Comment on Gu and Koenker, "Invidious Comparisons: Ranking and Selection as Compound Decisions"

Keisuke Hirano *

January 16, 2022

1 Introduction

In their stimulating and insightful paper, Jiaying Gu and Roger Koenker consider the problem of ranking and selecting units, adopting a nonparametric empirical Bayes (NPEB) perspective. Drawing on recent developments (including their own contributions) on empirical Bayes methodology, they illustrate the feasibility of the NPEB approach and its potential to improve upon conventional ranking methods. With the growing availability of large data sets with a hierarchical or grouped structure, and increasing interest in policies that use such data to rank and select units, this is a timely and important contribution. In this comment, I discuss some decision-theoretic aspects of Gu and Koenker's analysis, and the potential to use large-sample distributional approximations within their approach.

2 Statistical Decision Theory Perspective

Gu and Koenker study a type of random effects model where, for $i = 1, \dots, n$,

$$\theta_i \overset{\text{iid}}{\sim} G, \qquad y_i \mid \theta \overset{\text{ind}}{\sim} N(\theta_i, \sigma_i^2).$$

The σ_i^2 are known, but G and $\theta = (\theta_1, \dots, \theta_n)$ are unknown. (They also consider some settings where the σ_i^2 are not known but can be estimated.)

Gu and Koenker express the subset selection problem as a compound decision problem. In their formulation, the action space is taken to be the set $\{0,1\}^n$, and decisions $\delta = (\delta_1, \dots, \delta_n)$ are vectors of binary inclusion choices. They propose a compound loss function that is additive in unit losses

$$\mathbb{1}(\theta_i \geq \theta_\alpha)(1 - \delta_i), \quad \text{where} \quad \theta_\alpha = G^{-1}(1 - \alpha).$$

This penalizes choices where δ_i misclassifies unit i as being in the bottom $(1-\alpha)$ part of the distribution G. The condition that $\lceil \alpha n \rceil$ units are selected is incorporated by introducing the constraint $\sum_{i=1}^n \delta_i =$

*Department of Economics, Pennsylvania State University, 602 Kern Graduate Building, University Park PA 16802. Email: kuh237@psu.edu.

αn, and a desire to minimize false discovery rate is introduced by adding a constraint involving false discovery terms $\mathbb{1}(\theta_i < \theta_\alpha)\delta_i$.

Gu and Koenker's analysis falls somewhere between conventional statistical inference and a full-blown decision-theoretic analysis in the sense of Wald (1950) or Savage (1972).[1] This flexibility in perspective allows them to propose and explore novel ideas. There may be gains, however, from taking on the added burdens of a more complete decision-theoretic perspective. For example, Gu and Koenker set up their decision problem with parameter $\theta = (\theta_1, \dots, \theta_n)$, taking G as "known." The estimation of G is external to the formal decision-theoretic analysis, as is common in empirical Bayes theory. Gu and Koenker show that, provided G is estimated sufficiently well, the discrepancy between the idealized posterior analysis and the empirical Bayes procedure vanishes as $n \to \infty$. This is reassuring, but it does not allow us to make finer distinctions between different (consistent) estimators for G. In principle one could study the complete statistical decision problem that combines the estimation of G and inference about θ, and thereby explore the consequences of alternative methods for estimating G on the quality of the final ranking or selections. Of course, this is far from a trivial task, whether one adopts a subjective Bayesian perspective or a frequentist Wald-type perspective.

In their formulation, the close connection between selection problems and multiple testing emerges from the specification of the loss functions and constraints. There is a potential connection to another literature as well: the decision rules δ, being binary vectors, are equivalent to treatment assignment rules, which have been studied recently by a number of authors including Manski (2004), Dehejia (2005), Stoye (2009), Chamberlain (2011), Kitagawa and Tetenov (2018), and Athey and Wager (2021). For example, Gu and Koenker mention the possibility of weighting misclassifications by the magnitude of θ, through terms of the form

$$(1-\delta_i)\mathbb{1}(\theta_i \geq \theta_\alpha)\theta_i.$$

This is similar to the the asymmetric welfare regret loss analyzed in Tetenov (2012) and Hirano and Porter (2009).

Decision theory generally forces one to be explicit about the objective or loss function. In this regard, Gu and Koenker's exploration of alternative loss functions is useful and revealing. On the one hand, their class of loss functions connects nicely to classic multiple hypothesis testing; on the other hand, the loss functions feel somewhat reverse-engineered for this purpose, rather than being directly motivated by the empirical problem at hand. In practice, ranking and subset selection are often intermediate steps within larger policies (e.g. policies on retention or promotion of workers). For such applications, the decision-theoretic framework offers the possibility to tailor the loss function to the specific policy objective, rather than working with "general-purpose" loss functions (which still require user input, such as the choice of α and the false discovery rate constraint γ). Moving towards more economically motivated loss functions may be an interesting direction for future work.

[1] See Ferguson (1967) and Berger (1993) for overviews of classical statistical decision theory, and Manski (2021) for a discussion of statistical decision theory in econometrics.

3 Asymptotic Approximations

Gu and Koenker regard the normality assumption on y_i as an approximation. For example, in their application to U.S. dialysis centers, they start with a Poisson model for counts, and then take the variance-stabilized transformation of the count to be exactly normally distributed. In many cases, proceeding as if the data are normally distributed may be quite reasonable, but can this appeal to approximate normality be put on firmer analytical grounds?

Suppose for example that for each i, we observe $z_{i1},\ldots,z_{iT} \overset{\text{iid}}{\sim} F_{\theta_i}$, for some parametric family of distributions F_θ. Assume independence between units i, and let $y_i = \hat{\theta}_i$ be the maximum likelihood estimator based on $z_{i1},\ldots,z_{iT}$. Then, under the usual regularity conditions, we would have

$$T^{1/2}\left(y_i - \theta_i\right) \xrightarrow{d} N\left(0, I(\theta_i)^{-1}\right) \quad \text{as } T \to \infty,$$

where $I(\theta)$ is the Fisher information at θ. Let $\hat{I}_i$ be a consistent estimator of $I(\theta_i)$. We could interpret this as saying that the distribution of y_i is approximately normal, centered at θ_i with variance $\sigma_i^2 = \left(T\hat{I}_i\right)^{-1}$. Collecting units $i = 1,\ldots,n$, we would have a joint approximation of $(y_1,\ldots,y_n)$ by a vector of independent but heterogeneous normals with *known* variances.

This classic asymptotic approximation is not completely suitable for a formal decision-theoretic treatment. As $T \to \infty$, one can essentially perfectly rank the units, making the decision problem trivial in the limit. A solution is to model the unit means θ_i as being "local" to each other, and obtain limits of experiments representations in the spirit of Le Cam (1972), van der Vaart (1991), and others. Fix some θ_0 and let unit i's parameter be

$$\theta_i = \theta_0 + \frac{h_i}{\sqrt{T}},$$

where $h = (h_1,\ldots,h_n)$ are the local parameters. Under a local asymptotic normality condition, the original decision problem can be approximated by a decision problem based on observing a vector $(y_1,\ldots,y_n)$, where the y_i are independent with

$$y_i \sim N(h_i, \sigma_i^2),$$

and σ_i^2 equals the inverse of Fisher information for the ith model. Again we are in the known, but possibly heterogeneous variance case. Hirano and Porter (2016) study some treatment choice and shrinkage procedures in a similar setup.

This type of asymptotic approximation differs fundamentally from Gu and Koenker's asymptotic adaptivity theory. Here, $T \to \infty$ but n is held fixed, whereas Gu and Koenker take $n \to \infty$ but maintain the exact normality assumption on y_i. Moreover, the local asymptotics outlined above treat the h_i (and hence the θ_i) as "fixed effects," while in the adaptivity results, the θ_i are i.i.d. draws from some fixed G. Depending on the application, one, both, or neither of these types of asymptotic approximations may be useful. An open question is whether there is a limit experiment approximation as $n, T \to \infty$ jointly.

References

ATHEY, S. AND S. WAGER (2021): "Policy Learning with Observational Data," *Econometrica*, 89, 133–161.

BERGER, J. O. (1993): *Statistical Decision Theory and Bayesian Analysis*, New York: Springer, 2nd ed.

CHAMBERLAIN, G. (2011): "Bayesian Aspects of Treatment Choice," in *The Oxford Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. van Dijk, Oxford University Press, 11–39.

DEHEJIA, R. H. (2005): "Program Evaluation as a Decision Problem," *Journal of Econometrics*, 125, 141–173.

FERGUSON, T. S. (1967): *Mathematical Statistics: A Decision Theoretic Approach*, New York: Academic Press.

HIRANO, K. AND J. R. PORTER (2009): "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77, 1683–1701.

——— (2016): "Panel Asymptotics and Statistical Decision Theory," *The Japanese Economic Review*, 67, 33–49.

KITAGAWA, T. AND A. TETENOV (2018): "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86, 591–616.

LE CAM, L. M. (1972): "Limits of Experiments," in *Proceedings of the Sixth Berkeley Symposium of Mathematical Statistics*, vol. 1, 245–261.

MANSKI, C. F. (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 1221–1246.

——— (2021): "Econometrics for Decision Making: Building Foundations Sketched by Haavelmo and Wald," *Econometrica*, 89, 2827–2853.

SAVAGE, L. J. (1972): *The Foundations of Statistics*, New York: Dover, 2nd ed.

STOYE, J. (2009): "Minimax Regret Treatment Choice With Finite Samples," *Journal of Econometrics*, 151, 70–81.

TETENOV, A. (2012): "Statistical Treatment Choice Based on Asymmetric Minmax Regret Criteria," *Journal of Econometrics*, 166, 157–165.

VAN DER VAART, A. W. (1991): "An Asymptotic Representation Theorem," *International Statistical Review*, 59, 99–121.

WALD, A. (1950): *Statistical Decision Functions*, New York: Wiley.