# Main Figures

*Rachael Caelie (Rocky) Aikens*

*5/3/2019*

## 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

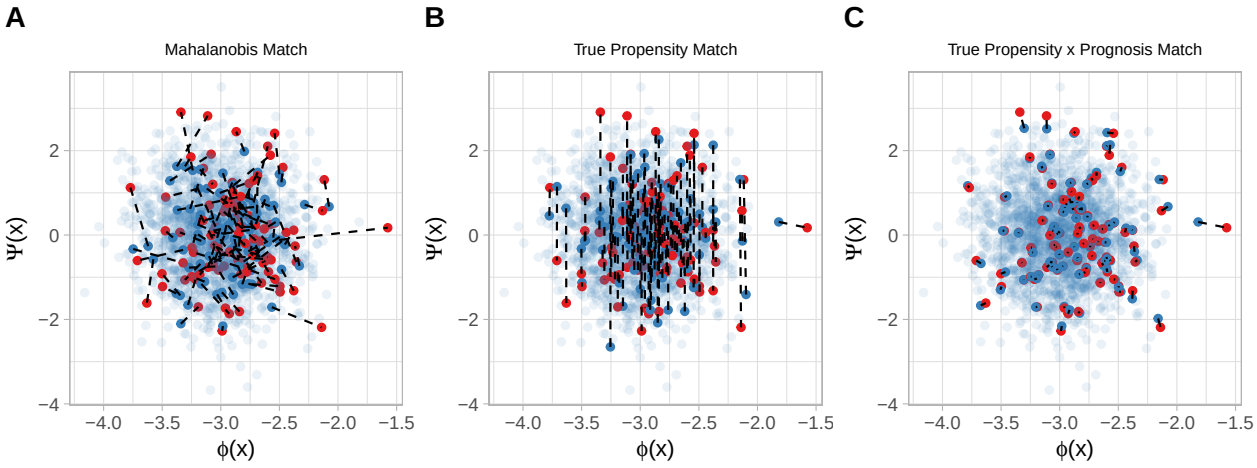
$$
\begin{aligned}
\phi(X_i) &= X_{i1}/3 - c, \\
\Psi(X_i) &= \rho X_{i1} + \sqrt{(1-\rho^2)} X_{i2},
\end{aligned}
$$

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. 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`.

# Motivation

The figure below gives a heuristic representation of what this algorithm is attempting to do. We imagine a scenario in which each individual in our data set, both treated and control, is represented in a reduced space of only two covariates: The variation determining the treatment assignment ($\phi(X_i)$), and the variation determining the outcome ($\Psi(X_i)$). These are the two features which are directly relevant to our matching: $\phi(X_i)$ balance (propensity balance) reduces bias, and $\Psi(X_i)$ balance (prognostic balance) reduces bias as well as variance and sensitivity to unobserved confounding.

In our simulation, $\phi(X_i)$ and $\Psi(X_i)$ are known linear combinations of the covariates (see set up), so we can visualize them directly. For simplicity, we assume that the prognosis and treatment assignment are entirely uncorrellated ($\rho = 0$), although this need not always be the case (See supplementary figures 1 and 2). Optimal mahalanobis distance matching (Figure 1A), pairs individuals who are closest in the full covariate space. However, since only $X_{i1}$ and $X_{i2}$ are important for prognosis and treatment assignment, individuals who are close in the full covariate space may be very distant in the feature space of $\phi(X_i)$ and $\Psi(X_i)$. Propensity score matching (Figure 1B) pairs individuals who are close in the axis important for treatment assigment, $\phi(X_i)$, but not for prognosis, $\Psi(X_i)$. This matching will reduce bias compared to the unmatched dataset, but will lose the protection from variance and unobserved confounding conferred by prognostic balance. In contrast, if we match jointly on $\phi(X_i)$ and $\Psi(X_i)$, we obtain individuals who are close together in the feature space below. This optimizes for both desirable types of covariate balance: prognostic and propensity.




```
## pdf
##   2
```

# 5 Results

## 5.1 Buffalo Performance with a large control reserve

Below, we use the simulation parameters described in the set up to estimate the bias, variance, mse, and median gamma sensitivity of effect estimates produced from Mahalanobis distance matching, propensity score matching, and Buffalo matching.

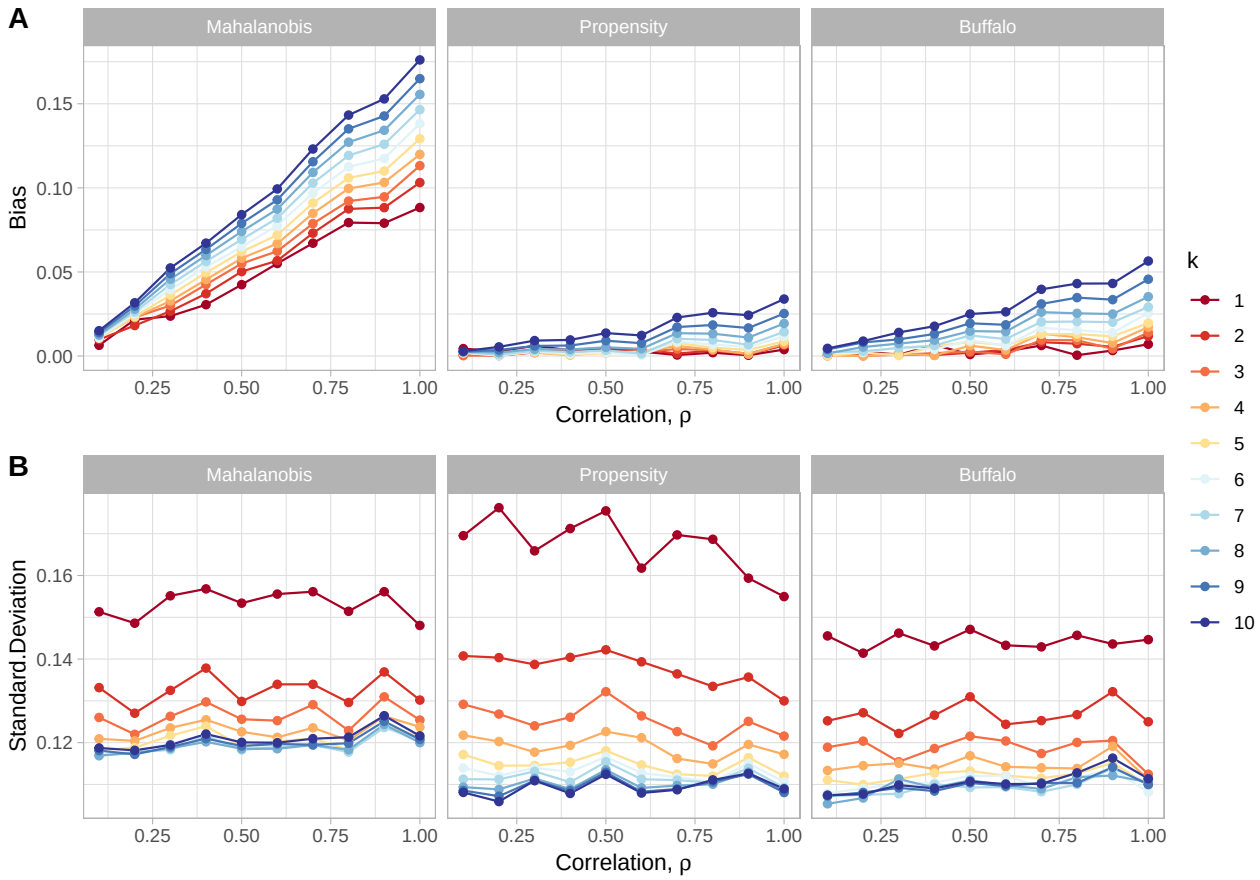



Figure 1 shows the bias and variance of $1:k$ matching for each method on the same simulated data set as $k$ increases (colored lines) and the correllation of propensity and prognosis, $\rho$ increases from left to right. When $\rho = 0$, the treatment effect is completely unconfounded, since treatment is entirely determined (up to randomness) by variation in $X_{i1}$, and outcome (under the control assignment) is entirely determined by variation in $X_{i2}$. When $\rho = 1$, the data is highly confounded, since outcome and treatment assignment are both determined solely by variation in $X_{i1}$.

A first observation is that the bias from Mahalanobis distance matching is large, and it increases with the correllation between prognosis and treatment assignment (Figure 1A). This is probably because, as $\rho$ approaches 1, only a single covariate, $X_{i1}$ is important to match on, yet Mahalanobis distance considers 9 other covariates which are entirely random noise unimportant to the problem. It thus selects worse matches with respect to the true covariate of interest. This problem is exacerbated when the number of uninformative covariates is increased (Supplementary Figure 3).

Figure 1B shows the protective effect of Buffalo matching against the variance of the estimator. Propensity score matching has highest variance, and this is worse when $\rho$ is close to zero. This is because, when $\rho$ is small, the propensity score contains no information about prognosis under the control assignment. This causes the propensity score method to match observations which may be very different in terms of their potential outcome under the control assignment, giving poor prognostic balance and thus larger variance in the estimate of the causal effect. While Mahalanobis distance performs slightly better, it is again choosing lower quality matches because of the uninformative covariates in the data. The lowest variance is achieved by Buffalo matching, since this algorithm optimizes for prognostic balance as well as propensity score balance.

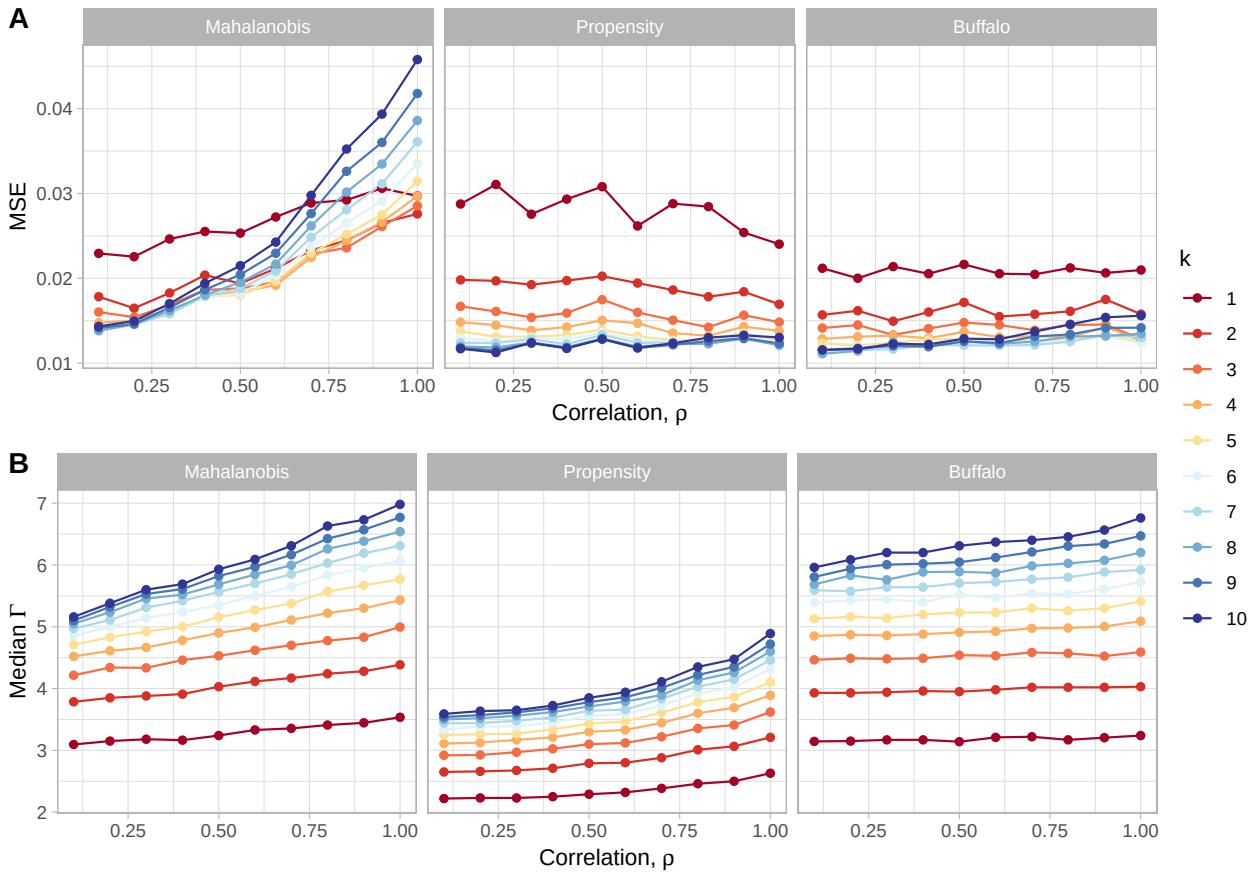



Figure 2 shows the mean squared error and median gamma sensitivity from the same set of simulations as above. Figure 2A demonstrates that Buffalo does well compared to the other two methods in terms of mean squared error, because of the protection against bias and variance shown in figure 1. Figure 2B shows the protective effect of prognostic balance against unobserved confounding. Propensity score matching gives lowest $\Gamma$ values, especially when $\rho$ is small. When this happens, the propensity score contains no prognostically relevant covariates, so the matches produced have poor prognostic balance. As $\rho$ increases, the prognostic score and the propensity score become highly correllated, so matching on propensity score starts to give some prognostic balance as well as propensity balance. While the Gamma sentistivity of Mahalanobis distance matching seems promising when $\rho$ is large, most of this is likely due to the large amount of bias in the effect estimate from this method (Figure 1A). Because Buffalo matching directly attempts to achieve prognostic balance in the matched sets, the Buffalo method has better protection against unobserved confounding, without giving a highly biased estimator.

## 5.2 Methodological Considerations

In the previous section, we illustrated a use case in which Buffalo Matching is very useful: There is an abundance of control individuals which overlap fairly well with the treated population, and the underlying processes dictating propensity and prognosis are easily fit with standard linear models. In this section, we consider four design considerations which are important to the selection of the method:

(1) Correlation of treatment and prognosis (confoundedness)
(2) Tradeoffs in sample size
(3) Tradeoffs in match quality
(4) Fitting the propensity and prognostic models

### 5.2.1 Correlation of treatment and prognosis

A first consideration is the confoundedness of the problem, modeled here as the correlation, $\rho$, between $\phi(X_i)$ and $\Psi(X_i)$. Figures 2 and 3 show the performance of each matching method as $\rho$, varies from 0 to 1. When this correllation is close to one, only the covariate $X_{i1}$ is important for treatment assignment ($\phi(X_i)$) and potential outcome under the control assignment ($\Psi(X_i)$). When this happens, all methods unsurprisingly have their greatest levels of bias (Figure 2A). Mahalanobis distance matching suffers the worst from this phenomemon because it is matching on several covariates which are unimportant to either treatment assignment or outcome.

One result worth remarking on is that propensity score matching is lowest variance when confoundedness is worse. This is because, when propensity and prognosis are highly correllated, matching on the covariates most important for treatment assignment also imposes balance in the covariate most important for prognosis. This imposes prognostic balance, protecting against the variance (Figure 2B) and increasing robustness to confounding (Figure 3B).

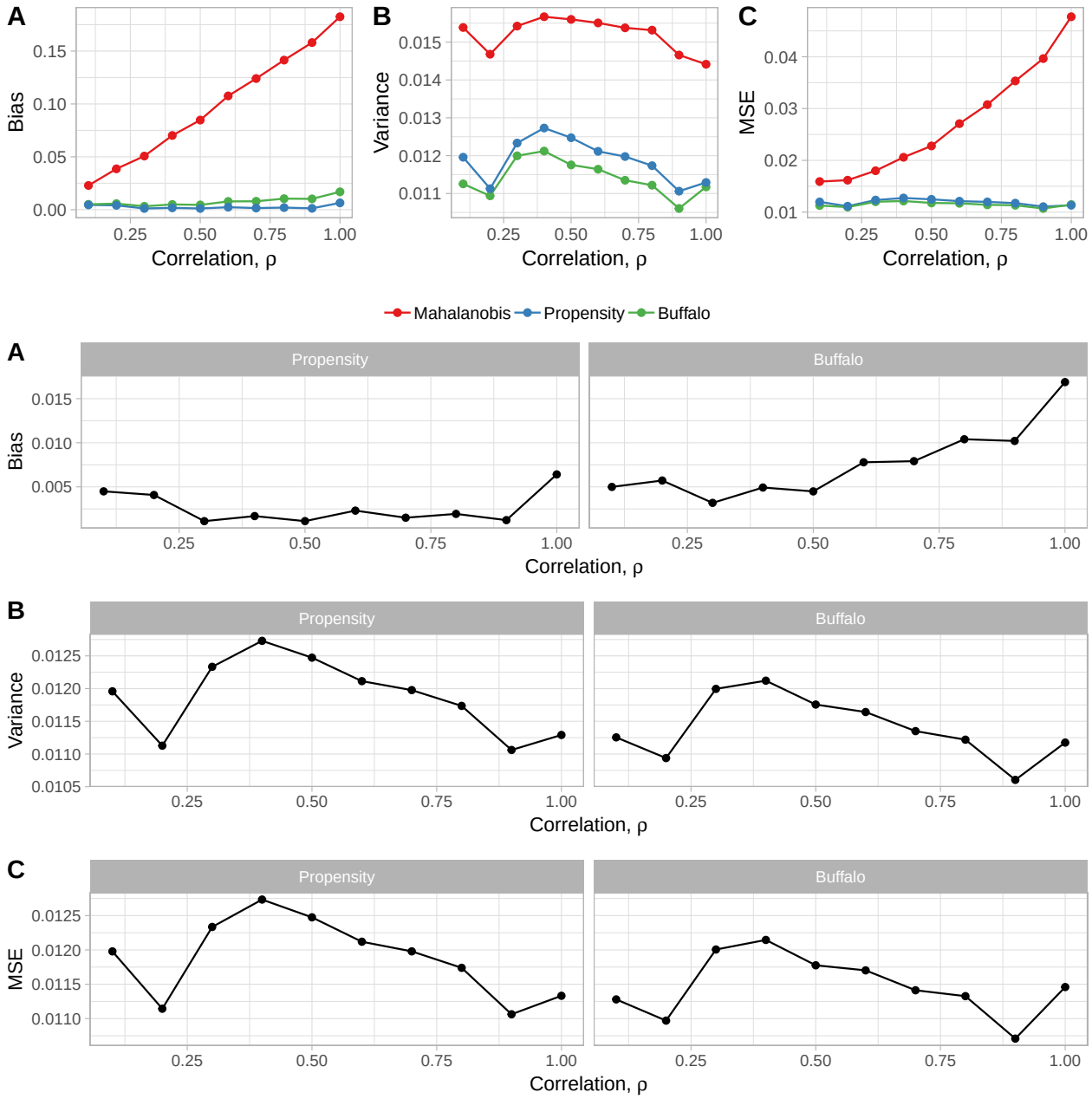
In contrast, when treatment assignment is less closely correlated with potential outcome, propensity score matching suffers most from issues of high variance in the effect estimate (Figure 2B), because ideal propensity score matches may be quite distant in terms of the covariates most important to outcome (Figure 1B). These are the scenarios in which buffalo matching is most protective against variance in estimation, since it imposes prognostic balance on the matched sets as well as propensity balance.

### 5.2.2 Tradeoff in sample size

A second consideration is the tradeoff in control sample size implicit in creating a held-aside set for fitting the prognostic score prior to matching. In the Buffalo Algorithm, one control individual for each treated observation is removed from the analysis data set for the purpose of fitting the prognostic score, amounting to a sacrifice of $\sim 100$ observations.

Section 5.2.3 considers the effect of this sacrifice on the quality of the available matches, and the influence that this has on the bias of the estimator. For the moment, let us consider the effect of this sacricice on the variance of the estimator, since removing this hold-out set means fewer control individuals may be matched to each treated individual (smaller $k$ in $1:k$ matching). Generally speaking, the central limit theorem implies that the variance of an estimator tends to decrease with $\frac{1}{n}$, where $n$ is the sample size. This means that decreasing the sample size by $\sim 100$ is likely to have a very large effect on variance when the original sample size is small, and a relatively small effect when the original sample size is much larger than 100. Thus, when the control reserve is much larger in size than the held aside sample, the loss of those observations from the analysis phase will have a small effect on sample variance.

Meanwhile, inducing prognostic balance in the analysis set has a variance-reducing effect for equivalent sample sizes (Figure 2B). This means that when the control reserve is very large (say, greater than 1500), it may in fact be variance-*reducing* to "sacrifice" some individuals from the analysis set so that the overall matched sets obtained in the analysis phase have better prognostic balance. This trade-off, of course, depends on the total size of the control reserve. In order to illustrate this, we simulated fullmatching on Mahalanobis Distance, Propensity score, and Buffalo Matching. Since fullmatch uses all of the treated and control individuals, the Buffalo fullmatch has ~100 fewer control individuals to use for estimation than the alternative methods. However, when the control reserve is large, the variance of the estimator from Buffalo fullmatching is actually smaller than that for the other methods (Figure 4).



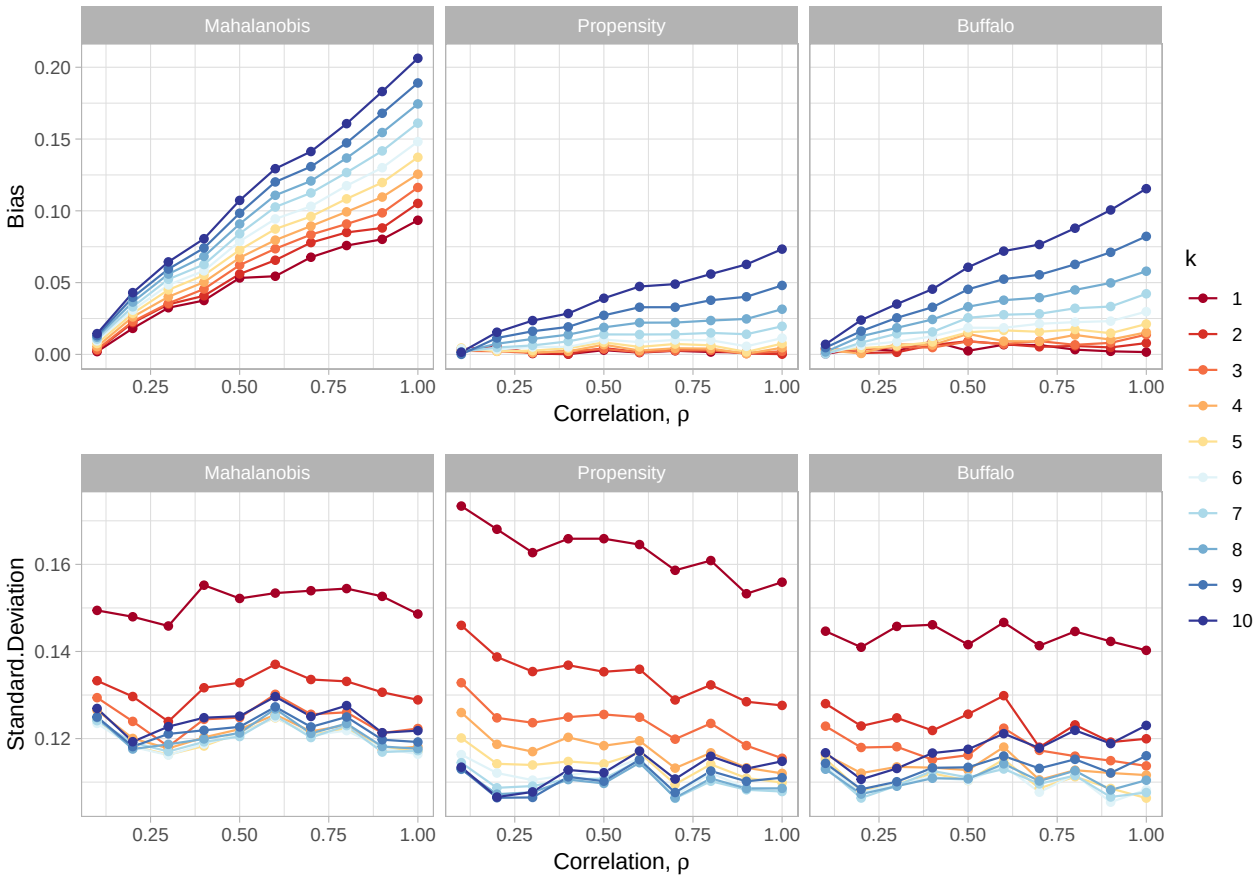

### 5.2.3 Tradeoff in match quality

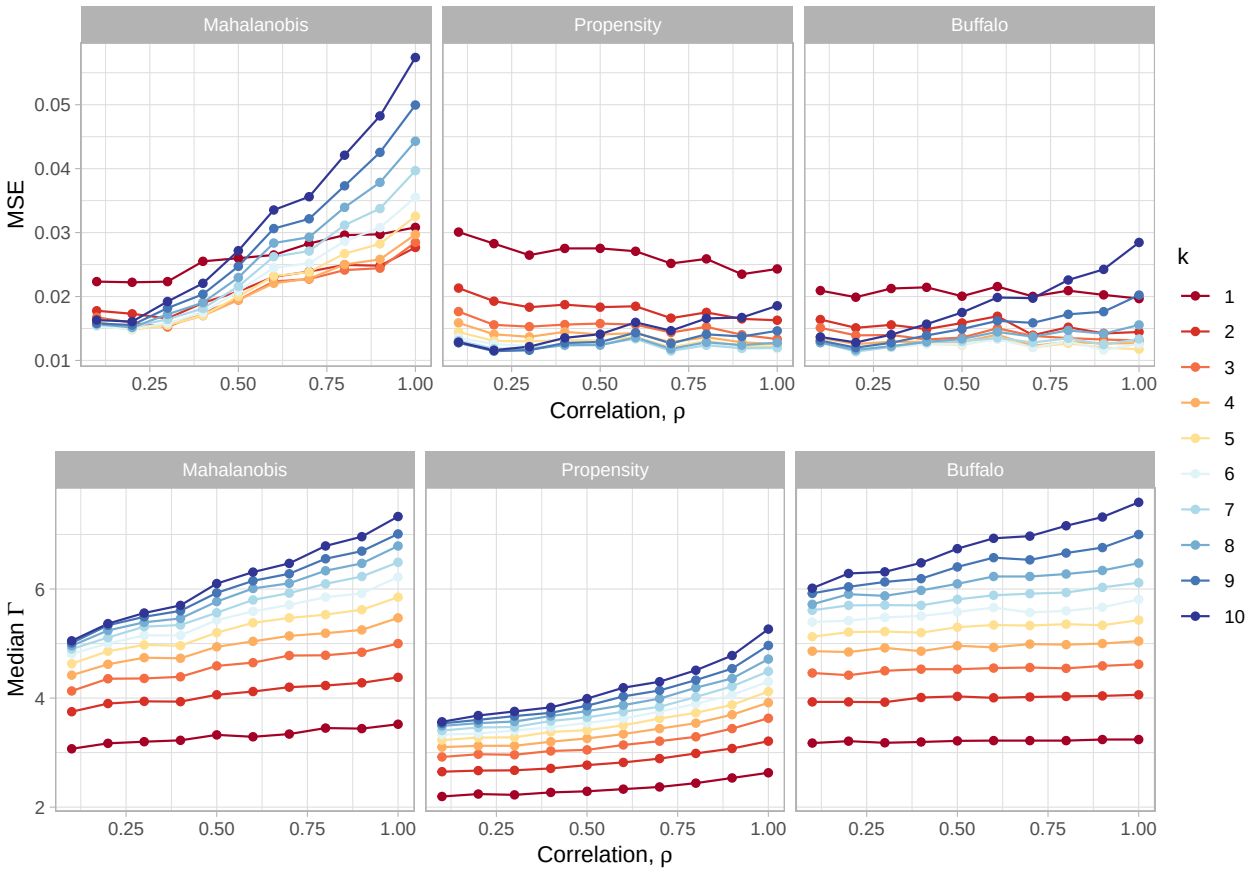
As opposed to the variance-reducing effect discussed above, Buffalo Matching does tend to have the effect of *increasing* bias because of the loss of individuals from the control reserve which might otherwise have been effective matches. In essence, some of the held-aside control individuals used to build the prognostic score may actually have been potential high-quality matches to the treated individuals. This can force the subsequent matching to compromise on lower quality matches, increasing bias and - in some cases - variance (see supplement). This effect is most pronounced when (1) k is large, (2) the control reserve is small, and (2) overlap of the treated and control individuals is poor (Figure 2A, Figure 5A). When $k$ is large, bias for all methods is worse because lower and lower quality matches are being carried over to the analysis phase. If the control reserve is small and/or there is little overlap between the treated and control individuals, every control individual which is close in covariate space to the treated individuals is a precious potential match, and sacrificing these individuals means that there are fewer good alternatives to choose from. Additionally,

we find that poor overlap combined with high confoundedness (large $\rho$) tends to increase the variance of the estimator for both Buffalo and propensity score matching (see Supplement). We hypothesize that this is because both methods are forced to chose matches which are more distant both in propensity and prognostic score. In particular, this prognostic imbalance induces increased variance in estimation.

As a result, we advise the following. First, the number, $k$, of treated individuals to be matched to each control individual, should always be chosen with the size of the control reserve and the overlap of treated and control observations, regardless of the particular matching distance specification used. When overlap is poor and the number of control individuals is sparse, increasing $k$ will increase the bias of all methods considered here. Second, the selection of the control individuals for the held-out set is a nuanced design decision (discussed further in the discussion section). Just as with the selection of $k$, the choice to use Buffalo Matching and the selection of the held-out set should depend on the data. Because of the sacrifice in match quality associated with discarding some individuals, Buffalo Matching is most appropriate when the control reserve is plentiful, particularly in the region of overlap between treated and control individuals.

One potential simplification of this issue is to use fullmatch rather than 1:k matching, since fullmatch deals elegantly with the problem of the ratios of treated and control individuals varying across the covariate space. Figure 5 shows the results of fullmatching using Mahalanobis distance, propensity score, and Buffalo.

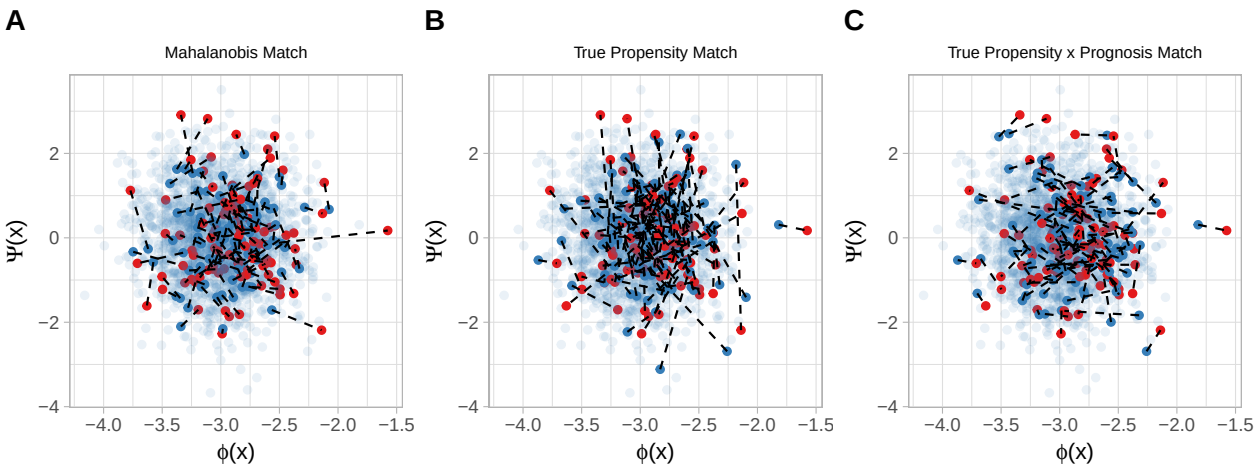





### 5.2.4 Fitting propensity and prognostic models

A final consideration is that of fitting the propensity and prognostic models themselves. This simulation was largely a proof of concept, with the true data generating processes for treatment assigment and prognosis fairly simple and easily fit by common linear modeling methodologies. In practice, the researcher may choose to draw from whatever model fitting methodologies they prefer to capture the prognostic score.

It is intuitive that the value of buffalo matching is greatest when the propensity and prognostic score models are most accurate. In contrast to figure 1, which shows the idealized matchings that would be selected if the propensity and prognostic scores were precisely known, figure 6 displays the matches which might be selected in practice. Both propensity and buffalo matching methods rely on the assumption that the prognosis and propensity for treatment are easy to estimate; in cases where the model is imperfect, imperfect matches may be selected. In fact, in simulations where more random variation is added to the outcome, we find that the performance of buffalo is diminished because the prognostic score is harder to fit (Supplement). In cases like these, it may be useful to budget more data for fitting the prognostic score (if able), or use more sophisticated modeling techniques. In cases where the control reserve is practically infinitely abundant, this may be an opportunity to leverage more data for the purposes of building a very accurate model for prognosis to be used in matching.


A
Mahalanobis Match
B
True Propensity Match
C
True Propensity x Prognosis Match
Ψ(x)
ϕ(x)
2
0
−2
−4
−4.0
−3.5
−3.0
−2.5
−2.0
−1.5

```
## pdf
##   2
```