# Structured Correlation Detection with Application to Colocalization Analysis in Dual-Channel Fluorescence Microscopic Imaging

Shulei Wang*, Jianqing Fan†, Ginger Pocock§,Ellen T. Arena§,
Kevin W. Eliceiri§ and Ming Yuan*,‡

*Morgridge Institute for Research*[*,§] *and University of Wisconsin-Madison*[*,§]
*and Princeton University*[†]

**Supplementary Material**

## S1 Structured Correlation Detection Algorithms and Auxiliary Figures

### S1.1 Structured Correlation Detection Algorithms

In this section, we present two algorithms of structured correlation detection.

### S1.2 Auxiliary Figures

In this section, we show some auxiliary figures.

**Algorithm S1** Structured Correlation Detection by $T^*$

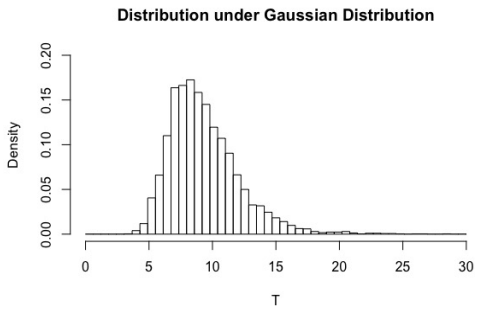
**Require:** Dual-channel Image $\{(X_i, Y_i)\}_{i\in I}$ and significance level $\alpha$.
**Ensure:** Decision on if colocalization happens.
  $T^* = 0$
  **for** $R \in \mathcal{R}$ **do**
    Calculate $T_R := \frac{1}{\log\log(n/A)} \left[\max_{R\in\mathcal{R}:|R|=A} L_R - 2\log(n/A)\right]$.
    **if** $T_R > T^*$ **then**
      $T^* = T_R$.
    **end if**
  **end for**
  **return** If $T^* > q_\alpha$, return "yes", else, return "no".

**Algorithm S2** Fast Structured Correlation Detection by $\tilde{T}^*$

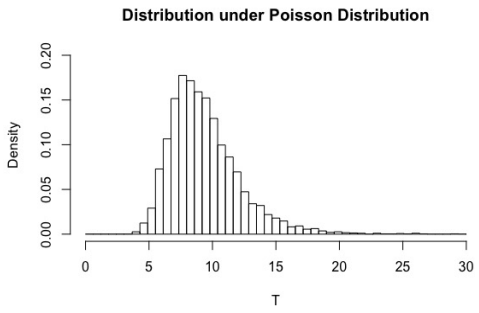
**Require:** Dual-channel Image $\{(X_i, Y_i)\}_{i\in I}$ and significance level $\alpha$.
**Ensure:** Decision on if colocalization happens.
  $\tilde{T}^* = 0$
  **for** $k = 1$ to $\lfloor \log_2 n \rfloor + 1$ **do**
    **if** $k > k_*$ **then**
      $\mathcal{R}' = \mathcal{R}_k$.
    **else**
      $\mathcal{R}' = \tilde{\mathcal{R}}_k$.
    **end if**
    **for** $R \in \mathcal{R}'$ **do**
      Calculate $T_R := \frac{1}{\log\log(n/A)} \left[\max_{R\in\mathcal{R}:|R|=A} L_R - 2\log(n/A)\right]$.
      **if** $T_R > \tilde{T}^*$ **then**
        $\tilde{T}^* = T_R$.
      **end if**
    **end for**
  **end for**
  **return** If $\tilde{T}^* > \tilde{q}_\alpha$, return "yes", else, return "no".

## S2 Proofs of Main Results

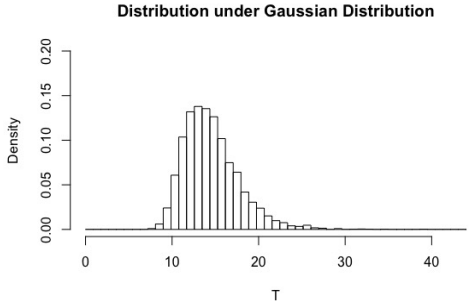
In this section, we present the proofs to our main results, namely Theorems 1, 2 and 3. Proofs of Propositions 1 and 2, as well as a number of auxiliary results, will be given in next section. To distinguish from the constants appeared in the previous sections, we shall use the capital letter $C$ to denote a generic positive constant that may take different values at each appearance.



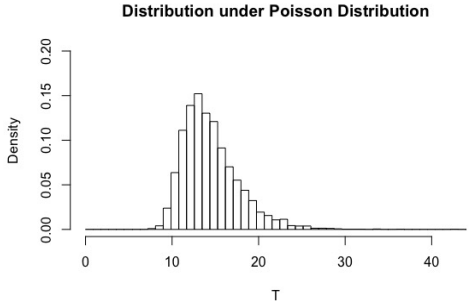

(a) Gaussian Distribution with A=50




(b) Poisson Distribution with A=50




(c) Gaussian Distribution with A=10




(d) Poisson Distribution with A=10

Figure S1: Simulated Distribution of $\max_{R\in\mathcal{R}(A)} L_R$

*Proof of Theorem 1.* We first prove the upper bound (2.7) under conditions (2.4) and (2.6). To this end, we shall establish a stronger result that there exists a constant $C > 0$ such that for any $0 < t < (\log n)^3$.

$$\mathbb{P}\left\{\max_{R\in\mathcal{R}(A)} L_R > 2\log n + C(\log\log n + t)\right\} \leq \exp(-t). \tag{S2.1}$$

It is clear that (2.7) follows immediately from (S2.1).

We now proceed to prove (S2.1). We shall consider the cases where $A \leq (\log n)^5$ and $A \geq (\log n)^5$ separately. First consider the situation when $A \leq (\log n)^5$. By Lemma 6,

there exists a constant $C > 0$ such that for any fixed $R \in \mathcal{R}(A)$

$$\mathbb{P}\{L_R > x\} \leq C \exp(-x/2).$$

Applying union bound yields

$$\begin{aligned} \mathbb{P}\left\{\max_{R\in\mathcal{R}(A)} L_R > x\right\} &\leq C|\mathcal{R}(A)| \exp(-x/2) \\ &\leq c_1 C n A^{c_2} \exp(-x/2) \\ &\leq c_1 C n (\log n)^{5c_2} \exp(-x/2), \end{aligned}$$

where the second inequality follows from (2.4). Equation (S2.1) then follows by taking

$$x = 2\log(c_1 C) + 2\log n + 10 c_2 \log\log n + 2t.$$

The treatment for $A \geq (\log n)^5$ is more involved and we apply a chaining argument. Let $\mathcal{R}_{\text{app}}(A, e^{-s})$ be an $e^{-s}$ covering set of $\mathcal{R}(A)$ so that

$$|\mathcal{R}_{\text{app}}(A, e^{-s})| = N(A, e^{-s}).$$

For any segment $R \in \mathcal{R}(A)$, denote by

$$\pi_s(R) = \operatorname*{argmin}_{R' \in \mathcal{R}_{\text{app}}(A, e^{-s})} d(R, R').$$

Of course, the minimizer on the right hand side may not be uniquely defined, in which case, we take $\pi_s(R)$ to be an arbitrarily chosen minimizer.

Write

$$L_R = \sum_{s=s_*}^{s^*-1} \left(L_{\pi_{s+1}(R)} - L_{\pi_s(R)}\right) + \left(L_R - L_{\pi_{s^*}(R)}\right) + L_{\pi_{s_*}(R)},$$

where $s^* > s_* \geq \log\log(n/A)$ are to be specified later. It is clear that

$$\max_{R\in\mathcal{R}(A)} L_R \leq \sum_{s=s_*}^{s^*-1} \max_{R\in\mathcal{R}(A)} \left|L_{\pi_{s+1}(R)} - L_{\pi_s(R)}\right| + \max_{R\in\mathcal{R}(A)} \left|L_R - L_{\pi_{s^*}(R)}\right| + \max_{R\in\mathcal{R}(A)} |L_{\pi_{s_*}(R)}|. \tag{S2.2}$$

We now bound the three terms on the right hand side of (S2.2) separately.

By definition,

$$d(R, \pi_s(R)) \leq e^{-s}, \qquad \text{and} \qquad d(R, \pi_{s+1}(R)) \leq e^{-(s+1)}.$$

Hence there exists a constant $C > 0$ such that

$$|\pi_s(R) \cap \pi_{s+1}(R)| \geq (1 - Ce^{-s})|R|, \qquad \text{and} \qquad d(\pi_s(R), \pi_{s+1}(R)) \leq Ce^{-s}.$$

Now by Lemma 7, for any fixed $R \in \mathcal{R}(A)$,

$$|L_{\pi_s(R)} - L_{\pi_{s+1}(R)}| \leq C\left(e^{-s/2}x + |R|^{-1/2}x^{3/2}\right)$$

with probability at least $1 - Ce^{-x}$. An application of the union bound yields

$$\begin{aligned}
& \mathbb{P}\left\{ \max_{R\in\mathcal{R}(A)} \left|L_{\pi_{s+1}(R)} - L_{\pi_s(R)}\right| > C\left(e^{-s/2}x + \sqrt{2}A^{-1/2}x^{3/2}\right)\right\} \\
\leq\ & CN(A, e^{-s})N(A, e^{-(s+1)})e^{-x} \\
\leq\ & C[N(A, e^{-(s+1)})]^2 e^{-x} \\
\leq\ & c_4^2 C\left(\frac{n}{A}\right)^2 \left(\log\frac{n}{A}\right)^{2c_5} e^{2c_6(s+1)}e^{-x},
\end{aligned}$$

where the last inequality follows from (2.6). In particular, taking

$$x = t + 2\log s + \log(c_4^2 C) + 2\log(n/A) + 2c_5\log\log(n/A) + 2c_6(s+1)$$

yields, with probability at least $1 - s^{-2}e^{-t}$,

$$\max_{R\in\mathcal{R}(A)} \left|L_{\pi_{s+1}(R)} - L_{\pi_s(R)}\right| \leq C\left((s + t + \log(n/A))e^{-s/2} + A^{-1/2}(s + t + \log(n/A))^{3/2}\right).$$

Here we used the fact that $s \geq s_* \geq \log\log(n/A)$. Now applying the union bound over all $s_* \leq s < s^*$, we get, with probability at least $1 - s_*^{-1}e^{-t} \geq 1 - e^{-t}$,

$$\begin{aligned}
\sum_{s=s_*}^{s^*-1} \max_{R\in\mathcal{R}(A)} \left|L_{\pi_{s+1}(R)} - L_{\pi_s(R)}\right| &\leq C\sum_{s=s_*}^{s^*-1}\left((s + t + \log(n/A))e^{-s/2} + A^{-1/2}(s + t + \log(n/A))^{3/2}\right) \\
&\leq C\left(s_* e^{-s_*/2} + A^{-1/2}(s^*)^{5/2}\right) \\
&\quad + C\left(e^{-s_*/2}(t + \log(n/A)) + A^{-1/2}s^*(t + \log(n/A))^{3/2}\right).
\end{aligned}$$

To bound the second term on the right hand side of (S2.2), we again apply Lemma

7. For any fixed $R \in \mathcal{R}(A)$, we get

$$\mathbb{P}\left\{\left|L_R - L_{\pi_{s^*}(R)}\right| \geq C\left(e^{-s^*/2}x + \sqrt{2}A^{-1/2}x^{3/2}\right)\right\} \leq Ce^{-x}.$$

Another application of the union bound yields,

$$\begin{aligned}\max_{R\in\mathcal{R}(A)}\left|L_R - L_{\pi_{s^*}(R)}\right| &\leq C\left(e^{-s^*/2}\log|\mathcal{R}(A)| + A^{-1/2}(\log|\mathcal{R}(A)|)^{3/2} + e^{-s^*/2}t + A^{-1/2}t^{3/2}\right)\\ &\leq c_2C\left(e^{-s^*/2}\log n + A^{-1/2}(\log n)^{3/2} + e^{-s^*/2}t + A^{-1/2}t^{3/2}\right),\end{aligned}$$

with probability at least $1 - Ce^{-t}$, where we used (2.4) in the last inequality.

Finally, for the third term on the right hand side of (S2.2), we have

$$\mathbb{P}\left\{\max_{R\in\mathcal{R}_{\text{app}}(A,e^{-s_*})}|L_R| \geq x\right\} \leq CN(A,e^{-s_*})e^{-x/2} \leq c_4C\left(\frac{n}{A}\right)\left(\log\frac{n}{A}\right)^{c_5}e^{c_6s_*}e^{-x/2}.$$

Taking

$$x = 2\log(c_4C) + 2\log\frac{n}{A} + c_5\log\log\frac{n}{A} + 2c_6s_* + t$$

yields, with probability at least $1 - Ce^{-t}$,

$$\max_{R\in\mathcal{R}_{\text{app}}(A,e^{-s_*})}|L_R| \leq 2\log(c_4C) + 2\log\frac{n}{A} + c_5\log\log\frac{n}{A} + 2c_6s_* + t.$$

In summary, we get, with probability at least $1 - Ce^{-t}$,

$$\begin{aligned}\max_{R\in\mathcal{R}(A)} L_R \;\leq\; & C\bigg(s_* e^{-s_*/2} + A^{-1/2}(s^*)^{5/2} + e^{-s_*/2}t + A^{-1/2}s^*t^{3/2} + e^{-s^*/2}\log n \\ & + A^{-1/2}(\log n)^{3/2} + e^{-s^*/2}t + A^{-1/2}t^{3/2} + e^{-s_*/2}\log\frac{n}{A} + A^{-1/2}s^*(\log(n/A))^{3/2}\bigg) \\ & + 2\log(c_4 C) + 2\log\frac{n}{A} + c_5\log\log\frac{n}{A} + 2c_6 s_* + t.\end{aligned}$$

Recall that $A \geq (\log n)^5$. If we take $s^* = 2\log n$ and $s_* = 2\log\log(n/A)$, then for any $t \leq (\log n)^3$, we can deduce from the above inequality that

$$\max_{R\in\mathcal{R}(A)} L_R \leq 2\log\frac{n}{A} + C\left(\log\log\frac{n}{A} + t\right), \tag{S2.3}$$

which implies (S2.1).

We now prove (2.8) if in addition, (2.5) holds. In the light of (2.6), we can find a subset $\tilde{\mathcal{R}}(A)$ of $\mathcal{R}(A)$ such that for any $R_1, R_2 \in \tilde{\mathcal{R}}(A)$, $R_1 \cap R_2 = \emptyset$ and

$$|\tilde{\mathcal{R}}(A)| \geq c_3\frac{n}{A}.$$

Obviously,

$$\max_{R\in\mathcal{R}(A)} L_R \geq \max_{R\in\tilde{\mathcal{R}}(A)} L_R.$$

If $A \le (\log n)^5$, then

$$\begin{aligned}
\mathbb{P}\left\{\max_{R\in\tilde{\mathcal{R}}(A)} L_R \le x\right\} &= \prod_{R\in\tilde{\mathcal{R}}(A)} \mathbb{P}\{L_R \le x\} \\
&\le \prod_{R\in\tilde{\mathcal{R}}(A)} \left(1 - C|R|^{-1/2}e^{-x/2}\right) \\
&\le \left[1 - CA^{-1/2}e^{-x/2}\right]^{c_3 n/A} \\
&\le \left[1 - C(\log n)^{-5/2}e^{-x/2}\right]^{c_3 n/A},
\end{aligned}$$

where the first inequality follows from the lower bound given by Lemma 6. It can then be derived that

$$\max_{R\in\tilde{\mathcal{R}}(A)} L_R \ge 2\log n + O_p(\log\log n). \tag{S2.4}$$

Together with (2.7), (S2.4) implies the desired claim when $A \le (\log n)^5$.

Next we consider the case when $A \ge (\log n)^5$. We proceed in a similar fashion as before but rely on the following tail bound of $L_R$: if $A \ge 24$, then there exists a constant $C > 0$ such that for any $R \in \mathcal{R}(A)$ and $0 < x < \sqrt{A}$,

$$\mathbb{P}\{L_R \ge x\} \le Cx^{-1/2}\exp(-x/2). \tag{S2.5}$$

If (S2.5) holds, then

$$\mathbb{P}\left\{\max_{R\in\tilde{\mathcal{R}}(A)} L_R \le x\right\} \ge \left(1 - Cx^{-1/2}e^{-x/2}\right)^{c_3 n/A},$$

which yields

$$\max_{R\in\mathcal{R}(A)} L_R \geq \max_{R\in\tilde{\mathcal{R}}(A)} L_R \geq 2\log(n/A) + O_p(\log\log(n/A)).$$

Together with (2.7), this concludes the proof.

It now remains to prove (S2.5). Write

$$T_R = (|R|-2)r_R^2.$$

Note that $\log(1+x) > x - x^2/2$ for any $x > 0$. We get

$$L_R \geq (|R|-2)\log(1+r_R^2) \geq T_R^2 - \frac{T_R^4}{2(|R|-2)} \geq T_R^2 - \frac{T_R^4}{A-4},$$

for any $A \geq 5$, where in the last inequality we used the fact that $|R| > A/2$ for any $R \in \mathcal{R}(A)$. This can be further lower bounded by $T_R^2 - 3T_R^4/A$ for any $A \geq 6$. Thus, for any $0 < x < A/24$,

$$\begin{aligned}
\mathbb{P}\{L_R \geq x\} &\geq \mathbb{P}\left\{T_R^2 - \frac{3T_R^4}{A} \geq x\right\} \\
&\geq \mathbb{P}\left\{T_R^2 - \frac{3T_R^4}{A} \in [x, 2x)\right\} \\
&\geq \mathbb{P}\left\{T_R^2 \in [x + 12x^2/A, 2x + 3x^2/A)\right\} \\
&\geq \mathbb{P}\left\{T_R \in [\sqrt{x + 12x^2/A}, \sqrt{2x + 3x^2/A})\right\}.
\end{aligned}$$

Because $T_R \sim t_{|R|-2}$, we have

$$\begin{aligned}
&\mathbb{P}\left\{T_R \in [\sqrt{x+12x^2/A}, \sqrt{2x+3x^2/A})\right\} \\
\geq\ & C\int_{\sqrt{x+12x^2/A}}^{\sqrt{2x+3x^2/A}} \left(1+\frac{u^2}{|R|-2}\right)^{-\frac{|R|-1}{2}} du \\
\geq\ & C\int_{\sqrt{x+12x^2/A}}^{\sqrt{2x+3x^2/A}} \exp\left[-\frac{|R|-1}{2}\log\left(1+\frac{u^2}{|R|-2}\right)\right] du \\
\geq\ & C\int_{\sqrt{x+12x^2/A}}^{\sqrt{2x+3x^2/A}} \exp\left(-\frac{|R|-1}{2(|R|-2)}u^2\right) du,
\end{aligned}$$

for some constant $C > 0$, where in the last inequality we used the fact that $\log(1+x) \leq x$ for all $x \geq 0$. Thus,

$$\begin{aligned}
&\mathbb{P}\left\{T_R \in [\sqrt{x+12x^2/A}, \sqrt{2x+3x^2/A})\right\} \\
\geq\ & C(2x+3x^2/A)^{-1/2}\int_{\sqrt{x+12x^2/A}}^{\sqrt{2x+3x^2/A}} u\exp\left(-\frac{|R|-1}{2(|R|-2)}u^2\right) du \\
=\ & C(2x+3x^2/A)^{-1/2}(1-(|R|-1)^{-1})\left[\exp\left(-\frac{|R|-1}{2(|R|-2)}(x+12x^2/A)\right)\right. \\
& \qquad \left. -\exp\left(-\frac{|R|-1}{2(|R|-2)}(2x+3x^2/A)\right)\right].
\end{aligned}$$

Recall that $0 < x < A/24$. We get

$$\begin{aligned} & \mathbb{P}\left\{T_R \in [\sqrt{x+12x^2/A}, \sqrt{2x+3x^2/A})\right\} \\ \geq\ & Cx^{-1/2}\exp\left(-\frac{|R|-1}{2(|R|-2)}(x+12x^2/A)\right) \\ \geq\ & Cx^{-1/2}\exp\left(-\frac{A-1}{2(A-2)}(x+12x^2/A)\right) \\ \geq\ & Cx^{-1/2}\exp(-x/2), \end{aligned}$$

where in the last inequality, we used the fact that $x \leq \sqrt{A}$. The proof is then completed.

□

*Proof of Theorem 2 (Consistency of $T^*$).* We first show that the claim is true for $T^*$. To this end, we begin by arguing that $q_\alpha = O(1)$, and then show that under $H_1$, $T^* \to \infty$. Note that

$$\begin{aligned} T^* &= \max_{R\in\mathcal{R}}\left\{\frac{1}{\log\log(n/|R|)}\left[L_R - 2\log\left(\frac{n}{|R|}\right)\right]\right\} \\ &= \max_{1\leq k\leq \log n}\max_{R\in\mathcal{R}(e^{-k+1}n)}\left\{\frac{1}{\log\log(n/|R|)}\left[L_R - 2\log\left(\frac{n}{|R|}\right)\right]\right\}. \end{aligned}$$

As shown in the proof of Theorem 1, there exists a constant $C > 0$ such that for any $0 < t < (\log n)^3$,

$$\mathbb{P}\left\{\max_{R\in\mathcal{R}(e^{-k+1}n)} L_R \geq 2k + C\left(\log k + t\right)\right\} \leq \exp(-t).$$

Taking $t = x + \log(2k^2)$ yields

$$\mathbb{P}\left\{\max_{R\in\mathcal{R}(e^{-k+1}n)}\left\{\frac{1}{\log\log(n/|R|)}\left[L_R - 2\log\left(\frac{n}{|R|}\right)\right]\right\} \geq C(x+1)\right\} \leq \frac{1}{2k^2}\exp(-x).$$

Applying union bound over all $k$, we get

$$\mathbb{P}\left\{T^* \geq C(x+1)\right\} \leq \sum_{1\leq k\leq \log n}\frac{1}{2k^2}\exp(-x) \leq \exp(-x),$$

which implies that $q_\alpha \leq C(1-\log(1-\alpha))$.

It now suffices to show that if (4.9) holds for some $R \in \mathcal{R}$, then $T^* \to \infty$. To this end, note that

$$T^* \geq \frac{1}{\log\log(n/|R|)}\left[L_R - 2\log\left(\frac{n}{|R|}\right)\right].$$

We treat the case $|R| \geq \log n$ and $|R| \leq \log n$ separately.

Consider first the situation when $|R| \leq \log n$. By Lemma 3,

$$\left(\frac{1-r_R^2}{1-\rho^2}\right)\sum_{i\in R}(Y_i - \bar{Y}_R)^2 \sim \chi^2_{|R|-2}.$$

Applying the $\chi^2$ tail bounds of Laurent and Massart (2000), we get, with probability at least $1-2e^{-x}$,

$$\left(\frac{1-r_R^2}{1-\rho^2}\right)\sum_{i\in R}(Y_i - \bar{Y}_R)^2 \leq (|R|-2) + 2\sqrt{x(|R|-2)} + 2x$$

and

$$\sum_{i\in R}(Y_i-\bar{Y}_R)^2 \geq (|R|-1)-2\sqrt{x(|R|-1)}.$$

Under this event,

$$\frac{1-r_R^2}{1-\rho^2} \leq \frac{(|R|-2)+2\sqrt{x(|R|-2)}+2x}{(|R|-1)-2\sqrt{x(|R|-1)}}.$$

Assuming that $x = o(|R|)$, this can be further simplified as

$$\frac{1-r_R^2}{1-\rho^2} \leq 1+o\left(\sqrt{\frac{x}{|R|}}\right).$$

If in addition, $x \to \infty$, then

$$\begin{aligned} -(|R|-2)\log\left(1-r_R^2\right) &\geq -|R|\log\left(1-\rho^2\right)+o\left(\sqrt{x|R|}\right) \\ &\geq 2\log(n/|R|)+\delta_n\log(n/|R|)+o\left(\sqrt{x|R|}\right), \end{aligned}$$

which diverges with $n$ because

$$\delta_n\log(n/|R|) \gg \sqrt{\log(n/|R|)} \gg |R| \gg \sqrt{x|R|}.$$

Since

$$\delta_n\log(n/|R|) \gg \sqrt{\log(n/|R|)} \gg \log\log(n/|R|),$$

this immediately suggests that

$$T^* \geq \frac{1}{\log\log(n/|R|)}\left[L_R-2\log\left(\frac{n}{|R|}\right)\right] \to_p \infty.$$

Next consider the case when $|R| \geq \log n$. Assume without loss of generality that $\rho > 0$. The treatment for $\rho < 0$ is identical. Following an argument similar to that for Lemma 4, we get

$$\sum_{i \in R}(X_i - \bar{X}_R)^2, \sum_{i \in R}(Y_i - \bar{Y}_R)^2 \leq (|R| - 1) + 2\sqrt{x(|R| - 1)} + 2x$$

and

$$\sum_{i \in R}(X_i - \bar{X}_R)(Y_i - \bar{Y}_R) \geq (|R| - 1)\rho - 2\sqrt{x(|R| - 1)} - 2x.$$

with probability at least $1 - 6e^{-x}$. Denote this event by $\mathcal{E}(x)$. We shall now proceed under $\mathcal{E}(x)$ with an appropriately chosen $x \to \infty$.

$$r_R \geq \frac{\rho - 2\sqrt{x/(|R| - 1)} - 2[x/(|R| - 1)]}{1 + 2\sqrt{x/(|R| - 1)} + 2[x/(|R| - 1)]}. \tag{S2.6}$$

It is not hard to see that under the condition (4.9), $|R|\rho^2 \to \infty$. Assuming that $x \to \infty$ such that $x = o(|R|\rho^2)$, we get

$$r_R \geq \rho + o\left(\sqrt{\frac{x}{|R|}}\right)$$

Then,

$$L_R \geq -(|R| - 2)\log\left[1 - \left(\rho + o\left(\sqrt{\frac{x}{|R|}}\right)\right)^2\right] \tag{S2.7}$$

Recall that

$$-|R|\log(1 - \rho^2) \geq (2 + \delta_n)\log\left(\frac{n}{|R|}\right).$$

Denote by $\rho_* > 0$ the solution to

$$-|R|\log(1-y^2) = (2+\delta_n)\log\left(\frac{n}{|R|}\right).$$

It is clear that $\rho \geq \rho_*$. Together with the fact that the right hand side of (S2.7) is an increasing function of $\rho$, we get

$$\begin{aligned} L_R &\geq -(|R|-2)\log\left[1-\left(\rho_* + o\left(\sqrt{\frac{x}{|R|}}\right)\right)^2\right] \\ &= -(|R|-2)\log(1-\rho_*^2) + o(\sqrt{x|R|}\rho_*) \\ &= (2+\delta_n)\log\left(\frac{n}{|R|}\right) + o(\sqrt{x|R|\rho_*^2}). \end{aligned}$$

Note that

$$|R|\rho_*^2 \leq 2|R|\log(1+\rho_*^2) \leq -2|R|\log(1-\rho_*^2) = 2(2+\delta_n)\log\left(\frac{n}{|R|}\right).$$

It is not hard to see that

$$\frac{\delta_n^2}{2+\delta_n}\log\left(\frac{n}{|R|}\right) \to \infty$$

if (4.9) holds. Assuming that

$$x = o\left(\frac{\delta_n^2}{2+\delta_n}\log\left(\frac{n}{|R|}\right)\right),$$

we get

$$T^* \geq (\log\log(n/|R|))^{-1}(L_R - 2\log(n/|R|)) \to \infty.$$

This concludes the proof of consistency of $T^*$ under (4.9). □

*Proof of Theorem 2 (Consistency of $\tilde{T}^*$).* We now consider the computationally efficient test based on $\tilde{T}^*$ is also consistent. As before, we begin by arguing that $\tilde{q}_\alpha = O(1)$, and then show that under $H_1$, $\tilde{T}^* \to \infty$. To show that $\tilde{q}_\alpha = O(1)$, it suffices to note that

$$\begin{aligned} T^* &= \max_{R \in \mathcal{R}} \left\{ (\log\log(n/|R|))^{-1} \left( L_R - 2\log(n/|R|) \right) \right\} \\ &\geq \max_{R \in \cup_k \tilde{\mathcal{R}}_k} \left\{ (\log\log(n/|R|))^{-1} \left( L_R - 2\log(n/|R|) \right) \right\} \\ &= \tilde{T}^*. \end{aligned}$$

Therefore, $\tilde{q}_\alpha \leq q_\alpha = O(1)$ following the argument before.

Next we show that under the alternative hypothesis where $X_i$ and $Y_i$ are correlated on a set $R \in \mathcal{R}_k$ for some $k$, $\tilde{T}^* \to \infty$. By definition, there exists a $\tilde{R} \in \tilde{\mathcal{R}}_k$ such that

$$d(R, \tilde{R}) \leq \frac{1}{4k^2}. \tag{S2.8}$$

Observe that

$$\tilde{T}^* \geq \tilde{T}^*_k \geq (\log\log(n/|\tilde{R}|))^{-1} \left( L_{\tilde{R}} - 2\log(n/|\tilde{R}|) \right).$$

It now suffices to show that the rightmost hand side is unbounded with probability approaching to 1 when $k \leq k_*$. To this end, we first consider the case when $\tilde{R} \subseteq R$.

Note that if $\tilde{R} \subseteq R$, then (1.2) holds for any $i \in \tilde{R}$. Following an identical argument

for consistency of $T^*$, it suffices to show that there exists a $\tilde{\delta}_n > 0$ such that

$$\tilde{\delta}_n \sqrt{\log(n/|\tilde{R}|)} \to \infty \tag{S2.9}$$

and

$$-|\tilde{R}| \log(1-\rho^2) \geq (2+\tilde{\delta}_n) \log\left(\frac{n}{|\tilde{R}|}\right). \tag{S2.10}$$

Observe that (S2.8) implies that

$$|\tilde{R}| \geq \left(1 - \frac{1}{4k^2}\right) |R|.$$

Thus

$$\begin{aligned} |\tilde{R}| \log \frac{1}{1-\rho^2} &\geq \left(1 - \frac{1}{4k^2}\right) |R| \log \frac{1}{1-\rho^2} \\ &\geq \left(1 - \frac{1}{4k^2}\right) (2+\delta_n) \log\left(\frac{n}{|R|}\right). \end{aligned}$$

Because

$$\log\left(\frac{n}{|R|}\right) = \log\left(\frac{n}{|\tilde{R}|}\right) + \log\left(\frac{|\tilde{R}|}{|R|}\right) \geq \log\left(\frac{n}{|\tilde{R}|}\right) + \log\left(1 - \frac{1}{4k^2}\right) \geq \log\left(\frac{n}{|\tilde{R}|}\right) - \frac{1}{4k^2},$$

we get

$$\begin{aligned} |\tilde{R}| \log \frac{1}{1-\rho^2} &\geq \left(1-\frac{1}{4k^2}\right)(2+\delta_n)\left[\log\left(\frac{n}{|\tilde{R}|}\right)-\frac{1}{4k^2}\right] \\ &\geq \left(1-\frac{1}{4k^2}\right)^2(2+\delta_n)\log\left(\frac{n}{|\tilde{R}|}\right) \\ &\geq \left(1-\frac{1}{2k^2}\right)(2+\delta_n)\log\left(\frac{n}{|\tilde{R}|}\right). \end{aligned}$$

Let

$$\tilde{\delta}_n = \left(1-\frac{1}{2k^2}\right)\delta_n - \frac{1}{k^2}.$$

Then (S2.10) holds. We now verify (S2.9). Recall that

$$\delta_n^2(k-1)\log 2 \leq \delta_n^2 \log\left(\frac{n}{|R|}\right) \to \infty,$$

we get, for sufficiently large $n$,

$$\tilde{\delta}_n \geq \frac{1}{4}\delta_n.$$

This implies that

$$\tilde{\delta}_n^2 \log\left(\frac{n}{|\tilde{R}|}\right) \geq \tilde{\delta}_n^2 \log\left(\frac{n}{|R|}\right) \geq \frac{1}{16}\delta_n^2 \log\left(\frac{n}{|R|}\right) \to \infty,$$

which completes the proof for the case $\tilde{R} \subseteq R$.

Now consider the case when $\tilde{R} \not\subseteq R$. By definition,

$$\frac{|\tilde{R} \cap R|}{\sqrt{|R||\tilde{R}|}} \geq 1 - \frac{1}{4k^2}.$$

Because $\tilde{R} \cap R \subseteq \tilde{R}$, we get

$$\frac{|\tilde{R}|}{|R|} \geq \left(1 - \frac{1}{4k^2}\right)^2. \tag{S2.11}$$

Thus,

$$|\tilde{R} \cap R| \geq \left(1 - \frac{1}{4k^2}\right)^{3/2} |R| \geq \left(1 - \frac{1}{3k^2}\right) |R|.$$

Similarly, we can derive that

$$|\tilde{R} \cap R| \geq \left(1 - \frac{1}{3k^2}\right) |\tilde{R}|. \tag{S2.12}$$

Following the same treatment as for the previous case, we can derive that

$$\frac{1}{\log\log(n/|\tilde{R} \cap R|)} \left[L_{\tilde{R}\cap R} - 2\log\left(\frac{n}{|\tilde{R} \cap R|}\right)\right] \to_p \infty.$$

Since $|\tilde{R} \cap R| \leq |\tilde{R}|$,

$$\frac{1}{\log\log \frac{n}{|\tilde{R}|}} \left[L_{\tilde{R}\cap R} - 2\log\left(\frac{n}{|\tilde{R}|}\right)\right] \geq \frac{1}{\log\log \frac{n}{|\tilde{R}\cap R|}} \left[L_{\tilde{R}\cap R} - 2\log\left(\frac{n}{|\tilde{R} \cap R|}\right)\right] \to \infty.$$

It now suffices to show that

$$|L_{\tilde{R}\cap R} - L_{\tilde{R}}| = O_p\left(\log\log\left(\frac{n}{|\tilde{R}|}\right)\right) \tag{S2.13}$$

In the light of (S2.11),

$$\begin{aligned}\log\log\left(\frac{n}{|\tilde{R}|}\right) &\geq \log\left[\log\left(\frac{n}{|R|}\right)+2\log\left(1-\frac{1}{4k^2}\right)\right]\\ &\geq \log\left[(k-1)\log 2-\frac{1}{2k^2}\right]=O(\log k).\end{aligned}$$

On the other hand, by Lemma 8,

$$|L_{\tilde{R}\cap R}-L_{\tilde{R}}|\leq C\left[\frac{1}{3k^2}x+|\tilde{R}|^{-1/2}x^{3/2}\right],$$

with probability at least $1-e^{-x}$. Observe that

$$|\tilde{R}|\geq\left(1-\frac{1}{4k^2}\right)^2|R|\geq\left(1-\frac{1}{2k^2}\right)|R|\geq n2^{-(k+1)}.$$

Equation (S2.13) then follows by taking

$$x=\min\left\{k^2\log k, 2^{-k/3}n^{1/3}(\log k)^{2/3}\right\}.$$

The proof is now completed. □

*Proof of Theorem 3.* Our argument is similar to those used earlier by Lepski and Tsybakov (2000) and Walther (2010). We shall outline only the main steps for brevity. Note first that a lower bound for a special case necessarily yields a lower bound for the general case. Thus it suffices to consider the case when $\mu_1=\mu_2=0$ and $\sigma_1=\sigma_2=1$. In the light of (2.5), for any $A$, we can find $\tilde{\mathcal{R}}(A)\subset\mathcal{R}(A)$ such that $|\tilde{\mathcal{R}}(A)|=c_3(n/A)$, and for

any $R_1, R_2 \in \tilde{\mathcal{R}}(A)$, $R_1 \cap R_2 = \emptyset$. For brevity, we shall assume that $c_3 = 1$ and for any $R \in \tilde{\mathcal{R}}(A)$, $|R| = A$. More general case can be treated in the same fashion albeit the argument becomes considerably more cumbersome.

Denote by $\mathbb{P}_0$ the joint distribution of $\{(X_i, Y_i) : i \in \mathbb{I}\}$ under null hypothesis, and by $\mathbb{P}_R$ the joint distribution under alternative hypothesis where $X_i$ and $Y_i$ are correlated on $R \in \tilde{\mathcal{R}}(A)$ so that (1.2) holds for $i \in R$ and (1.1) holds for $i \notin R$. The likelihood ratio between $\mathbb{P}_0$ and $\mathbb{P}_R$ can be computed:

$$W_R = \frac{d\mathbb{P}_R}{d\mathbb{P}_0} = \frac{1}{(1-\rho^2)^{A/2}} \exp\left\{ -\frac{\sum_{i \in R} (\rho^2 X_i^2 - 2\rho X_i Y_i + \rho^2 Y_i^2)}{2(1-\rho^2)} \right\}$$

To prove the first statement, we first show

$$\mathbb{E}_0(W_R^{1+\delta_n/4})/(\eta|\tilde{\mathcal{R}}(A)|)^{\delta_n/4} \to 0 \qquad \text{for any } 0 < \eta < 1,$$

where $\mathbb{E}_0$ stands for expectation taken with respect to $\mathbb{P}_0$.

It can be computed that

$$\mathbb{E}_0(W_R^{1+\delta_n/4}) = \frac{1}{(1-\rho^2\delta_n^2/16)^{A/2}(1-\rho^2)^{A\delta_n/8}}.$$

Recall that

$$A \log \frac{1}{1-\rho^2} \le (2-\delta_n) \log \frac{n}{A}.$$

We get

$$\begin{aligned}
&-\log\left[\mathbb{E}_0(W_R^{1+\delta_n/4})/(\eta|\tilde{\mathcal{R}}(A)|)^{\delta_n/4}\right] \\
\geq&\frac{A\delta_n}{8}\log(1-\rho^2)+\frac{A}{2}\log(1-\rho^2\delta_n^2/16)+\frac{\delta_n}{4}\log\frac{n}{A}+\frac{\delta_n}{4}\log\eta \\
\geq&\frac{\delta_n^2}{8}\log\frac{n}{A}-(1-\delta_n/2)\left(\log\frac{n}{A}\right)\frac{\log(1-\rho^2\delta_n^2/16)}{\log(1-\rho^2)}+\frac{\delta_n}{4}\log\eta \\
\geq&\frac{\delta_n^2}{8}\log\frac{n}{A}-\frac{\delta_n^2}{16}(1-\delta_n/2)\left(\log\frac{n}{A}\right)+\frac{\delta_n}{4}\log\eta \\
=&\frac{\delta_n^2}{16}(1+\delta_n/2)\log\frac{n}{A}+\frac{\delta_n}{4}\log\eta \\
\geq&\frac{\delta_n^2}{16}\log\frac{n}{A}\to\infty.
\end{aligned}$$

Thus,

$$\mathbb{E}_0(W_R^{1+\delta_n/4})/(\eta|\tilde{\mathcal{R}}(A)|)^{\delta_n/4}\to 0.$$

Next, we argue that

$$\mathbb{E}_0\left||\tilde{\mathcal{R}}(A)|^{-1}\sum_{R\in\tilde{\mathcal{R}}(A)}W_R-1\right|\to 0.$$

To this end, write

$$\bar{W}=|\tilde{\mathcal{R}}(A)|^{-1}\sum_{R\in\tilde{\mathcal{R}}(A)}(W_R-1),$$

$$\bar{W}_1=|\tilde{\mathcal{R}}(A)|^{-1}\sum_{R\in\tilde{\mathcal{R}}(A)}(W_R-1)\mathbf{1}_{(|W_R-1|>\eta|\tilde{\mathcal{R}}(A)|)},$$

and

$$\bar{W}_2=|\tilde{\mathcal{R}}(A)|^{-1}\sum_{R\in\tilde{\mathcal{R}}(A)}(W_R-1)\mathbf{1}_{|W_R-1|\leq\eta|\tilde{\mathcal{R}}(A)|}.$$

Observe that

$$\mathbb{E}_0 \left|\bar{W}\right| \leq \mathbb{E}_0 \left|\bar{W}_1\right| + \mathbb{E}_0 \left|\bar{W}_2\right| \leq \mathbb{E}_0 \left|\bar{W}_1\right| + \eta.$$

On the other hand,

$$\mathbb{E}_0 \left|\bar{W}_1\right| \leq \mathbb{E}_0(W_R \mathbf{1}_{(W_R > \eta|\tilde{\mathcal{R}}(A)|)}) \leq \mathbb{E}_0(W_R^{1+\delta_n/4})/(\eta|\tilde{\mathcal{R}}(A)|)^{\delta_n/4} \to 0.$$

We can take $\eta \downarrow 0$ to get

$$\mathbb{E}_0 \left| |\tilde{\mathcal{R}}(A)|^{-1} \sum_{R \in \tilde{\mathcal{R}}(A)} W_R - 1 \right| \to 0.$$

Finally, let $\mathbb{P}_1$ be the uniform mixture of $\mathbb{P}_R$ for $R \in \tilde{\mathcal{R}}(A)$, that is,

$$\mathbb{P}_1 = |\tilde{\mathcal{R}}(A)|^{-1} \sum_{R \in \tilde{\mathcal{R}}(A)} \mathbb{P}_R.$$

Then for any test $\Delta$,

$$\begin{aligned}
\mathbb{P}_0(\Delta = 1) + \mathbb{P}_1(\Delta = 0) =& \mathbb{E}_0(\Delta) + 1 - \min_{R \in \tilde{\mathcal{R}}(A)} \mathbb{E}_R(\Delta) \\
\geq& \mathbb{E}_0(\Delta) + 1 - |\tilde{\mathcal{R}}(A)| \sum_{R \in \tilde{\mathcal{R}}(A)} \mathbb{E}_R(\Delta) \\
\geq& 1 - \mathbb{E}_0(\Delta(1 - |\tilde{\mathcal{R}}(A)|^{-1} \sum_{R \in \tilde{\mathcal{R}}(A)} W_R)) \\
\geq& 1 - \mathbb{E}_0 \left| 1 - |\tilde{\mathcal{R}}(A)|^{-1} \sum_{R \in \tilde{\mathcal{R}}(A)} W_R \right| \to 1,
\end{aligned}$$

which completes the proof of the first statement.

To show the second statement, we assume the contrary that $c_n$ is bounded from above. Then $\{c_n\}$ must have a convergent subsequence. Without loss of generality, assume $c_n$ itself converges to some $b \in [0, \infty)$. Then

$$\log W_R \to_d N\left(-\frac{b}{2}, b\right),$$

which implies that

$$\limsup \mathbb{P}_R(\Delta = 1) < 1.$$

This contradicts with the fact that that the type II error of $\Delta$ goes to 0 as $n \to \infty$. The second statement is therefore proven. □

## S3 Auxiliary Results and Proofs

We first state tail bounds for $t$ and $F$ distributions necessary for our derivations.

**Lemma 1.** *Let $X$ be a random variable following a $t$ distribution with degree of freedom $n > 1$. There exists a numerical constant $0 < c_1 < c_2$ such that*

$$c_1 n^{-1/2} \left(1 + \frac{x^2}{n}\right)^{-\frac{n}{2}} \leq \mathbb{P}(|X| > x) \leq c_2 \left(1 + \frac{x^2}{n}\right)^{-\frac{n}{2}} \tag{S3.14}$$

*for any $x \geq 1$. In particular,*

$$c_1 n^{-1/2} e^{-u/2} \leq \mathbb{P}\left\{n \log\left(1 + \frac{X^2}{n}\right) \geq u\right\} \leq c_2 e^{-u/2}, \tag{S3.15}$$

*for any* $u \geq 1$.

*Proof of Lemma 1.* Recall that the density of a $t$ distribution with degree of freedom $n$ is

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \leq C\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},$$

for an absolute constant $C > 0$. Then, for any $u > 0$,

$$\begin{aligned}
\mathbb{P}(X > u) &\leq C\int_u^\infty \left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx \\
&\leq C\int_u^\infty \frac{x}{u}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx \\
&= \frac{nC}{2u}\int_u^\infty \left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} d\left(1+\frac{x^2}{n}\right) \\
&= \frac{nC}{(n-1)u}\left(1+\frac{x^2}{n}\right)^{-\frac{n-1}{2}}\Bigg|_u^\infty \\
&\leq 2C\frac{1}{u}\left(1+\frac{u^2}{n}\right)^{-\frac{n-1}{2}}.
\end{aligned}$$

The upper bound in (S3.14) follows immediately by taking $c = 4\sqrt{2}C$, by symmetry of $t$ distribution. On the other hand, observe that

$$f(x) \geq C\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},$$

for some constant $C > 0$. Thus,

$$\begin{aligned}
\mathbb{P}(X > u) &\geq C \int_u^\infty \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx \\
&\geq C \int_u^\infty \frac{x}{\sqrt{n}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n}{2}-1} dx \\
&= \frac{\sqrt{n}C}{2} \int_u^\infty \left(1 + \frac{x^2}{n}\right)^{-\frac{n}{2}-1} d\left(1 + \frac{x^2}{n}\right) \\
&= \frac{\sqrt{n}C}{(n-1)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n}{2}} \Bigg|_u^\infty \\
&\geq \frac{\sqrt{n}C}{(n-1)} \left(1 + \frac{u^2}{n}\right)^{-\frac{n}{2}}.
\end{aligned}$$

The lower bound in (S3.14) then follows immediately.

Now, taking

$$x = \sqrt{n(e^{u/n} - 1)}$$

in (S3.14) yields (S3.15). □

**Lemma 2.** *Let $U_1 \sim \chi^2_{n_1}$ and $U_2 \sim \chi^2_{n_2}$ be two independent random variables. Then for any $-1 < x < 1$,*

$$\mathbb{P}\left\{\left|\frac{n_1 + n_2}{n_1}\frac{U_1}{U_1 + U_2} - 1\right| \geq x\right\} \leq 2\exp\left(-\frac{n_1 x^2}{12}\right).$$

*Proof of Lemma 2.* As shown by Dasgupta and Gupta (2003), for any $x > 0$,

$$\mathbb{P}\left\{\frac{n_1 + n_2}{n_1}\frac{U_1}{U_1 + U_2} \leq 1 - x\right\} \leq \exp\left(\frac{n_1}{2}(x + \log(1 - x))\right),$$

and

$$\mathbb{P}\left\{\frac{n_1+n_2}{n_1}\frac{U_1}{U_1+U_2} \geq 1+x\right\} \leq \exp\left(\frac{n_1}{2}(-x+\log(1+x))\right).$$

The claim then follows from the fact that

$$\log(1+x) \leq x - \frac{x^2}{6}$$

for all $x$ such that $|x| < 1$. □

The following observation on the sample correlation coefficient is useful:

**Lemma 3.** *Assume that* $\{(X_i, Y_i) : i \in R\}$ *are iid copies of* $(X, Y) \sim N((\mu_1, \mu_2)^\top, \Sigma)$ *where*

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

*Then*

$$\sum_{i\in R}(Y_i - \bar{Y}_R)^2(1-r_R^2) \sim (1-\rho^2)\chi^2_{|R|-2}.$$

*Proof of Lemma 3.* Consider a linear regression of $Y$ over $X$:

$$Y = \beta_0 + \beta X + \epsilon.$$

Recall that

$$\widehat{\beta} = \frac{\sum_{i\in R}(X_i - \bar{X}_R)(Y_i - \bar{Y}_R)}{\sum_{i\in R}(X_i - \bar{X}_R)^2}$$

and $\widehat{\beta}_0 = \bar{Y}_R - \widehat{\beta}\bar{X}_R$ are the least squares estimate of of $Y$ over $X$ where

$$\bar{X}_R = \frac{1}{|R|}\sum_{i\in R} X_i \qquad \text{and} \qquad \bar{Y}_R = \frac{1}{|R|}\sum_{i\in R} Y_i.$$

The residual sum of squares of the regression can then be written as

$$\sum_{i\in R}(Y_i - \widehat{\beta}_0 - \widehat{\beta}X_i)^2 = \sum_{i\in R}(Y_i - \bar{Y}_R)^2(1 - r_R^2)$$

Conditioned on $X_i$, the residual sum of squares will follow $(1-\rho^2)\chi^2_{|R|-2}$. Thus the margin distribution of the residual sum of squares is also $(1-\rho^2)\chi^2_{|R|-2}$. □

Next we derive a tail bound for the sample correlation coefficient. For brevity, we work with the case when $(X, Y)$ are known to be centered so that

$$r_R = \frac{\sum_{i\in R} X_i Y_i}{\sqrt{\sum_{i\in R} X_i^2}\sqrt{\sum_{i\in R} Y_i^2}} \tag{S3.16}$$

where $(X_i, Y_i)$s are independent copies of $(X, Y)$. Treatment for the more general case is completely analogous, yet this simplification allows us to avoid lengthy discussions about the smaller order effects due to centering by sample means, and repeatedly switching between $|R| - 1$ or $|R| - 2$ as the appropriate degrees of freedom.

**Lemma 4.** *Assume that* $\{(X_i, Y_i) : i \in R\}$ *are iid copies of* $(X, Y) \sim N(0, I_2)$. *Then for any* $x > 0$,

$$\mathbb{P}\left\{\left|\sum_{i\in R} X_i Y_i\right| \geq 2\sqrt{x|R|} + 2x\right\} \leq 4e^{-x}.$$

*If in addition,* $0 < x < |R|/16$*, then*

$$\mathbb{P}\{|r_R| \geq x\} \leq 6\exp(-|R|x^2/64).$$

*Proof of Lemma 4.* Write

$$\sum_{i\in R} X_i Y_i = \frac{1}{2}\sum_{i\in R}\left(\frac{1}{\sqrt{2}}(X_i+Y_i)\right)^2 - \frac{1}{2}\sum_{i\in R}\left(\frac{1}{\sqrt{2}}(X_i-Y_i)\right)^2.$$

Then

$$\begin{aligned}\mathbb{P}\left\{\left|\sum_{i\in R} X_iY_i\right| \geq 2\sqrt{u|R|}+2u\right\} &\leq \mathbb{P}\left\{\left|\sum_{i\in R}\left(\frac{1}{\sqrt{2}}(X_i+Y_i)\right)^2 - |R|\right| \geq 2\sqrt{u|R|}+2u\right\} \\ &\quad + \mathbb{P}\left\{\left|\sum_{i\in R}\left(\frac{1}{\sqrt{2}}(X_i-Y_i)\right)^2 - |R|\right| \geq 2\sqrt{u|R|}+2u\right\} \\ &\leq 4e^{-u},\end{aligned}$$

where the second inequality follows from the $\chi^2$ upper and lower tail bound of Laurent and Massart (2000). Applying the $\chi^2$ lower tail bound from Laurent and Massart (2000), we can also derive that

$$\mathbb{P}\left\{\left|\sum_{i\in R} X_i^2\right| \leq |R| - 2\sqrt{u|R|}\right\} \leq e^{-u} \tag{S3.17}$$

and

$$\mathbb{P}\left\{\left|\sum_{i\in R} Y_i^2\right| \leq |R| - 2\sqrt{u|R|}\right\} \leq e^{-u} \tag{S3.18}$$

Therefore, for any $u < |R|/16$,

$$|r_R| \le \frac{2\sqrt{u|R|} + 2u}{|R| - 2\sqrt{u|R|}} \le \frac{2}{|R|}\left(2\sqrt{u|R|} + 2u\right) \le 8\sqrt{\frac{u}{|R|}}.$$

with probability at least $1 - 6e^{-u}$. The claim follows immediately. □

We are also interested in the difference in correlation coefficients between two different regions. The following lemma provides a useful probabilisitic tool for such purposes.

**Lemma 5.** *Assume that* $\{(X_i, Y_i) : i \in R_1 \cup R_2\}$ *are iid copies of* $(X, Y) \sim N(0, I_2)$*, and* $2|R_1 \cap R_2| \ge |R_1 \cup R_2|$*. Then there exist numerical constants* $c_0, c_1, c_2 > 0$ *such that for any* $x < c_0|R_1|$*,*

$$\mathbb{P}\left(\left||R_1|r_{R_1}^2 - |R_2|r_{R_2}^2\right| \ge x\right) \le c_1 \exp\left(-c_2 \min\left\{\left(\frac{|R_1 \cap R_2|}{|R_1 \cup R_2| - |R_1 \cap R_2|}\right)^{1/2} x, |R_1 \cap R_2|^{1/3} x^{2/3}\right\}\right).$$

*In particular, if*

$$\zeta := \frac{|R_1 \cap R_2|}{\sqrt{|R_1||R_2|}} \ge \frac{1}{4},$$

*then there exists a numerical constant* $c_3 > 0$ *such that for any* $x < c_0|R_1|$*,*

$$\mathbb{P}\left(\left||R_1|r_{R_1}^2 - |R_2|r_{R_2}^2\right| \ge x\right) \le c_1 \exp\left(-c_3 \min\left\{(1-\zeta)^{-1/2} x, |R_1 \cap R_2|^{1/3} x^{2/3}\right\}\right).$$

*Proof of Lemma 5.* We first consider the case when $R_2 \subseteq R_1$. Recall that

$$r_{R_1} = \frac{\sum_{i\in R_1} X_i Y_i}{\sqrt{\sum_{i\in R_1} X_i^2 \sum_{i\in R_1} Y_i^2}}, \qquad \text{and} \qquad r_{R_2} = \frac{\sum_{i\in R_1} X_i Y_i}{\sqrt{\sum_{i\in R_2} X_i^2 \sum_{i\in R_2} Y_i^2}}.$$

Therefore,

$$\begin{aligned} |R_2| r_{R_2}^2 - |R_1| r_{R_1}^2 &= \left(\frac{1}{\sqrt{|R_2|}} \sum_{i\in R_2} X_i Y_i\right)^2 \left(\frac{|R_2|^2}{\sum_{i\in R_2} X_i^2 \sum_{i\in R_2} Y_i^2} - \frac{|R_1|^2}{\sum_{i\in R_1} X_i^2 \sum_{i\in R_1} Y_i^2}\right) \\ &\quad + \frac{|R_1|^2}{\sum_{i\in R_1} X_i^2 \sum_{i\in R_1} Y_i^2} \left[\frac{1}{|R_2|}\left(\sum_{i\in R_2} X_i Y_i\right)^2 - \frac{1}{|R_1|}\left(\sum_{i\in R_1} X_i Y_i\right)^2\right] \end{aligned}$$

We now bound the terms on the right hand side separately.

Observe that

$$\begin{aligned} &\left|\left(\frac{1}{\sqrt{|R_2|}} \sum_{i\in R_2} X_i Y_i\right)^2 \left(\frac{|R_1|^2}{\sum_{i\in R_1} X_i^2 \sum_{i\in R_1} Y_i^2} - \frac{|R_2|^2}{\sum_{i\in R_2} X_i^2 \sum_{i\in R_2} Y_i^2}\right)\right| \\ = \; & |R_2| r_{R_2}^2 \left|1 - \left(\frac{|R_1|}{|R_2|}\frac{\sum_{i\in R_2} X_i^2}{\sum_{i\in R_1} X_i^2}\right)^{-1} \left(\frac{|R_1|}{|R_2|}\frac{\sum_{i\in R_2} Y_i^2}{\sum_{i\in R_1} Y_i^2}\right)^{-1}\right|. \end{aligned}$$

By Lemma 2,

$$\mathbb{P}\left\{\left|\frac{|R_1|}{|R_2|}\frac{\sum_{i\in R_2} X_i^2}{\sum_{i\in R_1} X_i^2} - 1\right| \geq x\right\} \leq 2\exp\left(-\frac{|R_2|}{12}x^2\right),$$

and

$$\mathbb{P}\left\{\left|\frac{|R_1|}{|R_2|}\frac{\sum_{i\in R_2} Y_i^2}{\sum_{i\in R_1} Y_i^2} - 1\right| \geq x\right\} \leq 2\exp\left(-\frac{|R_2|}{12}x^2\right).$$

We get, for any $x < 1/2$,

$$\left|1 - \left(\frac{|R_1|}{|R_2|}\frac{\sum_{i\in R_2} X_i^2}{\sum_{i\in R_1} X_i^2}\right)^{-1}\left(\frac{|R_1|}{|R_2|}\frac{\sum_{i\in R_2} Y_i^2}{\sum_{i\in R_1} Y_i^2}\right)^{-1}\right| \le 4x$$

with probability at least $1 - 4\exp\left(-|R_2|x^2/12\right)$. On the other hand, by Lemma 4,

$$\mathbb{P}\{|r_{R_2}| \ge x\} \le 6\exp(-|R_2|x^2/64).$$

Thus, by taking $u = 4|R_2|x^3$,

$$|R_2|r_{R_2}^2\left|1 - \left(\frac{|R_1|}{|R_2|}\frac{\sum_{i\in R_2} X_i^2}{\sum_{i\in R_1} X_i^2}\right)^{-1}\left(\frac{|R_1|}{|R_2|}\frac{\sum_{i\in R_2} Y_i^2}{\sum_{i\in R_1} Y_i^2}\right)^{-1}\right| \le u$$

with probability at least

$$1 - 10\exp\left(-\frac{|R_2|^{1/3}u^{2/3}}{64\cdot 4^{2/3}}\right).$$

Denote by $\mathcal{E}_1(u)$ the event that the above inequality holds.

To bound the second term, first note that

$$\begin{aligned}
&\frac{1}{|R_2|}\left(\sum_{i\in R_2} X_iY_i\right)^2 - \frac{1}{|R_1|}\left(\sum_{i\in R_1} X_iY_i\right)^2 \\
= \;& \left(\frac{1}{|R_2|} - \frac{1}{|R_1|}\right)\left(\sum_{i\in R_1} X_iY_i\right)^2 - \frac{1}{|R_2|}\left(\sum_{i\in R_1\setminus R_2} X_iY_i\right)^2 - \frac{2}{|R_2|}\left(\sum_{i\in R_2} X_iY_i\right)\left(\sum_{i\in R_1\setminus R_2} X_iY_i\right).
\end{aligned}$$

By Lemma 4,

$$\frac{|R_1|^2}{\sum_{i\in R_1} X_i^2 \sum_{i\in R_1} Y_i^2}\left(\frac{1}{|R_2|}-\frac{1}{|R_1|}\right)\left(\sum_{i\in R_1} X_iY_i\right)^2 = \left(\frac{|R_1|}{|R_2|}-1\right)|R_1| r_{R_1}^2 \le \left(\frac{|R_1|}{|R_2|}-1\right)|R_1|x^2,$$

with probability at least $1-6\exp(-|R_1|x^2/64)$. On the other hand, again by Lemma 4,

$$\mathbb{P}\left\{\left|\sum_{i\in R_1\setminus R_2} X_iY_i\right| \ge 2\sqrt{x(|R_1|-|R_2|)}+2x\right\} \le 4e^{-x},$$

Recall that, by $\chi^2$ lower tail bounds from Laurent and Massart (2000), we get

$$\mathbb{P}\left\{\sum_{i\in R_1} X_i^2 \le |R_1| - 2\sqrt{x|R_1|}\right\} \le e^{-x},$$

and

$$\mathbb{P}\left\{\sum_{i\in R_1} Y_i^2 \le |R_1| - 2\sqrt{x|R_1|}\right\} \le e^{-x}.$$

Thus, for any $x < |R_1|/16$,

$$\begin{aligned}\frac{|R_1|^2/|R_2|}{\sum_{i\in R_1} X_i^2 \sum_{i\in R_1} Y_i^2}\left(\sum_{i\in R_1\setminus R_2} X_iY_i\right)^2 &\le \frac{|R_1|^2/|R_2|}{\left(|R_1|-2\sqrt{x|R_1|}\right)^2}\left(2\sqrt{x(|R_1|-|R_2|)}+2x\right)^2 \\ &\le \frac{16}{|R_2|}\left(\sqrt{x(|R_1|-|R_2|)}+x\right)^2,\end{aligned}$$

with probability at least $1 - 6e^{-x}$. In other words,

$$\begin{aligned}\frac{|R_1|^2/|R_2|}{\sum_{i\in R_1} X_i^2 \sum_{i\in R_1} Y_i^2}\left(\sum_{i\in R_1\setminus R_2} X_iY_i\right)^2 &\leq \frac{1}{|R_2|}\left(\frac{1}{2}x\sqrt{|R_1|(|R_1|-|R_2|)}+\frac{|R_1|x^2}{16}\right)^2 \\ &\leq \frac{1}{2}\left(\frac{|R_1|}{|R_2|}-1\right)|R_1|x^2+\frac{|R_1|^2x^4}{128|R_2|},\end{aligned}$$

with probability at least $1 - 6\exp(-|R_1|x^2/64)$ for any $x < 2$. Following a similar argument, we can also show that

$$\begin{aligned}&\frac{2|R_1|^2/|R_2|}{\sum_{i\in R_1} X_i^2 \sum_{i\in R_1} Y_i^2}\left(\sum_{i\in R_2} X_iY_i\right)\left(\sum_{i\in R_1\setminus R_2} X_iY_i\right) \\ \leq\ &\frac{1}{|R_2|}\left(\frac{1}{2}x\sqrt{|R_1||R_2|}+\frac{|R_1|x^2}{16}\right)\left(\frac{1}{2}x\sqrt{|R_1|(|R_1|-|R_2|)}+\frac{|R_1|x^2}{16}\right),\end{aligned}$$

with probability at least $1 - 10\exp(-|R_1|x^2/64)$ for any $x < 2$. Note $2|R_2| \geq |R_1|$. In summary, we get

$$\begin{aligned}&\frac{|R_1|^2}{\sum_{i\in R_1} X_i^2 \sum_{i\in R_1} Y_i^2}\left|\frac{1}{|R_2|}\left(\sum_{i\in R_2} X_iY_i\right)^2-\frac{1}{|R_1|}\left(\sum_{i\in R_1} X_iY_i\right)^2\right| \\ \leq\ &2\left(\frac{|R_1|}{|R_2|}-1\right)^{1/2}|R_1|x^2+\frac{|R_1|x^3}{16}\end{aligned}$$

with probability at least $1 - 22\exp(-|R_1|x^2/64)$ for any $x < 2$. Hence, with probability at least

$$1-22\exp\left(-\frac{1}{256}\left(\frac{|R_1|}{|R_2|}-1\right)^{-1/2}u\right)-22\exp\left(-\frac{1}{16}|R_1|^{1/3}u^{2/3}\right)$$

we have

$$\frac{|R_1|^2}{\sum_{i\in R_1} X_i^2 \sum_{i\in R_1} Y_i^2}\left|\frac{1}{|R_2|}\left(\sum_{i\in R_2} X_iY_i\right)^2 - \frac{1}{|R_1|}\left(\sum_{i\in R_1} X_iY_i\right)^2\right| \le u.$$

Denote this event by $\mathcal{E}_2(u)$.

In summary, for any $u < |R_1|/256$,

$$||R_1|r_{R_1}^2 - |R_2|r_{R_2}^2| \le 2u$$

with probability at least

$$\begin{aligned}\mathbb{P}\left\{\mathcal{E}_1(u)\bigcap\mathcal{E}_2(u)\right\} &\ge 1 - 22\exp\left(-\frac{1}{256}\left(\frac{|R_1|}{|R_2|}-1\right)^{-1/2}u\right) - 22\exp\left(-\frac{1}{16}|R_1|^{1/3}u^{2/3}\right) \\ &\quad -10\exp\left(-\frac{|R_2|^{1/3}u^{2/3}}{64\cdot 4^{2/3}}\right) \\ &\ge 1 - 22\exp\left(-\frac{1}{256}\left(\frac{|R_1|}{|R_2|}-1\right)^{-1/2}u\right) - 32\exp\left(-\frac{1}{128}|R_2|^{1/3}u^{2/3}\right).\end{aligned}$$

The statement, when $R_2 \subseteq R_1$, then follows.

Now consider the general case when $R_2 \nsubseteq R_1$. In this case,

$$||R_1|r_{R_1}^2 - |R_2|r_{R_2}^2| \le ||R_1|r_{R_1}^2 - |R_1\cap R_2|r_{R_1\cap R_2}^2| + ||R_2|r_{R_2}^2 - |R_1\cap R_2|r_{R_1\cap R_2}^2|.$$

We can now appeal to the bounds we derived for nested sets before to get

$$\mathbb{P}(||R_1|r_{R_1}^2 - |R_1 \cap R_2|r_{R_1\cap R_2}^2| \geq x) \leq 22\exp\left(-\frac{1}{256}\left(\frac{|R_1 \cap R_2|}{|R_1| - |R_1 \cap R_2|}\right)^{1/2} u\right) + 32\exp\left(-\frac{1}{128}|R_1 \cap R_2|^{1/3}u^{2/3}\right),$$

and

$$\mathbb{P}(||R_2|r_{R_2}^2 - |R_1 \cap R_2|r_{R_1\cap R_2}^2| \geq x) \leq 22\exp\left(-\frac{1}{256}\left(\frac{|R_1 \cap R_2|}{|R_2| - |R_1 \cap R_2|}\right)^{1/2} u\right) + 32\exp\left(-\frac{1}{128}|R_1 \cap R_2|^{1/3}u^{2/3}\right).$$

The first claim then follows from an application of the union bound.

To show the second statement, assume that $|R_1| \geq |R_2|$ without loss of generality. Observe that

$$\rho = \frac{|R_1 \cap R_2|}{\sqrt{|R_1||R_2|}} \leq \sqrt{\frac{|R_2|}{|R_1|}},$$

which implies that $|R_2| \geq \rho^2|R_1|$. Therefore,

$$|R_1 \cap R_2| = \rho\sqrt{|R_1||R_2|} \geq \rho^2|R_1|,$$

and

$$|R_1 \cap R_2| = \rho\sqrt{|R_1||R_2|} \geq \rho|R_2|.$$

Thus,

$$\frac{|R_1 \cap R_2|}{|R_1 \cup R_2| - |R_1 \cap R_2|} \geq \frac{1}{\rho^{-2} + \rho^{-1} - 2} = \frac{1}{1-\rho}\frac{\rho^2}{2\rho+1} \geq \frac{1}{48}(1-\rho)^{-1},$$

where the last inequality follows from the fact that $1/4 \leq \rho \leq 1$. □

We are now in position to derive bounds for the likelihood ratio statistic $L_R$. Since we work with centered random variables as stated earlier, it is natural to redefine $L_R$ as:

$$L_R = -(|R| - 1)\log(1 - r_R^2).$$

where $r_R$ is given by (S3.16).

**Lemma 6.** *Assume that* $\{(X_i, Y_i) : i \in R\}$ *are iid copies of* $(X, Y) \sim N(0, I_2)$ *for some* $|R| > 1$. *Then there exists numerical constants* $0 < c_1 < c_2$ *such that for any* $x > 1$,

$$c_1|R|^{-1/2}e^{-x/2} \leq \mathbb{P}(L_R > x) \leq c_2 e^{-x/2}.$$

*Proof of Lemma 6.* Observe that

$$L_R = (|R| - 1)\log\left(1 + \frac{T_R^2}{|R| - 1}\right)$$

where

$$T_R = r_R\sqrt{\frac{|R| - 1}{1 - r_R^2}}$$

and

$$r_R = \frac{\sum_{i\in R} X_i Y_i}{\sqrt{\sum_{i\in R} X_i^2 \sum_{i\in R} Y_i^2}}$$

It is well known that, under the null hypothesis,

$$T_R \sim t_{|R|-1}.$$

See, e.g., Hotelling (1953). By Lemma 1,

$$c_1 |R|^{-1/2} e^{-x/2} \le \mathbb{P}(L_R > x) \le c_2 e^{-x/2},$$

for any $x > 1$. □

The following lemma bounds the change in the likelihood ration statistic due to a perturbation of the index set.

**Lemma 7.** *Assume that* $\{(X_i, Y_i) : i \in R_1 \cup R_2\}$ *are iid copies of* $(X, Y) \sim N(0, I_2)$*, and* $2|R_1 \cap R_2| \ge |R_1 \cup R_2|$*. Then there exist numerical constants* $c_0, c_1, c_2 > 0$ *such that for any* $x < c_0|R_1|$*,*

$$\mathbb{P}(|L_{R_1} - L_{R_2}| \ge x) \le c_1 \exp\left(-c_2 \min\left\{\left(\frac{|R_1 \cap R_2|}{|R_1 \cup R_2| - |R_1 \cap R_2|}\right)^{1/2} x, |R_1 \cap R_2|^{1/3} x^{2/3}\right\}\right).$$

*In particular, if*

$$\zeta := \frac{|R_1 \cap R_2|}{\sqrt{|R_1||R_2|}} \ge \frac{1}{4},$$

*then there exists a numerical constant* $c_3 > 0$ *such that for any* $x < c_0|R_1|$,

$$\mathbb{P}(|L_{R_1} - L_{R_2}| \geq x) \leq c_1 \exp\left(-c_3 \min\left\{(1-\zeta)^{-1/2}x, |R_1 \cap R_2|^{1/3}x^{2/3}\right\}\right).$$

*Proof of Lemma 7.* Similar to Lemma 5, it suffices to prove the first statement when $R_2 \subseteq R_1$. By the convexity of $-\log(1-x)$, we can ensure

$$L_{R_1} = -|R_1|\log(1-r_{R_1}^2) \geq -|R_1|\log(1-r_{R_2}^2) + \frac{|R_1|(r_{R_1}^2 - r_{R_2}^2)}{1-r_{R_2}^2}$$

and

$$L_{R_2} = -|R_2|\log(1-r_{R_2}^2) \geq -|R_2|\log(1-r_{R_1}^2) + \frac{|R_2|(r_{R_2}^2 - r_{R_1}^2)}{1-r_{R_1}^2}$$

Therefore,

$$|L_{R_1} - L_{R_2}| \leq (|R_2| - |R_1|)\log(1 - \max\{r_{R_1}^2, r_{R_2}^2\}) + \frac{|R_2||r_{R_2}^2 - r_{R_1}^2|}{1 - \max\{r_{R_1}^2, r_{R_2}^2\}}. \tag{S3.19}$$

We now bound the two terms on the right hand side separately.

Denote by $\mathcal{E}(\alpha)$ the event that

$$\max\{r_{R_1}^2, r_{R_2}^2\} < \alpha.$$

By Lemma 4,

$$\mathbb{P}\{\mathcal{E}(\alpha)\} \geq 1 - 12\exp(-|R_2|\alpha/64).$$

Note that, for any $0 < x < \alpha$,

$$-\log(1-x) \leq \frac{x}{1-\alpha}.$$

We can upper bound the first term on the right hand side of (S3.19) by

$$\frac{1}{1-\alpha}\left(|R_1| - |R_2|\right)\max\{r_{R_1}^2, r_{R_2}^2\}.$$

Therefore,

$$\mathbb{P}\{(|R_2| - |R_1|)\log(1 - \max\{r_{R_1}^2, r_{R_2}^2\}) \geq u\} \leq 12\exp\left(-\frac{1}{64}|R_2|\min\left\{\alpha, \frac{(1-\alpha)u}{|R_1| - |R_2|}\right\}\right). \tag{S3.20}$$

The second term of (S3.19) can be upper bounded by

$$\frac{1}{1-\alpha}\left(||R_2|r_{R_2}^2 - |R_1|r_{R_1}^2| + (|R_1| - |R_2|)r_{R_1}^2\right),$$

under the event $\mathcal{E}(\alpha)$. By Lemma 5, we get

$$\mathbb{P}\{||R_2|r_{R_2}^2 - |R_1|r_{R_1}^2| \geq x\} \leq c_1\exp\left(-c_2\min\left\{\left(\frac{|R_2|}{|R_1| - |R_2|}\right)^{1/2}x, |R_2|^{1/3}x^{2/3}\right\}\right).$$

And by Lemma 4,

$$\mathbb{P}\{(|R_1| - |R_2|)r_{R_1}^2 \geq x\} \leq 6\exp\left(-\frac{|R_1|x}{64(|R_1| - |R_2|)}\right)$$

Therefore,

$$\mathbb{P}\left\{\frac{|R_2||r_{R_2}^2 - r_{R_1}^2|}{1-\max\{r_{R_1}^2, r_{R_2}^2\}} \geq u\right\} \leq 1 - 12\exp(-|R_2|\alpha/64) - 6\exp\left(-\frac{(1-\alpha)|R_1|u}{128(|R_1|-|R_2|)}\right) \\ -c_1\exp\left(-\frac{1-\alpha}{2}c_2\min\left\{\left(\frac{|R_2|}{|R_1|-|R_2|}\right)^{1/2}u, |R_2|^{1/3}u^{2/3}\right\}\right).$$

Together with (S3.20), this implies the desired statement for $R_2 \subseteq R_1$. □

A careful inspection of the derivation of Lemma 7 suggests that it can be extended to a more general situation where $X$ and $Y$ are correlated for some indices.

**Lemma 8.** *Let $R_1 \subset R_2$ be two index sets. Assume that $\{(X_i, Y_i) : i \in R_1\}$ are independent observations so that $(X_i, Y_i) \sim N(0, I_2)$ for $i \in R_1$, and $X_i$, $Y_i$ are standard normal random variables with correlation coefficient $\rho$ for $i \notin R_1$. Then there exist numerical constants $c_0, c_1, c_2 > 0$ such that for any $x < c_0|R_1|$,*

$$\mathbb{P}(|L_{R_1} - L_{R_2}| \geq x) \leq c_1\exp\left(-c_2\min\left\{(1-\zeta)^{-1/2}x, |R_1|^{1/3}x^{2/3}\right\}\right).$$

*provided that $\zeta := |R_1|/|R_2| \geq 1/4$.*

Finally we derive a perturbation bounds for a polygon which is useful for our discussion in Section 3.1.

**Lemma 9.** *Let $K_1$ and $K_2$ be two polygons with vertices $u_1, u_2, \ldots, u_k$ and $v_1, v_2, \ldots, v_k$*

*respectively. Denote by $e_j$ the length of the edge between $u_j$ and $u_{(j \bmod k)+1}$. Then*

$$|K_1 \cap K_2^c| \leq r \sum_{j=1}^{k} (e_j + 2r),$$

*where $r$ is maximum distance between $u_j$ and $v_j$.*

*Proof.* Denote by $Q_i$ the polygon whose first $i$ vertices are the same with $K_2$ and whose remaining vertices are the same with $K_1$. In particular, $Q_0 = K_1$ and $Q_k = K_2$. It is not hard to see that the $j$th edge of $Q_i$ is no longer than $e_j + 2r$. If we compare $Q_0$ and $Q_1$, then only the first vertex might be different, as illustrated in Figure S2.

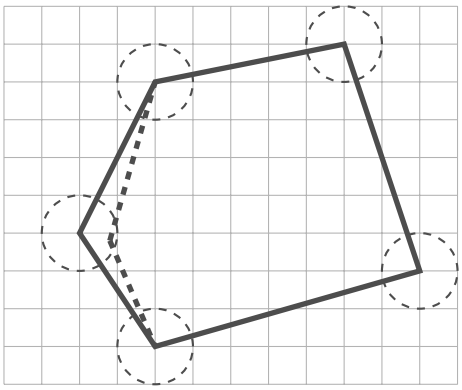


Figure S2: Effect of perturbation of vertices of a polygon.

Because $Q_0$ and $Q_1$ are different only in the first vertex, they can only be different in the two edges linked with the first index. It can then be computed that

$$|Q_0 \cap Q_1^c| \leq \frac{1}{2} r(e_1 + e_k + 4r)$$

Similarly,

$$|Q_i \cap Q_{i+1}^c| \leq \frac{1}{2} r(e_i + e_{i+1} + 4r), \qquad i = 1, 2, \ldots, k-1,$$

It is clear that

$$K_1 \cap K_2^c = Q_0 \cap Q_k^c \subset \cup_{i=0}^{k-1} (Q_i \cap Q_{i+1}^c)$$

Therefore,

$$|K_1 \cap K_2^c| \leq r \sum_{i=1}^{k} (e_i + 2r),$$

which completes the proof. □

*Proof of Proposition 1.* Write

$$\mathcal{C}_{p_1,\ldots,p_k} = \{K(\{(a_i, b_i) : 1 \leq i \leq k\}) : 2^{p_i} \leq r_i < 2^{p_i+1}, i = 1, \ldots, k\}.$$

It is clear that there exists a constant $C > 0$ such that

$$|\mathcal{C}_{p_1,\ldots,p_k}| \leq Cn2^{2\left(\sum_{i=1}^{k} p_i\right)}.$$

Note that there are constants $c_1, c_2 > 0$ depending on $k$ and $M$ only such that

$$\mathcal{R}_{\text{polygon}}(A; k, M) \subset \left\{K \in \mathcal{R}_{\text{polygon}}(k, M) : c_1 A^{1/2} \leq r_i \leq c_2 A^{1/2}, i = 1, 2, \ldots, k\right\}.$$

Therefore,

$$\mathcal{R}_{\text{polygon}}(A; k, M) \leq cnA^k$$

which completes the proof because $A \asymp r_i^2$. □

*Proof of Proposition 2.* Note that $\pi_s(K(\{(a_i, b_i) : 1 \leq i \leq k\}))$ is also a polygon. For brevity, we shall hereafter denote it by $K(\{(\tilde{a}_i, \tilde{b}_i) : 1 \leq i \leq k\})$. By Lemma 9, we get

$$|K(\{(a_i, b_i) : 1 \leq i \leq k\}) \setminus K(\{(\tilde{a}_i, \tilde{b}_i) : 1 \leq i \leq k\})| \leq C2^s \sum_i r_i \leq Ck2^s r_1.$$

Hence

$$\rho\left(K(\{(a_i, b_i) : 1 \leq i \leq k\}), K(\{(\tilde{a}_i, \tilde{b}_i) : 1 \leq i \leq k\})\right) \geq 1 - \frac{Ck2^s r_1}{\pi r_1^2} \geq 1 - \frac{C2^s}{r_1},$$

which completes the proof. □

## Bibliography

Dasgupta, S. and A. Gupta (2003). An elementary proof of a theorem of johnson and lindenstrauss. *Random structures and algorithms 22*(1), 60–65.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological) 15*(2), 193–232.

Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 1302–1338.

Lepski, O. and A. Tsybakov (2000). Asymptotically exact nonparametric hypothesis

testing in sup-norm and at a fixed point. *Probability Theory and Related Fields 117*(1), 17–48.

Walther, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *The Annals of Statistics 38*(2), 1010–1033.