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

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

so that $\text{Cor}(\phi(X_i), \Psi(X_i)) \propto \rho$. The propensity score formula was chosen such that there were approximately 100 treated observations in each dataset. We consider $p = 10$, $\rho = 0, 0.1, \dots, 0.9, 1.0$, and $k = 1, \dots, 10$. 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`

My Take

What we have now is not terrible. Sd is smaller than bias, but variance still factors somewhat into MSE. It would be nice if variance were slightly greater so that Buffalo would win on MSE. Also, something weird is happening where variance is actually increasing with k, and when this happens the variance of buffalo is also worse. We hypothesized that this was because we were choosing worse and worse matches in the prognostic dimension where overlap is bad (t > control). If this were true, it stands to reason that buffalo would be worse because we have discarded matches in this sparse area that's troubling us.

Here are the ways I tinkered with the simulation parameters and my thoughts on what happened:

- **Decreasing treatment effect:** weirdly didn't do much?
- **Increasing random variation in outcome:** Increased variance by a lot, but buffalo does worse relative to the other methods - probably because it is harder to fit a good prognostic model.
- **Decreasing the number of treated:** This is somewhat what Dylan did, and it works like a charm when $n_t = 30 - 40$, but this seems like too few. Right now I'm trying to keep it at about $n_t = 100$, which feels more reasonable. Dylan's formula also increased the randomness in the treatment assignment (see below), which has it's own pros and cons.
- **Increasing the randomness in treatment assignment:** This seems to work okay, and I'm running some more simulations for it. It also deals with the problem of variance increasing with k, I assume because there is better overlap between the treat and control populations. My hang-up is that this dimishes the amount of structure/confounding there is to the problem, which is not necessarily what we're trying to do.
- **Increasing the importance of covariates for prognosis:** This also isn't bad, but it has the weird side-effect that Γ values skyrocket, for reasons I don't thoroughly understand.

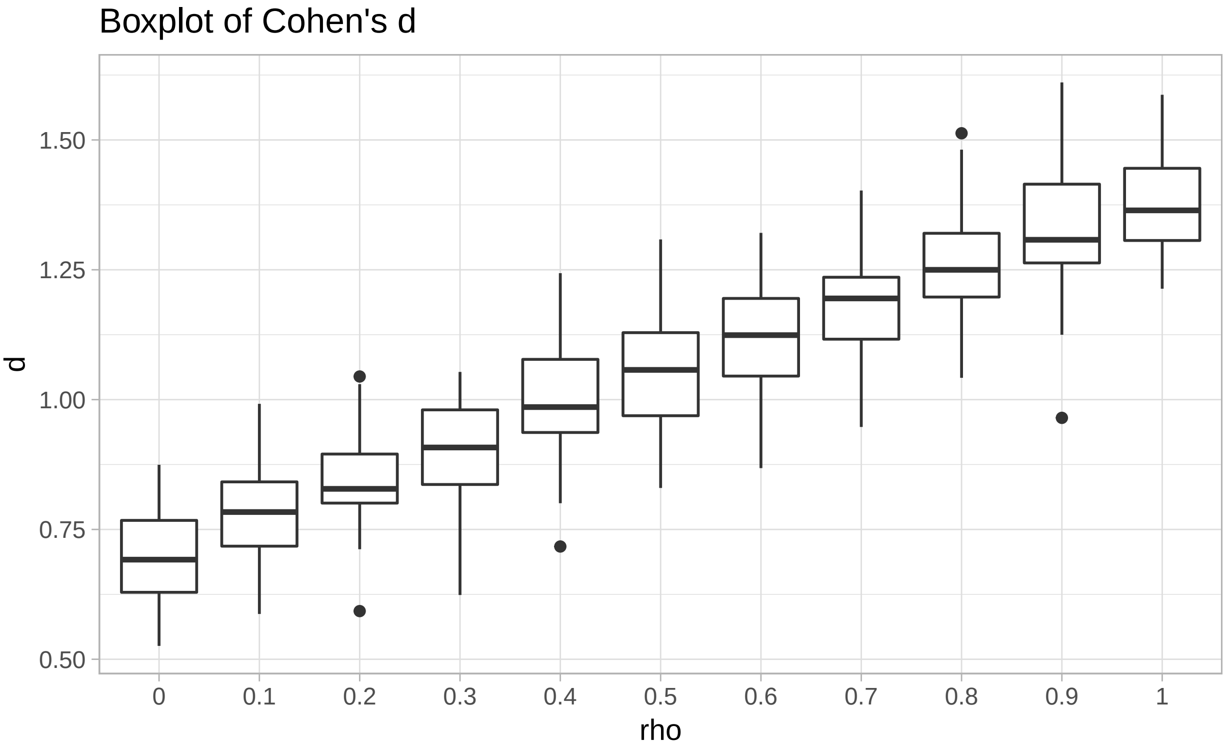
What we have already

Cohen's D

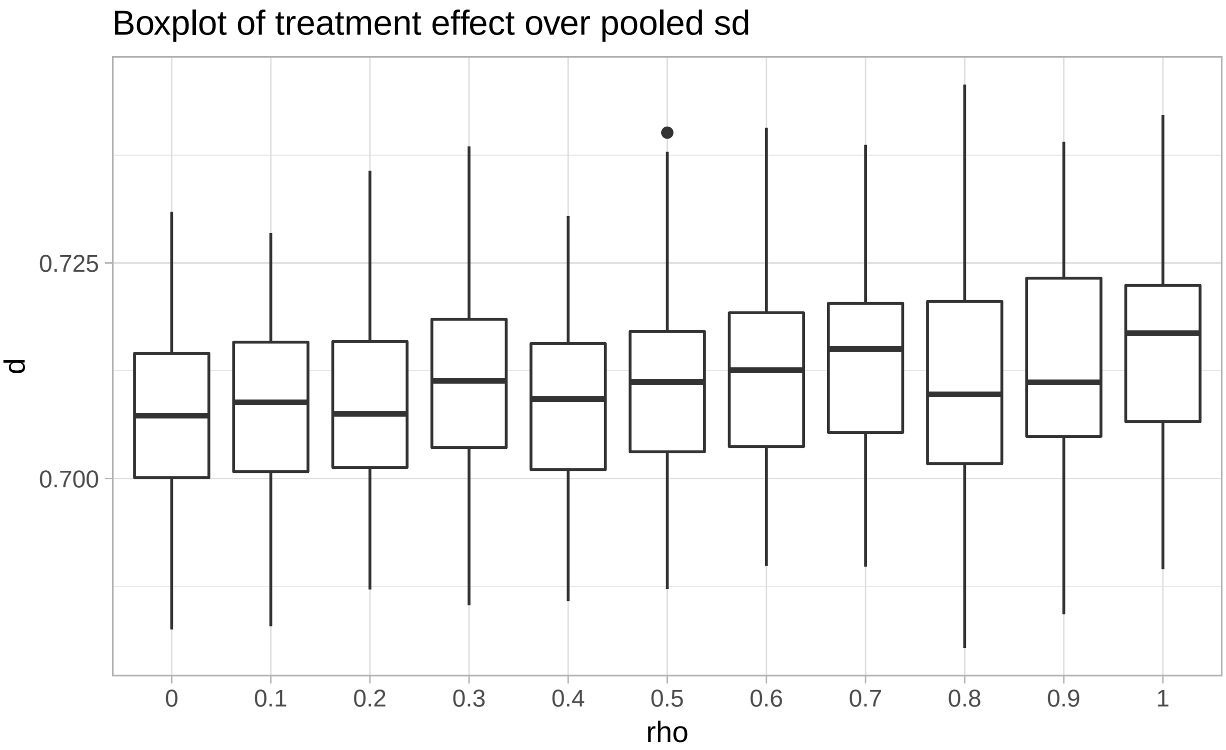
We'd ideally like simulation parameters that give a cohen's d of about 0.2.

This formulation of the problem has a cohen's d of about 0.69 (ranging about 0.4 - 0.9):

This set up has the following cohen's d distributions, which is probably too high.

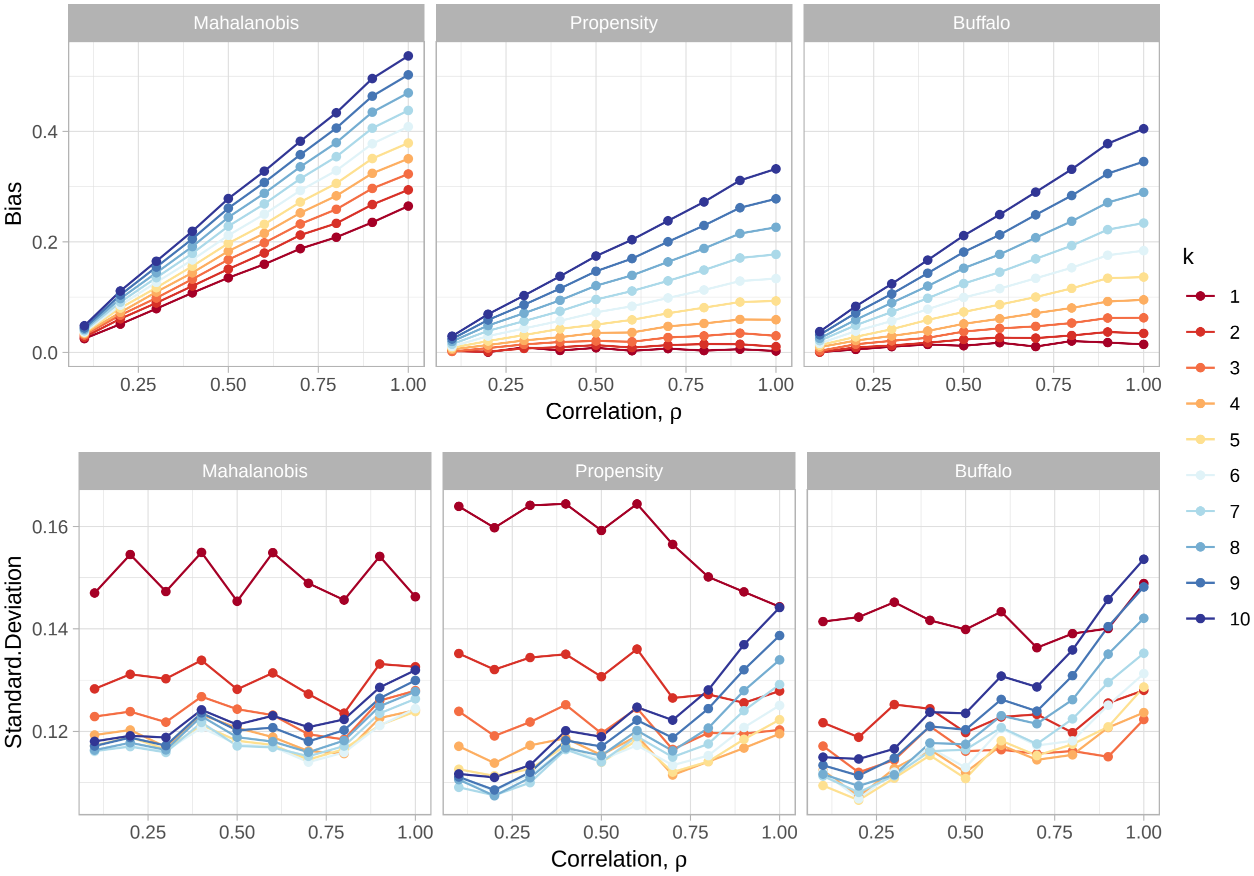


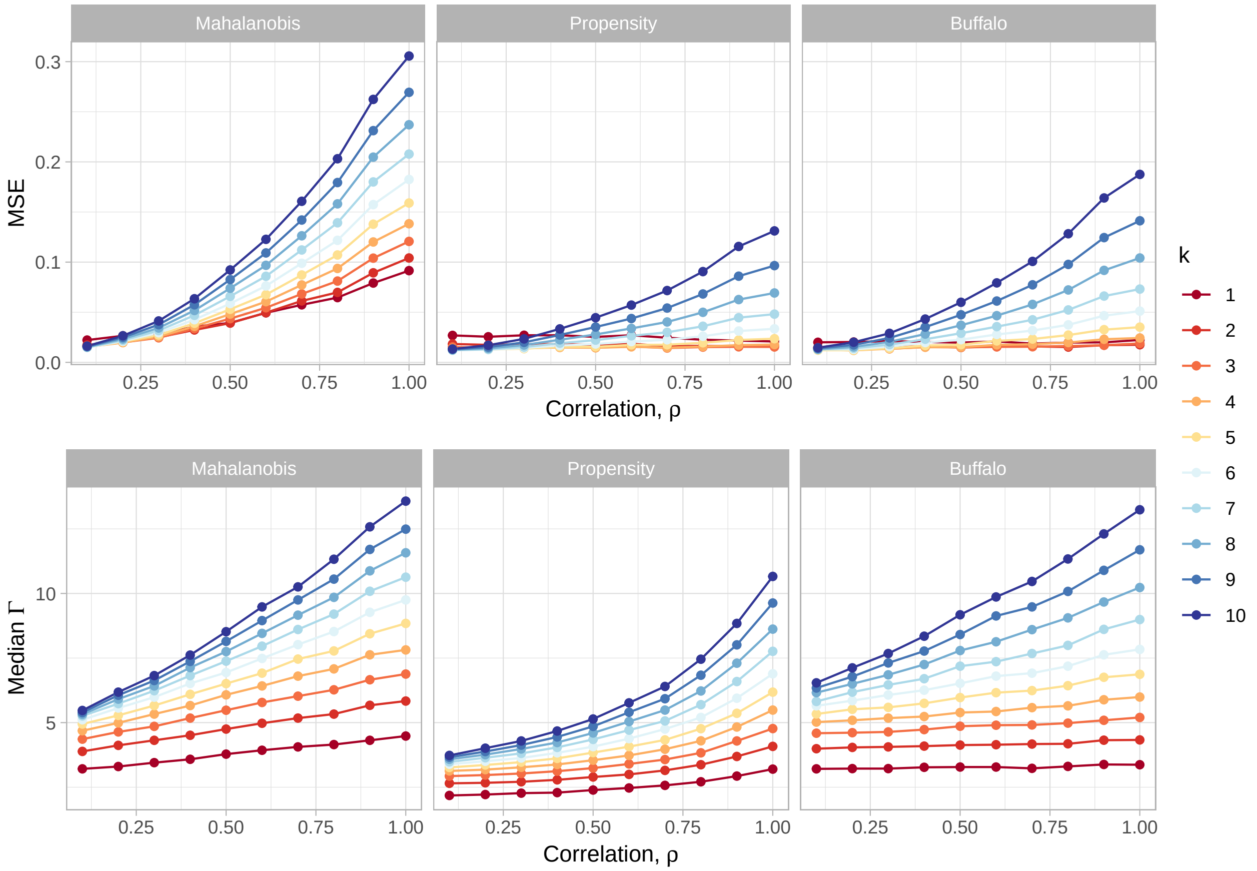
A lot of the magnitude of Cohen's D above is actually caused by confounding rather than the effect of treatment; when rho is close to 1, the outcomes in the treatment and control groups appear farther apart because the covariate which determines treatment assignment also strongly determines outcome. If, instead of Cohen's D, we calculate the true causal effect divided by the pooled standard deviation of the outcome across treatment and controlled individuals, we get a different result:



Performance

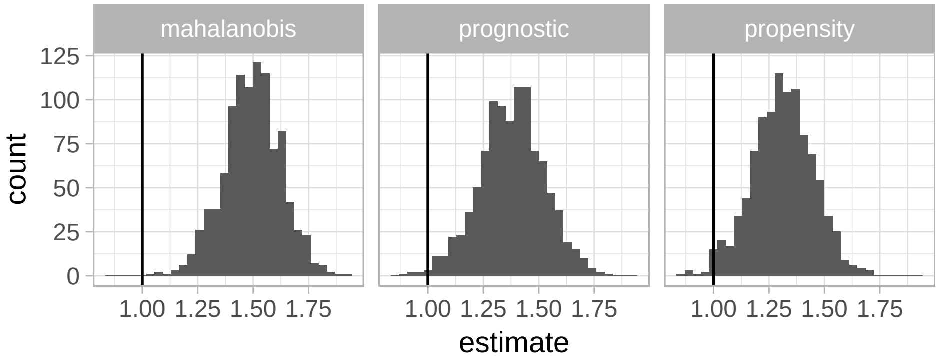
Below, we see the results of the latest complete batch of N = 1000 simulations, which use the exact set-up above. One issue with this set of results is that the bias is on such a larger scale than the variance that the bias dominates the MSE, and the bias/variance tradeoffs of the different approaches aren't highlighted.


Mahalanobis
Propensity
Buffalo
Bias
0.0
0.2
0.4
0.25
0.50
0.75
1.00
Correlation, ρ
k
1
2
3
4
5
6
7
8
9
10
Standard.Deviation
0.12
0.14
0.16



To visualize this a little differently, here's a histogram of the estimates made by each method when $k = 1$, $\rho = 0.9$.

```
## # A tibble: 3 x 2
##   method       mean
##   <fct>       <dbl>
## 1 mahalanobis  1.50
## 2 prognostic   1.38
## 3 propensity   1.31
```

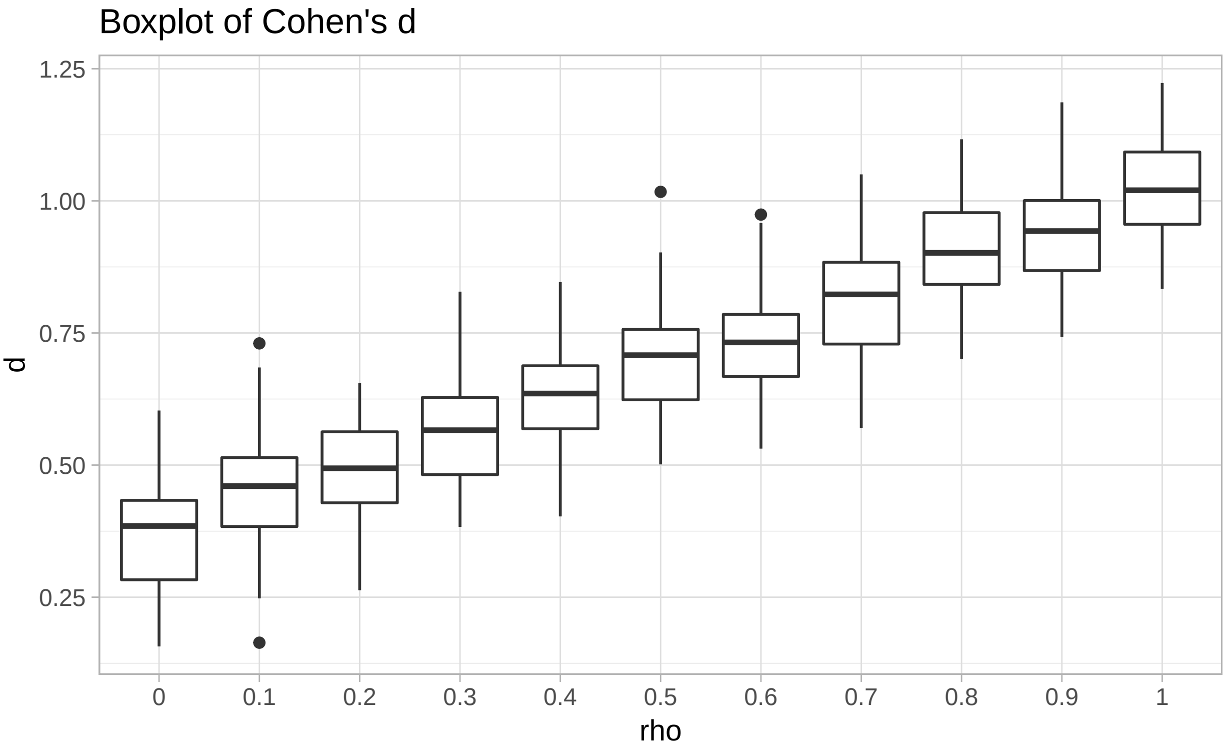


Playing with Tau

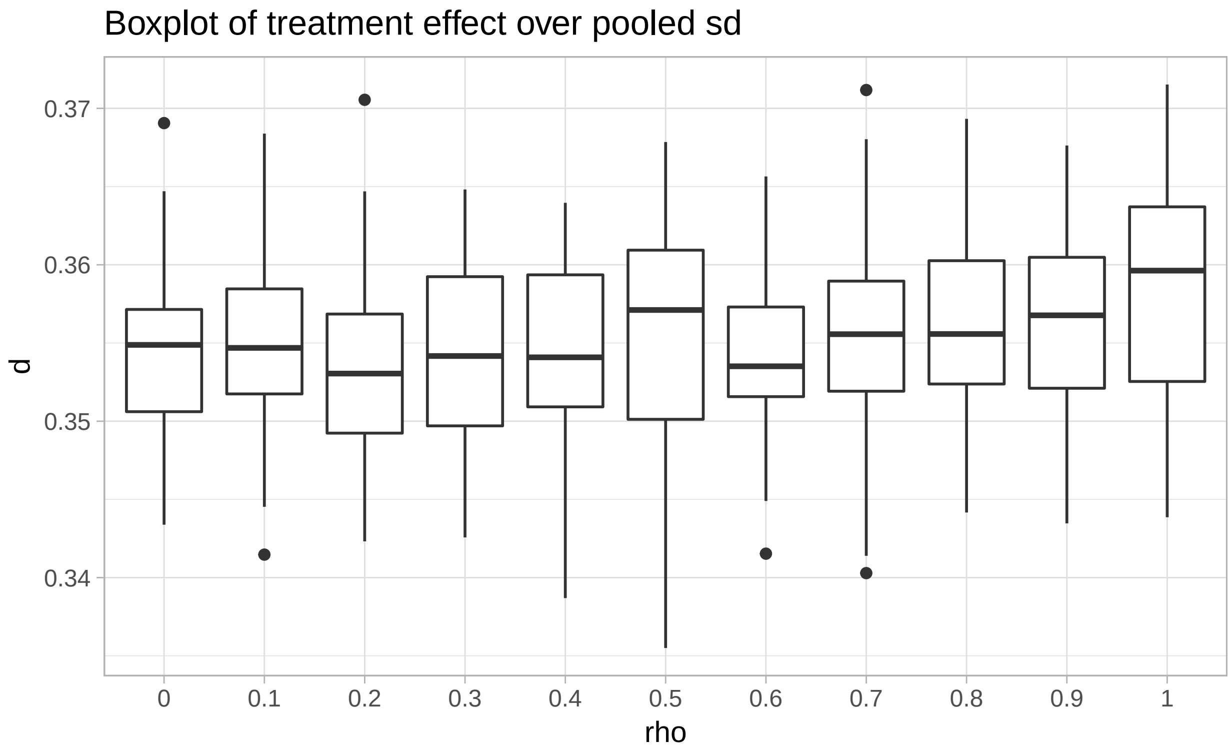
Suppose we switch to the following model:

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

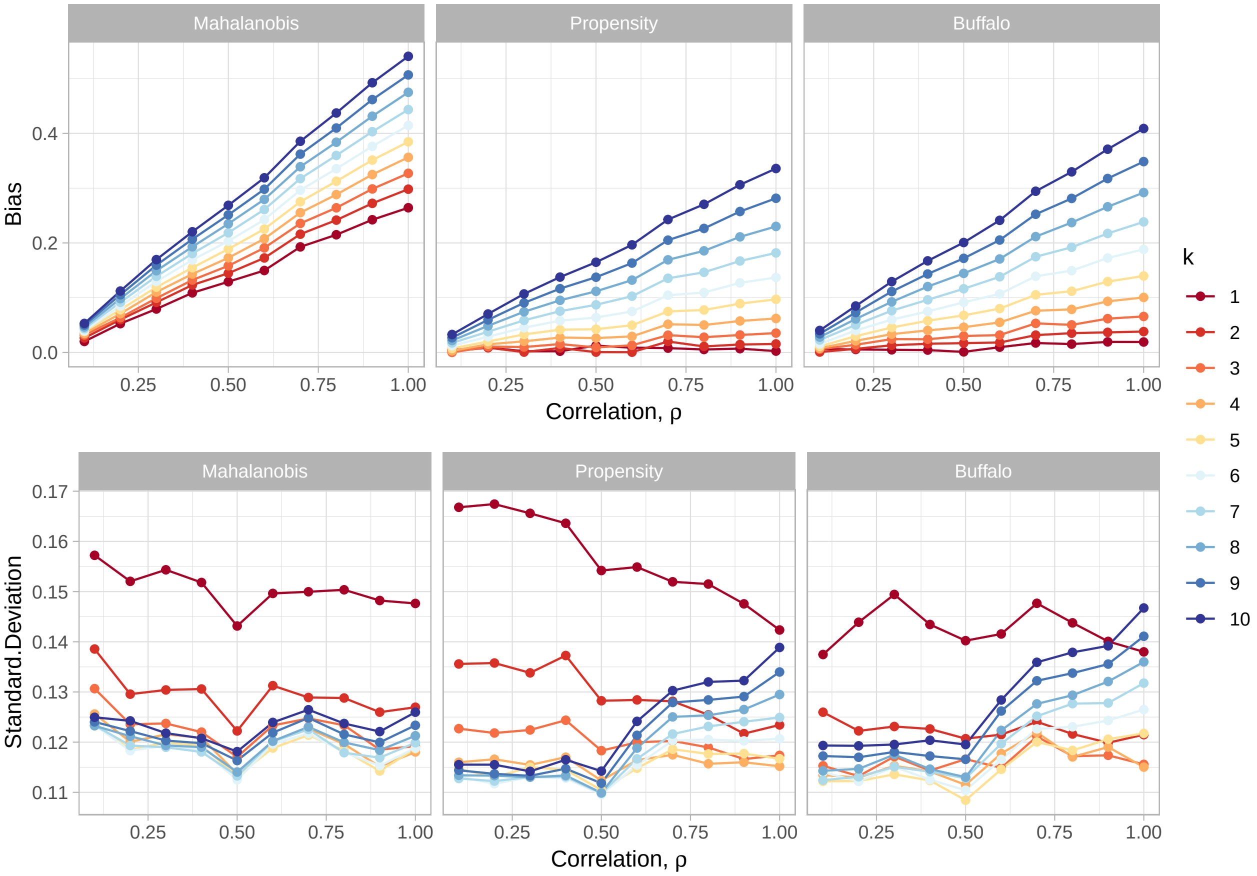
This gives lower cohen's d:

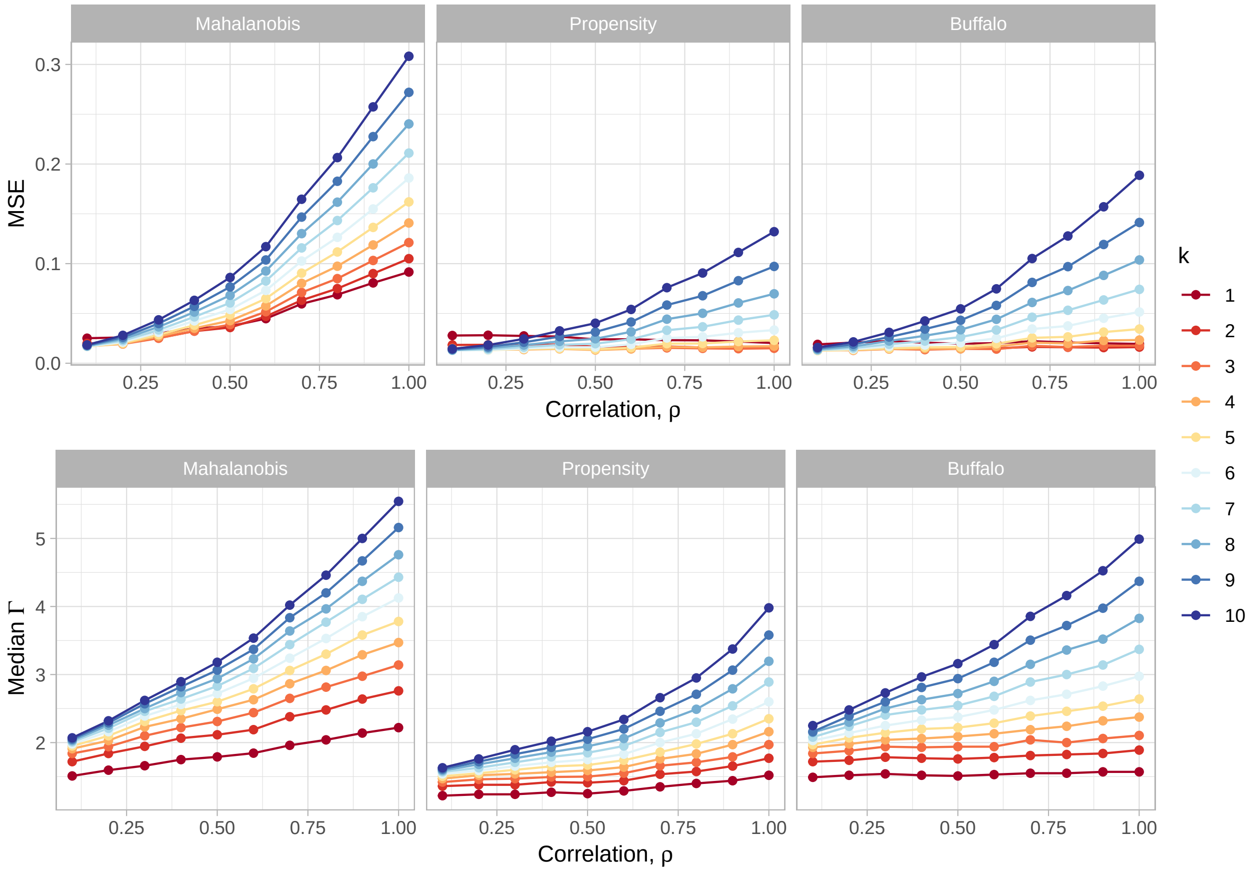


Also, the true effect divided by the pooled sd is lower:

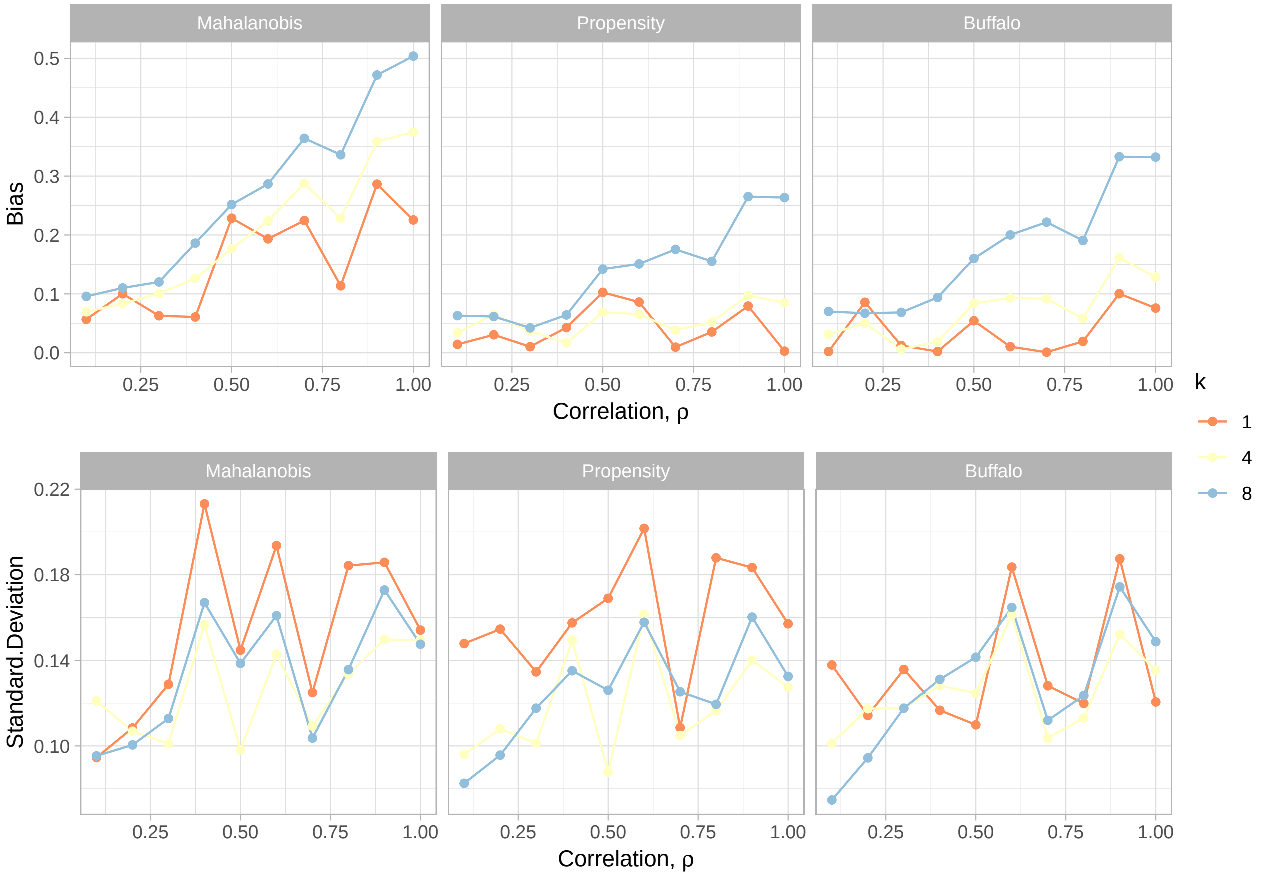


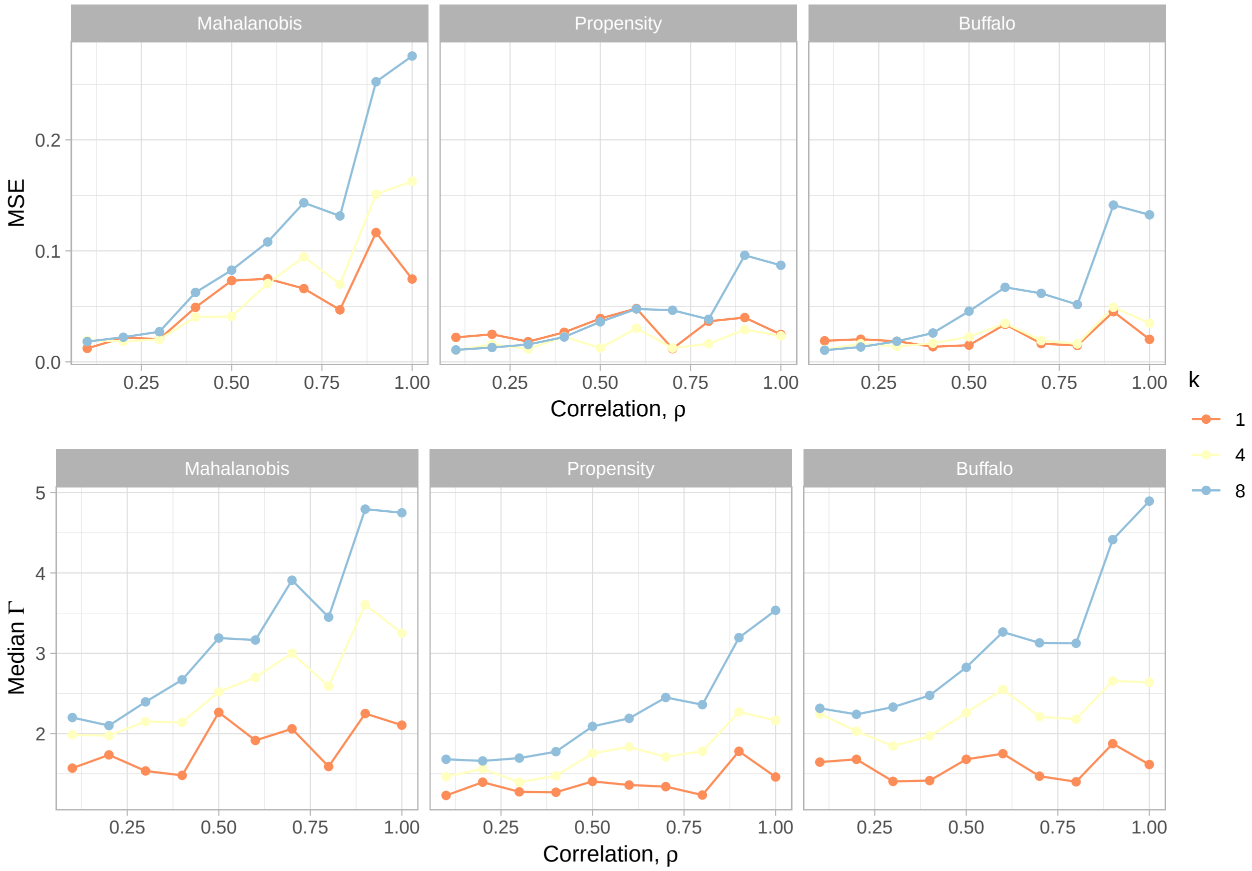
Here are some results from large-batch simulations. There is surprisingly little changed compared to the old simulation parameters:





Here's a smaller simulation showing more-or-less the same thing just to sanity check.


Mahalanobis
Propensity
Buffalo
Bias
0.5
0.4
0.3
0.2
0.1
0.0
0.25
0.50
0.75
1.00
Correlation, ρ
Mahalanobis
Propensity
Buffalo
Standard.Deviation
0.22
0.18
0.14
0.10
0.25
0.50
0.75
1.00
Correlation, ρ
k
1
4
8



Playing with Sigma

Sigma = 2

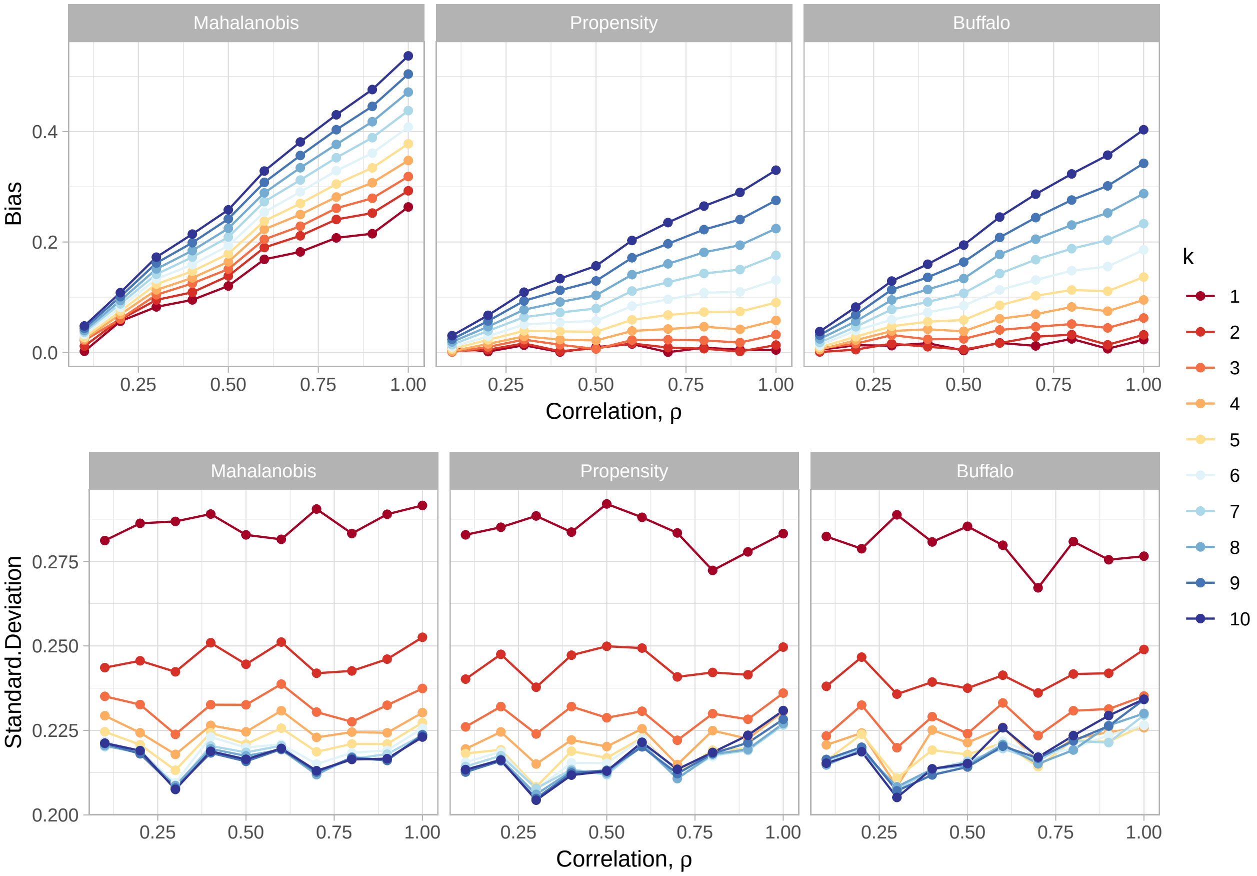
Another potential solution is increasing the noise. Suppose we have this model:

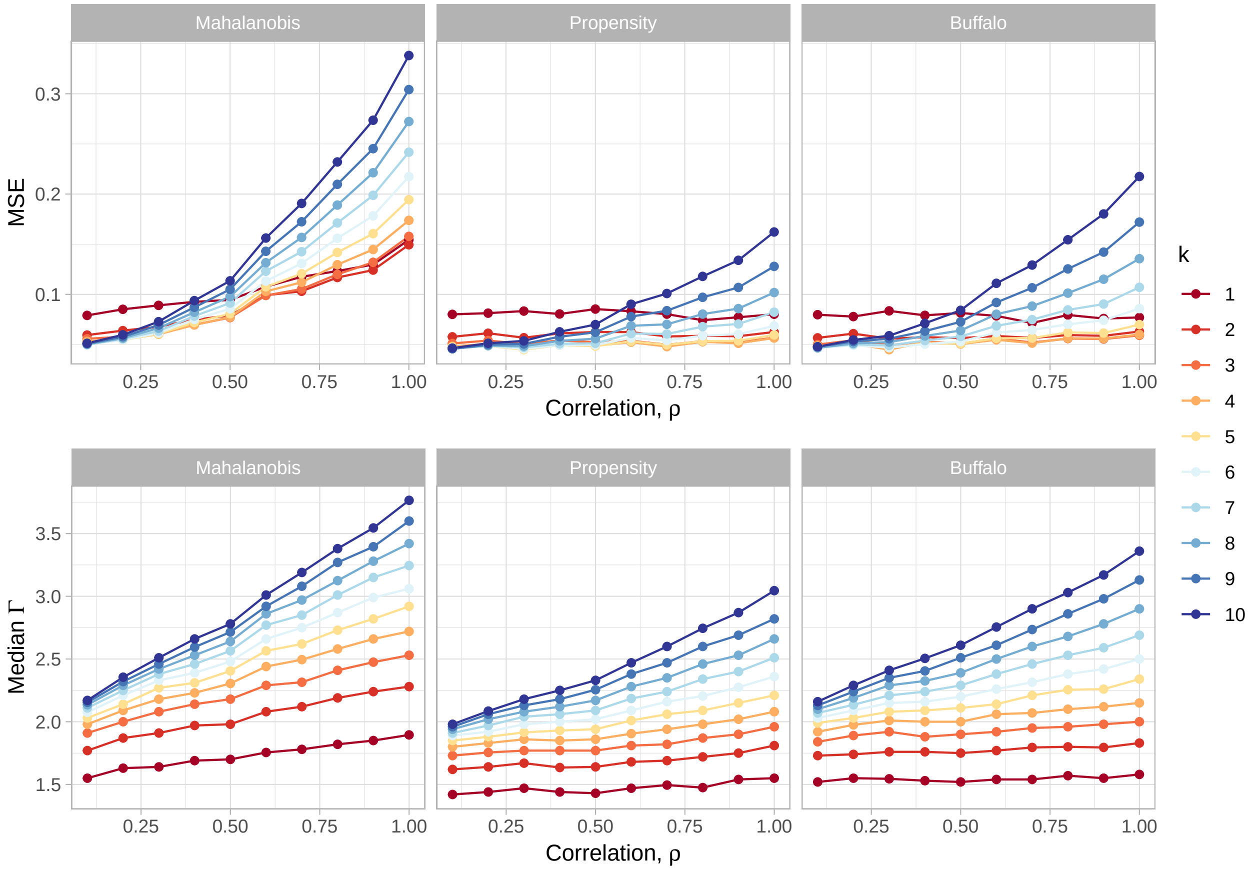
$$
\begin{aligned}
X_i &\sim_{iid} \text{Normal}(0, I_p),\\
T_i &\sim_{iid} \text{Bernoulli}\left(\frac{1}{1+\exp(-\phi(X_i))}\right),\\
Y_i &= \tau T_i + \Psi(X_i) + \epsilon_i,\\
\epsilon_i &\sim_{iid} N(0, \sigma^2),
\end{aligned}
$$

With the following formulae for propensity and prognosis

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

Except that we let $\sigma = 2$.


Mahalanobis
Propensity
Buffalo
Bias
0.0
0.2
0.4
0.25
0.50
0.75
1.00
Correlation, ρ
Mahalanobis
Propensity
Buffalo
Standard.Deviation
0.200
0.225
0.250
0.275
0.25
0.50
0.75
1.00
Correlation, ρ
k
1
2
3
4
5
6
7
8
9
10



This puts Bias and SD on the same scale, but there's so much noise that the prognostic score methods probably have a harder time fitting a good model.

Playing with Propensity

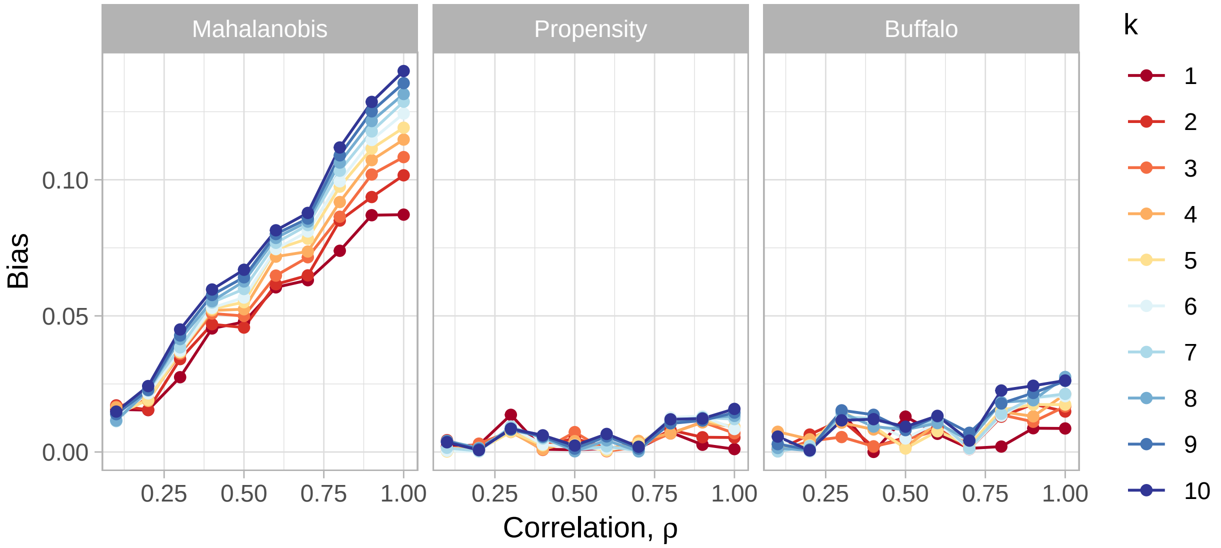
Phi = X1/3 - 4

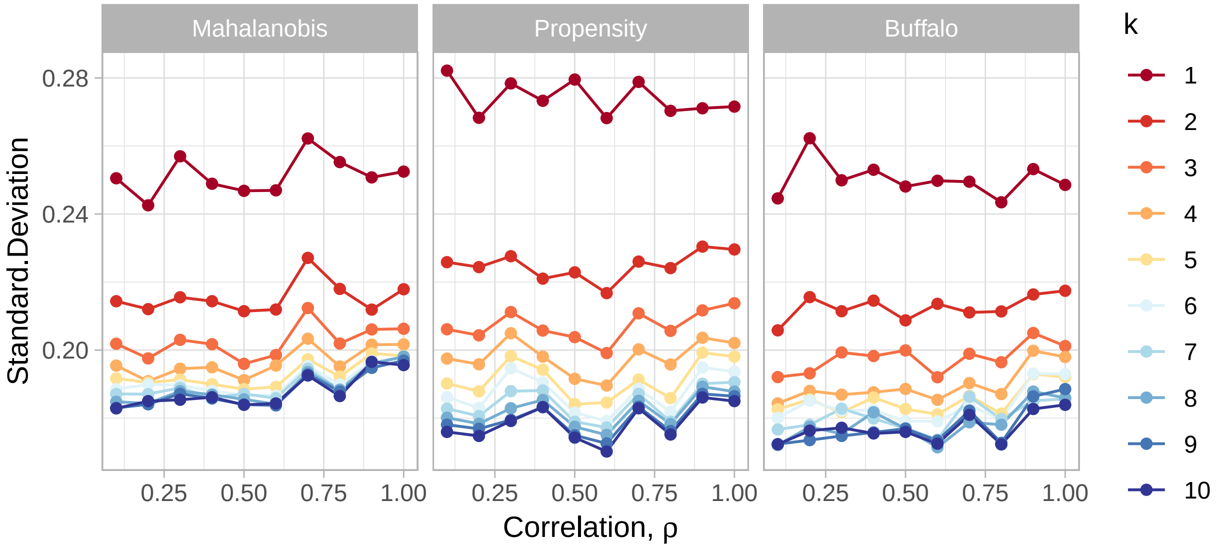
This is the model that Dylan was using when I took over:

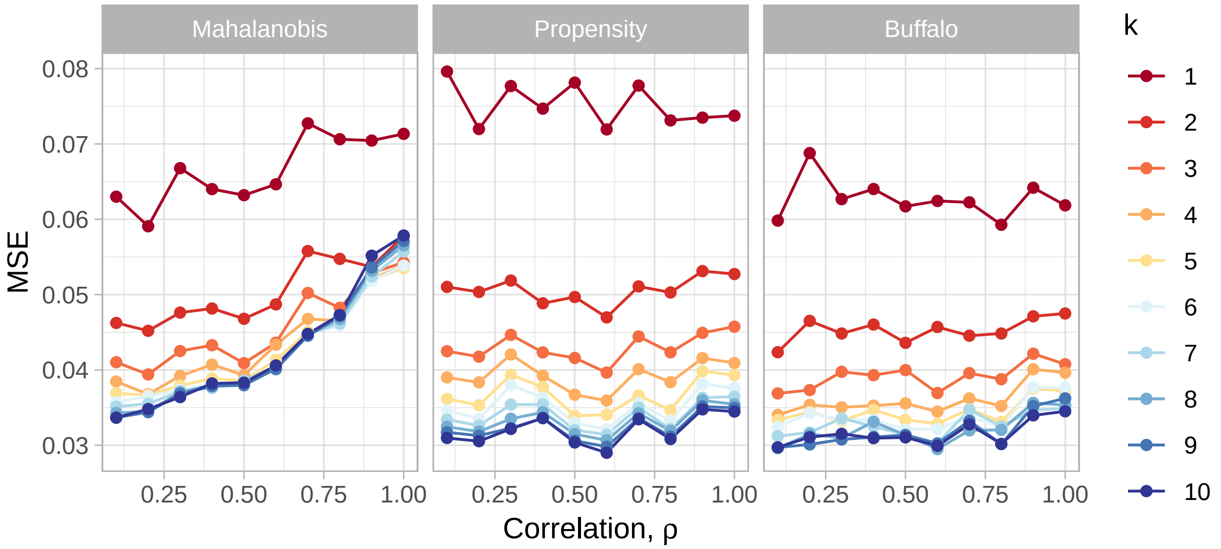
$$\phi(X_i) = X_{i1}/3 - 4,$$
$$\Psi(X_i) = \rho X_{i1} + \sqrt{(1-\rho^2)}X_{i2},$$

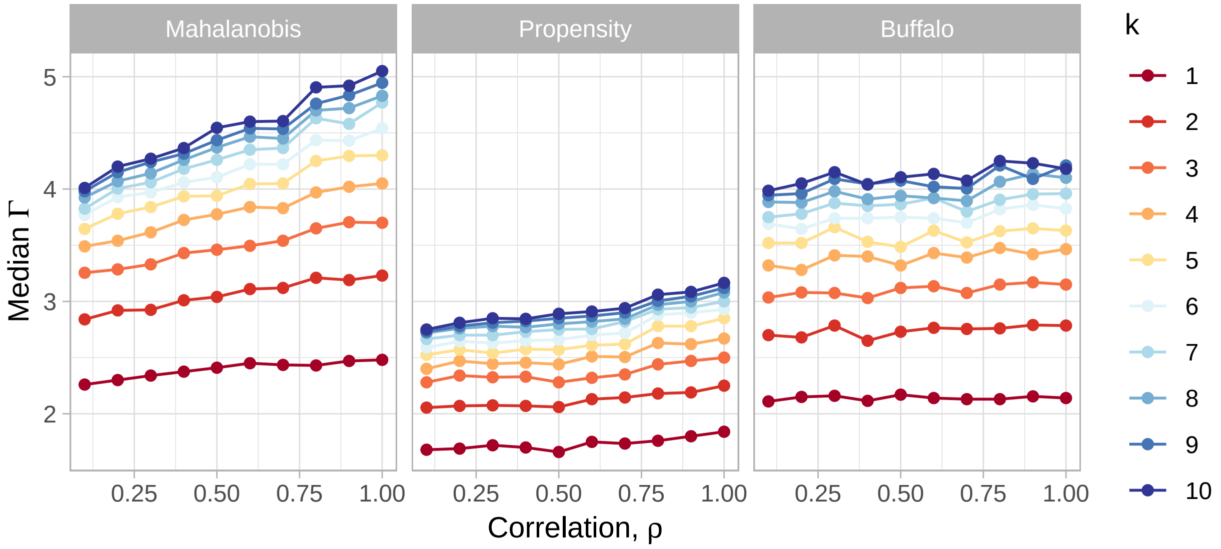
It's nice, because it puts bias and variance on the same approximate scale, but it allows the number of treated individuals to drop to 30-40, rather than 100. Buffalo also does weirdly poorly in terms of gamma compared to mahalanobis.

```
true_tau <- 1

dat <- mutate(dat,
              squared_err = (estimate-true_tau)**2,
              k = as.factor(k))

plt_data <- dat %>%
  group_by(method, k, rho) %>%
  summarize(Bias = abs(mean(estimate) - true_tau),
            median_gamma = median(gamma),
            Standard.Deviation = sd(estimate),
            MSE = Bias^2 + Standard.Deviation^2) %>%
  ungroup() %>%
  mutate(method = recode(method, propensity = "Propensity",
                           mahalanobis = "Mahalanobis",
                           prognostic = "Buffalo"))
```







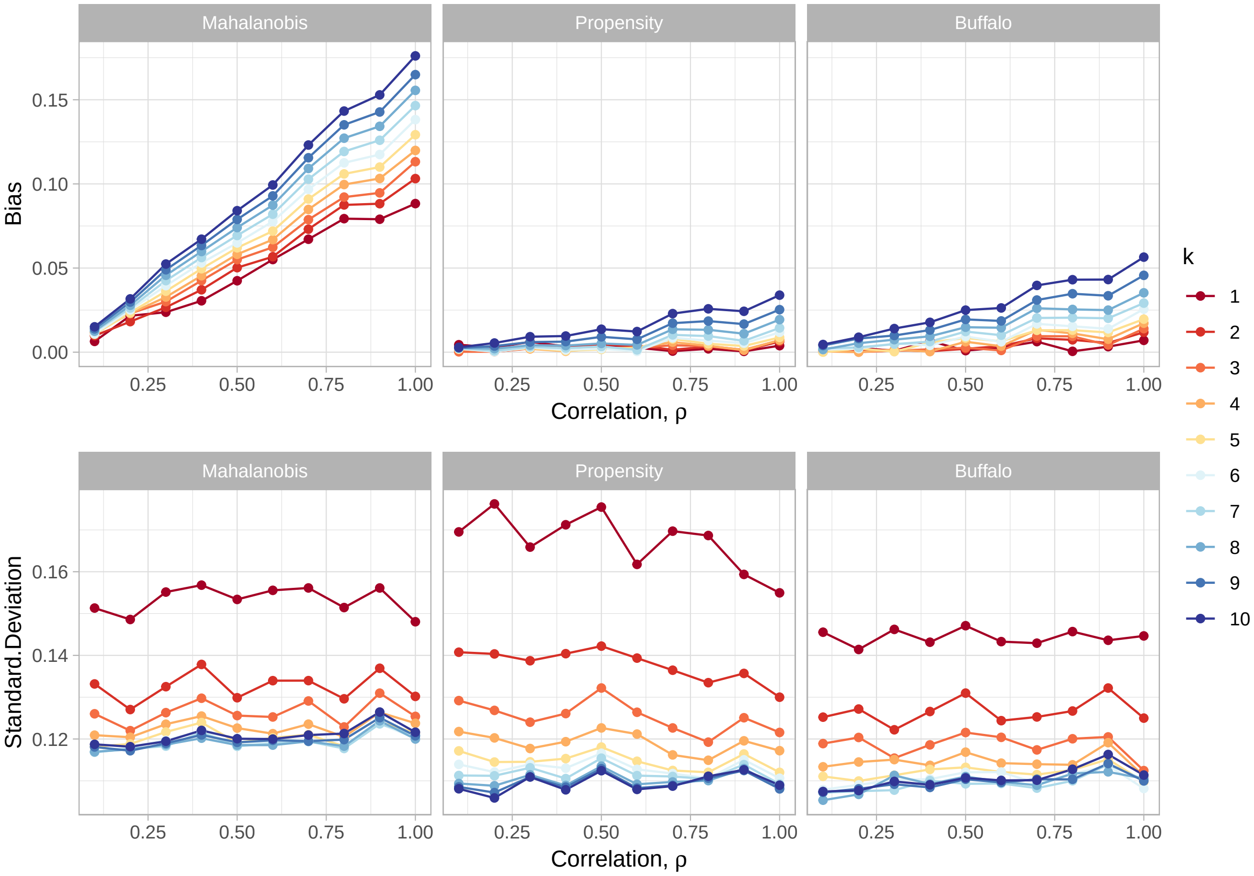


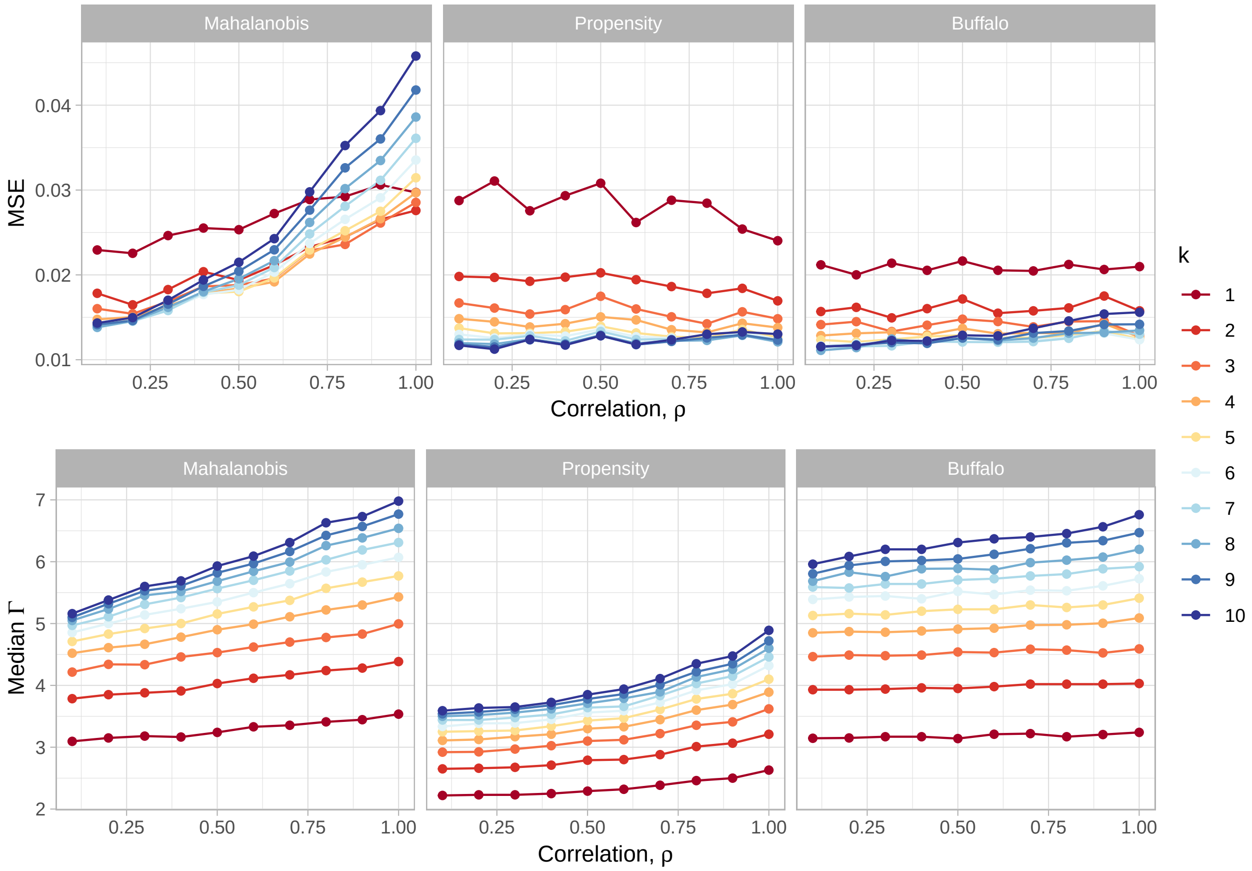
Phi = X1/3 - 3

Suppose we switch to the following model:

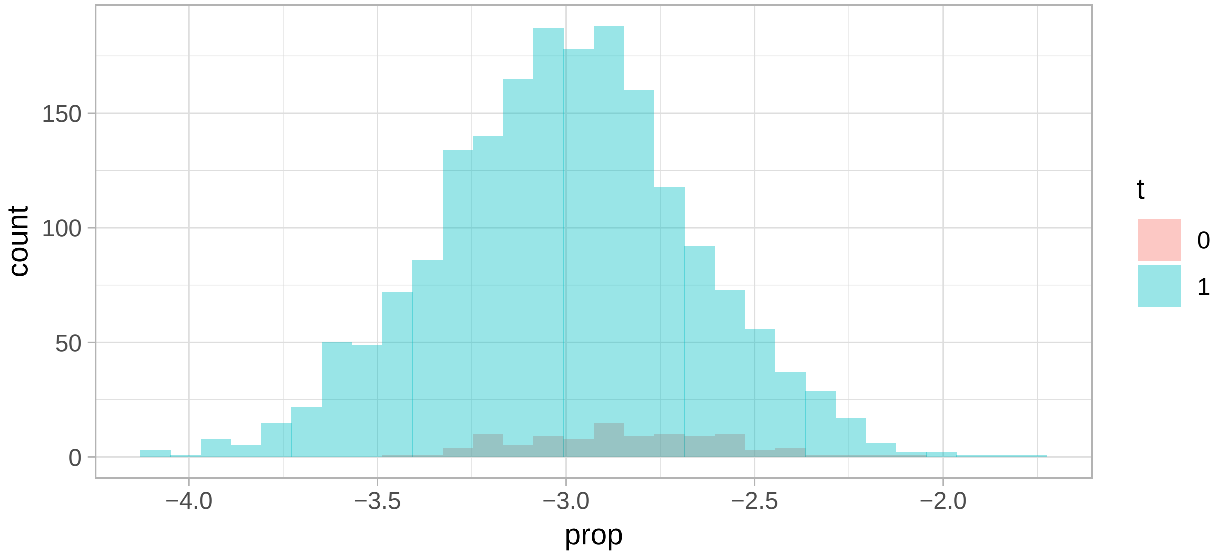
$$\phi(X_i) = X_{i1}/3 - 3,$$
$$\Psi(X_i) = \rho X_{i1} + \sqrt{(1-\rho^2)}X_{i2},$$

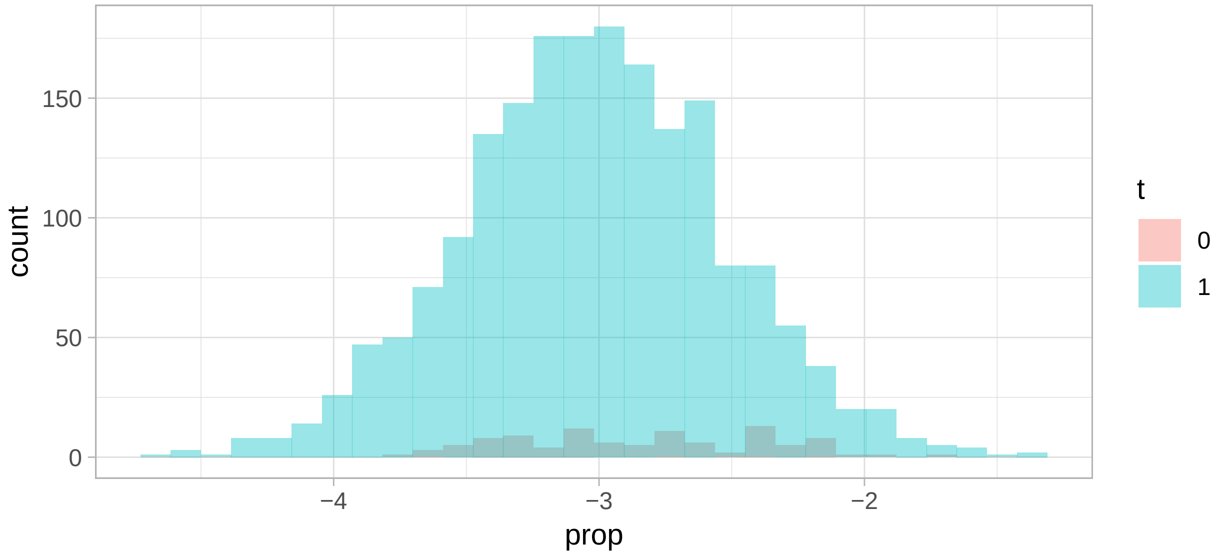
This decreases the importance of X_1 in determining treatment assignment, while keeping the number of treated individuals at approximately 100. Doing this decreases bias while increasing variance.

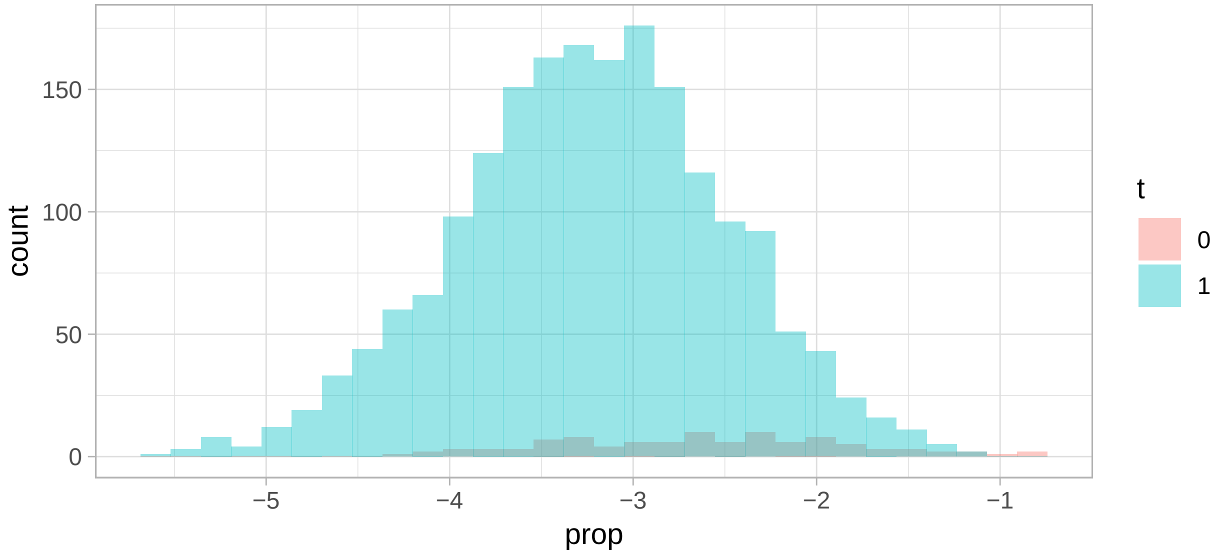

Mahalanobis
Propensity
Buffalo
Bias
0.00
0.05
0.10
0.15
0.25
0.50
0.75
1.00
Correlation, ρ
Standard.Deviation
0.12
0.14
0.16
k
1
2
3
4
5
6
7
8
9
10

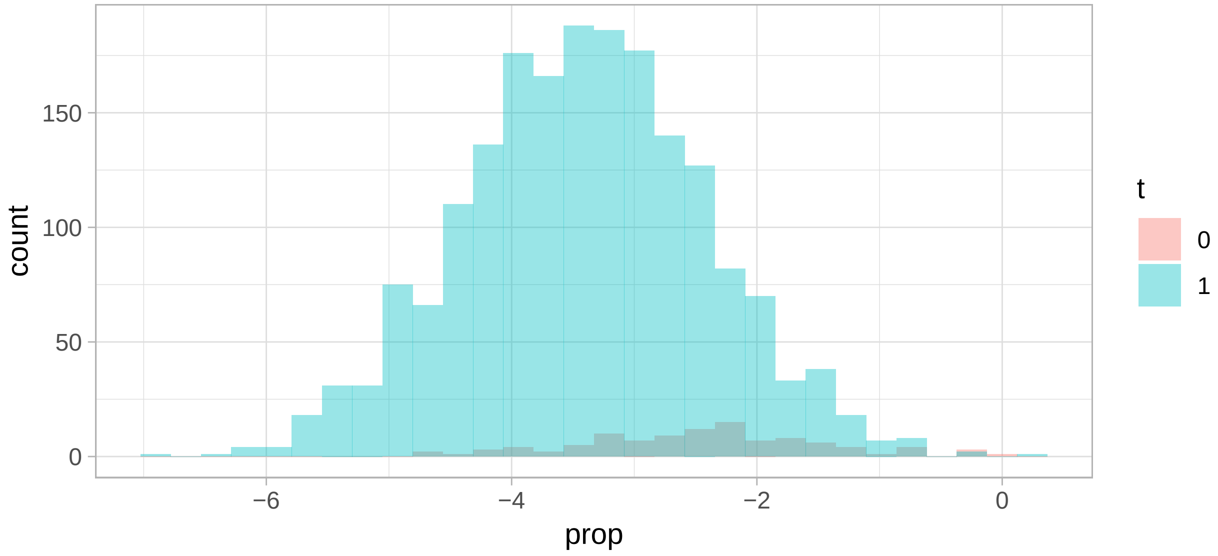


This seems pretty good, and I'm running a larger batch of simulations with this set-up now. One thing here that I'm not crazy about is that when we diminish the importance of X_1 in determining treatment assignment, we get closer and closer to random treatment assignment, which maybe isn't the use case that we're most interested in. Below, I've printed a histogram of ϕ for treated (red) and control (blue) individuals. On the left, we see what happen s when $\phi = X_1/3 - 3$, and on the right, we see what happens when $\phi = X1 - 10/3$.




count
150
100
50
0
−4
−3
−2
prop
t
0
1


count
150
100
50
0
−5
−4
−3
−2
−1
prop
t
0
1

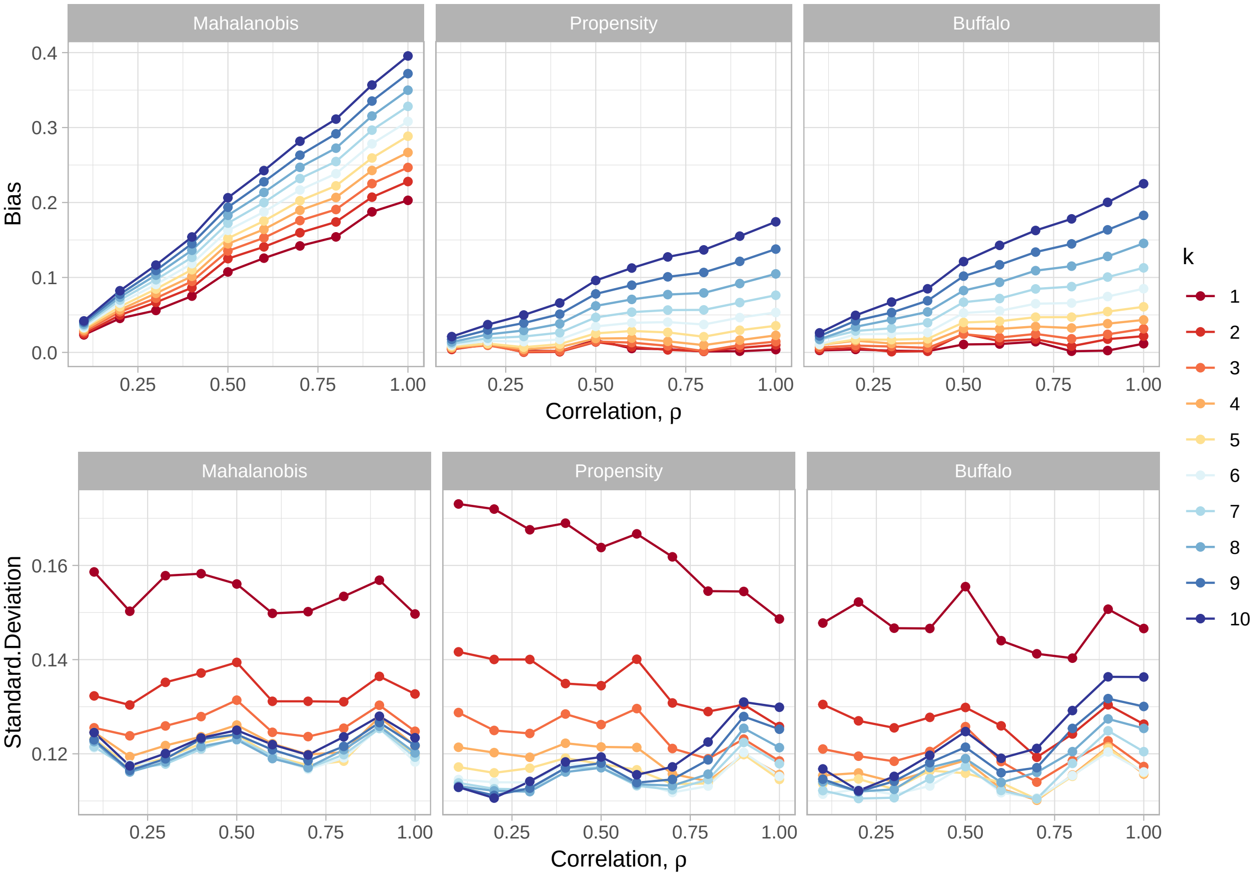

count
150
100
50
0
−6
−4
−2
0
prop
t
0
1

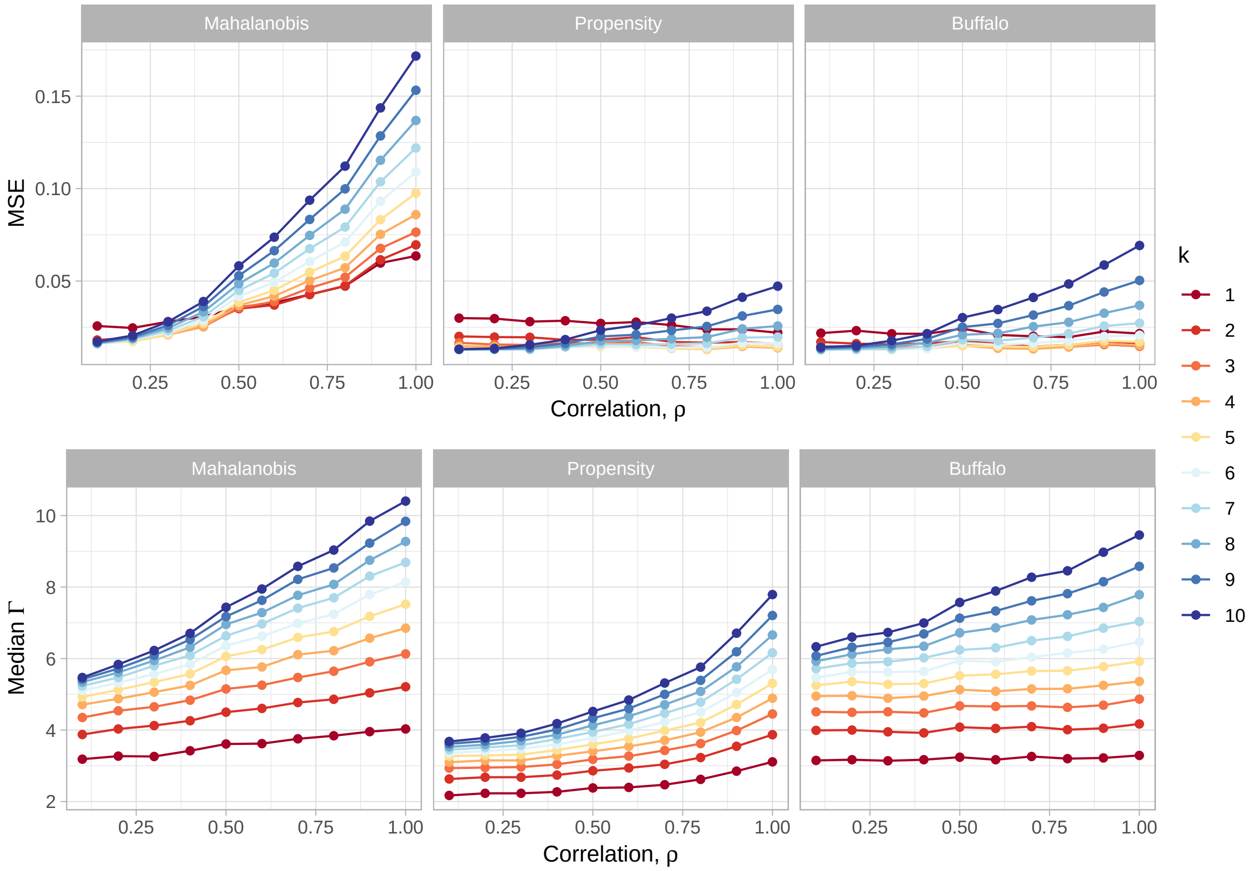
Phi = 3X1/4 - 3.2

Suppose we switch to the following model:

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

This decreases the importance of X_1 in determining treatment assignment, while keeping the number of treated individuals at approximately 100. Doing this decreases bias while increasing variance. It might be a happy medium between the two propensity score formulae above





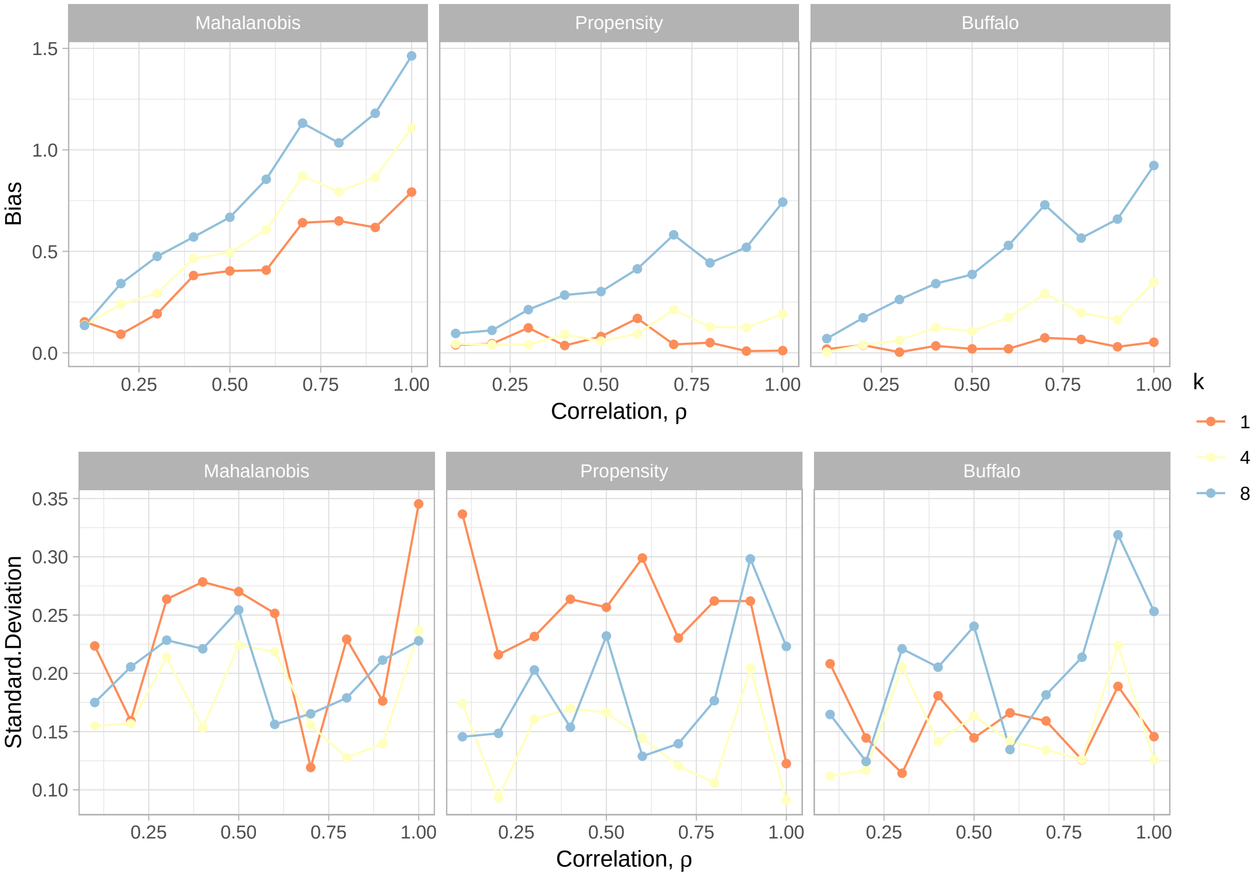
Playing with Prognosis

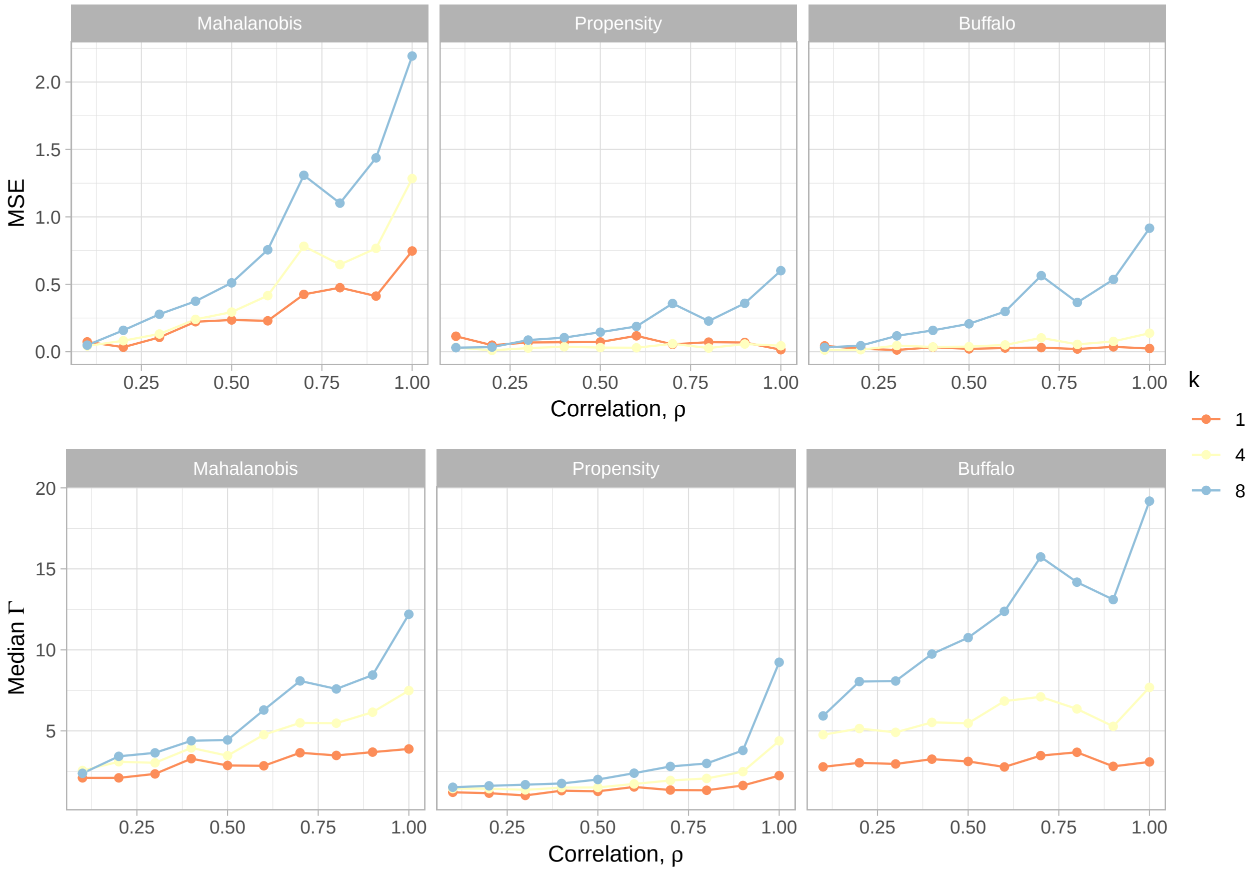
Psi times 3

Suppose we switch to the following model:

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

Then bias and variance both increase, but variance increases more than bias. This makes bias and variance closer to the same scale (but not quite the same). The maximum bias can be about 10 times the maximum variance. Also worth noting: Gamma for the buffalo goes *way* up.


Mahalanobis
Propensity
Buffalo
Bias
1.5
1.0
0.5
0.0
0.25
0.50
0.75
1.00
Correlation, ρ
k
1
4
8
Standard.Deviation
0.35
0.30
0.25
0.20
0.15
0.10



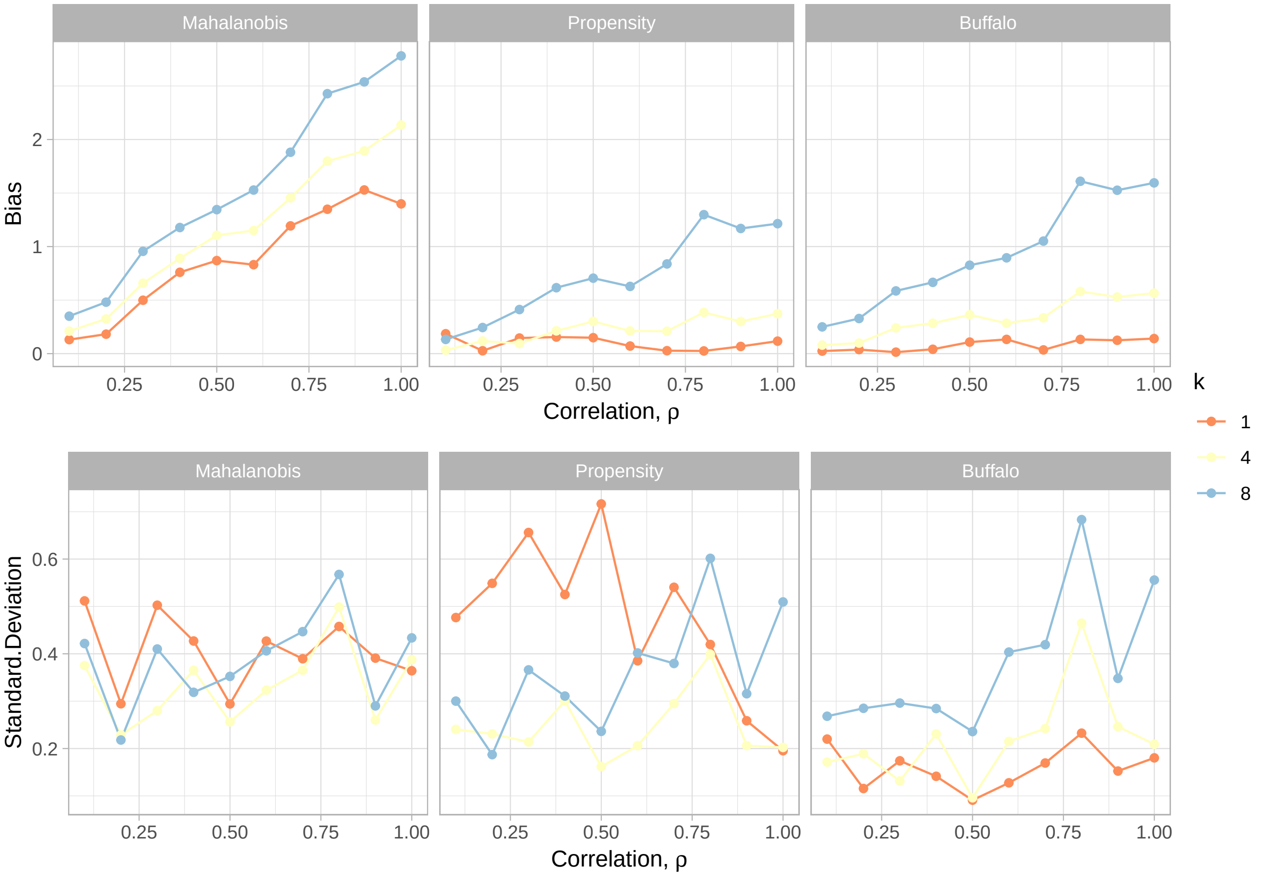
This is weird because gamma skyrockets.

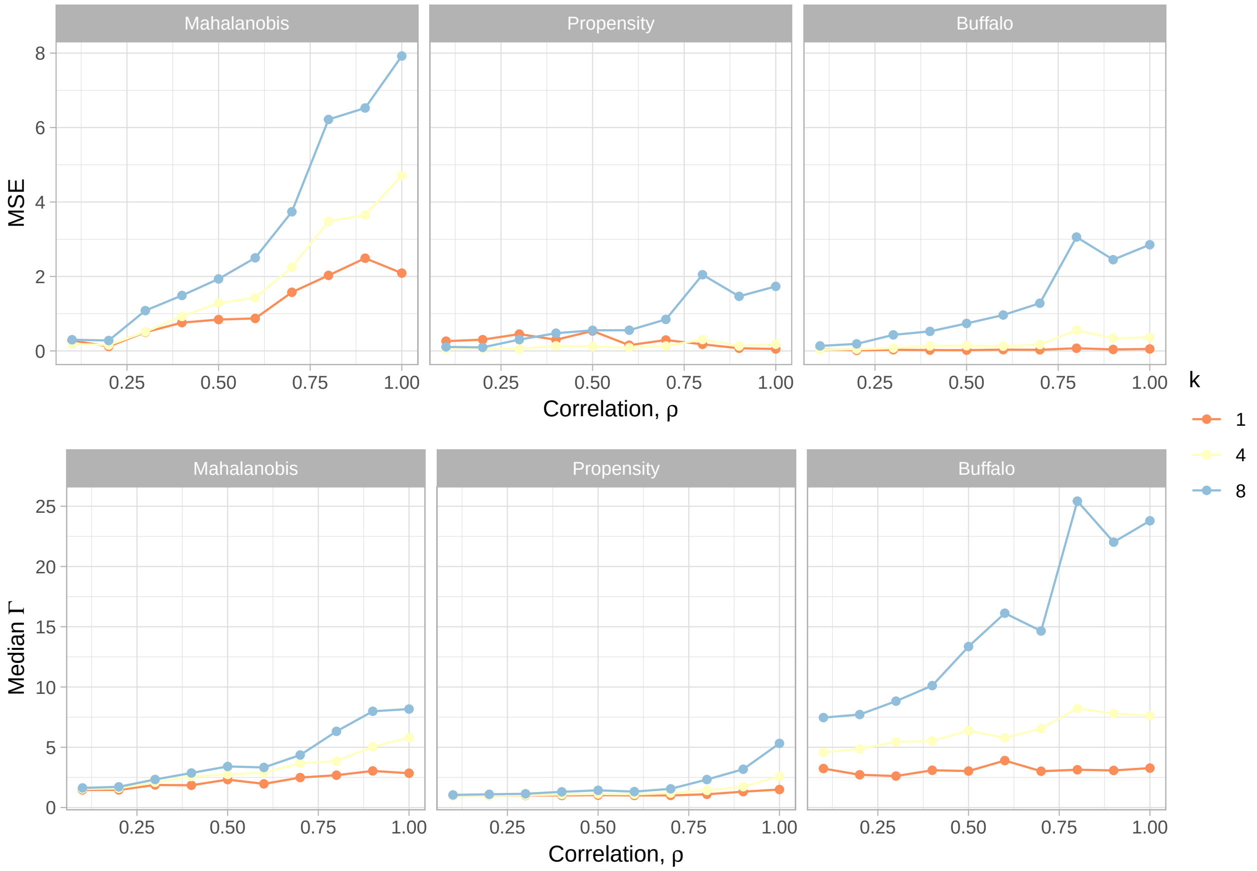
Psi times 6

Now suppose we emphasize these relationships even more:

$$\phi(X_i) = X_{i1} - 10/3,$$
$$\Psi(X_i) = 6\rho X_{i1} + 6\sqrt{(1-\rho^2)}X_{i2},$$

As we might expect, the behavior from the previous experiment is amplified. Bias goes up, but variance goes up more. Now the bias can be about 5 times the variance. Gamma for the buffalo skyrockets.


Mahalanobis
Propensity
Buffalo
Bias
0
1
2
0.25
0.50
0.75
1.00
Correlation, ρ
k
1
4
8
Standard.Deviation
0.2
0.4
0.6


Mahalanobis
Propensity
Buffalo
MSE
0
2
4
6
8
0.25
0.50
0.75
1.00
Correlation, ρ
k
1
4
8
Mahalanobis
Propensity
Buffalo
Median Γ
0
5
10
15
20
25
0.25
0.50
0.75
1.00
Correlation, ρ