*The Annals of Statistics*
2006, Vol. 34, No. 2, 584–653
DOI: 10.1214/009053606000000074


# SPECIAL INVITED LECTURE[1]

## ADAPTING TO UNKNOWN SPARSITY BY CONTROLLING THE FALSE DISCOVERY RATE

BY FELIX ABRAMOVICH,[2] YOAV BENJAMINI,[2] DAVID L. DONOHO[2,3,4] AND IAIN M. JOHNSTONE[2,3,5]

*Tel-Aviv University, Tel-Aviv University, Stanford University and Stanford University*

We attempt to recover an $n$-dimensional vector observed in white noise, where $n$ is large and the vector is known to be sparse, but the degree of sparsity is unknown. We consider three different ways of defining sparsity of a vector: using the fraction of nonzero terms; imposing power-law decay bounds on the ordered entries; and controlling the $\ell_p$ norm for $p$ small. We obtain a procedure which is asymptotically minimax for $\ell^r$ loss, simultaneously throughout a range of such sparsity classes.

The optimal procedure is a data-adaptive thresholding scheme, driven by control of the *false discovery rate* (FDR). FDR control is a relatively recent innovation in simultaneous testing, ensuring that at most a certain expected fraction of the rejected null hypotheses will correspond to false rejections.

In our treatment, the FDR control parameter $q_n$ also plays a determining role in asymptotic minimaxity. If $q = \lim q_n \in [0, 1/2]$ and also $q_n > \gamma/\log(n)$, we get sharp asymptotic minimaxity, simultaneously, over a wide range of sparse parameter spaces and loss functions. On the other hand, $q = \lim q_n \in (1/2, 1]$ forces the risk to exceed the minimax risk by a factor growing with $q$.

To our knowledge, this relation between ideas in simultaneous inference and asymptotic decision theory is new.

Our work provides a new perspective on a class of model selection rules which has been introduced recently by several authors. These new rules impose complexity penalization of the form $2 \cdot \log(\text{potential model size}/\text{actual model sizes})$. We exhibit a close connection with FDR-controlling procedures under stringent control of the false discovery rate.

Received March 2000; revised May 2005.

[1]Presented at the 1999 IMS Annual Meeting in Baltimore, Maryland.

[2]Supported by Israel–USA BSF Grant 1999441.

[3]Supported by NSF Grants DMS-95-05151 and DMS-00-77621.

[4]Supported by AFOSR Grant MURI 95-P49620-96-1-0028.

[5]Supported by NIH Grants CA72028 followed by EB001988, a Guggenheim Fellowship and an Adjunct Professorship at the Australian National University.

*AMS 2000 subject classifications.* Primary 62C20; secondary 62G05, 62G32.

*Key words and phrases.* Thresholding, wavelet denoising, minimax estimation, multiple comparisons, model selection, smoothing parameter selection.

**1. Introduction.** The problem of *model selection* has attracted the attention of both applied and theoretical statistics for as long as anyone can remember. In the setting of the standard linear model, we have noisy data on a response variable which we wish to predict linearly using a subset of a large collection of predictor variables. We believe that good parsimonious models can be constructed using only a relatively few variables from the available ones. In the spirit of the modern, computer-driven era, we would like a simple automatic procedure which is data adaptive, can find a good parsimonious model when one exists and is effective for very different types of data and models.

There has been an enormous range of contributions to this problem, so large in fact that it would be impractical to summarize here. Some key contributions, mentioned further below, include the AIC, BIC and RIC model selection proposals [4, 22, 31, 35]. Key insights from this vast literature are:

(a) the tendency of certain rules (notably AIC), when used in an exhaustive model search mode, to include too many irrelevant predictors—Breiman and Freedman [11];

(b) the tendency of rules which do not suffer from this problem (notably RIC) to place evidentiary standards for inclusion in the model that are far stricter than the time-honored "individually significant" single coefficient approaches.

In this paper we consider a very special case of the model selection problem in which a full decision-theoretic analysis of predictive risk can be carried out. In this setting, model parsimony can be concretely defined and utilized, and we exhibit a model selection method enjoying optimality over a wide range of parsimony classes. While the full story is rather technical, at the heart of the method is a simple practical method with an easily understandable benefit: the ability to prevent the inclusion of too many irrelevant predictors—thus improving on AIC—while setting lower standards for inclusion—thus improving on RIC. The optimality result assures us that in a certain sense the method is unimprovable.

Our special case is the problem of estimating a high-dimensional mean vector which is sparse, when the nature and degree of sparsity are unknown and may vary through a range of possibilities. We consider three ways of defining sparsity and will derive asymptotically minimax procedures applicable across all modes of definition.

Our asymptotically minimax procedures will be based on a relatively recent innovation—false discovery rate (FDR) control in multiple hypothesis testing. The FDR control parameter plays a key role in delineating superficially similar cases where one can achieve asymptotic minimaxity and where one cannot.

To our knowledge, this connection between developments in these two important subfields of statistics is new. Historically, the multiple hypothesis testing literature has had little to do with notions like minimax estimation or asymptotic minimaxity in estimation.

The procedures we propose will be very easy to implement and run quickly on computers. This is in sharp contrast to certain optimality results in minimaxity which exhibit optimal procedures that are computationally unrealistic. Finally, because of recent developments in harmonic analysis—wavelets, wavelet packets, and so on—these results are of immediate practical significance in applied settings. Indeed, wavelet analysis of noisy signals can result in exactly the kind of sparse means problem discussed here.

Our goal in this introduction is to make clear to the nondecision-theorist the motivation for these results, the form of a few select results and some of the implications. Later sections will give full details of the proofs and the methodology being studied here.

1.1. *Thresholding.* Consider the standard multivariate normal mean problem:

$$y_i = \mu_i + \sigma_n z_i, \qquad z_i \overset{\text{i.i.d.}}{\sim} N(0,1),\ i = 1, \dots, n. \tag{1.1}$$

Here $\sigma_n$ is known, and the goal is to estimate the unknown vector $\mu$ lying in a fixed set $\Theta_n$. The index $n$ counts the number of variables and is assumed large. The key extra assumption, to be quantified later, is that the vector $\mu$ is *sparse*: only a small number of components are significantly large, and the indices, or locations of these large components are not known in advance. In such situations, thresholding will be appropriate, specifically, hard thresholding at threshold $t\sigma_n$, meaning the estimate $\hat{\mu}$ whose $i$th component is

$$\hat{\mu}_i = \eta_H(y_i, t) = \begin{cases} y_i, & |y_i| \geq t\sigma_n, \\ 0, & \text{else.} \end{cases} \tag{1.2}$$

A compelling motivation for this strategy is provided by wavelet analysis, since the wavelet representation of many smooth and piecewise smooth signals is sparse in precisely our sense [18]. Consider, for example, the empirical wavelet coefficients in Figure 1(c). Model (1.1) is quite plausible if we consider the coefficients to be grouped *level by level*. Within a level, the number of large coefficients is small, though the relative number clearly decreases as one moves from coarse to fine levels of resolution.

1.2. *Sparsity.* In certain subfields of signal and image processing, the wavelet coefficients of a typical object can be modeled as a sparse vector; the interested reader might consult literature going back to Field [21], extending through DeVore, Jawerth and Lucier [12], Ruderman [33], Simoncelli [39] and Huang and Mumford [28]. A representative result was given by Simoncelli, who found that in looking at a database of images, the typical behavior of histograms of wavelet coefficients at a single resolution level of the wavelet pyramid was highly structured, with a sharp peak at the origin and somewhat heavy tails. In short, many coefficients are small in amplitude while a few are very large.

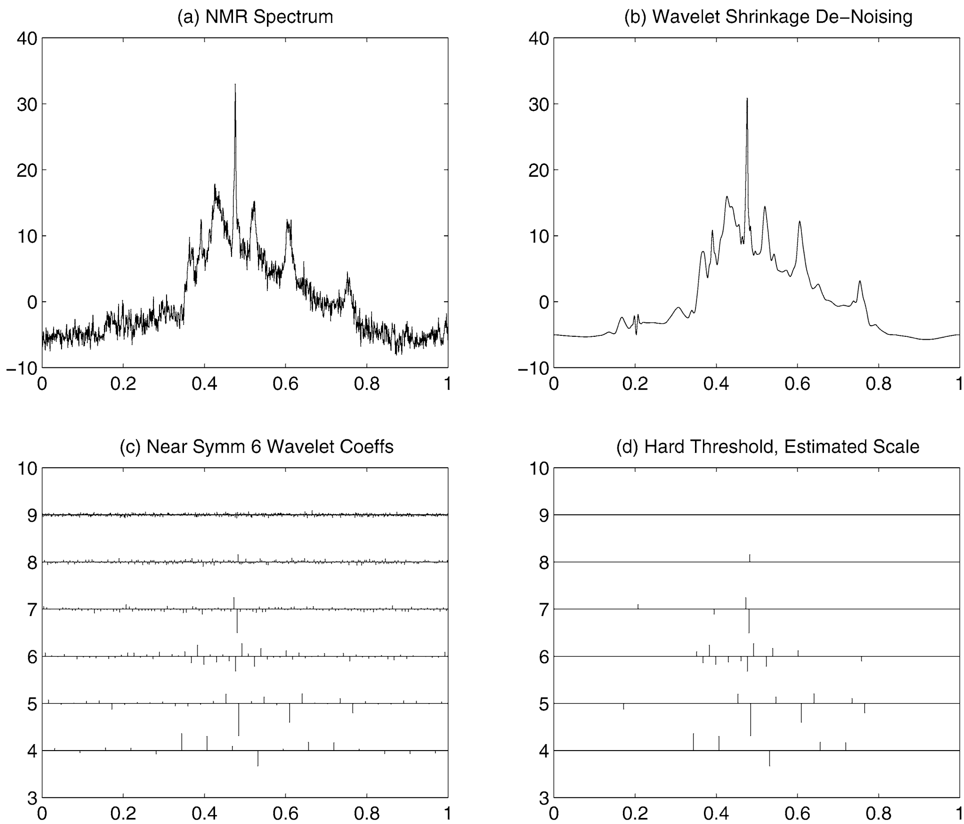


FIG. 1. (a) *Sample NMR spectrum provided by A. Maudsley and C. Raphael*, $n = 1024$, *and discussed in* [18]. (b) *Reconstruction using inverse discrete wavelet transform.* (c) *Empirical wavelet coefficients* $w_{jk}$ *displayed by nominal location and scale* $j$, *computed using a discrete orthogonal wavelet transform and the Daubechies near symmetric filter of order* $N = 6$. (d) *Wavelet coefficients after hard thresholding using the FDR threshold described at* (1.8), *with estimated scale* $\hat{\sigma} = med.abs.dev.(w_{9k})/0.6745$, *a resistant estimate of scale at level* 9*—for details on* $\hat{\sigma}$, *see* [18].

Wavelet analysis of images is not the only place where one meets transforms with sparse cofficients. There are several other signal processing settings—for example, acoustic signal processing—where, when viewed in an appropriate basis, the underlying object has sparse coefficients [5].

In this paper we consider several ways to define sparsity precisely.

The most intuitive notion of sparsity is simply that there is a *relatively small proportion of nonzero coefficients*. Define the $\ell_0$ quasi-norm by $\|x\|_0 = \#\{i : x_i \neq 0\}$. Fixing a proportion $\eta$, the collection of sequences with at most a proportion $\eta$ of nonzero entries is

$$\ell_0[\eta] = \{\mu \in \mathbb{R}^n : \|\mu\|_0 \leq \eta n\}. \tag{1.3}$$

By analogy with night-sky images, we will call *nearly black* a setting where the fraction of nonzero entries $\eta \approx 0$ [17].

Sparsity can also mean that there is a relatively small proportion of relatively large entries. Define the decreasing rearrangement of the amplitudes of the entries so that

$$|\mu|_{(1)} \geq |\mu|_{(2)} \geq \cdots \geq |\mu|_{(n)};$$

we control the entries by a termwise power-law bound on the decreasing rearrangements:

$$|\mu|_{(k)} \leq C \cdot k^{-\beta}, \qquad k = 1, 2, \ldots.$$

For reasons which will not be immediately obvious, we work with $p = 1/\beta$ instead, and call such a constraint a weak-$\ell_p$ constraint. The interesting range is $p$ small, yielding substantial sparsity. One can check whether a vector obeys such a constraint by plotting the decreasing rearrangement on semilog axes, and comparing the plot with a straight line of slope $-1/p$. Certain values of $p < 2$ provide a reasonable model for wavelet coefficients of real-world images [12].

Formally, a *weak-$\ell_p$ ball* of radius $\eta$ is defined by requiring that the ordered magnitudes of components of $\mu$ decay quickly:

$$m_p[\eta] = \{\mu \in \mathbb{R}^n : |\mu|_{(k)} \leq \eta n^{1/p} k^{-1/p} \text{ for all } k = 1, \ldots, n\}. \tag{1.4}$$

Weak $\ell_p$ has a natural "least-sparse" sequence, namely,

$$\bar{\mu}_k = \eta n^{1/p} k^{-1/p}, \qquad k = 1, \ldots, n \tag{1.5}$$

(and its permutations). We also measure sparsity using $\ell_p$ norms with $p$ small:

$$\|\mu\|_p = \left( \sum_{i=1}^{n} |\mu_i|^p \right)^{1/p}. \tag{1.6}$$

That small $p$ emphasizes sparsity may be seen by noting that the two vectors

$$(1, 0, \ldots, 0) \quad \text{and} \quad (n^{-1/p}, \ldots, n^{-1/p})$$

have equivalent $\ell_p$ norms, but when $p$ is small the components of the latter dense vector are all negligible. Strong-$\ell_p$ balls of small average radius $\eta$ are defined so:

$$\ell_p[\eta] = \left\{ \mu \in \mathbb{R}^n : \frac{1}{n} \sum_{i=1}^{n} |\mu_i|^p \leq \eta^p \right\}.$$

If we refer to $\ell_p$ without qualification—weak or strong—we mean strong $\ell_p$.

There are important relationships between these classes. Note that as $p \to 0$, the $\ell_p$ norms approach $\ell_0$: $\|\mu\|_p^p \to \|\mu\|_0$. Weak-$\ell_p$ balls contain the corresponding strong-$\ell_p$ balls, but only just:

$$\ell_p[\eta] \subset m_p[\eta] \not\subset \ell_{p'}[\eta], \qquad p' > p.$$

1.3. *Adapting to unknown sparsity.* Estimation of sparse normal means over $\ell_p$ balls has been carefully studied in [15], with the result that much is known

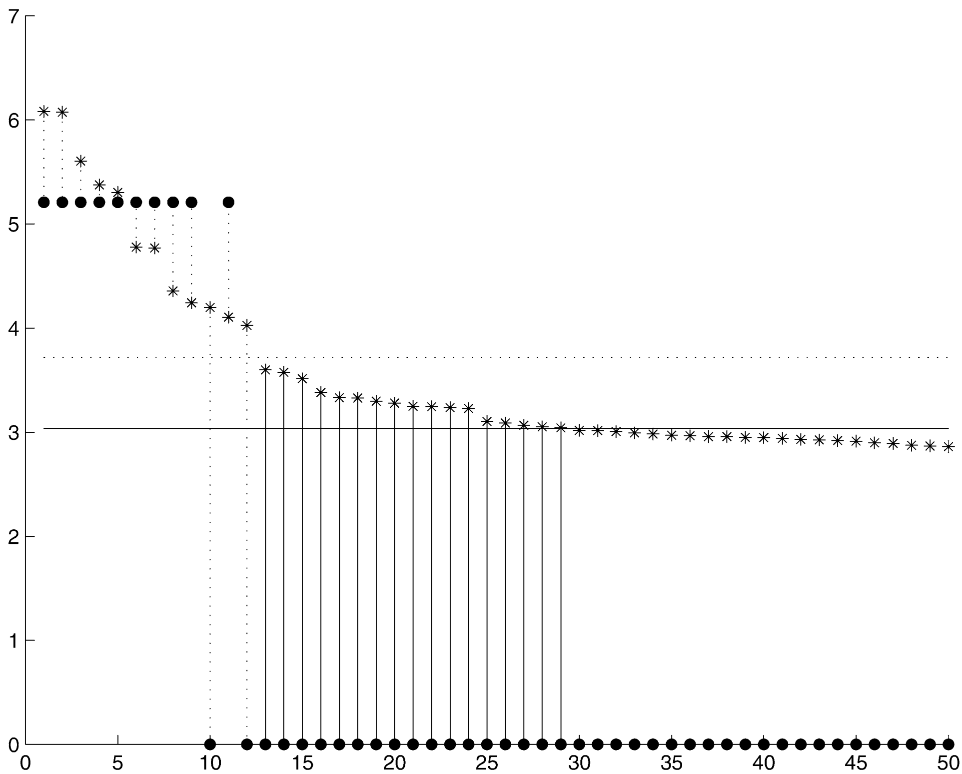


FIG. 2. *Gaussian shift model* (1.1) *with* $n = 10{,}000$ *and* $\sigma_n = 1$. *There are* $n_0 = n^{1/4} = 10$ *nonzero components* $\mu_i = \mu_0 = 5.21$. *Thus* $\beta = 1/4$. *Stars show ordered data* $|y|_{(k)}$ *and solid circles the corresponding true means. Dotted horizontal line is "correct" threshold* $t_{1/4} = \sqrt{2(1-\frac{1}{4})\log n} = 3.72$, *and dotted vertical lines show magnitude of the error committed with* $t_{1/4}$. *Solid horizontal line is a "misspecified" threshold* $t_{1/2} = \sqrt{2(1-\frac{1}{2})\log n} = 3.03$, *which would be the appropriate choice for* $n_0 = n^{1/2} = 100$ *nonzero components. Solid vertical lines show the additional absolute error suffered by using this misspecified threshold. Quantitatively, the absolute error* $\|\hat{\mu} - \mu\|_1$ *using the right threshold is* 14.4 *versus* 70.0 *for the wrong threshold. For* $\ell_2$ *error* $\|\hat{\mu} - \mu\|_2^2$, *the right threshold has error* 38.8 *and the wrong one has error* 221.1.

about asymptotically minimax strategies for estimation. In essence, *if we know* the degree of sparsity of the sequence, then it turns out that thresholding is indeed asymptotically minimax, and there are simple formulas for optimal thresholds.

Figure 2 gives an example. One simple model of varying sparsity levels sets $n_0 = n^{\beta}$ nonzero components out of $n$, $0 < \beta < 1$. Theory reviewed in Section 3 suggests that a threshold of about $t_\beta = \sigma_n\sqrt{2(1-\beta)\log n}$ is appropriate for such a sparsity level. Suppose that $\beta$ is unknown, and examine the consequences of using misspecified thresholds $t_\gamma$, $\gamma \neq \beta$. The solid lines in Figure 2 show the *increased* absolute error incurred using $t_{1/2}$ when $t_{1/4}$ is appropriate—the total absolute error is five times worse. For squared error, the misspecified threshold produces a discrepancy that is larger by nearly a factor of 6.

Typically we could not know in advance the degree of sparsity of the estimand, so we prefer methods adapting automatically to the unknown degree of sparsity.

1.4. *FDR-controlling procedures.* Benjamini and Hochberg [6] proposed a new principle for design of simultaneous testing procedures—control of the false discovery rate (FDR). In a setting where one is testing many hypotheses, the principle imposes control on the expected ratio of the number of erroneously rejected hypotheses to the total number rejected. The exact definition and basic properties of the FDR, as well as examples of procedures holding it below a specified level $q$, are reviewed in Section 2. In the context of *estimation*, a thresholding procedure, which reflects the step-up FDR controlling procedure in [6], was first proposed in [1]. The procedure is quite simple:

Form the order statistics of the magnitudes of the observed estimates,

$$|y|_{(1)} \geq |y|_{(2)} \geq \cdots \geq |y|_{(k)} \geq \cdots \geq |y|_{(n)}, \tag{1.7}$$

and compare them to the series of right tail Gaussian quantiles $t_k = \sigma_n z(q/2 \cdot k/n)$. Let $\hat{k}_F$ be the largest index $k$ for which $|y|_{(k)} \geq t_k$; threshold the estimates at (the data-dependent) threshold $t_{\hat{k}_F} = \hat{t}_F$,

$$\hat{\mu}_{F,k} = \begin{cases} y_k, & |y_k| \geq \hat{t}_F, \\ 0, & \text{else.} \end{cases} \tag{1.8}$$

The FDR threshold is inherently adaptive to the sparsity level: it is higher for sparse signals and lower for dense ones. In the context of model selection, control of the FDR means that when the model is discovered to be complex, so that many variables are needed, we should not be concerned unduly about occasional inclusion of unnecessary variables; this is bound to happen. Instead, it is preferable to control the expected proportion of erroneously included variables. In a limited simulation study in the context of wavelet estimation, Abramovich and Benjamini [2] demonstrated the good adaptivity properties of the FDR thresholding procedure as reflected in relative mean square error performance.

In order to demonstrate the adaptivity of FDR thresholding, Figure 3 illustrates the results of FDR thresholding at two different sparsity levels. In the first, sparser, case a higher threshold is chosen. Furthermore, the fraction of discoveries (coefficients above threshold) that are false discoveries (coming from coordinates with true mean 0) is roughly similar in the two cases. This is consistent with the fundamental result of Benjamini and Hochberg [6] that the FDR procedure described above controls the false discovery rate below level $q$, whatever be the configurations of means $\mu \in \mathbb{R}^n$, $n \geq 1$.

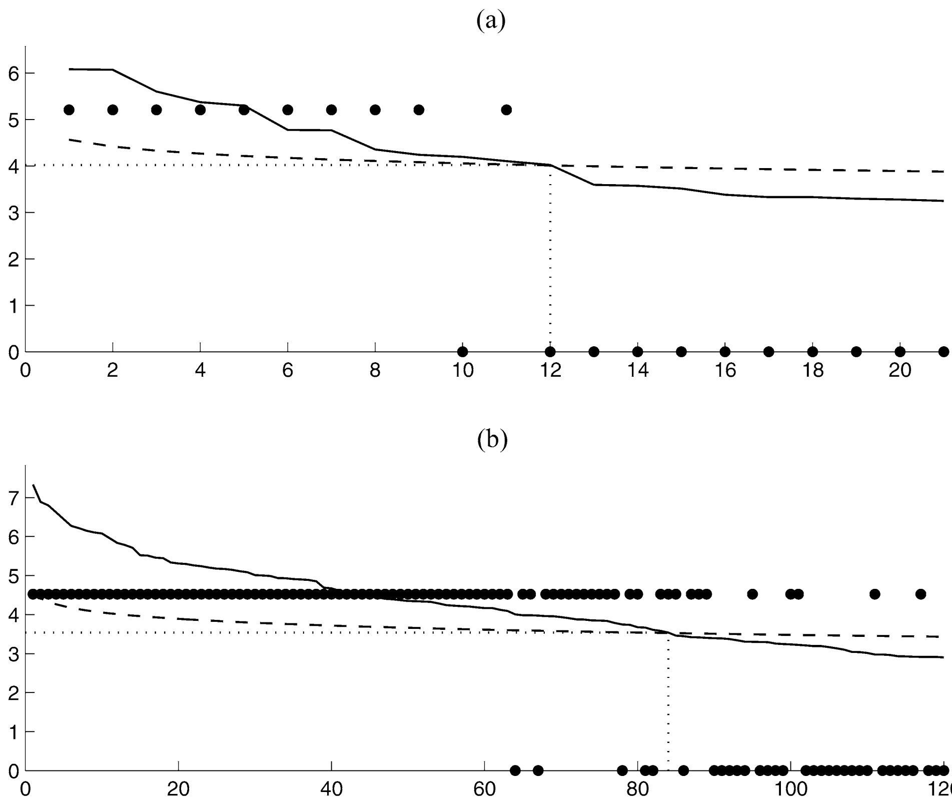


FIG. 3. (a) 10 *out of* 10,000. $\mu_i = \mu_0 \doteq 5.21$ *for* $i = 1, \dots, n_0 = 10$ *and* $\mu_i = 0$ *if* $i = 11, 12, \dots, n = 10{,}000$. *Data* $y_i$ *from model* (1.1), $\sigma_n = 1$. *Solid line*: *ordered data* $|y|_{(k)}$. *Solid circles*: *true unobserved mean value* $\mu_i$ *corresponding to observed* $|y|_{(k)}$. *Dashed line*: *FDR quantile boundary* $t_k = z(q/2 \cdot k/n)$, $q = 0.05$. *Last crossing at* $\hat{k}_F = 12$ *producing threshold* $\hat{t}_F = 4.02$. *Thus* $|y|_{(10)}$ *and* $|y|_{(12)}$ *are false discoveries out of a total of* $\hat{k}_F = 12$ *discoveries. The empirical false discovery rate* $\widehat{\mathrm{FDR}} = 2/12$. (b) 100 *out of* 10,000. $\mu_i = \mu_0 \doteq 4.52$ *for* $i = 1, \dots, n_0 = 100$; *otherwise zero. Same FDR quantile boundary*, $q = 0.05$. *Now there are* $\hat{k}_F = 84$ *discoveries*, *yielding* $\hat{t}_F = 3.54$ *and* $\widehat{\mathrm{FDR}} = 5/84$.

1.5. *Certainty-equivalent heuristics for FDR-based thresholding.* How can FDR multiple-testing ideas be related to the performance of the corresponding estimator? Here we sketch a simple heuristic.

Consider an "in-mean" analysis of FDR thresholding. In the FDR definition, replace the observed data $|y|_{(k)}$ by the mean values $\bar{\mu}_k$, assumed to be already decreasing. Consider a pseudo-FDR index $k_*(\bar{\mu})$, found assuming $\sigma_n = 1$, by solving for the crossing point

$$\bar{\mu}_{k_*} = t_{k_*}.$$

Consider the case where the object of interest obeys the weak-$\ell_p$ sparsity constraint $\mu \in m_p[\eta_n]$. Weak $\ell_p$ has a natural "extreme" sequence, namely (1.5).

Consider the "in-mean" behavior at this extremal sequence; the crossing point relation (1.5) yields

$$\eta_n (n/k_*)^{1/p} = t_{k_*}.$$

Using the relation $t_k \sim \sqrt{2 \log n/k}$, valid for $k = o(n)$, one sees quickly that

$$t_{k_*} \sim \sqrt{2 \log \eta_n^{-p}};$$

the right-hand side of this display is asymptotic to the correct minimax threshold for weak- and strong-$\ell_p$ balls of radius $\eta_n$!

Thus, FDR, in a heuristic certainty-equivalent analysis, is able to determine the threshold appropriate to a given signal sparsity. Further, this calculation makes no reference to the loss function, and so we might hope that the whole range $0 < r \le 2$ of $\ell_r$ error measures is covered.

1.6. *Main results.* Given an $\ell_r$ error measure and $\Theta_n \subset \mathbb{R}^n$, the worst-case risk of an estimator $\hat{\mu}$ over $\Theta_n$ is

$$\bar{\rho}(\hat{\mu}, \Theta_n) = \sup_{\mu \in \Theta_n} E_\mu \|\hat{\mu} - \mu\|_r^r. \tag{1.9}$$

The parameter spaces of interest to us will be those introduced earlier:

(i) $\Theta_n = \ell_0[\eta_n]$ ("nearly black"),
(ii) $\Theta_n = m_p[\eta_n], 0 < p < r$ (weak-$l_p$ balls), and
(iii) $\Theta_n = \ell_p[\eta_n], 0 < p < r$ (strong-$l_p$ balls).

In these cases we will need to have $\eta_n \to 0$ with increasing $n$, reflective of increasing sparsity.

For a given $\Theta_n$, the minimax risk is the best attainable worst-case risk,

$$R_n(\Theta_n) = \inf_{\hat{\mu}} \bar{\rho}(\hat{\mu}, \Theta_n); \tag{1.10}$$

the infimum covers all estimators (measurable functions of the data). Any particular estimator such as FDR must have $\bar{\rho}(\hat{\mu}_F, \Theta_n) \ge R_n(\Theta_n)$, but we might ask how inefficient $\hat{\mu}_F$ is relative to the "benchmark" for $\Theta_n$ provided by $R_n(\Theta_n)$.

THEOREM 1.1. *Let $y \sim N_n(\mu, \sigma_n^2 I)$ and the FDR estimator $\hat{\mu}_F$ be defined by* (1.8). *In applying the FDR estimator, the FDR control parameter* ($q_n$, *say*) *may depend on* $n$, *but suppose this has a limit* $q \in [0, 1)$. *In addition, suppose* $q_n \ge \gamma/\log(n)$ *for some* $\gamma > 0$ *and all* $n \ge 1$.

*Use the $\ell_r$ risk measure* (1.9) *where* $0 \le p < r \le 2$. *Let $\Theta_n$ be one of the parameter spaces detailed above with* $\eta_n^p \in [n^{-1} \log^5 n, n^{-\delta}]$, $\delta > 0$. *Then as* $n \to \infty$,

$$\sup_{\mu \in \Theta_n} \rho(\hat{\mu}_F, \mu) = R_n(\Theta_n) \left\{ 1 + u_{rp} \frac{(2q-1)_+}{1-q} + o(1) \right\},$$

*where $u_{rp} = 1$ and $u_{rp} = 1 - (p/r)$ for strong- and weak-$l_p$ balls, respectively.*

Hence, if the FDR control parameter $q \le 1/2$, $\bar{\rho}(\hat{\mu}_F, \Theta_n) \sim R_n(\Theta_n)$ in the sense that the ratio approaches 1 as $n \to \infty$. Otherwise, $\bar{\rho}(\hat{\mu}_F, \Theta_n) \sim c(q) R_n(\Theta_n)$ for an explicit $c(q) > 1$ growing with $q$.

In short, Theorem 1.1 establishes the asymptotic minimaxity of the FDR estimator in the setting of (1.1)—provided we control false discoveries so that there are more true discoveries than false ones. Moreover, this minimaxity is *adaptive* across various losses and sparse parameter spaces.

This exhibits a tighter connection between false discovery rate ideas and adaptive minimaxity than one might have expected. The key parameter in the FDR theory—the rate itself—seems to be diagnostic for performance.

1.7. *Interpretations.* Two remarks help place the above result in context.

1.7.1. *Comparison with other estimators.* The result may be compared to traditional results in the estimation of the multivariate normal mean. Summarizing results given in [15]:

(i) Linear estimators attain the wrong *rates* of convergence when $0 < p < r$ over these parameter spaces;

(ii) The James–Stein estimator, which is essentially a linear estimator with data-determined shrinkage factor, has the same defect as linear estimators;

(iii) Thresholding at a fixed level, say $\sigma_n\sqrt{2\log n}$, *does* attain the right rates, but with the wrong constants for $0 < p < r$;

(iv) Stein's unbiased risk estimator (SURE) directly optimizes the $\ell_2$ error, and is adaptive for $r = 2$ and $1 < p \le 2$ [16]. However, there appears to be a major technical (empirical process) barrier to extending this result to $p \le 1$, and indeed, instability has been observed in such cases in simulation experiments [16]. Further, there is no reason to expect that optimizing an $\ell_2$ criterion should also give optimality for $\ell_r$ error measures, $p < r < 2$.

In short, traditional estimators are not able to achieve the desired level of adaptation to unknown sparsity. On the other hand, recent work by Johnstone and Silverman [30], triggered by the present paper, exhibits an empirical Bayes estimator—EBayesThresh—which seems, in simulations, competitive with FDR thresholding, although the theoretical results for sparse cases are currently weaker.

1.7.2. *Validity of simultaneous minimaxity.* Minimax estimators are often criticized as being complicated, counterintuitive and distracted by irrelevant worst cases. An often-cited example is $\hat{p} = [x + \sqrt{n}/2]/[n + \sqrt{n}]$ for estimating a success probability $p \in [0, 1]$ from $X \sim \mathrm{Bin}(n, p)$. Although this estimator is minimax for estimating $p$ under squared-error loss, "everybody" agrees that the common-sense estimator $\bar{x} = x/n$ is "obviously better"—better at most $p$ and marginally worse only at $p$ near $1/2$.

Perhaps surprisingly, simultaneous (asymptotic) minimaxity seems to avoid such objections. Instead, to paraphrase an old dictum, it shifts the focus from an "exact solution to the wrong problem" to "an approximate solution to the right problem." To explain this, note that to develop a standard minimax solution, one starts with parameter space $\Theta$ and error measure $\|\cdot\|$ and finds a minimax estimator $\hat{\mu}_{\Theta,\|\cdot\|}$ attaining the minimum in (1.10). This estimator may indeed be unsatisfactory in practice, for example, because it may depend on aspects of $\Theta$ that will not be known, or may be incorrectly specified.

In contrast, we begin here with an a priori reasonable estimator $\hat{\mu}_F$ whose definition does not depend on the imposed $\|\cdot\|$ and the presumably unknown $\Theta_n$. Adaptive minimaxity for $q \le 1/2$—as established for $\hat{\mu}_F$ in Theorem 1.1—shows that, for a large class of relevant parameter spaces $\Theta_n$ and error measures $\|\cdot\|$, $\bar{\rho}(\hat{\mu}_n, \Theta_n) \sim R_n(\Theta_n)$. In other words, the prespecified estimator $\hat{\mu}_n$ is flexible enough to be approximately an optimal solution in many situations of very different type (varying sparsity degree $p$, sparsity control $\eta_n$ and error measure $r$ in the FDR example).

Using large $n$ asymptotics to exhibit approximately minimax solutions for finite $n$ also renders the theory more flexible. For example, in the binomial setting cited earlier, the standard estimator $\bar{x} = x/n$, while not exactly minimax for finite $n$, *is* asymptotically minimax. More: if we consider in the binomial setting the parameter spaces $\Theta_{[a,b]} = \{p : a \le p \le b\}$, then $\bar{x}$ is simultaneously asymptotically minimax for a very wide range of parameter spaces—each $\Theta_{[a,b]}$ for $0 < a < b < 1$—whereas $\hat{p}$ is asymptotically minimax only for special cases $a < 1/2 < b$. In short, whereas minimaxity violates common sense in the binomial case, simultaneous asymptotic minimaxity agrees with it perfectly.

1.8. *Penalized estimators.* At the center of our paper is the study, not of $\hat{\mu}_F$, but of a family of *complexity-penalized* estimators. These yield approximations to FDR-controlling procedures, but seem far more amenable to direct mathematical analysis. Our study also allows us to exhibit connections of FDR control to several other recently proposed model selection methods.

A penalized estimator is a minimizer of $\tilde{\mu} \mapsto K(\tilde{\mu}, y)$, where

$$K(\mu, y) = \|y - \mu\|_2^2 + \mathrm{Pen}(\mu). \tag{1.11}$$

If the *penalty term* $\mathrm{Pen}(\mu)$ takes an $\ell_p$ form, $\mathrm{Pen}(\mu) = \lambda\|\mu\|_p^p$, familiar estimators result: $p = 2$ gives linear shrinkage $\hat{\mu}_i = (1+\lambda)^{-1} y_i$, while $p = 1$ yields soft thresholding $\hat{\mu}_i = (\mathrm{sgn} y_i)(|y_i| - \lambda/2)_+$; for $p = 0$, $\mathrm{Pen}(\mu) = \lambda\|\mu\|_0$ gives hard thresholding $\hat{\mu}_i = y_i I\{|y_i| \ge \lambda\}$.

*Penalized FDR* results from modifying the penalty to

$$\mathrm{Pen}(\mu) = \sum_{l=1}^{\|\mu\|_0} t_l^2.$$

Denote the resulting minimizer of (1.11) by $\hat{\mu}_2$. For small $\|\mu\|_0$, $\mathrm{Pen}(\mu) \sim t_{\|\mu\|_0}^2 \cdot \|\mu\|_0$. It therefore has the flavor of an $\ell_0$ penalty, but with the regularization parameter $\lambda$ replaced by the squared Gaussian quantile appropriate to the complexity $\|\mu\|_0$ of $\mu$. Further, $\hat{\mu}_2$ is indeed a variable hard threshold rule. If $\hat{k}_2$ is a minimizer of

$$S_k = \sum_{l=k+1}^{n} y_{(l)}^2 + \sum_{l=1}^{k} t_l^2,$$

then

$$\hat{\mu}_{2,i} = y_i I\{|y_i| \geq t_{\hat{k}_2}\}.$$

The connection with original FDR arises as follows: $\hat{k}_2$ is the location of the *global* minimum of $S_k$, while the FDR index $\hat{k}_F$ is the rightmost *local* minimum. Similarly, we define $\hat{k}_G$ as the leftmost local minimum of $S_k$: evidently $\hat{k}_G \leq \hat{k}_2 \leq \hat{k}_F$. For future reference, we will call $\hat{k}_G$ the *step-down FDR* index. In practice, these indices are often identical. For theoretical purposes, we show (Proposition 5.1 and Theorem 9.3) that $\hat{k}_F - \hat{k}_G$ is uniformly small enough on our sparse parameter spaces $\Theta_n$ that asymptotic minimaxity conclusions for $\hat{\mu}_2$ can be carried over to $\hat{\mu}_F$.

To extend this story from $\ell_2$ to $\ell_r$ losses, we make a straightforward translation:

$$\hat{\mu}_r = \arg\min_{\mu} \|y - \mu\|_r^r + \sum_{l=1}^{\|\mu\|_0} t_l^r. \tag{1.12}$$

Again it follows that $\hat{k}_r \in [\hat{k}_G, \hat{k}_F]$. Our strategy is, first, to prove $\ell_r$-loss optimality results using $\hat{\mu}_r$, and later, to draw parallel conclusions for the original FDR rule $\hat{\mu}_F$.

Why is the penalized form helpful? In tandem with the definition of $\hat{\mu}_r$ as the minimizer of an empirical complexity $\tilde{\mu} \mapsto K(\tilde{\mu}, y)$, we can define the minimizer $\mu_0$ of the *theoretical* complexity $\tilde{\mu} \mapsto K(\tilde{\mu}, \mu)$ obtained by replacing $y$ by its expected value $\mu$. By the very definition of $\hat{\mu}_r$, we have $K(\hat{\mu}_r, y) \leq K(\mu_0, y)$, and by simple manipulations one arrives (in the $\ell_2$ case here) at the basic bound, valid *for all* $\mu \in \mathbb{R}^n$:

$$E\|\hat{\mu}_2 - \mu\|^2 \leq K(\mu_0, \mu) + 2E\langle \hat{\mu}_2 - \mu, z\rangle - E_\mu \operatorname{Pen}(\hat{\mu}_2). \tag{1.13}$$

Analysis of the individual terms on the right-hand side is very revealing. Consider the theoretical complexity term $K(\mu_0, \mu)$. For $\Theta_n$ of type (i)–(iii) in the previous section, it turns out that the worst-case theoretical complexity is asymptotic to the minimax risk! Thus

$$\sup_{\mu \in \Theta_n} K(\mu_0, \mu) \sim R_n(\Theta_n), \qquad n \to \infty. \tag{1.14}$$

The argument for this relation is rather easy, and will be given in Section 9.2. The remaining term $2E\langle \hat{\mu}_2 - \mu, z\rangle - E_\mu \operatorname{Pen}(\hat{\mu}_2)$ in (1.13) has the flavor of an error term of lower order. Detailed analysis is actually rather hard work, however. Section 9.3 overviews a lengthy argument, carried out in the immediately following sections, showing that this error term is indeed negligible over $\Theta_n$ if $q \le 1/2$, and of the order of $R_n(\Theta_n)$ otherwise.

Plausibility for simultaneous asymptotic minimaxity of FDR is thus laid out for us very directly within the penalized FDR point of view. A full justification requires study of the theoretical complexity and the error term, respectively. This fact permeates the architecture of the arguments to follow.

1.9. *Penalization by* $2k\log(n/k)$. Penalization connects our work with a vast literature on model selection. Dating back to Akaike [4], it has been popular to consider model selection rules of the form

$$\hat{k} = \arg\min_k \mathrm{RSS}(k) + 2\sigma^2 k\lambda,$$

where $\lambda$ is the penalization parameter and $\mathrm{RSS}(k)$ stands for "the best residual sum of squares $\|y - m\|_2^2$ for a model $m$ with $k$ parameters." The AIC model selection rule takes $\lambda = 1$. Schwarz's BIC model selection rule takes $\lambda = \log(n)/2$, where $n$ is the sample size. Foster and George's RIC model selection rule takes $\lambda = \log(p)$, where $p$ is the number of variables available for potential inclusion in the model.

Several independent groups of researchers have recently proposed model selection rules with variable penalty factors. For convenience, we can refer to these as $2\log(n/k)$ factors, yielding rules of the form

$$\hat{k} = \arg\min_k \mathrm{RSS}(k) + 2\sigma_n^2 k\log(n/k). \tag{1.15}$$

(a) Foster and Stine [23] arrived at a penalty $\sigma^2 \sum_{j=1}^k 2\log(n/j)$ from information-theoretic considerations. Along sequences of $k$ and $n$ with $n \to \infty$ and $k/n \to 0$, $2k\log(n/k) \sim \sum_{j=1}^k 2\log(n/j)$.

(b) For prediction problems, Tibshirani and Knight [42] proposed model selection using a covariance inflation criterion which adjusts the training error by the average covariance of predictions and responses on permuted versions of the dataset. In the case of orthogonal regression, their proposal takes the form of a complexity-penalized residual sum of squares, with the complexity penalty approximately of the above form, but larger by a factor of 2: $2\sigma_n^2 \sum_{j=1}^k 2\log(n/j)$. There are intriguing parallels between the covariance expression for the optimism [19] in [42], formula (6), and the complexity bound (1.13).

(c) George and Foster [25] adopted an empirical Bayes approach, drawing the components $\mu_i$ independently from a mixture prior $(1 - w)\delta_0 + wN(0, C)$ and then estimating the hyperparameters $(w, C)$ from the data $y$. They argued that the resulting estimator penalizes the addition of a $k$th variable by a quantity close to $2\log(\frac{n+1}{k} - 1)$.

(d) Birgé and Massart [10] studied complexity-penalized model selection for a class of penalty functions, including penalties of the form $2\sigma_n^2 k\log(n/k)$. They developed nonasymptotic risk bounds for such procedures over $\ell_p$ balls.

Evidently, there is substantial interest in the use of variable-complexity penalties. There is also an *extensive similarity* of $2k\log(n/k)$ penalties to FDR penalization. Penalized FDR $\hat{\mu}_2$ from (1.12) can be written in penalized form with a variable-penalty factor $\lambda_{k,n}$,

$$\hat{k}_2 = \arg\min_k \mathrm{RSS}(k) + 2\sigma_n^2 k\lambda_{k,n},$$

where

$$\lambda_{k,n} = \frac{1}{2k}\sum_{l=1}^{k} z^2\left(\frac{lq}{2n}\right) \sim z^2\left(\frac{kq}{2n}\right)\Big/2$$
$$\sim \log(n/k) - \frac{1}{2}\log\log(n/k) + c(q,k,n)$$

for large $n$, $k = o(n)$ and bounded remainder $c$ [cf. (12.7) below]. FDR penalization is thus slightly weaker than $2k\log(n/k)$ penalization. We could also say that $2k\log(n/k)$ penalties have a formal algebraic similarity to FDR penalties, but require a variable $q = q(k,n)$ that is both small and decreasing with $n$. This perspective on $2k\log(n/k)$ penalties suggests the following conjecture:

CONJECTURE 1.2. *In the setting of this paper, where "model selection" means adaptive selection of nonzero means, and the underlying estimand $\mu$ belongs to one of the parameter spaces as detailed in Theorem* 1.1, *the procedure* (1.15) *is asymptotically minimax, simultaneously over the full range of parameter spaces and losses covered in that theorem.*

In short, although the $2k\log(n/k)$ rules were not proposed from a formal decision-theoretic perspective, they might well exhibit simultaneous asymptotic minimaxity. We suspect that the methods developed in this paper may be extended to yield a proof of this conjecture.

1.10. *Take-away messages.* The theoretical results in this paper suggest the following two messages:

TAM 1. FDR-based thresholding gives an optimal way of adapting to unknown sparsity: choose $q \le 1/2$. In words, aiming for fewer false discoveries than true ones yields sharp asymptotic minimaxity.

TAM 2. Recently proposed $2k\log(n/k)$ penalization schemes, when used in a sparse setting, may be viewed as similar to FDR-based thresholding.

1.11. *Simulations.* We tested FDR thresholding and related procedures in simulation experiments. The outcomes support TAM's 1 and 2.

Table 1 displays results from simulations at the so-called least-favorable case $\mu_k = \min\{n^{-1/2}k^{-1/p}, \sqrt{(2-p)\log n}\}$ for the weak-$\ell_p$ parameter ball [cf. remark following (9.13)]. Here $p = 1.5$, $r = 2$, $n = 1024$ and $n = 65{,}536$, $\sigma = 1$.

TABLE 1
*Ratios of* MSE(FDR)/ MSE($t^*(p,n)$), $p = 3/2$

| | $q$ | Step-up FDR | Penalized FDR | Step-down FDR |
|---|---|---|---|---|
| $n = 1024$ | 0.01 | 1.3440 | 1.3440 | 1.3440 |
| | 0.05 | 1.3283 | 1.3293 | 1.3334 |
| | 0.25 | 1.2473 | 1.2482 | 1.2512 |
| | 0.40 | 1.2171 | 1.2171 | 1.2173 |
| | 0.50 | 1.2339 | 1.2335 | 1.2321 |
| | 0.75 | 1.4159 | 1.4132 | 1.4100 |
| | 0.99 | 1.9810 | 1.9744 | 1.9687 |
| $n = 65{,}536$ | 0.01 | 1.3370 | 1.3372 | 1.3374 |
| | 0.05 | 1.3178 | 1.3180 | 1.3183 |
| | 0.25 | 1.2276 | 1.2277 | 1.2277 |
| | 0.40 | 1.1889 | 1.1889 | 1.1890 |
| | 0.50 | 1.1937 | 1.1936 | 1.1936 |
| | 0.75 | 1.5122 | 1.5118 | 1.5114 |
| | 0.99 | 4.0211 | 4.0189 | 4.0174 |

The table records the ratio of squared-error risk of FDR to squared-error risk of the asymptotically optimal threshold $t^* = t^*(p,n) = \sqrt{(2-p)\log n}$ for that parameter ball (cf. Section 3.3 below). All results derive from 100 repeated experiments. The standard errors of the MSE's were between 0.001 and 0.003 for $n = 1024$ and between 0.0005 and 0.0007 for $n = 65{,}536$.

These results should be compared with the behavior of $2\log(n/k)$-style penalties. For the estimator of Foster and Stine [23], minimizing $\mathrm{RSS} + \sigma^2 \sum_{j=1}^{k} 2 \times \log(n/j)$, we have that for $n = 1024$, MSE / MSE($t^*$) $= 1.2308$ while for $n = 65{,}536$, MSE / MSE($t^*$) $= 1.2281$. This is consistent with behavior that would result from FDR control with $q = 0.3$ for $n = 1024$ and $q = 0.25$ for $n = 65{,}536$.

In Figure 4, we display simulation results under a range of sample sizes. Apparently the minimum MSE occurs somewhere below $q = 1/2$.

We propose the following interpretations:

INT 1. FDR procedures with $q \le 1/2$ have a risk which is a reasonable multiple of the "ideal risk" based on the threshold which would have been optimal for the given sparsity of the object. The ratios in Table 1 do not differ much for various $q \le 1/2$, which demonstrates robustness of the FDR procedures toward the choice of $q$.

INT 2. An FDR procedure with $q$ near $1/2$ appears to outperform $q$-small procedures at this configuration, achieving risks which are roughly comparable to the ideal risk.

INT 3. Avoid FDR procedures with large $q$, in favor of $q \le 1/2$.

1.12. *Contents.* The paper to follow is far more technical than the Introduction, in our view necessarily so, since much of the work concerns refined properties

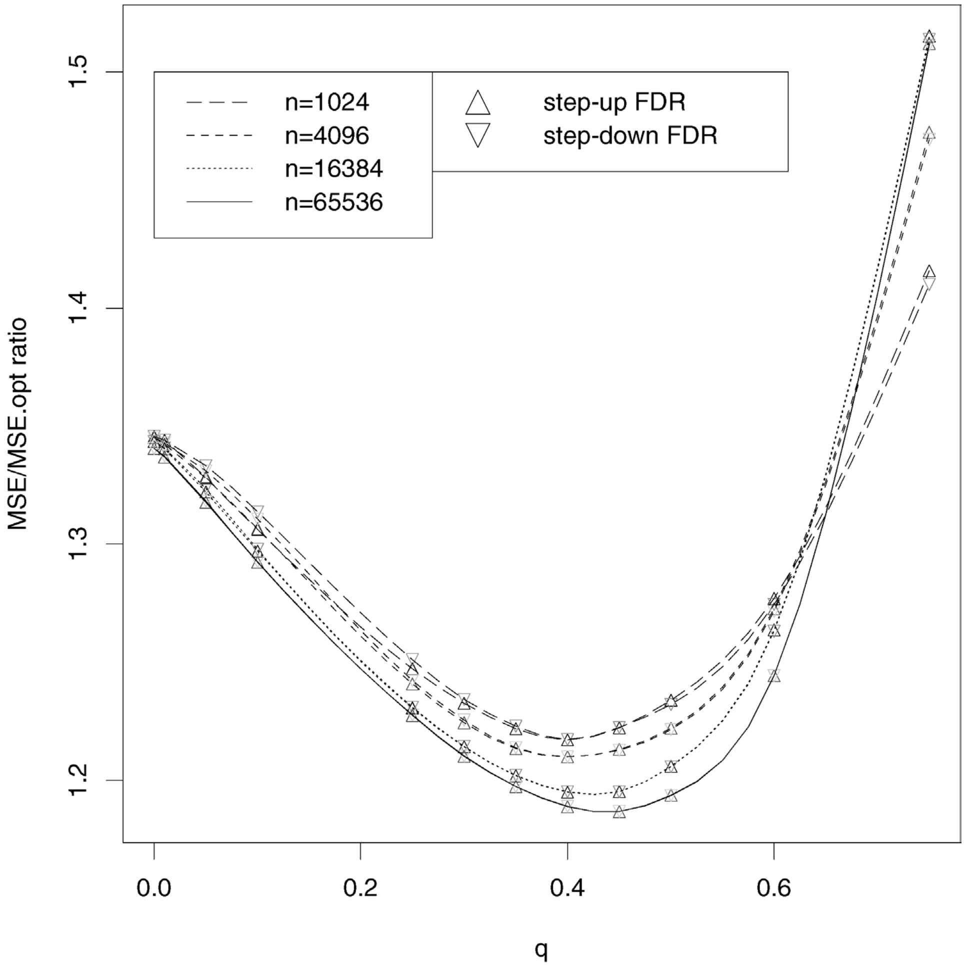


FIG. 4. *Ratios of* MSE(FDR)/MSE($t^*(p,n)$), $p = 3/2$.

of fluctuations in the extreme upper tails of the normal order statistics. However, Sections 2–4 should be accessible on a first reading. They review pertinent information about FDR-controlling procedures and about minimax estimation over $\ell_p$ balls, and parse our main result into an upper bound result and a lower bound result. Section 4 then gives an overview of the paper to follow, which carries out rigorous proofs of the lower bound (Sections 5–8, 13) and the upper bound (Sections 9–11).

**2. The false discovery rate.** The field of multiple comparisons has developed many techniques to control the increased rate of type I error when testing a family of $n$ hypotheses $H_{0i}$ versus $H_{1i}$, $i = 1, 2, \dots, n$. The traditional approach is to control the *familywise error rate* at some level $\alpha$, that is, to use a testing procedure that controls at level $\alpha$ the probability of erroneously rejecting even one true $H_{0i}$.

The venerable Bonferroni procedure test ensures this by testing each hypothesis at the $\alpha/n$ level.

The Bonferroni procedure is criticized as being too conservative, since it often lacks power to detect the alternative hypotheses. Much research has been devoted to devise more powerful procedures: tightening the probability inequalities, and incorporating the dependency structure when it is known. For surveys, see Hochberg and Tamhane [26] and Shaffer [37]. In one fundamental sense the success has been limited. Generally the power deteriorates substantially when the problem is large. As a result, many practitioners avoid altogether using any multiplicity adjustment to control for the increased type I errors caused by simultaneous inference.

Benjamini and Hochberg [6] argued that the control of the familywise error rate is a very conservative goal which is not always necessary. They proposed to control the expected ratio of the number of erroneously rejected hypotheses to the number rejected—the false discovery rate (FDR). Formally, for any fixed configuration of true and false hypotheses, let $V$ be the number of true null hypotheses erroneously rejected, among the $R$ rejected hypotheses. Let $Q$ be $V/R$ if $R > 0$, and 0 if $R = 0$; set $\mathrm{FDR} = E\{Q\}$, where the expectation is taken according to the same configuration. The FDR is equivalent to the familywise error rate when all tested hypotheses are true, so an FDR-controlling procedure at level $q$ controls the probability of making even one erroneous discovery in such a situation. Thus for many problems the value of $q$ is naturally chosen at the conventional levels for tests of significance. The FDR of a multiple-testing procedure is never larger than the familywise error rate. Hence controlling FDR admits more powerful procedures.

Here is a simple *step-up* FDR-controlling procedure. Let the individual $P$-values for the hypotheses $H_{0i}$ be arranged in ascending order: $P_{[1]} \le \cdots \le P_{[n]}$. Compare the ordered $P$-values to a linear boundary $i/nq$, and note the last crossing time,

$$\hat{k}_F = \max\{k : P_{[k]} \le (k/n)q\}. \tag{2.1}$$

The FDR multiple-testing procedure is to reject all hypotheses $H_{(0i)}$ corresponding to the indices $i = 1, \ldots, \hat{k}_F$. If $\hat{k}_B$ denotes the number of $P$-values below the Bonferroni cutoff $q/n$, it is apparent that $\hat{k}_F \ge \hat{k}_B$ and hence that the FDR test conducted at the same level is necessarily less conservative.

Benjamini and Hochberg [6] considered the above testing procedure in the situation of independent hypothesis tests on many individual means. They considered the two-sided $P$-values from testing that each individual mean was zero. They found that the false discovery rate of the above multiple-testing procedure is bounded by $q$ *whatever be the number of true null hypotheses* $n_0$ *or the configuration of the means under the alternatives*:

$$\mathrm{FDR} = E_\mu\{Q\} = qn_0/n \le q \qquad \text{for all } \mu \in \mathbb{R}^n. \tag{2.2}$$

The multiple-testing procedure (2.1) was proposed informally by Elkund, by Seeger [36] and much later independently by Simes [38]. Each time it was neglected because it was shown not to control the familywise error rate [27, 36]. In the absence of the FDR concept, it was not understood why this procedure could be a good idea. After introduction of the FDR concept, it was recognized that $\hat{k}_F$ had the FDR property, but also that other procedures offered FDR control—most importantly for us, the step-down estimator $\hat{k}_G$ [34]. This rule, introduced in Section 1, will also be referred to frequently below, and our theorems are also applicable to thresholding estimators based upon it.

As noted in the Introduction, Abramovich and Benjamini [1] adapted FDR testing to the setting of estimation, in particular of wavelet coefficients of an unknown regression function. In this setting, given $n$ data on a unknown function observed in Gaussian white noise, we are testing $n$ independent null hypotheses on a function's wavelet coefficients, $\mu_i = 0$. Using the above formulation with two-sided $P$-values, we obtain (1.8).

Previously in the same setting of wavelet estimation, Donoho and Johnstone [15] had proposed to estimate wavelet coefficients by setting to zero all coefficients below a certain "universal threshold" $\sqrt{2\log(n)}\sigma_n$. A key observation in [13] and [18] about this threshold is that, with high probability, every truly zero wavelet coefficient is estimated by zero.

Using ideas from simultaneous inference we can look at universal thresholding differently. The likelihood ratio test of the null hypothesis $H_{0i}: \mu_i = 0$ rejects if and only if $|y_i| > t\sigma$, and the Bonferroni method at familywise level $\alpha$ sets the cutoff for rejection $t$ at $t_{\mathrm{BON}} = \sigma z(\alpha/2n)$. Now very roughly, $z(1/n) \sim \sqrt{2\log(n)}$; much more precise results are derived below and lie at the center of our arguments. Hence, Bonferroni at any reasonable level $\alpha$ leads us to set a threshold not far from the universal threshold. Put another way, universal thresholding may be viewed as precisely a Bonferroni procedure, for $\alpha = \alpha_n^U$. We can derive $\alpha_n^U \asymp 1/\sqrt{\log(n)}$ as $n \to \infty$.

As was emphasized by Abramovich and Benjamini [2], the FDR estimator can choose lower thresholds than $\sigma_n\sqrt{2\log n}$ when $\hat{k}_F$ is relatively large. It thus offers the possibility of adapting to the unknown mean vector by adapting to the data, choosing less conservative thresholds when significant signal is present. It is this possibility we explore here.

*Pointers to the FDR literature more generally.* The above discussion of FDR thresholding emphasizes just that "slice" of the FDR literature needed for this paper, so it is highly selective. The literature of FDR methodology is growing rapidly, and is too diverse to adequately summarize here. Recent papers have illuminated the FDR from different points of view: asymptotic, Bayesian, empirical Bayes and as the limit of empirical processes [20, 24, 40].

Another line of work, starting with Benjamini and Hochberg [7], addresses the factor $n_0/n$ in (2.2); many methods have been offered to estimate this, followed by the step-up procedure with the adjusted (larger) $q$. The results of Benjamini, Krieger and Yekutieli [8] and Storey, Taylor and Siegmund [41] assure FDR control under independence. When $n_0/n$ is close to 1, as is our case in this paper, such methods are close to the original step-up procedure.

An immediate next step beyond this paper would be to study dependent situations. The FDR-controlling property of the step-up procedure under positive dependence has been established in [9], and similar results were derived for the step-down version in [34]. Since much of the formal structure below is based on marginal properties of the observations, this raises the possibility that our estimation results would extend to a broader class of situations involving dependence in the noise terms $z_i$.

**3. Minimax estimation on $\ell_0, \ell_p, m_p$.** As a prelude to the formulation of the adaptive minimaxity results, we review information [15, 17, 29] on minimax estimation over $\ell_0$, $\ell_p$ and weak-$\ell_p$ balls in the sparse case: $0 < p < 2$ and with normalized radius $\eta_n \to 0$ as $n \to \infty$. Throughout this section, we suppose a shift Gaussian model (1.1) with unit noise level $\sigma_n = 1$. We will denote the risk of an estimator $\hat{\mu}$ under $\ell_r$ loss by

$$\rho(\hat{\mu}, \mu) = E_\mu \|\hat{\mu} - \mu\|_r^r.$$

Particularly important classes of estimators are obtained by thresholding of individual coordinates: hard thresholding was defined at (1.2), while soft thresholding of a single coordinate $y_1$ is given by $\eta_S(y_1, t) = \mathrm{sgn}(y_1)(|y_1| - t)_+$. We use a special notation for the risk function of thresholding on a single scalar observation $y_1 \sim N(\mu_1, 1)$:

$$\rho_S(t, \mu_1) = E_{\mu_1} |\eta_S(y_1, t) - \mu_1|^r,$$

with an analogous definition of $\rho_H(t, \mu_1)$ for hard thresholding.

3.1. *$\ell_0$ balls.* Asymptotically least-favorable configurations for $\ell_0$ balls $\ell_0[\eta_n]$ can be built by drawing the $\mu_i$ i.i.d. from sparse two-point prior distributions

$$\pi = (1 - \eta_n)\delta_0 + \eta_n \delta_{\mu_n}, \qquad \mu_n \sim (2 \log \eta_n^{-1})^{1/2}.$$

The precise definition of $\mu_n$ is given in the remark below. The expected number of nonzero components $\mu_i$ is $k_n = n\eta_n$. The prior is constructed so that the corresponding Bayes estimator essentially estimates zero even for those $\mu_i$ drawn from the atom at $\mu_n$, and so the Bayes estimator has an $l_r$ risk of at least $k_n \mu_n^r$. A corresponding asymptotically minimax estimator is given by soft or hard thresholding at threshold $\tau_\eta = \tau(\eta_n) := (2 \log \eta_n^{-1})^{1/2} \sim \mu_n$ as $n \to \infty$. This estimator achieves the precise asymptotics of the minimax risk, namely:

$$R_n(\ell_0[\eta_n]) \sim k_n \mu_n^r = n\eta_n \mu_n^r \sim n\eta_n (2 \log \eta_n^{-1})^{r/2}. \tag{3.1}$$

REMARK 1. Given a sequence $a_n^2 = o(\log \eta_n^{-1})$ that increases slowly to $\infty$, $\mu_n$ is defined as the solution of the equation $\phi(a_n + \mu_n) = \eta_n \phi(a_n)$, where $\phi$ denotes the standard Gaussian density. Equivalently, $\mu_n^2 + 2a_n\mu_n = 2\log \eta_n^{-1} = \tau_\eta^2$, giving the more precise relation

$$\tau_\eta = \mu_n + a_n + o(a_n). \tag{3.2}$$

Thus $\tau_\eta - \mu_n \to \infty$, which for both soft and hard thresholding at $\tau_\eta$ indicates $\rho(\tau_\eta; \mu_n) \sim \mu_n^r$. (Note also that, to simplify notation, we are using $\mu_n$ to denote a sequence of constants rather than the $n$th component of the vector $\mu$.)

3.2. *$\ell_p$ balls.* Again, asymptotically least-favorable configurations for $\ell_p[\eta_n]$ are obtained by i.i.d. draws from $\pi = (1 - \beta_n)\delta_0 + \beta_n \delta_{\mu_n}$, where now the mass of the nonzero atom and its location are informally given by the pair of properties

$$\beta_n = \eta_n^p \mu_n^{-p}, \qquad \mu_n \sim (2\log \beta_n^{-1})^{1/2}. \tag{3.3}$$

More precisely, $\mu_n = \mu_n(\eta_n, a_n; p)$ is now the solution of $\phi(a_n + \mu_n) = \beta_n \phi(a_n)$, which implies that

$$\mu_n \sim \tau_\eta = (2\log \eta_n^{-p})^{1/2}, \qquad n \to \infty, \tag{3.4}$$

and then that (3.2) continues to hold for $\ell_p$ balls. The expected number of nonzero components $\mu_i$ is now $n\beta_n = n\eta_n^p \mu_n^{-p}$. For later use, we define

$$k_n = n\eta_n^p \tau_\eta^{-p}; \tag{3.5}$$

since $\mu_n \sim \tau_\eta$, we have $k_n \sim n\beta_n$, and so $k_n$ is effectively the nonzero number. With similar heuristics for the Bayes estimator, the exact asymptotics of minimax risk becomes

$$R_n(\ell_p[\eta_n]) \sim k_n \mu_n^r = n\eta_n^p \tau_\eta^{r-p} = n\eta_n^p (2\log \eta_n^{-p})^{(r-p)/2}. \tag{3.6}$$

Asymptotic minimaxity is had by thresholding at

$$t_{\eta_n} = (2\log \eta_n^{-p})^{1/2} \sim (2\log n/k_n)^{1/2}.$$

3.3. *Weak-$\ell_p$ balls.* The weak-$\ell_p$ ball $m_p[\eta_n]$ contains the corresponding strong-$\ell_p$ ball $\ell_p[\eta_n]$ with the same radius, and the asymptotic minimax risk is larger by a constant factor:

$$R_n(m_p[\eta_n]) \sim \bigl(r/(r-p)\bigr) R_n(\ell_p[\eta_n]), \qquad n \to \infty. \tag{3.7}$$

Let $F_p(x) = 1 - x^{-p},\ x \geq 1$, denote the distribution function of the Pareto($p$) distribution and let $X$ be a random variable having this law. Then an asymptotically least-favorable distribution for $m_p[\eta_n]$ is given by drawing $n$ i.i.d. samples from the univariate law

$$\pi_1 = \mathcal{L}\bigl(\min(\eta_n X, \mu_n)\bigr), \tag{3.8}$$

where $\mu_n$ is defined exactly as in the strong case. The mass of the prior probability atom at $\mu_n$ equals $\int_{\mu_n}^{\infty} F_p(dx/\eta_n) = \eta_n^p \mu_n^{-p} = \beta_n$, again as in the strong case. Thus, the weak prior can be thought of as being obtained from the strong prior by smearing the atom at 0 out over the interval $[\eta_n, \mu_n]$ according to a Pareto density with scale $\eta_n$. One can see the origin of the extra factor in the minimax risk from the following outline (for details when $r = 2$, see [29]). The minimax theorem says that $R(m_p[\eta_n])$ equals the Bayes risk of the least-favorable prior. This prior is roughly the product of $n$ copies of $\pi_1$, and the corresponding Bayes estimator is approximately (for large $n$) soft thresholding at $\tau_\eta$, so

$$R_n(m_p[\eta_n]) \sim n \int \rho_S(\tau_\eta, \mu)\pi_1(d\mu).$$

Now consider an approximation to the risk function of soft thresholding, again at threshold $t_\eta$. Indeed, using the estimate $\rho_S(t, \mu) \doteq \rho_S(t, 0) + |\mu|^r$, appropriate in the range $0 \le \mu \le \mu_n$, ignoring the term $\rho_S(t, 0)$ and reasoning as before (3.6), we find

$$R_n(m_p[\eta_n]) \sim n \int_{\eta_n}^{\mu_n} \rho_S(t_\eta, \mu) F_p(d\mu/\eta_n) + n\beta_n \rho_S(t_\eta, \mu_n) \tag{3.9}$$

$$\doteq n \int_{\eta_n}^{\mu_n} \mu^r p \eta_n^p \mu^{-p-1} \, d\mu + k_n \mu_n^r \tag{3.10}$$

$$\doteq \left[\frac{p}{r-p} + 1\right] n \eta_n^p \mu_n^{r-p} \sim \frac{r}{r-p} R(\ell_p[\eta_n]). \tag{3.11}$$

Comparison with (3.6) shows that the second term in (3.9)–(3.10) corresponds exactly to $R(\ell_p[\eta_n])$; the first term is contributed by the Pareto density in the weak-$\ell_p$ case.

**4. Adaptive minimaxity of FDR thresholding.** We now survey the path to our main result, providing in this section an overview of the remainder of the paper and the arguments to come.

What we ultimately prove is broader than the result given in the Introduction, and the argument will develop several ideas seemingly of broader interest.

4.1. *General assumptions.* Continually below we invoke a collection of assumptions (Q), (H), (E) and (A) defined here.

*False discovery control.* We allow false discovery rates to depend on $n$, but approach a limit as $n \to \infty$. Moreover, if the limit is zero, rates should not go to zero very fast. Formally define the assumption:

(Q) Suppose that $q_n \to q \in [0, 1)$. If $q = 0$, assume that $q_n \ge b_1/\log n$.

The constant $b_1 > 0$ is arbitrary; its value could be important at a specific $n$.

*Sparsity of the estimand.* We consider only parameter sets which are sparse, and we place quantitative upper bounds keeping them away from the "dense" case. Formally define:

(H) $\eta_n$ (for $\ell_0[\eta_n]$) and $\eta_n^p$ (for $m_p[\eta_n]$) lie in the interval $[n^{-1}\log^5 n, b_2 n^{-b_3}]$.

Here the constants $b_2 > 0$ and $b_3 > 0$ are arbitrary; their chosen values could again be important in finite samples.

*Diversity of estimators.* Our results apply not just to the usual FDR-based estimator $\hat{\mu}_F$ of (1.7) but also to the penalty-based estimator $\hat{\mu}_r$ of (1.11). More generally, recall the terms $[\hat{t}_F, \hat{t}_G]$ defined in Section 1.8. Under formal assumption (E), we consider any estimator $\hat{\mu}$ obeying

$$\hat{\mu} = \text{hard thresholding at } \hat{t} \in [\hat{t}_F, \hat{t}_G] \text{ w.p. } 1. \tag{4.1}$$

*Notation.* We introduce a sequence $\alpha_n$ which often appears in estimates in Sections 7 and 8 and in dependent material. We also define constants $q'$ and $q''$. Formally:

(A) Set $\alpha_n = 1/(b_4\tau_\eta)$, with $b_4 = (1-q)/4$. Also set $q' = (q+1)/2$ and $q'' = (1-q)/2 = 1-q'$.

Finally, as a global matter, we suppose that our observations $y \sim N_n(\mu, I)$; thus $\sigma_n^2 \equiv 1$. For estimation of $\mu$, we consider $\ell_r$ risk (1.9), $0 < r \le 2$, and minimax risk $R_n(\Theta_n)$ of (1.10). Here the parameter spaces are $\Theta_n = \ell_0[\eta_n]$ or $\ell_p[\eta_n]$ or $m_p[\eta_n]$ defined by (1.3), (1.6) and (1.4), respectively, with $0 < p < r$.

4.2. *Upper bound result.* Our argument for Theorem 1.1 splits into two parts, beginning with an upper bound on minimax risk.

THEOREM 4.1. *Assume* (H), (E), (Q). *Then, as $n \to \infty$,*

$$\sup_{\mu\in\Theta_n} \rho(\hat{\mu}, \mu) \le R_n(\Theta_n)\left\{1 + u_{rp}\frac{(2q_n-1)_+}{1-q_n} + o(1)\right\}, \tag{4.2}$$

*where $u_{rp} = 1-(p/r)$ if $\Theta_n = m_p[\eta_n]$ and $u_{rp} = 1$ if $\Theta_n = \ell_p[\eta_n]$ or $\ell_0[\eta_n]$.*

The bare bones of our strategy for proving the upper bound result were described in the Introduction. The global idea is to study the penalized FDR estimator $\hat{\mu}_2$ of (1.8) and then compare to the behavior of $\hat{\mu}_F$. To make this work, numerous technical facts will be needed concerning the behavior of hard thresholding, the mean and fluctuations of the threshold exceedance process, and so on. As it turns out, those same technical facts form the core of our lower bound on the risk behavior of $\hat{\mu}_F$. As a result, it is convenient for us to study the lower bound and associated technical machinery first, in Sections 5–8 (with some details deferred to Sections 12 and 13), and then later, in Sections 9–11, to prove the upper bound, using results and viewpoints established earlier.

4.3. *Lower bound result.* Theorem 1.1 is completed by a lower bound on the behavior of the FDR estimator.

THEOREM 4.2. *Suppose* (H), (Q). *With notation as in Theorem* 4.1,

$$\sup_{\mu\in\Theta_n} \rho(\hat{\mu}_F, \mu) \geq R_n(\Theta_n)\left\{1 + u_{rp}\frac{(2q_n-1)_+}{1-q_n} + o(1)\right\}, \qquad n\to\infty, \tag{4.3}$$

*where* $u_{rp} = 1-(p/r)$ *for* $\Theta_n = m_p[\eta_n]$, *and* $u_{rp} = 1$ *for* $\Theta_n = \ell_p[\eta_n]$.

This bound establishes the importance of $q$, showing that if $q > 1/2$, then certainly FDR cannot be asymptotically minimax. We turn immediately to its proof.

**5. Proof of the lower bound.** The proof involves three technical but significant ideas. First, it bounds the number of discoveries made by FDR, as a function of the underlying means $\mu$. Second, it studies the risk of ordinary hard thresholding with nonadaptive threshold in a specially chosen, (quasi-) least-favorable one-parameter subfamily of $\Theta_n$. Finally, it combines these elements to show that, on this least-favorable subfamily, $\hat{\mu}_F$ behaves like hard thresholding with a particular threshold. The lower bound result then follows. Unavoidably, the results in this section will invoke lemmas and corollaries only proven in later sections.

Beyond simply proving the lower bound, this section introduces some basic viewpoints and notions. These include:

(a) a threshold exceedance function $M$, which counts the number of threshold exceedances as a function of the underlying means vector,

(b) a special "coordinate system" for thresholds, mapping thresholds $t$ onto the scale of relative expected exceedances,

(c) a special one-parameter (quasi-) least-favorable family for FDR, at which the lower bound is established.

These notions will be used heavily in later sections.

5.1. *Mean exceedances and mean discoveries.* Define the exceedance number $N(t_k) = \#\{i : |y_i| \geq |t_k|\}$. Since $|y_{(k)}| \geq |t_k|$ if and only if $N(t_k) \geq k$, we are interested in the values of $k$ for which $N(t_k) \approx k$. (See Section 7 for details.)

Throughout the paper we will refer to the mean threshold exceedance function, counting the mean number of exceedances over threshold $t_k$ as $k$ varies,

$$M(k;\mu) = E_\mu N(t_k) = \sum_{l=1}^{n} P_\mu(|y_l| \geq t_k) = \sum_{l=1}^{n} \Phi([\mu_l - t_k, \mu_l + t_k]^c). \tag{5.1}$$

[Here $\Phi(A)$ denotes the probability of event $A$ under the standard Gaussian probability distribution, and $k$ is extended from positive integers to positive *real* values.] If $\mu = 0$, then $M(k;\mu) = 2n\tilde{\Phi}(t_k) = qk$, reflecting the fact that in the null case, the fraction of exceedances is always governed by the FDR parameter $q$. If $\mu \neq 0$, we expect that $\hat{k}_F$ will be close to the *mean discovery number*

$$k(\mu) = \inf\{k \in \mathbb{R}^+ : M(k; \mu) = k\}. \tag{5.2}$$

The existence and uniqueness of $k(\mu)$ when $\mu \neq 0$ follows from facts to be established in Section 6.3: that (taking $k$ as real-valued) the function $k \to M(k; \mu)/k$ decreases strictly and continuously from a limit of $+\infty$ as $k \to 0$ to a limit $< 1$, as $k \to n$.

A key point is that the mean discovery number is bounded over the parameter spaces $\Theta_n$. The mean discovery number is monotone in $\mu$: if $|\mu_1|_{(k)} \geq |\mu_2|_{(k)}$ for all $k$, then $k(\mu_1) > k(\mu_2)$. Thus, on $\ell_0[\eta_n]$, the largest mean discovery number $\overline{M}$, say, is obtained by taking $k_n = [n\eta_n]$ components to be very large. Writing this out,

$$\begin{aligned}\overline{M}_n(k) &= \sup_{\ell_0[\eta_n]} M(k; \mu) = k_n + 2(n - k_n)\tilde{\Phi}(t_k)\\ &= k_n + (1 - k_n/n)kq_n \sim k_n + kq_n,\end{aligned}$$

using the definition of $t_k = z(kq_n/2n)$ and $\eta_n \sim k_n/n \to 0$. The first term corresponds to "true" discoveries, and the second to "false" ones. Solving $\overline{M}(k) = k$ yields a solution

$$\tilde{k} = k_n/\big(1 - (1 - n^{-1}k_n)q_n\big) \sim k_n/(1 - q_n). \tag{5.3}$$

In particular, for all $\mu \in \ell_0[\eta_n]$, we have $k(\mu) \leq k_n/(1 - q_n(1 + o(1))$.

*Weak $\ell_p$.* On $\Theta_n = m_p[\eta_n]$, $E_\mu N(t_k)$ is maximized by taking the components of $\mu$ as large as possible—that is, at the coordinatewise upper bound $\bar{\mu}_l = \eta_n(n/l)^{1/p}$. Thus now

$$\overline{M}_n(k) = \sup_{m_n[\eta_n]} E_\mu N(t_k) = E_{\bar{\mu}} N(t_k).$$

To approximate $M(k; \bar{\mu})$, note first that the summands in (5.1) are decreasing from nearly 1 for $\mu_l$ large to $2\tilde{\Phi}(t_k)$ when $\mu_l$ is near 0. With $k$ held fixed, break the sum into two parts using $k_n = n\eta_n^p \tau_\eta^{-p}$. [This choice is explained in more detail after (9.13) below.] For $l \leq k_n$, the summands are mostly well approximated by 1, and for $l \geq k_n$ predominantly by $2\tilde{\Phi}(t_k) = qk/n$. Since $k_n/n \approx 0$, we have

$$\begin{aligned}M(\nu; \bar{\mu}) &\approx k_n + (n - k_n)q_n\nu/n\\ &\approx n\eta_n^p\tau_\eta^{-p} + q_n\nu.\end{aligned}$$

Again the first term tracks "true" discoveries and the second "false" ones. Solving $M(\nu; \bar{\mu}) = \nu$ based on this approximation suggests that, just as in (5.3),

$$k(\bar{\mu}) \leq k_n/(1 - q_n)\big(1 + o(1)\big). \tag{5.4}$$

The full proof is given in Section 6.4.4.

*Strong* $\ell_p$. Since $\ell_p[\eta_n] \subset m_p[\eta_n]$, (5.4) applies here as well.

5.2. *Typical behavior of* $\hat{k}_F$ *and* $\hat{k}_G$. We turn to the stochastic quantities $\hat{k}_F$ and $\hat{k}_G$. These are defined in terms of the exceedance numbers $N(t_k)$, which themselves depend on independent (and nonidentically distributed) Bernoulli variables. This suggests the use of bounds on $\hat{k}_F$ and $\hat{k}_G$ derived from large deviations properties of $N(t_k)$. Since we are concerned mainly with relatively high thresholds $t_k$, results appropriate to Poisson regimes are required. Details are in Section 7.

To describe the resulting bounds on $[\hat{k}_G, \hat{k}_F]$, we first introduce some terminology. We say that an event $A_n(\mu)$ is $\Theta_n$-likely if there exist constants $c_0, c_1$ not depending on $n$ and $\Theta$ such that

$$\sup_{\mu \in \Theta_n} P_\mu\{A_n^c(\mu)\} \le c_0 \exp\{-c_1 \log^2 n\}.$$

With $\alpha_n$ as in assumption (A), define

$$k_-(\mu) = \begin{cases} k(\mu) - \alpha_n k_n, & k(\mu) \ge 2\alpha_n k_n, \\ 0, & k(\mu) < 2\alpha_n k_n, \end{cases} \tag{5.5}$$

and

$$k_+(\mu) = k(\mu) \vee \alpha_n k_n + \alpha_n k_n. \tag{5.6}$$

PROPOSITION 5.1. *Assume* (Q), (H) *and* (A). *For each of the parameter spaces* $\Theta_n$, *it is* $\Theta_n$*-likely that*

$$k_-(\mu) \le \hat{k}_G \le \hat{k}_F \le k_+(\mu).$$

Thus all the penalized estimates $\hat{k}_r$ (and any $\hat{k} \in [\hat{k}_G, \hat{k}_F]$) are with uniformly high probability bracketed between $k_-(\mu)$ and $k_+(\mu)$. In particular, note that

$$k_+(\mu) - k_-(\mu) \le 3\alpha_n k_n, \tag{5.7}$$

and so the fluctuations in $\hat{k}_r$ are typically small compared to the maximal value over $\Theta_n$.

Here and below it is convenient to have a notational variant for $t_k$, used especially when the subscript would be very complicated; so define

$$t[k] = z(2n/kq);$$

keep in mind that $t$ depends implicitly on $q = q_n$ and $n$. We occasionally use $t_k$ when the subscript is very simple.

Giving this notation its first workout, the thresholds $\hat{t}_r$ are bracketed between

$$t_+(\mu) = t[k_-(\mu)] \quad \text{and} \quad t_-(\mu) = t[k_+(\mu)]. \tag{5.8}$$

Note that $t_+ > t_-$, but from (5.7) and (12.13), it follows that $t_+/t_- \le 1 + 3\alpha_n/t_-^2$.

5.3. *Risk of hard thresholding.* We now study the error of *fixed* thresholds as a prelude to the study of the data-dependent FDR thresholds. We define one-parameter families of configurations and of thresholds which exhibit key transitional behavior. As might be expected, these are concentrated around the critical threshold $\tau_\eta = \sqrt{2\log \eta_n^{-1}}$ corresponding to sparsity level $\eta_n$.

Consider first a family of (quasi-) least-favorable means $\mu_\alpha$. The coordinates take one of two values, most being zero, and the others amounting to a fraction $\eta$ with value roughly $\tau_\eta + \alpha$. Specifically, for $\alpha \in \mathbb{R}$, set

$$\mu_{\alpha,l} = \begin{cases} t[k_n] + \alpha, & l \le k_n, \\ 0, & k_n < l \le n. \end{cases} \tag{5.9}$$

In a sense $\mu_{0,k_n}$ is right at the FDR boundary, while with $\alpha > 0$, $\mu_{\alpha,k_n}$ is above the boundary and with $\alpha < 0$ it is below the boundary.

Next, consider a "coordinate system" for measuring the height of thresholds in the vicinity of the FDR boundary. Think of thresholds $\{t : t > 0\}$ as generated by $\{t[ak_n], a > 0\}$, with $a$ fixed while $n$ and $k_n$ increase. For $a = 1$, we are on the FDR boundary at $k_n$, so that $a < 1$ is above the boundary and $a > 1$ is below the boundary. The "coordinate" $a$ will be heavily used in what follows.

In fact, these thresholds vary only slowly with $a$: for $a$ fixed, as $n \to \infty$,

$$|t[ak_n] - t[k_n]| \le c(a)\tau_\eta^{-1}. \tag{5.10}$$

Nevertheless, the effect of $a$ is visible in the leading term of the risk:

PROPOSITION 5.2. *Let $\alpha \in \mathbb{R}$ and $a > 0$ be fixed. Let the configuration $\mu_\alpha \in \ell_0[\eta]$ be defined by* (5.9). *For $\ell_r$ loss, the risk of hard thresholding at $t[ak_n]$ is given, as $n \to \infty$, by*

$$\rho(\hat{\mu}_{H,t[ak_n]}, \mu_\alpha) = [\tilde{\Phi}(\alpha) + aq_n]k_n\tau_\eta^r(1 + o(1)). \tag{5.11}$$

Here $k_n\tau_\eta^r$ is asymptotic to the minimax risk for $\ell_0[\eta_n]$—compare (3.1)—and so defines the benchmark for comparison.

The two leading terms in (5.11) reflect false negatives and false positives, respectively. The proof is given in Section 13. Here we aim only to explain how these terms arise.

The false-negative term $\tilde{\Phi}(\alpha)k_n\tau_\eta^r$ decreases as $\alpha$ increases. This is natural, as the signals with mean $m_\alpha = t[k_n] + \alpha$ become easier to detect as $\alpha$ increases—whatever be the threshold $t[ak_n]$. More precisely, the $\ell_r$ error due to nondetection, $|y_l| \le t[ak_n]$, on each of the $k_n$ terms with mean $m_\alpha$ contributes risk

$$k_n m_\alpha^r P_{m_\alpha}(|y_l| < t[ak_n]) \sim k_n\tau_\eta^r\Phi(t[ak_n] - m_\alpha), \tag{5.12}$$

since $m_\alpha \sim \tau_\eta$ as $n \to \infty$. Finally, (5.10) shows that

$$m_\alpha - t[ak_n] = \alpha + t[k_n] - t[ak_n] = \alpha + O(\tau_\eta^{-1}), \tag{5.13}$$

so that (5.12) is approximately $\tilde{\Phi}(\alpha)k_n\tau_\eta^r$.

The false-positive term shows a relatively subtle dependence on threshold $a \to t[ak_n]$. There are $n - k_n$ means that are exactly 0, and so the risk due to false discoveries is

$$(n-k_n)E\{|z|^r : |z| > t[ak_n]\} \sim 2nt[ak_n]^r\tilde{\Phi}(t[ak_n]) \tag{5.14}$$

$$= ak_nq_nt[ak_n]^r \tag{5.15}$$

$$\sim aq_nk_n\tau_\eta^r. \tag{5.16}$$

(Equation (5.14) follows from (8.6) below, while (5.15) uses the definition of the FDR boundary, $t[k_n] = \tilde{\Phi}^{-1}(k_nq_n/2n)$, and finally (5.16) follows from (12.18) below.)

*Weak $\ell_p$.* In this case, we replace the two-point configuration by Winsorized analogs in the spirit of Section 3.3:

$$\mu_{\alpha,l} = \bar{\mu}_l \wedge m_\alpha, \qquad m_\alpha = t[k_n] + \alpha. \tag{5.17}$$

Now an extra term appears in the risk of hard thresholding when using thresholds $t[ak_n]$:

PROPOSITION 5.3. *Adopt the setting of Proposition 5.2, replacing only (5.9) by (5.17). Then*

$$\rho(\hat{\mu}_{H,t[ak_n]}, \mu_\alpha) = \left[\tilde{\Phi}(\alpha) + \frac{p}{r-p} + aq_n\right]k_n\tau_\eta^r(1+o(1)).$$

The same phenomena as for $\ell_0[\eta]$ apply here, *except* that the $p/(r-p)$ term arises due to the cumulative effect of missed detections of means $\bar{\mu}_l$ that are smaller than $m_\alpha$ but certainly not 0. This term decreases as $p$ becomes smaller, essentially due to the increasingly fast decay of $\bar{\mu}_l = c_nl^{-1/p}$. The term disappears in the $p \to 0$ limit, and we recover the $\ell_0$ risk (5.11). This result also is proved in Section 13.

5.4. *FDR on the least-favorable family.* To track the response of the FDR estimator to members of the family $\{\mu_\alpha, \alpha \in \mathbb{R}\}$, we look first at the mean discovery numbers. In Section 13 we prove:

PROPOSITION 5.4. *Assume* (Q), (H) *and* (A). *Fix $\alpha \in \mathbb{R}$ and define $\mu_\alpha$ by* (5.9) *and* (5.17) *for $\ell_0[\eta_n]$ and $m_p[\eta_n]$, respectively. Then as $n \to \infty$,*

$$k(\mu_\alpha) \sim (1-q_n)^{-1}\Phi(\alpha)k_n. \tag{5.18}$$

Heuristically, for $\ell_0[\eta_n]$, we approximate

$$M(k;\mu_\alpha) = k_n[\tilde{\Phi}(t_k - m_\alpha) + \Phi(-t_k - m_\alpha)] + 2(n-k_n)\tilde{\Phi}(t_k) \tag{5.19}$$
$$\sim k_n\Phi(m_\alpha - t_k) + q_n k \sim k_n\Phi(\alpha) + q_n k,$$

from (5.13). Solving $M(k;\mu_\alpha) = k$ based on this approximation leads to (5.18). The same approach works for $m_p[\eta_n]$, but with more attention needed to bounding the component terms in $M(k;\mu_\alpha)$ as detailed in Section 6.

Proposition 5.4 suggests that at configuration $\mu_\alpha$, FDR will choose a threshold close to $t[k(\mu_\alpha)]$, which is of the form $t[ak_n]$ with $a \sim (1-q)^{-1}\Phi(\alpha)$. Thus, as $\alpha$ increases, and with it the nonzero components of $\mu_\alpha$, the FDR threshold *decreases*, albeit modestly.

The risk incurred by FDR at $\mu = \mu_\alpha$ corresponds to that of hard thresholding at $t[k(\mu_\alpha)]$. In Section 13 below we prove:

PROPOSITION 5.5. *Assume* (Q), (H) *and* (A). *Fix* $\alpha \in \mathbb{R}$ *and consider* $\mu_\alpha$ *defined by* (5.9) *for* $\ell_0[\eta_n]$. *Then as* $n \to \infty$,

$$\rho(\hat{\mu}_F, \mu_\alpha) = \left[\tilde{\Phi}(\alpha) + \Phi(\alpha)\frac{q_n}{1-q_n}\right] k_n \tau_\eta^r \bigl(1 + o(1)\bigr). \tag{5.20}$$

*On the other hand, define* $\mu_\alpha$ *using* (5.17) *for* $m_p[\eta_n]$; *then*

$$\rho(\hat{\mu}_F, \mu_\alpha) = \left[\tilde{\Phi}(\alpha) + \frac{p}{r-p} + \Phi(\alpha)\frac{q_n}{1-q_n}\right] k_n \tau_\eta^r \bigl(1 + o(1)\bigr). \tag{5.21}$$

Formula (5.20) shows visibly the role of the FDR control parameter $q$. Note that

$$\sup_\alpha \tilde{\Phi}(\alpha) + \Phi(\alpha)\frac{q}{1-q} = \begin{cases} 1, & q \le 1/2, \\ \dfrac{2q-1}{1-q}, & q > 1/2. \end{cases} \tag{5.22}$$

Consider the implications of this in (5.20) in the $\ell_0[\eta_n]$ case. The minimax risk $\sim k_n\tau_\eta^r$, and so the minimax risk is exceeded asymptotically whenever $q > 1/2$.

We interpret this further. When $q < 1/2$, the worst configurations in $\{\mu_\alpha\}$ correspond to $\alpha$ large and negative, and yield essentially the minimax risk. Indeed, only $\Phi(\alpha)$ of the true nonzero means are discovered. Each missed mean contributes risk $\sim \mu_\alpha^r \sim \tau_\eta^r$ and so the risk due to missed means is given roughly by $\tilde{\Phi}(\alpha)k_n\tau_\eta^r$. The risk contribution due to false discoveries, being controlled by $\Phi(\alpha)$, is negligible in these configurations.

When $q > 1/2$, the worst configurations in $\{\mu_\alpha\}$ correspond to $\alpha$ large and positive. Essentially all of the $k_n$ nonzero components are correctly discovered, along with a fraction $q$ of the $k_n(\mu_\alpha) \sim (1-q)^{-1}\Phi(\alpha)k_n$ which are false discoveries. In the $\ell_0$ case, the false discoveries dominate the risk, yielding an error of order $\Phi(\alpha)[q/(1-q)]k_n\tau_\eta^r$.

When $q = 1/2$, the risk $\rho(\hat{\mu}_F, \mu_\alpha) \sim k_n \tau_\eta^r$ *regardless of* $\alpha$, so that all configurations $\mu_\alpha$ are equally bad, even though the fraction $\Phi(\alpha)$ of risk due to false discoveries changes from 0 to 1 as $m_\alpha = t[k_n] + \alpha$ increases from values below $t[k_n]$ through values above.

5.5. *Proof of the lower bound.* Our interpretation of Proposition 5.5 in effect gave away the idea for the proof of (4.3). We now fill in a few details.

In the $\ell_0[\eta_n]$ case, fix $\varepsilon > 0$. Choosing $\alpha = \alpha(\varepsilon; q)$ sufficiently large positive or negative according as $q > 1/2$ or $q < 1/2$, we get

$$\tilde{\Phi}(\alpha(\varepsilon)) + \frac{q}{1-q}\Phi(\alpha(\varepsilon)) > \left(1 - \frac{\varepsilon}{2}\right) \sup_\alpha \tilde{\Phi}(\alpha) + \frac{q}{1-q}\Phi(\alpha).$$

Equation (5.20) gives that for large $n$,

$$\rho(\hat{\mu}_F, \mu_{\alpha(\varepsilon)}) \geq \left[1 + \frac{(2q-1)_+}{1-q}\right] k_n \tau_\eta^r (1 - \varepsilon).$$

But the family $\mu_\alpha \in \ell_0[\eta_n]$, and $R_n(\ell_0[\eta_n]) \sim k_n \tau_\eta^r$, hence

$$\bar{\rho}(\hat{\mu}_F, \ell_0[\eta_n]) \geq \rho(\hat{\mu}_F, \mu_{\alpha(\varepsilon)}) \geq \left[1 + \frac{(2q-1)_+}{1-q}\right] R_n(\ell_0[\eta_n])(1 - \varepsilon).$$

As this is true for all $\varepsilon > 0$, the $\ell_0[\eta_n]$ case of (4.3) follows.

For the $\ell_p[\eta_n]$ case, fix $\varepsilon > 0$, and choose again $\alpha = \alpha(\varepsilon; q)$ as in the $\ell_0$ case. Note that $m_\alpha$ is implicitly a function $m_\alpha[k_n]$ of the number of nonzeros. Define the pair $\beta_n'$ and $k_n'$ informally as the joint solution of

$$\beta_n' = \eta_n^p (m_\alpha[k_n'])^{-p}, \qquad k_n' = n\beta_n'.$$

(A formal definition can be made using the approach in Section 3.2.) Now the mean vector $\mu_\alpha'$ with $k_n'$ nonzeros each taking value $m_\alpha[k_n']$ gives, again by (5.20), that for large $n$,

$$\rho(\hat{\mu}_F, \mu_\alpha) \geq \left[1 + \frac{(2q-1)_+}{1-q}\right] k_n' (m_\alpha[k_n'])^r (1 - \varepsilon).$$

Now from Section 3.2, $R_n(\ell_p(\eta_n)) \sim n\eta_n^p \tau_\eta^{r-p}$, while

$$k_n'(m_\alpha[k_n'])^r = n\beta_n'(m_\alpha[k_n'])^{-p} = n\eta_n^p (m_\alpha[k_n'])^{r-p} \sim n\eta_n^p \tau_\eta^{r-p}, \qquad n \to \infty.$$

At the same time, $\mu_\alpha \in \ell_p[\eta_n]$. Hence

$$\bar{\rho}(\hat{\mu}_F, \ell_p[\eta_n]) \geq \rho(\hat{\mu}_F, \mu_{\alpha(\varepsilon)}) \geq \left[1 + \frac{(2q-1)_+}{1-q}\right] R_n(\ell_p[\eta_n])(1 - \varepsilon).$$

Again this holds for all $\varepsilon > 0$, and the $\ell_p[\eta_n]$ case of (4.3) follows.

The argument for (4.3) in the $m_p[\eta_n]$ case is entirely parallel, only using (5.21) and the modified definition of $\mu_\alpha$ for the $m_p$ case.

5.6. *Infrastructure for the lower bound.* We look ahead now to the arguments supporting the propositions of this section.

Propositions 5.2–5.4 will be proved in Section 13 at the very end of the paper. The proofs exploit viewpoints and estimates set forth in Sections 6–8. Section 6 sets out properties of the mean detection function $\nu \to M(\nu; \mu)$ of (5.1) and its derivatives, with a view to deriving information and bounds on the discovery number $k(\mu)$ of Section 5.1.

Section 7 applies these bounds in combination with the large deviations bounds to prove Proposition 5.1 and show that $\hat{k}_F - \hat{k}_G \le 3\alpha_n k_n$. Section 8 collects various bounds on the risk of fixed thresholds, and the risk difference between two data dependent thresholds.

All these sections frequently invoke a very useful section, Section 12, which assembles needed properties of the Gaussian distribution, of the quantile function $z(\eta) = \tilde{\Phi}^{-1}(\eta)$ and of implications of the FDR boundary $t_k$.

## 6. The mean detection function.

6.1. *Comparing weak $\ell_p$ with $\ell_0$: the effective nonzero fraction.* A key feature of $\ell_0[\eta_n]$ is that only $k_n = n\eta_n$ coordinates may be nonzero. Consequently, the number of "discoveries" at threshold $t[\nu]$ from $n - k_n$ zero coordinates is at most linear in $\nu$ with slope $q_n$:

$$(n - k_n)p_\nu(0) = n(1 - \eta_n)\nu q_n/n \le q_n\nu, \tag{6.1}$$

since $p_\nu(0) = 2\tilde{\Phi}(t[\nu]) = q_n\nu/n$.

In the case of weak $\ell_p$, the discussion around (9.12)–(9.13) shows that for certain purposes, the index $k_n = n\eta_n^p\tau_\eta^{-p}$ counts the maximum number of "significantly" nonzero coordinates.

In this section we will see that an alternative, slightly larger index, $k_n' = n\eta_n^p\tau_\eta^p$, yields for weak $\ell_p$ the property analogous to (6.1): the number of discoveries at $t_\nu$ from the $n - k_n'$ smallest means $\mu_l$ is not essentially larger than $q_n\nu$. At least for the extremal configuration $(\bar{\mu}_l)$, the range of indices $[k_n, k_n']$ constitutes a "transition" zone between "clearly nonzero" means and "effectively zero" ones; this is discussed further in Section 6.4.

To state the result, we need a certain error-control function; let

$$\delta_p(\varepsilon) = p\varepsilon \int_\varepsilon^1 w^{p-2}\,dw \le \begin{cases} p|p-1|^{-1}\varepsilon^{p\wedge 1}, & p \ne 1, \\ (\log\varepsilon^{-1})\varepsilon, & p = 1. \end{cases} \tag{6.2}$$

LEMMA 6.1. *Assume hypotheses* (Q) *and* (H). *Let $\tau_\eta^2 = 2\log\eta_n^{-p}$ and $\varepsilon_n = \eta_n\tau_\eta$ and $\delta_p(\varepsilon)$ be defined as above. There exists $c = c(b_1, b_3) > 0$ such that for $\nu$ with $t_\nu^2 \ge 2$, we have, uniformly in $\mu \in m_p[\eta_n]$, that*

$$[1 - \varepsilon_n^p]q_n\nu \le \sum_{l > k_n'} p_\nu(\mu_l) \le [1 + c\delta_p(\varepsilon_n)]q_n\nu. \tag{6.3}$$

The proof is deferred to Section 6.4—compare the proof of (6.29) there.

REMARK. Suppose $q_n \to q \in [0, 1)$. Then for $n$ sufficiently large (i.e., $n$ larger than some $n_0$ depending on $p, q$, and the particular sequence $\eta_n$), it follows that

$$[1 + c\delta_p(\varepsilon_n)]q_n \le \tilde{q}_n := \begin{cases} (3/2)q_n, & \text{if } q_n \le 1/2, \\ (1+q)/2, & \text{if } 1/2 < q_n < 1; \end{cases} \tag{6.4}$$

in particular, the right-hand side is strictly less than 1.

We have just defined $\tilde{q}_n$ for the case of $m_p$. If in the nearly-black case ($\ell_0[\eta_n]$) we agree to define $\tilde{q}_n = q_n$, then we may write both conclusions (6.1) and (6.3) in one unified form,

$$\sum_{l > k'_n} p_\nu(\mu_l) \le \tilde{q}_n \nu. \tag{6.5}$$

*The "true positive" rate.* Adopt for a moment the language of diagnostic testing and call those means with $\mu_l \ne 0$ "positives," and those with $\mu_l = 0$ "negatives." In the nearly-black case there are typically $k_n$ positives out of $n$. In the weak-$\ell_p$ case, there are formally as many as $n$ positives, but as argued above, there are *effectively* $k'_n = n\eta_n^p t_\eta^p$ positives, and it is this interpretation that we take here. If it is assumed (without loss of generality) that $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_n \ge 0$, then the *true positive rate* using threshold $t[\nu]$ is defined by

$$\bar{\pi}_\nu(\mu) = (1/k'_n) \sum_1^{k'_n} p_\nu(\mu_l). \tag{6.6}$$

In our sparse settings, if the mean discovery number for $\mu$ exceeds $\nu$, then there is a lower bound on the true positive rate at $\mu$ using threshold $t[\nu]$.

COROLLARY 6.2. *Assume* (Q) *and* (H) *and define* $\tilde{q}_n$ *by* (6.4). *If $n$ is sufficiently large, then uniformly over $\mu$ in $m_p[\eta_n]$ or $\ell_o[\eta_n]$ for which $k(\mu) > \nu$, the true positive rate using threshold $t[\nu]$ satisfies*

$$\bar{\pi}_\nu(\mu) \ge (1 - \tilde{q}_n)(\nu/k'_n).$$

PROOF. From the definition of the mean exceedance number, we have

$$k'_n\bar{\pi}_\nu(\mu) = M(\nu, \mu) - \sum_{l > k'_n} p_\nu(\mu_l).$$

Since $\nu < k(\mu)$ we have $M(\nu; \mu) \ge \nu$, and to bound the sum we use (6.5). Hence $k'_n\bar{\pi}_\nu(\mu) \ge \nu - \tilde{q}_n\nu$, as required. $\square$

6.2. *Convexity properties of exceedances.* The goal of this subsection, Corollary 6.5, shows that a lower bound on the mean discovery number $k(\mu)$ forces a lower bound on the mean threshold function $\nu \to M(\nu; \mu)$ at least on sparse parameter sets. The idea is to establish convexity of a certain power function associated with testing individual components $\mu_l$ and then to use the convexity to construct two-point configurations providing the needed lower bounds.

Let $N(t_k) = \#\{l : |y_l| \geq t_k\}$, and as before $M(k; \mu) = E_\mu N(t_k) = \sum_{\ell=1}^n p_k(\mu_l)$. Here, the exceedance probability for threshold $t_k$ is given by

$$p_k(\mu) = P\{|Z + \mu| > t_k\} = \tilde{\Phi}(t_k - \mu) + \Phi(-t_k - \mu),$$

and we note that as $\mu$ increases from 0 to $\infty$, $p_k$ increases from $p_k(0) = 2\tilde{\Phi}(t_k) = qk/n$ to $p_k(\infty) = 1$. It has derivative

$$p_k'(\mu) = \phi(t_k - \mu) - \phi(t_k + \mu) > 0, \qquad \mu \in (0, \infty).$$

Since $\mu \to p_k(\mu)$ is strictly monotone, the inverse function $\mu_k(\pi) = \mu[\pi; k] = p_k^{-1}(\pi)$ exists for $qk/n \leq \pi \leq 1$. In the language of testing, consider the two-sided test of $H_0 : \mu = 0$ that rejects at $t_k$. Then $\mu_k(\pi)$ is that alternative $\mu$ at which the test has power $\pi$. In addition

$$\frac{d}{d\pi} p_k^{-1}(\pi) = \frac{1}{p_k'(p_k^{-1}(\pi))} = \frac{1}{p_k'(\mu)}.$$

*The bi-threshold function.* Given indices $\nu < k$, so that $t_\nu > t_k$, consider minimizing $M(\nu; \mu)$ over $\mu$ subject to the constraint that $M(k; \mu)$ stay fixed. Introduce variables $\pi_l = p_k(\mu_l)$; we wish to minimize

$$M(\nu; \mu) = \sum_l p_\nu(\mu_l) = \sum_l p_\nu(p_k^{-1}(\pi_l)) \quad \text{subject to} \quad \sum_l \pi_l = m.$$

Define a bi-threshold function

$$g(\pi) = g_{\nu,k}(\pi) = p_\nu(p_k^{-1}(\pi)), \qquad qk/n \leq \pi \leq 1.$$

Thus, $g_{\nu,k}(\pi)$ gives the power of the test based on the $t_\nu$-threshold at the alternative where the $t_k$-threshold has power $\pi$. As $\nu < k$, $g_{\nu,k}(\pi) \leq \pi$.

LEMMA 6.3. *If $\nu < k$, then $\pi \to g_{\nu,k}(\pi)$ is convex and increasing from $q\nu/n$ to 1 for $\pi \in [qk/n, 1]$.*

PROOF. Setting $\mu = p_k^{-1}(\pi)$, we have

$$g'(\pi) = \frac{p_\nu'(\mu)}{p_k'(\mu)} = \frac{\phi(t_\nu - \mu) - \phi(t_\nu + \mu)}{\phi(t_k - \mu) - \phi(t_k + \mu)} = e^{(t_k^2 - t_\nu^2)/2} \frac{\sinh t_\nu \mu}{\sinh t_k \mu} > 0.$$

To complete the proof, we show that if $t > u$, then the function $G(\mu) = f(t, \mu)/f(u, \mu)$ is increasing for $f(t, \mu) = \sinh t\mu$. First note that

$$G'(\mu) = G(\mu) \int_u^t D_s D_\mu (\log f)(s, \mu)\, ds,$$

and that, on setting $y = 2s\mu$,

$$D_\mu \log \sinh(s\mu) = s\frac{\cosh s\mu}{\sinh s\mu} = \frac{1}{2\mu}\left[y + \frac{2y}{e^y - 1}\right].$$

Finally, $D_y\{y + 2y/(e^y - 1)\}$ has numerator proportional to $(e^y - y)^2 - 1 - y^2 \geq y^4/4 > 0$. $\square$

For weak $\ell_p$, define $\tilde{q}_n$ as in (6.4), while for the nearly-black case, set $\tilde{q}_n = q_n$. Let $a_n$ be positive constants to be specified. Since (6.4) guarantees that $\tilde{q}_n < 1$, define

$$\pi_n = (1 - \tilde{q}_n)a_n.$$

PROPOSITION 6.4. *Assume* (Q) *and* (H). *As before, let* $k_n' = n\eta_n$ (*for* $\ell_0$) *or* $n\eta_n^p\tau_\eta^p$ (*for weak* $\ell_p$). *Define* $\pi_n = a_n(1 - \tilde{q}_n)$ *as above. Then, uniformly in* $\mu$ *for which* $k(\mu) \geq a_n k_n'$, *we have*:

(a) *For* $\nu \leq a_n k_n'$,

$$M(\nu; \mu) \geq k_n' \bar{\Phi}(t_\nu - \mu[\pi_n; a_n k_n']). \tag{6.7}$$

(b) *In particular, for* $\nu = 1$ *and* $a_n \geq b_5(\log n)^{-r}$,

$$M(1; \mu) \geq c(\log n)^{\gamma - r - 1}. \tag{6.8}$$

REMARK. 1. The lower bound of (6.8) is valid for all sparsities $\eta_n^p$ in the range $[n^{-1}\log^\gamma n, n^{-b_3}]$; clearly it is far from sharp for $\eta_n^p$ away from the lower limit. If needed, better bounds for specific cases would follow from (6.12) and (6.13) in the proof below.

2. We shall need only the lower bound for $\nu = 1$, but the methods used below would equally lead to bounds for larger $\nu$, working from the intermediate estimate (6.7).

COROLLARY 6.5. *Let* $k_n = n\eta_n^p\tau_\eta^{-p}$, *and take* $\alpha_n$ *as in assumption* (A); *then uniformly in* $\mu$ *for which* $k(\mu) \geq \alpha_n k_n$,

$$M(1; \mu) \geq c(\log n)^{\gamma - p - 3/2}.$$

PROOF OF PROPOSITION 6.4. For convenience, we abbreviate $a_n k_n'$ by $\kappa$. The bi-threshold function $g = g_{\nu,\kappa}$ is convex (Lemma 6.3), and so

$$M(\nu; \mu) = \sum_{l=1}^{n} g_{\nu,\kappa}(\pi_l) \geq \sum_{l=1}^{k_n'} g(\pi_l) \geq k_n' g(\bar{\pi}_\kappa(\mu)),$$

where $\bar{\pi}_\kappa(\mu) = (1/k_n') \sum_1^{k_n'} \pi_l$ is the true positive rate defined at (6.6). Since $k(\mu) \geq \kappa$, Corollary 6.2 bounds $\bar{\pi}_\kappa(\mu) \geq (1 - \tilde{q}_n)(\kappa/k_n') = \pi_n$ and so from the monotonicity of $g$,

$$g(\bar{\pi}_\kappa(\mu)) \geq g(\pi_n) = p_\nu(\mu_\kappa(\pi_n)) \geq \bar{\Phi}\big(t_\nu - \mu_\kappa(\pi_n)\big).$$

This establishes part (a). For (b), we seek an upper bound for

$$t_\nu - \mu_\kappa(\pi_n) = t_\nu - t_\kappa + t_\kappa - \mu_\kappa, \tag{6.9}$$

where we abbreviate $\mu_\kappa(\pi_n)$ by $\mu_\kappa$ for convenience. First note that

$$\pi_n = p_\kappa(\mu_\kappa) = \Phi(-t_\kappa - \mu_\kappa) + \bar{\Phi}(t_\kappa - \mu_\kappa),$$

which shows that $\bar{\Phi}(t_\kappa - \mu_\kappa) \leq \pi_n \leq 2\bar{\Phi}(t_\kappa - \mu_\kappa)$, from which we get

$$t_\kappa - \mu_\kappa \leq z(\pi_n/2) \leq \sqrt{2\log(2\pi_n^{-1})},$$

using (12.7). Since $\pi_n = (1 - \tilde{q}_n)a_n \geq c_3/(\log n)^{-r}$, we conclude that

$$t_\kappa - \mu_\kappa(\pi_n) \leq \sqrt{2r \log\log n} + c_4. \tag{6.10}$$

From (12.13), we have

$$0 \leq t_\nu - t_\kappa \leq \sqrt{2\log(n/\nu)} - \sqrt{2\log(n/\kappa)} + c(b_1, b_3). \tag{6.11}$$

Since $\kappa = a_n k_n'$ with $a_n \leq 1$, the right-hand side only increases if we replace $\kappa$ by $k_n'$.

At this point, we specialize to the case $\nu = 1$ and set $v_n = \sqrt{2\log n} - \sqrt{2\log n/k_n'}$. Combining (6.9), (6.10) and (6.11), we find that

$$t_\nu - \mu_\kappa(\pi_n) \leq v_n + c(b_1, b_3) + \sqrt{2r \log\log n} + c_4.$$

For $n \geq n(b)$, the last three terms are bounded by $s_n = \sqrt{(2r+1)\log\log n}$. So, from (6.7),

$$M(1; \mu) \geq k_n' \bar{\Phi}(v_n + s_n). \tag{6.12}$$

We may rewrite $k_n'$ in terms of $v_n$, obtaining

$$\log k_n' = v_n \sqrt{2\log n} - v_n^2/2.$$

The bound $\bar{\Phi}(w) \geq \phi(w)/(2w)$ holds for $w > \sqrt{2}$; applying this we conclude

$$k_n' \bar{\Phi}(v_n + s_n) \geq \frac{e^{-s_n^2/2}}{2(v_n + s_n)} \exp\big\{v_n\big(\sqrt{2\log n} - v_n - s_n\big)\big\}. \tag{6.13}$$

Since $e^{-s_n^2/2} = (\log n)^{-r-1/2}$ and $v_n \leq \sqrt{2\log n}$, the first factor is bounded below by $c_0(\log n)^{-r-1}$. To bound the main exponential term, set $g(v) = v(\sqrt{2\log n} -$

$v - s_n)$. We note that $n\eta_n^p \in [\log^\gamma n, n^{1-b_3}]$ and so $\tau_\eta^2 = 2\log \eta_n^{-p} \le 2\log n$ and so $\tau_\eta^p \in [1, 2\log n]$ and so $k_n' \in [\log^\gamma n, (2\log n)n^{1-b_3}]$. For $\ell_0[\eta_n]$, $k_n' \in [\log^\gamma n, n^{1-b_3}]$; we shall see shortly that the difference between the two cases does not matter here.

We now estimate the values of $v$ and $g(v)$ corresponding to these bounds on $k_n'$. At the lower end, $k_n' = \log^\gamma n$; then [using $\sqrt{a} - \sqrt{a-\varepsilon} \ge \varepsilon/(2\sqrt{a}\,)$]

$$v_n \ge \frac{\gamma \log\log n}{\sqrt{2\log n}} =: v_{1n}$$

and one checks that $g(v_{1n}) \ge \gamma \log\log n - 1$ for $n \ge n(b)$. At the upper end, if $k_n' = (2\log n)n^{1-b_3}$, then

$$v_n \le \bigl(1 - \sqrt{b_3}\bigr)\sqrt{2\log n} + c =: v_{2n}$$

and one checks that $g(v_{2n}) = (\sqrt{b_3} - b_3)(2\log n)(1 + o(1)) \ge g(v_{1n})$ for $n$ large.

Since $g(v)$ is a concave quadratic polynomial with maximum between $v_{1n}$ and $v_{2n}$, it follows that for $k_n'$ in the range indicated above,

$$e^{g(v_n)} \ge e^{g(v_{1n})} \ge e^{-1}(\log n)^\gamma.$$

Combined with the bound on the first factor in (6.13), we get

$$k_n' \tilde{\Phi}(v_n + s_n) \ge c_0 e^{-1}(\log n)^{\gamma - r - 1},$$

as required for part (b). $\square$

6.3. *Properties of the mean detection function.* This subsection collects some properties of

$$M(\nu; \mu) = \sum_l \tilde{\Phi}(t_\nu - \mu_l) + \Phi(-t_\nu - \mu_l)$$

as a function of $\nu$, considered as a real variable in $\mathbb{R}_+$. Writing $\dot{M}$, $\ddot{M}$ for partial derivatives w.r.t. $k$, calculus shows that

$$\partial t_\nu / \partial \nu = -q_n/(2n\phi(t_\nu)),$$

$$\dot{M}(\nu; \mu) = (-\partial t_\nu/\partial \nu) \sum_l \phi(t_\nu - \mu_l) + \phi(t_\nu + \mu_l) \tag{6.14}$$

$$= (q_n/n) \sum_l e^{-\mu_l^2/2} \cosh(t_\nu \mu_l) > 0, \tag{6.15}$$

$$\ddot{M}(\nu; \mu) = -q_n^2/(2n^2\phi(t_\nu)) \sum_l \mu_l e^{-\mu_l^2/2} \sinh(t_\nu \mu_l) \le 0, \tag{6.16}$$

with strict inequality unless $\mu \equiv 0$. Finally, since $M(0; \mu) = 0$, there exists $\tilde{\nu} \in [0, \nu]$ such that the threshold exceedance function $\nu^{-1} M(\nu; \mu) = \nu^{-1}(M(\nu; \mu) -$

$M(0;\mu)) = \dot{M}(\tilde{\nu},\mu)$, and hence, for each $\mu$ the exceedance function is decreasing in $\nu$:

$$\frac{\partial}{\partial\nu}\left(\frac{M(\nu;\mu)}{\nu}\right) = \frac{1}{\nu}[\dot{M}(\nu;\mu) - \dot{M}(\tilde{\nu},\mu)] \le 0. \tag{6.17}$$

Let us focus now on $\ell_0[\eta_n]$. In this case

$$M(\nu;\mu) = \sum_1^{k_n}[\tilde{\Phi}(t_\nu - \mu_l) + \Phi(-t_\nu - \mu_l)] + (1-\eta_n)q_n\nu,$$

and so, using (6.14) and (12.9),

$$\dot{M}(\nu;\mu) \le \frac{2\phi(0)k_n}{\nu t_\nu} + (1-\eta_n)q_n.$$

In particular, if $\nu = ak_n$, then

$$\dot{M}(ak_n;\mu) \le \frac{2\phi(0)}{at[ak_n]} + q_n.$$

Finally, if $a = \alpha_n = 1/(b_4\tau_\eta)$, then (12.15) shows that for $\eta_n$ sufficiently small,

$$\frac{2\phi(0)}{at[ak_n]} = b_4\sqrt{\frac{2}{\pi}}\frac{\tau_\eta}{t[\alpha_n k_n]} \le b_4. \tag{6.18}$$

As a result, uniformly in $\ell_0[\eta_n]$,

$$\dot{M}(\alpha_n k_n;\mu) \le b_4 + q_n < q', \tag{6.19}$$

by the definition of $b_4$; recall assumptions (A) and (Q) of Section 4.1.

6.4. *Weak $\ell_p$: bounds for the detection function.* For weak $\ell_p$, we do not have such a simple bound on $\dot{M}$ as (6.19). From the preceding calculations, we know that $\nu \to \dot{M}(\nu;\mu)$ is positive and decreasing. We will need now some sharper estimates, uniform over $m_p[\eta_n]$ (and $\ell_0[\eta_n]$) in the scaling $\nu = ak_n$, with $a$ regarded as variable. This will lead to bounds on the solution of $M(ak_n;\mu) = ak_n$ and hence to bounds on $k(\mu)$ (cf. Corollary 6.8).

The two key phenomena are:

(a) If $a_1$ is fixed, then for $\nu$ in intervals $[a_1k_n, a_1^{-1}k_n]$, the slope of $M$ is, for large $n$, essentially constant and equal to $q_n$. This reflects exclusively the effect of false detections.

(b) For small $a$ ($\sim 1/\tau_\eta$, say), the order of magnitude of $\dot{M}(ak_n;\mu)$ can be as large as $1/(at[ak_n])$. This reflects essentially the effect of true detections.

Since $\mu \to \dot{M}(k;\mu)$ is an even function of each $\mu_l$, we may assume without loss of generality that $\mu_l \ge 0$ for each $l$.

To bound $\dot{M}$, divide the range of summation into three regions, defined by the indices

$$k_n = n\eta_n^p \tau_\eta^{-p}, \qquad \bar{\mu}_{k_n} = \tau_\eta,$$
$$k_n' = n\eta_n^p \tau_\eta^{p}, \qquad \bar{\mu}_{k_n'} = \tau_\eta^{-1}.$$

Thus, we write

$$\dot{M} = \dot{M}_{\text{pos}} + \dot{M}_{\text{trn}} + \dot{M}_{\text{neg}},$$

where the sum in $\dot{M}_{\text{pos}}$ extends over the range $[1, k_n]$ of true "positives." The sum in $\dot{M}_{\text{trn}}$ ranges over $(k_n, k_n']$ and is "transitional," while the sum for $\dot{M}_{\text{neg}}$ ranges over $(k_n', n]$ and corresponds to means that are essentially true "negatives."

A rough statement of the results to follow is that for $a$ in the range $[\gamma\tau_\eta^{-1}, 1]$,

$$\sup_{m_p[\eta_n]} \dot{M}_{\text{pos}}(ak_n; \mu) \asymp \frac{1}{at[ak_n]},$$
$$\sup_{m_p[\eta_n]} \dot{M}_{\text{trn}}(ak_n; \mu) = O\left(\frac{1}{at^2[ak_n]}\right),$$
$$\dot{M}_{\text{neg}}(ak_n; \mu) \sim q_n.$$

Combining these will establish:

PROPOSITION 6.6. *Assume* (Q), (H) *and* (A). *For $n > n(b)$, $t_\nu \geq c_0, a \geq 1$ and all $\mu \in m_p[\eta_n]$,*

$$\begin{aligned} q_n[1-\varepsilon_n^p] &\leq \dot{M}(ak_n; \mu) \\ &\leq q_n[1 + c(b)\delta_p(\varepsilon_n)] + \frac{2\phi(0)}{at[ak_n]}\left[1 + \frac{c_0}{t[ak_n]}\right]. \end{aligned} \tag{6.20}$$

*If, in addition, $a \geq \gamma\tau_\eta^{-1}$, then for $\eta^p \leq \eta(\gamma, p, b_1, b_3)$ sufficiently small,*

$$\sup_{m_p[\eta_n]} \dot{M}(ak_n; \mu) \geq q_n[1-\varepsilon_n^p] + \frac{c_0}{at[ak_n]}. \tag{6.21}$$

The proof consists in building estimates for $\dot{M}$ in the positive, transition and negative zones. It will also be convenient, for Corollary 6.8 below, to obtain estimates at the same time for the corresponding components of the detection function $M$ itself.

6.4.1. *Positive zone.* $M_{\mathrm{pos}}(\nu;\mu)=\sum_{l=1}^{k_n} p_\nu(\mu_l)$ is, for $\mu=\bar\mu_l$, approximately constant on the interval $[a_1k_n, a_1^{-1}k_n]$: for $a_1\le a\le a_1^{-1}$,

$$M_{\mathrm{pos}}(ak_n;\bar\mu)\in[1-\varepsilon_{1n},1]k_n. \tag{6.22}$$

PROOF OF PROPOSITION 6.6. The upper bound follows from $p_\nu(\mu_l)\le 1$. For the lower bound, since $\tilde\Phi(t_\nu-\bar\mu_l)$ is decreasing in $l$ and increasing in $\nu$, we have for any $l_1\le k_n$,

$$M_{\mathrm{pos}}(\nu;\bar\mu)\ge\sum_1^{k_n}\tilde\Phi(t_\nu-\bar\mu_l)\ge l_1\tilde\Phi(t_\nu-\bar\mu_l)\ge l_1\tilde\Phi\big(t[\delta k_n]-\bar\mu_{l_1}\big).$$

Choose $\gamma_n=\tau_\eta^{-1}$ and define $l_1$ by the equation

$$\bar\mu_{l_1}=t[\delta k_n]+z(\gamma_n).$$

From (12.15) it is clear that $\bar\mu_{l_1}\ge\tau_\eta$, so that $l_1\le k_n$. Hence

$$M_{\mathrm{pos}}(\nu;\bar\mu)\ge l_1\tilde\Phi(-z(\gamma_n))=l_1(1-\gamma_n). \tag{6.23}$$

We have from (12.16) that $\bar\mu_l=\tau_\eta+\sqrt{2\log\tau_\eta}+c(b_1,b_3)$ for $\eta$ small, and hence

$$\frac{l_1}{k_n}=\left[\frac{\tau_\eta}{t[\delta k_n]+z(\gamma_n)}\right]^p\ge\left[1+\frac{\sqrt{2\log\tau_\eta}+c}{\tau_\eta}\right]^{-p}\ge 1-c\tau_\eta^{-1}\sqrt{2\log\tau_\eta}.$$

Since $\gamma_n=\tau_\eta^{-1}$, the last two displays imply (6.22) with $\varepsilon_{1n}=c\tau_\eta^{-1}\sqrt{2\log\tau_\eta}$.

*Bound for $\dot M_{\mathrm{pos}}$.* Since $M_{\mathrm{pos}}(\nu;\mu)=\sum_1^{k_n}\tilde\Phi(t_\nu-\mu_l)+\Phi(-t_\nu-\mu_l)$, we have

$$\dot M_{\mathrm{pos}}(\nu;\mu)=(-\partial t_\nu/\partial\nu)\sum_1^{k_n}\phi(t_\nu-\mu_l)+\phi(t_\nu+\mu_l). \tag{6.24}$$

From (12.9), we obtain $\dot M_{\mathrm{pos}}(\nu;\mu)\le 2k_n\phi(0)/(\nu t_\nu)$. Hence, for $\nu=ak_n$ with $t_\nu\ge 1$ and all $\mu$,

$$\dot M_{\mathrm{pos}}(ak_n;\mu)\le\frac{2\phi(0)}{at[ak_n]}. \tag{6.25}$$

Turning to the lower bound, we note that $\bar\mu_l\ge t_\nu$ if and only if $l\le n\eta_n^p t_\nu^{-p}=:l_\nu$. By setting all $\mu_l=t_\nu$ for $l\le l_\nu$, we find from (6.24) and (12.9) that

$$\sup_{m_p[\eta_n]}\dot M_{\mathrm{pos}}(\nu;\mu)\ge\frac{\phi(0)}{2}\frac{k_n\wedge l_\nu}{\nu t_\nu}.$$

(This also holds for $\ell_0[\eta_n]$.) If $\nu=ak_n$, then

$$\frac{k_n\wedge l_\nu}{\nu t_\nu}=\frac{1}{at[ak_n]}\left[\left(\frac{\tau_\eta}{t[ak_n]}\right)^p\wedge 1\right].$$

If $a \geq \gamma\tau_\eta^{-1}$, then for $\eta$ sufficiently small, (12.17) says $\tau_\eta/t[ak_n] \geq \frac{1}{2}$. Combining the last remarks, we conclude that for $\nu = ak_n$ and $a \geq \gamma\tau_\eta^{-1}$, then for $\eta$ sufficiently small,

$$\sup_{m_p[\eta_n]} \dot{M}_{\text{pos}}(ak_n; \mu) \geq \frac{c_0}{at[ak_n]}. \tag{6.26}$$

Here $c_0$ denotes an absolute constant.

6.4.2. *Transition zone.* For $a \leq a_1^{-1}$ and $\eta \leq \eta_0$ sufficiently small, we have, uniformly in $m_p[\eta_n]$:

$$0 \leq M_{\text{trn}}(ak_n; \mu) \leq c_0\tau_\eta^{-1}k_n, \tag{6.27}$$

$$\dot{M}_{\text{trn}}(ak_n; \mu) \leq \frac{c_0}{at^2[ak_n]}. \tag{6.28}$$

To bound $\dot{M}_{\text{trn}}(k; \mu)$, introduce

$$h(\mu, t) = e^{t\mu - \mu^2/2},$$

which increases from 1 to a global maximum of $e^{t^2/2} = \phi(0)/\phi(t)$ as $\mu$ grows from 0 to $t$. We have from (6.15) that $\dot{M}(k; \mu) \leq 2(q_n/n)\sum h(\mu_l, t_k)$. The arguments for the two bounds run in parallel. We have

$$M_{\text{trn}}(\nu; \mu) \leq \sum_{k_n}^{k_n'} H_1(\mu_l) \quad \text{and} \quad \dot{M}_{\text{trn}}(\nu; \mu) \leq (q_n/n)\sum_{k_n}^{k_n'} H_2(\mu_l),$$

where $H_1(\mu) = 2\tilde{\Phi}(t_\nu - \mu)$ and $H_2(\mu) = h(\mu \wedge t_\nu, t_\nu)$ are both increasing functions of $\mu \geq 0$. By integral approximation,

$$\sum_{l=k_n}^{k_n'} H(\mu_l) \leq \sum_{k_n}^{k_n'} H_1(\bar{\mu}_l) \leq n\eta_n^p \int_{\tau_\eta^{-1}}^{\tau_\eta} H(u)u^{-p-1}\,du.$$

For $a \leq a_1^{-1}$, we have $t = t_\nu \geq \tau_\eta - 3/2$ for $\eta$ sufficiently small by (12.15), while (12.17) shows that $\tau_\eta^{-1} \geq 1/(2t_\nu)$. Let $\bar{H} = \sup H$; we have

$$\int_{\tau_\eta^{-1}}^{\tau_\eta} H(u)u^{-p-1}\,du \leq \int_{1/(2t)}^{t} H(u)u^{-p-1}\,du + (3/2)\bar{H}t^{-p-1} \leq c_0\bar{H}t^{-p-1}$$

after using Lemma 6.7 below. For $M_{\text{trn}}$, $\bar{H} = 2$ and we have from the previous displays

$$M_{\text{trn}}(ak_n; \mu) \leq c_0 n\eta_n t[ak_n]^{-p-1} \leq c_0\tau_\eta^{-1}k_n,$$

since $t[ak_n]^{-1} \leq 2\tau_\eta^{-1}$.

For $\dot{M}_{\text{trn}}$, $\bar{H} = h(t_\nu, t_\nu) = \phi(0)/\phi(t_\nu)$, and since $\phi(t_\nu) \geq \frac{1}{2} t_\nu \tilde{\Phi}(t_\nu)$,

$$\dot{M}_{\text{trn}}(ak_n; \mu) \leq q_n \eta_n^p \frac{c_0 h(t_\nu, t_\nu)}{t_\nu^{p+1}} \leq \frac{c_0 q_n \eta_n^p}{t_\nu^{p+2} \tilde{\Phi}(t_\nu)} \leq c_0 \left(\frac{\tau_\eta}{t[ak_n]}\right)^p \frac{1}{at^2[ak_n]},$$

and so (6.28) follows from (12.15).

LEMMA 6.7. *For $p \leq 2$ and $t \geq 2$, and for $h(u,t)$ given by either $\tilde{\Phi}(t-u)$ or $\phi(t-u)/\phi(t)$, there is an absolute constant $c_0$ such that*

$$\int_{1/(2t)}^{t} \frac{h(u,t)\,du}{u^{p+1}} \leq \frac{c_0 h(t,t)}{t^{p+1}}.$$

PROOF. Writing $v$ for $t-u$, we find that in the two cases $h(u,t)/h(t,t)$ equals $2\tilde{\Phi}(v)$ or $e^{-v^2/2}$, respectively. By (12.3), $2\tilde{\Phi}(v) \leq e^{-v^2/2}$ for $v \geq 0$, so in either case the integral in question is bounded by

$$\frac{h(t,t)}{t^{p+1}} \int_0^{t-1/(2t)} \exp\{-v^2/2 + (p+1)g(v)\}\,dv,$$

where the convex function $g(v) = \log t - \log(t-v)$ is bounded for $0 < v < t-1$ by $4v(\log t)/t$. Completing the square in the exponent gives an integrand smaller than $\sqrt{2\pi}$ times a unit-variance Gaussian density centered at $\mu(p,t) = 4(p+1)(\log t)/t$. Since $\mu(p,t) \leq c_0$ for $p \leq 2$ and $t \geq 2$, the previous integral is bounded by $\sqrt{2\pi}\exp\{\frac{1}{2}\mu^2(p,t)\} \leq c_0$. $\square$

6.4.3. *Negative zone.* Under conditions described immediately below,

$$[1-\varepsilon_n^p]q_n\nu \leq M_{\text{neg}}(\nu;\mu) \leq [1 + c\delta_p(\varepsilon_n)]q_n\nu, \tag{6.29}$$

$$[1-\varepsilon_n^p]q_n \leq \dot{M}_{\text{neg}}(\nu;\mu) \leq [1 + c\delta_p(\varepsilon_n)]q_n. \tag{6.30}$$

The lower bound in (6.29) holds for all $n, \nu, \mu$. All other bounds require $\mu \in m_p[\eta_n]$ and $n \geq n(b)$. The upper bounds further require $\nu$ such that $t_\nu \geq c(b)$.

*Lower bounds.* For $M_{\text{neg}}(\nu;\mu) = \sum_{l \geq k_n'} p_\nu(\mu_l)$ this is simple because $p_\nu(\mu_l) \geq p_\nu(0) = q_n\nu/n$, so that $M_{\text{neg}}(\nu;\mu) \geq (n-k_n')q_n\nu/n = [1-\varepsilon_n^p]q_n\nu$.

For $\dot{M}_{\text{neg}}(\nu;\mu)$, we set $f(\mu,t) = e^{-\mu^2/2}\cosh(t\mu)$ and check that for given $c_1$, $\mu \to f(\mu,t)$ is increasing for $t\mu \leq c_1$ and $t \geq \sqrt{c_1}$. For $l \geq k_n'$ we have

$$\mu_l \leq \bar{\mu}_l \leq \bar{\mu}_{k_n'} = \tau_\eta^{-1} \leq c_1 t_1^{-1} \leq c_1 t_\nu^{-1}, \tag{6.31}$$

by (12.14), for $c_1 = c_1(b)$. Consequently $f(\mu_l, t_\nu) \geq f(0, t_\nu) = 1$ and so

$$\dot{M}_{\text{neg}}(\nu;\mu) = (q_n/n)\sum_{l \geq k_n'} f(\mu_l, t_\nu) \geq (q_n/n)(n-k_n') = [1-\varepsilon_n^p]q_n.$$

*Upper bounds.* The arguments run in parallel: we have

$$M_{\mathrm{neg}}(\nu;\mu) \le \sum_{l\ge k_n'} H_1(\mu_l) \quad \text{and} \quad \dot{M}_{\mathrm{neg}}(\nu;\mu) \le (q_n/n)\sum_{l\ge k_n'} H_2(\mu_l), \tag{6.32}$$

where $H_1(\mu) = 2\tilde{\Phi}(t_\nu - \mu)$ and $H_2(\mu) = h(\mu, t_\nu)$ are both increasing and convex functions of $\mu \in [0,1]$, at least when $t > 2$. Using (6.31) along with $t = t_\nu$, this convexity implies

$$H(\mu_l) \le (1 - t\mu_l/c_1)H(0) + (t\mu_l/c_1)H(c_1/t)$$

and hence, since $t \le c_b\tau_\eta$ from (12.14),

$$n^{-1}\sum_{l\ge k_n'} H(\mu_l) \le H(0) + c_b c_1^{-1} H(c_1/t)\tau_\eta n^{-1}\sum_{l\ge k_n'} \bar{\mu}_l. \tag{6.33}$$

By an integral approximation, since $\varepsilon_n = \eta_n\tau_\eta$ and $k_n'/n = \varepsilon_n^p$, we find

$$\tau_\eta n^{-1}\sum_{l=k_n'}^{n} \bar{\mu}_l \le \tau_\eta\eta_n \int_{k_n'/n}^{1} x^{-1/p}\,dx = p\varepsilon_n \int_{\varepsilon_n}^{1} s^{p-2}\,ds = \delta_p(\varepsilon_n). \tag{6.34}$$

To apply these bounds to $M_{\mathrm{neg}}(\nu,\mu)$, we note that $H_1(0) = q_n\nu/n$ while from (12.4), for $t \ge \sqrt{2c_1}$,

$$H(c_1 t^{-1}) \le 8e^{c_1}\tilde{\Phi}(t) = 4e^{c_1}(q_n\nu/n).$$

From (6.32), (6.33) and (6.34), we obtain (6.29).

Turning to $\dot{M}_{\mathrm{neg}}(\nu;\mu)$, we note that $H_2(0) = 1$ and $H_2(c_1/t) = \exp\{c_1 - c_1^2/(2t^2)\} \le e^{c_1}$ and so the same bounds combine to yield (6.30).

6.4.4. *Conclusion.* The upper bound in (6.20) follows by combining those in (6.25), (6.28) and (6.30). The lower bound (6.21) follows by combining (6.26) and (6.30).

COROLLARY 6.8. *Let $d_n = 2c_0\tau_\eta^{-1}$ [where $c_0$ is the constant in (6.27)]. Uniformly in $m_p[\eta_n]$,*

$$k(\mu) \le (1 - q_n - d_n)^{-1}k_n. \tag{6.35}$$

PROOF. Let $s = (1 - q_n - d_n)^{-1}$. Combining the bounds on $M_{\mathrm{pos}}$, $M_{\mathrm{trn}}$ and $M_{\mathrm{neg}}$ in (6.22), (6.27) and (6.29), we find for $n > n(b)$ and $\eta$ sufficiently small that

$$M(sk_n;\bar{\mu}) \le [1 + c_0\tau_\eta^{-1} + q_n s(1 + c_b\delta_p(\varepsilon_n))]k_n \le [1 + q_n s + r_n]k_n,$$

where, since $s \le (1 - q')^{-1}$,

$$r_n = c_0\tau_\eta^{-1} + c\delta_p(\varepsilon_n), \qquad c = c(b, q').$$

We have

$$M(sk_n;\bar{\mu})/(sk_n) \le 1 - q_n - d_n + q_n + r_n = 1 - d_n + r_n.$$

Since $\delta_p(\varepsilon_n) = o(\tau_\eta^{-1})$ (from the assumptions on $\eta_n$), clearly $r_n - d_n < 0$ for $n > n(b)$; for such $n$, $M(sk_n;\bar{\mu}) < sk_n$, and so $k(\bar{\mu}) < sk_n$, as required. $\square$

We draw a consequence for later use. Define

$$\kappa_n = \begin{cases} [\alpha_n + (1-q_n)^{-1}]k_n, & \text{for } \ell_0[\eta_n], \\ [\alpha_n + (1-q_n-d_n)^{-1}]k_n, & \text{for } m_p[\eta_n]. \end{cases} \tag{6.36}$$

Recall now the notational assumption (A). Clearly, for large $n$,

$$\kappa_n \sim (1-q_n)^{-1}k_n \quad \text{and} \quad \kappa_n \le k_n/q''. \tag{6.37}$$

From the remark after (5.3) (case $\Theta_n = \ell_0[\eta_n]$), and from Corollary (6.8) (case $\Theta_n = m_p[\eta_n]$),

$$\sup_{\mu\in\Theta_n} k_+(\mu) \le \kappa_n.$$

$\square$

**7. Large deviations bounds for $[\hat{k}_G, \hat{k}_F]$.** We now develop exponential bounds on the FDR interval $[\hat{k}_G, \hat{k}_F]$ that lead to a proof of Proposition 5.1. "Switching" inequalities allow the boundary-crossing definitions of $\hat{k}_G, \hat{k}_F$ to be expressed in terms of sums of independent Bernoulli variables for which large deviations inequalities in a "small numbers" regime can be applied.

7.1. *Switching inequalities.* We will write $Y_l$ for the absolute ordered values $|y|_{(l)}$. Let $1 \le \hat{k}_G \le \hat{k}_F \le n$ be, respectively, the smallest and largest local minima of $k \to S_k = \sum_{l=k+1}^n Y_l^2 + \sum_{l=1}^k t_l^2$ for $0 \le k \le n$. The possibility of ties in the sequence $\{S_k\}$ complicates the exact description of local minima. Since ties occur with probability zero, we will for convenience ignore this possibility in the arguments to follow, lazily omitting explicit mention of "with probability 1."

Define the exceedance numbers

$$N(t_k) = \#\{i : |y_i| > t_k\}, \qquad N_+(t_k) = \#\{i : |y_i| \ge t_k\}.$$

Clearly $N(t_k)$ and $N_+(t_k)$ have the same distributions. We now have

$$\hat{k}_F = \max\{l : Y_l > t_l\} = \max\{l : N(t_l) \ge l\}, \tag{7.1}$$

$$\hat{k}_G + 1 = \min\{l : Y_l < t_l\} = \min\{l : N_+(t_l) < l\}. \tag{7.2}$$

[We set $\hat{k}_F = 0$ or $\hat{k}_G = n$ if no such indices $l$ exist.] To verify the left-hand side inequalities, note that $S_k - S_{k-1} = t_k^2 - Y_k^2$, so that

$$S_k \ge S_{k-1} \iff Y_k \le t_k.$$

The largest local minimum of $S_k$ occurs at $k = \hat{k}_F$ exactly when $S_k < S_{k-1}$ but $S_l \geq S_{l-1}$ for all $l > k$. In other words $Y_k > t_k$ but $Y_l \leq t_l$ for all larger $l$, which is precisely (7.1). Similarly, the smallest local minimum of $S_k$ occurs at $k = \hat{k}_G$ exactly when $S_{k+1} > S_k$ but $S_l \leq S_{l-1}$ for all $l \leq k$, and this leads immediately to (7.2).

For the right-hand side inequalities, we simply note that

$$N(t_k) \geq k \qquad \text{iff } Y_k > t_k \quad \text{and} \quad N_+(t_k) < k \qquad \text{iff } Y_k < t_k.$$

7.2. *Exponential bounds.* First, recall Bennett's exponential inequality (e.g., [32], page 192) in the form which states that for independent, zero-mean random variables $X_1, \dots, X_n$ with $|X_i| \leq K$ and $V = \sum \mathrm{Var}(X_i)$,

$$P\{X_1 + \cdots + X_n \geq \eta\} \leq \exp\left\{-\frac{\eta^2}{2V} B\left(\frac{K\eta}{V}\right)\right\},$$

where $B(\lambda) = (2/\lambda^2)[(1+\lambda)\log(1+\lambda) - \lambda]$ for $\lambda > 0$ is decreasing in $\lambda$. The extended version [3] gives the following consequence, useful for settings of Poisson approximation.

LEMMA 7.1. *Suppose that $Y_l$, $l = 1, \dots, n$, are independent $0/1$ variables with $P(Y_l = 1) = p_l$. Let $N = \sum_1^n Y_l$ and $M = EN = \sum_1^n p_l$. Then*

$$P\{N \leq k\} \leq \exp\{-\tfrac{1}{4} M h(k/M)\}, \qquad \text{if } k < M, \tag{7.3}$$

$$P\{N \geq k\} \leq \exp\{-\tfrac{1}{4} M h(k/M)\}, \qquad \text{if } k > M, \tag{7.4}$$

*where $h(x) = \min\{|x-1|, |x-1|^2\}$.*

7.3. *Bounds on $k/M_k$.* The Lipschitz properties of $k \to \dot{M}(k; \mu)$ established in Section 6 are now applied to derive bounds on the ratios $k/M_k$ appearing in the exponential bounds (7.3)–(7.4). In the following, we use $b$ to denote the vector of constants $b = (b_1, \dots, b_4, q')$, and the phrase $n > n_0(b)$ to indicate that a statement holds for $n$ sufficiently large, depending on the constants $b$.

PROPOSITION 7.2. *Assume hypotheses* (Q), (H) *and* (A). *If $\alpha_n k_n \leq k_1$, then for $n > n(b)$, uniformly in $\mu \in \ell_0[\eta_n]$ and $m_p[\eta_n]$,*

$$M(k_1 + \alpha_n k_n; \mu) - M(k_1; \mu) \leq q' \alpha_n k_n. \tag{7.5}$$

(a) *If $k(\mu) \leq k_1 \leq (1-q')^{-1} k_n$ and $k_2 = k_1 + \alpha_n k_n$, then*

$$\frac{k_2}{M_{k_2}} - 1 \geq (1-q')^3 \alpha_n =: \alpha_n. \tag{7.6}$$

(b) *There is* $\zeta > 0$ *so that, if* $2\alpha_n k_n \le k(\mu) \le (1-q')^{-1}k_n$ *and* $k_1 = k(\mu) - \alpha_n k_n$, *then*

$$1 - \frac{k_1}{M_{k_1}} \ge (1-q')^2\alpha_n \ge \zeta\alpha_n. \tag{7.7}$$

PROOF. Formulas (6.15)–(6.16) show that $\nu \to \dot{M}(\nu;\mu)$ is positive and decreasing and so the left-hand side of (7.5) is positive and bounded above by $\dot{M}(\alpha_n k_n;\mu)$. For $\mu \in \ell_0[\eta_n]$, (6.19) shows that for $n > n(b)$, $\dot{M}(\alpha_n k_n;\mu) \le q'$ for all $\mu$ in $\ell_0[\eta_n]$. That the same bound holds uniformly over $m_p[\eta_n]$ also is a consequence of (6.20) and (6.18).

(a) To prove (7.6), note that the assumption $k(\mu) \le k_1$ entails $M_{k_1} \le k_1$, so from (7.5),

$$M_{k_2} = M_{k_1} + M_{k_2} - M_{k_1} \le k_1 + q'\alpha_n k_n.$$

Since $k_1 \le k_n/(1-q')$ and $q' < 1$, we have

$$\frac{M_{k_2}}{k_2} \le \frac{k_1 + q'\alpha_n k_n}{k_1 + \alpha_n k_n} \le \frac{1 + q'(1-q')\alpha_n}{1 + (1-q')\alpha_n}.$$

Thus, since $\alpha_n = O(1/\sqrt{\log n})$, for $n > n(b)$, we find

$$\frac{k_2}{M_{k_2}} - 1 \ge \frac{(1-q')^2\alpha_n}{1 + q'(1-q')\alpha_n} \ge (1-q')^3\alpha_n.$$

(b) The assumption that $k(\mu) \ge 2\alpha_n k_n$ yields $k_1 \ge \alpha_n k_n$, and so the Lipschitz bound (7.5) implies $M_{k(\mu)} - M_{k_1} \le q'\alpha_n k_n$. Hence, since $k(\mu) \le k_n/(1-q')$,

$$\frac{M_{k_1}}{k_1} \ge \frac{k(\mu) - q'\alpha_n k_n}{k(\mu) - \alpha_n k_n} \ge \frac{1 - q'(1-q')\alpha_n}{1 - (1-q')\alpha_n},$$

which leads to (7.7) by simple rewriting. □

PROOF OF PROPOSITION 5.1. 1. Let $k_1 = k(\mu) \vee \alpha_n k_n$ and $k_2 = k_1 + \alpha_n k_n$. From (7.1),

$$\{\hat{k}_F \ge k_2\} \subset \bigcup_{l \ge k_2} \{N(t_l) \ge l\}. \tag{7.8}$$

For $l > k_2 > k(\mu)$, we necessarily have $E_\mu N(t_l) = M(l;\mu) < l$, and so from Lemma 7.1

$$P_\mu\{N(t_l) \ge l\} \le \exp\{-\tfrac{1}{4} M_l h(l/M_l)\}, \qquad M_l = M(l;\mu). \tag{7.9}$$

For $x \ge 1$, the function $h(x)$ is increasing, and for $l > k_2$, $l \to l/M_l$ is increasing and so $h(l/M_l) \ge h(k_2/M_{k_2})$. Now $k_1$ and $k_2$ satisfy the assumptions of Proposition 7.2(a) and so from (7.6), $h(k_2/M_{k_2}) \ge \zeta^2\alpha_n^2$. Since $l \to M_l$ is increasing, we have from Proposition 6.4 that

$$M_l \ge M(1;\mu) \ge c(\log n)^{\gamma - 3/2}. \tag{7.10}$$

Combining (7.8), (7.9) and (7.10), we find

$$\begin{aligned} P_\mu\{\hat{k}_F > k_2\} &\le \sum_{l>k_2} \exp\{-\tfrac{1}{4} M_1 \zeta^2 \alpha_n^2\} \\ &\le n \exp\{-c\alpha_n^2 (\log n)^{\gamma-3/2}\} \\ &\le n \exp\{-c'(\log n)^{\gamma-5/2}\}, \end{aligned}$$

for $c'$ depending on $q'$ and $b_4$.

2. Now assume that $k(\mu) \ge 2\alpha_n k_n$; we establish a high probability lower bound for $\hat{k}_G$. Let $k_1 = k(\mu) - \alpha_n k_n$; from (7.2)

$$\{\hat{k}_G < k_1\} = \{\hat{k}_G + 1 \le k_1\} \subset \bigcup_{l \le k_1} \{N_+(t_l) < l\}.$$

For $l \le k_1 < k(\mu)$, we necessarily have $M_l > l$ and so

$$P\{N_+(t_l) < l\} = P\{N(t_l) < l\} \le \exp\{-\tfrac{1}{4} M_l h(l/M_l)\}.$$

Since $l \to l/M_l \le 1$ is increasing, and since $k_1$ and $k(\mu)$ satisfy the assumptions of Proposition 7.2(b), we obtain from (7.7) that

$$h(l/M_l) \ge \left(1 - \frac{k_1}{M_{k_1}}\right)^2 \ge \zeta^2 \alpha_n^2.$$

In addition $l \to M_l$ is increasing, and so $M_l \ge M_1$. Since $k(\mu) \ge 2\alpha_n k_n \ge \alpha_n k_n$, we have from Proposition 6.4 that (7.10) holds here also. Hence

$$P_\mu\{\hat{k}_G < k_1\} \le k_1 \exp\{-\tfrac{1}{4} M_1 \zeta^2 \alpha_n^2\} \le n \exp\{-c'(\log n)^{\gamma-5/2}\},$$

in the same way as for (7.11). $\square$

**8. Lemmas on thresholding.** This section collects some preparatory results on hard (and in some cases soft) thresholding with both fixed and data-dependent thresholds. These are useful for the analysis and comparison of the various FDR and penalized rules, and are perhaps of some independent utility.

8.1. *Fixed thresholds.* First, we give an elementary decomposition of the $\ell_r$ risk of hard thresholding.

LEMMA 8.1. *Suppose that $x \sim N(\mu, 1)$ and that $\eta_H(x,t) = xI\{|x| \ge t\}$. Then*

$$\begin{aligned} \rho_H(t,\mu) &= E|\eta_H(x,t) - \mu|^r \\ &= \int_{-t}^{t} |\mu|^r \phi(x-\mu)\,dx + \int_{|x|>t} |x-\mu|^r \phi(x-\mu)\,dx \qquad (8.1) \\ &= D(\mu,t) + E(\mu,t), \qquad (8.2) \end{aligned}$$

*where*

$$D(\mu, t) = |\mu|^r [\Phi(t-\mu) - \Phi(-t-\mu)], \tag{8.3}$$

$$\begin{aligned} E(\mu, t) = {} & |t-\mu|^{r-1}\phi(t-\mu) + |t+\mu|^{r-1}\phi(t+\mu) \\ & + \varepsilon(t-\mu) + \varepsilon(t+\mu), \end{aligned} \tag{8.4}$$

$$|\varepsilon(v)| = |r-1| \int_v^\infty z^{r-2}\phi(z)\,dz \le v^{r-3}\phi(v), \qquad v > 0,\ 0 < r \le 2. \tag{8.5}$$

*We note that for* $0 \le r \le 2$

$$\rho_H(t, 0) = 2\int_t^\infty z^r \phi(z)\,dz = 2t^r\tilde{\Phi}(t)(1 + \theta t^{-2}), \qquad 0 \le \theta \le r, \tag{8.6}$$

*and that* $E(\mu, t)$:

(i) *is globally bounded*: $0 \le E(\mu, t) \le c_r = \int |z|^r \phi(z)\,dz$,
(ii) *is increasing in* $\mu$, *at least for* $0 \le \mu \le t - \sqrt{2}$,
(iii) *satisfies* $E(1, t) \le c_0 t^r \tilde{\Phi}(t-1)$ *for* $t > 1$.

*A consequence of* (8.6) *is that for some* $|\theta_2| \le 1$,

$$\sum_{l=1}^{k} t_l^k = k t_k^r (1 + \theta t_k^{-2}) + \theta_2 t_1^r = k t_k^r \bigl(1 + o(1)\bigr), \tag{8.7}$$

*so long as* $k \to \infty$ *and* $k/n \to 0$.

This is proven in the extended version on which this article is based [3]; the same is true for the next lemma, which concerns covariance between the data and hard thresholding. This helps analyze $\ell_2$ loss; for the $\ell_r$ analog, see Section 11.3.

LEMMA 8.2. *Let* $x \sim N(\mu, 1)$. $\xi(t, \mu) = E_\mu(x - \mu)[\eta_H(x, t) - \mu]$ *has the properties*:

(i)

$$\xi(t, \mu) = t[\phi(t-\mu) + \phi(t+\mu)] + \tilde{\Phi}(t-\mu) + \Phi(-t-\mu), \tag{8.8}$$

(ii)

$$\xi(t, \mu) \le \begin{cases} 2, & \text{for } |\mu| \le t - \sqrt{2\log t}, \\ t+1, & \text{for all } \mu, \end{cases} \tag{8.9}$$

(iii)

$$\mu \to \xi(t, \mu) \text{ is symmetric about } 0, \text{ increasing for } 0 \le \mu \le t, \tag{8.10}$$

$$\text{and convex for } 0 \le \mu \le t/3 \text{ if } t \ge 3, \tag{8.11}$$

(iv)

$$\sup_{|\mu| \le t/3} |\xi_{\mu\mu}(t, \mu)| \le c_0. \tag{8.12}$$

8.2. *Data-dependent thresholds.*

LEMMA 8.3. *Let $x = \mu + z \sim N(\mu, 1)$ and $\eta(x, \hat{t})$ denote soft or hard thresholding at $\hat{t}$. For $r > 0$,*

$$|\eta(x, \hat{t}) - \mu|^r \le 2^{(r-1)_+}(|z|^r + |\hat{t}|^r). \tag{8.13}$$

PROOF. Check cases and use $|a + b|^r \le 2^{(r-1)_+}(|a|^r + |b|^r)$. $\square$

LEMMA 8.4. *Suppose that $y \sim N_n(\mu, I)$ and that $\hat{\mu}(y)$ corresponds to soft or hard thresholding at random level $\hat{t} : \hat{\mu}(y)_i = \eta(y_i, \hat{t})$. Suppose that $\hat{t} \le t$ almost surely on the event $A$ (with $t \ge [E|z|^{2r}]^{1/2r}$). Then for $r > 0$,*

$$E_\mu[\|\hat{\mu} - \mu\|^r, A] \le 2^{r \vee 1/2} t^r n P_\mu(A)^{1/2}. \tag{8.14}$$

REMARK. The notation $E[X, A]$ denotes $EXI_A$ where $I_A$ is the indicator function of the event $A$.

PROOF OF LEMMA 8.4. Rewrite the left-hand side and use Cauchy–Schwarz:

$$\sum_{i=1}^n E[|\eta(y_i, \hat{t}) - \mu_i|^r, A] \le P(A)^{1/2} \sum_1^n \{E[|\eta_i - \mu_i|^{2r}, A]\}^{1/2}.$$

Now (8.13) and the bound on $\hat{t}$ imply

$$E[|\eta_i - \mu_i|^{2r}, A] \le 2^{(2r-1)_+}[E|z_i|^{2r} + \hat{t}^{2r}] \le 2^{2r \vee 1} t^{2r}. \qquad \square$$

Continuing with $y \sim N_n(\mu, I)$, the next lemma matches the $\ell_r$ risks of two hard threshold estimators $\hat{\mu}(y)_i = \eta_H(y_i; \hat{t})$ and $\hat{\mu}'$ with data-dependent thresholds $\hat{t}$ and $\hat{t}'$ if those thresholds are close. Assume also that there is a nonrandom bound $t$ such that $\hat{t}, \hat{t}' \le t$ with probability 1. Then

$$|\eta_H(y_i, \hat{t}) - \eta_H(y_i, \hat{t}')| \le \begin{cases} t, & \text{if } |y_i| \text{ lies between } \hat{t}, \hat{t}', \\ 0, & \text{otherwise.} \end{cases}$$

Let $N' = \#\{i : |y_i| \in [\hat{t}, \hat{t}']\}$—clearly $\|\hat{\mu} - \hat{\mu}'\|_r^r \le t^r N'$. In various cases, $N'$ can be bounded on a high probability event, yielding:

LEMMA 8.5. *Let $\beta_n$ be a specified sequence, and with the previous definitions, set $B_n = \{N' \le \beta_n\}$. For $0 < r \le 2$,*

$$\begin{aligned} |\rho(\hat{\mu}, \mu) - \rho(\hat{\mu}', \mu)| \le 2\beta_n t^r + rI\{r > 1\}\rho(\hat{\mu}', \mu)^{1-1/r} \beta_n^{1/r} t \\ + 8t^r n P_\mu(B_n^c)^{1/2}. \end{aligned} \tag{8.15}$$

PROOF. To develop an $\ell_r$ analog of (9.21), we note a simple bound valid for all $a, z \in \mathbb{R}$:

$$||a+z|^r - |a|^r| \le \begin{cases} |z|^r, & 0 < r \le 1, \\ r(|a|+|z|)^{r-1}|z|, & 1 < r. \end{cases} \tag{8.16}$$

(For $r > 1$, use derivative bounds for $y \to |y|^r$.) We consider here only $1 < r \le 2$; the case $r \le 1$ is similar and easier. Thus, setting $\varepsilon = \hat{\mu} - \hat{\mu}'$, $\Delta = \hat{\mu} - \mu$ and similarly for $\Delta'$,

$$\left| E\left\{ \sum_i |\Delta_i|^r - |\Delta_i'|^r, B_n \right\} \right| \le r E\left\{ \sum_i |\Delta_i'|^{r-1}|\varepsilon_i| + |\varepsilon_i|^r, B_n \right\}.$$

Using Hölder's inequality and defining $\epsilon_n = E\{\|\hat{\mu} - \hat{\mu}'\|_r^r, B_n\}$, we obtain

$$|E\{\|\Delta\|_r^r - \|\Delta'\|_r^r, B_n\}| \le r\rho(\hat{\mu}', \mu)^{(r-1)/r}\epsilon_n^{1/r} + r\epsilon_n. \tag{8.17}$$

From the definition of event $B_n$ and the remarks preceding the lemma, $\epsilon_n \le \beta_n t^r$. On $B_n^c$, apply Lemma 8.4 to obtain (8.15). $\square$

**9. Upper bound result: $\ell_2$ error.** We now turn to the upper bound, Theorem 4.1. We begin with the simplest case: squared-error loss. Only the outline of the argument is presented in this section, with details provided in the next section. The extensions to $\ell_r$ error measures, of considerable importance to the conclusions of the paper, are not straightforward. The proofs are postponed until Section 11.

The approach taken in this section was sketched in the Introduction; see (1.11) and (1.13). We define certain empirical and theoretical complexity functions—the empirical complexity being minimized by $\hat{\mu}_2$. A basic inequality bounds the theoretical complexity of $\hat{\mu}_2$ by the minimal theoretical complexity plus an "error term" of covariance type. When maximized over a sparse parameter set $\Theta_n$, the minimum theoretical complexity has the same leading asymptotics as the minimax estimation risk for $\Theta_n$. To complete the proof, the error term is bounded. This analysis is sketched in Section 9.3; the detailed proof relies on an average case and large deviations analysis of the penalized FDR index $\hat{k}_2$. The upshot is that these terms are negligible if $q \le 1/2$, but add substantially to the maximum risk when $q > 1/2$; this was foreshadowed in Proposition 5.5 and its discussion. Finally, a risk comparison is used to extend the conclusion from the penalized estimate $\hat{\mu}_2$ to the original FDR estimate $\hat{\mu}_F$.

9.1. *Empirical and theoretical complexities.* In Section 1.8 we have defined $\hat{\mu}_2$ as the minimizer of the *empirical* complexity $K(\tilde{\mu}, y) = \|y - \tilde{\mu}\|^2 + \mathrm{Pen}(\tilde{\mu})$ (note that now we set $\sigma^2 = 1$). Substituting $y = \mu + z$ into $K(\hat{\mu}_2, y)$ yields the decomposition

$$K(\hat{\mu}_2, y) = K(\hat{\mu}_2, \mu) + 2\langle z, \mu - \hat{\mu}_2\rangle + \|z\|^2. \tag{9.1}$$

Now let $\mu_0 = \mu_0(\mu)$ denote the minimizer over $\tilde{\mu}$ of the *theoretical* complexity $K(\tilde{\mu}, \mu)$ corresponding to the unknown mean vector $\mu$,

$$K(\mu_0, \mu) = \inf_{\tilde{\mu}} \|\mu - \tilde{\mu}\|^2 + \mathrm{Pen}(\tilde{\mu}). \tag{9.2}$$

There is a decomposition for $K(\mu_0, y)$ that is exactly analogous to (9.1),

$$K(\mu_0, y) = K(\mu_0, \mu) + 2\langle z, \mu - \mu_0\rangle + \|z\|^2.$$

Since by definition of $\hat{\mu}_2$, $K(\mu_0, y) \geq K(\hat{\mu}_2, y)$, we obtain, after noting the cancellation of the quadratic error terms and rearranging,

$$K(\hat{\mu}_2, \mu) \leq K(\mu_0, \mu) + 2\langle z, \hat{\mu}_2 - \mu_0\rangle. \tag{9.3}$$

Thus the complexity of $\hat{\mu}_2$ is bounded by the minimum theoretical complexity plus an error term. Up to this point, the development is close to that of [14], as well as work of other authors (e.g., [43]). Since

$$K(\hat{\mu}_2, \mu) = \|\hat{\mu}_2 - \mu\|^2 + \mathrm{Pen}(\hat{\mu}_2), \tag{9.4}$$

we obtain a bound for $\rho(\hat{\mu}_2, \mu) = E_\mu \|\hat{\mu}_2 - \mu\|^2$ by taking expectations in (9.3). Since the error term has zero mean, we may replace $\mu_0$ by $\mu$ and obtain the basic bound

$$\rho(\hat{\mu}_2, \mu) \leq K(\mu_0, \mu) + 2E_\mu\langle z, \hat{\mu}_2 - \mu\rangle - E_\mu \mathrm{Pen}(\hat{\mu}_2). \tag{9.5}$$

We view the right-hand side as containing a "leading term" $K(\mu_0, \mu)$—the theoretical complexity—and an "error term,"

$$\mathrm{Err}_2(\mu, \hat{\mu}_2) \equiv 2E_\mu\langle z, \hat{\mu}_2 - \mu\rangle - E_\mu \mathrm{Pen}(\hat{\mu}_2). \tag{9.6}$$

We now claim that the maximum theoretical complexity over sparsity classes $\Theta_n$ is asymptotic to the minimax risk.

PROPOSITION 9.1. *Assume* (Q), (H). *Then*

$$\sup_{\mu \in \Theta_n} K\big(\mu_0(\mu), \mu\big) \sim R_n(\Theta_n), \qquad n \to \infty. \tag{9.7}$$

The same minimax risk bounds the error term:

PROPOSITION 9.2. *Assume* (Q), (H). *With $u_{2p} = 1$ for $\ell_0$ and strong $\ell_p$, and $u_{2p} = 1 - p/2$ for weak $\ell_p$,*

$$\sup_{\mu \in \Theta_n} \mathrm{Err}_2(\mu, \hat{\mu}_2) = \begin{cases} R_n(\Theta_n) u_{2p} \dfrac{(2q-1)}{1-q}, & q > 1/2, \\ o(R_n(\Theta_n)), & q \leq 1/2\,. \end{cases} \tag{9.8}$$

Together these propositions give the upper bound result in the squared-error case.

9.2. *Maximal theoretical complexity.* We prove Proposition 9.1, beginning with the nearly-black classes $\Theta_n = \ell_0[\eta_n]$.

As in Section 1.8, decompose the optimization problem (9.2) defining $K(\mu_0, \mu)$ over the number of nonzero components in $\tilde{\mu}$. Assign these to the largest components of $\mu$: hence

$$K(\mu_0, \mu) = \inf_k \sum_{l=k+1}^{n} \mu_{(l)}^2 + \sum_{l=1}^{k} t_l^2. \tag{9.9}$$

On $\ell_0[\eta_n]$, at most $k_n = [n\eta_n]$ components of $\mu$ can be nonzero. Hence the infimum over $k$ may be restricted to $0 \le k \le k_n$. This implies

$$\sup_{\mu \in \ell_0[\eta_n]} K(\mu_0, \mu) = \sum_1^{k_n} t_l^2. \tag{9.10}$$

Indeed, choosing $k = k_n$ in (9.9) shows the left-hand side to be smaller than the right-hand side in (9.10), while equality occurs for any $\mu$ with nonzero entries $\mu_1 = \cdots = \mu_{k_n} > t_1$. Noting

$$\sum_1^k t_l^2 \sim k t_k^2, \qquad t_k^2 \sim 2\log\left(\frac{n}{k}\frac{2}{q}\right) \qquad \text{if } k = o(n) \tag{9.11}$$

[cf. (8.7) and Lemma 12.3], along with $\eta_n = O(n^{-\delta})$, we get

$$\sum_1^{k_n} t_l^2 \sim k_n t_{k_n}^2 \sim n\eta_n 2\log\eta_n^{-1} \sim R_n(\ell_0[\eta_n]),$$

which establishes (9.7) in the $\ell_0$ case.

REMARK. Using a fixed penalty $\mathrm{Pen}_{\mathrm{fix}}(\mu) = t^2\|\mu\|_0$ in the above argument would yield $\sup K = k_n t^2 \approx n\eta_n t^2$, but the $t^2$ term is unable to adapt to varying signal complexity.

*Weak $\ell_p$.* The maximum of (9.9) over $\mu \in m_p[\eta_n]$ occurs at the extremal vector $\bar{\mu}_l = C_n l^{-1/p}$, where $C_n = n^{1/p}\eta_n$. Define $\underline{k}_n$ to be the solution of $C_n^2 \underline{k}_n^{-2/p} = t_{\underline{k}_n}^2$. Using (9.11), we obtain

$$\begin{aligned} \sup_{\mu \in m_p[\eta_n]} K(\mu_0, \mu) &= \inf_k C_n^2 \sum_{k+1}^{n} l^{-2/p} + \sum_1^k t_l^2 \\ &\sim C_n^2 \tau_p \underline{k}_n^{1-2/p} + \underline{k}_n t_{\underline{k}_n}^2, \qquad \tau_p = \frac{p}{2-p} \\ &= (1+\tau_p)\underline{k}_n t_{\underline{k}_n}^2. \end{aligned} \tag{9.12}$$

Thus $\underline{k}_n$ is the optimal number of nonzero components and may be rewritten as

$$\underline{k}_n = C_n^p t_{\underline{k}_n}^{-p} = n\eta_n^p t_{\underline{k}_n}^{-p}. \tag{9.13}$$

A little analysis using (9.11) and the equation for $\underline{k}_n$ shows that $t_{\underline{k}_n}^2 \sim 2\log \eta_n^{-p}$. (For this reason, we define $k_n = n\eta_n^p \tau_\eta^{-p}$.) From (3.4) we then conclude $t_{\underline{k}_n}^2 \sim \mu_n^2$, which via (3.6) and (3.7) shows that the right-hand side of (9.12) is asymptotically equivalent to $R(m_p[\eta_n])$, as claimed.

REMARK. The least-favorable configuration for $\mu$ is thus given by $\mu_l = \min(C_n l^{-1/p}, t_l) = \min(\eta_n (l/n)^{-1/p}, t_l)$, which, after noting (9.11), is essentially identical with the least-favorable distribution (3.8). In addition, the maximization has exactly the same structure as the approximate evaluation of the Bayes risk of soft thresholding over this least-favorable distribution; compare (3.9)–(3.11). Replacing the slowly varying boundary $l \to t_l$ by $m_k = t[k_n] + \alpha$ leads to the configurations (5.17).

*Strong* $\ell_p$. The maximal theoretical complexity is the value of the optimization problem

$$\big(Q(n, \eta_n^p)\big) \qquad \max \sum \min\big(\mu_{(l)}^2, t_\ell^2\big) \quad \text{subject to} \quad \sum_\ell \mu_{(\ell)}^p \le n\eta_n^p.$$

The change of variables $x_\ell = \mu_{(\ell)}^p$ allows to write this as

$$\max \sum_\ell \min(x_\ell^{2/p}, t_\ell^2) \quad \text{subject to} \quad \sum_\ell x_\ell \le n\eta_n^p, \quad x_1 \ge x_2 \ge \cdots.$$

Since $p < 2$, the objective function is strictly convex on $\Pi_\ell[0, t_\ell^2]$, and the constraint set is convex. The maximum will be obtained at an extreme point of the constraint set, that is, roughly at a sequence vanishing for $\ell > k$ (for some $k$) and equal to $t_\ell^{1/p}$ for $\ell \le k$. Let $\tilde{k}_n$ be the largest $k$ for which

$$\sum_{\ell=1}^{k} t_\ell^p \le n\eta_n^p.$$

Then the maximal theoretical complexity obeys

$$\sum_{\ell=1}^{\tilde{k}_n} t_\ell^2 \le \operatorname{val}\big(Q(n, \eta_n^p)\big) \le \sum_{\ell=1}^{\tilde{k}_n+1} t_\ell^2.$$

Again using (9.11), we get

$$\sup\{K(\mu_0, \mu) : \mu \in \ell_p[\eta_n]\} \sim \tilde{k}_n t[\tilde{k}_n]^2 \sim n\eta_n^p \tau_\eta^{2-p}.$$

So (9.7) follows in the $\ell_p[\eta_n]$ case.

9.3. *The error term.* We outline the proof of (9.8). Recall the definitions $k_\pm$ and $t_\pm$ from Section 5, at (5.6), (5.5), (5.8), and their use in Proposition 5.1. We will rely on the fact that it is $\Theta_n$-likely under $P_\mu$ that $\hat{k}_2 \le k_+(\mu)$ and hence that $\hat{t}_2 \ge t_-(\mu)$. First write

$$\langle z, \hat{\mu}_2 - \mu\rangle = \sum_1^n z_i[\eta_H(y_i, \hat{t}_2) - \mu_i]. \tag{9.14}$$

We exploit monotonicity of the error term for *small* components $\mu_i$. Indeed, if $|\mu_i| \le t_-(\mu) \le \hat{t}_2$, then (cf. Lemma 10.1)

$$z_i[\eta_H(y_i, \hat{t}_2) - \mu_i] \le z_i\big[\eta_H\big(y_i, t_-(\mu)\big) - \mu_i\big], \tag{9.15}$$

as may be seen by checking cases. This permits us to replace $\hat{t}_2$ by the fixed threshold value $t_-(\mu)$ for the vast majority of components $\mu_i$, with Proposition 5.1 providing assurance that $\hat{t}_2 \ge t_-(\mu)$ with high probability. We recall the definition of the covariance kernel $\xi$ in Lemma 8.2. For $z_1 \sim N(0, 1)$ and scalar mean $\mu_1$,

$$\xi(t, \mu_1) = Ez_1[\eta_H(z_1 + \mu_1, t) - \mu_1].$$

The function $\xi(t, \mu_1)$ is the covariance between $y_1$ and $\eta_H(y_1, t)$ when the data $y_1 \sim N(\mu_1, 1)$. Lemma 8.2 shows that $\xi$ is even in $\mu_1$, and has a minimum of $2t\phi(t)$ at $\mu_1 = 0$, rising to a maximum near $\mu_1 = t$ (though always bounded by $t + 1$), and dropping quickly to 1 for large $\mu_1$. It turns out that, uniformly on nearly-black sequences $\mu \in \Theta_n$, the main contribution to the sum comes from components $\mu_1$ near 0.

Similarly, it is $\Theta_n$-likely that

$$\mathrm{Pen}(\hat{\mu}_2) = \sum_1^{\hat{k}_2} t_l^2 \ge k_-(\mu)t_+^2(\mu).$$

We proceed heuristically here, leaving the (necessary!) careful verification to Section 10. Interpreting "$\approx$" to mean "up to terms whose positive part is $o(R_n(\Theta))$," we have, uniformly on $\Theta_n$,

$$\mathrm{Err}_2(\mu, \hat{\mu}_2) \approx 2\sum \xi\big(t_-(\mu), \mu_i\big) - k_-(\mu)t_+^2(\mu) \tag{9.16}$$

$$\approx 4nt_-(\mu)\phi(t_-(\mu)) - k_-(\mu)t_+^2(\mu) \tag{9.17}$$

$$\approx [4n\tilde{\Phi}(t_-(\mu)) - k_-(\mu)]t_+^2(\mu) \tag{9.18}$$

$$\approx (2q_n - 1)k_-(\mu)t_+^2(\mu). \tag{9.19}$$

At (9.17) we first use the fact that $\xi(t, 0) = 2t\phi(t)$. Second, for the at most $k_n$ nonzero terms, we use the bound $\xi(t, \mu) \le t + 1 \le t_1 + 1$—compare (8.9)—and note that their contribution is at most $O(k_n t_1) = o(R_n)$. At (9.18) we use $\phi(t) \sim t\tilde{\Phi}(t)$ as $t \to \infty$, and at (9.19) the definitions of $t_+(\mu)$, $k_-$ and the asymptotics of each.

Expression (9.19) is negative if $q_n < 1/2$, making $\mathrm{Err}_2$ for our purposes negligible. For $q_n \geq 1/2$, since $k \to k t_k^2$ is increasing [cf. (12.10)], we use the bound of (5.3), namely $k(\mu) \leq \tilde{k}_n$ on $\ell_0[\eta_n]$, along with (9.11) to conclude that (9.19) is not larger than

$$(2q_n - 1)\tilde{k} t_{\tilde{k}}^2 \sim \frac{2q_n - 1}{1 - q_n} R_n(\ell_0[\eta_n]). \tag{9.20}$$

This motivates (9.8) in the $\ell_0$ case.

*Weak* $\ell_p$. The outline is much as above, although there is detailed technical work since all means may be nonzero (subject to the weak-$\ell_p$ sparsity constraint), for example, in the transition (9.16) to (9.17). Again, after (9.19) we are led to maximize $k_-(\mu)$ over $m_p[\eta]$ and from (5.4), find $k_-(\mu) \leq k(\bar{\mu}) \leq k_n/(1 - q_n) = \tilde{k}$, say. Here $k_n = n\eta_n^p \tau_\eta^{-p}$ is the effective nonzero index for weak $\ell_p$ defined after (9.13).

Consequently, since $t_{\tilde{k}} \sim \tau_\eta$, and using the expressions (3.6) and (3.7) for minimax risks, we obtain

$$\begin{aligned}
(2q_n - 1)k_-(\mu)t_+^2(\mu) &\leq (2q_n - 1)\tilde{k} t_{\tilde{k}}^2 \bigl(1 + o(1)\bigr) \\
&\sim \frac{2q_n - 1}{1 - q} k_n \tau_\eta^2 \sim \frac{2q_n - 1}{1 - q} n\eta_n^p \tau_\eta^{2-p} \\
&\sim \frac{2q_n - 1}{1 - q} R_n(\ell_p[\eta_n]) \sim u_{2p} \frac{2q_n - 1}{1 - q} R(m_p[\eta_n]),
\end{aligned}$$

with $u_{2p} = (1 - p/2)$.

*Strong* $\ell_p$. The inclusion $\ell_p[\eta_n] \subset m_p[\eta_n]$ and the previous display give

$$(2q_n - 1)k_-(\mu)t_+^2(\mu) \leq \frac{2q_n - 1}{1 - q} R_n(\ell_p[\eta_n])\bigl(1 + o(1)\bigr).$$

If the above arguments were complete—rather than just sketches—we would now have the right-hand side of (9.8) in Proposition 9.2. Details will come in Section 10.

9.4. *From penalized to original FDR.* We extend the adaptive minimaxity result for the penalized estimator $\hat{\mu}_2$ which thresholds at $\hat{t}_2$ to any threshold $\hat{t}$ in the range $[\hat{t}_F, \hat{t}_G]$ defined in Section 1.8. In particular, the adaptive minimaxity will apply to the original FDR estimator $\hat{\mu}_F$.

First compare the squared error of a deviation $\hat{\delta}_2 = \hat{\mu}_2 - \mu$ with that of $\hat{\delta} = \hat{\mu} - \mu$:

$$\|\hat{\delta}\|_2^2 - \|\hat{\delta}_2\|_2^2 = \|\hat{\mu} - \hat{\mu}_2\|_2^2 + 2(\hat{\mu} - \hat{\mu}_2) \cdot (\hat{\mu}_2 - \mu). \tag{9.21}$$

Now suppose $\hat{\mu}_D$ ("$D$" for "data dependent") has the form (4.1). All such estimators differ from $\hat{\mu}_2$ at most in those coordinates $y_l$ with $\hat{k}_G \le l \le \hat{k}_F$, and on such coordinates the difference between the two estimates is at most $\hat{t}_G \le t_1 = z(q/2n)$. Hence

$$\|\hat{\mu} - \hat{\mu}_2\|_2^2 \le t_1^2(\hat{k}_F - \hat{k}_G).$$

Proposition 5.1 and (5.7) show that FDR control, combined with sparsity, forces the "crossover interval" $[\hat{k}_G, \hat{k}_F]$ to be relatively small, having length bounded by $\gamma\alpha_n k_n$.

On the event described in Proposition 5.1, we have

$$\|\hat{\mu} - \hat{\mu}_2\|_2^2 \le \gamma\alpha_n k_n t_1^2 = o(R_n(\Theta_n)). \tag{9.22}$$

We summarize, with remaining details deferred to Section 11.4.

THEOREM 9.3. *If $\hat{\mu}_D$ satisfies (4.1), then for each $r \in (0, 2]$*

$$\sup_{\mu\in\Theta_n} |\rho(\hat{\mu}_D, \mu) - \rho(\hat{\mu}_r, \mu)| = o(R_n(\Theta_n)),$$

*so that asymptotic minimaxity of $\hat{\mu}_r$ implies the same property for any such $\hat{\mu}_D$.*

**10. Error term: quadratic loss.** We now formalize the error term analysis of Section 9.3, collecting and applying the tools built up in earlier sections.

LEMMA 10.1. *If $|\mu| \le t^1 \le t^2$, then*

$$(x - \mu)[\eta_H(x, t^2) - \mu] \le (x - \mu)[\eta_H(x, t^1) - \mu].$$

PROOF. The difference RHS − LHS equals

$$(x - \mu)[\eta_H(x, t^1) - \eta_H(x, t^2)] = (x - \mu)xI\{t^1 \le |x| \le t^2\} \ge 0,$$

since $\operatorname{sgn} x = \operatorname{sgn}(x - \mu)$ if $|x| \ge t^1 \ge |\mu|$. $\square$

We proceed with the formal analysis of the error term (9.6). Set

$$\hat{e}_i = e_i(\hat{t}_2) = 2(y_i - \mu_i)[\eta_H(y_i, \hat{t}_2) - \mu_i]$$

and

$$A_n = A_n(\mu) = \{t_- \le \hat{t}_2 \le t_+\}, \qquad S_n(\mu) = \{i : |\mu_i| \le t_-\}.$$

We have

$$\begin{aligned} 2E\langle z, \hat{\mu}_2 - \mu\rangle &= E\sum \hat{e}_i \\ &= E\Bigg[\sum_{S_n(\mu)} \hat{e}_i, A_n\Bigg] + E\Bigg[\sum_{S_n^c(\mu)} \hat{e}_i, A_n\Bigg] + E\Big[\sum \hat{e}_i, A_n^c\Big] \end{aligned} \tag{10.1}$$

$$= D_{an} + T_{2n} + T_{3n}, \tag{10.2}$$

where we use $D_{an}$, $D_{bn}$ and so on to denote "dominant" terms, and $T_{jn}$ to denote terms that will be shown to be negligible.

Let $e_i = e_i(t_-)$: the monotonicity of errors for small components (shown in Lemma 10.1) says that the first term on the right-hand side is bounded above by

$$E\Bigg[\sum_{S_n(\mu)} e_i, A_n\Bigg] = E\Bigg[\sum_{S_n(\mu)} e_i\Bigg] - E\Bigg[\sum_{S_n(\mu)} e_i, A_n^c\Bigg] = D_{bn} + T_{4n}.$$

Recalling the definition of $\xi(t,\mu)$ from Section 8.2, we have $Ee_i = 2\xi(t_-, \mu_i)$ and

$$\begin{aligned} D_{bn} &= 2|S_n(\mu)|\xi(t_-, 0) + 2\sum_{S_n(\mu)} [\xi(t_-, \mu_i) - \xi(t_-, 0)] \\ &\le 2n\xi(t_-, 0) + T_{1n}(\mu), \end{aligned}$$

say. To summarize, we obtain the following decomposition for the error term (9.6):

$$\mathrm{Err}_2(\mu, \hat{\mu}_2) \le D_{cn}(\mu) + \sum_{j=1}^{4} T_{jn}(\mu),$$

where

$$D_{cn}(\mu) = 2n\xi(t_-, 0) - E_\mu \,\mathrm{Pen}(\hat{\mu}_2).$$

Recall that $R_n(\Theta_n) \asymp k_n\tau_\eta^2$ for both $\ell_0[\eta_n]$ and $m_p[\eta_n]$. In the following, we will show negligibility of error terms by establishing that they are $O(k_n\tau_\eta)$ [or, in one case, $o(k_n\tau_\eta^2)$] uniformly over $\ell_0[\eta_n]$ or $m_p[\eta_n]$, respectively.

*Dominant term.* Using (12.1),

$$\xi(t, 0) = 2t\phi(t) + 2\tilde{\Phi}(t) \le 2t^2\tilde{\Phi}(t) + 6\tilde{\Phi}(t).$$

Since $2\tilde{\Phi}(t_-) = q_n k_+ n^{-1}$, we obtain

$$2n\xi(t_-, 0) \le 2q_n k_+ t_-^2 + 6q_n k_+ \le 2q_n k_- t_-^2 + ck_n\tau_\eta, \tag{10.3}$$

after observing that $k_+ - k_- \le 3\alpha_n k_n \le ck_n\tau_\eta^{-1}$ by assumption (A), and that $t_- \le c\tau_\eta$ from (12.14).

For the penalty term in $D_{cn}$, we note that on $A_n$, $\hat{k}_2 \ge k_-$, and so $\mathrm{Pen}(\hat{\mu}_2) = \sum_1^{\hat{k}_2} t_l^2 \ge k_- t_+^2 \ge k_- t_-^2$. On the other hand, since $A_n$ is $\Theta_n$-likely,

$$E_\mu\Bigg[\sum_1^{\hat{k}_2} t_l^2, A_n^c\Bigg] \le nt_1^2 P_\mu(A_n^c) \le ck_n\tau_\eta.$$

As a result

$$E_\mu \,\mathrm{Pen}(\hat{\mu}_2) \ge k_- t_-^2 + O(k_n\tau_\eta). \tag{10.4}$$

Combining (10.3) and (10.4), we obtain

$$D_{cn}(\mu) \le (2q_n - 1)k_- t_-^2 + O(k_n \tau_\eta).$$

If $q_n \le 1/2$, then of course the leading term is nonpositive, while in the case $1/2 \le q_n < 1$, we note from the monotonicity of $k \to k t_k^2$ [cf. (12.10)] and the definition (6.37) of $\kappa_n$ that

$$k_- t_-^2 \le k_+ t^2[k_+] \le \kappa_n t^2[\kappa_n] \sim (1 - q_n)^{-1} k_n \tau_\eta^2,$$

which leads to the second term in the upper bound of (4.2).

*Negligibility of $T_{1n} - T_{4n}$.* Consider first the term $T_{1n}(\mu)$. For the nonzero $\mu_l$ (of which there are at most $n\eta_n$), use the bound (8.9) to get

$$T_{1n}(\mu) \le k_n(t_1 + 1) \le k_n \tau_\eta.$$

For the large signal component term $T_{2n}$, we have, using Lemma 8.4 and the bound $\hat{t}_2 \le t_1 = O(\log^{1/2} n)$,

$$\begin{aligned} T_{2n} &\le \sum_{S_n^c(\mu)} E[\hat{e}_i, A_n] \\ &\le 2 \sum_{S_n^c(\mu)} \{E[\eta(y_i, \hat{t}_2) - \mu_i]^2\}^{1/2} \\ &\le c_0 t_1 |S_n^c(\mu)| \\ &\le |S_n^c(\mu)|(c_0 \log n)^{1/2}. \end{aligned} \tag{10.5}$$

On $\ell_0[\eta_n]$, clearly $|S_n^c(\mu)| \le n\eta_n$ and so $T_{2n}(\mu) \le c_1 k_n \tau_\eta$.

For the small threshold term $T_{3n}$, note first that $\sum \hat{e}_i \le 2\|z\| \|\hat{\mu}_2 - \mu\|$, so that

$$T_{3n}(\mu) \le 2(E\|z\|^2)^{1/2} \{E[\|\hat{\mu}_2 - \mu\|^2, A_n^c]\}^{1/2}.$$

Now $E\|z\|^2 = n$, and since $A_n^c$ is a rare event, apply Lemma 8.4, noting the bound $\hat{t}_2 \le t_1$. Thus

$$T_{3n}(\mu) \le 8nt_1 P_\mu(A_n^c)^{1/4} \le c_1 n t_1 \exp\{-c_2 \log^2 n\} = o(R_n(\ell_0[\eta_n])),$$

uniformly on $\ell_0[\eta_n]$ after applying Proposition 5.1.

The remaining term $T_{4n}$ is handled exactly as was $T_{3n}$: if we let $\hat{\mu}_F$ denote hard thresholding at the (fixed) threshold $t_-$, then $\sum_{S_n(\mu)} e_i \le \|z\| \|\hat{\mu}_F - \mu\|$; and now Lemma 8.4 and Proposition 5.1 can be used as before.

*Weak $\ell_p$.* A little extra work is required to analyze $T_{1n}(\mu)$, so we first dispose of $T_{2n} - T_{4n}$. The analysis of $T_{3n}$ and $T_{4n}$ is essentially as above. For $T_{2n}$, we bound $|S_n^c(\mu)|$ using the extremal element of $m_p[\eta_n]$, namely $\bar{\mu}_l = \eta_n(n/l)^{1/p}$. Thus, for all $\mu \in m_p[\eta_n]$,

$$S_n^c(\mu) \subset \{l : \bar{\mu}_l > t_-(\mu)\},$$

and since, for $\eta$ sufficiently small,

$$t_-(\mu) = t[k_+(\mu)] \geq t[k_+(\bar{\mu})] \geq \tau_\eta - 3/2 = \tau_\eta\bigl(1 + o(1)\bigr), \tag{10.6}$$

we have

$$|S_n^c(\mu)| \leq n\eta_n^p t_-^{-p}(\mu) \leq n\eta_n^p \tau_\eta^{-p}\bigl(1 + o(1)\bigr). \tag{10.7}$$

From (3.6) we have $R_n = R(\ell_p[\eta_n]) \sim n\eta_n^p \tau_\eta^{2-p}$, and so from (10.5) and (10.7) we get

$$T_{2n}(\mu) \leq c_0 t_1 n\eta_n^p \tau_\eta^{-p} \leq ck_n\tau_\eta.$$

For the $T_{1n}$ term, we obtain from Lemma 8.2 that

$$\xi(t, \mu) - \xi(t, 0) \leq c(\mu^2 \wedge t),$$

and so $T_{1n}(\mu) \leq 2c \sum \bar{\mu}_l^2 \wedge t_1$. The negligibility of $T_{1n}$ is a consequence of the following.

LEMMA 10.2. *For* $0 < p < r \leq 2$, *we have*

$$\sum_l \bar{\mu}_l^r \wedge t_1^{(r-1)_+} = o(k_n\tau_\eta^r).$$

PROOF. Define $\tilde{k}$ by $\bar{\mu}_{\tilde{k}} = \tau_\eta / \log \tau_\eta$ so that $\tilde{k} = n\eta_n^p \tau_\eta^{-p} \log^p \tau_\eta$. We have

$$\sum_{l=1}^n \bar{\mu}_l^r \wedge t_1^{(r-1)_+} \leq \tilde{k} t_1^{(r-1)_+} + \sum_{l > \tilde{k}} \bar{\mu}_l^r.$$

Since $t_1 \leq c\tau_\eta$,

$$\tilde{k} t_1^{(r-1)_+} \leq ck_n \tau_\eta^{(r-1)_+} \log^p \tau_\eta = o(k_n\tau_\eta^r).$$

By integral approximation, since $0 < p < r$,

$$\begin{aligned}\sum_{l > \tilde{k}} \bar{\mu}_l^r &\leq n\eta_n^r \int_{\tilde{k}/n}^1 x^{-r/p}\, dx \leq c_{rp} n\eta_n^r (\tilde{k}/n)^{1-r/p} \\ &= c_{rp} n\eta_n^p \tau_\eta^{r-p} \log^{p-r} \tau_\eta = o(k_n\tau_\eta^r). \qquad \square\end{aligned}$$

**11. $\ell_r$ losses.** This section retraces for $\ell_r$ loss the steps used for squared error in Section 9, making adjustments for the fact that the quadratic decomposition (9.1) is no longer available. It turns out that this decomposition is merely a convenience—the asymptotic result of Theorem 4.1 is as sharp for all $0 < r \le 2$. However, the analysis of the error term is more complex than in Section 10, requiring bounds developed in Lemmas 11.1 and 11.2.

11.1. *Empirical complexity for $\ell_r$ loss.* For an $\ell_r$ loss function, we use a modified empirical complexity,

$$K(\tilde{\mu}, y; r) = \|y - \tilde{\mu}\|_r^r + \sum_{l=1}^{N(\tilde{\mu})} t_l^r.$$

The minimizers of empirical and theoretical complexity are defined, respectively, by

$$\hat{\mu}_r = \arg\min_{\tilde{\mu}} K(\tilde{\mu}, y; r),$$
$$\mu_0 = \arg\min_{\tilde{\mu}} K(\tilde{\mu}, \mu; r).$$

For $\ell_r$ loss, the quadratic decomposition of (9.1) is replaced by

$$K(\tilde{\mu}, \mu + z) = K(\tilde{\mu}, \mu) + \|\mu - \tilde{\mu} + z\|_r^r - \|\mu - \tilde{\mu}\|_r^r. \tag{11.1}$$

The key inequality

$$K(\hat{\mu}_r, y) \le K(\mu_0, y),$$

when combined with (11.1), applied to both $\tilde{\mu} = \hat{\mu}_r$ and $\mu_0$, yields the analog of (9.3),

$$K(\hat{\mu}_r, \mu) \le K(\mu_0, \mu) + D(\hat{\mu}_r, \mu_0, \mu, y).$$

Setting

$$\hat{\delta} = \mu - \hat{\mu}_r, \qquad \delta_0 = \mu - \mu_0, \qquad y = \mu + z, \tag{11.2}$$

we have for the error term

$$\hat{D} = D(\hat{\mu}_r, \mu_0, \mu, y) = \|\delta_0 + z\|_r^r - \|\delta_0\|_r^r - \|\hat{\delta} + z\|_r^r + \|\hat{\delta}\|_r^r = \sum_1^n \hat{d}_l. \tag{11.3}$$

Thus, with $\mathrm{Err}_r \equiv E_\mu \hat{D} - E_\mu \mathrm{Pen}_r(\hat{\mu}_r)$,

$$E\|\hat{\mu}_r - \mu\|_r^r \le K(\mu_0, \mu; r) + \mathrm{Err}_r(\mu, \hat{\mu_r}). \tag{11.4}$$

11.2. *Maximum theoretical complexity.* The theoretical complexity corresponding to $\mu$ is given by

$$K(\mu_0, \mu; r) = \inf_k \sum_{k+1}^{n} |\mu|_{(l)}^r + \sum_1^k t_l^r. \tag{11.5}$$

We may argue as in Section 9.2 that for $\Theta_n = \ell_0[\eta_n]$,

$$\sup_{\mu \in \Theta_n} K(\mu_0, \mu) \le \sum_1^{k_n} t_l^r \sim k_n t_{k_n}^r \sim n\eta_n (2 \log \eta_n^{-1})^{r/2} \sim R_n(\Theta_n; r), \tag{11.6}$$

and that for $\Theta_n = m_p[\eta_n]$,

$$\sup_{\mu \in \Theta_n} K(\mu_0, \mu) = \inf_k \ C_n^r \sum_{k+1}^{n} l^{-r/p} + \sum_1^k t_l^r \sim R_n(\Theta_n; r).$$

Finally, for $\Theta_n = \ell_p[\eta_n]$, we may argue that

$$\sup_{\mu \in \Theta_n} K(\mu_0, \mu) \sim \max_k \left\{ \sum_1^k t_l^r : \sum_1^k t_l^p \le n\eta_n^p \right\} \sim R_n(\Theta_n; r).$$

We remark that if $\underline{k}(\mu)$ is an index minimizing (11.5), then $\mu_{0i}$ is obtained from hard thresholding of $\mu_i$ at $t_0 = t[\underline{k}(\mu)]$ [interpreted as $t_1$ if $k(\mu) = 0$]. In any event, this implies

$$|\delta_{0i}| = |\mu_i - \mu_{0i}| \le |\mu_i| \wedge t_1. \tag{11.7}$$

11.3. *The $\ell_r$ error term.* There is an $\ell_r$ analog of bound (9.15); this allows us to replace the random threshold $\hat{t}_r$ by the fixed threshold value $t_\kappa$ for the most important cases.

Indeed, suppose that $|\mu_i| \le t_\kappa \le \hat{t}_r$. Let $\bar{\mu}_i(y) = \eta_H(y_i, t_\kappa)$ denote hard thresholding at $t_\kappa$, and let $\bar{\delta}_i = \mu_i - \bar{\mu}_i$ denote the corresponding deviation. We claim that

$$|\hat{\delta}_i|^r - |\hat{\delta}_i + z_i|^r \le |\bar{\delta}_i|^r - |\bar{\delta}_i + z_i|^r. \tag{11.8}$$

Indeed, $\hat{\delta}_i = \bar{\delta}_i$ unless $t_\kappa \le |y_i| \le \hat{t}_r$. In this case, we have $\hat{\mu}_i = 0$ so that $\hat{\delta}_i = \mu_i$, while $\bar{\mu}_i = y_i$ so that $\bar{\delta}_i = -z_i$. In this case, (11.8) reduces to

$$|\mu_i|^r - |y_i|^r \le |z_i|^r,$$

which is trivially true since $|\mu_i| \le t_\kappa \le |y_i|$.

We now derive the $\ell_r$ analog of the error decomposition (10.1). Recalling the notation (11.2)–(11.4), we have

$$\hat{d}_i = d_i(\hat{t}_r) = |\delta_{0i} + z_i|^r - |\delta_{0i}|^r - |\hat{\delta}_i + z_i|^r + |\hat{\delta}_i|^r. \tag{11.9}$$

Defining as in Section 9.3 the sets $A_n = \{t_- \le \hat{t}_r \le t_+\}$ and $S_n(\mu) = \{i : |\mu_i| \le t_-\}$, we obtain

$$E_\mu \hat{D} = E \sum \hat{d}_i = E\Biggl[\sum_{S_n(\mu)} \hat{d}_i, A_n\Biggr] + E\Biggl[\sum_{S_n^c(\mu)} \hat{d}_i, A_n\Biggr] + E\Bigl[\sum \hat{d}_i, A_n^c\Bigr]$$
$$= D_{an} + T_{2n} + T_{3n}.$$

Let $d_i = d_i(t_-)$; the monotonicity of errors for small components [cf. (11.8)] says that the leading term

$$D_{an} \le E\Biggl[\sum_{S_n(\mu)} d_i, A_n\Biggr] = E\Biggl[\sum_{S_n(\mu)} d_i\Biggr] - E\Biggl[\sum_{S_n(\mu)} d_i, A_n^c\Biggr] = D_{bn} + T_{4n}.$$

Consider first the dominant term $D_{bn}$. First, write

$$\begin{aligned} Ed_i &= E_\mu|\delta_{0i} + z_i|^r - |\delta_{0i}|^r - |\bar{\delta}_i + z_i|^r + |\bar{\delta}_i|^r \\ &= \psi_r(\delta_{0i}) + \xi_r(t_-, \mu_i), \end{aligned} \tag{11.10}$$

where, for $y = \mu + z,\ z \sim N(0, 1)$ and $0 < r \le 2$, we define

$$\psi_r(a) = E[|a + z|^r - |a|^r - |z|^r], \tag{11.11}$$

$$\xi_r(t, \mu) = E[|\eta_H(y, t) - \mu|^r - |\eta_H(y, t) - y|^r + |y - \mu|^r]. \tag{11.12}$$

[Note that a term $E|z|^r$ has been introduced in both $\psi_r$ and $\xi_r$; as a result $\psi_2(a) \equiv 0$ and $\xi_2(t, \mu) = 2\xi(t, \mu)$ as defined at (8.8).] The next lemmas, proved in the extended version [3], play the same role as Lemma 8.2 for the $\ell_2$ case.

LEMMA 11.1. *The function $\psi_r(a)$ defined at (11.11) is even in $a$ and*

$$|\psi_r(a)| \le \begin{cases} C_1|a|^r, & \textit{for all } a, \\ C_2|a|^{(r-1)_+}, & \textit{for } |a| \textit{ large}. \end{cases} \tag{11.13}$$

LEMMA 11.2. *The function $\xi_r(t, \mu)$ defined at (11.12) is even in $\mu$ and satisfies*

$$\xi_r(t, 0) = 2\int_{|z|>t} |z|^r \phi(z)\,dz, \tag{11.14}$$

$$|\xi_r(t, \mu)| \le C\bigl[t^{(r-1)_+} + 1\bigr], \qquad \mu \in \mathbb{R},\ t > 0. \tag{11.15}$$

With the preceding notation, we may therefore write

$$D_{bn}(\mu) = \sum_{S_n(\mu)} \psi_r(\delta_{0i}) + \xi_r(t_-, \mu_i) = |S_n(\mu)|\xi_r(t_-, 0) + T_{1n}(\mu),$$

say. To summarize, we obtain the following decomposition for the error term in (11.4):

$$E_\mu \hat{D} - E_\mu \operatorname{Pen}_r(\hat{\mu}) \le D_{cn}(\mu) + \sum_{j=1}^{4} T_{jn}(\mu),$$
$$D_{cn}(\mu) = n\xi_r(t_-, 0) - E_\mu \operatorname{Pen}_r(\hat{\mu}_r).$$

*Dominant term.* Using (8.6),

$$\xi_r(t, 0) \le 4t^r \tilde{\Phi}(t)[1 + 2t^{-2}].$$

Since $2\tilde{\Phi}(t_-) = q_n k_+ n^{-1}$, we obtain

$$n\xi_r(t_-, 0) \le 2q_n k_+ t_-^r + 8q_n k_+ t_-^{r-2} \le 2q_n k_- t_-^r + ck_n\tau_\eta^{r-1},$$

since $k_+ - k_- \le ck_n\tau_\eta^{-1}$ and $t_-(\mu) \asymp \tau_\eta$.

For the penalty term, arguing as before yields

$$E_\mu \operatorname{Pen}_r(\hat{\mu}_r) \ge k_- t_-^r + O(k_n\tau_\eta^{r-1}).$$

Combining the two previous displays, we obtain

$$D_{cn}(\mu) \le (2q_n - 1)k_- t_-^r + O(k_n\tau_\eta^{r-1}).$$

If $q_n \le 1/2$, of course the leading term is nonpositive, while in the case $1/2 \le q_n < 1$, we note from (12.10) and the definition (6.37) of $\kappa_n$ that

$$k_-(\mu)t_-^r(\mu) \le k_+ t^r[k_+] \le \kappa_n t^r[\kappa_n] \sim (1 - q_n)^{-1} k_n \tau_\eta^r,$$

which shows that $D_{cn}(\mu)$ is bounded by the second term in the upper bound of (4.2).

*Negligibility of $T_{1n} - T_{4n}$.* Since there are at most $k_n$ nonzero terms in $T_{1n}$, from Lemmas 11.1 and 11.2 we obtain

$$T_{1n}(\mu) \le Ck_n t_1^{(r-1)_+} = o(k_n\tau_\eta^r).$$

To bound $T_{2n}$, we first note from the properties of hard thresholding [cf. (8.13)] that

$$|\hat{\delta}_i| = |\hat{\mu}_{r,i} - \mu_i| \le \hat{t}_r + |z_i| \le t_1 + |z_i|.$$

Inequality (11.7) shows that $|\delta_{0i}| \le t_1$.

Combined with (11.9) and (8.16), this shows, for $1 < r \le 2$,

$$|\hat{d}_i| \le 3t_1^{r-1}|z_i| + 6|z_i|^r, \tag{11.16}$$

while for $0 < r \le 1$ only the $|z_i|^r$ term is needed. Consequently, there exist constants $C_i$ such that for $0 < r \le 2$

$$E|\hat{d}_l| \le C_1 t_1^{(r-1)_+} \quad \text{and} \quad E\hat{d}_i^2 \le C_2 t_2^{2(r-1)_+}. \tag{11.17}$$

Thus

$$|T_{2n}| \le \sum_{S_n^c(\mu)} E|\hat{d}_i| \le C|S_n^c(\mu)|t_1^{(r-1)_+}. \tag{11.18}$$

And so on $\ell_0[\eta_n]$,

$$|T_{2n}| \le Cn\eta_n t_1^{(r-1)_+} = o(k_n \tau_\eta^r).$$

To bound $T_{3n}(\mu)$, use (11.17) and Cauchy–Schwarz:

$$T_{3n}(\mu) \le P(A_n^c)^{1/2} \sum_i (E\hat{d}_i^2)^{1/2} \le cnt_1^{(r-1)_+} \exp\{-c_0(\log n)^2/2\} = o(R_n(\Theta_n)),$$

since $A_n$ is $\Theta_n$-likely. Argue similarly for $T_{4n}$, with threshold at $t_-(\mu)$ instead of $\hat{t}_r$.

*Weak* $\ell_p$. For the $T_{1n}$ term, we use a consequence of Lemmas 11.1 and 11.2, proved in the extended report [3], to bound the summands in $T_{1n}(\mu)$:

$$\psi_r(\delta_{0i}) + |\xi_r(t_-, \mu_i) - \xi_r(t_-, 0)| \le C\big[|\mu_i|^r \wedge t_1^{(r-1)_+}\big]. \tag{11.19}$$

Combined with Lemma 10.2, this shows that $\sup_{m_p[\eta_n]} T_{1n}(\mu) = o(k_n \tau_\eta^r)$.

For $T_{2n}$, we use (10.7) in (11.18) to obtain

$$|T_{2n}| \le Cn\eta_n^p t_\eta^{r-p} t_1^{(r-1)_+ - r} = o(R_n(\Theta_n)).$$

The analysis of $T_{3n}$ and $T_{4n}$ is as for the $\ell_0$ case.

11.4. *From penalized to original FDR.*

PROOF OF THEOREM 9.3. Apply Lemma 8.5 with $\hat{\mu} = \hat{\mu}_D$ and $\hat{\mu}'$ and $t = t_1$. We abbreviate the minimax risk $R_n(\Theta_n)$ by $R_n$. Theorem 4.1 shows that for sufficiently large $n$, $\sup_{\Theta_n} \rho(\hat{\mu}_r, \mu) \le c_0 R_n$, so that the bound established by Lemma 8.5 yields

$$\begin{aligned} &\sup_{\mu \in \Theta_n} |\rho(\hat{\mu}_D, \mu) - \rho(\hat{\mu}_r, \mu)| \\ &\quad \le 2\beta_n t_1^r + 2c_0 I\{r > 1\} R_n^{1-1/r} (\beta_n t_1^r)^{1/r} + 8nt_1^r \sup_{\Theta_n} P_\mu(B_n^c)^{1/2}. \end{aligned} \tag{11.20}$$

The thresholds $\hat{t}_D$ and $\hat{t}_r$ corresponding to $\hat{\mu}_D$ and $\hat{\mu}_r$ both lie in $[\hat{t}_G, \hat{t}_F]$, and so with probability 1,

$$N' = \#\{i : |y_i| \in [\hat{t}_D, \hat{t}_r]\} \;\le\; \#\{i : \hat{t}_F \le |y_i| < \hat{t}_G\} \le \hat{k}_F - \hat{k}_G.$$

[The first inequality is valid except possibly on a zero probability event in which some $|y_i| = \hat{t}_G$. To see the second inequality, note from (7.1) that $l > \hat{k}_F$ implies $Y_l \le t_l < \hat{t}_F$, while (7.2) entails that $l \le \hat{k}_G$ implies $Y_l \ge t_l \ge \hat{t}_G$. Consequently $\hat{t}_F \le Y_l < \hat{t}_G$ implies $\hat{k}_G < l \le \hat{k}_F$, which yields the required inequality.]

If we take $\beta_n = 3\alpha_n k_n$, then Proposition 5.1 implies that

$$P_\mu(B_n^c) = P_\mu(N' > \beta_n) \le P_\mu(\hat{k}_F - \hat{k}_G > 3\alpha_n k_n) \le c_0 \exp\{-c_1 \log^2 n\},$$

so that the third term in (11.20) is $o(R_n)$. Finally $\beta_n t_1^r \le c\alpha_n k_n \tau_\eta^r \le c\alpha_n R_n$ so that the first two terms are also $o(R_n)$, completing the proof. $\square$

**12. Gaussian tails and quantiles.** We collect in this section some results about the normal density $\phi$, the normal CDF $\Phi$ and the normal quantile function $z(\cdot)$; these have been used extensively above. All proofs are given in the extended report [3].

LEMMA 12.1. *We have*

$$\phi(v) \le (v + 2v^{-1})\tilde{\Phi}(v), \qquad v \ge \sqrt{2}. \tag{12.1}$$

*More generally, Mills' ratio $M(y) = y\tilde{\Phi}(y)/\phi(y)$ increases from $0$ to $1$ as $y$ increases from $0$ to $\infty$. In particular,*

$$\phi(v)/(2v) \le \tilde{\Phi}(v) \le \phi(v)/v, \qquad v \ge 1, \tag{12.2}$$

$$2\tilde{\Phi}(v) \le e^{-v^2/2}, \qquad v \ge 0, \tag{12.3}$$

$$\tilde{\Phi}(v - c/v) \le 4e^c \tilde{\Phi}(v), \qquad v \ge \sqrt{2c}. \tag{12.4}$$

LEMMA 12.2. *Suppose that $k$ and $\alpha$ are such that $\max\{\alpha, 1/\alpha\} \le C \log k$. Then*

$$\sqrt{2 \log k\alpha} = \sqrt{2 \log k} + \theta\sqrt{C}, \tag{12.5}$$

*and if $\alpha \ge 1$, then $0 \le \theta \le \sqrt{2}/e \le 1$, while if $\alpha \le 1$, then $-1.1 \le -\sqrt{8}/e \le \theta \le 0$.*

LEMMA 12.3. (1) *Let $z(\eta) = \tilde{\Phi}^{-1}(\eta)$ denote the upper $(1 - \eta)$th percentile of the Gaussian distribution. If $\eta \le 0.01$, then*

$$z^2(\eta) = 2 \log \eta^{-1} - \log\log \eta^{-1} - r_2(\eta), \qquad r_2(\eta) \in [1.8, 3], \tag{12.6}$$

$$z(\eta) = \sqrt{2 \log \eta^{-1}} - r_1(\eta), \qquad r_1(\eta) \in [0, 1.5]. \tag{12.7}$$

(2) *We have $z'(\eta) = -1/\phi(z(\eta))$, and hence if $\eta_2 > \eta_1 > 0$, then*

$$z(\eta_1) - z(\eta_2) \le \frac{\eta_2 - \eta_1}{\eta_1 z(\eta_1)}. \tag{12.8}$$

*In addition, if $t_\nu = z(\nu q/2n) \ge 1$, then*

$$-\partial t_\nu/\partial \nu = \theta/(\nu t_\nu), \qquad \theta \in \left[\tfrac{1}{2}, 1\right], \tag{12.9}$$

*and for $0 \le r \le 2$ and $t_\nu^2 > 2$,*

$$\partial(\nu t_\nu^r)/\partial\nu = t_\nu^{r-2}[1 - r\theta t_\nu^{-2}] > 0. \tag{12.10}$$

(3) *If* $n^{-1}\log^5 n \le \eta_n^p \le b_2 n^{-b_3}$, *then*

$$2b_3 \log n - 2\log b_2 \le \tau_\eta^2 \le 2\log n - 10\log\log n, \tag{12.11}$$

*and so for* $n > n(b)$, *we have*

$$\tau_\eta^2 = 2\gamma_n \log n, \qquad 0 < c(b) \le \gamma_n \le 1. \tag{12.12}$$

(4) *If* $q_n \ge b_1/\log n$ *and* $\nu \le b_2 n^{1-b_3}$, *then for* $n \ge n(b_2, b_3)$,

$$-3/2 \le t_\nu - \sqrt{2\log(n/\nu)} \le 2(b_1 b_3)^{-1/2}. \tag{12.13}$$

(5) (i) *For* $n > n(b)$,

$$t_1/\tau_\eta \le c(b). \tag{12.14}$$

(ii) *If* $a \le 1$ *and* $\eta_n^p \le e^{-1/2}$, *then*

$$t[ak_n] \ge \tau_\eta - 3/2. \tag{12.15}$$

*If* $a \le \delta^{-1}$, *the same inequality holds for* $\eta < \eta(p, \delta)$ *sufficiently small.*

(iii) *If* $a \ge \gamma\tau_\eta^{-1}$, *then for* $\eta_n^p \le \eta(\gamma, p, b_1, b_3)$ *sufficiently small* (*and not the same at each appearance*),

$$t[ak_n] \le \tau_\eta + c(b_1, b_3), \tag{12.16}$$

$$t[ak_n] \le 2\tau_\eta. \tag{12.17}$$

(iv) *In particular, for* $a \in [\gamma\tau_\eta^{-1}, \delta^{-1}]$, *then as* $\eta_n \to 0$,

$$t[ak_n] \sim \tau_\eta. \tag{12.18}$$

**13. Proofs of lower bounds.** This final section combines ideas from Sections 6–8 to finish the lower bound result.

PROOF OF PROPOSITION 5.2. From the structure (5.9) of the configuration $\mu_\alpha$, the total risk can be written in terms of the univariate component risks as

$$\rho(\hat{\mu}_{H,t}, \mu_\alpha) = k_n \rho_H(t, m_\alpha) + (n - k_n)\rho_H(t, 0).$$

From (8.6) together with the definition of $t = t[ak_n]$, we obtain

$$n\rho_H(t[ak_n], 0) = ak_n q_n t^r[ak_n](1 + \theta),$$

with $0 \le \theta \le 2t[ak_n]^{-2}$. Since $t[ak_n] \sim \tau_\eta$ by (12.18), we conclude that

$$(n - k_n)\rho_H(t[ak_n], 0) \sim aq_n k_n \tau_\eta^r.$$

In the notation of (8.2),

$$\rho_H(t, m_\alpha) = m_\alpha^r[\Phi(t - m_\alpha) - \Phi(-t - m_\alpha)] + E(m_\alpha, t),$$

and as noted there, $0 \le E(m_\alpha, t) \le c_r$. In addition, $m_\alpha^r \Phi(-t - m_\alpha) \le c'_r$, and as noted at (5.13), $m_\alpha - t = \alpha + O(\tau_\eta^{-1})$ and so $\Phi(t - m_\alpha) = \tilde{\Phi}(\alpha) + O(\tau_\eta^{-1})$. Consequently, since $m_\alpha \sim \tau_\eta$, we conclude that

$$k_n \rho_H(t, m_\alpha) \sim k_n \tau_\eta^r \tilde{\Phi}(\alpha)\bigl(1 + o(1)\bigr). \qquad \square$$

PROOF OF PROPOSITION 5.3. We use (8.2) to decompose the total risk

$$\rho(\hat{\mu}_{H,t}, \mu_\alpha) = \sum_l D(\mu_{\alpha l}, t) + E(\mu_{\alpha l}, t) = D + E,$$

say. To bound the "bias" or "false-negative" term $D$, we choose an index $l_\alpha = n\eta_n^p m_\alpha^{-p}$ so that $\bar{\mu}_{l_\alpha} = m_\alpha \sim \tau_\eta$. Now decompose $D$ into $D_1 + D_2$ according as $l \le l_\alpha$ or $l > l_\alpha$. For $l \le l_\alpha$, we have identically $\mu_{\alpha l} \equiv m_\alpha$, and so

$$D_1 = l_\alpha m_\alpha^r [\Phi(t - m_\alpha) - \Phi(-t - m_\alpha)] = l_\alpha m_\alpha^r \tilde{\Phi}(\alpha)\bigl(1 + o(1)\bigr)$$

by the same arguments as for Proposition 5.2. And since $m_\alpha \sim \tau_\eta$, we have

$$l_\alpha m_\alpha^r = n\eta_n^p m_\alpha^{r-p} \sim n\eta_n^p \tau_\eta^{r-p} = k_n \tau_\eta^r.$$

The novelty with the weak-$\ell_p$ risk comes in the analysis of

$$D_2 = \sum_{l > l_\alpha} \bar{\mu}_l^r [\Phi(t - \bar{\mu}_l) - \Phi(-t - \bar{\mu}_l)].$$

The second term is negligible, being bounded in absolute value by $\tilde{\Phi}(t) \sum_{l > l_\alpha} \bar{\mu}_l^r = o(k_n \tau_\eta^r)$. For the first term we have an integral approximation [with $\bar{\mu}(x) = \eta_n (n/x)^{1/p}$]

$$D_2 \sim \int_{l_\alpha}^n \bar{\mu}^r(x) \Phi\bigl(t - \bar{\mu}(x)\bigr)\, dx = p l_\alpha m_\alpha^r \int_{\eta_n/m_\alpha}^1 v^{r-p-1} \Phi(t - m_\alpha v)\, dv,$$

after setting $x = l_\alpha v^{-p}$.

REMARK. To bound the error in the integral approximation, observe that if $f'(x)$ is smooth with at most one zero in $[a, b]$, then the difference between $\sum_a^b f(l)$ and $\int_a^b f$ is bounded by $\sup_{[a,b]} |f|$.

For $0 < v < 1$, we have $t - m_\alpha v \sim (1 - v)\tau_\eta \to \infty$, and so from the dominated convergence theorem, the integral converges to $\int_0^1 v^{r-p-1}\, dv$, so

$$D_2 \sim [p/(r - p)] l_\alpha m_\alpha^r \sim [p/(r - p)] k_n \tau_\eta^r.$$

Putting together the analyses of $D_1$ and $D_2$, we find that the false-negative term

$$D = \bigl[\tilde{\Phi}(\alpha) + [p/(r - p)]\bigr] k_n \tau_\eta^r \bigl(1 + o(1)\bigr).$$

To bound the "variance" or "false-positive" term $E$, decompose the sum into three terms, corresponding to indices in the ranges $[1, l_1]$, $(l_1, l_2]$ and $(l_2, n]$, where

$$l_1 = n\eta_n^p \leftrightarrow \bar{\mu}_{l_1} = 1 \quad \text{and} \quad l_2 = n\eta_n^p t_\eta^{2p} \leftrightarrow \bar{\mu}_{l_2} = \tau_\eta^{-2}.$$

We use (i)–(iii) of Lemma 8.1 to show that terms $E_1$ and $E_2$ are negligible. For $E_1$, the global bound (i) gives $E_1 \le c_r l_1 = o(k_n \tau_\eta^r)$. For $E_2$, properties (ii) (monotonicity) and (iii) show that

$$E(\bar{\mu}_l, t) \le E(\bar{\mu}_{l_1}, t) = E(1, t) \le c_0 t^r \tilde{\Phi}(t-1) \le c_1 \tau_\eta^r \tilde{\Phi}(\tau_\eta - 5/2),$$

where the last inequality uses $t = t[ak_n] \sim \tau_\eta$ and that $t \ge \tau_\eta - 3/2$ from (12.15). Hence

$$E_2 \le l_2 E(1, t) \le c_0 k_n \tau_\eta^r \tau_\eta^{3p} \tilde{\Phi}(\tau_\eta - 5/2) = o(k_n \tau_\eta^r).$$

Finally, we focus on the dominant term $E_3 = \sum_{l > l_2} E(\bar{\mu}_l, t)$. For $l > l_2$, we have $\bar{\mu}_l \le \tau_\eta^{-2}$ and $t \le c\tau_\eta$ so that

$$\phi(t \pm \bar{\mu}_l) = \phi(t) \exp\{\mp t\bar{\mu}_l - \bar{\mu}_l^2/2\} = \phi(t)\big(1 + O(\tau_\eta^{-1})\big).$$

Now if $|\mu| \le \tau_\eta^{-1}$ and $t = t[ak_n] \le 2\tau_\eta$ [from (12.17)], then

$$\phi(t - \mu) = \phi(t) \exp(t\mu - \mu^2/2) = \phi(t)\big(1 + O(\tau_\eta^{-1})\big),$$

and so

$$\gamma(t - \mu) = |t - \mu|^{r-1} \phi(t - \mu) = t^{r-1} \phi(t)\big(1 + O(\tau_\eta^{-1})\big),$$

$$\varepsilon(t - \mu) \le |t - \mu|^{r-3} \phi(t - \mu) \le t^{r-3} \phi(t)\big(1 + O(\tau_\eta^{-1})\big).$$

Consequently, using (8.4) and $t^{-1}\phi(t) \sim \tilde{\Phi}(t)$, we get

$$E_3 = 2(n - l_2) t^{r-1} \phi(t)\big(1 + O(\tau_\eta^{-1})\big) = aq_n k_n \tau_\eta^r \big(1 + o(1)\big),$$

using the manipulations of (5.14)–(5.16). $\square$

PROOF OF PROPOSITION 5.4. Fix $a_1 > 0$. The idea, both for $\ell_0[\eta]$ and for $m_p[\eta_n]$, is to obtain bounds

$$M_-(k) \le M(k; \mu_\alpha) \le M_+(k), \qquad k \in [a_1 k_n, a_1^{-1} k_n],$$

for which solutions $k_\pm$ to $M_\pm(k) = k$ can be easily found. From monotonicity of $k \to M(k; \mu_\alpha)$, it then follows that $k_- \le k(\mu_\alpha) \le k_+$.

In both cases we establish a representation of the form

$$M(ak_n; \mu_\alpha) = k_n[\Phi(\alpha) + \theta_1 \tau_\eta^{-1}] + [1 + \theta_2 \delta(\eta_n)] q_n k, \tag{13.1}$$

where $|\theta_i| \le c(\alpha, a, b)$ and $\delta(\eta_n) \to 0$ as $\eta_n \to 0$. From this, expressions for $k_\pm$ are easily found and $k_\pm \sim k_n \Phi(\alpha)/(1 - q_n)$ is easily checked.

For $\ell_0[\eta_n]$, (13.1) follows from (5.19) and (5.13). For $m_p[\eta_n]$, we formulate a lemma, proven in the extended report [3].

LEMMA 13.1. *Let $a$ and $\alpha$ be fixed. If $|f|$ and $|f'|$ are bounded by* 1, *then*

$$k_n^{-1}\sum_{l=1}^{k_n} f(\mu_{\alpha l} - t[ak_n]) = f(\alpha) + \theta\tau_\eta^{-1}, \qquad |\theta| \le c(\alpha, a, b).$$

We exploit the division into the "positive," "transition" and "negative" zones defined in Section 6.4. Applying the preceding lemma with $f = \Phi$ yields

$$M_{\text{pos}}(k; \mu_\alpha) = k_n[\Phi(\alpha) + \theta_3\tau_\eta^{-1}],$$

while from (6.27) and (6.29) we obtain

$$M_{\text{trn}}(k; \mu_\alpha) = \theta_4 k_n \tau_\eta^{-1} \quad \text{and} \quad M_{\text{neg}}(k; \mu_\alpha) = [1 + \theta_5\delta_p(\varepsilon_n)]q_n k.$$

Putting together the last two displays, we recover (13.1) and the result. □

PROOF OF PROPOSITION 5.5. Let $t'$ denote the fixed threshold $t' = t[a_0 k_n]$, where $a_0 = (1-q)^{-1}\Phi(\alpha)$. We will use Lemma 8.5 to show that

$$|\rho(\hat{\mu}_F, \mu_\alpha) - \rho(\hat{\mu}_{H,t'}, \mu_\alpha)| = o(k_n \tau_\eta^r),$$

so that the conclusion will follow from Proposition 5.3.

To apply Lemma 8.5, set $\hat{\mu} = \hat{\mu}_F$ and $\hat{\mu}' = \hat{\mu}_{H,t'}$, so that the thresholds $\hat{t} = \hat{t}_F$ and $\hat{t}' = t'$, respectively, which are both bounded by $t_1$.

Let $k_\pm = k_\pm(\mu_\alpha) = k(\mu_\alpha) \pm \alpha_n k_n$ and recall that the event $A_n = \{k_- \le \hat{k}_F \le k_+\}$ is $m_p[\eta_n]$-likely. If we set $t_\pm = t[k_\mp]$, then on event $A_n$,

$$N' = \#\{i : |y_i| \in [\hat{t}_F, t']\} \le N'' = \#\{i : |y_i| \in [t_-, t_+]\}.$$

Hence $P(N' > \beta_n) \le P(A_n^c) + P(N'' > \beta_n)$.

We use the exponential bound of Lemma 7.1 to choose $\beta_n$ so that $P(N'' > \beta_n)$ is small. From the definition of the threshold function, $M(k) = M(k; \mu_\alpha)$, we have

$$M'' := EN'' = M(k_+) - M(k_-).$$

Since $k \to \dot{M}_k$ is decreasing, and using the derivative bounds of Proposition 6.6, we find that for $n$ sufficiently large,

$$(k_+ - k_-)\dot{M}(k_+) \le M'' \le (k_+ - k_-)\dot{M}(k_-). \tag{13.2}$$

CLAIM. $\dot{M}(k_\pm; \mu_\alpha) = \theta[q_n + c(\alpha)\tau_\eta^{-1}]$ *for some* $\theta \in [1/2, 2]$.

PROOF. We again use the positive-transition-negative decomposition, this time of $\dot{M}(ak_n; \mu_\alpha)$. Write, with $\nu = ak_n$,

$$\dot{M}_{\text{pos}}(ak_n; \mu_\alpha) = (-\partial t_\nu/\partial\nu)\sum_1^{k_n} \phi(t - \mu_{\alpha l}) + \phi(t + \mu_{\alpha l}).$$

From (12.8) and (12.9), we have $-k_n(\partial t_\nu/\partial\nu) \sim 1/(a\tau_\eta)$. Applying Lemma 13.1 to $f(x) = \phi(-x)$, we conclude that for $n$ sufficiently large,

$$\dot{M}_{\text{pos}}(ak_n; \mu_\alpha) = (\theta_1/(a\tau_\eta))[\phi(\alpha) + \theta_2\tau_\eta^{-1}].$$

Appealing to (6.28) and (6.30),

$$\dot{M}_{\text{trn}}(ak_n; \mu_\alpha) = \theta_3\tau_\eta^{-2} \quad \text{and} \quad \dot{M}_{\text{neg}}(ak_n; \mu_\alpha) = [1 + \theta_4\delta_p(\varepsilon_n)]q_n.$$

Combining the last two displays and noting that $k_\pm = k(\mu_\alpha) \pm \alpha_n k_n$ correspond to $a = \phi(\alpha)(1 - q_n)^{-1}$, we obtain the claim. $\square$

Now set $q_{\alpha n} = q_n + c(\alpha)\tau_\eta^{-1}$ and select $\beta_n = 8\alpha_n k_n q_{\alpha n}$. From (13.2) and the claim, we have

$$\alpha_n k_n q_{\alpha n} \le M'' \le 4\alpha_n k_n q_{\alpha n},$$

and so $\beta_n/M'' \ge 2$. Consequently,

$$\begin{aligned} P(N'' > \beta_n) &\le \exp\{-(1/4)M''h(\beta_n/M'')\} \\ &\le \exp\{-(1/4)\alpha_n k_n q_{\alpha n} h(2)\} \\ &\le c_0 \exp\{-c_1 \log^2 n\}. \end{aligned}$$

Now $\beta_n t_1^r \asymp \alpha_n q_{\alpha n} k_n \tau_\eta^r = o(k_n \tau_\eta^r)$, while Proposition 5.3 shows that $\rho(\hat{\mu}_{H,t'}, \mu_\alpha) = O(k_n \tau_\eta^r)$. So Lemma 8.5 applies and we are done. $\square$

**Acknowledgments.** D. L. Donoho would like to thank F. Abramovich and Yoav Benjamini for hospitality at Tel Aviv University during a sabbatical there. An early version of this paper was presented by Iain M. Johnstone as a Special Invited Lecture at the 1999 Annual Meeting of the IMS in Baltimore.

## REFERENCES

[1] ABRAMOVICH, F. and BENJAMINI, Y. (1995). Thresholding of wavelet coefficients as multiple hypotheses testing procedure. *Wavelets and Statistics*. *Lecture Notes in Statist.* **103** 5–14. Springer, Berlin.

[2] ABRAMOVICH, F. and BENJAMINI, Y. (1996). Adaptive thresholding of wavelet coefficients. *Comput. Statist. Data Anal.* **22** 351–361. MR1411575

[3] ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. and JOHNSTONE, I. M. (2005). Adapting to unknown sparsity by controlling the false discovery rate. (Extended version.) Available at www.arXiv.org/abs/math.st/0505374.

[4] AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (B. N. Petrov and F. Czáki, eds.) 267–281. Akadémiai Kiadó, Budapest. MR0483125

[5] BENEDETTO, J. and ANTHONY, T. (1993). A wavelet auditory model and data compression. *Appl. Comput. Harmon. Anal.* **1** 3–28.

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

[7] BENJAMINI, Y. and HOCHBERG, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educational and Behavioral Statistics* **25** 60–83.

[8] BENJAMINI, Y., KRIEGER, A. M. and YEKUTIELI, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*. To appear.

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

[10] BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3** 203–268. MR1848946

[11] BREIMAN, L. and FREEDMAN, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78** 131–136. MR0696857

[12] DEVORE, R. A., JAWERTH, B. and LUCIER, B. (1992). Image compression through wavelet transform coding. *IEEE Trans. Inform. Theory* **38** 719–746. MR1162221

[13] DONOHO, D. L. (1995). De-noising by soft-thresholding. *IEEE Trans. Inform. Theory* **41** 613–627. MR1331258

[14] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal denoising in an orthonormal basis chosen from a library of bases. *C. R. Acad. Sci. Paris Sér. I Math.* **319** 1317–1322. MR1310679

[15] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over $\ell_p$-balls for $\ell_q$-error. *Probab. Theory Related Fields* **99** 277–303. MR1278886

[16] DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464

[17] DONOHO, D. L., JOHNSTONE, I. M., HOCH, J. C. and STERN, A. S. (1992). Maximum entropy and the nearly black object (with discussion). *J. Roy. Statist. Soc. Ser. B.* **54** 41–81. MR1157714

[18] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: Asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 301–369. MR1323344

[19] EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81** 461–170. MR0845884

[20] EFRON, B., TIBSHIRANI, R. J., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571

[21] FIELD, D. J. (1987). Relation between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Amer. A* **4** 2379–2394.

[22] FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975. MR1329177

[23] FOSTER, D. P. and STINE, R. A. (1999). Local asymptotic coding and the minimum description length. *IEEE Trans. Inform. Theory* **45** 1289–1293. MR1686271

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

[25] GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747. MR1813972

[26] HOCHBERG, Y. and TAMHANE, A. (1987). *Multiple Comparison Procedures*. Wiley, New York. MR0914493

[27] HOMMEL, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75** 383–386.

[28] HUANG, J. and MUMFORD, D. (1999). Statistics of natural images and models. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **1** 541–547.

[29] JOHNSTONE, I. M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. In *Statistical Decision Theory and Related Topics V* (S. Gupta and J. Berger, eds.) 303–326. Springer, New York. MR1286310

[30] JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135

[31] MALLOWS, C. L. (1973). Some comments on $C_P$. *Technometrics* **15** 661–675.

[32] POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York. MR0762984

[33] RUDERMAN, D. L. (1994). The statistics of natural images. *Network*: *Computation in Neural Systems* **5** 517–548.

[34] SARKAR, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30** 239–257. MR1892663

[35] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

[36] SEEGER, P. (1968). A note on a method for the analysis of significances en masse. *Technometrics* **10** 586–593.

[37] SHAFFER, J. (1995). Multiple hypothesis testing. *Ann. Rev. Psychology* **46** 561–584.

[38] SIMES, R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754. MR0897872

[39] SIMONCELLI, E. P. (1999). Bayesian denoising of visual images in the wavelet domain. *Bayesian Inference in Wavelet Based Models*. *Lecture Notes in Statist.* **141** 292–308. Springer, New York.

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

[41] STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 187–205. MR2035766

[42] TIBSHIRANI, R. and KNIGHT, K. (1999). The covariance inflation criterion for adaptive model selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 529–546. MR1707859

[43] VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. MR1056343

F. ABRAMOVICH
Y. BENJAMINI
DEPARTMENT OF STATISTICS
AND OPERATIONS RESEARCH
SCHOOL OF MATHEMATICAL SCIENCES
THE SACKLER FACULTY OF EXACT SCIENCES
TEL AVIV UNIVERSITY
RAMAT AVIV, TEL AVIV 69978
ISRAEL
E-MAIL: felix@math.tau.ac.il
ybenja@post.tau.ac.il

D. L. DONOHO
I. M. JOHNSTONE
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065
USA
E-MAIL: donoho@stanford.edu
imj@stanford.edu