Vector-Based Kernel Weighting: A Simple Estimator for Improving Precision and Bias of Average Treatment Effects in Multiple Treatment Settings

Melissa M. Garrido\textsuperscript{1,2}  
Jessica Lum\textsuperscript{1}  
Austin B. Frakt\textsuperscript{1,2,3}  
Steven D. Pizer\textsuperscript{1,2}

1. Department of Veterans Affairs  
2. Boston University School of Public Health, Boston, MA  
3. Harvard University School of Public Health, Boston, MA

This version: May 23, 2019

This work was supported by VA HSR&D IIR 16-140 (PI Garrido). The views expressed are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the US government.

An earlier version of this paper was presented at the Annual Health Econometrics Workshop in Baltimore, MD on September 28, 2018.

This is a work in progress. Please contact the authors for the most recent version of the paper prior to citation (melissa.garrido@va.gov).
Abstract

Treatment effect estimation must account for endogeneity, in which factors affect treatment assignment and outcomes simultaneously. By ignoring endogeneity, we risk concluding that a helpful treatment is not beneficial or that a treatment is safe when actually harmful. Propensity score matching or weighting adjusts for observed endogeneity, but matching becomes impracticable with multiple treatments, and weighting methods are sensitive to propensity model misspecification in applied analyses. We used Monte Carlo simulations (1,000 replications) to examine sensitivity of multi-valued treatment inferences to propensity score weighting or matching strategies. We consider four variants of propensity score adjustment: inverse probability of treatment weights (IPTW), kernel weights, vector matching, and a new hybrid that is easily implemented — vector-based kernel weighting (VBKW). VBKW matches observations with similar propensity score vectors, assigning greater kernel weights to observations with similar probabilities within a given bandwidth. We varied degree of propensity score model misspecification, sample size, treatment effect heterogeneity, initial covariate imbalance, and sample distribution across treatment groups. We evaluated sensitivity of results to propensity score estimation technique (multinomial logit or multinomial probit). Across simulations, VBKW performed equally or better than the other methods in terms of bias, efficiency, and covariate balance measured via prognostic scores. For instance, we tabulated the number of scenarios in which each method led to estimates with less than 20% bias. Across 756 analytic scenarios with a sample size of n=1200 and 3 treatment groups, VBKW led to estimates with less than 20% bias in 96% of scenarios, vector matching in 85% of scenarios, kernel weights in 48% of scenarios, and IPTW in 42% of scenarios. Among the estimates with less than 20% bias, VBKW had the lowest root-mean-squared error (0.038), compared to vector matching (0.046), kernel weights (0.044), and IPTW (0.049). VBKW estimates also had the lowest median absolute standardized differences in prognostic score values, indicating good covariate balance after propensity score adjustment. Our simulations suggest that VBKW is less sensitive to PS model misspecification than other methods used to account for endogeneity in multi-valued treatment analyses.
1. Introduction

Most propensity score guidance is restricted to methods for matching individuals with similar propensity scores across two groups (treatment, no treatment). Many treatments, however, have multiple levels. They may be continuous (drug dose) or categorical (several drug types). Consider, for example, a comparative effectiveness study that compares the ability of three interventions to prevent nursing home placement among individuals with functional and/or cognitive impairment: participant-directed home and community-based services, monthly home health aide services, and adult day care programs. If longitudinal data are not available, and if a valid instrumental variable is not available, a propensity score analysis that reduces bias due to observed confounding may be the best analytic strategy. In a traditional propensity score analysis, a series of dichotomous comparisons would be made: home and community-based services versus the other programs, home health aide services versus the other programs, home health aide services versus adult day care programs, and so on.

However, restricting treatments to binary indicators obscures between-group differences, including nonlinear relationships between treatment level and outcome (Cattaneo 2010). For instance, a binary comparison of adult day care versus the other two programs may not show a large difference in nursing home placement rates. However, if the three programs were to be compared simultaneously, one might find that adult day care is superior to home health aide services but not to participant-directed home and community-based services. Accounting for all values of a treatment variable in a single equation (rather than using several equations to make binary comparisons) helps ensure that a propensity score for a multi-valued treatment leads to treatment effect estimation among patients who have a non-zero chance of receiving any of the values of the treatment (i.e., that the assumption of common support is valid) (Rassen et al. 2013). For instance, in our example, a binary model of the probability of receiving participant-directed services will not distinguish between patients who have a chance of receiving adult day health care or home health aide services from patients who have a chance of receiving adult day health care but zero chance of receiving home health aide services. Inclusion of patients who have zero chance of receiving home health aide services in the analysis violates the assumption of common support necessary for propensity scores to reduce observed selection bias.

As the number of treatment groups increases, the option to estimate a single multinomial model, a generalized propensity score model, versus multiple binary models becomes more attractive. A generalized propensity score is the probability of receiving one treatment level, conditional on observed covariates. Each level is represented by a different propensity score (Imai & Van Dyk 2004; Imbens 2000).

The best way to use propensity scores for multinomial treatments is unknown; it is unclear when choice of a different weighting or matching strategy leads to divergent inferences. A popular method is inverse probability of treatment weights (IPTWs) based on the propensity score (e.g., McCaffrey, Griffin et al. 2013). However, IPTWs are sensitive to extreme values of the propensity score. If the extreme weights are caused by misspecification of the estimated propensity score model, use of IPTWs will lead to a biased estimate of the treatment effect. As the number of treatment groups increases, so does the likelihood of obtaining extreme values of the propensity score for at least one treatment group (Lopez & Gutman 2017).
Other options for incorporating generalized propensity scores into analyses include subclassification, matching, or kernel weights. Subclassification may not reduce selection bias as much as weighting when traditional numbers of strata (often 5) are used, and the optimal number of strata required to reduce selection bias may vary with sample size (Lunceford & Davidian 2004; Stuart 2010). Matching across groups becomes increasingly impractical, though not impossible (Rassen et al. 2013; Ratkovic 2012; Lopez and Gutman 2017), as the number of treatment groups increases; covariate distributions may not overlap well across multiple groups. Kernel weights, in which weights are assigned to comparison observations within a given bandwidth of the treated observation’s propensity score, are less popular than IPTWs but minimize the influence of extreme weights (DiNardo & Tobias 2001; Garrido et al. 2014). However, there is little guidance to facilitate choice between IPTWs, matching, or kernel weighting.

In empirical analyses, investigators often encounter situations that may affect the ability of a propensity score to reduce selection bias. Nonlinearity in the data generating process (DGP) for the true propensity score, number of treatment groups, distribution of sample across treatment groups, and treatment effect heterogeneity will interact to influence bias and efficiency of treatment effect estimates. The degree to which different weighting or matching strategies lead to robust inferences in messy empirical analytic scenarios with multiple treatment groups is unknown.

Here, we examine the extent to which inferences are likely to diverge among four methods of matching or weighting on the propensity score. We examine IPTW and kernel weighting as well as one variant of matching (vector matching) that looks especially promising for reducing covariate imbalance in studies of treatments with multiple levels (Lopez and Gutman 2017). However, vector matching is relatively complex to implement. Finally, we introduce a hybrid of kernel weights and vector matching that is easier to implement, which we term vector-based kernel weighting (VBKW).

In the following sections, we describe treatment effects of interest, weighting and matching strategies in greater detail, reasons we might expect inferences to diverge based on choice of weighting or matching strategy, and our Monte Carlo simulation design. We show that VBKW is an easy-to-use weighting method that improves bias relative to existing methods across a wide range of true propensity scores and distributions of the sample across treatment groups.

2. Treatment effects of interest
We are interested in bias and efficiency of average treatment effects (ATEs) and average treatment effects on the treated (ATTs). Again, consider a treatment with three levels: A, B, and C, and let \(E[Y_A]\) represent the estimated outcome when everyone in the sample receives level A. For this treatment, 3 ATEs (\(E[Y_A] - E[Y_B]\); \(E[Y_A] - E[Y_C]\); \(E[Y_B] - E[Y_C]\)) and 9 ATTs (each ATE evaluated among individuals who received a single treatment) can be estimated (McCaffrey, Griffin, et al. 2013). If the treatment effect is homogenous across the sample, the ATE and ATTs will be equal. In the more likely empirical case where treatment effectiveness varies by individual characteristics, the ATE and ATTs will not be equivalent. More formal definitions of our treatment effects of interest are presented in the Appendix.
3. Weighting and Matching Methods

3.1 Inverse probability of treatment weights (IPTWs)

In IPTWs, observations receive weights equal to the inverse of their propensity scores (Hirano, Imbens, & Ridder 2003; Imbens 2004; see Appendix for more details). Incorrectly estimated IPTWs may have extreme values, however, increasing treatment effect estimate variance (Stuart 2010). IPTWs permit average treatment effect (ATE) calculation but can be modified to calculate average treatment effects on the treated (ATTs) by assigning treated individuals a weight of one (called “weighting by the odds” or standardized mortality/morbidity ratio weighting) (Hirano et al. 2003; Ellis et al. 2013; Stuart, DuGoff, et al. 2013). We examined normalized IPTWs (Hirano et al. 2003). Our primary analyses do not consider IPTW adjustments such as trimming or truncating, as those methods employ arbitrary cut-points and may lead to estimates that are difficult to interpret (Harder, Stuart, & Anthony 2010; Stuart 2010; Lee, Lessler, & Stuart 2011). However, we do explore sensitivity of our results when we trim extreme weights that occur when propensity score values are less than 0.01.

As in all propensity score analyses, IPTW analyses are restricted to the range of common support. For IPTWs, this is often operationalized broadly as including observations between the maximum of the minima and the minimum of the maxima of each treatment group’s propensity score (Caliendo & Kopeinig 2008). For instance, consider a treatment with three levels (A, B, C): 1) Find the maximum of the minima of the propensity score for treatment A (p(A)) across each treatment group, 2) Find the minimum of the maxima of p(A) across each treatment group, 3) Drop any observation with a value of p(A) outside of the region identified by steps 1 and 2, and 4) Repeat steps 1-3 for p(B) and p(C) (Lopez & Gutman 2017).

3.2 Kernel weights (KW)

KW$s permit ATT calculation by assigning treated observations a weight of one and assigning weights to comparison observations within a given bandwidth of the treated observation’s propensity score according to a kernel function (DiNardo & Tobias 2001). By doing this, KW$s may have fewer extreme values than IPTWs. ATTs from different treatment groups can be combined to calculate ATE$s. Treatment effect estimates from KW$s are sensitive to the bandwidth, the range from a treated observation’s propensity score, used to construct the weight (Caliendo & Kopeinig 2008). Smaller bandwidths lead to more exact matches, which reduces bias. However, if the bandwidth is too small, fewer observations will be included in the analytic sample and variance will increase. We use a bandwidth that is dependent on the standard deviation of the logit of the propensity score (.2* standard deviation of the logit of the propensity score). We also consider a bandwidth identified as optimal because it has been shown to minimize mean squared error (MSE) of the estimated propensity score model over the sample without sacrificing smoothness of the estimator: a constant of 0.06 (Heckman, Ichimura & Todd 1997). KW$s are assigned according to kernel functions, where higher weights are given to comparison individuals with propensity scores most similar to treated individuals within the bandwidth. Weights are not as sensitive to choice of kernel function as they are to bandwidth (Caliendo & Kopeinig 2008), and we used the commonly used Epanechnikov kernel for all KW calculations (DiNardo & Tobias 2001; Busso, DiNardo, & McCrary 2009). The Epanechnikov kernel is the “optimal kernel” in that it minimizes MSE well at both interior and boundary points (Fan et al. 1997). We explored sensitivity of results to weight choice and repeated simulations with a Gaussian kernel. Weights are normalized to sum to one in each treatment group (Imbens
KW's are composed from observations with similar probability of treatment and similar probability of non-treatment. This could be interpreted as including observations with similar probability of all treatment levels, or as including observations with similar probability of receiving each of the treatment levels being compared. For a binary treatment, these two interpretations are equivalent. However, consider a multiple treatment case where we are interested in the effect of treatment A vs treatment B (but where there are others who received treatment C). In this case, weights could be constructed among observations with similar probability of treatment A and similar probability of treatment B, within the range of common support (so all observations have a non-zero probability of treatment A, B, and C). Alternatively, weights could be constructed among observations with similar probability of treatment A, similar probability of treatment B, and similar probability of treatment C (Lopez and Gutman 2017). We implement the first, broader, option in our KW construction. The second option relies on identification of similar vectors of propensity scores, which is part of vector matching (Lopez and Gutman 2017) and which we add to traditional kernel weight construction to create a new hybrid strategy, vector-based kernel weighting (VBKW).

3.3 Vector matching (VM)
As its name suggests, VM creates matches after identifying vectors of similar propensity scores across groups (Lopez and Gutman 2017). VM lends itself to the calculation of an ATT, but ATT estimates can be combined to create ATEs. Lopez and Gutman recently developed this procedure and found that it leads to better matches (lowest bias in covariate distribution among treatment groups) than common referent matching or IPTWs (2017). Matching within vectors ensures treatment effect estimates are applicable to observations with similar probability of receiving any of the treatments. However, this process becomes more difficult to implement as number of treatment groups increases, which makes it less likely that there will be available matches.

VM identifies similar vectors of propensity scores two ways when creating a matched set: first, by clustering, and then, by 1:1 greedy matching with replacement (Lopez and Gutman 2017). The ability to create good matches relies on several steps. After dropping observations outside of the range of common support, Lopez and Gutman recommend refitting the propensity score model and using k-means clustering on the logit of the propensity score for each treatment group to create strata with similar vectors of propensity scores. Clustering and matching is repeated as many times as there are treatment groups (3 rounds of clustering and matching for a treatment with 3 groups). For a treatment with values A, B, and C, treatment group A serves as the first reference group. Within strata formed from clusters of observations with similar values of the logit of p(C), observations from treatment groups A and B are matched based on values of the logit of p(A). Matches occur within a caliper of 0.2*SD(logit(p(A))). Then, within strata formed from clusters of observations with similar values of the logit of p(B), observations from treatment groups A and C are matched based on values of the logit of p(A). The observations from treatment group A that matched to observations in both other treatment groups, as well as the matches from groups B and C are retained. Similar steps are repeated with treatment group B and treatment group C as the reference groups. VM creates different matched samples, depending on the treatment effects of interest. For instance, when the reference group is A, we can calculate the ATT of A vs B and the ATT of A vs C, among observations with treatment A.
3.4 Vector based kernel weighting (VBKW)

In VBKW, weights are assigned based on a kernel function (as described above), but the weights are assigned to observations that have similar vectors of propensity scores for each treatment. VBKW includes elements of VM and kernel weighting, but it is simpler to implement than VM. Recall that in traditional kernel weighting, weights for an observation in treatment group A equal one, and nonzero weights for observations in treatment group B are assigned if their value of \( p(A) \) is within a bandwidth of the treated observation’s value of \( p(A) \). Values of \( p(B) \) and \( p(C) \) are not considered. In contrast, in VBKW, to ensure that weights are created for observations with similar vectors of propensity scores, we assign nonzero weights to observations in treatment group B if their value of \( p(A) \) is within a bandwidth of .2 multiplied by the standard deviation of the logit transform of the treated observation’s value of \( p(A) \), and if their value of \( p(B) \) is within a bandwidth of .2 multiplied by the standard deviation of the logit transform of the treated observation’s value of \( p(B) \), and if their value of \( p(C) \) is within a bandwidth of .2 multiplied by the standard deviation of the logit transform of the treated observation’s value of \( p(C) \). This means that nonzero weights are assigned to controls with a similar propensity score vector instead of just being similar on \( p(A) \). Rather than creating several subsets of matches, as in VM, VBKW creates one single subpopulation, allowing for easier comparison of estimated treatment effects. As is the case in kernel weighting and VM, ATT estimates from VBKW can be combined to form ATEs.

4. Reasons we might expect inferences to diverge based on choice of weighting or matching strategy

Researchers need guidance for the use of propensity scores in finite empirical samples. We wish to identify scenarios in which inferences are most likely to diverge. For instance, we expect estimates from kernel weights (low emphasis on extreme weights) to be less biased than estimates from IPTWs when the true data generating process for the propensity score is nonlinear. We expect differences in inferences to be more likely when the presence of extreme weights is more likely or when identification of matches may be more difficult. We expect this to occur when: 1) the true propensity score includes more nonlinearity and nonadditivity, 2) the number of treatment groups increases, 3) the sample size decreases, 4) the sample is distributed more unevenly across treatment groups, 5) when there are heterogeneous treatment effects, and 6) when there is greater pre-weighting imbalance in observed covariates across groups (Lee, Lessler, & Stuart 2010; Lee et al. 2011; Rassen et al. 2013; Setoguchi et al. 2008). Because the methods we compare vary in the number of dimensions that are factored into weight construction, and the number of dimensions does not change with sample size, we do not expect the methods to be asymptotically equivalent when treatments have multiple values.

5. Methods

We used simulated datasets to understand the relative ability of different propensity score matching and weighting strategies to reduce selection bias in treatment effect estimation. We ranked weighting and matching strategies’ ability to produce unbiased treatment effects.

We began with a simulation design with a known data generating process so that we can isolate and identify the influence of weighting/estimation strategies and analytic scenarios on treatment
We based our definitions of true propensity scores on an established simulation protocol where covariates are correlated and the true propensity score exhibits varying degrees of nonlinearity and nonadditivity (Table 1) (Lee et al. 2010; Wyss et al. 2014; Lee et al. 2011; Setoguchi et al. 2008; Austin 2012).

The initial true propensity score model was a multinomial logistic model of covariates (X₁-X₇), where X₁-X₄ were confounders and X₅-X₇ were associated with treatment only (Setoguchi 2008). All covariates were drawn from the standard normal distribution. Our outcome was a linear function of X₁-X₄, X₈-X₁₀, and treatment assignment (Lee et al. 2010). The true ATEs, E[Yₐ] - E[Y₉], E[Yₐ] - E[Yₙ], and E[Y₉] - E[Yₙ], were set to have values of: -0.1, -0.2, and -0.1, respectively. The true ATTs were equal to the true ATEs when treatment effects were homogenous. Following Setoguchi et al. 2008’s protocol, we set X₁ and X₅ to have a correlation coefficient of 0.2, and X₃ and X₈ to have a correlation coefficient of 0.2. In addition, both X₂ and X₆ were set to have a correlation coefficient of 0.9, as were X₄ and X₉. After setting the correlation coefficients, X₁, X₃, X₅, X₆, X₈, X₉ were dichotomized (X₁ₙew = 0 if X₁ ≤ X̄₁, X₁ₙew = 1 if X₁ > X̄₁).

To generate the true propensity score and create treatment groups, we calculated a multinomial logit model:

\[
p(A) = \frac{1}{1 + e^{X\beta_B} + e^{X\beta_C}}
\]

\[
p(B) = \frac{e^{X\beta_B}}{1 + e^{X\beta_B} + e^{X\beta_C}}
\]

\[
p(C) = \frac{e^{X\beta_C}}{1 + e^{X\beta_B} + e^{X\beta_C}}
\]

where \(X\beta_B = .2*(-.2 + X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7)\)

and \(X\beta_C = -.9*(-.1 + X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7)\)

Coefficients in this model were selected to reflect magnitudes often present in empirical analyses, and that were similar to those used in other simulation studies involving treatments with multiple levels. We tested sensitivity of results when the coefficient magnitudes were increased from .2 to .8 (relative risk ratio (RRR) from 1.22 to 2.23) and from -.9 to -1 (RRR of .4 to .037) in the linear predictor functions of treatments B and C, respectively.

To assign observations to treatment groups, we generated a random number from the uniform distribution that represents probability of treatment (denoted by j). Observations were assigned to treatment group A if j < p(A), to treatment group B if p(A) ≤ j < p(A) + p(B), and to treatment group C if j ≥ p(A) + p(B). In order for us to vary distribution of the sample across treatment groups (Table 1, line 3), we began with a sample size of 400,000. After generating treatment...
group assignment for the entire simulated dataset, we randomly drew observations from each treatment group. For instance, in our simulations of n=1200 with equal treatment distribution across three treatment groups, we randomly drew 400 observations from treatment groups A, B, and C.

For each true propensity score (all possible combinations of Table 1 characteristics), we used each weighting and matching strategy to estimate all possible ATEs and ATTs in sample sizes expected to occur in prospective observational cohort studies (n=1200), and in administrative data analyses (n=9,600) of empirical health services research questions. Each simulation consisted of 1,000 replications. Analyses were conducted in Stata version 14 (StataCorp 2015).

Here, estimated propensity scores were calculated using maximum likelihood estimation (multinomial logit regression and multinomial probit [Tschernis, Horvitz-Lennon, & Normand 2005]) on the main effects of X₁-X₆ and X₈-X₁₀. Because most practitioners use multinomial logit models, we focus on the results from those models here. Differences in inferences from weighting strategies when other estimation strategies are employed are left for future work (covariate balancing propensity scores [Imai & Ratkovic 2014], generalized boosting methods [McCaffrey, Griffin, et al 2013]).

5.1 Outcomes
We present outcomes for 756 unique analytic scenarios. We compare performance of our estimators when we vary simulations by 7 propensity score model misspecifications, 9 estimands (excluding ATT A vs B | T = C, ATT A vs C | T = B, and ATT B vs C | T = A), 3 types of sample distributions across treatment groups, 2 treatment effect distributions, and 2 coefficient sets (7 x 9 x 3 x 2 x 2): For each ATE and ATT, we measured bias (distance between the true treatment effect and mean estimated treatment effect, where smaller distances represent less bias). We also calculated absolute mean relative bias (bias as percent of true treatment effect) (Austin & Stuart 2017). Efficiency of each treatment effect estimate was evaluated by interquartile range (IQR) magnitude, root-mean-squared error (RMSE), and median absolute error (MAE). We counted the number of analytic scenarios in which each matching and weighting strategy produced estimates where absolute mean relative bias was < 20% to understand the frequency with which biased estimates would be obtained. Among those scenarios, we ranked matching and weighting strategies by efficiency. We also include measures of the absolute mean standardized differences in prognostic score values, where lower values indicate better covariate balance (Harder et al. 2010).

6. Results
We start by presenting results from simulations where n=1200 and the number of treatments (denoted by k) =3. We also present results from simulations where n = 9600 and k = 3.

6.1 Overall performance of IPTW, KW, VM, and VBKW¹
When n=1200 and k=3, estimates based on IPTW were more likely to be biased and inefficient

¹ Performance measures in tables are calculated via the following steps:
than estimates based on KW, VM, or VBKW (Table 2). Of the 756 analytic scenarios we ran for IPTW, only 314 (42%) produced estimates where median absolute mean relative bias across all analytic scenarios were less than 20%. VBKW-based estimates, in contrast, are the most likely to be unbiased and efficient. VBKW produced estimates where bias was less than 20% in 96% of the analytic scenarios.

Across all estimates (Table 2) and across estimates with <20% bias (Table 3), VBKW produced the most efficient estimates (median RMSE among estimates with <20% bias = 0.038, median RMSE from other strategies ranged from 0.046 - .049). Across all estimates, VBKW produced estimates with the lowest median absolute mean standardized differences in prognostic scores (median = 0.027, median across other strategies ranged from .045-.102). Across estimates with < 20% bias, KW produced estimates with the lowest median differences in prognostic scores (0.022), and VBKW produced estimates with the second lowest median difference (0.026).

6.2 Sensitivity to misspecification of the propensity score model
Regardless of weight type, the least biased and most efficient estimates were observed when the true propensity score included only additive main effects. As expected, bias was greatest when the true propensity score included moderate nonlinearity and nonadditivity (scenario G) that were not captured in the estimated propensity score model.

IPTW produces similar results to the other techniques in ideal settings (true propensity score only includes additive main effects, homogenous treatment effect, equal distribution of sample across treatment groups), but it is more sensitive to propensity score model misspecification than any of the other methods we examined (Figure 1). The median absolute mean relative bias of estimates produced by VBKW is less sensitive to propensity score misspecification than in IPTW, KW, or VM (Figure 2).

6.3 Sensitivity to distribution of sample across treatment groups
Again, VBKW estimates were the least sensitive (smallest changes in bias, as well as the lowest overall bias) to variations in the distribution of the sample across treatment groups. The magnitude of IQRs changed similarly across all weighting and matching strategies as treatment distribution became more skewed (results available from authors).

6.4 Sensitivity to treatment effect heterogeneity
In the presence of homogenous treatment effects and heterogeneity due to a non-confounding variable, both VBKW and VM were more likely to produce estimates with bias < 20% of standard deviation than IPTW or KW (results available from authors).

---

a) From each simulation, we calculated the average bias, average IQR, average RMSE, average median absolute error (MAE), and average of the mean standardized differences in prognostic scores across 1,000 replications. This leaves us with one statistic per simulation.
b) We then took the absolute value of the average bias, and absolute value of the average of the mean standardized differences in prognostic scores.
c) To make comparisons across analytic scenarios, we took the median of the absolute value of the average bias, average IQR, average RMSE, average MAE, and absolute value of the average of the mean standardized differences in prognostic scores. This leaves us with one statistic per estimator and are the numbers reported in the tables.
6.5 Sensitivity to coefficient choice
We tested sensitivity of results when the coefficient magnitudes were increased from .2 to .8 (relative risk ratio (RRR) from 1.22 to 2.23) and from -0.9 to -1 (RRR of .4 to .0.37) in the linear predictor functions of treatments B and C, respectively. Results (available from authors) were qualitatively similar. To generate larger baseline levels of imbalance in covariates, we generated all covariates as continuous. Again, results were qualitatively similar.

6.6 Performance across different estimands
VBKW and VM produced estimates with the lowest median absolute mean relative bias, regardless of the estimand of interest (See Figure 3 for homogeneous effects). In addition, they always generated estimates with relative bias < 20%, which was not the case for IPTW or kernel weights. In turn, this allows VBKW and VM to produce less biased estimates of transitive treatment effects (i.e., calculating the ATT of B vs C among observations receiving treatment A from the ATTs of A vs B and A vs C among observations receiving treatment A).

6.7 Performance with larger sample size
We tested the performance of IPTW, KW, VM, and VBKW in a larger sample size (n= 9600) with three treatment groups. These simulations focused on a subset of 48 analytic scenarios (true propensity score models A, C, E, and G, homogeneous treatment effects, 2 coefficient sets, 2 distributions of the sample across treatment groups, and ATEs only) (Table 4). VBKW continued to be the most likely to produce estimates with <20% absolute mean relative bias (in 100% of scenarios), whereas IPTW produced estimates with <20% relative bias in only 35% of scenarios. VBKW continued to produce the most efficient estimates in the larger sample size. For ease of comparison to our original sample size, we present the performance of each method in the same 48 analytic scenarios when n=1200 (Table 5).

6.9 Other sensitivity analyses
We obtained similar patterns of results when we used a fixed bandwidth for VBKW and KW, a Gaussian rather than an Epanechnikov kernel, multinomial probit models for propensity score estimation, and trimming of IPTWs (results available from authors).

7. Discussion
We investigated bias and efficiency of IPTW, KW, VM, and VBKW in analytic scenarios likely to be encountered in empirical analyses, including misspecified estimated propensity score models, treatment effect heterogeneity, and sample distribution across treatment groups. The commonly used IPTW strategy led to biased estimates more often than any other strategy we investigated. In nearly all scenarios, VBKW led to the least biased and most efficient estimates of the true treatment effect. It was robust to estimation via a multinomial logit model, which are used more commonly in applied analyses than multinomial probit models. These patterns persisted when we increased the sample size. Preliminary results with 5 treatment groups show similar results as presented here. These results suggest that VBKW may be less sensitive to propensity score model misspecification and sample distribution across treatment groups than other methods used to account for endogeneity in multi-valued treatment analyses. VBKW’s performance was only slightly better than that of VM, but it is simpler to implement.
Estimates based on IPTW were especially sensitive to degree of propensity score model misspecification and skewed distribution of the sample across treatment groups. In addition, neither KW nor IPTW estimates appear well-suited to produce unbiased estimates of transitive ATTs. Transitive ATT estimates are less likely to be biased when weights are constructed among observations with similar vectors of propensity scores (VBKW, VM) than when they are constructed among observations within a range of common support defined by the maximum of minima and minimum of maxima of propensity scores (IPTW, KW) (Lopez and Gutman 2017).

In order to develop useful guidance for empirical analyses, where bias cannot be ascertained, we included a measure of covariate balance across treatment groups in our simulated data (Harder et al. 2010). This allowed us to verify that the patterns of superior performance of VBKW are reflected in measures of covariate balance.

7.1 Limitations
These results, while promising, need to be evaluated in light of several limitations. First, our simulations were based on an imposed DGP rather than an empirical one, potentially inaccurately reflecting scenarios likely to be encountered in applied analyses. We will verify our results with plasmode simulations based on DGPs present in empirical data (Franklin et al. 2014; see more details below). However, we obtained similar results when we used alternate coefficients in treatment models and alternate ways of generating treatment groups.

In addition, we deliberately estimated a misspecified propensity score model with only main effects. A well-done propensity score analysis should ensure that the propensity score is leading to adequate covariate balance (Garrido et al. 2014), but we wanted to understand the degree to which results are robust to misspecification (and to potential lapses in analytic quality). This follows the pattern of previous propensity score simulation studies (e.g., Setoguchi et al. 2008).

Relatedly, we do not test sensitivity of results to covariate measurement errors, nor do we test performance when propensity scores are combined with covariates in doubly-robust estimates. These important factors may affect estimates’ bias and efficiency in finite samples (Stuart et al. 2010; Kang and Schafer 2007; McCaffrey, Lockwood, & Setodji 2013; Pearl 2009; Shadish 2013; VanderWeele & Arah 2011) and should be addressed after we have a better understanding of the relative performance of weighting and matching strategies for a given set of confounders.

7.2 Future directions
Future work will allow us to verify our results in simulations based on DGPs present in empirical data (plasmode simulations). A plasmode is a dataset based on empirical data generating processes that “has been constructed so that at least some aspect of the ‘truth’ of the data generating process is known” (Vaughan et al. 2009). Plasmode simulations were developed for genome and microarray research and are now being applied to electronic health data (Franklin et al. 2014; Franklin et al. 2017). Traditional simulations are often criticized for their artificiality, and empirical data analyses are limited by analysts’ inability to observe the true treatment effect. Plasmode simulations overcome these limitations by combining the benefits of a traditional simulation (known treatment effect that enables calculation of bias in treatment effect estimates) with the benefits of empirical data (empirical values of covariates and relationships among covariates are preserved). Plasmode simulations have the benefit of being derived from an
empirical DGP while allowing us to observe a true treatment effect and thus the degree to which each weighting or matching strategy leads to biased treatment effect estimates.

In future work, we will also determine the degree to which inferences diverge when we use nonparametric and semiparametric methods of estimating propensity scores. Values of propensity score weights vary with propensity score estimation method. As a result, treatment effect estimates obtained after propensity score weighting are sensitive to propensity score estimation method; this is well-documented in studies of binary treatments (Harder et al. 2010; Stuart 2010; Imai & Ratkovic 2014; Kang & Schafer 2007). Covariate balancing propensity scores (estimated with generalized method of moments) and propensity scores created through generalized boosting methods rely less on investigator trial and error than maximum likelihood estimation methods to create a propensity score that achieves covariate balance across treatment groups (Harder et al. 2010; Garrido et al. 2014; Dehejia & Wahba 1999).

Our plasmode simulations will also include a measure of robustness to residual confounding. A limitation of propensity scores is that they only adjust for observed, not unobserved, confounding. To that end, we will identify how much unobserved confounding would need to be present for each strategy in each simulation scenario in order for the inference from the analysis to change. For each simulation scenario, we will rank the ATTs and ATEs produced by each strategy by degree of robustness to unobserved confounding. We will do this by following a weighted adaption of Rosenbaum’s sensitivity analysis methods. (Liu, Kuramoto, & Stuart 2013; Rosenbaum 2002). For each treatment effect estimate, we will identify the smallest amount of unobserved selection bias that would need to be present to change the inference from rejection to acceptance of the null hypothesis of no treatment effect. Strategies that require relationships between an unobserved confounder and the treatment and between an unobserved confounder and the outcome to be stronger before inferences change are considered more robust.

8. Conclusion
When propensity scores are used in analyses of binary treatments, vector matching and weighting are implicitly conducted. Matching on the probability of being treated leads to matching on the probability of not being treated. If a treatment has more than two values, vectors need to be explicitly included in the creation of propensity score matches or weights. If they are not included, the propensity score for only one treatment group will be balanced, and estimates are likely to be biased and inefficient. VM and VBKW both lead to less biased and more efficient estimates than IPTW or KW that do not include vectors when there are more than two treatment groups. VBKW is relatively simple to implement and creates a single weighted subpopulation, facilitating comparisons of ATTs and ATEs among observations eligible to receive any of the treatments under consideration.
9. References


Rassen JA, Shelat AA, Franklin JM, Glynn RJ, Solomon DH, Schneeweiss S. Matching by propensity score in cohort studies with three treatment groups. Epidemiology 2013; 24: 401-409.


Rosenbaum PR. Observational studies (2nd ed.). New York: Springer; 2002


StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.


Table 1. Characteristics varied in Monte Carlo simulation and rationale for inclusion

<table>
<thead>
<tr>
<th>Characteristic: Rationale</th>
<th>Possible Levels</th>
</tr>
</thead>
</table>
| 1) Nonlinearity/nonadditivity in covariates included in propensity score (Lee et al. 2010; Lee et al. 2011; Setoguchi et al. 2008): | True propensity score is function of the following terms: a) $X_1, \ldots, X_{10}$ b) $X_1, \ldots, X_{10}, X_2^2$ c) $X_1, \ldots, X_{10}, X_2, X_4^2, X_7^2$  
  d) $X_1, \ldots, X_{10}, X_1X_3$, $X_2X_4, X_2X_5, X_3X_6$ e) $X_1, \ldots, X_{10}, X_1X_3$, $X_2X_4, X_2X_5, X_3X_6, X_5^2$  
  f) $X_1, \ldots, X_{10}, X_1X_3$, $X_2X_4, X_2X_5, X_3X_6, X_4^2$  
  g) $X_1, \ldots, X_{10}, X_1X_3$, $X_2X_4, X_2X_5, X_3X_6, X_5X_7, X_1X_6, X_2X_3, X_3X_4$  
  h) $X_1, \ldots, X_{10}, X_1X_3$, $X_2X_4, X_2X_5, X_3X_6, X_5X_7, X_1X_8, X_3X_4, X_2X_3, X_3X_4$  
  i) $X_1, \ldots, X_{10}, X_1X_3$, $X_2X_4, X_2X_5, X_3X_6, X_5X_7, X_1X_8, X_3X_4, X_2X_3, X_3X_4$ |
| 2) Number of treatment groups: Evaluate performance in common numbers of treatment groups expected in empirical research (e.g., k=4: drug A + placebo, drug B + placebo, drugs A+B, placebo). We expect CBPS and kernel weights (less influence of extreme weights) to lead to the least biased estimates. We expect differences in inferences across strategies to be more likely as k increases (making covariate balance more difficult). | $k = 3, 5^*$                                                                 |
| 3) Distribution of sample across treatment groups (Rassen et al. 2013): Understand how well strategies approximate counterfactuals for treated individuals when there is relatively little information from comparison individuals. Unequal treatment group sizes often occur empirically. As group size decreases, weights will be constructed from fewer observations and will have greater variance. As the distribution becomes more skewed, we expect inferences from estimates based on non-stabilized IPTW (greater sensitivity to variance in weights) to diverge more than inferences from estimates based on other strategies. | a) Equal split across groups  
  b) One treated group = 50% of observations, other groups split the remaining 50% equally  
  c) One treated group = 10% of observations, other groups split the remaining 90% equally |
| 4) Treatment effect heterogeneity: Understand how well strategies reduce bias in ATTs when ATTs are not expected to equal the ATE. We expect that differences in inferences that will arise with characteristics in rows 1-3 will be exacerbated in the presence of treatment effect heterogeneity (Rassen et al. 2013). | Coefficient on treatment variable in outcome equation is: a) Constant (c)  
  b) $cX_{10}^*$ (associated with outcome) |
| 5) Coefficient sets: Understand whether strategies perform differently when coefficient sets (and covariate imbalance across treatment groups) vary. | Coefficient magnitudes in linear predictors of treatments B and C, respectively, are:  
  a) 0.2, -0.9  
  b) 0.8, -1.0 |

*Simulations with 5 groups are in progress
Table 2. Summary of bias and efficiency of estimates across all 756 analytic scenarios

<table>
<thead>
<tr>
<th>Weighting or matching strategy</th>
<th>Number (%) of analytic scenarios with &lt;20% absolute mean relative bias</th>
<th>Median (min, max) absolute mean relative bias</th>
<th>Median absolute bias</th>
<th>Median IQR</th>
<th>Median RMSE</th>
<th>Median MAE</th>
<th>Median (min, max) absolute mean standardized differences in prognostic scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPTW</td>
<td>314 (42%)</td>
<td>30.96 (0.02, 482.16)</td>
<td>0.040</td>
<td>0.075</td>
<td>0.078</td>
<td>0.052</td>
<td>0.102 (0, 1.30)</td>
</tr>
<tr>
<td>KW</td>
<td>365 (48%)</td>
<td>20.97 (0.02, 143.91)</td>
<td>0.024</td>
<td>0.056</td>
<td>0.059</td>
<td>0.042</td>
<td>0.079 (0, 0.46)</td>
</tr>
<tr>
<td>VM</td>
<td>646 (85%)</td>
<td>5.49 (0.002, 83.12)</td>
<td>0.007</td>
<td>0.061</td>
<td>0.049</td>
<td>0.033</td>
<td>0.045 (0, 0.22)</td>
</tr>
<tr>
<td>VBKW</td>
<td>728 (96%)</td>
<td>3.82 (0.016, 35.14)</td>
<td>0.005</td>
<td>0.050</td>
<td>0.039</td>
<td>0.026</td>
<td>0.027 (0, 0.14)</td>
</tr>
</tbody>
</table>

IQR = Interquartile range, RMSE = root-mean-squared error, MAE = median absolute error, IPTW = inverse probability of treatment weights, KW = kernel weights, SD = standard deviation, VM = vector matching, VBKW = vector-based kernel weights

a) One analytic scenario is one combination of the elements from Table 1 and one treatment effect estimate (e.g., one analytic scenario includes n=999, true propensity score includes mild nonlinearity, k = 3, evenly split sample distribution across treatment groups, a homogeneous treatment effect, and an estimate of the ATT of A vs B among observations receiving A)

Table 3. Summary of bias and efficiency of estimates across analytic scenarios that lead to <20% bias

<table>
<thead>
<tr>
<th>Weighting or matching strategy</th>
<th>Number of analytic scenarios with &lt;20% bias</th>
<th>Median (min, max) absolute mean relative bias</th>
<th>Median absolute bias</th>
<th>Median IQR</th>
<th>Median RMSE</th>
<th>Median MAE</th>
<th>Median (min, max) absolute mean standardized differences in prognostic scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPTW</td>
<td>314</td>
<td>7.82 (0.024, 19.80)</td>
<td>0.013</td>
<td>0.059</td>
<td>0.049</td>
<td>0.032</td>
<td>0.042 (0, 0.15)</td>
</tr>
<tr>
<td>KW</td>
<td>365</td>
<td>6.22 (0.019, 19.99)</td>
<td>0.007</td>
<td>0.055</td>
<td>0.044</td>
<td>0.029</td>
<td>0.022 (0, 0.17)</td>
</tr>
<tr>
<td>VM</td>
<td>646</td>
<td>4.40 (0.002, 19.92)</td>
<td>0.006</td>
<td>0.058</td>
<td>0.046</td>
<td>0.031</td>
<td>0.038 (0, 0.14)</td>
</tr>
<tr>
<td>VBKW</td>
<td>728</td>
<td>3.62 (0.016, 19.85)</td>
<td>0.005</td>
<td>0.050</td>
<td>0.038</td>
<td>0.025</td>
<td>0.026 (0, 0.14)</td>
</tr>
</tbody>
</table>

IQR = Interquartile range, RMSE = root-mean-squared error, MAE = median absolute error, IPTW = inverse probability of treatment weights, KW = kernel weights, SD = standard deviation, VM = vector matching, VBKW = vector-based kernel weights

a) One analytic scenario is one combination of the elements from Table 1 and one treatment effect estimate (e.g., one analytic scenario includes n=999, true propensity score includes mild nonlinearity, k = 3, evenly split sample distribution across treatment groups, a homogeneous treatment effect, and an estimate of the ATT of A vs B among observations receiving A)
<table>
<thead>
<tr>
<th>Weighting or matching strategy</th>
<th>Number (%) of analytic scenarios with &lt;20% absolute mean relative bias</th>
<th>Median (min, max) absolute mean relative bias</th>
<th>Median absolute bias</th>
<th>Median IQR</th>
<th>Median RMSE</th>
<th>Median MAE</th>
<th>Median (min, max) absolute mean standardized differences in prognostic scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPTW</td>
<td>17 (35%)</td>
<td>54.17 (3.96, 574.74)</td>
<td>0.056</td>
<td>0.031</td>
<td>0.072</td>
<td>0.050</td>
<td>0.147 (0.03, 1.29)</td>
</tr>
<tr>
<td>KW</td>
<td>8 (17%)</td>
<td>32.47 (8.33, 108.03)</td>
<td>0.046</td>
<td>0.015</td>
<td>0.047</td>
<td>0.046</td>
<td>0.121 (0.03, 0.31)</td>
</tr>
<tr>
<td>VM</td>
<td>42 (88%)</td>
<td>3.88 (0.09, 67.34)</td>
<td>0.006</td>
<td>0.015</td>
<td>0.014</td>
<td>0.010</td>
<td>0.029 (0.01, 0.15)</td>
</tr>
<tr>
<td>VBMW</td>
<td>48 (100%)</td>
<td>2.17 (0.004, 14.25)</td>
<td>0.004</td>
<td>0.011</td>
<td>0.009</td>
<td>0.006</td>
<td>0.020 (0.01, 0.05)</td>
</tr>
</tbody>
</table>

IQR = Interquartile range, RMSE = root-mean-squared error, MAE = median absolute error, IPTW = inverse probability of treatment weights, KW = kernel weights, SD = standard deviation, VM = vector matching, VBKW = vector-based kernel weights

a) One analytic scenario is one combination of the elements from Table 1 and one treatment effect estimate (e.g., one analytic scenario includes n=999, true propensity score includes mild nonlinearity, k = 3, evenly split sample distribution across treatment groups, a homogeneous treatment effect, and an estimate of the ATT of A vs B among observations receiving A)
Figure 1. IPTW estimates were much more likely than other estimates to be biased when the propensity score model was misspecified.

Results shown for a data generating process with true homogenous treatment effects, $E[Y(1) - Y(2)]$, $N=1200$, Equal distribution of sample across 3 treatment groups, and coefficient set 1, 1000 replications.

A through G represent different true propensity score models. A represents a correctly specified, fully saturated, propensity score model.

IPTW = inverse probability of treatment weighting, VBKW = vector-based kernel weighting, KW = kernel weighting, VM = vector matching
Figure 2. The median absolute mean relative bias of estimates produced by VBKW is less sensitive to propensity score misspecification than in IPTW, KW, or VM.

A through G represent different true propensity score models. A represents a correctly specified, fully saturated, propensity score model.

IPTW = inverse probability of treatment weighting, VBKW = vector-based kernel weighting, KW = kernel weighting, VM = vector matching

Results represent all 756 analytic scenarios, with 1000 replications each.
Figure 3. VBKW and VM produced estimates with the lowest median absolute mean relative bias, regardless of the estimand of interest.

IPTW = inverse probability of treatment weighting, VBKW = vector-based kernel weighting, KW = kernel weighting, VM = vector matching

Results represent 1006 scenarios (the 756 described in the text, with the addition of scenarios for transitive ATTs (e.g., ATT A vs B | T = C), with 1000 replications each.
Appendix

Setup and notation:

In a standard cross-sectional setting, we observe a sample of individuals $i = 1, 2, \ldots, N$ from a population.

Our sample size $N$ is the sum of each the treatment group sizes: $N = N_1 + N_2 + \ldots + N_Z$.

Each individual has been assigned one of $z$ possible treatment levels, where $z = 1, 2, \ldots, Z$. We observe the outcome variable $y_i$, the observed treatment level $t_i$, and a $k_x \times 1$ vector covariates, $x_i$. We also define an indicator variable $d_i(z) = 1(t_i=z)$ which is equal to 1 if unit $i$ received treatment $z$ and equal to 0 otherwise. We distinguish between the observed outcome $y_i$ and the $Z$ potential outcomes, $y_i(z)$. The observed outcome is given by

$$y_i = d(1)y_i(1) + d(2)y_i(2) + \ldots + d(Z)y_i(Z)$$

Only one of the $Z$ possible outcomes is observed for each individual in the sample. We observe a propensity score, defined as $p_i(t=z | x_i)$ and a propensity score vector

$$p_i(z, x_i) = \{ p_i(t=1 | x_i), p_i(t=2 | x_i), \ldots, p_i(t=Z | x_i) \},$$

for each unit $i$.

Estimands:

$\forall z \neq z'$,

$$\text{ATE}_{z, z'} = \frac{\sum_{i=1}^{N} y_i d_i(z)w_{i, \text{ATE}}}{\sum_{i=1}^{N} d_i(z)w_{i, \text{ATE}}} - \frac{\sum_{i=1}^{N} y_i d_i(z')w_{i, \text{ATE}}}{\sum_{i=1}^{N} d_i(z')w_{i, \text{ATE}}}$$

$$\text{ATT}_{z, z'} = \frac{\sum_{i=1}^{N} y_i d_i(z)w_{i, \text{ATT}}}{\sum_{i=1}^{N} d_i(z)w_{i, \text{ATT}}} - \frac{\sum_{i=1}^{N} y_i d_i(z')w_{i, \text{ATT}}}{\sum_{i=1}^{N} d_i(z')w_{i, \text{ATT}}}$$

$$\text{ATU}_{z, z'} = \frac{\sum_{i=1}^{N} y_i d_i(z)w_{i, \text{ATU}}}{\sum_{i=1}^{N} d_i(z)w_{i, \text{ATU}}} - \frac{\sum_{i=1}^{N} y_i d_i(z')w_{i, \text{ATU}}}{\sum_{i=1}^{N} d_i(z')w_{i, \text{ATU}}}$$
Weights:

Define $j$ as an index of observations in treatment group $z'$, where $j = \{1, 2, \ldots, N_{z'}\}$.

Define $l$ as an index of observations in treatment group $z$, where $l = \{1, 2, \ldots, N_z\}$.

Inverse Probability of Treatment Weights (IPTW):

$$w_{l,ATE} = \begin{cases} \frac{1}{p_l(t = z | x_i)}, & \forall \ i = l \\ \frac{1}{p_l(t = z' | x_i)}, & \forall \ i = j \end{cases}$$

$$w_{l,ATT} = \begin{cases} 1, & \forall \ i = l \\ \frac{p_l(t = z | x_i)}{p_l(t = z' | x_i)}, & \forall \ i = j \end{cases}$$

$$w_{l,ATU} = \begin{cases} 1, & \forall \ i = j \\ \frac{p_l(t = z' | x_i)}{p_l(t = z | x_i)}, & \forall \ i = l \end{cases}$$

Kernel Weights (KW):

$$w_{l,ATT} = \begin{cases} 1, & \forall \ i = l \\ k_l(D_{lz}), & \forall \ i = j \end{cases}$$

$$k_l(D_{lz}) = \begin{cases} \frac{3}{4} \left( 1 - \left( \frac{D_{lz}}{h} \right)^2 \right), & \text{if } D_{lz} < h \\ 0, & \text{otherwise} \end{cases}$$

$$D_{lz} = | p_l(t = z | x_i) - p_l(t = z | x_i) |$$

$$w_{l,ATU} = \begin{cases} 1, & \forall \ i = j \\ k_l(D_{jz'}), & \forall \ i = l \end{cases}$$

$$k_l(D_{jz'}) = \begin{cases} \frac{3}{4} \left( 1 - \left( \frac{D_{jz'}}{h} \right)^2 \right), & \text{if } D_{jz'} < h \\ 0, & \text{otherwise} \end{cases}$$

$$D_{jz'} = | p_l(t = z' | x_i) - p_l(t = z' | x_i) |$$

$$w_{l,ATE} = w_{l,ATT} + w_{l,ATU}$$
Vector-Based Kernel Weighting (VBKW):

\[ w_{i,\text{ATT}} = \begin{cases} 1, & \forall \ i = l \\ k_i(D_{lz}), & \forall \ i = j \end{cases} \]

\[ k_i(D_{lz}) = \begin{cases} \frac{3}{4} \left(1 - \left(\frac{D_{lz}}{h}\right)^2\right), & \text{if } D_{lz} < h \text{ and } D_{lm} \\ 0, & \text{otherwise} \end{cases} \]

\[ D_{lz} = | p_i(t = z | x_i) - p_j(t = z | x_i) | \]

\[ D_{lm} = | p_i(t = m | x_i) - p_i(t = m | x_i) | \quad \forall \ m \neq z \]

\[ w_{i,\text{ATU}} = \begin{cases} 1, & \forall \ i = j \\ k_i(D_{jz}), & \forall \ i = l \end{cases} \]

\[ k_i(D_{jz}) = \begin{cases} \frac{3}{4} \left(1 - \left(\frac{D_{jz}}{h}\right)^2\right), & \text{if } D_{jz} < h \text{ and } D_{jn} < h \\ 0, & \text{otherwise} \end{cases} \]

\[ D_{jz} = | p_i(t = z' | x_i) - p_j(t = z' | x_i) | \]

\[ D_{jn} = | p_i(t = n | x_i) - p_j(t = n | x_i) | \quad \forall \ n \neq z' \]

\[ w_{i,\text{ATE}} = w_{i,\text{ATT}} + w_{i,\text{ATU}} \]

Vector Matching (VM):

\[ w_{i,\text{ATT}} = n_{i,\text{matched}}, \text{where } n_{i,\text{matched}} \text{ is the number of times subject } i \text{ is part of a matched set of observations composed of at least one individual from each treatment group, when matching is implemented using reference group } z. \]

\[ w_{i,\text{ATU}} = n_{i,\text{matched}}, \text{where } n_{i,\text{matched}} \text{ is the number of times subject } i \text{ is part of a matched set of observations composed of at least one individual from each treatment group when matching is implemented using reference group } z'. \]

\[ w_{i,\text{ATE}} = w_{i,\text{ATT}} + w_{i,\text{ATU}} \]