Alma Mater Studiorum - Università di Bologna

Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XXI ciclo

A partial dependence factorial analysis to deal with selection bias in observational studies

Ida D'Attoma

Dipartimento di Scienze Statistiche "P. Fortunati"

Marzo 2009

Alma Mater Studiorum - Università di Bologna

Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XXI ciclo

A partial dependence factorial analysis to deal with selection bias in observational studies

Ida D'Attoma

Coordinatore
Prof.ssa Daniela Cocchi

Tutor
Prof. Furio Camillo

Settore Disciplinare
SECS-S/03

Dipartimento di Scienze Statistiche "P. Fortunati"

Marzo 2009

Preface

This thesis presents a creative and practical approach to dealing with the problem of selection bias.

Selection bias may be the most important vexing problem in program evaluation or in any line of research that attempts to assert causality. Some of the greatest minds in economics and statistics have scrutinized the problem of selection bias, with the resulting approaches-Rubin's potential outcome approach or Heckman's selection model-being widely accepted and used as the best fixes. That said, these solutions to the bias that arises in particular from self selection are imperfect, and many researchers, when feasible, reserve their strongest causal inference for data from experimental rather than observational studies.

The innovative aspect of this thesis is to propose a data transformation that allows measuring and testing in an automatic and multivariate way the presence of selection bias.

Specifically, the approach involves the construction of a multi-dimensional conditional space of the X matrix in which the bias associated with treatment assignment has been eliminated. This approach could be considered as a data pre-processing that allows us to measure selection bias in terms of variability of the original X-space that has been eliminated.

At the same time, this procedure allows testing if the balancing property is satisfied after a matching procedure or when the propensity score is used, by preserving the multivariate nature of data.

Further,we propose the use of a clustering procedure as a tool to find groups of comparable units on which estimate local causal effects, and the use of the multivariate test of imbalance as a stopping rule in choosing the best cluster solution set.

The method is non parametric and does not depend on knowing or estimating the propensity score.

The proposed approach does not call for *modeling* the data, based on some underlying theory or assumption about the selection process, but instead it calls for using the existing variability within the data and letting the data to speak.

The idea of proposing this multivariate approach to measure selection bias and test balance comes from the consideration that in applied research all aspects of multivariate balance, not represented in the univariate variable-by-variable summaries, are ignored.

Analysis have been obtained using the statistical softwares Spad and Sas. The remainder of this thesis presents the new approach, first by discussing our underlying paradigm,

then by explaining the problem of causal inference with attention to existing methods in dealing with it and by discussing when assumptions behind conventional methods break down; finally, by describing the proposed method theoretically and empirically.

Structure of thesis

An introduction to evaluation methods as part of public and private decision process with some considerations on the role of data mining is contained in *chapter 1.* The aim is to contextualize in a statistical vision the fundamental problem of causal inference in the presence of non-experimental data and clarify the perspective of data mining.

Chapter 2 concerns conventional statistical tools used in the evaluation context to draw causal inferential conclusions with particular attention to the Potential Outcome Approach (Rubin; 1983,1984,1988). The aim is to give an idea to the reader about the state of literature in the evaluation context by explaining methods used as the best fixes. This chapter represents a starting point for then highlight where these methods should break down (Chapter 3) if the assumptions on which they are based are not carefully checked.

Chapter 3 describes when conventional methods break down, with particular attention to the problem of how testing in the correct way balancing. We aim at highlighting the lack of a multivariate test of balancing in literature. Here we will discuss also remedies that have been proposed to address the resulting problems.

Chapter 4 contains the original contribution. This part presents theoretically the new multivariate approach. Particulary, we propose the use of a partial dependence analysis of the X-space as a tool for investigating the dependence relationship between a set of observable covariates X and a treatment indicator variable T in order to obtain a measure of imbalance according to their dependence structure. Then we propose an operative use of the method.

Chapter 5 aims at testing the new multivariate test of imbalance via simulated data. We check the performance of the method for a given dependence setting, in order to show some of its essential aspects.

Chapter 6 is dedicated to the application of the new method to a real data set. Particularly, we analyze the impact of PSA programs on the variation of the number of employees of handicraft firm in Tuscany region.

Acknowledgments

First of all, I would like to express my sincere gratitude to my advisor professor Furio Camillo, who inspired my work since the beginning, for his continuous support during the PhD program. He provided me with many helpful suggestions and important advices.

A special thanks goes to my co-advisor, professor Laura R.Peck, who is most responsible for helping me complete my knowledge of program evaluation. I am also grateful to Laura for giving me the possibility of attending the workshop on Quasi-Experimental Design at Northwestern University in Chicago. There, fundamental was the help of professors Tom Cook and William Shadish who gave me many suggestions. They showed me different ways to approach my research problem.

I am also grateful to Caterina, who was always ready to listen and give me advice.I thank her for helping me in data simulation. Without her guidance, may be I could not have finished this dissertation.

Furthermore, special thanks to Professor Fabrizia Mealli for providing me data on which test my approach.

I especially thank my colleagues at School Of Public Affairs of Arizona State University; particularly, Khalid Al-Yahya. I thank my friends Nora and Fasial for being in that period the surrogate family.

I am also grateful to all my Ph.D colleagues, with whom I have shared my Ph.D. experience: Massimo, Elena, Clarissa, Christian, Davide, Serena, Anna. I cannot forget to especially thank Anna, who has been my Ph.D. colleague but also a close friend. I thank her for her care, attention, support and for the great time I had with her.

My thanks to all my friends, my colleagues, my brother, my sister in law. I enjoyed their friendship, their support and their love.

Finally, I am forever indebted to my parents for giving me life in the first place, for their love, unconditional support and encouragement to pursue my interest when it was most required.

Last but not least, my boyfriend, who made me an happy person and gave me the extra strength, motivation and love necessary to get this done, especially in the final part of my Ph.D program. He had confidence in me when I doubted myself. I thank him for his continuous understanding and endless patience.

Ida D'Attoma

Bologna, March 17th 2009

Contents

Chapter 1

Data mining for causal inference

1.1 Introduction

In this chapter we first give an overview of fields of application in which instruments and methods for drawing causal conclusions are applied. Then, we underly the usefulness of evaluation methods for both public and private setting. Finally, we explain the fundamental problem of causal inference in observational studies and we introduce our underlying paradigm in proposing the use of data mining as an automatic instrument to detect selection bias and test the balancing property.

1.2 Evaluation methods and observational studies

The availability of information concerning the processes aimed at monitoring the activities of bodies, institutions, private and public companies has over the last decades increased. This phenomenon led to the proliferation of semi-automatic control processes which largely rely on the advances made by information technology and on the development of statistical techniques that are peculiar of modern data mining.
On many different fields, new demands arise of an evaluation of the impacts that large-scale actions and policies generate on the various stakeholders, users or managers, involved in the production of goods or services. Reference is made to the evaluation of the impact of social or economic policies on the individual citizens or businesses. The modern dataflow within the organizations has turned the monitoring processes into a step of ordinary production process. The assessment of the validity of the actions which are developed and implemented becomes part of tools which are available to private or public decision-makers. The evaluation process is, by definition, a *scientific* process which takes place following a large-scale action and provides a retrospective view of the events.
Bingham and Felbinger(2002) refers to the evaluation of agency programs or legislative policy as the use of scientific methods to estimate the successful implementation and resultant outcomes of programs or policies for decision-making purposes.

Evaluation methods represent a field in continuous development. The literature on evaluation methods is vast in many area such us economics,educations, and since few years also in marketing. Rubin and Watermann(2006), for example, presented the results of a project for a major pharmaceutical company concerned with their marketing interventions with doctors for the purpose of promoting a doctor to describe the details of the drug. The causal question they would like to answer is if the marketing intervention causes a difference. In particular,if the number of scripts written after being visited is more than the number of scripts without being visited. Other examples in the marketing field are in Schonlau, Soest, Kaptevn and Couper (2006), Mizuno and Hoshino (2006), Wangenhein and Bayòn (2007), Tripathi (2007), Mithas, Almirall and Krisnan (2006). The causal question in all considered fields is similar. In the program evaluation setting, for example, researchers need to know if social programs work; whereas, in marketing context, one of the fundamental question is if marketing interventions cause one-to-one marketing effectiveness; where marketing effectiveness concerns any change or improvement in a well defined target variable.

Further, all these fields used to work with observational data, where the lack of randomization represents the main characteristic.

Literature refers to several types of evaluations: process evaluations, impact evaluations, policy evaluations, meta-evaluations [1]. The focus here is only on *impact evaluations*. In fact, we will focus on the end results of programs or, more generally, of an action.

Typically we are interested on measuring outcomes by answering the question *What would have happened to target population in terms of outcomes in the absence of the program or of the specific action?*

In answering to this kind of questions researchers agree in considering the randomized experiments the *Gold Standard*; but in social, economic and marketing fields randomized experiments are not feasible, due to ethical considerations (for example, when treatment cannot be denied to needed units), budget constraints, and to retrospective nature of analysis, that is evaluation usually tends to occur after a program was in place. For example, it may happen that the treatment have already been implemented before researcher designed the study, or laws may entitle eligible participants to a treatment so that placing them in a control group at random is not legal.For all that reasons, in program evaluation randomization is uncommon.

When randomized experiments are infeasible, the logic reference framework is one well known in literature as quasi-experiments or observational studies. [2] Even in the presence of observational data, the purpose is to test causal hypotheses about a manipulable cause. Observational data do not represent a problem at all: on one hand,they lack random assignments,given that units self-select into treatments or are selected non randomly to receive treatment by an administrator; on the other hand, observational studies can have

[1] see Bingham and Felbinger, 2002 for more details

[2] Rosenbaum (1995)and Cochran (1965) refer to these as observational studies; Campbell and Stanley (1963) refer to these as quasi-experiments.

desirable features: study conditions may be more representative of real-world setting than randomized experiments to the extent that the latter use, for example, less representative participants, such as volunteers, or less representative setting, such as sites willing to accept random assignment (Luellen, Shadish and Clark, 2005).

1.3 Causal inference: basic concepts and the fundamental problem

Typically, to draw causal inferential conclusions, data and auxiliary information are used to learn what might happen if there is an intervention in some social, biological, physical or other kinds of process. In the last three decades statisticians have developed and applied, with the resulting success, a variety of formal frameworks for manipulating causal concepts and conducting causal inference. A concept common to all framework is that *Correlation does not prove or does not imply causation.* Correlation may occur when is not clear which variable come first and if alternative explanations for the presumed effect exist: that is a relationship may not be causal at all rather due to a third variable called confound.
As the philosopher John Stuart Mill formalized, a causal relationship exists if:

- The cause precede the effect
- The cause was related to the effect
- Researchers are able to rule out all plausible alternative explanations for the effect other than the cause.

The three characteristics of a casual relationship are usually matched by experiments in which researchers are able to manipulate the presumed cause and observe an outcome afterward; they can see whether variation in the cause is related to variation in the effect and they use various methods during the experiment to reduce the plausibility of other explanations for the effect.
Experimental data meet all this; whereas observational data are problematic on the third criterion when is difficult to make most other causes less likely.
A *cause* is viewed as a manipulation or treatment that brings about a change in the variable of interest, compared to some baseline, called the control(Cox,1992;Holland,1986).
After defining the outcome variable and the cause both measured at a unit level, the goal is to consider how the unit would be different if the cause is altered. Neyman (1923) and later Holland(1986), Rubin(1986), Rubin and Waterman (2006) formalized the definition of causal effects as the characterization of two different potential outcomes, one that would be observed with the interventions and one without the intervention.
Literature, generally refers to what would be observed as the *Counterfactual* .
In the definitions above emerge that causal effects are always comparative. It also follows that causal effects are defined in a *what if* manner and, as such, are hypothetical. In fact

the *what if* cannot be directly observed. The *what if* logic, formalized as the *Neyman-Rubin Model*, is pervasive in statistical discussions of cause and effect; but it is also pervasive within areas such as econometric and epidemiology. All these fields totally accept the potential outcome conception of causal inference as the best fixe(see for example Hoffer, 2005a; Winship and Morgan, 1999).
More formally, with the observed intervention as t that equals 1 if a unit gets assigned to treatment and 0 if not; t^* as the hypothetical intervention that equals 1 if the unit hypothetically gets assigned to intervention and 0 if not; with Y as the outcome of interest, the question is what would have happen as a result of intervention compared with the hypothetical intervention that has not actually introduced. The individual causal effect of interest (eq. 1.1) is defined as the difference between the potential outcomes.

$$\tau_i = [(Y(1)_i \mid t^* = 1] - [Y(0)_i \mid t^* = 0)] \tag{1.1}$$

where $Y(1)$ represents the potential outcome under treatment and $Y(0)$ the potential outcome under control. The causal effect of interest in 1.1 could never be directly observed given that after the experiment researchers can observe only one of the two potential outcomes. Berk(2004) has considered four possible pairing between the intervention that was received and the hypothetical intervention:

1. $Y(1) \mid (t^* = 1, t = 1)$:the outcome if, hypothetically,treatment were received and it actually was received

2. $Y(1) \mid (t^* = 1, t = 0)$:the outcome if, hypothetically,treatment were received and it was actually not received

3. $Y(0) \mid (t^* = 0, t = 1)$:the outcome if, hypothetically,treatment were not received but it actually was received

4. $Y(0) \mid (t^* = 0, t = 0)$:the outcome if, hypothetically,treatment were not received and it was not actually received

	$Y(1) \mid t^* = 1$	$Y(0) \mid t^* = 0$
$t = 1$	observable	Missing counterfactual
$t = 0$	Missing counterfactual	observable

Table 1.1: Observed and Missing Data in the Potential Outcome Framework

As shown in table 1.1 the fundamental problem of causal inference is essentially one of missing data. Given that researchers cannot observe the same unit under both treatment and control states, it becomes impossible to observe the causal effect of the treatment T on a specific unit i. As a consequence, researchers focus their attention on estimating

the average causal effect which is made particularly problematic when the assignment-to-treatment mechanism is not random and each potential outcome could belong to a different population. We define the average causal effect as in equation 1.2

$$E[\delta_i] = E[Y_i(1)] - E[Y_i(0)] \tag{1.2}$$

with $E[.]$ denoting the expectation operator. This changing of interest from individual level to average level is known as statistical solution [3]. In the example depicted in table 1.2 the highlighted cells represent the observed outcome and the remaining cells represent what we cannot directly observe. If we consider the mean of each potential outcome based

individual	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$	treatment
1	1	**6**	-5	$t = 0$ (operation B)
2	3	**12**	-9	$t = 0$ (operation B)
3	**9**	8	1	$t = 1$ (operation A)
4	**11**	10	1	$t = 1$ (operation A)
mean	6	9	-3	
observed mean	10	9	1	

Table 1.2: A teaching example adapted from Rubin(2004)

on the available information, then it seems that treatment A is better than B, because units under operation A will have an outcome one more than units under B. The achieved conclusion could be wrong, because by considering the counterfactual - which exists in some Platonic world - the average of the individual causal effects $(Y_i(1) - Y_i(0))$ favors B, giving an average benefit of three. The misunderstanding mentioned above is due to the fact that the assignment mechanism is not random, and each potential outcome could be related to a different population: the counterfactual component remains. In the chapter 2 we will show some ways in which the counterfactual problem can be addressed.

1.4 Conventional methods vs data mining: underlying paradigm

Although others most certainly provide both a more thorough and more nuanced discussion of the difference between the economic and statistical approaches, our attempt here is to make some observations about the two paradigms, and to discuss the paradigm that underlies our proposed approach to dealing with selection bias. Generally, the economic approach is one that rests on underlying economic theory to drive and test models of economic behavior and phenomena. For dealing with issues of selection bias in program evaluation setting, this generally means modeling the selection process as a function of

[3] Holland,1986

known variables. The persistent imperfection is the omnipresence of "unobservables" that one hopes are sufficiently dealt with by controlling for observables. Researchers acknowledge these shortcomings in their analysis and explore the implications of unobservables on the extent and direction of bias in results.

In contrast, a focus of statistics may be to fit the *best model*, but that model need not necessarily be based on some underlying theory about human behavior. According to Breiman (2001), 98 percent of statisticians engage in a *data modeling culture* that emphasizes model validation through goodness-of-fit tests and residual examination; whereas, the other two percent uses an *algorithmic modeling culture*, where predictive accuracy validates models. The technique examined in this thesis comes from the edge of the statistical perspective - Breiman's (2001) less common paradigm- where a fundamental underlying belief is that any research influence unduly affects the results, such that *multiple solutions* arrive simply by virtue of researchers' choice of model. More precisely, the paradigm we refer to is not only about statistics and economics, but about Breiman's (2001) two different cultures in statistical modeling: data modeling versus algorithmic modeling. The latter of these, the *Data Mining* perspective, can be thought of as *letting the data to speak*. This line of research compels questions about what the model is for a data miner, if the model suits the nature of the data, and if the model can represent correctly the real complexity of the data. Breiman's (2001) work is fundamental to understanding the role and the limitations of data models and the rationale for utilizing, and perhaps even preferring, algorithmic models. He asserts that *Approaching problems by looking for a data model imposes an a priori straight jacket that restricts the ability of statisticians to deal with a wide range of statistical problems.* Conclusions from the data modeling perspective *are about the model's mechanism, and not about nature's mechanism*, such that if a data model *is a poor emulation of nature, the conclusion may be wrong* (Breiman, 2001). If different models *give different pictures of the relation between the predictor and response variable* then *the question of which one most accurately reflects the data is difficult to resolve* and does not help for commercial or policy purposes (Breiman, 2001).

In brief, our underlying paradigm is that the problem at hand should define the approach. In response we propose to follow an algorithmic approach as appropriate to deal with the particular problem of bias in the selection to treatment. With reference to conventional propensity score methods for causal inference, that we will discuss in the next chapter, the subjectivity in choosing which variable and model to use for propensity score estimation, as well as subsequent choices about testing balance and stratifying scores, introduces important yet unnecessary bias into an analysis, which could much preferably be conducted an algorithmic modeling approach. [4] We favor the algorithmic approach because we think that literature on propensity score estimation lacks automatic tools for

[4]Stone et al.(1995) and Luellen, Shadish and Clark (2005), for example, used a classification tree procedure to perform propensity score analysis. From those studies emerged an important future that is the automatic future of the classification trees algorithm in selecting variable for the PS model, in detecting interaction in the data and in the automatic role of tree's terminal node, that eliminated the need to set stratification cut points.

testing if conditions on which propensity score is based hold. With the lack of objective criteria results could be a *multiplicity of good models* with potential for informing wrong decisions. The choice of the best model should be justified according to a rigid, unbreakable criterion, quantitatively defined prior the analysis, that is in a data mining sense the *score function*. *Score functions* quantify how well a model or parameter structure fits a given data set. Without some form of score function, we cannot tell whether one model is better than another or, indeed, how to choose a good set of values for the parameter of the model.
The criterion we propose is a measure of imbalance. Particularly, we propose a multivariate data mining approach to detect and measure the presence of selection bias and establish in an objective way if the analysis (i.e. propensity score or any kind of matching procedure)balances data.

1.5 Some definitions of Data Mining

In the literature on the assessment of the causal effect of interventions, is common to consider data mining as an inappropriate technique. Rubin and Watermann (2006), for example, asserted that *causal effect estimation is not generally accomplished by: regression, data mining, neural nets, CART, support vector machines, random forests, and so on*. They think that the techniques mentioned above could be useful only after the estimation of causal effects. They underly the importance of such techniques in a second stage of the analysis when the aim is, for example, *to classify units into subgroups based on background variables describing types of units, where the subgroups differ by the expected size of their causal effects*. When Rubin and Watermann speak about the inappropriateness of data mining, they refer to predictive data mining methods. A necessary distinction is that between explorative and predictive data mining. The explorative data mining is unsupervised: it uses descriptive algorithm to find structure in the data; whereas, predictive data mining is supervised:it aims to predict as much as possible future data. Here we consider descriptive DM as springboard. We consider the descriptive data mining in terms of the French School of *Analyse des données*, which was the first in dealing with statistical analysis by using statistical software.[5] Particularly, here we use DM as a powerful tool to check balance in a multivariate and automatic way, when we have not idea about the presence and the amount of selection bias.
To highlight the power of DM we refer to the definition of Hand et al. (2001). They define data mining as the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The definition above refers to *observational data* as opposed to *experimental data*. In fact, data mining typically deals with data that have already been collected for some purposes other than the data mining analysis. This means that

[5] L'Analyse des données est un outil pour dégager de la gangue des données le pur diamant de la véridique nature (J.P.Benzecri, 1973

the objectives of the data mining play no role in the data collection. This is one way in which data mining differs from much of statistics, in which data are often collected by using efficient strategies to answer specific questions. This characteristic of data mining matches the retrospective nature of evaluation analysis, where the rule of how units get assigned to treatment is lost, and researchers often don't know anything about the selection or self-selection mechanism.

In fact, we aim at discovering not a priori known dependence between observed covariates (potentially involved in the assignment mechanism) and treatment assignment variable.

Data mining is a process that aims to seek relationships within a data set involving a number of steps, such as: determining the nature and structure of the representation to be used, deciding how to quantify and compare how well different representations fit the data (that is, choosing a *score* function),choosing an algorithmic process to optimize the score function. That process could ensure objectivity in results, being not in contrast with statistics. In fact, it is an interdisciplinary process: statistics, database technology, machine learning, pattern recognition, artificial intelligence, and visualization, all play a role. On one hand, statistical techniques alone may not be sufficient to address some of the more challenging issues in data mining. On the other hand, statistics plays a very important role in data mining as a necessary component.

Usually, DM is helped by traditional statistical tools involving multivariate analysis such as: classification, clustering, contingency table analysis, principal component analysis, and so on. In understanding the role of both statistics and DM as an interdisciplinary process we refer to a special contribute of professor Bozdogan , one of the best mind of the DM fields. H. Bozdogan conied the term *Statistical Data Mining*. In the book edited in 2004 he defined statistical data mining as *the process of selecting and exploring large amounts of complex information and data using modern statistical techniques and new generation computer algorithms to discover hidden patterns in the data* . This definition is the testimony that statistics is undergoing a fundamental transformation and it is in an evolutionary stage (H.Bozdogan, 2004).

Researchers should consider that with high dimensionality and different data types traditional statistical methods at all are not sufficient.

A key difference between statistics and data mining is also related to the role of model. In economics and statistics the model follows the theory; whereas in a data mining setting, model follows the data exploration. More precisely, in a statistical setting questions comes before data and in DM setting data comes before questions.

Another important definition is that of U.M.Fayyad and G.Piatetski-Shapiro. They defines data mining as the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data. Based on their definition researchers should find structure from data.

Other definitions of data mining are from software companies which emphasize its practical usefulness in helping companies in solving their business problems, given that DM

provides information that helps business improve marketing, sales and customer services. From the definitions introduced above it emerges that a more general objective of a data mining process, that has some analogies with the aim of the evaluation process, is to describe the general process by which the data arose.

Further, it emerges that the usefulness of DM rather than statistics alone depends on the size of data sets and the curse of dimensionality. For example, in dealing with exact matching methods in finding clones, when units are not assigned randomly to the treatment, the main problem is represented by the curse of dimensionality: the exponential rate of growth of the number of unit cells in a space as the number of variable increases. Data mining could be useful in dealing with this issue because many data mining techniques are based on multi-dimensionality reduction methods and find groups based on similarity or distance measure between objects.

Finally, an other important characteristic of Data Mining is that it is an *automatic process.*This characteristic matches policy-makers requirements of automatic processes in answering to evaluation questions.

Chapter 2

Conventional statistical tools for causal inference in observational studies

2.1 Introduction

This chapter aims to give an idea to the reader about the state of literature for what concerns statistical tools used in the evaluation context in dealing with the causal effect estimation. The attention is focused on statistical techniques for solving and addressing the *counterfactual problem* in order to estimate the effect of an intervention on a well defined target variable of interest, when randomized experiments are infeasible. Our concern here is with the evaluation of an intervention at the individual level. At the heart of this kind of intervention evaluation is a missing data problem since, at any moment in time, a unit is either in the treatment state under consideration or not, never both. Thus, constructing the counterfactual is the central issue that the conventional evaluation methods we discuss address. We will briefly take into account various conventional methods (randomized experiments, propensity score, matching methods, the economic approach), but we will especially focus our attention on *The Potential Outcome Approach*, pioneered principally by Rubin (1974;1978). Implicitly, each approach provides an alternative way of constructing the counterfactual; different are also the assumptions on which they are based and different the methods to check if that assumptions are satisfied. At the same time, this chapter represents a starting point for then highlight some important drawbacks of the conventional methods (chapter 3) that have motivated our research project (chapter 4).

2.2 Experimental Data

Randomization represents one solution to the evaluation problem. Randomized experiments provide the missing counterfactual by ruling out the selection bias as units are

randomly assigned to the treatment state. Based on the definition of Shadish et al.(2002) random assignment means any procedure that assigns units to conditions based only on chance, where each unit has a non zero probability of being assigned to a condition. It doesn't mean that every unit have an equal probability of being assigned to conditions. Due to random assignment to treatment, the treated and control groups are drawn from the same population; thus, the estimator defined as in eq.2.1

$$\tau = E(Y_i^1) - E(Y_i^0) \tag{2.1}$$

will be an unbiased estimator of the average treatment effect; where i index the population under consideration, Y_i^1 the value of the variable of interest when unit i is subject to the treatment $t = 1$ and Y_i^0 is the value of the same variable when the unit is exposed to the treatment state $t = 0$. One simply compares the experience of the treated group with that of the untreated group; the effect in equation 2.1 is estimable in an experimental setting because observations in treatment and control groups are exchangeable. In fact, by design, the experiment will be independent of any kind of influence on outcome Y whether observed or unobserved. In fact, due to condition 2.2

$$Y_i^1, Y_i^0 \perp t_i \tag{2.2}$$

randomization equates groups on expectation: as the sample size grows, observed and unobserved baseline variables are balanced across treatment and control groups. The assumption 2.2 implies that for $j = 0, 1$

$$E(Y_{ij} \mid t_i = 1) = E(Y_{ij} \mid t_i = 0) = E(Y_i \mid t_i = j) \tag{2.3}$$

and

$$\begin{aligned} \tau &= E(Y_{i1} \mid t_i = 1) - E(Y_{i0} \mid t_i = 0) \\ &= E(Y_i \mid t_i = 1) - E(Y_i \mid t_i = 0) \end{aligned} \tag{2.4}$$

In an observational setting, researchers are usually interested on the treatment effect on the treated (eq.2.5)

$$\begin{aligned} \tau_{|t=1} &= E(\tau_i \mid t=1) \\ &= E(Y_i^1 \mid t=1) - E(Y_i^0 \mid t=1) \end{aligned} \tag{2.5}$$

but, without a random assignment to treatment,is possible to estimate $E(Y_i^1 \mid t_i = 1)$ and not $E(Y_i^0 \mid t_i = 1)$.

It clearly emerges that with randomization, covariates play no role in the estimation of treatment effects given that random assignment breaks the link between T and X. A different way to consider causal inference in Classical Randomized Experiments is through the potential outcomes notation introduced by Rubin.

The potential outcome notation implies that for each unit exist two potential outcomes $Y(1)$ in the presence of treatment and $Y(0)$ in the absence of treatment, even if each unit is observed in only one treatment state. Rubin (2005) takes into account three modes of causal inference in Classical Randomized experiments; one is Bayesian, which treats the potential outcomes as random variables, and two are based only on the assignment mechanism, which treat the potential outcomes as fixed but unknown quantities(Neyman,1923; Fisher, 1925). Each mode of inference shares a common framework that requires the consideration of a posited assignment mechanism.

For what concerns the Fisher's mode of inference an important aspect is represented by the *null hypothesis*, which is $Y(1) \equiv Y(0)$ for all units. Under this null hypothesis, all potential outcomes are known from the observed outcome Y_{obs} because $Y(1) \equiv Y(0) \equiv Y_{obs}$. It follows that, under this null hypothesis, the value of any statistics, S, such as the difference of the observed averages for units exposed to treatment 1 and units exposed to treatment 0, $\overline{y}_1$ - $\overline{y}_0$, is known not only for the observed assignment, but for all possible assignment T.

Thus, it is possible to calculate a significance level in order to assess how unusual the actual observed statistic is relative to all possible values of that statistic that might have been observed with these units.

Neyman's form of randomization-based inference can be viewed as drawing inferences by evaluating the expectations of statistics over the distribution induced by the assignment mechanism in order to calculate a confidence interval for the typical causal effect. In particular, an unbiased estimator of the causal estimand is created, and an unbiased estimator of the variance of that unbiased estimator is found. The *causal estimand* is the average causal effect $\overline{Y(1)}$ - $\overline{Y(0)}$, where the averages are over all units in the population being studied, and the traditional statistic for estimating this effect is the difference in sample averages for the two groups, $\overline{y}_1$ - $\overline{y}_0$, which can be shown to be unbiased for $\overline{Y(1)}$ - $\overline{Y(0)}$ in a completely randomized design.

Despite in addressing the causal effect estimation randomized experiments are considered the gold standard, the deriving causal effect estimation can be invalidated by some aspects:

such us, for example, dropout or partial compliance. Dropout implies that some units of treatment group does not receive the treatment, and, as a consequence, the experimental mean difference estimates the intent to treat and not the treatment effect. To obtain an estimate of the impact additional assumptions are required (see, for example, Heckman, LaLonde and Smith, 1999).

2.3 The potential outcome approach

This section gives an overview of the *Potential Outcome Approach* following the papers of Holland and Rubin (1988);Rubin (1991;2001;2004;2005;2007); Rosenbaum and Rubin (1983;1984);Holland (1986); Rubin and Waterman (2006); Frangakis and Rubin (2002); and the book of Rubin (1987).

From literature, we know that the Potential Outcome framework is principally due to Rubin, but the formal notation dates back to Neyman (1923). He was the first writer to use the potential outcome notation for randomized experiments. Only Rubin extended that notation to describe causal effects in non-randomized studies in 1974. The extension of Neyman's potential outcome notation to define causal effects in both non-randomized and randomized studies is called *Neyman-Rubin* model (Pearl, 1996) or the *Rubin Causal Model* (RCM) (Holland, 1986). The RCM is a counterfactual model of causation: it moves from the idea that much of researchers knowledge of causal effects in the evaluation context must come from non-randomized observational studies given the infeasibility of randomized experiments. Aiming at measuring a causal effect observational studies should be designed to approximate randomized experiments as closely as possible. It means that, with an observational data set, data should be conceptualized as having arisen from an underlying regular [1] assignment mechanism.

In particular, an observational study in the Rubin perspective, is conceptualized as a *broken randomized experiment*, in the sense that observed data are considered as *having arisen from an hypothetical complex randomized experiment with a lost rule for the propensity score, whose values we will try to reconstruct.* The RCM shares with randomized experiments an important future: the analysis in both randomized experiments and observational studies takes place before seeing any outcome data. The RCM is composed of two essential elements:

1. *the definition of science*: it represents the conceptual part defined before seing any data.

2. *the assignment mechanism*: a probabilistic model for the treatment each unit receives as a function of observed covariates and potential outcomes.

These two parts are fundamental in the design stage of an observational study, where *by design* Rubin means the collection, organization and analysis of data that takes place prior

[1] unconfounded and probabilistic

to seeing any outcome data in order to achieve objectivity in results.

2.3.1 *The definition of science*

In this framework what Rubin calls **The Science** is unaffected by whether researchers try to learn about a causal effect of interest (via experiments or observational studies or any kind of analysis). The Science represents the state of art before any causal analysis and is defined as composed of the following elements:

- the **units** of study
- the **treatments** (interventions, real or hypothetical)
- the **covariates** (i.e. background variables) that are presumed to be unaffected by the treatments
- the **potential outcome** variables

Table 2.1 summarizes the potential outcome notation. The N units i are considered as physical objects at a particular time t. The *treatment* is an action, or an intervention the effects of which the researcher wishes to asses relative to no intervention. The innovative aspect of this approach is that before an experiment starts, each unit has two potential outcomes: $Y(1)$ given treatment and $Y(0)$ without treatment.
In a philosophical sense both $Y(0)$ and $Y(1)$ are concerned as existing simultaneously, in some Platonic paradise ,even though, in the light of their interpretation, there is no world, actual or conceivable, in which both could be observed (Dawid, 2006), that is the fundamental problem of causal inference (Holland, 1986). The counterfactual is what we cannot observe: the outcome that would have happened $Y(0)$ if the unit had not received the treatment; whereas observed values of the potential outcomes are those revealed by the assignment mechanism.
Then causal effects are defined to be the comparisons of the potential outcomes that would have been observed under different exposures of units to treatments.[2] Finally, covariates $\mathbf{X}$ are variables that take the same value for each unit no matter which treatment is applied to the units, such as quantities measured before treatments are assigned, and as a consequence, the simply cannot be affected by the treatment. More precisely, a *covariate* is a special type of variable for which $X_t(i) = X_i$ for all $t \in T$.

The framework formally described above requires the plausibility of the Stable Unit-Treatment-Value Assumption (SUTVA). SUTVA means that the set of Y(0), Y(1) for each unit fully represents the possible values of the outcome Y under all pairings of $t \in T$ with $i \in N$. The SUTVA assumption comprises two sub-assumptions. First, it assumes

[2] At an average level, for example, the critical requirement is that the causal effect must be a comparison of $Y_i(1)$ and $Y_i(0)$ for a common set of unit (S), such that $\{Y_i(1), i \in S\}$ and $\{Y_i(0), i \in S\}$.

Units	Covariates X	Potential outcomes Treatment Y(1)	Potential outcomes Control Y(0)	Unit-level Causal effects	Summary Causal effects
1	X1	Y1(1)	Y1(0)	Y1(1) vs Y1(0)	Comparison of Yi(1) vs Yi(0) for a common set of units
I	Xi	Yi(1)	Yi(0)	Yi(1) vs Yi(0)	
N	Xn	Yn(1)	Yn(0)	Yn(1) vs Yn(0)	

Figure 2.1: RCM notation

that there is no interference between units(Cox, 1958); that is neither $Y_i(1)$ nor $Y_i(0)$ is affected by what action any other unit received. Second, it assumes that there are no hidden version of treatments. The values of $Y = (Y(1), Y(0))$ are not influenced by T. More precisely, the set of components of (Y(1),Y(0)) we observe is determined by the value of T, but the values of (Y(1),Y(0)) are the same, no matter what the value of T is. Only under stability (SUTVA) each unit has a potential outcome under treatment 1 and another potential outcome under treatment 0. The SUTVA assumption is essential in the sense that it allows to get a causal effect for each unit.

2.3.2 *The Assignment mechanism*

The main future of RCM approach is that it takes into account the assignment mechanism. The aim is to re-construct the missing counterfactual by explicitly defining a formal model for the assignment mechanism, the process that creates missing and observed potential outcomes. The assignment mechanism could be viewed as a *real* or *hypothetical* rule used to assign treatments to the units. With the assignment mechanism vector indicated as:

$$\underline{T} = (T_1, ..., T_i, ..., T_n)^T \tag{2.6}$$

where T_i equals 1 in the presence of an active treatment and equals 0 otherwise,the model for the assignment mechanism gives the probability of the vector T given fixed values of the Science $X, Y(1) and Y(0)$ (2.7).

$$Pr(T \mid X, Y(1), Y(0) \tag{2.7}$$

where the science (X, Y(1), Y(0)) is regarded as fixed and partially revealed by the assignment mechanism. In an observational setting the specification of an assignment mechanism is required in the sense that causal answers generally change if the posited assignment mechanism is changed. The assignment mechanism matters to valid inference:simply comparing observed values under the treatments only work if units are randomly assigned treatments. Without random assignment, given that half the potential outcomes (which define causal effects) are missing, the process that makes them missing must be part of the inferential model.(Rubin,1976,p.581). The most critical template for causal effect estimation from observational data is represented by *regular designs*[3] with unknown propensity scores. These designs are not common in practice because of the need to know all covariates used in the assignment mechanism. But it is possible to draw valid causal inference by assembling data with enough covariates that it becomes possible to claim that the unknown assignment mechanism is unconfounded given these covariates. It means to assume that the treatment assignment is *strongly ignorable* (Rosenbaum and Rubin, 1983). More precisely, an assignment mechanism is *strongly ignorable* when it is regular. A regular assignment mechanism is defined as both unconfounded (2.8) and probabilistic (2.9):

$$Pr(T_i \mid X_i, Y_i(0), Y_i(1)) = Pr(T_i \mid X_i) \tag{2.8}$$

$$0 < p(T_i = 1 \mid X_i) < 1 \tag{2.9}$$

The next step is represented by choosing a model for the unknown assignment mechanism.

2.4 The propensity score methodology

As introduced in the previous section, the main source of selection bias in observational studies is represented by self-selection or some systematic judgment by the researcher in selecting units to be assigned to the treatment. Many recent attempts to address such selection bias have focused on modeling the selection process as a means of removing bias in the estimation of treatment effects. Rosenbaum and Rubin (1983) presented an approach that involves propensity score. The propensity score (PS) represents the model for the assignment mechanism. It is widely applied in various fields such as: education (see for

[3] A regular design is like a completely randomized experiments except that the probabilities of treatment assignment are allowed to depend from covariates and can vary from unit to unit. These designs have two features: the assignment mechanism is unconfounded and they are probabilistic

example, Agodini and Dynarski,2001; Morgan, 2001), social (see for example, Peck and Scott, 2005; Peck [4], 2007), economic-econometric, medical (Connors et Al., 1996; Gum et al., 2001), marketing (see for example, Mithas, Almirall and Krishnan, 2006; Mizuno and Hoshino, 2006; Tripathi, 2007), web survey(S.Lee, 2206), epidemiology (Joffe et al., 1999; Normand et al., 2001). Given a matrix of observed variables, $\underline{X}$, considered as scientific entities, collected before the experiment takes place, the propensity score model allows researchers to reconstruct the missing counterfactual they were looking for, by modeling the selection process that has generated the missing data.The propensity score's knowledge and estimation allow researchers to achieve a randomized experiments approximation, by eliminating only part of selection bias due to the selection mechanism. When the propensity scores for each unit are known, then the assignment mechanism is essentially known, and no fundamental problem of causal inference will exist anymore. Whereas, when the propensity scores are unknown, but the assignment mechanism is *regular*, researchers have to estimate them. Given the estimated propensity scores, units under different treatment's levels, could be compared if their probability to get assigned to one treatment given the covariates is the same. If the assignment mechanism is *unconfounded* (2.8), then no dependence will exist between assignment to treatment and potential outcome. The unconfoundness implies that two subgroups, respectively treated and controls, with the same distribution of the covariates entered in the selection mechanism, will be comparable. Of course, propensity score methods can eliminate the overall *bias*, if the assignment mechanism is really *unconfounded given the observed covariates* X.
Controversy exists on the topic of whether propensity score sufficiently approximates experimental conditions, with some researchers concluding favorably and suggesting some operative procedures (e.g., Becker and Ichino, 2002; Dehejia and Wahba 1999,2002) and others unfavorably (e.g., Agodini and Dynarski 2001; Luellen, Shadish and Clark 2005; Wilde and Hollister, 2002).
Like any probabilities, a PS, defined as in equation 2.10, ranges from 0 to 1.

$$e_i \equiv e(\underline{X}_i) \equiv Pr(T_i = 1 \mid \underline{X}_i) \tag{2.10}$$

Rosenbaum and Rubin (1983) have demonstrated two key properties of propensity score:

- *The covariate balance property.*

They have demonstrated that the propensity scores are *balancing scores* ($b(\underline{X})$).
Balancing scores are function of observed covariates $\underline{X}$ such that the conditional distribution of $\underline{X}$ given $b(X)$ is the same for treated and controls. In particular,the treatment assignment vector and the observed covariates are conditionally independent given the propensity score (2.11)

[4]She uses PS to identify subgroups within both the treatment and control groups of social experiments

$$\underline{X} \perp t \mid e(\underline{X}) \tag{2.11}$$

If the property 2.11 holds, then treatment and control subgroups with the same scalar $e(\underline{X})$ will have the same distribution of all covariates entered in $e(\underline{X})$. Thus, matching on propensity score automatically controls for differences in outcomes between the treated and controls. As a consequence, observed differences in the outcomes cannot be due to those observed covariates.
Of course, propensity score methodology can only attempt to achieve balance in observed covariates whereas randomization in experiments can balance all covariates, both observed and unobserved.

- *The strong ignorability property*

The second property they have demonstrated is the ignorability. In literature, there are different versions of ignorability: unconfoundedness and ignorable treatment assignment (Rosenbaum and Rubin, 1983), selection on observables (Barnow, Cain and Goldberg, 1980), conditional independence (Lechner 1999, 2002), and exogeneity (Imbens, 2004).
Ignorability means that treatment assignment and unobserved potential outcomes are independent, after conditioning on X and the observed potential outcomes. Thus, all unobserved variables could be ignored. In particular, Rosenbaum and Rubin (1983) have demonstrated that if treatment assignment is strongly ignorable given $\underline{X}$, then it is strongly ignorable given $e(\underline{X})$ (2.12).

$$\text{if } (Y_i(1), Y_i(0)) \perp T_i \mid \underline{X}_i \text{ then } (Y_i(1), Y_i(0)) \perp T_i \mid e(\underline{X}_i) \tag{2.12}$$

As a consequence, at any value of a balancing score, the difference between the treatment and control means is an unbiased estimate of the average treatment effect. With strongly ignorable treatment assignment, pair matching on a balancing score, subclassification on a balancing score can all produce unbiased estimates of treatment effects, because by conditioning on observed covariates X_i, treatment and control groups are balanced [5].
Then, when the effect of interest is represented by the Average Treatment Effect (ATT), it could be estimated as in equation 2.13:

$$\tau \mid (T = 1) = E\{[E(Y_i \mid X_i, T_i = 1) - E(Y_i \mid X_i, T_i = 0)] \mid T_i = 1\} \tag{2.13}$$

[5]The most straightforward and non parametric way to condition on X is to exactly match on the covariates

It is possible that the PS approach could fail in achieving a comparison of treated with controls: this may occur when there is little or no overlap. In this circumstance no valid inferential conclusions could be achieved. The interpretation is that the characteristics (as measured by covariates) of the two groups are so dissimilar that no meaningful comparison is possible. We can consider it as an advantage of PS that reveals how much of the data provide information for causal effect estimation. Anyway, for drawing valid causal inference, researchers must check the existence of *overlap* in the estimated propensity scores and diagnostic analysis must be implemented in order to asses the resulting balance of covariate distribution. Unfortunately, as we will show in the next chapter, literature lacks of valid guideline and objective criteria to test balancing property.

2.4.1 A well-known algorithm to estimate PS and subclassification on PS

Rosenbaum and Rubin(1983) have demonstrated some important properties of PS by assuming that PS for each unit was known. In practice, however, propensities are often unknown and researchers have to estimate them. In literature and in practice, were proposed various methods dealing with the propensity score estimation, such as: discriminant analysis, logistic regression, probit, decision trees, classification trees (Stone et al., 1995; Luellen et al., 2005). Dehejia and Wahba (2002) and Beker and Ichino(2002)[6] have suggested an easy algorithm aiming at estimating propensity score.

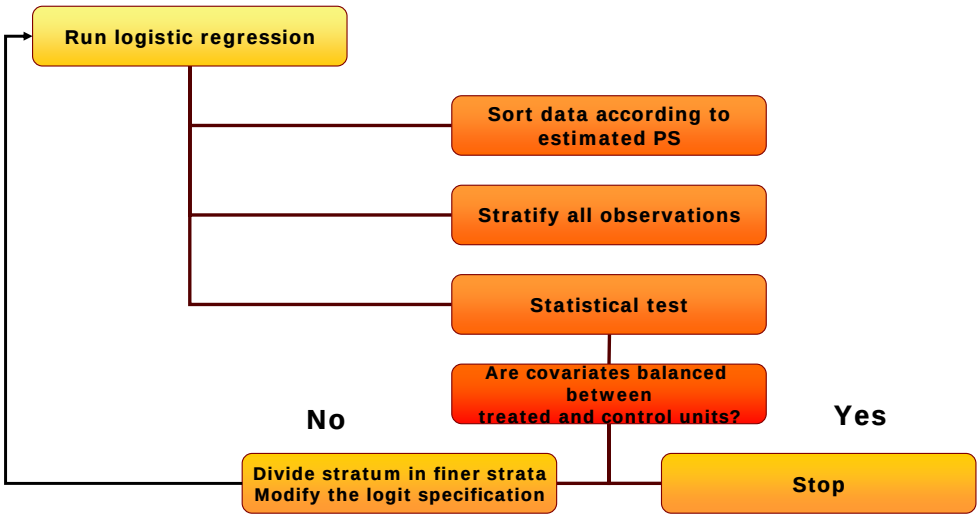

Figure 2.2: PS estimation algorithm

Figure 2.2. summarizes the steps of the algorithm described by Dehejia and Wahba (2002) and Becker and Ichino(2002). We highlight the Becker and Ichino procedure because it is

[6]They also published the stata code to implement the propensity score estimation widely used by evaluators in variuos fields

the widely applied in operative context, but also because it is the algorithm on which the pscore Stata procedure is based.
Usually, a PS analysis starts with the estimation, by probit or logit, of a treatment assignment equation. Logistic regression is the most commonly used method for computing PS: researchers used to start with a parsimonious model specification in which all observed covariates affecting assignment and outcome are included as predictors and the treatment assignment (dummy code, 0/1) as the dependent variable. After propensity scores are estimated for each unit and ranked from lowest to highest, some matching procedures are implemented. The most commonly employed is the subclassification on PS [7], that involves the stratification of all observations such that the estimated propensity scores within each stratum for treated and comparison units are close [8]. Analysts used to divide the PS in 5 bins according to Cochran(1968), who observed that subclassification with 5 subclasses is sufficient to remove at least 90% of the bias. Then the distribution of covariates for treated and controls within each bin are compared and if they still differ, the model specification is further developed, by adding interaction terms and/or higher-order terms of the covariates, until researchers can find a *good* model, where *good* means to achieve balance of ps and covariates within bins. Another method of computing propensity scores involved classification trees algorithms rather than logistic regression. Luellen, Shadish and Clark (2005) have underlined some advantages of classification tree approach: the algorithm automatically selects variables for the model, it automatically detects interactions in the data and tree's terminal nodes automatically supply the researcher with strata, eliminating the need to set stratification cut points.

2.5 Matching methods

Matching techniques have origins in experimental work from the first half of the twentieth century.
In the early 1980s, matching techniques were advanced in a set of papers by Rosenbaum and Rubin (1983a, 1984,1985).
In the late 1990s, economists joined in the development of matching techniques in the course of evaluating social programs (e.g. Heckman, Ichimura and Todd, 1997,1998; Heckman, Ichimura, Smith and Todd, 1998; Heckman, LaLonde, and Smith 1999). [9] Matching is a non parametric method that deals with the selection bias by constructing a comparison group of units with observable characteristics similar to the treated.
The main idea of this method is to replicate the condition of an experiment in the presence of observational data. This is possible by dropping, repeating, grouping observations from an observed dataset in order to reduce covariates imbalances between the treated

[7] *the binning procedure*
[8] When subclasses are perfectly homogeneous in $b(x)$ then $\underline{X}$ has the same distribution for treated and controls in each subclass (Rosenbaum and Rubin, 1983)
[9] For a complete discussion of matching methods, see for example, Greenwood 1945; S.L.Morgan and D.H. Harding, 2006.

and control groups that were not avoided during data collection (no random assignment to treatment).

The ultimate goal of matching is to achieve the best balance for a large number of observations, by pruning observations according to some metric, and by using any method of matching that is function of X without introducing outcome in the analysis. All treated units and matched control units are retained, and all non-matched control units are discarded.

Matching is not a method of estimation, and as a consequence, any application of it must be followed by a simple difference in means or some other method to estimate the causal effect. Matching is the beginning rather than the end of a causal analysis. Many researchers prefer matching to other methods because it allows not statisticians to easy understand the equivalence of treatment and control groups and to perform simple matched pair analysis which potentially adjust for confounding variables.

An important assumption required for all matching methods is the availability of a set of covariates, such that, conditioning on them, potential outcomes are independent of treatment status (2.14):

$$Y(0), Y(1) \perp T \mid X \tag{2.14}$$

We have to distinguish between exact matching on covariates and matching based on propensity score.

Exact matching on covariates represents a valid substitute for the absence of experimental control units. It assumes that a control group can be obtained for a set of potential comparison units, which are not necessarily drawn from the same population as the treated units, but for whom researchers observe the same set of pre-treatment covariates, X_i. Under the matching assumption the only remaining difference between the two groups is the treatment effect. But, it may occur some bias due to incomplete matching (failure to match all treated units) or inexact matching (failure to find exact matches, that is match treated-control pairs with different values of X) (Rosenbaum and Rubin, 1985).

Many researchers agree in considering that one limitation of exact matching on covariates is represented by the dimensionality of the vector of covariates $\underline{X}$. For example, if $\underline{X}$ is n-dimensional and if all n variables are dichotomous, the number of possible values for the vector $\underline{X}$ will be 2^n. Clearly, as the number of variables increases, the number of cells increases exponentially, increasing the difficulty of finding exact matches for each of the treated units, that is the matching problem.

An important distinction is that between exact matching and one-to-one exact matching. The exact matching uses all control units with exactly the same covariate values that match each control unit; whereas one-to-one matching uses only one control unit for each treated unit. The one-to-one exact matching estimates the counterfactual $Y_i(0)$, corresponding to

each observed treated unit i (with outcome value Y_i and covariates X_i) with the outcome value of a control unit (denote Y^*_{match} with covariate values X^*_{match}), chosen such that $X^*_{match} = X_i$ (Imai et al. 2008).Using all exact control matches for each treated unit rather only one, reduces variance without any increase of bias. When exact matching methods are used, some units could not have an exact clone especially when samples are small, variables are measured with many categories and the distribution of participants between groups is uneven (Shadish et al., 2002).
In the presence of problems mentioned above, the choice concerns one of approximate matching methods, that matches the treated unit to some control observations according to some metric.
Examples are the nearest neighbor, Mahalanobis matching, or matching methods based on the estimated propensity score.

2.5.1 Multivariate matching based on Mahalanobis distance

The most common used multivariate matching method is based on Mahalanobis distance (Cochran and Rubin, 1973; Rubin,1979,1980). The mahalanobis distance between any two units is given by:

$$md(X_i, X_j) = \{(X_i - X_j)' S^{-1} (X_i - X_j)\}^{\frac{1}{2}} \tag{2.15}$$

where S is the sample covariance matrix of the matching variables X, X_i and X_j are respectively the multivariate vectors of values of the matching variables for treated unit i and the untreated unit j. Commonly, this matching procedure first randomly orders units, then calculates the distance between the first treated unit and all untreated units. To estimate ATT by matching with replacement, one matches each treated unit with the M closest control units, as defined by this distance measure in equation 2.15. In particular, the untreated unit j, with the minimum distance is chosen as the match for the treated unit i, and both are removed from the pool. The analysis is repeated until matches are found for all treated. Under Mahalanobis distance matching, individual covariates are collapsed into a single scalar metric using Mahalanobis distance, which is defined as the generalization of the standardized distance from the origin of an n-dimensional space to a point where the coordinates represent the X values for a particular observation. Other multivariate matching methods are cited in Shadish et al. (2002): such as benchmark group matching, cluster group matching, index matching. Index matching selects multiple control units above and below a treatment unit; cluster group matching uses cluster analysis to embed the treatment group in a cluster of similar controls; benchmark group matching selects control units that falls close to the treatment unit on a multivariate distance measure. [10]

[10] In Henry and McMillan (1993), we can find a simulation that suggest cluster and benchmark methods may work better than other matching methods.

All multivariate matching start by considering a multidimensional space and aim to find clones similar with respect to their multidimensional characteristics. The main problem is represented by the high dimensionality of the considered space.

2.5.2 Propensity score matching algorithms

A matching estimator non parametrically balances the variables in X_i across T_i with the aim of obtaining the best possible estimate of the causal effect of T_i on Y_i. The most popular technique is to estimate the probability of T_i for each unit i as a function of X_i (i.e. the propensity score) and then to select for further analysis only matched sets of treated and controls that contain units with same values for the propensity scores. In the Rubin perspective, the propensity score could be also seen as a way of reducing a large space of covariates, $\underline{X}_i$, to a one-dimensional summary,the probability of treatment assignment, e_i. [11] Then the use of a one-dimensional summary for matching, when the dimensionality of the matching space [12] is high, could be considered as a key bridge between *matching* and *propensity score.* Propensity score matching is not new in literature and widely applied in various fields. We refers, for example, to papers of Dehejia and Wahba (2002), Rosenbaum and Rubin (1983), Heckman, Ichimura and Todd(1997), Morgan and Harding(2006).
For what concerns the causal effect estimation, propensity score matching could be viewed as a way to *correct* the estimation of the treatment effects controlling for the existence of uncontrolled factors, based on the idea that the bias is reduced when the comparison of outcomes is performed using treated and control cases who are as similar as possible with respect to their estimated propensity score.
According to this perspective, matching methods could be useful in all settings, where needed data are costly to obtain. Matching, more generally, represents an advantages because it allows to obtain the outcome variable of the relevant comparison units, after discarding the irrelevant potential comparison units.
Heckman, Ichimura and Todd (1997,1998) and Smith and Todd (2005) outline a general framework for representing alternative matching estimators.
All matching estimators could be defined as a *weighting scheme*, which determines what weights are placed on comparison units when computing the estimated treatment effect:

$$\hat{\tau}_{|T=1} = \frac{1}{N^T} \sum_{i \in T} [(Y_i \mid T_i = 1) - \sum_{j \in C} \omega_{ij} (Y_j \mid T_j = 0)] \tag{2.16}$$

where N^T is the number of the treatment group, i is the index over treatment group (T), j is the index over control group (C), and ω_{ij} represents a set of scaled weights that measure the distance between each control unit and the treated unit.
For example, exact matching cited in the previous paragraph uses weights equal to $\frac{1}{k}$ for

[11] (Rubin, 2001)
[12] observable characteristic

matched control units, where k is the number of matches selected for each target treatment unit. Weights of 0 are given to all unmatched control units. If only one match is chosen randomly from among possible exact matches, then $\omega_{i,j}$ is set to 1 for the randomly selected match and 0 for all other control units.
The difference in propensity scores is the most common distance measure used to construct weights.
Other measures of distance are available including the estimated odds of the propensity score, the difference in the index of the estimated logit, and the mahalanobis metric. The amount of bias reduction for each matching procedure depends on many aspects: one is represented by whether comparison units are matched with replacement or without replacement.

Matching with replacement minimizes the propensity score distance between the matched comparison units and the treatment unit: each treatment unit can be matched to the nearest comparison unit, even if a comparison unit is matched more than once, with resulting bias reduction.

Matching without replacement may force researchers to match treated to comparison units with a quite different propensity score with resulting increment of bias and improvement of estimation precision.

One simple algorithm to identify the most similar comparison units to be matched to the treated units is the *nearest neighbor matching*,which selects the m comparison units (the clones) whose propensity scores are closest to the treated unit in question, as a result of a distance metric minimization:

$$C(i) = \min_j \|e(i) - e(j)\| \tag{2.17}$$

The traditional algorithm randomly orders the treatment units and then selects for each treatment unit the control unit with the smallest distance. The algorithm can be ran with or without replacement. With Nearest Neighbor Matching all treated units find a match. However, it is obvious that some of these matches are fairly poor because for some treated units the nearest neighbor may have a different propensity score and nevertheless it would contribute to the estimation of the treatment effect independently of this difference.

Another possible procedure is the *radius matching*, which admits for each treated unit to be matched with more than one excluded unit. In this procedure the matches are made only if propensity score falls in a predefined neighborhood of the PS of the treated unit:

$$C(i) = \{p_j \mid \|e(i) - e(j)\| < r\} \tag{2.18}$$

where r is a tolerance level chosen by the researcher. If the dimension of the neighborhood (i.e. the radius or caliper) is set to be vary small it is possible that some treated units are not matched because the neighborhood does not contain control units. On the other hand, the smaller the size of the neighborhood the better is the quality of the matches. Similarly to the matching with replacement this method allows a given excluded unit to be matched more than one time. In both nearest neighbor and radius matching after defining as N_i^C the number of controls matched,the weights as $\omega_{ij} = \frac{1}{N_i^C}$ if $j \in C(i)$ and $\omega_{ij=0}$ otherwise, the matching estimator can be defined as follows:

$$\begin{aligned}
\tau &= \frac{1}{N^T}\sum_{i\in T}[(Y_i \mid T_i = 1) - \sum_{j\in C_i}\omega_{ij}(Y_j \mid T_j = 0)] \\
&= \frac{1}{N^T}[\sum_{i\in T}(Y_i \mid T_i = 1) - \sum_{i\in T}\sum_{j\in C(i)}\omega_{ij}Y_j^C] \\
&= \frac{1}{N^T}\sum_{i\in T}(Y_i \mid T_i = 1) - \frac{1}{N^T}\sum_{j\in C}\omega_j(Y_j \mid T_j = 0)
\end{aligned} \tag{2.19}$$

where $\omega_j = \sum_i \omega_{ij}$. If one wants to use the entire comparison sample, a possible solution is the *kernel matching*. Referring to Heckman, Ichimura, Smith and Todd(1988) and Heckman, Ichimura, and Todd (1997,1998) kernel matching constructs the counterfactual for each treatment case using all control units, but weights each control unit based on its distance from the treatment case. Weights are inversely proportional to the distance between the propensity scores of treated and control. When the estimated propensity score is used to measure the distance, kernel-matching estimators define the weight as:

$$\omega_{ij} = \frac{G[\frac{e(j)-e(i)}{h_n}]}{\sum_{k\in C} G(\frac{e(k)-e(i)}{h_n})} \tag{2.20}$$

where h_n is a bandwidth parameter that scales the difference in the estimated propensity scores based on the sample size, $e()$ is the estimated propensity score, and G is a Kernel function. The numerator of the expression 2.20 yields a transformed distance between each control case and the target treatment case. The denominator is a scaling factor equal to the sum of all the transformed distances across control cases, which is needed so that the sum of ω_{ij} equals 1 across all control cases when matched to each target treatment case. Then the kernel matching estimator is given by equation 2.21:

$$\tau^k = \frac{1}{N^T}\sum_{i\in T}\{Y_i^T - \frac{\sum_{j\in C} Y_j^C G(\frac{e(j)-e(i)}{h_n})}{\sum_{k\in C} G(\frac{e(k)-e(i)}{h_n})}\} \tag{2.21}$$

Under standard conditions on the bandwidth and kernel the estimator 2.22 is a consistent estimator of the counterfactual outcome $Y(0)$.

$$\frac{\sum_{j\in C} Y_j^C G(\frac{e(j)-e(i)}{h_n})}{\sum_{k\in C} G(\frac{e(k)-e(i)}{h_n})} \tag{2.22}$$

Finally, *stratification* divides units into strata so members of the treatment and control groups have similar propensity scores within strata. Rosenbaum and Rubin (1983) suggested using five equal-size strata as a convention. Their choice of five strata is based largely on Cochran (1968), who found that five strata are often sufficient to remove approximately 90% of the bias due to a single continuous covariate. Differences in outcome between the treatment and control group in each interval are then calculated. The average treatment effect is obtained as an average of outcome measure differences per block, weighted by the distribution of treated units across blocks.[13] By construction, in each block defined by this procedure the covariates are balanced and the assignment to treatment can be considered random. Within each block the average treatment effect is then computed as in equation 2.23.

$$\tau_q^S = \frac{\sum_{i\in I(q)} Y_j^T}{N_q^T} - \frac{\sum_{j\in I(q)} Y_j^C}{N_q^C} \tag{2.23}$$

where q index the blocks, $I(q)$ is the set of units in block q, while N_q^T and N_q^C are the number of treated and control units in block q. The ATT estimator is then computed as in equation 2.24.

$$\tau^S = \sum_{q=1}^{Q} \tau_q^S \frac{\sum_{i\in I(q)} D_i}{\sum_{\forall i} D_i} \tag{2.24}$$

where the weight for each block is given by the corresponding fraction of treated units and Q is the number of blocks. An important disadvantage of this procedure is that it discards observations in blocks where either treated or control units are absent.

2.6 Which of the existent matching algorithms work best?

Given the existence of many matching algorithms, in literature is open the debate about which method to select in practice. Clearly, the method to select depends on many aspects.

First of all, the kind of data to analyze, the degree of overlap between the comparison and

[13] for the variance formula of the considered estimators see Becker and Ichino, 2002

treatment groups with respect to estimated propensity score. When there is enough overlap in the distribution of propensity score between treated and controls, the considered matching algorithm will yield similar results. When there is not overlap the treatment effect could not be estimated. When the overlap is poor the choice of which methods to adopt depends on subjective choice of researchers for what concerns matching algorithms, use of replacement or not, range of ps. For good matching researchers have to solve the trade-off between finding matched for all treated units and to obtain matching pairs that are very similar to each other. Another important aspect to consider is that propensity score matching algorithms work only if the assumption of selection on observable covariates is valid. A key problem of the existing approximate matching methods (Iacus et al. 2008) is that, for example, the propensity score can be used to find the area of extrapolation only after we know that the correct propensity score model has been used. However, the only way to verify that the correct propensity score model has been specified is to check whether matching on it produces balance between the treated and control groups on the relevant covariates. But balance cannot be reliably checked until the region of extrapolation has been removed.

It clearly emerges that there are poor specific guidelines in the literature on which of these matching algorithms works best, and the answer depends especially on data. But, generally, if the point of matching estimator is to minimize bias by comparing target units to similar matched units, then methods that make it possible should be preferred. Matching is generally successful if, for both the treatment and matched control groups, the distribution of the matching variables is the same. When this result is achieved, the data are said to be balanced. Balance usually is assessed using pairing t-tests for differences in means of the matching variables across matched treatment and control cases. But to achieve full balance, the entire joint distribution of the matching variables must be the same, with all observed differences small enough to be attributable to random variation. To meet this standard, one must evaluate the equivalence of the full joint distributions, and more complicated tests are required. These complicated tests are often not implemented in practice.To facilitate the way of testing balance, we propose a method that automatically tests balance across a multivariate X-space. Then, if the covariates are not balanced, one can change the estimation model for the propensity score.

In literature, there are some contributes that favor the multivariate balance test. Rosenbaum (2002) reports on recent results for full optimal matching algorithms. His algorithm seeks to optimize balance and efficiency of estimation by searching through all possible matches that could be made, after stipulating the minimum and maximum number of matches for matched sets of treatment and control cases.

Diamond and Sekhon (2005) propose a general multivariate matching method that uses a genetic algorithm to search for the match that achieves the best possible balance. The quality of balance is specified as a standard set of t-tests of differences of means. Their technique is general and can remove the researchers from having to make any specification

choices other than designating the matching variable that one wishes to balance.
After using one of the matching estimators one should use thereafter some adjustment procedure. One is, for example, covariance adjustment (Rubin and Thomas, 2000). Other procedure are proposed in Heckman, Ichimura, and Todd (1997,1998); Heckman, Ichimura, Smith and Todd (1998); Abadie and Imbens (2004). Although these adjustment procedures may help to refine the balance of $\underline{X}$ across treatment and control cases, they do not help to address the problem of unobservable variables. In the presence of unobservable and when treatment assignment is not ignorable literature suggests to perform a sensitivity analysis.(Rosenbaum 1991,1992; Rosenbaum and Rubin 1983b; Ichino, Mealli, Nannicini, 2004).

2.7 The Economic Approach

This paragraph introduces the economic approach to causal inference by following the papers of Heckman, LaLonde and Smith (1999), Heckman (1979;1989).
The economic approach to program evaluation is based on estimating behavioral relationships that can be applied to evaluate policies yet implemented. The economic approach guided by economic model is, for some aspects, in contrast with statistics. Statisticians are interested in estimators that must be correct and efficient; whereas, economists are usually interested on the framework that motivates estimators.
In particular, they are interested on covariates involved in both outcome and participation equations. More precisely ,they suggest specific functional forms of estimating equations motivated by *a priori* theory.
In particular, outcomes under conditions $D = 1$ (eq. 2.25) and $D = 0$ (eq. 2.26) are defined as functions of observable ($\underline{X}$) and unobservable (U_1, U_0).

$$Y_1 = g_1(X) + U_1 \tag{2.25}$$

$$Y_0 = g_0(X) + U_0 \tag{2.26}$$

with the assumption that $E(U_1 \mid X) = 0$,$E(U_0 \mid X) = 0$ and that both g_1 and g_0 are non stochastic functions.
Then, the main parameter of interest is defined. It could be the average treatment effect on the population (ATE) or the average treatment effect on the treated units (ATTE) or the average treatment effect on the untreated (ATUE). The most commonly used evaluation parameters is *the average effect of treatment on the treated* (ATTE)(2.27)(Heckman and Robb, 1985; Heckman et al. 1997)

$$\begin{aligned} E(Y_1 - Y_0 \mid X, D = 1) &= E(\triangle \mid X, D = 1) \\ &= g_1(X) - g_0(X) + E(U_1 - U_0 \mid X, D = 1) \end{aligned} \tag{2.27}$$

In estimating the ATTE researchers have to deal with the presence of selection bias due to lack of random assignment. In Heckman's perspective selection bias is defined as a form of omitted variable bias(Heckman, 1979). In particular, he asserts (Heckman,1989) that selection bias exist if:

$$E(Y_{it}^* \mid D_i = 1) \neq E(Y_{it}^* \mid D_i = 0) \tag{2.28}$$

with Y_{it}^* as the outcome for those units in period t who do not receive the treatment.
Then:

$$Y_{it} = Y_{it}^* + D_i \alpha_{it} \text{ if } t > k \tag{2.29}$$

$$Y_{it} = Y_{it'}^* \text{ if } t < k \tag{2.30}$$

where the convention is that treatment occurs in period k.
The mean post-program of the outcome for treated is defined as:

$$E(Y_{it} \mid D_i = 1) = E(\alpha_{it} \mid D_i = 1) + E(Y_{it}^* \mid D_i = 1) \tag{2.31}$$

and the outcome for untreated as:

$$E(Y_{it} \mid D_i = 0) = E(Y_{it}^* \mid D_i = 0) \tag{2.32}$$

The difference in mean of outcomes between treated and not treated is:

$$E(Y_{it} \mid D_i = 1) - E(Y_{it} \mid D_i = 0) = E(\alpha_{it} \mid D_i = 1) + \{E(Y_{it}^* \mid D_i = 1) - E(Y_{it}^* \mid D_i = 0)\} \tag{2.33}$$

The last term in the equation 2.33 represents the selection bias term.
The selection bias problem does not exist in the presence of random assignment where the 2.34 holds.

$$E(Y_{it}^* \mid D_i = 1) = E(Y_{it}^* \mid D_i = 0) = E(Y_{it}^*) \tag{2.34}$$

In the absence of a random assignment, in order to draw causal inferential conclusions, one of the most used model in the economic setting is the Heckman's selection model (Heckman, 1979). The Heckman selection model deals with sample selection, but the same approach can be used in dealing with non-random assignment to treatment as well. The selection model takes into account two equations: a selection equation (a model of program participation) and an outcome equation.
For what concerns the outcome equation, here we consider the more general case in which:

- the outcome equation is a simple linear model

$$Y_{it} = X_{it}\beta + D_i\alpha_t + U_{it} \text{ with } t > k \tag{2.35}$$

$$Y_{it} = X_{it}\beta + U_{it} \text{ with } t \geq k \tag{2.36}$$

  with U as a random disturbance term, with $E(U_{it} \mid X_i) = 0$, with t as the number of periods of data on X available for each observation, α_t as the impact of the program under evaluation, and Y_{it} as the observed outcome.

- the treatment effect is invariant across individuals, such that

$$\alpha_{it} = \alpha_t \tag{2.37}$$

When assignment to treatment is non-random, selection bias in the estimation of α_t can arise because of dependence between d_i and U_{it}. That is, in a model without regressors, $E(U_{it} \mid d_i) \neq 0$, and with regressors $E(U_{it} \mid d_i, X_i) \neq 0$. So, $E(Y_{it} \mid d_i, X_i) \neq X_{it}\beta + d_i\alpha_t$. In this case, an ordinary least square regression of Y_{it} on X_{it} and d_i will not yield consistent estimates of α_t (or β). This is due to the fact that unit's participation decision is probably based on personal unobservable characteristics that may affect the outcome.
The baseline idea of the Heckman selection model, is to directly control for that part of the

error term (U) in the outcome equation that is correlated with the participation dummy variable (D).
The Heckman's selection model assumes that the participation decision can be parameterised in terms of an index function well known as the selection equation:

$$I_i = Z_i\gamma + V_i \tag{2.38}$$

where V_i represents a random disturbance for unit i for selection equation. Then the outcome Y_i is observed if I_i exceeds a particular threshold. In fact:

$$D_{it} = \begin{cases} 1 & \text{if } I_i > 0 \text{ and } t > k \\ 0 & \text{otherwise} \end{cases}$$

Z affects the outcome only through the participation status D. Then by imposing additional structures on the model it is possible to estimate the treatment effect of interest. [14] V_i is assumed to be independently and identically distributed across units. Assuming that V_i is independent of Z_i then $Pr(d_i = 1 \mid Z_i) = E(d_i \mid Z_i) = 1 - F(-Z_i\gamma)$ which Rosenbaum and Rubin call the propensity score (Heckman, 1989). Dependence between U_{it} and D_i can arise for one of two not necessarily mutually exclusive reasons: dependence between Z_i and U_{it} or dependence between V_i and U_{it}. Heckman (1989) refers to the first case as selection on observable and the second case as selection on unobservable. The source of selection bias for any particular problem depends on the actual process used to select units.
Selection on observable occurs when the dependence between U_{it} and D_i is due to a set of observed variables, Z_i, which influence selection into program. More formally, following the Heckman's notation: $E(U_{it} \mid D_i, X_i) \neq 0$ and $E(U_{it} \mid D_i, X_i, Z_i) \neq 0$; but $E(U_{it} \mid D_i, X_i, Z_i) = E(U_{it} \mid X_i, Z_i)$.
Controlling for the observed selection variables -Z_i- solves the selection bias problem.
Selection on unobservable may occur when the dependence between the treatment indicator variable, D_i, and U_{it} is not eliminated even after controlling for Z_i. That is: $E(U_{it} \mid D_i, X_i) \neq 0$ and $E(U_{it} \mid D_i, X_i, Z_i) \neq E(U_{it} \mid X_i, Z_i)$.
Selection is then said to depend on unobservable. Such selection bias estimators, when selection is on unobservable, are formed by invoking assumptions about the distribution of V_i, Z_i and U_{it}.

Shadish et al.(2002) asserted that the Heckman selection model has some analogies with the PS. As with PS models, the selection equation predicts actual group membership from a set of presumed determinants of selection into conditions, yielding a predicted

[14] For a comprehensive review of the estimation procedure for both homogeneous and heterogeneous treatment regimes see Blundell and Costa Dias(2002).

group membership score. This prediction is then included in the outcome equation.
In selection bias models, if the residual of the selection equation departs much from zero, then the selection bias may fail to yield unbiased estimates of treatment effects. The functional form of the selection equation must be correctly specified.
An advantage of the selection bias models is that they address the question of taking hidden bias into account rather than just adjusting for observed covariates. These models would probably work better if they used predictors that were selected to reflect theory and research about variables that affect selection into treatment, which requires studying the nature of selection bias as a phenomenon in its own right (e.g. Andrman, Cheadle, Curry, Diehr, Shultz, and Wagner,1995).
The selection estimator of α_t is only one of the existing set of estimators. Heckman and Robb (1985,1986) present a comprehensive summary of selection bias estimators which can be implemented in alternative types of data. All non-experimental estimators differ in the assumptions imposed, the data required to implement such estimators, and their robustness to alternative sampling plans and measurement errors.

2.8 Regression Discontinuity Design (RDD)

Another way to deal with selection bias is represented by the Regression Discontinuity Design (RDD). Campbell and Stanley,1966; Cook and Campbell, 1979; Cook and Shadish, 1994 categorize the regression discontinuity design as a quasi-experiment. This is partly due to their understanding of a quasi-experimental design that has structural features of an experiment but that lacks random assignment. Works on the RDD began in 1958 (Campbell, 1984) with the first published example being Thistlewaite and Campbell (1960).RDD could be defined as a special case of *Selection on observable*. It represents a design in which assignment is based on a cutoff score: the experimenter assigns units to conditions on the basis of a cutoff score on an assignment variable, not by coin toss or lottery as in a randomized experiment. That means the probability of assignment to treatment depends in a discontinuous way on some observable variable S.
The assignment variable can be any measure taken prior to treatment, where the units scoring on one side of the cutoff are assigned to one condition and those on the other side to another.
Examples of allocation variable can be , for example, in an education setting, merit score, need (or risk) score, first come, date of birth, and so on. RDD can be viewed as a randomized experiment at cutoff or as a completely known assignment process.
In most other quasi-experiments where assignment to treatment is uncontrolled, the selection process is sometimes totally unknown, often partially known, but almost never fully known. If the selection process could be completely known and perfectly measured, then one could adjust for differences in selection to obtain an unbiased estimate of treatment effect. In theory, these conditions are met in both RD and the randomized experiments,

and so both designs can be viewed as special cases of selection bias modeling. In a randomized experiment, the assignment mechanism is completely known and is equivalent to a coin toss. It is also fully known for RD, being whether the score on the assignment variable is above or below the cutoff. In neither case exists the problem of unobservable, that is the presence of unknown variables that influence the assignment mechanism. In both cases, the assignment mechanism can be perfectly measured and implemented, that is, the researcher records correctly whether the coin came up heads or tails, or whether a person's score is above or below the cutoff. When units get assigned to the treatment on the basis of a known and pre-established cutoff score on a pre-intervention covariate, the assignment variable cannot be caused by treatment. This requirement is met by an assignment variable that never changed, like the year of one's birth (Judd and Kenny, 1981). The assignment variable can even be totally unrelated to outcome and have no particular substantive meaning. The best assignment variable is a continuous variable that maximize the chance of correctly modeling the regression line for each group. It is possible to use many assignment variables simultaneously, and not just one. If several assignment variable are in different metrics, one could form a total score from them after first standardizing them and possibly weighting them differentially (Judd and Kenny, 1981; Trochim, 1984, 1990). Assignment to treatment must be controlled, which rules out most retrospective uses of the design. It is especially appropriate when decision makers wish to target an action or a program to those who most need or deserve it.
The basic analysis involves an Analysis of Covariance (ANCOVA) with the assignment variable as the covariate:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 Z_i + \hat{\beta}_2 (X_i - X_c) + e_i \tag{2.39}$$

where Y is the outcome, $\hat{\beta}_0$ is the intercept, Z is the treatment dummy variable $(1, 0)$, X is the assignment variable, X_c is the cutoff (to estimate the effects of treatment at the cutoff), $\hat{\beta}_2$ predicts outcome from assignment, $\hat{\beta}_1$ is the estimate of treatment effect, e is a random error term.If the outcome variable is continuous, then an ordinary regression equation can be used; whereas if the outcome is dichotomous, then should be used a logistic regression.[15] A big problem of RDD is represented by potentially misspecified functional form of assignment on outcome. In fact, with RDD we measure the size of the effect as the size of the discontinuity in regression lines at the cutoff. In doing this, we assume that relationship between assignment and outcome is linear. But functional forms can be non linear due to: nonlinear relationship between the assignment variable and the outcome; interactions between the assignment variable and treatment. Functional

[15]Subtracting the cutoff value from the assignment variable $(X_i - X_c)$, which is the same as centering the assignment variable if the cutoff is the mean, causes the equation to estimate the effect of treatment at the cutoff score, the point at which groups are most similar. One could estimate the effect anywhere on the range of the assignment variable by varying which value is subtracted, or estimate it at the intercept by subtracting zero.

form is an important aspect because effects are unbiased only if the functional form of the relationship between the assignment variable and the outcome variable is correctly modeled. A solution could be to include nonlinear functions of the assignment variable in the equation, as follows:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 Z_i + \hat{\beta}_2 (X_i - X_c) + \hat{\beta}_3 (X_i - X_c)^2 + e_i \tag{2.40}$$

One can also add interactions between treatment assignment (Z) and the assignment variable (X) as follows:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 Z_i + \hat{\beta}_2 (X_i - X_c) + \hat{\beta}_3 Z_i (X_i - X_c)^2 + e_i \tag{2.41}$$

or finally, one can add both nonlinear and interaction terms to the model.
Another shortcoming is that all participants must belong to one population prior to being assigned to conditions, though the RDD literature is unclear about how to define a population. A definition could be that, in RDD, it must have been possible for all units in the study to receive treatment had the cutoff been set differently. Ideally, as in a randomized experiment, those in the treatment group, all should receive the same amount of treatment, and those in control no treatment at all.
In addition, a disadvantage of RDD is that it only identifies the mean effect at the discontinuity point for selection. If the treatment effect is heterogeneous across units, RDD tells us nothing about units away from the threshold: it is able to identify only a local mean impact. Heterogeneity in the effects represents a vexing problem for many researchers(see for example Peck, 2003;2005). Heterogeneity may occur when treatment works better for some people than for others: for example, in the education setting, it is common to find that more advantaged children benefit more from treatment than do less advantaged children. In this circumstance, if the interaction between the assignment variable and treatment is not modeled correctly, a false discontinuity will appear. In RDD, an effective treatment will alter the slope or intercept of the regression line at the known cutoff point.

Chapter 3

Some drawbacks of conventional methods

3.1 Introduction

In the previous chapter we have introduced various methods that aim to construct the counterfactual dealing especially with the selection bias problem.
In this chapter we will discuss when and how the assumptions behind conventional methods break down in practice.
Our concern here is that different estimation methods, and different model specifications, potentially, and often in the real applications, led to different results.
This could especially happen when conditions on which estimation methods are based are not always really checked in the correct way.
This chapter represents a review of contributes that in literature have increased the debate about PS methods problems, but also about problems concerning the use of economic models for assert causality.
Here we will also discuss some remedies that have been proposed to address the resulting problems. We think that drawbacks of PS principally derive from the wrong practice and not from theory at all. Rubin(2007), for example, has emphasized some conflicts between the prescription of the potential outcome approach and the practice of observational studies in epidemiology and social science, where outcome data, Y_{obs} are used to fit various models, try transformations, improve p-values, in order to achieve publishable results.
We will take into account the three major open debates of conventional propensity score methods: the problem of variable selection in controlling for selection bias, the problem of how test balance and how to minimize the model dependence of ps methods. Finally, in explaining these problems, which are related each other,we will consider new research lines and contributes that overcome some limits and that represent a key bridge between conventional methods and our proposed method.

3.2 The hidden bias problem

A crucial step in modeling the unknown selection mechanism in observational studies, is to identify potentially relevant covariates to measure. Potentially relevant covariates are those expected to affect treatment selection and outcomes. Researchers agree in considering that the omission of such relevant covariates results in hidden bias that propensity score cannot adjust. In fact, PS analysis assumes that all variables related to both outcomes and treatment assignment are included in the vector of observed covariates(Rosenbaum and Rubin, 1983) that is the researcher knows and measures the selection model perfectly, as with the perfectly implemented regression discontinuity design (Cook, 2008; Shadish. Cook and Campbell,2002). Unfortunately, this assumption is not always realistic, and researchers have to consider how much the results are robust to departures from it. In fact, in an observational setting, selection process is complex and not perfectly known, usually involving some combination of self-,administrator-, or other third-person-selection. (Steiner et al, 2008). If the PS model is incorrect or the covariates are measured imperfectly, then hidden bias may exist that affects estimates. Hidden bias results when a covariate is signficantly related to treatment assignment and outcome, but has not been measured and included in the propensity score model. The selection of rights covariates affect also the plausibility of the strong ignorability assumption [1]. In fact, when selection process is not perfectly known, the strongly ignorability assumption may not hold, and often cannot be tested.

The hidden bias is strictly related to strongly ignorable treatment assignment. But, in practice, this assumption is not careful checked.

Many authors have taken into account the *hidden bias problem* (Rosenbaum,2002;Imbens,2002; Smith and Todd,2001; Peck, 2007; Steiner et al., 2008; Rosenbaum and Rubin, 1983b), and they have proposed different solutions to the problem. Rosenbaum and Rubin (1983), for example, consider that differences due to unobserved covariates should be addressed after the balancing of observed covariates in the initial design stage, using models for sensitivity analysis or models based on specific structural assumptions. Rosenbaum(2002) presented a detailed discussion of sensitivity analysis, that examines whether the qualitative conclusions of a study would change in response to hypothetical hidden bias of varying magnitudes. Imbens(2002), for example, has considered an alternative approach, where the unconfoundedness assumption is relaxed by allowing for a limited amount of correlation between treatment and unobserved components of the outcomes.In the Imbens's perspective, the starting point of sensitivity analysis is the assumption that the unconfoundedness is satisfied only conditional on an additional, unobserved covariate. The analysis is close to the practice of assessing sensitivity of estimates by comparisons with results obtained by discarding one more observed covariates (Heckman and Hotz,

[1] If all covariates $\underline{X}$ related to both treatment T and potential outcome are observed, then treatment assignment is said to be *strongly ignorable given* $\underline{X}$(Rosenbaum and Rubin, 1983). Then, potential outcomes are independent of treatment assignment conditional on $\underline{X}$,and the average treatment effect can be estimated without bias.

1989; Dehejia and Wahba,1999; Smith and Todd,2001). To simplify, sensitivity analysis concerns the analysis of bias that can occur when not all relevant covariates were observed. The hidden bias problems affects the assumptions behind propensity score but also behind matching estimators. The existence of no hidden bias implies that the strong ignorability property holds. In particular, it implies that:

$$\begin{aligned} E[Y_i(1) \mid T_i = 1] &= E[Y_i(1) \mid T_i = 0] \\ E[Y_i(0) \mid T_i = 1] &= E[Y_i(0) \mid T_i = 0] \end{aligned} \tag{3.1}$$

But, not in all situations the assumption 3.1 is satisfied. When it is not satisfied under such a perfect stratification of data,it is possible to assert a conditional variant of assumption 3.1.

$$\begin{aligned} E[Y_i(1) \mid T_i = 1, S_i] &= E[Y_i(1) \mid T_i = 0, S_i] \\ E[Y_i(0) \mid T_i = 1, S_i] &= E[Y_i(0) \mid T_i = 0, S_i] \end{aligned} \tag{3.2}$$

whit S as a perfect stratification variable, such that units within strata defined by values of S are similar each other in all aspects except for observed value of the treatment indicator variable.

The condition 3.3 could be satisfied only if there are not unobservable variables in S.

In the presence of an unobserved variable, it could be a differential growth rate for the outcome that is correlated with treatment assignment/selection (S.L. Morgan and D.J.Harding, 2006).

Sensitivity analysis could be also helpful in this task. Ichino et al. (2005) have proposed a sensitivity analysis for matching estimators aimed at assessing if estimates derived under the strong ignorability assumption are robust with respect to specific failures of this assumption.

They suppose that *strong ignorability* is not satisfied when another additional binary variable could be observed. They simulate this additional variable and used it as an additional matching variable. Then a comparison of the estimates obtained with and without matching on this simulated binary variable makes clear if the estimator is robust to the specific source of failure of the unconfoundedness assumption. More precisely, in the sensitivity analysis the unconfoundedness assumption requires independence of the potential outcomes and the treatment indicator only after conditioning on one additional, unobserved, covariate U_i [2]. In doing this, a parametric model is postulated and estimated. Then, the

[2] $Y_i(0), Y_i(1) \perp T_i \mid X_i, U_i$

focus of the sensitivity analysis is the representation of the estimated average treatment effect in terms of the sensitivity parameters.
Rosenbaum and Rubin (1983b) have proposed a method to assess the sensitivity of average treatment effect (ATE) estimates in parametric regression models (ATE). In particular, their sensitivity analysis consists of the estimation of the average effect of a treatment on a binary outcome variable after adjustment for observed categorical covariates and an unobserved binary covariate U, under several sets of assumptions about U. (Rosenbaum and Rubin, 1983b). They assume that treatment assignment is not strongly ignorable given X, but is strongly ignorable given X and U, such that:

$$\begin{aligned} Pr(T = 1 \mid Y(0), Y(1), X) &\neq Pr(T = 1 \mid X) \\ Pr(T = 1 \mid Y(0), Y(1), X, U) &= Pr(T = 1 \mid X, U) \end{aligned} \tag{3.3}$$

with X as observed covariates, and U as the unobserved covariate and where the unobservable U is usually assumed to be independent of the observed covariates.

$$Pr(U = 1 \mid X) = Pr(U = 1) \tag{3.4}$$

If conclusions are insensitive over a range of plausible assumptions about U, the number of interpretations of the data is reduced, and causal conclusions are more defensible (Rosenbaum and Rubin,1983b).
Sometimes, the wrong practice could produce biased result if sensitivity analysis is not carefully checked. Usually, researchers use only those covariates for which statistically significant differences between treatment and comparison groups are found. (Rosenbaum, 2002c) offered three cautions against doing so:

- the relationship between the covariates and outcome is not considered and is just as important in many respects
- statistical significance is not a prerequisite for practical relevance, especially because the former depends so heavily on sample size
- the covariates are considered in isolation, whereas adjustment consider them collectively.

It is not clear which rationale researchers should adopt in selecting the right covariates. One should include all variables that affect both treatment assignment and dependent variable in order to reduce bias and avoid omitted variable bias. The theoretical literature emphasizes that including variables only weakly related to treatment assignment usually

reduces bias more than it will increase variance (Rubin and Thomas 1966; Heckman et al. 1998). One should include all variables that play a role in selection process (including interactions and other nonlinear terms;(Rosenbaum and Rubin, 1984; Rubin and Thomas, 1996) and that are presumptively related to outcome, even if only weakly so (Rubin, 1997) unless a variable can be excluded because there is a consensus that it is unrelated to outcome or is not a proper covariate, it is advisable to include it in the PS model even if it is not statistically significant (Rubin and Thomas,1996,p.253). The idea is that to reduce hidden bias propensity scores should be constructed using as many predictors of group membership as possible.

In contrast, in economic literature is emphasized the importance of the trade-off between the bias of excluding relevant variables and the inefficiency of including irrelevant ones. Steiner et al. (2008) implemented a within-study comparison to check if selection bias can be reduced with a particular set of covariates or with some particular analytic model. They decompose the complete set of covariates $\underline{X}$ into smaller, more homogeneous sets in order to investigate how well they establish strong ignorability and reduce bias. The within-study design they have proposed permits comparing the adjusted results of the quasi-experiment to the results from the randomized experiment and, to directly assess how well the covariates succeed in reducing selection bias, and to test whether the strong ignorability assumption is met. They use a series of different selection models that systematically vary covariate sets. They assessed the percentage of bias reduction of each method and covariate set by the fraction of the initial selection bias remaining after adjustment:

$$b\% = (\tau_Q^a - \tau_E) \backslash (\tau_Q^u - \tau_E) * 100 \tag{3.5}$$

where τ_Q is the adjusted or unadjusted average treatment effect in the quasi-experiment and τ_E the estimated average treatment effect in the randomized experiment. A positive sign indicates an under-adjustment with respect to the experimental effect, and a negative sign over-adjustment.

In their study, Steiner et al. (2008) found that selection bias can be almost reduced when appropriate covariates are available; further, they found that the choice of covariates is more important than the choice of analytic method; third, adding different sets of covariates systematically improves bias reduction since they collectively increase the capacity to predict the assignment process and outcome. They found that some covariates, are more important than others, and without some *important variable* in the selection model, incorporating many variables into a seemingly *rich* covariate set is not sufficient to eliminate bias. Finally, they make strong ignorability assumption more transparent than usual and they show its crucial importance for causal inference.

3.3 Limitations of propensity score estimation

Propensity score methods differ from economic models in the sense that they do not require any model for outcome. But both PS method and economic selection models are model dependent: economists use a model for both the selection process and outcomes; whereas, PS methods use a model for the assignment mechanism.

King and Zang (2006) gave a formal definition of model dependence. They consider it at point X as the difference, or distance, between the predicted outcome values from any two plausible alternative models; where ,by plausible, they mean models that fit the data well. In practice, model dependence exist in all situations where a single correct model must be chosen between multiple candidates. As a consequence, a unique estimator is not even specified ex ante and thus not well defined.

Typically, the model dependence problem exists because in real applications, propensity scores are unknown. Many authors agree in considering that small variations in choice during the estimation stage could yield to different results for what concerns bias reduction and size of treatment effect estimation:these choices concern control variables, functional forms, model assumptions, (see for example, W.Shadish, M.H.Clark, P.M.Steiner, 2008). When researchers use parametric methods, they do not know the true parametric model, and many different specification could be plausible.

In this sense, Ho et al. (2007) consider the PS as a *tautology*. In fact, in order to use non parametric matching to avoid parametric modeling researchers must know the parametric functional form of the propensity score equation. PS is a tautology also in the sense that to be a balancing score, analysts must know a consistent estimate of the true PS; but researchers know to have a consistent estimate of the PS when matching on the PS balances the covariates. Obviously, a wrong or not unique PS estimate will affect all sub-sequent analysis based on the estimated PS.

Sekhon and Grieve (2008) noted that if the PS model is wrong then PS matching makes covariate balance worse, and as a consequence, increase the bias in the estimates even if the selection on observable assumption is satisfied. [3] Ho, Imai et al. (2007) considers that PS estimates depend on their underlying model assumptions and that different specifications can yield very different causal inference conclusions.

More generally, Ho et al. (2007) consider that it does not exist a right model specification if researchers cannot verify assumptions on which they are based. They offer a solution to the model dependence problem by introducing a preprocessing method that does not call for parametric assumptions. They propose to preprocess a data set with matching methods so that the treated group is similar as possible to the control group, and the treatment variable is closer to being independent of the background variables.In doing this,they ensure that any subsequent parametric adjustment will be irrelevant in the sense that with preprocessing estimates based on the subsequent parametric analysis will be less

[3] i.e. even if the conditional distribution of the outcomes given the observed covariate is indipendent of treatment assignment

dependent on modeling choices and specifications.
Further, Luellen (2007) has conducted a simulation study, which suggests that PS score adjustment may be also sensitive to which estimation method (logistic regression, classification trees, bosted regression, random forest and so on) is used. Heckman, Ichimura and Todd (1998) highlight that Rosenbaum and Rubin (1983) proved that matching on PS balances all covariates by assuming that PS is known exactly, but practically researchers have to estimate it. They argue that matching on all covariates rather than on the estimated PS, could be more efficient than matching on PS.

For what concerns PS estimation,in applied research, analysts and researchers use logistic regression as the best fixes method. An important drawback of logistic regression is, for example, that it can underestimate the probability of rare events (King et al., 2003). As Baser(2006) pointed out researchers used to keep the control data set as large as possible to increase the likelihood of finding better matches for the treatment group. The question behind the use of propensity scores estimation methods is about which criterion should be maximized in order to obtain the best model and avoid the model dependence. Shadish, Luellen and Clark(2006) gave an answer to this question by focusing their attention on the *rationale* the researchers should adopt in estimating PS. They refer to two possible rationales: the *balancing strata* rationale and the *maximum prediction* rationale. The *maximum prediction* rationale aims at obtaining the best possible prediction of group membership, that is predict as well as possible. In doing this, they adopt as maximization criterion the percentage of correct classification of participant into conditions. They think that applying the maximum prediction rationale PS analysis could yield to the creation of the equivalent of a perfect assignment variable in a regression discontinuity design, for example, by creating a set of PS in which all treated have a PS grater than .50 and all controls have a ps less than $< .50$. They argued that applying this logic, the better the prediction, the less overlap exists between ps of the two groups, but overlap is an essential condition that have to hold in using, for example, a subclassification on PS. Whereas, the *balancing strata* rationale, implies that any PS that balances predictors over groups will do. As a consequence, the criteria for a good set of PS should be the maximization of how well the propensity scores balance predictors over conditions. They conclude that the goal is not to get accurate prediction into groups, but is to create scores that, when used, create balance on predictors over groups within propensity score strata.
This is also the suggestion of Rosenbaum and Rubin. In fact, under Rosenbaum and Rubin (1984), any propensity score that balances predictors will do.
Against what Rosenbaum and Rubin (1984) suggest, in applied research, such tests aim at maximizing the prediction. The Hosmer-Lemeshow test, for example,is useful for detecting the classification power of the logistic regression. The test suggests regrouping the data according to predicted probabilities (PS) and then creating equal-size groups. The insignificant value of the test is needed for precise classification.

The area under the receive operator curve (ROC) value is another way to detect classification power. The ROC curve is a graph of sensitivity versus one minus specificity as the cutoff varies. The greater the predictive power, the more bowed the curve. Therefore, the area under the curve can be used to determine the predictive power of logistic regression. To classify group membership correctly, C-Statistics should be greater than 0.80.
From literature, it clearly emerges that some authors agree in considering that the rationale underlying the estimation of PS should be the maximization of prediction; on the other hand, other authors suggest the use of the maximum balancing rationale.
The debate remains open.

3.4 Testing the balance property

How to evaluate balance is at the center of a rich debate in literature. The balancing property is important for both PS and matching method. Shadish et al. (2008), for example, have shown that PS adjustment may be sensitive to which covariate balance criteria are used. The success of matching, for example, is based on reducing selection bias by generating as much balance as possible between the distribution of pre-treatment covariates in the treated and control groups. There is no consensus on how exactly matching ought to be done, how to measure the success of the matching procedure and whether or not matching estimators are sufficiently robust to mis-specifications. In recent years, researchers start to be interested in how to choose the best matching techniques for their data sets. (Baser,2006; King and Stuart, 2006; Iacus et al., 2000). Ho et al. (2007) consider balance as the main diagnostic of success, as well as the number of observations remaining after matching. They highlight the importance of balance by emphasizing it provides a straightforward objective function to maximize in order to choose matching solutions. In various academic fields researchers used to evaluate the degree of equivalence by conducting hypothesis tests, most commonly the t-test for the mean difference of each of the covariates in the two matched groups, but also, the chi-square test, the F and Kolmogorov-Smirnov tests. Imai, King and Stuart (2006) showed that the common approach used in evaluating the success of this method is invalid. Imai, King and Stuart (2006) suggest that balancing holds when the 3.6 is satisfied:

$$\hat{p}(X \mid T = 1) = \hat{p}(X \mid T = 0) \tag{3.6}$$

where $\hat{p}$ is the empirical density of the observed data, rather than the population density. If the above assumption holds then the average treatment effect can be estimated by a simple difference in means of Y between treated and controls. It is clear that the immediate goal of matching is to choose an algorithm that satisfies the equation 3.6 as best as possible. Ideally that would involve comparing the joint distribution of all covariates X between

the matched treated and controls. However, when X is high dimensional, this is generally infeasible and thus lower-dimensional measures of balance are used instead. The standard practice involves the evaluation of 3.6 for the chosen matching algorithm by conducting t-test for the difference in means for each variable in X between the matched treated and control groups. For what concerns t-test, tables of t and/or they p-values are used as a justification for the adequacy of the chosen matching method and statistically insignificant t-test are used as a stopping rule for maximizing balance in the search for the appropriate matched sample from which to draw inferences. Iacus et al.(2008), have considered a common mistake that researchers do in real applications: they used to ignore imbalance due to differences in variances, ranges, covariances, and higher order interactions. These contributes clarify that the goal of measuring imbalance is to summarize the difference between the multivariate empirical distribution of the pre-treatment covariates for the treated $\hat{p}(X \mid T = 1)$ and matched control $\hat{p}(X \mid T = 0)$ groups. Unfortunately, many matching applications do not check balance. Generally, as mentioned above, who checks balance used to compare only the univariate absolute difference in means in the treated and control groups as in equation 3.7.

$$I_1^{(j)} = |\overline{X}^{(j)}_{m,(T,W)} - \overline{X}^{(j)}_{m,(C,W)}|, \forall j = 1, ..., k \tag{3.7}$$

where $\overline{X}^{(j)}_{m,(T,W)}$ and $\overline{X}^{(j)}_{m,(T,W)}$ denote weighted means, with weights appropriate to each matching method; or measure the imbalance in univariate moments, univariate density plots, propensity score summary statistics, or the average of the univariate differences between the empirical quantile distributions (Austin and Mamdani,2006; Imai, King and Stuart, 2008; Rubin, 2001).

Iacus, King and Porro (2008) gave an innovative measure of imbalance. They measure the multivariate differences between $\hat{P}(X \mid T = 1)$ and $\hat{P}(X \mid T = 0)$ via an L_1-type distance. Their measure works for both categorical and continuous covariates.

In particular, they first choose the number of bins for each continuous covariate to be discretized. Then, they cross-tabulate the discretized covariates as $X_1 \times ... \times X_k$ for the treated and control groups separately, and record in each cell the k-dimensional relative frequency for the treated $f_{l_1,...,l_k}$ and control $g_{l_1,...,l_k}$. Then their measure of imbalance is represented by the absolute difference over all the cell values as in 3.8.

$$\ell_1(f, g) = \sum_{l_1,...,l_k} |f_{l_1,...,l_k} - g_{l_1,...,l_k}| \tag{3.8}$$

An important property of their approach is that the empty cells do not affect the measure of imbalance. Furthermore, the use of relative frequencies controls for potentially different sample sizes between the treated and the control groups.

King and Stuart (2006) have argued that the common practice of conducting matching is problematic for many reasons.

First, they have shown that randomly dropping observations can influence not only balance but also statistical power, and unfortunately the t-test, like most statistical tests, is a function of both. The more observations dropped, the less power the tests have to detect imbalance in observed covariates. The difference in sample means as a measure of balance is distorted in the t-test by the total number of remaining observations, the ratio of remaining treated units to the total number of remaining observations and the sample variance of $\underline{X}$ for the remaining treated and control units. Then they have argued that a difference in means is a fine way to start. Other options include higher order moments than the mean, nonparametric density plots, and propensity score summary statistics.

Sometimes software do not incorporate a correct balance test. In the Becker and Ichino procedure, for example, the pscore.ado program does not test the balancing property in the strict sense, but only one of its implications; i.e. the mean. Softwares should add tests for higher moments of the distribution of characteristics.

A more general approach to alleviate errors in balance testing is represented by the use of quantile-quantile plots that compare the empirical distribution of two variables, although statistics based on QQ plots can have higher variance (Ho, Imai et al., 2007).

In addition, Imai et al. (2006) suggest that the statistics chosen to assess balance should be characteristics of the sample and not some hypothetical population.

Steiner et al. (2008) assess balance in observables using Cohen's

$$d = \frac{(\overline{X}_t - \overline{X}_c)}{\sqrt{\frac{(s_t^2 + s_c^2)}{2}}} \tag{3.9}$$

and variance ratio $\nu = \frac{s_t^2}{s_c^2}$ between treatment and comparison group. After propensity score adjustment, standardized mean differences d should be close to zero, variance ratios ν close to one (Rubin, 2001). If it is not possible to obtain balance in the covariates, then perhaps the groups are so nonequivalent that they should not be compared.

A graphical analysis of the overlap in estimated PS could be also useful to examine whether groups overlap enough to be worth comparing.

To summarize, we think that to correctly implement propensity score and matching algorithm, instead of using hypothesis tests for assessing balance, we need to assess the difference in the multivariate empirical densities of X for the treatment and control groups. In the next chapter we will show how the use of a partial dependence analysis could be useful in testing in a multivariate way the balancing property.

3.4.1 Genetic matching algorithm (GM)

Here, following Sekhon and Grieve (2008);Diamond and Sekhon (2006); Sekhon and Mebane (2000) we consider an alternative matching method for causal inference, that automatically checks balance.

GM is a data mining method that searches the best solution within all possible solutions. The GM, in fact, uses an evolutionary algorithm which consists of a set of heuristic rules to modify a population of trial solutions in such a way that each generation of trial values tends to be, on average, better than its predecessor. The Genetic Matching (GM) is a new non parametric multivariate matching method for addressing covariate imbalance in observational studies.

We introduce GM here rather than in the section concerning matching methods, because we think that GM represent a first step in avoiding the balance test problem and that, for some aspects, aims at ensuring objectivity in results. It uses an evolutionary search algorithm to automatically determine the weight each covariate is given, that maximizes the balance of observed potential confounders across matched treated and control units. This method does not depend on whether PS is known or not, but it is improved when a propensity score is incorporated. The basic idea of Geneting Matching is that if matching using Mahalanobis distance is not optimal for achieving balance between treatment and controls, then it should be possible to search over the space of distance metrics and find something better by directly minimizing measures of covariate imbalance. One way of generalizing the Mahalanobis metric is to include an additional weight matrix as follows:

$$d(X_i, X_j) = \{(X_i - X_j)'(S^{-\frac{1}{2}})'WS^{-\frac{1}{2}}(X_i - X_j)\}^{\frac{1}{2}} \tag{3.10}$$

where W is a square weight matrix with rows and columns equal to the number of covariates in X, and $S^{\frac{1}{2}}$ is the Cholesky decomposition of S, the variance covariance matrix of X. GM is an invariant matching algorithm that uses distance measure d() in which all elements of W are zero except down the main diagonal. The main diagonal consists of k parameters that must be chosen. If each of these parameters are set equal to 1, $d()$ is the same as the Mahalanobis distance. An important issue is how to choose the free elements of W [4], due to the fact that the optimization problem grows exponentially with the number of free parameters. By default, geneting matching uses cumulative probability distribution functions of standardized statistics. The default standardized statistics are paired t-tests and non-parametric bootstrap Kolmogorov-Smirnov tests that compare the distribution of covariates across treatment and control groups. By default GM attempts to minimize a measure of the maximum observed discrepancy between the matched treated and control covariates, at each iteration of the optimization.

[4] W has an infinity of equivalent solutions because the matches produced are invariant to a constant scale change to the distance measure. The matched produced are the same for ever $W = cW$, with c¿0

For a given set of matches resulting from a given W, the loss is defined as the minimum p-value observed across a series of balance tests performed on distributions of matched baseline covariates. Usually the tests conducted are t-tests for the difference of means and non parametric (bootstrap) Kolmogorov-Smirnov distributional test. The algorithm attempts to maximize this loss function by minimizing the largest discrepancy at every step. As shown by Diamond and Sekhon (2006), the main advantage of GM is that covariate balance was much improved compared to using propensity score or Mahalanobis distance matching. Another advantage of GM is that it is an algorithm that by searching and finding relationships in the data, achieves excellent levels of balance that does not depend on PS estimation.

3.5 Some Problem of Heckman's selection model

The main problem of Heckman's selection model concerns how to choose among competing estimators(Heckman, 1989). When not all characteristics related to selection mechanism are controlled, then bias in the estimates of program impacts may occur. If selection bias exists, then different non-experimental estimators could lead to different results because of differences existing in the assumptions underlying each estimators. Many authors have verified that different estimators produce different estimates of the same program. Lalonde(1986) and Faker and Maynard (1984,1987), for example, using experimental data from the National Supported Work Demonstration, have found that non-experimental estimates vary widely and differ greatly from the experimental estimates. Other authors that have found such dependence on different estimators are, for example, Burtless and Orr (1986,p.613), Ashenfelter and Card (1985, p.648), Barnow (1987, p.190).
Two important features of economic models are the following: on one hand, alternative non-experimental estimation procedures should produce approximately the same program estimate,but this requirement is not always matched. On the other hand, there is no objective way to choose among alternative non-experimental estimates.
The first feature is not matched when there are systematic differences between treated and comparison group in observed and unobserved characteristics that affect outcome. This is due to the fact that different non-experimental estimators make different assumptions about the distribution of these differences.
In solving these problems Heckman(1989) has proposed some model specification tests. He has considered the problem of assessing the validity of alternative non-experimental evaluation models that do not produce estimated program impacts close to the experimental results: the model not rejected produced impacts that are close to the experimental results.

Chapter 4

A multivariate data mining approach to deal with selection bias

4.1 Introduction

In the previous section we presented some limitations of conventional methods in estimating causal effects when random assignment is not feasible. In particular, we have highlighted some limitations of existing methods in testing the balance property.In order to maximize balance across treatment and control groups, it is necessary to be able to measure and test for balance. There are many issues involved with choosing appropriate tests, but we noted that most researchers especially ignore all aspects of multivariate balance not represented in the well known variable-by-variable summaries. The concern here is to theoretically introduce the new approach to measure selection bias and test balance by preserving the multivariate nature of data.

The main idea lies in the use of the more general framework of the partial dependence analysis (Daudin,1981) as a tool for investigating the dependence relationship between a set of observable covariates $\underline{X}$ and a treatment indicator variable T in order to obtain a measure of imbalance according to their dependence structure.

Further, we propose the use of a clustering procedure as a tool to find groups of comparable units on which estimate local causal effects and we propose the multivariate test of imbalance as a stopping rule in choosing the best partition of the X-space.

4.2 Objectives

Usually, there are several pre-treatment covariates X along which balance ought to be checked, and a method of combining differences it needed. The question is how many tests should be performed: one for each pretreatment covariate or a single omnibus test?(Hansen and Bowers,2008).

The innovative aspect of this thesis try to answer to these questions by performing a multivariate approach that involves measuring selection bias under non-experimental conditions and testing the imbalance in a multivariate way. It implies that balance is checked not only on $x_1, ..., x_k$, but simultaneously on all covariates involved in the selection process.
Here we use Rubin's framework as our springboard. The idea is to consider the available information as starting point, being strictly interested in the current available sample and not in inference about a population. As in the Rubin's approach, potential outcomes and covariates are defined as scientific entities, no matter which design - experimental, observational or something else - researcher use. What Rubin calls *The Science*, in our approach is represented by the information matrix $\underline{X}$, and by observed potential outcomes (Y_{obs})(tab 4.1). Obviously, the procedure we propose has no magic, in the sense that it

	1 ... j ... Q
1	
...	
i	x_{ij}
...	
n	

T
1
...
0
...
1

Y(0)	Y(1)
missing	Y_{obs}
...	...
Y_{obs}	missing
...	...
missing	Y_{obs}

Table 4.1: Left: Information matrix; Center: assignment vector; Right: observed potential outcome

does not help us control for covariates involved in the selection process that are not available.
We assume to have sufficient information in the measured pre-treatment control variables X_i. The information matrix $\underline{X}$, must include all variables that are causally prior to the treatment assignment T_i and that affects Y_i conditional on T_i.
The method we propose has two main objectives:

First, given the information matrix it aims at measuring the global selection bias by a two-stage procedure involves the following:

1. Original pre-treatment covariates, without introducing outcome in the analysis, are transformed using a specific eigenvalues and eigenvector de-composition to derive a factorial conditional space, in which the inertia associated with treatment assignment has been eliminated.
   The eliminated inertia represents the global measure of selection bias existent in the information matrix.In this way we derive a bias elimination coefficient (BEC) that represents the measure of selection bias relative to the total amount of the inertia of the specific information matrix considered.

2. Then we test if the detected bias is important with respect to the hypothetical case of random partition (BIAS=0)

In essence, given an information matrix $\underline{X}$ of pre-treatment covariates, our analysis detects

the variability associated with the selection mechanism by creating a new space that is void of any variability associated with that mechanism.
The goal of measuring imbalance is to summarize the difference between the multivariate space of the pre-treatment covariates for the treated and the multivariate space of the pre-treatment covariates for the controls.

Second, we propose the use of a clustering procedure in order to find subspaces on which measure local causal effects. Then, the multivariate test of imbalance is used as a stopping rule in finding the best partition of the X-space on which we wish to measure unbiased local average treatment effects.

4.3 General framework: the partial dependence analysis

Our underlying paradigm here is that of French School of Analyse des Données. According to Benzécri the data are king, not the model one might want to propose for them (Greenacre, 2006). The philosophy of that school is to place data at the center of the researcher.
When there is not an a priori knowledge about the relationship between variables, displaying the existing relationship between variables on a factorial space is one of the most powerful tools for detecting the hidden information.
If there is dependence between covariates and treatment assignment any descriptive factorial analysis may exhibit this link. The aim is to implement a conditional analysis in order to find a new X-space free of any dependence from the treatment assignment: the part of variability of the original $\underline{X}$ space that has been eliminated will represent the measure of selection bias.
Here we propose to study the conditioning applied to a problem with qualitative variables[1] where all or some of them may be linked to the treatment assignment variable.
The problem of dependence of a set of qualitative variables from the influence of an external qualitative variable T was studies by B. Escofier (Escofier, 1987), who aimed at obtaining a factorial space by taking into account only the variability not dependent from T (Inertia Within) with the resulting CORCO model (Escofier 1987;1988). Escofier refers to the more general framework of partial dependence analysis due to Daudin (Daudin, 1981). Daudin has extended the concept of partial dependence first defined by J.N. Darroch (Darroch, 1979). Darroch has distinguished between two sources of partial dependence between two variables: the dependence due to T, called the *dependance attachée*, and the dependence not due to T, called the *dependance détachée*. The key contribute of Daudin was the transition from the definition of partial dependence in the analysis of probability tables in a probabilistic framework (Darroch, 1979) to the the analysis of contingency tables in the correspondence factorial analysis framework (Daudin, 1981). Starting from

[1] continuous covariates could be also introduced in the analysis if discretized

the decomposition of the marginal dependence in dependence not due to T e dependence due to T (4.1)

$$\underbrace{P_{ij.} - P_{i..}P_{.j.}}_{\text{marginal dependance}} = \underbrace{(P_{ij.} - \Pi_{ij})}_{\text{dependence not due to T}} + \underbrace{(\Pi_{ij} - P_{i..}P_{.j.})}_{\text{dependence due to T}} \tag{4.1}$$

where $P_{ij.} = \sum_t P_{ijt}$, $P_{i..} = \sum_j \sum_t P_{ijt}$, $P_{.j.} = \sum_i \sum_t P_{ijt}$ and $\pi_{ij} = \sum_t \frac{P_{i.t}P_{.jt}}{P_{..t}}$, with π_{ij} interpreted as the conjoint probability of $(X = i)$ and $(Y = j)$ when the two events are independent given T. [2]

Daudin has proposed two separated correspondence factorial analysis: one that analyze the dependence between variables due to T and another that analyzes the dependence between variables not due to T. To analyze the dépendance détachée he proposed to perform a correspondence factorial analysis of the contingency table with generic term N^*_{ij}:

$$N^*_{ij} = \frac{N_{i..}N_{.j.}}{N} + (N_{ij.} - M_{ij}) \tag{4.2}$$

with N^*_{ij} that aims at studying the dependence not due to T; $\frac{N_{i..}N_{.j.}}{N}$ as the generic term of the table that aims at studying the marginal dependence (X,Y) of the variable X and Y; and with $M_{ij} = \sum_t \frac{N_{i.t}N_{.jt}}{N_{..t}}$ that aims at studying the relationship between variables due to T. N_{ijt} indicates the absolute frequencies of the tridimensional table of $X = i$, $Y = j$ and $T = t$, $N_{i.t} = \sum_j N_{ijt}$, $N_{.jt} = \sum_i N_{ijt}$, $N_{..t} = \sum_{ij} N_{ijt}$, $N_{i..} = \sum_{jt} N_{ijt}$ and $N_{.j.} = \sum_{it} N_{ijt}$.

The factorial analysis of the N^*_{ij} table is defined as a factorial analysis with reference to a model (Escofier, 1984), where the object of the analysis is a table derived as the difference between the raw data and a model table, with the model corresponding to the structure induced by T on the data.

4.4 Notation

We aim at decomposing the original $\underline{X}$ space of analysis in two complementary spaces: one whose variability is only that due to the relationships between the covariates entered in the analysis (**X**), but not due to the selection mechanism (**T**); and another whose variability is only that due to the selection mechanism.

A matrix's overall variability $[Inertia(X;T)]$ can be decomposed into elements that are

[2] see appendix B for more details

independent of the selection-to-treatment mechanism [$Inertia(X \perp T)$] and dependent on that mechanism [$Inertia(X|T)$].

$$\underbrace{Inertia(X;T)}_{total} = \underbrace{Inertia(X|T)}_{between} + \underbrace{Inertia(X \perp T)}_{within} \tag{4.3}$$

According to a conventional data matrix decomposition in eigenvalues and eigenvectors, our approach involves decomposing the portion of the matrix that does not depend on the selection mechanism (inertia within) for then use the part of inertia that has been eliminated (inertia between) as a measure of selection bias.
Here we consider the problem with categorical variables ($\mathbf{X}$) where some of them may be linked to an external categorical variable ($\mathbf{T}$). The information matrix (table 4.1) could be set by two disjunctive tables: the $\mathbf{K}$ matrix that represents I on rows and J on columns ; and the $\mathbf{T}$ matrix that represents I on rows and T on column [3].
The indicator matrix $\mathbf{K}$ (represented in a disjunctive form) has generic term $k_{ij} = \{K_{ij} : i \in I_n, j \in J_Q\}$,whit I_n as the population of n units under consideration and J_Q as the set of all categories of the Q pre-treatment considered covariates.

$$\mathbf{K} = \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & k_{ij} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}_{n \times J_Q} \qquad \mathbf{T} = \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & k_{it} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}_{n \times t}$$

The rows ($k_{i.}$) and columns ($k_{.j}$) margins and grand total ($k_{..}$) of the $\mathbf{K}$ matrix are formally expressed as follows:

$$k_{i.} = \sum_{j=1}^{Q} k_{ij} = Q \tag{4.4}$$

$$k_{.j} = \sum_{i=1}^{n} k_{ij} = k_{.j} \tag{4.5}$$

$$k_{..} = \sum_{i=1}^{n} \sum_{j=1}^{p} k_{ij} = nQ \tag{4.6}$$

The $\mathbf{T}$ matrix has generic term $k_{it} = \{k_{it} : i \in I_n, t \in T\}$ whit T as the set of level of the treatment indicator variable. $\mathbf{T}$ takes into account the structure induced by the selection

[3] the number of column equals the number of level of T

mechanism on the population. The rows ($t_{i.}$) and columns ($t_{.t}$) margins and grand total ($t_{..}$) of the **T** matrix could be expressed as follows:

$$t_{i.} = \sum_{t=1}^{t} k_{it} = 1 \tag{4.7}$$

$$t_{.t} = \sum_{i=1}^{n} k_{it} = k_{.t} \tag{4.8}$$

$$t_{..} = \sum_{t=1}^{T} k_{.t} = n \tag{4.9}$$

The row margins equal one given that each unit can receive only one treatment's level. The column margin equals $k_{.t}$ and represents the size of the group corresponding to the level t of T. The T variable generates a partition on the population I_n, the classes of that partition are defined as $I_{n_{(t)}}$.To each class $I_{n_{(t)}}$ corresponds a sub-table of the K-matrix; in this sense we can consider the **K** matrix as the juxtaposition of those t sub-tables with dimension $k_{.t}$.

We will also consider the **B** matrix with generic term $b_{tj} = \{b_{tj} : t \in T, j \in J_Q\}$, obtained from the **K** matrix by collapsing the rows of the **K** matrix corresponding to the same level $t \in T$.

$$\mathbf{B} = \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & b_{tj} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}_{t \times J_Q}$$

We call the **B** table as the **sum** table, it represents the average profile of each group, and gives an idea of whether margins differ with respect to T. If we consider, for example, a treatment variable with two levels (treated/not treated), the **B** matrix will have two rows and the number of columns will equal the number of J_Q. The **B** table has generic term b_{tj}.

$$b_{tj} = \sum_{i \in I_n} k_{ij} k_{it} = \sum_{i \in I_t} k_{ij} \tag{4.10}$$

The rows ($b_{t.}$) and columns ($b_{.j}$) margins and grand total ($b..$) of the **B** matrix could be expressed as follows:

$$b = \sum_{t=1}^{T}\sum_{j=1}^{Q} b_{tj} = nQ \tag{4.11}$$

$$b_{t.} = \sum_{j=1}^{Q} b_{tj} = Qk_{.t} \tag{4.12}$$

$$b_{.j} = \sum_{t=1}^{T} b_{tj} = k_{.j} \tag{4.13}$$

4.4.1 Profiles, metrics and weights

Here, for each table introduced in the previous paragraph (K, T and B), we consider profiles, metrics and weights. Let $D_{i.}^{(k)}$ and $D_{.j}^{(k)}$ the diagonal matrix of the row and column margins of the **K** matrix:

$$\mathbf{D}_{i.}^{(k)} = \begin{bmatrix} Q & & \\ 0 & Q & \\ 0 & 0 & Q \end{bmatrix} \qquad \mathbf{D}_{.j}^{(k)} = \begin{bmatrix} k_{.1} & & \\ 0 & k_{.j} & \\ 0 & 0 & k_{.q} \end{bmatrix}$$

Let $D_n^{(k)}$ and $D_p^{(k)}$ the weights of units and modalities of the **K** matrix, defined by the ratio between the margins and grand totals:

$$\mathbf{D}_n^{(k)} = Q \begin{bmatrix} \frac{1}{nQ} & & \\ 0 & \frac{1}{nQ} & \\ 0 & 0 & \frac{1}{nQ} \end{bmatrix} = \frac{1}{n} I_n \qquad \mathbf{D}_p^{(k)} = \begin{bmatrix} \frac{k_{.1}}{nQ} & & \\ 0 & \frac{k_{.j}}{nQ} & \\ 0 & 0 & \frac{k_{.q}}{nQ} \end{bmatrix} = \frac{1}{nQ} D_{.j}^{(k)}$$

Let the metric induced by the **K** matrix on the variables defined as the inverse of the weights $D_n^{(k)^{(-1)}} = nI_n$ and the metric on the units defined as $D_p^{(k)^{-1}}$.
Let $D_{i.}^{(k)^{-1}}\mathbf{K}$ the row profile of the **K** matrix with generic term $\{\frac{k_{ij}}{Q}\}_{(j)}$ and $D_{.j}^{(k)^{-1}}\mathbf{K}$ the column profile of the **K** matrix with generic term $\{\frac{k_{ij}}{k_{.j}}\}_{(i)}$.
Let $D_{i.}^{(T)}$ and $D_{.j}^{(T)}$ the diagonal matrix of the row and column margins of the **T** matrix:

$$\mathbf{D}_{i.}^{(T)} = \begin{bmatrix} 1 & & \\ 0 & 1 & \\ 0 & 0 & 1 \end{bmatrix} = I_n \qquad \mathbf{D}_{.j}^{(T)} = \begin{bmatrix} k_{.1} & & \\ 0 & k_{.t} & \\ 0 & 0 & k_{.T} \end{bmatrix}$$

Let $D_n^{(T)}$ and $D_p^{(T)}$ the weights of units and modalities of the **T** matrix, defined by the ratio between the margins and grand totals:

$$\mathbf{D}_n^{(T)} = \begin{bmatrix} \frac{1}{n} & & \\ 0 & \frac{1}{n} & \\ 0 & 0 & \frac{1}{n} \end{bmatrix} = \frac{1}{n} I_n \qquad \mathbf{D}_p^{(T)} = \begin{bmatrix} \frac{k_{.1}}{n} & & \\ 0 & \frac{k_{.t}}{n} & \\ 0 & 0 & \frac{k_{.T}}{n} \end{bmatrix} = \frac{1}{n} D_{.j}^{(T)}$$

Let the metric induced by the $\mathbf{T}$ matrix on the variables defined as the inverse of the weights $D_n^{(T)^{(-1)}} = nI_n$ and the metric on the units defined as $D_p^{(T)^{-1}} = nD_{.j}^{(T)}$.
Let $D_{i.}^{(T)^{-1}}$ $\mathbf{T}$ the row profile of the $\mathbf{T}$ matrix with generic term $\{k_{it}\}_{(t)}$ and $D_{.j}^{(T)^{-1}}\mathbf{T}$ the column profile of the $\mathbf{T}$ matrix with generic term $\{\frac{k_{it}}{k_{.t}}\}_{(i)}$.
Let $D_{i.}^{(B)}$ and $D_{.j}^{(B)}$ the diagonal matrix of the row and column margins of the $\mathbf{B}$ matrix:

$$\mathbf{D}_{i.}^{(B)} = \begin{bmatrix} Qk_{.1} & & \\ 0 & Qk_{.t} & \\ 0 & 0 & Qk_{.T} \end{bmatrix} = \qquad \mathbf{D}_{.j}^{(B)} = \begin{bmatrix} k_{.1} & & \\ 0 & k_{.j} & \\ 0 & 0 & k_{.q} \end{bmatrix}$$

Let $D_n^{(B)}$ and $D_p^{(B)}$ the weights of units and modalities of the $\mathbf{B}$ matrix, defined by the ratio between the margins and grand totals:

$$\mathbf{D}_n^{(B)} = \begin{bmatrix} \frac{Qk_{.1}}{nQ} & & \\ 0 & \frac{Qk_{.t}}{nQ} & \\ 0 & 0 & \frac{Qk_{.T}}{nQ} \end{bmatrix} = \frac{1}{n} D_{i.}^{(B)} \qquad \mathbf{D}_p^{(B)} = \begin{bmatrix} \frac{k_{.1}}{nQ} & & \\ 0 & \frac{k_{.j}}{nQ} & \\ 0 & 0 & \frac{k_{.q}}{nQ} \end{bmatrix} = \frac{1}{nQ} D_{.j}^{(B)}$$

Let the metric induced by the $\mathbf{B}$ matrix on the variables defined as the inverse of the weights $D_n^{(B)^{(-1)}} = nD_{i.}^{(B)^{-1}}$ and the metric on the units defined as $D_p^{(B)^{-1}} = nQD_{.j}^{(B)^{-1}}$.
Let $D_{i.}^{(B)^{-1}}$ $\mathbf{B}$ the row profile of the $\mathbf{B}$ matrix with generic term $\{\frac{b_{tj}}{Qk_{.t}}\}_{(j)}$ and $D_{.j}^{(B)^{-1}}\mathbf{B}$ the column profile of the $\mathbf{B}$ matrix with generic term $\{\frac{b_{tj}}{k_{.j}}\}_{(t)}$.

4.5 The inertia decomposition

The term inertia in correspondence analysis is used by analogy with the definition in applied mathematics of moment of inertia, which stands for the integral of mass times the squared distance to the centroid (e.g. Greenacre, 1984,p.35). A default analysis dealing with the factorial decomposition of the inertia related to the juxtaposition of the $\mathbf{K}$ matrix and the $\mathbf{T}$ matrix, when variables are categorical , is the Multiple Correspondence Analysis (MCA) that has the purpose of studying the marginal links between pairs of categorical variables in a given table and studying the structure induced by these variables on the units (Estadella et al., 2005). Multiple Correspondence Analysis (MCA) was proposed by Benzècri (1973) in his seminal work,*Analyse des Données*. MCA is an explorative multivariate technique for the analysis of any kind of matrix with nonnegative entries, but it principally involves table of frequency or counts with two or more dimensions in which make sense the sum by rows or by columns. Because it is oriented toward categorical data,

it can be used to analyze almost any type of tabular data after suitable data transformation or recoding. Given that the variability of a data matrix can be decomposed in eigenvalues and eigenvectors, and referring to the MCA for the study of the relationship between variables and of the structure induced by variables on the population, the presence of a conditioning variable will strongly influence the structure of the matrix decomposition. MCA produces a decomposition of the overall variance in eigenvalues and eigenvectors by a transition from an indicator matrix **K** to the Burt table. The latter consists of $q \times q$ [4] partitions created by each variable being tabulated against itself, and against the categories of all other variables.

Usually the MCA is carried out on the overall inertia that is the sum of all non-trivial eigenvalues, as shown in equation 4.14:

$$Inertia(X;T) = (\frac{1}{q}\sum_{i=1}^{q} j_i) - 1 = \frac{J}{Q} - 1 \tag{4.14}$$

where q is the number of variables and j_i is the number of categories of a generic variable $i \in Q$.

When the generated factorial space shows dependence of $\underline{X}$ from T, then the information matrix appears divided in two sub-tables each corresponding to a different treatment level; further, the cloud of units will appear as divided in different sub-clouds, each corresponding to a different treatment level. The results of applying classical method such as MCA to the juxtaposition of two different sub-tables can be affected by some problems when row margins are different or not proportional.

There are differences in margins when the sub-tables arises from different samples or different time points, or different treatment levels. In those situations results can be particularly affected by the differences between the inertias of the sub-tables: the higher the table's inertia, the grater is its influence on the overall analysis.

The conditional method we propose can be viewed as a particular case of partial factorial analysis, when the variable that causes the structure in the data is qualitative (e.g. the treatment). Generally, when two continuous variables X_1 and X_2 are dependent from an exogenous variable T, the partial analysis aims to measure the correlation coefficient $r(X_1, X_2)$. In the partial analysis are considered two n-dimensional populations represented in a R^n space by the n-dimensional vectors $\underline{X}_1$ and $\underline{X}_2$. Then the correlation's coefficient is computed by eliminating the effect due to Z. [5] When the exogenous conditioning variable is qualitative and covariates are categorical, is more complicated to study the relations between variables without the effect of the conditioning variable(Lebart et al., 1997).

The difficulty arises from the fact that researchers have to consider row and columns pro-

[4] where q represent the number of variables considered in the analysis

[5] Other studies in the continuous case are present in Rao,1964; Nonell, Thiç and Aluja, 2000

file and they have to take into account a metric more complex than the Euclidean metric, that is the well-known chi-square metric. To measure the influence of an exogenous conditioning variable T on the overall variability of a data matrix $\underline{X}$ we refer to Huygens' overall inertia decomposition as within-groups and between-groups(4.15):

$$\begin{aligned} I_{total} &= \sum_t D_p^{(T)} \|g_t - g\|^2 + \sum_t \sum_{i \in I_{n_{(t)}}} m_i^t \|x_i^t - g_t\|^2 \\ &= \sum_t \sum_{i \in I_{n_{(t)}}} m_i^t (x_i^t - g)' D_p^{(k)^{-1}} (x_i^t - g) \\ &= \sum_t D_p^{(T)} (g_t - g)' D_p^{-1} (g_t - g) + \sum_t m_i \sum_{i \in I_{n_{(t)}}} m_i^t (x_i^t - g_t)' D_p^{-1} (x_i^t - g_t) \\ &= I_{between} + I_{within} \end{aligned} \tag{4.15}$$

Where x_i^t is the unit i belonging to group t and m_i^t its mass; g is the global centroid and g_t are the T subcentroids. The $D_p^{(k)^{-1}}$ is the diagonal metric of the Euclidean space as defined by the inverse of the weight of each category $k_{.j}$ over the global mass of the **K** matrix. [6].

The $D_p^{(T)}$ term represents the weight of the categories, given by the amount of each categories over the total of individual in the overall population. ($D_p^{(T)} = \frac{k_{.t}}{n}$). Thus,in the case of the structure induced by the selection into treatment mechanism the two centroids are defined as in equation 4.16 and 4.17.

$$g = (\frac{k_{.j}}{nQ})_{j=1,\ldots,J_Q} \tag{4.16}$$

and

$$g_t = (\frac{b_{tj}}{Qk_{.t}})_{j=1,\ldots,J_Q} \tag{4.17}$$

As a consequence, the inertia between groups is given by:

$$I_{between} = \sum_{t=1}^{T} D_p^{(T)} (g_t - g)' D_p^{(k)^{-1}} (g_t - g) \tag{4.18}$$

with the metric $D_p^{(k)^{-1}}$ and weights $D_p^{(T)}$, thus:

[6] we consider the terms mass and weight as interchangeable

$$\begin{aligned} I_{between} &= \sum_t \frac{k_{.t}}{n} \sum_j \frac{nQ}{k_{.j}} (\frac{b_{tj}}{Qk_t} - \frac{k_{.j}}{nQ})^2 \\ &= \frac{1}{Q} \sum_t \sum_j \frac{b_{tj}^2}{k_{.t}k_{.j}} - 1 \end{aligned} \tag{4.19}$$

Therefore, the inertia within group is:

$$\begin{aligned} I_{within} &= I_{total} - I_{between} \\ &= \frac{J}{Q} - 1 - \frac{1}{Q} \sum_t \sum_j \frac{b_{tj}^2}{k_{.t}k_{.j}} - 1 \\ &= \frac{J - \sum_t \sum_j \frac{b_{tj}^2}{k_{.t}k_{.j}}}{Q} - 2 \end{aligned} \tag{4.20}$$

When we deal with the construction of new spaces representative of the original variability, the analysis can be decomposed, as in the Huygens inertia decomposition, in two parts: an between-groups analysis that analyzes the relative position of groups and an within-groups analysis that detects and describes differences between units within each group by not considering the effect due to the partition's structure.
Usually, in the evaluation context, this structure is induced by the non-random selection mechanism.
Aiming at measuring how much of imbalance exist in the data (the selection bias), we propose a factorial transformation that works for both qualitative and quantitative pre-treatment covariates, taking into account only the within-inertia in the decomposition of the information matrix $\mathbf{X}$ in eigenvalues and eigenvectors.

4.6 The conditional analysis as an intra analysis: the CORCO model

The multivariate measure of selection bias is obtained referring to an existing method known as the Conditional Multiple Correspondence Analysis (CORCO model), whose aim is to obtain a factorial decomposition by taking into account the inertia-within of a given data matrix (Escofier, 1988).
The original version of CORCO model (Escofier,1988) aimed at decondition the data matrix variability from the influence of an exogenous qualitative variable.
The author refers to the questionnaire analysis framework, when the same survey has been made at different points in time and when analysts are interested in time stable links

between units and variables rather than in temporal evolution.
There are few applications of the CORCO model to problems arising from real data, (see for example Mercedes, 2002). The CORCO model is new for the purpose of constructing a conditional space with treatment indicator T as the conditioning variable. In this sense it represents a new key tool to deal with selection bias in observational studies
We can consider the CORCO model according to different point of view: as an intra analysis , as an extension of the conventional MCA when an external qualitative variable generates a structure in the data pattern or as a partial dependence analysis.
The conventional MCA decomposition model is symmetric, because of the transition formula. Symmetry implies that it is equivalent to read a matrix by rows or by columns. In fact, it has been demonstrated (Escofier, 1988)that transition formula hold also in the CORCO model. For the reasons explained above, we could perform two separated but equivalent analysis: an intra analysis in the space of units (R^P) and an intra-analysis in the space of variables (R^N).

4.6.1 The conditional analysis in the R^P space: a geometric point of view

When there is dependence between $\underline{X}$ and T, the unit X-space generated by $\underline{X}$ will appear as divided in T subspaces, also if T has not been introduced in the analysis.
Geometrically, as shown in figure 4.1, if the data pattern differs too much with respect to different levels of T, then we will see well separated pattern of points. The conditional unit space ($R^P_{conditional}$) is obtained by centering the t sub-spaces of units with the same category t of T on its own center: each subspace gets translated to the origin (fig. 4.2). According to the Huygens inertia decomposition, the translation to the origin eliminates

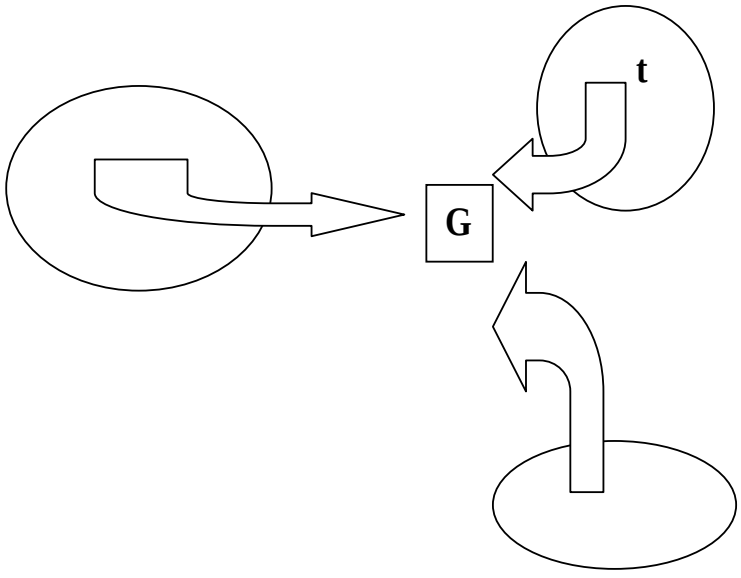

Figure 4.1: The unit space in the CORCO model

the inertia between. The unit i with category t of T will have the coordinates on the j-axe of the new conditional space as expressed in equation 4.21:

$$\frac{k_{ij}}{Q} - \frac{b_{tj}}{Qk_{.t}} = \frac{1}{Q}(k_{ij} - \frac{b_{tj}}{k_{.t}}) \tag{4.21}$$

where $\frac{k_{ij}}{Q}$ represents the row profile of the $K_{n\times Q}$ matrix and $\frac{b_{tj}}{Qk_{.t}}$ represents the row profile of the $B_{n\times Q}$ matrix. Particularly $\frac{b_{tj}}{Qk_{.t}}$ represents the average profile of the sub-cloud $I_{n_{(t)}}$ [7]in the space R^P.

$$\frac{1}{k_t}\sum_{i\in I_{n_{(t)}}}\frac{k_{ij}}{Q} = \frac{1}{QK_{.t}}\sum_{i\in I_{n_{(t)}}} k_{ij} = \frac{b_{tj}}{Qk_{.t}} \tag{4.22}$$

4.6.2 The conditional analysis in the R^n space: a geometric point of view

According to the MCA, the column profiles of the J categories of the Q variables in the K matrix, indicated as $\frac{k_{ij}}{k_{.j}}$, are located in the R^n space.
The R^n space could be decomposed in T orthogonal components whose number is the same as the number of treatment indicator variables (t of T)[8]. We indicate the t-dimensional subspace of R^n generated by the t indicator variables t as R^T. R^T explains the structure induced by the selection mechanism.
To obtain a conditional variable space we will project the original space R^n on the space orthogonal to R^T.

$$R^n \rightarrow^{\perp} R^T \tag{4.23}$$

It is a two-stage procedure that involves the following.
In the first stage, we project the column profile $\frac{k_{ij}}{k_{.j}}$ of the $\mathbf{K}$ matrix onto R^T, that is the subspace generated by the modalities of T being the indicator vectors of the modalities (k_{it}) are orthogonal. After some algebra, the coordinate of the $\frac{k_{ij}}{k_{.j}}$ profile on k_{it} is expressed as in equation 4.24:

$$\frac{\sum_i(\frac{k_{ij}k_{it}}{k_{.j}})}{\sum_i(k_{it})^2} = \frac{b_{tj}}{k_{.j}k_{.t}} \tag{4.24}$$

[7] $n_{(t)}$ indicates the number of units of the population under consideration that belongs to group t
[8] We construct a conditional variable space ($R^n_{conditional}$) voided of any influence of the selection mechanism referring to the space decomposition in orthogonal and supplementary subspaces and to the definition of direct sum of vector spaces.

Then the distance between j and j' projections is given by:

$$\begin{aligned} D^2(j_{proj}, j'_{proj}) &= \sum_t \sum_{i \in I_{n_{(t)}}} [(\frac{b_{tj}}{k_{.j}k_{.t}}) - (\frac{b_{j't}}{k_{.j'}k_{.t}})]^2 n \\ &= \sum_t (\frac{b_{jt}}{k_{.j}} - \frac{b_{j't}}{k_{.j'}})^2 \frac{n}{k_{.t}} \end{aligned} \tag{4.25}$$

The distance in 4.25 is exactly the chi-square distance between j and j' profiles in the $B_{n\times Q}$ table; it represents the distance induced by the selection mechanism on the J categories.

In the second stage, the structure induced by the selection mechanism is eliminated by making a projection on the space orthogonal to R^T ($^{\perp}R^T$).
The category j in the conditional space will have the coordinate expressed as in equation 4.26.

$$\frac{k_{ij}}{k_{.j}} - \frac{b_{jt}}{k_{.j}k_{.t}} = \frac{1}{k_{.j}}(k_{ij} - \frac{b_{jt}}{k_{.t}}) \text{ , i } \in I_{n_{(t)}} \tag{4.26}$$

Geometrically, the Huygens inertia decomposition in the variable space corresponds to the orthogonal projection of columns profile on two subspaces of R^n: the R^T subspace for what concerns the inertia *between* and the space orthogonal to R^T, for what concerns the inertia *within*. With the distance induced by the $K_{n\times Q}$ matrix as $D_{I_n}(j, j')$, with the distance induced by the $B_{n\times Q}$ table as $D_T(j, j')$, and with the distance considered in the CORCO model as $D_{I_n|T}(j, j')$, the Huygens Inertia decomposition could be rearranged in terms of distance as follows:

$$\begin{aligned} D^2_{I_n}(j, j') &= D^2_T(j, j') + D_{I_n|T}(j, j') \\ D^2_{TOTAL} &= D^2_{BETWEEN} + D^2_{WITHIN} \end{aligned} \tag{4.27}$$

It clearly emerges that we are able to measure how much of distance is due to the selection mechanism and how much is not.

4.7 The conditional analysis: an algebric point of view

Given the matrix $K_{n\times J_Q}$, $T_{n\times t}$, and $B_{t\times J_Q}$, the aim of the CORCO model is to analyze only the variability *within* after the elimination of the variability *between*.
The variability *between* is that associated to the structure induced by the selection into

treatment, we will indicate this structure as **model**.

The conventional Correspondence Analysis (CA), for example, analyzes the differences between a frequency table and a model defined as the product of the marginal distribution of the frequency table; that is an independence model between the two variables considered. The ACM analyzes the difference between each profile with respect to the theoric independence model.[9] Escofier(1987) has demonstrated that the analysis of the divergency between a given frequency table and an independence model could be generalized to the analysis of the differences between a generic data matrix and a generic model.

In doing the conditional analysis we consider as model table one that represents the structure induced by T that has the same margins as the disjunctive table $K_{n\times Q}$.

The model table is indicated as **M**.

$$\mathbf{M} = \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \frac{b_{tj}}{k_{t.}} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}_{n\times J_Q}$$

M has generic term $\frac{b_{tj}}{k_{t.}}$ with $\{\frac{b_{tj}}{k_{t.}} : i \in I, j \in J_Q\}$.

The inertia associated to the **M** matrix is the inertia between. The numerator (b_{tj}) represents the number of units with categories j in the treatment group t of T.

The denominator $(k_{t.})$ represents the number of unit in the group t of T. In **M** all rows related to units i in the same class t of T are identical.

These rows represent the profile of how the number of units in T are distributed along each categories J_Q of Q.

The **M** matrix is not a disjunctive table, but, for each row, the sum of values corresponding to the categories of the same variable equals 1. The row margins of **M** are constant and equals Q, as in the **K** matrix.

The column margins in **M** are the same as the column margin in **K**.

Specifically,

$$m_{i.} = \sum_{j=1}^{Q} \frac{b_{tj}}{k_{t.}} = Q \tag{4.28}$$

$$m_{.j} = \sum_{i=1}^{n} \frac{b_{tj}}{k_{t.}} = \sum_{t=1}^{T} \sum_{i\in I_{n_{(t)}}} \frac{b_{tj}}{k_{t.}} = \sum_{t} b_{tj} = \sum_{I\in I_n} k_{ij} = k_{.j} \tag{4.29}$$

The row profile of the **M** table is $\frac{b_{tj}}{Qk_{t.}}$ and the column profile is $\frac{b_{tj}}{k_{.j}k_{t.}}$. Given that both the **K** and **M** matrix have the same margins, both the metrics and weights are the same as those considered for **K**. To analyze only the variability within, that is the part independent

[9] In $R^n \rightarrow \frac{k_{ij}}{nQ} - \frac{k_{.j}}{nQ}$ and in $R^P \rightarrow \frac{k_{ij}}{k_{.j}} - \frac{1}{n}$

from the selection to treatment, could be performed a conventional correspondence analysis of the **K*** table derived as follows:

DATA-MODEL+MARGINS PRODUCT/POPULATION SIZE

The matrix to diagonalize, $\mathbf{K}^*$ has generic term k^*_{ij}.

$$\begin{aligned} k^*_{ij} &= k_{ij} - (\frac{b_j t}{k_t}) + \frac{k_{.j}Q}{nQ} \\ &= k_{ij} - (\frac{b_j t}{k_t}) + \frac{k_{.j}}{n} \end{aligned} \tag{4.30}$$

The $\mathbf{K}^*$ matrix with I on rows and J on columns has the same margins as both the $\mathbf{K}$ and $\mathbf{M}$ matrix.

$$\mathbf{K*} = \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & k_{ij*} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}_{n \times J_Q}$$

In particular:

$$\begin{aligned} K^*_{i.} &= Q \\ k^*_{.j} &= k_{.j} \\ k^*_{..} &= n \times Q \end{aligned} \tag{4.31}$$

Then, we consider both the row-profiles (4.32) and the column-profiles (4.33)

$$D^{(k)^{-1}}_{i.} K^* = \frac{k_{ij}}{Q} - \frac{b_{jt}}{Qk_{.t}} + \frac{k_{.j}}{nQ} \tag{4.32}$$

$$\begin{aligned} D^{(k)^{-1}}_{.j} K^* &= \frac{k_{ij}}{k_{.j}} - \frac{b_{jt}}{k_{.j}k_{.t}} + \frac{k_{.j}Q}{nQ}\frac{1}{k_{.j}} \\ &= \frac{k_{ij}}{k_{.j}} - \frac{b_{jt}}{k_{.j}k_{.t}} + \frac{1}{n} \end{aligned} \tag{4.33}$$

Thus, the object of analysis in the R^p space will be A (4.34), with generic term $a_{jj'}$ (4.35)

$$\begin{aligned} A &= K^{*\prime} D_n^{(k)} k_n^* Q D_{.j}^{(k)^{-1}} \\ &= K^{*\prime} \frac{1}{n} I_n K^* n Q D_{.j}^{(k)^{-1}} \\ &= Q K^{*\prime} K^* D_{.j}^{(k)^{-1}} \end{aligned} \tag{4.34}$$

$$a_{jj'} = Q \sum_i \frac{k_{ij}^* k_{ij'}^*}{k_{.j}} \tag{4.35}$$

In particular,

- K^* has generic term k_{ij}^* which indicates the coordinate of units in the conditional variable space
- $D_n^{(k)}$ is the diagonal matrix of weights with generic term $d_{ii} = \frac{1}{n} I_n$
- $D_p^{(k)^{-1}}$ is the diagonal metric with generic term $d_{jj} = \frac{nQ}{k_{.j}}$

Escofier (1988) has demonstrated that the transition formula hold: they link the R^P space to the dual R^n space.
It has also been demonstrated the equivalence between the analysis of the K^* table and the Burt table B^* with generic term $b_{jj'}^*$.

$$b_{jj'}^* = \sum_i k_{ij}^* k_{ij'}^* \tag{4.36}$$

By substituting the term $k*_{ij}$ in 4.36 then:

$$\begin{aligned} b_{jj'}^* &= \sum_i k_{ij} k_{ij'} - \sum_i \frac{b_{jt}}{k_t} \frac{b_{j't}}{k_t} + \sum_i \frac{k.j}{n} \frac{k_{.j'}}{n} \\ &= b_{jj'} - \sum_t \frac{b_{jt} bj't}{k_t} + \frac{1}{n} k_{.j} k_{.j'} \end{aligned} \tag{4.37}$$

with $b_{jj'}$ as the general term of un unconditional Burt table that crosses J and J'.

4.8 STRATEGY 1: Inference in the conditional analysis

Our aim is going deeper within the relationships existing between variables. We analyze the association among variables within the framework of multivariate descriptive analysis,

by using the inertia as a measure of association between variables.
Based on the decomposition of total inertia into between-inertia and within-inertia, we first compute a bias elimination coefficient (BEC), then we test the significance of the partition induced by the non random selection process using the asymptotical distribution of the between inertia (Estadella et al., 2005) based on a chi-square distribution.

4.8.1 The bias elimination coefficient(BEC)

Once obtained the new coordinates voided of any influence of the selection mechanism [10], we are able to derive a bias elimination coefficient (BEC) that tell us whether the influence of conditioning is important or not. In fact, the BEC will establish the dependence of a set of qualitative variables on a model generated by a conditioning variable. How much of the inertia between has been eliminated will be determined by one minus the ratio between the inertia-within relative to the total inertia (4.38).

$$BEC = 1 - \frac{I_{within}}{I_{total}} \tag{4.38}$$

with the total inertia as the inertia of the unconditional X space (MCA), and the within inertia as that of the conditional space obtained after inertia-between has been eliminated (CORCO model).

4.8.2 The multivariate test of imbalance

From literature we know that when the conditioning variable defines a random partition, the $I_{between}$ approaches zero and the $I_{within} \cong I_{total}$ (Estadella et al., 2005).
Then to determine how much important is the inertia between with respect to the hypothetical case of a random partition ($I_{between} = 0$), we need to perform an hypothesis test. We specify the null hypothesis as follows:

$$H_0 : I_{within} = I_{total} \Longrightarrow \text{ no dependence between X and T}$$

If we do not reject the null hypothesis then the observed covariates are not related to the selection into treatment.
If we have considered the right covariates involved in the assignment mechanism then we can consider the inertia between that has been eliminates as the correct global measure of imbalance.
In order to assess if the detected imbalance is significant we use results obtained by Estadella et al. (2005) in studying the distribution of inertia.

[10] the coordinates of the CORCO model

Estadella et al. (2005) have derived the distribution of inertia between, with the aim of assessing when the conditioning variable gives different results with respect to the unconditional analysis in order to determine whether conditioning is significant.
Specifically, Estadella et al. (2005) have derived the distribution of inertia between under the null hypothesis of a random partition. As starting point, they use the Burt band which may be considered as a contingency table with marginals: $(k_{.j})_{j=1,...,J}$, $(Qk_{.t})_{t=1,...,T}$ and grand totals nQ. The Burt Band has dimension $J \times T$, generic term $\{b_{jt}\}$ and it crosses the categories of the pre-treatment covariates considered with the modalities of the conditioning variable T.

$$\mathbf{B}_{Band} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1T} \\ b_{21} & b_{22} & \cdots & b_{2T} \\ b_{j1} & b_{j2} & \cdots & b_{jT} \\ b_{J1} & b_{J2} & \cdots & b_{JT} \end{bmatrix}_{J \times T}$$

The chi-square coefficient [11] (4.39) of the Burt Band is represented by equation 4.39:

$$\chi^2_{Burt_{band}} = \sum_t \sum_j \frac{(b_{jt} - \frac{k_{.j}Qk_{t.}}{nQ})^2}{\frac{k_{.j}Qk_{t.}}{nQ}} = n \sum_t \sum_j \frac{b^2_{jt}}{k_{.j}k_{t.}} - nQ \tag{4.39}$$

It clearly emerges from equation 4.39 that the chi-square coefficient is exactly nQ times the value of the inertia between (4.19).

$$\begin{aligned} \chi^2_{Burt_{band}} &= nQI_{between} \\ \chi^2_{Burt_{band}} &= nQ \underbrace{(\frac{1}{Q} \sum_t \sum_j \frac{b^2_{jt}}{k_{t.}k_{.j}} - 1)}_{I_{Between}} \\ &= n \sum_t \sum_j \frac{b^2_{jt}}{k_{t.}k_{.j}} - nQ \end{aligned} \tag{4.40}$$

As a consequence,

$$I_{between} = \frac{1}{nQ} \chi^2_{Burt_{band}} \tag{4.41}$$

[11] see Josep Daunis i Estadella PhD thesis pp.146, 2004

Therefore, since the $\chi^2_{burt_{band}}$ coefficient, under the assumption of independence, has asymptotically a χ^2 distribution function with $(T-1)(J-1)$ degrees of freedom [12], then the inter-groups inertia ($I_{between}$) , under the same assumption has a scaled χ^2 distribution function.

$$I_{between} \sim \frac{\chi^2}{nQ} \tag{4.42}$$

Thus, under the null hypothesis of a random partition they assume that:

$$I_{between} \sim \frac{\chi^2_{(T-1)(J-1)}}{nQ} \tag{4.43}$$

With moments:

$$\begin{aligned} E(I_{between}) &= \frac{(T-1)(J-1)}{nQ} \\ Var(I_{between}) &= \frac{2(T-1)(J-1)}{(nQ)^2} \end{aligned} \tag{4.44}$$

Estadella et al.(2005) have established the confidence intervals for inertia between actually obtained with conditional MCA. Particularly at the α value, the confidence interval for the inertia between will be determined as in equation 4.45.

$$I_{between} \in (0, \frac{\chi^2_{(T-1)(J-1),\alpha}}{nQ}) \tag{4.45}$$

If the $I_{between}$ calculated on the specific data set under analysis is out of the confidence interval, then the null hypothesis is rejected.
To reject the null hypothesis makes us sure that, given all covariates involved in the selection mechanism, it doesn't exist dependence between the information matrix $\underline{X}$ and T, or if it exist, it is not statistically significant.

4.8.3 How to measure imbalance: a toy example

We consider a simple case in which there are 18 units, and a treatment binary variable T. The information matrix $\mathbf{X}$ is composed of three categorical pre-treatment covariates: X_1 with two categories, X_2 with three categories and X_3 with four categories. First, we

[12] with T as the number of level of the treatment indicator and J as the number of levels of the Q variables considered

implement a conventional MCA, carried out on the overall inertia. According to 4.14, $I_{total} = 2$ with Q=3 and J=9. The MCA was carried out on the Burt table (tab. 4.2).

	$X_{1=1}$	$X_{1=2}$	$X_{2=1}$	$X_{2=2}$	$X_{2=3}$	$X_{3=1}$	$X_{3=2}$	$X_{3=3}$	$X_{3=4}$
$X_{1=1}$	8								
$X_{1=2}$	0	10							
$X_{2=1}$	3	4	7						
$X_{2=2}$	1	4	0	5					
$X_{2=3}$	4	2	0	0	6				
$X_{3=1}$	3	2	4	0	1	5			
$X_{3=2}$	0	3	0	1	2	0	3		
$X_{3=3}$	1	2	1	2	0	0	0	3	
$X_{3=4}$	4	3	2	2	3	0	0	0	7

Table 4.2: The Burt Table

Id	**t**	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**	**Factor 6**
1	1	1.08	0.58	0.25	-0.24	-0.32	-0.29
2	1	-0.66	-0.95	1.05	-0.29	0.28	-0.04
3	1	-1.22	0.79	-0.57	-0.49	0.05	0.32
4	1	0.09	0.24	0.02	0.98	0.54	-0.06
5	1	-0.72	0.00	-0.48	0.81	-0.46	0.21
6	1	0.51	-0.98	-0.60	-0.01	0.15	0.11
7	0	0.51	-0.98	-0.60	-0.01	0.15	0.11
8	0	0.18	0.79	-0.58	-0.82	0.80	-0.55
9	0	0.09	0.24	0.02	0.98	0.54	-0.06
10	0	-1.22	0.79	-0.57	-0.49	0.05	0.32
11	0	-1.29	-0.21	0.67	0.04	-0.59	-0.54
12	0	0.51	-0.98	-0.60	-0.01	0.15	0.11
13	0	1.08	0.58	0.25	-0.24	-0.32	-0.29
14	0	0.90	-0.39	0.14	-0.73	-0.45	0.48
15	0	-0.66	-0.95	1.05	-0.29	0.28	-0.04
16	0	0.48	0.83	0.76	0.26	-0.07	0.31
17	0	0.48	0.83	0.76	0.26	-0.07	0.31
18	0	-0.12	-0.24	-0.98	0.31	-0.72	-0.40

Figure 4.2: Units coordinates in the conventional MCA space

The coordinates of the new space generated by MCA (fig. 4.2)cannot be used for evaluation purposes given that this space has been generated by some individual characteristics associated with the assignment mechanism. In fact, by computing the means of each factor (tab. 4.3), they differ between treated and untreated.

Whereas, by implementing a conditional MCA we obtain the coordinates in fig. 4.3, voided of any dependence from T.

	mean(t=1)	**mean (t=0)**
factor 1	-0.15	0.08
factor 2	-0.05	0.03
factor 3	0.06	0.03
factor 4	0.13	-0.06
factor 5	0.04	-0.02
factor 6	0.04	-0.02

Table 4.3: means of factors in MCA

id	**t**	**factor 1**	**factor 2**	**fcator 3**	**factor 4**	**factor 5**	**factor 6**
1	1	0.32	-1.29	0.44	0.22	-0.28	-0.33
2	1	0.13	0.32	-1.33	0.51	0.42	-0.06
3	1	-1.21	0.33	0.17	-0.14	-0.32	0.23
4	1	0.17	-0.24	0.21	-0.70	0.60	-0.09
5	1	-0.18	0.56	0.10	-0.68	-0.50	0.19
6	1	0.76	0.32	0.42	0.79	0.08	0.06
7	0	0.71	0.56	0.37	0.64	0.13	0.13
8	0	-0.59	-0.31	0.75	0.36	0.47	-0.60
9	0	0.12	0.00	0.16	-0.85	0.66	-0.03
10	0	-1.26	0.57	0.12	-0.29	-0.26	0.29
11	0	-0.55	0.72	-1.18	-0.44	-0.46	-0.49
12	0	0.71	0.56	0.37	0.64	0.13	0.13
13	0	0.27	-1.05	0.39	0.07	-0.23	-0.27
14	0	0.54	-0.41	0.08	0.88	-0.39	0.50
15	0	0.08	0.56	-1.38	0.36	0.48	0.00
16	0	-0.05	-0.96	-0.13	-0.60	0.14	0.35
17	0	-0.05	-0.96	-0.13	-0.60	0.14	0.35
18	0	0.09	0.72	0.57	-0.16	-0.80	-0.36

Figure 4.3: Units coordinates in the conditional space

Table 4.4 shows that the means of each factor, do not differ between treated and untreated.

Once obtained the conditional space, we are able to assess how much of the inertia between has been eliminated. By considering the $Burt_{Band}$ (tab. 4.5) and according to 4.19

then,

$$I_{between} = 0.0242 \tag{4.46}$$

and

$$BEC = 1 - \frac{I_{within}}{I_{total}} = \frac{1.9758}{2} = 0.0121 \tag{4.47}$$

	mean(t=1)	**mean (t=0)**
factor 1	0	0
factor 2	0	0
factor 3	0	0
factor 4	0	0
factor 5	0	0
factor 6	0	0

Table 4.4: means of factors in conditional MCA

	$X_{1=1}$	$X_{1=2}$	$X_{2=1}$	$X_{2=2}$	$X_{2=3}$	$X_{3=1}$	$X_{3=2}$	$X_{3=3}$	$X_{3=4}$
$\mathbf{t=1}$	2	4	2	2	2	1	1	1	3
$\mathbf{t=0}$	6	6	5	3	4	4	2	2	4

Table 4.5: The Burt Band

Under the random partition hypothesis, with $\alpha = 0.05$ the confidence interval for the inertia between is:

$$I_{between} \in (0, \frac{\chi^2_{(2-1)(9-1),\alpha}}{18 * 3}) = (0; 0.28) \tag{4.48}$$

The resulting inertia of the conditioning by T (0.0242), remains inside the confidence interval, showing the independence of the variables $\mathbf{X}$ from the variable T.

4.9 Some properties of the conditional space

As Escofier (1988) has shown, and as mentioned before, the CORCO model has the same properties as the MCA:

- Constructing and projecting two spaces (R^n and R^p) on their main principal axes.
- Duality and transition formula from units space to variable space and vice versa (the conventional barycentric formula hold)
- Equivalence with the analysis of a table like a Burt table where the contingency tables are conditioned to T.

Further, Escofier (1988) has shown important guidelines for what concerns the interpretation of the distance in both units space and variable space.
A numeric example due to Escofier (1988) highlights how part of distance induced by the exogenous variable T could be eliminated by the construction of a conditional factorial space. The example concerns the questionnaire analysis, the field in which the CORCO model has been introduced and developed.
Suppose that a question has 4 items and that the interviewed population is equally partitioned over the 4 items.

According to MCA two units, i and i', that have chosen the first item will have as coordinates those represented in table 4.6.

	X			
	X_1	X_2	X_3	X_4
Unit i and i'	1	0	0	0
Total population	0.25	0.25	0.25	0.25
Units i and i' centered	0.75	-0.25	-0.25	0.25

Table 4.6: Both units have chosen the first item

If the population of i has chosen the first item, then according to the CORCO model, the i-coordinates will equal zero, due to the translation to the origin (tab. 4.7).

	X			
	X_1	X_2	X_3	X_4
Population of i	1	0	0	0
Unit i	0	0	0	0

Table 4.7: i has chosen the same item as the population

If the population of i' is partitioned with certain percentages, i.e. - 0.1, 0, 0.4, 0.5 - then the i'-coordinates are those represented in table 4.8 .

	X			
	X_1	X_2	X_3	X_4
Population of i'	0.1	0	0.4	0.5
Unit i'	0.9	0	-0.4	-0.5

Table 4.8: the i' coordinates

The higher is the difference between the unit i and the population to which it belongs to, the higher will be its distance from the origin. In the conditional space the units that have a different answer's profile with respect to the group to which they belong to will be located in the extremity of the axe in the light of the chi-square metric.
Two units member of the same sub-population (i.e. both members of the treatment group) will be close each other if the difference between their coordinates and the means of the reference group is the same; they will be distant if their answer's profile is different and rare.
Two units close each other in the MCA could be very distant in the CORCO model.
In fact, if two units have chosen the same item j, they could not be close if this item is the reference situation of an unit but not for both. For the reasons explained above the relative distance between units are different from the distance obtained by MCA.

4.10 STRATEGY 2: estimating local average causal effects

The quantity of interest is represented by the average treatment effect on local spaces. From the property of the distance in the conditional space, it clearly emerges that, in order to estimate the local causal effect of interest, comparable units cannot be found on the conditional space.
This is due to the fact that distance between units in the conditional space does not take into account the part of distance due to T.
Thus, we propose the use of Cluster Analysis (CA) on the coordinates obtained with a Multiple Correspondence Analysis (MCA) as a tool to find local groups of comparable units on which estimate local average causal effects.
CA is not new in the literature on evaluation. Henry and McMillan(1993), for example, compare three different matching techniques: index groupings, cluster groupings, benchmark groupings.Specifically, Cluster group matching uses cluster analysis to embed the treatment group in a cluster of similar controls. Their simulations suggest that cluster and benchmark methods work better than index matching(Nenry and McMillan,1993).
Another example of CA application is in Peck(2005). She proposes using cluster analysis to identify subgroups within experimental data, with the aim of understanding variation in program impacts that accrues across heterogeneous populations.
CA is an atheoretical, mathematically based technique that seeks to maximize heterogeneity between clusters while minimizing heterogeneity within clusters(Peck, 2005).
The result is groupings of like observations in terms of covariates, that are different from other groupings.
In particular, we use the hierarchical approach, in which the process of clustering proceeds sequentially such that at each step only one unit or group of units changes group membership and the group at each step are nested with respect to previous groups. Once, an unit has been assigned to a group it is never removed from that group. (Jobson, 1992). The clustering process will produce any number of clusters, ranging, in theory, from one cluster per observation (where each observation is its own cluster) to one cluster containing all n units, where all observations are in the same group.
We use an agglomerative hierarchical process meaning that as the process moves from n clusters to one cluster, the sizes of the clusters increase and the number of cluster decrease. Usually,the process begins with the Euclidean or standardized Euclidean distance matrix. At each step in the process the proximity measures are updated according to an optimality criterion value. The optimality criterion value is the closest proximity value among groups at that stage of process (Jobson, 1992). The proximity value is determined by the specific method used. Assuming that Euclidean distance is used as dissimilarity measure, then the single linkage approach uses the smallest possible Euclidean distance measure between objects in the two groups; the complete linkage uses the largest possible distance between objects in the two groups, and the average linkage approach uses the average distance.

Whereas, Ward's method uses an analysis of variance approach to evaluate the distances between clusters. In short, it attempts to minimize the sum of squares (SS) of any two hypothetical clusters that can be formed at each step. The hierarchical clustering process does not provide a single cluster solution: each step of the process represents a cluster solution. To determine the appropriate number of clusters we need to select one of the steps of the hierarchical process using a second optimality criterion.
Usually, the hierarchical process can be represented using a tree diagram (dendrogram), and the choice of how many cluster retain, depends on analysis purposes.
Empirically, the appropriate number of clusters can be identified by examining the cut points in groups depicted on a dendrogram.
Generally, the goal is to optimize the relative amounts of within- and between- clusters heterogeneity; but, for this particular problem - finding groups of comparable units- the goal is to generate a cluster solution that results in a number of subgroups that meets the criterion of achieving balance between treated and untreated as best as possible.
Thus, we propose the use of the multivariate test of imbalance as a stopping rule.
The basic idea is that, given a partition represented in a dendrogram, the lower is the cut level (maximum number of groups), the higher is the possibility of achieving balance within groups, and the achievement of balance is checked by performing the multivariate test of imbalance.
Then, if the test tell us that no dependence between X and T exists within the specific group, a local average treatment effect is estimated in an unbiased way.
Given a n-clusters solution set, if in some groups balance is not achieved, we can decide to perform a finer partition (more clusters) or stop the analysis by discarding units that belong to non balanced groups.

Chapter 5

Testing the new multivariate method via simulated data

5.1 Introduction

A simulation study is performed in order to evaluate the performance of the method in identifying selection bias according to the dependence structure of the data.
In particular, we first test the ability of the method to check selection bias when a data set is available before any analysis; second, we test how the method is able to check if balancing is achieved after a propensity score analysis is performed; finally, we propose a cluster analysis as a strategy to find groups of comparable units before the causal effect estimation, and we use the multivariate test of imbalance as a stopping rule in choosing the correct number of clusters.
The presence of selection bias is evaluated by establishing a confidence interval for inertia between under the null hypothesis of a random partition and by making tests for the values of inertia actually obtained.

5.2 Simulation:the assignment to treatment is not random but the selection process is perfectly known

5.2.1 Data and assumptions

The analysis could be considered as a toy example rather than a simulation in the strict sense.
We designed four qualitative [1] pre-treatment covariates: X_1 with two levels, X_2 with two levels, X_3 with three levels, and X_4 with two levels. We considered all 24 combinations of those covariates.
We assume that the variables in the data are measured without error. We adopt this general setup for expository reasons. Then, we consider a binary treatment variable T,

[1] continuous covariates could be discretized

that equals 1 for treated and equals 0 for untreated.
For each combination of covariates, units were assigned with different proportions (π) to different levels of treatment, in order to create dependence between X and T. (Appendix B, tab. B.1). For each unit i we have constructed the potential outcome as in equations 5.1 and 5.2

$$Y(1) = 2.2X_1 + 0.2X_2 + 1.8X_3 + 2.8X_4 \tag{5.1}$$

$$Y(0) = Y(1) - \overline{Y(1)} \tag{5.2}$$

We have chosen to construct the potential outcomes without error only for expository reasons, with the aim of exactly check if the method detects imbalance and ,as a consequence, the true average treatment effect. In the simulated data $\overline{Y(1)} = 10.8321$ and, by design, it represents the true average treatment effect. We assume no omitted variable bias, such that conditional on $\underline{X}$, the treatment variable indicator T is independent of the potential outcomes (5.3):

$$P(T \mid \underline{X}, Y(0), Y(1)) = P(T \mid \underline{X}) \tag{5.3}$$

The observed outcome is expressed as in equation 5.4 :

$$Y_{i,obs} = T_i Y_i(1) + (1 - T_i) Y_i(0) \tag{5.4}$$

The naive estimator of the average causal effect is then defined as in eq 5.5:

$$\widehat{\delta}_{naive} \equiv E_N[Y_{i,obs} \mid T_i = 1] - E_N[Y_{i,obs} \mid T_i = 0] \tag{5.5}$$

which is simply the difference in the means of the observed outcome variable $Y_{i,obs}$ for the observed treatment and control units in the full data set considered (N=764). With $E_N[Y_{i,obs} \mid T_i = 1]$ as the mean of the outcome for those observed in the treatment group; and with $E_N[Y_{i,obs} \mid T_i = 0]$ as the mean of the outcome for those observed in the control group. We are interested on the causal effect of a treatment indicator variable T_i on an observed outcome $Y_{i,obs}$.
But, the naive estimator is an inconsistent estimator of the average treatment effect ($\widehat{\delta}_{naive} = 8.4051$). It corresponds to the coefficient γ of T_i in a bivariate regression model:

$$Y_{i,obs} = \alpha + \gamma T_i + \varepsilon_i \tag{5.6}$$

that will yield an estimated coefficient $\widehat{\gamma}$ that represents a biased and inconsistent estimate of the causal effect of interest. This is due to the fact that, when the assignment to treatment is not random, the causal variable T_i, usually, is associated with variables involved in the selection process and embedded in the error term ε_i. In literature,the standard regression solution is to estimate an expanded regression equation, by considering the set of background covariates $\underline{X}$,assumed to predict both T_i and $Y_{i,obs}$,

$$Y_{i,obs} = \alpha + \gamma T_i + \beta' \underline{X}_i + \varepsilon_i \tag{5.7}$$

The γ coefficient of the expanded regression represents the unbiased causal effect.

Regression type	**causal effect**	
bivariate	γ=8.4051	biased
expanded	γ=10.8321	unbiased

Table 5.1:

5.2.2 The assessment of selection bias

To assess the level of selection bias that arises from the non random selection mechanism we have first performed a conventional MCA, then a conditional MCA (CORCO).
Conventional MCA was carried out on the overall inertia of the information matrix **X**. According to 4.14, I_{total}=1.25 with Q=4 and J=9. [2]
The resulting coordinates of the MCA space show that the space has been generated by some individual characteristics that are associated with the assignment mechanism. In fact, by computing the means of scores for each factor, we can note that some of them differ between treated and untreated (table 5.2).

factor	**mean(t=1)**	**mean(t=0)**	**t**	**pr > |t|**
factor1	0.5311	-0.271	25.71	< .0001
factor2	0.0572	-0.029	2.13	0.0333
factor3	0.0655	-0.033	2.54	0.0111
factor4	-0.183	0.0934	-8.05	< .0001
factor5	0.233	-0.119	11.74	< .0001

Table 5.2: Means of scores for each factor (MCA)

[2] where Q represents the number of covariates included in the analysis and J represents the number of categories of the Q covariates considered

Then, we performed the conditional MCA (CORCO), with T as the conditional variable and $\underline{X}$ as the covariates introduced in the analysis.
The resulting coordinates of the conditional space show no dependence from T. In fact, the means of each factor equals 0 for both treated and untreated (table 5.3). The level

factor	**mean(t=1)**	**mean(t=0)**	**t**	**pr > |t|**
factor1	0	0	0	1.000
factor2	0	0	0	1.000
factor3	0	0	0	1.000
factor4	0	0	0	1.000
factor5	0	0	0	1.000

Table 5.3: Means of scores for each factor (CORCO)

of selection bias is represented by the amount of the inertia between that has been eliminated in obtaining the conditional space. By considering both the Burt Table (table B.2, appendix B) and the Burt Band (table B.3, appendix B) of the conditional space and according to 4.19 it results $I_{between} = 0.1924$. The amount of inertia between that has been eliminated with respect to the total inertia is determined by computing the BEC according to 4.38 giving as result 15% of total inertia.
Under the random hypothesis, level $\alpha = 0.05$, $\chi^2_{8;0.05} = 15.51$, $n = 764$, the confidence interval for the $I_{between}$ is (0;0.0052). The resulting inertia of the conditioning by T remains outside the interval, showing the dependence of the $\underline{X}$-variables from the variable T. It means that the amount of conditioning by T is significant and that data are not balanced between treated and controls.

5.2.3 The propensity score model

After checked the existence of selection bias, we have specified a logit model to estimate the propensity score as in equation 5.8:

$$Pr(T_i = 1 \mid X_i) = \frac{exp(\alpha + \beta X_i)}{1 + exp(\alpha + \beta X_i)} \tag{5.8}$$

We performed a propensity score with the aim of showing how the multivariate test is able to check if the PS balances the data.
We specified the PS by considering all covariates involved in the selection process, without interaction terms or higher order terms. Once propensity scores were obtained, we performed a subclassification on the propensity score for finding balanced groups on which estimate the average causal effect.
We divided the estimated range of propensity score in 5 bins. Then we have tested if PS balances the data by performing the proposed multivariate test of imbalance, and we mea-

sured the causal effect in each bin. Results (Appendix B, table B.5) show that PS balances data, and in each obtained bin the true benchmark average causal effect (ATE=10.8321) is reproduced.

5.2.4 Find groups of comparable units before causal effect estimation

We propose the use of the cluster analysis on MCA coordinates, in finding groups of comparable units.
The idea is that, given a partition represented in a dendrogram (Appendix B. fig B.1), we should cut the dendrogram at the lowest level.
The data set used in this analysis is composed of 24 combinations of the X-variables (cells). By choosing a 24-clusters partition we will reproduce exactly the 24-combinations of the data design, in which all units are similar in terms of pre-treatment covariates and, as a consequence, does not exist dependence between X and T. In this circumstances, the causal effect estimation is allowed and the results are unbiased.
But, a good level of balance could be achieved also in the half of the dendrogram. We use the multivariate test of imbalance as criterion to choose how many groups are necessary to achieve balance and yield an unbiased treatment effect by group .
We performed the cluster-analysis procedures on the coordinates obtained with MCA (Appendix B, fig B.1).
The application of this analysis was carried out in Sas system that uses a hierarchical clustering method. The cluster procedure classifies units defined by the factorial coordinates obtained with the previous multiple correspondence analysis (MCA).It was first used the Ward Algorithm (B.7) and the Euclidean distance as its dissimilarity measure.
We most closely examine the two-,four-,six-,eight-,ten-,and twelve-cluster solutions to identify which one appears to meet the criteria of:

- achieving balance
- approximating as best as possible the benchmark true average treatment effect (ATE=10.8321)

Results (Tab B.7, Appendix B) show that going deeper in the cut of the dendrogram, i.e. moving from a 2-clusters partition to a 12-clusters partition, we really move from the situation of imbalance to that of balance.
Further, when balance is achieved , an unbiased estimation of the average causal effect of interest is obtained.
As in the subclassification based on the estimated propensity score, an important disadvantage of this procedure is that it discards observations in blocks where either treated or control units are absent (no common support).
Then, we rerun the analysis using the single linkage method (tab. B.8, Appendix B), the complete linkage method (tab. B.9, Appendix B) and the average linkage (tab. B.10, Appendix B).

5.3 Discussion

Using as criteria the achieved balance, the approximation of the true average causal effect and the number of discarded units, we prefer results obtained with the Ward's method.
Using the Ward's method, by moving from the 10-clusters solution to the 12-clusters solution, the number of discarded units remains invariated, meaning that the method is able to group similar units until very soon in the aggregation process.
Despite the single, complete and average linkage in the 12-clusters solution discarded less units than Ward's method, we think they are not preferable. In fact, moving from a 10-clusters solution to a 12-clusters solution, results in terms of discarded units and achieved balance are very different.
It seems (tab. B.12)that simple, complete and average linkage methods achieve balance until very late during the clustering process.
This could happen because those methods are sensitive to the nature of data.
It is well known from literature (Jobson, 1992, pp. 524-525), for example, that with the single linkage method outliers tends to remain as isolated points until very late in the hierarchical process.
The single linkage method is said to be *space conservative* (Jobson, 1992), meaning that it tends to produce long clusters in unevenly sized groups. This is in contrast, for example, with the complete linkage method which is called *space diluting*, meaning that it tends to result in compact clusters.
Both the single and complete linkage methods are sensitive to extreme observations. For example, with the single linkage an outlier between two clusters can result in the joining of the two groups. Whereas, with complete linkage, small changes in the location of some observations could have a big impact on the hierarchical cluster solution set.
For the reasons explained above, we consider the average linkage and Ward's method as more preferable to the single linkage and complete linkage methods. We conclude that depending on what one expects or believes about the nature of units being studied, a particular method might be more or less preferable than others. We suggest to perform different methods and then choose the one that meets the criteria of achieving balance until very soon in the clustering process, as the Ward's method in the simulated data.

Chapter 6

Applying the multivariate test of imbalance to real data

6.1 Introduction

This chapter is dedicated to the application of the proposed strategy to a real data set. In particular, for our intent, we replicated the analysis concerning data on subsidized and not subsidized handicraft firms of PSA programs in Tuscany region, discussed in the *IRPET*'s report (2007), with the aim of analyzing the impact of PSA programs on the performance of subsidized handicraft firms.

6.2 The Law 36/95

The legislation governing incentives for new businesses or those already in existence is growing.

More specifically, through Regional Law n° 36 of 1995, the Tuscany region has the aim of supporting small and medium handicraft enterprises. In implementation of the Regional Law *n°* 36 of 4 April 1995 *Financial intervention in favor of craftmanship and discipline of guarantee associations for craft*, art.3, para. II, the Tuscany region with the help of Artigiancredito and Fidi Toscana [1] delivers incentives to handicraft firms. The subsidies are allocated to handicraft firms on the basis of specific programs. Here we analyze the effects of two programs: the PSA 2001/2002 and the PSA 2003/2005. The main differences between the two programs concern years when incentives are delivered, the number of financed projects and the amount of the subsides allocated. Different is also the selection or auto-selection process.

Our intent here is to replicate the evaluation of the efficacy of both PSA 2001/2002 and PSA 2003/2005 programs, already performed by the joined work of Unioncamere Toscana, IRPET and the Statistical Department of Florence University. We are not interested in

[1] Fidi Toscana was set up on the initiative of the Tuscan Regional Authority and the major banks operating in the Region with the objective of facilitating access to credit for small and medium businesses featuring valid prospects of growth but without adequate guarantees

the design study, but only in the final dataset obtained with their work.
The causal question we would like to answer is if PSA programs work: particularly if the improvement in the performance of the subsidized handicraft firms could be attributed to the subsidies allocated rather than to the structural characteristics of firms or to the context in which they operate. In the rest of this paragraph we will describe the available information needed to analyze the problem: the treatment, the outcome and the covariates involved in the selection process.

6.3 Description of the data set

The final dataset arises from a complex work of data integration of different sources concerning administrative data and a field survey [2]. The final dataset consists of 266 subsidized handicraft firms and 721 not subsidized handicraft firms. Particularly, 147 are projects financed by the PSA 2001/2002 program and 119 projects financed by PSA 2003/2005 program. We perform two separated analysis: one analyzes the effect of PSA

	N	%
Subsidized firms: PSA 2001/2002	147	14
Subsidized firms: PSA 2003/2005	119	12
Not subsidized firms	721	74
Total	987	100

Table 6.1: The available sample

2001/2002 program on subsidized firms ; another analyzes the effect of PSA 2003/2005 program on subsidized firms. Then, we compare and discuss the results obtained in both analysis.
In the original study were considered 6 different potential control groups for the analysis[3].Here, we consider only the control group composed of all non subsidized handicraft firms.
We have considered the subsidized firms that have obtained at least one subside allocated by PSA 2001/2002 and/or by PSA 2003/2005 and that have realized the investment before 31 December 2005.

The covariates

As covariates involved in the selection process, causally prior to the treatment assignment T and that affect the outcome Y conditional on T, we have considered the following:

- number of employees in 2002

[2]The data integration process is described in the report *Analisi e Valautazione delle politiche di sostegno alle imprese artigiane della Toscana* (2007)
[3]The use of different control groups is discussed in the report *Analisi e Valutazione delle politiche di sostegno alle imprese artigiane della Toscana (2007)*

- legal form
- county code (geographical location)
- being or not in area objective 2
- sector
- start up date
- being or not female firm
- being or not young firm
- operate or not in local market
- operate or not in private market
- realize or not internal production
- turnover

Table C.1 shows that beneficiaries of both PSA programs are especially partnership companies. In particular, beneficiaries of PSA 2001/2002 program are mainly partnerships followed by individual and family firms; whereas, beneficiaries of PSA 2003/2005 are mainly partnerships followed by limited liability companies.

The distributions of the structural characteristics of handicraft firms show that participants of PSA 2003/2005 program have dimensions higher than participants of PSA 2001/2002 in terms of employees and turnover. The difference in the average number of employees (figure C.1) suggests that high dimensional firms are more likely to participate in PSA 2003/2005 program than PSA 2001/2002 program.
Similarly, the analysis suggests that high dimensional firms in terms of turnover (figure C.2) are more likely to participate in PSA 2003/2005 program than PSA 2001/2002. Figure C.4 shows that the 73% of beneficiaries of PSA 2001/2002 program sells their product in local markets. Similarly, the local market represents the main market for beneficiaries of PSA 2003/2005 program (57%); but, they appear more likely to sell their products also outside the local market (43%) than beneficiaries of PSA 2001/2002(27%).
The internal production is the main characteristic of all group considered; but, figure C.3 suggests that it is more typical of beneficiaries of PSA 2001/2002 program than other groups.

The presence of young or female firms is limited. Young firms represent the 15% in all group considered. Female firms represent the 14% of both not beneficiaries and beneficiaries of PSA 2001/2002 program. Their presence is a bit higher in the group of beneficiaries of PSA 2003/2005 program.

The outcome variable

In this section we analyze the performance of beneficiaries firms. More precisely, we analyze the impact of both PSA 2001/2002 program and PSA 2003/2005 program on the variation of the number of employees between 2005 and 2002 [4]. We will analyze if the number of employees in 2005 differs with respect to the number of employees in 2002 and if the detected difference, is attributable to the specific program or not.
At a descriptive level, figure 6.1 shows that beneficiaries have increased the number of employees (respectively, 27 % and 37%)more than not beneficiaries(17 %).
It clearly emerges that especially beneficiaries of PSA 2003/2005 program(37 %) engage new employees.

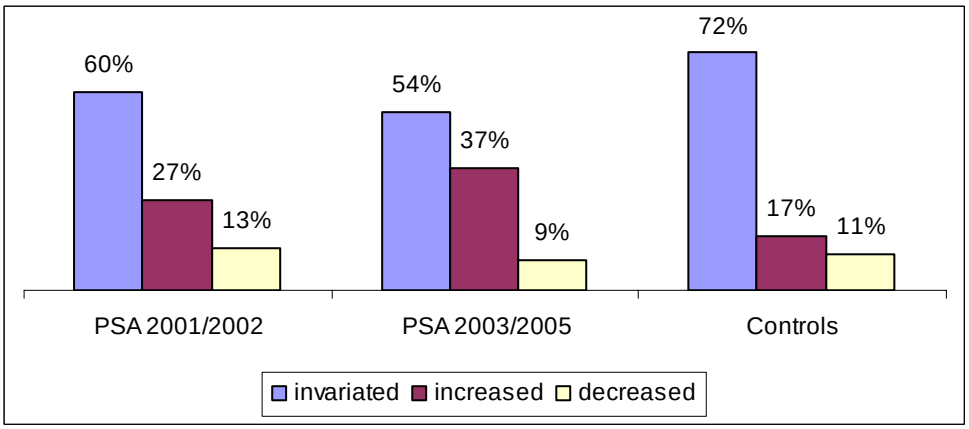

Figure 6.1: Difference of number of employees from 2002 to 2005

6.4 The impact analysis: PSA 2001/2002

In analyzing the impact of PSA 2001/2002 program on the variation of the number of employees we have first assumed to have sufficient information in the available pre-treatment covariates; then, we have assumed there was no confounding and that all variables were causally prior to the treatment assignment and that affect the outcome conditional on treatment. Finally, we have assumed that bias arises only due to difference in observed covariates.
In estimating the causal effect of PSA 2001/2002 program on beneficiaries we need to know what would have happened for the beneficiaries in the absence of the specific program. Given the evaluation problem - that is- only one of the two potential outcomes can be observed (i.e., the one corresponding to the treatment the unit received), and the consequent infeasibility of estimating the causal effect defined as the comparison $Y(1) - Y(0)$, we use not beneficiaries to approximate the counterfactual situation of beneficiaries in the absence of subsidies.

[4] In the original study the impact of both programs was evaluated on two different outcome variable: the turnover and the number of employees

Thus, the aim is to find groups of beneficiaries and not beneficiaries as similar as possible, on which estimate the causal effect of interest.
We will first check if the original data are balanced, meaning that the empirical distributions of the covariates in the groups are more similar. In doing this, we use the measure and the test of imbalance introduced in chapter 4.
Then, if imbalance is detected, we will try to balance data first by controlling for X with a model (i.e. , Propensity score); second, controlling for X by performing a cluster procedure to find local groups of balanced and comparable units.
Finally, we discuss which method appears to better improve balance in order to correctly estimate the causal effect of interest.

The assessment of selection bias

Before starting any analysis we assess the level of selection bias that arises from the non random selection mechanism.
In doing this, we have performed the multivariate test of imbalance [5]. In order to obtain a measure of selection bias, a conditional MCA was performed with participation into program indicator variable as the conditional variable and all available pre-treatment covariates introduced in the analysis (12 covariates).

$\mathbf{N_{T=1}}$	$\mathbf{N_{T=0}}$	**N**	**actual between**	**confidence interval**	α	**BEC**
147	721	868	0.0082	(0;0.0069)	0.05	1%

Table 6.2: The confidence interval for inertia between (PSA 2001/2002)

Results (table 6.2) show that the resulting inertia of the conditioning by T (participation indicator variable in PSA 2001/2002 program) remains outside the interval, showing the dependence of the X-variables from T. The significance of the conditioning means that data are not balanced between treated and controls.
The actual between represents the measure of absolute and global imbalance in the distributions of treated and control units in the original data with $n_{T=1} = 147$ and $n_{T=0} = 721$. Conversely, the variable-by-variable chi-square test does not represent an objective instrument to conclude if data are really imbalanced or not. In fact, results (table 6.3)show that three of twelve variables considered are imbalanced; but there is no objective way to know if the detected imbalance in that variables is dangerous or not, in order to correctly estimate an unbiased treatment effect.

Controlling for X with a model

After checked the presence of selection bias, we control for X with a model: the Propensity Score.

[5] All continuous pre-treatment covariates (number of employees in 2002, start up date, turnover) were previously discretized

Covariates	chi-square	p-value	Balance
employees in 2002	1.4552	0.8345	yes
start up date	5.4839	0.2412	yes
turnover	30.8210	0.0003	no
legal form	4.3689	0.3584	yes
county code	9.1862	0.4203	yes
realize or not internal production	13.3037	0.0003	no
operate or not in private market	6.0460	0.0139	no
operate or not in local market	0.4452	0.5046	yes
being or not young firm	0.0022	0.9626	yes
being or not female firm	0.0016	0.9678	yes
being or not in area ob2	2.3874	0.3031	yes
sector	5.0332	0.5396	yes

Table 6.3: The variable-by-variable chi-square test

To estimate the propensity score we have specified a logit model as in equation 5.8.
We have specified the model by considering all available covariates involved in the selection process, without introducing interaction terms or higher order terms.
Once obtained the estimated propensity scores, we have performed a subclassification on the propensity score for finding balanced groups on which estimate the average causal effect.
We divided the estimated range of propensity score in 5 bins [6]. Then, we tested if PS balances the data by performing the proposed multivariate test of imbalance, and we measured the average causal effect (ATE) in each balanced bin.

BIN	n	$n_{T=1}$	$n_{T=0}$	I_b	Interval for I_b	balance	ATE	err std
1	173	59	114	0.0183	(0;0.031)	yes	0.046	0.2544
2	174	31	143	0.0154	(0;0.032)	yes	0.608	0.3232
3	174	24	150	0.0136	(0;0.032)	yes	-0.039	0.5609
4	174	25	149	0.0183	(0;0.032)	yes	0.3434	1.1329
5	173	8	165	0.0227	(0;0.031)	yes	**-1.726**	0.8334

Table 6.4: Results of stratification on estimated propensity score

Table 6.4 shows that only in the bin number 5 the effect is statistically significant. The effect has minus sign showing a negative impact of the PSA 2001/2002 program on the variation of the number of employees.
Then, to obtain an overall estimate of the average causal effect, we have computed the Average Treatment Effect on the Treated (ATT) according to the following formula[7]:

[6] Based on Cochran's results (1968) we may expect a 90% bias reduction for each of the twelve variables when we subclassify at the quintile of the distribution of the population propensity score
[7] S.O.Becker and A.Ichino,2002

$$\tau^S = \sum_{q=1}^{Q} \tau_q^S \frac{\sum_{i \in I(q)} T_i}{\sum_{\forall i} T_i} \tag{6.1}$$

with Q as the number of bins, I(q) as the set of units in bin q, τ_q^S as the treatment effect in bin q, with the weight for each bin given by the corresponding fraction of treated units.

Then, we have computed the average causal effect (ATE) according to the following formula:

$$\tau^{S(p)} = \sum_{q=1}^{Q} \tau_q^S \frac{n_q}{N} \tag{6.2}$$

where n_q is the number of units (both treated and untreated) in bin q and N is the number of units in the sample considered. Table D.1 shows that both the ATT and the ATE have not a big impact on the group of firms considered; furthermore, the estimated effects are not statistically significant.

Estimate local causal effects via a non parametric method

Here we propose the use of the cluster analysis on MCA coordinates, with the aim of finding groups of comparable units on which estimate local causal effects.
The method non-parametrically controls for some or all of the pre-treatment control variables involved in the selection process.
Aiming at finding balanced groups, we will check balance via the multivariate test of imbalance proposed in chapter 4.
The main advantage of performing a cluster analysis on MCA coordinates is that it avoids both the problems of model dependence and of dimensionality of categorical variables,being the new MCA coordinates less in number and of continuous nature. The application of this analysis was carried out in Sas.
In the light of discussion presented in chapter 5, we have preferred to use a hierarchical clustering method, and the Ward's method as group proximity measure. Given the hierarchical clustering process represented in a dendrogram [8], that is a tree diagram used to keep track of the sequential clustering process, we have chosen a 18-clusters solution set. The basic idea is that going deeper in the cut of the dendrogram, is more plausible that groups are balanced in terms of pre-treatment covariates, and as shown in table 6.5, the multivariate test of imbalance gives rise to our idea. Once the specified cluster solution set is selected, the analysis estimates program impacts by cluster where balance is achieved and common support satisfied. The units belonging to non balanced groups were discarded.

[8] We omitted to insert the dendrogram here because the sample size was too large

We have chosen the 18-clusters solution set because it discards a small number of units with respect to other solutions; it clearly emerges (tab. B.11) there was a big jump in terms of discarded units moving from the 16-clusters solution to the 18-clusters solution. Going more deeper than 18-clusters in the cut of the dendrogram, the 20-,22-,24-,26- clusters solution are invariated in terms of discarded units with respect to the 18-clusters solution. Furthermore, by choosing more than 26 clusters, common support starts to be not satisfied in some clusters, giving as result an higher number of discarded units than the previous cluster solution.

It clearly emerges that, on one hand, if the clusters are too many, then too many observations may be discarded due to the lack of common support.On the other hand, if the number of clusters is chosen too small, then too many observations may be discarded due to the lack of balance.

Thus, we think that the 18-clusters solution solves the trade-off between the two problems. Then,despite our primary interest remains the estimation of local effects by group, to examine how much close or not are the results with respect to PS, we have computed the ATT according to formula 6.1 where bins are replaced by groups. Similarly, we have computed the ATE according to the formula 6.2. They are both positive but not statistically significant. It means that the effect of PSA 2001/2002 program had not a considerable impact on the variation of the number of employees.

Group	**n**	$\mathbf{n_{T=1}}$	$\mathbf{n_{T=0}}$	$\mathbf{I_b}$	**Interval for $\mathbf{I_b}$**	**balance**	**local effect**	**err std**
(1)	326	43	283	0.012	(0;0.0145)	**yes**	0.1227	0.3856
(2)	191	44	147	0.0118	(0;0.024)	**yes**	0.3912	0.2512
(3)	40	5	35	0.0842	(0;0.098)	**yes**	-0.2	0.2588
(4)	33	10	23	0.0291	(0;0.1135)	yes	0.9636	0.7352
(5)	20	4	16	0.1128	(0;0.2224)	yes	0.3125	1.1444
(6)	24	2	22	0.1874	(0;0.1812)	**no**	-	-
(7)	13	6	7	0.1805	(0;0.3038)	**yes**	**0.025**	0.0034
(8)	6	1	5	0.2533	(0;0.5056)	**yes**	1.2	1.2
(9)	6	2	4	0.215	(0;0.4884)	**yes**	**3**	0.07
(10)	14	2	12	0.1699	(0;0.2892)	**yes**	-4.333	13.956
(11)	29	5	24	0.1139	(0;0.143)	**yes**	0.8417	1.1023
(12)	41	9	32	0.0435	(0;0.1036)	**yes**	-0.014	0.1266
(13)	37	7	30	0.0645	(0;0.1094)	**yes**	-0.933	1.3384
(14)	34	5	29	0.0573	(0;0.1132)	**yes**	-0.091	1.9552
(15)	3	0	3	no common support				
(16)	11	3	8	0.2503	(0;0.2945)	**yes**	-2.25	1.4884
(17)	4	1	3	0.5648	(0;0.7843)	**yes**	-0.333	0.6667
(18)	36	8	28	0.1209	(0;0.1180)	**no**	-	-

Table 6.5: 18-Clusters solution set:PSA 2001/2002

n-clusters solution	discarded units	discarded units %
14	620	71 %
16	417	48 %
18	63	7%
20	63	7%
22	63	7%
24	63	7%
26	63	7%
28	82	9%
30	82	9%

Table 6.6: Discarded Units

6.5 The impact analysis: PSA 2003/2005

Analogous considerations hold for the impact analysis of PSA 2003/2005 program.We aim at analyzing the impact of PSA 2003/2005 program on the variation of the number of employees between 2005 and 2002.
The pre-treatment covariates are the same as those considered for the impact analysis of PSA 2001/2002.
We will first check if data are balanced, then we aim at finding local groups of balanced and comparable units on which estimate local causal effects.

The assessment of selection bias

Results (Tab. 6.7) show that the resulting inertia of the conditioning by T (participation indicator variable in PSA 2003/2005 program) remains outside the interval, showing the dependence of the X-variables from T.
The significance of the conditioning means that data are not balanced between treated and controls.

$N_{T=1}$	$N_{T=0}$	N	actual between	confidence interval	α	BEC
119	721	840	0.0159	(0; 0.007)	0.05	1%

Table 6.7: balance (psa 2003/2005)

Conversely, the chi-square variable-by-variable summary (Tab. 6.8) shows that six of twelve considered variables are imbalanced without giving a global measure of imbalance.

Controlling for X with a model

To estimate the propensity score we have specified a logit model as in equation 5.8. We have specified the model by considering all covariates involved in the selection process, without introducing interaction terms or higher order terms.
Once obtained the estimated PS we performed a subclassification on the PS for finding

Covariates	Chi-square	p-value	balance
section	19.6712	0.0032	no
being or not in area ob2	2.5490	0.2796	yes
being or not female firm	0.8183	0.3657	yes
being or not young firm	0.0500	0.8230	yes
operate or not in local market	8.2807	0.0040	yes
operate or not in private market	4.6086	0.0318	no
realize or not internal production	0.7659	0.3815	yes
county code	17.1003	0.0472	no
legal form	20.3648	0.0001	no
turnover	58.4026	¡.0001	no
birth date of firms	2.2852	0.6835	yes
employees in 2001	24.8567	¡.0001	no

Table 6.8: The variable by variable chi square test (PSA 2003/2005)

balanced groups on which estimate the average causal effect. We divided the estimated range of PS in 5 bins.
Then, we tested if PS balances the data by performing the proposed multivariate test of imbalance, and we measured the average causal effect (ATE) in each balanced bin.
Table 6.9 shows that in all bins the estimated effects are not statistically significant.

BIN	n	$n_{T=1}$	$n_{T=0}$	I_b	Interval for I_b	balance	ATE	err std
1	167	2	165	0.0187	(0;0.0331)	yes	-0.097	1.1055
2	167	12	155	0.0131	(0;0.0319)	yes	-0.13	0.5752
3	167	18	149	0.0162	(0;0.0331)	yes	0.104	0.4838
4	167	22	145	0.011	(0;0.0325)	yes	0.8342	0.4999
5	167	64	103	0.0172	(0;0.0336)	yes	2.1062	1.1314

Table 6.9: Results of stratification on estimated propensity score

Then, we have computed the overall Average Treatment Effect on the Treated (ATT) according to formula 6.1 and the ATE according to formula 6.2. Results in table D.1 show that they are both statistically significant. More precisely, table D.1 shows that according to the subclassification on the estimated PS, beneficiaries of PSA 2003/2005 program engage 1,29 employees more than not beneficiaries.

Estimate local causal effects via a non parametric model

Here we perform a cluster analysis on the pre-treatment covariates in order to find local groups of comparable units.
The application of this analysis was carried out in Sas and uses an hierarchical clustering method with the Ward's method as aggregation criterion. We most closely examine the 4-,5-,6-,8-,10-,12-,14-,16,18-,20-,22-,28- clusters solutions (6.10).

Table 6.10 shows that since the five-clusters solution data are balanced, with zero discarded units. Even if data are balanced since the five-clusters solution, we have chosen

n-clusters solution	discarded units	discarded units (%)
4	430	51 %
5	0	0%
6	0	0%
7	0	0%
8	0	0%
10	0	0%
12	0	0%
14	0	0%
16	0	0%
18	0	0%
20	0	0%
22	0	0%
⋮	⋮	⋮
28	44	5.2%

Table 6.10: Discarded units:PSA 2003/2005

the 14-clusters solution in order to obtain clusters more homogeneous (tab. 6.11) in terms of inertia. Table 6.11 shows that the inter-inertia increases with the number of groups. By choosing the 14-clusters solution the ratio between inter inertia and total inertia achieves an acceptable level, giving as result clusters more homogeneous than the previous cluster solution. Then we have computed the ATT and ATE (tab. D.1).

n-Clusters	**Total inertia**	**Inter inertia**	**Inter/Total**
5	1.3189	0.3716	0.2818
6	1.3189	0.4179	0.3168
7	1.3189	0.4587	0.3477
8	1.3189	0.4941	0.3746
9	1.3189	0.5287	0.4008
10	1.3189	0.5560	0.4215
11	1.3189	0.5816	0.4409
12	1.3189	0.6029	0.4571
13	1.3189	0.6238	0.4729
14	1.3189	0.6421	0.4868
15	1.3189	0.6564	0.4977
16	1.3189	0.6706	0.508
17	1.3189	0.684	0.510
18	1.3189	0.696	0.528
19	1.3189	0.709	0.537
⋮	⋮	⋮	⋮
28	1.3189	0.8334	0.6318

Table 6.11: inter inertia in n-clusters solutions

Table 6.12 shows that in three groups of the fourteen considered the treatment effects are significant.

Table 6.13 notes the characteristics associated with membership in each of the three

Group	n	$n_{T=1}$	$n_{T=0}$	I_b	Interval for I_b	balance	local effect	err std
(1)	65	16	49	0.0608	(0;0.080)	yes	2.023	0.5481
(2)	72	3	69	0.0276	(0;0.071)	yes	-0.188	0.5751
(3)	70	4	66	0.0419	(0;0.0691)	yes	-0.061	0.5646
(4)	20	6	14	0.1494	(0;0.2273)	yes	-5.071	7.401
(5)	44	4	40	0.0473	(0;0.11)	yes	-0.15	0.4045
(6)	69	5	64	0.064	(0;0.070)	yes	0.0094	1.086
(7)	38	13	25	0.0616	(0;0.1196)	yes	4.4215	2.5615
(8)	41	14	27	0.0641	(0;0.1036)	yes	0.7884	1.3597
(9)	24	3	21	0.059	(0;0.1812)	yes	-1.095	0.7022
(10)	92	17	75	0.0238	(0;0.0505)	yes	1.8071	0.5775
(11)	82	10	72	0.0387	(0;0.0387)	yes	-0.436	0.872
(12)	77	8	69	0.0455	(0;0.0616)	yes	4.212	1.2912
(13)	48	8	40	0.04143	(0;0.0947)	yes	1.5	1.0717
(14)	93	7	86	0.0377	(0;0.0531)	yes	0.9452	0.6416

Table 6.12: 14-Clusters solution set: PSA 2003/2005

clusters in which the estimated treatment effect is significant. To interpret this information, those characteristics represented in table 6.13 are the describing features of members in that clusters, for whom the absolute frequency in the cluster is above the absolute frequency in the overall sample [9].

Covariates	Cluster 1	Cluster 10	Cluster 12
employees in 2002		8-12	8-12
legal form	individual firms limited liability company	individual firms	partnership
County code	Massacarrara,Pistoia	Arezzo	
area objective 2	Ob2/Pashing Out		
sector		manifacturing	manifacturing
start up date		1995-1999	1979-1985
female firms			
young firm			no
local market		no	yes
private market		no	no
internal production			no
turnover		500000-1000000	25000-500000
n	65	92	77
effect	2.023	1.8071	4.212

Table 6.13: Clusters description

To focus more specifically on the meaningful story that these results suggest, consider the cluster 12. As shown in table 6.13 Cluster 12 is composed of a greater proportion of handicraft partnership company, who operate in the manifacturing sector, with a relevant

[9] With n_{jk} as the number of units with category j in the cluster k, with n_k as the number of units in cluster k, with n_j the number of units with category j in the overall sample of n units, then the difference between $\frac{n_{jk}}{n}$ and $\frac{n_j}{n}$ represents a measure of the importance of the category j in the cluster k

number of years of experience in the market (start up date between 1979 and 1985), and of those more likely to externalize production.
In addition, all firms in cluster 12 were not firm with a young management or young employees, and their turnover was not of considerable entity.
In brief, the PSA 2003/2005 program seems to have a big impact (effect=4.212) on units that come together in cluster 12. Units within cluster 12 are generally of medium dimensions in terms of employees and turnover but with many experience in terms of start up date and in terms of age of employees.
If we look at table D.1, both the estimated ATT and ATE are statistically significant. Against the PSA 2001/2002 program, the PSA 2003/2005 program have increased the number of employees for those beneficiaries. More precisely, table D.1 shows that, according to the ATT based on the clustering procedure, beneficiaries engage 1.22 employees more than not beneficiaries.

6.6 Discussion

Here, we have dealt with the problem of selection bias in a real and complex problem. Results in table D.1 confirm what observed in descriptive analysis (Fig 6.1). It emerges that PSA 2003/2005 program works better than PSA 2001/2002 program. More precisely, beneficiaries of PSA 2003/2005 program have increased the average number of employees 1.29 (according to subclassification on PS) and 1.22 (according to the clustering procedure) more than non beneficiaries.
We would like to highlight that the innovative aspect of the proposed analysis is represented by the ability of identifying subgroups within data and capitalizing on their heterogeneity. Our proposed strategy has the strength of allowing to answer the question *For what kinds of handicraft firms does PSA programs work?*
In turn, the analysis has the strength of eliminating model dependence, given that, against PS, the strategy completely controls for X in a non-parametric way.
In addition, if the considered X-variables are those involved in the selection process, the proposed strategy eliminates plausible alternative cause for why the program achieved a specific impact. The elimination of other plausible causes is allowed by the fact that groups are really balanced and the achieved balance is globally checked via the proposed multivariate test.
Furthermore, by using our proposed strategy rather than subclassification on PS, we have found groups on which the effect of PSA program was statistically significant.
We think that the resulting analytic advantage of performing a cluster analysis, when the aim is to analyze program impacts on local spaces, is represented by the ability of identifying which units fall into which specific groups well-defined in terms of baseline pre-treatment characteristics.
The same advantage arises by performing a subclassification on the estimated PS; but,

against the subclassification on the PS, our proposed strategy is not model dependent.

Conclusions and perspectives

The main goal in this Ph.D thesis has been to introduce a strategy for making causal inferences from observational data without model dependence.
As part of that strategy,we have proposed a new data driven procedure able to check and test the presence of selection bias by preserving the multivariate nature of data.
The procedure is also able to choose automatically the correct number of clusters on which estimate local causal effects in an unbiased way.
Our proposal originates from the intention to discover local groups of comparable units according to a test of balance that overcomes limits of other procedures not always able to check balance in a multivariate way.
We gave a measure of global imbalance (BEC) and we test it in a way not accomplished by the variable-by-variable t-test or chi-square test commonly used in applied research.
We have tested the performance of the proposed strategy on simulated data, which has shown that when the test detects the balance, then the true average causal effect is reproduced.
We think the proposed strategy outperforms other ways of drawing causal conclusions for the following reasons:

1. it uses all available information of the X matrix without problem of dimensionality.
2. the procedure could be useful for subgroup analysis by overcoming limits of the conventional way to measure program impacts- i.e. compute the overall average treatment effect - that may obscure impacts that accrue to subgroups (Peck, 2005). In this sense, it represents an effort to detect treatment group heterogeneity.
3. it non parametrically controls for X, with less resulting model dependence. In particular, it is not needed to specify *a priori* any model; but, it lets the data to speak
4. it is able to find clusters of comparable units according to the dependence structure of data
5. it allows to automatically check global imbalance
6. it can account for the dependence relationship of any number of covariates

The idea of measuring impacts on local spaces represents an initial stage in learning more from data, with the ultimate intent of estimating an unit level effect rather than the

average causal effect.

Future works in this area might concern other classification methods, being the cluster analysis sensitive to the nature of data, the method and the dissimilarity measure adopted. Further, future works might explore analytic properties of the conditional space in order to understand if the coordinates of the conditional space could be used in reconstruct the missing counterfactual at a unit level.

We will also write a Sas program able to perform the overall analysis in order to develop an automatic node of a DM process that automatically checks and tests balance of a given data set.

Finally, we will examine the sensitivity of the multivariate test of imbalance to specific failures of the unconfoundedness assumption.

Bibliography

[1] Roberto Agodini and Mark Dynarski (2004), *Are the experiments the only option? A look at Dropout Prevention Programs*, The Review of Economics and Statistics, MIT PRESS, vol.86(1), pp.180-194, 09.

[2] Onur Baser (2006), *Too Much Ado about Propensity Score Models? Comparing Methods of Propensity Score Matching*, Value in health, Vol.9, N°6.

[3] Sascha D.Becker and Andrea Ichino (2002), *Estimation of average treatment effects based on propensity scores*, 2(4), 358-377.

[4] Richard A.Berk (2004), *Regression Analysis*. Sage Publications.

[5] R.D. Bingham and C.L.Felbinger(2002), *Evaluation in practice: a methodological approach*, Chatham House.

[6] Richard Blundell and Monica Costa Dias (2002), *Alternative Approaches to Evaluation in Empirical Microeconomics*, working paper

[7] H.Bozdogan(2004), *Statistical Data Mining and Knowledge Discovery*, Chapman and Hall/CRC.

[8] Cochran, W.G. (1968), *The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies*, Biometrics, 24, 205-213.

[9] D. Conniffe, V. Gash and P.J. O'Connell (2000), *Evaluating State Programmes: Natural Experiments and Propensity Scores*, The Economic and Social Review, Vol.31, N° 4, pp. 283-308.

[10] Jean-Jacque Daudin (1981), *Analyse factorielle des dépendances partielles*, Revue de statistique appliquée, tome 29, n°2, pp. 15-29.

[11] Rajeev H. Dehejia and Sadek Wahba (2002), *Propensity score matching methods for non-experimental causal studies*, Review of Economics and Statistics, Vol.84,n° 1, pp.151-161

[12] A.Diamond and J.S.Sekhon (2006), *Geneting Matching for estimating Causal Effects: A general multivariate MAtching Method for Achieving Balance in Observational studies*, available at http://repositories.org/igs/WP2006-35

[13] B.Escofier(1984), *Analyse factorielle en référence à un modèle. Application à l'analyse de tableaux d'échanges*, INRIA, rapport de Recherche N° 337.

[14] B.Escofier (1988), *Analyse des correspondances multiples conditionelle*, La Revue de Modulad

[15] Josep Dunis i Estadella (2004), *Estudi de les inèrcies estructurals en anàlisis de correspondències.Aportacions per a una millora de les anàlisis*, PhD thesis, available online at http://www.tesisenxarxa.net/TDX-0411105-121820/index-an.html

[16] J.D.Estadella, T. Aluja-Banet and S. Thiò-Henestrosa (2005), *Distribution of the inter and intra inertia in conditional MCA*, Computational Statistics 20:449-463.

[17] C.E.Frangakis and D.B.Rubin(2002), *Principal stratification in Causal Inference*, Biometrics, 58, pp.21-29.

[18] Micheal J.Greenacre (1984), *Theory and Applications of Correspondence Analysis*

[19] D.Hand, H.Mannila and P.Smyth (2001) *Principles of Data Mining.*

[20] James J.Heckman (1979), *Sample selection bias as a specification error*, Econometrica, Vol.47, N°1.

[21] James J. Heckman(1989), *Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training*, Working paper N° 2861.

[22] James J.Heckman, Hidehiko Ichimura and Petra E.Todd(1997b), *Matching as an econometric evaluation estimator: Evidence from evaluating a job training Programme*, The Review of Economic Studies, Vol. 64, N°4, pp 605-654.

[23] James J. Heckman, Robert J. Lalonde, Jeffrey A. smith (1999), *The economics and econometrics of active labore market programs*, Handbook of Labor Economics, in: O. Ashenfelter & D.Card (ed.), 1, vol.3, ch. 31, pp. 1865-2097.

[24] Gary T. Henry and James H. McMillan (1993), *Performance Data: Three comparison Methods*, Evaluation Review; 17; pp. 643-652.

[25] D.E. Ho, K.Imai, G.King and E.A.Stuart(2007), *Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference*, Political Analysis, 15:199-236.

[26] P.W.Holland(1986), *Statistics and Causal Inference*, Journal of the American Statistical Association, Vol.81, n° 396, pp. 945-960.

[27] P.W.Holland and D.B.Rubin (1988), *Causal Inference in Retrospective Studies*, Evaluation Review, 12;203

[28] S.M.Iacus, G.King and G.Porrp (2008), *Matching for causal inference without balance checking*, working paper, available at http://gking.harvard.edu/files/abs/cem-abs.shtml

[29] A.Ichino, F.Mealli, T.Nannichini (2005), *Sensitivity of matching estimators to unconfoundedness. An application to the effect of temporary work on future employment*, available at:www.econ-pd.unisi.it/paperseminari/ichino.pdf

[30] K.Imai, G. King and E.A.Stuart (2006), *The balance test fallacy in matching methods for causal inference*, working paper.

[31] Guido W. Imbens (2002), *Sensitivity to exogeneity assumptions in program evaluation*, The american economic review, Vol.93, N° 2.

[32] IRPET in collaborazione con il dipartimento di statistica G.Parenti dell'Università di Firenze (2007), *Analisi e Valutazione delle Politiche di sostegno alle imprese artigiane della Toscana*, available at http://www.irpet.it/

[33] J.D.Jobson (1992), *Applied Multivariate Data Analysis*, Springer-Verlag

[34] S.Lee (2006), *Propensity score adjustment as a weighting scheme for volunteer panel web surveys.*

[35] , Jason K.Luellen, William R. Shadish and M.H. Clark (2005), *Propensity scores: an introduction and experimental test*, Evaluation Review; 29; 530.

[36] S.Mithas, D. Akmirall and M.S. Krishnan (2006) *Do CRM Systems Cause One-to-One Marketing Effectiveness?*, Statistical Science, Vol.21, n° 2, pp. 223-233.

[37] M.Mizuno and T.Hoshino (2006), *Assessing the short-term causal effect of TV advertising via the Propensity Score Method*, University of Tsukuba, Japan.

[38] S.L.Morgan and D.J.Harding (2006), *Matching Estimators of causal effects:Prospects and Pitfalls in theory and practice*, Sociological Methods Research, 35;3

[39] Laura R.Peck(2003), *Subgroup Analysis in social experiments: Measuring Program Impacts Based on Post-treatmnet choice*, American Journal of evaluation, 24(2).

[40] Laura R.Peck and Ronald J. Scott Jr.(2005), *Can Welfare Case Management Increase Employment? Evidence from a Pilot Program Evaluation*, The policy studies Journal, Vol.33, n°4.

[41] Laura R.Peck (2005), Using cluster analysis in program evaluation, Evaluation Review; 29; 178.

[42] Laura R.Peck (2007), *What are the effects of Welfare Sanction Policies?*, American Journal of Evaluation, Vol.28, n°3, pp. 256-274.

[43] P.R.Rosenbaum and D.B.Rubin (1983a), *The central Role of the Propensity Score in Observational Studies for Causal Effects*, Biometrika, Vol.70, n°1, pp. 41-55.

[44] P.R.Rosenbaum and D.B.Rubin (1983b), *Assessing sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome*, Journal of the Royal Statistical Society, Vol. 45, n°2, pp. 212-218.

[45] P.R. Rosenbaum and D.B. Rubin(1984), *Reducing Bias in Observational Studies Using Subclassification on the Propensity Score*, Journal of American Statistical Association, Vol.79, n°387, pp.516-524.

[46] P.R. Rosenbaum and D.B. Rubin(1985), *The bias due to incomplete matching*, Biometrics, Vol. 41, N° 1, pp. 103-116.

[47] D.B.Rubin (1987), *Multiple imputation for nonresponse in surveys*, New York:John Wiley.

[48] D.B.Rubin (1991), *Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism*, Biometrics, Vol.47, n°4, pp.1213-1234.

[49] D.B.Rubin (2001), *Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation*, Health Services and Outcomes Research Methodology 2, pp. 169-188.

[50] D.B.Rubin (2004), *Teaching Statistical Inference for Causal Effects in Experiments and Observational Studies*, Journal of Educational and Behavioral Statistics, Vol.29, n°3, pp. 343-367.

[51] D.B.Rubin (2005), *Causal Inference Using Potential Outcomes: Design, Modeling, Decisions*,Journal of the American Statistical Association, Vol.100, n°469.

[52] D.B.Rubin (2006), *The Design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials*, Statistics in medicine, 26: pp. 20-36.

[53] D.B. Rubin and R.P. Waterman (2006), *Estimating the Causal effects of marketing interventions using propensity score methodology.* Statistical Science, Vol.21, n° 2, pp. 206-222

[54] M.Schonlau, A.V. Soest, A. Kaptevn and M.P. Couper (2006), *Selection Bias in Web Surveys and the Use of Propensity Scores*, working paper.

[55] J.S.Sekhon and W.R.Mebane (2000), *Geneting Optimization Using Derivatives*, work in paper.

[56] J.S.Sekhon, *Alternative Balance Metrics for Bias reduction in Matching Methods for causal inference*, work in paper.

[57] J.S.Sekhon, *Multivariate and propensity Score matching Software with Automated Balance Optimization: The Matching package for R*, Journal of statistical Software,available at http://www.jstatsoft.org/

[58] J.S.Sekhon and R.Grieve (2008), *A new non-parametric matching method for bias adjustment with applications to economic evaluations*, work in paper.

[59] W.R.Shadish, Thomas D. Cook and Donald T. Campbell (2002), *Experimental and quasi-experimental designs for generalized causal inference*, Houghton Mifflin Company.

[60] W.R.Shadish, J.K. Luellen and M.H. Clark (2006), *Propensity Scores and quasi-experiments: A testimony to the practical side of lee sechrest*, in *Strenghtening Research Methodology: Psychological Measurement and Evaluation*, Edited by R.R. Bootzin and P.E. McKnight, part II, ch. 8,pp.143-157.

[61] W.R. Shadish, M.H. Clark (2008), *Can non randomized experiments yield accurate answers? A randomized experiment comparing Random to non Random Assignment*, working paper.

[62] P. M. Steiner, T.D.Cook,W.R.Shadish and M.H. Clark (2008), *the importance of covariate selection in controlling for selection bias in observationsl studies*, in press.

[63] M.Tripathi (2007), *Channel Spillovers from Offline Entry*, Kellogg School of Management. Northwestern University.

Appendix A

The concept of partial dependence

Definitions of J.N.Darroch

Let X, Y, T three discrete random variables and let:

$$\begin{aligned} P_{ijt} &= P(X=i, Y=j, T=t) \text{ with } i=1,...,I_n; j=1,...,J; t=1,...,T \\ P_{ij.} &= \sum_t P_{ijt} \\ P_{.jt} &= \sum_i P_{ijt} \\ P_{i.k} &= \sum_j P_{ijt} \\ P_{..t} &= \sum_{ij} P_{ijt} \end{aligned} \tag{A.1}$$

Then X and Y are conditionally independent given T if for each ijt:

$$P_{ijt} = \frac{P_{i.t}P_{.jt}}{P_{..t}} \tag{A.2}$$

Darroch measures the conditionally dependence of the events $(X = i)$ and $(Y = j)$ given $(T = t)$ as:

$$\frac{P_{ijt}}{P_{..t}} - (\frac{P_{i.t}}{P_{..t}})(\frac{P_{.jt}}{P_{..t}}) \tag{A.3}$$

and the average conditional dependence as:

$$\sum_t (\frac{P_{ijt}}{P_{..t}} - \frac{P_{i.t}}{P_{..t}})P_{..t} = P_{ij.} - \pi_{ij} \tag{A.4}$$

where

$$\pi_{ij} = \sum_t \frac{P_{i.t} P_{.jt}}{P_{..t}} \tag{A.5}$$

He measures the marginal dependence between two events $(X = i)$ and (Y=j) as:

$$P_{ij.} - P_{i..} P_{.j.} \tag{A.6}$$

Finally he measures the dependence due to T as:

$$\pi_{ij} - P_{i..} P_{.j.} \tag{A.7}$$

where π_{ij} could be interpreted as the conjoint probability of the events $(X = i)$ and $(Y = j)$ if these two events are conditionally independent given T. If there is not dependence due to T between the event (X=i) and (Y=j) then:

$$\pi_{ij} = P_{i..} P_{.j.} \tag{A.8}$$

According to Darroch the marginal dependence between two variables X and Y can be decomposed as follows:

$$\underbrace{P_{ij.} - P_{i..} P_{.j.}}_{\text{marginal dependance}} = \underbrace{(P_{ij.} - \Pi_{ij})}_{\text{dependence not due to T}} + \underbrace{(\Pi_{ij} - P_{i..} P_{.j.})}_{\text{dependence due to T}} \tag{A.9}$$

To the marginal probabilities $P_{ij.}$, $P_{i.t}$, $P_{.jt}$ correspond three different tables:

1. the table with generic term P_{ij}
2. the table π_{ij}
3. the table $P_{i..} P_{.j.} + (P_{ij.} - \pi_{ij})$

All the three tables have the same marginal distributions $P_{i..}$ and $P_{.j.}$. At the same way could be constructed the contingency tables corresponding to the table of probabilities 1,2 and 3. In particular, to study the marginal dependence (X,Y) between X and Y he performs a factorial analysis of the table 1; to study the dependence due to T he performs a factorial analysis of the table 2 and for the analysis of the dependence not due to T he performs a factorial analysis of the table 3.

Appendix B

Simulation

	$\mathbf{T}$	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$
$\mathbf{X_1}$	540.15 $< .0001$		1.2780 0.2583	39.2858 $< .0001$	4.2089 0.0402
$\mathbf{X_2}$	1.1086 0.2924				
$\mathbf{X_3}$	37.8922 $< .0001$		2.1787 0.3364		10.6046 0.0050
$\mathbf{X_4}$	8.9560 0.0028		6.4091 0.0114		

Table B.1: The dependence structure by design

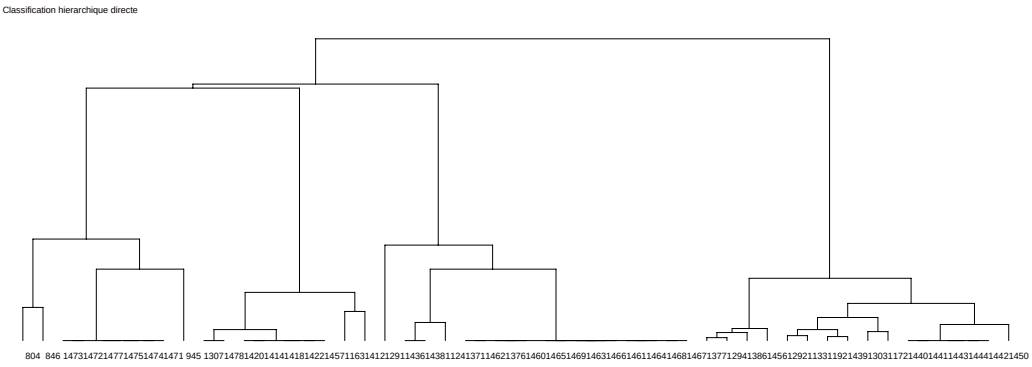

Figure B.1: dendrogram MCA

Combinations	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	**T=1**	**T=2**	$\mathbf{\Pi_{(T=1)}}$	$\mathbf{\Pi_{(T=2)}}$	**N**
1	1	1	1	1	40	1	97.5%	2.5%	41
2	1	1	1	2	5	0	100%	0 %	5
3	1	1	2	1	40	2	95.2%	4.8%	42
4	1	1	2	2	3	0	100%	0%	3
5	1	2	1	1	40	3	93.02%	6.98%	43
6	1	2	1	2	9	1	98.9%	1.1%	10
7	1	2	2	1	30	1	9.67%	90.33%	31
8	1	2	2	2	2	0	100%	0%	2
9	1	1	3	1	40	7	85.1%	14.9%	47
10	1	1	3	2	2	2	50%	50%	4
11	1	2	3	1	10	5	97.5%	2.5%	15
12	1	2	3	2	5	0	100%	0 %	5
13	2	1	1	1	7	93	7%	93%	100
14	2	1	1	2	0	10	0%	100%	10
15	2	2	1	1	4	80	4.76%	95.24%	84
16	2	2	1	2	0	10	0%	100%	10
17	2	1	2	1	3	30	9%	91%	33
18	2	1	2	2	0	10	0%	100%	10
19	2	2	2	1	1	8	11%	89%	9
20	2	2	2	2	0	20	0%	100%	20
21	2	1	3	1	7	93	7%	93%	100
22	2	1	3	2	0	20	0%	100%	20
23	2	2	3	1	10	90	10%	90%	100
24	2	2	3	2	0	20	0%	100%	20

Table B.2: The Data Design

	$\mathbf{X_{11}}$	$\mathbf{X_{12}}$	$\mathbf{X_{21}}$	$\mathbf{X_{22}}$	$\mathbf{X_{31}}$	$\mathbf{X_{32}}$	$\mathbf{X_{33}}$	$\mathbf{X_{41}}$	$\mathbf{X_{42}}$
$\mathbf{X_{11}}$	248								
$\mathbf{X_{12}}$	0	516							
$\mathbf{X_{21}}$	142	273	415						
$\mathbf{X_{22}}$	106	243	0	349					
$\mathbf{X_{31}}$	99	204	156	147	303				
$\mathbf{X_{32}}$	78	72	88	62	0	150			
$\mathbf{X_{33}}$	71	240	171	140	0	0	311		
$\mathbf{X_{41}}$	219	426	363	282	268	115	262	645	
$\mathbf{X_{42}}$	29	90	52	67	35	35	49	0	119

Table B.3: The Burt table of the conditional space

	T=1	**T=0**
$\mathbf{X_{11}}$	226	22
$\mathbf{X_{12}}$	32	484
$\mathbf{X_{21}}$	147	268
$\mathbf{X_{22}}$	111	238
$\mathbf{X_{31}}$	105	198
$\mathbf{X_{32}}$	79	71
$\mathbf{X_{33}}$	74	237
$\mathbf{X_{41}}$	232	413
$\mathbf{X_{42}}$	26	93

Table B.4: The Burt Band of the conditional space

N	$\mathbf{N_{T=1}}$	$\mathbf{N_{T=0}}$	**actual between**	**confidence interval**	α	**BEC**
764	258	506	0.1924	(0;0.0052)	0.05	15%

Table B.5: The confidence interval for inertia between

Group	**Q**	**J**	**n**	$\mathbf{n_{T=1}}$	$\mathbf{n_{T=0}}$	$\mathbf{I_b}$	**Interval**	α	**balance**	**ATE**
Subclassification on the propensity score										
BIN 1	3	7	190	7	183	0.0205	(0;0.22)	0.05	yes	10.8
BIN 2	4	9	100	10	90	0	(0;0.038)	0.05	yes	10.832
BIN 3	4	9	184	11	173	0	(0;0.021)	0.05	yes	10.812
BIN 4	4	9	133	80	53	0.0247	(0;0.029)	0.05	yes	10.657
BIN 5	4	9	157	150	7	0	(0;0,024)	0.05	yes	10.88

Table B.6: The subclassification on the estimated PS

Groups	n	$n_{T=1}$	$n_{T=0}$	I_b	Interval for I_b	achieved balance	ATE
2-clusters partition							
(Group 1)	645	232	413	0.2559	(0;0.006)	no	8.7634
(Group 2)	119	26	93	0.3170	(0;0.035)	no	7.8254
4-clusters partition							
(Group 1)	268	91	177	0.3823	(0;0.0145)	no	8.9507
(Group 2)	115	74	41	1.1465	(0;0.032)	no	8.9519
(Group 3)	262	67	195	0.2607	(0;0.0149)	no	9.288
(Group 4)	119	26	93	0.3170	(0;0.0352)	no	7.8254
6-clusters partition							
(Group 1)	84	80	4	0	(0;0.0356)	yes	10.782
(Group 2)	184	11	173	0	(0;0.0162)	yes	10.874
(Group 3)	115	74	41	0.397	(0;0.0320)	no	8.9519
(Group 4)	115	20	95	0.125	(0;0.026)	no	9.8479
(Group 5)	147	47	100	0.031	(0;0.0203)	no	9.1138
(Group 6)	119	26	93	0.317	(0;0.0352)	no	7.8254
8-clusters partition							
(Group 1)	84	80	4	0	(0;0.0356)	yes	10.782
(Group 2)	84	4	80	0	(0;0.0356)	yes	10.832
(Group 3)	100	7	93	0	(0;0.0299)	yes	10.832
(Group 4)	115	74	41	0.397	(0;0.0260)	no	8.9519
(Group 5)	100	10	9 0	0	(0;0.0299)	yes	10.832
(Group 6)	62	50	12	0	(0;0.048)	yes	10.789
(Group 7)	100	7	93	0	(0;0.0299)	yes	10.832
(Group 8)	119	26	93	0.31704	(0;0.0352)	no	7.8254
10-clusters partition							
(Group 1)	84	80	4	0	(0;0.0356)	yes	10.782
(Group 2)	84	4	80	0	(0;0.0356)	yes	10.832
(Group 3)	100	7	93	0	(0;0.0299)	yes	10.832
(Group 4)	42	4	38	0	(0;0.0878)	yes	10.84
(Group 5)	73	70	3	0	(0;0.0505)	yes	10.851
(Group 6)	100	10	90	0	(0;0.0299)	yes	10.832
(Group 7)	62	50	12	0	(0;0.0595)	yes	10.789
(Group 8)	100	7	93	0	(0;0.0299)	yes	10.832
(Group 9)	35	5	30	0.5186	(0;0.1054)	no	8.5788
(Group 10)	84	21	63	0.4885	(0;0.0356)	no	7.5718
12-clusters partition							
(Group 1)	41	40	1	0	(0;0.09)	yes	10.832
(Group 2)	43	40	3	0	(0;0.0858)	yes	10.832
(Group 3)	84	4	80	0	(0;0.0356)	yes	10.832
(Group 4)	100	7	93	0	(0;0.0299)	yes	10.832
(Group 5)	42	4	38	0	(0;0.0878)	yes	10.84
(Group 6)	73	70	3	0	(0;0.0505)	yes	10.851
(Group 7)	100	10	90	0	(0;0.0299)	yes	10.832
(Group 8)	62	50	12	0	(0;0.0595)	yes	10.789
(Group 9)	100	7	93	0	(0;0.0299)	yes	10.832
(Group 10)	35	5	30	0.5186	(0;0.1054)	no	8.5788
(Group 11)	44	21	23	0.4952	(0;0.0838)	no	9.8437
(Group 12)	40	0	40	no common support			

Table B.7: Clusters solution set: Ward's method

Groups	n	$n_{T=1}$	$n_{T=0}$	I_b	Interval for I_b	balanced	ATE	std err
2-clusters partition								
(Group 1)	303	105	198	0.1954	(0;0.0103)	no	9.0005	0.1192
(Group 2)	461	153	308	0.1966	(0;0.0076)	no	8.0884	0.141
4-clusters partition								
(Group 1)	268	91	177	0.1911	(0; 0,0103)	no	8.9507	0.0754
(Group 2)	35	14	21	0.2256	(0;0.0790)	no	8.7607	0.1278
(Group 3)	377	141	236	0.1934	(0;0.0083)	no	8.4518	0.1163
(Group 4)	84	12	72	0.2084	(0;0.0374)	no	8.6988	0.2861
6-clusters partition								
(Group 1)	268	91	177	0.1911	(0;0.0103)	no	8.9507	0.0644
(Group 2)	35	14	21	0.2256	(0;0.0790)	no	8.7607	0.1278
(Group 3)	115	74	41	0.1984	(0;0.0240)	no	8.9519	0.1034
(Group 4)	35	5	30	0.2593	(0;0.0790)	no	8.5788	0.0472
(Group 5)	262	67	195	0.1303	(0;0.1056)	no	9.2880	0.0976
(Group 6)	49	7	42	0.1921	(0;0.0564)	no	8.7845	0.1933
8-clusters partition								
(Group 1)	268	91	177	0.1911	(0;0.0103)	no	8.9507	0.0644
(Group 2)	35	14	21	0.2256	(0;0.0790)	no	8.7607	0.1278
(Group 3)	115	74	41	0.1984	(0;0.0240)	no	8.9519	0.1034
(Group 4)	35	5	30	0.2593	(0;0.0790)	no	8.5788	0.0472
(Group 5)	62	50	12	0.0099	(0;0.0380)	yes	10.789	0.0274
(Group 6)	9	7	2	0.0892	(0;0.2635)	yes	10.975	0.0724
(Group 7)	200	17	183	0.0007	(0;0.0211)	yes	10.851	0.0254
(Group 8)	40	0	40	no common support				
10-clusters partition								
(Group 1)	268	91	177	0.1911	(0;0.0103)	no	8.9507	0.0644
(Group 2)	15	14	1	0.0089	(0;0.1581)	yes	10.761	0.1029
(Group 3)	115	74	41	0.1984	(0;0.0240)	no	8.9519	0.1034
(Group 4)	5	5	0	no common support				
(Group 5)	62	50	12	0.0099	(0;0.038)	yes	10.789	0.0274
(Group 6)	9	7	2	0.0892	(0;0.2635)	yes	10.975	0.0724
(Group 7)	20	0	20	no common support				
(Group 8)	30	0	30	no common support				
(Group 9)	200	17	183	0.0007	(0;0.0211)	yes	10.831	0.0254
(Group 10)	40	0	40	no common support				
12-clusters partition								
(Group 1)	84	80	4	0.0028	(0;0.0029)	yes	10.782	0.0515
(Group 2)	15	14	1	0.0089	(0;0.2585)	yes	10.761	0.1029
(Group 3)	73	70	3	0.0003	(0;0.0579)	yes	10.851	0.0591
(Group 4)	5	5	0	no common support				
(Group 5)	62	50	12	0.0099	(0;0.068)	yes	10.789	0.0274
(Group 6)	9	7	2	0.0892	(0;0.4308)	yes	10.975	0.0724
(Group 7)	184	11	173	0.0005	(0;0.0244)	yes	10.812	0.0311
(Group 8)	20	0	20	no common support				
(Group 9)	42	4	38	0.0001	(0;0.1007)	yes	10.832	0.0442
(Group 10)	30	0	30	no common support				
(Group 11)	200	17	183	0.0007	(0;0.02115)	yes	10.832	0.0254
(Group 12)	40	0	40	no common support				

Table B.8: Clusters solution set: single linkage method

Groups	n	$n_{T=1}$	$n_{T=0}$	I_b	Interval for I_b	balanced	ATE	std err
2-clusters partition								
(Group 1)	303	105	198	0.1954	(0;0.0103)	no	9.0005	0.1192
(Group 2)	461	153	308	0.1966	(0;0.0076)	no	8.0884	0.141
4-clusters partition								
(Group 1)	268	91	177	0.1911	(0; 0,0103)	no	8.9507	0.0754
(Group 2)	35	14	21	0.2256	(0;0.0790)	no	8.7607	0.1278
(Group 3)	377	141	236	0.1934	(0;0.0083)	no	8.4518	0.1163
(Group 4)	84	12	72	0.2084	(0;0.0374)	no	8.6988	0.2861
6-clusters partition								
(Group 1)	268	91	177	0.1911	(0;0.0103)	no	8.9507	0.0644
(Group 2)	35	14	21	0.2256	(0;0.0790)	no	8.7607	0.1278
(Group 3)	115	74	41	0.1984	(0;0.0240)	no	8.9519	0.1034
(Group 4)	35	5	30	0.2593	(0;0.0790)	no	8.5788	0.0472
(Group 5)	262	67	195	0.1303	(0;0.1056)	no	9.2880	0.0976
(Group 6)	49	7	42	0.1921	(0;0.0564)	no	8.7845	0.1933
8-clusters partition								
(Group 1)	268	91	177	0.1911	(0;0.0103)	no	8.9507	0.0644
(Group 2)	15	14	1	0.0089	(0;0.1581)	yes	10.761	0.1029
(Group 3)	73	70	3	0.0003	(0;0.0579)	yes	10.851	0.0591
(Group 4)	35	5	30	0.2593	(0;0.0790)	no	8.5788	0.0472
(Group 5)	262	67	195	0.1303	(0;0.0105)	no	9.288	0.0976
(Group 6)	49	7	42	0.1921	(0;0.0564)	no	8.7845	0.1933
(Group 7)	20	0	20	no common support				
(Group 8)	42	4	38	0.0001	(0;0.1007)	yes	10.84	0.0442
10-clusters partition								
(Group 1)	268	91	177	0.1911	(0;0.0103)	no	8.9507	0.0644
(Group 2)	15	14	1	0.0089	(0;0.1581)	yes	10.761	0.1029
(Group 3)	73	70	3	0.0003	(0;0.0579)	yes	10.851	0.0591
(Group 4)	5	5	0	no common support				
(Group 5)	262	67	195	0.1303	(0;0.0105)	no	9.288	0.0976
(Group 6)	9	7	2	0.0892	(0;0.2635)	yes	10.975	0.0724
(Group 7)	20	0	20	no common support				
(Group 8)	42	4	38	0.0001	(0;0.1007)	yes	10.84	0.042
(Group 9)	30	0	30	no common support				
(Group 10)	40	0	40	no common support				
12-clusters partition								
(Group 1)	84	80	4	0.0028	(0;0.0029)	yes	10.782	0.0515
(Group 2)	15	14	1	0.0089	(0;0.2585)	yes	10.761	0.1029
(Group 3)	73	70	3	0.0003	(0;0.0579)	yes	10.851	0.0591
(Group 4)	5	5	0	no common support				
(Group 5)	62	50	12	0.0099	(0;0.068)	yes	10.789	0.0274
(Group 6)	9	7	2	0.0892	(0;0.4308)	yes	10.975	0.0724
(Group 7)	184	11	173	0.0005	(0;0.0244)	yes	10.812	0.0311
(Group 8)	20	0	20	no common support				
(Group 9)	42	4	38	0.0001	(0;0.1007)	yes	10.832	0.0442
(Group 10)	30	0	30	no common support				
(Group 11)	200	17	183	0.0007	(0;0.02115)	yes	10.832	0.0254
(Group 12)	40	0	40	no common support				

Table B.9: Clusters solution set: complete linkage method

Groups	n	$n_{T=1}$	$n_{T=0}$	I_b	Interval for I_b	balanced	ATE	std err
2-clusters partition								
(Group 1)	303	105	198	0.1954	(0;0.0103)	no	9.0005	0.1192
(Group 2)	461	153	308	0.1966	(0;0.0076)	no	8.0884	0.141
4-clusters partition								
(Group 1)	303	105	198	0.1954	(0; 0.0103)	no	9.0005	0.1192
(Group 2)	115	74	41	0.1984	(0;0.0249)	no	8.9519	0.1034
(Group 3)	84	12	72	0.2084	(0;0.0374)	no	8.6988	0.2861
(Group 4)	262	67	195	0.1303	(0;0.1056)	no	9.288	0.0976
6-clusters partition								
(Group 1)	268	91	177	0.1911	(0;0.0103)	no	8.9507	0.0644
(Group 2)	35	14	21	0.2256	(0;0.0790)	no	8.7607	0.1278
(Group 3)	115	74	41	0.1984	(0;0.0240)	no	8.9519	0.1034
(Group 4)	35	5	30	0.2593	(0;0.0790)	no	8.5788	0.0472
(Group 5)	262	67	195	0.1303	(0;0.1056)	no	9.2880	0.0976
(Group 6)	49	7	42	0.1921	(0;0.0564)	no	8.7845	0.1933
8-clusters partition								
(Group 1)	84	80	4	0.0028	(0;0.0029)	yes	10.782	0.0515
(Group 2)	35	14	21	0.2256	(0;0.0790)	no	8.7607	0.1278
(Group 3)	115	74	41	0.1984	(0;0.0240)	no	8.9519	0.1034
(Group 4)	5	5	0	no common support				
(Group 5)	262	67	195	0.1303	(0;0.0105)	no	9.288	0.0976
(Group 6)	49	7	42	0.1921	(0;0.0564)	no	8.7845	0.1933
(Group 7)	184	11	173	0.0005	(0; 0.0244)	yes	10.812	0.0311
(Group 8)	30	0	30	no common support				
10-clusters partition								
(Group 1)	268	91	177	0.1911	(0;0.0103)	no	8.9507	0.0644
(Group 2)	15	14	1	0.0089	(0;0.1581)	yes	10.761	0.1029
(Group 3)	73	70	3	0.0003	(0;0.0579)	yes	10.851	0.0591
(Group 4)	5	5	0	no common support				
(Group 5)	262	67	195	0.1303	(0;0.0105)	no	9.288	0.0976
(Group 6)	9	7	2	0.0892	(0;0.2635)	yes	10.975	0.0724
(Group 7)	20	0	20	no common support				
(Group 8)	42	4	38	0.0001	(0;0.1007)	yes	10.84	0.042
(Group 9)	30	0	30	no common support				
(Group 10)	40	0	40	no common support				
12-clusters partition								
(Group 1)	84	80	4	0.0028	(0;0.0029)	yes	10.782	0.0515
(Group 2)	15	14	1	0.0089	(0;0.2585)	yes	10.761	0.1029
(Group 3)	73	70	3	0.0003	(0;0.0579)	yes	10.851	0.0591
(Group 4)	5	5	0	no common support				
(Group 5)	62	50	12	0.0099	(0;0.068)	yes	10.789	0.0274
(Group 6)	9	7	2	0.0892	(0;0.4308)	yes	10.975	0.0724
(Group 7)	184	11	173	0.0005	(0;0.0244)	yes	10.812	0.0311
(Group 8)	20	0	20	no common support				
(Group 9)	42	4	38	0.0001	(0;0.1007)	yes	10.832	0.0442
(Group 10)	30	0	30	no common support				
(Group 11)	200	17	183	0.0007	(0;0.02115)	yes	10.832	0.0254
(Group 12)	40	0	40	no common support				

Table B.10: Clusters solution set: average linkage method

Method	10-clusters solution	12-clusters solution
Ward	119	119
Single linkage	478	95
Complete linkage	625	95
Average linkage	625	95

Table B.11: Discarded units

n-clusters	**Ward**	**Single**	**Complete**	**Average**
2	0	0	0	0
4	0	0	0	0
6	2	0	0	0
8	6	3	3	2
10	8	4	4	4
12	9	8	8	8

Table B.12: Balanced groups

Appendix C

Descriptive Analysis of real data

Legal Form	**Not subsidized**	**PSA 2001/2002**	**PSA 2003/2005**
Individual Firms	216 (30%)	53 (36%)	21 (19%)
Limited Liability Companies	111 (15%)	23 (16%)	37 (31%)
Partnerships	326 (45%)	61 (41%)	48 (40%)
Others	68 (10%)	10 (7%)	12 (10%)
Total	**721 (100%)**	**147(100%)**	**119 (100%)**

Table C.1: Legal form

County Code	**Not subsidized**	**PSA 2001/2002**	**PSA 2003/2005**
Arezzo	180	45	25
Firenze	268	45	44
Grosseto	5	1	1
Livorno	3	1	0
Lucca	61	20	4
Massa Carrara	31	8	13
Pisa	22	2	2
Prato	50	8	9
Pistoia	73	12	11
Siena	28	5	10
Total	**721**	**147**	**119**

Table C.2: County code

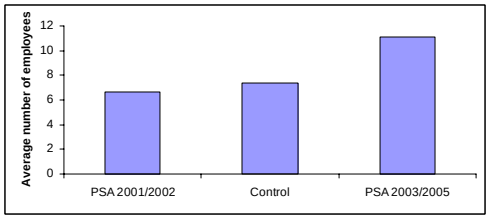

Figure C.1: employees

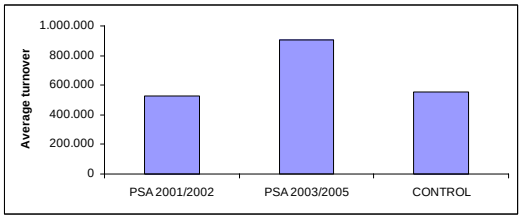

Figure C.2: The average turnover

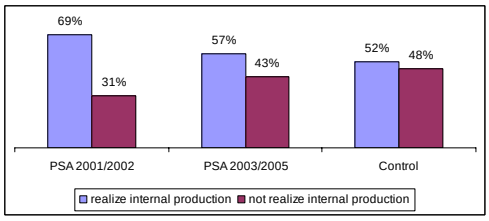

Figure C.3: Internal production

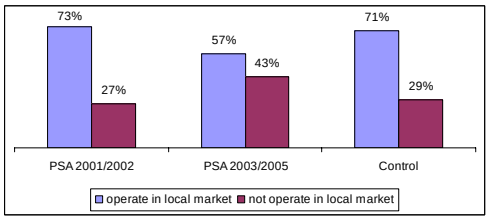

Figure C.4: operate in local market

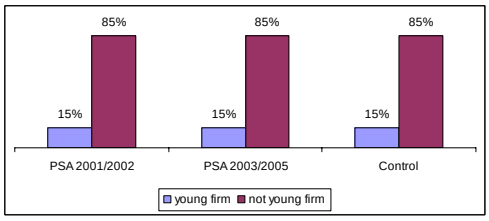

Figure C.5: Being or not young firm

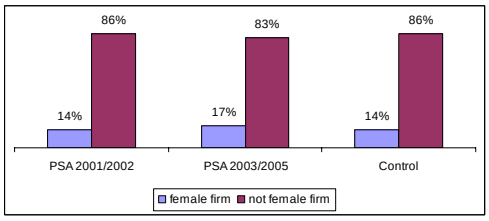

Figure C.6: being or not female firm

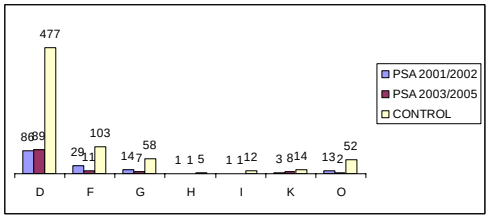

Figure C.7: section

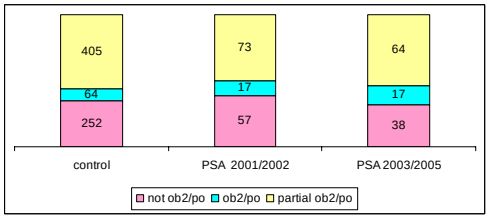

Figure C.8: area ob2

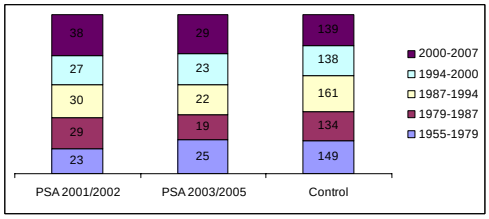

Figure C.9: birth date of firms

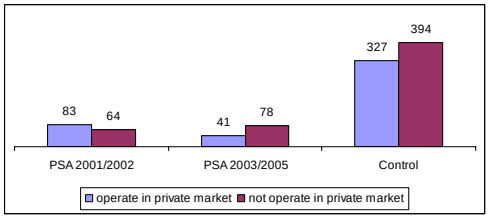

Figure C.10: Operate in private market

Appendix D

Impact Analysis of PSA programs

Method	Beneficiaries	Treated	Controls	ATT	ATE
Subclassification on PS	PSA 2001/2002	147	721	0.1054 (0.0954)	-0.1516 (0.0899)
18-Clusters (Ward Method)	PSA 2001/2002	147	721	0.14335 (0.5136)	0.0803 (0.08972)
Subclassification on PS	PSA 2003/2005	119	721	1.2988 (0.5552)	0.5653 (0.2027)
14-clusters (Ward Method)	PSA 2003/2005	119	721	1.2244 (0.5462)	0.9522 (0.2028)

Table D.1: Comparing results