

INVIDIOUS COMPARISONS: RANKING AND SELECTION AS COMPOUND DECISIONS

JIAYING GU AND ROGER KOENKER

ABSTRACT. 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.

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, $\theta^* = \max_i\{\theta_1, \dots, \theta_K\}$, was formulated as choosing weights z_1, \dots, z_K to minimize,

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

Bahadur showed that among “impartial decision rules,” i.e. 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 $\bar{X}_i = \max\{\bar{X}_1, \dots, \bar{X}_K\}$ 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 in Bahadur and Robbins (1950) who focused on the two-sample setting but relaxed the common variance

Version: September 15, 2021. 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 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.

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 Pan-chapakesan (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 non-parametric 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 (2016a), 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 employ 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 U.S. 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) propose 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) propose 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 study

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) study 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 focus 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 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 is 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.

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, $Y_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, \dots, n$ with means θ_i taking either the value $+1$ or -1 . We are asked to estimate the n -vector $\theta = (\theta_1, \dots, \theta_n)$ 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 shows that the minimax decision rule is $\delta(y) = \text{sgn}(y)$; 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 goes 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,

$$\delta_p(y) = \text{sgn}(y - \frac{1}{2} \log((1-p)/p)).$$

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

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

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),

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

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 non-decreasing 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, i.e. 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 address this concern: they note 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, e.g. 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 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) and studied in more detail in Jiang and Zhang (2021). 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,

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

Efron (2011) refers to this as Tweedie's formula, it appears in Robbins (1956) credited to M.C.K. Tweedie. Appendix A of Gu and Koenker (2016b) 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, $y_i : i = 1, 2, \dots, n$ 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 $H_0 : \theta < \theta_\alpha$ vs. $H_A : \theta \geq \theta_\alpha$. 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,

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

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_α can be found such that $\mathbb{P}(\delta(Y) = 1) = \mathbb{P}(v_\alpha(Y) \geq \lambda_\alpha/(1 + \lambda_\alpha)) = \alpha$.

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,

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

which is equivalent to

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

Ignoring the first term since it doesn't depend upon δ , the Oracle Bayes rule, becomes, choose $\delta(y) = 1$ if,

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

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 are 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, $\delta_{\alpha}(Y) = 1(Y \geq \eta_{\alpha})$, 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,

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

The marginal false non-discovery rate is,

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

Figure 3.1 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 discoveries. When $\alpha = 0.10$ the proportion of selected observations with θ below the threshold θ_{α} is slightly greater than 50 percent.

	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

TABLE 3.1. 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.

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 3.1 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 variances of alpha increases

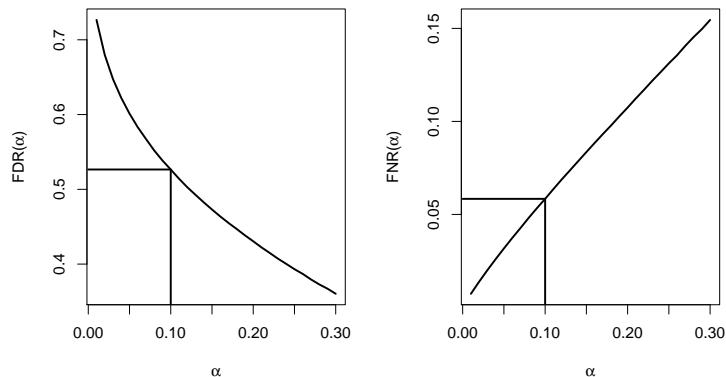


FIGURE 3.1. False Discovery Rates and False Non-Discovery Rates for a Standard Gaussian Mixing Distribution.

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.

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 3.1. 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,

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

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 over-looked 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,

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

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:

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

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,

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

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

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

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,

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

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.

Remark. 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 $\mathbb{P}[\delta_i = 0, \theta_i \geq \theta_\alpha]$ subject to the constraint that $\mathbb{P}[\delta_i = 1, \theta_i < \theta_\alpha] / \mathbb{P}[\delta_i = 1] \leq \gamma$ and $\mathbb{P}[\delta_i = 1] \leq \alpha$. 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 , $\mathbb{P}[\delta_i = 1 | \theta_i \geq \theta_\alpha]$, 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, $\delta_i^* = \mathbb{1}\{v_\alpha(y_i) \geq \lambda^*(\alpha, \gamma)\}$ where $\lambda^*(\alpha, \gamma) = v_\alpha(t^*)$ with $t^* = \max\{t_1^*, t_2^*\}$,*

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

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

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 is 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, Tibshirani, Storey, and Tusher (2001), Sun and Cai (2007) and Basu, Cai, Das, and Sun (2018), typically focuses on the null hypothesis of $H_{0i} : \theta_i = 0$. When variances are homogeneous, it doesn't 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,

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

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) ,

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

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

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

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

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

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

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

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

and $t_{\alpha}(\lambda, \sigma)$ 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,

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

Instead of transforming the data by v_{α} , we consider the transformation,

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

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 $\delta_i^{\mathbb{T}} = \mathbb{1}\{\mathbb{T}(y_i, \sigma_i) \geq \lambda\}$ then chooses the cutoff value λ that respects both the capacity constraint and the FDR control constraint for selecting the top α proportion.

Figure 4.1 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

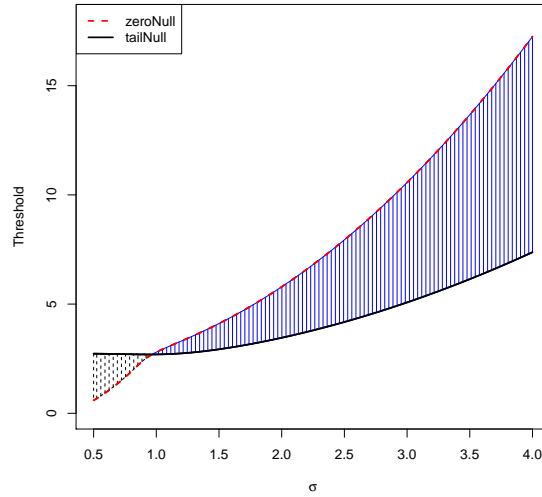


FIGURE 4.1. Selection boundaries based on the model (4.1) with $\alpha = 0.05$ and $\gamma = 0.1$. The solid black curve corresponds to the boundary of the selection region based on transformation v_α . The dash red curve corresponds to the boundary of the selection region based on transformation T .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

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 *can not* 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 transformation v_α and T differ 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 δ , then $\beta(\delta^T) = 39\%$ and $\beta(\delta^*) = 69\%$. 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, $\alpha_1 > \alpha_0$ 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 several subgroups of the centers.

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

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

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

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

It is also convenient for later discussion to define

$$\Omega_{\alpha, \gamma}^{FDR} := \{(y, \sigma) : v_{\alpha}(y, \sigma) \geq \lambda_1^*(\alpha, \gamma)\} = \{(y, \sigma) : y \geq t_{\alpha}(\lambda_1^*(\alpha, \gamma), \sigma)\}$$

$$\Omega_{\alpha}^C := \{(y, \sigma) : v_{\alpha}(y, \sigma) \geq \lambda_2^*(\alpha)\} = \{(y, \sigma) : y \geq t_{\alpha}(\lambda_2^*(\alpha), \sigma)\}$$

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

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

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

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 $\Omega_{\alpha_2, \gamma}^{FDR} \subseteq \Omega_{\alpha_1, \gamma}^{FDR}$ for any $\alpha_1 > \alpha_2$.*

Remark. The condition in Lemma 4.2 is sufficient but not necessary for nestedness of $\Omega_{\alpha, \gamma}^{FDR}$. 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 $\alpha_1 > \alpha_2$, $t_{\alpha_1}(\lambda_2^*(\alpha_1), \sigma) \leq t_{\alpha_2}(\lambda_2^*(\alpha_2), \sigma)$ for each σ , then $\Omega_{\alpha_2}^C \subseteq \Omega_{\alpha_1}^C$.*

Remark. The monotonicity here coincides with the condition in Theorem 3 of Henderson and Newton (2016), who demonstrate that it holds when G is Gaussian. However, it need not hold as shown in our counter-example in Section C. ■

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,

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

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,

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

Integrating out σ we have the marginal density of v ,

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

The capacity constraint is $\mathbb{P}(v \geq \lambda_2^*) = \alpha$, with cut-off 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 cut-off 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,

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

while FDR control requires a thresholding value that solves,

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

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 $\{(y, \sigma) : y \geq \omega^*/\rho\}$.

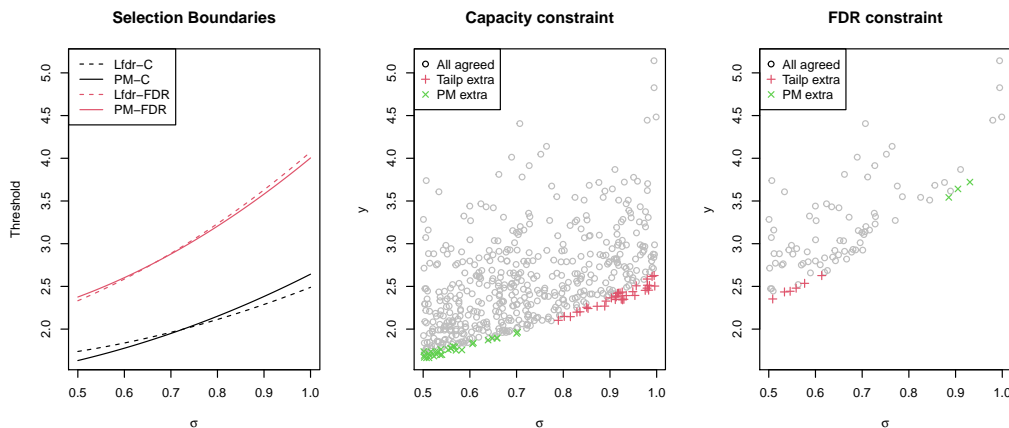


FIGURE 4.2. The left panel plots the selection boundaries for the normal-normal model with $\sigma_\theta^2 = 1$ and $\alpha = 0.05$ and $\gamma = 0.2$. The density of σ is assumed to be uniform on the range $[0.5, 1]$. Selected units must have (y_i, σ_i) above the curves. The red curves correspond to the selection region boundaries with FDR controlled at level 0.2; solid lines for posterior mean ranking and dash 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 grey 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.

Figure 4.2 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 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 3.1 that FDR was also very high when α is set at 0.05. Figure 4.2 also depicts the selected set with a realized sample of 10,000 from the normal-normal model. With capacity constraint alone, the posterior mean criteria favours 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 4.2. 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,

$$\begin{aligned} f(y|\sigma^2) &= \int \varphi(y|\theta, \sigma^2)dG(\theta) \\ &= (0.85\varphi(y|1, \sigma^2) + 0.1\varphi(y|2, \sigma^2) + 0.05\varphi(y|5, \sigma^2)). \end{aligned}$$

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 4.3 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 procedure, 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 4.3 again shows the selected observations and once more we see that the posterior mean criteria favours individuals with smaller variances, both under 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.

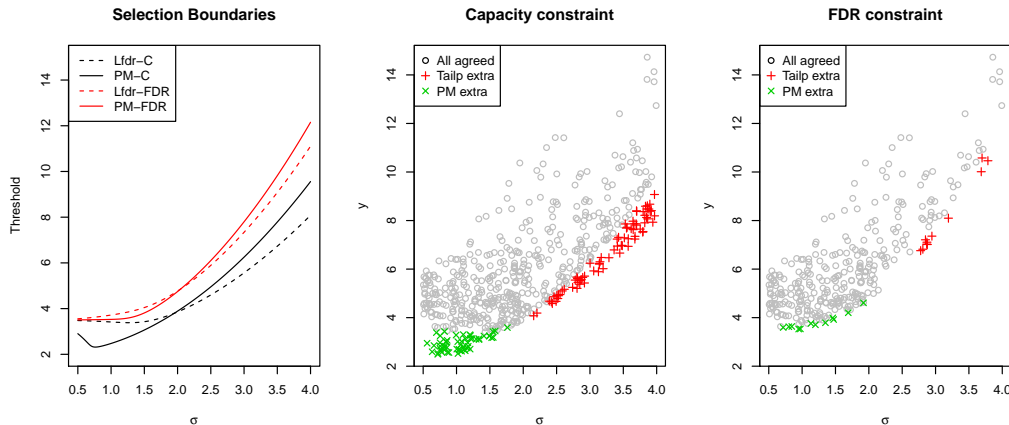


FIGURE 4.3. The left panel plots the selection boundaries for the normal-discrete model with $\theta \sim G = 0.85\delta_{-1} + 0.1\delta_2 + 0.05\delta_5$ and $\alpha = 0.05$ and $\gamma = 0.1$. The density of σ is uniform on the range $[0.5, 4]$. The red curves correspond to the selection region with 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.

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,

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

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,

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

with

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

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

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

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*

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

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

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

and

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

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

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

Remark. Note that for each prespecified 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, $\text{NIX}(\theta_0, \kappa_0, \nu_0, \sigma_0^2) = \mathcal{N}(\theta | \theta_0, \sigma^2 / \kappa_0) \chi^{-2}(\sigma^2 | \nu_0, \sigma_0^2)$. 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 $NIX(\theta_T, \kappa_T, \nu_T, \sigma_T^2) = \mathcal{N}(\theta | \theta_T, \sigma_T^2 / \kappa_T) \chi^{-2}(\sigma^2 | \nu_T, \sigma_T^2)$ with

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

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 5.1 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 behaviour 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 a increasing function of S , while for y_0 large, $v_\alpha(y_0, S)$ becomes a decreasing function of S .

A capacity constraint of size α implies the thresholding rule,

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

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

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

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 5.1. 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 5.2 further shows the comparison of the selected set based on a sample realization from the model.

In Appendix B 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 heteroskedasticity that should be allowed in ϵ_{it} . Here we consider a variant where

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

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”

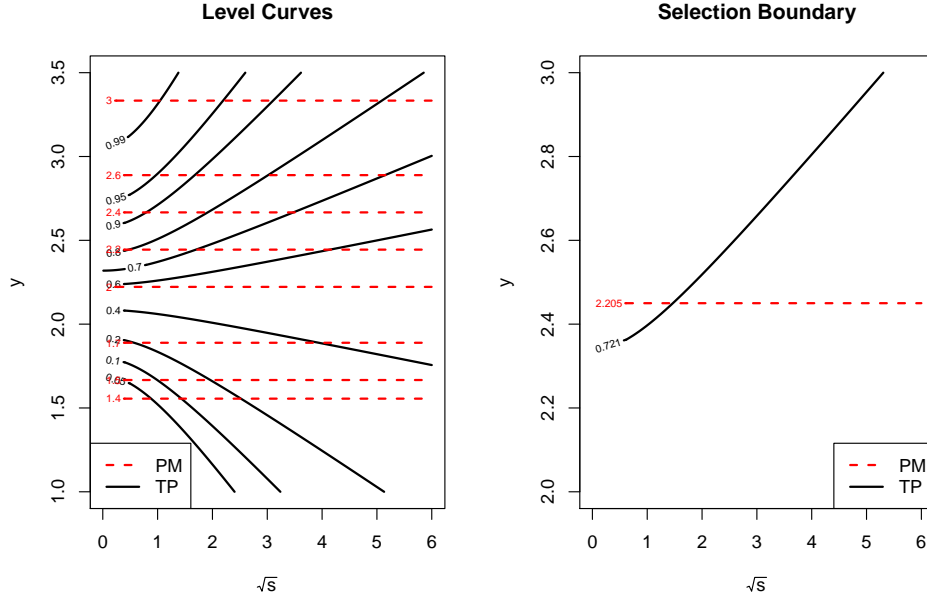


FIGURE 5.1. The left panel shows level curves of the posterior mean (marked as red dashed lines) and posterior tail probability (marked as black solid lines) for the normal model with $(\theta, \sigma^2) \sim NIX(0, 1, 6, 1)$ and panel time dimension $T = 9$. The right plot shows the boundary of the 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 $\gamma = 10\%$.

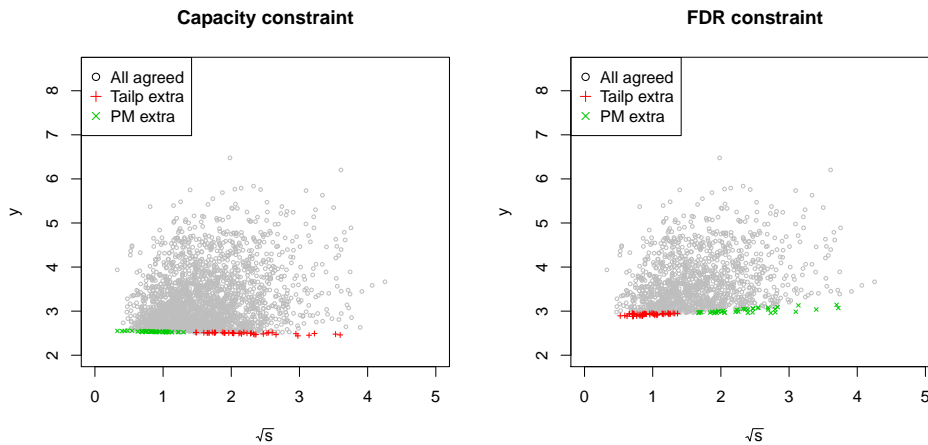


FIGURE 5.2. Selection set comparison for one sample realization from the normal model with $(\theta, \sigma^2) \sim NIX(0, 1, 6, 1)$ and panel time dimension $T = 9$: The left panel shows in grey circles the agreed selected elements by both the posterior mean and posterior tail probability criteria under 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 FDR constraint with $\alpha = 5\%$ and $\gamma = 10\%$.

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

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

and the posterior mean takes the form

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

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

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

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

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

which can be further represented as

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

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 region as those plotted in Figure 5.1 and Figure B.1 to appreciate how different decision criteria determines the selection. As soon as the ranking statistics depends 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 $\delta^* = \mathbb{1}\{v_\alpha \geq \lambda^*\}$ 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 $[\underline{\sigma}^2, \bar{\sigma}^2]$ respectively for some $\underline{\sigma}^2 > 0$ and $\bar{\sigma}^2 < +\infty$.

(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

$$\begin{aligned}\lambda_1^* &= \inf\{t \in (0, 1), H_v(t) \geq 1 - \alpha\} \\ \lambda_2^* &= \inf\{t \in (0, 1), Q(t) \leq \gamma\}\end{aligned}$$

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 discuss 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 easy to see that $\mathbb{P}[\delta_i = 1, \theta_i < \theta_\alpha] = (1 - \alpha) \int 1\{v_{\alpha,i} \geq t\} f_0(D_i) dD_i = \int 1\{v_{\alpha,i} \geq t\} (1 - v_{\alpha,i}) f(D_i) dD_i = \mathbb{E}[(1 - v_{\alpha,i})1\{v_{\alpha,i} \geq t\}]$. 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:

$$\begin{aligned}H_n(t) &= \frac{1}{n} \sum_{i=1}^n 1\{v_{\alpha,i} \leq t\} \\ Q_n(t) &= \frac{\sum_{i=1}^n (1 - v_{\alpha,i}) 1\{v_{\alpha,i} \geq t\}}{\sum_{i=1}^n 1\{v_{\alpha,i} \geq t\}}\end{aligned}$$

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

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

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 $n \rightarrow \infty$, more specifically*

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

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 , $\hat{G}_n(k) \rightarrow G(k)$ almost surely as $n \rightarrow \infty$. 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 is then defined respectively as

$$\begin{aligned}\hat{v}_\alpha(y_i, \sigma_i) &= \int_{\hat{\theta}_\alpha}^{+\infty} \varphi(y_i | \theta, \sigma_i^2) d\hat{G}_n(\theta) / \int_{-\infty}^{+\infty} \varphi(y_i | \theta, \sigma_i^2) d\hat{G}_n(\theta) \\ \hat{v}_\alpha(y_i, s_i) &= \int_{\hat{\theta}_\alpha}^{+\infty} f(y_i, s_i | \theta, \sigma) d\hat{G}_n(\theta, \sigma^2) / \int_{-\infty}^{+\infty} f(y_i, s_i | \theta, \sigma) d\hat{G}_n(\theta, \sigma^2)\end{aligned}$$

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,

$$\hat{\lambda}_{1n} = \inf\{t \in [0, 1] : \frac{1}{n} \sum_{i=1}^n 1\{\hat{v}_{\alpha,i} \leq t\} \geq 1 - \alpha\}$$

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

Theorem 6.2. *(Asymptotic validity of adaptive procedure) Under Assumptions 1 and 2, the adaptive procedure $\delta_i = 1\{\hat{v}_{\alpha,i} \geq \hat{\lambda}_n\}$ 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 $n \rightarrow \infty$, more specifically*

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

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

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} as in Jiang and Zhang (2021) with a bandwidth slowly tending to zero with n . As they show, the smoothed mixing distribution is also consistent, hence fulfilling Assumption 2 and therefore all our adaptivity results.

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

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

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 estimates 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 7.1, where we plot empirical power against degrees of freedom of the t distribution for a selected set of values for the capacity

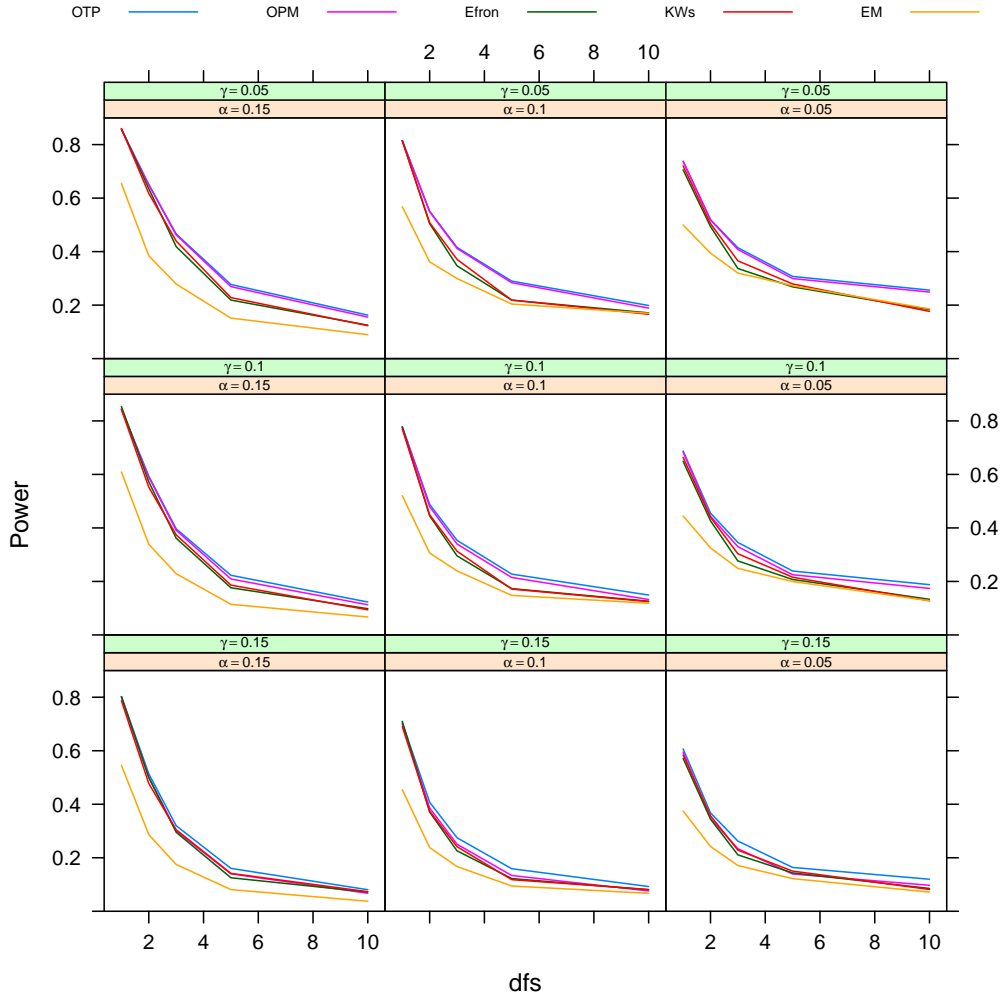


FIGURE 7.1. Power Performance for Several Selection Rules with Student t Signal. Capacity and FDR constraints are indicated at the top of each panel in the Figure.

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 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-added 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 this 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 7.2. Variances, in keeping with our hypothesis in Section 4, are drawn from a distribution with density illustrated in the right panel of Figure 7.2. 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)).

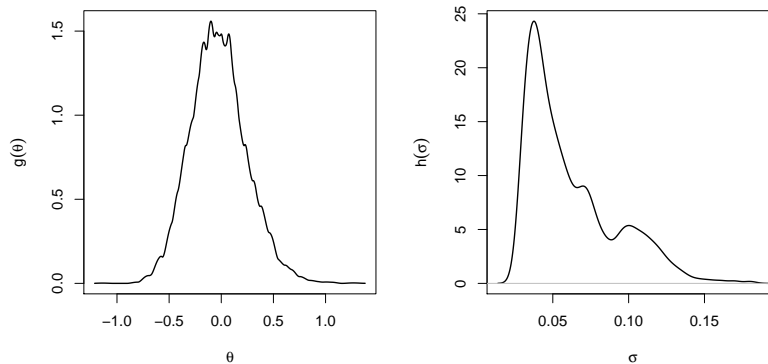


FIGURE 7.2. Densities of “Latent” (Mean) Ability and Standard Deviation for the Teacher Value Added Simulations

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 7.1 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.

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 U.S 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 2004–18 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

	$\gamma = 5\%$				$\gamma = 10\%$			
	$\alpha = 1\%$	$\alpha = 3\%$	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 1\%$	$\alpha = 3\%$	$\alpha = 5\%$	$\alpha = 10\%$
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

TABLE 7.1. Comparison of Performance of Several Selection Rules for the Teacher Value Added Simulation.

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

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

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 8.1 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 - 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 NPMLE, \hat{G} . The assessment of FDR reported in Table 8.1 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 8.1 the centers selected for the grade A (Five star, which consists 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 standardized mortality, T_i , and their estimate of the precision, w_i , of these mortality estimates. In each plot the solid curves 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

	$\alpha = 4\%$	$\alpha = 10\%$	$\alpha = 15\%$	$\alpha = 20\%$	$\alpha = 25\%$
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
Pvalue	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
Pvalue	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

TABLE 8.1. FDR Estimates: 2004-2006

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 selection 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 selected centers 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 longitudinal 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 8.2 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

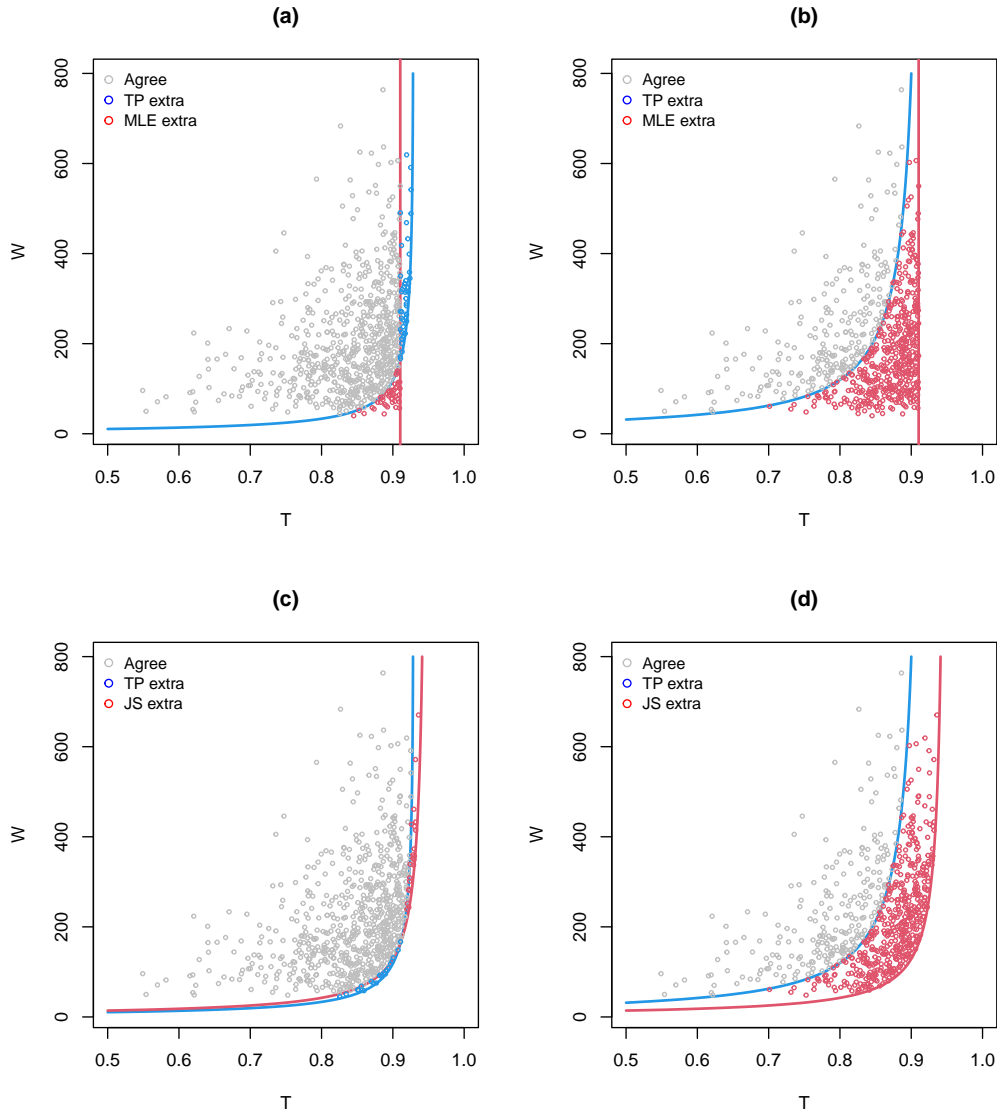


FIGURE 8.1. Contrasting Selections for A-rated centers: The two upper panels compare posterior tail probability selection with MLE (fixed effects) selection, while the lower panels compare TP selection with James-Stein (linear shrinkage) selection. Left panels impose capacity control only, while the 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.

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 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

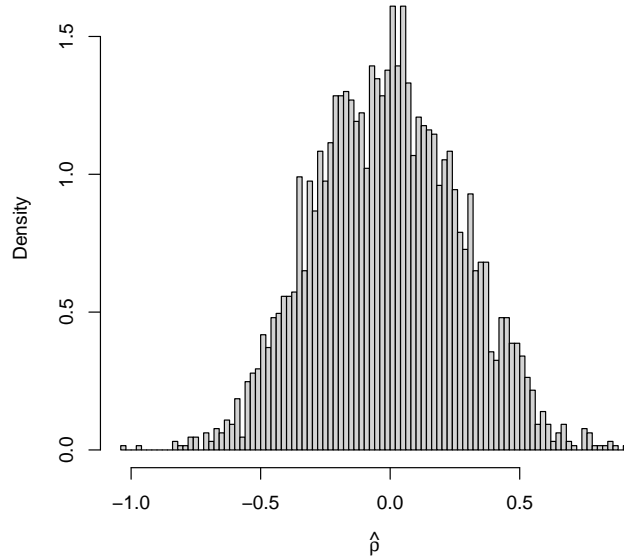


FIGURE 8.2. Histogram of estimated AR(1) coefficients for 3230 Dialysis centers based on annual data 2004-2017

	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

TABLE 8.2. 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.

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 three year windows. In each of the 5 3-year windows we assign centers letter grades, A-F, with proportions $\{0.22, 0.30, 0.35, 0.09, 0.04\}$ respectively. Table 8.2 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 reenforces 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 both in terms of power and false discovery rate.

APPENDIX A. PROOFS

Proof. [Lemma 3.1] We can write:

$$\begin{aligned}
\nabla_y \frac{\alpha f_1(y)}{f(y)} &= \frac{\int_{[\theta_\alpha, +\infty)} \nabla_y \log \varphi(y|\theta, \sigma^2) \varphi(y|\theta, \sigma^2) dG(\theta)}{\int_{(-\infty, +\infty)} \varphi(y|\theta, \sigma^2) dG(\theta)} \\
&\quad - \frac{\int_{[\theta_\alpha, +\infty)} \varphi(y|\theta, \sigma^2) dG(\theta)}{\int_{(-\infty, +\infty)} \varphi(y|\theta, \sigma^2) dG(\theta)} \frac{\int_{(-\infty, +\infty)} \nabla_y \log \varphi(y|\theta, \sigma^2) \varphi(y|\theta, \sigma^2) dG(\theta)}{\int_{(-\infty, +\infty)} \varphi(y|\theta, \sigma^2) dG(\theta)} \\
&= \mathbb{E} \left[\mathbb{1}\{\theta \geq \theta_\alpha\} \nabla_y \log \varphi(y|\theta, \sigma^2) | Y \right] - \mathbb{E} \left[\mathbb{1}\{\theta \geq \theta_\alpha\} | Y \right] \mathbb{E} [\nabla_y \log \varphi(y|\theta, \sigma^2) | Y] \\
&= \text{Cov} \left[\mathbb{1}\{\theta \geq \theta_\alpha\}, \nabla_y \log \varphi(y|\theta, \sigma^2) | Y \right] \geq 0
\end{aligned}$$

The last inequality holds since $\nabla_y \log \varphi(y|\theta, \sigma^2)$ is increasing in θ for each fixed σ^2 , by the Gaussian assumption and the fact that covariance of monotone functions of θ is non-negative (see Schmidt (2014)) assuming the existence of $\mathbb{E}_{\theta|Y}[\nabla_y \log \varphi(Y|\theta, \sigma^2)|Y]$ which we assume. Nesting follows from the monotonicity of the $v_\alpha(y)$ criterion: monotonicity of $v_\alpha(y)$ implies that there exists t_α such that $\mathbb{1}\{v_\alpha(y) \geq \lambda_\alpha/(1 + \lambda_\alpha)\} = \mathbb{1}\{y \geq t_\alpha\}$, hence if $\alpha_1 > \alpha_2$ and $\mathbb{P}(y \geq t_{\alpha_1}) = \alpha_1$ and $\mathbb{P}(y \geq t_{\alpha_2}) = \alpha_2$, then it must be that $t_{\alpha_1} \leq t_{\alpha_2}$, implying nestedness. ■

Proof. [Lemma 3.2] Denote the three decision criteria, $v_1(y) = \mathbb{E}(\theta|Y = y)$, $v_2(y) = \mathbb{P}(\theta \geq G^{-1}(1 - \alpha)|Y = y)$ and $v_3(y) = \mathbb{E}(\theta \mathbb{1}(\theta \geq G^{-1}(1 - \alpha))|Y = y)$. Assuming that $\mathbb{E}[\theta|Y] < \infty$, $\mathbb{E}_{\theta|Y}[\nabla_y \log \varphi(y|\theta, \sigma^2)|Y] < \infty$ and $\mathbb{E}_{\theta|Y}[\theta \nabla_y \log \varphi(y|\theta, \sigma^2)|Y] < \infty$, the calculation leading to the proof of Lemma 1 shows:

$$\begin{aligned}
\nabla_y v_1(y) &= \text{Cov}(\theta, \nabla_y \log \varphi(y|\theta)|Y = y) \\
\nabla_y v_2(y) &= \text{Cov}(\mathbb{1}(\theta \geq G^{-1}(1 - \alpha)), \nabla_y \log \varphi(y|\theta)|Y = y) \\
\nabla_y v_3(y) &= \text{Cov}(\theta \mathbb{1}(\theta \geq G^{-1}(1 - \alpha)), \nabla_y \log \varphi(y|\theta)|Y = y).
\end{aligned}$$

Thus, the monotonicity of $\nabla_y \log \varphi(y|\theta, \sigma^2)$ implies they all yield identical rankings. ■

Proof. [Proposition 3.3] The Bayes rule for the non-randomized selections can be characterized as,

$$\delta_i^* = \begin{cases} 1, & \text{if } v_\alpha(y_i) \geq \tau_1^*(1 - v_\alpha(y_i) - \gamma) + \tau_2^* \\ 0, & \text{if } v_\alpha(y_i) < \tau_1^*(1 - v_\alpha(y_i) - \gamma) + \tau_2^* \end{cases}$$

with Karush-Kuhn-Tucker conditions,

$$(A.1) \quad \tau_1^* \left(\mathbb{E} \left[\sum_{i=1}^n \left\{ (1 - v_\alpha(y_i)) \delta_i^* - \gamma \delta_i^* \right\} \right] \right) = 0$$

$$(A.2) \quad \tau_2^* \left(\mathbb{E} \left[\sum_{i=1}^n \delta_i^* \right] - \alpha n \right) = 0$$

$$(A.3) \quad \mathbb{E} \left[\sum_{i=1}^n \left\{ (1 - v_\alpha(y_i)) \delta_i^* - \gamma \delta_i^* \right\} \right] \leq 0$$

$$(A.4) \quad \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n \delta_i^* \right] - \alpha \leq 0$$

$$(A.5) \quad \tau_1^* \geq 0$$

$$(A.6) \quad \tau_2^* \geq 0$$

The Bayes rule takes the form of thresholding on the posterior tail probability $v_\alpha(y)$ and since $v_\alpha(y)$ is monotone in y as shown in Lemma 3.2, it is therefore a thresholding rule on Y , $\delta_i^* = \mathbb{1}\{y_i \geq t^*\}$ with cutoff t^* depending on the values of $(\tau_1^*, \tau_2^*, \alpha, \gamma)$. Condition (A.3) is equivalent to the

condition that the marginal false discovery rate, mFDR, since it requires

$$\mathbb{E}\left[\sum_{i=1}^n\{(1-v_\alpha(y_i))\delta_i^*\}\right]/\mathbb{E}\left[\sum_{i=1}^n\delta_i^*\right]\leq\gamma$$

and we can show that the left hand side quantity is precisely the mFDR since

$$\begin{aligned}\text{mFDR}(t^*) &= \mathbb{P}(\delta_i^* = 1, \theta_i \leq \theta_\alpha) / \mathbb{P}(\delta_i^* = 1) \\ &= (1 - \alpha) \int \mathbb{1}\{y \geq t^*\} f_0(y) dy / \int \mathbb{1}\{y \geq t^*\} f(y) dy \\ &= \int \mathbb{1}\{y \geq t^*\} (1 - v_\alpha(y)) f(y) dy / \int \mathbb{1}\{y \geq t^*\} f(y) dy \\ &= \mathbb{E}\left[\sum_{i=1}^n\{(1-v_\alpha(y_i))\delta_i^*\}\right]/\mathbb{E}\left[\sum_{i=1}^n\delta_i^*\right] \\ &= \frac{\int_{-\infty}^{\theta_\alpha} \tilde{\Phi}((t^* - \theta)/\sigma) dG(\theta)}{\int_{-\infty}^{+\infty} \tilde{\Phi}((t^* - \theta)/\sigma) dG(\theta)} \leq \gamma.\end{aligned}$$

For any mixing distribution G , as t^* increases, it becomes less likely for condition (A.3) to bind. And as t^* approaches $-\infty$, left side of (A.3) approaches $1 - \alpha - \gamma$, and hence we've restricted $\gamma < 1 - \alpha$ to avoid cases where the condition (A.3) never binds. On the other hand, condition (A.4) is equivalent to

$$\mathbb{P}(\delta_i^* = 1) - \alpha = \int_{-\infty}^{+\infty} \tilde{\Phi}((t^* - \theta)/\sigma) dG(\theta) - \alpha \leq 0.$$

As t^* increases, it also becomes less likely that condition (A.4) binds. Therefore, we can define,

$$\begin{aligned}t_1^* &= \min\{t : \text{mFDR}(t) - \gamma \leq 0\} \\ t_2^* &= \min\{t : \int_{-\infty}^{+\infty} \tilde{\Phi}((t - \theta)/\sigma) dG(\theta) - \alpha \leq 0\}\end{aligned}$$

When $t_1^* < t_2^*$, the feasible region for Y defined by inequality (A.4) is a strict subset of that defined by inequality (A.3). When $t_1^* > t_2^*$, then the feasible region defined by inequality (A.3) is a strict subset of that defined by inequality (A.4). When $t_1^* = t_2^*$, the feasible regions coincide. This case occurs when $\text{mFDR}(t_1^*) = \gamma$ and $\mathbb{P}(y \geq t_2^*) = \alpha$, so $\mathbb{E}[v_\alpha(Y)\mathbb{1}\{v_\alpha(Y) \geq \lambda^*\}] = \alpha - \alpha\gamma$ with $v_\alpha(t_1^*) = \lambda^*(\alpha, \gamma)$. Again, the strict thresholding enforced by the statement of the proposition can be relaxed slightly by randomizing the selection probability of the last unit so that the active constraint is satisfied exactly. \blacksquare

Proof. [Proposition 4.1] In the proof we will suppress the dependence of λ^* on the (α, γ) . The argument is very similar to the proof for Proposition 3.3, except that now the feasible region defined by constraint (A.3) and (A.4) is a two-dimensional region for (y_i, σ_i) . Since the posterior tail probability $v_\alpha(y, \sigma)$ is monotone in y for any fixed σ as a result of Lemma 3.2, the optimal rule can be reformulated as a thresholding rule on Y again, $\delta_i^* = \mathbb{1}\{y_i > t_\alpha(\lambda^*, \sigma_i)\}$ except now the threshold value also depends on σ_i .

Now consider the constraint (A.3) and (A.4). Condition (A.3) is equivalent to the condition that,

$$\frac{\mathbb{E}\left[\sum_{i=1}^n\{(1-v_\alpha(y_i, \sigma_i))\delta_i^*\}\right]}{\mathbb{E}\left[\sum_{i=1}^n\delta_i^*\right]} - \gamma = \frac{\int \int_{-\infty}^{\theta_\alpha} \tilde{\Phi}((t_\alpha(\lambda^*, \sigma) - \theta)/\sigma) dG(\theta) dH(\sigma)}{\int \int_{-\infty}^{+\infty} \tilde{\Phi}((t_\alpha(\lambda^*, \sigma) - \theta)/\sigma) dG(\theta) dH(\sigma)} - \gamma \leq 0.$$

For any marginal distributions G and H of (θ, σ) and for a fixed pair of (α, γ) , as λ^* increases, $t_\alpha(\lambda^*, \sigma)$ also increases for any $\sigma > 0$ and therefore it is less likely for condition (A.3) to bind. On the other hand, condition (A.4) is equivalent to,

$$\mathbb{P}(\delta_i^* = 1) - \alpha = \int \int \tilde{\Phi}((t_\alpha(\lambda^*, \sigma) - \theta)/\sigma) dG(\theta) dH(\sigma) - \alpha \leq 0.$$

So as λ^* increases, it is also less likely for condition (A.4) to bind. Thus, when $\lambda_1^* < \lambda_2^*$, the feasible region on (Y, σ) defined by inequality constraint (A.4) is a strict subset of that defined by inequality (A.3). When $\lambda_1^* > \lambda_2^*$, then the feasible region defined by inequality (A.3) is a strict subset of that defined by inequality (A.4). When $\lambda_1^* = \lambda_2^*$, the feasible regions coincide; this case occurs when

$$\mathbb{E}\left[v_\alpha(Y, \sigma)\mathbb{1}\{v_\alpha(Y, \sigma) \geq \lambda^*\}\right] = \alpha - \alpha\gamma$$

where the expectation is taken with respect to the joint distribution of (Y, σ) .

Finally regarding the existence of λ^* , note that existence of a solution for λ_2^* for any $\alpha \in (0, 1)$ follows from the fact that for any fixed α , $f_2(\alpha, \lambda) = \mathbb{P}(v_\alpha(y, \sigma) > \lambda) - \alpha$, is a decreasing function in λ and $\lambda_2^*(\alpha)$ is defined as the zero-crossing point of $f_2(\alpha, \lambda)$. Note that $f_2(\alpha, 0) = 1 - \alpha$ and $f_2(\alpha, 1) = -\alpha$. Therefore for any $\alpha \in (0, 1)$, we can always find a $\lambda_2^*(\alpha) \in (0, 1)$ such that $f_2(\alpha, \lambda_2^*(\alpha)) = 0$. Now consider $f_1(\alpha, \gamma, \lambda) = \mathbb{E}[(1 - v_\alpha(y, \sigma) - \gamma)\mathbb{1}\{v_\alpha(y, \sigma) > \lambda\}]$. For a fixed pair of (α, γ) , $\lambda_1^*(\alpha, \gamma)$ is defined as the zero crossing point of $f_1(\alpha, \gamma, \lambda)$. Note that $f_1(\alpha, \gamma, \lambda)$ decreases first and then increases in λ with its minimum achieved at $\lambda = 1 - \gamma$. We also know that $f_1(\alpha, \gamma, 0) = 1 - \gamma - \mathbb{E}[v_\alpha(y, \sigma)] = 1 - \gamma - \alpha$ and $f_1(\alpha, \gamma, 1) = 0$. Hence as long as $\gamma < 1 - \alpha$, the zero-crossing $\lambda_1^*(\alpha, \gamma)$ exists. The condition $\gamma < 1 - \alpha$ is imposed to rule out cases where FDR constraint never binds. ■

Proof. [Lemma 4.2] Note that for any cutoff value λ , the mFDR can be expressed as,

$$\text{mFDR}(\alpha, \lambda) = \mathbb{E}[(1 - v_\alpha(y_i, \sigma_i))\mathbb{1}\{v_\alpha(y_i, \sigma_i) \geq \lambda\}]/\mathbb{P}[v_\alpha(y_i, \sigma_i) \geq \lambda].$$

Thus, mFDR depends both on the cutoff value λ and on α since v_α , is a function of α , and consequently its density function is also indexed by α .

First, we will show that $\nabla_\lambda \text{mFDR}(\alpha, \lambda) \leq 0$ for all $\alpha \in (0, 1)$. Differentiating with respect to λ gives,

$$\begin{aligned} \nabla_\lambda \frac{\int_\lambda^1 (1-v)f_v(v; \alpha)dv}{\int_\lambda^1 f_v(v; \alpha)dv} &= \frac{-(1-\lambda)f_v(\lambda; \alpha) \int_\lambda^1 f_v(v; \alpha)dv + \int_\lambda^1 (1-v)f_v(v; \alpha)dv f_v(\lambda; \alpha)}{(\int_\lambda^1 f_v(v; \alpha)dv)^2} \\ &= \frac{f_v(\lambda; \alpha)}{(\int_\lambda^1 f_v(v; \alpha)dv)^2} \left(\int_\lambda^1 (1-v)f_v(v; \alpha)dv - \int_\lambda^1 (1-\lambda)f_v(v; \alpha)dv \right) \\ &\leq 0. \end{aligned}$$

Next, to establish that $\nabla_\alpha \text{mFDR}(\alpha, \lambda) \leq 0$, we differentiate with respect to α , to obtain,

$$\begin{aligned} \nabla_\alpha \frac{\int_\lambda^1 (1-v)f_v(v; \alpha)dv}{\int_\lambda^1 f_v(v; \alpha)dv} &= \frac{\int_\lambda^1 (1-v)\nabla_\alpha \log f_v(v; \alpha)f_v(v; \alpha)dv}{\int_\lambda^1 f_v(v; \alpha)dv} - \frac{\int_\lambda^1 (1-v)f_v(v; \alpha)dv \int_\lambda^1 \nabla_\alpha \log f_v(v; \alpha)f_v(v; \alpha)dv}{\int_\lambda^1 f_v(v; \alpha)dv \int_\lambda^1 f_v(v; \alpha)dv} \\ &= \frac{\mathbb{E}[(1-v)\nabla_\alpha \log f_v(v; \alpha)\mathbb{1}\{v \geq \lambda\}]}{\mathbb{P}(v \geq \lambda)} - \frac{\mathbb{E}[(1-v)\mathbb{1}\{v \geq \lambda\}]}{\mathbb{P}(v \geq \lambda)} \frac{\mathbb{E}[\nabla_\alpha \log f_v(v; \alpha)\mathbb{1}\{v \geq \lambda\}]}{\mathbb{P}(v \geq \lambda)} \\ &= \mathbb{E}[(1-v)\nabla_\alpha \log f_v(v; \alpha)|v \geq \lambda] - \mathbb{E}[1-v|v \geq \lambda]\mathbb{E}[\nabla_\alpha \log f_v(v; \alpha)|v \geq \lambda] \\ &= \text{cov}[1-v, \nabla_\alpha \log f_v(v; \alpha)|v \geq \lambda] \leq 0. \end{aligned}$$

where the last inequality holds because $\nabla_\alpha \log f_v(v; \alpha)$ is non-decreasing in v .

Now suppose we have the cutoff value $\lambda_1^*(\alpha_2, \gamma)$ such that,

$$\mathbb{E}[(1 - v_{\alpha_2}(y_i, \sigma_i))\mathbb{1}\{v_{\alpha_2}(y_i, \sigma_i) > \lambda_1^*(\alpha_2, \gamma)\}]/\mathbb{P}(v_{\alpha_2}(y_i, \sigma_i) > \lambda_1^*(\alpha_2, \gamma)) = \gamma.$$

If we maintain the same cutoff value for $v_{\alpha_1}(y_i, \sigma_i)$ with $\alpha_1 > \alpha_2$, given the second property of mFDR, we know,

$$\mathbb{E}[(1 - v_{\alpha_1}(y_i, \sigma_i))\mathbb{1}\{v_{\alpha_1}(y_i, \sigma_i) \geq \lambda_1^*(\alpha_2, \gamma)\}]/\mathbb{P}(v_{\alpha_1}(y_i, \sigma_i) \geq \lambda_1^*(\alpha_2, \gamma)) \leq \gamma.$$

If equality holds, then by definition we have $\lambda_1^*(\alpha_2, \gamma) = \lambda_1^*(\alpha_1, \gamma)$, if strict inequality holds, then by the first property of mFDR, in order to increase mFDR level to be equal to γ , we must have $\lambda_1^*(\alpha_1, \gamma) < \lambda_1^*(\alpha_2, \gamma)$. ■

Proof. [Corollary 4.3] For any $\alpha_1 > \alpha_2$, we have $v_{\alpha_1}(y, \sigma) \geq v_{\alpha_2}(y, \sigma)$ for all pair of $(y, \sigma) \in \mathbb{R} \times \mathbb{R}_+$. When the condition in Lemma 4.2 holds, then $v_{\alpha_1}(y, \sigma) \geq v_{\alpha_2}(y, \sigma) > \lambda_1^*(\alpha_2, \gamma) \geq \lambda_1^*(\alpha_1, \gamma)$, which implies $\Omega_{\alpha_2, \gamma}^{FDR} \subseteq \Omega_{\alpha_1, \gamma}^{FDR}$. ■

Proof. [Lemma 4.4] Since $t_\alpha(\lambda_2^*(\alpha), \sigma)$ defines the boundary of the selection region under the capacity constraint for a fixed level α . The condition imposed that as α increases, for each fixed σ , the thresholding value for Y decreases, hence nestedness of the selection region. ■

Proof. [Lemma 4.5] Based on results in Lemma 4.2 and Lemma 4.4 and the fact that $\Omega_{\alpha, \gamma} = \Omega_{\alpha, \gamma}^{FDR} \cap \Omega_{\alpha}^C$, we have nestedness of the selection set. ■

Proof. [Proposition C.1] The capacity constraint requires that

$$\alpha = \mathbb{P}(M(y, \sigma) \geq C_2^*(\alpha)) = \int \int \mathbb{1}\{M(y, \sigma) \geq C_2^*(\alpha)\} f(y|\theta, \sigma) dG(\theta) dH(\sigma)$$

For any $\alpha_1 > \alpha_2$, it is then clear that $C_2^*(\alpha_1) \leq C_2^*(\alpha_2)$. Given the monotonicity of $M(y, \sigma)$ for each fixed σ established in Lemma 3.2, the selection set based on capacity constraint is nested. For FDR constraint, we require

$$(A.7) \quad \gamma = \frac{\int \int_{-\infty}^{\theta_\alpha} \mathbb{1}\{M(y, \sigma) \geq C_1^*(\alpha, \gamma)\} f(y|\theta, \sigma) dG(\theta) dH(\sigma)}{\int \int \mathbb{1}\{M(y, \sigma) \geq C_1^*(\alpha, \gamma)\} f(y|\theta, \sigma) dG(\theta) dH(\sigma)}$$

Fix γ , it suffices to show that if $\alpha_1 > \alpha_2$, then $C_1^*(\alpha_1, \gamma) \leq C_1^*(\alpha_2, \gamma)$. First, solve for $C_1^*(\alpha_2, \gamma)$ from equation (A.7). Now suppose we use this same thresholding value when we increase capacity to $\alpha_1 > \alpha_2$, we evaluate the right hand side of equation (A.7). Since $\theta_{\alpha_1} \leq \theta_{\alpha_2}$, then the numerator decreases,

$$\begin{aligned} & \int \int_{-\infty}^{\theta_{\alpha_1}} \mathbb{1}\{M(y, \sigma) \geq C_1^*(\alpha_2, \gamma)\} f(y|\theta, \sigma) dG(\theta) dH(\sigma) \\ & \leq \int \int_{-\infty}^{\theta_{\alpha_2}} \mathbb{1}\{M(y, \sigma) \geq C_1^*(\alpha_2, \gamma)\} f(y|\theta, \sigma) dG(\theta) dH(\sigma), \end{aligned}$$

while the denominator does not change. The only way to satisfy the equality (A.7) again is to decrease the thresholding value, therefore $C_1^*(\alpha_1, \gamma) \leq C_1^*(\alpha_2, \gamma)$. The result in the Proposition is then reached since the selection set is the intersection of the selection set under capacity constraint and that under FDR constraint. ■

Proof. [Lemma 5.1] The logarithm of the Gamma density of S_i takes the form,

$$\log \Gamma(S_i | r_i, \sigma_i^2) = r_i \log(r_i / \sigma_i^2) - \log(\Gamma(r_i)) + (r_i - 1) \log S_i - S_i \frac{r_i}{\sigma_i^2},$$

hence $\nabla_s \Gamma(s|r, \sigma^2) = \Gamma(s|r, \sigma^2) \left(\frac{r-1}{s} - \frac{r}{\sigma^2} \right)$. Fixing y and differentiating with respect to s , we have,

$$\begin{aligned} \nabla_s v_\alpha(y, s) &= \frac{\int \int_{\theta_\alpha}^{+\infty} f(y|\theta, \sigma^2) \Gamma(s|r, \sigma^2) \left[\frac{r-1}{s} - \frac{r}{\sigma^2} \right] dG(\theta, \sigma^2)}{\int \int f(y|\theta, \sigma^2) \Gamma(s|r, \sigma^2) dG(\theta, \sigma^2)} \\ &\quad - \frac{\int \int_{\theta_\alpha}^{+\infty} f(y|\theta, \sigma^2) \Gamma(s|r, \sigma^2) dG(\theta, \sigma^2)}{\int \int f(y|\theta, \sigma^2) \Gamma(s|r, \sigma^2) dG(\theta, \sigma^2)} \frac{\int \int f(y|\theta, \sigma^2) \Gamma(s|r, \sigma^2) \left[\frac{r-1}{s} - \frac{r}{\sigma^2} \right] dG(\theta, \sigma^2)}{\int \int f(y|\theta, \sigma^2) \Gamma(s|r, \sigma^2) dG(\theta, \sigma^2)} \\ &= \mathbb{E} \left[\mathbb{1}\{\theta \geq \theta_\alpha\} \left(\frac{r-1}{s} - \frac{r}{\sigma^2} \right) | Y = y, S = s \right] - \mathbb{E} \left[\mathbb{1}\{\theta \geq \theta_\alpha\} | Y = y, S = s \right] \mathbb{E} \left[\frac{r-1}{s} - \frac{r}{\sigma^2} | Y = y, S = s \right] \\ &= -\text{Cov} \left[\mathbb{1}\{\theta \geq \theta_\alpha\}, \frac{r}{\sigma^2} | Y = y, S = s \right] \end{aligned}$$

The covariance term can take either sign since we do not restrict the distribution G , so $v_\alpha(Y, S)$ need not be monotone in S . On the other hand, if we fix s and differentiate with respect to y ,

$$\begin{aligned}\nabla_y v_\alpha(y, s) &= \mathbb{E}\left[\mathbb{1}\{\theta \geq \theta_\alpha\} \left[-\frac{y-\theta}{\sigma^2/T}\right] \middle| Y=y, S=s\right] - \mathbb{E}\left[\mathbb{1}\{\theta \geq \theta_\alpha\} \middle| Y=y, S=s\right] \mathbb{E}\left[-\frac{y-\theta}{\sigma^2/T} \middle| Y=y, S=s\right] \\ &= \text{Cov}\left[\mathbb{1}\{\theta \geq \theta_\alpha\}, \frac{\theta-y}{\sigma^2/T} \middle| Y=y, S=s\right].\end{aligned}$$

Again, the covariance term can take either sign, depending on the correlation of θ and σ^2 conditional on (Y, S) . Therefore, fixing S , $v_\alpha(Y, S)$ need not be a monotone function of Y . \blacksquare

Proof. [Proposition 5.2] The proof is very similar to that of Proposition 4.1, the only difference is that we can no longer formulate the decision rule by simply thresholding on Y because the transformation $v_\alpha(Y, S)$ need not be monotone in Y for fixed values of S as shown in Lemma 5.1, hence $\lambda_1^*(\alpha, \gamma)$ and $\lambda_2^*(\alpha)$ must now be defined directly through the random variable v_α . The first constraint states that

$$\mathbb{E}\left[\sum_{i=1}^n \{(1 - v_\alpha(y_i, s_i))\delta_i^*\} / \mathbb{E}\left[\sum_{i=1}^n \delta_i^*\right]\right] \leq \gamma$$

with $\delta_i^* = \mathbb{1}\{v_\alpha(y_i, s_i) \geq \lambda^*\}$. For each fixed α , let the density function for $v_\alpha(Y, S)$ be denoted as $f_v(\cdot; \alpha)$, then the constraints can be formulated as

$$\int_{\lambda^*}^1 (1-v)f_v(v; \alpha)dv / \int_{\lambda^*}^1 f_v(v; \alpha)dv$$

which is non-increasing in λ^* , hence the constraint becomes less likely to bind as λ^* increases. On the other hand the second constraint states that,

$$\mathbb{P}(\delta_i^* = 1) - \alpha = \int_{\lambda^*}^1 f_v(v; \alpha)dv - \alpha$$

For each fixed $\alpha \in (0, 1)$, this constraint also becomes less likely to bind as λ^* increases. \blacksquare

Proof. [Theorem 6.1] To prove Theorem 6.1, we first introduce some additional notation and prove several lemmas. Let

$$\begin{aligned}H_{n,0}(t) &= 1 - H_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{v_{\alpha,i} \geq t\} \\ H_{n,1}(t) &= \frac{1}{n} \sum_{i=1}^n (1 - v_{\alpha,i}) \mathbb{1}\{v_{\alpha,i} \geq t\} \\ Q_n(t) &= H_{n,1}(t) / H_{n,0}(t) \\ V_n(t) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{v_{\alpha,i} \geq t\} \mathbb{1}\{\theta_i \leq \theta_\alpha\} \\ H_0(t) &= 1 - H(t) = \mathbb{P}(v_{\alpha,i} \geq t) \\ H_1(t) &= \mathbb{E}\left[(1 - v_{\alpha,i}) \mathbb{1}\{v_{\alpha,i} \geq t\}\right] \\ Q(t) &= H_1(t) / H_0(t)\end{aligned}$$

Lemma A.1. *Under Assumption 1, as $n \rightarrow \infty$,*

$$\begin{aligned}\sup_{t \in [0,1]} |H_{n,0}(t) - H_0(t)| &\xrightarrow{P} 0 \\ \sup_{t \in [0,1]} |H_{n,1}(t) - H_1(t)| &\xrightarrow{P} 0\end{aligned}$$

Proof. [Proof of Lemma A.1] Under Assumption 1 and the fact that $v_{\alpha,i} \in [0, 1]$, the weak law of large numbers implies that we have for any $t \in [0, 1]$, as $n \rightarrow \infty$,

$$\begin{aligned} H_{n,0}(t) &\xrightarrow{P} H_0(t) \\ H_{n,1}(t) &\xrightarrow{P} H_1(t) \end{aligned}$$

By the Glivenko-Cantalli theorem, the first result is immediate. To prove the second result, it suffices to show that for any $\epsilon > 0$, as $n \rightarrow \infty$,

$$\mathbb{P}\left(\sup_{t \in [0,1]} |H_{n,1}(t) - H_1(t)| > \epsilon\right) \rightarrow 0.$$

It is clear that since $v_{\alpha,i}$ has a continuous distribution, $H_1(t)$ is a monotonically decreasing and bounded function in t with $H_1(0) = 1 - \alpha$ and $H_1(1) = 0$. It is also clear that the function $H_{n,1}(t)$ is monotonically decreasing in t , so we can find $m_\epsilon < \infty$ points such that $0 = t_0 < t_1 < \dots < t_{m_\epsilon} = 1$, and for any $j \in \{1, 2, \dots, m_\epsilon\}$, we have $H_1(x_j) - H_1(x_{j-1}) \leq \epsilon/2$. For any $t \in [0, 1]$, there exists a j such that $x_{j-1} \leq t \leq x_j$ and

$$\begin{aligned} H_{n,1}(t) - H_1(t) &\leq H_{n,1}(t_{j-1}) - H_1(t_j) \\ &= (H_{n,1}(t_{j-1}) - H_1(t_{j-1})) + (H_1(t_{j-1}) - H_1(t_j)) \\ &\leq |H_{n,1}(x_{j-1}) - H_1(x_{j-1})| + \epsilon/2 \leq \max_j |H_{n,1}(t_{j-1}) - H_1(t_{j-1})| + \epsilon/2 \end{aligned}$$

Likewise we can show that $H_{n,1}(t) - H_1(t) \geq -\max_j |H_{n,1}(t_j) - H_1(t_j)| - \epsilon/2$, hence

$$\sup_{t \in [0,1]} |H_{n,1}(t) - H_1(t)| \leq \max_j |H_{n,1}(t_j) - H_1(t_j)| + \epsilon/2.$$

Since m_ϵ is finite then for any $\delta > 0$, there exists N such that for all $n \geq N$,

$$\mathbb{P}\left(\max_j |H_{n,1}(t_j) - H_1(t_j)| \geq \epsilon/2\right) \leq \delta$$

which then implies that

$$\begin{aligned} \mathbb{P}\left(\sup_{t \in [0,1]} |H_{n,1}(t) - H_1(t)| \geq \epsilon\right) &\leq \mathbb{P}\left(\frac{\epsilon}{2} + \max_j |H_{n,1}(t_j) - H_1(t_j)| \geq \epsilon\right) \\ &= \mathbb{P}\left(\max_j |H_{n,1}(t_j) - H_1(t_j)| \geq \epsilon/2\right) \rightarrow 0 \end{aligned}$$

■

Lemma A.2. *Under Assumption 1 and $\alpha < 1 - \gamma$, $Q(1 - \gamma) < \gamma$.*

Proof. [Proof of Lemma A.2] Define $\bar{Q}(t) = \mathbb{E}\left[(1 - v_{\alpha,i} - \gamma)1\{v_{\alpha,i} \geq t\}\right]$, then $Q(t) = \gamma$ implies $\bar{Q}(t) = 0$. Since $Q(t)$ is monotonically decreasing in t as shown in the proof of Lemma 4.2, it suffices to prove that $\bar{Q}(1 - \gamma) < 0$. To this end, note that $\nabla_t \bar{Q}(t) < 0$ for $t < 1 - \gamma$ and $\nabla_t \bar{Q}(t) > 0$ for $t > 1 - \gamma$, hence $\bar{Q}(t)$ obtains its minimum value at $t = 1 - \gamma$. Note that $\bar{Q}(0) = 1 - \gamma - \alpha$ and $\bar{Q}(1) = 0$, thus $\bar{Q}(1 - \gamma) < 0$. ■

Now we are ready to prove Theorem 6.1. We prove the first statement, as the second statement can be shown with a similar argument. First we show that, $\sup_{t \leq 1 - \gamma} |Q_n(t) - Q(t)| \xrightarrow{P} 0$, since,

$$\begin{aligned} |Q_n(t) - Q(t)| &= \left| \frac{H_0(t)H_{n,1}(t) - H_1(t)H_{n,0}(t)}{H_{n,0}(t)H_0(t)} \right| = \left| \frac{H_0(t)(H_{n,1}(t) - H_1(t)) - H_1(t)(H_{n,0}(t) - H_0(t))}{H_{n,0}(t)H_0(t)} \right| \\ &\leq \frac{H_0(0)\sup_t |H_{n,1}(t) - H_1(t)| + H_1(0)\sup_t |H_{n,0}(t) - H_0(t)|}{H_0(1 - \gamma)(H_0(1 - \gamma) - \sup_t |H_{n,0}(t) - H_0(t)|)} \xrightarrow{P} 0 \end{aligned}$$

uniformly for any $t \leq 1 - \gamma$. The last inequality holds because $\min_{t \leq 1 - \gamma} H_0(t) = H_0(1 - \gamma)$ by monotonicity of $H_0(t)$. With a similar argument, we can also show that $\sup_{t \leq 1 - \gamma} \left| \frac{V_n(t)}{H_{n,0}(t)} - Q(t) \right| \xrightarrow{p} 0$. Using this result and the fact that $Q(1 - \gamma) < \gamma$ by Lemma A.2, we have $\mathbb{P}\left(|Q_n(1 - \gamma) - Q(1 - \gamma)| < \frac{\gamma - Q(1 - \gamma)}{2}\right) \rightarrow 1$ and therefore, $\mathbb{P}(Q_n(1 - \gamma) < \gamma) \rightarrow 1$ and $\mathbb{P}(\lambda_{2n} \leq 1 - \gamma) \rightarrow 1$ by the definition of λ_{2n} . Since $\lambda_n \leq \lambda_{2n}$ by definition, we also have $\mathbb{P}(\lambda_n \leq 1 - \gamma) \rightarrow 1$. On the other hand,

$$Q_n(\lambda_n) - \frac{V_n(\lambda_n)}{H_{n,0}(\lambda_n)} \geq \inf_{t \leq 1 - \gamma} \left(Q_n(t) - Q(t) + Q(t) - V_n(t)/H_{n,0}(t) \right) = o_p(1)$$

Since $Q_n(\lambda_n) \leq Q_n(\lambda_{2n}) \leq \gamma$, it follows that

$$\frac{V_n(\lambda_n)}{H_{n,0}(\lambda_n) \sqrt{1}} \leq \frac{V_n(\lambda_n)}{H_{n,0}(\lambda_n)} \leq \gamma + o_p(1).$$

Since $\frac{V_n(\lambda_n)}{H_{n,0}(\lambda_n) \sqrt{1}}$ is upper bounded by 1, by Fatou's lemma, we have

$$\limsup_{n \rightarrow \infty} \mathbb{E} \left[\frac{V_n(\lambda_n)}{H_{n,0}(\lambda_n) \sqrt{1}} \right] \leq \gamma$$

■

Lemma A.3. *Under Assumption 1 and 2, as $n \rightarrow \infty$,*

$$\hat{\theta}_\alpha \rightarrow \theta_\alpha \quad a.s.$$

Proof. [Proof of Lemma A.3] See Lemma 21.2 in van der Vaart (2000). ■

Lemma A.4. *Under Assumption 1 and 2, as $n \rightarrow \infty$,*

$$\sup_i |\hat{v}_{\alpha,i} - v_{\alpha,i}| \rightarrow 0 \quad a.s.$$

Proof. [Proof of Lemma A.3] Since

$$\hat{v}_{\alpha,i} = \frac{\int_{\hat{\theta}_\alpha}^{+\infty} f(D_i|\theta) d\hat{G}_n(\theta)}{\int_{-\infty}^{+\infty} f(D_i|\theta) d\hat{G}_n(\theta)}$$

where we denote D_i as data with a density function $f(D_i|\theta)$. When variances are known, then $D_i = \{y_i, \sigma_i\}$ and $f(D_i|\theta) = \frac{1}{\sigma_i} \varphi((y_i - \theta)/\sigma_i)$ and when variances are unknown, then $D_i = \{y_i, s_i\}$ and $f(D_i|\theta) = \frac{1}{\sqrt{\sigma^2/T}} \varphi((y_i - \theta)/\sqrt{\sigma^2/T}) \Gamma(s_i|r, \sigma^2/r)$ with $r = (T - 1)/2$ with $\varphi(\cdot)$ and $\Gamma(\cdot|\cdot, \cdot)$ being the standard normal and gamma density function, respectively.

We first analyze the denominator and prove

$$(A.8) \quad \sup_x \left| \int_{-\infty}^{+\infty} f(x|\theta) d\hat{G}_n(\theta) - \int_{-\infty}^{+\infty} f(x|\theta) dG(\theta) \right| \rightarrow 0 \quad a.s.$$

Let $f_n(x) = \int f(x|\theta) d\hat{G}_n(\theta)$ and $f(x) = \int f(x|\theta) dG(\theta)$. Under Assumption 2, we have $\frac{1}{2} \int \left(\sqrt{f_n(x)} - \sqrt{f(x)} \right)^2 d\mu(x) \rightarrow 0$ almost surely, which implies that $\int |f_n(x) - f(x)| dx \rightarrow 0$ almost surely. If $f_n(x)$ and $f(x)$ are Lipschitz continuous, we proceed by contradiction. Suppose (A.8) doesn't hold, then there exists $\epsilon > 0$ and a sequence $\{x_n\}_{n \geq 1}$ such that $|f_n(x_n) - f(x_n)| \geq \epsilon$ for all n . By Lipschitz continuity of f_n and f , there exists C such that

$$\begin{aligned} |f_n(x_n + \delta) - f_n(x_n)| &\leq C \|\delta\| \\ |f(x_n + \delta) - f(x_n)| &\leq C \|\delta\| \end{aligned}$$

And therefore there exists $\eta > 0$ and for all $\|y - x_n\| \leq \eta$, $|f_n(y) - f(y)| \geq \epsilon/2$, which then implies

$$\int |f_n(x) - f(x)| dx \geq \int \mathbf{1}\{\|y - x_n\| \leq \eta\} |f_n(y) - f(y)| dy \geq \frac{\epsilon}{2} \int \mathbf{1}\{\|y - x_n\| \leq \eta\} dy$$

which contradicts $\int |f_n(x) - f(x)|dx \rightarrow 0$ almost surely. To prove that the functions f_n and f are Lipschitz continuous. Note that it suffices to prove that for each fixed parameter θ , $|f(x|\theta) - f(y|\theta)| \leq C_\theta \|x - y\|$ and $\sup_\theta C_\theta < \infty$. This clearly holds for the Gaussian density since the Gaussian density is everywhere differentiable and has bounded first derivative. Under Assumption 1 with $T \geq 4$, the Gamma density is also everywhere differentiable and has bounded first derivative, and thus is Lipschitz continuous.

We next analyze the numerator and show

$$(A.9) \quad \sup_x \left| \int_{\hat{\theta}_\alpha}^{+\infty} f(x|\theta) d\hat{G}_n(\theta) - \int_{\theta_\alpha}^{+\infty} f(x|\theta) dG(\theta) \right| \rightarrow 0 \quad a.s.$$

Note that

$$\begin{aligned} & \left| \int_{\hat{\theta}_\alpha}^{+\infty} f(x|\theta) d\hat{G}_n(\theta) - \int_{\theta_\alpha}^{+\infty} f(x|\theta) dG(\theta) \right| \\ & \leq \left| \int_{\hat{\theta}_\alpha}^{+\infty} f(x|\theta) d\hat{G}_n(\theta) - \int_{\theta_\alpha}^{+\infty} f(x|\theta) d\hat{G}_n(\theta) \right| + \left| \int_{\theta_\alpha}^{+\infty} f(x|\theta) d\hat{G}_n(\theta) - \int_{\theta_\alpha}^{+\infty} f(x|\theta) dG(\theta) \right|. \end{aligned}$$

The first term converges to 0 uniformly due to Lemma A.3. To show the second term also converges to zero uniformly, we make use of the result that if \hat{G}_n weakly converges to G , which holds under Assumption 2, then $\sup_{g \in \mathcal{BL}} |\int g d\hat{G}_n - \int g dG| \rightarrow 0$ where \mathcal{BL} is the class of bounded Lipschitz continuous functions. Note that $f(x|\theta)1\{\theta \geq \theta_\alpha\}$ is bounded and continuous except at $\theta = \theta_\alpha$. So we construct a smoothed version of $f(x|\theta)1\{\theta \geq \theta_\alpha\}$, denoted as $g(x|\theta)$, by replacing $1\{\theta \geq \theta_\alpha\}$ by a piecewise linear function taking value zero for $\theta < \theta_\alpha$ and value 1 for $\theta \geq \theta_\alpha + \epsilon$ and taking the form $-\theta_\alpha/\epsilon + \theta/\epsilon$ for $\theta \in [\theta_\alpha, \theta_\alpha + \epsilon]$, then $g \in \mathcal{BL}$. The result (A.9) then holds by showing that

$$\sup_x \left| \int_{\theta_\alpha}^{\theta_\alpha + \epsilon} f(x|\theta) d\hat{G}_n(\theta) + \int_{\theta_\alpha}^{\theta_\alpha + \epsilon} f(x|\theta) dG(\theta) \right| \rightarrow 0 \quad a.s.$$

which holds by Assumptions 1 and 2. ■

Proof. [Proof of Theorem 6.2] Define analogously

$$\begin{aligned} \hat{H}_{n,0}(t) &= \frac{1}{n} \sum_{i=1}^n 1\{\hat{v}_{\alpha,i} \geq t\} \\ \hat{H}_{n,1}(t) &= \frac{1}{n} \sum_{i=1}^n (1 - \hat{v}_{\alpha,i}) 1\{\hat{v}_{\alpha,i} \geq t\} \\ \hat{Q}_n(t) &= \hat{H}_{n,1}(t) / \hat{H}_{n,0}(t) \end{aligned}$$

We first show

$$\begin{aligned} \sup_{t \in [0,1]} |\hat{H}_{n,0}(t) - H_0(t)| &\xrightarrow{P} 0 \\ \sup_{t \in [0,1]} |\hat{H}_{n,1}(t) - H_1(t)| &\xrightarrow{P} 0 \end{aligned}$$

We will prove the second statement now, and the first can be proved using a similar argument. To prove the second statement, it suffices to show that

$$\sup_{t \in [0,1]} \left| \hat{H}_{n,1}(t) - H_{n,1}(t) \right| \xrightarrow{P} 0$$

To this end, note

$$\begin{aligned}
& \sup_{t \in [0,1]} \left| \frac{1}{n} \sum_i (1 - \hat{v}_{\alpha,i}) 1\{\hat{v}_{\alpha,i} \geq t\} - \frac{1}{n} \sum_i (1 - v_{\alpha,i}) 1\{v_{\alpha,i} \geq t\} \right| \\
&= \sup_{t \in [0,1]} \left| \frac{1}{n} \sum_i (1 - \hat{v}_{\alpha,i}) 1\{\hat{v}_{\alpha,i} \geq t\} - \frac{1}{n} \sum_i (1 - v_{\alpha,i}) 1\{\hat{v}_{\alpha,i} \geq t\} \right| \\
&+ \sup_{t \in [0,1]} \left| \frac{1}{n} \sum_i (1 - v_{\alpha,i}) 1\{\hat{v}_{\alpha,i} \geq t\} - \frac{1}{n} \sum_i (1 - v_{\alpha,i}) 1\{v_{\alpha,i} \geq t\} \right| \\
&\leq \frac{1}{n} \sum_i |\hat{v}_{\alpha,i} - v_{\alpha,i}| + \sup_{t \in [0,1]} \frac{1}{n} \sum_i |1\{\hat{v}_{\alpha,i} \geq t\} - 1\{v_{\alpha,i} \geq t\}|
\end{aligned}$$

The first term is implied by the result in Lemma A.4. The second term can be written as

$$\begin{aligned}
& \sup_{t \in [0,1]} \frac{1}{n} \sum_i \left| 1\{\hat{v}_{\alpha,i} \geq t\} - 1\{v_{\alpha,i} \geq t\} \right| \\
&= \sup_{t \in [0,1]} \frac{1}{n} \sum_i \left[1\{\hat{v}_{\alpha,i} \geq t, v_{\alpha,i} < t\} + 1\{\hat{v}_{\alpha,i} < t, v_{\alpha,i} \geq t\} \right] \\
&= \sup_{t \in [0,1]} \frac{1}{n} \sum_i \left[1\{\hat{v}_{\alpha,i} \geq t, t - e < v_{\alpha,i} < t\} + 1\{\hat{v}_{\alpha,i} < t, t \leq v_{\alpha,i} < t + e\} \right] \\
&+ \frac{1}{n} \sum_i \left[1\{\hat{v}_{\alpha,i} \geq t, v_{\alpha,i} < t - e\} + 1\{\hat{v}_{\alpha,i} < t, v_{\alpha,i} \geq t + e\} \right] \\
&\leq \sup_{t \in [0,1]} \frac{1}{n} \sum_i 1\{t - e \leq v_{\alpha,i} \leq t + e\} + \frac{1}{ne} \sum_i |\hat{v}_{\alpha,i} - v_{\alpha,i}| \\
&\leq \sup_{t \in [0,1]} |H_0(t + e) - H_0(t - e)| + 2 \sup_{t \in [0,1]} |H_{n,0}(t) - H_0(t)| + \frac{1}{ne} \sum_i |\hat{v}_{\alpha,i} - v_{\alpha,i}|
\end{aligned}$$

for some $e > 0$ arbitrarily small and bounded away from zero, the right hand side converges to zero in probability by results in Lemma A.1 and the uniform continuity of H_0 . Using similar arguments in the proof of Theorem 6.1, we can establish that $\sup_{t \leq 1-\gamma} |\hat{Q}_n(t) - Q(t)| \xrightarrow{P} 0$ and $\sup_{t \leq 1-\gamma} \left| \frac{\hat{V}_n(t)}{\hat{H}_{n,0}(t)} - Q(t) \right| \xrightarrow{P} 0$ with $\hat{Q}_n(t) = \hat{H}_{n,1}(t)/\hat{H}_{n,0}(t)$ and $\hat{V}_n(t) = \frac{1}{n} \sum_{i=1}^n 1\{\hat{v}_{\alpha,i} \geq t\} 1\{\theta_i \leq \theta_\alpha\}$ and consequently $\limsup_{n \rightarrow \infty} \mathbb{E} \left[\frac{\hat{V}_n(\hat{\lambda}_n)}{\hat{H}_{n,0}(\hat{\lambda}_n) \sqrt{1}} \right] \leq \gamma$. \blacksquare

Proof. [Proof of Theorem 6.3]

We first show that $\hat{\lambda}_{1n} \xrightarrow{P} \lambda_1^*$ and $\hat{\lambda}_{2n} \xrightarrow{P} \lambda_2^*$, then by the continuous mapping theorem, we have $\hat{\lambda}_n = \max\{\hat{\lambda}_{1n}, \hat{\lambda}_{2n}\} \xrightarrow{P} \max\{\lambda_1^*, \lambda_2^*\} = \lambda^*$. The second statement follows from Lemma A.3. The first statement holds because by the argument for Theorem 6.2, we have

$$(A.10) \quad \sup_{t \geq 1-\gamma} \left| \hat{Q}_n(t) - Q(t) \right| \xrightarrow{P} 0$$

And therefore for any $\epsilon > 0$ not too large, we have $\inf_{t \leq \lambda^* - \epsilon} Q(t) > \gamma$ and $Q(\lambda^* + \epsilon) > \gamma$ by monotonicity of $Q(t)$. Combined with (A.10), we have $\hat{\lambda}_{1n} \xrightarrow{P} \lambda_1^*$.

Now define

$$\begin{aligned}
H_{n,2} &= \frac{1}{n} \sum_{i=1}^n v_{\alpha,i} \mathbb{1}\{v_{\alpha,i} \geq t\} \\
\hat{H}_{n,2}(t) &= \frac{1}{n} \sum_{i=1}^n \hat{v}_{\alpha,i} \mathbb{1}\{\hat{v}_{\alpha,i} \geq t\} \\
\hat{U}_n(t) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\theta_i \geq \theta_\alpha, \hat{v}_{\alpha,i} \geq t\} \\
U_n(t) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\theta_i \geq \theta_\alpha, v_{\alpha,i} \geq t\} \\
H_2(t) &= \mathbb{P}(\theta_i \geq \theta_\alpha, v_{\alpha,i} \geq t) = \alpha\beta(t)
\end{aligned}$$

It suffices to prove that $\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\theta_i \geq \theta_\alpha, \hat{v}_{\alpha,i} \geq \hat{\lambda}_n\} \xrightarrow{P} H_2(\lambda^*)$. Using a similar argument as for Theorem 6.2, we can show

$$\begin{aligned}
\sup_{t \in [0,1]} \left| \hat{H}_{n,2}(t) - H_2(t) \right| &\xrightarrow{P} 0 \\
\sup_{t \in [0,1]} \left| \hat{U}_n(t) - H_2(t) \right| &\xrightarrow{P} 0
\end{aligned}$$

Combining this with the result that $\hat{\lambda}_n \xrightarrow{P} \lambda^*$, we have $\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\theta_i \geq \theta_\alpha, \hat{v}_{\alpha,i} \geq \hat{\lambda}_n\} \xrightarrow{P} H_2(\lambda^*)$ by continuity of H_2 . ■

APPENDIX B. A DISCRETE BIVARIATE EXAMPLE

In this appendix we consider a case where G is a discrete distribution joint distribution in the pairs, (θ, σ) , in particular, $G(\theta, \sigma) = 0.85\delta_{(-1,6)} + 0.1\delta_{(4,2)} + 0.05\delta_{(5,4)}$. In contrast to the discrete example in Section 4.4, the unobserved variance σ^2 is now clearly informative about θ for this distribution G .

We will focus on the capacity constraint $\alpha = 0.05$ so $\theta_\alpha = 5$. For $T = 9$, the level curves for tail probability and posterior mean are shown in Figure B.1 and the selection set comparison for one sample realization in Figure B.2. The right panel of the figure plots the selection boundaries for the two ranking criteria for $\gamma = 10\%$. The non-monotonicity of $v_\alpha(y, s)$ in both y and s is apparent. The posterior mean criteria based on $\mathbb{E}(\theta|Y, S)$, prefers individuals with smaller variances compared to the rule based on the tail probability. Since sample variances S are informative about θ , when the sample variance is small and selection is based on posterior tail probability, the oracle is aware that such a small sample variance is only likely when $\theta = 4$, hence will only make a selection when we observe a very large y . As a result, the Oracle sets a higher selection threshold on y to avoid selecting individuals with true effect $\theta = 4$. On the other hand, the posterior mean criterion also tries to use information from the sample variance, but not as effectively for our selection objective. This can be seen in the level curves in the middle panel of Figure B.1. When the sample variance is small, the posterior mean shrinks very aggressively towards 4, thereby sacrificing valuable information from y . For a wide range of values for the sample mean y , the posterior mean delivers a value close to 4, thus failing to distinguish between those with $\theta = 5$ and those with $\theta = 4$. Consequently, the posterior mean rule sets a lower thresholding value on y for the selection region when sample variance is small resulting in inferior power performance, as shown in Table B.1.

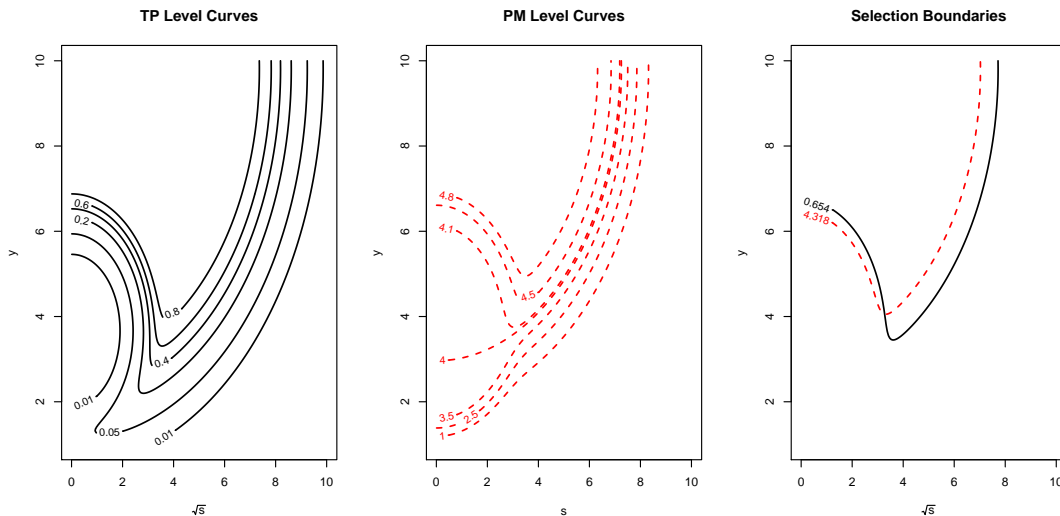


FIGURE B.1. The left panel plots the level curves for the posterior tail probability criterion and the middle panel depicts the level curves for posterior mean criterion. The right panel plots the selection boundary based on posterior mean ranking (shown as the red dashed lines) and the posterior tail probability ranking (shown as the black solid lines) for $\alpha = 5\%$ and $\gamma = 10\%$ with $G(\theta, \sigma^2)$ follows a three points discrete distribution.

Table B.1 reports several performance measures over 200 simulation repetitions with $n = 50,000$. There, we consider four additional methods for ranking:

- MLE: ranking of the maximum likelihood estimators, Y_i for each of the θ_i ,

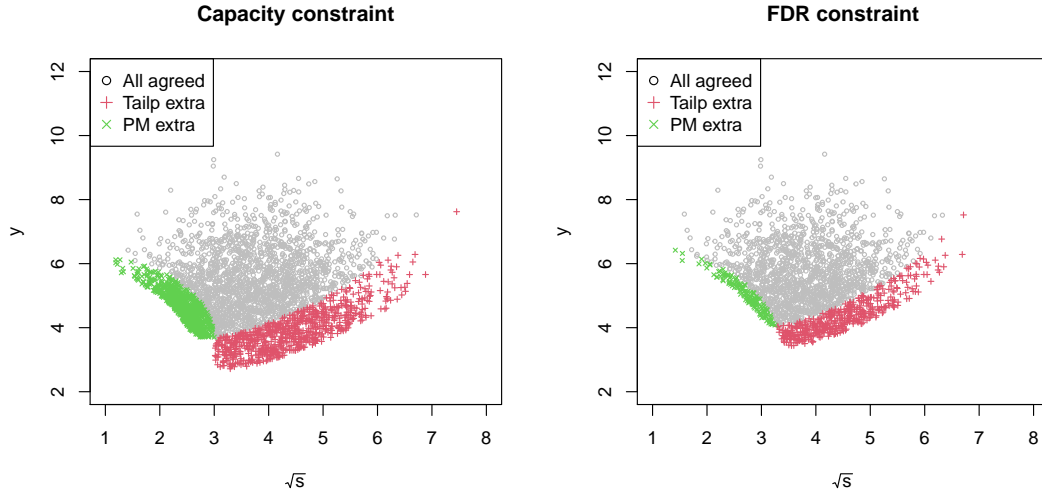


FIGURE B.2. Selection set comparison for one sample realization from the three points discrete distribution model: The left panel shows in black circles the agreed selected elements by both the posterior mean and posterior tail probability criteria under the capacity constraint, extra elements selected 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 FDR constraint with $\alpha = 5\%$ and $\gamma = 10\%$.

	$\gamma = 1\%$			$\gamma = 5\%$			$\gamma = 10\%$		
	Power	FDR	SelProp	Power	FDR	SelProp	Power	FDR	SelProp
PM	0.217	0.010	0.011	0.482	0.050	0.025	0.580	0.100	0.032
TP	0.252	0.010	0.013	0.561	0.050	0.030	0.697	0.100	0.039
P-value	0.651	0.349	0.050	0.651	0.350	0.050	0.651	0.349	0.050
MLE	0.611	0.390	0.050	0.611	0.390	0.050	0.610	0.390	0.050
PM-NIX	0.611	0.390	0.050	0.611	0.390	0.050	0.610	0.390	0.050
TP-NIX	0.619	0.382	0.050	0.619	0.382	0.050	0.618	0.382	0.050

TABLE B.1. Performance comparison for ranking procedures based on posterior mean, posterior tail probability, the P-value and the MLE of θ_i . All results are based on 200 simulation repetitions with $n = 50,000$ for G following the three point discrete distribution and $T = 9$ or when G is assumed to follow a normal-inverse-chi-square distribution. For the first two rows, number reported in the table correspond to performance when both capacity and FDR constraints are in place. For the last four rows, only capacity constraint is in place.

- P-values: ranking of the P-values of the conventional one-sided test of the null hypothesis $H_0 : \theta < \theta_\alpha$,
- PM-NIX: ranking of the posterior means based on the normal-inverse-chi-square (NIX) prior distribution,
- TP-NIX: ranking of the posterior tail probability based on NIX prior distribution

The first two of these selection rules ignore the compound decision perspective of the problem entirely. The other two ranking criteria we consider are based on posterior mean and tail probability assuming G follows a normal-inverse-chi-square (NIX) distribution (denoted as PM-NIX and TP-NIX in Table B.1). The parameters of the NIX distribution are estimated from the data, hence these rules can be viewed as generalization of James-Stein estimator for homogeneous variances and the Efron-Morris shrinkage estimator for known heterogeneous variances case. We refer the details of the NIX distribution and the posterior distribution of (θ, σ^2) to Example 5.1.

We report the power and false discovery rates for the posterior mean (denoted as PM) and posterior tail probability (denoted as TP) selection as well as the proportion of selected observations for $\alpha = 5\%$ and for several different γ under both capacity and FDR control. For all other four selection rules we only impose the capacity constraint, as how they are usually implemented in current practise. For the PM and TP rules, from the proportion selected observations we can infer whether the FDR constraint or the capacity constraint is binding in each configuration. Ranking based on the posterior tail probability clearly has better power performance for each of configurations when compared to the posterior mean ranking. When selecting as few as 5%, FDR constraints are binding for both PM and TP rule for all ranges of $\gamma \in \{1\%, 5\%, 10\%\}$. Among all the other rules, we see that the false discovery rate is around 40% and PM-NIX has identical performance as ranking based on the MLE for θ ; this can be understood by noting that the posterior mean of θ under the NIX prior is simply linear shrinkage of the MLE of θ , hence it does not alter individual rankings between the two methods. TP-NIX behaves similarly to PM-NIX, with slightly better power and slightly lower false discovery rate.

APPENDIX C. A COUNTEREXAMPLE: NON-NESTED SELECTION REGIONS

Thus far we have stressed conditions under which selection regions are nested with respect to α , that is, for $\alpha_1 < \alpha_2 < \dots < \alpha_m$, we have $\Omega_{\alpha_1} \subseteq \Omega_{\alpha_2} \subseteq \dots \subseteq \Omega_{\alpha_m}$, for the selection regions. However, this need not hold when there is variance heterogeneity and when nesting fails we can have seemingly anomalous situations in which units are selected by the tail probability rule at some stringent, low α , but are then rejected for some less stringent, larger α 's. To illustrate this phenomenon we will neglect the FDR constraint and focus on our discrete mixing distribution, $G = 0.85\delta_{-1} + 0.10\delta_2 + 0.05\delta_5$, with $\sigma \sim U[1/2, 4]$. The selection regions are depicted in Figure C.1 for $\alpha \in \{0.04, 0.05, 0.06, 0.08\}$. Units are selected when their observed pair, (y_i, σ_i) lies above these curves for various α 's. When σ is small we see, as expected, that selection is nested: if a unit is selected at low α it stays selected at larger α 's. However, when $\sigma = 3$, we see that there are units selected at $\alpha = 0.05$ and even $\alpha = 0.04$ and yet they are rejected for $\alpha = 0.06$. How can this be?

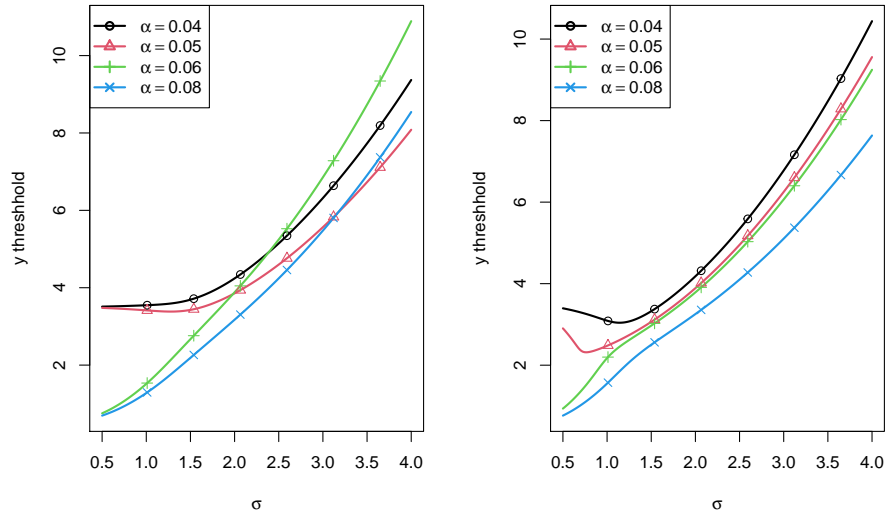


FIGURE C.1. Oracle selection boundaries (with just capacity constraint) for several α levels for the tail probability criterion (left panel) and posterior mean criterion (right panel) with a discrete example with $G = 0.85\delta_{-1} + 0.10\delta_2 + 0.05\delta_5$, and $\sigma \sim U[1/2, 4]$. Crossing of the boundaries implies that the selection regions are non-nested as explained in the text.

Imagine you are the oracle, so you know G , and when you decide to select with $\alpha = 0.06$ you know that you will have to select a few $\theta = 2$ types, since there are only 5 percent of the $\theta = 5$ types. Your main worry at that point is to try to avoid selecting any $\theta = -1$ types; this can be accomplished but only by avoiding the high σ types. In contrast when $\alpha = 0.05$ so we are trying to vacuum up all of the $\theta = 5$ types it is worth taking more of a risk with high σ types as long as their y_i is reasonably large.

The crossing of the selection boundaries and non-nestedness of the selection regions is closely tied up with the tail probability criterion and the α dependent feature of the hypothesis. If we repeat our exercise with the same G , and σ distribution, but select according to posterior means, we get the nested selection boundaries illustrated in Figure C.1. Proposition C.1 establishes this to be a general phenomenon for any distribution G .

It should be noted that the crossing of selection boundaries we have illustrated seems to have been anticipated in Henderson and Newton (2016), who consider similar tail criteria. They propose

a ranking scheme that assigns rank equal to the smallest α for which a unit would be selected as a way to resolve the ambiguities generated by crossing. We don't see a compelling decision theoretic rationale for this revised ranking rule, instead we prefer to maintain some separation between the ranking and selection problems and focus on risk assessment as a way to reconcile them.

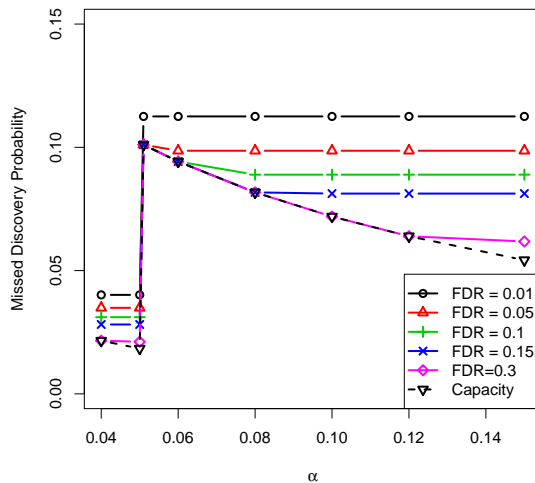


FIGURE C.2. Oracle risk evaluation for several α levels for the tail probability criterion and a discrete example with $G = 0.85\delta_{-1} + 0.10\delta_2 + 0.05\delta_5$, and $\sigma \sim U[1/2, 4]$. The solid lines correspond to the evaluation of the loss function specified in (3.1) with both capacity and FDR control constraints. The dotted line corresponds to oracle risk when only the capacity constraint is imposed.

The risk based on the loss function defined in (3.1) clearly depends on α and γ . More specifically it consists of three pieces, the leading term has the interpretation of “missed discovery” probability, which we try to minimize and the second and third pieces correspond to the FDR and capacity constraints respectively each weighted by a Lagrangian multiplier. Focusing on the first term, we have

$$\begin{aligned} \mathbb{E}[H_i(1 - \delta_i)] &= \mathbb{E}[\mathbb{1}\{\theta_i \geq \theta_\alpha\}(1 - \delta_i)] \\ &= \mathbb{P}[\theta_i \geq \theta_\alpha] - \mathbb{E}[\delta_i \mathbb{1}\{\theta_i \geq \theta_\alpha\}] \\ &= \mathbb{P}[\theta_i \geq \theta_\alpha] - \int \int_{\theta_\alpha}^{+\infty} \left[1 - \Phi((t_\alpha(\lambda, \sigma) - \theta)/\sigma)\right] dG(\theta) dH(\sigma) \end{aligned}$$

where λ , depends on α and γ , is determined by either the false discovery rate control or the capacity constraint, whichever binds.

A feature of discrete mixing distributions, G , is that the first term in the loss, $\mathbb{P}(\theta_i \geq \theta_\alpha)$, is piece-wise constant, with jumps occurring only at discontinuity points of G , while the second term depends on both θ_α and the cut-off values $t_\alpha(\lambda, \sigma)$. When the capacity constraint binds there exist ranges of α such that θ_α remains constant, while $t_\alpha(\lambda, \sigma)$ decreases for each σ , hence the risk with just capacity constraint binding is a decreasing function for α in the interval $(0.05, 0.15)$. On the other hand, when the FDR constraint binds, it can be shown that the cut-off value $t_\alpha(\lambda, \sigma)$ is constant. To see this recall that the cutoff λ determined by the FDR constraint is defined as $\mathbb{E}[(1 - v_\alpha(Y, \sigma)) \mathbb{1}\{v_\alpha(Y, \sigma) \geq \lambda\}] = \gamma \mathbb{P}(v_\alpha(Y, \sigma) \geq \lambda)$, so when θ_α is constant over a range of α the

distribution of $v_\alpha(Y, \sigma)$ does not change, and consequently the value λ is constant over that range of α .

Figure C.2 evaluates risk based on the optimal selection rule for various α and FDR levels, $\gamma \in \{0.01, 0.05, 0.1, 0.15, 0.3\}$. The solid curves correspond to risk evaluated at the optimal Bayes rule defined in Proposition 4.1. The dotted line corresponds to the risk evaluated at the Bayes rule when only the capacity constraint is imposed. As γ increases, the risk decreases as expected. For FDR levels as stringent as $\gamma = 0.01$, the FDR constraint binds and the risk is piece-wise constant. As the FDR level is relaxed, there are range of α such that capacity constraint becomes binding, and the risk decreases after the initial jump at $\alpha = 0.05$.

The evaluation of the risk for this particular example indicates that it is easier to select the top 5% individuals, those with $\theta = 5$. As we intend to select more in the right tail, we are facing more uncertainty. This also motivates a more systematic choice of (α, γ) . Although selection based on the tail probability criterion can lead to non-nested selection regions, we conclude this sub-section by demonstrating that posterior mean selection is necessarily nested.

Proposition C.1. *Let the density function of y conditional on θ and σ be denoted as $f(y|\theta, \sigma)$. If selection is based on the posterior mean, $\delta_i = \{M(y, \sigma) \geq c(\alpha, \gamma)\}$ with*

$$M(y, \sigma) := \frac{\int \theta f(y|\theta, \sigma) dG(\theta)}{\int f(y|\theta, \sigma) dG(\theta)}$$

and $c(\alpha, \gamma)$ is chosen to satisfy both the capacity constraint at level α and the FDR constraint at level γ , then the selection regions, defined as $\Lambda_{\alpha, \gamma} = \{(y, \sigma) : M(y, \sigma) \geq c(\alpha, \gamma)\}$, are nested, that is, for any $\alpha_1 > \alpha_2$, $\Lambda_{\alpha_2, \gamma} \subseteq \Lambda_{\alpha_1, \gamma}$.

REFERENCES

- ANDREWS, I., T. KITAGAWA, AND A. MCCLOSKEY (2020): “Inference on Winners,” Available from: <https://www.cemmap.ac.uk/publication/inference-on-winners-4>.
- ARMSTRONG, T. B., M. KOLESÁR, AND M. PLAGBORG-MØLLER (2020): “Robust Empirical Bayes Confidence Intervals,” Available from: <https://arxiv.org/abs/2004.03448>.
- ATHEY, S., G. IMBENS, J. METZGER, AND E. MUNRO (2019): “Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations,” Available from: <http://arxiv.org/pdf/1909.02210>.
- BAHADUR, R. (1950): “On a problem in the theory of k populations,” *Annals of Mathematical Statistics*, 21, 362–375.
- BAHADUR, R. R., AND H. ROBBINS (1950): “The Problem of the Greater Mean,” *The Annals of Mathematical Statistics*, 21, 469–487.
- BASU, P., T. CAI, K. DAS, AND W. SUN (2018): “Weighted False Discovery Rate Control in Large-Scale Multiple Testing,” *Journal of the American Statistical Association*, 113, 1172–1183.
- BECHHOFFER, R., J. KIEFER, AND M. SOBEL (1968): *Sequential Identification and Ranking Procedures*. University of Chicago Press.
- BECHHOFFER, R. 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.
- BENJAMINI, Y., AND Y. 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.
- BERGER, J. O., AND J. 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.
- BOYD, S., C. CORTES, M. MOHRI, AND A. RADOVANOVIC (2012): “Accuracy at the Top,” in *Advances in Neural Information Processing Systems 25*, ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, pp. 953–961. Curran Associates, Inc.
- CAO, H., W. SUN, AND M. R. KOSOROK (2013): “The optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing,” *Biometrika*, 100(2), 495–502.
- CHETTY, R., J. FRIEDMAN, AND J. ROCKOFF (2014a): “Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates,” *American Economic Review*, 104, 2593–2632.
- (2014b): “Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood,” *American Economic Review*, 104, 2633–2679.
- CHETTY, R., AND N. HENDREN (2018): “The impacts of neighborhoods on intergenerational mobility II: County-level estimates,” *The Quarterly Journal of Economics*, 133(3), 1163–1228.
- CHETTY, R., N. HENDREN, P. KLINE, AND E. SAEZ (2014): “Where is the land of opportunity? The geography of intergenerational mobility in the united states,” *The Quarterly Journal of Economics*, 129, 1553–1624.
- EFRON, B. (2011): “Tweedie’s Formula and Selection Bias,” *Journal of the American Statistical Association*, 106, 1602–1614.
- EFRON, B. (2016): “Empirical Bayes deconvolution estimates,” *Biometrika*, 103, 1–20.
- (2019): “Bayes, Oracle Bayes and Empirical Bayes,” *Statistical Science*, 34, 177–201.
- EFRON, B., AND C. MORRIS (1973): “Stein’s Estimation Rule and Its Competitors - An Empirical Bayes Approach,” *Journal of the American Statistical Association*, 68, 117–130.
- EFRON, B., R. TIBSHIRANI, J. STOREY, AND V. TUSHER (2001): “Empirical Bayes Analysis of Microarray Experiments,” *J. American Statistical Association*, 96, 1151–1160.
- GELMAN, A., AND P. N. PRICE (1999): “All maps of parameter estimates are misleading,” *Statistics in Medicine*, 18, 3221–3234.
- GENOVESE, C., AND L. WASSERMAN (2002): “Operating Characteristic and Extensions of the False Discovery Rate Procedure,” *Journal of the Royal Statistical Society, Series B*, 64, 499–517.
- GILRAINE, M., J. GU, AND R. McMILLAN (2020): “A New Method for Estimating Teacher Value-Added,” NBER Working Paper Series Number 27094.

- GOEL, P. K., AND H. RUBIN (1977): "On Selecting a Subset Containing the Best Population-A Bayesian Approach," *The Annals of Statistics*, 5, 969–983.
- GOLDSTEIN, H., AND D. 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.
- GU, J., AND R. KOENKER (2016a): "On a Problem of Robbins," *International Statistical Review*, 84, 224–244.
- GU, J., AND R. KOENKER (2016b): "Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective," *J. of Economic and Business Statistics*, forthcoming.
- (2017): "Empirical Bayesball Remixed: Empirical Bayes Methods for Longitudinal Data," *Journal of Applied Econometrics*, 32, 575–599.
- GUO, X., AND X. HE (2020): "Inference on Selected Subgroups in Clinical Trials," *Journal of the American Statistical Association*, pp. 1–19.
- GUPTA, S. (1956): "On a decision rule for a problem in ranking means," Mimeograph Series No. 150, Institute of Statistics, University of North Carolina, Chapel Hill.
- GUPTA, S. S., AND S. PANCHAPAKESAN (1979): *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*. Wiley.
- HANUSHEK, E. A. (2011): "The economic value of higher teacher quality," *Economics of Education review*, 30(3), 466–479.
- HECKMAN, J., AND B. SINGER (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52, 63–132.
- HENDERSON, N., AND M. 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.
- JIANG, W. (2020): "On general maximum likelihood empirical Bayes estimation of heteroscedastic iid normal means," *Electronic Journal of Statistics*, 14(1), 2272–2297.
- JIANG, W., AND C.-H. ZHANG (2021): "Compound Empirical Bayes Interval Estimation," preprint.
- 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.
- KIEFER, J., AND J. WOLFOWITZ (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics*, 27, 887–906.
- KLINE, P., AND C. WALTERS (2021): "Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination," *Econometrica*, 89, 565–592.
- KOENKER, R. (2020): "Empirical Bayes Confidence Intervals: An R Vinaigrette," available from <http://www.econ.uiuc.edu/~roger/research/ebayes/cieb.pdf>.
- KOENKER, R., AND J. GU (2015–2021): "REBayes: An R Package for Empirical Bayes Methods," Available from <https://cran.r-project.org/package=REBayes>.
- KOENKER, R., AND J. GU (2017): "REBayes: An R Package for Empirical Bayes Mixture Methods," *Journal of Statistical Software*, 82, 1–26.
- (2019): "Comment: Minimalist G -Modeling," *Statistical Science*, 34, 209–213.
- KOENKER, R., AND I. MIZERA (2014): "Convex Optimization, Shape Constraints, Compound Decisions and Empirical Bayes Rules," *J. of Am. Stat. Assoc.*, 109, 674–685.
- LAIRD, N. M., AND T. A. LOUIS (1989): "Empirical Bayes Ranking Methods," *Journal of Educational Statistics*, 14, 29–46.
- (1991): "Smoothing the non-parametric estimate of a prior distribution by roughening," *Computational Statistics & Data Analysis*, 12, 27–37.
- LIN, R., T. LOUIS, S. PADDOCK, AND G. RIDGEWAY (2006): "Loss Function Based Ranking in Two-Stage, Hierarchical Models," *Bayesian Analysis*, 1, 915–946.
- (2009): "Ranking USRDS provider specific SMRs from 1998-2001," *Health Service Outcomes Research Methodology*, 9, 22–38.
- LINDSAY, B. (1995): "Mixture Models: Theory, Geometry and Applications," in *NSF-CBMS regional conference series in probability and statistics*.

- MOGSTAD, M., J. ROMANO, A. SHAIKH, AND D. WILHELM (2020): “Inferences for ranks with applications to mobility across neighborhoods and academic achievement across countries,” preprint.
- POLYANSKIY, Y., AND Y. WU (2020): “Self-regularizing Property of Nonparametric Maximum Likelihood Estimator in Mixture Models,” preprint.
- PORTNOY, S. (1982): “Maximizing the probability of correctly ordering random variables using linear predictors,” *Journal of Multivariate Analysis*, 12, 256 – 269.
- ROBBINS, H. (1950): “A Generalization of the Method of Maximum Likelihood: Estimating a Mixing Distribution (Abstract),” *The Annals of Mathematical Statistics*, 21, 314–315.
- (1951): “Asymptotically Subminimax Solutions of Compound Statistical Decision Problems,” in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, vol. I, pp. 131–149. University of California Press: Berkeley.
- (1956): “An Empirical Bayes Approach to Statistics,” in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. I, pp. 157–163. University of California Press: Berkeley.
- SAHA, S., AND A. GUNTUBOYINA (2020): “On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising,” *Annals of Statistics*, 48(2), 738–762.
- SCHMIDT, K. D. (2014): “On inequalities for moments and the covariance of monotone functions,” *Insurance: Mathematics and Economics*, 55, 91–95.
- STOREY, J. D. (2002): “A direct approach to false discovery rates,” *Journal of the Royal Statistical B*, 64, 479–498.
- SUN, W., AND T. T. CAI (2007): “Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control,” *Journal American Statistical Association*, 102, 901–912.
- SUN, W., AND A. C. MCLAIN (2012): “Multiple Testing of Composite Null Hypotheses in Heteroscedastic Models,” *Journal of the American Statistical Association*, 107, 673–687.
- UNIVERSITY OF MICHIGAN KIDNEY EPIDEMIOLOGY AND COST CENTER (2009–2019): “Dialysis Facility Reports,” available from: <https://data.cms.gov/dialysis-facility-reports>.
- VAN DE GEER, S. (1993): “Hellinger-consistency of certain nonparametric maximum likelihood estimators,” *The Annals of Statistics*, pp. 14–44.
- VAN DER VAART, A. W. (2000): *Asymptotic statistics*. Cambridge University Press.
- WALD, A. (1950): *Statistical Decision Functions*. Wiley.