LUND UNIVERSITY

School of Economics and Management

Machine Learning as a Data Driven Approach to Automate Multivariate Matching Methods

Author: Sebastian Schmidt

Supervisor: Alessandro Martinello

MSc in Economics

Lund University

School of Economics and Management

Acknowledgement

First and foremost, I would like to thank my supervisor, Alessandro Martinello, for very helpful comments and insightful discussions in the process of conducting this thesis. I thank my family and friends from Lund, especially Team Alfa and the 💯 for highly appreciated support, when it was needed the most. A final and very special thanks goes to my significant other.

Abstract

This paper introduces machine learning to automate the coarsening choices in coarsened exact matching (CEM) as a monotonic imbalance bounding matching class. I suggest to replace the otherwise arbitrary multivariate stratification process with a binary classification tree. This way, I can minimise potential bias caused by subjective preferences. By using the LaLonde (1986) dataset, I systematically compare this novel approach with competing matching specifications, in particular arbitrary CEM and propensity score matching (PSM).

While the automated CEM returns more accurate results than PSM, coarsening arbitrarily performs best in terms of reducing imbalance as well as in the post-matching causal estimation.

Keywords: Matching, Machine learning, Coarsened exact matching, Causal inference

Contents

List of Figures

List of Tables

1 Introduction

Matching is a common method to preprocess observational data (Ho et al., 2007; Morgan and Winship, 2014; King and Nielsen, 2016). The aim of matching is to reduce model dependence by creating covariate balance between the treated and the control group (Diamond and Sekhon, 2013; King and Nielsen, 2016). Covariate balance implies to equalise the empirical distributions between both of these groups, which is accomplished by discarding ill-suited observations from the data (Diamond and Sekhon, 2013). If matching is successful, it can reduce inefficiencies and bias (King and Nielsen, 2016). Thus, matching helps to make the assumptions needed to make causal inference more credible.

In its basic form, the matching algorithm assigns each treated unit one observation from the control group with identical characteristics. On the one hand, using more covariates increases the accuracy of a match. On the other hand, it also becomes increasingly difficult to find identical – or exact – matches. This trade-off is called the "curse of dimensionality" (Bellman, 1961) and represents the fundamental problem of matching. A number of different matching methods have been proposed that aim to reduce dimensionality while trying to keep the variance increase under control (Rosenbaum and Rubin, 1983; Iacus et al., 2011; Diamond and Sekhon, 2013). Among the different matching methods, propensity score matching (PSM) is by far the most commonly applied (Pearl, 2010). PSM effectively reduces dimensionality by collapsing a hyper-dimensional space of conditioning variables and projecting it on a one-dimensional propensity score. An alternative matching method is coarsened exact matching (CEM), which first stratifies the dataset into (more) balanced subsets before finding matches (Iacus et al., 2011, 2012). Despite their popularity in applied research, matching methods are frequently criticised as not being suitable to effectively reduce the imbalance in observational data. King and Nielsen (2016) argue that the propensity score methodology is unsuitable for matching as it fails to accurately measure the multi-dimensional distance between observations from the treatment and the control group. Yet, CEM relies on manually or arbitrarily stratifying each covariate. Therefore the results may suffer from bias caused by the personal preferences of the researcher.

This paper proposes to integrate machine learning with non-parametric monotonic imbalance bounding matching methods such as CEM. I develop a data-driven method that automatically stratifies the data before applying CEM. This way, I offer a numeric solution to the problem of arbitrary coarsening, as the machine learning algorithm makes the stratification choice that has previously been made by the researcher. I systematically test and

compare this novel approach with the standard applications of PSM as well as CEM by using the seminal dataset of LaLonde (1986). In addition, I also test whether estimating the propensity scores non-parametrically, can improve the performance in PSM.

Machine learning are non-parametric methods for model prediction that have become increasingly popular to research in econometrics and applied economics (Varian, 2014; Athey and Imbens, 2017; Mullainathan and Spiess, 2017). In PSM, machine learning algorithms represent a non-parametric alternative to estimate the propensity score (McCaffrey et al., 2004; Setoguchi et al., 2008; Lee et al., 2010; Wyss et al., 2014). Indeed, estimating the propensity score parametrically – usually with a logistic regression – is a major point of critique to the PSM methodology. Since the true propensity score usually remains unknown, the results of a parametric approach may be subject to misspecification (Diamond and Sekhon, 2013; Imai and Ratkovic, 2014). Diamond and Sekhon (2013) use machine learning as an automation process for finding the optimal covariate balance before estimating the propensity score. This way, the researcher can circumvent the problem of having to know the actual propensity score. King and Nielsen (2016) criticise PSM to perform badly in particular once the data is balanced, because PSM can no longer differentiate between good and bad matches. Instead, the authors propose CEM as an alternative and more accurate methodology. CEM first stratifies the dataset into (more) balanced subsets. Next, the method matches the data within these subsets with exact matching or a statistical model. The latter now has to estimate much shorter distances than in the unstratified data (Iacus et al., 2011, 2012). I build on CEM and implement recursive binary classification trees as a data-driven method to automate the selection of the hyper-dimensional histograms on which the stratification is grounded.

Overall, I find that both CEM specifications outperform PSM. However, the automated CEM cannot reduce imbalance as effectively as matching arbitrarily with CEM. Moreover, the arbitrarily matched data returns the most accurate results in the post-matching causal estimations. I also find, that using classification trees to estimate the propensity score does not improve the performance in PSM.

The remainder of this thesis is organised as follows. Section 2 provides a brief overview on matching and introduces the two matching methods of PSM as well as CEM. Furthermore, I explain why propensity scores are unsuitable for matching and highlight potential shortcomings of the current coarsened exact matching methodology. In section 3, I develop a strategy how classification trees might be able to overcome the problem of coarsening ar-

bitrarily. Following this, section 4 introduces a multivariate imbalance measure that enables a comparison of the different methods. In section 5, I present the results and measure the performance of the methods by applying them to a real dataset. Section 6 summarises the findings, outlines the limitations and gives some ideas for future research.

2 Matching Methods

Matching may be defined as "simultaneously seeking to maximize covariate balance between the treated and control groups and the matched sample size" (King et al., 2011, p. 1). Matching is a popular method in many quantitative fields, including statistics (Rosenbaum, 2002; Rubin and Stuart, 2006), economics (Dehejia and Wahba, 1999; Imbens, 2004; Smith and Todd, 2005; Abadie and Imbens, 2006; VanderWeele and Hernan, 2013), sociology (Morgan and Harding, 2006) and political science (Ho et al., 2007).

In this section, I give a brief introduction to the notations and properties of matching. I introduce the two main classes of matching methods: Equal Percent Bias Reducing (EPBR) and Monotonic Imbalance Bounding (MIB) matching. Thereafter, I explain propensity score matching as an EPBR class method and why propensity scores are ineffective for matching. Next, I introduce coarsened exact matching as a MIB class method followed by possible shortcomings in the current CEM methodology.

2.1 The Properties of Matching

The goal of matching is to establish the same covariate balance that would otherwise be created by random assignment. This way, model dependence can be reduced by matching. Model dependence implies that the difference between the treatment and the control group cannot be isolated to the treatment effect alone (King and Nielsen, 2016). Instead, the control variables, which are included in the model, also have some effect on the outcome.

Assume a population N in which each unit $i = 1, ..., N$ is assigned into two groups by a treatment variable $T_i \in 0, 1$, where 1 denotes treatment and 0 control (VanderWeele and Hernan, 2013). Further assume a k-dimensional set of covariates X_i as well as some outcome Y_i. For each unit there are two potential outcomes, $Y_{1,i}$ for being treated and $Y_{0,i}$ for being assigned into the control group.

The resulting true causal effect is denoted as the difference $\tau = Y_{1,i} - Y_{0,i}$. Yet, the effect remains unobservable at the unit level as only one of the two outcomes is observed, such that $Y_i = T_i Y_{1,i} + (1 - T_i) Y_{0,i}$ (Holland, 1986; Iacus et al., 2011). Estimating the causal effects essentially implies to predict or approximate the unobserved potential outcomes (Rubin, 1976a; Stuart, 2010). Typically, we approximate the unobserved potential outcome Y_0 of a population with that of the untreated (control) population. If treatment is randomly assigned, for example in a natural experiment, the treatment assignment is known by the researcher. Then, the covariance between treatment and control group is balanced and any

difference between both groups is random (Stuart, 2010). In observational data, however, treatment assignment is non-random and not controlled or known by the researcher (King and Nielsen, 2016). Here, matching can help to balance the covariates between the treatment and control group before estimating the causal effect.

Formally, matching algorithms try to find for every treated unit with covariates X_i some control unit with covariates X_j such that X_i and X_j are as similar as possible, $d(X_i, X_j) \simeq 0$ (Iacus et al., 2011). Three assumptions must hold in order to estimate an unbiased causal effect with matched data: conditional independence, the overlap assumption and the stable unit treatment value assumption.

Conditional independence or unconfoundedness is defined as some outcome Y_i to be independent of being assigned treatment $Y_{1,i}$ (or control $Y_{0,i}$) in T_i conditional on the covariates included in X, such that $(Y_{1,i}, Y_{0,i}) \perp T_i|X$ (Rosenbaum and Rubin, 1983; Stuart, 2010). The conditional independence assumption is an identifying assumption and thus, cannot be tested. It is satisfied when X_i includes all variables that affect – or in practice are known to affect – the outcome Y_i (VanderWeele and Hernan, 2013; King and Nielsen, 2016). If the conditional independence assumption is violated, the result suffers from the omitted variable bias and it becomes infeasible to make causal inference with the matched data (Ho et al., 2007). Yet, it is impossible to be ultimately certain that all relevant covariates are included for sure (Rubin, 2008). Therefore, the researcher has to rely on economic reasoning as well as previous research. The overlap assumption, $0 < Pr(T_i = 1|X) < 1$, ensures that all observations i have a positive probability of being assigned a treatment (Smith and Todd, 2005). Finally, the stable unit treatment value assumption implies that the values of potential outcomes are fixed and do not change if treatment T_i changes from one to zero (VanderWeele and Hernan, 2013).

Assuming that these three assumptions hold, it is possible to overcome the counterfactual problem and to estimate the causal effect with the matched data using the average treatment effect on the treated (ATET) (Imbens, 2004; Angrist and Pischke, 2009). ATET is defined as

$$\hat{\tau}^{ATET} = E[Y_{1,i}|T_i = 1] - E[Y_{0,i}|T_i = 1] \tag{1}$$

and if the data is successfully balanced, it is an unbiased estimator for the true causal effect such that $E[\hat{\tau}^{ATET}] = \tau$ (Imbens, 2004; Angrist and Pischke, 2009; King and Nielsen, 2016).

In exact matching the potential outcomes of the control group are estimated at the unit

level, such that both, treatment and control group, have an identical set of covariates and $d(X_i, X_j) = 0$ (Iacus et al., 2011). Then $\hat{\tau}$ reduces to a simple difference in mean

$$\hat{\tau} = \frac{1}{N} \sum_{i \in T} Y_i^T - \frac{1}{N} \sum_{j \in C} \omega_j Y_j^C, \tag{2}$$

where ω is a weight that compares a treated individual with one from the control group (Smith and Todd, 2005; Iacus et al., 2012).

The accuracy of finding good matches increases with the number of covariates. Yet, the number of dimensions increases as well and local matching becomes infeasible. When exact matching is no longer possible, some form of model m_ℓ has to be applied in order to estimate the distance between X_i and X_j such that $\hat{Y}_0 = m_\ell(\tilde{X}_j)$, where $\tilde{X}_j$ denotes those values in the control group that are approximately similar to the treated units (Iacus et al., 2011). This estimation process increases imbalance and causes some model dependence. Therefore, the researcher faces a trade-off between reducing dimensionality in order to maintain the predictive power of the matched data and an increasing imbalance, since no existing matching method is able to optimally reduce both (King et al., 2011).

All matching methods address this trade-off by trying to reduce dimensionality without increasing variance too much. The literature defines two groups of matching methods: Equal Percent Bias Reducing (EPBR) and Monotonic Imbalance Bounding (MIB) (Iacus et al., 2011, 2012). PSM belongs the the first class and CEM to the latter. The fundamental difference between both is that EPBR fixes the number of matched control units before matching, while MIB fixes the maximum level of imbalance (Iacus et al., 2011). In EPBR all treated observations have to be matched, which implies that the number of treated observations that are matched is equal to the total number of treated units in the dataset, i.e. $m_T = n_T$. However, the number of controls can be $m_C \leq n_C$. Furthermore, denote $\bar{X}$ the sample mean of the relevant subgroup of the dataset. EPBR then solves the following equation

$$E[\bar{X}_{m_T} - \bar{X}_{m_C}] = \gamma E[\bar{X}_{n_T} - \bar{X}_{n_C}], \tag{3}$$

where $0 < \gamma < 1$ is a scalar that indicates the level of imbalance between the total dataset and the feasible matches (Rubin, 1976b; Iacus et al., 2011).

MIB on the contrary solves the inequality

$$|\bar{X}_{m_T j} - \bar{X}_{m_C j}| \leq \gamma |\bar{X}_{n_T j} - \bar{X}_{n_C j}|, \tag{4}$$

where $\bar{X}_{n_T j}$, $\bar{X}_{n_C j}$, $\bar{X}_{m_T j}$ and $\bar{X}_{m_C j}$ are the sample means before and after matching for treated and control units respectively (Iacus et al., 2011). Rewrite the right hand side of equation 4 as $\delta_j = \gamma |\bar{X}_{n_T j} - \bar{X}_{n_C j}|$ to obtain

$$|\bar{X}_{m_T j} - \bar{X}_{m_C j}| \leq \delta_j, \quad for \quad j = 1, ..., k. \tag{5}$$

Since δ_j is chosen before matching, m_T and m_c are defined through the matching process Iacus et al. (2011).

Recall, that the goal of matching was to minimise the distance between both sets of covariates X_i and X_j. Equation 5 can then be rewritten such that a matching method can be called monotonic imbalance bounding, MIB(f,d), if for some monotonically increasing function $\gamma_{f,d}(\pi)$ it holds that

$$d(f(\chi_{m_T(\pi)}), f(\chi_{m_C(\pi)})) \leq \gamma_{f,d}(\pi). \tag{6}$$

Here, π denotes a predefined vector of tuning parameters, that has the same number of rows as there are covariates in X (i.e. k) while χ_{m_T} and χ_{m_c} are subsets created by the matching process (Iacus et al., 2011). From equation 6 two essential properties follow for MIB class matching methods: firstly the function $\gamma_{f,d}(\pi)$ only depends on π, as d and f are predefined by design; and secondly changing one tuning parameter in π for its assigned covariate in X does not alter the defined maximum imbalance for the other covariates (Iacus et al., 2011).

2.2 Propensity Score Matching

Propensity score matching (PSM) seeks to achieve covariance balance such that $(Y_{1,i}, Y_{0,i}) \perp p(T_i|X)$ (Rosenbaum and Rubin, 1983; Dehejia and Wahba, 1999). Hereby, PSM addresses the problem of dimensionality by collapsing the k-dimensional matching process into a one-dimensional propensity score (King and Nielsen, 2016). Given the conditional independence assumption (potential outcomes are independent on treatment status conditional on X), potential outcomes are also independent on a scalar function of covariates defined as

$$p(x) = E[T_i|X] = Pr[T_i = 1|X] \tag{7}$$

which is the propensity score (Dehejia and Wahba, 2002; Angrist and Pischke, 2009). In other words, propensity score matching implies that it is enough to estimate the probability of being assigned to a treatment, given the characteristics in X for $\hat{\tau}$ to be an unbiased estimator for the true causal effect.

2.3 The Inefficiency of PSM and the PSM Paradox

In its attempt to reduce dimensionality, PSM assumes a fully randomised experiment within the observational dataset. Fully randomised means that PSM – as an EPBR matching class method – finds matches for all observations from the treatment group such that $m_T = n_T$. The propensity score methodology is therefore unsuitable in three ways for matching: first, PSM is inefficient. Second, PSM estimates the propensity score with a logistic regression model in its standard application. The logit specification, however, makes PSM sensitive to statistical bias as estimating the propensity score now relies on parametric assumptions. Third, collapsing a multi-dimensional problem into a one-dimensional propensity score may cause random pruning if the covariates are already balanced in the dataset. In fact, PSM begins earlier to prune randomly if the unmatched data is more balanced from the beginning. Since this phenomenon fundamentally contradicts with the idea of matching, it is called the "Propensity Score Paradox" in the literature (King et al., 2011; King and Nielsen, 2016).

To understand the inefficiency argument recall that under exact matching the matched subset is defined such that $d(X_i - X_j) = 0$. Therefore, imbalance is equal to zero by design. However, as the number of covariates in X grows in k, finding exact matches becomes an increasingly difficult task. PSM tries to circumvent this problem by reducing the dimensionality from a $k \times n$ matrix to an one-dimensional propensity score. Now, only the condition that $d(\hat{p}(x_i) - \hat{p}(x_j)) = 0$ must hold. This solution seems sensible, since it is computationally much more efficient to match two one-dimensional propensity scores, than two $k \times n$ matrices. However, this simplification is misleading and $d(X_i - X_j) \neq d(\hat{p}(x_i) - \hat{p}(x_j))$. In PSM, the treatment assignment only depends on the probability of actually being treated. Recall equation 7, where equality is achieved by conditioning on the propensity score such that $p(x) = Pr(T_i|X)$. This equalisation implies that $p(x)$ is in turn conditional on the covariates such that $p(x)|X$. While PSM may indeed estimate two balanced one-dimensional propensity scores, $\hat{p}(x_i) = \hat{p}(x_j)$, it does not necessarily imply that the underlying $k \times n$-dimensional covariate matrix is matched exactly as well and $X_i \neq X_j$. Consequently, the imbalance is not equal to zero and the assumption that $d(\hat{p}(x_i) - \hat{p}(x_j)) = 0$ does no longer hold true (King

and Nielsen, 2016).

Full randomisation as in the case for PSM requires some statistical model to estimate the propensity score. Usually $p(x)$ is estimated by a logit specification of the form $p(x)_i = (1 + e^{X_i\hat{\beta}})^{-1}$ before computing the distance (King and Nielsen, 2016). In practice, however, the true propensity score remains unknown to the researcher. Therefore, using a logistic regression makes PSM sensitive to misspecification and thus, a possible source of bias (Diamond and Sekhon, 2013). Furthermore, making parametric assumptions contradicts with the fundamental goal of matching to actually find the true data generating process (King and Nielsen, 2016).

The propensity score paradox implies that if PSM is applied to a randomised and balanced dataset, PSM leads to random pruning. Recall, that the aim of matching is to find the experimental data within an observational dataset. Therefore, unnecessary observations are pruned away in order to isolate the "experiment" and to achieve covariate balance. However, once PSM has established covariate balance, the one-dimensional propensity score can no longer differentiate between good and bad matches. This is because a one-dimensional vector does not depict the true multi-dimensional distance of two points. Consequently, PSM starts to match randomly which leads to random pruning. The imbalance increases again and eventually trends towards maximal imbalance. In real datasets the paradox causes random pruning as soon as PSM has established covariate balance. In an extreme case – i.e. when using data from a natural experiment – the propensity score paradox sets in immediately (King and Nielsen, 2016).

2.4 Coarsened Exact Matching

Coarsened exact matching (CEM) temporarily creates $s \in S$ strata or bins to place similar values in a single stratum. T^s denotes the treated units in stratum s and $m_T^s = \#T^s$ the number of treated units in s as well as C^s and $m_C^s = \#C^s$ for the control units respectively. Let $n_T = \cup_{s\in S} m_T^s$ denote the total number of all treated units in the dataset and the total number controls as $n_C = \cup_{s\in S} m_C^s$ respectively (Iacus et al., 2012). Next, CEM computes weights as

$$w_i = \begin{cases} 1, & i \in T^s \\ \frac{m_C}{m_T}\frac{m_T^s}{m_C^s}, & i \in C^s \end{cases}$$

which implies that the number of control units equals the number of treated units (which are assigned a weight of $w_i = 1$). Bins that do not contain at least one treated and control unit

are discarded ($w_i = 0$) (King et al., 2011; Iacus et al., 2011, 2012).

Recall equations 5 and 6 from the previous section. CEM as an MIB method fixes the level maximum of imbalance before the matching process. All bad matches that would cause a higher imbalance than the predefined threshold are automatically excluded from the matching process. Now, the remaining imbalance should be small enough to match at unit level or to apply some statistical model within each stratum in order to efficiently estimate the distance without causing too much model dependence (Iacus et al., 2012).

2.5 The Choice of Coarsening

CEM, as a MIB class matching method, chooses the number of included treated units according to the ex ante defined maximum level of imbalance such that $m_T \subseteq n_T$. Moreover, CEM only has to find matches within the predefined number of s subsets or strata. This way, the distances are short enough that CEM can apply exact matching and therefore it is a non-parametric matching method. Finally, the researcher has full control over the maximum imbalance of the matching process, as he or she chooses the coarsening level for every variable independently. Yet, these convenient properties of CEM come at a cost.

First, by fixing the maximum level of imbalance, CEM may prune away too many observations such that the matching process threatens the explanatory power of the post-matching causal estimations (Iacus et al., 2012). This threat is particularly present in large and high-dimensional datasets with many covariates. Already very few cut-points may result in large numbers of bins. Note that the number of strata increases in k with s^k. Yet, CEM only allows at most n strata, that is when each observation is placed in its own bin (Iacus et al., 2011; Abadie and Imbens, 2012).

Secondly, CEM relies in its standard application on economic reasoning or arbitrarily choosing the coarsening level for every variable.[1] However, human interference in any kind of estimation should be avoided as it is one of the prime sources of bias (Banaji and Greenwald, 2016). Indeed, replacing as much as possible of the human decision making process by non-parametric numerical approaches is the ultimate goal of statistical science (King and Nielsen, 2016).

[1] Although Iacus et al. (2009) and Iacus et al. (2012) offer some automating mechanisms. Ultimately, the final choice on coarsening (and the recommended best practice) requires the researcher to take an active role in the coarsening decision.

3 Proposing Classification Trees to Automate the Coarsening Choices in CEM

There are different types of tree based methods that are applicable to regression as well as classification problems (see Hastie et al., 2009, for details). Since treatment assignment is binary, I choose to use a recursive binary tree from the CART (Classification and Regression Tree) family. First, I introduce the method of classification trees. Second, I develop an approach how trees can be applied to automate the coarsening choices in CEM.

3.1 Classification Trees

Classification trees predict the likelihood that some observation belongs to a certain class of an event by stratifying the feature space, X_i (i.e. the $n \times k$ covariate matrix), into hyper-dimensional rectangles. In region $R_1, ..., R_m$, the tree fits a simple model or a constant denoted as c_m (Hastie et al., 2009; James et al., 2013). The tree predicts the outcome – here treatment – with the model

$$\hat{f}(X) = \sum_{m=1}^{M} c_m \mathbb{I}_{Rm}(X), \tag{8}$$

where $m = 1, ..., M$ denotes the number of nodes for regions $R_1, ..., R_m$ (Hastie et al., 2009). In order to optimally split the nodes, it is desired to place each unit in a region according to the most likely class (James et al., 2013). Let $\hat{p}_{mk}$ denote the proportion of observations that belong to class k in node m such that

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k). \tag{9}$$

The researcher can choose from several criteria that are used to define the splits in recursive binary classification trees. Here, I define splits according to the standard practice, by maximising the Gini index computed on the observed outcomes across m nodes (Hastie et al., 2009). The Gini index is defined as

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}). \tag{10}$$

The algorithm thus minimises the variance of observed outcomes within nodes while maximising the variance across K classes. Clearly, if many values of $\hat{p}_{mk}$ are either close to one or zero, the Gini index is low, indicating that a node m contains many values from one class

only (James et al., 2013).

Tree based methods tend to overfit the data. Overfitting, implies that the model loses prediction power by extracting information from the error or the noise and falsely follows that information too closely (James et al., 2013). Therefore, after having grown an extensive or "greedy" tree $\mathcal{T}_0$, some complexity criterion $C_\alpha(\mathcal{T})$ is applied to prune it in order to improve its predictive power (Breiman et al., 1984). In classification problems, trees are pruned according to minimising the misclassification rate over all splits (James et al., 2013). The optimal misclassification rate that minimises α is given as

$$C_\alpha(\mathcal{T}) = 1 - \max_k(\hat{p}_{mk}). \tag{11}$$

3.2 Using Classification Trees in Matching

The original purpose of decision trees as a form of supervised machine learning is to train the algorithm with the help of a teacher on a training dataset, in order to predict some outcome out of sample (Varian, 2014; Athey and Imbens, 2017). Here, I propose a binary tree because of its binning properties and not primarily for its predictive power. In fact, tree based methods work similarly to CEM as they also stratify the data into hyper-dimensional regions (Hastie et al., 2009). Therefore, I implement recursive binary trees as a data-driven method to automatise the stratification process in CEM. This way, human interference in the decision making process is reduced to a minimum.

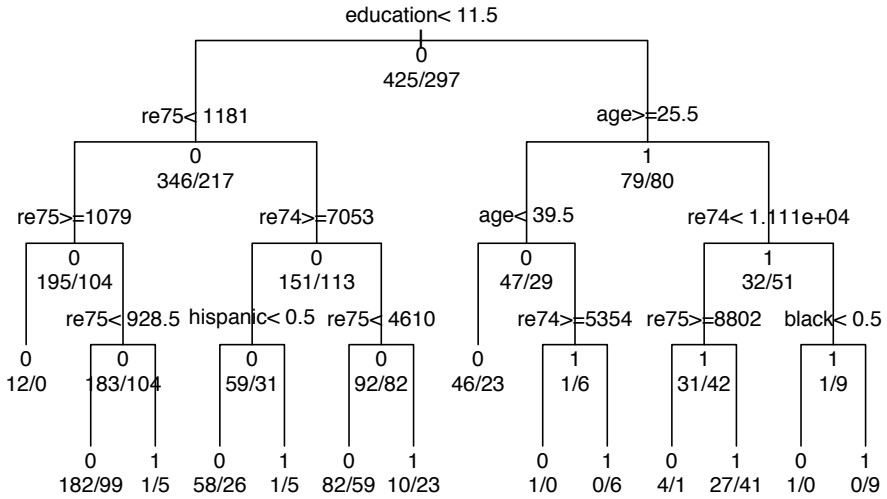

Figure 1: Visualisation of a recursive binary classification tree with four levels and 14 terminal nodes

I grow an extensive tree $\mathcal{T}_0$, where the number of terminal nodes is equal to the number of strata $|\mathcal{T}_0| = s$. Then, the usual CEM algorithm is applied in these $|\mathcal{T}|$ strata. Figure 1 presents an example of a tree with four levels and 14 terminal nodes estimated from the LaLonde (1986) data, which I present in section 5.

Automatising the multivariate stratification process with a classification tree is efficient but does not violate the properties of MIB class matching methods. The binary tree optimises the coarsening of each covariate in X according to equations 9 and 10. Thus, machine learning algorithms replace the manual and cumbersome process of coarsening every variable independently. Yet, the researcher remains in full control of the absolute imbalance level by defining the total depth of the tree or the maximum number of nodes since $|\mathcal{T}_0| = s$. In essence, this approach implies that one only has to control a single variable: the number of terminal nodes.

However, the matching process remains monotonic imbalance bounding. The decision tree stratifies the $k \times x$ covariate matrix – or feature space in statistical terms – into hyper-dimensional subregions. Therefore, applying machine learning algorithms in CEM does not violated the properties of equation 6.

As a result, binary trees represent a data-driven solution to automate the process of creating covariate balance and reducing model dependence.

4 Measuring Imbalance

To measure the performance of a particular matching technique in the context of reducing model dependence, a multivariate imbalance measure is needed, that incorporates as much of the distribution as possible and is able to capture discrete splitting (Iacus et al., 2012). Iacus et al. (2011) introduce a multivariate imbalance measure based on the L_1 norm regularisation. I follow the authors' advise and introduce the L_1 imbalance measure because it normalises the measurement of imbalance and thus allows to compare the different methods on a single scale.

$H(x_1)$ denotes the values resulting from binning the covariate x_1. The multidimensional histogram is the Cartesian product of all k covariates $H(X) = H(x_1) \times ... \times H(x_k)$. The relative frequency distributions for the units that are placed in the bin with coordinates $\ell_1, ..., \ell_k$ of the multivariate cross-tabulation for the treated and control group are denoted as $f_{\ell_1,...,\ell_k}$ and $g_{\ell_1,...,\ell_k}$ respectively (Iacus et al., 2011, 2012). The L_1 distance is then defined as

$$L_1(f, g, H) = \frac{1}{2} \sum_{\ell_1 ... \ell_k \in H(\mathbf{X})} |f_{\ell_1 ... \ell_k} - g_{\ell_1 ... \ell_k}|, \tag{12}$$

where $L_1 \in [0, 1]$. In principle, the L_1 norm compares two multi-dimensional histograms, one with the distribution of the treatment group, the other one from the control group. If both groups are perfectly balanced, the two histograms are identical and the imbalance is equal to zero. If both histograms are perfectly different L_1 is equal to one. For all other cases the L_1 imbalance returns the relative imbalance. If for example, $L_1 = 0.7$, 30 percent of the distributions overlap (Iacus et al., 2012). The L_1 distance has a number of useful properties: first, it can measure discrete variables. Second, the measurement is not affected by large numbers of strata that are weighted down to zero in case they only contain units from the control group. Third, the sum in equation 12 has at most as many non-zero terms as there are units from the treated group. That is when only one observation from the treated group is placed in each stratum (Iacus et al., 2012).

5 Evaluating the Matching Methods

In this section, I test the performance of the method that I have developed in section 3 by applying it to a real dataset. First, I evaluate the ability to reduce imbalance for a given level of pruning. Second, I estimate the causal effect using the matched data. According to the matching methodology, estimating the ATET with the difference-in-mean estimator from equation 2 returns an unbiased estimator for the true causal effect in case the data is successfully balanced through the matching process. This way, I can evaluate if a matching method is suitable for making causal inference.

Table 1: Summary statistics: NSW male subsample

	N	Mean	SD	Min	Max
1978 real Earnings	722.00	5454.64	6252.94	0.00	60307.93
Treated	722.00	0.41	0.49	0.00	1.00
Age	722.00	24.52	6.63	17.00	55.00
Education	722.00	10.27	1.70	3.00	16.00
1974 real earnings	722.00	3630.74	6220.64	0.00	39570.68
1975 real earnings	722.00	3042.90	5066.14	0.00	37431.66
Married	722.00	0.16	0.37	0.00	1.00
High school dropout	722.00	0.78	0.41	0.00	1.00
Black	722.00	0.80	0.40	0.00	1.00
Hispanic	722.00	0.11	0.31	0.00	1.00
Unemployed in 1974	722.00	0.45	0.50	0.00	1.00
Unemployed in 1975	722.00	0.40	0.49	0.00	1.00

I use the seminal dataset from LaLonde (1986). The dataset is used regularly in the literature to evaluate matching methods (Dehejia and Wahba, 1999, 2002; Smith and Todd, 2005; Iacus et al., 2012). The data includes in total 722 observations of which 297 were assigned a treatment and the rest belongs to the control group. Originally, the data represents the male subgroup from the National Supported Work Demonstration (NSW) and evaluated the success of a randomly assigned job training programme for low qualified and disadvantaged workers based on the salary after the programme had ended (re78) (LaLonde, 1986). The unit is 1982 USD. The variable "treated" is equal to one if the person participated in the job training programme and zero otherwise. The dataset includes the following control variables: age, years of education (education), 1974 real earnings (re74) and 1975 real earnings (re75). Moreover, a set of dummy variables assigns whether the participants were unemployed in 1974 (u74) or in 1975 (u75), are married (married), dropped out of high school (nodegree) and are of Afro-American or Hispanic descent. Table 1 presents the summary statistics.

5.1 Evaluating the Matching Performance

I match the LaLonde dataset four times. Firstly, I apply PSM with nearest neighbour matching, where the propensity score is estimated using a logistic regression. Secondly, I run PSM again, but estimate the propensity score with the pruned classification tree $\mathcal{T}_\alpha$. Thirdly, I apply CEM, where the coarsening is chosen arbitrarily. Finally, I match the data a fourth time with CEM. This time, the leafs of the classification tree decide the coarsening of each variable such that $s = |\mathcal{T}|$.

For both PSM specifications, I match the data and then exclude the two observations that were the worst match, i.e. the matched data pair, which lay the furthest apart on the propensity score, $\max_d |\hat{p}(x_i) - \hat{p}(x_j)|$. I measure the L_1 imbalance and repeat until all data is pruned away.

In the second PSM specification I replace the logistic regression with the same classification tree that I propose in CEM. Since trees are non-parametric it is no longer necessary to know the true propensity score and thus misspecification represents no threat to the validity of the PSM result. Estimating the propensity score is a prediction problem, therefore I use a subtree of the original tree, $\mathcal{T}_\alpha \subset \mathcal{T}_0$, that was pruned according to equation 11. This way, the binary tree minimises the misclassification rate $C_\alpha(\mathcal{T})$ and optimises the prediction accuracy.

When matching with CEM arbitrarily, I begin by the highest coarsening possible, which means that all data is placed in a single stratum. This initial point shows the same imbalance as if the data was not matched at all. Then, I randomly choose one covariate and add one cut-point such that the distance is split in equal shares (i.e. for the first cut-point in the ratio $1/2, 1/2$ for the second cut-point of that variable $1/3, 1/3, 1/3$ and so on). I measure the imbalance and repeat. As the number of cut-points rises, the data is divided into more and smaller bins. Thus, the likelihood increases that a certain bin only contains observations from the control group which is weighted to zero and discarded.

Finally, I automate the coarsening process by binning the data with the help of the classification tree. I grow an excessive tree $\mathcal{T}_0$ and leave it unpruned. As for the arbitrary CEM, I begin by sorting all data into a single stratum, such that both methods start at the same initial point. For the tree, this means that it collapses to a single root. Now instead of adding cut-points arbitrarily, I increase the depth of the binary tree by one level such that the tree has two terminal nodes which divide the data into two strata. I measure the imbalance and increase the depth of the tree to two levels. Now, the data is stratified into at least three but

at most four terminal nodes (for the third level, the number of strata lies between five and eight). Again, I measure the imbalance and continue adding levels until I reach the maximum depth of the estimated tree. This way, the number of terminal nodes (or strata) increases non-linearly. With an increasing number of leafs the likelihood that a leaf only contains observations from the control group rises as well. As for the arbitrary case, these strata are weighted to zero and pruned by the CEM algorithm.

Figure 2 plots the performance of all four methods tested. The vertical axis denotes the L_1 imbalance, while the horizontal axis shows the number of pruned observations. The green line represents the result for PSM with nearest neighbour matching using a logistic regression, while the red line depicts the PSM outcome when using the binary classification tree for estimating the propensity score. The blue line shows the result of CEM with stratifying arbitrarily and the black line visualises the performance of CEM if the stratification is automatised by the classification tree. The black points indicate the levels of the tree and the figures above the line denote the number of strata at that level of tree depth.

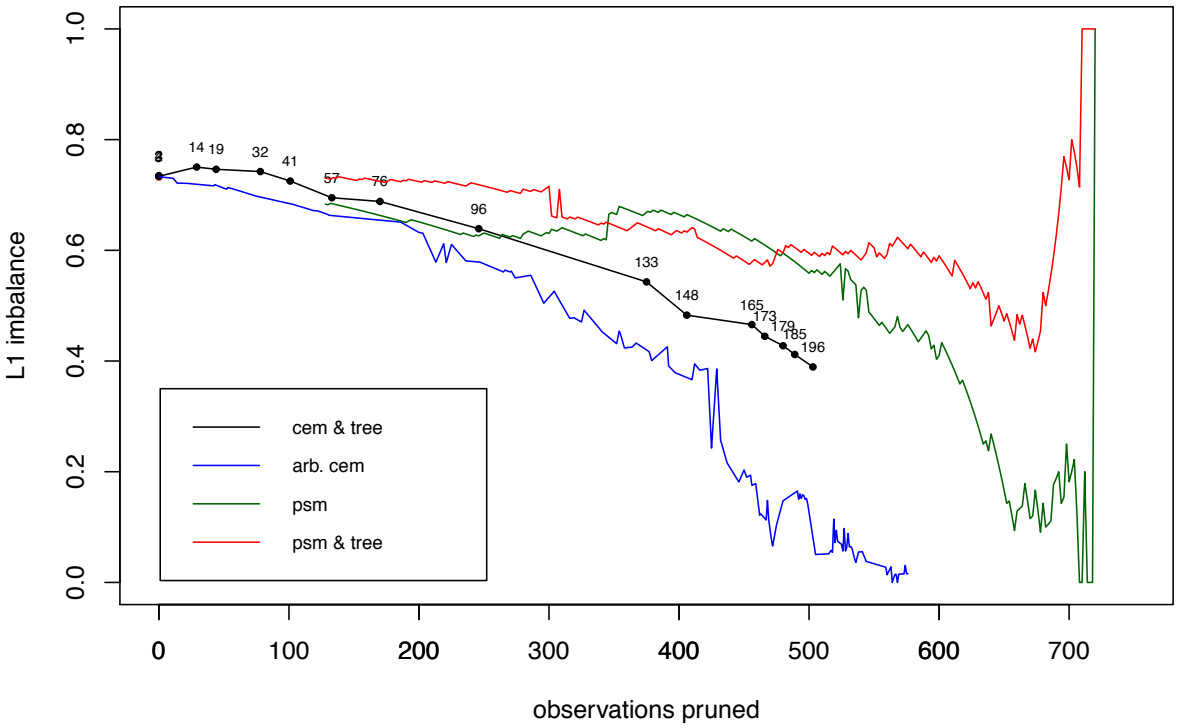

Figure 2: Imbalance-Matched Sample Size Graph, using data from LaLonde (1986)

Overall, both CEM specifications outperform the two PSM variations. The graphs for both CEM specifications are downwards sloping, indicating that the two methods reduce

imbalance as data is pruned. Although the arbitrary CEM performs best, the automated CEM is able to follow closely until more than half of the data (ca. 400 observations) is pruned. Thereafter, the blue line of the arbitrary CEM moves all the way down to complete balance. At this point, however, the prediction accuracy of the tree is already very low, considering that tree has grown to 12 levels with 148 terminal nodes, while the data includes only 722 observations in total. Note, that $\mathcal{T}_0$ has a maximal depth of 17 levels with 196 leafs. In addition, both graphs follow a relatively straight line which is a sign that both specification match the data well, considering that the data is experimental and treatment randomly assigned. Therefore, pruning data should not cause drastic movements in the imbalance.

The graphs of both PSM specifications begin after pruning the 128 excessive observations from the control group, such that the number of control units equals the number of treated units, $m_T = n_T$. Recall that there have to be at least as many treated units as there are control units for EPBR class matching methods. When matching the data with PSM, imbalance does not decline as continuously for either specification compared to CEM. Unlike identified by previous research, however, both methods manage to temporarily reduce some imbalance after 500 observations are pruned. Towards the end of the simulation the imbalance of both PSM specifications increases again and eventually reaches the maximum level of $L_1 = 1$. In addition, the logit specification temporarily reduces imbalance very drastically, but the graph does not stabilise. Both factors indicate that PSM matches and prunes randomly and the results confirm the findings from King and Nielsen (2016). Indeed, the one-dimensional projection of the hyper-dimensional distance estimated with a propensity score is unsuitable for matching. However, random matching and pruning does not begin immediately as it should for experimental data.[2] In addition, replacing the estimation of the propensity score with a non-parametric method does not improve the performance of PSM. On the contrary, applying a classification tree causes even higher levels of imbalance. Based on the evidence, I can confirm that it is the propensity score theory and not just misspecification alone which threatens the validity of PSM.

[2]When running the simulation multiple times, the results for both PSM methods are subject to extreme volatility especially towards the end. Indeed, applying the simulation in the framework of a Monte Carlo simulation would reduce the volatility, but this was not possible due to lack of computational power.

5.2 Comparing Post-Matching Results

Table 2 presents the post-matching estimations of the causal effect. Overall, the results from the matching simulation in figure 2 are confirmed. Both CEM specifications produce a better covariate balance than PSM.

I perform post-matching estimations at four different pruning stages which are summarised in panels B to E, while panel A shows the coefficient of the difference in mean estimator for the unmatched data as a baseline. Here, column three denotes the true causal effect of 886 USD that was originally estimated by LaLonde (1986). It states that male participants earned on average 886 USD more after completing the NSW training programme. When including the controls in the regression, the effect reduces to 824 USD as presented in column 4. Panel B presents a very low pruning stage. Yet, it is the coarsening level suggested by the pruned classification tree $\mathcal{T}_\alpha$ with $|\mathcal{T}| = s = 3$. Panel C shows the estimations after 128 observations are pruned, which is also the initial point for both PSM specifications. In panel D, I present the estimates after roughly pruning half of the data (361) and panel E shows the pruning stage close to the maximum depth of the tree.

Column one shows the L_1 measure of the matched data and column two depicts the number of pruned observations. Recall, that the exact number varies for both CEM variations in MIB class methods. Here, pruning is a result of the coarsening process and cannot be controlled. Therefore, I choose the value closest to the threshold. In column three, I present the results by only regressing the outcome (real earnings in 1978) on the treated variable, while in column four, I also include the controls in the regression. According to matching theory, the matching process was successful if the results between including and excluding the covariates in the regression do not vary.

The trade-off between variance and bias is observable for all matched subsets that were brought forward to the causal estimation. As more observations are pruned, the data becomes more balanced, however at the cost of higher standard errors. Yet again, the data matched with CEM returns the more accurate results than PSM also in the post-matching stage.

In panel B, the arbitrarily matched data returns coefficients very close to the true effect of the unmatched data from panel A. At this stage, the CEM has only pruned two observations. The coefficients for the automated CEM are also very close to panel A, with 838 (without controls) and 834 (including controls) respectively. Yet, the pruned tree underestimates the true causal effect. In fact, stratifying on the base of the pruned tree achieves this balance

Table 2: Post-matching estimation results

	(1)	(2)	(3)	(4)
			Outcomes	
	L1	obs.pruned	no controls	including controls
	Panel A: Pre-matching baseline			
diff. in mean	0.779	0	886.304	823.655
			(472.086)	(468.462)
	Panel B: Low pruning			
arbitrary CEM	0.736	2	890.512	820.488
			(471.666)	(468.945)
CEM & Tree	0.733	0	837.771	833.669
			(477.43)	(472.041)
	Panel C: Mid-low pruning			
arbitrary CEM	0.667	135	1217.289*	1209.339*
			(509.069)	(508.267)
CEM & Tree	0.695	133	1708.463*	1663.539*
			(493.403)	(488.431)
PSM	0.684	128	956.92	822.68
			(524.679)	(519.614)
PSM & Tree	0.731	128	774.05	723.419
			(531.332)	(528.702)
	Panel D: Mid-high pruning			
arbitrary CEM	0.444	363	1014.54	1033.955
			(593.236)	(591.98)
CEM & Tree	0.543	375	2226.156*	2265.272*
			(583.046)	(588.098)
PSM	0.674	360	1063.655	1001.091
			(674.522)	(670.335)
PSM & Tree	0.635	360	574.259	659.284
			(688.925)	(696.505)
	Panel E: High pruning			
arbitrary CEM	0.153	500	1202.711	1200.673
			(687.082)	(691.577)
CEM & Tree	0.412	489	2738.138*	2433.632*
			(754.916)	(758.758)
PSM	0.559	500	985.305	753.352
			(742.926)	(742.227)
PSM & Tree	0.595	500	1357.251	1160.862
			(780.802)	(786.88)
Controls			NO	YES

Notes: Bootstrap standard errors in parentheses. * $p < 0.05$. L_1 denotes the value of the L_1 imbalance. Panel B, denotes the pruning level, suggested by the pruned tree. Included controls in column 4 are: age, education, re74, re75, married, nodegree, black, hispanic, u74 and u75.

by simply reweighting the data as no units are discarded. For panels C to D the coefficients remain quite equal, indicating that the coarsening choices made by the classification tree balanced the treatment and control group well. However, when the tree depth increases to 16 levels in panel E, the coefficients become more unequal again, which agrees with the methodology of decision trees. The predictive power is very low so close to the maximum tree depth (17 levels) that would indicate a perfect fit. Despite these positive results in terms of balancing the data, the causal effect is strongly over-estimated in panels D and onwards, reaching nearly 3,000 USD in panel E. An explanation could be, that the binary tree sorts the data such that low values are discarded first. Likewise, the arbitrary CEM balances the data well throughout panels C to E. Moreover, covariate balance increases with the number of observations pruned. Yet, the arbitrary CEM also increasingly overestimates the causal effect, although not as extremely as the classification tree. Despite these obvious shortcomings of the automated CEM, the standard errors are lower (except for panel E) compared to coarsening arbitrarily.

The two PSM specifications, on the contrary, are not able to successfully match the data. While the logit specification, with an estimated effect of 957 (without control variables) and 923 (including controls variables), is still close to the true causal effect, the propensity score estimated with the classification tree underestimates the effect. Thereafter, the effects – although relatively balanced for the logit model – vary from panel to panel, once underestimating the causal effect and then again overestimating it in the next panel. Both PSM methods can reduce the imbalance, again with the logit model better than the tree. Yet, the imbalance returned by PSM is always higher than for both CEM specifications except in panel C. The fluctuations as well as the high levels of imbalance are an indicator for randomness in the pruning and matching process.

Overall, table 2 confirms what the graphs in figure 2 have indicated before: CEM reduces imbalance more effectively than PSM and also returns the better matching results. Thus, I can confirm the findings by King et al. (2011) and King and Nielsen (2016), that CEM is the more efficient method and that propensity scores are ineffective for matching. Estimating the propensity score with a binary tree instead of a logit model does not improve the performance of PSM, unlike claimed in previous research (McCaffrey et al., 2004; Setoguchi et al., 2008; Wyss et al., 2014). Finally, the arbitrary CEM specification creates the lowest levels of imbalance. The causal effects estimated on the basis of the arbitrarily matched data deviate

the least from the baseline. Using a binary classification tree for the stratification does balance the data well. However, the returned values in panels C to E question the binning properties of the tree after going beyond its optimally pruned size.

6 Conclusion

The thesis proposes a data-driven and automated method for selecting the hyper-dimensional histogram in coarsened exact matching. I systematically compare this novel approach with competing matching specifications. Overall, I show that the arbitrary as well as the automated CEM clearly outperform all tested PSM specifications, replicating on the LaLonde dataset the results by King and Nielsen (2016). Propensity scores in practice fail to reduce imbalance and cause model dependence. In conclusion therefore, I cannot recommend PSM, at least in medium-sized samples and with high-dimensional datasets.

Despite these encouraging results, the post-matching estimations question the suitability of the recursive binary classification trees for stratifying the data. In fact, classification trees have a number of shortcomings which is why they are not the first hand choice in modern statistical analysis any more. Trees have weak small sample properties which can be seen in the present case with less than 1000 observations. Moreover, binary trees are not very robust and small divergences in the data can cause large changes in the appearance of the predicted tree (James et al., 2013). Hence, the prediction power of trees is limited compared to more advanced machine learning methods such as bagging, boosting, random forests, or clustering. Yet, classification trees represent a starting point for automating the stratification process in CEM. Therefore, this paper should be seen as an encouragement to test other and more powerful machine learning methods as well. These methods can be easily implemented in CEM as I have demonstrated for recursive binary classification trees.

References

Abadie, A. and Imbens, G. W. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2012). A Martingale Representation for Matching Estimators. *Journal of the American Statistical Association*, 107(498):833–843.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press: Princeton.

Athey, S. and Imbens, G. W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Banaji, M. R. and Greenwald, A. G. (2016). *Blindspot: Hidden Biases of Good People*. Bantam: New York.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press: London.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks: London.

Dehejia, R. H. and Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American statistical Association*, 94(448):1053–1062.

Dehejia, R. H. and Wahba, S. (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics*, 84(1):151–161.

Diamond, A. and Sekhon, J. S. (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics*, 95(3):932–945.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer: New York.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, 15(3):199–236.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

Iacus, S., King, G., Porro, G., et al. (2009). CEM: Software for Coarsened Exact Matching. *Journal of Statistical Software*, 30(13):1–27.

Iacus, S. M., King, G., and Porro, G. (2011). Multivariate Matching Methods That Are Monotonic Imbalance Bounding. *Journal of the American Statistical Association*, 106(493):345–361.

Iacus, S. M., King, G., and Porro, G. (2012). Causal Inference Without Balance Checking: Coarsened Exact Matching. *Political Analysis*, 20(1):1–24.

Imai, K. and Ratkovic, M. (2014). Covariate Balancing Propensity Score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economics and Statistics*, 86(1):4–29.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer: New York.

King, G. and Nielsen, R. (2016). Why Propensity Scores Should Not Be Used for Matching. *Unpublished Manuscript*.

King, G., Nielsen, R., Coberley, C., Pope, J. E., and Wells, A. (2011). Comparative Effectiveness of Matching Methods for Causal Inference. *Unpublished Manuscript*.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs With Experimental Data. *The American Economic Review*, 76(4):604–620.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving Propensity Score Weighting Using Machine Learning. *Statistics in Medicine*, 29(3):337–346.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods*, 9(4):403.

Morgan, S. L. and Harding, D. J. (2006). Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice. *Sociological Methods & Research*, 35(1):3–60.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference*. Cambridge University Press: Cambridge.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106.

Pearl, J. (2010). The Foundations of Causal Inference. *Sociological Methodology*, 40(1):75–149.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer: New York.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1976a). Inference and Missing Data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1976b). Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples. *Biometrics*, 32(1):109–120.

Rubin, D. B. (2008). Comment: The Design and Analysis of Gold Standard Randomized Experiments. *Journal of the American Statistical Association*, 103(484):1350–1353.

Rubin, D. B. and Stuart, E. A. (2006). Affinely Invariant Matching Methods with Discriminant Mixtures of Proportional Ellipsoidally Symmetric Distributions. *The Annals of Statistics*, 34(4):1814–1826.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating Uses of Data Mining Techniques in Propensity Score Estimation: A Simulation Study. *Pharmacoepidemiology and Drug Safety*, 17(6):546–555.

Smith, J. A. and Todd, P. E. (2005). Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? *Journal of Econometrics*, 125(1-2):305–353.

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1):1–21.

VanderWeele, T. J. and Hernan, M. A. (2013). Causal Inference Under Multiple Versions of Treatment. *Journal of Causal Inference*, 1(1):1–20.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Wyss, R., Ellis, A. R., Brookhart, M. A., Girman, C. J., Jonsson Funk, M., LoCasale, R., and Stürmer, T. (2014). The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, bCART, and the Covariate-balancing Propensity Score. *American Journal of Epidemiology*, 180(6):645–655.