Journal: Political Analysis
MS: PA 70-188, revised for resubmission
Title: Matching as Nonparametric Preprocessing...

Revise-and-resubmit is usually an easy matter, but I still cannot unreservedly recommend publication of this ms. The revised ms is an improvement on the previous version, but some of the problems I had with the original paper remain unanswered by the authors. I have some specific criticisms, and one or two more major criticisms (to do with the notion of model dependence and its role in the current paper). I know the paper on this topic I'd like to see, but I don't think thats going to emerge from another round of revision-and-resubmission; thus, in the spirit of trying to bring the R-and-R cycle to a close at some point, I'd be inclined to sign off on a further revision that at least dealt with the point-by-point items I list below.

The concept of model dependence remains elusive, and, at least in my mind, an odd way to motivate this paper. The authors want us to do a simple thing: pre-process the data via some kind of matching procedure, and then run the same parametric analysis/analyses we would ordinarily run. Why should we do this? The answer, apparently, is that after the pre-processing via matching, we will produce estimates of causal effects that are less "model dependent". But, since when did this become the relevant criterion? What about bias? What about mean square error? Check the index of any major statistics/econometrics textbook for "model dependence". My cynical interpretation is that the focus on model dependence is there because its easy to demonstrate (e.g., Figures 2 and 4), while something like the MSE of matching-based estimators is much harder to characterize (i.e., witness the dearth of such results in statistics and econometrics).

One virtue of what the authors propose is that by staying in the parametric world, you may be well be hitting a "sweet spot" in terms of MSE (I wish we knew for sure, or could know). That is, matching reduces bias in estimates of causal effects, and, under circumstances that are not clear, maybe also reduce variance. If, after matching, I were to estimate causal effects non-parametrically, I may well be overshooting on further bias corrections at the expense of variance. On the other hand, "simple" parametric analysis is almost always "spartan" relative to non-parametric modeling, and if matching buys relatively low bias, maybe thats enough, and a "simple", parametric 2nd stage analysis could be enough (and – I think this is important – say what you like about a parametric analysis, but at least we think we know how to compute a standard error when we run a regression; contrast the matching literature *per se*). Now sure, the virtues of a matching-then-parametric analysis have a "it depends" quality, and oh how I wish the authors could say more on this. Instead, we're shown graphs demonstrating less "model dependence" after matching, and I'm still not entirely sure how to evaluate that result. I want "good" (low MSE) estimates of causal effects. Who wouldn't? How does low/less "model dependence" map onto that?

If model dependence is to remain the stalking horse (and, ok, the Carpenter example

is nothing if not vivid on this point), what I'd really like to see is a *proof* of the following general proposition: that point estimates of ATT based on data with good balance display less variation across all possible subsets of models than those based on all possible subsets of the raw data. That is, did the authors just "get lucky" with the Carpenter example (picking a data set with horrible balance, from the looks of it), or, under what conditions have they got a general result?

Of course, stating a general result would first require a *definition* of model dependence. On p9 of the ms the authors promise a "more formal definition of model dependence" but I don't see it. At the top p11, there is a quote from King and Zeng (2006b) that "defines"

> model dependence at point x as the difference, or distance, between the predicted outcome values from any two *plausible* alternative models... By 'plausible' alternative models, we mean models that fit the data reasonably well and, in particular, they fit about equally well around either the 'center' of the data (such as a multivariate mean or median) or the center of a sufficiently large cluster of data nearest the counterfactual x of interest.

This is hardly a "formal definition", and introduces all kinds of other concepts that are not well defined: e.g., "plausible alternative models", "fitting the data reasonably well", "a sufficiently large cluster of data", a "point" x, a "counterfactual" x. My point here is that if "minimizing model dependence" is the goal, then this needs to be *clearly and rigorously* defined. The current "definition" is simply too loose, and too vague. I see what the authors mean, but please tighten this up.

Formal definitions aside, the key point about "model dependence" does actually appear in the paper, but too late. I wonder if this formulation might be helpful to the authors, and needs to come earlier: if T and X can be made to be more or less uncorrelated, then the "risk" of omitted variable bias disappears, meaning that estimates about the effect of T will display less sensitivity as various combinations of X enter or exit any model(s) the researcher may care to estimate. Now this is a reasonably simple point, and any *PA* reader who has come through a typical PhD level quantitative methods sequence will follow. So I wonder about moving this take on "model dependence" forward, or some version/elaboration of it.

I have a number of specific queries and criticisms that I list below. The common thread is that the paper is too breezy. We're being asked to buy into an entire "inferential framework", a two-stage approach to estimating treatment effects. But aside from the (standard) definitions of treatment effects, the paper is an essentially an essay; there is not a single proposition or theorem stated in the entire paper. Re-engineering the paper to the level of the *Annals of Statistics* is not what I have in mind; but surely the authors can do a little better than what we have now. The ubiquitous bias-variance tradeoff crops up throughout the paper (and, most ominously, towards the end in section 8 on pp26-7); can the authors precisely tell us the conditions or provide an example as to when matching might make us

worse off on MSE?

Indeed, on p27 we have the provocative concession:

> In particular, the methodological literature offers no formal estimates of mean square error and so in marginal cases it can be difficult to know whether or how much preprocessing will help.

Wow. A less charitable reviewer might well stop there: i.e., here is a methodological procedure that works when it works, but doesn't when it doesn't, but we really don't know where the point of demarcation is, or even how to assess where that point is; oh, and by the way, by pre-processing the authors mean, "use your 'substantive knowledge' and try a whole bunch of matching procedures, picking the one that produces the best QQ plot". But hey, I'm not that reviewer... The point is that the ms is intended, presumably, to be a contribution to the "methodological literature". So, can't we say something general/formal about the conditions under which the procedures described here will outperform (some) alternatives? The discussion on the lower half of p13 is a promising start, if highly qualified.

Point-by-point criticisms/queries:

1. "young literature", Abstract and p1. Matching is not as old as, say, MLE, but is it "young"? Cochran was writing about this stuff a long time ago now.

2. p10, "...researchers...do not often go very far in portraying the sensitivity of their causal inferences to model specification, and conveying all the sensitivity is essentially impossible. Indeed, attempt to do in many situations lead to nihilistic conclusions." Cites? Examples?

3. Equation 11, p11; does this need to hold for all i?

4. p11 "The simplest (although not necessarily the best) way to understand how we can satisfy (11)....is via *one-to-one exact matching*." Why is this not necessarily "the best", whatever "best" might mean?

5. pp11-12 "The idea is to match each treated unit with one control unit....Our preprocessed dataset this is the same as the original dataset with any unmatched control units discarded..." What about unmatched treated units?

6. p12 "..our second stage parametric analysis" Surely you mean "the *researcher's* second stage parametric analysis", since "you" (the authors) provide a menu of techniques for matching, leaving the rest to the researcher.

7. p12, "For the same reason scholars rarely use an unadjusted difference in means to analyze randomized data..." Which scholars? What's the basis for this characterization? I'd actually disagree with the characterization.

8. p12, "...we recommend that scholars make use of our decades of experience with parametric models..." Again, surely "you" mean "...researchers' decades of experience...". The sentence reads presumptuously otherwise, but perhaps that is what the authors intended...!

9. p13, "In matching, efficiency gains occur when dropping data if heterogeneity is reduced." "Heterogeneity" in what, with respect to what? In Y, right? But we're not looking at Y when we match, so hence its an "if" as to whether "matching reduces heterogeneity"?

10. p13, "..and as a result helps parametric models fit better" How? See previous point re what is "heterogeneity".

11. p13, "If matching reduces heterogeneity, then σ^2 will become smaller". See previous point re what is "heterogeneity".

12. p13, highly qualified nature of the claim in the bottom of the page:

    > Thus, the resulting variance for the treatment effect *may* become smaller even if some observations are dropped. This *suggests* that so long as one keeps as many observations as possible, matching *will often* reduce bias and improve the efficiency of subsequent parametric analyses. (emphasis added).

    Can you state something stronger than this? That is, *when*, exactly, will matching reduce variance, or (even better) reduce MSE?

13. p14 I'd rename section 6 "How to Match", and write it much more tightly. The earlier part of the paper sets us up for matching as a solution to the problem of estimating causal effects; now, in section 6, we (finally!) learn "how to match." This section makes repeated reference to "steps" (e.g., p15, "we may skip the remaining steps"); as I asked in my original review, is there a recipe/algorithm the authors have in mind here? Could it be more clearly laid out. Like the matching literature itself, this section is perhaps the most confusing section of the present paper, a long laundry list of things one might do, or should do, and a thinly-veiled shot at Sekhon and Diamond on p18 (the "balance test fallacy"), but not clear specific recipes for the applied researcher.

    Jumping ahead, I note that the examples are actually a little more enlightening on this: e.g., in the Carpenter example, no control units lie in the convex hull of the covariate values observed among the treated, and then the authors resort to other matching methods. Some elaboration (and/or formalization) of the choices one makes when trying to match would be especially useful, since these choices seem important, but not automated (?) and open to lots of "judgment calls" by analysts.

14. p14, "The goal of matching is to improve *balance*... without losing too many observations in the process." Again, this raises the bias-variance issue. At this point it seems that matching is all about eliminating bias, possibly at the cost of variance (by eliminating cases). Which raises several questions: (1) is there any way to match with a view to minimizing "downstream" MSE?, or would that involve looking at Y, which we're not allowed to do when we match; (2) my earlier question comes back into play, i.e., where do we come out on MSE if matching is relentlessly geared towards bias minimization via balance. In short, although matching is considered "optimal" when it produces "balance", if the end goal is to "accurately" estimate causal effects shouldn't we be thinking about MSE of causal effects as we do the matching?

15. p17, does Mahalanobis distance need to be defined for the *PA* readership?

16. p19, "If meeting these criteria for balance..." Ok, but *which* criteria? Look at QQ plots, sure. But what, exactly, are the criteria for "balance" the authors would have us adopt?

17. Figure 2. The point estimates of the causal effects from the matched data display less variation across all 262,143 specifications. But so what? Remind me why, in light of Figure 2, I should now believe that the "true effect" is closer to -35 than -50? That is, make the jump from "less model dependent" to "less biased".

18. p26, "If we drop many observations during matching, do not in the process reduce heterogeneity, ... the mean square error... might actually increase." Please tighten this up. Heterogeneity in what, and how defined? Mean square error of what (estimated causal effects, I know, but the readers might not)?