Econometrica, Vol. 91, No. 1 (January, 2023), 61–66

REPLY TO: Comments on *"Invidious Comparisons: Ranking and Selection as Compound Decisions"*

JIAYING GU
Department of Economics, University of Toronto

ROGER KOENKER
Department of Economics, University College London

WE WOULD LIKE to 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 $\boldsymbol{\theta} = (\theta_i, \ldots, \theta_n)$ 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 restricted attention to simple decision rules, $\delta_i = \delta(Y_i)$; this seems natural since we are faced with n identical, but independent problems. Compound risk can then be written as

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

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 prior, G_n, the

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

 https://doi.org/10.3982/ECTA20537

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

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

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

$$\delta^*_{p_n}(y) = \operatorname{sgn}\left(y + \frac{1}{2}\log\big(p_n/(1-p_n)\big)\right).$$

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) showed 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 proved that $\sup_{\boldsymbol{\theta}} R(\boldsymbol{\delta}, \boldsymbol{\theta})$ is minimized with the naive rule $\tilde{\delta}(y) = \operatorname{sgn}(y)$, 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 tradeoff 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 qualify and guide their future exploration.

- Testing, applied mechanically, has many drawbacks and Kline is totally justified in questioning our interpretation of decision rules that may appear to be based on rather arbitrary α and γ choices for capacity and FDR constraints, respectively. In our defense, we would stress that, from its inception, the Neyman–Pearson testing apparatus emphasized that choices of test levels and their associated cut-offs should be based on careful consideration of the relative costs of mistakes. These costs would usually be directly tied to the *rating* evaluations that have been used to generate the rankings. Thus, for example, the choice of how many firms to prosecute for hiring discrimination should depend upon the absolute magnitude of the severity of discrimination and the precision with which it is measured. Relative rankings come into play only afterwards. This is in accord with the usual practice of assigning letter grades in academic testing: first, one looks for gaps in the score distribution to determine cut-offs, and only then are grades assigned according to the ranking of the scores.
- When rankings are determined by a scalar rating measurement with homogeneous precision, there can be little controversy about ranking, but the cut-off for selection may, as we have noted, be influenced by the estimation of ratings. In contrast, with heterogeneous precision of the ratings, the rankings themselves are called into question: should they be based on posterior mean ratings, posterior tail probabilities, or some other criterion? Such heteroscedastic environments offer more opportunity for introducing risk aversion into decision making; imagine selecting students for college admission, for example. We have proposed posterior tail probability selection rules as a way to balance capacity and FDR objectives, but we acknowledge that other loss functions may find favor in some applications.
- Our cautionary remarks about the challenges of ranking and selection in settings with Gaussian heterogeneity, G, should not be interpreted as an expression of the universal futility of the selection problem. On the contrary, we agree with Kline that the advantage of nonparametric methods of G-modeling, like the Kline–Walters method of moments approach to their paired binomial problem, or the Kiefer–Wolfowitz NPMLE we have used, is that they are able to reveal more general mixing distributions and lead to more informative shrinkage rules than are typically employed in parametric empirical Bayes applications.
- The use of variance stabilization transformations to achieve approximately Gaussian behavior for Poisson settings like our dialysis center observations or binomials for baseball batting averages is intended to alter only the base distribution of the mixture and leave the mixing distribution, G, intact. To what extent this intention is realized deserves further study, as Kline suggests. Clearly, it rests on reasonably large Poisson intensities and binomial sample sizes to justify the approximation.
- When one observes only a few distinct frequencies, as with low-dimensional binomials, identification becomes a paramount concern since only a few moments of G are identified, not the entire distribution as noted by Lindsay. But as shown by Kline and Walters, effective decision rules can still be crafted.
- Ranking is an inherently relative enterprise, while the legal system's quest for absolute standards may be quixotic in many circumstances. Raising awareness of the uncertainties associated with rankings seems a more feasible objective. We hope that nonparametric empirical Bayes methods can help achieve this.

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 I 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.

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 preferably 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 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

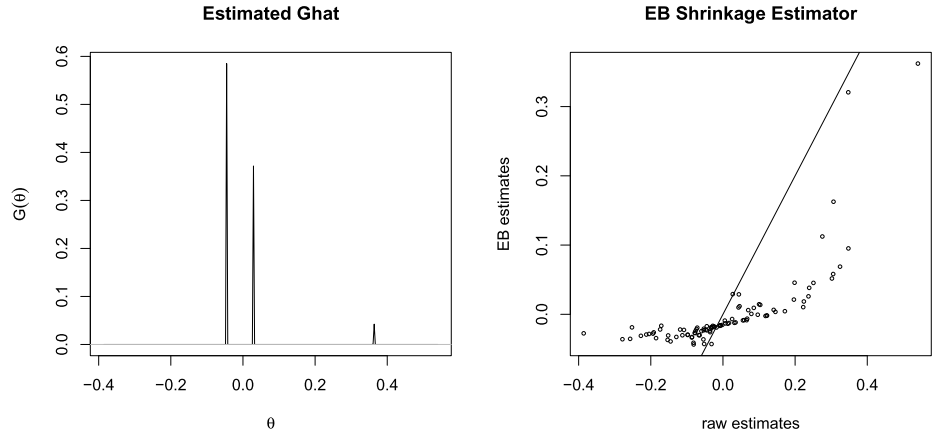

FIGURE 1.—The estimate $\hat{G}$ and $\mathbb{E}[\theta|y_i, s_i]$ from the county-level estimates with the movers' design. The solid line is the 45 degree line.

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 I 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 contained in the NPMLE, $\hat{G}$, can aid this process but it cannot help when the underlying data are too noisy, and it is superfluous when the underlying data are 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, an even more stringent criterion like that of MRSW may be preferred, at an inevitable cost of reduced power.

TABLE I

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.

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				

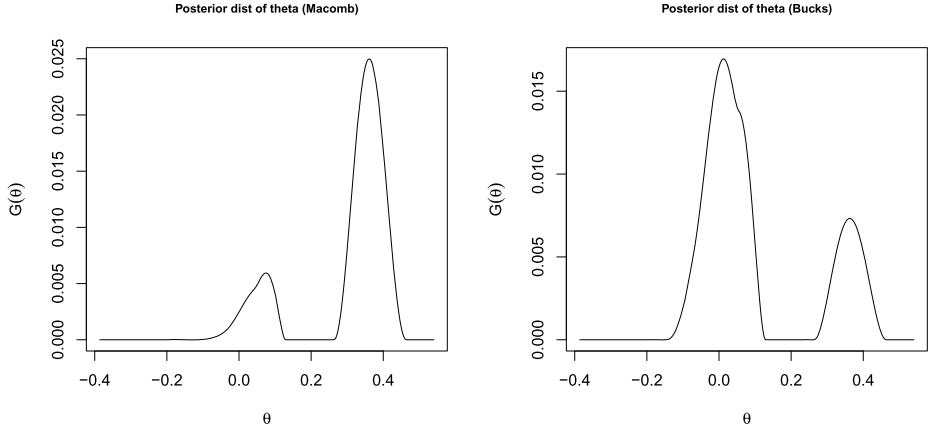

FIGURE 2.—The posterior distributions of θ for Macomb (left) and Bucks (right) counties based on the kernel smoothed $\tilde{G}$ with biweight kernel and bandwidth 0.10.

REFERENCES

CHETTY, RAJ, AND NATHANIEL HENDREN (2018): "The Impacts of Neighborhoods on Intergenerational Mobility i: Childhood Exposure Effects," *The Quarterly Journal of Economics*, 133, 1107–1162. [64,65]

CHETTY, RAJ, JOHN N. FRIEDMAN, NATHANIEL HENDREN, MAGGIE R. JONES, AND SONYA R. PORTER (2018): "The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility," Tech. rep, National Bureau of Economic Research. [64]

GU, JIAYING, AND ROGER KOENKER (2016): "On a Problem of Robbins," *International Statistical Review*, 84, 224–244. [62]

HANNAN, JAMES F., AND HERBERT ROBBINS (1955): "Asymptotic Solutions of the Compound Decision Problem for Two Completely Specified Distributions," *The Annals of Mathematical Statistics*, 26, 37–51. [62]

LE CAM, LUCIEN (1990): "Maximum Likelihood: An Introduction," *International Statistical Review*, 58, 153–171. [62]

MOGSTAD, MAGNE, JOSEPH ROMANO, AZEEM SHAIKH, AND DANIEL WILHELM (2020): "Inferences for Ranks With Applications to Mobility Across Neighborhoods and Academic Achievement Across Countries," Preprint. [64]

ROBBINS, HERBERT (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. [61]

——— (1990): "The Origins of Empirical Bayes: Butterflies, Oysters and Stars," 10th Annual Pfizer Lecture at the University of Connecticut, Available at https://www.youtube.com/watch?v=id6YSycD5lc. [61]

SAMUEL, ESTER (1955): "On Simple Rules for the Compound Decision Problem," *J. Royal Statistical Society (B)*, 27, 238–244. [62]

Editor Guido Imbens handled this manuscript.

Manuscript received 28 January, 2022; final version accepted 31 January, 2022; available online 16 June, 2022.