# False discovery rate smoothing

Wesley Tansey*
Oluwasanmi Koyejo†
Russell A. Poldrack‡
James G. Scott§

November 25, 2014

**Abstract**

We present false discovery rate smoothing, an empirical-Bayes method for exploiting spatial structure in large multiple-testing problems. FDR smoothing automatically finds spatially localized regions of significant test statistics. It then relaxes the threshold of statistical significance within these regions, and tightens it elsewhere, in a manner that controls the overall false-discovery rate at a given level. This results in increased power and cleaner spatial separation of signals from noise. The approach requires solving a non-standard high-dimensional optimization problem, for which an efficient augmented-Lagrangian algorithm is presented. We demonstrate that FDR smoothing exhibits state-of-the-art performance on simulated examples. We also apply the method to a data set from an fMRI experiment on spatial working memory, where it detects patterns that are much more biologically plausible than those detected by existing FDR-controlling methods. All code for FDR smoothing is publicly available in Python and R.[1]

# 1 Introduction

## 1.1 Spatial smoothing in the two-groups model

The traditional problem of multiple testing concerns a group of related null hypotheses $h_1, \ldots, h_n$ that are to be tested simultaneously. In the simplest form of the problem, a summary statistic $z_i$ is observed for each test. The goal is to decide which $z_i$ are signals ($h_i = 1$) and which are null ($h_i = 0$) while ensuring that some maximal error rate is not

*Department of Computer Science, University of Texas at Austin, tansey@cs.utexas.edu
†Department of Psychology, Stanford University, sanmi@stanford.edu
‡Department of Psychology, Stanford University, poldrack@stanford.edu
§Department of Information, Risk, and Operations Management; Department of Statistics and Data Sciences; University of Texas at Austin, james.scott@mccombs.utexas.edu (corresponding author)
[1]https://github.com/tansey/smoothfdr

exceeded. Standard approaches to multiple testing include Bonferroni correction, which controls the overall probability of a single false positive; and the Benjamini-Hochberg procedure, which controls the false discovery rate, or the expected fraction of false positives among all discoveries (Benjamini and Hochberg, 1995). These techniques have been successfully applied across many fields of science, most notably in the analysis of DNA microarrays and other sources of genomic data.

Many of today's large-scale multiple testing problems, however, exhibit clear spatial patterns that are not present in microarray analysis. Examples include: (1) fMRI studies, where significant test statistics tend to cluster in anatomically relevant parts of the brain; (2) studies of allele frequencies in biological populations, where genetic loci correspond to physical locations on the chromosome; (3) studies of variation in DNA methylation fraction across specific genomic regions; (4) neural spike-train data recorded from a multi-electrode array, in which electrodes fall at known locations on a two-dimensional lattice; and (5) environmental sensor networks designed to detect spatially localized anomalies.

This paper presents a new method called *false discovery rate smoothing* that can learn and exploit the underlying spatial structures in these multiple-testing problems. FDR smoothing finds spatially localized regions of significant test statistics by solving a specific optimization problem. It then relaxes the threshold of statistical significance within these regions in a manner that controls the global false-discovery rate at a pre-specified level. This results in increased power and cleaner spatial separation of signals from noise, without the need to pre-smooth the raw $z$ scores.

## 1.2 Connections with existing work

Our approach combines two previously distinct lines of research. First, we incorporate spatial smoothing directly into the "two-groups" model, a popular empirical-Bayes approach for controlling the false-discovery rate that has been advocated by Bradley Efron and many others (e.g. Efron, 2008a; Bogdan et al., 2008; Martin and Tokdar, 2012). This strategy leads to a non-standard high-dimensional optimization problem. As a result, we must also draw on recent advances in convex optimization for solving problems with composite penalty functions (e.g. the fused lasso of Tibshirani et al., 2005). Efron (2008a) provides a recent review on multiple testing under the two-groups model, while Tibshirani and Taylor (2011) describe a wide class of composite regularizers which they call "generalized lasso" problems. We recommend these two papers to readers who wish to get a deeper sense of these two areas of the literature.

Many authors have considered the problem of multiple testing when the test statistics have a complicated dependence structure (e.g. Leek and Storey, 2008; Clarke and Hall, 2009). The focus in these papers is to make conclusions robust in the presence of arbitrarily strong dependence among the test statistics. FDR smoothing explicitly uses known spatial structure to inform the outcome of each test, and therefore differs both conceptually and methodologically from these approaches.

Our approach is similar in spirit to the method of FDR regression recently proposed by Scott et al. (2014). That paper estimates a regression function relating local false discovery rates to a vector of covariates $x_i$ for each test statistic $z_i$. However, FDR smoothing addresses a fundamentally different problem in which we observe a spatial location, rather than a covariate, for each test statistic. In this sense, FDR smoothing differs from FDR regression in the same way that ordinary spatial smoothing differs from ordinary regression.

# 2 FDR smoothing: the basic approach

## 2.1 The two-groups model

The FDR-smoothing algorithm builds upon the two-groups model for multiple testing (Berry, 1988; Efron et al., 2001). Here, one assumes that the test statistics $z_1, \ldots, z_n$ arise from the mixture

$$z \sim c \cdot f_1(z) + (1 - c) \cdot f_0(z)\,, \tag{1}$$

where $c \in (0, 1)$ is an unknown mixing fraction, and where $f_0$ and $f_1$ describe the null ($h_i = 0$) and alternative ($h_i = 1$) distributions of the test statistics. For each $z_i$, one then reports the quantity

$$w_i = \mathrm{P}(h_i = 1 \mid z_i) = \frac{c \cdot f_1(z_i)}{c \cdot f_1(z_i) + (1 - c) \cdot f_0(z_i)}\,. \tag{2}$$

One of the many appealing features of (2) is that it offers a tentative methodological unification for the multiple-testing problem. Bayesians may interpret $w_i$ as the posterior probability that $z_i$ is a signal (e.g. Scott and Berger, 2006; Muller et al., 2006). Frequentists may interpret $1-w_i$ as a local false discovery rate, from which an estimate of the global false-discovery rate can be recovered (Efron et al., 2001). In the simple case, $f_0$ is known while $c$ and $f_1$ must be estimated. But in many applications $f_0$ also has unknown parameters, an important practical complication which Efron (2004) calls the "empirical null." In either scenario, the crucial parameter is $c$, which controls for multiplicity by adapting to the unknown fraction of true signals in the data (Scott and Berger, 2010).

## 2.2 Smoothing across a graph

**Formulating the optimization problem.** FDR smoothing involves a conceptually simple modification of the two-groups model that leads to a non-standard high-dimensional optimization problem. First we construct this problem, and then we show how the solution can be used to produce spatially adaptive false discovery rates. Our method for actually solving the optimization problem is deferred to Sections 4 and 5.

Let each $z_i$ be associated with a vertex $s_i \in \mathcal{S}$ in an undirected graph $\mathcal{G}$ with edge set $\mathcal{E}$. For example, in the fMRI problem shown in the next section, each $s_i$ is a voxel, and $\mathcal{E}$

encodes a three-dimensional adjacency structure in which $(i, j) \in \mathcal{E}$ if and only if sites $i$ and $j$ are adjacent on the 3D grid. Now suppose that the prior probability in (1) changes from site to site:

$$z_i \sim c_i \cdot f_1(z_i) + (1 - c_i) \cdot f_0(z_i) \tag{3}$$

$$c_i = \frac{e^{\beta_i}}{1 + e^{\beta_i}} . \tag{4}$$

Thus $e^{\beta_i}$ is the prior odds that site $s_i$ gives rise to a signal. Let $\beta = (\beta_1, \ldots, \beta_n)^T$ be the vector of log odds, and let $l(\beta)$ be the negative log likelihood assuming that $f_0$ and $f_1$ are fixed:

$$l(\beta) = -\sum_{i=1}^{n} \log \left[ \left( \frac{e^{\beta_i}}{1 + e^{\beta_i}} \right) f_1(z_i) + \left( 1 - \frac{e^{\beta_i}}{1 + e^{\beta_i}} \right) f_0(z_i) \right] .$$

This loss function corresponds to a saturated model in which every site has its own prior log odds. Thus the performance of the method will depend upon the prior or regularizer imposed on $\beta$.

We propose to estimate $\beta$ via a penalized-likelihood approach defined by the following optimization problem:

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \quad l(\beta) + \lambda \sum_{(i,j) \in \mathcal{E}} |\beta_i - \beta_j| . \tag{5}$$

The second term enforces smoothness on the underlying vector of probabilities by penalizing differences in log odds across edges in the graph. The penalty function is the same as the generalized lasso over a graph (Tibshirani and Taylor, 2011). When $\mathcal{G}$ is a grid graph, the penalty function is also the same as the one used in total-variation denoising (Rudin et al., 1992). But equation (5) differs mathematically from these applications, in that the loss function is not the sum of squared errors. It also differs conceptually, in that the object being smoothed ($\beta$) parametrizes a latent image of site-level latent indicators for signal versus noise, rather than the observations themselves.

Following Tibshirani and Taylor (2011), we rewrite (5) in the following way. Let $m = |\mathcal{E}|$ be the size of the edge set, and let $D$ be the oriented adjacency matrix of the graph $\mathcal{G}$, which is the $m \times n$ matrix defined as follows. If $(j, k), j < k$ is the $i$th edge in $\mathcal{E}$, then the $i$th row of D has a $1$ in position $j$, a $-1$ in position $k$, and a $0$ everywhere else. Thus the vector $D\beta$ encodes the set of pairwise first differences between adjacent sites in the log odds of being a signal, and $\sum_{(i,j) \in \mathcal{E}} |\beta_i - \beta_j| = \|D\beta\|_1$. We can therefore express the optimization problem as

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \quad l(\beta) + \lambda \|D\beta\|_1 . \tag{6}$$

This clarifies that the penalty is a composition of two functions applied to $\beta$: a linear transformation composed with the $\ell^1$ norm.

**Using the solution.** The solution to optimization problem (6) yields an estimated graph—or in the 2D case, an image—of log odds. The $\ell^1$ penalty encourages sparsity in the first differences of the log odds across the edges of the graph. As a result, the estimate will partition the nodes of the graph into regions where the log odds are locally constant. The nonzero elements in $D\beta$ correspond to jumps where the log odds change sharply. This builds spatial structure directly into the prior probabilities in Equations (3) and (4). These site-dependent prior probabilities are then used to compute the posterior probability

$$w_i = \mathrm{P}(h_i = 1 \mid z_i) = \frac{c_i \cdot f_1(z_i)}{c_i \cdot f_1(z_i) + (1 - c_i) \cdot f_0(z_i)}\,, \tag{7}$$

from which an estimate of frequentist FDR may be computed using the results in Efron et al. (2001).

Our formulation of the FDR-smoothing problem assumes that both $f_0(z)$ and $f_1(z)$ are known, which is obviously untrue in practice. However, estimating these two densities jointly with the underlying spatial pattern of log odds makes the problem dramatically more difficult. To keep the inference tractable, we employ the following two-stage approach, inspired by the FDR regression technique of Scott et al. (2014). First, we pre-compute estimates for $f_0$ and $f_1$ under the ordinary two-groups model. That is, we assume that the log odds of being a signal are unknown but spatially constant, and therefore described by a single quantity $\beta_s$. To arrive at these estimates, we combine techniques from Efron (2004) and Martin and Tokdar (2012). We then solve the optimization problem in (6) treating $f_0$ and $f_1$ as fixed. The regularization parameter $\lambda$ is chosen separately by a path-based procedure.

Sections through 4 through 7 describe this end-to-end process in detail. However, before turning to these finer points of the FDR smoothing problem, we present two examples of the method in action.

# 3 Examples

## 3.1 A toy one-dimensional problem

We first show the performance of FDR smoothing on a toy problem by simulating $z$ scores along a one-dimensional grid of sites $s_i \in \{1, \ldots, 5000\}$ according to the following model:

$$\begin{aligned} z_i &\sim c_i \cdot N(0, 3^2) + (1 - c_i) \cdot \mathrm{N}(0, 1) \\ c_i &= \begin{cases} 0.5 & \text{if } s_i \in [1501, 2000] \\ 0.02 & \text{otherwise.} \end{cases} \end{aligned}$$

Signals from the over-dispersed $\mathrm{N}(0, 3^2)$ component arise frequently at sites 1501 to 2000 and are rare elsewhere. This is a highly stylized version of a multiple-testing problem that

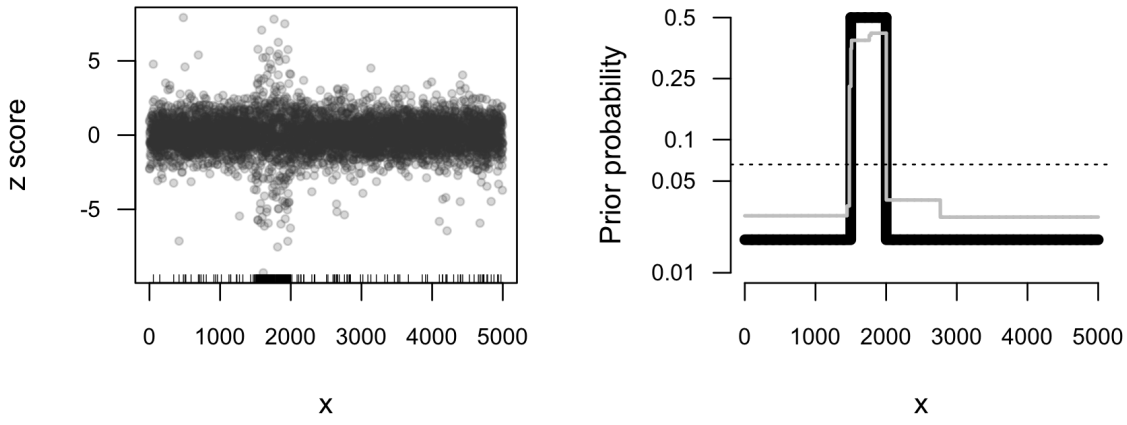


Figure 1: Left panel: raw $z$ scores from the toy example (Section 3.1). The rug at the bottom of the panel show the locations of the true signals, which concentrate heavily at sites 1501 to 2000. Right panel: the true $c_i$ as a function of site is shown as a thick black curve; the FDR-smoothing estimate, as a thinner grey curve. The dashed line shows the grand mean of the $c_i$'s across all sites, which the ordinary two-groups model attempts to recover.

might come up in analyzing allele frequencies or DNA methylation fraction across adjacent sites in the genome (e.g. Jaffe et al., 2012).

The left panel of Figure 1 shows the simulated data. The right panel shows the true versus reconstructed prior probabilities from FDR smoothing. The true prior probability $c_i$ as a function of site is shown as a solid black curve, and the average $c_i$ across the whole data set (the grand mean) is shown as a dashed line. This dashed line is what the ordinary two-groups model aims to recover. Compared with the grand mean, the FDR smoothing estimate (shown in grey) shows a favorable blend of adaptability and stability. For sites 1501 to 2000, the estimate is higher than average, though not as high as the truth—it is shrunk downwards to the mean. For all other sites, the estimate is lower than average, though not as low as the truth—it is shrunk upwards to the mean.

In our simulation studies described in Section 8, FDR smoothing consistently exhibits both of these features: it adapts to automatically to spatial patterns in the data, but it also shrinks toward the two-groups model. The site-level adaptation yields improved power, while the shrinkage yields stability, preventing the model from being too aggressive in isolating spurious groups of signals.

## 3.2 Finding significant regions in fMRI

To show the method in a more realistic scenario, we analyzed data from an fMRI experiment on spatial working memory. The experiment and analysis protocol are described in detail in Appendix A.

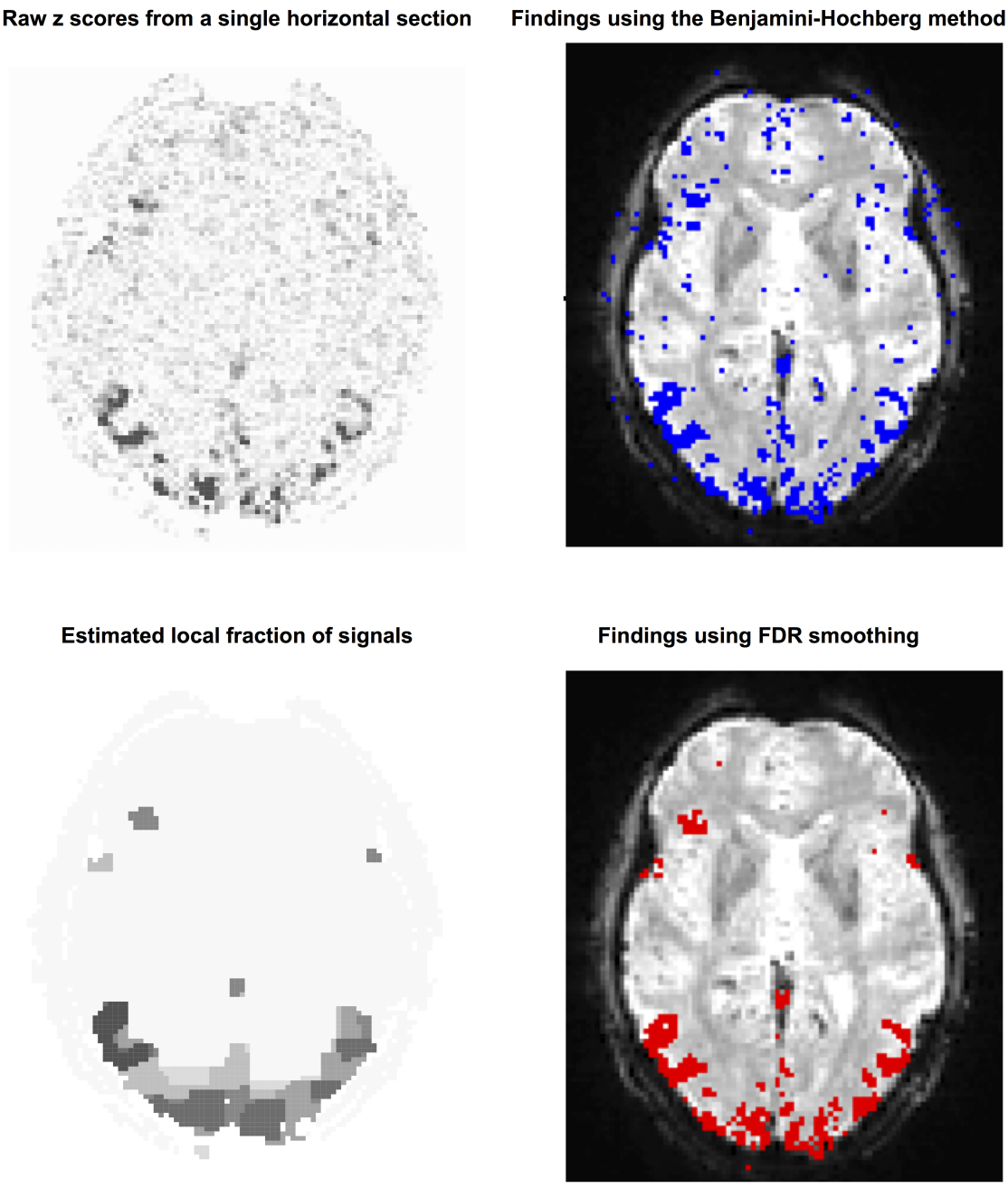


Figure 2: Top left panel: raw $z$ scores from a horizontal section in an fMRI experiment. Darker greys indicate $z$ scores that are larger in absolute value. There is obvious spatial clustering of the large $z$ scores. Top right: the significant discoveries that arise from controlling the global false discovery rate at 5% using the Benjamini-Hochberg procedure. Bottom left: the spatial pattern estimated by FDR smoothing. Darker greys correspond to areas of elevated signal density (i.e. a locally higher fraction of significant z scores). Lighter areas correspond to sparser signal density (a locally smaller fraction of significant z scores). Bottom right panel: discoveries from FDR smoothing at the 5% level.

The upper left panel of Figure 2 shows an image of $z$ scores, often called the statistical parametric map, arising from the experiment and model. Signals correspond to regions of the brain that exhibit systematically different levels of activity across the two experimental conditions (difficult versus easy spatial working memory tasks). The full 3D image has $128 \times 128 \times 75 \approx 1.23$ million voxels. The left panel of Figure 2 shows a single $128 \times 128$ horizontal section in which the large $z$ scores exhibit an obvious pattern of spatial clustering. The shapes of these clusters suggest the underlying brain regions associated with the specific cognitive task under study.

The upper-right panel shows the results of applying the Benjamini-Hochberg procedure to these $z$ scores at a 5% false discovery rate. The procedure clearly finds regions of adjacent points that are all significant. However, the edges of these regions are indistinct, and there are many spatially isolated discoveries that (presumably) are spurious.

Now consider the bottom two panels. The bottom left panel shows our procedure's estimated partition of the raw data shown in the left panel. Darker greys correspond to signal-dense areas containing locally higher fraction of significant $z$ scores; lighter areas correspond to signal-sparse areas containing a locally lower fraction of significant $z$ scores. The bottom right panel then shows the final output: the discovered signals at the 5% FDR level. Compared with the Benjamini-Hochberg procedure, the image reveals regions of significant signals that are more biologically plausible. This reflects the ability of FDR smoothing to be locally adaptive, loosening the threshold for significance in the apparently interesting regions and tightening it in the uninteresting regions.

The partitions shown in the right panel of Figure 1 and the bottom left panel of Figure 2 are estimated by a specialized adaptation of an edge-detection algorithm used to denoise images of natural scenes. However, we emphasize that FDR smoothing is *not* simply denoising the $z$ scores. That is, the estimated partition does not merely pick out areas in which the actual $z$ scores (raw pixel values) are locally constant or locally homoscedastic. Rather, it picks out areas in which the unknown true fraction of signals is locally constant. Unlike the pixel values, these local enrichment fractions are not actually observed by the experimenter. This is a significant complication not present in image denoising and is the fundamental statistical problem addressed by our approach.

# 4 An expectation-maximization algorithm

We now turn to the details of fitting the FDR smoothing model, beginning with the optimization problem in (6). This problem is hard for at least two reasons: the likelihood term $l(\beta)$ is nonconvex, and the penalty term is nonseparable in $\beta$. We are not aware of any algorithm that is guaranteed to find the global minimum efficiently, even for fixed $\lambda$, and the method we describe below finds only a local minimum. Nonetheless, this paper marshals evidence that the local solutions actually found by our algorithm yield good reconstructions of underlying spatial patterns and better power than existing FDR-controlling methods.

We handle the likelihood term with a simple data-augmentation step that leads to an expectation-maximization (EM) algorithm. For now, we assume that $\lambda$ is fixed; we describe our method for choosing this hyperparameter in Section 7. Introduce binary latent variables $h_i$ such that

$$\begin{aligned} z_i &\sim \begin{cases} f_1(z_i) & \text{if } h_i = 1\,, \\ f_0(z_i) & \text{if } h_i = 0 \end{cases} \\ \mathrm{P}(h_i = 1) &= \frac{e^{\beta_i}}{1 + e^{\beta_i}}\,. \end{aligned}$$

Marginalizing out the $h_i$ clearly gives us the original model (3). Treating $h$ as fixed gives the complete-data negative log likelihood:

$$l(\beta, h) = \sum_{i=1}^{n} \left\{ \log\left(1 + e^{\beta_i}\right) - h_i \beta_i \right\}\,. \tag{8}$$

With $h$ fixed, this is a convex function in $\beta$ and is equivalent to the negative log likelihood of a logistic-regression model with identity design matrix.

Therefore, a stationary point of (6) may be found via a conceptually simple EM algorithm. Suppose that the step-$k$ estimate for the underlying image of log odds is $\beta^{(k)}$. In the E step, we compute $q^{(k)}(\beta) = E\{l(\beta, h) \mid \beta^{(k)}\}$. Because the complete-data log likelihood is separable and linear in the $h_i$, we simply plug in the conditional expected value for $h_i$, given the current guess for $\beta_i$, into $l(\beta, h)$. Since $h_i$ is a binary random variable, this is just the conditional probability that $h_i = 1$:

$$w_i^{(k)} = E(h_i \mid \beta^{(k)}, z_i) = \frac{c_i \cdot f_1(z_i)}{c_i \cdot f_1(z_i) + (1 - c_i) \cdot f_0(z_i)}\,, \tag{9}$$

where $c_i$ is the prior probability that site $i$ produces a signal, given by the inverse logit transform of the current estimate $\beta_i^{(k)}$ from equation (4).

In the M step, we maximize the complete-data log likelihood. This requires solving the convex sub-problem

$$\underset{\beta \in \mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^{n} \left\{ \log\left(1 + e^{\beta_i}\right) - w_i \beta_i \right\} + \lambda \|D\beta\|_1\,, \tag{10}$$

where the $w_i$ are the complete-data sufficient statistics.

To solve this sub problem, we expand $l(\beta, w)$ in a second-order Taylor approximation at the current iterate $x$. This turns the $M$ step into a weighted least-squares problem with a generalized-lasso penalty. Thus up to a constant term not depending on $\beta$, the intermediate

problem to be solved is,

$$\underset{\beta\in\mathbb{R}^n}{\text{minimize}} \quad [\nabla\, l(x,w)]^T\,(\beta - x) + \frac{1}{2}(\beta - x)^T H(x,w)(\beta - x) + \lambda\|D\beta\|_1\,, \tag{11}$$

where $\nabla\, l(x,w)^T$ and $H(x)$ are the gradient and Hessian with respect to the first argument of the complete-data negative log likelihood $l(\beta, w)$, evaluated at the current iterate $\beta^{(k)}$ (denoted generically as $x$). These are simple to evaluate:

$$\begin{aligned}
[\nabla\, l(x,w)]_i &= \frac{e^{x_i}}{1+e^{x_i}} - w_i \\
[H(x,w)]_{i,j} &= \begin{cases} \frac{e^{x_i}}{(1+e^{x_i})^2} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}
\end{aligned}$$

The Hessian matrix is diagonal because the log likelihood is separable in $\beta_i$. Ignoring terms that are constant in $\beta$, the solution to (11) can be expressed as the solution of a penalized, weighted least-squares problem:

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} \frac{\eta_i(y_i - \beta_i)^2}{2} + \lambda\|D\beta\|_1 \right\}, \tag{12}$$

with working responses $y_i$ and weights $\eta_i$ given as follows:

$$\begin{aligned}
y_i &= x_i - \frac{c_i - w_i}{\eta_i} \\
\eta_i &= c_i(1 - c_i) \\
c_i &= \frac{e^{x_i}}{1 + e_i^x}\,.
\end{aligned}$$

Recall that $x$ is the point at which the Taylor expansion for the complete-data log likelihood is computed. In our EM algorithm, this is the current estimate $\beta^{(k)}$.

Thus the overall steps of algorithm can be summarized as follows.

**1) E-step:** Use formula (9) to form the complete-data sufficient statistics $w_i$, given the current estimate of $\beta$, to get the complete-data negative log likelihood $l(\beta, w)$ in (8).

**2) Quadratic approximation:** Expand $l(\beta, w)$ in a second-order Taylor series about the current iterate $x \equiv \beta^{(k)}$, thereby forming the "quadratic + penalty" surrogate sub-problem in (12).

**3) Penalized weighted least squares:** Solve the surrogate problem (12) using the augmented-Lagrangian method described in Section 5.

In principle, a full M step requires that steps 2 and 3 be interated until local convergence after each E step. In practice, we take a partial M step by iterating steps 2 and 3 only once.

This speeds the algorithm up: step 3 is by far the most computationally expensive, and we want it to be using sufficient statistics $w_i$ that are as up-to-date as possible. Moreover, as long as the complete-data objective function is improved at each step, the resulting sequence of iterates still converges to a stationary point of (6).

# 5 An augmented-Lagrangian method

The most computationally expensive part of the FDR-smoothing algorithm is the need to repeatedly solve (12). To do so, we formulate a chain of equivalent optimization problems starting from (12). The final problem in this chain can be solved efficiently using an augmented-Lagrangian method known as the alternating direction method of multipliers, or ADMM (for a review, see Boyd et al., 2011). This gives us the solution to the original problem. See Wahlberg et al. (2012) for a similar approach to total-variation denoising.

The first step is to re-express (12) in terms of the following equivalent problem:

$$\begin{aligned} \underset{x,r}{\text{minimize}} \quad & \sum_{i=1}^{n} \frac{\eta_i (y_i - x_i)^2}{2} + \lambda \|r\|_1 \\ \text{subject to} \quad & Dx = r \,. \end{aligned} \tag{13}$$

This is clearly equivalent to (12) due to the constraint on the slack variable $r$. While we could now directly solve (13) via ADMM, doing so would require costly matrix operations at every step of the $x$ update. To avoid this, we encode the slack-variable constraint as part of the objective, rewriting the problem in unconstrained form to yield the new problem

$$\underset{x,r}{\text{minimize}} \quad \sum_{i=1}^{n} \frac{\eta_i (y_i - x_i)^2}{2} + \lambda \|r\|_1 + I_C(x, r) \,, \tag{14}$$

where $I_C(x, r)$ takes the value 0 whenever $(x, r) \in C$, and $\infty$ otherwise. Here $C$ is the convex set

$$C = \{x, r : Dx = r\} \,.$$

This problem enforces the original constraint, because the objective is infinite whenever the constraint is violated. It is therefore also equivalent to the original problem (12).

Now we will introduce two sets of slack variables ($z$ for $x$, $s$ for $r$). This yields yet another new problem:

$$\begin{aligned} \underset{x,z,r,s}{\text{minimize}} \quad & \sum_{i=1}^{n} \frac{\eta_i (y_i - x_i)^2}{2} + \lambda \|r\|_1 + I_C(z, s) \\ \text{subject to} \quad & x = z \in \mathbb{R}^n \\ & r = s \in \mathbb{R}^m \,. \end{aligned} \tag{15}$$

We now have a constrained optimization in four sets of primal variables $x, z, r, s$. Let $u$ be the scaled dual variable corresponding to the constraint $x = z$, and let $t$ be the scaled dual variable corresponding to the constraint $r = s$. We can write the augmented Lagrangian of problem (15) in scaled form as

$$L_a(x, z, r, s, t, u) = \sum_{i=1}^{n} \frac{\eta_i(y_i - x_i)^2}{2} + \lambda\|r\|_1 + I_C(z, s) + \frac{a}{2}\|x - z + u\|_2^2 + \frac{a}{2}\|r - s + t\|_2^2 .$$

See Boyd et al. (2011) for an extensive discussion of the scaled form of the augmented Lagrangian.

ADMM proceeds by iteratively updating the primal and dual variables until a stationary point of the scaled augmented Lagrangian is reached. We describe each update step individually and then discuss the choice of the step-size parameter $a$.

**Updating $x$.** The $x$ update is

$$x^{(k+1)} = \arg\min_x \quad \sum_{i=1}^{n} \frac{\eta_i(y_i - x_i)^2}{2} + \frac{a}{2}\|x - z^{(k)} + u^{(k)}\|_2^2 .$$

This is separable in each component of $x$, and the minimizing value is simply

$$x_i^{(k+1)} = \frac{\eta_i y_i + a(z_i^{(k)} - u_i^{(k)})}{\eta_i + a} .$$

**Updating $r$.** The $r$ update is

$$r^{(k+1)} = \arg\min_r \lambda\|r\|_1 + \frac{a}{2}\|r - s^{(k)} + t^{(k)}\|_2^2 .$$

This is also separable in each component, with minimum given by the soft-thresholding operator:

$$r_j^{(k+1)} = S_{\lambda/a}(s_j^{(k)} - t_j^{(k)}) .$$

Recall that the soft-thresholding operator is $S_a(x) = \text{sign}(x) \cdot |x - a|_+$ with the subscript $+$ indicating the positive part.

**Updating $(z, s)$.** The update in $(z, s)$ must be done jointly. It is the only computationally demanding step of the algorithm. Specifically, we have

$$(z^{(k+1)}, s^{(k+1)}) = \arg\min_{(z,s)} \quad I_C(z, s) + \frac{a}{2}\|x^{(k+1)} - z + u^{(k)}\|_2^2 + \frac{a}{2}\|r^{(k+1)} - s + t^{(k)}\|_2^2 .$$

This is the Euclidean projection of the point $(w, v) = (x^{(k+1)} + u^{(k)}, r^{(k+1)} + t^{(k)})$ to the convex set $C = \{(z, s) : Dz = s\}$. We can equivalently write this sub-problem as

$$\begin{aligned} \underset{z,s}{\text{minimize}} \quad & \|z - w\|_2^2 + \|Dz - v\|_2^2 \\ \text{subject to} \quad & s = Dz\,. \end{aligned}$$

Since $s$ does not appear in the objective, we can just solve for $z$ and then set $s = Dz$. The ordinary first-order optimality condition for $z$ in the above optimization problem is

$$(I + D^T D)z^{(k+1)} = w + D^T v\,, \tag{16}$$

recalling that $w = x^{(k+1)} + u^{(k)}$ and $v = r^{(k+1)} + t^{(k)}$.

Thus to update $z$ (and therefore $s$), we must solve a linear system whose dimension is the size of the underlying vector of log odds. Two important facts are that (1) the matrix of coefficients $A = I + D^T D$ is very sparse, and (2) it never changes over the course of the algorithm. The best algorithm for solving the system will depend on context. For example, if $\mathcal{G}$ is a chain graph corresponding to a one-dimensional smoothing problem, (16) is a tri-diagonal system that can be solved in linear time.

For general graphs, we pre-compute a sparse Cholesky factorization of $A$ under a fill-reducing permutation. This is a significant up-front cost for large systems, but it must be paid only once. We then use this stored factorization to solve (16) for each new right-hand side.

Even greater speed-ups may be possible. Because $D$ is the oriented incidence matrix of a graph, $D^T D$ is the corresponding graph Laplacian matrix. Therefore the matrix of coefficients $A = I + D^T D$ is symmetric diagonally dominant. Amazingly, there exist nearly linear time solvers for such systems (see e.g. Vishnoi, 2013). However, existing practical implementations of these ideas have produced iterative solvers. Much like the conjugate-gradient method, they provide an approximate solution to the linear system after a finite number of steps. We do not know how the use of an iterative solver (as opposed to a direct solver) in the ADMM subroutine would affect the convergence of our algorithm. Therefore we use the sparse-Cholesky method for all analyses in this paper. In future research, we hope to exploit these highly specialized solvers for symmetric diagonally dominant linear systems.

**Updating the dual variables.** The updates for $u$ and $t$ are simply the ordinary scaled-dual variable updates in ADMM:

$$\begin{aligned} u^{(k+1)} &= u^{(k)} + x^{(k+1)} - z^{(k+1)} \\ t^{(k+1)} &= t^{(k)} + r^{(k+1)} - s^{(k+1)}\,. \end{aligned}$$

**Choosing the step-size parameter.** The step-size parameter $a$ greatly affects the speed of the algorithm. Boyd et al. (2011) discusses this issue and provides heuristics for choosing $a$. In our implementation, we use the starting value $a = 2\lambda$ and change $a$ dynamically to track the value of the primal and scaled dual residual. Let

$$e_{\text{pr}} = \begin{pmatrix} x^{(k)} - z^{(k)} \\ r^{(k)} - s^{(k)} \end{pmatrix} \quad \text{and} \quad e_{\text{du}} = \begin{pmatrix} a[z^{(k} - z^{(k-1)}] \\ a[s^{(k} - s^{(k-1)}] \end{pmatrix}$$

be the primal and scaled-dual residual vectors at step $k$, respectively. At each step we calculate the Euclidean norm of these two vectors, which should both be driven to zero over the course of the algorithm. We use the following checks to ensure that this happens quickly. If $\|e_{\text{pr}}\|_2 \geq 5\|e_{\text{du}}\|_2$ then $a$ is too small. Thus for the next step we double the value of $a$ and halve the values of the scaled dual variables $u$ and $t$. Likewise, if $\|e_{\text{du}}\|_2 \geq 5\|e_{\text{pr}}\|_2$ then $a$ is too large. Thus for the next step we halve the value of $a$ and double the values of the scaled dual variables $u$ and $t$. If neither condition is met, we leave $a$ as it is.

# 6 Estimating the null and alternative densities

As described earlier, we estimate $f_0$ and $f_1$ separately, in what amounts to a pre-processing step before solving the FDR-smoothing problem (6). This mirrors the approach of Scott et al. (2014), who also estimate $f_0$ and $f_1$ before fitting an FDR regression model. This can be thought of as a rough approximation to a more standard (but computationally intractable) empirical-Bayes approach in which the underlying spatial pattern is integrated or profiled out. The authors of the FDR-regression paper justify the two-stage approach at length by appealing to the properties of Gaussian convolutions. They also show empirically that (at least in the regression context) the two-stage approach yields answers that are nearly indistinguishable from the answers one gets when fitting the model in a fully Bayesian way. We do not repeat the same arguments here. Instead, we simply describe our approach and refer the interested reader to Scott et al. (2014). As described below, we estimate $f_0$ first. We then use this estimate, together with a simplifying assumption of spatial homogeneity in the prior log odds, to estimate $f_1$. These estimated densities are then used to formulate the FDR-smoothing optimization problem (6).

**Estimating $f_0$.** In some cases, $f_0$ is taken directly from the distributional theory of the test statistic in question (e.g. standard Gaussian) and therefore need not be estimated at all. For such problems where a theoretical null describes the data well, this step can be skipped.

But as Efron (2004) argues, in many multiple-testing problems, the data are poorly described by the theoretical null. In such cases, an empirical null hypothesis with known parametric form but unknown parameters must be estimated in order to produce reasonable results. For the problems considered in this paper, the null hypothesis is assumed to be a Gaussian with unknown mean and variance: $f_0(z) = \text{N}(z \mid \mu_0, \sigma_0^2)$. To estimate $\mu$

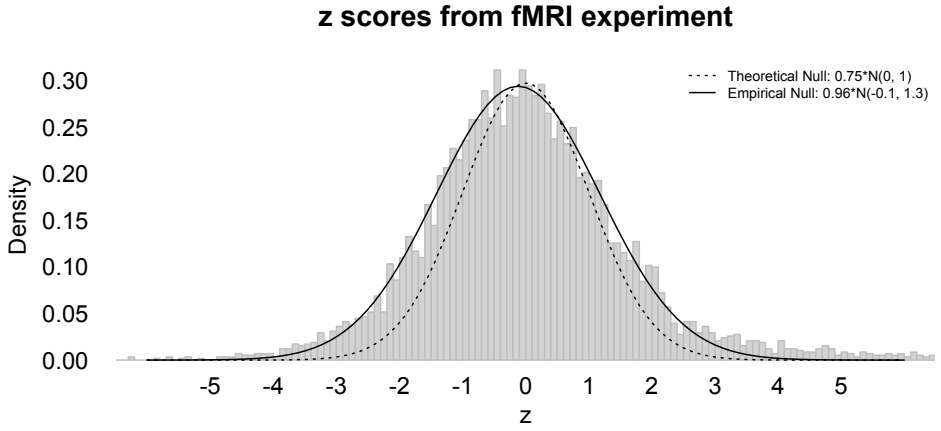


Figure 3: Theoretical versus empirical null for the fMRI data from Section 3.2.

and $\sigma$, we apply the central-matching method of Efron (2004), which uses the shape of the histogram of test statistics near zero (which come mostly or exclusively from the null distribution).

Specifically, let $Z_C$ be the subset comprising some central fraction (we use 1/3) of the $z$ scores. Central matching proceeds in three steps:

1. Construct a smooth estimate $\hat{g}(z)$ of the log density of the test statistics in $Z_C$, and let $z_0$ be the point where $\hat{l}(z)$ obtains its maximum.

2. Form a second-order Taylor approximation of $\hat{g}(z)$ about its maximum:

$$\hat{g}(z) \approx q(z) = \frac{d_2}{2}(z - z_0)^2 + d_1(z - z_0) + d_0 \,.$$

3. The quadratic approxation on the log scale corresponds to a Gaussian on the original scale. Therefore set the mean and standard deviation of the empirical null using the slope and curvature of $q(z)$:

$$\begin{aligned} \mu_0 &= z_0 - \frac{d_1}{2d_2} \\ \sigma_0 &= \sqrt{-\frac{1}{2d_2}} \,. \end{aligned}$$

This approach assumes a Gaussian null, although the approach of using the least interesting test statistics to estimate an empirical null applies more generally in other parametric families. For an alternative approach to estimating an empirical null hypothesis, see Martin and Tokdar (2012).

Figure 3 shows both the theoretical and empirical null for the fMRI data analyzed in Section 3.2. Our analysis uses false discovery rates estimated using the empirical null.

**Estimating $f_1$.** We estimate $f_1(z)$ under the ordinary two-groups model (1), assuming that the null distribution is known (or has been estimated first, as above), and that the prior probability $c$ is unknown but spatially invariant. This problem can be solved by any of several existing methods for one-dimensional Gaussian deconvolution, including finite mixture models or Dirichlet-process models (Do et al., 2005). We use and recommend the predictive-recursion algorithm of Newton (2002) because it is fast, flexible, and enjoys strong guarantees of accuracy (see Tokdar et al., 2009). Predictive recursion generates a nonparametric estimate $\hat{f}_1(z)$ for the marginal density under the alternative after a small number of passes through the data.[2] For further details, see Martin and Tokdar (2012); for pseudo-code, see Scott et al. (2014).

# 7 Choosing the regularization parameter

Once the null and alternative densities have been estimated, the only remaining tuning parameter in FDR smoothing is $\lambda$, the amount of regularization applied to the vector of first differences in log odds across the edges in $\mathcal{G}$. We now describe our method for choosing this number in a data-adaptive way.

Figure 4 illustrates the importance of choosing an appropriate $\lambda$ value. The top left panel depicts an underlying image of prior probabilities. We used this image to generate $z$ scores according to the spatially varying two-groups model in equations (3) and (4) with a specific choice of $f_1$. (For details, see the "small signal" experiment in Section 8.) Choosing $\lambda$ too small, as in the bottom left panel, produces a grainy reconstruction that overfits the data. Choosing $\lambda$ too large, as in the bottom right panel, results in oversmoothing and the loss of interesting spatial structure. Our procedure yields the choice of $\lambda$ shown in the top-right panel. The true regions are recovered with reasonable accuracy, and the graininess of the bottom left panel is avoided.

We avoid having to hand-tune $\lambda$ in an ad-hoc fashion by adopting the following approach, based on the same solution-path idea that is often used to set $\lambda$ in $\ell^1$ problems (e.g. Tibshirani and Taylor, 2011).

1. Calculate the FDR-smoothing solution $\beta(\lambda)$ across a decreasing grid of regularization parameters $\lambda_M > \lambda_{M-1} > \cdots > \lambda_1$, using the solution for $\lambda_s$ as a warm start to find the $\lambda_{s-1}$ solution.

2. For each solution $\beta_s$ corresponding to point $\lambda_s$ on the grid, calculate a relative quality measure $J(\beta_s)$.

[2] In our examples we use 50 passes, although in our experience 10 passes is virtually always sufficient to yield stable estimates.

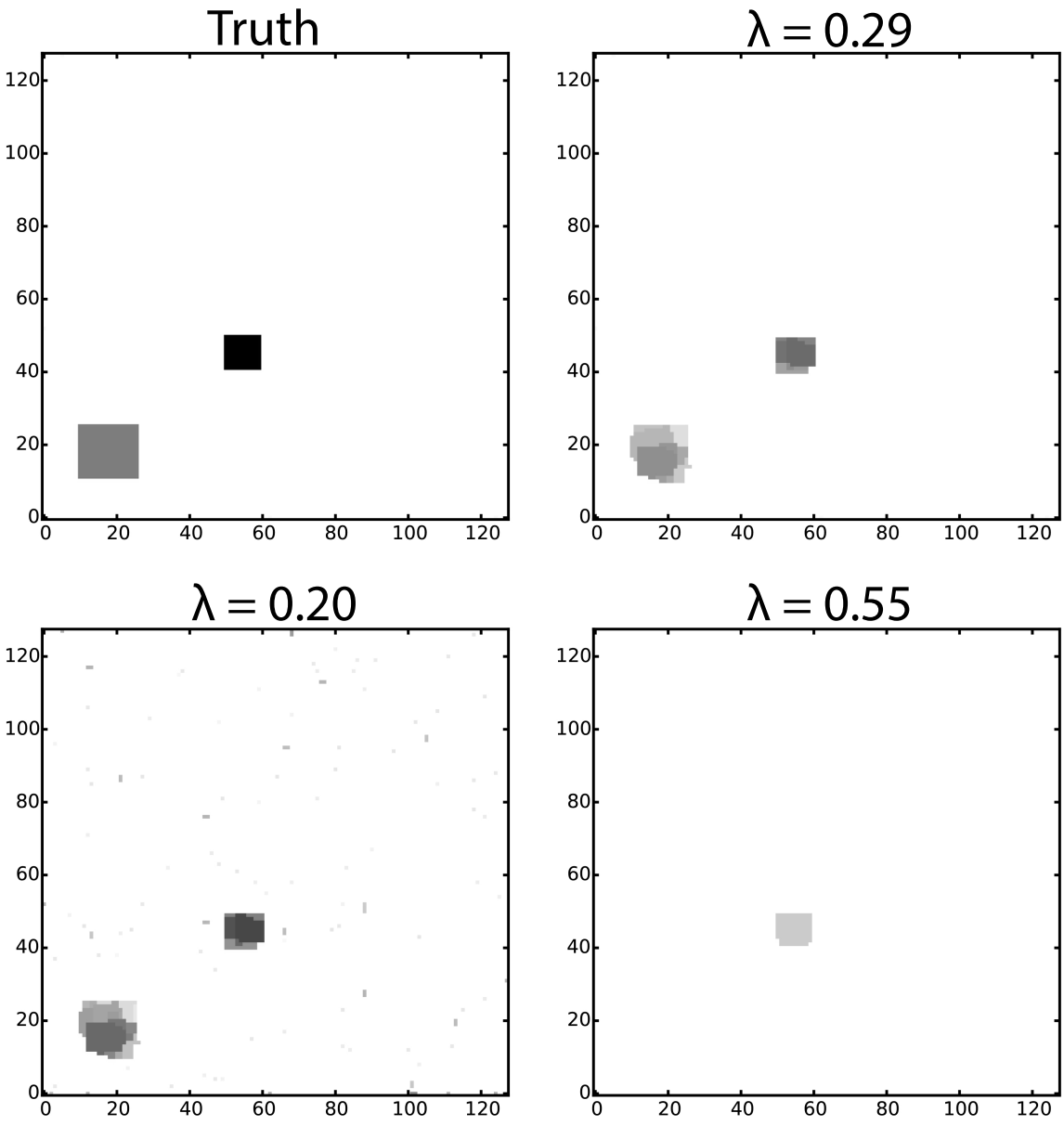


Figure 4: Comparisons of different choices of the $\lambda$ penalty parameter. Choosing $\lambda$ too small (bottom left) will produce a grainy reconstruction that overfits the data. Choosing $\lambda$ too large (bottom right) will oversmooth the data and potentially lose crucial structure. Our path-based method for choosing $\lambda$ results in the choice shown in the top right panel.

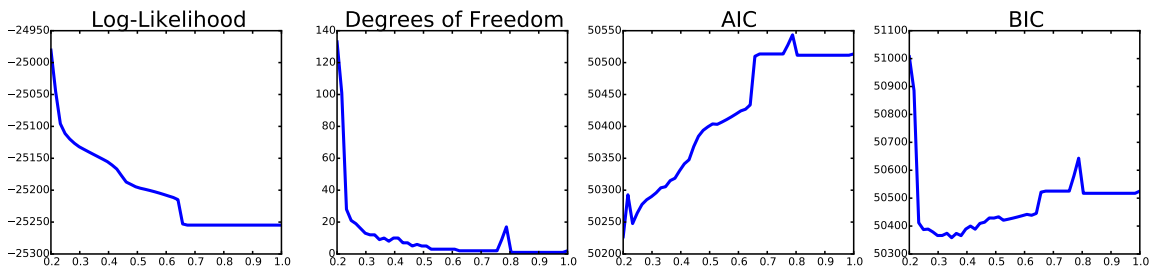


Figure 5: Left two panels: the change in log likelihood and degrees of freedom as the value of lambda changes from 1 to 0. Right two panels: change in AIC and BIC. The small (spurious) bump in the degrees of freedom near $\lambda = 0.8$ is due to the numerical precision of ADMM; see the appendix for details.

3. Choose $\lambda_s$ to be the point in the grid where the quality measure is smallest.

The choice of quality measure should enforce a compromise between the fit and complexity of the reconstructed image. Perhaps the two most common approaches are AIC and BIC. Let $k(\beta)$ be the degrees of freedom of the estimator and $l(\beta)$ the maximized value of the log likelihood. Then up to constants not depending on $\beta$,

$$\mathrm{AIC}(\beta) = -2\log l(\beta) + 2k(\beta) \tag{17}$$

$$\mathrm{BIC}(\beta) = -2\log l(\beta) + \log(n)k(\beta)\,. \tag{18}$$

For simple one-dimensional problems under squared error loss, calculating the degrees of freedom of the generalized lasso equates to counting the number of change points along the $x$ axis. The two-dimensional extension of this result appeals to Stein's lemma, and involves counting the number of distinct contiguous 2-d regions or *plateaus* in $\beta$ (Tibshirani and Taylor, 2012). For example, the true prior in the upper left panel of Figure 4 has three plateaus: the two darker squares, and the white background.

Unfortunately, this remarkable result on the degrees of the freedom of the generalized lasso applies only to problems where $l(\beta)$ is squared error loss. We are aware of no analogous results in more complicated situations involving mixture models such as ours. Therefore, we cannot plug in the true degrees of freedom when calculating AIC and BIC, because it is not known. In the absence of a better alternative, we use the number of plateaus as a surrogate for the degrees of freedom. This is a heuristic solution, but one that seems to yield good performance in practice. The upshot is that if a good estimator for the true degrees of freedom could be found, it is likely that a smarter $\lambda$ could be chosen automatically, and that our overall method could be improved.

Figure 5 shows a typical solution path trace for the log likelihood, surrogate degrees of freedom, AIC, and BIC. In FDR smoothing problems, the number of plateaus is typically much smaller than the number of data points, and the penalty that AIC places on the degrees of freedom is dominated by the log likelihood. As a result, AIC is a disaster in practice,

producing images that are far too grainy. On the other hand, BIC achieves a much better balance across a range of problems, and we recommend it as a criterion for choosing $\lambda$.

An additional practical complication is that it is non-trivial to compute the number of plateaus efficiently for large-scale problems. The naïve approach of counting the number of distinct values in $\hat{\beta}$ can fail badly if the estimate has multiple spatially-separated plateaus with the same estimated prior probability (up to the precision of the ADMM convergence criterion). Pseudo-code for our plateau-counting method is provided in the appendix.

# 8 Simulation experiments

To demonstrate the effectiveness of FDR smoothing, we conducted simulation experiments across eight different scenarios defined by a full cross of two factors:

- two different configurations of the site-specific prior signal probability $c_i$. In one configuration, the true plateaus (or regions where signals are common) are large; in the other configuration, the true plateaus are small. Figure 6 shows the true prior signal probabilities $c_i$ for the large (left) and small (right) scenarios.
- four different choices for $f_1(z)$, the distribution of test statistics under the alternative hypothesis. Each $f_1(z)$ is defined as the Gaussian convolution of some "noiseless" signal distribution $\pi(\theta)$, as described below.

In all eight scenarios, the spatial structure was a $128 \times 128$ two-dimensional grid graph, as in the fMRI example. We simulated 100 data sets in each scenario and set the desired false discovery rate at $10\%$.

For each data set, we simulated $z$ scores as follows. Let $\{c_i\}$ be the true image of prior probabilities, let $\pi(\theta)$ the true (noiseless) distribution of signals, and let $\delta_0$ be a Dirac measure at zero. Then $z_i$ is drawn from the mixture model

$$\begin{aligned} \theta_i &\sim c_i\pi(\theta) + (1 - c_i)\delta_0 \\ z_i &\sim \mathrm{N}(\theta_i, 1)\,. \end{aligned}$$

The null hypothesis is that $\theta_i = 0$, in which case $z_i \sim f_0(z) = \mathrm{N}(0, 1)$. The alternative hypothesis is that $\theta_i \neq 0$, in which case $z_i$ is drawn from the Gaussian convolution $f_1(z) = \int_{\mathbb{R}} \mathrm{N}(\theta_i, 1)\ \pi(\theta_i)\ d\theta_i$.

Figure 7 shows the true $\pi(\theta)$ (dotted grey curve) and the corresponding true convolution $f_1(z)$ (blue curve) in all eight scenarios. The left four panels show the nonparametric estimate of $f_1(z)$ for one data set in each of the four "large-signal" scenarios (dashed orange line). The right four panels have the same true $\pi(\theta)$ and $f_1(z)$ as the left four panels, but show the nonparametric estimate of $f_1(z)$ for one data set in each of the four "small-signal" scenarios. There are 100 data sets and thus 100 estimates for $f_1(z)$ corresponding to each scenario. We show only one estimate out of 100 in each panel, to convey the sense that

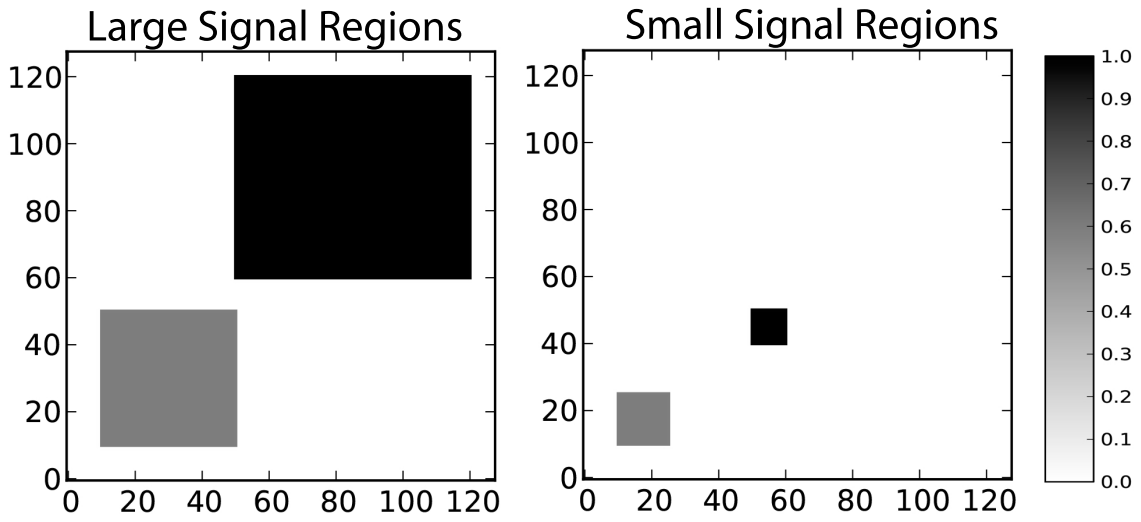


Figure 6: Large (left) and small (right) regions of elevated probability of a $z$ score being drawn from the alternative hypothesis (signal) distribution. The darker square is $p(h = 1) = 0.8$, the lighter square is $p(h = 1) = 0.5$, and the ambient probability is $p(h = 1) = 0.05$.

predictive recursion does a reasonable job at estimating $f_1(z)$ (which is an intrinsically difficult deconvolution problem).

We compare our approach against three other techniques for multiple testing: 1) the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995); 2) Efron's two-groups model (Efron, 2004, 2008b); and 3) FDR regression (Scott et al., 2014). The first two methods are well known, but the third requires some explanation. It is possible treat a spatial-smoothing problem as a regression problem by introducing a suitably rich set of basis functions as spatial "covariates." We have used this trick to shoehorn the FDR-smoothing problem into the FDR-regression framework. As a basis set, we used a tensor product of a standard $b$-spline basis (10 bases in each direction) over the two-dimensional grid. Finally, to establish a baseline we also present the results of an "oracle" model in which both the true $f_1(z)$ and the true underlying $\beta$ vector are assumed known. The represents the theoretical limit of performance of the two-groups model.

Results are presented in Table 1. FDR smoothing has the highest true positive rate (TPR) on seven of the eight examples. Moreover, although FDR regression has a slightly higher TPR on the *Small Alt 3* study, it does so at the cost of slightly overshooting the desired FDR for all of the small-signal examples. FDR smoothing instead errs on the conservative side of the FDR limit.

Compared to the other three techniques, FDR smoothing performs best. In the *Alt 1* and *Alt 4* cases, where the alternative hypothesis density is multimodal and peaked away from the origin, FDR smoothing comes close to the performance of the Bayes oracle model (in which $\beta$ and $f_1(z)$ are fixed at their true values). Overall the method is slightly conservative

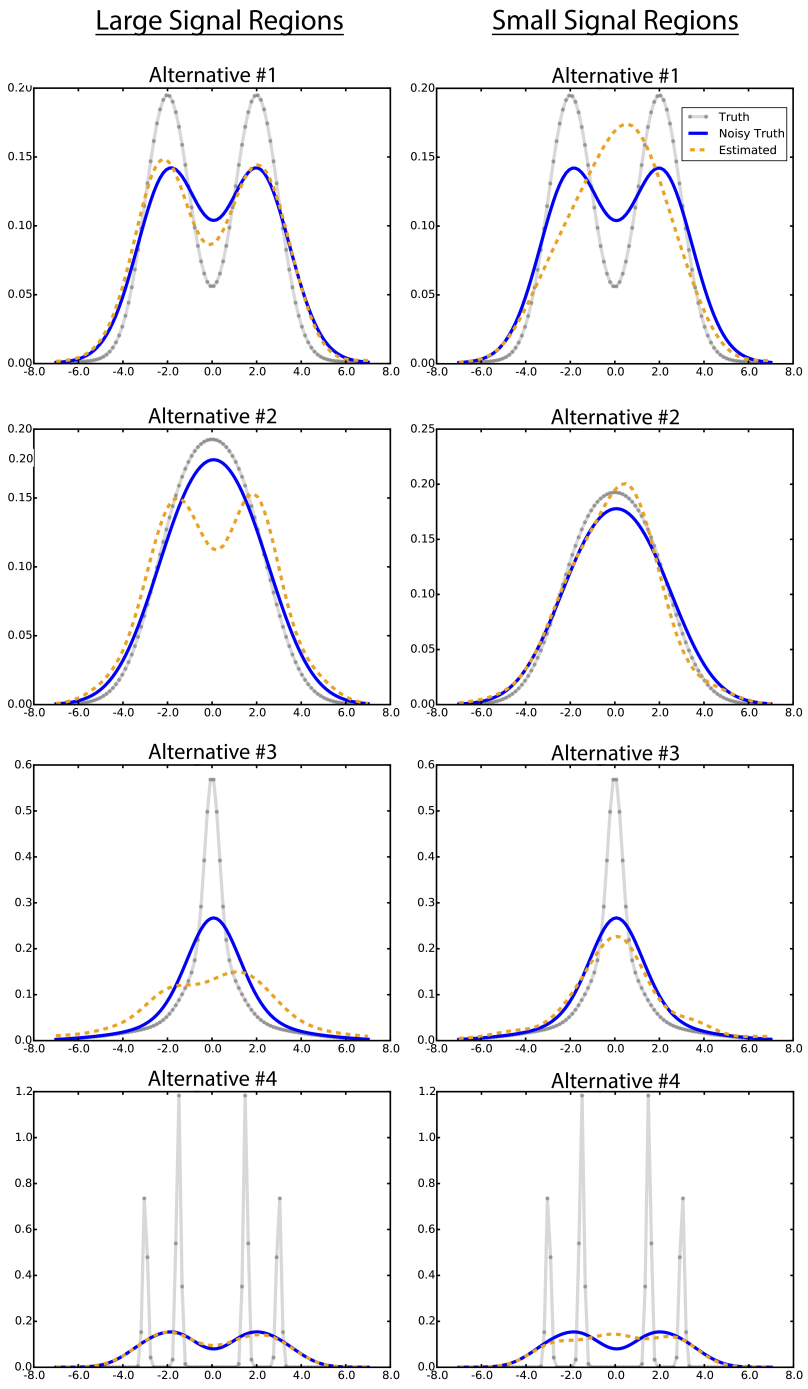


Figure 7: Settings for simulation experiment. The left four panels show the true $\pi(\theta)$ as dotted grey curve and the corresponding convolution $f_1(z)$ as a solid blue curve. The dashed orange curve shows one example (out of 100 simulated data sets) of a nonparametric estimate for $f_1(z)$ via predictive recursion when the true prior signal probabilities are from the large-signal case. The right four panels show the same four pictures, except that the dashed orange curves are estimates of $f_1(z)$ when the true signal probabilities are from the small-signal case.

and therefore loses power versus the oracle. The *Alt 2* and *Alt 3* cases involve much harder deconvolution problems, because these alternative densities have their mode at zero and therefore generate many non-null observations that are nearly indistinguishable from the null density. Even here, FDR smoothing comes closer to the performance of the Bayes oracle than the other methods.

FDR regression using a 100-dimensional $b$-spline basis comes close to the performance of FDR smoothing, but also has many conceptual and computational disadvantages. These are essentially the same disadvantages that one would face in treating *any* spatial smoothing problem in a regression framework. For example, to handle a smoothing problem using FDR regression, one must choose the basis set and the number of basis elements. This is implicitly a choice about the smoothness of the underlying prior image, and is not straightforward in large problems or problems over an arbitrary graph structure. FDR smoothing, on the other hand, has no tunable parameters once our path-based method for choosing $\lambda$ is used. Moreover, FDR regression cannot localize sharp edges in the underlying image of prior probabilities, unless these edges happen to coincide with any edges present in the basis set. FDR smoothing finds these edges automatically without requiring a clever choice of basis, and without having to tolerate undersmoothing in other parts of the image. Finally, at an algorithmic level, the important matrix operations in FDR smoothing involve very sparse matrices and benefit enormously from pre-caching. This is not true in FDR regression, which involves dense matrices and linear systems that change at every iteration.

As the table shows, FDR regression with basis functions does provide sensible answers and good FDR performance. However, the FDR-smoothing approach benefits greatly by exploiting the spatial structure of the problem, resulting in better power and more interpretable summaries at lower computational cost.

# 9 Discussion

Modern scientific analyses often involve large-scale multiple hypothesis testing. In many cases, such as fMRI experiments, these analyses exhibit spatial structure that is ignored by traditional methods for multiplicity adjustment. As we have shown, exploiting this spatial structure via FDR smoothing offers a simple way to increase the overall power of an experiment while maintaining control of the false discovery rate. Our method achieves this performance by automatically identifying spatial regions where the local fraction of signals is enriched versus the background.

While our results show strong statistical and computational performance, there are many areas in which our approach could be improved. We call attention to a few of these areas and suggest them as subjects for future research.

First, the choice of a constant $\ell^1$ penalty on the first differences $D\beta$ leads to some slight overshrinkage in the estimated prior probabilities. This is most evident in Figure 1, where the estimated $c_i$ are shrunk back to the grand mean (or equivalently, toward the

True positive rate (TPR)

| | Large Regions | | | | Small Regions | | | |
|---|---|---|---|---|---|---|---|---|
| | Alt 1 | Alt 2 | Alt 3 | Alt 4 | Alt 1 | Alt 2 | Alt 3 | Alt 4 |
| BH | 0.364 | 0.215 | 0.128 | 0.366 | 0.212 | 0.123 | 0.090 | 0.194 |
| 2G | 0.394 | 0.229 | 0.134 | 0.403 | 0.211 | 0.123 | 0.091 | 0.196 |
| FDR-R | 0.559 | 0.334 | 0.167 | 0.610 | 0.242 | 0.141 | **0.097** | 0.232 |
| FDR-S | **0.592** | **0.352** | **0.168** | **0.645** | **0.264** | **0.144** | 0.093 | **0.257** |
| Oracle | 0.688 | 0.524 | 0.332 | 0.718 | 0.298 | 0.193 | 0.139 | 0.292 |

False discovery rate (FDR)

| | Large Regions | | | | Small Regions | | | |
|---|---|---|---|---|---|---|---|---|
| | Alt 1 | Alt 2 | Alt 3 | Alt 4 | Alt 1 | Alt 2 | Alt 3 | Alt 4 |
| BH | 0.072 | 0.070 | 0.073 | 0.070 | 0.090 | 0.093 | 0.093 | 0.092 |
| 2G | 0.089 | 0.083 | 0.083 | 0.089 | 0.092 | 0.096 | 0.098 | 0.096 |
| FDR-R | 0.075 | 0.058 | 0.050 | 0.086 | 0.102 | 0.106 | 0.109 | 0.105 |
| FDR-S | 0.072 | 0.057 | 0.054 | 0.079 | 0.092 | 0.095 | 0.098 | 0.096 |
| Oracle | 0.101 | 0.100 | 0.100 | 0.101 | 0.097 | 0.101 | 0.101 | 0.098 |

Table 1: Results of the eight simulation studies. Each entry is an average error rate across 100 simulated data sets. FDR smoothing (FDR-S) results in the highest true-positive rate for all but one of the scenarios, consistently beating both the Benjamini–Hochberg procedure (BH) and the two-groups model (2G). FDR regression (FDR-R) comes close, but slightly overshoots the desired FDR limit of 10% in the small-signal examples. (Scott et al., 2014) also report this behavior. In contrast, FDR smoothing remains (on average) under the nominal FDR across all experiments.

estimate of the ordinary two-groups model) versus the true $c_i$. This reflects the well-known "non-diminishing bias" feature of the $\ell^1$ or lasso penalty, and is often mitigated in linear regression by using the adaptive lasso (Zou, 2006) or a concave penalty (Mazumder et al., 2011; Polson and Scott, 2012). Translating these ideas to the FDR smoothing problem presents a formidable algorithmic challenge and is an important area for future work. Nevertheless, even with this noticeable overshrinkage, FDR smoothing achieves state-of-the-art performance in our synthetic experiments. Moreover, it is possible that the overshrinkage is a feature rather than a deficit, in that it prevents the method from being too aggressive in hunting for very small regions of signals.

Second, computational efficiency is a chief concern for FDR smoothing. All of our examples were conducted on a $128 \times 128$ grid, yielding ~16K parameters to estimate. Full fMRI analysis in three dimensions involves over 1M parameters and will likely require additional speedups to our algorithm in order to produce results quickly and without the use of expensive clusters or specialty hardware. The Laplacian linear solvers mentioned in Section 5 and GPU programming are two approaches we plan to investigate.

Third, our method for choosing $\lambda$, the regularization parameter, is effective but *ad hoc*. Our path-based approach would benefit from new theory on the degrees of freedom of the generalized lasso in mixture-model settings, or from an entirely different principle for choosing the tuning parameter in sparse estimation.

Fourth, we have presented FDR smoothing as a general method and provided examples that suggest its wide potential for application. Perhaps the most obvious area in which it could be useful is in the analysis of fMRI data. The literature on fMRI data analysis is large and mature, including the literature on multiplicity correction (e.g. Hayasaka and Nichols, 2003; Poldrack, 2007; Nichols, 2012). We have not attempted to benchmark FDR smoothing against some of these specialized methods, which exploit specific features of fMRI problems that may not hold more generally. This comparison would be out of place in a paper intended for a general statistical audience, but we intend to pursue it in our future work.

Finally, both the lasso and the two-groups model sit (independently of one another) at the center of a large body of theoretical work. We cannot hope to summarize this literature and merely refer to reader to Bickel et al. (2009) for the lasso and Bogdan et al. (2011) for the two-groups model. Combining these two lines of work to produce a theoretical analysis of FDR smoothing represents a major research effort and is beyond the scope of this paper. Nonetheless, given the strong empirical performance of the method, we are hopeful that such an analysis will someday bear fruit.

All code for FDR smoothing is publicly available in Python and R[3].

[3] `https://github.com/tansey/smoothfdr`

# A Details of fMRI data set

The fMRI data set analyzed in Section 3 was acquired and processed as follows. A spatial working memory localizer (Fedorenko et al., 2013) was performed by a single subject. On each trial, a 4x2 spatial grid is presented, and locations in that grid are presented sequentially (1000 ms per location), followed by a forced-choice probe between two grids, one of which contained all of the locations presented in the preceding series. In the easy condition, one location is presented on each presentation, whereas in the hard condition two locations are presented on each presentation. Twelve 32-second experimental blocks were interspersed with 4 16-second fixation blocks (acquisition time = 7:28). The contrast presented in Figure 1 compares the hard versus easy conditions.

fMRI acquisition was peformed using a multi-band EPI (MBEPI) sequence (Moeller et al., 2010) (TR=1.16 ms, TE = 30 ms, flip angle = 63 degrees, voxel size = 2.4 mm X 2.4 mm X 2 mm, distance factor=20%, 64 slices, oriented 30 degrees back from AC/PC, 96x96 matrix, 230 mm FOV, MB factor=4, 10:00 scan length). fMRI data were preprocessed according to a pipeline developed at Washington University, St. Louis (Power et al., 2014), including realignment for motion correction, distortion correction using a field map, and registration to a 3-mm isotropic atlas space. Preprocessed task fMRI data were analyzed at the first level using the FSL Expert Analysis Tool (FEAT, version 5.0.6), using prewhitening and high-pass temporal filtering (100 second cutoff).

# B Finding plateaus in 2D images

Algorithm 1 outlines our approach to finding plateaus, which is needed in our path-based algorithm for choosing $\lambda$. Note that each point in the grid is touched at most $k$ times, where $k$ is the number of neighbors of that point. Thus the algorithm runs in $\mathcal{O}(kn)$, which is effectively linear time since $k \ll n$. The algorithm is mildly sensitive to underlying numerical inaccuracies in the ADMM solution for $\beta$. It is well known that finite-precision ADMM solutions tend to slightly "round off" sharp edges in the underlying image. This produces some slight numerical noise in the degrees of freedom estimate. In our experience, this is rarely a practical concern, and can always be corrected by tightening the convergence criterion for ADMM below the plateau tolerance in Algorithm 1.

# References

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

D. Berry. Multiple comparisons, multiple tests, and data dredging: A Bayesian perspective. In

**Algorithm 1** Our plateau-finding algorithm.

**Input:** grid of values $\beta$, plateau tolerance $\epsilon$
**Output:** list of plateaus and their values $\phi$

$tocheck \leftarrow coordinates(\beta)$
$checked \leftarrow \{\emptyset\}$
$\phi \leftarrow \{\emptyset\}$
**while** $tocheck$ not empty **do**
    $(x_0, y_0) \leftarrow$ pop $tocheck$ until $(x_0, y_0) \notin checked$
    $points \leftarrow \{(x_0, y_0)\}$
    $\beta_{min}, \beta_{max} \leftarrow \beta_{x_0,y_0} - \epsilon, \beta_{x_0,y_0} + \epsilon$
    $unchecked \leftarrow \{(x_0, y_0)\}$
    **while** $unchecked$ not empty **do**
        $(x, y) \leftarrow$ pop $unchecked$
        **for each** neighbor $(v, w)$ of $(x, y)$ **do**
            **if then**$(v, w) \notin checked$ and $\beta_{min} \leq \beta_{v,w} \leq \beta_{max}$
                Add $(v, w)$ to $points$, $unchecked$, and $checked$
            **end if**
        **end for**
    **end while**
    Add $points$ to $\phi$
**end while**
**return** $\phi$


J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, editors, *Bayesian Statistics 3*, pages 79–94. Oxford University Press, 1988.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Annals of Statistics*, 37:1705–32, 2009.

M. Bogdan, J. K. Ghosh, and S. T. Tokdar. A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, volume 1, pages 211–30. Institute of Mathematical Statistics, 2008.

M. Bogdan, A. Chakrabarti, F. Frommlet, and J. K. Ghosh. Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics*, 39(3):1551–79, 2011.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

S. Clarke and P. Hall. Robustness of multiple testing procedures against dependence. *The Annals of Statistics*, 37:332–58, 2009.

K.-A. Do, P. Muller, and F. Tang. A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society, Series C*, 54(3):627–44, 2005.

B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(96–104), 2004.

B. Efron. Microarrays, empirical Bayes and the two-groups model (with discussion). *Statistical Science*, 1(23):1–22, 2008a.

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

B. Efron, R. Tibshirani, J. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of American Statistical Association*, 96:1151–60, 2001.

E. Fedorenko, J. Duncan, and N. Kanwisher. Broad domain generality in focal regions of frontal and parietal cortex. *Proc Natl Acad Sci U S A*, 110(41):16616–21, Oct 2013. doi: 10.1073/pnas.1315235110.

S. Hayasaka and T. Nichols. Validating cluster size inference: random field and permutation methods. *Neuroimage*, 20(4):2343–56, 2003.

A. E. Jaffe, A. P. Feinberg, R. A. Irizarry, and J. T. Leek. Significance analysis and statistical dissection of variably methylated regions. *Biostatistics*, 13(1):166–78, 2012.

J. T. Leek and J. D. Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–23, 2008.

R. Martin and S. Tokdar. A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics*, 13(3):427–39, 2012.

R. Mazumder, J. Friedman, and T. Hastie. Sparsenet: coordinate descent with non-convex penalties. *Journal of the American Statistical Association*, 106(495):1125–38, 2011.

S. Moeller, E. Yacoub, C. A. Olman, E. Auerbach, J. Strupp, N. Harel, and K. Uğurbil. Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri. *Magn Reson Med*, 63(5):1144–53, May 2010. doi: 10.1002/mrm.22361.

P. Muller, G. Parmigiani, and K. Rice. FDR and Bayesian multiple comparisons rules. In *Proceedings of the 8th Valencia World Meeting on Bayesian Statistics*. Oxford University Press, 2006.

M. A. Newton. A nonparametric recursive estimator of the mixing distribution. *Sankhya, Series A*, 64:306–22, 2002.

T. Nichols. Multiple testing corrections, nonparametric methods, and random field theory. *Neuroimage*, 62(2):811–5, 2012.

R. A. Poldrack. Region of interest analysis for fMRI. *Social Cognitive & Affective Neuroscience*, 2 (1):67–70, 2007.

N. G. Polson and J. G. Scott. Local shrinkage rules, Lévy processes, and regularized regression. *Journal of the Royal Statistical Society (Series B)*, 74(2):287–311, 2012.

J. D. Power, A. Mitra, T. O. Laumann, A. Z. Snyder, B. L. Schlaggar, and S. E. Petersen. Methods to detect, characterize, and remove motion artifact in resting state fmri. *Neuroimage*, 84:320–41, Jan 2014. doi: 10.1016/j.neuroimage.2013.08.048.

L. Rudin, S. Osher, and E. Faterni. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(259–68), 1992.

J. G. Scott and J. O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162, 2006.

J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619, 2010.

J. G. Scott, R. C. Kelly, M. A. Smith, P. Zhou, and R. E. Kass. False discovery-rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 2014. to appear.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society (Series B)*, 67:91–108, 2005.

R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *Annals of Statistics*, 39: 1335–71, 2011.

R. J. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40 (2):1198–1232, 2012.

S. Tokdar, R. Martin, and J. Ghosh. Consistency of a recursive estimate of mixing distributions. *The Annals of Statistics*, 37(5A):2502–22, 2009.

N. K. Vishnoi. $Lx = b$ Laplacian solvers and their algorithmic applications. In *Foundations and Trends in Theoretical Computer Science*, volume 8, pages 1–141, 2013.

B. Wahlberg, S. Boyd, M. Annergren, and Y. Wang. An ADMM algorithm for a class of total variation regularized estimation problems. In *Proceedings 16th IFAC Symposium on System Identification*, volume 12, 2012.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.