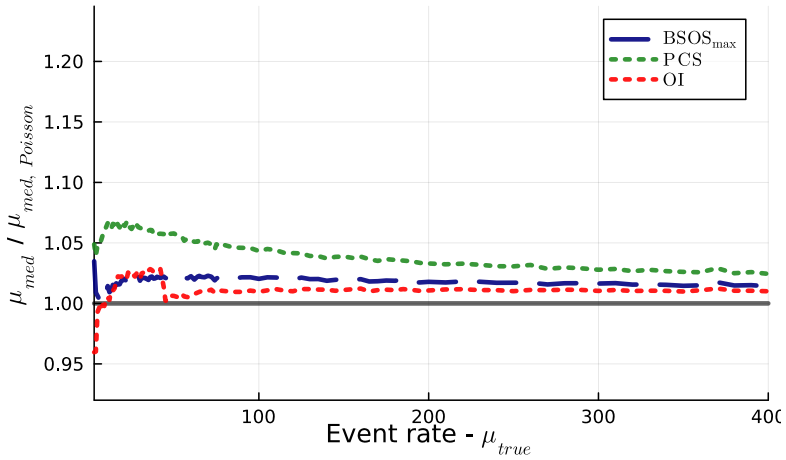



Figure 4.4.: Median $CL = 0.90$ upper limit normalized to the Poisson-test's result; data generated according to a purely Uniform distribution (background-free).

### 4.4.2. Exponential background-only experiment

Next we investigate the case in which a background is present in the simulations and the signal strength is negligible in comparison: this mimics a rare process search in which the signal might be absent. In the simulated experiments, I produce data directly in the cumulative space (hence the signal distribution is always assumed to be flat, i.e. the null-hypothesis) and I first consider an exponential background with rate 0.1, truncated on the unit interval $[0, 1]$. In dark matter search experiments, it is often the case that the distribution of events, after transforming to the cumulative space, is peaked at one end of the analysis window with rapidly decaying tails. Thus, the chosen

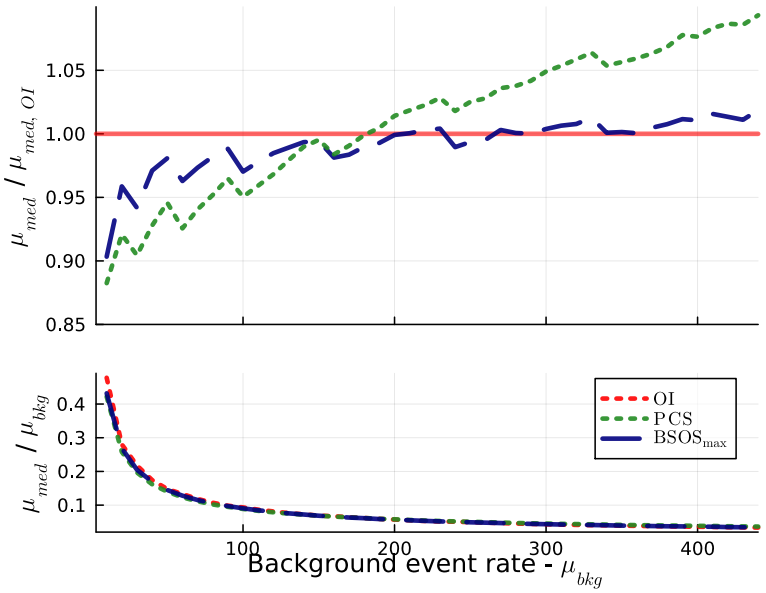



Figure 4.5.: Median $CL = 0.90$ upper limits for data generated according to an Exponential distribution of rate 0.1 (only-background): absolute results (bottom) and results normalized to the Optimum Interval limit (top).

exponential background example is representative of such a case. Fig. 4.5 reports the median $CL = 0.90$ upper limits of the measured event rate normalized to the injected background event rate (bottom), as well as the ratio of the results of all methods (minus Poisson) to the Optimum Interval test (top). The Poisson limit was omitted as it would simply scale with the total number of events, producing a correspondingly large upper limit. Analysing the results, we notice that when dealing with relatively peaked event distributions, all methods perform similarly, just like the background-free case. All methods are able to filter out most of the background contribution and reconstruct small overall event rates. For small injected background rates ($\leq 100$), the non-local methods (Sum of sorted spacings and product of complementary spacings) are able to set up to $5 - 10\%$ more stringent limits. As the background rate increases ($\geq 200$), the performance of the Sum of sorted spacings' test matches the Optimum Interval's one, while the Product of spacings's results are up to $5 - 10\%$ worse.

### 4.4.3. Mixing background and signal

Now we investigate the case in which a detectable signal distribution is contaminated by an unknown background. In the simulated experiments, I consider a background-to-signal ratio of 5, i.e the rate of background events is 5 times larger than the rate of signal events. Since we operate directly in the cumulative space, we always assume a uniform signal distribution. The total support of the background event distribution spans a quarter of the analysis window (meaning that the background events occupy a quarter of the

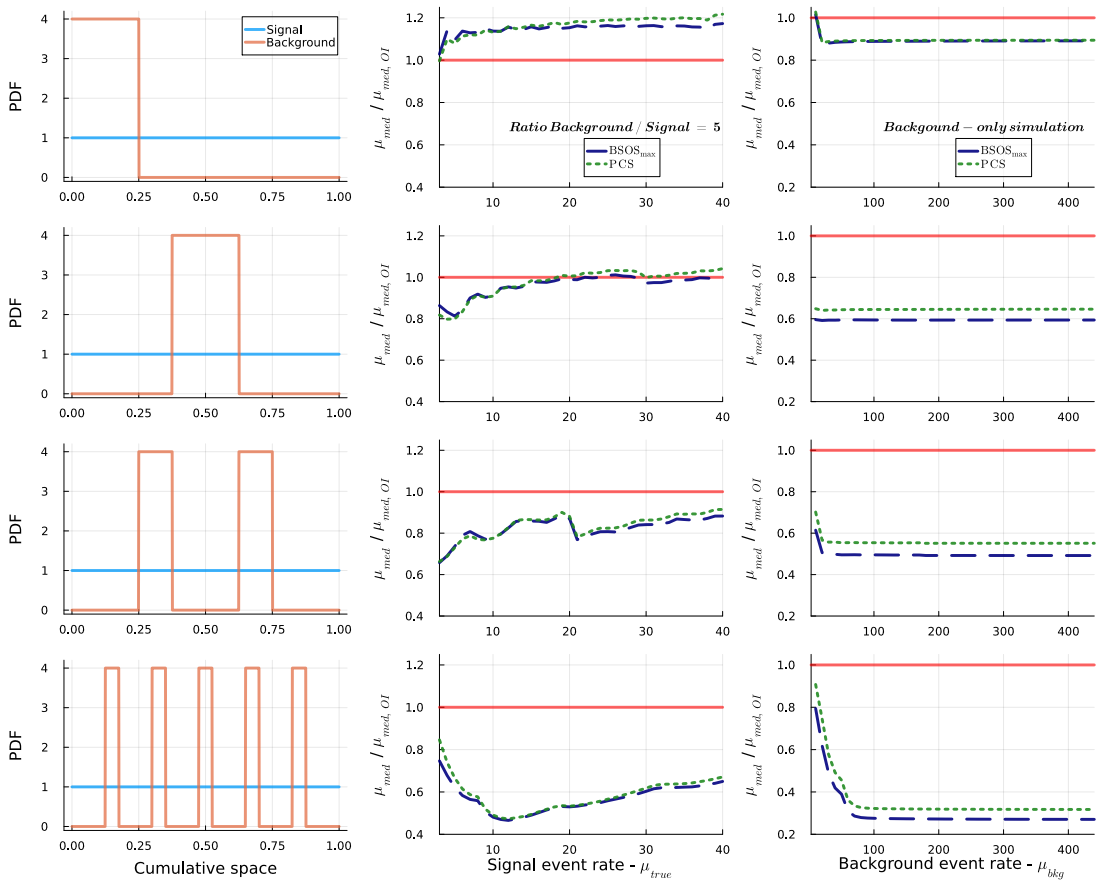



Figure 4.6.: Comparison of limit-setting methods depending on event distribution in the cumulative space: (left column) background and signal event distribution; (middle column) median $CL = 0.90$ upper limit normalized to the Optimum Interval's result for simulated event distributions with a background and signal mixing of $\mu_{bkg}/\mu_{sig} = 5$; (right column) median 90% CL upper limit normalized to the Optimum Interval's result for purely background-like event distributions ($\mu_{bkg}/\mu_{sig} = \infty$).

unit interval) but the shape of the distribution is varied: specifically, the background distribution is a mixture of one or more uniform distributions whose total width sums up to 0.25.

Fig. 4.6 shows different choices of background distributions in the left column and the resulting median $CL = 0.90$ upper limits obtained with different methods (normalized to the Optimum Interval's result) in the central column.

If the background distribution is fully concentrated in one region, localised at either end of the analysis window, as shown in the first row of Fig. 4.6, this creates an uninterrupted low-density region of the resulting event distribution. This is the best case scenario for the Optimum Interval method, as previously discussed. This expectation is reflected in

the results, where the Optimum Interval method's results are up to 20% better than the other methods.

As the background distribution is moved to the middle of the analysis window, or even split into two or more peaks, then we notice how the proposed tests are more sensitive, being able to set more competitive limits. For a bimodal background distribution, it is possible to set limits 20% lower than the Optimum Interval method on average, while the gain rises up to 40% for a pentamodal background distribution. The performance of the proposed tests (central column of Fig. 4.6) is due to their sensitivity to all large regions of low event density, regardless of their number or location, while the Optimum Interval method can only choose the most significant, ignoring the others.

The case of multimodal background distributions, especially when it presents well-defined and relatively narrow peaks, is interesting since it is similar to experimental scenarios in which the event rate of a three or multi-body decay is sought after: the expected spectrum of such a decay is relatively flat and could be contaminated by peaking background distributions which are representative of processes with a Standard-Model counterpart. The $\text{BSOS}_{\text{max}}$ and PCS methods would be well-suited to tackle these problems since they are able to filter out the contributions coming from these "peaks" and estimate the underlying "flat" event rate, without introducing additional parameters in the analysis (biasing the result) to modify, limit or segment the Region-of-Interest in order to exclude peaking backgrounds.

#### Vanishing signal

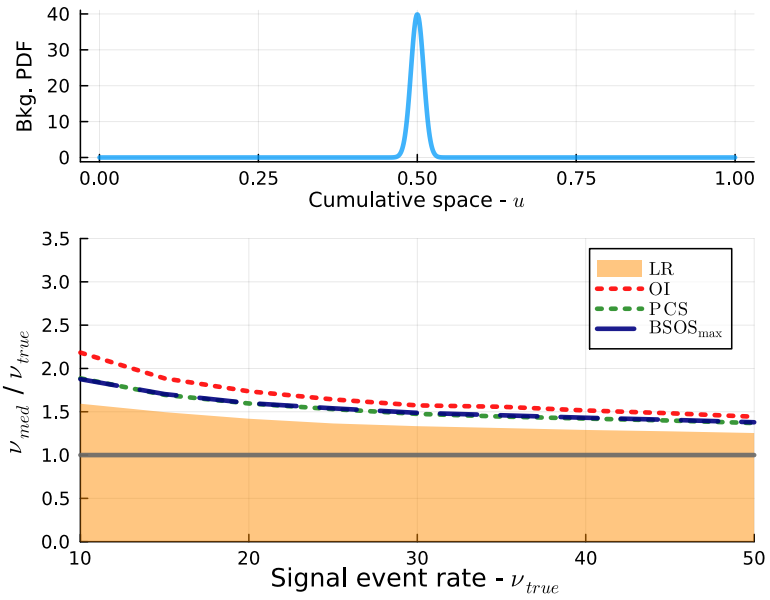
Considering the case of a very faint or absent signal, we can analyse the resulting limit if events were distributed only according to the background distribution. The results of these simulations are shown in the right column of Fig. 4.6, where we notice that, regardless of the shape of the background distributions, the 90% CL median upper limits of the $\text{BSOS}_{\text{max}}$ and PCS statistics are always smaller than the Optimum Interval counterpart, with limit gains increasing up to a factor of 3 as the number of event-free regions increases.

### 4.4.4. Comparison to a Likelihood-Ratio Test

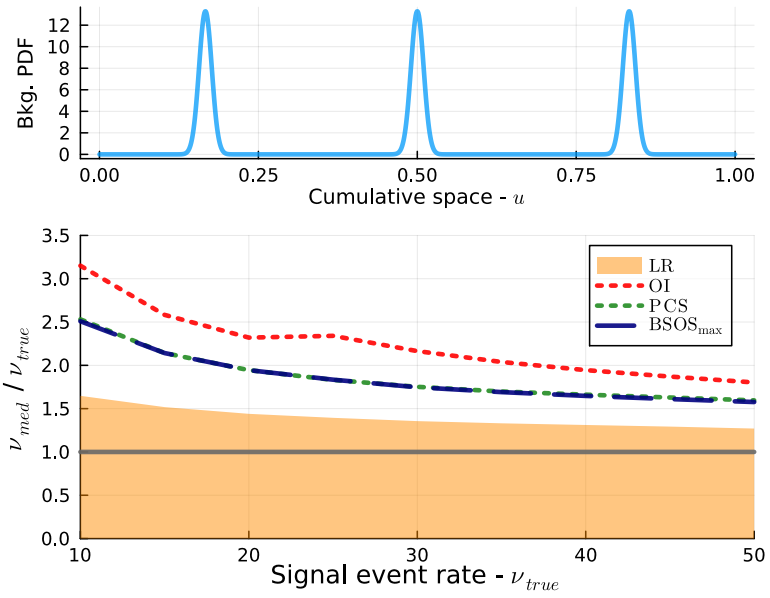
As a final example, I compare the efficiency of the non-parametric tests discussed so far against a Likelihood-Ratio (LR) test in the case of peaking backgrounds. I consider a mixture of Gaussian backgrounds with an associated event rate ten times stronger than the signal's event rate, $\nu_{true}$, which is varied from 10 up to 50.

Assuming $m$ distinct Gaussian background peaks are present, the distribution of events can be expressed as a mixture model:

$$H(\nu_{sig}, \boldsymbol{\nu}_{bkg}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{\nu_{sig} \cdot \mathcal{U}(0,1) + \sum_{j=1}^{m} \nu_{bkg,j} \cdot \mathcal{N}(\mu_j, \sigma_j)}{\nu_{sig} + \sum_{j=1}^{m} \nu_{bkg,j}} \tag{4.26}$$




(a)




(b)

Figure 4.7.: Background distribution in the cumulative space: mixture of Gaussian distribution with $\sigma = 0.01$; Median $CL = 0.90$ upper limit normalized to the signal event rate.

where $\nu_{sig}$ is the signal event rate, $\boldsymbol{\nu}_{bkg} = \{\nu_{bkg,1}, ..., \nu_{bkg,m}\}$ are the background event rates and $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the locations and standard deviations of the Gaussian peaks respectively. Given $H(\nu_{sig}, \boldsymbol{\nu}_{bkg}, \boldsymbol{\mu}, \boldsymbol{\sigma})$, the corresponding unbinned extended likelihood is simply the product of the probability density function evaluated at each sample $u_i$ times the Poisson probability of the observed number of samples:

$$\mathcal{L}(\nu_{sig}, \boldsymbol{\nu}_{bkg}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{\nu_{tot}^n e^{-\nu_{tot}}}{n!} \prod_{i=1}^{n} \left[ \frac{\nu_{sig} + \sum_{j=1}^{m} \nu_{bkg,j} \cdot f_{\mathcal{N}}(x_i|\mu_j, \sigma_j)}{\nu_{tot}} \right] \tag{4.27}$$

where:

$$\nu_{tot} = \nu_{sig} + \sum_{j=1}^{m} \nu_{bkg,j}. \tag{4.28}$$

To find a limit on the signal event rate, consider the reduced model $H(\boldsymbol{\nu}_{bkg}, \boldsymbol{\mu}, \boldsymbol{\sigma}|\nu_{sig} = \nu_0)$, where the value of $\nu_{sig}$ is fixed to a specific value $\nu_0$. Given these two models, the LR test statistic $\rho$ is defined as:

$$\rho(\nu_0) = -2 \log \left[ \frac{\sup \mathcal{L}(\boldsymbol{\nu}_{bkg}, \boldsymbol{\mu}, \boldsymbol{\sigma}|\nu_{sig} = \nu_0)}{\sup \mathcal{L}(\nu_{sig}, \boldsymbol{\nu}_{bkg}, \boldsymbol{\mu}, \boldsymbol{\sigma})} \right] \tag{4.29}$$

which is a test statistic that depends on the value of $\nu_0$. The distribution of $\rho(\nu_0)$ converges asymptotically to a $\chi^2_1$ distribution with one degree of freedom for any value of $\nu_0$. Given the relatively large number of samples involved ($n \sim 11 \cdot \nu_{true}$) and the small number of degrees of freedom of the asymptotic distribution, it would be reasonable to assume that $\rho(\nu_0)$ already follows a $\chi^2_1$ distribution. This assumption was verified numerically for different values of $\nu_0$ and allows us to estimate the 90% limits of the signal event rate, which are the values of $\nu_0$ such that $F_{\chi^2_1}[\rho(\nu_0)] = 0.9$. This usually results in two distinct values of $\nu_0$ that satisfy the equation, defining an interval $[\nu_{sig}^{min}, \nu_{sig}^{max}]$ that contains the true signal event rate with a 90% confidence level. In practice, $\nu_{sig}^{min}$ is almost always 0, or very close to being zero, hence $\nu_{sig}^{max}$ is the de-facto 90% upper limit.

Fig. 4.7 shows two examples, one in which the background comprises a single Gaussian peak (top) and another comprising three Gaussian peaks. Given these backgrounds, the median $CL = 0.90$ upper limits are obtained at different signal event rates. Inspecting the results, we notice that the most stringent limits are set by the LR method in all cases, which is hardly surprising since partial information of the background (shape and number of peaks) was folded into the analysis. The limits set by the non-parametric spacings-based tests, although more conservative, are still close enough to the LR ones, even without any assumption on the background shape. In the unimodal background case of Fig. 4.7 (top), spacings-based methods provide results no more than 20% larger than the LR for low event rates and up to 10% larger limits for higher event rates, with $\text{BSOS}_{\text{max}}$ and $PCS$ methods being slightly ahead of the Optimum Interval. In the

multimodal background case of Fig. 4.7 (bottom), the results provided by the $\mathrm{BSOS}_{\mathrm{max}}$ and $PCS$ limits range from being 50% higher than the LR one for low event rates, down to being 25% larger for higher $\nu_{true}$. The difference between the limits of the Optimum Interval method and the LR approach, in this case, is roughly twice as much as the other non-parametric tests, which is due to the presence of multiple background modes that split up the low event density regions in the analysis window.

# 5. Goodness-of-fit tests for arbitrary multivariate models

## 5.1. Introduction

So far I have discussed the use of goodness-of-fit tests as a discovery tool, in order to detect an unknown signal against a known background, as well as a tool to set limits on the event rate of proposed signal distributions in experiments contaminated by poorly understood backgrounds. All the models and data considered so far were univariate. Often enough, experiments collect multivariate samples, and having to resort only to one dimension in order to quickly sieve through their large datasets can be reductive.

In the following, I discuss how to build goodness-of-fit tests for arbitrary multivariate distributions or multivariate data generation models. The resulting tests perform an unbinned analysis and do not need any trials factor or look-elsewhere correction since the multivariate data can be analyzed all at once. The proposed distribution or generative model is used to transform the data to an uncorrelated space where the tests are developed. Depending on the complexity of the model, it is possible to perform the transformation analytically or numerically with the help of a Normalizing Flow algorithm.

The flexibility of targeting vastly different univariate distributions is made possible by the probability integral transformation [2, 3]. Building upon the univariate case, I start discussing how to perform this transformation in the multivariate case. I then discuss different ways of performing a multivariate uniformity test and how to adapt this tool in the case of signal discovery or setting upper limits.

Finally, I consider examples for each application in order to test the sensitivity of these methods.

The content of this chapter closely follows [54], where these results were first presented.

## 5.2. Multivariate probability integral transformation

In the univariate case, the probability integral transformation allows to develop tests in a standardized environment, where the null-hypothesis is represented by the standard Uniform distribution $\mathcal{U}(0,1)$. Concretely, given $m$ i.i.d. samples $x_i$ and a continuous distributions $f(x)$ with cumulative $F(x)$, we transform the samples onto the unit interval $[0,1]$ via $u_i = F(x_i)$.

Much like the univariate case, the goal with multivariate samples (in $n$ dimensions) is to develop uniformity tests in the unit hyper-cube $[0,1]_n$. In order to target any given multivariate distribution $\mathbf{M}$, we need to transform the probability space described by $\mathbf{M}$

into $[0,1]_n$. This transformation can be easy or difficult depending on the distribution $\mathbf{M}$, specifically, depending on the correlation among the dimensions of $\mathbf{M}$. In the following I show how to perform the transformation into the unit hyper-cube in three main cases: first, distributions comprised of uncorrelated dimensions are considered, moving then to distributions with correlated dimensions or sample-generating processes for which a probabilistic model is not available and finally hierarchical models are discussed.

### 5.2.1. Independent dimensions

If the dimensions of the proposed distribution $\boldsymbol{M}$ are all independent of each other, then $\boldsymbol{M}$ is just a composition of $n$ independent univariate distributions:

$$\boldsymbol{M} = [M_1, M_2, .., M_n] \tag{5.1}$$

where $M_j$ is the distribution of the $j$-th dimension. Much like the univariate case, it is possible to transform the $j$-th component of each sample using the corresponding cumulative distribution function $F_{M_j}$. Thus, the transformation of sample $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, .., x_{i,n})$ in $[0,1]_n$ is simply:

$$\boldsymbol{u}_i = [u_{i,1}, u_{i,2}, .., u_{i,n}] = [F_{M_1}(x_{i,1}), .., F_{M_n}(x_{i,n})] \tag{5.2}$$

### 5.2.2. Correlated dimensions and generative models

If the dimensions of the distribution to be compared to the data are not mutually independent, then it might be difficult to write down a transformation to the hyper-cube. This is still possible when dealing with nicely behaved distributions, such as a multivariate Normal distribution whose covariance matrix is not diagonal, but that might not be the case for a more complex distribution, such as a weighted sum of distributions. In such cases, it is possible to learn the transformation to the unit hyper-cube by using a Normalizing Flow (NF) which can perform a whitening of the distribution; i.e., transform the distribution so that it becomes a diagonal multivariate Normal distribution in the new coordinates. Once the original distribution is transformed in this way, it is then possible to further transform it to the unit hyper-cube one component at a time as shown earlier.

The Normalizing Flow (NF) is made up of a Neural Network which is trained using samples from the proposed distribution $\boldsymbol{M}$. The samples needed for training can be obtained from an associated generative model or by sampling $\boldsymbol{M}$ using a Markov chain Monte Carlo. The use of the generative model is particularly interesting because it allows to train the NF without having a normalized distribution or any model at all. In such cases, the NF is learning the associated distribution and the transformation all at once. [55, 56] offer a nice review of the theory and some of the many applications of Normalizing Flows. I briefly summarize this topic in Appendix D, where I report some

details of our NF implementation, developed in collaboration with M. Dudkowiak during his Bachelor thesis work, as well as with other collaborators from our group. In order to show the feasibility of this approach, a proof of principle example is presented where a Normalizing flow is used to whiten data sampled from a sum of three two-dimensional Normal distributions. A sampled distribution is depicted in Fig. 5.1 and the resulting marginal distributions of the whitened samples are shown in Fig. 5.2. The Normalizing Flow used for this example was adapted from [56].

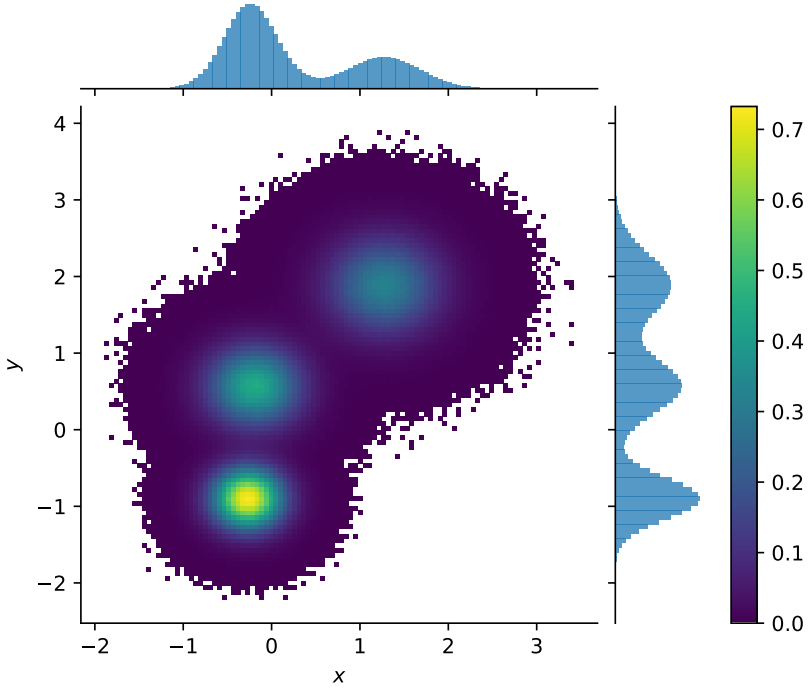



Figure 5.1.: Sample distribution of the sum of three two-dimensional Gauss distributions.

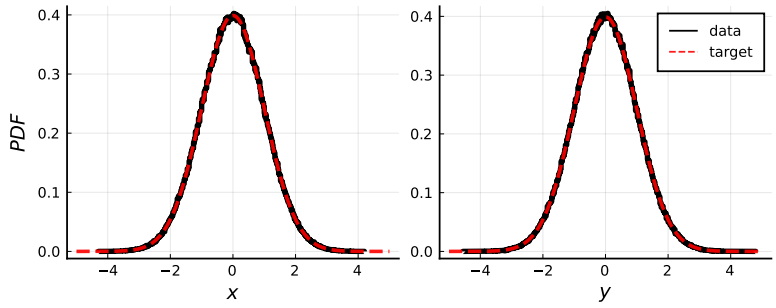



Figure 5.2.: Whitened marginal distributions after transforming with the Normalizing Flow.

### 5.2.3. Hierarchical models

Given a hierarchical model, the distribution of some components of the data is dependent on the values of other components, which are referred to as hyper-parameters of the model. If the hyper-parameters are mutually independent or if a transformation to the unit hyper-cube is available for their distribution and if the same is true for all the dependent parameters at each layer of depth of the hierarchical model, then it is possible to transform the whole distribution into the unit hyper-cube in stages.

Consider for example a 2 layer hierarchical model producing distributions $\boldsymbol{M} = [\boldsymbol{M}_1, \boldsymbol{M}_2(\boldsymbol{M}_1)]$. $\boldsymbol{M}_1$ models the distribution of the hyper-parameters $\boldsymbol{x}^{high}$ of the model and these components can be transformed to the corresponding uniform unit hyper-space using the associated function $T_{\boldsymbol{M}_1}$. The distribution of the dependent parameters $\boldsymbol{x}^{low}$ is affected by the observed value of the hyper-parameters $\boldsymbol{x}^{high}$:

$$\boldsymbol{x}_i^{low} \sim \boldsymbol{M}_2(\boldsymbol{x}_i^{high}) \tag{5.3}$$

For any given sample $\boldsymbol{x}_i$, the value of the hyper-parameters $\boldsymbol{x}_i^{high}$ is fixed, so the distribution $\boldsymbol{M}_2(\boldsymbol{x}_i^{high})$ is fully defined and it is possible to compute the corresponding transformation to the unit hyper-space. While $T_{\boldsymbol{M}_1}$ is sample-independent, $T_{\boldsymbol{M}_2}$ is sample-dependent. In case of hierarchical models with more layers, this staged transformation approach is to be repeated for each layer.

## 5.3. Uniformity tests in the unit hyper-cube

In the following, I discuss various methods that allow performing a multivariate uniformity test by reducing this task to a series of univariate uniformity tests. These tests are sensitive to non-uniformities in a transformed dataset and their application is twofold:

- detection of clustering of events against a uniform background, in a discovery scenario
- estimation of the upper limit on the rate of events corresponding to the uniform component of the data, representative of a proposed signal, against unknown backgrounds

For the latter, a desired confidence level is set in advance.

### 5.3.1. Projection - Discovery

Assume we have $m$ samples within a unit hypercube $\{\boldsymbol{u}_i\} \in [0,1]_n$. The $n$ components of each sample are assumed independent of one another after the necessary transformations. The projections of the samples along each axis of the hyper-cube, therefore, yield $n$ univariate uniformly distributed sets of data: $\{u_{i,j}\}$ for the $j$-th dimension. For each one of these projected datasets, $\{u_{i,j}\}$, it is possible to perform a uniformity test using a

test statistic of choice and condense the information for the $j$-th dimension in one scalar p-value $p_j$. A simple two-dimensional depiction of these projections is shown in Fig. 5.3.

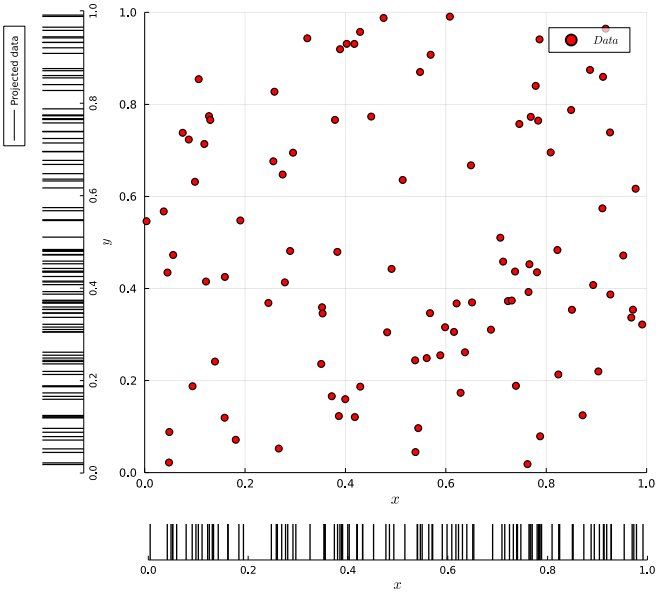



Figure 5.3.: Dataset consisting of two-dimensional samples distributed uniformly in the hypercube $[0,1]_2$ and projections.

Given our assumptions, the expected distribution of each p-value $p_j$ is uniform, and moreover, the p-values will be independent of one another. On this resulting dataset, $\{p_j\}$, it is possible to perform a uniformity test using a test statistic of choice in order to check whether there are any significant deviations from uniformity. The result of this last uniformity check results in one last p-value $p_{final}$ which is the overall p-value of the multivariate goodness-of-fit test.

As pointed out in the discussion above, in order to obtain the intermediate p-values, $\{p_j\}$, and then the final one, $p_{final}$, it is possible to use any test statistic of choice, as long as the chosen statistics preserve the non-correlation among dimensions (results from tests that have a Poisson dependent factor, for example, will be correlated, since the same number of samples is projected on all dimensions). What is important is that the distribution of the resulting p-values is uniform. This implies that the test statistic used for the evaluation of the intermediate p-values, $\{p_j\}$, does not have to be the same as the one used to evaluate $p_{final}$; as a matter of fact one could also use different tests for different dimensions in the evaluation of $\{p_j\}$, but it might be a more consistent approach to consider all dimensions equally and use the same test for all projections.

In the previous discussion, we considered a dataset of $m$ samples $\{\boldsymbol{u}_i\} \in [0,1]_n$. In such a case, if the number of events $m$ is large, it might be appropriate to use a test such as $\mathrm{BSS}_{\mathrm{min}}$, RPS or KS in order to pick up a signal in any of the projections. Afterwards, when considering the $n$ p-values $\{p_j\}$, it could be better to look for outliers, since already one of a few small $p_j$ could be indicative of the presence of a signal in the data. In this

case, especially when dealing with low-dimensionality spaces ($n$ small), instead of using RPS or KS on the set $\{p_j\}$ it might be more informative to look at the smallest p-value or rather their product in case we want to improve the sensitivity in the presence of multiple small p-values.

#### Minimum p-value

As discussed, observing one small p-value might already be enough to point to a possible signal in the data. Under the assumption of a uniform distribution of $\{p_j\}$, the distribution of $p_{min} = \min\{p_j\}$ is simply the first Order Statistic, and it follows a Beta distribution:

$$p_{min} = \min_j\{p_j\} \sim \text{Beta}(1, n) \tag{5.4}$$

where $n$ is the dimensionality of the original data. Thus the final p-value is:

$$p_{final} = n \cdot (1 - p_{min})^{n-1}. \tag{5.5}$$

#### Product of p-values

Given more than one small p-value $p_j$, looking only at the smallest one might be reductive and we could gain in sensitivity by combining the small p-values together. One way of doing so is to consider the product of all p-values:

$$p_{prod} = \prod_{j=1}^{n} p_j \tag{5.6}$$

Once again, we expect all $\{p_j\}$ to be uniformly distributed, and the distribution of $p_{prod}$ is known [57]:

$$\Pr\{p_{prod} = x; n\} = \frac{(-1)^{n-1}}{(n-1)!} [\ln(x)]^{n-1} \tag{5.7}$$

thus the final p-value ,$p_{final}$, is:

$$p_{final} = p_{prod} \cdot \sum_{j=1}^{n} \frac{(-1)^{j-1}}{(j-1)!} [\ln(p_{prod})]^{j-1} \tag{5.8}$$

### 5.3.2. Projection - Limit setting

Several spacings-based tests have been developed for this task in the univariate case, as we discussed in Ch. 4, but for the multivariate case, I do not know of any method being available.

It is noteworthy that Yellin reports in [52] an idea regarding the extension of the Optimum Interval method to multiple dimensions, by looking at the hyper-rectangular volumes containing a fixed number of samples. This proposal is computationally costly in calculating the test statistic value, and even more so in tabulating results for a large number of samples. The author hints at the possibility of developing asymptotic results in order to lessen the burden of the cumulative distribution parametrization. Although this idea has been presented, as far as I know, it was not developed further or implemented.

In this work, I decided to take a different approach from the one outlined by Yellin. Instead of tackling the problem in all its complexity, by considering hyper-volumes and the number of samples that might occupy them, we might rely on a simplification of the problem, by considering the projections of samples on the axis of the hypercube. This allows us to reduce the complexity of the problem and to obtain a solution, even at the cost of this method potentially being less powerful than a direct $n$-dimensional approach.

As discussed before, given $m$ uniformly distributed samples $\{\boldsymbol{u}_i\} \in [0,1]_n$, we consider the projection of the samples on the $n$ axes, knowing these will be uniformly distributed as well. For each one of these projected datasets $\{u_{i,j}\}$ it is possible to estimate an upper limit $\mu_j$ on the event rate with confidence level $L_1$.

Out of the upper limits $\{\mu_j\}, j = 1, .., n$ obtained from each projection, we can use a best-of-the-bunch approach and select the smallest one as the final limit:

$$\mu_{final} = \min_j \{\mu_j\} \tag{5.9}$$

At this point we must consider the confidence level $L_n$ associated with this estimate. If the projected limits $\{\mu_j\}$ were completely independent of one another, then we might consider that selecting the smallest limit amounts to a resulting confidence level $L_n$ equal to the product of $n$ Bernoulli variables with rate $L_1$, thus:

$$L_n = (L_1)^n \tag{5.10}$$

Under this assumption, we could easily select the confidence level $L_1$ of the individual projection limit estimations in order to ensure that $L_n$ is equal to a desired value.

This assumption, however, is not correct. Although the distribution of the projected events on each axis is independent, the number of samples projected on each axis is not: if there are $m$ samples in the multi-dimensional space then there will be $m$ samples on each projected dataset $\{u_{i,j}\}, j = 1, .., m$. In order to set a limit we consider both the distribution of events and the total number of events, merging a goodness-of-fit test with a Poisson test. Since all projected datasets $\{u_{i,j}\}$ share the same number of events, this introduces a correlation in the Poisson statistic part of each limit-setting estimation, rendering all resulting limits correlated.

Although the projection-independence assumption is not valid if applied after the Poisson-averaging, it is possible to calculate the corrections necessary to ensure the desired final confidence level $L_n$. We assume that $L_n$ is a function of the projection-specific

confidence level $L_1$ and that it is dependent on the value of the reconstructed limit $\mu_{final}$, for a given number of dimensions $n$: $L_n(\mu_{final}, L_1|n)$. If we seek a specific Confidence Level CL, then we need to find the value of $L_1$ that for the resulting best limit $\mu_{final}$ yields:

$$L_n(\mu_{final}(L_1), L_1|n) = \mathrm{CL} \tag{5.11}$$

This equation is just a one-dimensional root finding problem in $L_1$ which can be solved iteratively (for example using a Bisection algorithm) by estimating the error at $\mu_{final}(L_1)$ for a proposed value of $L_1$. The estimation of the error rate can be done via Monte-Carlo simulations, producing data according to a uniform distribution in the $n$-dimensional hypercube, since the Eq. 5.11 only needs to hold in this nominal case.

Although this procedure might seem complicated, it is easy to devise and can be performed well before any real analysis has to be run, during the method validation phase, allowing for the tabulation, interpolation and sharing of $L_n(\mu_{final}, L_1|n)$. I have calculated the exact correction for the $\mathrm{BSOS}_{\max}$ method and an approximate correction for the Optimum Interval method up to five dimensions.

### 5.3.3. Product of Complementary Spacings - Limit setting

#### Best projection

As discussed above, if one calculates the Poisson-averaged p-value on each projected dataset and then chooses the most significant value, a correction needs to be calculated to account for the correlation of these values due to the fixed number of samples on each axis. In order to avoid this problem, if the definition of the chosen test statistic allows it, it is possible to perform the selection of the best p-value before averaging with a Poisson distribution. In such a case it would be trivial to calculate the correct confidence level without having to resort to numerical corrections.

Consider the Product of Complementary Spacings (PCS) statistic from Eq. 4.17:

$$C(m) = -\sum_{i=1}^{m+1} \log(1 - S_i).$$

For each of the projected datasets, one can compute the corresponding value of the test $C_j$ and its p-value (here $p_j = F_C(C_j)$). The $n$ projected p-values, $\{p_j\}$, form order statistics with Uniform distribution. If we were to select the largest $F_C(C_j)$, its distribution would be simply:

$$\max_j\{F_C(C_j)\} \sim \mathrm{Beta}(n, 1). \tag{5.12}$$

Given the test-statistic values $C_j$ for each projection, the Poisson-averaged p-value of the largest one, $C_{max} = \max_j\{C_j\}$, is:

$$F_{C,Pois}(C_{max}|\mu) = \sum_{m=1}^{\infty} F_{Beta}\left[F_C(C_{max}|m)\ |n\right] \cdot \frac{\mu^m e^{-\mu}}{m!}. \tag{5.13}$$

It follows that the upper limit $\mu_{lim}$, with a confidence level CL, is such that:

$$F_{C,Pois}(T_{max}|\mu_{lim}) = \text{CL} \tag{5.14}$$

**Sum of projections**

Given the PCS test-statistic values $C_j$ on each projection, instead of selecting the largest, we can consider their sum:

$$C_{sum} = \sum_{j=1}^{n} C_j \tag{5.15}$$

which can be interpreted as a product of the product of complementary spacings. Assuming we know the distribution of $C_{sum}$ for a fixed number of events $m$, $F(C_{sum}|m)$, then we can compute the Poisson-averaged p-value of this test for a given event rate $\mu$:

$$F_{Pois}(C_{sum}|\mu) = \sum_{m=1}^{\infty} F(C_{sum}|m) \cdot \frac{\mu^m e^{-\mu}}{m!} \tag{5.16}$$

Given this definition, it is possible to invert the formula and find the upper limit on the event rate up to a desired confidence level. For example, the 90% confidence level upper limit $\mu_{lim}$ is such that:

$$F_{Pois}(C_{sum}|\mu_{lim}) = 0.9 \tag{5.17}$$

If $F(C_j|m)$ is known, it is rather easy to compute $F(C_{sum}|m)$. Since $C_j$ are all i.i.d., the distribution of $C_{sum}$ is just $f_{C,m}$ convolved $n-1$ times with itself:

$$f(C_{sum}|m) = \underbrace{f(C|m) * f(C|m) * .. * f(C|m)}_{n \text{ times}} \tag{5.18}$$

Since $F(C|m)$ has been tabulated in the Julia package SpacingStatistics.jl [47], and is available as a monotonic cubic spline polynomial function, it is possible to easily obtain its derivative $f(C|m)$, transform it to the Fourier space using a FFT, raise it to the power of $n$ and transform back to the real space to obtain $f(C_{sum}|m)$:

$$f(C_{sum}|m) = FFT^{-1}\left\{[FFT(f(C|m))]^n\right\} \tag{5.19}$$

This procedure is used for the tabulated $F_{PCS,m}$ ($m \leq 10^3$). For values of $m$ larger than $10^3$ one can use the asymptotic distribution of $F_{PCS,m}$, Eq. 4.18, which is a Gaussian distribution, thus rendering the calculation of the convolution much easier.

These two approaches show how to adapt the PCS test to a multivariate limit-setting scenario, similarly to how the minimum p-value and product of p-values were used in the multivariate discovery case. Although I discussed the PCS test specifically, these corrections apply in general to any test-statistic $T$ where the Poisson-averaging can be calculated as a final step.

### 5.3.4. Projection - Problematic configurations

Before moving on, I would like to point out a data configuration that might be difficult to analyse using projection methods.

Considering a simple two-dimensional sample distribution, such as the one shown in Fig. 5.4, we can see that if we were to take the projections of these samples along the $x$ and $y$ axis, these would all be uniform.

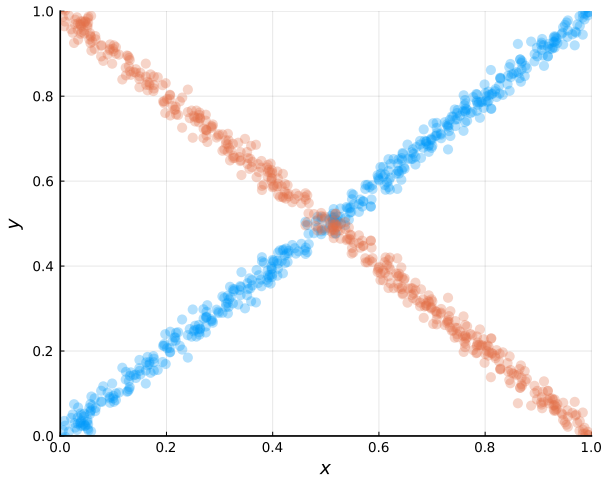



Figure 5.4.: Distributions of samples that are not easily detectable via projection on the axis.

We notice then that if served with such a distribution of data, we would not be able to distinguish it from a truly uniform one in the hypercube. For a signal to go undetected, it needs to be perfectly aligned along a diagonal of the hypercube, and this is a rather unlikely configuration. Nevertheless, in order to correct for this shortcoming, if visualization of the data is not possible, then one could think of testing the data twice: testing the original dataset (in the hypercube $[0, 1]_n$) then transforming the data into a Standard Multivariate Gaussian, perform a rotation and transform back into the uniform hypercube. Such a transformation would maintain the uniform distribution of data in the hypercube under the null-hypothesis, but would also break the alignment between possible problematic

signals and the hypercube's diagonals.
This operation produces two p-values, which will be correlated for small numbers of events $m$, but that will reach a stable distribution for large sample sizes. Out of these two p-values one could select the smallest and tabulate its distribution, in order to calculate from it a true p-value. This parametrization is not too computationally expensive and can be carried out in a preparatory step, before having to use the test on real data. Since this parametrization is transformation dependent, I stop here and do not produce examples, but I would like to bring to the reader's attention the need of considering such cases when using projection methods to perform goodness-of-fit tests.

### 5.3.5. Volume transformation

Finally, I consider a different dimensionality reduction strategy. Given $m$ samples $\{\boldsymbol{u}_i\} \in [0,1]_n$, instead of projecting them onto the axes and obtaining $n$ independent sets of univariate data, we can use a dimension-reducing transformation to map them all at once onto a single univariate dataset. To achieve this, calculate the volume contained in the hyper-rectangle defined by its projections simply by taking the product of its coordinates:

$$v_i = V(\boldsymbol{u}_i) = \prod_{j=1}^{n} u_{i,j} \ . \tag{5.20}$$

Calculating the volume this way for each multivariate sample, yields a simple univariate dataset: $\{\boldsymbol{u}_i\} \xrightarrow{V} \{v_i\}$. Since the $\{\boldsymbol{u}_i\}$ were i.i.d. samples, so are the $\{v_i\}$ (although not uniformly distributed). Since $v_i$ is the product of $n$ independent uniform variables, whose distribution is given by Eq. 5.7, its probability distribution is known:

$$\Pr\{v_i = x; n\} = \frac{(-1)^{n-1}}{(n-1)!} [\ln(x)]^{n-1}$$

Using the probability integral transformation, Eq. 5.8, we can therefore transform $\{v_i\}$ into a set of uniform i.i.d. samples $\{z_i\}$. We can then use these to perform a univariate uniformity test using a test statistic of choice; standard univariate discovery (Ch. 3) and limit-setting (Ch. 4) tests can be used in order to analyse the data.

#### Problematic configurations and origin selection

In the transformation described above, the volume calculation is not an injective function, meaning that multiple points in the hypercube $[0,1]_n$ can share the same volume, specifically, all points lying on one of the hyperboles shown in Fig. 5.5 (for a 2D example). This means that if all points lying on a given hyperbola were to be moved to a single point on the same hyperbole, this would not be detectable from the transformed set.

Additionally, in my studies, I also found out that the choice of the origin, for the volume calculation, might matter. Given a set of samples distributed according to a multivariate

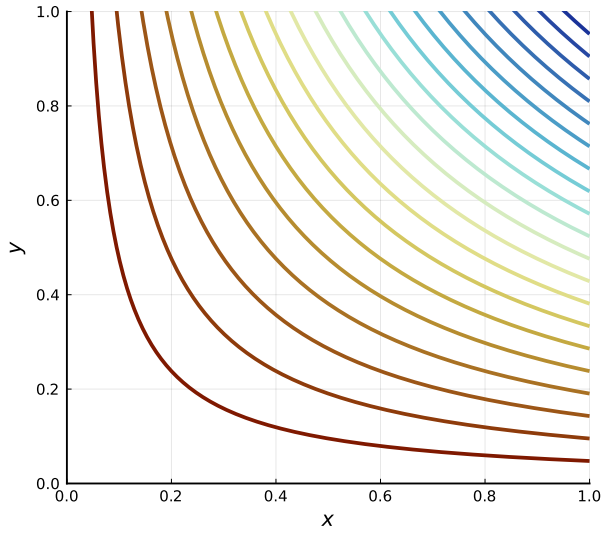



Figure 5.5.: Example of 2D surfaces (hyperbole) that share the same volume.

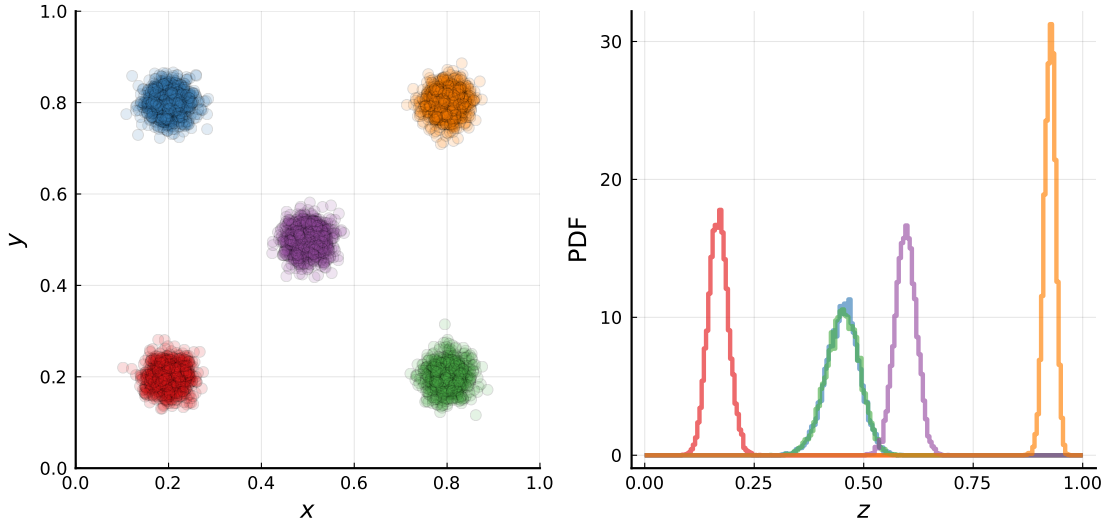



Figure 5.6.: Example of 2D Gaussian signal located in different positions inside the hypercube $[0, 1]_2$ (left) and the corresponding distribution of events after the volume transformation (right).

Normal distribution with a diagonal covariance matrix, the resulting distribution of events in the unit range $[0, 1]$ after the full volume transformation will depend on the location of the signal distribution in the hypercube. Considering a 2D example, shown in Fig. 5.6, we notice that signals located at the corners of the anti-diagonal (top-left and bottom-right positions) will have a wider distribution after the volume transformation compared to those present on the main diagonal, and of these, those furthest away from the origin yield the narrowest distribution after transforming. This will impact the

goodness-of-fit test, since narrower distributions tend to yield smaller p-values. If we were to change the vertex from which the volume is calculated, we would end up in the same situation, just shuffling the plots already shown.

Nevertheless, for a given choice of the origin used for the volume transformation, the resulting p-value is valid and correct, given that its distribution is flat under the null-hypothesis of a Uniform distribution of samples in the hypercube.

## 5.4. Example - $n$D Discovery

Here I illustrate how the proposed goodness-of-fit tests can be used in a scenario where a possible "new physics" model is searched for but it is not wished to specify how the new physics might populate the data space. It is then to be tested whether the data follows a known distribution, which is a "background" to a possible new signal. After having collected some data, one wants to quantify the goodness-of-fit of the background-only distribution to the data and a resulting low p-value could indicate the presence of events distributed according to an additional, previously unknown, signal distribution.

### 5.4.1. Multivariate Gaussian signal

In this example the background is modelled by a simple Uniform distribution in the 5-dimensional hyper-cube $[0, 1]_5$ and in order to illustrate how the presence of an actual signal (alternative hypothesis) would affect the outcome, additional events are injected, following a multivariate Normal distribution randomly positioned within the hyper-cube with an isotropic variance of either 0.01 or 0.1. The number of events is Poisson fluctuated for both background and signal populations, with respectively expected values of $\langle n_b \rangle = 10^4$ and expected values of $\langle n_s \rangle$ ranging up to $10^3$.

The p-value distributions under the assumption of $H^0$ (i.e. only background is present) are shown in Fig. 5.7: the results corresponding to the narrow signal ($\Sigma = I_5 \cdot 0.01$) are on the left column and those corresponding to the broad signal ($\Sigma = I_5 \cdot 0.1$) are on the right column; the first two rows present p-value distributions calculated using projection methods while the third row presents p-value distributions obtained with the volume transformation method; the fourth row presents the sensitivity of each test quantified as the median p-value for each distribution. Regarding the results of the projection method, the evaluation of the intermediate p-values was performed using the KS test, given the large count rates, while the evaluation of the final p-value, since there are only 5 dimensions, was performed using the two tests previously described, namely the minimum and the product of intermediate p-values, corresponding to the first and second rows respectively. Similarly, the KS test statistic was used in the final uniformity test after performing the volume transformation.

Distributions with no signal ($\langle n_s \rangle = 0$) show a flat p-value distribution, as expected. The distributions of trials with injected signals are trending towards smaller p-values, indicating the worsened goodness-of-fit for the background-only model. The distributions of trials where the signal has a smaller variance (left) are much more skewed towards

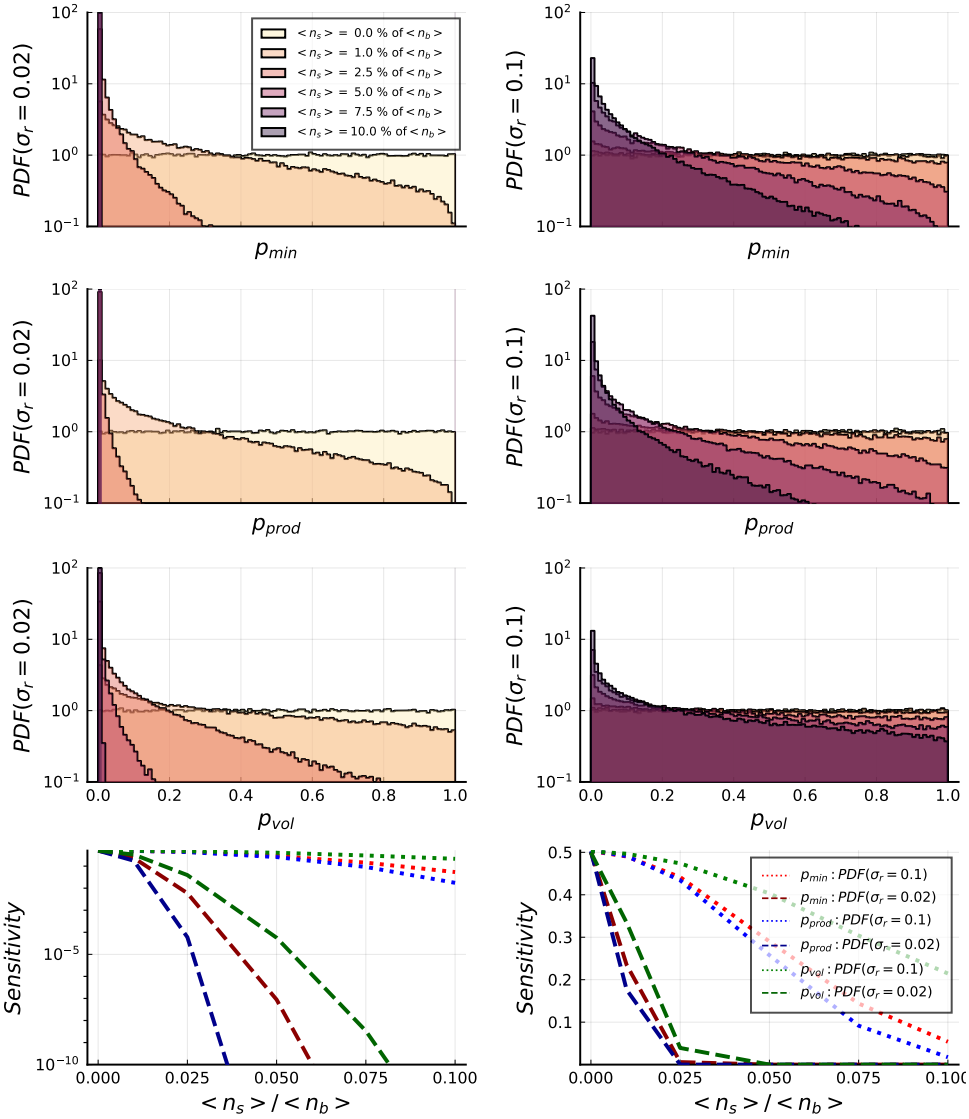



Figure 5.7.: Distributions of p-values for background only samples ($\langle n_s \rangle = 0$) and background plus randomised signal injections from a 5D Gaussian distribution: 'narrow' signal with random $\mu \in [0.2, 0.8]$, $\Sigma = 0.01 \cdot \mathrm{I}_5$ (left) and 'wide' signal with random $\mu \in [0.2, 0.8]$, $\Sigma = 0.1 \cdot \mathrm{I}_5$ (right) of varying strength; comparison to the background model for either the minimum p-value statistic (first row), the product of p-values statistic (second row) or the volume-transformed p-value (third row); median p-value (sensitivity) both in linear and logarithmic scale (fourth row).

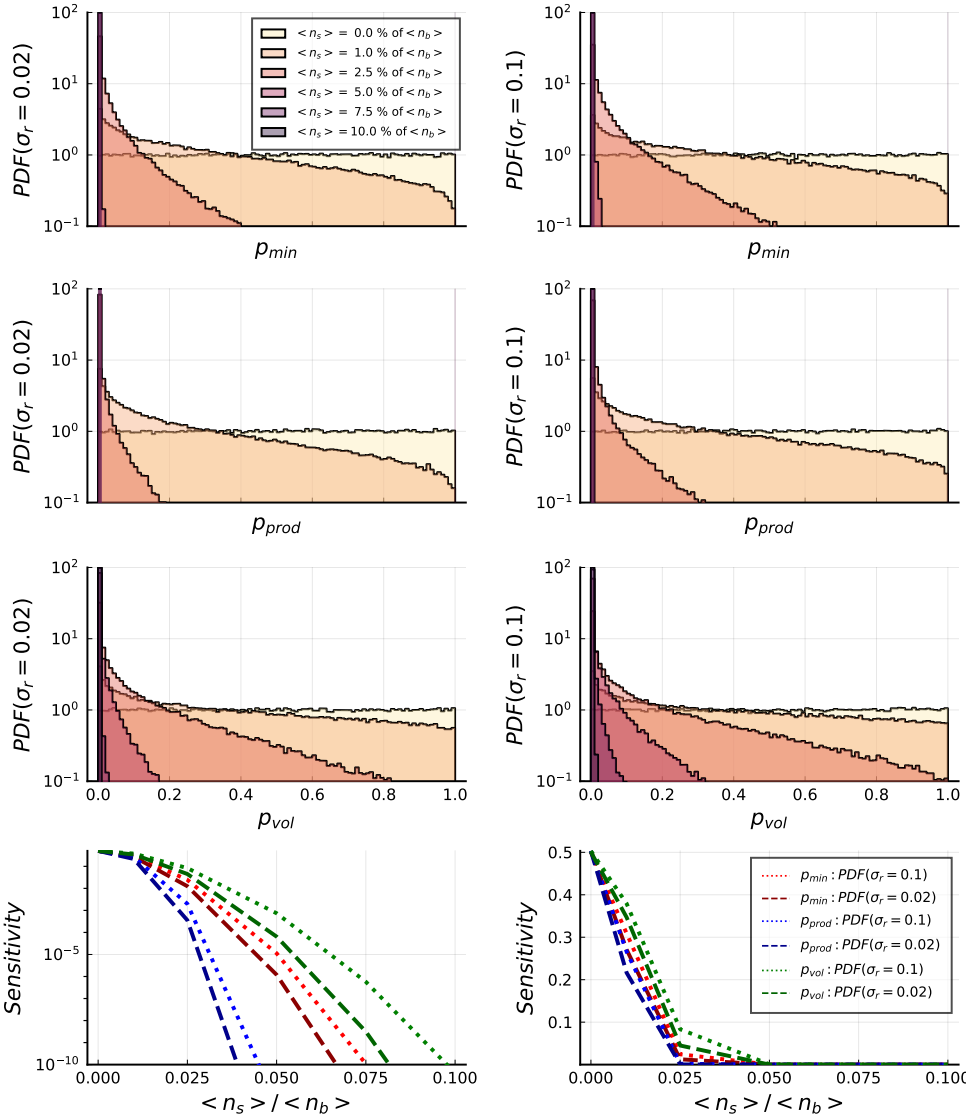



Figure 5.8.: Distributions of p-values for background only samples ($\langle n_s \rangle = 0$) and background plus randomised signal injections from a 5D Gaussian-shell distribution: 'narrow' signal with random $\mu \in [0.25, 0.75]$, $r = 0.25$, $\sigma_r = 0.02$ (left) and "wide" signal with random $\mu \in [0.25, 0.75]$, $r = 0.25$, $\sigma_r = 0.1$ (right) of varying strength; comparison to the background model for either the minimum p-value statistic (first row), the product of p-values statistic (second row) or the volume-transformed p-value (third row); median p-value (sensitivity) distribution both in linear and logarithmic scale (fourth row).

small p-value compared to those where a larger variance signal was injected (right). This shows how the sensitivity of the tests varies when targeting clusters of varying width and strength relative to the background.

In this example, since the signal can be spotted in the projection of multiple dimensions, the product of p-values test (second row) offers the largest rejection probability of the null hypothesis compared to the volume-transformed p-value (third row) or the minimum p-value test (first row).

### 5.4.2. Multivariate Gaussian-shell signal

Instead of injecting a clustered signal, we can assess the sensitivity of the methods for a Gaussian-shell signal. This signal is five-dimensional and characterized by a radius $r = 0.25$, a radial standard deviation of either $\sigma_r = 0.02$ or $\sigma_r = 0.1$ and the center of its distribution, $\mu$, chosen at random within the hypercube $[0.25, 0.75]_5$. The results of these trials are shown in Fig. 5.8. In this case, we notice that the sensitivity to either signal thickness, $\sigma_r$, is very similar, which shows that all methods are mostly sensitive to the shell-like structure and its radial extension. Of the three tested methods, the product of p-values shows the highest sensitivity, followed by the minimum p-value and then the volume transformed p-value.

Note that the data in the previous examples were analyzed all in one pass for each trial, meaning that the extracted p-values do not need any corrections for a 'trials effect' or 'look-elsewhere effect'. Of course, if one analyzes many separate sets of data, the resulting p-values will need to be corrected, as is usually done in the univariate case.

## 5.5. Example - $n$D Limit setting

The performance of the proposed methods for limit-setting is explored in a series of simulated experiments for multivariate sample distributions. Consider the case where a background model is absent, and only a distribution of counts according to a signal model is available. In this case, the task is to set a limit on the signal strength of the signal model.

### 5.5.1. Background-free experiment

To begin with, consider the case in which no background contaminates the experiment, in order to estimate the baseline of the different methods. Fig. 5.9 shows the median of the $CL = 0.90$ upper limits on the event rate normalized to the median limit of the Poisson test. We notice that in this baseline scenario, the Poisson test is the best of the bunch, as expected, but it does not drastically outperform the others, much like the univariate case.

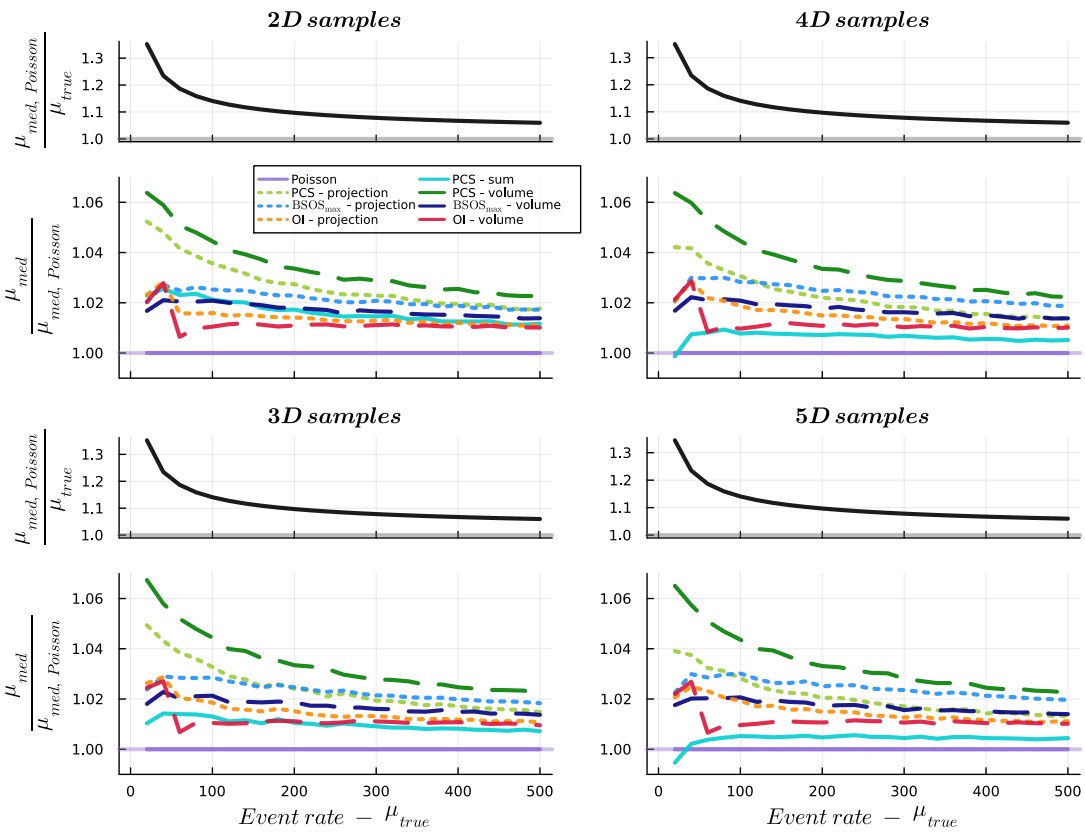



Figure 5.9.: Median $CL = 0.90$ upper limit for the Poisson test, upper panels, and for tests discussed in the text normalized to the limit from a standard Poisson probability test, lower panels, from 2D up to 5D uniform signal distributions and no background.

### 5.5.2. Background-only experiment

Next, we investigate the case in which a background is present in the simulations and the signal strength is negligible in comparison: this mimics a rare process search in which the signal is absent.

#### Exponential distribution

We first consider a background resulting from the product of $n$ independent Exponential distributions of rate 0.1 in each dimension. The exponential distributions peak at the origin, from which the volume is calculated, thus being a rather conservative location for the background distribution, referencing the discussion in Sec. 5.3.5.

Fig. 5.10 reports the median $CL = 0.90$ upper limits of the measured event rate normalized to the smallest median result for a specific background event rate $\mu_{bkg}$. Analysing these results, we notice that the volume transformation method provides the best limits, regardless of the test used. All other projection-based methods perform

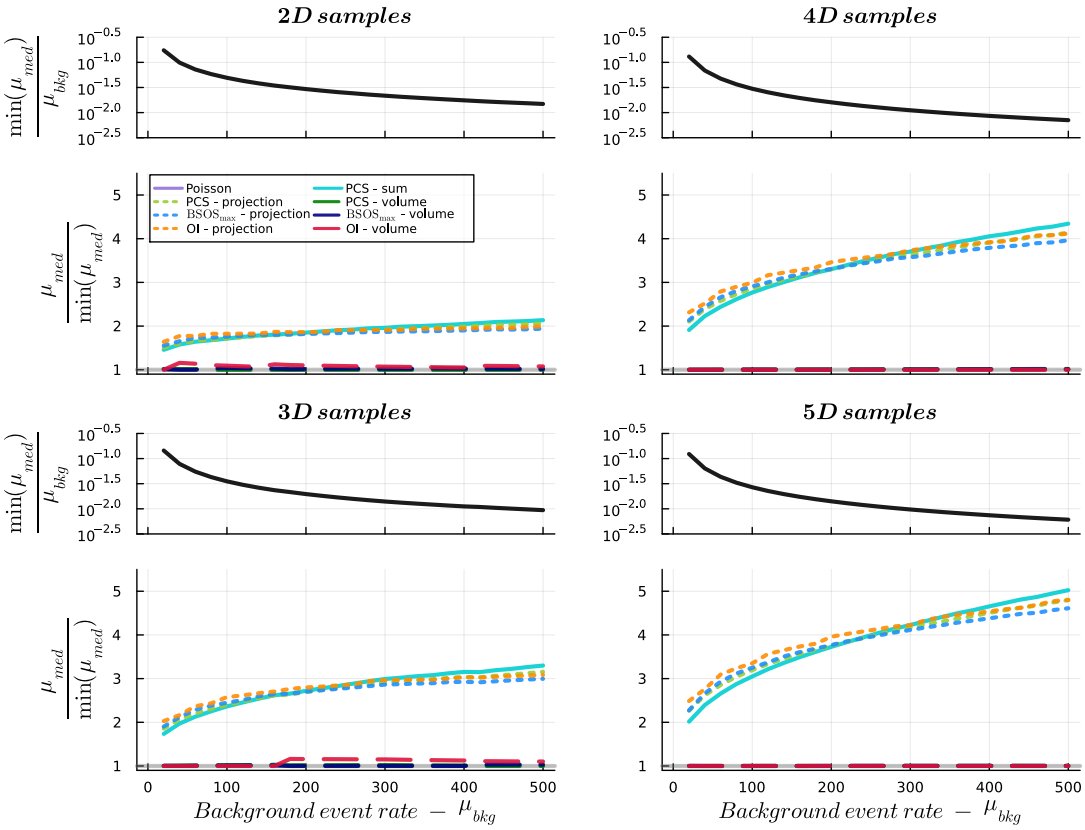



Figure 5.10.: Median $CL = 0.90$ signal upper limit for the best available test, normalized to the background strength, upper panels, and for tests discussed in the text normalized to the limit from the best test, lower panels, from 2D up to 5D distributions containing only an exponentially distributed background.

similarly: in the two-dimensional scenario, the limits are a factor $1.5 - 2$ worse than the volume transformation results, and in the case of a three-dimensional distribution, a factor $2 - 3$ worse. The limits for large event rates become up to 4 and 5 times worse in the respective number of dimensions. Overall though, the limits set by all methods are quite good, since they reject more than 90% of the background rate when it is small and able to reject up to 99% for large rates $\mu_{bkg} \geq 300$.

### Gaussian distribution

Next, we consider a background distributed according to a multivariate Gaussian centered in the middle of the hypercube and with covariance matrix $\Sigma = I \cdot 0.01$.

Fig. 5.11 reports the median $CL = 0.90$ upper limits of the measured event rate normalized to the smallest median result for a specific background event rate $\mu_{bkg}$. Once again, the volume transformation method provides the best limits, regardless of the test used. Out of these, the $\text{BSOS}_{\text{max}}$ test sets the lowest limits. This is to be expected since

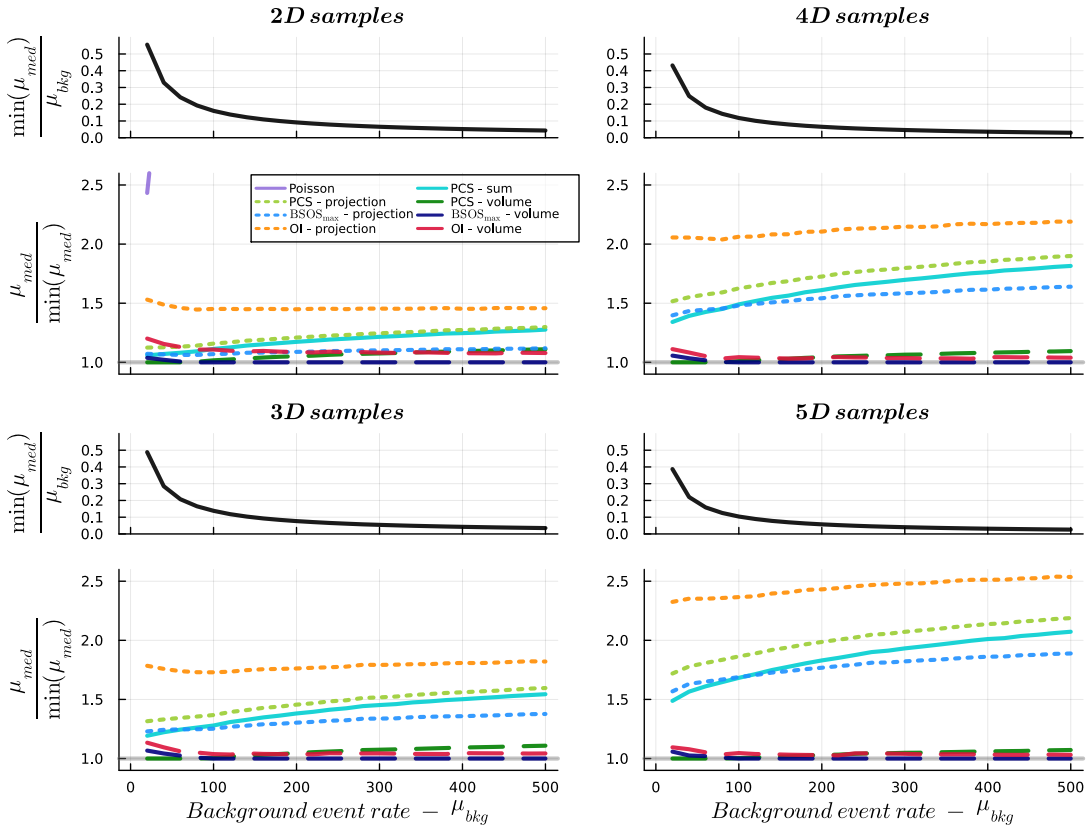



Figure 5.11.: $CL = 0.90$ upper limit normalized to minimum median result, from 2D up to 5D multivariate Normal distributions with $\Sigma = I \cdot 0.01$ centered in the middle of the hypercube. The upper panels in each case show the best limit result normalized to the background expectation.

it is better suited to analyse datasets that present multiple disconnected low-density regions.

The projection-based methods provide weaker limits: the $\text{BSOS}_{\text{max}}$ and PCS version being up to a factor $1.25(1.5, 1.75, 2)$ larger in the 2D (3D, 4D, 5D) case respectively; the Optimum Interval test's limits are weaker by a factor $1.5(1.75, 2, 2.5)$ in the 2D (3D, 4D, 5D) case respectively. This is understandable since this test relies only on one low-density region to estimate its limit.

Overall, all tests are able to reject more than 90% of the background rate for $\mu_{bkg} \geq 300$.

### Concave distribution

Finally, we can consider a bowl-shaped background, obtained by reversing the roles of signal and background distribution of the previous example: assuming a uniform background and a Gaussian signal in the real space (truncated to the unit interval $[0, 1]$ with $\mu = 0.5$ and $\sigma = 0.1$), if we performed the probability integral transformation with

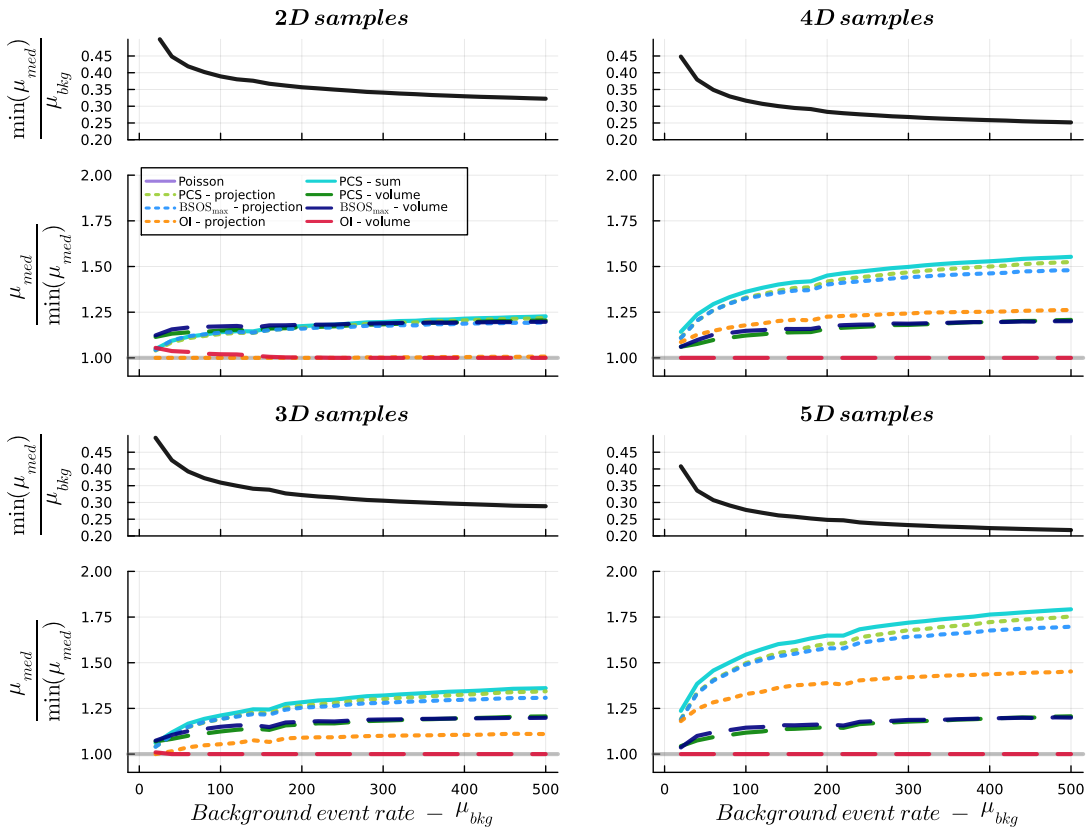



Figure 5.12.: $CL = 0.90$ upper limit normalized to minimum median result, from 2D up to 5D multivariate Normal distributions for the concave background model. The upper panels in each case show the best limit result normalized to the background expectation.

respect to the latter, we would obtain a bowl-shaped background distribution in the cumulative space. Fig. 5.12 reports the median $CL = 0.90$ upper limits of the measured event rate normalized to the smallest median result for a specific background event rate $\mu_{bkg}$. We notice that the best results, in this case, are set by the Optimum Interval test with volume transformation. This is reasonable since there is only one fully connected region of low event density, namely the basin of the bowl, thus being the best-suited case for the OI test. The next best results are obtained by the OI-projection method up to three dimensions and beyond by the $\text{BSOS}_{\text{max}}$ and PCS volume transformations, which yield no more than 25% larger limits. Finally, the remaining projection-based methods yield the most conservative limits. Overall we notice that as the number of dimensions grows, the tests are able to reject more of the background distribution, ranging from 60% up to 75% rejection for larger values of $\mu_{bkg}$.

These examples show how to apply the multivariate goodness of fit tests I introduce and give an idea of their performance in the case of realistic analysis scenarios.

# 6. Physics applications

In this chapter, I illustrate how the tests introduced above could be used in a physics scenario, both for signal discovery and limit setting, in the univariate case. I consider examples inspired by real or proposed experiments, either constructing the models based on realistic expectations of the background and signal involved or by replicating previous analysis using the newly proposed Spacings-based tests.

In order to show the versatility of the tests, I consider three examples: a hypothetical bump-hunting scenario, where the task is to identify a narrow peaked signal against a background; a limit setting example, in which I replicate the analysis of CRESST data in order to set an upper limit on the spin-independent WIMP-nucleus scattering cross-section; lastly, I develop an online trigger for neutrino-flare detection (hopefully due to Supernovae) to be used in the future by the RES-NOVA experiment.

The examples shown here have already been presented in [25, 50, 58], which I closely follow in the following for my discussion.

## 6.1. Example Particle Physics: Bump Hunting

I consider a detector that collects a number of events in an observable $x$, where $x$ could for example be the energy of an event, the detection time, or a reconstructed quantity like an invariant mass. I expect some or all of the observed events to follow a known background distribution $f_B(x)$, but there may be an additional contribution of events from an unknown signal distribution $f_S(x)$—such as a rare, exotic particle decay with unknown mass. Hence I want to quantify the goodness-of-fit of the background-only model to the data. A resulting low p-value could indicate the presence of events distributed according to an additional, unknown signal distribution.

In this example, I use an exponential distribution $f_B(x) = e^{-x}$ for the background model (null-hypothesis). In order to illustrate how the presence of an actual signal (alternative hypothesis) would affect the outcome, I also inject additional events following a normal distribution centred at $x = 1$ and standard deviation $\sigma = 0.05$. The number of events is Poisson fluctuated for both background and signal, with expected values of $\langle n_b \rangle = 100$ and $\langle n_s \rangle$ varied as specified. In Fig. 6.1, an example distribution of observed events is shown, together with the assumed background distribution, and the distribution with injected signal events (here $\langle n_s \rangle = 5$). Repeating this example multiple times and analysing each trial with all available test statistics, yields a p-value distribution for each of them, which can then be used to determine the sensitivity of each test as a function of the number of injected signal events. Apart from the non-parametric tests considered

in the general comparison of Sec. 3.8, I also discuss a Likelihood-Ratio approach and compare the results.

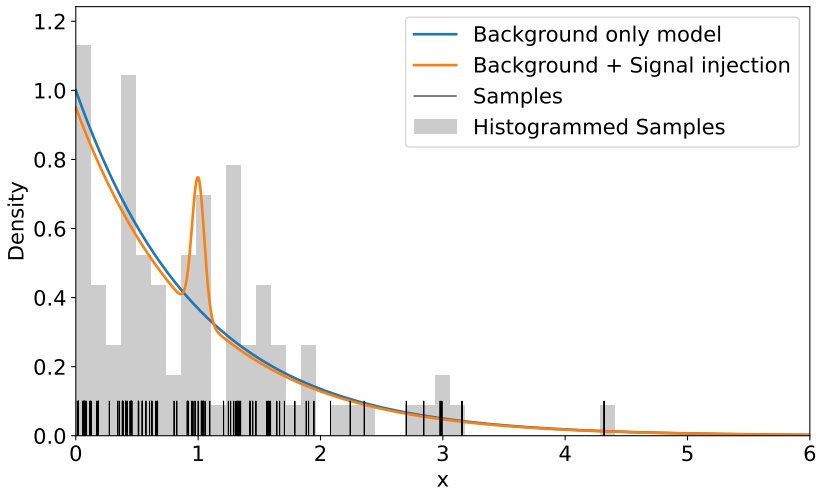



Figure 6.1.: Example of physics analysis problem, with observed events distributed in $x$. I test the goodness-of-fit of the background only model (blue) to the samples. Here the samples have been generated according to a different distribution with an injected signal (orange).

### 6.1.1. Non-parametric tests

No assumption is made on the rate of the underlying processes, meaning that the number of observed counts is not included in the analysis other than for the calculation of the test statistic. This means the "shape" of the distribution is tested for, not its normalization. The conversion of events via the CDF of the distribution function considered, $f_B$, (probability integral transformation) transforms the problem into a test of uniformity.

The p-value distributions under the assumption of $H_0$ (i.e. only background is present) for repeated trials with $\langle n_b \rangle = 100$, and various injected $\langle n_s \rangle = [0, 3, 6, 9, 12, 15]$ are shown in Fig. 6.2. All distributions with no signal ($\langle n_s \rangle = 0$) show a flat p-value distribution, as expected, since in that case all events are drawn from the background distribution $f_B$. For trials with injected signal, the distributions are trending towards smaller p-values, indicating the worsened goodness-of-fit for the background-only model. In the example, all tests exhibit this behaviour. The largest rejection of the null-hypothesis is offered by the RPS and $\mathrm{BSS}_{\mathrm{min}}$ tests, which present the most skewed p-value distributions for any number of injected signals $\langle n_s \rangle$.

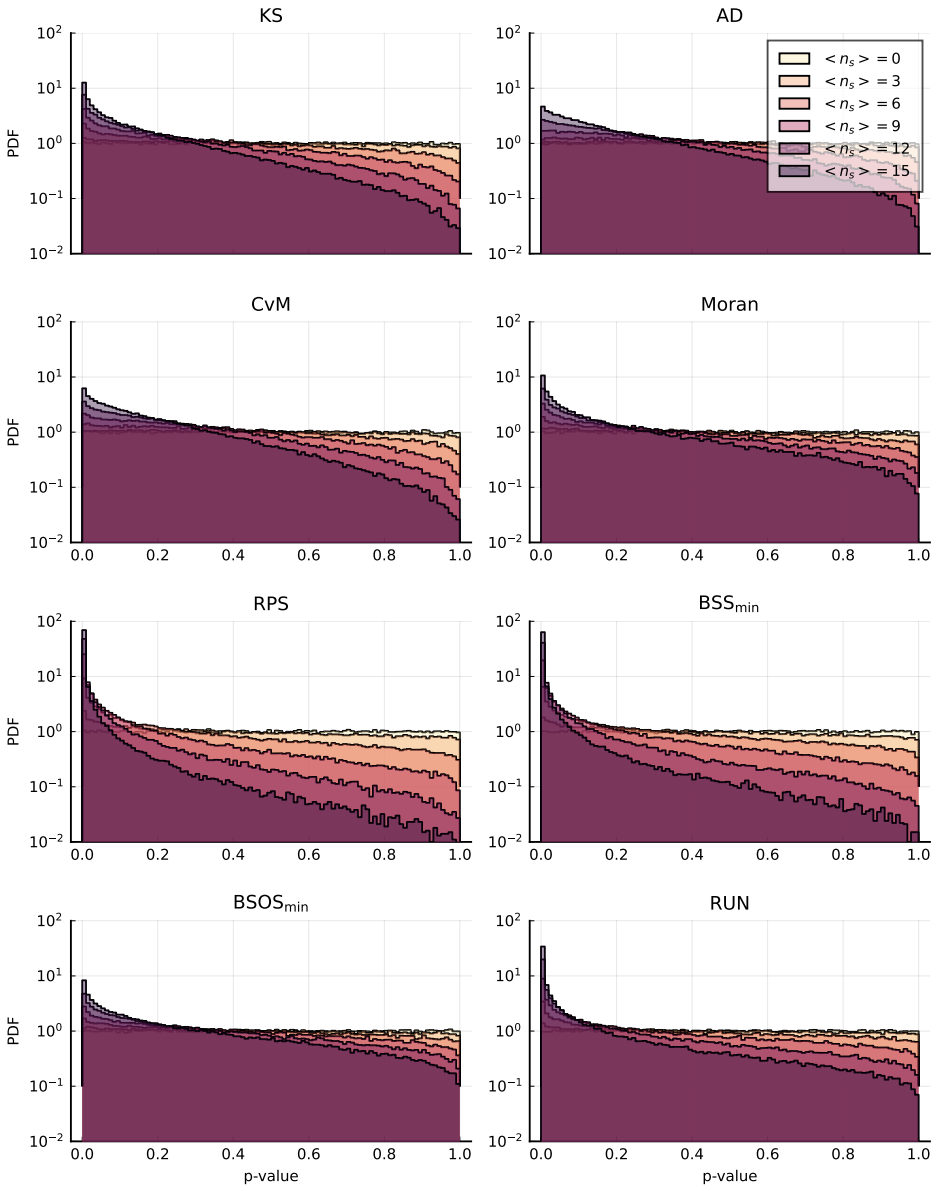



Figure 6.2.: p-value distributions for background only samples ($\langle n_s \rangle = 0$) and background plus randomised signal injections comparing to the background model for several choices of test statistics.

### 6.1.2. Likelihood Ratio test

An alternative approach to using non-parametric test statistics is the Likelihood-Ratio (LR), especially if partial information about the signal is known or assumed.

Assuming it is known that the signal has a Gaussian shape, the distribution of the events could be represented by a mixed distribution:

$$H(\lambda, c, \mu, \sigma) = c \cdot \text{Exp}(\lambda) + (1 - c) \cdot \mathcal{N}(\mu, \sigma) \tag{6.1}$$

where $c$ is the fraction of background, $\lambda$ the exponential rate of the background and $(\mu, \sigma)$ the position and standard deviation of the Gaussian signal.

The likelihood is simply defined as the product of the probability density function evaluated at each sample $x_i$:

$$\mathcal{L}(\lambda, c, \mu, \sigma) = \prod_{i=1}^{n} \left[ c \cdot f_{\text{Exp}}(x_i|\lambda) + (1 - c) \cdot f_{\mathcal{N}}(x_i|\mu, \sigma) \right]. \tag{6.2}$$

The null hypothesis corresponding to the background-only distribution is denoted by $H(\lambda) = H(\lambda|c = 1, \mu = 1, \sigma = 0.05)$, where the rate $\lambda$ is not fixed, but will be allowed to vary in order to maximize the corresponding likelihood given the data. For a given alternative hypothesis, such as the one shown in Eq. 6.1.2, the likelihood ratio test statistic is:

$$\rho = -2 \log \left[ \frac{\sup \mathcal{L}(\lambda)}{\sup \mathcal{L}(\lambda, c, \mu, \sigma)} \right]. \tag{6.3}$$

Given the current example, I consider three different alternative hypotheses: $H(\lambda, c, \mu, \sigma)$, where it is only assumed that the signal has a Gaussian shape, with very loose constraints on its location or width; $H(\lambda, c, \mu|\sigma = 0.05)$ where apart from the shape there is also an assumption on the width of the signal; $H(\lambda, c, \sigma|\mu = 1)$ where only the shape and location are assumed, but not the width of the signal.

Under the assumption that the null hypothesis is correct, the distribution of $\rho$ is not known a priori, but for large enough sample sizes, due to Wilk's theorem [59], we know it converges asymptotically to a $\chi^2_k$ distribution with $k$ degrees of freedom, where $k$ is the difference between the number of free parameters of the null and alternative hypotheses. In this case, for a given value of $\rho$, its p-value would be $1 - F_{\chi^2}(\rho|k)$.

Given that the rate of events drawn in each trial is $\sim 100$, the asymptotic regime might not approximate the distribution of $\rho$ correctly under the null hypothesis.

In order to test this assumption, I collected samples of $\rho$ for each alternative hypothesis and compared them to the corresponding asymptotic $\chi^2$ distributions. The results are shown in Fig. 6.3, where we notice that the empirical distribution of $\rho$ when generating data according to the null hypothesis is not correctly approximated in either case (empirical distribution in black and asymptotic distribution in red). The largest discrepancy can

be noticed for the alternative hypothesis with the most free parameters, $H(\lambda, c, \mu, \sigma)$: the tail of the $\chi^2_3$ distribution decays faster than that of the empirical one, meaning that if we were to use this one to calculate a p-value for a given $\rho$, the estimate would be too small, not close to the real p-value. In the case of $H(\lambda, c, \mu|\sigma = 0.05)$ or $H(\lambda, c, \mu|\mu = 1)$, the asymptotic distribution appears to be closer to the empirical distribution, nevertheless, it can be noticed that the empirical distributions are more peaked at the origin, exhibiting a slightly faster decay compared to the asymptotic distribution: in this case, the p-value calculated using the asymptotic distributions would be slightly larger than its actual value.

In order to estimate p-values correctly for the various likelihood ratio approaches, I use the empirical distributions instead of the asymptotic ones. For given background and signal rates, $\langle n_b \rangle$ and $\langle n_s \rangle$, the reference distribution is produced from $2 \cdot 10^7$ trials where samples are generated with a Poissosn rate of $< n_b + n_s >$ according to the background distribution.

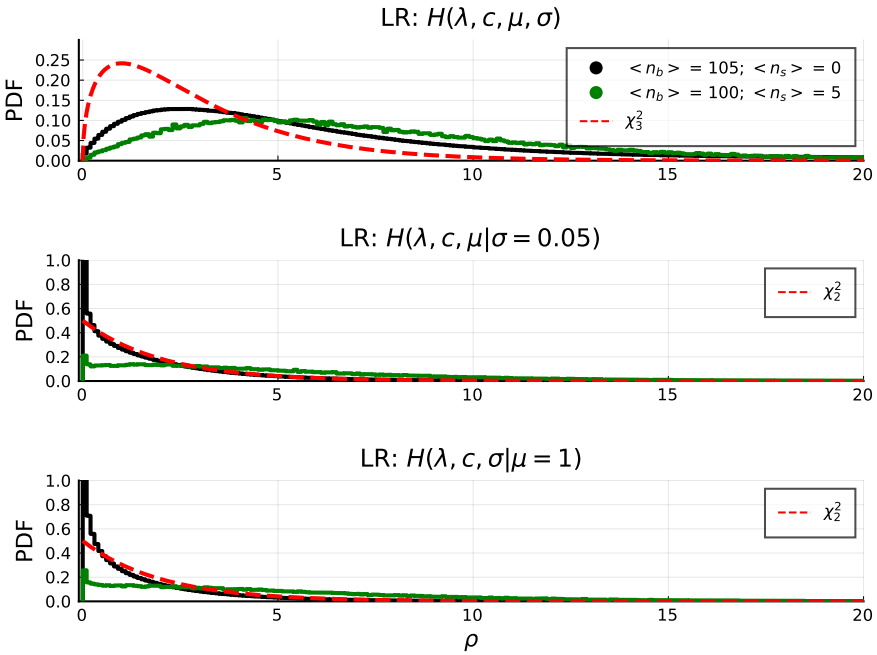



Figure 6.3.: Distribution of the likelihood ratio statistic, $\rho$, for the different choices of the alternative hypothesis, compared to the corresponding asymptotic distributions (red).

### 6.1.3. Results

I quantify the sensitivity of the analysis to reject the background-only model at different significance levels under the assumption of the presence of a signal. Therefore, I check the median p-value of repeated trials, and at what value of $\langle n_s \rangle$ it crosses specific critical

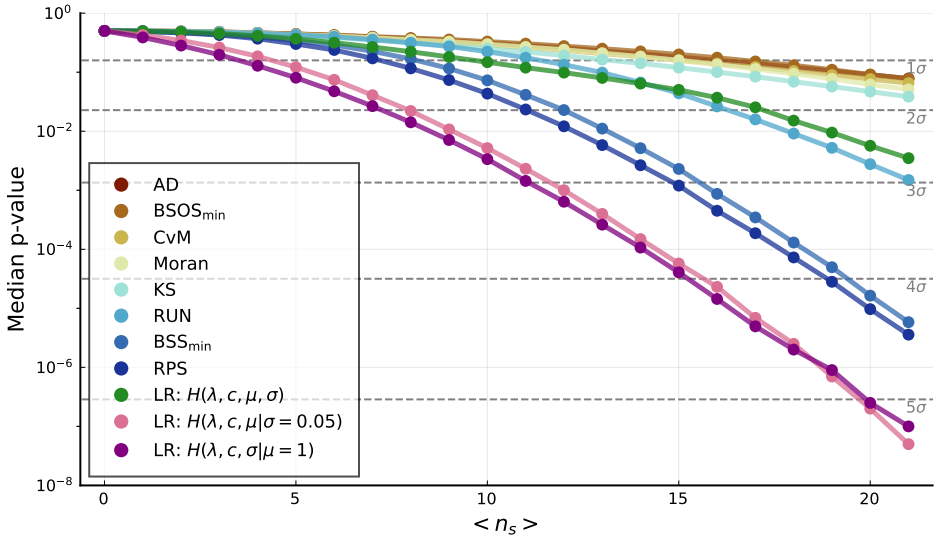



Figure 6.4.: The expected significance level at which the background model can be excluded under the assumption of a signal, as a function of $\langle n_s \rangle$ for the different tests.

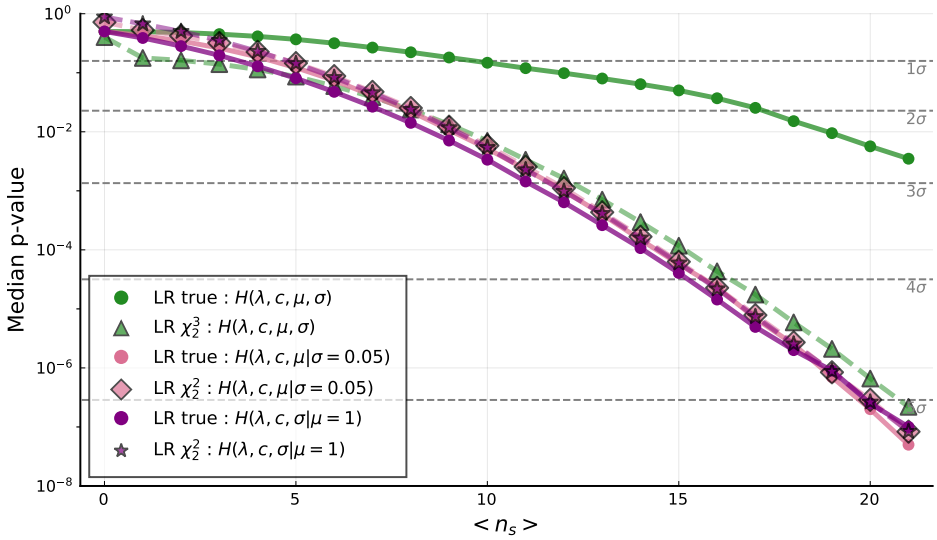



Figure 6.5.: Sensitivity of likelihood ratio alternatives, as a function of $\langle n_s \rangle$, calculated using the empirical and asymptotic distributions.

values, as shown in Fig. 6.4. In the chosen example, for a signal of strength $\langle n_s \rangle \sim 10$, I expect to reject the background-only model at the $2\sigma$ significance level[1] using RPS or $\mathrm{BSS}_{\mathrm{min}}$, whereas for the other non-parametric tests (apart from the RUN statistic), a signal rate of at least $\langle n_s \rangle \sim 20$, is needed to achieve the same exclusion level. Such a large signal of $\langle n_s \rangle \sim 20$ would allow rejecting the background-only model at $> 4\sigma$ significance with the RPS or $\mathrm{BSS}_{\mathrm{min}}$ tests while the RUN statistic achieves a significance of $3\sigma$. For lower signal rates, the RUN test reaches the $2\sigma$ significance level at approximately $\langle n_s \rangle = 16$, a rate at which the RPS or $\mathrm{BSS}_{\mathrm{min}}$ can already reject the null-hypothesis with a significance of $3\sigma$. Looking at the sensitivity of the LR statistics, we notice they are higher compared to the ECDF ones, but they show vastly different behaviour depending on the number of free parameters: considering $H(\lambda, c, \mu, \sigma)$, it performs similarly to the RUN statistic, but it does not reach a $3\sigma$ significance even for $\langle n_s \rangle \sim 20$; $H(\lambda, c, \mu|\sigma)$ and $H(\lambda, c, \sigma|\mu)$, on the other hand, reach a sensitivity of $3\sigma$ for $\langle n_s \rangle \sim 10$ and at $\langle n_s \rangle \sim 20$ can reject the null hypothesis with $> 5\sigma$ significance. The overall performance of the LR statistics is not surprising, since they are supplied with additional knowledge regarding the signal. Still, it is noteworthy the effect that fixing either $\mu$ or $\sigma$ has on the sensitivity, compared to allowing both to be free. As stated above, the sensitivity of the LR statistics was calculated using the empirical distributions. A comparison of the sensitivity assuming the asymptotic distributions to that from the empirical distributions is shown in Fig. 6.5. Since the asymptotic distributions of $H(\lambda, c, \mu|\sigma)$ and $H(\lambda, c, \sigma|\mu)$ were similar to the empirical ones, the sensitivity is reasonably accurate, but as expected, it is slightly worsened. The sensitivity of $H(\lambda, c, \mu, \sigma)$ calculated using the $\chi^2_3$ distribution shows a much larger discrepancy since it ends up being similar to the previous two cases, producing a wrong result. This example shows that unless one has reliable information on the detailed shape or location of an eventual signal, then spacings-based statistics such as RPS or $\mathrm{BSS}_{\mathrm{min}}$ are able to reach noticeably better sensitivities compared to a likelihood ratio approach.

## 6.2. CRESST Analysis Example

As a further example, I replicate the analysis of the CRESST Collaboration [48] to determine upper limits on the WIMP cross-section.

I begin with a brief introduction on the importance of dark matter searches and its experimental signature. For a comprehensive description of the theory behind direct search approaches and the CRESST experiments, I refer to Dr. Iachellini's thesis [60], from which I borrowed for the following introduction and theoretical background.

### 6.2.1. Introduction

Dark matter is a fundamental brick of modern day cosmology. Its significance in contemporary research is due to the great success of the ΛCDM model, which convinced

[1] A significance level in terms of numbers of $k$ standard deviations $\sigma$ can be translated to a p-value as one minus the integral over a unit normal distribution form $-k$ to $+k$.

the scientific community of the existence of a large fraction of non-luminous matter in the Universe. Dark matter has a profound impact on our understanding of the Universe: it plays a pivotal role in galaxy formation and evolution, its presence can help explain the observed motion of galaxies within galaxy clusters, and it even acts as the underlying framework for the large-scale structure of the Universe.

Popular dark matter candidates are Weakly Interacting, very Massive and stable Particles: WIMPs. Currently, the most favoured hypothesis regarding their origin assumes they are a thermal relic of an earlier epoch of the Universe. Their creation and annihilation, at the time of a hot and dense Universe, was analogous to all other particles, but the rapid expansion of the Universe lowered their density to the point of suppressing their annihilation, freezing the population of WIMPs.

Ongoing experiments and research projects are dedicated to uncovering the nature of dark matter, targeting various candidates apart from WIMPS, such as axions, and sterile neutrinos. Detection efforts include direct detection experiments, indirect detection through cosmic rays and gamma rays, and collider searches. The CRESST experiment is one of the most promising experimental efforts for the direct detection of dark matter particles present in the Milky Way using extremely sensitive cryogenic detectors, with great energy resolution and an event-by-event particle identification.

### 6.2.2. Experimental signature

The experimental signature of direct detection experiments is the differential event rate:

$$\frac{d\Gamma}{dE_R} = \frac{\rho_\chi}{m_N m_\chi} \int_{v_{min}}^{\infty} d^3\bar{v} f(\bar{v}) v \frac{d\sigma(v, E_R)}{dE_R} \tag{6.4}$$

where $m_N$ is the nuclear mass, $m_\chi$ the WIMP particle mass, $E_R$ the recoil energy, $\sigma$ the cross section for the scattering process, $\rho_\chi$ the local density of dark matter, $f(\bar{v})$ the dark matter particle velocity distribution and $v_{min}$ the lowest velocity that can transfer $E_R$ energy to the recoiling nucleus:

$$v_{min} = \sqrt{\frac{E_R m_N}{2\mu^2}} \tag{6.5}$$

where $\mu$ is the reduced mass of the combined system of dark matter particle and nucleus.

The differential cross-section $\sigma(v, E_R)$ contains the physics of the interaction between the nucleus and the dark matter particle. In general, it consists of both a scalar and a vector coupling, with the latter describing the interaction of the dark matter particle with the net spin of the target nucleus. Sensitivity to the spin-dependent interaction requires that target nuclei have a non-vanishing net spin, but the relevant target in this case ($CaWO_4$) does not have a significant nuclear spin. Therefore, we will neglect spin-dependent interactions.

The differential cross-section for spin-independent dark matter particle-nucleus scattering is:

$$\left(\frac{d\sigma}{dE_R}\right)_{SI} = \frac{m_N \sigma_0}{2\mu^2 v^2} F^2(E_R) \tag{6.6}$$

where $F(E_R)$ is the form factor and $\sigma_0$ is the point-like, zero-momentum cross-section for the scattering process:

$$\sigma_0 = \frac{4}{\pi} A^2 f^2 \mu^2 \tag{6.7}$$

where $f$ is the strength of the coupling (which is generally considered to be equal for protons and neutrons). $A^2$ represents the coherence-induced enhancement of the interaction and for this reason, heavy targets are preferred for direct detection of dark matter, because of the quadratic dependence on the atomic mass number $A$. Details regarding the parametrization and choice of form factor and velocity distribution are reported in [60].

Given the differential rate, Eq. 6.4, for a specific value of $m_\chi$, it is possible to predict the shape of a WIMP-induced nuclear recoil signal in a direct detection experiment, and since $CaWO_4$ is a composite material, the total event rate has to be computed for each nuclear species and the total rate is the sum of the single components.

In a real experiment, the theoretical event rate I just described has to be corrected in order to account for finite energy resolution, energy threshold, and cut-survival probability.

### 6.2.3. Cross section limit

The goal is to replicate the analysis of the most recent spin-independent public dataset [61] released from the CRESST collaboration and to compare the performance of the different test statistics I have studied. These data are accompanied by information regarding the energy resolution and efficiencies of their setup for $CaWO_4$ targets that allow calculating the corrected differential rate $\frac{dN}{dE}$ for a specific WIMP mass across an energy range of interest $[E_{min}, E_{max}]$.

The data I analyse are shown in Fig. 6.6 (top). Denoting the integral of the differential rate over the whole energy range as $\Lambda = \int_{E_{min}}^{E_{max}} \frac{dN}{dE} dE$, for a given set of ordered events $\{E_i\}$, the probability integral transformation is simply:

$$x_i = \frac{1}{\Lambda} \cdot \int_{E_{min}}^{E_i} \frac{dN}{dE} dE \tag{6.8}$$

yielding a set of ordered events $\{u_i\}$ in the unit interval $[0, 1]$.

Fig. 6.6 (bottom) shows the distribution of events in the cumulative space after transforming using the signal distributions calculated at two different WIMP masses. Apart from small differences, the two datasets are very similar, showing an extremely

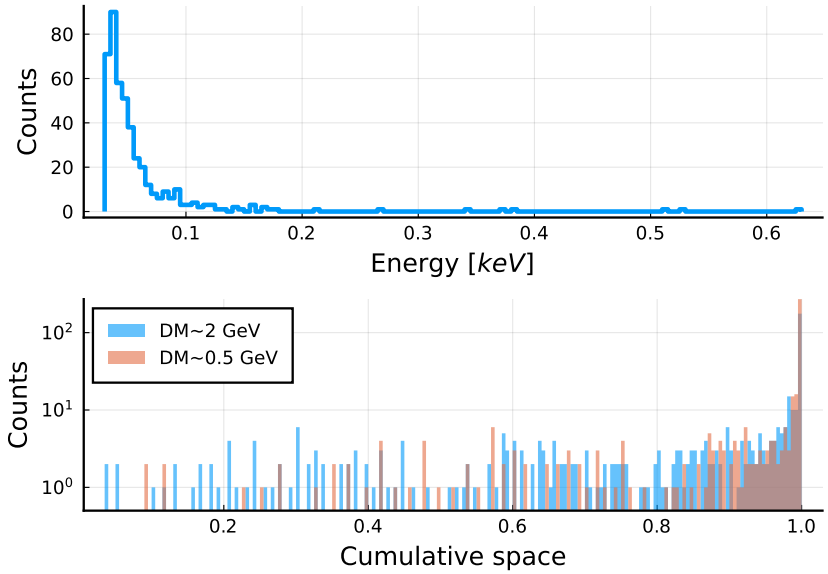



Figure 6.6.: (top) Histogram of CRESST data [61] consisting of energy deposition from an interaction of a particle in the $CaWO_4$ crystal; (bottom) Histogram of data transformed using the signal distribution for two proposed WIMP masses, 0.5 and 2 GeV/$c^2$

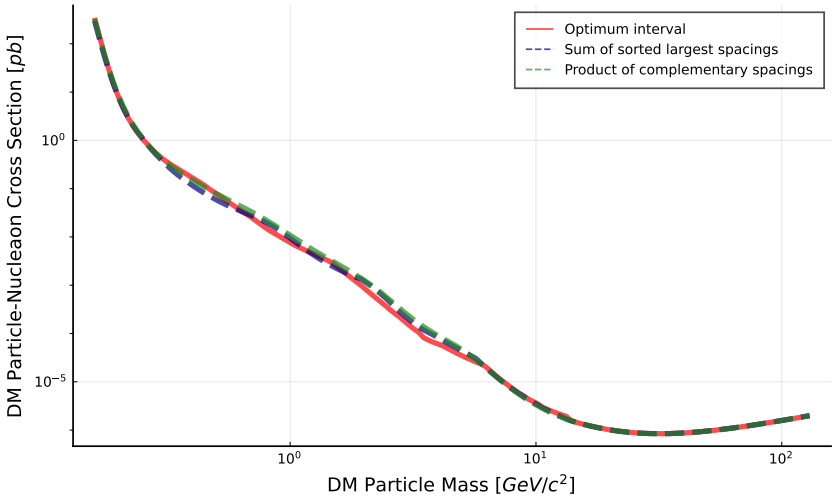



Figure 6.7.: $CL = 0.90$ upper limit on the WIMP-nucleon cross-section as a function of the WIMP mass calculated with different tests.

peaked distribution close to 1 and an almost linear distribution of events in the rest of the unit interval.

Fig. 6.7 shows the $CL = 0.90$ upper limits on the cross-section calculated using our methods as well as those computed with the Optimum Interval method, officially used by the CRESST collaboration, which match the officially published limits.

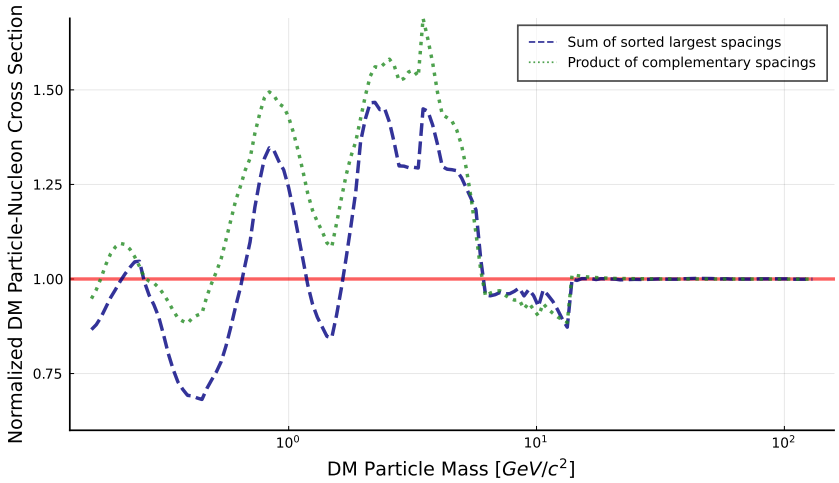



Figure 6.8.: $CL = 0.90$ upper limit on the WIMP-nucleon cross-section normalized to the Optimum Interval's result as a function of the WIMP mass calculated with our proposed methods.

Comparing the results of our calculations I do not notice any large deviations from one another. To better grasp the differences across results, I normalize the limits obtained with our methods to the official ones (obtained with the Optimum Interval method), as shown in Fig. 6.8. Here, I notice that, for the given data, the Product-of-Complementary-Spacings method yields 25% to 50% higher limits on average, whereas the Sum-of-Ordered-Spacings presents an oscillating behaviour, being able to provide up to 30% lower limits for low WIMP masses and up to 40% higher limits for masses of the order of $\sim 5\text{GeV/c}^2$. Since the CRESST experiment focuses on the low mass regime (and is less competitive at higher masses), this is a particularly interesting result and it aligns with the expectations shown in Sec. 4.4.2 - 4.4.3 for peaking backgrounds.

Finally, for WIMP masses $\geq 20\text{GeV/c}^2$, all methods saturate and yield the same result, reconstructing a signal event rate of $\sim 2.3$ events, corresponding to the $CL = 0.90$ limit of the Poisson test for an empty analysis window.

This example based on a published data set, as well as the results of our performance comparisons, shows that in general there is no "best" test statistic when it comes to setting upper limits in the presence of unknown backgrounds, but the results are highly dependent on the actual distribution of events.

## 6.3. Online trigger for supernova detection

Here I present an application of the RPS statistic to the design of an online trigger for Supernova detection for the RES-NOVA experiment. This work was done in collaboration with Dr. L. Pattavina, Dr. N. Ferreiro Iachellini and Dr. P. Eller and the material presented here closely follows extracts from [62, 58].

### 6.3.1. Introduction

Supernovae (SNe) are among the most energetic events in the Universe. They mark the end of a star's life with an intense burst of neutrinos [63, 64]. Why and how massive stars explode is one of the important long-standing unsolved mysteries in astrophysics. Neutrinos are known to play a crucial role in such events [65], nevertheless our understanding is still limited due to the lack of experimental observations. The knowledge we have relies mostly on hydro-dynamical simulations of the stellar matter, where also neutrino are propagated, but a direct validation of these simulations is still missing [66]. A timely, high resolution and high statistics detection of these neutrinos can be decisive for the understanding of the gravitational collapse and the connected neutrino emission [67]. In fact, neutrinos and gravitational waves (GWs), carry imprints of the explosion mechanism in real time, enabling a direct access to the inner stellar core [68]. A simultaneous detection of neutrinos and GWs is considered the *Holy Grail* of modern multi-messenger astronomy.

Multiple neutrino detectors are currently operating, and scrutinizing different region of the cosmos waiting for the next SN event. These experiments can be classified into three main categories: water-based Cherenkov (WBC) detectors [69, 70], liquid scintillator (LS) detectors [71, 72, 73] and liquid Ar (LAr) time projection chambers [74]. They all have two common features: they run detectors with active volumes ranging from few $m^3$ to several thousands $m^3$, and they are mostly sensitive only to $\overline{\nu}_e/\nu_e$.

Coherent Elastic neutrino-Nucleus Scattering (CE$\nu$NS), discovered few years ago [75], is an ideal channel for neutrino detection. In fact, it opens a window of opportunities for the study of neutrino properties [76, 77, 78], thanks to its high interaction cross-section and its equal sensitivity to all neutrino flavors. Currently, the SN neutrino community is lacking an experimental technique highly sensitive to the full SN neutrino signal. Recently, dark matter (DM) detectors, searching for nuclear recoils induced by galactic DM particles, were proposed to detect SN neutrinos via CE$\nu$NS [79, 80, 81], given the similarities in the expected signal (i.e. low energy nuclear recoils).

It is difficult to forecast when and where the next SN will occur. Though, some predictions can be made through the study of the stellar formation rate and the distribution of SN remnants in a galaxy [82]. In [83] it is shown that in the region around 1 kpc from the Sun the expected SN rate is 5-6 times greater than the galactic mean value. Furthermore, looking at the spatial distribution of all the past galactic SNe, they all occurred in a range between 1 kpc and 4 kpc [84] of the sun. Such proximity demands suitable detectors, able to tolerate high neutrino interaction rates. This requirement can be challenging for large-volume monolithic detectors, as the ones which are currently operated or planned in the near future. Compact and highly modular detectors are ideally suited to fulfil this requirement.

RES-NOVA is a recently proposed neutrino observatory that exploits CE$\nu$NS as detection channel and uses an array of archaeological Pb-based cryogenic detectors [85]. Pb is an ideal target for the detection of neutrinos from astrophysical sources via CE$\nu$NS. In fact, it is the only element of the periodic table that ensures simultaneously

the highest cross section, as this scales with the square of the neutron number of the target nucleus, and the highest nuclear stability, for achieving low-background levels. Furthermore, archaeological Pb promises unprecedented isotopic purity [86, 87], leading to low background levels in the region of interest (ROI) [88, 89].

### 6.3.2. Detector background model

In order to deliver a robust estimate of the experimental sensitivity to SN neutrinos, the development of a detailed background model is mandatory. For this reason, starting from the current knowledge on the concentration of radioactive impurities in cryogenic low-background experiments, a Monte Carlo tool was developed in order to simulate the energy spectra produced by the distributions of radioactive contamination in different detector components. It is possible to estimate the expected background level in the ROI, which lies between the detector energy threshold and 30 keV [85], using as input to the Monte Carlo the detector geometry and the concentration of background sources. For a detailed description of the detector background model and its response to SN neutrinos, the reader is referred to [62].

Nevertheless, it has to be pointed out that, in a real setup, especially when dealing with low-background experiments, the radioactive backgrounds are difficult to assess exactly and do show a time dependence that we do not account for. Uncertainties on these spoil the application of simple Poisson statistics for the determination of expected rates and the confidence on them. Furthermore, not all backgrounds can be attributed to radiogenic origin. An example of this comes from the CRESST experiment, where it is observed that there are time periods (minutes of duration) where the trigger rates are substantially higher than the standard operating conditions [90, 91], due to instabilities of the cryogenic system.

For what concerns the radioactive background, the best possible estimation for the background rate $r_{bkg}$ is its direct measurement and monitoring once the experiment is set in operation: the background rate can be extracted from a selection of collected data using Monte Carlo methods, and once an estimate is available, it can be used to define the parameters of the analysis, as discussed in Sec. 6.3.4. The situation is more complicated when dealing with the other sources and because of that, in the following, I present other test statistics beyond Poisson counting, when facing the problem of identifying a neutrino signal.

### 6.3.3. Early identification of neutrino signals

The next galactic SN will bring information on physics processes that cannot be studied in any terrestrial experiment, and the elusive rate of such an event makes this information extremely valuable. For the first time in history, technologies to detect neutrinos, gravitational waves, and electromagnetic radiation from SN events are in place. It is of uttermost importance to record all possible data in the best quality, and to do so, the SN event needs to be detected as early as possible.

The Supernova Early Warning System (SNEWS) [92] is an international group of neutrino sensitive experiments aiming at providing the astronomical community with an early alerts for SN events. The ability to combine the signals from experiments sited at different locations on the globe brings several advantages. Firstly, by integrating data from multiple observatories, it allows to increase the detection sensitivity, especially for weak signals coming from distant SNe, and improves the accuracy of source localization by analyzing the time delays and spatial distribution of neutrino detections across the participating observatories. Secondly, by operating a coincidence trigger between experiments, it effectively reduces the false alarms, minimizing the likelihood of erroneous alerts due to background noise or local interferences.

A multi-messenger observing strategy is key to fully exploit the wealth of information carried away by neutrinos. The neutrino emission starts before the core's collapse even begins, meaning that neutrinos can provide an early warning signal. Knowing when and possibly where to anticipate the signal dramatically improves detection prospects. During the stellar collapse of the core, the neutrino emission is accompanied by the emission of GWs. As discussed in [93], the arrival time of the neutrinos can also act as a trigger for SN, increasing the sensitivity of GW experiments. In addition, an early detection of neutrinos, and possibly pre-SN neutrinos, can anticipate the electromagnetic burst by several minutes or days [93].

### 6.3.4. Statistical tools and data processing

This section discusses approaches for building a triggering system to detect transient events—such as SNe—based on the real-time data stream of a neutrino detector. Detected signal events are interspersed with background events that, for now, are considered to be distributed according to a fixed-rate process, of which the true rate is known. The top panel of Fig. 6.9 shows an example of the expected count rate in a RES-NOVA like detector for neutrinos from a SN at a distance of 10 kpc, together with the uniform expectation of background events. The middle panel shows what an example data stream could look like, generated from random variates of the expected count rate. The goal is to find a statistical method that can deliver a yes/no answer in near real-time to decide whether there was a SN signal present in the data. The false alarm rate (FAR) (i.e., type I error rate)—as an external constraint to our system—cannot exceed 1 per week, as required for participation in SNEWS.

A standard way to analyze such time series data is to use fixed or variable length windows in time, and reduce the task at hand to decide whether the events inside a window exceed the expected Poisson count from the background rate. If the background rate is exactly known, and the windows are non-overlapping, the threshold for a given FAR can be directly calculated from the Poisson distribution. Since such a configuration depends on the location of the time window edges, often overlapping windows are used. For 50% overlapping windows, approximations for the calculation of the p-value can be used. Such overlapping time window Poisson tests represent the current state-of-the-art, as used, for example, in [95]. Alternative approaches have been proposed, for example,

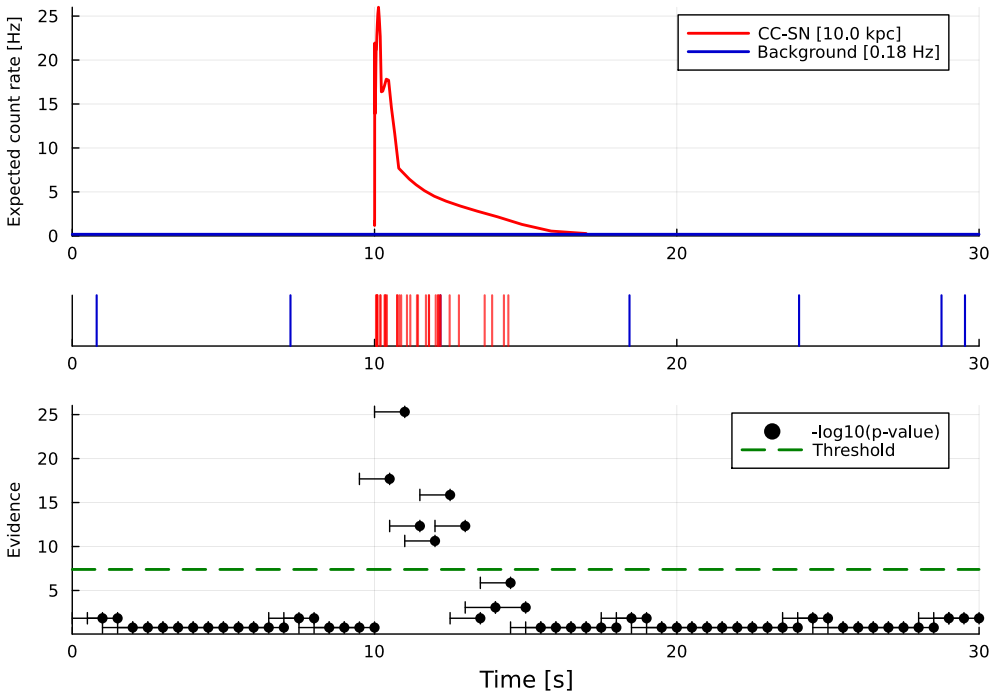



Figure 6.9.: Example of a CC-SN signal and Poisson analysis: (top) Expected count rate for a SN at $t = 10\,\mathrm{s}$ at distance of 10 kpc, plotted together with a constant background rate of 0.18 Hz. The considered CC-SN model is taken from hydrodynamical simulations performed by the Garching group [94] (ref. name `s27_ls220`). (middle) Random realization of counts as seen in the detector. (bottom) Analysis using 50% overlapping windows of 1 s length. For each window, the Poisson evidence is calculated and indicated with the black dot for the window extending 1 s into the past. The green line gives a threshold that would yield a FAR of 1 per 15 days.

with dynamic time windows [96]. The bottom panel of Fig. 6.9 shows the Poisson evidence for 1 s, 50% overlapping windows for our example, and a threshold value that results in a FAR of 1 per 15 days.

This type of system has a few parameters, including the background rate $r_{bkg}$ that can be estimated from background-only data, and the window configuration. There is a balance between choosing the window size $\omega$ large enough so that most of the signal is contained within the window and at the same time small enough so that signal events will not be washed out by an additional background contribution. The refresh rate, i.e., how often is a window analyzed, is another parameter of choice. A refresh rate of $1/\omega$ yields the configuration of non-overlapping windows, and $2/\omega$ would correspond to the 50% overlap. In order to retain the freedom of exploring more complicated window choices, and later also different statistical tests, I first introduce a simulation-based method to calculate critical values for a desired FAR.

### Computation of critical values

Given a custom test that operates at a certain refresh rate, we can calculate a test statistic value $TS$, which for example could be the Poisson p-value itself. However, for overlapping windows this quantity $TS$ can no longer be interpreted as the p-value of the test, and its distribution is in general unknown. To make a statistical statement about $TS$ we need to know, or rather estimate, its cumulative distribution $F_{TS}$.

The estimation $F_{TS}$ can be obtained using simulated data, producing values of $TS$ for the detection of neutrinos with a predetermined rate of the background-only scenario. The simulation cannot be done with independent runs, since this would remove the important correlations of successive values of $TS$ in consecutive time windows. Therefore, I simulate an extended run of the experiment and collect the values of $TS$ in a serial fashion, at least for time scales of the order of a day and below.

Since we are interested in using $F_{TS}$ to construct very low FAR thresholds for our analysis, we need a good approximation of its distribution for extreme values. This means in practice that we need to simulate and analyse a very long run of our experiment to produce enough statistics. In our simulations, I simulate between 25 and 100 years of background data.

Since the dataset modelling $F_{TS}$ was obtained through simulations, it means that it is completely dependent on the setup of the simulated experiment, namely the background rate $r_{bkg}$, the refresh time $t_r$, the window size $\omega$ and the definition of the test statistics $TS$. If any of the simulation parameters are changed, the dataset needs to be recalculated. Out of the parameters listed above only one of them will not be specified by us during the real operation of the experiment, and that is the background rate $r_{bkg}$. The estimation of this parameter is discussed in detail in Sec. 6.3.2, but for now, we can assume that it is known using a nominal value of $r_{bkg} = 0.18$ cts/s. Finally, given a model of $F_{TS}$ and a false alarm interval $\tau_{false}$, expressed in seconds just like the refresh time, I can derive the corresponding threshold value $TS^*$ for the trigger:

$$TS^* = F_{TS}^{-1}\left(\frac{t_r}{\tau_{false}}\right) \tag{6.9}$$

This threshold estimation holds for any possible test $TS$ that can be used to implement a trigger and is how the thresholds are calculated for all the test statistics I considered.

### Trigger evaluation and comparison

Given a specific setup of the experimental parameters $t_r$, $\omega$, and $TS$, the efficiency of the trigger can be assessed by evaluating the success rate when it comes to trigger activation in the presence of signal events. In order to test this, I simulate experiments with injected events that follow a possible model distribution of neutrinos after a SN explosion, as shown in Fig. 6.9. Here I present results pertaining to a small selection of neutrino flare models, although the number of proposed models is abundant in literature and unknown in reality [66].

The number of neutrino events $\lambda_{sig}$ is determined by the distance of the SN and after repeating these experiments a large number of times, I can estimate the fraction of successful trigger activations (i.e. 1−type II error). Fig. 6.10 shows the maximum distance to achieve a 95% rate of success for different choices of the window size parameter $\omega$, as a function of the time after the explosion. The refresh time $t_r$ is kept constant at 0.5 s, an indicative value that matches the expected overall throughput of the raw data processing rate.

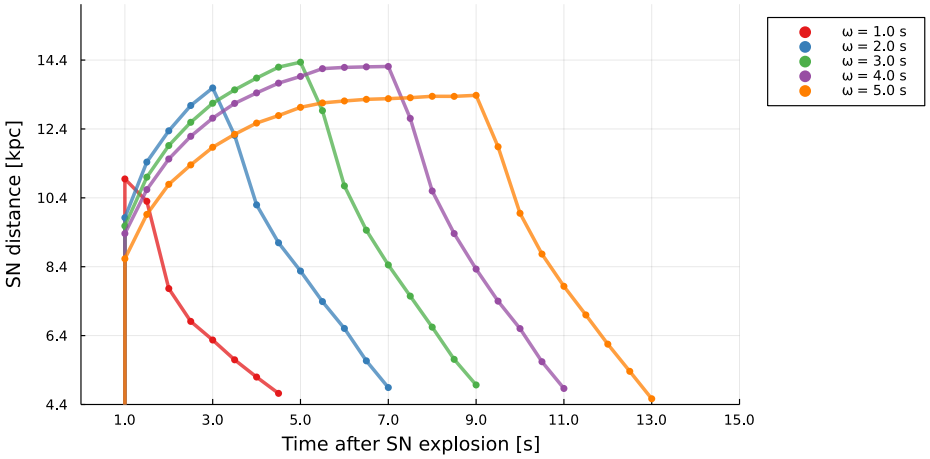



Figure 6.10.: The 95% quantile of successful SN detection distance, based on a FAR of 15 days, for different choices of the window size $w$. The refresh time is kept constant at 0.5 s.

Examining Fig. 6.10 we notice that around one second after the SN explosion the trigger starts to activate with a sharp turn on due to SN neutrinos. For very short windows, not all signal can be contained inside the window and the curve dies down rapidly again. For larger windows, further distances can be probed, since more of the signal can contribute to the statistic inside a window. For windows that are too large, however, more background events are being picked up that deteriorate the performance again. So there is an optimal window size, for the example in Fig. 6.10 this lies at around 5 s.

### Non-parametric Tests as Alternatives to Poisson

In the case of an optimal choice of window size and known background rate, the Poisson test will, in general, perform well. We have not found any alternative test outperforming an optimized Poisson test. However, the Poisson test relies on the fact that the background rate is known or can be reliably estimated from the data. This may not always be the case, or the background rate can even fluctuate. Furthermore, considering different signals of various time scales and shapes, or even a priori unknown transient signals, it is worthwhile to explore alternatives to the Poisson test.

I investigate non-parametric tests as a viable alternative to, or used in combination

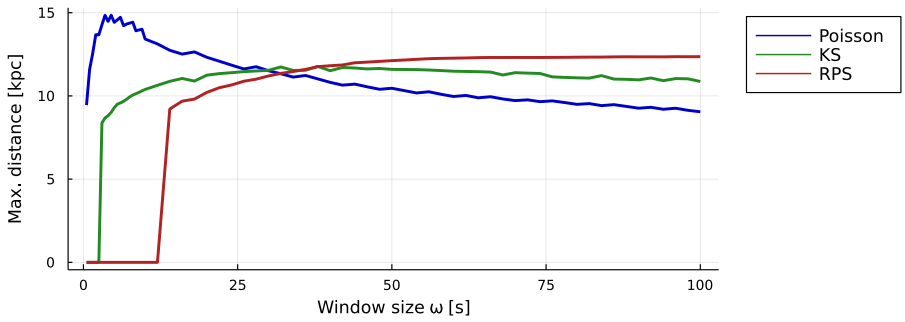



Figure 6.11.: The 95% quantile of successful detection distance for $r_{bkg} = 0.18$ cts/s, based on a FAR of 15 days, for different choices of the window size $\omega$. The refresh time is kept constant at 0.5 s.

with, Poisson. In our application, the null hypothesis is events only from background, resulting in a flat, i.e. uniform, distribution.

In this study, I have evaluated various statistical tests. To compare and evaluate the performance of different tests, I consider a simulated experiment with a background rate of $r_{bkg} = 0.18$ cts/s, signal expectation of $\lambda_{sig} = 29.65$ cts at a reference distance of 10 kpc and a refresh time $t_r = 0.5$ s. I inspect the sensitivity of analysis windows of various sizes and perform a first screening by filtering for the SNe farthest detected at the 95% success rate. To guarantee a fair comparison, the trigger thresholds for each test were evaluated in the same way as the one for the Poisson test, i.e., simulating an extended run of the experiment assuming a known background rate. For comparisons across different tests, I condense the information given by success rate curves as in Fig. 6.10 into a single number corresponding to the maximum distance that can be explored at the set success rate of 95% for a given test and window size. The results comparing the sensitivity through a selection of tests are shown in Fig. 6.11. In particular, here are shown the best-performing test based on the rate (Poisson), the EDF (KS) and spacings (RPS). As we can see, for short analysis windows, the Poisson test outperforms the others, but as the window size becomes larger, the KS and RPS tests become more sensitive and yield better results. Looking at the furthest distance probed by each test for any given window size, the Poisson test appears to be the most sensitive, with the RPS test as a close second. Out of the test statistics we studied, the Poisson and the RPS tests excelled due to their sensitivity, and in the following section we will show that for non-optimal signal shape, window size, or background rate choices, RPS can in fact outperform Poisson.

### Application to prompt detection of SN neutrino emission

Following the previous example of CC-SN neutrino detection, after analyzing different window sizes $\omega$ and different test statistics, the best choice in terms of the farthest successfully detected SN signal from the one shown in Fig. 6.9 is a window of 5 s analyzed with the Poisson test. The signal distribution used in this example is just one possible

signal that we would like to detect with our experiment. A very short list of possible SN models is available in [97], where 30 models are presented with time distributions spanning from 0.5 s to 15.4 s. When selecting the correct analysis scheme, i.e., the window sizes and tests to use, we should also study the robustness of our choice against multiple models. As an example, I consider an alternative signal distribution, modelling a neutrino burst coming from a failed CC-SN event that results in the formation of a black hole, as depicted in the right panel of Fig. 6.12.

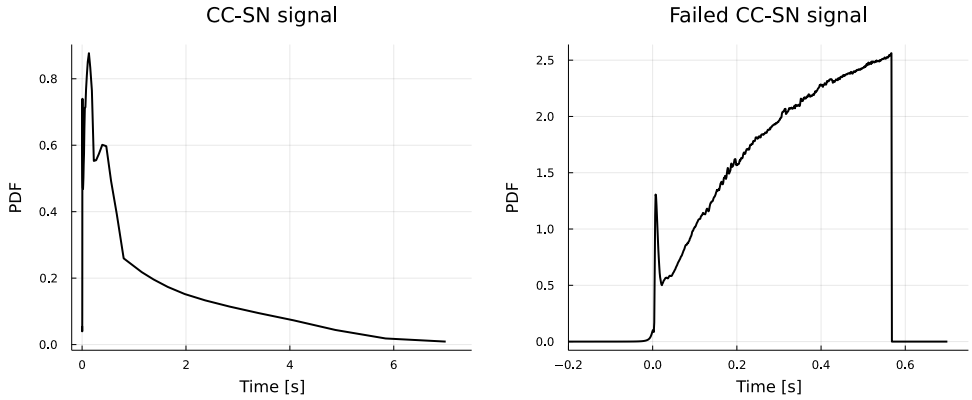



Figure 6.12.: Normalized signal distributions for a core-collapse SN (left) and failed CC-SN (right), for progenitors stars with 27 $M_\odot$ and 40 $M_\odot$ respectively. These are 1D hydrodynamical simulations performed by the Garching group [94] and named `s27_ls220` and `s40_s7b2c`.

These models are 1D hydrodynamical simulations performed by the Garching group [94]. They are the same adopted in [85] and named `LS220` and `failed-SN fast`, and they refer to progenitor stars with 27 $M_\odot$ and 40 $M_\odot$, respectively. In the latter case, the signal strength that is seen by the detector is $\lambda_{sig} = 16.39\,\text{cts}$, weaker than the one induced by the CC-SN `LS220` model and this will result in shorter distances that can be probed by the detector. I repeat the same analysis previously described, namely, estimating the maximum distance at which we can reliably detect a neutrino burst with a 95% success rate, while at the same time considering different values of the background rate, to account for possibly higher background levels in our experiment. The analysis of both the CC-SN and the failed CC-SN signals shown in Fig. 6.12 using both the Poisson test and the RPS test, and the results of this study are shown in Fig. 6.13.

Looking at these results, we notice that for both signals, as the background rate increases, the 95% detection horizon starts to decrease. Given a fixed background rate, we notice that, while using the Poisson test, it is possible to achieve the furthest detection only for a select few analysis windows, while the horizon probed via the RPS test appears to be much more robust to changes of the window size, an effect that is particularly visible in the case of the failed CC-SN signal. If we have detailed knowledge of the background affecting our experiment at any given time, and especially if we knew the time distributions of all the signals we might detect, then we could select the best combination

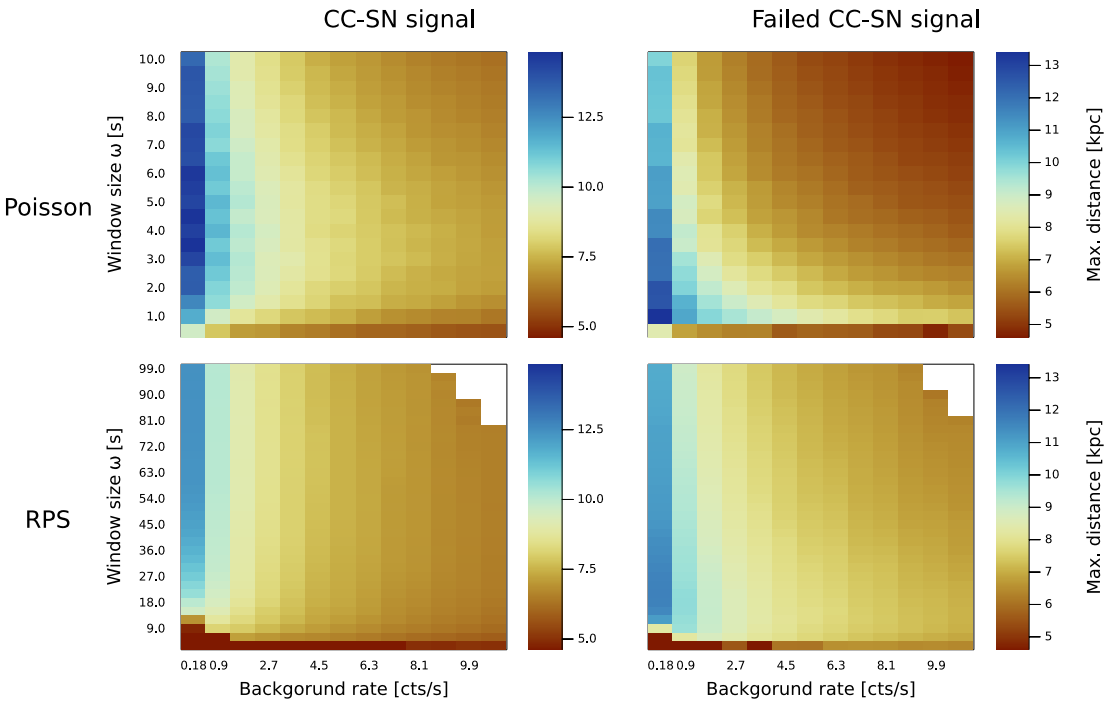



Figure 6.13.: Maximum distance probed at 95% success rate for different background rates and window sizes. The refresh time is kept constant at 0.5 s. The white corners in the bottom row plots are due to the high total event expectation surpassing $10^3$, which is currently the upper limit when it comes to the parametrization of the RPS test.

of test statistic and window sizes. Looking at the results of Fig. 6.13 it would appear that the Poisson test is the most sensitive, provided that we have detailed knowledge of both the background and the signal. Although there are lists of different possible signals, in order to maximize the detection of all signals we might run a dedicated analysis for each proposed model. Such an approach would translate in running a multitude of parallel analysis streams, each with their own optimized window size for a given background rate. Our objective is to integrate our analysis with the SNEWS alert system; thus, we must curtail the FAR of our final analysis. If we were to use independent analysis windows, then the FAR of each analysis stream would have to decrease to account for the total number of windows, which would discourage having too many of them. Additionally, such an approach may well not be the best suited one when we consider the sensitivity of our analysis to unknown signals whose model was not considered during the development of the analysis scheme. Furthermore, during the operation of our experiment, the background may realistically experience fluctuations in time, which could affect the sensitivity of the analysis windows. In order to limit the number of analysis windows needed, and in order to retain good sensitivity against background fluctuations or with respect to unknown signals, it is reasonable to consider using few

analysis streams, each with their own window size and using either the Poisson test or the RPS test. To gauge the potential and shortcomings of either test, I test the sensitivity of the analyses optimized against a known signal with a known background against another signal distribution at different background levels.

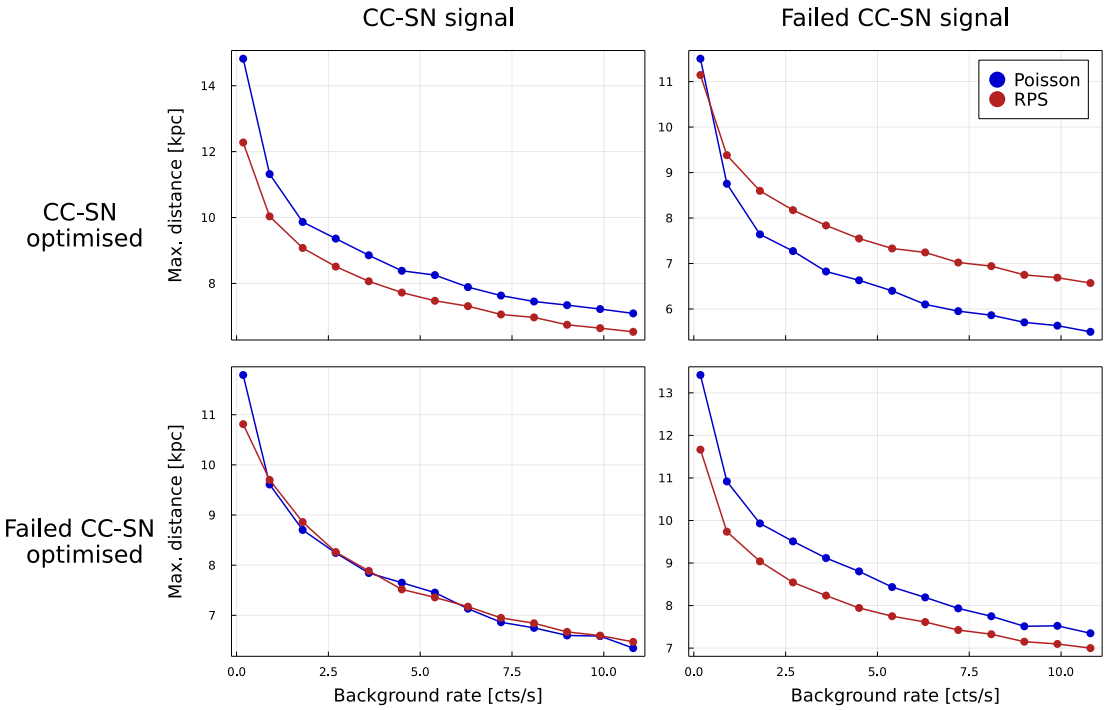



Figure 6.14.: Maximum distance probed at a 95% success rate as a function of time with respect to two sample signals, the Core-Collapse SN and the failed Core-Collapse SN, obtained using analysis windows optimised on each of the tested signals.

The results of this study are shown in Fig. 6.14. If we consider a CC-SN signal and a nominal background rate of 0.18 cts/s, we can select one of the best window sizes for each test, i.e., the window sizes that allow the detection of the furthest sources. Once a window has been selected for each test, the background rate is increased and I estimate the new horizon with 95% trigger success. These results are reported in Fig. 6.14 (top-left) where we notice that both horizons are decreasing, as expected, while the Poisson one remains dominant. If instead of the CC-SN signal we try to detect the failed CC-SN signal, while retaining the CC-SN-optimized windows, in Fig. 6.14 (top-right) we see that the horizon delivered by the RPS windows is roughly the same as the Poisson one for the nominal background rate and for higher backgrounds it becomes the better one, meaning that the RPS analysis proves to be more sensitive to narrower than expected signals compared to the Poisson analysis. Repeating the same analysis, this time optimizing with respect to the failed CC-SN signal and then testing against the CC-SN one, we notice that both analyses are equally sensitive to the latter signal.

These results, coupled with the overall picture presented in Fig. 6.13, show that the RPS analysis can be more sensitive than the Poisson one when considering different signal distributions, as in a real-case scenario, especially when the detector operates in high background conditions. This shows the advantages of non-parametric test statistics for the application in a real-case scenario, where the details of the sought-for signal are unknown and not all the experimental conditions are fully under control: under these conditions, non-parametric tests tend to be more robust. If on the other hand, the observations match the expectation of a Poisson process with a known rate and the correct signal model is considered, then it is hardly surprising that the Poisson test would be more sensitive, just like a likelihood-ratio approach that considers the correct model with few degrees of freedom shows the largest rejection of the null-hypothesis, as seen in Sec. 6.1.2.

Lastly, I would like to point out that these studies were conducted and published prior to the development of the $\mathrm{BSS}_{\mathrm{min}}$ statistic, hence the reason why it is not included. Nevertheless, looking at the result presented above for the RPS statistic, we notice that the width of the assumed signal distribution is narrow with respect to the size of the analysis window that maximizes its discovery. Knowing this, and referencing Fig. 3.11, we notice that we fall in the category of simulations where the RPS and $\mathrm{BSS}_{\mathrm{min}}$ are equipollent, thus we expect the $\mathrm{BSS}_{\mathrm{min}}$ results to be similar to the ones presented here.

# 7. Conclusions and future outlook

The main topic of this thesis was the study and development of non-parametric goodness of fit tests for signal discovery and limit setting in one and multiple dimensions.

In the univariate case, I developed new test statistics based on Order Statistics, more specifically based on spacings (or gaps), which are the intervals between samples in a given dataset. If an unknown signal is present in a dataset, it could produce clusters of samples which give rise to small spacings between events in the vicinity of the signal. The spacings' distribution can be used to provide more sensitive test statistics, such as the three novel ones presented in this thesis: the "Recursive Product of Spacings" (RPS), which at its core considers the product of spacings between samples; the "Best Sum of Spacings" ($\text{BSS}_{min}$) which considers the smallest distance comprising $k$ consecutive ordered samples for all values of $k$ and then selecting the one that is most significant, i.e. whose p-value is the smallest; the "Best Sum of Ordered Spacings" ($\text{BSOS}_{min}$) which considers the sorted set of spacings and selects the most significant sum of $k$ smallest such elements. Regarding the probability distribution of these test statistics, I derived analytic closed-form results for the Sum of Ordered Spacings (Appendix A) and provided a general integral solution for the others (Appendix B). The difficulty of deriving closed-form results stems from the use of higher-rank spacings (i.e. spacings between non-consecutive ordered samples), which entail a highly correlated variable set that proves very difficult to solve analytically as the overall number of samples grows. Ultimately, a numerical approximation of the cumulative distribution function is necessary: I discussed an approach based on simulations and a novel estimate of its statistical uncertainty based on Order Statistics (Appendix C). Given an approximate cdf of a test statistic, the uncertainty of a p-value estimate $p$ will depend on its absolute value: the smaller the value of $p$, the larger its expected relative error since it is more susceptible to statistical fluctuations during the sampling process. For the discovery-oriented tests, the approximate distributions I report reach a relative error of approximately 100% for p-values of the order of $10^{-7}$; such a high error is not alarming, since it does not alter the magnitude of such an extreme quantile.

Additionally, relying on the transformation of a Dirichlet random variable into a set of independent Beta variables [10], I showed how to transform spacings into a set of independent Gaussian variables, allowing to perform goodness of fit tests using tests statistics devoted to the analysis of time-series, such as the RUN statistic [45, 46].

The performance of these proposals was tested and compared against well-known ECDF-statistics, such as Kolmogorov-Smirnov (KS), Anderson-Darling (AD) and Cramér–von Mises (CvM). The results of these studies show that the $\text{BSS}_{min}$ test is the best perform-

ing one, improving the discovery sensitivity by several orders of magnitude for tested signals (Fig. 3.11). When dealing with narrow signals, the RPS test is as good as $\text{BSS}_{min}$, followed closely by the adapted RUN test statistic. Based on these results, I presented an application of the RPS test to an online trigger for supernova detection in the context of the RES-NOVA experiment: I discussed how to optimize the analysis in order to maximize the discovery success rate for the RPS and Poisson tests in the case of an expected signal distribution and background event rate ($\nu_{bkg}$) and then showed that RPS proves to be more robust than Poisson when analysing an unexpected signal, especially if affected by higher levels of $\nu_{bkg}$ (Fig. 6.14).

Apart from signal discovery, I also investigated the application of spacing-based test statistics to the problem of setting upper limits in the presence of unknown backgrounds. Interestingly enough, when it comes to spacings, discovery and limit setting appear to be symmetric problems: small spacings lead to small p-values in the former, while large spacings yield higher confidence levels to exclude a proposed rate in the latter. In this light, it is possible to interpret the Optimum Interval method (introduced by Yellin [1]) as the counterpart to $\text{BSS}_{min}$, where instead of considering the higher-rank spacing with the smallest p-value one chooses the one with the largest p-value. Similarly, in this thesis, I presented two new test statistics, namely the "Best Sum of Ordered Spacings" ($\text{BSOS}_{max}$), which is the counterpart of the discovery-oriented $\text{BSOS}_{min}$, and the "Product of Complementary Spacings" (PCS), which can be interpreted as the counterpart of the Moran statistic. Both these tests leverage the presence of regions in the analysis window with low event density, regardless of their number or location relative to one another, to estimate the underlying uniform event distribution in the cumulative space. These features allow these tests to be viable alternatives for the analysis of rare process searches that aim to set competitive limits on their parameters of interest, as shown by the analysis of the spin-independent WIMP-nucleus scattering cross-section obtained for all methods using the data released from CRESST [61] (Fig. 6.8). Additionally, the newly proposed tests are the best suited, among non-parametric methods, to setting the most competitive limits when faced with peaked multimodal backgrounds (Fig. 4.6).

Finally, I tackled the challenging problem of multivariate goodness of fit tests, presenting two main approaches for such tests: either by considering the volumes identified by each sample or by taking into account their projections. Based on these, it is possible to reduce the complexity of the problem and break it down into a series of independent univariate datasets, which can be analysed using univariate test statistics. The tests developed with these methods perform an unbinned analysis of the data and do not need any trials factor or look-elsewhere correction since the multivariate data is analyzed all at once. These novel methods allow testing for the presence of a signal beyond the known background expectation, as well as setting a limit on a signal's event rate in cases where the background is not well modelled. The sensitivity of these proposals was tested in the context of mock signal searches (Fig. 5.7-5.8). Similarly, the limit-setting capabilities of these methods were assessed in simulated rare process searches under

a variety of background behaviours (Fig. 5.10-5.12). The results of these comparisons suggest that projection-based tests are better suited for discovery applications, while volume transformation tests are preferable when seeking to set upper limits.

The use of these tests is made possible thanks to a multivariate probability integral transformation, which is achieved either analytically in the case of simple models or numerically, using a Normalizing-Flow for complex models.

The test statistics and methods developed in this thesis hold great promise for further refinement and broader adoption, even beyond the realm of physics. The development of user-friendly software tools and packages incorporating these non-parametric tests would greatly facilitate their adoption by researchers. Most of the tests described in this thesis are available through the SpacingStatistic.jl [47] software package, which will be continuously maintained and updated, alongside software implementing Normalizing-Flows which we are currently developing. Since simple models are difficult to come by, Normalizing-Flows will assume a critical role in being able to apply the methods I describe for multivariate goodness of fit tests. For this reason, this is one of my main research goals beyond the scope of this thesis.

# A. Appendix: Distribution of Sum of Ordered Spacings

Given $n$ i.i.d. samples $X_i$, $U_i$ denotes the variables after the probability integral transformation. Denoting by $U_{(i)}$ the corresponding ordered statistics, $S_i$ describes the first-rank spacings, i.e. the distance between consecutive $U_{(i)}$, and similarly $S_{(i)}$ is the ordered set of spacings.

Here I provide the analytic form of the distribution for the sum of extreme ordered spacings, i.e. either the $k$ smallest $S_{(i)}$ or the $k$ largest. This is done both for the case where the boundaries of the unit interval are included as fictitious order statistics ($U_{(0)} = 0$ and $U_{(n+1)} = 1$), or when only the inner spacings are considered ($\{S_2, S_3, ..., S_n$). This Appendix will closely follow [98], where this derivation was first presented.

The starting point of my derivation is the distribution of the smallest ordered spacing $S_{(1)}$:

$$\Pr\{S_{(1)} = x|n, L = 1\} = n(n+1)\left[1-(n+1)x\right]^{n-1} \tag{A.1}$$

$$\Pr\{S_{(1)} \leq x|n, L = 1\} = 1-\left[1-(n+1)x\right]^{n}. \tag{A.2}$$

where $n$ indicates the number of samples and $L$ the sum of all spacings, i.e. the length of the interval. This distribution can be easily derived by induction using a recursive formula, which allows calculating the probability density function of $S_{(1)}$ given $n$ samples as follows:

$$\begin{aligned}
\Pr\{S_{(1)} = x|n, L = 1\} &= \Pr\{S_1 = x|n, L = 1\} \cdot \Pr\left\{ S_{(1)} \geq \frac{x}{1-x} \middle| \, n-1, L = 1 \right\} + \\
&+ \int_x^{1-nx} \Pr\{S_1 = y|n, L = 1\} \cdot \Pr\left\{ S_{(1)} = x \middle| \, n-1, L = 1-y \right\} dy = \\
&= n(1-x)^{n-1}\left(1-\frac{nx}{1-x}\right)^{n-1} + \int_x^{1-nx} n(1-y)^{n-1}\frac{n(n-1)}{1-y}\left(1-\frac{nx}{1-y}\right)^{n-2} = \\
&= n\left[1-(n+1)x\right]^{n-1} + \int_x^{1-nx} n^2(n-1)\left[1-y-nx\right]^{n-2} dy \\
&= n\left[1-(n+1)x\right]^{n-1} + n^2\left[1-(n+1)x\right]^{n-1} \\
&= n(n+1)\left[1-(n+1)x\right]^{n-1}
\end{aligned} \tag{A.3}$$

where the distribution of $S_1$ is given by Eq. 2.26.

The sum of the smallest $k$ ordered spacings is defined in Eq. 3.27:

$$S_{(k)}^{min}(n) = \sum_{i=1}^{k} S_{(i)}.$$

The distribution of $S_{(k)}^{min}(n)$ can be derived by induction.

## A.1. Sum of minima: $k = 2$

Considering $S_{(2)}^{min}$, its joint distribution with the smallest ordered spacing is:

$$\begin{aligned}\Pr\{S_{(1)} = x, S_{(2)}^{min} = s|n, L = 1) =& \Pr\{S_{(1)} = x|n, L = 1\}\cdot \\ & \cdot \Pr\{S_{(2)}^{min} = s|n, L = 1, S_{(1)} = x\}\end{aligned} \tag{A.4}$$

In order to derive an expression for $\Pr\{S_{(2)}^{min}|S_{(1)}\}$ we can consider that once we have chosen the length on the smallest spacing, by definition all the other spacings need to be longer or equal to this minimum length. We can then proceed to subtract $S_{(1)}$ from the length of all the other spacings:

$$S_{(i)} - S_{(1)} = S_{(i-1)}^{*} \text{ , for } i = 2, ..., n+1 \tag{A.5}$$

This operation leaves a reduced set of spacings (since subtracting $S_{(1)}$ from itself results in 0, do it can be discarded), where the reduced spacings retain their ordering:

$$\{S_{(1)}, ..., S_{(n+1)}\} \rightarrow \{S_{(1)}^{*}, ..., S_{(n)}^{*}\} \tag{A.6}$$

and they sum up to:

$$\sum_{i=1}^{n} S_{(i)}^{*} = 1 - (n+1)S_{(1)}. \tag{A.7}$$

The set $\{S_{(1)}^{*}, \ldots, S_{(n)}^{*}\}$ can be interpreted as ordered uniform spacings determined by sampling $n-1$ values in an interval of length $1-(n+1)S_{(1)}$. Given this rearrangement, we can express the sum of $k$ minima using this new set of spacings:

$$S_{(k-1)}^{min*} = \sum_{i=1}^{k-1} S_{(i)}^{*} = \sum_{i=1}^{k} (S_{(i)} - S_{(1)}) = S_{(k)}^{min} - k \cdot S_{(1)} \tag{A.8}$$

This allows to rewrite the conditional distribution of $S^{min}_{(2)}$ as:

$$\begin{aligned}\Pr\{S^{min}_{(2)} = s|n, L = 1, S_{(1)} = x\} &= \Pr\{S^{min*}_{(1)} = s - 2x|n-1, L = 1-(n+1)x\} \\ &= \left(\frac{1}{1-(n+1)x}\right) n(n-1)\left[1 - \frac{n(s-2x)}{1-(n+1)x}\right]^{n-2}.\end{aligned} \tag{A.9}$$

Putting Eq. (A.1, A.4, A.9) together we obtain:

$$\begin{aligned}\Pr\{S_{(1)} = x, S^{min}_{(2)} = s) &= (n+1)n^2(n-1)\left[1-(n+1)x\right]^{n-2}\left[\frac{1+(n-1)x-ns}{1-(n+1)x}\right]^{n-2} \\ &= (n+1)n^2(n-1)\left[1-(n-1)x-ns\right]^{n-2}\end{aligned} \tag{A.10}$$

The support of $S^{min}_{(2)}$ is $\left[0, \frac{2}{n+1}\right]$ and the support of $S^{min*}_{(1)}$ with $n-1$ samples is $\left[0, \frac{1}{n}\right]$, thus the joint distribution is bound within a triangle as shown in Fig. A.1.

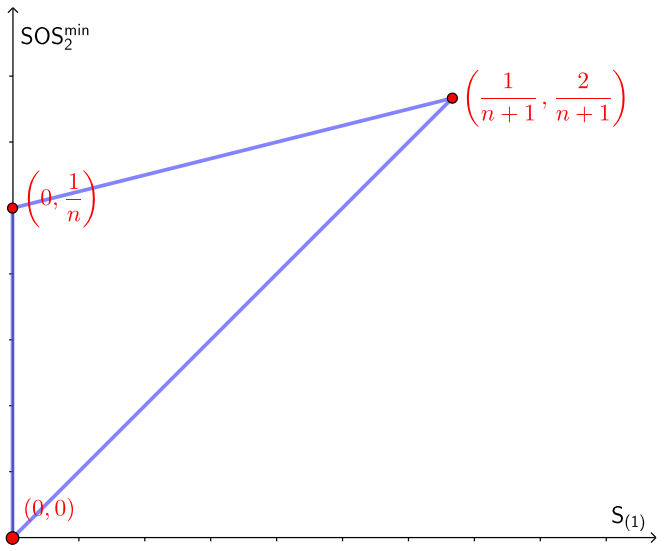



Figure A.1.: Support of the joint distribution $\Pr\{S_{(1)}, S^{min}_{(2)}\}$

The distribution of $S^{min}_{(2)}$ is obtained by marginalizing Eq.A.10 over $S_{(1)}$:

$$\begin{aligned}
\Pr\{S_{(2)}^{min} = s\} &= \int_0^{\frac{s}{2}} (n+1)n^2(n-1)\left[1+(n-1)x-ns\right]^{n-2} dx \qquad \text{for } 0 \le s \le \frac{1}{n} \\
&= \int_{\frac{ns-1}{n-1}}^{\frac{s}{2}} (n+1)n^2(n-1)\left[1+(n-1)x-ns\right]^{n-2} dx \quad \text{for } \frac{1}{n} \le s \le \frac{2}{n+1} \\
&= \frac{n(n+1)}{(n-1)}\left(\left[1-\left(\frac{n+1}{2}\right)s\right]^{n-1} - \left[1-ns\right]^{n-1}\right) \quad \text{for } 0 \le s \le \frac{1}{n} \\
&= \frac{n(n+1)}{(n-1)}\left[1-\left(\frac{n+1}{2}\right)s\right]^{n-1} \qquad \text{for } \frac{1}{n} \le s \le \frac{2}{n+1}
\end{aligned} \tag{A.11}$$

## A.2. Sum of minima: $k$

Given the exact distribution of $S_{(2)}^{min}$, it is possible to make a hypothesis regarding the general distribution of $S_{(k)}^{min}$:

$$p(S_{(k)}^{min} = s|n,1) = A(k,n)\sum_{i=1}^{k} a(i,k)\left[1-\left(\frac{n+2-i}{k+1-i}\right)s\right]^{n-1} \cdot H\left(s,0,\frac{k+1-i}{n+2-i}\right) \tag{A.12}$$

where $H(x,a,b) = 1$ if $a \le x \le b$ and 0 otherwise, while the coefficients $A(k,n)$ and $a(i,k)$ are given by:

$$A(k,n) = \frac{n(n+1)!}{(n+1-k)^{k-1}(n+1-k)!} \tag{A.13}$$

$$a(i,k) = \frac{(-1)^{i-1}(k+1-i)^{k-2}}{(k-i)!(i-1)!} \tag{A.14}$$

To prove the inductive step we start from the joint distribution of $S_{(k)}^{min}$ and $S_{(1)}$:

$$\begin{aligned}
&\Pr\{S^{min}_{(k)} = s, S_{(1)} = x|n, L = 1\} = \\
&\quad = \Pr\{S_{(1)} = x|n, L = 1\}\Pr\{s^*_{k-1} = s - kx|n-1, L = 1-(n+1)x\} \\
&\quad = \Pr\{S_{(1)} = x|n, L = 1\}\left(\frac{1}{1-(n+1)x}\right)\Pr\left\{s^*_{k-1} = \frac{s-kx}{1-(n+1)x}\middle| n-1, L = 1\right\} \\
&\quad = n(n+1)A(k-1,n-1)\cdot\left(\sum_{i=1}^{k-1} a(i,k-1)\cdot\right. \\
&\qquad \left.\cdot\left[1 + x\cdot\frac{i(n+1-k)}{k-i} - s\cdot\frac{(n+1-i)}{k-i}\right]^{n-2}\cdot H\left(\frac{s-kx}{1-(n+1)x}, 0, \frac{k-i}{n+1-i}\right)\right)
\end{aligned} \tag{A.15}$$

Marginalizing over $S_{(1)}$ we have:

$$\begin{aligned}
&\Pr\{S^{min}_{(k)} = s|n, L = 1\} = \int_0^{\frac{s}{k}} \Pr\{S^{min}_{(k)} = s, S_{(1)} = x|n, L = 1\} \\
&\quad = n(n+1)A(k-1,n-1)\cdot \\
&\qquad \cdot\sum_{i=1}^{k-1}\int_{\max\left(0,\frac{s(n+1-i)-k+i}{i(n+1-k)}\right)}^{\frac{s}{k}} a(i,k-1)\left[1 + x\cdot\frac{i(n+1-k)}{k-i} - s\cdot\frac{(n+1-i)}{k-i}\right]^{n-2} \\
&\quad = \frac{n(n+1)}{(n-1)(n+1-k)}A(k-1,n-1)\cdot\left(\left(\sum_{i=1}^{k-1} a(i,k-1)\cdot\frac{(k-i)}{i}\right)\cdot\left[1 - s\cdot\frac{(n+1)}{k}\right]^{n-1} - \right. \\
&\qquad \left. - \sum_{i=2}^{k} a(i-1,k-1)\cdot\frac{(k+1-i)}{i-1}\cdot\left[1 - s\cdot\frac{(n+2-i)}{k+1-i}\right]^{n-1}\cdot H\left(s, 0, \frac{k+1-i}{n+2-i}\right)\right)
\end{aligned} \tag{A.16}$$

Looking back at Eq. A.13 notice that:

$$\begin{aligned}
\frac{n(n+1)}{(n-1)(n+1-k)}A(k-1,n-1) &= \frac{n(n+1)}{(n-1)(n+1-k)}\cdot\frac{(n-1)n!}{(n-k-1)^{k-2}(n-k-1)!} \\
&= \frac{n(n+1)!}{(n+1-k)^{k-1}(n+1-k)!} \\
&= A(k,n)
\end{aligned} \tag{A.17}$$

and:

$$
\begin{aligned}
-a(i-1,k-1)\cdot\frac{(k+1-i)}{i-1} &= -\frac{(-1)^{i-2}(k+1-i)^{k-3}}{(k-i)!(i-2)!}\cdot\frac{(k+1-i)}{i-1} \\
&= \frac{(-1)^{i-1}(k+1-i)^{k-2}}{(k-i)!(i-1)!} \\
&= a(i,k) \qquad \text{for } 2\le i\le k
\end{aligned}
\tag{A.18}
$$

The result of Eq. A.18 implies a recursion formula for the coefficients $a(i,k) = f[a(i-1,k-1)]$. Making use of this recursion we can relate any $a(i,k)$ to $a(1,k+1-i)$:

$$
a(i,k) = \frac{(-1)^{i-1}(k+1-i)^{i-1}a(1,k+1-i)}{(i-1)!} \tag{A.19}
$$

Finally, we have that:

$$
\begin{aligned}
\sum_{i=1}^{k-1} a(i,k-1)\cdot\frac{(k-i)}{i} &= \sum_{i=1}^{k-1}\frac{(-1)^{i-1}(k-i)^{i-1}a(1,k-i)}{(i-1)!}\cdot\frac{(k-i)}{i} \\
&= -\sum_{i=1}^{k-1}\frac{(-1)^{i}(k-i)^{i}a(1,k-i)}{i!}
\end{aligned}
\tag{A.20}
$$

For Eq. A.16 to satisfy the hypothesis, it is necessary that:

$$
\sum_{i=1}^{k-1} a(i,k-1)\cdot\frac{(k-i)}{i} = a(1,k) \tag{A.21}
$$

Putting together Eq. A.20 and Eq. A.21, we find a recursion rule for the coefficients of $a(1,k)$. Using this recursion, we get:

$$
\begin{aligned}
-\sum_{i=1}^{k-1}\frac{(-1)^{i}(k-i)^{i}a(1,k-i)}{i!} &= -\sum_{i=1}^{k-2}\frac{ik}{(i+1)}\cdot\frac{(-1)^{i}(k-1-i)^{i}a(1,k-1-i)}{i!} \\
&= -\sum_{i=1}^{k-m}\frac{i\cdot k^{m-1}(-1)^{i}(k-m+1-i)^{i}a(1,k-m+1-i)}{(m-1)!(i+m-1)i!} \quad \text{for } 1\le m\le k-1 \\
&= \frac{k^{k-2}}{(k-1)!}\cdot a(1,1) = a(1,k)
\end{aligned}
\tag{A.22}
$$

where we have used Eq. A.21 to express the first factor in each of the sums, allowing us to reduce the limits of the sum by means of this recursion. The result we obtain proves

the consistency of the recursion relation, proving the consistency of the initial hypothesis.

The cumulative distribution function of $S^{min}_{(k)}$ is:

$$\begin{aligned}&\Pr\{S^{min}_{(k)} \leq x|n, L=1\} = \\ &= \frac{A(k,n)}{n}\sum_{i=1}^{k}\frac{a(i,k)(k+1-i)}{(n+2-i)}\left(1-\left[1-\left(\frac{n+2-i}{k+1-i}\right)x\right]^{n}H\left(x,0,\frac{k+1-i}{n+2-i}\right)\right)\end{aligned} \tag{A.23}$$

## A.3. Sum of largest ordered spacing

Since the sum of all the spacings is 1, knowing the sum of the $k$ smallest spacings allows us to know the value of the sum of the largest $(n+1-k)$ spacings:

$$S^{max}_{(k)} = \sum_{i=1}^{k} S_{(n+2-i)} = 1 - \sum_{i=1}^{n+1-k} S_{(i)} = 1 - S^{min}_{(n+1-k)} \tag{A.24}$$

which implies that:

$$\Pr\{S^{max}_{(k)} = s|n, L=1\} = \Pr\{S^{min}_{(n+1-k)} = 1-s|n, L=1\} \tag{A.25}$$

thus the probability and cumulative density functions are:

$$\begin{aligned}&\Pr\{S^{max}_{(k)} = s|n, 1\} = \\ &= A(n+1-k,n)\sum_{i=1}^{n+1-k} a(i,n+1-k)\left[\frac{s(n+2-i)-k}{n+2-k-i}\right]^{n-1}H\left(s,\frac{k}{n+2-i},1\right)\end{aligned} \tag{A.26}$$

$$\begin{aligned}\Pr\{S^{max}_{(k)} \leq s|n, L=1\} = \frac{A(n+1-k,n)}{n}\sum_{i=1}^{n+1-k}\frac{a(i,n+1-k)(n+2-k-i)}{n+2-i}\cdot \\ \cdot\left[\frac{s(n+2-i)-k}{n+2-k-i}\right]^{n}H\left(s,\frac{k}{n+2-i},1\right)\end{aligned} \tag{A.27}$$

where the coefficients $A$ and $a$ are the same as Eq. A.13-A.14.

## A.4. Excluding boundaries

So far, the boundaries of the unit interval, 0 and 1, have been included in the set of samples as fictitious events, $U_{(0)} = 0$ and $U_{(n+1)} = 1$. This means that in all previous derivations we considered an effective population of $n+2$ values, where $U_{(n+1)} - U_{(0)} = 1$ and the remaining $n$ values determine the spacings $S_i$.
This approach introduces some possibly unwelcome artifacts in the analysis of spacings data. For example, if we shift all the inner $n$ values slightly towards one of the boundaries, then:

$$\{U_{(1)}, ..., U_{(n)}\} \to \{U_{(1)} \pm \epsilon, ..., U_{(n)} \pm \epsilon\} \implies S_1 \pm \epsilon, \quad S_{n+1} \mp \epsilon \tag{A.28}$$

Depending on the application, it is possible that one is interested only in the spacings between the inner $n$ values, without considering how close this set is to either boundary. Therefore, distributions of known test statistics that exclude the boundaries might be of interest.

In this scenario, we will derive the spacing statistic of only the inner spacings $\{S_2, ..., S_n\}$, thus $U_{(1)}$ and $U_{(n)}$ become the new boundaries:

$$\{U_{(1)}, ..., U_{(n)}\} \to \{U^\dagger_{(0)}, ..., U^\dagger_{(n-1)}\} \to \{S^\dagger_1, ..., S^\dagger_{n-1}\} \to \{S^\dagger_{(1)}, ..., S^\dagger_{(n-1)}\}. \tag{A.29}$$

This means that given $n$ values in the no boundary scenario, we have an effective population of $n-1$ spacings in an interval with length $X_{(n)} - X_{(1)} = X^\dagger_{(n-1)} - X^\dagger_{(0)} = \ell$.

Given $\ell$, we can reuse the same distribution of a quantity we have studied with the presence of the boundaries decreasing the number of spacings from $n+1$ to $n-1$ and by rescaling the support of the distribution to an interval of length $\ell$ by means of the change of variable rule.
Looking at $\ell$ we notice that it is none other than the spacing between the extremes of the ordered values and its distribution is given by Eq. (2.24).

Given $n$ values, a quantity of interest $A$ and its distribution with boundaries $f_{w.b.}(A = x|n, L = \ell)$, in order to derive the distribution of $A$ without boundaries, $f_{n.b.}(A = x|n, L = 1)$, we have to marginalize over all possible values of $\ell$:

$$\begin{aligned} f_{n.b.}(A = x|n+1, L = 1) &= \int_0^1 p(\ell) \cdot f_{w.b.}(A = x|n-1, L = \ell) d\ell \\ &= \int_0^1 n(n-1)\ell^{n-3}(1-\ell) \cdot f_{w.b.}\left(A = \frac{x}{\ell} \middle| n-1, L = 1\right) d\ell \end{aligned} \tag{A.30}$$

# B. Appendix: Distribution of extreme sum of Spacings

Here I present a general integral solution to the distribution of test statistics based on spacings, including functions of spacings such as $\min_{i,j} f(S_{i,j})$ or $\max_{i,j} f(S_{i,j})$. In the example below, I consider the distribution of the smallest spacing of rank $k$, $S_k^{min}$, whose cumulative distribution can be expressed as:

$$\Pr\{S_k^{min} \leq x\} = 1 - \Pr\{S_{i,k} \geq x \,\forall i\} \tag{B.1}$$

where the RHS is a joint probability over all spacings of order $k$, which are not independent of one another.

To derive this distribution, we can resort to the transformation of spacings into independent random variables, in order to better express the constraint relative to the minimum or maximum of a population being respectively greater or less than a specific value. Referencing Sec. 2.5.3, we consider the transformation (Eq. 2.55) of the spacings, $S_i$, in independent Beta variates, $B_i$, which allows us to express the general spacings $S_{i,k}$ as follows:

$$\begin{aligned}
S_{1,k} &= \sum_{j=1}^{k} S_j = \sum_{j=1}^{k} (1 - B_j) \cdot \prod_{t=1}^{j-1} B_t = 1 - \prod_{j=1}^{k} B_j \\
S_{i,k} &= \sum_{j=i}^{i+k-1} S_j = \sum_{j=i}^{i+k-1} (1 - B_j) \cdot \prod_{t=1}^{j-1} B_t = \left( \prod_{j=1}^{i-1} B_j \right) \left( 1 - \prod_{j=i}^{i+k-1} B_j \right) \\
&\dots \\
S_{n+2-k,k} &= \prod_{j=1}^{n+1-k} B_j.
\end{aligned} \tag{B.2}$$

This allows us to rewrite the RHS of Eq. B.1 as a system of independent inequalities:

$$S_{i,k} \geq x \,\forall i \equiv \begin{cases} 1 - \prod_{j=1}^{k} B_j & \geq x \\ \left( \prod_{j=1}^{i-1} B_j \right) \left( 1 - \prod_{j=i}^{i+k-1} B_j \right) & \geq x \\ & \dots \\ \prod_{j=1}^{n+1-k} B_j & \geq x \end{cases} \tag{B.3}$$

In order to calculate $\Pr\{S_{i,k} \geq x\, \forall i\}$ we can simply integrate over the joint probability distribution of $\{B_i\}$ satisfying the constraints of Eq. B.3. Since the $B_i$ variables are independent, their joint probability distribution is just the product of the individual Beta distributions, thus:

$$\Pr\{S_{i,k} \geq x\, \forall i\} = \int_{constr.} \prod_{j=1}^{n} (n+1-j) b_j^{n-j}\, db_1 ... db_n \tag{B.4}$$

This integral is easy to calculate for reasonably small values of $n$, but it quickly becomes complicated as $n$ increases. The integrand of Eq. B.4 is polynomial, thus easy to integrate, but the system of constraints is not easy to linearize, especially since each $B_i$ variable is limited to the range $[0, 1]$, meaning that the extremes of integration will change depending on the value of $x$ and clipped to either 0 or 1 when outside the support of the Beta distribution.

One way of solving this integral would be to derive all intervals that satisfy the constraints of Eq. B.3 using Cylindrical algebraic decomposition and then solve the integral using symbolic integration, given that we only deal with polynomial functions.

A numerical approach to the estimation of Eq. B.4 would be that of using Monte-Carlo integration: we consider the joint distribution of $B_j$ as the prior distribution and use a simple likelihood equal to 1 when all proposed $B_j$ variables satisfy the constraints and 0 otherwise. Given this setup, we can use the Bayesian evidence to estimate the value of the integral, using MC integration routines such as Adaptive Harmonic Mean Algorithm [99], Bridge Sampling [100] or many others. If the value $x$ at which the CDF, thus the integral, needs to be calculated is small, then it might be difficult to randomly find a good point in the prior space that yields a likelihood of 1. In such cases, a good starting point for the MC chains can be the list of Beta variables corresponding to a set of $n+1$ equal spacings, which always satisfy the constraints of Eq. B.3.

An alternative approach to the estimation of the desired p-value, Eq. B.1, would be to consider the transformation of spacings $S_j$ into independent Exponential variates $Y_i$, as shown in Sec.2.5.1. Assuming we knew that the sum of all $Y_i$ was equal to $T$, we could start expressing the conditions of the RHS of Eq. B.1 in terms of $Y_i$:

$$S_{i,k} \geq x\, \forall i \equiv \begin{cases} \sum_{j=1}^{k} Y_j & \geq x \\ \sum_{j=i}^{i+k-1} Y_j & \geq x \\ & ... \\ \prod_{j=1}^{n+1-k} Y_j & \leq T - x \\ \sum_{j=1}^{n} Y_j & \geq T \end{cases} \tag{B.5}$$

where the last inequality is due to the constraint on the sum of all $Y_i$ being equal to $T$. Since the variables $Y_i$ are independent of one another, we could rewrite a similar integral to Eq. B.4:

$$\int_{constr.} n! t^{-n} \, dy_1 ... dy_n \tag{B.6}$$

where the integrand is the joint distribution of $Y_i$ conditioned on their sum, from Eq. 2.43. The integral of Eq. B.6 is not equal to the p-value we seek, but is necessary for its estimation, and appears easier to solve since the constraints that limit it are all linear and since the integrand is independent of the integration variables. This means that in order to solve the integral above we need "only" compute the volume of a $n$-dimensional convex polytope, which even if the faces are explicitly given, as is our case, is still a #P-hard problem.

This shows that solving Eq. B.6, or equivalently Eq. B.4, is not an easy task. Moreover, even if we were able to find a closed-form solution of Eq. B.4, this would end up being polynomial, given the nature of its integrated, and in such cases it is not assured that the solution can be computationally easy to use for large values of $n$, because it might rely on the difference of very large numbers, an operation that is prone to numerical cancellation when using floating-point arithmetic. This very case arises when using the exact distribution formulae of the Sum of Ordered Spacings derived in Appendix A: considering Eq. A.23, the sign of the factors inside the sum changes depending on the parity of the index, and the absolute value of the factors can grow as an exponential of $k$, the rank, which is $k \leq n$.

Since we do not have a closed-form solution for the p-value we seek, I built an approximation for the cumulative distribution of $S_k^{min}(n)$, precise enough to compute meaningful p-values up to relatively extreme values as $10^{-8}$, and sample sizes $n$ of up to 1000. Finally, the distribution of $\mathrm{BSS}_{\mathrm{min}}$ will also be approximated numerically, similarly to the distribution of the RPS test. The approximations are based on simulations, and details about the fitting procedure are described in Appendix C.

# C. Appendix: Numerical interpolation of approximate distributions

Most of the test statistics I consider in this thesis are functions of higher rank spacings $S_{i,k}$ ($k > 1$), which tend to produce a set of highly interdependent variables that renders the usual formulations of their cumulative distribution hard to untangle and solve for, as seen in Appendix B.

In order to use these tests, I resort to numerically approximating their cumulative distributions, with the goal of computing meaningful p-values of up to $10^{-7}$ or $10^{-8}$ for signal-discovery applications and up to $10^{-6}$ for limit setting applications.

In the following, I consider the RPS statistic to show the approximation and fitting procedure for test statistics that develop over only one dimension (RPS($n$) has only $n$, the number of observed samples, as a degree of freedom). The approximation and fitting procedures are discussed in Sec. C.1 and Sec. C.2 respectively. In Sec. C.3 I discuss the error of the p-value estimates produced by these approximate distributions. The same approximation and interpolation procedure of the RPS statistic is used to produce the distributions of $\text{BSS}_{min}$, $\text{BSOS}_{min}$ and $\text{BSOS}_{max}$, given the distributions of $S_k^{min}$, $S_{(k)}^{min}$ and $S_{(k)}^{max}$, which are discussed in Sec. C.4. This discussion will closely follow [25], where it was initially reported.

## C.1. Approximate Distribution

The approximation of the RPS statistic is based on simulations: events with uniform distribution in the $[0, 1]$ range are drawn for a given $n$, collecting $N = 2 \cdot 10^8$ samples of $RPS^*(n)$. Such simulation could be directly used to calculate p-value estimates by counting the fraction of trials below or above an observed $RPS^*$ value $x$ for a fixed $n$. However, the goal is to provide a continuous and smooth function valid for any $n \leq 1000$. For this, I use simulated data to infer the values $x$ of our test statistics corresponding to a discrete list of specific quantiles $p \in [10^{-7}, 1 - 10^{-7}]$. Taking the $i$-th element in the sorted simulation set gives an estimate for the value of $x(p = i/N)$. In order to improve this estimate, we could use bootstrapping [101], collecting different realisations of $x$ by resampling the original dataset with replacement, resulting in a distribution of values of $x$ for each $p$, from which we can then extract the mean and the standard deviation, indicative of the error (see Fig C.1). Instead of manually performing the bootstrapping, we can calculate the probability of each sample $x$ to represent a specific quantile $p$ if we were to sample randomly with replacement. For simplicity, let us consider rational

quantiles that can be expressed in the form $p = \frac{k}{N}$; the probability that the $i$-th sample could end up representing the $k$-th quantile is:

$$\pi_{k,i} = F_B\left(\frac{i}{N}\middle| k, N+1-k\right) - F_B\left(\frac{i-1}{N}\middle| k, N+1-k\right) \tag{C.1}$$

where $F_B(t|a,b)$ is the cumulative function of the Beta distribution with parameters $(a,b)$ estimated at $t$. The distribution $\text{Beta}(k, N+1-k)$ represents the $k$-th order statistic of the uniform distribution [4], i.e. the $k$-th smallest element of a set on $N$ uniformly distributed random variable. Eq. C.1 corresponds to the limiting case of performing an infinite number of bootstrapping steps and can be used to quickly estimate the mean and standard deviation of all $x(p)$ for a choice on $n$, especially when dealing with large datasets:

$$\text{E}\left[x\left(\frac{k}{N}\right)\right] = \sum_{i=1}^{N} x_i \cdot \pi_{k,i} \tag{C.2}$$

$$\text{Std}\left[x\left(\frac{k}{N}\right)\right] = \sqrt{\sum_{i=1}^{N}\left(x_i - \text{E}\left[x\left(\frac{k}{N}\right)\right]\right)^2 \cdot \pi_{k,i}}. \tag{C.3}$$

It would be inefficient to produce such simulation for any $n$, and hence I repeat the above procedure for only 180 different choices of $n$ between 2 and 1000 following approximately a logarithmic spacing.

## C.2. Fitting procedure

Using Eq. C.2 and Eq. C.3 we are able to define a grid of points with mean $\mu(n,p)$ and standard deviation $\sigma(n,p)$. The goal is to estimate a set of points $\hat{x}(n,p)$, used to interpolate and infer the distribution of the test statistic for all values of $n$ and $p$ defined above. The points $\hat{x}(n,p)$ are allowed to deviate from the means $\mu(n,p)$ within the uncertainties $\sigma(n,p)$, and can thereby provide a more accurate approximation by smoothing out stochastic noise. Additionally, points from the analytic solution for $n = 1$ (Eq. 3.23) are added to the list as anchor points at the boundary.

Given a trial set $\tilde{x}(n,p)$, I interpolate a cubic spline polynomial across the values of $n$ for each value of $p$, similarly to the fits shown in Fig C.1. Given one such cubic spline, I evaluate the third derivative on both sides of each node, calculating the square of their difference and summing up across all nodes. Since we are using cubic splines, the third derivative is not continuous, and the "size" of the discontinuity is indicative of the smoothness of the interpolation. Summing up the contributions form all nodes of all cubic splines constructs the smoothing cost function. The construction of this cost function is based on [102, 103, 104], where smoothness is treated very similarly. The estimation of the cubic spline coefficients and the evaluation of the smoothness cost function can be represented as a quadratic objective function, which we want to minimize:

$$G(\tilde{x}) \propto \frac{1}{2}\tilde{x}^T \cdot Q \cdot \tilde{x} + \bar{h}^T \cdot \tilde{x} \tag{C.4}$$

In addition to obtaining a smooth fit, there are also some additional constraints that need to be considered: monotonicity and sum of squared residuals.

Since the samples $\tilde{x}(p|n)$ should represent a cumulative density function, then it is important they are properly ordered, ensuring that $\tilde{x}(p_i|n) \leq \tilde{x}(p_j|n)$ for $i \leq j$. This is ensured by including a number of linear inequality constraints modelled as a linear constraint matrix:

$$A \cdot \tilde{x} \leq b \tag{C.5}$$

Lastly, we assume that the values $\tilde{x}(n,p)$ are normally distributed with means $\mu(n,p)$ and standard deviations $\sigma(n,p)$. Since we want to move away from the initial values $\mu(n,p)$ to obtain a smoother fit, it is important to limit this movement the further away we get. We do so by considering the sum of squared residuals, which is a typical measure to account for the global deviation from the mean. Since we assume gaussian deviations, the sum of all squared residuals can be modelled by a $\chi^2$ distribution with $m$ degrees of freedoms, where $m$ is the total number of parameters, i.e. the number of nodes. Given this distribution, we can estimate the value of the cost function to be limited to the mean $(m)$ plus one standard deviation $(\sqrt{2m})$ of the $\chi^2$ distribution, thus:

$$\sum_{i=1}^{m} \frac{(\tilde{x}_i - \mu_i)^2}{\sigma_i^2} \leq m + \sqrt{2 \cdot m} \tag{C.6}$$

Fig. C.1 shows a fitted spline representation of $\hat{x}(n|p)$ for different values of $p$. Based on the resulting list of corresponding $p$ and $\hat{x}$ values, that we obtained for any $n$, we generate another spline interpolation as the approximation of the desired cumulative distribution $F(\hat{x};n)$ for a given $n$. As the cumulative distribution function $F$ is strictly monotonous in $\hat{x}$, we use the [105] monotonic spline interpolation on the points $[\hat{x}(p|n), p]$ to produce the final CDFs, shown in Fig. 3.7 for a few values of $p$.

## C.3. Error estimation

Finally, we are also able to estimate the precision of our approximation. Given any set of i.i.d. random variables, such as $x$, the corresponding list of estimated quantiles $p$ represents a random set of uniform variates. For any rational quantile $p_{test} = \frac{k}{N}$ we can estimate it 98% credible interval $[p_{0.01}(p_{test}), p_{0.99}(p_{test})]$ using the distribution of the $k$-th order statistic Beta$(k, N+1-k)$:

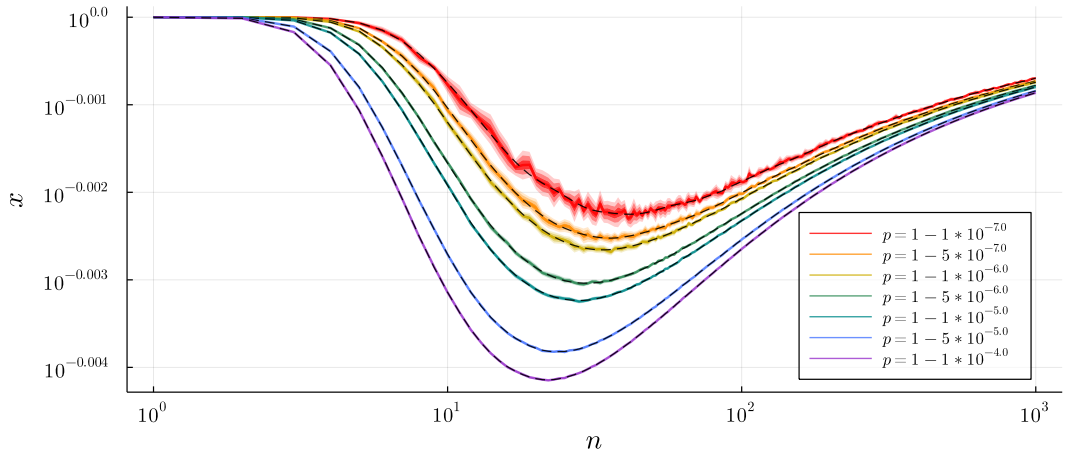



Figure C.1.: Example of spline fitted $x$-values across $n$ for a few extreme p-values. The colored bands show the 1, 2 and 3 sigma bands estimated via bootstrapping, the black, dashed lines show the approximations by the spline fits.

$$p_{0.01}\left(\frac{k}{N}\right) = F_B^{-1}\left(0.01|k, N+1-k\right)$$
$$p_{0.99}\left(\frac{k}{N}\right) = F_B^{-1}\left(0.99|k, N+1-k\right) \tag{C.7}$$

where $F_B^{-1}$ is the inverse of the cumulative distribution $\text{Beta}(k, N+1-k)$. Given this credible interval, we calculate the relative error of $p_{test}$ against the extrema of the interval, considering the largest value representative of the relative error of a random ECDF up to a specified credible level. The results of the estimated relative error for our choice of $N = 2 \cdot 10^8$ and for quantiles as low as $p = 10^{-7}$ are shown in Fig. C.2.

As expected, the errors are increasing towards smaller p-values and exhibit an approximately linear behaviour in the log-log plot. We see that the estimated upper bound of the relative error for a p-value of $10^{-3}$ is below 1%, while for a p-value of $10^{-5}$ it increases to $< 10\%$ and ultimately to $< 100\%$ for p-values of $10^{-7}$. Such a "large" relative error for small p-values may sound alarming at first, but estimating a p-value of $10^{-7}$ and knowing it could actually be closer to $2 \cdot 10^{-7}$ would hardly change the statistical interpretation of a result.

In order to show the validity of these results, we compute the relative error of our approximate distributions against a test dataset containing 10 times more samples using bootstrapping. We do so for a few choices of number of events $n$, and in Fig. C.3 it can bee seen that the behavior of the relative error is in complete agreement with our analytic estimates of Fig. C.2.

So defined, the relative error $\delta(p|N)$ is a function of the quantile $p$ and number of

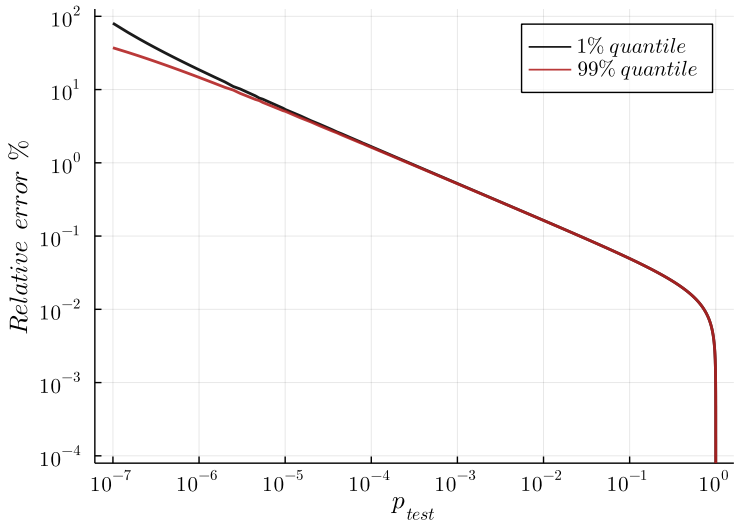



Figure C.2.: Estimated relative error of empirical p-value with respect to the 98 % credible interval and $2 \cdot 10^8$ samples. The vertical axis reports the scale of the relative error in percent for two extremes, the 1% and the 99% quantile of the order statistic distribution.

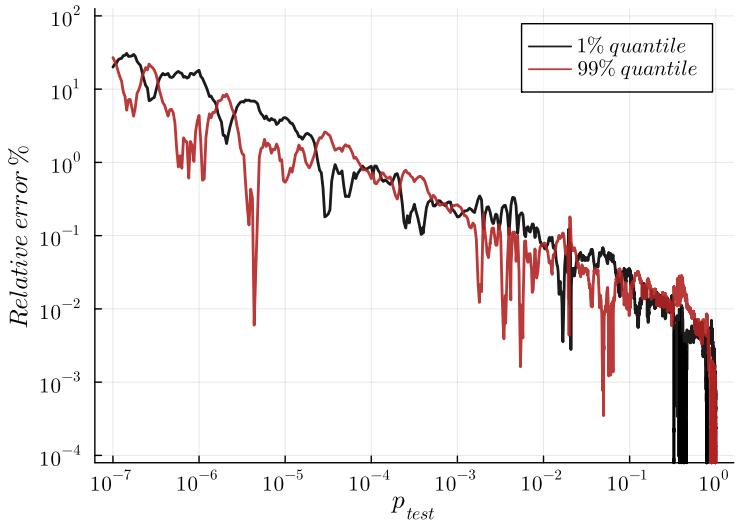



Figure C.3.: Estimated relative error of fitted p-value with respect to p-values obtained via bootstrapping. The vertical axis reports the scale of the relative error in percent for two extremes, the 1% and the 99% quantile of the bootstrapping distribution. Results for $n = 75$.

samples $N$, but this relationship can also be inverted in order to determine the number of samples necessary to achieve a desired relative error for a specific quantile: $N(p|\delta)$. Our choice of $N = 2 \cdot 10^8$ was in fact guided by the requirement of having a relative error lower than 100% for a p-value of $10^{-7}$ in at least 99% of cases.

It is worth stressing that these estimates of the relative error are accurate with respect to the ECDF that was sampled for each independent $n$, but might be subject to small changes after the smoothing fit we performed in order to regularize and infer the distributions for all missing values of $n$.

## C.4. Interpolation over $n$ and $k$

So far, I discussed the fitting procedure only in one dimension, i.e. for test statistics of the form $T(n)$, where only the number of samples $n$ is varied.

When considering the distribution of statistics such as the smallest or largest higher rank spacings, $S_r^{min}(n)$ and $S_r^{max}(n)$ respectively, we notice that they develop along two dimensions, the number of samples $n$ and the rank $r$, thus a slightly different fitting procedure is needed for these quantities.

For a given $n$ and $r$, the approximate distribution of $S_r^{min}(n)$ is obtained as described above, by means of repeated simulations, which yield an estimate of the average position of the test statistic value $\mu(n, r, p)$ and its standard deviation $\sigma(n, r, p)$ for different p-values $p$. Considering the subspace $(n, r)$, we notice that this region is not rectangular, but triangular, since $r \leq n$. The result of sampling this triangular region at selected values of $n$ and $r(n)$, produces an irregular triangular grid, such as the one shown in Fig. C.4. Due to this, we cannot resort to a simple bicubic interpolation scheme on an irregular grid as we did before but would need to resort to finite element interpolation over triangular elements. This approach would yield an approximate surface for each value of $p$, over which it would be possible to calculate the derivative at each node and produce a smooth fit using a cost function similar to Eq. C.4. This approach can be computationally intensive, and during testing, I noticed that high numerical precision is required not to produce artefacts in the final result. In order to simplify the problem, one could consider applying the smoothing only to surfaces corresponding to extreme p-values, since those are the ones where the statistical noise is more prominent.

In first approximation, I adopt the values $\mu(n, r, p)$ as our estimates of the test statistic value regarding the $S_r^{min}(n)$ statistic, whose approximations rely on $2 \cdot 10^9$ samples, allowing to estimate p-values as low as $10^{-8}$, and reducing the statistical uncertainty on p-values above $10^{-7}$.

Given a list of quantiles $\bar{\boldsymbol{p}} = \{p_1, ..., p_W\}$, a list of number of samples values $\bar{\boldsymbol{n}} = \{n_0, n_1, ..., n_V\}$ and for each $n_i$ a corresponding list of rank values $\bar{\boldsymbol{r}}_i = \{r_{i,1}, ..., r_{i,j}...\}$, the approximate distibution of $S_{r_{i,j}}^{min}(n_i)$ is collected. We refer to the estimated test statistic value of these approximations as $x_{node}(n, r, p)$. In order to construct the approximate distribution of $S_b^{min}(a)$, we need the test statistic values $x(A, B, p)$ for each quantile in $\bar{\boldsymbol{p}}$, indicated as $x_{res}$ in red in Fig. C.4. For a given value of $p$, for each value of $n_i$, the

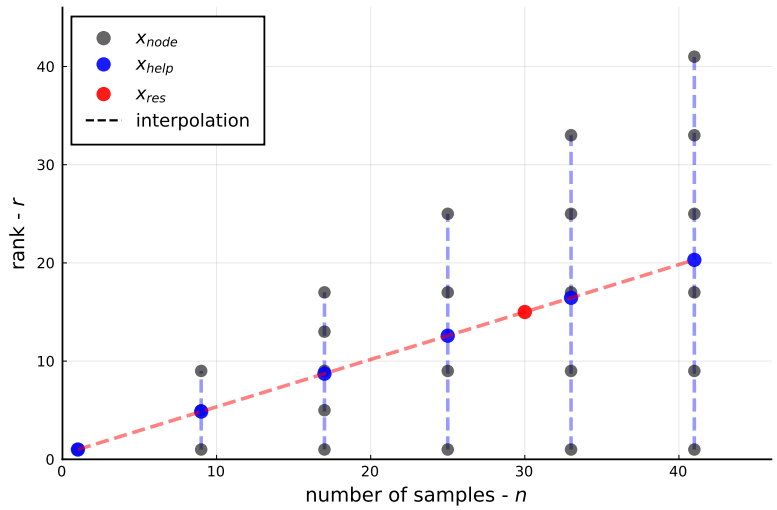



Figure C.4.: Interpolation of test statistic value for a given p-value based on a triangular grid.

values $x_{node}(n_i, \bar{\boldsymbol{r}}_i, p)$ can be use to calculate a set $x_{help}(n_i, r(n_i), p)$ such that all points $(n_i, r(n_i))$ lie on the same line as $(1, 1)$ and $(A, B)$, as shown in blue in Fig. C.4. The values $x_{help}$ are calculated interpolating along each row of ranks, $\bar{\boldsymbol{r}}_i$ for each $n_i \in \bar{\boldsymbol{n}}$; the interpolation is carried out with a cubic spline just like the described for the RPS statistic above. The values $r(n_i)$ need not be integers, since they are just auxiliary variables in this fit. Finally, the value of $x_{res}$ is obtained by interpolating across the auxiliary variables $x_{help}$. To recap, this triangular interpolation procedure relies on the initial grid of points $x_{node}(p)$ to calculate a list of auxiliary variables $x_{help}(p)$, which then is the base of a second interpolation that yields the value of $x_{res}(p)$, all of this for each quantile $p \in \bar{\boldsymbol{p}}$. As a note, the values of $x_{node}(n_i, 1, p)$ and $x_{node}(n_i, n_i, p)$ can be calculated analytically (the former from Eq. A.2, the latter can be easily derived) and used during the interpolation to improve the estimates for extreme ranks. This interpolation scheme is depicted in Fig. C.4.

Such an interpolation scheme is also used to produce an approximate distribution for the Sum of Ordered Spacings. Although I derived the exact distribution for $S^{min}_{(k)}$ and $S^{max}_{(k)}$, their cumulative distributions are not numerically stable for large values of $n$. For example, considering Eq. A.23, the sign of the factors inside the sum changes depending on the parity of the index, and the absolute value of the factors can grow as an exponential of $k$, the rank, which is $k \leq n$. This leads to numerical cancellation when $n$ and consequently $k$ become large. In order to accurately use this formula one needs to use increased precision floating-point arithmetic, which can become costly both in terms of memory and time when calculating a test statistic like $\text{BSOS}_{max}$, which entails calculating $S^{max}_{(k)}$ for all $k$. In the case of $S^{min}_{(k)}$ or $S^{max}_{(k)}$, we do not need to rely on simulations, but can use the exact distribution to find the exact value for $x_{node}$. Thus the "approximation" comes only from the interpolation procedure, not the sampling.

# D. Appendix: Normalizing Flows

A *Normalizing Flow* (NF) is a chain of transformations used to convert a complex probability distribution into a simpler one, for example, into a Standard Normal distribution, hence the name. Normalizing Flows, first introduced by Tabak and Vanden-Eijnden [106], are built using Deep Neural Networks and trained on samples of the target distribution, ensuring the following:

- the input and output dimensions of samples are the same
- the transformation is bijective and invertible
- efficient (and differentiable) computation of the determinant of the Jacobian of the transformation.

Normalizing Flows are mainly used as generative models, since drawing samples from the Standard Normal distribution is easy and transforming them (backwards) yields i.i.d. random variables distributed according to the target distribution.

Such a feature is particularly useful in the context of Bayesian inference when running an MCMC. Using a Normalizing Flow allows to run Markov Chains in the transformed space, where the distribution is already Gaussian, thus no particular tuning or special choice of proposal function is necessary. This would especially prove useful in simplifying integration problems, such as the Evidence estimation, necessary to calculate Bayes Factors.

Other common uses of Normalizing Flows include clustering and classification [107], density estimation [55, 108, 109, 110], and variational inference [111].

Density estimation is particularly interesting since it offers a way of learning the distribution of data as long as a generative model is available. Given a set of samples $\{x_1, ..., x_n\}$, one could resort to fitting the distribution if their probability density function $f_X$ is unknown. However, even in this case, one still needs to calculate the integral of the distribution to normalize $f_X$ properly. Such an integration is often a highly complex or impossible task to be carried out in the original space, but it becomes tractable using Normalizing Flows.

To achieve this, consider a random variable $Y$ with a known and tractable probability density function $f_Y$ and consider an invertible and differentiable bijection $T$ such that $y = T(x)$. Using the change of variable formula, $f_X(x)$ can be expressed as:

$$f_X(x) = f_Y(T(x)) \cdot |\det(J_T(x))| \tag{D.1}$$

where $J_T$ denotes the Jacobian of the transformation $T$.

Given a set of samples from the complex target distribution $f_X$ it is possible to train the Normalizing Flow and learn the transformation $T$, at which point the calculation of $f_X(x)$ becomes trivial given the expression above.

The limitations on the efficiency of Normalizing Flows in the applications described above come down to the efficacy and flexibility of the transformation $T$ used to transform the actual data. Research into Normalizing Flows is progressing at an accelerated pace, and many approaches are proposed in the literature to tackle this problem. For a detailed review of recent methods, see [112].

The method we decided to implement in our work is proposed by Papamakarios et al. [56]. In their method, they combine *coupling* transforms and special piece-wise defined Spline functions to achieve an element-wise data transformation. In the coupling transforms, the transformation of a subset of components of a sample $x_i$ depends on the value of another subset of components of $x_i$, thus embedding the inherent correlations among dimensions in the data transformation. Additionally, the transformations are defined as piece-wise functions, specifically monotonous rational quadratic splines, which are bijective and invertible. Such a definition of the transformation is ideally suited to be implemented as a Neural Network.

In our group, collaborating with Prof. Dr. A. Caldwell, Dr. O. Schulz, Dr. V. Hafych and M. Dudkowiak, we have produced a Julia implementation of a Spline Normalizing Flow. We are currently investigating their use to transform complex or multimodal probability distribution functions, using space-partitioning [113], and including the Normalizing Flow in an MCMC as a sampler, in order to train the NF and perform Bayesian inference at the same time. Once trained, the NF would contain the full description of the posterior distribution and could be made available or shared with other collaborators in order to include it as a prior in future analysis. More details on our implementation and use of Normalizing Flows for Bayesian inference will be available in M. Dudkowiak's Bachelor thesis.

# Bibliography


[1] S. Yellin. "Finding an upper limit in the presence of unknown background". In: *Phys. Rev. D* 66 (2002), p. 032005.

[2] K. Pearson. "Note on Francis Galton's problem". In: *Biometrika* 1.4 (1933), pp. 390–399.

[3] K. Pearson. "On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random". In: *Biometrika* 25.3/4 (1933), pp. 379–410.

[4] H. A. David and H. N. Nagaraja. *Order statistics*. Wiley, 2003.

[5] N. B. B. C. Arnold and H. N. Nagaraja. *A First Course in Order Statistics*. Society for Industrial and Applied Mathematics, 2008.

[6] R. Pyke. "Spacings". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 27.3 (Sept. 1965), pp. 395–436.

[7] P. V. Sukhatme. *On the analysis of k samples from exponential populations with especial reference to the problem of random intervals*. London University. University College, Department of Statistics, 1936.

[8] A. Rényi. "On the theory of order statistics". In: *Acta Mathematica Academiae Scientiarum Hungarica* 4.3 (Sept. 1953), pp. 191–231.

[9] R. C. H. Cheng. *Handbook of Simulation*. Ed. by J. Banks. John Wiley & Sons, Inc., Aug. 1998.

[10] M. Betancourt. "Cruising the simplex: Hamiltonian Monte Carlo and the Dirichlet distribution". In: *AIP Conference Proceedings*. AIP, 2012.

[11] S. Hassani. *Mathematical Physics*. Springer International Publishing, 2013.

[12] *Discrete Multivariate Analysis Theory and Practice*. Springer New York, 2007.

[13] L. H. C. Tippett. "On the Extreme Individuals and the Range of Samples Taken from a Normal Population". In: *Biometrika* 17.3/4 (1925), pp. 364–387.

[14] R. A. Fisher and L. H. C. Tippett. "Limiting forms of the frequency distribution of the largest or smallest member of a sample". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.2 (Apr. 1928), pp. 180–190.

[15] B. Gnedenko. "Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire". English. In: *Annals of Mathematics*. Second Series 44.3 (1943), pp. 423–453.

[16] M. Siddiqui. "Distribution of quantiles in samples from a bivariate population". In: *Journal of Research of the National Bureau of Standards Section B Mathematics and Mathematical Physics* 64B.3 (July 1960), p. 145.

[17] D. A. Darling. "On a Class of Problems Related to the Random Division of an Interval". In: *The Annals of Mathematical Statistics* 24.2 (1953), pp. 239–253.

[18] L. L. Cam. "The Publications and Writings of Lucien Le Cam". In: *Publications de l'Institut de statistique de l'Université de Paris* 7 (1958), pp. 7–16.

[19] K. A. L. "Sulla determinazione empirica di una legge di distribuzione". In: *G. Ist. Ital. Attuari* 4 (1933), pp. 83–91.

[20] N. Smirnov. "Table for Estimating the Goodness of Fit of Empirical Distributions". In: *Annals of Mathematical Statistics* 19 (1948), pp. 279–281.

[21] H. Cramér. "On the composition of elementary errors". In: *Scandinavian Actuarial Journal* 1928.1 (Jan. 1928), pp. 13–74.

[22] T. W. Anderson and D. A. Darling. "A Test of Goodness of Fit". In: *Journal of the American Statistical Association* 50.268 (1954), pp. 765–769.

[23] Y. Marhuenda, D. Morales, and M. C. Pardo. "A comparison of uniformity tests". In: *Statistics* 39.4 (2005), pp. 315–327.

[24] Z. W. Birnbaum. "Distribution-free Tests of fit for Continuous Distribution Functions". In: *The Annals of Mathematical Statistics* 24.1 (1953), pp. 1–8.

[25] P. Eller and L. Shtembari. "A goodness-of-fit test based on a recursive product of spacings". In: *Journal of Instrumentation* 18.03 (Mar. 2023), P03048.

[26] F. J. Massey. "A Note on the Estimation of a Distribution Function by Confidence Limits". In: *The Annals of Mathematical Statistics* 21.1 (Mar. 1950), pp. 116–119.

[27] F. J. Massey. "The Kolmogorov-Smirnov Test for Goodness of Fit". In: *Journal of the American Statistical Association* 46.253 (Mar. 1951), pp. 68–78.

[28] Z. W. Birnbaum. "Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size". In: *Journal of the American Statistical Association* 47.259 (Sept. 1952), pp. 425–441.

[29] P. A. W. Lewis. "Distribution of the Anderson-Darling Statistic". In: *The Annals of Mathematical Statistics* 32.4 (Dec. 1961), pp. 1118–1124.

[30] R. Pyke. "The Supremum and Infimum of the Poisson Process". In: *The Annals of Mathematical Statistics* 30.2 (1959), pp. 568–576.

[31] J. Durbin. "Tests for Serial Correlation in Regression Analysis Based on the Periodogram of Least-Squares Residuals". In: *Biometrika* 56.1 (1969), pp. 1–15.

[32] H. D. Brunk. "On the Range of the Difference between Hypothetical Distribution Function and Pyke's Modified Empirical Distribution Function". In: *The Annals of Mathematical Statistics* 33.2 (1962), pp. 525–532.

[33] M. Greenwood. "The Statistical Study of Infectious Diseases". In: *Journal of the Royal Statistical Society* 109.2 (1946), pp. 85–110.

[34] P. A. P. Moran. "The Random Division of an Interval". In: *Supplement to the Journal of the Royal Statistical Society* 9.1 (1947), p. 92.

[35] P. A. P. Moran. "The Random Division of an Interval–Part II". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 13.1 (1951), pp. 147–150.

[36] P. A. P. Moran. "The Random Division of an Interval-Part III". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 15.1 (Jan. 1953), pp. 77–80.

[37] B. F. Kimball. "On the Asymptotic Distribution of the Sum of Powers of Unit Frequency Differences". In: *The Annals of Mathematical Statistics* 21.2 (1950), pp. 263–271.

[38] B. F. Kimball. "Some Basic Theorems for Developing Tests of Fit for The Case of the Non-Parametric Probability Distribution Function, I". In: *The Annals of Mathematical Statistics* 18.4 (1947), pp. 540–548.

[39] B. Sherman. "A Random Variable Related to the Spacing of Sample Values". In: *The Annals of Mathematical Statistics* 21.3 (Sept. 1950), pp. 339–361.

[40] D. A. Darling. "On a the Test for Homogeneity and Extreme Values". In: *The Annals of Mathematical Statistics* 23.3 (1952), pp. 450–456.

[41] P. Lévy. "Sur la division d'un segment par des points choisis au hasard". In: *CR Acad. Sci. Paris* 208 (1939), pp. 147–149.

[42] J. G. Mauldon. "Random division of an interval". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 47.2 (1951), pp. 331–336.

[43] N. Cressie. "On the Logarithms of High-Order Spacings". In: *Biometrika* 63.2 (1976), pp. 343–355.

[44] N. Cressie. "An Optimal Statistic Based on Higher Order Gaps". In: *Biometrika* 66.3 (1979), pp. 619–627.

[45] F. Beaujean and A. Caldwell. "A test statistic for weighted runs". In: *Journal of Statistical Planning and Inference* 141.11 (2011), pp. 3437–3446.

[46] F. Beaujean, A. Caldwell, and O. Reimann. "Is the bump significant? An axion-search example". In: *The European Physical Journal C* 78.9 (Sept. 2018), p. 793.

[47] L. Shtembari. *SpacingStatistics.jl*. `github.com/bat/SpacingStatistics.jl/tree/dev`.

[48] M. Bravin et al. "The CRESST dark matter search". In: *Astropart. Phys.* 12 (1999), pp. 107–114.

[49] D. Abrams et al. "Exclusion Limits on the WIMP Nucleon Cross-Section from the Cryogenic Dark Matter Search". In: *Phys. Rev. D* 66 (2002), p. 122003.

[50] L. Shtembari and A. Caldwell. *Limit setting using spacings in the presence of unknown backgrounds*. 2023. arXiv: `2303.09520 [physics.data-an]`.

[51] R. A. Fisher. "Tests of Significance in Harmonic Analysis". In: *Proceedings of the Royal Society of London. Series A* 125.796 (1929), pp. 54–59.

[52] S. Yellin. "Extending the optimum interval method". In: *arXiv* (Feb. 2008). arXiv: 0709.2701 [physics.data-an].

[53] F. N. Fritsch and R. E. Carlson. "Monotone Piecewise Cubic Interpolation". In: *SIAM Journal on Numerical Analysis* 17.2 (1980), pp. 238–246.

[54] L. Shtembari and A. Caldwell. *On goodness-of-fit tests for arbitrary multivariate models*. 2023. arXiv: 2211.03478 [stat.ME].

[55] I. Kobyzev, S. J. Prince, and M. A. Brubaker. "Normalizing Flows: An Introduction and Review of Current Methods". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11 (2021), pp. 3964–3979.

[56] C. Durkan, A. Bekasov, I. M. 0001, and G. Papamakarios. "Neural Spline Flows". In: *Advances in Neural Information Processing Systems* 32 (2019). Ed. by H. M. Wallach, H. Larochelle, A. Beygelzimer, et al., pp. 7509–7520.

[57] M. Springer. *The Algebra of Random Variables*. Wiley, 1979.

[58] P. Eller, N. F. Iachellini, L. Pattavina, and L. Shtembari. "Online triggers for supernova and pre-supernova neutrino detection with cryogenic detectors". In: *Journal of Cosmology and Astroparticle Physics* 2022.10 (Oct. 2022), p. 024.

[59] S. S. Wilks. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". In: *The Annals of Mathematical Statistics* 9.1 (Mar. 1938), pp. 60–62.

[60] N. Ferreiro Iachellini. "Increasing the sensitivity to low mass dark matter in CRESST-III with a new DAQ and signal processing". PhD thesis. Munich U., 2019.

[61] A. H. Abdelhameed, G. Angloher, P. Bauer, et al. "First results from the CRESST-III low-mass dark matter program". In: *Phys. Rev. D* 100 (10 Nov. 2019), p. 102002.

[62] L. Pattavina, N. F. Iachellini, L. Pagnanini, et al. "RES-NOVA sensitivity to core-collapse and failed core-collapse supernova neutrinos". In: *Journal of Cosmology and Astroparticle Physics* 2021.10 (Oct. 2021), p. 064.

[63] W. Baade and F. Zwicky. "On Super-Novae". In: *Proceedings of the National Academy of Sciences* 20.5 (1934), pp. 254–259. eprint: https://www.pnas.org/content/20/5/254.full.pdf.

[64] A. Burrows. "Colloquium: Perspectives on core-collapse supernova theory". In: *Rev. Mod. Phys.* 85 (2013), p. 245. arXiv: 1210.4921 [astro-ph.SR].

[65] H.-T. Janka, T. Melson, and A. Summa. "Physics of Core-Collapse Supernovae in Three Dimensions: a Sneak Preview". In: *Ann. Rev. Nucl. Part. Sci.* 66 (2016), pp. 341–375. arXiv: 1602.05576 [astro-ph.SR].

[66] A. Mirizzi, I. Tamborra, H.-T. Janka, et al. "Supernova Neutrinos: Production, Oscillations and Detection". In: *Riv. Nuovo Cim.* 39.1-2 (2016), pp. 1–112. arXiv: 1508.00785 [astro-ph.HE].

[67] D. Vartanyan and A. Burrows. "Gravitational Waves from Neutrino Emission Asymmetries in Core-collapse Supernovae". In: *The Astrophysical Journal* 901.2 (Sept. 2020), p. 108.

[68] K. Nakamura, S. Horiuchi, M. Tanaka, et al. "Multimessenger signals of long-term core-collapse supernova simulations: synergetic observation strategies". In: *Monthly Notices of the Royal Astronomical Society* 461.3 (June 2016), pp. 3296–3313. eprint: https://academic.oup.com/mnras/article-pdf/461/3/3296/13773820/stw1453.pdf.

[69] Y. Suwa, K. Sumiyoshi, K. Nakazato, et al. "Observing Supernova Neutrino Light Curves with Super-Kamiokande: Expected Event Number over 10 s". In: *The Astrophysical Journal* 881.2 (Aug. 2019), p. 139.

[70] R. Abbasi et al. "IceCube Sensitivity for Low-Energy Neutrinos from Nearby Supernovae". In: *Astron. Astrophys.* 535 (2011). [Erratum: Astron. Astrophys.563,C1(2014)], A109. arXiv: 1108.0171 [astro-ph.HE].

[71] L. Cadonati, F. P. Calaprice, and M. C. Chen. "Supernova neutrino detection in borexino". In: *Astropart. Phys.* 16 (2002), pp. 361–372. arXiv: hep-ph/0012082 [hep-ph].

[72] K. Asakura et al. "KamLAND Sensitivity to Neutrinos from Pre-Supernova Stars". In: *Astrophys. J.* 818.1 (2016), p. 91. arXiv: 1506.01175 [astro-ph.HE].

[73] F. An et al. "Neutrino physics with JUNO". In: *Journal of Physics G: Nuclear and Particle Physics* 43.3 (Feb. 2016), p. 030401.

[74] J. Migenda. "Supernova Burst Observations with DUNE". In: *Proceedings, Prospects in Neutrino Physics (NuPhys2017): London, UK, December 20-22, 2017*. 2018, pp. 164–168. arXiv: 1804.01877 [physics.ins-det].

[75] D. Akimov et al. "Observation of Coherent Elastic Neutrino-Nucleus Scattering". In: *Science* 357.6356 (2017), pp. 1123–1126. arXiv: 1708.01294 [nucl-ex].

[76] D. Z. Freedman. "Coherent Neutrino Nucleus Scattering as a Probe of the Weak Neutral Current". In: *Phys. Rev. D* 9 (1974), pp. 1389–1392.

[77] D. Z. Freedman, D. N. Schramm, and D. L. Tubbs. "The Weak Neutral Current and Its Effects in Stellar Collapse". In: *Ann. Rev. Nucl. Part. Sci.* 27 (1977), pp. 167–207.

[78] A. Drukier and L. Stodolsky. "Principles and Applications of a Neutral Current Detector for Neutrino Physics and Astronomy". In: *Phys. Rev.* D30 (1984), p. 2295.

[79] R. F. Lang, C. McCabe, S. Reichard, et al. "Supernova neutrino physics with xenon dark matter detectors: A timely perspective". In: *Phys. Rev.* D94.10 (2016), p. 103009. arXiv: 1606.09243 [astro-ph.HE].

[80] D. Khaitan. "Supernova neutrino detection in LZ". In: *JINST* 13.02 (2018), p. C02024. arXiv: 1801.05651.

[81] P. Agnes et al. "Sensitivity of future liquid argon dark matter search experiments to core-collapse supernova neutrinos". In: *JCAP* 03 (2021), p. 043. arXiv: 2011.07819 [astro-ph.HE].

[82] K. Rozwadowska, F. Vissani, and E. Cappellaro. "On the rate of core collapse supernovae in the milky way". In: *New Astron.* 83 (2021), p. 101498. arXiv: 2009.03438 [astro-ph.HE].

[83] J. Schmidt, M. Hohle, and R. Neuhäuser. "Determination of a temporally and spatially resolved supernova rate from OB stars within 5 kpc". In: *Astron. Nachr.* 335 (2014), pp. 935–948. arXiv: 1409.3357 [astro-ph.SR].

[84] L.-S. The, D. D. Clayton, R. Diehl, et al. "Are ti-44 producing supernovae exceptional?" In: *Astron. Astrophys.* 450 (2006), p. 1037. arXiv: astro-ph/0601039.

[85] L. Pattavina, N. Ferreiro Iachellini, and I. Tamborra. "Neutrino observatory based on archaeological lead". In: *Phys. Rev. D* 102.6 (2020), p. 063001. arXiv: 2004.06936 [astro-ph.HE].

[86] P. Belli et al. "Search for $2\beta$ decay of $^{106}$Cd with an enriched $^{106}CdWO_4$ crystal scintillator in coincidence with four HPGe detectors". In: *Phys. Rev. C* 93.4 (2016), p. 045502. arXiv: 1603.06363 [nucl-ex].

[87] P. Belli et al. "Search for Double Beta Decay of $^{106}$Cd with an Enriched 106$CdWO_4$ Crystal Scintillator in Coincidence with $CdWO_4$ Scintillation Counters". In: *Universe* 6.10 (2020), p. 182. arXiv: 2010.08749 [nucl-ex].

[88] C. Alduino et al. "The projected background for the CUORE experiment". In: *Eur. Phys. J.* C77.8 (2017), p. 543. arXiv: 1704.08970 [physics.ins-det].

[89] L. Pattavina, J. W. Beeman, M. Clemenza, et al. "Radiopurity of an archeological Roman Lead cryogenic detector". In: *Eur. Phys. J.* A55 (2019), p. 127. arXiv: 1904.04040 [physics.ins-det].

[90] A. H. Abdelhameed et al. "First results from the CRESST-III low-mass dark matter program". In: *Phys. Rev.* D100.10 (2019), p. 102002. arXiv: 1904.00498 [astro-ph.CO].

[91] G. Angloher et al. "Results on light dark matter particles with a low-threshold CRESST-II detector". In: *Eur. Phys. J. C* 76.1 (2016), p. 25. arXiv: 1509.01515 [astro-ph.CO].

[92] S. Al Kharusi et al. "SNEWS 2.0: a next-generation supernova early warning system for multi-messenger astronomy". In: *New J. Phys.* 23.3 (2021), p. 031201. arXiv: 2011.00035 [astro-ph.HE].

[93] K. Nakamura, S. Horiuchi, M. Tanaka, et al. "Multimessenger signals of long-term core-collapse supernova simulations: synergetic observation strategies". In: *Mon. Not. R. Astron. Soc.* 461.3 (June 2016), pp. 3296–3313.

[94] MPA Supernova Archive, `https://wwwmpa.mpa-garching.mpg.de/ccsnarchive`. `https://wwwmpa.mpa-garching.mpg.de/ccsnarchive/data/`.

[95] N. Y. Agafonova et al. "On-line recognition of supernova neutrino bursts in the LVD detector". In: *Astropart. Phys.* 28 (2008), pp. 516–522. arXiv: `0710.0259 [astro-ph]`.

[96] M. Lamoureux. "Identification of neutrino bursts associated to supernovae with real-time test statistic (RTS2) method". In: *Astron. Astrophys.* 654 (2021), A95. arXiv: `2103.09733 [astro-ph.HE]`.

[97] O. I. Gonzalez-Reina et al. *EstrellaNueva: an open-source software to study the interactions and detection of neutrinos emitted by supernovae*. 2022.

[98] L. Shtembari and A. Caldwell. *On the sum of ordered spacings*. 2020. arXiv: `2008.02048 [math.ST]`.

[99] A. Caldwell, P. Eller, V. Hafych, et al. "Integration with an adaptive harmonic mean algorithm". In: *International Journal of Modern Physics A* 35.24 (Aug. 2020), p. 1950142.

[100] Q. F. Gronau, A. Sarafoglou, D. Matzke, et al. "A tutorial on bridge sampling". In: *Journal of Mathematical Psychology* 81 (2017), pp. 80–97.

[101] B. Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *Annals Statist.* 7.1 (1979), pp. 1–26.

[102] P. Dierckx. "An algorithm for smoothing, differentiation and integration of experimental data using spline functions". In: *Journal of Computational and Applied Mathematics* 1.3 (1975), pp. 165–184.

[103] P. Dierckx. "A Fast Algorithm for Smoothing Data on a Rectangular Grid while Using Spline Functions". In: *SIAM Journal on Numerical Analysis* 19.6 (1982), pp. 1286–1304. eprint: `https://doi.org/10.1137/0719093`.

[104] P. Dierckx. "Curve and surface fitting with splines". In: *Monographs on numerical analysis*. 1996.

[105] F. N. Fritsch and J. Butland. "A Method for Constructing Local Monotone Piecewise Cubic Interpolants". In: *SIAM Journal on Scientific and Statistical Computing* 5.2 (1984), pp. 300–304. eprint: `https://doi.org/10.1137/0905021`.

[106] E. G. Tabak and E. Vanden-Eijnden. "Density estimation by dual ascent of the log-likelihood". In: *Communications in Mathematical Sciences* 8.1 (2010), pp. 217–233.

[107] J. P. Agnelli, M. Cadeiras, E. G. Tabak, et al. "Clustering and Classification through Normalizing Flows in Feature Space". In: *Multiscale Modeling & Simulation* 8.5 (2010), pp. 1784–1802. eprint: `https://doi.org/10.1137/100783522`.

[108] P. Laurence, R. Pignol, and E. Tabak. "Constrained density estimation". In: *Proceedings of the 2011 Wolfgang Pauli Institute Conference on Energy and Commodity Trading*. Springer-Verlag, 2014, pp. 259–284.

[109] O. Rippel and R. P. Adams. *High-Dimensional Probability Estimation with Deep Density Models*. 2013. arXiv: `1302.5125 [stat.ML]`.

[110] L. Dinh, D. Krueger, and Y. Bengio. *NICE: Non-linear Independent Components Estimation*. 2015. arXiv: `1410.8516 [cs.LG]`.

[111] D. J. Rezende and S. Mohamed. *Variational Inference with Normalizing Flows*. 2016. arXiv: `1505.05770 [stat.ML]`.

[112] I. Kobyzev, S. J. Prince, and M. A. Brubaker. "Normalizing Flows: An Introduction and Review of Current Methods". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11 (Nov. 2021), pp. 3964–3979.

[113] V. Hafych, P. Eller, O. Schulz, and A. Caldwel. "Parallelizing MCMC sampling via space partitioning". In: *Statistics and Computing* 32.4 (June 2022).