New matching algorithm--: Outlier First Matching (OFM) and its performance on Propensity Score Analysis (PSA) under new Stepwise Matching Framework (SMF)

Yi Sun
University at Albany, State University of New York, njadwards@yahoo.com

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Recommended Citation
https://scholarsarchive.library.albany.edu/legacy-etd/1281

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive.
Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
NEW MATCHING ALGORITHM – OUTLIER FIRST MATCHING (OFM) AND ITS PERFORMANCE ON PROPENSITY SCORE ANALYSIS (PSA) UNDER NEW STEPWISE MATCHING FRAMEWORK (SMF)

by

Yi Sun

A Dissertation
Submitted to the University at Albany, State University of New York
in Partial Fulfillment of
the Requirements for the Degree of
Doctor of Philosophy

College of Public Health
Department of Epidemiology & Biostatistics
2014
New Matching Algorithm – Outlier First Matching (OFM) And Its
Performance On Propensity Score Analysis (PSA) Under New Stepwise
Matching Framework (SMF)

by

Yi Sun

COPYRIGHT 2014
Abstract
An observational study is an empirical investigation of treatment effect when randomized experimentation is not ethical or feasible (Rosenbaum 2009). Observational studies are common in real life due to the following reasons: a) randomization is not feasible due to the ethical or financial reason; b) data are collected from survey or other resources where the object and design of the study has not been determined (e.g. retrospective study using administrative records); c) little knowledge on the given region so that some preliminary studies of observational data are conducted to formulate hypotheses to be tested in subsequent experiments. When statistical analysis are done using observational studies, the following issues need to be considered: a) the lack of randomization may lead to a selection bias; b) representativeness of sampling with respect to the problem under consideration (e.g. study of factors influencing a rare disease using a nationally representative survey with respective to race, income, and gender but not with respect to the rare disease condition). We will use the following sample to illustrate the challenges of observational studies and possible mitigation measures.
Our example is based on the study by Lalonde (1986), which evaluated the impact of job training on the earnings improvement of low-skilled workers in 1970’s (In Paper 1 section 1.5.2, we will discuss this data set in more detail). The treatment effect estimated from the observational study was quite different from the one obtained using the baseline randomized “National Supported Work (NSW) Experiment” carried out in the mid-1970’s. Now we understand the treatment effect which is the impact of job training. Selection bias may contaminate the treatment effect, in other words, workers who receive
the job training may be fundamentally different from those who do not. Furthermore, the sample of control group selected for observational study by Lalonde may not represent the sample of control group from the original NSW experiment.

In this study, we address the issue of lack of randomization by applying a new matching algorithm (Outlier First Matching, OFM) which can be used in conjunction with the Propensity Score Analysis (PSA) or other similar methods to achieve the convincible treatment effect estimation in observational studies.

This dissertation consists of three papers.

Paper 1 proposes a new “Stepwise Matching Framework (SMF)” and rationalizes its usage in causal inference study (especially for PSA study using observational data). Furthermore, under the new framework of SMF, one new matching algorithm (Outlier First Matching or OFM in short) will be introduced. Its performance along with other well-known matching algorithms will be studied using the cross sectional data.

Paper 2 extends methods of paper 1 to correlated data (especially to longitudinal data). In the circumstance of correlated data (e.g. longitudinal data), besides the selection bias as in cross-sectional observational data, the repeated measures bring out the between-subject and within-subject correlation. Furthermore, the repeated measures can also bring out the missing value problem and rolling enrollment problem. All of above challenges from correlated data complexity the data structure and need to be addressed using more complex model and methodology. Our methodology calculate the variant p-score of control subjects at each time point and generate the p-score difference from each control subject to every treatment subject at treatment subject’s time point. Then such p-score differences are summarized to create the distance matrix for next step analysis. Once
again, the performance of OFM and other well-established matching algorithms are compared side by side and the conclusion will be summarized through simulation and real data applications.

Paper 3 handles missing value problem in longitudinal data. As we have mentioned in paper 2, the complexity of data structure of longitudinal data often comes with the problem of missing data. Due to the possibility of between subject and within subject correlation, the traditional imputation methodology will probably ignore the above two correlations so that it may lead to biased or inefficient imputation of missing data. We adopt one missing value imputation strategy introduced by Schafer and Yucel (2002) through one R package “pan” to handle the above two correlations. The “imputed complete data” will be treated using the similar methodology as paper 2. Then MI results will be summarized using Rubin’s rule (1987). The conclusion will be drawn based on the findings through simulation study and compared to what we have found in complete longitudinal data study in paper 2.

In last section, we conclude the dissertation with the discussion of preliminary results, as well as the strengths and limitations of the present research. Also we will point out the direction of the future study and provide suggestions to practice works.
Acknowledgements

I still have fresh memory of the moment that I was admitted to the Department of Epidemiology and Biostatistics in 2006 to have the opportunity to learn from so many cutting-edge professors on this challenging and excellent program. I am grateful to so many things but hard to find the exact word to express my feelings. But one thing that I am confident to say is that I would not change anything if I had the chance to start all over again.

I am full of gratitude deeply in my heart to Dr. Recai M. Yucel and Dr. Robert M. Pruzek, my advisors, for their strong supports, constant encouragements, outstanding guidance and tremendous helps during my study and authoring of this thesis over the years. I always remember their endless kindness and patience to guide me when I was in my very difficult time that I was almost thinking of giving up. What I learned from them is not only the research skill and idea, but also the optimistic attitude toward life, which will absolutely benefit me a lot in my life.

I am grateful to Dr. Gregory A DiRienzo and Dr. Tao Lu for accepting to be my committee member, and for their helps and encouragements to me to pursue my PhD study. I also appreciate it very much for their insightful comments and constructive endeavors on this work. Without their critical suggestions, this thesis would appear less convincible and cannot be read. I am so grateful to them that they are always available and always supportive when I come to them with questions even they are always busy, and all their comments are cherished.

I would also like to thank Dr. David Strogatz, Dr. Edward Fitzgerald, Dr. Howard Stratton, Dr. Larry Lessner, Dr. Igor Zurbenko, Mrs. Nikki Malachowski, Mrs. Judith L.
Moran and other staffs in School of Public Health at SUNY, Albany for their helps on my study. I would to express my special thanks to Dr. Charles J. Neighbors who trigger the idea of this dissertation. I would also like to cherish the encouragements from Dr. Jon Morgenstern, Dr. Rajeev Yerneni, Dr. Gerod Hall and other CASA Columbia Staffs, Constance Burke, Edward P. Tesiny, William Hogan and other OASAS/NYS staffs and RFMH staffs. Furthermore, I would like to partake the triumph with my friends, Ming Liu, Xuelin Wen, Changning Xu, Fazhan Li, Dr. Kuangnan Xiong, Jin Qu, Honglei Tang, Dr. Ying Guo, Meng Wu, Yan Wu, Xiaohang Wang, Dr. Yimin Dong, Dr. Lei Chen and many, many others.

I would also like to offer my sincerest thanks to my parents, Yuzhu Sun and Guojin Zhong, for their patience and tolerance over this lengthy process and the long absence. In addition, I thank my beloved son, Alfred Sun, and my beloved wife, Qiao Liang, for being next to me during this important part of my life. I apologize that I spent so many midnights, weekends and holidays on the dissertation when I was supposed to be with them. This dissertation is absolutely dedicated to them.

I fear any enumeration of the people who deserve acknowledgement will inevitably be incomplete. But I would like to share this joyful moment with all my friends. Without their encouragements, the completion of this dissertation cannot be reached.

Finally, thanks for reading this dissertation.
Contents

Abstract ................................................................................................................................. iii
Acknowledgements ........................................................................................................... vi

1 Applying Outlier First Matching (OFM) under new Stepwise Matching Framework (SMF) for
cross section data .............................................................................................................. 1

1.1 Introduction ................................................................................................................. 1

1.2 Literature Review ....................................................................................................... 2
   1.2.1 Causal inference and PSA study ........................................................................... 2
   1.2.2 Fundamental Questions about Matching .............................................................. 7
   1.2.3 Ration and selection of matched units ................................................................. 9
   1.2.4 Non-compatible values and “tied” values ............................................................ 11

1.3 New Stepwise Matching Framework (SMF) ................................................................. 12
   1.3.1 Traditional matching framework .......................................................................... 12
   1.3.2 The new stepwise matching framework (SMF) based on distance matrix .......... 13

1.4 Introduction to “Outlier First Matching” (OFM) .......................................................... 16
   1.4.1 What is OFM? ....................................................................................................... 16
   1.4.2 “Non Compatible Pattern Matrix” and “Determination inequality” .................... 17
   1.4.3 OFM algorithm .................................................................................................... 19

1.5 Performance of OFM under SMF on cross-sectional study ......................................... 22
   1.5.1 Simulated data study ............................................................................................ 22
   1.5.2 Application of “Lalonde” Data ........................................................................... 26

1.6 Discussion ..................................................................................................................... 29

2 Applying OFM under SMF for longitudinal data ............................................................ 31

2.1 Introduction ................................................................................................................. 31

2.2 Methodology ............................................................................................................. 33

2.3 Performance of OFM under SMF on longitudinal study ........................................... 38
   2.3.1 Simulated data study ........................................................................................... 38
   2.3.2 Application of “Care Management” Data ........................................................... 41

2.4 Discussion ................................................................................................................... 46
List of Tables

Table 1: List of different matching methods based on their targeting interests.................................61

Table 2: Comparison of estimated treatment effects (true parameter=1) from different matching methods for simulated cross-sectional data (multivariate-normal) using correct model (simulation times=1000) .................................................................................................................................62

Table 3: Comparison of estimated treatment effects (true parameter=1) from different matching methods for simulated cross-sectional data (multivariate-normal) using incorrect model a (simulation times=1000) .................................................................................................................................63

Table 4: Comparison of estimated treatment effects (true parameter=1) from different matching methods for simulated cross-sectional skew data (multivariate-skewnormal) using correct model (simulation times=1000) .................................................................................................................................64

Table 5: Comparison of estimated treatment effects (true parameter=1) from different matching methods for simulated cross-sectional data (multivariate-skewnormal) using incorrect model a (simulation times=1000) .................................................................................................................................65

Table 6: Actual Treatment Effect from NSW Experiment ........................................................................66

Table 7: Covariates description of "Lalonde" data from R package “MatchIt” ........................................67

Table 8: Comparison of estimated treatment effects among different matching methods of Lalonde Study .............................................................................................................................................68

Table 9: Comparison of estimated treatment effects (true parameter=1) from different matching methods for simulated longitudinal data using correct model (simulation times=1000) ............................................................69

Table 10: Comparison of estimated treatment effects (true parameter=1.0) from different matching methods for simulated longitudinal data using incorrect model (simulation times=1000) .................................................................................................................................70

Table 11: Covariates description of "Management Care system" data...............................................71

Table 12: Comparison of the balances from different matching algorithm for Management Care Study .............................................................................................................................................72

Table 13: Top 3 most skew matched pairs from different matching algorithms for Management Care Study .............................................................................................................................................73

Table 14: Comparison of estimated treatment effects among different matching methods for Management Care Study .............................................................................................................................................74
Table 15: Comparison of estimated treatment effects (true parameter=1) from different matching methods for simulated longitudinal data with missing value using correct model (simulation times=1000, missing rate = 0.3)..................................................................................................... 75

Table 16: Comparison of estimated treatment effects (true parameter=1) from different matching methods for simulated longitudinal data with missing value using incorrect model (simulation times=1000, missing rate = 0.3)..................................................................................................... 76
List of Figures

Figure 1: Examples of two matching methods using NN and RDM ........................................ 77
Figure 2: Graph for example 1 .................................................................................................. 78
Figure 3: Matching results based on different distance measures ......................................... 79
Figure 4: Flowchart of OFM algorithm .................................................................................. 80
Figure 5 Balance assessment of “p_score” distributions between treatment and control groups before / after matching (from different matching algorithms) for “Lalonde” data ............... 81
Figure 6: Balance assessment of “p_score” distributions between treatment and control groups using OFM within caliper for “Lalonde” data ........................................................................... 82
Figure 7: Histograms of absolute distance between treatment and control units of each matched pair for "Lalonde" data ........................................................................................................ 83
Figure 8: Back on back histograms of p-scores between engagement and control groups before / after matching (from different matching algorithms) for Management Care Study ................. 84
Figure 9: The distribution of absolute value of difference of p-score among all matched pairs for Management Care Study ........................................................................................................ 85
Figure 10: Quarterly Cost Difference between engagement and control groups for Management Care Study ........................................................................................................................................ 86
Applying Outlier First Matching (OFM) under new Stepwise Matching Framework (SMF) for cross section data

Abstract
Rubin suggested that the causal inference can be treated as one missing value problem with the challenge of finding the counterfactual unobserved outcome. The properties of randomization experiment will be studied and the similar properties can be achieved under the assumption of strong (weak) ignorability through propensity score analysis (PSA) for observational causal inference study. Matching algorithm is one widely used methodology in observational causal inference studies such as PSA study. New stepwise matching framework (SMF) will be introduced and compared to the traditional matching framework. Treatment effect estimated using new matching algorithm – Outlier First Matching (OFM) will be compared to ones estimated using other well-known matching algorithms under the new matching framework “SMF” using cross-sectional data. The simulation and real data analysis suggests that OFM may select the same units as other matching algorithms but generate the least skew matched pairs in cross-sectional PSA study.

1.1 Introduction
Rubin Causal Model treats the treatment effect estimation problem as the missing value problem. The property of randomized trial will be addressed and lead to the counterpart for observational studies (strong ignorability). The idea of matching will be traced back and its usage in treatment effect study (especially for PSA study using observational data) will be emphasized. Then the new “Stepwise Matching Framework (SMF)” and “Outlier First Matching (OFM)” algorithm will be introduced and the application of OFM on cross sectional data will be compared with other well-known matching algorithms.
1.2 Literature Review

1.2.1 Causal inference and PSA study

One fundamental problem which interests statisticians is to draw inference on relationships between treatment and outcome, sometimes referred as causal inference (Holland 1986). Based on work by Rubin (1974; 1978) that extends the idea originally proposed by Jerzy Neyman (1923), Holland (1986) coined the term “Rubin Causal Model”. From the view of Rubin, the causal inference (or inference treatment effect) can be conceptualized in terms of the difference between potential outcomes (i.e. $Y_i(1) − Y_i(0)$) under treatment (i.e. $Y_i(1)$) and control conditions (i.e. $Y_i(0)$). Note that only one of $Y_i(1)$ or $Y_i(0)$ is observed for each experimental unit (Rubin 1974; Rubin 1978; Rosenbaum and Rubin 1983; Rubin 1990). This Counterfactual logic implies that both of these outcomes are potentially available for all units, even though only one of them can be observed in practice. Thus, the estimation of causal effects can be treated as a missing data problem with the core challenge of predicting the unobserved potential outcome (Rubin 1976).

In randomized studies, one element will be enrolled into treatment or control group only by a random process, which means that the random assignment mechanism ensures “balance” of all covariates (observed or non-observed). Here “balance” is defined as the similarity of the empirical distributions of the full set of covariates in the treatment and control groups. Under randomization, treatment assignment (T) is not only independent of all covariates (X) (i.e. $T \perp X$) but also independent of potential outcomes (i.e. $T \perp (Y(0), Y(1))$). Under randomization, the natural estimator of treatment effect for unit i, $Y_i(1) − Y_i(0)$, will be systematically determined only by the treatment assignment (T),
not by the covariates (X). Therefore, the average (or expected) treatment effect is $\text{ATE} = E[Y_i(1) - Y_i(0)] = E[Y_i(1)|T_i=1] - E[Y_i(0)|T_i=0]$ (Bellinger, Leviton et al. 1987; Imbens 2004). For this reason Fisher (1935) called randomization ‘the reasoned basis for inference’ in applied research.

In observational studies, however, the direct or unadjusted estimator (of the treatment effect), $(Y_i(1) - Y_i(0))$ for unit $i$, is not feasible by design, since this estimator not only includes the effect of the treatment, but also the influence of confounders that are not balanced. In observational studies, individuals in treatment groups tend to be different from the ones in control group not by chance but in some characteristics that the investigator may or may not observe (i.e. $T \perp X$ does not hold). Potential selection bias can mislead one’s attempts to interpret treatment differences when such bias exists. Since the direct comparison of the treatment difference leads to biased results, investigators often want to mimic the comparison based on a randomization mechanism despite the possible effects of confounders. According to the key properties shown above of the randomization mechanism: (1) $T \perp X$ and (2) $T \perp (Y(0), Y(1))$ ($T$=treatment; $X$=covariates; $Y(1)$= outcome under active treatment; $Y(0)$=outcome under control treatment), if some method can be utilized to achieve (or at least mimic) these two properties, the impact of confounders could probably be reduced, or even eliminated. Following Rosenbaum and Rubin (1983), if we assume that conditional on $X$, (1) treatment assignment ($T$) is independent of the potential outcomes ($Y(0), Y(1)$) given the covariates ($X$): $T \perp (Y(0), Y(1))|X$, and (2) there is a positive probability of receiving each treatment for all values of $X$: $0 < P(T = 1|X) < 1$ for all $X$, we meet the criterion called “strong ignorability” (Rubin 1974). Under this circumstance, we can approximately achieve the similar
property \((T \perp (Y(0), Y(1))|x)\) as the point 2 in the randomization mechanism so that we obtain 
\(E[Y_{ij}|X_i, T_i=1] = E[Y_{ij}|X_i, T_i=0] = E[Y_{ij}|X_i, T_i=j, j=0,1]\). Therefore, the Average Effect for the Treatment (ATT) can be estimated by: 
\[
\text{ATT} = E[Y_i(1) - Y_i(0)|X_i, T_i=1] = E[Y_i(1)|X, T_i=1]-E[Y_i(0)|X_i, T_i=1]= E[Y_i(1)|X_i, T_i=1]-E[Y_i(0)|X_i, T_i=0]
\]
(Imbens 2004). If the research interest is ATT, the un-confoundedness assumption of “strong ignorability” can be weakened to mean independence: 
\[
E(Y_{ij}|X_i, T_i) = E(Y_{ij}|X_i).
\]
And the overlap assumption for ATT only requires that the support of \(X\) for the treatment be a subset of the support of \(X\) for control observations (Heckman, Ichimura et al. 1998).

The application of matching in causal inference study can be traced back to Cochran and Rubin (1973) with respect to their “multiple covariate matching”. According to Cochran, if we can control for each of the covariates (the possible confounders) separately (by finding the exact or similar values within defined boundaries), we may achieve near perfectly matched pairs with respect to all the controlled covariates. The problem of this method is that matching on each covariate may potentially decrease matched pairs and result in very few (possibly even none) matched pairs. To assure unbiased estimation of treatment effects and obtain a satisfactory number of matched pairs, statisticians have created many different methods to propose “dimension reduction”. When the covariates are continuous and have an ellipsoidal distribution, one can decrease the distance of each covariate between the treatment and control simultaneously. This is a property known as “equal percent bias reduction” (EPBR) (Rubin and Thomas 1992). Under this circumstance, a matching method using relative smaller dimensions (even scalar) can be applied to reduce the distances of all covariates and improve the similarity between the treatment and control units (Diamond and Sekhon 2006).
Rosenbaum and Rubin (1983) introduced the “Propensity score” (p-score), which is a scalar score defined as the probability of receiving the treatment given the observed covariates. Rosenbaum and Rubin (1983) further proved that under “strong ignorability”, we have $T \perp (Y(0), Y(1))|e(X)$, where $e(x)$ stands for the p-score. Therefore, given a well-constructed p-score, $e(x)$, we can use it instead of “multivariate distance” to match treatment and control groups. This means that given the “strong ignorability” assumption, we can (approximately) achieve the similar property $(T \perp (Y(0), Y(1)))$ as in a randomization mechanism to minimize the effect of confounders when comparing potential outcomes for the treatment and control groups.

Compared to the traditional multivariate regression process, PSA has the advantage in estimating the treatment effect when any one or more of the following criteria are met (Cochran and Rubin 1973; Baser 2006):

C1. The distribution of the covariates in both groups is not symmetrical.
C2. The sample sizes are quite different.
C3. The distributions of the covariates in the groups have different variances.
C4. The means of propensity scores in the two groups are more than half a standard deviation apart.
C5. The ratio of the variances of the propensity score in the two groups differs from one.
C6. The ratio of the variances of the residuals of the covariates after adjusting for the propensity score is not close to one (one-half or two are far too extreme).

In practice, PSA is usually applied in two phases: Phase I: Generate $e(x)$ (i.e. p-score); Phase II: Units with similar $e(x)$ are matched. In practice, more condensed version of $e(x)$
is used by creating strata based on e(x). In the later, matched units are found in each stratum. In phase II of PSA, three methods are usually used to estimate treatment effect:

Method 1 -- Match each unit from the treatment group with one or more units in control group using p-scores from phase I, and then compare matched individuals on the outcome measure.

Method 2 -- Stratify using p-scores and then compare treatment and control groups’ means on the outcome measure within each stratum and combine the treatment effect from each stratum to form the overall estimated treatment effect.

Method 3 -- Inverse probability of treatment weighting using the p-score (IPTW).

Note that method 2 and 3 can also be seen as the specified (n_i : m_i) matching methods using our new matching framework introduced below.

These three methods noted above represent different ways to estimate treatment effects, which (in the population) generally take two forms: The first is called the Average Treatment Effect (ATC) (Bellinger, Leviton et al. 1987), while the second is called the Average Effect for the Treatment (ATT). They can be defined as: \( \text{ATE} = E[Y_i(1) - Y_i(0)] \) (the population average treatment effect; i.e., the difference between population means for the treatment and control outcomes) and \( \text{ATT} = E[Y_i(1) - Y_i(0)|Z = 1] \) (population average treatment effect for the treatment) (Imbens 2004). Methods 2 and 3 (above) aim at estimation of ATE, while Method 1 estimates ATT.

As mentioned above, there are at least three different methods to choose to estimate treatment effect in Phase II of PSA. However, recent studies of PSA methods tend to prefer matching methodology over the other two methods since matching eliminates a greater proportion of the systematic differences between treatment and control groups.
than other methodologies (Austin 2009). Recall that the most important goal of PSA is to achieve the first property of “strong ignorability” (i.e. \( T \perp (Y(0), Y(1))|X \) or \( T \perp (Y(0), Y(1))|e(X) \)). Matching on p-score is the naturally straightforward method to achieve “\( T \perp (Y(0), Y(1))|e(X) \)” compared to the other two options since the other two options cannot guarantee the similarity of e(x) and then the similarity of covariates within each stratum. Adjustment by weights (in IPTW) can partially offset the impact from extreme values on potential outcome in each group, however since such “weight” is generated by using one specific method in Phase I so that even IPTW is one “robust” method but usually is not the best “efficient” method especially for small sample analysis.

As mentioned above, the un-confoundedness assumption of PSA analysis is “strong ignorability”. However, this ignorability cannot be tested. Violation almost certainly exist in practice since (1) The analyst almost never includes all relevant covariates, (2) The analyst almost never knows the interactions among covariates, and (3) The analyst almost never knows the true or fundamental model for constructing the phase I PSA model. Usually the analyst adopts the approach suggested by Rosenbaum (2002) to perform analyses of sensitivity to an unobserved variable. The idea is to find how strong the relationship between the non-ignorable covariate (confounder) and both treatment assignment and the outcome needed to wipe out or account for the observed treatment effect.

### 1.2.2 Fundamental Questions about Matching

Matching is a commonly used method in many divergent fields such as machine learning, game theory, statistics, etc. Although matching has different definitions in different regions, in mathematics, matching refers to one selecting process people select and group
individuals or objects according to some numerical measure of “distance” between them. These distances could be Euclidean, Hamming, Manhattan or any of several others and refer to some measurements describing “how far or close” individuals are to one another in some space using specific calculating methods.

In causal inference, matching refers to a statistical technique which is used to evaluate the effect of a treatment by combing and comparing the treatment and control units (Rubin 1973; Anderson, Kish et al. 1980; Kupper, Karon et al. 1981). Through matching, similarity can be ensured between the treated and non-treated units with regards to “nuisance” factors that might distort the relationship between treatment and outcome (Rubin, 1973). Matching can be used with experimental data, but it is most often recommended for observational studies where treatments are not randomly assigned. Matching is used to find, for any treated unit, at least one non-treated unit with similar observable characteristics, i.e., covariate values. By matching units and then comparing the average difference in outcomes, the analyst can evaluate a treatment’s effect, compared to a control, in such a way that covariate confounding is reduced, or possibly even eliminated. Given that matching is based on all available confounders, “strong ignorability” can become convincing assumption. And under this circumstance, we can assure no systematic differences in covariates’ distributions so that the outcome effect is systematically determined only by the treatment assignment (T).

There are two fundamental questions concerning matching. The first question is “What is the question we aim to answer, or what is the reason we want to match?” The answer to this question depends on specific research interests which will require different methodologies (and consequently lead to different results). Table 1 shows several
common goals in the first column, some of which will be explained below. The second fundamental question is closely connected to the first, viz., “How, i.e., by which method, can we most effectively answer the initial question(s)?” The second column in Table 1 presents the methods related to the answer to the corresponding question. This dissertation is less concerned with the first what question and more about the second, the how question. Nevertheless, the two fundamental questions are related and thus reflection and analysis about each is essential.

Let’s look more closely into the matching processes with respect to these two essential questions. Suppose we have two different groups of individuals (or units), one called a treatment group, the other a control group. We want to match one unit from the treatment group to another from the control group (one to one, without replacement). Given a choice of distance measures we can create a distance matrix where each entry stands for the distance between one element from the treatment group (in m rows) and the other from the control group (in n columns). Given any (m x n) rectangular distance matrix (m<=n), we can first calculate the $P_n^m$ permutation of selections and then compare them to pick up one (or some) that fulfills our particular interests (i.e. smallest distance first, shortest total distance, etc.). However, a straightforward approach like this is usually inefficient so that different methodologies have been introduced to help improve efficiency. Table 1 below includes some well-known methodologies in the second column that could be used to fulfill the interests in column 1.

1.2.3 Ration and selection of matched units

Besides these two fundamental research questions about matching mentioned above, there also exist some issues on matching in practice. One issue has to do with the ratio of
units been matched (1: 1, 1: n, or n1: m1). 1: 1 matching usually results in the best accuracy but sometimes suffers from a problem of small sample size (and hence low power). If we extend our interest from 1: 1 to 1: n (or n1: m1) matching, we usually need to deal with the trade-off between bias and variance. Selecting multiple controls for each treatment unit generally increases bias since the 2nd (3rd, 4th and so on) closest matches are, by definition, further away from the treatment unit than is the 1st or closest match. On the other hand, utilizing multiple matches can decrease the distance matrix variance due to the larger matched sample size, and therefore increase the power for the analysis (Stuart 2010). There are different perspectives on how to find the best treatment: control ratio and how to evaluate the contribution of units in each matched pair. Rubin and Thomas (1996) use an approximation method to determine the best ratio. Some other statisticians such as Rosenbaum (1987), Alberto, Joshua et al. (1998), Czajka (1992), Robins, Hernan et al. (2000), and Lunceford and Davidian (2004) introduced inverse probability of treatment weighting (IPTW) to address the above trade-off, which will be discussed in details later when applying OFM in propensity score analysis (PSA).

Another interesting question about matching is whether or not to allow replacement of units. Here the trade-off between with and without replacement is the number of matched pairs vs. the dependence, and hence the interpretation of results based on matching. Matching with replacement usually leads to more matched pairs since units (either control or treatment) can be used multiple times. This is particularly helpful in settings where there are fewer control individuals comparable to the treatment individuals (e.g., Dehejia (1999)). However, when using matching with replacement, our inference from the matched pairs could easily suffer from the small sample problem or weighting issues
(i.e., counting too much on some individuals). For this reason, *matching without replacement* is more commonly adopted for most analyses even though it may result in an insufficient number of matched units especially when the sample size is relatively small. There exist some other aspects of matching so that the combinations of all these aspects will lead to many distinctive matching methodologies in real life data analysis.

**1.2.4 Non-compatible values and “tied” values**

When matching is performed in practice, one is often faced with obstacles. One of them is “non-compatible values”. “Non-compatible value” problem corresponds to the incompatibility of the specific treatment and control units. Under this circumstance, according to the *distance matrix* mentioned earlier in the dissertation, a “Non-compatible value” can be understood as a “NC” entry in the *distance matrix* and the related non compatible problem can be resolved by examining the “Non Compatible Pattern Matrix” (NCPM) originated from the *distance matrix* (See 1.4.2 “Non Compatible Pattern Matrix” and “Determination inequality” for detail illustration).

Another common barrier for effective matching is the existence of “tied values”. There exist several options to deal with this situation. The first and the easiest option is just to randomly pick up one of equivalents s and use it in the matching. However, this option will generally not bring out the desired result. The second option is to use “jittering” where one adds a small component of random error to each *tied value* and then proceed with matching. Then, one may use bootstrapping to reduce or evaluate impacts of the random error. The investigator may also apply matching on each *tied value* to obtain the complete set of matching result candidates, and then select one (or some) that best meets specified needs. However, the implementation in reality can be much complicated,
especially when the number of candidates of tied values is very large or there exist
different ties at different stages of a matching process. There is no perfect solution for
handling tied values. Analysts usually balance the accuracy and efficiency of matching
before they conclude which methodology is more appropriate. With the new
interpretation of matching, which will be illustrated in the following part, a tied value
problem can be described as having the “same values in any single row or column” for
the original distance matrix.

1.3 New Stepwise Matching Framework (SMF)

1.3.1 Traditional matching framework
Traditionally in a matching process, after calculating the distance, the first thing to do is
to select one unit from the treatment group according to some order (random, sequential
or some other order) (Cochran and Rubin 1973). Figure 1 depicts two criteria for finding
matches: NN, RDM. Whatever order has been chosen, the units from the treatment group
will be picked up one by one, not simultaneously. Secondly, the selected unit will be
matched to one or more units from control group according to some specific rule, such as
NN, RDM, FM, and etc. Then the above steps will be continued for all available units in
the treatment group. Stuart (2010) summarized this in the following four steps, “with the
first three representing the ‘design’ and the fourth the ‘analyses’:

1. Defining “closeness”: the distance measure used to determine whether an individual is
a good match for another;

2. Implementing a matching method, given that measure of closeness: Usually use
specific order to pick up units from treatment group to match to control units;
3. Assessing the quality of the resulting matched samples, and perhaps iterating with Steps (1) and (2) until well-matched samples result, and

4. Analysis of the outcome and estimation of the treatment effect, given the matching done in Step (3).

However, the above steps still suffer the following shortcomings:

- The picking order (even the random order) usually matters in matching process;
- Hard to compare different matching algorithm results and hard to use mathematics language to describe the incompatible value issues (level 2 missing value), etc.;
- Easy to mix up step 1 methods of creating distances (and the distance matrix) to the step 2 methods of selecting units. (more detail will be given in the following section)

1.3.2 The new stepwise matching framework (SMF) based on distance matrix

\[
\begin{bmatrix}
A & a & \ldots & z

a_{11} & \ldots & a_{1n}

\ldots & \ldots & \ldots \\
Z & a_{m1} & \ldots & a_{mn}
\end{bmatrix}
\]

The new matching framework proceeds with its two steps: **Step I: Create the distance matrix using predetermined distance mechanism;** **Step II: Choose elements from this distance matrix and make groups.** Step II of the new matching framework consists of two sub-steps: 1) selecting units according to a certain criterion and 2) combining units into groups (or strata), possibly then employing different weights for the treatment and control elements within each group (or stratum). If we only consider 1:1 matching, each chosen cell from the distance matrix forms one stratum so that the corresponding treatment and control elements within such stratum will all have weights equal to 1. If we consider 1: \(n_i\) (where \(n_i\) generally varies across \(i\)) matching, the units chosen from the same row form one stratum and the treatment element (here only
one element) has weight equal to 1 and all elements from the control group have weights equal to 1/ni (Ming and Rosenbaum 2001). If we consider m_i:n_i matching (general stratification), each group (stratum) is formed using some band (e.g. quantile-based) so that each treatment element within this stratum has weight equal to 1/m_i and each control element within such stratum has weight equal to 1/ni.

In the narrow sense, matching usually refers to 1:1 or 1:n_i matching where we only concentrate on selecting entries from distance matrix (i.e. sub-step 1 of step II of the matching framework). Hereafter we will call this kind of matching process as “matching” to simplify the words and focus on it at this moment (we will call m_i:n_i matching as “stratification” hereafter). Given any distance matrix (m x n which means m treatment units and n control units), by the multiplication rule, we have at most 2^m\times n different selection processes from this distance matrix. We can define each of these 2^m\times n different selecting processes as one matching method so that one selecting process corresponds to one matching method (1:1). On the other side, each well-known matching method (of step II) also can build its corresponding selecting process upon this m x n (m \leq n) distance matrix. Here we create this 1:1 correspondence between a selection process and a matching method because we want to borrow strength (and also trigger intuition) based on the selecting process. Using this idea, it’s easy to interpret any known or unknown matching methods and easy to explore various matching properties.

First, we can use this new idea to interpret some well-known matching methods such as NN, optimal matching:

- NN: a selection process whose minimum is the smallest of all possible selecting processes from the same distance matrix;
• optimal: a selection process whose total distance is the smallest of all possible selection processes from the same distance matrix;

Second, we can use this new idea to interpret 1: $n_i$ matching and matching with/out replacement:

• $1: n_i$ match: select at most $n_i$ columns for row $i$;
• With replacement (at most $n_j$ times): select at most $n_j$ rows for column $j$;

Each specific matching method corresponds to one special selection process for this matrix and it can be easily and accurately defined. For example, NN (1:1, w/o replacement) can be described as one picking-up from the distance matrix with its minimum the smallest of all possible selection processes from the same matrix given that each row and column can be chosen at most once and each row at least once (if possible).

This new framework of matching can also be used to explore existing matching methods from a new aspect. For example, now we can interpret the difficulty in applying “multiple covariate matching” with high dimensional data set (Cochran and Rubin 1973) in much more articulate wording: Suppose we have only 3 covariates for which we aim to match using “multiple covariate matching”. This means that we should match separately on each covariate to generate the final matched pairs, so that the corresponding distance matrix for each of the covariates and the overall distance matrix can be expressed as below:

<table>
<thead>
<tr>
<th>Similarity on 1st cov</th>
<th>Similarity on 2nd cov</th>
<th>Similarity on 3rd cov</th>
<th>Overall similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 1 0 1</td>
<td>A 0 1 1</td>
<td>A 1 0 0</td>
<td>A 0 0 0</td>
</tr>
<tr>
<td>B 1 1 0</td>
<td>B 1 0 1</td>
<td>B 0 0 1</td>
<td>B 0 0 0</td>
</tr>
<tr>
<td>C 0 0 1</td>
<td>C 0 0 0</td>
<td>C 1 1 1</td>
<td>C 0 0 0</td>
</tr>
</tbody>
</table>
Here in each matrix, 1 stands for the “exact match” on one particular categorical variable (or “within the bin” for one particular continuous variable) between the treatment and control units and the overall distance equal to the production of all corresponding cells from all three distance matrices. With more than just a few covariates it becomes very difficult to find matches with close or exact values of all covariates (for this example, we can’t find any exact matched pair even for only 3 covariates). From this example, we can see that even though sometimes we consider more than 1 covariate in matching, technically we convert the multiple matrices (or array) into one distance matrix (for this example, we apply the Boolean operation (“all”) for each cell across all matrices). The interpretation of this “multiple covariate matching” challenge using the new matching idea is now much clearer and in a more formal format (in mathematical sense), ordinarily used in the past.

1.4 Introduction to “Outlier First Matching” (OFM)

1.4.1 What is OFM?

When we look closely at different selection processes (different matching methods) from a *distance matrix* in step II of SMF, our interest is triggered for creating another matching method that uses the distances to find the matches in unconventional way. We call it as “Outlier First Matching” (OFM). It pertains to a **selection process whose maximum distance yields the smallest maximum value of all possible selections from the same distance matrix**;

One example below will show the different results using different selecting methods (matching methods) upon different distance matrices.
Example 1: Let \( z \) denote a bivariate random vector of dimension 2 for subjects in two groups. We will denote \( Z_{ik} = (x_{ik}, y_{ik})^T \) the bivariate observations token for unit \( i \) in group \( k \), where \( i=1,2, k=0,1 \) (\( k=1 \) for treatment group and \( k=0 \) for control group). Figure 2 denotes 4 observations of \( Z \) (points A, B in the treatment group, and points C, D in the control group). Distances matrices 1 and 2 in Figure 3 are formed on Manhattan and Euclidean distance measures. Table A and B (inside Figure 3) presents side by side matching results using different matching algorithms. We see that the combination of step I and step II of SMF (page 12) will generate different matching results. Here NN matching searches for the units with smallest minimum distance; OFM searches for the units with smallest maximum distance; while OPM searches for the units with the smallest total distance.

1.4.2 “Non Compatible Pattern Matrix” and “Determination inequality”

Let \( y_k = (y_k)^T = (y_0, y_1)^T \) denotes a random vector of dimension \( t \) for subjects in two groups (\( k=0 \)-Control group; \( 1 \)-Treatment group). In other words, we observe subject’s \( t \) characteristics in treatment and control group. Then \( Y_k \) is a data matrix of dimension \( n_k \times t \), presenting multivariate observations for \( n_k \) subjects in group \( k \) representing treatment (\( k=1 \)) and control group (\( k=0 \)).

Let’s define \( d_{lm} \) as the distance between unit \( l \) (from treatment group) and \( m \) (from control group) using some distance measure; where \( l=1, \ldots, n_1; m=1,\ldots, n_0 \). The distance matrix can be expressed as

\[
D = \begin{bmatrix}
  d_{11} & d_{12} & \ldots & d_{1n_0} \\
  d_{21} & d_{22} & \ldots & d_{2n_0} \\
  \vdots & \vdots & \ddots & \vdots \\
  d_{n_11} & d_{n_12} & \ldots & d_{n_1n_0}
\end{bmatrix}
\]
If unit l cannot be matched to unit m, the corresponding entry will be indicated as “NC” (Non compatible). Let’s use “0” to stand for compatible value and “1” for non-compatible value, and define the corresponding matrix describing compatibility as “Non Compatible Pattern Matrix (NCPM)”. The NCPM can be re-arranged as:

\[
\begin{bmatrix}
1 & 1 & \ldots & 0 \\
1 & 1 & \ldots & \vdots \\
\vdots & & & \ddots \\
0 & 0 & \ldots & 0
\end{bmatrix}
\]

and can be partitioned into four regions as below:

\[
\begin{bmatrix}
I & \vdots & II \\
\vdots & \ddots & \vdots \\
III & \vdots & IV
\end{bmatrix}
\]

and we assume the elements in region I are all “NC” and call region I as “rectangular non-compatible pattern (RNCP)”. Let’s define the following terms to facilitate our interpretation:

m = the number of rows of the inspected matrix (i.e. # of rows of NCPM);

n = the number of columns of the inspected matrix (i.e. # of columns of NCPM);

p = the number of rows from RNCP (i.e. # of rows of RNCP);

q = the number of columns from RNCP (i.e. # of columns of RNCP);

n_int = the number of matched pairs the analyst is interested in.

Region III can provide at most (m-p) matched pairs (# of rows of part III) and region II can provide at most (n-q) matched pairs (# of columns of part II). If the sum of these two numbers is less than n_int, we will end up the situation of “not enough matched pairs (NEMP)”. We summarize our above illustration into one inequality and call it “determination inequality (DI)” (the detail proof of DI can be found in Appendix I: Proof of DI).

\[
\text{DI: } n_{\text{int}} = (<, >) \min \{(m-p)+(n-q), m, n\}
\]
If left side is larger than the right side, there exists the problem of NEMP. When DI holds equality (and assuming \( n_{\text{int}} \leq \min(m, n) \) ), it means the original matrix can hold exactly \( n_{\text{int}} = (m - p) + (n - q) \) matched pairs. Let’s define “order” as the maximum number of 1:1 non-replacement matched pairs the original matrix can support. Among all transformed matrix having the same “order” as the original matrix, the one having the smallest number of entries is defined as “smallest maximum order matrix” (SMOM); such matrix has the property that it meets the equality of above “DI”. How does the above process of checking NCPM facilitate the pursuing of OFM? Recall the goal of OFM is to find one matching result whose maximum value is not larger than the one from any other potential candidates. If the original matrix can be transformed to SMOM, its maximum entry will be the smallest maximum value among maximum ones from all potential matching results. Now the key point is how to get this SMOM. If some “threshold” can be set and used to transform all entries above this “threshold” to non-compatible value and the transformed matrix have the property of SMOM (i.e. the equality holds for “DI” with the smallest entries as possible), the result of OFM can be obtained.

1.4.3 OFM algorithm

Let \( y = (y_k)^T = (y_0, y_1)^T \) denotes a random vector of dimension \( t \) for subjects in two groups (k= 0-Control group; 1-Treatment group). In other words, we observe subject’s \( t \) characteristics in treatment and control group. Then \( Y_k \) is a data matrix of dimension \( n_k \times t \), presenting multivariate observations for \( n_k \) subjects in group k representing treatment (k=1) and control group (k=0).

**Step 1: Creating “distance matrix”**
Let’s define $d_{lm}$ as the distance between unit l (from treatment group) and m (from control group) using some distance measure; where $l=1, \ldots, n_1$; $m=1,\ldots, n_0$. The distance matrix can be expressed as

$$D =
\begin{bmatrix}
   d_{11} & d_{12} & \cdots & d_{1n_0} \\
   d_{21} & d_{22} & \cdots & d_{2n_0} \\
   \vdots & \vdots & \ddots & \vdots \\
   d_{n_11} & d_{n_12} & \cdots & d_{n_1n_0}
\end{bmatrix}.$$ 

**Step 2: “Simplifying”**

Order each row of the original distance matrix and pick the first “n_int” ordered (ascending) units for each row. Do the similar job for column. If any entry from distance matrix has not been selected in above two series, replace the value of this entry with “NC”. Then delete the rows and column where all entries inside all “NC” to simplify the original matrix.

**Step 3: “Beyond threshold value transformation”**

Then let’s define “ordered $i^{th}$ minimum”:

“Ordered $i^{th}$ row minimum”: First pick up the $i^{th}$ minimum for each row and then sort them ascending (the non-compatible value will be treated as “Inf” and put into the end of the sorted series).

“Ordered $i^{th}$ column minimum”: First pick up the $i^{th}$ minimum for each column and then sort them ascending (the non-compatible value will be treated as “Inf” and put into the end of the sorted series).

Order the original matrix by each row using “Ordered $i^{th}$ row minimum”, then do the similar job by each column ($i=1, \ldots, m$ for row and $1,\ldots,n$ for column) to create two different “ordered minimum matrices” (one called “ordered row min matrix” and the other called “ordered col min matrix”).
To achieve OFM, for any $i^{th}$ minimum, only $[n_{int}-j: m]$ can be candidates for “threshold”, where $j$ is the number of columns occupied by $(i-1)$ iteration (if $i=1, j=0$) (see Appendix II: Proof of Criterion I for interpretation).

If at any iteration $i$, the above criteria cannot be satisfied (i.e., NEMP arises), the current matrix including $1$-$i^{th}$ minimum is the “simplified matrix” and the current “order” of simplified matrix is the “order” of the original matrix.

All candidates from different minimums are sorted ascending to form the candidate set. We treat the candidates one by one as “threshold” and replace all the entries larger than “threshold” with “NC”.

**Step 4: “Pop up outlier iteratively”**

We use the “Hungarian algorithm” to determine the equality of DI for each transformed distance matrix. If for any “threshold”, the transformed distance matrix supports the equality of DI (the “order” of the transformed matrix equal to “n_int”). Then we pop out such threshold as result and delete the corresponding row or column for such threshold. For the left matrix we do the step 3 and 4 using “n_int-1” as the new interested number of matched pairs.

**Step 5: Lump together the result or report “no support result”**

We continue step 3 and 4 until all “n_int” matched pairs have been selected or we have the problem of NEMP. We either report the lump-up result or report the message of “no support result”.

The above 5 steps can be summarized using the follow flowchart Figure 4.
1.5 Performance of OFM under SMF on cross-sectional study

1.5.1 Simulated data study

1.5.1.a Simulation of multivariate normal distribution

We simulate a data set (n=100) of covariates, treatment indicator variable and outcome variable in the following fashion:

\[
\begin{pmatrix}
    x_{1i} \\
    x_{2i} \\
    x_{3i} \\
    x_{4i}
\end{pmatrix}
\sim MVN\left( \begin{pmatrix}
-1 \\
0 \\
0.5 \\
0.5
\end{pmatrix}, \begin{pmatrix}
    1 & 0.15 & 0.60 & 0 \\
    0.15 & 1 & 0.17 & 0.3 \\
    0.60 & 0.17 & 1 & 0 \\
    0 & 0.30 & 0 & 1
\end{pmatrix} \right)
\]

\[T_i \sim Bin(P(T_i = 1))\]

where

\[
\logit(P(T_i = 1)) = x_{1i} - 0.5x_{2i} + 0.5x_{3i};
\]

\[O_i = \gamma_1 T_i + 2x_{1i} - x_{3i} + 5x_{4i} + \varepsilon,\]

where \(\gamma_1 = 1, i = 1, \ldots, 100\) and \(\varepsilon \sim N(0,1)\)

Note that the simulation of \(O\) ignores \(x_2\) and incorporates \(x_4\) because the variables determining outcome may be not exactly the same variables determining the treatment.

At each iteration of simulation, two methods are used to estimate the treatment effect:

One is the un-adjusted crude mean difference of two treatment groups (method I) and the other is model adjusted regression estimated treatment effect (method II). Without losing generality, we assume two scenarios for constructing treatment effect estimation working models:

Correct treatment effect estimation working model:

\[O_i = \gamma_1 T_i + \gamma_2 x_{4i} + \gamma_3 x_{3i} + \gamma_4 x_{4i} \text{ (Model 1)}.\]

Incorrect treatment effect estimation working model where \(x_4\) is accidently replaced with \(x_2\) (\(cor(x_2, x_4) = 0.3\)):

\[O_i = \overline{\gamma}_1 T_i + \overline{\gamma}_2 x_{4i} + \overline{\gamma}_3 x_{2i} + \overline{\gamma}_4 x_{3i} \text{ (Model 2)}.\]
The simulation iterated 1000 times and on average led to 34 clients in treatment group \((n_T=34, 34\%)\) and 66 clients in control group \((n_c=66, 66\%)\). The treatment effect estimation with and without matching using two different methods (un-adjusted and model unadjusted) are summarized in Table 2 and Table 3 separately for model 1 and model 2.

If the covariates follow multivariate normal distribution, w/o matching and matching generate the similar regression adjusted treatment effect estimations \((1.01 (0.54, 1.47) vs. 1.01 (0.51, 1.50) for w/o matching vs. matching)\) but matching generates better crude mean difference treatment effect estimation \((1.49 (-0.69, 3.65) vs. 1.24 (-1.16, 3.61) for w/o matching vs. matching)\). Furthermore, under the circumstance of correct model (model 1), OFM (without caliper) obtains the same estimation of treatment effect \((\gamma_1=1.01, 95\% CI = (0.51, 1.50))\), same percentage bias \((1.07\%)\), same CR \((95.1\%)\), and same RMSE \((0.25)\) as other 1:1, non-replacement competing matching algorithms (NN, OPM). Under the circumstance of incorrect model (model 2), once again, OFM (without caliper) obtains the same estimation of treatment effect \((\gamma_1=0.98, 95\% CI = (-1.61, 3.45))\), same percentage bias \((1.52\%)\), same CR \((93.0\%)\), and same RMSE \((1.26)\) as other 1:1, non-replacement competing matching algorithms (NN, OPM).

1.5.1.b Simulation of multivariate skew-normal distribution

As noted in 1.2.1, under the circumstances of C1-C6, PSA method has the advantage to traditional regression methods. To test the above hypothesis, we simulate a skew date set \((n=100)\) of covariates, treatment indicator variable and outcome variable in the following fashion:
A k-dimensional random variable $Z$ is said to have a multivariate skew-normal distribution (MVSN) if it is continuous with density function

$$2\phi_K(Z; \Omega)\Phi(\alpha^T z), \quad z \in R^k$$

where $\phi_K(Z; \Omega)$ is the k-dimensional normal density with zero mean and correlation matrix $\Omega$, $\Phi(\cdot)$ is the N(0, 1) distribution function, and $\alpha$ is a k-dimensional vector. For simplicity, $\Omega$ is assumed to be of full rank. The meaning of parameter $\Omega$, $\alpha$ and background information of multivariate skew-normal (MVSN) distribution can be found in Azzalini, A. and Dalla Valle 1996 Biometrika paper (Azzalini and Dalla Valle 1996).

$$T_i \sim \text{Bin}(P(T_i = 1))$$

where

$$\logit(P(T_i = 1)) = x_{1i} - 0.5x_{2i} + 0.5x_{3i};$$

$$O_i = \gamma_1 T_i + 2x_{1i} - x_{3i} + 5x_{4i} + \epsilon,$$

where $\gamma_1 = 1$, $i = 1, \cdots, 100$ and $\epsilon \sim N(0,1)$.

Note that the simulation of $O$ ignores $x_2$ and incorporates $x_4$ because the variables determining outcome may be not exactly the same variables determining the treatment.

We use the same methods as 1.5.1.a and the following models to estimate the treatment effect.

**Correct treatment effect estimation working model:**

$$O_i = \gamma_1 T_i + \gamma_2 x_{1i} + \gamma_3 x_{3i} + \gamma_4 x_{4i} \quad \text{(Model 3)}.$$

**Incorrect treatment effect estimation working model where $x_1$ is accidently replaced with $x_2$ ($\text{cor}(x_1, x_2) = 0.15)$:**
\[ O_i = \tilde{\gamma}_1 T_i + \tilde{\gamma}_2 x_{2i} + \tilde{\gamma}_3 x_{3i} + \tilde{\gamma}_4 x_{4i} \] (Model 4).

The simulation iterated 1000 times and on average led to 30 clients in treatment group \((n_T = 30, 30\%)\) and 70 clients in control group \((n_c = 70, 70\%)\). The treatment effect estimation with and without matching using two different methods (un-adjusted and model unadjusted) are summarized in Table 4 and Table 5 separately for model 3 and model 4.

If the covariates follow multivariate skew-normal distribution, using method I (un-adjusted crude difference of mean), matching will generate better treatment effect estimation compared to w/o matching methods \((2.19 (-0.07, 4.58) \text{ vs. } 1.24 (-1.16, 3.61))\) for w/o matching vs. matching). If the treatment effect estimating model is correct (e.g. model 3), the treatment effect estimation through method II (model adjusted treatment effect estimation) will be similar to each other for with and w/o matching algorithms \((0.99 (0.46, 1.49) \text{ vs. } 0.99 (0.41, 1.54))\) for w/o matching vs. matching). However, if the treatment effect estimating model is incorrect (e.g. model 4), matching will achieve the better model based adjusted treatment effect estimation compared to w/o matching method \((1.72 (0.94, 2.48) \text{ vs. } 1.20 (0.43, 1.94))\) for w/o matching vs. matching).

Furthermore, under the circumstance of correct model (model 3), OFM (without caliper) obtains the same estimation of treatment effect \((\hat{\gamma}_1 = 0.99, 95\% \text{ CI } = (0.41, 1.54))\), same percentage bias \((0.73\%)\), same CR \((95.8\%)\), and same RMSE \((0.29)\) as other 1:1, non-replacement competing matching algorithms (NN, OPM). Under the circumstance of incorrect model (model 4), once again, OFM (without caliper) obtains the same estimation of treatment effect \((\hat{\gamma}_1 = 1.20, 95\% \text{ CI } = (0.43, 1.94))\), same percentage bias
(19.51), same CR (94.9%), and same RMSE (0.43) as other 1:1, non-replacement competing matching algorithms (NN, OPM).

In summary from these two simulations, first, matching will produce more convincible crude mean difference estimation of treatment effect (methods I) compared to non-matching methods. And if the inherent distributions of covariates do not follow MVN and/or any one of C1-C6 conditions is met, matching will provide more convincible model-based regression estimated treatment effect. Furthermore, if we concentrating on evaluating the performances of different matching algorithms, comparing to other competitive matching algorithms, OFM provides the same treatment effect estimations as other 1:1, non-replacement matching algorithms both in correct and in incorrect treatment effect estimation working models for cross sectional PSA study where the p-score is generated using the logistic function. Even though the selected units from three matching algorithms are the same, the matched pairs are different from each other. Among them, OFM generates the least skew matched pair (i.e. the largest distance difference from matched pairs is the smallest).

1.5.2 Application of “Lalonde” Data
Lalonde (1986) evaluated the performance of non-experimental estimators using experimental data as a benchmark. At first, Lalonde used experimental data called “National Supported Work (NSW) Experiment” carried out in the mid-1970’s to evaluate the impact of job training on the earnings improvement of low-skilled workers. In NSW experiment, A cohort of 6616 workers were randomly assigned to control and treatment groups, after 9 ~18 months of job training (for treatment group), their earnings in 1978
were compared to evaluated the impact of job training on the earnings improvement
(Treatment = Job training; Outcome = Earning in 1978).
Randomization ensures that not only the observed potential “confounders” but also the
unobserved ones are unrelated to the treatments so that they contribute minimal (if any)
effect to the outcomes. Therefore, the result from this experiment could be used as
benchmark to consistently identify the causal effect of receiving job training.
In non-experimental study, Lalonde kept the treatment cases from the NSW experiment
(those who participated in the job training program) and threw out the real experimental
controls. He constructed a new set of controls from the “Current Population Survey”
(CPS) and the “Panel Study on Income Dynamics” (PSID) data, which are commonly
used data for economists and sociologists. He then fit a variety of regression models to
estimate the effect of job training and compared his results to the initial randomized
experiment (NSW experiment). Two sets of analyses led to conflicting results — they
often had the wrong sign and were two to three times the size of the real effect. Such
confliction showed that in observational studies, simple regression without confounder
adjustment can lead to incorrect inference on treatment effect. A decade later two more
economists, Dehejia and Wahba (1999) used the “counterfactual causality and matching
methods” to re-analyze Lalonde’s non-experimental data and reported that for men the
treatment effect was close to what Lalonde found in NSW experiment (see result in Table
6)
In this paper the “Lalonde” non-experimental data set from R package “MatchIt” will be
used for illustration. This data set (n=614) is one of six data sets used by Lalonde (1986)
and Dehejia and Wahba (1999). Income in 1978 (one or two years after the job training)
is the outcome variable; and the treatment variable is whether the individual received job training or not. The data contain a number of demographic and work-related variables designed to control for other factors that might affect outcome (All variables are list in Table 7). PSA analysis is used here to minimize the confounder effect in estimating the treatment effect on outcome. Our ultimate interest is to study the estimations of treatment effect under different matching methods in phase II of PSA. Here, we estimate propensity score (i.e. p-score) using one simple logistic regression as:

$$\logit(P(T = 1)) = \alpha + \beta_1 * age + \beta_2 * educ + \beta_3 * black + \beta_4 * hisp + \beta_5 * married $$

$$+ \beta_6 * nodegree + \beta_7 * re74 + \beta_8 * re75$$

Where $\alpha, \beta_1 \sim \beta_8$ are parameters.

The back-to-back histogram of “p_score” before matching can be found in top left in Figure 5.

In the phase II of PSA, different matching methods (NN, OPM, radius and OFM matching) are applied to generate matched-pairs. We choose 1:1, non-replacement criteria for NN, OPM and OFM matching. For radius matching, 0.25 s.e. of the treatment “p_score” is used to generate the radius so that one treatment can be matched to multiple control elements and one control element can be matched more than once (i.e. 1: n with replacement). To simple the calculation, the treatment element in each matched pair is assigned the weight equal to one and the control element in each matched pair is assigned the weight equal to (1/number of total control elements in this matched pair). In the other words, the difference between the average of treatment and control elements in each matched pair is treated as the treatment effect for such matched pair and the average of those difference across all matched pairs is treated as the estimation of ATT.
The results coming from different matching algorithms are listed in Table 8. As we have seen in cross-sectional simulation study, NN and OFM generate the same model adjusted treatment estimations which are the better estimators compared to both the crude outcome mean difference (-9756) and the mis-specified model adjusted treatment effect (-886) without matching. The similarity between NN and OPM in cross-section study coincide the precedent study (Gu and Rosenbaum 1993). Such coincidence may suggest that in cross-section study we can choose any one (from NN, OPM, & OFM) to estimate the average treatment effect if we don’t consider the matched pairs in detail. Nevertheless, if we look at the detail of each matched pair, the situation will be different. As Figure 7 suggests, OFM achieve the least skew matched pair with the biggest “p-score” difference close to 0.5 compared to the ones from OPM and NN both around 0.8. Furthermore, OFM within the caliper report treatment effect estimation as -1753 with 95% CI (-2221, 283) which is very close to the model adjusted value -1672 (-2920, -423) from experimental data. And the back-to-back histogram (Figure 6) also suggests that OFM within the caliper achieve the better similarity of p-scores between treatment and control groups compared to any back-to-back histogram from matching algorithm without caliper. In a word, OFM with caliper may be a good candidate to estimate the treatment effect if we have relatively large sample size to create sufficient matched pairs.

1.6 Discussion
The simulations and real data analysis show some interesting properties of OFM compared to other matching methods under cross-sectional scenario: If the distance is Euclidean distance calculated from logistic model and fixed across the study (e.g. in cross-section study), NN, OPM and OFM all tend to select the same treatment and
control units. However, the matched pairs will be different from each other. Among them, OFM will generate the least skew matched pair whose largest measure of imbalance (in the sense of “p-score distance” generated in step I of our new matching framework (SMF) between treatment and control units) is the smallest. One of particular cases mentioned before which coincides our finding is the statement from Gu and Rosenbaum (1993), “...optimal matching picks about the same controls [as greedy matching (i.e. NN)] but does a better job of assigning them to treatment units.”

Due to the above finding, under the cross-sectional scenario, if the research interest is to study the treatment effect between treatment and control groups, we would suggest to use any one of above matching algorithms (NN, OPM, OFM). The only concern here is the calculating speed. However, if specific subgroup will be the secondary or even primary research interest, more appropriate matching algorithm will be chosen according to the research interest (e.g. if our research interest is to match the most costly clients first who are usually the most difficult to be matched to, OFM may turn to be the most appropriate matching method).
2 Applying OFM under SMF for longitudinal data

Abstract
Outlier First Matching (OFM) will be extended to longitudinal data where the entries into treatment and correlation between observations complex the study. Such longitudinal data arise from observational study where randomization cannot be applied. Treatment subjects are enrolled into the treatment group at different time points and stay there until the end of the study. At each time point, control subjects have the potential to be matched to any treatment subjects who are enrolled at such time point. Generalized Estimating Equations (GEE) will be used to adjust the inner-subject variation given that between-subject variation will be ignored due to the construct of data. Variant Propensity scores (p-score) are estimated by GEE model across the timeline and used for matching. Different matching algorithms will be used in the second phase of the two-stage longitudinal propensity score analysis (LPSA) to generate matched pairs. Treatment effect estimated using OFM will be compared to ones estimated using other well-known matching algorithms under the new matching framework “SMF”. The simulation and real data analysis suggests that OFM may provide more convincible treatment effect estimations compared to other matching algorithms in LPSA.

2.1 Introduction
In the past three decades, with the development of the computer and other new technology, more and more longitudinal data are collected in an effort to draw causal links between interventions and endpoints. Among these longitudinal data, a big proportion of data are prospective or retrospective longitudinal data sets where subjects are not initially randomized to the control and treatment conditions. Longitudinal studies
of the causal effect have widely spread on the regions of economics, epidemiology, sociology, and so on (Cochran 1968, Heckman 1979, Rubin 1974, 1977, Rosenbaum, and Rubin 1983, Angriest et al. 1996). There are some fundamental challenges for longitudinal observational causal effect studies: First, similar to the cross-sectional study, there exists selection bias between treatment and control subjects. However, since there are time variant and invariant covariates, the heterogeneity among the covariates will be across different time points. Second, a longitudinal observation study will have repeated measures on many of the subjects. In other words, there may exist different level of cluster (and corresponding correlation) and/or random effect within the data. Third, unlike randomized controlled clinical trials, the duration of treatment is not standardized in a research protocol but instead determined by the decision of the clinical and patients. In the other words, the enrollment time and duration of staying in the treatment maybe vary from patient to patient. Four, to add more complexity, even the treatment status can be switched during the study process. Five, there may exist the stopping of medication, quitting of clients, terminating the study issues (the censor problem). We will only address the first three cases and leave other issues out of our study for simplicity. Since the differences in background characteristics of subjects persist across the time, to draw the casual inference from the observation longitudinal data without the randomized allocation into treatment group, people need to use tools to adjust such differences’ impacts on the estimation of treatment effect for longitudinal analysis. Similar to the cross-sectional study, if heterogeneity exists among the covariates between treatment and control subjects, the traditional regression approaches cannot compensate the impact coming from such heterogeneity on the estimation of treatment effect (Lissner, Skoog et
al. 2003; van den Berg, van der Velden et al. 2007; Wolke, Waylen et al. 2009). In other words, selection bias will confound the estimation of treatment effect for heterogeneous observational longitudinal data. Using the similar idea of cross-sectional propensity score analysis, if we can achieve the balance across the covariates at the baseline points, and adjust the time variant covariates’ impact on the outcome, we can achieve the convincible estimation treatment effect across the duration of treatment period. The idea mentioned above attributes to Longitudinal Propensity Score Analysis (LPSA) (Austin 2008; Achy-Brou, Frangakis et al. 2010; Achy-Brou and Frangakis 2010). Under LPSA study framework, the treatment subjects can be enrolled into treatment group at different time but will stay in the treatment group for a while. After his disenrollment, the subject will be never enrollment again (If the subject can vary his enrollment status at different time points, the complexity will lead to Longitudinal Marginal Structure Model Analysis (LMSMA) (Faries and Kadziola 2010) (Bodnar, Davidian et al. 2004) (Mortimer, Neugebauer et al. 2005) or Structural Nested Model Analysis (SNMA) (Bray, Almirall et al. 2006; Almirall, Coffman et al. 2010; Almirall, Ten Have et al. 2010)). In this paper, we address the issue of adjusting the heterogeneity at baseline by extending our method of matching based on the OFM to observational non-randomized longitudinal data through a two phase LPSA framework.

2.2 Methodology
Two phases LPSA strategy incorporates a model of propensity for treatment (in phase I) and a model of treatment effectiveness (in phase II). In between phase I and II, matching methods will be used to create matched pairs based on p-score from phase I. Then
matched subjects will be used in phase II of LPSA to generate the estimation of treatment
effect.

In phase I of the LPSA, the propensity model examines the repeated observations of
binary treatment over time. The model can include multiple treatment intervals per
subject over time and variation in both between-subject propensity for treatment and
within-subject treatment over the course of the study. Among a wide variety of models
used in phase I of LPSA, Mixed Effect Model (MEM) (Laird and Ware 1982) and
are two commonly used ones. Even though both of methods use the similar approach to
adjust the within-subject correlation, the big difference between them lies on how they
construct between – subject variations. MEM approach relates the between-subject
variant to two sources: the inherent characteristics of the random effect of individual
which may be not observed from the covariates, and the invariant and variant covariates.
However, GEE model only addresses the between - subject variant from the observable
variant and invariant covariates, but not inherent characteristics (random effect).

From this paper prospective, we will choose GEE model to estimate the propensity score
due to the following reason: 1) we generate the p-score from a wide variety of
demographic and social economical covariates, the impact from the inherent
characteristics is relatively smaller compared to the covariates involved. 2) We are more
concerned about the estimation of treatment effect for the whole population instead of the
one for individual subject. 3) We will concentrate on the balance of the covariates
included on the baseline time point so that we do not want the impact from inherent
characteristics (random effect) contaminate the balance of those covariates. 4) Unlike
MEM model that uses “full-likelihood” method which is more model dependent, GEE use “partial-likelihood” method (Hedeker and Gibbons 2006) which can provide more robust and less model dependent estimation. In this scenario, GEE model play better than mixed model to generate a better balanced (in the sense of controlled variables) matched pairs and more robust (less model dependent) p-score for next step population based treatment effect estimation.

GEE framework accounts for the correlated recurrence times that represent the successive within-subject treatment intervals. A GEE logistic regression model examines treatment as a function of these characteristics, whether time-invariant or time-varying:

Conditional expectation or mean of each response: \( E(T_{ij}|X_{ij}) = \mu_{ij} \) is assumed to depend on the covariates through a known linking function:

\[
\logit(\mu_{ij}) = x_{ij}'\beta; \quad \text{where}
\]

\( \mu_{ij} = \text{marginal response} = E(T_{ij}|X_{ij}) \)

\( T_{ij} \) is the dummy variable standing for enrollment or not for subject i at time point j (i.e. recruited into treatment for client i at time point j or not)

\( X_{ij} \) is a \((P \times 1)\) vector of all controlled covariates (e.g. treatment history, social/demographic status, etc.)

\( \beta \) is a \((P \times 1)\) vector of unknown regression coefficients (average population effect)

The conditional variance of each \( T_{ij} \), given the covariates, is assumed to depend on the mean according to

\[
\text{Var}(T_{ij}|X_{ij}) = \phi \nu(\mu_{ij}), \quad \text{where}
\]

\( \phi \) is a scalar parameter that may be known or estimated

\( \nu(\mu_{ij}) \) is a known “variance function”.

35
Generalized Estimating Equation:

\[ U(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_{ij}}{\partial \beta_k} V_i^{-1} \{ T_i - \mu_i(\beta) \} \],

where

\( V_i \) (or \( V_{ij} \)) is a \((J \times J)\) variance matrix standing for correlation of dependent variables.

\( V_i = \phi A_i^{1/2} R(\alpha) A_i^{1/2} \), where \( \phi \) is a glm dispersion (scale) parameter, \( A \) is a diagonal matrix of variance functions, and \( R(\alpha) \) is the working correlation matrix of \( Y \) (independence, exchangeable, autoregressive, stationary m-dependent, and unstructured).

Solve \( U(\beta) = 0 \) using “Newton-Raphson” algorithm (semi-parametric) to achieve parameter estimators \( \hat{\beta} \).

The estimated propensity score (p-score) can be expressed using the logistic response function for subject \( i \) at time \( j \) as

\[ \text{P-score} = P(T_{ij} = 1 | x_{ij}, V_i) = \frac{\exp(x_{ij}' \hat{\beta})}{1 + \exp(x_{ij}' \hat{\beta})}. \]

After achieving p-score, we will use several matching algorithms to create matched pairs.

Similar to what we have chosen in cross sectional study, we choose OFM, NN, OPM, RDM matching without caliper, and one application of OFM with caliper. Under longitudinal data framework, at each time point, control subject has the potential to be matched to treatment subject who are enrolled in that time point. Concentrating on the distance matrix, for each treatment subject, the distances between such treatment subject and all control subjects will create the entries of one row for such treatment subject. Since all treatment subjects are enrolled into treatment at different time points, for any given column, the entries of this column stand for the p-score difference for this control subject and all treatment subjects at their different enrollment time points. Using the idea of
trajectory, we try to match the treatment and control subjects at the enrollment time point of treatment subject based on their traces from the first time point until the enrollment time points.

In the phase II of PSA, Gee model is used to estimate the treatment effect. For treatment subject, the follow-up period will be the time from the start of each treatment until the disenrollment from treatment. For control subject, the follow-up period will be enrollment period of the corresponding matched treatment subject. Based on the ignorability assumption, causal inference can be drawn regarding the estimation of treatment effect from GEE model. A GEE generalized normal regression model examines treatment as a function of these characteristics, whether time-invariant or time-varying:

Linking function:

\[ O_{ij} = \beta \times x_{ij}' + \gamma_1 \times T_{ij} \times \text{Time}_{ij}; \]

model is created at the cluster of subject, where

- \( O_{ij} \) is the outcome for subject \( i \) at time point \( j \) (i.e. recruited into treatment for client \( i \) at time point \( j \) or not)
- \( x_{ij} \) is a \((P \times 1)\) vector of all controlled covariates (e.g. treatment history, social/demographic status, etc.)
- \( T_{ij} \) is the dummy variable standing for enrollment or not for subject \( i \) at time point \( j \) (i.e. recruited into treatment for client \( i \) at time point \( j \) or not)
- \( \text{Time}_{ij} \) stands for the time period of enrollment after the beginning of treatment until time \( j \) for subject \( i \) for treatment subject (or the corresponding matched treatment subject’s enrollment period for control subject).
\( \beta \) is a \((P \times 1)\) vector of unknown regression coefficients for controlled covariates.

\( \gamma_1 \) is the coefficient standing for treatment effect increment per time point from baseline.

We use “exchangeable” working correlation matrix to construct the within-subject correlation to achieve the robust parameter estimation of \( \hat{\gamma}_1 \) using “Newton-Raphson” algorithm. And \( \hat{\gamma}_1 \) will be interpreted as the estimations of the treatment effect increment per time point from baseline point. The treatment effect estimations will be compared side by side, and the interpretation and suggestions will be given after simulating the above application many times.

2.3 Performance of OFM under SMF on longitudinal study

2.3.1 Simulated data study

Let’s simulate the longitudinal data with four time points for two different populations (0 - control; 1 – treatment). Each client has four observations and has the potential to be enrolled into treatment on 1\(^{st}\), 2\(^{nd}\), or 3\(^{rd}\) time points. And if the client is enrolled into treatment, he (she) will stay in the treatment group until the last observation (4\(^{th}\) time point). Since each treatment client is enrolled into treatment at different time point, his (her) enrollment and follow-up period after the treatment may be different from others.

To obtain the treatment effect across the time, we would be better to compare the treatment to control clients based on their treatment histories and enrollment times. Consequently, we would like to match the clients across the time to ensure the matched treatment and control clients are similar to each other at the time point when the treatment client is enrolled into the treatment.
We simulate a data set \((n=150)\) with 4 time points \((t = 1, 2, 3, 4)\) of covariates, treatment indicator variable \((T_{it})\) and outcome variable \((O_{it})\) in the following fashion:

At each time point \(t\), there exist the following covariates:

\[
x_{1it} = -1 \quad (if \ t = 1, 3) \quad or \quad 1 \quad (if \ t = 2, 4).
\]

\(x_{1it}\) simulates the vibrate variable;

\[
x_{2it} = t \cdot x_{1it}
\]

simulates the time variable;

\[
\begin{pmatrix}
x_{3i} \\
x_{4i}
\end{pmatrix} \sim \text{MVN} \left( \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix} \right)
\]

\(T_{it} \sim \text{Bin}(P(T_{it} = 1))\) where

\[
\text{logit}(P(T_{it} = 1)) = x_{1it} - 0.5x_{2it} + 0.5x_{3it};
\]

Let’s define \(S_{it}\) as how long the client has stayed in the treatment group. The outcome variable \((O_{it})\) is simulated as:

\[
O_{it} = \gamma_1 T_{it} S_{it} + \gamma_2 x_{1it} + \gamma_3 x_{3it} + \gamma_4 x_{4it} + \epsilon_{it},
\]

\(where \ \gamma_1 = 1, \ \gamma_2 = 5, \ \gamma_3 = -1, \ \gamma_4 = 4\); \(t = 1, 2, 3, 4; \ i = 1, \ldots, 150\) and

\[
\epsilon_{it} \sim N \left( \begin{pmatrix} 0.2 \\ 0 \\ -0.3 \\ 0 \end{pmatrix}, \Omega_i \right)
\]

\(where \ \Omega_i = \text{Toeep}(0.3).\)

Note that the simulation of \(O\) ignores \(x_2\) and incorporates \(x_4\) because the variables determining outcome may be not exactly the same variables determining the treatment.

The within-subject covariance from the simulation is \(V(y_{i,ij,y_{i,j'}} | x_i) = \Omega_i : \text{Toeep}(0.3).\)

The simulation iterated 1000 times and on average led to 100 controls \((T=0)\) and 50 treatment subjects \((T=1)\) \(\text{(i.e. } n_c=100 \quad (67\% \text{) and } n_T=50 \quad (33\%)).\) Since different treatment subjects are enrolled into treatment group at different time points, the crude difference of
the outcome between the treatment subjects and control subjects is not a good estimation for treatment effect increment per time ($\gamma_1$).

Without losing generality, we assume two scenarios for constructing treatment effect estimation working models.

The correct treatment effect estimation working model:

$$O_{it} = \gamma_1 T_{it} S_{it} + \gamma_2 x_{1it} + \gamma_3 x_{3it} + \gamma_4 x_{4it} \text{ (Model 5)}.$$

The incorrect treatment effect estimation working model where $x_{1it}$ is accidently replaced with $x_{2it}$:

$$O_{it} = \gamma_1 T_{it} S_{it} + \gamma_2 x_{2it} + \gamma_3 x_{3it} + \gamma_4 x_{4it} \text{ (Model 6)}.$$

For working model 5, the estimation of treatment effect increment per time point ($\hat{\gamma}_1$) is 0.93 (95% CI = (0.81, 1.04)), which is close to the true parameter $\gamma_1 = 1$. From working model 6, the estimation of treatment effect increment per time point ($\hat{\gamma}_1$) is 9.18 (95% CI = (-9.68, 16.18)). The simple model adjusted treatment effect estimation based on misspecified model is clearly not good estimators for “treatment effect increment per time point”. Let’s apply PSA (through Longitudinal GEE model) together with different matching methods to estimate the above treatment effect for these two models. For model 5, the results are summarized in Table 9. For model 6, the results are summarized in Table 10.

Under the circumstance of correct model (model 5), OFM (without caliper) obtains the closest estimation of treatment increment per time ($\hat{\gamma}_1$=0.98, 95% CI = (0.84, 1.12) with the smallest percentage bias (2.27%), highest CR (92.9%), and the smallest RMSE (0.07) among all competing matching algorithms without sacrificing validity (i.e. without substantially increasing s.e.). Under the circumstance of incorrect model (model 6), once
again, OFM (without caliper) obtains the closest estimation of treatment increment per time ($\hat{\gamma}_1=1.15$, 95% CI = (0.94, 1.35) with the smallest percentage bias (14.56%), 2nd highest CR (68.4%), and the smallest RMSE (0.18) among all competing matching algorithms without sacrificing validity (i.e. without substantially increasing s.e.). In summary from the estimation from model 5 and model 6, comparing to other competitive matching algorithms, OFM provides the best (smallest percentage bias, highest CR, and smallest RMSE) treatment effect estimations without sacrificing validity (i.e. without substantially increasing s.e.) both in correct and in incorrect treatment effect estimation working models for longitudinal PSA study. Furthermore, as we have seen in cross-sectional PSA study, OFM also generate the least skew matched pair among all competitive 1:1, non-replacement matching algorithms.

2.3.2 Application of “Care Management” Data
The data set used for longitudinal analysis comes from one analysis about “Care Management (CM) Effect on Cost Analysis”. Clients who have met one of multiple criteria have been enrolled into one “management care system”. The aim of this analysis is to identify whether such “management care (MC) system” has any short-term and long-term cost saving or not and the patterns of those costs after enrollment between treatment (here “CM”) and control groups.

The data set contains a number of demographic, social economical and treatment variables which are list in Table 11. Since each treatment client is enrolled into “care system” at different time, the corresponding follow-up period for each treatment client (and its matched control unit) is also different from each other. Not only that, each control client can be matched to different treatment clients at different time points.
Considering the variations of treatment history, demographic and social economic status across the time, we create our data set longitudinally with 3 months interval. At each time point, every client (treatment or control) has one observation with his treatment history and demographic, social economic status calculated based on such time point. We use GEE model (clustered at each client) to create the propensity score (hereafter referred to as “p-score”) for each observation. Then we create “distance matrix” based on the distance (difference of p-score) between each treatment client and control client at treatment client admission period. After matching, we will compare the outcome variables (long term cost and short term cost) between treatment and control clients either directly or with model adjustment. Our research interest is to compare the performances of different matching methods under longitudinal circumstance. We will look closely on their overall p-score distributions, the absolute value of p-score difference distribution for each matched pair, the most unbalanced pair (in the sense of p-score), the balances of key variables (through effect size), and their different outcome results. The detail steps are illustrated below:

- Phase I of PSA analysis: Create p-score

As we mentioned above, each client has one observation at 3 month interval. To match the most similar “treatment” and “control” client, we’d better compare the treatment client to control client based on their treatment histories. The following GEE model is used to calculated the p-score for each observation (clustered at client).
logit(Treat\(_{i,j}\)) = \alpha + \beta_{1-4} \ast age(5\text{ categories}_i \ast j) + \beta_5 \ast \text{gender}_i \ast \text{black}_i + \beta_6 \ast \text{Hisp}_i + \beta_7 \\
\ast \text{married}_i \ast j + \beta_8 \ast \text{nondegree}_i \ast j + \beta_9 \ast \text{Homeless}_i \ast j + \beta_{10} \ast \text{Severe}_{mh\_i \ast j} \\
+ \beta_{11} \ast \text{Hepatitis}_i \ast j + \beta_{12} \ast \text{AIDS}_i \ast j + \beta_{13} \ast \text{gender}_i \ast \text{Alcohol}_i \ast j + \beta_{14} \\
\ast \text{Sub. Daily Usg}_i \ast j + \beta_{15} \ast \text{AIDS} + \beta_{16} \ast (\text{AOD cost 1yr}_i) + \beta_{17} \\
\ast (\text{AOD cost 1yr}_i)^2 + \beta_{18} \ast (\text{AOD cost 1yr}_i)^3 + \beta_{19} \\
\ast (\text{AOD cost 3 months}_i) + \beta_{20-23} \ast \text{count}_{detox}(3\text{ dummies}_i,j) + \beta_{24-26} \\
\ast \text{count}_{rehab}(2\text{ dummies}_i,j) + \beta_{27-28} \\
\ast \text{count}_{outpatient}(6\text{ dummies}_i,j), \text{ for all } i = 1, 2, \ldots, n

Where within-subject covariance matrix is \(V(y_{i,j}|x_i) = \Sigma\): Unstructured

GEE model concentrates on estimating the overall population effect of each covariate on treatment variable without classifying the individual effect. Therefore, there are some assumptions based on this model specification:

- The propensity of enrolling into treatment is related to the covariates listed in the model so that the history longer than we controlled (such as any episodes happened before last year) will be ignored;

- The inherent distinction across different clients will be ignored, each client is only different from the covariates we controlled into the model (Since our research object is to define the treatment effect between two populations, such coarse assumption is valid);

- Connected to above point 2, since we do not specify the individual character but only population effect, the estimated p-score from GEE model is consistent even with the possible mis-specification of the structure of covariance matrix;

After achieving p-score, we create “distance matrix” based on the distance (difference of p-score) between each “treatment” client and “control” client at “treatment” client’s
admission time point. Since the treatment history and p-score varies across the time, the same “control” client will have different properties (also different “p-scores”) based on the time points when he is matched to particular “treatment” clients. Here we use several different matching methods (selecting processes) to select and group treatment and control clients. To assure the comparison fair, (except radius matching) we choose 1:1 without replacement for all matching methods.

The balances of some key variables (in the sense of Effect Size of Cohen’s d (for continuous variable) or ODDS Ratio (OR) for dummy variable) are shown in Table 12. Among all different matching algorithms, OFM seems to achieve the best balances of those key variables (especially for cost variables). The back-on-back histogram of p-scores between and control groups before / after matching (from different matching algorithms) is shown in Figure 8. OFM generates the most symmetric “back on back histogram”, which suggests that OFM creates the most similar distributions of p-scores between treatment and control subjects. If we concentrating on the similarity of each matched pair, Figure 9 illustrates the distribution of absolute p-score differences among all matched pairs. For the most extreme case, the largest absolute p-score difference from OFM (around 0.005) is much less the ones from NN (around 0.0125) and OPM (around 0.0225). Table 13 shows the top 3 most unbalanced pairs (in the sense of biggest p-score difference among all matched pairs) from different matching algorithms. Comparing to other matching algorithms, OFM generate relatively less unbalanced matched pairs in the sense of differences of p-score and cost variables.

- Phase II of PSA study: Outcome analysis based on different matching processes
Before matching, at baseline time period (when the “treatment” clients are enrolled into system), the average of long term cost (1 year) and short term cost (3 months) are listed below categorized by different treatment group.

<table>
<thead>
<tr>
<th></th>
<th>Last 1 year AOD cost</th>
<th>Last 3 month AOD cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control ($n_c=11827, 99.2%$)</td>
<td>5534</td>
<td>1369</td>
</tr>
<tr>
<td>Treatment ($n_T=100, 0.8%$)</td>
<td>17622</td>
<td>4659</td>
</tr>
</tbody>
</table>

Without the adjustment from matching or regression, the differences between the means of last 1 year and last 3 month AOD tx cost are 12088 and 3290 respectively. We can clearly see that “treatment” group have a much higher cost both in long term and short term history. After matching, the matched “treatment” client and “control” client have relatively similar distribution in p-score and therefore their observed potential confounders. The outcome results coming from different matching processes are listed in Table 14 for comparison. Since the ratio of (treatment: control) is very small (i.e. there are enough control subjects who can be matched at each time point for treatment subjects), the treatment effects estimated from NN, OPM, OFM are not quite different from each other. All these matching algorithms suggest that AOD cost decreases around 50–80 per month. Since OFM achieves the most balanced matched pairs and general balanced distributions of key variables, the treatment estimation from OFM is more convincible compared to other matching algorithms.

Figure 10 outlines the quarterly cost difference of AOD cost along with its components (detox, inpatient and outpatient cost) between engagement and control groups 1 years after the treatment enrollment. The graph suggests that “AOD 3 month cost” declines gradually. Looking inside its three components, we can see the decrease mostly comes from the drop of detox cost and we do see a climb of outpatient cost at the first 3 months
after the admission. All these findings suggest that such “management care system” has
an effect on cost saving, especially on the most expensive part (detox cost) and
encourages clients to be transferred to relatively mild treatment- “outpatient” within 3
months after the admission.

2.4 Discussion

The simulations and real data analysis show some interesting properties of OFM
compared to other matching methods under longitudinal scenario: Even though the
distance (e.g. p-score) is still Euclidean distance calculated from logistic model, it varies
across different time points (e.g. in longitudinal study). Under this circumstance, NN,
OPM and OFM select different control subjects to treatment ones. From simulation study,
among all competitive 1:1, non-replacement matching algorithms (NN, OPM, & OFM),
OFM achieves the best balances of the p-score and most of controlled covariates and
provides the best (smallest percentage bias, highest CR, and smallest RMSE) treatment
effect estimations without sacrificing validity (i.e. without substantially increasing s.e.).
In the same time, as we have seen in cross-sectional study, OFM generates the least skew
pair whose largest measure of imbalance (in the sense of “p-score difference” between
treatment and control units) is the smallest. Due to the above facts, OFM may be used in
longitudinal PSA study to achieve the robust and efficient treatment effect estimation
when the true treatment effect estimation model is hard to be identified and/or testified.
Abstract

This paper employs inference by multiple imputation (MI) to deal with arbitrary missing values. We assume MAR as the underlying missingness mechanism on the responses or time-varying covariates. Posterior predictive distribution of missing data is simulated using Markov chain Monte Carlo (MCMC) simulation techniques. These simulation values, i.e. multiple imputations, are used to estimate treatment effects under previously developed OFM algorithm. We also compare this technique with other estimation techniques under the same MI methodology.

3.1 Introduction

Occurrence of missing values in longitudinal applications is typically normal and not an exception. For example, in clinical trial, patients may drop out because treatment does not work or another example, in a survey, respondents may provide partial data to survey items. The reason statistical literature focused heavily on missing data problem is that if no action is taken, validity of the statistic inference is highly questionable (Rubin 1987; Schafer 1997; Little and Rubin 2002). Rubin (1977) use the following notation to define the three different missing data mechanisms:

For subject $i$ at occasion $j$, define the complete data $Y = (y_{ij}) = (Y_{obs}, Y_{mis})$ and the missing pattern indicator matrix $M = (m_{ij} = 1)$ (i.e. $Y_{obs}^{ij}$ is observed) or $(m_{ij} = 0)$ (i.e. $Y_{mis}^{ij}$ is unobserved). Let us use $f(M|Y, \theta)$ to denote the conditional distribution of $M$ given $Y$, where $\theta$ denotes unknown parameter. This is also called missingness mechanism, and Rubin (1977) defines the following conditions:
If \( f(M|Y, \theta) = f(M|\theta) \) for all \( Y, \theta \), the missingness mechanism is called missing completely at random (MCAR).

If \( f(M|Y, \theta) = f(M|Y_{obs}, \theta) \) for all \( Y, \theta \), the missingness mechanism is called missing at random (MAR).

If \( f(M|Y, \theta) = f(M|Y_{mis}, \theta) \) for all \( Y, \theta \), the missingness mechanism is called missing not at random (MNAR).

When the underlying missing mechanism is MCAR, one can perform Complete Case (CC) analysis without inferential harm with respect to bias. However, it can be inefficient especially in multivariate analysis.

When the underlying missing mechanism is MAR (or MNAR), CC methodology fails and there are two major methods of inference:

The first one is “Expectation-Maximization (EM)” algorithm: an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables.

The second one is “Multiple Imputation (MI)” algorithm (Rubin 1987) which allows complete data techniques to be followed while quantifying uncertainty associated with missing data.

When the underlying missing mechanism is MNAR, EM and MI could still be used as long as the missingness mechanism is correctly specified and incorporated into the likelihood/posterior distribution.

As the data design becomes more complex, conducting principled missing data inference also becomes more complex. In longitudinal applications, Schafer and Yucel (2002) applied a multivariate extension of a popular linear mixed-effects model and created
multiple imputations of missing values for subsequent analyses using Markov chain Monte Carlo (MCMC) procedure. Here we apply these computational techniques to conduct MI inference in studying causal inference using OFM with SMF.

3.2 Creating Multiple Imputations

The following methodology used by Schafer and Yucel (2002) listed here, we use the same format:

Let $y_i$ denote a $n_i \times r$ matrix of multivariate responses for sample unit $i$, $i = 1, 2, \ldots, m$, where each row of $y_i$ is a joint realization of variables $Y_{1i}, Y_{2i}, \ldots, Y_{ri}$. We consider situations where portions of $y_1, \ldots, y_m$ are MCAR. The model for the complete data is

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i \quad (1)$$

where $X_i$ ($n_i \times p$) and $Z_i$ ($n_i \times q$) are known covariate matrices, $\beta$ ($p \times r$) is a matrix of regression coefficients common to all subjects, and $b_i$ ($q \times r$) is a matrix of coefficients specific to subject $i$. Here $\beta$ and $b_i$ are called “fixed effect” and “random effects”, respectively. We assume that $n_i$ rows of $\varepsilon_i$ are distributed as $N(0, \Omega)$, and that the random effects are distributed as $\text{vec}(b_i) \sim N(0, \Psi)$ independently for $i = 1, \ldots, m$ (the “vec” operator vectorizes a matrix by stacking its columns). Without conditioning on $b_1, \ldots, b_m$, the implied model for $\text{vec}(y_i)$ is normal with mean $\text{vec}(X_i\beta)$ and covariance matrix

$$V(y_i|x_i) = Z_i \Psi Z_i' + \sigma^2_i \Omega_i \quad (2)$$

In longitudinal application, timing of repeated measures should be more or less the same for all subjects even though subjects do not have to be measured at all timepoints.

If there only exists one response variable, the simplified model is called univariate model that has been extensively treated by Laird and Ware (1982) and others. However, if there
are multiple responses and/or there are incomplete data in responses, the practice of the so-called “seemingly unrelated regression” (Zellner 1962) becomes impracticable (Verbeke and Molenberghs 2000). Furthermore, if there are missing data in predictors, the joint modeling of the multiple responses becomes helpful or even necessary (Schafer and Yucel 2002). Shafer and Yucel (2002) propose inferential techniques based on EM and MI under multivariate linear mixed effect model similar to equation (1).

Schafer and Yucel (2002) also implemented an iterative simulation algorithm called Gibbs Sampler. This algorithm is used to simulate posterior predictive distribution of missing data. It iteratively samples from the distribution parameters in (3) and (4) below. Assuming that the current values of the unknown parameters and missing data are

\[ \theta^{(t)} = (\beta^{(t)}, \Sigma^{(t)}, \psi^{(t)}) \text{ and } Y_{mis}^{(t)}, \] Gibbs Samples updates \( (\theta^{(t)}, Y_{mis}^{(t)}) \) as:

\[ \theta^{(t+1)} \sim P(\theta^{(t)} | Y_{obs}, Y_{mis}^{(t)}) \quad (3) \]

\[ Y_{mis}^{(t+1)} \sim P(Y_{mis}^{(t)} | Y_{obs}, \theta^{(t+1)}) \quad (4) \]

for \( i = 1, \ldots, m \). Given starting values \( \theta^{(0)} \) and \( Y_{mis}^{(0)} \), these steps define one cycle of a MCMC procedure called a Gibbs sampler. Executing the cycle repeatedly creates sequences \{\theta^{(1)}, \theta^{(2)}, \ldots\} and \{Y_{mis}^{(1)}, Y_{mis}^{(2)}, \ldots\\} whose limiting distributions \( P\{\theta^{(t)}, Y_{mis}^{(t)}\} \)

\[ \sim P(\theta, Y_{mis} | Y_{obs}). \]

These techniques have been implemented in an R package called “pan” (Schafer and Yucel 2002).
3.3 MI inference using OFM under SMF with missing data

Now we provide a simulation study to gauge the performance of the MI inference in studying the OFM. After creating MI, we apply methods of the section 2 to make inference on the treatment effect under OFM.

We simulate a data set (n=150) with 4 time points (t = 1, 2, 3, 4) of covariates, treatment indicator variable (T_{it}) and outcome variable (O_{it}) in the following fashion:

At each time point t, there are the following covariates:

x_{1it} = 0 (if t = 1,3) or 1 (if t = 2,4). x_{1it} simulates the dummy variable;

x_{2it} = t. x_{1it} simulates the time variable;

\begin{pmatrix} x_{3i} \\ x_{4i} \end{pmatrix} \sim MVN \left( \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix} \right);

T_{it} \sim Bin(P(T_{it} = 1)) where

\text{logit} \left( P(T_{it} = 1) \right) = x_{1it} - 0.5x_{2it} + 0.5x_{3it};

Let’s define S_{it} as how long the client has stayed in the treatment group. The outcome variable (O_{it}) is simulated as:

O_{it} = \gamma_1 T_{it} S_{it} + \gamma_2 x_{1it} + \gamma_3 x_{3it} + \gamma_4 x_{4it} + \epsilon_{it},

where \gamma_1 = 1, \gamma_2 = 5, \gamma_3 = -1, \gamma_4 = 4; t = 1, 2, 3, 4; and

\epsilon_{it} \sim N \left( \begin{pmatrix} 0.2 \\ 0 \\ -0.3 \\ 0 \end{pmatrix}, \Omega_i \right)

where \Omega_i = Toep(0.3).

Note that the simulation of O ignores x_2 and incorporates x_4 because the variables determining outcome may not be exactly the same variables determining the treatment.
The within-subject covariance (for subject i between time points j and j’) from the simulation is
\[ \text{Cov}(y_{i,j}y_{i,j’}|x_i) = (\Omega_i)_{j,j’}. \]

The simulation iterated 1000 times and on average led to 100 controls (T=0) and 50 treatment subjects (T=1) (i.e. \(n_c=100\) (67%) and \(n_T=50\) (33%)). Now let us apply missing value mechanism to the complete data. Suppose missing is MCAR and \(P(m_{ij}=1|\theta) = 0.3\) (where \(\theta\) is missingness parameter). Due to the complexity of between-subject and within-subject association of the longitudinal data, we use function “pan” from “pan” package in R (Schafer 1997; Schafer 1997; Schafer and Yucel 2002) to impute the missing data and use Rubin’s rule to aggregate the estimations of treatment effect and standard error from multiple imputation (MI, here iteration=10). Then we use the methodology similar to section 2.3.1 to do the simulation and calculate the statistics of the estimated treatment effects. Recall that two working models to estimate the treatment effects are:

Correct model:
\[ O_{it} = \gamma_1 T_{it}S_{it} + \gamma_2 x_{1it} + \gamma_3 x_{3it} + \gamma_4 x_{4it} \] (Model 7);

Incorrect model:
\[ O_{it} = \gamma_1’ T_{it}S_{it} + \gamma_2’ x_{2it} + \gamma_3’ x_{3it} + \gamma_4’ x_{4it} \] (Model 8).

For model 7, the results are summarized in Table 15. For model 8, the results are summarized in Table 16.

Under the circumstance of correct model (model 7), OFM (without caliper) obtains the closest estimation of treatment increment per time (\(\hat{\gamma}_1=1.00\), 95% CI = (0.87, 1.12) with the smallest percentage bias (0.42%), highest CR (96.2%), and the smallest RMSE (0.06).
among all competing matching algorithms without sacrificing validity (i.e. without substantially increasing s.e.). Under the circumstance of incorrect model (model 8), once again, OFM (without caliper) obtains the general best estimation of treatment effect increment per time ($\gamma_1 = 1.14$, 95% CI = (0.97, 1.30)) with smallest percentage bias (13.62%), 2nd highest CR (84.8%), and the smallest RMSE (0.17) among all competing matching algorithms without sacrificing validity (i.e. without substantially increasing s.e.). However, among all 1:1, non-replacement matching algorithms, OFM achieve the smallest percentage bias, highest CR, and the smallest RMSE (compared to NN, OPM). Note at this time, even though RDM provides the highest CR, it has the largest 95% CI and largest RMSE among all matching algorithms. From above two simulation studies, we can conclude that under MCAR missing mechanism in longitudinal data study, by using MI imputation through “pan” package in R, comparing to other competing 1:1, non-replacement matching algorithms, OFM provides the best (smallest percentage bias, highest CR, and smallest RMSE) treatment effect estimations without sacrificing validity (i.e. without substantially increasing s.e.) both in correct and in incorrect treatment effect estimation working models.

3.4 Discussion

The simulation studies implemented here extend the study of matching algorithm OFM and other competing matching algorithms in longitudinal PSA to the circumstance of longitudinal data with missing values. Very rarely, the performances of different matching algorithms of PSA study on the longitudinal missing data have been compared with the simulation using Rubin’s rule through MI.
With the first step of using “pan” package to simulate missing data, second step of applying different matching algorithms to achieve the estimations of treatment effect, and third step of aggregating these MI estimations using Rubin’s rule, it is straightforward to show that under both correct and incorrect estimating model circumstances, OFM provides the best (smallest percentage bias, highest CR, and smallest RMSE) treatment effect estimations without sacrificing validity (i.e. without substantially increasing s.e.) among all competitive 1:1, non-replacement matching algorithms (NN, OPM & OFM). In the same time, as we have seen in complete data longitudinal PSA study, OFM will also achieve the best balance of controlled covariates and generate least skew pair whose largest measure of imbalance (in the sense of “p-score difference” generated in step I of our new matching framework (SMF) between treatment and control units) is the smallest. Due to the above facts, OFM may be used in longitudinal PSA study with missing values to achieve the robust and efficient treatment effect estimation when the true estimating model is hard to be identified and/or testified.
4. Discussion

Stepwise Matching Framework (SMF) provides a framework where all matching algorithms can be defined and explained as either the different ways of creating “distance matrix” or different ways of selecting and grouping units corresponding to the different steps (or sub-steps) of SMF.

Furthermore, SMF also provide a mathematical tool to interpret some interesting topics and obstacles related to matching (e.g. 1:1 or n to m; with or without replacement; non compatible (NC) issue; a tie value; and etc.). Under SMF, one new matching algorithm – Outlier First Matching is defined and its performance is compared to other well-established matching algorithms for observational causal inference study.

In cross-sectional PSA study where p-score is estimated using logistic regression, all three 1:1, non-replacement matching algorithms (NN, OPM, and OFM) select the same treatment and control units. However, the matched pairs are different from these three matching algorithms. Among them, OFM generate the least skew matched pair (i.e. the largest distance difference from the matched pairs is the smallest). Since the units from different matching algorithms are same, whatever under the correct or incorrect outcome working model, the estimated treatment effects are the same among all 1:1, non-replacement matching algorithms (NN, OPM, & OFM).

On the other side, in longitudinal PSA study where p-score varies across the time, the three matching algorithms mentioned above not only select the different units but also generate different matched pairs. Among them, OFM again, generates the least skew matched pair. Through simulation study, at longitudinal data circumstance, among the above three algorithms, OFM tends to provide the best estimation of treatment effect (in
the sense of smallest bias, highest coverage rate, and smallest RMSE) both in correct and in incorrect outcome working models even with missing data values.

OFM tends to generate least matched pairs and provide the best estimation of treatment effect both in cross-sectional and longitudinal data circumstances. Besides its relatively good performance, its property of handling “NC”, “tie” values automatically and the readiness of potential use in wide regions such as “non-Euclidean distance” and “correlated data” can fulfill different research interests.

The future research interests of OFM will be on the following directions: First, we would like to increase the computation speed of OFM to meet the request of huge data set. Second, the performance of OFM will be tested using non-Euclidean distance (e.g. Mahalanobis distance) under SMF compared to other well-established matching algorithms. Third, some supplementary programs such as testing “balance”, checking “NC” pattern, and etc. will be added to the main program “OFM” to create one R package.
References


Analysis of observational health care data using SAS


### Tables

**Table 1: List of different matching methods based on their targeting interests**

<table>
<thead>
<tr>
<th>Object of interest, or Goal</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match the smallest distance first and keep looking for the next smallest distance, etc.</td>
<td>Nearest Neighbor Matching (NN)</td>
</tr>
<tr>
<td>Match to get shortest overall distance</td>
<td>Full Matching (Pirkle, Kaufmann et al.), Optimal Matching (OPM)</td>
</tr>
<tr>
<td>Match all units within some radius or caliper limit</td>
<td>Radius Matching (RDM), Caliper Matching</td>
</tr>
<tr>
<td>Match based on stratification</td>
<td>Stratification (on quantiles, or from leaves in classification tree)</td>
</tr>
<tr>
<td>Match based on selected covariates</td>
<td>Coarsened Exact Matching (CEM), Mahalanobis Distance Matching (MDM)</td>
</tr>
<tr>
<td>Match the largest distance first</td>
<td>Outlier First Matching (OFM)</td>
</tr>
</tbody>
</table>
Table 2: Comparison of estimated treatment effects (true parameter=1) from different matching methods for simulated cross-sectional data (multivariate-normal) using correct model (simulation times=1000)

<table>
<thead>
<tr>
<th>Method</th>
<th>w/o matching</th>
<th>NN</th>
<th>OPM</th>
<th>RDM $^a$</th>
<th>OFM</th>
<th>OFM w/ Caliper $^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1 or 1: n</td>
<td></td>
<td>1: 1</td>
<td>1: n</td>
<td>1: 1</td>
<td>1:1</td>
<td></td>
</tr>
<tr>
<td>w or w/o replacement</td>
<td>w/o</td>
<td>w/o</td>
<td>w</td>
<td>w</td>
<td>w/o</td>
<td>w/o</td>
</tr>
<tr>
<td>Crude dif between the mean of the treatment and control groups (unadjusted for any other variables)</td>
<td>1.49 (-0.69, 3.65)</td>
<td>1.24 (-1.16, 3.61) &amp; 1.11 (-1.60, 3.61) &amp; 1.24 (-1.16, 3.61) &amp; 1.10 (-1.54, 3.73)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression-adjusted difference between treatment and controls (95% CI)</td>
<td>1.01 (0.54, 1.47) &amp; 1.01 (0.51, 1.50) &amp; 1.01 (0.37, 1.68) &amp; 1.01 (0.51, 1.50) &amp; 1.01 (0.51, 1.50)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentage Bias (%)</td>
<td>0.69%</td>
<td>1.07%</td>
<td>1.07%</td>
<td>1.09%</td>
<td>1.07%</td>
<td>0.94%</td>
</tr>
<tr>
<td>Coverage Rate (%)</td>
<td>94.3%</td>
<td>95.1%</td>
<td>95.1%</td>
<td>84.6%</td>
<td>95.1%</td>
<td>95.0%</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.34</td>
<td>0.25</td>
<td>0.26</td>
</tr>
</tbody>
</table>

$^a$ Use 0.25 of the s.e. of the treatment group “p_score” to construct the radius.

$^b$ Use 0.20 of the s.e. of the treatment group “p-score” to construct the caliper.
Table 3: Comparison of estimated treatment effects (true parameter=1) from different matching methods for simulated cross-sectional data (multivariate-normal) using incorrect model \(^a\) (simulation times=1000)

<table>
<thead>
<tr>
<th>Method</th>
<th>w/o matching</th>
<th>NN</th>
<th>OPM</th>
<th>RDM (^b)</th>
<th>OFM</th>
<th>OFM w/ Caliper (^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1 or 1: n</td>
<td></td>
<td>1:1</td>
<td>1:1</td>
<td>1: n</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>w or w/o replacement</td>
<td></td>
<td>w/o</td>
<td>w/o</td>
<td>w</td>
<td>w/o</td>
<td>w/o</td>
</tr>
<tr>
<td>Crude diff between the</td>
<td>1.49</td>
<td>1.24</td>
<td>1.24</td>
<td>1.24</td>
<td>1.11</td>
<td>1.10</td>
</tr>
<tr>
<td>mean of the treatment</td>
<td>(-0.69,3.65)</td>
<td>(-1.16, 3.61)</td>
<td>(-1.16, 3.61)</td>
<td>(-1.60, 4.11)</td>
<td>(-1.16,3.61)</td>
<td>(-1.54, 3.73)</td>
</tr>
<tr>
<td>and control groups</td>
<td>(unadjusted for any other variables)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression-adjusted</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>difference between</td>
<td>(-1.41, 3.46)</td>
<td>(-1.61,3.45)</td>
<td>(-1.61,3.45)</td>
<td>(-2.41,4.13)</td>
<td>(-1.61,3.45)</td>
<td>(-1.56,3.64)</td>
</tr>
<tr>
<td>treatment and controls</td>
<td>(95% CI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentage Bias (%)</td>
<td>1.75%</td>
<td>1.52%</td>
<td>1.52%</td>
<td>2.30%</td>
<td>1.52%</td>
<td>1.04%</td>
</tr>
<tr>
<td>Coverage Rate (%)</td>
<td>92.9%</td>
<td>93.0%</td>
<td>93.0%</td>
<td>84.3%</td>
<td>93.0%</td>
<td>93.2%</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.23</td>
<td>1.26</td>
<td>1.26</td>
<td>1.64</td>
<td>1.26</td>
<td>1.33</td>
</tr>
</tbody>
</table>

\(^a\) The working model is “\(O_t = \gamma_1 x_{1t} + \gamma_2 x_{2t} + \gamma_3 x_{3t}\)” instead of “\(O_t = \gamma_1 x_{1t} + 2 \gamma_2 x_{3t} + \gamma_3 x_{4t}\)”.  

\(^b\) Use 0.25 of the s.e. of the treatment group “p_score” to construct the radius.  

\(^c\) Use 0.20 of the s.e. of the treatment group “p-score” to construct the caliper.
Table 4: Comparison of estimated treatment effects (true parameter=1) from different matching methods for simulated cross-sectional skew data (multivariate-skewnormal) using correct model (simulation times=1000)

<table>
<thead>
<tr>
<th>Method</th>
<th>w/o matching</th>
<th>NN</th>
<th>OPM</th>
<th>RDM</th>
<th>OFM</th>
<th>OFM w/ Caliper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1 or 1: n</td>
<td>1:1</td>
<td>1:1</td>
<td>1: n</td>
<td>1:1</td>
<td>1:1</td>
<td></td>
</tr>
<tr>
<td>w or w/o replacement</td>
<td>w/o</td>
<td>w/o</td>
<td>w</td>
<td>w/o</td>
<td>w/o</td>
<td></td>
</tr>
<tr>
<td>Crude dif between the mean of the treatment and control groups (unadjusted for any other variables)</td>
<td>2.19 (-0.07, 4.58)</td>
<td>1.31 (-0.91, 3.82)</td>
<td>1.31 (-0.91, 3.82)</td>
<td>1.27 (-1.04, 3.71)</td>
<td>1.31 (-0.91, 3.82)</td>
<td>1.14 (-1.14, 3.51)</td>
</tr>
<tr>
<td>Regression-adjusted difference between treatment and controls (95% CI)</td>
<td>0.99 (0.46, 1.49)</td>
<td>0.99 (0.41, 1.54)</td>
<td>0.99 (0.41, 1.54)</td>
<td>0.98 (0.18, 1.71)</td>
<td>0.99 (0.41, 1.54)</td>
<td>0.99 (0.39, 1.55)</td>
</tr>
<tr>
<td>Percentage Bias (%)</td>
<td>0.81%</td>
<td>0.73%</td>
<td>0.73%</td>
<td>1.54%</td>
<td>0.73%</td>
<td>1.11%</td>
</tr>
<tr>
<td>Coverage Rate (%)</td>
<td>94.6%</td>
<td>95.8%</td>
<td>95.8%</td>
<td>83.1%</td>
<td>95.8%</td>
<td>95.5%</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.26</td>
<td>0.29</td>
<td>0.29</td>
<td>0.38</td>
<td>0.29</td>
<td>0.30</td>
</tr>
</tbody>
</table>
Table 5: Comparison of estimated treatment effects (true parameter=1) from different matching methods for simulated cross-sectional data (multivariate-skewnormal) using incorrect model a (simulation times=1000)

<table>
<thead>
<tr>
<th>Method w/o matching</th>
<th>w/o replacement</th>
<th>w</th>
<th>w</th>
<th>w/o</th>
<th>w/o</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1 or 1: n</td>
<td>1:1</td>
<td>1:1</td>
<td>1:n</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>NN</td>
<td>w/o</td>
<td>w/o</td>
<td>w</td>
<td>w/o</td>
<td>w/o</td>
</tr>
</tbody>
</table>

Crude diff between the mean of the treatment and control groups (unadjusted for any other variables)

<table>
<thead>
<tr>
<th></th>
<th>w/o replacement</th>
<th>w</th>
<th>w</th>
<th>w/o</th>
<th>w/o</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression-adjusted difference between treatment and controls (95% CI)</td>
<td>1.72 (0.94, 2.48)</td>
<td>1.20 (0.43, 1.94)</td>
<td>1.20 (0.43, 1.94)</td>
<td>1.70 (0.54, 2.85)</td>
<td>1.20 (0.43, 1.94)</td>
</tr>
</tbody>
</table>

Percentage Bias (%)

<table>
<thead>
<tr>
<th></th>
<th>w/o replacement</th>
<th>w</th>
<th>w</th>
<th>w/o</th>
<th>w/o</th>
</tr>
</thead>
<tbody>
<tr>
<td>72.02%</td>
<td>19.51%</td>
<td>19.51%</td>
<td>69.99%</td>
<td>19.51%</td>
<td>9.48%</td>
</tr>
</tbody>
</table>

Coverage Rate (%)

<table>
<thead>
<tr>
<th></th>
<th>w/o replacement</th>
<th>w</th>
<th>w</th>
<th>w/o</th>
<th>w/o</th>
</tr>
</thead>
<tbody>
<tr>
<td>55.4%</td>
<td>94.9%</td>
<td>94.9%</td>
<td>57.3%</td>
<td>94.9%</td>
<td>97.9%</td>
</tr>
</tbody>
</table>

RMSE

<table>
<thead>
<tr>
<th></th>
<th>w/o replacement</th>
<th>w</th>
<th>w</th>
<th>w/o</th>
<th>w/o</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.83</td>
<td>0.43</td>
<td>0.43</td>
<td>0.91</td>
<td>0.43</td>
<td>0.37</td>
</tr>
<tr>
<td>Analysis used</td>
<td>Average Treatment Effect (s.e.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------------</td>
<td>---------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difference between the mean of the treatment and control groups (unadjusted for any other variables)</td>
<td>-1794 (633)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression-adjusted difference between treatment and controls</td>
<td>-1672 (637)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 7: Covariates description of "Lalonde" data from R package “MatchIt”

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>1 if in NSW Job Training Program; 0 if PSID control case</td>
</tr>
<tr>
<td>age</td>
<td>Age</td>
</tr>
<tr>
<td>educ</td>
<td>Education (# of years)</td>
</tr>
<tr>
<td>black</td>
<td>1 if black</td>
</tr>
<tr>
<td>hisp</td>
<td>1 if Hispanic</td>
</tr>
<tr>
<td>married</td>
<td>1 if married</td>
</tr>
<tr>
<td>nodegree</td>
<td>1 if no high school degree</td>
</tr>
<tr>
<td>re74</td>
<td>1974 earnings</td>
</tr>
<tr>
<td>re75</td>
<td>1975 earnings</td>
</tr>
<tr>
<td>re78</td>
<td>1978 earnings (outcome)</td>
</tr>
</tbody>
</table>
Table 8: Comparison of estimated treatment effects among different matching methods of Lalonde Study

<table>
<thead>
<tr>
<th>Method</th>
<th>NN</th>
<th>OPM</th>
<th>Radius (Caliper) a</th>
<th>OFM</th>
<th>OFM w/ Caliper b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1 or 1: n</td>
<td>1: 1</td>
<td>1: 1</td>
<td>1: n</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>w or w/t replacement</td>
<td>w/t</td>
<td>w/t</td>
<td>with</td>
<td>w/t</td>
<td>w/t</td>
</tr>
<tr>
<td>matched pairs</td>
<td>185</td>
<td>185</td>
<td>184</td>
<td>185</td>
<td>116</td>
</tr>
<tr>
<td>Dif between the mean</td>
<td>-908</td>
<td>620</td>
<td>-1206</td>
<td>-894</td>
<td>-1517</td>
</tr>
<tr>
<td>of the treatment and control groups (unadjusted for any other variables)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression-adjusted difference</td>
<td>-1351</td>
<td>-1176</td>
<td>-1249</td>
<td>-1345</td>
<td>-1753</td>
</tr>
<tr>
<td>between treatment and controls (95% CI)</td>
<td>(-2899,197)</td>
<td>(-3163,811)</td>
<td>(-2543,-155)</td>
<td>(-2893,203)</td>
<td>(-2211,283)</td>
</tr>
</tbody>
</table>

a Use 0.25 of the s.e. of the treatment group “p_score” to construct the radius.

b Use 0.20 of the s.e. of the treatment group “p-score” to construct the caliper
Table 9: Comparison of estimated treatment effects (true parameter=1) from different matching methods for simulated longitudinal data using correct model (simulation times=1000)

<table>
<thead>
<tr>
<th>Method</th>
<th>NN</th>
<th>OPM</th>
<th>RDM a</th>
<th>OFM</th>
<th>OFM w/ Caliper b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1 or 1: n</td>
<td>1: 1</td>
<td>1: 1</td>
<td>1: n</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>w or w/o replacement</td>
<td>w/o</td>
<td>w/o</td>
<td>w</td>
<td>w/o</td>
<td>w/o</td>
</tr>
<tr>
<td>Regression-adjusted Treatment effect increment per time point (95% CI)</td>
<td>0.95 (0.81, 1.09)</td>
<td>0.95 (0.80, 1.09)</td>
<td>0.89 (0.47, 1.34)</td>
<td>0.98 (0.84, 1.12)</td>
<td>0.98 (0.84, 1.13)</td>
</tr>
<tr>
<td>Percentage Bias (%)</td>
<td>5.44%</td>
<td>5.40%</td>
<td>6.87%</td>
<td>2.27%</td>
<td>2.27%</td>
</tr>
<tr>
<td>Coverage Rate (%)</td>
<td>84.9%</td>
<td>84.9%</td>
<td>66.3%</td>
<td>92.9%</td>
<td>92.7%</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.09</td>
<td>0.09</td>
<td>1.38</td>
<td>0.07</td>
<td>0.08</td>
</tr>
</tbody>
</table>

a Use 0.25 of the s.e. of the treatment group “p_score” to construct the radius.

b Use 0.20 of the s.e. of the treatment group “p-score” to construct the caliper.
Table 10: Comparison of estimated treatment effects (true parameter=1.0) from different matching methods for simulated longitudinal data using incorrect model (simulation times=1000)

<table>
<thead>
<tr>
<th>Method</th>
<th>NN</th>
<th>OPM</th>
<th>RDM (^a)</th>
<th>OFM</th>
<th>OFM w/ Caliper (^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1 or 1: n</td>
<td>1: 1</td>
<td>1: 1</td>
<td>1: n</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>w or w/o</td>
<td>w/o</td>
<td>w/o</td>
<td>w</td>
<td>w/o</td>
<td>w/o</td>
</tr>
<tr>
<td>replacement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression-adjusted Treatment effect increment per time point (95% CI)</td>
<td>1.32</td>
<td>1.29</td>
<td>0.72</td>
<td>1.15</td>
<td>1.15</td>
</tr>
<tr>
<td>(95% CI)</td>
<td>(0.88, 2.12)</td>
<td>(0.94, 1.91)</td>
<td>(-0.74, 2.09)</td>
<td>(0.94, 1.35)</td>
<td>(0.94, 1.35)</td>
</tr>
<tr>
<td>Percentage Bias (%)</td>
<td>31.78%</td>
<td>28.73%</td>
<td>28.49%</td>
<td>14.56%</td>
<td>14.62%</td>
</tr>
<tr>
<td>Coverage Rate (%)</td>
<td>44.8%</td>
<td>43.2%</td>
<td>78.5%</td>
<td>68.4%</td>
<td>68.2%</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.10</td>
<td>0.83</td>
<td>6.18</td>
<td>0.18</td>
<td>0.18</td>
</tr>
</tbody>
</table>

\(^a\) Use 0.25 of the s.e. of the treatment group “p_score” to construct the radius.

\(^b\) Use 0.20 of the s.e. of the treatment group “p-score” to construct the caliper.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>1 if ever enrolled into “management care system”; 0 otherwise</td>
</tr>
<tr>
<td>age</td>
<td>Age (5 categories)</td>
</tr>
<tr>
<td>black</td>
<td>1 if black</td>
</tr>
<tr>
<td>hisp</td>
<td>1 if Hispanic</td>
</tr>
<tr>
<td>married</td>
<td>1 if ever married</td>
</tr>
<tr>
<td>nodegree</td>
<td>1 if no high school degree</td>
</tr>
<tr>
<td>Homeless</td>
<td>1 if homeless at enrolled time</td>
</tr>
<tr>
<td>Severe_mh</td>
<td>1 if “severe mental health issue”</td>
</tr>
<tr>
<td>Hepatitis</td>
<td>1 if “Hepatitis”</td>
</tr>
<tr>
<td>AIDS</td>
<td>1 if AIDS</td>
</tr>
<tr>
<td>Alcohol_Prime</td>
<td>1 if “Alcohol as Prim. Substance usage”</td>
</tr>
<tr>
<td>Sub. Daily Usage</td>
<td>1 if “use substance daily”</td>
</tr>
<tr>
<td>AOD Cost_1yr</td>
<td>AOD cost of last year = “Detox cost 1 yr” + “Rehab cost 1 yr” + “Outp cost 1yr”</td>
</tr>
<tr>
<td>Detox cost 1 yr</td>
<td>Detox cost within last year</td>
</tr>
<tr>
<td>Rehab cost 1 yr</td>
<td>Rehab cost within last year</td>
</tr>
<tr>
<td>Outp cost 1 yr</td>
<td>Outpatient cost within last year</td>
</tr>
<tr>
<td>AOD Cost_3m</td>
<td>Cost in last 3 months = “Detox cost 3mh”+“Rehab cost 3mh”+“Outp cost 3mh”</td>
</tr>
<tr>
<td>Detox cost 3mh</td>
<td>Detox cost within last 3 months</td>
</tr>
<tr>
<td>Rehab cost 3mh</td>
<td>Rehab cost within last 3 months</td>
</tr>
<tr>
<td>Outp cost 3mh</td>
<td>Outpatient cost within last 3 months</td>
</tr>
<tr>
<td>Count of detox</td>
<td>Count of detox episodes within last year (4 categories: 0, 1, 2, 3+)</td>
</tr>
<tr>
<td>Count of rehab</td>
<td>Counts of inpatient rehabilitation episodes within last year (3 categories: 0, 1, 2+)</td>
</tr>
<tr>
<td>Count of Outp visit</td>
<td>Count of outpatient visits within last year (6 categories: 0, (0, 20%], (20%, 40%], (40%, 60%], (60%, 80%], (80%, 100%])</td>
</tr>
</tbody>
</table>
### Table 12: Comparison of the balances from different matching algorithm for Management Care Study

<table>
<thead>
<tr>
<th></th>
<th>Before Matching (100:)</th>
<th>NN (100: 100)</th>
<th>OPM</th>
<th>OFM w.o. cal (100:100)</th>
<th>OFM w. cal (98:98)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trt Effect Size^a</td>
<td>Ctrl Trt Effect Size^a</td>
<td>Ctrl Trt Effect Size^a</td>
<td>Ctrl Trt Effect Size^a</td>
<td>Ctrl Trt Effect Size^a</td>
</tr>
<tr>
<td>Black</td>
<td>37.0% 1.00^c</td>
<td>35.5% 0.63^c</td>
<td>27.0% 0.51^c</td>
<td>23.0% 0.66^c</td>
<td>28.0% 0.71^c</td>
</tr>
<tr>
<td>Less than high school</td>
<td>29.0% 1.14^c</td>
<td>31.8% 0.77^c</td>
<td>24.0% 0.91^c</td>
<td>27.0% 1.05^c</td>
<td>30.0% 1.10^c</td>
</tr>
<tr>
<td>Severe mental health</td>
<td>78.0% 0.22^c</td>
<td>43.7% 1.06^c</td>
<td>79.0% 0.94^c</td>
<td>77.0% 0.80^c</td>
<td>74.0% 0.80^c</td>
</tr>
<tr>
<td>AIDS</td>
<td>8.0% 0.45^c</td>
<td>3.8% 1.42^c</td>
<td>11.0% 1.00^c</td>
<td>8.0% 1.57^c</td>
<td>12.0% 1.56^c</td>
</tr>
<tr>
<td>Alcohol</td>
<td>44.0% 0.81^c</td>
<td>38.9% 0.78^c</td>
<td>38.0% 0.88^c</td>
<td>41.0% 0.72^c</td>
<td>36.0% 0.74^c</td>
</tr>
<tr>
<td>Sub.Daily Usage</td>
<td>31.0% 1.75^c</td>
<td>44.0% 1.15^c</td>
<td>34.0% 1.15^c</td>
<td>34.0% 1.15^c</td>
<td>34.0% 1.15^c</td>
</tr>
<tr>
<td>Cnt_Care</td>
<td>62.0% 0.21^c</td>
<td>25.3% 0.85^c</td>
<td>58.0% 1.00^c</td>
<td>62.0% 1.00^c</td>
<td>62.0% 1.00^c</td>
</tr>
<tr>
<td>AOD cost last year</td>
<td>17622 -0.97^b</td>
<td>5534 -0.05^b</td>
<td>16971 -0.09^b</td>
<td>16530 0.08^b</td>
<td>18589 0.09^b</td>
</tr>
<tr>
<td>AOD cost last 3 months</td>
<td>4659 -0.58^b</td>
<td>1370 -0.21^b</td>
<td>3460 -0.18^b</td>
<td>3640 -0.06^b</td>
<td>4303 0.06^b</td>
</tr>
<tr>
<td>Inp. Cost last year</td>
<td>7374 -0.66^b</td>
<td>1421 -0.22^b</td>
<td>5417 -0.19^b</td>
<td>5626 0.19^b</td>
<td>6187 0.19^b</td>
</tr>
</tbody>
</table>

^a Annotations regarding effect size.

^b Cohen’s d for mean comparison (continuous variable): \( d = \frac{\bar{x}_{trt} - \bar{x}_{ctl}}{s} \), where \( s = \sqrt{\frac{(n_{trt} - 1)s_{trt}^2 + (n_{ctl} - 1)s_{ctl}^2}{n_{trt} + n_{ctl}}} \) and \( s^2_{trt} = \frac{1}{n_{trt} - 1} \sum_{i=1}^{n_{trt}}(x_{trt,i} - \bar{x}_{trt})^2; \)

^c OR (odds ratio) for binary variable: \( OR = \frac{\text{odds of treatment group}}{\text{odds of control group}} = \frac{(s_{trt}^2)/(1-s_{trt}^2))}{(s_{ctl}^2)/(1-s_{ctl}^2)} \)
Table 13: Top 3 most skew matched pairs from different matching algorithms for Management Care Study

**NN top 3 outliers**

<table>
<thead>
<tr>
<th>Matched Pair</th>
<th>ID</th>
<th>Treatment</th>
<th>P score</th>
<th>Last year cost</th>
<th>Last 3 months cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0388</td>
<td>61317</td>
<td>23974</td>
</tr>
<tr>
<td>a</td>
<td>0</td>
<td>0.0249</td>
<td>23967</td>
<td>4558</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0.0208</td>
<td>12800</td>
<td>3359</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>0.0298</td>
<td>24575</td>
<td>1926</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0.0240</td>
<td>30670</td>
<td>2337</td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>0.0185</td>
<td>13603</td>
<td>5559</td>
<td></td>
</tr>
</tbody>
</table>

**OPM top 3 outliers**

<table>
<thead>
<tr>
<th>Matched Pair</th>
<th>ID</th>
<th>Treatment</th>
<th>P score</th>
<th>Last year cost</th>
<th>Last 3 months cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0388</td>
<td>61317</td>
<td>23974</td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>0.0175</td>
<td>19024</td>
<td>924</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0.0297</td>
<td>21067</td>
<td>12658</td>
</tr>
<tr>
<td>e</td>
<td>0</td>
<td>0.0100</td>
<td>14308</td>
<td>2072</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0.0283</td>
<td>17759</td>
<td>3697</td>
</tr>
<tr>
<td>f</td>
<td>0</td>
<td>0.0092</td>
<td>7086</td>
<td>1078</td>
<td></td>
</tr>
</tbody>
</table>

**OFM (w/o caliper) top 3 outliers**

<table>
<thead>
<tr>
<th>Matched Pair</th>
<th>ID</th>
<th>Treatment</th>
<th>P score</th>
<th>Last year cost</th>
<th>Last 3 months cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>1</td>
<td>0.0442</td>
<td>28473</td>
<td>12360</td>
</tr>
<tr>
<td>g</td>
<td>0</td>
<td>0.0400</td>
<td>43573</td>
<td>19572</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>1</td>
<td>0.0283</td>
<td>17759</td>
<td>3697</td>
</tr>
<tr>
<td>h</td>
<td>0</td>
<td>0.0317</td>
<td>40743</td>
<td>1695</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>1</td>
<td>0.0284</td>
<td>15791</td>
<td>2773</td>
</tr>
<tr>
<td>l</td>
<td>0</td>
<td>0.0264</td>
<td>12290</td>
<td>2456</td>
<td></td>
</tr>
</tbody>
</table>

**OFM (with caliper) top 3 outliers**

<table>
<thead>
<tr>
<th>Matched Pair</th>
<th>ID</th>
<th>Treatment</th>
<th>P score</th>
<th>Last year cost</th>
<th>Last 3 months cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>1</td>
<td>0.0078</td>
<td>11034</td>
<td>1032</td>
</tr>
<tr>
<td>j</td>
<td>0</td>
<td>0.0058</td>
<td>10770</td>
<td>1077</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>1</td>
<td>0.0064</td>
<td>27839</td>
<td>14160</td>
</tr>
<tr>
<td>k</td>
<td>0</td>
<td>0.0046</td>
<td>47281</td>
<td>3158</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>1</td>
<td>0.0121</td>
<td>33243</td>
<td>5007</td>
</tr>
<tr>
<td>l</td>
<td>0</td>
<td>0.0139</td>
<td>9210</td>
<td>3449</td>
<td></td>
</tr>
</tbody>
</table>
Table 14: Comparison of estimated treatment effects among different matching methods for Management Care Study

<table>
<thead>
<tr>
<th>Method</th>
<th>NN</th>
<th>OPM</th>
<th>RDM a</th>
<th>OFM</th>
<th>OFM w/ Caliper b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1 or 1: n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w or w/o replacement</td>
<td>w/o</td>
<td>w/o</td>
<td>w/o</td>
<td>w/o</td>
<td>w/o</td>
</tr>
<tr>
<td>matched pairs</td>
<td>100</td>
<td>100</td>
<td>98</td>
<td>100</td>
<td>98</td>
</tr>
<tr>
<td>(# of treatment elements: # of control elements)</td>
<td>(100: 100)</td>
<td>(100: 100)</td>
<td>(98: 11827)</td>
<td>(99: 99)</td>
<td>(98: 98)</td>
</tr>
<tr>
<td>Regression-adjusted estimation of treatment effect on AOD cost increment per month (95% CI)</td>
<td>-50</td>
<td>-47</td>
<td>-76</td>
<td>-58</td>
<td>-57</td>
</tr>
</tbody>
</table>

a Use 0.25 of the s.e. of the treatment group “p_score” to construct the radius.

b Use 0.20 of the s.e. of the treatment group “p-score” to construct the caliper.
<table>
<thead>
<tr>
<th>Method</th>
<th>NN</th>
<th>OPM</th>
<th>RDM (^a)</th>
<th>OFM</th>
<th>OFM w/ Caliper (^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1 or 1: n w or w/o replacement</td>
<td>1:1</td>
<td>1:1</td>
<td>1: n</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>Regression-adjusted Treatment effect increment per time point (95% CI)</td>
<td>0.97 (0.85, 1.10)</td>
<td>0.97 (0.85, 1.10)</td>
<td>1.03 (0.74, 1.12)</td>
<td>1.00 (0.87, 1.12)</td>
<td>1.00 (0.87, 1.12)</td>
</tr>
<tr>
<td>Percentage Bias (%)</td>
<td>2.70%</td>
<td>2.71%</td>
<td>2.63%</td>
<td>0.42%</td>
<td>0.44%</td>
</tr>
<tr>
<td>Coverage Rate (%)</td>
<td>94.1%</td>
<td>93.8%</td>
<td>82.5%</td>
<td>96.2%</td>
<td>96.5%</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.07</td>
<td>0.07</td>
<td>2.56</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>missInfo Rate (%)</td>
<td>18.89%</td>
<td>17.98%</td>
<td>24.46%</td>
<td>17.33%</td>
<td>17.56%</td>
</tr>
</tbody>
</table>

\(^a\) Use 0.25 of the s.e. of the treatment group “p_score” to construct the radius.

\(^b\) Use 0.20 of the s.e. of the treatment group “p-score” to construct the caliper.
Table 16: Comparison of estimated treatment effects (true parameter=1) from different matching methods for simulated longitudinal data with missing value using incorrect model (simulation times=1000, missing rate = 0.3)

<table>
<thead>
<tr>
<th>Method</th>
<th>NN</th>
<th>OPM</th>
<th>RDM a</th>
<th>OFM</th>
<th>OFM w/ Caliper b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1 or 1: n</td>
<td>1:1</td>
<td>1:1</td>
<td>1: n</td>
<td>1:1</td>
<td>1:1</td>
</tr>
<tr>
<td>w or w/o replacement</td>
<td>w/o</td>
<td>w/o</td>
<td>w</td>
<td>w/o</td>
<td>w/o</td>
</tr>
<tr>
<td>Regression-adjusted Treatment effect increment per time point (95% CI)</td>
<td>1.29 (0.95, 1.74)</td>
<td>1.30 (0.95, 1.76)</td>
<td>1.20 (-0.77, 3.23)</td>
<td>1.14 (0.92, 1.33)</td>
<td>1.14 (0.92, 1.33)</td>
</tr>
<tr>
<td>Percentage Bias (%)</td>
<td>28.54%</td>
<td>29.54%</td>
<td>19.75%</td>
<td>13.62%</td>
<td>13.67%</td>
</tr>
<tr>
<td>Coverage Rate (%)</td>
<td>83.4%</td>
<td>80.6%</td>
<td>99.7%</td>
<td>84.8%</td>
<td>85.0%</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.63</td>
<td>0.57</td>
<td>9.89</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>missInfo Rate (%)</td>
<td>34.95%</td>
<td>31.85%</td>
<td>47.26%</td>
<td>20.97%</td>
<td>20.99%</td>
</tr>
</tbody>
</table>

a Use 0.25 of the s.e. of the treatment group “p_score” to construct the radius.

b Use 0.20 of the s.e. of the treatment group “p-score” to construct the caliper
Figures

Nearest Neighbor

Radius matching

Figure 1: Examples of two matching methods using NN and RDM
Figure 2: Graph for example 1
distance matrix 1:
Manhattan distance
\[ ||x_{j1} - x_{j0}|| + ||y_{j1} - y_{j0}|| \]

distance matrix 2:
Euclidean distance
\[ \sqrt{(x_{j1} - x_{j0})^2 + (y_{j1} - y_{j0})^2} \]

Table A: Different matching results based on Manhattan distance.

<table>
<thead>
<tr>
<th>Matching Algorithm</th>
<th>NN</th>
<th>OFM</th>
<th>OPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result</td>
<td>[BD(3), AC(7)]</td>
<td>[BC(4), AD(6)]</td>
<td>either</td>
</tr>
</tbody>
</table>

Table B: Different matching results based on Euclidean distance.

<table>
<thead>
<tr>
<th>Matching Algorithm</th>
<th>NN</th>
<th>OFM</th>
<th>OPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result</td>
<td>[BD(2.2), AC(5)]</td>
<td>[BD(2.2), AC(5)]</td>
<td>[BD(2.2), AC(5)]</td>
</tr>
</tbody>
</table>

Figure 3: Matching results based on different distance measures
Figure 4: Flowchart of OFM algorithm
Figure 5 Balance assessment of “p_score” distributions between treatment and control groups before / after matching (from different matching algorithms) for “Lalonde” data.
Figure 6: Balance assessment of “p_score” distributions between treatment and control groups using OFM within caliper for “Lalonde” data
Figure 7: Histograms of absolute distance between treatment and control units of each matched pair for "Lalonde" data
Figure 8: Back on back histograms of p-scores between engagement and control groups before / after matching (from different matching algorithms) for Management Care Study.
Figure 9: The distribution of absolute value of difference of p-score among all matched pairs for Management Care Study
Figure 10: Quarterly Cost Difference between engagement and control groups for Management Care Study
Appendix

Appendix I: Proof of DI

[Proof] Suppose we have ordered “Non Compatible Pattern Matrix” (NCPM) as following and the elements in region I are all “non-compatible” (NC) values so that we define region I as “rectangular non-compatible pattern (RNCP)”.

\[
\begin{bmatrix}
X & \vdots & Z \\
\vdots & \ddots & \vdots \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
\begin{array}{ccc}
I & p & \vdots & II \\
q & \vdots & \vdots & \vdots \\
III & \vdots & \vdots & IV \\
\end{array}
\end{bmatrix}
\]

where

\begin{align*}
m &= \text{the number of rows of the inspected matrix (i.e. number of rows of NCPM)}; \\
n &= \text{the number of columns of the inspected matrix (i.e. number of columns of NCPM)}; \\
p &= \text{the number of rows from RMP (i.e. number of rows of RNCP)}; \\
q &= \text{the number of columns from RMP (i.e. number of columns of RNCP)}; \\
n_{\text{int}} &= \text{the number of matched pairs the analyst is interested in.}
\end{align*}

The object is to obtain at least “n_int” 1:1 non-replacement matched pairs. There are four occasions we need to take into considerations:

Scenario A: \((m-p) \geq q\) and \((n-q) \geq p\)

In this scenario, region II can at most provide \(p\) matched pairs and region III at most \(q\) matched pairs. So region V need to provide at least \((n_{\text{int}}-(p+q))\) matched pairs and its row and column should meet the following conditions:

\[
\begin{align*}
(m-p) - q &\geq n_{\text{int}} - (p + q) \quad \text{for row} \\
(n-q) - p &\geq n_{\text{int}} - (p + q) \quad \text{for column}
\end{align*}
\]
Simplifying the above conditions, we have \( n_{int} \leq \min(m, n) \)

Scenario B: \((m-p)\geq q\) and \((n-q) \leq p\)

In this scenario, region II can at most provide \((n-q)\) matched pairs and region III at most \(q\) matched pairs. Since region IV has already exhausted all columns so that we only need the matched pairs from region II and III to meet the following criteria: \((n - q) + q \geq n_{int}\). That’s \(n_{int} \leq n\).

Scenario C: \((m-p) \leq q\) and \((n-q) \geq p\)

Use the similar analysis as scenario B, we have \(n_{int} \leq m\).

Scenario D: \((m-p) \leq q\) and \((n-q) \leq p\)

Use the similar analysis as scenario B, we have \(n_{int} \leq (m - p) + (n - q)\).

Put in the results from all four scenarios together, we need

\[ n_{int} \leq \min\{[(m - p) + (n - q)], m, n\} \]
Appendix II: Proof of Criterion I

If 1\textsuperscript{st} ~ i\textsuperscript{th} minimums occupy j columns, we need at least (n\_int-j) entries for (i+1)\textsuperscript{th} minimum. Otherwise, “NEMP”

[Proof] Suppose after 1\textsuperscript{st} ~ i\textsuperscript{th} minimum iteration, we have j columns occupied and for (i+1)\textsuperscript{th} minimum, we have less than (n\_int-j) units. Suppose we only have (n\_int-j-k) units (k>0), then we have p=(m-(n\_int-j-k))=(m-n\_int)+(j+k) rows having all “NC” from the (i+1)th minimum iteration (i.e. for q=(n-j) columns).

So p+q=(m-n\_int)+(j+k)+(n-j)=(m+n)-n\_int+k>=(m+n)-n\_int \text{. We have problem of “NEMP”}.