Bias Variance Tuning

Rachael Caelie (Rocky) Aikens

5/1/2019

Set Up

We compare the performance of propensity score matching, Mahalanobis distance matching, and Buffalo Matching (described in the previous section) on simulated data, varying the dimensionality of the problem, the fixed treatment to control ratio during matching, and the correlation between the true propensity and prognostic score. 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

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

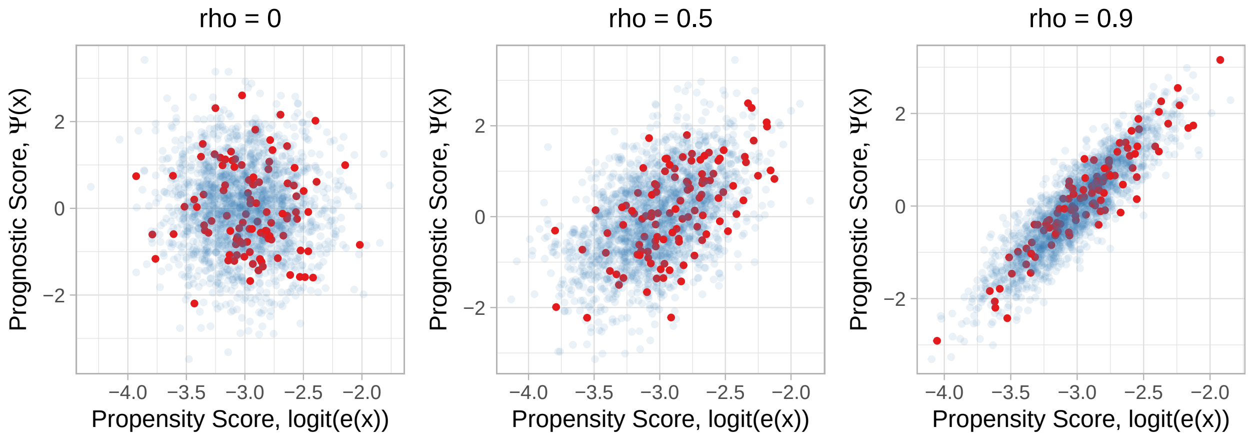
so that $\text{Cor}(\phi(X_i), \Psi(X_i)) \propto \rho$. The constant, c in the propensity score formula was chosen such that there were approximately 100 treated observations in each dataset. For the simulations reported in the main figures of the paper, we let $c = 3$. We consider $p = 10$, $\rho = 0, 0.1, \ldots, 0.9, 1.0$, and $k = 1, \ldots, 10$. Each simulation consisted of a dataset of size $n = 2000$ 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`

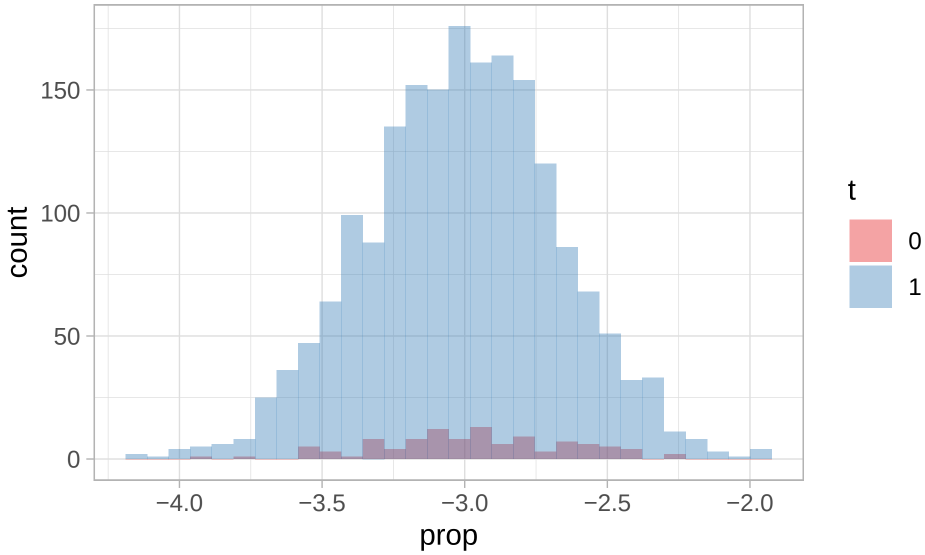
Primary simulations from paper

Just to be explicit, the primary simulations in our paper rely on the following formulations of ϕ and ψ.

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

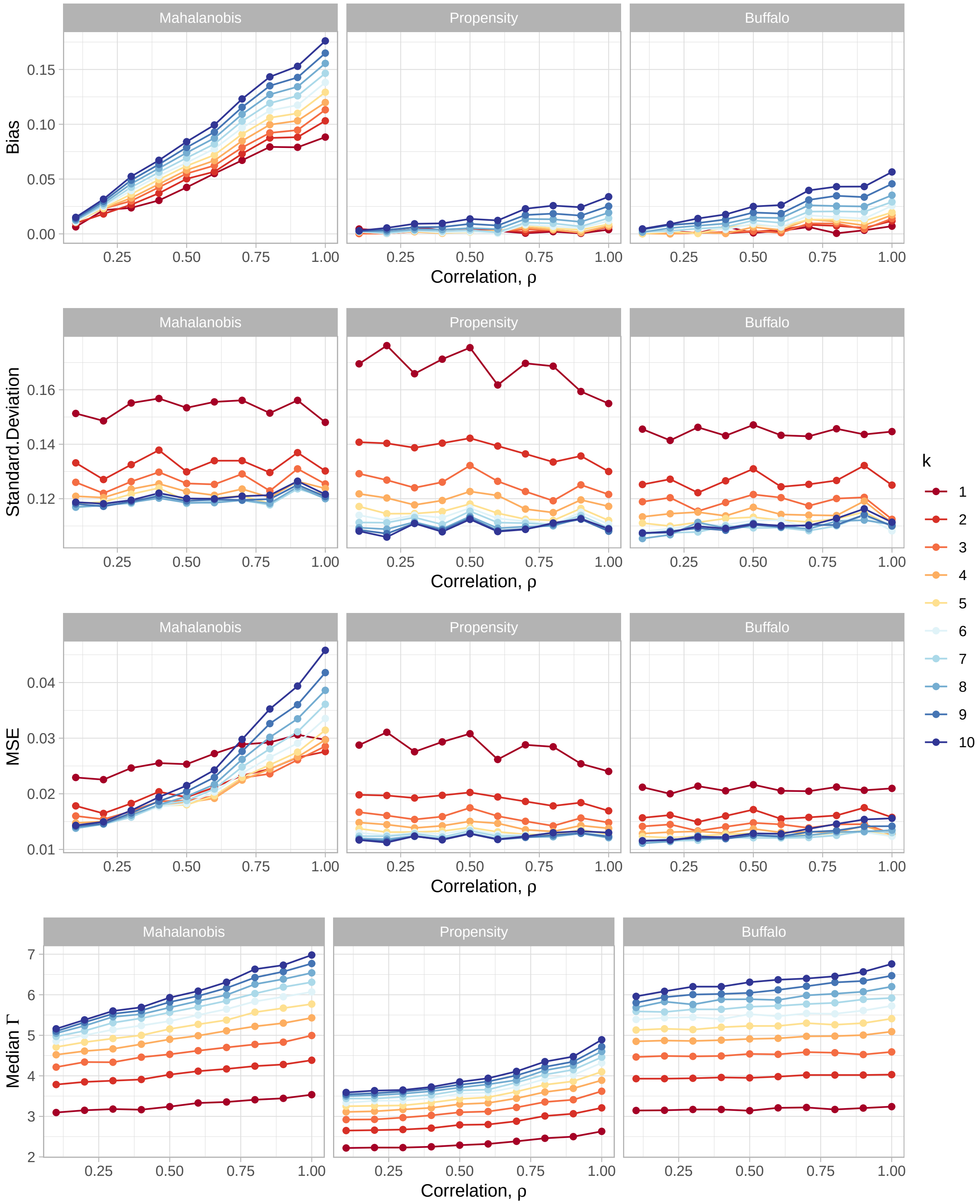
A valid question to ask is: what is the overlap between treated and control individuals in this set-up? We can interrogate that with Fisher-Mill plots and histograms





Inessence, overlap is quite good in this scenario.

For reference, the performance of each method under this formulation is given below. These are essentially figures 2 and 3 of the existing paper:

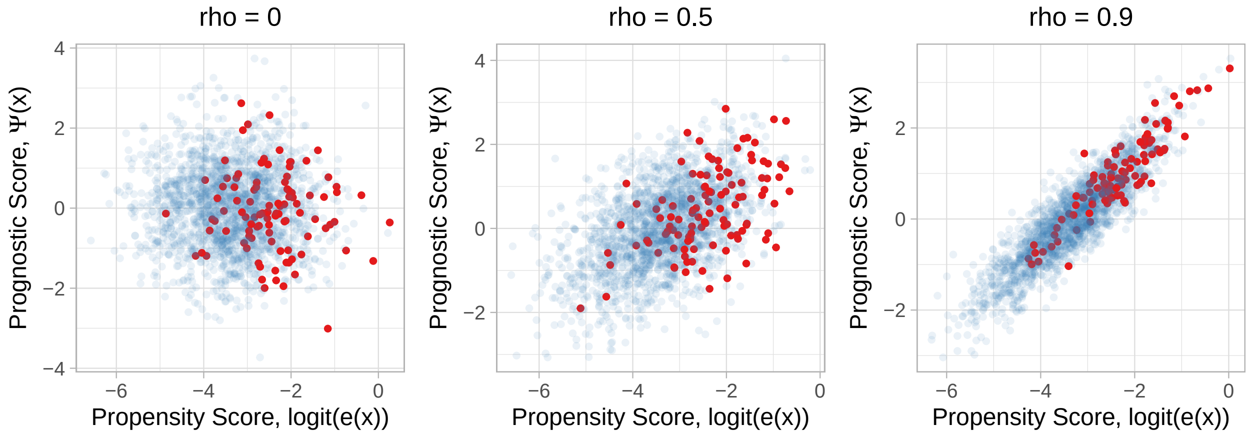

Mahalanobis
Propensity
Buffalo
Bias
0.15
0.10
0.05
0.00
0.25
0.50
0.75
1.00
Correlation, ρ
Standard.Deviation
0.16
0.14
0.12
MSE
0.04
0.03
0.02
0.01
Median Γ
7
6
5
4
3
2
k
1
2
3
4
5
6
7
8
9
10

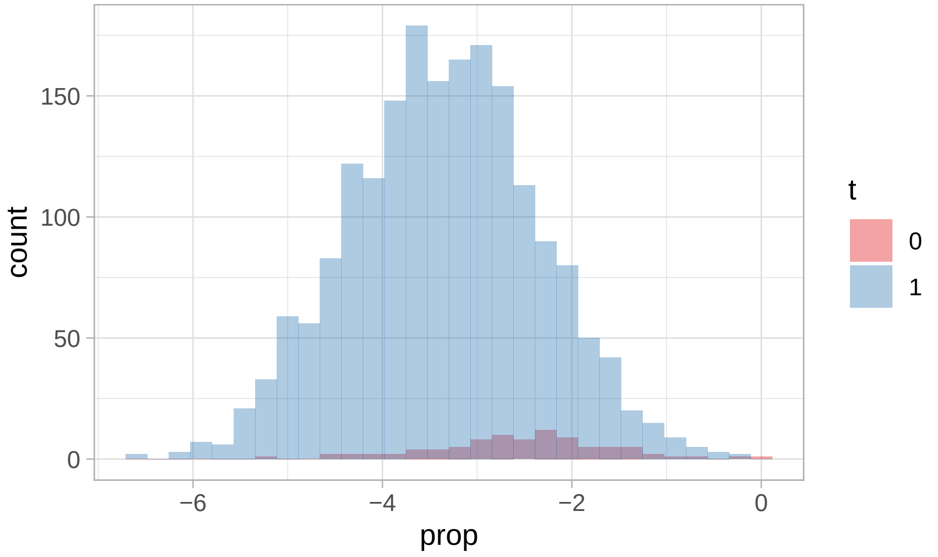
Phi = X1 - 10/3

As it turns out, we've already run a batch of simulations with worse overlap, and the standard deviations for each method under these simulation parameters is already in Supplementary figure 2. Inessence, all that was changed is that the weight of the covariate X_1 in the propensity formula was increased, meaning that the baseline covariates more strongly determine treatment effect:

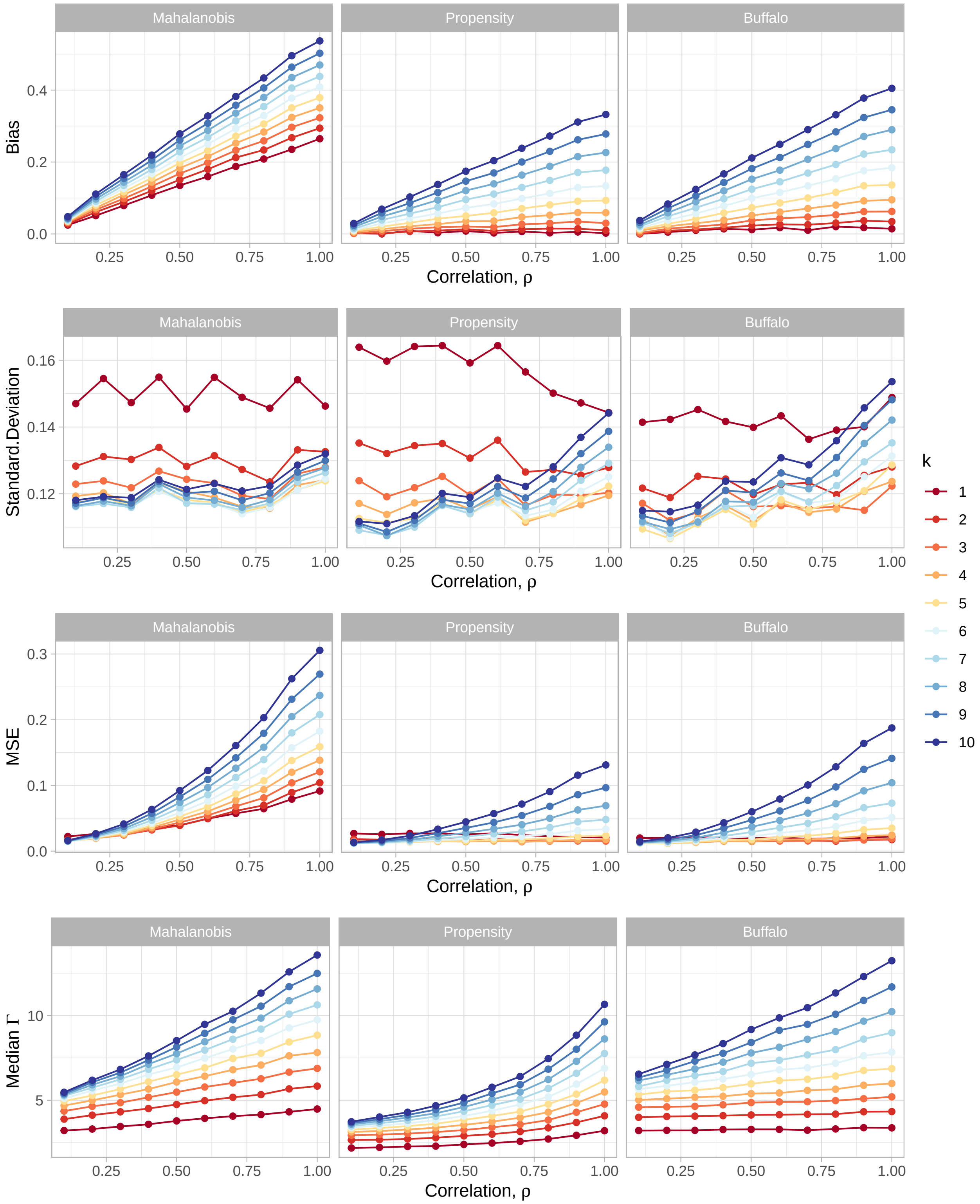
$$\phi(X_i) = X_{i1} - 10/3,$$
$$\Psi(X_i) = \rho X_{i1} + \sqrt{(1-\rho^2)}X_{i2},$$

Note that the constant in the propensity score formula was increased slightly to ensure that there were still approximately 100 treated individuals in each simulated data set. The FM plots and overlap histograms for some sample data sets are shown below:





And here we see the performance:



I make the following observations:

1. *Bias is much higher when rho is large* - bias goes up to 0.4 when rho is large, whereas in the main simulation bias never rose above 0.2, and was even lower (0.05) for propensity and joint matching. I suppose this is because when the overlap is worse the matches to the treatment group are worse in terms of propensity score (and prognostic score). This gives more severe bias, especially when there is

more confounding in the data set.

2. *Standard deviation gets much worse for larger rho and k* - we mention this in the main text, and speculate that when prognosis and treatment are highly correllated and overlap is poor, treated individuals are matched with worse prognostic matches. Since prognostic balance in the data set is then poor, you see higher variance in addition to bias.
3. *Gammas are much higher across the board, and they get increasingly large with rho* - This makes sense because bias is way higher. As it turns out, the difference in gamma between propensity score matching and joint matching is about the same as in our main simulations (at least for rho ~ 0). It just looks like the benefit is smaller because gamma is getting so big and changing the scale of the graphs.

Some intermediate simulations

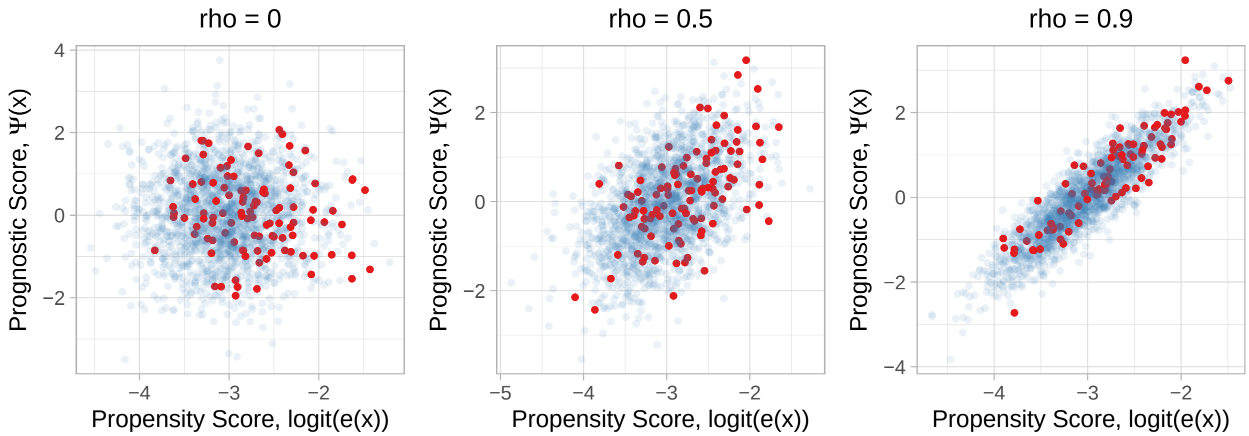
I happen to have run these at one point, so I thought I'd throw them in. Each one has some level of overlap between the simulations shown in the paper and the ones in the previous section.

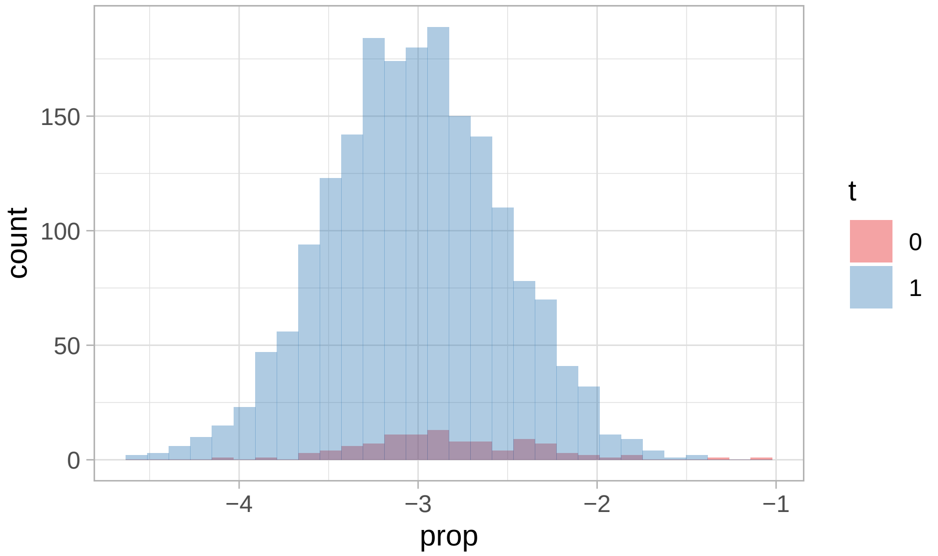
Phi = X1/2 - 3

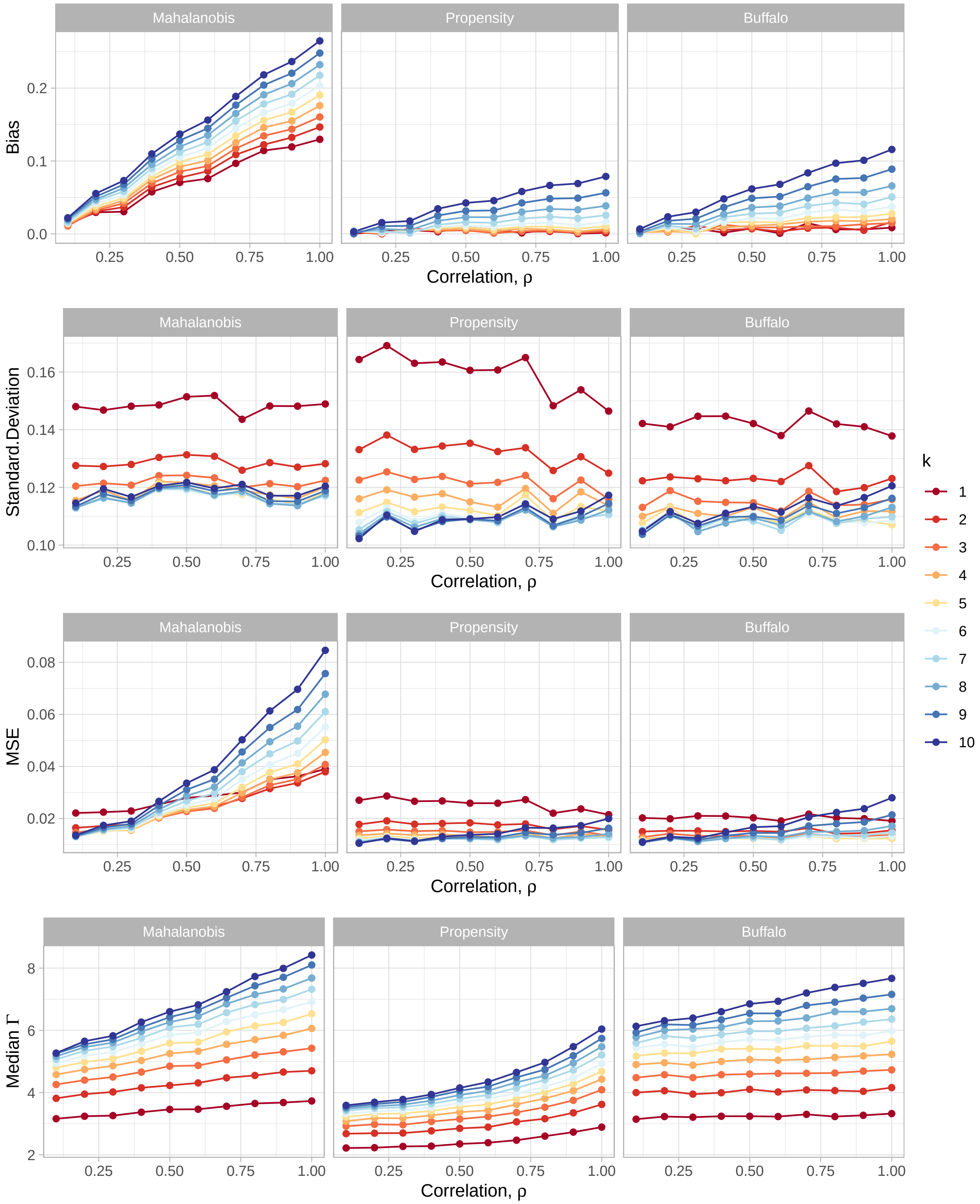
Suppose we switch to the following model:

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

Now our overlap is as below:





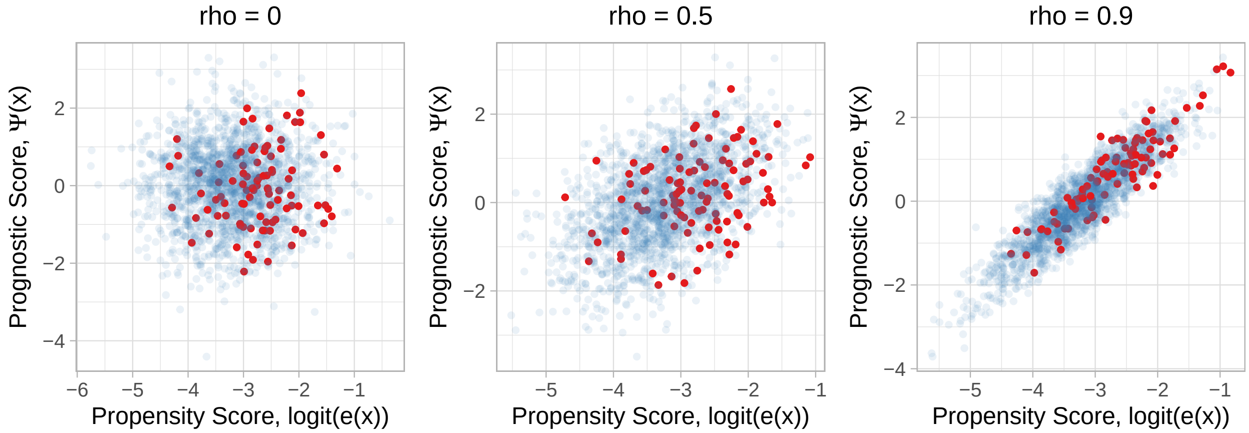

Mahalanobis
Propensity
Buffalo
Bias
0.0
0.1
0.2
0.25
0.50
0.75
1.00
Correlation, ρ
Standard.Deviation
0.10
0.12
0.14
0.16
MSE
0.02
0.04
0.06
0.08
Median Γ
2
4
6
8
k
1
2
3
4
5
6
7
8
9
10

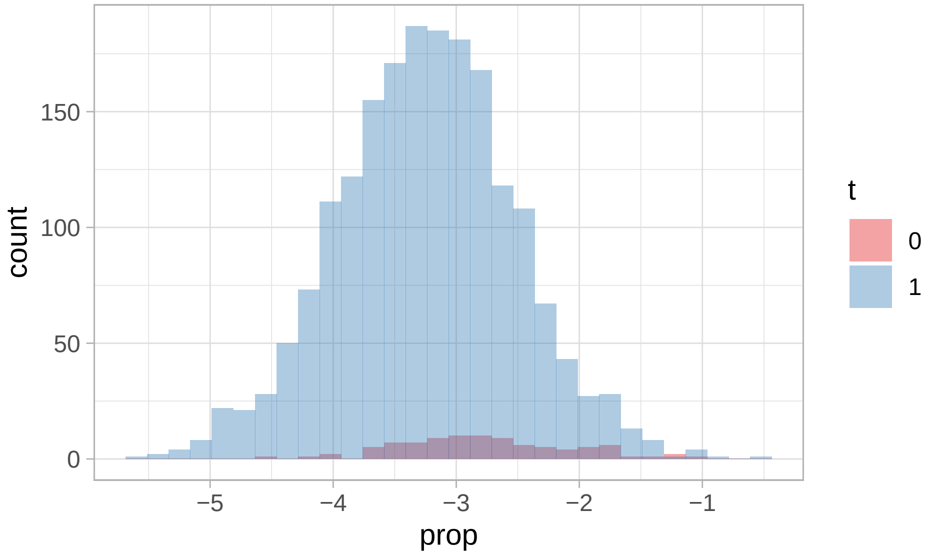
Phi = 3X1/4 - 3.2

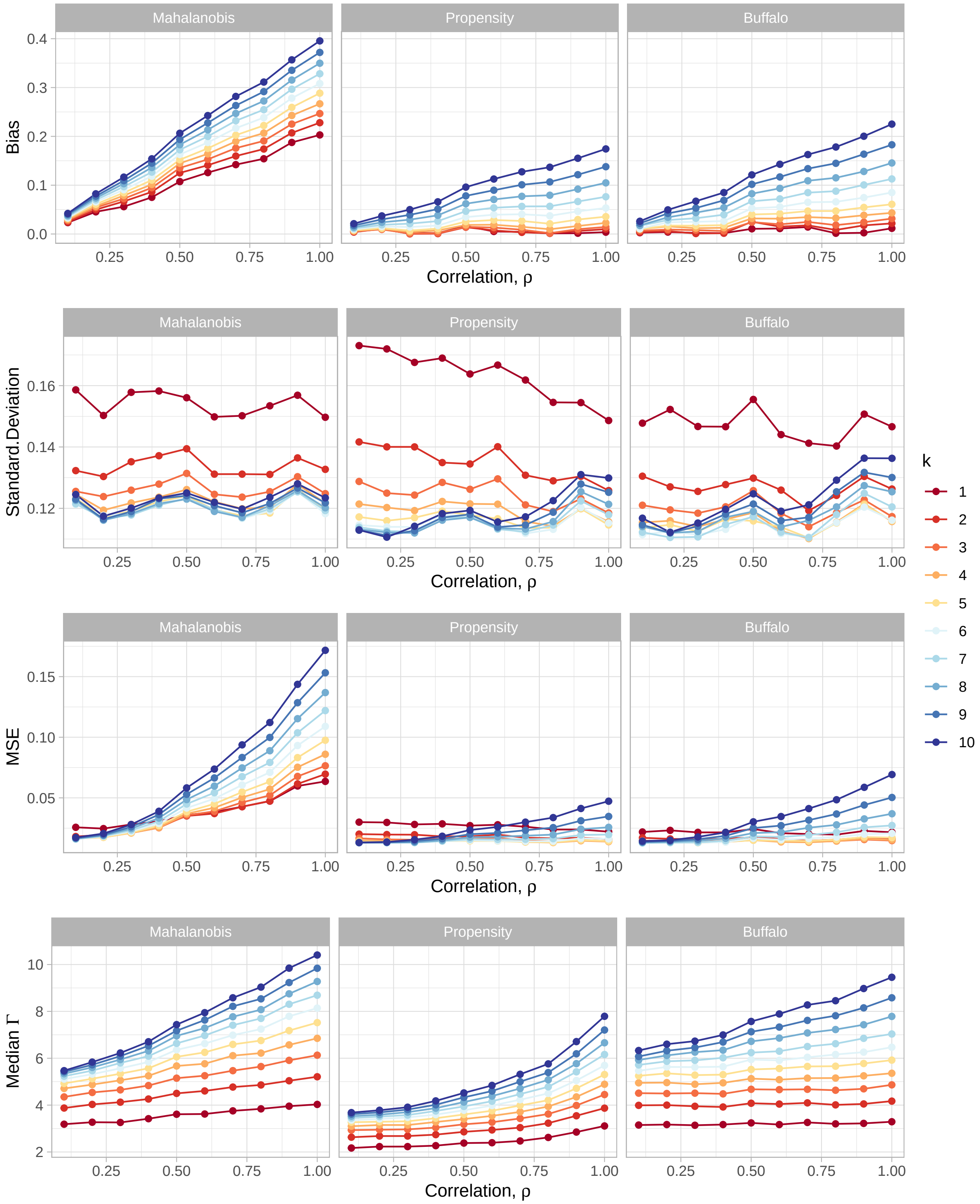
Suppose we switch to the following model:

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

Now this:






Mahalanobis
Propensity
Buffalo
Bias
0.0
0.1
0.2
0.3
0.4
0.25
0.50
0.75
1.00
Correlation, ρ
Standard.Deviation
0.12
0.14
0.16
MSE
0.05
0.10
0.15
Median Γ
2
4
6
8
10
k
1
2
3
4
5
6
7
8
9
10