

# Achieving Optimal Covariate Balance Under General Treatment Regimes\*

Marc Ratkovic

September 21, 2011

## Abstract

Balancing covariates across treatment levels provides an effective and increasingly popular strategy for conducting causal inference in observational studies. Matching procedures, as a means of achieving balance, pre-process the data through identifying a subset of control observations with similar background characteristics to the treated observations. Inference in a matched sample is unbiased and robust to model specification. The proposed method adapts the support vector machine (SVM) classifier to the matching problem. The SVM separates easy to classify observations from hard to classify observations, and only uses the hard to classify cases in estimating a decision boundary between two classes. The treatment levels for these hard to classify observations are estimated with some uncertainty, the hallmark of random assignment. A series of lemmas prove that these hard to classify observations are balanced, for both binary and continuous treatment regimes. Unlike existing methods, the proposed method maximizes balance across all covariates simultaneously, rather than along a summary measure of balance. The method accommodates both binary and continuous treatment regimes. The proposed method is applied to four prominent social science datasets: the effect of a job training program on income, the effect of UN interventions on conflict duration, the effect of changes in foreign aid on domestic insurgent conflict, and the effect of education on political participation. The method is shown to recover an experimental benchmark, retain more observations than its competitors, and avoid dichotomization of a continuous treatment.

---

\*I thank Kosuke Imai for continued support throughout this project. I also thank participants at Yale's ISPS Experiments Lunch seminar for useful comments and feedback.

# 1 Introduction

Researchers conducting causal inference with observational data often face two problems. First, observations may select into a particular treatment level, biasing the estimated effect of the treatment on the outcome. Second, results may be model-dependent, in that the inclusion of different sets of controls may lead to different inference. Matching is an increasingly popular method for addressing both concerns (Ho *et al.*, 2007). Matching identifies a set of untreated observations with similar observed pre-treatment characteristics to the treated observations. This is done in a pre-processing stage, prior to analysis, where a matched subset of the data is identified. Given this smaller, matched dataset, estimation of the treatment effect on the outcome is conducted as normal, perhaps with a regression model. Under assumptions of exogeneity, common support, and non-interference among units matching corrects for selection bias. Within a matched dataset, the treatment level is independent of the pre-treatment covariates, so the estimated treatment effect on the outcome remains stable across different control variable specifications. Modeling an outcome on a matched dataset offers the potential for unbiased causal inference that is robust to model specification.

All existing matching methods proceed in two steps. First, a measure of similarity is defined. Next, using this similarity measure, untreated observations are selected to maximize the similarity between the treated and control observations. The selected treated and untreated units are then used in the subsequent analysis. Common similarity measures include the difference in estimated probability of treatment, or propensity score (Rosenbaum and Rubin, 1983; Hansen, 2004); the Mahalanobis distance; the  $p$ -value of  $t$ - and  $KS$ -statistics for the distance between covariate distributions for the treated and untreated units (Diamond and Sekhon, 2005); and placement in the same bin of a histogram (Iacus *et al.*, 2011b).

Each similarity measure, and hence each matching method, faces a fundamental problem: matching along a similarity measure does not necessarily achieve covariate balance. The similarity measure *may* produce balance, but the result is not generally guaranteed. In practice, most matching methods require researchers to adjust and rerun the matching procedure several times, until a satisfactory degree of balance is achieved.

Furthermore, these methods cannot accommodate continuous treatments. Existing means for

handling a continuous treatment involve either dichotomizing the continuous treatment (Nielsen *et al.*, 2011; Kam and Palmer, 2011; Healy and Malhotra, 2009; Brady and McNulty, 2011) or parametric estimation of generalized propensity scores (Imai and van Dyk, 2004; Hirano and Imbens, 2005). Matching requires selecting a set of untreated observations that are similar to treated observations; with a continuous treatment, there is no natural baseline group with which to compare the selected observations.

Instead of attempting to achieve balance along a similarity measure, the proposed method directly maximizes covariate balance. The matched observations are balanced across either a binary or continuous treatment. Concerns over parametric assumptions are ameliorated through a nonparametric specification of the propensity function. The method is fully automated, so no user input or adjustments are necessary. I show that the resulting matched subset balances the joint, rather than simply marginal, distribution of covariates across treatment levels, without discarding treated units.

The proposed method works through adapting the support vector machine (SVM) technology to the matching problem. SVMs were originally developed as means of classifying a binary outcome (Cortes and Vapnik, 1995). Researchers commonly apply maximum likelihood (ML) methods, such as logistic regression, when the outcome is binary, but SVMs have been shown to outperform these ML methods in classification (Friedman and Fayyad, 1997; Scholkopf and Smola, 2001). SVMs achieve higher performance through discarding easy to classify cases when fitting the boundary between the two outcome classes, allowing a focus exclusively on the hard to classify cases. Easy to classify cases offer no empirical loss, as opposed to ML methods, where every observation offers *some* deviance.

The remaining, hard to classify cases have some uncertainty as to their predicted class. As a group, their covariates are independent of the treatment level, because any systematic dependence is used in classifying the observations. Uncertainty over treatment level and independence between treatment level and covariates are the defining characteristics of randomization. These hard to classify cases, as a group, are balanced. More formally, a series of lemmas are presented which prove that the parametric SVM balances the mean of the covariates across treatment level, while the nonparametric SVM balances the joint distribution of the covariates across treatment levels. The result holds for binary and continuous treatment regimes.

To illustrate the method, four different datasets are analyzed. The first two consider a binary treatment: a job-training program on income (LaLonde, 1986) and UN intervention during wartime on conflict duration (Gilligan and Sergenti, 2008). I show in the first example that, unlike its competitors, the proposed method is able to return experimental results when applied to experimental data. Second, the proposed method is shown to return a sufficiently large subset of the data to produce powerful results. In the case considered, there are only sixteen treated units, and methods that rely on one-to-one matching return only thirty-two observations, producing an underpowered analysis.

The next two analyses illustrate the method in the presence of a continuous treatment: shifts in foreign aid on the probability of conflict (Nielsen *et al.*, 2011) and education level on political participation in the United States (Kam and Palmer, 2008). Following current practice, Nielsen *et al.* (2011) dichotomize a continuous treatment. The proposed method is shown to identify, rather than assume, a threshold effect. The proposed method is then used to replicate and extend the original results of Kam and Palmer (2008). Attending college offers no significant effect, as Kam and Palmer find, but post-bachelor’s education does lead to significantly more political participation—a result lost through the dichotomization of the education treatment variable.

The paper consists of five sections. First, matching methods are introduced and placed within a causal framework. The most commonly used methods are discussed. Second, the support vector machine is introduced and related to the matching problem. Analytic results are obtained that illustrate each of the proposed method’s advantages. Third, the two binary-treatment analyses are presented. Fourth, the continuous treatment analyses are presented. Fifth, a conclusion follows.

## 2 Causal Inference, Matching, and Balance

This section situates causal inference within the Neyman-Rubin-Holland framework. A brief overview of current matching methods is provided.

### 2.1 The Neyman-Rubin-Holland Framework

The most common method of formalizing causal inference in political science is the Neyman-Rubin-Holland framework (Holland, 1986). In this framework, treatment effects are defined in terms of each individual observation’s difference in outcome under different treatment assignments.

Only one treatment can be administered to each observation, and hence only outcome can be observed per observation. This poses the “fundamental problem of causal inference.” Assuming no interference among units, and a correctly specified model of treatment assignment, the average causal effect can be identified.

More formally, denote the potential outcome of the  $i^{th}$  observation in a simple random sample as  $Y_i$ , with  $i \in \{1, 2, \dots, n\}$ . The potential outcome is a function of the treatment level,  $\tau_k$ , a random variable with distribution  $F_\tau$  and support  $\mathcal{T}$ . The outcome function maps a treatment to each observation’s potential outcome, denoted  $Y_i(\tau_k)$ . For a binary treatment,  $\mathcal{T} = \{-1, 1\}$ , with treated units assigned a value of 1 and untreated units assigned  $-1$ <sup>1</sup>. For a continuous treatment regime, the treatment may be either: the real number line,  $\mathcal{T} = \mathfrak{R}$ ; an interval of the number line ranging from  $a$  to  $b$ ,  $\mathcal{T} = [a, b]$ ; or a set of  $j$  ordered values,  $\mathcal{T} = \{1, 2, \dots, j\}$ . Though the potential outcome for each individual is defined for any treatment  $\tau_k$ , only a single value is observed for individual  $i$ , and this value is denoted  $\tau_i^k$ . Denote  $x_i \sim F_X$  as realizations of an  $m$ -dimensional random vector of pre-treatment covariates for individual  $i$ , observed for each individual.

The fundamental quantity of interest with a binary treatment is the treatment effect, given as  $TE_i = Y_i(1) - Y_i(-1)$ . The two most common estimands are the average treatment effect,  $ATE = E(TE_i)$ , and the average treatment effect on the treated,  $ATT = E(TE_i | \tau_i = 1)$ .

I make two common assumptions through this analysis. First, the Stable Unit Treatment Value Assumption posits non-interference among units, conditional on pre-treatment covariates. Next, the Strong Ignorability of Treatment Assignment Assumption assumes the potential outcomes independent of the treatment level, conditional on pre-treatment covariates:  $\tau_k \perp\!\!\!\perp Y_i(\cdot) | x_i$  (Rubin, 1990). Strong Ignorability also assumes every treatment has positive probability of occurring:  $0 < P(\tau_k | x_i) < 1$ . The ignorability assumption presumes no omitted variables and the model of treatment assignment is properly characterized. Under these non-interference and ignorability assumptions, the ATE and ATT are identified.

---

<sup>1</sup>Under a binary treatment regime, the treated units are normally assigned value 1 and the untreated units value 0. Support vector machines handle values of  $\{\pm 1\}$  more naturally, so I use these values for the treated and untreated units throughout.

## 2.2 Existing Methods for Achieving Balance

Three matching methods are most commonly used to identify a balanced subset of the data: propensity scores, genetic matching, and coarsened exact matching. The proposed method is evaluated in relation to these methods, so a brief overview of each existing method’s strengths and shortcomings follows.

The *propensity score* is the probability of receiving the treatment:  $e_i = P(\tau_k = 1|x_i)$ . Under non-interference and ignorability, the propensity score is a balancing score, in that  $\tau_k \perp\!\!\!\perp x_i | e_i$  (Rosenbaum and Rubin, 1983). The implementation of one-to-one matching on propensity scores follows directly from the analytic result. A logistic regression is used to estimate  $e_i$ , and each treated unit is paired with the control unit with the closest estimated propensity score. These pairs act as each others’ counterfactuals, and the subsequent analysis is conducted on the matched subset.

Propensity score matching suffers from three flaws. First, it does not accommodate continuous treatment regimes. Second, results may not be robust to different specifications of the propensity function (Smith and Todd, 2005). Third, balance along covariates implies balance along the estimated propensities, but not vice versa. The researcher does not know if the propensity function is properly specified without checking covariate balance, but balance cannot be checked without specifying a propensity function—this creates the “propensity score tautology (Ho *et al.*, 2007).”

*Genetic Matching* provides an alternative to propensity score methods. A genetic optimization algorithm is used to identify a subset of observations that minimize the discrepancy between the distribution of each covariate for the treated and control units (Sekhon and Mebane, Jr., 1998; Diamond and Sekhon, 2005). A weight is estimated for each covariate that maximizes the minimal  $p$ -value associated with the  $t$ - and  $KS$ -statistics associated with the discrepancy between treated and untreated marginal covariate distributions.

GenMatch outperforms propensity score matching in both simulations and on benchmark datasets (Diamond and Sekhon, 2005). Its primary shortcomings are twofold. First, it does not accommodate a continuous treatment regime. Second,  $p$ -values are sample-size dependent, biasing outcome towards smaller matched samples (Imai *et al.*, 2008).

*Coarsened Exact Matching (CEM)* has been proposed recently as an automatic matching method (Iacus *et al.*, 2011a,b). CEM “coarsens” the data, through binning along each covari-

ate. Within each bin of this multivariate histogram, control units are matched to treated units. The method is shown to be “Monotone Imbalance Bounding,” in that the maximal distance along any covariate between the treated and control groups is governed by a user-controlled (or default) value.

The method suffers from several shortcomings. First, as with propensity scores and GenMatch, continuous treatments are not accommodated. Second, some treated units may remain unmatched, and thus are subsequently dropped from the analysis. This shifts the estimand from an average treatment effect on the treated (ATT) to a local ATT. When the ATT is the estimand of interest, CEM fails. Second, the arbitrary nature of the binning may result in different estimated values. This is the method’s greatest strength and weakness. In the presence of substantive knowledge about appropriate coarsening for each variable, the method provides a rigorous means of matching. In the absence of such substantive knowledge, little concrete guidance is available beyond suggested defaults.

### 3 Support Vector Machines and Balance

This section presents the geometric intuition and analytic results for the proposed method. The method is shown to identify observations over which there is uncertainty as to treatment level—the hallmark of a randomized experiment. An objective function is provided, and the optimum is shown to identify a balanced subset of the data, for both continuous and binary treatment regimes.

#### 3.1 The Geometry of Support Vector Machines and Balance

Support Vector Machines (SVMs) are a powerful classification and regression technique (Scholkopf and Smola, 2001). The intuition relating SVMs to the matching problem is best considered in the case of a binary treatment, as illustrated in Figure 1. SVMs fit an optimal separating hyperplane between the two classes. Given the two classes,  $X$  and  $O$ , observations to the right of the separating hyperplane are estimated as class  $X$  and the rest as class  $O$ . Balanced cases lie near the separating hyperplane between the two classes (Rubin and Stuart, 2006; Crump *et al.*, 2006).

SVMs also generate a margin, which characterizes observations considered in fitting the separating hyperplane. Observations properly classified outside the margin are easy to classify, and therefore are assumed to carry no information about the shape of the boundary. They are liter-

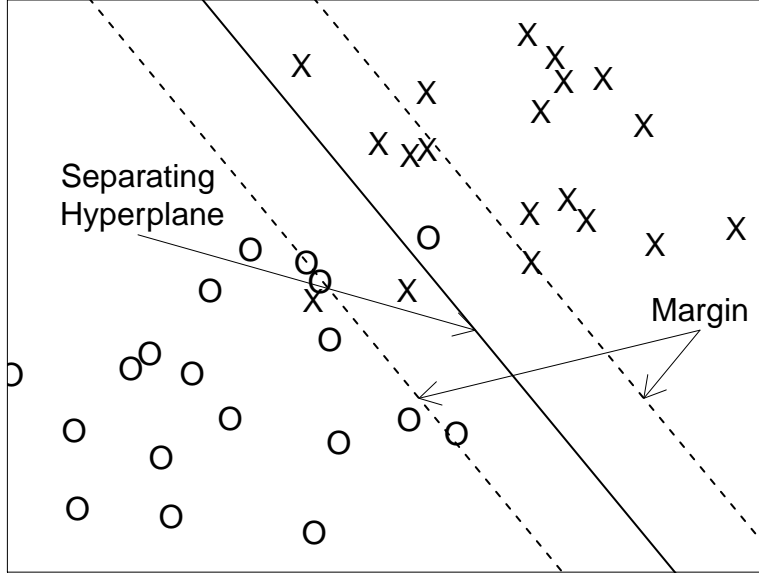


Figure 1: A support vector machine classifier generates a separating hyperplane between two classes, denoted  $X$  and  $O$  in this figure. Cases outside the margin are easy to classify, so they add no empirical loss. Only cases within the margin, i.e. those that are hard to classify, are considered in calculating the boundary. Cases within the margin have some uncertainty as to their estimated class assignment, which is the hallmark of random assignment. I show that these cases in the margin are balanced across the pre-treatment covariates.

ally dropped from the algorithm, and not considered in the fit. Observations correctly classified outside the margin add no empirical loss to the model; this is as opposed to logistic regression, where every observation contributes *some* deviance. This allows the algorithm to focus on hard to classify cases, and these marginal cases allow the hyperplane to not get distracted by the easy to classify cases.

The relation between a balanced subset of the data and hard to classify observations becomes clear. Observations for which the covariates  $x_i$  contain sufficient information are classified appropriately. The remaining cases are a subset of observations for which the covariates contain no information on the observed treatment level,  $\tau_k^i$ , else they would be easily classified. In other words, for these observations,  $\tau_k^i \perp\!\!\!\perp x_i$ , which is precisely the definition of balance.

SVMs separate cases into two different sets. Cases correctly classified outside the margin are estimated with no uncertainty. Cases incorrectly classified or within the margin are estimated with some uncertainty. Uncertainty over treatment level is the defining characteristic of random assignment. The cases identified by SVMs are balanced, in that they are a subset of the data that “looks like” random assignment to treatment level independent of observed pre-treatment



covariates has occurred. While there is no simple comparable figure for a continuous treatment regime, the results comes naturally from analyzing the SVM loss function.

## 3.2 The Analytics of Support Vector Machines

Support vector machines were originally formulated as a maximum margin classifier, which maximizes the distance between two classes in the parameter space (Cortes and Vapnik, 1995; Shawe-Taylor and Cristianini, 2004). SVMs produces a decision boundary between two classes, with observations on one side of the boundary classified as  $+1$ , and observations on the other side as  $-1$ . Two hyperplanes equidistant from the boundary are fit, with properly identified cases outside this area not considered in the fit. These properly identified cases outside the margin are easy to classify, and the remainder are hard to classify. The “margin” is the space between the two hyperplanes, and the width of the margin is maximized. This subsection first introduces the parametric SVM for the binary matching problem, extends it to the nonparametric binary matching problem, and then further extends the method to the continuous treatment regime.

### 3.2.1 The Parametric Balancing Problem with Binary Treatment

Assume a treatment,  $\tau_i \in \{\pm 1\}$ , and  $i \in \{1, 2, \dots, n\}$ . Each observation has a vector of  $m$  pre-treatment covariates  $x_i$ , with associated parameters  $\beta$ . The  $j^{th}$  elements of  $\beta$  and  $x_i$  are denoted  $\beta_j$  and  $x_{ij}$ .  $X$  is the  $n \times m$  observed matrix with rows  $x_i$ . Assume each covariate is centered on the values corresponding with treated units, so  $\sum_{\{i:\tau_i=1\}} x_{ij} = 0$  for all covariates,  $j$ . Taking the fitted values as  $\hat{\tau}_i = x_i' \hat{\beta}$ , the decision boundary and the set of observations in the margin,  $\mathcal{A}$ , are given as

$$\text{Decision Boundary:} \quad \hat{\tau}_i = 0 \quad (1)$$

$$\text{Marginal Observations:} \quad \mathcal{A} = \{\tau_i \hat{\tau}_i < 1\} \cup \{\tau_i = 1\} \quad (2)$$

Equation 1 gives the decision boundary. Observations with a positive  $\hat{\tau}_i$  are classified as  $+1$ ; those with a negative  $\hat{\tau}_i$  are classified as  $-1$ . The sets in equation 2 give the marginal, or hard to classify observations, denoted as the set  $\mathcal{A}$ . For example, consider an observation with  $\hat{\tau}_i = -2$  and  $\tau_i = -1$ . This observation is easy to classify, in that the observation’s fitted value is well below  $-1$ , and  $(-2) \cdot (-1) \not< 1$ . This observation is then disregarded when estimating the boundary.

The difference between the proposed method and the traditional SVM comes in the second

part of equation 2, where all treated units are maintained ( $\{\tau_i = 1\}$ ). An SVM will discard both treated and untreated units when fitting the classifier. The proposed method maintains all treated units so that the ATT may be estimated.

The loss function that generates a traditional SVM classifier is a “hinge-loss,” of the form  $|v|_+ = \max(v, 0)$  (Wahba, 2002). To adapt the support vector machine to that matching problem, I propose a “matching hinge-loss,” of the form

$$|1 - \tau_i \hat{\tau}_i|_+^{match} = \begin{cases} \max(1 - \tau_i \hat{\tau}_i, 0) , & \text{for all } i \text{ such that } \tau_i = -1 \\ 1 - \tau_i \hat{\tau}_i , & \text{for all } i \text{ such that } \tau_i = +1 \end{cases} \quad (3)$$

This forces each treated unit to be maintained in the fit, and produces an empirical loss function of the form

$$\mathcal{L}(\beta) = \sum_{i=1}^n |1 - \tau_i(x'_i \beta)|_+^{match} \quad (4)$$

The parametric SVM minimizes equation 4. The covariates of the marginal observations identified by the matching algorithm are balanced in means across treatment level, as given in Lemma 1:

LEMMA 1 *Mean-Balancing of the Parametric Support Vector Machine*<sup>2</sup>

*The minimizer of equation 4 selects a set of observations,  $\mathcal{A}$ , that balances the means of each covariate in expectation, such that*

$$E(x_{ij} | \tau_i = +1, i \in \mathcal{A}) = E(x_{ij} | \tau_i = -1, i \in \mathcal{A}) \quad (5)$$

*so long as there are treated and nontreated units in  $\mathcal{A}$ .*

Lemma 1 proves that, in expectation, the proposed method equates the means of the covariates in  $x_i$ . The next section extends the proposed method to a nonparametric SVM, such that matching occurs along the joint distribution of  $x_i$ , not simply the means.

### 3.2.2 The Nonparametric Balancing Problem with Binary Treatment

Extension of SVMs from a parametric to nonparametric, smoothing setting is straightforward (Hastie *et al.*, 2004). Given  $x_i$ , as above, an  $n$ -vector of smooth covariates are generated to

---

<sup>2</sup>A formal derivation of this lemma appears in Appendix A.

characterize the nonparametric component. These take the form of an  $n \times n$  matrix  $R$ , with rows  $r_i$  and elements  $r_{ii'}$  of the form

$$R = [r_{ii'}] = \phi(\|x_i - x_{i'}\|) \quad (6)$$

This nonparametric basis is composed of “radial basis functions,”  $\phi(\cdot)$ , where the value  $r_{ii'}$  is a function of the distance between  $x_i$  and  $x_{i'}$ . These bases characterize a smooth function given by the *joint* distribution of the  $x_i$ . As above, the covariates in  $x_i$  and  $r_i$  are centered on treated units, while the matrix  $R$  remains uncentered. Proper centering and characterizing the smooth function allows for the balance for the joint distribution of  $x_i$  across treatment levels.

The fitted values of the margin,  $\hat{\tau}_i = 0$ , now takes the form of  $\hat{\tau}_i = x_i' \hat{\beta} + r_i' \hat{c}$ , for  $n \times 1$  vector of coefficients  $\hat{c}$

The Representer Theorem (Kimeldorf and Wahba, 1971; Scholkopf and Smola, 2001) guarantees that the population minimizer of the matched hinge-loss can be written as

$$\underset{\hat{\tau}_i}{\operatorname{argmin}} E(|1 - \tau_i \hat{\tau}_i|_+^{match} | X) = x_i' \beta + r_i' c \quad (7)$$

The parameters  $(\beta, c)$  can be estimated through minimizing the objective

$$\mathcal{L}(\beta, c) = \sum_{i=1}^n |1 - \tau_i(x_i' \beta + r_i' c)|_+^{match} + \lambda c' R c \quad (8)$$

The final term in equation 8 serves to penalize the “curviness” of the resultant fit, and  $\lambda$  balances the tradeoff between a complex model that overfits the data ( $\lambda = 0$ ) and a too-simple parametric fit, that ignores the nonparametric component ( $\lambda \rightarrow \infty$ ). Wahba (1990) provides a full explication of this component.

With Lemma 1 in hand, proving mean-balance in covariates across treatment level, extending the claim to the nonparametric setting is straightforward. Balancing the joint distribution of  $x_i$  is equivalent to balancing in mean across the nonparametric bases. The covariates of the marginal observations identified by the matching algorithm are balanced in distribution across treatment level, as given in Lemma 2:

**LEMMA 2** *Distribution Balancing of the Joint Distribution of Pre-treatment Covariates by the Nonparametric Support Vector Machine*<sup>3</sup>

---

<sup>3</sup>A formal derivation of this lemma appears in Appendix A.

The minimizer of equation 4 selects a set of observations,  $\mathcal{A}$ , that balances the means of each covariate in expectation, such that

$$E_X(F(x_i)|\tau_i = +1, i \in \mathcal{A}, X) = E_X(F(x_i)|\tau_i = -1, i \in \mathcal{A}, X) \quad (9)$$

so long as there are treated and nontreated units in  $\mathcal{A}$ .

Lemma 2 differs from Lemma 1 in that it establishes balance across the joint distribution of covariates rather than their means, balancing  $E_X(F(x_i))$  rather than  $E_X(x_{ij})$ .

### 3.2.3 The Nonparametric Balancing Problem with Continuous Treatment

Balancing covariates nonparametrically along a continuous treatment can be accomplished with a slight adjustment to the loss function. The discussion and proof follow nearly identically from the case with the binary treatment. Since there is no referent group, covariates cannot be centered on the treated units. Instead of centering on treated units, covariates with a continuous treatment are centered on the sample mean of marginal observations. Specifically, define  $\tau_i^*$ ,  $x_i^*$ , and  $r_i^*$  the same as in the nonparametric binary case, except centered on the observations in  $\mathcal{A}$ . This gives fitted values of  $\hat{\tau}_i = x_i^* \beta + r_i^* c$ . Instead of the matching hinge-loss, define the continuous matching hinge-loss function as<sup>4</sup>

$$|(\tau_i^*)^2 - \tau_i^* \hat{\tau}_i^*|_+^{cont} = \max(0, (\tau_i^*)^2 - \tau_i^* \hat{\tau}_i^*) \quad (10)$$

By the Representer Theorem, the population minimizer of the continuous matching hinge-loss function can be written as

$$\operatorname{argmin}_{\hat{\tau}_i^*} E(|(\tau_i^*)^2 - \tau_i^* \hat{\tau}_i^*|_+^{cont} | X) = x_i^{*\prime} \beta + r_i^{*\prime} c \quad (11)$$

The parameters  $(\beta, c)$  can be estimated through minimizing the objective

$$\mathcal{L}(\beta, c) = \sum_{i=1}^n |(\tau_i^*)^2 - \tau_i^* (x_i^{*\prime} \beta + r_i^{*\prime} c)|_+^{cont} + \lambda c' R c \quad (12)$$

The covariates of the marginal observations identified by the matching algorithm are balanced in distribution across the levels of the continuous treatment regime, as given in Lemma 3:

---

<sup>4</sup>In the traditional SVM scenario, the continuous matching hinge-loss function is the  $\epsilon$ -insensitive hinge-loss, with  $\epsilon = 0$

LEMMA 3 *Balancing of the Joint Distribution of Pre-treatment Covariates by the Nonparametric Support Vector Machine with a Continuous Treatment*<sup>5</sup>

*The minimizer of equation 8 selects a set of observations,  $\mathcal{A}$ , that balances the means of each covariate in expectation, such that*

$$E_X(F(x_i)|i \in \mathcal{A}, X) = E_X(F(x_i)|\tau_i = E(\tau_i), i \in \mathcal{A}, X) \quad (13)$$

*for all values of  $\tau_i$  spanned by the observations in  $\mathcal{A}$ .*

These lemmas guarantee balance, across repeated sampling, for the cases identified by the proposed method. To show that these analytic results lead to improved performance on applied data, the proposed method is applied to a series of prominent social science datasets and compared to commonly used alternatives.

## 4 Empirical Applications with a Binary Treatment

The first two applied examples match along a binary treatment. In the first example, the proposed method is applied to a benchmark dataset, the National Supported Work program (LaLonde, 1986). The data come from a field experiment in which hard-to-employ individuals were randomly assigned to receive job support, or to a control group that received no support. Unlike existing studies, I apply each matching method to the randomized, experimental data. This is a baseline test of a matching method: pre-processing data with an already randomized treatment should not affect future inference. The proposed method is the only method that performs well by this measure.

In the second example, the proposed method is applied to a dataset from a recent study on the effect of UN intervention during wartime on conflict duration (Gilligan and Sergenti, 2008). The data contains only sixteen UN wartime interventions. GenMatch and propensity score matching return only sixteen controls observations, while CEM maintains only two treated observations. The subsequent estimation of the treatment effect is therefore statistically underpowered. Rather than sixteen control units, the proposed method returns seventy nine control observations, and statistically significant results are obtained.

---

<sup>5</sup>A formal derivation of this lemma appears in Appendix A.

## 4.1 Replicating Experimental Results: The National Supported Work Program

First, I apply the proposed methodology to the Manpower Demonstration Research Corporation’s National Supported Work (NSW) Program. The NSW study was conducted from 1975 to 1978 over 15 sites in the United States. Disadvantaged workers who qualified for this job training program consisted of welfare recipients, ex-addicts, young school dropouts, and ex-offenders. Participants were unemployed and had not maintained a job for more than three months of the past half year. The job training was randomly administered to 3,214 such workers while 3,402 belonged to the control group. This analysis focuses upon the subset of these individuals previously used by other researchers (LaLonde, 1986; Dehejia and Wahba, 1999). In this reduced sample, the size of the treatment and control groups is 185 and 260, respectively. The outcome of interest,  $Y_i(\cdot)$ , is the increase in earnings after the job training program (1978) compared to the earnings before the program (1975). The pre-treatment covariates include 1975 earnings, 1974 earnings, age, years of education, race (black, white, or Hispanic), marriage status (married or single), whether unemployed through 1974, whether unemployed through 1974, and whether a worker has a high school degree. The observational data used to generate a control comparison group is from the 1978 Panel Study for Income Dynamics, labeled PSID-1 in Dehejia and Wahba (1999). In the PSID-1 dataset,  $n = 2490$ , and all pre-treatment covariates in the experimental data are observed for each individual. All pre-treatment main effects are centered on the treated units, and square terms come from squaring the centered main effects.

In the first analysis, matching is conducted within the experimental data. This data does not need to be balanced, as treatment was randomly assigned. In practice, a matching algorithm should not greatly affect the characteristics of an already-balanced dataset. In the second analysis, the treated observations from the original study are matched to observations from the PSID-1 dataset.

Table 1 shows the  $t$ -statistics for the difference between treated and control observations in the experimental data (top) and with controls drawn from the observational PSID data (bottom). The first column contains the  $t$ -statistic within the randomized experiment, and the remaining columns contain these statistics for each matching method. The competing methods over-match

| <b>Controls Taken from Experimental Data</b>  |              |       |            |          |       |  |
|---|--------------|-------|------------|----------|-------|--|
|   | Experimental | SVM   | Prop Score | GenMatch | CEM   |  |
| Age   | -1.11        | -1.41 | -0.95      | -0.88    | 0.15  |  |
| Years of education                            | -1.44        | -1.31 | -0.41      | -0.23    | -0.99 |  |
| Black   | -0.46        | 0.64  | 0.14       | 0.00     | 0.53  |  |
| Hispanic                                      | 1.86         | 0.03  | 0.42       | 0.42     | -0.37 |  |
| Married                                       | -0.97        | -0.65 | 0.00       | -0.27    | -0.24 |  |
| No high school degree                         | 3.11         | 3.08  | 1.30       | 2.19     | 1.87  |  |
| 1974 earnings                                 | 0.02         | -1.01 | -0.77      | 0.18     | -0.44 |  |
| 1975 earnings                                 | -0.87        | -1.18 | -0.70      | 0.30     | -0.82 |  |
| Unemployed in 1974                            | 1.83         | 2.45  | 0.21       | 0.21     | 0.33  |  |
| Unemployed in 1975                            | 0.97         | 2.77  | 0.35       | -0.11    | 0.34  |  |
| Age <sup>2</sup>                              | -0.75        | -0.99 | -0.96      | -0.81    | 0.58  |  |
| Education <sup>2</sup>                        | -2.25        | -2.03 | -0.85      | -0.92    | -1.33 |  |
| 1974 earnings <sup>2</sup>                    | 0.66         | -0.02 | -1.06      | 0.01     | -0.04 |  |
| 1975 earnings <sup>2</sup>                    | -0.29        | -0.57 | -0.75      | 0.35     | -1.04 |  |
| Mean absolute $t$ -statistic                  | 1.18         | 1.30  | 0.63       | 0.49     | 0.65  |  |
| <b>Controls Taken from Observational Data</b> |              |       |            |          |       |  |
|   | Experimental | SVM   | Prop Score | GenMatch | CEM   |  |
| Age   | -1.11        | 1.86  | 4.15       | 6.77     | 0.81  |  |
| Years of Education                            | -1.44        | 1.18  | 0.02       | -0.23    | 0.54  |  |
| Black   | -0.46        | -1.75 | -3.61      | -4.85    | -1.31 |  |
| Hispanic                                      | 1.86         | 1.08  | 0.42       | -0.97    | 0.00  |  |
| Married                                       | -0.97        | 1.97  | 7.05       | 7.88     | 2.09  |  |
| No high school degree                         | 3.11         | -1.45 | -1.87      | -0.56    | -0.92 |  |
| 1974 earnings                                 | 0.02         | 1.43  | 5.46       | 3.94     | 2.86  |  |
| 1975 earnings                                 | -0.87        | 1.25  | 5.03       | 3.42     | 4.31  |  |
| Unemployed in 1974                            | 1.83         | -1.42 | -4.14      | -0.11    | -2.48 |  |
| Unemployed in 1975                            | 0.97         | -2.15 | -7.09      | -5.22    | -2.53 |  |
| Age <sup>2</sup>                              | -0.75        | 1.85  | 3.76       | 6.54     | 0.91  |  |
| Education <sup>2</sup>                        | -2.25        | 1.41  | 0.42       | 1.10     | 0.61  |  |
| 1974 earnings <sup>2</sup>                    | 0.66         | 1.05  | 2.92       | 1.95     | 2.37  |  |
| 1975 earnings <sup>2</sup>                    | -0.29        | 1.02  | 2.53       | 2.59     | 2.58  |  |
| Mean absolute $t$ statistic                   | 1.18         | 1.49  | 3.46       | 3.30     | 1.74  |  |

Table 1: Balance statistics for the National Supported Work data, with controls taken from the experimental group (top) and observational PSID data (bottom). Columns contain the  $t$ -statistic of the difference between the control and treatment groups. The bottom row of each section contains the mean absolute  $t$ -statistic across controls. The top section shows that SVMs are best able to replicate the mean absolute  $t$ -statistic, 1.30, versus 1.18 on the experimental data. The remaining methods overfit, since their mean absolute  $t$ -statistic is well below the experimental value of 1.18. On the observational data, the bottom row shows that SVMs are best able to replicate the mean absolute  $t$ -statistic, 1.49, versus 1.18 on the experimental data.

on the experimental data, returning  $t$ -statistics systematically below those in the experimental data. With the exception of CEM, these methods also under-match with the observational data, returning  $t$ -statistics well above the experimental benchmark.

In order to assess how well the matching algorithm solves the problems of selection bias and model dependence, the estimated treatment effect is collected for each method. The estimated effect is taken as the least squares coefficient on the treatment variable, in each dataset.<sup>6</sup> To assess model stability, all possible subsets of controls are included, and the densities plots of the estimated treatment effects across all control specifications appears in Figure 2.

The left half of Figure 2 contains the density plot of coefficient estimates for the experimental data. This is a baseline test for matching methods: pre-processing data with an already randomized treatment assignment mechanism should not affect future inference. The proposed method is the only method that performs well by this measure, capturing the distribution of the treatment effect across model specifications nearly perfectly. The right half of Figure 2 presents the same results, except with a control group taken from the observational PSID data. In this data, no method does particularly well; all are either biased (CEM, SVM) or estimated imprecisely (propensity score matching, GenMatch). The proposed method outperforms its competitors in the experimental data, and performs no worse in the observational data.

## 4.2 Keeping More Observations: UN Interventions during Wartime

This section replicates a portion of the analysis from a recent study that estimated the effect of United Nations interventions on civil war duration (Gilligan and Sergenti, 2008). The UN decision to intervene introduces an obvious and severe selection bias, and this bias must be accounted for before attempting causal inference. The original study provided separate estimates for peacetime and wartime interventions. Consistent with the broader literature (Fortna and Howard, 2008, e.g.), the authors find that the UN are better peacekeepers than peacemakers: interventions during peacetime increase time to the next conflict, while increase during conflict has no effect on duration.

I focus on the estimating the effect of UN intervention during wartime, as the original analysis suffered from a severe flaw induced by the matching method. There were only sixteen post Cold

---

<sup>6</sup>The exception is CEM. Since CEM discards treated units, the ATT estimates require a weighting of the selected treated observations. This is implemented through R library `cem` with command `att`.



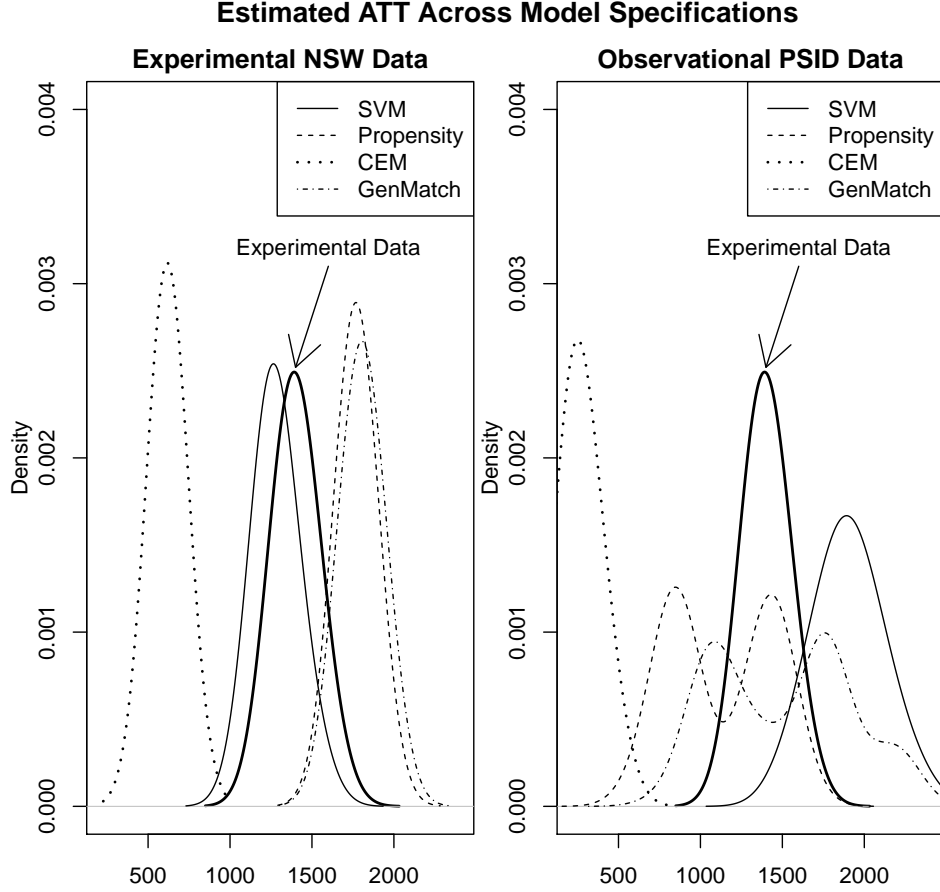


Figure 2: The density plots of estimated ATT for the benchmark National Supported Work study experiment, by matching method, for the experimental data (left) and observational data (right). The densities are generated by taking the point estimates of the ATT across all model specifications. Only the proposed method replicates experimental results when given experimental data (left), and the proposed method performs as well as its competitors when selecting untreated observations from the PSID survey (right).

War wartime interventions in the data, so the matched dataset analyzed by Gilligan and Sergenti contained only thirty two observations. The Cox proportional hazards model in the original study, with thirty two observations and fourteen covariates, is grossly underpowered. Gilligan and Sergenti find no significant effect, but report it as no effect.<sup>7</sup>

Statistical significance reflects sample size as well as effect size, and GenMatch returns a dataset with little power. The proposed method returns seventy-nine untreated units, allowing a more precise estimate of the effect of UN interventions. Both GenMatch and the proposed method have

<sup>7</sup>Gilligan and Sergenti (p. 111, 2008) state that “..within a matched sample, United Nations presence has no effect,” rather than reporting that they do not have sufficient data to identify the sign of the effect.

negative point estimates for the effect of UN interventions during wartime, but a larger matched set allows for significant results.

The full dataset consists of 1,227 post-Cold War civil wars, so  $n = 1,227$ . The treatment, UN intervention, occurs sixteen times, and the outcome of interest is the duration of the civil war. Because duration is right-censored for several observations, a Cox proportional hazards model is used to estimate effect size. Following Gilligan and Sergenti, each method is used to balance along six pre-treatment covariates: ethnic fractionalization, pre-observation war duration, log population, log mountainous area in each country, whether the war was funded with contraband, and indicator variables for Eastern Europe and Latin America. Additional confounders used in the subsequent analysis on war duration are: non-UN third party interventions, pre-intervention battle deaths, whether the war was a coup or revolution, whether the war is a “Sons-of-the-Soil” conflict between a state-supported migrant group and minority ethnic group at the country’s periphery, log number of military personnel, and regional indicator variables for Asia, Sub-Saharan Africa, and North Africa/Middle East. The theoretical motivation and complete description of which can be found in the original study (Gilligan and Sergenti (2008), pp. 106-107).

Results from the matching procedures are in Figure 3. The leftmost column contains the raw data, and severe imbalance appears across each variable except for contraband and the Latin America indicator. The proposed method returns a sample with reasonable balance along all covariates. All sixteen treated units and seventy nine untreated observations are retained, so  $n = 95$ . Propensity score matching and GenMatch both return balanced samples, each returning the sixteen treated observation and sixteen untreated observations, so  $n = 32$ . CEM retains only two treated and seventeen untreated observations,  $n = 19$ , rendering the subsequent analysis on conflict duration untenable.

Figure 4 presents the density estimates of the treatment effect for wartime UN interventions, for each matched dataset across all model specifications. CEM is not included in this figure, since the matched dataset contains nineteen observations, and the Cox proportional hazards model did not converge. Two effects stand out. First, in moving from the raw data to a matched sample, the civil war duration decreases. Second, across model specifications, the majority of estimates for both the proposed method and GenMatch are negative. In the model with all controls included, both methods agree on the negative point estimate, but only the proposed method returns a

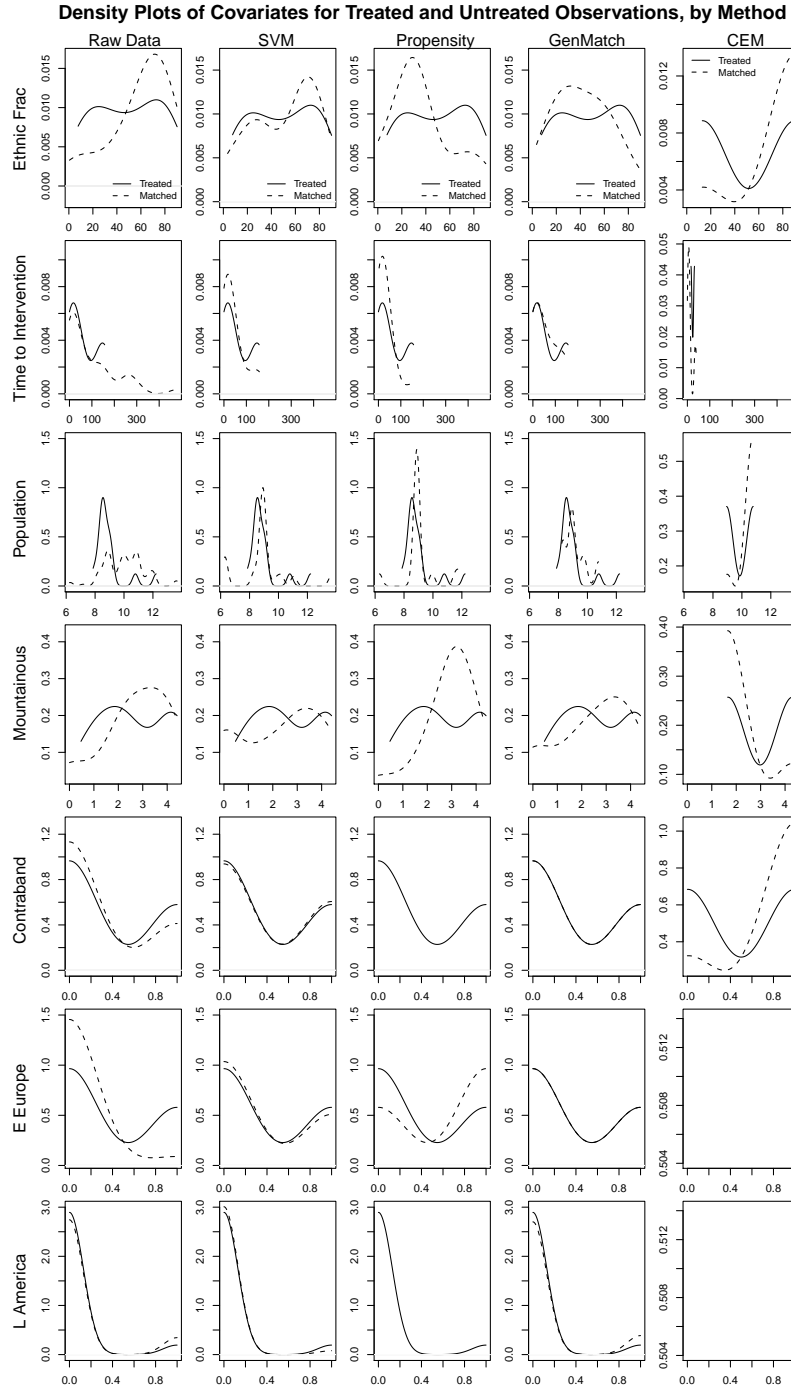


Figure 3: The marginal distributions of control variables for treated units (solid lines) and control observations selected by each method (dashed lines). There are only sixteen treated observations, so propensity scores and GenMatch identify only sixteen controls, leading to an under-powered analysis. CEM retains only two treated units. The proposed method selects seventy nine control observations, while maintaining all sixteen treated units. Since CEM selects two only treated observations, it is placed on a different  $y$ -axis than the remaining methods, which keep all sixteen treated units.

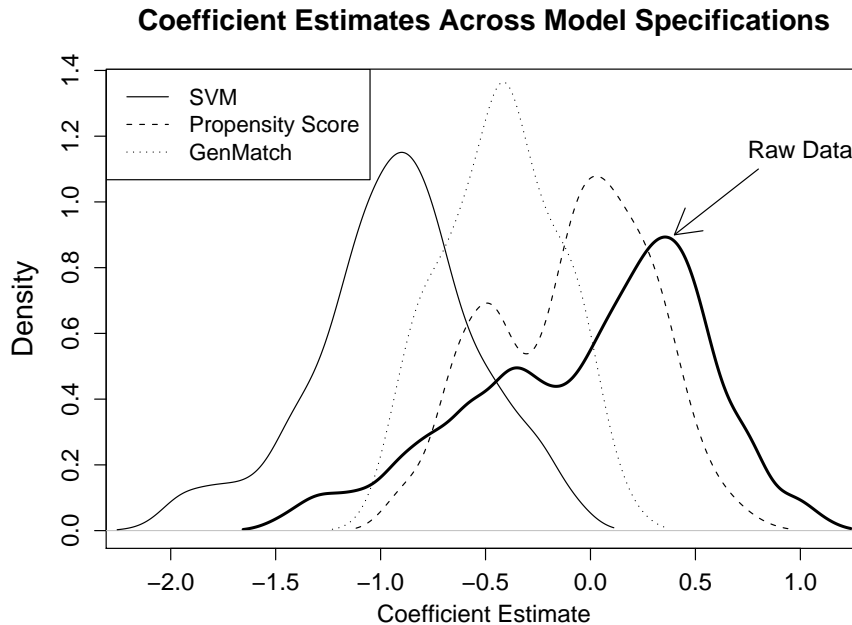


Figure 4: The density plots of estimated ATT for the UN intervention data, by matching method. The densities are generated by taking the point estimates of the ATT across all model specifications. CEM is not included, as only two treated units are selected and the remainder dropped. All of the matching methods shift the densities to the left of the raw data, suggesting a positive selection bias. Only the proposed method selects a sufficient number of control units, seventy nine, to produce statistically significant results.

statistically significant result (SVM:  $\hat{\beta} = -1.23, t = -2.60$ , GenMatch:  $\hat{\beta} = -0.18, t = .56$ ).

In the Gilligan and Sergenti data, the proposed method returns a matched sample of sufficient size as to allow significant results. In the NSW data, the method returns experimental results given experimental data. In both respects, it outperformed its competitors. The next section illustrates the proposed method’s strengths in balancing along a multi-valued treatment.

## 5 Empirical Applications with a Multi-valued Treatment

This section applies the proposed method to situations with a multi-valued treatment. When faced with a continuous or ordered treatment, current practice involves dichotomizing that treatment. Even when the dichotomization is motivated by theory, information is lost in reducing the multi-valued treatment to a binary variable. Both analyses in this section illustrate the value of retaining this information.

The first section analyzes data from a recent study on the effect of shifts in foreign aid on conflict (Nielsen *et al.*, 2011). Shifts in foreign aid are continuous; following common practice, the authors dichotomize this variable. Shifts below the fifteenth percentile are called “aid shocks.”

They then find that these negative aid shocks have an effect on insurgent conflict. Rather than assume a threshold and test its effect, the proposed method uncovers this threshold empirically.

The second analysis considers the effect of education on political participation in the United States. It has been long known that education level and political participation correlate, but the causal effect of education on participation is unclear. The effect may be due to those more likely to participate selecting into higher levels of education. A recent study dichotomizes the treatment, reducing the multi-valued education level to a binary variable for college attendance (Kam and Palmer, 2008). Like the existing study, I find that attending college has no discernible effect on political participation. Unlike the existing study, though, I find that post-bachelor’s education does have an effect on participation. This effect is lost through dichotomization, but recoverable through the proposed method.

## **5.1 Identifying a Treatment Threshold: Foreign Aid Shocks and the Probability of Conflict**

Shifts in foreign aid alter the balance of power between a domestic government and a rebel group. Not all such shifts are exogenous, as donors may choose to systematically adjust aid levels based on state characteristics. Common practice involves dichotomizing this continuous treatment, and a theoretical argument is needed to do so. Even a sound theoretical argument will only put a sign on the threshold, rather than give its precise location.<sup>8</sup> The proposed method, through not dichotomizing the continuous treatment, allows identification of the location of the threshold.

A recent study argues that aid shifts below a certain level, denoted aid shocks, are likely to increase the probability of conflict (Nielsen *et al.*, 2011). Shifts in aid change the relative bargaining power in favor of rebels, as the government has less money to spend on defense. Conflict in this scenario requires a bargaining failure as well as a shift in power (Fearon, 1995). The authors argue that, as rebels can demand more benefits in the presence of less aid, but governments cannot credibly commit to future aid streams, bargaining can break down and violence ensue. The authors posit a negative treatment threshold, a level of aid shift below which bargaining breaks down. Aid shifts are thus dichotomized, and renamed “aid shocks.” After balancing pre-

---

<sup>8</sup>In the absence of a method that balances along a continuous treatment, the authors demonstrate that their basic result is robust to different specifications of the assumed threshold.

treatment covariates across this binary variable, the authors find that negative aid shocks have a positive impact on the likelihood of rebel conflict. Rather than assume the location of a threshold, at the fifteenth percentile, the proposed method identifies the effect. This validates the major assumption underlying the analysis in Nielsen *et al.* (2011).

Data on conflict come from the UCDP/PRIO Armed Conflict Dataset (Gleditsch *et al.*, 2002), and data on foreign aid come from the AidData database (Nielson *et al.*, 2009). Data consist of 2,624 observations, taken across 139 countries, with observations annually from 1975 to 2009. Aid shift is the change in the ratio of aid commitments to country GDP, averaged over the previous two years. Aid shift is then lagged a year. Balancing covariates consist of measures of human rights violations, riots, general strikes, anti-government demonstrations, infant mortality, the number of adjacent countries with domestic conflict (“bad neighborhood”), log GDP per capita, log population, ethnic fractionalization, religious fractionalization, the nation having contiguous territory, logged mountainous area, and binary variables for partial autocracy, factional democracy, full democracy, and an observation during the Cold War. Cubic splines in time for each country are also included. For a full description of each variable, and the reason for its inclusion, see Nielsen (2011).

The results from a least squares regression, regressing aid shift on the balancing covariates in the original (left) and matched sample (right) can be found in Table 2. Using the unmatched data, regressing aid shift on the balancing covariates produces an  $R^2$  of 0.010 and a  $p$ -value of 0.31. Two of the balancing covariates are significant at the 5% level, factional and full democracy. Each type of democracy is more likely to receive a negative aid shift. In the matched dataset, neither covariate is significant. The  $R^2$  stays the same, 0.010, but the  $p$ -value moves to 0.99. As mentioned above, using  $p$ -values as a guide to balance confounds sample size and balance (King, Imai, Stuart 2008). I use it for two reasons. First, there is sufficient data left to maintain statistical power in the matched dataset— $n = 1,109$ . Second, there is no scholarly consensus for assessing balance with binary treatments, much less with a continuous treatment. Regardless, the large sample size and  $p$ -value close to 1 do indicate that the data is acceptably balanced along the continuous treatment.

As a second assessment of balance, I consider a quantile plot of  $p$ -values from regressing the treatment level, aid shifts, on the pre-treatment covariates in Figure 5. The reference distribution is a uniform random variable. The  $p$ -values that fall below the symmetry line indicate that the

|                         | Estimate | Pr(> t )            | Estimate | Pr(> t )   |
|-------------------------|----------|---------------------|----------|------------|
| Intercept               | 0.00     | 0.43                | 0.00     | 0.84       |
| Human Rights Violations | -0.97    | 0.02**              | -0.30    | 0.74       |
| Assassinations          | 0.37     | 0.37                | 0.22     | 0.76       |
| Riots                   | -0.13    | 0.70                | -0.05    | 0.95       |
| Strikes                 | 0.16     | 0.76                | 0.21     | 0.87       |
| Demonstrations          | -0.00    | 0.99                | -0.19    | 0.72       |
| Infant Mortality        | 0.00     | 0.98                | -0.00    | 0.94       |
| Bad Neighborhood        | -0.22    | 0.45                | -0.39    | 0.54       |
| Partial Autocracy       | -1.25    | 0.16                | -0.61    | 0.73       |
| Partial Democracy       | -1.50    | 0.12                | -1.29    | 0.52       |
| Factional Democracy     | -2.50    | 0.02**              | -1.57    | 0.48       |
| Full Democracy          | -2.62    | 0.02**              | -1.38    | 0.61       |
| Logged GDP per capita   | -0.22    | 0.66                | 0.09     | 0.94       |
| Logged Population       | 0.08     | 0.71                | -0.09    | 0.88       |
| Oil                     | -0.01    | 0.76                | 0.04     | 0.68       |
| Instability             | -0.32    | 0.71                | -1.07    | 0.52       |
| Ethnic Frac.            | 0.20     | 0.88                | 0.54     | 0.87       |
| Religious Frac.         | 0.12     | 0.93                | 0.89     | 0.79       |
| Non-contiguous          | 0.26     | 0.77                | 0.82     | 0.77       |
| Mountains               | 0.12     | 0.53                | 0.07     | 0.87       |
| Cold War                | -0.52    | 0.49                | 0.13     | 0.94       |
| n=2624, $R^2=0.010$     |          | n=1109, $R^2=0.010$ |          |            |
|                         |          | $p = .31$           |          | $p = 0.99$ |

Table 2: Results for regressing the level of aid shock on the controls used in Nielsen (2011), for the raw data (left) and on the observations selected using the proposed method (right). The  $R^2$  in the raw data is low: 0.01. This explains why matching does not change Nielsen’s, et al., results; they find the same results from matching as using a rare events logit. After matching, the  $p$ -value shifts dramatically, from 0.31 to 0.99, while maintaining a sufficient number of observations to maintain power in the subsequent analysis ( $n = 1109$ ). Cubic splines were included but not reported.

covariates are more informative than random noise, as they are smaller than we would expect if the  $p$ -values were draws from a uniform random variable (Benjamini and Hochberg, 1995). The raw data is almost balanced, though large parts lie slightly below the symmetry line. This is consistent with the low  $R^2$  on the left of Table 2. Since the matched data falls above the symmetry line, it is well-balanced.

In assessing the effect of aid shocks on conflict in the balanced data, Figure 6 shows the effect of shifts in aid on the probability of conflict, for the raw data (left) and matched data (right). A smoother was fit to the binary outcome, conflict, along the range of shifts in aid; dashed lines

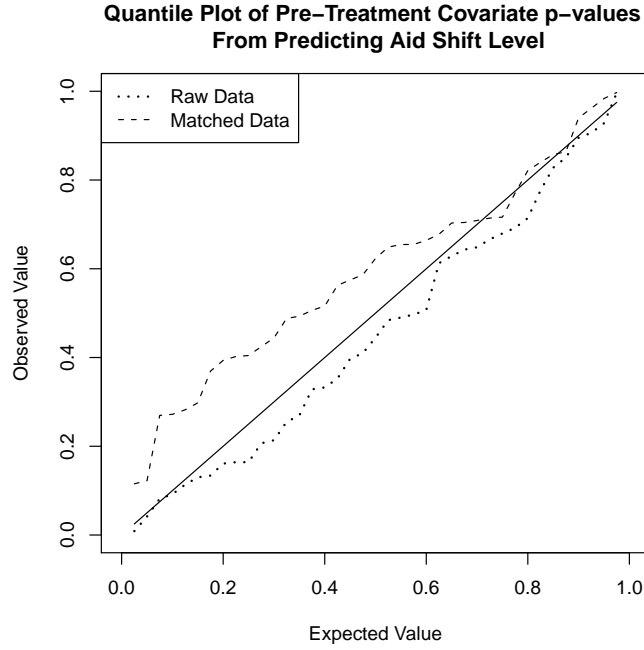


Figure 5: A quantile plot of  $p$ -values from regressing the treatment level, aid shifts, on the pre-treatment covariates. The reference distribution is a uniform random variable. The  $p$ -values that fall below the symmetry line indicate that the covariates are more informative than random noise. The raw data is almost balanced, though large parts of the distribution lie slightly below the symmetry line. Since the  $p$ -values from the matched data falls above the symmetry line, their covariates are well-balanced.

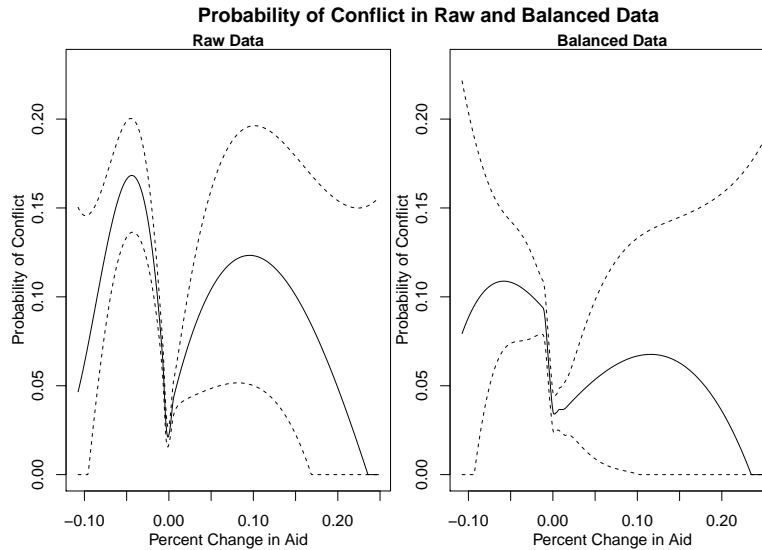


Figure 6: A smoother, comparing the probability of conflict for the raw data (left) and matched data (right). The discontinuity in aid shocks used by Nielsen *et al.* (2011) is apparent in the matched data (right).



present standard error estimates. The original analysis rests on a presumption of a threshold, and the results are conditional on the accuracy of this assumption. Rather than assume the threshold, the right side of Figure 6 displays a distinct threshold in the data.

The proposed method has been shown to reduce the number of assumptions in the analysis of a continuous treatment, lending further credence to the result that negative aid shocks increase the probability of conflict. While the proposed method helped to further reinforce the earlier results, and validate a key assumption in the original analysis, the next example extends the earlier findings of a prominent study in American political behavior.

## 5.2 Avoiding Dichotomization: The Effect of Education on Political Participation

That the higher-educated participate more in politics has long been a stylized fact in studies of American political behavior (Verba *et al.*, 1995, e.g.). This claim is not causal, as those more likely to participate may be selecting into higher levels of education. A recent study utilized propensity score matching in order to account for this selection bias and found no causal effect of attending college on political participation (Kam and Palmer, 2008). The result comes from dichotomizing the measure for education level at whether a respondent attended college. I show that this dichotomization causes a loss in information. Like Kam and Palmer, I show that attending college itself does not have a discernible effect on participation, but the proposed method uncovers a positive effect for post-Bachelor's degrees that is not discernible with current matching methods.

Kam and Palmer analyzed political participation in 1975 and 1982 for students who were high school seniors in 1965, with results from the Political Socialization Panel Study (Jennings and Niemi, 1991, HSSPS,). Kam and Palmer also present the results from a study independent of the first, the High School and Beyond survey results. In each instance, Kam and Palmer dichotomize the multi-valued education variable as whether college is attended or not. To illustrate the proposed method, I focus on political participation in the 1982 wave of the HSSPS survey, but I do not dichotomize the education variable. I show that, after using the proposed method, there is no significant effect between attending college and participation, but there is an effect for advanced post-bachelor's degrees. This insight is lost through the dichotomization of the original treatment variable.

The ninety seven high schools participating in the survey were drawn from a nationally representative sample. Seniors from the Class of 1965 were sampled in 1965, 1973, and 1982, and the parents were interviewed in 1973. The 1982 outcome, political participation, is an additive scale in eight political actions that consist of voting in 1976 or 1980, and additional means of involvement with a campaign or local official (see Kam and Palmer (2008), p. 624, for a complete description). The ordinal treatment variable takes on integer values from 0 to 7, with values corresponding to: no college, an associate's degree, a bachelor's degree, a master's degree, a theological degree, a law degree, a medical degree, and a doctorate. The original study used 203 covariates in the propensity function;<sup>9</sup> I reduce this number down to twenty five main effects and fifteen quadratic terms, for a total of forty covariates. The main effects include youth covariates for political knowledge, whether they plan to attend school next year, grade point average, their role as an officer, their role in a publication, hobbies clubs, occupational clubs, neighborhood clubs, religious clubs, youth organizations, and whether they have a phone. Parent covariates include education, father's income, household income, political knowledge, whether the father is employed, whether they own a home, and covariates for race (white and black). Quadratic terms for non-binary covariates are included.

Table 3, in Appendix B, contains the results from a regression of the continuous treatment, level of education, on the balancing covariates. In moving from the raw data to the matched data, the  $R^2$  is cut nearly in half, from 0.232 to 0.124, and the  $p$ -value moves from near zero to 0.189. As discussed with the Gilligan and Sergenti data, there is no accepted way to gauge balance with a continuous covariate. The proposed method, though, has selected a subset of the data with less of a systematic relationship between the treatment level and the balancing covariates.

Similar to the previous example, Figure 7 presents a quantile plot of  $p$ -values from regressing the treatment level, aid shifts, on the pre-treatment covariates. The raw data is poorly balanced, as the majority of its  $p$ -values are smaller than would be expected if there were no systematic relationship between the covariates and outcome. This is evidence by the  $p$ -values falling below the symmetry line. The  $p$ -values from the matched dataset fall above the symmetry line, so they are larger than would be expected under an assumption of random noise. The matched dataset is

---

<sup>9</sup>For a critique of the matching used by Kam and Palmer, see Henderson and Chatfield (2011) and Mayer (2011), with rejoinder (Kam and Palmer, 2011).

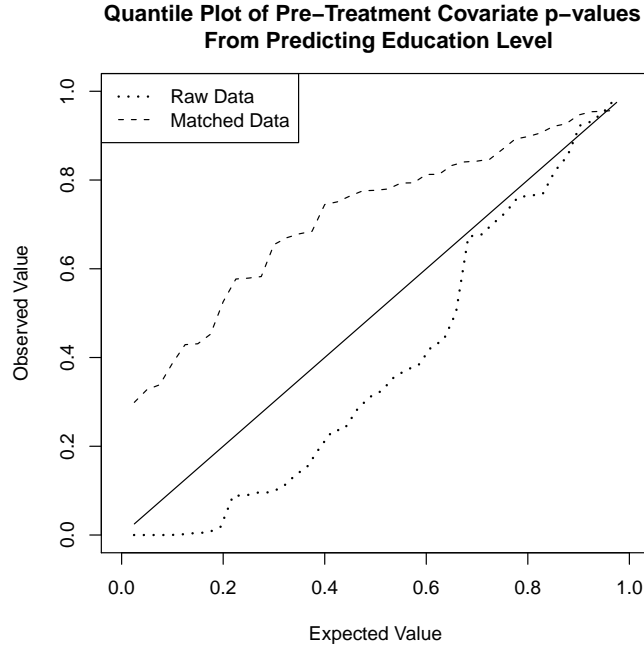


Figure 7: A quantile plot of  $p$ -values from regressing the treatment level, education, on the pre-treatment covariates. The reference distribution is a uniform random variable.  $p$ -values that fall below the symmetry line indicate that the covariates are more informative than random noise. The raw data is poorly balanced, since large parts of the distribution lie well below the symmetry line. Since the  $p$ -values from the matched data falls above the symmetry line, their covariates are well-balanced.

well-balanced.

Results for estimating the effect of education on participation can be found in Figure 8 and Table B. Figure 8 contains the mean participation level by education, with standard deviation bars. In the raw data (left), the stark linear relationship is clear. The more education one receives, the more they participate. The relationship between education level and participation in the matched data (right) provides more nuanced results. There is no discernible relationship among participation for the first four education levels: no college, an associate's degree, a bachelor's degree, and a master's degree (MBA, MS, MA). For the remaining levels, a theology degree, law degree, medical degree, or PhD, there is an effect.

The effect is borne out in the regression results from Table B, in Appendix B. The table contains regression results from regressing participation on the treatment level and controls, with the treatment characterized as the continuous index (left two columns) and dichotomized as in Kam and Palmer's study (right two columns). When left as a multi-valued index, education has a statistically significant linear relationship with participation levels ( $\hat{\beta} = 0.263$ ,  $p = 0.000$ ), but

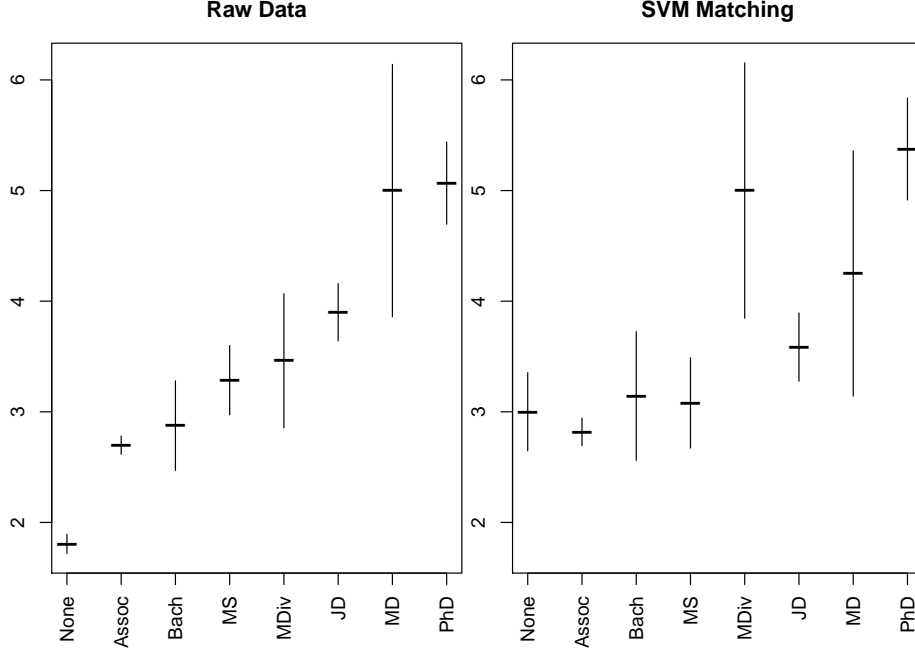


Figure 8: Political participation for the raw data (left) and matched data (right)

when the treatment is dichotomized, the effect disappears ( $\hat{\beta} = 0.170$ ,  $p = 0.606$ ). This effect of 0.17 is consistent with the estimated effect of 0.15 by Kam and Palmer (pg. 624, table 2).

The proposed method has been shown to replicate and extend the earlier analysis by Kam and Palmer. Like the original study, I find no causal relation between attending college and political participation. This result, though, depends on dichotomizing an ordinal variable, education level, so that existing matching methods may be accommodated. I show that leaving the variable as ordinal leads to a finer analysis. Advanced specialized degrees—divinity, legal, medical degrees, or a PhD—do have a causal effect that survives the matching procedure.

## 6 Conclusion

Matching methods have become an increasingly popular means for addressing selection bias and model-dependent inference. Existing matching methods suffer from two basic shortcomings. First, they do not guarantee balance of observed pre-treatment covariates along the treatment level. Second, they do not accommodate continuous treatment regimes. Common practice involves dichotomizing the treatment variable, which forces a loss of information.

The proposed method has been shown to account for both of these shortcomings. A series of lemmas show that matching through the proposed method produces, in expectation, a subset for

which the treatment level is independent of the pre-treatment covariates. The result extends to continuous as well as binary treatment regimes.

Applying the proposed method to a series of prominent social science datasets illustrates the method’s use and efficacy. The method is shown to recover an experimental benchmark, retain more observations than its competitors, and avoid dichotomization of a continuous treatment. The proposed method uncovered new results in the data that were masked due to the limitations of existing matching methods.

## References

- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 1, 289–300.
- Brady, H. E. and McNulty, J. E. (2011). Turning out to vote: The costs of finding and getting to the polling place. *American Political Science Review* **105**, 01, 115–134.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning* **20**, 3, 273–297.
- Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2006). Moving the goalposts: Addressing limited overlap in the estimation of Average Treatment Effects by changing the estimand. NBER Technical Working Papers 0330, National Bureau of Economic Research, Inc.
- 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**, 1053–1062.
- Diamond, A. and Sekhon, J. (2005). Genetic matching for estimating causal effects: A new method of achieving balance in observational studies. Tech. rep., Harvard University.
- Fearon, J. D. (1995). Rationalist explanations for war. *International Organization* **49**, 3, 379–414.
- Fortna, V. P. and Howard, L. M. (2008). Pitfalls and prospects in the peacekeeping literature. *Annual Review of Political Science* **11**, 283–301.
- Friedman, J. H. and Fayyad, U. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* **1**, 55–77.
- Gilligan, M. J. and Sergenti, E. J. (2008). Do UN interventions cause peace? Using matching to improve causal inference. *Quarterly Journal of Political Science* **3**, 2, 89–122.
- Gleditsch, N., Wallensteen, P., Eriksson, M., Sollenberg, M., and Strand, H. (2002). Armed conflict 1946-2001: A new dataset. *Journal of Peace Research* **39**, 5, 615–637.
- Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* **99**, 467, 609–618.

- Hastie, T., Rosset, S., Tibshirani, R., Zhu, J., and Cristianini, N. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research* **5**, 1391–1415.
- Healy, A. and Malhotra, N. (2009). Myopic voters and natural disaster policy. *American Political Science Review* **103**, 03, 387–406.
- Henderson, J. and Chatfield, S. (2011). Who matches? Propensity scores and bias in the causal effects of education on participation. *The Journal of Politics* **73**, 03, 646–658.
- Hirano, K. and Imbens, G. (2005). *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin’s Statistical Family*, chap. The Propensity Score with Continuous Treatments. John Wiley and Sons, Ltd, Chichester, UK.
- 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 (with discussion). *Journal of the American Statistical Association* **81**, 945–960.
- Iacus, S., King, G., and Porro, G. (2011a). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association* **106**, 189–213.
- Iacus, S. M., King, G., and Porro, G. (2011b). Causal inference without balance checking: Coarsened exact matching. *Political Analysis* .
- Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **171**, 2, 481–502.
- Imai, K. and van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* **99**, 467, 854–866.
- Jennings, M. K. and Niemi, R. G. (1991). Youth-parent socialization panel study, 1965-1973 (computer file). 2nd ICPSR ed. Ann Arbor: Inter-University Consortium for Political and Social Research.

- Kam, C. D. and Palmer, C. L. (2008). Reconsidering the effects of education on political participation. *The Journal of Politics* **70**, 03, 612–631.
- Kam, C. D. and Palmer, C. L. (2011). Rejoinder: Reinvestigating the causal relationship between higher education and political participation. *The Journal of Politics* **73**, 03, 659–663.
- Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* **33**, 1, 82–95.
- LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* **76**, 4, 604–620.
- Mayer, A. K. (2011). Does education increase political participation? *The Journal of Politics* **73**, 03, 633–645.
- Nielsen, R. A., Findley, M. G., Davis, Z. S., Candland, T., and Nielson, D. L. (2011). Foreign aid shocks as a cause of violent armed conflict. *American Journal of Political Science* **55**, 2, 219–232.
- Nielson, D., Powers, R., Tierney, M., Findley, M., Hawkins, D., Hicks, R., Parks, B., Roberts, J. T., and Wilson, S. (2009). AidData: Tracking development finance. Tech. rep. Presented at the PLAID Data Vetting Workshop, Washington, DC. <http://www.aiddata.org/research/releases> accessed February 15, 2010.
- 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. and Stuart, E. (2006). Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *Annals of Statistics* **34**, 1814–1826.
- Rubin, D. B. (1990). Comments on “On the application of probability theory to agricultural experiments. Essay on principles. Section 9” by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science* **5**, 472–480.
- Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.



- Sekhon, J. and Mebane, Jr., W. (1998). Genetic optimization using derivatives: Theory and application to nonlinear models. *Political Analysis* **7**, 189–213.
- Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.
- Smith, J. and Todd, P. (2005). Does matching overcome LaLonde’s critique of nonexperimental estimators? *Journal of Econometrics* **125**, 1-2, 305–353.
- Verba, S., Schlozman, K. L., and Brady, H. E. (1995). *Voice and Equality: Civic Voluntarism in American Democracy*. Cambridge: Harvard University Press.
- Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA, USA.
- Wahba, G. (2002). Soft and hard classification by reproducing kernel hilbert space methods. *Proceedings of the National Academy of Sciences* **99**, 26, 16524–16530.

## A Proof of Lemmas

This appendix derives the primary analytic results for the proposed method. It is proven that the proposed method identifies observations for which the treatment level is independent of the balancing covariates in distribution, for binary and continuous treatment regimes. The proofs are presented in order of increasing complexity. First, the parametric SVM with a binary treatment is shown to identify a subset of observations for which the mean of each covariate is, in expectation, the same for the treated and untreated units. Second, the result is extended to balance in mean for nonparametric radial basis functions. This is shown to be guarantee balance across the joint distribution of covariates. Finally, the result is extended to continuous treatment regimes.

Through the appendix, assume a simple random sample, with observations denoted  $i \in \{1, 2, \dots, n\}$ . A single actualization of a random treatment variable,  $\tau_k$  is assigned to each observation, with  $\tau_k \sim F_\tau$  and with support  $\mathcal{T}$ . Each observation has an  $m$ -vector of covariates,  $x_i$ , with  $j^{th}$  element  $x_{ij}$ , and  $x_i$  realizations from a distribution  $F_X$ , with compact support  $\mathcal{X}$ . Let  $\mathcal{A}$  denote observations not selected by the procedure. Finally, assume that all fourth moments of  $F_X$  and  $F_\tau$  are finite.

### A.1 The Parametric SVM with Binary Treatment

Denote fitted values from the proposed method as  $\hat{\tau}_i = x_i' \hat{\beta}$ . Assume a binary treatment, so  $\mathcal{T} = \{\pm 1\}$ , and covariates are centered on the treated units,  $E_X(x_{ij} | \tau_k = 1) = 0$ . For the parametric SVM with binary treatment, the minimized risk functional is

$$E_{X,\tau}(\mathcal{L}(\beta)) = E_{X,\tau}(|1 - \tau_i(x_i' \beta)|_+^{match}) \quad (14)$$

Taking the first derivative of equation 14 with respect to the  $j^{th}$  element of  $\beta$  gives the first order condition

$$E_{X,\tau}(x_{ij} \tau_i | i \in \mathcal{A}) = 0 \quad (15)$$

Conditioning on  $\tau_i$ , using  $E_X(x_{ij} | \tau_k = 1) = 0$ , and expanding results in

$$E_X(x_{ij} | \tau_i = 1, i \in \mathcal{A}) = E_X(x_{ij} | \tau_i = -1, i \in \mathcal{A}) = 0 \quad (16)$$

This proves Lemma 1, the mean-balancing of selected observations by the SVM.

## A.2 The Nonparametric SVM with Binary Treatment

Balance on the joint distribution follows from extending the results of Lemma 1 to a set of nonparametric bases that can be used to approximate  $F_X$ . Assume  $F_X(x_i|i \in \mathcal{A})$  is twice continuously differentiable, bounded away from zero and one, and its squared twice-iterated Laplacian integrated over its support is finite. Denote the Gaussian radial basis function  $\phi(x_i, x_j) = \exp(-\theta\|x_i - x_j\|^2)$ . Let  $r_{ij} = \phi(x_i, x_j) - E(\phi(x_i, x_j)|\tau_i = 1, i \in \mathcal{A})$ , which is simply  $\phi(\cdot)$  centered on the treated units. The Representer Theorem (Kimeldorf and Wahba, 1971) demonstrates that the conditional expectation of the multivariate distribution of  $x_i$  can be expressed as

$$E_X(F_X(x_i)|i \in \mathcal{A}, X) = \mu + \sum_{k=1}^n w_k r_{ik} \quad (17)$$

The elements of  $r_i$  in the fit are functionals of  $\phi(\cdot)$  evaluated at the data, centered on the treated units. Denote SVM fitted values  $\hat{\tau}_i = x_i' \hat{\beta} + r_i' c$ . Assume a binary treatment, so  $\mathcal{T} = \{\pm 1\}$ , and covariates and bases are centered on the treated units,  $E_X(x_{ij}|\tau_k = 1) = 0$  and  $E_X(r_{ij}|\tau_k = 1, X) = 0$ . For the nonparametric SVM with binary treatment, the minimized risk functional is

$$E_{X,\tau}(\mathcal{L}(\beta, c)) = E_{X,\tau}(|1 - \tau_i(x_i' \beta + r_i' c)|_+^{match} | X) \quad (18)$$

Taking the first derivative of equation 18 with respect to the  $j^{th}$  element of  $c$  gives the first order condition

$$E_{X,\tau}(r_{ik} \tau_i | i \in \mathcal{A}, X) = 0 \quad (19)$$

Expanding and using  $E_X(r_{ij}|\tau_k = 1) = 0$  gives

$$E_{X,\tau}(r_{ik} \tau_i | i \in \mathcal{A}) = 0 \quad (20)$$

$$\Rightarrow E_X(r_{ik} | \tau_i = 1, i \in \mathcal{A}, X) = E_X(r_{ik} | \tau_i = -1, i \in \mathcal{A}, X) = 0 \quad (21)$$

$$\Rightarrow E_X(w_k r_{ik} | \tau_i = 1, i \in \mathcal{A}, X) = E_X(w_k r_{ik} | \tau_i = -1, i \in \mathcal{A}, X) \quad (22)$$

$$\Rightarrow E_X\left(\mu + \sum_{k=1}^n w_k r_{ik} \middle| \tau_i = 1, i \in \mathcal{A}, X\right) = E_X\left(\mu + \sum_{k=1}^n w_k r_{ik} \middle| \tau_i = -1, i \in \mathcal{A}, X\right) \quad (23)$$

$$\Rightarrow E_X(F_X(x_i) | \tau_i = 1, i \in \mathcal{A}, X) = E_X(F_X(x_i) | \tau_i = -1, i \in \mathcal{A}, X) = E_X(F_X(x_i) | i \in \mathcal{A}, X) \quad (24)$$

$$\Rightarrow x_i \perp\!\!\!\perp \tau_i^k | i \in \mathcal{A}, X \quad (25)$$

This proves Lemma 2, that the SVM identifies observations for which the treatment level is independent of the covariates.

### A.3 The Nonparametric SVM with Continuous Treatment

This situation is nearly identical to the nonparametric case with binary treatment, except there is no natural referent group on which to center covariates. Instead, all variables, including the treatment, are centered on the selected observations. Let  $v^\star$  denote the variable  $v$  centered on its expectation among observations in  $\mathcal{A}$ . So, in the continuous case,  $r_{ij} = \phi(x_i, x_j)^\star$ . I assume that  $E(r_{ij}|\tau_i = E(\tau_i), i \in \mathcal{A}, X) = 0$ .

I make the same assumptions on  $F_X(x_i|i \in \mathcal{A})$  regarding the ability to approximate the distribution with radial basis functions as with the nonparametric binary case; all that changes is the nature of the centering. Any difference is absorbed into the mean term,  $\mu$ .

For the nonparametric SVM with continuous treatment, the minimized risk functional is

$$E_{X,\tau}(\mathcal{L}(\beta, c)) = E_{X,\tau}(|\tau^\star - \tau_i^\star(x_i^\star\beta + r_i'c)|_+^{cont}|X) \quad (26)$$

Taking the first derivative of equation 26 with respect to the  $j^{th}$  element of  $c$  gives the first order condition

$$E_{X,\tau}(r_{ij}(\tau_i - E(\tau_i))|i \in \mathcal{A}, X) = 0 \quad (27)$$

Expanding and using  $E(r_{ij}|\tau_i = E(\tau_i), i \in \mathcal{A}, X) = 0$  gives

$$E_{X,\tau}(r_{ij}(\tau_i - E(\tau_i))|i \in \mathcal{A}, X) = 0 \quad (28)$$

$$\Rightarrow E_X(r_{ij}|\tau_i, i \in \mathcal{A}, X) = E_X(r_{ij}|\tau_i = E(\tau_i), i \in \mathcal{A}, X) \quad (29)$$

$$\Rightarrow E_X(w_k r_{ij}|\tau_i, i \in \mathcal{A}, X) = E_X(w_k r_{ij}|\tau_i = E(\tau_i), i \in \mathcal{A}, X) \quad (30)$$

$$\Rightarrow E_X\left(\mu + \sum_{k=1}^n w_k r_{ij} \middle| \tau_i, i \in \mathcal{A}, X\right) = E_X\left(\mu + \sum_{k=1}^n w_k r_{ij} \middle| \tau_i = E(\tau_i), i \in \mathcal{A}, X\right) \quad (31)$$

$$\Rightarrow E_X(F_X(x_i)|\tau_i, i \in \mathcal{A}, X) = E_X(F_X(x_i)|\tau_i = E(\tau_i), i \in \mathcal{A}, X) = E_X(F_X(x_i)|i \in \mathcal{A}, X) \quad (32)$$

$$\Rightarrow x_i \perp\!\!\!\perp \tau_i^k | i \in \mathcal{A}, X \quad (33)$$

This proves Lemma 3, that the covariates of the selected observations are balanced across the continuous treatment.

## B Regression Tables

|                          | Estimate | Pr(> t )             | Estimate | Pr(> t ) |
|--------------------------|----------|----------------------|----------|----------|
| Intercept                | 1.313    | 0.007                | 1.962    | 0.039    |
| <i>Youth Covariates</i>  |          |                      |          |          |
| <i>Linear Terms</i>      |          |                      |          |          |
| Political Knowledge      | 0.122    | 0.009                | -0.101   | 0.269    |
| School Next Year         | 0.153    | 0.000                | 0.148    | 0.404    |
| GPA                      | -0.155   | 0.001                | 0.043    | 0.610    |
| School Officer           | -0.076   | 0.075                | 0.042    | 0.584    |
| School Publication       | 0.013    | 0.893                | -0.037   | 0.828    |
| Hobbies                  | -0.086   | 0.681                | 0.021    | 0.954    |
| School Club              | 0.166    | 0.039                | -0.099   | 0.451    |
| Occupational Club        | -0.034   | 0.628                | -0.043   | 0.739    |
| Neighborhood Club        | 0.041    | 0.617                | 0.022    | 0.879    |
| Relig Club               | -0.016   | 0.697                | -0.010   | 0.893    |
| Youth Org                | -0.028   | 0.765                | 0.096    | 0.545    |
| Misc Club                | -0.155   | 0.263                | 0.019    | 0.946    |
| Club Level               | 0.281    | 0.333                | 0.073    | 0.903    |
| Has Phone                | 0.007    | 0.864                | -0.080   | 0.429    |
| <i>Quadratic Terms</i>   |          |                      |          |          |
| Political Knowledge      | 0.020    | 0.594                | -0.069   | 0.359    |
| GPA                      | 0.006    | 0.879                | 0.024    | 0.758    |
| School Officer           | -0.117   | 0.010                | 0.047    | 0.532    |
| School Publication       | -0.039   | 0.692                | 0.026    | 0.874    |
| Hobby                    | -0.004   | 0.979                | 0.047    | 0.855    |
| School Club              | -0.106   | 0.172                | 0.072    | 0.552    |
| Occupational Club        | -0.043   | 0.573                | 0.072    | 0.602    |
| Neighborhood Club        | 0.000    | 0.998                | -0.009   | 0.954    |
| Relig Club               | 0.006    | 0.889                | -0.031   | 0.704    |
| Youth Org                | 0.099    | 0.236                | -0.155   | 0.280    |
| Club Level               | -0.076   | 0.693                | -0.125   | 0.756    |
| <i>Parent Covariates</i> |          |                      |          |          |
| <i>Linear Terms</i>      |          |                      |          |          |
| Education                | 0.070    | 0.059                | 0.023    | 0.739    |
| Father's Income          | 0.009    | 0.874                | -0.001   | 0.991    |
| Household Income         | 0.059    | 0.270                | -0.146   | 0.133    |
| Political Knowledge      | 0.001    | 0.967                | 0.034    | 0.632    |
| Employed                 | 0.042    | 0.384                | -0.043   | 0.641    |
| Father's Income          | -0.057   | 0.538                | 0.014    | 0.936    |
| Own Home                 | 0.084    | 0.038                | -0.010   | 0.903    |
| General Knowledge        | 0.014    | 0.782                | 0.052    | 0.559    |
| White                    | 0.069    | 0.549                | -0.063   | 0.754    |
| Black                    | 0.097    | 0.346                |          |          |
| <i>Quadratic Terms</i>   |          |                      |          |          |
| Education                | 0.159    | 0.002                | -0.262   | 0.006    |
| Knowledge                | 0.097    | 0.046                | -0.093   | 0.348    |
| Father's Income          | -0.057   | 0.538                | 0.014    | 0.936    |
| Household Income         | 0.135    | 0.134                | -0.092   | 0.578    |
| n=1051, $R^2 = 0.232$    |          | n=375, $R^2 = 0.124$ |          |          |
| $p < 2e - 16$            |          | $p = 0.189$          |          |          |

Table 3: Results from regressing the treatment, education level, on pre-treatment covariates in the raw data (left) and matched data (right). The  $p$ -value for the regression goes from near zero to 0.408, while still keeping over three hundred observations. This indicates that the treatment level is not systematically correlated with the pre-treatment covariates.

