Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

LAWS: A Locally Adaptive Weighting and Screening Approach To Spatial Multiple Testing

T. Tony Cai , Wenguang Sun & Yin Xia

To cite this article: T. Tony Cai , Wenguang Sun & Yin Xia (2020): LAWS: A Locally Adaptive Weighting and Screening Approach To Spatial Multiple Testing, Journal of the American Statistical Association, DOI: 10.1080/01621459.2020.1859379

To link to this article: https://doi.org/10.1080/01621459.2020.1859379

LAWS: A Locally Adaptive Weighting and Screening Approach To Spatial Multiple Testing

T. Tony Cai[1,*], Wenguang Sun[2,#], and Yin Xia[3,†]

[1]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.

[2]Department of Data Sciences and Operations, University of Southern California.

[3]Department of Statistics, School of Management, Fudan University.

*The research of Tony Cai was supported in part by NSF Grants DMS-1712735 and DMS-2015259 and NIH grants R01-GM129781 and R01-GM123056.

#The research of Wenguang Sun was supported in part by NSF grants DMS-CAREER 1255406 and DMS-1712983.

†The research of Yin Xia was supported in part by NSFC Grants 12022103, 11771094 and 11690013.

Corresponding author Yin Xia xiayin@fudan.edu.cn

Abstract

Exploiting spatial patterns in large-scale multiple testing promises to improve both power and interpretability of false discovery rate (FDR) analyses. This article develops a new class of locally–adaptive weighting and screening (LAWS) rules that directly incorporates useful local patterns into inference. The idea involves constructing robust and structure-adaptive weights according to the estimated local sparsity levels. LAWS provides a unified framework for a broad range of spatial problems and is fully data-driven. It is shown that LAWS controls the FDR asymptotically under mild conditions on dependence. The finite sample performance is investigated using simulated data, which demonstrates that LAWS controls the FDR and outperforms existing methods in power. The efficiency gain is substantial in many settings. We further illustrate the merits of LAWS through applications to the analysis of 2D and 3D images.

Keywords: adjusted *p*-value; covariate-assisted inference; dependent tests; false discovery rate; structured multiple testing

1 Introduction

1.1 Structural information in spatial multiple testing

Spatial multiple testing arises frequently from a wide variety of applications including functional neuroimaging, environmental studies, disease mapping and astronomical surveys. Intuitively, exploiting spatial structures can help identify signals more accurately and improve the interpretability of scientific findings. There are various ways of incorporating spatial information into the inferential process: the spatial structures and covariates may be utilized to form new hypotheses, define novel error rates, prioritize key tasks and construct new test statistics. For example, in Pacifico et al. (2004) and Heller et al. (2006), pre-determined spatial clusters are used to form new hypotheses with clusters as basic inference units. Benjamini and Heller (2007) suggested that aggregating the data from nearby locations can increase the signal-to-noise ratio and reduce the multiplicity. To reflect the relative importance of decision errors, Benjamini and Heller (2007), Sun et al. (2015), and Basu et al. (2018) proposed to take into account spatial covariates such as the size of a cluster when defining the error rates. Moreover, the prior knowledge on the relationship between individual locations and spatial clusters is highly informative and can be utilized to develop new hierarchical testing and selective inference procedures, which promise to improve both the power and interpretability (Yekutieli, 2008; Benjamini and Bogomolov, 2014).

1.2 Challenges of dependence in multiple testing

The localization of sparse signals from massive spatial data often involves conducting thousands and even millions of hypotheses tests. The false discovery rate (FDR; Benjamini and Hochberg, 1995) provides a powerful and practical criterion for multiplicity adjustment in large-scale testing problems. An important line of research is concerned with the impact of dependence on FDR procedures. The Benjamini-Hochberg (BH) method is shown to be valid for FDR control under a range of dependence settings (Benjamini and Yekutieli, 2001; Sarkar, 2002). In particular, Wu (2008) developed conditions under which the BH method controls the FDR for spatially correlated tests in a hidden Markov random field. Meanwhile, Efron (2007) argued that correlation may degrade statistical accuracy and should be accounted for when conducting simultaneous inference. Optimality under dependence has been

investigated in Sun and Cai (2009), which showed, in a hidden Markov model, that incorporating dependence structure into a multiple-testing procedure can greatly improve the efficiency of conventional approaches that ignore dependence. This idea has been further explored in a range of spatial settings, including the Gaussian random field models (Sun et al., 2015), Ising models (Shu et al., 2015), spatial change-point models (Cao and Wu, 2015), and graphical models (Liu et al., 2016a).

Most spatial multiple testing methods have assumed that clusters are known *a priori*, or the dependence structure can be well estimated from data. However, there are several practical issues. First, the spatial clusters, which are typically formed by aggregating nearby locations according to prior knowledge (e.g. Heller et al., 2006), can be mis-specified. In other works (e.g. Pacifico et al., 2004; Sun et al., 2015), the clusters are obtained by inspecting the testing results from a preliminary point-wise analysis, which can be subjective and highly sensitive to the choice of threshold in the tests at individual locations. Second, contiguous spatial clusters may not serve as an appropriate proxy of reality when signals appearing more frequently in a local area but do not form adjoining regions. Hence it is desirable to develop robust and fully data-driven procedures to capture local patterns in spatial data more accurately. Third, although existing spatial FDR methods have good performances when spatial models are estimated well, the commonly used computational algorithms may not produce desired estimates if assumptions on the underlying spatial process are violated, or the model/prior is misspecified. The poor estimates may lead to less powerful and even invalid FDR procedures. Finally, estimating/modeling spatial dependence structures is very challenging in high-dimensional settings, wherein strong regularity conditions and heavy computations greatly limit the scope and applicability of related works.

1.3 The main idea in our approach

The goal of the present paper is to develop simple and robust FDR methods for spatial analysis that are capable of adaptively learning the sparse structure

of the underlying spatial process without prior knowledge on clusters, parametric assumptions of the underlying model or intensive computation of posterior distributions. The main idea is to recast spatial multiple testing in the framework of simultaneous inference with auxiliary information. Under this framework, the *p*-values play primary roles for assessing the significance, while the spatial locations are viewed as auxiliary variables for providing important structural information to assist inference. We propose a locally adaptive weighting and screening (LAWS) approach that consists of three steps. LAWS first estimates the local sparsity structure using a screening approach, then constructs spatially adaptive weights to reorder the *p*-values, and finally chooses a threshold to adjust for multiplicity. The proposed method bypasses complicated spatial modeling and directly incorporates useful structures into inference. LAWS is nonparametric and assumption-lean – it only requires that the underlying spatial process is smooth at most locations. By capturing unknown spatial patterns adaptively, LAWS tends to up–weight/down–weight the *p*-values in neighborhoods where signals are abundant/sporadic. Our numerical results show that LAWS offers dramatic improvements in power over conventional methods in many settings.

1.4 Connection to existing works and our contributions

Large-scale inference with auxiliary/side information is an important topic that has received much recent attention. There are two lines of research, where the additional information is respectively (i) extracted from the same data set using carefully constructed auxiliary sequences (Liu, 2014; Cai et al., 2019), or (ii) gleaned from secondary data sources such as prior studies and external covariates (Scott et al., 2015; Fortney et al., 2015; Ignatiadis et al., 2016; Basu et al., 2018). Our work departs from these two lines of research in that the side information corresponds to the intrinsic ordering of spatial data. The spatial ordering, which encodes useful patterns such as local clusters and smoothness of the underlying process, is different from conventional auxiliary variables that are either quantitative or qualitative. For example, in the context of inference with side information, the qualitative and quantitative auxiliary variables are often used to create groups to reflect the

inhomogeneity among the hypotheses (Efron, 2008; Ferkingstad et al., 2008; Cai and Sun, 2009). The works on multiple testing with groups show that weighted *p*-values methods can be developed to improve the power of BH (Hu et al., 2010; Liu et al., 2016b; Barber and Ramdas, 2017; Xia et al., 2019). However, the grouping strategy is not suitable for spatial analysis because dividing a region into informative groups requires either good prior knowledge or intensive computation, which becomes infeasible in many scenarios, in particular when hypotheses are located on a two or three dimensional lattice. Moreover, as pointed out by Cai et al. (2019), grouping corresponds to discretizing a continuous variable, which often leads to substantial information loss. By contrast, LAWS directly incorporates the spatial structure into the weights and eliminates the need to define groups.

FDR control via LAWS offers a unified, principled and objective way for exploiting important spatial structures. It has several advantages over recent works on multiple testing with side information such as AdaPT (Lei and Fithian, 2018), SABHA (Li and Barber, 2019) and STAR (Lei et al., 2017). First, LAWS provides a general framework that is capable of handling a broad range of spatial settings. Concretely, SABHA only develops weights for grouped structure and ordered structure along a one-dimensional direction, whereas STAR only works when signals form *contiguous* clusters with convexity or other shape constraints. By contrast, LAWS is applicable to two or three-dimensional settings, and makes no assumption on the contiguity or convexity of the signal process as required by STAR. Second, LAWS is motivated by the optimality theory in Cai et al. (2019) and built upon solid theoretical foundations. We prove that the oracle LAWS method uniformly dominates BH in ranking and propose data-driven methods that asymptotically emulate the oracle. We present both intuitions and numerical results to demonstrate that the weights in LAWS are in general superior to the weights in SABHA. Finally, in contrast with AdaPT, SABHA and STAR whose performances heavily depend on the quality of prior information or human interactions, LAWS is fully data-driven and provides an objective and principled approach to incorporate side information. This feature is attractive

in many scenarios where investigators do not have much flexibility to control the study design or decision-making process. Finally, we develop new theories to prove that LAWS controls the FDR asymptotically under dependence. The theory only requires mild conditions that seem to be substantially weaker than existing results on spatial FDR analysis in the literature.

1.5 Organization

The article is organized as follows. Section 2 introduces the model and problem formulation. Section 3 develops structure–adaptive weights and illustrates its superiority in ranking. In Section 4, we propose the LAWS procedure for spatial multiple testing and study its theoretical properties. Simulation is conducted in Section 5 to investigate the finite sample performance of LAWS and compare it with existing methods. The merits of LAWS are further illustrated in Section 6 through applications to analyzing 2D and 3D images. The proofs are provided in the Appendix.

2 Model and Problem Formulation

Let $\mathcal{S} \subset R^d$ denote a d-dimensional spatial domain and s a location. We focus on a setting where hypotheses are located on a finite, regular lattice $\mathbb{S} \subset \mathcal{S}$ and data are observed at every location $s \in \mathbb{S}$. We consider the infill–asymptotics framework (Stein, 2012) and assume $\mathbb{S} \to \mathcal{S}$ in our theoretical analysis. The setup is suitable for analyzing, say, high-frequency linear network data and fine resolution images from satellite monitoring and neuroimaging[1].

Let $\theta(s)$ be a binary variable, with $\theta(s) = 1$ and $\theta(s) = 0$ respectively indicating the presence and absence of a signal of interest at location s. The identification of spatial signals can be formulated as a multiple testing problem:

$$H_0(s): \theta(s) = 0 \quad \text{versus} \quad H_1(s): \theta(s) = 1, \quad s \in \mathbb{S}. \quad (2.1)$$

Let $\{T(s): s \in \mathbb{S}\}$ be the summary statistic at location *s*. The common practice in multiple testing is to first convert *T*(*s*) to a *p*-value *p*(*s*) and then choose a threshold that corrects for multiplicity. The conditional cumulative distribution functions (CDF) of the *p*-values are given by

$$\mathbb{P}\{p(s) \le t \mid \theta(s)\} = \{1-\theta(s)\}t + \theta(s)G_1(t \mid s), \qquad (2.2)$$

where $t \in [0,1]$ and $G_1(t \mid s)$ is the non-null *p*-value CDF at *s*. The corresponding non-null density is denoted by $g_1(t \mid s)$. Define the sparsity level at location *s*

$$\pi(s) = \mathbb{P}\{\theta(s) = 1\}. \quad (2.3)$$

Due to the existence of spatial correlations and external covariates, signals may appear more frequently in certain regions, and the magnitude of non-null effects may also fluctuate across locations. Consequently we allow $\pi(s)$ and $G_1(t \mid s)$ to vary across the spatial domain to capture important local patterns. A mild condition in our methodological development, characterized precisely in Section 4, is that $\pi(s)$ varies smoothly as a continuous function of *s*. The smoothness in the sparsity levels provides the key structural information, which can be exploited to integrate information from nearby locations and construct more efficient spatial multiple testing procedures.

We focus on point-wise analysis where testing units are individual locations. The decision at location *s* is represented by a binary variable $\delta(s)$, where $\delta(s) = 1$ if $H_0(s)$ is rejected and $\delta(s) = 0$ otherwise. The widely used false discovery rate (Benjamini and Hochberg, 1995) is defined as

$$\mathrm{FDR} = \mathbb{E}\left\{\frac{\sum_{s \in \mathbb{S}} \{1-\theta(s)\}\delta(s)}{\max\{\sum_{s \in \mathbb{S}} \delta(s), 1\}}\right\}. \qquad (2.4)$$

The power of an FDR procedure $\boldsymbol{\delta} = \{\delta(s): s \in \mathbb{S}\}$ can be evaluated using the expected number of true positives:

$$\mathrm{ETP}(\boldsymbol{\delta}) \equiv \Psi(\boldsymbol{\delta}) = \mathbb{E}\left\{\sum_{s\in\mathbb{S}} \theta(s)\delta(s)\right\}. \quad (2.5)$$

It is important to note that although we only consider point-wise tests, the proposed LAWS procedure provides a particularly effective tool for revealing underlying spatial clusters. Hence it may be employed in the preliminary stage of a cluster-wise inference where spatial clusters need to be specified by investigators based on point-wise testing. Moreover, in contrast with existing methods which assume known spatial clusters, LAWS provides a fully data-driven approach to incorporate local structures and does not suffer from possible mis-specifications of the underlying model.

3 Structure–Adaptive Weighting and Its Properties

This section describes a weighted *p*-value approach to spatial FDR analysis. The key idea is to construct weights by exploiting the local sparsity structure in a spatial domain. A multiple testing procedure involves two steps: ranking and thresholding. It can be represented by a thresholding rule of the form $\delta(s,t) = \mathbb{I}\{T(s) \le t\}$, where $T(s)$ is the test statistic to order/rank the hypotheses and t is a threshold for adjusting multiplicity. In Section 3.1, we study how to improve the ranking by exploiting the spatial pattern and constructing structure–adaptive weights to adjust the *p*-values. Further intuition and connections to existing work are discussed in Section 3.2. In Section 3.3, we address the threshold issue and illustrate the superiority of the proposed weighting strategy. Throughout this section we assume that the local sparsity level $\pi(s)$ is known. The setting with unknown sparsity structure is considered in Section 4.

3.1 Incorporating sparsity structure by adjusting the *p*-values

To motivate our weighting strategy, consider the following covariate–adjusted mixture model under the *independence* assumption

$$X(s) \overset{ind}{\sim} f(x \mid s) = \{(1-\pi(s)\} f_0(x \mid s) + \pi(s) f_1(x \mid s), \quad (3.1)$$

where the covariate *s* encodes useful side information, $f_0(x \mid s)$ and $f_1(x \mid s)$ are the null and non-null densities, $\pi(s)$ is the sparsity level and $f(x \mid s)$ is the mixture density. Ignoring the inhomogeneity captured by the covariate *s*, Model (3.1) reduces to the widely used random mixture model (Efron et al., 2001; Newton et al., 2004; Sun and Cai, 2007)

$$X(s) \overset{iid}{\sim} f(x) = (1-\pi) f_0(x) + \pi f_1(x). \quad (3.2)$$

Define the conditional (or covariate–adjusted) local false discovery rate

$$\mathrm{CLfdr}(x \mid s) = \mathbb{P}\{\theta(s) = 0 \mid x, s\} = \frac{\{(1-\pi(s)\} f_0(x \mid s)}{f(x \mid s)}. \qquad (3.3)$$

It follows from the optimality theory in Cai et al. (2019) [Section 4.1] that under Model (3.1), the CLfdr thresholding rule is optimal in the sense that it maximizes the ETP subject to the constraint on FDR.

However, CLfdr cannot handle dependent tests. Under the spatial setting, we aim to develop weighted *p*-values to approximate the optimal ranking by CLfdr. Let

$$\Lambda(x \mid s) = \frac{1-\pi(s)}{\pi(s)} . \frac{f_0(x \mid s)}{f_1(x \mid s)}. \quad (3.4)$$

Then $\mathrm{CLfdr} = \Lambda / (\Lambda + 1)$ is monotone in Λ. The inspection of (3.4) reveals that whether $\theta(s) = 1$ should be decided based on two factors: (a) the information of the sparsity structure that reflects how frequently signals appear in the neighborhood, i.e. $\frac{1-\pi(s)}{\pi(s)}$; (b) the information exhibited by the data itself that indicates the strength of evidence against the null, i.e. $\frac{f_0(x \mid s)}{f_1(x \mid s)}$. The term $\frac{f_0(x \mid s)}{f_1(x \mid s)}$ is extremely difficult to model and calculate, we propose to replace it by the *p*-value, which also captures the evidence against the null in the data. Combining the above concerns, we define the weighted *p*-values:

$$p^w(s) = \min\left\{\frac{1-\pi(s)}{\pi(s)}\, p(s), 1\right\} = \min\left\{\frac{p(s)}{w(s)}, 1\right\},\ s \in \mathbb{S}, \qquad (3.5)$$

where $w(s) = \frac{\pi(s)}{1-\pi(s)}$. Similar to (3.4), the weighted *p*-values (3.5) combines the structural information in the neighborhood and evidence of the signal at a specific location *s*.

3.2 Intuitions and connections to existing weighting methods

Weighting is a widely used strategy for incorporating side information into FDR analyses (Benjamini and Hochberg, 1997; Genovese et al., 2006; Roquain and Van De Wiel, 2009; Basu et al., 2018). Unlike other methods where the side information is acquired *externally* through domain knowledge or prior data, our inference aims to utilize the spatial information, which encodes the *intrinsic structure* of the collected data. The spatial structure is effectively incorporated into inference via $w(s) = \frac{\pi(s)}{1-\pi(s)}$. The key structural assumption, which is suitable for a wide range of applications, is that the local sparsity level $\pi(s)$ varies smoothly in *s*. Our proposed LAWS procedure employs a kernel screening method to estimate $\pi(s)$ by pooling information from points close to *s*. It effectively takes into account important local patterns such as spatial clusters in a data-adaptive fashion. For example, suppose there are many signals in the neighborhood of *s*, then LAWS tends to produce a large estimate of $\pi(s)$, thereby up–weighting the *p*-values in the neighborhood.

The SABHA algorithm by Li and Barber (2019) adopts a different set of weights $w'(s) = \frac{1}{1-\pi(s)}$. Under Model (3.2), SABHA reduces to the methods in Benjamini and Hochberg (2000); Genovese and Wasserman (2002); Storey (2002), who suggested applying BH procedure to adjusted *p*-values $(1-\pi)p(s)$. These works showed that exploiting the *global sparsity structure* π can improve the power of BH by raising the FDR from $(1-\pi)\alpha$ to the nominal level α. The ideas in SABHA and LAWS further illustrate that exploiting the *local patterns* can improve the efficiency even

more dramatically; the idea is formalized in our theoretical analysis in Section 3.3. Compared to the SABHA weight $w'(s) = \frac{1}{1-\pi(s)}$, our weight $w(s) = \frac{\pi(s)}{1-\pi(s)}$ can separate clustered non-null *p*-values more effectively; this is intuitively justified by the connection to the optimality theory (3.3) and confirmed by our simulation studies. Moreover, the motivation, interpretation and estimation of our weighted *p*-values are all fundamentally different from the weights in Hu et al. (2010); Xia et al. (2019), which are developed under the group setting.

Finally, we stress that $w(s)$ only captures the sparsity structure, and the amplitude and variance structures of the underlying spatial process, which is subsumed in the ratio $\frac{f_0(x \mid s)}{f_1(x \mid s)}$, has been intentionally discarded when constructing our weights. This leads to a much simpler and theoretically sound methodology. It remains an open question regarding the information loss when suppressing other structural information in the proposed weights. The heterogeneity issue and the derivation of optimal weighting functions are highly nontrivial (Peña et al., 2011; Ignatiadis et al., 2016; Habiger, 2017; Habiger et al., 2017). Note that existing methods are already very complicated for the independent tests and it would require substantial efforts to extend these methods to the spatial setting.

3.3 A theoretical analysis of ranking

This section demonstrates the benefit of weighting. Let $\boldsymbol{\delta}^{v}(t) = \{\delta^{v}(s,t) : s \in S\}$ denote a class of testing rules where $\delta^{v}(s,t) = \mathbb{I}\{p^{v}(s) \le t\}$, and $p^{v}(s) = \min\left\{\frac{p(s)}{v(s)}, 1\right\}$ with $v(s)$ being the pre-specified weight. Consider the covariate–adjusted *p*-value mixture model (2.2). It is shown in Proposition 2 of Appendix C that, under mild conditions, the FDR of $\boldsymbol{\delta}^{v}(t)$ can be written as

$$\mathrm{FDR}\{\boldsymbol{\delta}^{v}(t)\} = Q^{v}(t) + o(1) = \frac{\sum_{s \in \mathcal{S}} \{1-\pi(s)\} v(s) t}{\sum_{s \in \mathcal{S}} \{1-\pi(s)\} v(s) t + \sum_{s \in \mathcal{S}} \pi(s) G_1\{v(s)t \mid s\}} + o(1). \quad (3.6)$$

The power of $\boldsymbol{\delta}^{v}(t)$ is evaluated using the ETP

$$\Psi\{\boldsymbol{\delta}^{v}(t)\} = \sum_{s\in\mathbb{S}} \pi(s) G_1\{v(s)t \mid s\}.$$

To focus on the main idea, we derive the oracle FDR procedure under an asymptotic setting, which uses the leading term $Q^{v}(t)$ in (3.6) to approximate the actual FDR. Define the oracle threshold $t^{v}_{OR} = \sup\{t : Q^{v}(t) \le \alpha\}$. Then the oracle procedure is

$$\boldsymbol{\delta}^{v}_{OR} \equiv \boldsymbol{\delta}^{v}(t^{v}_{OR}) = [\mathbb{I}\{p^{v}(s) \le t^{v}_{OR}\} : s \in \mathbb{S}].$$

Next we demonstrate that the weighted *p*-values $p^{w}(s)$ defined in (3.5) produces *better ranking* than the unweighted *p*-values. Our basic strategy is to show that at the same FDR level, thresholding (oracle) weighted *p*-value *always* yields larger ETP than unweighted *p*-values. Consider two sets of weights $\{v(s) = 1 : s \in \mathbb{S}\}$ and $\{v(s) = w(s) : s \in \mathbb{S}\}$. The asymptotic FDR and ETP of $\boldsymbol{\delta}^{1}(t)$ and $\boldsymbol{\delta}^{w}(t)$ are denoted by $Q^{1}(t), Q^{w}(t), \Psi^{1}(t)$ and $\Psi^{w}(t)$. The corresponding oracle procedures are defined as $\boldsymbol{\delta}^{1}_{OR} \equiv \boldsymbol{\delta}^{1}(t^{1}_{OR})$ and $\boldsymbol{\delta}^{w}_{OR} \equiv \boldsymbol{\delta}^{w}(t^{w}_{OR})$. The next theorem shows that $\boldsymbol{\delta}^{w}_{OR}$ uniformly dominates $\boldsymbol{\delta}^{1}_{OR}$.

Theorem 1. *Assume that* $\dfrac{\sum_{s\in\mathbb{S}} \pi(s)}{\sum_{s\in\mathbb{S}} \{1-\pi(s)\}} \le 1$. *For each* $s \in \mathbb{S}$, *if the function* $t \to G_1(t \mid s)$ *is concave and the function* $x \to G_1(t / x \mid s)$ *is convex for* $\min_{s\in\mathbb{S}} w^{-1}(s) \le x \le \max_{s\in\mathbb{S}} w^{-1}(s)$, *then we have*

$$(a) Q^{w}(t^{1}_{OR}) \le Q^{1}(t^{1}_{OR}) \le \alpha; \quad (b) \Psi^{w}(t^{w}_{OR}) \ge \Psi^{w}(t^{1}_{OR}) \ge \Psi^{1}(t^{1}_{OR}).$$

Condition $\dfrac{\sum_{s\in\mathbb{S}} \pi(s)}{\sum_{s\in\mathbb{S}} \{1-\pi(s)\}} \le 1$ in Theorem 1 is mild. It only requires that the expected number of alternative hypotheses is smaller than or equal to the expected number of null hypotheses. The condition corresponds to the notion of sparsity that holds trivially in most practical situations. By Theorem 1, we conclude that the superiority of $\boldsymbol{\delta}^{w}_{OR}$ over $\boldsymbol{\delta}^{1}_{OR}$ is due to the improved ranking

via weighted *p*-values since with the same threshold t^1_{OR}, we simultaneously have $Q^w(t^1_{OR}) \le Q^1(t^1_{OR})$ and $\Psi^w(t^1_{OR}) \ge \Psi^1(t^1_{OR})$.

4 Spatial Multiple Testing by LAWS

This section discusses a locally adaptive weighting and screening (LAWS) approach to spatial multiple testing. To emulate the oracle procedure $\boldsymbol{\delta}^w_{OR}$, we need to estimate two unknown quantities: the sparsity level $\pi(s)$ and threshold t^w_{OR}. We first develop a nonparametric screening approach for estimating $\pi(s)$ in Section 4.1, then propose a data-driven procedure to approximate t^w_{OR} in Section 4.2, and finally establish the theoretical properties of LAWS in Section 4.3.

4.1 Sparsity Estimation via Screening

The direct estimation of $\pi(s)$ is very difficult. We instead introduce an intermediate quantity to approximate $\pi(s)$:

$$\pi^\tau(s) = 1 - \frac{\mathbb{P}\{p(s) > \tau\}}{1-\tau}, \quad 0 < \tau < 1. \quad (4.1)$$

We first present some intuitions to explain why $\pi^\tau(s)$ provides a good approximation to $\pi(s)$, then describe a screening approach to estimate $\pi^\tau(s)$ and finally establish the theoretical properties of the proposed estimator.

The *relative bias* of the approximation can be calculated as

$$\frac{\pi^\tau(s) - \pi(s)}{\pi(s)} = -\frac{1 - G_1(\tau \mid s)}{1-\tau}.$$

This result has two implications. First, the bias is always negative, which desirably leads to conservative FDR control as we show in Theorem 2.4. Second, as *τ* becomes larger, we expect that the null *p*-values will become increasingly dominant in the right tail area $[\tau, 1)$ compared to the non-null *p*-values, making the ratio $\frac{1 - G_1(\tau \mid s)}{1-\tau}$ very small. Hence $\pi^\tau(s)$ provides a good approximation to $\pi(s)$ with a suitably chosen *τ*.

We now describe two key steps in estimating $\pi^{\tau}(s)$: smoothing and screening. In the smoothing step, we exploit the structural assumption that $\pi(s)$ [thus $\pi^{\tau}(s)$] varies as a smooth function of spatial location *s*. In reality we only have one observation at location *s*. To pool information from nearby locations, we use a kernel function to assign weights to observations according to their distances to *s*. Specifically, for any given grid $\mathbb{S}$ on $\mathcal{S} \subset \mathbb{R}^d$, let $K: \mathbb{R}^d \to \mathbb{R}$ be a positive, bounded and symmetric kernel function satisfying

$$\int_{\mathbb{R}^d} K(t)dt = 1, \int_{\mathbb{R}^d} tK(t)dt = 0, \int_{\mathbb{R}^d} t^{\mathsf{T}} tK(t)dt < \infty.$$

Denote by $K_h(t) = h^{-1}K(t/h)$, where *h* is the bandwidth. At location *s*, define

$$v_h(s, s') = \frac{K_h(s - s')}{K_h(0)}, \quad (4.2)$$

for all $s' \in \mathbb{S}$. Under the spatial setting, $K_h(s - s')$ is computed as a function of the Euclidean distance $\| s - s' \|$ and $h > 0$ is a scalar. Now consider the quantity $m_s = \sum_{s' \in \mathbb{S}} v_h(s, s')$. We can conceptualize m_s as the "total mass" (or "total number of observations") at location *s*. This is a key quantity in our methodological development. Thus, the smoothing step utilizes the spatial structure to calculate m_s by borrowing strength from points close to *s* while placing little weight on points far apart from *s*.

Next we explain the screening step. Motivated by (4.1), we first apply a screening procedure with threshold τ to obtain a subset $\mathcal{T}(\tau) = \{s \in \mathbb{S} : p(s) > \tau\}$. Suppose we are interested in counting how many *p*-values from the null are greater than τ among the m_s "observations" at *s*. The empirical count, which assumes that the majority cases in $\mathcal{T}(\tau)$ come from the null, is given by

$$\sum_{s' \in \mathcal{T}_{\tau}} v_h(s, s'). \quad (4.3)$$

By contrast, the expected count can be calculated theoretically as

$$\{\sum_{s'\in\mathbb{S}} v_h(s,s')\}\{1-\pi^\tau(s)\}(1-\tau). \qquad (4.4)$$

Setting Equations (4.3) and (4.4) equal, we obtain the following estimate

$$\hat{\pi}^\tau(s) = 1 - \frac{\sum_{s'\in\mathcal{T}_\tau} v_h(s,s')}{(1-\tau)\sum_{s'\in\mathbb{S}} v_h(s,s')}. \quad (4.5)$$

Next we justify the estimator (4.5) by showing that $\hat{\pi}^\tau(s)$ converges to $\pi^\tau(s)$ for every $s\in\mathbb{S}$ as $\mathbb{S}\to\mathcal{S}$ by appealing to the infill–asymptotics framework (Stein, 2012), where the grid $\mathbb{S}$ becomes denser and denser in a fixed and finite domain $\mathcal{S}\in\mathbb{R}^d$. For each $s\in\mathbb{S}$, let $\lambda_{\min}(s)$ and $\lambda_{\max}(s)$ respectively be the smallest and largest eigenvalues of the Hessian matrix $\mathbb{P}^{(2)}(p(s)>\tau)\in\mathbb{R}^{d\times d}$. We introduce the following technical assumptions.

(A1) Assume that $\pi^\tau(\cdot)$ has continuous first and second partial derivatives and there exists a constant $C>0$ that $-C\le\lambda_{\min}(s)\le\lambda_{\max}(s)\le C$ uniformly for all $s\in\mathbb{S}$.

(A2) Assume that $\mathrm{Var}(\sum_{s\in\mathbb{S}}\mathbb{I}\{p(s)>\tau\})\le C'\sum_{s\in\mathbb{S}}\mathrm{Var}(\mathbb{I}\{p(s)>\tau\})$ for some constant $C'>1$.

Proposition 1. *Under (A1) and (A2), if* $h\gg|\mathbb{S}|^{-1}$*, we have, uniformly for all* $s\in\mathbb{S}$*,*

$$\mathbb{E}\{\hat{\pi}^\tau(s)-\pi^\tau(s)\}^2\to 0, \text{ as } \mathbb{S}\to\mathcal{S}.$$

Remark 1. *Assumption (A1) is a mild regularity condition on the alternative CDF* $G_1(\tau\,|\,s)$*. (A2) assumes that most of the p-values are weakly correlated and it can be further relaxed with a larger choice of the bandwidth. For example, with the common choice of* $h\sim|\mathbb{S}|^{-1/5}$*, by the proof of Proposition 1, we can relax (A2) to "*$\mathrm{Var}(\sum_{s\in\mathbb{S}} I\{p(s)>\tau\})\le C'|\mathbb{S}|^c\sum_{s\in\mathbb{S}}\mathrm{Var}(I\{p(s)>\tau\})$ *for some constant* $c<4/5$*", which allows the p-values to be highly correlated.*

4.2 Data-driven procedure

This section describes the proposed LAWS procedure for FDR control. Define the locally adaptive weights

$$\hat{w}(s) = \frac{\hat{\pi}(s)}{1-\hat{\pi}(s)}, \quad s \in \mathbb{S}, \qquad (4.6)$$

where $\hat{\pi}(s)$ is estimated by the screening approach (4.5) [the tuning parameter r has been suppressed in the expression]. To increase the stability of the algorithm, we take $\hat{\pi}(s) = (1-\nu)$ if $\hat{\pi}(s) > 1-\nu$ and take $\hat{\pi}(s) = \nu$ if $\hat{\pi}(s) < \nu$ with $\nu = 10^{-5}$. Next we order the weighted p-values from the smallest to largest. If $\pi(s)$ is known and the threshold is given by t_w, then the expected number of false positives (EFP) can be calculated as

$$\mathrm{EFP} = \sum_{s \in \mathbb{S}} \mathbb{P}\left\{ p^w(s) \le t_w \mid \theta(s) = 0 \right\} \mathbb{P}\left\{ \theta(s) = 0 \right\} = \sum_{s \in \mathbb{S}} \pi(s) t_w. \quad (4.7)$$

It follows that if j hypotheses are rejected along the ranking, then we expect that $\sum_{s \in \mathbb{S}} \hat{\pi}(s) p^{\hat{w}}_{(j)}$ rejections are likely to be false positives. It follows that $j^{-1} \sum_{s \in \mathbb{S}} \hat{\pi}(s) p^{\hat{w}}_{(j)}$ provides a good estimate of the false discovery proportion (FDP). The following step-wise algorithm selects a threshold to maximize the number of rejections subject to the FDP constraint.

Algorithm 1 The LAWS Procedure

1:

Order the weighted p-values from the smallest to largest $p^{\hat{w}}_{(1)}, \cdots, p^{\hat{w}}_{(m)}$ and denote corresponding null hypotheses $H_{(1)}, \cdots, H_{(m)}$.

2:

Let $k^{\hat{w}} = \max\left\{ j : j^{-1} \sum_{s \in \mathbb{S}} \hat{\pi}(s) p^{\hat{w}}_{(j)} \le \alpha \right\}$.

3:

Reject $H_{(1)}, \cdots, H_{(k^{\hat{w}})}$.

Consider the special case where $\hat{\pi}(s) = \hat{\pi}$ for all $s \in \mathbb{S}$. Then LAWS coincides with the SABHA (Li and Barber, 2019), and both recover the methods in Benjamini and Hochberg (2000); Storey (2002); Genovese and Wasserman (2002), which are essentially equivalent to applying the BH algorithm to the adjusted *p*-values $(1-\hat{\pi})p(s)$. However, the ranking by LAWS is substantially different from SABHA when $\pi(s)$ are heterogeneous. Our simulations show that LAWS is more powerful than SABHA and the power gain can be substantial in many settings. Moreover, SABHA does not provide a systematic way to estimate $\pi(s)$. It also requires preordering or grouping of the hypotheses, which is not suitable for handling higher-dimensional spatial settings.

4.3 Theoretical properties

This section studies the theoretical properties of the LAWS procedure. Define the *z*-values by $z(s) = \Phi^{-1}(1-p(s)/2)$, for $s \in \mathbb{S}$, and let $m = |\mathbb{S}|$. Arrange $\{s \in \mathbb{S}\}$ in any pre-specified order $\{s_1, \ldots, s_m\}$ and denote the corresponding *z*-values $\boldsymbol{Z} = (z_1, \ldots, z_m)^{\mathrm{T}}$. We collect below several regularity conditions for the asymptotic error rates control. In spatial data analysis with a latent process $\{\theta(s): s \in \mathcal{S}\}$, the dependence among *p*-values may come from two possible sources: the correlations among *p*-values when $\theta(s)$ are given and the correlations among $\theta(s)$. Our conditions on these two types of correlations are respectively specified in (A3) and (A4).

(A3) Define $(r_{i,j})_{m\times m} = \boldsymbol{R} = \mathrm{Corr}(\boldsymbol{Z})$. Assume $\max_{1\le i<j\le m} |r_{i,j}| \le r < 1$ for some constant $r > 0$. Moreover, there exists $\gamma > 0$ such that $\max_{\{i:\theta(s_i)=0\}} |\Gamma_i(\gamma)| = o(m^{\kappa})$ for some constant $0 < \kappa < \frac{1-r}{1+r}$, where $\Gamma_i(\gamma) = \{j: 1 \le j \le m, |r_{i,j}| \ge (\log m)^{-2-\gamma}\}$.

(A4) Under Model (2.3), there exists a sufficiently small constant $\xi > 0$, such that $\pi^{\tau}(s) \in [\xi, 1-\xi]$, and that $\text{Var}\left[\sum_{s \in \mathcal{S}} \mathbb{I}\{\theta(s) = 0\}\right] = O(m^{1+\zeta})$ for some constant $0 \le \zeta < 1$.

(A5) Define $\mathcal{S}_{\rho} = \left\{i : 1 \le i \le m, |\mu_i| \ge (\log m)^{(1+\rho)/2}\right\}$, where $\mu_i = \mathbb{E}(z_i)$. For some $\rho > 0$ and some $\delta > 0, |\mathcal{S}_{\rho}| \ge [1/(\pi^{1/2}\alpha) + \delta](\log m)^{1/2}$, where $\pi \approx 3.14$ is a math constant.

Remark 2. Condition (A3) assumes that most of the null *p*-values [i.e. given that $\theta(s) = 0$] are weakly correlated. The condition can be fulfilled by a wide class of correlation structures because (i) it still allows each *p*-value to be highly correlated with polynomially growing number of other *p*-values under the null and (ii) we do not impose any conditions on the correlation structures of the *p*-values under the alternative. Condition (A4) only assumes that the latent variables $\{\theta(s) : s \in \mathcal{S}\}$ are not perfectly correlated. It allows highly correlated $\theta(s)$ so is a rather weak condition. In the case where $\theta(s)$ are mutually independent, (A4) is satisfied trivially with ζ = 0. Condition (A5) is mild, as it only requires that there exist a few spatial locations with mean effects of *z*-values exceeding $(\log m)^{(1+\rho)/2}$ for some $\rho > 0$.

Our theoretical analysis is divided into two steps. We first consider the setup where $\pi(s)$ is known (Theorem 2.4) and then turn to the case where $\pi(s)$ must be estimated (Theorem 3). Define the FDP of a decision rule $\boldsymbol{\delta}^{\nu}(t)$ by

$$\text{FDP}\{\boldsymbol{\delta}^{\nu}(t)\} = \frac{\sum_{s \in \mathcal{S}} \{1 - \theta(s)\}\delta^{\nu}(s,t)}{\max\{\sum_{s \in \mathcal{S}} \delta^{\nu}(s,t), 1\}}.$$

We first take $\nu(s)$ as $w(s) = \dfrac{\pi^{\tau}(s)}{1-\pi^{\tau}(s)}$ with known $\pi^{\tau}(s)$. Then similar to Algorithm 1, we order the weighted *p*-values from the smallest to largest $p^{w}_{(1)}, \cdots, p^{w}_{(m)}$, and calculate $k^{w} = \max\left\{j : j^{-1}\sum_{s \in \mathcal{S}} \pi^{\tau}(s)\, p^{w}_{(j)} \le \alpha\right\}$. The

corresponding decision rule, denoted $\boldsymbol{\delta}^w \equiv \boldsymbol{\delta}^w\left\{p^w_{(k^w)}\right\}$, is to reject $H_0(s)$ with $p^w(s) \le p^w_{(k^w)}$.

Let $\mathcal{H}_0$ be the set of null hypotheses and $\mathcal{H}_1$ be the set of alternatives. Without loss of generality, we assume that $m_0 = |\mathcal{H}_0| \ge cm$ for some $c > 0$. (Otherwise we could simply reject all the hypotheses, and the FDR would tend to zero.) The next theorem shows that $\boldsymbol{\delta}^w$ controls both the FDP and FDR at the nominal level asymptotically under dependency.

Theorem 2. *Under Conditions (A3) - (A5), we have for any* $\epsilon > 0$

$$\limsup_{m\to\infty} \mathrm{FDR}(\boldsymbol{\delta}^w) \le \alpha, \text{ and } \lim_{m\to\infty} \mathbb{P}\left\{\mathrm{FDP}(\boldsymbol{\delta}^w) \le \alpha + \epsilon\right\} = 1.$$

Remark 3. The decision rule $\boldsymbol{\delta}^w$ is defined based on the weight $w(s) = \dfrac{\pi^\tau(s)}{1-\pi^\tau(s)}$, where $\pi^\tau(s)$ is a conservative approximation of $\pi(s)$ as explained in Section 4.1. Theorem 2.4 shows that the use of $\pi^\tau(s)$ instead of $\pi(s)$ leads to conservative error rates control.

The next theorem establishes the theoretical properties of the data-driven LAWS procedure (Algorithm 1, with decision rule denoted by $\boldsymbol{\delta}^{\hat{w}} \equiv \boldsymbol{\delta}^{\hat{w}}\left\{p^{\hat{w}}_{(k^{\hat{w}})}\right\}$), which utilizes the estimated weights via (4.5).

Theorem 3. *Under the conditions in Proposition 1 and Theorem 2.4, we have for any* $\epsilon > 0$

$$\limsup_{\mathbb{S}\to\mathcal{S}} \mathrm{FDR}(\boldsymbol{\delta}^{\hat{w}}) \le \alpha, \text{ and } \lim_{\mathbb{S}\to\mathcal{S}} \mathbb{P}(\mathrm{FDP}(\boldsymbol{\delta}^{\hat{w}}) \le \alpha + \epsilon) = 1.$$

5 Simulation

This section conducts simulation studies to compare the proposed LAWS procedure with several competing methods. The implementation details are first described in Section 5.1. Sections 5.2 and 5.3 respectively consider linear block and triangle block patterns. The applications to higher dimensional

settings (2D and 3D) for identifying more complicated spatial patterns are illustrated in Section 6.

5.1 Estimating the conditional proportions

The proposed estimator (4.5) captures the sparsity structure and plays a key role in constructing the weights. This section first discusses its implementation and illustrates its effectiveness. To create the screening subset $\mathcal{T}$, we choose τ as the p-value threshold of the BH procedure at $\alpha = 0.9$. This ensures that the null cases are dominant in $\mathcal{T}$. See Appendix B for a more detailed discussion on the bias-variance tradeoff when calibrating τ. The bandwidth h is set using the "h.cvv" option in the R package `kedd`.

Next we investigate the performance of $\hat{\pi}$ using simulated data. We generate m = 5, 000 hypotheses from the following normal mixture model:

$$X_i \mid \theta_i \overset{ind}{\sim} (1-\theta_i)N(0,1) + \theta_i N(\mu,1), \quad \theta_i \sim \text{Bernoulli}(\pi_i). \qquad (5.1)$$

We consider two setups under which the signals appear with elevated frequencies in the following blocks $[1001,1200],[2001,2200],[3001,3200],[4001,4200]$. The patterns of $\pi(s)$, which are piecewise constants and triangle blocks, are shown in the top and bottom rows in Figure 1 (solid red lines), respectively. We can see that the varying sparsity structure of the spatial data can be reasonably captured by the estimated $\hat{\pi}(s)$ (dashed blue lines). As predicted by theory, our estimated $\hat{\pi}(s)$ tend to be smaller than true $\pi(s)$ within the blocks where signals are observed with elevated frequencies. The *underestimation* of $\pi(s)$ leads to conservative FDR levels. This is confirmed by the simulation in the next section.

5.2 The block–wise 1D setting with piece-wise constants

This section compares LAWS with competitive methods. Similar to the previous section, we generate data from (5.1) under the setup where $\pi(s)$ is a piecewise constant function (top row of Figure 1). The following methods are applied to the simulated data:

- Benjamini-Hochberg procedure (BH);
- SABHA with known $\pi(s)$ (SABHA.OR);
- data-driven SABHA with estimated $\hat{\pi}(s)$ (SABHA.DD);
- LAWS with known $\pi(s)$ (LAWS.OR); and
- data-driven LAWS with estimated $\hat{\pi}(s)$ (LAWS.DD).

We stress that our proposed estimator (4.5) has been used to implement SABHA. The SABHA paper does not provide an estimator of $\pi(s)$ with proven theoretical properties. The inclusion of SABHA is to illustrate the superiority of the LAWS weight $w(s) = \pi(s)/\{1-\pi(s)\}$ over the SABHA weight $1/\{1-\pi(s)\}$. The FDR and average power [defined as $E\{\sum_{s\in\mathcal{S}} \theta(s)\delta(s) / \sum_{s\in\mathcal{S}} \theta(s)\}$] of different methods are computed by averaging over 200 replications, and the nominal level is chosen at $\alpha = 0.05$. The simulation results are summarized in Figure 2.

In the top row, the signals appear with elevated frequencies in the following blocks:

$$\pi(s) = 0.9 \text{ for } s \in [1001,1200] \cup [2001,2200]; \pi(s) = 0.6 \text{ for } s \in [3001,3200] \cup [4001,4200].$$

For rest of the locations, we have $\pi(s) = 0.01$. We vary μ from 2 to 4 to investigate the impact of the signal strength. In the bottom row, we fix $\mu = 2.5$. We let $\pi(s) = \pi_0$ in the above specified blocks and $\pi(s) = 0.01$ elsewhere. Then π_0 is varied from 0.3 to 0.9 to investigate the impact of sparsity structure.

We can see from Figure 2 that all methods control the FDR at the nominal level, with LAWS.DD being conservative due to the underestimation of $\pi(s)$ in the linear blocks (see also Figure 1). LAWS.OR substantially outperforms SABHA.OR, showing the superiority of the LAWS weight. Similarly, LAWS.DD outperforms SABHA.DD. Both LAWS and SABHA, which exploit the varying sparsity structure, outperform BH. This illustrates the benefits of incorporating side information into inference. Finally, we can see that the efficiency gain of

LAWS over competing methods is more pronounced when the signals are relatively weak (top row of Figure 2). This shows the advantage of LAWS, which integrates information from nearby locations via the weighted kernel. Moreover, the power improvement by LAWS is greater when the signals are more concentrated in the designated blocks (bottom row of Figure 2). This is consistent with our intuition since larger π_0 indicates greater disparity among spatial locations (and hence more informative spatial structure).

5.3 The block–wise 1D setting with triangular patterns

We generate data from (5.1) under the setup where $\pi(s)$ follows a triangular block pattern; see the bottom row of Figure 1 for an illustration. We apply BH, SABHA.OR, SABHA.DD, LAWS.OR and LAWS.DD to the simulated data and summarize the results in Figure 3. Similar as before, in the top and bottom rows we respectively vary the signal strength and sparsity levels. We can see that the power of BH is improved by SABHA, which is further improved by LAWS. The proposed method is in particular useful when the signals are weak and the structural information is strong in the spatial data.

5.4 Simulation in 2D setting

This section presents simulation results in the 2D setting. We did not compare with SABHA and STAR because (i) the original SABHA algorithm cannot be implemented since it is unclear how to order the hypotheses as a fixed sequence or divide them into groups; (ii) the STAR algorithm does not work in one of our settings where the underlying shape is not a convex region. Instead, we compare with the FDR smoothing method proposed in Tansey et al. (2018) ("Tansey" in short) for exploiting spatial structure in large multiple-testing problems.

We generate the data by Model (5.1) on a 200 × 200 lattice, where the signals are more likely to be located on a double-triangle or a rectangle shape as shown in Figure 4. We let $\pi(s) = 0.9$ for the left triangle and left half of the rectangle respectively, $\pi(s) = 0.6$ for the right triangle and right half of the rectangle and let $\pi(s) = 0.01$ for the rest of the locations. Similarly as in the 1D

setting, we first vary μ from 2.5 to 4 to investigate the impact of the signal strength. We then fix μ = 3, let $\pi(s) = \pi_0$ in the triangle and rectangle patterns, $\pi(s) = 0.01$ for the rest, and vary π_0 from 0.6 to 0.9 to illustrate the impact of sparsity structure. The empirical FDR and power are computed over 200 replications with nominal level $\alpha = 0.05$.

We can see from Figures 5 and 6 that, all methods except Tansey controls the empirical FDR well and LAWS.DD is slightly more conservative than LAWS.OR due to the negative bias of π^{τ} as explained in Section 4.1. By successfully incorporating the spatial information, the empirical power of BH has been significantly improved by LAWS.OR and LAWS.DD for varying signal strengths and sparsity levels. The improvement is more significant when the signals are weaker or the sparse structure is more informative. Note that, due to the seriously inflated empirical FDRs, Tansey has higher empirical power than the competing methods.

5.5 Simulation in 3D setting

We compare in this section the numerical performance of LAWS.DD and LAWS.OR with Tansey and the BH method in 3D spatial settings. The data are generated by Model (5.1) on a 3D $20 \times 25 \times 30$ lattice, where the signals are located on a cubic with dimension $10 \times 10 \times 15$. We let $\pi(s) = 0.8$ within the cubic and let $\pi(s) = 0.01$ for the rest of the locations. To show the impact of the signal strength and the impact of sparsity structure, we vary μ and $\pi(s)$ (fix $\mu = 3.5$ for the latter) in the same way as the 2D settings. The empirical FDR and power are computed over 200 replications with nominal level $\alpha = 0.05$.

We see from Figure 7 that, similarly as 1D and 2D settings, all methods except Tansey control the FDR and LAWS.DD is slightly more conservative than LAWS.OR; the empirical power of BH is significantly improved by LAWS.OR and LAWS.DD for different signal strengths and sparsity levels.

6 Applications

This section applies LAWS to identify 2D spatial clusters (Section 6.1) and signal patterns in 3D image data (Section 6.2). LAWS has several advantages over existing structure–adaptive testing methods. For example, SABHA requires that the hypotheses can be divided into groups or should be ordered as a fixed sequence, which are not always feasible in 2D and 3D spatial applications. By contrast, LAWS constructs weights based on the distance between spatial locations (4.2) and can easily handle higher dimensional spatial settings. Unlike the STAR procedure (Lei et al., 2017) which requires that the spatial region must be contiguous and convex, LAWS is applicable to a wider types of settings where the local sparsity patterns are heterogeneous. We present two examples to show that LAWS is more accurate and effective in identifying and recovering specific patterns of interest in analysis of 2D and 3D image data.

6.1 The 2D setting with spatial clusters

We simulate data on a 200 × 200 lattice. The signals of interest form two spatial clusters respectively with donut and square shapes. The observations follow the random mixture model (5.1), where $\theta(s)=1$ if s is within the donut or square and $\theta(s)=0$ otherwise. We first obtain $\hat{\pi}(s)$ using (4.5), where $\| s-s' \|$ is calculated as the usual Euclidean distance. We then obtain two-sided p-values and finally apply both BH and LAWS to the simulated data set. From the first to last row, we vary the signal strength from 2.0 to 3.0. The true states, and the rejected locations by BH and LAWS are respectively displayed from Column 1 to Column 3.

We can see that LAW is more powerful than BH in uncovering the underlying truth. Both spatial patterns, namely the donut and square, can be more easily identified based on the results of LAWS. The key idea is that $\hat{\pi}(s)$ tend to be very large in the neighborhood of clustered signals (yellow areas) due to the strong spatial correlations. Therefore the p-values in these neighborhood are upweighted via data-driven weights. We conclude that by exploiting the local sparsity structures, LAWS is more effective in rejecting the hypotheses in

regions where signals appear in clusters. This property is in particular attractive in spatial data analysis.

6.2 The 3D setting: application to fMRI data

We further illustrate the LAWS procedure through a magnetic resonance imaging (MRI) data for a study of attention deficit hyperactivity disorder (ADHD). The dataset is available at `http://neurobureau.projects.nitrc.org/ADHD200/Data.html`. The images were produced by the ADHD-200 Sample Initiative, then preprocessed by the Neuro Bereau.

We first reduce the resolution of MRI images from $256 \times 198 \times 256$ to $30 \times 36 \times 30$ (Li and Zhang, 2017) by aggregating the corresponding pixels into blocks. This helps the analysis in several ways. First, the aggregation of pixels not only increases the signal to noise ratio but also effectively avoids misalignments of brain regions. Second, the *p*-values, which are calculated based on normal approximations, should satisfy the required accuracy needed in the large *p* small *n* paradigm (Fan et al., 2007; Liu and Shao, 2010; Chernozhukov et al., 2017). The downsizing helps to increase the precision of the approximations. Finally, the aggregation can effectively eliminate noises and make it easier to visualize interesting spatial patterns.

The dataset consists of 931 subjects, among whom 356 are combined ADHD subjects and 575 are normal controls. We conduct two-sample t-tests to compare the two groups and use normal approximation to obtain the *p*-values. Finally, we apply LAWS and BH procedures to identify brain regions that exhibit significant differences between subjects with and without ADHD.

Figure 9 displays the testing results from two different angles of the 3D image, with FDR level equal to 0.05. The significant brain regions identified by BH are a subset of those identified by LAWS. To be more specific, the LAWS procedure identifies 538 regions, while BH recovers 349. The graph further shows that LAWS has superior power performance over BH in identifying spatial signals.

7 Discussions

This paper develops a new locally-adaptive weighting approach that incorporates the spatial structure into statistical inference. It provides a unified framework for a broad range of spatial multiple testing problems and is fully data-driven. We show that LAWS controls the FDR asymptotically under dependence and outperforms existing methods in power.

LAWS is powerful yet simple. It is capable of adaptively learning the sparse structure of the underlying spatial process without prior knowledge. The spatial locations are viewed as auxiliary variables for providing important structural information to assist inference. However, as explained in Section 3.1 of the paper, there are two pieces of information that could potentially be useful in spatial setting: the varying sparsity structure that we have considered, and the varying distributional information that we have replaced by individual *p*-values. Such replacement may lead to certain information loss which requires further investigation. The estimation of the distributional information is challenging and computationally intensive. Substantial work is needed to extend existing methods to the spatial setting; such analysis is beyond the scope of the current paper. The development of more powerful weighting strategies to incorporate other types of side information is an interesting and important direction for future research.

References

Barber, R. F. and Ramdas, A. (2017). The p-filter: multilayer false discovery rate control for grouped hypotheses. *J. Roy. Statist. Soc. B*, 79(4):1247–1268.

Basu, P., Cai, T. T., Das, K., and Sun, W. (2018). Weighted false discovery rate control in large-scale multiple testing. *J. Am. Statist. Assoc.*, 113(523):1172–1183.

Benjamini, Y. and Bogomolov, M. (2014). Selective inference on multiple families of hypotheses. *J. Roy. Statist. Soc. B*, 76(1):297–318.

Benjamini, Y. and Heller, R. (2007). False discovery rates for spatial signals. *J. Amer. Statist. Assoc.*, 102(480):1272–1281.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B*, 57(1):289–300.

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scand. J. Stat.*, 24(3):407–418.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.*, 25(1):60–83.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188.

Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.*, 104(488):1467–1481.

Cai, T. T., Sun, W., and Wang, W. (2019). CARS: Covariate assisted ranking and screening for large-scale two-sample inference (with discussion). *J. Roy. Statist. Soc. B*, 81(2):187–234.

Cao, H. and Wu, W.-B. (2015). Changepoint estimation: another look at multiple testing problems. *Biometrika*, 102(4):974–980.

Chernozhukov, V., Chetverikov, D., Kato, K., et al. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.*, 45(4):2309–2352.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.*, 102(477):93–103.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23:1–22.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96(456):1151–1160.

Fan, J., Hall, P., and Yao, Q. (2007). To how many simultaneous hypothesis tests can normal, student's t or bootstrap calibration be applied? *J. Amer. Statist. Assoc.*, 102(480):1282–1288.

Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G., and Kong, A. (2008). Unsupervised empirical bayesian multiple testing with external covariates. *Ann. Appl. Stat.*, 2(2):714–735.

Fortney, K., Dobriban, E., Garagnani, P., Pirazzini, C., Monti, D., Mari, D., Atzmon, G., Barzilai, N., Franceschi, C., Owen, A. B., et al. (2015). Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLoS genetics*, 11(12):e1005728.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. B*, 64(3):499–517.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3):509–524.

Habiger, J. (2017). Adaptive false discovery rate control for heterogeneous data. *Stat. Sin.*, pages 1731–1756.

Habiger, J., Watts, D., and Anderson, M. (2017). Multiple testing with heterogeneous multinomial distributions. *Biometrics*, 73(2):562–570.

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., and Benjamini, Y. (2006). Cluster-based analysis of fmri data. *NeuroImage*, 33(2):599–608.

Hu, J. X., Zhao, H., and Zhou, H. H. (2010). False discovery rate control with groups. *J. Amer. Statist. Assoc.*, 105(491):1215–1227.

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods*, 13(7):577.

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *J. Roy. Statist. Soc. B*, 80(4):649–679.

Lei, L., Ramdas, A., and Fithian, W. (2017). Star: A general interactive framework for fdr control under structural constraints. *arXiv preprint arXiv:1710.02776*.

Li, A. and Barber, R. F. (2019). Multiple testing with the structure-adaptive benjamini–hochberg algorithm. *J. Roy. Statist. Soc. B*, 81(1):45–74.

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *J. Amer. Statist. Assoc.*, 112(519):1131–1146.

Liu, J., Zhang, C., Page, D., et al. (2016a). Multiple testing under dependence via graphical models. *Ann. Appl. Stat.*, 10(3):1699–1724.

Liu, W. (2014). Incorporation of sparsity information in large-scale multiple two-sample *t* tests. *arXiv preprint arXiv:1410.4282*.

Liu, W. and Shao, Q.-M. (2010). Cramer-type moderate deviation for the maximum of the periodogram with application to simultaneous tests in gene expression time series. *Ann. Statist.*, 38(3):1913–1935.

Liu, Y., Sarkar, S. K., and Zhao, Z. (2016b). A new approach to multiple testing of grouped hypotheses. *J. Stat. Plan. Inference*, 179:1–14.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176.

Pacifico, M., Genovese, C., Verdinelli, I., and Wasserman, L. (2004). False discovery control for random fields. *J. Amer. Statist. Assoc.*, 99(468):1002–1014.

Peña, E. A., Habiger, J. D., and Wu, W. (2011). Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *Ann. Statist.*, 39(1):556.

Roquain, E. and Van De Wiel, M. A. (2009). Optimal weighting for false discovery rate control. *Electron. J. Stat.*, 3:678–711.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.*, 30:239–257.

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P., and Kass, R. E. (2015). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *J. Amer. Statist. Assoc.*, 110(510):459–471.

Shu, H., Nan, B., and Koeppe, R. (2015). Multiple testing for neuroimaging via hidden markov random field. *Biometrics*, 71(3):741–750.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. B*, 64(3):479–498.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.*, 102(479):901–912.

Sun, W. and Cai, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. B*, 71(2):393–424.

Sun, W., Reich, B. J., Cai, T. T., Guindani, M., and Schwartzman, A. (2015). False discovery control in large-scale spatial multiple testing. *J. Roy. Statist. Soc. B*, 77(1):59–83.

Tansey, W., Koyejo, O., Poldrack, R. A., and Scott, J. G. (2018). False discovery rate smoothing. *Journal of the American Statistical Association*, 113(523):1156–1171.

Wu, W. B. (2008). On false discovery control under dependence. *Ann. Statist.*, 36(1):364–380.

Xia, Y., Cai, T. T., and Sun, W. (2019). GAP: A General Framework for Information Pooling in Two-Sample Sparse Inference. *J. Amer. Statist. Assoc.*, to appear.

Yekutieli, D. (2008). Hierarchical false discovery rate–controlling methodology. *J. Amer. Statist. Assoc.*, 103(481):309–316.

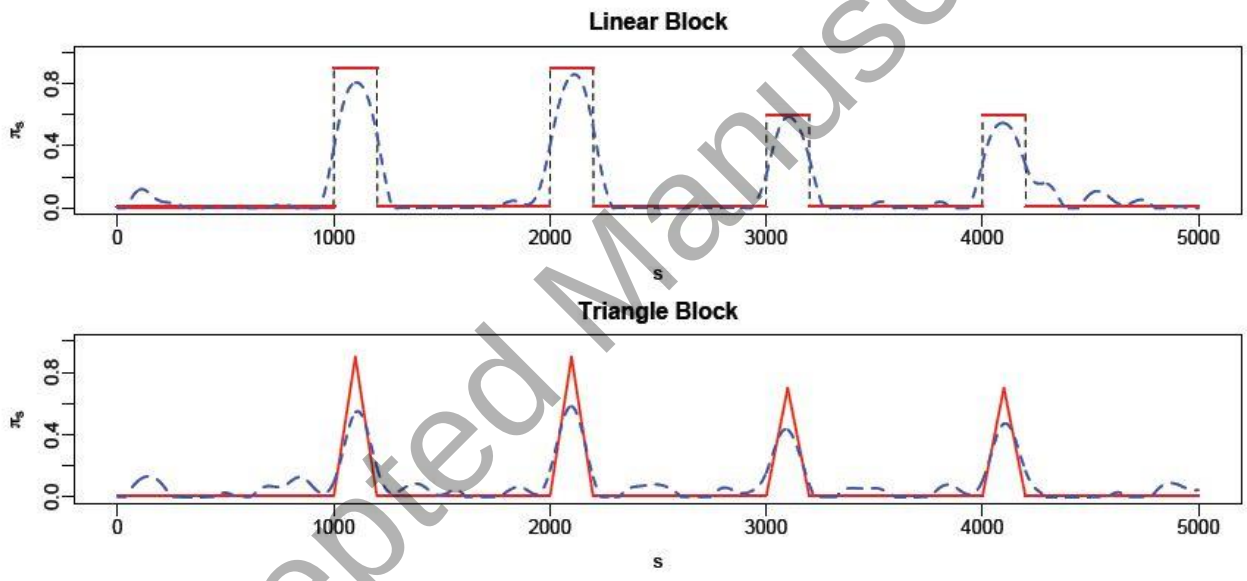

Fig. 1 True π_s (solid lines) vs estimated $\hat{\pi}_s$ (dashed lines). Top row: piecewise constants; bottom row: triangle blocks.

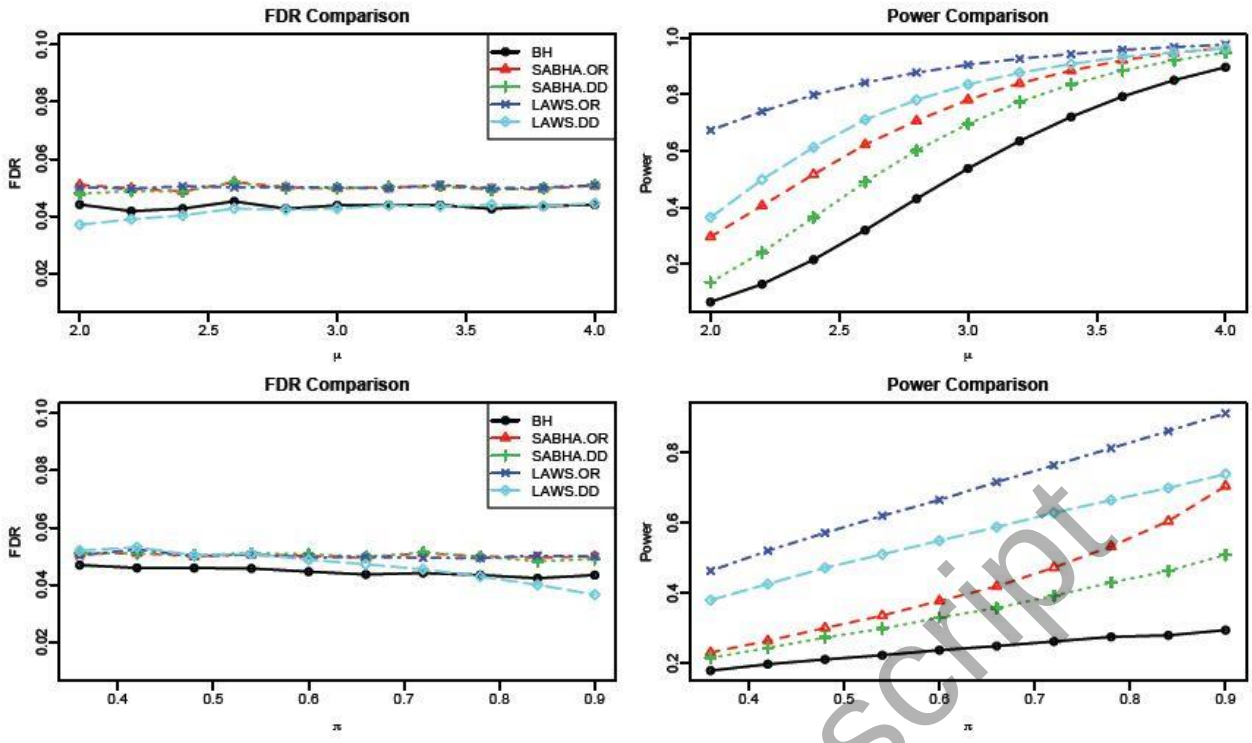

Fig. 2 FDR and Power comparisons: the linear block pattern.

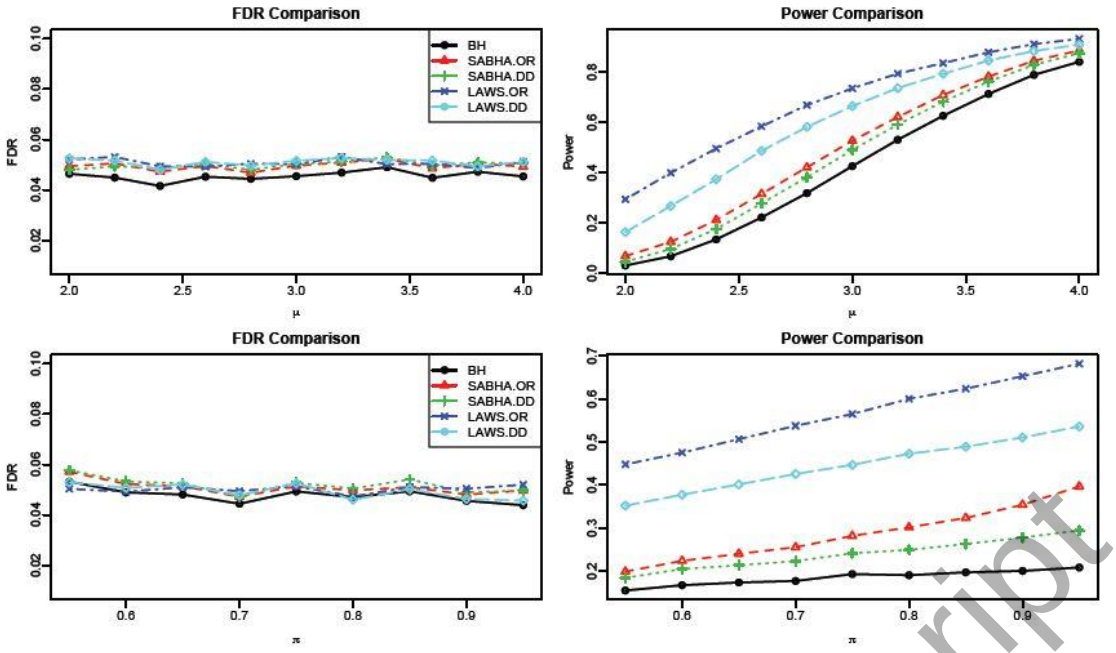

Fig. 3 FDR and Power comparisons: the triangular block pattern.

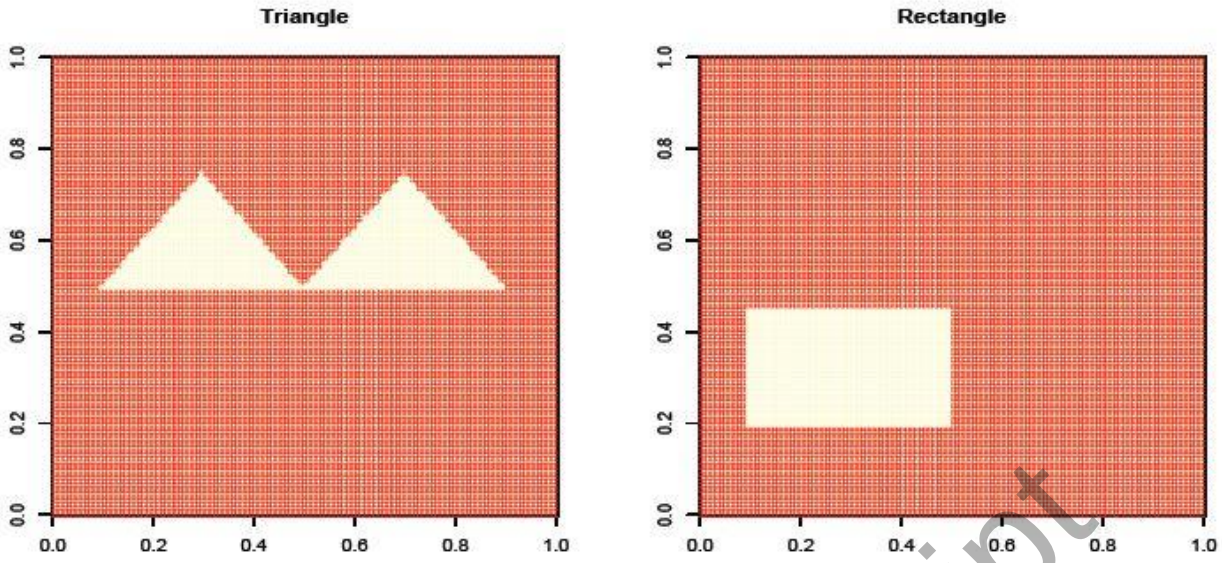

Fig. 4 2D triangle and rectangle pattern.

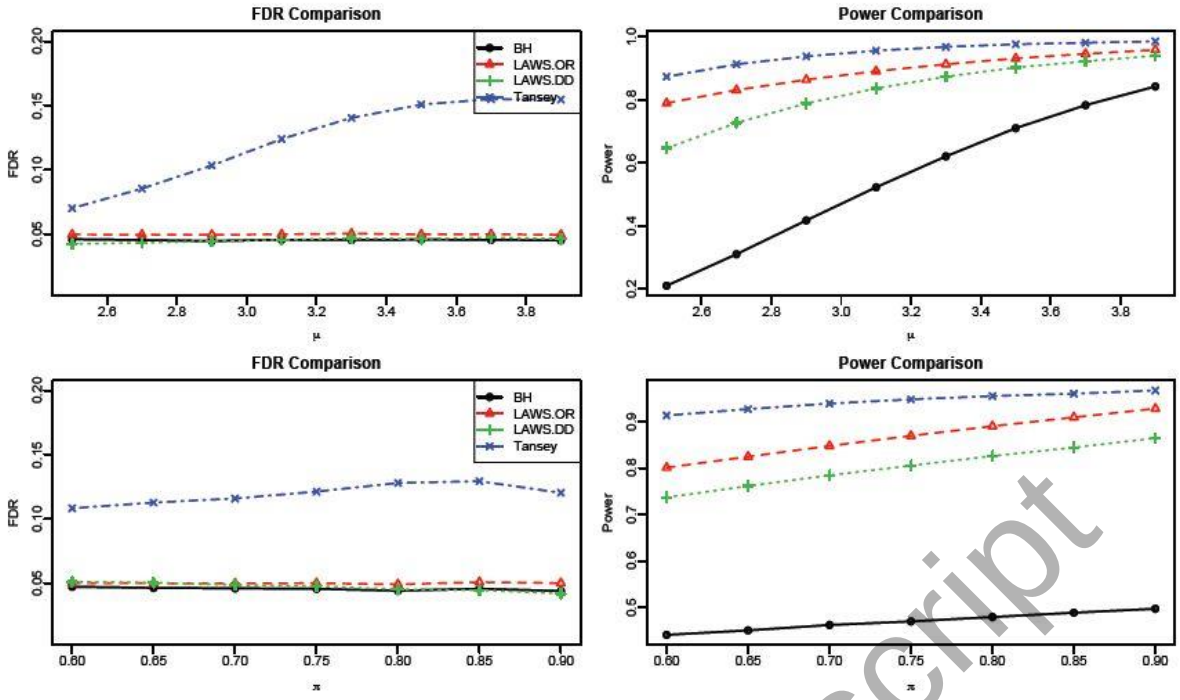

Fig. 5 FDR and Power comparisons: the 2D triangle pattern.

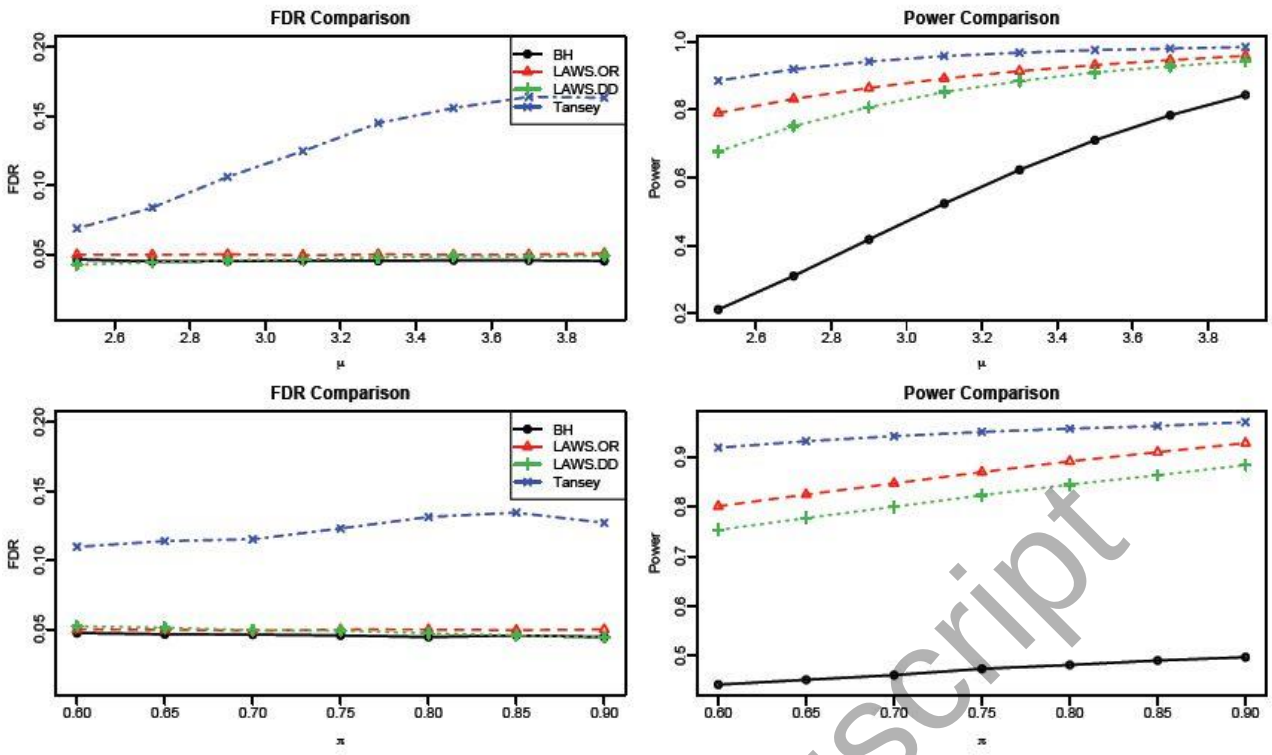

Fig. 6 FDR and Power comparisons: the 2D rectangle pattern.

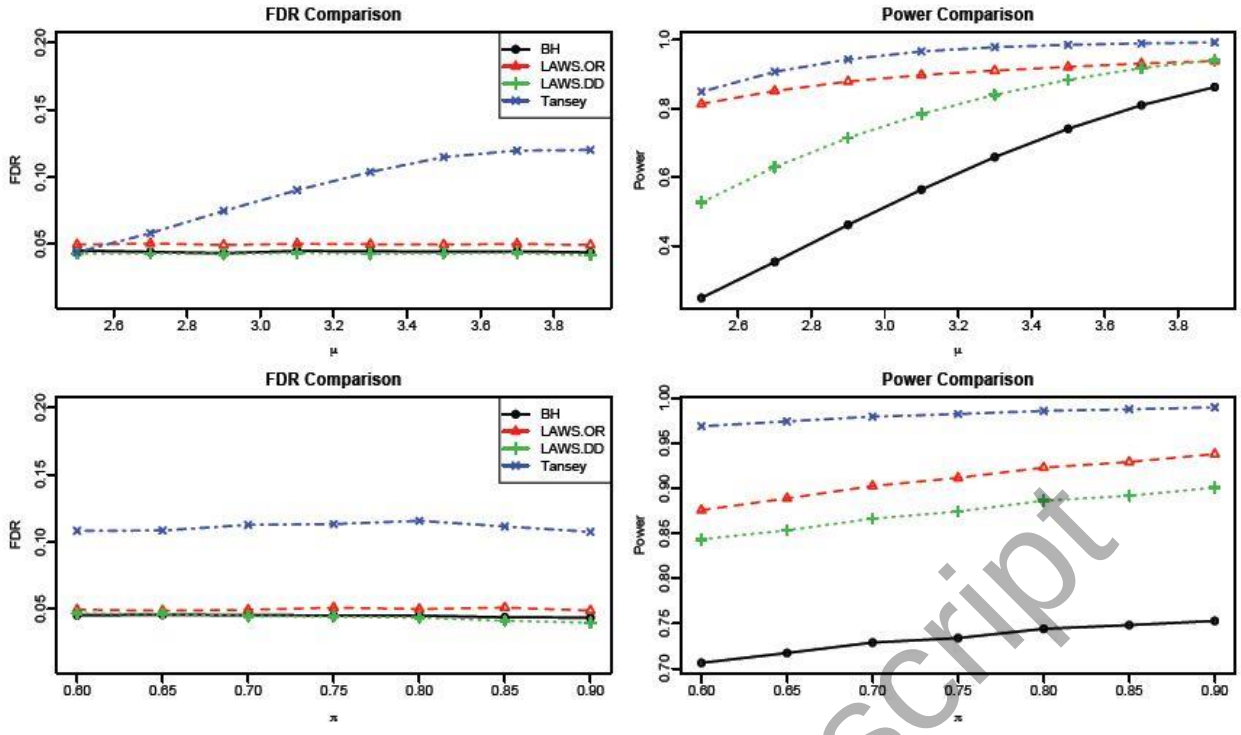

Fig. 7 FDR and Power comparisons: the 3D cubic pattern.

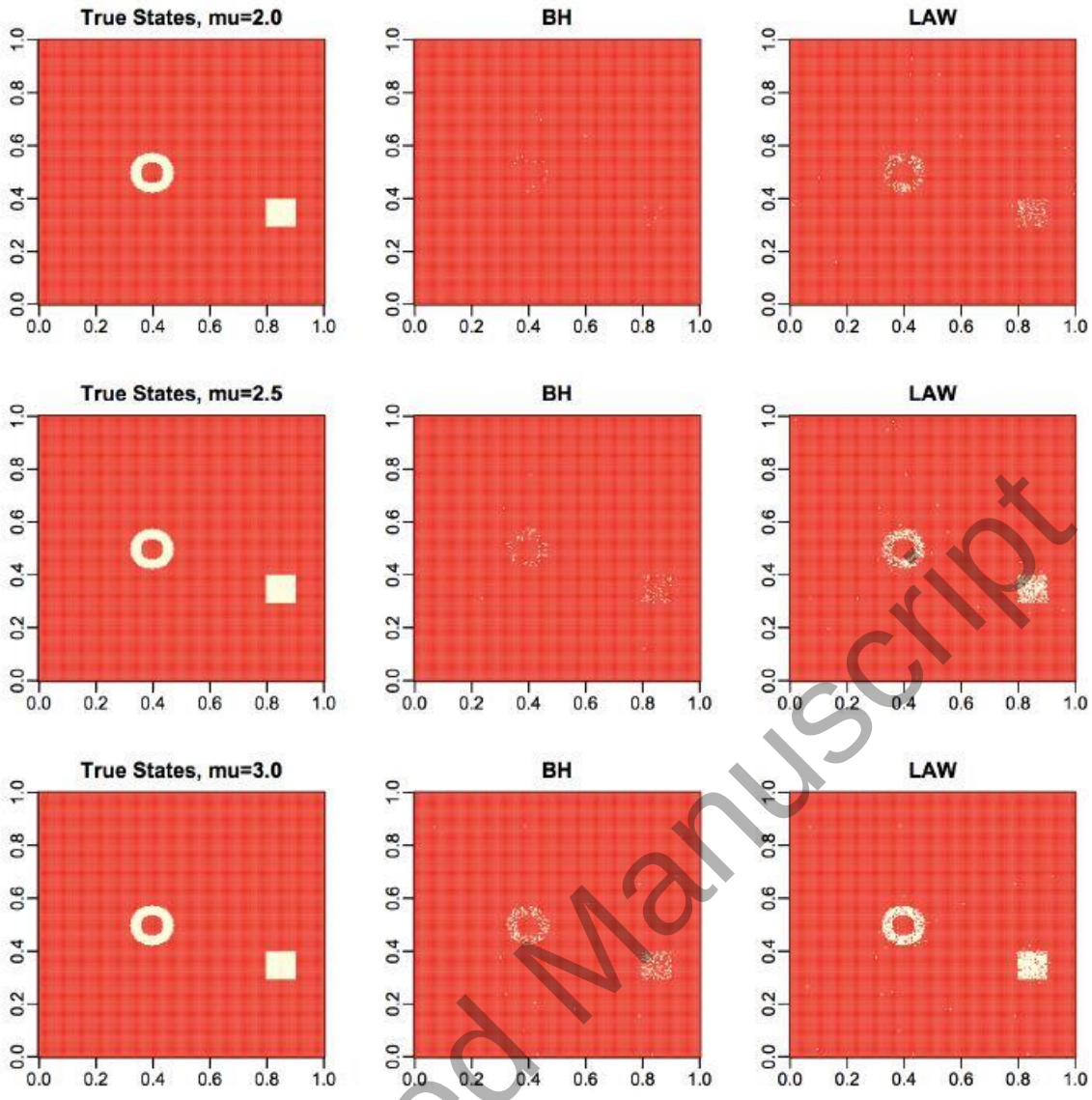

Fig. 8 Spatial FDR Analysis in 2D setting. LAWS is more effective in revealing the donut and square shapes by up-weighting the *p*-values in the regions where signal appear in clusters.

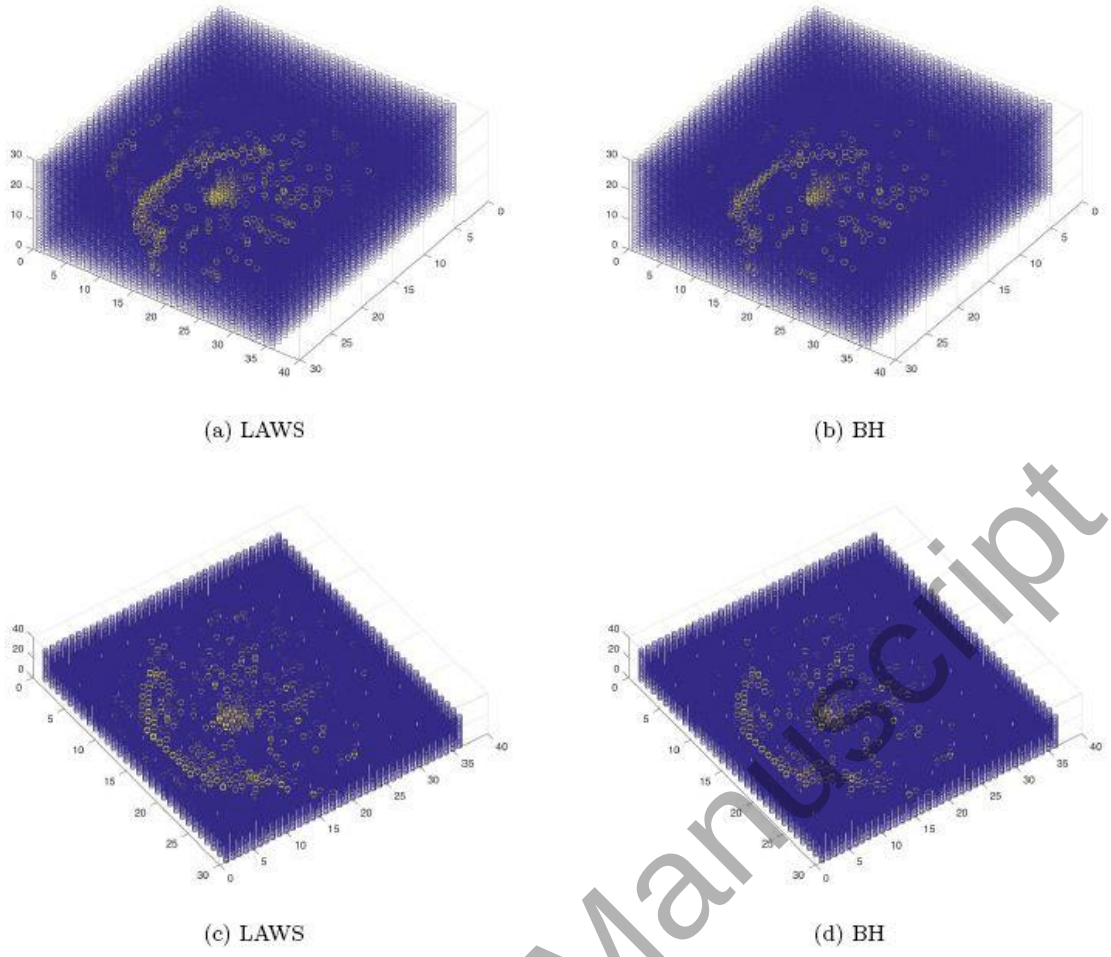

Fig. 9 Significant brain regions (yellow) after applying LAWS (left) and BH (right), view with azimuth and elevation angles $(-35, -65)$ on top row and $(35, -80)$ on bottom row. FDR level $\alpha = 0.05$.

Notes

[1] In other applications such as climate change analysis, one observes incomplete data points at irregular locations (e.g. weather monitoring stations) but needs to make inference at every point in the whole spatial domain. This setting goes beyond the scope of our work; see Sun et al. (2015) for related discussions.

[2] The actual order would not affect the methodology or theory as the weights are fully determined by the spatial structure. We only need an ordering for characterizing the dependence structure between all pairs of *p*-values.