Applied Microeconometrics

Selection on Observables: Part 5

Derya Uysal

IHS, Vienna

MSc Course IHS/TU Wien

Spring 2014

General Idea

- Find a “match” - or small number of “matches”- for each individual from the opposite group.
- A match will be determined in terms of the similarities of the observable characteristics.
- The matching estimators impute the missing potential outcome by using average outcomes for individuals with “similar” values for the covariates.

Simple Matching Estimator (Abadie and Imbens (2002))

- Let $\ell_m(i)$ be the index ℓ that satisfies $D_\ell \neq D_i$ and

$$\sum_{j|D_j \neq D_i} \mathbb{1}\{\| X_j - X_i \| \leq \| X_\ell - X_i \|\} = m,$$

- In other words, $\ell_m(i)$ is the index of the unit in the opposite treatment group that is the $m-$th closest to unit i in terms of the distance measure based on the norm $\|\|$.
- $\ell_1(i)$ is the nearest match for unit i.

Simple Matching Estimator: Distance Metric

- The usual distance metric is Standard Euclidian metric:

$$\| X_j - X_i \| = \sqrt{(X_j - X_i)'(X_j - X_i)} = \sqrt{\sum_{k=1}^{K}(X_{kj} - X_{ki})^2}$$

  where K is number of covariates.
- The Euclidean distance is not invariant to changes in the scale of the X's.
- There are alternative distances that are invariant to changes in scale.
  - Mahalanobis distance:
    $\| X_j - X_i \| = \sqrt{(X_j - X_i)'\Sigma_X^{-1}(X_j - X_i)}$, where Σ_X is covariance matrix of the covariates
  - Normalized Euclidean distance:
    $\| X_j - X_i \| = \sqrt{(X_j - X_i)'\text{diag}(\Sigma_X^{-1})(X_j - X_i)}$

Simple Matching Estimator

- Let $\mathcal{J}_M(i)$ denote the set of indices for the first M matches for unit i: $\mathcal{J}_M(i) = \{\ell_1(i), \ldots, \ell_M(i)\}$
- Define the imputed potential outcomes as:

$$\hat{Y}_{0i} = \begin{cases} Y_i & \text{if } D_i = 0 \\ \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } D_i = 1 \end{cases} \quad \text{and} \quad \hat{Y}_{1i} = \begin{cases} \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } D_i = 0 \\ Y_i & \text{if } D_i = 1 \end{cases}$$

- The simple matching estimator is then

$$\hat{\Delta}_M^{sm} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}_{1i} - \hat{Y}_{0i}\right)$$

Simple Matching Estimator

Table 1: A matching estimator with 7 observations, $m = 1$ (matching with replacement)

i	D_i	X_i	Y_i	$\mathcal{J}_1(i)$	$\hat{Y}_{0i}$	$\hat{Y}_{1i}$
1	0	2	7	$\{5\}$	7	**8**
2	0	4	8	$\{4, 6\}$	8	**7.5**
3	0	5	6	$\{4, 6\}$	6	**7.5**
4	1	3	9	$\{1, 2\}$	**7.5**	9
5	1	2	8	$\{1\}$	**7**	8
6	1	3	6	$\{1, 2\}$	**7.5**	6
7	1	1	5	$\{1\}$	**7**	5

Simple Matching Estimator

Asymptotic Properties of the M-Nearest Neighbor Matching estimator for fixed K (see Abadie and Imbens (2006)):

- For one continuously distributed conditioning variable, M-NN Matching is $\sqrt{n}$-consistent
- M-NN is in general not $\sqrt{n}$-consistent if the number of continuous conditioning variables is large.
- The M-NN estimator does not achieve the semiparametric efficiency bound
- The bootstrap for calculating the standard errors of the matching estimator is not valid
- Abadie and Imbens (2006) provide an estimator for the large sample variance of the M-NN Matching estimator

Matching on Propensity Score

- Since conditioning on all relevant covariates is limited in the case of a high dimensional vector X ('curse of dimensionality'), Rosenbaum and Rubin (1983b) suggest the use of so-called balancing scores $b(X)$
- One possible balancing score is the propensity score
- Instead of determining the closeness in terms of covariates we can measure the closeness in term of estimated propensity score

Matching on Propensity Score

- Since conditioning on all relevant covariates is limited in the case of a high dimensional vector X ('curse of dimensionality'), Rosenbaum and Rubin (1983b) suggest the use of so-called balancing scores $b(X)$
- One possible balancing score is the propensity score
- Instead of determining the closeness in terms of covariates we can measure the closeness in term of estimated propensity score
- Matching on the propensity score is essentially a weighting scheme, which determines what weights are placed on comparison units

Matching on Propensity Score

Dehejia and Wahba states three issues arising when implementing matching:

- whether or not to match with replacement,
- how many comparison units to match to each treated unit
- which matching method to choose

Matching on Propensity Score

- Matching with replacement minimizes the propensity score distance between the matched comparison units and the treatment unit: Bias reduction
- matching without replacement, when there are few comparison units similar to the treated units, we may be forced to match treated units to comparison units that are quite different in terms of the estimated propensity score. This increases bias, but it could improve the precision of the estimates.
- An additional complication of matching without replacement is that the results are potentially sensitive to the order in which the treatment units are matched.

Matching on Propensity Score

- By using more comparison units, one increases the precision of the estimates, but at the cost of increased bias.
- $m-$nearest-neighbor method or caliper matching (comparison units within a predefined propensity score radius)

Matching on Propensity Score

- Let $P = \Pr(D = 1|X)$.
- A typical propensity score matching estimator for ATT takes the form

$$\hat{\Delta}_{TT} = \frac{1}{N_1} \sum_{i \in I_1 \cap S_p} \left\{ Y_{1i} - \hat{\mathrm{E}}[Y_{0i}|D_i = 1, P_i] \right\},$$

where

$$\hat{\mathrm{E}}[Y_{0i}|D_i = 1, P_i] = \sum_{j \in I_0} W(i,j) Y_{0j}.$$

and where

  - I_1: set of program participants
  - I_0: set of non-participants
  - S_p: region of common support.
- The weights $W(i,j)$ depend on the distance between P_i and P_j.

M-Nearest Neighbor Matching

Let $\ell_k(i)$ be the index $l \in I_0$ that satisfies

$$\sum_{j \in I_0} \mathbb{1}\{\| P_l - P_i \| \leq \| P_j - P_i \|\} = m.$$

$\ell_m(i)$ is the index of the unit in the control group that is m^{th} closest to unit i in terms of the propensity score.
Let $J_M(i)$ denote the set of indices for the first K matches for unit i:

$$\mathcal{J}_M(i) = \{\ell_1(i), ..., \ell_M(i)\}$$

The M-Nearest Neighbor Matching estimator (with replacement) is

$$\hat{\Delta}_{TT}^{NN} = \frac{1}{N_1} \sum_{i \in I_1} \left\{ Y_{1i} - \frac{1}{K} \sum_{j \in \mathcal{J}_K(i)} Y_{0j} \right\}.$$

M-Nearest Neighbor Matching

Alternatively, the K-Nearest Neighbor Matching estimator (with replacement) can be written as

$$\hat{\Delta}_{TT}^{NN} = \frac{1}{N_1} \sum_{i \in I_1 \cap S_p} \left\{ Y_{1i} - \sum_{j \in I_0} \frac{\mathbb{1}\{j \in J_K(i)\}}{K} Y_{0j} \right\}.$$

$\Rightarrow W_{NN}(i,j) = \frac{\mathbb{1}\{j \in J_K(i)\}}{K}$ with $\sum_{j \in I_0} W_{NN}(i,j) = 1$.

M-Nearest Neighbor Matching

- Abadie and Imbens (2008) show that the bootstrap is not in general valid for matching estimators.
- Abadie and Imbens (2012) derive the large sample asymptotics for propensity score matching estimator of ATE and ATT with estimated propensity score
- They show that matching on the estimated propensity score is more efficient than matching on the true propensity score in large samples.
- The ATET estimator the sign of the adjustment term depends on the data generating process
- ignoring the estimation error in the propensity score may lead to confidence intervals that are either too large or too small.

Stratification or Interval Matching

Method:

- In this variant of matching, the common support of P is partitioned into a set of intervals.
- Average treatment impacts are calculated through simple averaging within each interval.
- Overall average impact estimate:
  - a weighted average of the interval impact estimates, using the fraction of the $D = 1$ population in each interval for the weights.
- Requires decision on how wide the intervals should be:
  - Dehejia and Wahba (1999) use intervals that are selected such that the mean values of the estimated P_i's and P_j's are not statistically different from each other within intervals.

Kernel Matching

The Kernel Matching estimator is defined as

$$\hat{\Delta}_{TT}^{KM} = \frac{1}{N_1} \sum_{i \in I_1 \cap S_p} \left\{ Y_{1i} - \sum_{j \in I_0} \frac{K(\frac{P_j - P_i}{h})}{\sum_{j=1}^{N_0} K(\frac{P_j - P_i}{h})} Y_{0j} \right\},$$

where $K(.)$ is a "Kernel Function" and h the bandwidth.

$$\Rightarrow W_{KM}(i,j) = \frac{K(\frac{P_j - P_i}{h})}{\sum_{j=1}^{N_0} K(\frac{P_j - P_i}{h})} \text{ with } \sum_{j \in I_0} W_{KM}(i,j) = 1.$$

Kernel Matching

- Uses a weighted average of all observations within the common support region: the farther away the comparison unit is from the treated unit the lower the weight.
- The Kernel Matching estimator is formally defined as

$$\hat{\Delta}_{TT}^{KM} = \frac{1}{N_1} \sum_{i \in I_1 \cap S_p} \left\{ Y_{1i} - \sum_{j \in I_0} \frac{K(\frac{P_j - P_i}{h})}{\sum_{j=1}^{N_0} K(\frac{P_j - P_i}{h})} Y_{0j} \right\},$$

where $K(.)$ is a "Kernel Function" and h the bandwidth.

$$\Rightarrow W_{KM}(i,j) = \frac{K(\frac{P_j - P_i}{h})}{\sum_{j=1}^{N_0} K(\frac{P_j - P_i}{h})} \text{ with } \sum_{j \in I_0} W_{KM}(i,j) = 1.$$

Kernel Matching

Table 2: Examples for Kernel Functions

Kernel	Kernel Function $K(z)$
Uniform	$\frac{1}{2}\mathbb{1}\{\lvert z\rvert < 1\}$
Epanechnikov	$\frac{3}{4}(1-z^2)\mathbb{1}\{\lvert z\rvert < 1\}$
Quartic	$\frac{15}{16}(1-z^2)^2\mathbb{1}\{\lvert z\rvert < 1\}$
Gaussian	$(2\pi)^{-\frac{1}{2}}\exp(-\frac{z^2}{2})$

Kernel functions are

- positive and integrate to unity over the band.
- are symmetric around zero, so that points below x get the same weight as those an equal distance above.
- are decreasing in the absolute value of its argument.

Local Linear Matching

- Similar to the kernel estimator but includes a linear term in the weighting function, which helps to avoid bias.
- A generalized version of Kernel Matching is Local Linear Matching:

$$\hat{\Delta}_{TT}^{LLM} = \frac{1}{N_1} \sum_{i \in I_1 \cap S_p} \left\{ Y_{1i} - \sum_{j \in I_0} W_{LLM}(i,j) Y_{0j} \right\},$$

where

$$W_{LLM}(j,i) = \frac{K_{ij} \sum_{k=1}^{N_0} K_{ik}(P_k - P_i)^2 - [K_{ij}(P_j - P_i)][\sum_{k=1}^{N_0} K_{ik}(P_k - P_i)]}{\sum_{j=1}^{N_0} K_{ij} \sum_{k=1}^{N_0} K_{ik}(P_k - P_i)^2 - \left(\sum_{k=1}^{N_0} K_{ik}(P_k - P_i)\right)}.$$

and where $K_{ij} \equiv K(\frac{P_j - P_i}{h})$. Again, $\sum_{j \in I_0} W_{LLM}(i,j) = 1$.

Local Linear Matching

Local regression interpretation of the Kernel and Local Linear Matching Estimator:
Let

$$\hat{\alpha}, \hat{\beta} = \arg\min_{\alpha,\beta} \sum_{j \in I_0} (Y_{0j} - \alpha - \beta(P_j - P_i))^2 K(\frac{P_j - P_i}{h}).$$

Then $\hat{\alpha} = \sum_{j \in I_0} W_{LLM}(i,j) Y_{0j}$.
If the locally weighted regression model includes only a constant then

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{j \in I_0} (Y_{0j} - \alpha)^2 K(\frac{P_j - P_i}{h}).$$

and $\hat{\alpha} = \sum_{j \in I_0} W_{KM}(i,j) Y_{0j}$.

Local Linear Matching

Asymptotic properties of KM and LLM (see Heckman, Ichimura and Todd (1997)):

- KM and LLM are $\sqrt{n}$ consistent and asymptotically normal when matching is with respect to X, the known propensity score or when the propensity score is estimated.
- KM- and LLM- estimators based on X or the known propensity score achieve the semiparametric efficieny bound.
- The propensity score matching estimator does not necessarily improve upon the variance of the matching estimator with respect to X
- Using the estimated propensity score increases the asymptotic variance due to the additional variability of the propensity score estimation step

Practical Guide to PSM
Caliendo and Kopeinig, Journal of Economic Surveys (2008)

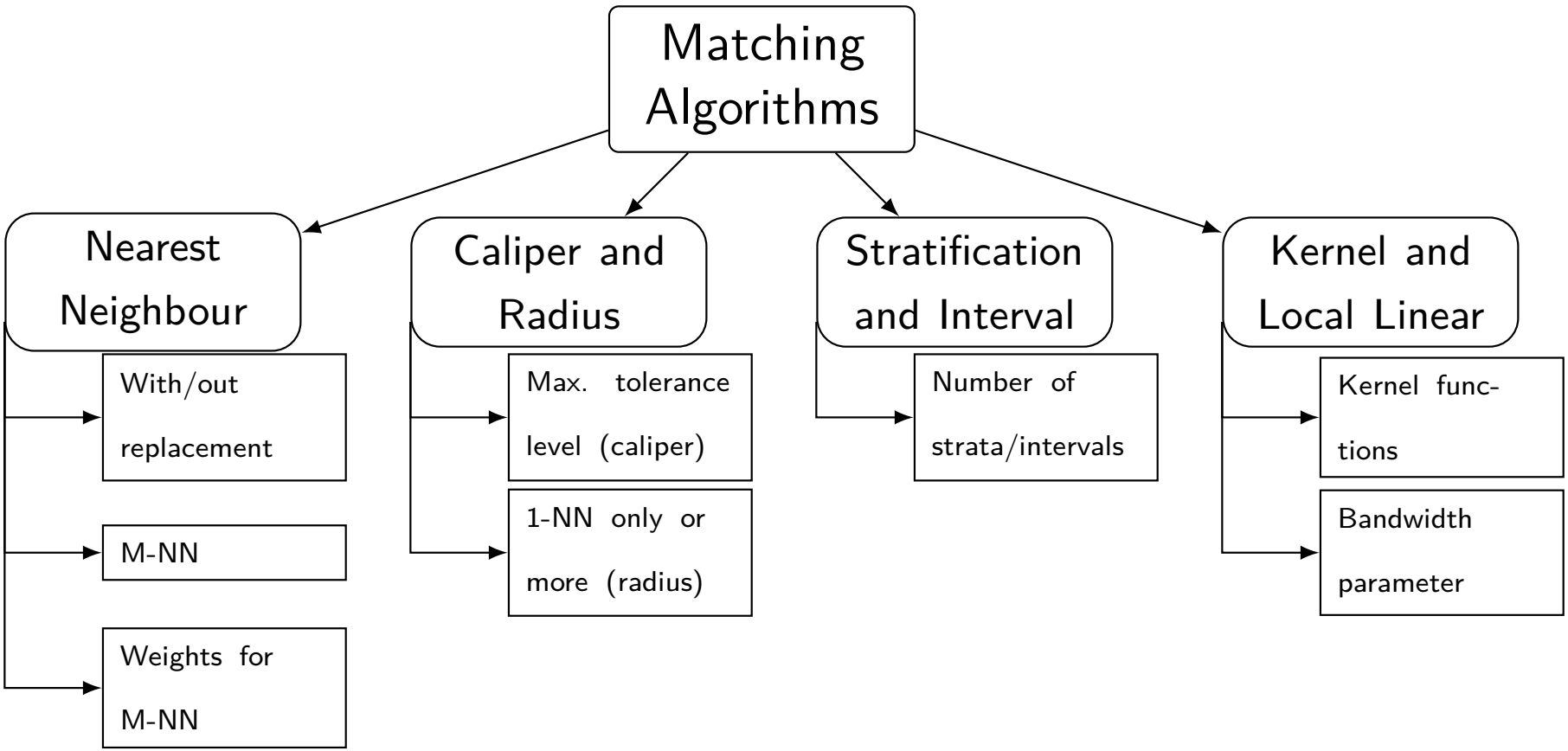
1. Propensity Score Estimation
2. Choose Matching Algorithm
3. Check Overlap/ Common Support
4. Matching Quality/Effect Estimation
5. Sensitivity Analysis

1. Propensity Score Estimation

Two choices have to be made:

- the model to be used for the estimation
  - For the binary treatment case, logit and probit models usually yield similar results.
  - Avoid LPM due to known reasons
- the variables to be included in this model.
  - Variables should not be influenced by participation (or anticipation) and must satisfy CIA
    - → Economic issues Choose variables by economic theory and previous empirical evidence
    - → Statistical issues 'Hit or miss' method, stepwise augmentation, leave-one-out cross-validation
    - → Key variables 'Overweighting' by matching on subpopulations or insisting on perfect match

2. Choosing a Matching Algorithm



Nearest Neighbour Matching

NN matching 'with replacement' and 'without replacement'.

- Matching with replacement involves a trade-off between bias and variance.
-
  - If we allow replacement, the average quality of matching will increase and the bias will decrease.
  - replacement reduces the number of distinct nonparticipants used to construct the counterfactual outcome and thereby increases the variance of the estimator (Smith and Todd, 2005).
  - A problem which is related to NN matching without replacement is that estimates depend on the order in which observations get matched.

Nearest Neighbour Matching

- M-NN: This form of matching involves a trade-off between variance and bias, too.
- It trades reduced variance, resulting from using more information to construct the counterfactual for each participant, with increased bias that results from on average poorer matches (see e.g. Smith, 1997).
- When using oversampling, one has to decide how many matching partners should be chosen for each treated individual and which weight (e.g. uniform or triangular weight) should be assigned to them.

Caliper and Radius Matching

- M-NN matching faces the risk of bad matches if the closest neighbour is far away.
- This can be avoided by imposing a tolerance level on the maximum propensity score distance (caliper).
- Bad matches are avoided and the matching quality rises.
- However, if fewer matches can be performed, the variance of the estimates increases.
- As Smith and Todd (2005) note, a possible drawback of caliper matching is that it is difficult to know a priori what choice for the tolerance level is reasonable.
- Dehejia and Wahba (2002) suggest a variant of caliper matching which is called radius matching.
- The basic idea of this variant is to use not only the NN within each caliper but all of the comparison members within the caliper.

Stratification and Interval Matching

- The idea of stratification matching is to partition the common support of the propensity score into a set of intervals (strata) and to calculate the impact within each interval by taking the mean difference in outcomes between treated and control observations.
- This method is also known as interval matching, blocking and subclassification (Rosenbaum and Rubin, 1984).
- How many strata should be used in empirical analysis.
- Cochran (1968) shows that five subclasses
- One way to justify the choice of the number of strata is to check the balance of the propensity score (or the covariates) within each stratum (see e.g. Aakvik, 2001).

Kernel and Local Linear Matching

- Kernel matching (KM) and local linear matching (LLM) are nonparametric matching estimators that use weighted averages of (nearly) all -depending on the choice of the kernel function- individuals in the control group to construct the counterfactual outcome.
- Advantage: lower variance, Disadvantage: higher bias
- KM can be seen as a weighted regression of the counterfactual outcome on an intercept with weights given by the kernel weights.
- The estimated intercept provides an estimate of the counterfactual mean.
- KM and LLM is that the latter includes in addition to the intercept a linear term in the propensity score of a treated individual.
- This is an advantage whenever comparison group observations are distributed asymmetrically around the treated observation, e.g. at boundary points, or when there are gaps in the propensity score distribution.

Kernel and Local Linear Matching

- When applying KM one has to choose the kernel function and the bandwidth parameter.
- Kernel function does not matter much
- bandwidth leads to bias-var trade off
- High bandwidth: lower variance but higher bias intercept provides an estimate of the counterfactual mean.
- KM and LLM is that the latter includes in addition to the intercept a linear term in the propensity score of a treated individual.
- This is an advantage whenever comparison group observations are distributed asymmetrically around the treated observation, e.g. at boundary points, or when there are gaps in the propensity score distribution.

Trade-offs in Terms of Bias and Efficiency

Table 3: Trade-offs in Terms of Bias and Efficiency.

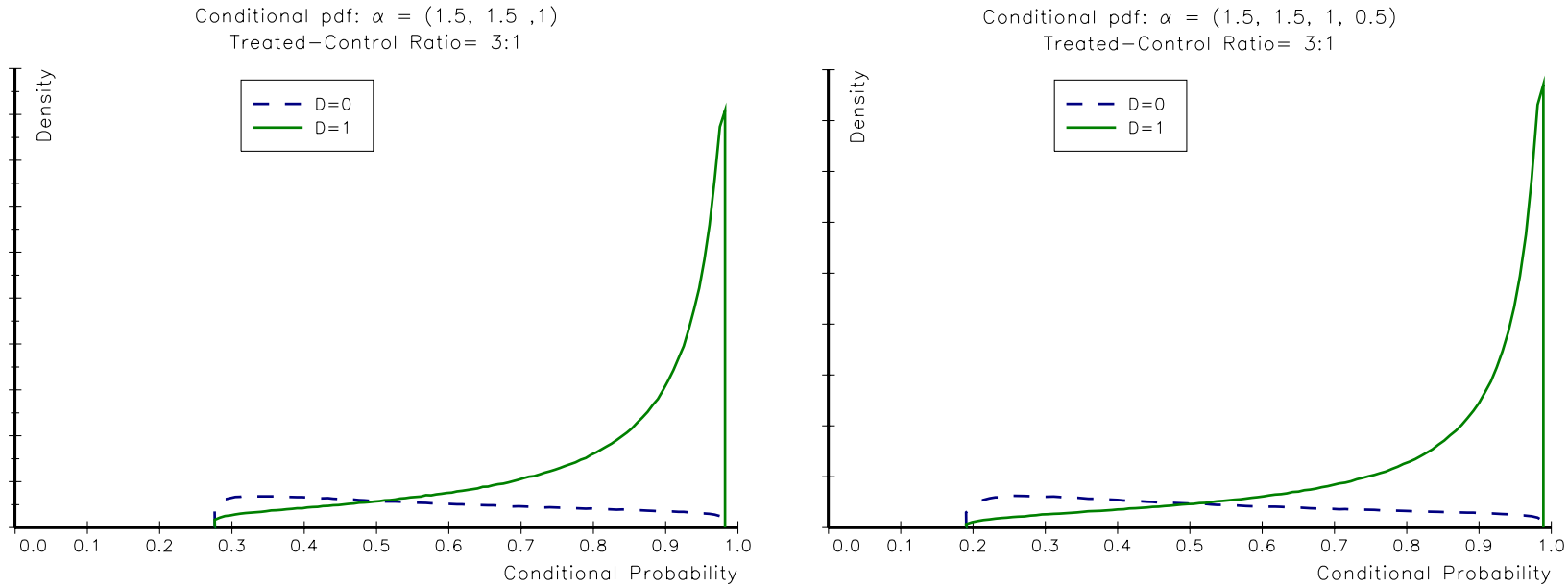
Decision	Bias	Variance
Nearest neighbour matching:		
multiple neighbours/single neighbour	$(+)/(-)$	$(-)/(+)$
with caliper/without caliper	$(-)/(+)$	$(+)/(-)$
Use of control individuals:		
with replacement/without replacement	$(-)/(+)$	$(+)/(-)$
Choosing method:		
NN matching/Radius matching	$(-)/(+)$	$(+)/(-)$
KM or LLM/NN methods	$(+)/(-)$	$(-)/(+)$
Bandwidth choice with KM:		
small/large	$(-)/(+)$	$(+)/(-)$

KM, kernel matching, LLM; local linear matching; NN, nearest neighbour; increase; $(+)$; decrease $(-)$.

Overlap and Common Support

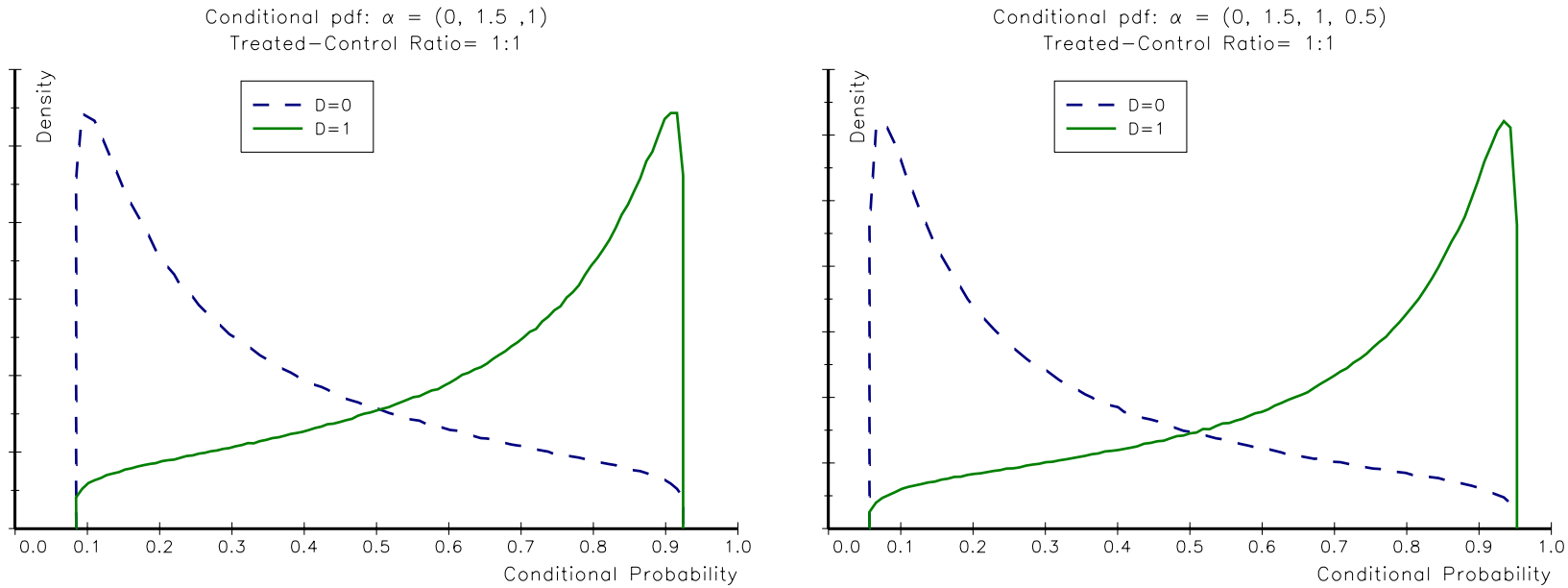
- Visual analysis of propensity score distributions
- Trimming as a way of sustaining common support:
  1. The first trimming rule goes back to a suggestion by Dehejia & Wahba (199). Let $T_i^{ATE} = \mathbb{1}(\hat{a} < \hat{p}_i < \hat{b})$ setting $\hat{b}$ to be the m^{th} largest propensity score in the control group and $\hat{a}$ to be the m^{th} smallest propensity score in the treatment group. Then the estimators are computed based on the subsample for which $T_i^{ATE} = 1$. Usually $m = 1$.
  2. In the second trimming rule suggested by Crump et al. (2009), all units with an estimated propensity score outside the interval $[0.1; 0.9]$ for the ATE are discarded.
  3. The third trimming method suggested by Imbens (2004) is setting an upper bound on the relative weight of each unit. We restrict the maximum relative weight by 4% as in Huber et al. (2013).

Figure 1: Overlap Plots with uniformly distributed X's



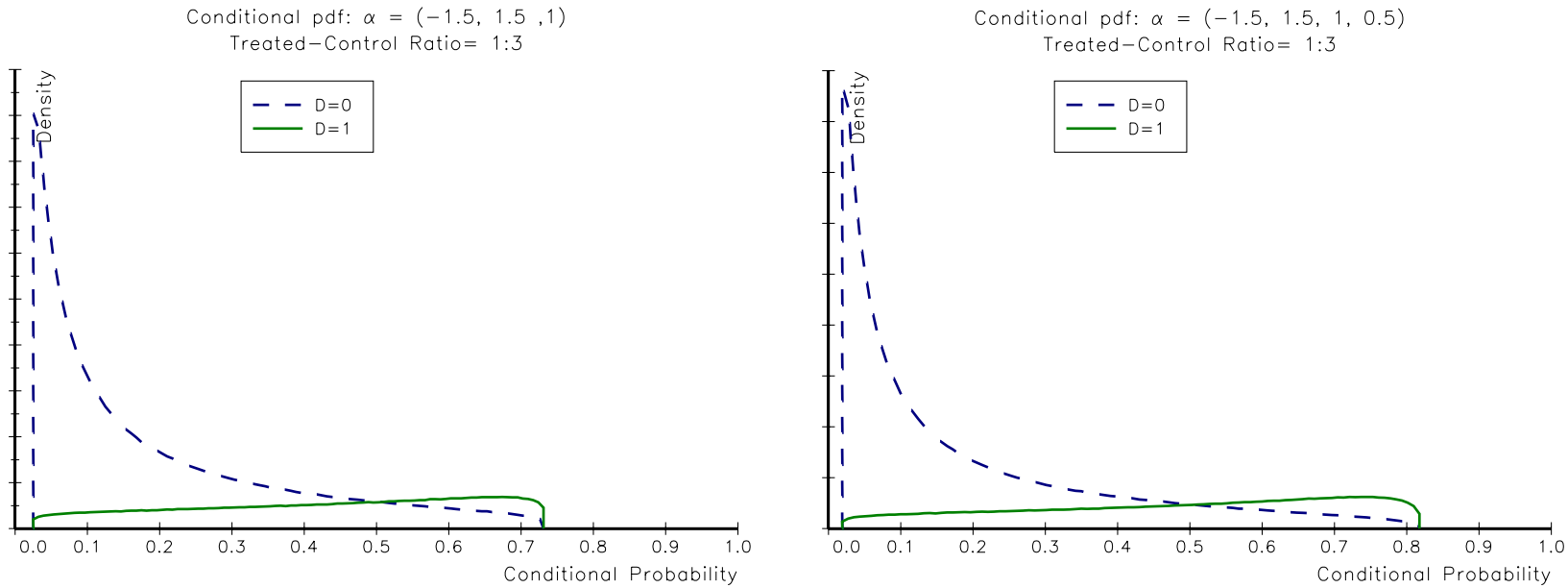
The graphs display estimated densities of conditional probabilities for treated (D=1, solid line) and control (D=0, dashed line) groups where Xs are drawn from a uniform distribution. Treated to control ratio: 3 to 1

Figure 2: Overlap Plots with uniformly distributed X's



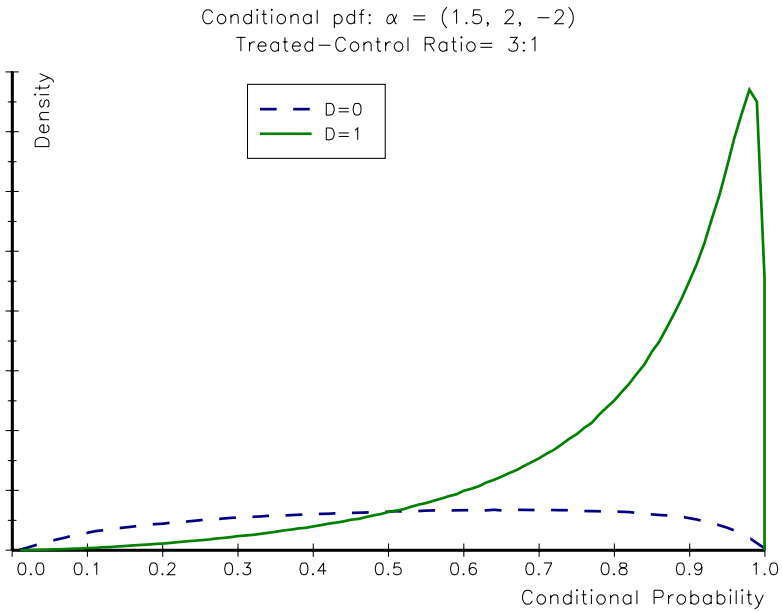
The graphs display estimated densities of conditional probabilities for treated (D=1, solid line) and control (D=0, dashed line) groups where Xs are drawn from a uniform distribution. Treated to control ratio: 1 to 1

Figure 3: Overlap Plots with uniformly distributed X's



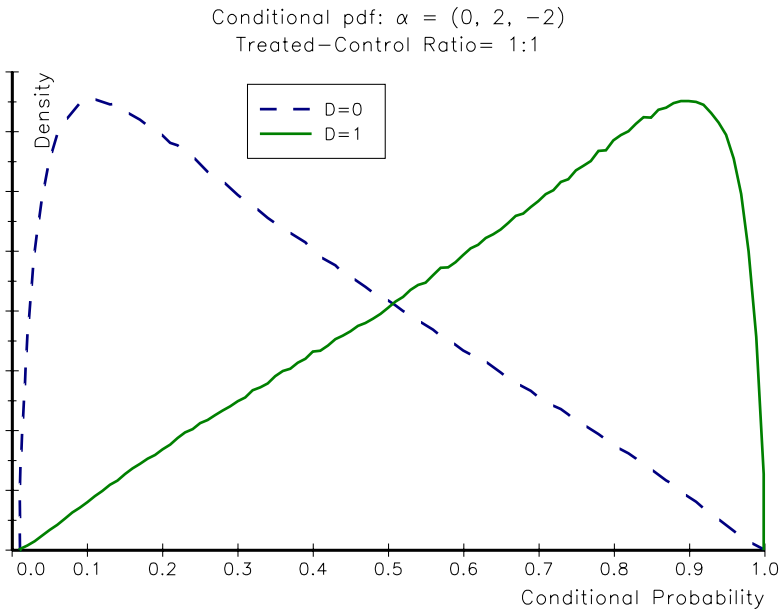
The graphs display estimated densities of conditional probabilities for treated (D=1, solid line) and control (D=0, dashed line) groups where Xs are drawn from a uniform distribution. Treated to control ratio: 1 to 3

Figure 4: Overlap Plots with normally distributed X's



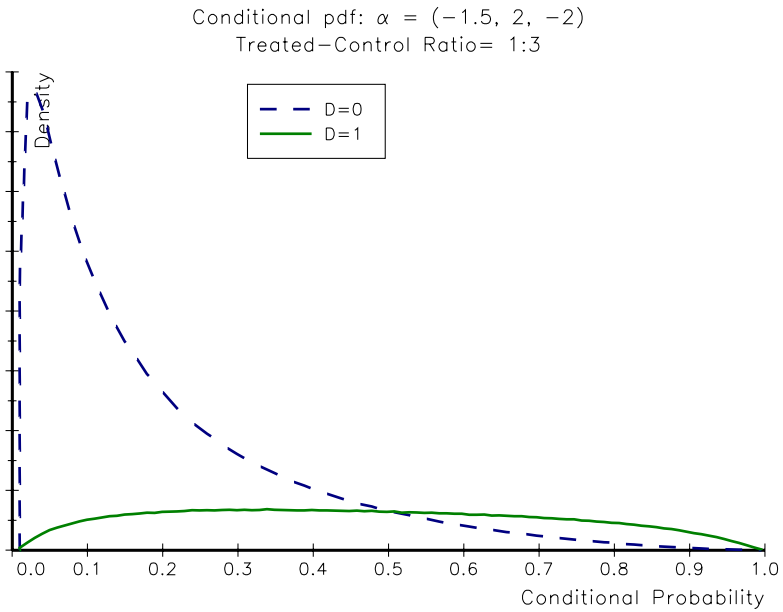
The graphs display estimated densities of conditional probabilities for treated (D=1, solid line) and control (D=0, dashed line) groups where X's are drawn from a normal distribution.

Figure 5: Overlap Plots with normally distributed X's



The graphs display estimated densities of conditional probabilities for treated (D=1, solid line) and control (D=0, dashed line) groups where X's are drawn from a normal distribution.

Figure 6: Overlap Plots with normally distributed X's



The graphs display estimated densities of conditional probabilities for treated (D=1, solid line) and control (D=0, dashed line) groups where X's are drawn from a normal distribution.

Assessing the Matching Quality

- Standardized Bias: For each covariate X it is defined as the difference of sample means in the treated and matched control subsamples as a percentage of the square root of the average of sample variances in both groups.
- t-Test
- Joint Significance and Pseudo-R^2

Sensitivity Analysis

- Rosenbaum (2002) and Aakvik (2001): The purpose of the sensitivity analysis is to ask whether inferences about causal effects may be altered by unobserved factors.
- Ichino et al. (2006): They derive point estimates of the ATT under different possible scenarios of deviation from unconfoundedness.