Comment by Bertrand S. Clarke[5] and Gregory E. Holt[5]

We argue that the authors' focus on nonparametric Bayes estimation, despite being well executed, has led them to neglect the topic of nonparametric Bayes testing – a topic many non-statisticians think is just as important as estimation. Leaving aside whether estimation or testing is more important, our point here is that the arguments in favor of NPB from a testing perspective appear to have been neglected in general. As noted by Tokdar et al. (2010) 'The Bayesian literature on these testing problems is still rather meagre, unlike the case of nonparametric estimation...' Despite Borgwardt and Ghahramani (2009) and Holmes et al. (2012) our literature search did not turn up much evidence to invalidate this observation. So, let us give a class of settings where NPB hypothesis testing is likely to be better than parametric Bayes testing or Frequentist testing. We will focus on testing the equality of two distributions.

Consider the following thought experiment. A scientist is interested in conducting a clinical trial enrolling patients with end stage cancer who are otherwise out of treatment options. Despite the need for comparisons to placebo based control groups, clinical trialists realize patients do not enroll in studies where they may receive a placebo and therefore most of these trials remain uncontrolled. Researchers often rely on historical controls despite their known deficiencies.

As an alternative, to study therapeutic modalities in patients with terminal diseases, researchers could enroll patients only seen in clinic on one defined day while creating a control group formed from patients satisfying the same inclusion/exclusion criteria but seen on an alternative clinic day. We refer to this sort of control group as 'virtual' since it is constructed artificially after the treatment group is enrolled. The dependence between the treatment group and the virtual control group only comes from the inclusion/exclusion criteria and from matching the distribution of the baseline variables (described below). Such virtual control groups should exhibit the same outcome variable, here overall survival denoted Y, and Y should be a function of the baseline variables for both the treatment and virtual control groups. In this procedure, the virtual control group corresponds to patients receiving standard of care therapy so any differences between treatment and control would suggest a treatment effect.

Although placebo controlled randomized trials would still be preferable, in settings involving patients who typically avoid placebo controlled trials, this clinical trial design may permit better comparisons than historical controls that do not take into account current treatment practices or characteristics of the local population and treating physicians. In these contexts, NPB testing of the equality of the distribution of the baseline variables would be a better way to verify that a candidate virtual control group will provide a suitable comparison for a treatment group than Frequentist or parametric Bayes testing would be. At root, this follows because Bayes testing is better than Frequentist testing, see Berger and Bayarri (2004), Berger (2003), and M. Eaton (2013) among others, and nonparametric testing is more flexible than parametric testing.

To set up this testing problem, let us assume that all patients seen by a physician

[5] Department of Medicine, University of Miami, Miami, FL

on a day of experimental enrollment (say Tuesday) or on a day of virtual control group formation (say Thursday) have had the same basline tests. Now, in principle, we can compare the baselines of the patients in the Tuesday group with a collection of Thursday patients that we can use to form a virtual control group. More formally, suppose the baseline measurements for the treatment group are represented as $\mathbf{X} = (X_1, \ldots, X_K)^T$ and we have n outcomes $\mathcal{D}_T = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. To form a 'virtual control group' let $\mathbf{X}'$ be the same variables as $\mathbf{X}$ but measured on the Thursday patients and let $\mathcal{D}_C = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_n\}$ be the resulting set of baseline measurements. The question is how to choose $\mathcal{D}_C$ so that we can compare the corresponding $Y_1, \ldots, Y_n$ from the treatment group with the $Y'_1, \ldots, Y'_n$ from the control group.

One way to formulate this is as a hypothesis test. Let P be the distribution of $\mathbf{X}$ and let Q be the distribution of $\mathbf{X}$. We want to test

$$\mathcal{H}_0 : P \neq Q \quad \text{vs.} \quad \mathcal{H}_1 : P = Q. \tag{1}$$

It is natural to use the Bayesian formulation. Foundationally, Bayesian techniques are not probabilistic in the data on which one conditions, see Chen (1985) Sec. 3.1. Specifically, the conditioning data need only form a well-defined deterministic sequence. So, it is legitimate to search the Thursday patients to find the ones that will give a $\mathcal{D}_C$ that lets us reject $\mathcal{H}_0$, i.e., mimics $\mathcal{D}_T$ well enough that the posterior probability of the null is small enough.

The NPB solution is clear: Find a nonparametric prior distribution for the pair (P, Q), for instance a bivariate DP as described in Walker and Muliere (2003) or a bivariate MDP as in the present paper. Now, reinterpeting (1) as

$$\mathcal{H}_0^* : d(P, Q) \geq \epsilon \quad \text{vs.} \quad \mathcal{H}_1^* : d(P, Q) < \epsilon, \tag{2}$$

for some distance d and writing the prior as W, the Bayes test is based on

$$\frac{W(d(P, Q) \leq \epsilon | \mathcal{D})}{W(d(P, Q) > \epsilon | \mathcal{D})} \tag{3}$$

where $\mathcal{D} = \mathcal{D}_T \cup \mathcal{D}_C$. If (3) is large enough then we are led to accept the alternative in (2) and therefore use $\mathcal{D}_C$ as a 'virtual control group' for inference on Y and Y'.

What would the nonparametric Frequentist solution be? First, (2) would be harder to test than (1), so let us focus on (1). Frequentist Neyman-Pearson testing treats the hypotheses asymmetrically and familiar two sample forms of tests such as Kolmogorov-Smirnov, the Anderson-Darling test, and the Cramer-von Mises test treat $\mathcal{H}_1$ vs. $\mathcal{H}_0$, the reverse of (1). To adapt such a test statistic to our present case requires that the null be decomposed into a series of nulls that can be tested separately and then put together by some kind of multiple comparisons procedure. That is, write

$$\{P \neq Q\} = \cup_{j=1}^J B((P_j, Q_j), \eta) \cup S \tag{4}$$

where $B((P_j, Q_j), \eta)$ is a collection of balls of radius $\eta > 0$ and $S = [\cup_{j=1}^J B((P_j, Q_j), \eta)]^c$ is a set of pairs of distributions deemed to be so far from the 'line' of distributions $P = Q$

that they can be ignored. Now it is enough to consider the J composite vs. composite tests $\mathcal{H}_{0,j} : (P, Q) \in B((P_j, Q_j), \eta)$ vs. $\mathcal{H}_1 : P = Q$. However, if η is small enough then

$$\mathcal{H}_{0,j} : (P, Q) \in B((P_j, Q_j), \eta) \approx \mathcal{H}^*_{0,j} : (P, Q) = (P_j, Q_j),$$

and for each j we can reduce $\mathcal{H}_1$ to $\mathcal{H}_{1,j} : (\tilde{P}_j, \tilde{Q}_j) = \arg\min_{P=Q} d((P_j, Q_j), (P, Q))$. So, to test (1), it is approximately enough to do the J simple vs. simple tests

$$\mathcal{H}^*_{0,j} : (P, Q) = (P_j, Q_j) \quad \text{vs.} \quad \mathcal{H}_{1,j} : (\tilde{P}_j, \tilde{Q}_j).$$

Now, if we can reject in all J tests under a multiple comparisons procedure we have a Frequentist test of (1). If we can't reject all J nulls, problems remain. Overall, in contrast to (3), Frequentist reasoning is too precious to be disturbed by refutation.

The Frequentist parametric approach will reduce J and so be simpler than the Frequentist nonparametric approach – at the cost of specifying a parametric family. The Bayes parametric approach is likewise simpler than the NPB approach but also has the cost of specifying a parametric family. Neither parametric reduction is persuasive.

Thus, the NPB prescription for finding a virtual control group is to find sets $\mathcal{D}_C$ that let us reject in (1) or (2). This is easier to implement and interpret than a Frequentist analysis and should also give better results – as Bayes tests commonly do.

References

Berger, J. (2003). "Could Fisher, Jerffreys, and Neyman have agreed on testing?" *Statistical Science*, 18: 1–32.

Berger, J. and Bayarri, S. (2004). "The interplay of Bayesian and Frequentist analysis." *Statistical Science*, 19: 58–80.

Borgwardt, K. and Ghahramani, Z. (2009). "Bayesian two-sample tests."
URL arXiv:0906.4032[cs.LG]

Chen, C.-F. (1985). "On asymptotic normality of limiting density functions with Bayesian implications." *Journal of the Royal Statistical Society Series B*, 47: 540–546.

Holmes, C., Caron, F., Griffin, J., and Stephens, D. (2012). "Two-sample Bayes nonparametric hypotheis tests."
URL arXiv:0910.5060v2[stat.ME]

M. Eaton, A. S., R. Muirhead (2013). "On the limiting behavior of the probability of claiming superiority in a Bayesian context." *Bayesian Analysis*, 8: 221–232.

Tokdar, S., Chakrabarti, A., and Ghosh, J. (2010). "Bayesian nonparametric goodness of fit tests." In Sun, M. C. D. D. P. M. D. and Ye, K. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, chapter 6.1. Springer.

Walker, S. and Muliere, P. (2003). "A bivariate Dirichlet process." *Statistics and Probability Letters*, 64: 1–7.