

國立臺灣大學社會科學院經濟學系
碩士論文



Department of Economics
College of Social Sciences
National Taiwan University
Master Thesis

多維模糊斷點迴歸設計下的新估計方法
New Estimation Method for Multidimensional
Fuzzy Regression Discontinuity Design

林真
Chen Lin

指導教授：管中閔 博士
Advisor: Chung-Ming Kuan, Ph.D

中華民國 107 年 7 月
July, 2018

國立臺灣大學碩士學位論文
口試委員會審定書

多維模糊斷點迴歸設計下的新估計方法
New Estimation Method for Multidimensional Fuzzy
Regression Discontinuity Design

本論文係 林真君 (學號 R05323002) 在國立臺灣大學經濟學系
完成之碩士學位論文，於民國一零七年七月二日承下列考試委員審查
通過及口試及格，特此證明

口試委員：

管中文

(指導教授)

許育達

楊子耀

江源芳

銘謝

本篇論文能順利付梓，首先當要感謝我的指導教授，管中閔教授，在我探索研究方向時給予我相當多寶貴的建議。還記得眾多午夜，和老師在通訊軟體上來回討論的情景。在專業知識上有所精進也好、初次撰寫論文，有許多規則不懂也好，老師總不厭其煩地引領我、指導我，讓我了解自己也有完成整篇研究的能力。

我也相當感謝諸位口試委員：江淳芳老師、許育進老師及楊子霆老師。除了相當有耐心地聽完我的報告外，也對本篇論文及我的口語表達指出不少應該修正和加強的部分，使其更臻完美。

此外，我也想向陳釗而老師、日本九州大學瀧本太郎老師及二位所屬眾指導學生致上謝意，讓我得以在二位主要召集的聯合研討會上報告本篇論文的雛形。在準備報告的過程中，使我好好審視了整篇論文的邏輯和架構。

研究和撰寫論文誠如管老師在確定指導我時，所曾向我提及，是份艱辛且寂寞的工作。然而，每當夜深人靜，懷疑自己之際，總有諸位師長、家人、同事、朋友的支持縈繞心頭。每段建議、每句評論，乃至每聲加油，共同成為這份作品的基石。沒有各位無條件的支持，本篇論文絕對無法問世。

林真 謹誌於國立臺灣大學

經濟學研究所

2018年7月18日

摘要

當政策、療程之施行與否取決於受試者是否通過某些特定標準時，研究者可以使用斷點迴歸(Regression discontinuity design; RDD)對局部平均處理效應(LATE)做不偏估計。在本篇論文中，我們將會回顧多維模糊斷點迴歸(Multidimensional RDD)之概念與假設，在其中處置(Treatment)施行與否取決於多個標準，而受試者也不盡然都遵從指示接受或不接受處置。本文第一個貢獻為推廣 Lo (2017)及 Hsu、Kuan 與 Lo (2018)文中概念，並指出傳統的估計方法未考慮資料中潛在的異質性，從而可能導致估計偏誤。此外，我們指出兩個異質性的潛在來源：指標變數(Assignment variable、Running variable)邊際效果不同，以及接受處置的機率不同。由此我們針對多維模糊斷點迴歸提出平均法(Average Method)以及交點法(Intersection Method)，成功克服資料中的異質性。在模擬中，我們發現我們提出的方法相較於傳統估計法確實能更準確地估計出處置效果，顯示我們的方法能夠在更普遍的環境下進行估計。

關鍵詞：異質性、局部平均處理效應(LATE)、局部多項式迴歸、斷點迴歸、二階段最小平方法(2SLS)。

JEL 分類：C21、C26、C90。



Abstract

Regression discontinuity design (RDD) is an easy, yet rigorous setting allowing researchers to unbiasedly estimate local average treatment effect, particularly when the treatment is determined by whether subjects pass certain pre-specified thresholds or not. In this thesis, we shall review basic concepts and assumptions of **multidimensional fuzzy RDD**, in which there are multiple thresholds, and we do not require all subjects to follow the assignment rule. As the first contribution, we generalize the idea in Lo (2017) and Hsu, Kuan, Lo (2018), pointing out traditional estimation methods fail to take potential heterogeneity in the dataset into account and hence induce biased estimates. In addition, we identify the two potential sources of heterogeneity: different marginal effect of running variables and different treatment probabilities. With this in mind, we propose *average method* and *intersection method* for multidimensional fuzzy RDD, overcoming potential heterogeneity in the dataset. In the simulation study, we find out that our methods do produce a more accurate estimate than traditional methods, showing that our methods can accommodate much more general settings than traditional ones can do.

Keywords: Heterogeneity, Local Average treatment effect, Local polynomial regression, Regression discontinuity design (RDD), Two stage least square estimation (2SLS)

JEL classification: C21, C26, C90



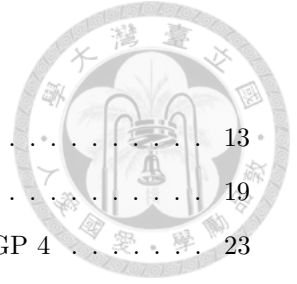
Contents



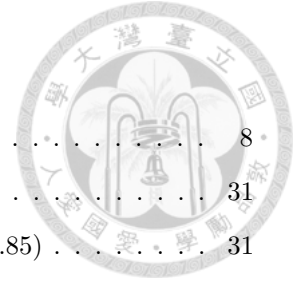
1	Introduction	4
2	Regression Discontinuity Design	6
3	1-dim Fuzzy RDD	8
3.1	Problem Formulation	8
3.2	Identification	10
3.3	Estimation Strategy	12
4	Multidimensional Fuzzy RDD	14
4.1	Problem Formulation and assumptions	15
4.2	Estimation Method	16
5	Simulation	21
5.1	Setup	21
5.2	Results	24
6	Conclusion	26
	Reference	27
	Appendix	30

List of Figures

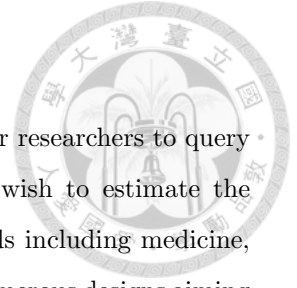
1	An illustration of one dimensional sharp RDD	13
2	Difference in sample points used in both methods	19
3	Distribution of (W_i, W_{1i}, W_{2i}) in the four quadrants for DGP 3 and DGP 4	23
4	Distribution of treatment effect assignment in the four quadrants for DGP 3	24
5	Distribution of treatment assignment in the four quadrants for DGP 4	30



List of Tables



1	Classification of individuals.	8
2	Simulation Result: DGP 1 with local linear fitting and $(a, b) = (0, 0.7)$	31
3	Simulation Result: DGP 1 with local linear fitting and $(a, b) = (0.15, 0.85)$	31
4	Simulation Result: DGP 1 with local linear fitting and $(a, b) = (0.3, 1)$	31
5	Simulation Result: DGP 1 with local linear fitting and $(a, b) = (0, 0.9)$	32
6	Simulation Result: DGP 1 with local linear fitting and $(a, b) = (0.05, 0.95)$	32
7	Simulation Result: DGP 1 with local linear fitting and $(a, b) = (0.1, 1)$	32
8	Simulation Result: DGP 2 with local linear fitting and $(a, b, c) = (0.15, 0.65, 1)$	33
9	Simulation Result: DGP 2 with local linear fitting and $(a, b, c) = (0.15, 0.5, 0.85)$	33
10	Simulation Result: DGP 2 with local linear fitting and $(a, b, c) = (0.15, 0.75, 1)$	33
11	Simulation Result: DGP 2 with local linear fitting and $(a, b, c) = (0.15, 0.75, 0.85)$	34
12	Simulation Result: DGP 3 with local linear fitting and $(a, b) = (0.15, 0.85)$	34
13	Simulation Result: DGP 3 with local linear fitting and $(a, b) = (0.05, 0.95)$	34
14	Simulation Result: DGP 3 with local quadratic fitting and $(a, b) = (0.15, 0.85)$	35
15	Simulation Result: DGP 3 with local quadratic fitting and $(a, b) = (0.05, 0.95)$	35
16	Simulation Result: DGP 4 with local quadratic fitting and $(a, b) = (0.15, 0.85)$	35
17	Simulation Result: DGP 4 with local quadratic fitting and $(a, b) = (0.05, 0.95)$	36
18	Statistics of standardized chosen bandwidth: DGP 3 with local linear fitting and $(a, b) = (0.15, 0.85)$	36
19	Statistics of standardized chosen bandwidth: DGP 3 with local quadratic fitting and $(a, b) = (0.15, 0.85)$	36



1 Introduction

After a new treatment has been introduced, it is common for policy makers or researchers to query whether the treatment is indeed effective or not; if possible, one may even wish to estimate the effect quantitatively. In fact, besides econometricians, many researchers in fields including medicine, pedagogy, and even politics have dug into treatment effect estimation. Among numerous designs aiming at this problem, **regression discontinuity design (RDD)** has gained much attention recently.

Originally proposed in Thistlethwaite and Campbell (1960), RDD is a facile way allowing for local average treatment effect estimation when the assignment rule is known to the researchers. More specifically, in an RDD, whether a subject is eligible for treatment depends on whether it passes some pre-specified threshold or not. For example, students may require additional math classes if they fail to pass an exam; patients, say, with high blood pressure or cholesterol may be diagnosed with certain disease and call for certain kinds of treatment. Despite its potential, however, it is not until four decades after RDD has been introduced, Hahn, Todd, and Van der Klaauw (2001) formalize the setting in the language of Rubin casual model (Rubin, 1974) that RDD begins to receive wide attention¹. Actually, RDD can be adopted in various fields including sociology (Hahn, Todd and Van der Klaauw, 1999), politics (Eggers, Fowler, Hainmueller, Hall, and Snyder, 2015) and epidemiology (Bor, Moscoe, Mutevedzi, Newell and Bärnighausen, 2014).

The main advantage of RDD is its facility. The basic idea of RDD is to exploit the continuity of the interested outcome variable. One simply fits polynomials for sample points just above and just below the threshold respectively; then attribute the difference of intercept at the cutoff to local treatment effect at the threshold for those who follow the rule. In addition, when randomized experiments are not available, RDD may serve as an alternative, especially in medical studies. As mentioned in Moscoe, Bor and Bärnighausen (2015), if a treatment has been ubiquitously accepted as indispensable in medical care, it would be hard to conduct randomized experiments. In this case, researchers can utilize data from medical records to do inference. Even when randomized experiment is a possible solution, using previously collected data saves time and cost.

Still another merit of RDD is its flexibility. Although treatment is assigned according whether the observations pass certain cutoff or not, we do not require all subjects to follow the rule. Specifically, if the assignment rule is enforced, then one faces a **sharp RDD**; otherwise one has to resort to a **fuzzy RDD**, which in fact includes the former as a special case. Therefore, we shall focus on fuzzy RDD, in which one may use whether the subject actually receives treatment or not as an instrument to unbiasedly estimate the desired local average treatment effect for those who comply with the rule

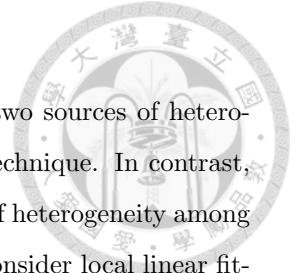
¹For a more detailed review of the history of RDD in the second half of the twentieth century, see Cook (2008).

at the cutoff.

In pioneering Thistlewaite and Campbell (1960), as well as many researches following its methodology, the treatment depends on solely one standard. In reality, however, there are also scenarios where multiple standards are present; for example, see Jacob and Lefgren (2004). Many works have hence considered RDD with multiple assignment variables and thresholds, or multidimensional RDD; see Imbens and Zajonc (2011) or Wong, Steiner and Cook (2013). When there is only one standard, it is natural to separate the observations according whether they pass the threshold or not; nonetheless, if there are, say, two standards, one cannot be sure whether those with no, one, or two passed standards have similar characteristics. Putting heterogeneous subjects in the same group naïvely would lead to a biased estimate. Regrettably, as indicated in Lo (2017) and Hsu, Kuan and Lo (2018), most existing estimation method fail to take heterogeneity in the dataset into consideration, thus producing a biased estimate. In contrast, Lo (2017) and Hsu et al. (2018) use information contained in observations carefully, proposing *intersection method* and *average method* for sharp RDD, both of which have the flexibility to tackle with heterogeneity in the dataset.

This thesis aims to go beyond Lo (2017) and Hsu et al. (2018), generalizing their idea from sharp RDD to **fuzzy RDD**. We indicate that the heterogeneity in the dataset may result from **different marginal effect of running variables** and/or **different treatment probabilities**, with the latter can only be detected in fuzzy RDD. To be specific, suppose students are required to attend additional classes if they fail one of their reading or math exams or both, and we are interested in the effect of those additional classes on these students' score on another exam two months later. It may be intuitive to place those fail one subject and those fail both in the same group since they are at least more likely to attend additional classes. However, there is no guarantee that the average marginal effect of original reading (or math) score on performance two months later is the same for both kinds of students; what's worse, if the additional classes are not compulsory, students failing one subject may have a different attendance rate from those failing two subjects, thus creating heterogeneity in students who are eligible for treatment and potential bias in traditional estimation methods.

To overcome such problem, we modify and generalize the idea in Lo (2017) and Hsu et al. (2018) to accommodate it in fuzzy RDD. Take the case where two standards are present for example, we treat observations with no, one and two passed standards differently. By comparing subjects with different number of passed standards respectively and average the results, we shall get an unbiased estimate of local average treatment effect. We shall refer to this procedure **average method**. On the other hand, to lessen computational burden, we can also drop observations which pass exactly one cutoff, and then compare the remaining subjects to get another unbiased estimate. The latter procedure, which we name **intersection method**, though easier to compute, but in most cases suffer from a



larger standard deviation due to information loss.

According to our simulation studies, we find out that the aforementioned two sources of heterogeneity do result in less accurate estimate when using traditional estimation technique. In contrast, intersection method and average method still acquire an unbiased estimate even if heterogeneity among dataset exists. On the other hand, as Lo (2017) and Hsu et al. (2018) mainly consider local linear fitting in their work, we explore whether local quadratic polynomial fitting can generate a more desirable estimate. Still, quadratic fitting still gives unbiased estimate, meaning that one can use higher-order polynomials to allow for more flexibility. Hopefully, this work can contribute to the emerging trend of researches into multidimensional RDD, calling attention to scrutinizing observations cautiously and using data at hand more wisely in order to accommodate more general settings.

The rest of this paper is arranged as follows: in section 2, we shall introduce the basic ideas and aims in RDD. In section 3 and section 4, we will formalize one dimensional RDD and multidimensional RDD respectively, presenting assumptions needed, as well as identifying the desired local average treatment effect at the threshold and introducing estimation methods. We will verify our argument by simulation in section 5. Finally, we will conclude our discussion in section 6.

2 Regression Discontinuity Design

Treatment effect estimation has always been of central importance not only in social science, but also in many other fields like medicine or pedagogy. Doctors, for example, may be curious about whether a new medical treatment is effective or not. Teachers may want to know whether a new teaching program really help students in test performance. To estimate treatment effect, ideally, we would like to conduct *randomized experiments*, in which individuals are randomly assigned being treated or not; then a comparison between those treated (treatment group) and those not treated (control group) gives the desired treatment effect. In reality, however, randomized experiments are not always feasible or they are just too costly. For example, in medical applications, patients may be ineligible for treatment because of random assignment, which sometimes is controversial.

When randomized experiments cannot be done, researchers turn to *quasi-experiment designs*. In quasi-experiment designs, samples are assigned to be treated not by randomness, but by arbitrariness of researchers. For instance, to evaluate whether a new teaching program is effective or not, researchers may consider implementing it in one class as treatment group and taking another class as control group. Unlike randomized experiments, students are not always divided into different classes randomly. In other words, there may be other underlying factors affecting the outcome. For example, students in the treatment group are originally doing better on tests than those in the control group. In such case,

even if students in the treatment group achieve better scores on exams, it may be hard to attribute their better performance to the new program.

Generally speaking, in a quasi-experiment design, if there exists underlying factors affecting the outcome between groups, or *selection bias* exists, the two groups would not be comparable, thus leading to a biased result. There have been numerous methods proposed to overcome selection bias according to the treatment assignment rule or features of the data, such as difference in differences or propensity score matching, just to name a few. If the treatment assignment rule is some thresholds based on a set of observable factors, then **regression discontinuity design (RDD)** may serve as an alternative method to correctly identify the treatment effect.

RDD is first introduced by Thistlethwaite and Campbell (1960) to evaluate a scholarship program. In an RDD, whether the individual is eligible to treatment is (partly) determined by pre-specified thresholds. Individuals have higher tendency to receive treatment if they pass these thresholds². According to the number of thresholds, we can define the dimension of RDD. If the treatment depends on solely one measure, then the RDD is of one dimension. Otherwise it is a multidimensional RDD. In this work, we shall focus on the latter more general case. In fact, as we will explain further, multidimensional RDD resembles much its one dimensional counterpart.

In reality, however, unless the assignment is enforced by law or regulation, there is no guarantee that every subject would follow the assignment rule. Depending on whether treatment assignment rule is perfectly followed, RDD can be categorized in two types. The first is called **sharp RDD**, in which all individuals passing the threshold receive treatment, while those failing to pass do not. The other is **fuzzy RDD**, in which there may be some individuals somehow do not follow the assignment rule. In most empirical studies, we do not observe perfect compliance of the assignment rule, hence we face a fuzzy RDD. Moreover, as sharp RDD can be seen as a special case of fuzzy RDD, we would put our emphasis on fuzzy RDD.

The main idea of RDD is that observations just around the threshold are nearly the same, except their treatment status. In other words, the two groups are comparable, without selection bias. Therefore, with proper continuity assumption, we may estimate the outcome just below and above the threshold, and then attribute the difference between the two estimates to the treatment. In this way, we may analyze the average treatment effect **at the threshold**.

However, as mentioned before, in fuzzy RDD we allow for some samples not complying assignment rule. We may group the individuals according to whether they follow the rule or not. Suppose the individual follows the rule, that is, he receives treatment if he passes the threshold, and does not

²Conceptually, our discussion may also be applied to the opposite scenario. That is, individuals receive treatment if they fail to pass the thresholds. For convenience and coherence, we will assume the former case.

receive treatment if he does not pass the threshold, then he is called a **complier**. In contrast, suppose the individual does not get treated if he passes the threshold and gets treated if not passing, he is named a **defier**. There may be some individuals always receiving treatment no matter he passes the threshold or not, who we shall call an **always-taker**; the last kind of individual, who never receive treatment even if he passes the threshold, is named a **never-taker**. We may summarize the four kind of individuals in table 1.

Table 1: Classification of individuals.

	If pass the threshold	If not pass the threshold
Complier	O	X
Always-taker	O	O
Never-taker	X	X
Defier	X	O

^a O means treated, while X means not treated.

^b Note we assume the treatment assignment rule as giving the treatment when the individual pass the threshold.

As always-taker and never-taker exhibit the same behavior no matter they pass the threshold or not, actually we cannot identify the treatment effect for them. Conceptually, we cannot observe, or even approximate their outcome if they are in the opposite treatment status since there is no such information. On the other hand, empirically it is very rare to have defiers, who is deliberately against the assignment rule. Therefore, we will impose the condition that there are no defiers (**No-defier condition**). What we are trying to estimate, as a result, is the average treatment effect for the compliers. To round up the above discussion, we try to estimate **local average treatment effect for the compliers at the threshold**.

In the following section, we shall formalize RDD and illustrate how to estimate the desired parameter. First we will start by discussing the case where only one standard is present (one dimensional RDD), and then move on to the more complex cases.

3 1-dim Fuzzy RDD

3.1 Problem Formulation

In (fuzzy) RDD, treatment assignment depends on a pre-specified criterion. A sample point has higher probability to get treated if its observable covariate x (running variable, or assignment variable) exceeds a known threshold value. Such threshold may be determined by regulations or a rule of thumb. Without loss of generality, we set this threshold value to 0 in this paper unless otherwise specified. Researchers can also observe interested outcome variable y , and whether or not the sample point

actually receives treatment, documented in a binary variable w . For instance, in Almond, Doyle, Kowalski and Williams (2011), the authors quantitatively estimate the effect of intensive care on very-low-birth weight newborns. In this case, the outcome variable y is one-year mortality of the newborn (or medical expenses), and the running variable x is the weight of the newborn with known threshold of very-low-birth-weight infant, 1500 gram.

In the language of Rubin casual model (Rubin, 1974), we may write the data generating process as follows:

$$y_i = y_i(1)w_i + y_i(0)(1 - w_i) = y_i(0) + w_i(y_i(1) - y_i(0)), \quad (3.1)$$

where $y_i(1)$ and $y_i(0)$ gives the status of outcome with and without treatment, respectively; w_i is the indicator of treatment status. On the other hand, the treatment status w_i is (at least partially) determined by the covariate x , we introduce another variable indicating whether x exceeds the threshold value or not. Namely, $z_i = \mathbb{1}(x_i \geq 0)$.³ If $z_i = 1$, we say that this individual is assigned to the treatment group; otherwise, it is assigned to the control group. Then w_i is determined through the following mechanism:

$$w_i = w_i(1)z_i + w_i(0)(1 - z_i) = w_i(0) + z_i(w_i(1) - w_i(0)), \quad (3.2)$$

where $w_i(1)$ and $w_i(0)$ are binary variables indicating the treatment status when the subject is assigned to the treatment group and the control group, respectively.

As mentioned before, in a fuzzy RDD, we do not require all individuals to follow the treatment assignment rule. There may exist some sample points in treatment group but does not receive treatment; there also may exist other observations in control group which indeed receive treatment. What we really want to find out is the local average treatment effect for those who truly follows the assignment rule, or the *compliers*. According to whether the samples follow the assignment or not, they can be divided into the four groups in table 1. Using the notation in the last paragraph, they can be defined as follows:

Definition 3.1 (Classification of individuals).

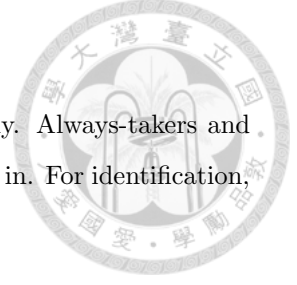
Observations can be categorized into only one of the following four groups depending on whether they follow the assignment rule or not.

1. Complier: $w_i(0) = 0, w_i(1) = 1$
2. Always-taker: $w_i(0) = 1, w_i(1) = 1$
3. Never-taker: $w_i(0) = 0, w_i(1) = 0$

³Note we have assumed that subjects have higher tendency to get treated if they pass the threshold. If one assumes the opposite case, then one should define $z_i = \mathbb{1}(x_i \leq 0)$

4. Defier: $w_i(0) = 1, w_i(1) = 0$

In other words, compliers are those who follow the assignment rule perfectly. Always-takers and never-takers always receive or do not receive treatment whichever group they fall in. For identification, we usually assume no-defier assumption:



Assumption 3.1 (No-defier Assumption). $w_i(\cdot)$ is a non-decreasing function.

In fuzzy RDD, we allow the presence of always-takers and never-takers. What we only need is different treatment probabilities of the two groups, at least for those observations near the threshold, or those whose covariate x satisfying $|x| < \epsilon$, where ϵ is a small positive number. Specifically, we make the following assumption:

Assumption 3.2 (Different treatment probabilities).

$$0 \leq \lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i | z_i = 0, |x_i| = \epsilon) < \lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i | z_i = 1, |x_i| = \epsilon) \leq 1.$$

That is, the probability of receiving treatment just above the threshold is different from that for just below the threshold. The covariate x has partial, but not necessarily full, impact on receiving treatment. It is also worth noting that with Assumption 3.1, Assumption 3.2 is equivalent to the existence of compliers.

In the special case that $\mathbb{E}(w_i | z_i = 1) = 1$ and $\mathbb{E}(w_i | z_i = 0) = 0$, whether or not an individual receives treatment totally depends on which group it lies in. In other words, every observation follows the assignment rule perfectly, or more simply, all observations are compliers. In this case, fuzzy RDD reduces to sharp RDD. In many empirical studies, if a treatment is compulsory by law, there should be nearly no ambiguity of receiving the treatment or not, therefore a sharp RDD can be applied.

3.2 Identification

Intuitively, sample points with running variable just below and just above the threshold should be nearly the same except their treatment status. To make treatment group and control group comparable, no other factors except the treatment should affect the outcome. This condition can be characterized by the following assumption:

Assumption 3.3 (Continuity Assumption).

$$\mathbb{E}(y_i(1)|x_i), \mathbb{E}(y_i(0)|x_i), \mathbb{E}(w_i(1)|x_i) \text{ and } \mathbb{E}(w_i(0)|x_i) \text{ should be continuous at } x_i = 0.$$

In other words, at least around the threshold, by the continuity assumption on y_i , the discontinuity

observed can be well ascribed to the treatment. With this assumption, even if we do not have sample points exactly at the threshold, we can use sample points whose covariate is near the threshold to estimate the outcome at the cutoff. This assumption would hold if the samples do not have perfect control over x , hence the group they fall in. For example, consider physiological measurements like blood pressure or heart rate. Alternatively, test scores, in most cases, cannot be fully controlled by subjects as well. On the other hand, the continuity assumption on w_i ensures that the proportion of compliers, always-takers and never-takers does not vary tremendously at the threshold.

However, in fuzzy RDD, not all subjects are compliers, to correctly identify the treatment effect for the compliers at the cutoff, we have to adjust $\lim_{\epsilon \rightarrow 0} \mathbb{E}(y_i | z_i = 1, |x_i| = \epsilon) - \lim_{\epsilon \rightarrow 0} \mathbb{E}(y_i | z_i = 0, |x_i| = \epsilon)$ by dividing the proportion of compliers, which can be estimated by the difference of proportions of samples treated just below and above the threshold, $\mathbb{E}(w_i | z_i = 1, |x_i| = \epsilon) - \mathbb{E}(w_i | z_i = 0, |x_i| = \epsilon)$, where ϵ is a small amount. We may summarize the above argument in the following theorem in Hahn (2001):

Theorem 3.1 (Identification). The local average treatment effect of the compliers at the threshold (τ_{FRD}) can be identified as follows:

$$\tau_{\text{FRD}} = \frac{\lim_{\epsilon \rightarrow 0} \mathbb{E}(y_i | z_i = 1, |x_i| = \epsilon) - \lim_{\epsilon \rightarrow 0} \mathbb{E}(y_i | z_i = 0, |x_i| = \epsilon)}{\lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i | z_i = 1, |x_i| = \epsilon) - \lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i | z_i = 0, |x_i| = \epsilon)} \quad (3.3)$$

Proof. First observe that $\mathbb{E}(y_i | z_i = 1, |x_i| = \epsilon) = \mathbb{E}(y_i(1)w_i + y_i(0)(1 - w_i) | z_i = 1, |x_i| = \epsilon)$
 $= \mathbb{E}(y_i(1)w_i(1) + y_i(0)(1 - w_i(1)) | z_i = 1, |x_i| = \epsilon)$.

Therefore, by continuity assumption,

$$\lim_{\epsilon \rightarrow 0} \mathbb{E}(y_i | z_i = 1, |x_i| = \epsilon) = \mathbb{E}(y_i(1)w_i(1) + y_i(0)(1 - w_i(1)) | x_i = 0).$$

$$\text{Similarly, } \lim_{\epsilon \rightarrow 0} \mathbb{E}(y_i | z_i = 0, |x_i| = \epsilon) = \mathbb{E}(y_i(1)w_i(0) + y_i(0)(1 - w_i(0)) | x_i = 0).$$

Hence, the numerator of τ_{FRD} can be simplified as:

$$\mathbb{E}((y_i(1) - y_i(0))(w_i(1) - w_i(0)) | x_i = 0) = \mathbb{E}(y_i(1) - y_i(0) | x_i = 0, \text{Complier}) \mathbb{P}(\text{Complier at } x_i = 0).$$

On the other hand, by the no-defier assumption,

$$\lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i | z_i = 1, |x_i| = \epsilon) = \lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i(1) | z_i = 1, |x_i| = \epsilon)$$

$$= \lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i(1) | x_i = 0) = \mathbb{P}(\text{Always Taker at } x_i = 0) + \mathbb{P}(\text{Complier at } x_i = 0).$$

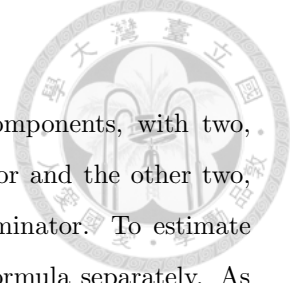
$$\text{Similarly, } \lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i | z_i = 0, |x_i| = \epsilon)$$

$$= \lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i(0) | x_i = 0) = \mathbb{P}(\text{Always Taker at } x_i = 0).$$

This shows that the denominator of τ_{FRD} is $\mathbb{P}(\text{Complier at } x_i = 0)$.

Finally, a slight algebra gives that

$$\tau_{\text{FRD}} = \mathbb{E}(y_i(1) - y_i(0) | x_i = 0, \text{Complier}), \text{ or the local average treatment effect for the compliers at the threshold.} \quad \square$$



3.3 Estimation Strategy

Further inspection into equation (3.2) shows that τ_{FRD} consists of four components, with two, $\lim_{\epsilon \rightarrow 0} \mathbb{E}(y_i | z_i = 1, |x_i| = \epsilon)$ and $\lim_{\epsilon \rightarrow 0} \mathbb{E}(y_i | z_i = 0, |x_i| = \epsilon)$, in the numerator and the other two, $\lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i | z_i = 1, |x_i| = \epsilon)$ and $\lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i | z_i = 0, |x_i| = \epsilon)$, in the denominator. To estimate τ_{FRD} , it is straightforward and tempting to estimate each component in the formula separately. As mentioned in Hahn, Todd, and Van der Klaauw (2001), we can estimate $\lim_{\epsilon \rightarrow 0} \mathbb{E}(y_i | z_i = 1, |x_i| = \epsilon)$ by doing the following (local) linear regression:

$$\min_{\beta_0, \beta_1} \sum_{i: 0 \leq x_i \leq h} (y_i - \beta_0 - \beta_1 x_i)^2 \kappa\left(\frac{x_i}{h}\right),$$

where h is a chosen bandwidth, and then take β_0 to be the estimate. Here $\kappa(\cdot)$ is a kernel function which the econometricians can freely choose. Note that we have to choose a bandwidth h , or keep only the sample points around the cutoff, since we want to estimate the average outcome and the proportion of treatment receiver for those individuals near the threshold. Other components in (3.2) can be estimated by similar methods.

However, as we do a total of four estimations, the estimate of τ_{FRD} would suffer from a huge amount of sampling variation. To overcome such difficulties, first one can observe that since z_i is binary, the denominator and numerator of (3.3) can be estimated by γ_1 and δ_1 in the following two regression models respectively:

$$w_i = \gamma_0 + \gamma_1 z_i + \gamma_2 x_i + \gamma_3 x_i z_i + \xi_i, \quad (3.4)$$

$$y_i = \delta_0 + \delta_1 z_i + \delta_2 x_i + \delta_3 x_i z_i + \nu_i, \quad (3.5)$$

where ξ_i and ν_i are error terms. In fact, in (3.4) and (3.5), we shall only consider sample points with $|x_i| < h$ to more precisely capture the local effect around the threshold. Also note that we allow the average marginal effect of x_i to be different for those subjects above the threshold and below by incorporating $x_i z_i$. Hahn et al. (2001) first notice the numerical equivalence between $\tau_{\text{FRD}} = \delta_1 / \gamma_1$ and the following two stage least square estimation (2SLS), with the first stage being (3.4), and the second stage being:

$$y_i = \beta_0 + \alpha w_i + \beta_1 x_i + \beta_2 x_i z_i + \epsilon_i. \quad (3.6)$$

One can observe that after inserting estimated w_i from (3.4) into (3.6), one have:

$$\begin{aligned} y_i &= \beta_0 + \alpha w_i + \beta_1 x_i + \beta_2 x_i z_i + \epsilon_i \\ &= \beta_0 + \alpha(\gamma_0 + \gamma_1 z_i + \gamma_2 x_i + \gamma_3 x_i z_i) + \beta_1 x_i + \beta_2 x_i z_i + \epsilon_i. \end{aligned}$$

Finally, comparison of coefficients with (3.5) gives $\delta_1 = \alpha\gamma_1$, or $\alpha = \delta_1/\gamma_1 = \tau_{\text{FRD}}$.

To sum up the above discussion, we do the following (local) instrumental regression:

$$\min_{\alpha, \beta_0, \beta_1, \beta_2} \sum_{i: -h \leq x_i \leq h} (y_i - \alpha w_i - \beta_0 - \beta_1 x_i - \beta_2 x_i z_i)^2 \kappa\left(\frac{x_i}{h}\right), \quad (3.7)$$

with z_i being the instrument of w_i . The estimated coefficient of w_i (α) is then equivalent to the 2SLS estimator above, which gives the desired τ_{FRD} .

It is also worth noting that in sharp RDD, (3.7) reduces to an ordinary linear regression since the instrumental variable z_i of w_i is exactly itself in this particular case. One may also explain this result by examining the 2SLS procedure. In the first stage (3.4), as $w_i = z_i$, the best fit $(\gamma_0, \gamma_1, \gamma_2, \gamma_3) = (0, 1, 0, 0)$, and hence the first stage is in fact trivial. Graphically, (3.7) is essentially to fit two different lines for those observations above and below the threshold respectively, which we illustrate by figure 1 below.

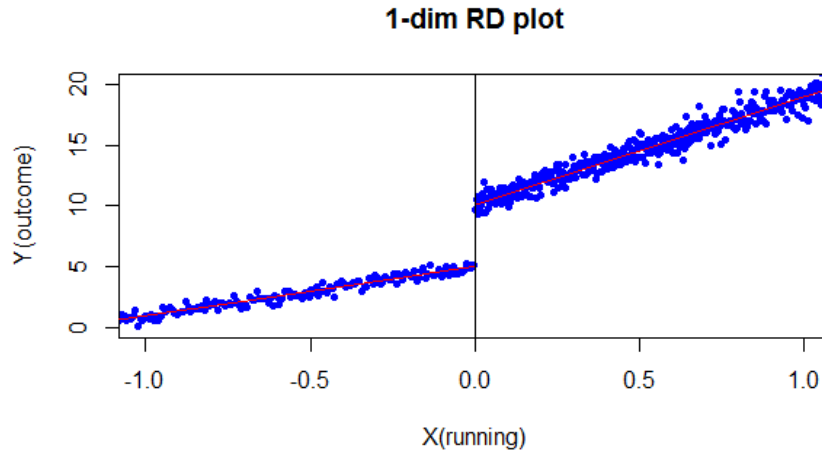


Figure 1: An illustration of one dimensional sharp RDD. We use two different lines to fit data below and above the threshold respectively.

By using 2SLS, which is incorporated in most statistical software, the standard error of the estimate τ_{FRD} can be easily attained. However, one should be aware that the standard error would still be high if the denominator of the estimate, namely $\lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i | z_i = 1, |x_i| = \epsilon) - \lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i | z_i = 0, |x_i| = \epsilon)$, is small. Numerically, this corresponds to the case with weak instruments, or small γ_1 in the first stage (3.4). Intuitively, the small difference in treated proportion in treatment group and control group means that the two groups are only slightly different, making it harder to estimate the treatment effect.

In fact, if one wish, one may also use higher order polynomial to fit the data. However, one should always be aware that higher order polynomial is not a guarantee for better estimate as the estimated coefficient and model would be highly unstable due to overfitting. As Gelman and Imbens (2017) pointed out, the estimate would be easily affected by the degree of the globally-fitted polynomial. In contrast, locally-fitted polynomial provides a relatively more stable estimate. Empirically, one may use the famous Akaike information criterion (AIC) to determine the order of fitted polynomial; namely,

$$AIC = N \ln(SSR/N) + 2p, \quad (3.8)$$

where N is the number of sample points used, SSR stands for residual sum of squares, and p is the number of estimated parameters in the model. Particularly, in one-dimensional fuzzy RDD, $p = 2(d + 1)$, where d is the degree of the fitting polynomial.

Since our main goal is to estimate the local treatment effect at the threshold, observations far away from the threshold may mess our estimation up. Therefore, it is better to choose a bandwidth h around the threshold and keep only the sample points in this band. As for existing bandwidth selection procedure, Imbens and Lemieux (2008) considers cross-validation. On the other hand, Imbens and Kalyanaraman (2012) proposes another method to directly estimate the optimal bandwidth. Basically, they tried to choose a bandwidth which minimizes the asymptotic mean squared error of the estimates. Calonico, Cattaneo and Titiunik (2014, CCT) argue that most methods for choosing bandwidth would actually produce a biased estimate and proposed a robust method to correct such bias. In this work, we mainly follow CCT to choose bandwidth.

4 Multidimensional Fuzzy RDD

In the former framework, we only allow a single running variable x . That is, whether or not an individual receives treatment depends solely on a single factor. In reality, however, there may be cases involving more factors. For example, a patient is diagnosed with hypertension if his systolic pressure **or** diastolic pressure exceed 140 and 90 mm-Hg, respectively. In Jacob and Lefgren (2004), the authors investigate a policy implemented in Chicago starting from 1996, in which students are required to attend summer school if their math **or** reading test score are below a certain cutoff.

Although the methods introduced later in this paper can be easily generalized to an arbitrary number of factors, for simplicity, we shall discuss the case where only two factors are present.



4.1 Problem Formulation and assumptions

Recall DGP (3.1) and (3.2). That is,

$$y_i = y_i(1)w_i + y_i(0)(1 - w_i) = y_i(0) + w_i(y_i(1) - y_i(0))$$

$$w_i = w_i(1)z_i + w_i(0)(1 - z_i) = w_i(0) + z_i(w_i(1) - w_i(0))$$

This time, however, treatment status w_i is related to two (or more) factors, documented in the vector of running variables X . Individuals may have a higher probability receiving treatment if X_1 and X_2 both exceed certain cutoff (and-rule) or either one of them pass some threshold (or-rule). More specifically, we set both the threshold for X_1 and X_2 to 0 in this work unless otherwise specified; then for an and-rule, we define z_i , the binary variable indicating whether the subject falls in the treatment group or not, as $z_i = \mathbf{1}(X_{1i} > 0)\mathbf{1}(X_{2i} > 0) = \min\{\mathbf{1}(X_{1i} > 0), \mathbf{1}(X_{2i} > 0)\}$. For an or-rule, we define $z_i = \max\{\mathbf{1}(X_{1i} > 0), \mathbf{1}(X_{2i} > 0)\}$. When $z_i = 1$, we say the sample lies in the treatment area; otherwise it lies in the control area.

In fact, both kinds of rule do not make much difference when dimension of the covariate vector X equals to two as the negation of an and-rule leads to an or-rule. In other words, if (z_i, X_{1i}, X_{2i}) follows an or-rule, then $(1 - z_i, -X_{1i}, -X_{2i})$ follows an and-rule in the sense that $1 - z_i = 1 - \max\{\mathbf{1}(X_{1i} > 0), \mathbf{1}(X_{2i} > 0)\} = \min\{\mathbf{1}(X_{1i} \leq 0), \mathbf{1}(X_{2i} \leq 0)\} = \min\{\mathbf{1}(-X_{1i} \geq 0), \mathbf{1}(-X_{2i} \geq 0)\}$. Therefore, in this work, unless otherwise specified, we assume that **the assignment follows an and-rule**. That is, the observation has a higher tendency to be treated if both X_1 and X_2 are greater than 0, or X lies in the first quadrant of the coordinate plane with X_1 and X_2 being the two axes.

For convenience, we say the observation **lies in the first quadrant** (of the plane spanned by X_{1i} and X_{2i}) if the covariate vector X of observation i satisfies $X_{1i} > 0$ and $X_{2i} > 0$. In similar ways, we can define what observations lie in the second, the third, or the fourth quadrants mean. One can observe that if we adopt an and-rule, then the treatment group is composed of subjects in the first quadrant, while the control group comprises subjects from all other quadrants⁴.

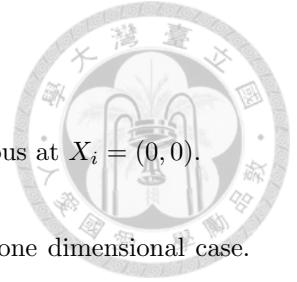
To estimate the local treatment effect for the compliers at the threshold, or at $(0,0)$, we again need the following assumptions:

Assumption 4.1 (No-defier Assumption). $w_i(\cdot)$ is a non-decreasing function.

Assumption 4.2 (Different treatment probabilities).

$$0 \leq \lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i | z_i = 0, |x_i| = \epsilon) < \lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i | z_i = 1, |x_i| = \epsilon) \leq 1.$$

⁴If we consider or-rule, then the control group is composed of subjects in the third quadrant, while subjects from all the other quadrants lie in the treatment group.



Assumption 4.3 (Continuity Assumption).

$\mathbb{E}(y_i(1)|X_i)$, $\mathbb{E}(y_i(0)|X_i)$, $\mathbb{E}(w_i(1)|X_i)$ and $\mathbb{E}(w_i(0)|X_i)$ should be continuous at $X_i = (0, 0)$.

As one can see, these assumptions are just analogues of their counterparts in one dimensional case. In fact, the identification formula of the treatment effect stays the same in high-dimensional case, as summarized in the following theorem:

Theorem 4.1 (Identification). The local average treatment effect of the compliers at the threshold can be identified as follows:

$$\tau_{\text{FRD}} = \frac{\lim_{\epsilon \rightarrow 0} \mathbb{E}(y_i|z_i = 1, |X_i| = \epsilon) - \lim_{\epsilon \rightarrow 0} \mathbb{E}(y_i|z_i = 0, |X_i| = \epsilon)}{\lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i|z_i = 1, |X_i| = \epsilon) - \lim_{\epsilon \rightarrow 0} \mathbb{E}(w_i|z_i = 0, |X_i| = \epsilon)}.$$

Recall that in the proof of one-dimensional case, we actually do not use any properties that only holds in one dimension. Thus the proof for this theorem is exactly the same as Theorem 3.1.

In one dimensional RDD, it is natural to separate the observations into two groups according to whether their running variables pass the threshold or not. In two dimensional case, following and-rule, we label those sample points in the second, the third, the fourth quadrant as control group. Implicitly, we have assumed samples in the control group have the same probability of receiving treatment, as well as the same marginal effect of x_1 and x_2 in the three quadrants. Nevertheless, empirically this may not be the case. For example, although only those whose systolic and diastolic blood pressure are over 140/90 mm-Hg are diagnosed with hypertension, individuals with only one passed standard may have a higher tendency receiving treatment than those who are healthy. The latter two kinds of individuals are both categorized into control group by definition, but obviously there are heterogeneity between them. As we shall see in the simulation, naïvely neglecting the heterogeneity in the control group would often lead to a biased estimate.

4.2 Estimation Method

There are numerous methods to estimate treatment effect at the threshold according to previous studies. They can be briefly summarized into the following two categories.

(A) Dimension Reduction

To estimate τ_{FRD} , one may consider compressing the information from multidimensional vector X into one dimension by taking a norm of it. For example, Reardon and Robinson (2012) consider ℓ_2

norm and introduce the following variable:

$$d_i = z_i \sqrt{X_{1i}^2 + X_{2i}^2} - (1 - z_i) \sqrt{X_{1i}^2 + X_{2i}^2} = (2z_i - 1) \sqrt{X_{1i}^2 + X_{2i}^2}.$$

In other words, if the sample falls into the treatment area, we attach a positive distance; otherwise, a negative distance is used. In this way, we may estimate a one-dimensional RDD model with outcome still being y_i , but running variable d_i with threshold 0. Generally, one may consider other norms such as ℓ_1 norm or maximum norm (ℓ_∞ norm). In such cases, the running variable $d_i = (2z_i - 1) \|X_i\|$ should be introduced, where $\|\cdot\|$ is the norm one wishes to use. This method can be similarly generalized to higher dimensional case.

On the other hand, Wong, Steiner and Cook (2013) consider taking $d_i = \min\{X_{1i}, X_{2i}\}$, which they named “centering approach”. In the same paper, Wong, Steiner and Cook also introduce “univariate method”, in which one simply focus on a single running variable, neglecting the other. Specifically, one discard observations with $X_{1i} < 0$, or observations in the second and the thrid quadrant. The remaining observations then are decided to be treated or not solely according to X_2 . Therefore, one dimensional RDD estimation strategies addressed in the previous section can be adopted with running variable being X_2 . Similarly, one can consider only the observations in the first and the second quadrant; in other words, one neglects those with $X_{2i} < 0$.

Nevertheless, we have to stress that by using dimension reduction method, we may falsely simplify the relation between X_{1i} and X_{2i} . For instance, if one wish to use ℓ_2 norm to compress the running vector X_i , then one basically has to (implicitly) assume equal treatment effect, equal marginal effect of (X_{1i}, X_{2i}) , as well as equal treatment probability for sample points with the same $\|X_i\|_2$ in treatment group and control group respectively. Otherwise, misspecification would lead to a biased estimate.

(B) Local polynomial fitting

Instead of compressing existing information, Imbens and Zajonc (2011) consider directly fitting polynomials near the threshold. Specifically, they estimate the following regression models.

$$\min_{\alpha, \beta} \sum_{i: X_i \in H} (y_i - \alpha w_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \beta_3 z_i X_{1i} - \beta_4 z_i X_{2i})^2 \kappa\left(\frac{X_{1i}}{h_1}\right) \kappa\left(\frac{X_{2i}}{h_2}\right), \quad (4.1)$$

with z_i being the instrument of w_i , $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ for brevity, $\kappa(\cdot)$ a kernel function free to choose, and H a chosen bandwidth around the threshold according to the data (we will discuss the selection procedure later). The estimated coefficient $\hat{\alpha}$ is then the desired τ_{FRD} . Graphically, similar to one dimensional case, they fit two different planes to treatment group and control group respectively.

As this method uses the whole sample in estimation, we shall refer to it as **union method** (Lo, 2017) in the rest of this paper. However, as suggested before, since union method implicitly assumes homogeneous effect of running variables among the control group by locally fitting a single polynomial. If heterogeneity among the control group exists, then as shown in Lo (2017), Hsu et al. (2018), and our simulation later, union method would produce a biased estimate.

To correctly estimate τ_{FRD} when heterogeneity is present, Lo and Hsu, Kuan and Lo note that since the bias of union method originates from neglecting the difference of sample points in the second, third, and fourth quadrants, one can simply tackle two quadrants, instead of four, at a time. Specifically, following their idea of dealing with sharp RDD, we may generalize their method to fuzzy design by implementing the following three instrumental least square regression:

$$\min_{\alpha_1, \beta} \sum_{i: X_i \in H, X_{2i} > 0} (y_i - \alpha_1 w_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \beta_3 z_i X_{1i} - \beta_4 z_i X_{2i})^2 \kappa\left(\frac{X_{1i}}{h_1}\right) \kappa\left(\frac{X_{2i}}{h_2}\right), \quad (4.2)$$

$$\min_{\alpha_2, \beta} \sum_{i: X_i \in H, X_{1i} X_{2i} > 0} (y_i - \alpha_2 w_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \beta_3 z_i X_{1i} - \beta_4 z_i X_{2i})^2 \kappa\left(\frac{X_{1i}}{h_1}\right) \kappa\left(\frac{X_{2i}}{h_2}\right), \quad (4.3)$$

$$\min_{\alpha_3, \beta} \sum_{i: X_i \in H, X_{1i} > 0} (y_i - \alpha_3 w_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \beta_3 z_i X_{1i} - \beta_4 z_i X_{2i})^2 \kappa\left(\frac{X_{1i}}{h_1}\right) \kappa\left(\frac{X_{2i}}{h_2}\right), \quad (4.4)$$

with z_i being the instrument of w_i in all three regressions. Finally, we take $\widehat{\tau_{\text{FRD}}} = (\widehat{\alpha}_1 + \widehat{\alpha}_2 + \widehat{\alpha}_3)/3$. One can observe that in (4.2), (4.3), (4.4), we regress with sample points in the first and the second, the first and the third, and finally the first and the fourth quadrants, respectively. Each of the regression gives an unbiased estimation of τ_{FRD} (namely $\widehat{\alpha}_i$), and hence the unbiasedness of $\widehat{\tau_{\text{FRD}}}$.

Lo (2017) and Hsu et al. (2018) name the above estimation procedure **average method**. In this procedure, one uses all observations near the threshold to estimate τ_{FRD} . However, as three instrumental regressions are required, there is considerable computational burden. As a modification, Lo again addresses that as our main goal is to estimate the average treatment effect at the threshold $(0, 0)$, one may consider only the observations in the first and the third quadrant since the two sets intersect at $(0, 0)$. Following his advice, we can do only regression (4.3), and take $\widehat{\tau_{\text{FRD}}} = \widehat{\alpha}_2$. This simplified procedure is named **intersection method** in Lo's work. As illustrated in figure 2, the only difference between union and intersection method is the scope where sample points are used in estimation.

It has been shown in Lo (2017) and Hsu et al. (2018) that for sharp RDD, intersection method and average method perform better over union method (and other dimension reduction methods) if heterogeneity among control group exists, in the sense that the former gives a unbiased estimate, while the latter does not. As for the comparison of the former two methods, although intersection method

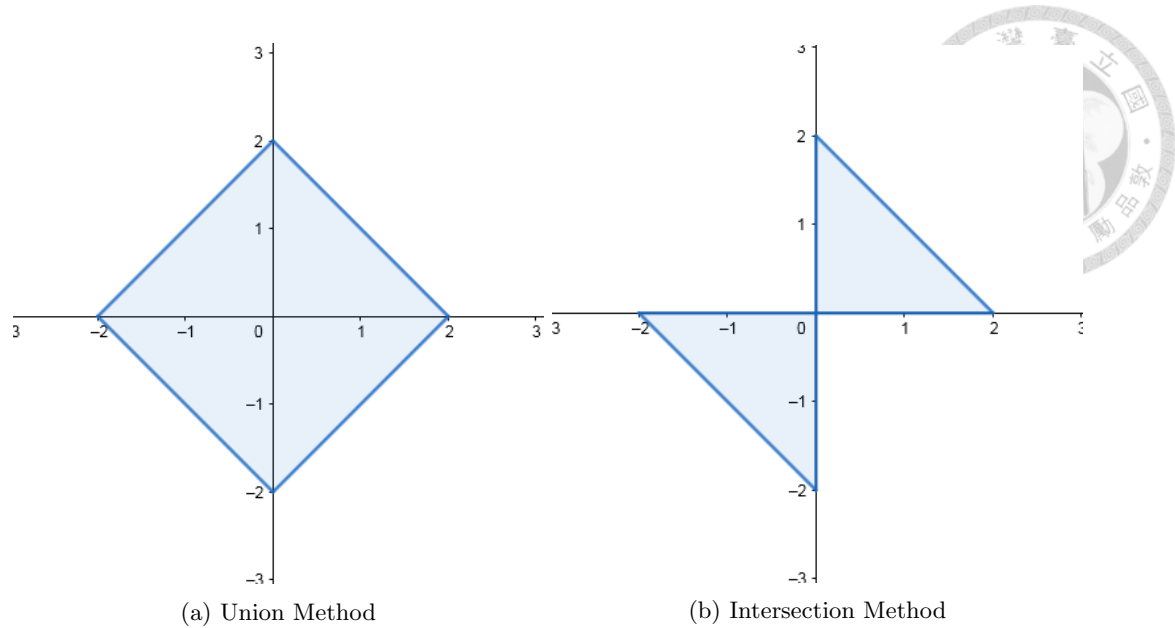
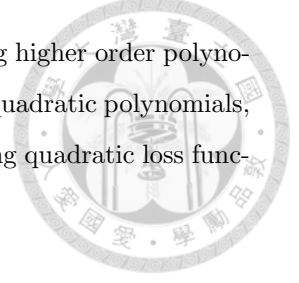


Figure 2: Difference in sample points used in both methods. Only the observations in the shaded area are taken into the regression. Note that here we adopt an ℓ_1 bandwidth.

makes less computation burben, it neglects the observations from the second and the fourth quadrants. This information loss leads to a higher standard error and mean square error (MSE) of the estimate compared to that acquired by average method.

Nevertheless, Lo (2017) and Hsu et al. (2018) only consider sharp design with local linear fitting. As we will further elaborate in the simulation section, for fuzzy RDD, there are actually two sources of bias of union method. The first one, also appearing in sharp design, is **heterogeneity of effects of running variables** among the control group. To be specific, the (average) marginal effect of X_{1i} and X_{2i} for sample points in different quadrants in the control group may not be equal. Union method fits only one polynomial to the whole control group, lacking the flexibility to tackle heterogeneity, hence its poor performance. On the other hand, intersection and average method fits each quadrant separately, allowing a wider class of setting. In particular, the case when heterogeneity of marginal effects is absent is also tractable by the latter two methods.

The other one, which only can be detected in fuzzy design, is **the difference of treatment probability** in the control group. Even if X_{1i} and X_{2i} have the same marginal effect on observations in differnt quadrants in the control group, different probabilities of exposure to treatment still contribute to heterogeneity. If the difference of probability of getting treated among quadrants in the control group is large, then the bias originated from misspecification would be amplified. It is also noteworthy that such difference would disappear in sharp design simply because all subjects in the control group are prohibited from being treated by definition.



On the other hand, just as one dimensional case, one may also consider fitting higher order polynomial locally instead of fitting linear ones. More specifically, say, if we want to fit quadratic polynomials, then we simply replace the linear loss functions in (4.1) to (4.4) with the following quadratic loss function:

$$(y_i - \alpha w_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \beta_3 z_i X_{1i} - \beta_4 z_i X_{2i} - \beta_5 X_{1i}^2 - \beta_6 X_{1i} X_{2i} - \beta_7 X_{2i}^2 - \beta_8 z_i X_{1i}^2 - \beta_9 z_i X_{1i} X_{2i} - \beta_{10} z_i X_{2i}^2)^2 \kappa\left(\frac{X_{1i}}{h_1}\right) \kappa\left(\frac{X_{2i}}{h_2}\right);$$

in other words, we include quadratic terms in regressions. As one may have observed, mere proceeding from linear fitting to quadratic fitting dramatically raises the number of coefficients to be estimated. When there is only one running variable, raising the order of fitted polynomial simply means including higher order term of that running variable. For higher dimensional case, however, not only do we have to include higher order terms of the running variables, but also have to consider the interaction terms, which leads to a much more complex fitting procedure.

As for determining the order of fitting polynomial, aforementioned criterions such as AIC, defined in (3.7), can also be applied here. That is,

$$AIC = N \ln(SSR/N) + 2p. \quad (3.7)$$

Specifically, the number of parameters in the model $p = (d + 1)(d + 2)$ for the case where two running variables are present and the degree of fitted polynomial equals d .⁵ Eventually, the fitted model with the least AIC would be preferred.

In the simulation section, we shall go beyond Lo (2017) and Hsu et al. (2018), exploring whether the predominance of average method and intersection method persists in fuzzy RDD. On the other hand, we will also extend the estimation procedure by fitting quadratic polynomials instead of the linear ones. Before we move on to simulation, however, we shall conclude this section by briefly discussing the bandwidth selection procedures. Unfortunately, there is little research in choosing bandwidth for RDD in multidimensional case. Therefore we have to resort to dimension reduction methods. We call the adopted method **separation method**. We simply perform two one-dimensional bandwidth selection procedures, one for X_{1i} , the other for X_{2i} . In each process, we choose bandwidth with respect to one running variable, pretending that the other one does not exist. After getting h_1, h_2 for X_{1i} and X_{2i} , we keep only the sample points with $|X_{1i}| < h_1$ and $|X_{2i}| < h_2$. This process gives a rectangular bandwidth.

In fact, originally we have tried another method named **norm method**. Similar to dimension

⁵More generally, if there are n running variables and the degree of fitted polynomial is d , then $p = 2C_n^{d+n}$

reduction method for RDD, we first standardize each argument in the covariate vector by dividing every component by its standard error. In other words, we introduce the standardized covariate vector $Xs_i = (X_{1i}/s_1, X_{2i}/s_2)$. By doing this, the bandwidth selection procedure would not be disturbed by the scale of X_{1i} and X_{2i} . Next we compress Xs_i by its norm, attaching a positive value if Xs_i falls into the treatment area, negative value otherwise. More specifically, we introduce $ds_i = (2z_i - 1)\|Xs_i\|$. Note that since standard error is positive, X_i falls into the treatment area if and only if Xs_i does. Finally, we choose the bandwidth h using one dimensional approach (in our work, CCT), and take only the observations such that $ds_i \leq h$ into regressions. However, we have tried ℓ_1 , ℓ_2 , and ℓ_∞ norm, and none of them gives credible estimate. We therefore mainly choose bandwidth according to separation method in the previous paragraph.

5 Simulation

5.1 Setup

To allow for different treatment probability as well as heterogeneous effect of X_i in the control groups, we consider the following four DGPs:

$$\text{DGP 1: } y_i = 5 + X_{1i} + X_{2i} + v_{1i}(5 + X_{1i} + 0.5X_{2i}) + v_{2i}(5 + 2X_{1i} + 2X_{2i}) + v_{3i}(5 + 3X_{1i} + 4X_{2i}) + \epsilon_i$$

$$\text{DGP 2: } y_i = 5 + 5w_i + X_{1i} + w_iX_{1i} + X_{2i} + 0.5w_iX_{2i} + \epsilon_i$$

$$\text{DGP 3: } y_i = 5 + 5w_i + X_{1i} + w_iX_{1i} + 0.3w_{1i}X_{1i} + X_{2i} + 0.5w_iX_{2i} + 0.3w_{2i}X_{2i} + \epsilon_i.$$

$$\begin{aligned} \text{DGP 4: } y_i = & 5 + 5w_i + X_{1i} + w_iX_{1i} + 0.3w_{1i}X_{1i} + X_{2i} + 0.5w_iX_{2i} + 0.3w_{2i}X_{2i} \\ & + X_{1i}^2 + 3X_{2i}^2 + w_iX_{1i}^2 + w_iX_{2i}^2 + \epsilon_i. \end{aligned}$$

We generate 500 samples with size 5000 for each DGP.

$X_i = (X_{1i}, X_{2i})$ are i.i.d. multivariate normal vectors following the distribution:

$$(X_{1i}, X_{2i}) \stackrel{iid}{\sim} N \left(\begin{bmatrix} \varphi(0.4) \cdot s_1 \\ \varphi(0.4) \cdot s_2 \end{bmatrix}, \begin{bmatrix} s_1^2 & 0 \\ 0 & s_2^2 \end{bmatrix} \right),$$

where φ is the quantile function of standard normal variables. Note that instead of setting the mean of X_{1i}, X_{2i} to be 0 (symmetric around the threshold), we adopt an asymmetric design, with respectively 40% of observations passing the threshold for each running variable. We do this because typically, subjects are not symmetrically distributed around the threshold; most of the time, observations in need of treatment is minority (compared to the whole population). Moreover, we follow Lo (2017)

and consider (s_1, s_2) being $(1, 1), (3, 3), (10, 10), (3, 1), (10, 1), (10, 3)$. On the other hand, ϵ_i are i.i.d. standard normal random variables.

In DGP 1, we try to explore whether different marginal effect of X_{1i} and X_{2i} in the control group really makes a difference. The three variables v_{1i}, v_{2i}, v_{3i} are defined as follows:

$$v_{1i} = 1(u_i \leq \Phi(a)) \cdot 1(X_{1i} < 0 \text{ and } X_{2i} < 0)$$

$$v_{2i} = 1(u_i \leq \Phi(a)) \cdot 1(X_{1i}X_{2i} < 0)$$

$$v_{3i} = 1(u_i \leq \Phi(b)) \cdot 1(X_{1i} \geq 0 \text{ and } X_{2i} \geq 0),$$

where $u_i \stackrel{iid}{\sim} N(0, 1)$ and $\Phi(\cdot)$ is the cumulative distribution function of standard normal variable. By v_{1i}, v_{2i}, v_{3i} , we set the treatment probability in the control group and the treatment group to be a and b respectively; however, the treatment effect in the third quadrant is $5 + X_{1i} + 0.5X_{2i}$, while the effect is $5 + 2X_{1i} + 2X_{2i}$ in the second and the fourth quadrant. In the simulation, we have considered $(a, b) = (0, 0.7), (0.15, 0.85), (0.3, 1), (0, 0.9), (0.05, 0.95), (0.1, 1)$. Moreover, in this DGP, we define the binary variable w_i indicating whether the subject receives treatment or not to be $w_i = v_{1i} + v_{2i} + v_{3i}$.

By DGP 2, we inspect the impact of different treatment probability in the control group. The treatment effect is the same for all subjects, which is $5 + X_{1i} + 0.5X_{2i}$. However, w_i is defined as follows:

$$w_i = 1(u_i \leq \Phi(a)) \cdot 1(X_{1i} < 0 \text{ and } X_{2i} < 0) + 1(u_i \leq \Phi(b)) \cdot 1(X_{1i} \geq 0 \text{ and } X_{2i} < 0) \\ + 1(u_i \leq \Phi(b)) \cdot 1(X_{1i} < 0 \text{ and } X_{2i} \geq 0) + 1(u_i \leq \Phi(c)) \cdot 1(X_{1i} \geq 0 \text{ and } X_{2i} \geq 0),$$

where again $u_i \stackrel{iid}{\sim} N(0, 1)$. In this way, the treatment probability in the third quadrant, a , would be different from that in the second and the fourth quadrant, b , thus introducing heterogeneity in the control group.

For DGP 3 and DGP 4, we examine the cases where both forces of heterogeneity in the control group are present. The three variables w_i, w_{1i}, w_{2i} are of key importance in inducing different treatment effect and different treatment probability among the control group. They are defined as follows:

$$w_i = 1\{u_i \leq \Phi(a)\} \cdot 1\{X_{1i} < 0 \text{ or } X_{2i} < 0\} + 1\{u_i \leq \Phi(b)\} \cdot 1\{X_{1i} \geq 0 \text{ and } X_{2i} \geq 0\} \\ w_{1i} = 1\{u_i \leq \Phi(a)\} \cdot 1\{X_{1i} < 0\} + 1\{u_i \leq \Phi(b)\} \cdot 1\{X_{1i} \geq 0\} \\ w_{2i} = 1\{u_i \leq \Phi(a)\} \cdot 1\{X_{2i} < 0\} + 1\{u_i \leq \Phi(b)\} \cdot 1\{X_{2i} \geq 0\},$$

where $u_i \stackrel{iid}{\sim} N(0, 1)$ and $\Phi(\cdot)$ is the cumulative distribution function of standard normal variable. By



w_i we have created different treatment probability between observations in the first quadrant (namely a) and the others (b). In the simulation, we consider (a, b) to be $(0.15, 0.85)$ or $(0.05, 0.95)$ to investigate whether the scale of the difference in treatment probability plays a role in estimation. The assignment rule is further complexified by w_{1i} and w_{2i} , which introduce difference around the y-axis and x-axis, respectively. In short, the distribution of (w_i, w_{1i}, w_{2i}) in each quadrant can be summarized in figure 3, which leads to the treatment effect assignment schedule of DGP 3 in figure 4⁶. Note that the only difference of DGP 3 and DGP 4 is the presence of quadratic terms. For both DGPs, the desired local treatment effect at the threshold $(0, 0)$ equals 5 (the coefficient of w_i).

For all the four DGPs, the desired local treatment effect at the threshold $(0, 0)$ equals 5. For the first two DGPs, we have used linear polynomials to do fitting, while for the latter two DGPs, we have considered locally fitting linear and quadratic polynomials, respectively. We also do a quadratic fitting to DGP 3 to observe the effect of redundant regressors. For bandwidth selection we follow CCT. As aforementioned, since CCT is originally designed for one-dimensional RDD, we have used separation method to adopt it in two-dimensional case. After the bandwidth is chosen, we conduct an instrumental regression with triangular kernel, which is $\kappa(x) = (1 - |x|)\mathbb{1}(|x| \leq 1)$.

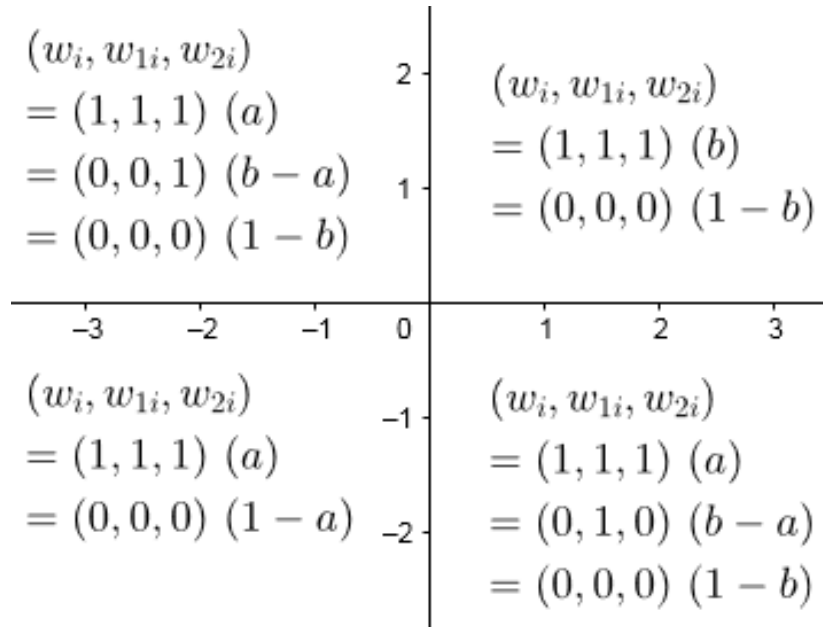


Figure 3: Distribution of (W_i, W_{1i}, W_{2i}) in the four quadrants for DGP 3 and DGP 4. The proportion in the parenthesis is the probability for each result.

⁶For brevity, corresponding figure for DGP 4 is postponed to the appendix.

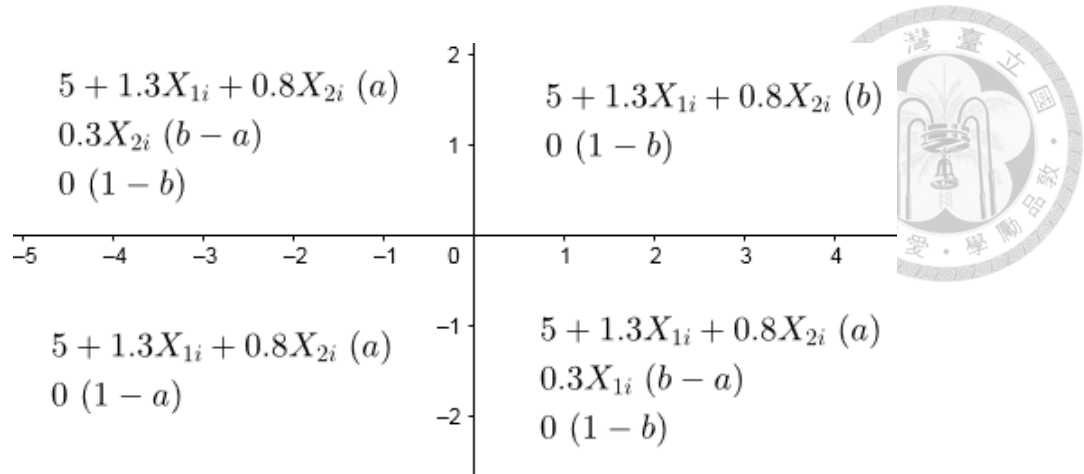


Figure 4: Distribution of treatment effect assignment in the four quadrants for DGP 3. The proportion in the parenthesis is the probability for each result. Note there is an underlying trend of $5 + X_{1i} + X_{2i}$ for all the observations.

5.2 Results

For each choice of (s_1, s_2) and (a, b, c) , we report the mean, the standard error and the mean square error (MSE) of estimate of union, intersection, and average method⁷. Moreover, we report the number of observations left in the bandwidth (out of total size 5000).

As the main result, we observe that intersection and average method provide a fairly unbiased estimate, whereas union method is less accurate. The bias is further exaggerated as (s_1, s_2) become larger. On the other hand, the standard deviation of estimates by intersection method and average method estimation is slightly bigger than that of union method. This happens due to less remained sample points in intersection method and the more complex estimation process for average method.

An inspection into the result of DGP 1 shows that different marginal effects of X_{1i} and X_{2i} do contribute to the bias in union method. The parameter controlling the scale of heterogeneity in DGP 1 is a . When a is small, even if the treatment effect is largely different for subjects lying in different quadrants, the effect of heterogeneity remains insignificant since not many observations in the control group actually receive treatment. As one may observe from table 2 to table 7, union method produces the largest bias when $(a, b) = (0.3, 1)$ (Table 4). On the other hand, in the extreme case $a = 0$ (Table 2 and 5), actually there is no heterogeneity in the control group since nobody gets treated. In this case, union method still produces a favorable result.

On the other hand, we observe that the overall standard deviation gets larger as $b - a$ shrinks. This happens since $b - a$, the difference in treatment probability, is the denominator of τ_{FRD} . One may imagine that as $b - a$ approaches 0, the estimate would be highly unstable. Numerically, this

⁷Other dimension reduction methods have been shown to produce a biased result even for sharp RDD in Lo; we therefore focus on the comparison of these three methods.

corresponds to the case where the instrumental variable is weak. From a different perspective, $b - a$ is the proportion of compliers at $(0, 0)$. Small $b - a$ means there are less compliers, hence the difficulty of estimating the local treatment effect for them.

DGP 2 shows that the impact of different treatment probability. We observe that even if the treatment effect is the same among the control group, difference in treatment probability does affect the accuracy of union method. Here the parameter controlling the heterogeneity is $b - a$. Holding a , $c - b$ fixed, one can see that the bias of union method rises as $b - a$ gets larger as in Table 8 to 10; on the contrary, intersection and average method produces a much more accurate estimate.

Interestingly, when $(a, b, c) = (0.15, 0.75, 0.85)$, not only union method, but also average method performs poorly (Table 11). This happens since $c - b$ is small. Recall that in average method, we do three local IV regressions respectively. However, as the difference between the first and the second quadrant, as well as the difference between the first and the fourth quadrant is not prominent, the estimator coming from regressions concerning those quadrants (namely $\hat{\alpha}_1$ and $\hat{\alpha}_3$ in (4.2) and (4.4)) would be highly unstable, thus the poor performance of average method. In contrast, since the difference of treatment probability between observations in the first and the third quadrant, $c - a$, is large enough, intersection method still produces a favorable result in this case.

Out of 5000 sample points, around 20% to 42% remain in the bandwidth. The proportion is highly related to the order of fitted polynomial. For instance, when we try to fit data generated from DGP 3 with quadratic polynomials, there are approximately 1.6 times observations left compared to when we fit with linear polynomials. The number of left observations is relatively stable. On the other hand, we have also documented the mean and standard error of the standardized chosen bandwidth (that is, the chosen bandwidth h_j divided by the standard deviation of the corresponding running variable X_{ji} , or X_{ji}/s_j). The result depends basically on the order of polynomial used to fit. If one uses higher order polynomial, then the standardized chosen bandwidth would be larger. It is also worth noting that if $s_1 > s_2$, then the standardized chosen bandwidth for X_{1i} (h_1/s_1) is smaller than h_2/s_2 on average. For brevity, we report the result of chosen bandwidths for DGP 3, $(a, b) = (0.15, 0.85)$, with linear fitting and quadratic fitting respectively.

DGP 3 and DGP 4 gives the result when both forces of heterogeneity are present. The overall trend for DGP 3 and DGP 4 remains the same when $b - a$ rises from 0.7 to 0.9. Moreover, the bias of union method becomes even larger as $b - a$ grows up. This is due to the fact that $b - a$ characterize the difference in treatment probability in the control group. If $b - a$ gets larger, then the difference in treatment effect and probability inflates, amplifying the heterogeneity in the control group. This again establishes the fact that intersection and average method can tackle the heterogeneity in the control group, hence its much better performance.

For DGP 3, we have tried both linear and quadratic fitting. One can observe that bias of union method slightly reduces (but the estimate itself is still biased) when one uses quadratic fitting. On the other hand, the estimate of intersection and average method remains more accurate, but has a higher standard deviation, which is a result of redundant explanatory variables, namely the quadratic terms.

Lastly, for the comparison of average method and intersection method, one can observe that in most cases, the standard deviation of estimates from intersection method is larger than that from average method. This may be a consequence of neglecting sample points in the second and the fourth quadrants in intersection method.

6 Conclusion

In this thesis, we review the basic concept and assumptions needed for multidimensional fuzzy RDD. Moreover, we show the identification formula is still valid when multiple running variables and thresholds are present. Most importantly, we generalize the idea mentioned in Lo (2017) and Hsu et al. (2018), proposing intersection method and average method for multidimensional fuzzy RDD, both of which gives unbiased estimate even if heterogeneity exists in control group. An implication of this is one should treat observations and data more cautiously and stay aware of potential heterogeneous structure in the dataset. With more flexibility in estimation procedure, average method and intersection method are still robust under a wide variety of scenarios compared with traditional methods. Moreover, we observe that one need not stick to local linear fitting when tackling RDD problem. If one wish, one may utilize local polynomial fitting to allow for more flexibility.

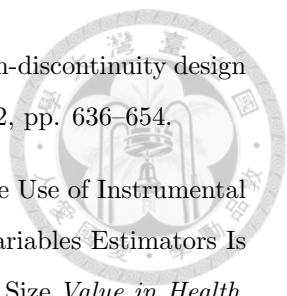
On the other hand, we note that heterogeneity among control group or treatment group in fuzzy RDD generally comes from two sources. The first is different marginal effect of running variables in different quadrants, while the other is different treatment probability. Both sources are important in empirical studies. For example, in medical studies concerning two or more indicators for one single disease, the willingness or urgency to receive treatment may be different for those with only one passed standard or multiple passed standards, leading to different treatment probability.

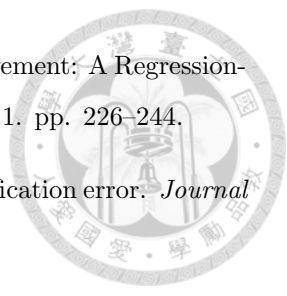
Considering the facility to do regressions nowadays, we highly recommend future researchers to adopt average method despite the slightly additional computational burden it induces. As the newly modified methods have a much more cautious and clever use of information at hand, we would be willing to see RDD utilized in a wide variety of fields, with the ability to deal with problems under much more general settings than ever before.

Reference

1. Almond, D., Doyle, J. J., Jr., Kowalski, A. E., & Williams, H. (2011) Estimating Marginal Returns to Medical Care: Evidence from At-risk Newborns. *The Quarterly Journal of Economics*, Vol 125, Issue 2, pp. 591–634.
2. Arai, Y. & Ichimura, H. (2015). Optimal Bandwidth Selection for the Fuzzy Regression Discontinuity Estimator. *Economics Letters*, Vol 141, pp. 103–106.
3. Banks, J., & Mazzonna, F. (2012). The Effect of Education on Old Age Cognitive Abilities: Evidence from a Regression Discontinuity Design. *The Economic Journal*, Vol 122 (May), pp. 418–448.
4. Bloom, H. S. (2012) Modern Regression Discontinuity Analysis, *Journal of Research on Educational Effectiveness*, Vol 5, Issue.1, pp. 43–82.
5. Bor, J., Moscoe, E., Mutevedzi, P., Newell, M.-L., & Bärnighausen, T. (2014). Regression Discontinuity Designs in Epidemiology: Causal Inference without Randomized Trials. *Epidemiology*, Vol 25, Issue 5, pp. 729–737.
6. Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica*, Vol 82, No.6, pp. 2295–2326.
7. Cattaneo, M. D., Frandsen, B. R., & Titiunik, R. (2015). Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate. *Journal of Causal Inference*, Vol 3, Issue 1, pp. 1–24.
8. Cattaneo, M. D., Keele, L., Titiunik, R., & Vazquez-Bare, G. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, Vol 78, No.4, pp. 1229–1248.
9. Caughey, D., & Sekhon, J. S. (2011). Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942-2008. *Political Analysis*, Vol 19, pp. 385–408.
10. Cellini, S. R., Ferreira, F., & Rothstein, J. (2010) The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design. *The Quarterly Journal of Economics*, Vol 125, Issue 1, pp. 215–261.
11. Cleveland, W. S. & Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, Vol 83, Issue 403. pp. 596–610.



- 
12. Cook, T. D. (2008). “Waiting for Life to Arrive”: A history of the regression-discontinuity design in Psychology, Statistics and Economics. *Journal of Econometrics*, Vol 142, pp. 636–654.
 13. Crown, W. H., Henk, H. J., & Vanness, D. J. (2011). Some Cautions on the Use of Instrumental Variables Estimators in Outcomes Research: How Bias in Instrumental Variables Estimators Is Affected by Instrument Strength, Instrument Contamination, and Sample Size. *Value in Health*. Vol 14, Issue 8, pp. 1078–1084.
 14. Dhrymes, P. J., & Lleras-Muney, A. (2006). Estimation of models with grouped and ungrouped data by means of “2SLS”. *Journal of Econometrics*, Vol 133, Issue 1, pp. 1–29.
 15. Eggers, A. C., Fowler, A., Hainmueller, J., Hall, A. B., & Snyder, J. M., Jr. (2015). On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from Over 40,000 Close Races. *American Journal of Political Science*, Vol 59, No.1, pp. 259–274.
 16. Fan, J. Q., & Gijbels, I. (1995). Adaptive Order Polynomial Fitting: Bandwidth Robustification and Bias Reduction. *Journal of Computational and Graphical Statistics*, Vol 4, No.3, pp. 213–227.
 17. Frölich, M. (2007) Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, Vol 139, pp. 35–75.
 18. Gelman, A., & Imbens, G. (2017). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, DOI: 10.1080/07350015.2017.1366909.
 19. Gelman, A., & Zelizer, A. (2015). Evidence on the deleterious impact of sustained use of polynomial regression on causal inference. *Research & Politics*, DOI: 10.1177/2053168015569830
 20. Hahn, J. Y., Todd, P., & Van der Klaauw, W. (1999). Evaluating the Effect of an Antidiscrimination Law Using a Regression-Discontinuity Design. *NBER Working Paper* No.7131.
 21. Hahn, J. Y., Todd, P., & Van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, Vol 69, No.1, pp. 201–209.
 22. Hsu, Y. C., Kuan, C. M., & Lo, G. T. Y. (2018). Estimating Treatment Effects in Regression Discontinuity Designs with Multiple Assignment Variables. *Working paper*.
 23. Imbens, G., & Zajonc, T. (2011). Regression Discontinuity Design with Multiple Forcing Variables. *Report*, Harvard University.

- 
24. Jacob, B. A., & Lefgren, L. (2004). Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. *Review of Economics and Statistics*. Vol 86, Issue 1. pp. 226–244.
25. Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*. Vol 142, pp.655–674.
26. Lee, D. S., & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, Vol 48, pp. 281–355.
27. Lo, G. T. Y. (2017). Estimating Treatment Effects in Regression Discontinuity Designs with Multiple Assignment Variables. *Master Thesis*, NTHU.
28. Moscoe, E., Bor, J., & Bärnighausen, T. (2015). Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. *Journal of Clinical Epidemiology*, Vol 68, pp. 132–143.
29. Murnane, R. J., & Papay, J. P. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, Vol 161, Issue 2, pp. 203–207.
30. Reardon, S. F., & Robinson, J. P. (2010). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness*, Vol 5, Issue 1, pp. 83–104.
31. Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), pp. 688–701.
32. Sekhon, J. S., & Titiunik, R. (2017). On Interpreting The Regression Discontinuity Design as a Local Experiment. *Advances in Econometrics*, Vol 38, pp. 1–28.
33. Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6), pp. 309–317.
34. Van der Klaauw, W. (2002). Estimating The Effect of Financial Aid Offer on College Enrollment: A Regression-Discontinuity Approach. *International Economic Review*, Vol 43, No.4, pp. 1249–1287.
35. Wong, V. C., Steiner, P. M., & Cook, T. D. (2013). Analyzing Regression-Discontinuity Designs With Multiple Assignment Variables: A Comparative Study of Four Estimation Methods. *Journal of Educational and Behavioral Statistics*, Vol 38, Issue 2, pp. 107–141.

Appendix

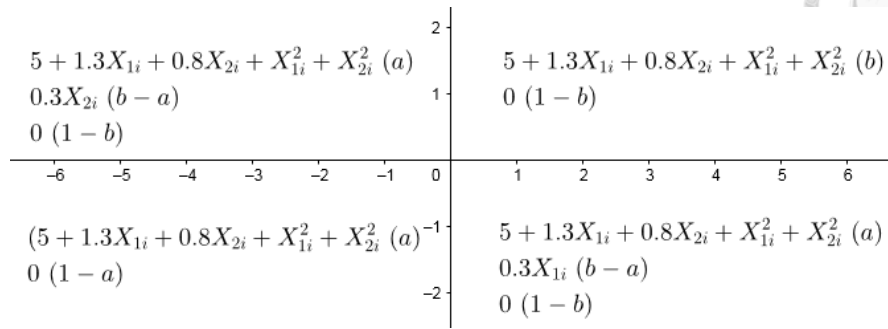
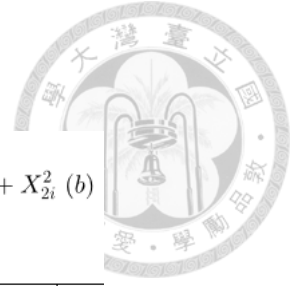


Figure 5: Distribution of treatment assignment in the four quadrants for DGP 4. The proportion in the parenthesis is the probability for each result. Note there is an underlying trend of $5 + X_{1i} + X_{2i} + X_{1i}^2 + 3X_{2i}^2$ for all the observations.

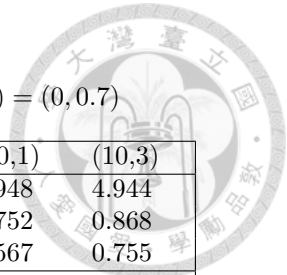


Table 2: Simulation Result: DGP 1 with local linear fitting and $(a, b) = (0, 0.7)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.979	4.916	4.920	4.966	4.948	4.944
	s.d.	0.374	0.566	1.351	0.433	0.752	0.868
	MSE	0.140	0.326	1.827	0.188	0.567	0.755
Intersection	Mean	4.969	4.914	4.927	4.997	4.966	4.958
	s.d.	0.464	0.622	1.386	0.521	0.831	0.916
	MSE	0.215	0.393	1.923	0.271	0.690	0.838
Average	Mean	4.972	4.907	4.918	4.975	4.947	4.946
	s.d.	0.408	0.590	1.348	0.466	0.790	0.886
	MSE	0.167	0.356	1.820	0.217	0.626	0.787
SampleRemained	Mean	1008.432	938.550	927.824	929.170	754.286	871.280
	s.d.	195.490	189.484	182.848	186.881	145.631	164.932

^a Sample remained gives the number of observations left in the bandwidth, or observations with $|X_{1i}| \leq h_1$ and $|X_{2i}| \leq h_2$.

Table 3: Simulation Result: DGP 1 with local linear fitting and $(a, b) = (0.15, 0.85)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.960	4.913	4.836	4.969	4.891	4.914
	s.d.	0.314	0.417	1.009	0.366	0.543	0.661
	MSE	0.100	0.181	1.044	0.135	0.306	0.444
Intersection	Mean	4.991	4.951	4.983	4.999	4.951	4.999
	s.d.	0.390	0.506	1.053	0.470	0.603	0.714
	MSE	0.152	0.258	1.106	0.221	0.365	0.509
Average	Mean	4.973	4.956	4.973	4.996	4.945	4.994
	s.d.	0.350	0.448	1.017	0.399	0.580	0.677
	MSE	0.123	0.202	1.033	0.159	0.338	0.457
Sample remained	Mean	1288.162	1254.992	1228.268	1228.382	1038.226	1153.720
	s.d.	275.506	261.905	247.481	257.727	238.677	221.687

^a Notations are defined as in table 2.

Table 4: Simulation Result: DGP 1 with local linear fitting and $(a, b) = (0.3, 1)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.959	4.918	4.754	4.958	4.928	4.870
	s.d.	0.302	0.301	0.315	0.309	0.363	0.334
	MSE	0.093	0.097	0.160	0.097	0.137	0.128
Intersection	Mean	4.982	5.008	5.009	5.007	5.061	5.019
	s.d.	0.396	0.410	0.462	0.413	0.475	0.439
	MSE	0.156	0.168	0.213	0.170	0.229	0.193
Average	Mean	4.985	4.999	5.000	5.007	5.034	5.021
	s.d.	0.339	0.349	0.374	0.360	0.413	0.391
	MSE	0.115	0.121	0.140	0.129	0.171	0.153
Sample remained	Mean	1279.150	1219.946	1210.248	1206.294	989.886	1158.996
	s.d.	252.224	222.004	234.789	244.770	211.477	231.304

^a Notations are defined as in table 2.

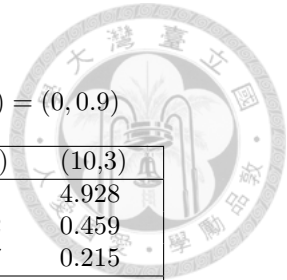


Table 5: Simulation Result: DGP 1 with local linear fitting and $(a, b) = (0, 0.9)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	5.020	4.992	4.948	5.016	5.001	4.928
	s.d.	0.287	0.342	0.684	0.297	0.432	0.459
	MSE	0.082	0.117	0.470	0.088	0.187	0.215
Intersection	Mean	5.021	4.978	4.934	5.012	5.008	4.945
	s.d.	0.367	0.417	0.712	0.370	0.500	0.521
	MSE	0.135	0.174	0.511	0.137	0.250	0.274
Average	Mean	5.020	4.981	4.943	5.023	5.010	4.935
	s.d.	0.310	0.371	0.693	0.322	0.460	0.474
	MSE	0.096	0.138	0.483	0.104	0.211	0.228
Sample remained	Mean	1013.306	940.402	936.252	933.782	751.420	863.276
	s.d.	190.187	186.075	179.178	187.837	142.953	166.353

^a Notations are defined as in table 2.

Table 6: Simulation Result: DGP 1 with local linear fitting and $(a, b) = (0.05, 0.95)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.998	4.994	4.930	5.002	4.984	4.972
	s.d.	0.222	0.247	0.517	0.246	0.350	0.329
	MSE	0.049	0.061	0.272	0.060	0.122	0.109
Intersection	Mean	5.005	4.988	4.965	5.001	4.995	5.002
	s.d.	0.291	0.313	0.554	0.316	0.409	0.392
	MSE	0.085	0.098	0.307	0.100	0.167	0.153
Average	Mean	4.998	5.000	4.964	5.001	5.000	4.994
	s.d.	0.245	0.273	0.533	0.272	0.379	0.357
	MSE	0.060	0.075	0.285	0.074	0.143	0.127
Sample remained	Mean	1320.746	1313.286	1290.748	1295.264	1094.296	1218.886
	s.d.	278.426	265.220	273.797	269.576	265.182	262.032

^a Notations are defined as in table 2.

Table 7: Simulation Result: DGP 1 with local linear fitting and $(a, b) = (0.1, 1)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.995	4.983	4.925	4.991	4.977	4.965
	s.d.	0.224	0.228	0.240	0.226	0.257	0.231
	MSE	0.050	0.052	0.063	0.051	0.066	0.054
Intersection	Mean	5.007	5.003	4.997	5.001	5.007	4.997
	s.d.	0.299	0.301	0.334	0.298	0.340	0.327
	MSE	0.090	0.090	0.111	0.089	0.115	0.107
Average	Mean	5.004	5.006	4.993	5.001	5.015	5.002
	s.d.	0.253	0.252	0.275	0.249	0.295	0.273
	MSE	0.064	0.063	0.076	0.062	0.087	0.074
Sample remained	Mean	1296.000	1300.386	1222.216	1269.518	1057.524	1194.468
	s.d.	261.402	257.657	237.635	263.374	244.688	256.852

^a Notations are defined as in table 2.

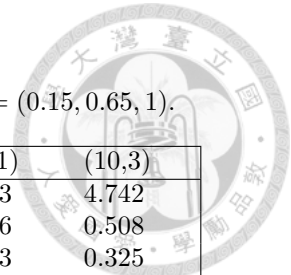


Table 8: Simulation Result: DGP 2 with local linear fitting and $(a, b, c) = (0.15, 0.65, 1)$.

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.919	4.862	4.541	4.900	4.763	4.742
	s.d.	0.448	0.465	0.502	0.473	0.526	0.508
	MSE	0.207	0.235	0.462	0.233	0.333	0.325
Intersection	Mean	4.979	5.000	4.988	4.991	5.034	5.030
	s.d.	0.314	0.339	0.361	0.325	0.387	0.345
	MSE	0.099	0.114	0.130	0.105	0.151	0.120
Average	Mean	4.932	4.989	4.959	4.982	5.018	5.055
	s.d.	0.607	0.623	0.634	0.599	0.852	0.635
	MSE	0.372	0.387	0.402	0.358	0.725	0.406
Sample remained	Mean	1308.878	1295.656	1312.146	1260.562	1091.402	1188.868
	s.d.	251.654	256.881	271.324	253.404	228.196	212.593

^a Notations are defined as in table 2.

Table 9: Simulation Result: DGP 2 with local linear fitting and $(a, b, c) = (0.15, 0.5, 0.85)$.

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.949	4.877	4.636	4.902	4.734	4.675
	s.d.	0.546	0.561	0.711	0.510	0.778	0.823
	MSE	0.300	0.329	0.638	0.269	0.675	0.782
Intersection	Mean	5.000	4.981	4.998	5.004	4.997	4.981
	s.d.	0.410	0.399	0.478	0.401	0.484	0.476
	MSE	0.168	0.159	0.228	0.160	0.234	0.227
Average	Mean	4.974	4.995	4.930	4.991	4.926	4.865
	s.d.	0.700	1.049	1.108	0.698	1.471	1.547
	MSE	0.489	1.097	1.229	0.486	2.165	2.407
Sample remained	Mean	1330.712	1293.502	1312.912	1296.510	1096.954	1205.984
	s.d.	270.761	246.957	246.545	260.741	229.062	246.276

^a Notations are defined as in table 2.

Table 10: Simulation Result: DGP 2 with local linear fitting and $(a, b, c) = (0.15, 0.75, 1)$.

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.921	4.833	4.402	4.848	4.628	4.554
	s.d.	0.598	0.578	0.597	0.565	0.629	0.626
	MSE	0.363	0.361	0.714	0.342	0.533	0.590
Intersection	Mean	4.996	5.012	5.035	4.991	5.009	4.999
	s.d.	0.311	0.317	0.357	0.311	0.360	0.336
	MSE	0.096	0.101	0.129	0.096	0.129	0.112
Average	Mean	4.976	4.943	4.983	4.982	4.974	5.008
	s.d.	0.881	2.508	0.975	0.879	1.116	1.220
	MSE	0.775	6.279	0.948	0.771	1.243	1.485
Sample remained	Mean	1321.252	1292.352	1276.594	1275.728	1098.172	1204.600
	s.d.	262.817	269.084	249.843	238.720	211.862	232.426

^a Notations are defined as in table 2.

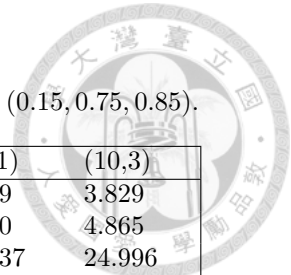


Table 11: Simulation Result: DGP 2 with local linear fitting and $(a, b, c) = (0.15, 0.75, 0.85)$.

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.604	2.145	3.375	4.454	4.129	3.829
	s.d.	6.690	53.472	4.889	2.825	5.490	4.865
	MSE	44.823	2861.702	26.492	8.264	30.837	24.996
Intersection	Mean	4.959	5.003	4.987	4.971	5.009	5.036
	s.d.	0.405	0.375	0.508	0.400	0.484	0.482
	MSE	0.166	0.141	0.258	0.161	0.234	0.233
Average	Mean	5.511	6.006	6.328	7.635	2.230	4.221
	s.d.	24.891	12.091	24.007	43.818	48.126	19.166
	MSE	618.560	146.912	576.945	1923.097	2319.149	367.194
Sample remained	Mean	1314.552	1269.826	1280.038	1273.130	1111.452	1191.682
	s.d.	276.747	244.002	257.313	258.333	232.654	219.924

^a Notations are defined as in table 2.

Table 12: Simulation Result: DGP 3 with local linear fitting and $(a, b) = (0.15, 0.85)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.940	4.809	4.345	4.867	4.691	4.566
	s.d.	0.309	0.322	0.462	0.304	0.378	0.444
	MSE	0.099	0.140	0.642	0.110	0.238	0.385
Intersection	Mean	5.001	5.013	5.031	4.997	4.973	4.984
	s.d.	0.388	0.424	0.565	0.422	0.506	0.527
	MSE	0.151	0.180	0.319	0.177	0.256	0.277
Average	Mean	5.011	5.014	5.009	4.995	4.980	4.982
	s.d.	0.342	0.351	0.458	0.342	0.417	0.445
	MSE	0.117	0.123	0.209	0.117	0.174	0.198
Sample remained	Mean	1358.054	1337.854	1313.052	1318.334	1128.498	1251.860
	s.d.	276.613	267.152	251.026	267.817	244.652	252.005

^a Notations are defined as in table 2.

Table 13: Simulation Result: DGP 3 with local linear fitting and $(a, b) = (0.05, 0.95)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.931	4.805	4.350	4.860	4.668	4.583
	s.d.	0.215	0.223	0.281	0.236	0.267	0.257
	MSE	0.051	0.087	0.501	0.075	0.181	0.240
Intersection	Mean	4.999	5.006	5.004	4.988	4.986	4.998
	s.d.	0.286	0.296	0.338	0.297	0.354	0.322
	MSE	0.082	0.087	0.114	0.088	0.125	0.104
Average	Mean	4.994	5.004	5.004	4.991	4.996	4.999
	s.d.	0.242	0.241	0.274	0.263	0.288	0.270
	MSE	0.058	0.058	0.075	0.069	0.083	0.073
Sample remained	Mean	1394.576	1356.994	1365.906	1370.580	1199.980	1314.644
	s.d.	287.258	282.219	275.110	285.922	269.809	285.504

^a Notations are defined as in table 2.

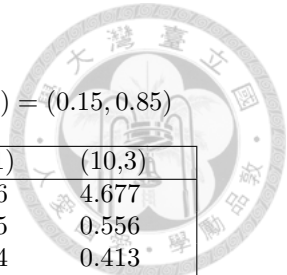


Table 14: Simulation Result: DGP 3 with local quadratic fitting and $(a, b) = (0.15, 0.85)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.958	4.850	4.540	4.928	4.716	4.677
	s.d.	0.405	0.433	0.702	0.419	0.595	0.556
	MSE	0.166	0.209	0.704	0.180	0.434	0.413
Intersection	Mean	5.010	4.999	4.954	5.009	4.928	4.996
	s.d.	0.568	0.583	0.825	0.556	0.779	0.749
	MSE	0.322	0.339	0.682	0.309	0.611	0.561
Average	Mean	4.994	4.986	4.952	5.019	4.913	4.942
	s.d.	0.457	0.495	0.708	0.483	0.651	0.627
	MSE	0.208	0.245	0.503	0.233	0.431	0.396
Sample remained	Mean	2137.120	2157.456	2092.726	2094.126	1865.108	2030.348
	s.d.	333.688	340.271	319.357	316.692	310.207	304.928

^a Notations are defined as in table 2.

Table 15: Simulation Result: DGP 3 with local quadratic fitting and $(a, b) = (0.05, 0.95)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.952	4.883	4.573	4.921	4.804	4.748
	s.d.	0.320	0.318	0.379	0.307	0.346	0.368
	MSE	0.105	0.114	0.326	0.100	0.158	0.198
Intersection	Mean	4.993	5.013	4.984	4.999	5.000	5.016
	s.d.	0.430	0.429	0.483	0.403	0.457	0.479
	MSE	0.185	0.184	0.233	0.162	0.208	0.230
Average	Mean	4.995	5.012	4.983	5.012	5.006	5.013
	s.d.	0.357	0.359	0.401	0.356	0.385	0.386
	MSE	0.127	0.129	0.161	0.127	0.148	0.149
Sample remained	Mean	2180.940	2141.522	2161.770	2135.266	1956.940	2088.624
	s.d.	342.482	341.904	338.409	336.649	343.685	363.869

^a Notations are defined as in table 2.

Table 16: Simulation Result: DGP 4 with local quadratic fitting and $(a, b) = (0.15, 0.85)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.946	4.866	4.471	4.896	4.793	4.595
	s.d.	0.436	0.499	2.258	0.489	0.910	1.428
	MSE	0.193	0.267	5.366	0.250	0.869	2.198
Intersection	Mean	5.000	5.014	4.851	4.999	4.938	4.809
	s.d.	0.586	0.594	2.821	0.629	1.164	1.703
	MSE	0.343	0.352	7.967	0.395	1.356	2.930
Average	Mean	4.995	4.996	4.794	4.982	5.052	4.824
	s.d.	0.497	0.549	2.583	0.556	2.852	1.537
	MSE	0.247	0.301	6.699	0.309	8.118	2.388
Sample remained	Mean	2121.978	2127.972	2151.772	2040.560	1435.292	2008.346
	s.d.	302.254	319.279	334.883	342.769	279.412	339.184

^a Notations are defined as in table 2.

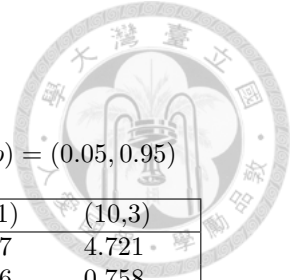


Table 17: Simulation Result: DGP 4 with local quadratic fitting and $(a, b) = (0.05, 0.95)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
Union	Mean	4.967	4.868	4.579	4.925	4.817	4.721
	s.d.	0.312	0.356	1.126	0.340	0.546	0.758
	MSE	0.098	0.144	1.443	0.121	0.331	0.651
Intersection	Mean	5.011	4.976	4.969	5.015	4.963	4.948
	s.d.	0.407	0.450	1.341	0.458	0.660	0.897
	MSE	0.166	0.203	1.795	0.210	0.436	0.806
Average	Mean	5.012	4.984	4.978	4.995	4.962	4.964
	s.d.	0.366	0.383	1.237	0.380	0.580	0.781
	MSE	0.134	0.147	1.527	0.144	0.337	0.610
Sample remained	Mean	2163.108	2158.510	2128.030	2009.734	1519.178	2070.430
	s.d.	354.902	349.850	361.360	337.632	340.813	341.996

^a Notations are defined as in table 2.

Table 18: Statistics of standardized chosen bandwidth: DGP 3 with local linear fitting and $(a, b) = (0.15, 0.85)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
h1ratio	Mean	0.736	0.734	0.718	0.711	0.600	0.678
	s.d.	0.121	0.124	0.113	0.123	0.114	0.119
h2ratio	Mean	0.736	0.727	0.725	0.736	0.728	0.727
	s.d.	0.126	0.122	0.119	0.133	0.113	0.120

^a h1ratio is defined as the bandwidth chosen for X_1 (h_1) divided by s_1 , or h_1/s_1 .
h2ratio is defined similarly.

Table 19: Statistics of standardized chosen bandwidth: DGP 3 with local quadratic fitting and $(a, b) = (0.15, 0.85)$

	(s_1, s_2)	(1,1)	(3,3)	(10,10)	(3,1)	(10,1)	(10,3)
h1ratio	Mean	0.994	0.986	0.985	0.957	0.867	0.946
	s.d.	0.155	0.154	0.153	0.142	0.151	0.153
h2ratio	Mean	0.998	0.984	0.998	1.006	0.992	0.990
	s.d.	0.149	0.158	0.151	0.158	0.152	0.158

^a h1ratio is defined as the bandwidth chosen for X_1 (h_1) divided by s_1 , or h_1/s_1 .
h2ratio is defined similarly.