TOPICS IN CAUSAL AND HIGH DIMENSIONAL INFERENCE

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Qingyuan Zhao

July 2016

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Trevor Hastie, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Art Owen

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Robert Tibshirani

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

Causality is a central concept in science and philosophy. With the ever increasing amount and complexity of data being collected, statistics is playing a more and more significant role in the inference of causes and effects.

This thesis consists of three Parts. Part I reviews the necessary mathematical languages so we can talk about causality. Three major approaches (by potential outcomes in Chapter 2, by graphs in Chapter 3, and by functions in Chapter 4) are described. They are in many senses equivalent or complementary to each other, and excel in different causal tasks. Part II considers the statistical inference of a single causal effect in the potential outcome framework. Chapter 5 reviews state-of-the-art matching and weighting methods. Chapter 6 proposes a new loss function tailored for propensity score estimation, which can boost the performance of the weighting methods. Chapter 7 reviews outcome regression and doubly robust inference and provides some insight to selecting propensity score models and constructing confidence intervals. Part III considers the statistical inference of multiple confounded effects. Chapter 8 introduces a confounding problem in linear model with latent variables. Two examples are given, one in genetics and one in finance. Chapter 9 proposes a twp-step procedure to adjust for the hidden confounding variables in high dimensional data. Chapter 10 presents the performance of this method in simulations and the two read data examples.

Acknowledgments

When writing this page of gratitude, I have physically left Stanford for a week, but I feel my heart is still with the Farm and all the people there. I would like to take this opportunity to express my deepest gratitude to the people who have ever offered help or brought joy to me in the last five years.

First and foremost the thanks go to my amazing advisor, Trevor Hastie, who guided me through the transition from a good student to an independent researcher. Trevor has been very generous with his time. Throughout so many long discussions we had together, I have benefited not only from his deep insights in statistics, but to a greater extent from his attitude towards work and how he collaborates with researchers in other fields. I am sure they will have a lasting influence on my future career as an applied statistician. Trevor has been supportive all through my graduate studies and has given me incredible freedom to pursue my own research ideas. I would like to thank both Lynda and Trevor for the enchanting dinners in their home, which make those Thanksgivings and the graduation especially cherishable.

Next I would like to thank Art Owen, Robert Tibshirani, Mike Baiocchi, Jonathan Taylor for serving my thesis proposal and defense committees. I am also very grateful to all the wonderful faculty and friendly staff at the Stanford Statistics Department. I thank all my teachers for the invaluable classes I have taken in Stanford (especially the first-year core courses).

I have been very fortunate to be able to collaborate with many Stanford and non-Stanford people. I am very grateful to Stanford for encouraging multi-disciplinary research, which is extremely important to statisticians. I thank Art Owen for our several collaborations and for his mentorship throughout my Ph.D. studies. I thank

Jingshu Wang, Murat Erdogdu, Hera He, Charles Zheng, and Kinjal Basu for the inspiring discussions we had. I thank Jure Leskovec, Anand Rajaraman, Andres Stapper, and Yang Song for bringing in exciting problems and trusting my knowledge in statistics. I am indebted to Shaji Sebastian from eBay, Daniel Percival, Jean Steiner, and Sangho Yoon from Google for being good mentors and friends during my summer internships.

The last five years could not be so delightful without the accompany of my friends and classmates. Special shout-out to the extended 206 family—Bhaswar Bhattacharya, Murat Erdogdu, Joshua Loftus and Pooja Loftus, with whom I have shared so many terrific memories. I am sure our friendship will last well beyond graduation. I am grateful to all the students who ever dropped by Sequoia 206 or stopped me in the hallway for some academic or non-academic chats, including Yi Liu, Shilin Zhu, Gourab Mukherjee, Sumit Mukherjee, Alex Chouldechova, Anirban Basak, Jian Li, Dennis Sun, Zhen Zhu, Yunjin Choi, Weijie Su, Stefan Wager, Edgar Dobriban, Ya Le, Charles Zheng, Subhabrata Sen, Rakesh Achanta, Feng Ruan, and Pragya Sur. Special thanks the fourteen USTCers, Zhouxiao Bao, Fang Liu, Xiaoge Liu, Yuling Liu, Shaoyu Lu, Xiaoxiong Lu, Weiluo Ren, Milan Shen, Haotian Wang, Danjie Wenren, Bing Wu, Fan Yang, Hongyuan Yuan, and Zhao Zhao, for starting our journey at Stanford and staying good friends together. I would like to thank the new friends I have made at Stanford, especially Chen Ran, Cong Meng, Yu He, Chunjing Jia, for sharing their enthusiasm in research and life.

Finally, I thank my parents for all the support, sacrifice, and everything they have ever done and continue to do for me. Because of them, I have never felt alone or discouraged facing all the challenges in life. Thank you to Hera He for her love and trust. This thesis is never possible without her being the emotional pillar.

Contents

List of Tables

List of Figures

Part I

MATHEMATICAL LANGUAGES OF CAUSALITY

Chapter 1

Introduction

Causality is one of oldest topics in philosophy and remains a staple in contemporary science. Causality or causation connects one process (the *cause*) with another (the *effect*) and establishes the relationship that the first is (partly) responsible for the second to happen. This concept is so fundamental that it is being used constantly and often unconsciously. "We think we have knowledge of a thing only when we have grasped its cause", said Aristotle in the *Posterior Analytics*.

Philosophers (Aristotle, Hume, Mill, etc.) laid the foundation for causality. But to use causality in our scientific studies on a daily basis, a formal mathematical language is required. Today, probability theory (formalized in 1930s by Kolmogorov) has become the dominant language in most disciplines that use causal modeling, including economics, epidemiology, and social sciences (Pearl, 2009a, Imbens and Rubin, 2015). Being the profession that applies probability theory to analyze data, statisticians have made some of the most important contributions to causality in the twentieth century. Fisher's monumental 1935 book *The Design of Experiments* recognized the indispensable place of randomization in experiments. Rubin and his coauthors' remarkable work in the 1970s and 1980s built a framework that allows us to study causality from observational studies.

Unlike randomized experiments, the second movement of causal inference using observational studies encountered substantial resistance inside the statistics community. Historically, although Fisher's argument against the causal relationship of smoking and lung cancer in the 1950s is now regarded erroneous, his concern about uncontrolled observational studies is certainly not dismissed. This dissertation takes a pragmatic empiricism position: since so many observational studies have already proved their ability of finding causation (Rosenbaum, 2002), I believe the studies and the statistical methods definitely have tremendous value. This view is extremely helpful as the ever-exploding data availability (mostly observational) is pushing statistics to an era of "big data". However, great deal of caution is vital in analyzing these observational datasets. Researchers must fully understand the assumptions and limitations of the statistical methods, otherwise enormous mistakes can be made.

Due to its vicinity to contemporary philosophy and science, causality has received some of the most fierce discussion in statistical journals. There are mainly two lasting debates:

1. Can we use observational studies to learn causal relationships?

2. What mathematical language (or model) should we use to study causality?

The two questions are of course related, but the discussants are often quite different.

Next, we shall turn our attention to the article "Causal Inference Without Counterfactuals" by Dawid (2000) and the discussion therein by Cox, Casella and Schwartz, Pearl, Robins and Greenland, Rubin, Shafer, and Wasserman. The long list of discussants include most of the important figures that have shaped the current landscape of causal inference. It is interesting that they (all the discussants excluding Shafer and perhaps Cox) seemed to reach a consensus facing Dawid's argument against the usage of counterfactuals in causal inference (a negative answer to question 1), but clashed heavily on question 2 in other debates, most noticeably between Pearl (2009b) and Rubin (2009).

A. P. Dawid is a renowned theoretical statistician who accomplished "path-breaking work on predictive probability", quoted from Glenn Shafer's comment. By examining his long publication list, it seems that Dawid did not work directly on causal inference prior to 2000. That makes his 2000 paper very intriguing and special—a respectful decision theorist took a stand against the use of counterfactuals, which are at the very core of causal inference. The article and all the comments really took the discussion of causality in statistics to a very advanced level, making it a perfect start for this dissertation.

Three levels of questions

Holland (1986), Dawid (2000), and Pearl (2009a) all suggest we should distinguish between the following three types of questions:

Associational: How many people take aspirin when they have a headache?

Interventional (effects of causes): "I have a headache. Will it help if I take aspirin?"

Counterfactual (causes of effects): "My headache has gone. Is it because I took aspirin?"

Notice that the first question is non-causal and does not appear in Holland (1986) or Dawid (2000). I add it to the list in order to distinguish associational inference from causal inference.

Classical statistics champion the first task. Take regression for example, we observe predictors X and responses Y and wish to infer the conditional distribution of Y given X. Commonly, we can embed the conditional distribution in a parametric family $\mathrm{P}(Y|X) = \mathrm{P}_\theta(Y|X)$ and use the likelihood principle to infer the parameter θ. A typical example is the (Gaussian) linear regression:

$$\mathrm{P}_{\beta,\sigma}(Y|X) \propto \exp\left\{-\frac{(y-\beta^T x)^2}{2\sigma^2}\right\}. \tag{1.1}$$

This type of inference is most useful to make predictions, where the probability distribution is assumed unchanged in the future. New statistical methods, such as the statistical learning algorithms described in Hastie et al. (2009) and the Bayesian methods used by Nate Silver to correctly predict U.S. 2012 election, have pushed the associational inference to a new level of sophistication.

But association does not imply causation, as warned by every instructor in their first class on regression. One of the most famous examples is a graph correlating global warming with the decline in the number of pirates.[1] Of course neither of them is the cause of the other. The real question is, can we use statistics to answer the causal questions (the interventional and counterfactual questions)?

Statisticians have mixed opinions about this, that is why Dawid (2000) decided to adopt the popular counterfactual language of causality to see how far he can get. Consider a large homogeneous population $\mathcal{U}$, to each of which we can choose to apply any one treatment out of $\mathcal{T} = 0, 1$ and observe the resulting response Y. The counterfactual approach focuses on the collection of potential outcomes $\{Y_u(i) : \ i \in \mathcal{T}, \ u \in \mathcal{U}\}$, where $Y_u(i)$ denotes "the response that would be observed if treatment i were assigned to unit u". Note that, for any unit u, one can observe $Y_u(i)$ for at most one treatment i. This is referred to as "the fundamental problem of causal inference" by Holland (1986).

One of the main concerns of Dawid (2000) is: should we or is it necessary to impose assumptions on the joint distribution of the potential outcomes $Y(0)$ and $Y(1)$, even if they are never jointly observed? The reader can bear this question in mind when reading the rest of this Part. Personally, I believe, after reading Richardson and Robins (2013), that the cross-world independencies are unnecessary to answer interventional queries and some counterfactual queries (as long as all the counterfactuals are in the same world). See Sections 3.3 and 4.3 for some discussion.

The rest of this Part describes three major mathematical languages designed to answer the "effects of causes" and "causes of effects" questions. The mathematical models are the basis of my doctoral research in Parts II and III. Personally, I agree with Lauritzen (2004)'s view that "different formalisms of causality are different 'languages' ", and "I have no difficulty accepting that potential responses, structural equations, and graphical models coexist as languages expressing causal concepts each with their virtues and vices." So I will try to maintain an objective attitude when introducing these languages.

[1] To see the graph, visit this webpage http://www.venganza.org/about/open-letter/.

Chapter 2

The Potential Outcome Approach

Among the causal languages described in Part I of this dissertation, potential outcome is perhaps the most widely adopted approach by applied researchers. The potential outcome language is generally attributed to Donald Rubin's work in 1970s, though Rubin (1990) himself thinks Neyman (1923) first used this language in randomized experiments. Due to this reason, this approach is commonly called the Rubin causal model or the Neyman-Rubin causal model. This Chapter first describes the mathematical model in Neyman (1923) and then discusses how that is related to the potential outcome language we use nowadays.

2.1 Neyman's model

In a Section of his doctoral thesis, Neyman (1923) studies agricultural field experiments with m plots and v varieties of crops. He begins with the notation that u_{ik} is the potential yield of the ith variety on the kth plot and then considers an urn model (sampling with replacement) to apply the varieties. This is equivalent to the randomized experiment with m/v plots exposed to each variety. Letting X_i be the sample mean of the n plots actually exposed to the i-th variety, Neyman shows that

$$\mathrm{E}[X_i - X_j | u] = \frac{1}{m}\sum_{k=1}^{m} u_{ik} - \frac{1}{m}\sum_{k=1}^{m} u_{jk},$$

and he also computes the variance of $X_i - X_j$ given u. Notice that the entire randomness of this hypothetical procedure comes from the random assignment (Neyman's urn model). The potential outcomes u_{ik} are assumed fixed. Interestingly, Neyman himself only considers this as a *theoretical* treatment and "the randomization was considered as a prerequisite to probabilistic treatment of the results" (Reid, 1982). He attributes the *physically* randomized experiments to Fisher and his followers.

Neyman's thesis was originally written in Polish and translated to English in 1990 by Dabrowska and Speed. Rubin (1990) provided some historic comments on the development of the potential outcome approach. According to Rubin (1990), "Neyman's notation seems to have formalized ideas that were relatively firmly in the minds of some experimenters, philosophers and scientists [including Fisher] prior to 1923". After Neyman (1923) and before Rubin's introduction of potential outcomes to observational studies, the standard approach used one variable to represent the observed outcome and an indicator to represent treatment assignment. This notation without counterfactuals limits the ability of applying statistical methods to observational studies.

2.2 Rubin's model

We now turn to Rubin's causal model of observational studies. Motivated by educational researches, Rubin (1974) argued that "the use of carefully controlled nonrandomized data to estimate causal effects is a reasonable and necessary procedure in many cases". Quite naturally and without referring to Neyman's work half a century ago, Rubin (1974) used the potential outcomes to define the causal effect of an educational treatment. This language clearly links observational studies to the more general "missing data" problem (Rubin, 1976, 1977).

In Rubin's view, the most important quantity about observational studies is the treatment *assignment mechanism*. Let $T_i \in \{0, 1\}$ indicate the assignment for unit i, where 1 implies the active treatment and 0 implies the control. Let X_i be some pretreatment covariates of unit i and $Y_i(0)$ and $Y_i(1)$ be the potential outcomes. Finally let T, X, $Y(0)$, $Y(1)$ be the vectors (or matrices) of the individual-level values. The assignment mechanism can be written as

$$\mathrm{P}(T|X, Y(0), Y(1)).$$

Rubin (2011) regards T as the only random variable and considers the other values (called "science" by Rubin) fixed, though he also allows a Bayesian framework that models the science. Rubin (1978) pointed out that all randomized experiments share a critical property called "ignorable",

$$\mathrm{P}(T|X, Y(0), Y(1)) = \mathrm{P}(T|X, Y). \tag{2.1}$$

Under this assumption, it is justifiable to "ignore" the missing values (unobserved potential outcomes). Here $Y = (Y_1, \ldots, Y_n)$ and $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. A stronger and theoretically more convenient assumption is "strongly ignorable" (Rosenbaum and Rubin, 1983) or "unconfounded",

$$\mathrm{P}(T|X, Y(0), Y(1)) = \mathrm{P}(T|X). \tag{2.2}$$

Sequentially experiments were described with (2.1), whereas classical randomized experiments were described with (2.2) (Rubin, 2011). Another important property of randomized experiment is that the assignment probability is between 0 and 1,

$$0 < \mathrm{P}(T|X, Y) < 1, \text{ or } 0 < \mathrm{P}(T|X) < 1. \tag{2.3}$$

Intuitively this says every unit has a positive chance of being assigned to treatment or control. In the context of observational studies, (2.3) is also called the overlap assumption.

One of the fundamental benefits of adopting a counterfactual language is that it allows easy definition of *causal effect*. This is the interventional question described in Chapter 1 (effects of causes). The individual causal effect can be represented as $Y_i(1) - Y_i(0)$. Holland (1986) would say the treatment causes the effect $Y_i(1) - Y_i(0)$. However, this individual treatment effect is never observable. In many cases (actually most of the traditional observation studies), the quantity of interest (estimand) is the *average treatment effect (ATE)*,

$$\tau_{\mathrm{ATE}} = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0), \tag{2.4}$$

or its variants, such as the *average treatment effect on the treated (ATT)*,

$$\tau_{\mathrm{ATT}} = \mathrm{E}\left[\frac{1}{n_1}\sum_{T_i=1} Y_i(1) - Y_i(0)\right],\ n_1 = \#\{1 \le i \le n : T_i = 1\}. \tag{2.5}$$

The expectation in (2.5) is taken over the treatment assignment, since that is the only random quantity in Rubin's potential outcome framework.

In the definitions above, selecting units occurs before the treatment assignment so the the estimands in (2.4) and (2.5) are defined with respect to the fixed n units. It is often more practical to view the n units as random samples from a large population. Theoretically, it is more convenient to assume the population is infinite, so the units are i.i.d. draws. Taking this perspective, we can define the estimands as

$$\tau_{\mathrm{ATE}} = \mathrm{E}[Y(1) - Y(0)],\ \tau_{\mathrm{ATT}} = \mathrm{E}[Y(1) - Y(0)|T = 1]. \tag{2.6}$$

In (2.6), the expectations are taken over the joint distribution of $(T, Y(1), Y(0))$, or in other words over a random draw from the infinite population and a random treatment assignment. By adopting this view, we also implicitly assumed the stable unit treatment value assumption (SUTVA) of Rubin (1980), which roughly says there is no interference between units.

Of course we do not know if ignorability (2.1) or strong ignorability (2.2) is true for an observational study. But what can be said if we are willing to assume those conditions? This shall be discussed in Part II of the dissertation.

2.3 Criticisms

Dawid (2000) made several criticisms against the usage of potential outcomes in causal inference.

2.3.1 Metaphysical and untestable assumptions

Dawid (2000) views the existence of the potential outcomes (u in Neyman's model and $(Y(0), Y(1))$ in Rubin's model) as metaphysical and unjustified. To Dawid, the first concerning fact is that the inference (more specifically, the variance) of τ_{ATE} indeed depends on the model we use. This disobeys the principle that "mathematically distinct models that cannot be distinguished on the basis of empirical observation should lead to indistinguishable inferences", what Dawid (2000) calls *Jefferey's law*. Then Dawid (2000) turns to the constant treatment effect assumption that is commonly imposed in causal inference. To Dawid, this is a untestable and dangerous assumption.

Dawid (2000) summarizes his criticisms by classifying causal analyses into *sheep* (those who obey Jefferey's law) and *goats* (the rest). In his opinion, the inference based on randomized experiments and decision theory is a sheep, and the potential outcome approach has the potential to generate goats. But unlike the discussion by Shafer (2000), Dawid (2000) holds a more neutral attitude towards to usage of potential outcome. He admits that "specific inferential uses of counterfactual models may turn out to be sheep".

2.3.2 Fatalism

A even harsher criticism of Dawid (2000) is that many counterfactual analyses are based on an attitude he terms *fatalism*. This considers the various potential responses $Y_i(t)$ as predetermined attributes of unit i, waiting only to be uncovered by suitable experimentation. One example of fatalism that Dawid (2000) gives is the counterfactual analyses of treatment non-compliance in Imbens and Rubin (1997), where each patient is supposed categorizable as a complier, a defier, an always taker, or a never taker. Complier means the person who would take the treatment if prescribed, and not take it if not prescribed. Other categories can be defined in a similar manner. Dawid (2000, Section 7.1) argues that "it is only under the unrealistic assumption of fatalism that this group has any meaningful identity, and thus only in this case could such inferences even begin to have any useful content".

This fatalistic worldview is of course dangerous, but as argued by Casella and Schwartz (2000) in the comment, it is perhaps a "straw man" view that very few statisticians hold. A possible reason of this hollow attack is that cross-world assumptions may look like fatalistic. For example, if we use $P \in \{0, 1\}$ to indicate the prescription, $T(p) \in \{0, 1\}$ to indicate taking the prescription and $Y(p, t)$ to indicate the potential outcome, then the average treatment effect for compliers (one version) can be defined as $\mathrm{E}[Y(1, 1) - Y(0, 0)|T(0) = 0, T(1) = 1]$. Of course we can never observe $T(0)$ and $T(1)$ together, but this usage of potential outcome is not fatalistic and should in fact belong to the metaphysical criticism.

2.3.3 Determinism

Dawid (2000) explains the popularity of counterfactual models by an implicit view that all problems of causal inference can be cast in the deterministic paradigm, which he believes is only rarely appropriate. This point is further discussed in Section 4.1 and an alternative view based on predictability is described in Section 4.3. Determinism is deeply connected with our understanding of the physical sciences and their explanatory ambitions, on the one hand, and with our views about human free action on the other. Currently, there is no agreement over whether determinism is true or even whether it can be known true or false (Hoefer, 2016).

Chapter 3

The Graphical Approach

This Chapter is based on Spirtes et al. (2000), Pearl (2009a, Chapters 2, 3) (1st edition in 2000), and Koller and Friedman (2009, Chapters 3, 21).

3.1 Conditional dependence and Bayesian network

Although it is natural for humans to interpret causality by a graph (represent "X causes Y" by an arrow from X to Y), the formal graphical approach was first developed to describe associational knowledge rather than causal knowledge. We shall first briefly review the representation of conditional dependence by a graph—Bayesian network.

Suppose we have a joint distribution P over some set of random variables $X = (X_1, \ldots, X_d)$. The core of the Bayesian network representation is a directed acyclic graph (DAG) $\mathcal{G}$, whose nodes are the random variables in X. This graph $\mathcal{G}$ can be viewed in two very different but indeed equivalent ways (Koller and Friedman, 2009, Section 3.2): first, $\mathcal{G}$ provides the skeleton for a way of factorizing the joint distribution; second, it represents a set of conditional independence assumptions.

Let's define the Bayesian network in the first way. Let $\mathrm{Pa}_{\mathcal{G}}(X_i)$ be the parents of

X_i in graph $\mathcal{G}$. A Bayesian network is a pair $(\mathcal{G}, P)$ such that P factorizes over $\mathcal{G}$, i.e.

$$P(X_1, \ldots, X_d) = \prod_{j=1}^{d} P(X_i | \mathrm{Pa}_{\mathcal{G}}(X_i)).$$

Next we turn to the second representation. Let $\mathcal{I}(P)$ denote the set of conditional independence relationships of the form $X \perp\!\!\!\perp Y|Z$ in P (X, Y, and Z can be multivariate). Further denote $\mathcal{I}_l(\mathcal{G})$ be the local independence relationships in $\mathcal{G}$: $\mathcal{I}_l(\mathcal{G}) = \big\{X \perp\!\!\!\perp \text{non-descendants of } X \,\big|\, \mathrm{pa}_{\mathcal{G}}(X)\big\}$. Then P factorizes over $\mathcal{G}$ if and only if $\mathcal{I}_l(\mathcal{G}) \subseteq \mathcal{I}(P)$.

Given this equivalence, the natural question is: what is $\mathrm{I}(\mathcal{G})$, the set of all (not just local) conditional independence relationships for all P that factorize over $\mathcal{G}$? This can be understood using the ideas of *active path* and *active vertex* on a path. Intuitively, a path is active if it carries information or dependence. Two variables X and Y might be connected by lots of paths in $\mathcal{G}$, where all, some, or none of the paths are active. The variables X and Y are called *d-separated* by another set of variables Z, if all the active paths connecting them are blocked by some variable in Z. To understand what active path means, consider all the possible undirected path between X and Y that go through a third variable Z:

1. $X \rightarrow Z \rightarrow Y$ or $X \leftarrow Z \leftarrow Y$: this is called a chain or causal trail. It is active if and only if Z *is not* included.

2. $X \leftarrow Z \rightarrow Y$: this is called a fork or common cause. It is active if and only if Z *is not* included.

3. $X \rightarrow Z \leftarrow Y$: this is called a collider or common effect. It is active if and only if Z *is* included.

This definition can be further generalized to multivariate X and Y by asking Z to deactivate all the trails connecting one variable in X with another variable in Y. Now we can call $\mathcal{I}(\mathcal{G})$ to be all the independencies that correspond to d-separation: $\mathcal{I}(\mathcal{G}) = \big\{(X \perp\!\!\!\perp Y|Z) : X \text{ and } Y \text{ is d-separated by } Z\big\}$. Notice that at this point $\mathcal{I}(\mathcal{G})$ is solely determined by the graph $\mathcal{G}$ and we have not yet linked it to the distribution P. The connection is elegant: whenever P factorizes over $\mathcal{G}$, we have $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$. The converse of this result is not always true. When it is true, i.e. all the conditional independencies of P can be read from d-separations in $\mathcal{G}$, the distribution P is called *faithful* to $\mathcal{G}$. With probability 1, a distribution factorizes over $\mathcal{G}$ is faithful to G (see Spirtes et al., 2000). Uhler et al. (2013) studied the geometry of the faithfulness assumption and suggested that due to sampling error the faithfulness condition alone is not sufficient for statistical estimation.

3.2 Causal Bayesian network

So far we are just using the graphical structure to represent a probability distribution, i.e. storing associational knowledge. This correspondence is elegant, but an important question remaining to be answer is which graph we should use for a given distribution. For example, the complete graph always provide a factorization for any distribution, but it does not reveal any of the conditional independence structure in the distribution. A "good" Bayesian network structure should be sparser, more natural, and robust to changes such as adding or removing a variable.

A causal Bayesian network represents or tries to represent a stable and autonomous physical mechanism. It allows us to predict the effect of *interventions* (interfering in the natural course of events). This is much more informative than probabilistic models because the network also stores causal knowledge, allowing us to answer the second question in Chapter 1.

Intuitively, we want to represent a interventional distribution by a subgraph that deletes all the arrows pointing to the intervention node(s). Formally, a (discrete) probability distribution P is assumed to have many interventional distributions $P_{\mathrm{do}(X=x)}$. The interventional distributions must be compatible with the original distribution, in the sense that $P_{\mathrm{do}(X=x)}(Y = y) = 1$ if $Y \in X$ and $Y = y$ is consistent with $X = x$, and $P_{\mathrm{do}(X=x)}(Y = y \,|\, \mathrm{pa}_{\mathcal{G}}(Y)) = z = P(Y = y \,|\, \mathrm{pa}_{\mathcal{G}}(Y) = z)$ whenever $\mathrm{pa}_{\mathcal{G}}(Y) = z$ is consistent with $X = x$. A DAG $\mathcal{G}$ is said to be *causal Bayesian network* compatible with P and its compatible interventional distributions if all $P_{\mathrm{do}(X=x)}$ factorize over $\mathcal{G}$.

The main benefit of using a causal Bayesian network is that it gives a efficient representation of P and all its vast interventional distributions. Let V be all the random variables and we can factorize every $P_{\mathrm{do}(X=x)}$ over the edge-deleted subgraph $\mathcal{G}_x = \mathcal{G}_{\mathrm{do}(X=x)}$ by treating $X = x$ as given:

$$\begin{aligned} P_{\mathrm{do}(X=x)}(V=v) &= \prod_{i:V_i \notin X} P(V_i = v_i | \mathrm{pa}_{\mathcal{G}}(V_i) = v_{\mathrm{pa}_{\mathcal{G}}(V_i)}) \\ &= \prod_{i:V_i \notin X} P(V_i = v_i | \mathrm{pa}_{\mathcal{G}_x}(V_i) = v_{\mathrm{pa}_{\mathcal{G}_x}(V_i)}, X = x), \text{ if } X = x \text{ is consistent with } V = v. \end{aligned}$$

This also implies two properties which match our intuition: $\mathrm{P}(X|\mathrm{pa}_{\mathcal{G}}(X)) = \mathrm{P}_{do(\mathrm{pa}_{\mathcal{G}}(X))}(X)$, and $\mathrm{P}_{do(\mathrm{pa}_{\mathcal{G}}(X),Y)}(X)$ does not depend on Y that is disjoint of X and $\mathrm{pa}_{\mathcal{G}}(X)$.

We have laid out the causal Bayesian network in a very condensed way. More detail about this approach can be found in Pearl (2009a).

3.3 Graphical identification of causal effect

Besides its easy visual representation of causality, what are the other reasons to use the graphical approach? Recall that one of the main goals of causal inference is to establish causal relationships from observational studies. In the potential outcome approach described in Chapter 2, this relies on the ignorability assumption (2.1) that is empirically untestable. In contrast, the causal Bayesian network provides a more structured way to choose the variables appropriate for adjustment.

Suppose we are given a causal diagram $\mathcal{G}$ and observational data on a subset of the variables V in $\mathcal{G}$. Our goal is to estimate the causal effect of T on Y. This is the second level question (effects of causes) in Chapter 1. Before estimation, one question that must be answered is: is this causal effect identifiable? In other words, do there exist two distributions P and P (and their interventional distributions) that are compatible with $\mathcal{G}$ and have the same marginal distributions of V, yet the interventional distributions $P_{\mathrm{do}(T=t)}(Y)$ and $P'_{\mathrm{do}(T=t)}(Y)$ are different?

Pearl (1993) first gave the graphical criterion for the strong ignorability assumption (2.2) due to Rosenbaum and Rubin (1983). This graphical criterion is called the *back-door* criterion, meaning X blocks all the "back-door" trails of T to Y. Formally, this means that no variable in X is a descendant of T, and X blocks every undirected trail between T and Y that contains an arrow *into* T. If X satisfies the back-door criterion relative to T and Y, then the interventional distribution is given by

$$\begin{aligned} P_{\mathrm{do}(T=t)}(Y) &= \sum_x P(Y|T, X = x)P(X = x) \\ &= \sum_x \frac{P(X = x, T, Y)}{P(T|X = x)} \end{aligned}$$

The second formula is called "inverse probability weighting" (see Section 5.4) which makes use of the treatment assignment mechanism (propensity score) $P(T|X = x)$ that is crucial in the potential-outcome analysis.

Putting it in the bigger picture, the back-door condition generalizes the strong ignorability condition (2.2). Recall that strong ignorability says that given X, the value that Y would obtain had the treatment T been t is independent of T. Vaguely speaking, if the counterfactual $Y(t)$ can be represented by the edge-deleted subgraph $\mathcal{G}_{\mathrm{do}(X=x)}$, this statement amounts to saying $Y(t)$ and T are d-separated by X in the subgraph $\mathcal{G}_{\mathrm{do}(X=x)}$. However, this reasoning is not quite rigorous as the subgraph $\mathcal{G}_{\mathrm{do}(X=x)}$ does not contain any node corresponding to the counterfactual variable.

This issue is solved in Richardson and Robins (2013) by changing the formulation of $\mathcal{G}_{\mathrm{do}(T=t)}$. In their single world intervention graph, $\mathcal{G}_{\mathrm{do}(T=t)}$ is obtained by, instead of deleting incoming edges, splitting the node T into two halfs. One half (call it T) inherits the incoming edges of X in $\mathcal{G}$, another half (call it $t = 0$ for example) inherits the outgoing edges of X in $\mathcal{G}$, and the two half nodes are not connected. All the descendants Y of T in $\mathcal{G}$ now become counterfactual variables $Y(t = 0)$, as they are descendants of the half node $t = 0$ in the new graph $\mathcal{G}_{\mathrm{do}(T=t)}$. The conditional independencies involving counterfactuals can be read from the new $\mathcal{G}_{\mathrm{do}(T=t)}$ via d-separation. The back-door criterion is a special case of this approach.

The back-door condition reflects the common advise that the covariates should be unaffected by the treatment. However, this is not necessary. Pearl (1995) gives the *front-door* criterion, in which the adjustment variables X can be the descendants of T. This approach essentially uses the backdoor criterion twice, first computes the causal effect of T on X and then computes the causal effect of X on Y. Based on these two criteria, Pearl (1995) derived a system of rules, called *causal calculus*, to determine the identifiability of a general interventional distribution given the causal diagram $\mathcal{G}$. See Pearl (2009a, Sections 3.3–3.4) for more detail.

Chapter 4

The Functional Approach

4.1 Laplacian determinism

The functional approach originates from Laplace's demon or Laplacian determinism in the history of science. Laplace (1814) presented the following articulation of causal or scientific determinism:

> We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.

To Laplace, the whole universe can be described by all its atoms (variables) and the physical laws (functions).

Pearl (2000) provides the following interpretation of Laplacian determinism: "The essence of any scientific law lies in the claim that certain relationships among observable variables remain invariant when the values of those variables change relative to our immediate observations." Pearl (2000) elaborates on this point with the following example:

> For example, Ohm's law ($V = IR$) asserts that the ratio between the current (I) and the voltage (V) across a resistor remains constant for all values of I, including yet-unobserved values of I. We usually express this claim in a function or a hypothetical sentence: "Had the current in the resistor been I (instead of the observed value I_0) the voltage would have been $V = I(V_0/I_0)$," knowing perfectly well that there is no way to simultaneously measure I and I_0.

Building on his work of causal diagrams (see Chapter 3), Pearl (2009a, Chapter 5) suggests a quasi-deterministic functional language to describe causality. Pearl's nonparametric structural equation model (SEM) assumes that every node (random variable) X in the graph $\mathcal{G}$ can be written as a function of its parents and noise

$$X = f_X(\mathrm{pa}_{\mathcal{G}}(X), \epsilon_X). \tag{4.1}$$

In most cases, Pearl assumes the noise variables ϵ_X are mutually independent. Pearl (2000) views (4.1) as an approximation of Laplace's conception of nature: "randomness surfaces merely due to our ignorance of the underlying boundary conditions".

Structural model is a synonym of functional model. Let's consider the following linear structural equation model (Wright, 1934) which is a special case of (4.1): for a graph $\mathcal{G}$ with nodes $X_1, \ldots, X_d$,

$$X_i = \sum_{j \in \mathrm{pa}_{\mathcal{G}}(X_i)} \alpha_{ij} X_j + \epsilon_i. \tag{4.2}$$

The word *structural* is used to emphasize that the equation (4.2) is different from a linear regression, in the sense that (4.2) also predicts interventional settings. Therefore, a unit (interventional) change of X_j for some $j \in \mathrm{pa}_{\mathcal{G}}(X_i)$ from $X_j = x_j$ to $X_j = x_j + 1$ will result in the same increase of X_i as a unit change from $X_j = x'_j$ to $X_j = x'_j + 1$. Also, once we hold $\mathrm{pa}_{\mathcal{G}}(X_i)$ as constant, changing all other variables in the model will not affect X_i. The same interpretation holds for the Ohm's law (a nonparametric SEM).

I want to make several remarks about Laplacian determinism and structural models. First, SEMs are widely used in many scientific disciplines, including genetics, economics, epidemiology, and education. Criticisms of SEM exist from its first proposal (Wright, 1921) and last till today. Bollen and Pearl (2013) argue that much of the controversy is due to misunderstanding.

Second, determinism is fundamentally connected to the objectivity of probability. If our world is truly deterministic, some philosophers argue that there is no room for objective probability. Third, determinism is also different from predictability, a concept that statisticians are more used to. Hoefer (2016) gives a fantastic discussion on this:

> Laplace probably had God in mind as the powerful intelligence to whose gaze the whole future is open. If not, he should have: 19th and 20th century mathematical studies showed convincingly that neither a finite, nor an infinite but embedded-in-the-world intelligence can have the computing power necessary to predict the actual future, in any world remotely like ours. But even if our aim is only to predict a well-defined subsystem of the world, for a limited period of time, this may be impossible for any reasonable finite agent embedded in the world, as many studies of chaos (sensitive dependence on initial conditions) show. Conversely, certain parts of the world could be highly predictable, in some senses, without the world being deterministic. When it comes to predictability of future events by humans or other finite agents in the world, then, predictability and determinism are simply not logically connected at all.

We will go back to this point later in Section 4.3.

4.2 Functional model and counterfactuals

At first glance, the functional approach, for example the nonparametric SEM in (4.1), may look very different from the counterfactual approach described in Chapter 2. However, the two approaches are indeed equivalent in a strong sense.

One of the motives when Pearl developed nonparametric SEM is to represent the imaginative counterfactuals in terms of mathematical expressions. If we assume the model (4.1) for variables T, X (treating as exogenous), and Y, i.e.

$$T = f(X, \epsilon_T), \ Y = g(T, X, \epsilon_Y),$$

then we can express the counterfactual $Y(0)$ as $g(0, X, \epsilon_Y)$ and $Y(1)$ as $g(1, X, \epsilon_Y)$, two random variables well-defined in the model. To check conditional independence for counterfactual variables, Balke and Pearl (1994) developed an approach called the twin network, which augments the original graph $\mathcal{G}$ with its counterfactuals and connect the two counterparts through the common noise variables. Conditional independence of counterfactuals can be read from the twin network via d-separation.

Richardson and Robins (2013, Section 4.2.3) presents an example where the twin network method fails. Richardson and Robins (2013) propose to use d-separation in their single-world intervention graphs to check counterfactual queries. Instead of doubling the original graph, they rely on a node-splitting operation described in Section 3.3. Counterfactuals are now represented by variables in the intervention graph. They also show that Pearl's functional representation of counterfactuals assume many more cross-world independence assumptions (which are empirically unverifiable) than what is necessary (just single-world independence assumptions).

So far we have not talked about statistical inference of causal models. Recall that there are two different types of causal queries in Chapter 1: "effects of causes" and "causes of effects". As discussed earlier, the three mathematical languages described in Chapters 2 to 4 are essentially equivalent to each other. Nevertheless, some language is more powerful and convenient in answering certain questions:

1. Potential outcomes are designed to describe "effects of causes". In fact, Holland (1986) said that "...an emphasis on the effects of causes rather than on the causes of effects is, in itself, an important consequence of bringing statistical reasoning to bear on the analysis of causation...". In the same article, Holland also argues that only variables with "potential exposability" can be a cause. If we inspect this statement today, this view is rather narrow and is largely due to

the limitation of the potential outcome language. However, there is not reason to dismiss this approach due to the limitation. In fact, the entire Part II of this dissertation is devoted to the potential outcome approach, as it is still the most convenient language for "effects of causes".

2. The functional approach handles large and complex causal networks the best. Statistical inference is also tractable: in principle, one can use maximum likelihood or method of moments to solve structural equation models like (4.1) and (4.2). Numerous methods are developed to test the goodness of fit and implications of SEMs (see Bollen, 2014). With the estimated functions, it is also more convenient to plan for interventions optimally.

3. The graphical approach is by far the clearest and the most concise way of summarizing causal knowledge. It allows easy interpretation of causal models and identification of causal effects. However, the statistical inference of causal Bayesian network is not straightforward. The next subsection describes a recently proposed approach to infer the causal diagram by using data from multiple interventional settings.

4.3 Invariant causal prediction

As noticed by Dawid (2000) in his rejoinder, "all of the discussants [of his paper] except Shafer and Robins and Greenland seem to be out-and-out Laplacian determinists, for who nothing short of a functional model relating outputs to inputs will do as a description of nature". Adopters of this functional model such as (4.1) or (4.2) are surely determinists. How about people who use potential outcomes?

The answer is not obvious. Consider the strong ignorability (2.2) that is commonly assumed. It contains the joint distribution of the potential outcomes $Y(0)$ and $Y(1)$, which are never observed under any experimental setting. However, the causal targets such as $\mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0)]$ are identifiable if we assume marginally $Y(0)$ and $Y(1)$ are independent of T conditional on X. Borrowing terms from Richardson and Robins (2013), this means many potential-outcome adopters assume extra cross-world independence assumptions. It seems to me that this is a consequence of being a determinist.

One advantage of determinism is that the statistical inference is more convenient. To fit a functional model like (4.1) or (4.2), one only needs to postulate the function forms and noise distributions and apply standard statistical tools (e.g. maximum likelihood).

A natural question after Dawid's observation is: Can we talk about causality without being a Laplacian determinist? To clarify, "taking about causality" means to answer the interventional (effects of causes) and counterfactual (causes of effects) questions in Chapter 1. This is an important question because many statisticians[1] and scientists do not believe in this philosophy.

Recently, Peters et al. (2015) propose an alternative approach that can potentially allow the disbelievers of determinism to work on causality. This approach views causality as *invariant prediction* under different environments. An environment can be the observational distribution or any interventional setting. In this definition, causality is a consequence of *predictability* instead of *determinism*. As discussed earlier in Section 4.1, predictability and determinism are different concepts and not logically connected.

Peters et al. (2015) consider the setting where we have different experimental conditions $e \in \mathcal{E}$ and have i.i.d. sample of (X^e, Y^e) in each environment. Peters et al. (2015) are interested in discovering the subset $\mathcal{S}^*$ of variables in X that can "causally" predict Y, in the sense that $Y^e | X^e_{\mathcal{S}^*}$ and $Y^f | X^f_{\mathcal{S}^*}$ are identical for all environments $e, f \in \mathcal{E}$. When $Y^e | X^e_{\mathcal{S}^*}$ satisfies the linear structural equation model (4.2), they provide identification conditions and a systematic procedure to infer this set from empirical data.

Once the causal parents of Y are found, they can be used to answer the causal queries. For interventional queries, Peters et al. (2015) give confidence intervals of the structural coefficients. For counterfactual queries, the answer is also immediate from the estimated prediction formula. Notice that some modularity assumptions are

[1] For example my Ph.D. advisor, Trevor Hastie, said determinism is "a little nutty" to him.

necessary to extend the discovery to an environment $e' \notin \mathcal{E}$ and additional independency assumptions are necessary to answer queries with cross-world counterfactuals. Both points are not adequately discussed in the Peters et al. (2015) paper.

Part II

INFERRING A SINGLE EFFECT

Chapter 5

Matching and Weighting

5.1 Raw matching

Historically, randomized experiments (e.g. Fisher, 1935) are studied before observational studies. The apparent benefit of a randomized experiment is that all pretreatment covariates, observed or unobserved, are always stochastically balanced. If the outcomes of treatment and control populations are different, the only possible cause is the treatment itself. This isolates the treatment effect and allows us to use standard inference tools in statistics. In observational studies, the stochastic covariate balance in general does not hold. For example, patients in worse condition may be more likely to choose certain treatment, making that treatment look worse than the others. Simply ignoring this fact can lead to significant confounding bias.

Due to this reason, it is often desirable to mimic a randomized experiment using observational data. In randomized experiments, one of the most common design technique is *blocking* to remove the effect of nuisance factors. This motivates the raw matching methods described in this Section. The term “raw matching” means that these methods only use the raw pretreatment covariates and do not attempt to model the assignment mechanism (the propensity score).

Any matching algorithm must first specify a distance metric d on the covariates.

One example is exact matching:

$$d(X_i, X_j) = \begin{cases} 0, & \text{if } X_i = X_j, \\ \infty, & \text{if } X_i \neq X_j. \end{cases}$$

In other words, two units are matched only if they have exactly the same covariates. This is often too stringent. In practice, a widely used metric is the Mahalanobis distance:

$$d(X_i, X_j) = \sqrt{(X_i - X_j)^T \Sigma^{-1} (X_i - X_j)}.$$

If the estimand is ATT, Σ is the covariance matrix of X in the control group; if the estimand is ATE, Σ is chosen to be the variance matrix in the pooled treatment and control groups. This distance metric is motivated by the multivariate normal distribution. The Mahalanobis distance can also be generalized with additional weight parameter W (Diamond and Sekhon, 2013a). Formally,

$$d(X_i, X_j) = \sqrt{(X_i - X_j)^T \left(\Sigma^{-\frac{1}{2}}\right)^T W \Sigma^{-\frac{1}{2}} (X_i - X_j)}. \tag{5.1}$$

Once the distance metric d is selected, we can apply a matching algorithm. One of the most common, and easiest to implement and understand, methods is $k:1$ nearest neighbor matching (Rubin, 1973). In this algorithm, each treated unit is matched to k control units that are closest in terms of d, and the control units that are not selected are discarded. Therefore the nearest neighbor matching estimates the ATT. When $k = 1$, matching is the counterpart of paired design in randomized experiments. The user can also choose if a control unit is allowed to be used multiple times as a match (matching with replacement). In matching without replacement, the algorithm usually proceeds in a greedy fashion and the output is sensitive to which treatment units are matched first. It is also possible to avoid this approach and instead minimize a global distance measure, which picks about the same controls but does a better job of assigning matched controls to treated units (Gu and Rosenbaum, 1993).

The state-of-the-art matching algorithm is *Genetic Matching* developed by Diamond and Sekhon (2013a). It is available in the R package `Matching` (Sekhon, 2011).

Given a user-specified criterion of covariate imbalance, it uses a genetic search algorithm to choose the weights W in the generalized Mahalanobis distance, so the matched samples optimize the specified imbalance criterion.

One drawback of the nearest neighbor matching is that some control units are discarded and not used in the analysis. *Subclassfication* and *full matching* instead use all individuals and can estimate either the ATE or the ATT. Another motivation of these approaches is the randomized block design, where each block typically contains more than one treated unit. Subclassification forms groups of units who have similar pretreatment covariates. Full matching is more sophisticated and creates a series of matched sets, where each matched set contains at least one treated unit and at least one control unit. See Hansen (2004) and Stuart (2010) for more detail of these methods.

5.2 Propensity score matching

As mentioned earlier in Section 2.2, the treatment assignment mechanism plays a key role in observational studies. If we are willing to assume the assignment mechanism $\mathrm{P}(T|X)$ is ignorable (2.1) or unconfounded (2.2), the observational study resembles a randomized experiment. How should we then proceed?

In a seminal work, Rosenbaum and Rubin (1983) highlight the role of "propensity score" $p(X) = \mathrm{P}(T = 1|X)$ in observational studies with binary T. Their work is motivated by the difficulty of extending raw matching methods (see section 5.1) to high dimensional covariates. They call a function $b(X)$ of the observed covariates a *balancing score* if $X \perp\!\!\!\perp T \mid b(X)$. In other words, $\mathrm{P}(T \,|\, X) = \mathrm{P}(T \,|\, b(X))$, so the multivariate matching problem is reduced to a univariate matching problem. In this sense, a balancing score is the "sufficient statistic" of non-random treatment assignment. Rosenbaum and Rubin (1983) prove that any function $b(X)$ is a balancing score if and only if it is finer than $p(X)$ in the sense that $b(X) = f(p(X))$ for some function f. In this sense, propensity score is the most basic tool to adjust for covariate imbalance, and sometimes it is more advantageous to match or subclassify not only for $p(X)$ but for other functions of X as well.

In randomized experiments, the propensity score is usually specified before the experiment (e.g. a randomized block design). In observational studies, $p(X)$ is unknown and needs to be estimated from the data. For this purpose, the most commonly used method is the logistic regression solved by maximum likelihood. Here we make two remarks about propensity score estimation: First, propensity score modeling is a means to an end (covariate balance), not an end in itself; Second, Chapter 6 shows that, in order to achieve better covariate balance, certain tailored loss functions should be used instead of the Bernoulli likelihood to estimate the propensity score.

After the propensity scores are estimated, we can in principle apply any matching methods described in Section 5.1. Notice that propensity score is a scalar, so this is a univariate matching problem. The most commonly used distance metrics are

$$\begin{aligned} d(X_i, X_j) &= \left|\hat{p}(X_i) - \hat{p}(X_j)\right|, \text{ and} \\ d(X_i, X_j) &= \left|\mathrm{logit}(\hat{p}(X_i)) - \mathrm{logit}(\hat{p}(X_j))\right|. \end{aligned}$$

Once the distance metric is chosen, one can apply nearest neighbor matching (with or without replacement), subclassification, or full matching described in Section 5.1.

Notice that the distance metric is defined by the *estimated* propensity scores which are not necessarily close to the true ones. The most common reason for this is model misspecification. In the standard practice, the matches/subclasses obtained by propensity score matching must go through a diagnostic step to ensure all covariates are well balanced. If not, one needs to explore other specifications until satisfactory covariate balance is achieved. This cyclic procedure, certainly not very pleasant for applied researchers, is sometimes called the “propensity score tautology” in the literature (Ho et al., 2007).

5.3 Empirical calibration weighting

We will discuss *weighting* methods in the next two Sections. In some sense, the matching methods described in Sections 5.1 and 5.2 are also weighting methods with discrete weights. For example, in 1:1 nearest neighbor matching without replacement, treated units always have weight 1 and control units have weight either 0 or 1. With subclassification or full matching, the weights can be more complicated, but still they are only finite many of possible weights. The weighting methods described in the next two Sections do not have this constraint and thus is inherently different from matching.

In general, weighting methods seek non-negative weights w such that the weighted empirical distributions $F_w(0) = \sum_{T_i=0} w_i \cdot \delta_{X_i}$ and $F_w(1) = \sum_{T_i=1} w_i \cdot \delta_{X_i}$ (δ_x is the point mass at x) are as close as possible. Typically, the distance between these weighted distributions is measured by the standardized difference (Rosenbaum and Rubin, 1985, Austin and Stuart, 2015) with respect to some covariate function $\phi(\cdot)$,

$$d_{\mathrm{sd},\phi(\cdot)}(F_w(0), F_w(1)) = \frac{\left|\mathrm{E}_{F_w(1)}[\phi(X)] - \mathrm{E}_{F_w(0)}[\phi(X)]\right|}{\mathrm{Var}_{F_w(0)+F_w(1)}(\phi(X))}, \tag{5.2}$$

or the univariate Kolmogorov-Smirnov statistics,

$$d_{\mathrm{KS},j} = \max_x |F_{w,j}(1,x) - F_{w,j}(0,x)|, \ j = 1, \ldots, p, \tag{5.3}$$

where $F_{w,j}(t,x)$ ($t = 0$ or 1) is the marginal cumulative distribution function of $F_w(t)$ for the j-th variable evaluated at x.

However, these distance metrics are too complicated and in general non-convex in w, thus it is difficult to minimize them over w directly. To remedy this, we need to find convex alternatives. The genetic matching of Diamond and Sekhon (2013b) is an example of this idea applied to matching. The user usually specifies some combination of the distance metrics in (5.2) and (5.3), but genetic matching instead search over the weighted Mahalanobis distance metrics (5.1), which are much easier to handle.

Now we turn to empirical calibration weighting, which seek weights w such that the standardized difference (5.2) is zero or small for some pre-specified functions ϕ_k, $k = 1, \ldots, m$. Zero standardized difference implies exact sample balance, i.e. the weights w solve

$$\sum_{T_i=1} w_i \phi_k(X_i) = \sum_{T_i=0} w_i \phi_k(X_i), \ k = 1, \ldots, m. \tag{5.4}$$

We should also avoid the trivial solution to (5.4) by asking

$$\sum_{T_i=1} w_i = \sum_{T_i=0} w_i = 1. \tag{5.5}$$

When k is not too large, there are usually infinite number of solutions to (5.4) and (5.5). To pick one of them, we can ask w as close to uniform as possible. The practical reason for this is that the weighted difference estimator

$$\hat{\tau} = \sum_{T_i=1} w_i Y_i - \sum_{T_i=0} w_i Y_i$$

has variance $\sum_{i=1}^n w_i^2 \mathrm{Var}(Y_i|X,T)$ conditional on X and T. If we assume homoskedastic noise $\mathrm{Var}(Y_i|X) = \sigma^2$, then we should pick w to minimize its squared norm $\sum_{i=1}^n w_i^2$.

In general, the empirical calibration method minimizes $\sum_{i=1}^n D(w_i, v_i)$ subject to exact balance (5.4) and the normalization (5.5), where $D(w,v)$ is a function of w that achieves its minimum at v and $\{v_i\}_{i=1}^n$ is a set of uniform weights. In Deville and Särndal (1992), calibration estimator is originally used in survey sampling with non-random samples. The entire population is unobserved, but we know some population moments (expectation, variance, etc.) of the covariates. To estimate the population average of the response, Deville and Särndal (1992) construct weighted survey samples that are calibrated to the known population moments. This the same as estimating ATT in observational studies if we view the treated units as the population whose $Y(0)$'s are missing.

Although calibration estimation is commonly used in survey sampling (Kim and Park, 2010), it is not considered in observational studies until recently. Hainmueller (2011) considers estimating ATT with D being the negative Shannon entropy. Zubizarreta (2015) uses the squared norm of w as the objective but allows inexact balance. Chan et al. (2015) prove that the empirical calibration estimators for ATT and ATE can achieve the semiparametric efficiency bound.

5.4 Propensity score weighting

In this Section we describe propensity score weighting, the continuous counterpart of propensity score matching in Section 5.2. One can draw a similar comparison between empirical calibration weighting in Section 5.3 and raw matching Section 5.1. This is summarized in Table 5.1 below.

	Discrete weights	Continuous weights
By raw covariates	Raw matching	Empirical calibration weighting
By propensity scores	Propensity score matching	Inverse probability weighting

Table 5.1: Matching and weighting methods

Propensity score weighting is commonly called inverse probability weighting (IPW) in the observational study literature, because that is the form of weighting to estimate the ATE. It is first developed by Horvitz and Thompson (1952) to estimate the mean of a population from a stratified random sample (a survey sampling problem), so it is also called Horvitz-Thompson estimator. IPW is applied to account for different proportions of observations within strata in the target population. If p_i is the inclusion probability of the sample Y_i, the Horvitz-Thompson estimator is given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} p_i^{-1} Y_i.$$

In observational studies, the ATE, $\mathrm{E}[Y(1) - Y(0)]$, can be viewed as estimating two population means. Therefore, the IPW estimator is given by the difference of two Horvitz-Thompson estimators

$$\hat{\tau}_{\mathrm{ATE}} = \frac{1}{n}\sum_{T_i=1} \frac{Y_i}{\hat{p}(X_i)} - \frac{1}{n}\sum_{T_i=0} \frac{Y_i}{1-\hat{p}(X_i)}, \tag{5.6}$$

where $\hat{p}(\cdot)$ is the estimated propensity score. This method can also be extended to other estimands. For the ATT, the IPW estimator is given by

$$\hat{\tau}_{\mathrm{ATT}} = \frac{1}{n}\sum_{T_i=1} Y_i - \frac{1}{n}\sum_{T_i=0} \frac{\hat{p}(X_i)}{1-\hat{p}(X_i)} Y_i. \tag{5.7}$$

Chapter 6 considers IPW estimators for more general estimands.

Compared to matching methods, IPW is a "cleaner" and more efficient approach, because the discrete matches are replaced by continuous weights. In a key paper, Hirano et al. (2003) showed that IPW paired with sieve propensity score model can achieve the semiparametric efficiency bound, which gives the theoretical reason to prefer weighting over matching (see Section 6.5.1 for more detail). However, this also comes with a price. The inverse probability weights are more volatile and sensitive to model misspecification. If some estimated propensity score $\hat{p}(X_i)$ is close to 0 (for a treated unit) or 1 (for a control unit), its inverse weight can become very large and unstable, so the IPW estimator may perform poorly in finite sample. One way to mitigate this is to normalize the weights within each treatment group. For example, the normalized IPW estimator of ATE is

$$\hat{\tau}^*_{\text{ATE}} = \left(\sum_{T_i=1} \frac{1}{\hat{p}(X_i)}\right)^{-1} \left(\sum_{T_i=1} \frac{Y_i}{\hat{p}(X_i)}\right) - \left(\sum_{T_i=0} \frac{1}{1-\hat{p}(X_i)}\right)^{-1} \left(\sum_{T_i=0} \frac{Y_i}{1-\hat{p}(X_i)}\right). \tag{5.8}$$

Still, this does not completely solve the instability issue that the estimator could be largely decided by just a few observations.

The instability issue was officially brought up by Kang and Schafer (2007), but it was perhaps well known by practitioners before that. Kang and Schafer (2007) constructed an artificial example in which inverse probability weights are very unstable and the IPW estimator performs poorly. Moreover, they showed that if we further augment IPW by an outcome regression (see Chapter 7), the estimator could perform even worse though it has the theoretical "double robustness" property.

Of course, when inverse probability weights are unstable, they usually do not balance the covariates very well. This example motivated many empirical calibration weighting methods (e.g. Tan, 2010, Hainmueller, 2011, Zubizarreta, 2015, Chan et al., 2015) and also the approach described in the next Chapter that estimates the propensity score by minimizing loss functions tailored for the objective of covariate balance.

Chapter 6

Tailoring the Propensity Score Model

This Chapter is based on Zhao (2016).

6.1 Motivation

In the last Chapter, we introduced several methods to estimate causal effects from observational studies. As summarized in Table 5.1, some of these methods try to directly balance the raw covariates, while others resort to the propensity score. In general, propensity score is a more principled approach and easier to implement, because all the confounding information is summarized in a single number. However, this approach is not robust to model specification. Consider the following three main steps that most of the propensity-score based methods share:

Step 1 Estimate a propensity score model, most commonly by maximizing some scoring rule such as the Bernoulli likelihood. A scoring rule is a negative loss function and the two terms will be used interchangeably in this Chapter. The generalized linear model (McCullagh and Nelder, 1989) has been a workhorse in practice, but more sophisticated alternatives such as nonparametric regression (e.g. Hirano et al., 2003) and machine learning methods (e.g. McCaffrey et al., 2004, Lee et al., 2010, Wager and Athey, 2015) have also been suggested in the literature.

Step 2 Adjust for covariate imbalance by using the estimated propensity scores from Step 1. Numerous methods have been proposed, including: matching (e.g. Rosenbaum and Rubin, 1985, Abadie and Imbens, 2006), subclassification (e.g. Rosenbaum and Rubin, 1984), and inverse probability weighting (e.g. Robins et al., 1994, Hirano and Imbens, 2001). The reader is referred to Lunceford and Davidian (2004), Imbens (2004), Caliendo and Kopeinig (2008), Stuart (2010) for some comprehensive reviews.

Step 3 Choose a weighted average treatment effect as the estimand and estimate it by using the matches/strata/weights generated in Step 2. Report the point estimate, a confidence interval, evidence of sufficient covariate balance in Step 2 and sensitivity results if necessary.

A leading concern of the propensity-score based methods is that the eventual estimator in Step 3 can be highly sensitive to the outcome of Step 1—the estimated propensity score model (see e.g. Smith and Todd, 2005, Kang and Schafer, 2007). In fact, all the adjustment methods in Step 2 assume that the estimated propensity scores are very close to the truth. This generally requires a correctly specified model or an effective nonparametric regression. In practice, correct model specification is often unrealistic, and nonparametric regression, due to the curse of dimensionality, is a sensible choice only if the sample size is large and the covariates are few. To alleviate the concern of model misspecification, a commonly adopted strategy is to gradually increase the model complexity by forward stepwise regression (Imbens and Rubin, 2015, Section 13.3–13.4). The first two steps described above are usually repeated for several times until satisfactory covariate balance is achieved.

In the standard practice, maximum likelihood is used to fit the propensity score model in Step 1. However, maximum likelihood is suboptimal at balancing covariates. Figure 6.1 implements the aforementioned forward stepwise strategy with logistic regression and inverse probability weighting (IPW). More detail about this simulation example due to Kang and Schafer (2007) can be found in Section 6.7.1. In this Figure, covariate imbalance is measured by the standardized difference of each predictor between the two treatment groups (precise definition in Section 10.2). A widely used criterion is that a standardized difference above 10% is unacceptable (Normand et al., 2001, Austin and Stuart, 2015), which is the dashed line in Figure 6.1. The left panel of Figure 6.1 uses the Bernoulli likelihood to fit and select logistic regression models. The standardized difference paths are not monotonically decreasing and never achieve satisfactory level (10%) for more than half of the predictors. This certainly creates inconvenience for applied researchers, and more importantly, limits our understanding of the fundamental bias-variance trade-off in selecting a propensity score model. In contrast, the right panel of Figure 6.1 uses the covariate balancing scoring rule (CBSR) proposed in this Chapter and all 8 predictors are well balanced after 4 steps. As another highlighting feature, all active predictors (i.e. variables in the selected model) are exactly balanced using inverse probability weights derived by CBSR.

Why doesn't maximum likelihood always generate covariate balancing weights? Let's review the three-step procedure above, in which the user has the freedom to choose: in Step 1, a form of propensity score model (e.g. certain link function in the GLM) and a scoring rule to fit the model; in Step 2, a propensity-score based adjustment method; in Step 3, a weighted average treatment effect as the estimand. It is understandably tempting to fit a single propensity score model and use it to infer multiple estimands. Along this road, maximum likelihood most efficiently estimates the propensity scores. However, the propensity score model is a means to an end, not an end in itself. The most accurate (or even the true) propensity scores do not necessarily produce the best sample balance.

The central message of this Chapter is that we should tailor the scoring rule according to the estimand, since ultimately we are interested in the estimate in Step 3. CBSR views the three-step procedure as a whole and provides a systematic approach to obtain balancing weights. The CBSR-maximizing propensity scores in Step 1 are best paired with inverse probability weighting (IPW) in Step 2 for its algebraically tractability. As a side note, IPW is also quickly gaining popularity in the literature

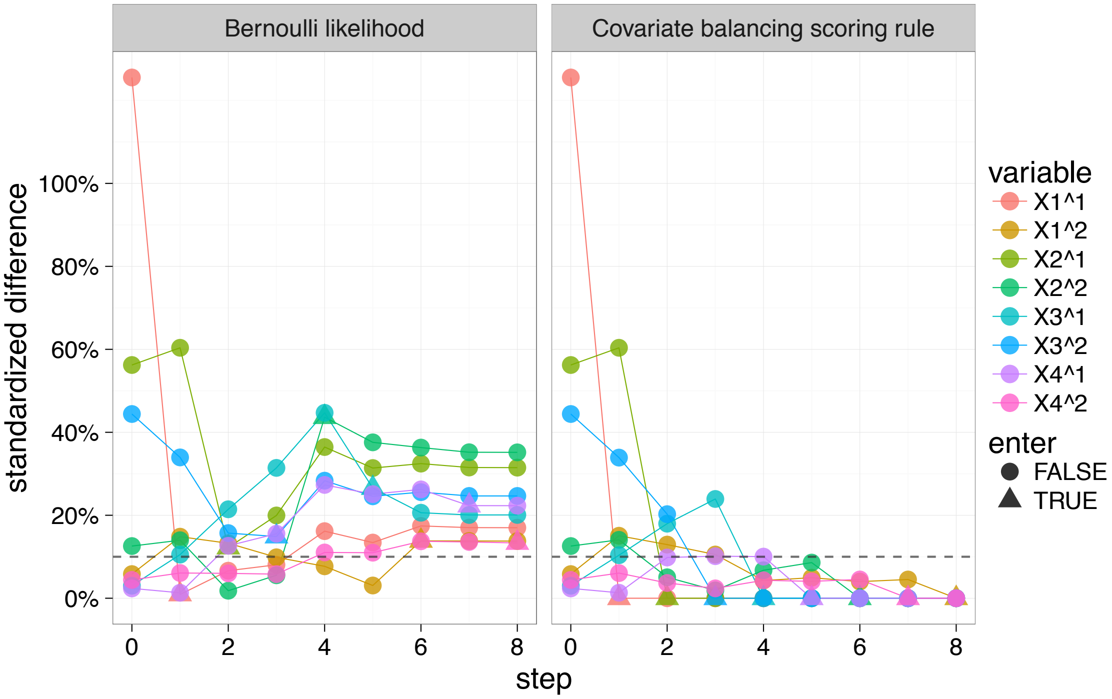


Figure 6.1: The covariate balancing scoring rule (CBSR) proposed in this Chapter is much better than Bernoulli likelihood at reducing covariate imbalance. Propensity score is modeled by logistic regression and fitted by CBSR or Bernoulli likelihood. Standardized difference is computed using inverse probability weighting (IPW) and pooled variance for the two treatment groups as in Rosenbaum and Rubin (1985), see equation (6.25) in Section 6.4.5. A standardized difference above 10% is viewed unacceptable by many practitioners. More detail of the forward stepwise regression and this simulation example can be found in Sections 6.4.1 and 6.7.1.

(Austin and Stuart, 2015) and is more more efficient than matching and subclassification. After obtaining the specific form of IPW from the estimand, the covariate balancing score rule can be uniquely determined from the link function of the GLM in Step 1.

6.2 Background on statistical decision theory

Propensity score estimation is a decision problem, though an unusual one. In a typical decision problem of making probabilistic forecast, the decision maker needs to pick an element as the prediction from $\mathcal{P}$, a convex class of probability measures on some general sample space Ω. For example, a weather forecaster needs to report the chance of rain tomorrow, so the sample space is $\Omega = \{\text{rain}, \text{no rain}\}$ and the prediction is a Bernoulli distribution. Propensity score is also a (conditional) probability measure, but the goal is to achieve satisfactory covariate balance rather than best predictive power. This marks a clear difference to the prediction problem. Nevertheless, statistical decision theory provides a general framework and effective tools to fit a covariate balancing propensity score model.

6.2.1 Proper scoring rules

Let's first review some useful concepts. At the core of statistical decision theory is the *scoring rule*, which can be any extended real-valued function $S : \mathcal{P} \times \Omega \to [-\infty, \infty]$ such that $S(P, \cdot)$ is $\mathcal{P}$-integrable for all $P \in \mathcal{P}$ (Gneiting and Raftery, 2007). If the decision is P and ω materializes, the decision maker's reward or utility is $S(P, \omega)$. An equivalent but more pessimistic terminology is *loss function*, which is just the negative scoring rule. These two terms will be used interchangeably in this Chapter.

If the outcome is probabilistic in nature and the actual probability distribution is Q, the expected score of forecasting P is

$$S(P, Q) = \int S(P, \omega) Q(d\omega).$$

To encourage honest decisions, we generally require the scoring rule S to be *proper* with respect to $\mathcal{P}$ that is defined by

$$S(Q, Q) \geq S(P, Q), \quad \forall P, Q \in \mathcal{P}. \tag{6.1}$$

The rule is called *strictly proper* with respect to $\mathcal{P}$ if (6.1) holds with equality if and only if $P = Q$. In estimation problems, strictly proper scoring rules provide appealing loss functions that can be tailored according to the scientific problem.

In observational studies, the sample space is commonly dichotomous $\Omega = \{0, 1\}$ (two treatment groups: 0 for control and 1 for treated), though there is no essential difficulty to extend the approach in this Chapter to $|\Omega| > 2$ (multiple treatments) or $\Omega \subset \mathbb{R}$ (continuous treatment). In the binary case, Savage (1971) showed that if $S(\cdot, 0)$ and $S(\cdot, 1)$ are real-valued except for possibly $S(0, 1) = \infty$ or $S(1, 0) = -\infty$, every proper scoring rule S can be characterized by

$$S(p, 1) = G(p) + (1 - p)G'(p) = \int (1 - p)G''(p)dp,$$
$$S(p, 0) = G(p) - pG'(p) = - \int pG''(p)dp,$$

where $G : [0, 1] \to \mathbb{R}$ is a convex function and $G'(p)$ is a subgradient of G at the point $p \in [0, 1]$. When G is second-order differentiable, an equivalent but useful representation is

$$\frac{\partial}{\partial p} S(p, t) = (t - p)G''(p), \ t = 0, 1. \tag{6.2}$$

Since the function G defines an equivalent class of scoring rule, we shall also call G a scoring rule.

A useful class of proper scoring rules is the following Beta family

$$G''_{\alpha,\beta}(p) = p^{\alpha-1}(1 - p)^{\beta-1}, \ -\infty < \alpha, \beta < \infty. \tag{6.3}$$

Notice that unlike the Beta family of distributions, the parameters here can be negative. These scoring rules were first introduced by Buja et al. (2005) to approximate the weighted misclassification loss by taking the limit $\alpha, \beta \to \infty$ and $\alpha/\beta \to c$. For example, if $c = 1$, the score $G_{\alpha,\beta}$ converges to the zero-one misclassification loss. Many important scoring rules belong to this family. For example, the Bernoulli log-likelihood function or the logarithmic score $S(p, t) = t \log p + (1 - t) \log(1 - p)$ corresponds to $\alpha = \beta = 0$, and the Brier score (or equivalently the squared error loss when flipping the sign) $S(p, t) = -(t - p)^2$ corresponds to $\alpha = \beta = 1$. For our purpose of estimating propensity score, it will be shown later that the subfamily $-1 \leq \alpha, \beta \leq 0$ is especially useful.

6.2.2 Propensity score estimation by maximizing score

Given i.i.d. observations $(X_i, T_i) \in \mathbb{R}^d \times \{0, 1\},\ i = 1, 2, \ldots, n$ where T_i is the binary treatment assignment and X_i is a vector of d pre-treatment covariates, the goal is to fit a model for the propensity score $p(X) = \mathrm{P}(T|X)$. Suppose we are willing to use a parametric model that belongs to the family $\mathcal{P} = \{p_\theta(X) : \theta \in \Theta\}$. Given a strictly proper scoring rule S, the goodness-of-fit of θ can be measured by the average score

$$\mathcal{S}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} S(p_\theta(X_i), T_i),$$

The *optimum score estimator* is obtained by the unique maximizer of the average score:

$$\hat{\theta}_n = \arg\max_{\theta} \mathcal{S}_n(\theta) \tag{6.4}$$

Notice that the affine transformation $S(p, t) \mapsto aS(p, t) + b(t)$ for any $a > 0$ and $-\infty < b(t) < \infty$ results in the same estimator $\hat{\theta}_n$, so we shall not differentiate between these equivalent scoring rules and use a single function $S(p, t)$ to represent all equivalent ones.

In view of the population identity

$$\mathrm{E}\left[\mathcal{S}_n(\theta)\right] = \mathrm{E}_{X,T}[S(p_\theta(X), T)] = \mathrm{E}_X[\mathrm{E}_{T|X}[S(p_\theta(X), T)]],$$

the optimum score estimator is Fisher-consistent. Fisher-consistency means that the true value of the parameter θ would be obtained if the true propensity score is $p(x) = p_\theta(x)$ and the estimator were calculated using the entire population rather than a sample. In many cases, including the problem considered in this Chapter, this property also leads to asymptotic consistency: $\hat{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$.

This Chapter focuses on using the generalized linear models (McCullagh and Nelder, 1989)

$$p_\theta(X) = l^{-1}(f_\theta(X)) = l^{-1}(\theta^T \phi(X)) \tag{6.5}$$

to model the propensity score. Here l is the *link function*, $f_\theta(X)$ is the *canonical parameter* which is modeled by a linear combination of the m-dimensional *predictors*

$\phi(X) = (\varphi_1(X), \ldots, \varphi_m(X))^T$. The covariate balancing scoring rule derived in this Chapter depends on the link function l. The most common choice is the logistic link:

$$l(p) = \log \frac{p}{1-p}, \; l^{-1}(f) = \frac{e^f}{1+e^f}. \tag{6.6}$$

This will be the choice for all the numerical examples in this Chapter.

When S is differentiable and assuming exchangeability of taking expectation and derivative, the maximizer of $\mathrm{E}[\mathcal{S}_n(\theta)]$, which is indeed θ if $p(x) = p_\theta(x)$ by Fisher-consistency, is characterized by the following estimating equations

$$\nabla_\theta \, \mathrm{E}[\mathcal{S}_n(\theta)] = \mathrm{E}[\nabla_\theta \, \mathcal{S}_n(\theta)] = \mathrm{E}_{X,T}[\nabla_\theta \, S(l^{-1}(\theta^T \phi(X)), T)] = 0. \tag{6.7}$$

Using the representation (6.2) and the inverse function theorem, we have

$$\nabla_\theta \, S(l^{-1}(\theta^T \phi(X)), T) = (T - p_\theta(X)) G''(p_\theta(X)) \frac{1}{l'(p_\theta(X))} \cdot \phi(X).$$

Therefore the condition (6.7) can be written as

$$\mathrm{E}_{X,T} \left\{ \frac{G''(p_\theta(X))}{l'(p_\theta(X))} \left[T(1 - p_\theta(X)) - (1-T) p_\theta(X) \right] \cdot \phi(X) \right\} = 0. \tag{6.8}$$

The optimum score estimator, $\hat{\theta}_n$, can be determined from (6.8) by taking the expectation over the empirical distribution of (X, T), provided that S is strictly proper so the solution to (6.8) is unique.

6.3 Covariate balancing scoring rules

The covariate balancing scoring rules (CBSR) are motivated by the estimating equations (6.8), which can be interpreted as weighted differences of $\phi(X)$ between the treatment ($T = 1$) and the control ($T = 0$). The weights are given by, for $t = 0, 1$,

$$w(x, t) = \frac{G''(p(x))}{l'(p(x))} [t(1 - p(x)) + (1-t) p(x)]. \tag{6.9}$$

Equation (6.8) can now be rewritten as stochastic balance of the predictors

$$\mathrm{E}[(T-(1-T))w(X,T)\cdot\phi(X)]=0, \tag{6.10}$$

The question is: Are these weights meaningful? In other words, do they correspond to some form of inverse probability weighting (IPW)?

6.3.1 Covariate balancing scoring rules

The answer to this question is, of course, positive. In short, every convex function G defines a weighted average treatment effect via (6.9). To see this we need to introduce some notation. Following the Neyman-Rubin causal model, let $Y(t),\ t=0,1$ be the potential outcomes and $Y=TY(1)+(1-T)Y(0)$ be the observed outcome. Throughout this Chapter we assume the strong ignorability (2.2) of treatment assignment (Rosenbaum and Rubin, 1983), so there is no hidden bias:

Assumption 6.1. *$T \perp\!\!\!\perp (Y(0),Y(1))|X$.*

First, we define a population parameter by replacing $\phi(X)$ in (6.10) with the outcome Y

$$\tau_w=\mathrm{E}_{X,T,Y}\{(T-(1-T))w(X,T)Y\},$$

Under Assumption 6.1, τ_w is indeed an (unnormalized) weighted average treatment effect

$$\tau_w=\mathrm{E}_{X,Y}\left[w(X)(Y(1)-Y(0))\right], \tag{6.11}$$

where

$$w(X)=p(X)w(X,1)=(1-p(X))w(X,0)=\frac{G''(p(X))\,p(X)\,(1-p(X))}{l'(p(X))}.$$

In practice, it is usually more meaningful to consider the normalized version of τ_w:

$$\tau_w^*=\tau_w\Bigg/\mathrm{E}_{X,T,Y}\left[\frac{G''(p(X))\,p(X)\,(1-p(X))}{l'(p(X))}\right]. \tag{6.12}$$

α	β	estimand	$S(p,1)$	$S(p,0)$
-1	-1	$\tau = \tau^* = \mathrm{E}[Y(1) - Y(0)]$	$\log \frac{p}{1-p} - \frac{1}{p}$	$\log \frac{1-p}{p} - \frac{1}{1-p}$
-1	0	$\tau^* = \mathrm{E}[Y(1) - Y(0) \vert T = 0]$	$-\frac{1}{p}$	$\log \frac{1-p}{p}$
0	-1	$\tau^* = \mathrm{E}[Y(1) - Y(0) \vert T = 1]$	$\log \frac{p}{1-p}$	$-\frac{1}{1-p}$
0	0	$\tau = \mathrm{E}[p(X)(1-p(X)) \cdot (Y(1) - Y(0))]$	$\log p$	$\log(1-p)$

Table 6.1: Estimands and scoring rules in the Beta family.

The question now becomes: is τ_w^* an interesting estimand in observational studies? The answer to this question is, again, positive. Consider the following Beta family of weighted average treatment effects

$$\tau_{\alpha,\beta} = \mathrm{E}[p(X)^{\alpha+1}(1-p(X))^{\beta+1}(Y(1) - Y(0))], \ -1 \leq \alpha, \beta \leq 0. \tag{6.13}$$

Several important estimands belong to this family, including the average treatment effect (ATE), the average treatment effect on the untreated (ATUT), the average treatment effect on the treated (ATT), and the optimally weighted average treatment effect under homoscedasticity (Crump et al., 2006). See the third column of Table 6.1 for the definitions of these estimands.

The next Proposition shows an exact correspondence between the Beta family of estimands (6.13) and the Beta family of scoring rules (6.3).

Proposition 6.1. *Under Assumption 6.1, if $G = G_{\alpha,\beta}$ and l is the logistic link function, then $\tau_w = \tau_{\alpha,\beta}$.*

Proof. Use equations (6.3), (6.6), (6.9), (6.11) and (6.13). □

Therefore, some of the most important estimands in observational studies can be defined by (6.12). Proposition 6.1 also suggests a general strategy to estimate average causal effect:

1. Pick a weighted average treatment effect $\tau = \tau_{\alpha,\beta}$ from the Beta family (6.13) as the estimand.

2. Compute its corresponding scoring rule using (6.9) or find it from Table 6.1 below.

3. Using the scoring rule, fit a logistic regression $\hat{p}(X) = l^{-1}(\hat{\theta}^T\phi(X))$ for the propensity score.

4. Estimate τ and its normalized version τ^* defined in (6.12) by

$$\hat{\tau} = \sum_{i:\,T_i=1} \hat{w}_i Y_i - \sum_{i:\,T_i=0} \hat{w}_i Y_i \text{ and } \hat{\tau}^* = \sum_{i:\,T_i=1} \hat{w}_i^* Y_i - \sum_{i:\,T_i=0} \hat{w}_i^* Y_i, \tag{6.14}$$

where

$$\hat{w}_i = p_{\hat{\theta}}(X_i)^{\alpha}(1 - p_{\hat{\theta}}(X_i))^{\beta}[T_i(1 - p_{\hat{\theta}}(X_i)) + (1 - T_i)p_{\hat{\theta}}(X_i)] \tag{6.15}$$

and the normalized weights are $\hat{w}_i^* = \hat{w}_i / \sum_{j:\,T_j=T_i} \hat{w}_j$, $i = 1, \ldots, n$.

A main advantage of this approach is that the weights automatically balance the predictors $\phi(X)$ in the logistic regression, as indicated by the next theorem.

Theorem 6.1. *Given a scoring rule $S_{\alpha,\beta}$ in the Beta family and a logistic regression model $p_\theta(X) = l^{-1}(\theta^T\phi(X))$, suppose $\hat{\theta}$ is obtained by maximizing the average score as in (6.4). Then the weights $\hat{w}_i$, $i = 1, \ldots, n$, exactly balance the sample predictors*

$$\sum_{i:\,T_i=1} \hat{w}_i \phi(X_i) = \sum_{i:\,T_i=0} \hat{w}_i \phi(X_i). \tag{6.16}$$

Furthermore, if the predictors include an intercept term (i.e. 1 is in the linear span of $\phi(X)$), then $\hat{w}^*$ also satisfies (6.16).

Proof. This theorem is a simple corollary of the estimating equations (6.10). □

Because of Theorem 6.1, $G_{\alpha,\beta}$ or the resulting $S_{\alpha,\beta}$ will be called the *covariate balancing scoring rule* (CBSR) with respect to the estimand $\tau_{\alpha,\beta}$ and the logistic link function.

6.3.2 A closer look at the Beta family

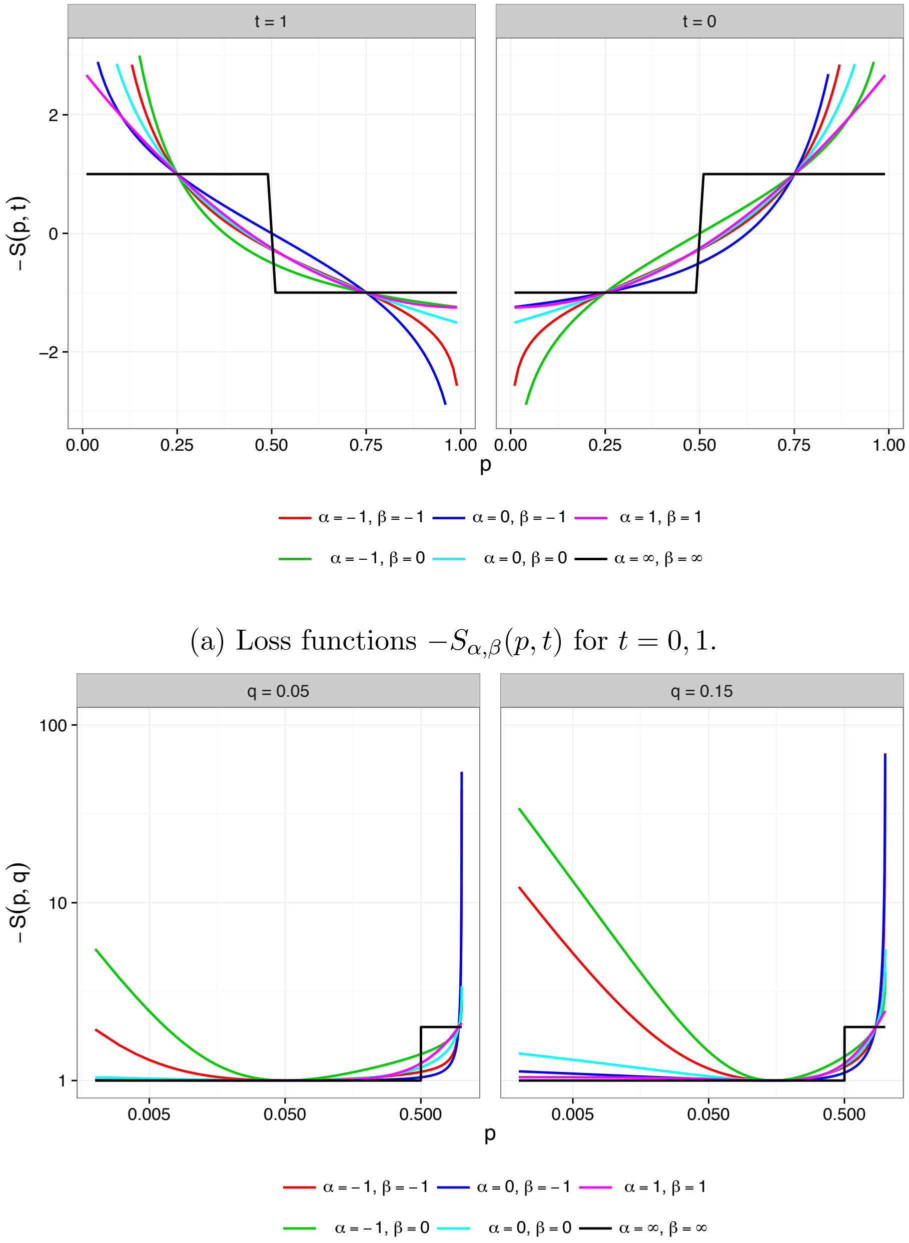
One may wonder why the estimands in (6.13) are restricted to the subfamily $-1 \leq \alpha, \beta \leq 1$. There are at least two reasons. First, as mentioned earlier, this subfamily already contains most of the important estimands that are meaningful to observational studies (see Table 6.1). Second, as shown in Proposition 6.2 below, this is the only region such that the maximum score problem (6.4) is convex when $p_\theta(X)$ is modeled by logistic regression. Therefore the optimization problem (6.4) has no local maximum and can be solved efficiently (e.g. by Newton's method).

Proposition 6.2. *For the Beta family of scoring rules defined in equations* (6.2) *and* (6.3) *and the logistic link function* $l^{-1}(f) = e^f/(1+e^f)$*, the score functions* $S(l^{-1}(f), 0)$ *and* $S(l^{-1}(f), 1)$ *are both concave functions of* $f \in \mathbb{R}$ *if and only if* $-1 \leq \alpha, \beta \leq 1$*. Moreover, if* $(\alpha, \beta) \neq (-1, 0)$*,* $S(l^{-1}(f), 0)$ *is strongly concave; if* $(\alpha, \beta) \neq (0, -1)$*,* $S(l^{-1}(f), 1)$ *is strongly concave.*

Proof. See Section 6.8.1. □

Figure 6.2 plots the scoring rules $S_{\alpha,\beta}$ for some combinations of α and β. The top panels show the score function $S(p, 0)$ and $S(p, 1)$ for $0 < p < 1$, which are normalized so that $S(1/4, 1) = S(3/4, 0) = -1$ and $S(1/4, 0) = S(3/4, 1) = 1$. By a change of variable, one can show $S_{\alpha,\beta}(p, 1) = S_{\beta,\alpha}(1-p, 0)$. This is the reason that the two subplots in Figure 6.2a are essentially reflections of each other. The bottom panels show the induced scoring rule $S(p, q)$ defined by section 6.2.1 or more specifically $S(p, q) = qS(p, 1) + (1-q)S(p, 0)$ at two different values of $q = 0.05, 0.15$. For aesthetic purposes, the scoring rules in Figure 6.2b are normalized such that $-S(p, q) = 1$ and $-S(p, 1-q) = 2$.

Figure 6.2 shows that the scoring rules $S_{\alpha,\beta}$, when $-1 \leq \alpha, \beta \leq 0$, are highly sensitive to small differences of small probabilities. For example, in Figure 6.2a the loss function $-S_{\alpha,\beta}(p, 1)$ is unbounded above when $\alpha, \beta \in \{-1, 0\}$, hence a small change of p near 0 may have a big impact on the score. In Figure 6.2b, the averaged scoring rules $S_{\alpha,\beta}(p, q)$, when $(\alpha, \beta) = (-1, -1)$ or $(-1, 0)$, are also unbounded near $p = 0$. Due to this reason, Selten (1998, Section 2.6) argued that these scoring rules are inappropriate for probability forecast problems.



(a) Loss functions $-S_{\alpha,\beta}(p, t)$ for $t = 0, 1$.

(b) Loss functions $-S_{\alpha,\beta}(p, q)$ for $q = 0.05$ and 0.15.

Figure 6.2: Graphical illustration of the Beta-family of scoring rules defined in (6.3).

On the contrary, the unboundedness is actually a desirable feature for propensity score estimation, as the goal is to avoid extreme probabilities. Consider the standard inverse probability weights (IPW)

$$\hat{w}_i = \begin{cases} \hat{p}_i^{-1} & \text{if } T_i = 1, \\ (1-\hat{p}_i)^{-1} & \text{if } T_i = 0, \end{cases} \tag{6.17}$$

where $\hat{p}_i = p_{\hat{\theta}}(X_i)$ is the estimated propensity score for the i-th data point. This corresponds to $\alpha = \beta = -1$ in the Beta family and estimates ATE. Several previous articles (e.g. Robins and Wang, 2000, Kang and Schafer, 2007, Robins et al., 2007) have pointed out the hazards of using large inverse probability weights. For example, if the true propensity score is $p(X_i) = q = 0.05$ and it happens that $T_i = 1$, we would want $\hat{p}_i$ not too close to 0 so $\hat{w}_i$ is not too large. Conversely, we also want $\hat{p}_i$ not too close to 1, so in the more likely event that $T_i = 0$ the weight $\hat{w}_i$ is not too large either. In an *ad hoc* attempt to mitigate this issue, Lee et al. (2011) studied weight truncation (e.g. truncate the largest 10% weights). They found that the truncation can reduce the standard error of the estimator $\hat{\tau}$ but also increases the bias.

The covariate balancing scoring rules provide a more systematic approach to avoid large weights. For example, the scoring rule $S_{-1,-1}$ precisely penalizes large inverse probability weights as $-S_{-1,-1}(p,q)$ is unbounded above when p is near 0 or 1 (see the left plot in Figure 6.2b). Similarly, when estimating the ATUT $\tau_{-1,0}$, the weighting scheme would put $\hat{w}_i \propto (1-\hat{p}_i)/\hat{p}_i$ if $T_i = 1$ and $\hat{w}_i \propto 1$ if $T_i = 0$. Therefore we would like $\hat{p}_i$ to be not close to 0, but it is acceptable if $\hat{p}_i$ is close to 1. As shown in in Figure 6.2b, the curve $-S_{-1,0}(p,q) = q/p + (1-q)\log(p/(1-p))$ precisely encourages this behavior, as it is unbounded above when p is near 0 and grows slowly when p is near 1.

6.4 Adaptive strategies

So far we have only considered a fixed GLM to model the propensity score. This Section discusses some adaptive extensions motivated by popular machine learning algorithms. In order to achieve the best predictive performance, most machine learning methods prespecify a loss function to train the model. For the purpose of obtaining covariate balancing weights, we only need to replace the loss function by the covariate balancing scoring rule (CBSR) introduced in this Chapter. This is indeed a major advantage of using scoring rules instead of estimating equations.

6.4.1 Forward Stepwise

Let's start with the forward stepwise regression which is already widely used in observational studies (Imbens and Rubin, 2015). The notation $\phi(x) = (\phi_1(x), \ldots, \phi_m(x))$ is used to indicate all the potential linear predictors. This entire Section allows $m > n$, but it is not necessary to include all the predictors in the model.

Algorithm 6.1 Forward stepwise regression for propensity score

```
Input data: $(T_i, X_i),\ i = 1, \ldots, n$.
Input arguments: predictors $\{\phi_1(x), \ldots, \phi_m(x)\}$, link function $l(\cdot)$, proper scoring rule $S(p, t)$.
Notation: $\mathcal{F}_{\mathcal{A}} = \mathrm{span}(\{\phi_k(x) | k \in \mathcal{A}\})$.
Algorithm:
Initialize active set $\mathcal{A} = \emptyset$.
for $j = 1, \ldots, m$ do
    Compute $S_{jk} = \max_{f \in \mathcal{F}_{\mathcal{A} \cup \{k\}}} \sum_{i=1}^{n} S(l^{-1}(f(X_i)), T_i)$ for $k \in \mathcal{A}^c$.
    Update $\mathcal{A}_k = \mathcal{A}_k \cup \{\arg\max_k S_{jk}\}$.
end for
Output:
$\mathcal{A}^*$ from $\mathcal{A}_k,\ k = 1, \ldots, m$ that optimizes some criterion (e.g. AIC, BIC, least covariate imbalance).
$f^* = \arg\max_{f \in \mathcal{F}_{\mathcal{A}^*}} \sum_{i=1}^{n} S(l^{-1}(f(X_i)), T_i)$.
```

In Algorithm 6.1, the predictors are added one by one in a forward stepwise regression. After choosing a scoring rule, the algorithm in each step fits a GLM using all the selected predictors and each unselected predictor. The unselected predictor that increases the score $\mathcal{S}_n$ the most is added to the active set. This procedure is repeated until no new predictor can be added or the current score $\mathcal{S}_n$ is already ∞. Figure 6.1 demonstrates this adaptive algorithm with a simulation example described in Section 6.7.1. There is no need to reiterate that CBSR is much better at reducing covariate imbalance than Bernoulli likelihood.

6.4.2 Regularized Regression

Another widely-used adaptive method is the following regularized solution of the GLM (6.5):

$$\hat{\theta}_\lambda = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} S(p_\theta(X_i), T_i) - \lambda J(\theta), \tag{6.18}$$

where $J(\cdot)$ is a regularization term that penalizes large θ (complicated model) and λ controls the degree of regularization. This estimator reduces to the optimum score estimator (6.4) when $\lambda = 0$. For simplicity, this Chapter only considers penalty of the form

$$J(\theta) = \frac{1}{a} \sum_{k=1}^{m} |\theta_k|^a \text{ for some } a \geq 1. \tag{6.19}$$

Some typical choices are the l_1 norm $J(\theta) = \|\theta\|_1$ (lasso) and the squared l_2 norm $J(\theta) = \|\theta\|_2^2$ (ridge regression).

An important advantage of the regularized regression (6.18) is that it allows high dimensional predictors $\phi(X)$. This is useful to propensity score estimation for at least three reasons:

1. The pre-treatment covariates X can be high dimensional, especially if we wish to follow Rubin (2009)'s advice that "we should strive to be as conditional as is practically possible".

2. Even if X is relatively low dimensional, we may still want to use a high dimensional $\phi(X)$ to essentially build a nonparametric propensity score model.

3. The Beta family of scoring rules (6.3) with $-1 \leq \alpha, \beta \leq 0$ are unbounded above, so $\sup_\theta \mathcal{S}_n(\theta)$ can easily be infinity if ϕ is high dimensional, making the optimum score problem (6.4) infeasible. The Bernoulli likelihood ($\alpha = \beta = 0$) also suffers from this. In this case, it is necessary to add some regularization as in (6.18) to obtain any propensity score model.

6.4.3 Kernel method

The predictors $\phi(X)$ can even be infinite dimensional via a popular nonparametric regression method in machine learning (Wahba, 1990, Hofmann et al., 2008, Hastie et al., 2009). This method models the propensity score $p(x)$ by $l^{-1}(f(x))$ with f in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$, where the kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ describes the similarity between two observations of pre-treatment covariates. Suppose that K has an eigen-expansion

$$K(x, x') = \sum_{k=1}^{\infty} c_k \phi_k(x) \phi_k(x')$$

with $c_k \geq 0$, $\sum_{k=1}^{\infty} c_k^2 < \infty$. Elements of $\mathcal{H}_K$ have an expansion in terms of these eigen-functions,

$$f(x) = \sum_{k=1}^{\infty} \theta_k \phi_k(x).$$

The standard generalized linear model (6.5) corresponds to a finite-dimensional linear reproducing kernel $K(x, x') = \sum_{k=1}^{m} \phi_k(x) \phi_k(x')$, but the eigen-functions (i.e. predictors) $\{\phi_k\}_{k=1}^{\infty}$ can easily be infinite-dimensional. Since the RKHS is usually a very rich function space, it is common to the regularize the score as in (6.18) with penalty $J(\theta) = \|f\|_{\mathcal{H}_K}^2 = \sum_{k=1}^{\infty} \theta_k^2 / c_k$.

Although RKHS incorporates potentially infinite-dimensional predictors, the numerical problem (6.18) is computationally feasible via the so-called “kernel trick”. The representer theorem (c.f. Wahba, 1990) states that the solution to (6.18) is indeed finite-dimensional and has the form $\hat{f}(x) = \sum_{i=1}^{n} \hat{\gamma}_i K(x, X_i)$. Consequently, the optimization problem (6.18) can be solved with the n-dimension parameter vector γ.

Kernels are not newcomers to the toolbox for observational studies. Most of the previous literature (e.g. Heckman et al., 1997, 1998) uses kernel as a smoothing technique for propensity score estimation (i.e. a generalization of the nearest neighbor matching) rather than generating a RKHS, but the kernel function K is the same. The tuning parameters in RKHS are the kernel bandwidths and the amount of smoothness penalty. Sensitivity analysis may be carried out with little extra effort by varying over different kernel forms and bandwidths.

The RKHS approach has another practical benefit: the modeling process is free from guessing model specifications. The user only needs to choose a kernel that measures the closeness of two units based on pre-treatment covariates X. It is arguably much easier for a field expert to answer questions like "is patient A or patient B more similar to patient C based on their age and education?" than to speculate and make sense of a model like "the logit of the propensity score is linear in age and years of education".

6.4.4 Gradient boosting

The gradient boosting machine of Friedman (2001) and Mason et al. (1999) is one of the best performing supervised learning method (Caruana and Niculescu-Mizil, 2006). Gradient boosting works particularly well when the model is intrinsically nonlinear, an appealing feature for researchers concerned with model misspecification.

The idea behind gradient boosting is quite simple. We first start with the population-level estimation. Let $\mathcal{P}$ denote the joint distribution of (X, T) and $\mathcal{F}$ be a normed model space of the canonical parameter $f(x) = l(p(x))$. If $\hat{f}(x)$ is the current guess, the next guess is given by the steepest (gradient) ascent

$$\hat{f}^{\text{new}}(x) = \hat{f}(x) + \hat{\eta}\hat{g}(x) \tag{6.20}$$

where

$$\hat{g} = \operatorname*{arg\,max}_{\|g\|=1} \frac{\partial}{\partial g} S_{\alpha,\beta}(\hat{f}, \mathcal{P}) = \lim_{\epsilon \to 0} \frac{\mathrm{d}}{\mathrm{d}\epsilon} S_{\alpha,\beta}(\hat{f} + \epsilon g, \mathcal{P}) = \operatorname*{arg\,max}_{\|g\|=1} \mathrm{E}_{\mathcal{P}}[(2T-1)\hat{w}(X,T)g(X)], \text{ and} \tag{6.21}$$

$$\hat{\eta} = \operatorname*{arg\,max}_{\eta \geq 0} S_{\alpha,\beta}(\hat{f} + \eta\hat{g}, \mathcal{P}). \tag{6.22}$$

Here $\hat{w}(X, T)$ is computed from $\hat{f}$ via (6.15). Intuitively, at each step we find the current most unbalanced covariate function $\hat{g}$ and move as far as we can towards that direction.

With finite sample, suppose we observe i.i.d. data (X_i, T_i), $i = 1, \ldots, n$. $T_i \in \{-1, 1\}$. Let $\mathcal{P}_n$ be its empirical distribution. Now we wish to maximize $S_{\alpha,\beta}(\mathcal{P}_n, f)$. However, this is not an easy task as it looks, because

$$\mathrm{E}_{\mathcal{P}_n}[(2T-1)\hat{w}(X,T)g(X)] = \sum_{i=1}^{n}(2T_i - 1)w_{\alpha,\beta}(T_i, \hat{f}(X_i))g(X_i)$$

only depends on the value of g at no more than n points. The definition (6.21) does not quite make sense because the sample imbalance can be infinity. For example, if all the X_i, $i = 1, \ldots, n$ are distinct, we can take

$$g(x) = \begin{cases} c(2T_i - 1) & x = X_i \text{ for some } i, \\ 0 & x \neq X_i, \ \forall\, i, \end{cases} \tag{6.23}$$

and let $c \to \infty$. The norm of g is always equal to 0, but its sample imbalance can be arbitrarily large.

Of course we wouldn't model the propensity score by some function like (6.23). To solve this issue, we need to constrain the model space $\mathcal{F}$ so that the functions f and g are "nice". The most common model space for gradient boosting is the decision tree. To be more precise, the approximate functional gradient $\hat{g}$ is obtained by maximizing (6.21) in the space of decision trees, and the eventual estimator $\hat{f}$ is the sum of many trees.

Notice that the finite sample solution to (6.21) admits a simple solution. Let $\mathcal{F}_{k\text{-tree}}$ be the space of all trees with depth at most k. Then

$$\begin{aligned} \hat{g} &= \mathop{\arg\max}_{\|g\|=1,\, g\in\mathcal{F}_{k\text{-tree}}} \sum_{i=1}^{n}(2T_i - 1)w_{\alpha,\beta}(T_i, \hat{f}(X_i))g(X_i) \\ &\propto \mathop{\arg\max}_{\|\{g(X_i)\}_{i=1}^{n}\|=1,\, g\in\mathcal{F}_{k\text{-tree}}} \sum_{i=1}^{n}(2T_i - 1)w_{\alpha,\beta}(T_i, \hat{f}(X_i))g(X_i) \\ &\propto \mathop{\arg\min}_{g\in\mathcal{F}_{k\text{-tree}}} \sum_{i=1}^{n}[(2T_i - 1)w_{\alpha,\beta}(T_i, \hat{f}(X_i)) - g(X_i)]^2 \end{aligned} \tag{6.24}$$

Algorithm 6.2 BalanceBoost for propensity score

```
**Input data:** $(T_i, X_i),\ i = 1, \ldots, n$.
**Input arguments:** $-1 \leq \alpha, \beta \leq 0$, tree depth $k$, shrinkage rate $\nu \leq 1$, subsampling rate $\gamma \leq 1$, maximum number of steps $N$.
**Algorithm:**
Initialize $\hat{f}(x) \equiv c$ that maximizes the average score $S_{\alpha,\beta}(f(x), \mathcal{P}_n)$.
**while** haven't reached $N$ steps **do**
    Compute the current weights $\hat{w}$ corresponding to $\hat{f}(x)$ by equation (6.15).
    Randomly generate a subsample $\mathcal{I}$ that $|\mathcal{I}| = \lfloor \gamma n \rfloor$.
    Grow a depth-$k$ regression tree $\hat{g}$ according to equation (6.24) by treating signed weights $\{T_i \hat{w}_i\}_{i \in \mathcal{I}}$ as responses.
    Choose $\eta$ to maximize $S_{\alpha,\beta}(\hat{f}(x) + \eta \hat{g}(x), \mathcal{P}_n)$.
    Choose $c$ to maximize $S_{\alpha,\beta}(\hat{f}(x) + c + \nu \eta \hat{g}(x), \mathcal{P}_n)$.
    Update $\hat{f}^{\mathrm{new}}(x) = \hat{f}(x) + c + \nu \eta \hat{g}(x)$.
**end while**
**return** $\hat{f}$.
```

In other words, we need to grow a depth-k tree that it's closest to the signed weights (i.e. the gradient) in squared error loss. This is the standard problem of regression tree.

Algorithm 6.2 provides the pseudo-code for the boosting procedure described above. Besides the regression tree heuristic, Algorithm 6.2 contains three other tweaks that are useful in practice:

1. Each gradient step is shrinked by a factor $\nu \leq 1$. This avoids overfitting the model and usually greatly improves the performance of gradient boosting. The shrinkage factor ν is usually chosen to be very small, e.g. $\nu < 0.01$ (Ridgeway et al., 2006, Hastie et al., 2009).

2. Each tree $\hat{g}$ is built by using a subsample of the observed data. The subsampling rate $\gamma \leq 1$. When $\gamma = 1$, no subsampling is used.

3. The estimated weights are most useful if they balance the constant function. Due to this reason, each gradient step is followed by a update of the intercept in Algorithm 6.2.

The boosting algorithm is closely related to forward stagewise regression and l_1-regularized regression (Friedman et al., 2000). In fact, BalanceBoost can be viewed as a path algorithm to minimize the largest imbalance for functions in $\mathcal{F}_{k\text{-tree}}$. When $k = 1$, the largest imbalance is in fact the largest Kolmogorov-Smirnov test statistics of any pretreatment covariate. This observation is illustrated in Figure 6.3. This figure is similar to the residual-correlation paths in forward stagewise regression (Hastie et al., 2009, Figure 3.14).

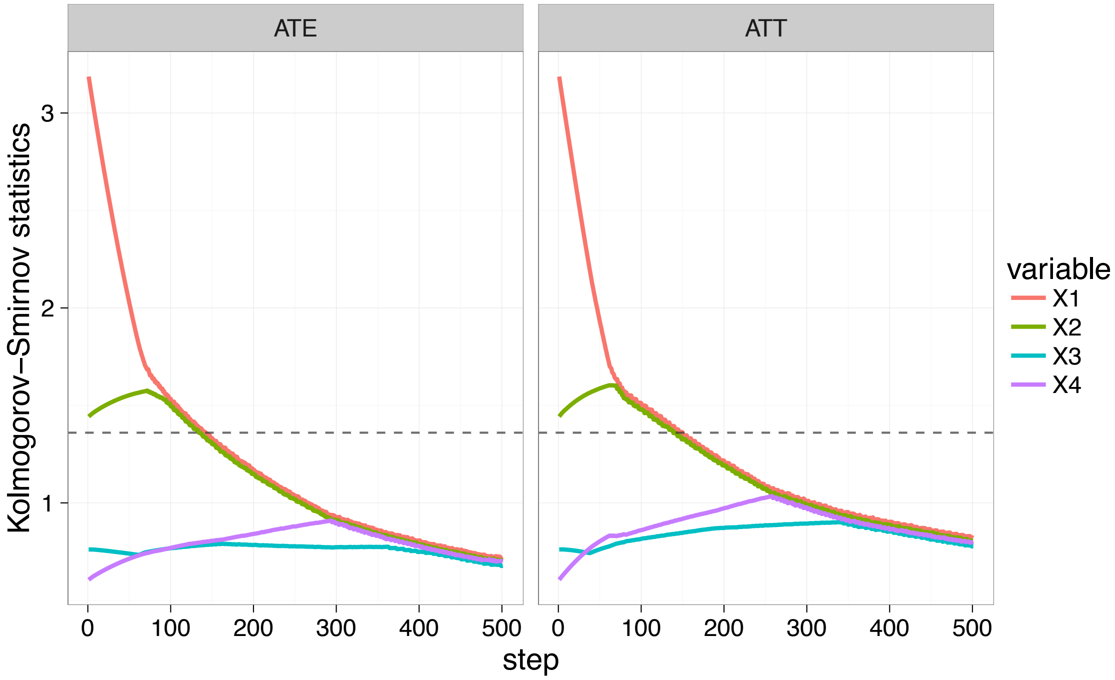


Figure 6.3: Kolmogorov-Smirnov statistics along the BalanceBoost path ($k = 1$, $\nu = 0.01$, $\eta = 1$). Dashed line is the 95% asymptotic rejection threshold.

6.4.5 Model selection and inference

After a series of propensity score models are fitted by forward stepwise regression or regularized regression as described earlier, the remaining question is to select one model for further statistical inference. This is a standard task in propensity-score based approaches. The selected model should best balance all the pretreatment covariates. One measurement of covariate imbalance is the absolute standardized difference (Rosenbaum and Rubin, 1985). For the estimated weights $\hat{w}$ and each predictor ϕ_k, $k = 1, \ldots, m$, it is defined as

$$d_k = \frac{\left|(1/n_1)\sum_{i:T_i=1}\hat{w}_i\phi_k(X_i) - (1/n_0)\sum_{j:T_j=0}\hat{w}_j\phi_k(X_j)\right|}{s_w}, \tag{6.25}$$

where s_w^2 is the sample variance of the numerator in (6.25). We can also use the t-test based on (6.25) to verify the means are not significantly different. Another widely used criterion is the nonparametric Kolmogorov-Smirnov test. The reader is referred to the review articles by Caliendo and Kopeinig (2008), Austin and Stuart (2015) for more practical guidance. In the simulation example in Section 6.7.1, we choose the propensity score model that has the smallest number of significant two-sample t-tests. When the covariate balancing scoring rule is used, the selected model is usually close to the end of the path.

Given a propensity score model, the weighted average treatment effect τ or its normalized τ^* can be estimated by inverse probability weighting described in (6.14) and (6.15). To obtain a confidence interval for τ or τ^*, we adopt a general method for estimating sampling variances in Imbens and Rubin (2015, Chapter 19). Let $\mathrm{Var}(Y_i) = \sigma_i^2$. Conditioning on the covariates X and the estimated weights $\hat{w}$, the sampling variance of $\hat{\tau}$ is given by

$$\mathrm{Var}(\hat{\tau}|X, w) = \sum_{i=1}^{n} \hat{w}_i^2\sigma_i^2. \tag{6.26}$$

Imbens and Rubin (2015, Section 19.6) described several ways to estimate σ_i^2 for all units. In the numerical examples in Section 6.7, we assume additive homoskedastic noise $\sigma_i^2 = \sigma^2$ and use a pilot outcome regression to estimate the noise variance σ^2.

6.5 Theoretical aspects

This Section discusses the following four theoretical aspects about CBSR. A first-time reader more interested in the empirical performance can skip the next two Sections and go to the numerical examples in Section 6.7.

1. With increasingly complex propensity score model as the sample size grows, any strongly concave proper scoring rule can provide semiparametrically efficient estimate of the weighted average treatment effect (Section 6.5.1).

2. Even if the propensity score model is misspecified, CBSR can still reduce the bias and the variance of $\hat{\tau}$ due to the covariate balancing weights (Section 6.5.2).

3. The Lagrangian dual of the CBSR maximization problem is an entropy maximization problem with covariate balancing constraints. This observation connects IPW estimators with calibration estimators in survey sampling (Section 6.5.3).

4. The Lagrangian duality also allows us to study the bias-variance tradeoff in selecting propensity score models (Section 6.5.4).

6.5.1 Global efficiency by sieve regression

If a statistician is asked about why maximum likelihood is the predominantly used scoring rule, most likely he/she will refer to its attractive limiting properties—consistency, asymptotic normality, and most importantly, efficiency, i.e. maximum likelihood can reach the Cramér-Rao bound. However, as mentioned in Section 6.1, the ultimate goal in an observational study is to infer some average treatment effect. Propensity score model, no matter fitted by maximizing Bernoulli likelihood or CBSR, is just a means to this end. A natural question is: is it necessary or even beneficial to fit the propensity score model most efficiently by maximum likelihood?

Here we study the efficient estimation of weighted average treatment effects in the setting of nonparametric sieve regression. As the sample size n grows, a *sieve* estimator uses progressively more complex models to estimate the unknown propensity score. For example, we can increase the dimensional of the predictors in $\phi(x)$ in the GLM (6.5). This approach is used in Hirano, Imbens, and Ridder (2003) to estimate the propensity score by maximum likelihood. Their renowned results claim that the resulting IPW estimator is globally efficient for estimating ATE, ATT and other weighted average treatment effects. It is shown below that the global efficiency still holds if the Bernoulli likelihood is changed to the Beta family of scoring rules $G_{\alpha,\beta}$, $-1 \leq \alpha, \beta \leq 0$ in (6.3) or essentially any strongly concave scoring rule. Therefore there is no efficiency gain by sticking to the likelihood criterion.

First, let's briefly review the sieve logistic regression in Hirano et al. (2003). For $m = 1, 2, \ldots$, let $\phi_m(x) = (\varphi_{1m}(x), \varphi_{2m}(x), \ldots, \varphi_{mm}(x))^T$ be a triangular array of orthogonal polynomials, which are obtained by orthogonalizing the power series: $\psi_{km}(x) = \prod_{j=1}^{d} x_j^{\gamma_{kj}}$, where $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kd})^T$ is an d-dimensional multi-index of nonnegative integers and satisfies $\sum_{j=1}^{d} \gamma_{kj} \leq \sum_{j=1}^{d} \gamma_{k+1,j}$. Let l be the logistic link function (6.6). Hirano et al. (2003) estimated the propensity score by the following maximum likelihood rule

$$\hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \sum_{i=1}^{n} T_i \log\left(l^{-1}(\phi_m(X_i)^T\theta)\right) + (1 - T_i)\log\left(1 - l^{-1}(\phi_m(X_i)^T\theta)\right).$$

This is a special case of the proper scoring rule maximization (6.4) when the rule S is $S_{0,0}$ in the Beta family.

Besides Assumption 6.1 (strong ignorability), the other technical assumptions in Hirano et al. (2003) are listed below.

Assumption 6.2. *(Distribution of X) The support of X is a Cartesian product of compact intervals. The density of X is bounded, and bounded away from* 0.

Assumption 6.3. *(Distribution of $Y(0)$, $Y(1)$) The second moments of $Y(0)$ and $Y(1)$ exist and $g(X,0) = \mathrm{E}[Y(0)|X]$ and $g(X,1) = \mathrm{E}[Y(1)|X]$ are continuously differentiable.*

Assumption 6.4. *(Propensity score) The propensity score $p(X) = \mathrm{P}(T = 1|X)$ is continuously differentiable of order $s \geq 7d$ where d is the dimension of X, and $p(x)$*

is bounded away from 0 *and* 1*.*

Assumption 6.5. *(Sieve estimation) The nonparametric sieve logistic regression uses a power series with* $m = n^{\nu}$ *for some* $1/(4(s/d-1)) < \nu < 1/9$*.*

The most notable assumptions are the compactness of the support of X (Assumption 6.2) and the smoothness of $p(X)$ (Assumption 6.4), which are generally required in nonparametric regression, and the strong overlap assumption that $p(X)$ is bounded away from 0 and 1 (Assumption 6.4), which is necessary to ensure generalized inverse probability weight (6.9) is bounded. Another important assumption is the rate $m = n^{\nu}$ as $n \to \infty$ (Assumption 6.5).

Theorem 6.2 below is an extension to the main theorem of Hirano et al. (2003). Compared to the original theorem which always uses the maximum likelihood for any weighted average treatment effect, the scoring rule is now tailored according to the estimand as described in Section 6.3.1.

Theorem 6.2. *Suppose we use the Beta-family of covariate balancing scoring rules defined by equations* (6.2) *and* (6.3) *with* $-1 \le \alpha, \beta \le 0$ *and the logistic link* (6.6)*. Under Assumptions 6.1 to 6.5, the propensity score weighting estimator* $\hat{\tau}_{\alpha,\beta}$ *and its normalized version* $\hat{\tau}^{*}_{\alpha,\beta}$ *are consistent for* $\tau_{\alpha,\beta}$ *and* $\tau^{*}_{\alpha,\beta}$*. Moreover, they reach the semiparametric efficiency bound for estimating* $\tau_{\alpha,\beta}$ *and* $\tau^{*}_{\alpha,\beta}$*.*

Proof. See Section 6.8.2. □

6.5.2 Implications of Covariate Balance

If there is no efficiency gain in using maximum likelihood, what are the benefits of using a CBSR so the predictors are automatically balanced? One benefit is that the inverse weights w are less volatile, thanks to the observation in Section 6.3.2 that CBSR penalizes extreme inverse probability weights. This Section discusses another advantage, namely the bias reduction of $\hat{\tau}$ when $p_{\theta}(X)$ is misspecified. This is perhaps more important in practice, as Box (1976) once said: "all models are wrong, but some are useful".

Here we investigate the bias of $\hat{\tau}$ under the global null model. Denote the true outcome regression functions by $g(X,t) = \mathrm{E}[Y(t)|X]$, $t = 0, 1$. In the global null model model, $g(x,1) = g(x,0)$ for all x, so there is no treatment effect whatsoever. By definition (6.11) and (6.12), the weighted average treatment effects $\tau = \tau^*$ are always equal to 0.

Suppose the propensity score model is specified by $p_\theta(X)$ and the corresponding weights (6.9) are $w_\theta(X)$. Let $\tilde{\theta} = \arg\max_\theta S(p_\theta(X), p(X))$, so $p_{\tilde{\theta}}(x)$ is the best approximation of $p(x)$ with respect to the scoring rule S. Furthermore, define

$$\tilde{w}(x) = \frac{e(x)w_{\tilde{\theta}}(x,1)}{\mathrm{E}[e(X)w_{\tilde{\theta}}(X,1)]} - \frac{(1-e(x))w_{\tilde{\theta}}(x,0)}{\mathrm{E}[(1-e(X))w_{\tilde{\theta}}(X,0)]}.$$

The asymptotic bias of $\hat{\tau}^*$ is given by

$$\mathrm{bias}(\hat{\tau}^*) = \mathrm{E}[\hat{\tau}^*] = \mathrm{E}\left[\tilde{w}(X)g(X)\right].$$

When $p_\theta(X)$ is correctly specified (i.e. $p_{\tilde{\theta}}(x) = p(x)$), by the definition of GIPW (6.9), $\tilde{w}(x)$ is always zero. Therefore $\hat{\tau}$ is asymptotically unbiased under correctly specified propensity score model. When $p_\theta(X)$ is not correctly specified, the bias of $\hat{\tau}$ heavily depends on the covariate balance under the weight $w_{\tilde{\theta}}(X)$. To see this, notice that the covariate balancing property (6.10) can be written as $\mathrm{E}[\tilde{w}(X)\phi(X)] = 0$. Therefore, for any $\eta \in \mathbb{R}^m$,

$$\begin{aligned}\mathrm{bias}(\hat{\tau}^*) &= \mathrm{E}\left[\tilde{w}(X)(g(X) - \eta^T\phi(X))\right] \\ &\leq \mathrm{E}|\tilde{w}(X)| \cdot \left(\sup_x \left|g(x) - \eta^T\phi(x)\right|\right) \\ &= 2\sup_x \left|g(x) - \eta^T\phi(x)\right|.\end{aligned} \tag{6.27}$$

The last inequality is true if $\phi(x)$ includes an intercept term, since by (6.10),

$$\mathrm{E}[e(x)w_{\tilde{\theta}}(X,1)] = \mathrm{E}[Tw_{\tilde{\theta}}(X,1)] = 1 = \mathrm{E}[(1-T)w_{\tilde{\theta}}(X,0)] = \mathrm{E}[(1-e(X))w_{\tilde{\theta}}(X,1)].$$

Equation (6.27) leads to the next result:

Theorem 6.3. *Under Assumption 6.1 and the global null that $g(x, 0) = g(x, 1) = g(x)$ for all x, the estimator $\hat{\tau}^*$ is asymptotically unbiased if*

(i) A covariate balancing scoring rule is used and $\phi(x)$ includes an intercept term, and

(ii) $g(x)$ is in the linear span of $\{\varphi_1(x), \ldots, \varphi_m(x)\}$, or more generally $\inf_\eta \|g(x) - \eta^T \phi_m(x)\|_\infty \to 0$ as $n, m(n) \to \infty$.

The last condition says that $g(x)$ can be uniformly approximated by functions in the linear span of $\phi_1(x), \ldots, \phi_m(x)$ as $m \to \infty$. This holds under very mild assumption of g. For example, if the support of X is compact and $g(x)$ is continuous, the Weierstrass approximation theorem ensures that $g(x)$ can be uniformly approximated by polynomials. Theorem 6.3 can also be easily extended to the constant treatment effect model $g(x, 1) = g(x, 0) + c$. In this case, $\tau^* = c$ under any weighting and one can verify that the upper bound in (6.27) still holds.

Finally we compare the results in Theorem 6.3 and Theorem 6.2. The main difference is that Theorem 6.2 uses *propensity score* models with increasing complexity, whereas Theorem 6.3 assumes uniform approximation for the *outcome regression* function. Since the unbiasedness in Theorem 6.3 does not presume any assumption on the propensity score, the estimator $\hat{\tau}$ obtained by CBSR is more robust to misspecified or overfitted propensity score model.

6.5.3 Langrangian Duality

To understand the fundamental connection between propensity score weighting and empirical calibration in survey sampling (Deville and Särndal, 1992), here we present an alternative way to derive CBSR through Lagrangian duality. First, let's rewrite the score optimization problem (6.4) by introducing new variables f_i for each observation i:

$$\begin{aligned} \underset{f,\theta}{\text{maximize}} \quad & \frac{1}{n} \sum_{i=1}^{n} S(l^{-1}(f_i), T_i) \\ \text{subject to} \quad & f_i = \theta^T \phi(X_i),\ i = 1, \ldots, n. \end{aligned} \tag{6.28}$$

Let the Lagrangian multiplier associated with the i-th constraint be $(2T_i - 1)w_i/n$. The notation w indicates inverse probability weights in the last section. The reason of this abuse of notation will become clear in a moment. The Lagrangian of (6.28) is given by

$$Lag(f, \theta; w) = \frac{1}{n} \sum_{i=1}^{n} S(l^{-1}(f_i), T_i) + (2T_i - 1)w_i \left[f_i - \theta^T \phi(X_i) \right].$$

By setting the partial derivatives of the Lagrangian equal to 0, we obtain

$$\frac{\partial Lag}{\partial \theta_k} = \frac{1}{n} \sum_{i=1}^{n} (2T_i - 1) w_i \phi_k(X_i) = 0, \ k = 1, \ldots, m. \tag{6.29}$$

$$\frac{\partial Lag}{\partial f_i} = \frac{1}{n} \left(\frac{\partial S(l^{-1}(f_i), T_i)}{\partial f_i} + (2T_i - 1) w_i \right) = 0, \ i = 1, \ldots, n, \tag{6.30}$$

Equation (6.29) is the same as (6.16), meaning the optimal dual variables w balance the predictors $\phi_1, \ldots, \phi_m$. Equation (6.30) determines w from f. By using (6.2) and the logistic link (6.6), it turns out that $w_i = w(X_i, T_i)$ is exactly the GIPW weights defined in (6.9). In conclusion, the weights w are the dual variables of the score optimization problem (6.4) and are required to balance the predictors ϕ.

The benefit of this derivation is that we can write down the Lagrangian dual problem of (6.28). In general, there is no explicit form for $-1 < \alpha, \beta < 0$ because it is difficult to invert (6.9), but in the particularly interesting cases $\alpha = 0, \beta = -1$ (corresponding to ATT) and $\alpha = -1, \beta = -1$ (corresponding to ATE), the dual problems are algebraically tractable. When $\alpha = 0, \beta = -1$, the treated units are weighted by 1 and the control units are weighted by $\hat{p}/(1 - \hat{p})$. In this case, the Lagrangian dual optimization problem is given by

$$\begin{aligned} \underset{w \geq 0}{\text{minimize}} \quad & \sum_{i:T_i=0} w_i \log w_i - w_i \\ \text{subject to} \quad & \sum_{i:T_i=0} w_i \phi_k(X_i) = \sum_{j:T_j=1} \phi_k(X_j), \ k = 1, \ldots, m. \end{aligned} \tag{6.31}$$

In most cases an intercept term is included in the GLM, so the constraints in (6.31) imply that $\sum_{T_i=0} w_i$ is equal to the number of treated units (a fixed value). Therefore the dual optimization problem is equivalent to the following maximum entropy problem

$$\begin{aligned} \underset{w \geq 0}{\text{minimize}} \quad & \sum_{i:T_i=0} w_i \log w_i \\ \text{subject to} \quad & \sum_{i:T_i=0} w_i \phi_k(X_i) = \sum_{j:T_j=1} \phi_k(X_j),\ k = 1, \ldots, m. \end{aligned} \tag{6.32}$$

When $\alpha = \beta = -1$, the inverse probability weights are always greater than 1. It turns out that the Lagrangian dual problem in this case is given by

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} (w_i - 1) \log(w_i - 1) - w_i \\ \text{subject to} \quad & \sum_{i:T_i=0} w_i \phi_k(X_i) = \sum_{j:T_j=1} w_j \phi_k(X_j),\ k = 1, \ldots, m. \\ & w_i \geq 1,\ i = 1, \ldots, n. \end{aligned} \tag{6.33}$$

The objective functions in (6.31) and (6.33) encourage w to be close to uniform. They belong to a general distance measure $\sum_{i=1}^{n} D(w_i, v_i)$ in Deville and Särndal (1992), where $D(w, v)$ is a continuously differentiable and strongly convex function in w and achieves its minimum at (the limit) $w = v$. When the estimand is ATT (or ATE), the target v is equal to 1 (or 2). The average treatment effect estimators of this kind are often called "calibration estimators" in the survey sampling literature, because the weighted sample averages are empirically calibrated to some known unweighted population averages.

The maximum entropy problem (6.32) appeared first in Hainmueller (2011) to estimate ATT and is called "Entropy Balancing". In Section 7.2, we use the primal-dual connection described above to show Entropy Balancing is doubly robust, which is stronger than Theorems 6.2 and 6.3. Unfortunately, the double robustness only holds when the estimand is ATT. This section generalizes the primal-dual connnection to other weighted average treatment effects. Chan et al. (2015) studied the calibration estimators with the general distance D and showed the estimator $\hat{\tau}$ is globally semiparametric efficient. When the estimand is ATE, Chan et al. (2015) require the weighted sums of ϕ_k in (6.33) to be calibrated to $\sum_{i=1}^n \phi_k(X_i)/n$, too. It is shown earlier in Section 6.5.1 that this extra calibration is not necessary for semiparametric efficiency.

In an extension to Entropy Balancing, Hazlett (2013) proposed to empirically balance kernel representers instead of ordinary predictors. This corresponds to unregularized ($\lambda = 0$) RKHS regression introduced in Section 6.4.3. The unregularized problem is unfeasible if the RKHS is rich, so Hazlett (2013) tweaked the objective in order to find a usable solution.

6.5.4 Bias-variance tradeoff

The results in Sections 6.5.2 and 6.5.3 allow us to study the fundamental bias-variance in selecting a propensity score model. Consider the regularized regression approach introduced in Section 6.4.2. By the Karush-Kuhn-Tucker conditions of the regularized score maximization problem (6.4), the solution $\hat{\theta}_\lambda$ satisfies (for any $a \geq 1$ and $\lambda \geq 0$)

$$\left| \sum_{T_i=1} w_{\hat{\theta}_\lambda}(X_i,T_i)\phi_k(X_i) - \sum_{T_i=0} w_{\hat{\theta}_\lambda}(X_i,T_i)\phi_k(X_i) \right| \leq \lambda \cdot |(\hat{\theta}_\lambda)_k|^{a-1}, k = 1, \ldots, m. \tag{6.34}$$

The equality in (6.34) holds if $(\hat{\theta}_\lambda)_k \neq 0$, which is true unless $a = 1$ and λ is large. This suggests that in general the predictors $\phi(x)$ are not exactly balanced when $\lambda > 0$.

Following Section 6.5.2, we assume the null model $\mathrm{E}[Y(1)|X] = \mathrm{E}[Y(0)|X] = g(x)$ to study how covariate imbalance affects the bias of $\hat{\tau} = \hat{\tau}_\lambda$. Moreover, let's assume the outcome regression function is in the linear span of the predictors, i.e. $g(x) =$

$g_\eta(x) = \sum_{j=1}^m \eta_k \phi_k(x)$. The finite sample bias of $\hat{\tau}_\lambda$ under the null is given by

$$\begin{aligned}\mathrm{bias}_\eta(\hat{\tau}_\lambda) &= \left|\sum_{T_i=1} w_{\hat{\theta}_\lambda}(X_i,T_i) g_\eta(X_i) - \sum_{T_i=0} w_{\hat{\theta}_\lambda}(X_i,T_i) g_\eta(X_i)\right| \\ &= \left|\sum_{j=1}^m \eta_k \left(\sum_{T_i=1} w_{\hat{\theta}_\lambda}(X_i,T_i)\phi_k(X_i) - \sum_{T_i=0} w_{\hat{\theta}_\lambda}(X_i,T_i)\phi_k(X_i)\right)\right| \\ &\le \lambda \left|\sum_{k=1}^m \eta_k |(\hat{\theta}_\lambda)_k|^{a-1}\right| \le \lambda \|\eta\|_a \|\hat{\theta}_\lambda\|_a^{a-1}.\end{aligned}$$

The last inequality is due to Hölder's inequality and is tight. Hence we have

$$\max_\eta \frac{\mathrm{bias}_\eta(\hat{\tau}_\lambda)}{\|\eta\|_a} = \lambda \|\hat{\theta}_\lambda\|_a^{a-1}. \tag{6.35}$$

The next proposition says that the right hand side of equation (6.35) is decreasing as the degree of regularization λ becomes smaller. This is consistent with our intuition that the more we regularize the propensity score model, the more bias we get.

Proposition 6.3. *Given a strictly proper scoring rule S and a link function l such that $S(l^{-1}(f),t)$ is strongly concave and second order differentiable in $f \in \mathbb{R}$ for $t = 0,1$, let $\hat{\theta}_\lambda$ be the solution to (6.18) and (6.19) for a given $a \ge 1$. Then $\lambda\|\hat{\theta}_\lambda\|_a^{a-1}$ is a strictly increasing function of $\lambda > 0$.*

Proof. See Section 6.8.3. □

The bias-variance tradeoff is more apparent in the Lagrangian dual problem of (6.18). Consider the case that the estimand is ATT and the corresponding scoring rule is $\alpha = 0$ and $\beta = -1$. When $\phi_1(x) = 1$ and $J(\theta) = \sum_{k=2}^m |\theta_k|^a/a$ so we do not penalize the intercept, the dual problem of (6.18) is given by

$$\begin{aligned} \underset{w \geq 0}{\text{minimize}} \quad & \sum_{i:T_i=0} w_i \log w_i \\ \text{subject to} \quad & b_k = \sum_{i:T_i=0} w_i \phi_k(X_i) - \sum_{j:T_j=1} \phi_k(X_j),\ k = 1, \ldots, m. \\ & b_1 = 0,\ \|b\|_{a/(a-1)} \leq r(\lambda), \end{aligned} \tag{6.36}$$

where $r(\lambda)$ is some increasing function of λ. In (6.36), the objective function measures the closeness between w and the uniform weights and the constraints bound the covariate imbalance with respect to the functions ϕ. They are related, respectively, to the variance and bias of the estimator $\hat{\tau}$. When $\lambda \to 0$, the solution of (6.36) converges to the weights w that minimizes the $a/(a-1)$-norm of covariate imbalance. The limit of $r(\lambda)$ when $\lambda \to 0$ can be 0 or some positive value, depending on if the unregularized score maximization problem (6.4) is feasible or not. When $\lambda \to \infty$, the solution of (6.36) converges to uniform weights (i.e. no adjustment at all) whose estimator $\hat{\tau}$ has smallest variance. Similar arguments hold if we change the estimand (e.g. to ATE) and the scoring rule accordingly.

The kernel method introduced in Section 6.4.3 is a special case of the regularized regression with potentially infinite-dimensional predictors. For RKHS regressions, the maximum bias (6.35) under the null model is given by

$$\max_{g \in \mathcal{H}_K} \frac{\text{bias}_g(\hat{\tau}_\lambda)}{\|g\|_{\mathcal{H}_K}} = \lambda \|f\|_{\mathcal{H}_K}.$$

Therefore, the bias of $\hat{\tau}$ is controlled for a rich class of outcome regression functions.

6.6 Discussions

The covariate balancing scoring rule (CBSR) is largely inspired by some recent approaches that directly incorporate covariate balance in propensity score estimation. Motivated by Graham et al. (2012), Imai and Ratkovic (2014) proposed to augment the Bernoulli likelihood with the covariate balancing estimating equations, hoping they can robustify the propensity score model. These estimating equations are exactly the first-order conditions of maximizing CBSR. However, our derivation of CBSR shows that different estimating equations correspond to different estimands and there is little reason to combine them. In fact, Imai and Ratkovic (2014) also found in their simulation study that just using the covariate balancing estimating equations usually performs better. Another distinction is that Imai and Ratkovic (2014) solved the estimating equations by generalized method of moments or empirical likelihood. Those generic methods for over-identified estimating equations are generally not convex. In this Chapter, we identify the scoring rules corresponding to the estimating equation (6.10) and Proposition 6.2 shows that they are concave for estimating ATE and ATT with the logistic link function. Convex optimization methods can be used to solve the score maximization problem very efficiently.

Another related approach is Hainmueller (2011)'s Entropy Balancing which specializes in estimating ATT. It operates by maximizing the Shannon entropy of sample weights subject to exact covariate balance. Zhao and Percival (2015) found that the Lagrangian dual of Entropy Balancing fits a logistic propensity score model with a loss function different from the Bernoulli likelihood. We generalize this approach to general estimands.

To summarize, the decision theoretical approach we take has a number of advantages:

1. A proper scoring rule generates Fisher consistent estimates of the propensity score, allowing us to study the asymptotic properties of the IPW estimators.

2. The Lagrangian duality connects IPW estimators with the calibration estimators in survey sampling (Deville and Särndal, 1992). It also demonstrates an explicit bias-variance trade-off with regularized propensity score models.

3. The scoring rules (loss functions) can be plotted and interpreted visually, showing how propensity score estimation should be treated differently than a standard classification problem. CBSR penalizes more heavily on larger inverse probability weights hence generates a more stable estimator.

4. The convex loss function opens up numerous opportunities to use machine learning algorithms to estimate the propensity score. These algorithms are usually designed to optimize predictive performance. With the covariate balancing scoring rules as the objective, the machine learning algorithms now try to optimize covariate balance between the treatment groups.

6.7 Numerical examples

We use two examples (one simulation and one real data) to demonstrate the effectiveness of CBSR and the proposed adaptive methods.

6.7.1 A simulation example

This example due to Kang and Schafer (2007) is also used to generate Figure 6.1 in the Introduction. The artificial dataset consists of i.i.d. random variables (X_i, Z_i, T_i, Y_i), $i = 1, \dots, n$, where X_i, Y_i and T_i are always observed and Z_i is never observed. To generate this data set, X_i is a 4-dimensional vector distributed as $\mathrm{N}(0, I_4)$; Z_i is computed by first applying the following transformation:

$$\begin{aligned}
Z_{i1} &= \exp(X_{i1}/2),\\
Z_{i2} &= X_{i2}/(1+\exp(X_{i1})) + 10,\\
Z_{i3} &= (X_{i1}X_{i3} + 0.6)^3,\\
Z_{i4} &= (X_{i2} + X_{i4} + 20)^2,
\end{aligned}$$

and then normalizing individual variables of Z to have sample mean 0 and variance 1.

There are in total four settings in this example. In the first setting (top-left panel in Figure 6.4), Y_i is generated by $Y_i = g(X, T_i)$ without any additional noise, where

$$g(X, 0) = 210 + 27.4X_{i1} + 13.7X_{i2} + 13.7X_{i3} + 13.7X_{i4},$$

and $g(X, 1)$ is either equal to $g(X, 0)$ (column “zero” in Figure 6.4) or $g(X, 0) + 10$

(column "constant" in Figure 6.4). The true propensity scores are generated by the logistic model $p(X_i) = l^{-1}(f(X_i)),\ f(X_i) = -X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.1X_{i4}$. In this setting, both Y and T can be correctly modeled by (generalized) linear model of the observed covariates X. In the other settings, at least one of the propensity score model and the outcome regression model is non-linear in X. In order to achieve this, the data generating process described above is altered such that Y or T (or both) is linear in the unobserved Z instead of the observed X, though the parameters are kept the same. In these three scenarios, at least one of the two functions f and g are nonlinear in X.

For each setting, 200 replicas of dataset of size $n = 200$ are drawn. The logistic link function is always used and different scoring rules in the Beta-family (6.3) are applied. The predictor vector ϕ used is $\phi(X) = (X_1, X_1^2, X_2, X_2^2, X_3, X_3^2, X_4, X_4^2)$. After an estimated propensity score model is obtained, we use the normalized IPW estimator $\hat{\tau}^*_{-1,-1}$ to estimate ATE and $\hat{\tau}^*_{0,-1}$ to estimate ATT. The covariate imbalance with respect to ϕ is shown earlier in Figure 6.1.

Figure 6.4 shows the boxplots of these estimates under different settings. It is clear that the covariate balancing scoring rules (CBSR) generate much more stable estimates than the Bernoulli likelihood (MLE). Furthermore, in the two left panels the true logit f is linear in X so the propensity score model is correctly specified. In the two top panels the true outcome regression function g_0 is linear in X so the unbiasedness is guaranteed by Theorem 6.3. As expected, the weighting estimators given by CBSR are unbiased across these three panels (besides the bottom-right panel). If instead the Bernoulli likelihood criterion is used to estimate the propensity score model, the weighting estimator is biased when f is non-linear in X even if g is linear in X (top-right panel). Even if f is linear in X so the propensity score model is correctly specified, the CBSR estimators have much smaller variance than MLE. Lastly, in the bottom-right panel where both f and g are non-linear, CBSR still has smaller bias and variance.

Next we test the adaptive strategies described in Section 6.4. Here we consider three adaptive strategies—forward stepwise regression and two reproducing kernel Hilbert space (RKHS) regressions. In the forward stepwise regression, we use all

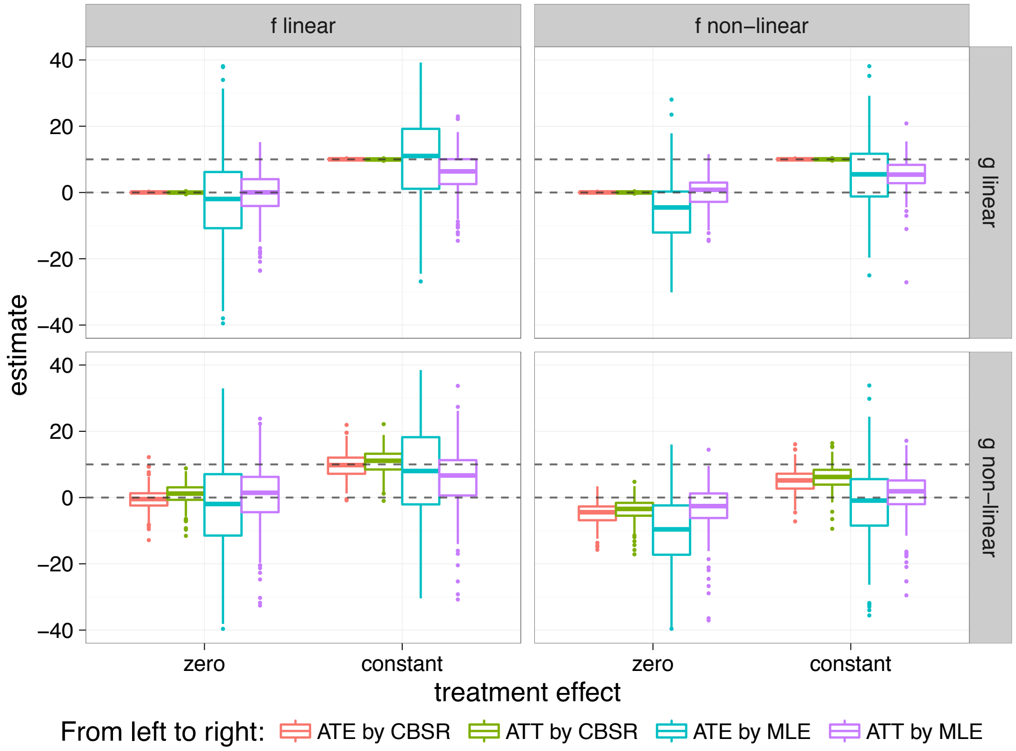


Figure 6.4: Estimate of average treatment effects (ATE or ATT) using different scoring rules under the four settings. The four boxes in each group with different colors correspond to the Beta scoring rule (6.3) with 1. $\alpha = \beta = -1$; 2. $\alpha = 0, \beta = -1$; 3. $\alpha = \beta = 0$; 4. $\alpha = \beta = 0$ (Bernoulli likelihood). In the first and third boxes, the inverse probability weights corresponding to ATE ($\alpha = \beta = -1$ in equation (6.15)) is used. In the second and fourth boxes, the inverse probability weights corresponding to ATT ($\alpha = 0, \beta = -1$) is used. The gray dashed lines correspond to the true treatment effect: 0 for group “zero” and 10 for group “constant”.

the two-way interactions of degree-two polynomials (in total 32 predictors) to allow sophisticated propensity score models. In the two RKHS regressions, we use the Gaussian kernel

$$K(x, x') = \exp(-\gamma \|x - x'\|^2),\ x, x' \in \mathbb{R}^4,$$

with γ equal to 0.2 and 0.5. After fitting a path of propensity score models (indexed by step for forward stepwise and regularization parameter λ for RKHS), for each strategy we choose the model that has the smallest number of significant two-sample t-tests as described in Section 6.4.5. Finally, the standard errors and confidence intervals by assuming homoscedasticity in (6.26) and using a pilot outcome regression (predictors are the observed X). Since a fairly sophisiticated propensity score model can be fitted, we use $n = 1000$ samples and set $g(X, 1) = g(X, 0)$ to test all the methods.

Table 6.2 shows the performance of the six different combinations of loss function and adaptive strategy in the four simulation settings. CBSR clearly outperforms Bernoulli likelihood. In almost all scenarios and no matter what adaptive strategy is used, the root mean squared error (RMSE) of CBSR is less than half of the RMSE of Bernoulli likelihood. The confidence intervals obtained by using Bernoulli likelihood also perform poorly. In many scenarios the actual coverage is less than 50%, whereas the nominal coverage is 95%. CBSR's confidence intervals have close to or over the nominal 95% coverage in almost all scenarios.

The two adaptive strategies (forward stepwise and RKHS) perform similarly. When using CBSR as the loss function, forward stepwise seems to have slightly smaller RMSE, but kernel methods can have the better coverage in some scenarios. In practice, the user may want to choose an adaptive strategy that is most convenient for the target application. The strength and weakness of these methods are discussed in Section 6.4.

6.7.2 A Real Data Example

This Section studies the National Supported Work (NSW) Demonstration which was previously analyzed by LaLonde (1986), Dehejia and Wahba (1999), Smith and Todd

f (true PS)	g (true OR)	estimand	loss	strategy	bias	RMSE	coverage
linear	linear	ATE	Bernoulli	forward stepwise	−1.27	2.43	50.5
				kernel (0.2)	−2.77	3.18	22.5
				kernel (0.5)	−3.05	3.45	17.5
			CBSR	forward stepwise	−0.24	0.98	90.0
				kernel (0.2)	−0.50	1.16	90.5
				kernel (0.5)	−1.41	1.77	63.0
		ATT	Bernoulli	forward stepwise	−2.92	6.28	55.5
				kernel (0.2)	−7.30	11.08	29.0
				kernel (0.5)	−2.70	3.34	27.5
			CBSR	forward stepwise	−0.24	1.19	91.5
				kernel (0.2)	−1.09	2.78	90.5
				kernel (0.5)	−1.91	2.49	63.5
	nonlinear	ATE	Bernoulli	forward stepwise	−1.05	2.17	82.5
				kernel (0.2)	−1.90	2.52	74.0
				kernel (0.5)	−2.13	2.77	68.0
			CBSR	forward stepwise	−0.46	1.17	99.5
				kernel (0.2)	−0.46	1.17	100.0
				kernel (0.5)	−1.10	1.62	96.0
		ATT	Bernoulli	forward stepwise	−0.93	3.87	87.5
				kernel (0.2)	−3.16	6.64	65.5
				kernel (0.5)	−0.05	1.82	94.0
			CBSR	forward stepwise	−0.49	1.37	99.0
				kernel (0.2)	−0.27	2.05	99.0
				kernel (0.5)	−0.59	1.83	99.5
nonlinear	linear	ATE	Bernoulli	forward stepwise	−1.55	2.07	45.5
				kernel (0.2)	−2.28	2.75	31.0
				kernel (0.5)	−2.54	2.97	24.5
			CBSR	forward stepwise	−0.27	1.02	86.5
				kernel (0.2)	−0.40	1.10	92.5
				kernel (0.5)	−1.19	1.61	64.5
		ATT	Bernoulli	forward stepwise	−0.45	1.94	78.5
				kernel (0.2)	−0.88	2.43	64.0
				kernel (0.5)	−0.92	1.84	62.0
			CBSR	forward stepwise	−0.13	1.14	89.0
				kernel (0.2)	−0.36	1.46	93.5
				kernel (0.5)	−0.83	1.61	82.5
	nonlinear	ATE	Bernoulli	forward stepwise	−2.25	2.75	64.5
				kernel (0.2)	−2.90	3.37	51.0
				kernel (0.5)	−3.29	3.69	41.5
			CBSR	forward stepwise	−0.61	1.01	100.0
				kernel (0.2)	−0.73	1.26	100.0
				kernel (0.5)	−1.88	2.19	86.5
		ATT	Bernoulli	forward stepwise	−0.12	1.82	96.5
				kernel (0.2)	−0.02	2.34	96.0
				kernel (0.5)	−0.37	1.64	95.5
			CBSR	forward stepwise	−0.39	1.12	100.0
				kernel (0.2)	−0.35	1.46	100.0
				kernel (0.5)	−1.01	1.74	99.5

Table 6.2: Performance of different loss functions combined with adaptive strategies—forward stepwise and kernel method (Gaussian kernel with bandwidth parameter 0.2 and 0.5). In each case, the propensity score model is selected to minimize the number of significant covariate imbalance tests. Compared to the Bernoulli likelihood, maximizing the covariate balancing scoring rule (CBSR) reduces the root mean square error (RMSE) by more than a half for most settings. CBSR's confidence intervals also have the superior coverage (nominal level is 95%).

(2005) and many other authors. The NSW Demonstration was a federally and privately funded program implemented in the 1970s, which program provided transitional and subsidized work experience for a period of 6–18 months to individuals who had faced economic and social problems prior to the enrollment in the program. The pre-treatment covariates include earnings, education, age, ethnicity, and marital status, and the outcome of interest in LaLonde (1986) is the post-intervention earnings in 1978. We use the experimental subsample taken by Dehejia and Wahba (1999) to demonstrate our methods, which include 185 treated and 260 control observations that joined the program early enough for the retrospective earnings information in the year 1974. To evaluate the non-experimental methods, we use the Current Population Survey (CPS) data extracted by LaLonde (1986) as the control group, which contain 15992 observations. The reader is referred to the previous articles listed in this paragraph for more detailed information on this dataset.

The observational methods are evaluated in two scenarios:

1. Compare the experimental treated group with the non-experimental control group. The average treatment effect estimators can be compared with the experimental benchmark, which is 1794.3 (standard error 632.9) by a linear regression of the earnings in 1978 on the treatment assignment.

2. Compare the experimental control group with the non-experimental control group. Since both groups did not receive treatment, the treatment effect is always zero (the null case in Section 6.5.2).

Since the non-experimental control group is very large and has very different covariate distribution to the experimental group, we only consider the average treatment effect on the treated (ATT) in this example. Using Table 6.1, the CBSR rule in this case is $S_{0,-1}$.

First, we apply the forward stepwise regression in Algorithm 6.1. To generate predictors $\phi(X)$, we use all the discrete covariates (race, married, no degree, no earning in 1974, no earning in 1975), all the continuous covariates (age, year of education, earning in 1974, earning in 1975) and their squares, and the first-order interactions of all these variables. This results in a 94-dimensional vector ϕ of predictors. In each scenario, two scoring rules, the Bernoulli likelihood $S_{0,0}$ and the covariate balancing scoring rule $S_{0,-1}$, are used, and the covariate imbalance and the estimator $\hat{\tau}$ are plotted along the stepwise path.

The results of the forward stepwise regressions can be found in Figure 6.5. Compared to the Bernoulli likelihood, CBSR requires a stronger condition for the existence of the solution (Zhao and Percival, 2015), so in both scenarios CBSR stops early (71 steps in scenario 1 and 23 steps in scenario 2). Nevertheless, CBSR is much better at reducing covariate imbalance as shown in Figures 6.5a and 6.5b. In fact, at least 10 predictors have standardized difference greater than 20% across the entire path when the Bernoulli likelihood is used, while some authors have suggested that a standardized difference above 10% can be deemed substantial (Normand et al., 2001, Austin and Stuart, 2015). When a two-sample t-test is used to compare the mean of the predictors, less than 20% of the 94 tests are insignificant with the Bernoulli likelihood, also implying insufficient covariate balance. On the contrary, CBSR successfully balances most predictors in both scenarios.

Figures 6.5c and 6.5d show the estimate of ATT along the path. Interesting, by just including the first predictor most of the bias of estimating ATT is corrected. Both scoring rules give similar estimates and are consistent with the experimental benchmarks. However, as discussed above, the weights generated by maximizing the Bernoulli likelihood are unacceptable to many applied researchers. Switching to CBSR solves this problem, though the ATT estimates are not very different in this particular example. Additionally, when using CBSR the standard error of $\hat{\tau}$ is smaller. This can be understood from the remark in Section 6.3.2 that CBSR tries to avoid large weights.

Next, we apply the kernel method in Section 6.4.3 and the results are presented in Figure 6.6. We use the Gaussian kernel $K(x, x') = \exp(-\sigma \|x - x'\|^2)$ with $\sigma = 0.15$ and $x = (\texttt{black}, \texttt{hispanic}, \texttt{no degree}, \texttt{married}, \texttt{age}/5, \texttt{education}/3, \texttt{re74}/4000, \texttt{re75}/4000)$. The first four entries in x are indicator variables, and `re74` (`re75`) stands for the annual earning of the person in the year 1974 (1975). Because the kernel matrix is a large $n \times n$ matrix, we use the subsample CPS2 extracted by Dehejia and Wahba (1999) that contains $n = 2369$ non-experimental controls. Overall, these two plots are similar to those for the forward stepwise regressions. Notice that the confidence intervals of ATT are wider when using the kernel method. This loss of efficiency is compensated by the improved robustness, as the propensity score weights approximately balance infinite many covariate functions (Section 6.4.3).

In conclusion, the two adaptive strategies work very well in the NSW job training example and CBSR continues to balance the sample covariates better than the Bernoulli likelihood.

6.8 Theoretical proofs

6.8.1 Proof of Proposition 6.2

The same result can be found in Buja et al. (2005, Section 15). For completeness we give a direct proof here. Denote $p = l^{-1}(f) \in (0,1)$ and notice that $\mathrm{d}f/\mathrm{d}p = (l^{-1})'(f) = p(1-p)$. By (6.2), we have

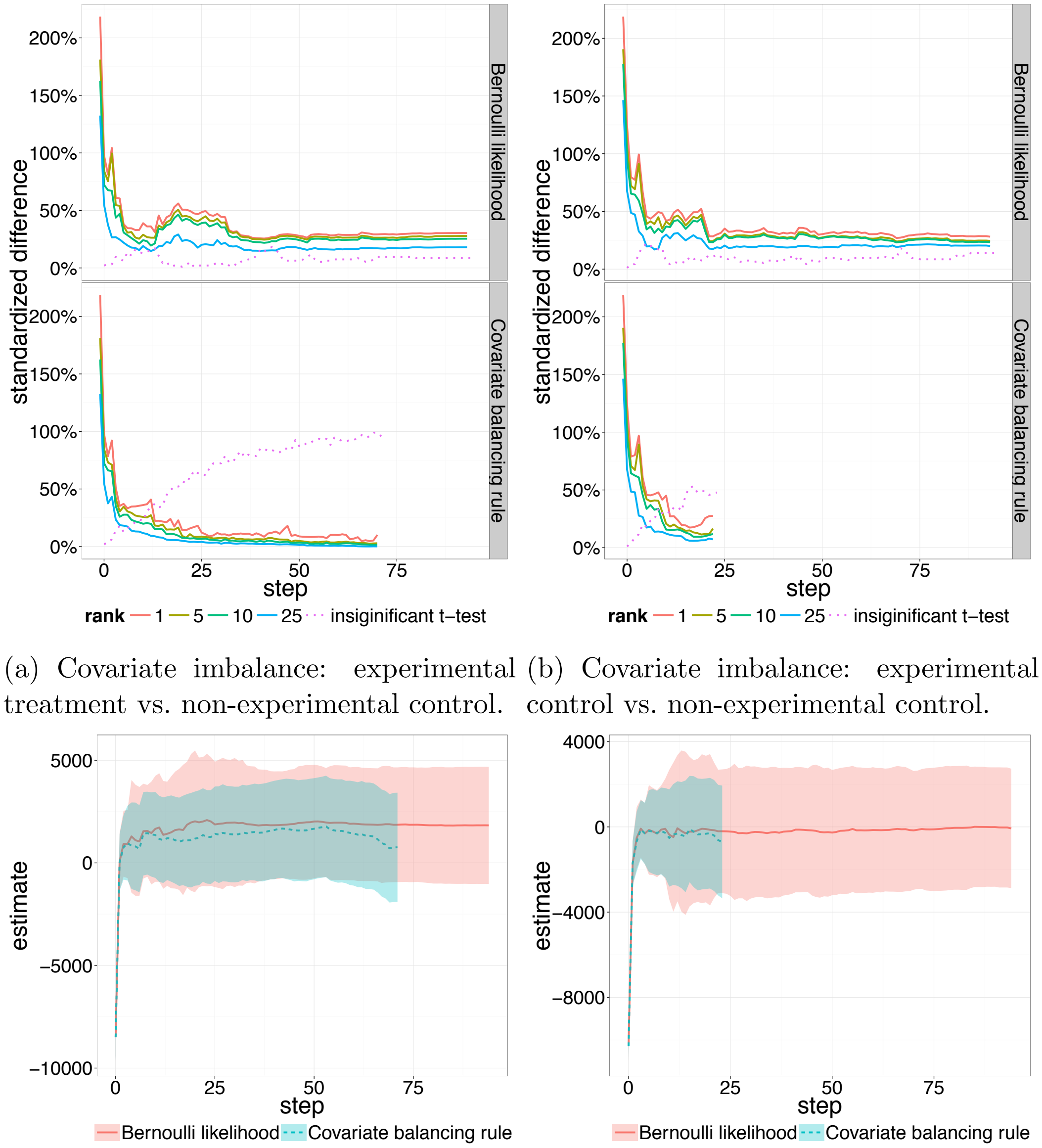
$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}f} S(l^{-1}(f),1) &= (1-p)G''(p)(l^{-1})'(f) = p^{\alpha}(1-p)^{\beta+1}, \\
\frac{\mathrm{d}}{\mathrm{d}f} S(l^{-1}(f),0) &= -pG''(p)(l^{-1})'(f) = -p^{\alpha+1}(1-p)^{\beta}, \text{ and}
\end{aligned}$$

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}f^2} S(l^{-1}(f),1) &= \alpha p^{\alpha}(1-p)^{\beta+2} - (\beta+1)p^{\alpha+1}(1-p)^{\beta+1}, \\
\frac{\mathrm{d}^2}{\mathrm{d}f^2} S(l^{-1}(f),0) &= -(\alpha+1)p^{\alpha+1}(1-p)^{\beta+1} + \beta p^{\alpha+2}(1-p)^{\beta}.
\end{aligned}$$

The conclusions immediate follow by letting the second order derivatives be less than or equal to 0.

6.8.2 Proof of Theorem 6.2

The proof is a simple modification of the proof in Hirano et al. (2003). In fact, Hirano et al. (2003) only proved the convergence of the estimated propensity score up to certain order. This essentially suggests that the semiparametric efficiency of $\hat{\tau}$ does not heavily depend on the accuracy of the sieve logistic regression.



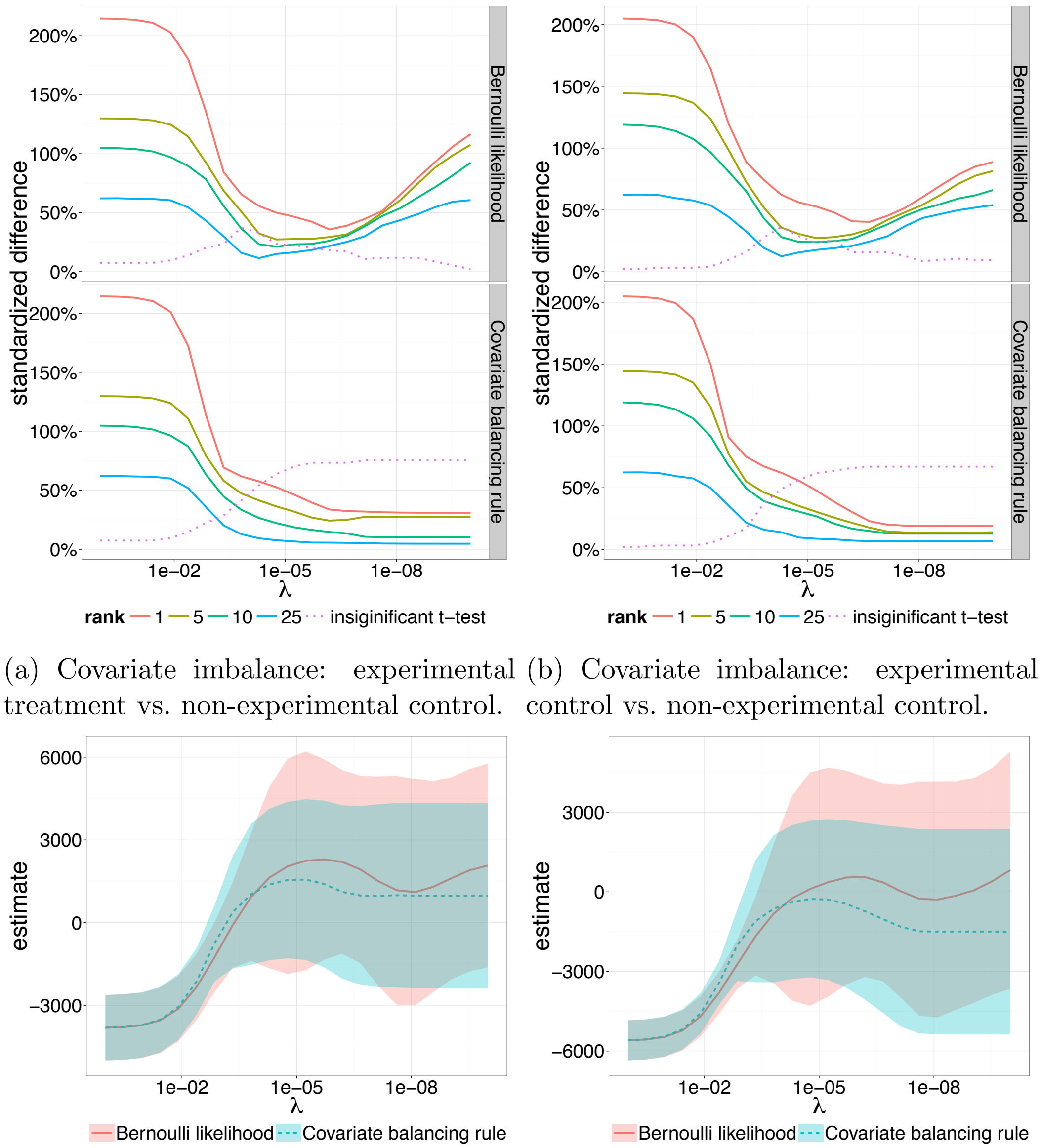
(a) Covariate imbalance: experimental treatment vs. non-experimental control.

(b) Covariate imbalance: experimental control vs. non-experimental control.

(c) Estimate of ATT: experimental treatment vs. non-experimental control.

(d) Estimate of ATT: experimental control vs. non-experimental control.

Figure 6.5: Forward stepwise regressions for the LaLonde (1986) dataset. Top panels: Covariate imbalance is in terms of standardized difference. The curves in the plot are the 1st, 5th, 10th, and 25th largest standardized difference among the 94 predictors (solid lines), and the percentage of insignificant two-sample t-tests comparing the mean of each predictor in the treatment and the weighted control (dotted line). Bottom panels: estimated ATT with 95% confidence interval.



(a) Covariate imbalance: experimental treatment vs. non-experimental control.

(b) Covariate imbalance: experimental control vs. non-experimental control.

(c) Estimate of ATT: experimental treatment vs. non-experimental control.

(d) Estimate of ATT: experimental control vs. non-experimental control.

Figure 6.6: Reproducing kernel method for the LaLonde (1986) dataset. Top panels: Covariate imbalance is in terms of standardized difference. The curves in the plot are the 1st, 5th, 10th, and 25th largest standardized difference among the 94 predictors (solid lines), and the percentage of insignificant two-sample t-tests comparing the mean of each predictor in the treatment and the weighted control (dotted line). Bottom panels: estimated ATT with 95% confidence interval.

To be more specific, only three properties of the maximum likelihood rule $S = S_{0,0}$ are used in Hirano et al. (2003, Lemmas 1,2):

1. $\tilde{\theta} = \arg\max_{\theta} S(p_\theta, p_{\tilde{\theta}})$ (line 5, page 19), this is exactly the definition of a strictly proper scoring rule (6.1);

2. The Fisher information matrix

$$\frac{\partial^2}{\partial\theta\partial\theta^T} S(p_\theta, p_{\tilde{\theta}}) = \mathrm{E}_{\tilde{\theta}} \left\{ \left[\frac{\mathrm{d}^2}{\mathrm{d}f^2} S(l^{-1}(f), T) \Big|_{f=\phi(X)^T\theta} \right] \phi(X)\phi(X)^T \right\}$$

   has all eigenvalues uniformly bounded away from 0 for all θ and $\tilde{\theta}$ in a compact set in $\mathbb{R}^m$, where the expectation on the right hand side is taken over X and $T|X \sim p_{\tilde{\theta}}$.

3. As $m \to \infty$, with probability tending to 1 the observed Fisher information matrix

$$\frac{\partial^2}{\partial\theta\partial\theta^T} \frac{1}{n} \sum_{i=1}^{n} S(p_\theta(X_i), T_i) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\mathrm{d}^2}{\mathrm{d}f^2} S(l^{-1}(f), T_i) \Big|_{f=\phi(X_i)^T\theta} \right] \phi(X_i)\phi(X_i)^T$$

   has all eigenvalues uniformly bounded away from 0 for all θ in a compact set of $\mathbb{R}^m$ (line 7–9, page 21).

Because the approximating functions ϕ are obtained through orthogonalizing the power series, we have $\mathrm{E}[\phi(X)\phi(X)^T] = I_m$ and one can show its finite sample version has eigenvalues bounded away from 0 with probability going to 1 as $n \to \infty$. Therefore a sufficient condition for the second and third properties above is that $S(l^{-1}(f), t)$ is strongly concave for $t = 0, 1$. In Proposition 6.2 we have already proven the strong concavity for all $-1 \leq \alpha, \beta \geq 1$ except for $\alpha = -1, \beta = 0$ and $\alpha = 0, \beta = -1$. In these two boundary cases, among $S(l^{-1}(f), 0)$ and $S(l^{-1}(f), 1)$ one score function is strongly concave and the other score function is linear in f. One can still prove the second and third properties by using Assumption 6.4 that the propensity score is bounded away from 0 and 1.

6.8.3 Proof of Proposition 6.3

The conclusion is trivial for $a = 1$. Denote

$$h(f,t) = \frac{\mathrm{d}}{\mathrm{d}f} S(l^{-1}(f),t) \text{ and } h'(f,t) = \frac{\mathrm{d}}{\mathrm{d}f} h(f,t),\ t = 0,1.$$

Because $S(l^{-1}(f),t)$ is concave in f, we have $h'(f,t) < 0$ for all f. The first-order optimality condition of (6.18) is given by

$$\frac{1}{n}\sum_{i=1}^{n} h(\hat{\theta}_\lambda^T \phi(X_i), T_i)\phi_k(X_i) + \lambda|(\hat{\theta}_\lambda)_k|^{a-1}\mathrm{sign}((\hat{\theta}_\lambda)_k) = 0,\ k = 1,\dots,m.$$

Let $\nabla\hat{\theta}_\lambda$ be the gradient of $\hat{\theta}_\lambda$ with respect to λ. By taking derivative of the identity above, we get

$$\left[\frac{1}{n}\sum_{i=1}^{n} h'(\hat{\theta}_\lambda^T \phi_i, T_i)\phi_i\phi_i^T + \lambda(a-1)\mathrm{diag}(|\hat{\theta}_\lambda|^{a-2})\right]\nabla\hat{\theta}_\lambda = -|\hat{\theta}_\lambda|^{a-1}\mathrm{sign}(\hat{\theta}_\lambda),$$

where we used the abbreviation $\phi_i = \phi(X_i)$ and $\theta^a = (\theta_1^a,\dots,\theta_m^a)$. For brevity, let's denote

$$H = \frac{1}{n}\sum_{i=1}^{n} h'(\hat{\theta}_\lambda^T \phi_i, T_i)\phi_i\phi_i^T \prec 0 \text{ and } G = \lambda(a-1)\mathrm{diag}(|\hat{\theta}_\lambda|^{a-2}).$$

For $a > 1$, the result is proven by showing the derivative of $\lambda\|\hat{\theta}_\lambda\|_a^{a-1}$ is greater than 0.

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\lambda\|\hat{\theta}_\lambda\|_a^{a-1}\right) &= \|\hat{\theta}_\lambda\|_a^{a-1} + \lambda\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\sum_{j=1}^{m}\left|(\hat{\theta}_\lambda)_k\right|^a\right]^{(a-1)/a} \\
&= \|\hat{\theta}_\lambda\|_a^{a-1} + \lambda(a-1)\|\hat{\theta}_\lambda\|_a^{-1}\sum_{j=1}^{m}\left|(\hat{\theta}_\lambda)_k\right|^{a-1}(\nabla\hat{\theta}_\lambda)_k\,\mathrm{sign}((\hat{\theta}_\lambda)_k) \\
&= \|\hat{\theta}_\lambda\|_a^{a-1} - \lambda(a-1)\|\hat{\theta}_\lambda\|_a^{-1}(|\hat{\theta}_\lambda|^{a-1})^T(H+G)^{-1}|\hat{\theta}_\lambda|^{a-1} \\
&> \|\hat{\theta}_\lambda\|_a^{a-1} - \lambda(a-1)\|\hat{\theta}_\lambda\|_a^{-1}(|\hat{\theta}_\lambda|^{a-1})^T G^{-1}|\hat{\theta}_\lambda|^{a-1} \\
&= 0.
\end{aligned}$$

Chapter 7

Outcome Regression and Doubly Robust Inference

So far we have given a comprehensive review of methods based on the treatment assignment mechanism in Chapter 5 and provided some new approaches in Chapter 6. These methods only model the relation between the covariates X and the treatment assignment T and ignores the relation between X and the response Y. This is in some sense an attractive property because we are only interested in the causal effect of T. Once a propensity score model is obtained, the inference for the average causal effect is relatively simple. Since we have not looked at Y yet, confidence interval can be constructed by estimating the variance of Y as in (6.26). This Chapter considers to augment this inference by modeling the relation between X and Y.

7.1 Outcome regression

Suppose we make the structural assumption that

$$\mathrm{E}[Y(t)|X] = \beta_0 + \tau t + \beta_1^T X + \beta_2^T tX,\ t = 0, 1, \tag{7.1}$$

and suppose $\mathrm{E}[X] = 0$. Then the conditional average treatment effect $\mathrm{E}[Y(1)|X] - \mathrm{E}[Y(0)|X] = \tau + \beta_2^T X$ and the average treatment effect (ATE) is τ. In principle, the parameters in (7.1) can be estimated by linear regression such as the ordinary least squares. This technique is called *outcome regression*, in contrast to the *propensity score* methods described in the last two Chapters.

The model (7.1) can be rewritten as two separate regressions:

$$\mathrm{E}[Y(0)|X] = g_0(X) \text{ and } \mathrm{E}[Y(1)|X] = g_1(X). \tag{7.2}$$

Once the regression functions g_0 and g_1 are estimated, the ATE can be estimated by

$$\hat{\tau}_{\mathrm{OR}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_1(X_i) - \hat{g}_0(X_i)). \tag{7.3}$$

It is easy to verify that this is equivalent to $\hat{\tau}$ obtained by fitting (7.1) jointly when both g_0 and g_1 are linear in x.

Consistency of the outcome regression estimator (7.2) depends on the consistency of $\hat{g}_0$ and $\hat{g}_1$. When the regression functions are modeled parametrically, $\hat{\tau}_{\mathrm{OR}}$ is consistent only if the model specification is correct. This is similar to propensity score methods in Chapter 5 whose consistency relies on correctly specifying the propensity score model. Practically, it is beneficial to fit a fairly complicated model for g_0 and g_1. as commonly done for the propensity score model too (Chapter 6).

7.2 Doubly robust estimation

Robins et al. (1994) introduce a method called *augmented inverse probability weighting* to combine propensity score and outcome regression in estimating the average treatment effect. Given an (estimated) propensity score model $\hat{p}(x)$ and an estimated outcome regression model $\hat{g}_t(x)$, $t = 0, 1$, the ATE can be estimated by

$$\hat{\tau}_{\mathrm{DR}} = \hat{\tau}_{\mathrm{OR}} + \sum_{i=1}^{n}(2T_i - 1)\hat{w}(X_i, T_i)\left(Y_i - \hat{g}_{T_i}(X_i)\right), \tag{7.4}$$

where $\hat{w}(X_i, T_i)$ is computed from $\hat{p}$ by (6.17). This estimator can be viewed as an improvement over the outcome regression estimator. The second term in (7.4) replaces the response Y_i in the IPW estimator $\hat{\tau}$ by the residual $Y_i - \hat{g}_{T_i}(X_i)$. Therefore, it estimates the bias of $\hat{\tau}_{\mathrm{OR}}$ due to model misspecification. Following this reasoning, it is easy to verify that $\hat{\tau}_{\mathrm{DR}}$ is consistent if $\hat{p}$ is consistent or $\{\hat{g}_t,\ t = 0, 1\}$ are consistent, or both. This property is called "double robustness" by Robins et al. (1994). Numerous doubly robust estimators have been proposed since then, see Bang and Robins (2005), Kang and Schafer (2007), Tan (2006, 2010) for some review.

Double robustness is also closely related to covariate balance weights discussed in Section 5.3 and Chapter 6. Next we present the main result of Zhao and Percival (2015) which shows that the Entropy Balancing estimator given by (6.32) is doubly robust. Notice that Hainmueller (2011)'s original proposal does not explicitly model propensity score or outcome regression.

Theorem 7.1. *Let Assumption 6.1 (strong ignorability) and the overlap assumption* $0 < \mathrm{P}(T = 1|X) < 1$ *be given. Additionally, assume the expectation of* $\phi(x)$ *(the covariate function being balanced) exists and* $\mathrm{Var}(Y(0)) < \infty$*. Then Entropy Balancing is doubly robust in the sense that*

1. *If* $\mathrm{logit}(p(x))$ *or* $g_0(x)$ *is linear in* $\phi_k(x),\ k = 1, \ldots, m$*, then* $\hat{\gamma}^{\mathrm{EB}}$ *is statistically consistent.*

2. *Moreover, if* $\mathrm{logit}(e(x))$*,* $g_0(x)$ *and* $g_1(x)$ *are all linear in* $\phi_k(x),\ k = 1, \ldots, R$*, then* $\hat{\gamma}^{\mathrm{EB}}$ *reaches the semiparametric variance bound of* γ *derived in Hahn (1998, Theorem 1) with unknown propensity score.*

We give two proofs of the first claim in Theorem 7.1. The first proof reveals an interesting connection between the primal-dual optimization problems and the statistical property double robustness. The second proof uses a stabilization trick in Robins et al. (2007). The reader is referred to Zhao and Percival (2015) for the proof of the second claim in Theorem 7.1.

First proof (sketch). The consistency under the linear model of $\mathrm{logit}(p(x))$ is a consequence of the Fisher consistency of the scoring rule $S_{0,-1}$. See Section 6.3.1. The consistency under the linear model of $Y(0)$ can be proved by expanding $g_0(X)$ and $\sum_{T_i=0} w_i Y_i$. Here we provide an indirect proof by showing that augmenting Entropy

Balancing with a linear outcome regression as in Equation (7.4) does not change the estimator, and hence Entropy Balancing is doubly robust. This fact can be proved by a few lines of algebra: let the estimated outcome regression model be $\hat{g}_0(x) = \sum_{j=1}^{p} \hat{\beta}_j c_j(x)$, then

$$\begin{aligned}
\hat{\tau}_{\mathrm{DR}} - \hat{\tau}_{\mathrm{EB}} &= \sum_{T_i=0} \hat{w}_i \hat{g}_0(X_i) - \frac{1}{n_1} \sum_{T_i=0} \hat{g}_0(X_i) \\
&= \sum_{T_i=0} \hat{w}_i \sum_{k=1}^{p} \hat{\beta}_k \phi_k(X_i) - \frac{1}{n_1} \sum_{T_i=1} \sum_{j=1}^{p} \hat{\beta}_j \phi_j(X_i) \\
&= \sum_{k=1}^{p} \hat{\beta}_k \left(\sum_{T_i=0} \hat{w}_i \phi_j(X_i) - \frac{1}{n_1} \sum_{T_i=1} \phi_j(X_i) \right) \\
&= 0.
\end{aligned}$$

Therefore, by enforcing covariate balancing constraints, Entropy Balancing implicitly fits a linear outcome regression model and is consistent for τ under that working model. □

Second proof. This proof is pointed out by an anonymous reviewer. In a discussion of Kang and Schafer (2007), Robins et al. (2007) indicated that one can stabilize the standard doubly robust estimator in a number of ways. Specifically, one trick suggested by Robins et al. (2007, Section 4.1.2) is to estimate the propensity score, say $\tilde{p}(x)$, by the following estimating equation

$$\sum_{i=1}^{n} \left[\frac{(1-T_i)\tilde{p}(X_i)/(1-\tilde{p}(X_i))}{\sum_{i=1}^{n}(1-T_i)\tilde{p}(X_i)/(1-\tilde{p}(X_i))} - \frac{T_i}{\sum_{i=1}^{n} T_i} \right] \hat{g}_0(X_i) = 0. \tag{7.5}$$

Then one can estimate the ATT by the usual IPW estimator with $\hat{p}(X_i)$ replaced by $\tilde{p}(X_i)$. This estimator is sample bounded (the estimator is always within the range of observed values of Y) and doubly robust with respect to the parametric specifications of $\tilde{p}(x) = \tilde{p}(x;\theta)$ and $\hat{g}_0(x) = \hat{g}_0(x;\beta)$. The only problem with (7.5) is it may not have a unique solution. However, when $\mathrm{logit}(p(x))$ and $g_0(x)$ are modeled linearly in $\phi(x)$, (7.5) corresponds to the first order condition of the optimization problem maximizing the score $S_{0,-1}$ (the dual of Entropy Balancing). Since Entropy Balancing is a strictly convex optimization problem, it has an unique solution and $\tilde{p}(X;\theta)$ is the same as the Entropy Balancing estimate $\hat{p}(X;\theta)$. As a consequence, $\hat{\tau}_{\mathrm{EB}}$ is also doubly robust. □

Notice that the reasoning in the first proof would work for any empirically calibrated weighting estimator. That is, if we have weights $\{\hat{w}_i, i = 1,\ldots,n\}$ that empirically balance $\phi(x)$, then augmenting the linear outcome regression estimator with $\phi(x)$ being the predictor does not change anything. However, not all empirically calibrated weights have the propensity score interpretation as Entropy Balancing.

Finally, it is easy to extend Entropy Balancing to accommodate arbitrary outcome regression model. The trick is to include the estimated $g_0(x)$ in the covariate functions to be balanced. Then Theorem 7.1 implies that the extended Entropy Balancing estimator is doubly robust with respect to logistic propensity score model and the given outcome regression model $\hat{g}_0$.

7.3 Robust inference of ATT

We can improve the adaptive procedures in Section 6.4 by an outcome regression model. For simplicity, let's focus on estimating the ATT, so the corresponding scoring rule is $S_{0,-1}$. In this case, it is not difficulty to show (notice that $\hat{w}_i = \hat{w}(X_i, T_i) = 1/n_1$ if $T_i = 1$)

$$\hat{\tau}_{\mathrm{DR}} = \sum_{i=1}^{n} (2T_i - 1)\hat{w}_i (Y_i - \hat{g}_0(X_i)).$$

Therefore, we have the following decomposition of estimation error

$$
\begin{aligned}
&\left|\hat{\tau}_{\mathrm{DR}} - \frac{1}{n_1}\sum_{T_i=1}(g_1(X_i) - g_0(X_i))\right| \\
=&\left|\sum_{i=1}^{n}(2T_i-1)\hat{w}(X_i,T_i)(Y_i - \hat{g}_0(X_i)) - \frac{1}{n_1}\sum_{T_i=1}(g_1(X_i)-g_0(X_i))\right| \\
\leq&\left|\sum_{i=1}^{n}(2T_i-1)\hat{w}_i\epsilon_i\right| + \left|\sum_{T_i=1}(2T_i-1)\hat{w}_i(g_0(X_i)-\hat{g}_0(X_i))\right| \\
\leq&\mathrm{N}\left(0, \sum_{i=1}^{n}\hat{w}_i^2\sigma_i^2\right) + \|\hat{g}_0 - g_0\| \cdot \sup_{\|g\|=1}\left|\sum_{T_i=1}(2T_i-1)\hat{w}_i g(X_i)\right|.
\end{aligned}
\tag{7.6}
$$

The last supremum term is estimated along the forward stepwise or boosting algorithm. For kernel regression, it can also be estimated by an unbiased estimator in Gretton et al. (2012, Lemma 6).

To use (7.6), we need to find sample estimates or upper bounds of σ_i^2 and $\|\hat{g}_0 - g_0\|$. If we assume homoskedastic Gaussian noise $\sigma_i^2 = \sigma^2$, then

$$
n\sigma^2 + \|\hat{g}_0 - g_0\|^2 = \text{prediction error}(\hat{g}_0, \{X_i, i = 1, \ldots, n\}).
$$

The prediction error of $\hat{g}_0$ can be estimated relatively easily (e.g. by cross validation). Let $\mathrm{MMD}(\hat{w})$ denote the maximum mean discrepancy

$$
\mathrm{MMD}(\hat{w}) = \sup_{\|g\|=1}\left|\sum_{T_i=1}(2T_i-1)\hat{w}_i g(X_i)\right|. \tag{7.7}
$$

This is a quantity determined by when our adaptive procedure of propensity score estimation stops. Our discussion above motivates the following two model selection criteria:

1. Minimize the following upper bound of the MSE

$$\begin{aligned}\mathrm{MSE}\left(\hat{\tau}_{\mathrm{DR}}\right) \leq & \sum_{i=1}^{n} \hat{w}_i^2 \sigma^2 + \mathrm{MMD}(\hat{w})^2 \|\hat{g}_0 - g_0\|^2 \\ \leq & \max\left(\frac{1}{n}\sum_{i=1}^{n} \hat{w}_i^2,\ \mathrm{MMD}(\hat{w})^2\right) \cdot \text{prediction error}(\hat{g}_0).\end{aligned}$$

2. Minimize the length of the (conservative) level-α confidence interval

$$\begin{aligned}|\mathrm{CI}(\hat{\tau}_{\mathrm{DR}})| =& 2 z_{1-\frac{\alpha}{2}} \sqrt{\sum_{i=1}^{n} \hat{w}_i^2 \sigma^2} + 2\|\hat{g}_0 - g_0\| \mathrm{MMD}(\hat{w}) \\ \leq & 2\sqrt{z_{1-\frac{\alpha}{2}}^2 \left(\frac{1}{n}\sum_{i=1}^{n} \hat{w}_i^2\right) + \mathrm{MMD}(\hat{w})^2} \cdot \sqrt{\text{prediction error}(\hat{g}_0)}.\end{aligned}$$

   The last inequality is due to Cauchy-Schwarz.

Compared to selecting model according to some covariate balance measure as suggested in Equation (6.25), the procedure described in this Section also takes the bias-variance tradeoff into account.

Part III

INFERRING MULTIPLE EFFECTS

Chapter 8

A Common Confounding Problem

In the last Part of this thesis, the focus will be shifted to inferring multiple (e.g. thousands of) causal effects simultaneously. The single-effect inference discussed in Part II is built on the unconfoundedness (ignorability) assumption. However, it is impossible to empirically validate this assumption because it involves joint distribution of the counterfactuals. This Part presents an alternative approach without the unconfoundedness assumption. This approach requires multiple confounded effects to be revealed at the same and to be confounded in a structured way. When this is the case, *sparsity* can be used to identify the true effects in absence of ignorability. Chapters in this Part are based on Wang, Zhao, Hastie, and Owen (2015) and Song and Zhao (2016).

8.1 Linear model with latent variables

We start with a general linear model with latent variables and discuss what confounding means in this model. Let $Y \in \mathbb{R}^{n\times p}$ be an observed response matrix and $X \in \mathbb{R}^{n\times d}$ be some predictors. Consider the following linear model between X and Y:

$$Y_{n\times p} = X_{n\times d}\, \alpha^T_{p\times d} + Z_{n\times r}\, \beta^T_{p\times r} + E_{n\times p}. \tag{8.1a}$$

Here Z contains r unmeasured factors or latent variables and E is the noise matrix:

$$E \perp\!\!\!\perp (X, Z),\ E \sim \mathrm{MN}(0, I_n, \Sigma). \tag{8.1b}$$

In many applications (two detailed later in Sections 8.2 and 8.3), we are interested in only certain components of α, so it is useful to differential between primary variables $X_1 \in \mathbb{R}^{n\times d_1}$ and nuisance variables $X_0 \in \mathbb{R}^{n\times d_0}$ satisfying $X = (X_0, X_1)$ and $d_0 + d_1 = d$, and the coefficient vector α is correspondingly splitted as $\alpha = (\alpha_0, \alpha_1)$. We are only interested in inferring the primary coefficients α_1.

The latent variables Z play a crucial role in the inference of α_1. To see this, let's first ignore Z and consider the standard approach that runs a least squares (OLS) regression for each column of Y. This gives an unbiased estimate of the marginal effects of X. Are these marginal effects the same as the α in (8.1)? The answer depends on the relationship between the confounders Z and the primary variable X_1. We assume a linear relationship

$$Z = X\gamma^T + W,\ W \sim \mathrm{MN}(0, I_n, I_r),\ W \perp\!\!\!\perp X. \tag{8.1c}$$

If we plug (8.1c) into (8.1a), it is easy to see that the marginal effects of X_1 (let's call them τ_1) and the primary effects α_1 satisfy

$$\tau = \alpha_1 + \beta\gamma_1. \tag{8.2}$$

Here we partition $\gamma \in \mathbb{R}^{p\times d}$ into $\gamma = (\gamma_0, \gamma_1)$. Therefore $\tau_1 = \alpha_1$ if and only if $\beta\gamma_1 = 0$. This is the "unconfounded" case because the OLS estimate is unbiased for α_1. Otherwise we shall call the problem "confounded".

The parameters in the model equation (8.1) are $\alpha \in \mathbb{R}^{p\times d}$, which contain the primary effects we are interested in, $\beta \in \mathbb{R}^{p\times r}$, $\gamma \in \mathbb{R}^{r\times d}$, and $\Sigma \in \mathbb{R}^{p\times p}$. We assume Σ is diagonal $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, so the noise of different outcome variables are independent but possibly heteroskedastic.

In (8.1), X_i is not required to be Gaussian or even continuous. For example, a binary or categorical variable after normalization also meets this assumption. The parameter vector γ measures how severely the data are confounded. For a more intuitive interpretation, consider a simple case that $X = X_1 \in \mathbb{R}^{n\times 1}$ is one dimensional and an oracle procedure of estimating α when the confounders Z in equation (8.1a) are observed. The best linear unbiased estimator in this case is the ordinary least squares $(\hat{\alpha}_j^{\text{OLS}}, \hat{\beta}_j^{\text{OLS}})$, whose variance is $\sigma_j^2 \text{Var}(X_i, Z_i)^{-1}/n$. Using equation (8.1c), it is easy to show that $\text{Var}(\hat{\alpha}_j^{\text{OLS}}) = (1 + \|\gamma\|_2^2)\sigma_j^2/n$ and $\text{Cov}(\hat{\alpha}_j^{\text{OLS}}, \hat{\alpha}_k^{\text{OLS}}) = 0$ for $j \neq k$. In summary,

$$\text{Var}(\hat{\alpha}^{\text{OLS}}) = \frac{1}{n}(1 + \|\gamma\|_2^2)\Sigma. \tag{8.3}$$

Notice that in the unconfounded linear model in which $Z = 0$, the variance of the OLS estimator of α is Σ/n. Therefore, $1 + \|\gamma\|_2^2$ represents the relative loss of efficiency when we add observed variables Z to the regression which are correlated with X. In Section 9.3.2, we show that the oracle efficiency (8.3) can be asymptotically achieved even when Z is unobserved.

At this point the linear model (8.1) may seem a little abstract. In the following two Sections, two distinct applications are described to motivate the confounding problem in this model. Then Section 8.4 discusses the connection of model (8.1) to the linear structural equation model described in Chapter 4.

8.2 Example 1: Batch effects in microarray experiments

Our first example is multiple hypothesis testing in genomics. In this example, the response matrix Y is the gene expression levels. Each row of Y is a sample (patient or cell) and each column of Y is a gene. Typically the primary variable X_1 is some condition or treatment indicator, and we want to know which genes are related to or affected by this treatment. The nuisance variable X_0 usually includes an intercept term.

In the simplest genomics testing, each column of Y is regressed on $X = (X_0, X_1)$ and a t-statistic corresponding to X_1 is computed. Then we can apply a multiple testing correction procedure (e.g. Bonferroni correction, Benjamini-Hochberg procedure) to control the familywise error rate (FWER) or false discovery rate (FDR). This is the standard practice in many genomics studies.

Traditionally the tests are assumed to be independent of each other, so the multiple testing errors can be easily controlled. Recent years have witnessed an extensive investigation of multiple hypothesis, ranging from permutation tests (Tusher et al., 2001, Korn et al., 2004), positive dependence (Benjamini and Yekutieli, 2001), weak dependence (Storey et al., 2004, Clarke and Hall, 2009), accuracy calculation under dependence (Owen, 2005, Efron, 2007) to mixture models (Efron, 2010, Sun and Cai, 2009) and latent factor models (Fan et al., 2012, Fan and Han, 2013, Lan and Du, 2014). Many of these works provide theoretical guarantees for FDR control under the assumption that the individual test statistics are valid and may even be correlated.

Many of the listed references above can be understood through the unconfounded scenario in (8.1) where $\gamma = 0$. In this case, the OLS estimate of α_1 is unbiased as mentioned earlier, but its sample variance is larger than Σ (for simplicity we assume $\Sigma = \sigma^2 I$ is homoskedastic) due to the variability of the additional latent variables. Therefore the regression t-statistics are less efficient than the oracle case that Z is known. More importantly, the statistics for different genes are dependent with each other, so there is no guarantee that the Benjamini-Hochberg procedure can control FDR. To correct for this, it is useful to include the latent variables in the analysis. The surrogate variable analysis (SVA) of Leek and Storey (2007, 2008) takes this approach and is quite successful in practice.

A more challenging problem is the confounded scenario, where $\gamma \neq 0$ (more precisely $\beta\gamma \neq 0$). In this case, the OLS estimate of α_1 is biased and the so do the corresponding t-statistics. Therefore this problem is fundamentally different from the literature in the previous paragraph and poses an immediate threat to the reproducibility of the discoveries.

To summarize the existing approaches to handle dependent/confounded multiple hypothesis testing, Table 8.1 summarizes some related publications with more detailed discussion in Section 9.6.

Next we use three real microarray datasets to illustrate the confounding problem.:

- The first dataset (Singh et al., 2011) tries to identify candidate genes associated

	Noise conditional on latent factors	
	Independent	Correlated
Positive or weak dependence	Benjamini and Yekutieli (2001) Storey et al. (2004) Clarke and Hall (2009)	
Unconfounding factors or other structure	Friguet, Kloareg, and Causeur (2009) Desai and Storey (2012)	Fan, Han, and Gu (2012) Lan and Du (2014) *Discussed in Section 9.6.1*
Confounding factors	Leek and Storey (2007, 2008) Gagnon-Bartsch and Speed (2012) Gagnon-Bartsch, Jacob, and Speed (2013) Sun, Zhang, and Owen (2012) *Studied in Sections 9.3 and 9.5* *Discussed in Section 9.6.2*	*Discussed in Section 9.6.3* *(future research)*

Table 8.1: Selected literature in multiple hypothesis testing under dependence. The categorization is partially subjective as some authors do not use exactly the same terminology as us.

with the extent of emphysema and can be downloaded from the GEO database (Series GSE22148). We preprocessed the data using the standard Robust Multi-array Average (RMA) approach (Irizarry et al., 2003). The primary variable of interest is the severity (moderate or severe) of the Chronic Obstructive Pulmonary Disease (COPD). The dataset also include age, gender, batch and date of the 143 sampled patients which are served as nuisance covariates.

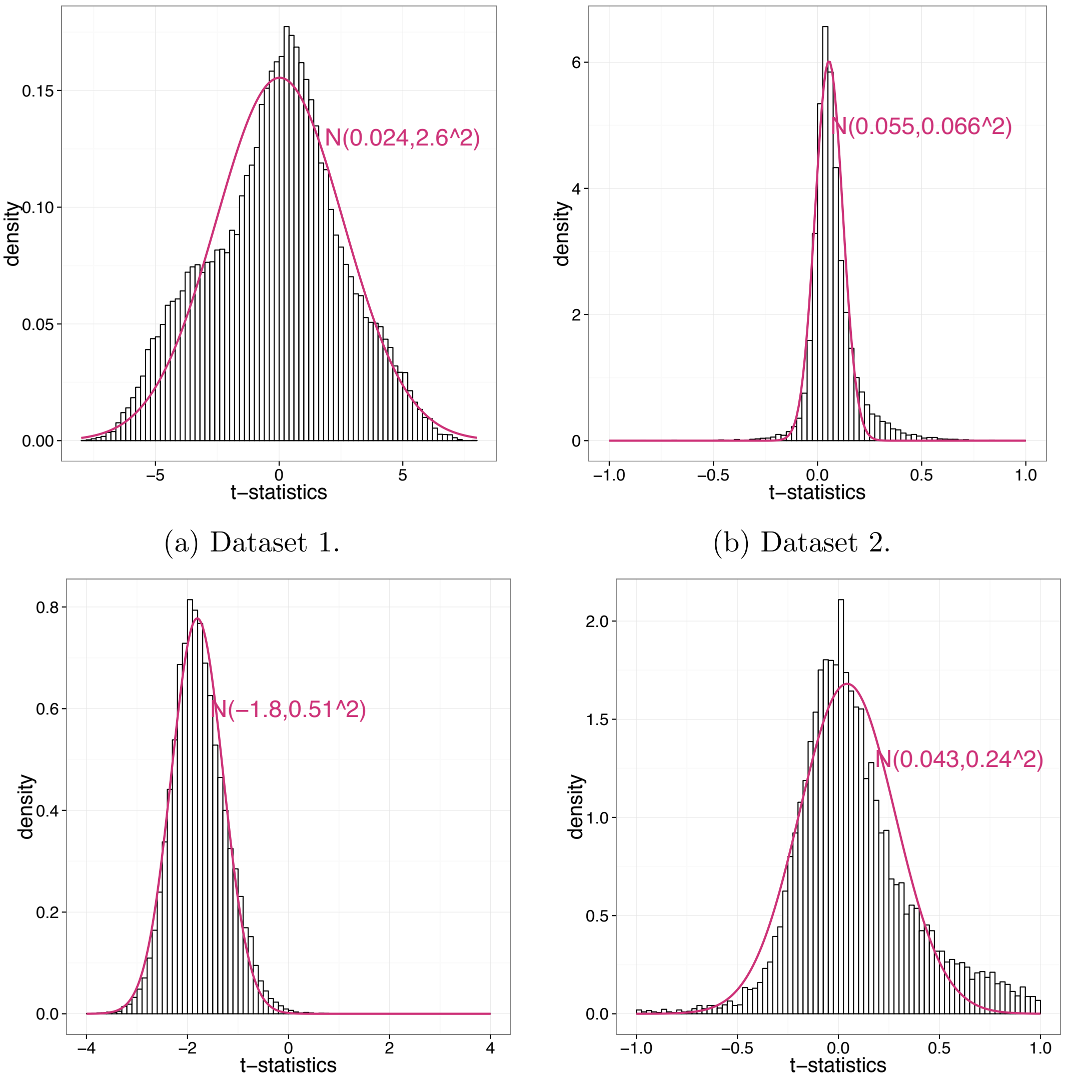
- The second and third datasets are taken from Gagnon-Bartsch et al. (2013) where they used them to compare RUV methods with other methods such as SVA and LEAPP. The original scientific studies are Vawter et al. (2004) and Blalock et al. (2004), respectively. The primary variable of interest is gender in both datasets, though the original objective in Blalock et al. (2004) is to identify genes associated with Alzheimer's disease. Gagnon-Bartsch et al. (2013) switch the primary variable to gender in order to have a gold standard: the differentially expressed genes should mostly come from or relate to the X or Y chromosome. We follow their suggestion and use this standard to study the performance of our RR estimator. In addition, as the first COPD dataset also

> contains gender information of the samples, we apply this suggestion and use gender as the primary variable for the COPD data as a supplementary dataset. Notice that the second dataset comes with batch and microarray platform labels (possible confounders). However, this dataset has repeated samples from the same patients but the individual information is lost.

In Figure 8.1, we plot the histogram of t-statistics of a simple linear model that regresses the gene expression on the variable of interest (disease status for the first and gender for the second and third datasets). The histograms clearly depart from the approximate theoretical null distribution $\mathrm{N}(0,1)$. The bulk of the test statistics can be skewed (Figures 8.1a and 8.1b), overdispersed (Figure 8.1a), underdispersed (Figures 8.1b and 8.1d), or noncentered (Figure 8.1c). In these cases, neither the theoretical null $\mathrm{N}(0,1)$, nor even the empirical null as shown in the histograms, look appropriate for measuring significance. Schwartzman (2010) proved that a largely overdispersed histogram like Figure 8.1a cannot be explained by correlation alone, and is possibly due to the presence of confounding factors. The p-values of our test of confounding (Section 9.4.2) in Table 10.1 indicate that all the three datasets suffer from confounding latent factors.

The most widely noted confounding variables are batch effects. For example, Leek et al. (2010) described three possible batch effects: 1. a subset of experiments was run on Monday and another set on Tuesday; 2. two technicians were responsible for different subsets of the experiments; 3. two different lots of reagents, chips or instruments were used. When batch effects are correlated with an treatment/outcome of interest, they become confounding variables and lead to incorrect conclusions.

Other common sources of confounding in gene expression profiling include systematic ancestry differences (Price et al., 2006), environmental changes (Gasch et al., 2000, Fare et al., 2003) and surgical manipulation (Lin et al., 2006). See Lazar et al. (2013) for a survey. In many studies, especially for observational clinical research and human expression data, the latent factors, either genetic or technical, are confounded with primary variables of interest due to the observational nature of the studies and heterogeneity of samples (Ransohoff, 2005, Rhodes and Chinnaiyan, 2005). Similar confounding problems also occur in other high-dimensional datasets such as brain



(a) Dataset 1.

(b) Dataset 2.

(c) Dataset 3.

(d) Dataset 2 after known batch correction.

Figure 8.1: Dataset 1 is the emphysema dataset (Singh et al., 2011). Dataset 2 and 3 are from Gagnon-Bartsch et al. (2013). Histograms of regression t-statistics in three microarray studies show clear departure from the theoretical null distribution N$(0, 1)$. The mean and standard deviation of the normal approximation are obtained from the median and median absolute deviation of the statistics. See Figure 10.4 for the empirical distributions after confounder adjustment.

imaging (Schwartzman et al., 2008) and metabonomics (Craig et al., 2006).

8.3 Example 2: Unknown risk factors in stock market

Next we turn to an example in an entirely different field—finance. Since the capital asset pricing model (CAPM) was first introduced in the 1960s (Markowitz, 1952, Treynor, 1961, Sharpe, 1964, Lintner, 1965, Mossin, 1966), much of the empirical research tries to find systemic risk factors that describe stock returns. This amounts to find a linear model (8.1) for the stock returns without any unobserved variables. In this example, the response matrix Y is the stock returns (rows are time and columns are stocks) and the predictor X contains intercept and known systemic factors. As of 2016, one of the most widely accepted models is the Fama-French-Carhart (FFC) four factor model (Fama and French, 1992, Carhart, 1997), where the systemic risk factors are: 1. Market return minus Risk free return (Mkt-Rf); 2. Small market capitalization Minus Big (SMB); 3. High book-to-market ratio Minus Low (HML); 4. Momentum (MOM). In other words, the returns of a stock j is modeled by

$$Y_j = \alpha_0 + \alpha_1 X_{\text{Mkt-Rf}} + \alpha_2 X_{\text{SMB}} + \alpha_3 X_{\text{HML}} + \alpha_4 X_{\text{MOM}} + \epsilon_j. \tag{8.4}$$

This model explains over 90% of the diversified portfolios returns, which was a huge success in financial economics. In this factor model, α_0 is usually called the *alpha* or *risk-adjusted return* of the stock. Since Fama (1970)'s introduction of efficient market hypothesis, many academics believe financial markets are too efficient to allow for repeatedly earning positive alpha, unless by chance. Nevertheless, the risk-adjusted return is still widely used to evaluate mutual fund and portfolio manager performance. In fact, and rather surprisingly, some recent research find that the majority of mutual fund investors allocate their savings to funds who generated superior CAPM-alpha (rather than the FFC-alpha) in the past (Barber et al., 2014, Berk and Van Binsbergen, 2016).

The Fama-French-Carhart model (8.4) belongs to our general linear model (8.1), where the primary variable $X_1 = 1$, the nuisance variables

$$X_0 = (X_{\text{Mkt-Rf}}, X_{\text{SMB}}, X_{\text{HML}}, X_{\text{MOM}}),$$

and the latent variable Z is non-existent. Considering the vast amount of research after Fama and French (1992) and Carhart (1997) that tries to identify additional factors in (8.4), it is reasonable to postulate that (8.1) with latent factors may describe the stock returns better. More importantly, it is very likely that the latent factors have non-zero mean (i.e. "correlated" with the primary variable—intercept) and confound the alpha. Intuitively, this means that the stock return may depend on other unknown systemic risk factors which may have positive mean return. A mutual fund with large positive CAPM-alpha or FFC-alpha may load on these unknown factors, and the true risk-adjusted alpha could be much smaller or even negative.

8.4 Connection to linear structural equation model

When model (8.1) is interpreted as the linear structural equations model in Chapter 4, our parameter of interest α is indeed the direct causal effect of X on Y (Pearl, 2009a). In fact, Gagnon-Bartsch et al. (2013, Section 3.3) use all the heuristics in structural equation models without realizing it. This partially motivates the proposal in the next Chapter.

It is not necessary to make the structural assumptions to use the linear model (8.1). For the microarray application described in Section 8.2, the model (8.1) is used to describe the marginal screening procedure commonly applied in high throughput data analysis. For the finance application described in (8.3), we are interested in the intercept, a somewhat non-causal parameter. The linear model (8.1) provides by far the clearest description of confounded effects in these applications (see Section 9.6 for some additional discussion). On the other hand, the asymptotic setting $(n, p \to \infty)$ and sparsity assumptions in the next Chapter are newcomers in structural equation modeling.

Notice that if the linear relationship between Z and X in (8.1c) is structural, the latent variables Z should be interpreted as mediators instead of confounders. However, in the linear setting they are almost indistinguishable and do not affect the inference. The practical interpretation should depend on the application. In microarray experiments, batch effects are commonly regarded as mediators, but there may exist other confounders as well.

The model (8.1) is also related to the common shock model that is widely used in actuarial science and economics (Bai and Li, 2014, e.g.). In the common shock model, both X and Y are modeled as linear functions of Z (Y also depends on X). When the common shock model is assumed structural, the common shocks Z are actually the confounders between X and Y.

Chapter 9

Cross-Sectional Regression After Factor Analysis

This Chapter proposes a two-step procedure to solve the confounding problem introduced in the last Chapter. The statistical method is implemented in an R package `cate` (short for "confounder ddjusted testing and estimation") that is available on CRAN. The reader is referred to the supplementary file of Wang, Zhao, Hastie, and Owen (2015) for technical proofs of the theorems in this Chapter.

9.1 Rotation

For simplicity, we start with the case that $X = X_1 \in \mathbb{R}^{n \times 1}$, so the only known dependent variable is the primary variable of interest. In Section 9.5, the statistical method and theory are extended to multiple regression, the original form in (8.1).

Following Sun et al. (2012), we introduce a transformation of the data to make the confounding problem clearer. Consider the Householder rotation matrix $Q^T \in \mathbb{R}^{n \times n}$ such that $Q^T X = \|X\|_2 e_1 = (\|X\|_2, 0, 0, \dots, 0)^T$. Left-multiplying Y by Q^T, we get $\tilde{Y} = Q^T Y = \|X\|_2 e_1 \alpha^T + \tilde{Z}\beta^T + \tilde{E}$, where

$$\tilde{Z} = Q^T Z = Q^T (X\gamma^T + W) = \|X\|_2 e_1 \gamma^T + \tilde{W}, \tag{9.1}$$

and $\tilde{W} = Q^T W \overset{d}{=} W$, $\tilde{E} = Q^T E \overset{d}{=} E$. As a consequence, the first and the rest of the rows of $\tilde{Y}$ are

$$\tilde{Y}_1 = \|X\|_2 \alpha^T + \tilde{Z}_1 \beta^T + \tilde{E}_1 \sim \mathrm{N}(\|X\|_2(\alpha + \beta\gamma)^T, \beta\beta^T + \Sigma), \tag{9.2}$$

$$\tilde{Y}_{-1} = \tilde{Z}_{-1} \beta^T + \tilde{E}_{-1} \sim \mathrm{MN}(0, I_{n-1}, \beta\beta^T + \Sigma). \tag{9.3}$$

Here $\tilde{Y}_1$ is a $1 \times p$ vector, $\tilde{Y}_{-1}$ is a $(n-1) \times p$ matrix, and the distributions are conditional on X.

The parameters α and γ only appear in equation (9.2) (a finite sample version of equation (8.2)), so their inference (step 1 in our procedure) can be completely separated from the inference of β and Σ (step 2 in our procedure). In fact, $\tilde{Y}_1 \perp\!\!\!\perp \tilde{Y}_{-1} | X$ because $\tilde{E}_1 \perp\!\!\!\perp \tilde{E}_{-1}$, so the two steps use mutually independent information. This in turn greatly simplifies the theoretical analysis.

We intentionally use the symbol Q to resemble the QR decomposition of X. In Section 9.5 we show how to use the QR decomposition to separate the primary effects from confounder and nuisance effects when X has multiple columns. Using the same notation, we discuss how SVA and RUV decouple the problem in a slightly different manner in Section 9.6.2.

9.2 Identifiability

Let $\theta = (\alpha, \beta, \gamma, \Sigma)$ be all the parameters and Θ be the parameter space. Without any constraint, the model equation (8.1) is not identifiable. In this Section, we show how to restrict the parameter space Θ to ensure identifiability.

9.2.1 Identifiability of β

Equation (9.3) is just the exploratory factor analysis model, thus β can be easily identified up to some rotation under some mild conditions. Here we assume a classical sufficient condition for the identification of β (Anderson and Rubin, 1956, Theorem 5.1)

Lemma 9.1. *Let $\Theta = \Theta_0$ be the parameter space such that*

1. *If any row of β is deleted, there remain two disjoint submatrices of β of rank r;*

2. *$\frac{1}{p}\beta^T\Sigma^{-1}\beta$ is diagonal and the diagonal elements are distinct, positive, and arranged in decreasing order.*

Then β and Σ are identifiable in the model equation (8.1).

In Lemma 9.1, condition (1) requires that $p \geq 2r+1$. Condition (1) identifies β up to a rotation which is sufficient to identify α. To see this, we can reparameterize β and γ to βU and $U^T\gamma$ using an $r \times r$ orthogonal matrix U. This reparameterization does not change the distribution of $\tilde{Y}_1$ in equation (9.2) if α remains the same. Condition (2) identifies the rotation uniquely but is not necessary for the theoretical analysis in later sections.

9.2.2 Identifiability of α

The parameters α and γ cannot be identified from (9.2) because they have in total $p + r$ parameters while $\tilde{Y}_1$ is a length p vector. If we write $\mathcal{P}_\beta$ and $\mathcal{P}_{\beta^\perp}$ as the projection onto the column space and orthogonal space of β so that $\alpha = \mathcal{P}_\beta\alpha + \mathcal{P}_{\beta^\perp}\alpha$, it is impossible to identify $\mathcal{P}_\beta\alpha$ from equation (9.2).

This suggests that we need to further restrict the parameter space Θ. We will reduce the degrees of freedom by restricting at least r entries of α to equal 0. We consider two different sufficient conditions to identify α:

Negative control $\Theta_1 = \{(\alpha, \beta, \gamma, \Sigma) : \alpha_{\mathcal{C}} = 0,\ \operatorname{rank}(\beta_{\mathcal{C}}) = r\}$ for a known negative control set $|\mathcal{C}| \geq r$.

Sparsity $\Theta_2(s) = \{(\alpha, \beta, \gamma, \Sigma) : \|\alpha\|_0 \leq \lfloor (p-s)/2 \rfloor,\ \operatorname{rank}(\beta_{\mathcal{C}}) = r,\ \forall \mathcal{C} \subset \{1, \ldots, p\},\ |\mathcal{C}| = s\}$ for some $r \leq s \leq p$.

Proposition 9.1. *If $\Theta = \Theta_0 \cap \Theta_1$ or $\Theta = \Theta_0 \cap \Theta_2(s)$ for some $r \leq s \leq p$, the parameters $\theta = (\alpha, \beta, \gamma, \Sigma)$ in the model equation (8.1) are identifiable.*

Proof. Since $\Theta \subset \Theta_0$, we know from Lemma 9.1 that β and Σ are identifiable. Now consider two combinations of parameters $\theta^{(1)} = (\alpha^{(1)}, \beta, \gamma^{(1)}, \Sigma)$ and $\theta^{(2)} = (\alpha^{(2)}, \beta, \gamma^{(2)}, \Sigma)$ both in the space Θ and inducing the same distribution in the model equation (8.1), i.e. $\alpha^{(1)} + \beta\gamma^{(1)} = \alpha^{(2)} + \beta\gamma^{(2)}$.

Let $\mathcal{C}$ be the set of indices such that $\alpha^{(1)}_{\mathcal{C}} = \alpha^{(2)}_{\mathcal{C}} = 0$. If $\Theta = \Theta_0 \cap \Theta_1$, we already know $|\mathcal{C}| \geq r$. If $\Theta = \Theta_0 \cap \Theta_2(s)$, it is easy to show that $|\mathcal{C}| \geq s$ is also true because both $\alpha^{(1)}$ and $\alpha^{(2)}$ have at most $\lfloor (p-s)/2 \rfloor$ nonzero entries. Along with the rank constraint on $\beta_{\mathcal{C}}$, this implies that $\beta_{\mathcal{C}}\gamma^{(1)} = \beta_{\mathcal{C}}\gamma^{(2)}$. However, the conditions in Θ_1 and Θ_2 ensure that $\beta_{\mathcal{C}}$ has full rank, so $\gamma^{(1)} = \gamma^{(2)}$ and hence $\alpha^{(1)} = \alpha^{(2)}$. □

We make four remarks regarding the identification conditions in Proposition 9.1:

Remark 1. The condition (2) in Lemma 9.1 that uniquely identifies β is not necessary for the identification of β. This is because for any set $|\mathrm{C}| \geq r$ and any orthogonal matrix $U \in \mathbb{R}^{r\times r}$, we always have $\mathrm{rank}(\beta_{\mathcal{C}}) = \mathrm{rank}(\beta_{\mathcal{C}})U$. Therefore β only needs to be identified up to a rotation.

Remark 2. Almost all dense matrices of $\beta \in \mathbb{R}^{p\times r}$ satisfy the conditions. However, for $\Theta_2(s)$ the sparsity of β allowed depends on the sparsity of β. The condition $\Theta_2(s)$ rules out some too sparse β. In this case, one may consider using confirmatory factor analysis instead of exploratory factor analysis to model the relationship between confounders and outcomes. For some recent identification results in confirmatory factor analysis, see Grzebyk et al. (2004), Kuroki and Pearl (2014).

Remark 3. The maximum allowed $\|\alpha\|_0$ in Θ_2, $\lfloor (p-r)/2 \rfloor$, is exactly the maximum breakdown point of a robust regression with p observations and r fixed predictors of full rank (Maronna et al., 2006, Section 4.6). Indeed, we use robust regression to estimate α in this case in Section 9.3.2.

Remark 4. To the best of our knowledge, the only existing literature that explicitly addresses the identifiability issue for the confounder problem is Sun (2011, Chapter 4.2), where the author gives sufficient conditions for *local* identifiability of α by viewing equation (8.1a) as a "sparse plus low rank" matrix decomposition problem. See Chandrasekaran et al. (2012, Section 3.3) for a more general discussion of the local and global identifiability for this problem. Local identifiability refers to identifiability of the parameters in a neighborhood of the true values. In contrast, the conditions in Proposition 9.1 ensure that α is *globally* identifiable within the restricted parameter space.

9.3 Estimation

We consider a two-step procedure, called cross-section regression (Section 9.3.2) after factor analysis (Section 9.3.1), to make statistical inference for the model (8.1).

9.3.1 Estimating β and Σ

The most popular approaches for factor analysis are principal component analysis (PCA) and maximum likelihood (ML). Bai and Ng (2002) derived a class of estimators of r by principal component analysis using various information criteria. The estimators are consistent under Assumption 9.3 in this section and some additional technical assumptions in Bai and Ng (2002). Due to this reason, we assume the number of confounding factors r is known in this section. See Owen and Wang (2016, Section 3) for a comprehensive literature review of choosing r in practice.

We are most interested in the asymptotic behavior of factor analysis when both $n, p \to \infty$. In this case, PCA cannot consistently estimate the noise variance Σ (Bai and Li, 2012). For theoretical analysis, we use the quasi maximum likelihood estimate in Bai and Li (2012) to get $\hat{\beta}$ and $\hat{\Sigma}$. This estimator is called "quasi"-MLE because it treats the factors $\tilde{Z}_{-1}$ as fixed quantities. Since the confounders Z in our model equation (8.1) are random variables, we introduce a rotation matrix $R \in \mathbb{R}^{r \times r}$ and let $\tilde{Z}_{-1}^{(0)} = \tilde{Z}_{-1}(R^{-1})^T$, $\beta^{(0)} = \beta R$ be the target factors and factor loadings that are studied in Bai and Li (2012).

To make $\tilde{Z}_{-1}^{(0)}$ and $\beta^{(0)}$ identifiable, Bai and Li (2012) consider five different identification conditions. However, the parameter of interest in model equation (8.1) is α instead of β or $\beta^{(0)}$. As we have discussed in Section 9.2.2, we only need the column space of β to estimate α, which gives us some flexibility of choosing the identification condition. In our theoretical analysis we use the third condition (IC3) in Bai and Li (2012), which imposes the constraints that $(n-1)^{-1}(\tilde{Z}_{-1}^{(0)})^T\tilde{Z}_{-1}^{(0)} = I_r$ and $p^{-1}\tilde{\beta}^{(0)T}\beta^{-1}\beta^{(0)}$ is diagonal. Therefore, the rotation matrix R satisfies $RR^T = (n-1)^{-1}\tilde{Z}_{-1}^T\tilde{Z}_{-1}$.

The quasi-loglikelihood being maximized in Bai and Li (2012)is

$$-\frac{1}{2p}\log\det\left(\beta^{(0)}(\beta^{(0)})^T+\Sigma\right) - \frac{1}{2p}\mathrm{tr}\left\{S\left[\beta^{(0)}(\beta^{(0)})^T+\Sigma\right]^{-1}\right\} \tag{9.4}$$

where S is the sample covariance matrix of $\tilde{Y}_{-1}$.

The theoretical results in this section rely heavily on recent findings in Bai and Li (2012). They use these three assumptions.

Assumption 9.1. *The noise matrix E follows the matrix normal distribution $E \sim \mathrm{MN}(0, I_n, \Sigma)$ and Σ is a diagonal matrix.*

Assumption 9.2. *There exists a positive constant D such that $\|\beta_j\|_2 \le D$, $D^{-2} \le \sigma_j^2 \le D^2$ for all j, and the estimated variances $\hat{\sigma}_j^2 \in [D^{-2}, D^2]$ for all j.*

Assumption 9.3. *The limits $\lim_{p\to\infty} p^{-1}\beta^T\Sigma^{-1}\beta$ and $\lim_{p\to\infty}\sum_{j=1}^p \sigma_j^{-4}(\beta_j\otimes\beta_j)(\beta_j^T\otimes\beta_j^T)$ exist and are positive definite matrices.*

Bai and Li (2012) prove the consistency and asymptotic normality of $\hat{\beta}$ and $\hat{\Sigma}$:

Lemma 9.2. *Under Assumptions 9.1 to 9.3, the maximizers $\hat{\beta}$ and $\hat{\Sigma}$ of the quasi-loglikelihood (9.4) satisfy*

$$\sqrt{n}(\hat{\beta}_j - \beta_j^{(0)}) \xrightarrow{d} \mathrm{N}(0, \sigma_j^2 I_r), \quad \textit{and} \quad \sqrt{n}(\hat{\sigma}_j^2 - \sigma_j^2) \xrightarrow{d} \mathrm{N}(0, 2\sigma_j^4).$$

Here we prove uniform convergence of the estimated factors and noise variances based on the proof in Bai and Li (2012), which are needed to prove subsequently results for $\hat{\alpha}$.

Lemma 9.3. *Under Assumptions 9.1 to 9.3, for any fixed index set S with finite cardinality,*

$$\sqrt{n}(\hat{\beta}_S - \beta_S^{(0)}) \xrightarrow{d} \mathrm{N}(0, \Sigma_S \otimes I_r) \tag{9.5}$$

where Σ_S is the noise covariance matrix of the variables in S. Further, if there exists $k > 0$ such that $p/n^k \to 0$ when $p \to \infty$, then

$$\max_{j=1,2,\cdots,p} |\hat{\sigma}_j^2 - \sigma_j^2| = O_p(\sqrt{\log p/n}), \quad \max_{j=1,2,\cdots,p} |\hat{\beta}_j - \beta_j^{(0)}| = O_p(\sqrt{\log p/n}), \text{ and} \tag{9.6}$$

$$\max_{j=1,2,\cdots,p} \left| \hat{\beta}_j - \beta_j^{(0)} - \frac{1}{n-1} \sum_{i=2}^{n} \tilde{Z}_i^{(0)} \tilde{E}_{ij} \right| = o_p(n^{-\frac{1}{2}}). \tag{9.7}$$

Remark 5. Assumption 9.2 is Assumption D from Bai and Li (2012). It requires that the diagonal elements of the quasi-MLE $\hat{\Sigma}$ be uniformly bounded away from zero and infinity. We would prefer boundedness to be a consequence of some assumptions on the distribution of the data, but at present we are unaware of any other results like Lemma 9.2 which do not use this assumption. In practice, the quasi-likelihood problem (9.4) is commonly solved by the Expectation-Maximization (EM) algorithm. Similar to Bai and Li (2012, 2014), we do not find it necessary to impose an upper or lower bound for the parameters in the EM algorithm in the numerical experiments.

9.3.2 Estimating α and γ

The estimation of α and γ is based on the first row of the rotated outcome $\tilde{Y}_1$ in (9.2), which can be rewritten as

$$\tilde{Y}_1^T/\|X\|_2 = \alpha + \beta(\gamma + \tilde{W}_1/\|X\|_2) + \tilde{E}_1^T/\|X\|_2 \tag{9.8}$$

where $\tilde{W}_1 \sim \mathrm{N}(0, I_p)$ is from equation (9.1) and $\tilde{W}_1$ is independent of $\tilde{E}_1 \sim \mathrm{N}(0, \Sigma)$. Note that $\tilde{Y}_1/\|X\|_2$ is proportional to the sample covariance between Y and X.

To reduce variance, we choose to estimate (9.8) conditional on $\tilde{W}_1$. Also, to use the results in Lemma 9.2, we replace β by $\beta^{(0)}$. Then, we can rewrite (9.8) as

$$\tilde{Y}_1^T/\|X\|_2 = \alpha + \beta^{(0)}\gamma^{(0)} + \tilde{E}_1^T/\|X\|_2 \tag{9.9}$$

where $\beta^{(0)} = \beta R$ and $\gamma^{(0)} = R^{-1}(\gamma + \tilde{W}_1/\|X\|_2)$. Notice that the random R only depends on $\tilde{Y}_{-1}$ and thus is independent of $\tilde{Y}_1$. In the proof of the results in this section, we first consider the estimation of α for fixed $\tilde{W}_1$, R and X, and then show the asymptotic distribution of $\hat{\alpha}$ indeed does not depend on $\tilde{W}_1$, R or X, and thus also holds unconditionally.

All the methods described in this section first try to find a good estimator $\hat{\gamma}$, then use $\hat{\alpha} = \tilde{Y}_1^T/\|X\|_2 - \hat{\beta}\hat{\gamma}$ to estimate α. Equation (9.9) can be viewed as a cross-sectional regression, meaning the "observations" $\tilde{Y}_1$ correspond to the columns of the original data matrix Y. In other words, we treat the marginal effects of X for each column of Y as the response, and try to explain them by the common pattern $\hat{\beta}$ we obtained from factor analysis. Whatever left unexplained is perhaps the real/direct effect of X.

Negative control scenario

If we know a set $\mathcal{C}$ such that $\alpha_{\mathcal{C}} = 0$ (so $\Theta \subset \Theta_1$), then $\tilde{Y}_1$ can be correspondingly separated into two parts:

$$\begin{aligned} \tilde{Y}_{1,\mathcal{C}}^T/\|X\|_2 &= \beta_{\mathcal{C}}^{(0)}\gamma^{(0)} + \tilde{E}_{1,\mathcal{C}}^T/\|X\|_2, \quad \text{and} \\ \tilde{Y}_{1,-\mathcal{C}}^T/\|X\|_2 &= \alpha_{-\mathcal{C}} + \beta_{-\mathcal{C}}^{(0)}\gamma^{(0)} + \tilde{E}_{1,-\mathcal{C}}^T/\|X\|_2. \end{aligned} \tag{9.10}$$

The number of negative controls $|\mathcal{C}|$ may grow as $p \to \infty$. We impose an additional assumption on the latent factors of the negative controls.

Assumption 9.4. $\lim_{p\to\infty} |\mathcal{C}|^{-1}\beta_{\mathcal{C}}^T\Sigma_{\mathcal{C}}^{-1}\beta_{\mathcal{C}}$ *exists and is positive definite.*

We consider the following negative control (NC) estimator where $\gamma^{(0)}$ is estimated by generalized least squares:

$$\hat{\gamma}^{\mathrm{NC}} = (\hat{\beta}_{\mathcal{C}}^T\hat{\Sigma}_{\mathcal{C}}^{-1}\hat{\beta}_{\mathcal{C}})^{-1}\hat{\beta}_{\mathcal{C}}^T\hat{\Sigma}_{\mathcal{C}}^{-1}\tilde{Y}_{1,\mathcal{C}}^T/\|X\|_2, \ \text{and} \tag{9.11}$$

$$\hat{\alpha}_{-\mathcal{C}}^{\mathrm{NC}} = \tilde{Y}_{1,-\mathcal{C}}^T/\|X\|_2 - \hat{\beta}_{-\mathcal{C}}\hat{\gamma}^{\mathrm{NC}}. \tag{9.12}$$

This estimator matches the RUV-4 estimator of Gagnon-Bartsch et al. (2013) except that it uses quasi-maximum likelihood estimates of Σ and β instead of using PCA, and generalized linear squares instead of ordinary linear squares regression. The details are in Section 9.6.2.

Our goal is to show consistency and asymptotic variance of $\hat{\alpha}^{\mathrm{NC}}_{-\mathcal{C}}$. Let $\Sigma_{\mathcal{C}}$ represents the noise covariance matrix of the variables in $\mathcal{C}$. We then have

Theorem 9.1. *Under Assumptions 9.1 to 9.4, if $n, p \to \infty$ and $p/n^k \to 0$ for some $k > 0$, then for any fixed index set $\mathcal{S}$ with finite cardinality and $\mathcal{S} \cap \mathcal{C} = \emptyset$, we have*

$$\sqrt{n}(\hat{\alpha}^{\mathrm{NC}}_{\mathcal{S}} - \alpha_{\mathcal{S}}) \xrightarrow{d} \mathrm{N}(0, (1 + \|\gamma\|_2^2)(\Sigma_{\mathcal{S}} + \Delta_{\mathcal{S}})) \tag{9.13}$$

where $\Delta_{\mathcal{S}} = \beta_{\mathcal{S}}(\beta_{\mathcal{C}}^T \Sigma_{\mathcal{C}}^{-1} \beta_{\mathcal{C}})^{-1} \beta_{\mathcal{S}}^T$.

If in addition, $|\mathcal{C}| \to \infty$, the minimum eigenvalue of $\beta_{\mathcal{C}}^T \Sigma_{\mathcal{C}}^{-1} \beta_{\mathcal{C}} \to \infty$ by Assumption 9.4, then the maximum entry of $\Delta_{\mathcal{S}}$ goes to 0. Therefore in this case

$$\sqrt{n}(\hat{\alpha}^{\mathrm{NC}}_{\mathcal{S}} - \alpha_{\mathcal{S}}) \xrightarrow{d} \mathrm{N}(0, (1 + \|\gamma\|_2^2)\Sigma_{\mathcal{S}}). \tag{9.14}$$

The asymptotic variance in (9.14) is the same as the variance of the oracle least squares in (8.3). Comparable oracle efficiency statements can be found in the econometrics literature (Bai and Ng, 2006, Wang et al., 2015). This is also the variance used implicitly in RUV-4 as it treats the estimated Z as given when deriving test statistics for α. When the number of negative controls is not too large, say $|\mathcal{C}| = 30$, the correction term $\Delta_{\mathcal{S}}$ is nontrivial and gives more accurate estimate of the variance of $\hat{\alpha}^{\mathrm{NC}}_{-\mathcal{C}}$. See Section 10.1 for more simulation results.

Sparsity scenario

When the zero indices in α are unknown but sparse (so $\Theta \subseteq \Theta_2$), the estimation of α and γ from $\tilde{Y}_1^T/\|X\|_2 = \alpha + \beta^{(0)}\gamma^{(0)} + \tilde{E}_1^T/\|X\|_2$ can be cast as a robust regression by viewing $\tilde{Y}_1^T$ as observations and $\beta^{(0)}$ as design matrix. The nonzero entries in α correspond to outliers in this linear regression.

The problem here has two nontrivial differences compared to classical robust regression. First, we expect some entries of α to be nonzero, and our goal is to make inference on the outliers; second, we don't observe the design matrix $\beta^{(0)}$ but only have its estimator $\hat{\beta}$. In fact, if $\alpha = 0$ and $\beta^{(0)}$ is observed, the ordinary least squares estimator of $\gamma^{(0)}$ is unbiased and has variance of order $1/(np)$, because the noise in equation (9.8) has variance $1/n$ and there are p observations. Our main conclusion is that $\gamma^{(0)}$ can still be estimated very accurately given the two technical difficulties.

Given a robust loss function ρ, we consider the following estimator:

$$\hat{\gamma}^{\mathrm{RR}} = \arg\min \sum_{j=1}^{p} \rho\left(\frac{\tilde{Y}_{1j}/\|X\|_2 - \hat{\beta}_j^T\gamma}{\hat{\sigma}_j}\right), \text{ and} \tag{9.15}$$

$$\hat{\alpha}^{\mathrm{RR}} = \tilde{Y}_1/\|X\|_2 - \hat{\beta}\hat{\gamma}^{\mathrm{RR}}. \tag{9.16}$$

For a broad class of loss functions ρ, estimating γ by equation (9.15) is equivalent to

$$(\hat{\gamma}^{\mathrm{RR}}, \tilde{\alpha}) = \arg\min_{\gamma,\alpha} \sum_{j=1}^{p} \frac{1}{\hat{\sigma}_j^2}(\tilde{Y}_{1j}/\|X\|_2 - \alpha_j - \hat{\beta}_j^T\gamma)^2 + P_\lambda(\alpha), \tag{9.17}$$

where $P_\lambda(\alpha)$ is a penalty to promote sparsity of α (She and Owen, 2011). However $\hat{\alpha}^{\mathrm{RR}}$ is not identical to $\tilde{\alpha}$, which is a sparse vector that does not have an asymptotic normal distribution. The LEAPP algorithm (Sun et al., 2012) uses the form (9.17). Replacing it by the robust regression equations (9.15) and (9.16) allows us to derive significance tests of $H_{0j} : \alpha_j = 0$.

We assume a smooth loss ρ for the theoretical analysis:

Assumption 9.5. *The penalty $\rho : \mathbb{R} \to [0, \infty)$ with $\rho(0) = 0$. The function $\rho(x)$ is non-increasing when $x \le 0$ and is non-decreasing when $x > 0$. The derivative $\psi = \rho'$ exists and $|\psi| \le D$ for some $D < \infty$. Furthermore, ρ is strongly convex in a neighborhood of* 0*.*

A sufficient condition for the local strong convexity is that $\psi' > 0$ exists in a neighborhood of 0. The next theorem establishes the consistency of $\hat{\gamma}^{\mathrm{RR}}$.

Theorem 9.2. *Under Assumptions 9.1 to 9.3 and 9.5, if $n, p \to \infty$, $p/n^k \to 0$ for some $k > 0$ and $\|\alpha\|_1/p \to 0$, then $\hat{\gamma}^{\mathrm{RR}} \xrightarrow{p} \gamma$. As a consequence, for any j, $\hat{\alpha}_j^{\mathrm{RR}} \xrightarrow{p} \alpha_j$.*

To derive the asymptotic distribution, we consider the estimating equation corresponding to (9.15). By taking the derivative of (9.15), $\hat{\gamma}^{\mathrm{RR}}$ satisfies

$$\Psi_{p,\hat{\beta},\hat{\Sigma}}(\hat{\gamma}^{\mathrm{RR}}) = \frac{1}{p}\sum_{j=1}^{p} \psi\left(\frac{\tilde{Y}_{1j}/\|X\|_2 - \hat{\beta}_j^T\hat{\gamma}^{\mathrm{RR}}}{\hat{\sigma}_j}\right)\hat{\beta}_j/\hat{\sigma}_j = 0. \tag{9.18}$$

The next assumption is used to control the higher order term in a Taylor expansion of Ψ.

Assumption 9.6. *The first two derivatives of ψ exist and both $|\psi'(x)| \leq D$ and $|\psi''(x)| \leq D$ hold at all x for some $D < \infty$.*

Examples of loss functions ρ that satisfy Assumptions 9.5 and 9.6 include smoothed Huber loss and Tukey's bisquare.

The next theorem gives the asymptotic distribution of $\hat{\alpha}^{\mathrm{RR}}$ when the nonzero entries of α are sparse enough. The asymptotic variance of $\hat{\alpha}^{\mathrm{RR}}$ is, again, the oracle variance in (8.3).

Theorem 9.3. *Under Assumptions 9.1 to 9.3, 9.5 and 9.6, if $n, p \to \infty$, with $p/n^k \to 0$ for some $k > 0$ and $\|\alpha\|_1\sqrt{n}/p \to 0$, then*

$$\sqrt{n}(\hat{\alpha}_{\mathcal{S}}^{\mathrm{RR}} - \alpha_{\mathcal{S}}) \xrightarrow{d} \mathrm{N}(0, (1 + \|\gamma\|_2^2)\Sigma_{\mathcal{S}})$$

for any fixed index set S with finite cardinality.

If $n/p \to 0$, then a sufficient condition for $\|\alpha\|_1\sqrt{n}/p \to 0$ in Theorem 9.3 is $\|\alpha\|_1 = O(\sqrt{p})$. If instead $n/p \to c \in (0, \infty)$, then $\|\alpha\|_1 = o(\sqrt{p})$ suffices.

9.4 Hypotheses testing

In this section, we construct significance tests for α and γ based on the asymptotic normal distributions in the previous section.

9.4.1 Test of the primary effects

We consider the asymptotic test for $H_{0j} : \alpha_j = 0,\ j = 1, \ldots, p$ resulting from the asymptotic distributions of $\hat{\alpha}_j$ derived in Theorems 9.1 and 9.3.

$$t_j = \frac{\|X\|_2 \hat{\alpha}_j}{\hat{\sigma}_j \sqrt{1 + \|\hat{\gamma}\|^2}}, \quad j = 1, \ldots, p \tag{9.19}$$

Here we require $|\mathcal{C}| \to \infty$ for the NC estimator. The null hypothesis H_{0j} is rejected at level-q if $|t_j| > z_{q/2} = \Phi^{-1}(1 - q/2)$ as usual, where Φ is the cumulative distribution function of the standard normal.

The next theorem shows that the overall type-I error and the family-wise error rate (FWER) can be asymptotically controlled by using the test statistics $t_j, j = 1, \ldots, p$.

Theorem 9.4. *Let $\mathcal{N}_p = \{j | \alpha_j = 0, j = 1, \ldots, p\}$ be all the true null hypotheses. Under the assumptions of Theorem 9.1 or Theorem 9.3, $|\mathcal{C}| \to \infty$ for the NC scenario, as $n, p, |\mathcal{N}_p| \to \infty$*

$$\frac{1}{|\mathcal{N}_p|} \sum_{j \in \mathcal{N}_p} I(|t_j| > z_{q/2}) \xrightarrow{p} q, \text{ and} \tag{9.20}$$

$$\limsup \mathrm{P}\Big(\sum_{j \in \mathcal{N}_p} I(|t_j| > z_{q/(2p)}) \geq 1 \Big) \leq q. \tag{9.21}$$

Although the individual test is asymptotically valid as $t_j \xrightarrow{d} \mathrm{N}(0, 1)$, Theorem 9.4 is not a trivial corollary of the asymptotic normal distribution in Theorems 9.1 and 9.3. This is because $t_j, j = 1, \ldots, p$ are not independent for finite samples. The proof of Theorem 9.4 investigates how the dependence of the test statistics diminishes when $n, p \to \infty$. The proof of Theorem 9.4 already requires a careful investigation of the convergence of $\hat{\beta}$ in Theorem 9.3. It is more cumbersome to prove FDR control using our test statistics. In the simulations in Section 10.1 we show that FDR is usually well controlled for the Benjamini-Hochberg procedure when the sample size is large enough.

Remark 6. We find a calibration technique in Sun et al. (2012) very useful to improve the type I error and FDR control for finite sample size. Because the asymptotic variance used in equation (9.19) is the variance of an oracle OLS estimator, when the sample size is not sufficiently large, the variance of $\hat{\beta}^{\mathrm{RR}}$ should be slightly larger than this oracle variance. To correct for this inflation, one can use median absolute deviation (MAD) with customary scaling to match the standard deviation for a Gaussian distribution to estimate the empirical standard error of $t_j, j = 1, \ldots, p$ and divide t_j by the estimated standard error. The performance of this empirical calibration is studied in the simulations in Section 10.1.

9.4.2 Test of confounding

We also consider a significance test for $H_{0,\gamma} : \gamma = 0$, under which the latent factors are not confounding.

Theorem 9.5. *Let the assumptions of Theorem 9.1 or Theorem 9.3 and $|\mathcal{C}| \to \infty$ for the NC scenario be given. Under the null hypothesis that $\gamma = 0$, for $\hat{\gamma} = \hat{\gamma}^{\mathrm{NC}}$ in (9.11) or $\hat{\gamma} = \hat{\gamma}^{\mathrm{RR}}$ in (9.15), we have*

$$n \cdot \hat{\gamma}^T \hat{\gamma} \xrightarrow{d} \chi^2_r$$

where χ^2_r is the chi-square distribution with r degree of freedom.

Therefore, the null hypothesis $H_{0,\gamma} : \gamma = 0$ is rejected if $n \cdot \hat{\gamma}^T \hat{\gamma} > \chi^2_{r,q}$ where $\chi^2_{r,q}$ is the upper-q quantile of χ^2_r. This test, combined with exploratory factor analysis, can be used as a diagnosis tool for practitioners to check whether the data gathering process has any confounding factors that can bias the multiple hypothesis testing.

9.5 Extension to multiple regression

Next, the confounder adjustment procedure is extended the multiple regression problem originally proposed in Chapter 8. In this general setting, we observe in total $d = d_0 + d_1$ predictors that can be separated into two groups:

1. X_0: $n \times d_0$ nuisance covariates that we would like to include in the regression model, and

2. X_1: $n \times d_1$ primary variables whose effects we want to study.

Leek and Storey (2008) consider the case $d_0 = 0$ and $d_1 \geq 1$ for SVA and Sun et al. (2012) consider the case $d_0 \geq 0$ and $d_1 = 1$ for LEAPP. Here we study the confounder adjusted multiple regression in full generality, for any $d_0 \geq 0$ and $d_1 \geq 1$.

In addition to the modeling assumptions in (8.1), it is necessary to consider the joint distribution of X. We assume

$$\begin{pmatrix} X_{0i} \\ X_{1i} \end{pmatrix} \text{ are i.i.d. with } \mathrm{E}\left[\begin{pmatrix} X_{0i} \\ X_{1i} \end{pmatrix}\begin{pmatrix} X_{0i} \\ X_{1i} \end{pmatrix}^T\right] = \Sigma_X, \text{ and } \Sigma_X \text{ is invertible.} \tag{9.22}$$

The model does not specify means for X_{0i} and X_{1i}; we do not need them. The parameters in this model are, for $i = 0$ or 1, $\alpha_i \in \mathbb{R}^{p \times d_i}$, $\beta \in \mathbb{R}^{p \times r}$, $\Sigma_X \in \mathbb{R}^{d \times d}$, and $\gamma_i \in \mathbb{R}^{r \times d_i}$. To clarify our purpose, we are primarily interested in estimating and testing for the significance of α_1.

For the multiple regression model, we again consider the rotation matrix Q^T that is given by the QR decomposition $\begin{pmatrix} X_0 & X_1 \end{pmatrix} = QU$ where $Q \in \mathbb{R}^{n \times n}$ is an orthogonal matrix and U is an upper triangular matrix of size $n \times d$. Therefore we have

$$Q^T \begin{pmatrix} X_0 & X_1 \end{pmatrix} = U = \begin{pmatrix} U_{00} & U_{01} \\ 0 & U_{11} \\ 0 & 0 \end{pmatrix}$$

where U_{00} is a $d_0 \times d_0$ upper triangular matrix and U_{11} is a $d_1 \times d_1$ upper triangular matrix. Now let the rotated Y be

$$\tilde{Y} = Q^T Y = \begin{pmatrix} \tilde{Y}_0 \\ \tilde{Y}_1 \\ \tilde{Y}_{-1} \end{pmatrix} \tag{9.23}$$

where $\tilde{Y}_0$ is $d_0 \times p$, $\tilde{Y}_1$ is $d_1 \times p$ and $\tilde{Y}_{-1}$ is $(n-d) \times p$, then we can partition the model into three parts: conditional on both X_0 and X_1 (hence U),

$$\tilde{Y}_0 = U_{00}\alpha_0^T + U_{01}\alpha_1^T + \tilde{Z}_0\beta^T + \tilde{E}_0, \tag{9.24}$$

$$\tilde{Y}_1 = U_{11}\alpha_1^T + \tilde{Z}_1\beta^T + \tilde{E}_1 \sim \mathrm{MN}(U_{11}(\alpha_1 + \beta\gamma_1)^T, I_{d_1}, \beta\beta^T + \Sigma) \tag{9.25}$$

$$\tilde{Y}_{-1} = \tilde{Z}_{-1}\beta^T + \tilde{E}_{-1} \sim \mathrm{MN}(0, I_{n-d}, \beta\beta^T + \Sigma) \tag{9.26}$$

where $\tilde{Z} = Q^T Z$ and $\tilde{E} = Q^T E \overset{d}{=} E$. Equation equation (9.24) corresponds to the nuisance parameters α_0 and is discarded according to the ancillary principle. Equation equation (9.25) is the multivariate extension to equation (9.2) that is used to estimate α_1 and equation equation (9.26) plays the same role as equation (9.3) to estimate β and Σ.

We consider the asymptotics when $n, p \to \infty$ and d, r are fixed and known. Since d is fixed, the estimation of β is not different from the simple regression case and we can use the maximum likelihood factor analysis described in Section 9.3.1. Under Assumptions 9.1 to 9.3, the precision results of $\hat{\beta}$ and $\hat{\Sigma}$ (Lemma 9.3) still hold.

Let $\Sigma_X^{-1} = \Omega = \left(\begin{smallmatrix}\Omega_{00} & \Omega_{01} \\ \Omega_{10} & \Omega_{11}\end{smallmatrix}\right)$. In the proof of Theorems 9.1 and 9.3, we consider a fixed sequence of X such that $\|X\|_2/\sqrt{n} \to 1$. Similarly, we have the following lemma in the multiple regression scenario:

Lemma 9.4. *As $n \to \infty$, $\frac{1}{n}U_{11}^T U_{11} \overset{a.s.}{\to} \Omega_{11}^{-1}$.*

Similar to (9.8), we can rewrite (9.25) as

$$\tilde{Y}_1^T U_{11}^{-T} = \alpha_1 + \beta(\gamma_1 + \tilde{W}_1 U_{11}^{-T}) + \tilde{E}_1 U_{11}^{-T}$$

where $\tilde{W}_1 \sim \mathrm{MN}(0, I_{d_1}, I_p)$ is independent from $\tilde{E}_1$. As in Section 9.3.2, we derive statistical properties of the estimate of α_1 for a fixed sequence of X, $\tilde{W}_1$ and Z, which also hold unconditionally. For simplicity, we assume that the negative controls are a known set of variables $\mathcal{C}$ with $\alpha_{1,\mathcal{C}} = 0$. We can then estimate each column of γ_1 by applying the negative control (NC) or robust regression (RR) we discussed in

Sections 9.3.2 and 9.3.2 to the corresponding row of $\tilde{Y}_1 U_{11}^{-T}$, and then estimate α_1 by

$$\hat{\alpha}_1 = \tilde{Y}_1^T U_{11}^{-T} - \hat{\beta}\hat{\gamma}_1.$$

Notice that $\tilde{E}_1 U_{11}^{-T} \sim \mathrm{MN}\big(0, \Sigma, U_{11}^{-1} U_{11}^{-T}\big)$. Thus the "samples" in the robust regression, which are actually the p variables in the original problem are still independent within each column. Though the estimates of each column of γ_1 may be correlated, we will show that the correlation won't affect inference on α_1. As a result, we still get asymptotic results similar to Theorem 9.3 for the multiple regression model (8.1):

Theorem 9.6. *Under Assumptions 9.1 to 9.6, if $n, p \to \infty$, with $p/n^k \to 0$ for some $k > 0$, and $\|\mathrm{vec}(\alpha_1)\|_1 \sqrt{n}/p \to 0$, then for any fixed index set $\mathcal{S}$ with finite cardinality $|\mathcal{S}|$,*

$$\sqrt{n}(\hat{\alpha}_{1,\mathcal{S}}^{\mathrm{NC}} - \alpha_{1,S}) \xrightarrow{d} \mathrm{MN}(0_{|\mathcal{S}|\times k_1}, \Sigma_{\mathcal{S}} + \Delta_{\mathcal{S}}, \Omega_{11} + \gamma_1^T\gamma_1), \quad \textit{and} \tag{9.27}$$

$$\sqrt{n}(\hat{\alpha}_{1,\mathcal{S}}^{\mathrm{RR}} - \alpha_{1,S}) \xrightarrow{d} \mathrm{MN}(0_{|\mathcal{S}|\times k_1}, \Sigma_{\mathcal{S}}, \Omega_{11} + \gamma_1^T\gamma_1) \tag{9.28}$$

where $\Delta_{\mathcal{S}}$ is defined in Theorem 9.1.

As for the asymptotic efficiency of this estimator, we again compare it to the oracle OLS estimator of α_1 which observes confounding variables Z in (8.1). In the multiple regression model, we claim that $\hat{\alpha}_1^{\mathrm{RR}}$ still reaches the oracle asymptotic efficiency. In fact, let $\alpha = (\alpha_0, \alpha_1, \beta)$. The oracle OLS estimator of α, $\hat{\alpha}^{\mathrm{OLS}}$, is unbiased and its vectorization has variance $V^{-1} \otimes \Sigma/n$ where

$$V = \begin{pmatrix} \Sigma_X & \Sigma_X \gamma^T \\ \gamma\Sigma_X & I_r + \gamma\Sigma_X\gamma^T \end{pmatrix}, \text{ for } \gamma = (\gamma_0, \gamma_1).$$

By the block-wise matrix inversion formula, the top left $d \times d$ block of V^{-1} is $\Sigma_X^{-1} + \gamma^T\gamma$. The variance of $\hat{\alpha}_1^{\mathrm{OLS}}$ only depends on the bottom right $d_1 \times d_1$ sub-block of this $d \times d$ block, which is simply $\Omega_{11} + \gamma_1^T\gamma_1$. Therefore $\hat{\alpha}_1^{\mathrm{OLS}}$ is unbiased and its vectorization has variance $(\Omega_{11} + \gamma_1^T\gamma_1) \otimes \Sigma/n$, matching the asymptotic variance of $\hat{\alpha}_1^{\mathrm{RR}}$ in Theorem 9.6.

9.6 Discussions

9.6.1 Confounding vs. unconfounding

The issue of multiple testing dependence arises because Z in the true model (8.1) is unobserved. We have focused on the case where Z is confounded with the primary variable. Some similar results were obtained earlier for the unconfounded case, corresponding to $\alpha = 0$ in our notation. For example, Lan and Du (2014) used a factor model to improve the efficiency of significance test of the regression intercepts. Jin (2012), Li and Zhong (2014) developed more powerful procedures for testing β while still controlling FDR under unconfounded dependence.

In another related work, Fan et al. (2012) imposed a factor structure on the unconfounded test statistics, whereas this Chapter and the articles discussed later in Section 9.6.2 assume a factor structure on the raw data. Fan et al. (2012) used an approximate factor model to accurately estimate the false discovery proportion. Their correction procedure also includes a step of robust regression. Nevertheless, it is often difficult to interpret the factor structure of the test statistics. In comparison, the latent variables Z in our model (8.1), whether confounding or not, can be interpreted as batch effects, laboratory conditions, or other systematic bias. Such problems are widely observed in genetics studies (see e.g. the review article by Leek et al., 2010).

As a final remark, some of the models and methods developed in the context of unconfounded hypothesis testing may be useful for confounded problems as well. For example, the relationship between Z and X needs not be linear as in (8.1c). In certain applications, it may be more appropriate to use a time-series model (e.g. Sun and Cai, 2009) or a mixture model (e.g. Efron, 2010).

9.6.2 Comparison with existing confounder adjustment methods

In this section we discuss in more detail how previous methods of confounder adjustment, namely SVA, RUV-4 and LEAPP, fit in the framework equation (8.1).

SVA

There are two versions of SVA: the reduced subset SVA (subset-SVA) of Leek and Storey (2007) and the iteratively reweighted SVA (IRW-SVA) of Leek and Storey (2008). Both of them can be interpreted as the two-step statistical procedure in the framework equation (8.1). In the first step, SVA estimates the confounding factors by applying PCA to the residual matrix $(I - H_X)Y$ where $H_X = X(X^TX)^{-1}X^T$ is the projection matrix of X. In contrast, we applied factor analysis to the rotated residual matrix $(Q^TY)_{-1}$, where Q comes from the QR decomposition of X in Section 9.5.

To see why these two approaches lead to the same estimate of β, we introduce the block form of $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$ where $Q_1 \in \mathbb{R}^{n\times d}$ and $Q_2 \in \mathbb{R}^{n\times(n-d)}$. It is easy to show that $(Q^TY)_{-1} = Q_2^TY$ and $(I - H_X)Y = Q_2Q_2^TY$. Thus our rotated matrix $(Q^TY)_{-1}$ decorrelates the residual matrix by left-multiplying by Q_2 (because $Q_2^TQ_2 = I_{n-d}$). Because $(Q_2^TY)^TQ_2^TY = (Q_2Q_2^TY)^TQ_2Q_2^TY$, $(Q^TY)_{-1}$ and $(I - H_X)Y$ have the same sample covariance matrix, they will yield the same factor loading estimate under PCA and also under MLE. The main advantage of using the rotated matrix is theoretical: the rotated residual matrices have independent rows.

Recall that the second step is to estimate α based on the confounding factors estimated in the first step. Because SVA doesn't assume an explicit relationship between the primary variable X and the confounders Z, it cannot use the regression equation (9.8) to estimate γ (not even defined) and α. Instead, the two SVA algorithms try to reconstruct the surrogate variables, which are essentially the confounders Z in our framework. Assuming the true primary effect α is sparse, the subset-SVA algorithm finds the outcome variables Y that have the smallest marginal correlation with X and uses their principal scores as Z. Then, it computes the p-values by F-tests comparing the linear regression models with and without Z. This procedure can easily fail because a small marginal correlation does not imply no real effect of X due to the confounding factors. For example, most of the marginal effects in the gender study in Figure 8.1b are very small, but after confounding adjustment we find some are indeed significant (see Section 10.2).

The IRW-SVA algorithm modifies subset-SVA by using an iterative procedure in the second step. The subset is chosen iteratively. At each step, IRW-SVA gives a weight to each outcome variable based on how likely it is that $\alpha_j = 0$, given the current estimate of surrogate variables. These weights are then used in a weighted PCA algorithm to update the estimated surrogate variables. IRW-SVA may be related to our robust regression estimator in Section 9.3.2 in the sense that an M-estimator is commonly solved by Iteratively Reweighted Least Squares (IRLS) and the weights also represents how likely the data point is an outlier. However, unlike IRLS, the iteratively reweighted PCA algorithm used in IRW-SVA has no theoretical guarantee of performance. It does not even have a guarantee of convergence. Some previous literature (Sun et al., 2012, Gagnon-Bartsch et al., 2013) and our experiments in Section 10.1 show that SVA is outperformed by the competitors in most cases and performs slightly better when confounding is a minor issue.

RUV

Gagnon-Bartsch et al. (2013) derived RUV-4 estimator of α via a sequence of heuristic calculations. In Section 9.3.2, we derived an analytically more tractable estimator $\hat{\alpha}^{\mathrm{NC}}$ which is actually the same as RUV-4, with the only difference being that we use MLE instead of PCA to estimate the factors and GLS instead of OLS in equation (9.11). To see why $\hat{\alpha}^{\mathrm{NC}}$ is essentially the same as $\hat{\alpha}^{\mathrm{RUV-4}}$, in the first step RUV-4 used residual matrix to estimate α and Z, which yields the same estimate as using rotated matrix (Section 9.6.2). In the second step, RUV-4 estimated α via a regression on X and $\hat{Z} = Q\begin{pmatrix}\tilde{Z}_{-1}^T & \hat{\gamma}^T\end{pmatrix}^T$. The regression would estimate α the same as $\hat{\alpha}^{\mathrm{PCA}}$ in the first step, thus estimate α the same as using (9.12). Based on more heuristic calculations, the authors claim the RUV-4 estimator has approximately the oracle variance in Section 8.1. We rigorously prove this statement in Theorem 9.1 when the number of negative controls is large and give a finite sample correction when the negative controls are few. In Section 10.1 we show this correction is very useful to control the type I error and FDR in simulations.

LEAPP

We follow the two-step procedure and robust regression framework in LEAPP in this Chapter, thus the test statistics t_j^{RR} are very similar to the test statistics t_j^{LEAPP} in LEAPP. The difference is that LEAPP uses the Θ-IPOD algorithm of She and Owen (2011) for outlier detection, which is robust against outliers at leverage points but is not easy to analyze. Indeed Sun et al. (2012) replace it by the Dantzig selector in their theoretical section. Here we use a classical M-estimator, which allows us to study the theoretical properties more easily. In practice, LEAPP and RR estimator usually produce very similar results; see Section 10.1 for a numerical comparison.

9.6.3 Inference when Σ is nondiagonal

Our analysis is based on the assumption that the noise covariance matrix Σ is diagonal, though in many applications, the researcher might suspect that the outcome variables Y in model equation (8.1) are still correlated after conditioning on the latent factors. Typical examples include gene regulatory networks (De La Fuente et al., 2004) and cross-sectional panel data (Pesaran, 2004), where the variable dependence sometimes cannot be fully explained by the latent factors or may simply require too many of them. Bai and Li (2015) extend the theoretical results in Bai and Li (2012) to approximate factor models allowing for weakly correlated noise. Approximate factor models have also been discussed in Fan and Han (2013).

Chapter 10

Numerical Examples

10.1 Simulations

We have provided theoretical guarantees of confounder adjusting methods in various settings and the asymptotic regime of $n, p \to \infty$ (e.g. Theorems 9.1 to 9.4 and 9.6). Now we use numerical simulations to verify these results and further study the finite sample properties of our estimators and tests statistics.

The simulation data are generated from the single primary variable model (8.1). More specifically, X_i is a centered binary variable $(X_i + 1)/2 \overset{\text{i.i.d.}}{\sim}$ Bernoulli(0.5), and Y_i, Z_i are generated according to equation (8.1).

For the parameters in the model, the noise variances are generated by $\sigma_j^2 \overset{\text{i.i.d.}}{\sim}$ InvGamma$(3, 2)$, $j = 1, \ldots, p$, and so $\mathbb{E}(\sigma_j^2) = \mathrm{Var}(\sigma_j^2) = 1$. We set each $\gamma_k = \|\gamma\|_2/\sqrt{r}$ equally for $k = 1, 2, \cdots, r$ where $\|\gamma\|_2^2$ is set to 0, 1, or 1/19, so the variance of X_i explained by the confounding factors is $R^2 = 0\%$, 5%, or 50%. The primary effect α has independent components α_i taking the values $3\sqrt{1 + \|\alpha\|_2^2}$ and 0 with probability $\pi = 0.05$ and $1 - \pi = 0.95$, respectively, so the nonzero effects are sparse and have effect size 3. This implies that the oracle estimator has power approximately $\mathrm{P}(\mathrm{N}(3, 1) > z_{0.025}) = 0.85$ to detect the signals at a significance level of 0.05. We set the number of latent factors r to be either 2 or 10. For the latent factor loading matrix β, we take $\beta = \tilde{\beta} D$ where $\tilde{\beta}$ is a $p \times r$ orthogonal matrix sampled uniformly from the Stiefel manifold $V_r(\mathbb{R}^p)$, the set of all $p \times r$ orthogonal matrix. Based on

Assumption 9.3, we set the latent factor strength $D = \sqrt{p} \cdot \mathrm{diag}(d_1, \cdots, d_r)$ where $d_k = 3 - 2(k-1)/(r-1)$ thus d_1 to d_r are distributed evenly inside the interval $[3, 1]$. As the number of factors r can be easily estimated for this strong factor setting (more discussions can be found in Owen and Wang (2015)), we assume that the number r of factors is known to all of the algorithms in this simulation.

We set $p = 5000$, $n = 100$ or 500 to mimic the data size of many genetic studies. For the negative control scenario, we choose $|\mathcal{C}| = 30$ negative controls at random from the zero positions of α. We expect that negative control methods would perform better with a larger value of $|\mathcal{C}|$ and worse with a smaller value. The choice $|\mathcal{C}| = 30$ is around the size of the spike-in controls in many microarray experiments (Gagnon-Bartsch and Speed, 2012). For the loss function in our sparsity scenario, we use Tukey's bisquare which is optimized via IRLS with an ordinary least-square fit as the starting values of the coefficients. Finally, each of the four combinations of n and r is randomly repeated 100 times.

We compare the performance of nine different approaches. There are two baseline methods: the "naive" method estimates α by a linear regression of Y on just the observed primary variable X and calculates p-values using the classical t-tests, while the "oracle" method regresses Y on both X and the confounding variables Z as described in Section 8.1. There are three methods in the RUV-4/negative controls family: the RUV-4 method (Gagnon-Bartsch et al., 2013), our "NC" method which computes test statistics using $\hat{\alpha}^{\mathrm{NC}}$ and its variance estimate $(1+\|\hat{\gamma}\|_2^2)(\hat{\Sigma}+\hat{\Delta})$, and our "NC-ASY" method which uses the same $\hat{\alpha}^{\mathrm{NC}}$ but estimates its variance by $(1+\|\hat{\gamma}\|_2^2)\hat{\Sigma}$. We compare four methods in the SVA/LEAPP/sparsity family: these are "IRW-SVA" (Leek and Storey, 2008), "LEAPP" (Sun et al., 2012), the "LEAPP(RR)" method which is our RR estimator using M-estimation at the robustness stage and computes the test-statistics using (9.19), and the "LEAPP(RR-MAD)" method which uses the median absolute deviation (MAD) of the test statistics in (9.19) to calibrate them. (see Section 9.4)

To measure the performance of these methods, we report the type I error (Theorem 9.4), power, false discovery proportion (FDP) and precision of hypotheses with the smallest 100 p-values in the 100 simulations. For both the type I error and power, we set the significance level to be 0.05. For FDP, we use Benjamini-Hochberg procedure with FDR controlled at 0.2. These metrics are plotted in Figures 10.1 to 10.3 under different settings of n and r.

First, from these figures, we see that the oracle method has exactly the same type I error and FDP as specified, while the naive method and SVA fail drastically when the latent variables are confounding. SVA performs performs better than the naive method in terms of the precision of the smallest 100 p-values, but is still much worse than other methods. Next, for the negative control scenario, as we only have $|\mathcal{C}| = 30$ negative controls, ignoring the inflated variance term Δ_S in Theorem 9.1 will lead to overdispersed test statistics, and that's why the type I error and FDP of both NC-ASY and RUV-4 are much larger than the nominal level. By contrast, the NC method correctly controls type I error and FDP by considering the variance inflation, though as expected it loses some power compared with the oracle. For the sparsity scenario, the "LEAPP(RR)" method performs as the asymptotic theory predicted when $n = 500$, while when $n = 100$ the p-values seem a bit too small. This is not surprising because the asymptotic oracle variance in Theorem 9.3 can be optimistic when the sample size is not sufficiently large, as we discussed in Remark 6. On the other hand, the methods which use empirical calibration for the variance of test statistics, namely the original LEAPP and "LEAPP(RR-MAD)", control both FDP and type I error for data of small sample size in our simulations. The price for the finite sample calibration is that it tends to be slightly conservative, resulting in a loss of power to some extent.

In conclusion, the simulation results are consistent with our theoretical guarantees when p is as large as 5000 and n is as large as 500. When n is small, the variance of the test statistics will be larger than the asymptotic variance for the sparsity scenario and we can use empirical calibrations (such as MAD) to adjust for the difference.

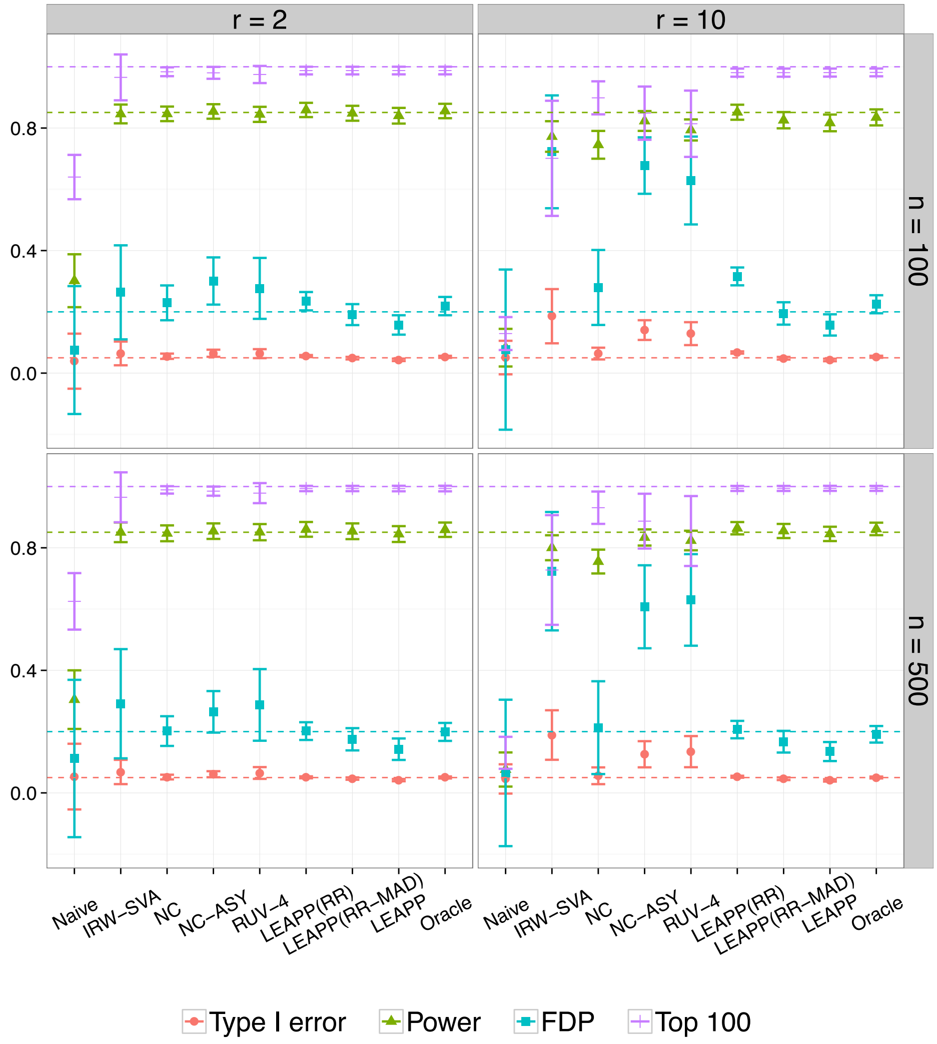


Figure 10.1: Compare the performance of nine different approaches in unconfounded scenario. From left to right: naive regression ignoring the confounders (Naive), IRW-SVA, negative control with finite sample correction (NC) in equation (9.13), negative control with asymptotic oracle variance (NC-ASY) in equation (9.14), RUV-4, robust regression (LEAPP(RR)), robust regression with calibration (LEAPP(RR-MAD)), LEAPP, oracle regression which observes the confounders (Oracle). The error bars are one standard deviation over 100 repeated simulations. The three dashed horizontal lines from bottom to top are the nominal significance level, FDR level and oracle power, respectively.

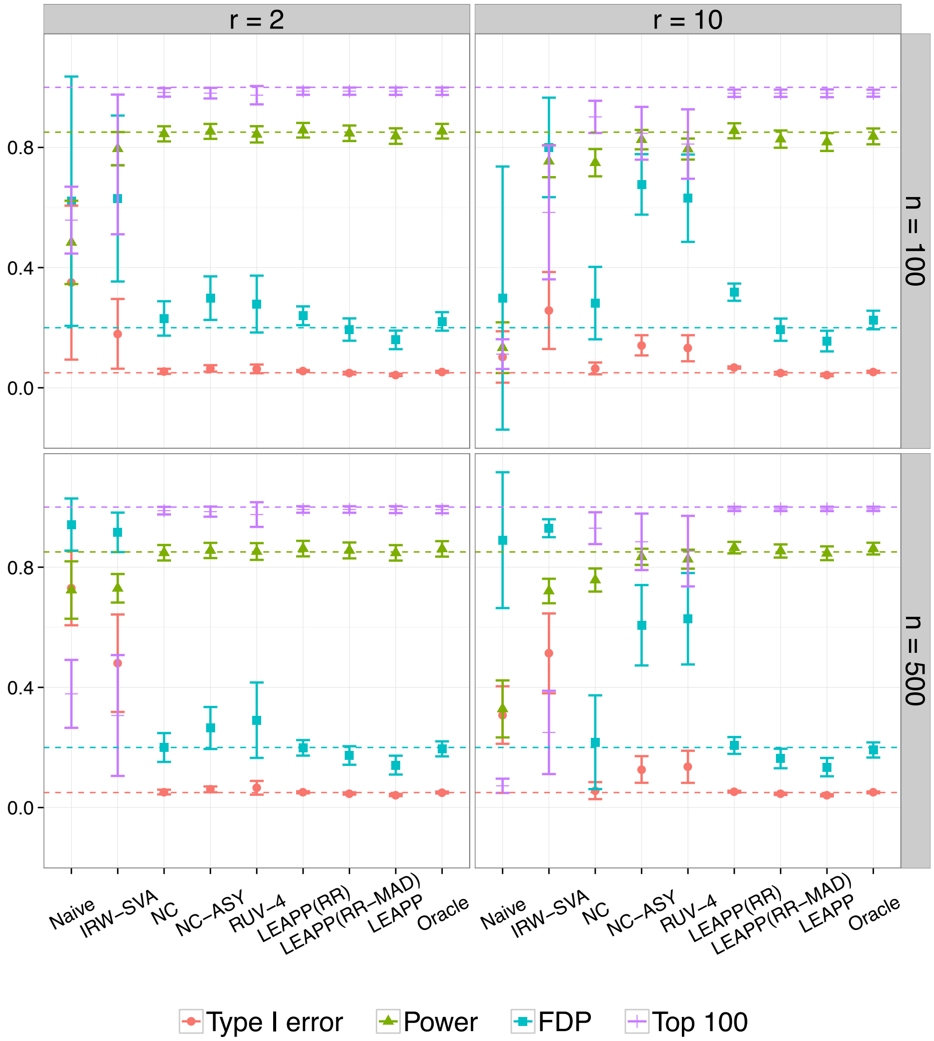


Figure 10.2: Compare the performance of nine different approaches in the confounded scenario ($\|\gamma\|_2^2 = 1/19$, i.e. the confounders explain 5% of the variation of X). From left to right: naive regression ignoring the confounders (Naive), IRW-SVA, negative control with finite sample correction (NC) in equation (9.13), negative control with asymptotic oracle variance (NC-ASY) in equation (9.14), RUV-4, robust regression (LEAPP(RR)), robust regression with calibration (LEAPP(RR-MAD)), LEAPP, oracle regression which observes the confounders (Oracle). The error bars are one standard deviation over 100 repeated simulations. The three dashed horizontal lines from bottom to top are the nominal significance level, FDR level and oracle power, respectively.

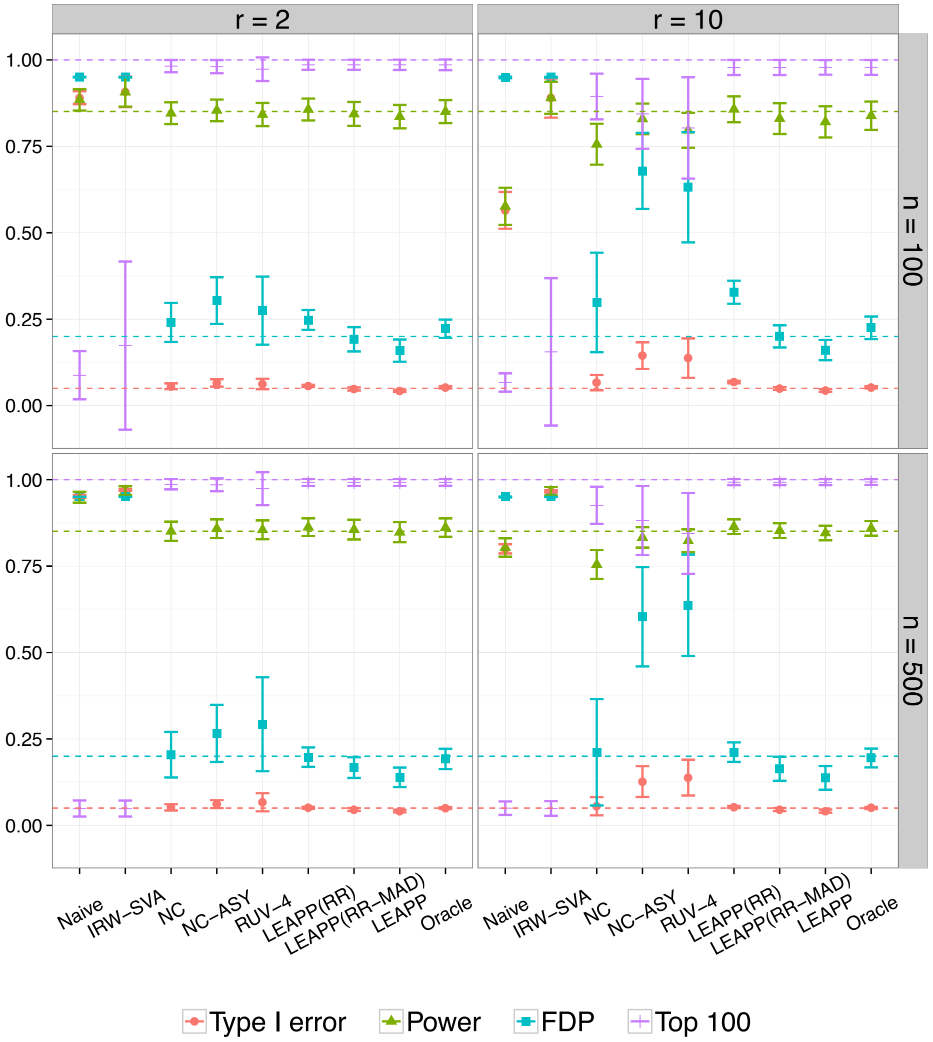


Figure 10.3: Compare the performance of nine different approaches in the confounded scenario ($\|\gamma\|_2^2 = 1$, i.e. the confounders explain 50% of the variation of X). From left to right: naive regression ignoring the confounders (Naive), IRW-SVA, negative control with finite sample correction (NC) in equation (9.13), negative control with asymptotic oracle variance (NC-ASY) in equation (9.14), RUV-4, robust regression (LEAPP(RR)), robust regression with calibration (LEAPP(RR-MAD)), LEAPP, oracle regression which observes the confounders (Oracle). The error bars are one standard deviation over 100 repeated simulations. The three dashed horizontal lines from bottom to top are the nominal significance level, FDR level and oracle power, respectively.

10.2 Real data example: Batch effects in microarray experiments

In this section, we return to the three motivating real data examples in Section 8.2. The main goal here is to demonstrate a practical procedure for confounder adjustment and show that our asymptotic results are reasonably accurate in real data. In an open-source R package `cate` (available on CRAN), we also provide the necessary tools to carry out the procedure.

Recall that without the confounder adjustment, the distribution of the regression t-statistics in these datasets can be skewed, noncentered, underdispersed, or overdispersed as shown in Figure 8.1. The adjustment method used here is the maximum likelihood factor analysis described in Section 9.3.1 followed by the robust regression (RR) method with Tukey's bisquare loss described in Section 9.3.2. Since the true number of confounders is unknown, we increase r from 1 to $n/2$ and study the empirical performance. We report the results without empirical calibration for illustrative purposes, though in practice we suggest using calibration for better control of type I errors and FDP.

In Table 10.1 and Figure 10.4, we present the results after confounder adjustment for the three datasets. We report two groups of summary statistics in Table 10.1: the first group is several summary statistics of all the z-statistics computed using equation (9.19), including the mean, median, standard deviation, median absolute deviation (scaled for consistency of normal distribution), skewness, and the medcouple. The medcouple (Brys et al., 2004)) is a robust measure of skewness. After subtracting the median observation some positive and some negative values remain. For any pair of values $x_1 \geq 0$ and $x_2 \leq 0$ with $x_1 + |x_2| > 0$ one can compute $(x_1 - |x_2|)/(x_1 + |x_2|)$. The medcouple is the median of all those ratios. The second group of statistics has performance metrics to evaluate the effectiveness of the confounder adjustment. See the caption of Table 10.1 for more detail.

In all three datasets, the z-statistics become more centered at 0 and less skewed as we include a few confounders in the model. Though the standard deviation (SD)

r	mean	median	sd	mad	skewness	medc.	#sig.	p-value
0	-0.16	0.024	2.65	2.57	-0.104	-0.091	164	NA
1	-0.45	-0.39	2.85	2.52	-0.25	0.00074	1162	0.0057
2	0.012	-0.039	1.35	1.33	0.139	0.042	542	<1e-10
3	0.014	-0.05	1.43	1.41	0.169	0.048	552	<1e-10
5	-0.029	-0.11	1.52	1.48	0.236	0.057	647	<1e-10
7	-0.1	-0.14	1.42	1.35	0.109	0.027	837	<1e-10
10	-0.06	-0.085	1.13	1.12	0.103	0.022	506	<1e-10
20	-0.083	-0.095	1.2	1.19	0.0604	0.0095	479	<1e-10
33	**-0.099**	**-0.11**	**1.33**	**1.3**	**0.0727**	**0.0056**	**579**	**<1e-10**
40	-0.1	-0.12	1.43	1.4	0.0775	0.0072	585	<1e-10
50	-0.16	-0.17	1.58	1.53	0.0528	0.0032	678	<1e-10

(a) Dataset 1 ($n = 143$, $p = 54675$). Primary variable: severity of COPD.

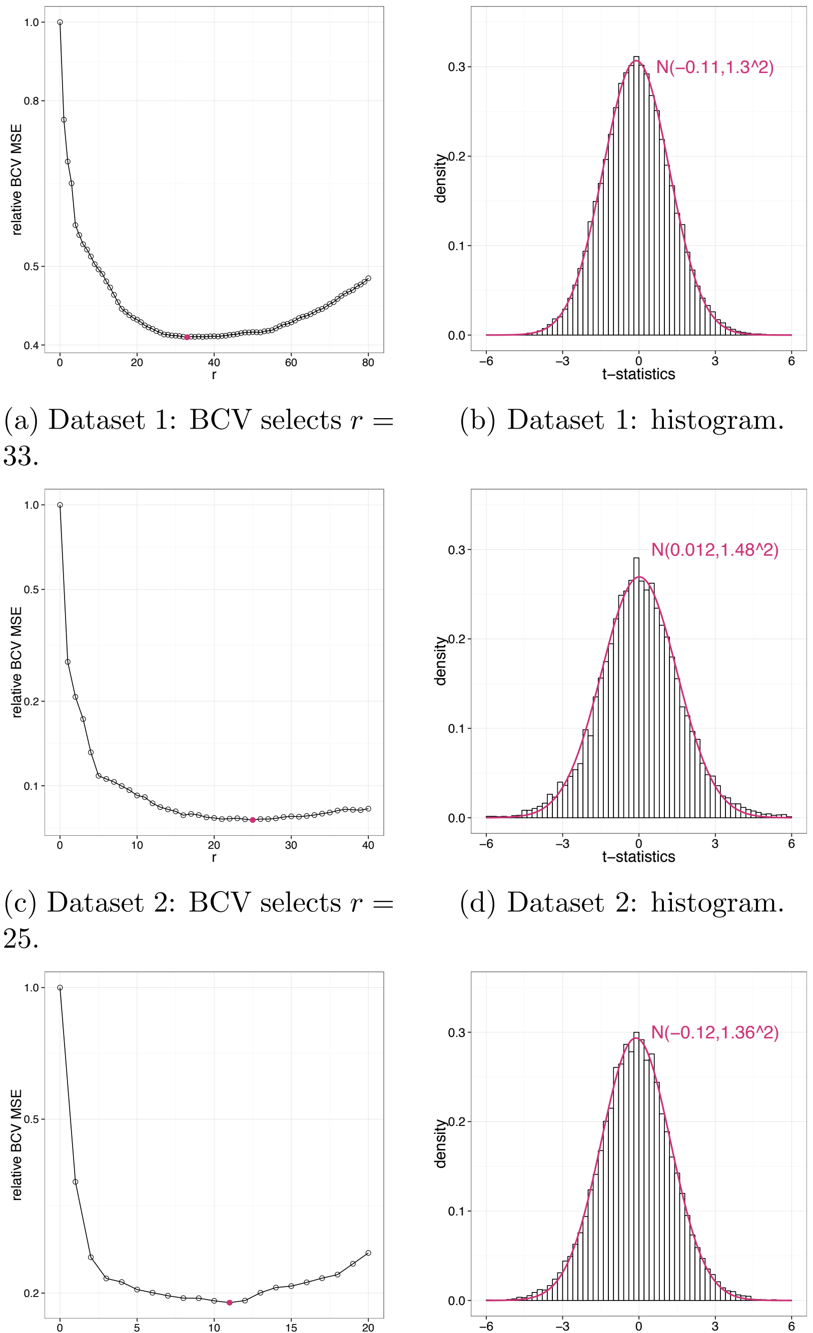
r	mean	median	sd	mad	skewness	medc.	#sig.	X/Y	top 100	p-value
0	0.11	0.043	0.36	0.237	2.99	0.2	1036	58	11	NA
1	-0.44	-0.47	1.06	1.04	0.688	0.035	108	20	20	0.74
2	-0.14	-0.15	1.15	1.13	0.601	0.015	113	21	21	0.31
3	0.013	0.012	1.13	1.08	0.795	-0.01	168	34	28	0.03
5	0.044	0.019	1.18	1.08	0.878	0.017	238	32	27	0.0083
7	0.03	0.012	1.26	1.15	0.784	0.0062	269	35	25	0.006
10	0.023	0.00066	1.36	1.24	0.661	0.011	270	38	27	0.019
15	0.049	0.022	1.46	1.31	0.584	0.012	296	36	29	0.00082
20	0.029	-0.0009	1.53	1.36	0.502	0.019	314	36	28	7.2e-07
25	**0.048**	**0.012**	**1.68**	**1.48**	**0.452**	**0.026**	**354**	**37**	**27**	**1.1e-06**
30	0.026	0.012	1.82	1.61	0.436	0.0068	337	40	27	8.7e-08
40	0.061	0.046	2.07	1.79	0.642	0.0028	363	41	27	7.7e-10

(b) Dataset 2 ($n = 84$, $p = 12600$). Primary variable: gender.

r	mean	median	sd	mad	skewness	medc.	#sig.	X/Y	top 100	p-value
0	-1.8	-1.8	0.599	0.513	-3.46	0.082	418	39	20	NA
1	-0.55	-0.56	1.09	1.01	-1.53	0.01	261	29	23	0.00024
2	-0.2	-0.22	1.2	1.11	-0.99	0.014	320	38	22	0.00014
3	-0.096	-0.12	1.27	1.18	-0.844	0.017	311	42	25	0.00014
5	-0.33	-0.32	1.31	1.22	-1.29	-0.011	305	35	23	2.1e-07
7	-0.37	-0.36	1.46	1.36	-0.855	-0.0099	300	38	23	4.0e-07
11	**-0.13**	**-0.12**	**1.51**	**1.36**	**-0.601**	**-0.0051**	**432**	**48**	**31**	**1.8e-09**
15	-0.12	-0.13	1.83	1.62	-0.341	0.013	492	54	25	2.3e-08
20	-0.13	-0.14	2.61	2.23	-0.327	0.0045	613	50	26	4.0e-06

(c) Dataset 3 ($n = 31$, $p = 22283$). Primary variable: gender.

Table 10.1: Summary of the adjusted z-statistics. The first group is summary statistics of the z-statistics before the empirical calibration. The second group is some performance metrics after the empirical calibration, including total number of significant genes of p-value less than 0.01 in Remark 6 (#sig.), number of the genes on X/Y chromosome that have p-value less than 0.01 (X/Y), the number among the 100 most significant genes that are on the X/Y chromosome (top 100) and the p-value of the confounding test in Section 9.4.2. The bold row corresponds to the r selected by BCV (Figure 10.4).



(a) Dataset 1: BCV selects $r =$ 33.

(b) Dataset 1: histogram.

(c) Dataset 2: BCV selects $r =$ 25.

(d) Dataset 2: histogram.

(e) Dataset 3: BCV selects $r =$ 11.

(f) Dataset 3: histogram.

Figure 10.4: Histograms of z-statistics after confounder adjustment (without calibration) using the number of confounders r selected by bi-cross-validation.

suggests overdispersed variance, the overdispersion will go away if we add MAD calibration as SD and MAD have similar values. The similarity between SD and MAD values also indicates that the majority of statistics after confounder adjustment are approximately normally distributed. Note that the medcouple values shrink towards zero after adjustment, suggesting that skewness then only arises from small fraction of the genes, which is in accordance with our assumptions that the primary effects should be sparse.

In practice, some latent factors may be too weak to meet Assumption 9.3 (i.e. $d_j \ll \sqrt{p}$) , making it difficult to choose an appropriate r. A practical way to pick the number of confounders r with presence of heteroscedastic noise we investigate here is the bi-cross-validation (BCV) method of Owen and Wang (2015), which uses randomly held-out submatrices to estimate the mean squared error of reconstructing factor loading matrix. It is shown in Owen and Wang (2015) that BCV outperforms many existing methods in recovering the latent signal matrix and the number of factors r, especially in high-dimensional datasets ($n, p \to \infty$). In Figure 10.4, we demonstrate the performance of BCV on these three datasets. The r selected by BCV is respectively 33, 25 and 11 (Figures 10.4a, 10.4c and 10.4e), and they all result in the presumed shape of z-statistics distribution (Figures 10.4b, 10.4d and 10.4f). For the second and the third datasets where we have a gold standard, the r selected by BCV has near optimal performance in selecting genes on the X/Y chromosome (columns 3 and 4 in Tables 10.1b and 10.1c). Another method we applied is proposed by Onatski (2010) based on the empirical distribution of eigenvalues. This method estimates r as 2, 9 and 3 respectively for the three datasets. Table 3 of Gagnon-Bartsch et al. (2013) has the "top 100" values for RUV-4 on the second and third dataset. They reported 26 for LEAPP, 28 for RUV-4, and 27 for SVA in the second dataset, and 27 for LEAPP, 31 for RUV-4, and 26 for SVA in the third dataset. Notice that the precision of the top 100 significant genes is relatively stable when r is above certain number. Intuitively, the factor analysis is applied to the residuals of Y on X and the overestimated factors also have very small eigenvalues, thus they usually do not change $\hat{\alpha}$ a lot. See also Gagnon-Bartsch et al. (2013) for more discussion on the robustness of the negative control estimator to overestimating r.

Lastly we want to point out that both the small sample size of the datasets and presence of weak factors can result in overdispersed variance of the test statistics. The BCV plots indicate presence of many weak factors in the first two datasets. In the third dataset, the sample size n is only 31, so the adjustment result is not ideal. Nevertheless, the empirical performance (e.g. number of X/Y genes in top 100) suggests it is still beneficial to adjust for the confounders.

10.3 Real data example: Mutual fund selection

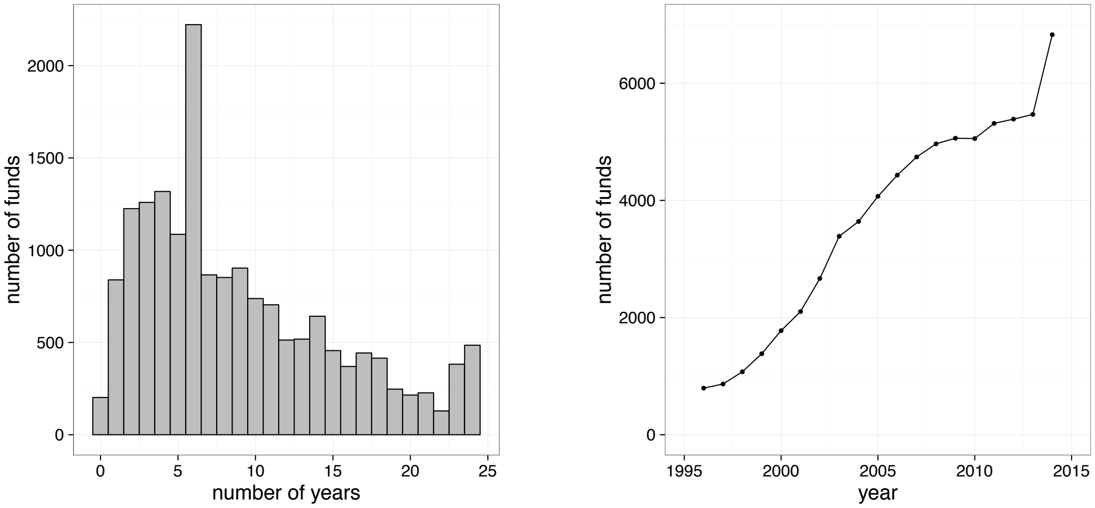
we use the Center for Research in Security Prices (CRSP) Survivorship Bias Free Mutual Fund Database first complied in Carhart (1997). Our data set contains montly return of the mutual funds from January 1991 to December 2014.[1] To focus our analysis on actively-managed open-ended U.S. domestic equity mutual funds, we eliminate balanced, bond, money market, international, as well as index funds. We also excluded funds which never managed more than 5 million dollars. In total, this leaves us with 17256 distinct mutual funds. Figure 10.5a shows the histogram of the number of observations of these mutual funds.

The empirical exercise is implemented as follows. At the beginning of every year from 1996 to 2013, we obtain a subsample of mutual funds whose monthly returns are fully available in the last 5 years. This means the eligible funds are at least 5 years old and have no missing observations in the CRSP database. The number of eligible mutual funds are plotted in Figure 10.5b. For every eligible fund, we model the monthly returns (in total 60 observations) by augmenting the standard Fama-French-Carhart four factor model (8.4) with $r = 3$ unobserved factors:

$$Y = \tilde{\alpha} + \alpha_1 X_{\text{Mkt-Rf}} + \alpha_2 X_{\text{SMB}} + \alpha_3 X_{\text{HML}} + \alpha_4 X_{\text{MOM}} + \beta^T Z + \epsilon_j. \tag{10.1}$$

Then we estimate the risk-adjusted return $\tilde{\alpha}_0$ using the robust regression method

[1] The CRSP data starts from 1962, but we concentrate on the period after 1991 bceause CRSP reports monthly total net asset since 1991.



(a) Histogram of the number of observations for each mutual fund (in year).

(b) Number of eligible mutual funds (no missing observations in the last 5 years).

Figure 10.5: Summary of the sample.

described in Chapter 9. Next, we compute the *smart alpha* of mutual fund by

$$SA \equiv \frac{\tilde{\alpha}}{\mathrm{sd}(\alpha_1 X_{\text{Mkt-Rf}} + \alpha_2 X_{\text{SMB}} + \alpha_3 X_{\text{HML}} + \alpha_4 X_{\text{MOM}} + \beta^T Z)}. \tag{10.2}$$

The eligible funds are sorted based on their smart alpha or CAPM alpha (model (10.1) with only the first two terms), and then evaluated based on their returns in the following year.

CAPM alpha is one of the most widely used skill measure of mutual fund managers (Berk and Van Binsbergen, 2016, Barber et al., 2014). However, several empirical studies (Kosowski et al., 2006, Fama and French, 2010) suggest that CAPM alpha is not persistent, i.e. if we select the mutual funds with high CAPM based on their performance in the last few years, these funds usually perform poorly later. Next we shall show that the smart alpha we propose is a much more persistent measure of mutual fund managers' skill.

Table 10.2: **Rankings of Mutual Funds Sorted on 5-year Past CAPM Alpha and Smart Alpha** Mutual funds are ranked on January 1 each year from 1996 to 2015 based on their CAPM alpha and smart alpha over the prior five years. The smart alpha is defined in (10.2). Row and column correspond to rankings based on CAPM alpha and smart alpha, respectively. $(0, 50\%]$, $(50\%, 70\%]$, $(70\%, 80\%]$, $(80\%, 90\%]$, and $(90\%, 100\%]$ indicate a fund's CAPM alpha or smart alpha that belongs to bottom five deciles, 5th to 7th deciles, 8th decile, 9th decile, and 10th decile, respectively.

CAPM-α \ smart-α	(0,50%]	(50%,70%]	(70%,80%]	(80%,90%]	(90%,100%]
(0,50%]	36.97%	8.04%	2.64%	1.64%	0.70%
(50%,70%]	8.40%	5.95%	2.72%	2.08%	0.84%
(70%,80%]	2.54%	2.94%	1.77%	1.74%	1.02%
(80%,90%]	1.53%	2.10%	1.79%	2.46%	2.12%
(90%,100%]	0.55%	0.97%	1.09%	2.08%	5.34%

10.3.1 Mismatch of smart alpha and CAPM alpha

We first study how CAPM alpha and smart alpha of the same mutual fund differ. Table 10.2 reports the results. A large number of mutual funds are ranked differently by CAPM alpha and smart alpha. For example, the last column of Table 10.2 indicates that among the funds in top smart alpha decile, on average $0.70\%/10\% = 7\%$ of them have CAPM alpha below the median, $(0.70\% + 0.84\% + 1.02\%)/10\% = 25.6\%$ of them have CAPM alpha below the top quintile, and $(0.70\%+0.84\%+1.02\%+2.12\%)/10\% = 46.6\%$ of them have CAPM alpha below the top decile. In other words, a large portion of the high smart alpha funds have produced relatively low CAPM alpha. Similarly, the last row of Table 10.2 implies that among funds in top CAPM alpha decile, on average $0.55\%/10\% = 5.5\%$ of them have smart alpha below the median, $(0.55\% + 0.97\% + 1.09\%)/10\% = 26.1\%$ of them have smart alpha below the top quintile, and $(0.55\% + 0.97\% + 1.09\% + 2.08\%)/10\% = 46.6\%$ of them have smart alpha below the top decile.

10.3.2 Persistence in smart-alpha-sorted mutual fund portfolios

To set the stage, we firstly examine the performance of mutual fund portfolios based on lagged return adjusted by market risk. i.e., CAPM alpha. This exerise is similar to Carhart (1997) and is used to compare with portfolios sorted by smart alpha later. The left panel of Table 10.3 reports the results. First, we find that the higher CAPM alpha deciles have larger total net asset (TNA) at the portfolio formation year. The top CAPM alpha decile manages \$1303 million assets on average, which is \$800 million higher than AUM of the bottom five deciles. This pattern is consistent with the conclusions in Berk and Van Binsbergen (2016), Barber et al. (2014) that aggregate investors chase CAPM alpha so that the highest CAPM alpha funds attract largest fund flows. On the other hand, the top CAPM alpha decile on average earns 1.0% less than the bottom five deciles in the next one year.

Next, we examine the performance of mutual fund portfolios based on their lagged smart alpha. The results are reported in the right panel of Table 10.3. First, we find smart alpha well predicts the fund's future performance: funds with high lagged smart alpha significantly outperform funds with low lagged smart alpha. For example, the value-weighted net-of-fee returns of top smart alpha decile are higher than returns of the bottom five smart alpha deciles by 2.0%.

Second, in contrast to the CAPM alpha sorted decile portfolios, fund portfolios with higher smart alpha don't necessarily have larger assets under management (AUM). For example, among ten smart alpha portfolios, the 7th decile has the largest AUM, which is \$ 200 million higher than AUM of the top smart alpha decile. When comparing to the top CAPM alpha decile, the top decile smart alpha funds manage assets that are on average 13% lower. This indicates that funds with superior smart alpha don't necessarily attract more fund flows from investors.

We also study persistence of fund performance at longer horizon. Figure 10.6 shows the average excess return of each decile portfolio in the first five years after funds are ranked based on CAPM alpha and smart alpha, respectively. The top smart alpha decile maintains a persistently higher mean return a full five years after

Portfolio	CAPM Alpha Sorted					Smart Alpha Sorted				
	TNA	Ret	SR	CAPM-α	FFC-α	TNA	Ret	SR	CAPM-α	FFC-α
1 (low)	251.0	6.1	32.5	-2.02	-1.47	239.2	5.6	36.2	-1.23	-1.72**
				(-1.55)	(-1.24)				(-1.07)	(-2.15)
2	397.0	6.4	34.9	-1.58	-1.35	352.4	6.0	36.3	-1.38**	-1.50**
				(-1.38)	(-1.47)				(-2.30)	(-2.53)
3	475.9	6.7	40.5	-0.64	-0.67	523.6	5.4	33.0	-1.90***	-1.93***
				(-0.81)	(-0.84)				(-3.32)	(-3.53)
4	592.0	6.4	39.0	-0.92	-0.94	706.7	5.8	36.7	-1.23*	-1.06
				(-1.55)	(-1.59)				(-1.76)	(-1.58)
5	769.8	6.6	42.0	-0.36	-0.51	739.1	5.3	33.6	-1.74***	-1.94***
				(-0.48)	(-0.74)				(-2.93)	(-3.48)
6	964.5	6.8	44.4	-0.03	-0.19	964.9	6.1	38.1	-1.07**	-1.16**
				(-0.06)	(-0.35)				(-2.19)	(-2.37)
7	1030.1	6.2	40.7	-0.62	-0.83	1311.7	5.8	37.5	-1.13**	-1.18**
				(-1.17)	(-1.62)				(-2.01)	(-2.49)
8	1103.9	6.7	43.9	-0.09	-0.43	1065.9	6.5	41.9	-0.44	-0.78
				(-0.14)	(-0.76)				(-0.74)	(-1.36)
9	1235.8	6.9	39.2	0.28	-0.52	1084.9	7.0	44.2	-0.02	-0.43
				(0.33)	(-0.52)				(-0.02)	(-0.61)
10 (high)	1303.9	5.5	32.2	-1.88	-2.11*	1056.2	7.6	48.7	0.73	0.51
				(-1.59)	(-1.86)				(0.84)	(0.61)
Top 5%	1469.0	4.4	23.9	-3.12*	-3.53**	1056.2	8.0	51.8	1.37	1.08
				(-1.70)	(-2.05)				(1.21)	(0.98)
10-1	1052.9	-0.6	-8.3	0.14	-0.63	924.9	2.1	47.3	1.97**	2.24**
				(0.08)	(-0.38)				(1.99)	(2.30)
10-1:5	803.7	-1.0	-16.8	-0.90	-1.21	645.1	2.0	56.4	2.23***	2.06***
				(-0.69)	(-1.21)				(2.78)	(2.72)

Table 10.3: **Portfolios of Mutual Funds Formed on Lagged 5-year CAPM Alpha and Lagged 5-year Smart Alpha.** Mutual funds are sorted on January 1 each year into value-weighted portfolios based on their CAPM alpha and smart alpha over the prior five years, respectively. Funds with the highest past alpha comprise decile 10 and funds with the lowest past five-year alpha comprise decile 1. TNA is total asset under management at the time of portfolio formation; Ret is the annualized monthly return of fund portfolio; SR is the Sharpe ratio; CAPM-α and FFC-α are the monthly excess return on the market portfolio MKT and the four factor portfolios MKT,SMB, HML, and UMD defined in Carhart (1997). t-statistics are reported in parentheses. ***,**, and * denote statistical significance at 1, 5, and 10% level.

the portfolio is initially formed. Besides, the ranks are quite persistent over the next five years except for the lowest decile. Apparently, a relatively high smart alpha is a reasonably good indicator of the relative long-term expected return on a mutual fund. On the contrary, for CAPM alpha deciles, the mean returns of the ten deciles converge after one year and the top CAPM alpha decile earns lowest average return in the next year.

10.3.3 Top-ranked mutual funds

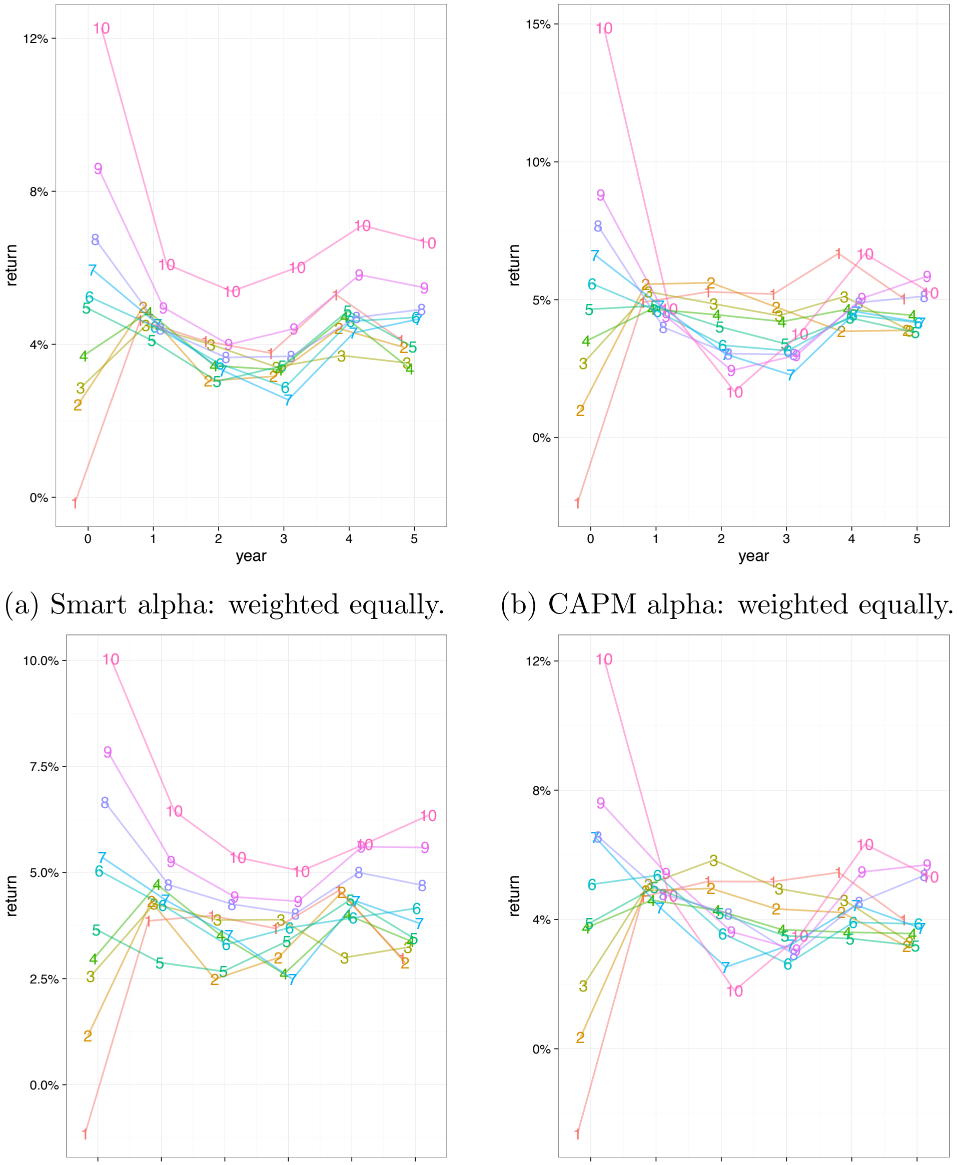
Roughly speaking, we can categorize the mutual funds in two ways: skilled funds (high smart alpha) vs. unskilled funds (low smart alpha), and appealing funds (high CAPM alpha) vs. unappealing funds (low CAPM alpha). Based on this, we can classifies top-ranked mutual funds (either by smart alpha or CAPM alpha) into three types:

1. Overestimated: those funds with high CAPM alpha but low smart alpha.

2. Skilled: those funds with high CAPM alpha and high smart alpha.

3. Underestimated: those funds with high smart alpha but low CAPM alpha.

Table 10.4 shows the performance of each of these three types of funds. For example, if we only consider the top decile as high and all the other nine deciles as low (Panel A), overestimated funds have mean return 5.1%, skilled funds have mean return 6.2%, and underestimated funds have mean return 9.1%. The order remains when we change the definition of high alpha to two deciles (Panel B) or three deciles (Panel C). The most probable reason of phenomenon is the irrational cash flow into the high CAPM alpha funds. For an economic-theory explanation and more results, we refer the reader to Song and Zhao (2016).



(a) Smart alpha: weighted equally.

(b) CAPM alpha: weighted equally.

(c) Smart alpha: weighted by value.

(d) CAPM alpha: weighted by value.

Figure 10.6: **Post-formation returns on portfolios of mutual funds sorted on lagged CAPM alpha and smart alpha**. In each year from 1996 to 2015, funds are ranked into value-weight and equal weight decile portfolios based on lagged CAPM alpha and smart alpha. The lines in the group represent the excess returns on the decile portfolios in the year subsequent to the formation year and in each of the next five years after formation.

	TNA	Proportion	One-year			Two-year		
			ExRet	CAPM-α	FFC-α	ExRet	CAPM-α	FFC-α
Panel A: Top Decile								
CAPM $\setminus$ Smart	1259.3	4.66%	5.1	-2.75	-2.98	3.5	-3.83	-3.98
				(-1.96)	(-2.06)		(-2.35)	(-2.54)
CAPM $\cap$ Smart	1279.5	5.34%	6.2	-0.86	-1.16	5.2	-1.57	-1.85
				(-0.69)	(-1.00)		(-1.21)	(-1.51)
Smart $\setminus$ CAPM	1097.5	4.66%	9.1	2.42	2.00	8.7	2.15	2.03
				(2.41)	(1.97)		(2.20)	(2.03)
Panel B: Top Two Deciles								
CAPM $\setminus$ Smart	1157.1	8.00%	5.5	-1.68	-2.24	3.4	-3.50	-3.81
				(-1.49)	(-2.29)		(-2.76)	(-3.15)
CAPM $\cap$ Smart	1260.4	12.00%	6.9	0.05	-0.41	6.0	-0.70	-1.06
				(0.05)	(-0.49)		(-0.74)	(-1.20)
Smart $\setminus$ CAPM	937.9	8.00%	7.8	0.72	0.70	8.8	2.09	1.83
				(0.78)	(0.76)		(1.99)	(1.71)
Panel C: Top Three Deciles								
CAPM $\setminus$ Smart	1053.1	10.59%	6.3	-0.66	-1.40	5.0	-1.62	-2.15
				(-0.68)	(-1.63)		(-1.60)	(-2.36)
CAPM $\cap$ Smart	1242.8	19.41%	6.9	0.08	-0.33	6.1	-0.51	-0.93
				(0.10)	(-0.47)		(-0.63)	(-1.24)
Smart $\setminus$ CAPM	851.3	10.59%	7.1	-0.11	-0.13	8.5	1.51	1.35
				(-0.13)	(-0.15)		(1.51)	(1.36)

Table 10.4: **Post-formation returns on mutual funds sorted on 5-year past CAPM alpha and 5-year past smart alpha.** Mutual funds are ranked on January 1 each year from 1994 to 2015 based on their CAPM alpha and smart alpha over the prior five years. Panel A, B, and C select top 10%, top 20%, and top 30% CAPM alpha funds and smart alpha funds, respectively. In each panel, CAPM $\setminus$ Smart includes funds that have CAPM alpha in the top group and smart alpha out of the top group; CAPM $\cap$ Smart includes funds that have both CAPM alpha and smart alpha in the top group; Smart $\setminus$ CAPM includes funds that have smart alpha in the top group and CAPM alpha out of the top group. TNA is total asset under management at the time of portfolio formation (in millions); ExRet is the annualized monthly return of value-weighted portfolio; CAPM-α and FFC-α are the monthly excess return within the next one year and two years on the market portfolio MKT and the four factor portfolios MKT,SMB, HML, and UMD defined in Carhart (1997).

Bibliography

Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica 74*(1), 235–267.

Anderson, T. W. and H. Rubin (1956). Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, Volume 5.

Austin, P. C. and E. A. Stuart (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine 34*(28), 3661–3679.

Bai, J. and K. Li (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics 40*(1), 436–465.

Bai, J. and K. Li (2014). Theory and methods of panel data models with interactive effects. *The Annals of Statistics 42*(1), 142–170.

Bai, J. and K. Li (2015). Maximum likelihood estimation and inference for approximate factor models of high dimension. *The Review of Economics and Statistics to appear.*

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bai, J. and S. Ng (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica 74*(4), 1133–1150.

Balke, A. and J. Pearl (1994). Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Volume 1, pp. 320–237. MIT press.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*(4), 962–973.

Barber, B. M., X. Huang, and T. Odean (2014). Which risk factors matter to investors? evidence from mutual fund flows. *ssrn.2408231*.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics 29*(4), 1165–1188.

Berk, J. B. and J. H. Van Binsbergen (2016). Assessing asset pricing models using revealed preference. *Journal of Financial Economics 119*(1), 1–23.

Blalock, E. M., J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield (2004). Incipient alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences of the United States of America 101*(7), 2173–2178.

Bollen, K. A. (2014). *Structural equations with latent variables.* John Wiley & Sons.

Bollen, K. A. and J. Pearl (2013). Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research*, pp. 301–328. Springer.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association 71*(356), 791–799.

Brys, G., M. Hubert, and A. Struyf (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics 13*(4), 996–1017.

Buja, A., W. Stuetzle, and Y. Shen (2005). Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft*.

Caliendo, M. and S. Kopeinig (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys 22*(1), 31–72.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance 52*(1), 57–82.

Caruana, R. and A. Niculescu-Mizil (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pp. 161–168. ACM.

Casella, G. and S. R. Schwartz (2000). Comment on "causal inference without counterfactuals". *Journal of the American Statistical Association 95*(450), 425–427.

Chan, K. C. G., S. C. P. Yam, and Z. Zhang (2015). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of Royal Statistical Society, Series B (Methodology) 78.*

Chandrasekaran, V., P. A. Parrilo, and A. S. Willsky (2012, 08). Latent variable graphical model selection via convex optimization. *Ann. Statist. 40*(4), 1935–1967.

Clarke, S. and P. Hall (2009). Robustness of multiple testing procedures against dependence. *The Annals of Statistics 37*(1), 332–358.

Craig, A., O. Cloarec, E. Holmes, J. K. Nicholson, and J. C. Lindon (2006). Scaling and normalization effects in nmr spectroscopic metabonomic data sets. *Analytical Chemistry 78*(7), 2262–2267.

Crump, R., V. J. Hotz, G. Imbens, and O. Mitnik (2006). Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical Report 330, National Bureau of Economic Research.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association 95*(450), 407–424.

De La Fuente, A., N. Bing, I. Hoeschele, and P. Mendes (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics 20*(18), 3565–3574.

Dehejia, R. H. and S. Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association 94*(448), 1053–1062.

Desai, K. H. and J. D. Storey (2012). Cross-dimensional inference of dependent high-dimensional data. *Journal of the American Statistical Association 107*(497), 135–151.

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*(418), 376–382.

Diamond, A. and J. S. Sekhon (2013a). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics 95*(3), 932–945.

Diamond, A. and J. S. Sekhon (2013b). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics 95*(3), 932–945.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association 102*, 93–103.

Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association 105*(491), 1042–1055.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance 25*(2), 383–417.

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *the Journal of Finance 47*(2), 427–465.

Fama, E. F. and K. R. French (2010). Luck versus skill in the cross-section of mutual fund returns. *The journal of finance 65*(5), 1915–1947.

Fan, J. and X. Han (2013). Estimation of false discovery proportion with unknown dependence. *arXiv:1305.7007*.

Fan, J., X. Han, and W. Gu (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association 107*(499), 1019–1035.

Fare, T. L., E. M. Coffey, H. Dai, Y. D. He, D. A. Kessler, K. A. Kilian, J. E. Koch, E. LeProust, M. J. Marton, M. R. Meyer, et al. (2003). Effects of atmospheric ozone on microarray data quality. *Analytical chemistry 75*(17), 4672–4675.

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd Edinburgh.

Friedman, J., T. Hastie, and R. Tibshirani (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics 28*(2), 337–407.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Friguet, C., M. Kloareg, and D. Causeur (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association 104*(488), 1406–1415.

Gagnon-Bartsch, J., L. Jacob, and T. Speed (2013). Removing unwanted variation from high dimensional data with negative controls. Technical report, Technical Report 820, Department of Statistics, University of California, Berkeley.

Gagnon-Bartsch, J. A. and T. P. Speed (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics 13*(3), 539–552.

Gasch, A. P., P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell 11*(12), 4241–4257.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Graham, B. S., C. C. D. X. Pinto, and D. Egel (2012). Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies 79*(3), 1053–1079.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). A kernel two-sample test. *The Journal of Machine Learning Research 13*(1), 723–773.

Grzebyk, M., P. Wild, and D. Chouanière (2004). On identification of multi-factor models with correlated residuals. *Biometrika 91*(1), 141–151.

Gu, X. S. and P. R. Rosenbaum (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics 2*(4), 405–420.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*(2), 315–332.

Hainmueller, J. (2011). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis 20*, 25–46.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association 99*(467), 609–618.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *Elements of Statistical Learning.* Springer.

Hazlett, C. (2013). A balancing method to equalize multivariate densities and reduce bias without a specification search. *Working draft*.

Heckman, J. J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies 65*(2), 261–294.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of conomic Studies 64*(4), 605–654.

Hirano, K. and G. Imbens (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology 2*, 259–278.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Ho, D. E., K. Imai, G. King, and E. A. Stuart (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis 15*(3), 199–236.

Hoefer, C. (2016). Causal determinism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016 ed.).

Hofmann, T., B. Schölkopf, and A. J. Smola (2008). Kernel methods in machine learning. *The Annals of Statistics*, 1171–1220.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*, 945–960.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*.

Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 243–263.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics 86*(1), 4–29.

Imbens, G. W. and D. B. Rubin (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 305–327.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, T. P. Speed, et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics 4*(2), 249–264.

Jin, J. (2012). Comment. *Journal of the American Statistical Association 107*(499), 1042–1045.

Kang, J. D. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Kim, J. K. and M. Park (2010). Calibration estimation in survey sampling. *International Statistical Review 78*(1), 21–39.

Koller, D. and N. Friedman (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Korn, E. L., J. F. Troendle, L. M. McShane, and R. Simon (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference 124*(2), 379–398.

Kosowski, R., A. Timmermann, R. Wermers, and H. White (2006). Can mutual fund stars really pick stocks? new evidence from a bootstrap analysis. *The Journal of finance 61*(6), 2551–2595.

Kuroki, M. and J. Pearl (2014). Measurement bias and effect restoration in causal inference. *Biometrika 101*, 423–437.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 604–620.

Lan, W. and L. Du (2014). A factor-adjusted multiple testing procedure with application to mutual fund selection. *arXiv:1407.5515*.

Laplace, P.-S. (1814). *A philosophical essay on probabilities*. Courcier. English translation by Truscott, Frederick Wilson and Emory, Frederick Lincoln.

Lauritzen, S. L. (2004). Discussion on causality. *Scandinavian Journal of Statistics 31*(2), 189–193.

Lazar, C., S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y. Weiss-Solís, R. Duque, H. Bersini, and A. Nowé (2013). Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in bioinformatics 14*(4), 469–490.

Lee, B. K., J. Lessler, and E. A. Stuart (2010). Improving propensity score weighting using machine learning. *Statistics in medicine 29*(3), 337–346.

Lee, B. K., J. Lessler, and E. A. Stuart (2011). Weight trimming and propensity score weighting. *PloS one 6*(3), e18174.

Leek, J. T., R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics 11*(10), 733–739.

Leek, J. T. and J. D. Storey (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics 3*(9), 1724–1735.

Leek, J. T. and J. D. Storey (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences 105*(48), 18718–18723.

Li, J. and P.-S. Zhong (2014). A rate optimal procedure for sparse signal recovery under dependence. *arXiv preprint arXiv:1410.2839*.

Lin, D. W., I. M. Coleman, S. Hawley, C. Y. Huang, R. Dumpit, D. Gifford, P. Kezele, H. Hung, B. S. Knudsen, A. R. Kristal, et al. (2006). Influence of surgical manipulation on prostate gene expression: implications for molecular correlates of treatment effects and disease prognosis. *Journal of clinical oncology 24*(23), 3763–3770.

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The Journal of Finance 20*(4), 587–615.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine 23*(19), 2937–2960.

Markowitz, H. (1952). Portfolio selection. *The journal of finance 7*(1), 77–91.

Maronna, R. A., D. R. Martin, and V. J. Yohai (2006). *Robust statistics: Theory and Methods*. John Wiley & Sons, Chichester.

Mason, L., J. Baxter, P. Bartlett, and M. Frean (1999). Boosting algorithms as gradient descent in function space. NIPS.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods 9*(4), 403.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models*. CRC press.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica: Journal of the econometric society*, 768–783.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472. Translated and edited by Dabrowska, DM and Speed, TP from the Polish original.

Normand, S.-L. T., M. B. Landrum, E. Guadagnoli, J. Z. Ayanian, T. J. Ryan, P. D. Cleary, and B. J. McNeil (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of clinical epidemiology 54*(4), 387–398.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics 92*(4), 1004–1016.

Owen, A. B. (2005). Variance of the number of false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(3), 411–426.

Owen, A. B. and J. Wang (2015). Bi-cross-validation for factor analysis. *arXiv:1503.03515*.

Owen, A. B. and J. Wang (2016). Bi-cross-validation for factor analysis. *Statistical Science (to appear)*.

Pearl, J. (1993). Comment: Graphical models, causality and intervention. *Statistical Science 8*(3), 266–269.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*(4), 669–688.

Pearl, J. (2000). Comment on "causal inference without counterfactuals". *Journal of the American Statistical Association 95*(450), 428–431.

Pearl, J. (2009a). *Causality*. Cambridge university press.

Pearl, J. (2009b). Letter to the editor: Remarks on the method of propensity score. *Department of Statistics, UCLA*.

Pesaran, M. (2004). General Diagnostic Tests for Cross Section Dependence in Panels. Cambridge Working Papers in Economics 0435, Faculty of Economics, University of Cambridge.

Peters, J., P. Bühlmann, and N. Meinshausen (2015). Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint arXiv:1501.01332*.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics 38*(8), 904–909.

Ransohoff, D. F. (2005). Bias as a threat to the validity of cancer molecular-marker research. *Nature Reviews Cancer 5*(2), 142–149.

Reid, C. (1982). *Neyman from life*. Springer.

Rhodes, D. R. and A. M. Chinnaiyan (2005). Integrative analysis of the cancer transcriptome. *Nature genetics 37*, S31–S37.

Richardson, T. S. and J. M. Robins (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper 128*(30), 2013.

Ridgeway, G. et al. (2006). gbm: Generalized boosted regression models. *R package version 1*(3).

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment: Performance of double-robust estimators when inverse probability weights are highly variable. *Statistical Science 22*(4), 544–559.

Robins, J. M., A. Rotnitzky, and L. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*, 846–866.

Robins, J. M. and N. Wang (2000). Inference for imputation estimators. *Biometrika 87*(1), 113–124.

Rosenbaum, P. and D. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rosenbaum, P. and D. Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association 79*, 516–524.

Rosenbaum, P. R. (2002). *Observational studies.* Springer.

Rosenbaum, P. R. and D. B. Rubin (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician 39*(1), 33–38.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 159–183.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioral statistics 2*(1), 1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34–58.

Rubin, D. B. (1980). Comment on "randomization analysis of experimental data: The fisher randomization test". *Journal of the American Statistical Association 75*(371), 591–593.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science 5*(4), 472–480.

Rubin, D. B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine 28*(9), 1420–1423.

Rubin, D. B. (2011). Causal inference using potential outcomes. *Journal of the American Statistical Association*.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association 66*(336), 783–801.

Schwartzman, A. (2010). Comment. *Journal of the American Statistical Association 105*(491), 1059–1063.

Schwartzman, A., R. F. Dougherty, and J. E. Taylor (2008). False discovery rate analysis of brain diffusion direction maps. *The Annals of Applied Statistics 2*(1), 153–175.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The Matching package for R. *Journal of Statistical Software 42*(7), 1–52.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics 1*(1), 43–62.

Shafer, G. (2000). Comment on “causal inference without counterfactuals”. *Journal of the American Statistical Association 95*(450), 438–442.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance 19*(3), 425–442.

She, Y. and A. B. Owen (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association 106*(494), 626–639.

Singh, D., S. M. Fox, R. Tal-Singer, J. Plumb, S. Bates, P. Broad, J. H. Riley, and B. Celli (2011). Induced sputum genes associated with spirometric and radiological disease severity in COPD ex-smokers. *Thorax 66*(6), 489–495.

Smith, J. A. and P. E. Todd (2005). Does matching overcome lalonde’s critique of nonexperimental estimators? *Journal of econometrics 125*(1), 305–353.

Song, Y. and Q. Zhao (2016). On the persistence of mutual fund skills. *working draft*.

Spirtes, P., C. N. Glymour, and R. Scheines (2000). *Causation, prediction, and search.* MIT press.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 187–205.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science 25*(1), 1–21.

Sun, W. and T. Cai (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 393–424.

Sun, Y. (2011). *On latent systemic effects in multiple hypotheses.* Ph. D. thesis, Stanford University.

Sun, Y., N. R. Zhang, and A. B. Owen (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data. *The Annals of Applied Statistics 6*(4), 1664–1688.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association 101*, 1619–1637.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika 97*(3), 661–682.

Treynor, J. L. (1961). Toward a theory of market value of risky assets. *Unpublished manuscript 6*.

Tusher, V. G., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences 98*(9), 5116–5121.

Uhler, C., G. Raskutti, P. Bühlmann, B. Yu, et al. (2013). Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics 41*(2), 436–463.

Vawter, M. P., S. Evans, P. Choudary, H. Tomita, J. Meador-Woodruff, M. Molnar, J. Li, J. F. Lopez, R. Myers, D. Cox, et al. (2004). Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes. *Neuropsychopharmacology 29*(2), 373–384.

Wager, S. and S. Athey (2015). Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*.

Wahba, G. (1990). *Spline models for observational data*, Volume 59. SIAM.

Wang, J., Q. Zhao, T. Hastie, and A. B. Owen (2015). Confounder adjustment in multiple hypothesis testing. *under revision for Annals of Statistics*.

Wang, S., G. Cui, and K. Li (2015). Factor-augmented regression models with structural change. *Economics Letters 130*, 124–127.

Wright, S. (1921). Correlation and causation. *Journal of agricultural research 20*(7), 557–585.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics 5*(3), 161–215.

Zhao, Q. (2016). Covariate balancing propensity score by tailored loss functions.

Zhao, Q. and D. Percival (2015). Primal-dual Covariate Balance and Minimal Double Robustness via Entropy Balancing. *ArXiv e-prints 1501.03571.*

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association 110*(511), 910–922.