Distribution-free detection of structured anomalies: permutation and rank-based scans

Citation for published version (APA):

Document status and date:
Published: 01/01/2015

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Distribution-Free Detection of Structured Anomalies: Permutation and Rank-Based Scans

Ery Arias-Castro$^1$, Rui M. Castro$^2$, Ervin Tánzos$^2$, and Meng Wang$^1$

$^1$University of California, San Diego  
$^2$Technische Universiteit Eindhoven

Abstract

The scan statistic is by far the most popular method for anomaly detection, being popular in syndromic surveillance, signal and image processing, and target detection based on sensor networks, among other applications. The use of the scan statistics in such settings yields an hypotheses testing procedure, where the null hypothesis corresponds to the absence of anomalous behavior. If the null distribution is known, then calibration of a scan-based test is relatively easy, as it can be done by Monte-Carlo simulation. When the null distribution is unknown, it is not clear what the best way to proceed is. We propose two procedures. One is a calibration by permutation and the other is a rank-based scan test, which is distribution-free and less sensitive to outliers. Furthermore, the rank-scan test requires only a one-time calibration for a given data size making it computationally more appealing. In both cases, we quantify the performance loss with respect to an oracle scan test that knows the null distribution, and show one incurs only a very small loss in the context of a natural exponential family. These results include the classical normal location model, as well as Poisson model popular in syndromic surveillance. We perform numerical experiments on simulated data further supporting our theory, and also experiments with a real dataset from genomics.

1 Introduction

Signal detection (and localization) is important in a large variety of applications, encompassing any situation where the goal is to discover patterns or detect/locate anomalies. When these problems are casted in a statistical way this leads to multiple testing situations. We focus here on settings where one desires to detect or locate anomalous behavior endowed with some structure. A prototypical example is that of detecting a time interval with an unusually high concentration of certain events being monitored. This setting might have been the motivation for the work on scan statistics (Wallenstein, 2009). To move the setting closer to the one we consider, assume we bin the data, so that we observe a sequence of Poisson-distributed random variables — this assumes that the original point-process was a Poisson process. The scan statistic in this particular case combines sums of these values over (discrete) intervals of different sizes and location, with possible normalization — see (1) further down. In this scenario we want to perform a hypotheses test, where the null hypothesis is that no anomaly is present (so we are in the presence of a homogenous Poisson process) versus an inhomogenous process where some time intervals have an elevated rate of events. If the null distribution is known then calibrating a test based on a scan statistic is relatively easy, and can done either analytically (in some specific cases) or by Monte-Carlo simulation. In this example this amounts to having knowledge of the density of the Poisson process when no anomaly is present.
More generally the scan tests can be used in a variety of applications where intervals/regions of unusually high amplitude/actively are sought (Balakrishnan and Koutras, 2002). For instance in (Cheung et al., 2013) the scan statistics was used to detect small geographic areas with large suicide rates. Guerriero et al. (2009) used the scan-statistics for target detection using distributed sensors in a two dimensional region.

As stated above, if the null distribution is known it is relatively easy to properly design and calibrate a scan test. In many applications, however, the null distribution is unknown and this is the situation that we study in this paper. Seeing the scan of a certain interval as comparing the values in that interval with those outside, we are confronted with a two-sample problem for each interval, which is then followed by some form of multiple testing, since we scan many intervals. Thus drawing from the classical literature on the two-sample problem, we propose two approaches:

- **Calibration by permutation.** This amounts to using the permutation distribution of the scan statistic for inference (detection/estimation).

- **Scanning the ranks.** This amounts to replacing each observation with its rank before scanning. As any rank-based method, it leads to a calibration by permutation, but the advantage is that the calibration must be done only once for each sample size.

Both approaches are quite natural. For example, a calibration by permutation is suggested in (Kulldorff et al., 2005) in the context of syndromic surveillance. See also (Flenner and Hewer, 2011) for a recent example in the context of detecting a change in a sequence of images. Although quite natural to a statistician, the method based on ranks appears to be new, to the best knowledge of the authors.

In this paper, we analyze both methods in a standard mathematical framework and evaluate their finite-sample performance in some numerical experiments. We show that even when the null distribution is unknown the performance of the tests based on those two approaches is very close to that of the oracle test (with clairvoyant knowledge of the distribution) in the context of a natural exponential family. This includes the classical normal location model and the Poisson example above, among many other scenarios. We perform numerical experiments on simulated data, confirming our theory, and also some experiments using a real dataset from genomics. As specified below, we focus on a “static” setting, where the length of the signal being monitored is fixed a priori. Adding time is typically done by adding one ‘dimension’ to the framework, as done for example in (Kulldorff et al., 2005).

### 1.1 General setting

A typical framework for static anomaly detection — which includes detection in digital signals and images, sensor networks, biological data, and more — may be described in general terms as follows. We observe a set of random variables, denoted \((X_v : v \in \mathcal{V})\), where \(\mathcal{V}\) is a finite index set of size \(N\). This is a snapshot of the state of the environment, where each element of \(\mathcal{V}\) corresponds to an element of the environment (e.g., these correspond to nodes of a network, pixel, genes, etc...). In this work we take an hypothesis testing point of view. Under the null hypothesis, corresponding to the nominal state when no anomalies are present, these random variables are Independent and Identically Distributed (IID) with some null distribution \(F_0\). Alternatively some of these random variables will have a different distribution. Let \(\mathcal{S} \subseteq 2^{\mathcal{V}}\) denote a class of possibly anomalous subsets, corresponding to the anomalous patterns we expect to encounter. Under the alternative hypothesis there is a subset \(S \in \mathcal{S}\) such that, for each \(v \in S\), \(X_v \sim F_v\), for some distribution \(F_v \neq F_0\), and independent of \((X_v : v \in \mathcal{V} \setminus S)\), which are still IID with distribution \(F_0\). In a number of
important applications, the variables are real-valued and the anomalous variables take larger-than-usual values, which can be formalized by assuming that each \( F_\nu \) stochastically dominates\(^1 \) \( F_0 \). We assume this is the case throughout the paper.

### 1.2 Exponential models

An important case is that of a one-parameter exponential model. Without loss of generality we assume the natural parameterization, since we can otherwise pre-process the data to ensure this. Consider a probability measure \( \nu \) on the real line with finite second moment. We assume that either \( \nu \) is continuous (i.e., diffuse) or discrete (i.e., with discrete support). For \( \theta \in (0, \theta_*), \) define \( f_\theta(x) = \exp(\theta x - \log \varphi_0(\theta)) \), where \( \varphi_0(\theta) = \int e^{\theta x} \nu(dx) \) and \( \theta_* = \sup\{ \theta > 0 : \varphi_0(\theta) < \infty \} \), assumed to be strictly positive (and possibly infinite). Note that in this model \( F_0 \) coincides with \( \nu \) (and in particular \( f_0 \equiv 1 \)). In this setting, we replace \( F_\nu \) with \( F_{\theta_*} \), where \( \theta_* > 0 \). Since a natural exponential family as the monotone likelihood ratio property\(^2 \), it follows that \( (F_\theta) \) is stochastically increasing in \( \theta \) (Lehmann and Romano, 2005, Lem 3.4.2). In particular, we do have \( F_\theta \geq F_0 \) for all \( \theta > 0 \). Important special cases of such an exponential model includes the normal location model — with \( F_\theta \) corresponding to \( \mathcal{N}(\theta, 1) \) — standard in many signal and image processing applications; and the Poisson model — with \( F_\theta \) corresponding to Pois(\( \theta \)) — popular in syndromic surveillance (Kulldorff et al., 2005).

Note that in the formulation above the alternative hypothesis is composite. Tackling this problem using a generalized likelihood ratio approach is popular in practice (Kulldorff, 1997) and often referred to as the scan test, as it works by scanning over the possible anomalous sets to determine if there is such a set that is able to “explain” the observed data. Assuming the nonzero \( \theta_* \)'s are all equal to \( \theta \) under the alternative, some simplifications lead to considering the test that rejects for large values of

\[
\max_{S \in \mathcal{S}} \frac{1}{\sqrt{|S|}} \sum_{v \in S} (X_v - \mathbb{E}_\theta(X_v)) \quad (1)
\]

(\( \mathbb{E}_\theta \) denotes the expectation with respect to \( F_\theta \). And for a discrete set \( S, |S| \) denotes its cardinality.) As argued in (Arias-Castro and Grimmett, 2013), this test is in a certain sense asymptotically equivalent to the generalized likelihood ratio test, but simpler.

### 1.3 Structured anomalies

Naturally, the intrinsic difficulty of the detection task depends not only on the data distribution, but also on the complexity of the class of anomalous sets. In several cases one can make some structural assumptions on such classes. For instance, grid-like networks are an important special case, arising in applications such as signal and image processing (where the signals are typically regularly sampled) and sensor networks deployed for the monitoring of some geographical area, for example. This situation is considered in great generality and from different perspectives in (Arias-Castro et al., 2011, 2005; Cai and Yuan, 2014; Desolneux et al., 2003; Hall and Jin, 2010; Perone Pacifico et al., 2004; Walther, 2010). Also, the distribution of the corresponding scan statistic (1) and variants has been studied in a number of places (Boutsikas and Koutras, 2006; Jiang, 2002; Kabluchko, 2011; Sharpnack and Arias-Castro, 2014; Siegmund and Venkatraman, 1995). We first detail the simplest and most emblematic setting, that of detecting an interval in a

---

1. For two distribution (functions) on the real line, \( F \) and \( G \), we say that \( G \) stochastically dominates by \( F \) if \( G(t) \leq F(t) \) for all \( t \in \mathbb{R} \). We denote this by \( G \geq F \).

2. A family of densities \( \{f_\theta : \theta \in \Theta \} \), where \( \Theta \subset \mathbb{R} \), has the monotone likelihood ratio property if \( f_{\theta'}(x)/f_\theta(x) \) is increasing in \( x \) when \( \theta' > \theta \).
one-dimensional regularly sampled signal, and then explain how this is generalized to more complex settings.

1.3.1 Detection of intervals

Let \( V = \{1, \ldots, N\} \) represent a one-dimensional lattice and \( S \) be the class of all discrete intervals of \( V \), meaning \( S = \{a, \ldots, b : 1 \leq a \leq b \leq N\} \). If one assumes a normal location model the scan test rejects the null for large values of

\[
\max_{1 \leq a \leq b \leq N} \frac{1}{\sqrt{b-a+1}} \sum_{v=a}^{b} X_v .
\]

Following the parameterization in (Arias-Castro et al., 2005) one assumes \( \theta_v = \mu / \sqrt{|S|} \) for all \( v \in S \), where \( S \) denotes the anomalous set (this ensures that all intervals are roughly equally difficult to detect). In a minimax setting, it can be shown that the detection boundary is at \( \mu = \sqrt{2 \log N} \), meaning that, for any \( \varepsilon > 0 \) fixed, no test can simultaneously attain an arbitrarily small type I and type II error probabilities in the large sample limit \( N \to \infty \) when \( \mu \leq (1-\varepsilon) \sqrt{2 \log N} \), while there is such a test when \( \mu \leq (1+\varepsilon) \sqrt{2 \log N} \) (meaning such a test has risk tending to 0). In fact, the scan test (2) is one of them. It is important to notice that in this model the short intervals (the most numerous in \( S \)) drive the difficulty of the problem and a refinement is possible. See (Arias-Castro et al., 2011; Walther, 2010) for details.

1.3.2 Detection of other objects

The principles underlying the detection of intervals can be used for the detection of much more general anomaly classes. As shown in (Arias-Castro et al., 2011), similar results apply to a general (nonparametric) class \( S \) of blob-like (‘thick’) sets \( S \) when \( V \) is a grid-like set of arbitrary finite dimension, although the scanning is done over an appropriate approximating net for \( S \) (instead of the entire class \( S \)). Furthermore, these results generalize to one-parameter exponential models, beyond the commonly assumed normal location model, as long as the sets \( S \in S \) are sufficiently large (poly-logarithmic in \( N \)). Other papers that develop theory for other environments include (Addario-Berry et al., 2010; Arias-Castro et al., 2008; Sharpnack and Singh, 2010; Sharpnack et al., 2013; Zhao and Saligrama, 2009). Other variants of this detection problem have been suggested. Also, the applied literature is quite extensive. We refer the reader to (Arias-Castro et al., 2011) and references therein.

In this paper we focus exclusively on the detection of intervals, for the sake of clarity and simplicity. We also study a scan test restricted to an approximating net, because of the following advantages: the analysis is simpler as it does not require chaining to attain tight results; it is applicable in more general settings, in particular, when the class \( S \) is nonparametric; and it is computationally faster to implement.

1.4 Calibration by permutation

Suppose we are considering a test that rejects the null for large values of a test statistic \( T(X) \) where \( X = (X_v, v \in V) \). Let \( x = (x_v, v \in V) \) the observed value of \( X \). If we were to know the null distribution \( F_0 \), we would return the P-value as \( P_0(T(X) \geq T(x)) \). In practice, more than knowing the null distribution, what really matters is that we can (efficiently) simulate from it, so that we can estimate this P-value by Monte-Carlo simulation. Ignoring computational constraints for the moment, a calibration by permutation amounts to computing \( T(x_\pi) \) for all \( \pi \in \mathcal{V}! \), where
\( \mathcal{V}! \) denotes the set of all permutations of \( \mathcal{V} \) and \( x_\pi = (x_{\pi(v)}, v \in \mathcal{V}) \) is the permuted data. We then return the P-value \( |\{ \pi \in \mathcal{V}! : T(x_\pi) \geq T(x)\}|/|\mathcal{V}!| \) and the rejection decision is based on this value. Let \( M = |\{T(x_\pi) : \pi \in \mathcal{V}!\}|. \) If there are no multiplicities, meaning \( M = \mathcal{V}! \), it can be shown such tests are exact and that under the null the P-value has a (discrete) uniform distribution on \( \{1/M, 2/M, \ldots, 1\} \). Otherwise the test will be conservative (Lehmann and Romano, 2005).

In practice, the number of permutations is very large (as \( |\mathcal{V}!| = |\mathcal{V}|! \)) and the P-value is estimated by simulation (by uniform sampling of permutations). Our first contribution is the following:

**Contribution 1:** We establish the performance of a scan test calibrated by permutation.

We are only aware of one other paper that does that (Walther, 2010). In that paper, the data are binary, making the analysis somewhat simpler, while we consider exponential models as described in Section 1.2. We carry our analysis in detail for the problem of detecting an interval of unknown length, and explain how to generalize this to more complex settings, such as considered in (Arias-Castro et al., 2011).

### 1.5 Scanning the ranks

When calibrating by permutation, the scan is performed on the permuted data. Even though this is done for only a relatively small number of permutations, that number is often chosen in the hundreds, if not thousands, meaning that the procedure requires the computation of that many scans. This can be rather time consuming. Furthermore, when new data becomes available, the whole procedure is undertaken anew. The computational burden of doing so may be prohibitive in some practical situations.

We propose a variant which is based on replacing the observations with their rank (in increasing order) and scanning the ranks. The resulting procedure is distribution-free and therefore only needs to be calibrated once as long as the data remain of same size, meaning that \( \mathcal{V} \) does not change. Such a procedure is very natural given the classical literature on nonparametric tests (Hettmansperger, 1984). It is in fact directly inspired by the Wilcoxon rank-sum test (Wilcoxon, 1945). As any other rank-based method, scanning ranks comes with two additional advantages: its calibration by permutation remains valid as long as the variables are exchangeable under the null; and the method is robust to outliers in the form of scattered observations with very large values.

In detail, let \( R_v \) denote the rank (in increasing order) of \( X_v \) among \( X \). If there are ties, they can be dealt with in any of the classical ways, for instance, given the average rank. For technical reasons however, our results are proven under the assumption that ties are broken randomly. If \( T(X) \) is a form of scan statistic, we then consider the rank scan, defined as \( T(R) \), where \( R = (R_v, v \in \mathcal{V}) \). For example, the rank variant of (1) is

\[
\max_{S \subseteq \mathcal{S}} \frac{1}{\sqrt{|S|}} \sum_{v \in S} (R_v - \frac{N+1}{2}),
\]

since \( E_0(R_v) = \frac{N+1}{2} \).

**Contribution 2:** We establish the performance of a rank scan test.

The test rejects the null hypothesis if the above test statistic is large. Since the rank’s null distribution is precisely the uniform distribution over the permutations of \( \mathcal{V} \), this test is equivalent to the permutation test when the original data is replaced by the ranks.
1.6 Content and notation

The rest of the paper is organized as follows. In Section 2 we consider the case when the null distribution is known. This section is expository, introducing the reader to the basic proof techniques that are used, for example in (Arias-Castro et al., 2011), to establish the performance of the scan statistic when calibrated with full knowledge of the null distribution. To keep the exposition simple, and to avoid repeating the substantially more complex arguments detailed in that paper and others, we focus on the problem of detecting an interval in a one-dimensional lattice. This allows us to set the foundation and discover what the performance bounds for the scan test in this case rely on. In Section 3 we consider the same setting and instead calibrate the scan statistic by permutation. In Section 4 we consider the same setting and instead scan the ranks. In both cases, our analysis relies on concentration inequalities for sums of random variables obtained from sampling without replacement from a finite set of reals, already established in the seminal paper of Hoeffding (1963). In the same section, we also consider the problem of detecting very small anomalous sets by scanning ranks. In Section 5 we perform some simulations to numerically quantify how much is lost in finite samples when calibrating by permutation or when using ranks. We also compare our methodology with the method of Cai et al. (2012), on simulated data, and also on a real dataset from genomics. We conclude the paper with a brief discussion in Section 6.

2 When the null distribution is known

This section is meant to introduce the reader to the techniques underlying the performance bounds developed in (Arias-Castro et al., 2011, 2005) for the scan statistic (and variants) when the null distribution is known. These provide a stepping stone for our results in regards to permutation and rank scan tests. We detail the setting of detecting an interval of unknown length in a one-dimensional lattice. Therefore, as in Section 1.3, consider the setting where

\[ V = \{1, \ldots, N\} \text{ and } S = \{\{a, \ldots, b\} : 1 \leq a \leq b \leq N\} . \]

We begin by considering the normal model — \( X_v \sim \mathcal{N}(\theta_v, 1) \) are independent — and explain later on how to generalize the arguments to an arbitrary exponential model as described in Section 1.2. For each \( S \in S \), take \( \theta_S > 0 \). We are interested in testing

\[
H_0 : \theta_v = 0, \forall v \in V \text{ versus } H_1 : \exists S \in S \text{ such that } \theta_v \geq \theta_S, \forall v \in S .
\]

We consider this problem from a minimax perspective. It is shown in (Arias-Castro et al., 2005) that if \( \theta_S \leq \sqrt{2(1 - \varepsilon) \log(N)/|S|} \), with \( \varepsilon > 0 \) fixed, then any test with level \( \alpha \) has power at most \( \beta(\alpha, N) \to \alpha \) as \( N \to \infty \). In other words, in the large-sample limit, no test can do better than random guessing, meaning, than the test that rejects with probability \( \alpha \) regardless of the data. On the other hand, if \( \theta_S \geq \sqrt{2(1 + \varepsilon) \log(N)/|S|} \), then for any level \( \alpha > 0 \) there exists a test with level \( \alpha \) with power \( \beta(\alpha, N) \to 1 \) as \( N \to \infty \). In particular, such a test can be constructed using a form a scanning.

Remark 1. As mentioned in Section 1.3, a refinement is possible (Arias-Castro et al., 2011; Walther, 2010), by implementing a form of scale-dependent multiple testing. The resulting procedure achieves the same detection boundary as the scan test knowing the length of the anomalous interval, which is at \( \theta_S = \sqrt{2 \log(N/|S|)/|S|} \). Although we could consider this refinement here as well, we choose not to for the sake of simplicity.
2.1 Scanning over an approximating net

Instead of considering a test that scans over all elements in $S$, as in (2), we describe a variant that consists in scanning an approximating net for the class $S$. This brings both computational and analytical advantages over scanning all sets in $S$ as discussed in Section 1.3. We use the approximating net of (Arias-Castro et al., 2005); see (Sharpnack and Arias-Castro, 2014) for an alternative construction. The underlying metric on $S$ is given by

$$\delta(S, S') := \sqrt{E_0(Y_S - Y_{S'})^2} = \sqrt{2 - 2\rho(S, S')} ,$$

where

$$Y_S := \frac{1}{\sqrt{|S|}} \sum_{v \in S} X_v \quad \text{and} \quad \rho(S, S') := \frac{|S \cap S'|}{\sqrt{|S||S'|}} .$$

STEP 1: Construction of an approximating net Instead of scanning over $S$ we will scan over a subclass of intervals $S_b$, where $0 \leq b \leq N$ is an integer to be specified later on. Such a subclass must satisfy two important properties, namely have cardinality significantly smaller than $S$, and be such that any element $S \in S$ can be well approximated by an element of $S_b$, in terms of the metric $\delta$ defined above. We consider a construction based on dyadic intervals and their extensions by concatenation of at most $b$ dyadic length intervals at each end. Details of the construction are presented in Appendix A. For simplicity we assume $N = 2^q$ for some integer $q$ (this is not restrictive, but facilitates the presentation). The proposed construction has the following properties.

Lemma 1. There is a class $S_b$ of intervals of cardinality at most $2N(\log_2(4N))^{2b}$ with the property that for any element $S \in S$ there is an element $S^* \in S_b$ such that $S^* \subset S$ and

$$\rho(S^*, S) \geq \sqrt{\max(1 - 2^{2-b}, 0)} .$$

STEP 2: Definition of the scan test We consider a test based on scanning only the intervals in $S_b$. This test rejects the null if

$$\max_{S \in S_b} Y_S \geq \sqrt{2(1 + \varepsilon_0) \log N} , \quad (3)$$

where $\varepsilon_0 > 0$ satisfies $\varepsilon_0 \to 0$ and $\varepsilon_0 \log(N)/\log\log(N) \to \infty$ (the reason for the last condition becomes apparent from the analysis in the next step).

STEP 3: Under the null hypothesis By the union bound, we have

$$P_0 \left( \max_{S \in S_b} Y_S \geq \sqrt{2(1 + \varepsilon_0) \log N} \right) \leq \sum_{S \in S_b} P_0 \left( Y_S \geq \sqrt{2(1 + \varepsilon_0) \log N} \right) \leq |S_b| \bar{\Phi} \left( \sqrt{2(1 + \varepsilon_0) \log N} \right) ,$$

where $\Phi$ denotes the standard normal distribution function and $\bar{\Phi} = 1 - \Phi$ denotes the corresponding survival function. We have the well-known bound on Mill’s ratio:

$$\bar{\Phi}(x) \leq e^{-x^2/2}, \quad \forall x \geq 0 . \quad (4)$$

Therefore we get

$$P_0 \left( \max_{S \in S_b} Y_S \geq \sqrt{2(1 + \varepsilon_0) \log N} \right) \leq 2N(\log_2(4N))^{2b} N^{-(1-\varepsilon_0)} .$$
We then see that choosing \( b \ll \varepsilon_0 \log(N)/\log(\log(N)) \) makes the last expression tend to zero as \( N \to \infty \). Since we want \( b \to \infty \) this motivates a choice of \( \varepsilon_0 \) satisfying \( \varepsilon_0 \log(N)/\log(\log(N)) \to \infty \), and taking, for instance \( b = \frac{1}{4} \varepsilon_0 \log(N)/\log(\log(N)) \), we conclude the test in (3) has level tending to 0 as \( N \to \infty \).

**STEP 4: Under the alternative** We now show that the power of this test tends to 1 when \( \theta_v = \sqrt{2(1 + \varepsilon)} \log N/\sqrt{|S|} \) for all \( v \in S \), where \( S \in \mathbb{S} \) is the anomalous interval under the alternative and \( \varepsilon > 0 \) is fixed. Referring to Lemma 1, there is a set \( S^* \in \mathcal{S}_0 \) such that \( \rho(S^*, S) \geq \sqrt{\max(0, 1 - 2^{2-b})} \). Furthermore \( Y_{S^*} \) is normal with mean \( \rho(S^*, S) \sqrt{2(1 + \varepsilon)} \log N \) and variance 1. We thus have

\[
\mathbb{P}(Y_{S^*} \geq \sqrt{2(1 + \varepsilon_0)} \log N) = \Phi(\xi),
\]

where

\[
\xi := \sqrt{2(1 + \varepsilon_0)} \log N - \rho(S^*, S) \sqrt{2(1 + \varepsilon)} \log N
\]

\[
= \sqrt{2(1 + \varepsilon_0)} \log N \left( 1 - \rho(S^*, S) \sqrt{\frac{1 + \varepsilon}{1 + \varepsilon_0}} \right)
\]

\[
\leq \sqrt{2(1 + \varepsilon_0)} \log N \left( 1 - \sqrt{\max(0, 1 - 2^{2-b})} \frac{1 + \varepsilon}{1 + \varepsilon_0} \right) \to -\infty
\]

as \( N \to \infty \), where we used the fact that \( \varepsilon > 0 \) is fixed, \( \varepsilon_0 \to 0 \) and \( b \to \infty \). We conclude that the test in (3) has power tending to 1 as \( N \to \infty \). In conclusion, we have shown the following result.

**Proposition 1.** *(Arias-Castro et al., 2005)* The test defined in (3) has level converging to 0 as \( N \to \infty \), and has power converging to 1 as \( N \to \infty \) provided

\[
\theta_S \geq \sqrt{2(1 + \varepsilon)} \log N/|S| \quad \text{for some } \varepsilon > 0.
\]

(5)

**2.2 Generalizations**

**Other exponential models** To extend the result to an arbitrary (one-parameter, natural) exponential model, we require the equivalent of the tail-bound (4). While such a bound may not apply to a particular exponential model, it does apply asymptotically to large sums of IID variables from that model by Chernoff’s bound and a Taylor development of the rate function.

Indeed, recalling the notation introduced in Section 1.2, let \( \psi_0(t) = \sup_{\lambda \geq 0} (\lambda t - \log \varphi_0(\lambda)) \), which is the rate function of \( F_0 \). By Chernoff’s bound, we have

\[
\mathbb{P}_0(Y_S \geq y) \leq \exp \left( -|S|\psi_0(y)|S|^{-1/2} \right).
\]

(6)

Assuming without loss of generality that \( F_0 \) has zero mean and unit variance, we have

\[
\psi_0(t) \geq \frac{1}{2} t^2 + O(t^3), \quad t \to 0.
\]

(7)

To see this, note that \( \varphi_0(\lambda) \) is infinitely many times differentiable when \( \lambda \geq 0 \), with \( \varphi_0'(0) = \mathbb{E}_0(X) = 0 \) and \( \varphi_0''(0) = \mathbb{E}_0(X^2) = 1 \), by assumption. Therefore, there is a constant \( c > 0 \) such that, when \( \lambda \in [0, 1] \), \( \varphi_0(\lambda) \leq 1 + \frac{1}{2} \lambda^2 - c\lambda^3 \). For \( t \in [0, 1] \), we then have

\[
\psi_0(t) \geq \sup_{\lambda \in [0, 1]} \left[ \lambda t - \log(1 + \frac{1}{2} \lambda^2 - c\lambda^3) \right] \geq t^2 - \log(1 + \frac{1}{2} t^2 - ct^3),
\]
which yields (7). From this we see that our derivations for the normal model apply essentially verbatim if, for some constant \(c > 0\), \(|S| \geq c(\log N)^3\) for all \(S \in \mathcal{S}\). Furthermore, it can be seen the test in (3) is essentially optimal for exponential models, as its performance matches the lower bounds in (Arias-Castro et al., 2011).

**Other detection settings** For a specific detection problem, specified by a set of nodes and a class of subsets \(\mathcal{S} \subset 2^\mathcal{V}\), and an exponential model, the arguments above continue to apply if one is able to construct an appropriate approximating net, as we did for the class of intervals. This is done for instance in (Arias-Castro et al., 2011, 2005) for a wide range of settings.

### 3 Calibration by permutation

Having described in detail how a performance bound is established for the scan test variant (3) for the problem of detecting an interval of unknown length, and its extensions to other detection problems, we now clearly see that the key to adapting this analysis to a calibration by permutation is a concentration of measure bound to replace (4) and (6). Since this is the same in any detection setting, we consider as in Section 2 the problem of detecting an interval of unknown length. This time, we impose a minimum and maximum length on the intervals

\[
\mathcal{S} = \{\{a, \ldots, b\} : 1 \leq a < b \leq N, 2^q_l \leq b - a \leq 2^q_u\}.
\]

Indeed, when calibrating the scan test by permutation, we necessarily have to assume nontrivial upper and lower bounds on the size of an anomalous interval. To see this consider intervals of length one. Then the value of the scan for any permutation of the data is the same. By symmetry the same reasoning applies for intervals of length \(N - 1\).

We consider essentially the same scan statistic (3) as before. The two main differences are that we scan over a smaller approximating net \(\bar{\mathcal{S}}_b := \bigcup_{q_l - 1}^{q_u} \mathcal{S}_{q_l}\), and do a “centering” of the statistic prior to the scan. Namely, let \(x = (x_v, v \in \mathcal{V})\) denote the observed data and define

\[
\text{SCAN}(x) = \max_{S \in \bar{\mathcal{S}}_b} \left( Y_S(x) - \sqrt{|S|} \bar{x} \right), \quad Y_S(x) := \frac{1}{\sqrt{|S|}} \sum_{v \in S} x_v,
\]

where \(\bar{x} = \frac{1}{N} \sum_{v \in \mathcal{V}} x_v\) is the overall average. The test rejects the null when

\[
\mathcal{P}(x) := \frac{1}{|\mathcal{V}|!} \left| \left\{ \pi \in \mathcal{V}^\pi : \text{SCAN}(x_\pi) \geq \text{SCAN}(x) \right\} \right| \leq \alpha,
\]

where \(\mathcal{P}(x)\) is the permutation P-value, and \(\alpha \in (0, 1)\) is the desired level.

Recall the definition of \(\theta\), in Section 1.2.

**Theorem 1.** Let \(0 < \alpha < 1\) and consider the test that rejects the null if \(\mathcal{P}(X) \leq \alpha\), where \(\mathcal{P}(X)\) is defined in (10). Assume that the anomalous set \(S\) belongs to \(\mathcal{S}\) defined in (8) with \(q_l - 3 \log_2 \log N \rightarrow +\infty\) and \(q_u - \log_2 N \rightarrow -\infty\) as \(N \rightarrow \infty\). In the definition of \(\bar{\mathcal{S}}_b\), take \(b = b_N \rightarrow \infty\) such that \(b_N = O(\log \log N)\). Then the test has level at most \(\alpha\) and power converging to 1 as \(N \rightarrow \infty\) provided that (5) holds and, unless \(\nu\) has compact support, \(\max_v \theta_v \leq \bar{\theta} < \theta_*\), for some fixed \(\theta > 0\).

The headline here is that a calibration by permutation has as much asymptotic power as a calibration by Monte-Carlo with full knowledge of the null distribution (to first-order accuracy). This is (qualitatively) in line with what is known in classical settings (Lehmann and Romano, 2005, Chap 15).
Note that the conditions required here allow \( S \) to be any class of intervals of lengths between \((\log N)^{3+\eta} \) and \( o(N) \), for any \( \eta > 0 \) fixed. This includes the most interesting cases of intervals not too short and also not too long. (The detection threshold for detecting intervals of length \( \propto N \) is in fact different. This relates to Remark 1). Indeed, for certain families of distributions removing from consideration very small interval sizes is essential and cannot be avoided. For instance consider a setup where \( X_v \sim \text{Bernoulli}(1/2) \), for all \( v \in \mathcal{V} \) under the null and \( X_v \sim \text{Bernoulli}(1) \), for all \( v \in S \) when \( S \) is anomalous. Even under the null we will encounter a run of ones of length \( \propto \log_2 N \) (the famous Erdős–Rényi Law) with positive probability. Therefore in this case the scan test, calibrated by Monte-Carlo or permutation, is powerless at detecting intervals of length \( \frac{1}{2} \log_2 N \). In fact, it can be shown that no test has any power here.

We place an upper bound on the nonzero \( \theta_v \)’s to streamline the proof arguments and also avoid special cases we were not able to rule out. For example, an open question is whether the power of this permutation test is monotone increasing in each of \( \{\theta_v : v \in S\} \) where \( S \) is in the anomalous set. If this is true, then obviously the upper bound (by \( \bar{\theta} \)) can be removed.

We note that when \( \nu \) does not have compact support, this can be enforced by applying a censoring. See Section 6.

**Proof.** Suppose we are under the null hypothesis. Note that \( X = (X_v, v \in \mathcal{V}) \) are IID under the null, and therefore exchangeable. This means that, for any permutation \( \pi \) the marginal distributions of \( \text{scan}(X) \) and \( \text{scan}(X_\pi) \) are the same. This implies that \( \text{scan}(X) \) is uniformly distributed on the set \( \{\text{scan}(X_\pi), \pi \in \mathcal{V}!\} \) (with multiplicities). With this we have

\[
\mathbb{P}(|\{\pi \in \mathcal{V}! : \text{scan}(X_\pi) \geq \text{scan}(X)\}| \geq \alpha \mathcal{V}!) \leq \frac{[\alpha \mathcal{V}]!}{\mathcal{V}!} \leq \alpha ,
\]

where \( [z] \) denotes the integer part of \( z \). If there were no ties, the first inequality above would be an equality, but with ties present the test becomes more conservative. For more details on permutation tests the reader is referred to (Lehmann and Romano, 2005).

All that remains to be done is to study the permutation test under the alternative hypothesis. This requires two main steps. First we need to control the randomness in the permutation, conditionally on the observations \( \mathbf{x} \). Once this is done we remove the conditioning on the observed data.

The key to the first step is the following Bernstein’s inequality for sums of variables sampled without replacement from a finite population.

**Lemma 2** (Bernstein’s inequality for sampling without replacement). Let \( (Z_1, \ldots, Z_m) \) be obtained by sampling without replacement from a given a set of real numbers \( \{z_1, \ldots, z_J\} \subset \mathbb{R} \). Define \( z_{\text{max}} = \max_j z_j, \bar{z} = \frac{1}{J} \sum_j z_j, \text{ and } \sigma_z^2 = \frac{1}{J} \sum_j (z_j - \bar{z})^2. \) Then the sample mean \( \bar{Z} = \frac{1}{m} \sum_i Z_i \) satisfies

\[
\mathbb{P} \left( \bar{Z} \geq \bar{z} + t \right) \leq \exp \left( - \frac{mt^2}{2\sigma_z^2 + \frac{2}{3}(z_{\text{max}} - \bar{z})t} \right), \quad \forall t \geq 0.
\]

This result is a consequence of (Hoeffding, 1963, Th. 4) and Chernoff’s bound, from which Bernstein’s inequality is derived, as in\(^3\) (Shorack and Wellner, 1986, p 851). See (Bardenet and Maillard, 2013; Boucheron et al., 2013; Dembo and Zeitouni, 2010) for a discussion of the literature on concentration inequalities for sums of random variables sampled without replacement from a finite set.

\(^3\)There is a typo in the statement of the result in (Shorack and Wellner, 1986, p 851), but following the proof one can find the correct result. Where the statement of the result reads \( -\frac{\lambda^2}{2\sigma^2} \) we should have \( -\frac{\lambda^2}{2\sigma^2} \) instead
Applying this result for a fixed (but arbitrary) set $S \in \tilde{S}_b$ when $\pi$ is uniformly drawn from $\mathcal{V}$! and $x$ is given, we get

$$\mathbb{P}\left( Y_S(x) - \sqrt{|S|} \bar{x} \geq t \right) \leq \exp \left[ - \frac{t^2}{2\sigma_x^2 + \frac{2}{3}(x_{\text{max}} - \bar{x})t/\sqrt{|S|}} \right], \quad \forall t \geq 0,$$

using the same notation as in Lemma 2. Plugging in $t = \text{scan}(x)$, noting that $|S| \geq 2^n$, and using this together with a union bound, we get

$$\mathbb{P}(x) \leq |\tilde{S}_b| \exp \left[ - \frac{\text{scan}(x)^2}{2\sigma_x^2 + \frac{2}{3}(x_{\text{max}} - \bar{x})2^{-n/2}\text{scan}(x)} \right]. \quad (11)$$

Now we proceed by upper bounding the right-hand side of the above inequality by assuming we are under the alternative, which yields an upper bound for the P-value $\mathbb{P}(X)$. This amounts to controlling the terms $X_{\text{max}} - \bar{X}, \sigma_X^2$ and $\text{scan}(X)$ under the alternative. (The upper-case $X$ relates to the random quantities.)

Assume that, without loss of generality, $f_0$ has zero mean and unit variance. Let $S$ denote the anomalous interval under the alternative and assume that $\theta_1 \leq \theta_v \leq \theta$ for all $v \in S$ with $\theta_1 = \sqrt{2(1+\varepsilon)} \log N/\sqrt{|S|}$ and $\varepsilon > 0$ fixed as before. Note that, by the assumption on $q_1$, we have $\theta_1 \to 0$ as $N \to \infty$.

We begin by controlling $X_{\text{max}} - \bar{X}$. For notational convenience define $\zeta_\theta = E_{\theta}(X)$ and $\sigma_\theta^2 = \text{Var}_{\theta}(X)$ for all $\theta \geq 0$. Note that $\zeta_\theta$ is increasing in $\theta$ and $\sigma_\theta$ is continuous in $\theta$, and in particular bounded on $[0, \theta]$. Let $\bar{X}_S = \frac{1}{|S|} \sum_{v \in S} X_v$ denote the sample mean over $S$. We have

$$\bar{X} = \frac{1}{N} \sum_{v} \zeta_{\theta_v} + \frac{1}{N} \sum_{v} (X_v - E(X_v)) = O(|S|/N) + o_P(1) = o_P(1),$$

as $N \to \infty$, since $|S| = o(N)$, using Chebyshev’s inequality in the second equality. Furthermore, let $X_{\text{max}, S} = \max_{v \in S} X_v$ be the maximum over $S$. We have $X_{\text{max}} = X_{\text{max}, S} \vee X_{\text{max}, \bar{S}}$, which by the union bound implies

$$\mathbb{P}(X_{\text{max}} > x) \leq \mathbb{P}(X_{\text{max}, S} > x) + \mathbb{P}(X_{\text{max}, \bar{S}} > x) \leq |S| \bar{F}_\theta(x) + |S^c| \bar{F}_0(x),$$

where $\bar{F}_\theta(x) = \mathbb{P}(X > x)$ and we used the fact that $\bar{F}_\theta(x)$ is monotone increasing in $\theta$ — see Section 1.2. For $c \in (0, \theta_* - \theta)$, we have

$$\bar{F}_\theta(x) = \int_x^\infty \frac{\phi_u}{\varphi_0(\theta)} f_0(u) \nu(du)$$

$$= \frac{1}{\varphi_0(\theta)} \int_x^\infty e^{-cu} e^{(\theta+c)u} f_0(u) \nu(du) \leq \frac{\varphi_0(\theta + c)}{\varphi_0(\theta)} e^{-cx}.$$

This exponential tail bound and the union bound that precedes it imply $\mathbb{P}(X_{\text{max}} > (2/c) \log N) \to 0$ as $N \to \infty$. This and the bound on $\bar{X}$ imply that

$$\mathbb{P}(X_{\text{max}} - \bar{X} > (3/c) \log N) \to 0.$$

We now consider $\sigma_X^2$. Similarly as before, we have

$$\sigma_X^2 = \frac{1}{N} \sum_{v \in \mathcal{V}} (X_v - \bar{X})^2 \leq \frac{1}{N} \sum_{v \in \mathcal{V}} X_v^2 = \frac{1}{N} \sum_{v} \sigma_{\theta_v}^2 + \frac{1}{N} \sum_{v \in \mathcal{V}} (X_v^2 - E(X_v^2))$$

$$= \left(1 - \frac{|S|}{N}\right) \sigma_0^2 + O(|S|/N) + o_P(1) = 1 + o_P(1),$$

where $\sigma_0^2$ is the variance of the model under the null hypothesis.
using the fact that $\max_{\theta \in [0, \theta]} \sigma_\theta^2 < \infty$ and $\max_{\theta \in [0, \theta]} \mathbb{E}_\theta(X^4) < \infty$ combined with Chebyshev’s inequality. In particular

$$
\mathbb{P}(\sigma_X^2 \leq 1 + \epsilon/4) \to 1.
$$

Finally we need to use a result analogous to Lemma 1 that applies to $\tilde{S}_b$. It is easy to see from the proof of Lemma 1 that there is a set $S^* \in \tilde{S}_b$ such that $S^* \subseteq S$ and $\rho(S^*, S) \geq \sqrt{\max(0, 1 - 2^{2-b})} = 1 - o(1)$ by the fact that $b \to \infty$. We then have

$$
\text{SCAN}(X) \geq Y_{S^*}(X) - \sqrt{|S^*|}X = \sqrt{|S^*|}(\bar{X}_{S^*} - \bar{X}) \geq \sqrt{\max(0, 1 - 2^{2-b})|S|}(\bar{X}_{S^*} - \bar{X}).
$$

By Chebyshev’s inequality again,

$$
\bar{X}_{S^*} = \frac{1}{|S^*|} \sum_{v \in S^*} \zeta_{\theta_v} + o_P(1) \geq \zeta_{\theta_1} + o_P(1).
$$

Furthermore, as seen in Section 2.2, $\zeta_{\theta} \geq \theta$ for $\theta \geq 0$. Hence,

$$
\text{SCAN}(X) \geq (\theta_1 + o_P(1))\sqrt{(1 - o(1))|S|} \geq \sqrt{2(1 + \epsilon/2) \log N},
$$

with probability tending to one as $N \to \infty$. We are now ready to evaluate the upper bound on the P-value given by (11). We have

$$
\log |\tilde{S}_b| \leq \log \left[2N/(\log_2 N)^{2b}\right] = \log N + O(\log \log N)^2 = (1 + o(1)) \log N, \quad (12)
$$

by our assumption on $b$, while the exponent in (11) is bounded from below by

$$
-\frac{2(1 + \epsilon/2) \log N}{2(1 + \epsilon/4) + (3/c)(\log N)^{2-\theta_1/2}\sqrt{2(1 + \epsilon/2) \log N}}
$$

with probability tending to one. Assuming $q_1 - 3\log_2 \log N \to +\infty$ guarantees that the second term in the denominator is a $o(1)$. Thus, eventually, this exponent is bounded from above by $-\left[(1 + \epsilon/2)/(1 + \epsilon/3)\right] \log N$. Combining this and the upper bound (12) allows us to conclude that $\log \mathbb{P}(X) \to -\infty$ with probability tending to one, meaning $\mathbb{P}(X) \to 0$, implying that the test as power tending to 1 as $N \to \infty$. 

\hfill \Box

4 Scanning the ranks

Having observed $x = (x_v, v \in V)$, scanning the ranks amounts to replacing every observation with its rank among all the observations, and computing the scan (9). We call this the rank scan. As for all rank-based methods, the null distribution is the permutation distribution when there are no ties.

- When there are no ties with probability one we calibrate the test by permutation, and this is done only once.

- When there are ties the rank scan test is also calibrated by permutation. If one breaks ties using the average rank then the calibration must be done every time as for the permutation test. Alternatively, one can break the ties randomly and the test is calibrated by permutation only once. The latter option is computationally superior and is the one we analyze.
See Section 5 for implementation issues.

Formally, let \( x = (x_v, v \in \mathcal{V}) \) denote the observations as before, and for every \( v \in \mathcal{V} \), let \( r_v \) be the rank (in increasing order) of \( x_v \) in \( x \), where ties are broken randomly, and let \( r = (r_v, v \in \mathcal{V}) \) be the vector of ranks. The rank scan test returns the P-value \( \mathcal{P}(r) \) defined in (10).

As we mentioned in the Introduction, an important advantage of the rank scan over the permutation scan is the fact that the former only requires calibration once, while the latter requires a new calibration with each dataset. This assumes that the size of \( \mathcal{V} \) of the node set remains the same.

An additional advantage of the rank scan is its robustness to outliers — although the permutation scan after censoring (discussed in Section 6) is also robust to outliers.

Because the rank scan test is a special case of the permutation scan test, we assume similarly upper and lower bounds on the size of the anomalous set as in Section 3. Define

\[
\Upsilon_v = \mathbb{E}[X \mathbf{1}_{\{X \geq Y\}}] + \frac{1}{2} \mathbb{E}[X \mathbf{1}_{\{X = Y\}}],
\]

where \( X, Y \sim F_0 \) and independent.

**Theorem 2.** Let \( 0 < \alpha < 1 \) and consider the test that rejects the null if \( \mathcal{P}(R) \leq \alpha \), as defined above. Assume that the anomalous set \( \mathcal{S} \) belongs to \( \mathcal{S} \) defined in (8) with \( q_0 - \log_2 N \to +\infty \) and that \( q_0 - \log_2 N \to -\infty \) as \( N \to \infty \). In the definition of \( \hat{S}_0 \), take \( b = b_N \to \infty \) such that \( b_N = O(\log \log N) \). Then the test has level at most \( \alpha \) and power converging to 1 as \( N \to \infty \) provided \( \theta_\mathcal{S} \geq (1+\varepsilon)C_0 \log(N)/|\mathcal{S}| \), with \( \varepsilon > 0 \) fixed and \( C_0 := 1/(6\Upsilon_0^2) \).

The headline here is that rank scan requires a signal amplitude which is \( \sqrt{C_0/2} \) larger than what is required of the regular scan test calibrated by Monte-Carlo with full knowledge of the null distribution. This is (qualitatively) in line with similar results in more classical settings (Hettmansperger, 1984). For the normal location model, we find that \( \sqrt{C_0/2} \leq \sqrt{\pi/3} \approx 1.023 \), so the detection threshold of rank scan is almost the same as that of the regular scan test. See the Appendix B for details.

**Proof.** The arguments used for the general permutation test apply verbatim under the null hypothesis, so all that remains to be done is to study the performance of the rank scan test under the alternative.

We may directly apply (11), to obtain

\[
\mathcal{P}(r) \leq |\hat{S}_0| \exp \left( -\frac{\text{SCAN}(r)^2}{\frac{N^2}{6} + \frac{N}{3} 2^{-(q-1)/2} \text{SCAN}(r)} \right),
\]

where we used \( \sigma_v^2 = (N^2 - 1)/12 < N^2/12, r_{\max} = N \) and \( \bar{r} = (N + 1)/2 \), so that \( r_{\max} - \bar{r} < N/2 \). The previous bounds can be directly computed when there are no ties in the ranks, and it is easy to verify that they also hold if ties are dealt with in any of the classical ways (assigning the average rank, randomly breaking ties, etc.). As before, this is a result conditional on the observations \( X = x \) and hence the ranks \( R = r \). The next step is to remove this conditioning, which now amounts to controlling the term \( \text{SCAN}(R) \).

Let \( \mathcal{S} \) denote the anomalous interval under the alternative and first assume that \( \theta_v = \theta_+ \) for all \( v \in \mathcal{S} \) with \( \theta_+ := \sqrt{(1+\varepsilon)C_0 \log \bar{N}/|\mathcal{S}|} \) and \( \varepsilon > 0 \) fixed. Note that since \( F_0 \) is assumed to have zero mean we have \( \mathbb{E}(X \mathbf{1}_{\{X > y\}}) \geq 0 \) and \( \mathbb{E}(X \mathbf{1}_{\{X \geq y\}}) \geq 0 \) for all \( y \). Hence one can easily check that \( \Upsilon_0 \geq 0 \). The reason behind the choice of the constant \( C_0 \) will become apparent at the end of the analysis. By our assumptions on the size of the anomalous set we have \( \theta_+ \to 0 \) as \( N \to \infty \).
As in the proof of Theorem 2, a result analogous to Lemma 1 applies. Namely, there is a set $S^* \in \mathbb{S}_b$ such that $S^* \in S$ and $\rho(S^*, S) \geq \sqrt{\max(0, 1 - 2^{-b})}$. Since

$$\text{SCAN}(R) \geq Y_{S^*}(R) - \sqrt{|S^*| N + 1},$$

we focus on obtaining a lower bound on $Y_{S^*}(R)$ that applies with high probability.

Take any $v, w \in S^*$ distinct and define

$$\zeta_\theta = E_\theta(R_v), \quad \sigma_\theta^2 = \text{Var}_\theta(R_v), \quad \xi_\theta = \text{Cov}_\theta(R_v, R_w).$$

Note that

$$E_\theta[Y_{S^*}(R)] = \sqrt{|S^*|} \zeta_\theta, \quad \text{Var}_\theta[Y_{S^*}(R)] = \sigma_\theta^2 + (|S^*| - 1)\xi_\theta,$$

and so, by Chebyshev’s inequality,

$$Y_{S^*}(R) = \sqrt{|S^*|} \zeta_\theta + O_P(\sqrt{\sigma_\theta^2 + |S^*|\xi_\theta}). \quad (14)$$

As in (Hettmansperger, 1984) we can derive the first two moments of the ranks.

**Lemma 3.** Suppose $A_1, \ldots, A_m$ are i.i.d and independent of $B_1, \ldots, B_n$ which are i.i.d, all being real-valued random variables. Let $R_i$ denote the rank (in increasing order) of $A_i$ in the combined sample, and suppose ties are broken randomly. Then, for any $i \neq i'$,

$$E(R_i) = \frac{m + n + 1}{2} + n(\lambda_1 - 1/2),$$

$$\text{Var}(R_i) = \frac{m - 1}{4} + n(\lambda_1 (1 - \lambda_1) + \frac{(m - 1)(m - 2)}{12} + n(n - 1)\lambda_3 - \lambda_1^2) + 2n(m - 1)(\lambda_4 - \lambda_1/2),$$

$$\text{Cov}(R_i, R_{i'}) = -\frac{1}{2} - \frac{m - 2}{12} + n(\lambda_2 - 2\lambda_1^2 + 2\lambda_5 - \lambda_1),$$

where

$$\lambda_1 := P(A_1 > B_1) + \frac{1}{2} P(A_1 = B_1),$$

$$\lambda_2 := P(\{A_1 > B_1\} \cap \{A_2 > B_1\}) + P(A_1 > A_2 = B_1) + \frac{1}{3} P(A_1 = A_2 = B_1),$$

$$\lambda_3 := P(\{A_1 > B_1\} \cap \{A_1 > B_2\}) + P(A_1 = B_1 > B_2) + \frac{1}{3} P(A_1 = B_1 = B_2),$$

$$\lambda_4 := P(\{A_1 > A_2\} \cap \{A_1 > B_1\}) + \frac{1}{2} (P(A_1 > A_2 = B_1) + P(A_1 = A_2 > B_1)) + \frac{1}{3} P(A_1 = A_2 = B_1),$$

$$\lambda_5 := P(\{A_2 > A_1\} \cap \{A_1 > B_1\}) + \frac{1}{2} (P(A_2 > A_1 = B_1) + P(A_1 = A_2 > B_1)) + \frac{1}{3} P(A_1 = A_2 = B_1).$$

For the sake of completeness we sketch a proof of Lemma 3 in Appendix C. What we directly use from the result above is the formula for the expectation, and the order of magnitude the variance and covariance terms. To emphasize the dependence on the underlying parameter $\theta$ we define

$$p_\theta = P(Y > X) + \frac{1}{2} P(Y = X), \quad \text{where } X \sim F_0 \text{ and } Y \sim F_\theta \text{ are independent.} \quad (15)$$

Using Lemma 3, we have

$$\zeta_\theta = \frac{N + 1}{2} + (N - |S|)(p_\theta - \frac{1}{2}), \quad \sigma_\theta^2 \leq 4N^2, \quad \xi_\theta \leq 2N.$$

Hence, together with (14), we obtain

$$Y_{S^*}(R) - \sqrt{|S^*| \frac{N + 1}{2}} = \sqrt{|S^*| (N - |S|)(p_\theta - \frac{1}{2}) + O_P(N)} \quad (16)$$

We continue by bounding $p_\theta$ by using a Taylor expansion. When $\nu$ is discrete, we have

$$p_\theta = \int_R (F_\theta(x) + \frac{1}{2} f_\theta(x)) \nu(dx).$$
We expand the integrand seen as a function of $\theta$, around $\theta = 0$. We have

$$\partial_{\theta=0} f_{\theta}\!(x) = x f_0(x), \quad \partial_{\theta=0} \bar{F}_\theta(x) = \int_{(x,\infty)} u f_0(u) \nu(du),$$

where the second one comes from differentiating inside the integral defining $\bar{F}_\theta$, justified by dominated convergence. Similarly, we have

$$c'_0 := \sup_x |\partial^2_{\theta=0} f_{\theta}(x)| < \infty, \quad c_0 := \sup_x |\partial^2_{\theta=0} \bar{F}_\theta(x)| < \infty.$$  

Hence,

$$p_\theta \geq \int_{\mathbb{R}} \left[ \bar{F}_0(x) + \frac{1}{2} f_0(x) + \theta \left( \int_{(x,\infty)} u f_0(u) \nu(du) + \frac{1}{2} x f_0(x) \right) - \frac{1}{2} (c_0 + c'_0/2) \theta^2 \right] f_0(x) \nu(dx)$$

$$= p_0 + \theta \left( \mathbb{E}_0[X 1_{(X-Y)}] + \frac{1}{2} \mathbb{E}_0[X 1_{(X=Y)}] \right) - \frac{1}{2} (c_0 + c'_0/2) \theta^2$$

$$= \frac{1}{2} \theta \mathcal{Y}_0 - \frac{1}{2} (c_0 + c'_0/2) \theta^2.$$  

When $\nu$ is continuous, we have

$$p_\theta = \int_{\mathbb{R}} \bar{F}_\theta(x) \nu(dx),$$

and the same calculations lead to

$$p_\theta \geq \frac{1}{2} \theta \mathcal{Y}_0 - \frac{1}{2} c_0 \theta^2.$$  

In any case, $p_\theta \geq \frac{1}{2} \theta \mathcal{Y}_0 - O(\theta^2).$

This lower bound, combined with (16), the fact that $|\mathcal{S}| = o(N)$, and the definitions of $C_0$ and $\mathcal{Y}_0$, yields

$$Y_{\mathcal{S}^*}(R) - \sqrt{|\mathcal{S}^*|} \frac{N}{2} = \rho(\mathcal{S}^*, \mathcal{S}) \sqrt{|\mathcal{S}|} (N - |\mathcal{S}|)(\theta_1 \mathcal{Y}_0 - O(\theta_1^2)) + O_P(N)$$

$$= (1 + o_P(1)) N \sqrt{\frac{1}{6}(1 + \varepsilon) \log N}$$

so that

$$\text{SCAN}(R) \geq N \sqrt{\frac{1}{6}(1 + \varepsilon/2) \log N}$$

with probability going to one as $N \to \infty$. The remaining arguments are exactly parallel to those at the end of Section 3. We have

$$\log \mathfrak{F}(R) \leq \log \left( 2N \log_2 N \right)^{2b} - \frac{N^2(1 + \varepsilon/2)(1 + \varepsilon/2) \log N}{N^2 / 3 + N^2 2^{-(q-1)/2} \sqrt{(1 + \varepsilon/2) \log N}}$$

$$\leq \log 2 + \log N + 2b \log \log_2 N - \frac{(1 + \varepsilon/2) \log N}{1 + o(1)},$$

with probability going to one as $N \to \infty$. Note that the RHS tends to $-\infty$ by our assumption of $b$.

We conclude that the rank scan test has power going to 1 as $N \to \infty$ when $\theta_v = \theta_1$ for all $v \in \mathcal{S}$ (the anomalous set).

Finally, to arrive at the same conclusion when $\theta_v \geq \theta_1$ for $v \in \mathcal{S}$ (as we assume in the statement of the theorem), note that $p_\theta$ is increasing in $\theta$ by virtue of the fact that $(F_\theta, \theta \geq 0)$ has monotone likelihood ratio.  

$\square$
The conditions of Theorem 2 allow for dealing with intervals of length of order (strictly) larger than \( \log N \). We give here results that encompass the scenario where the interval might be of smaller length. To keep the discussion simple we consider the class of intervals of a fixed size \( |S| = k \) under the alternative. In this situation, we simply scan over this class, denoted by \( S \).

Recall the definitions of \( p_\theta \) in (15) and the P-value of the permutation test (10). We can state the following result.

**Proposition 2.** Let \( 0 < \alpha < 1 \) and consider the test that rejects the null if \( \Psi(R) \leq \alpha \). Then, in the present context, the test has level at most \( \alpha \) and power converging to 1 as \( N \to \infty \) provided \( p_{\theta_0} \geq p_{\theta} \) for all \( \theta \in \Theta \) with:

\[
(i) \quad p_{\theta} = 1 - o(N^{-2/k}) \quad \text{when} \quad 2 < k = o(\log N); \quad \text{or}
(ii) \quad p_{\theta} > 1 - \exp(-\frac{c\theta}{\varepsilon})/2 \quad \text{when} \quad k = c \log N \quad \text{with an arbitrary constant} \quad c > 0.
\]

Theorem 2 and Proposition 2 cover essentially all interval sizes which are \( o(N) \). Theorem 2 covers the case of larger intervals, in which case \( p_{\theta} \) can go to 1/2 provided it does not converge too fast, and the test is still powerful asymptotically. In Proposition 2, a sufficient condition for an asymptotically powerful test is that \( p_{\theta} \) goes to 1 at a certain rate when the size of the anomalous interval is \( o(\log N) \). If the interval size is \( c \log N \) with \( c > 0 \) arbitrary the rank test is asymptotically powerful when \( p_{\theta} \) is greater than a constant (strictly larger than 1/2) depending on \( c \).

Unlike in Theorem 2, the statement of Proposition 2 is in terms of \( p_\theta \), and not \( \theta \) itself. This is due to the fact that, for small intervals, the signal magnitude must necessarily be large, implying that \( \theta \) is bounded away from zero. In such situations, one can only relate \( p_\theta \) and \( \theta \) with further knowledge about the family of distributions.

As an example, consider the normal means model when \( k = o(\log N) \). In this case, we have

\[
p_\theta = \Phi(-\theta/\sqrt{2}) \geq 1 - \frac{1}{2} e^{-\theta^2/4},
\]

where \( \Phi(x) \) is the CDF of the standard normal distribution. Hence, whenever \( \frac{1}{2} e^{-\theta^2/4} = o(N^{-2/k}) \), the condition in the proposition is met. This is satisfied when

\[
\theta = 2\sqrt{2(1+\varepsilon)(\log N)/k}, \tag{17}
\]

where \( \varepsilon > 0 \) is fixed. In view of (5), this means that in this case the rank scan requires an amplitude at most two times larger than the regular scan test calibrated with full knowledge of the null distribution.

Finally note that the condition \( p_{\theta_0} \to 1 \) or \( p_{\theta_0} > 1 - p_\theta \) might not be possible to meet for certain distributions of the exponential family. Recall the example of Bernoulli random variables discussed Section 3. In this setup \( p_{\theta_0} = 3/4 \), a case not covered by Proposition 2 when the interval size is smaller than \( c \log N \) and \( c \) is small enough. But this is expected since no test has any power if \( c \) is sufficiently small.

**Proof.** We treat each case separately.

**Condition (i).** The same arguments hold as before under the null, so again we are left with studying the alternative. To deal with smaller intervals, we need a slightly different concentration inequality than before.

**Lemma 4** (Chernoff’s inequality for ranks). In the context of Lemma 2, assume that \( z_j = j \) for all \( j \). Then

\[
P \left( \bar{Z} \geq \bar{z} + t \right) \leq \exp \left[ -m \sup_{\lambda \geq 0} \psi(t, \lambda) \right], \quad \forall t \geq 0, \quad \psi(t, \lambda) := \lambda t - \log \left( \frac{\sinh(\lambda N/2)}{N \sinh(\lambda/2)} \right).
\]
Similarly to Lemma 2 this result is also a consequence of (Hoeffding, 1963, Th. 4) and Chernoff’s bound. However, with the assumption on $z_j$ in the lemma above we can directly compute the moment generating function of $Z_j$ after using Chernoff’s bound instead of upper bounding it as is classically done to obtain Bernstein’s inequality.

In the present context, this yields

$$
\Psi(r) \leq |S| \exp \left[ -k \psi(\text{scan}(r)/\sqrt{\kappa}, \lambda) \right] \quad \forall \lambda > 0.
$$

As before, the next step is to bound $\text{scan}(R)$. Recall $\text{scan}(R) = \max_{S \subseteq \nu} Y_S(R) - \sqrt{k} N + 1$ as now $k$ is the common interval size. Note that we keep the centering to stay consistent with the previous definitions. Let $S$ denote the anomalous interval. As in the proof of Theorem 2, by monotonicity, we may assume that $\theta_v = \theta_1$ for all $v \in S$, with $\theta_1$ satisfying

$$
p_{\theta_1} = 1 - \frac{N - 2/k}{\omega_N}.
$$

For the sake of simplicity we present the proof when $\omega_N \to \infty$ not too fast. In particular assume $\omega_N = O(\log N)$. The proof can be carried out regardless of the choice of $\omega_N \to \infty$. Note that $Y_S(R)$ takes values in the interval $[0, \sqrt{k} N/2]$ because of the centering and the fact that we are considering only intervals of length $k$. For convenience we use the parametrization $Y_S(R) = \frac{1}{2} \sqrt{k} N (1 - \gamma_S(R))$ and control the term $\gamma_S(R)$. Thus

$$
\gamma_S(R) = 1 + \frac{N + 1}{N} - \frac{2Y_S(R)}{\sqrt{k} N}.
$$

By Lemma 3,

$$
\mathbb{E}_{\theta_1}[Y_S(R)] = \sqrt{k}(N + 1)/2 + \sqrt{k}(N - k) (p_{\theta_1} - 1/2)
$$

and

$$
\text{Var}_{\theta_1}[Y_S(R)] \leq p_{\theta_1} (1 - p_{\theta_1}) N^2 + O(kN),
$$

where for the variance we used the fact that $\lambda_3 \leq \lambda_1$ in Lemma 3. We thus have

$$
\mathbb{E}_{\theta_1}[\gamma_S(R)] \leq 2(1 - p_{\theta_1}) + O(k/N) \quad \text{and} \quad \text{Var}_{\theta_1}[\gamma_S(R)] \leq (1 - p_{\theta_1})/k + O(1/N).
$$

Therefore, using (18) and Chebyshev’s inequality we have

$$
P\left( \gamma_S(R) \geq \omega_N^{-1/4} N^{1/k} \right) = P\left( |\gamma_S(R) - \mathbb{E}(\gamma_S(R))| \leq \omega_N^{-1/4} N^{1/k} - 2(1 - p_{\theta_1}) - O(k/N) \right)
\leq P\left( \left| \gamma_S(R) - \mathbb{E}(\gamma_S(R)) \right| \geq 2 \omega_N^{-1/4} N^{1/k} \right)
\leq \frac{N^{-2/k}(k \omega_N) + O(1/N)}{4 N^{-2/k} \omega_N^{1/2}} \to 0,
$$

where the first inequality holds for $N$ large enough when $k = O(\log N)$. We conclude that when $k > 2$ we have $\gamma_S(R) \leq \tilde{\gamma} := \omega_N^{-1/4} N^{-1/k}$ with probability tending to 1. When this is the case, we have

$$
\psi(\text{scan}(R)/\sqrt{k}, \lambda) \geq \psi(\frac{1}{2} N (1 - \gamma_S(R)), \lambda)
\geq \lambda^2 / N (1 - \tilde{\gamma}) - \log \left( \frac{\sinh(\lambda N/2)}{N \sinh(\lambda/2)} \right),
$$
Choosing $\lambda = 1/(N\tilde{\gamma})$ and using the fact that $x \leq \sinh(x) \leq \frac{1}{2} e^x$ for all $x \geq 0$, we obtain

$$\psi(\text{SCAN}(R)/\sqrt{k}, \lambda) \geq \frac{1}{2} \frac{1 - \tilde{\gamma}}{\tilde{\gamma}} - \log \left( \frac{\sinh(1/(2\tilde{\gamma}))}{N\sinh(1/(2N\tilde{\gamma}))} \right)$$

$$= -\log \tilde{\gamma} - \frac{1}{2}.$$

Hence, using the fact that $|S| \leq N$, with probability tending to 1,

$$\log \mathcal{P}(R) \leq \log N + k \log \tilde{\gamma} + k/2 \tag{19}$$

$$= \log N + k \log(\omega^{-1/4} N^{-1/k}) + k/2$$

$$= -\frac{1}{4} k \log \omega_N + k/2 \to -\infty,$$

so that $\mathcal{P}(R) \to 0$ in probability, which is what we needed to prove.

**Condition (ii).** We can mimic the arguments above. Suppose $k = c \log N$ with arbitrary $c > 0$ and $p_{\theta_1} = 1 - \exp(-\frac{c+1}{c})/(2(1+\delta))$ with some $\delta > 0$. We can show that

$$\gamma(R) \leq \tilde{\gamma} := \frac{1+\delta/2}{1+\delta} \exp(-\frac{c+1}{c})$$

with probability going to 1 as $N \to \infty$. Plugging this new $\tilde{\gamma}$ and $k$ into (19) we see that the log of the P-value goes to $-\infty$, which is what we wanted to show. \qed

**Remark 2.** We considered the case when the size of the anomalous interval is known. However, we could consider the class of intervals of length greater than 2 and at most $\bar{k}$ for some given $\bar{k} = O(\log N)$. In this case we would simply perform a rank-scan for every fixed interval size up to $\bar{k}$ and apply a Bonferroni correction to the P-values. Following through the steps, one can see that the rank scan test would be asymptotically powerful when

(i') $p_{\theta_1} = 1 - o(\log N)^{-2/k}$ when $2 < k = o(\log N)$;

or when (ii) above holds. (This time $k$ is the size of the anomalous set under the alternative.) For the normal location model and considering $\bar{k} = o(\log N)$, we can see that this is satisfied when (17) holds.

## 5 Numerical experiments

### 5.1 Simulations

We present the results of some basic numerical experiments that we performed to corroborate our theoretical findings in finite samples. We generated the data from the normal location model — where $F_{\theta} = N(\theta, 1)$ — which is arguably the most emblematic one-parameter exponential family and a popular model in signal and image processing. We used the regular scan test, calibrated with full knowledge of the null distribution, as benchmark. The permutation scan test and rank scan test were calibrated by permutation.

To ease the computational and implementation burden, we only scan intervals of dyadic length. This allows for a very simple and fast implementation via one application of the Fast Fourier Transform per dyadic length. Letting $S$ denote the class of all discrete intervals in $\mathcal{V}$, this amounts to taking as approximating set

$$\mathcal{S}_{\text{dyad}} = \{ S \in \mathcal{S} : |S| = 2^j \text{ for some } j \in \mathbb{N} \}.$$
It is easy to see that, for each $S \in \mathcal{S}$, there is $S^* \in \mathcal{S}_{\text{dyad}}$ with $S^* \subset S$ and $|S^*| > |S|/2$. Hence,

$$\max_{S \in \mathcal{S}} \min_{S^* \in \mathcal{S}_{\text{dyad}}} \rho(S, S^*) \geq 1/\sqrt{2}.$$

A priori, this implies that scanning over $\mathcal{S}_{\text{dyad}}$ requires an amplitude $\sqrt{2}$ larger to achieve the same (asymptotic) performance as scanning over $\mathcal{S}$ or a finer approximating set as considered previously. To simplify things, however, in our simulations we took an anomalous interval of dyadic length, so that the detection threshold is in fact the same as before.

We set $N = 2^{15}$ and tried two different lengths for the anomalous interval $|S| \in \{2^7, 2^{10}\}$. All the nonzero $\theta_s'$s were taken to be equal to

$$\theta_S = t \sqrt{2 \log N / \sqrt{|S|}} \quad (20)$$

with $t$ varying. The critical values and power are based on 1000 repeats in each case. A level of significance of 0.05 was used. Also, 200 permutations were used for the permutation scan test. The results are presented in Figure 5.1. At least in these small numerical experiments, the three tests behave comparably, with the rank scan slightly dominating the others. Although the last finding is somewhat surprising, this is a finite-sample effect and is localized in the intermediate power range (around a power of 0.5) and so does not contradict the theory developed earlier. In fact, the three tests achieve power 1 at roughly the same signal amplitude, confirming the theory.

**Computational complexity** In terms of computational complexity, the regular scan test and the rank scan test are comparable, while the permutation test requires many more computations:

- The scan test is calibrated by Monte-Carlo with full knowledge of the null distribution.
- The rank scan test is calibrated by permutation and this is done only once. We mentioned in Section 4 that this calibration is exact when there are no ties. Although ties are common in practice, it is routine to apply the same protocol even if ties are present or expected. This is what is traditionally done with rank-based tests. And when breaking the ties randomly we recover the exactness of the test — at the expense of extrinsic randomization.
- The permutation test is much more demanding. Indeed, the mere computation (in fact, estimation) of its test statistic (10) requires the permutation of the data multiple times, each time followed by scanning.

### 5.2 Comparison with RSI

Next, we compare our rank scan with the robust segment identifier (RSI) of Cai et al. (2012). This is a recent method based taking the median over bins of a certain size (a tuning parameter of the method) and then scanning over intervals. Because the median is asymptotically normal, it allows for a calibration that only requires the value of the null density at 0. In turn, one can try to estimate this parameter. Although the method is not distribution-free proper, it appears to be the main contender in the literature. We first compare the two methods on simulated data, for in the context of detection (the problem we considered so far) and in the context of identification (a problem considered in that paper).
Figure 1: Power curves (with 95% margin of error) for the three tests (all set at level 0.05) as a function of the parameter \( t \) in (20): the scan test calibrated with knowledge of the null distribution (black); the permutation scan test (blue); and the rank scan test (red). On the left are the results for \(|S| = 2^7\) and on the right for \(|S| = 2^{10}\). \( N = 2^{15}\) in both cases. Each situation was repeated 1,000 times and each time 200 permutations were drawn for calibration. The vertical black dashed line is the minimax boundary for \( t \). The horizontal black dashed line is the significance level 0.05.

Detection  In the problem of detection, we compare the performance of the rank scan test and RSI with bin size \( m \in \{10, 20\} \) in normal data. To turn RSI into a test, we reject if it detects any anomalous interval. In the simulation, we set sample size \( N = 50,000 \) and considered the case where there is only one signal interval with known length \(|S| \in \{100, 1000\}\). The amplitude satisfy (20) as before. We report the empirical power curves (based on 100 repeats) in Figure 2.

To be fair, both methods only scan candidate signal intervals of length \(|S|\). The rank scan is calibrated as before. For RSI, we set the threshold to \( \sqrt{2 \log(N)} \) for the normalized data after localization to better control the family-wise type I error as explained in (Cai et al., 2012). From Figure 2, we can see that RSI is a bit more conservative. In fact, a drawback of RSI is the difficulty to calibrate it correctly.\(^4\) In any case, the rank scan test outperforms RSI in these simulations.

Identification  In the problem of identification, we compare the rank scan and RSI. Although we focused on the problem of detection so far, a scan can be as easily used for testing as for estimation (i.e., identification). Indeed, one sets an identification threshold and extract all the intervals that exceed that threshold. Some post-processing — such as merging significant intervals that intersect or keeping the most significant among significant intervals that intersect — is often applied.

\(^4\)Of course, it could be calibrated by permutation, but this would make the procedure much more like the permutation scan test (with the same high-computational burden), somewhat far from the intentions of (Cai et al., 2012).
Figure 2: Power curves (with 95% margin of error) for the three tests as a function of the parameter \( t \) in (20): the rank scan test (red); RSI with bin size 10 (solid green); and RSI with bin size 20 (dashed green). The rank scan test is set at level 0.05 and its critical value is from 1000 repeats. On the left are the results for \( |S| = 100 \) and on the right for \( |S| = 1000 \). \( N = 50,000 \) in both cases. Each situation was repeated 100 times. The vertical black dashed line is the minimax threshold for \( t \). The horizontal black dashed line is the significance level 0.05.

Here, in an effort to be fair, we simply took the procedure of (Cai et al., 2012) — which is essentially the procedure of (Jeng et al., 2010) — but scanning ranks and calibrating as we did for testing. Note that this implies a very stringent false identification rate. (At the 0.05 testing level, this means that the chances that one or more intervals are identified by mistake is 0.05.)

Following (Cai et al., 2012), in the simulation, we set the sample size to \( N = 10^4 \). We consider a range of null distributions: the standard normal distribution, the \( t \)-distribution with 15 degrees of freedom and that with one degree of freedom. In each case, we set the signal mean to \( \theta_s \in \{1, 1.5, 2\} \).

There are three signal intervals, \( S_1, S_2, S_3 \), starting at positions 1000, 2000, 3000, and having lengths \( 2^4, 2^5, 2^6 \), respectively. We set the threshold for the rank scan test by simulation at a significance level of 0.05. For RSI, we tried several bin sizes, \( m \in \{2^3, 2^5\} \). To simplify the computation, both methods only scan dyadic intervals of length at most \( 2^6 \). As in (Cai et al., 2012), we compare their performance in terms of the following dissimilarities

\[
D_j = \min_{\hat{S} \in \hat{S}} \{1 - \rho(S_j, \hat{S})\},
\]

and the number of false positives, namely

\[
O = \{\hat{S} \in \hat{S} : \hat{S} \cap S = \emptyset, \forall S \in S\},
\]

where \( \hat{S} \) are the estimated signal intervals.

We report the average and standard deviation (in the parenthesis in the tables below) based on 200 repeats in Tables 1, 2, and 3. We can see that the rank scan method performs better than RSI in when the null distribution is normal and \( t(15) \), and it performs similarly to RSI with bin size \( m = 2^3 \) in \( t(1) \). However, when the bin size of RSI is not properly chosen, RSI can perform poorly.
Table 1: Dissimilarity and number of over-selected intervals in $\mathcal{N}(0,1)$

| $\theta_S$ | Method      | $D_1(|S_1| = 2^2)$ | $D_2(|S_2| = 2^3)$ | $D_3(|S_3| = 2^6)$ | #O   |
|------------|-------------|----------------------|----------------------|----------------------|------|
| 1          | Rank Scan   | 0.734 (0.421)        | 0.148 (0.284)        | 0.031 (0.049)        | 0.000 (0.000) |
|            | RSI($m = 2^3$) | 0.916 (0.235) | 0.420 (0.406) | 0.095 (0.091) | 0.065 (0.267) |
|            | RSI($m = 2^5$) | 0.998 (0.029) | 0.959 (0.144) | 0.326 (0.278) | 0.130 (0.337) |
| 1.5        | Rank Scan   | 0.167 (0.326)        | 0.019 (0.044)        | 0.008 (0.012)        | 0.000 (0.000) |
|            | RSI($m = 2^3$) | 0.593 (0.391) | 0.132 (0.033) | 0.069 (0.029) | 0.080 (0.272) |
|            | RSI($m = 2^5$) | 0.980 (0.087) | 0.729 (0.284) | 0.204 (0.044) | 0.025 (0.157) |
| 2          | Rank Scan   | 0.018 (0.051)        | 0.006 (0.024)        | 0.004 (0.008)        | 0.000 (0.000) |
|            | RSI($m = 2^3$) | 0.277 (0.226) | 0.128 (0.021) | 0.064 (0.013) | 0.065 (0.247) |
|            | RSI($m = 2^5$) | 0.960 (0.122) | 0.476 (0.162) | 0.193 (0.032) | 0.010 (0.100) |

Table 2: Dissimilarity and number of over-selected intervals in $t(15)$

| $\theta_S$ | Method      | $D_1(|S_1| = 2^2)$ | $D_2(|S_2| = 2^3)$ | $D_3(|S_3| = 2^6)$ | #O   |
|------------|-------------|----------------------|----------------------|----------------------|------|
| 1          | Rank Scan   | 0.806 (0.369)        | 0.223 (0.354)        | 0.029 (0.048)        | 0.000 (0.000) |
|            | RSI($m = 2^3$) | 0.926 (0.223) | 0.436 (0.406) | 0.106 (0.099) | 0.050 (0.218) |
|            | RSI($m = 2^5$) | 0.996 (0.041) | 0.944 (0.168) | 0.336 (0.278) | 0.125 (0.332) |
| 1.5        | Rank Scan   | 0.232 (0.378)        | 0.026 (0.079)        | 0.010 (0.017)        | 0.000 (0.000) |
|            | RSI($m = 2^3$) | 0.554 (0.391) | 0.143 (0.112) | 0.069 (0.031) | 0.075 (0.282) |
|            | RSI($m = 2^5$) | 0.992 (0.057) | 0.732 (0.286) | 0.199 (0.042) | 0.020 (0.140) |
| 2          | Rank Scan   | 0.034 (0.097)        | 0.009 (0.019)        | 0.005 (0.014)        | 0.000 (0.000) |
|            | RSI($m = 2^3$) | 0.277 (0.220) | 0.128 (0.022) | 0.063 (0.013) | 0.060 (0.238) |
|            | RSI($m = 2^5$) | 0.968 (0.107) | 0.521 (0.214) | 0.192 (0.030) | 0.010 (0.100) |

5.3 Application to the real data

In this section, we apply the methods to the problem of detecting the copy number variant (CNV) in the context of next generation sequencing data. We compare the rank scan method and RSI on the task of identifying short reads on chromosome 19 of a HapMap Yoruban female sample (NA19240) from the 1000 genomes project (http://www.1000genomes.org), which is the same data set used in (Cai et al., 2012). Following standard protocols (Ernst et al., 2011), we extend all the reads to 100 base pairs (BPs). We take $10^6$ reads from the whole data set for comparison purpose and get 1,281,502 genomic locations.

We tune RSI as done in (Cai et al., 2012), setting the bin size to $m = 400$ and the maximum BPs in a possible CNV to $L = 2^{16}$. Note that (Cai et al., 2012) took $L = 60,000$, which is a bit smaller than $2^{16}$. (We chose the latter because we have the methods scan intervals of dyadic length.) To save computational time, in the implementation of the rank scan method in this experiment we group read depths in every 200 positions and take the summation of the read depths for each bins.
then apply the rank scan method on the sums. However, one can definitely apply the rank scan method on the original scale with a proper scanning range. We get the critical value for the rank scan method under the significance level 0.05 from 1000 repeats. In the experiment, we let RSI and the rank scan method only scan dyadic intervals of lengths from $2^1$ to $2^{16}$.

After merging the contiguous selected segments, RSI found 30 possible CNVs and the rank scan method selected 34. Figure 3 shows the histograms of the read depths of the selected CNVs. We can see the read depth in the rank scan method is generally larger than that in RSI. In the end, it is difficult to conclude which is method comes out best without confirmation from biological experiments.

| $\theta_S$ | Method       | $D_1(|S_1| = 2^3)$ | $D_2(|S_2| = 2^3)$ | $D_3(|S_3| = 2^6)$ | #O |
|-----------|--------------|---------------------|---------------------|---------------------|----|
| 1         | Rank Scan    | 0.989 (0.082)       | 0.878 (0.305)       | 0.461 (0.448)       | 0.000 (0.000) |
|           | RSI($m = 2^3$) | 0.950 (0.186)       | 0.764 (0.370)       | 0.332 (0.358)       | 4.305 (5.653) |
|           | RSI($m = 2^5$) | 0.998 (0.022)       | 0.982 (0.098)       | 0.609 (0.392)       | 0.520 (0.501) |
| 1.5       | Rank Scan    | 0.922 (0.251)       | 0.542 (0.455)       | 0.067 (0.132)       | 0.000 (0.000) |
|           | RSI($m = 2^3$) | 0.843 (0.307)       | 0.342 (0.354)       | 0.104 (0.080)       | 3.920 (2.082) |
|           | RSI($m = 2^5$) | 0.983 (0.079)       | 0.877 (0.236)       | 0.225 (0.111)       | 0.055 (0.229) |
| 2         | Rank Scan    | 0.763 (0.410)       | 0.206 (0.333)       | 0.043 (0.093)       | 0.000 (0.000) |
|           | RSI($m = 2^3$) | 0.619 (0.382)       | 0.154 (0.121)       | 0.089 (0.063)       | 3.945 (2.385) |
|           | RSI($m = 2^5$) | 0.978 (0.090)       | 0.667 (0.280)       | 0.208 (0.05)        | 0.060 (0.238) |

Figure 3: Histogram of the read depths of the selected CNVs in log scale (base 10). Both methods only scan dyadic intervals of lengths from $2^1$ to $2^{16}$. The RSI used a bin size $m = 400$, while the rank scan was calibrated as for testing.
6 Discussion

In this paper we consider distribution-free detection problems, in particular using tests based on scan statistics. In many contexts these are known to be optimal when one has full knowledge of the null distribution. When such knowledge is not available one can alternatively calibrate the test by permutation. Theorem 1 shows that calibration by permutation has asymptotically the same power as the regular scan test calibrated with the full knowledge of the distribution (to first-order accuracy). Along this line of reasoning we also propose a rank-scan test, where the original observations are replaced by the corresponding rank statistics. In this case Theorem 2 guarantees that the rank-scan test is powerful provided the signal magnitude is slightly larger than that required by the regular scan test. In our empirical experiments all three methods perform comparably. Interestingly, in our empirical experiments the rank-scan test appears to be slightly more powerful than regular-scan test. Moreover, by and large, the rank-scan methodology improves on the RSI methodology of Cai et al. (2012) at least on simulated data where we now the ground truth.

Censoring before permutation When \( \nu \) is not of compact support, we can enforce it by applying a censoring of the form \( \tilde{X}_v = X_v 1\{X_v \leq t\} + t \text{sign}(X_v) 1\{X_v > t\} \). With a choice of threshold \( t = t_N \to \infty \) slowly (e.g., \( t_N = \log \log N \)), Theorem 1 applies unchanged and without an upper bound on the \( \theta'_v \)'s, and the proof is parallel with only minor modifications. This censoring has the added advantage of making the method more robust to possible outliers.

Other scoring functions Although rank-sums are intuitive and classically used, any scan based on \( h(r_v) \), where \( h \) is increasing, is valid. (Recall that \( r_v \) is the rank of \( x_v \) in the sample.) In the classical setting, it is known that there is no uniformly better choice of function \( h \). See (Lehmann and Romano, 2005, Sec 6.9), where it is shown that choosing \( h(r) = \mathbb{E}(Z(r)) \) — where \( Z_1 < \cdots < Z_N \) are the order statistics of a standard normal sample — is optimal in the normal location model. Our method of proof applies to a general \( h \).

Acknowledgments

The basic concept of this paper arose out of conversations between some of the authors at the 2013 Mathematical Statistics conference in Luminy, France. We would like to acknowledge Dr. Bing Zhou (in the Department of Cellular and Molecular Medicine at UCSD) for help with the real data applications of Section 5.3. This work was partially supported by a grant from the US National Science Foundation (DMS 1223137) and a grant from the Nederlandse organisatie voor Wetenschappelijk Onderzoek (NWO 613.001.114).

A Construction and properties of the approximation net \( S_b \)

For notational convenience represent the discrete interval \( \{a, a + 1, \ldots, b\} \) as \([a, b]\). Assume, for simplicity, that \( N = 2^q \) for some integer \( \tilde{q} \). Let \( 0 \leq q \leq \tilde{q} - 1 \) and \( 0 \leq b \leq \tilde{q} \) be integers and define \( S_{q,b} \) to be the class of intervals of the form

\[
[1 + r2^q - \sum_{i=1}^\ell 2^{q_i}, (r + 1)2^q + 2^q + \sum_{i=1}^{\ell'} 2^{q'_i}],
\]

where \( r = 0, \ldots, 2^{\tilde{q} - q} - 1 \), \( 0 \leq \ell, \ell' \leq b \), \( q \geq q_1 \geq q_2 \geq \cdots \geq q_\ell \geq 0 \) and \( q \geq q'_1 \geq q'_2 \geq \cdots \geq q'_{\ell'} \geq 0 \). In words, the intervals in \( S_{q,b} \) are obtained by taking the dyadic interval \([1 + r2^q, (r + 1)2^q]\) (of length \( 2^q \)) and
appending to both sides at most \( b \) intervals of dyadic lengths no greater than \( 2^q \). In what follows we assume the parameters \( \ell, q, q_1, \ldots, q_1', \ldots, q_n' \) are taken so that the interval endpoints are in the range \( \{1, \ldots, N\} \).

Finally, the class \( S \) is defined as

\[
S = \bigcup_{q=0}^{\bar{q}-1} S_{q,b}.
\]

**Proof of Lemma 1.** By construction, the cardinality of the class \( S_{q,b} \) is upper bounded by \( 2^{q-q}(q + 2)^{2b} \), where the term \( (q + 2) \) accounts for the possible \( (q + 1) \) values for \( q_1 \) and \( q_1' \) and its presence or absence in the summation in the definition (21). Therefore the cardinality of \( S \) satisfies

\[
|S_b| \leq \sum_{q=0}^{\bar{q}-1} |S_{q,b}|
\]

\[
\leq \sum_{q=0}^{\bar{q}-1} 2^{q-q}(q + 2)^{2b} \leq 2 \cdot 2^\bar{q}(q + 2)^{2b} = 2N(\log_2(4N))^{2b}.
\]

This concludes the proof of the first claim in the lemma. For the second claim let \( S \in S \) be an arbitrary interval of the form \([a + 1, a + k]\). If \( k = 1 \) then \( S \in S_{0,b} \), and therefore taking \( S^* = S \) guarantees that \( \rho(S^*,S) = 1 \). For the case \( k > 1 \) let \( q \) be such that \( k \in [2^{q+1}, 2^{q+2}] \) (such an integer always exists). There is an interval of the form \([1 + r2^q, (r + 1)2^q]\) contained in \( S \). Now take this interval and \( \ell, \ell', q, q_1, \ldots, q_1', \ldots, q_n' \) so that the interval in (21) is still contained in \( S \) and has maximal length. This defines the interval \( S^* \equiv [1 + a', a' + k'] \in S_b \). The next step is to relate \( k \) to \( k' \). Note that \( (1 + r2^q) - (1 + a) \leq k - 2^q \leq 2^{q+2} - 2^q \). This means that, in the worst possible case \( a' - a \leq 2^{q+2} - b \) (this is the case where \( (1 + r2^q) - (1 + a) = 2^{q+2} - 2^q - 1 \), and so we must take \( q_1 = q, q_2 = q, q_3 = q - 1, q_4 = q - 2 \) and so on). Using an analogous argument we conclude also that \( (a + k) - (a' + k') \leq 2^{q+2} - b \). This means that \( k' - k \leq 2 \cdot 2^{q+2} - b = 2^{q+3} - b \). Finally

\[
\rho(S^*,S) = \sqrt{\frac{k}{k'}} \geq \sqrt{\max \left( 1 - \frac{2^{q+3} - b}{k'}, 0 \right)} \geq \sqrt{\max(1 - 2^{q+2}, 0)},
\]

where we used the fact that \( S^* \subset S \), and \( k \geq 2^q + 1 \), concluding the proof.

**B Derivation of \( C_0 \) in the normal location model**

Assume the normal model where \( F_0 = \mathcal{N}(\theta, 1) \). For this case we can derive a very good bound for \( \Upsilon_0 \) (confirmed by numerical integration). Since there are no ties with probability 1, we have

\[
\Upsilon_0 = \mathbb{E}(X \mathbb{1}_{(X > Y)}) = \int_{-\infty}^{\infty} \int_{\mathbb{R}} u f_0(u) du f_0(x) dx = \int_{\mathbb{R}} u f_0(u) F_0(u) du.
\]

Here \( F_0 \) is the standard normal distribution, which can be expressed as the following infinite sum

\[
F_0(u) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \sum_{j=0}^{\infty} \frac{u^{2j+1}}{(2j+1)!!} = \frac{1}{2} + f_0(u) \sum_{j=0}^{\infty} \frac{u^{2j+1}}{(2j+1)!!}.
\]

The expression above can be upper-bounded by \( 1/2 + u f_0(u) \) when \( u < 0 \) and lower-bounded by the same expression when \( u \geq 0 \). Hence we have

\[
\Upsilon_0 \geq \int_{\mathbb{R}} u f_0(u)(1/2 + u f_0(u)) du = \int_{\mathbb{R}} u^2 f_0^2(u) du = \frac{1}{2\sqrt{\pi}},
\]

where the last step follows from integration by parts. This yields \( C_0 = 1/(6\Upsilon_0^2) \leq 4\pi/6 \approx 2.094395 \). (Numerical integration yields the same value up to this precision!)
C Sketch proof of Lemma 3

First, assume that there are no ties in the ranks, with probability one. Note that we can write

\[ R_i = 1 + \sum_{j=1}^{m} 1_{A_i > A_j} + \sum_{j=1}^{n} 1_{A_i > B_j} \]

Taking expectation yields

\[ \mathbb{E}(R_i) = \frac{m+1}{2} + n \mathbb{P}(A_1 > B_1) = \frac{m+n+1}{2} + n(p_1 - 1/2), \]

since when there are no ties \( \mathbb{P}(A_1 = B_1) = 0 \). The variance and covariance terms can be worked out using the same representation of the ranks as above, but we omit these straightforward computations so that the presentation does not become cluttered.

In case of ties, to keep the presentation simple, assume that the distributions of \( A_1 \) and \( B_1 \) are supported on \( \mathbb{Z} \). Then randomly breaking ties in the ranks amounts to using the following procedure. Let \( \epsilon_1, \ldots, \epsilon_m \) and \( \delta_1, \ldots, \delta_n \) be independent and uniformly distributed on \( (-c, c) \) with \( c \leq 1/2 \), also independent from \( A_1, \ldots, A_m, B_1, \ldots, B_n \). Consider \( A'_i = A_i + \epsilon_i, \ i = 1, \ldots, m \) and \( B'_i = B_i + \delta_i, \ i = 1, \ldots, n \), and let \( R'_i \) be the rank of \( A'_i \) in the combined sample \( A'_1, \ldots, A'_m, B'_1, \ldots, B'_n \). Then the joint distribution of \( (R'_i, i = 1, \ldots, m) \) is the same as that of \( (R_i, i = 1, \ldots, m) \) when ties are broken randomly. However, for the ranks of the modified random variables, we have that there are no ties with probability one, hence we can apply the previous part of the proof. Hence we have

\[ \mathbb{E}(R'_i) = \frac{m+n+1}{2} + n(\mathbb{P}(A'_1 > B'_1) - 1/2), \]

where

\[ \mathbb{P}(A'_1 > B'_1) = \mathbb{P}(A_1 > B_1) + \mathbb{P}(\{A_1 = B_1\} \cap \{\epsilon_1 > \delta_1\}) = \mathbb{P}(A_1 > B_1) + \frac{1}{2} \mathbb{P}(A_1 = B_1) = p_1. \]

The variance and covariance can be dealt with in the same way.

Finally, when \( A_1, B_1 \) have arbitrary distributions a similar method can be applied, although it requires a bit more care and one needs to take \( c \) approaching zero.

References


