Statistica Sinica **35** (2025), 2233-2257
doi:https://doi.org/10.5705/ss.202022.0314

NAPA: NEIGHBORHOOD-ASSISTED AND POSTERIOR-ADJUSTED TWO-SAMPLE INFERENCE

Li Ma[1], Yin Xia*[1] and Lexin Li[2]

[1]*Fudan University and* [2]*University of California at Berkeley*

Abstract: Two-sample multiple testing problems of sparse spatial data are frequently arising in a variety of scientific applications. In this article, we develop a novel neighborhood-assisted and posterior-adjusted (NAPA) approach to incorporate both the spatial smoothness and sparsity type side information to improve the power of the test while controlling the false discovery of multiple testing. We translate the side information into a set of weights to adjust the p-values, where the spatial pattern is encoded by the ordering of the locations, and the sparsity structure is encoded by a set of auxiliary covariates. We establish the theoretical properties of the proposed test, including the guaranteed power improvement over some state-of-the-art alternative tests, and the asymptotic false discovery control. We demonstrate the efficacy of the test through intensive simulations and two neuroimaging applications.

Key words and phrases: False discovery rate, multiple testing, side information, sparsity, spatial smoothness, weighted p-values.

1. Introduction

Two-sample hypothesis testing of sparse spatial data is a fundamental problem in a wide variety of scientific applications. It manifests in numerous forms. One example is to compare the cerebral white matter tracts between multiple sclerosis (MS) patients and healthy controls (Goldsmith et al., 2011). The data records the fractional anisotropy measure along the right corticospinal tract, and takes the form of one-dimensional (1D) function. The scientific interest is to compare two sets of fractional anisotropy profiles and locate the tract regions that distinguish cases from controls. Another example is to compare the brain grey matter cortical thickness between subjects diagnosed with attention deficit hyperactivity disorder (ADHD) and typically developing controls (Bellec et al., 2017). The data records the volume of grey matter at different brain locations in a three-dimensional (3D) space. The scientific interest is to compare two sets of brain structural images and identify differentiating brain regions. In addition to these examples, similar problems arise in many other applications, for instance, astronomical surveys (Czakon et al., 2009), disease mapping (Sun et al., 2000), ecology (Bini et al., 2009), and genomics (Sun and Wei, 2011).

*Corresponding author. E-mail: xiayin@fudan.edu.cn

All these examples can be formulated as a two-sample testing problem, where the data reside in some spatial domain. More specifically, let $\mathbb{S} \subset \mathbb{R}^b$ denote a b-dimensional spatial domain, where $b = 1, 2, 3, \ldots$. Let $\mathcal{S} \subset \mathbb{S}$ denote a finite, regular lattice in $\mathbb{S}$, and $\boldsymbol{s} \in \mathcal{S}$ the coordinate of the location. For the 1D MS example, $b = 1$ and $\boldsymbol{s}$ is a scalar, whereas for the 3D ADHD example, $b = 3$ and $\boldsymbol{s}$ is a three-variate coordinate. Later in our theoretical analysis, we consider the infill-asymptotic framework (Stein, 1999) and assume $\mathcal{S} \to \mathbb{S}$. Suppose the data $Y_d(\boldsymbol{s}) \in \mathbb{R}$ is observed at every location $\boldsymbol{s} \in \mathcal{S}$ for two groups $d = 1, 2$, and write $\boldsymbol{Y}_d = \{Y_d(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}$. Suppose $\boldsymbol{Y}_d$ follows a probability distribution $\mathcal{P}_{\boldsymbol{\beta}_d, \boldsymbol{\eta}_d}$, where $\boldsymbol{\beta}_d = \{\beta_d(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}$ denotes the parameters of interest, and $\boldsymbol{\eta}_d$ collects all the nuisance parameters. Suppose we observe two sets of independent samples, $\{\boldsymbol{Y}_{i,d}\}_{i=1}^{n_d}$, where n_d is the sample size for group d, $d = 1, 2$. Our goal is to carry out multiple hypothesis testing given the observed data,

$$H_0(\boldsymbol{s}) : \ \beta_1(\boldsymbol{s}) = \beta_2(\boldsymbol{s}) \quad \text{versus} \quad H_1(\boldsymbol{s}) : \ \beta_1(\boldsymbol{s}) \neq \beta_2(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{S}. \tag{1.1}$$

We call the location $\boldsymbol{s}$ a signal location or a non-null location if $H_1(\boldsymbol{s})$ holds, and call it a null location otherwise. We comment that (1.1) covers a range of testing problems. In this article, we mostly illustrate with the problem of comparing two multivariate means (Cai, Liu and Xia, 2014; Xu et al., 2016), where $\boldsymbol{\beta}_d$ represents the mean of $\boldsymbol{Y}_d$. Meanwhile, our proposal is equally applicable to the problems of comparing large correlation or covariance structures (Li and Chen, 2012; Cai, Liu and Xia, 2013; Cai and Liu, 2016; Zheng et al., 2019), detecting differential networks (Xia, Cai and Cai, 2015; Chen et al., 2015), or identifying gene-environment interactions (Caspi and Moffitt, 2006; Xia, Cai and Cai, 2018). Next, we recognize that, since the data resides in a spatial domain, there exists naturally some form of spatial smoothness in the data. Additionally, there is likely sparsity in (1.1), in that the alternative hypothesis holds at only a small subset of locations $\boldsymbol{s}$ in the entire $\mathcal{S}$. Sparsity is a common phenomenon in scientific applications and is frequently encountered in multiple testing. The goal of this article is to effectively incorporate both spatial smoothness and sparsity information into the multiple testing problem (1.1), while controlling the false discovery rate (FDR) and improving the power of the test.

Both smoothness and sparsity can be viewed as some forms of side information, and there have been a large number of proposals to incorporate side information in multiple testing; see Cai and Sun (2017) for a review and references therein. In this article, we focus on the strategy of p-value weighting, which has been widely used for FDR control and power enhancement (Benjamini and Hochberg, 1997; Storey, 2002; Genovese, Roeder and Wasserman, 2006; Roeder and Wasserman, 2009). In particular, Hu, Zhao and Zhou (2010) adopted the prior knowledge that the hypotheses belong to a known number of groups, and weighed the p-values for the hypotheses in each group by $\pi_g/(1 - \pi_g)$, where

π_g is the non-null proportion for the group that needs to be estimated. Zhang, Fan and Yu (2011) proposed to smooth and aggregate the p-values in a local neighborhood to accommodate the spatial information of the neighboring p-values. Liu (2014) utilized the sparsity information in the mean vectors and developed an uncorrelated screening-based FDR control procedure. Li and Barber (2019) weighed the p-values by a heterogeneous weight $1/\{1-\hat{\pi}(s)\}$, where $\hat{\pi}(s)$ is the estimated probability of a hypothesis being a non-null. Ignatiadis et al. (2016), Lei and Fithian (2018), and Lei, Ramdas and Fithian (2021) incorporated generic side information through secondary data or external covariates, and constructed the p-value-based thresholding procedures adaptively. Xia, Cai and Sun (2020) proposed the grouping, adjusting, and pooling (GAP) method that exploits the sparsity information, where they adaptively constructed a set of auxiliary statistics, based on which they identified clusters of hypotheses, then weighed the p-values by some discrete group-wise weights. Cai, Sun and Xia (2022) proposed the locally adaptive weighting and screening (LAWS) method that utilizes the smoothness information, where they constructed a set of robust and structure-adaptive weights based on the estimated local sparsity levels, and then weighed the p-values by these continuous weights.

In this article, we propose a **n**eighborhood-**a**ssisted and **p**osterior-**a**djusted test (NAPA) for the two-sample multiple testing problem (1.1), whereas we aim to control the FDR and improve the power by incorporating both spatial and sparsity information. Our key idea is to translate the spatial and sparsity information to construct a set of weights to adjust the p-values, where the spatial pattern is encoded by the ordering of the locations, and the sparsity structure is encoded by a set of auxiliary statistics. Specifically, we first construct a set of test statistics $\{T(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}$, which contain useful information about the true signal locations $\theta(\boldsymbol{s}) = \mathsf{I}\{\beta_1(\boldsymbol{s}) \neq \beta_2(\boldsymbol{s})\}$, and $\mathsf{I}(\cdot)$ is the indicator function. We then construct a set of auxiliary statistics $\{U(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}$, which take the form of $\beta_1(\boldsymbol{s}) + \kappa(\boldsymbol{s})\beta_2(\boldsymbol{s})$ for some weight function $\kappa(\boldsymbol{s})$, and contain useful information about $\gamma(\boldsymbol{s}) = \mathsf{I}\{\beta_1(\boldsymbol{s}) \neq 0 \text{ or } \beta_2(\boldsymbol{s}) \neq 0\}$. We note that $\gamma(\boldsymbol{s}) = 0$ implies $\theta(\boldsymbol{s}) = 0$. This implication means that, if both $\beta_1(\boldsymbol{s})$ and $\beta_2(\boldsymbol{s})$ are zero, then the null hypothesis must be true, and at least one of $\beta_1(\boldsymbol{s})$ and $\beta_2(\boldsymbol{s})$ must be nonzero for the alternative hypothesis to hold. In other words, the locations where $\gamma(\boldsymbol{s}) = 1$ capture useful information about the locations of the true signals. Moreover, the location $\boldsymbol{s}$ itself contains useful information about the smoothness embedded in the data. Therefore, we propose the posterior non-null probability, $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\} = \Pr\{\theta(\boldsymbol{s}) = 1 | U(\boldsymbol{s})\}$, and weigh the p-value by the weight, $w\{\boldsymbol{s}, U(\boldsymbol{s})\} = \pi\{\boldsymbol{s}, U(\boldsymbol{s})\}/[1 - \pi\{\boldsymbol{s}, U(\boldsymbol{s})\}]$, which integrates both the spatial information encoded in $\boldsymbol{s}$ and the sparsity information encoded in $U(\boldsymbol{s})$. Recognizing that $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ is a continuous function of $\boldsymbol{s}$ and $U(\boldsymbol{s})$, we propose to estimate this posterior weight by pooling the information from its neighbors using a smoothing kernel approach. The neighborhood is defined through both

$\boldsymbol{s}$ and $U(\boldsymbol{s})$, in that a close spatial location in $\boldsymbol{s}$ and a similar value of $U(\boldsymbol{s})$ both indicate a similar likelihood for the hypothesis at $\boldsymbol{s}$ to be null or alternative. Finally, given the estimated posterior weights, we choose a proper threshold to adjust the multiplicity for FDR control, and further show that the new test enjoys some guaranteed power improvement.

Our proposed NAPA test is built on the recent proposals of multiple testing utilizing the side information. Particularly, our test combines the ideas of the GAP method of Xia, Cai and Sun (2020) and the LAWS method of Cai, Sun and Xia (2022). On the other hand, it is far from a straightforward extension, and is substantially different from both GAP and LAWS, as well as a simple combination of the two. More specifically, GAP translates the sparsity information embedded in the auxiliary statistic into several discrete groups; then constructs a weight for each group. However, the number of groups is generally unknown, and searching among all possible groupings is computationally expensive. Besides, such a discretization may lead to potential power loss. By contrast, our method does not weigh the p-values by groups, but instead weighs in a continuous fashion, which is computationally more efficient and can further improve the power. This new weighting strategy, nevertheless, induces new challenges. A key result in GAP, such that the grouping using auxiliary statistic does not distort the null distribution of the p-values, is no longer sufficient in our setting. To employ the auxiliary sequence continuously and to ensure the statistical properties, we derive a new conditional normal approximation, a result that is not available in the literature. Next, LAWS focuses on the one-sample testing problem and assumes that $\theta(\boldsymbol{s})$ is fully determined by the location $\boldsymbol{s}$, but ignores additional sparsity information. By contrast, we introduce an auxiliary variable $U(\boldsymbol{s})$ that is constructed adaptively from the data, and explore the posterior binomial variable $\theta(\boldsymbol{s})|U(\boldsymbol{s})$. The new approach, nevertheless, leads to a more involved theoretical development. This is because we have to tackle the correlations among the spatial locations and the covariates, as well as the dependencies among the primary and auxiliary statistics, and we derive new and sophisticated technical tools to address those challenges. Finally, our proposed test is far from a simple combination of GAP and LAWS. In Section 5, we carry out a simulation experiment and show that our new test is much more powerful than naively combining GAP and LAWS. In summary, we believe our proposal fills an important gap in two-sample inference that utilizes both spatial and sparsity information, and thus helps address a range of scientific questions in areas such as neuroimaging analyses that involve spatial data. Moreover, we develop new technical tools that are potentially useful for general inference problems of complex dependent data.

The rest of the article is organized as follows. Our proposed weight $w\{\boldsymbol{s}, U(\boldsymbol{s})\}$ and the posterior probability $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ hinge on both the location $\boldsymbol{s}$ and the auxiliary statistic $U(\boldsymbol{s})$. In Section 2, we first study this weighting scheme in the oracle setting, where $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ is known, the p-value obtained from the

test statistic $T(\boldsymbol{s})$ is uniformly distributed, and $U(\boldsymbol{s})$ is independent of $T(\boldsymbol{s})$ under the null. We establish the guaranteed power gain under this setting. In Section 3, we estimate the weight using smoothing kernel, and develop a multiple testing procedure based on the weighted p-values. We further illustrate the construction of the test and auxiliary statistics $\{T(\boldsymbol{s}), U(\boldsymbol{s})\}$ in a multivariate mean comparison problem. In Section 4, we show that our weight estimator is consistent, and the proposed test controls the FDR asymptotically. We use the mean comparison as an example and show that the p-value from $T(\boldsymbol{s})$ is asymptotically uniform, and $T(\boldsymbol{s})$ and $U(\boldsymbol{s})$ are asymptotically independent under the null. In Section 5, we study the finite-sample performance, and in Section 6, we illustrate with two neuroimaging data examples. We relegate all technical proofs and additional simulations to the online Supplementary Material.

2. Oracle Weighting and Power Gain

In this section, we first derive and motivate our proposed weight function. We then study the power gain under the oracle setting.

2.1. Posterior-adjusted weighting

Define the local non-null probability $\pi(\boldsymbol{s})$ at location $\boldsymbol{s}$ as $\pi(\boldsymbol{s}) = \Pr\{\theta(\boldsymbol{s}) = 1\}$, the local posterior non-null probability $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ at $\boldsymbol{s}$, and the weight function as,

$$\pi\{\boldsymbol{s}, U(\boldsymbol{s})\} = \Pr\{\theta(\boldsymbol{s}) = 1 | U(\boldsymbol{s})\} = \frac{\pi(\boldsymbol{s})\, q_1\{U(\boldsymbol{s})|\boldsymbol{s}\}}{\{1-\pi(\boldsymbol{s})\}\, q_0\{U(\boldsymbol{s})|\boldsymbol{s}\} + \pi(\boldsymbol{s})\, q_1\{U(\boldsymbol{s})|\boldsymbol{s}\}},$$
$$w\{\boldsymbol{s}, U(\boldsymbol{s})\} = \frac{\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}}{1-\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}}, \tag{2.1}$$

where $q_0(\cdot|\boldsymbol{s})$ and $q_1(\cdot|\boldsymbol{s})$ are the null $\theta(\boldsymbol{s}) = 0$ and non-null $\theta(\boldsymbol{s}) = 1$ density functions of $U(\boldsymbol{s})$, respectively. We call $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ a posterior probability as it is conditioning on the auxiliary variable $U(\boldsymbol{s})$. We propose to weigh the p-value $p(\boldsymbol{s})$ as,

$$p_w(\boldsymbol{s}) = \frac{p(\boldsymbol{s})}{w\{\boldsymbol{s}, U(\boldsymbol{s})\}} = \frac{1-\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}}{\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}} p(\boldsymbol{s}). \tag{2.2}$$

2.2. Theoretical improvement

We next compare our testing method with LAWS of Cai, Sun and Xia (2022) under the oracle setting. The two methods mainly differ in the weight used for the p-value, i.e., we use $w\{\boldsymbol{s}, U(\boldsymbol{s})\}$ in (2.1), whereas LAWS uses $w(\boldsymbol{s}) = \pi(\boldsymbol{s})/\{1-\pi(\boldsymbol{s})\}$. Under some mild conditions, we show our method that incorporates both spatial and sparsity information is guaranteed to improve the power over LAWS that only utilizes the spatial information. We do not analytically compare with GAP of Xia, Cai and Sun (2020) here, because GAP uses a discrete weighting

scheme and is not directly comparable. Nevertheless, we numerically compare with LAWS and GAP in Section 5.2.

We first formally define the evaluation criteria in terms of false discovery and power. Consider a sequence of weighted p-values $\{p_w(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}$ and a given threshold t. When there is no weighting, we set all weights equal to one. Let $\delta^w(\boldsymbol{s}, t) = \mathsf{I}\{p_w(\boldsymbol{s}) \leq t\}$ denote the decision rule for the hypotheses in (1.1), in that $\delta^w(\boldsymbol{s}, t) = 1$ if we reject the null, and $\delta^w(\boldsymbol{s}, t) = 0$ otherwise. Let $\boldsymbol{\delta}_t^w = \{\delta^w(\boldsymbol{s}, t) : \boldsymbol{s} \in \mathcal{S}\}$ be the collection of all decision rules for $\boldsymbol{s} \in \mathcal{S}$ under the threshold t. We define the FDR and the marginal FDR (mFDR) of the test $\boldsymbol{\delta}_t^w$ as,

$$\mathrm{FDR}(\boldsymbol{\delta}_t^w) = \mathsf{E}\left(\frac{\sum_{\boldsymbol{s}\in\mathcal{S}}\left[\{1-\theta(\boldsymbol{s})\}\delta^w(\boldsymbol{s},t)\right]}{\max\left\{\sum_{\boldsymbol{s}\in\mathcal{S}}\delta^w(\boldsymbol{s},t), 1\right\}}\right),$$
$$\mathrm{mFDR}(\boldsymbol{\delta}_t^w) = \frac{\mathsf{E}\left(\sum_{\boldsymbol{s}\in\mathcal{S}}\left[\{1-\theta(\boldsymbol{s})\}\delta^w(\boldsymbol{s},t)\right]\right)}{\mathsf{E}\left\{\sum_{\boldsymbol{s}\in\mathcal{S}}\delta^w(\boldsymbol{s},t)\right\}}.$$

Cai, Sun and Xia (2022) showed that $\mathrm{FDR}(\boldsymbol{\delta}_t^w) = \mathrm{mFDR}(\boldsymbol{\delta}_t^w) + o(1)$ under some mild conditions. Therefore, we can use the leading term $\mathrm{mFDR}(\boldsymbol{\delta}_t^w)$ to approximate $\mathrm{FDR}(\boldsymbol{\delta}_t^w)$ asymptotically. In addition, we define the power of the test as

$$\Psi(\boldsymbol{\delta}_t^w) = \mathsf{E}\left[\sum_{\boldsymbol{s}\in\mathcal{S}}\{\theta(\boldsymbol{s})\delta^w(\boldsymbol{s},t)\}\right].$$

We further define the oracle threshold values of the two methods as,

$$t_{\mathrm{LAWS}} = \sup_t\left\{0 \leq t \leq 1 : \mathrm{mFDR}(\boldsymbol{\delta}_t^{\mathrm{LAWS}}) \leq \alpha\right\},$$
$$t_{\mathrm{NAPA}} = \sup_t\left\{0 \leq t \leq 1 : \mathrm{mFDR}(\boldsymbol{\delta}_t^{\mathrm{NAPA}}) \leq \alpha\right\},$$

where $\boldsymbol{\delta}_t^{\mathrm{LAWS}}$ and $\boldsymbol{\delta}_t^{\mathrm{NAPA}}$ represent the decision rules based on the LAWS and NAPA weights, respectively, and α is a pre-specified significance level.

Let F_1 denote the non-null conditional cumulative distribution function (CDF) of the unweighted p-value, and F_1' its first derivative. Let $g_{\boldsymbol{s}}(x) = xF_1\{xt_{\mathrm{LAWS}}/(1-x)|\boldsymbol{s}\}$ and $g_{\boldsymbol{s}}'(x)$ its first derivative. Let $\pi_1 = \mathrm{Card}\left(\{\boldsymbol{s} \in \mathcal{S} : \pi(\boldsymbol{s}) > 0.5\}\right)/\mathrm{Card}(\mathcal{S})$, where $\mathrm{Card}(\cdot)$ is the cardinality of a set. The next theorem characterizes the theoretical gain of our NAPA method compared to LAWS.

Theorem 1. *For each $\boldsymbol{s} \in \mathcal{S}$, suppose $U(\boldsymbol{s})$ and $T(\boldsymbol{s})$ are independent under the alternative. Suppose $F_1(y|\boldsymbol{s})$ is concave in y, $g_{\boldsymbol{s}}(x)$ is convex in x for $x \leq 1/(1+t_{LAWS})$, and $yF_1'(y|\boldsymbol{s})$ is non-decreasing in y. Suppose $\pi(\boldsymbol{s}) \in [\zeta, 1-\zeta]$ for some small constant $\zeta > 0$. If there exists some constant $0 < \varrho \leq 1$, such that $\pi_1 \leq \varrho$ and $\{g_{\boldsymbol{s}}'(1-\zeta) - g_{\boldsymbol{s}}'(0.5)\}/\{1 - g_{\boldsymbol{s}}'(0.5)\} \leq 1/\varrho$, then*

$$\begin{aligned}\mathrm{mFDR}\left(\boldsymbol{\delta}_{t_{LAWS}}^{NAPA}\right) &\le \mathrm{mFDR}\left(\boldsymbol{\delta}_{t_{LAWS}}^{LAWS}\right) \le \alpha,\\ \Psi\left(\boldsymbol{\delta}_{t_{NAPA}}^{NAPA}\right) &\ge \Psi\left(\boldsymbol{\delta}_{t_{LAWS}}^{NAPA}\right) \ge \Psi\left(\boldsymbol{\delta}_{t_{LAWS}}^{LAWS}\right).\end{aligned}$$

We make a few remarks. First, Theorem 1 shows that, when using the oracle threshold t_{LAWS}, our NAPA method achieves an mFDR that is no greater than that of LAWS and a power that is no smaller than that of LAWS. Based on this result and the construction of the oracle threshold, t_{NAPA} is thus no smaller than t_{LAWS}. Therefore, using the threshold t_{NAPA} leads to an additional power gain of NAPA, and thus it establishes the guaranteed power gain of NAPA over LAWS. Second, Cai, Sun and Xia (2022) showed that LAWS dominates the classical Benjamini and Hochberg (BH) method (Benjamini and Hochberg, 1995), and therefore, NAPA dominates BH too. Third, this theorem does *not* require any spatial dependence condition, as the power of the test is mainly reflected through the ranking of weighted p-values and the testing errors are evaluated by marginal FDR under the oracle setting when the weights are known. However, spatial dependence plays a crucial role in subsequent FDP and FDR control, as we show later in Section 4. Finally, the conditions in Theorem 1 are all reasonably mild. In particular, as we show later when constructing the test statistic $T(\boldsymbol{s})$ and the auxiliary variable $U(\boldsymbol{s})$, they are asymptotically independent under both the null and alternative. The four distributional conditions can be easily verified for commonly used p-value distributions (Sellke, Bayarri and Berger, 2001; Held and Ott, 2018; Zhang and Chen, 2022). Among them, similar concavity and convexity conditions have been commonly imposed in the FDR literature (Storey, 2002; Genovese, Roeder and Wasserman, 2006; Hu, Zhao and Zhou, 2010; Xia, Cai and Sun, 2020; Cai, Sun and Xia, 2022). The condition $\pi_1 \le \varrho$ aligns with the sparsity framework we consider. We also give more discussion of the conditions in Theorem 1 in Section S3.5 of the online Supplementary Material.

2.3. An illustration

Next, we consider a covariate-adjusted mixture model to offer an intuitive explanation of both our proposed weight, and the comparison with the case without using $U(\boldsymbol{s})$. We clarify that we do *not* impose such a model in our test, but only use it for illustration. Similar model has been frequently studied in the two-sample inference literature (Efron et al., 2001; Newton, 2004; Sun and Cai, 2007; Efron, 2008).

$$\begin{aligned}T(\boldsymbol{s})|U(\boldsymbol{s}) &\sim f\{t|\boldsymbol{s},U(\boldsymbol{s})\}\\ &= [1-\pi\{\boldsymbol{s},U(\boldsymbol{s})\}]f_0\{t|\boldsymbol{s},U(\boldsymbol{s})\} + \pi\{\boldsymbol{s},U(\boldsymbol{s})\}f_1\{t|\boldsymbol{s},U(\boldsymbol{s})\},\end{aligned}$$

where f_0 and f_1 are the null and non-null conditional density function of $T(\boldsymbol{s})$ given $U(\boldsymbol{s})$, respectively. For the oracle setting, we have $f_0\{t|\boldsymbol{s},U(\boldsymbol{s})\} = f_0(t|\boldsymbol{s})$. Following Cai, Sun and Wang (2019), when the tests are independent, the optimal

test threshold for the above model is based on the ranking of the conditional local false discovery rate,

$$\mathrm{CLfdr}\{t|\boldsymbol{s}, U(\boldsymbol{s})\} = \Pr\{\theta(\boldsymbol{s}) = 0|t, \boldsymbol{s}, U(\boldsymbol{s})\} \propto \frac{1-\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}}{\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}} \times \frac{f_0(t|\boldsymbol{s})}{f_1\{t|\boldsymbol{s}, U(\boldsymbol{s})\}}.$$

If we ignore $U(\boldsymbol{s})$, the CLfdr reduces to the local false discovery rate,

$$\mathrm{Lfdr}(t|\boldsymbol{s}) = \Pr\{\theta(\boldsymbol{s}) = 0|t, \boldsymbol{s}\} \propto \frac{1-\pi(\boldsymbol{s})}{\pi(\boldsymbol{s})} \times \frac{f_0(t|\boldsymbol{s})}{f_1(t|\boldsymbol{s})}.$$

When the tests are dependent, we consider weighing the p-value to approximate CLfdr or Lfdr. We observe that, in both CLfdr and Lfdr, the first term reflects the sparsity structure, whereas the second term reflects the strength of evidence against the null. However, the second term is usually difficult to estimate, so we replace it with the p-value. This essentially leads to our proposed weight $w\{\boldsymbol{s}, U(\boldsymbol{s})\}$ in (2.1), and the weight $w(\boldsymbol{s}) = \pi(\boldsymbol{s})/\{1-\pi(\boldsymbol{s})\}$ used in LAWS.

We also observe that, when comparing the two weights, a large $|U(\boldsymbol{s})|$ usually provides a strong evidence that $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ is larger than $\pi(\boldsymbol{s})$. This can be seen through the calculation, $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\} = [\pi(\boldsymbol{s})q_1\{U(\boldsymbol{s})|\boldsymbol{s}\}]/[\{1-\pi(\boldsymbol{s})\}$ $q_0\{U(\boldsymbol{s})|\boldsymbol{s}\} + \pi(\boldsymbol{s})q_1\{U(\boldsymbol{s})|\boldsymbol{s}\}]$, where $q_0(\cdot|\boldsymbol{s})$ and $q_1(\cdot|\boldsymbol{s})$ are the null and non-null density functions of $U(\boldsymbol{s})$. Therefore, by incorporating $U(\boldsymbol{s})$, we may obtain a smaller weighted p-value for the alternatives and a better ranking of the tests, compared to LAWS that does not utilize $U(\boldsymbol{s})$. Theorem 1 then formally justifies such an intuition.

3. Two-sample Testing Procedure

In this section, we first discuss how to estimate the posterior non-null probability $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$. We then develop a general multiple testing procedure based on the weighted p-values. Finally, we illustrate the testing procedure with the problem of comparing two multivariate means, with a concrete construction of the test and auxiliary statistics $\{T(\boldsymbol{s}), U(\boldsymbol{s})\}$.

3.1. Neighborhood-assisted weight estimation

Recognizing that it is rather difficult to directly estimate the posterior non-null probability $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ in (2.1), we first propose an intermediate quantity $\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$. We show that $\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ provides a good approximation of $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$, and the weight constructed based on $\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ has the desired theoretical guarantees. A similar approximation has also been used in Schweder and Spjotvoll (1982), Storey (2002), and Cai, Sun and Xia (2022). Specifically, define

$$\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\} = 1 - \frac{\Pr\left\{p(\boldsymbol{s}) > \tau | U(\boldsymbol{s})\right\}}{1-\tau}, \quad \text{for some} \quad 0 < \tau < 1. \tag{3.1}$$

To justify the use of $\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$, consider

$$\begin{aligned}\Pr\{p(\boldsymbol{s}) \le t | U(\boldsymbol{s})\} &= [1 - \pi\{\boldsymbol{s}, U(\boldsymbol{s})\}]F_0\{t|\boldsymbol{s}, U(\boldsymbol{s})\} + \pi\{\boldsymbol{s}, U(\boldsymbol{s})\}F_1\{t|\boldsymbol{s}, U(\boldsymbol{s})\} \\ &\approx [1 - \pi\{\boldsymbol{s}, U(\boldsymbol{s})\}]t + \pi\{\boldsymbol{s}, U(\boldsymbol{s})\}F_1\{t|\boldsymbol{s}, U(\boldsymbol{s})\},\end{aligned}$$

where $t \in [0,1]$, F_0, F_1 are respectively the null and non-null conditional CDFs of the p-value, and the approximation comes from the fact that the null p-value is asymptotically uniform and it is asymptotically independent of the auxiliary statistic $U(\boldsymbol{s})$. Then the difference between $\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ and $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ can be approximated by $[\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\} - \pi\{\boldsymbol{s}, U(\boldsymbol{s})\}]/\pi\{\boldsymbol{s}, U(\boldsymbol{s})\} \approx -[1 - F_1\{\tau|\boldsymbol{s}, U(\boldsymbol{s})\}]/(1-\tau)$. Therefore, the difference between $\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ and $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ is small with a properly chosen τ, and it is asymptotically negative, which in turn would yield an asymptotically conservative FDR control. We discuss the choice of τ in Section 5.1.

Next, we develop a neighborhood-assisted approach to estimate $\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$. Intuitively, the estimator can be obtained by counting the proportion of p-values that are greater than τ among all p-values at the location $\boldsymbol{s}$ and with the same auxiliary covariate value $U(\boldsymbol{s})$. However, there is only one p-value at each $\{\boldsymbol{s}, U(\boldsymbol{s})\}$ pair. This prompts us to use a smoothing kernel approach to borrow information from the neighborhood of $\{\boldsymbol{s}, U(\boldsymbol{s})\}$. Specifically, consider a positive, bounded, unimodal kernel function $K(\boldsymbol{x}, y) : \mathbb{R}^{b+1} \to \mathbb{R}$ that is symmetric about zero in each dimension. Let $\boldsymbol{H} \in \mathbb{R}^{(b+1)\times(b+1)}$ be a positive definite bandwidth matrix, and write $K_{\boldsymbol{H}}\{\boldsymbol{s}, U(\boldsymbol{s})\} = |\boldsymbol{H}|^{-1/2} K(\boldsymbol{H}^{-1/2}\{\boldsymbol{s}^\intercal, U(\boldsymbol{s})\}^\intercal)$, where $|\cdot|$ is the determinant. We briefly remark that, the bandwidth matrix $\boldsymbol{H}$ is not diagonal, because of the dependency between $\boldsymbol{s}$ and $U(\boldsymbol{s})$ and possible correlations among the entries of $\boldsymbol{s}$. If we set $\boldsymbol{H}$ as a diagonal matrix with the same magnitude along the diagonal and ignore $U(\boldsymbol{s})$, then our estimator reduces to that in Cai, Sun and Xia (2022). For a given $\{\boldsymbol{s}, U(\boldsymbol{s})\}$, we assign the weight to the "pseudo-observation" of the p-value at $\{\boldsymbol{s}', U(\boldsymbol{s}')\}$ as

$$v_{\boldsymbol{H}}(\{\boldsymbol{s}, U(\boldsymbol{s})\}, \{\boldsymbol{s}', U(\boldsymbol{s}')\}) = \frac{K_{\boldsymbol{H}}\{\boldsymbol{s} - \boldsymbol{s}', U(\boldsymbol{s}) - U(\boldsymbol{s}')\}}{K_{\boldsymbol{H}}(\boldsymbol{0}, 0)}.$$

Then, the number of "pseudo-observations" that are greater than τ at each $\{\boldsymbol{s}, U(\boldsymbol{s})\}$ can be approximated by $\sum_{\boldsymbol{s}' \in \mathcal{I}(\tau)} v_{\boldsymbol{H}}\left(\{\boldsymbol{s}, U(\boldsymbol{s})\}, \{\boldsymbol{s}', U(\boldsymbol{s}')\}\right)$, where $\mathcal{I}(\tau) = \{\boldsymbol{s}' \in \mathcal{S} : p(\boldsymbol{s}') > \tau\}$. Meanwhile, the expectation of the p-values greater than τ can be calculated by $[\sum_{\boldsymbol{s}' \in \mathcal{S}} v_{\boldsymbol{H}}(\{\boldsymbol{s}, U(\boldsymbol{s})\}, \{\boldsymbol{s}', U(\boldsymbol{s}')\})][1 - \pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}]$ $(1-\tau)$. Setting the two equal, we obtain an estimator of $\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ as

$$\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\} = 1 - \frac{\sum_{\boldsymbol{s}' \in \mathcal{I}(\tau)} v_{\boldsymbol{H}}\left(\{\boldsymbol{s}, U(\boldsymbol{s})\}, \{\boldsymbol{s}', U(\boldsymbol{s}')\}\right)}{(1-\tau)\sum_{\boldsymbol{s}' \in \mathcal{S}} v_{\boldsymbol{H}}\left(\{\boldsymbol{s}, U(\boldsymbol{s})\}, \{\boldsymbol{s}', U(\boldsymbol{s}')\}\right)}. \tag{3.2}$$

Algorithm 1 The multiple testing procedure of NAPA.

Step 1. Calculate the weights $\hat{w}\{\boldsymbol{s}, U(\boldsymbol{s})\}$ as in (3.3), and then adjust p-values by $p_{\hat{w}}(\boldsymbol{s}) = p(\boldsymbol{s})/\hat{w}\{\boldsymbol{s}, U(\boldsymbol{s})\}$ for $\boldsymbol{s} \in \mathcal{S}$.

Step 2. Obtain the data-driven threshold

$$t_{\hat{w}} = \sup_t \left\{ 0 \le t \le 1 : \frac{\sum_{\boldsymbol{s}\in\mathcal{S}} \hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\} t}{\max\left\{\sum_{\boldsymbol{s}\in\mathcal{S}} \mathsf{I}\left\{p_{\hat{w}}(\boldsymbol{s}) \le t\right\}, 1\right\}} \le \alpha \right\}.$$

Step 3. Reject $H_0(\boldsymbol{s})$ if $p_{\hat{w}}(\boldsymbol{s}) \le t_{\hat{w}}$, $\boldsymbol{s} \in \mathcal{S}$.

We obtain the neighborhood-assisted and posterior-adjusted weight estimator as

$$\hat{w}\{\boldsymbol{s}, U(\boldsymbol{s})\} = \frac{\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}}{1 - \hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}}. \tag{3.3}$$

3.2. Multiple testing procedure

We next develop a general multiple testing procedure. We first observe that the expected number of false rejections with a known $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ at a given threshold t can be computed as

$$\begin{aligned}
&\mathsf{E}_{\{p(\boldsymbol{s}),\theta(\boldsymbol{s}),U(\boldsymbol{s})\},\boldsymbol{s}\in\mathcal{S}} \left[\sum_{\boldsymbol{s}\in\mathcal{S}} \mathsf{I}\left\{p_w(\boldsymbol{s}) < t, \theta(\boldsymbol{s}) = 0\right\}\right] \\
&= \sum_{\boldsymbol{s}\in\mathcal{S}} \mathsf{E}_{U(\boldsymbol{s})} \left(\mathsf{E}_{p(\boldsymbol{s}),\theta(\boldsymbol{s})} \left[\mathsf{I}\left\{p_w(\boldsymbol{s}) < t, \theta(\boldsymbol{s}) = 0\right\} | U(\boldsymbol{s})\right]\right) \\
&= \mathsf{E}_{U(\boldsymbol{s}),\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{s}\in\mathcal{S}} \left[\Pr\{\theta(\boldsymbol{s}) = 0 | U(\boldsymbol{s})\} \Pr\{p_w(\boldsymbol{s}) \le t | \theta(\boldsymbol{s}) = 0, U(\boldsymbol{s})\}\right] \\
&= \mathsf{E}_{U(\boldsymbol{s}),\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{s}\in\mathcal{S}} \left([1 - \pi\{\boldsymbol{s}, U(\boldsymbol{s})\}] w\{\boldsymbol{s}, U(\boldsymbol{s})\} t\right) = \mathsf{E}_{U(\boldsymbol{s}),\boldsymbol{s}\in\mathcal{S}} \left[\sum_{\boldsymbol{s}\in\mathcal{S}} \pi\{\boldsymbol{s}, U(\boldsymbol{s})\} t\right],
\end{aligned}$$

in the oracle setting. If the p-value is uniformly distributed asymptotically, and $U(\boldsymbol{s})$ is asymptotically independent of $T(\boldsymbol{s})$ under the null, then for a given estimate $\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ and the decision rule $\mathsf{I}\{p_{\hat{w}}(\boldsymbol{s}) \le t\}$, we can approximate the number of false rejection by $\sum_{\boldsymbol{s}\in\mathcal{S}} \hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\} t$. We aim to reject as many hypotheses as possible, while controlling the estimated false discovery proportion (FDP) not to exceed the pre-specified significance level. This leads to the proposed testing procedure as summarized in Algorithm 1.

3.3. Comparison of multivariate means

The problem of comparing two multivariate means has been widely studied, especially under the high-dimensional sparse setting (Cai, Liu and Xia, 2014; Liu,

2014; Xu et al., 2016). We illustrate the above general testing procedure with such a mean testing problem, and give a concrete construction of the test statistic $T(\boldsymbol{s})$ and the auxiliary statistic $U(\boldsymbol{s})$. Our method also applies to numerous other testing problems as well.

Specifically, given the observed data $\{Y_{i,d}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}_{i=1}^{n_d}$, where $Y_{i,d}(\boldsymbol{s}) \sim \mathcal{P}_{\beta_d(\boldsymbol{s}),\eta_d(\boldsymbol{s})}$, $\beta_d(\boldsymbol{s}) = \mathsf{E}\{Y_{i,d}(\boldsymbol{s})\}$, $d = 1, 2$, we aim to test the hypotheses,

$$H_0(\boldsymbol{s}) : \beta_1(\boldsymbol{s}) = \beta_2(\boldsymbol{s}) \quad \text{versus} \quad H_1(\boldsymbol{s}) : \beta_1(\boldsymbol{s}) \neq \beta_2(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{S}.$$

We construct the primary test statistic as

$$T(\boldsymbol{s}) = \frac{\bar{Y}_1(\boldsymbol{s}) - \bar{Y}_2(\boldsymbol{s})}{(\hat{\sigma}^2_{\boldsymbol{s},1}/n_1 + \hat{\sigma}^2_{\boldsymbol{s},2}/n_2)^{1/2}}, \quad \boldsymbol{s} \in \mathcal{S},$$

where $\bar{Y}_d(\boldsymbol{s}) = n_d^{-1} \sum_{i=1}^{n_d} Y_{i,d}(\boldsymbol{s})$ is the group sample mean, and $\hat{\sigma}^2_{\boldsymbol{s},d} = n_d^{-1} \sum_{i=1}^{n_d} \{Y_{i,d}(\boldsymbol{s}) - \bar{Y}_d(\boldsymbol{s})\}^2$ is the sample variance, $d = 1, 2$. Next, we construct the auxiliary statistic in the form of $\beta_1(\boldsymbol{s}) + \kappa(\boldsymbol{s})\beta_2(\boldsymbol{s})$, and for our mean comparison problem, we consider,

$$U(\boldsymbol{s}) = \frac{\bar{Y}_1(\boldsymbol{s}) + \hat{\kappa}(\boldsymbol{s})\bar{Y}_2(\boldsymbol{s})}{\{\hat{\sigma}^2_{\boldsymbol{s},1}/n_1 + \hat{\kappa}^2(\boldsymbol{s})\hat{\sigma}^2_{\boldsymbol{s},2}/n_2\}^{1/2}}, \quad \boldsymbol{s} \in \mathcal{S},$$

where $\hat{\kappa}(\boldsymbol{s}) = (n_2\hat{\sigma}^2_{\boldsymbol{s},1})/(n_1\hat{\sigma}^2_{\boldsymbol{s},2})$.

4. Theoretical Properties

In this section, we establish the theoretical properties of the NAPA testing procedure. We first show the estimated posterior non-null probability $\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ is a consistent estimator of $\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$. We then establish the asymptotic error rate control of NAPA under some conditions. Finally, we illustrate with the mean comparison problem again, and show the required conditions of NAPA are satisfied. Specifically, under the null, the test statistic $T(\boldsymbol{s})$ is asymptotically normally distributed, and thus the corresponding p-value is asymptotically uniformly distributed, and $T(\boldsymbol{s})$ is independent of the auxiliary statistic $U(\boldsymbol{s})$ asymptotically. Throughout our asymptotic analysis, we consider the infill-asymptotic framework (Stein, 1999) that $\mathcal{S} \to \mathbb{S}$.

4.1. Estimation consistency of the posterior probability

We begin with some notations. Let $\mathcal{S}_0 = \{\boldsymbol{s} \in \mathcal{S} : \theta(\boldsymbol{s}) = 0\}$ and $\mathcal{S}_1 = \{\boldsymbol{s} \in \mathcal{S} : \theta(\boldsymbol{s}) = 1\}$ denote the set of null locations and non-null locations, respectively, and let $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1$. Let $m = \text{Card}(\mathcal{S})$, $m_0 = \text{Card}(\mathcal{S}_0), m_1 = \text{Card}(\mathcal{S}_1)$, and $n = n_1 + n_2$. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \leq C|b_n|$ for any sufficiently large n, write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, and write $a_n \asymp b_n$ if there exists constants

$C > c > 0$ such that $c|b_n| \le |a_n| \le C|b_n|$ for any sufficiently large n. Let $\lambda_i(\cdot)$ and $\mathsf{tr}(\cdot)$ denote the ith eigenvalue and the trace of a matrix, respectively.

Next, we show that the estimator $\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ in (3.2) converges to the truth $\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ for all $\boldsymbol{s} \in \mathcal{S}$ as $\mathcal{S} \to \mathbb{S}$. Let $\boldsymbol{A} \in \mathbb{R}^{(b+2)\times(b+2)}$ denote the Hessian matrix of $\Pr\{p(\boldsymbol{s}') > \tau | U(\boldsymbol{s}'), U(\boldsymbol{s})\}$ with respect to $\{\boldsymbol{s}'^{\intercal}, U(\boldsymbol{s}'), U(\boldsymbol{s})\}^{\intercal}$. Partition the bandwidth matrix into $\boldsymbol{H} = \begin{pmatrix} \boldsymbol{H}_S & \boldsymbol{a} \\ \boldsymbol{a}^{\intercal} & h_U^2 \end{pmatrix}$, where $\boldsymbol{H}_S \in \mathbb{R}^{b\times b}$, $\boldsymbol{a} \in \mathbb{R}^{b\times 1}$, and $h_U^2 \in \mathbb{R}$. Let $\tilde{h} = h_U^2 - \boldsymbol{a}^{\intercal}\boldsymbol{H}_{\boldsymbol{S}}^{-1}\boldsymbol{a} \neq 0$. We introduce the following regularity conditions.

(C1) Suppose the kernel function $K(\boldsymbol{x}, y) : \mathbb{R}^{b+1} \to \mathbb{R}$ satisfies

$$\int_{\mathbb{R}^b} \boldsymbol{x}^{\intercal}\boldsymbol{x} K(\boldsymbol{x}, 0) d\boldsymbol{x} < \infty, \int_{\mathbb{R}^b} g^2(\boldsymbol{x}) K\{\boldsymbol{Q}\,[\boldsymbol{x}^{\intercal}, g(\boldsymbol{x})]^{\intercal}\} d\boldsymbol{x} < \infty,$$

for any orthogonal matrix $\boldsymbol{Q}$ and any function $g : \mathbb{R}^b \to \mathbb{R}$.

(C2) Let $\Lambda_{\boldsymbol{s}} = \{U(\boldsymbol{s})$: with probability $1 - O(m^{-1})$, uniformly for all $\boldsymbol{s}' \in \mathcal{S}$, $\Pr\{p(\boldsymbol{s}') > \tau | U(\boldsymbol{s}'), U(\boldsymbol{s})\}$ has continuous first and second partial derivatives at $\{\boldsymbol{s}'^{\intercal}, U(\boldsymbol{s}'), U(\boldsymbol{s})\}^{\intercal}$ and $\lambda_i(\boldsymbol{A}) = O(1)$ for $i = 1, \ldots, b+2\}$. Suppose $\Pr(\Lambda_{\boldsymbol{s}}) \to 1$ uniformly for all $\boldsymbol{s} \in \mathcal{S}$ as $\mathcal{S} \to \mathbb{S}$.

(C3) Suppose, uniformly for all $\boldsymbol{s} \in \mathcal{S}$,

$$\begin{aligned}
&\mathsf{Var}_{U(\boldsymbol{s}'), \boldsymbol{s}' \in \mathcal{S}} \left(\sum_{\boldsymbol{s}' \in \mathcal{S}} [K_{\boldsymbol{H}}\{\boldsymbol{s} - \boldsymbol{s}', U(\boldsymbol{s}) - U(\boldsymbol{s}')\}] \, | U(\boldsymbol{s}) \right) \\
&= O\left(\sum_{\boldsymbol{s}' \in \mathcal{S}} \mathsf{Var}_{U(\boldsymbol{s}')} \left[K_{\boldsymbol{H}}\{\boldsymbol{s} - \boldsymbol{s}', U(\boldsymbol{s}) - U(\boldsymbol{s}')\} | U(\boldsymbol{s}) \right] \right), \\
&\mathsf{Var}_{\{p(\boldsymbol{s}'), U(\boldsymbol{s}')\}, \boldsymbol{s}' \in \mathcal{S}} \left(\sum_{\boldsymbol{s}' \in \mathcal{S}} [K_{\boldsymbol{H}}\{\boldsymbol{s} - \boldsymbol{s}', U(\boldsymbol{s}) - U(\boldsymbol{s}')\} \mathsf{I}\{p(\boldsymbol{s}') > \tau\}] \, | U(\boldsymbol{s}) \right) \\
&= O\left(\sum_{\boldsymbol{s}' \in \mathcal{S}} \mathsf{Var}_{p(\boldsymbol{s}'), U(\boldsymbol{s}')} \left[K_{\boldsymbol{H}}\{\boldsymbol{s} - \boldsymbol{s}', U(\boldsymbol{s}) - U(\boldsymbol{s}')\} \mathsf{I}\{p(\boldsymbol{s}') > \tau\} | U(\boldsymbol{s}) \right] \right)
\end{aligned}$$

hold with probability tending to 1 as $\mathcal{S} \to \mathbb{S}$.

(C4) Suppose $\boldsymbol{H}_{\boldsymbol{S}}$ is nonsingular, $\tilde{h} \neq 0$, and uniformly for all $\boldsymbol{s} \in \mathcal{S}$,

$$\frac{1}{m} \sum_{\boldsymbol{s}' \in \mathcal{S}} \mathsf{E}_{U(\boldsymbol{s}')} \left[\{\boldsymbol{s} - \boldsymbol{s}', U(\boldsymbol{s}) - U(\boldsymbol{s}')\}^{\intercal} K_{\boldsymbol{H}}\{\boldsymbol{s} - \boldsymbol{s}', U(\boldsymbol{s}) - U(\boldsymbol{s}')\} | U(\boldsymbol{s}) \right] \to \boldsymbol{0}$$

holds with probability tending to 1 as $\mathcal{S} \to \mathbb{S}$. Furthermore, suppose that

$$\tilde{h} = O\{\mathsf{tr}(\boldsymbol{H})\}, \quad \mathsf{tr}(\boldsymbol{H}) = o(1), \quad m^{-1}|\boldsymbol{H}|^{-1/2} = o(\tilde{h}^{1/2}).$$

We make a few remarks about these conditions. Condition (C1) holds for

commonly used multivariate kernels, e.g., the standard normal kernel, the uniform kernel, among others. Condition (C2) regulates the first and second derivatives of the conditional CDF of the p-values, and is mild. Condition (C3) assumes that most of the auxiliary statistics $\{U(\boldsymbol{s})\}$ and most of the pairs $\{p(\boldsymbol{s}), U(\boldsymbol{s})\}$ are weakly correlated across $\boldsymbol{s}$. This condition holds for some of common spatial structures. We give more discussion of this condition in Section S3.6 of the online Supplementary Material. Condition (C4) generally states the symmetry of the kernel function and is mild too. It can also be verified numerically. Besides, the requirement $m^{-1}|\boldsymbol{H}|^{-1/2} = o(\tilde{h}^{1/2})$ reduces to Condition (A2) in Lemma 1 of Duong and Hazelton (2005) when ignoring the additional covariate $U(\boldsymbol{s})$.

Theorem 2. *Suppose Conditions* (C1) *to* (C4) *hold. Then, uniformly for all* $\boldsymbol{s} \in \mathcal{S}$,

$$\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\} \to \pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\} \text{ in probability, as } \mathcal{S} \to \mathbb{S}.$$

We remark that, LAWS (Cai, Sun and Xia, 2022) only considered the spatial information and applied a diagonal smoothing kernel with a homogenous bandwidth to estimate the weight. In comparison, to integrate the neighborhood information encoded by both $\boldsymbol{s}$ and $U(\boldsymbol{s})$, we develop a more sophisticated kernel estimation procedure that allows both non-orthogonal kernel components and heterogeneous bandwidth magnitudes.

4.2. Asymptotic error rate control of NAPA

Next, we show the NAPA procedure controls both the FDR and the FDP asymptotically, where we define the FDP of the test $\boldsymbol{\delta}_t^w = \{\delta^w(\boldsymbol{s}, t) : \boldsymbol{s} \in \mathcal{S}\}$ as,

$$\mathrm{FDP}\left(\boldsymbol{\delta}_t^w\right) = \frac{\sum_{\boldsymbol{s}\in\mathcal{S}}\left[\{1-\theta(\boldsymbol{s})\}\delta^w(\boldsymbol{s}, t)\right]}{\max\left\{\sum_{\boldsymbol{s}\in\mathcal{S}}\delta^w(\boldsymbol{s}, t), 1\right\}}.$$

We again begin with some regularity conditions.

(C5) Suppose that $n_1 \asymp n_2$, $\log m = o(n^{1/8})$. Let $\psi\{Y_{i,d}(\boldsymbol{s})\}$ denote the influence function of $\beta_d(\boldsymbol{s})$ at $Y_{i,d}(\boldsymbol{s})$. Let $Z_k(\boldsymbol{s}) = (n_2/n_1)\psi\{Y_{i,1}(\boldsymbol{s})\}$, for $k = i, i = 1, \ldots, n_1$, and $Z_k(\boldsymbol{s}) = -\psi\{Y_{i,2}(\boldsymbol{s})\}$, for $k = n_1 + i, i = 1, \ldots, n_2$. Suppose that $\mathsf{E}\{\exp(C_1|Z_k(\boldsymbol{s})|/[\mathsf{Var}\{Z_k(\boldsymbol{s})\}]^{1/2})\} < \infty$ for some $C_1 > 0$, and that there exists some $\mu(\boldsymbol{s}) = (1 + o\{(\log m)^{-1}\})\,\mathsf{E}\{U(\boldsymbol{s})\}$, such that

$$\Pr_{H_0(\boldsymbol{s})}\left\{\left|T(\boldsymbol{s}) - \frac{\sum_{k=1}^{n_1+n_2} Z_k(\boldsymbol{s})}{\mathsf{Var}\{\sum_{k=1}^{n_1+n_2} Z_k(\boldsymbol{s})\}^{1/2}}\right| \geq b_m\right\} = O(m^{-C_2}),$$

$$\Pr_{H_0(\boldsymbol{s})}\left\{\left|[U(\boldsymbol{s}) - \mu(\boldsymbol{s})] - \frac{\sum_{k=1}^{n_1} Z_k(\boldsymbol{s}) - \vartheta(\boldsymbol{s})\sum_{k=n_1+1}^{n_1+n_2} Z_k(\boldsymbol{s})}{\mathsf{Var}\{\sum_{k=1}^{n_1} Z_k(\boldsymbol{s}) - \vartheta(\boldsymbol{s})\sum_{k=n_1+1}^{n_1+n_2} Z_k(\boldsymbol{s})\}^{1/2}}\right| \geq b_m\right\}$$
$$= O(m^{-C_2}),$$

where $\vartheta(\boldsymbol{s}) = [n_1\mathsf{Var}\{Z_1(\boldsymbol{s})\}] / [n_2\mathsf{Var}\{Z_n(\boldsymbol{s})\}]$, for some constant $C_2 > 5$

and $b_m = o\{(\log m)^{-1/2}\}$.

(C6) For $\boldsymbol{Z}_k = \{Z_k(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}$ as defined in (C5), let $\boldsymbol{R}_1 = \mathsf{Corr}(\boldsymbol{Z}_k) = (r_{\boldsymbol{s},\boldsymbol{l};1})_{m\times m}$ for $1 \le k \le n_1$, $\boldsymbol{R}_2 = \mathsf{Corr}(\boldsymbol{Z}_k) = (r_{\boldsymbol{s},\boldsymbol{l};2})_{m\times m}$ for $n_1 + 1 < k \le n_1 + n_2$, and suppose that $\max\{\lambda_1(\boldsymbol{R}_1), \lambda_1(\boldsymbol{R}_2)\} = o(m)$. Furthermore, let $\Gamma_{\boldsymbol{s}}(\gamma) = \{\boldsymbol{l} : \boldsymbol{l} \in \mathcal{S}, |r_{\boldsymbol{s},\boldsymbol{l};d}| \ge (\log m)^{-2-\gamma},\ d = 1 \text{ or } 2\}$, and suppose that there exists some $\gamma > 0$, such that $\max_{\boldsymbol{s}\in\mathcal{S}_0} \mathrm{Card}\,\{\Gamma_{\boldsymbol{s}}(\gamma)\} \asymp 1$.

(C7) Suppose, with probability tending to 1, uniformly for all $\boldsymbol{s} \in \mathcal{S}$, $\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\} \in [\xi, 1-\xi]$ for some sufficiently small constant $\xi > 0$, and $\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ has bounded first derivatives with respect to $U(\boldsymbol{s})$. Furthermore, suppose that

$$\mathsf{Var}_{\theta(\boldsymbol{s}),\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{s}\in\mathcal{S}} \left(\mathsf{E}_{U(\boldsymbol{s})}\left[w\{\boldsymbol{s}, U(\boldsymbol{s})\}\middle|\{\theta(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}\right] \mathsf{I}\{\theta(\boldsymbol{s}) = 0\}\right) = o\left(m^2\right).$$

(C8) Let $\mathcal{S}_\nu = \{\boldsymbol{s} \in \mathcal{S} : \{\beta_1(\boldsymbol{s}) - \beta_2(\boldsymbol{s})\}/[\mathsf{Var}\{\sum_{k=1}^n Z_k(\boldsymbol{s})\}^{1/2}/n_2] \ge (\log m)^{1/2+\nu}\}$ for $\boldsymbol{Z}_k = \{Z_k(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}$ as defined in (C5). Suppose $\mathrm{Card}\,(\mathcal{S}_\nu) \ge \{1/(c_\pi^{1/2}\alpha) + \varepsilon\}(\log m)^{1/2}$ for some $\varepsilon, \nu > 0$, and c_π is the ratio of a circle's circumference to its diameter.

Condition (C5) assumes the asymptotic normality of $T(\boldsymbol{s})$ and $U(\boldsymbol{s})$ under the null, which is easily attainable, as we illustrate with the problem of mean comparison in Section 4.3. It also implies the asymptotic independence between $T(\boldsymbol{s})$ and $U(\boldsymbol{s})$ under the null, as we show in Lemma 1 of the online Supplementary Material. Condition (C6) requires that not too many variables have strong correlations that exceed $(\log m)^{-2-\gamma}$. This condition holds for some of common spatial structures, but may not hold for some other spatial covariance functions. This condition is imposed mainly due to the technical challenges in analyzing the correlation structures of two sequences of statistics. We give more discussion of this condition in Section S3.7 of the online Supplementary Material. Additionally, we show in Section S3.7 that, when Condition (C6) does not hold, our method can still maintain a reasonably good empirical performance. Condition (C7) requires $\pi_\tau(\boldsymbol{s}, U(\boldsymbol{s}))$ to vary smoothly with respect to $U(\boldsymbol{s})$, and not to be exactly 0 or 1 to ensure theoretical stability. It also assumes the latent variables $\theta(\boldsymbol{s})$ is not perfectly correlated, which ensures that $m_0 \asymp m$ has the probability tending to 1. Condition (C8) requires a few spatial locations to have the standardized signal magnitude exceeding $(\log m)^{1/2+\nu}$, which avoids an overly conservative FDR. In general, these conditions are mild, and similar conditions of (C6) to (C8) have been imposed in Cai, Sun and Xia (2022).

The next two theorems establish the asymptotical control of FDR and FDP at the nominal level, first for a known $\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ in Theorem 3, and then for the estimated $\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ in Theorem 4.

Theorem 3. *Suppose Conditions* (C5) *to* (C8) *hold. Then,*

$$\overline{\lim_{\mathcal{S}\to\mathbb{S}}} \mathrm{FDR}\left(\boldsymbol{\delta}_{t_w}^{NAPA}\right) \le \alpha, \quad \textit{and}$$
$$\lim_{\mathcal{S}\to\mathbb{S}} \Pr\left\{\mathrm{FDP}\left(\boldsymbol{\delta}_{t_w}^{NAPA}\right) \le \alpha+\epsilon\right\} = 1, \ \textit{for any} \ \ \epsilon>0,$$

where $t_w = \sup_t\{0 \le t \le 1 : [\sum_{\boldsymbol{s}\in\mathcal{S}} \pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}t]/\max\{\sum_{\boldsymbol{s}\in\mathcal{S}} \mathsf{I}\{p_w(\boldsymbol{s}) \le t\}, 1\} \le \alpha\}$*, and* $w\{\boldsymbol{s}, U(\boldsymbol{s})\} = \pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}/[1-\pi_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}]$.

Theorem 4. *Suppose Conditions* (C1) *to* (C8) *hold. Then,*

$$\overline{\lim_{\mathcal{S}\to\mathbb{S}}} \mathrm{FDR}\left(\boldsymbol{\delta}_{t_{\hat{w}}}^{NAPA}\right) \le \alpha, \quad \textit{and}$$
$$\lim_{\mathcal{S}\to\mathbb{S}} \Pr\left\{\mathrm{FDP}\left(\boldsymbol{\delta}_{t_{\hat{w}}}^{NAPA}\right) \le \alpha+\epsilon\right\} = 1, \ \textit{for any} \ \ \epsilon>0.$$

4.3. Asymptotic properties of mean comparison

Finally, we revisit the example of comparing multivariate means in Section 3.3, and show that the required asymptotic normality and independence both hold. For other testing problems such as comparing the networks and detecting interactions, similar properties can be established accordingly; see Xia, Cai and Sun (2020). We introduce two additional regularity conditions. Recall that, in this setting, $\beta_d(\boldsymbol{s}) = \mathsf{E}\{Y_{i,d}(\boldsymbol{s})\}$. Let $\sigma^2_{\boldsymbol{s},d} = \mathsf{Var}\{Y_{i,d}(\boldsymbol{s})\}$, and $\kappa(\boldsymbol{s}) = (n_2\sigma^2_{\boldsymbol{s},1})/(n_1\sigma^2_{\boldsymbol{s},2})$, for all $\boldsymbol{s}\in\mathcal{S}$.

(C9) Suppose $\log m = o(n^{1/5})$, $n_1 \asymp n_2$, and $\sigma^2_{\boldsymbol{s},1} \asymp \sigma^2_{\boldsymbol{s},2}$ for all $\boldsymbol{s}\in\mathcal{S}$.

(C10) There exists some constant $C_1 > 0$, such that $\mathsf{E}[\exp\{C_1|Y_{i,d}(\boldsymbol{s}) - \beta_d(\boldsymbol{s})|/\sigma_{\boldsymbol{s},d}\}] < \infty$, for $d = 1, 2$ and all $\boldsymbol{s}\in\mathcal{S}$.

Condition (C9) allows the total number of hypotheses to test m to grow exponentially with the total sample size n, while requiring the sample size and the variance of each group to be of the same order. Condition (C10) holds for a broad family of distributions with an exponential tail and is similarly assumed in various testing literatures (Cai, Liu and Xia, 2013; Cai and Liu, 2016; Guo and Tang, 2021; He et al., 2021). Both conditions are mild.

Proposition 1. *Suppose Conditions* (C9) *and* (C10) *hold. Then under the null hypothesis, for any constant* $C_3 > 0$*, there exists* $b_m = o\{(\log m)^{-1/2}\}$*, such that*

$$\Pr\left\{\left|T(\boldsymbol{s}) - \frac{\bar{Y}_1(\boldsymbol{s}) - \bar{Y}_2(\boldsymbol{s})}{(\sigma^2_{\boldsymbol{s},1}/n_1 + \sigma^2_{\boldsymbol{s},2}/n_2)^{1/2}}\right| \ge b_m\right\} = O\left(m^{-C_3}\right),$$
$$\Pr\left\{\left|\{U(\boldsymbol{s}) - \mu(\boldsymbol{s})\} - \frac{\bar{Y}_1(\boldsymbol{s}) - \beta_1(\boldsymbol{s}) + \kappa(\boldsymbol{s})\{\bar{Y}_2(\boldsymbol{s}) - \beta_2(\boldsymbol{s})\}}{\{\sigma^2_{\boldsymbol{s},1}/n_1 + \kappa^2(\boldsymbol{s})\sigma^2_{\boldsymbol{s},2}/n_2\}^{1/2}}\right| \ge b_m\right\} = O(m^{-C_3}),$$

uniformly for $\boldsymbol{s}\in\mathcal{S}_0$*, where* $\mu(\boldsymbol{s}) = [1+o\{(\log m)^{-1}\}]\,\mathsf{E}\{U(\boldsymbol{s})\}$*. This further*

implies that, for any constant $C_4 > 0$,

$$\Pr\{|T(\boldsymbol{s})| \geq t | U(\boldsymbol{s})\} = \{1 + o(1)\}G(t) + O(m^{-C_4}),$$

uniformly in $t = O\{(\log m)^{1/2}\}$, $|U(\boldsymbol{s}) - \mu(\boldsymbol{s})| = O\{(\log m)^{1/2}\}$, *and all* $\boldsymbol{s} \in \mathcal{S}_0$, *where* $G(t) = 2\{1 - \Phi(t)\}$, *and* $\Phi(\cdot)$ *is the CDF of a standard normal random variable.*

We make some remarks. First, Xia, Cai and Sun (2020) illustrated their GAP test with the problem of comparing two multivariate means too, but they only studied the multivariate normal distribution, while Proposition 1 extends to the family of distributions with an exponential tail. Second, compared to the asymptotic independence result in GAP, Proposition 1 establishes the exact conditional probability tail of $T(\boldsymbol{s})$ given $U(\boldsymbol{s})$. As a result, the proof of Proposition 1 is technically much more involved. Toward our goal, we obtain a conditional normal approximation result in the proof of Proposition 1, which to our knowledge is not available in the literature. Finally, because we incorporate the auxiliary statistic in the construction of the weight, Proposition 1 ensures that the null distribution of $T(\boldsymbol{s})$ is not to be affected by the observed $U(\boldsymbol{s})$. By contrast, the auxiliary statistic is only used for the grouping purpose in GAP, and hence their asymptotic independence result can be viewed as a simpler discretized version of Proposition 1.

5. Simulations

In this section, we first study the capability of NAPA in recovering the posterior non-null probability $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$, then the finite-sample performance of NAPA, and compare with BH (Benjamini and Hochberg, 1995), GAP (Xia, Cai and Sun, 2020), LAWS (Cai, Sun and Xia, 2022), and a simple combination of GAP and LAWS. This combination method first applies GAP with three groups to obtain the group-wise reweighted p-values; then feeds into the LAWS method for the second-stage reweighing. We present additional simulations in Section S3.1, and study irregular domain and lattice in Section S3.3, and heavy-tailed distribution in Section S3.4 of the online Supplementary Material.

5.1. Posterior probability estimation

Given the key role the posterior non-null probability plays, we first evaluate the capability of NAPA in recovering $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ through the estimator $\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ in (3.2). We simulate two groups of independent samples $\{Y_{i,1}(\boldsymbol{s})\}_{i=1}^{n_1}$ and $\{Y_{i,2}(\boldsymbol{s})\}_{i=1}^{n_2}$,

$$\begin{aligned} Y_{i,1}(\boldsymbol{s}) \mid \theta(\boldsymbol{s}) &\sim \{1 - \theta(\boldsymbol{s})\}\ \text{Normal}(0, 1) \ + \ \theta(\boldsymbol{s})\ \text{Normal}\big\{\beta_1(\boldsymbol{s}), 1\big\}, \\ Y_{i,2}(\boldsymbol{s}) \mid \theta(\boldsymbol{s}) &\sim \{1 - \theta(\boldsymbol{s})\}\ \text{Normal}(0, 4) \ + \ \theta(\boldsymbol{s})\ \text{Normal}\big\{\beta_2(\boldsymbol{s}), 4\big\}, \end{aligned} \tag{5.1}$$

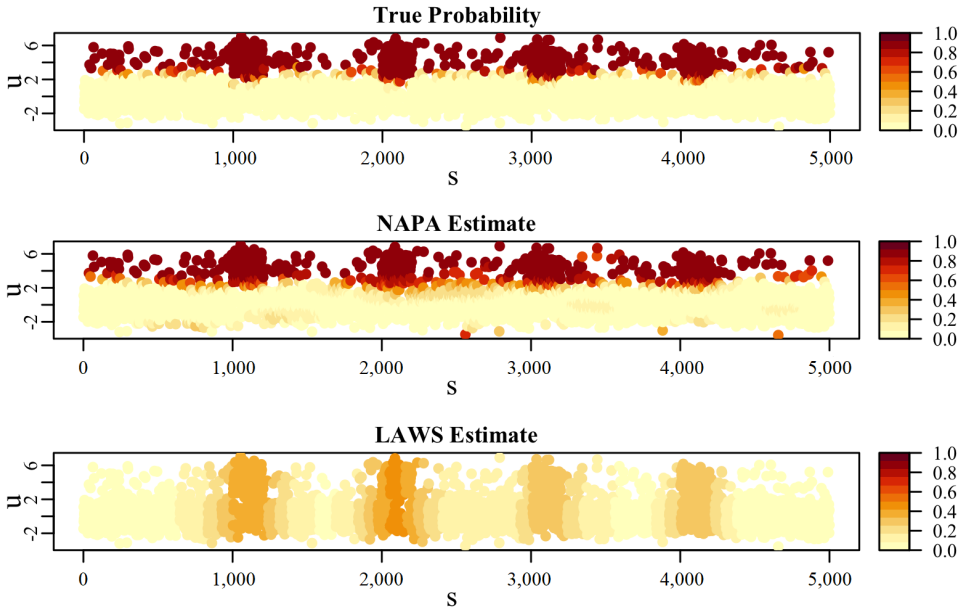


Figure 1. Estimation of the posterior non-null probability $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ for the 1D example. From top to bottom: the true probability, the estimated probability by NAPA, and the estimated probability by LAWS.

where $\theta(\boldsymbol{s}) \sim \text{Bernoulli}\{1, \pi(\boldsymbol{s})\}$, $\beta_1(\boldsymbol{s}) = 1/\sqrt{20}$, and $\beta_2(\boldsymbol{s}) = 2/\sqrt{5}$. Note that $\pi(\boldsymbol{s})$ specifies the likelihood of possible signal locations. We consider three examples of generating the signal regions: a 1D example of a piecewise constant-shaped signal, a 2D example of two rectangular-shaped signals, and a 3D example of a cubic-shaped signal. For the 1D case, we consider $s = 1, 2, \ldots, 5000$, and we set $\pi(s) = 0.8$ for $s \in [1001, 1200] \cup [2001, 2200]$, and $\pi(s) = 0.6$ for $s \in [3001, 3200] \cup [4001, 4200]$. For the 2D case, we consider $\boldsymbol{s} = (s_1, s_2)$, with $s_1 = 1, 2, \ldots, 100$, $s_2 = 1, 2, \ldots, 50$, and we set $\pi(\boldsymbol{s}) = 0.8$ for the left signal rectangle when $s_1 \in [20, 40]$, $s_2 \in [10, 30]$, and $\pi(\boldsymbol{s}) = 0.6$ for the right signal rectangle when $s_1 \in [60, 80]$, $s_2 \in [10, 30]$. For the 3D case, we consider $\boldsymbol{s} = (s_1, s_2, s_3)$, with $s_1 = 1, 2, \ldots, 20$, $s_2 = 1, 2, \ldots, 25$, $s_3 = 1, 2, \ldots, 15$, and we set $\pi(\boldsymbol{s}) = 0.7$ for the signal cube when $s_1 \in [5, 15]$, $s_2 \in [5, 15]$, $s_3 \in [1, 10]$. We set $\pi(\boldsymbol{s}) = 0.05$ for all the rest of locations. We set the sample size at $n_1 = n_2 = 100$. Given the generative model (5.1), we can derive the explicit distribution of $U(\boldsymbol{s})|\theta(\boldsymbol{s})$, and plugging it into (2.1) yields the true posterior non-null probability $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$.

Next, we estimate $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ using a bivariate Gaussian kernel function with the two-dimensional bandwidth matrix,

$$\boldsymbol{H} = \begin{pmatrix} h_{\boldsymbol{s}}^2 & \rho h_{\boldsymbol{s}} h_U \\ \rho h_{\boldsymbol{s}} h_U & h_U^2 \end{pmatrix}, \tag{5.2}$$

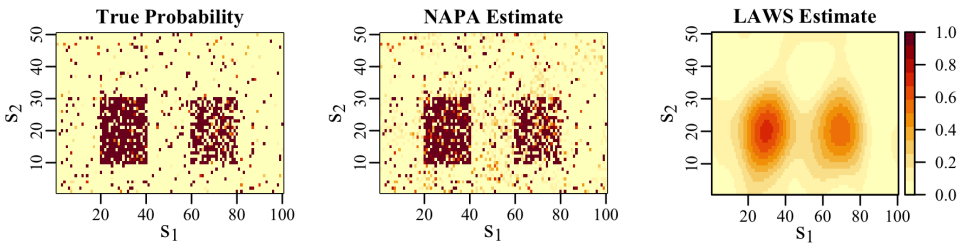


Figure 2. Estimation of the posterior non-null probability $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ for the 2D example. From left to right: the true probability, the estimated probability by NAPA, and the estimated probability by LAWS.

where $h_{\boldsymbol{s}}$ and h_U are the bandwidths for $\|\boldsymbol{s}-\boldsymbol{s}'\|$ and $\|U(\boldsymbol{s})-U(\boldsymbol{s}')\|$, respectively, ρ is the correlation between $\|\boldsymbol{s}-\boldsymbol{s}'\|$ and $\|U(\boldsymbol{s})-U(\boldsymbol{s}')\|$, and $\|\cdot\|$ denotes the Euclidean norm. We use the plug-in selector (Sheather and Jones, 1991) and the normal-scale selector (Chacón, Duong and Wand, 2011) to obtain $h_{\boldsymbol{s}}$ and h_U, respectively. Specifically, we employ the `R` package `ks` (Duong, 2007; Duong et al., 2022), in which we use the function `hpi` to select the bandwidth for $\|\boldsymbol{s}-\boldsymbol{s}'\|$, and the function `hns` to select the bandwidth for $\|U(\boldsymbol{s})-U(\boldsymbol{s}')\|$. We estimate ρ by the sample correlation. We further carry out a sensitivity analysis for the bandwidth selection in Section S3.2 of the online Supplementary Material. Moreover, we follow Cai, Sun and Xia (2022) and choose τ in (3.2) as the cutoff p-value when applying the BH procedure to the sequence of all the unweighted p-values at the significance level 0.9. This ensures that the null cases are dominant in the set $\mathcal{I}(\tau) = \{\boldsymbol{s}' \in \mathcal{S} : p(\boldsymbol{s}') > \tau\}$. Finally, to stabilize the probability estimation, we truncate $\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\} = \xi$ if $\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\} < \xi$, and $\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\} = 1-\xi$ if $\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\} > 1-\xi$, where we set $\xi = 10^{-5}$.

We compare our posterior probability estimator $\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ in (3.2) that utilizes both the smoothness and sparsity information with the truth $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$. We also compare with the corresponding estimator $\hat{\pi}_\tau(\boldsymbol{s})$ used in LAWS that only utilizes the smoothness information alone. Figures 1, 2 and 3 report the results based on a single data replication for the 1D, 2D and 3D examples, respectively. It is clearly seen that our posterior probability estimator $\hat{\pi}_\tau\{\boldsymbol{s}, U(\boldsymbol{s})\}$ is much closer to the truth than the LAWS estimator $\hat{\pi}_\tau(\boldsymbol{s})$.

5.2. FDR and power comparison

Next, we evaluate the empirical FDR and power of the proposed NAPA method, and compare it with BH, GAP, LAWS, and a simple combination of GAP and LAWS.

We continue to simulate the data from model (5.1) with the 1D, 2D and 3D examples. Denote by $\Delta_\beta = \beta_2(\boldsymbol{s}) - \beta_1(\boldsymbol{s})$ and it has the same value across all signal locations. Note that Δ_β controls the strength of the signal, whereas $\pi(\boldsymbol{s})$

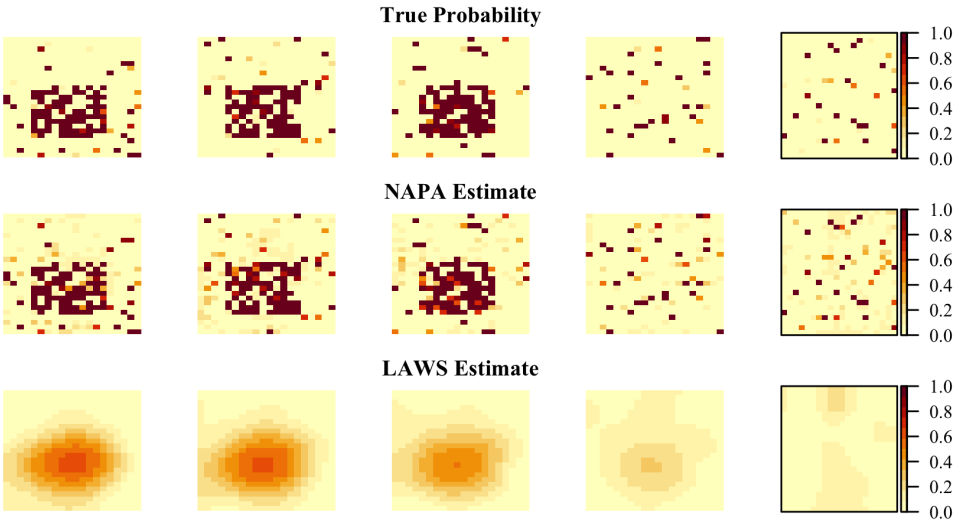


Figure 3. Estimation of the posterior non-null probability $\pi\{\boldsymbol{s}, U(\boldsymbol{s})\}$ for the 3D example. (Left to right) Five selected slices with $s_3 = 3, 6, 9, 12, 15$. (Top to bottom) The true probability, the estimated probability by NAPA, and the estimated probability by LAWS.

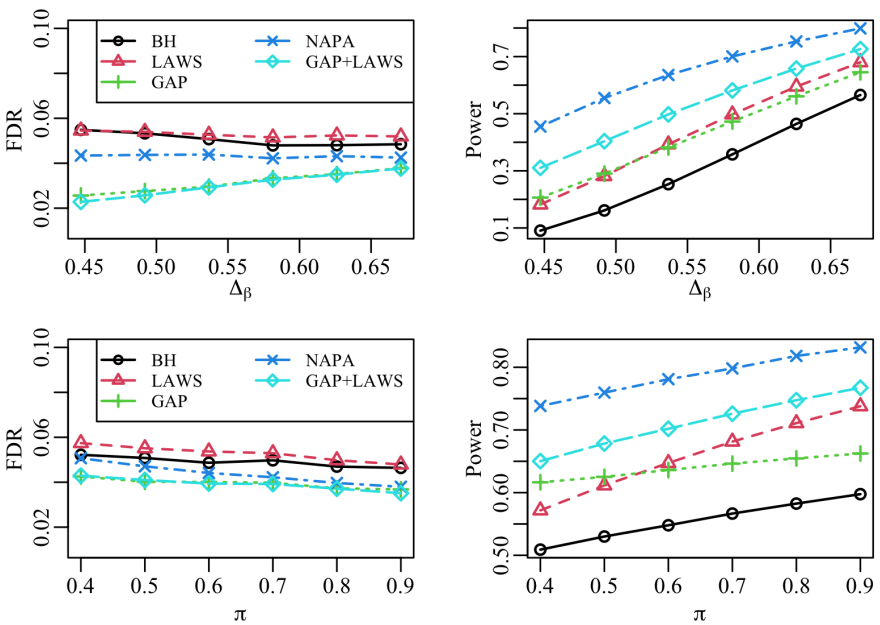


Figure 4. Empirical FDR and power for the 1D example. Top panels: varying Δ_β in scenario 1, and bottom panels: varying π in scenario 2. Five methods are compared: the proposed method (NAPA), the GAP method (Xia, Cai and Sun, 2020), the LAWS method (Cai, Sun and Xia, 2022), the simple combination of GAP and LAWS, and the BH method (Benjamini and Hochberg, 1995).

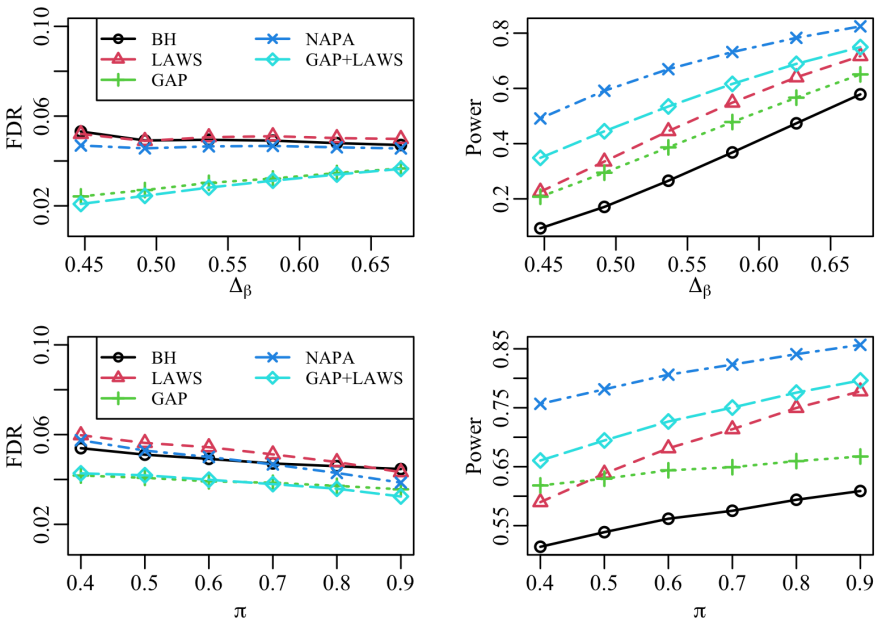


Figure 5. Empirical FDR and power for the 2D example, with same legend as Figure 4.

specifies the likelihood of possible signal locations. We consider two scenarios: vary Δ_β from $1/\sqrt{5}$ to $3/\sqrt{20}$, while fixing $\pi(\boldsymbol{s})$ in the same way as in Section 5.1; vary $\pi(\boldsymbol{s}) = \pi$ in all signal regions from 0.4 to 0.9, while fixing $\Delta_\beta = 3/\sqrt{20}$. Let $\beta_1(\boldsymbol{s}) \sim \text{Uniform}(-1, 1)/\sqrt{5}$ in both scenarios. We set the sample size at $n_1 = n_2 = 100$, and set the nominal level at $\alpha = 0.05$.

Figures 4, 5, and 6 report the empirical FDR and power of various testing methods based on 200 data replications. It is clearly seen that, in all three examples, while all methods can control the FDR around the nominal level, our proposed NAPA method achieves the most power gain compared to all the alternative methods. These results agree with our theory as well as our intuition that the NAPA method that utilizes both spatial and sparsity information outperforms the GAP and LAWS methods that utilize only one type of side information alone. Moreover, our method clearly outperforms the simple combination of GAP and LAWS. This simple combination has no theoretical guarantee. Besides, it loses information when applying the GAP method that reweighs the p-values in a discrete fashion. Through these examples, we see that our proposed NAPA test is more than just a simple combination of GAP and LAWS.

6. Real Data Applications

In this section, we illustrate our proposed test with two neuroimaging applications.

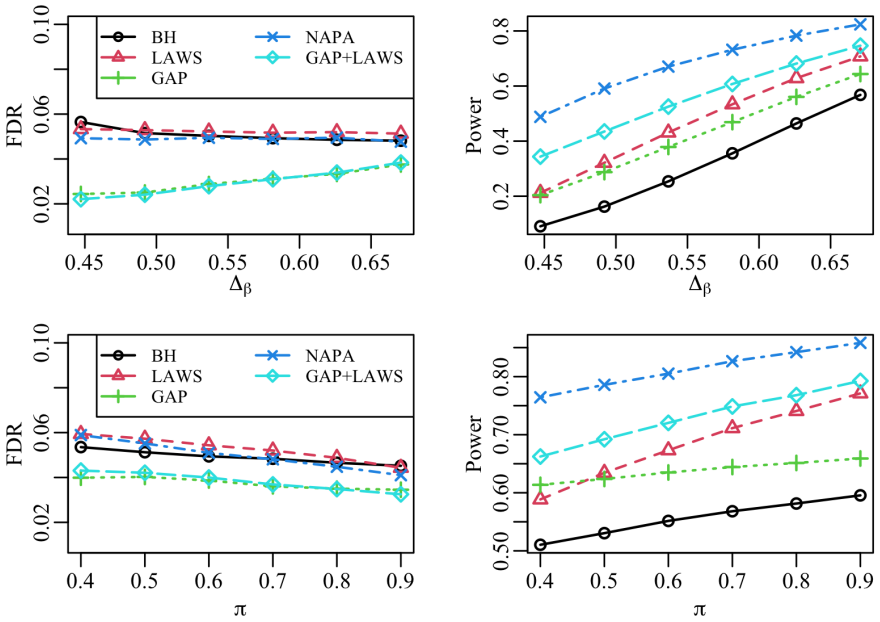


Figure 6. Empirical FDR and power for the 3D example, with same legend as Figure 4.

6.1. Multiple sclerosis study

The first study is to compare the cerebral white matter tracts between multiple sclerosis (MS) patients and healthy controls (Goldsmith et al., 2011). MS is a demyelinating autoimmune disease that causes lesions in the white matter tracts of a patient and results in severe disability. Diffusion tensor imaging (DTI) is a magnetic resonance imaging (MRI) technique that studies white matter tractography by measuring the diffusivity of water in the brain. The data records the fractional anisotropy measure, which describes the degree of diffusion anisotropy, along the right corticospinal tract for $n_1 = 340$ multiple sclerosis patients and $n_2 = 42$ healthy controls. The tract data are generally modeled as 1D functions, and there are in total $|\mathcal{S}| = 43$ locations for each tract. The dataset is available in the `R` library `refund`, and the data processing information can be found in Luo and Qi (2017). The scientific interest here is to compare the two mean functional profiles of diffusivity and locate the tract locations that distinguish cases from controls.

We apply the NAPA test to this dataset, and also compare it with the alternative tests, all under the nominal level $\alpha = 0.05$. The number of identified differential locations by NAPA, BH, GAP, LAWS, and their simple combination is 30, 0, 7, 25, and 17, respectively. Besides, the set of locations found by NAPA is a superset of those found by GAP and the simple combination of GAP and LAWS, and is also a superset of those found by LAWS except for two locations.

Together with our simulation studies, it seems to suggest that our proposed NAPA test manages to achieve the best power. It is also interesting to note that the locations identified by NAPA concentrate on the regions with distances 13 to 20, 22 to 32, and 37 to 47 along the tract. Such a finding warrants additional scientific validation.

6.2. Attention deficit hyperactivity disorder study

The second study is to compare the brain grey matter cortical thickness between subjects diagnosed with attention deficit hyperactivity disorder (ADHD) and typically developing controls (Bellec et al., 2017). ADHD is one of the most common child-onset neurodevelopmental disorders. Anatomical MRI is an imaging technique that studies brain anatomical structures. The data records the volume of grey matter at different brain locations in a 3D space for $n_1 = 356$ ADHD subjects and $n_2 = 575$ normal controls. The dataset is available at `http://neurobureau.projects.nitrc.org/ADHD200/Data.html`. The MRI images were preprocessed by the Neuro Bureau using the burner pipeline (Bellec et al., 2017). To reduce the dimensionality of the problem, we further downsize the image resolution from $256 \times 198 \times 256$ to $30 \times 36 \times 30$, following the same data reduction strategy as in Li and Zhang (2017) and Cai, Sun and Xia (2022). The scientific interest here is to compare the mean of two sets brain structural images and identify differentiating brain regions.

We apply the proposed NAPA test to this dataset, and also compare it with the alternative tests, all under the nominal level $\alpha = 0.05$. The number of identified differential locations by NAPA, BH, GAP, LAWS, and their simple combination is 1193, 349, 641, 539, and 948, respectively. Besides, the set of locations found by NAPA contains the majority of those found by BH, GAP, LAWS, and their simple combination, with the overlapping percentage equal to 97.1% of BH, 90.8% of GAP, 95.4% of LAWS, and 82.4% of the simple combination. Again, together with our simulation studies, it seems to suggest that our proposed NAPA test manages to achieve the best power. Comparing the identified locations with the Desikan-Killiany brain atlas (Desikan et al., 2006), a number of brain regions stand out, including the left and right entorhinal cortex, the left and right posterior cingulate cortex, and left precuneus, among others. These findings generally agree with the current literature on ADHD. Particularly, the posterior cingulate cortex forms a central node in the default mode network of the brain, and has been shown to communicate with various brain networks. ADHD has been suggested as a disorder of the default mode network, and there has been evidence showing that abnormalities in the posterior cingulate cortex may disrupt the default mode network that leads to attentional lapses (Nakao et al., 2011).

Supplementary Material

The online Supplementary Material contains all proofs and additional numerical results and discussions.

Acknowledgments

This work was supported by National Natural Science Foundation of China grants 12022103 and 12331009, NSF grant CIF-2102227, and NIH grants R01AG061303 and R01AG062542.

References

Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D. S. and Craddock, R. C. (2017). The neuro bureau ADHD-200 preprocessed repository. *NeuroImage* **144**, 275–286.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300.

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics* **24**, 407–418.

Bini, L. M., Diniz-Filho, J. A. F., Rangel, T. F. L. V. B., Akre, T. S. B., Albaladejo, R. G., Albuquerque, F. S. et al. (2009). Coefficient shifts in geographical ecology: An empirical evaluation of spatial and non-spatial regression. *Ecography* **32**, 193–204.

Cai, T., Liu, W. and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association* **108**, 265–277.

Cai, T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **76**, 349–372.

Cai, T. T. and Liu, W. (2016). Large-scale multiple testing of correlations. *Journal of the American Statistical Association* **111**, 229–240.

Cai, T. T. and Sun, W. (2017). Large-scale global and simultaneous inference: Estimation and testing in very high dimensions. *Annual Review of Economics* **9**, 411–439.

Cai, T. T., Sun, W. and Wang, W. (2019). Covariate-assisted ranking and screening for large-scale two-sample inference. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **81**, 187–234.

Cai, T. T., Sun, W. and Xia, Y. (2022). LAWS: A locally adaptive weighting and screening approach to spatial multiple testing. *Journal of the American Statistical Association* **117**, 1370–1383.

Caspi, A. and Moffitt, T. E. (2006). Gene-environment interactions in psychiatry: Joining forces with neuroscience. *Nature Reviews. Neuroscience* **7**, 583–590.

Chacón, J. E., Duong, T. and Wand, M. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica* **21**, 807–840.

Chen, S., Kang, J., Xing, Y. and Wang, G. (2015). A parsimonious statistical method to detect groupwise differentially expressed functional connectivity networks. *Human Brain Mapping* **36**, 5196–5206.

Czakon, N. G., Vayonakis, A., Schlaerth, J., Hollister, M. I., Golwala, S., Day, P. K. et al. (2009). Microwave Kinetic Inductance Detector (MKID) camera testing for submillimeter astronomy. In *AIP Conference Proceedings*, 172–175.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D. et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968–980.

Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of statistical software* **21**, 1–16.

Duong, T. and Hazelton, M. L. (2005). Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *Journal of Multivariate Analysis* **93**, 417–433.

Duong, T., Wand, M., Chacon, J. and Gramacki, A. (2022). `ks`: Kernel smoothing. `R` package, version 1.14.0.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **23**, 23–28.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

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

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B. and Reich, D. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics* **20**, 830–851.

Guo, X. and Tang, C. (2021). Specification tests for covariance structures in high-dimensional statistical models. *Biometrika* **108**, 335–351.

He, Y., Xu, G., Wu, C. and Pan, W. (2021). Asymptotically independent U-statistics in high-dimensional testing. *The Annals of Statistics* **49**, 154–181.

Held, L. and Ott, M. (2018). On p-values and Bayes factors. *Annual Review of Statistics and Its Application* **5**, 393–419.

Hu, J. X., Zhao, H. and Zhou, H. H. (2010). False discovery rate control with groups. *Journal of the American Statistical Association* **105**, 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. *Nature Methods* **13**, 577–580.

Lei, L. and Fithian, W. (2018). AdaPT: An interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **80**, 649–679.

Lei, L., Ramdas, A. and Fithian, W. (2021). A general interactive framework for false discovery rate control under structural constraints. *Biometrika* **108**, 253–267.

Li, A. and Barber, R. F. (2019). Multiple testing with the structure-adaptive Benjamini-Hochberg algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **81**, 45–74.

Li, J. and Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics* **40**, 908–940.

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association* **112**, 1131–1146.

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

Luo, R. and Qi, X. (2017). Function-on-function linear regression by signal compression. *Journal of the American Statistical Association* **112**, 690–705.

Nakao, T., Radua, J., Rubia, K. and Mataix-Cols, D. (2011). Gray matter volume abnormalities in ADHD: Voxel-based meta-analysis exploring the effects of age and stimulant medication. *American Journal of Psychiatry* **168**, 1154–1163.

Newton, M. A. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155–176.

Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical Science* **24**, 398–413.

Schweder, T. and Spjotvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika* **69**, 493–502.

Sellke, T., Bayarri, M. and Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician* **55**, 62–71.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* **53**, 683–690.

Stein, M. L. (1999). *Interpolation of Spatial Data*. Springer Series in Statistics. Springer-Verlag, New York.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**, 479–498.

Sun, D., Tsutakawa, R. K., Kim, H. and He, Z. (2000). Spatio-temporal interaction with disease mapping. *Statistics in Medicine* **19**, 2015–2035.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association* **102**, 901–912.

Sun, W. and Wei, Z. (2011). Multiple testing for pattern identification, with applications to microarray time-course experiments. *Journal of the American Statistical Association* **106**, 73–88.

Xia, Y., Cai, T. and Cai, T. T. (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika* **102**, 247–266.

Xia, Y., Cai, T. and Cai, T. T. (2018). Two-sample tests for high-dimensional linear regression with an application to detecting interactions. *Statistica Sinica* **28**, 63–92.

Xia, Y., Cai, T. T. and Sun, W. (2020). GAP: A general framework for information pooling in two-sample sparse inference. *Journal of the American Statistical Association* **115**, 1236–1250.

Xu, G., Lin, L., Wei, P. and Pan, W. (2016). An adaptive two-sample test for high-dimensional means. *Biometrika* **103**, 609–624.

Zhang, C., Fan, J. and Yu, T. (2011). Multiple testing via FDR_L for large-scale imaging data. *The Annals of Statistics* **39**, 613–642.

Zhang, X. and Chen, J. (2022). Covariate adaptive false discovery rate control with applications to omics-wide multiple testing. *Journal of the American Statistical Association* **117**, 411–427.

Zheng, S., Cheng, G., Guo, J. and Zhu, H. (2019). Test for high dimensional correlation matrices. *The Annals of Statistics* **47**, 2887–2921.

(Received September 2022; accepted November 2023)