REPLY

JIAYING GU AND ROGER KOENKER

We would like express our appreciation to the discussants for their engaging and astute comments. We will begin by briefly addressing Kei Hirano's queries about links to classical decision theory, then touch on several extensions suggested in the comment by Pat Kline, and conclude with a brief excursion into data analysis to respond to the comments of Mogstad, Romano, Shaikh and Wilhelm.

1. What Is Bayesian about Empirical Bayes?

We are happy to concede that our analysis "falls somewhere between conventional statistical inference and a full blown decision theoretic analysis of Wald or Savage." This is the inevitable fate of the empirical Bayesian. From its inception Robbins's intention, as expressed in (Robbins 1990), was to *épater les bourgeois* of statistical orthodoxy. Empirical Bayes is neither Bayesian nor frequentist, and certainly not Neyman-Pearsonian, but it shares features of all of these. Our exposition in Section 2 was perhaps more Bayesian than really necessary, so we would like to take this opportunity to redress this imbalance with a somewhat more frequentist interpretation.

The example from (Robbins 1951) that we sketch in our Section 2 can be made to look very frequentist. We need not posit the existence of a prior distribution G from which the $\boldsymbol{\theta} = (\theta_i, \cdots, \theta_n)$ are drawn iidly, instead we can take the θ_i's as a fixed, deterministic binary sequence from $\Theta = \{-1, 1\}^n$. More important is that the Y_i are assumed to have identical conditional densities, $\varphi(y|\theta)$, and that loss is additively separable, $\bar{L}(\boldsymbol{\theta}, \boldsymbol{\delta}) = n^{-1} \sum |\theta_i - \delta_i|$. Robbins restricts attention to simple decision rules, $\delta_i = \delta(Y_i)$; this seems natural since the we are faced with n identical, but independent problems. Compound risk can then be written as,

$$\begin{aligned} R_n(\boldsymbol{\theta}, \boldsymbol{\delta}) &= n^{-1} \mathbb{E}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \bar{L}(\theta_i, \delta(Y_i)) \\ &= \sum_{i=1}^{n} n^{-1} \mathbb{E}_{\theta_i} L(\theta_i, \delta(Y_i)) \\ &= \int \int L(\theta, \delta(y)) \varphi(y|\theta) dy dG_n(\theta), \end{aligned}$$

where $G_n(A) = n^{-1} \sum \mathbb{1}\{\theta_i \in A\}$ for any Borel set from Θ. Thus, compound risk is equivalent to the Bayes risk of a single component of the compound problem with

Version: June 12, 2022.

prior, G_n, the empirical distribution function of the θ_i's. When the θ_i's take only two values G_n reduces to a scalar parameter and risk becomes,

$$R_n(\boldsymbol{\theta}, \boldsymbol{\delta}) = p_n(\boldsymbol{\theta}) \int L(1, \delta(y))\varphi(y|1)dy + q_n(\boldsymbol{\theta}) \int L(-1, \delta(y))\varphi(y|-1)dy$$

where $p_n(\boldsymbol{\theta}) = n^{-1} \sum \mathbb{1}\{\theta_i = 1\}$ and $q_n(\boldsymbol{\theta}) = 1 - p_n(\boldsymbol{\theta})$. Were $p_n = p_n(\boldsymbol{\theta})$ known, the optimal decision rule would be,

$$\delta^*_{p_n}(y) = \mathrm{sgn}(y + \tfrac{1}{2}\log(p_n/(1-p_n)))$$

Of course we probably don't "know" p_n, how could we? But many candidate estimators of p_n present themselves, of which Robbins's method of moments choice $\hat{p}_n = (\bar{y} + 1)/2$ is simplest. But is it really a simple rule? We promised to use only simple rules of the form, $\hat{\theta}_i = \delta(Y_i)$ and $\delta^*_{p_n}(y)$ is surely like that, but once we put a hat on $\hat{p}_n$ the rabbit is poised to make an appearance. Yet nothing is lost as (Hannan and Robbins 1955) show that the risk of $\delta^*_{\hat{p}_n}(y)$ uniformly approximates the risk of $\delta^*_{p_n}(y)$.

How does this relate to Wald's minimax proposal? Robbins proves that $\sup_{\boldsymbol{\theta}} R(\boldsymbol{\delta}, \boldsymbol{\theta})$ is minimized with the naive rule $\tilde{\delta}(y) = \mathrm{sgn}(y)$, which is equivalent to $\delta^*_{1/2}(y)$. However, it is easy to verify that for any $p_n \neq 1/2$, $R(\delta^*_{1/2}, \boldsymbol{\theta}) \geq R(\delta^*_{p_n}, \boldsymbol{\theta})$ and furthermore that for any $\epsilon > 0$ there exists $n(\epsilon)$ such that for $n > n(\epsilon)$, $R(\delta^*_{\hat{p}_n}, \boldsymbol{\theta}) - R(\delta^*_{1/2}, \boldsymbol{\theta}) < \epsilon$ for any $\boldsymbol{\theta}$. Thus, although not an admissible rule – the naive rule is always superior when $p_n = 1/2$ – the compound decision rule is only an asymptotically negligible bit worse at $p_n = 1/2$, and potentially much better elsewhere. See (Hannan and Robbins 1955) and (Samuel 1955) for further formal details, and (Gu and Koenker 2016) for some numerical comparisons.

The foregoing example may seem overly simplified, after all our prior only required estimation of a single parameter, however similar structure arises in many other settings such as our ranking and selection problems where the prior can be much more complex. The crucial feature of such compound decision problems is the permutation invariance of both the probabilistic structure of the problem and the loss function being considered. And as we have argued elsewhere, estimation of the mixing distribution whether it is viewed as G_n or G is often a relatively benign convex optimization problem.

Regarding our loss function, there is more than a whiff of Neyman-Pearson about our α and γ. No doubt that it would be better to have loss defined on a more explicit action space, but like priors loss functions are difficult to elicit. By accentuating the connection to multiple testing, we have tried to highlight the balance that must be struck between the intended size of the selected population and the accuracy of the selection. This trade-off seems inherent in any ranking and selection problem. At a more fundamental level one may object to the nature of compound loss itself; why should component losses be aggregated in such a symmetric fashion? To this, our only answer is: why not? If the model is permutation invariant shouldn't the loss be as well?

The Le Cam limit experiment perspective has proven to be a powerful device in many decision theoretic settings and could do so in ranking applications provided we adhere to the (Le Cam 1990) "Principle 7: If you need to use asymptotic arguments, don't forget to let your number of observations tend to infinity," while maintaining the heterogeneity of the latent structure of the problem. Whether it can be deployed effectively in the compound decision framework to justify forms of shrinkage like those we have considered is, indeed, a very intriguing open question.

2. Challenges and Opportunities

Pat Kline has raised many important issues that deserve an extended response; we are only able to offer some superficial hints that might help guide future exploration.

- The form of G is a critical determinant of the difficulty of the ranking and selection problem. The atomic form of the Kiefer-Wolfowitz NPMLE is especially well suited to discrete G, but even more critical is its tail behavior. Heavier tails, like those of the lognormal or Student with low degrees of freedom make selection easier, lighter tails like the uniform or Gaussian make it more difficult. How variance stabilizing transformations influence this is delicate since such transformations affect both location and scale of the observations.
- Independence of latent location and scale parameters is a critical assumption. In prior work on income dynamics (Gu and Koenker 2017) we have relaxed this assumption and modeled location and scale of the log income process jointly and finding a negative association. This entails bivariate gridding to compute the NPMLE of the joint mixing distribution, but imposes no new difficulties in principle. In many applications this approach will seem quite natural, in others where scale is determined by variation in exogenous sample size, it may be unnecessary.
- Multinomial selection as in dialysis center grading again raises the questions, how is selection to be used and how should the loss function be structured?
- Posterior means and posterior tail probabilities are only two of many possible criteria. An advantage of the empirical Bayes formulation we have proposed is that ultimately the decision maker is confronted with the entire posterior distribution for each individual or firm being evaluated and more sophisticated forms of risk aversion can be contemplated.
- Ranking is inherently relative so the legal systems quest for absolute standards will seem quixotic in many circumstances. Raising awareness of the uncertainties associated with rankings is a more feasible objective. We hope that empirical Bayes methods can help achieve this.

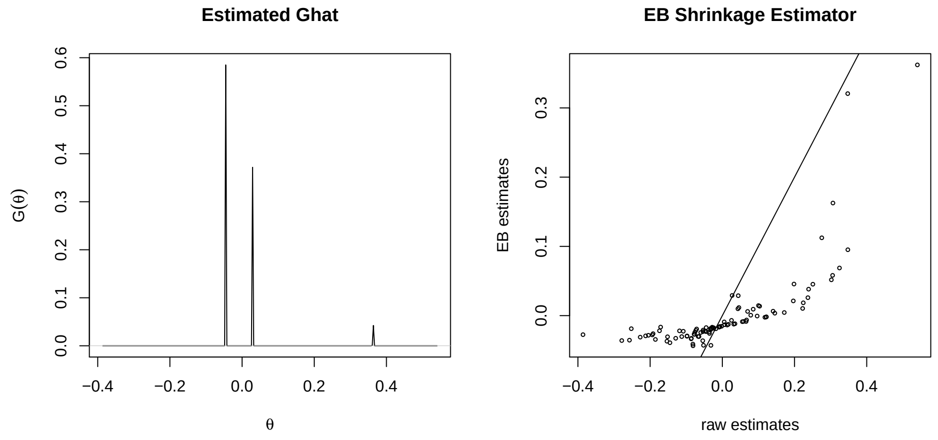

FIGURE 1. The estimate $\hat{G}$ and $\mathbb{E}[\theta|y_i, s_i]$ from the county level estimates with the movers' design. The solid line is the 45 degree line.

3. SOME COMPARATIVE DATA ANALYSIS

Mogstad, Romano, Shaikh and Wilhelm, hereafter MRSW, have provided a very valuable comparison of the ranking methods they have proposed in their 2020 paper with our empirical Bayes procedures. We will try to draw out a few more implications from these comparisons. We have already noted that the distinction between fixed and random θ_i's is perhaps not quite as essential as it might seem. More salient is the way ranks are constructed and their precision evaluated by the two approaches. Our empirical Bayes relies on an estimate, $\hat{G}$, of the distribution of the latent θ_i's to construct posterior distributions of each θ_i, and thereby posterior means and posterior tail probabilities. So the burden of the ranking exercise is borne by the way that the observed Y_i's and their associated σ_i's get baked into the "prior" $\hat{G}$ pie. In contrast, MRSW employ resampling and multiple testing methods to control family-wise error for the $\binom{n}{2}$ pairs, resulting in a much more stringent selection criterion.

As an initial comparison, consider the "correlational" estimates of intergenerational mobility and their standard errors from (Chetty, Friedman, Hendren, Jones, and Porter 2018). Restricting to the top 100 commuting zones, as in (Mogstad, Romano, Shaikh, and Wilhelm 2020), we see that these effect sizes are very precisely estimated: point estimates are all in the interval, $[0.325, 0.457]$, while standard errors are all the interval $[0.00035, 0.0025]$. The consequence of this is that the NPMLE, $\hat{G}$ assigns positive mass to almost all of the initial estimates, and posterior mean and posterior tail probability rankings are essentially the same as just ranking the initial estimates. FDR control is non-binding and selection under our EB approach would confidently just take the top α commuting zones as revealed by the raw estimates.

If we now consider the stricter "mover" design of (Chetty and Hendren 2018) intended to identify causal effects of mobility we see that the point estimates are much less precisely estimated since they are based on much smaller sample sizes. Focusing on the most populous 100 commuting zones and counties, the NPMLE, illustrated in Panel (a) of Figure 1 has only three distinct mass points. Unlike in the "correlational" design where the Bayes rule did essentially no shrinkage, now there is considerable shrinkage as shown in panel (b) of the figure that plots the raw Y_i estimates against their posterior means. With these considerably more noisy estimates, FDR control becomes again relevant. For the commuting zone data, setting capacity constraint at $\alpha = 0.10$, and the FDR control parameter $\gamma = 0.30$ our posterior tail probability criterion selects no commuting zones for "top 10" status. Similarly, the MRSW procedure with $\alpha = 0.10$ places all 100 CZs into an uninformative category covering all possible ranks from 1 to 100. The situation changes somewhat when we consider counties rather than CZs. Maintaining the capacity constraint at $\alpha = 0.10$, Table 1 reports the counties selected into the top 10 at several different FDR control levels. When γ is set at 0.30, our posterior tail probability rule selects four counties for the top 10; tightening γ to 0.05 reduces the number selected to two. The more stringent procedure of MRSW still produces intervals that cover the entire support of the ranks from 1 to 100 for all the counties.

County	y	s	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$
Dupage	0.540	0.123	×	×	×	×
Bucks	0.348	0.176				
Macomb	0.347	0.109	×	×	×	×
Hartford	0.325	0.182				
Contra Costa	0.306	0.129			×	×
Ventura	0.306	0.181				
Bergen	0.302	0.186				
Pinellas	0.276	0.127				×
Snohomish	0.251	0.154				
Providence	0.239	0.153				

Table 1. Selection of the top 10 counties based on the causal estimates for the 100 most populous counties with a mover's design in Chetty and Hendren (2018). The order of the 10 counties appearing here is based on their raw point estimates y.

In this more uncertain setting, we can also see how variances play a role in our EB procedure. If we compare Bucks with Macomb, the two counties have almost identical point estimates, however Macomb is more precisely estimated. Given our $\hat{G}$, or preferable a smoothed version, $\tilde{G}$ we can easily compute the whole posterior distribution of θ updated for any observed pair of (y, s). For Macomb the updated

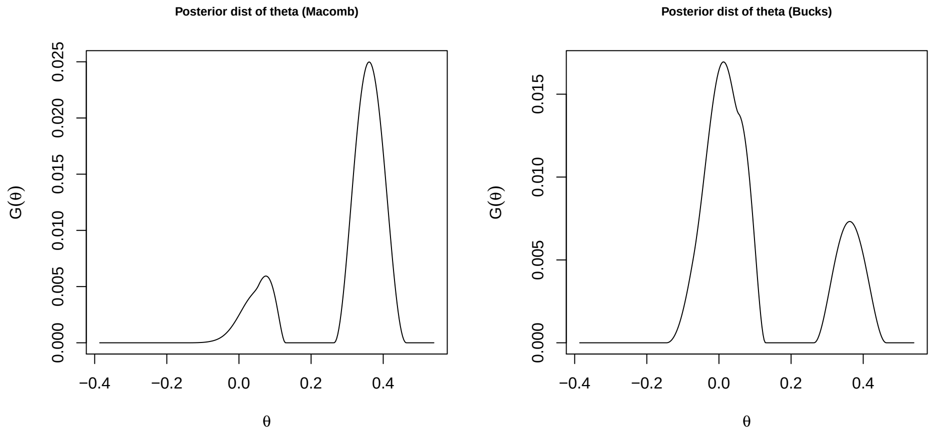

Figure 2. The posterior distributions of θ for Macomb (left) and Bucks (right) counties based on the kernel smoothed $\tilde{G}$ with biweight kernel and bandwidth 0.10.

posterior puts most of its weight on the rightmost mode near 0.4. For Bucks, because its point estimate is less precise most of its weight is attracted to the mass point near 0. This reduces the likelihood that Bucks will have a posterior tail probability for its θ to be in the upper 90% quantile, and helps to explain why it is never selected in Table 1 even though it is ranked second by observed y_i's. The posteriors for these two counties are illustrated in Figure 2.

The foregoing comparisons illustrate why it is difficult to construct reliable rankings and make credible selection decisions. The information congealed in the NPMLE, $\hat{G}$, can aid this process but it cannot help when the underlying data is too noisy, and it is superfluous when the underlying data is too precise. In between these extremes there is room for improvement in current ranking and selection practices. In some settings, like the county level mobility example we have described, balancing FDR control with reasonable capacity constraint using our empirical Bayes procedures may prove useful. In high stakes situations like teacher evaluation even more stringent criterion like that of MRSW may be preferred, at an inevitable cost of reduced power.

References

Chetty, R., J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter (2018): "The opportunity atlas: Mapping the childhood roots of social mobility," Discussion paper, National Bureau of Economic Research.

Chetty, R., and N. Hendren (2018): "The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects," *The Quarterly Journal of Economics*, 133, 1107–1162.

Gu, J., and R. Koenker (2016): "On a Problem of Robbins," *International Statistical Review*, 84, 224–244.

Gu, J., and R. Koenker (2017): “Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective,” *Journal of Business and Economic Statistics*, 35, 1–16.

Hannan, J. F., and H. Robbins (1955): “Asymptotic Solutions of the Compound Decision Problem for Two Completely Specified Distributions,” *The Annals of Mathematical Statistics*, 26, 37 – 51.

Le Cam, L. (1990): “Maximum Likelihood: An Introduction,” *International Statistical Review*, 58, 153–171.

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. (1951): “Asymptotically subminimax solutions of compound statistical decision problems,” in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, vol. I. University of California Press: Berkeley.

Robbins, H. (1990): “The Origins of Empirical Bayes: Butterflies, Oysters and Stars,” Available at https://www.youtube.com/watch?v=id6YSycD5lc, 10th Annual Pfizer Lecture at the University of Connecticut.

Samuel, E. (1955): “On Simple Rules for the Compound Decision Problem,” *J. Royal Statistical Society (B)*, 27, 238–244.