*The Annals of Statistics*
2011, Vol. 39, No. 1, 613–642
DOI: 10.1214/10-AOS848


# MULTIPLE TESTING VIA FDR$_L$ FOR LARGE-SCALE IMAGING DATA

BY CHUNMING ZHANG[1], JIANQING FAN[2] AND TAO YU

*University of Wisconsin-Madison, Princeton University and National University of Singapore*

The multiple testing procedure plays an important role in detecting the presence of spatial signals for large-scale imaging data. Typically, the spatial signals are sparse but clustered. This paper provides empirical evidence that for a range of commonly used control levels, the conventional FDR procedure can lack the ability to detect statistical significance, even if the $p$-values under the true null hypotheses are independent and uniformly distributed; more generally, ignoring the neighboring information of spatially structured data will tend to diminish the detection effectiveness of the FDR procedure. This paper first introduces a scalar quantity to characterize the extent to which the "*lack of identification phenomenon*" (LIP) of the FDR procedure occurs. Second, we propose a new multiple comparison procedure, called FDR$_L$, to accommodate the spatial information of neighboring $p$-values, via a local aggregation of $p$-values. Theoretical properties of the FDR$_L$ procedure are investigated under weak dependence of $p$-values. It is shown that the FDR$_L$ procedure alleviates the LIP of the FDR procedure, thus substantially facilitating the selection of more stringent control levels. Simulation evaluations indicate that the FDR$_L$ procedure improves the detection sensitivity of the FDR procedure with little loss in detection specificity. The computational simplicity and detection effectiveness of the FDR$_L$ procedure are illustrated through a real brain fMRI dataset.

**1. Introduction.** In many important applications, such as astrophysics, satellite measurement and brain imaging, the data are collected at spatial grid points, and a large-scale multiple testing procedure is needed for detecting the presence of spatial signals. For example, functional magnetic resonance imaging (fMRI) is a recent and exciting imaging technique that allows investigators to determine which areas of the brain are involved in a cognitive task. Since an fMRI dataset contains time-course measurements over voxels, the number of which is typically of the order of $10^4$–$10^5$, a multiple testing procedure plays an important role in detecting the regions of activation. Another example of important application of

Received May 2010; revised August 2010.
[1]Supported by NSF Grant DMS-07-05209 and Wisconsin Alumni Research Foundation.
[2]Supported by NIH R01-GM072611 and NSF Grant DMS-07-14554.
*AMS 2000 subject classifications.* Primary 62H35, 62G10; secondary 62P10, 62E20.
*Key words and phrases.* Brain fMRI, false discovery rate, median filtering, $p$-value, sensitivity, specificity.

multiple testing is to the diffusion tensor imaging, which intends to identify brain white matter regions [Le Bihan et al. (2001)].

In the seminal work, Worsley et al. (2002) proposed a Gaussian random field method which approximates the family-wise error rate (FWER) by modeling test statistics over the entire brain as a Gaussian random field. It has been found to be conservative in some cases [Nichols and Hayasaka (2003)]. Nichols and Hayasaka (2003) also discussed the use of permutation tests and their simulation studies showed that permutation tests tended to be more sensitive in finding activated regions. The false discovery rate (FDR) approach has become increasingly popular. The conventional FDR procedure offers the advantage of overcoming the conservativeness drawback of FWER, requiring fewer assumptions than random field based methods and being computationally less intensive than permutation tests.

Nevertheless, in practical applications to imaging data with a spatial structure, even if the $p$-values corresponding to the true null hypotheses are independent and uniformly distributed, the conventional FDR procedure may lack the ability to detect statistical significance, for a range of commonly used control levels $\alpha$. It will be seen, in the left panels of Figure 2, that the FDR procedure for a 2D simulated data declares only a couple of locations to be significant for $\alpha$ ranging from 0 to about 0.4. That is, even if we allow FDR to be controlled at the level 40%, one cannot reasonably well identify significant sites. The empirical evidence provided above for the standard FDR procedure is not pathological. Indeed, similar phenomena arise from commonly used signals plus noise models for imaging data, as will be exemplified by extensive studies in Section 4.2. In statistical literature, while some useful finite-sample and asymptotic results [Storey, Taylor and Siegmund (2004)] have been established for the FDR procedure, the results could not directly quantify the loss of power and "*lack of identification phenomenon*" (LIP).

More generally, for spatially structured imaging data, the significant locations are typically sparse, but clustered rather than scattered. It is thus anticipated that a location and its adjacent neighbors fall in a similar type of region, either significant (active) or nonsignificant (inactive). As will be seen in the simulation studies (where the LIP does not occur) of Section 5, the existing FDR procedure tends to be less effective in detecting significance. This lack of detection efficiency is due to the information of $p$-values from adjacent neighbors not having been fully taken into account. Due to the popularity of the FDR procedure in research practices, it is highly desirable to embed the spatial information of imaging data into the FDR procedure.

This paper aims to quantify the LIP and to propose a new multiple testing procedure, called $\mathrm{FDR}_L$, for imaging data, to accommodate the spatial information of neighboring $p$-values, via a local aggregation of $p$-values. Main results are given in three parts.

- In the first part, statistical inference for the null distribution of locally aggregated $p$-values is studied. See Method I proposed in Section 3.2 and Method II in Section 3.3.

- In the second part, asymptotic properties of the FDR$_L$ procedure are investigated under weak dependence (to be defined in Section 4.1) of $p$-values. See Theorems 4.1–4.3.
- The third part intends to provide a more in-depth discussion of why the LIP occurs and the extent to which the FDR$_L$ procedure alleviates the LIP. In particular, we introduce a scalar $\alpha_\infty$ to quantify the LIP: the smaller the $\alpha_\infty$, the smaller control level can be adopted without encountering LIP; $\alpha_\infty = 0$ rules out the possibility of the LIP. In the particular case of i.i.d. $p$-values, Theorem 4.4 provides verifiable conditions under which $\alpha_\infty = 0$ and under which $\alpha_\infty > 0$. Theorem 4.5 demonstrates that under mild conditions, $\alpha_\infty$ of the FDR$_L$ procedure is lower than the counterpart of the FDR procedure. These theoretical results demonstrate that the FDR$_L$ procedure alleviates the extent of the LIP, thus substantially facilitates the selection of user-specified control levels. As observed from the middle and right panels of Figure 2, for control levels close to zero, the FDR$_L$ procedure combined with either Method I or Method II identifies a larger number of true significant locations than the FDR procedure.

The rest of the paper is arranged as follows. Section 2 reviews the conventional FDR procedure and introduces $\alpha_\infty$ to characterize the LIP. Section 3 describes the proposed FDR$_L$ procedure. Its theoretical properties are established in Section 4, where Section 4.2 explores the extent to which the FDR$_L$ procedure alleviates LIP. Sections 5 and 6 present simulation comparisons of the FDR and FDR$_L$ procedures in 2D and 3D dependent data, respectively. Section 7 illustrates the computational simplicity and detection effectiveness of the proposed method for a real brain fMRI dataset for detecting the regions of activation. Section 8 ends the paper with a brief discussion. Technical conditions and detailed proofs are deferred to the Appendix.

## 2. FDR and lack of identification phenomenon.

2.1. *Conventional* FDR *procedure.* We begin with a brief overview of the conventional FDR procedure that is of particular relevance to the discussion in Sections 3 and 4. For testing a family of null hypotheses, $\{H_0(i)\}_{i=1}^n$, suppose that $p_i$ is the $p$-value of the $i$th test. Table 1 summarizes the outcomes.

TABLE 1
*Outcomes from testing n (null) hypotheses $H_0(i)$ based on a significance rule*

| | $H_0(i)$ retained | $H_0(i)$ rejected | Total |
|---|---|---|---|
| $H_0(i)$ true | $U$ | $V$ | $n_0$ |
| $H_0(i)$ false | $T$ | $S$ | $n_1$ |
| Total | $W$ | $R$ | $n$ |

Benjamini and Hochberg (1995) proposed a procedure that guarantees the False Discovery Rate (FDR) to be less than or equal to a pre-selected value. Here, the FDR is the expected ratio of the number of incorrectly rejected hypotheses to the total number of rejected hypotheses with the ratio defined to be zero if no hypothesis is rejected, that is, $\mathrm{FDR} = E(\frac{V}{R \vee 1})$ where $R \vee 1 = \max(R, 1)$. A comprehensive overview of the development of the research in the area of multiple testing can be found in Benjamini and Yekutieli (2001), Genovese and Wasserman (2002), Storey (2002), Dudoit, Shaffer and Boldrick (2003), Efron (2004), Storey, Taylor and Siegmund (2004), Genovese and Wasserman (2004), Lehmann and Romano (2005), Lehmann, Romano and Shaffer (2005), Genovese, Roeder and Wasserman (2006), Sarkar (2006), Benjamini and Heller (2007) and Wu (2008), among others. Fan, Hall and Yao (2007) addressed the issue on the number of hypotheses that can be simultaneously tested when the $p$-values are computed based on asymptotic approximations.

Storey, Taylor and Siegmund (2004) gave an empirical process definition of FDR, by

$$\mathrm{FDR}(t) = E\left\{\frac{V(t)}{R(t) \vee 1}\right\}, \tag{2.1}$$

where $t$ stands for a threshold for $p$-values. For realistic applications, Storey (2002) proposed the point estimate of $\mathrm{FDR}(t)$ by

$$\widehat{\mathrm{FDR}}(t) = \frac{W(\lambda)t}{\{R(t) \vee 1\}(1-\lambda)}, \tag{2.2}$$

where $\lambda \in (0, 1)$ is a tuning constant, and $W(t)$ is the number of nonrejections with a threshold $t$. The intuition of this will be explained in Section 3.4. The pointwise limit of $\widehat{\mathrm{FDR}}(t)$ under assumptions (7)–(9) of Storey, Taylor and Siegmund (2004) is

$$\widehat{\mathrm{FDR}}^{\infty}(t) = \frac{[\pi_0\{1 - G_0(\lambda)\} + \pi_1\{1 - G_1(\lambda)\}]t}{\{\pi_0 G_0(t) + \pi_1 G_1(t)\}(1-\lambda)}, \tag{2.3}$$

where $\pi_0 = \lim_{n\to\infty} n_0/n$, $\pi_1 = 1 - \pi_0$, and $\lim_{n\to\infty} V(t)/n_0 = G_0(t)$ and $\lim_{n\to\infty} S(t)/n_1 = G_1(t)$ are assumed to exist almost surely for each $t \in (0, 1]$. For a pre-chosen level $\alpha$, a data-driven threshold for $p$-values is determined by

$$t_\alpha(\widehat{\mathrm{FDR}}) = \sup\{0 \le t \le 1 : \widehat{\mathrm{FDR}}(t) \le \alpha\}. \tag{2.4}$$

A null hypothesis is rejected if the corresponding $p$-value is less than or equal to the threshold $t_\alpha(\widehat{\mathrm{FDR}})$. Methods (2.2) and (2.4) form the basis for the conventional FDR procedure.

2.2. *Proposed measure for lack of identification phenomenon.* Recall that the FDR procedure is essentially a threshold-based approach for multiple testing problems, where the data-driven threshold $t_\alpha(\widehat{\mathrm{FDR}})$ plays a key role. It is clearly seen from (2.4) that $t_\alpha(\widehat{\mathrm{FDR}})$ hinges on both the estimates $\widehat{\mathrm{FDR}}(t)$ devised, as well as the control level $\alpha$ specified.

Using (2.2), we observe that the corresponding $t_\alpha(\widehat{\mathrm{FDR}})$ is a nondecreasing function of $\alpha$. This indicates that for the FDR procedure, as $\alpha$ decreases below $\inf_{0<t\le 1}\widehat{\mathrm{FDR}}(t)$, the threshold $t_\alpha(\widehat{\mathrm{FDR}})$ will drop to zero and accordingly, the FDR procedure can only reject those hypotheses with $p$-values exactly equal to zero. We call this phenomenon "*lack of identification*."

To better quantify the "*lack of identification phenomenon*" (LIP), the limiting forms of $\widehat{\mathrm{FDR}}(t)$ as $n \to \infty$ will be examined.

DEFINITION 1. For estimation methods $\widehat{\mathrm{FDR}}(t)$ in (2.2), define

$$\alpha_\infty^{\mathrm{FDR}} = \inf_{0<t\le 1} \widehat{\mathrm{FDR}}^\infty(t),$$

where $\widehat{\mathrm{FDR}}^\infty(t)$ is defined in (2.3). Define the endurance by $E_{\mathrm{FDR}} = 1 - \alpha_\infty^{\mathrm{FDR}}$.

Notice that the existence of $\alpha_\infty^{\mathrm{FDR}} > 0$ implies the occurrence of the LIP: in real data applications with a moderately large number $n$ of hypotheses, the FDR procedure loses the identification capability when the control level $\alpha$ is close to or smaller than $\alpha_\infty^{\mathrm{FDR}}$. On the other hand, the case $\alpha_\infty^{\mathrm{FDR}} = 0$ rules out the possibility of the LIP. Henceforth, the smaller the $\alpha_\infty^{\mathrm{FDR}}$, the higher endurance of the corresponding $\widehat{\mathrm{FDR}}$, and the less likely the LIP happens. In other words, an FDR estimation approach with a higher endurance is more capable of adopting a smaller control level, thus reducing the extent of the LIP problem. We will revisit this issue in Section 4.2 after introducing the proposed FDR$_L$ procedure.

**3. Proposed FDR$_L$ procedure for imaging data.** Consider a set of spatial signals $\{\mu(v) : v \in \mathcal{V} \subseteq \mathbb{Z}^d\}$ in a 2D plane ($d = 2$) or a 3D space ($d = 3$), where $\mu(v) = 0$ for $v \in \mathcal{V}_0$, $\mu(v) \neq 0$ for $v \in \mathcal{V}_1$ and $\mathcal{V}_0 \cup \mathcal{V}_1 = \mathcal{V}$. Here $\mathcal{V}_0$ and $\mathcal{V}_1$ are unknown sets. A common approach for detecting the presence of the spatial signals consists of two stages. In the first stage, test the hypothesis

$$H_0(v) : \mu(v) = 0 \quad \text{versus} \quad H_1(v) : \mu(v) \neq 0$$

at each location $v$. The corresponding $p$-value is denoted by $p(v)$. In the second stage, a multiple testing procedure, such as the conventional FDR procedure, is applied to the collection, $\{p(v) : v \in \mathcal{V} \subseteq \mathbb{Z}^d\}$, of $p$-values.

In the second stage, instead of using the original $p$-value, $p(v)$, at each $v$, we propose to use a local aggregation of $p$-values at points located adjacent to $v$. We summarize the procedure as follows.

STEP 1. Choose a local neighborhood with size $k$.

STEP 2. At each grid point $v$, find the set $N_v$ of its neighborhood points, and the set $\{p(v') : v' \in N_v\}$ of the corresponding $p$-values.

STEP 3. At each grid point $v$, apply a transformation $f : [0,1]^k \mapsto [0,1]$ to the set of $p$-values in Step 2, leading to a "locally aggregated" quantity, $p^*(v) = f(\{p(v') : v' \in N_v\})$.

STEP 4. Determine a data-driven threshold for $\{p^*(v) : v \in \mathcal{V} \subseteq \mathbb{Z}^d\}$.

For notational clarity, we denote by $\{p_i^*\}_{i=1}^n$ the collection of "locally aggregated" $p^*$-values, $\{p^*(v) : v \in \mathcal{V} \subseteq \mathbb{Z}^d\}$. Likewise, the notation $U^*(t)$, $V^*(t)$, $T^*(t)$, $S^*(t)$, $W^*(t)$ and $R^*(t)$ can be defined as in Section 2, with $p_i$ replaced by $p_i^*$. For instance, $V^*(t) = \sum_{i=1}^n \mathrm{I}\{H_0(i) \text{ is true, and } p_i^* \le t\}$ and $R^*(t) = \sum_{i=1}^n \mathrm{I}(p_i^* \le t)$, with $\mathrm{I}(\cdot)$ an indicator function. Accordingly, the false discovery rate based on utilizing the locally aggregated $p_i^*$-values becomes

$$\mathrm{FDR}_L(t) = E\left\{\frac{V^*(t)}{R^*(t) \vee 1}\right\}. \tag{3.1}$$

As a comparison, $\mathrm{FDR}(t)$ in (2.1) corresponds to the use of the original $p$-values.

3.1. *Choice of neighborhood and choice of $f$.* As in Roweis and Saul (2000), the set of neighbors for each data point can be assigned in a variety of ways, by choosing the $k$ nearest neighbors in Euclidean distance, by considering all data points within a ball of fixed radius or by using some prior knowledge.

For the choice of the transformation function, $f$, one candidate is the median filter, applied to the neighborhood $p$-values, without having to specify particular forms of spatial structure. A discussion on other options for $f$ can be found in Section 8. Unless otherwise stated, this paper focuses on the median filtering.

3.2. *Statistical inference for $p^*$-values: Method* I. Let $G^*(\cdot)$ be the cumulative distribution function of a "locally aggregated" $p^*$-value corresponding to the true null hypothesis. Let $\widetilde{G}^*(\cdot)$ be the sample distribution of $\{p^*(v) : v \in \mathcal{V}_0\}$. Recall that the original $p$-value corresponding to the true null hypothesis is uniformly distributed on the interval $(0,1)$. In contrast, the distribution $G^*(\cdot)$ for a "locally aggregated" $p^*$-value is typically nonuniform. This indicates that a significance rule based on $p$-values is not directly applicable to the significance rule based on $p^*$-values. For the median operation $f$, we propose two methods for estimating $\widetilde{G}^*(\cdot)$. Method I is particularly useful for large-scale imaging datasets, whereas Method II is useful for data of limited resolution.

Method I is motivated from the observation: if the original $p$-values are independent and uniformly distributed on the interval $(0,1)$, then the median aggregated

$p^*$-value follows a Beta distribution. More precisely, if the neighborhood size $k$ is an odd integer, then the median aggregated $p^*$-value conforms to the

$$\text{Beta}\big((k+1)/2, (k+1)/2\big) \tag{3.2}$$

distribution [Casella and Berger (1990)]. If $k$ is an even integer, the median aggregated $p^*$-value is distributed as a random variable $(X+Y)/2$, where $(X, Y)$ has the joint probability density function $k!/\{(k/2-1)!\}^2 x^{k/2-1}(1-y)^{k/2-1}\mathrm{I}(0 < x < y < 1)$. Thus, as long as the resolution of the experiment data and imaging technique keeps improving, so that the proportion of boundary grid points (corresponding to those with neighborhood intersected with both $\mathcal{V}_0$ and $\mathcal{V}_1$) decreases and eventually shrinks to zero, $G^*(\cdot)$ will tend to the Beta distribution in (3.2).

Following this argument, if the original $p$-values corresponding to the true null hypotheses are independent and uniformly distributed [see, e.g., van der Vaart (1998), page 305], the median aggregated $p^*$-values corresponding to the true null hypotheses will approximately be symmetrically distributed about 0.5. Thus, assuming that the number of false null hypothesis with $p_i^* > 0.5$ is negligible, the total number of true null hypotheses, $n_0$, is approximately $2\sum_{i=1}^n \mathrm{I}(p_i^* > 0.5) + \sum_{i=1}^n \mathrm{I}(p_i^* = 0.5)$, and the number of true null hypotheses with $p^*$-values smaller than or equal to $t$ could be estimated by $\sum_{i=1}^n \mathrm{I}\{p_i^* \geq (1-t)\}$, for small values of $t$. Here, owing to the symmetry, we use the upper tail to compute the proportion to mitigate the bias caused by the data from the alternative hypotheses. Hence, $\widetilde{G}^*(t)$ can be estimated by the empirical distribution function,

$$\widehat{G}^*(t) = \begin{cases} \dfrac{\sum_{i=1}^n \mathrm{I}\{p_i^* \geq (1-t)\}}{2\sum_{i=1}^n \mathrm{I}(p_i^* > 0.5) + \sum_{i=1}^n \mathrm{I}(p_i^* = 0.5)}, \\ \qquad \text{if } 0 \leq t \leq 0.5, \\ 1 - \dfrac{\sum_{i=1}^n \mathrm{I}(p_i^* > t)}{2\sum_{i=1}^n \mathrm{I}(p_i^* > 0.5) + \sum_{i=1}^n \mathrm{I}(p_i^* = 0.5)}, \\ \qquad \text{if } 0.5 < t \leq 1. \end{cases} \tag{3.3}$$

A modification of the Glivenko–Cantelli theorem shows that $\sup_{0\leq t\leq 1}|\widehat{G}^*(t) - G^*(t)| = o(1)$ almost surely as $n \to \infty$. This method is distribution free, computationally fast and applicable when the $p^*$-values under the null hypotheses are not too skewedly distributed.

An alternative approach for approximating $\widetilde{G}^*(\cdot)$ is inspired by the central limit theorem. If the neighborhood size $k$ is reasonably large (e.g., $k \geq 5$ if the original $p$-values corresponding to the true null hypotheses are independent and uniformly distributed), then $\widetilde{G}^*(\cdot)$ could be approximated by a normal distribution centered at 0.5. This normal approximation scheme may be exploited in the situation (which rarely occurs, though) when the original $p$-values corresponding to the true null hypotheses are independent but asymmetric about 0.5 (when the null distribution function of the test statistic is discontinuous).

3.3. *Refined method for estimating* $\widetilde{G}^*(\cdot)$: *Method* II. More generally, we consider spatial image data of limited resolution. Recall the neighborhood size $k$ of a voxel $v$ in the paper includes one for $v$ itself. Let $\mathsf{n}_1(v)$ denote the number of points in $N_v$ that belong to $\mathcal{V}_1$. Thus for any grid point $v \in \mathcal{V}_0$, $\mathsf{n}_1(v)$ takes values $\{0, 1, \ldots, k-1\}$. Set

$$\theta_{n,j} = P\{\mathsf{n}_1(v) = j\}, \qquad Q_j^*(t) = P\{p^*(v) \le t | \mathsf{n}_1(v) = j\}.$$

Clearly, $\sum_{j=0}^{k-1} \theta_{n,j} = 1$. Therefore, the C.D.F. of $p^*(v)$ for a grid point $v \in \mathcal{V}_0$ is given by

$$G^*(t) = \theta_{n,0} Q_0^*(t) + \theta_{n,1} Q_1^*(t) + \cdots + \theta_{n,k-1} Q_{k-1}^*(t), \tag{3.4}$$

where $Q_0^*(t)$ corresponds to, for independent tests, the Beta distribution function in (3.2).

Likewise, we obtain

$$\widetilde{G}^*(t) = \sum_{j=0}^{k-1} \widetilde{\theta}_{n,j} \widetilde{Q}_j^*(t),$$

where $\widetilde{\theta}_{n,j} = \#\mathcal{V}_0^{(j)}/n_0$ is the proportion of $v \in \mathcal{V}_0$ with $j$ neighboring grid points in $\mathcal{V}_1$, and $\widetilde{Q}_j^*(t) = \sum_{v \in \mathcal{V}_0^{(j)}} \mathrm{I}\{p^*(v) \le t\}/\#\mathcal{V}_0^{(j)}$ is the sample distribution of $\{p^*(v) : v \in \mathcal{V}_0^{(j)}\}$, with $\#A$ denoting the number of elements in a set $A$ and $\mathcal{V}_0^{(j)} = \{v \in \mathcal{V}_0 : \mathsf{n}_1(v) = j\}$. Clearly, if the original $p$-values corresponding to the true null hypotheses are block dependent, then, by the Glivenko–Cantelli theorem, $\sup_{0 \le t \le 1} |\widetilde{G}^*(t) - G^*(t)| = o(1)$ almost surely, as $n \to \infty$.

We propose the following Method II to estimate $\widetilde{G}^*(t)$:

1. Obtain estimates $\widehat{n}_0$ and $\widehat{n}_1 = n - \widehat{n}_0$ of $n_0$ and $n_1$, respectively. One possible estimator of $n_0$ is $\widehat{n}_0 = \sum_{i=1}^n \mathrm{I}(p_i^* > \lambda)/\{1 - \widehat{G}^*(\lambda)\}$, for some tuning parameter $\lambda$.
2. Define $\widehat{\mathcal{V}}_1 = \{v \in \mathcal{V} : p^*(v) \le p^*_{(\widehat{n}_1)}\}$, where $\{p^*_{(i)}\}_{i=1}^n$ denote the order statistics of $\{p_i^*\}_{i=1}^n$. Define $\widehat{\mathcal{V}}_0 = \{v \in \mathcal{V} : p^*(v) > p^*_{(\widehat{n}_1)}\}$.
3. Set $\widehat{\mathcal{V}}_0^{(j)} = \{v \in \widehat{\mathcal{V}}_0 : \mathsf{n}_1(v) = j\}$. Estimate $\widetilde{\theta}_{n,j}$, $j = 0, \ldots, k-1$, by $\widehat{\theta}_{n,j} = \#\widehat{\mathcal{V}}_0^{(j)}/\widehat{n}_0$.
4. For $j = 0$, estimate $\widetilde{Q}_0^*(t)$ by $\widehat{Q}_0^*(t) = \widehat{G}^*(t)$, the estimator of $\widetilde{G}^*(t)$ by Method I in Section 3.2. To estimate $\widetilde{Q}_j^*(t)$, $j = 1, \ldots, k-1$, for each $v \in \widehat{\mathcal{V}}_0^{(0)}$, collect its neighborhood $p$-values, randomly exclude $j$ of them and obtain the set $D_j(v)$ for the remaining neighborhood $p$-values. Randomly sample $j$ grid points from $\widehat{\mathcal{V}}_1$ and collect their corresponding $p$-values in a set $A_j(v)$. Compute the median, $\widehat{p}_j^*(v)$, of $p$-values in $D_j(v) \cup A_j(v)$. Estimate $\widetilde{Q}_j^*(t)$ by $\widehat{Q}_j^*(t) = \sum_{v \in \widehat{\mathcal{V}}_0^{(0)}} \mathrm{I}\{\widehat{p}_j^*(v) \le t\}/\#\widehat{\mathcal{V}}_0^{(0)}$.

5. Combining (3.4), $\widetilde{G}^*(t)$ is estimated by $\widehat{G}_c^*(t) = \sum_{j=0}^{k-1} \widehat{\theta}_{n,j} \widehat{Q}_j^*(t)$.

3.4. *Significance rule for $p^*$-values.* Using the locally aggregated $p^*$-values, we can estimate FDR$_L(t)$ defined in (3.1) by either

$$\widehat{\text{FDR}}_L(t) = \frac{W^*(\lambda)\widehat{G}^*(t)}{\{R^*(t) \vee 1\}\{1 - \widehat{G}^*(\lambda)\}}, \tag{3.5}$$

using Method I, or

$$\widehat{\text{FDR}}_L(t) = \frac{W^*(\lambda)\widehat{G}_c^*(t)}{\{R^*(t) \vee 1\}\{1 - \widehat{G}_c^*(\lambda)\}}, \tag{3.6}$$

using Method II. The logic behind this estimate is the following. If we choose $\lambda$ far enough from zero, then the number of nonrejections, $W^*(\lambda)$, is roughly $U^*(\lambda)$. Using this, we have

$$V^*(\lambda) \approx n_0 \widetilde{G}^*(\lambda) \approx \{V^*(\lambda) + W^*(\lambda)\}\widetilde{G}^*(\lambda).$$

Solving the above equation suggests an estimate of $V^*(\lambda)$ by $W^*(\lambda)\widetilde{G}^*(\lambda)/\{1 - \widetilde{G}^*(\lambda)\}$. Now, using $V^*(t)/V^*(\lambda) \approx \widetilde{G}^*(t)/\widetilde{G}^*(\lambda)$, we obtain that at a threshold $t$, $V^*(t)$ can be estimated by $W^*(\lambda)\widetilde{G}^*(t)/\{1 - \widetilde{G}^*(\lambda)\}$. This together with the definition of FDR$_L(t)$ in (3.1) suggests the estimate in (3.5). Interestingly, in the particular case of $p_i^* \equiv p_i$ and $\widehat{G}^*(t) = t$ [or $\widehat{G}_c^*(t) = t$], $\widehat{\text{FDR}}_L(t)$ coincides with $\widehat{\text{FDR}}(t)$ defined in (2.2).

For a given control level $\alpha$, a null hypothesis is rejected if the associated $p^*$-value is smaller than or equal to the threshold,

$$t_\alpha(\widehat{\text{FDR}}_L) \equiv \sup\{0 \le t \le 1 : \widehat{\text{FDR}}_L(t) \le \alpha\}. \tag{3.7}$$

This data-driven threshold for $p^*$-values together with the point estimation method (3.5) [or (3.6)] for the false discovery rates comprises the proposed FDR$_L$ procedure.

## 4. Properties of the FDR$_L$ procedure.

4.1. *Asymptotic behavior.* This section explores the asymptotic behavior of the FDR$_L$ procedure under weak dependence of $p$-values. Technical assumptions are given in Condition A in the Appendix, where Conditions A1–A3 are similar to assumptions (7)–(9) of Storey, Taylor and Siegmund (2004). Thus the type of dependence in Condition A2 includes finite block dependence, and certain mixing dependence. Theorems 4.1–4.3 can be considered a generalization of Storey, Taylor and Siegmund (2004) from a single $p$-value to locally aggregating a number $k$ of $p$-values with $k > 1$.

Theorem 4.1 below reveals that the proposed estimator $\widehat{\mathrm{FDR}}_L(t)$ controls the $\mathrm{FDR}_L(t)$ simultaneously for all $t \geq \delta$ with $\delta > 0$, and in turn supplies a conservative estimate of $\mathrm{FDR}_L(t)$.

THEOREM 4.1. *Assume Condition A in Appendix A. For each* $\delta > 0$,

$$\lim_{n\to\infty} \inf_{t\geq\delta} \left\{ \widehat{\mathrm{FDR}}_L(t) - \frac{V^*(t)}{R^*(t) \vee 1} \right\} \geq 0$$

*and*

$$\lim_{n\to\infty} \inf_{t\geq\delta} \{\widehat{\mathrm{FDR}}_L(t) - \mathrm{FDR}_L(t)\} \geq 0$$

*with probability one.*

To show that the proposed $\widehat{\mathrm{FDR}}_L(t)$ asymptotically provides a strong control of $\mathrm{FDR}_L(t)$, we define

$$\widehat{\mathrm{FDR}}_L^{\infty}(t) = \frac{[\pi_0\{1 - G_0^*(\lambda)\} + \pi_1\{1 - G_1^*(\lambda)\}]G^{*\infty}(t)}{\{\pi_0 G_0^*(t) + \pi_1 G_1^*(t)\}\{1 - G^{*\infty}(\lambda)\}}, \tag{4.1}$$

which is the pointwise limit of $\widehat{\mathrm{FDR}}_L(t)$ under Condition A in Appendix A, where it is assumed that $\pi_0 = \lim_{n\to\infty} n_0/n$, and $\lim_{n\to\infty} V^*(t)/n_0 = G_0^*(t)$ and $\lim_{n\to\infty} S^*(t)/n_1 = G_1^*(t)$ exist almost surely for each $t \in (0, 1]$, and $G^{*\infty}(t) = \lim_{n\to\infty} G^*(t)$.

THEOREM 4.2. *Assume Condition A in Appendix A. If there is a* $t \in (0, 1]$ *such that* $\widehat{\mathrm{FDR}}_L^{\infty}(t) < \alpha$, *then* $\limsup_{n\to\infty} \mathrm{FDR}_L(t_\alpha(\widehat{\mathrm{FDR}}_L)) \leq \alpha$.

Theorem 4.3 states that the random thresholding rule $t_\alpha(\widehat{\mathrm{FDR}}_L)$ converges to the deterministic rule $t_\alpha(\widehat{\mathrm{FDR}}_L^{\infty})$.

THEOREM 4.3. *Assume Condition A in Appendix A. If* $\widehat{\mathrm{FDR}}_L^{\infty}(\cdot)$ *has a nonzero derivative at the point* $t_\alpha(\widehat{\mathrm{FDR}}_L^{\infty}) \in (0, 1)$, *then* $\lim_{n\to\infty} t_\alpha(\widehat{\mathrm{FDR}}_L) = t_\alpha(\widehat{\mathrm{FDR}}_L^{\infty})$ *holds almost surely.*

4.2. *Conditions for lack of identification phenomenon.*

DEFINITION 2. For estimation methods $\widehat{\mathrm{FDR}}_L(t)$ in (3.5) [or (3.6)], define

$$\alpha_{\infty}^{\mathrm{FDR}_L} = \inf_{0<t\leq 1} \widehat{\mathrm{FDR}}_L^{\infty}(t),$$

where $\widehat{\mathrm{FDR}}_L^{\infty}(t)$ is defined in (4.1).

Theorem 4.4 establishes conditions under which the LIP does or does not take place with the FDR and FDR$_L$ procedures. It will be seen that the conditions are characterized by the null and alternative distributions of the test statistics, without relying on the configuration of the neighborhood used in the FDR$_L$ procedure. Theorem 4.5 demonstrates that $\alpha_\infty^{\mathrm{FDR}} \geq \alpha_\infty^{\mathrm{FDR}_L}$ under mild conditions, thus the FDR$_L$ procedure reduces the extent of the LIP. For expository brevity, we assume the test statistics are independent, which can be relaxed.

THEOREM 4.4. *Let $\{T(v): v \in \mathcal{V} \subseteq \mathbb{Z}^d\}$ be the set of test statistics for testing the presence of the spatial signals $\{\mu(v): v \in \mathcal{V} \subseteq \mathbb{Z}^d\}$. Consider the one-sided testing problem,*

$$H_0(v): \mu(v) = 0 \quad \textit{versus} \quad H_1(v): \mu(v) > 0. \tag{4.2}$$

*For $j = 0$ and $j = 1$, respectively, assume that $T(v)$, corresponding to the true $H_j(v)$, are i.i.d. random variables having a cumulative distribution function $F_j$ with a probability density function $f_j$. Assume that the neighborhood size $k \geq 3$ used in the* FDR$_L$ *procedure is an odd integer and that the proportion of boundary grid points within $\mathcal{V}_0$ shrinks to zero, as $n \to \infty$, that is, $\lim_{n\to\infty} \#\mathcal{V}_0^{(0)}/n_0 = 1$, where $\mathcal{V}_0^{(0)} = \{v \in \mathcal{V}: \mu(v') = 0$ for any $v' \in N_v\}$. Assume Condition A1 in Appendix A. Let $x_0 = F_0^{-1}(1) = \inf\{t: F_0(t) = 1\}$.*

I. *If $\lim_{x\to x_0-} \frac{f_1(x)}{f_0(x)} = \infty$, then $\alpha_\infty^{\mathrm{FDR}} = 0$ and $\alpha_\infty^{\mathrm{FDR}_L} = 0$.*

II. *If $\limsup_{x\to x_0-} \frac{f_1(x)}{f_0(x)} < \infty$, then $\alpha_\infty^{\mathrm{FDR}} > 0$ and $\alpha_\infty^{\mathrm{FDR}_L} > 0$.*

THEOREM 4.5. *Assume the conditions in Theorem 4.4. Suppose that $f_0(\cdot)$ is supported in an interval; $f_1(x) \leq f_0(x)$ for any $x \leq F_0^{-1}(0.5)$; $1 - F_0(F_1^{-1}(0.5)) \leq \lambda \leq 0.5$. Then $\alpha_\infty^{\mathrm{FDR}} \geq \alpha_\infty^{\mathrm{FDR}_L}$.*

Corollaries 1 and 2 below provide concrete applications of Theorems 4.4 and 4.5. The detailed verifications are omitted.

COROLLARY 1. *Assume the conditions in Theorem 4.4. Suppose that the distribution $F_0$ is $N(0,1)$ and the distribution $F_1$ is $N(C, \sigma^2)$, where $\sigma \in (0, \infty)$ and $C \in (0, \infty)$ are constants.*

I. *If $\sigma \geq 1$, then $\alpha_\infty^{\mathrm{FDR}} = 0$ and $\alpha_\infty^{\mathrm{FDR}_L} = 0$.*

II. *If $0 < \sigma < 1$, then $\alpha_\infty^{\mathrm{FDR}} > 0$ and $\alpha_\infty^{\mathrm{FDR}_L} > 0$. Moreover, if $\exp\{-(C/\sigma)^2/2\}/\sigma \leq 1$ and $1 - F_0(C) \leq \lambda \leq 0.5$, then $\alpha_\infty^{\mathrm{FDR}} \geq \alpha_\infty^{\mathrm{FDR}_L}$.*

COROLLARY 2. *Assume the conditions in Theorem 4.4. Suppose that the distribution $F_0$ is that of a Student's $t_{d_0}$ variate with $d_0$ degrees of freedom and the distribution $F_1$ is that of $C$ plus a Student's $t_{d_1}$ variate with $d_1$ degrees of freedom, where $C \in (0, \infty)$ is a constant.*

I. *If $d_0 > d_1$, then $\alpha_\infty^{\mathrm{FDR}} = 0$ and $\alpha_\infty^{\mathrm{FDR}_L} = 0$.*

II. *If* $1 \le d_0 \le d_1$, *then* $\alpha_\infty^{\mathrm{FDR}} > 0$ *and* $\alpha_\infty^{\mathrm{FDR}_L} > 0$. *Moreover, if* $d_0 = d_1$ *and* $1 - F_0(C) \le \lambda \le 0.5$, *then* $\alpha_\infty^{\mathrm{FDR}} \ge \alpha_\infty^{\mathrm{FDR}_L}$.

REMARK 1. For illustrative simplicity, a one-sided testing problem (4.2) is focused upon. Two-sided testing problems can similarly be treated and we omit the details.

4.3. *An illustrative example of* $\alpha_\infty^{\mathrm{FDR}} > \alpha_\infty^{\mathrm{FDR}_L} > 0$. Consider a pixelated 2D image dataset consisting of $n = 50 \times 50$ pixels, illustrated in the left panel of Figure 1, where the black rectangles represent the true significant regions $\mathcal{V}_1$ with $n_1 = 0.16 \times n$ pixels and the white background serves as the true nonsignificant regions $\mathcal{V}_0$ with $n_0 = n - n_1$ pixels. The data are simulated from the model,

$$Y(i, j) = \mu(i, j) + \varepsilon(i, j), \qquad i, j = 1, \ldots, 50,$$

where the signals are $\mu(i, j) = 0$ for $(i, j) \in \mathcal{V}_0$, and $\mu(i, j) = C$ for $(i, j) \in \mathcal{V}_1$ with a constant $C \in (0, \infty)$, and the error terms $\{\varepsilon(i, j)\}$ are i.i.d. following the centered Exp(1) distribution. At each site $(i, j)$, the observed data $Y(i, j)$ is the (shifted) survival time and used as the test statistic for testing $\mu(i, j) = 0$ versus $\mu(i, j) > 0$. Clearly, all test statistics corresponding to the true null hypotheses are i.i.d. having the probability density function $f_0(x) = \exp\{-(x + 1)\}\mathrm{I}(x + 1 > 0)$; likewise, all test statistics in accordance with the true alternative hypotheses are i.i.d. having the density function $f_1(x) = \exp\{-(x + 1 - C)\}\mathrm{I}(x + 1 > C)$. It is easily seen that $x_0 = \infty$, and $\limsup_{x \to \infty} f_1(x)/f_0(x) = \exp(C) < \infty$. An appeal to Theorem 4.4 yields $\alpha_\infty^{\mathrm{FDR}} > 0$ and $\alpha_\infty^{\mathrm{FDR}_L} > 0$, and thus both the FDR and $\mathrm{FDR}_L$ procedures will encounter the LIP. Moreover, if $C > \log(2)$, $\exp(-C)/2 \le \lambda \le 0.5$ and the neighborhood size $k \ge 3$ is an odd integer, then sufficient conditions in Theorem 4.5 are satisfied and hence $\alpha_\infty^{\mathrm{FDR}} \ge \alpha_\infty^{\mathrm{FDR}_L}$.

Actual computations indicate that in this example, as long as $C > \log(4)$, $\alpha_\infty^{\mathrm{FDR}_L}$ is considerably smaller than $\alpha_\infty^{\mathrm{FDR}}$, indicating that the $\mathrm{FDR}_L$ procedure can adopt

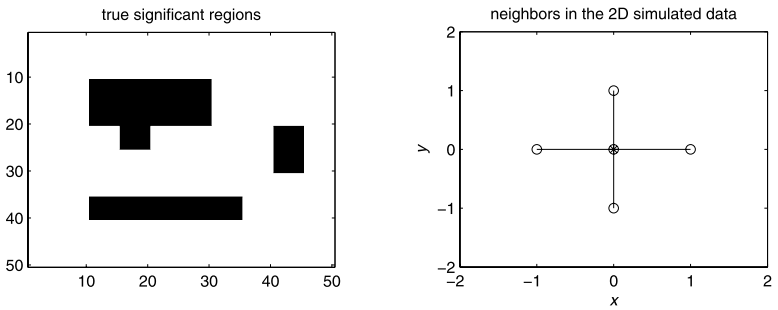


FIG. 1. *Left panel*: *the true significant regions for the* 2D *simulated data sets*. *Right panel*: *neighbors of a point at* $(x, y)$ *used in the* $\mathrm{FDR}_L$ *procedure for* 2D *simulated data*.

TABLE 2
*Comparing* $\alpha_\infty^{\mathrm{FDR}}$ *and* $\alpha_\infty^{\mathrm{FDR}_L}$

| $C$ | log(8) | log(12) | log(16) | log(20) | log(24) | log(28) | log(32) | log(36) |
|---|---|---|---|---|---|---|---|---|
| $\alpha_\infty^{\mathrm{FDR}}$ | 0.4130 | 0.3043 | 0.2471 | 0.2079 | 0.1795 | 0.1579 | 0.1409 | 0.1273 |
| $\alpha_\infty^{\mathrm{FDR}_L}$ | 0.0103 | 0.0030 | 0.0013 | 0.0007 | 0.0004 | 0.0002 | 0.0002 | 0.0001 |

a control level much smaller than that of the conventional FDR procedure without excessively encountering the LIP. For example, set $\lambda = 0.1$; assume that the neighborhood in the FDR$_L$ procedure is depicted in the right panel of Figure 1, that is, $k = 5$. Table 2 compares values of $\alpha_\infty^{\mathrm{FDR}}$ and $\alpha_\infty^{\mathrm{FDR}_L}$ for $C = \log(4j)$, $j = 2, \ldots, 9$. Refer to (C.2) and (C.5) in Appendix C for detailed derivations of $\alpha_\infty^{\mathrm{FDR}}$ and $\alpha_\infty^{\mathrm{FDR}_L}$, respectively.

To better visualize the LIP from limited data, Figure 2 compares the regions detected as significant by the FDR and FDR$_L$ procedures for $C = \log(8)$ based on one realization of the simulated data. It is observed from Figure 2 that for $\alpha$ between 0 and 0.4, the FDR procedure lacks the ability to detect statistical signif-

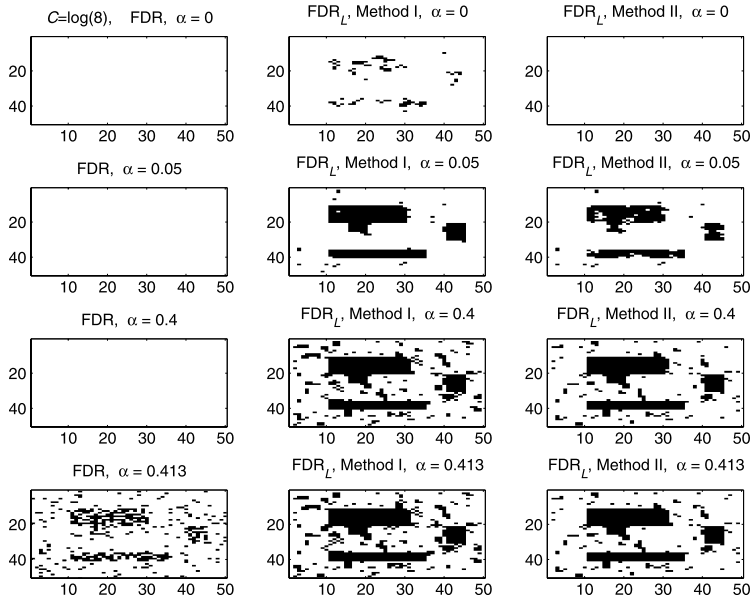


FIG. 2. *Lack of identification phenomenon when* $\alpha$ *varies from* 0 *to* $\alpha_\infty^{\mathrm{FDR}} = 0.4130$. *The sites that are called statistically significant based on the realization are shown in black. Left panels*: *the* FDR *procedure. Middle panels*: *the* FDR$_L$ *procedure using Method* I. *Right panels*: *the* FDR$_L$ *procedure using Method* II.

icance; as $\alpha$ increases to 0.413 (which is the limit $\alpha_\infty^{\mathrm{FDR}} = 0.413$ as calculated in Table 2) and above, some significant results emerge. In contrast, for $\alpha$ close to 0, both Method I and Method II for the $\mathrm{FDR}_L$ procedure are able to deliver some significant results. Similar plots to those in Figure 2 are obtained with other choices of $C$ and hence are omitted for lack of space.

## 5. Simulation study: 2D dependent data.

5.1. *Example* 1. To illustrate the distinction between the $\mathrm{FDR}_L$ and the conventional FDR procedures, we present simulation studies. The true significant regions are displayed as two black rectangles in the top left panel of Figure 3. The data are generated according to the model

$$Y(i,j) = \mu(i,j) + \varepsilon(i,j), \qquad i,j = 1,\ldots,258, \tag{5.1}$$

where the signals are $\mu(i,j) = 0$ for $(i,j) \in \mathcal{V}_0$, $\mu(i,j) = 4$ in the larger black rectangle and $\mu(i,j) = 2$ in the smaller black rectangle. The errors $\{\varepsilon(i,j)\}$ have zero-mean, unit-variance and are *spatially dependent*, by taking $\varepsilon(i,j) = \{e(i-1,j) + e(i,j) + e(i+1,j) + e(i,j-1) + e(i,j+1)\}/\sqrt{5}$, where $\{e(i,j)\}_{i,j=0}^{259}$ are i.i.d. $N(0,1)$. At each pixel $(i,j)$, $Y(i,j)$ is used as the test statistic for testing $\mu(i,j) = 0$ against $\mu(i,j) > 0$.

Both FDR and $\mathrm{FDR}_L$ procedures are preformed at a common control level 0.01, with the tuning constant $\lambda = 0.1$. In the $\mathrm{FDR}_L$ procedure, the neighborhood of a point at $(x,y)$ is taken as in the right panel of Figure 1. The histogram of the original $p$-values plotted in Figure 3(a) is flat except a sharp rise on the left border. The flatness is explained by the uniform distribution of the original $p$-values corresponding to the true null hypotheses, whereas the sharp rise is caused by the small $p$-values corresponding to the true alternative hypotheses. The histogram of the median aggregated $p^*$-values in Figure 3(c) shows a sharp rise at the left end and has a shape symmetric about 0.5. The approximate symmetry arises from the limit distribution of $p^*$-values corresponding to the true null hypotheses [see (3.2)], whereas the sharp rise is formed by small $p^*$-values corresponding to the true alternative hypotheses. Figures 3(b), (d) and (d′) manifest that the FDR procedure diminishes the effectiveness in detecting the significant regions than the $\mathrm{FDR}_L$ procedure, demonstrating that the $\mathrm{FDR}_L$ procedure more effectively increases the true positive rates. As a comparison, Figures 3(e), (f) and (f′) correspond to using the mean (other than median) filter for aggregating $p$-values. It is seen that the detections by the median and mean filters are very similar; but compared with the mean, the median better preserves the edge of the larger black rectangle between significant and nonsignificant areas. This effect gets more pronounced when $\alpha$ increases, lending support to the "edge preservation property" of the median.

To evaluate the performance of Method I and Method II in estimating $\widetilde{G}^*(t)$, the bottom panels of Figure 3 display the plots of $\widehat{G}^*(t)$ versus $\widetilde{G}^*(t)$ and $\widehat{G}_c^*(t)$

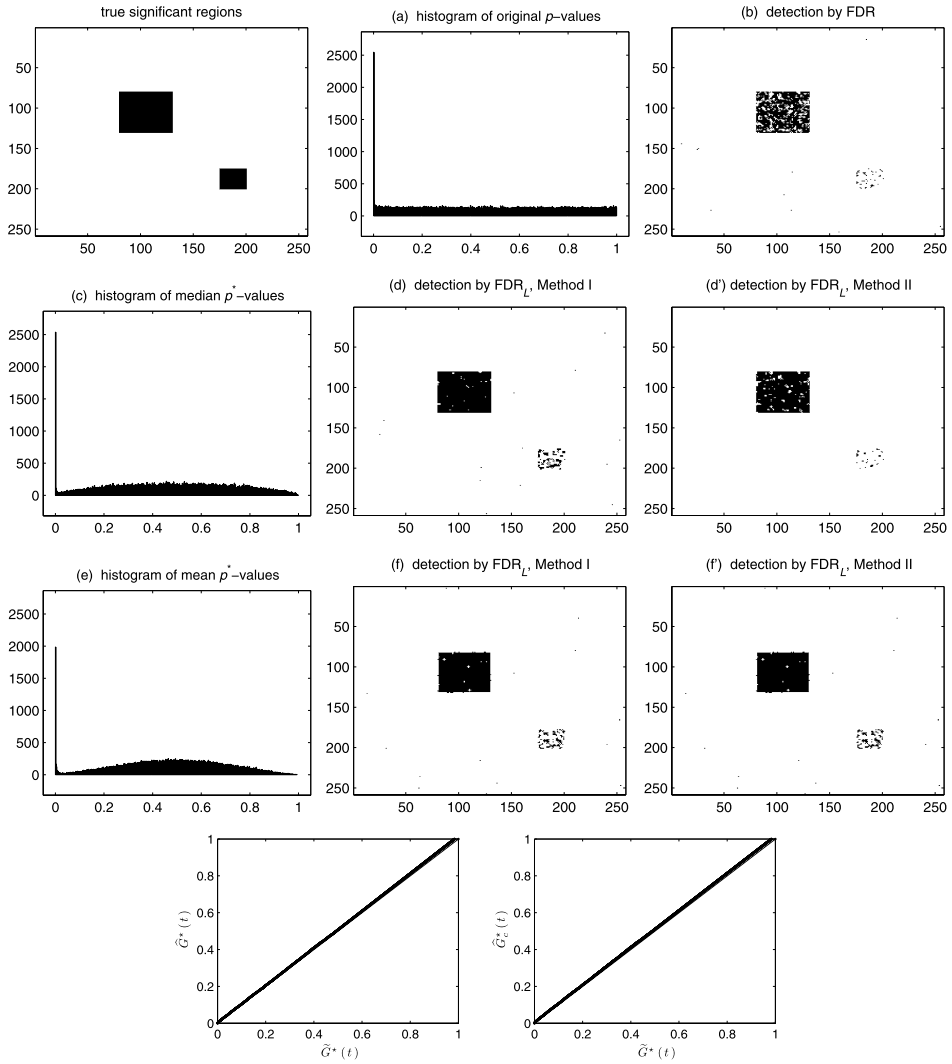


FIG. 3. *Comparison of the* FDR *and* FDR$_L$ *procedures for Example* 1. *In the first row, left*: *true significant regions shown in black*; *middle*: *histogram of the original p-values*; *right*: *significant regions detected by the* FDR *procedure. In the second row, left*: *histogram of the* $p^*$*-values using the median filter*; *middle and right*: *significant regions detected by the* FDR$_L$ *procedure using Methods* I *and* II, *respectively. In the third row, left*: *histogram of the* $p^*$*-values using the mean filter*; *middle and right*: *significant regions detected by the* FDR$_L$ *procedure using Methods* I *and* II, *respectively. In the bottom row, left*: $\widehat{G}^*(t)$ *versus* $\widetilde{G}^*(t)$; *right*: $\widehat{G}^*_c(t)$ *versus* $\widetilde{G}^*(t)$; *straight line*: *the* 45 *degree reference line. Here* $\alpha = 0.01$ *and* $\lambda = 0.1$.

versus $\widetilde{G}^*(t)$. The agreement with 45 degree lines well supports both estimation methods.

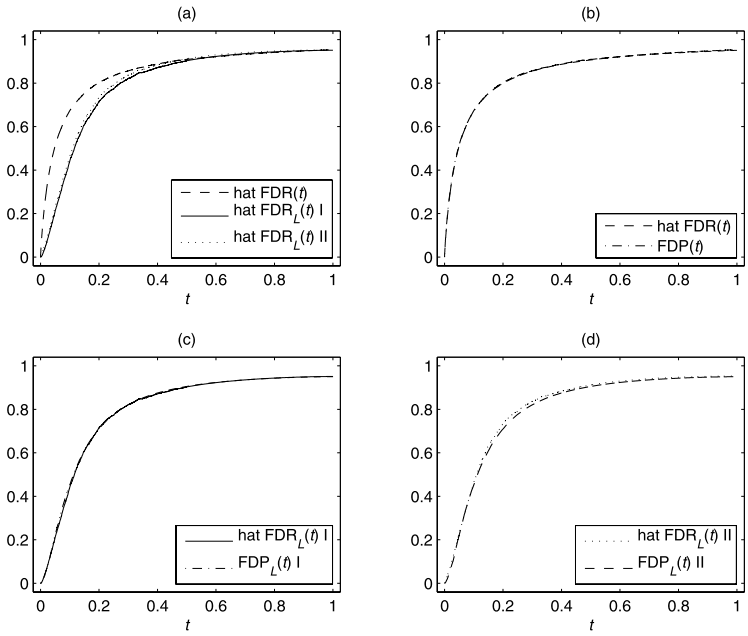


FIG. 4. *Panel* (a): *compare the average values of* $\widehat{\mathrm{FDR}}(t)$ *and those of* $\widehat{\mathrm{FDR}}_L(t)$ *using Methods* I *and* II. *Panel* (b): *compare the average values of* $\widehat{\mathrm{FDR}}(t)$ *and those of* $\mathrm{FDP}(t)$. *Panel* (c): *compare the average values of* $\widehat{\mathrm{FDR}}_L(t)$ *using Method* I *and those of* $\mathrm{FDP}_L(t)$. *Panel* (d): *compare the average values of* $\widehat{\mathrm{FDR}}_L(t)$ *using Method* II *and those of* $\mathrm{FDP}_L(t)$. *Here* $\lambda = 0.1$.

To examine the overall performance of the estimated $\mathrm{FDR}(t)$ and $\mathrm{FDR}_L(t)$ for a same threshold $t \in [0, 1]$, we replicate the simulation 100 times. For notational convenience, denote by $\mathrm{FDP}(t) = V(t)/\{R(t) \vee 1\}$ and $\mathrm{FDP}_L(t) = V^*(t)/\{R^*(t) \vee 1\}$ the false discovery proportions of the FDR and $\mathrm{FDR}_L$ procedures, respectively. The average values (over 100 data) of $\widehat{\mathrm{FDR}}(t)$ and $\widehat{\mathrm{FDR}}_L(t)$ at each point $t$ are plotted in Figure 4(a). It is clearly observed that $\widehat{\mathrm{FDR}}_L(t)$ using both Methods I and II is below $\widehat{\mathrm{FDR}}(t)$, demonstrating that the $\mathrm{FDR}_L$ procedure produces the estimated false discovery rates lower than those of the FDR procedure. Meanwhile, Figure 4 compares the average values of $\mathrm{FDP}(t)$ and those of $\widehat{\mathrm{FDR}}(t)$ in panel (b), and the average values of $\mathrm{FDP}_L(t)$ using Methods I and II and those of $\widehat{\mathrm{FDR}}_L(t)$ in panels (c) and (d), respectively. For each procedure, the two types of estimates are very close to each other, lending support to the estimation procedure in Section 3.4.

5.1.1. *Sensitivity and specificity.* To further study the relative performance of the FDR and $\mathrm{FDR}_L$ procedures, we adopt two widely used performance measures,

$$\text{sensitivity} \equiv \begin{cases} S(t_\alpha(\widehat{\mathrm{FDR}}))/n_1, & \text{for the FDR procedure,} \\ S^*(t_\alpha(\widehat{\mathrm{FDR}}_L))/n_1, & \text{for the } \mathrm{FDR}_L \text{ procedure,} \end{cases}$$

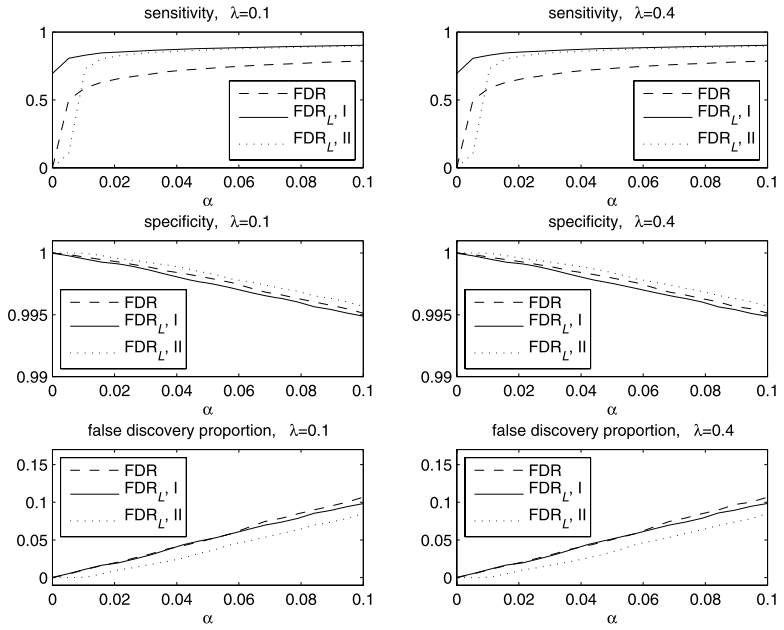


FIG. 5. *Comparison of the average sensitivity* (*top panels*), *average specificity* (*middle panels*) *and average false discovery proportion* (*bottom panels*). *Left panels*: $\lambda = 0.1$. *Right panels*: $\lambda = 0.4$.

$$\text{specificity} \equiv \begin{cases} U(t_\alpha(\widehat{\text{FDR}}))/n_0, & \text{for the FDR procedure,} \\ U^*(t_\alpha(\widehat{\text{FDR}}_L))/n_0, & \text{for the FDR}_L \text{ procedure,} \end{cases}$$

for summarizing the discriminatory power of a diagnosis procedure, where $S(t) = \sum_{i=1}^n \text{I}\{H_0(i) \text{ is false, and } p_i \le t\}$, $U(t) = \sum_{i=1}^n \text{I}\{H_0(i) \text{ is true, and } p_i > t\}$, $S^*(t) = \sum_{i=1}^n \text{I}\{H_0(i) \text{ is false, and } p_i^* \le t\}$ and $U^*(t) = \sum_{i=1}^n \text{I}\{H_0(i) \text{ is true, and } p_i^* > t\}$. Here, the sensitivity and specificity measure the strengths for correctly identifying the alternative and the null hypotheses, respectively.

Following Section 5.1, we randomly generate 100 sets of simulated data and perform FDR and FDR$_L$ procedures for each dataset, with the control levels $\alpha$ varying from 0 to 0.1. The left panel of Figure 5 corresponds to $\lambda = 0.1$, whereas the right panel corresponds to $\lambda = 0.4$. In either case, we observe that the average sensitivity (over the datasets) of the FDR$_L$ procedure using Method I is consistently higher than that of the FDR procedure, whereas the average specificities of both procedures approach one and are nearly indistinguishable. In addition, the bottom panels indicate that the FDR procedure yields larger (average) false discovery proportions than the FDR$_L$ procedure. It is apparent that the results in Figure 5 are not very sensitive to the choice of $\lambda$. Unless otherwise stated, $\lambda = 0.1$ will be used throughout the rest of the numerical work.

5.2. *Example* 2: *More strongly correlated case.* We consider a dataset generated according to the same model (5.1) as in Example 1, but with more strongly

correlated errors, by taking $\varepsilon(i, j) = \sum_{i=0}^{6} \sum_{j=0}^{6} e(i, j)/7$, where $\{e(i, j)\}_{i,j=0}^{264}$ are i.i.d. $N(0, 1)$. As seen from the figure in Zhang, Fan and Yu (2010), both FDR and $\text{FDR}_L$ (using Methods I and II) procedures perform worse with strongly-correlated data than with low-correlated data (given in Figure 3). However, there are no adverse effects by applying $\text{FDR}_L$ to more strongly correlated data, and Method I continues to be comparable with Method II for the $\text{FDR}_L$ procedure.

5.3. *Example* 3: *Large proportion of boundary grid points*. The efficacy of the $\text{FDR}_L$ procedure is illustrated in the figure of Zhang, Fan and Yu (2010) by a simulated dataset generated according to the same model (5.1) as in Example 1, but with a large proportion of boundary grid points, where $\mu(i, j) = 0$ for $(i, j) \in \mathcal{V}_0$ and $\mu(i, j) = 4$ for $(i, j) \in \mathcal{V}_1$. Similar plots using $\mu(i, j) = 2$ for $(i, j) \in \mathcal{V}_1$ are obtained and thus omitted. Again, there is no adverse effect of using $\text{FDR}_L$ to detect dense or weak signals.

**6. Simulation study: 3D dependent data.** We apply the FDR and $\text{FDR}_L$ procedures to detect activated brain regions of a simulated brain fMRI dataset, which is both spatially and temporally correlated. The experiment design, timings and size are exactly the same as those of the real fMRI dataset in Section 7. The data are generated from a semi-parametric model similar to that in Section 5.2 of Zhang and Yu (2008). (They demonstrated that the semi-parametric model gains more flexibilities than existing parametric models.) The left panel of Figure 6 contains 9 slices (corresponding to the 2D axial view) which highlight two activated brain regions involving 91 activated brain voxels. The neighborhood used in the $\text{FDR}_L$ procedure is illustrated in the right panel of Figure 6.

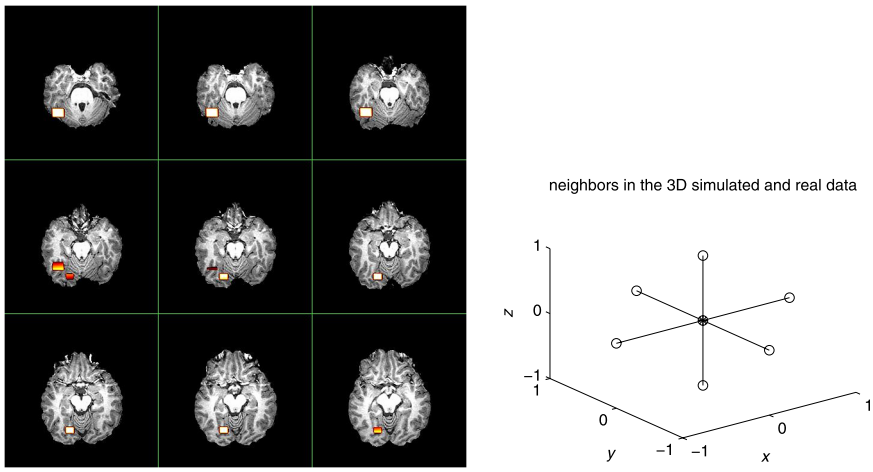


FIG. 6. *Left panel*: *true activated brain regions* (*denoted by hot color*) *for the simulated* fMRI *dataset*. *Right panel*: *neighbors of a point at* $(x, y, z)$ *used in the* $\text{FDR}_L$ *procedure for* 3D *simulated and real data*.

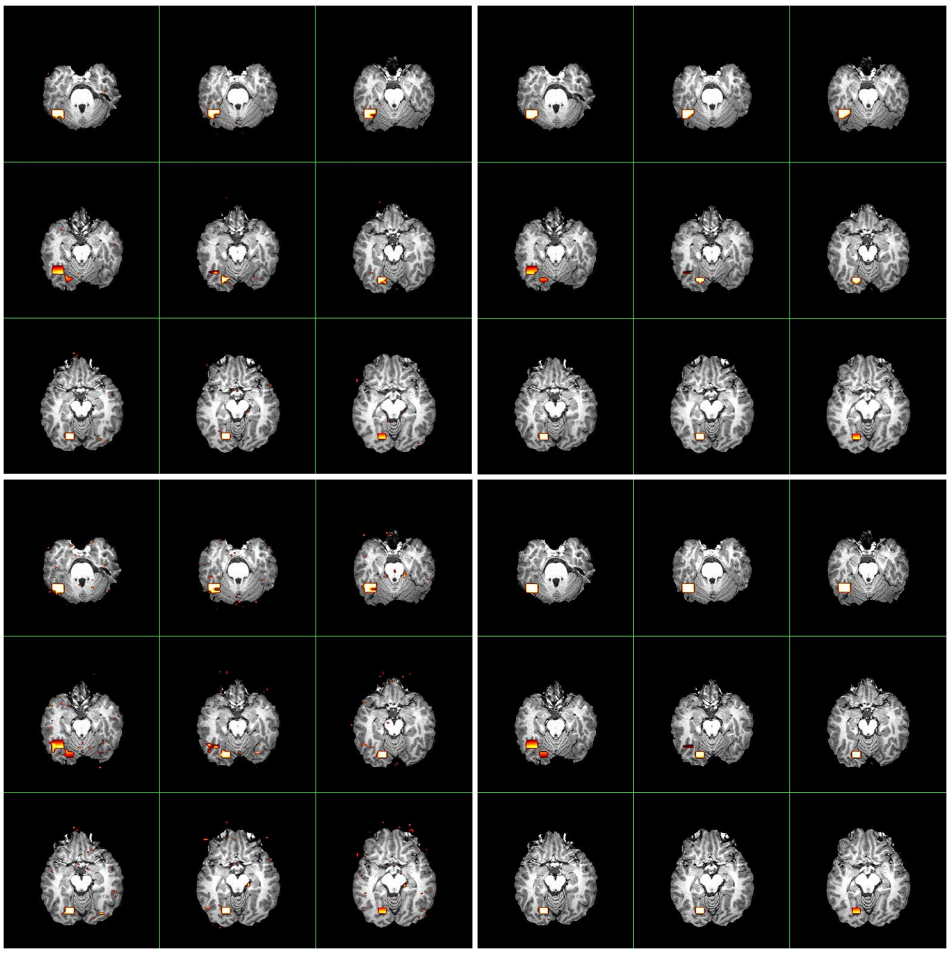

FIG. 7. *Comparison of activated brain regions detected for the simulated* fMRI *dataset using the conventional* FDR *approach* (*on the left*) *and the proposed* FDR$_L$ *procedure* (*on the right*) *using Method* I. *Top panels*: $\mathbb{K}$. *Bottom panels*: $\mathbb{K}_{\mathrm{bc}}$. *Here* $\alpha = 0.05$.

Figure 7 compares the activated brain regions identified by the FDR (in the left panels) and FDR$_L$ (in the right panels) procedures. Owing to the wealth of data, and for purposes of computational simplicity, results using Method I of FDR$_L$ are presented. Voxel-wise inactivity is tested with the semi-parametric test statistics $\mathbb{K} = (A\widehat{\mathbf{h}})^T\{A(\widetilde{\mathbf{S}}^T\widehat{R}^{-1}\widetilde{\mathbf{S}})^{-1}A^T\}^{-1}(A\widehat{\mathbf{h}})/\{\widehat{\mathbf{r}}^T\widehat{R}^{-1}\widehat{\mathbf{r}}/(n-rm)\}$ (in the top panels) and $\mathbb{K}_{\mathrm{bc}} = (A\widehat{\mathbf{h}}_{\mathrm{bc}})^T\{A(\widetilde{\mathbf{S}}^T\widehat{R}^{-1}\widetilde{\mathbf{S}})^{-1}A^T\}^{-1}(A\widehat{\mathbf{h}}_{\mathrm{bc}})/\{\widehat{\mathbf{r}}_{\mathrm{bc}}^T\widehat{R}^{-1}\widehat{\mathbf{r}}_{\mathrm{bc}}/(n-rm)\}$ (in the bottom panels) whose notation was given and asymptotic $\chi^2$ distributions were derived in Zhang and Yu (2008). The control level is 0.05. Inspection of Figure 7 reveals that $\mathbb{K}$ and $\mathbb{K}_{\mathrm{bc}}$ locate both active regions. In particular, using the FDR procedure, both methods detect more than 200 voxels (which are visible when

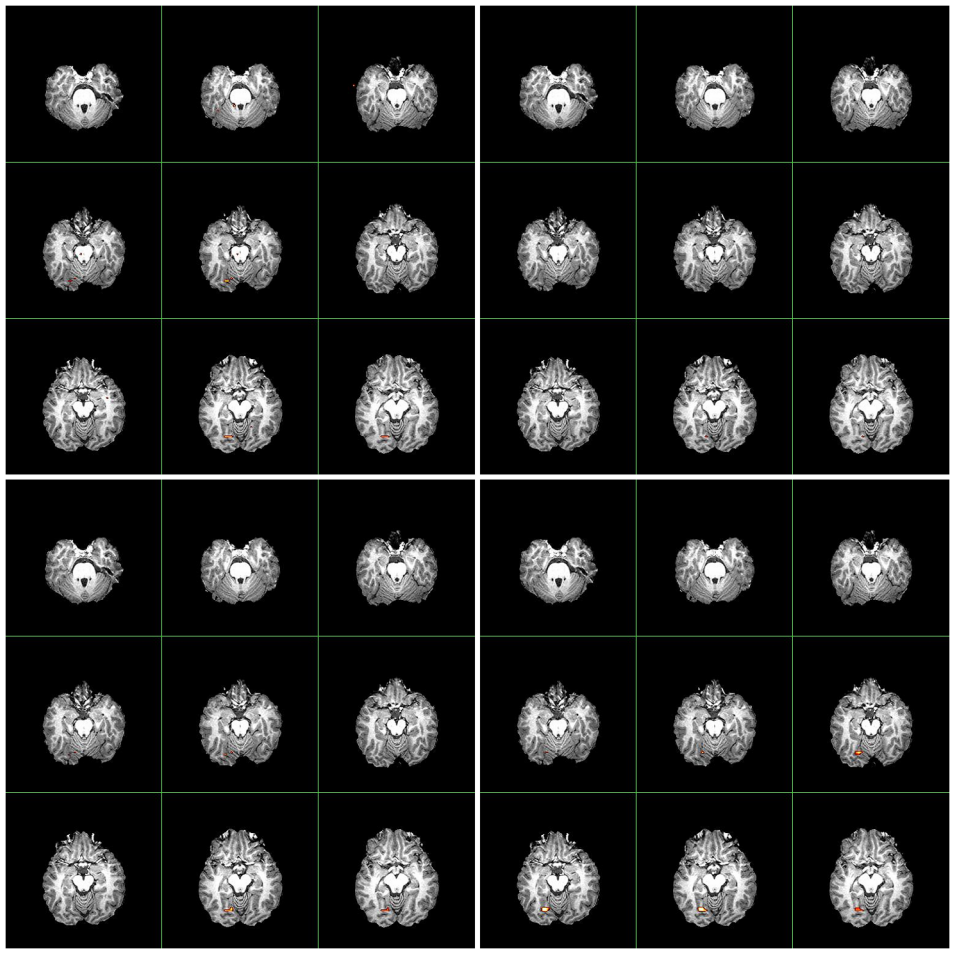

FIG. 8. *Comparison of activated brain regions detected for the simulated* fMRI *dataset using the conventional* FDR *approach* (*on the left*) *and the proposed* $\mathrm{FDR}_L$ *procedure* (*on the right*) *using Method* I. *Top panels*: AFNI. *Bottom panels*: FSL. *Here* $\alpha = 0.05$.

zooming the images), many of which are falsely discovered. When applying the $\mathrm{FDR}_L$ procedure, $\mathbb{K}$ detects 82 voxels, whereas $\mathbb{K}_{\mathrm{bc}}$ detects 90 voxels. Thus the $\mathrm{FDR}_L$ procedure reduces the number of tiny scattered false findings, gaining more accurate detections than the FDR procedure.

As a comparison, the detection results by popular software AFNI [Cox (1996)] and FSL [Smith et al. (2004) and Woolrich et al. (2001)] are given in Figure 8. We observe that both AFNI and FSL fail to locate one activated brain area, and that the other region, though correctly detected, has appreciably reduced size relative to the actual size. This detection bias is due to the stringent assumptions underlying AFNI and FSL in modeling fMRI data: the Hemodynamic Response Function

TABLE 3
*Comparing* FDR *and* FDR$_L$ *procedures*

| | | Test methods | | | |
|---|---|---|---|---|---|
| | **Multiple comparison** | $\mathbb{K}$ | $\mathbb{K}_{\mathbf{bc}}$ | **AFNI** | **FSL** |
| # of detected voxels | FDR | 276 | 870 | 16 | 6 |
| | FDR$_L$, Method I | 82 | 90 | 2 | 11 |
| False discovery proportion | FDR | 0.6993 | 0.9000 | 0.5625 | 0 |
| | FDR$_L$, Method I | 0 | 0 | 0.5000 | 0 |
| Sensitivity | FDR | 0.9121 | 0.9560 | 0.0769 | 0.0659 |
| | FDR$_L$, Method I | 0.9011 | 0.9890 | 0.0110 | 0.1209 |
| Specificity | FDR | 0.9921 | 0.9678 | 0.9996 | 1.0000 |
| | FDR$_L$, Method I | 1.0000 | 1.0000 | 0.9997 | 1.0000 |

(HRF) in FSL is specified as the difference of two gamma functions, and the drift term in AFNI is specified as a quadratic polynomial. As anticipated, applying the $F$ distributions restricted to parametric models to specify the distributions of test statistics in AFNI and FSL leads to bias, which in turn gives biased calculations of $p$-values and $p^*$-values. In this case, the detection performances of both the FDR and FDR$_L$ procedures deteriorate, and the FDR$_L$ procedure does not improve the performance of the FDR procedure. See Table 3 for a more detailed comparison.

To reduce modeling bias, for applications to the real fMRI dataset in Section 7, we will only employ the semi-parametric test statistics $\mathbb{K}$ and $\mathbb{K}_{\mathrm{bc}}$. It is also worth distinguishing between the computational aspects associated with the FDR$_L$ procedure: this paper uses (3.3) for the null distribution of $p^*$-values, whereas Zhang and Yu (2008) used the normal approximation approach in Section 3.2.

**7. Functional neuroimaging example.** In an emotional control study, subjects saw a series of negative or positive emotional images, and were asked to either suppress or enhance their emotional responses to the image, or to simply attend to the image. The sequence of trials was randomized. The time between successive trials also varied. The size of the whole brain dataset is $64 \times 64 \times 30$. At each voxel, the time series has 6 runs, each containing 185 observations with a time resolution of 2 seconds. For details of the dataset, please refer to Zhang and Yu (2008). The study aims to estimate the BOLD (Blood Oxygenation Level-Dependent) response to each of the trial types for 1–18 seconds following the image onset. We analyze the fMRI dataset containing one subject. The length of the estimated HRF is set equal to 18. Again, the neighborhood used in the FDR$_L$ procedure is illustrated in the right panel of Figure 6.

A comparison of the activated brain regions using the FDR and FDR$_L$ procedures is visualized in Figure 9. The level 0.001 is used to carry out the multiple comparisons. The conventional FDR procedure finds more tiny scattered active

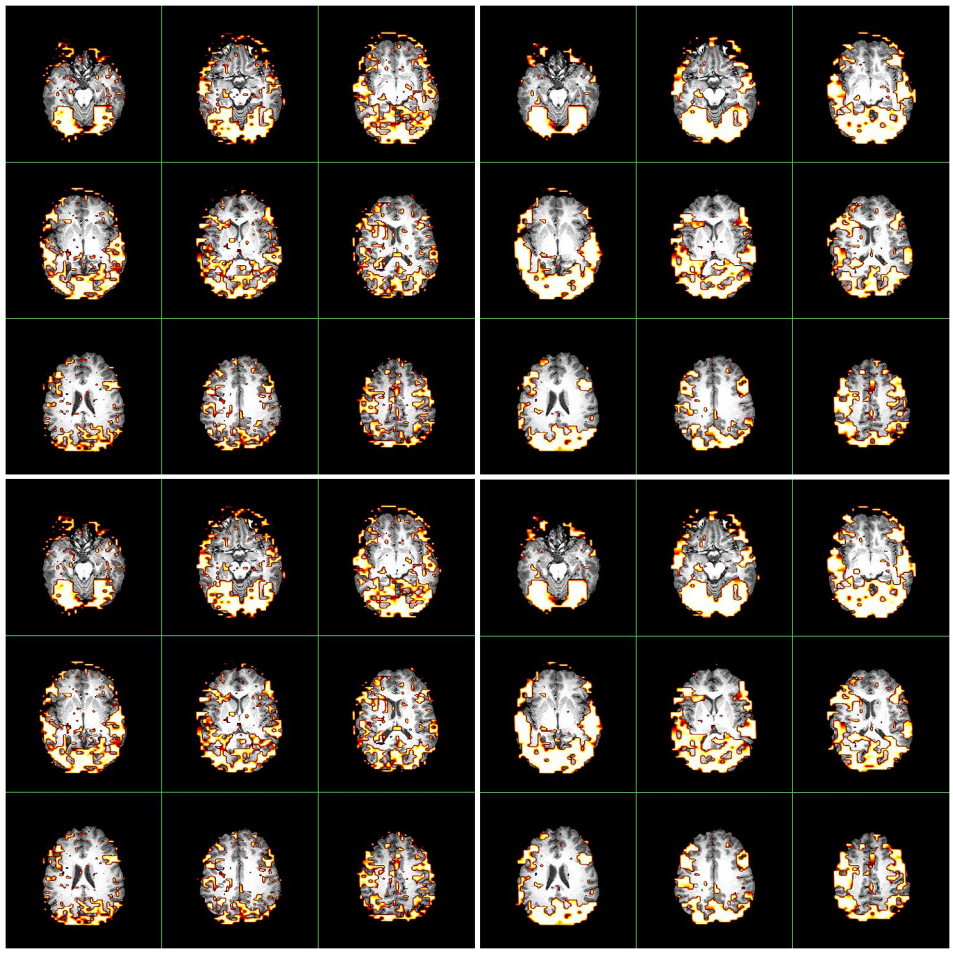

FIG. 9. *Comparison of activated brain regions detected for the real* fMRI *dataset using the conventional* FDR *approach* (*on the left*) *and the proposed* $\mathrm{FDR}_L$ *procedure* (*on the right*) *using Method* I. *Top panels*: $\mathbb{K}$. *Bottom panels*: $\mathbb{K}_{\mathrm{bc}}$. *Here* $\alpha = 0.001$.

voxels, which are more likely to be falsely discovered. In contrast, the $\mathrm{FDR}_L$ procedure finds activation in much more clustered regions of the brain.

**8. Discussion.** This paper proposes the $\mathrm{FDR}_L$ procedure to embed the structural spatial information of $p$-values into the conventional FDR procedure for large-scale imaging data with a spatial structure. This procedure provides the standard FDR procedure with the ability to perform better on spatially aggregated $p$-values. Method I and Method II have been developed for making statistical inference of the aggregated $p$-values under the null. Method I gains remarkable computational superiority, particularly for large/huge imaging datasets, when the

$p^*$-values under the null are not too skewed. Furthermore, we provide a better understanding of a "*lack of identification phenomenon*" (LIP) occurring in the FDR procedure. This study indicates that the FDR$_L$ procedure alleviates the extent of the problem and can adopt control levels much smaller than those of the FDR procedure without excessively encountering the LIP, thus substantially facilitating the selection of more stringent control levels.

As discussed in Owen (2005) and Leek and Storey (2008), a key issue with the dependencies between the hypotheses tests is the inflation of the variance of significance measures in FDR-related work. Indeed, similar to FDR, the FDR$_L$ procedure (using Methods I and II) performs less well with highly-correlated data than with the low-correlated data. Detailed investigation of the variance of FDR$_L$ will be given in future study.

Other ways of exploring spatially neighboring information are certainly possible in multiple comparison. For example, the median operation applied to $p$-values can be replaced by the averaging, kernel smoothing, "majority vote" and edge preserving smoothing techniques [Chu et al. (1998)]. Hence, taking the median is not the unique way to aggregate $p$-values. On the other hand, compared with the mean, the median is more robust, computationally simpler and does not depend excessively on the spatial co-ordinates, especially on the boundaries between significant and nonsignificant regions, as observed in Figures 3(d) and (f). An exhaustive comparison is beyond the scope of the current paper and we leave this for future research.

## APPENDIX A: PROOFS OF THEOREMS 4.1–4.3

We first impose some technical assumptions, which are not the weakest possible. Detailed proofs of Theorems 4.1–4.3 are given in Zhang, Fan and Yu (2010).

CONDITION A.

A0. The neighborhood size $k$ is an integer not depending on $n$.

A1. $\lim_{n\to\infty} n_0/n = \pi_0$ exists and $\pi_0 < 1$.

A2. $\lim_{n\to\infty} V^*(t)/n_0 = G_0^*(t)$ and $\lim_{n\to\infty} S^*(t)/n_1 = G_1^*(t)$ almost surely for each $t \in (0, 1]$, where $G_0^*$ and $G_1^*$ are continuous functions.

A3. $0 < G_0^*(t) \le G^{*\infty}(t)$ for each $t \in (0, 1]$.

A4. $\sup_{t\in(0,1]} |\widehat{G}^*(t) - G^{*\infty}(t)| = o(1)$ almost surely as $n \to \infty$.

## APPENDIX B: PROOFS OF THEOREMS 4.4 AND 4.5

**B.1. Proof of Theorem 4.4.** By the assumptions and $H_1(v)$, we see that the $p$-value has the expression, $p(v) = 1 - F_0(T(v))$. Thus, the distribution function of $p(v)$ corresponding to the true $H_0(v)$ is $G_0(t) = t$ for $0 < t < 1$ and (2.3) gives $\widehat{\mathrm{FDR}}^{\infty}(t) = \frac{\pi_0+\pi_1\{1-G_1(\lambda)\}/(1-\lambda)}{\pi_0+\pi_1 G_1(t)/t}$. Also, the distribution function of $p(v)$ corresponding to the true $H_1(v)$ is given by

$$G_1(t) = 1 - F_1\big(F_0^{-1}(1-t)\big). \tag{B.1}$$

Likewise, using (3.2), it follows that with probability one,

$$
\begin{aligned}
G_0^*(t) &= \lim_{n\to\infty} \frac{V^*(t)}{n_0} \\
&= \lim_{n\to\infty} \frac{\sum_{v\in\mathcal{V}_0^{(0)}} \mathrm{I}\{p^*(v)\le t\}}{\#\mathcal{V}_0^{(0)}} \cdot \lim_{n\to\infty} \frac{\#\mathcal{V}_0^{(0)}}{n_0} \\
&\quad + \lim_{n\to\infty} \frac{\sum_{v\in\mathcal{V}_0\setminus\mathcal{V}_0^{(0)}} \mathrm{I}\{p^*(v)\le t\}}{n_0} \qquad \text{(B.2)} \\
&= P\{p^*(v)\le t\} \qquad \text{with } v\in\mathcal{V}_0^{(0)} \\
&= G^{*\infty}(t) = B_{(k+1)/2,(k+1)/2}(t),
\end{aligned}
$$

the cumulative distribution function of a Beta$((k+1)/2, (k+1)/2)$ random variable and

$$
G_1^*(t) = \lim_{n\to\infty} S^*(t)/n_1 = B_{(k+1)/2,(k+1)/2}(G_1(t)). \qquad \text{(B.3)}
$$

Applying (B.2) and (4.1) gives $\widehat{\mathrm{FDR}}_L^{\infty}(t) = \frac{\pi_0+\pi_1\{1-G_1^*(\lambda)\}/\{1-G_0^*(\lambda)\}}{\pi_0+\pi_1 G_1^*(t)/G_0^*(t)}$.

Part I. For the FDR procedure, note that $\widehat{\mathrm{FDR}}^{\infty}(t)$ is a decreasing function of $G_1(t)/t$. Applying L'Hospital's rule and the fact $\lim_{t\to 0+} G_1(t) = 0$,

$$
\lim_{t\to 0+} \frac{G_1(t)}{t} = \lim_{t\to 0+} \frac{f_1(F_0^{-1}(1-t))}{f_0(F_0^{-1}(1-t))} = \lim_{x\to x_0-} \frac{f_1(x)}{f_0(x)} = \infty, \qquad \text{(B.4)}
$$

where $x = F_0^{-1}(1-t)$. Thus, $\sup_{0<t\le 1} G_1(t)/t = \infty$, which together with $\widehat{\mathrm{FDR}}^{\infty}(t)$ shows $\alpha_{\infty}^{\mathrm{FDR}} = 0$ for the FDR procedure.

For the $\mathrm{FDR}_L$ procedure, applying (B.2) and (B.3), we get

$$
\frac{dG_0^*(t)}{dt} = \frac{dG^{*\infty}(t)}{dt} = \frac{k!}{[\{(k-1)/2\}!]^2} t^{(k-1)/2}(1-t)^{(k-1)/2}, \qquad \text{(B.5)}
$$

$$
\frac{dG_1^*(t)}{dt} = \frac{k!}{[\{(k-1)/2\}!]^2} G_1(t)^{(k-1)/2}\{1-G_1(t)\}^{(k-1)/2} \frac{dG_1(t)}{dt}. \qquad \text{(B.6)}
$$

Note that $\widehat{\mathrm{FDR}}_L^{\infty}(t)$ is a decreasing function of $G_1^*(t)/G_0^*(t)$. Since $\lim_{t\to 0+} G_1^*(t) = 0$ and $\lim_{t\to 0+} G_0^*(t) = 0$,

$$
\begin{aligned}
\lim_{t\to 0+} \frac{G_1^*(t)}{G_0^*(t)} &= \lim_{t\to 0+} \frac{dG_1^*(t)/dt}{dG_0^*(t)/dt} \\
&= \lim_{t\to 0+} \left\{ \frac{G_1(t)}{t} \cdot \frac{1-G_1(t)}{1-t} \right\}^{(k-1)/2} \frac{dG_1(t)}{dt}, \qquad \text{(B.7)}
\end{aligned}
$$

which together with (B.4) shows $\lim_{t\to 0+} G_1^*(t)/G_0^*(t) = \infty$. Thus,

$$
\sup_{0<t\le 1} G_1^*(t)/G_0^*(t) = \infty,
$$

that is, $\alpha_\infty^{\mathrm{FDR}_L} = 0$ for the FDR$_L$ procedure.

Part II. Following $\widehat{\mathrm{FDR}}^\infty(t)$ and $\widehat{\mathrm{FDR}}_L^\infty(t)$, we immediately conclude that $\alpha_\infty^{\mathrm{FDR}} \neq 0$ if

$$\sup_{0<t\le 1} G_1(t)/t < \infty, \tag{B.8}$$

and that $\alpha_\infty^{\mathrm{FDR}_L} \neq 0$ if

$$\sup_{0<t\le 1} G_1^*(t)/G_0^*(t) < \infty. \tag{B.9}$$

We first verify (B.8) for the FDR procedure. Assume (B.8) fails, that is, $\sup_{0<t\le 1} G_1(t)/t = \infty$. Note that for any $\delta > 0$, the function $G_1(t)/t$, for $t \in [\delta, 1]$, is continuous and bounded away from $\infty$, thus, $\sup_{0<t\le 1} G_1(t)/t = \infty$ only if there exists a sequence $t_1 > t_2 > \cdots > 0$, such that $\lim_{m\to\infty} t_m = 0$ and $\lim_{m\to\infty} G_1(t_m)/t_m = \infty$. For each $m$, recall that both $G_1(t)$ and $t$ are continuous on $[0, t_m]$, and differentiable on $(0, t_m)$. Applying Cauchy's mean-value theorem, there exists $\xi_m \in (0, t_m)$ such that $G_1(t_m)/t_m = \{G_1(t_m) - G_1(0)\}/(t_m - 0) = \frac{dG_1(t)}{dt}|_{t=\xi_m}$. Since $\lim_{m\to\infty} G_1(t_m)/t_m = \infty$, it follows that

$$\limsup_{t\to 0+} \frac{dG_1(t)}{dt} = \infty. \tag{B.10}$$

On the other hand, the condition $\limsup_{x\to x_0-} \frac{f_1(x)}{f_0(x)} < \infty$ indicates that

$$\limsup_{t\to 0+} \frac{dG_1(t)}{dt} = \limsup_{t\to 0+} \frac{f_1(F_0^{-1}(1-t))}{f_0(F_0^{-1}(1-t))} = \limsup_{x\to x_0-} \frac{f_1(x)}{f_0(x)} < \infty, \tag{B.11}$$

where $x = F_0^{-1}(1-t)$. Clearly, (B.11) contradicts (B.10).

Next, we show (B.9) for the FDR$_L$ procedure. Combining (B.7), (B.8) and (B.11), the result follows. This completes the proof.

**B.2. Proof of Theorem 4.5.** We first show Lemma 1.

LEMMA 1. *Let $B(t)$ be the cumulative distribution function of a* Beta$(a, a)$ *random variable, where $a > 1$ is a real number. Then* I *for $t \in (0, 0.5)$, $B(t)/t$ is a strictly increasing function and $B(t) < t$;* II *for $t \in (0.5, 1)$, $B(t) > t$;* III *for $t_1 \in (0, 0.5]$ and $t_2 \in [t_1, 1]$, $B(t_1)/t_1 \le B(t_2)/t_2$.*

PROOF. Let $\Gamma(\cdot)$ denote the Gamma function. It is easy to see that

$$B''(t) = \Gamma(2a)/\{\Gamma(a)\}^2(a-1)t^{a-2}(1-t)^{a-2}(1-2t). \tag{B.12}$$

To show part I, define $F_1(t) = B(t)/t$. Then $F_1'(t) = \{B'(t)t - B(t)\}/t^2$, where $\frac{d\{B'(t)t - B(t)\}}{dt} = B''(t)t$. For $t \in (0, 0.5)$, (B.12) indicates $B''(t) > 0$, that

is, $B'(t)t - B(t)$ is strictly increasing, implying $B'(t)t - B(t) > B'(0)0 - B(0) = 0$. Hence for $t \in (0, 0.5)$, $B(t)/t$ is strictly increasing, and therefore $B(t)/t < B(0.5)/0.5 = 1$.

For part II, define $F_2(t) = B(t) - t$. Then $F_2''(t) = B''(t)$. By (B.12), $B''(t) < 0$ for $t \in (0.5, 1)$, thus $F_2(t)$ is strictly concave, giving $F_2(t) > \max\{F_2(0.5), F_2(1)\} = 0$.

Last, we show part III. For $t_2 \in [t_1, 0.5]$, part I indicates that $B(t_1)/t_1 \le B(t_2)/t_2$; for $t_2 \in [0.5, 1]$, part II indicates that $B(t_2)/t_2 \ge 1$ which, combined with $B(t_1)/t_1 \le 1$ from part I, yields $B(t_1)/t_1 \le B(t_2)/t_2$. □

We now prove Theorem 4.5. It suffices to show that

$$\{1 - G_1(\lambda)\}/(1 - \lambda) \ge \{1 - G_1^*(\lambda)\}/\{1 - G_0^*(\lambda)\}, \tag{B.13}$$

$$\sup_{0<t\le 1} G_1(t)/t \le \sup_{0<t\le 1} G_1^*(t)/G_0^*(t). \tag{B.14}$$

Following (B.5) and (B.6), for $0 \le t \le 1$,

$$G_1^*(t) = G_0^*(G_1(t)). \tag{B.15}$$

Applying (B.15), (B.1), $1 - F_0(F_1^{-1}(0.5)) \le \lambda$ and part II of Lemma 1 yields $G_1(\lambda) \le G_1^*(\lambda)$; applying $\lambda \le 0.5$ and part I of Lemma 1 implies $\lambda \ge G_0^*(\lambda)$. This shows (B.13).

To verify (B.14), let $M = \sup_{0<t\le 1} G_1(t)/t$. Since $G_1(1)/1 = 1$, we have $M \ge 1$ which will be discussed in two cases. Case 1: if $M = 1$, then

$$\sup_{0<t\le 1} \frac{G_1^*(t)}{G_0^*(t)} \ge \frac{G_1^*(1)}{G_0^*(1)} = 1 = \sup_{0<t\le 1} \frac{G_1(t)}{t}. \tag{B.16}$$

Case 2: if $M > 1$, then there exists $t_0 \in [0, 1]$ and $t_n \in (0, 1)$ such that $\lim_{n\to\infty} t_n = t_0$, and

$$\lim_{n\to\infty} G_1(t_n)/t_n = \sup_{0<t\le 1} G_1(t)/t = M > 1. \tag{B.17}$$

Thus, there exists $N_1$ such that for all $n > N_1$,

$$G_1(t_n) > t_n. \tag{B.18}$$

Cases of $t_0 = 1$, $t_0 = 0$ and $t_0 \in (0, 1)$ will be discussed separately. First, if $t_0 = 1$, then $M = \lim_{n\to\infty} G_1(t_n)/t_n = \lim_{n\to\infty} G_1(t_n) \le 1$, which contradicts (B.17). Thus $t_0 < 1$. Second, if $t_0 = 0$, then there exists $N_2$ such that $t_n < 0.5$ for all $n > N_2$. Thus for all $n > N \equiv \max\{N_1, N_2\}$, applying (B.15), (B.18) and part III of Lemma 1, we have that

$$\frac{G_1^*(t_n)}{G_1(t_n)} = \frac{G_0^*(G_1(t_n))}{G_1(t_n)} \ge \frac{G_0^*(t_n)}{t_n}.$$

This together with (B.17) shows

$$\sup_{0<t\le 1}\frac{G_1^*(t)}{G_0^*(t)} \ge \limsup_{n\to\infty}\frac{G_1^*(t_n)}{G_0^*(t_n)} \ge \lim_{n\to\infty}\frac{G_1(t_n)}{t_n} = M = \sup_{0<t\le 1}\frac{G_1(t)}{t}. \tag{B.19}$$

Third, for $t_0 \in (0,1)$, since both $F_0$ and $F_1$ are differentiable and $f_0$ is supported in a single interval, $G_1(t)/t = \{1 - F_1(F_0^{-1}(1-t))\}/t$ is differentiable in $(0,1)$. Thus,

$$\sup_{0<t\le 1} G_1(t)/t = G_1(t_0)/t_0 = M \tag{B.20}$$

and $\frac{d\{G_1(t)/t\}}{dt}|_{t=t_0} = 0$. Notice

$$\begin{aligned}\left.\frac{d\{G_1(t)/t\}}{dt}\right|_{t=t_0} &= \frac{(dG_1(t)/dt)|_{t=t_0} - G_1(t_0)/t_0}{t_0} \\ &= \frac{(dG_1(t)/dt)|_{t=t_0} - M}{t_0} = 0.\end{aligned} \tag{B.21}$$

If $t_0 > 0.5$, then $F_0^{-1}(1-t_0) \le F_0^{-1}(0.5)$. By (B.4) and the assumption on $f_0$ and $f_1$, $\frac{dG_1(t)}{dt}|_{t=t_0} = f_1(F_0^{-1}(1-t_0))/f_0(F_0^{-1}(1-t_0)) \le 1$, which contradicts (B.21). Thus, $0 < t_0 \le 0.5$. This together with (B.15), (B.20), and part III of Lemma 1 gives

$$\frac{G_1^*(t_0)}{G_1(t_0)} = \frac{G_0^*(G_1(t_0))}{G_1(t_0)} \ge \frac{G_0^*(t_0)}{t_0}.$$

This, together with (B.20), shows

$$\sup_{0<t\le 1}\frac{G_1^*(t)}{G_0^*(t)} \ge \frac{G_1^*(t_0)}{G_0^*(t_0)} \ge \frac{G_1(t_0)}{t_0} = M = \sup_{0<t\le 1}\frac{G_1(t)}{t}. \tag{B.22}$$

Combining (B.16), (B.19) and (B.22) completes the proof.

## APPENDIX C: $\alpha_\infty^{\mathrm{FDR}}$ AND $\alpha_\infty^{\mathrm{FDR}_L}$ IN TABLE 2 OF SECTION 4.3

Before calculating $\alpha_\infty^{\mathrm{FDR}}$ and $\alpha_\infty^{\mathrm{FDR}_L}$, we first present two lemmas.

LEMMA 2. *Let $f(x)$ and $g(x)$ be differentiable functions in $x \in (a,b) \subseteq \mathbb{R}$. Suppose that $g(x) \ne 0$ for $x \in (a,b)$, and $f(x)/g(x)$ is a nonincreasing function of $x$. For any $C \in (0,\infty)$ such that $g(x) + C \ne 0$, if $df(x)/dx \le dg(x)/dx$ for all $x \in (a,b)$, then $\{f(x)+C\}/\{g(x)+C\}$ is a decreasing function in $x \in (a,b)$.*

PROOF. The proof is straightforward and is omitted. □

LEMMA 3. *The function $h(x) = (10 - 15e^C x + 6e^{2C}x^2)/(10 - 15x + 6x^2)$ is decreasing in $x \in (0, e^{-C})$, for any constant $C \in (\log(4), \infty)$.*

PROOF. The function $h(x)$ can be rewritten as $h(x) = \{6(-e^C x + 5/4)^2 + 5/8\}/\{6(-x + 5/4)^2 + 5/8\}$. Note that $(-e^C x)/(-x) = e^C$ is nonincreasing in $x$ and $e^C > 1$ for $x > 0$. Applying Lemma 2, $(-e^C x + 5/4)/(-x + 5/4)$ is decreasing in $x \in (0, e^{-C})$, so is $(-e^C x + 5/4)^2/(-x + 5/4)^2$. When $C > \log(4)$, $d\{(-e^C x + 5/4)^2\}/dx \le d\{(-x + 5/4)^2\}/dx$. This together with Lemma 2 verifies that $h(x)$ is decreasing in $x \in (0, e^{-C})$. □

First, we evaluate $\alpha_\infty^{\mathrm{FDR}}$. From (B.1) and the conditions in Section 4.3,

$$\text{(C.1)} \qquad G_0(t) = t \quad \text{for } t \in [0,1] \quad \text{and} \quad G_1(t) = \begin{cases} te^C, & \text{if } t \in [0, e^{-C}], \\ 1, & \text{if } t \in (e^{-C}, 1]. \end{cases}$$

Thus $\sup_{0<t\le 1} G_1(t)/t = e^C$. By $\widehat{\mathrm{FDR}}^\infty(t)$ in Appendix B,

$$\text{(C.2)} \qquad \alpha_\infty^{\mathrm{FDR}} = \frac{\pi_0 + \pi_1\{1 - \lambda e^C \mathrm{I}(\lambda < e^{-C}) - \mathrm{I}(\lambda \ge e^{-C})\}/(1-\lambda)}{\pi_0 + \pi_1 e^C}.$$

Next, we compute $\alpha_\infty^{\mathrm{FDR}_L}$. Recall from Appendix B that the distribution $G_0^*(t)$ with $k = 5$ is that of a Beta(3, 3) random variable. Similarly, by (C.1), the distribution $G_1^*(t)$ is that of a Beta$(3,3)/e^C$ random variable. By $\widehat{\mathrm{FDR}}_L^\infty(t)$ in Appendix B, $\widehat{\mathrm{FDR}}_L^\infty(t)$ is a decreasing function of $G_1^*(t)/G_0^*(t)$, for which two cases need to be discussed. In the first case, $t \in (0, e^{-C}]$, it follows that

$$G_1^*(t)/G_0^*(t) = e^{3C}\frac{10 - 15 \cdot e^C t + 6 \cdot e^{2C} t^2}{10 - 15t + 6t^2},$$

which according to Lemma 3 is a decreasing function of $t$. Thus,

$$\sup_{t\in(0,e^{-C}]} G_1^*(t)/G_0^*(t) = \lim_{t\to 0+} G_1^*(t)/G_0^*(t) = e^{3C}$$

and

$$\text{(C.3)} \qquad \inf_{t\in(0,e^{-C}]} \widehat{\mathrm{FDR}}_L^\infty(t) = \frac{\pi_0 + \pi_1\{1 - G_1^*(\lambda)\}/\{1 - G_0^*(\lambda)\}}{\pi_0 + \pi_1 e^{3C}}.$$

In the second case, $t \in (e^{-C}, 1]$, since $G_1^*(t) = 1$, we observe from $\widehat{\mathrm{FDR}}_L^\infty(t)$ in Appendix B that $\widehat{\mathrm{FDR}}_L^\infty(t)$ is an increasing function of $G_0^*(t)$, and thus

$$\text{(C.4)} \qquad \inf_{t\in(e^{-C},1]} \widehat{\mathrm{FDR}}_L^\infty(t) = \frac{\pi_0 + \pi_1\{1 - G_1^*(\lambda)\}/\{1 - G_0^*(\lambda)\}}{\pi_0 + \pi_1/G_0^*(e^{-C})}.$$

Note that for $C > 0$, we have

$$\frac{1}{G_0^*(e^{-C})} = \frac{e^{3C}}{6(e^{-C} - 5/4)^2 + 5/8} \le \frac{e^{3C}}{6(1 - 5/4)^2 + 5/8} = e^{3C}.$$

Combining (C.3) and (C.4) gives

$$\text{(C.5)} \qquad \alpha_\infty^{\mathrm{FDR}_L} = \frac{\pi_0 + \pi_1\{1 - G_1^*(\lambda)\}/\{1 - G_0^*(\lambda)\}}{\pi_0 + \pi_1 e^{3C}}.$$

This completes the proof.

**Acknowledgments.** The comments of the anonymous referees, the Associate Editor and the Co-Editors are greatly appreciated.

## SUPPLEMENTARY MATERIAL

**Proofs and figures** (DOI: 10.1214/10-AOS848SUPP; .pdf). Section 1 gives detailed proofs of Theorems 4.1–4.3, Section 2 gives the figure in Section 5.2, and Section 3 gives the figure in Section 5.3.

## REFERENCES

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BENJAMINI, Y. and HELLER, R. (2007). False discovery rates for spatial signals. *J. Amer. Statist. Assoc.* **102** 1272–1281. MR2412549

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

CASELLA, G. and BERGER, R. L. (1990). *Statistical Inference*. Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, CA. MR1051420

CHU, C. K., GLAD, I., GODTLIEBSEN, F. and MARRON, J. S. (1998). Edge preserving smoothers for image processing (with discussion). *J. Amer. Statist. Assoc.* **93** 526–556. MR1631321

COX, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29** 162–173.

DUDOIT, S., SHAFFER, J. P. and BOLDRICK, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18** 71–103. MR1997066

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. MR2054289

FAN, J., HALL, P. and YAO, Q. (2007). To how many simultaneous hypothesis tests can normal, Student's $t$ or bootstrap calibration be applied? *J. Amer. Statist. Assoc.* **102** 1282–1288. MR2372536

GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 499–517. MR1924303

GENOVESE, C. R. and WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32** 1035–1061. MR2065197

GENOVESE, C. R., ROEDER, K. and WASSERMAN, L. (2006). False discovery control with $p$-value weighting. *Biometrika* **93** 509–524. MR2261439

LE BIHAN, D., MANGIN, J. F., POUPON, C., CLARK, C. A., PAPPATA, S., MOLKO, N. and CHABRIAT, H. (2001). Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging* **13** 534–546.

LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105** 18718–18723.

LEHMANN, E. L. and ROMANO, J. P. (2005). Generalizations of the familywise error rate. *Ann. Statist.* **33** 1138–1154. MR2195631

LEHMANN, E. L., ROMANO, J. P. and SHAFFER, J. P. (2005). On optimality of stepdown and stepup multiple test procedures. *Ann. Statist.* **33** 1084–1108. MR2195629

NICHOLS, T. and HAYASAKA, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Stat. Methods Med. Res.* **12** 419–446. MR2005445

OWEN, A. B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 411–426. MR2155346

ROWEIS, S. and SAUL, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326.

SARKAR, S. K. (2006). False discovery and false nondiscovery rates in single-step multiple testing procedures. *Ann. Statist.* **34** 394–415. MR2275247

SMITH, S., JENKINSON, M., WOOLRICH, M., BECKMANN, C. F., BEHRENS, T. E. J., JOHANSEN-BERG, H., BANNISTER, P. R., DE LUCA, M., DROBNJAK, I. FLITNEY, D. E., NIAZY, R. K., SAUNDERS, J., VICKERS, J., ZHANG, Y., DE STEFANO, N., BRADY, J. M. and MATTHEWS, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23** 208–219.

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 479–498. MR1924302

STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (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** 187–205. MR2035766

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge Univ. Press, Cambridge. MR1652247

WOOLRICH, M. W., RIPLEY, B. D., BRADY, M. and SMITH, S. M. (2001). Temporal autocorrelation in univariate linear modelling of FMRI data. *NeuroImage* **14** 1370–1386.

WORSLEY, K. J., LIAO, C. H., ASTON, J., PETRE, V., DUNCAN, G., MORALES, F. and EVANS, A. C. (2002). A general statistical analysis for fMRI data. *NeuroImage* **15** 1–15.

WU, W. B. (2008). On false discovery control under dependence. *Ann. Statist.* **36** 364–380. MR2387975

ZHANG, C. M. and YU, T. (2008). Semiparametric detection of significant activation for brain fMRI. *Ann. Statist.* **36** 1693–1725. MR2435453

ZHANG, C. M., FAN, J. and YU, T. (2010). Supplement to "Multiple testing via $\mathrm{FDR}_L$ for large scale imaging data." DOI: 10.1214/10-AOS848SUPP.

C. ZHANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
MADISON, WISCONSIN 53706
USA
E-MAIL: cmzhang@stat.wisc.edu

J. FAN
DEPARTMENT OF OPERATION RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: jqfan@princeton.edu

T. YU
DEPARTMENT OF STATISTICS
AND APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE 117546
E-MAIL: stayt@nus.edu.sg