Model Specification

Rachael Caelie (Rocky) Aikens

2/26/2020

Set Up

We compare the performance of propensity score matching, Mahalanobis distance matching, and pilot matching (described in the manuscript) 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:

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

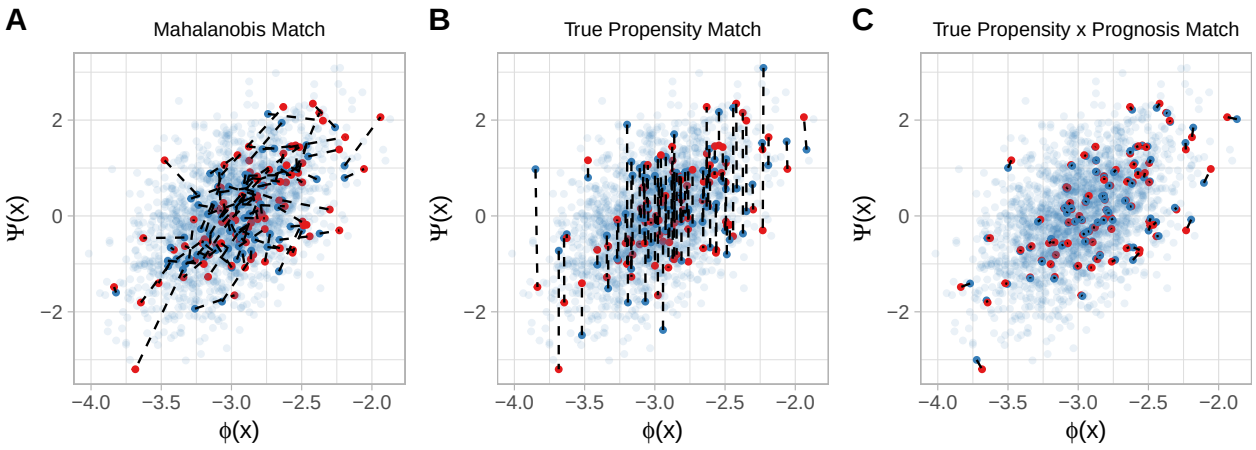
where the true propensity and prognositic scores are given by the linear combinations

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

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`.

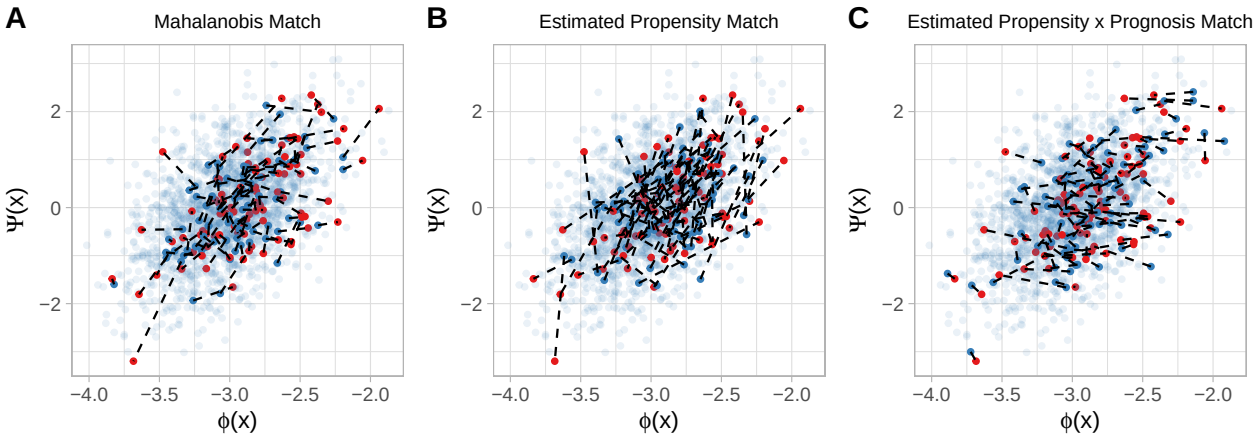
Fisher-Mill Plots

We're fairly familiar with this vizualization at this point. The figure below supposes that we knew the propensity and prognostic score for each individual in our data set, and shows - in Fisher-Mill space - what the optimal matches for each approach would be.



An overspecified model

Now lets suppose that we actually have to fit the propensity and prognostic score (i.e. we don't magically know them). In the simulations we've run thus far, we've regressed on all 10 covariates for the prognostic and propensity score models. This gives us substantially worse matches.



But, okay, a reasonable question to ask is: just how badly did we do in the model fitting? Keep in mind that, with $\rho = 0.5$ as it does in these demonstrations, the correct models are:

$$\phi(X_i) = X_{i1}/3 - c,$$
$$\Psi(X_i) = 0.5X_{i1} + 0.866X_{i2},$$

Below is a summary of our propensity and prognosis models. One thing that's interesting is that the propensity model is - like - kind of awful and the prognostic model is quite good, in spite of the fact that the propensity model is fit on the entire data set while the prognostic model is fit on only 100 controls. This may have something to do with the fact that the treatment is binary while the outcome is continuous.

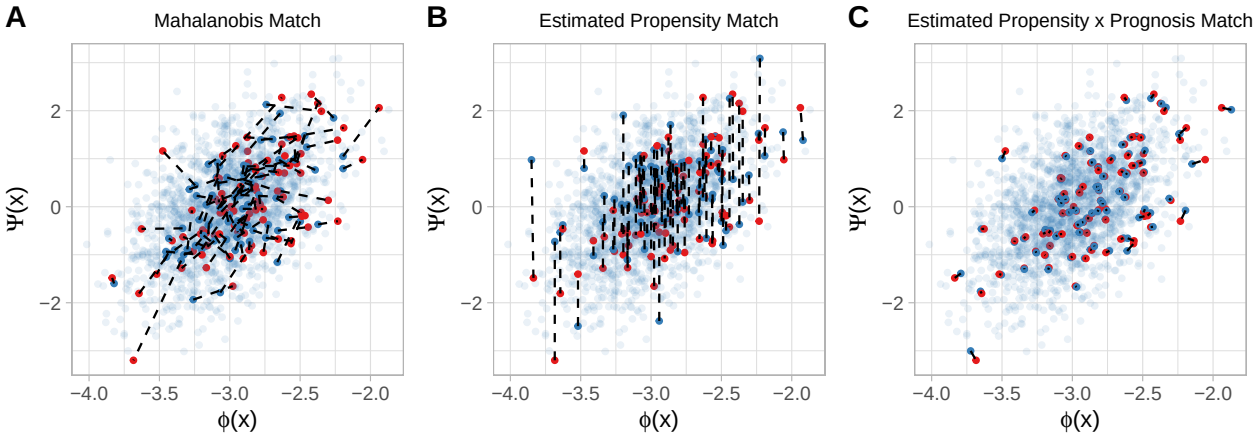
```
##
## Call:
## lm(formula = prog_model, data = dplyr::select(selected, -c(row,
##     m)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.87111 -0.57724  0.05925  0.49315  2.26086
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04049    0.10763  -0.376    0.708
## X1           0.55423    0.11890   4.661 1.26e-05 ***
## X2           0.89876    0.13528   6.644 3.56e-09 ***
## X3           0.11196    0.10603   1.056    0.294
## X4           0.05634    0.14660   0.384    0.702
## X5           0.02048    0.11171   0.183    0.855
## X6          -0.06831    0.13178  -0.518    0.606
## X7          -0.06422    0.13106  -0.490    0.626
## X8          -0.10345    0.10328  -1.002    0.320
## X9          -0.08580    0.13191  -0.650    0.517
```

```
## X10         -0.04072    0.12069  -0.337    0.737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9324 on 79 degrees of freedom
## Multiple R-squared:  0.5463, Adjusted R-squared:  0.4889
## F-statistic: 9.512 on 10 and 79 DF,  p-value: 3.471e-10
```

```
##
## Call:
## glm(formula = prop_model, family = binomial(), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6402  -0.3362  -0.2776  -0.2222   2.9320
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.219299   0.124208 -25.919  < 2e-16 ***
## X1           0.423163   0.110271   3.837 0.000124 ***
## X2          -0.095852   0.111160  -0.862 0.388531
## X3          -0.166837   0.110241  -1.513 0.130180
## X4          -0.061486   0.110594  -0.556 0.578238
## X5           0.130048   0.107966   1.205 0.228386
## X6           0.002099   0.108142   0.019 0.984512
## X7           0.237647   0.110517   2.150 0.031530 *
## X8          -0.040052   0.110051  -0.364 0.715906
## X9          -0.148727   0.108799  -1.367 0.171631
## X10         -0.062572   0.110450  -0.567 0.571039
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 734.08  on 1999  degrees of freedom
## Residual deviance: 706.71  on 1989  degrees of freedom
## AIC: 728.71
##
## Number of Fisher Scoring iterations: 6
```

A correctly specified model

Let's do something crazy and give the propensity and prognosis models the correct specification. That is, the propensity model should regress `t ~ X1` and the prognostic model should regress `t ~ X1 + X2`



But, okay, a reasonable question to ask is: just how badly did we do in the model fitting? Keep in mind that, with $\rho = 0.5$ as it does in these demonstrations, the correct models are:

$$\phi(X_i) = X_{i1}/3 - c,$$
$$\Psi(X_i) = 0.5X_{i1} + 0.866X_{i2},$$

Below is a summary of our propensity and prognosis models. One thing that's interesting is that the propensity model is - like - kind of awful and the prognostic model is quite good, in spite of the fact that the propensity model is fit on the entire data set while the prognostic model is fit on only 100 controls. This may have something to do with the fact that the treatment is binary while the outcome is continuous.

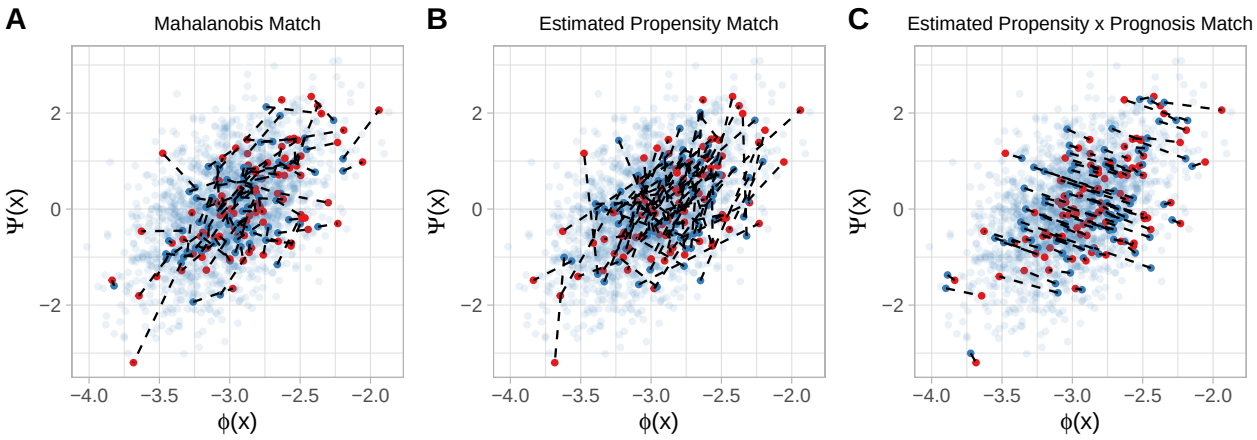
```
##
## Call:
## lm(formula = prog_model, data = dplyr::select(selected, -c(row,
##     m)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8215 -0.7511 -0.0364  0.6312  2.0764
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0961     0.1047  -0.918    0.361
## X1            0.6864     0.1181   5.811 1.00e-07 ***
## X2            0.8337     0.1269   6.570 3.54e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9552 on 87 degrees of freedom
## Multiple R-squared:  0.5223, Adjusted R-squared:  0.5114
## F-statistic: 47.57 on 2 and 87 DF,  p-value: 1.1e-14

##
## Call:
```

```
## glm(formula = t ~ X1, family = binomial(), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5799  -0.3317  -0.2864  -0.2457   2.9121
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.1508     0.1173 -26.850  < 2e-16 ***
## X1            0.4286     0.1089   3.936 8.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 734.08  on 1999  degrees of freedom
## Residual deviance: 718.41  on 1998  degrees of freedom
## AIC: 722.41
##
## Number of Fisher Scoring iterations: 6
```

Just for fun: Double-robustness

We have the theoretical result that matching jointly on propensity and prognosis is doubly robust. That is, as long as at least one of the models is correctly specified, the joint-matching approach will give you consistent inference.

This is kind of a can of worms (since we don't really consider different model mis-specification scenarios in our simulations), but let's see if we can visualize this with our Fisher-Mill plots. Below is a visualization of the matches we would have chosen if our propensity score model was fit on all 10 covariates, but our prognostic model was correctly specified.



And below is the reverse scenario: in which our propensity score model is right on the money and our prognostic score model is heavily overspecified.

```
##
## Call:
## lm(formula = prog_model, data = dplyr::select(selected, -c(row,
##     m)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56877 -0.55603 -0.09361  0.54247  2.43525
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.16585    0.10419  -1.592  0.11543
## X1           0.46325    0.10234   4.526 2.09e-05 ***
## X2           0.74733    0.12324   6.064 4.30e-08 ***
## X3           0.06494    0.09637   0.674  0.50238
## X4           0.01167    0.13132   0.089  0.92938
## X5          -0.02261    0.10928  -0.207  0.83664
## X6           0.11569    0.11370   1.017  0.31203
## X7          -0.35574    0.11981  -2.969  0.00395 **
## X8          -0.07555    0.10088  -0.749  0.45613
## X9           0.13574    0.11384   1.192  0.23666
## X10         -0.32640    0.11334  -2.880  0.00512 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8997 on 79 degrees of freedom
```

```
## Multiple R-squared:  0.5462, Adjusted R-squared:  0.4887
## F-statistic: 9.507 on 10 and 79 DF,  p-value: 3.513e-10
```