Causal Inference and Missing Values: Multiple treatment or Continuous treatment

Ayoub Tabaai, Benjamin Cohen, Oumaima BOUTHER, Jean Pachebat, Hamza Oulhaj

April 22, 2022

Abstract

Several methods to estimate the average effect (IPW, g-estimators, AIPW) and heterogeneous effects in the context of binary treatment already exist. The objective of this project is to make a state of the art of the available methods to manage multiple treatments or even continuous treatments (doses). On the one hand we will try to list the main estimators and their characteristics. On the other hand we will list their implementations to compare them empirically by reproducing simulations of a real dataset. It will help us to create recommendations to the users. We will prove some results empirically using notebook written in Python and also try to suggest improvements if needed.

1 Introduction, notations and reminders on Binary Treatment

For each subject i $\in$ (1,...,N), (Y_i, X_i, T_i) will denote respectively the observed outcome, set of covariates and binary treatment assignment. N is less than the total cardinal of people in the superpopulation. We denote τ the treatment space. For a binary treatment, τ = t_1,t_2.
(n_{t_1}, n_{t_2}) will denote the number of subjects receiving treatments t_1 and t_2 respectively.

The Rubin Causal Model (RCM) needs Stable Unit Treatment Value Assumption (SUTVA) to define the potential outcomes ($Y_i(t_1)$, $Y_i(t_2)$). SUTVA implements no interference between subjects and no hidden treatment versions, representing the non-variation of the set of potential outcomes for a subject with the treatment assignment of the other subjects. Every subject get a unique treatment at a specific point in time. This means for us that we observe either $Y_i(t_1)$ or $Y_i(t_2)$ for each subject.

Imputation or IPW on the propensity score can reduce the initial covariate bias between the treatment and control groups. Several estimands of superpopulation effects exist as the population average treatment effect $PATE_{t1,t2}$ and the population average treatment effect among those receiving t_1 $PATT_{t1,t2}$:

1. $PATE_{t1,t2}$ = E($Y_i(t_1)$ - $Y_i(t_2)$)
2. $PATT_{t1,t2}$ = E($Y_i(t_1)$ - $Y_i(t_2)|T_i = t_1$)

$I_{(T_i=t_1)}$ will denote the indicator function for an individual receiving treatment t_1.
$PATE_{t_1,t_2}$ and $PATT_{t_1,t_2}$ are generally approximated by the sample average treatment effects:

1. $SATE_{t_1,t_2} = \frac{1}{N} \sum_{i=1}^{N}(Y_i(t_1)$ - $Y_i(t_2))$
2. $SATT_{t_1,t_2} = \frac{1}{n_{t_1}} \sum_{i=1}^{N}(Y_i(t_1)$ - $Y_i(t_2)) \times I_{(T_i=t_1)}$

2 Multiple Treatments

2.1 Notation

The number of possible estimands grows with increasing treatment options. We denote here τ = t_1,t_2,...,t_Z the treatment support for Z treatments, with Y_i = $Y_i(t_1)$, $Y_i(t_2)$,...,$Y_i(t_Z)$ the set of potential outcomes for subject i.

In the beginning, the SUTVA expands across a subject's vector of potential outcomes. To apply regular treatments assignment we need unconfoundedness and other probabilistic properties for multiple exposures as in binary treatment, sampling N people from an infinite super-population results in an individualistic assignment mechanism.
Definition: Assignment mechanisms are super-population probabilistic if:

$$\forall(X_i, Y_i(t_1), ..., Y_i(t_Z)) \; and \; \forall t \in t_1, ..., t_Z, \; 0 < f_{T|Y(t_1),...,Y(t_Z),X}(t|Y_i(t_1), ..., Y_i(t_Z), X_i, \phi) < 1.$$

It means that for multiple treatments, super-population unconfounded assignment mechanism needs:

$$\forall(y_{t_1}, ..., y_{t_Z}, x, \phi) \; and \; t \in t_1, ..., t_Z, f_{T|y(t_1),...,y(t_Z),X}(t|Y_i(t_1), ..., Y_i(t_Z), x, \phi) = f_{T|X}(t|x, \phi)$$

2.2 Estimands of superpolutation effects

(w_1, w_2): two subgroups of treatments such that (w_1, w_2) $\subseteq \tau^2$ and $w_1 \cap w_2 = \emptyset$.
($|w_1|$, $|w_2|$): cardinalities of w_1 and w_2.
We define new estimands for multiple treatments as PATEw_1, w_2 and PATT$w_1|w_1, w_2$:

$$PATE_{w_1,w_2} = E\left[\frac{\sum_{t\in w_1} Y_i(t)}{|w_1|} - \frac{\sum_{t\in w_2} Y_i(t)}{|w_2|}\right]; \; PATT_{w_1|w_1,w_2} = E\left[\frac{\sum_{t\in w_1} Y_i(t)}{|w_1|} - \frac{\sum_{t\in w_2} Y_i(t)}{|w_2|}|T_i \in w_1\right]$$

Note that we have : PATE_{w_1,w_3} - PATE_{w_1,w_2} = PATE_{w_2,w_3}
Those estimands for multiple treatments compare all treatments using simultaneous pairwise comparisons. For $w_1 = t_1$ the reference group, we study Z - 1 pairwise PATT's, one for each of the treatments that the reference group did not receive. To compare the Z - 1 treatments, the PATT's must be assumed transitive: $\text{PATT}_{w_1|w_1,w_3}$ - $\text{PATT}_{w_1|w_1,w_2}$ = $\text{PATT}_{w_1|w_2,w_3}$. It is not true in general unless the super populations of the people affected to treatments w_1 and w_2 are the same.

2.3 IPW for multiple treatments

To estimate the PATE and PATT with IPW for multiple treatments, we can use a relaxed version of the assumption of a regular treatment assignment: weak unconfoundedness.
IPW needs that $\forall t \in \tau$, $\text{P}(\text{I}_i(t) = 1|\text{Y}_i(t),\text{X}_i) = \text{P}(\text{I}_i(t) = 1|\text{X}_i)$ to estimate PATE_{t_1,t_2} and PATT_{t_1,t_2}.
We then get the following estimands:

$$PATE_{t_1,t_2} = E[Y_i\hat{(}t_1)] - E[Y_i\hat{(}t_2)]$$

with

$$E[Y_i\hat{(}t_1)] = (\sum_{i=1}^{N} \frac{Y_i I(T_i = t_1)}{r(t_1, X_i)}) \times (\sum_{i=1}^{N} \frac{I(T_i = t_1)}{r(t_1, X_i)})^{-1}$$

and

$$E[Y_i\hat{(}t_2)] = (\sum_{i=1}^{N} \frac{Y_i I(T_i = t_2)}{r(t_2, X_i)}) \times (\sum_{i=1}^{N} \frac{I(T_i = t_2)}{r(t_2, X_i)})^{-1}$$

The weights almost equal to 0 leads to estimates with large variances. One way to solve this problem in binary treatment is to prune subjects with extreme weights. It reduces the variance but increases the bias. Doubly robust methods, covariate balancing propensity score, generalized boosted models are other ways to work against those extreme weights.

2.4 CRM model: Common referent matching

We work here with only 3 treatments, $\tau = (\text{t}_1, \text{t}_2, \text{t}_3)$ with $\text{n}_{t_1} = \min(\text{n}_{t_1}, \text{n}_{t_2}, \text{n}_{t_3})$, with t_1 the reference group.

1. $\forall$ subject receiving (t_1, t_2) or (t_1, t_3) we use logistic regression to estimate respectively $\text{e}_{t_1,t_2}(\text{X})$ and $\text{e}_{t_1,t_3}(\text{X})$.

2. Then we do a matching step: each pair of units receiving t_1 or t_2 are matched using $\hat{e}_{t_1,t_2}(X)$ and pairs of units receiving t_1 or t_3 are matched using $\hat{e}_{t_1,t_3}(X)$.

3. Construction of matched triplets with the patients receiving t_1 matched to both a unit receiving t_2 and a unit receiving t_3, with their matches. Matched pairs from treatments t_1 and t_3 are dropped if the unit getting t_1 did not match a unit on treatment t_2, and pairs of units receiving t_1 and t_2 are dropped when there is no match for the reference unit to a unit receiving t_3.

Notation:
E_{3i}: indicator of getting two pairwise treatments.

$$E_{3i} = \begin{cases} 1 \; if \; E_{2i}(t_1, t_2) = 1 \; and \; E_{2i}(t_1, t_3) = 1 \\ 0 \; otherwise \end{cases}$$

We then estimate the average difference in the potential outcomes:

$$PATT_{E_3(t_1|t_1,t_2)} = E(Y_i(t_1) - Y_i(t_2)|T_i = t_1, E_{3i} = 1)$$

$$PATT_{E_3(t_1|t_1,t_3)} = E(Y_i(t_1) - Y_i(t_3)|T_i = t_1, E_{3i} = 1)$$

$$PATT_{E_3(t_1|t_2,t_3)} = E(Y_i(t_2) - Y_i(t_3)|T_i = t_1, E_{3i} = 1)$$

This method can underestimate the sampling variance, because we do not take into account the variability caused by the matching procedure.

2.5 Common support

The goal of this method is to find the subject that are eligible to all the treatments and to estimate the effect only with this group of subject :

1. we estimate our vector of propensity scores $R(X)$ using a multinomial regression for example

2. We compute

$$r(t, X)^{(low)} = maxmin(r(t, X|T = t_1)), ..., min(r(t, X|T = t_Z))$$
$$r(t, X)^{(high)} = minmax(r(t, X|T = t_1)), ..., max(r(t, X|T = t_Z))$$

3. we eliminate all the subjects that have a propensity score outside the interval $[r(t, X)^{(low)}, r(t, X)^{(high)}]$

4. we do once again the previous step with the the subjects left

Let E_{4i} be the indicator for all treatments eligibility, where

$$E_{4i} = \begin{cases} 1 & if r(t, X_i) \in r(t, X)^{(low)}, (r(t, X)^{(high)}) \forall t \in T \\ otherwise \end{cases}$$

Using t1 as a reference treatment, PATT 's among subjects eligible for all treatments are defined as follows.

$$PATT_{E4}(t_1|t_1, t_2) = E[Y_i(t_1)Y_i(t_2)|T_i = t_1, E_{4i} = 1]$$
$$PATT_{E4}(t_1|t_1, t_3) = E[Y_i(t_1)Y_i(t_3)|T_i = t_1, E_{4i} = 1]$$
$$\ldots = \ldots$$
$$PATT_{E4}(t_1|t_1, t_Z) = E[Y_i(t_1)Y_i(t_Z)|T_i = t_1, E_{4i} = 1]$$

This approach has benefits regarding its definition of eligibility. All the estimates defined above are transitive so we can compare for instance the difference of effect between two treatment using only the Z-1 estimators defined above.

2.6 Vector Matching

Methods of matching are quite good providing good estimator for treatment effect. As for the binary case, the goal is to find some groups of subject receiving different treatment but with similar propensity scores. In the case of multiple treatment, the idea will be to create a number of sets corresponding to intervals of propensity scores so that there is at least one unit for each treatment. We can reduce step by step the size of the intervals to get more precise matching. However such an approach is very sensitive to the order of subclassification and can result to a number of data points without matching. An alternative to this naive matching is the following procedure:

1. Classify all units using KMC on the logit transform of $\hat{R}_{t,t'}(X)$, where $\hat{R}_{t,t'}(X) = (r(l,X), \forall l \neq t, t')$. This forms K strata of subjects, with similar $Z-2$ GPS scores (not including $r(t,X)$ or $r(t',X)$in each $k \in K$).
2. Within each strata $k \in K$, use 1:1 matching to match those receiving t to those receiving t' on $logit(r(t,X))$. Matching is performed with replacement using a caliper of $\epsilon SD(logit(r(t,X)))$,where $\epsilon = 0.25$.
3. Subjects receiving t who were matched to subjects receiving all treatments $l \neq t$, along with their matches receiving the other treatments, compose the final matched cohort.

At the end of this vector matching procedure, we are left with an optimal number of sets of unit from the reference treatment and matched units from each of the other Z 1 treatments. we have ensured that matched units are close on one component (step 2) of the GPS and roughly similar on the other components (step 1)

Finally we can use this matched sets to do the inference. Indeed, inferences using vector matching can be obtained by contrasting those matched using a weighted average, with weights proportional to ψ_i, where ψ_i is the number of times subject i is part of a matched set. Let $n_t rip$ be the number of matched sets:

$$
\begin{aligned}
SATT_{E_4(t_1|t_1,t_2)} &= \frac{\sum_{i\in E_4} Y_i I(T_i = t_1)\phi_i - Y_i I(T_i = t_2)\phi_i}{n_t rip} \\
SATT_{E_4(t_1|t_1,t_2)} &= \frac{\sum_{i\in E_4} Y_i I(T_i = t_1)\phi_i - Y_i I(T_i = t_3)\phi_i}{n_t rip} \\
\dots &= \dots \\
SATT_{E_4(t_1|t_1,t_2)} &= \frac{\sum_{i\in E_4} Y_i I(T_i = t_1)\phi_i - Y_i I(T_i = t_Z)\phi_i}{n_t rip}
\end{aligned}
$$

2.7 Propensity Score Subclassification for Multiple Treatment:

The subclassification method has an advantage of being much easier to implement than the matching method in the case of multiple treatment. The estimation of the average treatment effects needs an adjustment for differences in pre-treatment variables. This adjustment uses the propensity score proposed by Rosenbaum Rubin (1983, 1984).

The goal of the subclassification method used is to divide the sample into a number of subclasses by the value of the propensity score $r(t,x)$ such that $r(t,x) = P(T = t|X = x)$. r is called the generalized propensity score. Based on Cochran (1968) who shows that this removes much of the bias,and we often use five subclasses.

We are interested in $\tau(t,t')$ for some pair of treatment levels t and t'. $\tau(t,t') = E[Y_i(t)] - E[Y_i(t')]$ To estimate these expectations we construct subclasses or **strata** based on $r(t,x)$.

We estimate the first term by :

Step 1: we estimate the value of $Y_i(t)$ on the **jth strata**:

$$\hat{\mu}_{jt} = \frac{1}{N_{jt}} \sum_{i: q_{j-1}^{p(t|x)} \leq p(t|X_i) \leq q_j^{p(t|x)}, T_i = t} Y_i^{obs} \tag{1}$$

Where $q_j^{r(t,x)}$ is the quintile of the empirical distribution $p(t|X_i)$ in the sample, and N_{jt} is the number of units with $q_{j-1}^{p(t|x)} \leq p(t|X_i) \leq q_j^{p(t|x)}$ and $T_i = t$.

Step 2: The overall average of $Y_i(t)$ is then estimated as

$$\hat{E}[Y_i(t)] = \sum_{j=1}^{5} \frac{N_t}{N} . \hat{\mu}_{jt}$$

In the binary treatment case, we do not construct subclasses defined by similar values for the T 1 propensity scores such that we can estimate causal effects within the subclasses. Instead we construct subclasses defined by similar values for a single propensity score at a particular treatment level so that we can estimate the average potential outcome for that treatment level within the subclasses, and we do so separately for each treatment level, with different subclasses for each treatment level.

3 Continous Treatement

3.1 Covariate balancing propensity score for continuous treatment:

Propensity score methods are popular among researchers who wish to infer causal effects in observational studies. Under the assumption of unconfoundedness that $T_i \perp\!\!\!\perp Y_i|X_i$, propensity score matching and weighting methods aim to balance observed covariates across different values of a treatment variable. Despite the popularity of propensity score methods, the vast majority of their applications have been confined to a binary treatment.

In this part, we will present a new methods on the propensity score. Once we obtain the estimated propensity score, we can employ a variety of methods including regression adjustment and subclassification to estimate causal effects. In the setting with continuous treatment, the causal effect of treatment can be captured by the Average Dose Response Function (ADRF). $ADRF(t) = E[Y_i(t)]$.

The covariate balancing generalized propensity score (CBGPS) methodology Based on Generalized Propensity Score adapts covariate balancing condition for continuous treatment that $E[P(T_i|X_i)T_iX_i] = E(T_i)E(X_i) = 0$, where X and T are centralized and orthogonalized in preprocessing as we will see.

3.1.1 Parametric Covariate balancing propensity score for continuous treatment

For the parametric CBGPS, we follow a common practice of assuming a homoskedastic linear model. Then the generalized propensity score as a density function is given by conditional normal density:

$$f_\theta(T_i^*|X_i^*) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(T_i^* - X_i^{*\top}\beta)^2) \tag{2}$$

Where $T_i^* = s_T^{-1/2}(T_i - \bar{T})$ where s_T and $\bar{T}$ are the sample mean and the sample covariance matrix of T_i. Also, $X_i^* = S_X^{-1/2}(X_i - \bar{X})$ and $\theta = (\beta, \sigma^2)$. Note that the scaling by the covariance matrix makes the covariates indepenedent.

We supposed that the generalised propensity score has the last form in a parametric setting. In addition, we follow a typical parametric modeling approach and assume the marginal distribution of

T_i^* to be standard normal (due to centering and scaling), that is, T_i^* is $\overset{iid}{\sim} N(0,1)$. Then the stabilizing weight is given by :

$$w_i = \frac{f(T_i^*)}{f_\theta(T_i^*|X_i^*)} = \sigma exp\left[-\frac{1}{2\sigma^2}(T_i^* - X_i^{*\top}\beta)^2 - \frac{T_i^2}{2}\right]$$

Under the method of moments framework, we can estimate of $\theta = (\beta, \sigma^2)$ using the following equation:

$$E\begin{pmatrix} \frac{1}{\sigma^2}(T_i^* - X_i^{*\top}\beta)^2 - 1 \\ \sigma exp\left[\frac{1}{2\sigma^2}(T_i^* - X_i^{*\top}\beta)^2 - \frac{T_i^2}{2}\right] T_i^* X_i^* \end{pmatrix} = 0$$

The estimate of , which we denote by $\hat{\theta}$, is obtained by numerically solving the last equation.

One advantage of this parametric approach is that we can derive the asymptotic variance of the estimated causal effects by taking into account the estimation uncertainty of the generalized propensity score. This avoids the use of a more computationally intensive procedure such as bootstrap.

3.1.2 Non-parametric Covariate balancing propensity score for continuous treatment

In the parametric setting we supposed a specific form of the conditional density of the T_i^* and X_i^*. Now we will consider a nonparametric extension of the CBGPS we refer to as npCBGPS.
This method does not involve direct estimation of the generalized propensity score, and thus does not require the model to be correctly specified. Rather, we use an empirical likelihood approach to chose weights that represent the stabilizing inverse generalized propensity score, and simultaneously ensure balancing conditions (zero correlation with the treatment) are met in the sample.

The weights that we have previously defined as $w_i = \frac{f(T_i^*)}{f_\theta(T_i^*|X_i^*)}$ has a mean of 1:

$$E(w_i) = \int\int \frac{f(T_i^*)}{f_\theta(T_i^*|X_i^*)} f(T_i^*, X_i^*) dT_i^* dX_i^* = 1$$

In the same manner we can prove that :

$$E(w_i T_i^* X_i^*) = E(T_i^*)E(X_i^*) = 0$$

Also, we can prove that:

$$E(w_i T_i^*) = E(T_i^*) = 0,\ E(w_i X_i^*) = E(X_i^*) = 0 \tag{3}$$

Altogether, the constraints on the mean of w_i, on the marginal means of X and T, and on the cross-products X^T give rise to the sample conditions,

$$\sum_{i=1}^{N} w_i g(X_i^*, T_i^*) = 0,\ \sum_{i=1}^{N} w_i - N = 0 \tag{4}$$

where $g(X_i^*, T_i^*) = (X_i^*, T_i^*, X_i^* T_i^*)^\top$

In this setting, we can express the joint density of each observation in relation to the weights as:$f(T_i^*, X_i^*) = \frac{1}{w_i} f(T_i^*) f(X_i^*)$. Now, our goal is to maximize the empirical likelihood of the data by choosing w_i, but also require w_i to satisfy the constraints listed above. Thus we maximize :

$$\prod_{i=1}^{N} f(T_i^*, X_i^*) = \prod_{i=1}^{N} \frac{1}{w_i} f(T_i^*) f(X_i^*)$$

Subject to :

$$\sum_{i=1}^{N} w_i g(X_i^*, T_i^*) = 0,\ \sum_{i=1}^{N} w_i - N = 0$$

Which is equivalent to maximizing:

$$arg \min_{w \in R^N} \sum_{i=1}^{N} log(w_i)$$

subject to the same above constraints.

Numerical algorithm to find w:

The problem we have is a maximization problem under constraints, thus we can follow an approach similar to the standard Lagrange multiplier technique for numerically solving this optimization problem.

The Lagrangian is : $\mathcal{L}(w_i, \lambda, \gamma) = \sum_{i=1}^{N} log(w_i) + \lambda(N - \sum_{i=1}^{N} w_i) + \gamma^\top \sum_{i=1}^{N} w_i g(X_i^*, T_i^*)$

Using the first order conditions we find that $\lambda = 1$ and $w_i = \frac{1}{1-\gamma^\top g(X_i^*, T_i^*)}$.
Therefore, our constrained optimization problem is solved by the unconstrained maximization,

$$arg \max_{\gamma \in R^K} \sum_{i=1}^{N} log(1 - \gamma^\top g(X_i^*, T_i^*)) \tag{5}$$

where K is the covariate space dimension. This problem could be solved efficiently with one algorithm of unconstrained optimization like gradient descent or BFGS.

In many practical situations, the numerical algorithm described above fails to find a solution, this comes from our forcing on the covariate balancing conditions.

3.2 Generative Adversarial De-confounding (GAD) algorithm

The goal of this algorithm could be seen as two steps: first make the covariates X independent from the treatment T by randomly shuffling the value of each covariate $X_{.,i}$ over all samples in observed data $D_{obs} = \{T, X\}$, the shuffled covariates would become independent with the treatment T if sample size $n \to \infty$.
Second, develop a Generative Adversarial Network to learn a sample weight w on the observed data D_{obs} such that the distribution of weighted observed data would be similar even identical with the "calibration" data D_{cal}, formally $wP(T, X) = P(T, X')$.

Inspired by the immense success of Generative Adversarial Network (GAN), in producing simulated data that highly resembles the distribution of real-world samples, we propose a novel framework that leverages the objective of GAN to the task of generating weight for ensuring the distribution of adjusted observed data has the identical distribution of the "calibration" one.

Our loss is :

$$L(\mathbf{w}, d) = E_{(t,x)\sim D_{cal}}[l(d(t, x), 1)] \tag{6}$$

$$+ E_{(t,x)\sim D_{obs}}[w_{(t,x)} l(d(t, x), 0)], \tag{7}$$

$$s.t; E_{(t,x)\sim D_{cal}}[w_{(t,x)}] = 1, \mathbf{w} \succeq 0 \tag{8}$$

Where $d(.)$ is our descriminator. The term $E_{(t,x)\sim D_{cal}}[w_{(t,x)}] = 1$ avoids all sample weight to be zero, and $\mathbf{w} \succeq 0$ constrains each sample weight to be non-negative.

Now, we can get our weight by solving the following optimization problem:

$$\hat{\mathbf{w}} = arg \max_{\mathbf{w}} (\min_{d} L(\mathbf{w}, d))$$

In practice, we switch 0/1 labels for two data distributions, resulting the following loss functions for both w and discriminator $d()$ to minimize alternately:$L_d(\mathbf{w}, d)$.

Algorithm :

Input: Observed Data $D_{obs} = \{T, X\}$, stopping criterion $h(D_{obs}, D_{target}, w)$, optimizer for discriminator, $SGD(, L_d(w, d))$, and optimizer for w, $Ranger(w, L_w(w, d))$

Output:sample weight **w**

for $i = 1, 2, , p$ do:

Generating shuffled covariate $X'_{.,i}$ by randomly permuting the elements in $X_{.,i}$ **end for**
Generate target data $D_{cal} = \{T, X'\}$
Initialize sample weight $w^0 = [1, 1, ..., 1]$
Initialize discriminator d$(\cdot)$ with parameter $_0$
Initialize the iteration variable $t = 0$

Repeat:

$\theta^t \overset{S}{\leftarrow} GD(\theta^{t-1}, L_d(w, d))$
Update sample weight $w^t \overset{R}{\leftarrow} ange(w^{t-1}, L_w(w^{t-1}, d))$
Limit mean of sample weight $w_i^t \overset{n}{\leftarrow} w_i^t / \sum_{i=1}^n w_i^t,\ i = 1, ..., n$
until $h(D_{obs}, D_{cal}, w^t)$ satisfied or max iteration is reached
return sample weight w

4 Conclusion

The main methods for estimating the average treatment effect in the Multiple treatment or the average dose response function in the continuous treatments listed in our reports with their characteristics are IPW, Common referent matching,common support, vector matching and propensity score subclassification and for the continuous settings we have Covariate balancing propensity score(parametric and non parametric) and Generative adversarial De-confounding(GAD) the last one outperforms the other one, also we have listed the implementations of some of them in python and R.

References

https://imai.fas.harvard.edu/research/files/CBGPS.pdf

http://proceedings.mlr.press/v127/li20a/li20a.pdf

https://drum.lib.umd.edu/bitstream/handle/1903/18170/Galagate$_u md_0 117 E_1 6898.pdf?sequence =$ 1

Estimation of causal effects with multiple treatments: A review and new ideas. Statistical Science. Lopez, M. J. and Gutman, R. (2017). 32 432?454

The role of the propensity score in estimating dose-response functions. Imbens 2000. Biometrika 87 706?710.

Generalized propensity score for estimating the average treatment effect of multiple treatments. Feng, P., Zhou, X. H., Zou, Q. M., Fan, M. Y. and Li, X. S. (2012). Statistics in Medicine 31 681?697

A tutorial on propensity score estimation for multiple treatments using generalized boosted models. McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R. and Burgette, L. F. (2013).Statistics in Medicine 32 3388?3414

Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. Zanutto, E., Lu, B. and Hornik, R. (2005). Journal of Educational and Behavioral Statistics 30 59?73.