Statistica Sinica Preprint No: SS-2024-0152	
Title	Powerful Spatial Multiple Testing via Borrowing Neighboring Information
Manuscript ID	SS-2024-0152
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0152
Complete List of Authors	Linsui Deng, Kejun He and Xianyang Zhang
Corresponding Authors	Kejun He
E-mails	kejunhe@ruc.edu.cn
Notice: Accepted author version.	

Powerful Spatial Multiple Testing via Borrowing Neighboring Information

Linsui Deng, Kejun He, and Xianyang Zhang

Renmin University of China and Texas A&M University

Abstract: Clustered effects are often encountered in multiple hypothesis testing of spatial signals. In this paper, we propose a new method, termed *two-dimensional spatial multiple testing* (2d-SMT) procedure, to control the false discovery rate (FDR) and improve the detection power by exploiting the spatial information encoded in neighboring observations. The proposed method provides a novel perspective of utilizing spatial information by gathering signal patterns and spatial dependence into an auxiliary statistic. 2d-SMT rejects the null when a primary statistic at the location of interest and the auxiliary statistic constructed based on nearby observations are greater than their corresponding cutoffs. 2d-SMT can also be combined with different variants of the weighted BH procedures to improve the detection power further. A fast algorithm is developed to accelerate the search for optimal cutoffs in 2d-SMT. In theory, we establish the asymptotic FDR control of 2d-SMT under weak spatial dependence. Extensive numerical experiments demonstrate that the 2d-SMT method combined with various weighted BH procedures achieves the most competitive performance in FDR and power trade-off.

Key words and phrases: Empirical Bayes; False discovery rate; Near epoch dependence; Side information.

1. Introduction

Large-scale multiple testing with spatial structure has become increasingly important in various areas, e.g., Functional Magnetic Resonance Imaging research, genome-wide association studies,

Kejun He (Email: kejunhe@ruc.edu.cn) and Xianyang Zhang (Email: zhangxiany@stat.tamu.edu) are the corresponding authors.

environmental studies, and astronomical surveys. The essential task is identifying locations that exhibit significant deviations from the background to build scientific interpretations. Since thousands or even millions of spatially correlated hypotheses tests are often conducted simultaneously, incorporating informative spatial patterns to provide a powerful multiplicity adjustment for dependent multiple testing is becoming a significant challenge.

There has been a growing literature on spatial signal detection with false discovery rate control (FDR, Benjamini and Hochberg, 1995). Heller et al. (2006) and Sun et al. (2015) proposed to perform multiple testing on cluster-wise hypotheses by aggregating location-wise hypotheses to increase the signal-to-noise ratio. Benjamini and Heller (2007), Sun et al. (2015) and Basu et al. (2018) defined new error rates to reflect the relative importance of hypotheses associated with different clusters, e.g, a hypothesis related to a larger cluster is more important than the one associated with a smaller cluster. Scott et al. (2015) considered a two-group mixture model with the prior null probability dependent on the auxiliary spatial information. Yun et al. (2022) proposed a spatial-adaptive FDR-controlling procedure by exploiting the mirror conservatism of the null p-values and the spatial smoothness under the alternative. Tansey et al. (2018) enforced spatial smoothness by imposing a penalty on the pairwise differences of log odds of hypotheses being signals between adjacent locations. Along a related line, Genovese et al. (2006) suggested to weight p-values or equivalently assign location-specific cutoffs by leveraging the tendency of hypotheses being null. This idea has been further developed in some recent papers to include different types of structural and covariate information. See, e.g., Ignatiadis et al. (2016), Li and Barber (2019), Cai et al. (2022), Zhang and Chen (2022), and Cao et al. (2022).

In many applications, signals tend to exhibit in clusters. As a result, hypotheses around a non-null location are more likely to be under the alternative than under the null. One way to account for spatially clustered signals is to screen out the locations where the average signal strength of the neighbors captured by an auxiliary statistic is weak (Shen et al., 2002). The locations passing the screening step are subjected to further analysis. This procedure suffers from the so-called selection bias as the downstream statistical inference needs to account for the selection effect from the screening step. A simple remedy is sample splitting (Wasserman and Roeder, 2009; Liu et al., 2022), where a subsample is used to perform screening, and the remaining samples are utilized for the downstream inference. Sample splitting is intuitive and easy to implement, but it inevitably sacrifices the detection power because it does not excavate complete information. Furthermore, there are often no completely independent observations to conduct sample splitting in spatial settings.

We propose a new method named the two-dimensional spatial multiple testing (2d-SMT) procedure that fundamentally differs from the existing spatial multiple testing procedures. 2d-SMT consists of a two-dimensional rejection region built on two statistics, an auxiliary statistic $T_1(s)$ and a primary statistic $T_2(s)$, for each hypothesis. The first dimension utilizes the auxiliary information constructed from the neighbors of a location of interest to perform feature screening, which helps to increase the signal density and lessen the multiple testing burdens in the second dimension. The second dimension then uses a statistic computed from the data at the location of interest to pick out signals. 2d-SMT declares the hypothesis at location s to be non-null if $T_1(s) \geq t_1$ and $T_2(s) \geq t_2$. The optimal cutoffs $t_1^\star$ and $t_2^\star$ are chosen to achieve the maximum number of discoveries while controlling the FDR at the desired level. 2d-SMT involves

three main ingredients, designed to improve its robustness and efficiency: (1) accounting for the dependence between the auxiliary statistic and the primary statistic, which alleviates the selection bias; (2) borrowing spatial signal information through an empirical Bayes approach; and (3) accelerating the search for the bivariate cutoff through an efficient algorithm. In a related study, Yi et al. (2021) proposed the 2dFDR approach to detect the association between omics features and covariates of interest in the presence of confounding factors, borrowing information from confounder-unadjusted test statistics to boost the power in testing with confounder-adjustment. In contrast, 2d-SMT is designed for the spatial multiple testing by borrowing information from neighboring observations.

The contribution of this work lies in its innovative methodology, theoretical analysis, and a new searching algorithm. First, 2d-SMT explores spatial information from a completely different perspective compared to the existing weighted procedures. It thus can be combined with these methods to improve power further. Examples include the group BH procedure (GBH, Hu et al., 2010), independent hypothesis weighting (IHW, Ignatiadis et al., 2016), structure adaptive BH algorithm (SABHA, Li and Barber, 2019), and locally adaptive weighting and screening approach (LAWS, Cai et al., 2022). The readers are referred to Section 2.7 for more details. Second, our asymptotic analysis allows weak spatial dependence, which goes beyond the independence assumption required by the existing empirical Bayes theory, thereby broadening its application scope. To the best of our knowledge, this is the first analytical framework where dependent observations are allowed within the context of empirical Bayes theory. Third, we develop an algorithm to overcome the computational bottleneck in finding the 2d cutoff values without sacrificing accuracy, which can be applied to procedures using two-dimensional rejection

regions, including 2dFDR and 2d-SMT.

The rest of the paper is organized as follows. Section 2 develops the 2d-SMT procedure, including the oracle procedure, the feasible procedure with estimated covariance structure, and the extension by combining it with various weighted BH procedures. Section 3 discusses some implementation details. Section 4 establishes the asymptotic FDR control of the 2d-SMT procedure. In Sections 5 and 6, extensive simulation studies and an analysis of ozone data demonstrate the effectiveness of the 2d-SMT procedure. Section 7 concludes and points out a few future research directions. Some additional details of the numerical experiments and technical proofs are presented in an online supplementary file.

2. Method

Consider a random field $\{X(s) : s \in \mathcal{S}\}$ defined on a spatial domain $\mathcal{S} \subseteq \mathbb{R}^K$ with $K \geq 1$ that takes the form of $X(s) = \mu(s) + \epsilon(s)$, where $\mu(s)$ is an unobserved process of interest and $\epsilon(s)$ is a mean-zero Gaussian process. The model is prevalent in spatial multiple testing across various domains, such as fMRI (Heller et al., 2006), environment study (Sun et al., 2015), temperature data analysis (Huang et al., 2021). We are interested in examining whether $\mu(s)$ belongs to an indifference region $\mathcal{A}$. For example, $\mathcal{A} = \{\mu \in \mathbb{R} : \mu \leq \mu_0\}$ for a one-sided test and $\mathcal{A} = \{\mu \in \mathbb{R} : |\mu| \leq \mu_0\}$ for a two-sided test, where μ_0 is some pre-specified value. The unobserved process $\mu(s)$ and the indifference region $\mathcal{A}$ induce a background statement $\theta(s) = \mathbf{1}\{\mu(s) \notin \mathcal{A}\}$ on the spatial domain $\mathcal{S}$. We define $\mathcal{S}_0 = \{s \in \mathcal{S} : \theta(s) = 0\}$ and $\mathcal{S}_1 = \{s \in \mathcal{S} : \theta(s) = 1\}$ as the sets of null and non-null locations respectively.

We focus on the point-wise analysis, testing the hypothesis $\mathcal{H}_{0,s} : \mu(s) \in \mathcal{A}$ versus $\mathcal{H}_{a,s} : \mu(s) \notin \mathcal{A}$. At each location $s \in \mathcal{S}$, we make a decision $\delta(s)$, where $\delta(s) = 1$ if $\mathcal{H}_{0,s}$ is rejected

and $\delta(s) = 0$ otherwise. Let $\Delta_1 = \{s \in \mathcal{S} : \delta(s) = 1\}$ be the set of rejections associated with the decision rule δ. The false discovery rate (FDR) is defined as

$$\text{FDR} = E\left(\frac{|\Delta_1 \cap \mathcal{S}_0|}{1 \vee |\Delta_1|}\right),$$

where $|\cdot|$ denotes the cardinality of a set and $\Delta_1 \cap \mathcal{S}_0$ is the set of false discoveries. We next present a spatial multiple testing procedure that borrows neighboring information to improve the signal detection power without sacrificing the FDR control.

2.1 Motivation

For clarity, we focus our discussions on the one-sided test for the rest of the paper, i.e.,

$$\mathcal{H}_{0,s} : \mu(s) \leq 0 \quad \text{versus} \quad \mathcal{H}_{a,s} : \mu(s) > 0.$$

For each location $s \in \mathcal{S}$, we define a set of its neighbors as $\mathcal{N}(s) \subseteq \mathcal{S}\backslash\{s\}$. Because of the spatial dependence and smoothness encountered in many real applications, the set of neighboring observations $\{X(v) : v \in \mathcal{N}(s)\}$ is expected to provide useful side information on determining the state of $\theta(s)$. To formalize this idea, we consider two statistics, namely the auxiliary statistic

$$T_1(s) = \frac{1}{\tau(s)} \sum_{v \in \mathcal{N}(s)} X(v) \tag{2.1}$$

based on the averaged observed values in the neighborhood of s and the primary statistic $T_2(s) = \sigma^{-1}(s)X(s)$ based on the observation from the location of interest, where $\sigma^2(s) = \text{Var}\{\epsilon(s)\}$ and $\tau^2(s) = \sum_{v,v' \in \mathcal{N}(s)} \text{cov}\{\epsilon(v), \epsilon(v')\}$. For $s \in \mathcal{S}$, we have $T_1(s) = \xi(s) + V_1(s)$ and

$T_2(s) = \sigma^{-1}(s)\mu(s) + V_2(s)$ where $\xi(s) = \tau^{-1}(s)\sum_{v\in\mathcal{N}(s)}\mu(v)$ and

$$\begin{pmatrix} V_1(s) \\ V_2(s) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho(s) \\ \rho(s) & 1 \end{pmatrix}\right)$$

with $\rho(s) = \{\sigma(s)\tau(s)\}^{-1}\sum_{v\in\mathcal{N}(s)}\mathrm{cov}\{\epsilon(s),\epsilon(v)\}$. Our method is motivated by the following two-stage procedure. At stage 1, we use the auxiliary statistic $T_1(s)$ to screen out the nulls based on the belief that observations around a non-null location tend to take larger values on average. At stage 2, we use $T_2(s)$ to pick out signals among those who survive from stage 1. Given the cutoffs (t_1, t_2) of these two stages, the procedure can be described as follows:

Stage 1. Use the auxiliary statistic $T_1(s)$ to determine a preliminary set of signals $\mathcal{D}_1 = \{s \in \mathcal{S} : T_1(s) \geq t_1\}$.

Stage 2. Reject $\mathcal{H}_{0,s}$ for $T_2(s) \geq t_2$ and $s \in \mathcal{D}_1$. As a result, the final set of discoveries is given by $\mathcal{D}_2 = \{s \in \mathcal{S} : T_1(s) \geq t_1, T_2(s) \geq t_2\}$.

Since the screening step reduces the multiple testing burden for the second stage, we expect the above method to be more powerful than the traditional method based only on the primary statistic. Indeed, the 1d rejection region is a special case of the 2d rejection region $\{s \in \mathcal{S} : T_1(s) \geq t_1, T_2(s) \geq t_2\}$ by setting $t_1 \leq \min_{s\in\mathcal{S}} T_1(s)$, i.e., the first stage preserves all locations. If we select the two cutoffs one by one, then the choice of t_2 should consider the selection effect from the first stage. Here we propose a new method to address this issue by simultaneously selecting the two cutoffs, which we name the *2-dimensional* (2d) procedure.

2.2 Approximation of the false discovery proportion

Note that $\mathcal{H}_{0,s}$ is rejected when $T_1(s) \geq t_1$ and $T_2(s) \geq t_2$. Recalling that $\mathcal{S}_0$ is the set of true nulls, the false discovery proportion (FDP) is then given by

$$\mathrm{FDP}(t_1, t_2) = \frac{\sum_{s\in\mathcal{S}_0} \mathbf{1}\left\{T_1(s) \geq t_1, T_2(s) \geq t_2\right\}}{1 \vee R(t_1, t_2)} \leq \frac{\sum_{s\in\mathcal{S}_0} \mathbf{1}\left\{V_1(s) + \xi(s) \geq t_1, V_2(s) \geq t_2\right\}}{1 \vee R(t_1, t_2)},$$

where $R(t_1, t_2) = \sum_{s\in\mathcal{S}} \mathbf{1}\left\{T_1(s) \geq t_1, T_2(s) \geq t_2\right\}$ corresponds to the total number of rejections, and the inequality holds because $\mu(s) \leq 0$ for $s \in \mathcal{S}_0$. Motivated by the law of large numbers, we follow the approaches in multiple testing (Benjamini and Hochberg, 1995; Storey, 2002), further developed for detecting spatial signals (Benjamini and Heller, 2007; Sun et al., 2015; Cai et al., 2022), to substitute the numerator of the right-hand side (RHS) of the above inequality by its expected value. This substitution leads to an asymptotic upper bound of FDP as

$$\begin{aligned} \mathrm{FDP}(t_1, t_2) &\lesssim \frac{\sum_{s\in\mathcal{S}_0} P\left\{V_1(s) + \xi(s) \geq t_1, V_2(s) \geq t_2\right\}}{1 \vee R(t_1, t_2)} \\ &:= \frac{\sum_{s\in\mathcal{S}_0} L\left\{t_1, t_2, \xi(s), \rho(s)\right\}}{1 \vee R(t_1, t_2)}, \end{aligned} \tag{2.2}$$

where $L\left\{t_1, t_2, \xi(s), \rho(s)\right\} = P\left\{V_1(s) + \xi(s) \geq t_1, V_2(s) \geq t_2\right\}$. The major challenge here is the estimation of the expected number of false rejections given by $\sum_{s\in\mathcal{S}_0} L\left\{t_1, t_2, \xi(s), \rho(s)\right\}$, which involves a large number of nuisance parameters $\xi(s)$. To overcome this difficulty, we adopt an empirical Bayes viewpoint to borrow spatial information across different locations and directly estimate the expected number of false rejections without estimating individual $\xi(s)$ at each location explicitly.

2.3 Nonparametric empirical Bayes

Let $G_{\mathcal{S}_0}$ be the empirical distribution of $\{\xi(s) : s \in \mathcal{S}_0\}$. The expected number of false rejections in (2.2) can be approximated by $\sum_{s\in\mathcal{S}_0} \int L\{t_1, t_2, x, \rho(s)\}\, dG_{\mathcal{S}_0}(x)$. However, directly estimating $G_{\mathcal{S}_0}$ is challenging as the auxiliary statistics $\{T_1(s) : s \in \mathcal{S}\}$ blend information from both the null and alternative hypotheses. To overcome this difficulty, we observe that $\xi(s)$ typically takes greater values under the alternative than under the null and $L\{t_1, t_2, x, \rho(s)\}$ is a monotonically increasing function of x. Thus, we have

$$\begin{aligned}\sum_{s\in\mathcal{S}_0} \int L\{t_1, t_2, x, \rho(s)\}\, dG_{\mathcal{S}_0}(x) &\lesssim \sum_{s\in\mathcal{S}_0} \int L\{t_1, t_2, x, \rho(s)\}\, dG_{\mathcal{S}}(x) \\ &\le \sum_{s\in\mathcal{S}} \int L\{t_1, t_2, x, \rho(s)\}\, dG_{\mathcal{S}}(x),\end{aligned} \tag{2.3}$$

where $G_{\mathcal{S}}$ is the empirical distribution $\{\xi(s) : s \in \mathcal{S}\}$. Consequently, we aim to estimate $G_{\mathcal{S}}$ based on the whole set of auxiliary statistics $\{T_1(s) : s \in \mathcal{S}\}$ through the nonparametric empirical Bayes (NPEB) approach.

The estimation in NPEB can be achieved by maximizing the marginal distribution of $T_1(s) = \xi(s)+V_1(s)$, which is given by $f_{G_{\mathcal{S}}}(x) = \int \phi(x-u)dG_{\mathcal{S}}(u)$ with ϕ denoting the density function of the standard normal distribution. Classical empirical Bayes methods often assume independence among the observations, which is violated in our case due to spatial dependence. To reduce the dependence, we select a subset $\widetilde{\mathcal{S}}$ of $\mathcal{S}$ such that any two points in $\widetilde{\mathcal{S}}$ have a distance larger than some cutoff c_0 (so that the dependence between $T_1(v)$ and $T_1(v')$ for any $v, v' \in \widetilde{\mathcal{S}}$ is sufficiently weak). Following Kiefer and Wolfowitz (1956) and Zhang (2009), we consider the

general maximum likelihood estimator (GMLE) defined as

$$\widetilde{G}_{\widetilde{\mathcal{S}}} = \arg\max_{G\in\mathcal{G}} \sum_{s\in\widetilde{\mathcal{S}}} \log f_G\{T_1(s)\}, \tag{2.4}$$

where $\mathcal{G}$ represents the set of all probability distributions on $\mathbb{R}$ and $f_G(x) = \int \phi(x-u)dG(u)$ is the convolution between G and ϕ. Our theoretical analysis in Lemma S.8 of the supplement shows that the estimated GMLE is close to the limit of the empirical distribution of $\{\xi(s) : s \in \mathcal{S}\}$, denoted by G_0; see Assumption 7 for a formal definition of G_0. The optimization in (2.4) can be cast as a convex optimization problem that can be efficiently solved by modern interior point methods (Koenker and Mizera, 2014).

2.4 2d spatial multiple testing procedure

We now describe a procedure to select the two cutoffs simultaneously. In view of (2.2)–(2.4), we consider an approximated upper bound for $\mathrm{FDP}(t_1, t_2)$ given by

$$\widetilde{\mathrm{FDP}}(t_1, t_2) := \frac{\sum_{s\in\mathcal{S}} \int L\left\{t_1, t_2, x, \rho(s)\right\} d\widetilde{G}_{\widetilde{\mathcal{S}}}(x)}{1 \vee R(t_1, t_2)}. \tag{2.5}$$

As shown in Lemma S.1 of the supplement, $\sum_{s\in\mathcal{S}} \int L\left\{t_1, t_2, x, \rho(s)\right\} dG_0(x)$ can be consistently estimated by the numerator of (2.5). For a desired FDR level $q \in (0, 1)$, the 2d-SMT procedure chooses the optimal cutoff such that $(\widehat{t}_1^{\star}, \widehat{t}_2^{\star}) = \arg\max_{(t_1,t_2)\in\mathcal{F}_q} R(t_1, t_2)$, where $\mathcal{F}_q = \{(t_1, t_2) \in \mathbb{R}^2 : \widetilde{\mathrm{FDP}}(t_1, t_2) \le q\}$.

We argue that the 2d-SMT procedure is generally more powerful than the classical BH procedure based on the primary statistics alone. When setting $t_1 = -\infty$, 2d-SMT is equivalent to BH, as it fails to exclude any hypotheses at the first stage so that signal detection relies solely

on the second stage, and our FDP estimator in (2.5) is equivalent to the FDP estimator in the BH procedure, i.e.,

$$\frac{\sum_{s\in\mathcal{S}}\int P\left\{V_1(s)+x\geq-\infty, V_2(s)\geq t_2\right\}d\widetilde{G}_{\widetilde{\mathcal{S}}}(x)}{1\vee R(-\infty,t_2)}=\frac{\sum_{s\in\mathcal{S}}P\left\{V_2(s)\geq t_2\right\}}{1\vee\sum_{s\in\mathcal{S}}\mathbf{1}\left\{T_2(s)\geq t_2\right\}}.$$

The 2d-SMT procedure has the flexibility to choose an additional cutoff t_1 to maximize the number of rejections and guarantees to make more rejections. Section 4 will show 2d-SMT has asymptotic FDR control; thus more rejections typically translate into a higher power.

Similar to the FDP estimator of the BH procedure, the conservatism of (2.5) arises partly from expanding the index set $\mathcal{S}_0$ of the summation in (2.2) to $\mathcal{S}$ in (2.3). For a target FDR level q, the realized FDR level of the BH procedure is approximately $\pi_0 q$, where $\pi_0=|\mathcal{S}_0|/|\mathcal{S}|$ is the null proportion. The conservatism motivates us to estimate the null proportion to improve power; see Section 2.5. Theorem 2 rigorously proves that our procedure is more powerful than the one using primary statistics alone.

Remark 1. Shen et al. (2002) and Huang et al. (2021) considered the case where μ has a sparse wavelet representation. One of the goal is to detect the significant wavelet coefficients while controlling the FDR. Observing that the wavelet coefficients within each scale and across different scales are related, they screened the wavelet coefficients based on the largest adjacent wavelet coefficients to gain more power. The generalized degrees of freedom determine the number of locations for the subsequent FDR-controlling procedure. Our approach can be potentially applied to their settings by exploring the signal structure encoded in the neighboring wavelet coefficients.

2.5 Estimating the null proportion

It is well known that when the number of signals is a substantial proportion of the total number of hypotheses, the BH procedure will be overly conservative. We develop a modification of Storey's approach (Storey, 2002; Storey et al., 2004) to incorporate the estimation of the null proportion. As a motivation, we assume that $T_2(s)$ approximately follows the mixture model $T_2(s) \sim \pi_0 \mathcal{N}(\mu_0(s), 1) + (1-\pi_0)\mathcal{N}(\mu_1(s), 1)$, where $\mu_0(s) \leq 0$, $\mu_1(s) > 0$, and π_0 is the prior probability that $s \in \mathcal{S}_0$. Let Φ be the cumulative distribution function of $\mathcal{N}(0,1)$. Fixing some $\lambda \in \mathbb{R}$, we have

$$\begin{aligned} P\{T_2(s) < \lambda\} &= \pi_0 P\{\mathcal{N}(\mu_0(s), 1) < \lambda\} + (1-\pi_0) P\{\mathcal{N}(\mu_1(s), 1) < \lambda\} \\ &\geq \pi_0 P\{\mathcal{N}(\mu_0(s), 1) < \lambda\} \geq \pi_0 \Phi(\lambda), \end{aligned}$$

where the first inequality in the second line is tighter if $(1-\pi_0)P\{\mathcal{N}(\mu_1(s), 1) < \lambda\}$ is closer to zero and the second inequality becomes an equality when $\mu_0(s) = 0$. The above derivation suggests a conservative estimator for π_0 given by

$$\widehat{\pi}_0 := \frac{\sum_{s \in \mathcal{S}} \mathbf{1}\{T_2(s) < \lambda\}}{|\mathcal{S}|\, \Phi(\lambda)} \approx \frac{P\{T_2(s) < \lambda\}}{\Phi(\lambda)} \geq \pi_0. \tag{2.6}$$

2.6 A feasible procedure

So far we have assumed that the spatial covariance function $k(s, s') = \mathrm{cov}\{\epsilon(s), \epsilon(s')\}$ of the error process is known. In practice, we need to estimate the spatial covariance function $\mathrm{cov}\{\epsilon(s), \epsilon(s')\}$, which has been widely investigated in the literature (Sang and Huang, 2012; Katzfuss and Guinness, 2021); see Section S.VI of the supplement for more details. Given the

estimated covariance function, we let $\widehat{T}_1(s)$ and $\widehat{T}_2(s)$ respectively denote the feasible statistics of $T_1(s)$ and $T_2(s)$ by replacing $(\sigma(s), \tau(s), \rho(s))$ with their estimates $(\widehat{\sigma}(s), \widehat{\tau}(s), \widehat{\rho}(s))$. Let $\widehat{G}_{\widetilde{\mathcal{S}}}(u)$ denote the nonparametric empirical Bayes estimate of $G_{\mathcal{S}}$ based on $\{\widehat{T}_1(s) : s \in \widetilde{\mathcal{S}}\}$. We propose the following FDP estimate which accounts for the null proportion using the idea in Section 2.5,

$$\widehat{\mathrm{FDP}}_{\lambda,\widetilde{\mathcal{S}}}(t_1, t_2) := \frac{\sum_{s\in\mathcal{S}} \mathbf{1}\{\widehat{T}_2(s) < \lambda\} \sum_{s\in\mathcal{S}} \int L\{t_1, t_2, x, \widehat{\rho}(s)\}\, d\widehat{G}_{\widetilde{\mathcal{S}}}(x)}{m\Phi(\lambda)} \Big/ \frac{}{1 \vee \widehat{R}(t_1, t_2)}, \tag{2.7}$$

where $\widehat{R}(t_1, t_2) = \sum_{s\in\mathcal{S}} \mathbf{1}\big\{\widehat{T}_1(s) \geq t_1, \widehat{T}_2(s) \geq t_2\big\}$. Thus, given the desired FDR level $q \in (0, 1)$, the optimal rejection cutoffs are defined as

$$(\widetilde{t}_1^{\star}, \widetilde{t}_2^{\star}) = \underset{(t_1,t_2)\in\mathcal{F}_q}{\arg\max}\, \widehat{R}(t_1, t_2), \tag{2.8}$$

where $\mathcal{F}_q = \{(t_1, t_2) \in \mathbb{R}^2 : \widehat{\mathrm{FDP}}_{\lambda,\widetilde{\mathcal{S}}}(t_1, t_2) \leq q\}$. The left panel in Figure 1 exemplifies the cutoffs for the BH and 2d-SMT procedures with the target FDR level at 10%. Compared to the BH, the 2d-SMT realizes a lower cutoff for $T_2(s)$ (the red vertical line) as it excludes locations exhibiting weak neighboring signals with a cutoff for $T_1(s)$ (the red horizontal line). The lower cutoff for $T_2(s)$ in 2d-SMT leads to more true rejections in this example.

2.7 Spatial varying null proportions and cutoffs

Our proposed 2d-SMT is a flexible framework that can accommodate weighted BH procedures (wBH), a broad class of multiple testing procedures. The wBH method leverages the hypothesis heterogeneity by assigning location-specific cutoffs. According to wBH, $\mathcal{H}_{0,s}$ will be rejected with the rule $p(s) := 1 - \Phi(\widehat{T}_2(s)) \leq \min\{\tau, w(s)t\}$, $\tau \in (0, 1]$, where τ is the censoring level for all p-values and $w(s)$ is a location-specific weight that encodes external information for

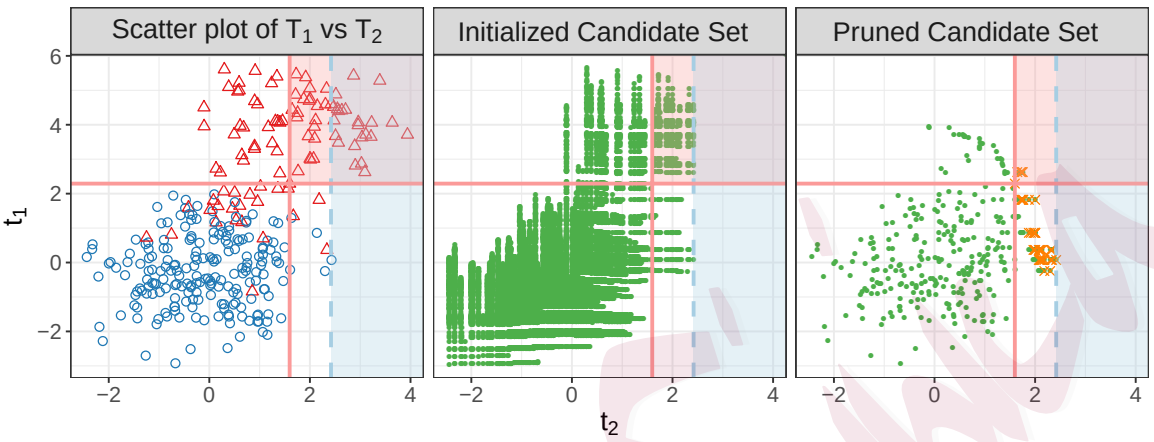


Figure 1: An illustration of 2d-SMT and the fast searching algorithm. The spatial domain $\mathcal{S}$ includes 300 points. The red triangle ($\triangle$) and blue circle ($\circ$) in the left panel denote non-null and null locations, respectively. The solid and dashed lines are the cutoffs for the 2d-SMT procedure and the BH procedure, respectively. The orange cross ($\times$) in the right panel corresponds to the cutoff whose corresponding FDP is less than 0.1. The set $\mathcal{T}'$ in the middle panel contains 14,534 points, whereas the set after our proposed pruning steps in the right panel significantly reduces the number of candidates to 345.

location s. Apparently, the rejection rule is equivalent to assigning location-specific cutoffs to the primary statistic, i.e., $\widehat{T}_2(s) \geq \Phi^{-1}(1 - \min\{\tau, w(s)t\})$. Inspired by the idea behind wBH, we extend 2d-SMT by allowing location-specific cutoffs to further incorporate external information on the prior null probability and signal distribution. Specifically, we reject $\mathcal{H}_{0,s}$ whenever $\widehat{T}_1(s) \geq c(t_1, s)$ and $\widehat{T}_2(s) \geq c(t_2, s)$, where $c(t, s) = \Phi^{-1}(1 - \min\{\tau, w(s)t\})$. With some abuse of notation, we let

$$\widehat{\mathrm{FDP}}_{\lambda,\widetilde{\mathcal{S}}}(t_1, t_2) := \frac{\sum_{s\in\mathcal{S}} \widehat{\pi}_0(s) \int L\left\{c(t_1, s), c(t_2, s), x, \widehat{\rho}(s)\right\} d\widehat{G}_{\widetilde{\mathcal{S}}}(x)}{1 \vee \widehat{R}(t_1, t_2)}, \tag{2.9}$$

where $\widehat{\pi}_0(s)$ is an estimate of the null proportion $\pi_0(s) = P\{\theta(s) = 0\}$ at location s. We reject $\mathcal{H}_{0,s}$ if $T_1(s) \geq c(\widetilde{t}_1^\star, s)$ and $T_2(s) \geq c(\widetilde{t}_2^\star, s)$, where $(\widetilde{t}_1^\star, \widetilde{t}_2^\star)$ is the solution to (2.8) with the FDP estimate given in (2.9).

There have been extensive recent studies on the choice of $w(s)$ in wBH. Examples include GBH (Hu et al., 2010), IHW (Ignatiadis et al., 2016; Ignatiadis and Huber, 2021), SABHA (Li and Barber, 2019) and LAWS (Cai et al., 2022). The weights in these examples are either proportional to $1/\pi_0(s)$ or $\{1-\pi_0(s)\}/\pi_0(s)$, where $\pi_0(s)$ can be estimated using various approaches. In the spatial setting, we often use the location associated with each hypothesis as the external covariate to estimate $\pi_0(s)$.

If additional types of covariate information are available, the proposed 2d-SMT framework is able to leverage both covariate and spatial information. Some covariate-adaptive FDR procedures can be utilized by assigning weights based on an estimation of covariate-specific null proportions; and the spatial information is, again, captured by the auxiliary test statistics. In Section S.I.6 of the supplement, we present a simulation experiment that uses the group information as the covariate. The simulation results show that integrating both the group covariate and the spatial information can further improve the detection power.

3. Implementation Details

In this section, we discuss a few crucial points for the implementation of our method. First of all, the auxiliary statistic $\widehat{T}_1(s)$ requires the specification of a pre-chosen neighborhood $\mathcal{N}(s)$ for each location. In our implementation, we let $\mathcal{N}(s)$ be the set of the κ-nearest neighbors around location s, and find that 2d-SMT is not quite sensitive to the choice of κ and shows satisfactory power improvement provided $2 \leq \kappa \leq 7$; see Section S.I.4 of the supplement for more details. Alternatively, one can also choose the locations within a certain distance from the location s of interest as its neighbors. Furthermore, the neighborhood set of location s can also be adaptively determined when external information that is independent of $\{X(s) : s \in \mathcal{S}\}$ is provided; see

Section S.I.5 of the supplement for more details.

Second, the auxiliary statistics used in NPEB estimation, $\{\widehat{T}_1(s) : s \in \widetilde{\mathcal{S}}\}$, should be from far enough spatial locations to weaken their spatial dependency. In practice, we choose $\widetilde{\mathcal{S}} \subset \mathcal{S}$ whose neighbors have no overlaps.

Third, when the signal is sparse, the estimation of the number of false rejections in the 2d-SMT procedure may be unstable. Inspired by the idea of Knockoff+, we add a small offset to improve the selection stability. More precisely, we replace $\sum_{s\in\mathcal{S}} \int L\{t_1, t_2, x, \widehat{\rho}(s)\}\, d\widehat{G}_{\widetilde{\mathcal{S}}}(x)$ in (2.7) by $\sum_{s\in\mathcal{S}} \int L\{t_1, t_2, x, \widehat{\rho}(s)\}\, d\widehat{G}_{\widetilde{\mathcal{S}}}(x) + q$, where q is the target FDR level. This replacement improves the selection stability for sparse signals but does not influence the power and FDR control for dense signals.

Finally, finding the optimal cutoffs in 2d-SMT requires solving the discrete constrained optimization problem (2.8). Due to the discrete nature of (2.8), the solution can be obtained if we replace $\mathcal{F}_q$ by $\{(t_1, t_2) \in \mathcal{T} : \widehat{\mathrm{FDP}}_{\lambda,\widetilde{\mathcal{S}}}(t_1, t_2) \le q\}$, where $\mathcal{T} = \{(\widehat{T}_1(s), \widehat{T}_2(s')) : s, s' \in \mathcal{S}\}$ is the set of all candidate cutoffs. A naive grid search algorithm would evaluate $\widehat{\mathrm{FDP}}_{\lambda,\widetilde{\mathcal{S}}}$ at $|\mathcal{S}|^2$ values, which is computationally prohibitive for a large number of spatial locations. To overcome the computational bottleneck, we propose a fast algorithm to utilize the specific structure of (2.8) through the following three steps. We briefly introduce the basic idea below and defer the comprehensive discussion to Section S.VII of the supplement.

Step 1. We maintain the cutoffs that achieve the minimum FDP among all the cutoffs realizing the same rejection sets. The derivation in Section S.VII of the supplement suggests that we

Algorithm 1 Fast Searching Algorithm.

Input: Test Statistics $\left\{\left(\widehat{T}_1(s), \widehat{T}_2(s)\right) : s \in \mathcal{S}\right\}$; target FDR level q;

Initialization:

Initialize $(\widetilde{t}_1^\star, \widetilde{t}_2^\star, R_{\max}, \mathrm{FDP}_{\min}) = \left(-\infty, \widetilde{t}_2^\#, \widehat{R}(-\infty, \widetilde{t}_2^\#), \widehat{\mathrm{FDP}}_{\lambda,\widetilde{\mathcal{S}}}(-\infty, \widetilde{t}_2^\#)\right)$

Search Step:

for $i = 1, 2, \ldots, \breve{m}$ **do**
  Calculate $R = \widehat{R}(t_{1,i,1}, t_{2,i})$;
  Set $j = 1 + \max(0, R_{\max} - R)$;
  while $j \le m_i$ **do**
    Calculate $R = \widehat{R}(t_{1,i,j}, t_{2,i})$ and $\widehat{\mathrm{FDP}} = \widehat{\mathrm{FDP}}_{\lambda,\widetilde{\mathcal{S}}}(t_{1,i,j}, t_{2,i})$ according to (2.7);
    if $R = R_{\max}$ and $\widehat{\mathrm{FDP}} < \mathrm{FDP}_{\min}$ or $R > R_{\max}$ and $\widehat{\mathrm{FDP}} \le q$ **then**
      $(\widetilde{t}_1^\star, \widetilde{t}_2^\star, R_{\max}, \mathrm{FDP}_{\min}) = (t_{1,i,j}, t_{2,i}, R, \widehat{\mathrm{FDP}})$;
      Update $j = j + 1$;
    else
      Calculate $R_{\mathrm{req}} = \lceil \widehat{\mathrm{FDP}} \times R/q \rceil$;
      Update $j = j + \max(1, R_{\mathrm{req}} - R)$;
    end if
  end while
end for

Output: Rejection cutoff $(\widetilde{t}_1^\star, \widetilde{t}_2^\star)$.

only need to consider the following set of candidate cutoffs:

$$\mathcal{T}' = \left\{(\widehat{T}_1(s_l), \widehat{T}_2(s_k)) : \widehat{T}_1(s_l) \le \widehat{T}_1(s_k) \text{ and } l \le k,\ k = 1, 2, \ldots, m\right\} \cup \{(\infty, \infty)\}.$$

Step 2. Let $\widetilde{t}_2^\#$ be the minimum value satisfying $\widehat{\mathrm{FDP}}_{\lambda,\widetilde{\mathcal{S}}}(-\infty, t_2) \le q$. It is not hard to see that the optimal cutoff for the primary statistic in 2d-SMT is no more than $\widetilde{t}_2^\#$. Hence we can reduce the set of candidate cutoffs to $\mathcal{T}'' = \left\{(t_1, t_2) \in \mathcal{T}' : t_2 \le \widetilde{t}_2^\#\right\}$.

Step 3. Denote the elements in $\mathcal{T}''$ by $(t_{1,i,j}, t_{2,i})$ for $i = 1, 2 \ldots, \breve{m}$ and $j = 1, \ldots, m_i$, where $\breve{m}$ is the number of $\widehat{T}_2(s)$'s that are smaller than or equal to $\widetilde{t}_2^\#$. Suppose the points are sorted in the following way: (1) $t_{2,1} > t_{2,2} > \cdots > t_{2,\breve{m}}$; (2) $t_{1,i,1} > t_{1,i,2} > \cdots > t_{1,i,m_i}$ for all $1 \le i \le \breve{m}$. Our algorithm involves two loops. In the outer loop, we search the cutoff for the primary statistic,

while in the inner loop, we search the cutoff for the auxiliary statistic. The key idea here is to skip those cutoffs in the inner loop that are impossible to procedure a value of $\widehat{\text{FDP}}$ equal to or below the level q. For example, consider a cutoff $(t_{1,i,j}, t_{2,i})$ which induces R rejections with $\widehat{\text{FDP}}_{\lambda,\widetilde{\mathcal{S}}}(t_{1,i,j}, t_{2,i}) = 2q$. Then the next cutoff, denoted as $(t_{1,i,j'}, t_{2,i})$, needs to induce at least $2R$ rejections to ensure that $\widehat{\text{FDP}}_{\lambda,\widetilde{\mathcal{S}}}(t_{1,i,j'}, t_{2,i}) \leq q$. When there is no tie, increasing j by k brings exactly k more rejections. Therefore, the next possible cutoff to be examined is $(t_{1,i,j+R}, t_{2,i})$. The middle and left panels in Figure 1 illustrate how the set of candidate cutoff values can be reduced by Step 3.

4. Asymptotic Results

In this section, we investigate the asymptotic property of the 2d-SMT procedure. We observe $\{X(s) : s \in \mathcal{S}_m\}$, where $\mathcal{S}_m = \{s_1, s_2, \cdots, s_m\} \subseteq \mathcal{U} \subseteq \mathbb{R}^K$, $\mathcal{U}_0 = \{s \in \mathcal{U}_0 : \theta(s) = 0\}$ and $\mathcal{U}_1 = \{s \in \mathcal{S} : \theta(s) = 1\}$. We denote by $\mathcal{S}_{0,m} = \mathcal{S}_m \cap \mathcal{U}_0$ with $m_0 = |\mathcal{S}_{0,m}|$ be the set of null locations and let $\mathcal{S}_{1,m} = \mathcal{S}_m \cap \mathcal{U}_1$ with $m_1 = |\mathcal{S}_{1,m}|$ be the set of non-null locations. We further let $\widetilde{\mathcal{S}}_m \subseteq \mathcal{S}_m$ with $|\widetilde{\mathcal{S}}_m| = \tilde{m}$ be the set of randomly selected locations for implementing the NPEB. Our asymptotic analysis requires the following regularity conditions.

Assumption 1 (Spatial Domain)**.** The spatial domain $\mathcal{U} \subset \mathbb{R}^K$ is infinitely countable. There exist $0 < \Delta_l < \Delta_u < \infty$, such that for every element in $\mathcal{U}$, the distance to its nearest neighbor is bounded from below and above respectively by Δ_l and Δ_u, i.e., $\Delta_l \leq \inf_{s' \neq s} \text{dist}(s, s') \leq \Delta_u$, for all $s \in \mathcal{U}$, where $\text{dist}(\cdot, \cdot)$ is the Euclidean distance of two points.

Assumption 2 (Neighborhood)**.** For each location s in $\mathcal{U}$, its neighborhood $\mathcal{N}(s) \subset \mathcal{S} \backslash \{s\}$ used in 2d-SMT is its nearest neighbors with size uniformly upper bounded by some positive integer $N_{nei} \in \mathbb{N}^+$, i.e., $0 < |\mathcal{N}(s)| < N_{nei}, \ \forall s \in \mathcal{U}$.

Assumption 1 states that the distance between any location in $\mathcal{U}$ and its nearest neighbor is moderately uniform (Cressie, 1993; Jenish and Prucha, 2012). One example satisfying Assumption 1 is the lattice $\mathcal{U} = \mathbb{Z}^K$ where $\Delta_l = \Delta_u = 1$. Particularly, the lower bound ensures $\max_{s,s' \in \mathcal{S}_m} \text{dist}(s, s')$ tend to infinity as m increases. Assumption 2 specifies the neighborhood of each location and requires the number of neighbors to be bounded for each location.

Assumption 3 (Second-order Structure)**.** The variance and covariance of the error process $\epsilon(s)$ satisfy: (a) There exist positive constants $B_{ud,\sigma}$, $B_{up,\sigma}$, $B_{ud,\tau}$, and $B_{up,\tau}$, such that $B_{ud,\sigma} \leq \inf_{s\in\mathcal{U}} \sigma(s) \leq \sup_{s\in\mathcal{U}} \sigma(s) < B_{up,\sigma}$ and $B_{ud,\tau} \leq \inf_{s\in\mathcal{U}} \tau(s) \leq \sup_{s\in\mathcal{U}} \tau(s) < B_{up,\tau}$ for all $s \in \mathcal{U}$; (b) The estimated covariance of X are uniformly weakly consistent with a polynomial rate, i.e., $\sup_{s,s'\in\mathcal{S}_m} |\widehat{\text{cov}}\{\epsilon(s), \epsilon(s')\} - \text{cov}\{\epsilon(s), \epsilon(s')\}| = o_P(m^{-q})$ for some $q > 0$.

Assumptions 3(a) requires the variance of the error process to be bounded away from zero and infinity. Assumption 3(b) imposes condition on the convergence rate of the covariance estimate, which is satisfied by many commonly-used estimators. For example, the maximum likelihood estimator achieves the desired convergence rate with $q = 1/2$ when the parametric covariance function is locally Lipschitz continuous (Mardia and Marshall, 1984).

To describe the spatial dependence structure, we adopt the near epoch dependency (NED), which has been extensively studied in time series analysis (Davidson, 1994) and first introduced to the spatial analysis by Jenish and Prucha (2012). The NED is satisfied by many classical models in spatial statistics, e.g., spatial autoregression models (Cliff and Ord, 1981). Denote by $\mathcal{V}$ a spatial domain such that $\mathcal{U} \subseteq \mathcal{V} \subset \mathbb{R}^K$. Let $\{Y(v), v \in \mathcal{V}\}$ be a random field and set $\mathfrak{F}(S) = \sigma\{Y(s), s \in S\}$ as the σ-field generated by $Y(s)$ for $s \in S \subset \mathcal{V}$. In reality, $\mathcal{V}$ represents a physical spatial domain in consideration, which might encompass a wide geographical area;

$\mathcal{S}$, on the other hand, is a set of locations to collect the data.

Definition 1 (NED)**.** Let $X = \{X(s), s \in \mathcal{S}_m\}$ be a random field with $\|X(s)\|_p < \infty$ where $\|X(s)\|_p = (E|X(s)|^p)^{1/p}$, $p \geq 1$. Define $Y = \{Y(s), s \in \mathcal{V}_m\}$ as a random field where $\mathcal{S}_m \subseteq \mathcal{V}_m$ and let $\mathbf{d} = \{d_m(s), s \in \mathcal{S}_m\}$ be a set of finite positive constants. Then, X is said to be $L_p(\mathbf{d})$-near-epoch dependent on Y if

$$\|X(s) - E\{X(s) \mid \mathfrak{F}(S)\}\|_p \leq d_m(s)\psi(r),$$

where $S \subset \mathcal{V}_m$, $r = \max\{t \geq 0 : B(s;t) \subseteq S\}$ with $B(s;t)$ being a ball centered around s with radius t, and $\psi(r) \geq 0$ is some non-increasing sequence with $\lim_{r\to\infty} \psi(r) = 0$. Here $\psi(r)$ and $d_m(s)$ are called NED coefficients and NED scaling factors, respectively. We say X is $L_p(\mathbf{d})$-NED of size $-\lambda$ if $\psi(r) = O(r^{-\mu})$ for some $\mu > \lambda > 0$. Furthermore, if $\sup_m \sup_{s\in\mathcal{S}_m} d_m(s) < \infty$, we say X is uniformly L_p-NED on Y.

Assumption 4 (Uniform NED)**.** $X = \{X(s), s \in \mathcal{S}_m\}$ is uniformly L_p-NED on $Y = \{Y(s) : s \in \mathcal{V}_m\}$ of size $-\lambda$, for $\lambda > 0$, $p \geq 2$, where $Y(s)$ are independently distributed.

In Definition 1, $B(s;t)$ is the maximum ball that is completely included in S. The quantity $\|X(s) - E\{X(s) \mid \mathfrak{F}(S)\}\|_p$ satisfies a generalized non-decreasing property with respect to S. Specifically, if $\|X(s) - E\{X(s) \mid \mathfrak{F}(S)\}\|_p \leq d_m(s)\psi(r)$, we then have $\|X(s) - E\{X(s) \mid \mathfrak{F}(V)\}\|_p \leq d_m(s)\psi(r)$ for any $S \subseteq V \subseteq \mathcal{S}$. We can choose $d_m(s) = 2\|X(s)\|_p$ as the scaling factor such that $0 \leq \psi(r) \leq 1$. In addition, if X is L_p-NED on Y, then X is also L_q-NED on Y with the same $\{d_m(s)\}$ and $\psi(r)$ for any $q \leq p$. The condition $p \geq 2$ in Assumption 4 guarantees that the variance of $E\{X(s) \mid \mathfrak{F}(S)\}$ converges to that of $X(s)$ as S expands. Jenish and

Prucha (2012) demonstrated that the NED property can be validated in examples from both Gorodetskii (1978) and Andrews (1984). The former showed that strong mixing might fail in linear processes with normal innovations and slowly declining coefficients, while the latter illustrated the absence of α-mixing in a simple first-order autoregressive process with independent Bernoulli innovations. Furthermore, for infinity-order moving average random fields, verifying the NED property involves checking the smoothness of the functional form and the absolute summability of coefficients, which is an easier task compared to verifying mixing conditions.

Under the NED assumption, we can approximate $X(s)$ with $E[X(s) \mid \mathfrak{F}\{B(s;r)\}]$ and in turn approximate $T_1(s)$ with

$$T_1^*(s;r) = E\left[T_1(s) \mid \mathfrak{F}\left\{\cup_{v\in\mathcal{N}(s)} B(v;r)\right\}\right]/\zeta_r(s), \tag{4.1}$$

where $\zeta_r^2(s) = \mathrm{Var}\left(E\left[T_1(s) \mid \mathfrak{F}\left\{\cup_{v\in\mathcal{N}(s)} B(v;r)\right\}\right]\right)$. We further require the conditional statistics to be normally distributed as required by the theory for NPEB.

Assumption 5 (Normality)**.** For any $S \subset \mathcal{V}$, $E\{X(s) \mid \mathfrak{F}(S)\}$ is normal.

Spatial linear autoregression models with Gaussian white noise satisfy Assumptions 4 and 5 (Jenish, 2012). Under Assumptions 4 and 5, $\{T_1^*(s;r) : s \in \widetilde{\mathcal{S}}_m\}$ defined in (4.1) are independent and normally distributed random variables with unit variance if $r > 0$ satisfies

$$\cup_{v\in\mathcal{N}(s)} B(v;r) \cap \cup_{v\in\mathcal{N}(s')} B(v;r) = \emptyset, \quad s, s' \in \widetilde{\mathcal{S}}_m. \tag{4.2}$$

Under Assumptions 1 and 2, setting $r = \widetilde{\Delta}_{l,m}/2 - N_{nei}\Delta_u$ indicates (4.2), where

$$\widetilde{\Delta}_{l,m} = \inf_{s,s'\in\widetilde{\mathcal{S}}_m} \mathrm{dist}(s, s'). \tag{4.3}$$

Assumption 6 (Size of Subset)**.** The subset for implementing NPEB, $\widetilde{\mathcal{S}}_m \subset \mathcal{S}_m$, satisfies $\tilde{m}^{1/(\lambda p)}\{\log(\tilde{m})\}^{-1/(2\lambda)} = o(\widetilde{\Delta}_{l,m})$ and $\tilde{m} \to \infty$ as $m \to \infty$, where $\widetilde{\Delta}_{l,m}$ is defined as in (4.3).

Let $G_m(u) = m^{-1}\sum_{s\in\mathcal{S}_m} \mathbf{1}\{\xi(s) \le u\}$ and $G_{\tilde{m}}(u) = \tilde{m}^{-1}\sum_{s\in\widetilde{\mathcal{S}}_m} \mathbf{1}\{\xi(s) \le u\}$ be the empirical distributions of the unknown means for $s \in \mathcal{S}_m$ and $s \in \widetilde{\mathcal{S}}_m$, respectively.

Assumption 7 (Limiting Distribution)**.** There exists a limiting distribution G_0 on $[-\nu_0, \nu_0]$ for some positive constant ν_0 such that $d_H(f_{G_m}, f_{G_0}) = o_P(1)$, as $m \to \infty$, and the subset $\widetilde{\mathcal{S}}_m$ satisfies $d_H(f_{G_{\tilde{m}}}, f_{G_0}) = o_P(1)$, as $\tilde{m} \to \infty$, where $f_G(x) = \int \phi(x-u)dG(u)$ and $d_H^2(f,g) = 2^{-1}\int\left\{\sqrt{f(x)} - \sqrt{g(x)}\right\}^2 dx$ is the squared Hellinger distance between densities f and g.

The condition about $\tilde{m}$ and $\widetilde{\Delta}_{l,m}$ in Assumption 6 trades off the number of $\widetilde{\mathcal{S}}_m$ and the distance between them, which becomes milder when decreasing the strength of near epoch dependency, i.e., increasing λ or p. In fact, we can choose $\widetilde{\mathcal{S}}_m$ satisfying Assumption 6 under Assumptions 1 and 2; see Section S.IV.1 of the supplement for more details. Assumption 7 states that the empirical distributions of $\xi(s)$ using $s \in \mathcal{S}_m$ and $s \in \widetilde{\mathcal{S}}_m$ have the same limiting distribution G_0, as the number of spatial samples goes to infinity.

Assumption 8 (Null Proportion)**.** $m_0/m \to \pi_0 \in (0,1)$, as $m \to \infty$.

Assumption 9 (Asymptotic True/False Rejection Proportion)**.** As m_0 and m_1 tend to infinity, for every $(t_1, t_2) \in \mathbb{R} \times \mathbb{R}$, we have

$$\frac{\sum_{s\in\mathcal{S}_{0,m}} P\left\{T_1(s) \ge t_1, T_2(s) \ge t_2\right\}}{m_0} := E\left\{\frac{V_m(t_1,t_2)}{m_0}\right\} \to K_0(t_1,t_2),$$
$$\frac{\sum_{s\in\mathcal{S}_{1,m}} P\left\{T_1(s) \ge t_1, T_2(s) \ge t_2\right\}}{m_1} := E\left\{\frac{S_m(t_1,t_2)}{m_1}\right\} \to K_1(t_1,t_2),$$

where $K_0(t_1,t_2) \le \lim_{m\to\infty}\sum_{s\in\mathcal{S}_m}\int L\{t_1,t_2,x,\rho(s)\}\,dG_0(x)/m$, and both $K_0(t_1,t_2)$ and $K_1(t_1,t_2)$ are non-negative continuous functions of (t_1,t_2).

Let $K_1(-\infty, t_2)$ be the limit of $\sum_{s\in\mathcal{S}_{1,m}} P\{T_2(s) \geq t_2\}/m_1$ as m_1 goes to infinity. Define

$$\mathrm{FDP}^{\infty}_{\lambda}(t_1, t_2) := \lim_{m\to\infty} \frac{F(\lambda)\sum_{s\in\mathcal{S}_m}\int L(t_1, t_2, x, \rho(s))dG_0(x)}{m\Phi(\lambda)K(t_1, t_2)}, \tag{4.4}$$

where $K(t_1, t_2) = \pi_0 K_0(t_1, t_2) + (1-\pi_0)K_1(t_1, t_2)$ and $F(\lambda) = \sum_{s\in\mathcal{S}_m} \pi_0\Phi(\lambda) + (1-\pi_0)\{1 - K_1(-\infty, \lambda)\}$.

Assumption 10 (Existence of Cutoffs)**.** There exist $t_1^\star$ and $t_2^\star$ such that $\mathrm{FDP}^{\infty}_{\lambda}(t_1^\star, 0) < q$, $\mathrm{FDP}^{\infty}_{\lambda}(0, t_2^\star) < q$, and $K(t_1^\star, t_2^\star) > 0$.

Assumption 8 requires the asymptotic null proportion to be strictly between zero and one, which rules out the case of very sparse signals. Assumption 9 characterizes the expected proportions of true and false rejections among the alternative and null hypotheses, respectively, as functions with respect to threshold (t_1, t_2). The upper bound condition of $K_0(t_1, t_2)$ appeared in Assumption 9 is valid in many examples. For instance, it is fulfilled when (a) the signal strength around the null locations is weaker than that around the non-null locations on average; and (b) the error process is stationary, $\mathcal{S}_m$ is observed at the lattice $\mathcal{S} = \mathbb{Z}^K$, and $\mathcal{N}(s)$ is selected as the κ-nearest neighbors, so that $\rho(s) \equiv \rho$ for some $\rho \in (0, 1)$. Assumption 10 reduces the searching region for the optimal cutoff to a rectangle. Assumptions 8, 9, and 10 align with existing conditions for proving asymptotic FDR control (Storey et al., 2004; Ferreira and Zwinderman, 2006; Benjamini and Heller, 2007), in which just the primary statistic is considered.

These assumptions are in general mild conditions for asymptotic FDR control in spatial multiple testing. Assumptions 1, 3, 8, 9, and 10 are widely used in spatial multiple testing (Cressie, 1993; Storey et al., 2004; Ferreira and Zwinderman, 2006; Benjamini and Heller, 2007;

Jenish and Prucha, 2012); and Assumptions 2, 4, 5, 6, and 7 are specifically required for the 2d-SMT method. In particular, the condition on distances between locations in Assumption 1 is requisite for increasing domain asymptotics (Cressie, 1993). Assumptions 2 and 6 regularize the sizes of subsets for constructing auxiliary statistics and implementing NPEB. When a practitioner implements the 2d-SMT method, these two assumptions can be directly satisfied by choosing suitable neighbors and a subset used for NPEB. Assumption 3 imposes regular conditions on the moment of the error process and requires the covariance of the error process can be consistently estimated (Sun et al., 2015). Assumptions 4, 5, and 7 are technical conditions for showing the convergence of the GMLE under certain dependencies: Assumption 4 imposes some specific weak dependence on the random field; Assumption 5 is crucial as we utilize the theory of NPEB developed for Gaussian location model; Assumption 7 requires that the empirical distribution of means possesses a limit. Assumptions 8, 9, and 10 naturally generalize the conditions for asymptotic FDR control in the Storey's procedure (Storey et al., 2004; Ferreira and Zwinderman, 2006; Benjamini and Heller, 2007), where the rejection regions transit from one-dimensional to two-dimensional.

The FDR of the 2d-SMT procedure is given by

$$\widetilde{\mathrm{FDR}}_m = E\left\{\mathrm{FDP}\left(\widetilde{t}_1^{\star}, \widetilde{t}_2^{\star}\right)\right\}, \quad \text{with} \quad \mathrm{FDP}(t_1,t_2) = \frac{\widehat{V}_m(t_1,t_2)}{\widehat{V}_m(t_1,t_2) + \widehat{S}_m(t_1,t_2)}, \tag{4.5}$$

where $(\widetilde{t}_1^{\star}, \widetilde{t}_2^{\star})$ is defined in (2.8), and $\widehat{V}_m(t_1,t_2) = \sum_{s\in\mathcal{S}_{0,m}} \mathbf{1}\{\widehat{T}_1(s) \geq t_1, \widehat{T}_2(s) \geq t_2\}$ and $\widehat{S}_m(t_1,t_2) = \sum_{s\in\mathcal{S}_{1,m}} \mathbf{1}\{\widehat{T}_1(s) \geq t_1, \widehat{T}_2(s) \geq t_2\}$ are respectively the numbers of false and true rejections.

Theorem 1. Under Assumptions 1–10, we have $\limsup_{m\to\infty} \widetilde{\mathrm{FDR}}_m \le q$.

Theorem 1 states that 2d-SMT procedure asymptotically controls the FDR. The proof of Theorem 1, which is deferred to Section S.II of the supplement, relies on two facts: (i) (2.7) uniformly converges to (4.4) and $\mathrm{FDP}(t_1, t_2)$ in (4.5) satisfies the uniform law of large numbers over the rectangle encompassed by $|t_1| \le t_1^\star$ and $|t_2| \le t_2^\star$; (ii) (4.4) is asymptotically larger than $\mathrm{FDP}(t_1, t_2)$ in (4.5). We make two technical innovations. First, we establish $d_H\left(f_{\widehat{G}_{\tilde{m}}}, f_{G_0}\right) = o_P(1)$ in Lemma S.8 of the supplement, which is the first result, to our knowledge, for the convergence of GMLE estimated from dependent observations. Second, we address the challenge posed by the non-Lipschitz nature of $\mathbf{1}\{T_1(s) \ge t_1, T_2(s) \ge t_2\}$ in Lemma S.9 of the supplement, which expands the applications of the NED-based law of large numbers.

We next formalize the power improvement of 2d-SMT over 1d-SMT (setting t_1 as $-\infty$) and defer the proof to Section S.V.1 of the supplement. Denote the thresholds for 1d- and 2d-SMT as $t_2^{1d} = \arg\max_{t_2 \in \mathcal{F}_{q,\infty}^{1d}} K(-\infty, t_2)$ and $(t_1^{2d}, t_2^{2d}) = \arg\max_{(t_1,t_2)\in\mathcal{F}_{q,\infty}^{2d}} K(t_1, t_2)$, respectively, where $\mathcal{F}_{q,\infty}^{1d} = \{t_2 : \mathrm{FDP}_\lambda^\infty(-\infty, t_2) \le q\}$ and $\mathcal{F}_{q,\infty}^{2d} = \{(t_1, t_2) : \mathrm{FDP}_\lambda^\infty(t_1, t_2) \le q\}$. The corresponding percentages of true discoveries in the asymptotic sense are respectively defined as $\mathrm{PTD}^{1d} = K_1(-\infty, t_2^{1d})$ and $\mathrm{PTD}^{2d} = K_1(t_1^{2d}, t_2^{2d})$.

Theorem 2. Under Assumptions 8 and 9, we have (i) $\mathrm{PTD}^{2d} \ge \mathrm{PTD}^{1d}$ and (ii) $\mathrm{PTD}^{2d} > \mathrm{PTD}^{1d}$ if additionally $K_0\left(t_1^{2d}, t_2^{2d}\right) < \lim_{m\to\infty} \sum_{s\in\mathcal{S}_m} \int L\{t_1^{2d}, t_2^{2d}, x, \rho(s)\}\, dG_0(x)/m$.

Theorem 2 shows that 2d-SMT offers superior power compared to 1d-SMT. Interestingly, 2d-SMT can achieve a strictly higher power than 1d-SMT, even when both methods yield the same total number of discoveries, i.e., $K(t_1^{2d}, t_2^{2d}) = K(-\infty, t_2^{1d})$. This power boost, as discussed alongside Assumption 9, stems not only from more discoveries but also because signals near null

locations tend to be weaker than those near non-null locations.

5. Simulation Studies

We conduct extensive simulations to evaluate the performance of the proposed 2d-SMT procedure. We consider various simulation settings, including (1) 1d and 2d spatial domains; (2) known and unknown covariance structures; (3) different signal shapes; and (4) simple and composite nulls. We investigate the specificity and sensitivity of different methods under different settings by varying the signal intensities, magnitudes, and degrees of dependency.

We compare the 2d procedure with the following competing methods: Storey's procedure (Storey, 2002, ST); Independent hypothesis weighting (Ignatiadis et al., 2016; Ignatiadis and Huber, 2021, IHW); Structure adaptive BH procedure with the stepwise constant weights (Li and Barber, 2019, SABHA); Locally adaptive weighting and screening (Cai et al., 2022, LAWS); Adaptive p-value thresholding procedure (Lei et al., 2018, AdaPT); Dependence-adjusted BH procedure (Fithian and Lei, 2022, dBH). As discussed in Section 2.7, our idea can combine with the ST, SABHA, and IHW methods to further enhance their power by borrowing neighboring information. We denote the corresponding procedures by 2D (ST), 2D (SA), and 2D (IHW), respectively, and include them in the numerical comparisons.

Throughout, we focus on testing the one-sided hypotheses $\mathcal{H}_{0,s} : \mu(s) \leq 0$ versus $\mathcal{H}_{a,s} : \mu(s) > 0$. We set the target FDR level at $q = 0.1$ and report the FDP and power (the number of true discoveries divided by the total number of signals) by averaging over 100 simulation runs. For the set of neighbors $\mathcal{N}(s)$ of the location s, we use the κ-nearest neighbors for each s. A sensitivity analysis of κ is conducted in Section S.I.4 of the supplement and empirically suggests κ to be an integer between 2 and 7. In this section, we use $\kappa = 4$.

We consider the process $X(s) = \mu(s) + \epsilon(s)$ defined on the one-dimensional domain $\mathcal{S} = [0, 30]$. We observe the process $X(s)$ at 900 locations that are evenly distributed over $\mathcal{S}$. We introduce three *data generating mechanisms* for the signal process $\mu(s)$ and consider three *signal sparsity levels* within each mechanism.

- **Setup I**: $\mu(s) = \gamma\mu_0(s)$, where γ determines the magnitude and $\mu_0(s)$ is generated from B-spline basis functions to control the signal densities and locations. Three different shapes of $\mu_0(s)$ are considered, which correspond to the sparse, medium, and dense signal cases, respectively.
- **Setup II**: $\mu(s) = \gamma\delta(s)$ with $\delta(s) \sim \text{Bernoulli}(\bar{\pi}_0(s))$. The non-null probability functions $\bar{\pi}_0(s)$ exhibit similar patterns as those of $\mu_0(s)$ described in Setup I.
- **Setup III**: $\mu(s) = \gamma G(s)$, where $G(s)$ follows a Gaussian process with a constant mean $\bar{\mu}$ and the covariance function $k_\mu(s, s') = \sigma_\mu^2 \exp\{-\left(\|s - s'\|/\rho_\mu\right)^k\}$ with $k = 1$, $\sigma_\mu^2 = 3$ and $\rho_\mu = 0.3$. We set $\bar{\mu} = -2.5, -2, -1$ for the sparse, medium, and dense signal cases, respectively. Their non-null proportions are nearly 4%, 8%, and 25%.

The shapes of $\mu_0(s)$ in Setup I, the generated signals $\delta(s)$ in Setup II, and the simulated signals $G(s)$ in Setup III are depicted in Figures S.1(a)–S.1(c) of the supplement, respectively. For the magnitude γ, we considered $\gamma \in \{2, 3, 4\}$ in Setups I and III, and $\gamma \in \{1, 1.5, 2\}$ in Setup II. We generated the noise process $\epsilon(s)$ from a mean-zero Gaussian process with the covariance function $k_\epsilon(s, s'; r, k, \rho_\epsilon) = (1 - r)\mathbf{1}\{s = s'\} + r \exp\{-\left(\|s - s'\|/\rho_\epsilon\right)^k\}$. Here, r determines the relative percentage of nugget effect, ρ_ϵ measures the strength of dependency, and k controls the decay rate of dependence. We demonstrated three different degrees of spatial dependence through the following choices of (r, k, ρ_ϵ): (1) $r = 0.5$, $k = 1$, $\rho_\epsilon = 0.05$ (exponential kernel);

(2) $r = 0.8$, $k = 1$, $\rho_\epsilon = 0.1$ (exponential kernel); and (3) $r = 0.6$, $k = 2$, $\rho_\epsilon = 0.2$ (Gaussian kernel). The above (1)–(3) combinations of (r, k, ρ_ϵ) represent the weak, medium, and strong correlation among locations, respectively; see Figure S.3 of the supplement. Here we assume only one observation is available at each location and the covariance is known.

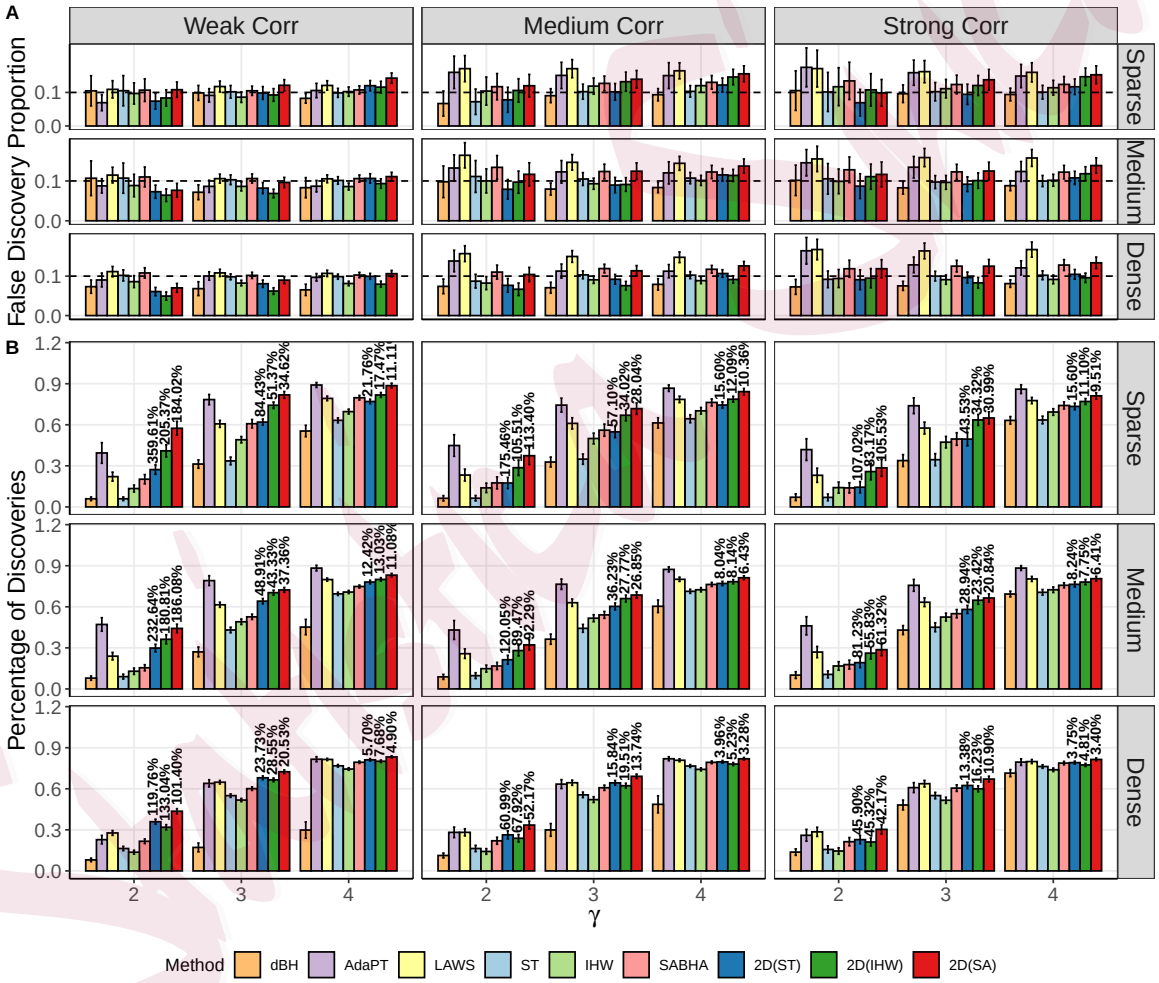


Figure 2: The mean and (1.96 multiple of) the standard error of FDP (Panel A) and power (Panel B) under Setup I with $\gamma \in \{2, 3, 4\}$. The percentages on the top of bars represent the power improvement of 2d procedures compared to their 1d counterparts.

We applied the competing methods to the generated datasets, and the empirical FDP and power under Setups I and III are summarized in Figures 2 and 3 based on 100 replicates.

Under Setup I, ST, IHW, SABHA, 2D (ST), and dBH controlled the FDR reasonably well across all cases. LAWS, AdaPT, 2D (IHW), and 2D (SA) were inflated for the medium and strong correlation cases, with LAWS being the worst. 2D (SA) and AdaPT were generally more powerful than the other methods, while dBH was quite conservative. As expected, the 2d procedures outperformed their 1d counterparts in terms of power. The results for Setup II were displayed in Figure S.4 of the supplement and generally similar to those in Setup I. Under Setup III, the empirical FDPs were close to zero, indicating all methods were conservative due to the composite null effect. The 2d-SMT procedures provided the highest power compared to the other approaches. Overall, the 2d-SMT procedures achieved remarkable improvements in power for either the weak correlation, the sparse signal, or the feeble magnitude cases. In Section S.I.2 of the supplement, we also performed simulations where the location sizes are increased to $m = 2000$. With larger location sizes, the 2d-SMT procedures exhibited more reliable FDR control under Setups I–II and demonstrated a significant improvement in power under Setup III. In Section S.I.3 of the supplement, we further considered a spatial process defined on a two-dimensional domain with multiple observations at each location to estimate the unknown covariance. The 2d-SMT procedures generally provided the best trade-off between FDR and power, especially when the correlation was weak. LAWS achieved the highest power at the cost of higher FDR. AdaPT provided reliable FDR control in all cases but their power were dominated by 2D (ST), 2D (IHW), and 2D (SA) for sparse signals and weak correlation. Additionally, 2d-SMT appeared to be robust to covariance function misspecifications, as we used an exponential kernel to estimate the covariance which was indeed generated from a Gaussian kernel.

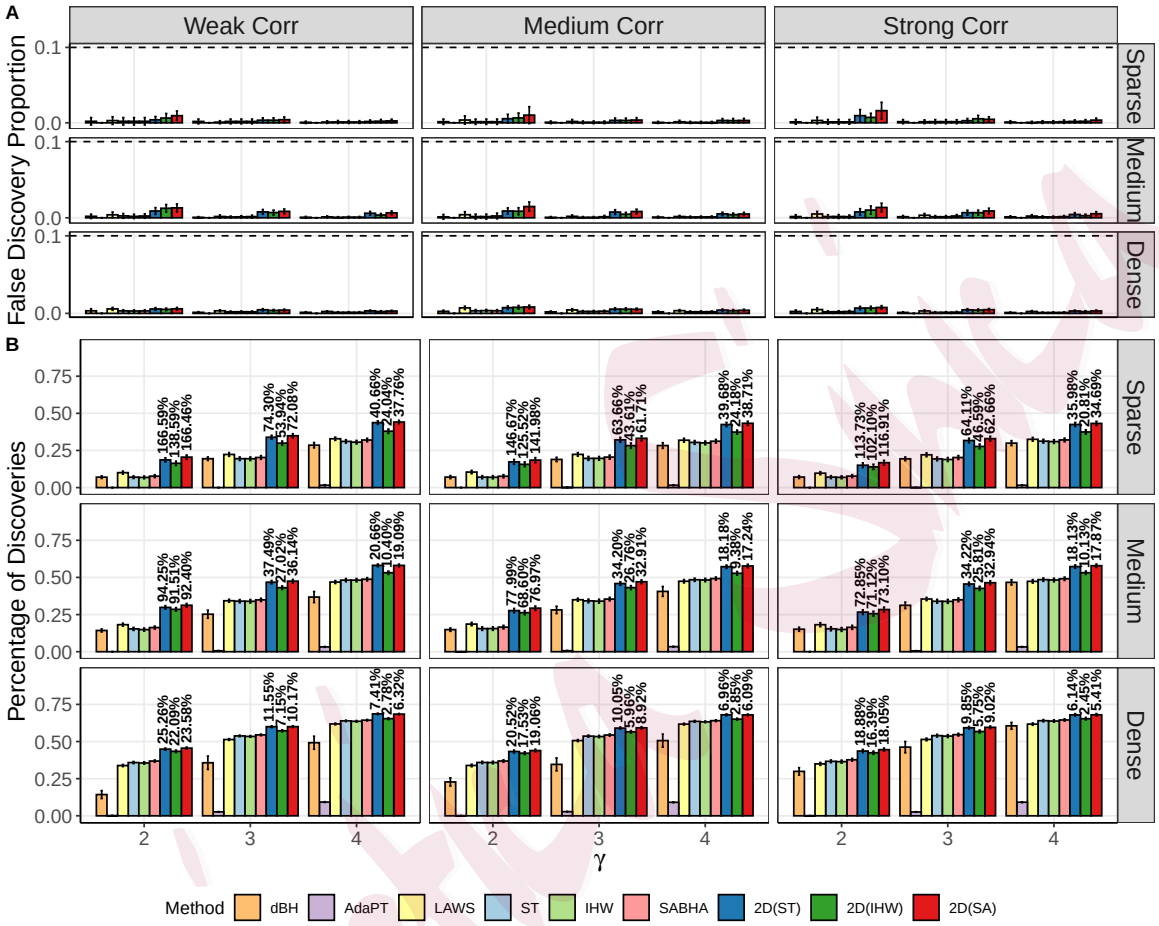


Figure 3: The mean and (1.96 multiple of) the standard error of FDP (Panel A) and power (Panel B) under Setup III with $\gamma \in \{2, 3, 4\}$. The percentages on the top of bars represent the power improvement of 2d procedures compared to their 1d counterparts.

6. Ozone Data Analysis

Ozone has double-edged effects on human health: ozone in the upper atmosphere (stratospheric ozone) shields humans from harmful ultraviolet (UV) radiation; while ozone at ground level (tropospheric ozone) triggers a variety of adverse health effects on human, sensitive vegetation and ecosystems (Weinhold, 2008; Liu et al., 2022). The US Environmental Protection Agency (EPA) formulates regulations to reduce tropospheric ozone levels in outdoor air. The majority of tropospheric ozone occurs through the reaction of nitrogen oxides (NO_x), carbon monoxide

(CO), and volatile organic compounds (VOCs) in the atmosphere when exposed to sunlight, particularly under the UV spectrum (Warneck, 2000).

We applied our proposed 2d procedures and their 1d alternatives to identify the locations where the decreasing trend is below a pre-specified level for the Contiguous United States from 2010 to 2021. The data were the annual averages of the fourth-highest daily maximum 8-hour ozone concentrations (see http://www.epa.gov/airexplorer/index.htm). To facilitate the analysis, we retained 697 stations (i) having a single site, (ii) having full records across the years, and (iii) being recorded by the World Geodetic System (WGS84). A regression model with mean-zero stationary Gaussian process of error has been widely used to analyze spatial data, e.g., for temperature (French and Sain, 2013) and for ozone (Sun et al., 2015). We followed the model in Sun et al. (2015) to obtain the test statistics. In particular, we first fitted the following linear model for each location

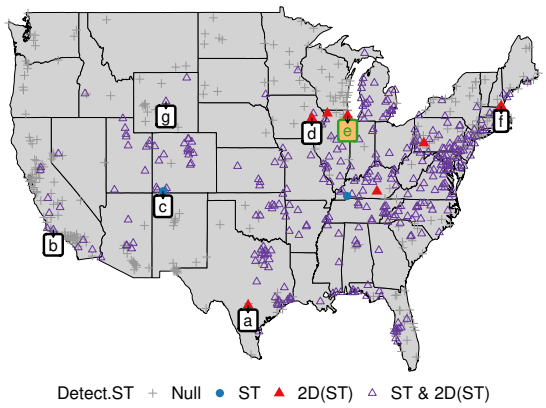
$$\mathbf{X}(s) = \mu_0(s) + \beta(s)\mathbf{t} + \sigma_\epsilon(s)\boldsymbol{\epsilon}(s), \tag{6.1}$$

where $\mathbf{X}(s) = (X_{2010}(s), \cdots, X_{2021}(s))^\top$ was the observed ozone level measured in parts per billion (ppb), $\mathbf{t} = (2010, 2011, \cdots, 2021)^\top$ was the predictor capturing the time trend, $\beta(s)$ was the slope at site s, $\boldsymbol{\epsilon}(s) = (\epsilon_{2010}(s), \cdots, \epsilon_{2021}(s))^\top$ was assumed to follow a mean-zero Gaussian process with the exponential kernel function $k_\epsilon(\cdot, \cdot; r, 1, \rho_\epsilon)$, and $\sigma_\epsilon(s)$ was the standard deviation of noise at site s. For each $\beta_0 \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, we were interested in testing whether the ozone level declined more than β_0 ppb per year at each site, i.e., $\mathcal{H}_{0,s} : \beta(s) \geq -\beta_0$ versus $\mathcal{H}_{1,s} : \beta(s) < -\beta_0$. We first conducted simple linear regression and obtained the OLS

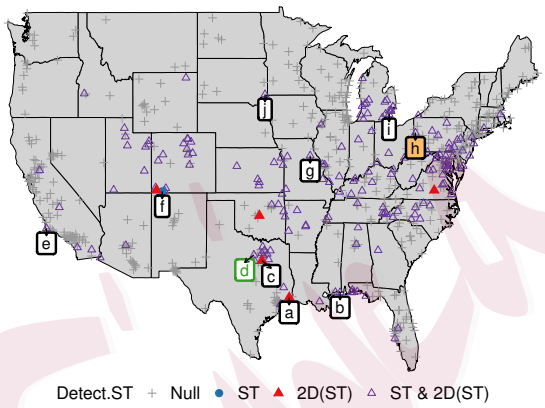
estimates of $(\mu_0(s), \beta(s), \sigma_\epsilon(s))$, denoted as $(\widehat{\mu}_0(s), \widehat{\beta}(s), \widehat{\sigma}_\epsilon(s))$. Then, we obtained $(\widehat{r}, \widehat{\rho}_\epsilon)$ by fitting the kernel function to the residuals $\widehat{\boldsymbol{\epsilon}}(s) := \{\mathbf{X}(s) - \widehat{\beta}(s)\mathbf{t} - \widehat{\mu}_0(s)\}/\widehat{\sigma}_\epsilon(s)$. Finally, the proposed 2d procedures are conducted with the target FDR level at 10% using the primary test statistic calculated as $\widehat{T}_2(s) = \{\widehat{\beta}(s) + \beta_0\}/\widehat{\sigma}_{\widehat{\beta}}(s)$, and the auxiliary test statistic given by $\widehat{T}_1(s) = \sum_{v\in\mathcal{N}(s)}\{\widehat{\beta}(v) + \beta_0\}/\widehat{\tau}(s)$, where $\mathcal{N}(s)$ was the set containing the two-nearest neighbors of s and $\widehat{\tau}(s) = \sum_{v,v'\in\mathcal{N}(s)} \widehat{\text{cov}}\{\widehat{\beta}(v), \widehat{\beta}(v')\}$.

Locations with significant ozone level decline. We applied 2D (ST), 2D (IHW), 2D (SA), and their 1d alternatives to identify the non-null locations. We trisected the ranges of the latitude and longitude, which divided the whole area into nine different regions and allocated each site a categorical variable; see Figure S.19 of the supplement for the division. Our analysis employed the categorical variable as the covariate in IHW and as the group indicator in SABHA. As shown in Figure 4, 2d procedures generally discovered more locations with significant decreasing ozone levels than their 1d counterparts did.

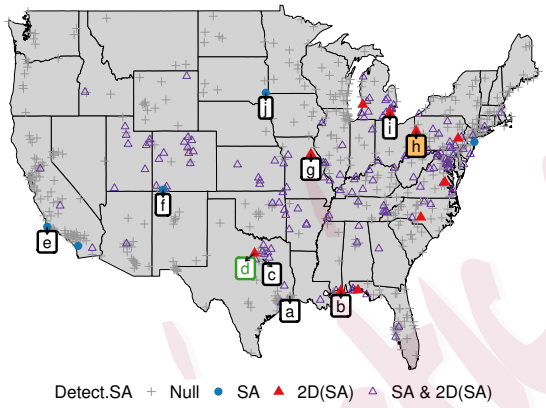
Ozone precursor. The EPA has been making efforts to reduce tropospheric ozone by executing air pollution control strategies, including formulating vehicle and transportation standards, regional haze and visibility rules, and regularly reviewing the National Ambient Air Quality Standards. The universal ozone precursors (NO_x, CO, and VOCs) first respond to these strategies and then influence the ozone levels. Indeed, some studies found the emissions of NO_2 and CO account for the increase in background ozone levels (Chin et al., 1994; Vingarzan, 2004; Han et al., 2011). Motivated by these findings, we collected the contemporaneous CO and NO_2 data from EPA and focused on the locations detected by either the 1d procedures or the 2d procedures but not both. We aimed to scrutinize the trends of CO and NO_2 at these locations



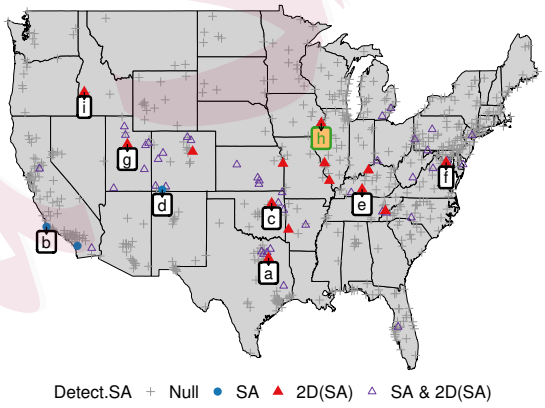
(a) $\beta_0 = 0.2$, ST versus 2D (ST)



(b) $\beta_0 = 0.3$, ST versus 2D (ST)



(c) $\beta_0 = 0.3$, SABHA versus 2D (SA)



(d) $\beta_0 = 0.5$, SABHA versus 2D (SA)

Figure 4: Results for ozone data analysis. The red triangles (▲), blue circles (●), purple triangles (△), and grey plus (+) signs represent the locations detected by the 2d procedure only, the 1d procedure only, both procedures, and neither one of the procedures, respectively. The labeled locations: (i) are detected by either 1d or 2d procedures but not both; and (ii) possess either CO or NO_2 records across 2010 to 2021. In each sub-figure, the location with an orange background/green typeface indicates the greatest decline in CO/NO_2 among the labeled locations.

and explain our findings.

To this end, we regressed the CO and NO_2 levels on $\mathbf{t}$ separately and recorded the slopes to understand the increasing/decreasing trends of the CO and NO_2 levels. We summarized the major findings in Figure 4 and Tables S.1–S.2 (of the supplement). First, the locations

detected only by the 2d procedures always included the ones with the most significant decline in the CO or NO_2 levels (i.e., the locations with the orange background or green typeface labels in Figure 4); see Table S.1 of the supplement. Second, the average decline (measured by the average of the standardized slopes) of the locations detected only by the 2d procedures was larger than that of the locations detected only by the 1d procedures. Take Figure 4(d) as an example, where we had NO_2 records at locations a, b, c, d, f, g, h and i, and CO records at locations b, c, e, f and h. SABHA detected the locations b and d, while 2D (SA) identified the locations a, c, e, f, g, h, and i. The average decline in the NO_2 levels was -3.63 for the locations detected by SABHA as compared to -4.68 for the locations detected by 2D (SA). As for CO, the average standardized slope was -0.96 for locations detected by SABHA in comparison to -1.40 for the locations identified by 2D (SA). In general, the CO and NO_2 levels tended to decrease more rapidly for those locations detected by 2D (SA) except for the case with $\beta_0 = 0.3$; see Table S.2 of the supplement.

Ozone data simulation. To further validate our findings and to demonstrate the effectiveness of the 2d procedures, we conducted a simulation where we generated data mimicking the structure of the original data. Specifically, we generated the ozone level data from 2010 to 2021 through (6.1) by setting $\beta(s) = \widehat{\beta}(s)$, $\mu(s) = \widehat{\mu}(s)$, $\sigma_\epsilon(s) = \widehat{\sigma}_\epsilon(s)$, $r = \widehat{r}$, and $\rho_\epsilon = \widehat{\rho}_\epsilon$. We processed the data and conducted multiple testing in the same way as discussed before. Table 1 shows that the 2d procedures achieved equal or higher power compared to the 1d alternatives while controlling FDR under 10%. To assess the robustness of our method, we followed Sun et al. (2015) to conduct additional simulations by using Gaussian kernel and empirical covariance matrix to generate synthetic data based on (6.1) and remained to use the exponential kernel

to analyze the synthetic data. The results are summarized in Section S.I.7 of the supplement, which is consistent with the findings reported in Table 1.

Table 1: Mean and standard deviation of FDPs and percentage of true discoveries (PTDs) for the simulated ozone data. The results are based on 100 simulation runs.

Criterion	β_0	ST	IHW	SABHA	2D (ST)	2D (IHW)	2D (SA)
FDP	0.5	0.021(0.028)	0.021(0.027)	0.038(0.036)	0.022(0.029)	0.021(0.028)	0.052(0.040)
	0.4	0.019(0.021)	0.018(0.019)	0.024(0.020)	0.021(0.022)	0.018(0.019)	0.028(0.020)
	0.3	0.014(0.015)	0.012(0.013)	0.012(0.012)	0.015(0.015)	0.012(0.013)	0.013(0.012)
	0.2	0.014(0.012)	0.008(0.009)	0.010(0.009)	0.015(0.013)	0.008(0.009)	0.010(0.009)
	0.1	0.014(0.012)	0.006(0.007)	0.007(0.007)	0.014(0.012)	0.007(0.007)	0.007(0.007)
PTD	0.5	0.287(0.123)	0.287(0.120)	0.416(0.150)	0.292(0.125)	0.287(0.120)	**0.474(0.151)**
	0.4	0.409(0.153)	0.393(0.139)	0.474(0.111)	0.414(0.154)	0.393(0.139)	**0.506(0.106)**
	0.3	0.545(0.150)	0.508(0.137)	0.520(0.125)	**0.552(0.151)**	0.507(0.137)	0.532(0.121)
	0.2	0.691(0.125)	0.627(0.112)	0.632(0.114)	**0.697(0.125)**	0.627(0.112)	0.638(0.113)
	0.1	0.796(0.090)	0.714(0.087)	0.721(0.088)	**0.799(0.090)**	0.714(0.087)	0.725(0.087)

7. Discussion

This paper proposes a new FDR-controlling procedure, 2d-SMT, to improve the signal detection power by incorporating the spatial information encoded in neighboring observations. It provides a unique perspective on utilizing spatial information, which is fundamentally different from the existing covariate and structural adaptive multiple testing procedures. The spatial information is gathered through an auxiliary statistic, which is used to screen out the noise. A primary statistic from the location of interest is then used to determine the existence of the signal. 2d-SMT is particularly effective when the signals exhibit in clusters. We demonstrate the usefulness of 2d-SMT through simulation studies and the analysis of an ozone data set. We recommend 2D (ST) among different variants of 2d-SMT because it provides the most stable FDR control performance in numerical experiments and enjoys proven asymptotic FDR control under weak spatial dependence.

To conclude, we point out a few future research directions. First, as discussed in Section 2.7, the 2d-SMT is flexible to combine with various weighted BH procedures. One challenge is, however, to establish a rigorous FDR control theory for the resulting weighted 2d-SMT procedures. Second, we use the κ-nearest neighbors to construct the auxiliary statistic in implementation. A more delicate strategy is to apply a weighting scheme to pool sufficient information from nearby locations. Third, extending the idea in 2d-SMT to other statistical problems, such as mediation analysis in causal inference, is of interest.

Supplementary Material

The online Supplementary Material contains our proofs of Theorems 1 and 2, additional numerical results, some discussions about the estimation for the covariance of noises, the details of Algorithm 1, and the address to download the reproducible code of this work.

Acknowledgments

The authors thank the editor, the associate editor, and anonymous reviewers for their helpful comments and suggestions. This research was supported by the Public Computing Cloud, Renmin University of China. The research of Zhang was partially supported by NSF DMS-2113359, NIH 1R01GM144351-01 and NIH 1R21HG011662.

References

Andrews, D. W. K. (1984). Non-strong mixing autoregressive processes. *J. Appl. Probab. 21*(4), 930–934.

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

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

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc., Ser. B, Methodol. 57*(1), 289–300.

Cai, T. T., W. Sun, and Y. Xia (2022). Laws: A locally adaptive weighting and screening approach to spatial multiple testing. *J. Amer. Statist. Assoc. 117*(539), 1370–1383.

Cao, H., J. Chen, and X. Zhang (2022). Optimal false discovery rate control for large scale multiple testing with auxiliary information. *Ann. Stat. 50*(2), 807–857.

Chin, M. et al. (1994). Relationship of ozone and carbon monoxide over north america. *J. Geophys. Res. Atmos. 99*(D7), 14565–14573.

Cliff, A. D. and J. K. Ord (1981). *Spatial Processes: Models and Applications*. London, UK: Pion Ltd.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley-Interscience. Part II: Lattice Data.

Davidson, J. (1994, 10). Near-epoch dependence. In *Stochastic Limit Theory: An Introduction for Econometricians*, pp. 241–278. Oxford University Press.

Ferreira, J. A. and A. H. Zwinderman (2006). On the benjamini–hochberg method. *Ann. Stat. 34*(4), 1827–1849, 23.

Fithian, W. and L. Lei (2022). Conditional calibration for false discovery rate control under dependence. *Ann. Stat. 50*(6), 3091–3118.

French, J. P. and S. R. Sain (2013). Spatio-temporal exceedance locations and confidence regions. *Ann. Appl. Stat. 7*(3), 1421–1449, 29.

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

Gorodetskii, V. (1978). On the strong mixing property for linear sequences. *Theory Probab. Its Appl. 22*(2), 411–413.

Han, S. et al. (2011). Analysis of the relationship between o3, no and no2 in tianjin, china. *Aerosol. Air Qual. Res. 11*(2),

128–139.

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

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

Huang, H.-C., N. Cressie, A. Zammit-Mangion, and G. Huang (2021). False discovery rates to detect signals from incomplete spatially aggregated data. *J. Comput. Graph. Stat. 30*(4), 1081–1094.

Ignatiadis, N. and W. Huber (2021). Covariate powered cross-weighted multiple testing. *J. R. Stat. Soc., Ser. B, Stat. Methodol. 83*(4), 720–751.

Ignatiadis, N., B. Klaus, J. B. Zaugg, and W. Huber (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods 13*(7), 577–580.

Jenish, N. (2012). Nonparametric spatial regression under near-epoch dependence. *J. Econom. 167*(1), 224–239.

Jenish, N. and I. R. Prucha (2012). On spatial processes and asymptotic inference under near-epoch dependence. *J. Econom. 170*(1), 178–190.

Katzfuss, M. and J. Guinness (2021). A general framework for vecchia approximations of gaussian processes. *Stat Sci 36*(1), 124–141.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat. 27*(4), 887–906.

Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *J. Amer. Statist. Assoc. 109*(506), 674–685.

Lei, J., M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman (2018). Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc. 113*(523), 1094–1111.

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

Liu, W. et al. (2022). Stratospheric ozone depletion and tropospheric ozone increases drive southern ocean interior warming. *Nat. Clim. Chang. 12*(4), 365–372.

Liu, W., Y. Ke, J. Liu, and R. Li (2022). Model-free feature screening and fdr control with knockoff features. *J. Amer. Statist. Assoc. 117*(537), 428–443.

Mardia, K. V. and R. J. Marshall (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika 71*(1), 135–146.

Sang, H. and J. Z. Huang (2012). A full scale approximation of covariance functions for large spatial data sets. *J. R. Stat. Soc., Ser. B, Stat. Methodol. 74*(1), 111–132.

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

Shen, X., H.-C. Huang, and N. Cressie (2002). Nonparametric hypothesis testing for a spatial signal. *J. Amer. Statist. Assoc. 97*(460), 1122–1140.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc., Ser. B, Methodol. 64*(3), 479–498.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc., Ser. B, Stat. Methodol. 66*(1), 187–205.

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

Tansey, W., O. Koyejo, R. A. Poldrack, and J. G. Scott (2018). False discovery rate smoothing. *J. Amer. Statist.*

Assoc. 113(523), 1156–1171.

Vingarzan, R. (2004). A review of surface ozone background levels and trends. *Atmospheric Environ. 38*(21), 3431–3442.

Warneck, P. (2000). Chapter 5 ozone in the troposphere. In P. Warneck (Ed.), *Chemistry of the Natural Atmosphere*, Volume 71 of *International Geophysics*, pp. 211–263. Academic Press.

Wasserman, L. and K. Roeder (2009). High dimensional variable selection. *Ann. Stat. 37*(5A), 2178–2201.

Weinhold, B. (2008). Ozone nation: Epa standard panned by the people. *Environ. Health Perspect. 116*(7), A302–A305.

Yi, S., X. Zhang, L. Yang, J. Huang, Y. Liu, C. Wang, D. J. Schaid, and J. Chen (2021). 2dfdr: A new approach to confounder adjustment substantially increases detection power in omics association studies. *Genome Biol. 22*(1), 208.

Yun, S., X. Zhang, and B. Li (2022). Detection of local differences in spatial characteristics between two spatiotemporal random fields. *J. Amer. Statist. Assoc. 117*(537), 291–306.

Zhang, C.-H. (2009). Generalized maximum likelihood estimation of normal mixture densities. *Stat. Sin. 19*(3), 1297–1318.

Zhang, X. and J. Chen (2022). Covariate adaptive false discovery rate control with applications to omics-wide multiple testing. *J. Amer. Statist. Assoc. 117*(537), 411–427.

Linsui Deng (denglinsui@ruc.edu.cn)

Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China.

Kejun He (kejunhe@ruc.edu.cn)

Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China.

Xianyang Zhang (zhangxiany@stat.tamu.edu)

Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.