International Statistical Review

International Statistical Review (2015), 0, 0, 1–21 doi:10.1111/insr.12098

On a Problem of Robbins

Jiaying Gu and Roger Koenker

Department of Economics, University of Illinois, Urbana, IL 61801, US
E-mail: rkoenker@uiuc.edu

Summary

An early example of a compound decision problem of Robbins (1951) is employed to illustrate some features of the development of empirical Bayes methods. Our primary objective is to draw attention to the constructive role that the nonparametric maximum likelihood estimator for mixture models introduced by Kiefer & Wolfowitz (1956) can play in these developments.

Key words: Empirical Bayes; mixture models; Kiefer–Wolfowitz nonparametric maximum likelihood estimator; classification; multiple testing; false discovery rate.

1 Introduction

Herbert Robbins's Second Berkeley Symposium paper, Robbins (1951), introduced the following (deceptively) simple 'compound decision' problem, we observe

$$Y_i = \theta_i + u_i, \quad i = 1, \cdots, n, \tag{1}$$

with $\{u_i\}$ i.i.d. standard Gaussian and assume that the θ_i take values ± 1. The objective was to estimate the n-vector, $\theta \in \{-1, 1\}^n$ subject to ℓ_1 loss,

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

Robbins's visionary 1951 paper can be seen as an exercise in binary classification, but also as a precursor to the outpouring of recent work on high-dimensional data analysis and multiple testing. It can also be seen as the birth of empirical Bayes methods.

Our objective in the present note is to use this problem and several variants of it to provide a glimpse into the evolution of empirical Bayes methods. Much more comprehensive surveys of empirical Bayes methods and their modern relevance are provided by Zhang (2003) and Efron (2010); here, we aspire only to tell a more condensed version of the story, but one that highlights the critical role that the nonparametric maximum likelihood estimator (NPMLE) of Kiefer & Wolfowitz (1956) can play. Recent developments in convex optimization, as argued in Koenker & Mizera (2014), have greatly expanded the applicability of the Kiefer–Wolfowitz estimator and thereby increased the potential scope of nonparametric empirical Bayes methods.

In prior work, Koenker & Mizera (2014), Koenker (2014), Koenker & Gu (2013), and Gu & Koenker (2014), we have emphasized the role of the Kiefer–Wolfowitz NPMLE in various estimation problems typically under squared-error loss. In this paper, in contrast, we will stress its potential usefulness mainly in classification and multiple testing.

2 The Robbins Solution

Robbins begins by observing that for $n = 1$, the least favorable version of his problem occurs when we assume that the θ_i's are drawn as independent Bernoulli's with probability $p = 1/2$ that $\theta_i = \pm 1$. He then proceeds to show that this remains true for the general 'compound decision' problem with $n \geq 1$. The minimax decision rule is thus

$$\delta_{1/2}(y) = \operatorname{sgn}(y)$$

and yields constant risk,

$$R(\delta_{1/2}, \theta) = \mathbb{E}L(\delta_{1/2}(Y), \theta) = \Phi(-1) \approx 0.1586,$$

irrespective of p. And yet, something seems wrong with this procedure.

Faced with this problem, suppose we observed a sample with 'mostly positive' y_i's: would we not want to conclude that p is likely to be greater than 1/2, and having drawn this conclusion, consider modifying our cutoff strategy for estimating the θ_i's? Robbins proposes a new strategy designed, as he puts it, to 'lift ourselves by our own bootstraps.' Exploiting the common structure of the model, he proposes to *estimate* p using the method of moments (MoM) estimator, $\hat{p} = (\bar{y} + 1)/2$. Given an estimate of p, he suggests plugging it into the decision rule,

$$\delta_p(y) = \operatorname{sgn}(y - 1/2 \log((1 - p)/p)),$$

a procedure that follows immediately from the requirement that

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

exceeds one half, that is, that the posterior median of θ be 1. Of course, combining the problems in this way is not an entirely obvious move, and Robbins himself jokes that it may seem odd if some coordinates describe oysters in Maryland and others butterflies in Ecuador. Efron (2010) refers to this paradox as the problem of 'relevance' and notes that it featured prominently in early discussions of Stein shrinkage. Robbins takes a firm stand asserting the irrelevance of relevance in the context of his original problem.

Robbins's MoM approach puts us well on the way toward empirical Bayes methodology. How does it perform compared with the minimax procedure? In Figure 1, we plot empirical risk for various settings of p, against the constant risk of the minimax rule, and the oracle risk achievable when p is known. When the sample size is modest, there is a small price to pay near $p = 1/2$ for using the somewhat inaccurate MoM $\hat{p}$, but this is compensated in the tails where the empirical Bayes risk is much lower than that of the minimax risk. Asymptotically, of course, as Robbins stresses, $\hat{p} \to p$ and the small advantage of the minimax rule vanishes, and the empirical Bayes rule dominates. It is clear that the Robbins solution constituted a direct challenge to the Wald minimax view of decision theory.

2.1 A Hierarchical Bayes Variation

One way to attenuate the modest disadvantage of the $\hat{p}$ rule when p is near $1/2$ would be to employ some form of (Bayesian) shrinkage strategy. For example, we may consider replacing the MoM $\hat{p}$ procedure by a more formal Bayes procedure that concentrates prior mass for p

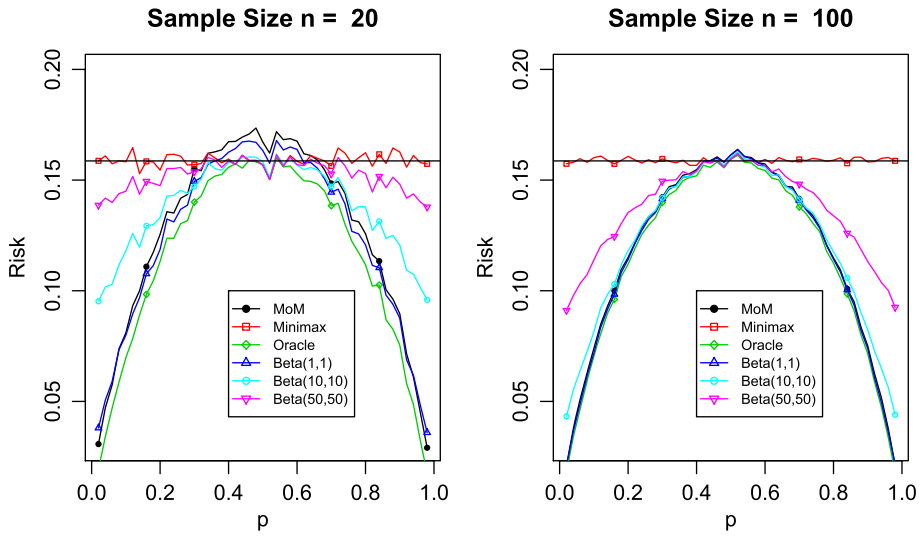

Figure 1. *Empirical risk of various decision rules for the original Robbins problem. Mean loss is computed over 1000 replications.*

near $p = 1/2$. A natural prior for p would thus be $\mathcal{B}(a, a)$, a beta distribution that becomes more concentrated near 1/2 as $a \to \infty$. Given our log likelihood,

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

adding the log prior,

$$f(p) = a \log(p) + a \log(1 - p) - \log B(a, a),$$

has the effect of concentrating the posterior distribution of p toward 1/2. As a side benefit, the beta prior acts as a log barrier penalty for the unconstrained maximum likelihood estimator and thus avoids the potential embarrassment of the MoM estimator when $\hat{p} \notin [0, 1]$. In Figure 1, we have included three variants of this beta prior rule with $a \in \{1, 10, 50\}$ to illustrate different degrees of shrinkage. For $n = 20$, it can be seen that they deliver better performance than the MoM procedure while sacrificing some of its advantage when p is near 0 or 1. When n is 100, the differences are almost imperceptible near $p = 1/2$, but the cost in the tails is still apparent for the two largest values of a.

To be more explicit about the beta prior procedure, for the Robbins (1951) setup, we have $(\theta_1, \ldots, \theta_n)$, each taking values in $\{1, -1\}$, with probability p and $(1 - p)$. We do not know p, so we assign a prior distribution for p with density function $f(p)$. The observables are $y_i|\theta_i \overset{iid}{\sim} \mathcal{N}(\theta_i, 1)$. So the posterior for θ is

$$p\left(\theta_i = 1 \mid y_i, y^{(i)}\right) = \frac{\int h\left(y_i, y^{(i)} \mid \theta_i = 1\right) pf(p)dp}{g\left(y_i, y^{(i)}\right)}$$

where $y^{(i)}$ is the observed sample deleting the i^{th} observation. The denominator is

$$g(y_i, y^{(i)}) = \int \prod_{i=1}^{n} (p\varphi(y_i - 1) + (1 - p)\varphi(y_i + 1)) f(p)dp$$

and the numerator is

$$\int h(y_i, y^{(i)} \mid \theta_i = 1) p f(p) dp = \int \varphi(y_i - 1) f(y^{(i)}) p f(p) dp$$
$$= \varphi(y_i - 1) \int p \prod_{j \neq i} (p\varphi(y_j - 1) + (1-p)\varphi(y_j + 1)) f(p) dp.$$

Hence, the posterior probability for $\theta_i = 1$ given the data is

$$p(\theta_i = 1 \mid y_1, \ldots, y_n) = \frac{\varphi(y_i - 1)\bar{p}}{\varphi(y_i - 1)\bar{p} + \varphi(y_i + 1)(1 - \bar{p})},$$

where $\bar{p}$ is the posterior mean of p given the data $y^{(i)}$.

$$\bar{p} = \frac{\int p \prod_{j \neq i} (p\varphi(y_j - 1) + (1-p)\varphi(y_j + 1)) f(p) dp}{\int \prod_{j \neq i} (p\varphi(y_j - 1) + (1-p)\varphi(y_j + 1)) f(p) dp}.$$

The Bayes rule under ℓ_1 loss leads to $\hat{\theta}_i = 1$ if $P(\theta_i = 1 \mid y_1, \ldots, y_n) > 1/2$, which gives a cut-off rule. In the simulation conducted for Figure 1, we have ignored the dependence of $\bar{p}$ on i. It is straightforward to construct a Gibbs sampler for this problem and it is reassuring to find agreement with the foregoing approach is excellent.

2.2 A Combinatorial Interpretation

A combinatorial interpretation of the foregoing hierarchical approach was already anticipated by Robbins (1951). He partitions the sample space Ω of $(\theta_1, \ldots, \theta_n)$ containing 2^n possible elements into Ω_k with $k = 0, 1, \ldots, n$. We say $\theta = (\theta_1, \ldots, \theta_n) \in \Omega_k$ if exactly k out of n elements in θ equal 1. Each partition Ω_k contains $\binom{n}{k}$ numbers of different θ's. Let $h(\theta)$ be a probability mass function for θ, for example, one such $h(\theta)$ could attach weights $b_k = \binom{n}{k}^{-1}/(n+1)$ to each element in a partition Ω_k. The interpretation of this weighting is that each element within a partition is treated equally, and each partition is also given equal weight.

The Bayes rule under ℓ_1 loss and a particular $h(\theta)$ asserts that $\theta_i = 1$ when

$$\sum_{k=0}^{n} b_k \left[\sum_{\Omega_{k,i}^+} \mathcal{L}(y; \theta) - \sum_{\Omega_{k,i}^-} \mathcal{L}(y; \theta) \right] \geq 0$$

where $\Omega_{k,i}^+ := \{\theta \in \Omega_k, \theta_i = 1\}$ and $\Omega_{k,i}^- := \{\theta \in \Omega_k, \theta_i = -1\}$ and $\mathcal{L}(y; \theta)$ is the likelihood of observing $y = (y_1, \ldots, y_n)$ given θ.

To be more explicit, consider the case $n = 3$. We have four partitions of Ω, that is, $\Omega_{0,1}^+ = \emptyset$, $\Omega_{0,1}^- = \{(-1,-1,-1)\}$; $\Omega_{1,1}^+ = \{(1,-1,-1)\}$, $\Omega_{1,1}^- = \{(-1,1,-1), (-1,-1,1)\}$; $\Omega_{2,1}^+ = \{(1,1,-1), (1,-1,1)\}$, $\Omega_{2,1}^- = \{(-1,1,1)\}$ and $\Omega_{3,1}^+ = \{(1,1,1)\}$, $\Omega_{3,1}^- = \emptyset$. Focusing on $i = 1$ and abbreviating $\varphi_i^{\pm} = \varphi(y_i \pm 1)$, the Bayes rule estimates $\theta_1 = 1$ if

$$b_3 \varphi_1^+ \varphi_2^+ \varphi_3^+ + b_2 \left(\varphi_1^+ \varphi_2^+ \varphi_3^- + \varphi_1^+ \varphi_2^- \varphi_3^- - \varphi_1^- \varphi_2^+ \varphi_3^+ \right)$$
$$+ b_1 \left(\varphi_1^+ \varphi_2^- \varphi_3^- - \varphi_1^- \varphi_2^+ \varphi_3^- - \varphi_1^- \varphi_2^- \varphi_3^+ \right) - b_0 \varphi_1^- \varphi_2^- \varphi_3^- \geq 0$$

Returning to the hierarchical Bayes rule, it estimates $\theta_1 = 1$ if

$$\int p^3 f(p)dp\varphi_1^+\varphi_2^+\varphi_3^+ + \int p^2(1-p)f(p)dp\left(\varphi_1^+\varphi_2^+\varphi_3^- + \varphi_1^+\varphi_2^-\varphi_3^- - \varphi_1^-\varphi_2^+\varphi_3^+\right) + \int p(1-p)^2 f(p)dp\left(\varphi_1^+\varphi_2^-\varphi_3^- - \varphi_1^-\varphi_2^+\varphi_3^- - \varphi_1^-\varphi_2^-\varphi_3^+\right) - \int (1-p)^3 f(p)dp\varphi_1^-\varphi_2^-\varphi_3^- \geq 0$$

For $b_k = \binom{3}{k}^{-1}/4$ with $k = 0, 1, 2, 3$, Robbins's Bayes rule is equivalent to the hierarchical Bayes rule with $f(p) = 1$, that is, the prior distribution for the proportion p is uniform on $[0, 1]$. More generally, by induction, we conjecture that the equivalence holds for any n with $b_k = \binom{n}{k}^{-1}/(n+1)$. Indeed, the factor $1/(n+1)$ indicates that equal weight is associated with each partition of the sample space, which is nothing but a discretization $\left[0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{k}{n}, \ldots, 1\right]$ of the $[0, 1]$ interval of the proportion p. Within each partition, there are $\binom{n}{k}$ elements, and the weights $\binom{n}{k}^{-1}$ again treats them equally. Given this connection, it is easy to generalize to cases where $f(p)$ is taken to be the density of $\mathcal{B}(a, b)$. The corresponding $b_k = \int p^k(1-p)^{n-k} f(p)dp = B(a+k, b+n-k)/B(a, b)$. To fix ideas, consider $a = b = 2$, then $b_k = \binom{n}{k}^{-1} \frac{6(k+1)(n-k+1)}{(n+1)(n+2)(n+3)}$. For $n = 3$, it approximates the $\mathcal{B}(2, 2)$ prior by probability weights $[1/5, 3/10, 3/10, 1/5]$ at atoms $[0, 1/3, 2/3, 1]$.

2.3 A Multiple Testing Perspective

The link to the multiple testing literature for the Robbins problem is immediately clear because estimation of $\theta \in \{-1, 1\}^n$ is essentially a testing problem in which we have weighed false discovery and false non-discovery equally. If we treat $\theta = -1$ as the null hypothesis and $\theta = 1$ as the alternative, a p-value procedure based on $T_i = 1 - \Phi(Y_i + 1)$ with cutoff $\Phi(-1)$ the decision rule,

$$\delta_p(T) = \text{sgn}(\Phi(-1) - T)$$

is equivalent to the minimax rule, $\delta(y) = \text{sgn}(y)$. If, instead, we would like to fix the marginal false discovery rate (mFDR) at some level and optimize marginal false nondiscovery rate (mFNR) a modified p-value cutoff can be constructed, and this would be equivalent to replacing our symmetric ℓ_1 loss for the estimation/classification problem by an asymmetric linear loss.

A p-value testing procedure that is equivalent to the empirical Bayes rule estimator described earlier for the Robbins problem can also be constructed. Under the null that $Y_i \sim \mathcal{N}(-1, 1)$, $T_i = 1 - \Phi(Y_i + 1) \sim U[0, 1]$, while if $Y_i \sim \mathcal{N}(1, 1)$,

$$\mathbb{P}(T_i < u) = \mathbb{P}(Y_i + 1 > \Phi^{-1}(1-u)) = 1 - \Phi(\Phi^{-1}(1-u) - 2).$$

Thus, under the null, the density of T is $f_0(t) \equiv 1$, and under the alternative,

$$f_1(t) = \varphi(\Phi^{-1}(1-t) - 2)/\varphi(\Phi^{-1}(1-t)),$$

and the posterior probability of $\theta_i = 1$ given t_i and assuming for the moment that the unconditional probability, $p = \mathbb{P}(\theta_i = 1)$ is known, is given by,

$$\mathbb{P}(\theta = 1|t, p) = \frac{pf_1(t)}{pf_1(t) + (1-p)f_0(t)}.$$

Under symmetric loss we were led to the posterior median so $\hat{\theta}_i = 1$ if $\mathbb{P}(\theta_i = 1|T_i, p) > 1/2$, which is equivalent to the p-value rule,

$$T_i < 1 - \Phi(1 + 0.5 \log((1-p)/p)).$$

Again, we are led back to the problem of estimating p. In these two point mixture problems ℓ_1 loss is equivalent to $0-1$ loss since the median and the mode are identical.

The special structure of the Robbins problem with its two point mixture ensures a strong equivalence between so-called 'p-value' and 'z-value' multiple testing methods. However, when this model is relaxed to allow more general mixtures, this equivalence breaks down as recent work by Sun & Cai (2007) and Efron (2008a) has pointed out. In particular, when variances are heterogeneous, the p-value approach also breaks down as Cao *et al.* (2013) have recently pointed out. We will return to this point in Section 4.2 in the succeeding text; however, before doing so, we would like to briefly consider a grouped version of the original Robbins problem, which can be viewed as an extension of the original Robbins problem with an additional level of hierarchy.

3 A Grouped Robbins Problem

A natural generalization of Robbins's original problem considers a grouped setting in which

$$Y_{ij} = \theta_{ij} + u_{ij}, \quad i = 1, \cdots, n, \quad j = 1, \cdots, m,$$

with $\{u_{ij}\}$ i.i.d. standard Gaussian as before, and $\theta_{ij} = 1$ with probability p_i and $\theta_{ij} = -1$ with probability $1-p_i$, and independent over $j = 1, \cdots, m$. In this framework, we can consider 'group specific' p_i that vary within the full sample yielding a nonparametric mixture problem. In the multiple testing context, this grouped model has been considered by Efron (2008b), Cai & Sun (2009), and Muralidharan (2010) among others.

Remarkably, Robbins (1951) anticipated formulations like this as well. In the final section of that paper, Robbins considers a general mixture problem in which we have observations from a density $\varphi(y|\theta)$, with θ drawn from a distribution, F, so the observations come from the mixture density,

$$g_F(y) = \int \varphi(y|\theta) dF(\theta).$$

He describes a 'generalized maximum likelihood estimator' for the mixing distribution F, and the corresponding Bayes rule for estimating the θs. Robbins mentions an abstract, Robbins (1950) in which he had announced earlier the finding that 'under certain conditions this method is consistent as $n \to \infty$.' This abstract is referred to again in Robbins (1956), we have found no further elaboration of the result by Robbins. A formal treatment seems to have appeared only with the paper of Kiefer & Wolfowitz (1956), who mention Robbins' abstract and comment that they found no further elaboration of these ideas.

Almost another 20 years elapsed before Laird (1978) provided a viable computational method for such generalized MLEs employing the EM algorithm. Laird's EM implementation rekindled considerable interest in the general mixture problem, notably in the work of Heckman & Singer (1984). More recently, Zhang (2003) and Jiang & Zhang (2009) have demonstrated the effectiveness of the Kiefer–Wolfowitz approach for the classical Gaussian compound decision

problem, again relying on the EM algorithm for simulation results. Unfortunately, the notoriously slow convergence of EM for mixture problems of this type seems to have impeded further progress.

Recent developments in convex optimization have, however, dramatically reduced the computational effort required for the Kiefer–Wolfowitz MLE. Motivated by the Jiang & Zhang (2009) results, Koenker & Mizera (2014) describe implementations for binomial and Gaussian location mixtures that employ modern interior point methods drastically improving both accuracy and speed over prior EM methods. Gu & Koenker (2014) describe several extensions of this approach to longitudinal models. In our longitudinal Robbins setting, denoting $g_i = g(y_{i1}, \cdots, y_{im})$, we can formulate the variational problem as follows:

$$\max_{F \in \mathcal{F}} \left\{ \sum_{i=1}^{n} \log g_i \mid \int_0^1 \prod_{j=1}^{m} (p\varphi(y_{ij} - 1) + (1-p)\varphi(y_{ij} + 1)) dF(p) = g_i,\ i = 1, \cdots, n \right\}$$

As noted by Laird and elaborated by Lindsay (1995) solutions, $\hat{F}$, in the space, $\mathcal{F}$, of distribution functions are discrete with $k \leq n$ mass points. The problem is strictly convex because we are maximizing a sum of strictly convex functions subject to linear equality and inequality constraints, so solutions are unique. Uniqueness is all the more remarkable given the notorious multimodality of finite mixture models.

We can discretize the problem by letting p take values $\{p_1, \cdots, p_K\}$ on a relatively fine grid of [0, 1], and write,

$$\max_{f} \left\{ \sum_{i=1}^{n} \log(g_i) \mid Af = g, f \in \mathcal{S} \right\}$$

where $g_i = g(y_{i1}, \cdots, y_{1m})$, A denotes the n by K matrix with typical element

$$A_{ik} = \prod_{j=1}^{m} (p_k \varphi(y_{ij} - 1) + (1 - p_k)\varphi(y_{ij} + 1))$$

and f is an K-vector in the $K - 1$ dimensional simplex, $\mathcal{S}$. It proves convenient to solve the dual problem,

$$\min_{\nu} \left\{ \sum_{i=1}^{n} \nu_i \mid A^\top \nu \leq n 1_K, \nu \geq 0 \right\}$$

where 1_K denotes an K vector of ones. Our implementation of the estimator relies on the open source R package REBayes, Koenker (2012), which relies in turn on the convex optimization package Rmosek, Friberg (2012) and Andersen (2010).

The crucial advantage of the group structure is that it permits the groups to have different p_i's and the Kiefer–Wolfowitz procedure enables us to estimate the points of support and the associated mass of these points. Within groups, the decision rule operates as we have described earlier in Section 2, but the proposed compound decision rules borrow strength across groups to produce improved estimates of the group specific mixture probabilities and thereby better estimates of the θ_{ij}'s. We consider three variants of the compound decision procedure for the grouped setting, each adapted to the application specific problem dimensions m and n. For small m, say $m < 15$, the exact likelihood can be used labeled as "Robbins" in Table 1. However,

Table 1. *Mean absolute error of several methods of estimation relative to oracle performance.*

n	m	Minimax	MoM	ECF	CLT	Bin	Robbins
200	5	1.668	1.599	1.472	1.357	1.344	1.343
200	10	1.300	1.290	1.224	1.043	1.043	1.043
200	100	1.305	1.036	1.048	1.011	1.011	

We compare three variants of the Bayes rule constructed from the Kiefer-Wolfowitz MLE procedure with the minimax procedure, the Robbins's method of moments (MoM) procedure and an empirical characteristic function (ECF) approach suggested by Jin and Cai (2007). Performance is measured by mean absolute error relative to the oracle performance of the procedure described in the text. The exact likelihood procedure labeled Robbins in the table is computationally prohibitive when m is large, accounting for missing entry of the table, but in those cases the approximate likelihood methods labeled CLT and Bin are adequate substitutes.

for larger m, this becomes numerically unstable and we propose using $\bar{y}_i = m^{-1}\sum_{j=1}^{m} y_{ij}$ as an 'almost' sufficient statistic for the ensemble $(y_{i1}, \cdots, y_{im})$. Two variants of the latter approach are considered, one that simply adopts the normal approximation for $\bar{y}_i \sim \mathcal{N}(2p_i - 1, 1 + 4p_i(1-p_i)/m)$, labeled CLT in Table 1, the second that employs a normal-binomial mixture density labeled Bin in Table 1 that represents the exact likelihood of the $\bar{y}_i$'s for the present problem. The normal approximation is expected to be adequate for $m \gg 30$, while the normal-binomial model offers a useful intermediate approach.

We would now like to compare performance of the empirical Bayes rules corresponding to these procedures for several instances of the grouped Robbins problem. In addition to our three variants based on the Kiefer–Wolfowitz procedure, we consider four other estimators of the θ_{ij}'s:

- The original (naive) minimax procedure: $\hat{\theta}_{ij} = \operatorname{sgn}(y_{ij})$;
- A within group Robbins MoM procedure with $\hat{\theta}_{ij} = \operatorname{sgn}\left(y_{ij} - \frac{1}{2}\log\left((1-\hat{p}_i)/\hat{p}_i\right)\right.$ for $\hat{p}_i = (\bar{y}_i + 1)/2$;
- An empirical characteristic function procedure proposed by Cai & Sun (2009) employing a group specific $\hat{p}_i$ proposed by Jin (2008);
- An oracle procedure based on the Bayes rule $\hat{\theta}_{ij} = \operatorname{sgn}\left(y_{ij} - \frac{1}{2}\log\left((1-p_i)/p_i\right)\right)$ with known p_i's.

We will focus on a simple special case in which $\mathbb{P}(\theta_{ij} = 1) \equiv p_i \sim \frac{1}{4}\delta_{0.1} + \frac{3}{4}\delta_{0.3}$, so unconditionally there is a 0.75 probability of a $\theta_{ij} = -1$, the notional 'null' case, versus a 0.25 probability of the alternative, $\theta_{ij} = 1$. The number of groups, n, and the number of members of each group, m, are crucial in determining relative performance.

As can be seen from the Table, the three Kiefer–Wolfowitz Bayes rule procedures perform well relative to the prior proposals. As long as m is moderately large they also perform nearly as well as the oracle procedure. Not surprisingly, settings with small m provide the all-knowing oracle with more of an advantage, but we would stress that even in those cases, there is benefit in the NPMLE methods when compared with earlier proposed procedures.

Having restricted attention to the original Robbins problem with only two *known* values of θ until now, it is finally time to relax this condition and consider what can be done in models with more general mixtures. Chekhov's well known maxim of dramatic economy maintains that

if there is a gun in the first act it should be fired in the final act. We have seen the Kiefer–Wolfowitz nonparametric MLE in the context of the grouped version of the Robbins problem, let us see what it can do in more general mixture settings.

4 The Robbins Problem with Unknown θ's

A focus of all empirical Bayes compound decision problems, no matter what loss function is used, is estimation of the mixing distribution we have denoted F. When the support of θ is restricted to $\{-1, 1\}$ this requires only the estimation of a single probability, p, while in our grouped version of the original Robbins problem we need a distribution that assigns mass to p's in $[0, 1]$ corresponding to the various groups. In this section we consider the more general case of estimating a mixing distribution, F for a real valued θ with general support. This opens the way to consideration of composite testing problems.

Once we abandon the assumption that there are only two known points of support for the mixing distribution of θ, we are faced with what appears to be a general Gaussian deconvolution problem. We observe i.i.d. $Y_1, Y_2, \cdots, Y_n$ from the mixture density,

$$g(y) = \int \varphi(y - \theta) dF(\theta),$$

but rather than focus on the notoriously difficult problem of estimating the distribution F, we will instead continue to focus on the prediction of the θ_i's given the data. For prediction, it suffices to find the posterior of each θ_i given the Y_i's. If we maintain our ℓ_1 loss criterion, our $\hat{\theta}_i$ should be the median of this posterior. Replacing ℓ_1 by ℓ_2 loss would lead us to posterior means instead of medians. For ℓ_2 loss, the Bayes rule, or posterior mean, is given by Robbins (1956) as,

$$\delta(y) = y + g'(y)/g(y).$$

Efron (2011) refers to this expression as the Tweedie formula. Tukey (1974) provides an earlier attribution to Arthur Eddington appearing in Dyson (1926). If we were so fortunate as to know the mixing distribution F, we would, having seen $Y_i = y$, and adhering to the Bayes rule, predict $\hat{\theta}_i = \delta(y)$. Knowing F seems a bit implausible, but because we only need g, the marginal density of the Y_i's, it is tempting to simply plug-in a reasonable estimator of g and use,

$$\hat{\delta}(y) = y + \hat{g}'(y)/\hat{g}(y).$$

In the Gaussian case, and more generally in other exponential family settings, we should be aware that the Bayes rule must be monotone in y whatever F might be. This constraint restricts the class of reasonable estimators of g: Koenker & Mizera (2014) describe two general approaches to this problem both involving a penalized maximum likelihood strategy. The first imposes the monotonicity constraint directly by maximizing the log likelihood,

$$\ell(g) = \sum_{i=1}^{n} \log g(Y_i)$$

subject to a convexity constraint on the function,

$$K(y) = \tfrac{1}{2} y^2 + \log g(y).$$

The convexity constraint can be formulated as a cone constraint in a discretized version of the problem and solved by interior point methods. In more general settings, especially those involving multiple testing with composite null and alternatives, we may require more explicit estimation of F. This leads us back to Robbins and the Kiefer–Wolfowitz MLE.

In variational form the Kiefer–Wolfowitz estimator solves,

$$\max_{F \in \mathcal{F}} \left\{ \sum_{i=1}^{n} \log g(Y_i) \mid g(y) = \int \varphi(y-\theta) dF(\theta) \right\}$$

where $\mathcal{F}$ denotes the set of distributions on $\mathbb{R}$. In this form, it looks very much like we are back to deconvolution, Efron (2014) refers to such mixture problems as Bayesian deconvolution, and adopts a parametric specification of the mixing distribution F. Rather than relying on a selection of a parametric model or empirical characteristic function methods, Koenker & Mizera (2014) propose a simple discretization that yields yet another convex optimization problem. Taking a relatively fine, equally spaced grid for the support of F, say $\{t_1, \cdots, t_m\}$ containing the empirical support of the sample, we can write an approximate version the variational problem as,

$$\min_{f} \left\{ -\sum_{i=1}^{n} \log g_i \mid g = Af, \ f \in \mathcal{S}_m \right\}$$

where $g_i = g(Y_i)$ denotes the ith element of the n vector g, A denotes the n by m matrix with typical element, $A_{ij} = \varphi(y_i - t_j)$ and $\mathcal{S}_m = \{s \in \mathbb{R}^m \mid s \geq 0, 1^\top s = 1\}$ denotes the $(m-1)$ dimensional simplex in $\mathbb{R}^m$. This problem is evidently convex, a convex objective to be minimized subject to linear equality and inequality constraints, and again the problem can be efficiently solved by interior point methods. Accuracy of the solution can be controlled by refining the grid and convergence tolerance of the optimization algorithm.

Problems with sample sizes up to a few thousand and m around 300 can be accurately solved in less than a second or two, while earlier EM methods require several minutes to achieve an even less reliable solution. For larger sample sizes, we have found it expedient to bin the Y_i's to further accelerate the estimation process. For observations with Gaussian tail behavior, we have found that binning large samples into a few hundred bins substantially reduces cpu effort without materially sacrificing accuracy.

4.1 Estimation

There is an extensive recent literature on variants of the Robbins problem that assume that the model (1) holds, with F assigning $\theta_i = 0$ with high probability, the 'haystack,' and with lesser probability a few needles, $\hat{\theta}_i \neq 0$, are hidden in this 'haystack.' An influential early paper in this line is Johnstone & Silverman (2004) that compares a variety of hard and soft thresholding procedures with several parametric empirical Bayes procedures. Under squared error loss we can compare the non-parametric shape constrained estimator and the Kiefer–Wolfowitz estimator of g described earlier to construct estimates of the θ_i's using Tweedie's formula. The same methods are immediately relevant for testing problems when considering composite null and alternative hypotheses.

Table 2, reproduced from Koenker & Mizera (2014), reports mean squared error results from a small simulation experiment following the design of Johnstone & Silverman (2004). In each

Table 2. *Comparison of several procedures for the Johnstone and Silverman 'Needles and Haystack' design.*

	$k = 5$				$k = 50$				$k = 500$			
Estimator	$\theta = 3$	$\theta = 4$	$\theta = 5$	$\theta = 7$	$\theta = 3$	$\theta = 4$	$\theta = 5$	$\theta = 7$	$\theta = 3$	$\theta = 4$	$\theta = 5$	$\theta = 7$
$\hat{\delta}_M$	37	34	21	11	173	121	63	16	488	310	145	22
$\hat{\delta}_{KW}$	33	30	16	8	153	107	51	11	454	276	127	18
J-S Min	34	32	17	7	201	156	95	52	829	730	609	505

The best of the 18 procedures considered by Johnstone and Silverman for each column of the table is reported in the last row of the table. The first row reports performance of the Tweedie formula estimator based on the monotonized Bayes rule maximum likelihood estimator (MLE), and the second row reports the Tweedie estimator based on the Kiefer–Wolfowitz MLE. Each table entry reports sum of squared errors over the sample of $\boldsymbol{n} = 1000$ observations averaged over 1000 replications for the first two rows. The last row is taken directly from the table of Johnstone and Silverman. As a benchmark, the naive MLE, $\hat{\delta}(y) = y$, would have expected loss of 1000 in all settings.

cell of the table, we report a sum of squared errors over the $n = 1000$ observations, averaged over 1000 replications, for the $\hat{\delta}_M$, the monotone Bayes rule estimator described at the beginning of Section 4, and Kiefer–Wolfowitz MLE. The last row of the table reports the performance of the best of the 18 procedures considered by Johnstone & Silverman (2004). Cells of the table differ in the number of non-zero θ_i's, denoted by k, and the value of the non-zeros, denoted θ in the table headings. Some further simulation evidence involving similar models is presented in Koenker (2014).

When the non-null θ_i's are truly rare, so $k = 5$, the parametric empirical Bayes procedures of Johnstone & Silverman (2004) are quite effective, but when the proportion of non-null θ_i is larger the Kiefer–Wolfowitz method is clearly superior. One explanation for this is that all of the Johnstone and Silverman procedures are good at shrinking the observed y_i's toward zero, but not so good at shrinking toward the non-null value θ_A. In two point mixture problems like those of the Johnstone and Silverman design, the Kiefer–Wolfowitz estimator is remarkably good at identifying that there are two points of support and estimating their locations. Hence, the Tweedie formula based on the Kiefer–Wolfowitz estimator is, at least when the sample size is reasonably large, quite good at shrinking the non-null y_i's toward an accurate estimate of θ_A. When we spread out the non-null θ_i's, this advantage is attenuated, and we will explore this further in the next subsection. It may also be worth noting that the Kiefer–Wolfowitz Bayes rule places no special significance on likelihood that $\theta_i = 0$, it simply estimates a few points of support for the mixing distribution and one of these estimated points of support is generally close enough to zero to produce good performance.

4.2 Classification and Multiple Testing

Suppose, instead of estimating the θ_i's we were only required to 'classify' them, that is, given the Y_i's, we must decide whether their associated θ_i's belong to a specified set A, or not. The set A is interpreted in such a way that $\theta_i \in A$ are deemed 'uninteresting,' while $\theta_i \notin A$ 'merit further investigation.' Typically, A contains zero. This brings us quite close to the realm of multiple testing with composite null and alternative. We will consider a heterogeneous Gaussian framework introduced by Sun & McLain (2012),

$$Y_i = \theta_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right), \quad \theta_i \sim G(\theta),$$

with σ_i's known constants. Let $H_i = 0$ if $\theta_i \in A$, and $H_i = 1$, otherwise, and denote our decision rule $\delta = \delta(y|\sigma)$ taking values 0 or 1, with the loss function,

$$L(\delta, H) = \begin{cases} 1-\tau & \text{if } \delta = 1, \text{ and } H = 0, \\ 0 & \text{otherwise,} \\ \tau & \text{if } \delta = 0, \text{ and } H = 1. \end{cases}$$

Thus, $1-\tau$ and τ denote the relative costs of type I and type II error, respectively. We will assume that the H_i are i.i.d. Bernoulli with probability p, so,

$$\{Y_i|H, \sigma\} \sim (1-H_i)F_0 + H_i F_1,$$

where F_0 and F_1 have densities

$$f_0(y|\sigma) = (1-p)^{-1} \int_A \varphi((y-\theta)/\sigma)/\sigma dG(\theta),$$

and

$$f_1(y|\sigma) = p^{-1} \int_{A^c} \varphi((y-\theta)/\sigma)/\sigma dG(\theta).$$

The marginal density of the Y_i's is given by $f(y|\sigma) = (1-p)f_0(y|\sigma) + pf_1(y|\sigma)$. Expected loss, or Bayes risk, is

$$\begin{aligned} R(\delta) &= (1-\tau)\mathbb{P}(\delta = 1, H = 0) + \tau\mathbb{P}(\delta = 0, H = 1) \\ &= (1-p)(1-\tau)\int \delta dF_0 + \tau p(1 - \int \delta dF_1) \\ &= \tau p + \int \delta[(1-p)(1-\tau)f_0 - \tau p f_1]dy. \end{aligned}$$

Minimizing R, we obtain a likelihood ratio criterion, which after transformation can be formulated in terms of local false discovery rate,

$$\text{Lfdr}(y|\sigma) = (1-p)f_0(y|\sigma)/f(y|\sigma),$$

rejecting when Lfdr is sufficiently small, so

$$\delta(y|\sigma) = \mathbb{I}(\text{Lfdr}(y|\sigma) < \lambda).$$

Thus, our objective is to find a cutoff value λ_α for the ordered Lfdr values so that the marginal false discovery rate, mFDR, is controlled at a prescribed value of α. The mFDR has a nice Bayesian interpretation, for example, Storey (2002), as a multiple testing analogue of the Type I error in classical hypothesis testing,

$$\begin{aligned} \text{mFDR}(\lambda) &= \mathbb{P}\{\theta \in A|\text{Lfdr}(y|\sigma) < \lambda\} \\ &= \frac{\mathbb{P}\{\theta \in A, \text{Lfdr}(y|\sigma) < \lambda\}}{\mathbb{P}\{\text{Lfdr}(y|\sigma) < \lambda\}} \\ &= \frac{\int\int \mathbb{I}(\text{Lfdr}(y|\sigma) < \lambda)\text{Lfdr}(y|\sigma)f(y|\sigma)f(\sigma)d\sigma dy}{\int\int \mathbb{I}(\text{Lfdr}(y|\sigma) < \lambda)f(y|\sigma)f(\sigma)d\sigma dy} \end{aligned}$$

mFDR can therefore be estimated by the empirical analogue,

$$\widehat{\text{mFDR}} = \frac{\sum_{i=1}^{m} \mathbb{I}\left(\widehat{\text{Lfdr}}(y|\sigma) < \lambda\right) \widehat{\text{Lfdr}}(y|\sigma)}{\sum_{i=1}^{m} \mathbb{I}\left(\widehat{\text{Lfdr}}(y|\sigma) < \lambda\right)},$$

and justifies the data driven procedure for choosing λ, for example, Sun & Cai (2007) and Sun & McLain (2012), as the kth order statistic of the $\widehat{\text{Lfdr}}$ where,

$$k = \max\left\{ i \,|\, i^{-1} \sum_{j=1}^{i} \widehat{\text{Lfdr}}_{(j)}(y|\sigma) < \alpha \right\},$$

Heterogeneity of variances introduces some potential anomalies in this mFDR criterion. To illustrate this, Cao *et al.* (2013) consider an example with observations from the mixture density,

$$f(y) = (1-p)\varphi(y) + p\varphi((y-\mu)/\sigma)/\sigma$$

with $p = 0.1, \mu = 2.5$, and $\sigma = 0.5$. For this seemingly one-sided testing problem, $\text{Lfdr}(y|\sigma)$ is however not monotone in x and mFDR thresholding leads to a closed interval rejection region for the Y_i. More importantly, it is no longer possible to achieve certain levels of mFDR. For example, in the model earlier, $\text{Lfdr}(y|\sigma) > 0.06$ so any mFDR level below this is unachievable. This is hardly surprising given that it is obviously difficult to distinguish the two components of this mixture; inevitably any collection of rejections will be marred by a substantial number of 'non-discoveries' because the two components of the mixture overlap substantially. It is worth mentioning that the p-value approach for thresholding the one-sided p-values $\mathbb{P}\{\mathcal{N}(0,1) > Y_i\}$ is ill-behaved in that it fails to satisfy the monotone likelihood ratio condition illustrated in Cao *et al.* (2013) and thresholding procedures controlling false discovery rate exactly at level α may no longer be optimal in the sense of minimizing false non-discovery rate.

In the foregoing discussion, we have assumed that the non-null value of σ was known, and fixed in repeated sampling, however, it is plausible that in many applications, the Y_i may have distinct σ_i's, as long as these are still known, we can consider procedures that control overall mFDR by pooling the resulting Lfdr statistics and computing a universal cutoff. It might seem that controlling the mFDR level for each σ value has some appeal, but in some circumstances if overall false discovery rate control is all that is desired, better power, that is, better false non-discovery rate, may be achieved by a universal cutoff. When the σ_i are unknown, there may be an opportunity to estimate their distribution and integrate them out, but we will not pursue this here.

4.3 Implementation of the Oracle Rules

The oracular setting of the previous subsection provides a useful benchmark for the more pragmatic procedures we will now consider. Most existing implementations of false discovery control for these models rely on some form of deconvolution method based on the empirical characteristic function. Given the success of the Kiefer–Wolfowitz MLE in closely related problems, it seems worthwhile to explore its performance in the multiple testing arena. We will focus attention on simulation settings employed by Sun & McLain (2012); their composite null and non-null behavior arising from a mixture of beta densities provides an especially challenging environment for Kiefer–Wolfowitz methods that would seem to favor discrete alternatives.

4.3.1 *Some Simulation Evidence: Sun and McLain Model 1*

In their initial simulation settings, Sun and McLain consider models with non-null density,

$$f(\theta) = q\beta(\theta, 3, 3) + (1-q)\tilde{\beta}(\theta, 3, b),$$

where $\beta(\cdot, a, b)$ denotes a β density with parameters a and b, and $\tilde{\beta}(\cdot, a, b)$ denotes a reversed β density supported on $[0, 2]$ and $[-2, 0]$, respectively. We illustrate a family of these densities for $b \in \{1, 2, 3, 4, 5\}$ in Figure 2. Observations $Y_i : i = 1, \cdots, n$ are generated as

$$Y_i = \theta_i + u_i$$

with u_i i.i.d. $\mathcal{N}(0, \sigma^2)$. With probability $\omega = 0.2, \theta_i$ is drawn from the density f, and with probability $1 - \omega$, we have $\theta_i = 0$. The composite null hypothesis is $\theta_i \in A \equiv [-1, 1]$.

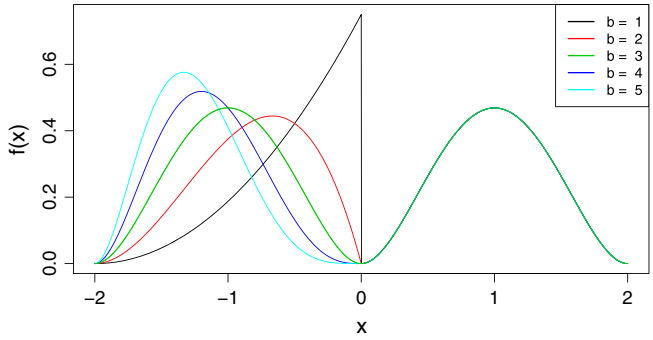

Figure 2. *Sun and McLain β model of the non-null density.*

Table 3. *Realised false discovery rate and false non-discovery rates: Sun-McLain Model 1 with $\omega = 0.2$ and 500 replications.*

	FDR					FNR				
	$b = 1$	$b = 2$	$b = 3$	$b = 4$	$b = 5$	$b = 1$	$b = 2$	$b = 3$	$b = 4$	$b = 5$
$n = 500$										
OR	0.0723	0.0695	0.0893	0.0909	0.0878	0.124	0.152	0.161	0.162	0.158
KW	0.1026	0.1091	0.1100	0.1139	0.1132	0.122	0.150	0.160	0.161	0.157
SM	0.1108	0.1059	0.1039	0.0950	0.0944	0.122	0.148	0.158	0.160	0.157
$n = 1000$										
OR	0.0922	0.0887	0.0937	0.0913	0.0962	0.123	0.151	0.161	0.161	0.158
KW	0.1020	0.1024	0.1069	0.1026	0.1064	0.122	0.151	0.161	0.162	0.159
SM	0.1318	0.1240	0.1158	0.1042	0.1048	0.120	0.147	0.157	0.158	0.156
$n = 5000$										
OR	0.0993	0.0989	0.0996	0.0974	0.1005	0.121	0.150	0.161	0.161	0.156
KW	0.0958	0.1002	0.0986	0.0962	0.0977	0.122	0.151	0.162	0.162	0.158
SM	0.1494	0.1410	0.1312	0.1216	0.1185	0.118	0.145	0.155	0.156	0.152
$n = 10000$										
OR	0.0972	0.1003	0.0990	0.0996	0.0997	0.122	0.150	0.160	0.160	0.157
KW	0.0949	0.0992	0.0963	0.0936	0.0951	0.122	0.151	0.162	0.162	0.158
SM	0.1501	0.1444	0.1319	0.1228	0.1177	0.118	0.145	0.155	0.156	0.153

We compare performance of an oracle (OR) who is aware of all of this with the empirical characteristic function approach of Sun and McLain (SM) and our Kiefer–Wolfowitz (KW) approach for four different sample sizes: 500, 1000, 5000, 10,000; 500 replications are done for each experimental setting.

Given the $T_i = \widehat{Lfdr}_i$'s, we need a threshold. This is performed precisely as mentioned earlier for all three methods, given their respective T_i's. In Table 3, we report results of the experiment for both FDR control and achieved FNR. It will be seen that the KW procedure has some what better FDR control than the SM procedure especially for the $b = 1$ setting that places quite a substantial amount of non-null mass in the interval $[-1, 1]$. FNR performance is quite comparable for all three methods.

4.3.2 Some Simulation Evidence: Sun and McLain Model 2

In our second simulation setting, again drawn from Sun & McLain (2012), Section 5.2, we have a similar non-null density,

$$f(\theta) = q\beta(\theta, 2, 2) + (1 - q)\tilde{\beta}(\theta, 2, 2),$$

where $\beta(\cdot, a, b)$ denotes a β density with parameters a and b, and $\tilde{\beta}(\cdot, a, b)$ denotes a reversed β density supported on $[0.5, 3.5]$ and $[-3.5, -0.5]$, respectively. The mixture proportion, $q = 0.5$ throughout, but we vary ω from 0.05 to 0.35. In Figure 3, we illustrate a single realization of this experiment, with $n = 5000$ and $\omega = 0.2$. The shaded rectangle is the null region, A, the solid (red) curves indicate the beta mixture non-null density known to the oracle, the dashed (green) curves depict the Sun and McLain empirical characteristic function estimate of the non-null density, and the dark spikes represent the mass points estimated by the NPMLE. The 'rug' plot below the x-axis shows the locations of the Y_i observations that correspond to

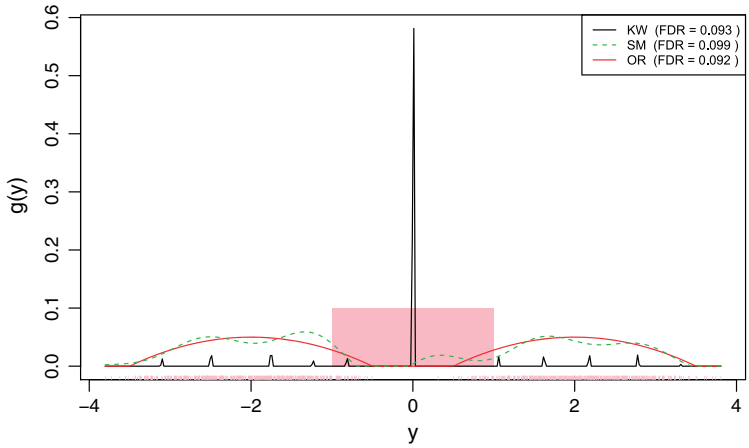

Figure 3. *One realization of Sun-McLain Model 2. The non-null density is composed of two symmetric, rescaled beta densities supported on* $\{(-3.5, -0.5) \cup (0.5, 3.5)\}$ *as shown by the solid (red) curve. The rejection region* A *appears as the shaded rectangle, the Sun and McLain estimate of the non-null density appears as the dashed (green) curve, and the nonparametric maximum likelihood estimator estimate of the mixing distribution, including the mass point near zero representing the null distribution, appears as the set of (black) spikes. The sample size for this realization is* $n = 5000$ *and* $\omega = 0.2$.

$\theta_i \notin A$, that is observations that should be rejected. Both the oracle and the NPMLE are a bit conservative in this example, but the discrete non-null distribution delivered by the Kiefer–Wolfowitz procedure does a remarkably good job of mimicking the smooth density generating the alternative.

In Figure 4, we report FDR and FNR results for three distinct sample sizes: 1000, 5000, and 10000; replications are again 500 for each instance. Again, we see that the KW procedure

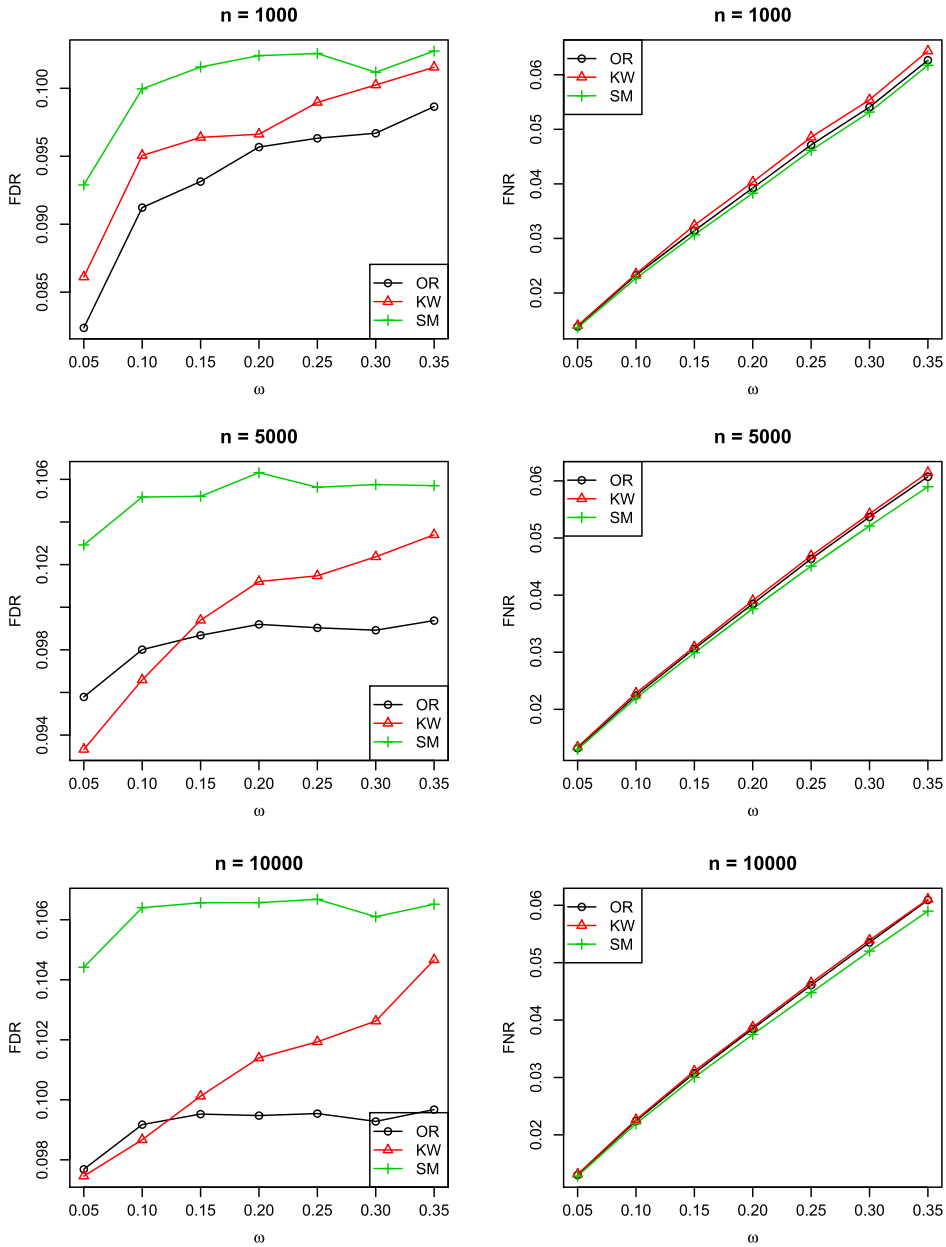

Figure 4. *False discovery rate and false non-discovery rate comparison for Sun-McLain Model 2 with symmetric non-null density and homogeneous scale.*

controls FDR somewhat better than SM, and the results for FNR are almost indistinguishable for the three procedures. Repeating this exercise for an asymmetric version of the non-null density with

$$f(\theta) = q\beta(\theta, 5, 2) + (1 - q)\tilde{\beta}(\theta, 5, 2),$$

and $q = 0.3$, yields very similar results depicted in Figure 5.

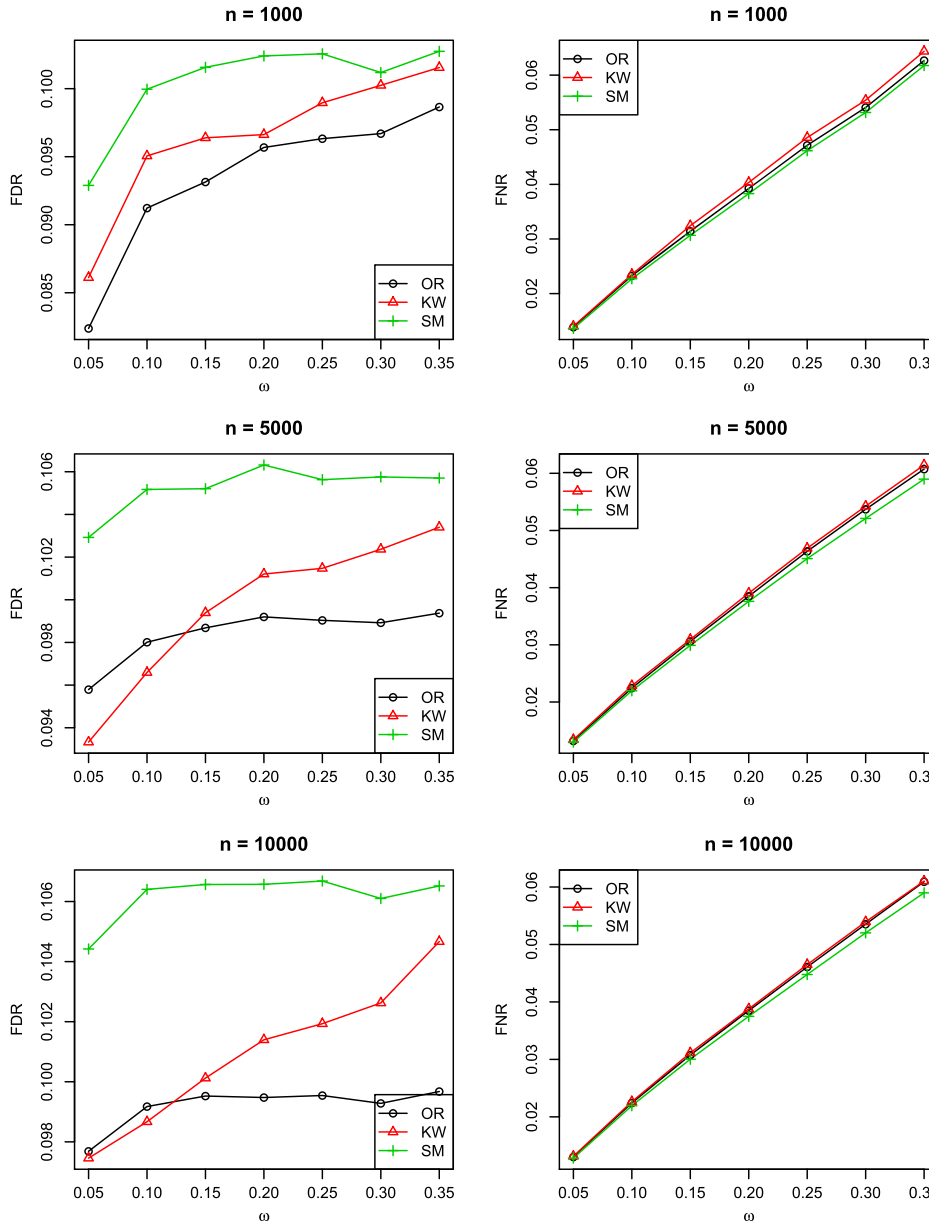

Figure 5. *False discovery rate and false non-discovery rate comparison for Sun-McLain Model 2 with asymmetric non-null density and homogeneous scale.*

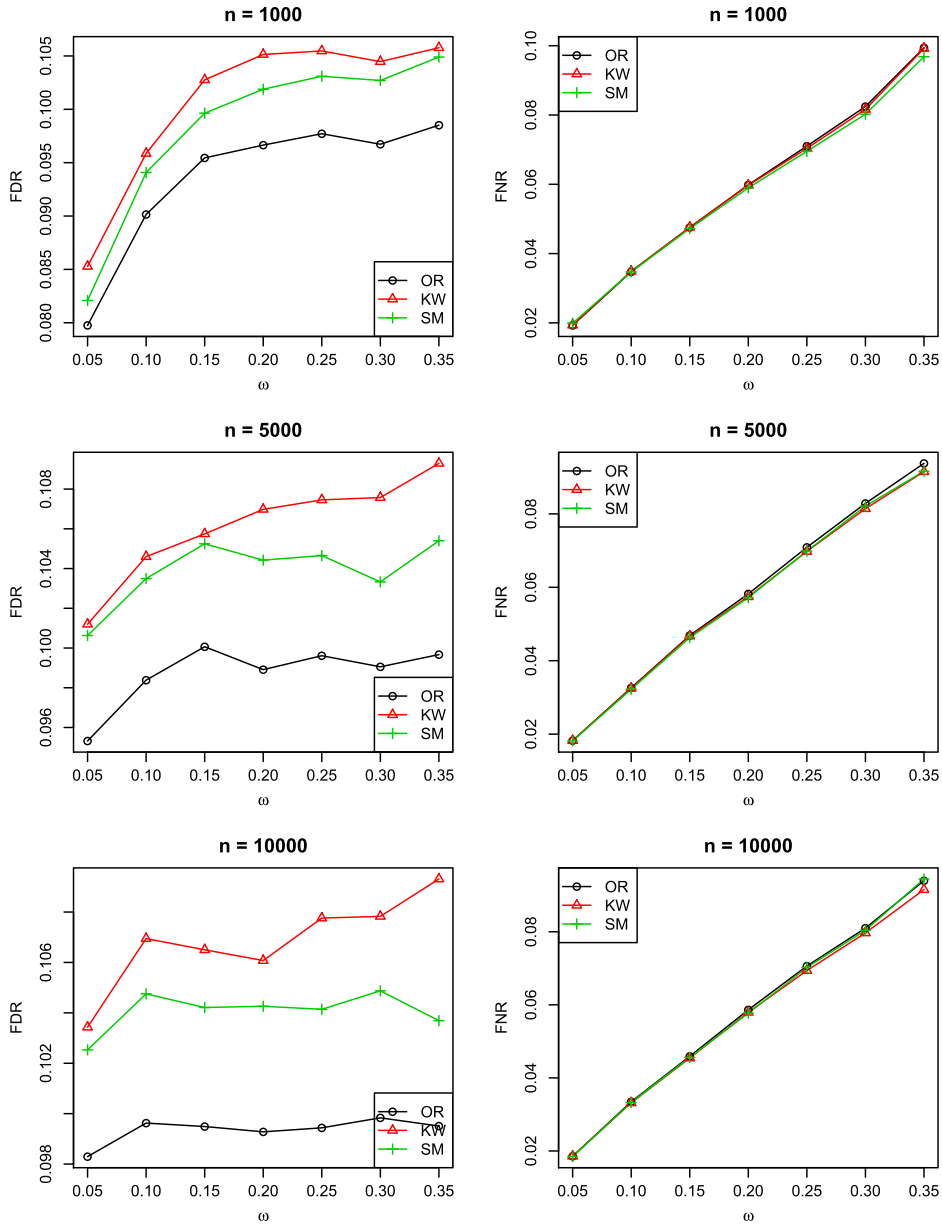

Figure 6. *False discovery rate and false non-discovery rate comparison for Sun-McLain Model 3 with symmetric non-null density and heterogeneous scale.*

4.3.3 Some Simulation Evidence: Sun and McLain Model 3

Finally, we conclude our simulation exercise with the Sun and McLain model of their Section 5.3 that involves heterogeneous scale parameters. The non-null densities are the same as in the previous subsection, but now, instead of a fixed $\sigma = 2/\sqrt{10}$, we have $\sigma \in \left\{1/\sqrt{10}, 2/\sqrt{10}, 3/\sqrt{10}\right\}$ with equal probability. Results are shown in Figures 6 and 7, for the symmetric and asymmetric cases, respectively. In this setting of the simulation binning of the observations is carried out separately for each distinct realization of the scale parameter.

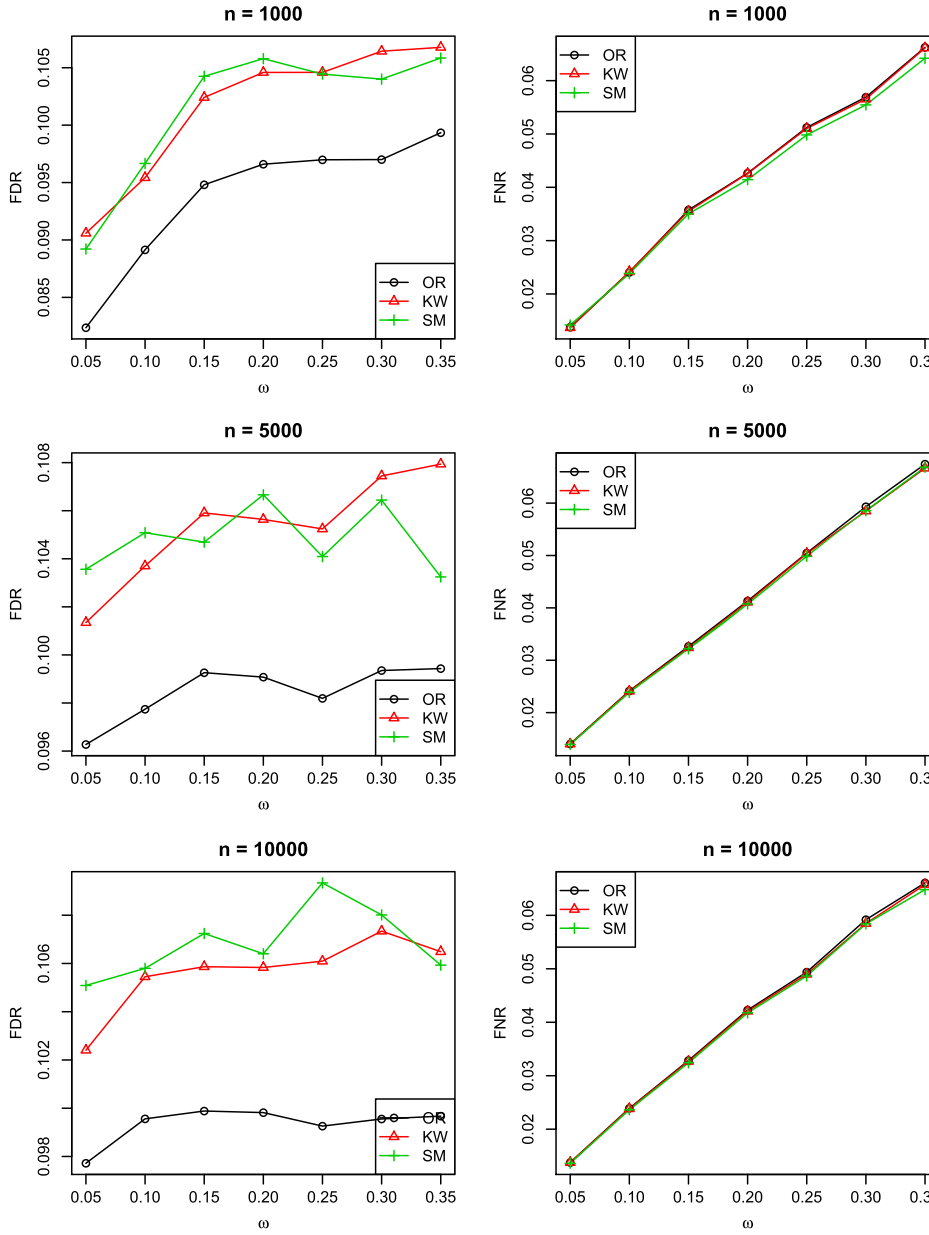

Figure 7. *False discovery rate and false non-discovery rate comparison for Sun-McLain Model 3 with asymmetric non-null density and heterogeneous scale.*

4.3.4 Discussion

The overall message of our FDR-FNR simulations is that even in composite null settings with non-null densities that spread alternative mass over a wide region, the Kiefer–Wolfowitz Lfdr procedure is highly effective. One might expect that the discrete mixing distributions delivered by the Kiefer–Wolfowitz MLE would have difficulties in such environments, but their FDR

performance is at least comparable with Sun and McLain's empirical characteristic function approach, and sometimes considerably better. False non-discovery rates are essentially similar for both methods and very close to what is achievable by the oracle.

5 Conclusion

Robbins (1951) presented a challenge to the emerging Wald minimax view of decision theory. Robbins showed that exploiting multiple instances of related problems could lead to dramatic improvements over minimax risk. This insight underlay much of Robbins subsequent empirical Bayes work and still offers tremendous potential for constructive future developments. Nonparametric maximum likelihood estimation of mixture models, as we have argued, can play an essential role in many facets of these developments as both data and computational resources improve. While we have focused mainly on Gaussian mixture models where deconvolution methods are also applicable, we would like to stress that the nonparametric MLE methods we have described, as foreseen by Robbins and introduced by Kiefer and Wolfowitz, are broadly applicable to the full range of mixture models with parametric base distribution, as illustrated by the binomial mixture of Section 3. This point is also stressed by Efron (2014). In repeated measurement, or longitudinal data settings there is some scope for nonparametric estimation of the base distribution and this is an interesting avenue for future research.

Acknowledgments

This research was partially supported by NSF grant SES-11-53548. We are grateful to a referee for comments that led to clarification of several points. Complete code for all the numerical results presented below is available from the second author.

References

Andersen, E.D. (2010). The MOSEK Optimization Tools Manual, Version 6.0. Available from http://www.mosek.com [Accessed on 2015].

Cai, T.T. & Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Am. Stat. Assoc.*, **104**, 1467–1481.

Cao, H., Sun, W. & Kosorok, M.R. (2013). The optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing. *Biometrika*, **100**, 495–502.

Dyson, F. (1926). A method for correcting series of parallax observations. *Mon. Not. Roy. Astron. Soc.*, **86**, 686–706.

Efron, B. (2008a). Microarrays, empirical bayes and the two-groups model. *Stat. Sci.*, **23**, 1–22.

Efron, B. (2008b). Simultaneous inference: When should hypothesis testing problems be combined? *The Ann. Appl. Stat.*, 197–223.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge: Cambridge U. Press.

Efron, B. (2011). Tweedie's formula and selection bias. *J. Am. Stat. Assoc.*, **106**, 1602–1614.

Efron, B. (2014). Bayesian Deconvolution. preprint.

Friberg, H.A. (2012). Rmosek: R-to-MOSEK Interface. Available from http://cran.r-project.org [Accessed on 2015].

Gu, J. & Koenker, R. (2014). Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective. preprint.

Heckman, J. & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, **52**, 63–132.

Jiang, W. & Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Stat.*, **37**, 1647–1684.

Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *J. Roy. Stat. Soc.: Series B (Stat. Method.)*, **70**(3), 461–493.

Johnstone, I. & Silverman, B. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Stat.*, **32**, 1594–1649.

Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Ann. Math. Stat.*, **27**, 887–906.
Koenker, R. (2012). REBayes: An R package for empirical Bayes methods. Available from http://cran.r-project.org [Accessed on 2015].
Koenker, R. (2014). A Gaussian compound decision bakeoff. *Stat.*, **3**, 12–16.
Koenker, R. & Gu, J. (2013). Frailty, profile likelihood and medfly mortality. In *Contemporary Developments in Statistical Theory: A Festschrift for Hira Lal Koul*, Eds. S. Lahiri, A. Schick, A. Sengupta & T. Sriram, pp. 227–237. New York: Springer.
Koenker, R. & Mizera, I. (2014). Convex optimization, shape constraints, compound decisions and empirical Bayes rules. *J. Am. Stat. Assoc.*, **109**, 674–685.
Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Am. Stat. Assoc.*, **73**, 805–811.
Lindsay, B. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, Hayward, California, IMS.
Muralidharan, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *The Ann. Appl. Stat.*, **4**, 422–438.
Robbins, H. (1950). A generalization of the method of maximum likelihood: Estimating a mixing distribution. *The Ann. Math. Stat.*, 314.
Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I. pp. 131–149. Berkeley: University of California Press.
Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I. Berkeley: University of California Press.
Storey, J.D. (2002). A direct approach to false discovery rates. *J. Roy. Stat. Soc.: B*, **64**, 479–498.
Sun, W. & Cai, T.T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Stat. Assoc.*, **102**, 901–912.
Sun, W. & McLain, A.C. (2012). Multiple testing of composite null hypotheses in heteroscedastic models. *J. Am. Stat. Assoc.*, **107**, 673–687.
Tukey, J. (1974). Named and faceless values: An initial exploration in memory of Prasanta C. Mahalanobis. *Sankhyā*, **36**, 125–176.
Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods. *Ann. Stat.*, 379–390.

[*Received August 2014, accepted February 2015*]