Spatially Adaptive False Discovery Rate Thresholding for Sparse Estimation

Jiajun Luo[1,7], Zeyu Yao[2,7], Yunjin Choi[3], Bowen Gang[4],
Gourab Mukherjee[5], Wenguang Sun[6]

Abstract

Thresholding is an essential tool for sparse estimation, but the optimal threshold depends on the unknown sparsity level. Existing sparsity-adaptive thresholding strategies, exemplified by the false discovery rate (FDR) thresholding approach in Abramovich et al. (2006), can adapt between sparse and dense regimes. However, they suffer from using a universal threshold for all coordinates, which may not capture varying sparsity levels across different spatial regions. We propose a spatially adaptive FDR thresholding (SAFT) approach, which is capable of adapting to *the local sparsity level* via the *weighted FDR analysis*. SAFT is simple and intuitive, and allows the flexibility to set varied thresholds adaptively by accounting for the spatial inhomogeneity in sparsity levels. The advantages of SAFT are illustrated through numerical analyses.

Keywords: Structured multiple testing; Spatial analysis; FDR-thresholding; weighted BH procedure.

[1]Department of Data Sciences and Operations, University of Southern California.
[2]Center for Data Science, School of Mathematical Sciences, Zhejiang University.
[3]Department of Statistics, University of Seoul.
[4]Department of Statistics and Data Science, Fudan University.
[5]Department of Data Sciences and Operations, University of Southern California.
[6]Center for Data Science, Zhejiang University. Corresponding Email: wgsun@zju.edu.cn.
[7]Joint first authors

1 Introduction

Thirty years after its introduction, the false discovery rate (FDR) framework (Benjamini and Hochberg, 1995) has become an indispensable tool for scientific discovery in large-scale data analysis. By controlling the expected proportion of false positives among rejected hypotheses, the FDR offers a practical and powerful alternative to the familywise error rate, enabling researchers to explore massive datasets without being overwhelmed by false discoveries. The influence of FDR methodology is evident across diverse domains: identifying differentially expressed genes and regulatory elements in genomics, detecting activated brain regions in neuroimaging, pinpointing celestial objects from noisy surveys in astronomy, and locating pollution hotspots or disease clusters in environmental studies.

From a decision-theoretic perspective (Abramovich et al., 2006; Sun and Cai, 2007; Xiang et al., 2026), the FDR criterion provides a principled trade-off between the cost of false positives and the benefit of true discoveries. It controls the rate of false positives while simultaneously maintaining power to detect true signals—a balance that is essential for exploratory analysis. The legacy of the FDR, however, extends far beyond multiple testing. In this article, we focus on how its methodological principles have influenced *high-dimensional sparse estimation*. One of the most remarkable features of the FDR is its *inherent adaptivity* in the decision process: the tolerance for false positives is implicitly calibrated by the number of discoveries made. This self-regulating mechanism is precisely what makes the FDR a powerful tool for sparse estimation.

1.1 FDR: an Adaptive Criterion for Sparse Inference

Sparsity—the phenomenon that only a small fraction of coordinates contain meaningful signals while the vast majority are negligible or zero—is a fundamental characteristic of modern data-intensive applications. In genomic studies, only a limited number of genes are associated with a particular disease; in image processing, most pixels are background noise, with only a few carrying essential information; in environmental studies and disease mapping,

elevated risk or pollution levels are typically concentrated in localized areas, leaving most of the domain at baseline.

Coordinate-wise thresholding is a standard approach for estimating sparse mean vectors (Donoho and Johnstone, 1994; Johnstone and Silverman, 2004; Abramovich et al., 2006; Cai and Zhou, 2009). Classical results show that the optimal threshold must scale with the unknown sparsity level to achieve the minimax risk bound. When the degree of sparsity is known, a threshold calibrated to this parameter attains the optimal rate. In practice, however, the sparsity level is unknown. Conventional data-driven methods, such as minimizing Stein's unbiased risk estimator (SURE) (Donoho and Johnstone, 1995), can exhibit high variance in finite samples. Alternative sparsity-adaptive strategies leverage distributional structure to avoid manual tuning; for example, Johnstone and Silverman (2004) proposed an empirical Bayes thresholding rule that adapts between sparse and dense regimes.

The FDR criterion offers a particularly elegant solution to this challenge. Its inherent adaptivity automatically adjusts the stringency of the testing rule in response to the underlying sparsity level. When true signals are abundant, many hypotheses are rejected, and the FDR procedure permits a higher number of false positives, leading to a less stringent threshold; when signals are sparse, the rejection count remains small, and the threshold tightens accordingly. The seminal work of Abramovich et al. (2006) formalized the bridge between testing and estimation by demonstrating that an FDR-driven thresholding rule adapts automatically to the unknown degree of sparsity, achieving asymptotic minimaxity across a broad range of parameter spaces without requiring manual specification of the sparsity parameter. The same idea has been explored in the context of high-dimensional sparse regression. For example, Bogdan et al. (2015) introduced the sorted ℓ_1 penalized estimation procedure, which employs a decreasing sequence of penalty weights to adaptively shrink regression coefficients. The adaptivity embedded in the FDR definition is therefore not only a property of multiple testing but also a powerful principle for constructing flexible and effective estimators in high-dimensional sparse settings.

1.2 Sparse Inference with Side Information

The prevalence of massive datasets in modern data-intensive disciplines—genomics, neuroimaging, and signal processing, among others—is frequently accompanied by the availability of rich auxiliary side information. This information provides critical contextual knowledge and can be extracted from a wide array of sources. For instance, intrinsic data characteristics such as temporal or spatial ordering (Benjamini and Heller, 2007; Sun and Wei, 2011) and grouping or hierarchical structures (Efron, 2008; Yekutieli, 2008; Sun and Wei, 2015) are valuable sources. Additionally, external resources, including prior studies and domain-specific expertise, can be leveraged to obtain further insights (Roeder and Wasserman, 2009; Dobriban et al., 2015; Li et al., 2023).

Various approaches have been proposed to incorporate side information into FDR analysis. This extensively studied field has explored several important directions, including grouping-based methods (Cai and Sun, 2009; Barber and Ramdas, 2017; Ramdas et al., 2019), weighting-based methods (Genovese et al., 2006; Roquain and van de Wiel, 2009; Durand, 2019; Basu et al., 2018; Gang et al., 2026), as well as covariate-adaptive methods (Lei and Fithian, 2018; Li and Barber, 2019; Ignatiadis and Huber, 2021; Cai et al., 2019; Leung and Sun, 2022; Ren and Candès, 2023). In the specific context of spatial data, weighted p-values and corresponding FDR methods have been developed to enhance power by borrowing information from neighboring locations (Cai et al., 2022).

Returning to the problem of sparse estimation, we note that the FDR-thresholding frameworks described above (e.g., Abramovich et al., 2006) operate under the assumption of *homogeneous sparsity* across all coordinates, leading to a single, universal threshold. This assumption is frequently violated in spatial and structured data applications, where the sparsity level varies substantially across regions. In neuroimaging, for example, task-related activation is often concentrated in specific anatomical clusters, while other regions contain predominantly noise. In remote sensing, signal concentrations can range from dense to nearly absent across a landscape. Under such heterogeneity, a global threshold derived from overall

signal density is statistically mismatched: it tends to overshrink coordinates in dense regions (attenuating true effects) and undershrink coordinates in sparse regions (retaining excessive noise). Consequently, the resulting estimator incurs a higher mean squared error than a procedure that accounts for local variations in sparsity. This motivates the central goal of this article: to exploit recent advances in structured FDR analysis with side information into improved methods for sparse estimation. Specifically, we show how weighted p-value strategies, originally developed for multiple testing, can be adapted to construct spatially adaptive thresholds that yield more efficient estimation than global thresholding.

1.3 Our Proposal and Contributions

This article proposes a spatially adaptive FDR thresholding (SAFT) procedure for high-dimensional sparse estimation. SAFT adapts to local sparsity levels through a weighted FDR framework. By assigning location-dependent weights to the marginal p-values, the procedure translates the rejection count from a weighted Benjamini–Hochberg (BH) procedure into a set of coordinate-specific cutoffs. Specifically, SAFT applies lower thresholds in dense signal regions and higher thresholds in sparse noise regions. This spatial calibration attenuates noise in inactive areas while reducing bias in active regions, leading to improved mean squared error performance.

The SAFT framework integrates spatial information into a coordinate-wise shrinkage estimator without requiring parametric assumptions on the prior distribution of nonzero effects. Unlike block-thresholding methods (Cai and Zhou, 2009), it does not rely on pre-specified cluster boundaries or contiguous region definitions. Local sparsity is estimated via kernel smoothing, which borrows strength from neighboring coordinates while preserving spatial boundaries. The nominal FDR level is treated as a tuning parameter and selected through a modified cross-validation risk criterion.

Simulation studies in one and two dimensions indicate that SAFT reduces mean squared error relative to global thresholding and competing spatial estimators, particularly when the local sparsity levels vary across the spatial domain. The method is applied to a spatial

analysis of shared bicycle trips in Seoul, where it yields the lowest test-set prediction error among the procedures considered.

The remainder of the paper is organized as follows. Section 2 details the problem formulation and the connection between sparse estimation and FDR analysis. Section 3 introduces the SAFT procedure, including weight construction, local sparsity estimation, and parameter selection. Section 4 presents simulation studies in one and two dimensions. Section 5 describes the real-data application to Seoul bike-sharing data.

2 Sparse Estimation and FDR Analysis

2.1 Problem Formulation

Let $\mathcal{S} \subseteq \mathbb{R}^d$ denote a continuous spatial domain, and let $\mathbb{S} \subset \mathcal{S}$ be a finite discrete lattice of size $n = |\mathbb{S}|$ where observations are collected. We assume the observed data follow the normal means model

$$Y_s = \mu_s + \sigma z_s, \quad z_s \overset{i.i.d.}{\sim} N(0,1), \quad s \in \mathbb{S}. \tag{2.1}$$

It is conventionally assumed that $\sigma > 0$ is either known or can be accurately estimated from available data (Abramovich et al., 2006; Brown and Greenshtein, 2009; Weinstein et al., 2018; Xie et al., 2012). Without loss of generality, we set $\sigma = 1$.

To capture spatially varying sparsity, we adopt a location-specific hierarchical prior for the mean vector. For each $s \in \mathbb{S}$,

$$\theta_s \overset{ind}{\sim} \mathrm{Ber}(\pi_s), \quad \mu_s \mid \theta_s \sim (1-\theta_s)\delta_0(\cdot) + \theta_s g_{1s}(\cdot), \tag{2.2}$$

where $\pi_s = \mathbb{P}(\mu_s \neq 0)$ denotes the local sparsity level, δ_0 is a point mass at zero, and g_{1s} is an unspecified density for nonzero signals. Although the data is collected on the discrete set $\mathbb{S}$, we assume that π_s and g_{1s} are smooth in s over the continuous domain $\mathcal{S}$. This smoothness assumption implies that proximate locations share similar distributional characteristics, enabling the estimation of local sparsity via pooling of neighboring observations.

If spatial variation is ignored by assuming $\pi_s \equiv \pi$ and $g_{1s} \equiv g_1$ for all $s \in \mathbb{S}$, the marginal distribution for μ_s reduces to the standard two-group mixture $g(\mu) = (1-\pi)\delta_0(\mu) + \pi g_1(\mu)$. This formulation recovers the empirical Bayes two-group model (Efron et al., 2001), which underlies numerous compound estimation and multiple testing procedures (Storey, 2002; Johnstone and Silverman, 2004; Sun and Cai, 2007; Brown and Greenshtein, 2009; Jiang and Zhang, 2009).

The objective is to estimate the unknown mean vector $\boldsymbol{\mu} = (\mu_s)_{s\in\mathbb{S}}$ under the average mean squared error loss

$$\rho(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \frac{1}{n}\sum_{s\in\mathbb{S}} \mathbb{E}_{\mu_s}(\hat{\mu}_s - \mu_s)^2,$$

where $\hat{\boldsymbol{\mu}} = (\hat{\mu}_s)_{s\in\mathbb{S}}$ is an estimator constructed from the observed data $\{Y_s\}_{s\in\mathbb{S}}$.

2.2 Connection to FDR Analysis

Sparse estimation is intrinsically connected to signal detection. A natural strategy for recovering a high-dimensional mean vector is to identify coordinates with nonzero effects and shrink the remaining coordinates toward zero. This objective translates directly to a simultaneous multiple testing problem. For each location $s \in \mathbb{S}$, we test

$$H_{0,s} : \mu_s = 0 \quad \text{versus} \quad H_{1,s} : \mu_s \neq 0.$$

Under the normal means model with unit variance, the two-sided p-value for the s-th coordinate is $p_s = 2\tilde{\Phi}(|Y_s|)$, where $\tilde{\Phi}(u) = \int_u^\infty \phi(z)\, dz$ denotes the standard Gaussian survival function.

Abramovich et al. (2006) formalized the connection between this testing framework and sparse estimation by applying the Benjamini-Hochberg (BH) method to the collection of p-values. Let R denote the total number of rejected null hypotheses and V the number of false rejections. The FDR is defined as $\mathrm{FDR} = \mathbb{E}[V/\max(R,1)]$. The BH procedure controls the FDR at a nominal level q by comparing the ordered p-values $p_{(1)} \leq \cdots \leq p_{(n)}$ to the

linear boundary jq/n. The number of rejections is determined by

$$k = \max\left\{1 \le j \le n : p_{(j)} \le \frac{j}{n}q\right\},$$

with the convention that $\max(\emptyset) = 0$. The BH procedure is inherently adaptive to the unknown degree of sparsity. When signals are abundant, the empirical distribution of p-values places more mass near zero, causing the step-up boundary to intersect the ordered statistics at a larger index k. Conversely, in sparse configurations, k remains small. This data-driven scaling automatically adjusts the stringency of the testing rule without requiring prior knowledge of the global sparsity level.

The BH procedure determines a data-dependent p-value cutoff $(k/n)q$. Under the normal means model, the rejection rule $p_s \le (k/n)q$ is equivalent to $|Y_s| \ge \tilde{\Phi}^{-1}\{\frac{1}{2}(k/n)q\}$. This establishes a direct correspondence between the FDR-controlled p-value cutoff and a threshold on the observation scale. Translating this cutoff into the soft-thresholding framework, Abramovich et al. (2006) proposed the coordinate-wise estimator

$$\hat{\mu}_s[t_F] = Y_s + \eta(Y_s; t_F), \quad s \in \mathbb{S}, \tag{2.3}$$

where the soft-thresholding operator is $\eta(y;t) = \text{sign}(y)(|y| - t)_+$, and the FDR-driven threshold is

$$t_F = \tilde{\Phi}^{-1}\left\{\frac{1}{2}\min\left(\frac{k}{n}q, 1-\varepsilon\right)\right\}. \tag{2.4}$$

Here $\varepsilon > 0$ is a small stabilization constant. Theoretical analysis in Abramovich et al. (2006) established that this estimator achieves asymptotic minimaxity simultaneously over a broad range of sparse parameter spaces.

Despite the strong theoretical guarantees, the FDR-thresholding procedure in Abramovich et al. (2006) imposes a single, universal threshold t_F across all coordinates. This construction implicitly assumes that the underlying sparsity level is homogeneous throughout the domain.

In many spatial and structured data applications, signal density varies substantially across regions. A global threshold calibrated to the average sparsity level tends to overshrink coordinates in dense signal regions and undershrink coordinates in sparse noise regions. This mismatch degrades estimation accuracy when spatial heterogeneity is pronounced. To address this limitation, we introduce a spatially adaptive extension that replaces the universal cutoff with coordinate-specific thresholds calibrated to local sparsity estimates.

3 Methodology

3.1 Preview of the SAFT Estimator

We return to the multiple testing problem $H_{0,s} : \mu_s = 0$ versus $H_{1,s} : \mu_s \neq 0$ for each $s \in \mathcal{S}$. Under the modeling assumption that π_s varies smoothly across $\mathcal{S}$, the location index s provides substantive side information about the prior probability of a nonzero signal. Ignoring this structure forfeits potential gains in estimation efficiency. The procedure in Abramovich et al. (2006) employs the standard BH procedure to test these hypotheses, which treats all coordinates symmetrically and discards the location information.

One strategy to incorporate spatial side information is to use the weighted BH procedure of Genovese et al. (2006). Let $w_s > 0$ denote a location-specific weight, standardized such that $\sum_{s\in\mathcal{S}} w_s = n$. The weighted procedure operates on the transformed p-values $p_s^w = p_s/w_s$ and applies the standard step-up rule to the ordered sequence $p_{(1)}^w \leq \cdots \leq p_{(n)}^w$. The number of rejections is determined by

$$k_w = \max\left\{1 \leq j \leq n : p_{(j)}^w \leq \frac{j}{n}q\right\}.$$

The weighted BH procedure implies a location-dependent p-value cutoff. A hypothesis at location s is rejected if and only if $p_s^w \leq k_w q/n$, which is equivalent to $p_s \leq w_s k_w q/n$. Recall that under the normal means model, the two-sided p-value is $p_s = 2\tilde{\Phi}(|Y_s|)$. The rejection condition $p_s \leq w_s k_w q/n$ translates directly to $|Y_s| \geq \tilde{\Phi}^{-1}\{\frac{1}{2}w_s k_w q/n\}$. Thus, the

weighted p-value cutoff maps to a coordinate-specific threshold on the observation scale. Translating this boundary into the shrinkage framework yields the Spatially-Adaptive FDR Thresholding (SAFT) estimator. For each $s \in \mathcal{S}$, we define SAFT estimator as the soft-thresholding estimator $\hat{\mu}_s = Y_s + \eta(Y_s; \hat{t}_s)$ with threshold

$$\hat{t}_s(w_s, k_w, q) = \tilde{\Phi}^{-1}\left\{\frac{1}{2}\min\left(w_s \frac{k_w}{n} q, 1-\varepsilon\right)\right\}, \tag{3.5}$$

where $\varepsilon > 0$ is a stabilization constant. The SAFT estimator depends on two key components: the location-specific weights w_s and the nominal FDR level q. The weights govern the spatial adaptation by modulating thresholds according to local signal density, while q controls the overall stringency of the procedure. This parallels compound decision frameworks in which heterogeneous penalties for Type II errors naturally induce coordinate-specific rejection boundaries (He et al., 2015). We discuss the construction of w_s in Section 3.2 and the selection of q via modified cross-validation in Section 3.3.

3.2 Construction of Sparsity-adaptive Weights

The spatial adaptation in the SAFT estimator is governed by the location-specific weights w_s in (3.5). Since $\tilde{\Phi}^{-1}(\cdot)$ is a decreasing function, increasing the weight w_s reduces the resulting threshold $\hat{t}_s$. To apply smaller thresholds in dense signal regions and larger thresholds in sparse noise regions, the weights w_s must increase with the local sparsity level π_s.

Two established weighting schemes satisfy this monotonicity requirement. The first, employed in the SABHA procedure by Li and Barber (2019), takes the form

$$\tilde{w}_s^{sab} = \frac{1}{1-\pi_s}. \tag{3.6}$$

The second, motivated by an empirical Bayes formulation and used in the LAWS procedure by Cai et al. (2022), is given by

$$\tilde{w}_s^{law} = \frac{\pi_s}{1-\pi_s}. \tag{3.7}$$

Both weights are strictly increasing in π_s for $\pi_s \in (0,1)$. Numerical studies in structured multiple testing indicate that the LAWS weight often yields higher detection power (Cai et al., 2022), though neither formulation strictly dominates across all spatial configurations. We consider both schemes in the empirical analysis.

The local sparsity levels π_s are unknown and must be estimated from the data. We employ a kernel-smoothing estimator that borrows strength from neighboring coordinates (Cai et al., 2022). Let K be a positive, bounded, and symmetric kernel function, and let $K_h(u) = h^{-1}K(u/h)$ for a bandwidth $h > 0$. Define the normalized kernel weights $v_h(s, s') = K_h(\|s - s'\|)/K_h(0)$. Let $\mathcal{T}(\tau) = \{s \in \mathbb{S} : p_s > \tau\}$ denote the set of locations with p-values exceeding a screening threshold τ, typically set to 0.5. Under the assumption that null p-values are uniformly distributed, the expected proportion of nulls in $\mathcal{T}(\tau)$ within a local neighborhood of s is approximately $(1 - \pi_s)(1 - \tau)$. Equating the empirical and expected proportions yields the estimator

$$\hat{\pi}_s^{\tau} = 1 - \frac{\sum_{s' \in \mathcal{T}(\tau)} v_h(s, s')}{(1-\tau)\sum_{s' \in \mathbb{S}} v_h(s, s')}. \tag{3.8}$$

Substituting $\hat{\pi}_s^{\tau}$ for π_s in (3.6) produces data-driven initial weights $\tilde{w}_s$. To ensure compatibility with the weighted BH procedure, these weights are standardized so that $\sum_{s \in \mathbb{S}} w_s = n$. The final standardized weights are $w_s = n\tilde{w}_s / \sum_{s' \in \mathbb{S}} \tilde{w}_{s'}$, which are then substituted into (3.5) to complete the spatial calibration.

3.3 Selection of the FDR Level via Cross-Validation

In the sparse estimation framework, the nominal FDR level q functions as a tuning parameter rather than a target for inference. To select q adaptively, we employ a modified cross-validation (MCV) approach (Brown et al., 2013).

Let $\zeta_s \overset{i.i.d.}{\sim} N(0,1)$ for $s \in \mathbb{S}$ denote independent auxiliary variables, and let $\alpha > 0$ be a

fixed constant. Construct the paired observations

$$a_s = Y_s + \alpha\zeta_s \quad \text{and} \quad b_s = Y_s - \alpha^{-1}\zeta_s.$$

Conditional on μ_s, the variables a_s and b_s are independent, with $a_s \sim N(\mu_s, 1+\alpha^2)$ and $b_s \sim N(\mu_s, 1+\alpha^{-2})$. The sample $\boldsymbol{a} = \{a_s\}_{s\in\mathbb{S}}$ is used to construct the SAFT estimator, while $\boldsymbol{b} = \{b_s\}_{s\in\mathbb{S}}$ serves as an independent validation set. Let $\hat{\mu}_s(\boldsymbol{a}; q)$ denote the estimator computed from $\boldsymbol{a}$ using FDR level q. Following Brown et al. (2013), we define the validation loss

$$\hat{L}(q,\alpha) = \frac{1}{n}\sum_{s\in\mathbb{S}} \left(\hat{\mu}_s(\boldsymbol{a}; q) - b_s\right)^2. \tag{3.9}$$

Intuitively, $\mathbb{E}[\hat{L}(q,\alpha)]$ approximates the true mean squared error of $\hat{\boldsymbol{\mu}}$ up to an additive constant that does not depend on q. We select the data-driven FDR level $\hat{q}$ by minimizing $\hat{L}(q,\alpha)$ over a grid of candidate values in $(0, 1/2]$. This optimization is computationally efficient, as the weighted BH step and threshold calculations require only sorting and elementary arithmetic operations. The SAFT algorithm is summarized in Algorithm 1.

Algorithm 1 The SAFT algorithm

Require: $\{Y_s : s \in \mathbb{S}\}$, $\alpha > 0$, $\tau = 0.5$, a positive bounded symmetric kernel K and bandwidth $h > 0$.

Ensure: $\hat{\boldsymbol{\mu}} = \{\hat{\mu}_s : s \in \mathbb{S}\}$

1: Calculate $\hat{\pi}_s^{\tau}$ by (3.8) using the given kernel.

2: Obtain initial weights $\tilde{w}_s$ by substituting $\hat{\pi}_s^{\tau}$ into (3.6). Compute standardized weights $w_s = n\tilde{w}_s / \sum_{s'\in\mathbb{S}} \tilde{w}_{s'}$.

3: Obtain the rejection count k_w by the weighted BH-procedure.

4: Generate independent auxiliary variables $\zeta_s \sim N(0,1)$ for $s \in \mathbb{S}$. Construct $a_s = Y_s + \alpha\zeta_s$ and $b_s = Y_s - \alpha^{-1}\zeta_s$.

5: Obtain the FDR level $\hat{q}$ by minimizing

$$\hat{L}(q,\alpha) = \frac{1}{n}\sum_{s\in\mathbb{S}} \left(\hat{\mu}_s(\boldsymbol{a}; q) - b_s\right)^2,$$

where $\hat{\mu}_s(\boldsymbol{a}; q)$ are computed by substituting w_s, k_w, and q into (3.5) and (2.3).

6: Calculate thresholds $\hat{t}_s = \hat{t}_s(w_s, k_w, \hat{q})$ for $s \in \mathbb{S}$ according to (3.5).

7: Calculate the final estimate $\hat{\mu}_s = Y_s + \eta(Y_s; \hat{t}_s)$ for $s \in \mathbb{S}$.

3.4 Heuristic Justification via a Block-sparse Model

We now present a toy example to illustrate the benefits of location-specific thresholding under spatial heterogeneity. The construction highlights how a global threshold, calibrated to an average sparsity level, becomes suboptimal when signal density varies across regions, and how coordinate-specific thresholds adapt to these variations to reduce estimation error.

Consider the normal means model with unit noise. Let $\Theta_n(\pi) = \{\boldsymbol{\mu} : \|\boldsymbol{\mu}\|_0 \leq n\pi\}$. Classical theory (Abramovich et al., 2006) establishes that as $\pi \to 0$ and $n \to \infty$, the minimax risk under average mean squared error satisfies

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \Theta_n(\pi)} \rho(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) \sim 2n\pi \log(\pi^{-1}). \tag{3.10}$$

An estimator attaining this rate applies a threshold of $\sqrt{2\log(\pi^{-1})}$. When the sparsity parameter π is unknown, a data-driven procedure such as the FDR-thresholding rule of Abramovich et al. (2006) can be used to estimate the global cutoff.

To examine the impact of spatial structure, suppose the index set $\mathbb{S}$ is partitioned into M disjoint blocks $\{B_m\}_{m=1}^M$, with block sizes $n_m = |B_m|$. Within each block, the local sparsity level is constant, denoted $\pi^{(m)}$. The corresponding block-sparse parameter space is

$$\Theta_n^b = \left\{\boldsymbol{\mu} : \|\boldsymbol{\mu}^{(m)}\|_0 \leq n_m \pi^{(m)},\, 1 \leq m \leq M\right\}, \tag{3.11}$$

where $\boldsymbol{\mu}^{(m)}$ contains the means in B_m. Applying the global minimax result blockwise suggests that the optimal risk over Θ_n^b is asymptotically

$$R_n(\Theta_n^b) \sim \sum_{m=1}^{M} 2n_m \pi^{(m)} \log\left((\pi^{(m)})^{-1}\right). \tag{3.12}$$

If $\pi^{(m)} \equiv \pi$ for all m, the spatial structure provides no additional information and the global threshold is appropriate. When the $\pi^{(m)}$ differ, however, a single threshold calibrated

to the average sparsity $\pi = n^{-1}\sum_m n_m \pi^{(m)}$ becomes mismatched to local conditions.

We illustrate this mismatch using the four-block configuration shown in Figure 1. The block sizes are $n_1 = n - 3n^{9/10}$ and $n_2 = n_3 = n_4 = n^{9/10}$. The local sparsity levels are $\pi^{(1)} = n^{-1/5}$, $\pi^{(2)} = n^{-1/10}$, $\pi^{(3)} = n^{-1/5}$, and $\pi^{(4)} = n^{-1/6}$. The global sparsity parameter is $\pi \sim 2n^{-1/5}$. In an oracle setting that knows the local sparsity levels, it is natural to apply block-specific thresholds of order $\sqrt{2\log((\pi^{(m)})^{-1})}$. The dominant contributions to the asymptotic risk come from blocks 1 and 2, yielding

$$\sup_{\boldsymbol{\mu}\in\Theta_n^b} \rho(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) \sim 2n_1\pi^{(1)} \log\left((\pi^{(1)})^{-1}\right) + 2n_2\pi^{(2)} \log\left((\pi^{(2)})^{-1}\right) \sim \frac{3}{5} n^{4/5} \log n, \tag{3.13}$$

since the terms for blocks 3 and 4 are of lower asymptotic order.

A non-adaptive procedure calibrated to the global sparsity π employs a uniform cutoff $t_{\text{global}} \sim \sqrt{2\log(\pi^{-1})}$. Its worst-case risk over Θ_n^b scales as

$$\sup_{\boldsymbol{\mu}\in\Theta_n^b} \rho(\hat{\boldsymbol{\mu}}_{\text{global}}, \boldsymbol{\mu}) \sim 2n\pi \log(\pi^{-1}) \sim \frac{4}{5} n^{4/5} \log n.$$

The spatially adaptive procedure reduces the asymptotic risk by approximately 25%. This gain arises primarily in the second block, where $\pi^{(2)} = n^{-1/10}$ is substantially larger than the global sparsity level. The global threshold $t_{\text{global}} \approx \sqrt{0.4 \log n}$ is excessively conservative for this dense region, overshrinking true signals. The locally optimal threshold for block 2 is $t^{(2)} \sim \sqrt{2\log(\pi^{(2)})^{-1}} = \sqrt{0.2\log n} \approx 0.707\, t_{\text{global}}$. By applying a lower threshold where the signal density is higher, a spatial adaptive procedure recovers a larger fraction of nonzero coordinates in block 2 while maintaining appropriate stringency in the remaining blocks. This example illustrates how incorporating local sparsity information yields more precise estimation under heterogeneous spatial structures.

The preceding analysis offers insights under ideal conditions: when local sparsity levels can be accurately estimated, the performance of SAFT approaches the risk given in (3.13), thereby illustrating the potential benefits of sparsity-adaptive thresholding. In practice,

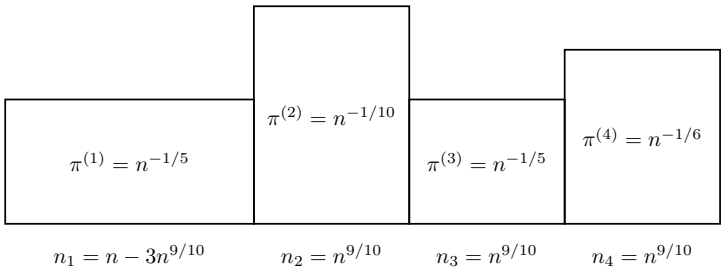


Figure 1: An illustration of the toy example.

precise estimation of local sparsity is rarely attainable. Nevertheless, SAFT remains capable of capturing the underlying structural patterns in the data. By learning these patterns through appropriately chosen weights, the procedure can still employ adaptive thresholds to enhance estimation efficiency. Consequently, even when the resulting weights deviate from their oracle counterparts, the method delivers improved performance, as confirmed by the numerical results in Section 4.

4 Numerical Experiments

This section presents simulation studies comparing the proposed SAFT procedure with established alternatives. Results for the one-dimensional setting are reported in Section 4.1, while two-dimensional simulations are presented in Section 4.2.

The following procedures are evaluated:

- **BH**: the SAFT estimator with uniform weights $w_s = 1$. This specification recovers the global FDR-thresholding procedure of Abramovich et al. (2006).
- **SAFT.SAB.OR**: the SAFT method with weights $w_s = (1 - \pi_s)^{-1}$, computed using the true local sparsity levels.
- **SAFT.SAB.DD**: the SAFT method with weights $\hat{w}_s = (1 - \hat{\pi}_s)^{-1}$, computed using kernel-smoothed sparsity estimates.

- **SAFT.LAW.OR**: the SAFT method with weights $w_s = \pi_s(1-\pi_s)^{-1}$, computed using the true local sparsity levels.
- **SAFT.LAW.DD**: the SAFT method with weights $\hat{w}_s = \hat{\pi}_s(1-\hat{\pi}_s)^{-1}$, computed using kernel-smoothed sparsity estimates.

All methods determine q via the MCV criterion described in Section 3.3. The kernel function and bandwidth follow the specifications in Cai et al. (2022). Estimation accuracy is measured by the average MSE across 100 independent replications. The modified cross-validation procedure uses $\alpha = 1$.

4.1 One-dimensional Block-wise Simulations

We evaluate the SAFT procedure under a piecewise-constant block configuration. Signals are concentrated in four disjoint intervals, with local sparsity levels specified as

$$\pi_s = \begin{cases} \pi, & s \in [1001, 1300] \cup [2001, 2300], \\ 0.7, & s \in [3001, 3300] \cup [4001, 4300], \\ 0.01, & \text{otherwise.} \end{cases} \tag{4.14}$$

Data are generated from the normal mixture model

$$X_s \mid \theta_s \overset{\text{ind}}{\sim} (1-\theta_s)N(0,1) + \theta_s N(\mu, 1), \quad \theta_s \sim \text{Ber}(\pi_s). \tag{4.15}$$

The SAFT weights depend on the local sparsity estimator (3.8). We set the screening parameter τ to 0.5.

We first assess the accuracy of the kernel-smoothed sparsity estimator. Figure 2 displays the true sparsity levels (red) and the estimated values (blue) for a representative replication with $\pi = 0.9$ and $\mu = 3$. The estimator tracks the underlying piecewise-constant structure closely, confirming that kernel smoothing effectively borrows strength from neighboring coordinates while preserving spatial boundaries.

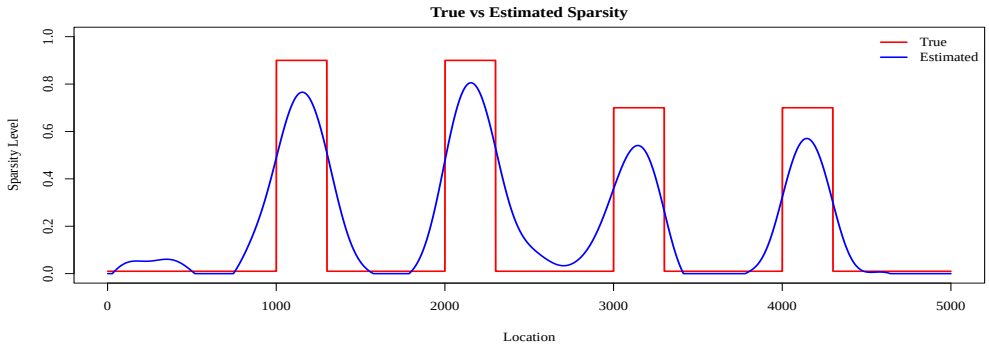


Figure 2: True and estimated local sparsity levels. Red: true π_s; blue: kernel-smoothed estimates $\hat{\pi}_s^{\tau}$. Setting: $\pi = 0.9$, $\mu = 3$.

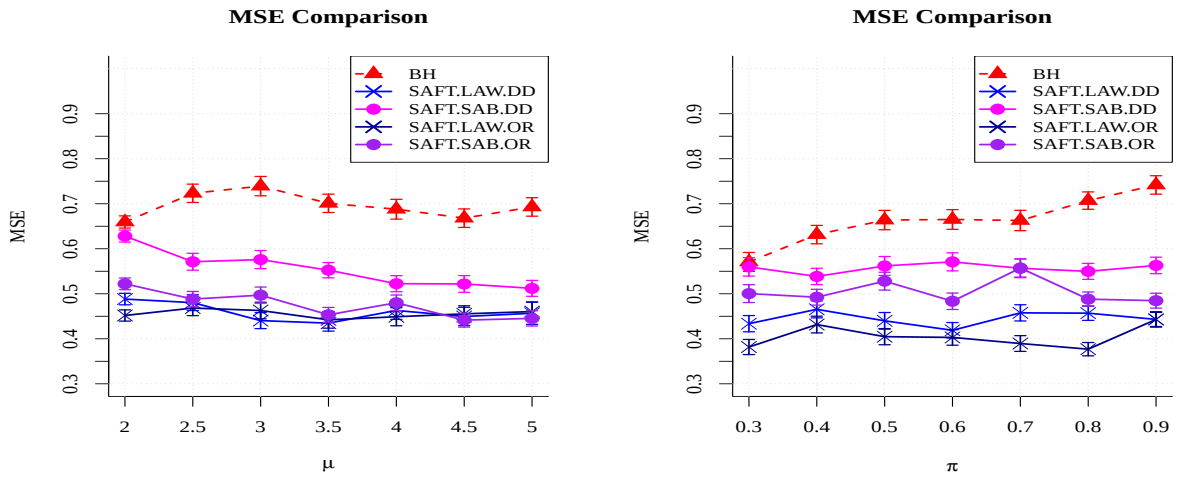


Figure 3: Left: varying signal strength μ with $\pi = 0.9$. Right: varying sparsity level π with $\mu = 3$

We next compare SAFT to competing procedures under two experimental settings. In the first setting, we fix $\pi = 0.9$ and vary the signal strength μ from 2 to 5 to examine the effect of amplitude. In the second setting, we fix $\mu = 3$ and vary the within-block sparsity level π from 0.3 to 0.9 to assess sensitivity to signal density.

The results in Figure 3 show that SAFT achieves lower mean squared error than the global BH procedure across a range of signal strengths and sparsity levels, reflecting the benefit of incorporating local spatial information. The performance gap widens as signals become more concentrated within the specified blocks (right panel), consistent with the intuition that stronger spatial heterogeneity provides more structure for adaptive thresholding. Among

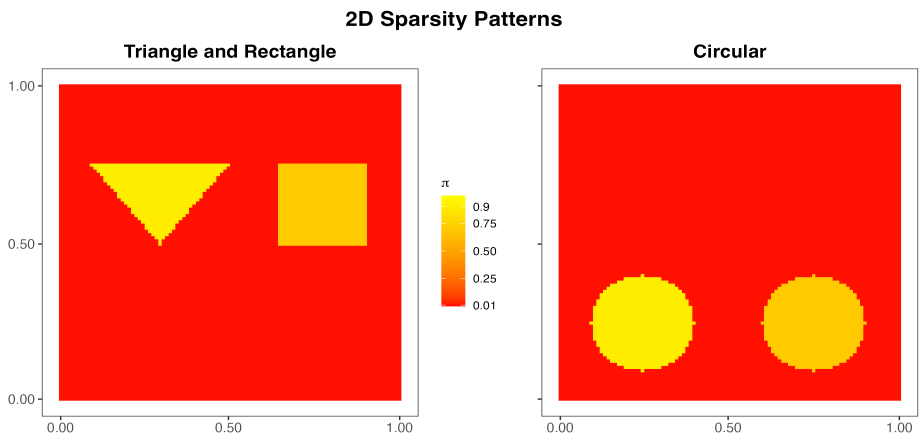


Figure 4: Two-dimensional sparsity patterns. Left: triangle-rectangle configuration; right: double-circular configuration. Colors indicate local sparsity levels π_s.

the weighting schemes, LAWS weights yield better performance than SABHA weights across configurations. The data-driven SAFT procedures approximate the oracle versions with reasonable accuracy, suggesting that the kernel-smoothed sparsity estimates are sufficient for practical use.

4.2 Two-dimensional Simulations

We evaluate the SAFT procedure under two-dimensional spatial configurations. Data are generated on a 100×100 lattice from the normal mixture model (4.15). We consider two spatial patterns for the local sparsity levels π_s: a triangle-rectangle configuration and a double-circular configuration, illustrated in Figure 4. In both patterns, the sparsity structure is defined as

$$\pi_s = \begin{cases} \pi, & s \text{ in left triangular or left circular region,} \\ \pi - 0.2, & s \text{ in right rectangular or right circular region,} \\ 0.01, & \text{otherwise.} \end{cases} \tag{4.16}$$

This construction creates heterogeneous signal density across the domain, with denser signals in the left regions and sparser signals in the right regions.

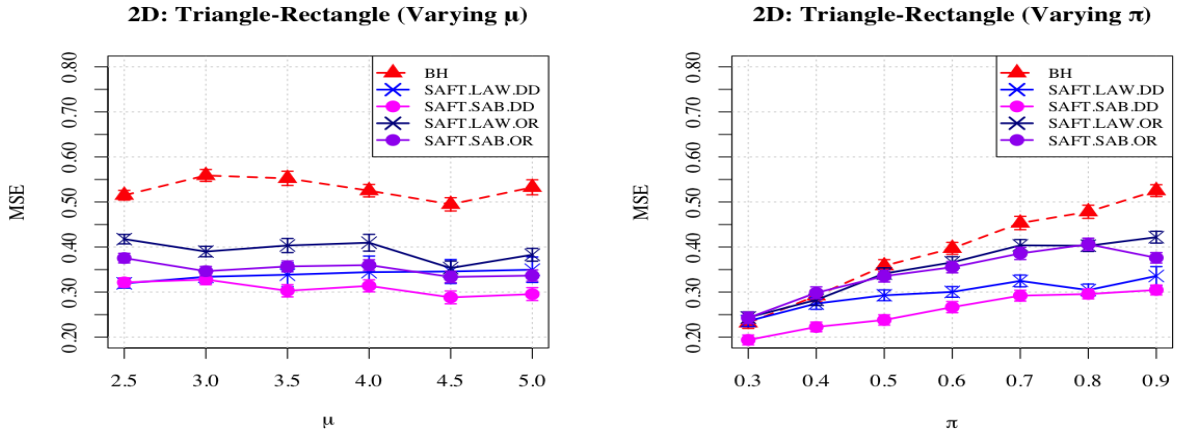


Figure 5: Mean squared error comparisons for the triangle-rectangle configuration. Left: varying signal strength μ with $\pi = 0.9$. Right: varying sparsity level π with $\mu = 3$.

As in the one-dimensional setting, we conduct two experimental studies. In the first, we fix $\pi = 0.9$ and vary the signal strength μ from 2.5 to 5 to examine the effect of amplitude. In the second, we fix $\mu = 3$ and vary the baseline sparsity level π from 0.3 to 0.9 to assess sensitivity to signal density.

Figures 5 and 6 summarize the two-dimensional results. Across both geometric configurations, the spatially adaptive thresholds consistently reduce mean squared error relative to the uniform BH cutoff. This advantage persists at lower signal amplitudes, demonstrating that kernel-based borrowing of information remains effective even when individual coordinates are weakly separated from noise. As the disparity between active and inactive regions grows, the relative improvement of SAFT becomes more pronounced, underscoring how marked spatial inhomogeneity creates additional leverage for location-specific shrinkage. Consistent with the one-dimensional experiments, the LAWS weighting scheme yields lower estimation error than the SABHA alternative across all tested regimes. Moreover, the estimated thresholds closely track the oracle benchmarks, confirming that the kernel-smoothed sparsity estimates maintain their accuracy and utility when extended to lattice-based spatial domains.

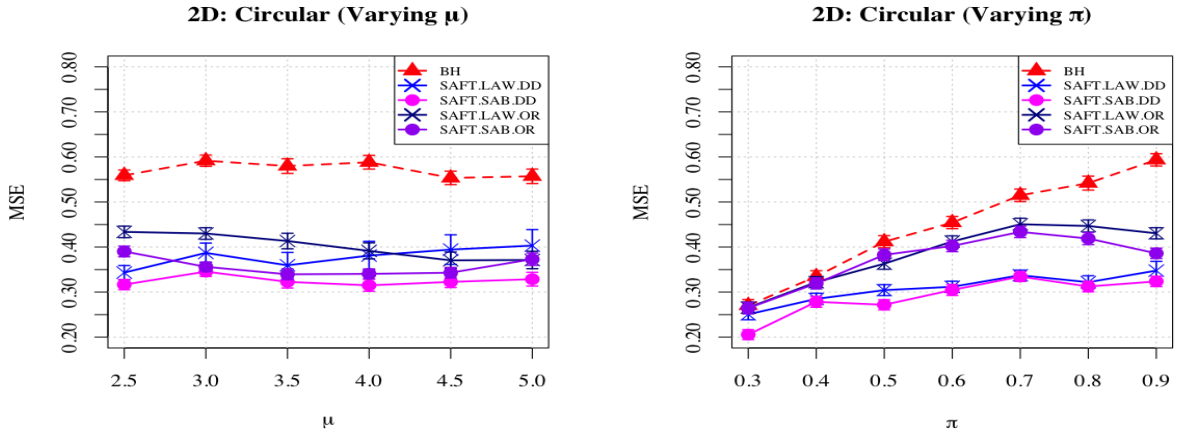
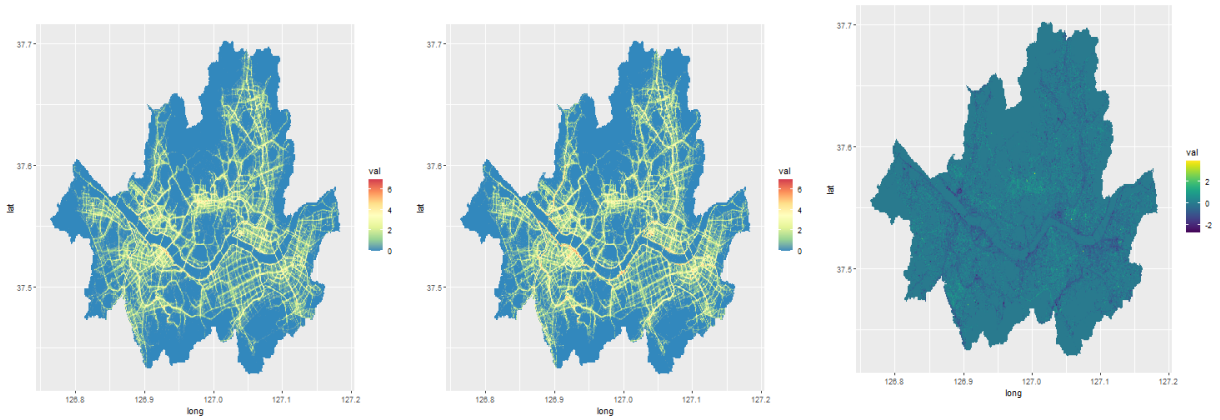


Figure 6: Mean squared error comparisons for the circular configuration. Left: varying signal strength μ with $\pi = 0.9$. Right: varying sparsity level π with $\mu = 3$.

5 Application to Seoul Bike-Sharing Data

Efficient allocation of shared bicycles requires accurate characterization of spatial usage patterns. In many metropolitan areas, trip demand varies systematically between weekdays and weekends due to distinct land-use characteristics. Business districts typically experience higher weekday demand, whereas recreational areas show increased weekend activity. Identifying locations with substantial weekday–weekend asymmetry informs targeted rebalancing strategies and operational planning. This task can be formulated as a high-dimensional estimation problem. The underlying signal is sparse, as most grid points exhibit comparable usage profiles across the week, while only a subset of locations contains meaningful discrepancies. Furthermore, the signal density is spatially heterogeneous, with active regions forming localized clusters rather than scattering uniformly. These structural features make the problem well-suited to spatially adaptive thresholding.

We analyze trip records from the Seoul public bicycle system collected between April 1 and April 28, 2019. Each record includes GPS coordinates, timestamps, and transaction identifiers. To construct a spatial lattice, we overlay an 860×580 grid on the city map with a resolution of 0.0005 degrees in both longitude and latitude. Grid cells corresponding to



(a) Weekdays (b) Weekends (c) Differences between (a) and (b)

Figure 7: Heatmaps of log-scale daily counts of shared bicycle trips in Seoul during April 2019.

impassable terrain, such as rivers, are excluded, leaving $n = 182,099$ valid locations. For each location $s \in \mathcal{S}$, we compute the average daily log-transformed trip counts for weekdays and weekends separately. The target parameter μ_s represents the difference in log-counts between the two periods. Figure 7 displays heatmaps of the weekday averages, weekend averages, and their raw differences. The visual contrast confirms that discrepancies are concentrated in specific districts, while the majority of the spatial domain contains negligible variation.

The estimation procedure is evaluated using a temporal hold-out design. Observations from April 1 to April 14 constitute the training set, and observations from April 15 to April 28 form the test set. Both sets are standardized by their respective empirical standard deviations to stabilize the noise variance across locations. The conditional sparsity levels π_s are estimated via the two-dimensional kernel smoothing method of Cai et al. (2022) with bandwidth $h = 2$ and screening threshold $\tau = 0.032$. The resulting estimates are displayed in Figure 8. We compare the SAFT procedure under LAWS and SABHA weighting schemes to the standard BH-based thresholding estimator. The nominal FDR level q is selected for each procedure via the modified cross-validation criterion with $\alpha = 1$. Estimation accuracy is measured by the mean squared error on the test set, averaged over 200 independent repetitions of the MCV optimization.

Table 1 summarizes the out-of-sample performance. The SAFT estimator with LAWS weights achieves the lowest mean squared error, followed by the SABHA-weighted variant and the global BH procedure. The improvement over the uniform thresholding baseline reflects the benefit of adapting to the spatially varying signal density visualized in Figure 8. By assigning location-specific weights derived from kernel-smoothed sparsity estimates, SAFT applies more appropriate shrinkage in active districts while maintaining stringent noise control in inactive regions. These empirical findings align with the simulation results presented in Section 4.2, confirming that spatial adaptation yields measurable gains in structured high-dimensional estimation.

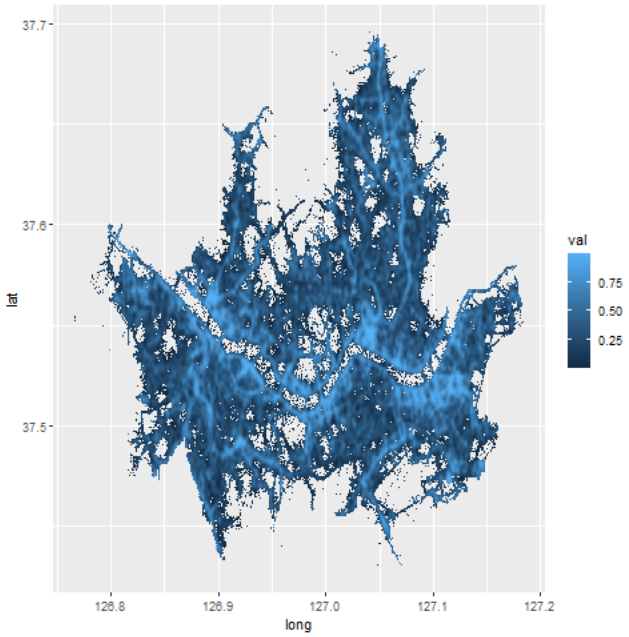


Figure 8: Heatmap of the estimated conditional probabilities π_s.

Method	SAFT.LAW	SAFT.SAB	BH
MSE	<u>0.920</u> (0.017)	0.954 (0.015)	1.049 (0.011)

Table 1: Average mean squared errors over 200 repetitions of the modified cross-validation procedure. Standard deviations are reported in parentheses. The best performing method is underlined.

References

Abramovich, F., Y. Benjamini, D. L. Donoho, I. M. Johnstone, et al. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics 34*(2), 584–653.

Barber, R. F. and A. Ramdas (2017). The p-filter: Multilayer False Discovery Rate Control for Grouped Hypotheses. *Journal of the Royal Statistical Society Series B: Statistical Methodology 79*(4), 1247–1268.

Basu, P., T. T. Cai, K. Das, and W. Sun (2018). Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association 113*(523), 1172–1183.

Benjamini, Y. and R. Heller (2007). False discovery rates for spatial signals. *Journal of the American Statistical Association 102*(480), 1272–1281.

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

Bogdan, M., E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès (2015). Slope—adaptive variable selection via convex optimization. *The Annals of Applied Statistics 9*(3), 1103.

Brown, L. D. and E. Greenshtein (2009). Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics 37*, 1685–1704.

Brown, L. D., E. Greenshtein, and Y. Ritov (2013). The poisson compound decision problem revisited. *Journal of the American Statistical Association 108*(502), 741–749.

Cai, T. T. and W. Sun (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association 104*(488), 1467–1481.

Cai, T. T., W. Sun, and W. Wang (2019). Covariate-Assisted Ranking and Screening for Large-Scale Two-Sample Inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology 81*(2), 187–234.

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

Cai, T. T. and H. H. Zhou (2009). A data-driven block thresholding approach to wavelet estimation. *The Annals of Statistics 37*(2), 569–595.

Dobriban, E., K. Fortney, S. K. Kim, and A. B. Owen (2015). Optimal multiple testing under a Gaussian prior on the effect sizes. *Biometrika 102*(4), 753–766.

Donoho, D. L. and I. M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association 90*(432), 1200–1224.

Donoho, D. L. and J. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika 81*(3), 425–455.

Durand, G. (2019). Adaptive p-value weighting with power optimality. *Electronic Journal of Statistics 13*(2), 3336 – 3385.

Efron, B. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *The Annals of Applied Statistics 2*(1), 197 – 223.

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

Gang, B., L. Fu, G. M. James, and W. Sun (2026). Ranking and selection in large-scale inference of heteroscedastic units. *The Annals of Applied Statistics 20*(1), 789–808.

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

He, L., S. K. Sarkar, and Z. Zhao (2015). Capturing the severity of type ii errors in high-dimensional multiple testing. *Journal of Multivariate Analysis 142*, 106–116.

Ignatiadis, N. and W. Huber (2021). Covariate Powered Cross-Weighted Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology 83*(4), 720–751.

Jiang, W. and C.-H. Zhang (2009). General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics 37*(4), 1647–1684.

Johnstone, I. M. and B. W. Silverman (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics 32*(4), 1594 – 1649.

Lei, L. and W. Fithian (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society Series B: Statistical Methodology 80*(4), 649–679.

Leung, D. and W. Sun (2022). ZAP: Z-Value Adaptive Procedures for False Discovery Rate Control with Side Information. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(5), 1886–1946.

Li, A. and R. F. Barber (2019). Multiple testing with the structure-adaptive benjamini–hochberg algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology 81*(1), 45–74.

Li, S., T. T. Cai, and H. Li (2023). Transfer learning in large-scale gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association 118*(543), 2171–2183.

Ramdas, A. K., R. F. Barber, M. J. Wainwright, and M. I. Jordan (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics 47*(5), 2790 – 2821.

Ren, Z. and E. Candès (2023). Knockoffs with side information. *The Annals of Applied Statistics 17*(2), 1152 – 1174.

Roeder, K. and L. Wasserman (2009). Genome-Wide Significance Levels and Weighted Hypothesis Testing. *Statistical Science 24*(4), 398 – 413.

Roquain, E. and M. A. van de Wiel (2009). Optimal weighting for false discovery rate control. *Electronic Journal of Statistics 3*(none), 678 – 711.

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

Sun, W. and Z. Wei (2011). Multiple testing for pattern identification, with applications to microarray time-course experiments. *Journal of the American Statistical Association 106*(493), 73–88.

Sun, W. and Z. Wei (2015). Hierarchical recognition of sparse patterns in large-scale simultaneous inference. *Biometrika 102*(2), 267–280.

Weinstein, A., Z. Ma, L. D. Brown, and C.-H. Zhang (2018). Group-linear empirical bayes estimates for a heteroscedastic normal mean. *Journal of the American Statistical Association 113*(522), 698–710.

Xiang, D., J. A. Soloff, and W. Fithian (2026). A frequentist local false discovery rate. *Biometrika 113*(1), asaf083.

Xie, X., S. Kou, and L. D. Brown (2012). Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association 107*(500), 1465–1479.

Yekutieli, D. (2008). Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association 103*(481), 309–316.