Causal Inference Final Project:

CBPS and Entropy Balancing - A Simulation Study

Chansoo Song

December 2018

Motivation:

A key challenge in the application of propensity scores for matching is that the propensity score is unknown and must be estimated. To make matters worse, slight misspecification of the propensity score model can lead to substantial biases in treatment effects. This has led to researchers iteratively re-estimating the propensity score model, subsequently checking the resulting covariate balance, then repeating over and over until they are satisfied. Imai et al (2008) [1] calls this the 'propensity score tautology': the estimated propensity score is appropriate if it balances covariates.

In this simulation study, I analyze two approaches that seek to bypass this 'propensity score tautology': Covariate Balancing Propensity Score and Entropy Balancing. Each method obviates the need for iteratively re-estimating the propensity score model and checking balance on the covariate moments. That is, a single model is used to estimate both the treatment assignment mechanism and the covariate balancing weights.

Matching, Propensity Scores and Assumptions:

In an observational study setting where the confounding covariates (variables correlated with both treatment and outcome) are known and measured, we may use matching methods to ensure that there is sufficient overlap and balance on these covariates. Then, we can estimate the treatment effect using a simple difference in means or regression methods.

Overlap is important because we want to make sure that for each treated or control subject in the study, there exists an empirical counterfactual (this criteria varies depending on the

[1] Imai, K., King, G. and Stuart, E. A. (2008)

estimand of interest, i.e. to estimate the ATT it is sufficient to have empirical counterfactuals for just the treated subjects in the study). **Balance** on the covariates is important because imbalance would force us to rely more on the correct functional form of the model.

There are many different matching methods, but the driving principle is to identify observations that are "most similar" based on some distance metric. Methods include: K-nearest-neighbor, caliper-matching, kernel-matching, Mahalanobis matching, Genetic Matching, Optimal Matching.

A **propensity score** is a one-number summary of the covariates. Rosenbaum and Rubin (1983) define the propensity score for participant *i* as the conditional probability of treatment assignment $(Z_i = 1)$ given a vector of observed covariates: $e(X_i) = Pr(Z_i = 1|X_i)$. The most common traditional approaches to estimating the propensity score are logistic regression and probit regression.

If strong **ignorability** holds after conditioning on the propensity score, that is:

$$y_0, y_1 \perp Z_i|e(X_i), \quad 0 < e(X_i) < 1$$

Then we may obtain an **unbiased** estimate of the treatment effect by either matching or weighting using just the propensity score instead of the vector of covariates.

After using either matching, propensity scores, or both to obtain a subset of the data that exhibits sufficient overlap, simple mean differences or a linear regression using weights can be used to estimate the treatment effect (ATE, ATC, or ATT). In all cases, ignorability, sufficient overlap, appropriate specification of the propensity score model / good balance, and SUTVA [2] are all important assumptions to obtain unbiased estimates of the treatment effect.

To summarize assumptions, propensity score and matching methods require that the **structural** assumptions of ignorability and SUTVA are met. And to a lesser degree make **parametric** assumptions: correct specification of the propensity score model. Theoretically, in some cases, sufficient overlap and balance may make the outcome estimation model robust to misspecification and thereby helps to relax the parametric assumptions.

[2] The Stable Unit Treatment Value Assumption (SUTVA) states that the potential outcomes are independent of the particular configuration of treatment assignment. In other words, there there is no dilution or concentration of treatment effects.

Describe the Designs / Estimators:

Covariate Balancing Propensity Score (CBPS)[3]:

The CBPS exploits the dual characteristics of the propensity score as a covariate balancing score and the conditional probability of treatment assignment. First, consider a commonly used model for estimating propensity scores: logistic regression (point of this part is to show the dual characteristics!):

$$e_B(X_i) = \frac{exp(X_i^T \beta)}{1+exp(X_i^T \beta)}$$

We typically estimate the unknown parameters by maximum likelihood:

$$\hat{\beta}_{MLE} = arg\max_{\beta} \sum_{i=1}^{N} Z_i \ \ log\{e_B(X_i) \ \ + (1-Z_i) \ \ log\{1-e_B(X_i)\}$$

And we get the ML estimates by differentiating the log likelihood with respect to the parameters then setting the derivative to zero. So differentiating with respect to β, we get:

$$\frac{1}{N} \sum_{i=1}^{N} \frac{Z_i \ e'_B(X_i)}{e_B(X_i)} - \frac{(1-Z_i) \ e'_B(X_i)}{1-e_B(X_i)}$$

Then, they operationalize the covariate balancing property by using inverse propensity score weighting:

$$E \left\{ \frac{Z_i \ \tilde{X}_i}{e_B(X_i)} - \frac{(1-Z_i)\tilde{X}_i}{1-e_B(X_i)} \right\} = 0$$

where $\tilde{X}_i = f(X_i)$, a function of X_i specified by the researcher. **Which happens to look a lot like the difference between treatment weights and control weights under inverse propensity score weighting (IPSW)!** (If we substitute Y_i for $\tilde{X}_i$, we get exactly the difference between the inverse propensity score weighted treated and control outcomes.) Recall that the inverse propensity score weights are used to make the treated group "look like" the control group. And here, the weighting provides a condition that balances a particular function of covariates (i.e. the mean or variance). Setting $\tilde{X}_i = e'_B(X_i)$ gives more weights to covariates that are predictive of

[3] Imai, Kosuke, and Marc Ratkovic. (2014)

treatment assignment according to the logistic regression propensity score model. Setting $\tilde{X}_i = X_i$ ensures the first moment of each covariate is balanced. Setting $\tilde{X}_i = (X_i^T X_i^{2T})^T$ ensures the first and second moment of each covariate is balanced. Hence, we've established the "dual" characteristics of the propensity score as a covariate balancing score and conditional probability of assignment.

To estimate the CBPS, Imai uses the GMM or EL framework. For more details please see Imai and Ratkovic (2014).

Entropy Balancing[4]:

Entropy balancing similarly involves a reweighting scheme that directly incorporates covariate balance into the weight function. To do this, entropy balancing searches for a set of weights that satisfies the balance constraints, while trying to keep the distribution of weights as uniform as possible (i.e. minimizing the divergence of distribution of weights from a uniform distribution). Thus, entropy balancing (1) allows us to obtain a high degree of covariate balance (using balance constraints that can involve the first, second, and possibly higher moments of the covariate distributions as well as interactions). And (2) allows for a more flexible reweighting scheme that seeks to retain as much information as possible.

Consider the reweighting scheme used to estimate the Average Treatment Effect on the Treated (ATT). We would want to estimate the counterfactual mean by:

$$E[Y(0)\hat{|}Z = 1] = \frac{\sum_{i|Z=0} Y_i w_i}{\sum_{i|Z=0} w_i}$$

where w_i is a weight for each control unit.

The weights are chosen by the following reweighting scheme:

$$H(w) = \sum_{i|D=0} h(w_i)$$

[4] Hainmueller, Jens. 2012.

where $h(.)$ is a distance metric and $c_{ri}(X_i) = m_r$ describes a set of R balance constraints imposed on the covariate moments of the reweighted control group.

Minimize $H(w)$ subject to the balance and normalizing constraints:

$$\sum_{i|D=0} w_i c_{ri}(X_i) = m_r$$

with $r \in 1, ..., r$

$$\sum_{i|Z=0} w_i = 1$$

and $w_i \geq 0$ for all i such that $T = 0$.

Comparison and implementation:

Both methods may be used to estimate either the ATT, ATC, or the ATT, and the two methods are very similar. The key difference is that entropy balancing bypasses the 'propensity score tautology' by ignoring the propensity score model estimation step. Instead, it looks for weights that achieve the best balance, subject to a constraint that seeks to retain as much information in the data as possible. In contrast, covariate balancing directly exploits the dual characteristic to estimate propensity scores AND balance covariates simultaneously.

For implementation, I use the 'ebal' and 'CBPS' packages in R to implement Entropy Balancing and Covariate Balancing Propensity Score, respectively.

Simulation Set Up:

Features:

In this section, I examine whether the CBPS or Entropy Balancing methods improve upon the performance of baseline approaches to both (1) achieving balanced covariates and (2) estimating the treatment effect. The baseline approach estimates include (1) propensity scores using logistic regression and matches using 1-1 matching with replacement and (2) mahalanobis matching with replacement.

Though ignorability may be the most crucial assumption, I assume that all of the confounders are known to the researcher for all simulations. I believe that testing the sensitivity to the ignorability assumption would be more interesting when comparing propensity score and matching methods to other causal inference models. Since one of the more interesting aspect of these models is that they are less dependent on correct model specification compared to traditional propensity score approaches, I'm most interested in:

(1) reliance on the correct specification of the propensity score model

(2) reliance on the correct specification of the outcome model

(3) reliance on the ellipsoidally symmetric shape of covariate distributions

It's clear from earlier discussion why reliance on correct specification is important. I add feature (3) because Mahalanobis distance and propensity score matching may make balance worse if the methods are not EPBR (equal percent bias reducing). But both methods are equal percent bias reducing if all of the covariates used have ellipsoidal distributions. [5]

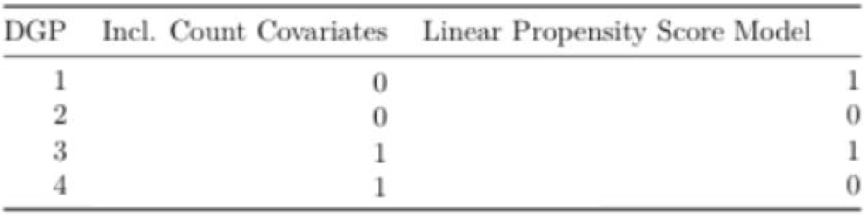
Estimand:

The estimand of interest is the ATT: Average Effect of Treatment on the Treated. We estimate the ATT by comparing the observed outcomes to the counterfactual outcomes that we would have measured had this group of subjects not received treatment. But since we are not able to observe this counterfactual state, we match each of these treated individuals to a control subject. In general, since the treatment group usually represents a specific subset of the general population, we would expect that the range of each covariate for the treated is a subset of the range of the same covariate for the general population. Then there will likely be some overlap so it makes sense to try to use matching and propensity score techniques to achieve good balance. On the other hand, estimating the treatment effect on the control may present a greater challenge. In general, most studies are concerned with a specific treatment on a specific group of people. There would be a higher risk of overlap issues and this would prevent extrapolating inferences based on this specific group to entirely different groups of people.

[5] Diamond, A. and Sekhon, J. (2012) and Rubin (1976)

DGPs and other simulation parameters:

I consider four data generating processes:



DGP	Incl. Count Covariates	Linear Propensity Score Model
1	0	1
2	0	0
3	1	1
4	1	0

(1) Standard Normally Distributed Covariates: The pre-treatment covariates $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})$ are four independent and identically distributed random variables following a standard normal distribution. The true propensity score model is a logistic regression whose linear predictor is a linear transform of the pre-treatment covariates.

(2) Standard Normally Distributed Covariates (non-linear propensity score model): The pre-treatment covariates are the same as Simulation #1; however, the true propensity score model is a logistic regression whose linear predictor are non-linear transforms of the pre-treatment covariates: $X_i^* = (-exp(X_{i1}/2),\ -X_{i2}/(1 + exp(X_{i1})),\ X_{i3},\ -sqrt(X_{i4}^2))$

(3) Standard Normally Distributed Covariates + 3 count covariates: The pre-treatment covariates $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6}, X_{i7})$ consist of four independent and identically distributed random variables following a standard normal distribution, a random variable following a poisson distribution with $\lambda = 1$, the negative values of a random variable following a binomial distribution with $n = 3$ and $p = 0.8$, and a random variable following a chi-squared distribution with $df = 1$.

(4) Standard Normally Distributed Covariates + 3 count covariates (non-linear propensity score model): The pre-treatment covariates are the same as Simulation (5); however, the true propensity score model is a logistic regression whose linear predictor are non-linear transforms of the pre-treatment covariates:

$$X_i^* = (0.5 * exp(X_{i1}/2),\ \ X_{i2}/(1 + exp(X_{i1})),\ \ -.2 * X_{i3}^2,\ \ X_{i1} * X_{i4},\ \ -0.4 * sqrt(X_{i5} - X_{i6}),\ \ 0.2 * (X_{i1} + 1.2 * X_{i6})^2,\ \ 0.5 * X_{i7})$$

I run each DGP twice, for a total of 8 simulations. For the first set of four simulations, the true outcome model is a linear regression with the pre-treatment covariates as predictors. For the second set of four simulations, the true outcome model is a linear regression with non-linear transformations of the pre-treatment covariates as predictors. I use the following non-linear model:

$$y_i = x_1^2 + x_1 x_2 + x_3^2 + \sqrt{x_4}$$

Baseline Matching Methods

Here, I briefly review the baseline models, then the specific specifications of the CBPS and EB models used in this simulation. For all six methods, the target estimand is the ATT and I match accordingly (that is, treated subjects receive weights equal to one and control subjects receive adjusted weights).

1. Baseline: Propensity Score using Logistic Regression:
2. Baseline: Mahalanobis Matching
3. CBPS (1)
4. CBPS (2)
5. EB (1)
6. EB (2)

The first baseline model uses logistic regression (without any interactions or transformations) to estimate propensity scores. I match using 1-1 nearest neighbor matching using the propensity scores.

The second baseline model uses Mahalanobis matching. Mahalanobis calculates distance as $m^2 = (x_T - x_C)'\Sigma_{CR}^{-1}(x_t - x_C)$ and it is equivalent to Euclidean matching based on standardized and orthogonalized X. To estimate the ATT: for each treatment subject, I match with replacement the control subject with the smallest Mahalanobis distance. Mahalanobis was intended for use with multivariate normally distributed data. When some covariates exhibit extreme outliers or very skewed distributions, Mahalanobis distance will place less weight on that covariate. On the other hand, a binary variable with a .99 probability of one would have low standard deviation and the Mahalanobis distance would give greater weight to this variable.[6] One way to address these concerns would be to use a rank-based Mahalanobis distance. For this simulation study, I use the standard Mahalanobis distance.

The third and fourth models use CBPS: an over-identified model and a just-identified model. The over-identified model (#3) combines the propensity score AND covariate balancing conditions. The just-identified model (#4) only contains covariate balancing conditions.

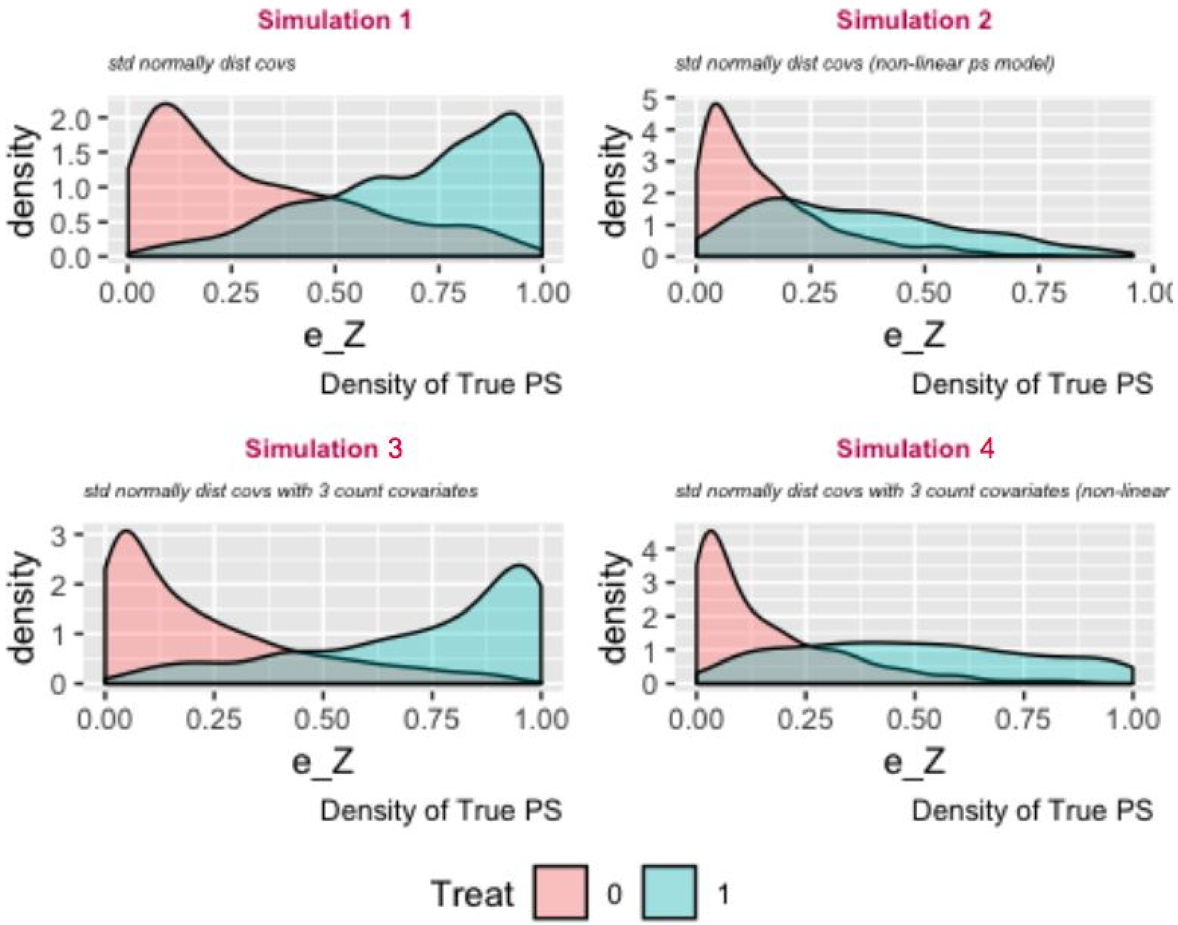
The fifth and sixth models use Entropy Balancing: one that achieves balance on just the first moment (#5) and one that achieves balance on both first and second moments (#6).

ATT Estimation Method

For all 6 models, I use a linear regression using (1) weights to reflect the restricted dataset of the corresponding matching method and (2) all observed covariates (without any interactions or transformations) to estimate the ATT.

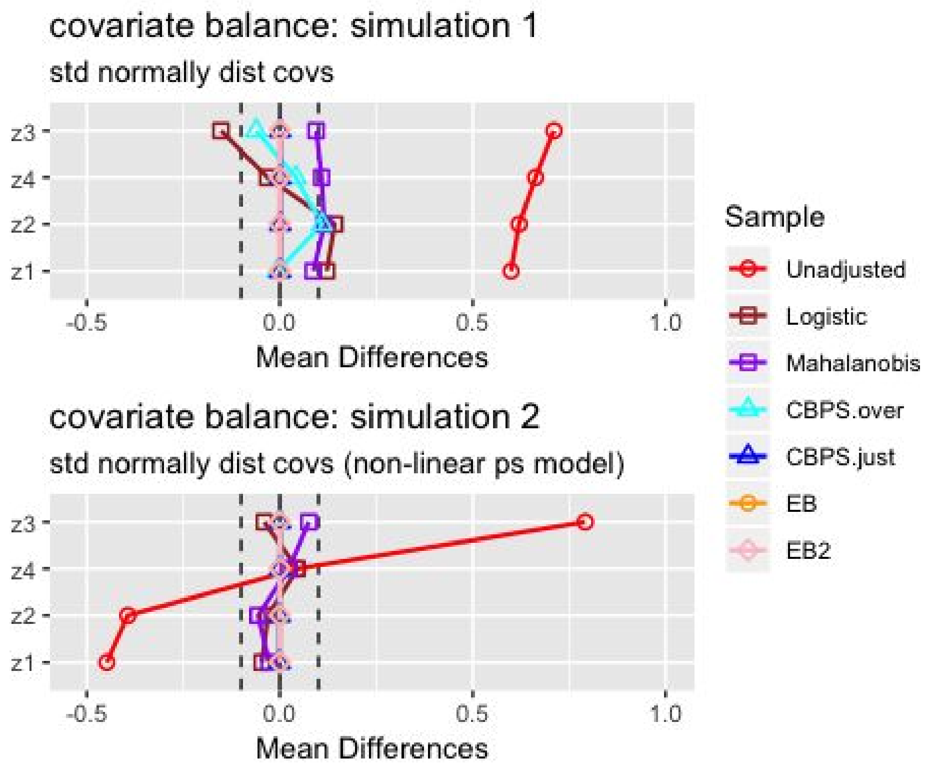
[6] Lecture notes from our causal class + lecture notes from this stats class at Penn: "www-stat.wharton.upenn.edu/~dsmall/stat921-f09/handouts/notes14_stat921_f09.doc"

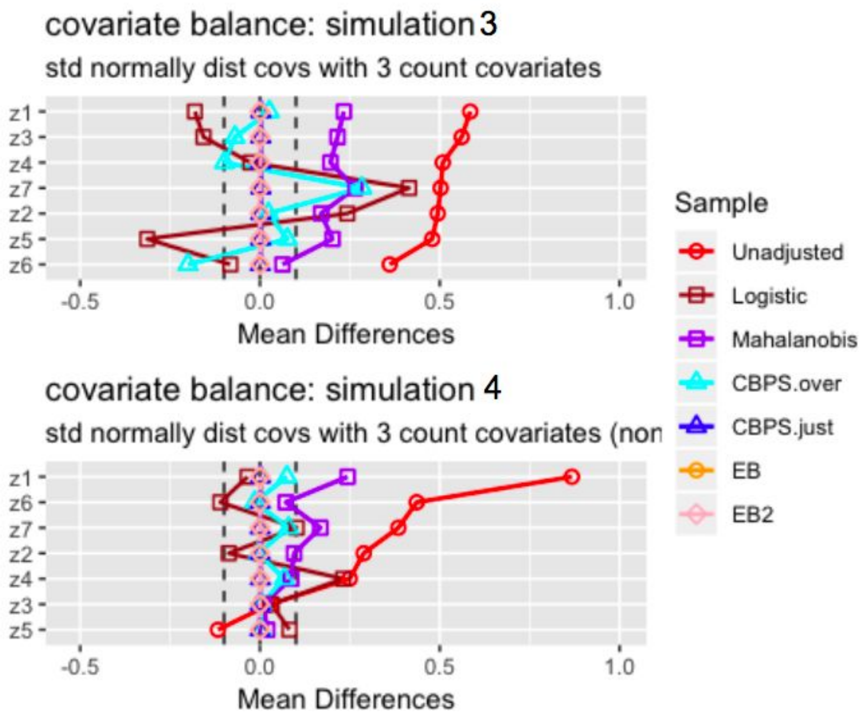
Simulation Results:



The above plots show the density of the true propensity score in treatment and control groups for each simulation. There is a misleading pattern: the linear propensity score models (#1 and #3) have strong separation, whereas the non-linear propensity score models (#2 and #4) show a platykurtic treatment distribution with positively skewed control density. However, this is totally arbitrary and is reflective of the specification of the true propensity score model. But this is ok because I’m interested in comparing the *relative* performances of the models *given* a simulation.

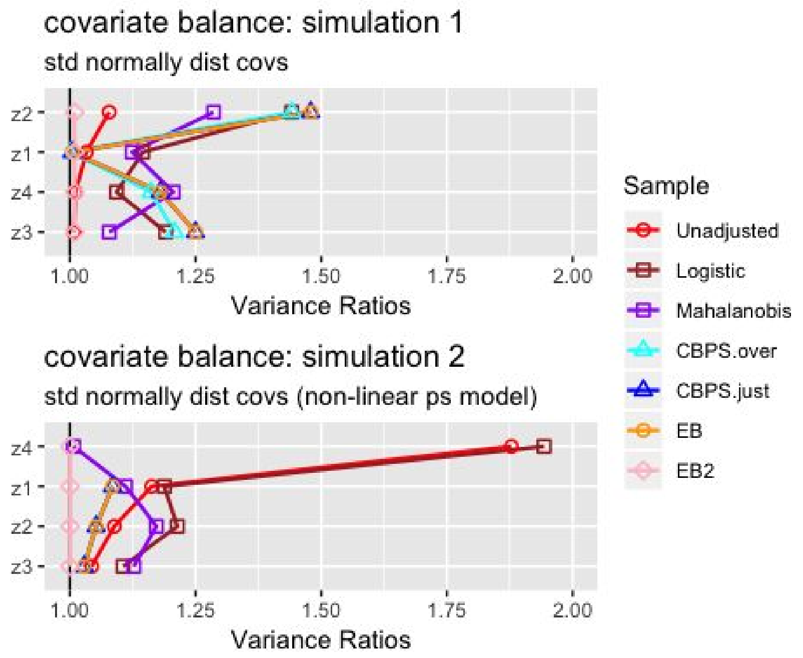
Examine Overlap of Propensity Score and Covariates (Before Matching)





Analysis of Mean Differences:

- Entropy balancing shows the best performance with respect to balancing covariate means. For all 4 simulations, the standardized mean difference is approximately zero for all covariates.
- The CBPS just-identified model similarly achieves perfectly balanced covariate means.
- The CBPS over-identified model performs significantly better in simulations where the true propensity score model is non-linear and worse in simulations where the true propensity score model is linear, which seems counter-intuitive. In fact, for both simulations with a non-linear true propensity score model, the CBPS over-identified model achieves nearly 0 mean difference for all covariates, where as mean differences remain large for both simulations with linear true propensity score model. One observation is that when the true propensity score model is linear, the covariates' mean differences are similar to the covariates' mean differences under the logistic model. Recall that the over-identified model combines the propensity score and covariate balancing conditions whereas the just-identified model only contains covariate balancing conditions. It seems likely that when the true propensity score model is linear in the covariates, the propensity score condition "dominates" the covariate balancing conditions, so the CBPS over-identified model's performance resembles the logistic baseline model's results. For the simulations with a non-linear propensity score model, the propensity score condition no longer dominates, so the CBPS over-identified model's performance resembles the just-identified model's results.
- The logistic regression and mahalanobis matching methods show small mean differences in simulation 1, where the pre-treatment covariates have standard normal distributions. Performance appears to weaken after including count variables and when the true propensity score model is non-linear.

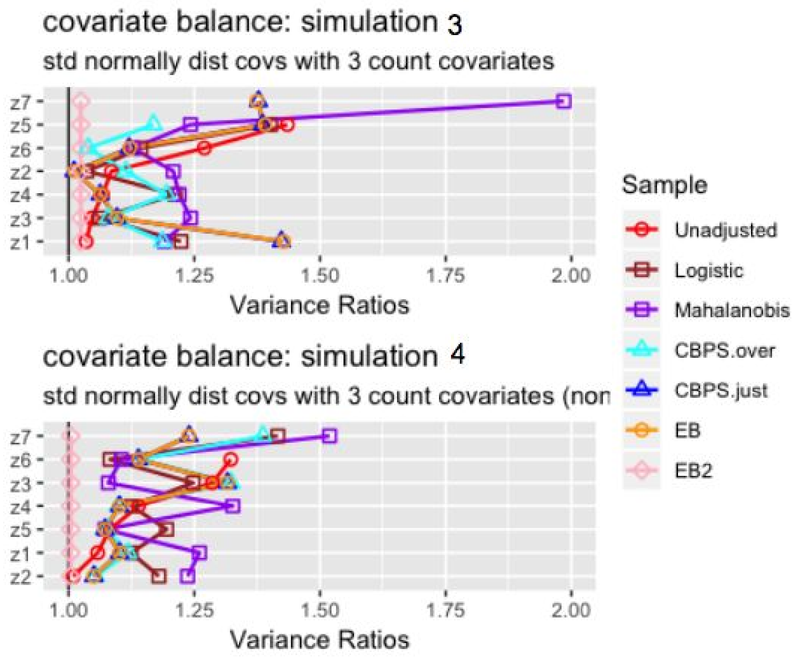


covariate balance: simulation 1

std normally dist covs

covariate balance: simulation 2

std normally dist covs (non-linear ps model)



covariate balance: simulation 3

std normally dist covs with 3 count covariates

covariate balance: simulation 4

std normally dist covs with 3 count covariates (non

Analysis of Variance Ratios:

- The Entropy Balancing model where I have set both first and second moment conditions (EB 2) is the only model that consistently achieves variance ratios = 1. The other models' ability to obtain similar variances of matched samples across treatment groups is rather sporadic.

Results:

Below, the first three tables display results from the 4 simulations for which the *true outcome model is linear*. The latter three tables show the equivalent results, except with a *non-linear true outcome model*. I also plot these tables so that it's easier to visually inspect model results across simulations.

Caution: we should only compare models (columns) given a row (simulation). For example, based on Table 3, it is incorrect to conclude that models do better under conditions for Simulation #2 compared to Simulation #1, because "Simulation #2 (multivariate normal covariates with misspecified propensity score model) has lower RMSE than Simulation #1 (multivariate normal covariates with correctly specified propensity score model)". This result can quickly be reversed by changing the specification of the true propensity score model. Rather, we are interested in how the models' performances given a simulation (e.g. CBPS 1 vs EB-1 when pre-treatment variables come from multivariate normal distribution and propensity score model is incorrectly specified).

Linear Outcome Model:

For this first set of simulations, Mahalanobis has the lowest RMSE for 2 out of 4 simulations. Both of these simulations have a linear true propensity score model. When the true propensity score model is non-linear (#2 and #4), Mahalanobis does better than Logit but worse than CBPS and EB.

All models do better than the baseline logit model (keeping in mind that this baseline model made no attempt to improve the propensity score estimation model).

Comparing CBPS and EB, the over-identified CBPS model (CBPS 1) either ties or out-performs the other specifications of CBPS and EB. CBPS does better when it exploits the dual specification (over-identified) than when it solely balances covariates (just-identified).

As cautioned above, these plots do not suggest that CBPS and EB models underperform for Simulation #3. Simulation #3 is equivalent to #4, except that the true propensity score model is LINEAR for #3 and NON-LINEAR for #4. So the expectation (all else equal) would be that the models would perform better when the true propensity score model is linear. But all else is NOT equal, the DGP (i.e. distributions of the true propensity scores) differ across simulations.

The third graph shows the percentage of iterations in which the true SATT (Sample Average Treatment Estimate on the Treated) falls inside the 95% confidence interval of estimated SATT. These results are much more discouraging for CBPS and EB. They appear to trivially improve upon the baseline logit model (if at all) and for simulations #1 and #3, perform much worse than the Mahalanobis estimate.

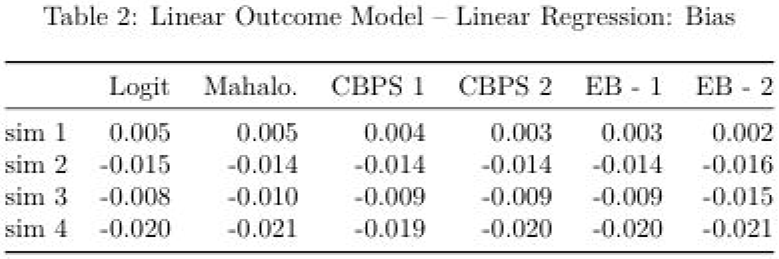


Table 2: Linear Outcome Model – Linear Regression: Bias

	Logit	Mahalo.	CBPS 1	CBPS 2	EB - 1	EB - 2
sim 1	0.005	0.005	0.004	0.003	0.003	0.002
sim 2	-0.015	-0.014	-0.014	-0.014	-0.014	-0.016
sim 3	-0.008	-0.010	-0.009	-0.009	-0.009	-0.015
sim 4	-0.020	-0.021	-0.019	-0.020	-0.020	-0.021

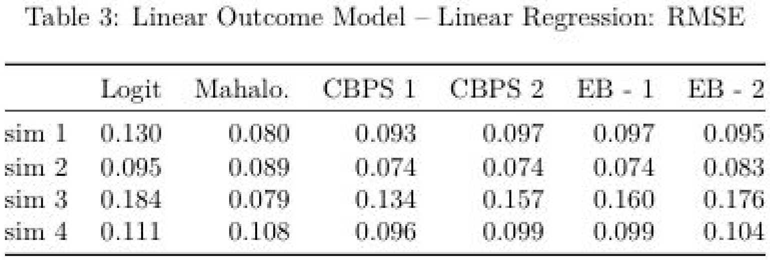


Table 3: Linear Outcome Model – Linear Regression: RMSE

	Logit	Mahalo.	CBPS 1	CBPS 2	EB - 1	EB - 2
sim 1	0.130	0.080	0.093	0.097	0.097	0.095
sim 2	0.095	0.089	0.074	0.074	0.074	0.083
sim 3	0.184	0.079	0.134	0.157	0.160	0.176
sim 4	0.111	0.108	0.096	0.099	0.099	0.104

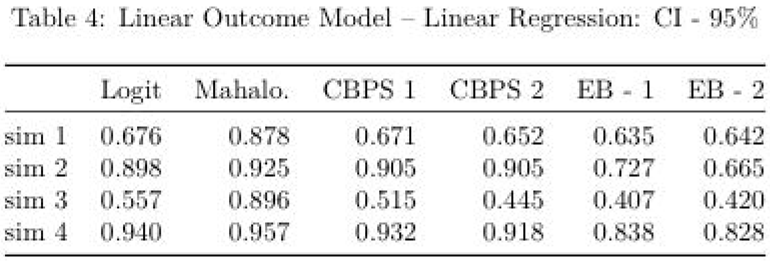
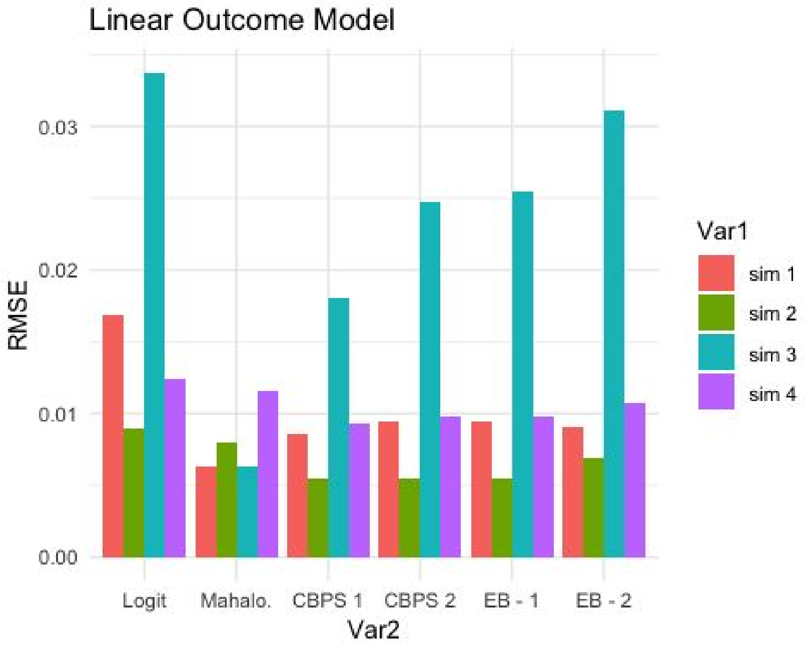
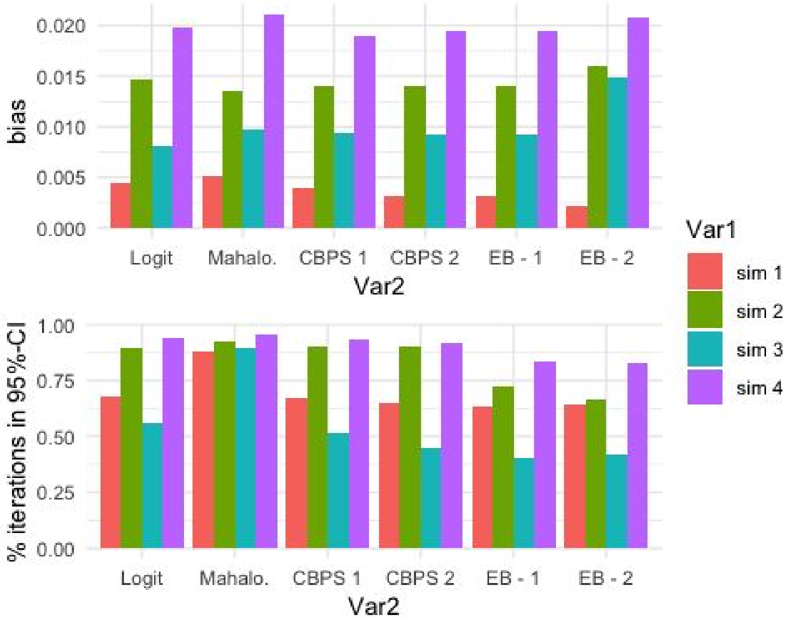


Table 4: Linear Outcome Model – Linear Regression: CI - 95%

	Logit	Mahalo.	CBPS 1	CBPS 2	EB - 1	EB - 2
sim 1	0.676	0.878	0.671	0.652	0.635	0.642
sim 2	0.898	0.925	0.905	0.905	0.727	0.665
sim 3	0.557	0.896	0.515	0.445	0.407	0.420
sim 4	0.940	0.957	0.932	0.918	0.838	0.828





Non-Linear Outcome Model:

Comparison of RMSE results in the same patterns as the Linear-Outcome case above. Mahalanobis does better the other models when the true propensity score model is linear. With a non-linear true propensity score model, CBPS and EB both achiever lower RMSE than Logit and

Mahalanobis. Once again, the CBPS-1 model does at least as well as any of the other CBPS or EB models.

Similar to the previous result, CPBS and EB do not appear to do much better than Logit with respect to % of iterations capturing true SATT within a 95% confidence interval of estimated SATT.

Table 5: Non-Linear Outcome Model – Linear Regression: Bias

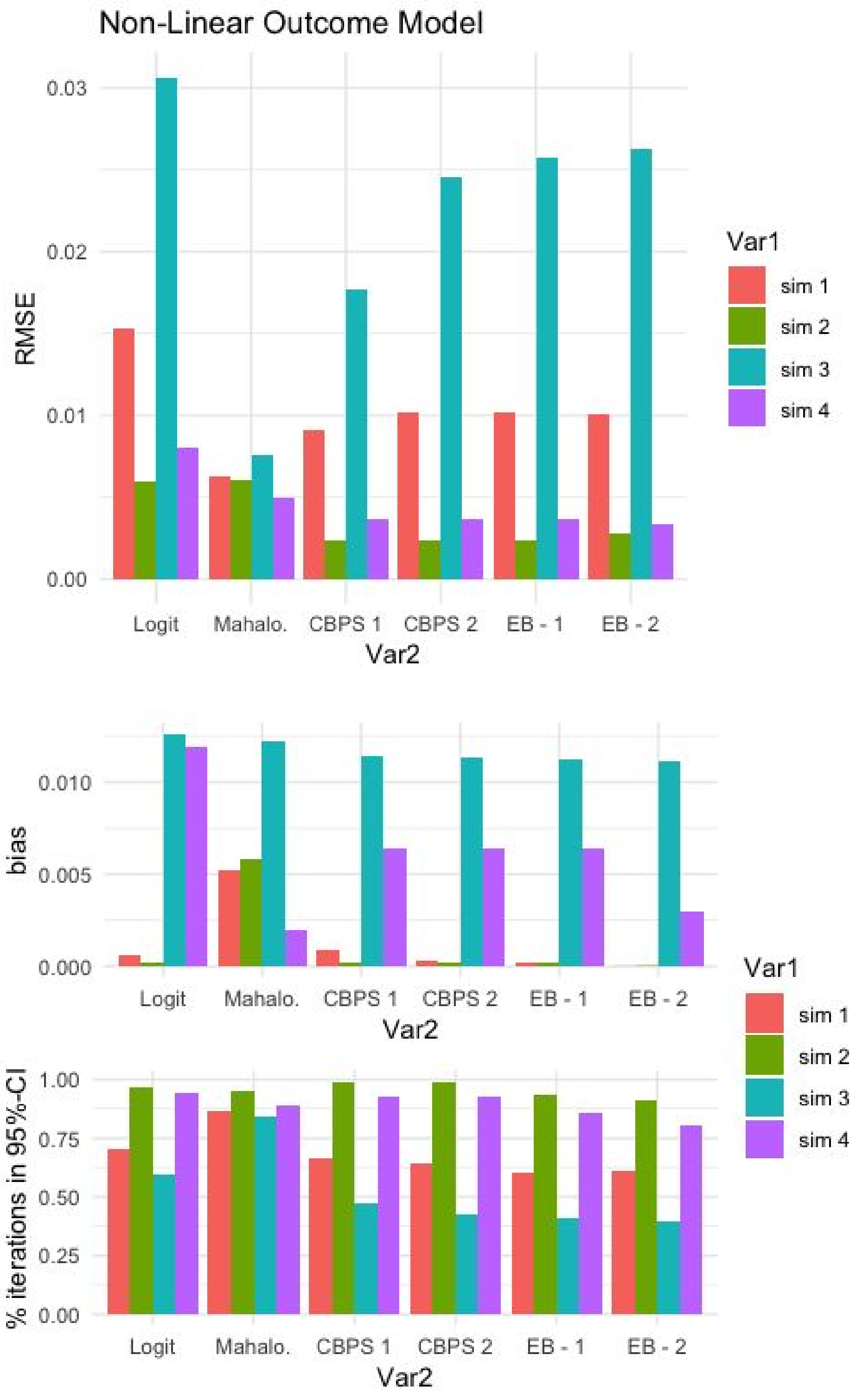
	Logit	Mahalo.	CBPS 1	CBPS 2	EB - 1	EB - 2
sim 1	0.001	0.005	0.001	0.000	0.000	0.000
sim 2	0.000	0.006	0.000	0.000	0.000	0.000
sim 3	-0.013	-0.012	-0.011	-0.011	-0.011	-0.011
sim 4	-0.012	0.002	-0.006	-0.006	-0.006	-0.003

Table 6: Non-Linear Outcome Model – Linear Regression: RMSE

	Logit	Mahalo.	CBPS 1	CBPS 2	EB - 1	EB - 2
sim 1	0.124	0.079	0.096	0.101	0.101	0.100
sim 2	0.077	0.078	0.049	0.049	0.049	0.053
sim 3	0.175	0.087	0.133	0.157	0.160	0.162
sim 4	0.090	0.070	0.061	0.061	0.061	0.058

Table 7: Non-Linear Outcome Model – Linear Regression: CI - 95%

	Logit	Mahalo.	CBPS 1	CBPS 2	EB - 1	EB - 2
sim 1	0.705	0.866	0.668	0.638	0.603	0.612
sim 2	0.965	0.949	0.991	0.991	0.934	0.909
sim 3	0.596	0.839	0.473	0.426	0.409	0.391
sim 4	0.940	0.888	0.927	0.926	0.855	0.804

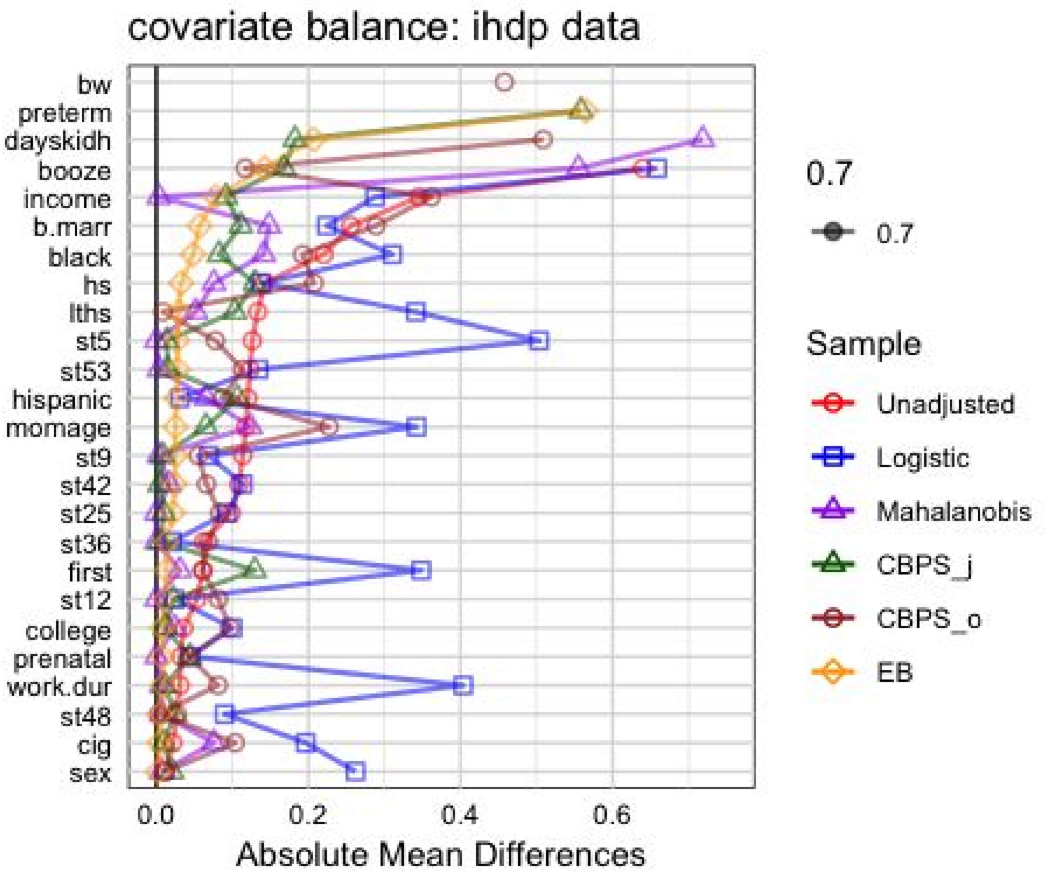

Non-Linear Outcome Model
RMSE
0.03
0.02
0.01
0.00
Logit
Mahalo.
CBPS 1
CBPS 2
EB - 1
EB - 2
Var2
Var1
sim 1
sim 2
sim 3
sim 4
bias
0.010
0.005
0.000
Logit
Mahalo.
CBPS 1
CBPS 2
EB - 1
EB - 2
Var2
Var1
sim 1
sim 2
sim 3
sim 4
% iterations in 95%-CI
1.00
0.75
0.50
0.25
0.00
Logit
Mahalo.
CBPS 1
CBPS 2
EB - 1
EB - 2
Var2

Brief Application to IHDP Data:

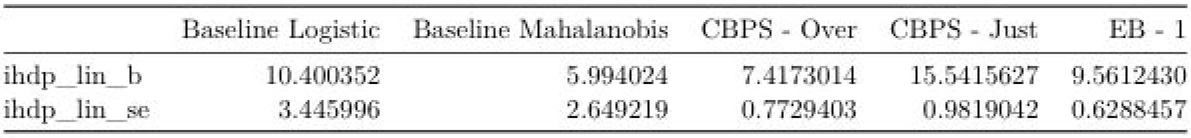
In this section, I apply the matching and re-weighting methods to evaluate the Infant Health and Development Program (IHDP).

The dataset for this section is from homework 4. The dataset contains personal details about approximately 4500 children born in the 1980s and their mothers. Of these 4500, 290 received the treatment: IHDP. IHDP provides special services for the children who receive treatment, such as high-quality child care in the second and third years of life. Treatment was not randomly assigned, but rather, to children who were born (1) prematurely, (2) with low birth weight (1500-2500 grams), and (3) lived in one of the eight cities where the intervention took place. The outcome of interest is a test score conducted at age 3 (similar to an IQ measure).

```
print(ihdp_plot)
```



Similar to the results from the previous simulations, the entropy balancing model achieves the best balance on the covariates.



	Baseline Logistic	Baseline Mahalanobis	CBPS - Over	CBPS - Just	EB - 1
ihdp_lin_b	10.400352	5.994024	7.4173014	15.5415627	9.5612430
ihdp_lin_se	3.445996	2.649219	0.7729403	0.9819042	0.6288457

Conclusion and Limitations:

These results suggest that CBPS and EB may not be a better method for using matching to estimate treatment effects compared to using Mahalanobis distance. It's likely that this result is due to the fact that all four simulations use data generating processes that do not cause major problems for the Mahalanobis Matching method. Even so, the DGP of this study are not radical by any means and these results are significantly different from Hainmueller's results: MSE from entropy balancing that was 3.4 times lower than Mahalanobis matching and 4.6 times lower than matching on the estimated propensity score.

It would be worthwhile to further study the discrepancies between this simulation study and Imai (CBPS) and Hainmueller (EB) results. There is a lot of arbitrariness to setting up the DGP for these simulations. We know that slight misspecifications of the propensity score model can lead to severe biases in treatment effect estimation. Similarly, it seems that slight tweaks in the data generating processes can lead to different outcomes. Perhaps it would be a good idea to design a simulation study that analyzes thousands of iterations of thousands of DGPs. Though I considered this, each simulation already took ~1.5 hours each using NYU's high performance clusters. Repeating each of these simulations ~1000 times was possible but I didn't want to risk waiting to the last minute to write up results.

Next, I would also like to spend some more time gaining a deeper understanding of Equal Percentage Bias Reduction and its implications. Also, for CBPS, I still need to learn how the model actually estimates the weights (generalized method of moments). When does it favor the covariate balancing constraint over the propensity score estimation and vice versa?

References:

- Diamond, A. and Sekhon, J. 2012. Genetic matching for estimating causal effects: a new method of achieving balance in observational studies.
- Hainmueller, Jens. 2012. “Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies.” Political Analysis 20(1): 25–46.
- Imai, K., King, G. and Stuart, E. A. 2008. Misunderstandings between experimentalists and observationalists about causal inference. J. R. Statist. Soc. A, 171, 481–502.
- Imai, Kosuke, and Marc Ratkovic. 2014. “Covariate Balancing Propensity Score.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76(1): 243–63.
- Rosenbaum, Paul R. and Donald B. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” Biometrika 70 (1): 41–55.
- Rubin, Donald B. 1976a. “Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples.” Biometrics 32 (1): 109–120.