Review of:

"Matching as Nonparametric Preprocessing for Reducing Model Dependence
in Parametric Causal Inference," Political Analysis Manuscript

****

The authors are talented and productive researchers who often produce innovative and influential research. Unfortunately, the current manuscript is unacceptable for publication. The manuscript raises more questions than it answers and is often erroneous. The authors have failed to respond in a meaningful way to my previous criticisms. Here is a point-by-point reply to the authors' reply note. Additional comments follow:

**
"1. Post-treatment bias. "
**

Originally I noted that the Rubin causal model is a radical departure from current practice. In their paper the authors repeatedly note that what they are doing is easily adopted by current researchers. Their abstract says that after matching researchers can "then...apply whatever familiar parametric techniques they would have used anyway." This is incorrect. As they concede in their response memo, "Our recommendation only applies to causal models that are valid in this [Rubin causal model] sense." Therefore, researchers cannot simply continue to do what they were doing, because the Rubin causal model is far from the norm in the social sciences. The paper's repeated claim of ease of adoption and consistency with current practice is profoundly wrong, and the statement in the response memo indicates that the authors know this.

As I noted before, they had to change Carpenter's model before they were willing to estimate it. This is something that most researchers would have to do. In current practice, researchers estimate kitchen sink models with k causal covariates of interest and t controls, where k is almost always far larger than 1. Researchers ignore issues related to post-treatment bias, in part because what is post-treatment for one causal variable of interest may not be for another. This is current practice. A world in which researchers estimate a separate model for each causal covariate of interest is a sharp departure from the status quo. And any such change raises numerous issues which the authors do not discuss.

For example: should researchers do a separate matching exercise for each causal covariate of interest? Assume we are estimating ATT. If so, the causal estimate for the first covariate may not be comparable to the second because the distribution of X covariates for both treatment groups will not in general be the same.

Semiparametric estimation is much closer to current research norms. But the authors claim in the conclusion that:

"Although when used properly each of these [semi-parametric] approaches reduce model dependence, matching is simpler to use and understand, and would work without modification to improve all the parametric models now used in the social sciences."

They claim that their methods work "without MODIFICATION to improve all the parametric models now used in the social sciences." (emphasis added). But in the response memo the authors agree that they had to change models so as to be consistent with the Rubin causal model. The Rubin model is a sharp departure from current practice.

This point comes home in point 9 of their response memo:

"A key point in our paper is to take seriously current practices, to change them relatively little (so it is easy for scholars to adopt our approach), but to offer ways of greatly improving how these techniques work (via preprocessing)." And then point 10 of the response memo:

"Semiparametric methods do not seem to be well known by many social scientists. The vast majority of articles in our journals use parametric regression. Perhaps this should change."

Changing to semiparametric methods is vastly easier than switching to the Rubin causal model. Going from least squares to GAM is a trivial change. Going from current practice of kitchen sink models to a world where one is careful about post-treatment bias is a huge change. It's not just about functional specification. It reaches to thinking differently about research design. If the purpose of this paper is not to rely on functional forms, then GAMs/kernels are an easier change than switching to matching.

In short, the authors' proposal should be measured against the current solution in the literature to the problem they raise--i.e., if functional form is the question, then try out semiparametric alternatives.

**

"2. The contributions of the paper."

**

Post-matching bias adjustment is not a new idea and the authors do not discuss well the pros and cons of such adjustments. The balance tests fallacy argument is a new one, but they don't offer much of a defense of it but cite a brief working paper instead.

There is a section and footnote (#2) where the authors cite some previous work. But they don't cite enough of it, and the authors try to claim credit for something that is not their idea. For example, they fail to cite Rubin's 1973 article entitled "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies" (Biometrics 29: 185--203). In that article Rubin says, "the combination of regression adjustment in matched samples generally produces the least biased estimate." The authors should also note Cochran's seminal work on the subject.

The balance hypothesis test argument in the paper is indeed a new one, but it is not adequately defended or supported but instead they offer a citation to a working paper which is the subject of controversy.

**
"3. Connections to missing data and ecological inference."
**

This writeup is better than it was, but it is at the very least incomplete. They should write a separate paper because they may have an interesting idea here. But there are too many complicated issues to address in a brief appendix.

In ecological inference when the bounds are not informative, the inference problem is indeed very difficult. In matching, correcting for observables is possible when the data are good enough. Maybe this (good data for matching) is the equivalent of the situation when the bounds are informative in ecological inference? That is an interesting idea, and it would be interesting if they flushed it out more.

But the authors' claim is more general. They note that causal inference as an "especially difficult case of ecological inference". Is this true for well done field experiments? I don't think so.

**
"4. Random vs. fixed causal effects."
**

The paper's concept of "random causal effects" is too vague. Instead of language like "average over repeated hypothetical draws" (page 3), which seems to be alluding to superpopulation sampling theory, there should be a clearer and more focused discussion using concepts like iid. Their language invites them to step into doing averages over different realized observations when it is not clear whether such averages are relevant.

Moreover, they largely ignore my point regarding valid SEs. They instead make assertions about how using the parametric models we currently use we can still obtain valid SEs post matching. I do not believe this. I don't see the argument. It is not enough to simply assert that the parametric models often used already condition on X. Yes, but which X? And what if we estimate a logit to get pscores to do the matching? Where does the uncertainty in the predicted probabilities in the pscore model go? If it is ignored, then why isn't this just as bad as ignoring the uncertainty in first stage estimates in 2SLS?

May be I'm being too much of an economist, but I just don't see how this proposed preprocessing step gets around this problem. Traditional econometric techniques have to deal with this problem.

**
"5. Doubly robust."
**

"In addition, the reviewer gives an example, stating "Imagine that the parametric model is correct. We are coming from the opposite perspective, where we are not willing to assume the model is correct."

But the literature on doubly robust inference allows one to assume that either of the two models is correct. So I don't understand the authors' response. Matching could drop the observations needed to identify a nonlinear functional form or simply drop observations that would make the nonlinear functional form change shape. The authors in the response memo go further than the text of their paper. In the paper they only concede "ruling out extreme cases where all data points are dropped during matching stage even though a correct parametric model is specified" (page 13). But in the response memo they concede that dropping the observations needed to identify a model would be a problem. I find this slippage disturbing. In any case, neither concession goes far enough. A polynomial may radically change shape post-matching due to a change of, for example, inliers (which are often just as dangerous as outliers).

And if the model is not correct, why use it at all? Is it locally correct only? How does this work? How local is local? Please discuss.

More generally, in the world of finite data: if matching is imperfect and the model is incorrect even locally, then the bias adjustment will make the situation worse and not better. For example, going back to an old example discussed by Cochran, it has been known that that regression adjustment done for control and treatment groups separately may be preferable to the same adjustment done together (this occurs when the mapping between a confounder x1 and y is not parallel in both treatment and control groups). So if matching is not perfect and the model not locally correct, we are in worse shape than if we only did matching.

**
"7. Balance test fallacy. We agree with the reviewer that we need to check the balance beyond the first moment. The revised manuscript includes this point"
**

This is better, but the authors need to defend their argument much better. Currently they mainly cite a working paper. But the working paper is brief, leaves many issue unaddressed and has been challenged. Particularly important is the issue that people do matching to minimize the bias of a causal estimate. They are only indirectly interested in the gap in covariates. The relationship between these two quantities is not straightforward, and I imagine this has implications for what measures of balance we should use.

**
"8. Differences between matching and TSLS. While both are two-step estimation procedures, matching and TSLS rely on different assumptions. Matching controls for potential confounding variables by matching on them, while TSLS controls for them by finding an instrument that is independent of them."
**

Yes, I realize this. But the question remains: how do I take into account the uncertainty in two step procedures? As I noted above, if we estimate a logit to get pscores to do the matching, where does the uncertainty in the predicted probabilities in the pscore model go? If it is ignored, then why isn't this just the same as ignoring the uncertainty in first stage estimates in 2SLS?

**
"9. Sensitivity analysis. If balance isn't perfect after matching, then checking the degree of model dependence after matching is well worthwhile, we agree. The examples we include in the paper provide some ways of doing sensitivity analyses to verify how much model dependence is left."
**

Since this is the ultimate focus of the paper, the authors' discussion of this issue is far too limited.

**
"10. Semiparametric methods. Semiparametric methods do not seem to be well known by many social scientists. The vast majority of articles in our journals use parametric regression. Perhaps this should change."
**

Again, such methods are easier to adopt than the Rubin causal model. I have long seen semiparametric models used in the literature. Perhaps not often enough, but sometimes. Until recently, I never saw matching. More social scientists know about semiparametric methods than matching. The scientific issue is that the authors main claim is that preprocessing limits model dependence. The preprocessing carries with it a radical change to current social science practice (the Rubin causal model). Semiparametric models are a far smaller change. So they are the natural comparison. Without such a comparison, it is impossible to evaluate the authors' suggestion.

**
Additional Comments:
**

In Figure 1 there is a small thing that to me speaks volumes. In the plot on the right ("After Matching"), how come they draw the fitted line to span the entire range of X? The figure should make it clear that in the After Matching world, the estimation has nothing to say outside of the span of X among the treated. But then, in fact, the estimation has nothing to say about any values for which there is no matched value of X. I.e., we get a finite set of causal effect estimates. Also see my inliers point above. "Don't extrapolate" is a good point, well established long before this paper came along.

Much of the text in the paper is far too imprecise given the subject matter and filled with issues both small and large. It is difficult for a reviewer to discuss all of these issues. For example the key paragraph that starts at the top of page 5 is highly problematic.

"These conditional ATE and ATT quantities are reasonable alternatives that we often use."

"Alternatives"? The other choice is what, not to use any of the observations on y?

"In addition, to change sample to population causal effects for the ATE, we can average it over the population distribution of Xi, and for the ATT we average it over the conditional distribution of Xi given Ti = 1."

They don't mean formally integrating out X, so I'm at a loss to know what they mean by "average it over the population distribution of Xi."

"Point estimates for the population and sample are normally identical, although the variances for the population estimates would be larger."

If Xi is formally integrated out, then point estimates will typically differ. But that's not what they mean. So, I don't understand what they mean.

"If instead one or more of the three conditions do not hold, then Xi may be omitted without any resulting bias (although the variance may increase)."

I hope that this sentence is a typo because the authors have it exactly backwards.

"Since maximum likelihood is invariant to reparameterization---meaning, for example, that the maximum likelihood estimate (MLE) of alpha is the same as the square root of the MLE of alpha^2"

No it isn't---e.g., positive or negative square root?

I could go on like this for the rest of the paper, but I think I've made my case that this paper should not be published. I'm happy to extend my remarks if the editor wishes me to.