Anomaly Detection for a Large Number of Streams: A Permutation-Based Higher Criticism Approach

Ivo V. Stoepker* a, Rui M. Castro a, Ery Arias-Castro b, and Edwin van den Heuvel a

aDepartment of Mathematics and Computer Science, Technische Universiteit Eindhoven, Eindhoven, The Netherlands; bDepartment of Mathematics and Halicioğlu Data Science Institute, University of California, San Diego, La Jolla, CA, USA

ABSTRACT
Anomaly detection when observing a large number of data streams is essential in a variety of applications, ranging from epidemiological studies to monitoring of complex systems. High-dimensional scenarios are usually tackled with scan-statistics and related methods, requiring stringent modeling assumptions for proper calibration. In this work we take a nonparametric stance, and propose a permutation-based variant of the higher criticism statistic not requiring knowledge of the null distribution. This results in an exact test in finite samples which is asymptotically optimal in the wide class of exponential models. We demonstrate the power loss in finite samples is minimal with respect to the oracle test. Furthermore, since the proposed statistic does not rely on asymptotic approximations it typically performs better than popular variants of higher criticism that rely on such approximations. We include recommendations such that the test can be readily applied in practice, and demonstrate its applicability in monitoring the content uniformity of a batch-produced drug product. Supplementary materials for this article are available online.

1. Introduction
We study the problem of anomaly detection when a large number of data streams is observed. The streams themselves are not structured in any special way. In particular, there is no spatial or other proximity measure between streams, unlike in some areas like syndromic surveillance (Kulldorff et al. 2005). We are interested in sparse alternatives; that is, anomalies affect only a small fraction of the streams. Our analysis assumes an anomalous stream is affected in its entirety. This is common in literature (see for instance Thomopoulos, Viswanathan, and Bougoulias 1989; Patwari and Hero 2003; Aldosari and Moura 2004; Lexa et al. 2004; Yu et al. 2006; Mei 2008) and greatly simplifies the presentation of power guarantees. Note, however, that the proposed methodology is naturally applicable also when the streams are partially affected, and will still have some power.

Formally we consider $n$ streams indexed by $[n] = \{1, \ldots, n\}$, that are observed over a time period $[t] = \{1, \ldots, t\}$. Specifically we observe $X \equiv (X_{ij} : i \in [n], j \in [t])$, where $X_{ij}$ is the value from stream $i \in [n]$ collected at time $j \in [t]$. Stream $i \in [n]$ corresponds to a vector denoted by $X_i \equiv (X_{ij} : j \in [t])$.

Following standard practice, we place ourselves in a statistical decision theory framework and cast detection as a hypothesis testing problem. For simplicity, we assume the variables to be real valued and independent.

In this article, we consider a stationary setting where the distribution of the observations remain unchanged over time. In detail, under the null hypothesis all $X_{ij}$ are independent and identically distributed (iid) with common distribution function denoted by $F_0$. The alternative hypothesis is similar, but there is a subset of anomalous streams $S \subset [n]$ such that for each $i \in S$ the distribution of $X_{ij}$, denoted by $F_i$, stochastically dominates $F_0$. In words, larger-than-usual values are observed in the streams indexed by $S$. The subset $S$ is unknown and has no particular structure. Note that we are still assuming all the random variables $X_{ij}$ are independent. The sparsity assumption means the number of affected streams is believed to be small in comparison with the total number of streams, that is, $|S| \ll n$.

Succinctly, our hypothesis testing problem is therefore:

$$H_0 : \forall_{i \in [n]} X_{ij} \overset{iid}{\sim} F_0$$

$$H_1 : \exists_{S \subset [n]} : \forall_{i \in S} X_{ij} \overset{iid}{\sim} F_i \text{ and } \forall_{i \notin S} X_{ij} \overset{iid}{\sim} F_0,$$

where for all $i \in S$, distribution $F_i$ stochastically dominates $F_0$.

Despite its apparent simplicity, the above formulation and ensuing hypothesis test is quite suitable in a number of scenarios. For instance, in industrial settings the stream observations $X_i$, $i \in [n]$ might correspond to a failure metric measured from $n$ machines. Under nominal conditions we expect these to be iid samples from some distribution, while under the anomalous conditions some of these machines will occasionally display higher-than-normal values for the failure metric. In more complex settings the observations might correspond to residual values, obtained by fitting a model to the streams, as a way to ensure the streams can be meaningfully compared. For instance, in gene expression studies the streams might correspond to the normalized gene expression level of a number of replicates, so...
that under nominal conditions the expression levels are "equalized," and one is trying to determine if genes are differentially expressed under a specific environmental change. Obviously in this and similar settings some pre-processing of the data is needed to ensure all the streams are comparable. In Section 2 in the supplementary materials, we consider yet a different application, where the proposed methodology is applied to the daily counts of diagnosed COVID-19 cases in different municipalities in the Netherlands. This application also showcases how our methodology can be used when observations are dependent, by modeling the dependence and applying our methodology on the corresponding residuals.

**Contributions.** Although a thorough characterization of the above testing scenario is far from trivial, it has been addressed quite extensively when the distribution $F_0$ is assumed known, and in particular when this is a normal distribution (Ingster 1997; Donoho and Jin 2004). Knowledge of that distribution is required to calibrate the proposed tests (e.g., by Monte Carlo simulation) and ensure the resulting p-value computation is done properly. When $F_0$ is unknown there are various ways one can proceed. A natural nonparametric approach is calibration by permutation. This approach is, in fact, quite standard in syndromic surveillance (Kulldorff et al. 2005; Huang, Kulldorff, and Gregorio 2007; Kulldorff, Huang, and Konty 2009). It is also common in neuroimaging (Nichols and Holmes 2002) and other contexts (Walther 2010; Flenner and Hewer 2011).

In this work we propose a variant of the higher criticism statistic that relies on permutations in both its definition and its calibration, leading to an exact test. Furthermore, we show that, in the context of a one-parameter exponential family this test is asymptotically optimal, in the sense that it has essentially the same asymptotic performance as the best possible test with full knowledge of the underlying model—referred to as the oracle test. We provide practical recommendations so that the test is readily applicable in practice and demonstrate it has good performance in finite sample scenarios. In terms of testing power our procedure compares quite favorably with other proposed variants of higher criticism, particularly when streams are short (i.e., small $t$), since in its definition it does not rely on asymptotic approximations.

**Related work.** Our methodology has two properties rarely encountered jointly in related literature; it can be employed in a distribution-agnostic manner, and it is properly calibrated regardless of sample size. We compare our work with related literature pertaining methodologies with either one of these properties.

A distribution-agnostic approach which is properly calibrated regardless of sample size is considered in Wu et al. (2014). The setting is different from ours, but with some important similarities: in an association setting in genomics the higher criticism is used to test if some covariates (gene SNP’s) have non-null association to the outcome variables (traits). A rare-weak assumption on the number and strength of the association is made. Calibration by permutation of the statistic is not the author’s main interest, but is mentioned in the numerical section. The higher criticism statistic itself is computed based on asymptotic approximations. A similar approach has previously been used in applied settings, for example in Sabatti et al. (2009). The review in Donoho and Jin (2015) discusses other works in the genomics domain. Calibration by permutation is mentioned here as well. In a different setting, Arias-Castro et al. (2018) proposes a permutation-based scan statistic for the detection of structured anomalies (e.g., an interval) in a single stream. This test is shown to match the first order performance of the scan statistic calibrated with knowledge of the null distribution.

Various authors have proposed distribution-agnostic methodologies which are asymptotically well-calibrated. In Delaigle and Hall (2009) and Delaigle, Hall, and Jin (2011) methodologies based on the higher criticism are proposed. However, in Delaigle and Hall (2009), the setting considered is significantly distinct from ours. The authors consider observations $Z_j$ (with $j = 1, \ldots, p$) with unknown distribution, and test if each of these have mean equal to the empirical mean of auxiliary observed variables $W_i = (W_{i1}, \ldots, W_{ip})$ with $i = 1, \ldots, n_W$. The distribution of the $W_{ij}$ variables is assumed unknown and equal to the distribution of $Z_j$ under the null hypothesis. In that context, we are thus interested in the distribution of $V_j = Z_j - \frac{1}{n_W} \sum W_{ij}$. The distribution of $V_j$ (under the null) may be approximated based on the repeated observations, and normal approximations naturally arise, and the accuracy of this approximation is characterized through an asymptotic lens. The authors also study the behavior of the higher criticism statistic under dependencies of the component statistics $V_j$, and discuss a classification problem. Conversely, Delaigle, Hall, and Jin (2011) consider a setting closer to ours. The higher criticism statistic is applied to studentized stream means. To compute the higher criticism statistic, the authors estimate the distribution of these studentized stream means through bootstrapping. An asymptotic characterization for the statistic is obtained. For both works there are only asymptotic calibration guarantees.

Moving away from higher criticism, in Arias-Castro and Wang (2017) the problem of detecting sparse heterogeneous mixtures from a nonparametric perspective is considered. The null distribution is assumed to be symmetric and the anomalous observations have a shift in mean, and the proposed tests are calibrated asymptotically. In Zou et al. (2017) a setting similar to ours is considered, where $n$ data streams of length $t$ are observed, and no knowledge on the data distribution is assumed. However, the inference goal is different: instead of hypothesis testing, their goal is to identify the set of anomalous streams. As a performance guarantee, the authors bound the probability of misidentifying the set of anomalous sequences. Their estimator is based on maximum-mean discrepancy and methodologically it is very different from our permutation approach. In Kurt, Yilmaz, and Wang (2020) the focus is on nonparametric changepoint detection in a real-time setting and the authors assume that there is a period of time guaranteed without anomalies. The authors show their decision statistics are asymptotically bounded under the null. Power properties are shown empirically.

Finally, Hall and Jin (2008, 2010) show how the higher criticism statistic behaves under dependencies of the component statistics, assuming distributional knowledge. Our methodology can be applied in ways such that it can deal with some
amount of dependencies, though the dependence must then be modeled and our methodology applied on residuals.

**Organization.** Section 1 introduces and motivates the problem. Section 2 briefly reviews known results for the normal model and motivates the use of a simple quantized variant of the higher criticism statistic. In Section 3 we introduce the class of distributional models we use as a benchmark for our theoretical power analysis and provide lower bounds on the anomalous signal strength for any test to be asymptotically powerful in the context of any one-parameter exponential family. Section 4 proposes a novel permutation higher criticism test, and we show that it is essentially asymptotically optimal in the context of the one-parameter exponential family. We also propose a permutation max test and establish its power properties. In Section 5 we examine the finite-sample performance of our methodology on simulated data, and also apply our methodology to a real dataset. All the proofs are provided in the supplementary materials.

**Notation.** Throughout the article we use standard asymptotic notation. Let \( n \to \infty \), then \( a_n = O(b_n) \) when \( |a_n/b_n| \) is bounded, \( a_n = o(b_n) \) when \( a_n/b_n \to 0 \), and \( a_n = o_2(b_n) \) when \( b_n = o(a_n) \). We also use probabilistic versions: \( a_n = O_P(b_n) \) when \( |a_n/b_n| \) is stochastically bounded\(^1\) and \( a_n = O_P(b_n) \) when \( a_n/b_n \) converges to 0 in probability. Unless otherwise stated, we consider asymptotic behavior with respect to \( n \to \infty \).

2. An Important Case: The Normal Location Model

This section considers a specific location model, which serves both as a benchmark and provides important insights needed for generalizations. These are well known results, but serve as a stepping stone to propose our methodology in the coming sections. Consider the setting where \( F_0 \) is the standard normal distribution with zero mean and unit variance. Under the alternative, there is a set \( S \subset [n] \) indexing the anomalous streams, and the corresponding observations have an elevated mean \( \mu > 0 \) corresponding to the signal strength. Specifically, consider the following hypothesis test:

\[
H_0 : \forall_{i \in [n]} X_{ij} \overset{iid}{\sim} \mathcal{N}(0,1).
\]

\[
H_1 : \exists_{S \subset [n]}: \forall_{i \in S} X_{ij} \overset{iid}{\sim} \mathcal{N}(\mu,1), \text{ and } \forall_{i \notin S} X_{ij} \overset{iid}{\sim} \mathcal{N}(0,1).
\]

A natural question to ask is how large does \( \mu \) need to be to ensure one can distinguish the two hypotheses. In this model it is obvious that the collection of stream means

\[
Y_i(X) \equiv \frac{1}{t} \sum_{j \in [t]} X_{ij}
\]

are jointly sufficient. The setting then reduces to that of the normal means model (Ingster 1997; Donoho and Jin 2004). Note that in those works the alternative hypothesis is a bit different, and often \( X_{ij} \) are assumed to be iid samples from a sparse normal mixture. Nevertheless the asymptotic characterization of the optimal tests can be easily translated to the setting we are considering.

We take an asymptotic point of view. Let \( \psi_n(X) : \mathbb{R}^n \to \{0,1\} \) be a sequence of tests, where the outcome 1 indicates rejection of the null hypothesis. Consider a scenario where, under the alternative hypothesis, the size of \( S \) is \( s \) (i.e., there are exactly \( s \) anomalous streams). The risk of a test is simply the sum of Type I and worst-case Type II error, namely

\[
R(\psi_n) \equiv \mathbb{P}_0(\psi_n(X) \neq 0) + \max_{S:|S|=s} \mathbb{P}_S(\psi_n(X) \neq 1),
\]

where we denote \( \mathbb{P}_S \) the probability under the alternative, with anomalous set \( S \) having signal strength \( \mu \). A sequence of tests is said to be asymptotically powerful if \( R(\psi_n) \to 0 \) as \( n \to \infty \), and it is said to be asymptotically powerless if \( R(\psi_n) \to 1 \).

To present an asymptotic characterization of the above hypothesis testing problem it is convenient to introduce the following parameterization. Let \( \beta \in (0,1) \) be arbitrary (but fixed) and suppose \( S \) has size

\[
s = |S| = \lfloor n^{1-\beta} \rfloor,
\]

where \([z]\) denotes the smallest integer larger or equal to \( z \). With this parameterization, we can identify two main regimes:

- **The dense regime:** \( 0 < \beta \leq 1/2 \). In this scenario a simple test based on \( \sum_{i \in [n]} Y_i(X) \) is optimal. As stated in the beginning we are interested in sparse alternatives, therefore, we do not focus much on the dense regime. Whenever appropriate we make comments on how our results and methods extend to this regime.

- **The sparse regime:** \( 1/2 < \beta < 1 \). In this case a test based on \( \sum_{i \in [n]} Y_i(X) \) will be asymptotically powerless. Let \( r > 0 \) be fixed and parameterize \( \mu \) as

\[
\mu = \sqrt{\frac{2r}{t}} \log(n).
\]

Define the detection boundary

\[
\rho^*(\beta) \equiv \begin{cases} 
\beta - 1/2, & 1/2 < \beta \leq 3/4, \\
(1-\sqrt{1-\beta^2})^2, & 3/4 < \beta < 1.
\end{cases}
\]

It can be shown that, if \( r < \rho^*(\beta) \) all tests are asymptotically powerless (see Ingster (1997) and Theorem 1). Furthermore, if \( r > \rho^*(\beta) \) there are tests (introduced below) that are asymptotically powerful. Note that there are actually two sub-regimes within the sparse regime, namely:

- **The moderately sparse regime:** \( 1/2 < \beta \leq 3/4 \). Here Donoho and Jin (2004) show that the higher criticism test, based on the maximum of a normalized empirical process of the \( (Y_i(X), t \in [n]) \) is asymptotically optimal. In fact, this is so in all regimes. Define the stream \( p \)-values as

\[
p_i \equiv 1 - \Phi \left( \sqrt{t} Y_i(X) \right),
\]

where \( \Phi \) denotes the cumulative distribution function of a standard normal random variable. One variant of the higher criticism rejects the null hypothesis for large values of

\[
\max_{i \in [1:n]} \sqrt{n} \frac{\frac{1}{n} - p_i(0)}{p_i(1) - p_i(0)},
\]

where \( \alpha_0 \in (0,1) \) and \( p_1 \leq \cdots \leq p_n \) are the ordered \( p \)-values.

\(^1\)That is, for any \( \varepsilon > 0 \), there is a \( C_\varepsilon \) and \( n_\varepsilon \) such that \( \forall n > n_\varepsilon \), 
\[ P(|a_n/b_n| > C_\varepsilon) < \varepsilon. \]
The very sparse regime: $3/4 < \beta < 1$. Here Donoho and Jin (2004) show that the test that rejects for large values of $\max_{i \in [n]} Y_i(X)$ achieves the detection boundary. This is not the case in the moderately sparse regime, as shown in Arias-Castro, Candès, and Plan (2011). Expressed in terms of $p$-values, the max test is entirely equivalent to multiple testing with Bonferroni correction, rejection the null hypothesis for small values of $p_i = \min_{i \in [n]} p_i$.

Remark 1. Note that the parameterizations in Equations (4) and (5) can be presented in a more general way using asymptotic notation. In particular, we can simply assume $|S| = n^{1-\beta+o(1)}$ as $n \to \infty$. For the sparse regime, it suffices to assume $\mu = \sqrt{2(r + o(1)) \log(n)/t}$.

### 2.1. A Quantized Higher Criticism Statistic

In this section we show that a quantized higher criticism statistic is also optimal in the normal location model. Quantizing the higher criticism statistic has been done before in different ways—for example, in Arias-Castro, Candès, and Plan (2011), Wu et al. (2014) and others, mainly to facilitate the analysis. We discuss a quantization approach because the adaption and analysis of the higher criticism statistic, when using calibration by permutation, is easier for the quantized form introduced here. In contrast, adaptation to the permutation setting of the original higher criticism statistic in (7) is not straightforward as the computation of the test statistic itself requires knowledge of the null distribution—through the computation of the $p$-values in Equation (6).

We quantize a different variant of higher criticism. This variant was already introduced for analytical purposes in Donoho and Jin (2004). Let $q \geq 0$ and define

$$N_q(X) \equiv \sum_{i \in [n]} \mathbb{I} \left( Y_i(X) \geq \sqrt{2q \log(n)} \right).$$

This is simply the number of streams whose average is larger or equal to $\sqrt{2q/t} \log(n)$ and $\mu$. We refer to the latter value as the $q$-threshold. Define also $p_q$, the probability that a stream mean from the null exceeds that same threshold:

$$p_q = 1 - \Phi \left( \sqrt{2q \log(n)} \right).$$

Clearly, under the null hypothesis $N_q(X)$ is binomial with parameters $n$ and $p_q$, therefore, we can standardize $N_q(X)$ and define the statistic

$$V_q(X) \equiv \frac{N_q(X) - np_q}{\sqrt{np_q(1-p_q)}}.$$  

The $V_q(X)$ statistic, when maximized over $q \in [0, \infty)$, is closely related to the higher criticism statistic (see Remark 4).

The following result summarizes several analytical steps in Donoho and Jin (2004).

**Proposition 1** (Donoho and Jin (2004)). Let $q \in [0, \infty)$ be fixed but arbitrary, and $h_n > 0$ with $h_n \to \infty$. Then $P_{\theta}(V_q(X) \geq h_n) \leq \frac{1}{h_n^2} \to 0$. Furthermore, for $S$ as in (4), when $1/2 < \beta < 1$ (sparse regime), $h_n = n^{o(1)}$, and $\mu = \sqrt{2(r/t) \log(n)}$ we have:

- If $r > (1 - \sqrt{1-\beta})^2$ then $P_S(V_1(X) \geq h_n) \to 1$.
- If $r < 1/4$ and $r > \beta - 1/2$ then $P_S(V_q(X) \geq h_n) \to 1$.

The proof of Theorem 3, which is critical for our main result, hinges crucially in showing that important quantities inside the proof of Proposition 1 have the same asymptotic characterization when considering calibration by permutation. The heart of the proof of Proposition 1 is the realization that, under the alternative, $N_q(X)$ is the sum of two independent binomial random variables; $N_q(X) \sim Bin(n - s, p_q) + Bin(s, v_q)$, where

$$p_q = n^{-q+o(1)}$$

and

$$v_q = \begin{cases} 
\frac{1}{2} & \text{if } r < q \\
1 - n^{-\sqrt{2-\sqrt{q}}+o(1)} & \text{if } r = q \\
1 - n^{-\sqrt{2-\sqrt{q}}+o(1)} & \text{if } r > q 
\end{cases}.$$ 

We will show a similar result in our setting, implying the result in this proposition can be used to complete the proof of Theorem 3.

**Remark 2.** In the dense regime, it is also possible to show that, when $\mu = \sqrt{t/r}(1-q)^{1/2}$ with $r > \beta$, we have $P_S(V_q(X) \geq h_n) \to 1$. This implies the higher criticism statistic is also optimal in that regime under the normal model.

**Proposition 1** shows that, depending on the combination of sparsity $|S|$ and signal strength $\mu$ we might want to consider a test based on $V_q(X)$ for different values of $q$. Interestingly, the value $q = 1$ suffices for the very sparse regime. For the moderately sparse regime the situation is more intricate, and it turns out the choice $q = 4r$ is essentially optimal (unless $r \geq 1/4$, as then $q = 1$ suffices). Obviously we do not know $r$, so the only way we can capitalize on this knowledge is to scan over a suitable selection of values for $q$, that hopefully include (or approximate) the optimal value. To keep the analysis simple and avoid the use of advanced tools in empirical processes we scan only over a coarse subset of values for $q$. The following result is an almost immediate consequence of Proposition 1.

**Proposition 2** (Quantized HC). Let $Q \subseteq [0,1]$ be the set

$$Q = \left\{ 0, \frac{1}{k_n}, \frac{2}{k_n}, \ldots, 1 \right\},$$

where $k_n \to \infty$ and $k_n = n^{o(1)}$. Define the test statistic

$$T(X) = \max_{q \in Q} V_q(X).$$

Finally, let $h_n = \omega(\sqrt{k_n})$ and $h_n = n^{o(1)}$, and consider the test that rejects the null hypothesis when $T(X) \geq h_n$. This test has vanishing Type I error. When $\beta > 1/2$ and $\mu = \sqrt{2(r/t) \log(n)}$ this test is asymptotically powerful provided $r > \rho^*(\beta)$.

The proof follows almost immediately by using a union of events bound over the grid elements. Under the alternative, the only intricate case is the moderately sparse regime. However, since the grid $Q$ with $k_n + 1$ elements becomes denser as $n$ grows, it will contain a value $q$ that differs at most $1/k_n$ from the optimal value $4r$, which suffices for our result.
Remark 3. For the dense regime, one can prove that, if \( \mu = (1/\sqrt{t})n^{r^{-1/2}} \) the test in Proposition 2 is asymptotically powerful provided \( r > \beta \). This follows immediately using the results from Proposition 1, since the optimal value 0 is guaranteed to be in the grid.

Remark 4. The \( V_q(X) \) statistic is closely related to the more familiar higher criticism statistic as in (7) as

\[
\sup_{q \in [0, \infty)} \left\{ V_q(X) \right\} = \max_{i \in \{1, \ldots, n\}} \left\{ \sqrt{n} \frac{\frac{1}{n} - p_i}{\sqrt{p_i(1 - p_i)}} \right\},
\]

where \( i^+ \) is the largest value of \( i \) for which \( p_i < 1/2 \).

3. Generalization to Exponential Families

The results of the previous section were specific for the normal location model. Nevertheless one can envision somewhat natural extensions of the approach and test statistics to the more general context of the exponential family. When studying distribution-free tests, it is customary to compare them with parametric tests (Hettmansperger 1984). As in Arias-Castro, Candès, and Durand (2011) and Arias-Castro et al. (2018), we consider one-parameter exponential models (in natural form) as parametric models, as these play an important role in the related literature (Donoho and Jin 2004).

One-parameter exponential models in natural form. Let \( F_0 \) be a probability distribution on the real line, such that all moments are finite. Let \( \mu_0 \) and \( \sigma_0^2 \) denote, respectively, the mean and variance of \( F_0 \). The distribution might be continuous, discrete or a combination of the two. In the exponential model there is a parameter \( \theta_i \) associated with each \( i \in [n] \), and the distribution \( F_1 \equiv F_{\theta_1} \) is defined through its density \( f_{\theta_1} \) with respect to \( F_0 \): for \( \theta \in [0, \theta_\alpha) \), define \( f_{\theta}(x) = \exp(\theta x - \log \phi_0(\theta)) \), where \( \phi_0(\theta) = \int e^{\theta x} dF_0(x) \) and \( \theta_\alpha = \sup\{\theta > 0 : \phi_0(\theta) < \infty\} \), assumed to be strictly positive (and possibly infinite). In other words, \( f_\theta \) denotes the Radon-Nykodym derivative of \( F_{\theta} \) with respect to \( F_0 \). Since a natural exponential family has the monotone likelihood ratio property it follows that \( F_0 \) is stochastically increasing in \( \theta \) (Lehmann and Romano 2005, Lemma 3.4.2). Particular cases of this model are used in a variety of settings. For example, in many signal and image processing applications, a model with \( F_0 \) corresponding to normal distribution with mean \( \theta \) and a fixed variance is common. In syndromic surveillance (Kulldorf et al. 2005), a model with \( F_0 \) corresponding to a Poisson distribution is popular. Bernoulli models (Walther 2010) are also a particular case of this class, with \( F_0 \) corresponding to a Bernoulli distribution.

Let \( X_{ij} \) be independent observations, where all observations within stream \( i \) are iid with distribution function \( F_{\theta_0} \), as defined above. Under the null hypothesis we have \( \theta_0 = 0 \) for all \( i \). Under the alternative, there is a subset of streams for which \( \theta_1 > 0 \). Our hypothesis testing problem is therefore given by

\[
H_0 : \quad \forall_{i \in [n]} \quad \theta_i = 0 ,
H_1 : \quad \exists_{S \subseteq [n]} : \forall_{i \in S} \quad \theta_i > 0 \quad \text{and} \quad \forall_{i \notin S} \quad \theta_i = 0 . \quad (11)
\]

Keeping the same minimax stance as before, in the remainder, our analysis assumes the anomalous observations have the same distribution so \( \theta_i > 0 \) for \( i \in S \). We keep the parameterization of the anomalous set as in (4), namely \( |S| = n^{1-\beta} \) for \( \beta \in (0, 1) \), and use the parameterization \( \theta = \sqrt{2r/(\sigma_0^2) \log n} \).

Like in the normal model, a detection boundary defines the region in the \((r, \beta)\) plane where any test will be asymptotically powerless:

**Theorem 1.** Refer to the hypothesis testing problem in (11) and the parameterization \( |S| = n^{1-\beta} \) with \( \beta \in (1/2, 1) \) and \( \theta = \sqrt{2r/(\sigma_0^2) \log n} \) with \( r > 0 \). Provided \( t = o(\log^3(n)) \) any test will be asymptotically powerless when \( r < \rho^*(\beta) \).

Remark 5. Perhaps surprisingly, if the stream lengths are large enough, the detection boundary in this broad family precisely matches the boundary of the normal location model discussed in Section 2. The reason is twofold: the collection of stream means is (in our exponential family) a sufficient statistic, and the stream lengths are (with respect to the tails of the exponential family) not too short. This combination results in a setting where an optimal test considers the collection of stream means which are jointly normal enough, such that one can asymptotically do no better than if the stream means were exactly normal. If the stream means are not sufficiently normal, the detection boundary may differ from \( \rho^*(\beta) \) (e.g., Donoho and Jin 2004 present a characterization for the Subbotin family of distributions).

In the next section, we develop a permutation-based test that is able to attain optimal asymptotic performance, giving a converse to the statement of Theorem 1.

4. Calibration by Permutation

The quantized higher criticism statistic proposed in Section 2 is asymptotically optimal in the normal location model, but to compute it and calibrate it (i.e., compute a test \( p \)-value) one needs to assume knowledge of the null distribution. With that knowledge, one could simply calibrate the test by Monte Carlo simulation. The goal in this section is to construct a test (and test statistic) that can be computed and properly calibrated without the knowledge of the null distribution. A natural and appealing idea is to calibrate these tests by permutation. The idea of using permutations is not new in the context of comparing multiple groups (e.g., consider the classical Kruskal–Wallis test or any test based on runs). However, such approaches cannot deal with sparse alternatives, and as in the setting we are considering (particularly when \( \beta > 1/2 \)), note that the validity of calibration by permutation critically depends on the independence between our observations under the null. Nonetheless, if one is willing to model the dependencies, then our methodology could be applied on the model residuals to draw inference. We exemplify this in Section 2 in the supplementary materials.

Note that, in the sparse regime we are most interested in, the distribution of the permutation stream means should intuitively follow the true null distribution quite closely. Indeed, in the sparse regime it is well known that the small fraction of non-null signals get “washed-out” in the overall mean (Donoho and Jin 2004). Heuristically one could similarly argue here that all permutation streams consist of mostly null observations, such
that the permutation stream means "washes-out" the nonnull signals. That being said, it is crucial to thoroughly quantify the extent of the contamination and account for dependencies induced by the permutation process as well. The dense regime is different, and one likely requires much longer streams (e.g., \( t \) growing linearly with \( n \)) to obtain asymptotic power characteristics. Nevertheless, in the dense regime it may be more fruitful to consider instead other statistics to calibrate by permutation (like the Fisher's combined probability test (Fisher 1934, sec. 21.1) since it is known the higher criticism statistic performs poorly in finite samples even with distribution knowledge. These research directions are out of the scope of this article and remains an interesting avenue for future work.

4.1. Permutation-based Max Test

As can be expected, using calibration by permutation to calibrate a test based on the value of the maximum stream mean yields a test that cannot be optimal for all values of \( \beta \). Nevertheless it is insightful to consider this option for its simplicity.

The max test rejects the null hypothesis for large values of

\[
\max_{i \in [n]} Y_i(X), \quad \text{where} \quad Y_i(X) = \frac{1}{t} \sum_{j=1}^{t} X_{ij}.
\]

Let \( \Pi \) denote the set of permutations of \(((i,j) : i \in [n], j \in [t])\). Note that there are \((nt)!\) distinct permutations. Let \( \pi \in \Pi \) and \( X^\pi \equiv (X_{\pi(i,j)}, i \in [n], j \in [t]) \). In words, \( X^\pi \) corresponds to the permuted data over both streams and time. Calibration by permutation amounts to computing a \( p \)-value as

\[
P_{\text{max-perm}}(X) = \frac{1}{(nt)!} \left| \left\{ \pi \in \Pi : \max_{i} \left\{ Y_i(X^\pi) \right\} \geq \max_{i} \{ Y_i(X) \} \right\} \right|.
\]

(12)

In words, this is the proportion of times the value of the test statistic computed with permuted data is at least as large as the value of the test statistic computed on the original data. Although the computation of the \( p \)-value might seem computationally prohibitive, we can approximate it accurately by considering instead a uniform sample from the set of permutations. The statistical properties of this test can be characterized as follows:

**Theorem 2 (Permutation max test).** Let \( \alpha \in (0,1) \) be arbitrary and refer to the hypothesis testing problem in (11). Let \( \theta = \sqrt{2r/(\sigma^2 t) \log n} \) with \( r > 0 \). Consider the test that rejects \( H_0 \) if \( P_{\text{max-perm}}(X) \leq \alpha \). This test has level at most \( \alpha \). Furthermore let \( t = \omega(\log^3(n)) \) and \( \beta > 0 \). This test has power converging to one provided \( r > (1 - \sqrt{T - \beta})^2 \).

This theorem shows the max test calibrated by permutation is only guaranteed to be optimal when \( \beta \geq 3/4 \), as expected. It is important to note the stream length \( t \) starts playing a more prominent role when considering calibration by permutation. Clearly, if \( t = 1 \) it is impossible to calibrate such a test by permutation, as the test statistic is actually invariant under permutations in that case. One expects that, for very small values of \( t \) there might therefore be a loss of power in comparison to an oracle test calibrated with knowledge of the null distribution. The requirement for \( t \) in the above theorem is needed to ensure the individual stream means have approximately Gaussian tails. For exponential tails this scaling seems necessary. We conjecture that if the tail of \( F_0 \) is sub-Gaussian the requirement can be relaxed to \( t = \omega(\log^3(n)) \) and if it has bounded support \( t = \omega(\log n) \) will suffice.

It might seem the above theorem should follow directly from the results in Arias-Castro et al. (2018) as one can interpret the max test as a scan test over intervals of length \( t \) within \( nt \) observations, but a more careful analysis is needed to get a sharp characterization of the detection boundary.

4.2. Permutation-based Higher Criticism Test

Although the simplicity of the max test is appealing, the test is far from optimal in the moderately sparse regime (or the dense regime for that matter). Therefore, we turn our attention to the possibility of calibrating the higher criticism statistic by permutation. Note that the statistic in Proposition 2 critically depends on the location and scale parameters in our normal setting, and since the mean and variance of the null distribution are unspecified in the general setting (11), this needs to be taken into account. For simplicity, we first introduce a statistic dependent on the null mean and variance, and motivate later how our methodology can be used without such knowledge.

4.2.1. Known First Two Null Moments

Assuming that the first two null moments are known, we define:

\[
\tilde{N}_q(X) \equiv \sum_{i \in [n]} \mathbb{1} \left\{ Y_i(X) - \mu_0 \geq \sqrt{2\sigma^2 q/(t \log(n))} \right\}, \quad (13)
\]

where \( Y_i(X) \) is as before. We must also adapt the normalizing term \( p_q \) which was previously based on the normal setting. In principle, the computation of this probability requires full knowledge of \( F_0 \). Instead, we obtain a surrogate value for \( p_q \) by permutation as follows:

\[
\tilde{P}_q(X) \equiv \frac{1}{(nt)!} \left| \left\{ \pi \in \Pi : \max_{i} \{ Y_i(X^\pi) \} \geq \max_{i} \{ Y_i(X) \} \right\} \right|.
\]

(14)

Finally, define the counterpart of \( V_q(X) \) as

\[
\tilde{V}_q(X) \equiv \frac{N_q(X) - np_q(X)}{\sqrt{np_q(X)(1 - p_q(X))}}, \quad (15)
\]

where we take the convention that \( 0/0 = 0 \). For this statistic, we have the following result:

**Theorem 3 (Permutation-based higher criticism test with known null mean and variance).** Let \( Q \subseteq [0,1] \) be the set defined in (10) with \( k_n \rightarrow \infty \) and \( |Q| = k_n = n^{o(1)} \). Define the test statistic:

\[
\tilde{T}(X) \equiv \max_{q \in Q} \tilde{V}_q(X). \quad (16)
\]

Finally, define the higher criticism permutation \( p \)-value for known null mean and variance as

\[
\tilde{P}_{\text{perm-hc}}(X) = \frac{1}{(nt)!} \left| \left\{ \pi \in \Pi : \tilde{T}(X^\pi) \geq \tilde{T}(X) \right\} \right|.
\]
Let $\alpha \in (0,1)$ be arbitrary and refer to the hypothesis testing problem in (11). Consider the test that rejects $H_0$ if $\tilde{P}_{\text{perm-hc}}(X) \leq \alpha$. This test has level at most $\alpha$. Furthermore, let $\beta > 1/2$ and $\theta = \sqrt{2r/(\sigma_0^2 t) \log(n)}$ with $r > 0$, $t = \omega(\log^3(n))$ and $t = n^{\omega(1)}$. This test has power converging to one provided $r > \rho^*(\beta)$.

Under the conditions of the theorem we see that the proposed test is optimal for all values $\beta > 1/2$, unlike the max test characterized in Theorem 2. It is important to reflect on what impractical as it requires knowledge of the mean and variance of the null distribution. However, this can be viewed as stating in Corollary 1.

The result in Theorem 3 is theoretically appealing, but somewhat impractical as it requires knowledge of the mean and variance of the null distribution. However, this can be viewed as an artifact of scanning the values of $\tilde{V}_t(X)$ over too small a grid $Q$. To ensure our test has the right power characterization, all that is needed is to ensure the optimal value of $q$ is adequately approximated in the grid $Q$. It turns out it is possible to construct a grid that can account for lack of knowledge of the null mean and variance, while still having adequate approximation properties.

To streamline the presentation we restate our test statistic unencumbered by the $q$-threshold parameterization as

$$\tilde{N}_t = \sum_{i \in [n]} 1 \{ Y_i(X) \geq \tau \},$$

$$\tilde{P}_t(X) \equiv \frac{1}{(nt)!} \left| \left\{ \pi \in \Pi : Y_1(X^\pi) \geq \tau \right\} \right|,$$

and

$$\tilde{V}_t(X) \equiv \frac{\tilde{N}_t(X) - n\tilde{P}_t(X)}{\sqrt{n\tilde{P}_t(X)(1 - \tilde{P}_t(X))}}.$$

Consider the grid

$$R \equiv \left\{ i + \sqrt{\frac{2r^2}{t}} \log(n) : i \in \left\{ \frac{k}{\sqrt{t}} \right\}_{k=\cdot}^{1/\sqrt{t}}, j \in \left\{ \frac{k}{\sqrt{\log(n)}} \right\}_{k=0}^{\infty} \right\}. \quad (19)$$

Note the expanding nature of this grid; as $n$ increases, the minimum and maximum gridpoint expand. Furthermore, due to the increasing gridsize, the grid elements approximate points within the grid elements with increasing accuracy. This grid is therefore sure to contain or approximate the asymptotically optimal threshold as $n$ increases, while the computation of the statistic $\max_{r \in R}(\tilde{V}_t(X))$ does not require any knowledge of the null distribution. There are other grid constructions that share similar properties, and for which our results hold, but we chose this for concreteness. We can state similar results for this statistic as in Theorem 3:

**Theorem 4.** Let $R$ be the set defined in (19). Define the test statistic:

$$\tilde{T}(X) \equiv \max_{r \in R} \tilde{V}_t(X). \quad (20)$$

Define the higher criticism permutation $p$-value as

$$\tilde{P}_{\text{perm-hc}}(X) \equiv \frac{1}{(nt)!} \left| \left\{ \pi \in \Pi : \tilde{T}(X^\pi) \geq \tilde{T}(X) \right\} \right|.$$

Let $\alpha \in (0,1)$ be arbitrary and refer to the hypothesis testing problem in (11). Consider the test that rejects $H_0$ if $\tilde{P}_{\text{perm-hc}}(X) \leq \alpha$. This test has level at most $\alpha$. Furthermore, let $\beta > 1/2$ and $\theta = \sqrt{2r/(\sigma_0^2 t) \log(n)}$ with $r > 0$, $t = \omega(\log^3(n))$ and $t = n^{\omega(1)}$. This test has power converging to one provided $r > \rho^*(\beta)$.

Note that no further assumptions are needed to attain the same results of Theorem 3. Importantly, the assumption $t = n^{o(1)}$ now plays two different roles: on the one hand it is needed for technical reasons, as explained earlier. On the other hand, it is required to ensure the grid is small enough (namely of size $n^{o(1)}$). In any case, the “short” streams requirement can be avoided by a simple modification of the methodology, leading to the following result:

**Corollary 1.** Consider the hypothesis testing problem in (11). Let $\alpha \in (0,1)$. There is a test not requiring knowledge of $F_0$ that has level at most $\alpha$. Furthermore it has power converging to one under the same conditions as in Theorem 4 except that $t$ may be arbitrarily large.

The corollary is proved in the supplementary materials. The modified methodology in Corollary 1 simply maps the original dataset to a new dataset by aggregating small groups of observations within each stream. This effectively creates a scenario with short streams. The methodology described in Section 4.2.2 can then be used and its performance characterized by using Theorem 4, leading to the stated result.
4.2.3. A Practical Alternative

Instead of considering the grid in (19), a more natural idea would be to approximate the null mean and variance by overall sample estimates. This would correspond to maximizing in (20) over the following data dependent grid:

\[
\hat{R}(X) \equiv \{ \hat{X} + \frac{2\sigma^2 q}{t} \log(n) : q \in Q \}, \tag{21}
\]

with Q the grid as in (10), and

\[
\hat{X} \equiv \frac{1}{nt} \sum_{i \in [n], j \in [t]} X_{ij}, \quad \sigma^2 \equiv \frac{1}{nt} \sum_{i \in [n], j \in [t]} (X_{ij} - \hat{X})^2. \tag{22}
\]

Given the efficiency of the sample estimates above, as well as the relatively coarse approximation required to the optimal gridpoint for our test in Theorem 4 to work, there is little reason to suspect the data dependent grid approach would incur any loss of power. However, proving this rigorously greatly and nontrivially complicates the analysis; due to the grid being now dependent on the data, there is a need to carefully address the dependencies between the summands of \( \hat{N}_t(X) \), as well as carefully characterize the extra source of randomness introduced in \( \hat{P}_t(X) \). Such an analysis seems to require very refined integro-local results that, to the best of our knowledge, are currently not existing in the literature. Ultimately, the requirement for an approximation for the sample mean and variance stem from the discreteness of the grid Q (so one can use a relatively crude union bound). Dropping the discreteness of the grid would require advanced results from empirical process theory that are beyond the scope of this article.

**Remark 6.** It has been suggested in related works (Arias-Castro, Candès, and Durand 2011; Wu et al. 2014) that scanning over a discrete grid for this type of statistic is merely a proof artifact, and the results are also valid when q in (16) is optimized over the continuous interval \([0, \infty)\), which would incorporate the true value of the null statistics automatically. We have conducted extensive numerical experiments regarding the choice of grid and observed no deteriorating performance for extremely fine grid choices. However, calibration by permutation without a discrete grid leads to an optimization problem that is very demanding computationally. So, from a practical standpoint the discretization is quite advantageous as well.

**Remark 7.** Instead of computing \( \tilde{P}_q(X) \) by permutation, one might consider an asymptotic normal approximation as used in Sabatti et al. (2009) and Wu et al. (2014). However, our numerical experiments (see Section 5.5) indicate the permutation-based approach in (14) leads to tests with higher power.

**Remark 8.** For calibration by permutation of the \( \tilde{T}(X) \) statistic, we encounter a minor complication; the test statistic itself depends on a term \( \tilde{P}_q(X) \) that must be computed by sampling permutations as well. Recognizing that \( \tilde{P}_q(X) \) is invariant under permutations prevents the procedure from being computationally prohibitive; details are deferred to Section 5.1.

**Remark 9.** For the dense case (i.e., \( \beta \leq 1/2 \)) our preliminary results indicate that when \( \theta = (\sigma_0/\sqrt{t})n^{-1/2} \) with \( r > 0 \) and \( t = o(n) \) the permutation higher criticism test has power converging to one provided \( r > \beta \). However, the analysis of this regime is quite different and out of the scope of this article.

5. Experimental Results

This section thoroughly explores the potential of the proposed procedure. We begin by making some important computational remarks, followed by results on simulated data, and concluding with the application of the methodology to a real dataset.

5.1. Computational Complexity

The p-value of the permutation test in Theorem 4 is calculated by considering all possible data permutations. Naturally, this becomes prohibitive even for moderately large datasets, and instead a sample of the permutation distribution can be used to estimate \( \tilde{P}_{\text{perm-hc}}(X) \).

At first glance computation and calibration of this test statistic might still seem prohibitive: for each permutation \( \pi \) taken from \( \Pi \), one computes \( \tilde{T}(X^{\pi}) \), which requires us to compute \( \tilde{P}_q(X^{\pi}) \). Naturally, we can estimate this quantity in a similar fashion using a random set of permutations. However, this would mean that every permutation sample \( \pi \) used to compute \( \tilde{T}(X^{\pi}) \) requires us to take a new set of permutation samples, which is disastrous computationally. However, note that \( \tilde{P}_q(X) \) is invariant under permutations, namely \( \tilde{P}_q(X) = \tilde{P}_q(X^\pi) \). Thus, we can estimate \( \tilde{P}_q(X) \) using a random sample of permutations, and then use these estimates to compute \( \tilde{T}(X^\pi) \). Moreover, numerical experiments indicate we can use the same set of permutations to both estimate \( \tilde{P}_q(X) \) and compute \( \tilde{T}(X) \), without loss of performance or raising the Type I error. This approach has therefore been used in the ensuing numerical experiments.

Finally, note that our test statistic is invariant under time-permutations within individual streams. So, although the size of \( \Pi \) is \( (nt)! \), there is an equivalence class of permutations that is much smaller, namely of size \( \approx (nt!)/(t!)^n \). This further motivates why a Monte Carlo approach works well even with a relatively small number of permutations.

5.2. Simulation Setting

For our experimental results, we consider two different models: the normal location model, discussed in Section 2 which is commonly used as a benchmark, and a model where the underlying distributions are exponential. In the second model, the null distribution \( F_0 \) is exponential with mean \( 1/\lambda_0 \) (where we chose \( \lambda_0 = 1.5 \)). Following the parameterization set up in Section 3 the observations in the anomalous streams are exponentially distributed with mean \( 1/(\lambda_0 - \theta) \), where \( \theta \in [0, \theta_0) \) where \( \theta_0 = \lambda_0 \). This second setting allows us to explore the influence of heavier tails. In fact, the exponential tails are the heaviest tails for which we provide asymptotic guarantees.

In all simulations considered we assume the observations in the anomalous streams have distribution \( F_0 \), where

\[
\theta_t = \tau \sqrt{\frac{2 \sigma^2 (\beta) \log(n)}{\sigma_0^2 t}}, \tag{23}
\]
with $\tau \geq 0$. In words, we consider a signal strength that is a multiple of the asymptotically minimal signal strength $\rho^*(\beta)$ needed for detection. Asymptotically we know that when $\tau < 1$ all tests are powerless, and when $\tau > 1$ the permutation higher criticism test is powerful. In all simulations, test are calibrated at a significance level of 5%.

### 5.3. Grid Choice

Although the approach characterized in Theorem 4 is asymptotically sensible, it is somewhat cumbersome in practice and can potentially ignore important information in nonasymptotic settings. The data dependent grid $\tilde{R}(X)$ in (21) is easier to employ, and extensive numerical experiments show there is no observable performance loss when it is used. However, even this data dependent grid may leave out important information, if the global maximum of $\tilde{V}_r(X)$ is attained for $\tau > \bar{X} + \sqrt{2\sigma_X^2 \log(n)}/t$. To safeguard the approach against this issue as well, one can extend the data dependent grid $\tilde{R}(X)$, such that the grid contains a sufficiently large point $\tau^*$ for which $N_{\tau^*}(X) = 0$. A specific choice is

$$
\tilde{R}(X) \equiv \left\{ \bar{X} + \sqrt{2\sigma_X^2 q \log(n)}/t : q \in \tilde{Q}(X) \right\}, \quad \text{with}
$$

$$
\tilde{Q}(X) \equiv \left\{ \frac{1}{d_n} \frac{2}{d_n}, \ldots, \left( \frac{\max_{ij} |X_{ij}| - \bar{X}}{\sigma_X} \right)^2 \frac{t}{2 \log(n)} \right\}, \quad (24)
$$

with $d_n \to \infty$. Note that Lemma 4 in the supplementary materials ensures that $O(d_n \log^2(nt)/\log(n))$, and therefore, provided $d_n = n^{o(1)}$, the grid-size is still small enough to ensure a union-bound argument is valid.

Due to the data dependency of this grid a statement similar to Theorem 4 is currently unavailable. Extensive numerical experiments show there is no reason to suspect the data dependent grid performs poorer than its counterpart in Theorem 4, and the shown numerical experiments therefore use the extended data dependent grid in (24).

To recommend the reader a choice for $d_n$, we simulate data from both models outlined in Section 5.2 and compare the performance of the tests for different choices of $d_n$ as in (24). We use $n = 10^3$ streams of length $t = \lceil (\log n)^2 \rceil = 48$. For the experiments shown we took $|S| = 12$, (i.e., $\beta \approx 0.64$, moderately sparse regime), but through extensive experimentation we observed our conclusions are not sensitive to the sparsity. The results are given in Figure 1. We observe that a spacing of $1/d_n = 1/\log(n)$ is a reasonable choice. In all remaining experiments, we therefore extend the grid as in (24) with $d_n = \log(n)$.

### 5.4. Comparison with an Oracle Test

We now compare our methodology with the oracle test making full use of the null distribution. The oracle test statistic is defined in (9) for the normal model. For the exponential model the stream averages are Gamma-distributed, and so one uses the quantiles of that distribution to compute $p_\Phi$ instead. The oracle-test is then calibrated by Monte Carlo simulation. For this calibration, $10^4$ simulations were used. In the presentation we use HC-permutation and HC-oracle to refer to the two tests. We compare the oracle test with our methodology with respect to two aspects:

- **Stream lengths.** For $t = 1$ calibration by permutation leads to a powerless test. Therefore, we expect the performance gap between an oracle and the proposed test to be large for short streams. However, we observe the finite-sample performance is already quite comparable to the oracle test even for very short streams.

- **Performance loss.** Our test makes no use of knowledge of the null distribution. While asymptotically equivalent to first order, we expect some performance loss in finite samples compared to the oracle.

We use $n = 10^3$ streams in our simulations. To study the influence of stream length we considered $|S| = 12$ (i.e., $\beta \approx 0.64$, moderately sparse regime) but the results are qualitatively identical for other choices of sparsity. The signal strength was chosen slightly above the asymptotic detectability threshold, but ensuring the difference in performance was highlighted. For the exponential model very short streams would lead to a value $\theta \geq \lambda_0$ in the parameterization in (23), so the shortest stream considered is $t = 4$. Figure 2 shows that the performance is comparable with respect to the oracle tests even at very short lengths, suggesting that finite-sample performance is much more forgiving than what our theoretical characterization might suggest, as there is only a very small loss of power in comparison to an oracle. Naturally, the permutation test has no power for streams of length $t = 1$. Note that the strength of the anomalous signal is decreasing in $t$, such that in the normal model the oracle test performs identically for any value of $t$. In contrast, for the exponential model, the difference in null and alternative means is $\theta/(\lambda_0 (\lambda_0 - \theta))$ and is larger for smaller streams, so all tests perform worse at larger values of $t$.

We now turn our attention to the power characteristics of the proposed tests. We fix $n = 10^3$ and $t = \lceil (\log n)^2 \rceil = 48$ and consider $|S| = 12$ (i.e., $\beta \approx 0.64$, moderately sparse regime), and $|S| = 3$ (i.e., $\beta \approx 0.84$, very sparse regime). The results are plotted in Figure 3, and we see that the proposed permutation higher criticism test has only minimal loss of power when compared to the oracle test.

### 5.5. Comparison with Approximation Methods

As mentioned earlier, several authors have considered practical variants of higher criticism where the permutation stream means are assumed to be approximately normal, and therefore the test statistic is computed based on a normal approximation (e.g., see Sabatti et al. (2009) and Wu et al. (2014) and references therein). In essence, instead of computing (or, more appropriately, estimating) $\tilde{Y}(X)$ as we do in Equation (17), one could instead compute:

$$
P_\Phi = 1 - \Phi(\sqrt{2q \log(n)}),
$$

which arises from the assumption that the standardized permutation stream means $Y(X^\tau)$ are approximately standard normal. The statistic can then be computed like before, and
Figure 1. Comparison of power of our permutation-based higher criticism test under different choices of $d_n$ as in (24). The power of the test is depicted as a function of a $\tau$ in the parameterization in (23). We took $n = 10^3$ streams with length $t = 48$ and set $|S| = 12$. For each test $10^3$ permutations were used. Each level was repeated $10^3$ times, resulting in the 95% confidence bars depicted.

Figure 2. Comparison of power of our permutation-based higher criticism test and the oracle test, as a function of a stream length $t$. We used the parameterization in (23) with a fixed level for $\tau$. Note that the strength of the anomalous signal is decreasing in $t$ as discussed in the text. For the exponential model, the parameterization of the alternative is undefined for $t \leq 3$, so the $t$-axis starts at $t = 4$. We took $n = 10^3$ streams and set $|S| = 12$. For the permutation tests, $10^3$ permutations were used. For the oracle test, the Monte Carlo calibration is based on $10^4$ samples. Each level was repeated $10^3$ times, resulting in the 95% confidence bars depicted.

Figure 4 depicts the results. We see that there is a clear improvement in finite-sample performance when computing $\hat{P}_q(X)$ by permutation, as opposed to the normal approximation. As expected, the difference in power decreases as the streams get longer, since the normal approximation becomes more accurate. In any case, the approximation $P_q^\Phi$ underestimates the true value of $p_q$ when $q$ is large, as the tail of a Gamma distribution is much heavier than that of a normal distribution. Since $\hat{P}_q(X)$ does not rely on asymptotic approximations it more accurately approximates $p_q$, resulting in a more powerful test. Extensive experimentation showed that the permutation-based approach appears to be always superior when the underlying model is not normal.

5.6. An Analysis of Content Uniformity of a Batch-Produced Drug Product

To showcase a possible application of our methodology, we consider a retrospective monitoring of the content uniformity of an active ingredient for a batch-produced drug product. Pharmaceutical companies are required to report on the quality of the production batches in an annual product quality review in line with good manufacturing practice. One aspect of this
Figure 3. Comparison of power of our permutation-based higher criticism test and the oracle test, as a function of a multiplicative factor \( \tau \) in the parameterization in (23). We took \( n = 10^3 \) streams with length \( t = 48 \). For the permutation tests, \( 10^3 \) permutations were used. For the oracle test, the Monte Carlo calibration is based on \( 10^4 \) samples. Each level was repeated \( 10^3 \) times, resulting in the 95% confidence bars depicted.

Figure 4. Comparison of power of our permutation-based higher criticism test and a variant using asymptotic approximations. The power of the test is depicted as a function of \( \tau \) in the parameterization in (23). We took \( n = 10^2 \) streams and set \(|S| = 12\). For each test \( 10^3 \) permutations were used. Each level was repeated \( 10^3 \) times, giving rise to the 95% confidence bars depicted.
product quality review is the identification of potential production issues, which may be identified by a proper anomaly detection approach. Batch production processes are well-controlled due to stringent quality requirements, but slight variability in the active ingredient within batches of the product may be expected. In the absence of production anomalies, this variability should be homogeneous across batches. Therefore, under nominal settings, batch-to-batch variation should come from the sampling variation within batches. Our methodology is therefore suitable to detect potential anomalies during the production process.

We consider data containing 242 production batches, each containing 10 measurements of the active ingredient, given as a percentage of the target concentration. Details pertaining to the drug or the active ingredient are not reported here due to the confidential nature of the data. The anonymized data is published as supplementary materials. A histogram of the average active ingredient concentration per batch is given in Figure 5(a), and a histogram of the sample standard deviation of the active ingredient concentration per batch is given in Figure 5(b).

Clearly, the batch with average concentration exceeding 106%, corresponding to batch number 64, is remarkably large in light of the other observations, and a powerful test is unnecessary to warrant further investigation of this batch and the production process. Unexpectedly, application of our methodology on this data results in a p-value of 0. There fore, it is more interesting to see if this clear anomaly is solely responsible for the detection, or if other signs of anomalies remain in the data after clear outliers are removed.

We may formalize the notion of “clear” outliers by using the max test statistic as described in Theorem 2. We use the 95% quantile of the permutation maximum batch average active ingredient distribution, and identify all batches with average active ingredient exceeding this quantile as “clear” outliers. For our data, this 95% quantile corresponds to an active ingredient concentration of 103.25%, which leads to identifying batch 64 as a “clear” outlier, but also two other batches with large averages; batch 68 and 242.

Now, the application of our methodology on the remaining 239 batches leads to a p-value of 0.027, indicating that, after removal of “clear” outliers, there are still signs of anomalies present in the data. Note that the data is reasonably normal, and the normal approximation alternative as proposed in Section 5.5 leads to a p-value of 0.029; indicating that our test may not lead to much power loss compared to normality-based methods when such assumptions are reasonable. Application of the max test as described in Theorem 2 leads to a p-value of 0.095, showcasing the benefit of using the higher criticism statistic in this context.

When the p-values of our test are small as above, one would ideally like to subsequently identify the anomalous batches. This requires significantly stronger signals than needed for detection alone. In Zou et al. (2017) a method for identification is given, with a bound on the probability of misidentifying the set of anomalous streams. Alternatively, in our context, a natural approach would be to compute p-values of each batch by permutation as

$$p_i^\pi = \frac{1}{(m)!} \left| \left\{ \pi \in \Pi : Y_i(X^\pi) \leq Y_i(X) \right\} \right|,$$  \hspace{1cm} (25)

and apply suitable multiple-testing methods to control the family-wise error rate (FWER) or the false discovery rate (FDR). A sensible way to control the FWER at level $\alpha$ is to use a Bonferroni correction and identify anomalous streams as those with p-values lower than $\alpha/n$. This method is intimately related to the method using the $(1 - \alpha)$-quantile of the max test statistic presented above. For the FDR control one needs to take into consideration that the permutation p-values in (25) are dependent. Standard methods as Benjamini and Hochberg (1995) have been adapted to work under arbitrarily dependent p-values (Benjamini and Yekutieli 2001), but are known to be conservative.

Note, however, that there are situations where detection of anomalies is possible, but identification is not (see Donoho and
Jin 2004). In the application we consider, knowledge of the existence of anomalous batches can be of use in the quality control process to gather additional data (such as retesting of batches) to fix production anomalies.

Remark 10. Note that the data in the analysis above seems to follow a normal distribution. In Section 2 in the supplementary materials, we include an analysis of COVID-19 cases in the Netherlands where such normality is questionable, and in this context our methodology exhibits more power than the methodology based on normal approximations. However, to apply our methodology in this context one needs to address the serial dependencies present in the data. We discuss this in the supplementary materials, noting that the validity of the conclusions depends naturally on the appropriateness of the model capturing such dependencies.

6. Discussion

This article considered the problem of detecting anomalous streams of data among many observed streams. This problem is closely related to the detection of sparse mixtures, where it is known that a test based on a higher criticism statistic is asymptotically optimal. Computation of this test statistic and calibration of the ensuing test requires the knowledge of the null (nominal) distribution. In this work we propose a distribution-free version of the statistic that is based on permutation, and we show that it is also asymptotically optimal in the context of one-parameter exponential families. The proposed test is exact, and appears to be superior in the sparse settings considered to alternative proposals that make use of asymptotic approximations.

The numerical results in Section 5.4 show that the calibration by permutation results in nearly no loss of power in finite samples compared to an oracle test constructed and calibrated with full distribution knowledge. Its usefulness on real statistical problems is showcased in Section 5.6, and Section 2 in the supplementary materials.

Our asymptotic results are proven for the class of exponential models. In Delaigle, Hall, and Jin (2011) a less restrictive class of models is considered, for which the proposed testing methodology attains a similar detection boundary when streams are of length $t = n$. We conjecture that, when we consider a similar class of models (i.e., a location family where the base distribution satisfies some tail-decaying condition) the detection boundary $\rho^*$ has a similar shape as well, provided the streams are long enough. Specifically, when the tails of the base distribution are decaying exponentially, a stream length $t = \omega(\log^3 n)$ suffices, as shown in our work. For Pareto tails, we conjecture that at least $t = \omega(n)$ is needed to attain the $\rho^*(\beta)$ detection boundary, using large and moderate deviations theory results from Mikosch and Nagaev (1998). One should keep in mind, however, that in an arbitrary location family the stream means are in general not a sufficient statistic for the location parameter, therefore, the optimal detection boundary might not be characterized in the same way as for the normal model. For example, if $F_0$ is the uniform distribution over $[\theta, \theta + 1]$ a test based on the stream means will surely be suboptimal.

Further research could discuss the possibility of adapting the Berk-Jones statistic (Berk and Jones 1979) in a similar fashion. Alternatively, one could investigate the possibility of replacing the observations by their ranks. This has numerical advantages and in some cases leads only to a minor loss of power (Arias-Castro et al. 2018). A thorough analysis of this proposal is challenging as one must carefully account for the dependencies induced by the use of ranks, and remains an interesting avenue for future work. Another area of future research could discuss the identification of anomalous streams. In case permutation $p$-values are used, their dependence must be carefully characterized and accounted for. Such research could characterize the signal regimes in which anomalies are identifiable, merely detectable, or neither. In high-dimensional problems, the FWER may be a too stringent measure to control and other measures, such as FDR control, may be more appropriate. Work related to this problem can be found, for example, in Romano and Wolf (2007) and Romano, Shaikh, and Wolf (2008).

Supplementary Materials

The online supplementary materials contain proofs of the main results and an application of the methodology on COVID-19 cases in the Netherlands. Please refer to https://arxiv.org/abs/2009.03117 for proofs of the supporting results.

Acknowledgments

I.V.S. would like to thank Richard A.J. Post for interesting discussions on applications of our methods. R.M.C. and E.A-C. would like to thank Ervin Táncozs and Meng Wang for helpful discussions in the early stages of this project.

ORCID

Ivo V. Stoepker https://orcid.org/0000-0001-9579-7259
Rui M. Castro https://orcid.org/0000-0003-4398-0718
Edwin van den Heuvel https://orcid.org/0000-0001-9157-7224

References

Aldosari, S. A., and Moura, J. M. F. (2004), “Detection in Decentralized Sensor Networks,” in 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 2), pp. ii–277. [461]

Arias-Castro, E., Candès, E. J., and Durand, A. (2011), “Detection of an Anomalous Cluster in a Network,” Annals of Statistics, 39, 278–304. [465,468]

Arias-Castro, E., Candès, E. J., and Plan, Y. (2011), “Global Testing under Sparse Alternatives: Ano, Multiple Comparisons and the Higher Criticism,” Annals of Statistics, 39, 2533–2556. [464]

Arias-Castro, E., Castro, R. M., Táncozs, E., and Wang, M. (2018), “Distribution-Free Detection of Structured Anomalies: Permutation and Rank-Based Scans,” Journal of the American Statistical Association, 113, 789–801. [462,465,466,473]

Arias-Castro, E., and Wang, M. (2017), “Distribution-Free Tests for Sparse Heterogeneous Mixtures,” TEST, 26, 71–94. [462]

Benjamini, Y., and Hochberg, Y. (1995), “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” Journal of the Royal Statistical Society, Series B, 57, 289–300. [472]

Benjamini, Y., and Yekutieli, D. (2001), “The Control of the False Discovery Rate in Multiple Testing Under Dependence,” The Annals of Statistics, 29, 1165–1188. [472]

Berk, R. H., and Jones, D. H. (1979), “Goodness-of-Fit Test Statistics that Dominate the Kolmogorov Statistics,” Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 47, 47–59. [473]
