Econometrica, Vol. 91, No. 1 (January, 2023), 47–52

A COMMENT ON: *"Invidious Comparisons: Ranking and Selection as Compound Decisions" by Jiaying Gu and Roger Koenker*

PATRICK KLINE
Department of Economics, UC Berkeley and NBER

GU AND KOENKER (2020, HENCEFORTH GK) OFFER EMPIRICISTS BESET by the "league table mentality" a powerful suite of tools for selecting extreme performers based on noisy measurements. GK's basic insight is to reformulate tail selection as a large scale inference problem in which the selection of a unit failing to exceed some tail performance threshold constitutes a false discovery. Utilizing nonparametric empirical Bayes (EB) methods to estimate the latent distribution of performance, units are ranked according to their posterior tail probability, which plays the role of a test statistic. A critical value is then chosen to trade off the False Discovery Rate (FDR) against the False Non-discovery Rate (FNR).

Chastened by the difficulties of discerning tail performance, GK warn that "there is something inherently futile about many ranking and selection problems." To appreciate their angst, it is instructive to consider GK's example of selecting teachers with extremely low value added. In an empirical calibration, they find that even an oracle who knows the true value added distribution, when searching for teachers in the bottom centile of value added, is able to select only 0.4% of teachers before a 5% FDR constraint binds. Hence, the "price" paid for ensuring a low FDR is a high FNR: at best, only 40% of teachers in the bottom centile are selected, implying an FNR of at least 60%.

Mistakes are inevitable in statistical ranking exercises. Yet even very noisy rankings, when accompanied by appropriate measures of uncertainty, can be quite valuable to audiences who might otherwise be forced to rely on anecdotes to make targeting decisions. Together, the FDR and FNR offer a transparent assessment of the sorts of ranking mistakes expected to arise from repeated application of a tail selection rule. GK make a compelling case that EB estimates of these error rates should become standard diagnostics in econometric selection exercises.

Below, I review the mechanics of GK's proposal, highlighting some implementation decisions researchers are likely to encounter when seeking to apply these tools. After offering some thoughts on why tail selection exercises may eventually prove less futile than GK fear, I consider some limitations of the analogy between testing and selection problems, concluding with a reminder that league table reasoning, alluring though it may be, can prove counterproductive when interest centers on absolute rather than relative standards of performance.

EMPIRICAL BAYES POSTERIORS

Suppose we have run an experiment at a set of "sites" indexed by $i = 1, \ldots, n$ and collected estimates $\{Y_i\}_{i=1}^n$ of site-specific treatment effects $\{\theta_i\}_{i=1}^n$. These sites might correspond to actual locations, across which we expect treatment effect heterogeneity, or to

Patrick Kline: pkline@econ.berkeley.edu

I am grateful to Evan Rose, Andres Santos, Ben Scuderi, and Chris Walters for helpful feedback on an earlier draft of this comment.

 https://doi.org/10.3982/ECTA20322

particular individuals or firms, whose aptitude or conduct we seek to evaluate. Treating the sites as exchangeable, each Y_i is modeled as an i.i.d. draw from the compound density

$$f_G(y) = \int p(y|t)\, dG(t),$$

where $G(t) = \Pr(\theta_i < t)$ is the unknown distribution of treatment effects and $p(y|t) = \frac{d}{dy}\Pr(Y_i < y \mid \theta_i = t)$ is a known density governing noise in the estimates. For example, Y_i might give the experimental contrast in mean outcomes at site i scaled by its standard error estimate, in which case it is natural to appeal to a central limit theorem in choosing $p(y|t) = \varphi(y - t)$, where φ is the standard normal density. Alternately, when working with count outcomes, it is common to employ a concave variance stabilizing transformation $y \mapsto v(y)$ that yields the approximation $v(Y_i) \mid \theta_i \sim \mathcal{N}(v(\theta_i), 1)$ (Bartlett (1947), Anscombe (1948)). When the estimate Y_i has been variance stabilized, $p(y|t)$ is again a standard normal and G may be redefined as giving the distribution of $v(\theta_i)$.

Consider now an "oracle" who knows G with certainty. After observing the vector of estimates $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, her posterior density over any vector $\boldsymbol{t} = (t_1, \ldots, t_n)'$ of treatment effects is

$$\rho_G(\boldsymbol{t}|\boldsymbol{Y}) = \prod_{i=1}^{n} \{p(Y_i|t_i)\, dG(t_i)/f_G(Y_i)\},$$

which can be used to assess the likelihood of any ranking of the sites. The EB ethos is to "borrow strength" across sites by constructing an estimate $\hat{G}$ of G. The plug-in predictive density $\rho_{\hat{G}}(\boldsymbol{t}|\mathbf{Y})$ can then be used to make selection decisions that minimize expected loss or to construct a posterior credible interval for the rank of any site.

Many approaches to estimating G have been proposed. GK favor nonparametric maximum likelihood estimation (NPMLE), which yields a $\hat{G}$ with mass points, while Efron (2016) suggested a penalized maximum likelihood approach generating a smooth $\hat{G}$ in the exponential family. In economics, one typically expects the true G to be continuous; moreover, a jumpy $\hat{G}$ will yield a posterior $\rho_{\hat{G}}$ with mass points, which can complicate inference (Koenker (2020)). Mindful of these considerations, GK opt to smooth their NPMLE estimates in a second step. In practice, decision rules based on the smoothed NPMLE and the exponential approximation tend to exhibit comparable performance. Whichever approach is used, the tuning parameters employed to construct $\hat{G}$ should be documented clearly. For example, Kline, Rose, and Walters (2021) choose the penalization parameter in an exponential approximation to ensure the variance of $\hat{G}$ matches unbiased variance estimates.

MAKING SELECTION DECISIONS

GK motivate their approach to tail selection with a discrete loss function balancing type I selection errors ("false discoveries") against type II errors ("false non-discoveries") and capacity constraints. The empirical loss of an n-vector of binary selection indicators $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$ given the latent vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ is

$$L(\boldsymbol{\delta}, \boldsymbol{\theta}) = \underbrace{\sum_{i=1}^{n} h_i(1 - \delta_i)}_{\text{False non-discoveries}} + \tau_1 \Bigg\{ \underbrace{\sum_{i=1}^{n} (1 - h_i)\delta_i}_{\text{False discoveries}} - \gamma \underbrace{\sum_{i=1}^{n} \delta_i}_{\text{Total discoveries}} \Bigg\} + \tau_2 \Bigg(\underbrace{\sum_{i=1}^{n} \delta_i}_{\text{Total discoveries}} - \alpha n \Bigg),$$

where $h_i = 1\{\theta_i \geq G^{-1}(1-\alpha)\}$ is an indicator for θ_i residing in the relevant α-tail of G. The first term captures that the researcher does not want to miss extreme performers. When τ_1 is positive, she is additionally concerned with the False Discovery Proportion (FDP): $\sum_{i=1}^n (1-h_i)\delta_i / \sum_{i=1}^n \delta_i$. Finally, when τ_2 is positive, selection itself is costly.

To estimate the risk of a decision rule, each unobserved h_i is replaced by its EB posterior expectation $\hat{\nu}_\alpha(Y_i) = \int_{\hat{G}^{-1}(1-\alpha)}^{\infty} p(Y_i|t)\, d\hat{G}(t)/f_{\hat{G}}(Y_i)$, which serves as a ranking device. Further integrating Y_i against $f_{\hat{G}}$, GK treat (τ_1, τ_2) as Lagrange multipliers on ex ante FDP and capacity constraints. Consequently, the optimal strategy is to select as many highly ranked sites as possible, subject to the FDR—that is, the posterior expected FDP (Storey (2003))—falling below a pre-specified level γ and the expected number of selections failing to exceed αn.[1]

By constraining the FDR, GK depart somewhat from the tenets of statistical decision theory, mimicking the familiar Neyman–Pearson testing paradigm and its emphasis on size control. In applications, the choice of the FDR threshold γ should be motivated by a careful description of the steps that are likely to follow selection. If mistakes regarding which sites are in the α-tail are not particularly costly, 20% may constitute a reasonable FDR threshold. For example, in adaptive experimentation schemes (e.g., Avivi, Kline, Rose, and Walters (2021)), an initial screen to find sites satisfying a pilot FDR constraint can be followed by a subsequent confirmatory analysis exhibiting a more stringent threshold. By contrast, a decision-maker who withholds pay from underperforming teachers selected based upon an FDR criterion may well find themselves in court, in which case it seems prudent to use a small γ from the outset.

The researcher must also choose the tail parameter α, which defines the null hypothesis under consideration. As Efron (2004) notes, the choice of α can be viewed as establishing which values of θ_i are deemed "interesting." For a fixed γ, the number of selected sites will tend to increase with α. To enumerate the different interpretations that can be attached to the selection of s top-ranked sites, one can plot the FDR/tail threshold frontier $\mathcal{R}_s = \{(\gamma, \alpha) : \sum_{i=1}^n \delta_i^*(Y_i) = s\}$, where δ_i^* is the optimized selection function.

IS TAIL SELECTION FUTILE?

GK note that "if the latent measure of true quality is Gaussian, as assumed in virtually all of the econometric applications of the selection problem, and we wish to select the top ten percent of individuals given that their true quality is contaminated by Gaussian noise, accurate selection can be very challenging when the signal to noise ratio is low." While teacher value added may be reasonably well approximated by a Gaussian, in part because test scores are designed to exhibit normal distributions, there is no particular reason to expect other forms of heterogeneity to exhibit Gaussian tails.

An illustrative example comes from Kline and Walters (2021), who analyzed correspondence experiments of employer discrimination that randomly assign racially distinctive names to job applications. Modeling applications to each job i as Bernoulli trials with race-specific employer contact probabilities (p_{iw}, p_{ib}), they studied the distribution G of racial contact gaps $\theta_i = p_{iw} - p_{ib}$ in an experiment devised by Nunley, Pugh, Romero, and Seals (2015) that sent to each job four applications with a randomly determined mix of distinctively white and Black names. Though G is only partially identified when a small

[1] As GK note, technically the constraint is on the marginal FDR $\sum_{i=1}^n \mathbb{E}[(1-h_i)\delta_i] / \sum_{i=1}^n \mathbb{E}[\delta_i]$, which is asymptotically equivalent to the traditional FDR notion $\mathbb{E}[\sum_{i=1}^n (1-h_i)\delta_i / \sum_{i=1}^n \delta_i]$ as n grows large (Genovese and Wasserman (2002)).

number of applications are sent to each job, method of moments estimates reveal that G exhibits substantial skew and kurtosis, suggesting a long tail of heavy discriminators with $\theta_i \gg 0$.

Even so, one might think that trying to select jobs with $\theta_i > 0$ from such an experiment is a fool's errand. Consider, for example, a job that contacts all three of the seemingly white applications it is sent but declines to contact the one fictitious Black application it receives. The Fisher (1922) p-value for the null hypothesis that race is independent of contacts at this job is $\binom{4}{3}^{-1} = 0.25$. However, if the job in question is a random draw from a population where discrimination is known to be rampant, the posterior probability that this null is true may be much lower. Searching across the space of distributions consistent with the Nunley et al. experiment, Kline and Walters found that at least 82% of the jobs that contact three of four applicants in such settings discriminate against Black applicants. Using this lower bound rate of prevalence, they estimated that at least 90% of the jobs contacting three white and no Black applicants exhibit $\theta_i > 0$. That the FDR can be controlled to 10% with only four experimental observations per site illustrates the potentially enormous value of borrowing strength from the rest of an experiment.

Of course, tail selection can be substantially more difficult than the problem of selecting sites where an effect of any sort is present. The power of any selection procedure will invariably depend on both the shape of G and the precision of the measurements $\{Y_i\}_{i=1}^n$. The application of variance stabilizing transforms will typically reduce the skew in G, which may make tail selection more difficult. An interesting topic for future research is to compare the power of methods that jointly estimate the distribution of scale and location parameters to those that transform scale parameters away.

When EB methods fail to produce sufficiently many "interesting" sites subject to desired FDR restrictions, the onus should be on researchers to seek out or create new data sets that measure up to the task at hand. Motivated by the limitations of existing audit studies, Kline, Rose, and Walters (2021) devised a correspondence experiment of large employers, each of which was sent up to 1000 fictitious applications, with the preregistered intention of applying EB methods to detect discriminating firms. Funding agencies typically ask that experiments be designed to achieve 80% power against alternatives of interest. Researchers planning experiments aimed at large scale selection may instead wish to target particular FNRs— perhaps 20%, in analogy with the traditional 80% rule of thumb—subject to a fixed FDR requirement. The design of experiments intended to achieve a targeted mix of FDRs and FNRs can be facilitated by the use of pilot studies; see, for example, the discussion in Avivi et al. (2021).

BEYOND TESTING

While the analogy between tail selection and multiple testing can be helpful for building intuition, it is important to remember that the ultimate goal of EB selection exercises is to facilitate decisions that minimize expected losses. As Manski (2021) noted, "Decision theory does not restrict attention to tests that yield a predetermined upper bound on the probability of a Type I error. Nor does it aim to minimize the maximum value of the probability of a Type II error when more than a specified minimum distance from the null hypothesis."

Consider the dialysis centers studied by GK, which they note are assigned letter grades A–F in fixed proportions. It is not obvious which notion of FDR control would be relevant for such a scoring exercise as the decision space of letter grades is multinomial rather than binary. Treating the grade C as the null hypothesis, for example, would raise the

thorny question of how to direct power against each of the alternative grades. Devising a continuous loss function that treats these hypotheses symmetrically, while penalizing larger mistakes more heavily, seems more natural than attempting to squeeze the ranking problem into the Neyman–Pearson straightjacket.

Even in problems featuring a binary selection decision, adhering to a selection rule devised to minimize the FNR subject to an FDR constraint can prove detrimental when loss is continuous in performance levels. Kline, Rose, and Walters (2021) considered the case of a hypothetical auditor seeking to investigate firms that are heavily biased against Black applicants. This example is not (entirely) contrived: the Office of Federal Contract Compliance annually audits thousands of federal contractors, regularly sanctioning firms found to be in violation of equal employment opportunity laws (Maxwell, Moorthy, Francis, and Ellis (2013)). Suppose the auditor can launch as many investigations as desired and has preferences described by the following loss function:

$$L(\boldsymbol{\delta}, \boldsymbol{\theta}) = -\sum_{i=1}^{n} \delta_i \left\{ {\theta_i}^{1-r} - c \right\},$$

where $\theta_i \geq 0$ measures the level of discrimination against Black applicants at firm i, $c > 0$ is the cost of conducting an investigation, and $r < 1$ indexes the auditor's degree of risk aversion.

The auditor would like to investigate all firms with $\theta_i^{1-r} > c$ but must decide which investigations to launch based on her posterior beliefs given the evidence at hand. When $r = 0$, the auditor is risk-neutral and minimizes expected loss by investigating firms with posterior mean estimates of θ_i exceeding c. As $r \to 1$, the loss function approaches $-\sum_{i=1}^{n} \delta_i (1\{\theta_i \geq 0\} - c)$ and the auditor launches investigations only when the posterior probability that $\theta_i > 0$ exceeds c. A Bayesian oracle who follows this posterior selection rule will exhibit FDR control at level $1 - c$ or greater.

Though the risk-neutral auditor is more likely to mistakenly investigate firms that do not discriminate, her investigations should reveal higher average levels of discrimination than a risk-averse auditor who launches the same number of investigations based upon posterior tail probabilities. From this vantage, league table reasoning appears difficult to defend, requiring an extreme form of risk aversion on the part of the decision-maker. While FDRs and FNRs may provide transparent assessments of the expected frequency of ranking mistakes generated by a selection rule, in most settings these rates are unlikely to characterize the optimal decision rule itself.

Similar considerations arise in the domain of education policy. Reassigning students to teachers with higher estimated value added should raise test scores on average even if it lowers the scores of some due to misclassification of teacher ranks. Trading off expected gains against possible losses requires careful contemplation of societal objectives. Prudent decision-making will, in general, tend to require consideration of the *level* of θ_i rather than just its expected rank.

CONCLUSION

As economists gain access to increasingly granular tomes of data, the temptation to score the performance of particular agents and organizations will only intensify. The league table mentality, it seems, is here to stay. Leveraging recent developments in EB estimation, GK offer a sophisticated, yet practical, approach to tail selection that blends control over false discoveries with near-oracle levels of power.

While interest in relative comparisons is inevitable, it is worth reiterating that the goal of large scale selection exercises is typically not to test hypotheses about ranks but rather to make decisions that minimize expected loss. Outside of settings featuring rigid capacity constraints, it is far from clear that perfect knowledge of ranks would enable decisions that are first-best. Indeed, many organizations, not to mention the U.S. legal system, specify absolute standards of behavior which, if found to be violated, will trigger remedial action. Tying selection problems more closely to these absolute standards may serve to simultaneously simplify, and raise the social relevance of, quantitative selection exercises.

REFERENCES

ANSCOMBE, FRANCIS J. (1948): "The Transformation of Poisson, Binomial and Negative-Binomial Data," *Biometrika*, 35 (3/4), 246–254. [48]

AVIVI, HADAR, PATRICK KLINE, EVAN ROSE, AND CHRISTOPHER WALTERS (2021): "Adaptive Correspondence Experiments," *AEA Papers and Proceedings*, 111, 43–48. doi: 10.1257/pandp.20211079. URL https://www.aeaweb.org/articles?id=10.1257/pandp.20211079. [49,50]

BARTLETT, MAURICE S. (1947): "The Use of Transformations," *Biometrics*, 3 (1), 39–52. [48]

EFRON, BRADLEY (2004): "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99 (465), 96–104. [49]

——— (2016): "Empirical Bayes Deconvolution Estimates," *Biometrika*, 103 (1), 1–20. [48]

FISHER, RONALD A. (1922): "On the Interpretation of χ^2 From Contingency Tables, and the Calculation of P," *Journal of the Royal Statistical Society*, 85 (1), 87–94. [50]

GENOVESE, CHRISTOPHER, AND LARRY WASSERMAN (2002): "Operating Characteristics and Extensions of the False Discovery Rate Procedure," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64 (3), 499–517. [49]

GU, JIAYING, AND ROGER KOENKER (2020): "Invidious Comparisons: Ranking and Selection as Compound Decisions," arXiv preprint. arXiv:2012.12550. [47]

KLINE, PATRICK M., AND CHRISTOPHER R. WALTERS (2021): "Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination," *Econometrica*, 89 (2), 765–792. [49,50]

KLINE, PATRICK M., EVAN K. ROSE, AND CHRISTOPHER R. WALTERS (2021): "Systemic Discrimination Among Large us Employers," Technical report, National Bureau of Economic Research. [48,50,51]

KOENKER, ROGER (2020): "Empirical Bayes Confidence Intervals: An R Vinaigrette." [48]

MANSKI, CHARLES F. (2021): "Econometrics for Decision Making: Building Foundations Sketched by Haavelmo and Wald," *Econometrica*, 89 (6), 2827–2853. [50]

MAXWELL, NAN, ARAVIND MOORTHY, CAROLINE MASSAD FRANCIS, DYLAN ELLIS (2013): "Using Administrative Data to Address Federal Contractor Violations of Equal Employment Opportunity Laws," Technical report, Mathematica Policy Research. [51]

NUNLEY, JOHN M., ADAM PUGH, NICHOLAS ROMERO, AND R. ALAN SEALS (2015): "Racial Discrimination in the Labor Market for Recent College Graduates: Evidence From a Field Experiment," *The BE Journal of Economic Analysis & Policy*, 15 (3), 1093–1125. [49,50]

STOREY, JOHN D. (2003): "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value," *The Annals of Statistics*, 31 (6), 2013–2035. [49]

Co-editor Guido Imbens handled this manuscript.

Manuscript received 12 November, 2021; final version accepted 15 January, 2022; available online 16 June, 2022.