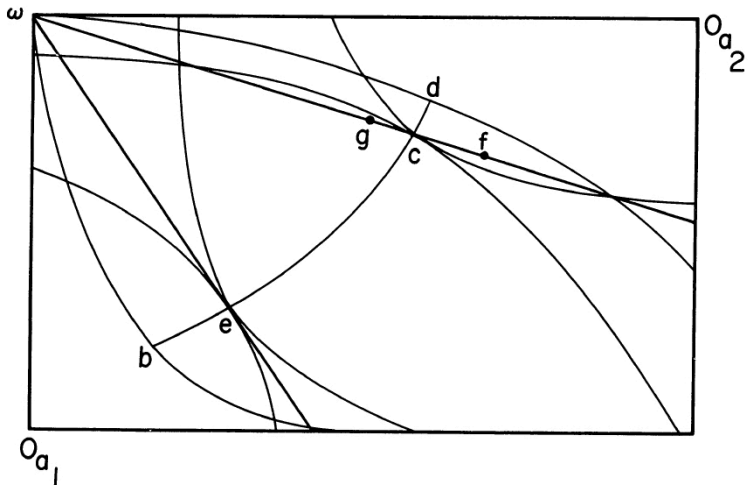
Invidious Comparisons: Ranking and Selection as Compound Decisions

Roger Koenker

University College London

Walras-Bowley Lecture: Milano* 19 August 2020



Bewley, T. (1973, E'a) Edgeworth's Conjecture

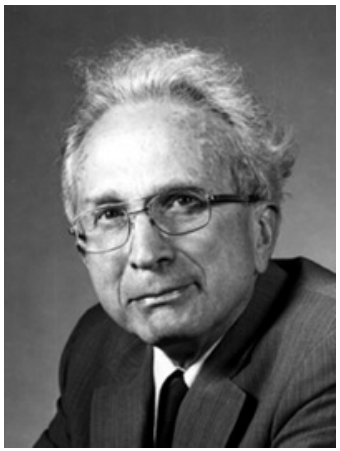
An Outline

1. Compound Decisions, the Ur Example, or
   Why minimax procedures can be suboptimal.
2. Empirical Bayes Deconvolution and Stein Shrinkage, or
   Borrowing strength, or why it takes a village.
3. Nonparametric Maximum Likelihood for Mixtures, or
   Convex optimization rules the waves.
4. Ranking and Selection as Compound Decisions, or
   How to do ranking and selection, if you must.

Credits

Jiaying Gu

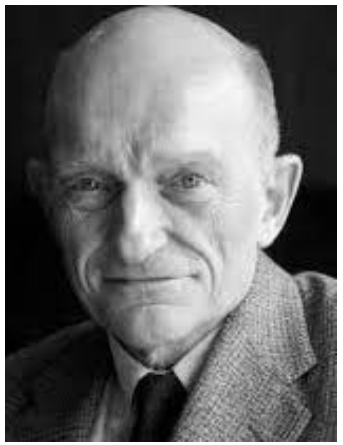
Jack Kiefer



Jack Wolfowitz

Herbert Robbins

Larry Brown



Gary Chamberlain

Robbins (1951) The Ur Compound Decision Problem

Suppose we observe, $y = (y_1, \cdots, y_n)$ from,

$$Y_i = \theta_i + u_i, \quad \theta_i \in \{-1, 1\}, \quad u_i \sim \mathcal{N}(0, 1)$$

and we would like to estimate $\theta \in \{-1, 1\}^n$ under loss,

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

Robbins shows that for $n = 1$ the minimax procedure is,

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

and he shows that this rule remains minimax for $n > 1$.

Let's be Bayesian

Lacking further information we may be willing to assume that the Y_i are exchangeable, and thus that the θ_i are iid Bernoulli (p). The minimax principle presumes that malevolent nature has chosen $p = 1/2$.

Let's be Bayesian

Lacking further information we may be willing to assume that the Y_i are exchangeable, and thus that the θ_i are iid Bernoulli (p). The minimax principle presumes that malevolent nature has chosen $p = 1/2$. Robbins observes that if we knew p,

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

we should guess $\hat{\theta}_i = 1$ if this probability exceeds 1/2, or equivalently, with a tiny bit of algebra, we obtain this elegant logistic shrinkage formula,

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

Let's be Bayesian

Lacking further information we may be willing to assume that the Y_i are exchangeable, and thus that the θ_i are iid Bernoulli (p). The minimax principle presumes that malevolent nature has chosen $p = 1/2$. Robbins observes that if we knew p,

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

we should guess $\hat{\theta}_i = 1$ if this probability exceeds 1/2, or equivalently, with a tiny bit of algebra, we obtain this elegant logistic shrinkage formula,

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

But we don't know the "prior" p. Let's try to estimate it.

Hierarchical Bayes Methods

We have the log likelihood,

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

a symmetric beta prior is convenient, like a flips of a fair coin,

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

Hierarchical Bayes Methods

We have the log likelihood,

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

a symmetric beta prior is convenient, like a flips of a fair coin,

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

The posterior for θ_i is,

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

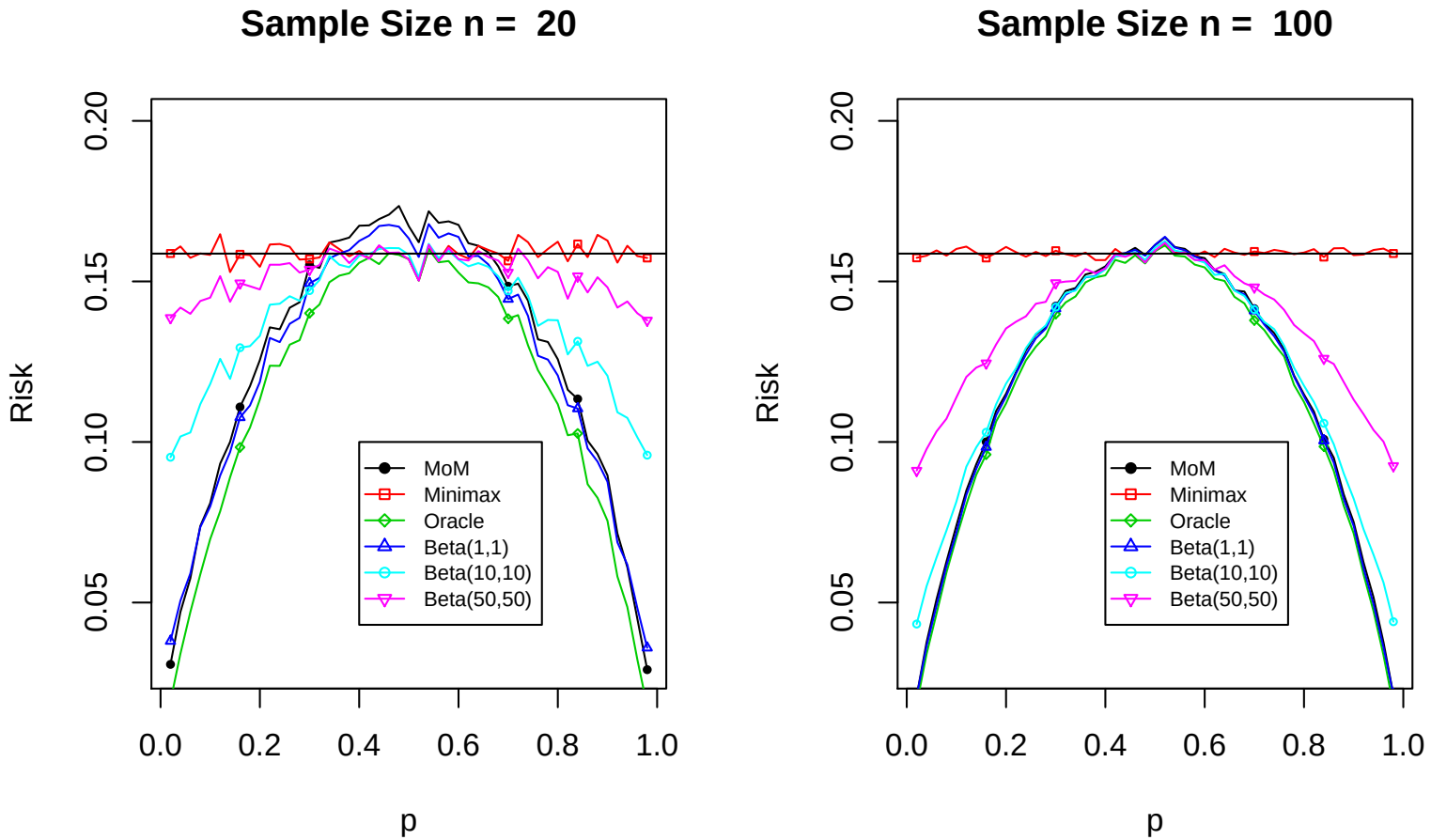
where $\bar{p}$ is the posterior mean of p given all the data y.

$$\bar{p} = \frac{\int p \prod (p\varphi(y_j - 1) + (1-p)\varphi(y_j + 1))\pi(p)dp}{\int \prod (p\varphi(y_j - 1) + (1-p)\varphi(y_j + 1))\pi(p)dp}.$$

and we have a plug-in Bayes rule,

$$\delta_{\bar{p}}(y_i) = \text{sgn}(y_i - \tfrac{1}{2} \log((1-\bar{p})/\bar{p})).$$

Empirical Risk for Minimax vs. Empirical Bayes Rules



Mean absolute loss over 1000 replications.

Free the θ's: The Gaussian Sequence Model

Restricting the θ_i's to live in $\{-1, 1\}$ seems a bit cruel, why not let them roam free? Suppose that, in the simplest case, we have

$$Y_i = \theta_i + u_i, \quad \theta_i \sim G, \quad u_i \sim \mathcal{N}(0, 1)$$

so marginally $Y_i \sim f(y) = \int \varphi(y - \theta) dG(\theta)$. Under squared error loss Robbins (1956) shows that the posterior mean, the optimal Bayes rule estimator, of the θ's is given by,

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

Efron (2011) calls this Tweedie's formula; it provides a general shrinkage strategy for Gaussian noise models, encompassing various parametric Stein rule procedures. When G is known we're good to go, otherwise we need to estimate our prior, G.

Tweedie's formula becomes a Stein Rule for Gaussian G

When the mixing distribution, G, is itself Gaussian, then the mixture density f is also Gaussian, and the Tweedie log derivative term simplifies to a linear (or perhaps affine) function and the Tweedie formula becomes linear shrinkage à la Stein. If $\theta_i \sim G = \mathcal{N}(\mu_0, \sigma_0^2)$ we shrink toward the prior mean,

$$\delta(y) = \mu_0 + \left(1 - \frac{1}{1+\sigma_0^2}\right)(y - \mu_0),$$

Tweedie's formula becomes a Stein Rule for Gaussian G

When the mixing distribution, G, is itself Gaussian, then the mixture density f is also Gaussian, and the Tweedie log derivative term simplifies to a linear (or perhaps affine) function and the Tweedie formula becomes linear shrinkage à la Stein. If $\theta_i \sim G = \mathcal{N}(\mu_0, \sigma_0^2)$ we shrink toward the prior mean,

$$\delta(y) = \mu_0 + \left(1 - \frac{1}{1+\sigma_0^2}\right)(y - \mu_0),$$

Estimating the prior mean parameter costs us one degree of freedom, and we obtain the celebrated James-Stein (1960) estimator,

$$\hat{\delta}(y) = \bar{Y}_n + \left(1 - \frac{n-3}{S}\right)(y - \bar{Y}_n),$$

with $\bar{Y}_n = n^{-1}\sum Y_i$ and $S = \sum (Y_i - \bar{Y}_n)^2$.

Tweedie's formula becomes a Stein Rule for Gaussian G

When the mixing distribution, G, is itself Gaussian, then the mixture density f is also Gaussian, and the Tweedie log derivative term simplifies to a linear (or perhaps affine) function and the Tweedie formula becomes linear shrinkage à la Stein. If $\theta_i \sim G = \mathcal{N}(\mu_0, \sigma_0^2)$ we shrink toward the prior mean,

$$\delta(y) = \mu_0 + \left(1 - \frac{1}{1+\sigma_0^2}\right)(y - \mu_0),$$

Estimating the prior mean parameter costs us one degree of freedom, and we obtain the celebrated James-Stein (1960) estimator,

$$\hat{\delta}(y) = \bar{Y}_n + \left(1 - \frac{n-3}{S}\right)(y - \bar{Y}_n),$$

with $\bar{Y}_n = n^{-1}\sum Y_i$ and $S = \sum (Y_i - \bar{Y}_n)^2$.

This is the original Bayesian sin: we have estimated the prior!

Fisherian Deconvolution

Having observed a random sample on $Y \sim f(y) = \int \varphi(y - \theta) dG(\theta)$, we would like to recover an estimate of the mixing distribution G. This is generally viewed as deconvolution and attacked with Fourier methods. Robbins thought differently, and asked, "why not maximum likelihood?" This yields a nice convex optimization problem:

$$\min_{G \in \mathcal{G}} \{ - \sum_{i=1}^{n} \log f(y_i) \mid f(y_i) = \int \varphi(y_i - \theta) dG(\theta) \}$$

- Works for any mixture problem, not just classical deconvolution
- Given a G we can compute posterior means, posterior quantiles, posterior whatevers for within-sample or out-of-sample observations.
- Location shift equivariant, unlike typical Bayesian shrinkage, e.g. spike and slab or horseshoe priors.

Minimalist G-Modeling and Alternatives

When φ is Gaussian we have a classical deconvolution problem, but Fourier methods perform poorly, while maximum likelihood in several forms performs quite brilliantly.

- Efron's logspline approach expresses $g = G'$ as a natural spline:

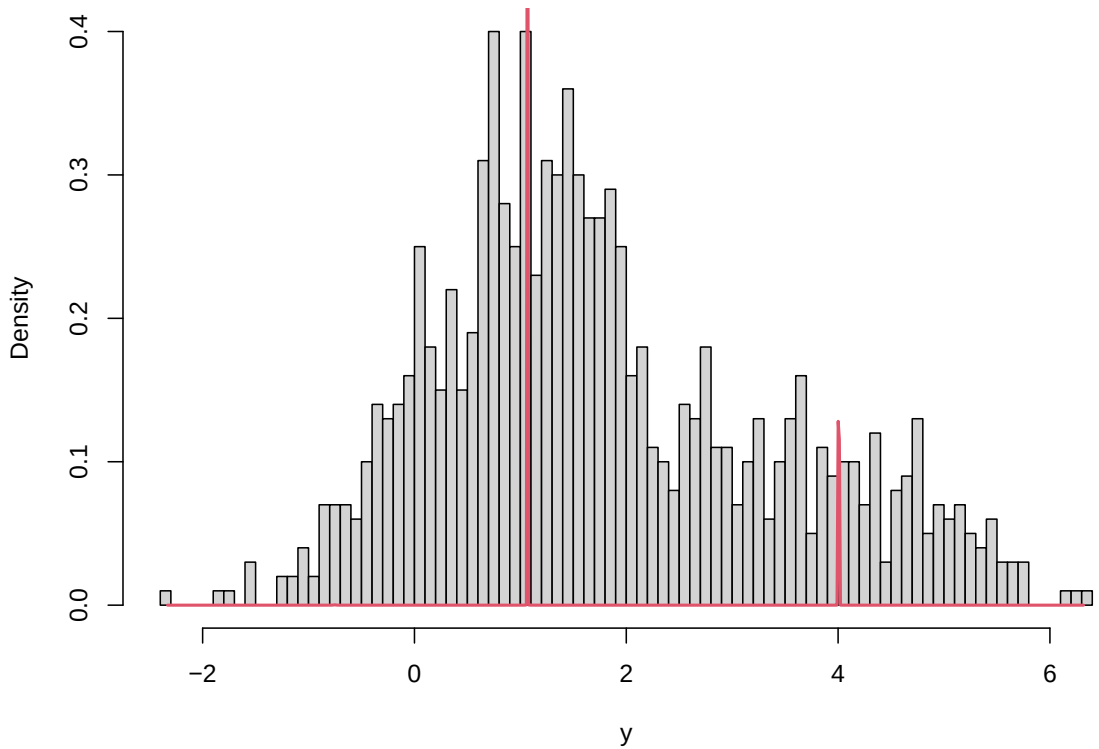
$$\log g(\theta) = \sum_{j=1}^{p} \alpha_j \psi_j(\theta),$$

  and estimates the parameters $\alpha \in \mathbb{R}^p$ by penalized maximum likelihood.
- The Kiefer and Wolfowitz NPMLE yields a discrete G typically with only a few atoms, and has the advantage that it is tuning parameter free.

Both approaches share the advantage that they are applicable to the general class of mixture problems, not only to Gaussian deconvolution.

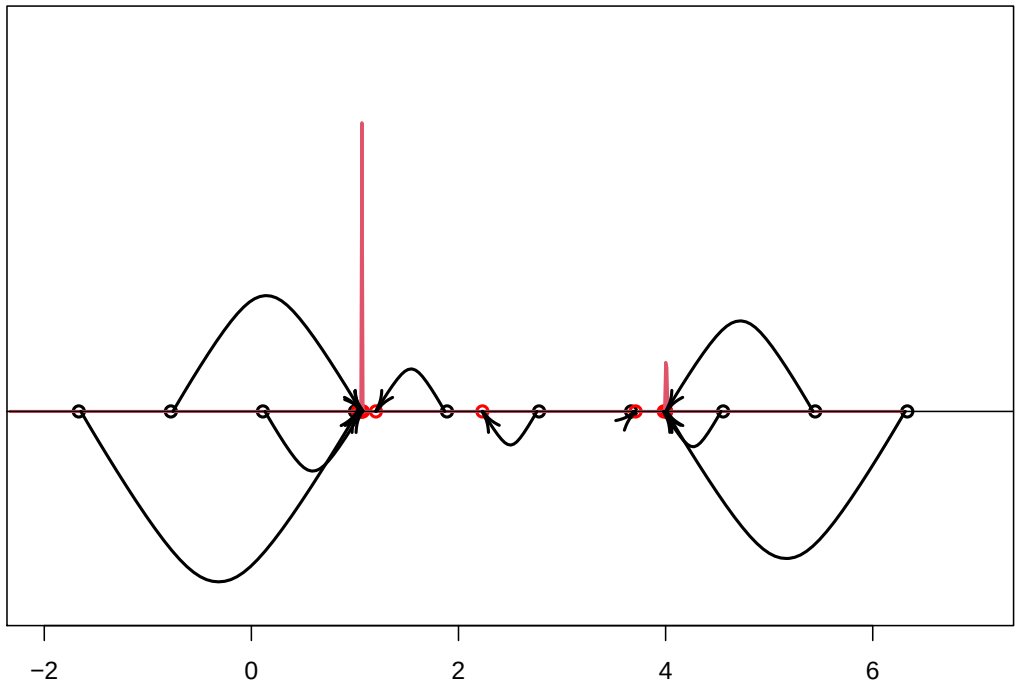
A Simple Discrete Mixture Example

Consider the simple model, $Y_i \sim \mathcal{N}(\theta_i, 1)$, with $\theta \in \{1, 4\}$ with probabilities $(0.75, 0.25)$ respectively. We draw a sample of $n = 1000$, Y's, plot their histogram, and then overplot the Kiefer-Wolfowitz NPMLE in red.



Tweedie Shrinkage for Posterior Means

Given our $\hat{\mathrm{G}}$ we can compute a posterior mean estimate for any value of y. What does this look like?



Tweedie shrinkage is quite smart about adapting shrinkage to the form of the G. No longer are we simply shrinking all observations toward the same fixed value.

Needles and Haystacks

It is commonly assumed that G contains a large mass point concentrated at zero, the haystack, and a smaller mass well separated from zero, i.e. the needles. Castillo and van der Vaart (2012) compare several Bayes and empirical Bayes procedures in this setting.

	s = 25			**s = 50**			**s = 100**		
	3	4	5	3	4	5	3	4	5
PM1	111	96	94	176	165	154	267	302	307
PM2	106	92	82	169	165	152	269	280	274
EBM	103	96	93	166	177	174	271	312	319
PMed1	129	83	73	205	149	130	255	279	283
PMed2	125	86	68	187	148	129	273	254	245
EBMed	110	81	72	162	148	142	255	294	300
HT	175	142	70	339	284	135	676	564	252
HTO	136	92	84	206	159	139	306	261	245
NPMLE	80	57	30	122	81	40	174	112	53

Mean squared error of several estimators considered by Castillo and van der Vaart and the NPMLE procedure of Robbins. Sample size $n = 500$ throughout, with s non-null observations concentrated at $\theta \in \{3, 4, 5\}$. Based on 100 replications for the first eight Castillo and van der Vaart procedures, and 1000 replications for the NPMLE.

Ranking and Selection

Ranking and selection are inextricably linked in applications; we rank because we want to select "the best" or "the worst."
We see noisy measurements, y_k of some latent quality, θ_k and would like to select the α-best, $\{k : \theta_k > \theta_\alpha \equiv G^{-1}(1-\alpha)\}$.
The optimal procedure is to compute posterior tail probabilities for each unit,

$$\nu_\alpha(y_k) = \mathbb{P}(\theta_k \geqslant \theta_\alpha | Y = y_k),$$

and then rank the the $\nu_\alpha(y_k)$ and select according to the rule,

$$\delta(y_k) = \mathbb{1}\{\nu_\alpha(y_k) \geqslant \lambda_\alpha\},$$

where λ_α satisfies, $\mathbb{P}\{\nu_\alpha(Y_k) \geqslant \lambda_\alpha\} = \alpha$,
This is compound decision rule since all the observations contribute to the determination of the $\nu_\alpha(y_k)$ through the estimation of G. It can be quite different than ranking on the basis of posterior means which is typically employed.

Isn't Selection a lot like Multiple Testing?

Indeed! In fact, $1 - \nu_\alpha(y)$ is the local false discovery rate for testing,

$$H_o : \theta_k < \theta_\alpha \quad \text{vs.} \quad H_a : \theta_k \geqslant \theta_\alpha$$

as in Efron, Tibshirani, Storey, and Tusher (2001). In our simplest Gaussian sequence setting,

$$\nu_\alpha(y) = \alpha f_1(y)/f(y) = \mathbb{P}\{\theta_k \geqslant \theta_\alpha | Y = y\},$$

where,

$$f_1(y) = \alpha^{-1} \int_{\theta_\alpha}^{\infty} \varphi(y - \theta) dG(\theta)$$

and

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

Note that this depends on knowing the mixing distribution, G, or being able to estimate it. Given a $\hat{G}$ we can easily compute ν_α's and do selection.

Selection with $\delta(y)$ is a Bayes Rule

Two preliminary lemmas are relatively easy:

Lemma 1

Let $h_k = \mathbb{1}(\theta_k \geqslant \theta_\alpha)$, *then* $\delta_k = \delta(y_k)$ *minimizes the Bayes (compound) risk of misclassification,*

$$R(\delta) = \mathbb{E}\Big[\sum_k L(\delta_k, h_k) = \lambda \mathbb{1}\{h_k = 0, \delta_k = 1\} + \mathbb{1}\{h_k = 1, \delta_k = 0\}\Big]$$

Selection with $\delta(y)$ is a Bayes Rule

Two preliminary lemmas are relatively easy:

Lemma 1

Let $h_k = \mathbb{1}(\theta_k \geqslant \theta_\alpha)$, *then* $\delta_k = \delta(y_k)$ *minimizes the Bayes (compound) risk of misclassification,*

$$R(\delta) = \mathbb{E}\big[\sum_k L(\delta_k, h_k) = \lambda \mathbb{1}\{h_k = 0, \delta_k = 1\} + \mathbb{1}\{h_k = 1, \delta_k = 0\}\big]$$

Lemma 2

When the variances of the y_k *are homogeneous, ranking by posterior means or posterior tail expectations (shortfall) are identical to the posterior tail probability ranking, so their selection rules are also equivalent.*

Guarding Against False Discoveries

Rather than simply trying to control the size of the selected set, we may also wish to constrain the false discovery rate, that is the proportion of selected observations that fail to meet our standard,

$$\mathrm{FDR} = \mathbb{P}(\theta < \theta_\alpha \mid \delta_\alpha(Y) = 1)$$

To control both the proportion selected and false discovery rate we propose the loss,

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

where $h_i = \mathbb{1}\{\theta_i \geqslant \theta_\alpha\}$, and the Lagrange multipliers τ_1 and τ_2 are chosen to control FDR at γ and capacity at α. When the problem is difficult, so FDR is high this formulation tends to reduce the proportion of selected units below the initially specified capacity level α.

Guarding Against False Discoveries

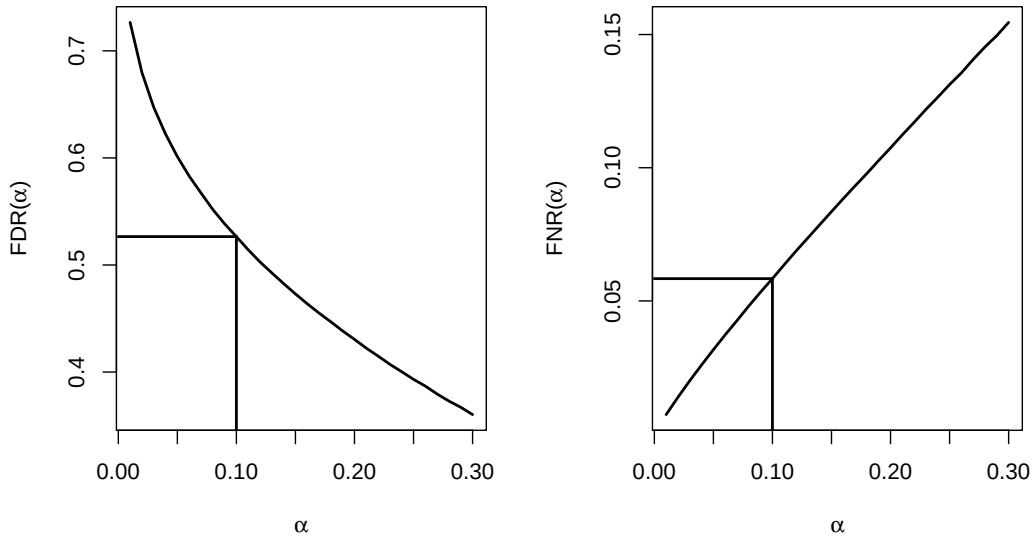
To gauge the difficulty of a particular selection problem it is important to know:

- The proportion of targeted "true" discoveries that are missed,
- The proportion selected units that are "false" discoveries

We will call these proportions FNR and FDR respectively. We may want to control not only the proportion selected, α, but the proportion of false discoveries.
A crucial advantage of estimating the mixing distribution, G, is that it enables us to estimate FDR and thereby compute thresholds required to construct Bayes selection rules for our augmented loss function.

A Normal Example that Gives the Oracle a Headache

Consider the simplest Gaussian model $Y_i \sim \mathcal{N}(\theta_i, 1)$ and $\theta_i \sim \mathcal{N}(0, 1)$, we would like to select the best α proportion. An oracle possessing full knowledge of this setting, knowing that G is standard normal and therefore the marginal distribution of the Y_k's is $\mathcal{N}(0, 2)$, incurs the errors appearing below.



So, for example, when $\alpha = 0.10$, more than half of the Oracle's selections are "false" and about six percent of those not selected should have been selected.

Heterogeneous Precision of the Y_k

Often it is important to account for observed differences in the precision of the measurements, Y_k, the number of "at bats" in the baseball batting average applications that pervade the early empirical Bayes literature, or the number of student test takers in the teacher value added applications.

$$Y_k \sim \mathcal{N}(\theta_k, \sigma_k^2), \quad \text{and} \quad \theta_k \sim G, \quad \sigma_k^2 \sim H, \quad \sigma_k \perp\!\!\!\perp \theta_k$$

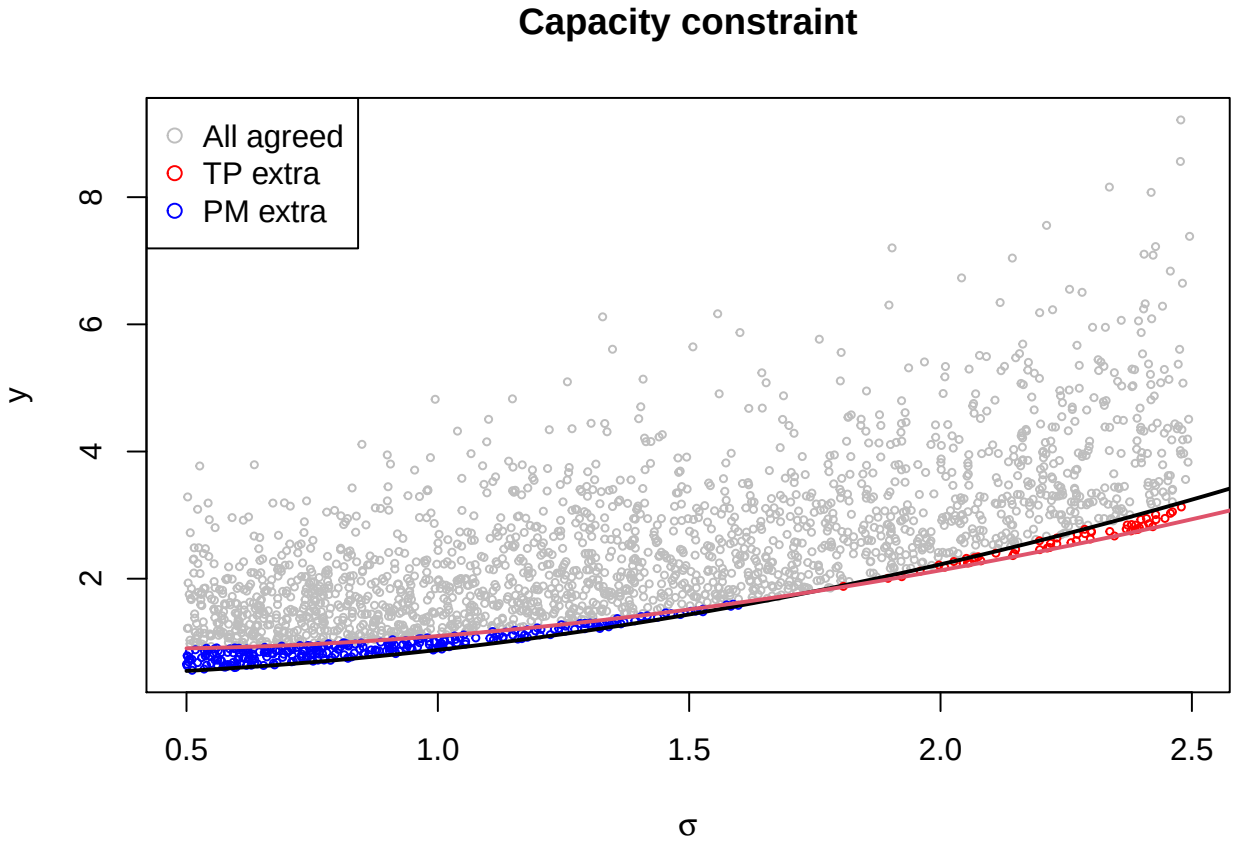
The posterior tail probability criterion becomes,

$$\nu_\alpha(y_k, \sigma_k) = \mathbb{P}(\theta_k \geqslant \theta_\alpha | y_k, \sigma_k) = \frac{\int_{\theta_\alpha}^{+\infty} \varphi_{\sigma_k}(y_k - \theta) dG(\theta)}{\int_{-\infty}^{+\infty} \varphi_{\sigma_k}(y_k - \theta) dG(\theta)}$$

where $\varphi_\sigma(u) = \varphi(u/\sigma)/\sigma$.
Now we have a "selection boundary" a curve in the (y, σ) plane that separates regions of selected from unselected observations.

Selection Boundaries and Conflicts Between Rules

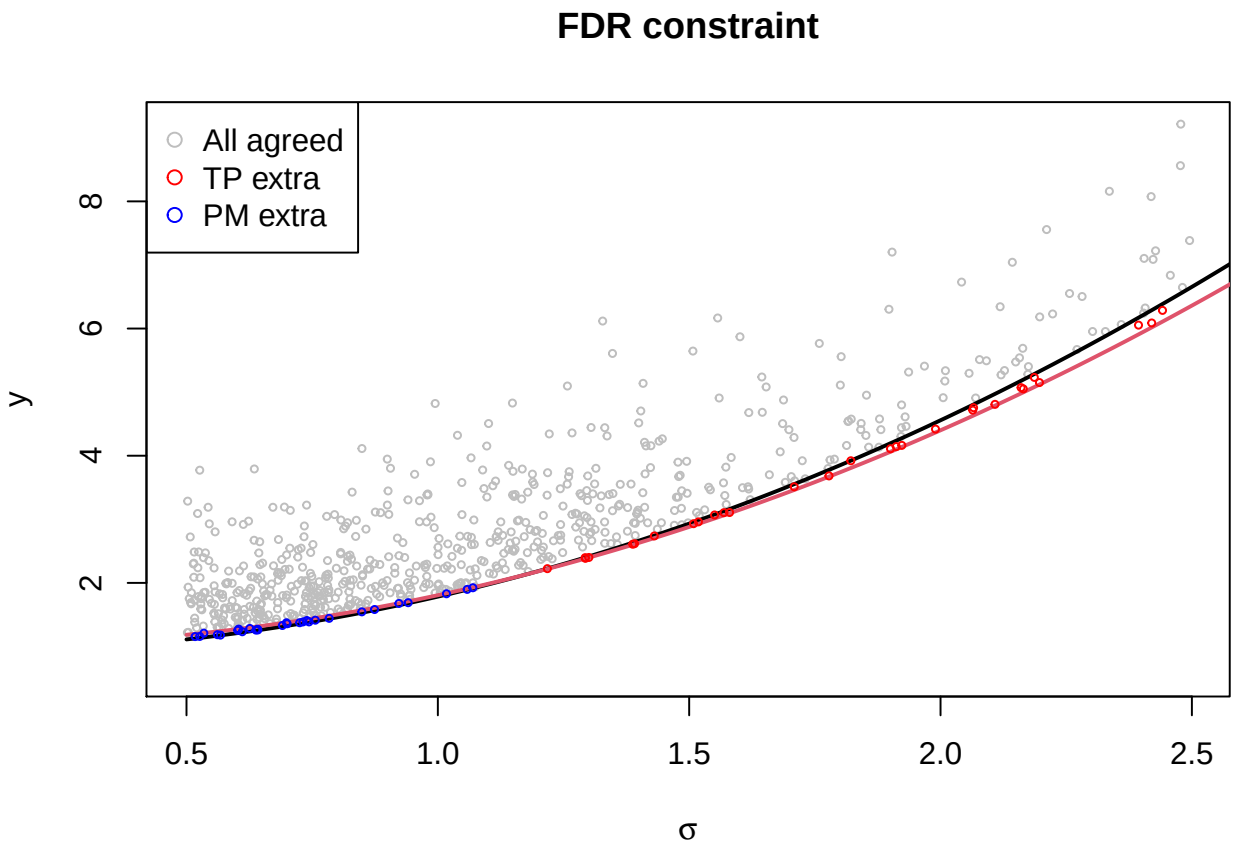
When only capacity is constrained, the tail probability rule is willing to select some high variance units that the posterior mean rule would reject.



A simulation of 10,000 observations with G standard Gaussian and $\sigma_i \sim U[0.5, 2.5]$. Grey points indicate units selected by both posterior tail probability and posterior mean rules, blue points are those selected by PM but not TP, and red are those selected by TP but not PM.

Selection Boundaries and Conflicts Between Rules

When both capacity and FDR are constrained, many fewer units are selected, but still the posterior tail probability rule selects a few more high variance units.



Even a very lax FDR constraint with $\gamma = 0.40$ dramatically reduces the number of selected units.

Nested Selection

Generally, we would expect that relaxing the capacity constraint, letting α to be larger for any fixed FDR control γ, would enlarge the selection regions,

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

Lemma 3

Let $f_v(v; \alpha)$ *denote the density function of* $v_\alpha(y_k, \sigma_k)$*, if* $\nabla_\alpha \log f_v(v; \alpha)$ *is non-decreasing in* v*, then the selection regions are nested, that is, for fixed* γ*,* $\alpha_1 \leqslant \alpha_2$ *implies* $\Omega_{\alpha_1,\gamma} \subseteq \Omega_{\alpha_2,\gamma}$

The monotonicity condition can be interpreted as a variant of the classical monotone likelihood ratio condition.

A Taste of Simulation

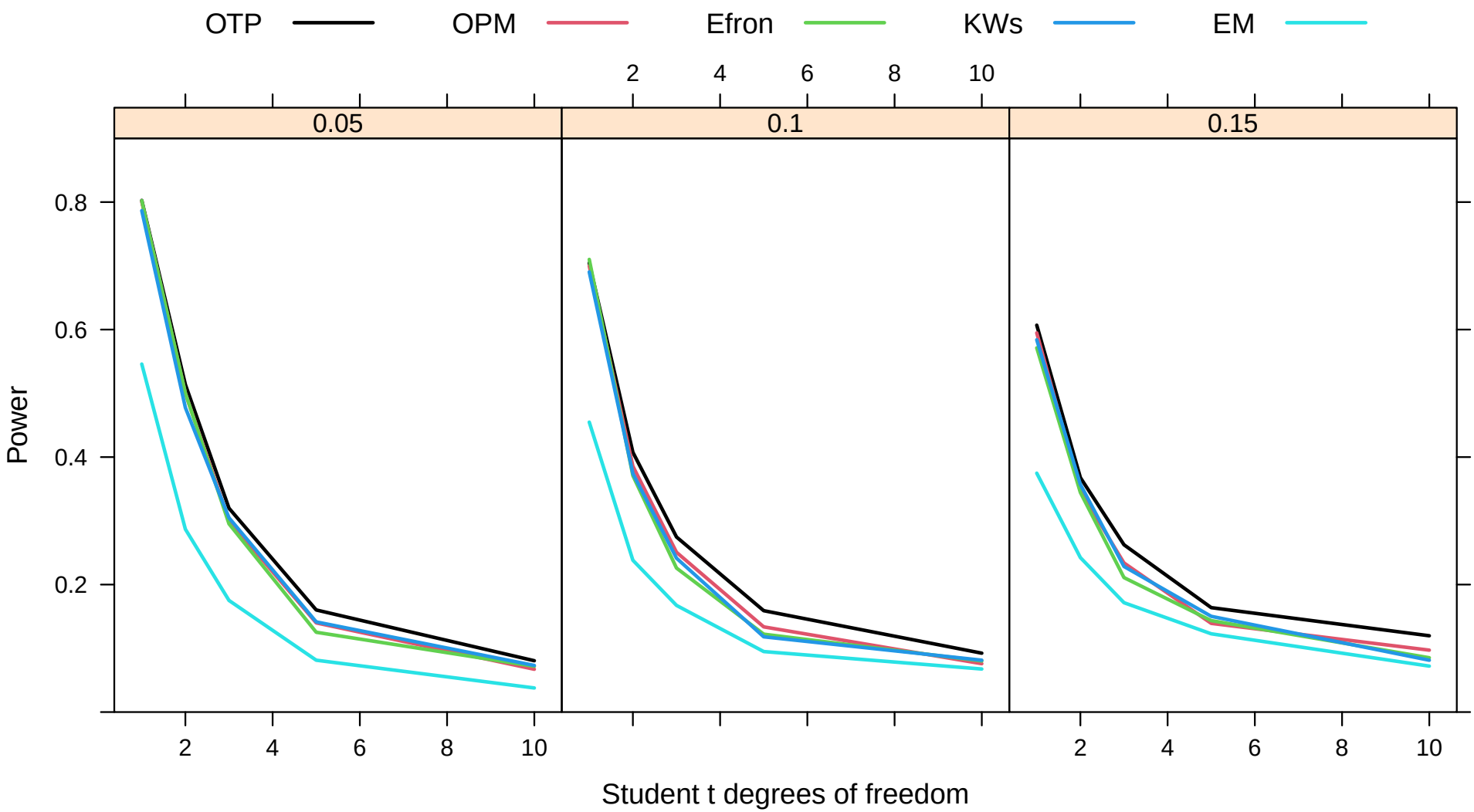
Not suprisingly tail behavior of the true mixing distribution G has an important impact on the performance of ranking and selection methods. To explore this we consider several Student t choices for G, with independent and obsesrvable $\sigma^2 \sim U[0.5, 1.5]$. Five selection procedures are compared:

OTP	Oracle posterior tail probability ranking
OPM	Oracle posterior mean ranking
Efron	Efron posterior tail probability ranking
KWs	Smoothed Kiefer-Wolfowitz posterior tail probability ranking
EM	Efron & Morris (1973) linear shrinkage posterior mean ranking

Performance is measured by Power, the proportion of the true top α of the θ_k's that are selected, an estimate of:

$$\beta(\delta) = \mathbb{P}(\theta_k \geqslant \theta_\alpha, \delta_k = 1)/\mathbb{P}(\theta_k \geqslant \theta_\alpha)$$

The Closer to Normality the Harder Selection Becomes



Ranking and Selection of U.S. Dialysis Centers

There is a well established data collection and analysis system for ranking and rating the performance of U.S. kidney dialysis centers. We use panel data on 3230 of these centers from 2004-2017 to illustrate our methods.

Ranking and Selection of U.S. Dialysis Centers

There is a well established data collection and analysis system for ranking and rating the performance of U.S. kidney dialysis centers. We use panel data on 3230 of these centers from 2004-2017 to illustrate our methods.

Centers have different mixes of patients, so the primary measure of center performance, patient mortality, is adjusted for "expected mortality" as estimated by a Cox proportional hazard model that captures variation in the patient mix. Observed deaths, denoted y_{it} for center i in year t are modeled as Poisson,

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

where μ_{it} is center i's expected deaths as predicted by the Cox model in year t and λ_i is the center's unobserved mortality rate, assumed constant over 3 to 5 year time horizons.

Back to Normality

The classical variance stabilizing transformation for the Poisson takes us back to the Gaussian model,

$$z_{it} = \sqrt{y_{it}/\mu_{it}} \sim \mathcal{N}(\theta_i, \sigma_i^2/w_{it}),$$

where $\theta_i = \sqrt{\lambda_i}$ and $w_{it} = 4\mu_{it}$. Exchangeability of the centers yields a mixture model in which the parameters (θ_i, σ_i^2), are assumed to be drawn iidly from the joint distribution, G. The σ_i^2 account for overdispersion of the Poisson. We have sufficient statistics:

$$T_i = \sum_{t=1}^{m} w_{it} z_{it}/W_i \sim \mathcal{N}(\theta_i, \sigma_i^2/W_i),$$

and

$$S_i = (m-1)^{-1} \sum_{t=1}^{m} (z_{it} - T_i)^2/w_{it} \sim \Gamma(r, \sigma_i^2/r),$$

where $W_i = \sum_t w_{it}$ and $r = (m-1)/2$, which give us an explicit likelihood for G.

Ranking Dialysis Centers

Given an estimate, $\hat{G}$, we can use posterior tail probability to rank:

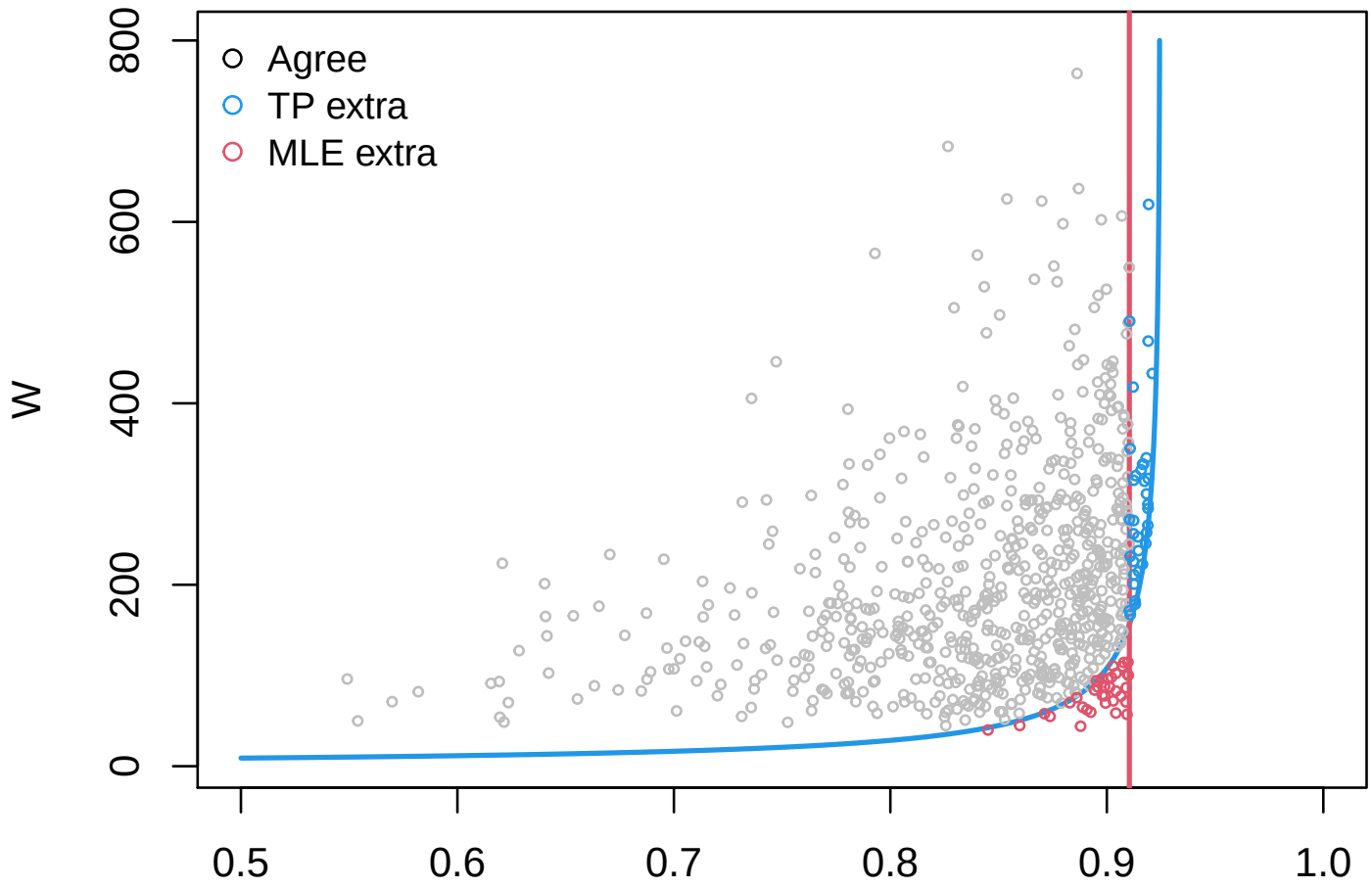
$$v_i = \mathbb{P}(\theta_i \geqslant \theta_\alpha | t_i, s_i, w_i) \approx \frac{\int_{\theta_\alpha}^{+\infty} \varphi(t_i|\theta_i, \sigma_i^2/w_i)\Gamma(s_i|r, \sigma_i^2/r) d\hat{G}(\theta, \sigma^2)}{\int_{-\infty}^{+\infty} \varphi(t_i|\theta_i, \sigma_i^2/w_i)\Gamma(s_i|r, \sigma_i^2/r) d\hat{G}(\theta, \sigma^2)}.$$

and thresholds may be computed to control capacity and the false discovery rate. To simplify the exposition, we will assume $\sigma_i \equiv 1$, so there is no over-dispersion, then,

$$v_i = \mathbb{P}(\theta_i \geqslant \theta_\alpha | t_i, w_i) \approx \frac{\int_{\theta_\alpha}^{+\infty} \varphi(t_i|\theta_i, 1/w_i) d\hat{G}(\theta)}{\int_{-\infty}^{+\infty} \varphi(t_i|\theta_i, 1/w_i) d\hat{G}(\theta)}.$$

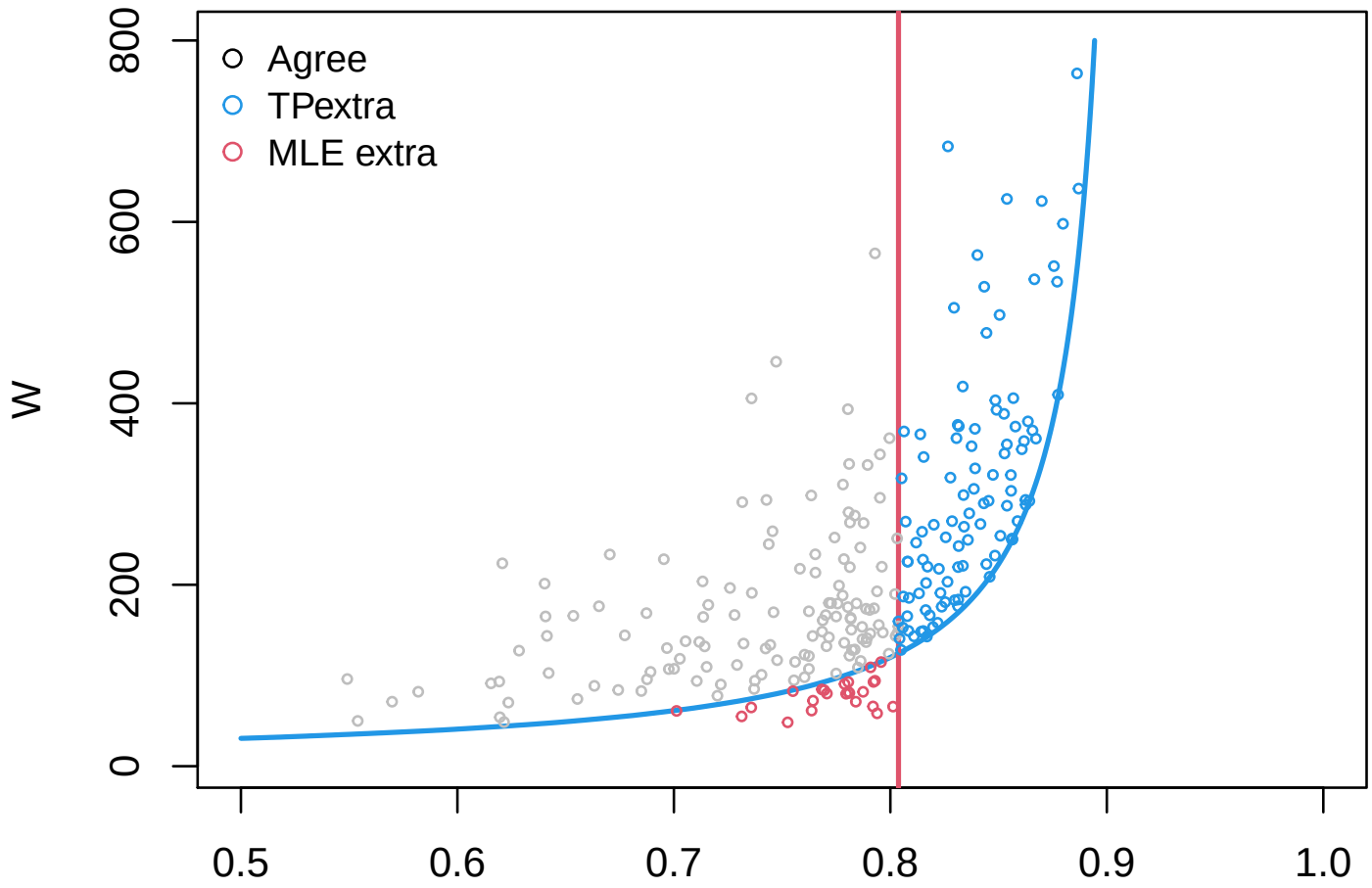
But bivariate heterogeneity is entirely feasible, as in our previous work on PSID income dynamics.

Comparison of TP Selection with MLE Selection



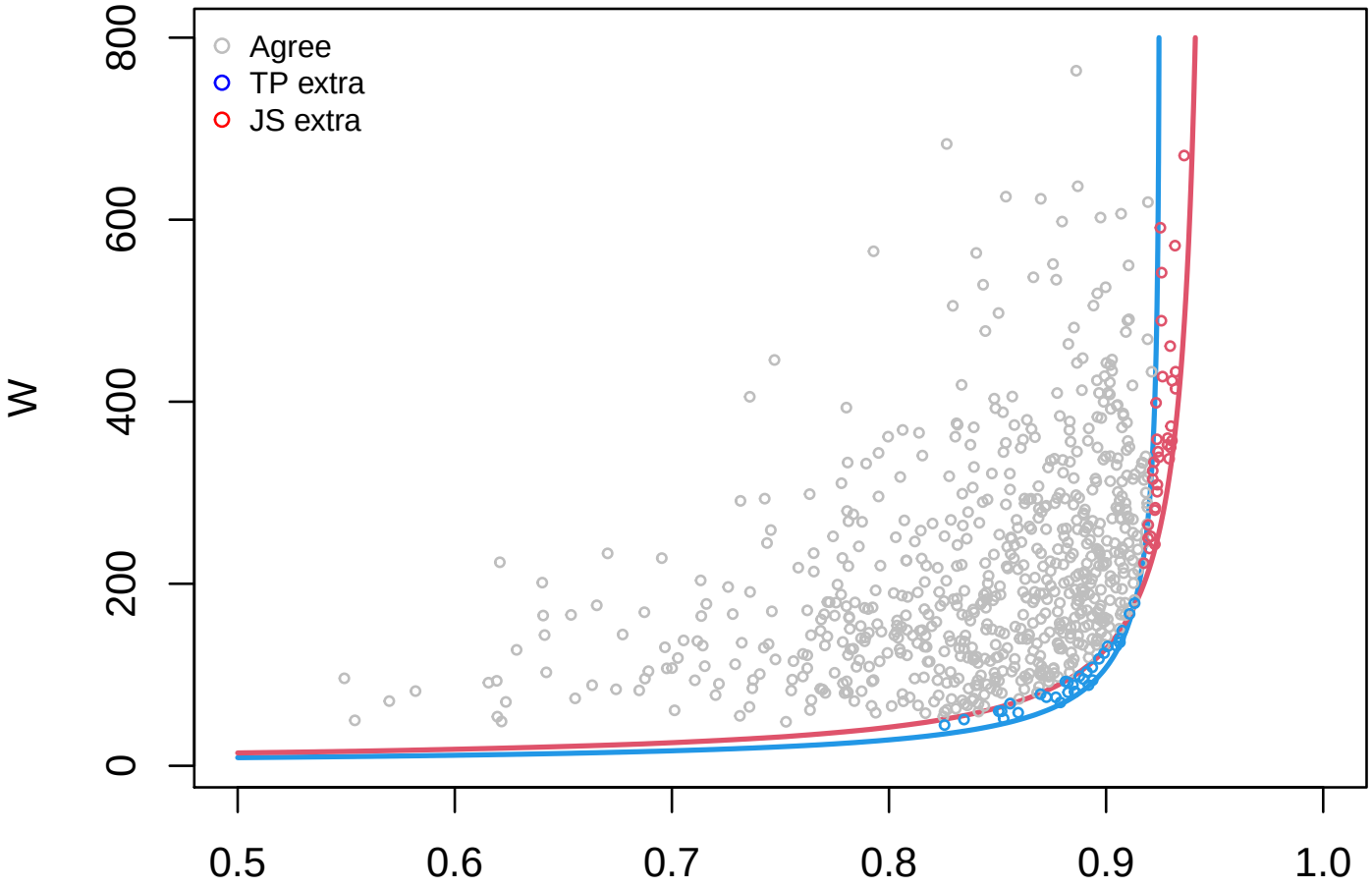
Centers selected as best (22% lowest mortality). Blue curve is selection boundary for posterior tail probability rule, vertical red line is MLE selection boundary. Red points are selected by MLE, but not TPR, blue points by TPR not MLE.

TP vs MLE Selection with $\gamma = 0.20$ FDR Control



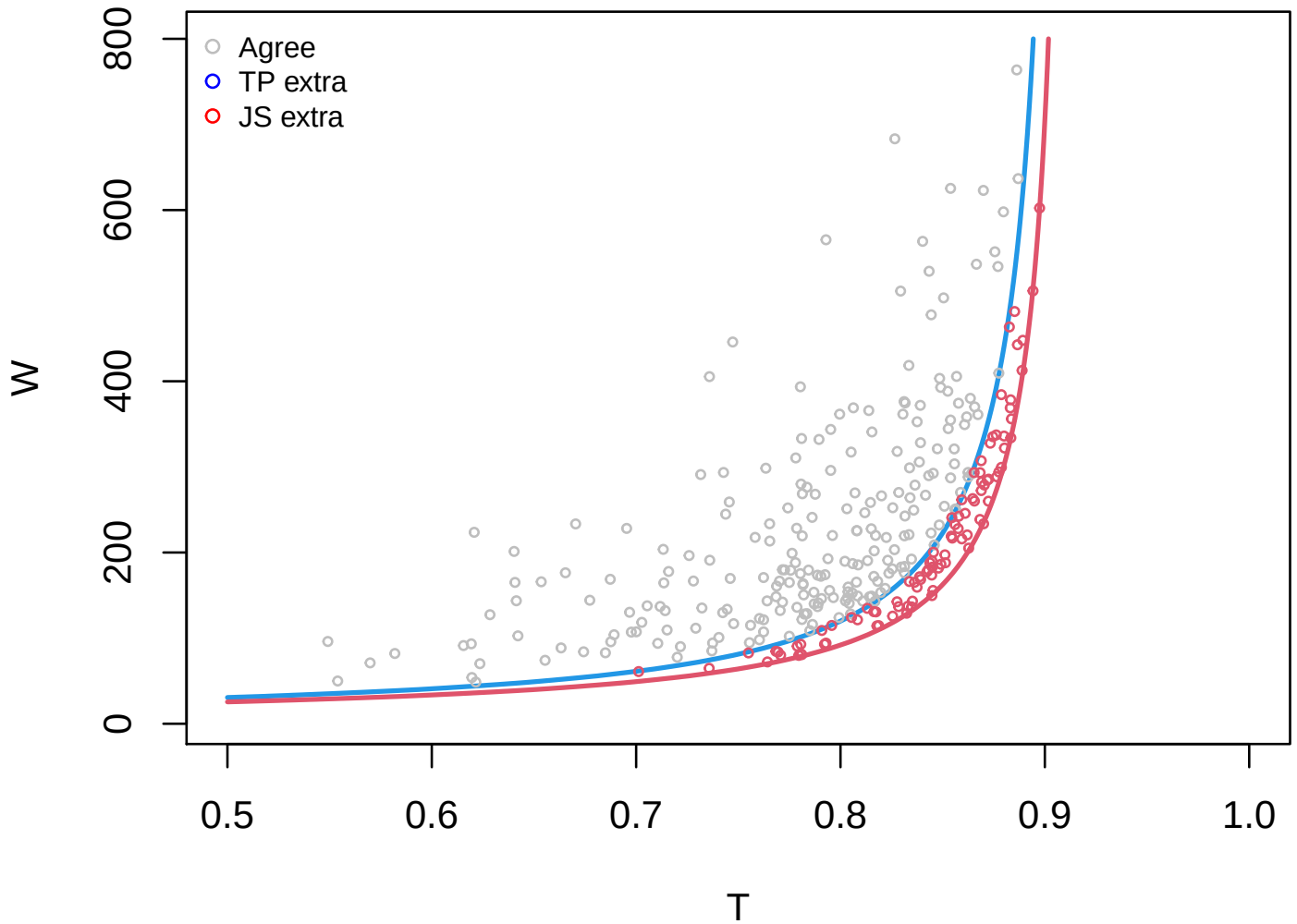
Centers selected as best (22% lowest mortality) with FDR constrained to 20%. Blue curve is selection boundary for TP rule, vertical red line is MLE selection boundary. Red points are selected by MLE, but not TPR, blue points by TPR not MLE.

Comparison of TP Selection with James-Stein Selection



Centers selected as best (22% lowest mortality). Blue curve is selection boundary for posterior tail probability rule, red curve is the James-Stein selection boundary. Red points are selected by James-Stein, but not TPR, blue points by TPR not James-Stein.

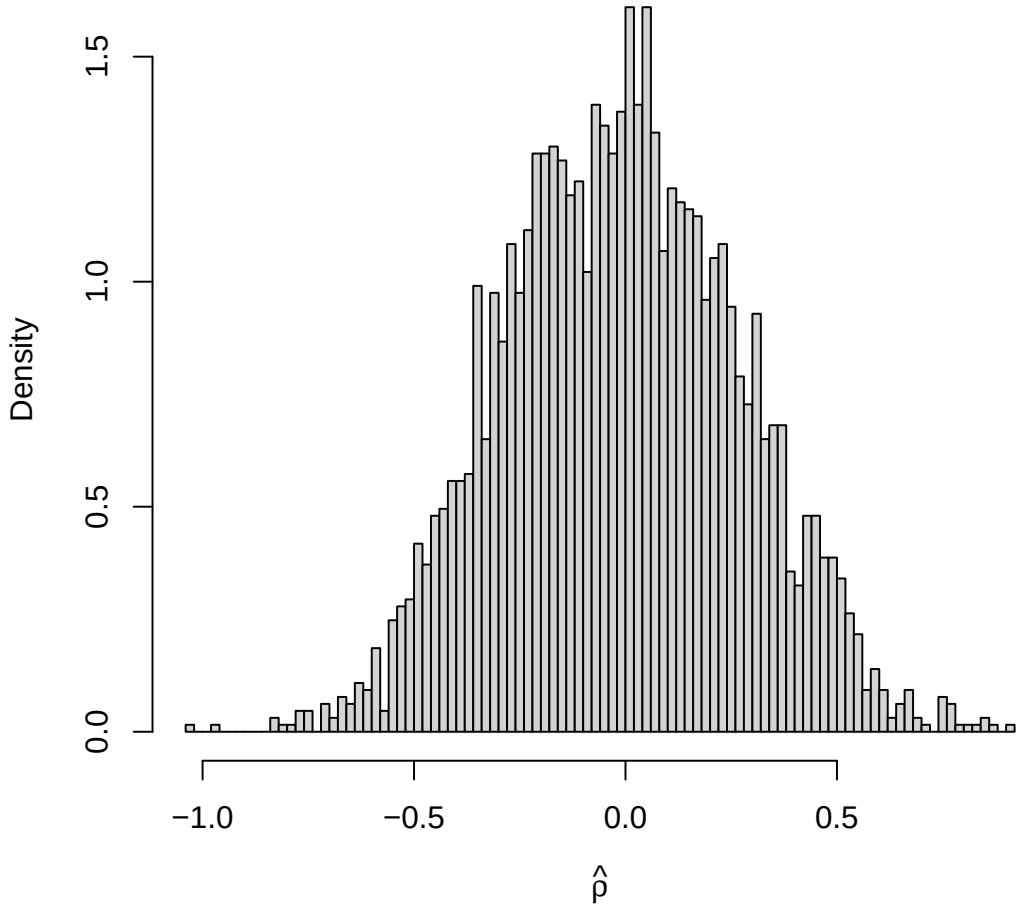
TP Selection vs James-Stein with FDR Control



Centers selected as best (22% lowest mortality). Blue curve is selection boundary for posterior tail probability rule, red curve is the James-Stein selection boundary. Red points are selected by James-Stein, but not TPR.

Temporal Stability of Ranking and Selection

One might wonder whether there was "significant" autocorrelation in the standardized mortality ratio that we have denoted by z_{it}. A histogram of the estimated AR(1) coefficients for the 3230 centers suggests considerable heterogeneity, but little systematic persistence.



Temporal Stability of Ranking and Selection II

Another way to explore temporal stability is to select centers into rating categories and estimate transition probabilities between categories. To implement this we compute posterior tail probability rankings for each of 5 3-year periods from 2004-2017. Centers are selected into one of 5 rating groups 22% A's, 30% B's, 35% C's, 9% D's and 4% F's, in each of these periods. The estimated Markov transition matrix looks like this:

	A	B	C	D	F
A	0.440	0.330	0.200	0.024	0.006
B	0.248	0.357	0.328	0.059	0.007
C	0.122	0.286	0.440	0.113	0.039
D	0.060	0.188	0.436	0.208	0.108
F	0.021	0.081	0.352	0.217	0.329

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

Conclusions

- Robbins's compound decision framework is well suited to the ranking and selection problem,
- The nonparametric MLE of Kiefer and Wolfowitz is an essential tool for compound decision making, but may need a little smoothing,
- Ranking and selection is difficult even for an Oracle who knows the probabilistic structure of the problem,
- Ranking and selection is especially difficult in Gaussian settings where conventional linear shrinkage methods are most appropriate.
- Nonparametric empirical Bayes methods are still somewhat mysterious from a formal theoretic viewpoint, so there are many important open questions.

Selected References I

BAHADUR, R. R., AND H. ROBBINS (1950): "The Problem of the Greater Mean," *The Annals of Mathematical Statistics*, 21, 469–487.

CHETTY, R., J. N. FRIEDMAN, AND J. E. ROCKOFF (2014): "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates," *American Economic Review*, 104, 2593–2632.

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 Competitiors - 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.

GILRAINE, M., J. GU, AND R. MCMILLAN (2020): "A New Method for Estimating Teacher Value-Added," NBER Working Paper Series Number 27094.

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 (2016): "Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective," *J. of Economic and Business Statistics*, forthcoming.

Selected References II

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.

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.

KOENKER, R., AND J. GU (2015): "REBayes: An R Package for Empirical Bayes Methods," Available from `https://cran.r-project.org/package=REBayes`.

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. (1978): "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Association*, 73, 805–811.

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*.

Selected References III

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.

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.

UNIVERSITY OF MICHIGAN KIDNEY EPIDEMIOLOGY AND COST CENTER (2009–2019): "Dialysis Facility Reports," available from: `https://data.cms.gov/dialysis-facility-reports`.