Distance Simulations

Rachael Caelie (Rocky) Aikens

7/11/2019

Set Up

As a different measure of performance, we chose to consider the distance between matched pairs in the propensity-by-prognostic feature reduction space shown in figure 1 of the manuscript. For this analysis, we simulated data sets of varying size and measured the mean Mahalanobis distance (in the prognosis-propensity feature space) between matched observations for Buffalo, Propensity, and Mahalanobis Distance matching. The generative model for all of our simulations is the following:

$$
\begin{aligned}
X_i &\sim_{iid} \text{Normal}(0, I_p), \\
T_i &\sim_{iid} \text{Bernoulli}\left(\frac{1}{1+\exp(-\phi(X_i))}\right), \\
Y_i &= \tau T_i + \Psi(X_i) + \epsilon_i, \\
\epsilon_i &\sim_{iid} N(0, \sigma^2),
\end{aligned}
$$

where the true propensity and prognositic scores are given by the linear combinations

$$
\begin{aligned}
\phi(X_i) &= X_{i1}/3 - c, \\
\Psi(X_i) &= \rho X_{i1} + \sqrt{(1-\rho^2)} X_{i2},
\end{aligned}
$$

so that $\text{Cor}(\phi(X_i), \Psi(X_i)) \propto \rho$. The constant c in the propensity score formula was chosen such that there were approximately 100 treated observations in each dataset. We consider $p = 10$. Unlike in the set-up for other simulations, we consider here only $\rho = 0.5$ and $k = 3$. Each simulation consisted of a dataset of size $n = 2000, 1600, 1000$ and was repeated $N = 1000$ times. We fix the treatment effect to be constant with $\tau = 1$ and the noise to be $\sigma = 1$. For a given matching, we estimate ATT and design sensitivity $\tilde{\Gamma}$ using the permutation t-statistic from the package `sensitivtymv`.

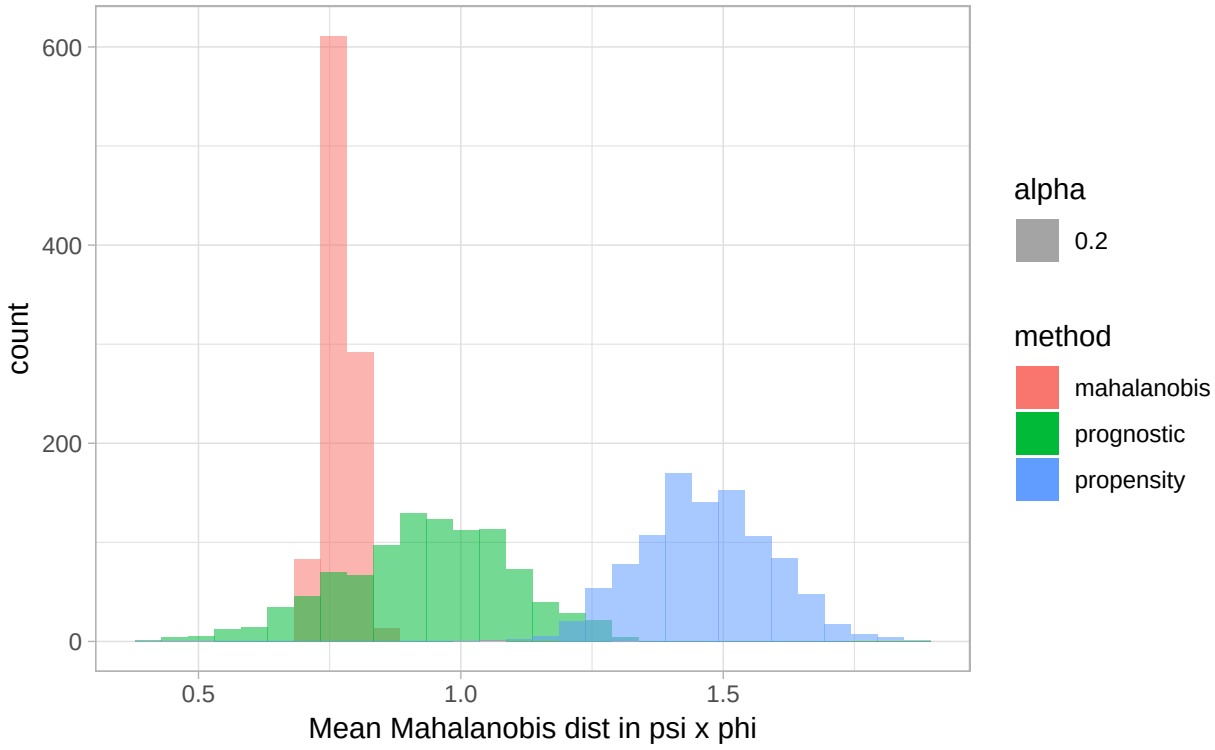
Results

Main Simulations: Mahalanobis distance in Propensity x Prognosis

First, I measured the average mahalanobis distance between treated and control individuals in the reduced feature space of prognostic score by propensity score. I was expecting that prognostic methods would perform better when sample size was large, and mahalanobis methods would perform better when sample size was small. Surprisingly, I found that Mahalanobis was almost always the most effective. Below is a table of the mean distance between matched pairs in terms of the true prognosis and log propensity.

n	mahalanobis	prognostic	propensity
2000	0.7699326	0.9328629	1.464651
1800	0.7787157	0.9381703	1.466500
1600	0.7917392	0.9231895	1.455737
1400	0.8063296	0.9353546	1.464461
1200	0.8215251	0.9449819	1.469888
1000	0.8421539	0.9558340	1.476926

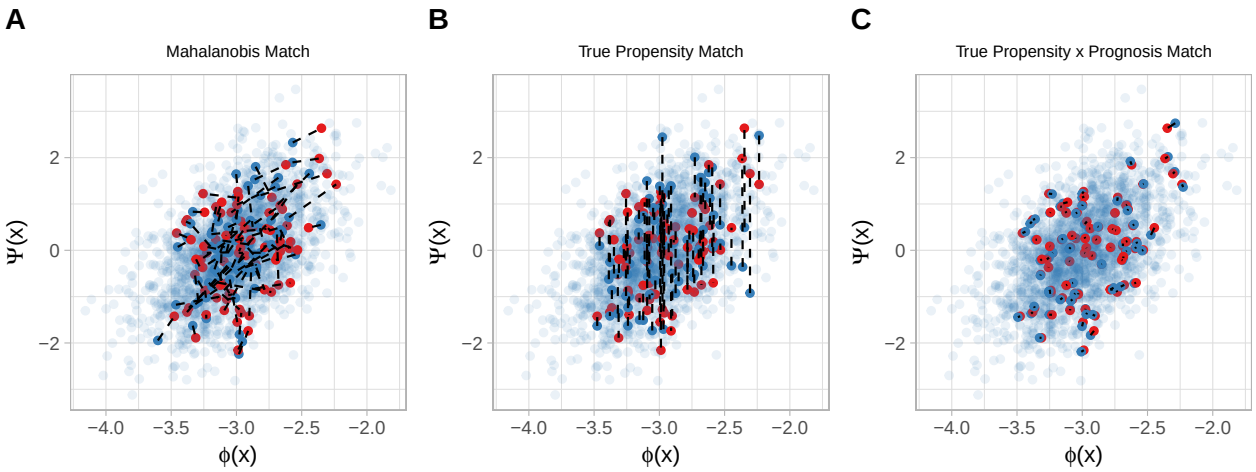
Here, you can see a histogram of the mean distances across simulations. Mahalanobis consistently has a lower mean distance.



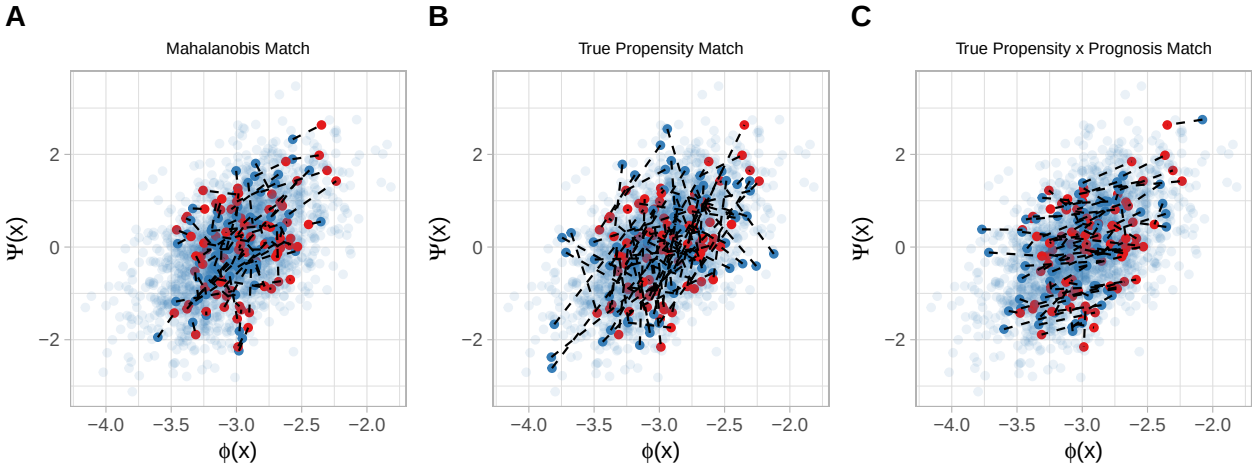
Potential Explanation? Model fitting

It may be that one of the reasons Buffalo and propensity score matching are selecting more distant matches on average is that the prognostic and propensity models are difficult to fit. For example, we can see in simulation that a lot of the drop in performance is eliminated when you assume the true model is given.

Below, we show Figure 1 of the Buffalo manuscript, where we find optimal matchings given the true scores:



In contrast, here are the matches that we might pick if we had to estimate the propensity and prognostic models ourselves rather than relying on an oracle.



The table below shows the avrage distances between matched pairs assuming (Oracle) the true propensity and prognostic scores are provided by an oracle, and (Estimated) the propensity and prognostic scores are estimated.

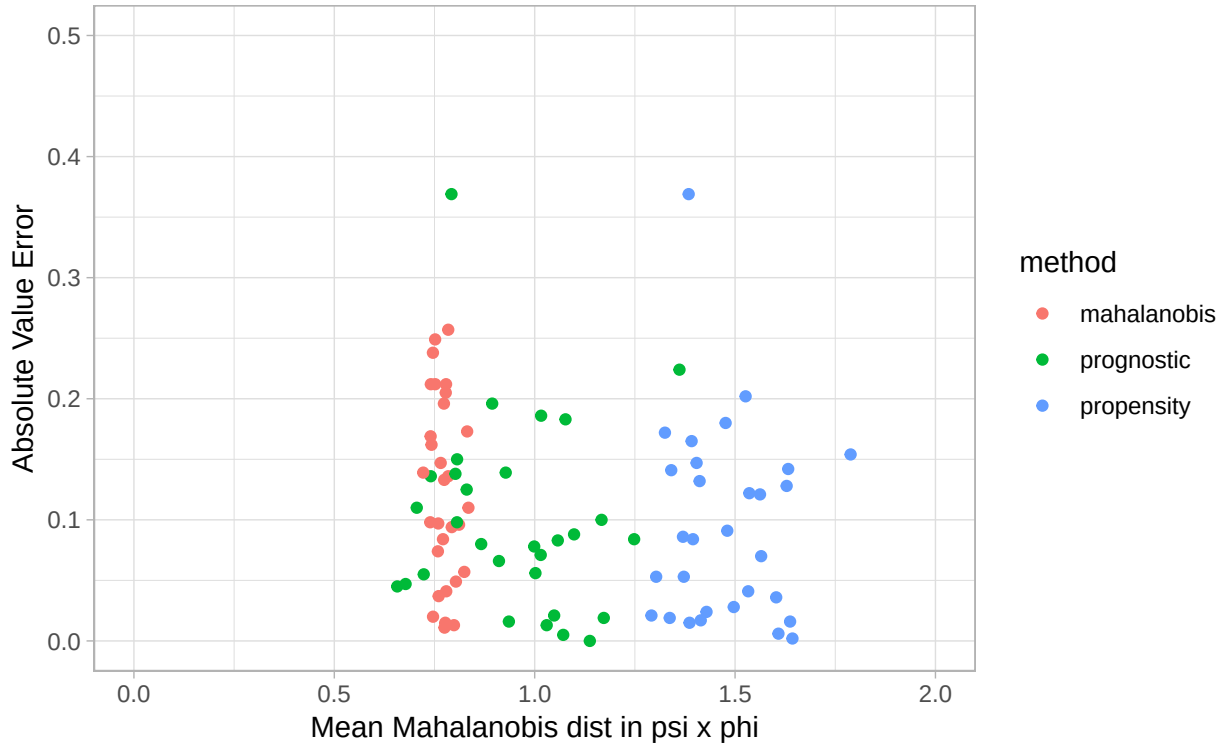
	Mahalanobis	Propensity	Buffalo
Oracle	0.69	1.05	0.05
Estimated	0.69	1.79	0.87

Smaller Simulations

I tried to better understand this with smaller simulations. From this point onward, all results are the product of N=30 simulations.

Estimates

As a sanity check, I looked at the correspondence between mean distance and estimate for a smaller batch of simulations with the same parameters. They are not closely correlated. Prognostic methods tend to be closer to the correct estimate than Mahalanobis, in spite of the fact that Mahalanobis matches are closer in the reduced feature space



Estimated distances

As a further sanity check, I wanted to make sure propensity and prognostic models were performing well on the estimated propensity and prognostic scores. They are; buffalo matching at least thinks it's doing a great job.

method	Mean Estimated Distance
mahalanobis	0.7752796
prognostic	0.0953247
propensity	1.1270248

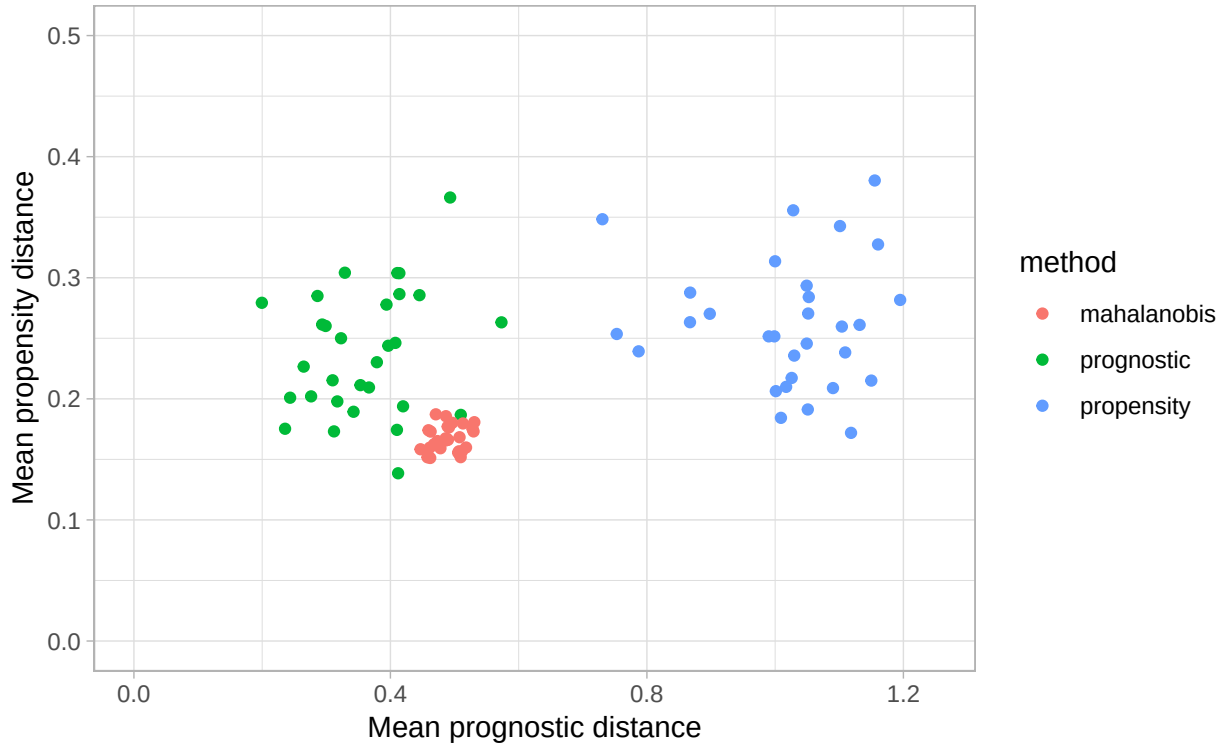
Absolute propensity and prognostic differences

Next, I asked what the distances were between matches sets just in terms of log propensity or just in terms of prognosis.

method	Prognostic distance	Log propensity distance
mahalanobis	0.4888172	0.1669668
prognostic	0.3610627	0.2380657
propensity	1.0190568	0.2620084

What's most surprising to me here is that propensity score matching actually does worst at minimizing log propensity distances. What?

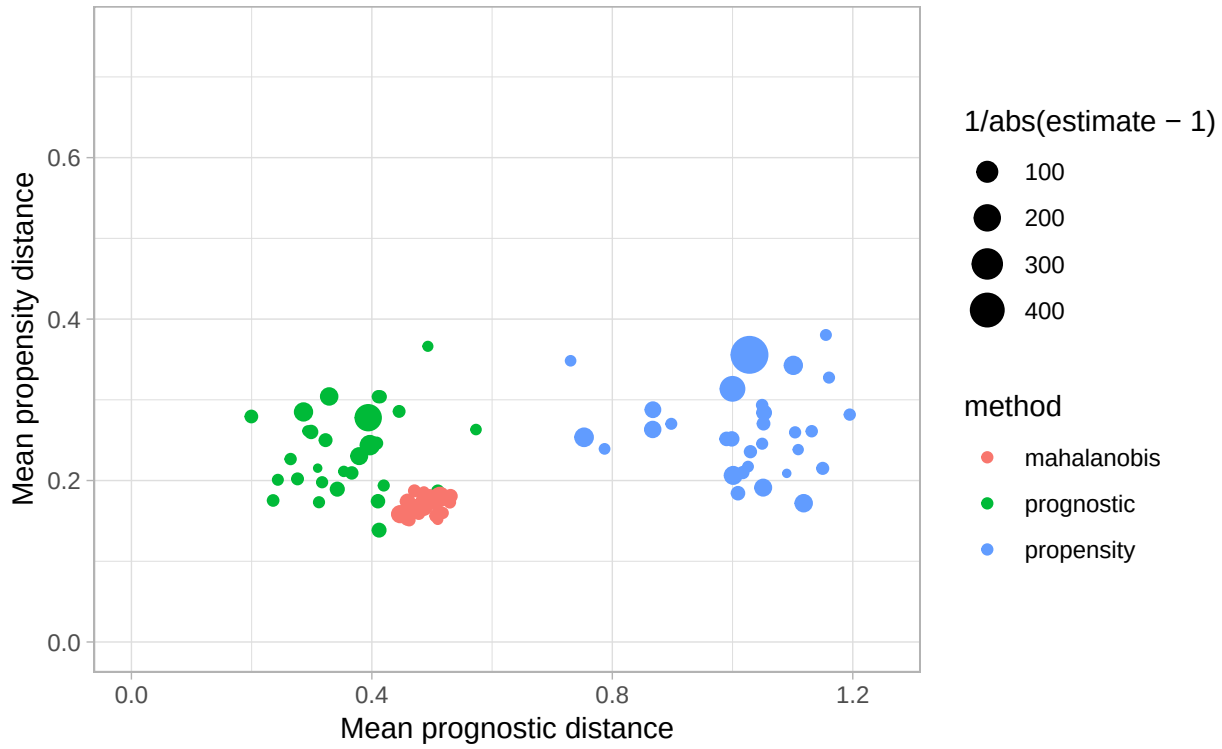
I can plot these distances together like so:



A couple observations here:

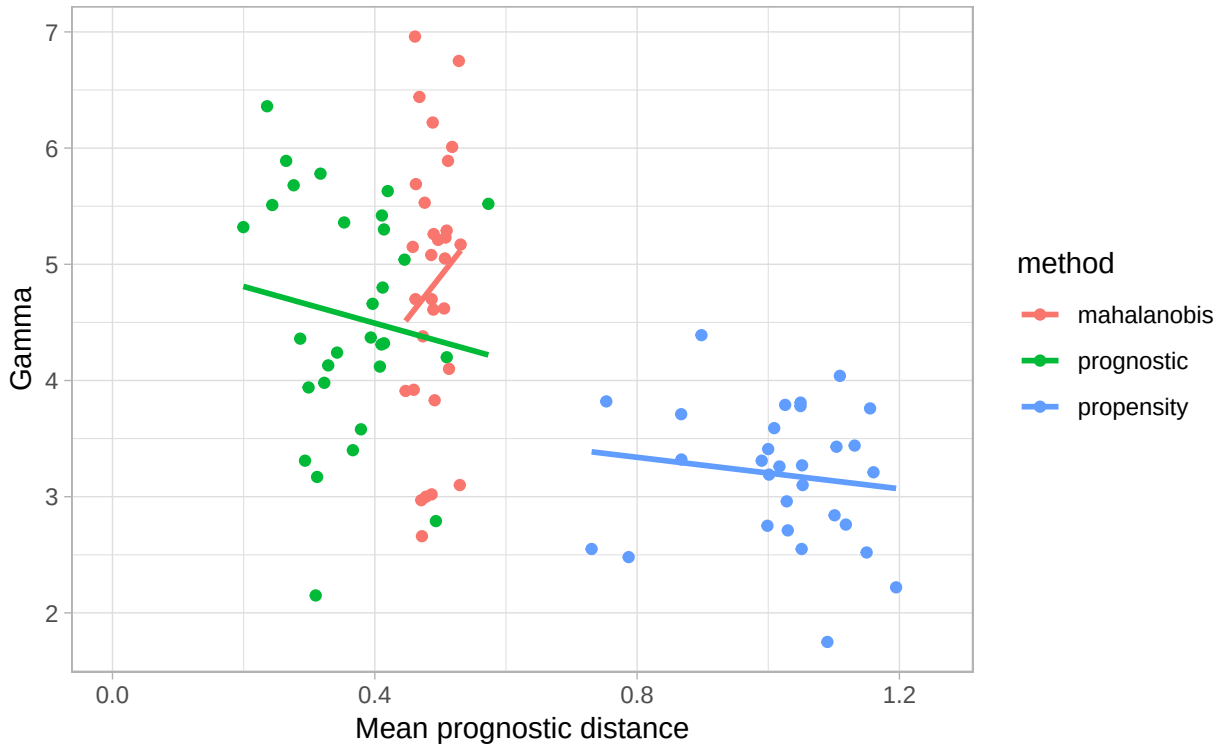
- Buffalo matching does better than all other methods at minimizing prognostic differences.
- Mahalanobis is most consistent.
- Propensity score matching is doing far worse than I would expect at minimizing propensity score distances.

Below, I set the size of the points so that estimates closer to the true effect are larger. Somehow propensity score matching does well even though the distances between matched pairs are the worst.

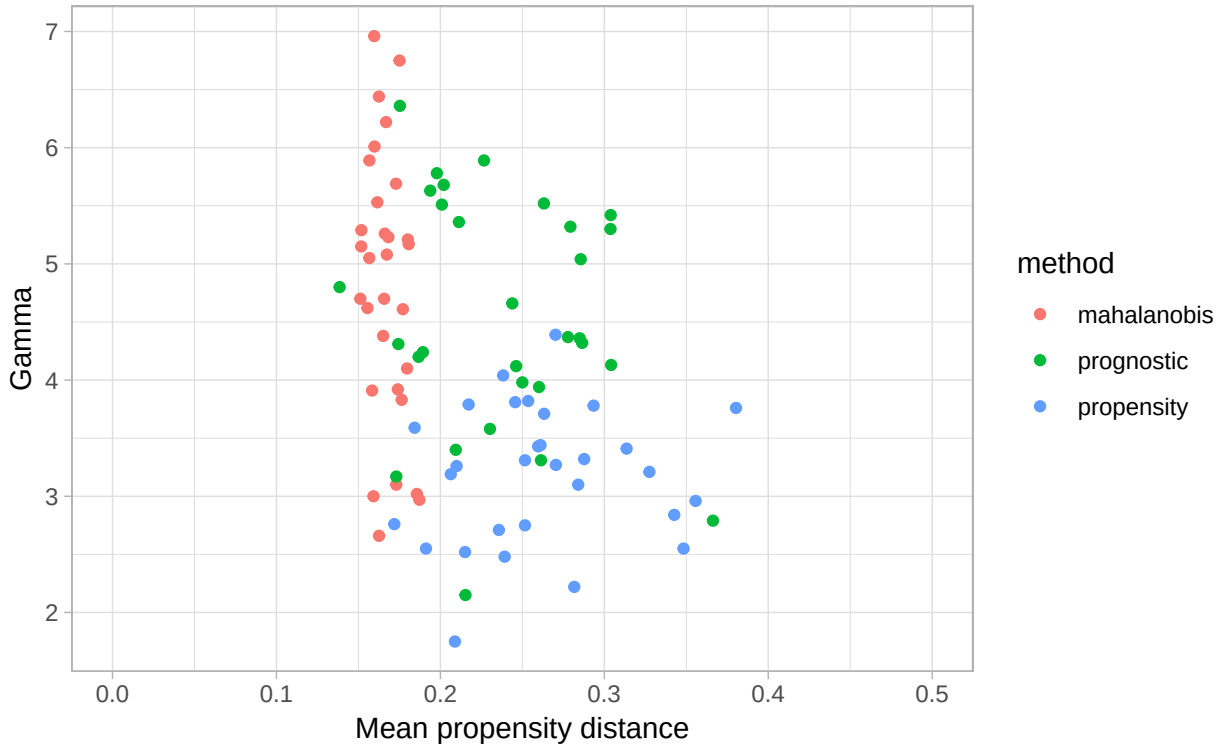


Gamma

I also ran simulations which measured gamma design sensitivity. The plots below show propensity and prognostic distances versus gamma. I overlayed a linear model on this so that you can see the trends between prognostic balance and gamma. I hypothesize that the positive correllation between gamma and prognostic imbalance for mahalanobis distance reflects the fact that mahalanobis matchings give biased estimates, which increase gamma in an unconstructive manner; increasingly poor matches give increasing bias which gives increasing gamma.



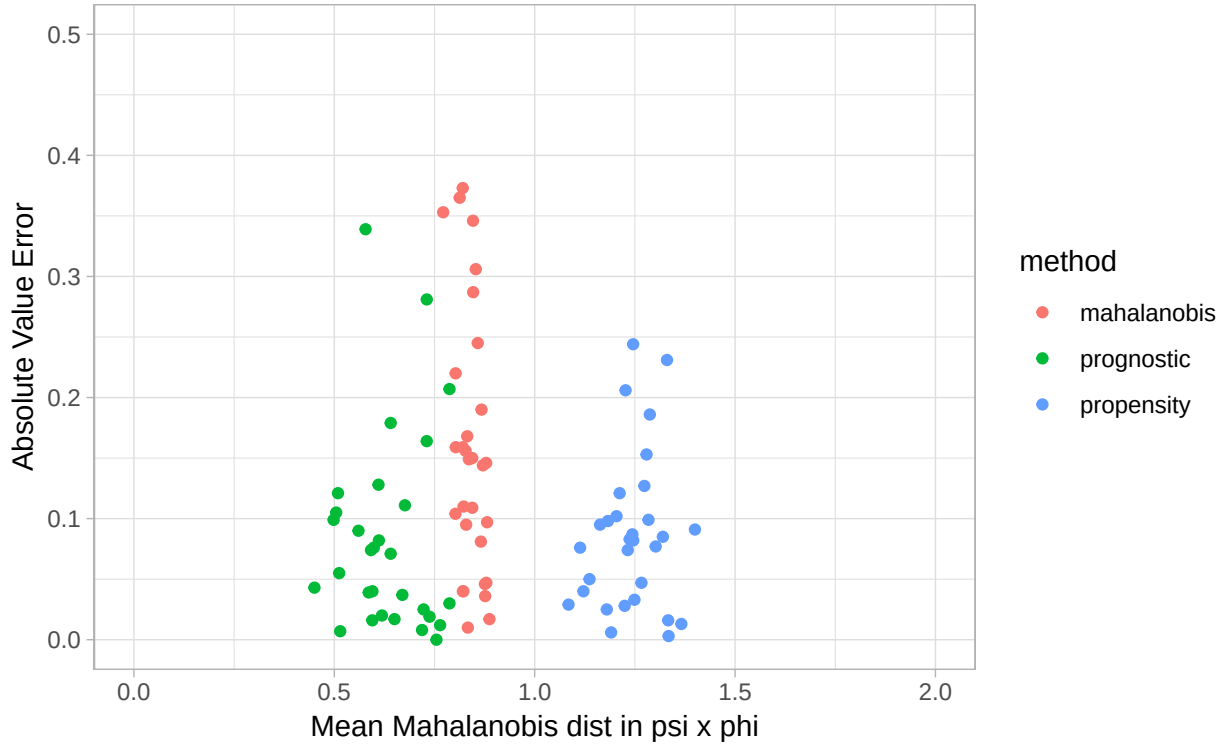
Below is the same chart but with propensity distance on the x axis.

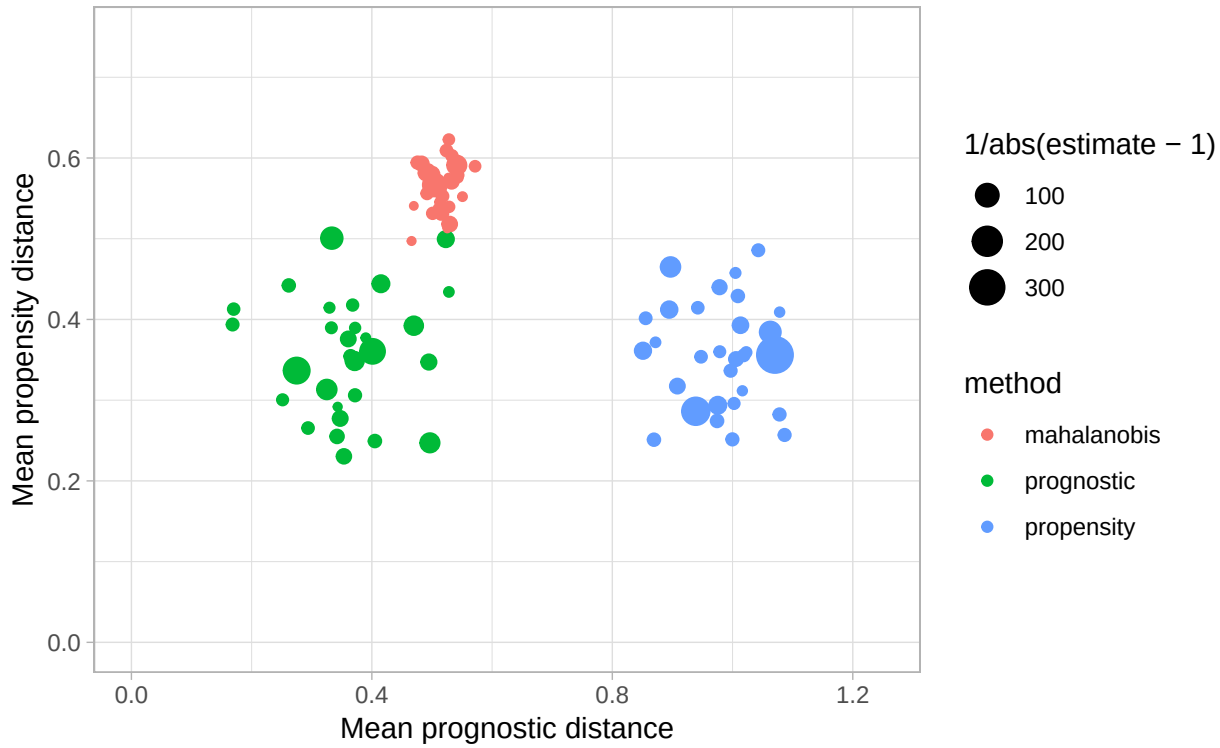


Tinkering with simulation parameters

Less randomness in treatment assignment

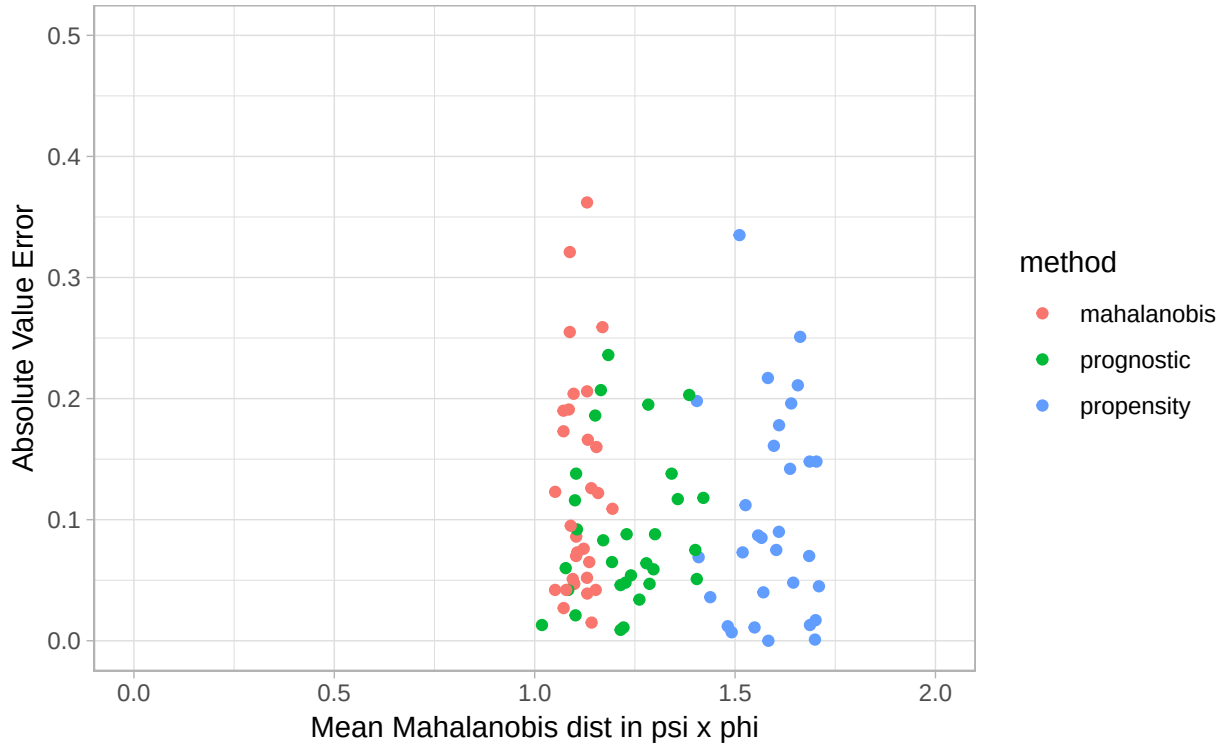
For reasons I don't understand, running the simulations with true log propensity "X1-10/3" rather than "X1/3-3" changes things. Below, we see that buffalo matching is now finding better matches than mahalanobis. This is especially interesting because we found that, in terms of MSE, Buffalo matching doesn't do as well with log propensity "X1-10/3". In essense, switching to log propensity "X1-10/3" means that treatment assignment is more systematic and less random than in our usual simulations. Below are some of the same diagnostic plots as above, with the altered simulation parameters.

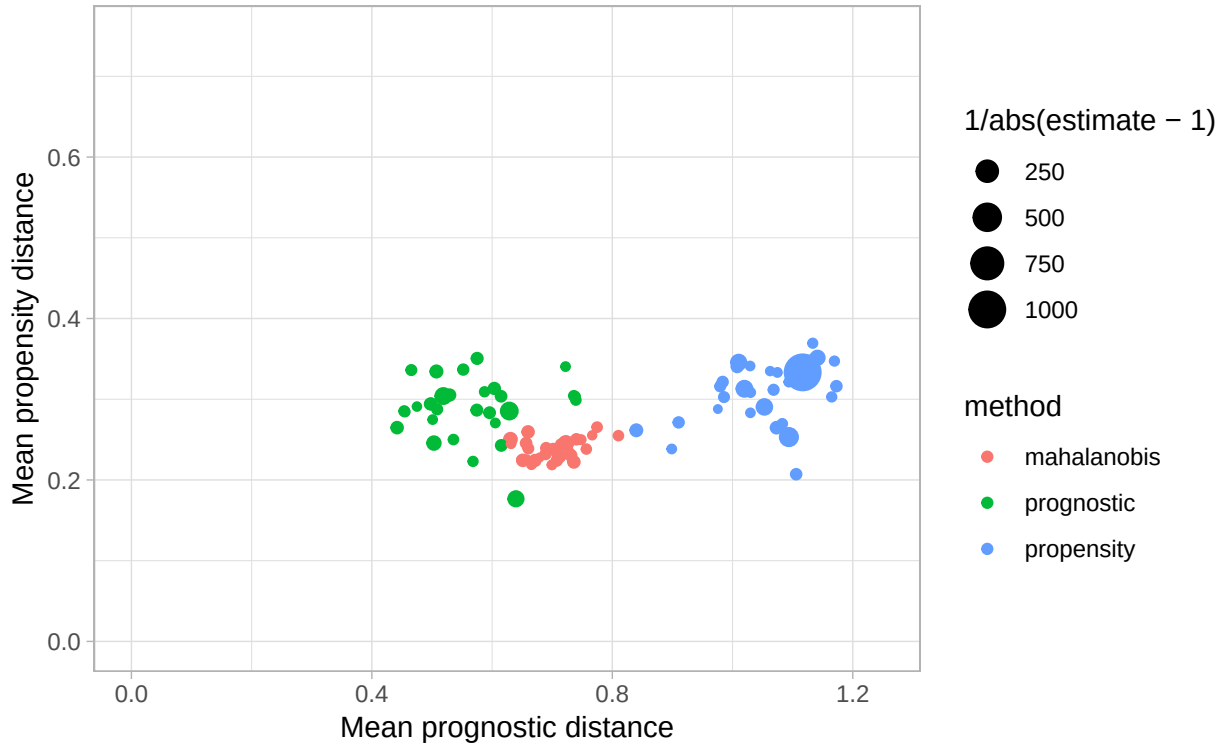




More noise variables

I also reran the simulations with p = 50 features instead of p = 10. Increasing the number of features makes things more difficult for all methods. Mahalanobis has more noise covariates to match on, but propensity and prognostic methods have to deal with a more complex feature space reduction. As p increases, the models we fit for propensity and prognosis become increasingly overspecified.





Future Directions

What this may be getting at is the larger question: "What defines a good matched set?" All three methods considered here work very well under the assumption that a perfect match can be found for each treated individual. In reality, matches are almost always "wrong" by some amount. The question is: how do we judge between imperfect matches to select the one most likely to give us a correct answer?

Here are some things we can consider trying:

- Restrict the problem to 1:1 matching for simplicity.
- Fit propensity and prognostic scores with a lasso to try and reduce the problem of model fitting.
- Make a literature search to see who else has thought about this problem. People to check: Sam Pimentel, Jas Sekhon, Kosuke Imai, Jose Zubizarreta, Ben Hansen