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

Ivo V. Stoepker[1], Rui M. Castro[1], Ery Arias-Castro[2], and Edwin van den Heuvel[1]

[1]Technische Universiteit Eindhoven

[2]University of California, San Diego

## 1 Proof of the main results

In this section we provide the proofs of our main results. We begin with a simple technical lemma that greatly facilitates the presentation. The proof is a trivial consequence of standard tail-bounds for the normal approximation (Feller, 1968, Section 7.1, Lemma 2).

**Lemma 1.** *Let $\Phi$ be the cumulative distribution function of the standard normal distribution and $x \geq 0$. Then*

$$1 - \Phi\left(\sqrt{2(x + o(1))\log(n)}\right) = n^{-x+o(1)} \text{ as } n \to \infty \ .$$

Next, we argue that for the proofs of our main results, we can assume $F_0$ has zero mean and unit variance.

## 1.1 Simplifying assumption

To prove the results in this paper it suffices to consider the case where the nominal distribution $F_0$ has zero mean and unit variance. To see this suppose $F_0$ has arbitrary mean and variance $\mu_0$ and $\sigma_0^2$. Define $\tilde{F}_0(x) = F_0(\mu_0 + \sigma_0 x)$. It is easy to see the distribution $\tilde{F}_0$ has zero mean and unit variance. Using this we can easily re-parameterize the hypothesis test in (11).

Let $X$ be a random variable with distribution $F_\theta$ for some $\theta \in [0, \theta_*)$ and define $\tilde{X} = \frac{X-\mu_0}{\sigma_0}$. Define also $\tilde{\varphi}_0(\tilde{\theta}) = \int e^{\tilde{\theta}x} \mathrm{d}\tilde{F}_0(x)$, the moment generating function of $\tilde{F}_0$. It is easy to check that $\tilde{X}$ has density with respect to $\tilde{F}_0$ given by $\exp\left(\tilde{\theta}x - \log(\tilde{\varphi}(\tilde{\theta}))\right)$ where $\tilde{\theta} = \sigma_0\theta$ (equivalently $\theta = \frac{1}{\sigma_0}\tilde{\theta}$). Therefore, statements in $\tilde{\theta}$ pertaining a zero mean and unit variance distribution can be translated to a general distribution by simple multiplication by a factor $1/\sigma_0$.

## 1.2 Proof of Theorem 1

*Proof.* Without loss of generality and as explained in Section 1.1 we assume that $F_0$ has mean zero and variance one, as this makes the arguments easier and less cluttered.

Let $\psi(\mathbf{X}) : \mathbb{R}^{nt} \to \{0, 1\}$ denote an arbitrary test function. We begin by bounding the worst case risk of this test by the average risk, namely

$$\begin{aligned} R(\psi) &= \mathbb{P}_\emptyset\left(\psi(\mathbf{X}) \neq 0\right) + \max_{\mathcal{S}:|\mathcal{S}|=s} \mathbb{P}_\mathcal{S}\left(\psi(\mathbf{X}) \neq 1\right) \\ &\geq \mathbb{P}_\emptyset\left(\psi(\mathbf{X}) \neq 0\right) + \frac{1}{\binom{n}{s}} \sum_{\mathcal{S}:|\mathcal{S}|=s} \mathbb{P}_\mathcal{S}\left(\psi(\mathbf{X}) \neq 1\right) \ . \end{aligned}$$

The average risk can naturally be interpreted as the risk of testing the simple null hypothesis against a simple alternative, where $\mathcal{S}$ is chosen uniformly at random over the class of all subsets of $[n]$ with cardinality $s$. Since we are doing a test between two simple hypotheses the optimal test (i.e., the test minimizing the average risk) is given by the Neyman-Pearson lemma, namely $\psi(\mathbf{X}) = \mathbb{1}\left\{L \geq 1\right\}$ where $L$ is the likelihood ratio given by

$$L \equiv \frac{1}{\binom{n}{s}} \sum_{\mathcal{S}:|\mathcal{S}|=s} \exp\left(\theta X_\mathcal{S} - ts\log(\varphi_0(\theta))\right) \text{ with } X_\mathcal{S} \equiv \sum_{i\in\mathcal{S}, j\in[t]} X_{ij} \ .$$

The risk of this test can be easily expressed as $1 - \frac{1}{2}\mathbb{E}\left(|L-1|\right)$, where the expectation is with respect to the null hypothesis (so all $X_{ij}$ are i.i.d. with distribution $F_0$). To proceed we need to get an upper bound on $\mathbb{E}\left(|L-1|\right)$. A simple, but often useful way to proceed is to use Jensen's inequality to get

$$\mathbb{E}\left(|L-1|\right) \leq \sqrt{\mathbb{E}\left((L-1)^2\right)} = \sqrt{\mathbb{E}\left(L^2\right) - 1}\ ,$$

where the equality above follows since $\mathbb{E}\left(L\right) = 1$. This approach is generally referred to as the *second moment method*. To show any test is asymptotically powerless it suffices therefore to show that $\mathbb{E}\left(L^2\right)$ converges to one as $n \to \infty$.

To simplify the presentation let $\mathcal{S}$ and $\mathcal{S}'$ denote two independent random variables, and both independent from $\mathbf{X}$. Both $\mathcal{S}$ and $\mathcal{S}'$ are sampled uniformly from the set $\{\mathcal{S} \subset [n] : |\mathcal{S}| = s\}$. Then clearly $L = \mathbb{E}\left(\exp\left(\theta X_{\mathcal{S}} - ts\log(\varphi_0(\theta))\right)|\mathbf{X}\right)$ and therefore

$$\begin{aligned}
\mathbb{E}\left(L^2\right) &= \mathbb{E}\left(\exp\left(\theta X_{\mathcal{S}} - ts\log(\varphi_0(\theta))\right)\exp\left(\theta X_{\mathcal{S}'} - ts\log(\varphi_0(\theta))\right)\right) \\
&= \mathbb{E}\left(\exp\left(\theta(X_{\mathcal{S}} + X_{\mathcal{S}'}) - 2ts\log(\varphi_0(\theta))\right)\right) \\
&= \mathbb{E}\left(\exp\left(t|\mathcal{S}\cap\mathcal{S}'|(\log\varphi_0(2\theta) - 2\log\varphi_0(\theta))\right)\right)\ .
\end{aligned}$$

For the last equality we used the fact that for all $\mathcal{S}$ and $\mathcal{S}'$ we have $X_{\mathcal{S}} + X_{\mathcal{S}'} = 2X_{\mathcal{S}\cap\mathcal{S}'} + X_{\mathcal{S}\triangle\mathcal{S}'}$ (in the previous expression $\triangle$ denotes the symmetric set difference).

The beauty of the above result is that is reduces quantification of the risk to a statement about the moment generating function of the random variable $K \equiv |\mathcal{S}\cap\mathcal{S}'|$. Given the distribution $\mathcal{S}$ and $\mathcal{S}'$ we conclude that $K$ has an hypergeometric distribution with parameters $(n, s, s)$, and therefore $K$ is stochastically bounded from above by the binomial distribution with parameters $(s, \frac{s}{n-s})$. Using the well-know expression for the moment generating function of a binomial distribution we conclude that

$$\mathbb{E}\left(L^2\right) \leq \left(1 - \tfrac{s}{n-s} + \tfrac{s}{n-s}\kappa(\theta)^t\right)^s\ ,$$

where $\kappa(\theta) \equiv \varphi_0(2\theta)/\varphi_0(\theta)^2$. Therefore $\mathbb{E}\left(L^2\right) \to 1$ provided

$$\frac{s^2}{n-s}(\kappa(\theta)^t - 1) \to 0\ .$$

Consider now the specific parameterizations of $s$ and $\theta$ in the theorem statement. Note that when $t = \omega(\log^3 n)$ then necessarily $\theta \to 0$, so we can conveniently use a Taylor expansion of the moment generating function $\varphi_0(\theta)$ around $\theta = 0$:

$$\varphi_0(\theta) = \varphi_0(0) + \theta\varphi_0'(0) + \frac{\theta^2}{2}\varphi_0''(0) + \mathcal{O}(\theta^3) \ , \tag{1}$$

as $\theta \to 0$. Using the fact that $F_0$ has mean zero and unit variance we get $\varphi_0(\theta) = 1 + \frac{1}{2}\theta^2 + \mathcal{O}(\theta^3)$ as $\theta \to 0$. Simple asymptotic algebra yields that $\kappa(\theta) = 1 + \theta^2 + \mathcal{O}(\theta^3)$. Since $1 + x \leq e^x$ we conclude that $\kappa(\theta) \leq \exp\left(\theta^2 + \mathcal{O}(\theta^3)\right)$.

When $\beta > 1/2$ we conclude that

$$\begin{aligned}
\frac{s^2}{n-s}(\kappa(\theta)^t - 1) &= (1 + o(1))n^{1-2\beta}(\kappa(\theta)^t - 1) \\
&\leq (1 + o(1))n^{1-2\beta}\left(\exp\left(t\theta^2 + \mathcal{O}(t\theta^3)\right) - 1\right) \\
&= (1 + o(1))\exp\left((1-2\beta)\log n\right)\left(\exp\left(2r\log n + o(1)\right) - 1\right) \ .
\end{aligned}$$

The last expression converges to 0 provided $1 - 2\beta + 2r < 0$ meaning that when $r < \beta - 1/2$ any test is asymptotically powerless. This lower bound is tight when $\beta \in (1/2, 3/4]$ (the moderately sparse regime) but it is a bit loose for the very sparse regime. However, a modification of the above argument allows us to get a tight lower bound when $\beta > 3/4$.

**The very sparse regime:** the main limitation of the second moment method as presented above has to do with the fact that the likelihood ratio statistic $L$ might take rather large values. Although this might be a rare occurrence, it can be enough to ensure the second moment is much larger than the first moment. A way to mitigate this issue is to consider a so-called truncated second moment method. Let $\Omega$ denote an arbitrary event and define the truncated likelihood ratio $\tilde{L} \equiv L\mathbb{1}\{\Omega\}$. Clearly $\tilde{L} \leq L$ and therefore

$$\begin{aligned}
\mathbb{E}\left(|L - 1|\right) &= \mathbb{E}\left(\left|L - \tilde{L} + \tilde{L} - 1\right|\right) \\
&\leq \mathbb{E}\left(\left|\tilde{L} - 1\right|\right) + 1 - \mathbb{E}\left(\tilde{L}\right) \\
&\leq \sqrt{\mathbb{E}\left(\tilde{L}^2\right) - 2\mathbb{E}\left(\tilde{L}\right) + 1} + 1 - \mathbb{E}\left(\tilde{L}\right) \ ,
\end{aligned}$$

where we used the triangle inequality and the fact that $\mathbb{E}(L) = 1$, followed by Jensen's inequality. Therefore, to show a test is powerless it suffices to show that both $\mathbb{E}\left(\tilde{L}\right)$ and $\mathbb{E}\left(\tilde{L}^2\right)$ converge to one as $n \to \infty$. The choice of event $\Omega$ is therefore quite crucial. In the present context we are going to consider the event

$$\Omega = \left\{ \max_{i\in[n]} Y_i < \underbrace{\sqrt{\frac{2(1+\eta)\log n}{t}}}_{\equiv\tau(\eta)} \right\}, \tag{2}$$

where $\eta > 0$ must be carefully chosen.

**Truncated first moment:** Note first that $\mathbb{E}\left(\tilde{L}\right) = \mathbb{E}\left(L\mathbb{1}\left\{\Omega\right\}\right)$ is the probability of $\Omega$ under the alternative hypothesis (where there is a set $\mathcal{S}$ of anomalous streams and $\mathcal{S}$ is chosen uniformly at random over the subsets of $[n]$ with cardinality $s$). Given the symmetry of the definition of $\Omega$ we see that $\mathbb{E}\left(\tilde{L}\right) = \mathbb{P}_{\mathcal{S}}(\Omega)$ where $\mathcal{S}$ is an arbitrary set with cardinality $s$. Without loss of generality let $\mathcal{S} = [s]$. Then

$$\begin{aligned}
\mathbb{E}\left(\tilde{L}\right) &= \mathbb{P}_{\mathcal{S}}(\Omega) \\
&= 1 - \mathbb{P}_{\mathcal{S}}\left(\max_{i\in[n]} Y_i \geq \tau(\eta)\right) \\
&= 1 - s\mathbb{P}_{\mathcal{S}}\left(Y_1 \geq \tau(\eta)\right) - (n-s)\mathbb{P}_{\emptyset}\left(Y_1 \geq \tau(\eta)\right) \ .
\end{aligned}$$

using the union bound in the last line. Using Lemma 6 we conclude that

$$\mathbb{P}_{\emptyset}\left(Y_1 \geq \tau(\eta)\right) = n^{-1+\eta+o(1)} \ ,$$

and provided $r \leq 1+\eta$

$$\mathbb{P}_{\mathcal{S}}\left(Y_1 \geq \tau(\eta)\right) = n^{-(\sqrt{1+\eta}-\sqrt{r})^2+o(1)} \ .$$

Therefore, when $r \leq 1+\eta$

$$\begin{aligned}
\mathbb{E}\left(\tilde{L}\right) &= 1 - n^{1-\beta}n^{-(\sqrt{1+\eta}-\sqrt{r})^2+o(1)} - (n-s)n^{-1+\eta+o(1)} \\
&= 1 - n^{1-\beta-(\sqrt{1+\eta}-\sqrt{r})^2+o(1)} - \mathcal{O}(1) \to 1 \ ,
\end{aligned}$$

provided $r < (\sqrt{1+\eta} - \sqrt{1-\beta})^2$. This means the first truncated moment converges to one for any $\eta > 0$, provided $r < (1-\sqrt{1-\beta})^2$.

**Truncated second moment:** bounding this term requires significantly more work. Begin by noting that

$$\begin{aligned}
\mathbb{E}\left(\tilde{L}^2\right) &= \mathbb{E}\left(\exp\left(2\theta X_{\mathcal{S}\cap\mathcal{S}'} + \theta X_{\mathcal{S}\triangle\mathcal{S}'} - ts\log(\varphi_0(\theta))\right)\mathbb{1}\left\{\Omega\right\}\right) \\
&= \varphi_0(\theta)^{-st}\mathbb{E}\left(\exp\left(2\theta X_{\mathcal{S}\cap\mathcal{S}'}\right)\exp\left(\theta X_{\mathcal{S}\triangle\mathcal{S}'}\right)\prod_{i\in[n]}\mathbb{1}\left\{Y_i < \tau(\eta)\right\}\right) \\
&\leq \varphi_0(\theta)^{-st}\mathbb{E}\left(\exp\left(2\theta X_{\mathcal{S}\cap\mathcal{S}'}\right)\exp\left(\theta X_{\mathcal{S}\triangle\mathcal{S}'}\right)\prod_{i\in\mathcal{S}\cup\mathcal{S}'}\mathbb{1}\left\{Y_i < \tau(\eta)\right\}\right) \\
&\leq \varphi_0(\theta)^{-st}\mathbb{E}\left(\exp\left(t|\mathcal{S}\cap\mathcal{S}'|(\log\tilde{\varphi}_0(2\theta)) - t|\mathcal{S}\triangle\mathcal{S}'|(\log\tilde{\varphi}_0(\theta))\right)\right) ,
\end{aligned}$$

where $\tilde{\varphi}_0(\theta)^t \equiv \mathbb{E}\left(\exp\left(\theta t Y_1\right)\mathbb{1}\left\{Y_1 < \tau(\eta)\right\}\right)$. The steps above mimic the derivation for the regular second moment, and the main difference is that we now need to consider the moment generating function of a truncated distribution, instead of the original distribution. Clearly $\tilde{\varphi}_0(\theta) \leq \varphi_0(\theta)$ and so we conclude that

$$\mathbb{E}\left(\tilde{L}^2\right) \leq \mathbb{E}\left(\exp\left(t|\mathcal{S}\cap\mathcal{S}'|(\log\tilde{\varphi}_0(2\theta) - 2\log\varphi_0(\theta))\right)\right) .$$

Define $\tilde{\kappa}(\theta) \equiv \tilde{\varphi}_0(2\theta)/\varphi_0(\theta)^2$. As before, to show the truncated moment converges to zero it suffices to show that

$$\frac{s^2}{n-s}(\tilde{\kappa}(\theta)^t - 1) \to 0 .$$

Note that, the argument based on the untruncated second moment method indicates all tests are powerless if $r < \beta - 1/4$. Since we are considering the case $\beta \geq 3/4$ this means that it suffices to treat only the case where $r \geq 3/4 - 1/2 = 1/4$. To get an upper bound on $\tilde{\varphi}_0(2\theta)^t$ we make use of the following technical result.

**Lemma 2.** *Let $X$ be a real-valued random variable and let $f : \mathbb{R} \to [0,\infty)$ be one-to-one increasing and differentiable. Then, for any $\tau \in \mathbb{R}$,*

$$\mathbb{E}\left(f(X)\mathbb{1}\left\{X \leq \tau\right\}\right) = \int_{-\infty}^{\tau}\mathbb{P}(X > x)f'(x)\mathrm{d}x . \tag{3}$$

To use this lemma we must get a good upper bound on $\mathbb{P}_\emptyset\left(Y_1 > x\right)$ for $x \leq \tau(\eta)$. When $x \leq 0$ we trivially bound this probability by one, and for $0 \leq x < \theta_*$ we make use of a simple

Chernoff bound. In a similar fashion to the proof of Lemma 6 we have

$$\begin{aligned}
\mathbb{P}_{\emptyset}\left(Y_1 > x\right) &\leq \mathbb{P}_{\emptyset}\left(\sum_{j\in[t]} X_{1j} \geq xt\right) \\
&\leq \exp\left(-t\left[\sup_{\lambda\in[0,\theta_*)}\left\{\lambda x - \log(\varphi_0(\lambda))\right\}\right]\right) \\
&= \exp\left(-t\left[\sup_{\lambda\in[0,\theta_*)}\left\{\lambda x - \log(1+\lambda^2/2+\mathcal{O}(\lambda^3))\right\}\right]\right) \\
&\leq \exp\left(-t\left[x^2 - \log(1+x^2/2+\mathcal{O}(x^3))\right\}\right]\right) \\
&\leq \exp\left(-t(x^2/2+\mathcal{O}(x^3))\right) \ .
\end{aligned}$$

Let $n$ be large enough so that $\tau(\eta) < \theta_*$. Applying the above result and Lemma 2 we get

$$\begin{aligned}
\tilde{\varphi}_0(2\theta)^t &= \int_{-\infty}^{\tau(\eta)} \mathbb{P}\left(Y_1 > x\right) 2\theta t \exp\left(2\theta t x\right) dx \\
&\leq \int_{-\infty}^{0} 2\theta t \exp\left(2\theta t x\right) dx \ + \ \int_{0}^{\tau(\eta)} \exp\left(-t(x^2/2+\mathcal{O}(x^3))\right) 2\theta t \exp\left(2\theta t x\right) dx \\
&\leq 1 \ + \ 2\theta t \int_{0}^{\tau(\eta)} \exp\left(-t(x^2/2+\mathcal{O}(x^3))\right) \exp\left(2\theta t x\right) dx \\
&= 1 \ + \ 2\theta t \exp\left(2\theta^2 t\right) \exp\left(\mathcal{O}(t\tau^3(\eta))\right) \int_{0}^{\tau(\eta)} \exp\left(-t(x-2\theta)^2/2)\right) dx \\
&= 1 \ + \ \sqrt{8\pi}(1+o(1))(\theta\sqrt{t}) \exp\left(2\theta^2 t\right) \int_{-2\theta\sqrt{t}}^{(\tau(\eta)-2\theta)\sqrt{t}} \frac{1}{\sqrt{2\pi}} \exp\left(-y^2/2\right) dy \\
&\leq 1 \ + \ \sqrt{8\pi}(1+o(1))(\theta\sqrt{t}) \exp\left(2\theta^2 t\right) \Phi((\tau(\eta)-2\theta)\sqrt{t}) \ ,
\end{aligned}$$

where in the second to last step we used the fact that $t = \omega(\log^3 n)$. At this point note that

$$(\tau(\eta)-2\theta)\sqrt{t} = -\sqrt{2\left(\sqrt{4r}-\sqrt{1+\eta}\right)^2 \log n} < 0$$

when $r \geq 1/4$, provided we choose $\eta > 0$ sufficiently small. Therefore using Lemma 1 we conclude that

$$\begin{aligned}
\tilde{\varphi}_0(2\theta)^t &\leq 1 \ + \ \sqrt{8\pi}(1+o(1))\sqrt{2r\log n} n^{4r} n^{-(\sqrt{4r}-\sqrt{1+\eta})^2+o(1)} \\
&= 1 \ + \ \sqrt{8\pi}(1+o(1))\sqrt{2r\log n} \ n^{4r-(\sqrt{4r}-\sqrt{1+\eta})^2+o(1)} \\
&= n^{4r-(\sqrt{4r}-\sqrt{1+\eta})^2+o(1)} \ .
\end{aligned}$$

With an analogous argument to the one used for $\tilde{\varphi}(2\theta)^t$ one can show that $(\varphi_0(\theta))^{2t} =$

$n^{2r+o(1)}$. Therefore

$$\frac{s^2}{n-s}(\tilde{\kappa}(\theta)^t - 1) = o(1) + n^{1-2\beta+2r-(\sqrt{4r}-\sqrt{1+\eta})^2+o(1)} \ .$$

The above converges to zero when $1-2\beta+2r-(\sqrt{4r}-\sqrt{1+\eta})^2 < 0$. This is the case $\eta$ is small enough and $r < (1-\sqrt{1-\beta})^2$, since in that case $1-2\beta+2r-(\sqrt{4r}-1)^2 < 0$, completing the proof. □

## 1.3 Proof of Theorem 2

*Proof.* By the arguments in Section 1.1, the proof continues under the assumption that $F_0$ has zero mean and unit variance, without loss of generality.

Under the null and for a given (but arbitrary) permutation $\pi \in \Pi$ it is clear that $\mathbf{X}^\pi$ and $\mathbf{X}$ have exactly the same distribution. Therefore $\max_i \{Y_i(\mathbf{X})\}$ is uniformly distributed on the set $\{\max_i\{Y_i(\mathbf{X}^\pi\}, \pi \in \Pi\}$ (with multiplicities) conditionally on the order statistics of $\mathbf{X}$. So, for a given $\alpha > 0$

$$\begin{aligned}\mathbb{P}_\emptyset\left(\mathcal{P}_{\text{max-perm}}(\mathbf{X}) \leq \alpha\right) &= \mathbb{P}_\emptyset\left(\left|\{\pi \in \Pi : \max_i\{Y_i(\mathbf{X}^\pi)\} \geq \max_i\{Y_i(\mathbf{X})\}\}\right| \leq \alpha(nt)!\right) \\ &\leq \frac{\lfloor \alpha(nt)!\rfloor}{(nt)!} \leq \alpha \ ,\end{aligned}$$

If there are no ties, the first inequality above is an equality, but with ties present the test becomes slightly more conservative. This argument is completely standard and for more details on permutation tests the reader is referred to (Lehmann and Romano, 2005).

What remains to be proven is the behavior of the test under the alternative. Namely we must show that, provided $r$ is large enough (as stated in the theorem) then for any $\alpha > 0$

$$\mathbb{P}_\emptyset\left(\mathcal{P}_{\text{max-perm}}(\mathbf{X}) > \alpha\right) \longrightarrow 0 \ .$$

For convenience, let $\pi$ be a uniformly distributed permutation of $\Pi$ and let this be independent from $\mathbf{X}$. We can rewrite our permutation $p$-value as a conditional probability:

$$\mathcal{P}_{\text{max-perm}}(\mathbf{X}) = \mathbb{P}\left(\max_i Y_i^\pi \geq \max_i Y_i \,\middle|\, \mathbf{X}\right) \ .$$

To get a good upper-bound on the permutation $p$-value we use the following concentration inequality (see Shorack and Wellner (1986) and Arias-Castro et al. (2018), for instance).

**Lemma 3 (Bernstein bound for sampling without replacement).** *Let* $(Z_1, \dots, Z_m)$ *be sampled without replacement from the set* $\{z_1, \dots, z_n\}$. *Define* $z_{max} = \max_j\{z_j\}$, $\overline{z} = \frac{1}{n}\sum_{j=1}^n z_j$, $\overline{Z} = \frac{1}{m}\sum_{j=1}^n Z_j$ *and* $\sigma_z^2 = \frac{1}{n}\sum_{j=1}^n (z_j - \overline{z})^2$. *Then, for all* $\tau \geq 0$*:*

$$\mathbb{P}\left(\overline{Z} \geq \overline{z} + \tau\right) \leq \exp\left(-\frac{m\tau^2}{2\sigma_z^2 + \frac{2}{3}(z_{max} - \overline{z})\tau}\right).$$

Using this lemma, we find that

$$\begin{aligned}
\mathcal{P}_{\text{max-perm}}(\mathbf{X}) &= \mathbb{P}\left(\max_i Y_i(\mathbf{X}^\pi) \geq \max_i Y_i(\mathbf{X}) \;\middle|\; \mathbf{X}\right) \\
&\leq \sum_{k\in[n]} \mathbb{P}\left(Y_k(\mathbf{X}^\pi) \geq \max_i Y_i(\mathbf{X}) \;\middle|\; \mathbf{X}\right) \\
&= \sum_{k\in[n]} \mathbb{P}\left(\frac{1}{t}\sum_{j\in[t]} X_{k,j}^\pi \geq \overline{X} + \left(\max_i Y_i(\mathbf{X}) - \overline{X}\right) \;\middle|\; \mathbf{X}\right) \\
&\leq \sum_{k\in[n]} \exp\left(-\frac{t\left(\max_i Y_i(\mathbf{X}) - \overline{X}\right)^2}{2\sigma_X^2 + \frac{2}{3}(\max_{i,j} X_{ij} - \overline{X})\left(\max_i Y_i(\mathbf{X}) - \overline{X}\right)}\right) \\
&= n \cdot \exp\left(-\frac{t\left(\max_i Y_i(\mathbf{X}) - \overline{X}\right)^2}{2\sigma_X^2 + \frac{2}{3}(\max_{i,j} X_{ij} - \overline{X})\left(\max_i Y_i(\mathbf{X}) - \overline{X}\right)}\right).
\end{aligned}$$

The first inequality is a consequence of a simple union bound, and we used Lemma 3 in the second inequality.

At this point it is clear that, to control the $p$-value of our test we need to characterize the behavior of $\overline{X}$, $\sigma_X^2$, $\max_{i,j} X_{ij}$ and $\max_i Y_i(\mathbf{X})$ under the alternative hypothesis. Since $|\mathcal{S}| = o(n)$ most of the elements of $\mathbf{X}$ are samples from the null distribution. Therefore we intuitively expect that $\mathbf{X}$ and $\sigma_X^2$ should be good estimators for the mean and variance of $F_0$. The behavior of the term $\max_{i,j} X_{ij}$ is a bit more delicate, but one can see that for the given parameterization of the null the dominant contribution is still given by the null distribution. In contrast, the term $\max_i Y_i(\mathbf{X}) - \overline{X}$ really depends on the alternative - the largest stream mean is surely driven by the anomalous observations. Formally, we can show the following result.

**Lemma 4.** *Let* $\beta \in (\frac{1}{2}, 1)$, $\theta = \sqrt{2r(\log n)/t}$ *with* $r > 0$ *and consider the alternative hypothesis in* (11). *Assume* $F_0$ *has zero mean and variance one and* $t = \omega(\log(n))$. *Then*

*(i)* $\overline{X} = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{nt}}\right)$ *and* $\sigma_X^2 = 1 + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{nt}}\right)$

*(ii) Let* $c \in (0, \theta_* - \theta)$*. Then,*

$$\mathbb{P}_{\mathcal{S}}\left(\max_{i,j} X_{ij} - \overline{X} \leq \frac{3}{c}\log(nt)\right) \to 1 \text{ as } n \to \infty \ .$$

*(iii) Assume further that* $t = \omega(\log^3(n))$ *and let* $\varepsilon > 0$*. Provided* $r > (\sqrt{1+\varepsilon} - \sqrt{1-\beta})^2$ *we have*

$$\mathbb{P}_{\mathcal{S}}\left(\max_i Y_i(\mathbf{X}) - \overline{X} \geq \sqrt{\frac{2(1+\varepsilon)}{t}\log(n)}\right) \to 1$$

*as* $n \to \infty$*.*

The first result of the lemma provides a rate at which the bound on our sample variance can decrease. For the analysis of the max test a much simpler result (already proved in Arias-Castro et al. (2018)) suffices: for any $\varepsilon > 0$ (i) implies that

$$\mathbb{P}_{\mathcal{S}}\left(\sigma_X^2 \leq (1 + \varepsilon/2)\right) \to 1 \ .$$

Note also that the bound in Lemma 3 is monotonically decreasing in $\tau$. This ensures that for $\varepsilon > 0$ and with probability tending to one under the alternative, provided $r > (\sqrt{1+\varepsilon} - \sqrt{1-\beta})^2$, the overall $p$-value of test satisfies

$$\mathcal{P}_{\text{max-perm}}(\mathbf{X}) \leq n \cdot \exp\left(-\frac{2(1+\varepsilon)\log(n)}{2(1+\varepsilon/2) + 2\varepsilon + \frac{2}{c}\log(nt)\sqrt{2(1+\varepsilon)\frac{1}{t}\log(n)}}\right) \ .$$

Written differently, with probability tending to 1 under the alternative:

$$\log \mathcal{P}_{\text{max-perm}}(\mathbf{X}) \leq \log(n)\left(1 - \frac{1+\varepsilon}{1 + \varepsilon/2 + \frac{1}{c}(\log(n) + \log(t))\sqrt{2(1+\varepsilon)\frac{\log(n)}{t}}}\right) \ .$$

To ensure $\mathcal{P}_{\text{max-perm}}(\mathbf{X}) \to 0$, or equivalently, $\log \mathcal{P}_{\text{max-perm}}(\mathbf{X}) \to -\infty$ it suffices to ensure

$$\frac{1}{c}(\log(n) + \log(t))\sqrt{2(1+\varepsilon)\frac{\log(n)}{t}} < \varepsilon/2 \ .$$

However, since we assume $t = \omega(\log^3(n))$ this is immediately satisfied, since the l.h.s. converges to zero.

We have just proved that, for $\varepsilon > 0$ and $r > (\sqrt{1+\varepsilon} - \sqrt{1-\beta})^2$, $\mathcal{P}_{\text{max-perm}}(\mathbf{X}) \to 0$ as $n \to \infty$. Since $\varepsilon > 0$ is arbitrary this implies the result in the theorem, concluding the

proof. □

## 1.4 Proof of Theorem 3

*Proof.* By the arguments in Section 1.1, the proof continues under the assumption that $F_0$ has zero mean and unit variance. Furthermore the conservativeness of this test follows by the standard argument already presented in the proof of Theorem 2.

For the rest of the proof consider alternative hypothesis. We must show that

$$\mathbb{P}_{\mathcal{S}}\left(\tilde{\mathcal{P}}_{\text{perm-hc}}(\mathbf{X}) \leq \alpha\right) \to 1$$

as $n \to \infty$. Like before, we can write our permutation $p$-value as a conditional probability as follows:

$$\tilde{\mathcal{P}}_{\text{perm-hc}}(\mathbf{X}) = \mathbb{P}_{\mathcal{S}}\left(\tilde{T}(\mathbf{X}^{\pi}) \geq \tilde{T}(\mathbf{X}) \,\middle|\, \mathbf{X}\right) ,$$

where $\pi$ is independent from $\mathbf{X}$ and uniformly distributed over $\Pi$. The first step is to understand and simplify the role of $\pi$ in the above expression. This mirrors the analysis under the null hypothesis in the proof of Theorem 2, as we need to show that the values of the test statistic computed with permuted data are a good surrogate for the values of the test statistic under the null. However, the argument becomes more complex due to the dependencies introduced by the permutation. Like before, we will use the union bound for the max-operator in the permuted statistic, inducing a multiplicity by the grid-size:

$$\tilde{\mathcal{P}}_{\text{perm-hc}}(\mathbf{X}) = \mathbb{P}_{\mathcal{S}}\left(\max_{q\in Q} \tilde{V}_q(\mathbf{X}^{\pi}) \geq \tilde{T}(\mathbf{X}) \,\middle|\, \mathbf{X}\right) \leq \sum_{q\in Q} \mathbb{P}_{\mathcal{S}}\left(\tilde{V}_q(\mathbf{X}^{\pi}) \geq \tilde{T}(\mathbf{X}) \,\middle|\, \mathbf{X}\right) . \quad (4)$$

To proceed recall that quantifying $\tilde{V}_q(\mathbf{X}^{\pi})$ requires the quantification of two terms: $\tilde{N}_q(\mathbf{X}^{\pi})$ and $\tilde{P}_q(\mathbf{X}^{\pi})$. Note, however, that $\tilde{P}_q(\mathbf{X})$ is invariant under permutations of $\mathbf{X}$, and therefore $\tilde{P}_q(\mathbf{X}^{\pi}) = \tilde{P}_q(\mathbf{X})$, as explained before. As such the only random quantity inside the probability operator above (conditionally on $\mathbf{X}$) is $\tilde{N}_q(\mathbf{X}^{\pi})$. Noting that $\mathbb{E}\left(\tilde{N}_q(\mathbf{X}^{\pi}) \,\middle|\, \mathbf{X}\right) = n\tilde{P}_q(\mathbf{X})$ we have

$$\tilde{\mathcal{P}}_{\text{perm-hc}}(\mathbf{X}) \leq \sum_{q\in Q} \mathbb{P}_{\mathcal{S}}\left(\tilde{N}_q(\mathbf{X}^{\pi}) - \mathbb{E}\left(\tilde{N}_q(\mathbf{X}^{\pi})\right) \geq \tilde{T}(\mathbf{X})\sqrt{n\tilde{P}_q(\mathbf{X})(1-\tilde{P}_q(\mathbf{X}))} \,\middle|\, \mathbf{X}\right) . \quad (5)$$

To apply Chebyshev's inequality we need the right-hand-side of the inequality inside the

probability to be positive, and $\tilde{T}(\mathbf{X})$ might be negative. In the latter case we simply bound the probability by one. Therefore, using Chebyshev's inequality we get

$$\tilde{\mathcal{P}}_{\text{perm-hc}}(\mathbf{X}) \leq \mathbb{1}\left\{\tilde{T}(\mathbf{X}) \leq 0\right\} + \frac{\mathbb{1}\left\{\tilde{T}(\mathbf{X}) > 0\right\}}{\tilde{T}(\mathbf{X})^2} \sum_{q \in Q} \frac{\text{Var}\left(\tilde{N}_q(\mathbf{X}^\pi) \,\middle|\, \mathbf{X}\right)}{n\tilde{P}_q(\mathbf{X})(1-\tilde{P}_q(\mathbf{X}))} \,, \tag{6}$$

where we convention that $0/0 = 0$. To continue, we must quantify the conditional variance of $\tilde{N}_q(\mathbf{X}^\pi)$. The permutation on $\mathbf{X}$ causes dependencies, but these are benign when realizing the conditional permuted stream means are negatively associated conditional on the data. Using Theorem 2.11 and the properties $\text{P}_6$ and $\text{P}_4$ in Joag-Dev and Proschan (1983), we find that $Y_i(\mathbf{X}^\pi)|\mathbf{X}$ and $Y_j(\mathbf{X}^\pi)|\mathbf{X}$ are negatively associated if $i \neq j$. For ease of notation, define

$$z_q \equiv \sqrt{\frac{2q}{t}\log(n)} \,. \tag{7}$$

Then,

$$\begin{aligned}
\text{Var}\left(\tilde{N}_q(\mathbf{X}^\pi) \,\middle|\, \mathbf{X}\right) &= \sum_{i \in [n]} \text{Var}\left(\mathbb{1}\left\{Y_i(\mathbf{X}^\pi) \geq z_q\right\} \mid \mathbf{X}\right) \\
&\qquad + \sum_{i \in [n]} \sum_{j \neq i} \text{Cov}\left(\mathbb{1}\left\{Y_i(\mathbf{X}^\pi) \geq z_q\right\}, \mathbb{1}\left\{Y_j(\mathbf{X}^\pi) \geq z_q\right\} \mid \mathbf{X}\right) \\
&\leq n\tilde{P}_q(\mathbf{X})(1-\tilde{P}_q(\mathbf{X})) \,,
\end{aligned}$$

where we used the definition of negative association (Definition 2.1 from Joag-Dev and Proschan (1983)). In conclusion we get the following simple bound for the $p$-value:

$$\begin{aligned}
\tilde{\mathcal{P}}_{\text{perm-hc}}(\mathbf{X}) &\leq \frac{1}{\tilde{T}(\mathbf{X})^2}\left(\sum_{q \in Q} 1\right) \mathbb{1}\left\{\tilde{T}(\mathbf{X}) > 0\right\} + \mathbb{1}\left\{\tilde{T}(\mathbf{X}) \leq 0\right\} \\
&= \frac{k_n + 1}{\tilde{T}(\mathbf{X})^2} \mathbb{1}\left\{\tilde{T}(\mathbf{X}) > 0\right\} + \mathbb{1}\left\{\tilde{T}(\mathbf{X}) \leq 0\right\} \,.
\end{aligned} \tag{8}$$

To continue the proof we must show that $\tilde{T}(\mathbf{X})$ is of order larger than $k_n = n^{o(1)}$. This mimics the approach in Proposition 1 and Theorem 2 under the alternative. Recall that $\tilde{T}(\mathbf{X}) = \max_{q \in Q} \tilde{V}_q(\mathbf{X})$, so it suffices to show that $\tilde{V}_q(\mathbf{X})$ is larger than $k_n$ with high probability for particular values of $q \in Q$. At the final stretch of the proof, it will become clear that one only needs to consider sequences of values $q_n \in Q$ which converge to a fixed value $q$, so let

$q_n \in Q$ with $q_n \to q$ and $q > 0$. To start, note that:

$$\mathbb{P}_{\mathcal{S}}\left(\tilde{V}_{q_n}(\mathbf{X}) \geq k_n\right) = \mathbb{P}_{\mathcal{S}}\left(\tilde{N}_{q_n}(\mathbf{X}) - \mathbb{E}\left(\tilde{N}_{q_n}(\mathbf{X})\right) \geq A_{q_n}(\mathbf{X})\right) \ , \tag{9}$$

where we have defined for convenience

$$A_{q_n}(\mathbf{X}) \equiv k_n\sqrt{n\tilde{P}_{q_n}(\mathbf{X})(1-\tilde{P}_{q_n}(\mathbf{X})} + \left(n\tilde{P}_{q_n}(\mathbf{X}) - \mathbb{E}\left(\tilde{N}_{q_n}(\mathbf{X})\right)\right) \ .$$

To bound the above probability using Chebyshev's inequality, we first need to find a high-probability upper bound for the random quantity $A_{q_n}(\mathbf{X})$. Note that the second term in $A_{q_n}(\mathbf{X})$ will typically be negative when anomalies are present. To characterize this quantity, define:

$$w_{i,q} \equiv \mathbb{P}_{\mathcal{S}}\left(Y_i(\mathbf{X}) \geq \sqrt{\frac{2q}{t}\log(n)}\right) \ ,$$

and let $\tilde{p}_q \equiv w_{i,q}$ if $i \notin \mathcal{S}$ and $\tilde{v}_q \equiv w_{i,q}$ if $i \in \mathcal{S}$. Now, note that under the alternative $\mathbb{E}\left(\tilde{N}_q(\mathbf{X})\right) = (n-s)\tilde{p}_q + s\tilde{v}_q$, such that

$$A_{q_n}(\mathbf{X}) = k_n\sqrt{n\tilde{P}_{q_n}(\mathbf{X})(1-\tilde{P}_{q_n}(\mathbf{X}))} + s(\tilde{p}_{q_n} - \tilde{v}_{q_n}) + n\left(\tilde{P}_{q_n}(\mathbf{X}) - \tilde{p}_{q_n}\right) \ . \tag{10}$$

Note that using Lemma 6, we can easily characterize $\tilde{p}_{q_n}$ and $\tilde{v}_{q_n}$, and conclude that $\tilde{p}_{q_n} = n^{-q+o(1)}$ and

$$\tilde{v}_{q_n} = \begin{cases} n^{-(\sqrt{q}-\sqrt{r})^2+o(1)} & \text{if } r < q \\ n^{o(1)} & \text{if } r \geq q \end{cases} \ .$$

At this point, the expression above looks remarkably similar to the critical terms encountered in the proof of Proposition 1. However, we have an extra term $n\left(\tilde{P}_{q_n}(\mathbf{X}) - \tilde{p}_{q_n}\right)$ that also needs to be controlled. If we ignore that term then it would suffice to show that $\tilde{P}_{q_n}(\mathbf{X}) \approx n^{-q+o(1)}$ to complete the proof. However, such guarantee is not enough to control the last term, and a much more refined result is required to ensure $\tilde{P}_{q_n}(\mathbf{X})$ is a sufficiently accurate surrogate for $\tilde{p}_{q_n}$. In detail, we require the first term in $A_{q_n}(\mathbf{X})$ to be at most $n^{(1-q)/2+o(1)}$ with high probability, and the third term cannot outweigh the preceding two. The accuracy of the approximation $\tilde{P}_{q_n}(\mathbf{X})$ to $\tilde{p}_{q_n}$ is captured in the following lemma:

**Lemma 5.** *Consider the setting of Lemma 4 and let $q_n \to q$ with $q \in (0,1]$, $t = \omega(\log^3(n))$ and $t = n^{o(1)}$. Then, for any $\varepsilon > 0$, there exists sequence $g_n \to 0$ such that under both the*

*null and alternative hypothesis*

$$\mathbb{P}\left(\tilde{P}_{q_n}(\mathbf{X}) - \tilde{p}_{q_n} \leq n^{\max\{-\frac{1+q}{2}, -\beta-q\}+\varepsilon+g_n}\right) \to 1 \ .$$

Note that this lemma, together with the characterization of $\tilde{p}_q$, implies that there exists a sequence $g_n \to 0$ such that

$$\mathbb{P}_{\mathcal{S}}\left(\tilde{P}_{q_n}(\mathbf{X}) \leq n^{-q+g_n}\right) \to 1 \ .$$

Putting all the facts together, we conclude that for any $\varepsilon > 0$, there exists a deterministic sequence $a_n$ with characterization

$$a_n \equiv \begin{cases} n^{\max\{\frac{1-q}{2}, 1-\beta-q\}+\varepsilon+o(1)} - n^{1-\beta-(\sqrt{q}-\sqrt{r})^2+o(1)} \text{ if } r < q \ , \\ n^{\max\{\frac{1-q}{2}, 1-\beta-q\}+\varepsilon+o(1)} - n^{1-\beta+o(1)} \text{ if } r \geq q \ , \end{cases}$$

such that for the event $\Omega \equiv \{A_{q_n}(\mathbf{X}) \leq a_n\}$ we have $\mathbb{P}(\Omega) \to 1$. Note that $a_n$ is nearly the same term as encountered in the proof of Proposition 1 - although there we were able to characterize the counterpart of $\tilde{v}_q$ in a sharper way, but this does not affect the final result. We can now proceed as follows:

$$\begin{aligned} \mathbb{P}_{\mathcal{S}}\left(\tilde{V}_{q_n}(\mathbf{X}) \leq k_n\right) &= \mathbb{P}_{\mathcal{S}}\left(\tilde{N}_{q_n}(\mathbf{X}) - \mathbb{E}\left(\tilde{N}_{q_n}(\mathbf{X})\right) \leq A_{q_n}(\mathbf{X})\right) \\ &\leq \mathbb{P}_{\mathcal{S}}\left(\tilde{N}_{q_n}(\mathbf{X}) - \mathbb{E}\left(\tilde{N}_{q_n}(\mathbf{X})\right) \leq a_n \ \Big| \ \Omega\right) + \mathbb{P}_{\mathcal{S}}(\Omega^c) \\ &\leq \mathbb{P}_{\mathcal{S}}(\Omega)^{-1}\,\mathbb{P}_{\mathcal{S}}\left(\tilde{N}_{q_n}(\mathbf{X}) - \mathbb{E}\left(\tilde{N}_{q_n}(\mathbf{X})\right) \leq a_n\right) + \mathbb{P}_{\mathcal{S}}(\Omega^c) \\ &= \mathbb{P}_{\mathcal{S}}(\Omega)^{-1}\,\mathbb{P}_{\mathcal{S}}\left(-\left(\tilde{N}_{q_n}(\mathbf{X}) - \mathbb{E}\left(\tilde{N}_{q_n}(\mathbf{X})\right)\right) \geq -a_n\right) + \mathbb{P}_{\mathcal{S}}(\Omega^c) \\ &\leq \mathbb{P}_{\mathcal{S}}(\Omega)^{-1}\,\mathbb{P}_{\mathcal{S}}\left(\left|\tilde{N}_{q_n}(\mathbf{X}) - \mathbb{E}\left(\tilde{N}_{q_n}(\mathbf{X})\right)\right| \geq -a_n\right) + \mathbb{P}_{\mathcal{S}}(\Omega^c) \\ &\leq \mathbb{P}_{\mathcal{S}}(\Omega)^{-1}\,a_n^{-2}\mathrm{Var}\left(\tilde{N}_{q_n}(\mathbf{X})\right) + \mathbb{P}_{\mathcal{S}}(\Omega^c) \ , \end{aligned} \tag{11}$$

where the last inequality follows from Chebyshev's inequality provided $a_n < 0$. Note that $a_n < 0$ for $n$ sufficiently large provided:

$$\begin{cases} 1-\beta-(\sqrt{q}-\sqrt{r})^2 - \max\left\{\frac{1-q}{2}, 1-\beta-q\right\} - \varepsilon > 0 & \text{if } r < q \ , \\ 1-\beta-\max\left\{\frac{1-q}{2}, 1-\beta-q\right\} - \varepsilon > 0 & \text{if } r \geq q \ . \end{cases} \tag{12}$$

Recall that $\mathbb{P}_{\mathcal{S}}(\Omega) \to 1$, and therefore for $n$ sufficiently large we have $\mathbb{P}_{\mathcal{S}}(\Omega) \geq 1/2$. Fur-

thermore, the summands in $\tilde{N}_q(\mathbf{X})$ are independent, such that

$$\mathrm{Var}\left(\tilde{N}_{q_n}(\mathbf{X})\right) = (n-s)\tilde{p}_{q_n}(1-\tilde{p}_{q_n}) + s\tilde{v}_{q_n}(1-\tilde{v}_{q_n}) \ .$$

Assuming (12) and using (11) we conclude that for large enough $n$

$$\mathbb{P}_{\mathcal{S}}\left(\tilde{V}_{q_n}(\mathbf{X}) \leq k_n\right) \leq 2a_n^{-2}\Big((n-s)\tilde{p}_{q_n}(1-\tilde{p}_{q_n}) + s\tilde{v}_{q_n}(1-\tilde{v}_{q_n})\Big) + \mathbb{P}\left(\Omega^c\right) \ .$$

Note that $\mathbb{P}_{\mathcal{S}}\left(\Omega^c\right) \to 0$. Using the asymptotic characterization of $\tilde{p}_{q_n}$ and $\tilde{v}_{q_n}$, the first term converges to 0 provided:

$$\begin{cases} \max\{1-q, 1-\beta-(\sqrt{q}-\sqrt{r})^2\} - 2(1-\beta-(\sqrt{q}-\sqrt{r})^2) < 0 & \text{if } r < q \ , \\ \max\{1-q, 1-\beta\} - 2(1-\beta) < 0 & \text{if } r \geq q \ . \end{cases} \tag{13}$$

Note that the conditions in (12) and (13) are nearly identical to those obtained when proving Proposition 1, with the former holding for any $\varepsilon > 0$. Now, similar algebra as used in the proof of that proposition boils down to the same resulting requirements in the statement of that proposition, i.e. if

$$\begin{cases} r > (1-\sqrt{1-\beta})^2 & \text{if } q = 1 \ , \\ r < 1/4 \text{ and } r > \beta - 1/2 & \text{if } q = 4r \ , \end{cases} \tag{14}$$

then there exists an $\varepsilon > 0$ such that (12) and (13) hold, and thus $\mathbb{P}_{\mathcal{S}}\left(\tilde{V}_{q_n}(\mathbf{X}) \leq k_n\right) \to 0$.

At this point we can simply follow the arguments of the proof of Proposition 2 almost verbatim. Suppose that $r \leq 1/4$ and $r \geq \beta - 1/2$. Consider the gridpoint $q_n^* \equiv \min_{q \in Q}|q - 4r|$. Since the size of the grid is increasing with $n$, we have $q_n^* = 4r + o(1)$. Therefore:

$$\mathbb{P}_{\mathcal{S}}\left(\max_{q \in Q}\left\{\tilde{V}_q(\mathbf{X})\right\} \leq k_n\right) \leq \mathbb{P}_{\mathcal{S}}\left(\tilde{V}_{q_n^*}(\mathbf{X}) \leq k_n\right) \to 0 \ .$$

The other case, when $r > (1-\sqrt{1-\beta})^2$, follows analogously with $q_n = 1$, since this value is included in the grid $Q$. Then, this result trivially implies that $\tilde{T}(\mathbf{X}) \to 1$ and $(k_n + 1)/\tilde{T}^2(\mathbf{X}) \to 0$ with probability tending to one, and therefore

$$\mathbb{P}_{\mathcal{S}}\left(\tilde{\mathcal{P}}_{\text{perm-hc}}(\mathbf{X}) \leq \alpha\right) \to 1 \ ,$$

completing the proof. □

## 1.5 Proof of Theorem 4

*Proof.* The conservativeness of this test follows by the standard argument already presented in the proof of Theorem 2. For the alternative, the proof relies on the results shown in Theorem 3. The proof of that theorem requires a grid no larger than $n^{o(1)}$, which contains two sequences $q_{1,n}, q_{2,n} \in Q$, such that $q_{1,n} = 4r + o(1)$ if $r < 1/4$ and $q_{2,n} = 1 + o(1)$ otherwise. In the context of our restated statistic, this means our result follows if there exists two sequences $\tau_{1,n}, \tau_{2,n} \in R$ such that:

$$\tau_{1,n} = \mu_0 + \sqrt{\frac{2\sigma_0^2}{t}(1 + o(1))\log(n)}\ ,$$
$$\tau_{2,n} = \mu_0 + \sqrt{\frac{2\sigma_0^2}{t}(4r + o(1))\log(n)}\ .$$

We show that the first sequence exists in $R$; the second sequence then follows analogously by replacing $\sigma_0^2$ by $4r\sigma_0^2$. Note that the requirement for $\tau_{1,n}$ can be rewritten as:

$$\tau_{1,n} - \left(\mu_0 + \sqrt{\frac{2\sigma_0^2}{t}\log(n)}\right) = o\left(\sqrt{\frac{\log(n)}{t}}\right)\ .$$

Since $\sqrt{\log(n)} \to \infty$, there exists an $n_0$ such that for $n \geq n_0$ there exists sequences $i_n^* \in \left\{\frac{k}{\sqrt{t}}\right\}_{k=-\sqrt{t\log(n)}}^{\sqrt{t\log(n)}}$ and $j \in \left\{\frac{k}{\sqrt{\log(n)}}\right\}_{k=0}^{\log(n)}$ such that

$$|i_n^* - \mu_0| \leq \frac{1}{\sqrt{t}}\ , \qquad |j_n^* - \sigma_0| \leq \frac{1}{\sqrt{\log(n)}}\ .$$

Now, defining $\tau_{1,n} = i_n^* + \sqrt{\frac{2(j_n^*)^2}{t}\log(n)}$, we have that:

$$\left|\tau_{1,n} - (\mu_0 + \sqrt{\frac{2}{t}\sigma_0^2\log(n)})\right| \leq |i_n^* - \mu_0| + |j_n^* - \sigma_0|\sqrt{\frac{2}{t}\log(n)} \leq \frac{1+\sqrt{2}}{\sqrt{t}} = o\left(\sqrt{\frac{\log(n)}{t}}\right)\ ,$$

so $\tau_{1,n}$ is sufficiently close to the optimal value. Now, the size of the grid is of order $\mathcal{O}\left(\sqrt{t\log^3(n)}\right)$, and since $t$ is of order $n^{o(1)}$, the grid is not too large. □

## 1.6 Proof of Lemma 2

*Proof.* We have

$$\begin{aligned}\mathbb{E}\left(f(X)\mathbb{1}\left\{X \leq \tau\right\}\right) &= \int_0^\infty \mathbb{P}(f(X)\mathbb{1}\left\{X \leq \tau\right\} > t)\mathrm{d}t \\ &= \int_0^{f(\tau)} \mathbb{P}(f(X) > t)\mathrm{d}t \\ &= \int_0^{f(\tau)} \mathbb{P}(X > f^{-1}(t))\mathrm{d}t \\ &= \int_{-\infty}^{\tau} \mathbb{P}(X > x)f'(x)\mathrm{d}x,\end{aligned}$$

where in the last line we changed the integration variable to $x := f^{-1}(t)$. □

## 1.7 Proof of Lemma 4

*Proof.* Let us start by characterizing the overall sample mean $\overline{X}$. By Chebyshev's inequality we have

$$\overline{X} = \mathbb{E}\left(\overline{X}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{nt}}\right) ,$$

as $n \to \infty$. Now note that, under the alternative hypothesis

$$\mathbb{E}\left(\overline{X}\right) = \frac{|\mathcal{S}|}{n}\mathbb{E}\left(X\right) = \frac{|\mathcal{S}|}{n}\int \frac{x\exp\left(\theta x\right)}{\varphi(\theta)}dF_0(x) ,$$

where $X \sim F_\theta$. A Taylor expansion of the function inside the integral around $\theta = 0$ yields

$$\begin{aligned}\int \frac{x\exp\left(\theta x\right)}{\varphi(\theta)}dF_0(x) &= \int x + x^2\theta + \mathcal{O}(\theta^2)dF_0(x) \\ &= \theta + \mathcal{O}(\theta^2)\end{aligned}$$

Finally $\theta = \mathcal{O}\left(\sqrt{\log(n)/t}\right) \to 0$ since $t = \omega(\log(n))$. Putting all this together yield the first result stated in (i).

For the second result in (i) note that

$$\begin{aligned}\sigma_X^2 = \frac{1}{nt}\sum_{i,j}(X_{ij} - \overline{X})^2 &= \frac{1}{nt}\sum_{i,j}X_{ij}^2 - \overline{X}^2 \\ &= \left(\frac{1}{nt}\sum_{i,j}\mathbb{E}\left(X_{ij}^2\right)\right) + \left(\frac{1}{nt}\sum_{i,j}X_{ij}^2 - \mathbb{E}\left(X_{ij}^2\right)\right) - \overline{X}^2\end{aligned}$$

For the first term we see that

$$\begin{aligned}\frac{1}{nt}\sum_{i,j}\mathbb{E}\left(X_{ij}^2\right) &= \frac{1}{nt}\sum_{i\notin\mathcal{S},j\in[t]}\mathrm{Var}\left(X_{ij}\right)+\frac{1}{nt}\sum_{i\in\mathcal{S},j\in[t]}\mathrm{Var}\left(X_{ij}\right)+\mathbb{E}\left(X_{ij}\right)^2\\ &= \frac{n-|\mathcal{S}|}{n}+\frac{|\mathcal{S}|}{n}\left(1+\mathcal{O}(\theta)\right)\\ &= 1+\mathcal{O}\left(n^{-\beta}\sqrt{\frac{\log n}{t}}\right)\ ,\end{aligned}$$

as $n\to\infty$. In the above the variance of the anomalous streams was characterized with a Taylor expansion of $\theta$ around 0, similarly to what was done for the average term.

For the second term note first that all the moments of $F_0$ are finite, in particular the fourth moment. Therefore by Chebyshev's inequality we have

$$\frac{1}{nt}\sum_{i,j}X_{ij}^2-\mathbb{E}\left(X_{ij}\right)=\mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{nt}}\right)\ .$$

Finally, we know that $\overline{X}^2=\mathcal{O}_{\mathbb{P}}(1/nt)$ when $\beta>1/2$. Putting everything together yields the second result stated in (i).

The argument needed to prove (ii) is the same already used in Arias-Castro et al. (2018), and presented here for completeness.

Letting $x>0$, a union bound gives

$$\begin{aligned}\mathbb{P}_{\mathcal{S}}\left(\max_{i,j}X_{ij}>x\right) &\le \mathbb{P}_{\mathcal{S}}\left(\max_{i\in\mathcal{S},j\in[t]}X_{ij}>x\right)+\mathbb{P}_{\mathcal{S}}\left(\max_{i\notin\mathcal{S},j\in[t]}X_{ij}>x\right)\\ &\le |\mathcal{S}|t(1-F_\theta(x))+(n-|\mathcal{S}|)t(1-F_0(x))\ .\end{aligned}\tag{15}$$

Now, let $c\in(0,\theta_*-\theta)$. We have that:

$$\begin{aligned}1-F_\theta(x) &= \frac{1}{\varphi_0(\theta)}\int_x^\infty \exp\left(\theta u\right)dF_0(u)\\ &= \frac{1}{\varphi_0(\theta)}\int_x^\infty \exp\left((\theta+c)u\right)\exp\left(-cu\right)dF_0(u)\\ &\le \frac{1}{\varphi_0(\theta)}\exp\left(-cx\right)\int_x^\infty \exp\left((\theta+c)u\right)dF_0(u)\\ &\le \frac{\varphi_0(\theta+c)}{\varphi_0(\theta)}\exp\left(-cx\right)\ .\end{aligned}$$

Therefore, with considerable slack, we can take $x=\frac{2}{c}\log(nt)$ guarantee that both terms

in (15) converge to zero. Together with the characterization of $\overline{X}$ in (i) we conclude that

$$\mathbb{P}_{\mathcal{S}}\left(\max_{i,j} X_{ij} - \overline{X} \leq \frac{3}{c}\log(nt)\right) \to 1 \ .$$

To show part (iii) very different argument is needed as this is a lower-bound on the tail probability, rather than an upper bound. The following lemma gives a precise characterization of the tail probability.

**Lemma 6.** *Consider the setting of Lemma 4. Let $i \in \mathcal{S}$ and let $q_n \to q > r$ as $n \to \infty$ and $t = \omega(\log^3 n)$. Then*

$$\mathbb{P}_{\mathcal{S}}\left(Y_i(\mathbf{X}) \geq \sqrt{\frac{2q_n \log n}{t}}\right) = n^{-(\sqrt{q}-\sqrt{r})^2+o(1)} \ .$$

*If $q \leq r$ we have $\mathbb{P}_{\mathcal{S}}\left(Y_i(\mathbf{X}) \geq \sqrt{\frac{2q_n \log n}{t}}\right) = n^{o(1)}$.*

To show (iii) begin by noting that

$$\begin{aligned}
&\mathbb{P}_{\mathcal{S}}\left(\max_{i\in[n]} Y_i(\mathbf{X}) \geq \sqrt{\frac{2(1+\varepsilon)}{t}\log(n)}\right) \\
&\geq \mathbb{P}_{\mathcal{S}}\left(\max_{i\in\mathcal{S}} Y_i(\mathbf{X}) \geq \sqrt{\frac{2(1+\varepsilon)}{t}\log(n)}\right) \\
&= 1 - \left(1 - \mathbb{P}_{\mathcal{S}}\left(Y_i(\mathbf{X}) \geq \sqrt{\frac{2(1+\varepsilon)}{t}\log(n)}\right)\right)^{|\mathcal{S}|} \qquad (16)
\end{aligned}$$

Consider the case where $(\sqrt{1+\varepsilon} - \sqrt{1-\beta})^2 < r < 1+\varepsilon$. Note that, in that case, we can use Lemma 6 with (16) which gives:

$$\begin{aligned}
\mathbb{P}_{\mathcal{S}}\left(\max_{i\in[n]} Y_i(\mathbf{X}) \geq \sqrt{\frac{2(1+\varepsilon)\sigma_0^2}{t}\log(n)}\right) &= 1 - \left(1 - n^{-(\sqrt{1+\varepsilon}-\sqrt{r})^2+o(1)}\right)^{|\mathcal{S}|} \\
&= 1 - \exp\left(n^{1-\beta}\log\left(1 - n^{-(\sqrt{1+\varepsilon}-\sqrt{r})^2+o(1)}\right)\right) \\
&\geq 1 - \exp\left(-n^{1-\beta}n^{-(\sqrt{1+\varepsilon}-\sqrt{r})^2+o(1)}\right) \ ,
\end{aligned}$$

where in last inequality we simply used the fact that $\log(1+x) \geq x$. Finally, provided $(\sqrt{1+\varepsilon} - \sqrt{1-\beta})^2 < r < 1+\varepsilon$ we guarantee that $1 - \beta - (\sqrt{1+\varepsilon} - \sqrt{r})^2 + o(1) > 0$. The statement for $r > 1+\varepsilon$ follows immediately since this probability is monotonically increasing in $r$. To get the statement in (iii) we just need to use this result together with

the characterization of $\overline{X}$, concluding the proof. □

## 1.8 Proof of Lemma 5

*Proof.* Note that $\tilde{p}_q$ and $\tilde{P}_q(\mathbf{X})$ differ in two major aspects; first, $\tilde{p}_q$ depends on the null distribution, while $\tilde{P}_q(\mathbf{X})$ depends on the distribution of permutation stream means - which may be "contaminated" by anomalous observations. Secondly, $\tilde{P}_q(\mathbf{X})$ is random, while $\tilde{p}_q$ is deterministic.

To characterize the "contamination" effect, we define the following quantity corresponding to the probability of a permutation stream exceeding $z_q \equiv \sqrt{2q(\log n)/t}$ when precisely $k$ anomalous observations are sampled in the permutation stream:

$$\tilde{P}'_{k,q}(\mathbf{X}) \equiv \mathbb{P}\left(Y_1(\mathbf{X}^{\pi_k}) \geq z_q \mid \mathbf{X}\right) \ ,$$

where $\pi_k$ is uniformly distributed over the set $\Pi^{(k)}$ independent from $\mathbf{X}$, and the set $\Pi^{(k)} \subseteq \Pi$ is defined as the set of permutations with exactly $k$ observations sampled from anomalous streams in permutation stream with index 1. Specifically:

$$\Pi^{(k)} \equiv \left\{\pi \in \Pi : \sum_{i\in\mathcal{S}} \sum_{j\in[t]} \sum_{h\in[t]} \mathbb{1}\left\{\pi(i,j) = (1,h)\right\} = k\right\} .$$

Note that $\Pi^{(k)}$ does not depend on the data, but merely on the index set $\mathcal{S}$. By definition $\mathbb{E}\left(\tilde{P}'_{0,q}(\mathbf{X})\right) = \tilde{p}_q$. Now consider the following decomposition of our quantity of interest:

$$\tilde{P}_{q_n}(\mathbf{X}) - \tilde{p}_{q_n} = \left(\tilde{P}_{q_n}(\mathbf{X}) - \tilde{P}'_{0,q_n}(\mathbf{X})\right) + \left(\tilde{P}'_{0,q_n}(\mathbf{X}) - \mathbb{E}\left(\tilde{P}'_{0,q_n}(\mathbf{X})\right)\right) \ . \tag{17}$$

Intuitively, bounding the first term amounts to characterizing the effect of "contamination" by anomalous observations when computing $\tilde{P}_{q_n}(\mathbf{X})$. The second term focusses mainly on the random fluctuations (when contamination is not present).

We start with the first term in (17). Intuitively, with high probability, the permutation stream $Y_1(\mathbf{X}^\pi)$ consists solely of nominal observations, especially for small $t$. With modest probability, a few anomalous observations are sampled, but their influence on the ensuing distribution of the stream mean $Y_1(\mathbf{X}^\pi)$ should be minimal, especially in light of the exponential tails of the distributions in question. Finally, sampling a large number of anomalies

might unduly influence the distribution of $Y_1(\mathbf{X}^\pi)$, but the probability of this happening is very small. For our purposes, the effect of the anomalous observations is modest on $Y_1(\mathbf{X}^\pi)$ provided there are no more than $\log(n)$ contaminating samples.

To proceed, we first condition on the number of anomalous observations sampled in the first permutation stream, which allows us to bound the first component in (17) as:

$$\tilde{P}_{q_n}(\mathbf{X}) - \tilde{P}'_{0,q_n}(\mathbf{X}) \leq \sum_{k=1}^{\lfloor \log(n) \rfloor} \tilde{P}'_{k,q_n}(\mathbf{X}) \mathbb{P}\left(\pi \in \Pi^{(k)}\right) + \sum_{k=\lceil \log(n) \rceil}^{t} \mathbb{P}\left(\pi \in \Pi^{(k)}\right) \ . \tag{18}$$

Note that the bound above is only sensible when $\mathbb{P}\left(\pi \in \Pi^{(0)}\right)$ is close to 1. By assuming $t = n^{o(1)}$ this is indeed the case.

To characterize the bound in (18), we start by characterizing the probability $\tilde{P}'_{k,q_n}(\mathbf{X})$. For small enough $k$, we can bound this term in a nontrivial way through the use of Lemma 3 and Lemma 4. Note that $Y_1(\mathbf{X}^{\pi_k})$ arises from two sampling processes; $t-k$ samples from the null streams, and $k$ samples from the anomalous streams. We will bound the contribution of the anomalous streams crudely by their maximum. Define $U_m(\mathbf{X})$ as the sum of a sample of size $m$ without replacement from the nominal observations of $\mathbf{X}$, i.e. from the set $\mathbf{X}_0 \equiv \{X_{ij} : i \notin \mathcal{S}, j \in [t]\}$. Now, we can bound:

$$\begin{aligned} \tilde{P}'_{k,q_n}(\mathbf{X}) &\equiv \mathbb{P}\left(Y_1(\mathbf{X}^{\pi_k}) \geq z_{q_n} \mid \mathbf{X}\right) \\ &\leq \mathbb{P}\left(\frac{1}{t}\left(U_{t-k}(\mathbf{X}) + k \max_{i,j}\{X_{ij}\}\right) \geq z_{q_n} \,\middle|\, \mathbf{X}\right) \ . \end{aligned} \tag{19}$$

To continue, we will use the Bernstein bound of Lemma 3. Define $\overline{X}_0$ and $\sigma^2_{\mathbf{X}_0}$ the sample mean and variance of $\mathbf{X}_0$. Direct application of the lemma results in complicated expressions, so we first define for convenience:

$$d_k(\mathbf{X}) = \frac{\sigma_{\mathbf{X}_0}}{\sigma_{\mathbf{X}}}\left(\sqrt{\frac{t-k}{t}} - \overline{X}_0\sqrt{\frac{t-k}{2\sigma_{\mathbf{X}_0} q \log(n)}} + \sqrt{\frac{t-k}{2\sigma_{\mathbf{X}_0} q \log(n)}}\frac{k}{t-k}\max_{i,j}\{X_{ij}\} + \mathcal{O}\left(\frac{1}{t}\right)\right) \ .$$

While this term is complex, it is ultimately a nuisance term as one should note that, since $t = \omega(\log(n)^3)$ and $k \leq \log(n)$, Lemma 4 applied to the set $\mathbf{X}_0$ implies there exists a sequence $g_n \to 0$ such that

$$\mathbb{P}\left(|d_k(\mathbf{X})| \leq 1 + g_n\right) \to 1 \ . \tag{20}$$

Now, we can continue from (19) as:

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{t}\left(U_{t-k}(\mathbf{X}) + k\max_{i,j}\{X_{ij}\}\right) \geq z_{q_n} \;\middle|\; \mathbf{X}\right) \\
&= \mathbb{P}\left(\frac{U_{t-k}(\mathbf{X})}{t-k} \geq \frac{t}{t-k}z_{q_n} - \frac{k}{t-k}\max_{i,j}\{X_{ij}\} \;\middle|\; \mathbf{X}\right) \\
&= \mathbb{P}\left(\frac{U_{t-k}(\mathbf{X})}{t-k} \geq \overline{X}_0 + d_k(\mathbf{X})\sqrt{\frac{2\sigma^2_{\mathbf{X}_0}q_n}{t-k}\log(n)} \;\middle|\; \mathbf{X}\right) \\
&\leq \exp\left(-\frac{2d_k^2(\mathbf{X})\sigma^2_{\mathbf{X}_0}q_n\log(n)}{2\sigma^2_{\mathbf{X}_0} + \frac{2}{3}(\max_{i\notin\mathcal{S},j\in[t]}\{X_{ij}\} - \overline{X}_0)\sqrt{\frac{2d_k^2(\mathbf{X})\sigma^2_{\mathbf{X}_0}q_n}{t-k}\log(n)}}\right) \\
&= \exp\left(-\frac{d_k^2(\mathbf{X})q_n\log(n)}{1 + \frac{1}{3}(\max_{i\notin\mathcal{S},j\in[t]}\{X_{ij}\} - \overline{X}_0)\sqrt{\frac{2d_k^2(\mathbf{X})q_n}{(t-k)\sigma^2_{\mathbf{X}_0}}\log(n)}}\right) \equiv B(\mathbf{X}) \;, \qquad (21)
\end{aligned}$$

where the inequality is due to Lemma 3. Now, due to the result in (20), as well as application of Lemma 4 to the set $\mathbf{X}_0$, we have that:

$$\mathbb{P}_{\mathcal{S}}\left(B(\mathbf{X}) \leq \exp\left(-\frac{q_n(1+o(1))\log(n)}{1+o(1)}\right)\right) \to 1 \;.$$

Since this high probability upper bound can be restated as $n^{-q+o(1)}$, we ultimately obtain that

$$\mathbb{P}_{\mathcal{S}}\left(\tilde{P}_{k,q_n}(\mathbf{X}) \leq n^{-q+o(1)}\right) \to 1 \;. \qquad (22)$$

Next, we characterize the probability $\mathbb{P}\left(\pi \in \Pi^{(k)}\right)$. Note that $\mathbb{P}\left(\pi \in \Pi^{(k)}\right) = \mathbb{P}\left(H = k\right)$, with $H$ a hypergeometric distribution with population size $nt$, number of success states $st$ and sample size $t$. A coupling argument can be used to show $H$ is stochastically dominated by a binomial random variable with $t$ trials with success probability $st/nt$, and $H$ stochastically dominates a binomial random variable with $t$ trials with success probability $(st-t)/nt$. For

convenience, denote $p_L = (st-t)/nt$, and $p_U = st/nt$. We have that:

$$\begin{aligned}\sum_{k=1}^{\lfloor \log(n) \rfloor} \mathbb{P}\left(\pi \in \Pi^{(k)}\right) &= \mathbb{P}\left(H \leq \lfloor \log(n) \rfloor\right) - \mathbb{P}\left(H = 0\right) \\ &\leq \mathbb{P}\left(\text{Binomial}(t, p_L) \leq \lfloor \log(n) \rfloor\right) - \mathbb{P}\left(\text{Binomial}(t, p_U) = 0\right) \\ &\leq (1-p_L)^t - (1-p_U)^t + \sum_{k=1}^{\lfloor \log(n) \rfloor} \binom{t}{k} p_L^k \ .\end{aligned}$$

Now, the first part of the expression above can be upper bounded as:

$$\begin{aligned}(1-p_L)^t - (1-p_U)^t &= \left(1 - \frac{st-t}{nt}\right)^t - \left(1 - \frac{s}{n}\right)^t \\ &= \left(1 - \frac{s}{n}\right)^t \left(\left(1 + \frac{1}{n-s}\right)^t - 1\right) \\ &\leq \left(1 + \frac{1}{n-s}\right)^t - 1 \\ &\leq \frac{1}{1 - \frac{t}{n-s}} - 1 \\ &= \frac{t}{n-s-t} = n^{-1+o(1)} \ ,\end{aligned}$$

where the second inequality is due to Bernoulli's inequality. The second part of the expression can be upper bounded as:

$$\begin{aligned}\sum_{k=1}^{\lfloor \log(n) \rfloor} \binom{t}{k} p_L^k &\leq \sum_{k=1}^{\lfloor \log(n) \rfloor} \left(\frac{et}{k}\right)^k p_L^k \\ &\leq \sum_{k=1}^{\lfloor \log(n) \rfloor} n^{-k\beta + k\frac{\log(t)}{\log(n)} + o(1)} \\ &\leq n^{-\beta + \frac{\log(t)}{\log(n)} + o(1)} + \sum_{k=2}^{\lfloor \log(n) \rfloor} n^{-k\beta + k\frac{\log(t)}{\log(n)} + o(1)} \\ &\leq n^{-\beta + o(1)} + \sum_{k=2}^{\lfloor \log(n) \rfloor} n^{-k\frac{\beta}{2} + o(1)} \\ &\leq n^{-\beta + o(1)} \ ,\end{aligned}$$

where the first inequality is due to Stirling's inequality, and the fourth inequality is due to $t = n^{o(1)}$ such that $\log(t)/\log(n) = o(1)$ and for sufficiently large $n$, we thus have $\log(t)/\log(n) \leq$

$\beta/2$. We can conclude that for any $\varepsilon > 0$,

$$\sum_{k=1}^{\lfloor \log(n) \rfloor} \mathbb{P}\left(\pi \in \Pi^{(k)}\right) \leq n^{-\beta+o(1)} \ . \tag{23}$$

For the other probability term in (18), we have that:

$$\begin{aligned}
\sum_{k=\lceil \log(n) \rceil}^{t} \mathbb{P}\left(\pi \in \Pi^{(k)}\right) &= \mathbb{P}\left(H \geq \lceil \log(n) \rceil\right) \leq \mathbb{P}\left(\text{Binomial}(t, p_U) \geq \lceil \log(n) \rceil\right) \\
&\leq \sum_{k=\lceil \log(n) \rceil}^{t} \left(\frac{et}{k}\right)^k \left(\frac{s}{n}\right)^k \\
&= \sum_{k=\lceil \log(n) \rceil}^{t} n^{-k\beta + k\frac{\log(t)}{\log(n)} + o(1)} \\
&\leq \sum_{k=\lceil \log(n) \rceil}^{t} n^{-k\beta + k\frac{\beta}{2} + o(1)} = \sum_{k=\lceil \log(n) \rceil}^{t} n^{-k\frac{\beta}{2} + o(1)} \\
&\leq t n^{-\log(n)\frac{\beta}{2}} \leq n^{-\beta - q} \ ,
\end{aligned} \tag{24}$$

where we have assumed that $t = n^{o(1)}$, and subsequently that the fraction $\frac{\log(t)}{\log(n)} \leq \frac{\beta}{2}$ for $n$ sufficiently large. We have also used that $\beta > 1/2$ and $q \leq 1$.

Putting the results from (22), (23), and (24) in (18), we have that there exists a deterministic sequence $g_n \to 0$ such that:

$$\mathbb{P}_{\mathcal{S}}\left(\tilde{P}_{q_n}(\mathbf{X}) - \tilde{P}'_{0,q_n}(\mathbf{X}) \leq n^{-\beta-q+g_n}\right) \to 1 \ . \tag{25}$$

For the second term in (17), we must show that $\tilde{P}'_{0,q}(\mathbf{X})$ concentrates well around its mean. We use Chebyshev's inequality, requiring a characterization of the variance of $\tilde{P}'_{0,q}(\mathbf{X})$, obtained by carefully characterizing the dependency between two permutation streams. In many cases, these streams do not share any observations, and thus are independent. To characterize this rigorously, let $\pi_*$ be some arbitrary fixed permutation from $\Pi^{(0)}$. We then partition $\Pi^{(0)}$ in sets $\{\Pi_k^{(0)}(\pi_*)\}_{k=0}^t$ as follows: we have $\pi \in \Pi_k^{(0)}(\pi_*)$ if $\pi$ permutes precisely $k$ of the same coordinates as $\pi_*$ to the first stream. Specifically:

$$\Pi_k^{(0)}(\pi_*) \equiv \left\{ \pi \in \Pi : \sum_{j=1}^{t} \mathbb{1}\left\{\pi^{-1}(1,j) = \pi_*^{-1}(1,j)\right\} = k \right\} \ .$$

In particular, note that if $\pi \in \Pi_0^{(0)}(\pi_*)$ with $\pi_*$ some fixed permutation, this means that the

random variables $Y_1(\mathbf{X}^{\pi})$ and $Y_1(\mathbf{X}^{\pi_*})$ are independent (given the two permutations). Note that, since $n \gg t$, intuitively $\Pi \approx \Pi_0(\pi_*)$ for any permutation $\pi_*$, such that one should expect a very weak dependency between $Y_1(\mathbf{X}^{\pi})$ and $Y_1(\mathbf{X}^{\pi_*})$ when $\pi \in \Pi$ uniformly at random.

Now, we first find the second moment of $\tilde{P}'_{0,q}(\mathbf{X})$. Let $\pi_1$ and $\pi_2$ be permutations uniformly and independently at random from $\Pi^{(0)}$, such that:

$$\begin{aligned}
\mathbb{E}\left(\tilde{P}'_{0,q}(\mathbf{X})^2\right) &= \mathbb{E}\Big[\mathbb{E}\Big(\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi_1}) \geq z_q\right\} \,\Big|\, \mathbf{X}\Big)\mathbb{E}\Big(\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi_2}) \geq z_q\right\} \,\Big|\, \mathbf{X}\Big)\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big(\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi_1}) \geq z_q\right\}\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi_2}) \geq z_q\right\} \,\Big|\, \mathbf{X}\Big)\Big] \\
&= \mathbb{E}\Big[\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi_1}) \geq z_q\right\}\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi_2}) \geq z_q\right\}\Big] \\
&= \sum_{\pi\in\Pi^{(0)}}\sum_{k=0}^{t}\sum_{\xi\in\Pi_k^{(0)}(\pi)} \mathbb{E}\Big[\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi_1}) \geq z_q\right\}\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi_2}) \geq z_q\right\} \,\Big|\, \pi_1 = \xi, \pi_2 = \pi\Big] \\
&\qquad \cdot \mathbb{P}\left(\pi_1 = \xi\right)\mathbb{P}\left(\pi_2 = \pi\right) \ . \quad (26)
\end{aligned}$$

At this point, it is convenient to look at the summands of $k$ in (26) individually. First, note that for $k = 0$ we have that, due to independence:

$$\begin{aligned}
&\sum_{\pi\in\Pi^{(0)}}\sum_{\xi\in\Pi_0^{(0)}(\pi)} \mathbb{E}\Big[\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi_1}) \geq z_q\right\}\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi_2}) \geq z_q\right\} \,\Big|\, \pi_1 = \xi, \pi_2 = \pi\Big]\mathbb{P}\left(\pi_1 = \xi\right)\mathbb{P}\left(\pi_2 = \pi\right) \\
&= \sum_{\pi\in\Pi^{(0)}}\sum_{\xi\in\Pi_0^{(0)}(\pi)} \mathbb{E}\Big[\mathbb{1}\left\{Y_1(\mathbf{X}^{\xi}) \geq z_q\right\}\Big]\mathbb{E}\Big[\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi}) \geq z_q\right\}\Big]\mathbb{P}\left(\pi_1 = \xi\right)\mathbb{P}\left(\pi_2 = \pi\right) \\
&\leq \sum_{\pi\in\Pi^{(0)}}\left\{\mathbb{E}\Big[\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi}) \geq z_q\right\}\Big]\mathbb{P}\left(\pi_2 = \pi\right)\sum_{\xi\in\Pi^{(0)}}\mathbb{E}\Big[\mathbb{1}\left\{Y_1(\mathbf{X}^{\xi}) \geq z_q\right\}\Big]\mathbb{P}\left(\pi_1 = \xi\right)\right\} \\
&= \sum_{\pi\in\Pi^{(0)}}\left\{\mathbb{E}\Big[\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi}) \geq z_q\right\}\Big]\mathbb{P}\left(\pi_2 = \pi\right)\mathbb{E}\left(\tilde{P}'_{0,q}(\mathbf{X})\right)\right\} \\
&= \mathbb{E}\left(\tilde{P}'_{0,q}(\mathbf{X})\right)^2 \ .
\end{aligned}$$

Now, before proceeding to bound the term in (26) for $k \geq 1$, we first define

$$\rho_k \equiv \mathbb{P}\left(\xi \in \Pi_k^{(0)}(\pi)\right) \ ,$$

when $\xi$ is sampled uniformly at random from $\Pi^{(0)}$, and $\pi \in \Pi^{(0)}$ arbitrarily and fixed.

Note that this is a hypergeometric probability; the probability corresponds to choosing $t$ indexes, of which $k$ indexes should match the $t$ indexes placed in the first stream by the permutation $\pi$, out of the $(n-s)t$ indexes in total, without replacement. To characterize this probability, we use a stochastic domination argument like before; note that a hypergeometric random variable $H$ with $(n-s)t$ total states, with $t$ success states and a sample size of $t$, is stochastically dominated by a binomial random variable with $t$ draws with success probability $t/((n-s)t)$. Therefore:

$$\begin{aligned}\sum_{k=1}^{t}\rho_k = \mathbb{P}\left(H \geq 1\right) &\leq \mathbb{P}\left(\text{Bin}\left(t, \frac{1}{n-s}\right) \geq 1\right) \leq \sum_{k=1}^{t}\binom{t}{k}\left(\frac{1}{n-s}\right)^k \\ &\leq \sum_{k=1}^{t}\left(\frac{et}{nk}\right)^k\left(1+\mathcal{O}\left(\frac{s}{n}\right)\right)^k \leq \sum_{k=1}^{t} n^{-k+k\frac{\log(t)}{\log(n)}+o(1)} \\ &\leq n^{-1+o(1)} + \sum_{k=2}^{t} n^{-k(1-\varepsilon)+o(1)} \leq n^{-1+o(1)}\ , \end{aligned} \tag{27}$$

where the fifth inequality holds for any $\varepsilon > 0$ since $t = n^{o(1)}$, and thus for sufficiently large $n$ we have $\log(t)/\log(n) \leq \varepsilon$. Then, for the summands $k \geq 1$ in (26), we now have that:

$$\begin{aligned}&\sum_{\pi\in\Pi^{(0)}}\sum_{\xi\in\Pi_k^{(0)}(\pi)}\mathbb{E}\Big[\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi_1}) \geq z_q\right\}\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi_2}) \geq z_q\right\} \,\Big|\, \pi_1 = \xi, \pi_2 = \pi\Big]\mathbb{P}\left(\pi_1 = \xi\right)\mathbb{P}\left(\pi_2 = \pi\right) \\ &\leq \sum_{\pi\in\Pi^{(0)}}\mathbb{E}\Big[\mathbb{1}\left\{Y_1(\mathbf{X}^{\pi_2}) \geq z_q\right\} \,\Big|\, \pi_2 = \pi\Big]\mathbb{P}\left(\pi_2 = \pi\right)\sum_{\xi\in\Pi_k^{(0)}(\pi)}\mathbb{P}\left(\pi_1 = \xi\right) \\ &= \rho_k\mathbb{E}\left(\tilde{P}'_{0,q}(\mathbf{X})\right)\ .\end{aligned}$$

For the second moment, we therefore have that:

$$\mathbb{E}\left(\tilde{P}'_{0,q}(\mathbf{X})^2\right) \leq \mathbb{E}\left(\tilde{P}'_{0,q}(\mathbf{X})\right)^2 + \mathbb{E}\left(\tilde{P}'_{0,q}(\mathbf{X})\right)\sum_{k=1}^{t}\rho_k = \mathbb{E}\left(\tilde{P}'_{0,q}(\mathbf{X})\right)^2 + \tilde{p}_q\sum_{k=1}^{t}\rho_k\ ,$$

where the equality holds by definition of $\tilde{P}'_{0,q}(\mathbf{X})$. Now, letting $q_n \to q$, the variance of $\tilde{P}'_{0,q_n}(\mathbf{X})$ can be bounded by:

$$\text{Var}\left(\tilde{P}'_{0,q_n}(\mathbf{X})\right) \leq \tilde{p}_{q_n}\sum_{k=1}^{t}\rho_k \leq n^{-1-q+o(1)}\ ,$$

where the second inequality is due to Lemma 6 and the result in (27). Now, Chebyshev's

inequality implies that, for any $\varepsilon > 0$, we have:

$$\mathbb{P}\left(\tilde{P}'_{0,q_n}(\mathbf{X}) - \mathbb{E}\left(\tilde{P}'_{0,q_n}(\mathbf{X})\right) \geq n^{-\frac{1+q}{2}+\varepsilon}\right) \leq \frac{\mathrm{Var}\left(\tilde{P}'_{0,q_n}(\mathbf{X})\right)}{n^{-1-q+2\varepsilon}} \leq n^{-\varepsilon} \to 0 \ . \tag{28}$$

We have now bounded both components of Equation (17). Combining our results of (25) and (28) implies that, for any $\varepsilon > 0$, there exists a sequence $g_n \to 0$ such that:

$$\mathbb{P}\left(\tilde{P}_q(\mathbf{X}) - \tilde{p}_q \leq n^{\max\{-\beta-q, n^{-\frac{1+q}{2}}\}+\varepsilon+g_n}\right) \to 1 \ ,$$

concluding the proof. □

## 1.9 Proof of Lemma 6

To streamline the presentation let $W_1, \ldots, W_t$ be i.i.d. with distribution $F_\theta$ and denote by $\varphi_\theta(x)$ the moment generating function of $F_\theta$. Define also $\tau = \sqrt{(2q_n/t)\log n}$. We start by getting an upper bound for the said probability when $r < q$. A simple Chernoff bounding argument yields

$$\mathbb{P}_{\mathcal{S}}\left(\frac{1}{t}\sum_{j\in[t]} W_j \geq \tau\right) \leq \exp\left(-t\left[\sup_{\lambda\in[0,\theta_*-\theta)}\{\lambda\tau - \log(\varphi_\theta(\lambda))\}\right]\right) \ . \tag{29}$$

We must now characterize $\varphi_\theta(\lambda)$. First note that $\varphi_\theta(\lambda) = \varphi_0(\lambda+\theta)/\varphi_0(\theta)$. Now, we develop a Taylor expansion of $\varphi_0(\lambda)$ around $\lambda = 0$, as we did in Equation (1). Note that $F_0$ has zero mean and unit variance. We obtain:

$$\varphi_0(\lambda) = 1 + \frac{\lambda^2}{2} + \mathcal{O}(\lambda^3) \ , \tag{30}$$

as $\lambda \to 0$. A similar expansion can be developed for $\varphi_0(\lambda+\theta)$ around $\lambda+\theta = 0$. Combining all this yields

$$\varphi_\theta(\lambda) = \frac{1 + 1/2(\lambda+\theta)^2 + \mathcal{O}((\lambda+\theta)^3)}{1 + \theta^2/2 + \mathcal{O}(\theta^3)} = 1 + \theta\lambda + \frac{\lambda^2}{2} + \mathcal{O}((\lambda+\theta)^3) \ ,$$

as both $\lambda, \theta \to 0$, where we used a Taylor expansion for the fraction around $\theta^2/2 + \mathcal{O}(\theta^3) = 0$. This suggests the choice $\lambda^* = \tau - \theta$, which is positive provided $n$ is large enough since $r < q$.

This choice yields the bound

$$\sup_{\lambda\in[0,\theta_*-\theta)}\{\lambda\tau-\log(\varphi_\theta(\lambda))\}\geq\tau^2-\log(\varphi_\theta(\lambda))$$
$$\geq\frac{1}{2}(\tau-\theta)^2+\mathcal{O}\left(\tau^3\right)\ ,$$

since $\tau>\theta$ and we used the basic inequality $\log(1+x)\leq x$. When $t=\omega(\log^3 n)$ the first term dominates, and therefore we conclude that

$$\mathbb{P}_{\mathcal{S}}\left(\frac{1}{t}\sum_{j\in[t]}W_j\geq\tau\right)\leq\exp\left(-(\sqrt{q_n}-\sqrt{r})^2(\log n)+o(1)\right)$$
$$=n^{-(\sqrt{q}-\sqrt{r})^2+o(1)}\ .$$

To lower-bound the probability in (29) we use a tilting argument. Let $\theta_\tau$ be such that $F_{\theta_\tau}$ has mean $\tau$. Such a choice exists for $n$ large enough and necessarily $\theta_\tau>\theta$ for large $n$, since $r<q$. Define $\tilde{W}_1,\ldots,\tilde{W}_t$ to be i.i.d. with distribution $F_{\theta_\tau}$. Then

$$\mathbb{P}_{\mathcal{S}}\left(\frac{1}{t}\sum_{j\in[t]}W_j\geq\tau\right)=\int\mathbb{1}\left\{\frac{1}{t}\sum_{j\in[t]}w_j\geq\tau\right\}dF_\theta(w_1)\cdots dF_\theta(w_t)$$
$$=\int\mathbb{1}\left\{\frac{1}{t}\sum_{j\in[t]}w_j\geq\tau\right\}\prod_{j=1}^{t}\exp\left(\theta w_j-\log\varphi_0(\theta)\right)dF_0(w_1)\cdots dF_0(w_t)$$
$$=\int\mathbb{1}\left\{\frac{1}{t}\sum_{j\in[t]}w_j\geq\tau\right\}\prod_{j=1}^{t}\exp\left((\theta-\theta_\tau)w_j-\log\left(\frac{\varphi_0(\theta)}{\log\varphi_0(\theta_\tau)}\right)\right)dF_{\theta_\tau}(w_1)\cdots dF_{\theta_\tau}(w_t)$$
$$=\left(\frac{\varphi_0(\theta_\tau)}{\varphi_0(\theta)}\right)^t\mathbb{E}\left(\mathbb{1}\left\{\frac{1}{t}\sum_{j\in[t]}\tilde{W}_j\geq\tau\right\}\exp\left(-(\theta_\tau-\theta)\sum_{j\in[t]}\tilde{W}_j\right)\right)\ .$$

With this change of measure we can conveniently use the central limit theorem to get a

meaningful bound. Begin by noting that

$$\begin{aligned}
&\mathbb{P}_{\mathcal{S}}\left(\frac{1}{t}\sum_{j\in[t]} W_j \geq \tau\right) \\
&\geq \left(\frac{\varphi_0(\theta_\tau)}{\varphi_0(\theta)}\right)^t \mathbb{E}\left(\mathbb{1}\left\{0 \leq \frac{1}{\sqrt{t}}\sum_{j\in[t]}\frac{\tilde{W}_j - \tau}{\sigma_{\theta_\tau}} \leq 1\right\}\exp\left(-(\theta_\tau-\theta)\sum_{j\in[t]}\tilde{W}_j\right)\right) \\
&\geq \left(\frac{\varphi_0(\theta_\tau)}{\varphi_0(\theta)}\right)^t \exp\left(-(\theta_\tau-\theta)(t\tau+\sqrt{t}\sigma_{\theta_\tau})\right)\mathbb{P}\left(0 \leq \frac{1}{\sqrt{t}}\sum_{j\in[t]}\frac{\tilde{W}_j - \tau}{\sigma_{\theta_\tau}} \leq 1\right) \ ,
\end{aligned}$$

where $\sigma^2_{\theta_\tau}$ denotes the variance of $F_{\theta_\tau}$. By the central limit theorem we know that

$$\frac{1}{\sqrt{t}}\sum_{j\in[t]}\frac{\tilde{W}_j - \tau}{\sigma_{\theta_\tau}}$$

converges in distribution to a standard normal distribution and therefore the probability in the expression above converges to $\Phi(1) - \Phi(0) \approx 0.34 > 1/4$. We conclude that, for $n$ large enough

$$\mathbb{P}_{\mathcal{S}}\left(\frac{1}{t}\sum_{j\in[t]} W_j \geq \tau\right) \geq \frac{1}{4}\left(\frac{\varphi_0(\theta_\tau)}{\varphi_0(\theta)}\right)^t \exp\left(-(\theta_\tau-\theta)(t\tau+\sqrt{t}\sigma_{\theta_\tau})\right) \ .$$

To control the remaining terms recall that $\varphi_0(\lambda) = 1 + \lambda^2/2 + \mathcal{O}(\lambda^3)$ as $\lambda \to 0$ (see Equation (30)). Note also that $\tau = \theta_\tau + \mathcal{O}(\theta_\tau^2)$, which implies (after some manipulation) that $\theta_\tau = \tau + \mathcal{O}(\tau^2)$. Finally, note that both $\tau$ and $\theta$ have the same order of magnitude. Putting all this together we conclude that

$$\log\left(\left(\frac{\varphi_0(\theta_\tau)}{\varphi_0(\theta)}\right)^t\right) = \frac{t}{2}\left(\tau^2 - \theta^2 + \mathcal{O}(\theta^3)\right) \ .$$

For the other term note that $\sigma_{\theta_\tau} = 1 + o(1)$, and therefore $\sigma_{\theta_\tau}/\sqrt{t} = o(\theta)$. This implies that

$$-(\theta_\tau-\theta)(t\tau+\sqrt{t}\sigma_{\theta_\tau}) = -t\left(\tau(\tau-\theta) + \mathcal{O}(\theta^3)\right) \ .$$

In conclusion

$$\mathbb{P}_{\mathcal{S}}\left(\frac{1}{t}\sum_{j\in[t]} W_j \geq \tau\right) \geq \frac{1}{4}\exp\left(-\frac{t}{2}\left((\tau-\theta)^2 + \mathcal{O}(\theta^3)\right)\right) \ .$$

When $t = \omega(\log^3 n)$ the term $(\tau-\theta)^2$ term dominates, and we see we get the asymptotic behavior as in the upper bound, concluding the proof.

For the case $r \geq q$ we see that necessarily $\mathbb{P}\left(\frac{1}{t}\sum_{j\in[t]} W_j \geq \tau\right) \geq n^{o(1)}$, so the tail probability

cannot be extremely small. In fact, when $r > q$ this probability will be lower bounded by a constant.

### 1.10 Proof of Corollary 1

*Proof.* To prove this corollary we simply map the original dataset to a new dataset with short streams, for which we can apply Theorem 4. Partition the set $[t]$ into $\tilde{t}$ sets of size $k$ such that $\tilde{t} \equiv \frac{t}{k} = n^{o(1)}$ and $\tilde{t} = \omega(\log^3(n))$. For simplicity, we assume $t$ is divisible by $k$ and define

$$\tilde{X}_{ij} \equiv \frac{1}{\sqrt{k}} \sum_{\ell=(j-1)k+1}^{jk} X_{i\ell} \ .$$

Recall that $X_{ij}$ belongs to a natural exponential family with natural parameter $\theta_i$. Note that the distribution of $\tilde{X}_{ij}$ also belongs to a natural exponential family. Particularly, let $\tilde{F}_0$ be the distribution of $\tilde{X}_{ij}$ when $\theta_i = 0$. Then the density of $\tilde{X}_{ij}$ with respect to $\tilde{F}_0$ is given by $\exp\left(\tilde{\theta}_i x - \log(\tilde{\varphi}_0(\tilde{\theta}_i))\right)$ where $\tilde{\varphi}_0(\theta) \equiv \varphi_0(\theta/\sqrt{k})^k$ is the moment generating function of $\tilde{F}_0$ and $\tilde{\theta}_i \equiv \theta_i\sqrt{k}$. Note that $\tilde{F}_0$ has variance $\sigma_0^2$ and so, using the parameterization in Theorem 4 we have for $i \in \mathcal{S}$

$$\tilde{\theta}_i = \theta\sqrt{k} = \sqrt{2r/(\sigma_0^2 t)\log(n)}\sqrt{k} = \sqrt{2r/(\sigma_0^2\tilde{t})\log(n)} \ .$$

We can now apply the test as in Theorem 4 on the set of observations $\tilde{\mathbf{X}} \equiv \{\tilde{X}_{ij} : i \in [n], j \in [\tilde{t}]\}$. Since $\tilde{t} = n^{o(1)}$ and $\tilde{t} = \omega(\log^3(n))$, Theorem 4 implies the test has power converging to one provided $r > \rho^*(\beta)$. If $t$ is not divisible by $k$ one can simply ignore the last observations in each stream of the original data and proceed as above. □

# 2 An analysis of daily COVID-19 diagnoses across municipalities in the Netherlands

To showcase another possible application of our methodology we consider a way to monitor the number of new daily diagnoses of COVID-19, aiming to quickly identify localized outbreaks. During the current COVID-19 pandemic countries experienced successive waves

with large numbers of cases, interspersed with periods with more stable disease dynamics, typically due to measures put in place to limit its spread. In the latter regime it is of high importance for policy makers to quickly detect signs of new impending outbreaks.

One possible way to proceed is to monitor the daily number of diagnoses per capita in each separate municipality of the country over a short time frame, and test if a (small) subset of municipalities has a higher-than-usual number of cases. As such, the framework we have introduced in this paper can be useful here to detect signs of local outbreaks, as these would lead to rejection of our null hypothesis (1). A municipality may turn anomalous during the observed stream and not right from the start. However, recall that our methodology still has some power even when anomalous streams are only partially affected.

Within small time windows, it may be sensible (to some extent) to apply the proposed methodology directly to the raw data. Nevertheless, we also consider a more sophisticated approach by first fitting a model to small time windows of the raw data and then analyzing the residuals instead, as explained in Section 1. Note, however, that our methodology is in principle not suitable for serially dependent data, and validity of the conclusions based on this residual analysis hinge crucially on the validity of the fitted model. In addition, estimated residuals are dependent (but usually only weakly provided an adequate model is chosen). In rigor, one should carefully address the influence of this dependency on the validity of the test conclusions, but this is outside of the scope of this manuscript.

We consider data from The Netherlands. In this country, it is not unreasonable to assume municipalities are sufficiently comparable: the country is very small and population density is not too different among municipalities. We use data on newly diagnosed COVID-19 cases per 100.000 inhabitants from 13th of March up until 10th of August 2020, for each of the $n = 355$ municipalities of the Netherlands. This data was processed from two sources; the data on the number of diagnoses (uncorrected for municipality population) was retrieved from the Dutch national institute for public health and the environment (RIVM) at https://data.rivm.nl[1],

[1]The specific hyperlink to this data is https://data.rivm.nl/geonetwork/srv/dut/catalog.search#

and the data on municipality population was retrieved from the Dutch central agency for statistics (CBS) at https://opendata.cbs.nl[2]. These were combined to obtain the number of newly diagnosed COVID-19 cases per 100.000 inhabitants. Let $i \in \{1, \ldots, 355\}$ denote a municipality and $j \in \{1, \ldots, 151\}$ denote a day in the period above. Then $y_{i,j}$ denotes the *daily rate of new cases per municipality* (daily-rate, for short). Specifically, the ratio between the number of new positive diagnoses on that day divided by the number of inhabitants of the municipality (as a multiplicative factor of 100.000). To give some insight into this data, we depict the number of new cases per municipality in each month in Figure 1. As can be seen, some municipalities clearly have a large number of cases when compared to the others, but they do not stay so large consistently throughout the months considered.

As we are interested in quickly detecting outbreaks our analysis focuses on very short time frames, namely windows of five consecutive days. This is motivated by the knowledge of the incubation time of the disease, believed to be on average around 5 days, and typically between two and fourteen days (Lauer et al., 2020). Within such a short time frame both the independence and stationarity assumptions might not be terribly unreasonable under the null; in a stable regime without local outbreaks, relatively few cases are distributed somewhat evenly over the population, and those infectious individuals tend not to infect many others (within that limited time frame). For larger windows of time one naturally expects the validity of such assumptions to be more questionable. That being said, within a five-day window there might be some amount of dependency and non-stationarity across time, even within the null streams. With this in mind an option is to attempt to capture such global trends and dependencies, and monitor the residual errors of such a model instead, as suggested in Section 1. Both approaches are discussed below.

Formally, our analysis and results pertain a window of $t = 5$ consecutive days, starting on day $w \in \{1, \ldots, 147\}$. We present results for all the possible windows. In short, the raw

/metadata/1c0fcd57-1102-4620-9cfa-441e93ea5604

[2]The specific hyperlink to this data is https://opendata.cbs.nl/statline/?dl=2096B#/CBS/nl/dataset/70072NED/table

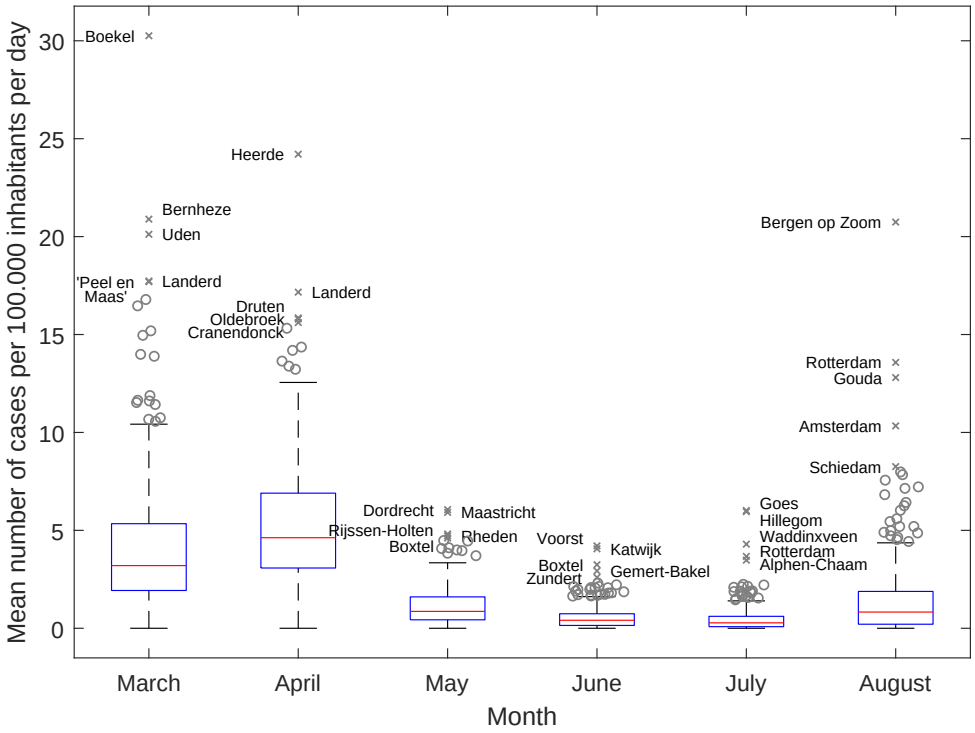



Figure 1: Boxplots depicting the normalized number of monthly cases per municipality. For each municipality, the total number of monthly cases per 100.000 inhabitants is normalized by the number of days in the month. The five municipalities with the highest mean number of cases are depicted with a cross and labeled. Data in March pertains only the 13th up until the 31st of March, and August only pertains the 1st up until the 10th of August.

observations for the window indexed by $w$ are $\left(x_{ij}^{(w)} : i \in [n], j \in [t]\right)$ where $x_{i,j}^{(w)} = y_{i,j+w-1}$. Taking into account the short time-windows and the first order epidemic dynamics a simple but sensible model to consider for this data is an AR(1) model, as suggested in Shtatland (2007, 2008). Note, however, that application of our methodology is crucially dependent on the validity of the model, but for presentation purposes we stick with this relatively simple model. Specifically, the model assumes that for null streams the observations $x_{i,j}^{(w)}$ are obtained as a sample from

$$X_{i,j}^{(w)} - \mu^{(w)} = a^{(w)} \left( X_{i,j-1}^{(w)} - \mu^{(w)} \right) + \varepsilon_{i,j}^{(w)} \ ,$$

where $i \in [n]$, $j \in [t]$, $a^{(w)}, \mu^{(w)} \in \mathbb{R}$ are (unknown) parameters of the model (common to all null streams), and $\varepsilon_{i,j}^{(w)}$ are i.i.d. samples from an unknown zero-mean distribution. Despite its simplicity, this model can capture some of the epidemic dynamics when applied to very

short time frames, in contrast with more sophisticated epidemiological models (like the ones described in Held et al. (2019)).

The first step in this approach is to estimate the unknown parameters of the model. Given that we do not have knowledge of the distribution of the errors a natural choice is to use the ordinary least squares estimator

$$(\hat{a}^{(w)}, \hat{\mu}^{(w)}) = \arg\min_{a,\mu\in\mathbb{R}} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{t} \left( (x_{i,j}^{(w)} - \mu) - a(x_{i,j-1}^{(w)} - \mu) \right)^2 \right\} .$$

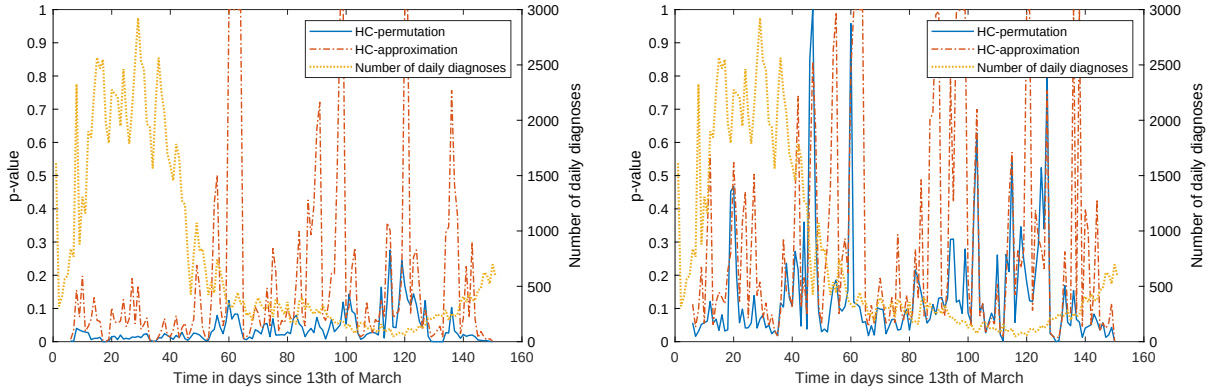
Finally, the methodology proposed in this paper is then applied to the residual errors, namely $(\tilde{x}_{i,j}^{(w)} : i \in [n], j \in [t])$ where

$$\tilde{x}_{i,j}^{(w)} \equiv x_{i,j}^{(w)} - \hat{\mu}^{(w)} - \hat{a}^{(w)} \left( x_{i,j-1}^{(w)} - \hat{\mu}^{(w)} \right) ,$$

We apply our testing methodology both to the raw data $(x_{i,j}^{(w)})_{i\in[n],j\in[t]}$ and to the residuals $(\tilde{x}_{i,j}^{(w)})_{i\in[n],j\in[t]}$, and contrast the obtained results. Note that often there are few very large observations. While these cases are important in the context of the application, one needs no powerful test to mark them as anomalous as their abnormality is so clear. Especially in the context of COVID-19, these "clear" outliers will be investigated regardless. A more interesting question is then if, apart from these "clear" outliers, we can still detect higher-than-usual values among the other municipalities.

To remove "clear" outliers in a rigorous way we use the max test as described in Theorem 2. First we obtain the 95% quantile of the permutation maximum stream mean distribution. We then mark all streams with means exceeding this threshold as "clear" anomalies. Finally, we apply our permutation higher criticism test as in Theorem 4 on the remaining data to detect possible signals. We also use the higher criticism test using a normal approximation as in Section 5.5 for a comparison. See also Remark 2 for a discussion on possible alternatives to consider when fitting the AR(1) model.

We present the results obtained for each possible window of 5 consecutive days in Figure 2. The $p$-value obtained by the testing procedures along with the virus' nationwide progression is depicted in that figure. Note that the time-windows are indexed by their starting day,




(a) Tests applied on the raw data.

(b) Tests applied on the residuals.

Figure 2: The $p$-values of our permutation higher criticism test and the higher criticisms test using normal approximations for a window of the previous five days, along with the total number of daily diagnoses in the Netherlands. For each test $10^5$ permutations were used.

so neighboring windows consider overlapping periods of time. Obviously, the obtained $p$-values are dependent and care must be taken if trying to interpret them jointly - the figure is given merely to aid the presentation. As can be seen in Figure 2, our test nearly always results in much smaller $p$-values than the approximation method at each window, both when using the raw data and the residuals of the AR(1) model. One can also see that, when the method is used on the residuals, larger $p$-values are typically observed. This indicates that in many of the windows considered, the use of the AR(1) model mitigates the effect of some dependencies and global trends that may unduly influence the conclusions. However, the autocorrelation parameter estimates were frequently small - across all windows considered, the estimates had a median value of 0.2243, and were smaller than 0.3 in 75% of the windows. At 5% significance, our test on the raw data rejects the null a total of 113 times out of 146, while the approximation test rejects a total of 49 times. On the residuals of our AR(1) model, our methodology rejects 43 times, while the approximation method rejects 20 times. Nevertheless, these figures should be interpreted with care, as the resulting $p$-values across different windows are dependent.

There are cases where our test indicates anomalies in seemingly stable periods. In these periods, while hard to see in aggregated data, we thus have some evidence that some mu-

nicipalities have larger-than-usual values. This does not necessarily lead to a nationwide outbreak, since local measures, such as a restricting access to specific nursery homes, might have been taken to prevent further spread. As our data is aggregated per municipality and local measures are very hard to identify, we cannot to take this into account in our analysis.

*Remark* 1. When the $p$-values of our test are small as above, one would ideally like to subsequently identify the anomalous municipalities. We refer to the ending of Section 5.6 for a discussion on the possibilities when one would like to identify anomalous municipalities.

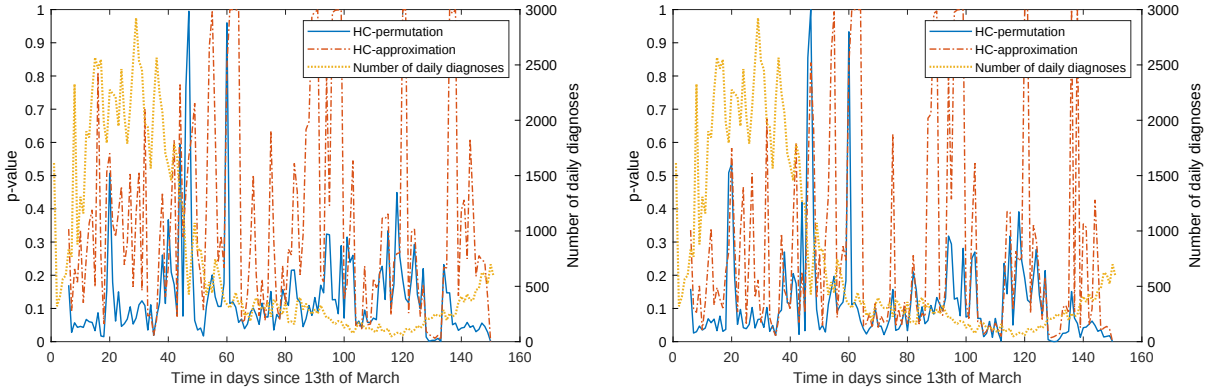
*Remark* 2. With respect to the exclusion of "clear" outliers, there are some natural alternatives to the approach above:

A. Remove obvious anomalies using the permutation distribution of the maximum stream mean on the raw data. Next, fit the AR(1) model on the remaining raw data, and apply the methodology on the residuals. In the results, we refer to this approach as "approach A".

B. First, fit an AR(1) model on the data. Then, identify obvious anomalous streams using the permutation distribution of the maximum stream mean on the residuals. If the stream mean of the residuals is larger than the 95% quantile, we remove the corresponding stream from our original raw data. Fit a new AR(1) model on the remaining raw data streams. Use our methodology on the residuals following from the second model fit. In the results, we refer to this approach as "approach B".

Compared to the previous approach, the first option avoids labeling streams as "obvious" anomalies based on the AR(1) model. The second option avoids using our methodology on residuals that arose from an AR(1) model fit which was unduly influenced by the presence of "obvious" anomalies.

The results do not qualitatively change compared to the results in the main text when these variations are considered. The results are presented in Figure 3.




(a) Results following approach A.

(b) Results following approach B.

Figure 3: The $p$-values of our permutation higher criticism test and the higher criticisms test using normal approximations for a window of the previous five days, along with the total number of daily diagnoses in the Netherlands. For each test $10^5$ permutations were used.

# References


Arias-Castro, E., Castro, R. M., Tánczos, E., and Wang, M. (2018). Distribution-free detection of structured anomalies: Permutation and rank-based scans. *Journal of the American Statistical Association*, 113(522):789–801.

Feller, W. (1968). *An introduction to probability theory and its applications*, volume 1. John Wiley & Sons, 3rd edition.

Held, L., Hens, N., D O'Neill, P., and Wallinga, J. (2019). *Handbook of infectious disease data analysis*. CRC Press.

Joag-Dev, K. and Proschan, F. (1983). Negative association of random variables with applications. *Annals of Statistics*, 11(1):286–295.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., and Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9):577–582.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, 3rd edition.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

Shtatland, E. S. (2007). Low-Order Autoregressive Models in Early Detection of Epidemic Outbreaks and Explosive Behaviors in Economic and Financial Time Series. In *NESUG*.

Shtatland, E. S. (2008). Another Look at Low-Order Autoregressive Models in Early Detection of Epidemic Outbreaks and Explosive Behaviors in Economic and Financial Time Series. In *SAS Global Forum*.