Econometrica, Vol. 91, No. 1 (January, 2023), 1–41

INVIDIOUS COMPARISONS: RANKING AND SELECTION AS COMPOUND DECISIONS

JIAYING GU
Department of Economics, University of Toronto

ROGER KOENKER
Department of Economics, University College London

There is an innate human tendency, one might call it the "league table mentality," to construct rankings. Schools, hospitals, sports teams, movies, and myriad other objects are ranked even though their inherent multi-dimensionality would suggest that—at best—only partial orderings were possible. We consider a large class of elementary ranking problems in which we observe noisy, scalar measurements of merit for n objects of potentially heterogeneous precision and are asked to select a group of the objects that are "most meritorious." The problem is naturally formulated in the compound decision framework of Robbins's (1956) empirical Bayes theory, but it also exhibits close connections to the recent literature on multiple testing. The nonparametric maximum likelihood estimator for mixture models (Kiefer and Wolfowitz (1956)) is employed to construct optimal ranking and selection rules. Performance of the rules is evaluated in simulations and an application to ranking U.S. kidney dialysis centers.

KEYWORDS: Empirical Bayes, compound decisions, multiple testing.

1. INTRODUCTION

IN THE WAKE OF WALD'S seminal monograph on statistical decision theory, there was a growing awareness that the Neyman–Pearson testing apparatus was inadequate for many important statistical tasks. Ranking and selection problems featured prominently in this perception. Motivated by a suggestion of Harold Hotelling, Bahadur (1950) studied selection of the best of several Gaussian populations. Assuming that sample means were observed for each of K populations with means θ_k and common variance, the problem of selecting the best population, $\theta^* = \max_i\{\theta_1, \ldots, \theta_K\}$, was formulated as choosing weights $z_1, \ldots, z_K$ to minimize

$$L(\theta, z) = \theta^* - \sum_{k=1}^{K} z_k \theta_k / \sum_{k=1}^{K} z_k.$$

showed that among "impartial decision rules," that is, permutation equivariant rules, it was uniformly optimal to select only the population with the largest sample mean, that is, to choose $z_i^* = 1$ if $\bar{X}_i = \max\{\bar{X}_1, \ldots, \bar{X}_K\}$ and $z_i^* = 0$ otherwise, thereby clearly demonstrating that procedures that did preliminary tests of equality of means and then chose $z_i > 0$ for several or even all of the populations when tests failed to reject were inadmissible. This finding was reinforced by Bahadur and Robbins (1950), who focused on the

Jiaying Gu: jiaying.gu@utoronto.ca
Roger Koenker: r.koenker@ucl.ac.uk

This paper was presented as the Walras–Bowley Lecture at the 2020 World Congress of the Econometric Society, and is dedicated to the memory of Larry Brown who introduced us to empirical Bayes methods. We thank the anonymous referees and Michael Gilraine, Keisuke Hirano, Robert McMillan, Stanislav Volgushev, and Sihai Dave Zhao for useful discussions. Jiaying Gu acknowledges financial support from Social Sciences and Humanities Research Council of Canada.

two-sample setting but relaxed the common variance assumption. In related work, Bechhofer (1954) and Gupta (1956) sought to optimize the number of selected populations as well as their identities; see Gupta and Panchapakesan (1979) and Bechhofer, Kiefer, and Sobel (1968) for extensive reviews of subsequent developments.

Goel and Rubin (1977) pioneered the hierarchical Bayesian approach to selection that has been adopted by numerous authors in the ensuing decades, early on by Berger and Deely (1988) and Laird and Louis (1989). Portnoy (1982) showed that rankings based on best linear predictors were optimal in Gaussian multivariate variance components models, but cautioned that departures from normality could easily disrupt this optimality. A notable feature of the hierarchical model paradigm is the recognition that sample observations may exhibit heterogeneous precision; this is typically accounted for by assuming known variances for observed sample means. As ranking and selection methods became increasingly relevant in genomic applications, there has been renewed interest in loss functions and linkages to the burgeoning literature on multiple testing. Our perspective is informed by recent developments in the nonparametric estimation of mixture models and its relevance for a variety of compound decision problems. This approach seeks to reduce the reliance on Gaussian distributional assumptions that pervades the earlier literature. As we have argued elsewhere (Gu and Koenker (2016b) and Koenker and Gu (2019)), nonparametric empirical Bayes methods offer powerful complementary methods to more conventional parametric hierarchical Bayes for multiple testing and compound decision problems. Our primary objective in this paper is to elaborate this assertion for ranking and selection applications. Throughout, we try to draw parallels and contrasts with the literature on multiple testing. We will restrict our attention to settings where we observe a scalar estimate of an unobserved latent quality measure accompanied by some measure of its precision, thereby evading more complex multivariate settings, as in Boyd, Cortes, Mohri, and Radovanovic (2012), who employed quantile regression methods.

An important motivation for revived interest in ranking and selection problems in econometrics has been the influential work of Chetty and his collaborators on teacher evaluation and geographic mobility in the United States. This has stimulated the important recent work of Mogstad, Romano, Shaikh, and Wilhelm (2020) proposing new resampling methods for constructing confidence sets for ranking and selection for a finite population. Armstrong, Kolesár, and Plagborg-Møller (2020) proposed an innovative approach to the construction of confidence intervals for classical, linear shrinkage, empirical Bayes estimators of the type used by Chetty. Recent work by Andrews, Kitagawa, and McCloskey (2020) and Guo and He (2020) proposed new confidence interval constructions for highly ranked individuals or treatments influenced by recent contributions to the "inference after model selection" literature. In contrast to these inferential approaches, we focus instead on the complementary perspective of compound decision making, constructing decision rules for selecting the best, or worst, populations subject to control of the expected number of elements selected, and among those selected, the expected proportion of false discoveries. Rather than treating each selection decision in isolation, the compound decision framework tries to exploit their common structure to produce improved *collective* performance. Our approach is thus more closely aligned to that of Kline and Walters (2021), who studied decision rules for assessing employer discrimination from experiments involving fictitious job applications using closely related GMM methods for binomial mixture models.

Gilraine, Gu, and McMillan (2020) studied teacher value-added estimation employing nonparametric maximum likelihood methods for estimating Gaussian mixture models as we advocate below. Their analysis of data from both North Carolina and Los Angeles

illustrates the advantages of more flexible mixture models for latent value added. In contrast to the present work, they focused on Bayes rules for posterior means that are often used to study teachers' influence on students' future outcomes. These more flexible nonparametric empirical Bayes methods improve upon traditional linear shrinkage rules especially in the tails of the distribution where policy attention is usually focused. This is a valuable, complementary perspective to the ranking and selection objectives of the present work.

Before proceeding, it is important to acknowledge that despite its universal appeal and application, there is something inherently futile about many ranking and selection problems, as intimated by our title. If the latent measure of true quality is Gaussian, as assumed in virtually all of the econometric applications of the selection problem, and we wish to select the top ten percent of individuals given that their true quality is contaminated by Gaussian noise, accurate selection can be very challenging when the signal-to-noise ratio is low. We will see that conventional linear shrinkage as embodied in the classical James–Stein formula can improve performance considerably over naive maximum likelihood (fixed effects) procedures, and some further improvement is possible by carefully tailoring the decision rules for tail probability loss. However, we find that even oracle decision rules that incorporate complete knowledge of the precise distributional features of the problem may not be able to achieve better than about even odds that selected individuals have latent ability above the selection thresholds when measurement error is comparable in magnitude to Gaussian variability in latent ability. When the latent distribution of ability is heavier tailed, then selection becomes somewhat easier, and more refined selection rules are more advantageous, but as we will show, the selection problem still remains quite challenging.

Thus, a secondary objective of the paper is to add another cautionary voice to those who have already questioned the reliability of existing ranking and selection methods. A critical overview of the role of ranking and selection in public policy applications was provided by Goldstein and Spiegelhalter (1996). It is widely acknowledged that league tables as currently employed can be a pernicious influence on policy, a viewpoint underscored in Gelman and Price (1999). While much of this criticism can be attributed to inadequate data collection and inherently low signal-to-noise ratios, we believe that there is also room for methodological improvements.

Section 2 provides a brief overview of compound decision theory and describes nonparametric methods for estimation of Gaussian mixture models. Section 3 introduces a basic framework for our approach to ranking and selection in a setting with homogeneous precision of the observed measurements. In Section 4, we introduce heterogeneous precision of known form, and Section 5 considers settings in which the joint distribution of the observed measurements and their precision determines the form of the ranking and selection rules. Optimal ranking and selection rules are derived in each of these sections under the assumption that the form of the mixing distribution of the unobserved, latent quality of the observations is known. Section 6 introduces feasible ranking and selection rules and conditions under which they attain the same asymptotic performance as the optimal rules. Section 7 then compares several *feasible* ranking and selection methods, some that ignore the compound decision structure of the problem, some that employ parametric empirical Bayes methods, and some that rely on nonparametric empirical Bayes methods. Finally, Section 8 describes an empirical application on evaluating the performance of medical dialysis centers in the United States. Proofs of all formal results are collected in Appendix A of the Supplemental Material (Gu and Koenker (2023)).

2. THE COMPOUND DECISION FRAMEWORK

Robbins (1951) posed a challenge to the nascent minimax decision theory of Wald (1950): Suppose we observe independent Gaussian realizations, $Y_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, \ldots, n$, with means θ_i taking either the value $+1$ or -1. We are asked to estimate the n-vector $\theta = (\theta_1, \ldots, \theta_n)$ subject to mean absolute error loss,

$$L(\hat{\theta}, \theta) = n^{-1} \sum_{i=1}^{n} |\hat{\theta}_i - \theta_i|.$$

When $n = 1$, Robbins showed that the minimax decision rule is $\delta(y) = \operatorname{sgn}(y)$; in the least favorable variant of the problem, malevolent nature chooses ± 1 with equal probability, and the optimal response is to estimate $\theta_i = +1$ when Y_i is positive, and $\theta_i = -1$ otherwise. Robbins went on to show that when $n > 1$, this rule remains minimax; each coordinate is treated independently as if viewed in complete isolation. This is also the maximum likelihood estimator, and may be viewed in econometrics terms as a classical fixed-effects estimator. But is it at all reasonable?

Doesn't our sample convey information about the relative frequency of ± 1 that might potentially contradict the pessimistic presumption of the minimax rule? If we happened to know the unconditional probability, $p = \mathbb{P}(\theta_i = 1)$, then the conditional probability that $\theta = 1$ given $Y_i = y$ is given by

$$\mathbb{P}(\theta = 1|y) = \frac{p\varphi(y-1)}{p\varphi(y-1) + (1-p)\varphi(y+1)},$$

where φ denotes the standard Gaussian density. We should guess $\hat{\theta}_i = 1$ if this probability exceeds 1/2, giving us the revised decision rule

$$\delta_p(y) = \operatorname{sgn}\left(y - \frac{1}{2}\log\big((1-p)/p\big)\right).$$

Each observed y_i is modified by a simple logistic perturbation before computing the sign. Our observed random sample, $\boldsymbol{y} = (y_1, \ldots, y_n)$, is informative about p. We have the log likelihood

$$\ell_n(p|y) = \sum_{i=1}^{n} \log\big(p\varphi(y_i - 1) + (1-p)\varphi(y_i + 1)\big),$$

which could be augmented by a prior of some form, if desired, to obtain a posterior mean for p and a plug-in Bayes rule for estimating each of the θ_i's. The Bayes risk of this procedure is substantially less than the minimax risk when $p \neq 1/2$ and is asymptotically equivalent to the minimax risk when $p = 1/2$. This is the first principle of compound decision theory: borrowing strength across an entire ensemble of related decision problems yields improved collective performance.

What happens when we relax the restriction on the support of the θ's and allow support on the whole real line? We now have a general Gaussian mixture setting where the observed Y_i's have marginal density given by the convolution, $f = \varphi * G$, that is,

$$f(y) = \int \varphi(y - \theta)\, dG(\theta),$$

and instead of merely needing to estimate one probability, we need an estimate of an entire distribution function, G. Kiefer and Wolfowitz (1956), anticipated by an abstract of Robbins (1950), established that the nonparametric maximum likelihood estimator (NPMLE),

$$\hat{G} = \underset{G \in \mathcal{G}}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} \log f(y_i) | f(y_i) = \int \varphi(y_i - \theta)\, dG(\theta) \right\},$$

where $\mathcal{G}$ is the space of probability measures on $\mathbb{R}$, is a consistent estimator of G. This is an infinite-dimensional convex optimization problem with a strictly convex objective subject to linear constraints. See Lindsay (1995) and Koenker and Mizera (2014) for further details on the geometry and computational aspects of the NPMLE problem. Heckman and Singer (1984) pioneered this approach in econometrics to argue that more flexible models of heterogeneity were needed to get reliable estimates of duration dependence in survival models.

A powerful consequence of the seemingly innocuous condition that G must be nondecreasing is that $\hat{G}$ must be atomic, a discrete distribution with fewer than n atoms. A secondary consequence is that the NPMLE is "self-regularizing," that is, the number, locations, and mass of the atoms are all determined jointly by the optimization without any recourse to auxiliary tuning parameters. This is all a consequence of the classical Carathéodory theorem, but until quite recently little was known about the precise growth rate of the number of atoms characterizing the solutions, although empirical experience suggested it was quite slow. Polyanskiy and Wu (2020) have recently established that, for G with sub-Gaussian tails, the cardinality of its support, that is, the number of atoms, of $\hat{G}$ grows like $\mathcal{O}(\log n)$. Thus, without any further penalization, maximum likelihood automatically selects a highly parsimonious $\hat{G}$. This is in sharp contrast to the notorious difficulties with maximum likelihood for finite-dimensional mixture models, or with Gaussian deconvolution employing Fourier methods.

Having seen that the upper bound on the complexity of the NPMLE $\hat{G}$ was only $\mathcal{O}(\log n)$, one might wonder whether $\mathcal{O}(\log n)$ mixtures are "complex enough" to adequately represent the process that generated our observed data. Polyanskiy and Wu (2020) also addressed this concern: they noted that for any sub-Gaussian G, there exists a discrete distribution, G_k, with $k = \mathcal{O}(\log n)$ atoms, such that for $f_k = \varphi * G_k$, the total variation distance $TV(f, f_k) = o(1/n)$, and consequently there is no statistical justification for considering estimators of G whose complexity grows more rapidly than $\mathcal{O}(\log n)$. This observation is related to recent literature on generative adversarial networks, for example, Athey, Imbens, Metzger, and Munro (2019), that target models and estimators that, when simulated, successfully mimic observed data.

Other nonparametric maximum likelihood estimators for G are potentially also of interest. Efron (2016) has proposed an elegant log-spline sieve approach that yields smooth estimates of G; this has advantages especially from an inferential perspective, at the cost of reintroducing the task of selecting tuning parameters. An early proposal of Laird and Louis (1991) merged parametric empirical Bayes estimation of G with an EM step that pulled the parametric estimate back toward the NPMLE.

Given an estimate $\hat{G}$, it is straightforward to compute posterior distributions for each sample observation, or for that matter, for out-of-sample observations. In effect, we have estimated the prior, as in the Robbins (1951) binary means problem, but we have ignored the variability of $\hat{G}$ when we adopt plug-in procedures that use it. This may account for

the improved performance of *smoothed* estimates of G in certain inferential problems, as conjectured in Koenker (2020). In the sequel, we will compare ranking and selection procedures based on various functionals of these posterior distributions. A leading example is the posterior mean, but ranking and selection problems suggest other functionals of potential interest. If we are asked to estimate the θ_i's subject to quadratic loss, and assuming standard Gaussian noise, the Bayes rule is given by the posterior mean,

$$\delta(y) = \mathbb{E}(\theta|y) = y + f'(y)/f(y). \tag{2.1}$$

Efron (2011) referred to this as Tweedie's formula; it appears in Robbins (1956) credited to M.C.K. Tweedie. Appendix A of Gu and Koenker (2016a) provides an elementary derivation. The nonlinear shrinkage term takes a particularly simple affine form when G happens to be Gaussian, since in this case f is itself also Gaussian and the formula reduces to well-known linear shrinkage variants of classical Stein rules.

We have focused in this brief overview on compound decision problems for Gaussian location mixtures and posterior means; however, the NPMLE is adaptable to a wide variety of other mixture problems and other loss functions that imply other posterior functionals, as we will see in the next section. Efron (2019) and the discussion thereof offers a broader perspective on related methods. Implementation of several NPMLE options are described in Koenker and Gu (2017) and are available in the R package REBayes of Koenker and Gu (2015–2021).

3. HOMOGENEOUS VARIANCES

Suppose that you are given real-valued measurements, $y_i : i = 1, 2, \ldots, n$, of some attribute like test score performance for students or their teachers, survival rates for hospital surgical procedures, etc., and are told that the measurements are exchangeable and approximately Gaussian with unknown means θ_i and known variances σ_i^2 assumed provisionally to take the same value σ^2. Your task, should you decide to accept it, is to choose a group of size not to exceed αn of the elements with the largest θ_i's. One's first inclination might be to view each y_i as the maximum likelihood estimate for the corresponding θ_i, and select the αn largest observed values, but the compound decision framework suggests that it would be better to treat the problems as an ensemble. A second natural inclination might be to compute posterior means of the θ's with some linear or nonlinear shrinkage rule, rank them, and select the α best, but we will see that this, too, may be questionable.

3.1. *Posterior Tail Probability*

A natural alternative to ranking by the posterior means is to rank by posterior tail probabilities. Let $\theta_\alpha = G^{-1}(1-\alpha)$, and define $v_\alpha(y) := \mathbb{P}(\theta \geq \theta_\alpha | Y = y)$; then ranking by posterior tail probability gives the decision rule

$$\delta(y) = \mathbb{1}\{v_\alpha(y) \geq \lambda_\alpha\},$$

where the threshold λ_α is chosen so that $\mathbb{P}(v_\alpha(Y) \geq \lambda_\alpha) = \alpha$. This ranking criterion has been proposed by Henderson and Newton (2016) motivated as a ranking device for a fixed quantile level α. It can be interpreted in multiple testing terms: $1 - v_\alpha(y)$ is the local false discovery rate of Efron, Tibshirani, Storey, and Tusher (2001) and Storey (2002), for

testing the hypothesis $H_0 : \theta < \theta_\alpha$ versus $H_A : \theta \geq \theta_\alpha$. To see this, let h_i be a binary random variable $h_i = \mathbb{1}\{\theta_i \geq \theta_\alpha\}$; the loss function for observation i is

$$L(\delta_i, \theta_i) = \lambda \mathbb{1}\{h_i = 0, \delta_i = 1) + \mathbb{1}\{h_i = 1, \delta_i = 0\}$$

for a generic Lagrange multiplier, λ. The compound Bayes risk is

$$\mathbb{E}\left[\sum_{i=1}^{n} L(\delta_i, \theta_i)\right] = n\left[\alpha + \int \delta(y)\big[(1-\alpha)\lambda f_0(y) - \alpha f_1(y)\big]\, dy\right],$$

where $f_0(y) = (1-\alpha)^{-1} \int_{-\infty}^{\theta_\alpha} \varphi(y|\theta, \sigma^2)\, dG(\theta)$ and $f_1(y) = \alpha^{-1} \int_{\theta_\alpha}^{+\infty} \varphi(y|\theta, \sigma^2)\, dG(\theta)$, $\varphi(y|\theta, \sigma^2) = \varphi((y-\theta)/\sigma)/\sigma$. The Bayes rule for a fixed λ is

$$\delta(y_i) = \mathbb{1}\left\{v_\alpha(y_i) \geq \frac{\lambda}{1+\lambda}\right\},$$

where $v_\alpha(y) = \alpha f_1(y)/f(y) = \mathbb{P}(\theta \geq \theta_\alpha | Y = y)$ and $f(y) = (1-\alpha) f_0(y) + \alpha f_1(y)$. Provided that $v_\alpha(y)$ is monotone in y, a unique λ^* can be found such that $\mathbb{P}(\delta(Y) = 1) = \mathbb{P}(v_\alpha(Y) \geq \lambda^*/(1+\lambda^*)) = \alpha$.

LEMMA 3.1: *For fixed α, assuming $\mathbb{E}_{\theta|Y}[\nabla_y \log \varphi(y|\theta, \sigma^2)|Y] < \infty$, $v_\alpha(y)$ is monotone in y, and the sets $\Omega_\alpha := \{Y : v_\alpha(Y) \geq \lambda_\alpha/(1+\lambda_\alpha)\}$ have a nested structure, that is, if $\alpha_1 > \alpha_2$, then $\Omega_{\alpha_2} \subseteq \Omega_{\alpha_1}$.*

Any implementation of such a Bayes rule requires an estimate of the mixing distribution, G, or something essentially equivalent that would enable us to compute the local false discovery rates $v_\alpha(y)$ and the cut-off θ_α. The NPMLE, or perhaps a smoothed version of it, will provide a natural $\hat{G}$ for this task.

3.2. *Posterior Tail Expectation and Other Losses*

Rather than assessing loss by simply counting misclassifications, we might consider weighting such misclassifications by the magnitude of θ, for example,

$$L(\delta_i, \theta_i) = \sum_{i=1}^{n} (1 - \delta_i) \mathbb{1}(\theta_i \geq \theta_\alpha) \theta_i.$$

This presumes, of course, that we have centered the distribution G in some reasonable way, perhaps by forcing the mean or median to be zero. Minimizing with respect to δ subject to the constraint that $\mathbb{P}(\delta(Y) = 1) = \alpha$ leads to the Lagrangian

$$\min_{\delta} \int \int \big(1 - \delta(y)\big) \mathbb{1}\{\theta \geq \theta_\alpha\} \theta \varphi\big(y|\theta, \sigma^2\big)\, dG(\theta)\, dy$$
$$+ \lambda\left[\int \int \delta(y) \varphi\big(y|\theta, \sigma^2\big)\, dG(\theta)\, dy - \alpha\right],$$

which is equivalent to

$$\min_{\delta} \int \int \mathbb{1}\{\theta \geq \theta_\alpha\}(\theta - \lambda)\varphi\big(y|\theta, \sigma^2\big)\, dG(\theta)\, dy$$
$$- \int \delta(y)\Big[\int \mathbb{1}\{\theta \geq \theta_\alpha\}(\theta - \lambda)\varphi\big(y|\theta, \sigma^2\big)\, dG(\theta) - \int \lambda \mathbb{1}\{\theta < \theta_\alpha\}\varphi\big(y|\theta, \sigma^2\big)\, dG(\theta)\Big]\, dy.$$

Ignoring the first term since it does not depend upon δ, the oracle Bayes rule becomes: choose $\delta(y) = 1$ if

$$\frac{\int \mathbb{1}\{\theta \geq \theta_\alpha\}\theta\varphi\big(y|\theta, \sigma^2\big)\, dG(\theta)}{\int \varphi\big(y|\theta, \sigma^2\big)\, dG(\theta)} \geq \lambda,$$

with λ chosen so that $\mathbb{P}(\delta(Y) = 1) = \alpha$. Such criteria are closely related to expected shortfall criteria appearing in the literature on risk assessment. Again, the NPMLE can be employed to construct feasible posterior ranking criteria.

Several other loss functions were considered by Lin, Louis, Paddock, and Ridgeway (2006), including some based on global alignment of the ranks. While intuitively appealing, such loss functions are considerably less tractable than those we consider in the remainder of the paper.

3.3. *False Discovery and the α-Level*

Although our loss functions yield distinct criteria for ranking, their decision rules lead to the same selections when the precision of the measurements is homogeneous. When variances are homogeneous, there is a global cut-off, η_α, and a decision rule, $\delta_\alpha(Y) = \mathbb{1}(Y \geq \eta_\alpha)$, determining a common selection for all decision rules.

LEMMA 3.2: *For fixed α and homogeneous variance, posterior mean, posterior tail probability and posterior tail expectation all yield the same ranking and therefore the same selection.*

The marginal false discovery rate for selection in our Gaussian mixture setting is

$$\text{mFDR} = \mathbb{P}\big(\theta < \theta_\alpha | \delta_\alpha(Y) = 1\big) = \alpha^{-1} \int_{-\infty}^{\theta_\alpha} \Phi\big((\theta - \eta_\alpha)/\sigma\big)\, dG(\theta).$$

The marginal false non-discovery rate is

$$\text{mFNR} = \mathbb{P}\big(\theta \geq \theta_\alpha | \delta_\alpha(Y) = 0\big) = (1 - \alpha)^{-1} \int_{\theta_\alpha}^{\infty} \Phi\big((\eta_\alpha - \theta)/\sigma\big)\, dG(\theta).$$

Figure 1 shows the false discovery rate and false non-discovery rate for a range of capacity constraints, α, when the mixing distribution, G, is standard Gaussian and $\sigma^2 = 1$. In this low signal-to-noise ratio case, the cut-off value η_α is the $(1 - \alpha)$ quantile of $\mathcal{N}(0, 2)$, and it is very difficult to distinguish the meritorious from the merely lucky. For selecting individuals at the top α quantile, the false discovery rate is alarmingly high especially for smaller α, implying that the selected set may consist of a very high proportion of false

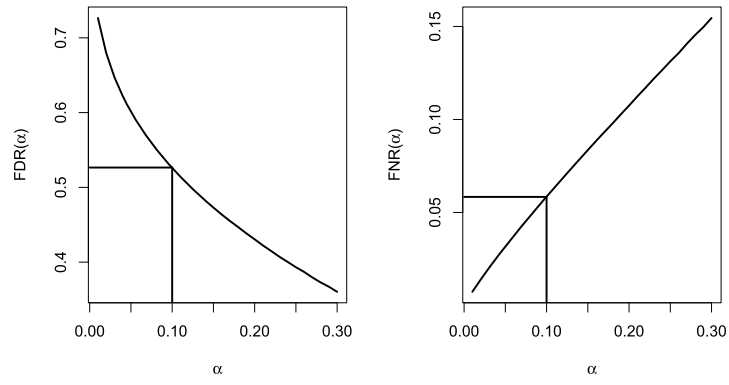

FIGURE 1.—False discovery rates and false non-discovery rates for a standard Gaussian mixing distribution.

discoveries. When $\alpha = 0.10$, the proportion of selected observations with θ below the threshold θ_α is slightly greater than 50 percent.

When the variance of the θ's, the signal-to-noise ratio, increases from 1 to 5, the selection problem becomes somewhat easier. This is reflected not only by the false discovery rate decreasing from above one half to about a third, but is also reflected in results in Table I that show that the wrongly selected individuals have their true values of θ clustered closer to the thresholding value θ_α measured in terms of standard deviation.

When $\sigma^2 = 1$, we have about 50% falsely selected, and the θ's among the correctly-selected and the wrongly-selected individuals are roughly symmetrically distributed around the thresholding value. In such a case, even oracle decision rules that incorporate complete knowledge of the precise distributional features of the problem may not be able to achieve better than about even odds that selected individuals have latent ability above the selection thresholds when measurement error is comparable in magnitude to Gaussian variability in latent ability. As variance of alpha increases to 5, then selection becomes somewhat easier; only 1/3 among the selected are falsely selected, and most (80%) of these θ values are within 0.5 standard deviations away from the selection threshold.

TABLE I

FDR IMPROVES AS THE SIGNAL BECOMES MORE DISPERSED. IN SETTINGS WITH STANDARD GAUSSIAN MEASUREMENT ERROR AND GAUSSIAN DISTRIBUTION, G, FOR THE θ'S, THE VARIANCE OF G CAN BE INTERPRETED AS A SIGNAL-TO-NOISE RATIO. AS THE VARIANCE OF G INCREASES, SELECTION BECOMES EASIER AND FDR IS REDUCED. THE WRONGLY SELECTED UNITS BECOME MORE CONCENTRATED NEAR THE SELECTION THRESHOLD. COLUMNS 2–6 OF THE TABLE REPORT QUANTILES OF THE WRONGLY SELECTED UNITS MEASURED IN STANDARD DEVIATIONS FROM THE THRESHOLD. COLUMNS 7–11 REPORT CORRESPONDING QUANTILES FOR THE CORRECTLY SELECTED UNITS.

	FDR	Wrongly Selected					Correctly Selected				
		20%	40%	50%	60%	80%	20%	40%	50%	60%	80%
$\sigma^2 = 1$	0.526	0.205	0.414	0.531	0.658	0.990	0.188	0.391	0.504	0.632	0.966
$\sigma^2 = 2$	0.421	0.146	0.302	0.391	0.486	0.743	0.197	0.394	0.503	0.625	0.951
$\sigma^2 = 3$	0.361	0.120	0.248	0.320	0.402	0.620	0.197	0.393	0.500	0.618	0.931
$\sigma^2 = 4$	0.319	0.103	0.212	0.275	0.347	0.540	0.198	0.392	0.495	0.609	0.915
$\sigma^2 = 5$	0.296	0.093	0.192	0.247	0.313	0.487	0.196	0.383	0.484	0.596	0.897

It is perhaps worth stressing that at the margin, near the decision boundary, it will always be difficult to distinguish true from false discoveries, but FDR measures the proportion of all selections that are incorrect, not just those near the threshold. Other loss functions that penalize in a more continuous way may be considered to reflect information in Table I. For example, losses that weight the classification error by the magnitude of the discrepancy between the latent effect and the threshold could be considered. Such losses, however, make it more difficult to incorporate conventional forms of error control.

Thus far, we have implicitly assumed that the size of the selected set is predetermined by the parameter α. Having established a ranking based on a particular loss function, we simply select a subset of size $\lceil \alpha n \rceil$ consisting of the highest ranked observations. In the next subsection, we begin to consider modifying this strategy by constraining the probability of false discoveries. This will allow the size of the selected set to adapt to the difficulty of the selection task.

3.4. *Guarding Against False Discovery*

Recognizing the risk of false "discoveries" among those selected, we will consider an expanded loss function,

$$L(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} h_i(1-\delta_i) + \tau_1\left(\sum_{i=1}^{n}\{(1-h_i)\delta_i - \gamma\delta_i\}\right) + \tau_2\left(\sum_{i=1}^{n}\delta_i - \alpha n\right), \tag{3.1}$$

where $h_i = \mathbb{1}\{\theta_i \geq \theta_\alpha\}$. If we set τ_1 to zero, then minimizing the expected loss leads to the Bayes rule discussed in Section 3.1. On the other hand, if we set τ_2 to zero, then minimizing expected loss leads to a decision rule that is equivalent to a multiple testing problem with null hypothesis $H_{0i} : \theta_i \leq \theta_\alpha$; the goal is to minimize the expected number of overlooked discoveries subject to the constraint that the marginal FDR rate is controlled at level γ, that is, $\mathbb{E}[\sum_{i=1}^{n}(1-h_i)\delta_i]/\mathbb{E}[\sum_{i=1}^{n}\delta_i] \leq \gamma$. When $\tau_1 = 0$, the risk can be expressed as

$$\mathbb{E}_{\boldsymbol{\theta}|Y}\big[L(\boldsymbol{\delta}, \boldsymbol{\theta})\big] = \sum_{i=1}^{n}(1-\delta_i)v_\alpha(Y_i) + \tau_2\left(\sum_{i=1}^{n}\delta_i - \alpha n\right),$$

where $v_\alpha(y_i) = \mathbb{P}(\theta_i \geq \theta_\alpha | Y_i = y_i)$. Taking another expectation over Y and minimizing over both $\boldsymbol{\delta}$ and τ_2 leads to the decision rule

$$\delta_i^* = \begin{cases} 1 & \text{if } v_\alpha(y_i) \geq \tau_2^*, \\ 0 & \text{if } v_\alpha(y_i) < \tau_2^*. \end{cases}$$

The Lagrange multiplier is chosen so that the constraint $\mathbb{P}(\delta_i = 1) \leq \alpha$ holds with equality:

$$\tau_2^* = \min\big\{\tau_2 : \mathbb{P}\big(v_\alpha(y_i) \geq \tau_2\big) \leq \alpha\big\}.$$

Each selection improves the objective function by $v_\alpha(y_i)$, but incurs a cost of τ_2. Since all selections incur the same cost, we may rank according to $v_\alpha(y_i)$, selecting units until the capacity constraint αn is achieved. Selection of the last unit may need to be randomized to exactly satisfy the constraint, as we note below.

When $\tau_2 = 0$, the focus shifts to the marginal FDR, the ratio of the expected number of false discoveries to the expected number of selections. This is slightly different from the

original FDR as defined in Benjamini and Hochberg (1995). However, when n is large, the two concepts are asymptotically equivalent as shown by Genovese and Wasserman (2002). Our objective becomes

$$\mathbb{E}_{\boldsymbol{\theta}|Y}\big[L(\boldsymbol{\delta},\boldsymbol{\theta})\big] = \sum_{i=1}^{n}(1-\delta_i)v_\alpha(Y_i) + \tau_1\left(\sum_{i=1}^{n}\big\{\delta_i\big(1-v_\alpha(Y_i)\big)-\gamma\delta_i\big\}\right).$$

Taking expectations again over Y and minimizing over both $\boldsymbol{\delta}$ and τ_1 yields

$$\delta_i^* = \begin{cases} 1 & \text{if } v_\alpha(y_i) > \tau_1^*\big(1-v_\alpha(y_i)-\gamma\big), \\ 0 & \text{if } v_\alpha(y_i) \le \tau_1^*\big(1-v_\alpha(y_i)-\gamma\big), \end{cases}$$

and the Lagrange multiplier takes a value τ_1^* to make the marginal FDR constraint hold with equality.

When both constraints are incorporated, we must balance the power gain from more selections and the cost that occurs from both the capacity constraint and FDR control. The Bayes rule solves

$$\min_{\boldsymbol{\delta}} \mathbb{E}\left[\sum_{i=1}^{n}(1-\delta_i)v_\alpha(y_i)\right] + \tau_1\left(\mathbb{E}\left[\sum_{i=1}^{n}\big\{\big(1-v_\alpha(y_i)\big)\delta_i-\gamma\delta_i\big\}\right]\right) + \tau_2\left(\mathbb{E}\left[\sum_{i=1}^{n}\delta_i\right]-\alpha n\right).$$

Given the discrete nature of the decision function, this problem appears to take the form of a classical knapsack problem; however, following the approach of Basu, Cai, Das, and Sun (2018), we will consider a relaxed version of the problem in which units are selected sequentially until one or the other constraint would be violated, with the final selection randomized to satisfy the constraint exactly.

EXAMPLE: Given the Lagrangian form of our loss function, it is natural to consider an optimization perspective for the selection problem. Minimizing the expectation of the loss defined in (3.1) is equivalent to minimizing $\mathbb{P}[\delta_i = 0, \theta_i \ge \theta_\alpha]$ subject to the constraint that $\mathbb{P}[\delta_i = 1, \theta_i < \theta_\alpha]/\mathbb{P}[\delta_i = 1] \le \gamma$, and $\mathbb{P}[\delta_i = 1] \le \alpha$. So we are looking for a thresholding rule that minimizes the expected number of missed discoveries subject to the capacity constraint and the constraint that the marginal FDR rate of the decision rule is below level γ. This minimization problem is also easily seen, from a testing perspective, to be equivalent to maximizing power of the decision rule δ, $\mathbb{P}[\delta_i = 1|\theta_i \ge \theta_\alpha]$, subject to the same two constraints.

PROPOSITION 3.3: *For any pair* (α,γ) *such that* $\gamma < 1-\alpha$, *the optimal Bayes rule takes the form* $\delta_i^* = \mathbb{1}\{v_\alpha(y_i) \ge \lambda^*(\alpha,\gamma)\}$, *where* $\lambda^*(\alpha,\gamma) = v_\alpha(t^*)$ *with* $t^* = \max\{t_1^*, t_2^*\}$,

$$t_1^* = \min\left\{ t : \frac{\displaystyle\int_{-\infty}^{\theta_\alpha} \tilde{\Phi}\big((t-\theta)/\sigma\big)\,dG(\theta)}{\displaystyle\int_{-\infty}^{+\infty} \tilde{\Phi}\big((t-\theta)/\sigma\big)\,dG(\theta)} - \gamma \le 0 \right\},$$

$$t_2^* = \min\left\{ t : \int_{-\infty}^{+\infty} \tilde{\Phi}\big((t-\theta)/\sigma\big)\,dG(\theta) - \alpha \le 0 \right\},$$

and $\tilde{\Phi}$ *denoting the survival function of a standard normal random variable*.

REMARK: The optimal cutoff t^* depends on the data generating process and also the choice of α and γ. When data are noisy, the FDR control constraint may be binding before the capacity constraint is reached, and consequently the selected set may be strictly smaller than the pre-specified α proportion. On the other hand, when the signal is strong, the FDR control constraint is unlikely to bind before the capacity constraint is reached.

We have seen that when variances are homogeneous, the optimal selection rule thresholds on Y, so it is clear then that any ranking that is based on a monotone transformation of Y will lead to an equivalent selected set. We should also stress that we have focused on a null hypothesis that depends on α, while the multiple testing literature, for example Efron et al. (2001), Sun and Tony Cai (2007), and Basu et al. (2018), typically focuses on the null hypothesis of $H_{0i} : \theta_i = 0$. When variances are homogeneous, it does not matter whether we use an α dependent null or the conventional zero null, because the transformation based on the conventional null, $\mathbb{P}(\theta > 0|Y = y)$, is also a monotone function of Y, and therefore yields an equivalent decision rule. However, when variances are heterogeneous, this invariance no longer holds; different transformations of the pair (y, σ) lead to distinct decision rules that lead to distinct performance, and using the conventional null hypothesis is no longer advisable for the ranking and selection problem, as we will show in the next section.

4. HETEROGENEOUS KNOWN VARIANCES

The homogeneous variance assumption of the preceding section is unsustainable in most applications. Batting averages are accompanied by a number of "at bats" and mean test score performances are accompanied by student sample sizes. In this section, we will consider the expanded model,

$$Y_i \sim \mathcal{N}\big(\theta_i, \sigma_i^2\big) \quad \text{and} \quad \theta_i \sim G, \qquad \sigma_i \sim H, \qquad \sigma_i \perp\!\!\!\perp \theta_i.$$

We will assume that we observe σ_i, an assumption that will be relaxed in the next section.

4.1. *Posterior Tail Probability*

With the same alternative hypothesis as $H_A : \theta \geq \theta_\alpha$, it is natural to consider the posterior tail probability again, now as a function of the pair (y_i, σ_i):

$$v_\alpha(y_i, \sigma_i) = \mathbb{P}(\theta_i \geq \theta_\alpha | y_i, \sigma_i) = \frac{\int_{\theta_\alpha}^{+\infty} \varphi\big(y_i|\theta, \sigma_i^2\big)\, dG(\theta)}{\int_{-\infty}^{+\infty} \varphi\big(y_i|\theta, \sigma_i^2\big)\, dG(\theta)}.$$

Solving the same decision problem with the loss function specified in (3.1), we have the conditional risk,

$$\mathbb{E}_{\boldsymbol{\theta}|Y,\boldsymbol{\sigma}}\big[L(\boldsymbol{\delta}, \boldsymbol{\theta})\big] = \sum_{i=1}^{n}(1 - \delta_i)v_\alpha(Y_i, \sigma_i) + \tau_1\left(\sum_{i=1}^{n}\big\{\delta_i\big(1 - v_\alpha(Y_i, \sigma_i)\big) - \gamma\delta_i\big\}\right) + \tau_2\left(\sum_{i=1}^{n}\delta_i - \alpha n\right).$$

Taking another expectation with respect to the joint distribution of the (Y_i, σ_i)'s, the Bayes rule solves

$$\min_{\boldsymbol{\delta}} \mathbb{E}\left[\sum_{i=1}^{n}(1-\delta_i)v_\alpha(y_i,\sigma_i)\right] + \tau_1\left(\mathbb{E}\left[\sum_{i=1}^{n}\{(1-v_\alpha(y_i,\sigma_i))\delta_i - \gamma\delta_i\}\right]\right) + \tau_2\left(\mathbb{E}\left[\sum_{i=1}^{n}\delta_i\right] - \alpha n\right).$$

The optimal selection rule can again be characterized as a thresholding rule on $v_\alpha(y_i, \sigma_i)$.

PROPOSITION 4.1: *For a pre-specified pair (α, γ) such that $\gamma < 1-\alpha$, the Bayes rule takes the form $\delta^*(y,\sigma) = \mathbb{1}\{v_\alpha(y,\sigma) \geq \lambda^*(\alpha,\gamma)\}$, where $\lambda^*(\alpha,\gamma) = \max\{\lambda_1^*(\alpha,\gamma), \lambda_2^*(\alpha)\}$,*

$$\lambda_1^*(\alpha,\gamma) = \min\left\{\lambda : \frac{\int\int_{-\infty}^{\theta_\alpha}\tilde{\Phi}\big((t_\alpha(\lambda,\sigma)-\theta)/\sigma\big)\,dG(\theta)\,dH(\sigma)}{\int\int_{-\infty}^{+\infty}\tilde{\Phi}\big((t_\alpha(\lambda,\sigma)-\theta)/\sigma\big)\,dG(\theta)\,dH(\sigma)} - \gamma \leq 0\right\},$$

$$\lambda_2^*(\alpha) = \min\left\{\lambda : \int\int_{-\infty}^{+\infty}\tilde{\Phi}\big((t_\alpha(\lambda,\sigma)-\theta)/\sigma\big)\,dG(\theta)\,dH(\sigma) - \alpha \leq 0\right\},$$

and $t_\alpha(\lambda,\sigma)$ is defined as $v_\alpha(t_\alpha(\lambda,\sigma),\sigma) = \lambda$ for all $\lambda \in [0,1]$.

REMARK: Note that although the thresholding value λ^* does not depend on the value of σ, the ranking does depend on σ. One way to see this is that since $v_\alpha(y,\sigma)$ is monotone in y for all $\sigma > 0$, the optimal rule is equivalent to $\mathbb{1}\{y_i > t_\alpha(\lambda^*,\sigma)\}$, where $t_\alpha(\lambda,\sigma)$ is a function of σ. For a fixed value of λ^*, the selection region for Y depends on σ in a nonlinear way. Comparing individuals i and j, it may be the case that $y_i > y_j$, but y_j belongs to the selection region while y_i does not. An example to illustrate this appears below. It should also be emphasized that when variances are heterogeneous, different loss functions need not lead to equivalent selections.

4.2. *The Conventional Null Hypothesis*

The posterior tail probability criterion is motivated by viewing the ranking and selection problems as hypothesis testing while allowing the null hypothesis to be α dependent. The particular construction of the null hypothesis turns out to be critical for the ranking exercise. In this subsection, we present a simple example to illustrate that tail probability based on the conventional null hypothesis of zero effect does not lead to a powerful ranking device. Consider data generated from a three-component normal mixture model,

$$Y_i|\sigma_i \sim 0.85\mathcal{N}(-1,\sigma_i^2) + 0.1\mathcal{N}(0.5,\sigma_i^2) + 0.05\mathcal{N}(5,\sigma_i^2), \quad \sigma_i \sim U[0.5,4]. \tag{4.1}$$

Instead of transforming the data by v_α, we consider the transformation

$$\mathrm{T}(y_i,\sigma_i^2) = \mathbb{P}(\theta_i > 0|y_i,\sigma_i^2) = \frac{\int_0^{+\infty}\varphi(y_i|\theta,\sigma_i^2)\,dG(\theta)}{\int_{-\infty}^{+\infty}\varphi(y_i|\theta,\sigma_i^2)\,dG(\theta)}$$

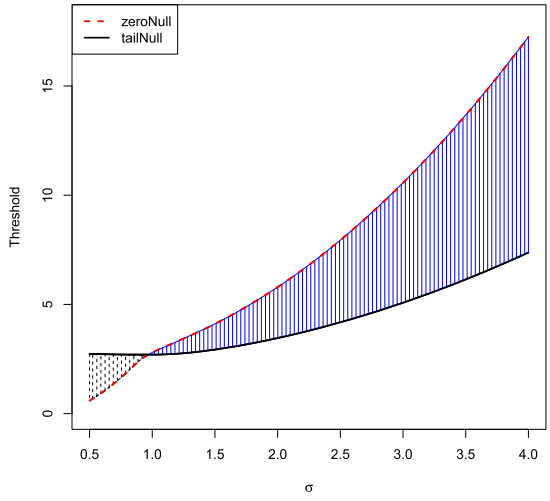

FIGURE 2.—Selection boundaries based on the model (4.1) with $\alpha = 0.05$ and $\gamma = 0.1$. The solid black curve corresponds to the boundary of the selection region based on transformation v_α. The dashed red curve corresponds to the boundary of the selection region based on transformation T. Density of σ is assumed to be uniform on the interval [0.5, 4].

and rank individuals accordingly. This transformation corresponds to the procedure proposed in Sun and McLain (2012), and is motivated for multiple testing problems under the conventional null hypothesis $H_0 : \theta \leq 0$. The decision rule $\delta_i^{\mathrm{T}} = \mathbb{1}\{\mathrm{T}(y_i, \sigma_i) \geq \lambda\}$ then chooses the cutoff value λ that respects both the capacity constraint and the FDR control constraint for selecting the top α proportion.

Figure 2 compares the selection region for the two ranking procedures with $\alpha = 5\%$ and marginal FDR control at level 10%. The solid black line corresponds to the selection boundary using ranking based on transformation v_α and the dashed red line corresponds to the selection boundary using ranking based on the transformation T. The black highlighted area below the black selection boundary corresponds to a region where the ranking method based on T will select but the ranking method based on v_α does not. On the other hand, the blue highlighted area corresponds to a region selected by v_α, but not for T. The transformation T ranks those in the black region higher than those in the blue region because although they have a relatively smaller mean effect y, their associated variances are also smaller, indicating stronger evidence that such individuals have a positive θ than those located in the blue area. However, our task is to find individuals with true effects, θ_i, in the upper tail. For $\alpha = 5\%$, we aim to select all individuals with $\theta = 5$; individuals in the black region present strong evidence that their true effect *cannot* be too large because their observed effect y is small and their associated variance is also small, while those in the blue region, although their observed mean effects are associated with larger variances, offer reasonable evidence that their associated true effect θ may be large. This evidence is not apparent in transformation T, but is captured in the transformation v_α.

Indeed, the average power of ranking based on the two different transformations v_α and T differs significantly. Defining the power of the selection rule as $\beta(\delta) := \mathbb{P}(\theta_i \geq \theta_\alpha, \delta_i = 1)/\mathbb{P}(\theta_i \geq \theta_\alpha)$, the proportion of true top α cases selected based on decision rule $\boldsymbol{\delta}$, then $\beta(\boldsymbol{\delta}^{\mathbf{T}}) = 39\%$ and $\beta(\boldsymbol{\delta}^*) = 69\%$. Thus, although much of the literature relies on ranking and selection rules based on some form of posterior means and conventional

hypothesis testing apparatus, we would caution that such methods can be quite misleading and inefficient.

4.3. *Nestedness of Selection Sets*

If we were to relax the capacity constraint to allow a larger proportion, $\alpha_1 > \alpha_0$, to be selected, while maintaining our initial false discovery control, we would expect that members selected under the more stringent capacity constraint should remain selected under the relaxed constraint. We now discuss sufficient conditions under which we obtain this nestedness of the selection sets when using the posterior tail probability rule. This is a natural condition in applications like our analysis of ranking and selection of dialysis centers especially because we would like to assign "letter grades" to several subgroups of the centers.

The optimal Bayes rule defines the selection set for each pair of (α, γ) as

$$\Omega_{\alpha,\gamma} := \big\{(y,\sigma) : v_\alpha(y,\sigma) \geq \lambda^*(\alpha,\gamma)\big\},$$

and when σ is known, $v_\alpha(y,\sigma)$ is monotone in y as shown in Lemma 3.2 for each fixed σ; hence the selection set can also be represented as

$$\Omega_{\alpha,\gamma} = \big\{(y,\sigma) : y \geq t_\alpha\big(\lambda^*(\alpha,\gamma),\sigma\big)\big\}.$$

It is also convenient for later discussion to define

$$\Omega^{\mathrm{FDR}}_{\alpha,\gamma} := \big\{(y,\sigma) : v_\alpha(y,\sigma) \geq \lambda^*_1(\alpha,\gamma)\big\} = \big\{(y,\sigma) : y \geq t_\alpha\big(\lambda^*_1(\alpha,\gamma),\sigma\big)\big\},$$
$$\Omega^{C}_{\alpha} := \big\{(y,\sigma) : v_\alpha(y,\sigma) \geq \lambda^*_2(\alpha)\big\} = \big\{(y,\sigma) : y \geq t_\alpha\big(\lambda^*_2(\alpha),\sigma\big)\big\},$$

which are respectively the selection sets when the false discovery rate constraint or the capacity constraint is binding. It is easy to see that $\Omega_{\alpha,\gamma} = \Omega^{\mathrm{FDR}}_{\alpha,\gamma} \cap \Omega^{C}_{\alpha}$.

LEMMA 4.2: *Let the density function of $v_\alpha(y_i,\sigma_i)$ be denoted as $f_v(v;\alpha)$, and let*

$$\lambda^*_1(\alpha,\gamma) = \min\left\{\lambda : \frac{\displaystyle\int\int_{-\infty}^{\theta_\alpha} \tilde{\Phi}\big(\big(t_\alpha(\lambda,\sigma)-\theta\big)/\sigma\big)\,dG(\theta)\,dH(\sigma)}{\displaystyle\int\int_{-\infty}^{+\infty} \tilde{\Phi}\big(\big(t_\alpha(\lambda,\sigma)-\theta\big)/\sigma\big)\,dG(\theta)\,dH(\sigma)} - \gamma \leq 0\right\}$$

with $t_\alpha(\lambda,\sigma)$ defined as $v_\alpha(t_\alpha(\lambda,\sigma),\sigma) = \lambda$ and $\tilde{\Phi}$ be the survival function of the standard normal random variable. If $\nabla_\alpha \log f_v(v;\alpha)$ is non-decreasing in v, then for fixed γ, if $\alpha_1 > \alpha_2$, we have $\lambda^*_1(\alpha_1,\gamma) \leq \lambda^*_1(\alpha_2,\gamma)$.

REMARK: The density function $f_v(v;\alpha)$ can be viewed as a function of v indexed by the parameter α. An explicit form for $f_v(v;\alpha)$ appears in Section 4.4 for the normal-normal model. The condition imposed in Lemma 4.2 is equivalent to a monotone likelihood ratio condition, that is, that the likelihood ratio $f_v(v;\alpha_1)/f_v(v;\alpha_2)$ is non-decreasing in v if $\alpha_1 > \alpha_2$.

COROLLARY 4.3: *If the condition in Lemma 4.2 holds, then $\Omega^{\mathrm{FDR}}_{\alpha_2,\gamma} \subseteq \Omega^{\mathrm{FDR}}_{\alpha_1,\gamma}$ for any $\alpha_1 > \alpha_2$.*

REMARK: The condition in Lemma 4.2 is sufficient but not necessary for nestedness of $\Omega^{\mathrm{FDR}}_{\alpha,\gamma}$. Even when $\lambda_1^*(\alpha_1,\gamma) > \lambda_1^*(\alpha_2,\gamma)$, we can still have $t_{\alpha_1}(\lambda_1^*(\alpha_1,\gamma),\sigma) < t_{\alpha_2}(\lambda_1^*(\alpha_2,\gamma),\sigma)$ because the function $v_\alpha(Y,\sigma)$ depends on α, as does its inverse function t_α.

LEMMA 4.4: *Let $\lambda_2^*(\alpha)$ be defined as in Proposition 4.1. If, for any $\alpha_1 > \alpha_2$, $t_{\alpha_1}(\lambda_2^*(\alpha_1),\sigma) \le t_{\alpha_2}(\lambda_2^*(\alpha_2),\sigma)$ for each σ, then $\Omega^C_{\alpha_2} \subseteq \Omega^C_{\alpha_1}$.*

REMARK: The monotonicity here coincides with the condition in Theorem 3 of Henderson and Newton (2016), who demonstrated that it holds when G is Gaussian. However, it need not hold, as shown in our counterexample in Appendix C of the Supplemental Material.

LEMMA 4.5: *If $\nabla_\alpha \log f_v(v;\alpha)$ is non-decreasing in v and the condition in Lemma 4.4 holds, then for a fixed γ, the selection region has a nested structure: if $\alpha_1 > \alpha_2$, then $\Omega_{\alpha_2,\gamma} \subseteq \Omega_{\alpha_1,\gamma}$.*

4.4. *Examples*

In this section, we consider several examples beginning with the simplest classical case in which the θ_i constitute a random sample from the standard Gaussian distribution. This Gaussian assumption on the form of the mixing distribution G underlies almost all of the empirical Bayes literature in applied economics; it is precisely what justifies the linear shrinkage rules that are typically employed.

EXAMPLE—Gaussian G: Consider the normal-normal model, where $y|\theta,\sigma^2 \sim \mathcal{N}(\theta,\sigma^2)$ and $\theta \sim \mathcal{N}(0,\sigma_\theta^2)$ and $\sigma \sim H$ with density function $h(\sigma)$. The marginal distribution of y given σ^2 is $\mathcal{N}(0,\sigma^2+\sigma_\theta^2)$ and the joint density of (y,σ) takes the form

$$f(y,\sigma) = \frac{1}{\sqrt{2\pi(\sigma^2+\sigma_\theta^2)}} \exp\left\{-\frac{y^2}{2(\sigma^2+\sigma_\theta^2)}\right\} h(\sigma).$$

Given the normal conjugacy, the posterior distribution of $\theta|y,\sigma^2$ follows $\mathcal{N}(\rho y,\rho\sigma^2)$ where $\rho = \sigma_\theta^2/(\sigma_\theta^2+\sigma^2)$. The random variable v is thus a transformation of the pair (Y,σ^2), defined as

$$v = \psi(y,\sigma^2) := \mathbb{P}(\theta \ge \theta_\alpha | y,\sigma^2) = \Phi\big((\rho y - \theta_\alpha)/\sqrt{\rho\sigma^2}\big).$$

For fixed σ^2, ψ is monotone increasing in y and $\psi^{-1}(v) = \theta_\alpha/\rho + \sqrt{\sigma^2/\rho}\,\Phi^{-1}(v)$ with $\nabla_v\psi^{-1}(v) = \sqrt{\sigma^2/\rho}/\varphi(\Phi^{-1}(v))$. The joint density of v and σ is thus

$$\begin{aligned} g(v,\sigma) &= f\big(\psi^{-1}(v),\sigma\big)\big|\nabla_v\psi^{-1}(v)\big| \\ &= \frac{1}{\sqrt{2\pi(\sigma^2+\sigma_\theta^2)}} \exp\left\{-\frac{\big(\theta_\alpha/\rho + \sqrt{\sigma^2/\rho}\,\Phi^{-1}(v)\big)^2}{2(\sigma^2+\sigma_\theta^2)}\right\} \frac{\sqrt{\sigma^2/\rho}}{\varphi\big(\Phi^{-1}(v)\big)} h(\sigma). \end{aligned}$$

Integrating out σ, we have the marginal density of v:

$$f_v(v;\alpha)=\int \frac{1}{\sqrt{2\pi(\sigma^2+\sigma_\theta^2)}}\exp\left\{-\frac{\left(\theta_\alpha/\rho+\sqrt{\sigma^2/\rho}\Phi^{-1}(v)\right)^2}{2(\sigma^2+\sigma_\theta^2)}\right\}\frac{\sqrt{\sigma^2/\rho}}{\varphi(\Phi^{-1}(\tilde{v}))}\,dH(\sigma).$$

The capacity constraint is $\mathbb{P}(v\geq\lambda_2^*)=\alpha$, with cutoff value λ_2^* satisfying

$$\alpha=\mathbb{P}(v\geq\lambda_2^*)=1-\int\Phi\left(\theta_\alpha\frac{\sqrt{\sigma^2+\sigma_\theta^2}}{\sigma_\theta^2}-\Phi^{-1}(1-\lambda_2^*)\sqrt{\sigma^2/\sigma_\theta^2}\right)dH(\sigma).$$

To find λ_1^*, we can use the formula provided in Proposition 4.1. A more direct approach is to recognize (see Section 6) that the FDR control constraint is defined as $\gamma=\mathbb{E}[(1-v)\mathbb{1}\{v\geq\lambda_1^*\}]/\mathbb{P}(v\geq\lambda_1^*)$, where the cutoff value λ_1^* is defined through

$$\gamma=\int_{\lambda_1^*}^1(1-v)f_v(v;\alpha)\,dv/\int_{\lambda_1^*}^1 f_v(v;\alpha)\,dv.$$

Let $\lambda^*=\max\{\lambda_1^*,\lambda_2^*\}$; the selection region is then $\{(y,\sigma):y\geq t_\alpha(\lambda^*,\sigma)\}$ with

$$t_\alpha(\lambda^*,\sigma)=\theta_\alpha/\rho-\Phi^{-1}(1-\lambda^*)\sqrt{\sigma^2/\rho}.$$

Suppose we use the posterior mean of θ as a ranking device, so $\delta_i^{\text{PM}}=\mathbb{1}\{y\rho\geq\omega^*\}$ for some suitably chosen ω^* that guarantees both capacity and FDR control. For the capacity constraint, the thresholding value solves

$$\begin{aligned}1-\alpha&=\int\mathbb{P}(y\rho<\omega_2^*)\,dH(\sigma)\\&=\int\Phi\left(\omega_2^*/\left(\sigma_\theta^2\sqrt{\sigma_\theta^2+\sigma^2}\right)\right)dH(\sigma),\end{aligned}$$

while FDR control requires a thresholding value that solves

$$\begin{aligned}\gamma&=\int\mathbb{P}(y\geq\omega_1^*/\rho,\theta<\theta_\alpha)\,dH(\sigma)/\int\mathbb{P}(y\geq\omega_1^*/\rho)\,dH(\sigma)\\&=\int\int_{[\omega_1^*/\rho,+\infty)}(1-\alpha)f_0(y)\,dy\,dH(\sigma)/\int 1-\Phi(\omega_1^*/\left(\sigma_\theta^2\sqrt{\sigma_\theta^2+\sigma^2}\right)dH(\sigma),\end{aligned}$$

with

$$f_0(y)=\frac{1}{1-\alpha}\frac{1}{\sqrt{2\pi(\sigma_\theta^2+\sigma^2)}}\exp\left\{-\frac{y^2}{2(\sigma_\theta^2+\sigma^2)}\right\}\Phi\left(\frac{(\theta_\alpha-y\rho)}{\sqrt{\rho\sigma^2}}\right),$$

denoting the density of y under the null $\theta<\theta_\alpha$. Setting $\omega^*=\max\{\omega_1^*,\omega_2^*\}$, the selection region is then $\{(y,\sigma):y\geq\omega^*/\rho\}$.

Figure 3 plots the selection boundaries for both constraints with $\theta\sim\mathcal{N}(0,1)$ and $\sigma\sim U[0.5,1]$. With $\alpha=0.05$ and $\gamma=0.2$, the FDR constraint is binding, but not the

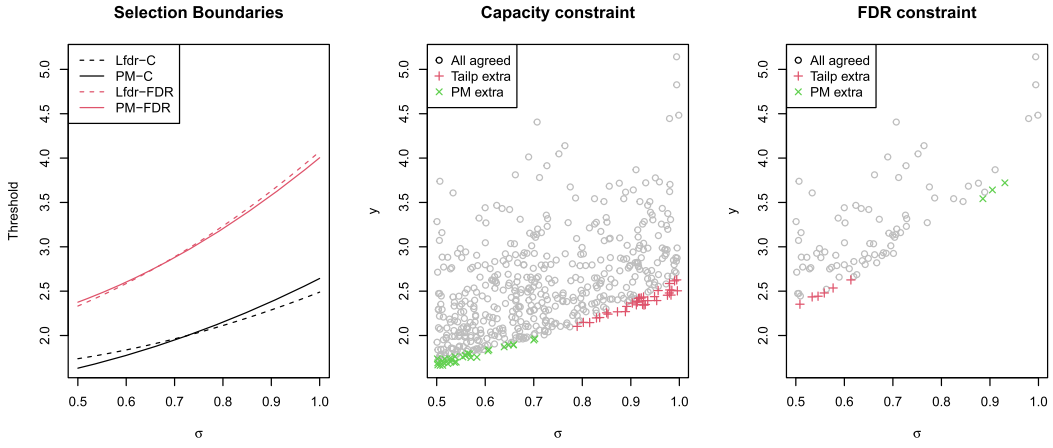

FIGURE 3.—The left panel plots the selection boundaries for the normal-normal model with $\sigma_\theta^2 = 1$ and $\alpha = 0.05$ and $\gamma = 0.2$. The density of σ is assumed to be uniform on the range $[0.5, 1]$. Selected units must have (y_i, σ_i) above the curves. The red curves correspond to the selection region boundaries with FDR controlled at level 0.2; solid lines for posterior mean ranking and dashed line for posterior tail probability ranking. The black curves correspond to the selection boundaries with capacity control at level 0.05. The middle and right panels illustrate the selected set from a realized sample of size 10,000. The gray circles correspond to individuals selected by both the posterior tail probability rule and the posterior mean rule. The green crosses depict individuals selected by the posterior mean rule but not the tail probability, and the red crosses indicate individuals selected by the tail probability rule but not by the posterior mean rule.

capacity constraint. In this example, if we only impose the capacity constraint to be 5 percent, even an oracle totally aware of G will face a false discovery rate of nearly 52 percent. In other words, more than half of those selected to be in the right tail will be individuals with $\theta < \theta_\alpha$ rather than from the intended $\theta \geq \theta_\alpha$ group. This fact motivates our more explicit incorporation of FDR into the selection constraints. We may recall that in the homogeneous variance Gaussian setting, we saw in Figure 1 that FDR was also very high when α is set at 0.05. Figure 3 also depicts the selected set with a realized sample of 10,000 from the normal-normal model. With capacity constraint alone, the posterior mean criterion favors individuals with smaller variances. When the FDR constraint is implemented, with $\gamma = 0.2$, it becomes the binding constraint in this setting, both criteria become more stringent and only a much smaller set of individuals are selected, and there is less conflict in the selections. The corresponding selected sets are plotted in the right panel of Figure 3. When the variance parameter σ_θ^2 in G is not observed, we can estimate it via the MLE based on the marginal likelihood of Y. This leads to a generalized James–Stein estimator of the type proposed in Efron and Morris (1973).

EXAMPLE—Discrete G: Suppose $\theta \sim 0.85\delta_{-1} + 0.1\delta_2 + 0.05\delta_5$. Then the marginal density of y given σ^2 takes the form

$$f(y|\sigma^2) = \int \varphi(y|\theta, \sigma^2)\, dG(\theta)$$
$$= \big(0.85\varphi(y|1, \sigma^2) + 0.1\varphi(y|2, \sigma^2) + 0.05\varphi(y|5, \sigma^2)\big).$$

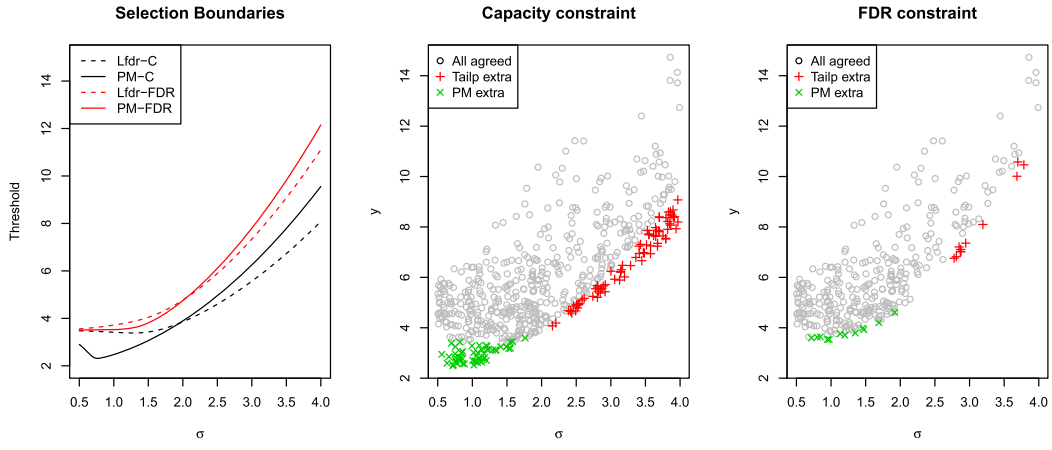

FIGURE 4.—The left panel plots the selection boundaries for the normal-discrete model with $\theta \sim G = 0.85\delta_{-1} + 0.1\delta_2 + 0.05\delta_5$ and $\alpha = 0.05$ and $\gamma = 0.2$. The density of σ is uniform on the range $[0.5, 4]$. The red curves correspond to the selection region with FDR controlled at level 0.2, solid lines for posterior mean ranking and dashed lines for posterior tail probability ranking. The black curves correspond to the selection region with capacity control at level 0.05. The other panels are structured as in the previous figure.

And the random variable v is a transformation of the pair (y, σ^2), defined as

$$v = \psi(y, \sigma^2) := \mathbb{P}(\theta \geq \theta_\alpha | y, \sigma^2) = \frac{\int_{\theta_\alpha}^{+\infty} \varphi(y|\theta, \sigma^2)\, dG(\theta)}{\int_{-\infty}^{\infty} \varphi(y|\theta, \sigma^2)\, dG(\theta)}.$$

The capacity constraint leads to a thresholding rule on v such that $\mathbb{P}(v \geq \lambda_2^*) = \alpha$, while the FDR control leads to a cutoff value λ_1^*, defined through $\gamma = \mathbb{E}[(1-v)\mathbb{1}\{v \geq \lambda_1^*\}]/\mathbb{P}(v \geq \lambda_1^*)$. Let $\lambda^* = \max\{\lambda_1^*, \lambda_2^*\}$; the selection region is then defined by $\{(y, \sigma) : y \geq t_\alpha(\lambda^*, \sigma)\}$, and can be found easily numerically.

Figure 4 plots the selection boundaries for both constraints when θ follows this discrete distribution. We again set $\alpha = 0.05$ and $\gamma = 0.2$, so we would like to select all the individuals associated with the largest effect size, $\{\theta = 5\}$, while controlling the FDR rate below 20%. The red curves again correspond to FDR control with the two ranking procedures, while the black curves correspond to capacity control. For the two regions to overlap with α fixed at 0.05, we must be willing to tolerate $\gamma \approx 37\%$. In this case, we see that the posterior probability ranking procedure prefers individuals with larger variances, while the posterior mean ranking procedure prefers smaller variances. Based on a realized sample of 10,000, Figure 4 again shows the selected observations, and once more we see that the posterior mean criterion favors individuals with smaller variances, under both the capacity constraint and the FDR constraint. In contrast to the normal-normal setting, now the FDR constraint is much less severe and allows us to select considerably more individuals.

5. HETEROGENEOUS UNKNOWN VARIANCES

Assuming that the σ_i's are known, up to a common scale parameter, may be plausible in some applications such as baseball batting averages, but it is frequently more plausible to adopt the view that we are simply confronted with estimates of scale available perhaps

from longitudinal data. In such cases, we need to consider the pairs, (y_i, S_i) as potentially jointly dependent random variables arising from the longitudinal model,

$$Y_{it} = \theta_i + \sigma_i \epsilon_{it}, \qquad \epsilon_{it} \sim_{iid} \mathcal{N}(0,1), \qquad (\theta_i, \sigma_i^2) \sim G,$$

with sufficient statistics, $Y_i = T_i^{-1} \sum_{t=1}^{T_i} Y_{it}$ and $S_i = (T_i - 1)^{-1} \sum_{t=1}^{T_i} (Y_{it} - Y_i)^2$, for (θ_i, σ_i^2). Conditional on (θ_i, σ_i^2), we have $Y_i | \theta_i, \sigma_i^2 \sim \mathcal{N}(\theta_i, \sigma_i^2/T_i)$ and $S_i | \sigma_i^2$ is distributed as Gamma with shape parameter $r_i = (T_i - 1)/2$, scale parameter, σ_i^2/r_i, and density function denoted as $\Gamma(S_i | r_i, \sigma_i^2/r_i)$.

Given the loss function (3.1) and defining θ_α as $\alpha = \mathbb{P}(\theta_i \geq \theta_\alpha) = \int \int_{\theta_\alpha}^{+\infty} dG(\theta, \sigma^2)$, the conditional risk is

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}|Y,S}\big[L(\boldsymbol{\delta}, \boldsymbol{\theta})\big] = &\sum_{i=1}^{n} (1 - \delta_i) v_\alpha(Y_i, S_i) \\ &+ \tau_1 \left(\sum_{i=1}^{n} \big\{ \delta_i \big(1 - v_\alpha(Y_i, S_i)\big) - \gamma \delta_i \big\} \right) + \tau_2 \left(\sum_{i=1}^{n} \delta_i - \alpha n \right),\end{aligned}$$

with

$$\begin{aligned} v_\alpha(y_i, s_i) &= \mathbb{P}(\theta_i \geq \theta_\alpha | Y_i = y_i, S_i = s_i) \\ &= \frac{\int \int_{\theta_\alpha}^{+\infty} \Gamma\big(s_i | r_i, \sigma_i^2/r_i\big) \varphi\big(y_i | \theta, \sigma^2/T_i\big)\, dG\big(\theta, \sigma^2\big)}{\int \int \Gamma\big(s_i | r_i, \sigma_i^2/r_i\big) \varphi\big(y_i | \theta, \sigma^2/T_i\big)\, dG\big(\theta, \sigma^2\big)}. \end{aligned}$$

Taking expectations with respect to (Y, S), the Bayes rule solves

$$\begin{aligned}\min_{\boldsymbol{\delta}} &\mathbb{E}\left[\sum_{i=1}^{n} (1 - \delta_i) v_\alpha(y_i, s_i) \right] + \tau_1 \left(\mathbb{E}\left[\sum_{i=1}^{n} \big\{ \big(1 - v_\alpha(y_i, s_i)\big) \delta_i - \gamma \delta_i \big\} \right] \right) \\ &+ \tau_2 \left(\mathbb{E}\left[\sum_{i=1}^{n} \delta_i \right] - \alpha n \right).\end{aligned}$$

Before characterizing the Bayes rule any further, we should observe that when variances σ^2 are not directly observed, the tail probability $v_\alpha(Y, S)$ may no longer have the monotonicity property we have described above.

LEMMA 5.1: *Consider the transformation $v_\alpha(Y, S) = \mathbb{P}(\theta \geq \theta_\alpha | Y, S]$; then for fixed $S = s$, the function $v_\alpha(Y, s)$ may not be monotone in Y; and for fixed $Y = y$, the function $v_\alpha(y, S)$ may not be monotone in S.*

PROPOSITION 5.2: *For pre-specified (α, γ) such that $\gamma < 1 - \alpha$, the Bayes selection rule takes the form*

$$\delta_i^* = \mathbb{1}\big\{ v_\alpha(Y, S) \geq \lambda^*(\alpha, \gamma) \big\},$$

where $\lambda^*(\alpha,\gamma)=\max\{\lambda_1^*(\alpha,\gamma),\lambda_2^*(\alpha)\}$ *with*

$$\lambda_1^*(\alpha,\gamma)=\min\big\{\lambda:\mathbb{E}\big[\big(1-v_\alpha(Y,S)-\gamma\big)\mathbb{1}\big\{v_\alpha(Y,S)\geq\lambda\big\}\big]\leq 0\big\}$$

and

$$\lambda_2^*(\alpha)=\min\big\{\lambda:\mathbb{P}\big(v_\alpha(Y,S)\geq\lambda\big)-\alpha\leq 0\big\}.$$

Based on the Bayes rule, the selected set is defined as

$$\Omega_{\alpha,\gamma}=\big\{(Y,S):v_\alpha(Y,S)\geq\lambda^*(\alpha,\gamma)\big\}.$$

REMARK: Note that for each pre-specified pair (α,γ), $\Omega_{\alpha,\gamma}$ is just the $\lambda^*(\alpha,\gamma)$-superlevel set of the function $v_\alpha(Y,S)$. For any $\alpha_1>\alpha_2$, nestedness of the selected sets would mean that the $\lambda^*(\alpha_2,\gamma)$-superlevel set of the function v_{α_2} must be a subset of the $\lambda^*(\alpha_1,\gamma)$-superlevel set of the function v_{α_1}. The construction and the form of the optimal selection rule may appear to be very similar to the case where σ_i^2 is observed. However, the crucial difference is that we no longer require the independence between θ and σ^2 in this section. In contrast, when σ_i^2 is assumed to be directly observed, the independence assumption is critical for all the derivations. For instance, the non-null proportion, defined as $\mathbb{P}(\theta_i\geq\theta_\alpha)$, must change for different values of σ_i if we allow the distribution of θ to depend on σ.

5.1. *A Conjugate Gaussian Example*

Suppose we have balanced panel data $y_{i1},\dots,y_{iT}\sim\mathcal{N}(\theta,\sigma^2)$ with sample means $Y_i=\frac{1}{T}\sum_t y_{it}$ and sample variances $S_i=\frac{1}{T-1}\sum_t(y_{it}-Y_i)^2$. Further, suppose that $G(\theta,\sigma^2)$ takes the normal-inverse-chi-squared form, $\mathrm{NIX}(\theta_0,\kappa_0,\nu_0,\sigma_0^2)=\mathcal{N}(\theta|\theta_0,\sigma^2/\kappa_0)\chi^{-2}(\sigma^2|\nu_0,\sigma_0^2)$. Integrating out σ^2, the marginal distribution of θ becomes a Student t-distribution,

$$\frac{\theta-\theta_0}{\sigma_0/\sqrt{\kappa_0}}\sim t_{\nu_0},$$

where t_{ν_0} is the t-distribution with degree of freedom ν_0. Therefore, the $1-\alpha$ quantile of θ, denoted θ_α, is simply

$$\theta_\alpha=\theta_0+\frac{\sigma_0}{\sqrt{\kappa_0}}F_{t_{\nu_0}}^{-1}(1-\alpha),$$

where $F_{t_{\nu_0}}^{-1}$ denotes the quantile function of t_{ν_0}.

Conjugacy of the distribution G implies that the posterior distribution of $(\theta,\sigma^2|Y,S)$ follows $\mathrm{NIX}(\theta_T,\kappa_T,\nu_T,\sigma_T^2)=\mathcal{N}(\theta|\theta_T,\sigma_T^2/\kappa_T)\chi^{-2}(\sigma^2|\nu_T,\sigma_T^2)$ with

$$\begin{aligned}
\nu_T&=\nu_0+T,\\
\kappa_T&=\kappa_0+T,\\
\theta_T&=\frac{\kappa_0\theta_0+TY}{\kappa_T},\\
\sigma_T^2&=\frac{1}{\nu_T}\left(\nu_0\sigma_0^2+(T-1)S+\frac{T\kappa_0}{\kappa_0+T}(\theta_0-Y)^2\right).
\end{aligned}$$

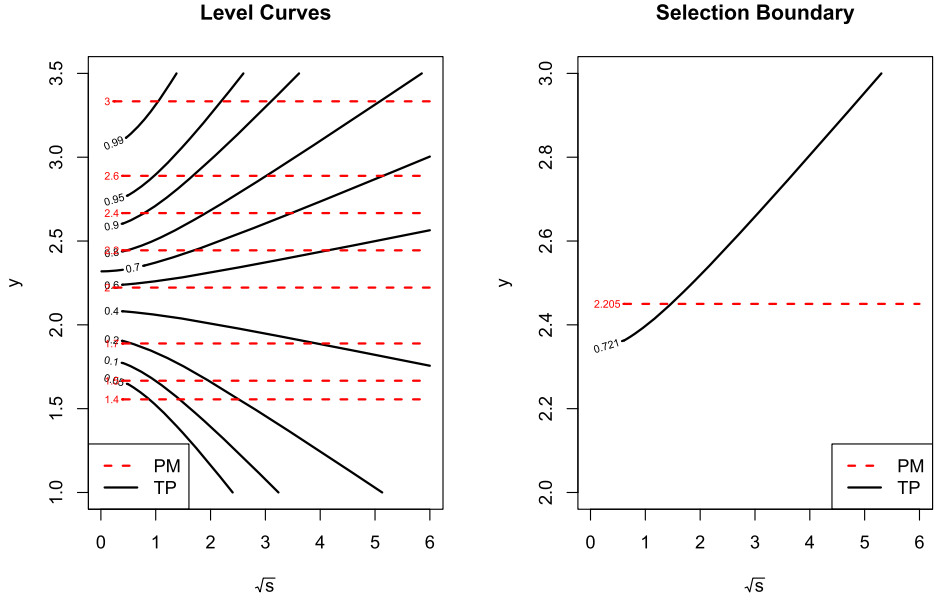

FIGURE 5.—The left panel shows level curves of the posterior mean (marked as red dashed lines) and the posterior tail probability (marked as black solid lines) for the normal model with $(\theta, \sigma^2) \sim \mathrm{NIX}(0, 1, 6, 1)$ and panel time dimension $T = 9$. The right plot shows the boundary of the selection region based on posterior mean ranking (marked as the red dashed line) and the posterior tail probability ranking (marked as the solid black line) with $\alpha = 5\%$ and $\gamma = 10\%$.

Integrating out σ^2, the marginal posterior of θ again follows a t-distribution,

$$\frac{\theta - \theta_T}{\sigma_T/\sqrt{\kappa_T}} \sim t_{\nu_T}.$$

It is thus clear that the posterior mean of θ is simply a linear function of Y and independent of S,

$$\mathbb{E}[\theta|Y, S] = \theta_T = \frac{\kappa_0\theta_0 + TY}{\kappa_T},$$

and the posterior tail probability is given by

$$v_\alpha(Y, S) = \mathbb{P}(\theta \geq \theta_\alpha|Y, S) = \mathbb{P}\left(\frac{\theta - \theta_T}{\sigma_T/\sqrt{\kappa_T}} \geq \frac{\theta_\alpha - \theta_T}{\sigma_T/\sqrt{\kappa_T}}|Y, S\right) = 1 - F_{t_{\nu_T}}\left(\frac{\theta_\alpha - \theta_T}{\sigma_T/\sqrt{\kappa_T}}\right).$$

To illustrate this case, suppose $\theta_0 = 0$, $\kappa_0 = 1$, $\sigma_0^2 = 1$ and $\nu_0 = 6$ and $T = 9$. It can be verified that $v_\alpha(Y, S)$ is in fact a monotone function of Y for each fixed S and any $\alpha > 0$; hence, in this example, we can invert the function $v_\alpha(y, s)$ to obtain the level curves. The left panel of Figure 5 shows the level curves for $v_\alpha(Y, S)$ and $\mathbb{E}(\theta|Y, S)$ for $\alpha = 5\%$. It is clear that the posterior mean is a constant function of S, while the posterior tail probability exhibits more exotic behavior with respect to S, especially for more extreme values of Y. If we fix $S = s_0$, then $v_\alpha(Y, s_0)$ is an increasing function of Y. On the other hand, fixing $Y = y_0$ for small y_0 implies that $v_\alpha(y_0, S)$ is an increasing function of S, while for y_0 large, $v_\alpha(y_0, S)$ becomes a decreasing function of S.

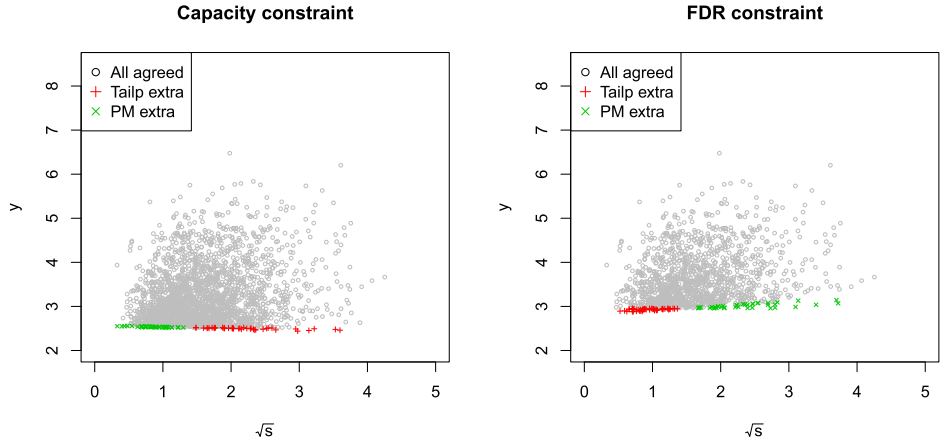

FIGURE 6.—Selection set comparison for one sample realization from the normal model with $(\theta, \sigma^2) \sim \mathrm{NIX}(0, 1, 6, 1)$ and panel time dimension $T = 9$. The left panel shows in gray circles the agreed selected elements by both the posterior mean and the posterior tail probability criteria under the capacity constraint; extra elements selected by the posterior mean are marked in green and extra elements selected by the posterior tail probability rule are marked in red. The right panel shows the comparison of the selected sets under both the capacity and the FDR constraint with $\alpha = 5\%$ and $\gamma = 10\%$.

A capacity constraint of size α implies the thresholding rule,

$$\mathbb{P}\big(v_\alpha(Y, S) \geq \lambda_2^*\big) = \alpha,$$

while FDR control at level γ leads to a cutoff value λ_1^* defined as

$$\gamma = \mathbb{E}\big[\big(1 - v_\alpha(Y, S)\big)\mathbb{1}\big\{v_\alpha(Y, S) \geq \lambda_1^*\big\}\big]/\mathbb{P}\big(v_\alpha(Y, S) \geq \lambda_1^*\big).$$

The larger of the two thresholds, denoted $\lambda^* = \max\{\lambda_1^*, \lambda_2^*\}$, defines the selection region based on posterior tail probability ranking $\Omega_{\alpha,\gamma} = \{(Y, S) : v_\alpha(Y, S) \geq \lambda^*\}$. For $\alpha = 5\%$ and $\gamma = 10\%$, the selection region based on the tail probability rule is $\{(Y, S) : v_\alpha(Y, S) \geq 0.72\}$. The posterior mean ranking is defined as $\{(Y, S) : \mathbb{E}[\theta|Y, S] \geq 2.2\}$. These selection boundaries are depicted as the red dashed line and black solid line respectively in the right panel of Figure 5. In this case, the FDR constraint binds. If only the capacity constraint were in place, we would have a cutoff for tail probability at 0.40 and the cutoff for the posterior mean at 1.84. Figure 6 further shows the comparison of the selected set based on a sample realization from the model.

In Appendix B of the Supplemental Material, we consider a more complex bivariate discrete example that illustrates somewhat more exotic behavior of the decision boundaries and compares performance of several different ranking and selection rules.

5.2. *Variants of the Unknown Variance Model*

We have assumed that the only scale heterogeneity is driven by σ_i in the above model, but often there may be more heteroscedasticity that should be allowed in ϵ_{it}. Here we consider a variant where

$$Y_{it} = \theta_i + \sigma_i \epsilon_{it}, \qquad \epsilon_{it} \sim \mathcal{N}(0, 1/w_{it}), \qquad \big(\theta_i, \sigma_i^2\big) \sim G.$$

We will assume that $w_{it} \sim H$ are known quantities and are independent from (θ_i, σ_i^2). Denoting $w_i = \sum_{t=1}^{T_i} w_{it}$, the sufficient statistics now take the form $Y_i = \sum_{t=1}^{T_i} w_{it} Y_{it}/w_i$ and $S_i = (T_i - 1)^{-1} \sum_{t=1}^{T_i} (Y_{it} - Y_i)^2$. In Gu and Koenker (2017), we have illustrated this formulation for predicting baseball batting averages; in that setting, "at bats" for player i in year t are given by the w_{it}, but there is still some player-specific heterogeneity in the σ_i's representing consistency of batting performance. It is easy to show that $Y_i|\theta_i, \sigma_i^2 \sim \mathcal{N}(\theta_i, \sigma_i^2/w_i)$ and $S_i|\sigma_i^2$ follows Gamma distribution with shape parameter $r_i = (T_i - 1)/2$ and scale parameter σ_i^2/r_i. The decision rules now become a function of the tuple (Y_i, S_i, w_i); for instance, the tail probability can be specified as

$$v_\alpha(y_i, s_i, w_i) = \mathbb{P}(\theta_i \geq \theta_\alpha | y_i, s_i, w_i) = \frac{\int_{\theta_\alpha}^{+\infty} f\big(y_i|\theta, \sigma^2/w_i\big)\Gamma\big(s_i|r_i, \sigma^2/r_i\big)\, dG\big(\theta, \sigma^2\big)}{\int_{-\infty}^{+\infty} f\big(y_i|\theta, \sigma^2/w_i\big)\Gamma\big(s_i|r_i, \sigma^2/r_i\big)\, dG\big(\theta, \sigma^2\big)},$$

and the posterior mean takes the form

$$\mathbb{E}[\theta_i|y_i, s_i, w_i] = \int \theta f\big(y_i, s_i|\theta, \sigma^2, w_i\big)\, dG\big(\theta, \sigma^2\big).$$

The threshold values under either the capacity or the FDR constraint can be worked out in a similar fashion. For any ranking statistics $\delta(Y_i, S_i, w_i)$ together with a decision rule $\mathbb{1}\{\delta(Y_i, S_i, w_i) \geq \lambda\}$, the capacity constraint requires choosing a thresholding value $\lambda_2^*(\alpha)$ such that

$$\alpha = \int\int \mathbb{1}\big\{\delta(y, s, w) \geq \lambda_2^*(\alpha)\big\} f\big(y, s|\theta, \sigma^2, w\big)\, dG\big(\theta, \sigma^2\big)\, dH(w),$$

while the thresholding value to control in addition the FDR rate under size γ requires solving for $\lambda_1^*(\alpha, \gamma)$ such that

$$\gamma = \frac{\mathbb{P}\big(\delta(y, s, w) \geq \lambda_1^*(\alpha, \gamma);\, \theta < \theta_\alpha\big)}{\mathbb{P}\big(\delta(y, s, w) \geq \lambda_1^*(\alpha, \gamma)\big)},$$

which can be further represented as

$$\gamma = \frac{\int\int \mathbb{1}\big\{\delta(y, s, w) \geq \lambda_1^*(\alpha, \gamma)\big\}(1 - \alpha) f_0\big(y, s|\theta, \sigma^2, w\big)\, dG\big(\theta, \sigma^2\big)\, dH(w)}{\int\int \mathbb{1}\big\{\delta(t, s, w) \geq \lambda_1^*(\alpha, \gamma)\big\} f\big(y, s|\theta, \sigma^2, w\big)\, dG\big(\theta, \sigma^2\big)\, dH(w)},$$

where $f_0(y, s|\theta, \sigma^2, w)$ is the density of (y, s) under the null hypothesis $\theta < \theta_\alpha$.

We can again consider selection regions as those plotted in Figure 5 and Figure S.1 to appreciate how different decision criteria determine the selection. As soon as the ranking statistics depend on w, the selection region of the thresholding rule $\mathbb{1}\{\delta(y, s, w) \geq \lambda^*\}$ will also depend on the magnitude of w.

6. ASYMPTOTIC ADAPTIVITY

The previous sections propose Bayes rules for minimizing the expected number of missed discoveries subject to both capacity and FDR constraints under several modeling

environments. In each of these environments, the Bayes rule takes the form $\delta^* = 1\{v_\alpha \geq \lambda^*\}$, where v_α is defined as the posterior probability of $\theta \geq \theta_\alpha$ conditional on the data. The thresholding value λ^* is defined to satisfy both the capacity and FDR constraints. The Bayes rule involves several unknown quantities, in particular the v_α's and the optimal thresholding value, λ^*, that require knowledge on the distribution of θ_i or the joint distribution of (θ_i, σ_i^2) when the variances are latent variables. For estimating this distribution of the latent variables, we propose a plug-in procedure that is very much in the spirit of empirical Bayes methods pioneered by Robbins (1956). In this section, we also establish that the resulting feasible rules achieve asymptotic validity and asymptotically attain the same performance as the infeasible Bayes rule.

We begin by discussing properties of the oracle procedure assuming that v_α is known and we only need to estimate the optimal thresholding value. We establish asymptotic validity of this oracle procedure and then propose a plug-in method for both v_α and the thresholding value thereby establishing the asymptotic validity of the empirical rule. Before presenting the formal results, we introduce regularity conditions that will be required. We distinguish two cases depending on whether the σ_i^2's are observed.

ASSUMPTION 1: 1. (*Variances observed*) *$\{Y_i, \sigma_i^2, \theta_i\}$ are independent and identically distributed with a joint distribution with σ_i^2 and θ_i independent. The random variables θ_i and σ_i^2 have positive densities with respect to Lebesgue measure on a compact set $\Theta \subset \mathbb{R}$ and $[\underline{\sigma}^2, \overline{\sigma}^2]$ respectively for some $\underline{\sigma}^2 > 0$ and $\overline{\sigma}^2 < +\infty$.*

2. (*Variance unobserved*) *Let S_i be an individual sample variance based on T repeated measurements and Y_i be the sample means with $T \geq 4$. Suppose further that $\{Y_i, S_i, \theta_i, \sigma_i\}$ are independent and identically distributed and that the random variables $\{\theta_i, \sigma_i^2\}$ have a joint distribution G with a joint density positive everywhere on its support.*

6.1. *Optimal Thresholding*

Whether σ_i^2 is observed or estimated, the optimal thresholding value can be defined in a unified manner by $\lambda^* = \max\{\lambda_1^*, \lambda_2^*\}$ with

$$\lambda_1^* = \inf\big\{t \in (0,1), H_v(t) \geq 1 - \alpha\big\},$$

$$\lambda_2^* = \inf\big\{t \in (0,1), Q(t) \leq \gamma\big\},$$

where H_v denotes the cumulative distribution of either $v_\alpha(y_i, \sigma_i)$ or $v_\alpha(y_i, s_i)$, induced by the marginal distribution of the data, either as the pair $\{y_i, \sigma_i\}$ when variances are observed or the pair $\{y_i, s_i\}$ otherwise. Hence λ_1^* is the $1 - \alpha$ quantile of H_v.

The function $Q(t)$ is defined as $Q(t) = \mathbb{E}[(1 - v_\alpha)1\{v_\alpha \geq t\}]/\mathbb{E}[1\{v_\alpha \geq t\}]$. Its formulation recalls Proposition 5.2 and the existence of λ_2^* is guaranteed as long as $\alpha < 1 - \gamma$. The thresholding value is also equivalent to those defined in Proposition 3.3 and Proposition 4.1. In particular, the thresholding values t_1^* and t_2^* in Proposition 3.3 are cast in terms of Y directly and it is easy to see $\lambda_j^* = v_\alpha(t_j^*)$ for $j = 1, 2$ when variances are homogeneous. In a similar spirit, the explicit formulae for λ_1^* and λ_2^* in Proposition 4.1 are a result of invoking the monotonicity of $v_\alpha(y, \sigma)$ with respect to y for each fixed value of σ. The function $Q(t)$ is the mFDR of the procedure $\delta = 1\{v_\alpha \geq t\}$ for any $\alpha \in (0, 1)$, and is monotonically decreasing in t. Monotonicity of $Q(t)$ is crucial to justify this thresholding procedure insuring that either the capacity constraint or the mFDR constraint must be binding. Cao, Sun, and Kosorok (2013) have observed that a sufficient condition for monotonicity for a broad class of multiple testing procedures is that the ratio of densities

under the null and alternative of the test statistics employed for ranking be monotone and they discussed the consequences of the violation of this condition. For the posterior tail probability criterion, this monotone likelihood ratio condition, as we will see, can be verified directly.

Recall that mFDR is defined as $\sum_{i=1}^n \mathbb{P}[\delta_i = 1, \theta_i < \theta_\alpha]/\sum_{i=1}^n \mathbb{P}(\delta_i = 1)$. It suffices to show that $\mathbb{P}[\delta_i = 1, \theta_i < \theta_\alpha] = \mathbb{E}[(1 - v_{\alpha,i})\delta_i]$. Since $v_{\alpha,i} = P[\theta_i \geq \theta_\alpha | D_i] = \alpha f_1(D_i)/f(D_i)$, where D_i is the individual data being either $\{y_i, \sigma_i\}$ or $\{y_i, s_i\}$ depending on the model and f_1 is the marginal density of the data when $\theta_i \geq \theta_\alpha$ and f is the marginal density of D_i, then it is clear that $\mathbb{P}[\delta_i = 1, \theta_i < \theta_\alpha] = (1 - \alpha) \int \mathbb{1}\{v_{\alpha,i} \geq t\} f_0(D_i)\, dD_i = \int \mathbb{1}\{v_{\alpha,i} \geq t\}(1 - v_{\alpha,i}) f(D_i)\, dD_i = \mathbb{E}[(1 - v_{\alpha,i})\mathbb{1}\{v_{\alpha,i} \geq t\}]$. Then $Q(t) = \int_t^1 (1 - v) h_v\, dv / \int_t^1 h_v\, dv$, where h_v is the density function of v_α. Monotonicity of $Q(t)$ can be easily verified by showing that the derivative with respect to t of the right-hand-side quantity is nonpositive.

6.2. *Oracle Procedures*

The only unknown quantity in the oracle procedure is the thresholding value and we now discuss how to estimate it to achieve asymptotic validity. H_v and Q can be estimated by the following quantities:

$$H_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{v_{\alpha,i} \leq t\},$$

$$Q_n(t) = \frac{\sum_{i=1}^n (1 - v_{\alpha,i}) \mathbb{1}\{v_{\alpha,i} \geq t\}}{\sum_{i=1}^n \mathbb{1}\{v_{\alpha,i} \geq t\}},$$

and the associated thresholding values are then defined as $\lambda_n = \max\{\lambda_{1n}, \lambda_{2n}\}$, with

$$\lambda_{1n} = \inf\big\{t \in [0, 1], H_n(t) \geq 1 - \alpha\big\},$$
$$\lambda_{2n} = \inf\big\{t \in [0, 1], Q_n(t) \leq \gamma\big\}.$$

THEOREM 6.1—Asymptotic Validity of the Oracle Procedure: *Under Assumption* 1, *the procedure* $\delta_i = 1\{v_{\alpha,i} \geq \lambda_n\}$ *asymptotically controls the false discovery rate below* γ *and the expected proportion of rejections below* α *for any* $(\alpha, \gamma) \in [0, 1]^2$ *and* $\gamma < 1 - \alpha$ *when* $n \to \infty$; *more specifically*,

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{1}\{\theta_i < \theta_\alpha, v_{\alpha,i} \geq \lambda_n\}}{\sum_{i=1}^n \mathbb{1}\{v_{\alpha,i} \geq \lambda_n\} \vee 1}\right] \leq \gamma,$$

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{v_{\alpha,i} \geq \lambda_n\}\right] \leq \alpha.$$

6.3. *Adaptive Procedures*

In practise, the posterior tail probability also involves the unknown quantity $\theta_\alpha = G^{-1}(1-\alpha)$ that needs to be estimated. We propose a plug-in estimator in the spirit of the empirical Bayes method: estimating G by its nonparametric maximum likelihood estimator $\hat{G}_n$ and estimating θ_α as its $1-\alpha$ quantile.

Consistency of the nonparametric maximum likelihood estimator, $\hat{G}_n$, was first proven by Kiefer and Wolfowitz (1956) using Wald type arguments. A Hellinger risk bound for the associated marginal density estimate and adaptivity of $\hat{G}_n$ and a self-regularization property have been recently established in Saha and Guntuboyina (2020) and Polyanskiy and Wu (2020). In particular, the following established result, stated here as an assumption, is crucial for establishing the asymptotic validity of the adaptive procedure.

ASSUMPTION 2: *The nonparametric maximum likelihood estimator $\hat{G}_n$ is strongly consistent for G. That is, for all continuity points k of G, $\hat{G}_n(k) \to G(k)$ almost surely as $n \to \infty$. Furthermore, the estimated marginal (mixture) density converges almost surely in Hellinger distance.*

When variances are homogeneous or when variances are unknown but we have longitudinal data so that we have a mixture model for the pair $\{Y_i, S_i\}$, the Hellinger convergence is established in van de Geer (1993). When variances are heterogeneous but known, the Hellinger bound for marginal density has been established recently in Jiang (2020).

The plug-in estimators for the posterior tail probability, $v_\alpha(y_i, \sigma_i)$ when variances are known or $v_\alpha(y_i, s_i)$ when variances are unknown, are then defined respectively as

$$\hat{v}_\alpha(y_i, \sigma_i) = \int_{\hat{\theta}_\alpha}^{+\infty} \varphi\big(y_i|\theta, \sigma_i^2\big)\, d\hat{G}_n(\theta) \Big/ \int_{-\infty}^{+\infty} \varphi\big(y_i|\theta, \sigma_i^2\big)\, d\hat{G}_n(\theta),$$

$$\hat{v}_\alpha(y_i, s_i) = \int_{\hat{\theta}_\alpha}^{+\infty} f(y_i, s_i|\theta, \sigma)\, d\hat{G}_n\big(\theta, \sigma^2\big) \Big/ \int_{-\infty}^{+\infty} f(y_i, s_i|\theta, \sigma)\, d\hat{G}_n\big(\theta, \sigma^2\big),$$

where f is the density function for (y_i, s_i) which is a product of Gaussian and gamma densities. Abbreviating the estimated posterior tail probability by $\hat{v}_{\alpha,i}$, we mimic the oracle procedure and estimate the thresholding value by $\hat{\lambda}_n = \max\{\hat{\lambda}_{1n}, \hat{\lambda}_{2n}\}$, where

$$\hat{\lambda}_{1n} = \inf\left\{ t \in [0,1] : \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\hat{v}_{\alpha,i} \le t\} \ge 1-\alpha \right\},$$

$$\hat{\lambda}_{2n} = \inf\left\{ t \in [0,1] : \frac{\sum_{i=1}^{n}(1-\hat{v}_{\alpha,i})\mathbb{1}\{\hat{v}_{\alpha,i} \ge t\}}{\sum_{i=1}^{n}\mathbb{1}\{\hat{v}_{\alpha,i} \ge t\}} \ge \gamma \right\}.$$

THEOREM 6.2—Asymptotic Validity of Adaptive Procedure: *Under Assumptions 1 and 2, the adaptive procedure $\delta_i = 1\{\hat{v}_{\alpha,i} \ge \hat{\lambda}_n\}$ asymptotically controls the false discovery rate below γ and the expected proportion of rejections below α for any $(\alpha, \gamma) \in [0,1]^2$ with $\alpha <$*

$1-\gamma$ *when* $n\to\infty$; *more specifically*,

$$\limsup_{n\to\infty}\mathbb{E}\left[\frac{\sum_{i=1}^{n}\mathbb{1}\{\theta_i<\theta_\alpha,\hat{v}_{\alpha,i}\geq\hat{\lambda}_n\}}{\sum_{i=1}^{n}\mathbb{1}\{\hat{v}_{\alpha,i}\geq\hat{\lambda}_n\}\vee 1}\right]\leq\gamma,$$

$$\limsup_{n\to\infty}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{v}_{\alpha,i}\geq\hat{\lambda}_n\}\right]\leq\alpha.$$

It is clear that given the Lagrangian formulation of the compound decision problem, it can be viewed equivalently as a constrained optimization problem. See also the discussion in Remark 3.4. We seek to maximize power defined as $\beta(t):=\mathbb{P}(\theta_i\geq\theta_\alpha,\delta_i=1)/\alpha$ subject to two constraints: the first is the marginal FDR rate and the other is the selected proportion. For each fixed pair of $\{\alpha,\gamma\}$, the Bayes rule achieves the best power among all thresholding procedures that respect the two constraints. The next theorem establishes that our feasible, adaptive procedure achieves the same power as the oracle rule asymptotically. It is supported by the simulation evidence presented in the next section. In practice, we suggest convolution smoothing of the discrete $\hat{G}$ with a bandwidth slowly tending to zero with n. The resulting smoothed mixing distribution is also consistent, hence fulfilling Assumption 2 and therefore all our adaptivity results.

THEOREM 6.3: *Under Assumptions 1 and 2, the adaptive procedure* $\delta_i=\mathbb{1}\{\hat{v}_{\alpha,i}\geq\hat{\lambda}_n\}$ *attains the same power as the optimal Bayes rule asymptotically. In particular, as* $n\to\infty$,

$$\frac{\sum_{i=1}^{n}\mathbb{1}\{\theta_i\geq\theta_\alpha,\hat{v}_{\alpha,i}\geq\hat{\lambda}_n\}}{\sum_{i=1}^{n}\mathbb{1}\{\theta_i\geq\theta_\alpha\}}\xrightarrow{p}\beta(\lambda^*).$$

7. SIMULATION EVIDENCE

In this section, we describe two small simulation exercises designed to illustrate performance of several competing methods for ranking and selection. As a benchmark for evaluating performance, we consider several oracle methods that presume knowledge of the true distribution, G, generating the θ's as well as several feasible methods that rely on estimation of G. These are contrasted with more traditional methods that are based on linear shrinkage rules of the Stein type. The linear shrinkage rule is the posterior mean of θ under the assumption that G follows a Gaussian distribution with unknown mean and variance parameters. This is the commonly used estimator for ranking and selection in applied work, notably Chetty, Friedman, and Rockoff (2014a, 2014b) for teacher evaluation and Chetty and Hendren (2018) for studying intergenerational mobility.

Typically the linear shrinkage estimator is used in the context of heterogeneous known variances; this will be the model we focus on in our simulation experiments. The linear shrinkage formula defined in (2.1) easily adapts to the heterogeneous variances case and

leads to the James–Stein shrinkage rule with heterogeneous known variances. Efron and Morris (1973) introduced some further modifications. As we have already demonstrated, when variances are heterogeneous, the linear shrinkage estimator provides a different ranking than the posterior tail probability rules. Further complications arise when we seek procedures that also control false discovery. To estimate the false discovery rate for different thresholding values, we requires knowledge of G. If the Gaussian assumption on G underlying the linear shrinkage rules is misplaced, it may lead to an inaccurate estimate of FDR, and consequently to procedures that fail to control for false discovery.

Performance will be evaluated primarily on the basis of power, which we define as the proportion of individuals whose true θ_i exceeds the cutoff $\theta_\alpha = G^{-1}(1-\alpha)$, who are actually selected. This is the sample counterpart of $\mathbb{P}(\delta_i = 1, \theta_i \geq \theta_\alpha)/\mathbb{P}(\theta_i \geq \theta_\alpha)$. FDR is calculated as the sample counterpart of $\mathbb{P}(\delta_i = 1, \theta_i < \theta_\alpha)/\mathbb{P}(\delta_i = 1)$, that is, the proportion of selected individuals whose true θ_i falls below the threshold. While our selection rules are *intended* to constrain FDR below the γ threshold, as in other testing problems, they are not always successful in this objective in finite samples so empirical power comparisons must be interpreted cautiously in view of this. Nonetheless, asymptotic validity is assured by the results in Section 6. We compare performance for three distinct α levels, $\{0.05, 0.10, 0.15\}$, and three γ levels, $\{0.05, 0.10, 0.15\}$.

7.1. *The Student t Setting*

Our first simulation setting focuses on the effect of tail behavior of the distribution on performance of competing rules. For these simulations, we take G to be a discrete approximation to Student t distributions with degrees of freedom in the set $\{1, 2, 3, 5, 10\}$, and supported on the interval $[-20, 20]$. The scale parameters of the Gaussian noise contribution are independent and uniformly distributed on the interval $[0.5, 1.5]$. We report power performance for several alternative ranking and selection rules:

OTP Oracle Tail Probability Rule
OPM Oracle Posterior Mean Rule
Efron Efron Tail Probability Rule
KWs Kiefer–Wolfowitz Smoothed Tail Probability Rule
EM Efron and Morris (1973) Linear Shrinkage Rule

The KWs rule uses $\tilde{G} = \hat{G} * K_h$, with biweight kernel K and bandwidth h equal to half the mean absolute deviation from the median of $\hat{G}$. The Efron rule uses his suggested default of a natural spline basis with five degrees of freedom and penalty parameter 0.1.

We illustrate the results in Figure 7, where we plot empirical power against degrees of freedom of the t distribution for a selected set of values for the capacity constraint, $\alpha \in \{0.05, 0.10, 0.15\}$, and FDR constraint, $\gamma \in \{0.05, 0.10, 0.15\}$, as indicated at the top of each panel of the figure. The most striking conclusion from this exercise is the dramatic decrease in power as we move toward the Gaussian distribution. At the Cauchy, t_1, power is quite respectable for all choices of α and γ, but power declines rapidly as the degrees of freedom increases, reinforcing our earlier conclusion that the Gaussian case is extremely difficult. We would stress, in view of this finding, that classical linear shrinkage procedures designed for the Gaussian setting are poorly adapted to heavy tailed settings in which the reliability of selection procedures is potentially greatest.

Careful examination of this figure also reveals that there is a slight advantage to the posterior tail probability rules over the posterior mean procedures, both for the oracle rules and for our feasible procedures. There is surprisingly little sacrifice in power in

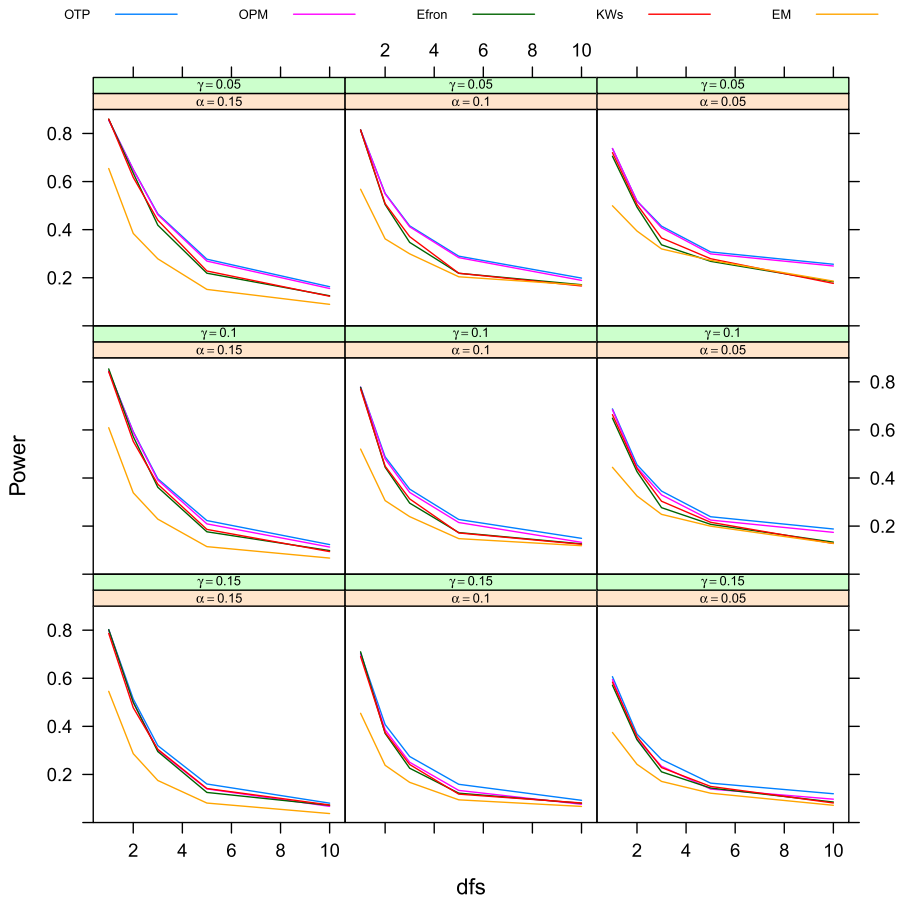

FIGURE 7.—Power performance for several selection rules with Student t signal. Capacity and FDR constraints are indicated at the top of each panel in the figure.

moving from the oracle methods to the Efron or Kiefer–Wolfowitz rules. The Efron and Morris selection rule is very competitive in the almost Gaussian, t_{10} setting but sacrifices considerable power in the lower degrees of freedom settings due to the misspecification of the distribution G and consequent inaccurate estimation of the false discovery rate.

7.2. *A Teacher Value-Added Setting*

Our second simulation setting is based on a discrete approximation of the data structure employed in Gilraine, Gu, and McMillan (2020) to study teacher value-addded methods. Several longitudinal waves of student test scores from the Los Angeles Unified School District were combined in this study. Here we abstract from many features of the full longitudinal structure of these data, and focus instead on comparing performance of several selection methods. We maintain our standard known variance model in which we observe $Y_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$ with θ_i's drawn iidly from a distribution $\tilde{G}$ estimated by Gilraine, Gu, and McMillan (2020). This distribution was estimated from the full longitudinal LA sample using the nonparametric maximum likelihood estimator of Kiefer and Wolfowitz and then smoothed slightly by convolution with a biweight kernel and illustrated in the left panel of

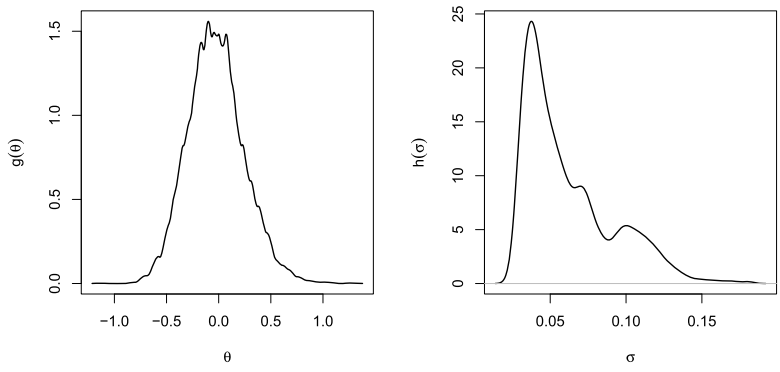

FIGURE 8.—Densities of "latent" (mean) ability and standard deviation for the teacher value-added simulations.

Figure 8. Variances, in keeping with our hypothesis in Section 4, are drawn from a distribution with density illustrated in the right panel of Figure 8. We focus on selection from the left tail of the resulting distribution since it is those teachers whose jobs are endangered by recent policy recommendations in the literature (see, for instance, Hanushek (2011)).

We draw samples of size 10,000 from the foregoing distribution and compute performance measures based on 100 replications. The fitted densities for this simulation exercise are based on a sample of roughly 11,000 teachers, so the simulation sample size is chosen to be commensurate with this. In Table II, we report power, FDR, and the proportion selected by ten selection rules. The oracle rules, OTP and OPM, based ranking by the tail probability, and posterior mean criteria can be considered benchmarks for the remaining feasible procedures. Only the oracle procedures can be considered reliable from the perspective of adhering to the capacity and FDR constraints. Consequently, some caution is required in the interpretation of the power comparisons since feasible procedures can exhibit good power at the expense of violating these constraints. This is analogous to the common difficulty in interpreting power in testing problems when different procedures have differing size. When FDR is constrained to 5%, even the oracle is only able to select about half of the deserving individuals; OTP is consistently preferable to OPM as expected and power performance improves somewhat as the capacity constraint is relaxed. Among the feasible G-modeling selection procedures, the Efron rules have good power performance, but fail to meet the FDR constraints. We conjecture that somewhat less aggressive smoothing than the default, $df = 5$, $c_0 = 0.1$, might help to rectify this. In contrast, the smoothed Kiefer–Wolfowitz rules are somewhat overly conservative in meeting the FDR constraints and might benefit from somewhat more aggressive smoothing.

Among the other procedures, the linear posterior mean rule, LPM, as employed by Chetty, Friedman, and Rockoff (2014a, 2014b), and the linear posterior mean rule, EM, of Efron and Morris (1973) behave identically and exhibit somewhat erratic FDR control due to the misspecified Gaussian assumption on G; this leads to weaker power performance. As a further comparison, when the linear shrinkage rules are implemented without any FDR constraint, denoted LPM* and EM* in the table, as they typically would be used in practice, the false discovery proportion is considerably higher than the targeted γ. We also report the performance of MLE and P-value rules, implemented without FDR control; again, both yield a higher FDR rate, making it difficult to evaluate their power performance.

TABLE II

COMPARISON OF PERFORMANCE OF SEVERAL SELECTION RULES FOR THE TEACHER VALUE-ADDED SIMULATION.

	$\gamma = 5\%$				$\gamma = 10\%$			
	$\alpha = 1\%$	$\alpha = 3\%$	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 1\%$	$\alpha = 3\%$	$\alpha = 5\%$	$\alpha = 10\%$
Power								
OTP	0.394	0.520	0.554	0.626	0.494	0.625	0.661	0.733
OPM	0.365	0.492	0.521	0.599	0.484	0.620	0.654	0.731
ETP	0.435	0.540	0.580	0.657	0.540	0.647	0.688	0.759
KWTP	0.355	0.477	0.521	0.614	0.452	0.583	0.631	0.723
EPM	0.398	0.511	0.552	0.632	0.528	0.642	0.683	0.758
KWPM	0.325	0.447	0.492	0.588	0.440	0.576	0.624	0.719
LPM	0.162	0.341	0.418	0.689	0.246	0.460	0.542	0.805
LPM*	0.726	0.781	0.796	0.829	0.726	0.781	0.796	0.829
EM	0.162	0.341	0.418	0.689	0.246	0.460	0.542	0.805
EM*	0.726	0.781	0.796	0.829	0.726	0.781	0.796	0.829
MLE	0.699	0.768	0.787	0.824	0.699	0.768	0.787	0.824
P-val	0.374	0.478	0.535	0.635	0.374	0.478	0.535	0.635
FDR								
OTP	0.050	0.050	0.051	0.051	0.103	0.103	0.100	0.101
OPM	0.047	0.050	0.051	0.053	0.103	0.101	0.101	0.102
ETP	0.070	0.059	0.061	0.064	0.128	0.115	0.117	0.119
KWTP	0.035	0.037	0.041	0.048	0.082	0.081	0.085	0.096
EPM	0.063	0.057	0.059	0.062	0.129	0.115	0.116	0.118
KWPM	0.038	0.040	0.045	0.051	0.081	0.084	0.087	0.097
LPM	0.016	0.025	0.033	0.083	0.031	0.048	0.061	0.151
LPM*	0.276	0.226	0.207	0.172	0.276	0.226	0.207	0.172
EM	0.016	0.025	0.033	0.083	0.031	0.048	0.061	0.151
EM*	0.276	0.225	0.207	0.172	0.276	0.225	0.207	0.172
MLE	0.304	0.238	0.216	0.177	0.304	0.238	0.216	0.177
P-val	0.627	0.526	0.467	0.365	0.627	0.526	0.467	0.365
Selected								
OTP	0.004	0.016	0.029	0.066	0.006	0.021	0.037	0.082
OPM	0.004	0.015	0.027	0.063	0.005	0.021	0.036	0.081
ETP	0.005	0.017	0.031	0.070	0.006	0.022	0.039	0.086
KWTP	0.004	0.015	0.027	0.064	0.005	0.019	0.034	0.080
EPM	0.004	0.016	0.029	0.067	0.006	0.022	0.038	0.086
KWPM	0.003	0.014	0.026	0.062	0.005	0.019	0.034	0.080
LPM	0.002	0.010	0.022	0.075	0.003	0.014	0.029	0.095
LPM*	0.010	0.030	0.050	0.100	0.010	0.030	0.050	0.100
EM	0.002	0.010	0.022	0.075	0.003	0.014	0.029	0.095
EM*	0.010	0.030	0.050	0.100	0.010	0.030	0.050	0.100
MLE	0.010	0.030	0.050	0.100	0.010	0.030	0.050	0.100
P-val	0.010	0.030	0.050	0.100	0.010	0.030	0.050	0.100

8. RANKING AND SELECTION OF U.S. DIALYSIS CENTERS

Motivated by important prior work on ranking and selection by Lin, Louis, Paddock, and Ridgeway (2006, 2009) illustrated by applications to ranking U.S. dialysis centers, we have chosen to maintain this focus to illustrate our own approach. Kidney disease is a growing medical problem in the United States and considerable effort has been devoted to data collection and evaluation of the relative performance of the more than 6000 dialysis centers serving the afflicted population. Centers are evaluated on multiple criteria, but the primary focus of center ranking is their standardized mortality rate, or SMR, the

ratio of observed deaths to expected deaths for center patients. Allocating patients to centers is itself a complex task since patients may move from one center to another in the course of a year. Centers also vary considerably in the mix of patients they serve. Predictions from an estimated Cox proportional hazard model that attempts to account for this heterogeneity are employed to estimate expected deaths for each center.

Our analysis focuses exclusively on the SMR evaluation of centers using longitudinal data from 2004–2018 as reported in University of Michigan Kidney Epidemiology and Cost Center (2009–2019). We restrict attention to 3230 centers that have consistently reported SMR data over this sample period. Observed deaths, denoted y_{it} for center i in year t, are conventionally modeled as Poisson,

$$y_{it} \sim \text{Pois}(\rho_i \mu_{it}),$$

where μ_{it} is center i's expected deaths as predicted by the Cox model in year t and ρ_i is the center's unobserved mortality rate. We view μ_{it} as the effective sample size for the center, after adjustment for patient characteristics of the center. Center characteristics are explicitly excluded from the Cox model. The classical variance stabilizing transformation for the Poisson brings us back to the Gaussian model,

$$z_{it} = \sqrt{y_{it}/\mu_{it}} \sim \mathcal{N}(\theta_i, 1/w_{it}),$$

where $\theta_i = \sqrt{\rho_i}$ and $w_{it} = 4\mu_{it}$. Exchangeability of the centers yields a mixture model in which the parameter θ_i is effectively assumed to be drawn iidly from a distribution, G. The predictions of expected mortality, μ_{it}, are assumed to be sufficiently accurate that we treat w_{it} as known, and independent of $\theta_i \sim G$.

Over short time horizons like 3 years, we assume that centers have a fixed draw of θ_i from G, and thus we have sufficient statistics for θ_i as

$$T_i = \sum_{t \in \mathcal{T}} w_{it} z_{it} / w_i \sim \mathcal{N}(\theta_i, 1/w_i),$$

where the set $\mathcal{T}$ is the corresponding 3-year window and $w_i = \sum_t w_{it}$. Given these ingredients, it is straightforward to construct a likelihood for the mixing distribution, G, and proceed with estimation of it.

Our objective is then to select centers based on the posterior distributions of their θ_i's. For example, the posterior tail probability of center i is given by

$$v_\alpha(t_i, w_i) = \mathbb{P}(\theta_i \geq \theta_\alpha | t_i, w_i) = \frac{\int_{\theta_\alpha}^{+\infty} f(t_i|\theta, w_i)\, dG(\theta)}{\int_{-\infty}^{+\infty} f(t_i|\theta, w_i)\, dG(\theta)},$$

where f is the density function of T_i conditional on θ_i and w_i. The capacity constraint requires choosing a thresholding value $\lambda_2^*(\alpha)$ such that

$$\alpha = \int \int \mathbb{1}\{v_\alpha(t, w) \geq \lambda_2^*(\alpha)\} \varphi(t|\theta, w)\, dG(\theta)\, dH(w),$$

which can be approximated by $\frac{1}{n}\sum_i \mathbb{1}\{v_\alpha(t_i, w_i) \geq \lambda_2^*(\alpha)\}$, and inverted to obtain the threshold. Based on the discussion in Section 6, for the FDR constraint we choose a

thresholding value $\lambda_1^*(\alpha,\gamma)$ such that

$$\gamma = \frac{\int\int \mathbb{1}\{v_\alpha(t,w) \geq \lambda_1^*(\alpha,\gamma)\}(1-v_\alpha(t,w))f(t|\theta,w)\,dG(\theta)\,dH(w)}{\int\int \mathbb{1}\{v_\alpha(t,w) \geq \lambda_1^*(\alpha,\gamma)\}f(t|\theta,w)\,dG(\theta)\,dH(w)}, \tag{8.1}$$

where H is the marginal distribution of the observed portion of the variance effect. The numerator can be approximated by $\frac{1}{n}\sum_i (1-v_\alpha(t_i,w_i))\mathbb{1}\{v_\alpha(t_i,w_i) \geq \lambda_1^*(\alpha,\gamma)\}$ and the denominator can be approximated by $\frac{1}{n}\sum_i \mathbb{1}\{v_\alpha(t_i,w_i) \geq \lambda_1^*(\alpha,\gamma)\}$.

The posterior mean ranking, in contrast, is based on

$$\delta(t_i,w_i) = \mathbb{E}[\theta_i|t_i,w_i] = \int \theta f(t_i|\theta,w_i)\,dG(\theta).$$

For the capacity constraint, we choose a thresholding value $C_2^*(\alpha)$ such that

$$\alpha = \int\int \mathbb{1}\{\delta(t,w) \geq C_2^*(\alpha)\}f(t|\theta,w)\,dG(\theta)\,dH(w).$$

For FDR constraint, we pick a thresholding value $C_1^*(\alpha,\gamma)$ such that

$$\gamma = \frac{\mathbb{P}(\delta(t,w) \geq C_1^*(\alpha,\gamma);\, \theta < \theta_\alpha)}{\mathbb{P}(\delta(t,w) \geq C_1^*(\alpha,\gamma))}.$$

The right-hand side of the FDR constraint can be approximated by

$$\frac{1}{n}\sum_i \mathbb{1}\{\delta(t_i,s_i,w_i) \geq C_1^*(\alpha,\gamma)\}(1-v_\alpha(t_i,w_i))/\frac{1}{n}\sum_i \mathbb{1}\{\delta(t_i,w_i) \geq C_1^*(\alpha,\gamma)\},$$

while the right-hand side of the capacity constraint can be approximated by

$$\frac{1}{n}\sum_i \mathbb{1}\{\delta(t_i,w_i) \geq C_2^*(\alpha)\},$$

so $C_2^*(\alpha)$ is simply the empirical quantile of the $\delta(t_i,w_i)$.

We will compare the foregoing ranking and selection rules with more naive rules based upon the Poisson and Gaussian MLEs, $\sum_{t\in\mathcal{T}} y_{it}/\sum_{t\in\mathcal{T}} \mu_{it}$, and T_i, respectively, a variant of the much maligned P-value, as well as a linear shrinkage procedure. For these rules, we do not attempt to control for FDR since this is how they are typically implemented in practice.

To help appreciate the difficulty of the selection task, Table III reports estimated FDR rates for several selection rules under a range of capacity constraints α for both right and left tail selection based on the data from 2004 to 2006. Right tail selection corresponds to identifying centers whose mortality rate is higher than expected; left tail selection to centers with mortality lower than expected. To estimate FDR, we require an estimate of the distribution of distribution, G. For this purpose, we use the smoothed version of the Kiefer–Wolfowitz NPMLE introduced in Section 2. The biweight bandwidth for the smoothing was chosen as the mean absolute deviation from the median of the discrete

TABLE III

FDR ESTIMATES: 2004–2006.

	$\alpha = 4\%$	$\alpha = 10\%$	$\alpha = 15\%$	$\alpha = 20\%$	$\alpha = 25\%$
Right Selection					
MLE	0.544	0.481	0.436	0.403	0.352
Poisson-MLE	0.545	0.485	0.440	0.406	0.355
P-value	0.532	0.475	0.432	0.399	0.349
Efron–Morris	0.521	0.473	0.429	0.398	0.349
James–Stein	0.521	0.473	0.429	0.398	0.349
PM	0.517	0.472	0.428	0.398	0.349
TP	0.517	0.471	0.428	0.397	0.349
Left Selection					
MLE	0.611	0.565	0.481	0.449	0.393
Poisson-MLE	0.600	0.561	0.478	0.448	0.393
P-value	0.620	0.565	0.477	0.450	0.393
Efron–Morris	0.595	0.552	0.472	0.445	0.391
James–Stein	0.595	0.552	0.472	0.445	0.391
PM	0.592	0.552	0.473	0.445	0.391
TP	0.589	0.550	0.471	0.444	0.390

NPMLE, $\hat{G}$. The assessment of FDR reported in Table III reflects the considerable uncertainty associated with the selected set of centers deemed by the capacity constraint to be in the upper (or lower) α quantile based upon our estimate of the distribution, G, of unobserved quality.

The MLE rule ranks centers based on their Gaussian MLE, T_i, while the Poisson-MLE rule ranks on $\sum_t y_{it} / \sum_t \mu_{it}$, which is the MLE of ρ_i from the Poisson model. *P*-value ranks centers based on the variance stabilizing transformation from the Poisson model under the null hypothesis $\rho_i = 1$ and $\rho_i > 1$ as the alternative hypothesis for right selection and $\rho_i < 1$ for the left selection. All these rules ignore the compound decision perspective of the problem entirely.

Among the compound decision rules, we consider the linear (James–Stein) shrinkage rule, $\hat{\mu}_\theta + (T_i - \hat{\mu}_\theta)\hat{\sigma}_\theta^2/(\hat{\sigma}_\theta^2 + 1/w_i)$, which is the posterior mean of θ_i based on the model $T_i \sim \mathcal{N}(\theta_i, 1/w_i)$ assuming that the latent variable θ_i follows a Gaussian distribution with mean μ_θ and variance σ_θ^2. We also consider the Efron and Morris (1973) estimator, which is a slight modification of the James–Stein estimator.

Finally, PM and TP are the posterior mean of θ and posterior tail probability of $\theta \geq \theta_\alpha$, for right selection, and $\theta \leq \theta_\alpha$ for left selection based on our estimated $\hat{G}$. For both left and right tail selection, as α increases, the FDR rate decreases, indicating the selection task becomes easier. All rules that account for the compound decision perspective of the problem have slightly lower FDRs than those that consider each center individually.

The Kidney Epidemiology and Cost Center (2018) assigns ratings of five stars down to one star to centers in the proportions $\{0.22, 0.30, 0.35, 0.09, 0.04\}$, respectively. We will abbreviate these ratings to the conventional academic scale of A–F. To illustrate the conflict between the selection criteria, we plot in Figure 9 the centers selected for the grade A (five stars, which consists of 22% of the centers that suppose to have their true mortality rate being the lowest) category with and without FDR control. Centers are characterized by pairs, (T_i, w_i), consisting of their weighted mean standarized mortality, T_i, and their estimate of the precision, w_i, of these mortality estimates. In each plot, the solid curves

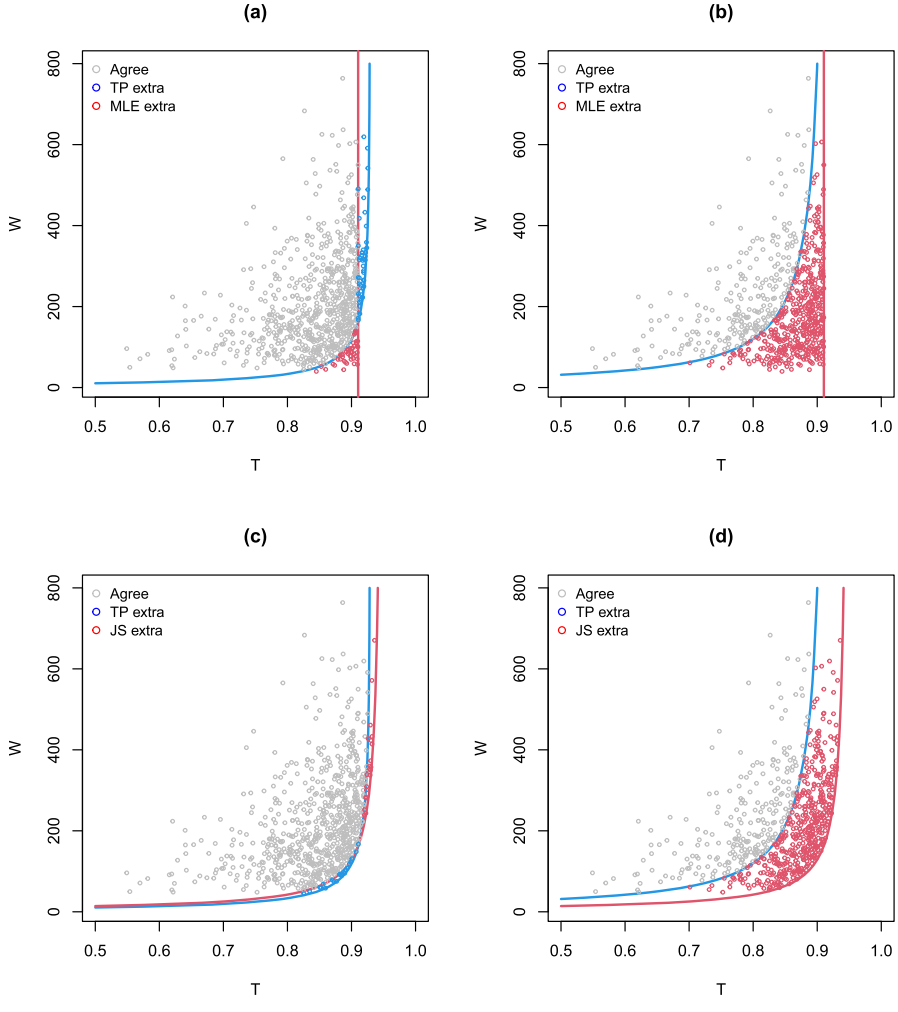

FIGURE 9.—Contrasting selections for A-rated centers: The two upper panels compare posterior tail probability selection with MLE (fixed effects) selection, while the lower panels compare TP selection with James–Stein (linear shrinkage) selection. Left panels impose capacity control only, while the right panels impose 20 percent FDR control for the TP rule. The estimated FDR rate for both the MLE and James–Stein selection under capacity constraint, using the smoothed NPMLE estimator for G, is 0.431. Comparisons are based on the 2004–2006 data.

represent the decision boundaries of the selection rule under comparison. Centers with low mortality and relatively high precision appear toward the northwest in each figure.

Panel (a) of the figure compares the posterior tail probability selection with the MLE, or fixed effect, selection. The selection boundary for the MLE is the (red) vertical line, since the MLE ignores the precision of the estimates entirely. The selection boundary for the tail probability rule is indicated by the (blue) curve. A few centers with high precision excluded by the MLE rule are selected by the TP rule, and on the contrary, a few centers with low precision are selected by the MLE rule but excluded by the TP criterion. Panel (b) imposes FDR control with $\gamma = 0.20$ on the TP selection with an estimated thresholding value implied by the FDR constraint using the smoothed NPMLE. The MLE selec-

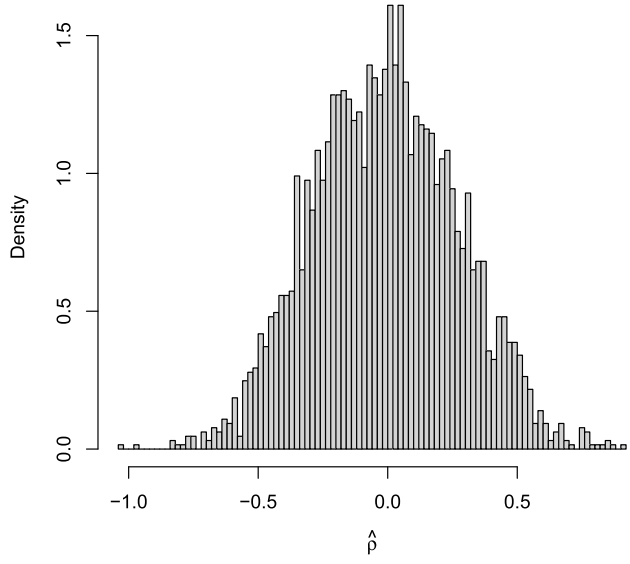

FIGURE 10.—Histogram of estimated AR(1) coefficients for 3230 dialysis centers based on annual data 2004–2017.

tion is the same as in Panel (a) without the FDR control. We see that under TP rule with FDR control, the number of selected centers is reduced considerably. Instead of selecting 711 centers allowed by the capacity constraint, it selects only 230 centers. In comparison, the MLE rule under capacity constraint has an estimated FDR rate at 0.431. Panel (c) compares centers selected by the TP rule with those selected by a James–Stein linear shrinkage rule. Now the TP rule tolerates a few more low precision centers, while it is the James–Stein rule that demands higher precision to be selected. Finally, in Panel (d), we again subject the TP rule to FDR control of 20 percent, while the James–Stein rule continues to adhere only to the capacity constraint. The TP boundary scales back substantially, suggesting that a large proportion of the extra selections made by James–Stein linear shrinkage rules are likely to be false discoveries. In fact, the estimated FDR rate of the James–Stein rule under just capacity constraint is also 0.431, the same as that of the MLE rule.

Given the longtitudinal structure of the dialysis data, it would be possible to consider the models in Section 5 that allow for unobserved variance heterogeneity. We refrain from doing so partly due to space considerations and because we are reluctant to assume stationarity of random effects over longer time horizons.

8.1. *Temporal Stability, Ranking, and Selection*

Given the longitudinal nature of the data, it is natural to ask, "How stable are rankings over time, and isn't there some temporal dependence in the observed data that should be accounted for?" Perhaps surprisingly, the year-to-year dependence in the observed mortality is quite weak. In Figure 10, we plot a histogram of estimated AR(1) coefficients for the 3230 centers; it is roughly centered at zero and slightly skewed to the left. We do not draw the conclusion from this that there is no temporal dependence in the observed y_{it}, but only that there is considerable heterogeneity in the nature of this dependence, with roughly as many centers exhibiting negative serial dependence as those with positive

TABLE IV

ESTIMATED FIRST-ORDER MARKOV TRANSITION MATRIX: ENTRY i, j OF THE MATRIX ESTIMATES THE PROBABILITY OF A TRANSITION FROM STATE i TO STATE j BASED ON POSTERIOR TAIL PROBABILITY RANKINGS FOR 3-YEAR LONGITUDINAL GROUPING OF THE CENTER DATA.

	A	B	C	D	F
A	0.441	0.328	0.201	0.024	0.006
B	0.247	0.360	0.327	0.059	0.007
C	0.122	0.286	0.440	0.112	0.040
D	0.062	0.181	0.441	0.210	0.106
F	0.021	0.085	0.346	0.219	0.329

dependence. Our approach of considering brief, 3–5-year, windows of presumed stability in center performance is consistent with the procedures of the official ranking agency. In each of these windows, we can compute a ranking according to one of the criteria introduced above, and it is of interest to see how much stability there is in these rankings.

To address this question, we consider rankings based on the posterior tail probability criterion for 3-year windows. In each of the five 3-year windows, we assign centers letter grades, A–F, with proportions $\{0.22, 0.30, 0.35, 0.09, 0.04\}$, respectively. Table IV reports the estimated transition matrix between these categories, so entry i, j in the matrix represents the estimated probability of a center in state i moving to state j in the next period.

It is obviously difficult to maintain an "A" rating for more than a couple of periods, but centers with poor performance are also likely to move into the middle of the rankings. Although, as we have seen, there is no guarantee that the posterior tail probability criterion yields a nested ranking, nestedness does hold in this particular application. Posterior mean ranking yields similar transition behavior. The high degree of mobility between rating categories reinforces our conclusion that ranking and selection into rating categories is subject to considerable uncertainty.

9. CONCLUSIONS

Robbins's compound decision framework is well suited to ranking and selection problems, and nonparametric maximum likelihood estimation of mixture models offers a powerful tool for implementing empirical Bayes rules for such problems. Posterior tail probability selection rules perform better than posterior mean rules when precision is heterogeneous. Ranking and selection is especially difficult in Gaussian settings where classical linear shrinkage methods are most appropriate. Nonparametric empirical Bayes methods can substantially improve upon selection methods based on linear shrinkage and traditional p-values when the latent mixing distribution is not Gaussian in terms of both power and false discovery rate.

REFERENCES

ANDREWS, ISAIAH, TORU KITAGAWA, AND ADAM MCCLOSKEY (2020): "Inference on Winners," Available at https://www.cemmap.ac.uk/publication/inference-on-winners-4. [2]

ARMSTRONG, TIMOTHY B., MICHAL KOLESÁR, AND MIKKEL PLAGBORG-MØLLER (2020): "Robust Empirical Bayes Confidence Intervals," Available at https://arxiv.org/abs/2004.03448. [2]

ATHEY, SUSAN, GUIDO IMBENS, JONAS METZGER, AND EVAN MUNRO (2019): "Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations," Available at http://arxiv.org/pdf/1909.02210. [5]

BAHADUR, RAGHU RAJ (1950): "On a Problem in the Theory of k Populations," *Annals of Mathematical Statistics*, 21, 362–375. [1]

BAHADUR, RAGHU RAJ, AND HERBERT ROBBINS (1950): "The Problem of the Greater Mean," *The Annals of Mathematical Statistics*, 21, 469–487. [1]

BASU, PALLAVI, T. TONY CAI, KIRANMOY DAS, AND WENGUANG SUN (2018): "Weighted False Discovery Rate Control in Large-Scale Multiple Testing," *Journal of the American Statistical Association*, 113, 1172–1183. [11, 12]

BECHHOFER, ROBERT E. (1954): "A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations With Known Variances," *The Annals of Mathematical Statistics*, 25, 16–39. [2]

BECHHOFER, ROBERT E., JACK KIEFER, AND MILTON SOBEL (1968): *Sequential Identification and Ranking Procedures*. University of Chicago Press. [2]

BENJAMINI, YOAV, AND YOSEF HOCHBERG (1995): "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of Royal Statistical Society, Series B*, 57, 289–300. [11]

BERGER, JAMES O., AND JOHN DEELY (1988): "A Bayesian Approach to Ranking and Selection of Related Means With Alternatives to Analysis-of-Variance Methodology," *Journal of the American Statistical Association*, 83, 364–373. [2]

BOYD, STEPHEN, CORINNA CORTES, MEHRYAR MOHRI, AND ANA RADOVANOVIC (2012): "Accuracy at the Top," in *Advances in Neural Information Processing Systems 25*, ed. by Fernando Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger. Curran Associates, Inc., 953–961. [2]

CAO, HONGYUAN, WENGUANG SUN, AND MICHAEL R. KOSOROK (2013): "The Optimal Power Puzzle: Scrutiny of the Monotone Likelihood Ratio Assumption in Multiple Testing," *Biometrika*, 100, 495–502. [25]

CHETTY, RAJ, AND NATHANIEL HENDREN (2018): "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates," *The Quarterly Journal of Economics*, 133, 1163–1228. [28]

CHETTY, RAJ, JOHN N. FRIEDMAN, AND JONAH E. ROCKOFF (2014a): "Measuring the Impacts of Teachers i: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104, 2593–2632. [28, 31]

——— (2014b): "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104, 2633–2679. [28,31]

EFRON, BRADLEY (2011): "Tweedie's Formula and Selection Bias," *Journal of the American Statistical Association*, 106, 1602–1614. [6]

——— (2016): "Empirical Bayes Deconvolution Estimates," *Biometrika*, 103, 1–20. [5]

——— (2019): "Bayes, Oracle Bayes and Empirical Bayes," *Statistical Science*, 34, 177–201. [6]

EFRON, BRADLEY, AND CARL MORRIS (1973): "Stein's Estimation Rule and Its Competitiors—an Empirical Bayes Approach," *Journal of the American Statistical Association*, 68, 117–130. [18,29,31,35]

EFRON, BRADLEY, ROBERT TIBSHIRANI, JOHN STOREY, AND VIRGINIA TUSHER (2001): "Empirical Bayes Analysis of Microarray Experiments," *J. American Statistical Association*, 96, 1151–1160. [6,12]

GELMAN, ANDREW, AND PHILLIP N. PRICE (1999): "All Maps of Parameter Estimates Are Misleading," *Statistics in Medicine*, 18, 3221–3234. [3]

GENOVESE, CHRISTOPHER, AND LARRY WASSERMAN (2002): "Operating Characteristic and Extensions of the False Discovery Rate Procedure," *Journal of the Royal Statistical Society, Series B*, 64, 499–517. [11]

GILRAINE, MICHAEL, JIAYING GU, AND ROBERT MCMILLAN (2020): "A New Method for Estimating Teacher Value-Added," NBER Working Paper Series Number 27094. [2,30]

GOEL, PREM K., AND HERMAN RUBIN (1977): "On Selecting a Subset Containing the Best Population—a Bayesian Approach," *The Annals of Statistics*, 5, 969–983. [2]

GOLDSTEIN, HARVEY, AND DAVID J. SPIEGELHALTER (1996): "League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance, (With Discussion)," *Journal of the Royal Statistical Society: Series A*, 159, 385–443. [3]

GU, JIAYING, AND ROGER KOENKER (2017): "Empirical Bayesball Remixed: Empirical Bayes Methods for Longitudinal Data," *Journal of Applied Econometrics*, 32, 575–599. [24]

——— (2016a): "Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective," *Journal of Economic and Business Statistics* (forthcoming). [6]

——— (2016b): "On a Problem of Robbins," *International Statistical Review*, 84, 224–244. [2]

——— (2023): "Supplement to 'Invidious Comparisons: Ranking and Selection as Compound Decisions'," *Econometrica Supplemental Material*, 91, https://doi.org/10.3982/ECTA19304. [3]

GUO, XINZHOU, AND XUMING HE (2020): "Inference on Selected Subgroups in Clinical Trials," *Journal of the American Statistical Association*, 1–19. [2]

GUPTA, SHANTI (1956): "On a Decision Rule for a Problem in Ranking Means," Mimeograph Series No. 150, Institute of Statistics, University of North Carolina, Chapel Hill. [2]

GUPTA, SHANTI S., AND SUBRAMANIAN PANCHAPAKESAN (1979): *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*. Wiley. [2]

HANUSHEK, ERIC A. (2011): "The Economic Value of Higher Teacher Quality," *Economics of Education review*, 30, 466–479. [31]

HECKMAN, JAMES, AND BART SINGER (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52, 63–132. [5]

HENDERSON, NICHOLAS C., AND MICHAEL A. NEWTON (2016): "Making the Cut: Improved Ranking and Selection for Large-Scale Inference," *Journal of the Royal Statistical Society, Series B*, 78 (4), 781–804. [6, 16]

JIANG, WENHUA (2020): "On General Maximum Likelihood Empirical Bayes Estimation of Heteroscedastic iid Normal Means," *Electronic Journal of Statistics*, 14, 2272–2297. [27]

KIDNEY EPIDEMIOLOGY AND COST CENTER (2018): "Technical Notes on the Dialysis Facility Compare Quality of Patient Care Star Rating Methodology for the October 2018 Release," University of Michigan, School of Public Health. [35]

KIEFER, JACK, AND JACK WOLFOWITZ (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics*, 27, 887–906. [1,5, 27]

KLINE, PATRICK, AND CHRISTOPHER WALTERS (2021): "Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination," *Econometrica*, 89, 565–592. [2]

KOENKER, ROGER, AND JIAYING GU (2015–2021): "REBayes: An R Package for Empirical Bayes Methods," Available at https://cran.r-project.org/package=REBayes. [6]

KOENKER, ROGER (2020): "Empirical Bayes Confidence Intervals: An R Vinaigrette," Available at http://www.econ.uiuc.edu/~roger/research/ebayes/cieb.pdf. [6]

KOENKER, ROGER, AND JIAYING GU (2017): "REBayes: An R Package for Empirical Bayes Mixture Methods," *Journal of Statistical Software*, 82, 1–26. [6]

——— (2019): "Comment: Minimalist *G*-Modeling," *Statistical Science*, 34, 209–213. [2]

KOENKER, ROGER, AND IVAN MIZERA (2014): "Convex Optimization, Shape Constraints, Compound Decisions and Empirical Bayes Rules," *J. of Am. Stat. Assoc.*, 109, 674–685. [5]

LAIRD, NAN M., AND THOMAS A. LOUIS (1989): "Empirical Bayes Ranking Methods," *Journal of Educational Statistics*, 14, 29–46. [2]

——— (1991): "Smoothing the Non-Parametric Estimate of a Prior Distribution by Roughening," *Computational Statistics & Data Analysis*, 12, 27–37. [5]

LIN, RONGHENG, THOMAS A. LOUIS, SUSAN M. PADDOCK, AND SUSAN M. RIDGEWAY (2009): "Ranking US-RDS Provider Specific SMRs From 1998–2001," *Health Service Outcomes Research Methodology*, 9, 22–38. [32]

——— (2006): "Loss Function Based Ranking in Two-Stage, Hierarchical Models," *Bayesian Analysis*, 1, 915–946. [8,32]

LINDSAY, BRUCE G. (1995): "Mixture Models: Theory, Geometry and Applications," in *NSF-CBMS Regional Conference Series in Probability and Statistics*. [5]

MOGSTAD, MAGNE, JOSEPH ROMANO, AZEEM SHAIKH, AND DANIEL WILHELM (2020): "Inferences for Ranks With Applications to Mobility Across Neighborhoods and Academic Achievement Across Countries," Preprint. [2]

POLYANSKIY, YURY, AND YIHONG WU (2020): "Self-Regularizing Property of Nonparametric Maximum Likelihood Estimator in Mixture Models," Preprint. [5,27]

PORTNOY, STEPHEN (1982): "Maximizing the Probability of Correctly Ordering Random Variables Using Linear Predictors," *Journal of Multivariate Analysis*, 12, 256–269. [2]

ROBBINS, HERBERT (1950): "A Generalization of the Method of Maximum Likelihood: Estimating a Mixing Distribution (Abstract)," *The Annals of Mathematical Statistics*, 21, 314–315. [5]

——— (1951): "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems," in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I. University of California Press: Berkeley, 131–149. [4,5]

——— (1956): "An Empirical Bayes Approach to Statistics," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I. University of California Press: Berkeley, 157–163. [1,6, 25]

SAHA, SUJAYAM, AND ADITYANAND GUNTUBOYINA (2020): "On the Nonparametric Maximum Likelihood Estimator for Gaussian Location Mixture Densities With Application to Gaussian Denoising," *Annals of Statistics*, 48, 738–762. [27]

STOREY, JOHN D. (2002): "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical B*, 64, 479–498. [6]

SUN, WENGUANG, AND ALEXANDER C. MCLAIN (2012): "Multiple Testing of Composite Null Hypotheses in Heteroscedastic Models," *Journal of the American Statistical Association*, 107, 673–687. [14]

SUN, WENGUANG, AND T. TONY CAI (2007): "Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control," *Journal American Statistical Association*, 102, 901–912. [12]

UNIVERSITY OF MICHIGAN KIDNEY EPIDEMIOLOGY AND COST CENTER (2009–2019)): "Dialysis Facility Reports," Available at https://data.cms.gov/dialysis-facility-reports. [33]

VAN DE GEER, SARA (1993): "Hellinger-Consistency of Certain Nonparametric Maximum Likelihood Estimators," *The Annals of Statistics*, 14–44. [27]

WALD, ABRAHAM (1950): *Statistical Decision Functions*. Wiley. [4]

Editor Guido Imbens handled this manuscript as an invited Walras–Bowley lecture. The invitation to deliver the Walras–Bowley lecture is also an invitation to publish a suitable version of the lecture in Econometrica.

Manuscript received 28 December, 2020; final version accepted 30 August, 2021; available online 16 June, 2022.