Statistica Sinica Preprint No: SS-2018-0230	
Title	Structured Correlation Detection with Application to Colocalization Analysis in Dual-Channel Fluorescence Microscopic Imaging
Manuscript ID	SS-2018-0230
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0230
Complete List of Authors	Shulei Wang Jianqing Fan Ginger Pocock Ellen T. Arena Kevin W. Eliceiri and Ming Yuan
Corresponding Author	Shulei Wang
E-mail	Shulei.Wang@pennmedicine.upenn.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

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[†]

Abstract: Current workflows for colocalization analysis in fluorescence microscopic imaging introduce significant bias in terms of region of interest (ROI) selection by the user. In this work, we introduce an automatic, unbiased structured detection method of correlated region detection between two random processes observed on a common domain. We argue that although intuitive, direct use of the maximum log-likelihood statistic suffers from potential bias and substantially reduced power and therefore introduce a simple size-based normalization to overcome this problem. We show that scanning with the proposed size-corrected likelihood ratio statistic leads to optimal correlated region detection over a large collection of structured correlation detection problems.

Key words and phrases: optimal rate, scan statistics, signal detection, structured signal, colocalization analysis.

1. Introduction

Most, if not all, biological processes are characterized by complex interactions among bio-molecules. A common way to decipher such interactions is through multichannel fluorescence microscopic imaging where each molecule is labeled with fluorescence of a unique emission wavelength and their biological interactions can be measured by correlation between fluorescently-labeled proteins in user-selected regions of interest (ROIs). Although an ad hoc approach, visual inspection of the overlayed image from both channels is a common first step in determining colocalization in multichannel fluorescence microscopy especially in terms of the spatial location of the colocalization. However, potential pitfalls of this naïve strategy are well-documented, as merged images are heavily influenced by factors such as bleed-through, cross-talk, and relative intensities between different channels. See, e.g., Bolte and Cordeliéres (2006), Comeau et al. (2006) and Dunn et al. (2011).

Since the pioneering work of Manders and his collaborators in early the 1990s, quantitative methods have also been introduced to colocalization analysis. See, e.g., Manders et al. (1992) and Manders et al. (1993). These approaches typically proceed by first manually selecting a region of interest (ROI) where the two molecules are considered likely to colocalize. The degree of colocalization is then determined through various measures of correlation coefficients, most notably Pearson's correlation coefficient or Manders' correlation coefficients, computed specifically within the chosen ROI. See Manders et al. (1993), Costes et al. (2004), Adler et al. (2008), Herce et al. (2013) among others. Obviously, the calculated outcomes of these approaches depends critically on the manually-selected ROI, which not only makes the analysis subjective, but also creates a bottleneck for high-

throughput microscopic image processing. Moreover, even if the region is selected in a principled way, colocalization may not be directly inferred from the value of the correlation coefficient computed within the ROI because the value of the coefficient itself does not translate into statistical significance. This problem could be alleviated using permutation tests, as suggested by Costes et al. (2004). However, one still neglects the fact that the ROI is selected based upon its plausibility of colocalization, introducing significant bias, and the resulting p-value may appear significant merely because of our failure to adjust for the selection bias. The present work is motivated by a clear need for an automated, objective, and statistically valid way to detect regions of colocalization.

Colocalization analysis can naturally be formulated as an example of a broad class of problems that we shall refer to as 'structured correlation detection' where we observe collections of random variables within a common domain and want to determine if there is a region where a subset of these variables are correlated. These types of problems arise naturally in many different fields. For example, in finance, detecting time periods where two common stocks show unusual correlation is essential to the so-called pairs trading strategy (see, e.g. Vidyamurthy, 2004). Other potential examples of structured correlation detection problems can also be found in Chen and Gupta (1997), Robinson et al. (2008), Wieda et al. (2011), and Rodionov (2015), among many others. We shall focus our work on the building of a novel mathematical model for structured correlation detection within the context of colocalization analysis; more specifically, we shall denote the index set of all pixels in the field of view by $\mathbb{I}$. In a typical two or three dimensional image, $\mathbb{I}$ could be a lattice of the corresponding dimension. In practice, it is also possible that $\mathbb{I}$ is a certain subset of a lattice. For example, when investigating intracellular activities, $\mathbb{I}$ only

includes pixels that correspond to the interior of a cell or a compartment (e.g. nucleus) within the cell. For each location $i \in \mathbb{I}$, let X_i and Y_i be the intensities measured at the two channels respectively, as illustrated in the left panel of Figure 1. Hereafter, to fix ideas, (X_i, Y_i)s are assumed to be independent across different pixel i.

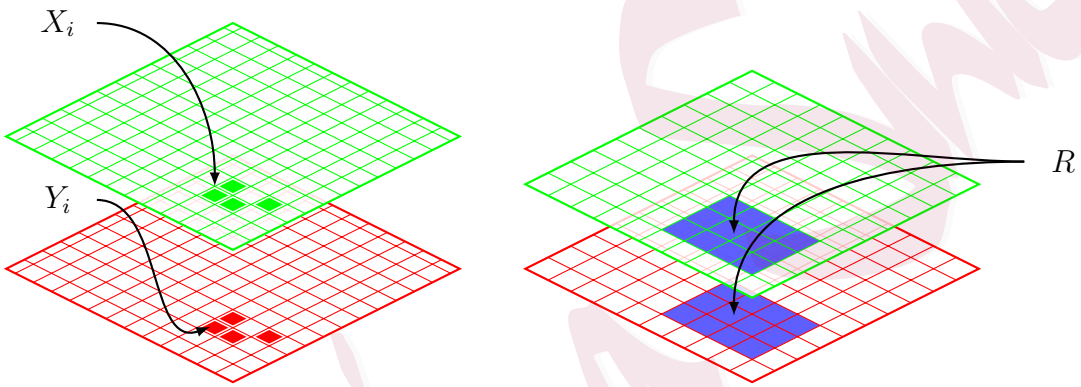


Figure 1: Pixel view of dual channel images.

In the absence of colocalization, we assume that X_i and Y_i are uncorrelated and can be modeled as

$$\begin{bmatrix} X_i \\ Y_i \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right), \tag{1.1}$$

where the marginal means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 may be unknown. In the presence of colocalizaton, X_i and Y_i are correlated, and we therefore treat them as observations from a correlated bivariate normal distribution

$$\begin{bmatrix} X_i \\ Y_i \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right). \tag{1.2}$$

When colocalization occurs, it typically does not occur at isolated locations. As a result, the region R of colocalization is more structured than an arbitrary subset of $\mathbb{I}$. For example, colocalization can frequently be observed within a contiguous region R, as illustrated in the right panel of Figure 1. Let $\mathcal{R}$ be a library containing all possible regions where correlation may be present. For example, $\mathcal{R}$ could be the collection of all ellipses or polygons on a two dimensional lattice ($\mathbb{I}$). The primary goal of correlation detection in general, and colocalization analysis in particular, is to examine whether or not there is an unknown region $R \in \mathcal{R}$ such that (1.1) holds for all $i \in \mathbb{I} \setminus R$, and (1.2) holds for all $i \in R$ and for some $\rho \neq 0$.

The fact that we do not know within which region $R \in \mathcal{R}$ correlation is present naturally brings about the issue of structured multiple testing. Various aspects of structured multiple testing have been studied in recent years. See, e.g., Lepski and Tsybakov (2000), Dümbgen and Spokoiny (2001), Desolneux et al. (2003), Pacifico et al. (2004), Arias-Castro et al. (2005), Dümbgen and Walther (2008), Hall and Jin (2010), Walther (2010), Arias-Castro et al. (2011), Fan et al. (2012), Chan and Walther (2013), Cai and Yuan (2014), and Enikeeva et al. (2015), among many others. However, the problem of colocalization analysis is unique in at least two aspects. Firstly, most if not all existing works focus exclusively on signals at the mean or variance with a single observation at every location. Our interest here, on the other hand, is on the correlation coefficient between two observations at each pixel. Not only do we want to detect signals in terms of the correlation, but we also want to do so in the presence of unknown marginal means and variances as nuisance parameters. Secondly, prior work typically deals with situations where $\mathbb{I}$ is one-dimensional, and $\mathcal{R}$ is a collection of segments, which is amenable

to statistical analysis and sometimes also allows for fast computation. Among the few exceptions are Arias-Castro et al. (2005) who studied several classes of geometrical shapes on a lattice and Walther (2010) who considered rectangles on a two-dimensional lattice. In the case of colocalization analysis, however, the index set $\mathbb{I}$ is multidimensional, and the set $\mathcal{R}$ usually contains more complex geometric shapes. To address both challenges, we have developed a general methodology for correlation detection on a broad domain that is readily applicable to colocalization analysis.

Our method is motivated by an observation that, for a fairly general family of $\mathcal{R}$, the likelihood ratio statistics exhibit a subtle dependence on the size of a candidate region. As a result, their direct use for correlation detection may lead to nontrivial bias, and substantially reduced power. Similar observations have also been made earlier in detecting signals at the mean level in the literature (e.g., Dümbgen and Spokoiny, 2001; Dümbgen and Walther, 2008; Walther, 2010; Chan and Walther, 2013). To overcome this problem, we introduce a size-corrected likelihood ratio statistic and show that scanning with the corrected likelihood ratio statistic yields optimal correlation detection for a large family of $\mathcal{R}$ in the sense that it can detect elevated correlation at a level no other detectors could improve upon significantly. We show that the corrected likelihood ratio statistic based scan can also be computed efficiently for a large collection of geometric shapes in arbitrary dimensions, characterized by their covering numbers under a suitable semimetric. This includes, among others, convex polygons or ellipses, arguably two of the most commonly encountered ROI shapes in practice.

The rest of the paper is organized as follows. In the next section, we introduce a size-corrected likelihood ratio statistic for a general index set $\mathbb{I}$ and collection $\mathcal{R}$ and

discuss how it can be used to automatically detect regions of colocalization. We shall also investigate its efficient implementation, as well as theoretical properties of the proposed method. Section 3 gives several concrete examples of $\mathbb{I}$ and $\mathcal{R}$ and show how the general methodology can be applied to these specific situations. Numerical experiments are presented in Section 5 to further illustrate the merits of the proposed methods. All proofs are relegated to the Supplement Materials due to the limit in space. We feel this new method will have the potential to greatly improve current colocalization analysis workflows, removing the biased pre-selection of ROIs and replacing it with an automatic, robust means of selecting regions of colocalization.

2. Structured Correlation Detection

In a general correlation detection problem, $\mathbb{I}$ can be an arbitrary index set, and $\mathcal{R} \subset 2^{\mathbb{I}}$ is a given collection of subsets of $\mathbb{I}$. We are interested in testing the null hypothesis H_0 that (1.1) holds for all $i \in \mathbb{I}$ against a composite alternative H_a that (1.2) holds for all $i \in R$ whereas (1.1) holds for all $i \notin R$, for some $R \in \mathcal{R}$. We argue here that the usual maximum log-likelihood ratio statistic may not be suitable for correlation detection, and introduce a size-based correction to address the problem.

2.1 Likelihood ratio statistics

A natural test statistic for our purpose is the scan, or maximum log-likelihood ratio statistic:

$$L^* = \max_{R \in \mathcal{R}} L_R,$$

where L_R is the log-likelihood ratio statistic for testing H_0:

$$L_R = -(|R| - 2)\log(1 - r_R^2). \tag{2.3}$$

Here $|R|$ is the cardinality of R and r_R is Pearson correlation within R:

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

where

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

It is worth noting that strictly speaking, L_R defined by (2.3) is not the genuine likelihood ratio statistic, which would replace the factor $|R| - 2$ on the right hand side of (2.3) by $|R|$. Our modification accounts for the correct degrees of freedom so that, for a fixed uncorrelated region R,

$$L_R \approx (|R| - 2)\frac{r_R^2}{1 - r_R^2} \sim t^2_{|R|-2}.$$

See, e.g., Muirhead (2008). Obviously, when $|R|$ is large, L_R approximately follows a χ^2_1 distribution, and the effect of such correction becomes negligible.

The use of scan or maximum log-likelihood ratio statistics for detecting spatial clusters or signals is very common across a multitude of fields. See, e.g., Fan (1996), Fan et al. (2001) and Glaz et al. (2001) and references therein. Their popularity is also justified, as it is well known that scan statistics are minimax optimal if $|R|$ is small when compared with $|\mathbb{I}|$. See, e.g., Lepski and Tsybakov (2000). However, it is also known that when con-

sidering changes in the mean, these methods may lead to nontrivial bias (e.g., Dümbgen and Spokoiny, 2001; Dümbgen and Walther, 2008). We show here that this is also the case for our task, and such a strategy may not be effective for correlation detection unless $|R|$ is very small. In particular, we show that, in the absence of a correlated region, the magnitude of L_R depends critically on its size $|R|$, and therefore, the maximum of L_R's over regions of different sizes is typically dominated by those evaluated on smaller regions. As a result, direct use of L^* for correlation detection could be substantially conservative in detecting larger correlated regions.

We now examine the behavior of the maximum of L_R for $R \in \mathcal{R}$ of a particular size. Note that it is possible that there is no element in $\mathcal{R}$ that is of a particular size. To avoid lengthy discussion to account for such trivial situations, we shall consider instead the subset

$$\mathcal{R}(A) = \{R \in \mathcal{R} : |R| \in (A/2, A]\},$$

for some positive A. In other words, $\mathcal{R}(A)$ is the collection of all possible correlated regions of size between $A/2$ and A. The factor of $1/2$ is chosen arbitrarily and can be replaced by any constant in $(0, 1)$. Basically, $\mathcal{R}(A)$ includes elements of $\mathcal{R}$ that, roughly speaking, are of size A. It is clear that

$$L^* = \max_A \left\{ \max_{R \in \mathcal{R}(A)} L_R \right\}.$$

We shall argue that $\max_{R \in \mathcal{R}(A)} L_R$ may have different magnitudes for different As under the null hypothesis. In particular, we shall show that for a large collection of $\mathcal{R}(A)$, $\max_{R \in \mathcal{R}(A)} L_R$ can be characterized precisely.

Obviously, the behavior of $\max_{R\in\mathcal{R}(A)} L_R$ depends on the complexity of $\mathcal{R}(A)$. More specifically, we shall first assume that the possible correlated regions are indeed more structured than arbitrary subsets of $\mathbb{I}$ in that there exist constants $c_1, c_2 > 0$ independent of A and $n := |\mathbb{I}|$ such that

$$|\mathcal{R}(A)| \leq c_1 n A^{c_2}. \tag{2.4}$$

In other words, (2.4) dictates that $|\mathcal{R}(A)|$ increases with A only polynomially. This is to be contrasted with the completely unstructured setting where $\mathcal{R} = 2^{\mathbb{I}}$, the collection of all subsets of $\mathbb{I}$, and the number of all subsets of $\mathbb{I}$ of size A is of the order n^A, which depends on A exponentially. Condition (2.4) essentially requires that $\mathcal{R}$ is a much smaller subset of $2^{\mathbb{I}}$ and therefore indeed imposes structures on the possible regions of correlation.

Naïve counting of the size of $\mathcal{R}(A)$ as above, however, may not reflect its real complexity. To this end, we also need to characterize the dissimilarity of elements of $\mathcal{R}(A)$. For any two sets $R_1, R_2 \in 2^{\mathbb{I}}$, write

$$d(R_1, R_2) = 1 - \frac{|R_1 \cap R_2|}{\sqrt{|R_1||R_2|}}.$$

It is easy to see that $d(\cdot,\cdot)$ is a semimetric on $2^{\mathbb{I}}$. We now consider the covering number of sets of a particular size in $\mathcal{R}$ under d. Let $N(A,\epsilon)$ be the smallest integer such that there is a subset, denoted by $\mathcal{R}_{\text{app}}(A,\epsilon)$, of $\mathcal{R}$ with

$$|\mathcal{R}_{\text{app}}(A,\epsilon)| = N(A,\epsilon)$$

and

$$\sup_{R_1 \in \mathcal{R}(A)} \inf_{R_2 \in \mathcal{R}_{\text{app}}(A,\epsilon)} d(R_1, R_2) \leq \epsilon.$$

It is worth emphasizing that we require the covering set $\mathcal{R}_{\text{app}}(A, \epsilon) \subset \mathcal{R}$. It is clear that $N(A, \epsilon)$ is a decreasing function of ϵ and $N(A, 0) = |\mathcal{R}(A)|$. We shall also adopt the convention that $N(A, 1)$ represents the largest number of non-overlapping elements from $\mathcal{R}(A)$. Clearly, without any structural assumption, we can always divide $\mathbb{I}$ into n/A subsets of size A. We shall assume that the collection $\mathcal{R}(A)$ is actually rich enough that

$$N(A, 1) \geq c_3 \frac{n}{A}, \tag{2.5}$$

for some constant $c_3 > 0$. Conversely, we shall assume also that there are not too many "distinct" sets in $\mathcal{R}(A)$ in that there are certain constants $c_4, c_5, c_6 > 0$ independent of A and N such that

$$N(A, \epsilon) \leq c_4 \frac{n}{A} \left(\log \frac{n}{A} \right)^{c_5} \left(\frac{1}{\epsilon} \right)^{c_6}. \tag{2.6}$$

Conditions (2.4), (2.5) and (2.6) are fairly general and hold for many common choices of $\mathcal{R}$. Consider, for example, the case when $\mathbb{I} = \{1, 2, \ldots, n\}$ is a one-dimensional sequence and

$$\mathcal{R} = \{(a, b] : 0 \leq a < b \leq n\}$$

is the collection of all possible segments on $\mathbb{I}$. It is clear that there are at most $n - \ell$ segments of length ℓ for any $\ell \in (A/2, A]$, which means

$$|\mathcal{R}(A)| \leq \frac{1}{2} nA.$$

In addition, for any A, there are at least $\lfloor n/A \rfloor$ distinct segments

$$\{((i-1)A, iA] : i = 1, \ldots, \lfloor n/A \rfloor\},$$

of length A, implying that (2.5) also holds. On the other hand, it is not hard to see that the collection of all segments starting at $(i-1)\epsilon A/2$ $(i = 1, 2, \ldots)$ of length between $A/2$ and A can approximate any segment of length between $A/2$ and A with approximation error ϵ. Therefore,

$$N(A, \epsilon) \leq \left(\frac{A/2}{\epsilon A/2}\right)\left(\frac{n}{\epsilon A/2}\right) = 2\frac{n}{A}\left(\frac{1}{\epsilon}\right)^2,$$

so that (2.6) also holds. In the next section, we shall consider more complex examples motivated by colocalization analysis and show that these conditions are expected to hold in fairly general settings.

We now show that if $\mathcal{R}(A)$ satisfies these conditions, $\max_{R\in\mathcal{R}(A)} L_R$ concentrates sharply around $2\log(n/A)$.

Theorem 1. *Suppose that (1.1) holds for all $i \in \mathbb{I}$. Assume also that (2.4) and (2.6) hold. Then*

$$\max_{R\in\mathcal{R}(A)} L_R \leq 2\log(n/A) + O_p(\log\log(n/A)), \qquad \text{as } n \to \infty. \tag{2.7}$$

If in addition, (2.5) holds, then

$$\max_{R\in\mathcal{R}(A)} L_R = 2\log(n/A) + O_p(\log\log(n/A)), \qquad \text{as } n \to \infty. \tag{2.8}$$

We adopted a generic chaining (see, e.g., Talagrand, 2000) argument for the proof of Theorem 1. A similar technique was used previously by Dümbgen and Spokoiny (2001) to establish bounds for likelihood ratio statistics in detecting mean shifts in a sequence. One of the main difficulties in using this type of argument is to quantify the dependence between the likelihood ratio statistics evaluated on two overlapping regions, which is considerably more involved for correlation coefficients than for normal means. More recently, Rivera and Walther (2013) argued that, instead of generic chaining, one could also take advantage of the classical result on the maximum of subgaussian random variables by considering the square root of the likelihood ratio statistics. Moreover, they show that it may also be possible, if $\mathcal{R}$ consists of one dimensional segments, to simplify the technical argument by making explicit use of the properties of an approximation set of $\mathcal{R}$. Similar arguments are also made by Walther (2010) to treat rectangles on a two dimensional lattice. It is not immediately clear to what extent their techniques could be applied in our setting because of, again, the difficulty in characterizing the dependence structure among L_Rs and the generality of the library $\mathcal{R}$.

2.2 Size-corrected likelihood ratio statistics

An immediate consequence of Theorem 1 is that the value of L^* alone may not be a good measure of the evidence of correlation. It also depends critically on the size of R for which L_R is maximized. As such, when using L^* as a test statistic, the critical value is largely driven by $\max_{R\in\mathcal{R}(A)} L_R$ corresponding to smaller A's. Therefore, a test based on L^* could be too conservative when correlation is present on a region with a large cardinality. Several remedies have been proposed in the literature to overcome

this hurdle when considering detecting mean shifts (e.g., Dümbgen and Spokoiny, 2001; Dümbgen and Walther, 2008; Chan and Walther, 2013). Following a similar spirit, we now consider normalizing $\max_{R\in\mathcal{R}(A)} L_R$, leading to a size-corrected log-likelihood ratio statistic:

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

For brevity, we shall hereafter assume that $\max_{R\in\mathcal{R}} |R| \le n/4$. In general, we can always replace $\log x$ by $\log_+(x) := \log(\max\{x, 1\})$ to avoid the trivial cases where the logarithms may not be well defined. After size correction, Theorem 1 suggests that T^* is bounded almost sure when $n \to \infty$ and is not dominated by statistics evaluated on small regions.

It is clear that under the null hypothesis, the distribution of T^* is invariant to the nuisance parameters and therefore can be readily evaluated through Monte Carlo simulation. More specifically, one can simulate $(X_i^*, Y_i^*)^\top \sim N(0, I_2)$ independently for $i \in \mathbb{I}$, and compute T^* for the simulated data. The distribution of T^* can be approximated by the empirical distribution of the test statistics estimated by repeating this process. Denote by q_α the $(1-\alpha)$-quantile of T^* under the null hypothesis. We shall then proceed to reject H_0 if and only if $T^* > q_\alpha$. The whole test procedure is summarized in Algorithm **??** in Supplement Materials. This clearly is an α-level test by construction. We shall show in Section 4, it is also a powerful test for detecting correlation.

One of the potential challenges for scan statistics is computation. To compute T^*, we need to enumerate all elements in $\mathcal{R}$, which could be quite burdensome. To reduce the

computational cost, Arias-Castro et al. (2005) suggested to evaluate L_R on a carefully chosen approximation set of $\mathcal{R}$ for several specific examples of $\mathcal{R}$. See also Walther (2010), where $\mathcal{R}$ is a collection of rectangles on a two dimensional lattice. A key insight obtained from studying T^* however suggests an alternative to T^* that is more amenable for computation. More specifically, it is noted that although numerous, regions of large size, namely $\mathcal{R}(A)$ with a large A, may have fewer "distinct" elements. As such, we do not need to evaluate L_R on each $R \in \mathcal{R}(A)$, but rather on a smaller covering set $\mathcal{R}(A)$.

With slight abuse of notation, write

$$\mathcal{R}_k = \{R \in \mathcal{R} : |R| \in (2^{-k}n, 2^{-(k-1)}n]\}, \qquad k = 2, \ldots, \lfloor \log_2 n \rfloor + 1.$$

It is clear that $T^* = \max_k T_k^*$ where

$$T_k^* = \max_{R \in \mathcal{R}_k} \left\{ \frac{1}{\log\log(n/|R|)} \left[L_R - 2\log\left(\frac{n}{|R|}\right) \right] \right\}.$$

It turns out that for

$$k \le k_* := \lfloor \log_2 n - 2\log_2 \log n \rfloor,$$

we can approximate T_k^* very well by scanning through only a small number of Rs from $\mathcal{R}_k$. In particular, let $\tilde{\mathcal{R}}_k$ be a $1/(4k^2)$ covering set of $\mathcal{R}_k$ with

$$|\tilde{\mathcal{R}}_k| = N\left(2^{-(k-1)}n, \frac{1}{4k^2}\right).$$

We shall proceed to approximate T_k^* by

$$\tilde{T}_k^* = \max_{R \in \tilde{\mathcal{R}}_k} \left\{ \frac{1}{\log\log(n/|R|)} \left[L_R - 2\log\left(\frac{n}{|R|}\right) \right] \right\},$$

when $k \le k_*$. Denote by

$$\tilde{T}^* = \max_k \tilde{T}_k^*,$$

where, with slight abuse of notation, $\tilde{T}_k^* = T_k^*$ for $k > k_*$. Instead of using T^*, we shall now consider $\tilde{T}^*$ as our test statistic, which is summarized in Algorithm **??** in Supplement Materials. As before, we can compute the $1-\alpha$ quantile $\tilde{q}_\alpha$ of $\tilde{T}^*$ under the null hypothesis by the Monte Carlo method and proceed to reject H_0 if and only if $\tilde{T}^* > \tilde{q}_\alpha$.

Compared with T^*, the new statistic $\tilde{T}^*$ is much more computationally friendly. More specifically, under the complexity condition (2.6), it amounts to computing the corrected likelihood ratio statistic on a total of

$$\begin{aligned} &\sum_{k \le k_*} N\left(2^{-(k-1)}n, \frac{1}{4k^2}\right) + \sum_{k > k_*} N\left(2^{-(k-1)}n, 0\right) \\ \le\ & c_4 (\log 2)^{c_5} 4^{c_6} n (\log n)^{c_5+2c_6+1} + c_1 n (\log n)^{2c_2+1} \end{aligned}$$

sets. In other words, the number of size-corrected likelihood ratio statistics we need to evaluate in computing $\tilde{T}^*$ is linear in n, up to a certain polynomial of logarithmic factor.

3. Correlation Detection on a Lattice

While a general methodology was presented for correlation detection under a generic domain in the previous section, we now examine more specific examples motivated by colocalization analysis in microscopic imaging, and discuss further the operating characteristics of the proposed approach. In particular, we shall focus on correlation detection in a two-dimensional lattice where $\mathbb{I} = \{(i, j) : 1 \leq i, j \leq m\}$ so that $n = m^2$, for concreteness, although the discussion can be extended straightforwardly to more general situations, such as rectangular or higher order lattices.

Most imaging tools allow users to visually identify areas of colocalization, allowing either a convex polygonal or ellipsoidal ROI to be selected by the user prior to colocalization calculations. Motivated by this, we shall consider specifically in this section, the automatic, objective detection of correlated ROIs on either an unknown convex polygonal or ellipsoidal region on a two-dimensional lattice. We show that in both cases, the collection $\mathcal{R}$ of all possible correlated areas satisfies conditions (2.4), (2.5) and (2.6), and therefore the size-corrected scan statistic $\tilde{T}^*$ can be efficiently computed.

3.1 Polygons

We first treat convex k-polygons. Any k-polygon can be indexed by its vertices $\{(a_i, b_i) : 1 \leq i \leq k\}$, and will therefore be denoted by $K(\{(a_i, b_i) : 1 \leq i \leq k\})$. For expositional ease, we focus on the case when the vertices are located on the lattice, although the general case can also be treated with further care. The convexity of a polygon allows us

to define its center as $(\bar{a}, \bar{b})$ where

$$\bar{a} = \frac{1}{k}\sum_{i=1}^{k} a_k, \qquad \text{and} \qquad \bar{b} = \frac{1}{k}\sum_{i=1}^{k} b_i.$$

Denote by

$$r_i = \sqrt{(a_i - \bar{a})^2 + (b_i - \bar{b})^2}$$

the distance from the ith vertex to the center. To fix ideas, we will focus attention on nearly regular polygons, where r_is are of the same order. In this case, the collection of possible correlated regions is:

$$\mathcal{R}_{\text{polygon}}(k, M) = \left\{ K(\{(a_i, b_i) : 1 \le i \le k\}) : \max_i r_i / \min_i r_i \le M \right\}.$$

Recall that

$$\mathcal{R}_{\text{polygon}}(A; k, M) = \{R \in \mathcal{R}_{\text{polygon}}(k, M) : |R| \in (A/2, A]\}.$$

The following result states that (2.4) holds for $\mathcal{R}_{\text{polygon}}(k, M)$.

Proposition 1. *There exists a constant $c > 0$ depending on k and M only such that*

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

We now verify (2.5) for $\mathcal{R}_{\text{polygon}}(k, M)$. To this end, we note that any convex k-

polygon can be identified with a minimum bounding circle as shown in Figure 2. Clearly if two polygons intersect, so do their minimum bounding circles. This immediately implies that (2.5) holds, because we can always place $\lfloor m/r \rfloor^2$ mutually exclusive circles of radius r over an $m \times m$ lattice.

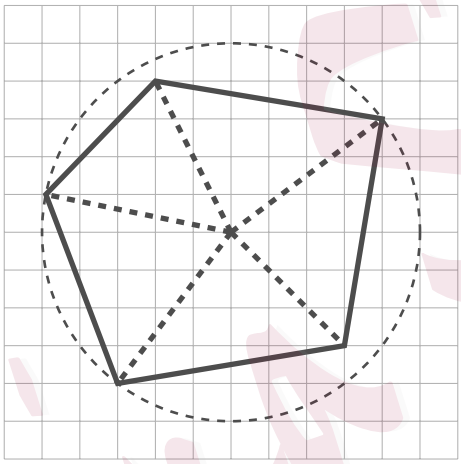


Figure 2: Convex polygon and its minimum bounding circle.

Finally, we show (2.6) also holds for $\mathcal{R}_{\text{polygon}}(k, M)$ by constructing an explicit covering set. The idea is fairly simple, we apply a local perturbation to each vertex:

$$\pi_s(K(\{(a_i, b_i) : 1 \le i \le k\})) = K(\{(2^s \lfloor 2^{-s} a_i \rfloor, 2^s \lfloor 2^{-s} b_i \rfloor) : 1 \le i \le k\}).$$

It can be shown that

Proposition 2. *Let π_s be defined above. Then there exists an absolute constant $c > 0$*

such that

$$d(K(\{(a_i, b_i) : 1 \le i \le k\}), \pi_s(K(\{(a_i, b_i) : 1 \le i \le k\}))) \le c(\min_i r_i)^{-1} 2^s.$$

It is clear that there exist constants $0 < c_7 < c_8$ depending on k and M only such that

$$\mathcal{R}_{\text{polygon}}(A; k, M) \subset \left\{K \in \mathcal{R}_{\text{polygon}}(k, M) : c_7 A^{1/2} \le r_i \le c_8 A^{1/2}, i = 1, 2, \ldots, k\right\}.$$

Therefore, by taking $s = \log_2\left(\epsilon A^{1/2}\right)$, we get

$$N(A, \epsilon) \le c_9 \frac{n}{A} \left(\log\left(\frac{n}{A}\right)\right)^{k-1} \left(\frac{1}{\epsilon}\right)^{2k+2}.$$

In addition, this argument suggests a simple strategy by *digitalization* (π_s) to construct a covering set for $\mathcal{R}$.

From this particular case, we can see the tremendous computational benefit of $\tilde{T}^*$ over T^*. To evaluate T^*, we need to compute the size-corrected likelihood ratio statistics for a total of $|\mathcal{R}| = O(n^k)$ possible regions. In contrast, computing $\tilde{T}^*$ only involves $O\left(n\text{polylog}(n)\right)$ regions as shown in the previous section. Here polylog$(\cdot)$ stands for a certain polynomial of $\log(\cdot)$.

3.2 Ellipses

Next, we consider the case when $\mathcal{R}$ is a collection of ellipses on a two-dimensional lattice. Recall that any ellipse can be indexed by its center $(\tau_1, \tau_2)^\top$, and a positive definite matrix $\Sigma \in \mathbb{R}^{2\times 2}$:

$$\mathcal{E}((\tau_1,\tau_2)^\top,\Sigma) = \left\{ (x_1,x_2)^\top \in \mathbb{R}^2 : (x_1-\tau_1, x_2-\tau_2)\,\Sigma^{-1} \begin{pmatrix} x_1-\tau_1 \\ x_2-\tau_2 \end{pmatrix} \le 1 \right\}.$$

For brevity, we shall consider the case when Σ is well conditioned in that its condition number, that is the ratio between its eigenvalues, is bounded to avoid lengthy discussion about the effect of discretization. In this case,

$$\mathcal{R}_{\text{ellipse}} = \left\{ \mathcal{E}((\tau_1,\tau_2)^\top,\Sigma) \cap \mathbb{I} : 1 \le \tau_1,\tau_2 \le m, \Sigma \succ 0, \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma) \le M \right\}.$$

We first note that any ellipse can be identified with its circumscribing rectangle as shown in Figure 3. Therefore, immediately following the bound on the number of rectangles on a lattice, for example by Proposition 1 with $k = 4$, we get

$$\mathcal{R}_{\text{ellipse}} \le cnA^4,$$

for some constant $c > 0$. Similarly, if two ellipses intersect, then so do their minimum bounding rectangles. By the argument for polygons, we therefore know that (2.5) and (2.6) also hold for $\mathcal{R}_{\text{ellipse}}$.

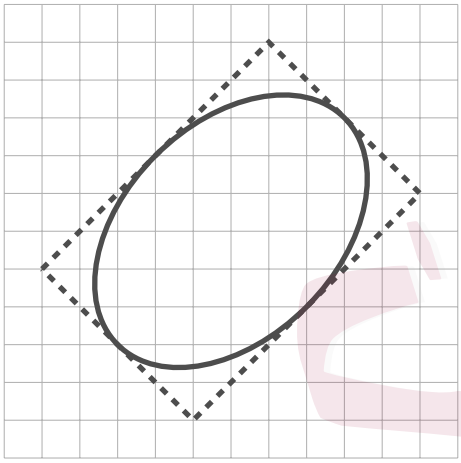


Figure 3: Circumscribing rectangle of an ellipse

4. Optimality

We now study the power of the proposed test T^* and its variant $\tilde{T}^*$. We shall first investigate the required strength of correlation so it can be detected using the proposed tests.

Theorem 2. *Assume that (2.4) and (2.6) hold. If there exists a correlated region $R \in \mathcal{R}$, with $|R| \to \infty$, such that (1.1) holds for $i \notin R$ and (1.2) holds for $i \in R$, and*

$$|R| \log\left(\frac{1}{1-\rho^2}\right) \geq (2+\delta_n) \log\left(\frac{n}{|R|}\right) \tag{4.9}$$

for some $\delta_n > 0$ such that $\delta_n \log^{1/2}(n/|R|) \to \infty$, then $T^* > q_\alpha$ and $\tilde{T}^* > \tilde{q}_\alpha$ with probability tending to one as $n \to \infty$.

Theorem 2 shows that whenever correlation on a region R satisfies (4.9), our tests will consistently reject the null hypothesis and have power tending to one. The detection

boundary of the proposed tests for a correlated region R can therefore be characterized by (4.9). More specifically, depending on the cardinality $|R|$, there are three different regimes.

- For large regions where $|R| \asymp n$, correlation is detectable if $|R|\rho^2 \to \infty$. Recall that, from Neyman-Pearson Lemma, even if the correlated region R is known in advance, we can only consistently detect it under the same requirement. Put differently, the proposed method is as powerful as if we knew the region in advance.

- For regions of intermediate sizes such that $\log n \ll |R| \ll n$, the detection boundary becomes $\rho^2 \geq (2+\delta_n)|R|^{-1}\log(n/|R|)$, provided that $\delta_n\sqrt{\log(n/|R|)} \to \infty$. Here, we can see that weaker correlation can be detected over larger regions.

- And finally for small regions where $|R| \ll \log(n)$, detection is only possible for nearly perfect correlation in that $\rho^2 \geq 1-\exp(-(2+\delta_n)\log(n)/|R|)$ where $\delta_n\sqrt{\log n} \to \infty$.

It turns out that the detection boundary achieved by T^* and $\tilde{T}^*$ as shown in Theorem 2 is indeed sharply optimal following arguments similar to that from Dümbgen and Spokoiny (2001); Dümbgen and Walther (2008); Walther (2010).

Theorem 3. *Assume that (2.5) holds. For any α-level test Δ, there exists an instance where correlation occurs on some $R \in \mathcal{R}$ obeying*

$$|R|\log\left(\frac{1}{1-\rho^2}\right) \geq (2-\delta_n)\log\left(\frac{n}{|R|}\right) \tag{4.10}$$

for a certain $\delta_n > 0$ with $\delta_n \log^{1/2}(n/|R|) \to \infty$, such that the type II error of Δ converges to $1-\alpha$ as $n \to \infty$. Moreover, if there exists some α-level test Δ for which the type II

error converges to 0 as $n \to \infty$ *on any instance where correlation occurs on some* $R \in \mathcal{R}$ *obeying*

$$|R| \log\left(\frac{1}{1-\rho^2}\right) \geq c_n \qquad \textit{and} \qquad |R| \to \infty, \tag{4.11}$$

then it is necessary to have $c_n \to \infty$ *as* $n \to \infty$.

In other words, Theorem 3 shows that any test is essentially powerless for detecting correlation with

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

for any $\delta_n > 0$ such that $\delta_n \log^{1/2}(n/|R|) \to \infty$. Together with Theorem 2, we see that, when $n/|R| \to \infty$, the optimal detection boundary for colocalization for a general index set $\mathbb{I}$ and a large collection of $\mathcal{R}$'s that satisfy certain complexity requirements is

$$|R| \log\left(\frac{1}{1-\rho^2}\right) = 2 \log\left(\frac{n}{|R|}\right);$$

and the size-corrected scan statistic is sharply optimal.

The second statement of Theorem 3 deals with the case when $\limsup n/|R|$ is finite. Together with Theorem 2, (4.11) implies that in this case, the correlated region can be detected if and only if

$$\rho^2 |R| \to \infty$$

and size-corrected scan statistic is again optimal.

To better appreciate the effect of the size of a correlated region on its detectability, it is instructive to consider the cases where $|R| = n^\alpha$ for some $0 < \alpha < 1$ or $|R| = (\log n)^\alpha$

for some $\alpha > 1$. In the former case, when $|R| = n^{\alpha}$, the detection boundary is

$$\rho^2 = 2(1-\alpha)n^{-\alpha}\log n.$$

In the latter case, when $|R| = (\log n)^{\alpha}$, the detection boundary is

$$\rho^2 = 2(\log n)^{1-\alpha}.$$

In both cases, it is clear that much weaker correlation can be detected on larger regions.

5. Numerical Experiments

We now conduct numerical experiments to further demonstrate the practical merits of the proposed methodology.

5.1 Simulation

We begin with a series of four sets of simulation studies focused on two dimensional lattices. The first set of simulations was designed to show the flexibility of the general method by considering a variety of different shapes of correlated regions, namely the choice of the library $\mathcal{R}$, including axis-aligned rectangles, triangles, and axis-aligned ellipses. We compare the performance of size-corrected likelihood ratio statistic and the uncorrected likelihood ratio statistic to demonstrate the necessity and usefulness of the proposed correction. The second set was carried out to compare the full scan statistic T^* and the nearly linear time scan $\tilde{T}^*$ and illustrate similar performance between the two methods

yet considerable computation gain by using $\tilde{T}^*$. The third and fourth sets of simulation studies were conducted to confirm qualitatively our theoretical findings about the effect of the size $|\mathbb{I}|$ of the lattice and the area A of the correlated region on its detectability. In each case, we shall assume that only the shape of the correlated region is known, and therefore $\mathcal{R}$ is the collection of all regions of a particular shape. In addition, we simulate the null distribution and identify the upper 5% quantile of the null distribution based on 1000 Monte Carlo simulations. We reject the null hypothesis for a simulation run if the corresponding test statistic, T^*, $\tilde{T}^*$, or L^*, exceeds their respective upper quantile. This ensures that each test is at level 5% up to Monte Carlo simulation error.

As argued in the previous sections, our methods can handle a variety of geometric shapes. We now demonstrate this versatility through simulation where we consider detecting a correlated region in the form of a triangle, an ellipse, or a rectangle. In particular, we simulated data on a 32×32 lattice. Correlation was imposed on a right triangle with side length 10, 20, and $10\sqrt{5}$, or an axis-aligned ellipse with short axis 4.94 and long axis 6.36, or a rectangle of size 10×10. The location of these correlated regions was selected uniformly over the lattice.

To assess the power of T^*, we considered two relatively small values of correlation coefficient ρ: 0.2 and 0.4. For comparison purposes, we computed for each simulation run both T^* and the uncorrected maximum likelihood ratio statistic L^*. The experiment was repeated 500 times for each combination of shape and correlation coefficient. The results are summarized in Table 1. These results not only show the general applicability of our method, but also demonstrate the improved power of the size correction we apply.

We now compare the full scan statistic T^* with its more computationally efficient

Shape	Rectangle		Ellipse		Triangle	
ρ	0.2	0.4	0.2	0.4	0.2	0.4
T^*	0.16	0.42	0.25	0.6	0.21	0.58
L^*	0.04	0.20	0.03	0.51	0.03	0.26

Table 1: Power comparison between T^* and L^* for different combinations of shape and correlation coefficient.

variant $\tilde{T}^*$. We focus on the case when the correlated region is known to be an axis-aligned rectangle. The true correlated region is a randomly selected 10×10 rectangle on a 64×64 lattice. We consider a variety of different correlation coefficients 0.2, 0.4, 0.6, and 0.8. The performance and computing time (all tests are implemented in Java and the experiments are run on a computer (Intel Core i7 @2.2 GHz/16GB)) of both tests are reported in Table 2, which is also based on 500 runs for each value of the correlation coefficient. It is clear from Table 2 that the two tests enjoy similar performance, with T^* being slightly more powerful. Yet $\tilde{T}^*$ is much more efficient to evaluate as expected.

	Correlation Coefficient	0.2	0.4	0.6	0.8
Power	T^*	0.108	0.228	0.502	0.708
	$\tilde{T}^*$	0.106	0.214	0.410	0.606
Time (ms)	T^*	444.084	447.236	452.634	453.064
	$\tilde{T}^*$	139.026	139.344	140.554	142.144

Table 2: Comparison between T^* and $\tilde{T}^*$.

We note that the computing gain of $\tilde{T}^*$ over T^* becomes more significant for larger images. In particular, we ran similar scans over lattices of size 256×256, 256×512 and 512×512. The computing time for a typical dataset is presented in Table 3.

We now evaluate the effect of the size of a correlated region on its detectability. In

Size of Lattice	256×256	256×512	512×512
Computing time of T^* (s)	129.942	487.238	1934.996
Computing time of $\tilde{T}^*$ (s)	16.59	45.117	144.206

Table 3: Comparison of computing times for T^* and $\tilde{T}^*$.

the light of the observations made in the previous set of experiments, we focus on using $\tilde{T}^*$ to detect a correlated rectangle on a 64×64 lattice. We consider four different sizes of the correlated rectangle: 5×5, 10×10, 20×20, and 40×40. For each given size of the correlated region, we varied the correlation coefficient to capture the relationship between the power of our detection scheme and ρ. The results summarized in Figure 4 are again based on 500 runs for each combination of size and correlation coefficient of the correlated region. The observed effect of A on its detectability is consistent with the results established in Theorem 2 and Theorem 3: larger regions are easier to detect with the same correlation coefficient.

Our final set of simulations was designed to assess the effect of $\mathbb{I}$. To this end, we consider identifying a 10×10 correlated rectangle on a squared lattice of size 32×32, 64×64, or 128×128. As in the previous example, we repeated the experiment 500 times for each combination of $\mathbb{I}$ and a variety of values of ρ. The results are presented in Figure 5. The observed effect of $|\mathbb{I}|$ is again consistent with our theoretical developments: as the size of lattice increases, detection becomes harder for a region of the same size and correlation.

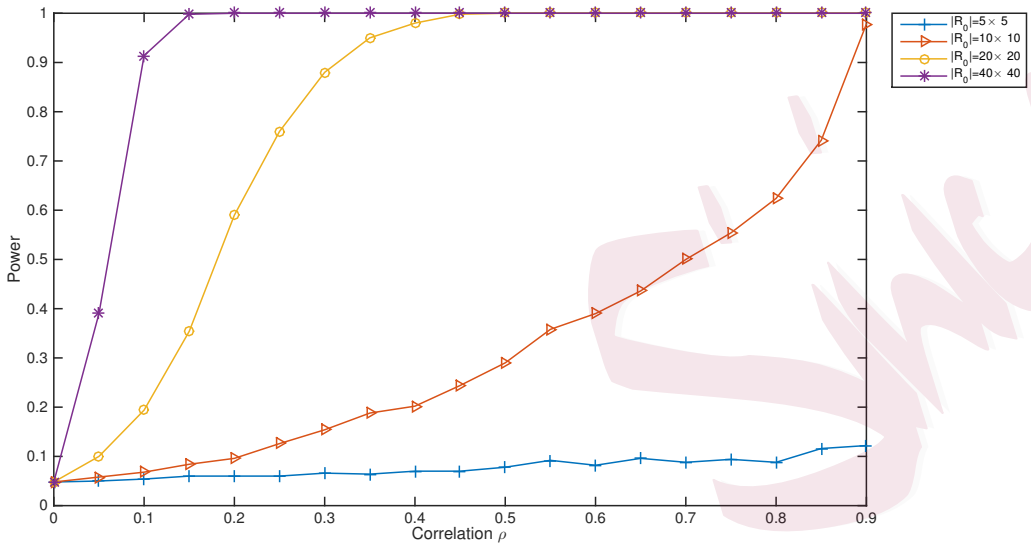


Figure 4: Power plot for detecting a correlated rectangle of different sizes on a 64×64 lattice.

5.2 Real data example

For illustration purposes, we now consider a specific biological dataset examining the post-transcriptional process of human immunodeficiency virus type 1 (HIV-1) using imaging based approaches. HIV uses the host cellular factor chromosome region maintenance 1 (CRM1) mRNA nuclear export pathway to initiate the post-transcriptional stages of the viral life cycle. It is well established that a viral Rev trafficking protein recruits CRM1 nuclear export receptor (Fukuda et al., 1997), having high levels of colocalization during the viral life cycle (Daelemans et al., 2005). HIV-1 genomic RNAs (gRNAs) frequently exhibit burst nuclear export kinetic events (Pocock et al., 2016) that are characterized by en masse evacuations of gRNAs from the nucleus to the cytoplasm; burst nuclear ex-

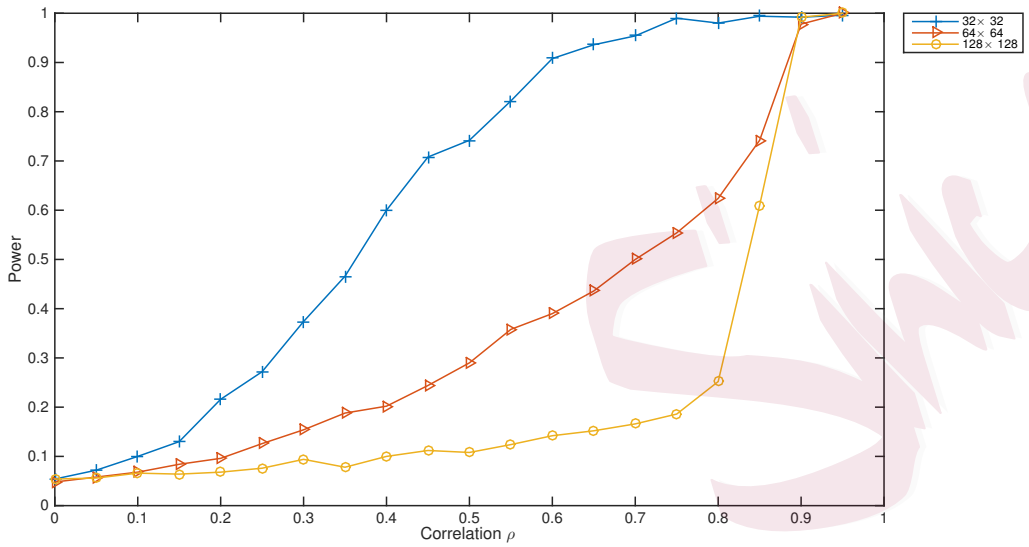


Figure 5: Power plot for detecting a 10×10 correlated rectangle on squared lattices of different sizes.

port is regulated through interactions between Rev and CRM1. Therefore, colocalization analyses of Rev and CRM1 binding can provide insight into the role of Rev in viral gene expression and virus particle assembly.

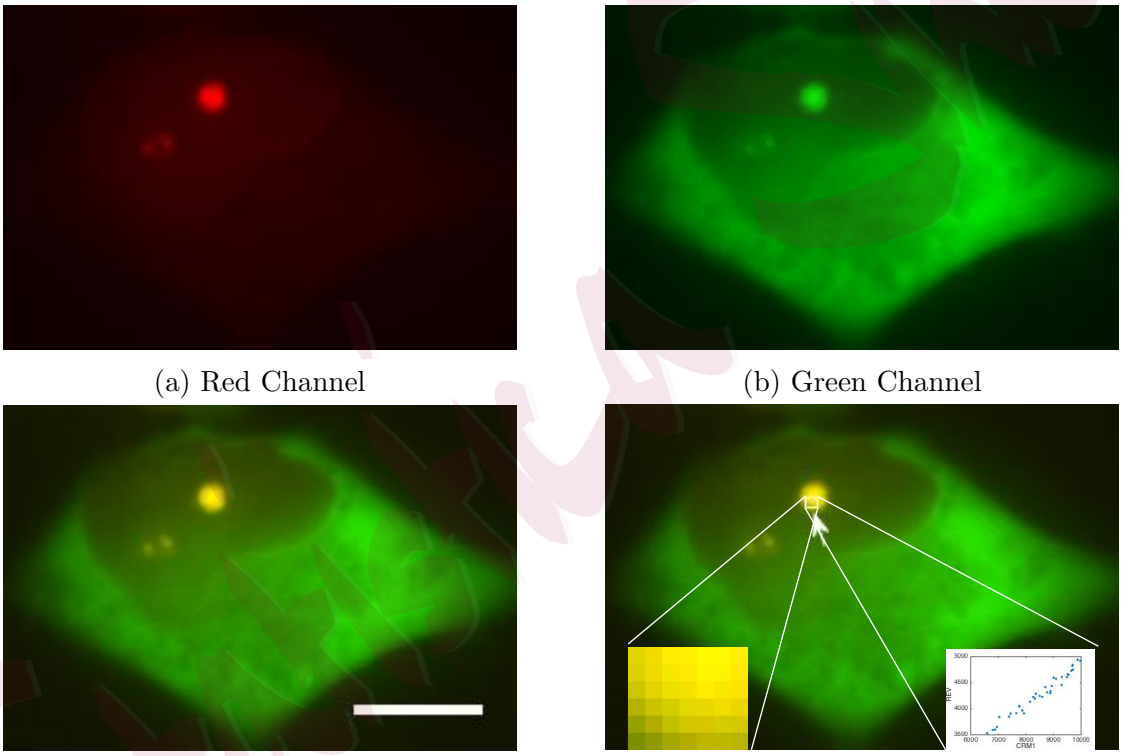
Previous studies have shown a strong association between the viral protein Rev and the CRM1 in the nucleolus (Adler et al., 2008; Daelemans et al., 2005). Therefore, the colocalization between Rev and CRM1 in the nucleolus was compared to a mutant form of Rev (Rev M10) that cannot bind CRM1. This method would provide a measurable way to describe the degree of association between the viral protein Rev and the host protein CRM1 in order to help ascertain their combined roles in the nuclear export of viral genomic RNA (Pocock et al., 2016).

A specific data example is provided in Figure 6 and 7 as dual-channel images of

a cell expressing wild type (WT) Rev (Figure 6) and a cell expressing the Rev M10 mutant, which is unable to bind CRM1 (Figure 7). Imaging experiments were performed on a Nikon Ti-Eclipse inverted wide-field epifluorescent deconvolution microscope (Nikon Corporation) using a 40x Plan Apo (N.A. 0.95) objective with a pixel size of 0.16 μm per pixel. Single images were typically acquired either every 30 minutes using the following excitation/emission filter sets (wavelengths in nm): 490-520/520-550 (YFP) and 565-590/590-650 (mCherry). Sizes of them are 172×255 and 201×281, respectively.

CRM1 is represented as red and Rev by green. While the "burst" gRNA nuclear export phenotype occurs for the WT Rev condition (Figure 6), it does not occur for the condition in which Rev can no longer bind CRM1 (Figure 7). Therefore, the ability of Rev to bind to CRM1 is essential for "burst" nuclear export. To show that degree of association between Rev or RevM10 and CRM1, we applied our method to this particular example following standard pre-processing steps, which included applying a threshold using Otsu's method for each channel to segment the cell and then identify the spatial compartments within where both channels are significantly expressed. On the post-processed images, we computed the test statistic T^* and evaluated its corresponding p-value by simulating the null distribution through 1000 Monte Carlo experiments. For the wild type cell, we obtained $T^* = 3.93 \times 10^3$, which is larger than any of the 1000 values from the Monte Carlo simulations under the null hypothesis, suggesting a p-value $< 0.1\%$, up to a Monte Carlo simulation error. In Figure 6d, we display the region with the largest log-likelihood ratio statistics, its zoomed-in version (left bottom corner) and corresponding scatter plot in this region (right bottom corner). The pixel intensities within the region showed a clear linear relationship. On the other hand, the test statistic for the mutant cell was 77.53, which

corresponds to a p-value of 0.664. This data aligns with the expected levels and, more importantly, the spatial location of colocalization between Rev/Rev M10 and CRM1, confirming the applicability of this region-finding method on biological datasets. It is worth noting that no existing method is able to help identify the location of colocalization automatically.



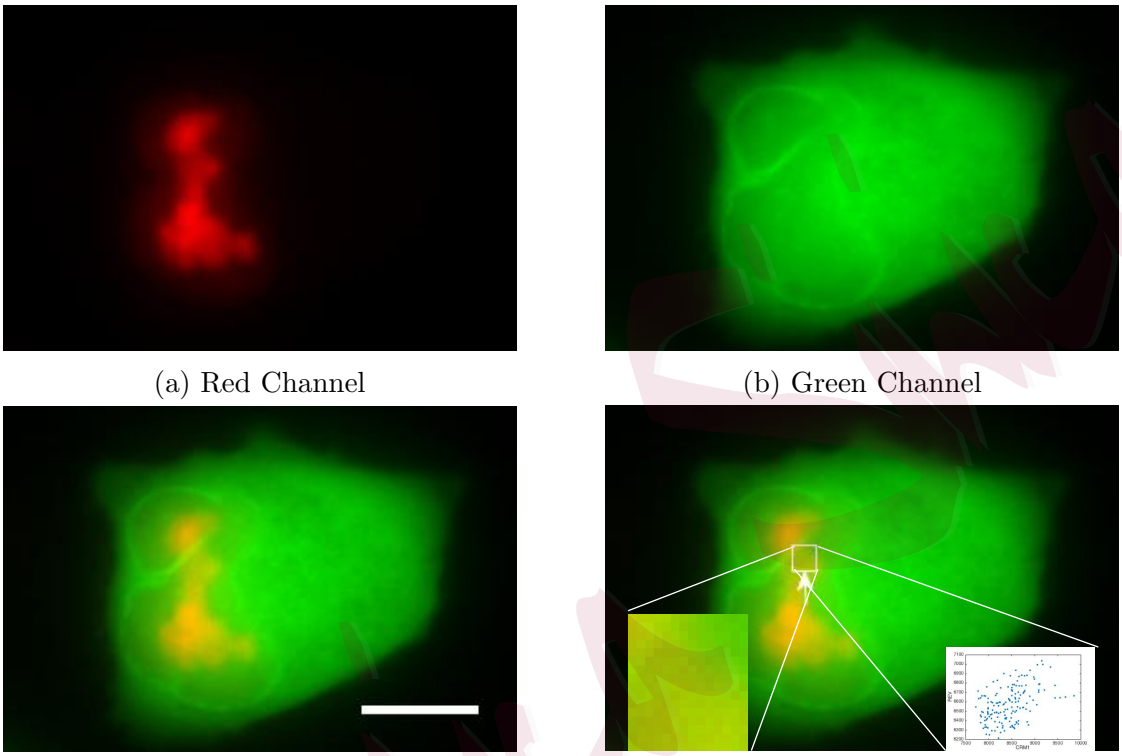
(a) Red Channel (b) Green Channel

(c) Overlay Image (scale bar: 10μm) (d) Region with the Largest Statistics

Figure 6: Colocalization between CRM1 and wild type Rev

6. Discussion

In this paper, we propose a new automated, objective colocalized region detection method for colocalization analysis on dual-channel fluorescence microscopic imaging. When colo-



(a) Red Channel (b) Green Channel

(c) Overlay Image (scale bar: 10μm) (d) Region with the Largest Statistics

Figure 7: No colocalization between CRM1 and Rev M10 mutant

calized region detection is formulated as a structure correlation detection problem, our investigation shows that the maximum of log-likelihood ratio statistics is dominated by those evaluated on small regions and thus is conservative when detecting large correlated regions. To overcome this problem, a size-corrected log-likelihood ratio statistic was proposed to yield optimal correlation detection. The optimal detection statistic can be computed very efficiently, as long as some mild complexity conditions on the shape of correlated regions are satisfied.

The formulation of colocalization analysis we consider in this paper can be seen as a generalization of formulation by traditional methods. More specifically, most existing

colocalization analysis methods, such as Pearson correlation coefficient and Manders' split coefficients, can be cast as a statistics for testing hypothesis within a fixed region of interest in image (see, e.g. Wang et al., 2018). In contrast, the goal of structure correlation detection problem is to test existence of a small colocalized region without any location information input. Because of the new formulation, the newly proposed method does not need input of region of interest (ROI) and thus avoids the selection bias brought by subjective ROI. Moreover, the proposed method is also able to provide the unique information on location of colocalization, which no existing methods can provide.

Although the theoretical analysis in this paper only focuses on the detection of a single correlated region, the analysis can also be extended to multiple regions if some regularity assumptions on regions are satisfied (see e.g. Jeng et al., 2010). In practice, we would recommend adopting the multiple regions detection strategy in Jeng et al. (2010). Let $\mathcal{R}_1$ be the collection of all significant regions, i.e. the regions statistics which are larger than the critical value, q_α. Firstly, we pick up the most significant region R_s from $\mathcal{R}_1$ (i.e. the region with the largest statistics). Secondly, we remove all regions overlapping with R_s from $\mathcal{R}_1$. The two steps above can be repeated until $\mathcal{R}_1$ is empty, i.e. no significant regions available. In this way, multiple regions can be detected.

The results of this paper are mainly presented under Gaussian distribution assumption. However, when the Gaussian assumption is violated, the proposed method is still applicable. More specifically, under non-Gaussian distribution, the parameter of interest ρ is no longer a parameter of bivariate gaussian distribution, but the linear correlation

coefficient between X_i and Y_i

$$\rho := \frac{\mathbb{E}(X_i - \mu(X))(Y_i - \mu(Y))}{\sqrt{\mathbb{E}(X_i - \mu(X))^2 \mathbb{E}(Y_i - \mu(Y))^2}},$$

where $\mu(X)$ and $\mu(Y)$ are expectation of X_i and Y_i, respectively. The concentration inequalities suggest that the key lemmas, including Lemma **??** and Lemma **??**, still hold on large enough region when the underlying distributions are sub-Gaussian distribution or sub exponential distribution (see e.g. Vershynin, 2010). Thus, we can still apply the same size correction technique and derive similar detection upper bound in Theorem 1 and Theorem 2, up to a constant, by the generic chaining. To illustrate this, we conducted a small experiment to compare $\max_{R \in \mathcal{R}(A)} L_R$ when distributions of (X_i, Y_i)s are Gaussian and Poisson distribution, respectively. Specifically, we generate 2000 pairs of random variables on a line and let $\mathcal{R}$ be collection of segments. (X_i, Y_i)s are generated from independent standard Gaussian distribution or independent Poisson distribution with mean 10. We repeat the simulation 5000 times and summarize the distribution of $\max_{R \in \mathcal{R}(A)} L_R$ when $A = 50$ in Figure **??** and **??**. The figures show the distributions are almost the same and confirm our arguments. On the other hand, when the region size is small enough, the form of L_R is specifically designed for Gaussian distribution and Lemma **??** does not always hold. In Figure **??** and **??**, we repeat the above simulation to compare distribution of $\max_{R \in \mathcal{R}(A)} L_R$ when $A = 10$. The figures suggest that there is only a little difference between two distributions. Hence, our newly proposed method is a robust approach to detect linear correlation on large regions.

In this paper, our focus is to detect the existence of colocalization in microscopic

image, which can be seen as one sample hypothesis test problem. However, in many applications, the interest of biologists lies in determining if the level of colocalization differs under different conditions (e.g. experiment group v.s. control group), which is basically a two samples hypothesis test problem. Applying the technique in this paper to two samples problem is not straightforward as registration issues between cells under different conditions arises in scanning. Nevertheless, extending the application to two samples case poses a promising direction for future research.

Supplementary Materials In supplemental materials, we provide structure correlation detection algorithm and the detailed proofs of theoretical results.

Acknowledgements The research of Shulei Wang and Ming Yuan was supported in part by NSF FRG Grant DMS-1265202 and NIH Grant 1U54AI117924-01. The research of Jianqing Fan was supported in part by NSF Grants DMS-1206464 and DMS-1406266 and NIH grants R01-GM072611-11. The research of Kevin W. Eliceiri was supported in part by NIH R01CA185251. Ming Yuan wishes to thank Paul Ahlquist and Nathan Sherer for introducing him to colocalization analysis in microscopic imaging, and Richard Samworth for helpful discussions and careful reading of an earlier draft that have led to much improved presentation.

References

Adler, J., S. Pagakis, and I. Parmryd (2008). Replicate-based noise corrected correlation for accurate measurements of colocalization. *Journal of Microscopy 230*(1), 121–133.

Arias-Castro, E., E. Candès, and A. Durand (2011). Detection of an anomalous cluster

in a network. *The Annals of Statistics*, 278–304.

Arias-Castro, E., D. Donoho, and X. Huo (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Transactions on Information Theory 51*(7), 2402–2425.

Bolte, S. and F. P. Cordeliéres (2006). A guided tour into subcellular colocalization analysis in light microscopy. *Journal of Microscopy 224*(3), 213–232.

Cai, T. and M. Yuan (2014). Rate-optimal detection of very short signal segments. *arXiv preprint arXiv:1407.2812*.

Chan, H. and G. Walther (2013). Detection with the scan and the average likelihood ratio. *Statistica Sinica 23*, 409–428.

Chen, J. and A. Gupta (1997). Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association 92*(438), 739–747.

Comeau, J., S. Costantino, and P. Wiseman (2006). A guide to accurate fluorescence microscopy colocalization measurements. *Biophysical Journal 91*(12), 4611–4622.

Costes, S., D. Daelemans, E. Cho, Z. Dobbin, G. Pavlakis, and S. Lockett (2004). Automatic and quantitative measurement of protein-protein colocalization in live cells. *Biophysical Journal 86*(6), 3993–4003.

Daelemans, D., S. Costes, S. Lockett, and G. Pavlakis (2005). Kinetic and molecular analysis of nuclear export factor crm1 association with its cargo in vivo. *Molecular and cellular biology 25*(2), 728–739.

Desolneux, A., L. Moisan, and J. Morel (2003). Maximal meaningful events and applications to image analysis. *The Annals of Statistics*, 1822–1851.

Dümbgen, L. and V. Spokoiny (2001). Multiscale testing of qualitative hypotheses. *The Annals of Statistics*, 124–152.

Dümbgen, L. and G. Walther (2008). Multiscale inference about a density. *The Annals of Statistics*, 1758–1785.

Dunn, K. W., M. M. Kamocka, and J. H. McDonald (2011). A practical guide to evaluating colocalization in biological microscopy. *American Journal of Physiology-Cell Physiology 300*(4), 723–742.

Enikeeva, F., A. Munk, and F. Werner (2015). Bump detection in heterogeneous gaussian regression. *arXiv preprint arXiv:1504.07390*.

Fan, J. (1996). Test of significance based on wavelet thresholding and neyman's truncation. *Journal of American Statistical Association 91*, 674–688.

Fan, J., X. Han, and W. Gu (2012). Control of the false discovery rate under arbitrary covariance dependence (with discussions). *Journal of American Statistical Association 107*, 1019–1045.

Fan, J., C. Zhang, and J. Zhang (2001). Generalized likelihood ratio statistics and wilks phenomenon. *The Annals of Statistics 29*, 153–193.

Fukuda, M., S. Asano, T. Nakamura, M. Adachi, M. Yoshida, M. Yanagida, and N. E (1997). Crm1 is responsible for intracellular transport mediated by the nuclear export signal. *Nature 390*(6657), 308–311.

Glaz, J., J. Naus, and S. Wallenstein (2001). *Scan statistics*. New York: Springer.

Hall, P. and J. Jin (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics 38*(3), 1686–1732.

Herce, H., C. Casas-Delucchi, and M. Cardoso (2013). New image colocalization coefficient for fluorescence microscopy to quantify (bio-) molecular interactions. *Journal of Microscopy 249*(3), 184–194.

Jeng, J., T. Cai, and H. Li (2010). Optimal sparse segment identification with application in copy number variation analysis. *Journal of the American Statistical Association 105*(491), 1156–1166.

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.

Manders, E., J. Stap, G. Brakenhoff, R. V. Driel, and J. Aten (1992). Dynamics of three-dimensional replication patterns during the s-phase, analysed by double labelling of dna and confocal microscopy. *Journal of Cell Science 103*(3), 857–862.

Manders, E., F. Verbeek, and J. Aten (1993). Measurement of co-localization of objects in dual-colour confocal images. *Journal of Microscopy 169*(3), 375–382.

Muirhead, R. (2008). *Aspects of Multivariate Statistical Theory*. New York, NY: John Wiley & Sons.

Pacifico, M., C. Genovese, I. Verdinelli, and L. Wasserman (2004). False discovery control for random fields. *Journal of the American Statistical Association 99*(468), 1002–1014.

Pocock, G., J. Becker, C. Swanson, P. Ahlquist, and N. Sherer (2016). Hiv-1 and m-pmv rna nuclear export elements program viral genomes for distinct cytoplasmic trafficking behaviors. *PLoS Pathog. 12*(4), e1005565.

Rivera, C. and G. Walther (2013). Optimal detection of a jump in the intensity of a poisson process or in a density with likelihood ratio statistics. *Scandinavian Journal*

of Statistics 40(4), 752–769.

Robinson, L., V. de la Pena, and Y. Kushnir (2008). Detecting shifts in correlation and variability with applications to enso-moonsoon rainfall relationships. *Theoretical and Applied Climatology 94*, 215–224.

Rodionov, S. (2015). Sequential method of detecting abrupt changes in the correlation coefficient and its application to bering sea climate. *Climate 3*, 474–491.

Talagrand, M. (2000). *The Generic Chaining*. New York, NY: Springer-Verlag.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vidyamurthy, G. (2004). *Pairs Trading: Quantitative Methods and Analysis*. New York, NY: Wiley.

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

Wang, S., E. T. Arena, K. W. Eliceiri, and M. Yuan (2018). Automated and robust quantification of colocalization in dual-color fluorescence microscopy: A nonparametric statistical approach. *IEEE Transactions on Image Processing 27*(2), 622–636.

Wieda, D., W. Krämera, and H. Dehling (2011). Testing for a change in correlation at an unknown point in time using an extended functional delta method. *Econometric Theory 28*, 570–589.