ESSAYS IN SOCIAL CHOICE AND ECONOMETRICS

ZHUZHU ZHOU

A dissertation
submitted to the Faculty of
the department of Economics
in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Boston College
Morrissey College of Arts and Sciences
Graduate School

August, 2021

ESSAYS IN SOCIAL CHOICE AND ECONOMETRICS

Zhuzhu Zhou

Advised by Professor Uzi Segal, Professor Arthur Lewbel,
and Professor M. Bumin Yenmez

Abstract

The dissertation studies the property of transitivity in the social choice theory. I explain why we should care about transitivity in decision theory. I propose two social decision theories: redistribution regret and ranking regret, study their properties of transitivity, and discuss the possibility to find a best choice for the social planner. Additionally, in the joint work, we propose a general method to construct a consistent estimator given two parametric models, one of which could be incorrectly specified.

In "Why Transitivity", to explain behaviors violating transitivity, e.g., preference reversals, some models, like regret theory, salience theory were developed. However, these models naturally violate transitivity, which may not lead to a best choice for the decision maker. This paper discusses the consequences and the possible extensions to deal with it.

In "Redistribution Regret and Transitivity", a social planner wants to allocate resources, e.g., the government allocates fiscal revenue or parents distribute toys to children. The social planner cares about individuals' feelings, which depend both on their assigned resources, and on the alternatives they might have been assigned. As a result, there could be intransitive cycles. This paper shows that the preference orders are generally non-transitive but there are two exceptions: fixed total resource and one extremely sensitive individual, or only two individuals with the same non-linear individual regret function.

In "Ranking Regret", a social planner wants to rank people, e.g., assign airline passengers a boarding order. A natural ranking is to order people from most to least sensitive to their rank. But people's feelings can depend both on their assigned rank, and on the alternatives they might have been assigned. As a result, there may be no best ranking, due to intransitive cycles. This paper shows how to tell when a best ranking exists, and that when it exists, it is indeed the natural ranking. When this best does not exist, an alternative

second-best group ranking strategy is proposed, which resembles actual airline boarding policies.

In "Over-Identified Doubly Robust Identification and Estimation", joint with Arthur Lewbel and Jinyoung Choi, we consider two parametric models. At least one is correctly specified, but we don't know which. Both models include a common vector of parameters. An estimator for this common parameter vector is called Doubly Robust (DR) if it's consistent no matter which model is correct. We provide a general technique for constructing DR estimators (assuming the models are over identified). Our Over-identified Doubly Robust (ODR) technique is a simple extension of the Generalized Method of Moments. We illustrate our ODR with a variety of models. Our empirical application is instrumental variables estimation, where either one of two instrument vectors might be invalid.

Contents

List of Tables

List of Figures

No Regrets.

1 Chapter 1: Why Transitivity?

1.1 Introduction

The widely accepted model about people's decision making under uncertainty is the expected utility model, proposed by von Neumann and Morgenstern [42]. The basic assumptions for the preferences in this model include completeness, transitivity, independence, and continuity.

However, we can observe some behaviors violating transitivity, for example, preference reversals, die examples. These phenomenons cannot be explained by the expected utility model. Some alternative models were developed to explain these behaviors, e.g., regret theory, salience theory.

As a result, these models naturally predict violations of transitivity, which also cause problems. One of the problems is there may not exist a best choice on the basis of binary preferences (see Fishburn and Lavalle [13]). This paper reviews the drawbacks of each model, and discusses some possible changes to deal with the consequences.

The next section describes some behaviors violate transitivity. Section 3 reviews some models explaining behaviors violating transitivity, and discusses the consequences and possible extensions. The last section concludes.

1.2 Behaviors Violate Transitivity

Fishburn and Lavalle [13] summarized that the violations of transitivity are often related to preference comparisons based on multi-dimensional features of the alternatives under consideration. This section reviews several examples which violate transitivity, including preference reversals, die example, marriage partners, social choice, and coffee and sugar.

For decision under risk, preference reversals involve two features: monetary outcomes and probabilities; die example involves monetary outcomes and states. For decision under certainty, marriage partner example considers three attributes, and social choice example considers n individuals. I will focus on the four kinds of behaviors. But here I still introduce the last example, coffee and sugar, which violates the transitivity of indifference. It is different from the previous four examples and is caused by unnoticeable difference (see Gilboa and Lapson [14]).

1.2.1 Preference Reversals

Preference reversal is a special case of the violations of transitivity (see Tversky [40], Lichtenstein and Slovic [21], Grether and Plott [15], Bell [2], etc.).

Consider two lotteries L_1 and L_2. L_1 has a higher payoff but lower probability to win while L_2 has a higher probability of winning but lower payoff. Preference reversal phenomenon normally describes that the certainty equivalent of L_1 is higher than the certainty equivalent of lottery L_2, however, in a direct comparison of the lotteries, L_2 is preferred. In other words, one individual prefers a lottery L_2 to L_1, but the individual, in possession of one or the other, would sell L_2 for less. When this happens, let $CE(L_1)$ and $CE(L_2)$ be the certainty equivalent of L_1 and L_2, we can get a preference cycle

$$L_2 \succ L_1 \sim (CE(L_1), 1) \succ (CE(L_2), 1) \sim L_2$$

Therefore, it shows a violation of transitivity. Here is a specific example provided by Fishburn and Lavalle [13]. $L_1 = (10,000, 0.3; 0, 0.7)$, $L_2 = (3,000, 0.9; 0, 0.1)$. An individual may prefer L_2 to L_1 because L_2 has a good chance for a nice outcome. However, this individual may not be willing to sell L_1 for less than say \$2800 because of the potential to get a high outcome while at the same time being willing to sell L_2 for \$2600.

Experimental results of preference reversal phenomenon was reported by Lichtenstein and Slovic [21] and further studied by Grether and Plott [15]. According to Grether and Plott [15], after controlling all the economic-theoretic explanations of the phenomenon which they could find, the preference reversal phenomenon which is inconsistent with the traditional statement of preference theory remains.

1.2.2 Die Example

Consider two lotteries L_1 and L_2. Assume that the states are determined by a fair die. The number of the die and the payoffs are shown in Table 1.1. As the probability for each state is 1/6, the two lotteries should be indifferent by expected utility theory. However, people may prefer L_1 to L_2 as L_1 pays \$100 more than L_2 for five of the six states.

Some papers, e.g., Tversky [41], Loomes and Sugden [24], Bell [2] discussed such preferences by considering regret or elation. They argue that it is reasonable for people to think L_1 and L_2 are not indifferent, e.g., $L_1 \succ L_2$. For the same reason, they will think $L_1 \succ L_2 \succ L_3 \succ L_4 \succ L_5 \succ L_6 \succ L_1$, which violates transitivity.

Table 1.1: Lotteries

	1	2	3	4	5	6
L_1	\$100	\$200	\$300	\$400	\$500	\$600
L_2	\$600	\$100	\$200	\$300	\$400	\$500
L_3	\$500	\$600	\$100	\$200	\$300	\$400
L_4	\$400	\$500	\$600	\$100	\$200	\$300
L_5	\$300	\$400	\$500	\$600	\$100	\$200
L_6	\$200	\$300	\$400	\$500	\$600	\$100

1.2.3 Marriage Partners

When people evaluate their choices by pairwise comparing more than two attributes, preference cycles are easily observed. I replicated the experiment mentioned in May [28] and included five questions:

1. You are making comparisons among hypothetical marriage partners characterized by three attributes. Please select all the attributes you care about: intelligence, looks, and wealth.

2. x is very intelligent, plain-looking, and well off; y is intelligent, very good looking, and poor. Which one do you prefer?

3. x is intelligent, very good looking, and poor; y is fairly intelligent, good looking, and rich. Which one do you prefer?

4. x is very intelligent, plain-looking, and well off; y is fairly intelligent, good looking, and rich. Which one do you prefer?

5. x is very intelligent, plain-looking, and well off; y is intelligent, very good looking, and poor; z is fairly intelligent, good looking, and rich. Which one do you prefer?

6 out of 40 participants show intransitive cycles. Only 3 of them claim that they care about all the three attributes. The interesting thing is that all of the six students can make decisions for question 5. Meanwhile, 4 out of 40 students do not have an answer for question 5 although they did not show intransitive cycles. Additionally, when we compare the best choice implied by questions 2, 3, and 4 to the choice in question 5, 16 out of 40 students show contradictions.

This replicated experiment shows that people occasionally violate transitivity when they evaluate their outcomes in multiple dimensions.

1.2.4 Social Choice

Consider a social planner does pairwise comparisons among three social policies: A, B, and C. The social planner collects people's opinions about each pair and follows majority rule. Suppose there are three individuals $\{1, 2, 3\}$ in the society, they have the following preferences:

$$A \succ^1 B \succ^1 C$$

$$B \succ^2 C \succ^2 A$$

$$C \succ^3 A \succ^3 B$$

When the social planner compares policies A to B, individuals 1 and 3 thinks $A \succ B$, by majority rule, the social planner also thinks $A \succ B$. Similarly, by majority rule, the social planner gets $B \succ C$ and $C \succ A$. Therefore, the social preference with pairwise comparisons and majority rule violates transitivity.

Consider a more specific example. Here is a current income distribution X, and two public policies A and B. If policy A is implemented, the new income distribution will be Y. If policy B, the income distribution will be Z. The social planner needs to decide whether to launch policy A or B or do nothing, which is essentially comparing income distributions X, Y, and Z.

Suppose that there are 3 individuals $\{I_1, I_2, I_3\}$ in the society. Traditionally, we evaluate a policy independently by its outcome. People's identities or feelings are often ignored. For example, we may consider the following three income distributions as the same.

Table 1.2: Income Distributions

	I_1	I_2	I_3
X	1	2	3
Y	2	3	1
Z	3	1	2

However, this is not intuitive. As distribution Y improves two persons' allocations relative to X, Y could be preferred to X. Similarly, a social planner may think X is preferred to Z and Z is preferred to Y, which forms a preference cycle.

1.2.5 Coffee and Sugar

Suppose a person prefers more sugar in his coffee, but cannot taste the difference of less than 2 tablespoons sugar. That says, this person thinks a cup of coffee with

1 tablespoon sugar and a cup of coffee with 2 tablespoons sugar are indifferent. Similarly, a cup of coffee with 2 tablespoons sugar and with 3 tablespoons sugar are also indifferent. However, a cup of coffee with 3 tablespoons sugar is preferred to the coffee with 1 tablespoon sugar. This example is a violation of indifference preference relations. The reason is that people sometimes cannot notice small differences but can identify accumulated small differences.

To explain this kind of behaviors, we can use semi-order (see Gilboa and Lapson [14], a binary relation P for which there is a utility U representing it in the sense that xPy iff $U(x) - U(y) > 1$. I will focus on the previous four kinds of behaviors and not discuss this category of behaviors in the following sections.

1.3 Models

To explain the behaviors violating transitivity, we need to use reference-dependent models. For example, prospect theory and regret theory can be used to explain preference reversals. For the die example and marriage partners, regret theory provides explanations. Redistribution regret was proposed to explain the violations of transitivity regarding social choice.

1.3.1 Prospect Theory

Prospect theory proposed by Kahneman and Tversky [18] is a decision theory under risk. It says that people have two phases in the choice process: editing and evaluation. Editing mainly includes coding, which locates the reference point and codes the outcome as gains or losses. Evaluation computes a value and chooses a higher value.

The overall value of an edited prospect, denoted V, is expressed in terms of two scales, weight function π and value function v. π is an increasing function of p, with $\pi(0) = 0$, $\pi(1) = 1$, and $\pi(p) + \pi(1-p)$ is typically less than unity. v measures the value of deviations from the reference point, i.e., gains and losses. It is normally concave above the reference point and often convex below it. Additionally, the value function for losses is steeper than the value function for gains.

The formula that Kahneman and Tversky assume for the evaluation phase is

$$V = \sum_{i=1}^{n} \pi(p_i) v(x_i)$$

This model can explain preference reversals. Take lotteries $L_1 = (10,000, 0.3; 0, 0.7)$

and $L_2 = (3,000, 0.9; 0, 0.1)$ for example. Suppose people's initial endowment is 0. If the individual buys L_1 and L_2, the prices are p_1^b and p_2^b. If the individual sells L_1 and L_2, the prices are p_1^s and p_2^s. We have

$$v(10,000 - p_1^b)\pi(0.3) + v(-p_1^b)\pi(0.7) = 0 \tag{1.1}$$

$$v(3,000 - p_2^b)\pi(0.9) + v(-p_2^b)\pi(0.1) = 0 \tag{1.2}$$

$$v(p_1^s - 10,000)\pi(0.3) + v(p_1^s)\pi(0.7) = 0 \tag{1.3}$$

$$v(p_2^s - 3,000)\pi(0.9) + v(p_2^s)\pi(0.1) = 0 \tag{1.4}$$

As $v(\cdot)$ is asymmetric, by equations (1.1) and (1.3), $p_1^b \neq p_1^s$. Similarly, by equations (1.2) and (1.4), $p_2^b \neq p_2^s$. It is possible to have $p_1^b < p_2^b$ and $p_1^s > p_2^s$ according to the assumptions.

However, prospect theory has some problems. First, its discussion relies on a reference point, but this model did not specify what the reference point is. Second, there are too many assumptions about the weight function and value function, which arises problems. For example, if the weight function is continuous, it implies a linear weight function (see Fishburn [12]).

1.3.2 Regret Theory

Regret theory was independently proposed by Bell [2] and by Loomes and Sugden [24]. It suggests that the utility from an outcome of a random variable depends not only on the outcome itself but also on the outcome that could have been obtained had the decision-maker chosen another random variable.

Let $\psi(x, y)$ measure the regret or elation a person feels when observing that he received x while the alternative choice would have yielded him y. This function satisfies skew symmetric $\psi(x, y) = -\psi(y, x)$, monotonocity $\frac{\partial\psi(x,y)}{\partial x} > 0$ and $\frac{\partial\psi(x,y)}{\partial y} < 0$, and regret aversion $\psi(x, z) > \psi(x, y) + \psi(y, z)$ for $x > y > z$. For two lotteries $X = (x_1, s_1; ...; x_n, s_n)$ and $Y = (y_1, s_1; ...; y_n, s_n)$, regret theory suggests that

$$X \gtrsim Y \Longleftrightarrow \sum \Pr(s_i)\psi(x_i, y_i) \geqslant 0$$

Regret theory was initially developed to explain preference reversals (see Lichtenstein and Slovic [21], Lindman [22], Grether and Plott [15]). The idea can also be used to explain die example, marriage partners, and social choice. I will discuss social choice further in the next section.

Recently, regret theory has been modified for use in other situations[1] or revised to a new model. For example, Bikhchandani and Segal [4] introduce distribution regret, where each outcome is compared to the entire alternative distribution instead of a single outcome under the same state.

As regret theory naturally predicts violations of transitivity (see Bikhchandani and Segal [3]), it generally cannot provide a best choice when there are more than two choices. But some extensions of standard regret theory can.

For example, Bell [2] mentioned that given an original status quo, for each alternative, we can get a level of regret and make decisions by comparing these levels. This modification leads to transitivity. Later Buturak and Evren [9] focus on a situation with more options than can be considered and proposes asymmetric regret, which assumes that decision-makers do not consider the possible regret for choosing the default option but consider the regret for choosing some other options. The default option can be treated as a fixed reference point, therefore, asymmetric regret satisfies transitivity by implementing the idea.

Alternatively, some other papers (see Luce and Raiffa [26]) suggest to compare one outcome to the best of the others and use minimax regret to make decisions among multiple choices. For instance, Sarver [34] proposes anticipating regret, which studies preferences over menus. This model assumes that a decision-maker feels regret as the realized outcome is compared to the best alternative outcome in a given menu. The best outcome is defined as a fixed reference point, therefore, anticipating regret also satisfies transitivity.

Additionally, Loomes and Sudgen [24] assigns action weights to each action in the choice set S. For each action i, define the aggregate modefied utility as

$$E_i^S = \sum_{k \in S} \frac{a_k^S}{1 - a_i^S} E_i^k \quad (k \neq i)$$

where E_i^k represents the expected modified utility of choosing action A_i in a situation where the only alternative is action A_k. The individual's decision rule, as in the case of pairwise choice, would be to maximize expected modified utility. Hence, this model admits transitive preference orders.

[1]Further studies can be found in Loomes and Sugden [25], Sugden [39], Quiggin [33], Starmer and Sugden [35], Hayashi [16], Bleichrod, Cillo and Diecidue [5], Stoye [38], Diecidue and Somasundaram [10], Levy [20]. And other extensions can be found in Braun and Munermann [8], Muermann, Mitchell and Volkman [32], Filiz-Ozba and Ozbay [11], Michenaud and Solnik [29], Maccheron, Marinacci and Rustichini [27], Bleichrodt and Wakker [6].

1.3.3 Redistribution Regret and Ranking Regret

Redistribution regret is designed to compare resource allocations, which involve individuals and one social planner. Although redistribution regret and regret theory are similar in the key assumption, they have significant differences. First, "regret" in redistribution regret is produced by individuals and aggregated by a social planner, who is the decision maker. Second, redistribution regret assumes that different individuals can have different regret functions, which is the case that the agent has different regret functions under different states in regret theory.

Suppose there are $n \geq 2$ individuals in the society, and two allocations $x = (x_1, ..., x_n)$, $y = (y_1, ..., y_n)$. Each individual uses his individual regret function $\psi_i(x_i, y_i)$ to measure regret (or elation) if she gets x_i instead of y_i. This non-linear individual regret function satisfies skew symmetric, which says that $\psi_i(x_i, y_i) = -\psi_i(y_i, x_i)$ and $\psi_i(x_i, x_i) = 0$; monotonicity, $\frac{\partial \psi_i(x_i, y_i)}{\partial x_i} > 0$ and $\frac{\partial \psi_i(x_i, y_i)}{\partial y_i} < 0$; normalization, $\psi_i(m, 0) = 1$ for any i, where $m > 0$ is a fixed number of resource; non-linearity, for some $x_i > y_i > z_i \geq 0$, $\psi_i(x_i, y_i) + \psi_i(y_i, z_i) \neq \psi_i(x_i, z_i)$.

The social planner uses a social regret function

$$W(x, y) = V(\psi_1(x_1, y_1), ..., \psi_n(x_n, y_n))$$

to aggregate individual regrets, and decide which allocation is better.

Redistribution regret is saying that

$$x \succsim y \quad \Leftrightarrow \quad W(x, y) = V(\psi_1(x_1, y_1), ..., \psi_n(x_n, y_n)) \geq 0$$

Zhou [43] shows that the preference orders are almost never transitive except one case where the resources are fixed and one individual out of three is extremely sensitive.

When we only consider people's rank in the framework of redistribution regret — ranking regret (see Zhou [44]), the property of transitivity changes. Like redistribution regret, ranking regret too can lead to preference cycles, where the social planner may find out that there is a positive aggregate satisfaction when ranking X is replaced with Y, positive aggregate satisfaction when Y is replaced by Z, and again a positive aggregation of satisfaction when Z is replaced with X. However, since different individuals have different sensitivity functions, the social planner may find a socially optimal ranking. Although the social preference order may not be transitive, this ranking may nonetheless be optimal, if it is not involved in any preference cycle. For

example, preferences over $\{X, Y, Z\}$ may form a non-transitive cycle, yet X^*, which orders individuals by their sensitivities, may be better than all of them.

Zhou [44] outlines conditions under which X^* is best, even if these conditions do not eliminate violations of transitivity. Once the best ranking is chosen, the social planner will not have any incentive to switch to another ranking. Regarding the case where X^* is not best, Zhou [44] proposes group ranking to find a second-best choice where the social planner can always achieve his goal by constructing groups appropriately and applying the natural ordering at the group level.

To sum up, when we apply the idea of regret theory to social choice, the violations of transitivity is essentially the same, but slightly mitigated because different individuals can have different regret functions. The real difference is that best choice and non-transitivity can exist at the same time, which makes transitivity less necessary to guarantee a best choice.

1.3.4 Salience Theory

Salience theory is proposed by Bordalo, Gennaioli, and Shleifer [7]. It shares the same assumption of pairwise comparisons with regret theory. But salience theory has different psychological motivations. It assumes that the decision maker's attention is drawn to salient payoffs, and the true probabilities are replaced by decision weights distorted in favor of salient payoffs. In other words, the decision maker overweights the salient attribute, which is the dimension in which the option is most different relative to alternatives of choice or expectations.

Consider two lotteries, L_1 and L_2. For each state $s \in S$, the probability π_s is known and satisfies $\sum_{s\in S} \pi_s = 1$. The monetary payoffs in each state s is x_s^i, where $i = 1, 2$. Let x_s^{-i} be the payoff in s of lottery L_j where $j \neq i$.

Based on the payoffs, salience theory defines a salience function $\sigma(x_s^i, x_s^{-i})$ to transform the object probability when they evaluate the lotteries. This function satisfies ordering $\sigma(x_s^i, x_s^{-i}) > \sigma(x_s^{i\prime}, x_s^{-i\prime})$ if $[\min(x_s^{i\prime}, x_s^{-i\prime}), \max(x_s^{i\prime}, x_s^{-i\prime})]$ is a subset of $[\min(x_s^i, x_s^{-i}), \max(x_s^i, x_s^{-i})]$, diminishing sensitivity $\sigma(x_s^i + \varepsilon, x_s^{-i} + \varepsilon) < \sigma(x_s^i, x_s^{-i})$ if x_s^i, x_s^{-i}, $\varepsilon > 0$.

According to Herweg and Muller [17], we can write salience theory as the following:

$$L_1 \succ L_2 \Longleftrightarrow \sum_{s\in S} f(\sigma(x_s^1, x_s^2))[v(x_s^1) - v(x_s^2)] \geq 0$$

Mathematically, this function form is less general than the standard regret theory as we can consider $\psi(x, y) = f(\sigma(x, y))[v(x) - v(y)]$ as a special case of regret function.

Salience theory can be used to explain preference reversals. It also violates transitivity if we directly apply the pairwise comparisons to many lotteries. To fix this problem, in the online appendix of Bordalo, Gennaioli, and Shleifer [7], they define a new salience function $\sigma(x_s^i, f(x_s^{-i}))$, where $f(x_s^{-i})$ is an average of the payoffs of all the other lotteries under s state.

1.4 Conclusions

People violate transitivity occasionally, especially when they evaluate their outcomes in multiple dimensions. So it is necessary to consider models implementing these intuitions. Most of the models are based on reference dependent preferences. As these models may violate transitivity, the decision maker may not be able to find a best choice. Some papers solve this problem by fixing the reference point or considering all possible alternatives, and some papers find a way to accommodate both non-transitivity and best choice.

2 Chapter 2: Redistribution Regret and Transitivity

2.1 Introduction

Consider a current income distribution X, and two public policies A and B. If policy A is implemented, the new income distribution will be Y. If policy B, the income distribution will be Z. The social planner needs to decide whether to launch policy A or B or do nothing, which is essentially comparing income distributions X, Y, and Z.

Suppose that there are three individuals $\{I_1, I_2, I_3\}$ in the society. Traditionally, we evaluate a policy independently by its outcome. People's identities or feelings are often ignored. For example, we may consider the following three income distributions as the same.

Table 2.1: Income Distributions

	I_1	I_2	I_3
X	1	2	3
Y	2	3	1
Z	3	1	2

However, this is not intuitive. As distribution Y improves two persons' allocations relative to X, Y could be preferred to X. Similarly, a social planner may think X is preferred to Z and Z is preferred to Y, which forms a preference cycle.

If the social planner maximizes a social welfare function which assumes heterogeneous individuals, then it can explain why the three distributions are different. However, it cannot explain why there exists a preference cycle.

To accommodate the two intuitions, this paper proposes redistribution regret to model the social planner's preferences over distributions of resources. It assumes that the social planner cares about people's feelings which depend both on the current allocation and on the alternative allocation they might have got had the social planner chosen differently. Additionally, the social planner admits that people are different.

Redistribution regret is an extension and application of regret theory, independently proposed by Bell [2] and by Loomes and Sugden[24]. It suggests that the utility from an outcome depends not only on the outcome itself but also on the outcome that could have been obtained had the decision maker chosen differently. Let $\psi(x, y)$ measure the regret or elation a person feels when observing that he won x while the

alternative choice would give him y. It satisfies skew symmetric in the sense that $\psi(x,y) = -\psi(y,x)$ and monotonicity in the sense that $\frac{\partial\psi(x,y)}{\partial x} > 0$ and $\frac{\partial\psi(x,y)}{\partial y} < 0$. It may also satisfy regret aversion: $\psi(x,y) + \psi(y,z) < \psi(x,z)$ for every $x > y > z$. For $X = (x_1, s_1; ...; x_n, s_n)$ and $Y = (y_1, s_1; ...; y_n, s_n)$ with the same set of states $(s_1, ..., s_n)$, regret theory suggests that

$$X \succsim Y \Longleftrightarrow \sum p_i \psi(x_i, y_i) \geq 0$$

As regret theory was developed to explain preference reversals, naturally, it predicts violations of transitivity, see Bikhchandani and Segal [3]. Formally, if the order

$$x \succsim y \Leftrightarrow \sum p_i \psi(x_i, y_i) \geq 0$$

is transitive, then

$$\psi(x,y) = u(x) - u(y)$$

which is just expected utility.

Although redistribution regret and regret theory are similar in the key assumption, they have significant differences. First, "regret" in redistribution regret is produced by individuals and aggregated by a social planner, who is the decision maker. Second, redistribution regret assumes that different individuals can have different regret functions, which is the case that the agent has different regret functions under different states in regret theory.

Most decision theories assume transitivity to guarantee an optimized decision. Due to pairwise comparisons, redistribution regret may predict a violation of transitivity as regret theory. However, as different individuals can have different regret functions, redistribution regret may also be different. This paper studies the property of transitivity of redistribution regret.

The results show that redistribution regret is generally non-transitive. However, there are some exceptions. First, when the total resource is fixed, if there is one very sensitive or insensitive person with the others following expected utility, then the social planner's preference order is transitive. Second, when there are only two persons in the society, and they have the same special individual regret function, then no matter whether the total resource is fixed or not, the preference order could be transitive.

The next section offers preliminary definitions. Section 3 and 4 propose theorems. And the last section concludes.

2.2 Preliminaries

Suppose there are $n \geq 2$ individuals in the society. Each individual uses his individual regret function $\psi(x, y)$ to measure regret (or elation) if he gets x instead of y. Here I define non-linear individual regret function.

Definition 2.1 *The non-linear individual regret function $\psi(x, y)$ satisfies:*

1. *Skew symmetric: $\psi(x, y) = -\psi(y, x)$; it also implies that $\psi(x, x) = 0$.*
2. *Monotonicity: $\frac{\partial \psi(x,y)}{\partial x} > 0$ and $\frac{\partial \psi(x,y)}{\partial y} < 0$.*
3. *Normalization: $\psi(m, 0) = 1$, where $m > 0$ is a fixed number of resource.*
4. *Non-linearity: for some $x > y > z \geq 0$, $\psi(x, y) + \psi(y, z) \neq \psi(x, z)$.*

Skew symmetry says that for $x > y$, the elation one feels from getting x instead of y is quantitatively the same as the regret one feels from getting y instead of x. Monotonicity suggests that elation increases as one's outcome improves, and regret increases with an improvement in the alternative option that was not received. Normalization allows interpersonal comparisons of people's regret or elation. Generally, when we talk about redistribution regret, it refers to non-linear individual regret functions. This condition is more general than "regret aversion". If an individual regret function $\psi(x, y)$ does not satisfy non-linearity, then it is called linear individual regret function, which is basically expected utility theory.

Consider two distributions of resources $X = (x_1, ..., x_n)$ and $Y = (y_1, ..., y_n)$. The social planner decides which distribution is preferred by using a social regret function

$$W(X, Y) = V(\psi_1(x_1, y_1), ..., \psi_n(x_n, y_n))$$

to aggregate individual regrets. This paper assumes that the social regret function is additive, where

$$V(\psi_1(x_1, y_1), ..., \psi_n(x_n, y_n)) = \sum_{i=1}^{n} \psi_i(x_i, y_i)$$

Redistribution regret is saying that

$$X \succsim Y \quad \Leftrightarrow \quad W(X, Y) = V(\psi_1(x_1, y_1), ..., \psi_n(x_n, y_n)) \geq 0$$

Definition 2.2 *Preferences over distributions of resources are transitive if whenever $X \succsim Y$ and $Y \succsim Z$, then $X \succsim Z$.*

To discuss the transitivity of redistribution regret, I consider two factors. First, whether the total resource is fixed or not. Second, how many persons in the society. On the basis of the first factor, the next section considers a fixed budget, and section 4 considers a variable budget. In each section, I discuss different numbers of individuals.

2.3 Fixed Budget

This section assumes that the total resource is fixed as $m = 1$, and individuals' regret functions are normalized by assuming $\psi_i(1,0) = 1$ for each person i.

When the budget is fixed, it is possible to find transitive preference orders. Proposition 2.1 and Proposition 2.2 provide two particular cases.

Proposition 2.1 *Suppose $m = 1$ and $\psi_i(x,y) = x - y$ for $i > 1$. If the non-linear regret function $\psi_1(x,y)$ satisfies*

$$\partial\left(\frac{\psi_1(x,y)}{x-y}\right)/\partial x > 0 \quad \& \quad \partial\left(\frac{\psi_1(x,y)}{x-y}\right)/\partial y > 0$$

or

$$\partial\left(\frac{\psi_1(x,y)}{x-y}\right)/\partial x < 0 \quad \& \quad \partial\left(\frac{\psi_1(x,y)}{x-y}\right)/\partial y < 0$$

for $x \neq y$. Then the preference order

$$(x_1, ..., x_n) \succsim (y_1, ..., y_n) \Longleftrightarrow \sum \psi_i(x_i, y_i) \geq 0$$

is transitive.

Proposition 2.1 says that if the resource is limited and there is only one person who is very sensitive or insensitive to the differences between his outcomes, then the social planner's preference order is the same or opposite to this person's preference order, therefore, it is transitive. The proposition provides a special case where the social planner only needs to consider a special person's preference. This person has to be very sensitive or insensitive, while the others have to follow expected utility theory, which is an extreme case. However, if the social planner does not have complete information and there is such a person or group, then it could be simple and reasonable for the social planner to follow this strategy to make decisions.

Here is an example satisfying the condition mentioned in Proposition 2.1. This function form implies that there is a very insensitive person in the society, so the social preference order is opposite to this insensitive person's preference order.

Example 2.1 *Let total resource* $m = 1$, $\psi_1(x, y) = (x + y - \alpha xy)(x - y)$, *where* $0 < \alpha < 1$, *and* $\psi_i(x, y) = x - y$ *for* $i > 1$, *then the order*

$$(x_1, ..., x_n) \succsim (y_1, ..., y_n) \Longleftrightarrow \sum \psi_i(x_i, y_i) \geq 0$$

is transitive. □

On the basis of Proposition 2.1, Proposition 2.2 provides another transitive case, where there are only two persons in the society, and they have the same non-linear individual regret function as in Proposition 2.1.

Proposition 2.2 *Suppose the total resource* $m = 1$ *and the population* $n = 2$. *If the two individuals have the same non-linear individual regret function* $\psi_1(x, y) = \psi_2(x, y) = \psi(x, y)$, *which satisfies the same condition in Proposition 2.1, then the order*

$$(x_1, x_2) \succsim (y_1, y_2) \Longleftrightarrow \sum_{i=1}^{2} \psi_i(x_i, y_i) \geq 0$$

is transitive.

Proposition 2.1 and Proposition 2.2 provide two transitive cases but with strict conditions. We can expect that these conditions are very delicate as they require strict form of each person's individual regret function, and also the number of persons who have non-linear regret functions. Proposition 2.3 show that if there are more than three persons in the society, and more than two persons have non-linear individual regret functions, the preference order is not transitive.

Proposition 2.3 *Given* $n \geq 3$, *if at least three persons have the same non-linear regret function, then the order*

$$(x_1, ..., x_n) \succsim (y_1, ..., y_n) \Longleftrightarrow \sum \psi_i(x_i, y_i) \geq 0$$

is NOT transitive.

The intuition is that if we shuffle the three persons' outcomes as (x_1, x_2, x_3), (x_2, x_3, x_1), and (x_3, x_1, x_2), while keeping the other persons' outcomes the same, then the three distributions form a non-transitive preference cycle.

Recall that regret theory naturally predicts violations of transitivity even when the expected values of all lotteries are the same. But Proposition 2.1 and Proposition 2.2 show that redistribution regret can be transitive when the total resources of all the

distributions are the same. Therefore, redistribution regret shows different property of transitivity. Meanwhile, the proof in the appendixes imply that the differences are caused by allowing different people have different individual regret functions.

2.4 Variable Budget

Instead of fixed budget, when the total resource is variable, redistribution regret has different properties of transitivity. This section assumes that the total resource m is variable. And normalizes individual regret functions by $\psi_i(1,0) = 1$ for each person i. Additionally, let $lim_{x_i \to \infty} \psi_i(x_i, y_i) = \infty$ and $lim_{y_i \to \infty} \psi_i(x_i, y_i) = -\infty$.

Under these settings, redistribution regret is generally not transitive. Formally, we have Theorem 2.1.

Theorem 2.1 *Suppose the total resource is variable and the population $n \geq 3$. Each person has a general regret function $\psi_i(x, y)$ where $i = 1, ..., n$. If the order*

$$(x_1, ..., x_n) \succ (y_1, ..., y_n) \Longleftrightarrow \sum_{i=1}^{n} \psi_i(x_i, y_i) \geq 0$$

is transitive, then for all $i = 1, ..., n$,

$$\psi_i(x, y) = u_i(x) - u_i(y)$$

which is expected utility.

The proof includes two parts. First, I show that to guarantee transitivity, either everyone follows expected utility theory, or everyone has a non-linear individual regret function. Because if at least one person follows expected utility theory, and the others have non-linear individual regret functions, then it has to violate transitivity. Second, if everyone has a non-linear individual regret function, then I show that we can always find three persons or two persons to form a non-transitive cycle.

Theorem 2.1 says that redistribution regret is generally not transitive when the budget is variable. Example 2.2 shows that whether the budget is fixed or not matters for transitivity. When $m_x = m_y = m_z = 1$, this example is transitive, while it is non-transitive when $m_x \neq m_y \neq m_z$.

Example 2.2 *Let $\psi_1(x, y) = (x + y - 0.5xy)(x - y)$, $\psi_2(x, y) = \psi_3(x, y) = x - y$. Consider the two different situations.*

First, when the budget is fixed, by Proposition 2.1, the social planner's preference order is transitive when $m_x = m_y = m_z = 1$.

Second, when the budget is variable, let $(m_x, m_y, m_z) = (0.876, 0.947, 1)$. *Consider* $(x_1, y_1, z_1) = (0.1, 0.2, 0.3)$, *then we have*

$$W(Y, X) = (0.947 - 0.876) + (0.3 - 0.01 - 1) \times 0.1 = 0$$

$$W(Z, Y) = (1 - 0.947) + (0.5 - 0.03 - 1) \times 0.1 = 0$$

$$W(X, Z) = (0.876 - 1) + (0.4 - 0.015 - 1) \times (-0.2) = -0.001$$

Therefore, $X \sim Y$, $Y \sim Z$ *but* $Z \succ X$, *which is not transitive.* □

Although variable budget generally predicts non-transitivity, when there are only two persons in the society, the results could change, as is shown in Example 2.3.

Example 2.3 *Suppose the total resource is variable and the population* $n = 2$. *If* $\psi_1(x, y) = \psi_2(x, y) = (x - y)^3$, *then the order*

$$(x_1, x_2) \succsim (y_1, y_2) \Longleftrightarrow \sum \psi_i(x_i, y_i) \geq 0$$

is transitive. □

2.5 Conclusions

Redistribution regret compares distributions of resources by applying the idea of regret theory, which assumes that people feel regret or elation when they get one outcome instead of an alternative outcome they might have got had the social planner chosen differently. Additionally, redistribution regret assumes that people have different regret functions, which is different from the homogeneous regret function assumption in regret theory. Due to the different assumptions, regret theory predicts a violation of transitivity in general while redistribution regret tells a different story.

It shows that if the budget is variable and there are more than three persons, then despite the extra flexibility derived from the fact that different people have different regret functions, redistribution regret is still impossible to be transitive. However, If the budget is fixed, and there exists only one very sensitive or insensitive person, then the social planner could follow this person's preference order and redistribution regret is transitive. Additionally, if there are only two homogenous persons in the society, it could also be transitive.

To sum up, redistribution regret provides a modified regret theory which could be transitive, to study a social choice problem. Therefore, it is possible for the social planner to find a best choice and make decisions under those circumstances.

3 Chapter 3: Ranking Regret

3.1 Introduction

A social planner wants to rank individuals. Examples are assigning seats to fans at sporting events, where the rank is seat quality, or assigning a boarding order to airplane passengers. In order to improve passengers' experience, some airlines in the U.S. switched in recent years from boarding according to passenger's assigned seats to group boarding. For example, handicapped passengers, families with infants, etc., may be granted the option to preboard. Suppose an agency knows subjects' preferences over their rank and is trying to generate a queuing process. Moreover, the agency only cares about subjects' feelings, which depend on both their current positions in the queue and the positions they would have been given under an alternative procedure. This paper investigates conditions under which giving priority to individuals according to their sensitivity to the changes in their positions is optimal.

If the agency maximizes a social welfare function based on direct utilities individuals receive from their rank, for example, a weighted sum of such utilities, then a best ranking always exists. However, individuals do not consider only their current positions. They may feel satisfaction or dissatisfaction by comparing their present rank to the alternative rank they might have got had the agency used a different procedure. For example, consider an airline's choice of whether to allow passengers with infants the option of preboarding. While these passengers will appreciate the option, they may feel acute dissatisfaction if they are not offered preboarding precisely because it is an accommodation offered by other airlines. Meanwhile, some other passengers may feel moderate dissatisfaction if their positions become worse due to this preboarding policy. Note that the levels of dissatisfaction of different passengers may be different. In such a case, airlines need to do pairwise comparisons of waiting lines. But there is a serious problem here. Such a process could lead to preference cycles. Therefore, the mentioned boarding procedure may not be the best and there may not necessarily exist a best design for a queue to eliminate aggregate feelings of dissatisfaction.

Consider a social planner who needs to rank individuals. This planner cares about people's feelings and has complete information about them. Assume that individuals know the social planner's choice set and that they are either satisfied or dissatisfied with his choice, where their feelings depend not only on the rank they get but also on the rank they might have got had the social planner used a different queue-generating process. People's feelings are subjective and as in the above example, may vary from

one person to another. The social planner has preferences over rankings based on an aggregation of people's feelings of satisfaction and dissatisfaction. Formally, given two rankings X and Y, we say that Y is socially preferred to X if the aggregation of people's feelings of satisfaction is positive when we switch from X to Y. For simplicity, assume that individuals care only about their rank and differ only in their degree of sensitivity. Given a sequence of sensitivities, a natural ranking would be to have the most sensitive person first, proceeding along with the sensitivity levels to the least sensitive one. Denote this ranking X^*. This paper investigates conditions under which X^* is a best way to rank individuals.

The proposed "ranking regret" of this paper is an extension and application of regret theory, independently proposed by Bell [2] and by Loomes and Sugden [24]. It suggests that the utility from an outcome of a random variable depends not only on the outcome itself but also on the outcome that could have been obtained had the decision-maker chosen another random variable. Let $\psi(x, y)$ measure the regret or elation a person feels when observing that he received x while the alternative choice would have yielded him y. For two random variables $X = (x_1, s_1; ...; x_n, s_n)$ and $Y = (y_1, s_1; ...; y_n, s_n)$, regret theory suggests that

$$X \succsim Y \Longleftrightarrow \sum \Pr(s_i)\psi(x_i, y_i) \geqslant 0$$

Regret theory was initially developed to explain preference reversals (see Lichtenstein and Slovic [21], Lindman [22], Grether and Plott [15]). Recently, it has been modified for use in other situations.[2]

There are several differences between ranking regret and traditional regret theory. Ranking regret replaces "states of nature" with "individuals." In ranking regret, "regret" is felt by individuals but is aggregated by a social planner, who is the decision-maker, while in standard regret theory, regret is felt and aggregated by the same agent. Ranking regret assumes that different individuals can have different regret functions $\psi_i(x_i, y_i)$, while in standard regret theory the same regret function is used for all states. Another technical difference is that the domain of ranking regret is over a finite number of individuals, while the domain in regret theory is a non-finite σ-algebra of events.

[2]Further studies can be found in Loomes and Sugden [25], Sugden [39], Quiggin [33], Starmer and Sugden [35], Hayashi [16], Bleichrod, Cillo and Diecidue [5], Stoye [38], Diecidue and Somasundaram [10], Levy [20]. Other extensions can be found in Braun and Munermann [8], Muermann, Mitchell and Volkman [32], Filiz-Ozba and Ozbay [11], Michenaud and Solnik [29], Maccheron, Marinacci and Rustichini [27], Sarver [34], Bikhchandani and Segal [4], Buturak and Evren [9], and Bleichrodt and Wakker [6].

Most decision theories assume transitivity, as it guarantees a best choice in compact sets. But regret theory predicts violations of transitivity (see Bell [2], Loomes and Sugden [24], see also Bikhchandani and Segal [3]). To find a best choice in a general choice set, Loomes and Sugden [25], Sugden [39], and Quiggin [33] suggest assuming transitivity in each feasible set, which changes the basic assumption of pairwise comparisons. This may be necessary for regret theory, but as I show in this paper, ranking regret may accommodate a best choice without such an assumption.

Like standard regret theory, ranking regret too can lead to preference cycles, where the social planner may find out that there is a positive aggregate satisfaction when ranking X is replaced with Y, positive aggregate satisfaction when Y is replaced by Z, and again a positive aggregation of satisfaction when Z is replaced with X. However, since different individuals have different sensitivity functions, the social planner may find a socially optimal ranking. Although the social preference order may not be transitive, this ranking may nonetheless be optimal, if it is not involved in any preference cycle. For example, preferences over $\{X, Y, Z\}$ may form a non-transitive cycle, yet X^*, which orders individuals by their sensitivities, may be better than all of them. This paper outlines conditions under which X^* is best, even if these conditions do not eliminate violations of transitivity. Once the best ranking is chosen, the social planner will not have any incentive to switch to another ranking.

This paper provides an algorithm for telling whether X^* is the best ranking or not. If people have sufficiently different levels of sensitivities to changes in their positions from one possible ranking to another, then X^* is best; otherwise, it is not. The paper analyzes the optimality of X^* when the number of people in the relevant group grows. The larger the number, the lower are the levels of critical values of sensitivities are needed for X^* to be optimal. Moreover, as the number of individuals grows, the critical values of all individuals converge to the same limit.

Regarding the case where X^* is not best, this paper proposes group ranking to find a second-best choice. In this part, I relax the goal of eliminating aggregate dissatisfaction and let the social planner ignore "small dissatisfaction" people feel relative to others within a group of similarly sensitive individuals. Then the social planner can always achieve his goal by constructing groups appropriately and applying the natural ordering at the group level. Given a sequence of sensitivities and a structure that tells the number of individuals in each group, I provide an algorithm for telling whether the natural ordering under the structure is a best group-ranking. If so, when people are more homogeneous, the intuition is that the maximum number of groups among such structures is smaller. This paper takes airlines boarding queue

as an example and shows that the intuition agrees with the empirical evidence.

The next section offers preliminary definitions. Section 3 proposes theorems about the best ranking. Section 4 discusses the second-best choice — group ranking. Section 5 takes boarding queue as an example to illustrate the intuition behind group ranking. Section 6 discusses several further topics. And the last section concludes.

3.2 Preliminaries

Let $N = \{1, \ldots, n\}$ be a group of n individuals who need to be ranked. Each of these individuals has preferences for a higher rank, but the intensities of these preferences may differ both among individuals and over specific rank.

An individual's absolute rank is an integer number A_i between 1 and n, where a smaller number indicates a higher (hence better) rank. The absolute ranking is a list $A = (A_1, \ldots, A_n)$, which is a permutation of $(1, \ldots, n)$. This setup corresponds to cases where individuals care only about the number of people in front of them, but not about the number of people behind them. Examples that are likely to fit this setup include boarding queues, waiting lines for cashiers or tickets, etc.

Let $R_i = A_i/n$ be the relative rank of person i. The relative rank of the top-ranked individual is $\frac{1}{n}$, while that of the bottom-ranked person is 1. The relative ranking is a list $R = (R_1, \ldots, R_n)$, which is a permutation of $(\frac{1}{n}, \ldots, 1)$. This ranking fits cases when people care about their relative positions in the line, such as where a larger number of individuals behind them makes people feel better. For example, when it comes to income distributions, people care about their income, but also about their relative positions in society. In this paper, I discuss only the ranking but not the actual income.

Given two rankings $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$, individual i evaluates his feelings by comparing his current rank x_i to the alternative rank y_i, using an individual function $\psi_i(x_i, y_i)$. This function measures the satisfaction or dissatisfaction felt by individual i from receiving x_i, knowing that the alternative is y_i.

Definition 3.1 *The individual functions $\psi(x, y)$ satisfy the following properties.*

1. *Skew symmetry: $\psi(x, y) = -\psi(y, x)$.*
2. *Monotonicity: $\frac{\partial \psi(x,y)}{\partial x} < 0$ and $\frac{\partial \psi(x,y)}{\partial y} > 0$.*
3. *Regret aversion: for all $0 < x < y < z$, $\psi(x, y) + \psi(y, z) \leqslant \psi(x, z)$, and for some $0 < x < y < z$, $\psi(x, y) + \psi(y, z) < \psi(x, z)$.*

4. Normalization: $\psi(1,n) = 1$ *for absolute ranks or* $\psi(\frac{1}{n},1) = 1$ *for relative ranks.*

Skew symmetry says that for $x < y$, the satisfaction one feels from getting x instead of y is quantitatively the same as the dissatisfaction one feels from getting y instead of x. (Recall that a lower value of x indicates a higher ranking). Monotonicity suggests that satisfaction increases as one's outcome improves, and dissatisfaction increases with an improvement in the alternative option that was not received. According to regret aversion, individuals do not pay much attention to a small difference between x and y, while they pay much more attention to a large difference. Normalization allows interpersonal comparisons of people's regret or elation.

A social planner wants to compare two rankings X and Y. Assume a social aggregation function of the form

$$W(X,Y) = V(\psi_1(x_1,y_1),\ldots,\psi_n(x_n,y_n))$$

The social planner prefers X to Y, denoted $X \succsim Y$, iff:

$$W(X,Y) \geqslant 0$$

If the individual regret functions are $\psi_i(x_i,y_i) = u_i(x_i) - u_i(y_i)$, then $W(X,Y) = \sum_{i=1}^{n}[u_i(x_i) - u_i(y_i)]$ and $X \succsim Y$ iff $\sum_{i=1}^{n} u_i(x_i) \geqslant \sum_{i=1}^{n} u_i(y_i)$. In other words, the utilitarian social welfare function is a special case of our model, where the function u_i is interpreted as the utility person i obtains from the received rank. This utility does not depend on the alternative rank this person might have reached.

Definition 3.2 *Preferences over rankings are transitive if whenever* $X \succsim Y$ *and* $Y \succsim Z$*, then* $X \succsim Z$*.*

Definition 3.3 *Ranking* X^* *is defined as a best choice in the preference relation if* $X^* \succsim Y$ *for all* $Y \neq X^*$*.*

Kreps [19] shows that transitivity and best choice are two different concepts. Transitivity implies the existence of a best choice on a finite domain. However, a best choice may still exist even if there are preference cycles as long as the preference cycles do not involve the best choice. This paper shows that aggregations of ranking regret are rarely transitive, but under some conditions, they still admit a best choice.

3.3 Best Choice

To simplify people's characteristics and further guarantee the validity of interpersonal comparisons, consider a set Ψ of non-linear continuous functions of the form $\psi = \psi(x, y, \alpha)$, uniquely defined by one parameter α, where a smaller value of α indicates a higher degree of sensitivity: if $\alpha_i < \alpha_j$, then for all $x < y$, $\psi(x, y, \alpha_i) > \psi(x, y, \alpha_j)$. In other words, $\partial\psi(x, y, \alpha_i)/\partial\alpha_i < 0$ for all $x < y$. The parameter α can be thought of as a measure of sensitivity, and can be interpreted in two ways.

First, individuals have the same preferences regarding rank, while the social planner assigns them different "sensitivity levels," which may depend on some objective, observable criteria. For example, airlines can observe whether passengers are disabled, traveling with children, or are pregnant. The US government may recognize that some asylum seekers are in greater danger than others (and should therefore prioritize their cases). Under this interpretation, the parameter α captures the social planner's view of the world.

Alternatively, individuals have different (observable) degrees of sensitivity due to some explicit and implicit reasons, such as personality, health condition, education, income, etc. In this case, the source of the social planner's differential treatment is the heterogeneity of the relevant population and not his views regarding the different types.

Regardless of the interpretation, whether α is assigned by the social planner or is part of the individuals' characteristics, this paper regards these parameters as given.

Given $n \geqslant 3$ individuals with a sequence of sensitivities $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_n$, consider the absolute ranking $X^A = (1, 2, \ldots, n)$ and the relative ranking $X^R = (\frac{1}{n}, \frac{2}{n}, \ldots, 1)$, giving a smaller number, which is a higher rank, to a more sensitive individual. I will use X^* for both X^A and X^R hereafter, and the related context applies to both the absolute and the relative definitions.

If the social planner maximizes an aggregation of individuals' utilities from their rank, X^* would intuitively be the best way to rank individuals. However, the social planner of this paper evaluates the aggregation of people's feelings from pairwise comparisons, which depend not only on the rank they get but also on the alternative rank they might have got had the social planner chosen differently. The social planer may find out that there is a positive aggregate satisfaction when ranking X is replaced with Y, positive aggregate satisfaction when Y is replaced by Z, and again a positive aggregation of satisfaction when Z is replaced with X. Therefore, there may exist preference cycles. Example 3.1 shows that ranking regret could be transitive in some cases, while it could also have preference cycles.

Example 3.1 Let $\psi_i(x,y) = sgn(y-x)\,|y-x|^{\alpha_i}$ and $W(X,Y) = V(\psi_1(x_1,y_1), \ldots, \psi_n(x_n,y_n)) = \sum_i \psi_i(x_i,y_i)$. Consider $n=3$ and relative ranking. If the sequence of sensitivities is $(\alpha_1,\alpha_2,\alpha_3) = (2,3,5)$, then the preference order over rankings is transitive, and $X^R = \left(\frac{1}{3},\frac{2}{3},1\right)$ is the best way to rank individuals. If $(\alpha_1,\alpha_2,\alpha_3) = (2,3,4)$, then it is not transitive, and X^R is not the best way as $\left(1,\frac{1}{3},\frac{2}{3}\right) \succ X^R \succ \left(\frac{2}{3},\frac{1}{3},1\right) \succ \left(1,\frac{1}{3},\frac{2}{3}\right)$. □

The following claim shows that for $n=3$ and either relative or absolute ranking, if the least sensitive individual is sufficiently non-sensitive, then the preference order over the six possible rankings is transitive.

Claim 3.1 *Let $n=3$. For any $\alpha_2 \geqslant \alpha_1 > 0$ there is $\alpha_3^* \geqslant \alpha_2$ such that for all $\alpha_3 > \alpha_3^*$, the preference order $X \succsim Y$ iff $W(X,Y) \geqslant 0$ is transitive.*

As noted above, transitive orders over a finite set admit a best element, but the existence of a best element does not imply transitivity. The rest of this section discusses the more general conditions for the existence of a best choice.

Suppose that the individual levels of sensitivities are $\alpha_1 < \ldots < \alpha_n$. If each person is significantly less sensitive than the person before him, that is, if for all i, α_{i+1} is significantly larger than α_i, then it is reasonable to expect that the ranking X^* is a best choice for the social planner. Theorem 3.1 below formalizes these intuitive claims, and moreover, offers an exact algorithm to tell whether X^* is a best choice given the sequence $\{\alpha_1,\ldots,\alpha_n\}$. Formally, it shows how to create a sequence $\{\alpha_1^*,\ldots,\alpha_n^*\}$ such that if for all i, $\alpha_i \geqslant \alpha_i^*$, then X^* is a best choice, but if for even one person i, $\alpha_i < \alpha_i^*$, then X^* is not best choice.

Theorem 3.1 *Given individual functions $\psi \in \Psi$ and a sequence of sensitivities $\{\alpha_1,\ldots,\alpha_n\}$, where $\alpha_1 < \ldots < \alpha_n$, there are critical value functions*

$$\alpha_i^* = f_i(\alpha_1,\ldots,\alpha_{i-1})$$

such that X^* is a best ranking if and only if $\alpha_i \geqslant \alpha_i^*$ for $i = \{3,\ldots,n\}$. Moreover, the functions $f_3,\ldots,f_n$ can be explicitly constructed given the profile of individual functions $\psi_1,\ldots,\psi_n$ and the profile of sensitivities $\alpha_1,\ldots,\alpha_n$.

Individual i's critical value function $f_i(\alpha_1,\ldots,\alpha_{i-1})$ is implied by the form of individual functions and the values of actual sensitivities. A specific description of the function form is provided in the Appendix. Note that the critical value of

sensitivity of person i depends only on the actual sensitivities of persons $\{1,\ldots,i-1\}$. Its calculation needs neither the critical sensitivities of persons $\{1,\ldots,i-1\}$, nor any information about the actual sensitivities of persons $\{i,\ldots,n\}$. If the actual sensitivity of person i is smaller than the critical value, then it is not necessary to evaluate the sensitivities of the persons behind person i.

Theorem 1 is proved in the Appendix, but I provide here an intuitive explanation of it. For simplicity, the explanation assumes the relative definition, but the same intuition applies to the absolute definition as well. Let Ω be the set of all possible rankings, and $\mathcal{Y}$ be the set $\Omega \setminus X^*$. There are $n!-1$ elements in $\mathcal{Y}$. Let $\mathcal{Y}_k = \{X : x_k \neq x_k^* \text{ and } \forall i > k, x_i = x_i^*\}$, $k = 2,\ldots,n$. There are $k! - (k-1)!$ elements in $\mathcal{Y}_k$. Then $\{\mathcal{Y}_2,\ldots,\mathcal{Y}_n\}$ are disjoint sets and $\mathcal{Y} = \mathcal{Y}_2 \cup \mathcal{Y}_3 \cup \ldots \cup \mathcal{Y}_n$. We need to compare X^* to some critical rankings[3] in each set by the method of induction. If $k = 2$, there is one element, $Y_2^1 = (\frac{2}{n}, \frac{1}{n}, \frac{3}{n}, \ldots, 1)$, in $\mathcal{Y}_2$. The social planner prefers X^* to Y_2^1 as person 1 is more sensitive than person 2. If $k = 3$, there are $3! - 2! = 4$ alternative rankings in $\mathcal{Y}_3$. I show that one can find α_3^* such that $X \succ Y_3^i$ for any $Y_3^i \in \mathcal{Y}_3$ if $\alpha_3 > \alpha_3^*$. I then show, by induction, that one can find α_i^* such that X^* is preferred to any ranking in $\mathcal{Y}_i$ provided $\alpha_i > \alpha_i^*$. Therefore, ranking X^* is a best choice if we can create a sequence $\{\alpha_3^*,\ldots,\alpha_n^*\}$ and $\alpha_i > \alpha_i^*$ for $i \in \{3,\ldots,n\}$.

Theorem 3.1 implies that X^* may not be best when people are too similar. Consider a case in which a social planner needs to decide whether or not to allow some individuals to cut in line. Sometimes, a social planner may permit an individual, who is not the most sensitive person, to cut in line and take the top position, because the social planner believes that the more sensitive individuals only get a little worse due to a small negative change in their positions while the less sensitive individual improves a lot owing to a large positive change. As a result, the ranking after cutting in line is preferred to the natural ranking X^*.

Theorem 3.1 offers an algorithm for identifying whether X^* is the best choice given the above sequence of sensitivities.

Step 1 : set $\alpha_1^* = \alpha_1$.

Step 2 : set $\alpha_2^* = \alpha_2$.

Step 3 : calculate $\alpha_3^* = f_3(\alpha_1, \alpha_2)$ and compare α_3 to α_3^*. If $\alpha_3 < \alpha_3^*$, then there is no best choice. If $\alpha_3 \geqslant \alpha_3^*$, continue to the next step.

[3]The way to find the critical rankings is shown in the proof together with the way to find the critical value functions.

Step $k = 3, \ldots, n$: By now, $\alpha_i \geqslant \alpha_i^*$ for $i \in \{3, \ldots, k-1\}$. Calculate $\alpha_k^* = f_k(\alpha_1, \ldots, \alpha_{k-1})$ and compare α_k to α_k^*. If $\alpha_k < \alpha_k^*$, there is no best choice. If $\alpha_k \geqslant \alpha_k^*$, continue to the next step, or terminate if $k = n$, and X^* is a best choice. □

To illustrate the above theorem and algorithm, consider the individual functions $\psi_i(x, y) = sgn(y - x)\,|y - x|^{\alpha_i}$ and use the relative definitions.

Example 3.2 Given six individuals, define $X^* = \left(\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, \frac{6}{6}\right)$. The sequence of sensitivities is given by the top part of Table 3.1.

To calculate α_3^*, consider the ranking set $\mathcal{Y}_3 = \{Y_3^1, Y_3^2, Y_3^3, Y_3^4\}$, where

$$Y_3^1 = \left(\frac{1}{6}, \frac{3}{6}, \frac{2}{6}, \frac{4}{6}, \frac{5}{6}, \frac{6}{6}\right)$$

$$Y_3^2 = \left(\frac{3}{6}, \frac{2}{6}, \frac{1}{6}, \frac{4}{6}, \frac{5}{6}, \frac{6}{6}\right)$$

$$Y_3^3 = \left(\frac{2}{6}, \frac{3}{6}, \frac{1}{6}, \frac{4}{6}, \frac{5}{6}, \frac{6}{6}\right)$$

$$Y_3^4 = \left(\frac{3}{6}, \frac{1}{6}, \frac{2}{6}, \frac{4}{6}, \frac{5}{6}, \frac{6}{6}\right)$$

Note that $\left(\frac{2}{6}, \frac{1}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, \frac{6}{6}\right)$ is not included in $\mathcal{Y}_3$ as it belongs to $\mathcal{Y}_2$. Recall that $X^* \succsim Y_3^1$ if $W(X^*, Y_3^1) \geqslant 0$; and $Y_3^1 \succ X^*$ if $W(X^*, Y_3^1) < 0$. Let $W(X^*, Y_3^1) = 0$, solve α_3, and define it as α_3^1. Observe that $W(X^*, Y_3^1)$ could always be positive given $\alpha_3 \geqslant \alpha_2$. If so, let $\alpha_3^1 = \alpha_2$. Similarly, compute α_3^2, α_3^3 and α_3^4. As the value of $W(X^*, Y_3^1)$ increases in α_3, we have $\alpha_3^* = \max\{\alpha_3^1, \alpha_3^2, \alpha_3^3, \alpha_3^4, \alpha_2\}$.

Following the rest of the steps of the algorithm, we can get α_4^*, α_5^* and α_6^*, which are shown in Table 3.1. Observe that X^* is a best choice as $\alpha_k > \alpha_k^*$ for $k \in \{3, 4, 5, 6\}$. Note that we can still find preference cycles which do not include X^*, e.g., $Y_1 \succ Y_2 \succ Y_3 \succ Y_1$, where $Y_1 = \left(\frac{6}{6}, \frac{5}{6}, \frac{3}{6}, \frac{4}{6}, \frac{2}{6}, \frac{1}{6}\right)$, $Y_2 = \left(\frac{6}{6}, \frac{5}{6}, \frac{4}{6}, \frac{3}{6}, \frac{1}{6}, \frac{2}{6}\right)$, and $Y_3 = \left(\frac{6}{6}, \frac{5}{6}, \frac{4}{6}, \frac{2}{6}, \frac{3}{6}, \frac{1}{6}\right)$. Because $W(Y_1, Y_2) = 0.0185$, $W(Y_2, Y_3) = 0.0219$, and $W(Y_3, Y_1) = 0.0616$.

Table 3.1: X^* is A Best Choice

α_1	α_2	α_3	α_4	α_5	α_6
1.1	1.2	1.4410	1.7300	2.4588	5
α_1^*	α_2^*	α_3^*	α_4^*	α_5^*	α_6^*
1.1	1.2	1.3410	1.6300	2.3588	4.9268

However, if the sequence of actual sensitivities is given in Table 3.2, following the steps in the algorithm shows there is no best choice, since $\alpha_5 < \alpha_5^*$, and therefore, the value of α_6^* does not exist. Specifically, we can find preference cycles which include X^*, e.g., $X^* \succ Y_4 \succ Y_5 \succ X^*$, where $Y_4 = \left(\frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{1}{6}, \frac{5}{6}, \frac{6}{6}\right)$ and $Y_5 = \left(\frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, \frac{1}{6}, \frac{6}{6}\right)$. Because $W(X^*, Y_4) = 0.0300$, $W(Y_4, Y_5) = 0.0860$, and $W(X^*, Y_5) = -0.0333$.

Table 3.2: No Best Choice

α_1	α_2	α_3	α_4	α_5	α_6
1.1	1.2	1.4410	1.7300	2.2	6

□

If the social planner cares about people's feelings, which depend both on their current rank and their alternative rank, then X^* may not be the best choice as the preference order could be non-transitive. Despite the existence of preference cycles, Theorem 3.1 shows that under some conditions, X^* could still be a best choice when X^* is not part of the non-transitive cycle.

Absolute definitions and relative definitions have different implications on the relationship between rank and population, therefore, the relationship between critical values and population. Regarding absolute definitions, rank is not correlated with population. For example, the individual, who takes the i^{th} position, always has a rank equals to i regardless of the change of population. Then the critical value functions of absolute ranking do not depend on population. However, if we assume relative definitions, rank change with population. For example, if there are n persons, the i^{th} individual has a rank $\frac{i}{n}$, which decreases with population. Then the critical value functions could depend on the size of population.

In the case of relative definitions, Theorem 3.2 studies how the critical value α^* changes with the size of the population n. The theorem assumes an additive social aggregation function and some less general individual functions, $\psi(y - x, \alpha)$ instead of the functions, $\psi(x, y, \alpha_i)$ used in Theorem 3.1. Functions like $\psi(x, y, \alpha) = (y - x)(x + y - \alpha xy)$ where $\alpha \in (0, 1)$ are thus excluded.

Consider an alternative ranking Y_k^j, where only the individuals from $k - j$ to k have different rank from X. Specifically, individual k moves upward j steps and individuals from $k - j$ to $k - 1$ move downward one step. Given the individual functions $\psi(y - x, \alpha)$, define α_k^j by letting $W(X, Y_k^j) = 0$. We have

$$\psi(\frac{j}{n}, \alpha_k^j) = \sum_{i=k-j}^{k-1} \psi(\frac{1}{n}, \alpha_i)$$

Denote $\psi_1(y-x,\alpha)=\frac{\partial\psi(y-x,\alpha)}{\partial(y-x)}$. Define $\bar{\alpha}_{k,j}$ by

$$\psi_1(\frac{1}{n},\bar{\alpha}_{k,j})=\frac{1}{j}\sum_{i=k-j}^{k-1}\psi_1(\frac{1}{n},\alpha_i)$$

Theorem 3.2 *Given social aggregate function W(X,Y)=$\sum_i\psi(x_i,y_i)$ and individual functions $\psi(y-x,\alpha_i)$, consider an infinite sequence of individuals with descending sensitivities. For each sequence of the first n individuals, calculate $(\alpha^*_{n,3},\ldots,\alpha^*_{n,n})$ as if there are only n individuals in society. For any $x<y$, let the individual functions $\psi(y-x,\alpha_i)$ satisfy*

$$\psi_1(\frac{j}{n},\alpha_k^j)\geqslant\psi_1(\frac{1}{n},\bar{\alpha}_{k,j})$$

for each $j\in\{1,\ldots,k-1\}$. Then $\alpha^*_{n,k}$ is weakly decreasing in n for any $k\in\{3,\ldots,n\}$. Moreover, there exists k for which $\alpha^*_{n,k}$ is not constant in n.

Let $\alpha^*_{n,i}=f_i(\alpha_1,\alpha_2,\alpha^*_{n,3},\ldots,\alpha^*_{n,i-1},n)$. Then for $k=3,\ldots,n$, $\lim_{n\to\infty}\alpha^*_{n,i}=\alpha_2$.

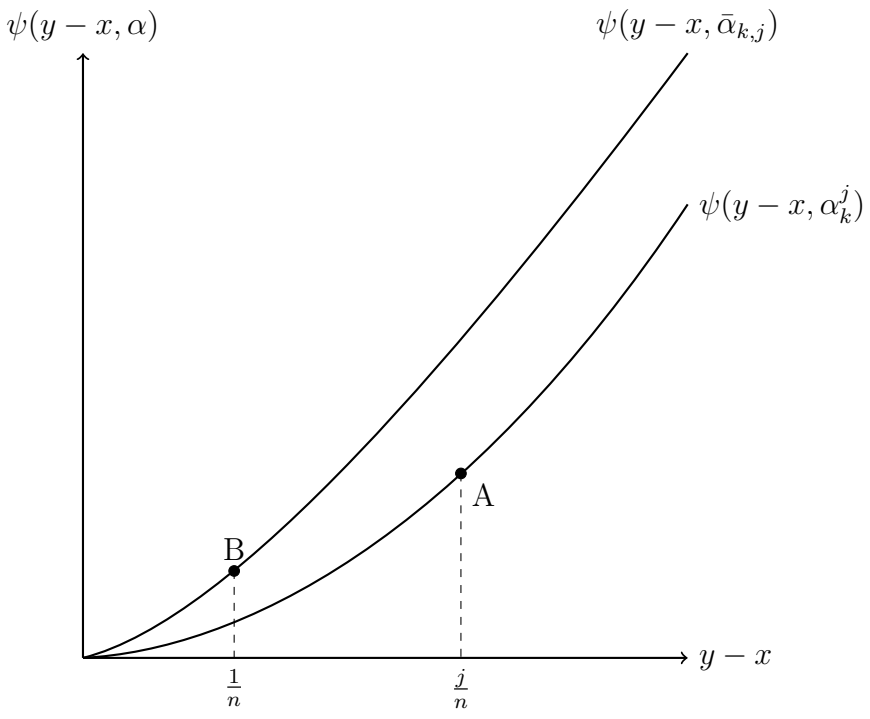

Figure 3.1: Theorem 3.2

The graphs of the functions $\psi(y-x,\alpha_k^j)$ and $\psi(y-x,\bar{\alpha}_{k,j})$ are depicted in Figure 3.1. The assumption says that the slope of $\psi(y-x,\alpha_k^j)$ when $y-x=\frac{j}{n}$ (point A) is

greater than the slope of $\psi(y-x,\bar{\alpha}_{k,j})$ when $y-x=\frac{1}{n}$ (point B).[4]

Theorem 3.2 implies that if there are many people, the conditions required to guarantee a best choice are weaker than those we obtained in Theorem 3.1. Although it applies to a smaller domain than Theorem 3.1, the set of individual functions satisfying the assumption is not empty. For example, $\psi(y-x,\alpha)=sgn(y-x)\,|y-x|^{\alpha}$.[5]

3.4 Second-Best Choice: Group Ranking

Theorem 3.1 implies that if people are too similar, there does not exist a best choice. That is, for any ranking X there is an alternative ranking Y such that $W(X,Y)\geqslant 0$. This is not an ideal outcome, as the social planner cannot optimize.

Consider ten individuals. The first five have similar degrees of sensitivity while the latter five have another similar level of sensitivity. By Theorem 3.1, there does not exist a best choice. However, the social planner may still prefer to put the five sensitive individuals before the other five less sensitive individuals, although it may not be clear how to rank individuals within each of these groups.

To accommodate this intuition, this section discusses how to find the second-best choice where the social planner can divide people into ranked groups. Specifically, the social planner needs to make two decisions: to choose the best ranking given a structure of groups and to choose the optimal structure of groups.

3.4.1 Choice of Ranking

Definition 3.4 *Consider a set of n individuals. A structure of their partition into groups is a vector $\nu=[n_1,\ldots,n_I]$ such that $\sum_{i=1}^{I}n_i=n$. In this partition, n_1 of the individuals are in the first group and have the top n_1 positions, and for $i=2,\ldots,I$, n_i of the individuals are in group i and take the positions from $\sum_{j=1}^{i-1}n_j+1$ to $\sum_{j=1}^{i-1}n_j+n_i$.*

All individuals in group i consider their rank to be the average rank of their group. For example, for the case of absolute ranking, the rank of each person in group i is the average rank of the "first" and "last" person in this group, that is, $\frac{1}{2}[(\sum_{j=1}^{i-1}n_j+1)+\sum_{j=1}^{i}n_j]$. Based on ν, define $N=[N_1,N_2,\ldots,N_I]$, where N_j is the total number of people in the first j groups. That is, $N_j=\sum_{i=1}^{j}n_i$ for $j=1,\ldots,I$.

[4]Given $\psi(\frac{j}{n},\alpha_k^j)=\sum_{i=k-j}^{k-1}\psi(\frac{1}{n},\alpha_i)$, we have $\psi(\frac{1}{n},\alpha_k^1)=\psi(\frac{1}{n},\alpha_{k-1})$ when $j=1$. By monotonicity, $\alpha_k^1=\alpha_{k-1}$. In that case, the two curves for $\psi(y-x,\alpha_{k-1})$ and $\psi(y-x,\alpha_k^1)$ in Figure 3.1 are the same, and $\frac{j}{n}=\frac{1}{n}$, therefore, the two slopes of A and B are equal. The assumption is satisfied. Note that point A does not have to be above point B.

[5]That the assumption is satisfied, is proved in the Appendix.

Rank the n individuals $\{1, \ldots, n\}$ in a descending order of sensitivity, assume absolute ranking, and define

$$X_G^A = (\underbrace{0.5(n_1+1), \ldots}_{\text{Individuals 1 to } N_1}, \underbrace{N_1 + 0.5(n_2+1), \ldots}_{N_1+1 \text{ to } N_2}, \ldots, \underbrace{N_{I-1} + 0.5(n_I+1), \ldots}_{N_{I-1}+1 \text{ to } N_I})$$

It specifies a rank for each group and gives a higher rank to groups with more sensitive people. Similarly, in the case of relative ranking, we have $X_G^R = \frac{1}{n} X_G^A$.

Consider the case where ν is given, thus, the alternative rankings only re-order individuals without changing the structure of the groups. Based on this assumption, I define a ranking to be best if it is preferred to any alternative rankings which re-order individuals into the same structure of groups. Similarly to Theorem 3.1, Theorem 3.1* describes an algorithm that can be used to determine whether given ν, the ranking X_G^A or X_G^R (hereafter X_G^*) is a best way to rank people and allocate them into the given group structure.

Theorem 3.1* *Consider individual functions $\psi(x, y, \alpha)$ and n individuals with a sequence of sensitivities $\{\alpha_1, \ldots, \alpha_n\}$ where $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_n$. Given the structure of groups $\nu = [n_1, \ldots, n_I]$, there are $I - 1$ functions f_k, $k = 2, \ldots, I$, where*

$$\alpha^{k*} = f_k(\alpha_{N_1}, \ldots, \alpha_{N_{k-1}}, n_1, \ldots, n_k)$$

such that X_G^* is the best ranking iff $\alpha_{N_{k-1}+1} > \alpha^{k*}$ for $k = \{2, \ldots, I\}$.

Theorem 3.1* is similar to Theorem 3.1. It also leads to a similar algorithm, but the critical value functions are different. First, we only need to consider the critical value of the most sensitive individual in group k. If this individual has a degree of sensitivity $\alpha_{N_{k-1}+1}$ smaller than the critical value α^{k*}, then X_G^* is not best. If it is greater, then by the order of individuals, the entire group must satisfy the critical condition. Second, when we calculate the critical value for group k, instead of considering all individuals in earlier groups, we only need to consider the least sensitive person in each of the first $k - 1$ groups. The specific functional form is provided in the appendix.

By Theorem 3.1*, given a structure of groups, we can tell whether there exists a best ranking or not, and if a best ranking exists, we can also tell what it is. Here I describe the choice of ranking given a structure.

Definition 3.5 *A structure ν is stable if it leads to a best choice which is preferred to any other rankings with the same structure.*

Note that stability is with respect to a given structure, which means that all rankings in the choice set have the same structure. Given a ranking X_G, the alternative rankings can have a different order of individuals but not a different order of groups.

Claim 3.2 *If structure $\nu = [n_1, \ldots, n_I]$ is stable, then the best choice is X_G^* which is*

$$X_G^A = (\underbrace{0.5(n_1+1), \ldots}_{\textit{Individuals 1 to } N_1}, \underbrace{N_1 + 0.5(n_2+1), \ldots}_{N_1+1 \textit{ to } N_2}, \ldots, \underbrace{N_{I-1} + 0.5(n_I+1), \ldots}_{N_{I-1}+1 \textit{ to } N_I})$$

or $X_G^R = \frac{1}{n} X_G^A$.

Claim 3.2 says that given a stable structure, the ranking X_G^* which puts more sensitive individuals into a group with a higher rank is the best choice. That is, for any stable structure, the social planner knows the corresponding best choice. Thus, we can define a stable structure as its best ranking.

3.4.2 Choice of Structure

The last section discusses the best choice of ranking given a structure. But sometimes the structure is not given, instead, the social planner needs to design one. This section studies how to find the choice of structure.

Claim 3.3 *For any sequence of $n \geqslant 3$ individuals, there are at least n stable structures. One of them includes one group and the other $n-1$ stable structures include two groups.*

The social planner would not like to have an unstable structure since in such a case he will not be able to optimize. By Claim 3.3, there are more than one stable structures, and the social planner needs to choose from the stable structures. The criteria depend on the social planner. It can be the stable structure with the largest number of groups, or evenly distributed groups, or the one which implies a best group ranking among all stable structures. Here I discuss only the last criterion.

Consider n individuals with their sensitivities. Find all stable structures $\nu^S = \{\nu_1, \ldots, \nu_m\}$ and their corresponding best rankings $X^S = \{X_{\nu_1}, \ldots, X_{\nu_m}\}$.

Definition 3.6 *Let $\nu \in \nu^S$ be a stable structure. It is an optimal stable structure if there is no structure $\nu' \in \nu^S$ such that $X_{\nu'} \succ X_\nu$.*

Definition 3.7 *Let $\nu \in \nu^S$ be a non-optimal stable structure. It is an inferior stable structure if for every structure $\nu' \in \nu^S$ such that $X_{\nu'} \succ X_\nu$, there are no structures $(\nu_1, \ldots, \nu_\ell) \in \nu^S$ such that $X_\nu \succ X_{\nu_1} \succ \ldots \succ X_{\nu_\ell} \succ X_{\nu'}$.*

Note that the social planner uses pairwise comparisons to evaluate stable structures, which are represented by their corresponding best rankings. As a result, there may not exist an optimal stable structure. In this case, the second-best choice is a set of stable but non-inferior structures, and the social planner can choose any one of them. Example 3.3 explains the second-best choice and shows that it can include more than one structure.

Example 3.3 Assume that the individual functions are $\psi_i(x,y) = sgn(y-x)\,|y-x|^{\alpha_i}$ and the social aggregation function is additive. Suppose five individuals with a sequence of sensitivities

$$\alpha = (1.01, 1.02, 1.03, 1.031, 2)$$

By Theorem 1*, there are seven stable structures of groups:

$$\nu_1 = [3,1,1]; \quad \nu_2 = [1,3,1]; \quad \nu_3 = [4,1]; \quad \nu_4 = [1,4]$$

$$\nu_5 = [2,2,1]; \quad \nu_6 = [2,3]; \quad \nu_7 = [3,2]$$

Comparing the best rankings of all stable structures, we have the following relationship:

$$(X_{\nu_1}, X_{\nu_2}, X_{\nu_3}, X_{\nu_5}) \succ X_{\nu_7} \succ X_{\nu_6} \succ X_{\nu_4}$$

$$X_{\nu_1} \succ X_{\nu_5} \succ X_{\nu_2} \succ X_{\nu_3} \succ X_{\nu_1}$$

Therefore, $\{\nu_4, \nu_6, \nu_7\}$ are inferior structures, and the social planner's second-best choice is the set $\{\nu_1, \nu_2, \nu_3, \nu_5\}$, which is not a unique structure. □

Suppose the social planner always puts the most sensitive individual into the first group (or if there are several identical such individuals, he puts all of them into the first group), and the second most sensitive individual(s) into the second group. Given a sequence of sensitivities, the numbers of individuals in the first two groups are thus fixed. In this case, if there are at least three different individuals, we cannot use any of the stable structures mentioned in Claim 3.3. For any other structure, given some sequences of sensitivities, Claim 3.4 shows that it could be the optimal stable structure. It implies that no structures should be excluded before analyzing the sequence of sensitivities.

Claim 3.4 *Consider a structure $\nu = [n_1, \ldots, n_I]$ and suppose that the social aggregation function is additive. Suppose further that the individual functions $\psi(y-x,\alpha)$ are strictly convex in $y-x$ and satisfy $\psi_1(y-x,\alpha) < \psi_1(z-x,\alpha')$, where α' is given*

by $\psi(y-x,\alpha)+\psi(z-y,\alpha)=\psi(z-x,\alpha')$ *for* $0<x<y<z$. *Then there is a sequence of sensitivities* α^* *such that* ν *is the optimal stable structure.*[6]

3.5 Application: Boarding Queues

Section 3.4 tells how to verify whether a structure is stable or not and specifies the best choice given a stable structure. Although the results there do not mention a particular way to find an optimal stable structure, they imply that an optimal stable structure should have fewer groups when people are more homogeneous. This section considers airline boarding queues as an empirical example.

Profit maximization for airlines includes both minimizing costs and improving passengers' experience. Efficiency during boarding reduces airlines' costs. Some papers study theoretically optimal boarding orders for minimizing boarding times (See Steffen [36] , Milne and Kelly [30], Milne and Salari [31]). However, in practice, most airlines do not adopt these optimal boarding strategies. One reason may be that airlines consider their passengers' feelings about the boarding experience to be more important than efficiency.

To provide passengers a good boarding experience, airlines must consider passengers' sensitivities about boarding orders. This paper's results suggest that airlines, in the role of social planners, will divide people into groups that are likely to have similar sensitivities, and give boarding priority to the more sensitive groups. A further prediction of the theory is that the more similar the passengers on a flight are to each other (in terms of sensitivity), the fewer boarding groups we should expect. Is this what airlines do in practice?

In the US, most airlines give boarding priority to frequent flyers, first-class passengers, passengers who pay extra for early boarding, passengers with disabilities, families with children, and military personnel. Some of these passenger types are explicitly revealing greater sensitivity (e.g., those who pay for priority boarding), while other early groups seem likely to be sensitive (e.g., those buying first-class tickets and passengers with disabilities). It seems reasonable to conclude that airlines do consider sensitivity to boarding order in their allocation of passengers into boarding groups.

Now consider the number of boarding groups. There is considerable variation across airlines in the number of groups, even for planes of similar sizes. For example, Frontier Airlines has just one main boarding group, American Airlines uses as many as ten.

[6]In the proof, I find α^* such that there is no other stable structure $[n_1, n_2, n_3', \ldots, n_J']$ given α^*. So ν is the only stable structure, therefore, the optimal stable structure.

Is the number of groups related to heterogeneity in sensitivity? It is reasonable to assume that, ceteris paribus, sensitivity positively correlates with the fare one pays. If so, then the previously derived theory suggests that, after controlling for factors like route and number of seats on the plane, flights with a wider range of fares charged to passengers should have more boarding groups than flights with more homogeneous fares. This implies that the coefficient α in the following regression should be positive.

$$group = \alpha \cdot range + \beta \cdot capacity + \gamma \cdot X$$

Data were collected from eight US airlines (Alaska, American, Delta, Frontier, JetBlue, Southwest, Spirit, and United) on the 43 most popular routes in the United States. The raw data includes round-trip ticket prices and airplane capacity for each flight.[7] Constructing averages for each airline on each route, there were 194 airline/route combinations when only non-stop flights are considered, and 275 observations when both non-stop and one-stop routes are included.

Averages are constructed for each airline on each route. The dependent variable *group* is the number of boarding groups as reported on the airlines' official websites. This number is a low integer, so the model is estimated using Poisson regression. The variable *range* is a measure of the range of fares charged by the airline on that route, *capacity* is the average number of seats per airplane flown by the airline on that route, and, depending on the specification, X is either just a constant or a constant plus a set of dummy variables for each route.

Four different measures of the variable *range* are considered: $range_1$ is the difference between the maximum and minimum ticket prices divided by the mean; $range_2$ is the difference between maximum and minimum divided by the median; $range_3$ is the standard deviation of ticket prices divided by the mean; $range_4$ is the interquartile range of ticket prices divided by the median.

In addition to the above model, estimates are also provided for the alternative specification:

$$group/cap = \alpha \cdot range + \beta \cdot X$$

where *group/cap* is *group* divided by *capacity*. This regression is estimated by Ordinary Least Squares. Summary statistics of all the data are provided in Table 3.3.

For each of the two model specifications above, a total of 16 different estimates

[7]The data were collected during June 6 – June 8, 2020, for flights scheduled from August 3 to August 7, 2020.

Table 3.3: Statistics

Variables	Obs	Mean	Max	Min	Std
	Non-stop & One-stop				
group	275	7.1345	10	4	2.2065
$range_1$	275	0.9891	2.1308	0	0.5678
$range_2$	275	1.1741	3.6692	0	0.8279
$range_3$	275	0.4200	0.8480	0	0.2348
$range_4$	275	0.5896	1.8346	0	0.4218
capacity	275	155.3382	230	72.5	31.5967
group/cap	275	0.0490	0.1379	0.0174	0.0217
	Non-stop				
group	194	7.1546	10	4	2.1967
$range_1$	194	1.0063	1.9093	0	0.5823
$range_2$	194	1.1811	3.6692	0	0.8279
$range_3$	194	0.4247	0.8101	0	0.2416
$range_4$	194	0.5959	1.8346	0	0.4332
capacity	194	157.1262	230	0.0174	33.9506
group/cap	194	0.0492	0.1379	72.5	0.0230

of α are reported. Using each of the four different measures of *range*, regression R_1 includes *route* dummies and uses both non-stop and one-stop flights; R_2 omits route dummies and uses both non-stop and one-stop flights; R_3 includes route dummies and only uses non-stop flights; and R_4 omits the route dummies and only uses non-stop flights. All 16 estimates for the Poisson regression model of *group* are shown in Table 3.4, while those for the OLS model of *group/cap* are shown in Table 3.5. Standard errors are in parentheses.

Table 3.4: Poisson Regressions

	R_1	R_2	R_3	R_4
$range_1$	0.2482^{***}	0.2229^{***}	0.2514^{***}	0.2245^{***}
	(0.0310)	(0.0284)	(0.0348)	(0.0313)
$range_2$	0.2201^{***}	0.1835^{***}	0.2332^{***}	0.2020^{***}
	(0.0195)	(0.0168)	(0.0216)	(0.0176)
$range_3$	0.5107^{***}	0.4546^{***}	0.5355^{***}	0.4690^{***}
	(0.0762)	(0.0675)	(0.0859)	(0.0746)
$range_4$	0.3079^{***}	0.2619^{***}	0.3480^{***}	0.2966^{***}
	(0.0396)	(0.0337)	(0.0451)	(0.0372)

To interpret the range of the magnitudes of these α estimates, in Table 3.4, at the mean of the data the number of groups is around 7, and the estimated coefficients imply that it would take roughly 20% to 25% increase in range to increase the number

of groups to 8. A comparable but larger range of responses is implied by the estimates in Table 3.5,

All the estimates of α in all the specifications are positive and statistically significant, implying that airlines do employ more groups on flights where the range of fares is larger, consistent with the theory that the planner should divide passengers into a greater number of groups when there is greater variation in sensitivity to boarding order.

Table 3.5: OLS Regressions

	R_1	R_2	R_3	R_4
$range_1$	0.0171^{***}	0.0156^{***}	0.0176^{***}	0.0161^{***}
	(0.0022)	(0.0021)	(0.0027)	(0.0026)
$range_2$	0.0139^{***}	0.0116^{***}	0.0149^{***}	0.0130^{***}
	(0.0014)	(0.0014)	(0.0018)	(0.0017)
$range_3$	0.0379^{***}	0.0332^{***}	0.0395^{***}	0.0349^{***}
	(0.0054)	(0.0052)	(0.0067)	(0.0064)
$range_4$	0.0233^{***}	0.0190^{***}	0.0253^{***}	0.0210^{***}
	(0.0030)	(0.0029)	(0.0037)	(0.0035)

3.6 Discussions

3.6.1 Best Choice and Transitivity

As mentioned above, transitivity implies the existence of best choice, but best choice may exist even if transitivity is not satisfied. That is, preference cycles and best choice can exist at the same time, as long as the best choice is not involved in any preference cycle. The following example illustrates a situation where there exists a best choice and at least one preference cycle.

Example 3.4 Suppose $n = 4$, $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (1.5, 2, 2.5150, 5.3933)$, and individual functions are $\psi(x, y, \alpha_i) = sgn(y - x)|y - x|^{\alpha_i}$. There are 24 relative rankings in total $\{X^*, Y_1, \ldots, Y_{23}\}$. Here we only consider four of them $X^* = (\frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1)$, $Y_1 = (\frac{1}{4}, \frac{3}{4}, \frac{2}{4}, 1)$, $Y_2 = (\frac{1}{4}, \frac{2}{4}, 1, \frac{3}{4})$, $Y_3 = (\frac{2}{4}, \frac{1}{4}, \frac{3}{4}, 1)$.

Comparing X^* to Y_i, where $i \in \{1, \ldots, 23\}$, we have $W(X^*, Y_i) > 0$. Therefore, X^* is a best choice. Meanwhile, we have $W(Y_1, Y_2) = 0.1119$, $W(Y_2, Y_3) = 0.0325$ and $W(Y_3, Y_1) = 0.0944$. Hence, we have at least one preference cycle $Y_1 \succ Y_2 \succ Y_3 \succ Y_1$ while X^* is a best choice. □

3.6.2 Regret Theory and Ranking Regret

The violation of transitivity is an issue both in regret theory and in ranking regret as both preferences are based on pairwise comparisons. However, there is an important difference between the two, which is shown in Example 3.1. Ranking regret is compatible with transitive orders, while in regret theory there are always non-transitive cycles. The reason for this difference is that ranking regret aggregates different individual functions while in regret theory there is only one regret function, which is used across all states.

Note that the domain in regret theory is a non-finite σ-algebra of events, and the domain of ranking regret is a finite set of individual preferences. To prove some theorems related to regret theory it is common to randomly rewrite the outcomes of events in a convenient way. However, to prove the theorems in this paper, I cannot rewrite the weight of individuals; therefore, the current paper could not rely on formal results from regret theory.

3.6.3 Ranking Regret and Income Inequality

Ranking regret studies situations where the social planner wants to rank individuals. To a certain extent, this resembles the analysis of income inequality. However, the income inequality problem considers both rank and income level. Different sequences of income can generate the same ranking but individuals may not be indifferent between them as one sequence yields different income levels. Therefore, ranking regret that only considers rankings cannot fully describe income inequality. For this, we should evaluate individuals' feelings in two dimensions: rank and income.

3.7 Conclusions

This paper proposes ranking regret. It studies the best way for a social planner to rank individuals under the assumption that individuals' feelings depend not only on their current rank but also on the alternative rank, which they might have got had the social planner chosen differently. Naturally, giving higher rank to more sensitive individuals, denoted by X^*, seems to be the best choice for the social planner. For example, our usual notions of a planner maximizing some social welfare function, i.e., some function of individual utilities, gives full preference orderings and shows that X^* is best. However, under ranking regret, we have binary preferences but not necessarily a full ordering. As a result, intransitivity and hence cycles can arise,

and when they do, the planner may not be able to achieve an optimizing goal like eliminating aggregate dissatisfaction, and therefore, X^* may not be best.

This paper provides an algorithm to tell whether X^* is a best choice or not, and if so, what conditions guarantee it. It shows that if individuals are very different, then X^* is a best choice; if people are too similar, a best choice may not exist. Regarding the latter case, this paper discusses group ranking to find a second-best choice, which is a set of non-inferior stable structures. The structures specify the number of individuals in each group, and the second-best ranking gives a higher rank to a more sensitive group. Intuitively and empirically, the second-best choice has fewer groups if people are more homogeneous.

Ranking regret provides a possibility of the co-existence of best choice and non-transitivity. It indicates that the social planner can still make decisions in a general choice set even though pairwise comparisons are assumed and the preference orders are not transitive. Additionally, even if there does not exist a best choice, group ranking still provides a way to rebuild the choice set and yield a second-best choice.

4 Chapter 4: Over-Identified Doubly Robust Identification and Estimation

4.1 Introduction

Consider two different parametric models, which we will call G and H. One of these models is correctly specified, but we don't know which one (or both could be right). Both models include the same parameter vector α. An estimator $\widehat{\alpha}$ is called *Doubly Robust* (DR) if $\widehat{\alpha}$ is consistent no matter which model is correct. The term double robustness was coined by Robins, Rotnitzky, and van der Laan (2000), but is based on Scharfstein, Rotnitzky, and Robins (1999) and the augmented inverse probability weighting average treatment effect estimator introduced by Robins, Rotnitzky, and Zhao (1994). In their application α is a population Average Treatment Effect (ATE).

We provide a general technique for constructing doubly robust (DR) estimators. The main requirements for applying our method is that models G and H each be characterized by a set of moment conditions, and each is over identified. We therefore call our method Over-identified Doubly Robust (ODR) estimation. Our ODR takes the form of a weighted average of Hansen's (1982) Generalized Method of Moments (GMM) based estimates of α, and has similar root-n asymptotics to GMM.

The main drawback of existing DR estimators is that they are not generic, meaning that for each problem, one needs to find a DR estimator, which can then be used only for that one specific application. No general method exists for finding or constructing DR estimators, and only a few examples of such models are known in the literature. Perhaps the closest thing to a general method is Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2018). These authors derive a set of locally robust estimators, provide a characterization result showing when these estimators will also be DR and thereby provide some new examples of constructing DR estimators.[8] In contrast, our ODR provides a simple general method of constructing DR estimators for a very wide class of models.

Most existing applications of DR methods, like ATE estimation, have models G and H that are exactly identified rather than overidentified. In such cases, it

[8]Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2018) also show that their DR estimators possess some additional useful asymptotic properties that the ODR estimators we construct may not possess. Ideally, some different terminology would distinguish between estimators that just have the DR property (including ours and theirs) vs. estimators that have the additional properties, including local robustness, that they document.

may be possible to add additional overidentifying moments, and thereby apply our ODR (e.g., in a online supplemental appendix, we provide details for doing so in the ATE application). However, we do not advise using our ODR for applications where DR methods already exist, particularly when existing DR methods do not require overidentification. Instead, the main virtue of our ODR is its widespread potential application to situations where there are *not* already existing DR estimators. We provide some examples in section 4.3 below.

Suppose we have data consisting of n observations of a random vector Z. Assume that the true value of α satisfies either $E\left[G\left(Z,\alpha,\beta\right)\right]=0$ or $E\left[H\left(Z,\alpha,\gamma\right)\right]=0$ (or both) for some known vector valued functions G and H, and some unknown additional parameter vectors β and γ. Our ODR estimator then consistently estimates α, despite not knowing which of these two sets of equalities actually holds, for any G and H that satisfy some regularity and identification conditions.

Consider three different possible estimators for the vector α, called $\widehat{\alpha}_g$, $\widehat{\alpha}_h$, and $\widehat{\alpha}_f$. The estimator $\widehat{\alpha}_g$ is a GMM estimator of α that is asymptotically efficient if just the model G is correctly specified, i.e., if $E\left[G\left(Z,\alpha_0,\beta_0\right)\right]=0$ at the true α_0 and β_0. Similarly, let $\widehat{\alpha}_h$ be an asymptotically efficient GMM estimator if $E\left[H\left(Z,\alpha_0,\gamma_0\right)\right]=0$, and let $\widehat{\alpha}_f$ be a GMM estimator based on both sets of moments, which would be asymptotically efficient if both sets of moments hold at α_0, β_0, and γ_0.

One possible approach to estimation of α would be to engage in some form of model selection. Under our assumptions, model selection would be relatively straightforward. However, model selection has some disadvantages relative to DR methods, e.g., one needs to correct limiting distributions for pretest bias, and tests for which model is superior can be inconclusive. In the context of GMM based models, selection methods like Andrews and Lu (2001), Caner (2009), and Liao (2013) use test-based methods or shrinkage penalties to select moments that are most likely to be valid.

Another alternative would be model averaging, which is generally not consistent unless both G and H happen to be correctly specified. Like DR, our ODR avoids these issues. However, our ODR estimator does take the form of a weighted average of $\widehat{\alpha}_g$, $\widehat{\alpha}_h$, and $\widehat{\alpha}_f$, and so closely resembles GMM model averaging. A number of model averaging estimators exist for GMM and related models. Kuersteiner and Okui (2010) apply Hansen's (2007) model averaging criterion for instruments in linear instrumental variables models. Averaging across instruments or moments in GMM models is also considered by Martins and Gabriel (2014), Sueishi (2013), and DiTraglia (2016). Unlike these papers, we do not use typical model averaging criteria like mean squared error, Bayes weights, or information criteria to choose weights. Instead, we construct

weights to yield the DR consistency property and for relative efficiency.

In the next section, we describe our ODR estimator. Section 3 then gives examples of potential applications of our ODR estimator (additional examples, including showing how existing DR applications could have alternatively been estimated using our ODR, are provided in an online supplemental appendix). In section 4 we show consistency and provide limiting distribution theory for our ODR. Section 5 provides Monte Carlo simulations and Section 6 gives an empirical application. In Section 7 we analyze properties of our estimator when the models G and H may be locally misspecified, i.e., where the parameter α_0 in the data generating process is replaced with $\alpha_0 + \delta n^{-s}$ for a constant δ and some $s > 0$. Section 8 considers extensions to more than two competing models, and Section 9 concludes. Proofs and additional results are provided in the Appendices.

4.2 The ODR Estimator

Let Z be a vector of observed random variables, let α, β and γ be vectors of parameters, and assume G and H are known functions. Assume a sample consisting of n independent, identically distributed (iid) observations z_i of the vector Z.[9] The goal is root-n consistent, asymptotically normal estimation of α. Let α_0 denote the true value of α. Define model G to be 'correct,'or 'true,' if $E\left[G\left(Z, \alpha_0, \beta_0\right)\right] = 0$ for some unique β_0. Similarly, define model H to be true if $E\left[H\left(Z, \alpha_0, \gamma_0\right)\right] = 0$ for some unique γ_0. Define model F to consist of both sets of moments, and model F is true if both models G and H are true.

As discussed in the introduction, we begin with three different possible estimators for the vector α, called $\widehat{\alpha}_g$, $\widehat{\alpha}_h$, and $\widehat{\alpha}_f$. The estimator $\widehat{\alpha}_g$ is a GMM estimator of α that would be asymptotically efficient if model G is true and model H is not true. Specifically, $\widehat{\alpha}_g$ (along with $\widehat{\beta}_g$) minimizes the Hansen (1982) two-step quadratic GMM objective function, which we will call $\widetilde{Q}^g(\alpha, \beta)$. This $\widehat{\alpha}_g$ will generally be inconsistent if G is not true. If model G is true, then $n\widetilde{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$ is asymptotically chi-squared. But more importantly for us, if model G is true then $\widetilde{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$ itself will converge to zero in probability, and (under our assumptions) not converge to zero otherwise. We use this property to construct our ODR estimator.

Analogous to $\widehat{\alpha}_g$, let $\widehat{\alpha}_h$ denote the estimator of α based on the moments $E\left[H\left(Z, \alpha_0, \gamma_0\right)\right] = 0$, so $\widehat{\alpha}_h$ and $\widehat{\gamma}_h$ minimize a quadratic GMM objective function $\widetilde{Q}^h(\alpha, \gamma)$, and are

[9] We assue iid data mainly for convenience. Our ODR is a straightforward generalization of GMM, so it should be applicable under more general conditions. We mainly require that the GMM estimators and associated objective functions satisfy some standard properties.

asymptotically efficient if model H is true and model G is not true. Finally, let $\widehat{\alpha}_f$ be the GMM estimator of α based on assuming both sets of moments $E\left[G\left(Z,\alpha_0,\beta_0\right)\right]=0$ and $E\left[H\left(Z,\alpha_0,\gamma_0\right)\right]=0$ hold. This $\widehat{\alpha}_f$ along with $\widehat{\beta}_f$ and $\widehat{\gamma}_f$ minimizes a GMM objective function $\widetilde{Q}^f(\alpha,\beta,\gamma)$, and is asymptotically efficient (generally more efficient than either $\widetilde{Q}^g$ or $\widetilde{Q}^h$) if both models G and H are true, but will otherwise generally be inconsistent.

Our proposed ODR estimator is a weighted average of $\widehat{\alpha}_g$, $\widehat{\alpha}_h$, and $\widehat{\alpha}_f$, taking the form

$$\widehat{\alpha}=\hat{W}_f\hat{W}_g\widehat{\alpha}_h+\hat{W}_f\left(1-\hat{W}_g\right)\widehat{\alpha}_g+(1-\hat{W}_f)\widehat{\alpha}_f \tag{4.1}$$

The novelty in our estimator relative to existing model averaging estimators is in the construction of the weights $\hat{W}_g$ and $\hat{W}_f$, given below in equations (4.3) and (4.5). In particular, we construct these weights so that, asymptotically, $\widehat{\alpha}$ becomes arbitrarily close to $\widehat{\alpha}_f$ if both models G and H are true, and otherwise becomes arbitrarily close to either $\widehat{\alpha}_g$ or $\widehat{\alpha}_h$, depending on which model is true. So, instead of the typical model averaging criteria such as minimizing mean squared error, we assume at least one of the models is correctly specified, and choose weights for efficiency, while satisfying the DR criterion.

4.2.1 Starting Assumptions

Let $g_0(\alpha,\beta)\equiv E\{G(Z,\alpha,\beta)\}$, $h_0(\alpha,\gamma)\equiv E\{H(Z,\alpha,\gamma)\}$, $\theta_0\equiv\{\alpha_0,\beta_0,\gamma_0\}$, and $\theta\equiv\{\alpha,\beta,\gamma\}$.

Assumption A1: For compact sets Θ_α, Θ_β, and Θ_γ, $\alpha_0\in\Theta_\alpha$, $\beta_0\in\Theta_\beta$, and $\gamma_0\in\Theta_\gamma$. Let $\Theta=\Theta_\alpha\times\Theta_\beta\times\Theta_\gamma$.

Assumption A2: Either 1) $g_0(\alpha_0,\beta_0)=0$, or 2) $h_0(\alpha_0,\gamma_0)=0$, or both hold.

Assumption A2 says that, for some unknown true coefficient values α_0, β_0, and γ_0, either model G is true, or model H is true, or both are true. This is a defining feature of DR estimators, and hence of our ODR estimator.

Assumption A3: The vector $G(Z,\alpha,\beta)$ has more elements than the set of elements in α and β. The vector $H(Z,\alpha,\gamma)$ has more elements than the set of elements in α and γ. For any $\{\alpha,\beta,\gamma\}\in\Theta$, if $g_0(\alpha,\beta)=0$ then $\{\alpha,\beta\}=\{\alpha_0,\beta_0\}$, and if $h_0(\alpha,\gamma)=0$ then $\{\alpha,\gamma\}=\{\alpha_0,\gamma_0\}$.

Assumptions A2 and A3 are identification assumptions. They imply that if G is the true model, then the true values of the coefficients $\{\alpha_0,\beta_0\}$ are identified by

$g_0(\alpha_0, \beta_0) = 0$, and if H is the true model, then the true values of the coefficients $\{\alpha_0, \gamma_0\}$ are identified by $h_0(\alpha_0, \gamma_0) = 0$. Assumption A3 rules out the existence of alternative pseudo-true values satisfying the 'wrong' moments, e.g., this assumption rules out having both $g_0(\alpha_0, \beta_0) = 0$ and $g_0(\alpha_1, \beta_1) = 0$ for some $\alpha_1 \neq \alpha_0$.

Note that Assumption A3 is a potentially strong restriction, and is not required by other DR estimators. Satisfying this assumption essentially implies that models G and H are each over identified. The first part of Assumption A3 is typically necessary to satisfy the second part, since if G contained the same number of elements as the set $\{\alpha, \beta\}$, then the equation $g_0(\alpha, \beta) = 0$ would have as many equations as unknowns, and so typically a pseudo-true solution α_1, β_1 would exist satisfying $g_0(\alpha_1, \beta_1) = 0$ even if G were misspecified.

Define the following functions:

$$\widehat{g}(\alpha, \beta) \equiv \frac{1}{n} \sum_{i=1}^{n} G(Z_i, \alpha, \beta), \qquad \widehat{h}(\alpha, \gamma) \equiv \frac{1}{n} \sum_{i=1}^{n} H(Z_i, \alpha, \gamma),$$

$$\widetilde{Q}^g(\alpha, \beta) \equiv \widehat{g}(\alpha, \beta)' \hat{\Omega}_g \widehat{g}(\alpha, \beta), \qquad \widetilde{Q}^h(\alpha, \gamma) \equiv \widehat{h}(\alpha, \gamma)' \hat{\Omega}_h \widehat{h}(\alpha, \gamma),$$

where $\hat{\Omega}_g$ and $\hat{\Omega}_h$ are estimates of the usual weighting matrices obtained in two step GMM, which under correct specification yields asymptotic efficiency of GMM. In the above definition, $\widetilde{Q}^g(\alpha, \beta)$ is the standard Hansen (1982) and Hansen and Singleton (1982) Generalized Method of Moments (GMM) objective function, which the GMM estimator minimizes to estimate α and β. Similarly, minimizing $\widetilde{Q}^h(\alpha, \gamma)$ is the standard GMM estimator for model H. Define $\widehat{\alpha}_g$, $\widehat{\beta}_g$, $\widehat{\alpha}_h$, and $\widehat{\gamma}_h$ by

$$\{\widehat{\alpha}_g, \widehat{\beta}_g\} = \arg \min_{\{\alpha,\beta\} \in \Theta_\alpha \times \Theta_\beta} \widetilde{Q}^g(\alpha, \beta) \quad \text{and} \quad \{\widehat{\alpha}_h, \widehat{\gamma}_h\} = \arg \min_{\{\alpha,\gamma\} \in \Theta_\alpha \times \Theta_\gamma} \widetilde{Q}^h(\alpha, \gamma). \tag{4.2}$$

So $\{\widehat{\alpha}_g, \widehat{\beta}_g\}$ is the standard GMM estimate of model G, and $\{\widehat{\alpha}_h, \widehat{\gamma}_h\}$ is the standard GMM estimate of model H. In our applications, we likewise use the standard efficient two step GMM method for estimating the matrices $\hat{\Omega}_g$ and $\hat{\Omega}_h$.

Define $\widetilde{Q}_0^g(\alpha, \beta)$ and $\widetilde{Q}_0^h(\alpha, \gamma)$ by

$$\widetilde{Q}_0^g(\alpha, \beta) \equiv g_0(\alpha, \beta)' \Omega_g g_0(\alpha, \beta) \quad \text{and} \quad \widetilde{Q}_0^h(\alpha, \gamma) \equiv h_0(\alpha, \gamma)' \Omega_h h_0(\alpha, \gamma)$$

for positive definite matrices Ω_g and Ω_h, where $\hat{\Omega}_g \to^p \Omega_g$ and $\hat{\Omega}_h \to^p \Omega_h$.

Assumption A4: Assume there exists $\{\alpha_g, \beta_g\} \in \Theta_\alpha \times \Theta_\beta$ such that $\widetilde{Q}_0^g(\alpha_g, \beta_g) < \widetilde{Q}_0^g(\alpha, \beta)$ for all $\{\alpha, \beta\} \in \Theta_\alpha \times \Theta_\beta \backslash \{\alpha_g, \beta_g\}$ and there exists $\{\alpha_h, \gamma_h\} \in \Theta_\alpha \times \Theta_\gamma$ such

that $\widetilde{Q}_0^h(\alpha_h, \gamma_h) < \widetilde{Q}_0^h(\alpha, \gamma)$ for all $\{\alpha, \gamma\} \in \Theta_\alpha \times \Theta_\gamma \backslash \{\alpha_h, \gamma_h\}$.

Given Assumptions A2 and A3, Assumption A4 will automatically be satisfied for model G when G is correctly specified, with $\{\alpha_g, \beta_g\} = \{\alpha_0, \beta_0\}$, and similarly for $\{\alpha_h, \gamma_h\}$ when H is correctly specified, by Lemma 2.3 of Newey and McFadden (1994). Together with Assumptions A1 to A3, Assumption A4 implies that GMM estimators of G or H will also converge to unique values (pseudo-true values) when they are misspecified. Assumption A4 is also imposed by Hall (2000) and Hall and Inoue (2003) for misspecified GMM models.

Our main reason for having Assumption A4 is to ensure that the weights $\hat{W}_f$ and $\hat{W}_g$ are asymptotically well behaved, which simplifies derivation of limiting distributions (and asymptotics under local misspecification). However, some of our results (like consistency of the SODR estimator defined below) will not require Assumption A4.

4.2.2 The SODR and ODR estimators

Let $c_g \equiv g_0(\alpha_g, \beta_g)$. Under minimal, standard regularity conditions (see details in the next section), we have $\widetilde{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) \to^p c_g' \Omega_g c_g$. If G is correctly specified, then $\alpha_g = \alpha_0$ and $\beta_g = \beta_0$, which makes $c_g = 0$, so $c_g' \Omega_g c_g = 0$. What is important for our ODR estimator is that the probability limit of $\widetilde{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$ is zero if G is correctly specified, and positive otherwise.

Having G correctly specified also means (again with minimal regularity), that $n^{1/2}\widehat{g}(\widehat{\alpha}_g, \widehat{\beta}_g)\Omega_g^{1/2} \to_d N\left(0, I_{k_g}\right)$ so $n\widetilde{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) \to_d \chi^2_{k_g}$. However, if G is incorrectly specified, then $c_g \neq 0$, so $c_g' \Omega_g c_g > 0$ and $n\widetilde{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$ does not follow the chi-squared distribution asymptotically. Analogous statements hold for model H.

Let $\hat{Q}^g(\alpha, \beta) \equiv \widetilde{Q}^g(\alpha, \beta)/k_g$ and $\hat{Q}^h(\alpha, \gamma) \equiv \widetilde{Q}^h(\alpha, \gamma)/k_h$, where the integer k_g is the degrees of freedom of the chi-squared statistic that $n\widetilde{Q}^g$ converges to if the G model is true. This is the number of moments in G minus the number of elements in α and β, which is positive as discussed earlier. Similarly, k_h is the degrees of freedom of the chi-squared statistic that $n\widetilde{Q}^h$ equals if the H model is true. This scaling by k_g and k_h is not necessary for our estimator, but improves its finite sample performance (see below for details).

Define $\hat{W}_g$ by

$$\hat{W}_g \equiv \frac{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)}{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) + \hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)}. \tag{4.3}$$

From the above derivations, we have that, if G is correctly specified and H is not,

$$\hat{W}_g \to^p \frac{0}{0 + c_h^{'}\Omega_h c_h/k_h} = 0,$$

while if H is correctly specified and and G is not,

$$\hat{W}_g \to^p \frac{c_g^{'}\Omega_g c_g/k_g}{c_g^{'}\Omega_g c_g/k_g + 0} = 1.$$

Before getting to our ODR estimator given by equation (4.1), consider the simpler estimator $\widetilde{\alpha}$ defined by

$$\widetilde{\alpha} = \hat{W}_g \widehat{\alpha}_h + \left(1 - \hat{W}_g\right) \widehat{\alpha}_g. \tag{4.4}$$

So $\widetilde{\alpha}$ is simply a weighted average of the GMM estimates $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$, where the weights are proportional to $\hat{Q}^g$ and $\hat{Q}^h$. We will call $\widetilde{\alpha}$ the SODR (simpler ODR) estimator.

The intuition behind $\widetilde{\alpha}$ is straightforward (the asymptotic statements in this paragraph are proved formally in the next section). Suppose model H is wrong and model G is right, so $E\left[H(Z, \alpha, \gamma)\right] \neq 0$ for any α and γ, and $E\left[G(Z, \alpha_0, \beta_0)\right] = 0$. Then $\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$ goes in probability to zero while the limiting value of $\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$ is nonzero, so $\hat{W}_g$, the weight on $\widehat{\alpha}_h$ in equation (4.4) will go to zero, and $\left(1 - \hat{W}_g\right)$, the weight on $\widehat{\alpha}_g$, will go to one. As a result, $\widetilde{\alpha}$ will have the same probability limit as $\widehat{\alpha}_g$, and since model G is right, this probability limit will be α_0. The same logic applies if model H is right and G is wrong, switching the roles of g and h, and the roles of β and γ. Finally, if both models are right, then $\widetilde{\alpha}$ is just a weighted average of consistent estimators of α_0, and so is consistent no matter what values the weights take on. We therefore obtain the double robustness property that, whichever model is right, $\widetilde{\alpha} \to^p \alpha_0$.[10]

We could have defined the weight $\hat{W}_g$ without scaling each GMM objective function by its degrees of freedom. Asymptotically, the estimator would still be doubly robust. The reason we scale is because, even when a model is correctly specified, in finite samples the greater is the degrees of freedom of a model, the larger its GMM objective function is likely to be. Asymptotically, the mean of $n\widetilde{Q}^g$ converges to k_g when g is correctly specified, and similarly for h. So, by scaling, when both models are correctly specified, both $n\hat{Q}^g$ and $n\hat{Q}^h$ will asymptotically have mean one. Other-

[10]Notice that when both G and H are correctly specified, $\hat{W}_g$ converges to a ratio of correlated chi-squared distributions, not to a constant. Nevertheless, $\widetilde{\alpha}$ is still consistent because $\widetilde{\alpha} = \widehat{\alpha}_g + \left(\widehat{\alpha}_h - \widehat{\alpha}_g\right)\hat{W}_g$, and when both are correctly specified, $\widehat{\alpha}_g \to_p \alpha_0$ and $\widehat{\alpha}_h - \widehat{\alpha}_g \to_p 0$.

wise, if we didn't scale, whichever model has more moments will tend to have a larger GMM objective function, which would then undesirably penalize that model in finite samples.

Although the SODR $\widetilde{\alpha}$ has the desired DR property, it also has two drawbacks. First, when G and H are both correct, the ratio $\hat{W}_g$ converges to a random variable rather than a constant, which complicates the limiting distribution of $\widetilde{\alpha}$. Second, when both G and H are correct, $\widetilde{\alpha}$ may be inefficient, relative to a GMM estimator that efficiently combines the moments from both models.

To address both of these issues, reconsider now the third model F, defined as the union of moments of the models G and H. Specifically, let $F(Z, \alpha, \beta, \gamma)$ be the vector valued function consisting of the union of elements of $G(Z, \alpha, \beta)$ and $H(Z, \alpha, \gamma)$. Then, letting $\widehat{f}(\alpha, \beta, \gamma) \equiv \frac{1}{n}\sum_{i=1}^{n} F(Z_i, \alpha, \beta, \gamma)$, we can define a third GMM estimator

$$\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} = \arg \min_{\{\alpha,\beta,\gamma\} \in \Theta_\alpha \times \Theta_\beta \times \Theta_\gamma} \widetilde{Q}^f(\alpha, \beta, \gamma)$$

where $\widetilde{Q}^f(\alpha, \beta, \gamma) \equiv \widehat{f}(\alpha, \beta, \gamma)' \hat{\Omega}_f \widehat{f}(\alpha, \beta, \gamma)$. This is efficient GMM assuming both specifications are correct, and so uses all the moments from both. If models G and H are correctly specified, then $\widehat{\alpha}_f$ is at least as asymptotically efficient, and generally much more asymptotically efficient, than $\widehat{\alpha}_g$, $\widehat{\alpha}_h$, or $\widetilde{\alpha}$. Let $c_f \equiv f_0(\alpha_f, \beta_f, \gamma_f) \equiv E\{F(Z, \alpha_f, \beta_f, \gamma_f)\}$. Then $\widetilde{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f) \to^p c_f' \Omega_f c_f$, which equals zero if both models G and H are correctly specified, and is positive otherwise.

We again scale by the degrees of freedom (number of moments in F minus number of elements of α, β, and γ), denoted k_f, defining $\hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f) \equiv \widetilde{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)/k_f$. We then define the weight $\hat{W}_f$ by

$$\hat{W}_f \equiv 1 - \frac{1}{n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f) + 1} \tag{4.5}$$

for some τ having $0 < \tau < 1$. Later we discuss selection of the tuning parameter τ, but for consistency we only require that τ lie between zero and one. Our ODR estimator, given by equation (4.1), can be equivalently written as

$$\widehat{\alpha} = \hat{W}_f \widetilde{\alpha} + \left(1 - \hat{W}_f\right) \widehat{\alpha}_f. \tag{4.6}$$

The intuition now is, if both G and H are correctly specified, then $\hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f) \to^p 0$ and $n\hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$ converges in distribution to a chi-squared statistic (divided by its degrees of freedom), which means that $n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$ for $0 < \tau < 1$ converges

in probability to zero. Alternatively, if either G or H is incorrectly specified, then $\hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$ converges in probability to a positive value, so $n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$ diverges to infinity. Therefore, if both G and H are correctly specified then $\hat{W}_f \to^p 0$ and so $\widehat{\alpha}$ has the same limiting value as $\widehat{\alpha}_f$, while if either G or H is incorrectly specified, then $\widehat{\alpha}$ has the same limiting value as $\widetilde{\alpha}$, which as shown earlier has the same limiting value as either $\widehat{\alpha}_g$ or $\widehat{\alpha}_h$, depending on which is correctly specified.

The estimator $\widehat{\alpha}$ therefore, like $\widetilde{\alpha}$, has the desired DR property. We show later that $\widehat{\alpha}$ avoids the asymptotic issues $\widetilde{\alpha}$ has when both G and H are correctly specified, and that $\widehat{\alpha}$ generally performs better than $\widetilde{\alpha}$ in finite samples. This is why $\widehat{\alpha}$ is our preferred ODR estimator. However $\widehat{\alpha}$ has the disadvantages of being a little more complicated to estimate (since it requires estimating the third model F), and it requires selection of a tuning parameter τ.

4.2.3 Tuning Parameters

One tuning parameter is τ, which for consistency can take any value between zero and one. The larger τ is, the less weight is put on $\widehat{\alpha}_f$ in any given sample. So for efficiency, the more likely it is that both models G and H are correct, the smaller one would want τ to be. Based on this observation, a choice of τ that we find works well in Monte Carlo simulations is to let $\tau = 1 - p$, where p is the p-value of the Wald statistic testing the null hypothesis that $\widehat{\alpha}_g = \widehat{\alpha}_h$.[11] [12]

Another potential tuning parameter is as follows. Let Λ be any strictly monotonically increasing function such that $\Lambda(0) = 0$ and $\Lambda(\cdot) \to \infty$ when $\cdot \to \infty$. Then $n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$, $n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$, and $n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$ can be replaced with $\Lambda\left(n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\right)$, $\Lambda\left(n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\right)$, and $\Lambda\left(n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)\right)$ in the definitions of the weights $\hat{W}_g$ and $\hat{W}_f$ in equations (4.3) and (4.5). The main asymptotic properties of the ODR estimator are preserved by any such choice of Λ, but finite sample properties of the estimator might be improved by different choices of the function Λ. For example, $\Lambda(z) = \exp(\lambda z) - 1$ for some $\lambda > 0$ resembles exponential tilting. Equation (4.4) already somewhat resembles Bayesian model averaging, and this choice of Λ would make that resemblance stronger.[13] See e.g., Kim (2002) and Martins and Gabriel

[11] Our derivation of the limiting distribution of $\widehat{\alpha}$ assumes $\tau > 1/2$, however, this restriction is only required to handle cases where $\alpha_g \neq \alpha_h$, and we $\tau = 1 - p$ will asymptotically increase to over $1/2$ in those cases.

[12] Under possible local misspecification, which we consider in section 7 below, choice of τ becomes more complicated, for two reasons. First, under local misspecification, having a random τ can affect the limiting distribution of $\widehat{\alpha}$. And second, for some range of rates of local misspecification parameter drift, a relatively large value of τ is needed to avoid complications in limitation distributions.

[13] A key difference with Bayesian or information based weighting is that we weight model G based

(2014).[14] Another choice for Λ would be a simple power transform $\Lambda(z) = z^{\lambda}$ for $\lambda > 0$. We consider different choices of Λ in our applications. Overall, we found that the exponential Λ works well, though choice of Λ had only modest effects on our monte carlo simulations, and virtually no effect on our empirical estimates.

Finally, we require estimators for the GMM weighting matrices $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$. As discussed later in section 4.3, these are the standard estimated weighting matrices used in two step GMM, but recentered. See in particular equation (4.11).

4.3 ODR Examples

Before proceeding to show consistency and deriving the limiting distribution of the ODR estimator, we present two example applications. Both are new applications for which no existing DR estimators are known. One concerns estimation of preference parameters in consumption Euler equations and asset pricing kernels. The second is alternative sets of instruments for linear model estimation.

In an Online Supplemental Appendix, we provide two additional examples, comparing the requirements of our ODR estimator to existing DR applications. The first discusses average treatment effect estimation, while the second concerns additive regression models.

4.3.1 Preference Parameter Estimates

One of the original applications of GMM estimation, Hansen and Singleton (1982), was the estimation of marginal utility parameters and of pricing kernels. Consider a lifetime utility function of the form

$$u_\tau = E\left(\sum_{t=0}^{T} b^t R_t U\left(C_t, X_t, \rho\right) \mid W_\tau\right)$$

where u_τ is expected discounted lifetime utility in time period τ, b is the subjective rate of time preference, R_t is the time t gross returns from a traded asset, U is the single period utility function, C_t is observable consumption expenditures in time t, X_t is a vector of other observable covariates that affect utility, ρ is a vector of utility parameters, and W_τ is a vector of variables that are observable in time period τ. Maximization of this expected utility function under a lifetime budget constraint

on the model H objective function, and vice versa, instead of weighting each model by its own objective function.

[14] We discuss comparisons of our estimator with Martins and Gabriel (2014) in more detail later, in sections 4.4.4 and 4.5.0.

yields Euler equations of the form

$$E\left(bR_{t+1}\frac{U'\left(C_{t+1},X_{t+1},\rho\right)}{U'\left(C_t,X_t,\rho\right)}-1\mid W_\tau\right)=0 \tag{4.7}$$

where $U'\left(C_t,X_t,\rho\right)$ denotes $\partial U\left(C_t,X_t,\rho\right)/\partial C_t$. If the functional form of U' is known, then this equation provides moments that allow b and ρ to be estimated using GMM. But suppose we have two different possible specifications of U', and we do not know which specification is correct. Then our ODR estimator can be immediately applied, replacing the expression in the inner parentheses in equation (4.7) with $G(Z,\alpha,\beta)$ or $H(Z,\alpha,\gamma)$ to represent the two different specifications. Here α would represent parameters that are the same in either specification, including the subjective rate of time preference b.

To give a specific example, a standard specification of utility is constant relative risk aversion with habit formation, where utility takes the form

$$U\left(C_t,X_t,\rho\right)=\frac{\left[C_t-M\left(X_t\right)\right]^{1-\rho}-1}{1-\rho}$$

where X_t is a vector of lagged values of C_t, the parameter ρ is the coefficient of relative risk aversion, and the function $M\left(X_t\right)$ is the habit function. See, e.g., Campbell and Cochrane (1999) or Chen and Ludvigson (2009). While this general functional form has widespread acceptance and use, there is considerable debate about the correct functional form for M, including whether X_t should include the current value of C_t or just lagged values. See, e.g., the debate about whether habits are internal or external as discussed in the above papers. Rather than take a stand on which habit model is correct, we could estimate the model by ODR.

To illustrate, suppose that with internal habits the function $M\left(X_t\right)$ would be given by $\widetilde{G}\left(X_t,\beta\right)$, where $\widetilde{G}$ is the internal habits functional form. Similarly, suppose with external habits $M\left(X_t\right)$ would be given by $\widetilde{H}\left(X_t,\gamma\right)$ where $\widetilde{H}$ is the external habits specification. Then, based on equation (4.7), we could define $G(Z,\alpha,\beta)$ and $H(Z,\alpha,\gamma)$ by

$$G(Z,\alpha,\beta)=\left(bR_{t+1}\frac{\left(C_{t+1}-\widetilde{G}\left(X_{t+1},\beta\right)\right)^{-\rho}}{\left(C_t-\widetilde{G}\left(X_t,\beta\right)\right)^{-\rho}}-1\right)W_\tau$$

and

$$H(Z,\alpha,\gamma) = \left(bR_{t+1} \frac{\left(C_{t+1} - \widetilde{H}\left(X_{t+1},\gamma\right)\right)^{-\rho}}{\left(C_t - \widetilde{H}\left(X_t,\gamma\right)\right)^{-\rho}} - 1 \right) W_\tau.$$

In this example, we would have $\alpha = (b, \rho)$, and so would consistently estimate the discount rate b and the coefficient of relative risk aversion ρ, no matter which habit model is correct. To satisfy the required overidentification (Assumption A3), we would want W_τ to have more elements than (α, β) and more than (α, γ). This would generally be the case, because the potential information set of consumers at time t is large relative the the number of parameters in the model.

4.3.2 Alternative Sets of Instruments

Consider a parametric model

$$Y = M(W,\alpha) + \epsilon$$

where Y is an outcome, W is a vector of observed covariates, M is a known functional form, α is a vector of parameters to be estimated, and ϵ is an unobserved error term. The errors ϵ may be correlated with W, so to estimate the model we wish to find instruments that are uncorrelated with ϵ. Let R and Q denote two different vectors of observed covariates that are candidate sets of instruments. One may be unsure if either R or Q are valid instrument vectors or not, where validity is defined as being uncorrelated with ϵ.

We may then define model G by $E(\epsilon R) = 0$, so $G(Z,\alpha) = [Y - M(W,\alpha)]R$ and define model H by $E(\epsilon Q) = 0$, so $H(Z,\alpha) = [Y - M(W,\alpha)]Q$. With these definitions we can then immediately apply the ODR estimator. In this case both β and γ are empty, but more generally, the variables R and Q could themselves be functions of covariates and of parameters β and γ, respectively.

A simple example that we consider in our Monte Carlo analysis is where $M(W,\alpha) = \alpha' W$, so the G model consists of the moments $E[(Y - \alpha' W)R] = 0$ and the H model is the moments $E[(Y - \alpha' W)Q] = 0$. The overidentification condition, Assumption A3, is generally satisfied when Q and R each have more elements than W.

Next consider a richer example, which we later empirically apply, based on a model of Lewbel (2012). Suppose $Y = X'\alpha_x + S\alpha_s + \epsilon$, where X is a K-vector of observed exogenous covariates (including a constant term) satisfying $E(\epsilon X) = 0$, and S is an endogenous or mismeasured scalar covariate that is correlated with ϵ. The goal is

estimation of the set of coefficients $\alpha = \{\alpha_x, \alpha_s\}$.

The standard instrumental variables based estimator for this model would consist of finding one or more covariates L such that $E(\epsilon L) = 0$. Then the set of instruments R would be defined by $R = \{X, L\}$. The resulting GMM (or linear two stage least squares) estimator would be based on the moments $E[G(Z, \alpha)] = 0$ where $G(Z, \alpha)$ is given by the stacked vectors

$$G(Z, \alpha) = \left\{ \begin{array}{c} X\left(Y - X'\alpha_x - S\alpha_s\right) \\ L\left(Y - X'\alpha_x - S\alpha_s\right) \end{array} \right\}. \tag{4.8}$$

The main difficulty with applying this two stage least squares or GMM estimator is that one must find one or more covariates L to serve as instruments.

Lewbel (2012) proposes an alternative estimator that, rather than requiring that one find instruments L, instead constructs instruments based on assumptions regarding heteroscedasticity. This estimator consists of first linearly regressing S on X, and obtaining the residuals from that regression. Then a vector of instruments P is constructed by setting P equal to demeaned X (excluding the constant) times these residuals. This constructed vector P is then used instead of L above as instruments.[15] As shown in Lewbel (2012), one set of conditions under which the vector P can be a valid set of instruments is when the endogeneity in S is due to classical measurement error in S.

Let X_c denote the vector X with the constant removed. Algebraically, we can write the instruments obtained in this way as $R = \{X, P\}$ where $P = (X_c - \gamma_1)(S - X'\gamma_2)$, and where the vectors γ_1 and γ_2 in turn satisfy $E(X_c - \gamma_1) = 0$ and $E[X(S - X'\gamma_2)] = 0$. An efficient estimator based on this construction would be standard GMM using the moments $E[H(Z, \alpha, \gamma)] = 0$ where $H(Z, \alpha, \gamma)$ is a vector that consists of the stacked vectors

$$H(Z, \alpha, \gamma) = \left\{ \begin{array}{c} X_c - \gamma_1 \\ X(S - X'\gamma_2) \\ X\left(Y - X'\alpha_x - S\alpha_s\right) \\ (X_c - \gamma_1)(S - X'\gamma_2)\left(Y - X'\alpha_x - S\alpha_s\right) \end{array} \right\}. \tag{4.9}$$

The moments given by $E[G(Z, \alpha)] = 0$ or $E[H(Z, \alpha, \gamma)] = 0$ correspond to two very different sets of identifying conditions. ODR estimation based on these moments therefore allows for consistent estimation of α if either one of these sets of conditions

[15] This estimator is implemented in the STATA module IVREG2H by Baum and Schaffer (2012).

hold. To satisfy the over identification Assumption A3, X_c and L must each have two or more elements.

As a motivating example, consider the following application involving Engel curve estimation (see Lewbel 2008 for a short survey, and references therein). Suppose Y is a consumer's expenditures on food, X is a vector of covariates that affect the consumer's tastes, and S is the consumer's total consumption expenditures (i.e., their total budget, which must be allocated between food and non-food expenditures). Suppose, as is commonly the case, that S is observed with some measurement error. To deal with this budget measurement error, a commonly employed set of instruments L consists of functions of the consumer's income. However, validity of functions of income as instruments for total consumption in a food Engel curve assumes separability between the consumer's decisions on savings and their within period food expenditure decision, and this behavioral assumption may or may not be valid. It is therefore useful to consider the alternative set of potential instruments P defined above. Use of P does not require finding covariates from outside the model, like income, to use as instruments, but does require that certain measurement error assumptions hold. Our later empirical application applies ODR to this application, thereby obtaining consistent estimates of α if either L or P are valid instruments.

4.4 The ODR Estimator Asymptotics

In this section we show consistency of our ODR estimator $\widehat{\alpha}$, and then derive its limiting distribution, which is root n consistent and asymptotically normal. We make the following additional assumptions. What these assumptions mostly do is make GMM estimation of models G, H, and F asymptotically normal around either the true values when correctly specified, or around pseudo-true values when misspecified, and ensure that the models are over identified.

Assumption A5: $G(Z, \alpha, \beta)$, $H(Z, \alpha, \gamma)$ and $F(Z, \alpha, \beta, \gamma)$ are continuous at $\{\alpha, \beta\} \in \Theta_\alpha \times \Theta_\beta$, $\{\alpha, \gamma\} \in \Theta_\alpha \times \Theta_\gamma$, and $\{\alpha, \beta, \gamma\} \in \Theta_\alpha \times \Theta_\beta \times \Theta_\gamma$ respectively, with probability one.

Assumption A6: With $||A|| \equiv \{trace(A'A)\}^{1/2}$ for a matrix A, $E[\sup_{\{\alpha,\beta\}\in\Theta_\alpha\times\Theta_\beta} ||G(Z, \alpha, \beta)||] < \infty$, $E[\sup_{\{\alpha,\gamma\}\in\Theta_\alpha\times\Theta_\gamma} ||H(Z, \alpha, \gamma)||] < \infty$, and $E[\sup_{\{\alpha,\beta,\gamma\}\in\Theta_\alpha\times\Theta_\beta\times\Theta_\gamma} ||F(Z, \alpha, \beta, \gamma)||] < \infty$.

Taken together Assumptions A1, A2, A3, A5, and A6, are standard conditions that suffice for consistency of the GMM estimators of models G, H, and F when

they are correctly specified. See, e.g., Theorem 2.1 in Newey and McFadden (1994). Let $\nabla_\theta(\cdot) \equiv \partial(\cdot)/\partial\theta$ be arranged such that its row dimension is that of θ and let $\nabla_{\theta'}(\cdot) \equiv \{\nabla_\theta(\cdot)\}'$. Define $\theta_0^g \equiv \{\alpha_0, \beta_0\}$, $\theta_0^h \equiv \{\alpha_0, \gamma_0\}$, $\theta_0^f \equiv \{\alpha_0, \beta_0, \gamma_0\}$, $\theta^g \equiv \{\alpha_g, \beta_g\}$, $\theta^h \equiv \{\alpha_h, \gamma_h\}$, and $\theta^f \equiv \{\alpha_f, \beta_f, \gamma_f\}$.

Assumption A7: With probability one, $G(Z, \alpha, \beta)$, $H(Z, \alpha, \gamma)$, and $F(Z, \alpha, \beta, \gamma)$ are twice continuously differentiable in a neighborhood $\aleph^g$ of θ^g, $\aleph^h$ of θ^h, and $\aleph^f$ of θ^f, respectively.

Assumption A8: $\nabla_\theta g_0(\theta_0^g)\Omega_g\nabla_{\theta'} g_0(\theta_0^g)$, $\nabla_\theta h_0(\theta_0^h)\Omega_h\nabla_{\theta'} h_0(\theta_0^h)$, and $\nabla_\theta f_0(\theta_0^f)\Omega_f\nabla_{\theta'} f_0(\theta_0^f)$ are non-singular.

Assumption A9: $\{\alpha_g, \beta_g\}$, $\{\alpha_h, \gamma_h\}$, and $\{\alpha_f, \beta_f, \gamma_f\}$ lie in the interior of $\Theta_\alpha \times \Theta_\beta$, $\Theta_\alpha \times \Theta_\gamma$, and $\Theta_\alpha \times \Theta_f \times \Theta_\gamma$.

Assumption A10: $E[||G(Z, \alpha, \beta)||^2] < \infty$, $E[||H(Z, \alpha, \gamma)||^2] < \infty$, and $E[||F(Z, \alpha, \beta, \gamma)||^2] < \infty$.

Assumption A11: $E[\sup_{\{\alpha,\beta\}\in\aleph^g} ||\nabla_{\theta^g} G(Z, \alpha, \beta)||] < \infty$, $E[\sup_{\{\alpha,\gamma\}\in\aleph^h} ||\nabla_{\theta^h} H(Z, \alpha, \gamma)||] < \infty$, and $E[\sup_{\{\alpha,\beta,\gamma\}\in\aleph^f} ||\nabla_{\theta^f} F(Z, \alpha, \beta, \gamma)||] < \infty$.

Assumption A7, A9, A10, and A11 are regularity conditions for a uniform weak law of large numbers and the asymptotic normality of GMM. Assumption A8 rules out perfect collinearity in linearized moment conditions. Assumption A11 gives interchangeability of $\nabla(\cdot)$ and $E(\cdot)$ so that

$$\nabla_\theta g_0(\theta^g) = E\{\nabla_{\theta^g} G(Z, \alpha_g, \beta_g)\},\ \nabla_\theta h_0(\theta^h) = E\{\nabla_{\theta^h} H(Z, \alpha_h, \gamma_h)\},\ \nabla_\theta f_0(\theta^f) = E\{\nabla_{\theta^f} F(Z, \alpha_f, \beta_f, \gamma_f)\}$$

Assumption A12: $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$ are $\sqrt{n}$-consistent, asymptotically normal estimators of Ω_g, Ω_h and Ω_f, respectively, where $\Omega_g^{-1} = Var\,[G(Z, \alpha_g, \beta_g)]$, $\Omega_h^{-1} = Var\,[H(Z, \alpha_h, \gamma_h)]$, and $\Omega_f^{-1} = Var\,[F(Z, \alpha_f, \beta_f, \gamma_f)]$.

Assumption A13: $E[||\nabla_{\theta^g} G(Z, \alpha, \beta)||^2] < \infty$, $E[||\nabla_{\theta^h} H(Z, \alpha, \gamma)||^2] < \infty$, and $E[||\nabla_{\theta^f} F(Z, \alpha, \beta, \gamma)||^2] < \infty$.

Assumption A14: Letting $\nabla_{\theta^g\theta^{g'}}(\cdot) \equiv \partial(\cdot)/\partial\theta^g\partial\theta^{g'}$, $E[\sup_{\{\alpha,\beta\}\in\aleph^g} ||\nabla_{\theta^g\theta^{g'}} G(Z, \alpha, \beta)||] < \infty$, $E[\sup_{\{\alpha,\gamma\}\in\aleph^h} ||\nabla_{\theta^h\theta^{h'}} H(Z, \alpha, \gamma)||] < \infty$, and $E[\sup_{\{\alpha,\beta,\gamma\}\in\aleph^f} ||\nabla_{\theta^f\theta^{f'}} F(Z, \alpha, \beta, \gamma)||] < \infty$.

Assumption A15: $plimVar\left[\frac{1}{\sqrt{n}}\sum_i G(Z_i,\theta^g)\right]$, $plimVar\left[\frac{1}{\sqrt{n}}\sum_i H(Z_i,\theta^h)\right]$, and $plimVar\left[\frac{1}{\sqrt{n}}\sum_i F(Z_i,\theta^F)\right]$ exist and are positive definite.

Assumption A12 strengthens the standard assumption for asymptotically efficient GMM estimation in requiring that the estimated weighting matrices converge at rate $\sqrt{n}$. This assumption is satisfied by the standard two-step GMM estimators for $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$, provided that the sample moments are demeaned, e.g., Ω_g is based on $Var\left[G(Z,\alpha_g,\beta_g)\right]$ rather than $E\left[G(Z,\alpha_g,\beta_g)G(Z,\alpha_g,\beta_g)'\right]$. The strengthening of Assumption A12 over the standard assumptions for GMM estimation ensures that the probability limits of $\hat{W}_g$ and $\hat{W}_g\hat{W}_f$ remain well behaved when either model G or H is misspecified.

Assumptions A13, A14, and A15 are for the asymptotic normality of the normalized sum of derivatives of G, H, and F. These assumptions are to ensure asymptotic normality of the GMM estimators when model G or H is misspecified. Assumptions A12 to A14 above are adapted from Hall and Inoue (2003), who use them to derive asymptotics for possibly misspecified GMM estimation.

4.4.1 ODR Consistency

Lemma 4.1: Suppose Assumptions A1 to A15 hold. Then, for any τ with $0<\tau<1$, $\hat{W}_f$ and $\hat{W}_f\hat{W}_g$, defined in equations (4.5) and (4.3), have finite probability limits. Specifically,

$$\begin{aligned}
&\text{Case 1) } G \text{ and } H \text{ are correctly specified} && \Longrightarrow \hat{W}_f \to^p 0 \text{ and } \hat{W}_f\hat{W}_g \to^p 0,\\
&\text{Case 2) } G \text{ is correctly specified but } H \text{ is not} && \Longrightarrow \hat{W}_f \to^p 1 \text{ and } \hat{W}_f\hat{W}_g \to^p 0,\\
&\text{Case 3) } H \text{ is correctly specified but } G \text{ is not} && \Longrightarrow \hat{W}_f \to^p 1 \text{ and } \hat{W}_f\hat{W}_g \to^p 1.
\end{aligned}$$

Lemma 4.1 is proved in Appendix I, but the intuition is as follows. When either G or H is misspecified, we have $\hat{Q}^f \to^p c_f'\Omega_f c_f/k_f > 0$, so $n^\tau\hat{Q}^f$ diverges to infinity and $\hat{W}_f \to^p 1$. If G is correct but H is not, then $\hat{Q}^g \to^p 0$ while the limiting value of $\hat{Q}^h$ is nonzero. Thus, $\hat{W}_g \to^p 0$ and so $\hat{W}_g\hat{W}_f \to^p 0$. If H is correct but G is not, following the same logic but switching the roles of g and h, $\hat{W}_g \to^p 1$ and so $\hat{W}_g\hat{W}_f \to^p 1$. When both G and H are correctly specified, so F is correctly specified, we have $\hat{Q}^f \to^p c_f'\Omega_f c_f/k_f = 0$, so $n^\tau\hat{Q}^f \to^p 0$ and therefore $\hat{W}_f \to^p 0$, and in this case both $n\hat{Q}^g$ and $n\hat{Q}^h$ converge to chi-squared distributions so $\hat{W}_g$ converges to a ratio of possibly dependent chi-squares, which is bounded in probability, making $\hat{W}_g\hat{W}_f \to^p 0$.

The following theorem shows consistency of the ODR estimator $\widehat{\alpha}$ in equation (4.1). We will further discuss construction of $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$ later, but note for now that these are recentered GMM weight matrix estimates using the sample moments in mean deviation form.

Theorem 4.1: Under Assumptions A1 to A15, for $\widehat{\alpha}$ given by equation (4.1), $\widehat{\alpha} \to^p \alpha_0$.

Proof of Theorem 4.1: By A1, A2, A3, A5, and A6, the conditions of Theorem 2.1 of in Newey and McFadden (1994) (uniqueness, compactness, continuity, and uniform convergence) hold for GMM based on model G, model H, or both when these moments are correctly specified. Therefore, if $g_0(\alpha_0, \beta_0) = 0$ then the GMM estimator of model G is consistent, if $h_0(\alpha_0, \gamma_0) = 0$ holds then the GMM estimator of model H is consistent, and if both the equalities hold parts hold then the GMM estimator of F is consistent.

For simplicity, let $\hat{Q}^g \equiv \hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$, $\hat{Q}^h \equiv \hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$, $\hat{Q}^f \equiv \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$, $Q_0^g \equiv c_g^{'}\Omega_g c_g/k_g$, $Q_0^h \equiv c_h^{'}\Omega_h c_h/k_h$, and $Q_0^f \equiv c_f^{'}\Omega_f c_f/k_f$. Assumption A2 says that either $g_0(\alpha_0, \beta_0) = 0$, $h_0(\alpha_0, \gamma_0) = 0$, or both. Consider each of these three cases.

Case 1) Suppose both $g_0(\alpha_0, \beta_0) = 0$ and $h_0(\alpha_0, \gamma_0) = 0$. Then $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_0, \beta_0\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_0, \gamma_0\}$, and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_0, \beta_0, \gamma_0\}$, so $\hat{Q}^g \to^p 0$, $\hat{Q}^h \to^p 0$, and $\hat{Q}^f \to^p 0$. By Lemma 1, $\hat{W}_f$ and $\hat{W}_f\hat{W}_g$ both converge to zero, and the consistency of $\widehat{\alpha}$ therefore follows from consistency of $\widehat{\alpha}_f$.

Case 2) Suppose that $g_0(\alpha_0, \beta_0) = 0$ and $h_0(\alpha_0, \gamma_0) \neq 0$. Then $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_0, \beta_0\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_h, \gamma_h\}$, and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_f, \beta_f, \gamma_f\}$. By Lemma 1, $\hat{W}_g$ converges to zero and $\hat{W}_f$ converges to one in probability. The consistency of $\widehat{\alpha}$ then follows from consistency of $\widehat{\alpha}_g$.

Case 3) Suppose that $g_0(\alpha_0, \beta_0) \neq 0$ and $h_0(\alpha_0, \gamma_0) = 0$. Then $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_g, \beta_g\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_0, \gamma_0\}$, and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_f, \beta_f, \gamma_f\}$. By Lemma 1, $\hat{W}_g$ and $\hat{W}_f$ both converge to one in probability, so consistency of $\widehat{\alpha}$ follows from consistency of $\widehat{\alpha}_h$. Q.E.D.

4.4.2 Limiting Distribution

We now provide the asymptotic distribution of $\widehat{\alpha}$, and a simple consistent estimator of its limiting variance. Let $\widehat{\eta}_i^g$, $\widehat{\eta}_i^h$ and $\widehat{\eta}_i^f$ be consistent estimators of the GMM influence functions for $\widehat{\alpha}_g$, $\widehat{\alpha}_h$ and $\widehat{\alpha}_f$, the details of which are in Appendix III.

Theorem 4.2: Suppose Assumptions A1 to A15 hold. Then, for $1/2 < \tau < 1$, there exists a matrix $\widetilde{V}$ such that

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(0, \widetilde{V}),$$

and

$$\frac{1}{n}\sum\nolimits_{i=1}^{n} \widehat{\eta}_i \widehat{\eta}_i' \to^p \widetilde{V} \tag{4.10}$$
$$\text{where } \widehat{\eta}_i \equiv \hat{W}_f \hat{W}_g \widehat{\eta}_i^h + \hat{W}_f(1 - \hat{W}_g)\widehat{\eta}_i^g + (1 - \hat{W}_f)\widehat{\eta}_i^f.$$

The first part of Theorem 4.2 states that the ODR estimator $\widehat{\alpha}$ is root n consistent and asymptotically normal, while the second part gives a consistent estimator for the limiting variance of $\widehat{\alpha}$. The proof of Theorem 4.2 is given in the Appendix I. The basic structure of the proof follows Newey and McFadden (1994) for multistep parametric estimators.

Note that while consistency only requires $0 < \tau < 1$, Theorem 4.2 assumes $\tau > 1/2$ to ensure $\sqrt{n}$-consistency of $\widehat{\alpha}$. This condition is only required for the case where $\alpha_g \neq \alpha_h$.

The estimator of $\widetilde{V}$ given in equation (4.10) does not require knowing which of the models G or H is correct. Nevertheless, as shown in Appendix I, $\widetilde{V}$ will either equal a matrix $\widetilde{V}^g$ or $\widetilde{V}^h$ or $\widetilde{V}^f$, depending on whether models G, H, or both are correctly specified.

A complication in the derivation of Theorem 4.2 is that, if model H is wrong, then we cannot consistently estimate the influence function η_i^h for model H. However, in the limiting variance formula for $\widehat{\alpha}$, the function η_i^h is multiplied by $\hat{W}_f \hat{W}_g$, so if model H is wrong then $\hat{W}_f \hat{W}_g$ goes to zero. We therefore only need an estimate for η_i^h that is consistent when model H is right, and that estimate is the standard GMM influence function $\widehat{\eta}_i^h$. A similar analysis applies to the influence function $\widehat{\eta}_i^g$ for model G when model G is wrong.

4.4.3 Efficiency and Numerical Issues

For asymptotic efficiency of α, we could consider estimating the weighting matrices $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$ to minimize the variance given by equation (4.10). However, the standard two step GMM estimators of $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$ should be at least close to efficient for $\widehat{\alpha}$. This is because the ODR objective function is asymptotically dominated by the GMM objective function of the correct model when either G or H is correct,

and dominated by the GMM objective function of model F when both models are correct.

The scaling of moments affects the relative magnitudes of $\hat{Q}^g$, $\hat{Q}^h$, and $\hat{Q}^f$ (and hence the estimated weights $\hat{W}_g$ and $\hat{W}_f$). It is therefore numerically desirable in finite samples to have these matrices be comparable in magnitude. The standard two step GMM estimators of $\hat{\Omega}_g$, $\hat{\Omega}_h$, and $\hat{\Omega}_f$ help make $\hat{Q}^g$, $\hat{Q}^h$, and $\hat{Q}^f$ comparable. Specifically, standard two step GMM makes $n\hat{Q}^g$ have a mean of one asymptotically when model G is right, and similarly for $n\hat{Q}^h$ and $n\hat{Q}^f$ (this is also the role of scaling each by the degrees of freedom k_g, k_h, and k_f, respectively). We therefore find it desirable to use the standard GMM estimates of $\hat{\Omega}_g$ and $\hat{\Omega}_h$ (as in Assumption A12) even if that possibly sacrifices a small amount of efficiency. In particular, we let

$$\hat{\Omega}_g \equiv \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(G(Z_i,\widehat{\alpha}_{1g},\widehat{\beta}_{1g})-\overline{G}(Z,\widehat{\alpha}_{1g},\widehat{\beta}_{1g})\right)\left(G(Z_i,\widehat{\alpha}_{1g},\widehat{\beta}_{1g})-\overline{G}(Z,\widehat{\alpha}_{1g},\widehat{\beta}_{1g})\right)' \tag{4.11}$$

where $\widehat{\alpha}_{1g}$ and $\widehat{\beta}_{1g}$ are first step GMM estimates based on a constant weighting matrix such as the identity matrix, and $\overline{G}$ is the sample average of $G(Z_i,\widehat{\alpha}_{1g},\widehat{\beta}_{1g})$. Analogous formulas apply for $\hat{Q}^h$ and $\hat{Q}^f$.

4.4.4 Comparison to Model Averaging

The weights in our SODR and ODR estimators can be compared to more traditional model averaging methods. An example of GMM model averaging (for instrument selection in linear instrumental variables models) is Martins and Gabriel (2014), who construct weights based on Andrews (1999)'s J-statistic based GMM model selection criteria. To most readily compare their weights to ours, consider the special case of our ODR in which the candidate models G and H are linear regressions with different sets of instruments. This comparison is particularly apt because our simulations and empirical application are choice of instruments in linear models.

Martins and Gabriel (2014) provide a variety of estimators, but the one that is closest to our model is

$$\widetilde{\alpha}^{MG} \equiv \hat{W}_g^{MG}\widehat{\alpha}_h + \left(1-\hat{W}_g^{MG}\right)\widehat{\alpha}_g$$
$$\text{where } \hat{W}_g^{MG} \equiv \frac{\exp\left(-\frac{1}{2}(n\widetilde{Q}^h-\kappa_n k_h)\right)}{\exp\left(-\frac{1}{2}(n\widetilde{Q}^h-\kappa_n k_h)\right)+\exp\left(-\frac{1}{2}(n\widetilde{Q}^g-\kappa_n k_g)\right)}$$

and $\kappa_n = o(n)$ is a sequence depending on the selection criteria, e.g. $\kappa_n = \ln(n)$ for

a Bayesian Information Criterion. This estimator is similar to our SODR with an exponential tuning function Λ.

One difference between $\widetilde{\alpha}^{MG}$ and SODR (with exponential Λ) is in the degrees of freedom term κ_n. Another important difference is that, in $\widetilde{\alpha}^{MG}$, as in other model averaging methods, the numerator of the weight on each model depends on the criterion for that model, while in our estimator, the numerator of the weight on model H depends on the criterion for model G (i.e., on $\widetilde{Q}^g$) and vice versa. This is because, for the DR property, we asymptotically need to put all weight on model H when model G is wrong, and vice versa. Note that Martins and Gabriel (2014) assume both models are correctly specified, and they do not account for the weight $\hat{W}_g^{MG}$ having a possibly random probability limit.

In contrast to SODR, our preferred ODR estimator differs more substantially from $\widetilde{\alpha}^{MG}$ in its construction of weights. We compare the finite sample performance of both our SODR and ODR estimators to $\widetilde{\alpha}^{MG}$ in the next section.

4.5 Simulation Results

Here we do some Monte Carlo analyses to investigate small sample properties of our estimator. Our design is two competing sets of instruments as in section 3.2. For each simulation, we draw $n = 100$ or $n = 500$ independent, identically distributed observations of the random vector $(Y, W, R_1, R_2, Q_1, Q_2)$. We generate data from the model

$$Y = \alpha_0 + \alpha_1 W + \epsilon.$$

The goal is estimation of $\alpha = (\alpha_0, \alpha_1) = (1, 1)$. The regressor W is endogenous (correlated with ϵ), so estimation is by instrumental variables. Model G assumes $E(\epsilon) = E(\epsilon R_1) = E(\epsilon R_2) = 0$, meaning that $R = (1, R_1, R_2)'$ is a vector of valid instruments for instrumental variables estimation. Model H assumes $E(\epsilon) = E(\epsilon Q_1) = E(\epsilon Q_2) = 0$, making $Q = (1, Q_1, Q_2)'$ be a vector of valid instruments. Here $Z = (Y, W, R, Q)$, $G(Z, \alpha) = (Y - \alpha_0 - \alpha_1 W) R$, and $H(Z, \alpha) = (Y - \alpha_0 - \alpha_1 W) Q$. In this application there is no β or γ.

We let $W = 1 + 4R_1 + R_2 + 2Q_1 + Q_2 + \epsilon$. Having the 4 and 2 in this equation means that model G has stronger instruments (i.e., instruments more highly correlated with the endogenous regressor W) than model H, and that R_1 and Q_1 are stronger instruments than R_2 and Q_2.

We let R_1, R_2, Q_1, Q_2, and ϵ be standard normals, with $corr(R_j, \epsilon) = \rho_{Rj}$, $corr(Q_j, \epsilon) = \rho_{Qj}$, for $j = 1, 2$, and all the other correlations among these normals are zero. We

consider three different simulation designs, that vary by correlations ρ_{Rj} and ρ_{Qj}. The first design takes $\rho_{Rj} = \rho_{Qj} = 0$, which makes both models right (both sets of instruments are valid). The second takes $\rho_{R1} = \rho_{R2} = 0$, $\rho_{Q1} = 0.4$, and $\rho_{Q2} = 0.6$, which makes model G right (i.e., R are valid instruments so G is correctly specified) and model H be wrong (i.e., Q are not valid instruments, because they correlate with the model error ϵ). The third takes $\rho_{R1} = 0.4$, $\rho_{R2} = 0.6$ and $\rho_{Q1} = \rho_{Q2} = 0$, which makes model H right and model G wrong.

For the tuning function Λ discussed in sections 2.3 and 4.4, we consider two different choices; $\Lambda_1\left(n\hat{Q}\right) = \exp\left(n\hat{Q}\right) - 1$ and $\Lambda_2(n\hat{Q}) = (n\hat{Q})^2$ so the weighting functions $\hat{W}_g$ and $\hat{W}_f$ are

$$\Lambda_1 : \hat{W}_g = \frac{\exp\{n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\} - 1}{\exp\{n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\} + \exp\{n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\} - 2}, \hat{W}_f = 1 - \frac{1}{\exp\{n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)\}}, \tag{4.12}$$

$$\Lambda_2 : \hat{W}_g = \frac{\{n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\}^2}{\{n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\}^2 + \{n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\}^2}, \hat{W}_f = 1 - \frac{1}{\{n^\tau \hat{Q}^f(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)\}^2 + 1}. \tag{4.13}$$

For the tuning parameter τ, we use $\tau = 1 - p$, where p is the p-value of the Wald statistic as discussed in section 4.2.3.

We report eight estimates of α_1 and α_0 for each simulation. First is GMM based on the model G moments, denoted by GMM_g (which is only consistent if model G is right). Second is GMM based on the H moments, denoted by GMM_h (which is only consistent if model H is right). Third is GMM based on both sets of moments, denoted by GMM_f (which is consistent, and more efficient than either the first or second set of estimates, only if both models are right). Fourth is the model averaging estimator provided by Martins and Gabriel (2014) and discussed in section 4.4.4, denoted by MG. Fifth and sixth are our ODR estimators in equation (4.1) using tuning functions Λ_1 and Λ_2, respectively, denoted by ODR_{Λ_1} and ODR_{Λ_2} (which are consistent for all designs). Seventh and eighth are our simpler estimators in equation (4.4), denoted by $SODR_{\Lambda_1}$ and $SODR_{\Lambda_2}$ (which are consistent for all designs, but asymptotically less efficient than ODR when both sets of moments are valid).

For each of the eight estimators, Tables 4.1 and 4.2 present simulation results of $n = 100$ observations, and Tables 4.3 and 4.4 present simulation results of $n = 500$ observations. All tables are based on 2000 Monte Carlo simulations. The reported summary statistics on the estimated parameters are, respectively, the bias (Bias), me-

Table 4.1: Simulation Results of α_1 ($n = 100$)

	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
Both correct										
GMM_g	0.0008	0.0015	0.0006	0.0170	0.0247	0.0819	3.1153	0.9390	0.0236	0.0038
GMM_h	-0.0010	0.0010	0.0023	0.0302	0.0480	0.2350	2.8924	0.9520	0.0455	0.0138
GMM_f	0.0012	0.0018	0.0005	0.0159	0.0222	0.0833	3.0008	0.9290	0.0202	0.0030
MG	-0.0010	-0.0004	0.0008	0.0184	0.0288	0.1681	2.9945	0.9535	0.0268	0.0084
ODR_{Λ_1}	0.0004	0.0012	0.0006	0.0164	0.0255	0.0829	3.0406	0.9250	0.0214	0.0049
ODR_{Λ_2}	0.0006	0.0011	0.0005	0.0149	0.0232	0.1840	3.3537	0.9285	0.0210	0.0040
$SODR_{\Lambda_1}$	-0.0016	-0.0003	0.0012	0.0200	0.0348					
$SODR_{\Lambda_2}$	-0.0011	-0.0003	0.0012	0.0201	0.0342					
G correct										
GMM_g	0.0007	0.0022	0.0006	0.0169	0.0248	0.2852	3.2493	0.9380	0.0237	0.0046
GMM_h	0.1991	0.1951	0.0413	0.1951	0.0408	0.3284	3.3054	0.0000	0.0348	0.0100
GMM_f	0.0731	0.0725	0.0059	0.0725	0.0244	0.2104	3.1962	0.0540	0.0166	0.0023
MG	0.0372	0.0284	0.0038	0.0317	0.0487	0.8547	3.0249	0.5760	0.0201	0.0054
ODR_{Λ_1}	0.0229	0.0114	0.0038	0.0207	0.0570	1.4912	4.8689	0.7730	0.0230	0.0061
ODR_{Λ_2}	0.0247	0.0130	0.0038	0.0223	0.0563	1.3465	4.3800	0.7560	0.0229	0.0059
$SODR_{\Lambda_1}$	0.0229	0.0114	0.0038	0.0207	0.0570					
$SODR_{\Lambda_2}$	0.0242	0.0122	0.0038	0.0223	0.0569					
H correct										
GMM_g	0.1123	0.1121	0.0130	0.1121	0.0201	0.3521	3.4293	0.0000	0.0163	0.0025
GMM_h	0.0003	0.0069	0.0025	0.0308	0.0498	0.6336	3.3317	0.9220	0.0465	0.0238
GMM_f	0.0938	0.0939	0.0092	0.0939	0.0193	0.3582	3.3534	0.0015	0.0145	0.0021
MG	0.0009	0.0075	0.0025	0.0309	0.0499	0.6559	3.3722	0.9165	0.0462	0.0238
ODR_{Λ_1}	0.0025	0.0080	0.0024	0.0317	0.0494	1.2264	5.4509	0.8925	0.0449	0.0214
ODR_{Λ_2}	0.0047	0.0110	0.0024	0.0320	0.0489	1.5845	7.7317	0.8800	0.0437	0.0204
$SODR_{\Lambda_1}$	0.0003	0.0070	0.0025	0.0308	0.0499					
$SODR_{\Lambda_2}$	0.0001	0.0073	0.0026	0.0311	0.0509					

dian error (MdE), root mean-squared error (RMSE), median absolute error (MAE), and the standard deviation (SD). To check the quality of our limiting distribution, we also calculate the estimated t-statistic $\widehat{\alpha}_j - 1$ divided by the estimated standard error of $\widehat{\alpha}_j$ for $j = 0, 1$ in each simulation. We report skewness (Skew) and kurtosis (Kurt) of these t-statistics across simulations, and the frequency (Freq) that these t-statistics are less than 2 in magnitude, corresponding to the frequency with which a ± 2 estimated standard error confidence interval contains the true parameter value. Also, to check the accuracy of the standard error estimates, we report the average of the estimated standard errors (SE), and standard deviation of the estimated standard errors (SD_{SE}), across the simulations. The last five summary statistics are not reported for $SODR$, because we do not consider its limiting distribution due to the random probability limit of $\hat{W}_g$.

Table 4.2: Simulation Results of α_0 ($n = 100$)

	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
Both correct										
GMM_g	-0.0038	-0.0048	0.0112	0.0687	0.1058	0.0005	3.1738	0.9415	0.1009	0.0089
GMM_h	-0.0024	-0.0090	0.0134	0.0757	0.1157	-0.0131	2.9788	0.9490	0.1115	0.0182
GMM_f	-0.0046	-0.0073	0.0113	0.0688	0.1063	0.0212	3.1124	0.9350	0.0981	0.0085
MG	-0.0022	-0.0063	0.0115	0.0697	0.1071	0.0291	3.0642	0.9440	0.1022	0.0110
ODR_{Λ_1}	-0.0039	-0.0062	0.0113	0.0686	0.1063	0.0583	3.0524	0.9370	0.0989	0.0092
ODR_{Λ_2}	0.0001	-0.0017	0.0105	0.0687	0.1025	-0.0532	3.1744	0.9525	0.0990	0.0088
$SODR_{\Lambda_1}$	-0.0016	-0.0067	0.0120	0.0703	0.1097					
$SODR_{\Lambda_2}$	0.0014	0.0023	0.0108	0.0707	0.1041					
G correct										
GMM_g	-0.0038	-0.0060	0.0112	0.0683	0.1060	-0.0390	3.1287	0.9395	0.1009	0.0108
GMM_h	-0.2005	-0.1977	0.0554	0.1977	0.1234	0.1485	3.0509	0.5750	0.1103	0.0179
GMM_f	-0.0744	-0.0737	0.0219	0.0999	0.1280	-0.0354	3.1266	0.7540	0.0867	0.0074
MG	-0.0401	-0.0396	0.0140	0.0774	0.1115	-0.1154	3.1855	0.8885	0.0954	0.0109
ODR_{Λ_1}	-0.0258	-0.0198	0.0147	0.0722	0.1186	-0.2332	3.2476	0.9010	0.0996	0.0120
ODR_{Λ_2}	-0.0245	-0.0198	0.0136	0.0744	0.1139	-0.2004	3.0110	0.9065	0.0995	0.0114
$SODR_{\Lambda_1}$	-0.0258	-0.0198	0.0147	0.0722	0.1186					
$SODR_{\Lambda_2}$	-0.0240	-0.0194	0.0136	0.0745	0.1142					
H correct										
GMM_g	-0.1151	-0.1166	0.0230	0.1198	0.0989	0.0139	2.8983	0.6735	0.0808	0.0069
GMM_h	-0.0028	-0.0088	0.0133	0.0722	0.1153	-0.2405	2.9748	0.9530	0.1123	0.0344
GMM_f	-0.0963	-0.0966	0.0203	0.1039	0.1050	-0.0085	2.9169	0.7095	0.0791	0.0068
MG	-0.0035	-0.0094	0.0133	0.0720	0.1151	-0.2389	2.9660	0.9515	0.1120	0.0343
ODR_{Λ_1}	-0.0051	-0.0105	0.0131	0.0725	0.1146	-0.2535	2.9609	0.9475	0.1109	0.0320
ODR_{Λ_2}	-0.0084	-0.0187	0.0135	0.0753	0.1159	-0.1964	3.0290	0.9380	0.1095	0.0287
$SODR_{\Lambda_1}$	-0.0029	-0.0089	0.0133	0.0722	0.1153					
$SODR_{\Lambda_2}$	-0.0038	-0.0144	0.0138	0.0760	0.1176					

Table 4.3: Simulation Results of α_1 ($n = 500$)

	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
Both correct										
GMM_g	-0.0001	0.0001	0.0001	0.0074	0.0108	0.0652	2.8494	0.9565	0.0108	0.0008
GMM_h	-0.0005	-0.0004	0.0004	0.0131	0.0199	0.0602	2.9137	0.9565	0.0200	0.0023
GMM_f	0.0000	0.0001	0.0001	0.0066	0.0096	0.0227	2.7919	0.9495	0.0094	0.0006
MG	-0.0004	-0.0007	0.0001	0.0081	0.0120	0.0009	2.7642	0.9525	0.0119	0.0024
ODR_{Λ_1}	-0.0001	0.0001	0.0001	0.0069	0.0106	0.0145	2.7364	0.9390	0.0097	0.0013
ODR_{Λ_2}	-0.0004	-0.0006	0.0001	0.0067	0.0109	0.1468	3.1553	0.9415	0.0097	0.0013
$SODR_{\Lambda_1}$	-0.0005	-0.0006	0.0002	0.0091	0.0142					
$SODR_{\Lambda_2}$	-0.0007	-0.0004	0.0002	0.0089	0.0149					
G correct										
GMM_g	-0.0001	0.0002	0.0001	0.0073	0.0108	0.1479	2.8751	0.9560	0.0108	0.0009
GMM_h	0.1990	0.1986	0.0399	0.1986	0.0177	0.1549	3.0003	0.0000	0.0155	0.0018
GMM_f	0.0729	0.0728	0.0054	0.0728	0.0109	0.1287	3.0088	0.0000	0.0077	0.0005
MG	0.0001	0.0004	0.0001	0.0073	0.0110	0.3379	4.0679	0.9535	0.0108	0.0009
ODR_{Λ_1}	-0.0001	0.0002	0.0001	0.0073	0.0108	0.1480	2.8743	0.9560	0.0108	0.0009
ODR_{Λ_2}	0.0010	0.0010	0.0001	0.0076	0.0115	1.2373	10.9036	0.9425	0.0107	0.0009
$SODR_{\Lambda_1}$	-0.0001	0.0002	0.0001	0.0073	0.0108					
$SODR_{\Lambda_2}$	0.0009	0.0009	0.0001	0.0076	0.0115					
H correct										
GMM_g	0.1124	0.1125	0.0127	0.1125	0.0091	0.1833	2.9687	0.0000	0.0074	0.0005
GMM_h	-0.0004	0.0006	0.0004	0.0132	0.0201	0.3063	3.0314	0.9580	0.0201	0.0034
GMM_f	0.0939	0.0937	0.0089	0.0937	0.0088	0.1908	3.0597	0.0000	0.0067	0.0004
MG	-0.0004	0.0006	0.0004	0.0132	0.0201	0.3063	3.0314	0.9580	0.0201	0.0034
ODR_{Λ_1}	-0.0004	0.0006	0.0004	0.0132	0.0201	0.3063	3.0314	0.9580	0.0201	0.0034
ODR_{Λ_2}	0.0002	0.0018	0.0004	0.0131	0.0203	0.3885	3.1614	0.9475	0.0201	0.0035
$SODR_{\Lambda_1}$	-0.0004	0.0006	0.0004	0.0132	0.0201					
$SODR_{\Lambda_2}$	0.0001	0.0017	0.0004	0.0131	0.0203					

Table 4.4: Simulation Results of α_0 ($n = 500$)

	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
Both correct										
GMM_g	-0.0010	-0.0002	0.0021	0.0315	0.0458	-0.1391	2.9732	0.9565	0.0459	0.0018
GMM_h	-0.0008	0.0005	0.0024	0.0328	0.0492	-0.1701	3.0631	0.9500	0.0491	0.0030
GMM_f	-0.0011	0.0000	0.0021	0.0311	0.0458	-0.1335	2.9799	0.9550	0.0454	0.0017
MG	-0.0007	0.0004	0.0021	0.0311	0.0463	-0.1527	3.0340	0.9570	0.0462	0.0021
ODR_{Λ_1}	-0.0010	0.0000	0.0021	0.0310	0.0459	-0.1327	3.0063	0.9540	0.0455	0.0018
ODR_{Λ_2}	0.0009	-0.0005	0.0022	0.0315	0.0471	0.0061	2.9664	0.9445	0.0455	0.0019
$SODR_{\Lambda_1}$	-0.0005	0.0003	0.0022	0.0321	0.0468					
$SODR_{\Lambda_2}$	0.0010	0.0003	0.0023	0.0334	0.0483					
G correct										
GMM_g	-0.0010	-0.0003	0.0021	0.0314	0.0458	-0.1566	2.9735	0.9570	0.0459	0.0021
GMM_h	-0.2000	-0.2000	0.0428	0.2000	0.0529	0.0663	3.1573	0.0225	0.0495	0.0033
GMM_f	-0.0732	-0.0731	0.0084	0.0739	0.0554	0.0501	2.9813	0.5400	0.0402	0.0014
MG	-0.0012	-0.0004	0.0021	0.0314	0.0458	-0.1553	2.9705	0.9570	0.0458	0.0021
ODR_{Λ_1}	-0.0010	-0.0003	0.0021	0.0314	0.0458	-0.1563	2.9744	0.9570	0.0459	0.0021
ODR_{Λ_2}	-0.0020	-0.0011	0.0021	0.0315	0.0459	-0.1685	2.9918	0.9550	0.0457	0.0021
$SODR_{\Lambda_1}$	-0.0010	-0.0003	0.0021	0.0314	0.0458					
$SODR_{\Lambda_2}$	-0.0020	-0.0011	0.0021	0.0315	0.0459					
H correct										
GMM_g	-0.1122	-0.1121	0.0146	0.1121	0.0448	-0.0037	3.0575	0.1945	0.0367	0.0013
GMM_h	-0.0007	-0.0007	0.0024	0.0329	0.0494	-0.2688	3.0914	0.9480	0.0492	0.0048
GMM_f	-0.0938	-0.0948	0.0111	0.0948	0.0481	-0.0661	2.9792	0.3445	0.0366	0.0013
MG	-0.0007	-0.0007	0.0024	0.0329	0.0494	-0.2688	3.0914	0.9480	0.0492	0.0048
ODR_{Λ_1}	-0.0007	-0.0007	0.0024	0.0329	0.0494	-0.2688	3.0914	0.9480	0.0492	0.0048
ODR_{Λ_2}	-0.0011	-0.0038	0.0025	0.0340	0.0500	-0.1804	2.9318	0.9555	0.0491	0.0049
$SODR_{\Lambda_1}$	-0.0007	-0.0007	0.0024	0.0329	0.0494					
$SODR_{\Lambda_2}$	-0.0011	-0.0037	0.0025	0.0340	0.0500					

When both sets of instruments are valid, ODR estimates are almost as precise as GMM_f, and when either set of instruments is invalid, ODR estimates are more precise than inconsistent GMM estimators. The $SODR$ estimates are found to be less efficient than ODR when both G and H models are valid (as expected), but when one model is invalid, $SODR$ is similar to ODR. In this application, the cost in efficiency of choosing the simpler $SODR$ seems small[16]. Presumably the gains to ODR would have been larger in a simulation design where the efficiency of GMM_f more greatly exceeded that of GMM_g.

Despite the fact that MG is specifically designed for instrument selection in linear models, while our ODR is a generic estimator for arbitrary moment based models, the finite sample performance of ODR is close to, and in some cases slightly better than, MG, particularly when both models are correctly specified.

Our simulation results also show that the limiting distributions provide reasonably good approximations to their finite sample counterparts, and these approximations improve substantially when going from the sample size $n = 100$ to $n = 500$. In particular, the quality of ODR estimated standard errors and confidence intervals is similar to that of the corresponding correctly specified GMM standard errors and confidence intervals. This can be seen by comparing the SE and SD columns, and comparing how close Freq is to .95 in the ODR rows, relative to same comparisons in the correctly specified GMM rows. Indeed, at $n = 500$ almost all of the summary statistics of ODR become close to those of the most efficient correctly specified GMM in each block. One exception is ODR_{Λ_2} when the model H is invalid. In this case, there were a few large outlier ODR_{Λ_2} estimates, resulting in substantial nonnormal skewness and kurtosis in the t-statistic distribution. But other summary statistics are still similar to those of ODR_{Λ_1} and correctly specified GMM. This suggests a modest advantage of the exponential tuning function Λ_1.

One should expect correctly specified GMM estimators to be more efficient than ODR, and that is indeed the case. But in many of the simulations, the loss in efficiency from using ODR is very low. In particular, when model G is invalid, so only the weaker instruments are valid, the precision of ODR is almost identical to that of the efficient GMM_h. So, using our ODR, there is little loss in efficiency from not knowing which specification is correct. In summary, we conclude that our proposed ODR works well, even at low sample sizes.

[16]However, $SODR$ incurs the additional cost of possibly not having a normal limiting distribution when both G and H are correctly specified.

4.6 Empirical Application: Engel Curve Estimation

Here we empirically estimate the Engel curve example discussed in section 4.3.2. Y is the food budget share, S is log real total consumption expenditures, and X is a vector of other covariates that serve as controls[17]. The goal is estimation of the coefficient of S in a regression of Y on S and X. Total consumption S is observed with measurement error, so instrumental variables estimation is used to correct for the resulting endogeneity. The vector L consists of two candidate external instrument variables, real total income and real total income squared. Model G assumes these external instruments are valid. Model H instead assumes that constructed instruments based on heteroscedasticity as described by Lewbel (2012) and summarized in section 4.3.2 above are valid. Model F assumes boths sets of instruments are valid.

The data consist of 854 households collected from the UK Family Expenditure Survey 1980-1982 as studied by Banks, Blundell, and Lewbel (1997), Lewbel (2012), and Baum and Schaffer (2012). The sample means are $\overline{Y} = 0.285$ and $\overline{S} = 0.599$, and the standard deviations are 0.106 for Y and 0.410 for S.

The parameter of interest is the coefficient of log real total expenditure α_s. Table 5 summarizes estimates of α_s and of the constant term α_0. GMM_{g0} is the estimate reported in Lewbel (2012) and Baum and Schaffer (2012). GMM_g is the GMM estimator using the moments in equation (4.8), which makes use of the external instruments L.[18] GMM_h is the GMM estimator that uses the moments in equation (4.9), which are heteroscedasticity based constructed instruments. GMM_f is the GMM estimator that uses both sets of instruments, and $SODR$ and ODR are our new estimators given in equations (4.4) and (4.1) with the tuning functions Λ_1 and Λ_2..

The estimated results show that the external instruments of model G are much stronger than the constructed instruments of model H. This is not surprising since the constructed instruments are based on higher moments of the data. This difference in strength can be seen in the standard errors of $\hat{\alpha}_s$, which are much lower in model G than in model H, and also in model GMM_f which gives estimates much closer to GMM_g than GMM_h.

The point estimates of GMM_g and GMM_h are substantially different, which could

[17]These covariates are a constant, age, spouse's age, squared ages, seasonal dummies, and dummies for spouse working, gas central heating, ownership of a washing machine, one car, and two cars.

[18]The estimates of GMM_{g0} and GMM_g are not identical because we use the two external instruments income and income squared, instead of just using income. There's a similar small difference between GMM_f and the models based on both sets of moments reported in Lewbel (2012) and Baum and Schaffer (2012), for the same reason.

Table 4.5: Engel Curve Estimates

	GMM_{g0}	GMM_g	GMM_h	GMM_f	$SODR_{\Lambda_1}$	ODR_{Λ_1}	$SODR_{\Lambda_2}$	ODR_{Λ_2}
$\hat{\alpha}_s$	-0.0859 (0.0198)	-0.0840 (0.0197)	-0.0521 (0.0546)	-0.0862 (0.0177)	-0.0812	-0.0862 (0.0192)	-0.0831	-0.0862 (0.0192)
$\hat{\alpha}_0$	0.336 (0.0122)	0.335 (0.0120)	0.317 (0.0328)	0.337 (0.0109)	0.333	0.337 (0.0118)	0.335	0.337 (0.0118)
χ^2		0.191	12.91	15.94				
$d.f.$		1	11	13				
p-value		0.662	0.299	0.252				
$\hat{Q}$		0.0002	0.0014	0.0014				
$\hat{W}_g, \hat{W}_f, p$					0.09, 0.004, 0.86		0.03, 0.000, 0.86	

19

be due to having one of these sets of instruments be invalid. However, this difference could also just be due to imprecision, particularly of GMM_h. This illustrates the usefulness of our ODR, which does not require resolving which set of instruments is valid, or if both are valid.

The estimated weight $\hat{W}_g$ is 0.09 with the tuning function Λ_1 and 0.03 with Λ_2, so $SODR$ puts over ten times as much weight on model G as on model H. However, in ODR the weight on model F, $1 - \hat{W}_f$, is 0.996 with Λ_1 and one to three decimal places with Λ_2. The very small difference in $\hat{W}_f$ between Λ_1 and Λ_2 is why both of the ODR estimates appear the same in Table 4.5 (they actually differ in the fourth significant digit: -0.08617 vs. -0.08619 for $\hat{\alpha}_s$).

The very high weight on model F strongly suggests that both models are likely to be correctly specified. This therefore implies that the difference between GMM_g and GMM_h is likely due to imprecision of GMM_h rather than misspecification of the constructed instruments in model H. Further evidence that both are correctly specified is given by the chi-squared statistics in Table 4.5, which test validity of the moments comprising each of the GMM estimates. This situation, where both models appear to be correctly specified, is when we would expect ODR to perform better than $SODR$.

Lewbel (2012) observes that a virtue of the constructed instruments is that they are valid under very different conditions than those required for validity of the external instruments, and suggests that they therefore are useful for testing overidentification. Our proposed ODR estimator makes further use of these instruments, by delivering estimates that are consistent if either (or both) sets of instruments are valid.

4.7 Local Misspecification

Consider the case where model G or H is locally misspecified with the parameter in the data generating process being $\theta^g = \theta_0^g + \delta_g n^{-s}$ or $\theta^h = \theta_0^h + \delta_h n^{-s}$ for constants δ_g and δ_h, and $s > 0$. Note $s = 0$ is equivalent to global misspecification, while $s = \infty$ is equivalent to correct specification, which are the cases we have already considered in our previous theorems. Pitman (1949) drift corresponds to the case of $s = 1/2$. This model is used by, e.g. Newey and West (1987), Bera and Yoon (1993) and Newey and McFadden (1994) to develop local power analyses. Here we summarize the asymptotic properties of our ODR estimator under local misspecification, with formal results provided in Appendix II.

The asymptotic distribution of $\sqrt{n}(\hat{\alpha} - \alpha_0)$ depends on the value of s. We show in Appendix II that the influence function of our ODR estimator consists of three terms; the first is the weighted sum of three different well behaved influence functions, the second converges to zero in probability for all $s \geq 0$, and the third either converges to a constant or diverges depending on s (and sometimes τ) as discussed below.[20]

First suppose model G is locally misspecified with $s > 1/2$. Then $n\widetilde{Q}^g\left(\widehat{\alpha}_g, \widehat{\beta}_g\right) \to^d \chi^2_{k_g}(0)$, which is the same limit as when G is correctly specified, and similarly for H. As a result, in this case the SODR and ODR estimators have the same $\sqrt{n}$ consistent, asymptotically normal limiting distribution as they have when G is correctly specified, and similarly for H. Note this means that instead of requiring that either G or H (or both) be correctly specified, it is sufficient to assume that either G or H (or both) are locally misspecified with $s > 1/2$, noting that correct specification is the special case of $s = \infty$.

If model G is locally misspecified with $s < 1/2$, then $n\widetilde{Q}^g\left(\widehat{\alpha}_g, \widehat{\beta}_g\right)$ diverges, and the SODR has the same $\sqrt{n}$ consistent, asymptotically normal limiting distribution as when G is globally misspecified. The ODR will also have the same limiting distribution as when G is globally misspecified, as long as the tuning parameter τ has $\tau > s + 0.5$. This then guarantees that model G will asymptotically have zero weight. Since these cases are equivalent asymptotically to G being globally misspecified, we need to assume that H is either correctly specified, or locally misspecified with its $s > 1/2$. This generalizes our original theorems that simply assumed either G or H is correctly specified.

[20] In Appendix II we also explicitly derive the implications of these results for the limiting distribution of the ODR estimator when one model is correctly specified and the other is locally misspecified for varying values of s. The results summarized in this subsection are all either directly verified in Appendix II, or are immediate extentions.

Finally, suppose model G is locally misspecified with $s = 1/2$. Then $n\widetilde{Q}^g$ converges to a noncentral chi-squared distribution. Specifically, $n\widetilde{Q}^g\left(\widehat{\alpha}_g, \widehat{\beta}_g\right) \to^d \chi^2_{k_g}(\omega_g'\Omega_g^{1/2}\Pi_g^*\Omega_g^{1/2}\omega_g)$, where the object in parentheses is the noncentrality parameter and the definitions of Π_g^* and ω_g are given in equation (C.3) and at the beginning of Appendix II, respectively. In this case the GMM estimator of model G is consistent but not $\sqrt{n}$ consistent, as established in, e.g., Newey and McFadden (1994). Here $n\widetilde{Q}^g$ is still bounded in probability, so ODR will asymptotically put weight on either model G or, if H is correctly specified (or locally misspecified with its $s > 1/2$) on model F, which then is consistent but may not be $\sqrt{n}$ consistent. As a result, in this knife edge case, ODR will be consistent, but not $\sqrt{n}$ consistent.

The main results here can be summarized as follows. If both G and H are locally misspecified, each with $s > 1/2$ (including the special case where one or both is correctly specified, corresponding to $s = \infty$), then ODR will have the same limiting distribution as efficient GMM with both G and H correctly specified. If just G is locally misspecified with $s > 1/2$ (again including as a special case having G be correctly specified by $s = \infty$), and H is either misspecified or locally misspecified with $s < 1/2$, then (assuming $\tau > s + 0.5$) ODR will have the same limiting distribution as efficient GMM based just on model G (and vice versa, exchanging the roles of G and H). Equivalently we can say that our earlier Theorem 4.2 still holds, replacing "correctly specified model" with "locally misspecified model having any $s > 1/2$, including $s = \infty$" and replacing "incorrectly specified model" with "locally misspecified model having any $s < 1/2$, including $s = 0$."

We conclude this section with some Monte Carlo results (reported in Tables 4.6 to 4.7 below), which we find support these conclusions. We use the same simulation designs and estimators as in section 4.5 but with a drift parameter s for the locally misspecified cases. Since ODR performed better with the tuning function Λ_1 in section 5, to save space we only report ODR_{Λ_1}, along with GMM_g, GMM_h, and GMM_f for comparison. In all these tables, model H is either globally mispecified, or locally misspecified with s equal to 0.25, 0.50, or 0.75. In Tables 4.6-1 and 4.6-2 model G is correctly specified, while in Tables 4.7-1 and 4.7-2, G is locally misspecified with $s = 0.75$.

The finite sample results in these tables largely accord with asymptotic theory, with one interesting difference. When model H is locally misspecified with $s = 0.5$ (Pitman drift) our ODR should be comparable to GMM_f, but actually performs slightly better than GMM_f. This is due to our use of the Wald statistic to select τ. With $s = 0.5$, the Wald statistic over-rejects the null, making τ large and therefore

Table 4.6-1: Model G is Correctly Specified and Model H is Misspecified ($n = 500$)

α_1	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
s=0.25										
GMM_g	0.0002	0.0006	0.0001	0.0075	0.0111	0.2310	3.1966	0.9465	0.0108	0.0011
GMM_h	0.2374	0.2367	0.0566	0.2367	0.0157	0.1558	3.1392	0.0000	0.0139	0.0016
GMM_f	0.1094	0.1094	0.0121	0.1094	0.0112	0.0817	3.0557	0.0000	0.0068	0.0005
ODR_{Λ_1}	0.0002	0.0006	0.0001	0.0075	0.0111	0.2311	3.1963	0.9460	0.0108	0.0011
s=0.5										
GMM_g	0.0002	0.0006	0.0001	0.0075	0.0110	0.1255	3.0813	0.9535	0.0108	0.0008
GMM_h	0.0827	0.0822	0.0071	0.0822	0.0174	-0.0439	3.1104	0.0045	0.0175	0.0019
GMM_f	0.0220	0.0223	0.0006	0.0223	0.0093	0.0825	3.0819	0.3365	0.0090	0.0006
ODR_{Λ_1}	0.0128	0.0058	0.0008	0.0102	0.0259	0.9210	3.2417	0.7455	0.0109	0.0019
s=0.75										
GMM_g	-0.0001	0.0001	0.0001	0.0074	0.0108	0.0707	2.8505	0.9570	0.0108	0.0008
GMM_h	0.0181	0.0180	0.0007	0.0194	0.0192	0.0233	2.9125	0.8355	0.0193	0.0021
GMM_f	0.0044	0.0045	0.0001	0.0074	0.0095	0.0275	2.7750	0.9270	0.0094	0.0006
ODR_{Λ_1}	0.0058	0.0052	0.0002	0.0081	0.0123	0.0457	2.7501	0.8905	0.0099	0.0016
Global										
GMM_g	-0.0001	0.0002	0.0001	0.0073	0.0108	0.1479	2.8751	0.9560	0.0108	0.0009
GMM_h	0.1990	0.1986	0.0399	0.1986	0.0177	0.1549	3.0003	0.0000	0.0155	0.0018
GMM_f	0.0729	0.0728	0.0054	0.0728	0.0109	0.1287	3.0088	0.0000	0.0077	0.0005
ODR_{Λ_1}	-0.0001	0.0002	0.0001	0.0073	0.0108	0.1480	2.8743	0.9560	0.0108	0.0009

pulling the ODR estimator towards to GMM_g, which is better behaved than GMM_f with Pitman drift.

4.8 Extension: Multiple Robustness

It is possible to construct triply and higher multiply robust estimators that are similar to $SODR$. Suppose we have a third model, called model L, with GMM objective function $\hat{Q}^l(\alpha, \lambda)$. The GMM estimator of model L is $\{\widehat{\alpha}_l, \widehat{\lambda}_l\} = \arg\min_{\{\alpha,\lambda\}\in\Theta_\alpha\times\Theta_\lambda} \hat{Q}^l(\alpha, \lambda)$. A possible formula for triply robust estimation of α would then be the weighted average

$$\widetilde{\alpha} = \frac{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\widehat{\alpha}_l + \hat{Q}^l(\widehat{\alpha}_l, \widehat{\lambda}_l)\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\widehat{\alpha}_g + \hat{Q}^l(\widehat{\alpha}_l, \widehat{\lambda}_l)\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\widehat{\alpha}_h}{\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h) + \hat{Q}^l(\widehat{\alpha}_l, \widehat{\lambda}_l)\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h) + \hat{Q}^l(\widehat{\alpha}_l, \widehat{\lambda}_l)\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)}. \tag{4.14}$$

In equation (4.14), the weight on $\widehat{\alpha}_l$ is proportional to the product of objective functions for the other models, $\hat{Q}^g\hat{Q}^h$, and similarly for the weights on $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$.

The logic of this estimator is the same as for our $SODR$ estimator. For example, if model G is right and models L and H are wrong, then only $\widehat{\alpha}_g$ will get a nonzero

Table 4.6-2: Model G is Correctly Specified and Model H is Misspecified ($n = 500$)

α_0	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
s=0.25										
GMM_g	-0.0005	-0.0008	0.0022	0.0319	0.0467	0.0210	2.8965	0.9530	0.0459	0.0025
GMM_h	-0.2385	-0.2364	0.0598	0.2364	0.0544	-0.0138	2.8922	0.0030	0.0504	0.0033
GMM_f	-0.1108	-0.1099	0.0166	0.1099	0.0654	-0.0509	2.9047	0.2860	0.0369	0.0013
ODR_{Λ_1}	-0.0005	-0.0008	0.0022	0.0319	0.0467	0.0210	2.8964	0.9530	0.0459	0.0025
s=0.5										
GMM_g	-0.0016	-0.0012	0.0022	0.0324	0.0467	-0.0374	2.9635	0.9515	0.0459	0.0019
GMM_h	-0.0838	-0.0827	0.0092	0.0827	0.0470	-0.0797	2.9642	0.5840	0.0465	0.0026
GMM_f	-0.0234	-0.0226	0.0028	0.0359	0.0473	-0.0714	2.9394	0.8995	0.0441	0.0017
ODR_{Λ_1}	-0.0141	-0.0131	0.0027	0.0346	0.0503	-0.1517	3.0821	0.9175	0.0455	0.0020
s=0.75										
GMM_g	-0.0010	-0.0002	0.0021	0.0314	0.0458	-0.1404	2.9735	0.9565	0.0459	0.0018
GMM_h	-0.0193	-0.0186	0.0027	0.0350	0.0483	-0.1429	3.0438	0.9355	0.0482	0.0028
GMM_f	-0.0054	-0.0044	0.0021	0.0313	0.0456	-0.1263	2.9815	0.9540	0.0452	0.0017
ODR_{Λ_1}	-0.0069	-0.0055	0.0022	0.0314	0.0461	-0.1310	3.0076	0.9490	0.0453	0.0017
Global										
GMM_g	-0.0010	-0.0003	0.0021	0.0314	0.0458	-0.1566	2.9735	0.9570	0.0459	0.0021
GMM_h	-0.2000	-0.2000	0.0428	0.2000	0.0529	0.0663	3.1573	0.0225	0.0495	0.0033
GMM_f	-0.0732	-0.0731	0.0084	0.0739	0.0554	0.0501	2.9813	0.5400	0.0402	0.0014
ODR_{Λ_1}	-0.0010	-0.0003	0.0021	0.0314	0.0458	-0.1563	2.9744	0.9570	0.0459	0.0021

weight asymptotically. Now suppose two but not all three models are right, e.g., suppose models G and H are right and L is wrong. Then all the weights in both the numerator and denominator of equation (4.14) go to zero. However, in this case we can divide the numerator and denominator by $\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$. Both $\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)$ and $\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$ converge to zero, but $n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)/n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$ is finite and nonzero, so the limiting weights on $\widehat{\alpha}_g$ and $\widehat{\alpha}_h$ will be nonzero while the limiting weight on $\widehat{\alpha}_l$ will be zero, as desired.

As with $SODR$, the limiting distribution of the triply robust estimator $\widetilde{\alpha}$ in equation (4.14) is complicated by the potential limiting randomness of ratios like $n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)/n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)$ in the weights. In the doubly robust case, we avoided this problem in ODR by using the additional weight W_f for when both models are correctly specified. An analogous construction for triply robust estimation would be more complicated, since we would also need to consider the cases where any pair of models is correct, and when all three are correct. This would require at least constructing an ODR for each of the three possible pairs of models, and for the model that combines all three.

Table 4.7-1: Model G is Misspecified with $s = 0.75$ and Model H is Misspecified ($n = 500$)

α_1	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
s=0.25										
GMM_g	0.0088	0.0093	0.0002	0.0102	0.0104	0.1033	3.0590	0.8355	0.0102	0.0010
GMM_h	0.2297	0.2292	0.0530	0.2292	0.0148	0.2036	2.9334	0.0000	0.0130	0.0015
GMM_f	0.1112	0.1108	0.0125	0.1108	0.0108	-0.1158	3.3286	0.0000	0.0065	0.0004
ODR_{Λ_1}	0.0088	0.0093	0.0002	0.0102	0.0104	0.1064	3.0616	0.8355	0.0102	0.0010
s=0.5										
GMM_g	0.0086	0.0087	0.0002	0.0098	0.0109	0.1853	3.1186	0.8600	0.0105	0.0008
GMM_h	0.0807	0.0805	0.0068	0.0805	0.0178	-0.0272	2.9853	0.0090	0.0171	0.0017
GMM_f	0.0276	0.0277	0.0009	0.0277	0.0095	0.1450	3.2462	0.1590	0.0088	0.0006
ODR_{Λ_1}	0.0222	0.0155	0.0011	0.0161	0.0254	0.6447	2.7902	0.6020	0.0108	0.0021
s=0.75										
GMM_g	0.0090	0.0090	0.0002	0.0101	0.0105	0.0550	3.0231	0.8565	0.0106	0.0008
GMM_h	0.0181	0.0185	0.0007	0.0198	0.0198	0.0769	2.8749	0.8185	0.0192	0.0022
GMM_f	0.0113	0.0113	0.0002	0.0115	0.0092	0.0496	3.1639	0.7720	0.0092	0.0006
ODR_{Λ_1}	0.0125	0.0120	0.0003	0.0123	0.0115	0.0250	3.0848	0.7445	0.0098	0.0018
Global										
GMM_g	0.0089	0.0092	0.0002	0.0102	0.0106	0.1271	3.0891	0.8520	0.0103	0.0009
GMM_h	0.1939	0.1926	0.0379	0.1926	0.0167	0.1383	3.1092	0.0000	0.0146	0.0017
GMM_f	0.0768	0.0766	0.0060	0.0766	0.0101	0.0503	2.8409	0.0000	0.0075	0.0005
ODR_{Λ_1}	0.0089	0.0092	0.0002	0.0102	0.0106	0.1241	3.0766	0.8500	0.0103	0.0009

Table 4.7-2: Model G is Misspecified with $s = 0.75$ and Model H is Misspecified ($n = 500$)

α_0	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD_{SE}
s=0.25										
GMM_g	-0.0083	-0.0071	0.0022	0.0315	0.0458	-0.1606	2.8672	0.9475	0.0445	0.0023
GMM_h	-0.2309	-0.2292	0.0560	0.2292	0.0524	-0.0060	3.0920	0.0030	0.0485	0.0032
GMM_f	-0.1115	-0.1098	0.0166	0.1098	0.0647	-0.0952	2.9328	0.2735	0.0363	0.0013
ODR_{Λ_1}	-0.0083	-0.0071	0.0022	0.0315	0.0458	-0.1605	2.8666	0.9475	0.0445	0.0023
s=0.5										
GMM_g	-0.0090	-0.0087	0.0022	0.0317	0.0455	-0.0878	2.9419	0.9485	0.0453	0.0018
GMM_h	-0.0811	-0.0807	0.0087	0.0807	0.0457	-0.0785	2.9850	0.5940	0.0459	0.0024
GMM_f	-0.0281	-0.0278	0.0029	0.0369	0.0455	-0.0613	2.9619	0.8930	0.0437	0.0016
ODR_{Λ_1}	-0.0225	-0.0204	0.0030	0.0351	0.0497	-0.2171	3.1523	0.9000	0.0449	0.0019
s=0.75										
GMM_g	-0.0100	-0.0091	0.0021	0.0321	0.0448	-0.0071	2.9514	0.9535	0.0455	0.0018
GMM_h	-0.0189	-0.0199	0.0027	0.0346	0.0481	-0.0238	2.9757	0.9310	0.0481	0.0029
GMM_f	-0.0122	-0.0122	0.0021	0.0318	0.0446	-0.0059	2.9843	0.9450	0.0449	0.0017
ODR_{Λ_1}	-0.0133	-0.0130	0.0022	0.0330	0.0453	0.0016	2.9476	0.9410	0.0450	0.0018
Global										
GMM_g	-0.0106	-0.0117	0.0021	0.0319	0.0450	-0.0511	2.9491	0.9475	0.0448	0.0020
GMM_h	-0.1952	-0.1941	0.0407	0.1941	0.0513	0.1066	3.2575	0.0200	0.0479	0.0031
GMM_f	-0.0785	-0.0784	0.0092	0.0785	0.0549	-0.0661	2.9737	0.5035	0.0396	0.0014
ODR_{Λ_1}	-0.0106	-0.0118	0.0021	0.0319	0.0450	-0.0508	2.9492	0.9475	0.0448	0.0020

4.9 Conclusions

In this paper, we provide a general technique for constructing doubly robust estimators. Our Over-identified Doubly Robust (ODR) technique is a simple extension of the Generalized Method of Moments. It takes the form of a weighted average of Hansen's (1982) Generalized Method of Moments (GMM) based estimators, and has similar associated root-n asymptotics. The proposed estimator appears to work well in a small Monte Carlo study and in an empirical application to instrumental variables estimation, where either one of two sets of instrument vectors might be invalid.

Our estimator requires that the candidate models be over-identified, having more moments than parameters. Ideally the number of moments should not greatly exceed the number of parameters, because GMM can suffer from well known finite sample biases when models have many more moments than parameters, and particularly when some moments might be weak. In such cases, it may be desirable to let models G and H equal just a subset of the available moments for each. Existing moment selection methods such as Andrews and Lu (2001), Caner (2009), or Liao (2013) might be used prior to applying ODR, though this then introduces pretest bias that ODR is intended to avoid. A potential subject for future work could be more formally modifying ODR to deal with many moments and/or with weak moments.

Another potential extension for future work is to consider cases where β and γ are infinite dimensional, e.g., where models G and H may contain unknown functions, perhaps replacing unconditional expectations with conditional expectations as in Ai and Chen (2003). One difficulty in such extensions is guaranteeing that the model is still over-identified regarding α, or more precisely, ensuring that no solution to all the moment conditions exists if the model is misspecified. Chen and Santos (2018) might be helpful regarding this point. Another issue would be ensuring that the objective functions used in constructing weights remain comparable and well behaved.

References

[1] Aczél, J., 1966: *Lectures on Functional Equations and Their Applications*. New York: Academic Press.

[2] Bell, D., 1982, "Regret in decision making," *Operations Research*, **30**, 961–981.

[3] Bikhchandani, S. and Segal, U., 2011, "Transitive regret," *Theoretical Economics*, **6**, 95–108.

[4] Bikhchandani, S. and Segal, U., 2014, "Transitive regret over statistically independent lotteries," *Journal of Economics Theory*, **152**, 237–248.

[5] Bleichrodt, H., Cillo, A. and Diecidue, E., 2010, "A quantitative measurement of regret theory," *Management Science*, **56**, 161–175.

[6] Bleichrodt, H. and Wakker, P., 2015, "Regret theory: a bold alternative to the alternatives," *The Economic Journal*, **125**, 493–532.

[7] Bordalo, P., Gennaioli, N. and Shleifer, A., 2012, "Salience Theory of Choice Under Risk," *The Quarterly Journal of Economics*, **127 (3)**, 1243–1285.

[8] Braun, M. and Munermann, A., 2004, "The impact of regret on the demand for insurance," *Journal of Risk and Insurance*, **71**, 737–767.

[9] Buturak, G. and Evren, O., 2017, "Choice overload and asymmetric regret," *Theoretical Economics*, **12**, 1029–1056.

[10] Diecidue, E. and Somasundaram, J., 2017, "Regret theory: a new foundation," *Journal of Economic Theory*, **172**, 88–119.

[11] Filiz-Ozbay, E. and Ozbay, E., 2017, "Auctions with anticipated regret: theory and experiment," *American Economic Review*, **97**, 1407.

[12] Fishburn, P., 1978, "Axioms for approval voting: Direct proof," *Journal of Economic Theory*, **19**, 180–185.

[13] Fishburn, P. and LaValle, I., 1988, "Context-Dependent Choice with Nonlinear and Nontransitive Preferences," *Econometrica*, **56**, 1221–1239.

[14] Gilboa, I. and Lapson, R., 1990, "Aggregation of Semiorders: Intransitive Indifference Makes a Difference," *Working Paper*.

[15] Grether, D. and Plott, C., 1979, "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review*, **69**, 623–638.

[16] Hayashi, T., 2008, "Regret aversion and opportunity-dependence," *Journal of Economic Theory*, **139**, 242–268.

[17] Herweg, F. and Muller, D., 2020, "A Comparison of Regret Theory and Salience Theory for Decisions under Risk," *CESifo Working Papers*.

[18] Kahneman, D. and Tversky, A., 1979, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, **47**, 263–292.

[19] Kreps, D., 1988: *Notes on The Theory of Choice*. Colorado: Westview Press.

[20] Levy, H., 2017, "Regret theory: state dominance and expected utility," *Journal of Mathmatical Psychology*, **79**, 1–12.

[21] Lichtenstein, S. and Slovic, P., 1971, "Reversal of Preference Between Bids and Choices in Gambling Decisions," *Journal of Experimental Psychology*, **89**, 46–55.

[22] Lindman, H., 1971, "Inconsistent preference among gambles," *Journal of Experimental Psychology*, **89**, 390–397.

[23] Loomes, G., Starmer, C., and Sugden, R., 1991, "Observing Violations of Transitivity by Experimental Methods," *Econometrica*, **59**, 425–439.

[24] Loomes, G. and Sugden, R. (1982), "Regret theory: an alternative theory of rational choice under uncertainty," *The Economic Journal*, **92**, 805–824.

[25] Loomes, G. and Sugden, R. (1987), "Some Implications of a More General Form of Regret Theory," *Journal of Economic Theory*, **41**, 270–287.

[26] Luce, D. and Raiffa H., 1956: *Games and Decisions*. New York: John Wiley & Sons.

[27] Maccheroni, F., Marinacci, M. and Rustichini, A., 2012. "Social decision theory: Choosing within and between groups," *The Review of Economic Studies*, **79**, 1591–1636.

[28] May, K., 1954, "Intransitivity, Utility, and Aggregation of Preference Patterns," *Econometrics*, **22**, 1–13.

[29] Michenaud, S. and Solnik, S. (2008), "Applying regret theory to investment choices: currency hedging decisions," *Journal of International Money and Finance*, **27**, 677–694.

[30] Milne, J. and Kelly, A. (2014), "A New Method for Boarding Passengers onto an Airplane," *Journal of Air Transport Management*, **34**, 93–100.

[31] Milne, J. and Salari, M. (2016), "Optimization of Assigning Passengers to Seats on Airplanes Based on Their Carry-on Luggage," *Journal of Air Transport Management*, **54**, 104–110.

[32] Muermann, A., Mitchell, O., and Volkman, J., 2006, "Regret, portfolio choice, and guarantees in defined contribution schemes," *Insurance Mathematics and Economics*, **39**, 219-229.

[33] Quiggin, J., 1994, "Regret theory with general choice set," *Journal of Risk and Uncertainty*, **8**, 153–165.

[34] Sarver, T., 2008, "Anticipating reget: why fewer options may be better," *Econometrica*, **76**, 263–305.

[35] Starmer, C. and Sugden, R. (1993), "Testing for juxtaposition and event-splitting effects," *Insurance Mathematics and Economics*, **13**, 165.

[36] Steffen, J. (2008), "Optimal Boarding Method for Airline Passengers," *Journal of Air Transport Management*, **14**, 146–150.

[37] Steffen, J. and Hotchkiss, J. (2012), "Experimental Test pf Airplane Boarding Methods," *Journal of Air Transport Management*, **18**, 64–67.

[38] Stoye, J., 2011, "Axioms for minimax regret choice correspondences," *Journal of Economic Theory*, **146**, 2226–2251.

[39] Sugden, R., 1993, "An axiomatic foundation for regret theory," *Journal of Economic Theory*, **60**, 159–180.

[40] Tversky, A., 1969, "Intransitivity of Preferences," *Psychological Review*, **76**, 31–48.

[41] Tversky, A., 1975, "A critique of Expected Utility Theory: Descriptive and Normative Considerations," *Erkenntnis*, **9**, 163–173.

[42] von Neumann, J. and Morgenstern O., 1944: *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

[43] Zhou, Z., 2021, "Redistribution Regret," *Working Paper*.

[44] Zhou, Z., 2021, "Ranking Regret," *Working Paper*.

[45] Ai, C. and Chen, X., 2003, "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions", *Econometrica*, 71(6), 1795-1843.

[46] Andrews, D.W.K., 1999, "Consistent moment selection procedures for generalized method of moments estimation", *Econometrica,* 67(3), 543-564.

[47] Andrews, D.W.K. and Lu, B., 2001, "Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models", *Journal of Econometrics*, 101(1), 123-164.

[48] Banks, J., Blundell, R., and Lewbel, A., 1997, "Quadratic Engel Curves and Consumer Demand", *Review of Economics and Statistics*, 79(4), 527-539.

[49] Bang, H., and Robins, J., 2005, "Doubly Robust Estimation in Missing Data and Causal Inference Models", *Biometrics*, 61(4), 962-973.

[50] Baum, C., and Schaffer, M., 2012, "IVREG2H: Stata Module to Perform Instrumental Variables Estimation Using Heteroskedasticity-based Instruments", Statistical Software Components S457555, Boston College Department of Economics, revised 18 Feb 2018.

[51] Bera, A. and Yoon, M., 1993, "Specification Testing with Locally Misspecified Alternatives", *Econometric Theory*, 9(4), 649-658.

[52] Campbell, J., and Cochrane, J., 1999, "By Force of Habit: A Consumption Based Explanation of Aggregate Stock Market Behavior", *Journal of Political Economy*, 107(2), 205-251.

[53] Caner, M., 2009, "Lasso-type GMM Estimator", *Econometric Theory*, 25(1), 270-290.

[54] Chen, X., and Ludvigson, S., 2009, "Land of Addicts? an Empirical Investigation of Habit-based Asset Pricing Models", *Journal of Applied Econometrics*, 24(7), 1057-1093.

[55] Chen, X. and Santos, A., 2018, "Overidentification in Regular Models", *Econometrica*, 86(5), 1771-1817.

[56] Chernozhukov, V., Escanciano, J.C., Ichimura, H., Newey, W. and Robins, J., 2018, "Locally Robust Semiparametric Estimation", Unpublished Manuscript.

[57] DiTraglia, F., 2016, "Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM", *Journal of Econometrics*, 195(2), 187-208.

[58] Funk, M., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M., and Davidian, M., 2011, "Doubly Robust Estimation of Causal Effects", *American Journal of Epidemiology*, 173(7), 761-7.

[59] Hall, A.R., 2000, "Covariance Matrix Estimation and the Power of the Overidentifying Restrictions Test", *Econometrica,* 68(6), 1517-1528.

[60] Hall, A.R. and Inoue, A., 2003, "The Large Sample Behaviour of the Generalized Method of Moments Estimator in Misspecified Models", *Journal of Econometrics,* 114(2), 361-394.

[61] Hansen, B., 2007, "Least Squares Model Averaging", *Econometrica*, 75(4), 1175-1189.

[62] Hansen, L., 1982, "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50(4), 1029-1054.

[63] Hansen, L., and Singleton, K., 1982, "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models", *Econometrica*, 50(5), 1269-1286.

[64] Kim, J-Y, 2002, "Limited information likelihood and Bayesian analysis", *Journal of Econometrics*, 107(1-2), 175-193.

[65] Kuersteiner, G. and Okui, R., 2010, "Constructing Optimal Instruments by First-Stage Prediction Averaging", *Econometrica*, 78(2), 697-718.

[66] Lee, M.J., and Lee, S., 2019, "Double Robustness Without Weighting", *Statistics and Probability Letters*, 146, 175-180.

[67] Lewbel, A., 2008, "Engel curves", entry for The New Palgrave Dictionary of Economics, 2nd Edition, MacMillan Press.

[68] Lewbel, A., 2012, "Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models", *Journal of Business and Economic Statistics*, 30(1), 67-80.

[69] Liao, Z., 2013, "Adaptive GMM Shrinkage Estimation With Consistent Moment Selection", *Econometric Theory*, 29(5), 857-904.

[70] Lunceford, J.K., and Davidian, M., 2004, "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: a Comparative Study", *Statistics in Medicine*, 23(19), 2937–2960.

[71] Martins, L.F., and Gabriel, V.J., 2014, "Linear Instrumental Variables Model Averaging Estimation", *Computational Statistics and Data Analysis*, 71, 709-724.

[72] Newey, W. and McFadden, D., 1994, "Chapter 36 Large Sample Estimation and Hypothesis Testing", in Handbook of Econometrics, 4, 2111-2245.

[73] Newey, W. and West, K., 1987, "Hypothesis Testing with Efficient Method of Moments Testing", *International Economic Review*, 28(3), 777-787.

[74] Okui, R., Small, D., Tan, Z., and Robins, J., 2012, "Doubly Robust Instrumental Variable Regression", *Statistica Sinica*, 22(1), 173-205.

[75] Pitman, E.T.G., 1949, "Notes on Nonparametric Statistical Inference", Manuscript.

[76] Robins, J., Rotnitzky, A., and Van Der Laan, M., 2000, "On Profile Likelihood: Comment", *Journal of the American Statistical Association*, 95(450), 477-482.

[77] Robins, J., Rotnitzky, A., and Zhao, L., 1994, "Estimation of Regression Coefficients When Some Regressors are not Always Observed", *Journal of the American Statistical Association*, 89(427), 846-866.

[78] Rose, S., and Van der Laan, M., 2014, "A Double Robust Approach to Causal Effects in Case-Control Studies", *American Journal of Epidemiology*, 179(6), 663-669.

[79] Scharfstein, D., Rotnitzky, A., and Robins, J., 1999, "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models", *Journal of the American Statistical Association*, 94(448), 1096-1120.

[80] Słoczyński, T., and Wooldridge, J., 2018, "A General Double Robustness Result for Estimating Average Treatment Effects", *Econometric Theory*, 34(01), 112-133.

[81] Sueishi, M., 2013, "Generalized Empirical Likelihood-Based Focused Information Criterion and Model Averaging", *Econometrics*, 1(2), 141-156.

[82] Wooldridge, J., 2007, "Inverse Probability Weighted Estimation for General Missing Data Problems", *Journal of Econometrics*, 141(2), 1281-1301.

A Proofs of Chapter 2

A.1 Lemma A.1

Lemma A.1 *A linear individual regret function implies that individual i follows expected utility theory, which is saying that the following two statements are equal.*

Statement 1: $\psi(x, y) = u(x) - u(y)$ *for some* $u(\cdot)$.

Statement 2: $\psi(x, y) + \psi(y, z) = \psi(x, z)$ *for any* $x > y > z \geq 0$.

Proof[21]**:**

1. Statement 1 $\Rightarrow$ Statement 2.

   Since Statement 1 is true, then the LHS of statement 2 equals to

$$\psi(x, y) + \psi(y, z)$$

$$= u(x) - u(y) + u(y) - u(z)$$

$$= u(x) - u(z) = \psi(x, z)$$

2. Statement 2 $\Rightarrow$ Statement 1.

   Since $\psi(x, y) + \psi(y, z) = \psi(x, z)$, take derivative of y on both sides, we have

$$\frac{\partial \psi(x, y)}{\partial y} + \frac{\partial \psi(y, z)}{\partial y} = 0$$

$$\Rightarrow \frac{\partial \psi(x, y)}{\partial y} = -\frac{\partial \psi(y, z)}{\partial y}$$

   Take derivative of x on both sides, we have

$$\Rightarrow \frac{\partial \psi(x, y)}{\partial y \partial x} = 0$$

$$\Rightarrow \int \frac{\partial \psi(x, y)}{\partial x \partial y} dx = c(y)$$

$$\Rightarrow \int \frac{\partial \psi(x, y)}{\partial y} dy = \int c(y) dy + d(x) = C(y) + d(x)$$

   Therefore, $\psi(x, y) = u_1(x) - u_2(y)$ for some $u_1(\cdot)$ and $u_2(\cdot)$. Plug this form into

[21] A shorter proof can be found in [1], page 223.

2, we have

$$\psi(x,y)+\psi(y,z)=\psi(x,z)$$

$$u_1(x)-u_2(y)+u_1(y)-u_2(z)=u_1(x)-u_2(z)$$

$$\Rightarrow u_1(y)-u_2(y)=0$$

for any $y \in [0,1]$, which implies $u_1(\cdot)=u_2(\cdot)=u(\cdot)$. Therefore, $\psi(x,y)=u(x)-u(y)$.

■

A.2 Proof of Proposition 2.1

Proof: Let $X=(x_1,x_2,\cdots,x_n)$, $Y=(y_1,y_2,\cdots,y_n)$, $Z=(z_1,z_2,\cdots,z_n)$, $\sum_i x_i=1$, $\sum_i y_i=1$ and $\sum_i z_i=1$. WLOG, let $0\le x_1<y_1<z_1\le 1$, and each value in x, y, z is in $[0,1]$. Then

$$W(Y,X)=\sum_i \psi_i(y_i,x_i)=\left[\frac{\psi_1(y_1,x_1)}{y_1-x_1}-1\right](y_1-x_1) \tag{A.1}$$

$$W(Z,Y)=\sum_i \psi_i(z_i,y_i)=\left[\frac{\psi_1(z_1,y_1)}{z_1-y_1}-1\right](z_1-y_1) \tag{A.2}$$

$$W(X,Z)=\sum_i \psi_i(x_i,z_i)=\left[\frac{\psi_1(x_1,z_1)}{x_1-z_1}-1\right](x_1-z_1) \tag{A.3}$$

Since $0\le x_1<y_1<z_1\le 1$, the signs of equations (1), (2), (3) depend on the signs of

$$\frac{\psi_1(y_1,x_1)}{y_1-x_1}-1 \tag{A.4}$$

$$\frac{\psi_1(z_1,y_1)}{z_1-y_1}-1 \tag{A.5}$$

$$\frac{\psi_1(x_1,z_1)}{x_1-z_1}-1 \tag{A.6}$$

The condition in Proposition 2.1 implies that for $x_1\neq y_1$, there are two cases: $\frac{\partial\frac{\psi_1(x_1,y_1)}{x_1-y_1}}{\partial x_1}>0$ and $\frac{\partial\frac{\psi_1(x_1,y_1)}{x_1-y_1}}{\partial y_1}>0$; or $\frac{\partial\frac{\psi_1(x_1,y_1)}{x_1-y_1}}{\partial x_1}<0$ and $\frac{\partial\frac{\psi_1(x_1,y_1)}{x_1-y_1}}{\partial y_1}<0$.

Case 1 implies that (4) $\le$ (6) since $\frac{\psi_1(y_1,x_1)}{y_1-x_1}=\frac{\psi_1(x_1,y_1)}{x_1-y_1}$, $\frac{\partial\frac{\psi_1(x_1,y_1)}{x_1-y_1}}{\partial y_1}>0$ and $y_1<z_1$. Similarly, (6) $\le$ (5). Therefore, (4) $\le$ (6) $\le$ (5).

To test whether it is transitive under any circumstances, consider the following situations:

1. None of (4), (5), (6) equal to 0

   (a) $(4) \leq (6) \leq (5) < 0$.

   The signs of equation (1), (2), (3) are $-\ -\ +$, which is transitive.

   (b) $0 < (4) \leq (6) \leq (5)$.

   The signs of equation (1), (2), (3) are $+\ +\ -$, which is transitive.

   (c) $(4) < 0 < (6) \leq (5)$.

   The signs of equation (1), (2), (3) are $-\ +\ -$, which is transitive.

   (d) $(4) \leq (6) < 0 < (5)$.

   The signs of equation (1), (2), (3) are $-\ +\ +$, which is transitive.

2. One of them equals to 0

   (a) $(4) = 0 < (6) \leq (5)$.

   The signs of equation (1), (2), (3) are $0\ +\ -$, which is transitive.

   (b) $(4) < (6) = 0 < (5)$.

   The signs of equation (1), (2), (3) are $-\ +\ 0$, which is transitive.

   (c) $(4) \leq (6) < (5) = 0$.

   The signs of equation (1), (2), (3) are $-\ 0\ +$, which is transitive.

If $x_1 = y_1$, we have (1) = 0, (2) and (3) have opposite signs, which is transitive. If $y_1 = z_1$, we have (2) = 0, (1) and (3) have opposite signs, which is transitive. If $x_1 = y_1 = z_1$, then (1) = (2) = (3) = 0, which is also transitive.

As we have exhausted all possibilities, case 1 is a sufficient condition. Similarly, we can also show that case 2 is a sufficient condition.

■

A.3 Proof of Proposition 2.2

Proof: Suppose there are three allocations $X = (x_1, x_2)$, $Y = (y_1, y_2)$ and $Z = (z_1, z_2)$, where $0 \leq x_1 \leq y_1 \leq z_1 \leq 1$, then

$$W(Y, X) = \psi(y_1, x_1) + \psi(y_2, x_2)$$

$$W(Z,Y) = \psi(z_1, y_1) + \psi(z_2, y_2)$$

$$W(X,Z) = \psi(x_1, z_1) + \psi(x_2, z_2)$$

Since $m = 1$, we have

$$W(Y,X) = \psi(y_1, x_1) + \psi(1 - y_1, 1 - x_1)$$

$$W(Z,Y) = \psi(z_1, y_1) + \psi(1 - z_1, 1 - y_1)$$

$$W(X,Z) = \psi(x_1, z_1) + \psi(1 - x_1, 1 - z_1)$$

1. If $x_1 < y_1 < z_1$.

   Consider the relationships among $x_1,\ 1-x_1,\ y_1,\ 1-y_1$, there are three situations:

   (a) $x_1 = 1 - y_1$ (or $y_1 = 1 - x_1$);

   (b) $x_1 > 1 - y_1$ (or $y_1 > 1 - x_1$);

   (c) $x_1 < 1 - y_1$ (or $y_1 < 1 - x_1$).

   Combining $x_1 < y_1 < z_1$, we have the following conclusions:

   (a) $x_1 = 1 - y_1$ and $y_1 = 1 - x_1$;

   By skew symmetric, we have

$$W(Y,X) = psi(1 - x_1, x_1) + \psi(x_1, 1 - x_1) = 0$$

   By $x_1 < z_1$, we have $1 - z_1 < 1 - x_1$, combining the condition mentioned in Proposition 2.1, we have

$$W(Z,Y) = psi(z_1, 1 - x_1) + \psi(1 - z_1, x_1)$$

$$= \left[\frac{\psi(z_1, 1 - x_1)}{z_1 - (1 - x_1)} - \frac{\psi(1 - z_1, x_1)}{(1 - z_1) - x_1}\right](z_1 - y_1) > 0$$

   Since $y_1 < z_1$, we have $z_1 > 1 - x_1$ and $x_1 > 1 - z_1$, by skew symmetric and the condition in Proposition 2.1,

$$W(X,Z) = psi(x_1, z_1) + \psi(1 - x_1, 1 - z_1)$$

$$= -\left[\frac{\psi(z_1, x_1)}{z_1 - x_1} - \frac{\psi(1 - z_1, 1 - x_1)}{(1 - z_1) - (1 - x_1)}\right](z_1 - x_1) < 0$$

   Therefore, it is transitive.

(b) $x_1 > 1 - y_1$ and $y_1 > 1 - x_1$;

Similarly, we have

$$W(Y, X) = psi(y_1, x_1) + \psi(1 - y_1, 1 - x_1)$$

$$= \left[\frac{\psi(y_1, x_1)}{y_1 - x_1} - \frac{\psi(1 - y_1, 1 - x_1)}{(1 - y_1) - (1 - x_1)} \right] (y_1 - x_1) > 0$$

By $z_1 > y_1$, we have $z_1 > 1 - x_1$ and $x_1 > 1 - z_1$, then

$$W(X, Z) = psi(x_1, z_1) + \psi(1 - x_1, 1 - z_1)$$

$$= - \left[\frac{\psi(z_1, x_1)}{z_1 - x_1} - \frac{\psi(1 - z_1, 1 - x_1)}{(1 - z_1) - (1 - x_1)} \right] (z_1 - x_1) < 0$$

As $x \to y$ and $z \to x$ have different signs, it is transitive.

(c) $x_1 < 1 - y_1$ and $y_1 < 1 - x_1$.

Similarly, $W(Y, X)$ is negative. Considering the relationship between z_1 and $1 - x_1$, there are two situations:

i. $1 - x_1 > z_1$.

We have $z_1 < 1 - x_1$ and $x_1 < 1 - z_1$, then $W(X, Z)$ is positive, hence, it is transitive.

ii. $z_1 > 1 - x_1$.

By $x_1 < y_1$, we have $z_1 > 1 - y_1$ and $y_1 > 1 - z_1$, then $W(Z, Y)$ is positive, hence, it is transitive.

2. Suppose $x_1 = y_1 < z_1$.

By skew symmetric, $y \to z$ and $z \to x$ have different signs, hence, it is transitive.

3. Suppose $x_1 < y_1 = z_1$.

By skew symmetric, $x \to y$ and $z \to x$ have different signs, hence, it is transitive.

4. Suppose $x_1 = y_1 = z_1$.

We have $x \sim y \sim z$, which is also transitive.

As we have considered all situations, and all of them are transitive, proved.

■

A.4 Proof of Proposition 2.3

Proof: Suppose the three persons have the same non-linear regret functions, $\psi_1(x,y) = \psi_2(x,y) = \psi_3(x,y) = \psi(x,y)$. And there are three allocations $X = (x_1, x_2, x_3, ...)$, $Y = (y_1, y_2, y_3, ...)$ and $Z = (z_1, z_2, z_3, ...)$. We have

$$W(Y,X) = \psi_1(y_1,x_1) + \psi_2(y_2,x_2) + \psi_3(y_3,x_3) + (y_4 - x_4) + ... + (y_n - x_n)$$

$$= \psi_1(y_1,x_1) + \psi_2(y_2,x_2) + \psi_3(y_3,x_3) + (1 - y_1 - y_2 - y_3) - (1 - x_1 - x_2 - x_3)$$

$$= [\psi(y_1,x_1) - (y_1 - x_1)] + [\psi(y_2,x_2) - (y_2 - x_2)] + [\psi(y_3,x_3) - (y_3 - x_3)]$$

Similarly, we have

$$W(Z,Y) = [\psi(z_1,y_1) - (z_1 - y_1)] + [\psi(z_2,y_2) - (z_2 - y_2)] + [\psi(z_3,y_3) - (z_3 - y_3)]$$

$$W(X,Z) = [\psi(x_1,z_1) - (x_1 - z_1)] + [\psi(x_2,z_2) - (x_2 - z_2)] + [\psi(x_3,z_3) - (x_3 - z_3)]$$

We can always find $x_1 \leq y_1 \leq z_1$, $x_1 = z_2 = y_3$, $y_1 = x_2 = z_3$ and $z_1 = y_2 = x_3$, which leads to the same equations for $W(Y,X)$, $W(Z,Y)$ and $W(X,Z)$.

$$[\psi(y_1,x_1) - (y_1 - x_1)] + [\psi(z_1,y_1) - (z_1 - y_1)] + [\psi(x_1,z_1) - (x_1 - z_1)]$$

Hence, they have the same number which is not necessarily 0, and it means non-transitivity.

■

A.5 Lemma A.2

Lemma A.2 *The following two statements are equal.*
Statement 1*: For any $0 \leq x_0 < x_1$, if x_2 is such that $f(x_2,x_1) = f(x_1,x_0)$, then $f(x_2,x_0) = 2f(x_1,x_0)$. Create a sequence $\{x_i\}$, where $i \in \{0,1,2,3,...,2^n\}$, $x_0 < x_1 < x_2 < ... < x_{2^n}$, satisfying*

$$f(x_1,x_0) = f(x_2,x_1)... = f(x_{2^n},x_{2^n-1})$$

If for some k, $0 \leq k \leq 2^n$, such that $f(x_k,x_0) = kf(x_1,x_0)$, then $f(x_{2^n},x_k) = (2^n - k)f(x_1,x_0)$.
Statement 2*: $f(x,y) = u(x) - u(y)$.*

Proof:

1. Statement 1 $\Rightarrow$ Statement 2.

   First, I want to show that, for any n, we have $f(x_k, x_0) = kf(x_1, x_0)$ for any $0 < k \leq 2^n$.

   We prove it by induction on n.

   (a) For $n = 1$, then $k = 0, 1, 2$, it is true as $f(x_0, x_0) = 0f(x_1, x_0)$, $f(x_1, x_0) = 1f(x_1, x_0)$, and $f(x_2, x_0) = 2f(x_1, x_0)$.

   (b) Want to show that if the statement is true for $n-1$, it is also true for n.

   By induction, if $n-1$ is true, $f(x_k, x_0) = kf(x_1, x_0)$ for $k = 1, 2, 3, ..., 2^{n-1}$, we want to show that $f(x_k, x_0) = kf(x_1, x_0)$ for $k = 2^{n-1} + 1, ..., 2^n$.

   Consider a sequence $\{y_k\}$, such that $y_0 = x_{2^n}$, $y_1 = x_{2^n-1}$, ..., $y_k = x_{2^n-k}$, ..., $y_{2^n-1} = x_1$, $y_{2^n} = x_0$.

   Since $n-1$ is true for any sequence, we have $f(y_k, y_0) = kf(y_1, y_0)$ for $k = 1, 2, 3, ..., 2^{n-1}$.

   Then for $k = 1, 2, 3, ..., 2^{n-1}$,

$$f(y_{2^n}, y_k) = (2^n - k)f(y_1, y_0)$$

$$\Rightarrow f(x_0, x_{2^n-k}) = (2^n - k)f(x_{2^n-1}, x_{2^n})$$

$$\Rightarrow f(x_{2^n-k}, x_0) = (2^n - k)f(x_1, x_0)$$

   Please note that $2^n - k \in \{2^{n-1}, ..., 2^n - 1\}$, hence, the statement is also true for n.

Therefore, for any n, we have $f(x_k, x_0) = kf(x_1, x_0)$ for $0 < k \leq 2^n$.

Pick any $0 \leq x < y$ and fix them. For any $z \in [x, y]$, create a function:

$$g(z) = f(z, x) + f(y, z) - f(y, x)$$

We can find $z_{\frac{1}{2}}$ such that

$$f(z_{\frac{1}{2}}, x) = f(y, z_{\frac{1}{2}})$$

then

$$g(z_{\frac{1}{2}}) = 0$$

In the meantime, we have $g(x) = 0$, $g(y) = 0$.

If I can find a dense set M, such that $g(z) = 0$ for any $z \in M$, then $g(z) = 0$ for any $z \in [x, y]$. As x and y are arbitrary, by Lemma A.1, it is proved.

Define $z_{\frac{k}{2^n}}$, where n is any positive integer, $0 < k \leq 2^n$, such that

$$f(z_{\frac{k+1}{2^n}}, z_{\frac{k}{2^n}}) = f(z_{\frac{1}{2^n}}, x)$$

$$= \frac{1}{2^n} f(y, x)$$

By the conclusion of Statement 1, we have

$$g(z_{\frac{k}{2^n}}) = f(z_{\frac{k}{2^n}}, x) + f(y, z_{\frac{k}{2^n}}) - f(y, x)$$

$$= kf(z_{\frac{1}{2^n}}, x) + (2^n - k)f(z_{\frac{1}{2^n}}, x) - 2^n f(z_{\frac{1}{2^n}}, x) = 0$$

then it is proved, as $\{z_{\frac{k}{2^n}}\}$ is a dense set.

2. Statement 2 $\Rightarrow$ Statement 1.

   As $f(x, y) = u(x) - u(y)$, we have

$$f(x_2, x_0) = u(x_2) - u(x_0)$$

$$= u(x_2) - u(x_1) + u(x_1) - u(x_0)$$

$$= f(x_2, x_1) + f(x_1, x_0)$$

$$= 2f(x_1, x_0)$$

   As $f(x, y) = u(x) - u(y)$, we have

$$f(x_{2^n}, x_k) = u(x_{2^n}) - u(x_k)$$

$$= u(x_{2^n}) - u(x_{2^n-1}) + ... + u(x_{k+1}) - u(x_k)$$

$$= f(x_{2^n}, x_{2^n-1}) + ... + f(x_{k+1}, x_k)$$

$$= (2^n - k)f(x_1, x_0)$$

■

A.6 Proposition A.1

Proposition A.1 *Suppose the budget is variable, and the population $n \geq 2$. For any $1 \leq m < n$, if person i has general regret function $\psi_i(x, y)$ for any $i \in \{1, .., m\}$ while*

$\psi_j(x,y) = x - y$ *for any* $j \in \{m+1, ..., n\}$. *If the order*

$$(x_1, ..., x_n) \succsim (y_1, ..., y_n) \Longleftrightarrow \sum \psi_i(x_i, y_i) \geq 0$$

is transitive, then $\psi_i(x,y) = u_i(x) - u_i(y)$ *for any* $i \in \{1, ..., n\}$.

Proof: If the budgets for the distributions

$$X = (x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_m, x_{m+1}, ...x_n)$$

$$Y = (x_1, ..., x_{i-1}, y_i, x_{i+1}, ..., x_m, y_{m+1}, ...y_n)$$

$$Z = (x_1, ..., x_{i-1}, z_i, x_{i+1}, ..., x_m, z_{m+1}, ...z_n)$$

are $m_x,\ m_y,\ m_z$, then

$$W(Y,X) = \psi_i(y_i, x_i) + (m_y - y_i) - (m_x - x_i)$$

$$W(Z,Y) = \psi_i(z_i, y_i) + (m_z - z_i) - (m_y - y_i)$$

$$W(X,Z) = \psi_i(x_i, z_i) + (m_x - x_i) - (m_z - z_i)$$

which can be written as:

$$W(Y,X) = (m_y - m_x) + \psi_i(y_i, x_i) - (y_i - x_i) \tag{A.7}$$

$$W(Z,Y) = (m_z - m_y) + \psi_i(z_i, y_i) - (z_i - y_i) \tag{A.8}$$

$$W(X,Z) = (m_x - m_z) + \psi_i(x_i, z_i) - (x_i - z_i) \tag{A.9}$$

For any $(x_i,\ y_i,\ z_i)$, we can find $m_y - m_x$ and $m_z - m_y$ such that equations (7) and (8) equal 0. To guarantee transitivity, the third equation has to be 0, which means (7) + (8) + (9) = 0, then we have

$$(m_x - m_z) + \psi_i(x_i, z_i) - (x_i - z_i)$$

$$+(m_y - m_x) + \psi_i(y_i, x_i) - (y_i - x_i)$$

$$+(m_z - m_y) + \psi_i(z_i, y_i) - (z_i - y_i) = 0$$

$$\Rightarrow \psi_i(y_i, x_i) + \psi_i(z_i, y_i) = \psi_i(z_i, x_i)$$

for any (x_i, y_i, z_i) and for any $i \in \{1, ..., m\}$. By Lemma A.1, $\psi_i(x,y) = u_i(x) - u_i(y)$

for any $i \in \{1, ..., n\}$.

■

A.7 Proof of Theorem 2.1

Proof: Suppose not, by Proposition A.1, every person i has a non-linear regret function. There are two cases.

Case 1: there exist some persons i, some $a > 0$, and some triples (x, y, z), where $z > y > x \geq 0$, such that $\psi_i(y, x) = \psi_i(z, y) = a$ and $\psi_i(z, x) \neq 2a$.

Suppose there exist some a and some triples (x_1, y_1, z_1), where $z_1 > y_1 > x_1 > 0$ such that $\psi_1(y_1, x_1) = \psi_1(z_1, y_1) = a$ and $\psi_1(z_1, x_1) > (<)2a$. By continuity, we can find two other triples (x_2, y_2, z_2) and (x_3, y_3, z_3) such that $\psi_2(y_2, x_2) = \psi_2(z_2, y_2) = \psi_3(y_3, x_3) = \psi_3(z_3, y_3) = a$. There are two cases:

1. Three persons are the same.

   WLOG, suppose $\psi_2(z_2, x_2) \geq (\leq)2a$ and $\psi_3(z_3, x_3) \geq (\leq)2a$.

   Let three allocations be $X = (x_1, z_2, y_3, x_4, ..., x_n)$, $Y = (y_1, x_2, z_3, x_4, ..., x_n)$ and $Z = (z_1, y_2, x_3, x_4, ..., x_n)$, we have

   $$W(Y, X) = \psi_1(y_1, x_1) + \psi_2(x_2, z_2) + \psi_3(z_3, y_3)$$

   $$W(Z, Y) = \psi_1(z_1, y_1) + \psi_2(y_2, x_2) + \psi_3(x_3, z_3)$$

   $$W(X, Z) = \psi_1(x_1, z_1) + \psi_2(z_2, y_2) + \psi_3(y_3, x_3)$$

   Then

   $$W(Y, X) = 2a - \psi_2(z_2, x_2) \leq (\geq)0$$

   $$W(Z, Y) = 2a - \psi_3(z_3, x_3) \leq (\geq)0$$

   $$W(X, Z) = 2a - \psi_1(z_1, x_1) < (>)0$$

   Hence, the signs for $W(Y, X)$, $W(Z, Y)$, and $W(X, Z)$ indicate a violation of transitivity.

2. Two persons are different.

   WLOG, suppose $\psi_2(z_2, x_2) \leq (\geq)2a$. Again, suppose there are three allocations $X = (x_1, z_2, x_3, ..., x_n)$, $Y = (y_1, y_2, x_3, ..., x_n)$ and $Z = (z_1, x_2, x_3, ..., x_n)$, we

have

$$W(Y,X) = \psi_1(y_1,x_1) + \psi_2(y_2,z_2) = 0$$

$$W(Z,Y) = \psi_1(z_1,y_1) + \psi_2(x_2,y_2) = 0$$

$$W(X,Z) = \psi_1(x_1,z_1) + \psi_2(z_2,x_2) < (>) - 2a + 2a = 0$$

Then the signs for $W(Y,X)$, $W(Z,Y)$, and $W(X,Z)$ also indicate a violation of transitivity.

Case 2: for each person, we cannot find any triple (x,y,z), where $z > y > x \geq 0$ such that $\psi(y,x) = \psi(z,y) = a$ and $\psi(z,x) \neq 2a$.

If so, their regret functions satisfy the first part of the Statement 1 in Lemma A.2. By Lemma A.2, to guarantee non-linearity, they have to violate the second part of the Statement 1.

Create a sequence $\{x_1^i\}$, where $i = 1, 2, ...$, such that $f(x_1^1, x_1) = ... = f(x_1^{i+1}, x_1^i) = a$. By Lemma 2, there exist k and $0 \leq x_1 < y_1 = x_1^k < z_1 = x_1^{2^n}$, such that $f(y_1,x_1) = ka$, $f(z_1,x_1) = 2^n a$ and $f(z_1,y_1) > (<)(2^n - k)a$. By continuity, we can find two other triples (x_2,y_2,z_2) and (x_3,y_3,z_3) such that $\psi_2(y_2,x_2) = \psi_3(y_3,x_3) = ka$, $\psi_2(z_2,x_2) = \psi_3(z_3,x_3) = 2^n a$. There are two cases:

1. Three persons are the same.

   WLOG, suppose $\psi_2(z_2,y_2) \geq (\leq)(2^n - k)a$ and $\psi_3(z_3,y_3) \geq (\leq)(2^n - k)a$.

   Let three allocations be $X = (x_1, z_2, y_3, x_4, ..., x_n)$, $Y = (y_1, x_2, z_3, x_4, ..., x_n)$ and $Z = (z_1, y_2, x_3, x_4, ..., x_n)$, we have

   $$W(Y,X) = \psi_1(y_1,x_1) + \psi_2(x_2,z_2) + \psi_3(z_3,y_3)$$

   $$W(Z,y) = \psi_1(z_1,y_1) + \psi_2(y_2,x_2) + \psi_3(x_3,z_3)$$

   $$W(X,Z) = \psi_1(x_1,z_1) + \psi_2(z_2,y_2) + \psi_3(y_3,x_3)$$

   Then

   $$W(Y,X) = -(2^n - k)a + \psi_3(z_3,y_3) \geq (\leq)0$$

   $$W(Z,y) = -(2^n - k)a + \psi_1(z_1,y_1) > (<)0$$

   $$W(X,Z) = -(2^n - k)a + \psi_2(z_2,y_2) \geq (\leq)0$$

   Hence, the signs for $W(Y,X)$, $W(Z,Y)$, and $W(X,Z)$ indicate a violation of transitivity.

2. Two persons are different.

   WLOG, suppose $\psi_2(z_2, y_2) \leq (\geq)(2^n - k)a$.

   We can find $x_1' = x_1^{2^n+k}$ such that $f(x_1', z_1) = ak$, then $f(x_1', y_1) = 2^n a$. Suppose there are three allocations $X = (x_1', x_2, x_3, ..., x_n)$, $Y = (z_1, y_2, x_3, ..., x_n)$ and $Z = (y_1, z_2, x_3, ..., x_n)$, we have

$$W(Y, X) = \psi_1(z_1, x_1') + \psi_2(y_2, x_2) = 0$$

$$W(Z, Y) = \psi_1(y_1, z_1) + \psi_2(z_2, y_2) < (>)0$$

$$W(X, Z) = \psi_1(x_1', y_1) + \psi_2(x_2, z_2) = 0$$

   Then the signs for $W(Y, X)$, $W(Z, Y)$, and $W(X, Z)$ also indicate a violation of transitivity.

Therefore, to guarantee transitivity, no persons have non-linear regret functions.

■

A.8 Proof of Example 2.3

Proof: Suppose there are three allocations $X = (x_1, x_2)$, $Y = (y_1, y_2)$ and $Z = (z_1, z_2)$. WLOG, we assume $x_1 \leq y_1 \leq z_1$, then

$$W(Y, X) = (y_1 - x_1)^3 + (y_2 - x_2)^3$$

$$W(Z, Y) = (z_1 - y_1)^3 + (z_2 - y_2)^3$$

$$W(X, Z) = (x_1 - z_1)^3 + (x_2 - z_2)^3$$

Since the relationship among (x_2, y_2, z_2) is arbitrary, we need to consider the following six conditions:

1. $z_2 \geq y_2 \geq x_2$.

   The signs for $W(Y, X)$, $W(Z, Y)$ and $W(X, Z)$ will be $+, +, -$, which means it is transitive.

2. $z_2 \geq x_2 \geq y_2$.

   The signs will be $?, +, -$, which means it is transitive.

3. $y_2 \geq z_2 \geq x_2$,

   The signs will be $+, ?, -$, which means it is transitive.

4. $x_2 \geq y_2 \geq z_2$.

   Consider the following three possibilities:

   (a) The signs of $W(Y, X)$ and $W(Z, Y)$ are different. Then it is transitive.

   (b) The signs are $+, +$. Then we have $y_1 - x_1 > x_2 - y_2$ and $z_1 - y_1 > y_2 - z_2$, which indicates $z_1 - x_1 > x_2 - z_2$, therefore, the sign of $W(X, Z)$ is $-$. In sum, the signs are $+, +, -$, which means it is transitive.

   (c) The signs are $-, -$. Similarly, we have $y_1 - x_1 < x_2 - y_2$ and $z_1 - y_1 < y_2 - z_2$, which indicates $z_1 - x_1 < x_2 - z_2$, therefore, the sign of $W(X, Z)$ is $+$. In sum, the signs are $-, -, +$, which means it is transitive.

5. $x_2 \geq z_2 \geq y_2$.

   So far, we know the signs are $?, +, ?$. Consider the following two possibilities:

   (a) The signs are $-, +, ?$. Then it is transitive.

   (b) The signs are $+, +, ?$. We can write

$$W(X, Z) = (x_1 - y_1 + y_1 - z_1)^3 + (x_2 - y_2 + y_2 - z_2)^3$$

$$= (x_1 - y_1)^3 + (y_1 - z_1)^3 + (x_2 - y_2)^3 + (y_2 - z_2)^3$$

$$+3(x_1 - y_1)(y_1 - z_1)(x_1 - z_1) + 3(x_2 - y_2)(y_2 - z_2)(x_2 - z_2)$$

   As $(x_1 - y_1)^3 + (y_1 - z_1)^3 + (x_2 - y_2)^3 + (y_2 - z_2)^3$ is the opposite of $W(Y, X)$ plus $W(Z, Y)$, it is negative. In addition, $(x_1 - y_1)(y_1 - z_1)(x_1 - z_1)$ is negative and $(x_2 - y_2)(y_2 - z_2)(x_2 - z_2)$ is also negative, so the signs are $+, +, -$, which is transitive.

6. $y_2 \geq x_2 \geq z_2$.

   So far, we know the signs are $+, ?, ?$. Consider the following two possibilities:

   (a) The signs are $+, -, ?$. Then it is transitive.

   (b) The signs are $+, +, ?$. Similarly, as the sign of $(x_2 - y_2)(y_2 - z_2)(x_2 - z_2)$ is negative, the signs are $+, +, -$, which is transitive.

As we have exhausted the possibilities, it is transitive.

■

B Proofs of Chapter 3

B.1 Lemmas

B.1.1 Lemma B.1

Lemma B.1 If

$$\begin{aligned} W(X,Y) &= V(\ldots,0,\psi_{i1}(x_{i1},y_{i1}),\ldots,\psi_{ig}(x_{ig},y_{ig}),0,\ldots,0, \\ &\qquad \psi_{j1}(x_{j1},y_{j1}),\ldots,\psi_{jg}(x_{jg},y_{jg}),0,\ldots) \end{aligned}$$

and $\psi_{ik}(x_{ik},y_{ik}) > 0 > \psi_{jk}(x_{jk},y_{jk})$ for any $k \in \{1,\ldots,g\}$, then

$$|\psi_{ik}(x_{ik},y_{ik})| > |\psi_{jk}(x_{jk},y_{jk})| \text{ for any } k \in \{1,\ldots,g\} \Longleftrightarrow X \succ Y$$

and

$$|\psi_{ik}(x_{ik},y_{ik})| < |\psi_{jk}(x_{jk},y_{jk})| \text{ for any } k \in \{1,\ldots,g\} \Longleftrightarrow Y \succ X$$

Proof: Given

$$W(X,Y) = V(\ldots,0,\psi_{i1}(x_{i1},y_{i1}),\ldots,\psi_{ig}(x_{ig},y_{ig}),0,\ldots,0,$$

$$\psi_{j1}(x_{j1},y_{j1}),\ldots,\psi_{jg}(x_{jg},y_{jg}),0,\ldots)$$

we have

$$W(Y,X) = V(\ldots,0,\psi_{i1}(y_{i1},x_{i1}),\ldots,\psi_{ig}(y_{ig},x_{ig}),0,\ldots,0,$$

$$\psi_{j1}(y_{j1},x_{j1}),\ldots,\psi_{jg}(y_{jg},x_{jg}),0,\ldots)$$

As the functions $\psi(x,y)$ are skew symmetric,

$$W(Y,X) = V(\ldots,0,-\psi_{i1}(x_{i1},y_{i1}),\ldots,-\psi_{ig}(x_{ig},y_{ig}),0,\ldots,0,$$

$$-\psi_{j1}(x_{j1},y_{j1}),\ldots,-\psi_{jg}(x_{jg},y_{jg}),0,\ldots)$$

If $\psi_{ik}(x_{ik}, y_{ik}) = -\psi_{jk}(x_{jk}, y_{jk})$ for any $k \in \{1, \ldots, g\}$,

$$W(Y, X) = V(\ldots, 0, \psi_{j1}(x_{j1}, y_{j1}), \ldots, \psi_{jg}(x_{jg}, y_{jg}), 0, \ldots, 0,$$

$$\psi_{i1}(x_{i1}, y_{i1}), \ldots, \psi_{ig}(x_{ig}, y_{ig}), 0, \ldots)$$

By anonymity,

$$W(Y, X) = V(\ldots, 0, \psi_{i1}(x_{i1}, y_{i1}), \ldots, \psi_{ig}(x_{ig}, y_{ig}), 0, \ldots, 0,$$

$$\psi_{j1}(x_{j1}, y_{j1}), \ldots, \psi_{jg}(x_{jg}, y_{jg}), 0, \ldots) = W(X, Y)$$

Therefore,

$$\psi_{ik}(x_{ik}, y_{ik}) = -\psi_{jk}(x_{jk}, y_{jk}) \text{ for any } k \in \{1, \ldots, g\} \Longleftrightarrow X \sim Y$$

By monotonicity and continuity, if $\psi_{ik}(x_{ik}, y_{ik}) > 0 > \psi_{jk}(x_{jk}, y_{jk})$ for any $k \in \{1, \ldots, g\}$, we have

$$|\psi_{ik}(x_{ik}, y_{ik})| > |\psi_{jk}(x_{jk}, y_{jk})| \text{ for any } k \in \{1, \ldots, g\} \Longleftrightarrow X \succ Y$$

$$|\psi_{ik}(x_{ik}, y_{ik})| < |\psi_{jk}(x_{jk}, y_{jk})| \text{ for any } k \in \{1, \ldots, g\} \Longleftrightarrow Y \succ X$$

■

B.1.2 Lemma B.2

Lemma B.2 Suppose $0 < \alpha_1 \leqslant \ldots \leqslant \alpha_n$. Given two rankings and $(\alpha_1, \ldots, \alpha_k)$, if $V(\psi_1, \ldots, \psi_k, 0, \ldots, 0) > 0$, $\psi_{k+1} < 0, \ldots,$ and $\psi_{k+j} < 0$, where $1 \leqslant j \leqslant n - k$ then there exists α^*_{k+1}, such that

$$V(\psi_1, \ldots, \psi_k, \psi_{k+1}(\alpha_{k+1}), \ldots, \psi_{k+j}(\alpha_{k+1}), 0, \ldots, 0) > 0$$

for any $\alpha_{k+1} > \alpha^*_{k+1}$.

Proof: By monotonicity and continuity, either for all $\alpha_{k+1} \geqslant \alpha_k$,

$$V(\psi_1, \ldots, \psi_k, \psi_{k+1}(\alpha_{k+1}), \ldots, \psi_{k+j}(\alpha_{k+1}), 0, \ldots, 0) > 0$$

or there exists $\alpha^*_{k+1,2} \geqslant \alpha_k$, such that

$$V(\psi_1, \ldots, \psi_k, \psi_{k+1}(\alpha^*_{k+1,2}), \ldots, \psi_{k+j}(\alpha^*_{k+1,2}), 0, \ldots, 0) = 0$$

If it is the first case, let $\alpha^*_{k+1} = \alpha_k$; if it is the second case, let $\alpha^*_{k+1} = \alpha^*_{k+1,2}$. Then for any $\alpha_{k+1} > \alpha^*_{k+1}$, we have

$$V(\psi_1, \ldots, \psi_k, \psi_{k+1}(\alpha_{k+1}), \ldots, \psi_{k+j}(\alpha_{k+1}), 0, \ldots, 0) > 0$$

■

B.2 Proof of Claim 3.1

There are six possible rankings, $\{D_1, \ldots, D_6\}$, as is shown in the table.

Table B.1: Possible Rankings

$D_1 = (\frac{1}{3}, \frac{2}{3}, 1)$	$D_2 = (\frac{1}{3}, 1, \frac{2}{3})$	$D_3 = (\frac{2}{3}, \frac{1}{3}, 1)$
$D_4 = (\frac{2}{3}, 1, \frac{1}{3})$	$D_5 = (1, \frac{1}{3}, \frac{2}{3})$	$D_6 = (1, \frac{2}{3}, \frac{1}{3})$

Consider a 6×6 matrix. Each element (i, j) in the matrix tells the relationship between D_i and D_j. "+" in (i, j) says $D_i \succ D_j$; "0" means $D_i \sim D_j$; "−" indicates $D_j \succ D_i$; and "?" stands for an unknown relationship between D_i and D_j.

Given a matrix, if there is no "?", we can tell whether the preference order is transitive by the following steps.

Step 1: Check each row. If there are only 0 or + in the i row, then D_i is the best ranking. If multiple rows satisfy this condition, and all of the rows are exactly the same, then the corresponding rankings are the best and they are indifferent. If no row satisfies or multiple satisfying rows are different, then the preference orders are not transitive.

Step 2: If i row satisfies the above condition, then remove i row and i column, and get a 5×5 matrix. Repeat checking and removing. If we can remove till there is no element left, then the preference orders are transitive. Otherwise, they are not.

Consider relative definitions. By Lemma 1, we have

$$W(D_1, D_2) = V\left(\psi\left(\frac{2}{3}, 1, \alpha_2\right), \psi\left(1, \frac{2}{3}, \alpha_3\right)\right) > 0$$

Similarly, we can get Matrix 1.

For those ? in the matrix, we have

$$W(D_1, D_4) = V\left(\psi\left(\frac{1}{3}, \frac{2}{3}, \alpha_1\right), \psi\left(\frac{2}{3}, 1, \alpha_2\right), \psi\left(1, \frac{1}{3}, \alpha_3\right)\right)$$

$$W(D_3, D_6) = V\left(\psi\left(\frac{2}{3}, 1, \alpha_1\right), \psi\left(\frac{1}{3}, \frac{2}{3}, \alpha_2\right), \psi\left(1, \frac{1}{3}, \alpha_3\right)\right)$$

Table B.2: Matrix 1

	1	2	3	4	5	6
1	0	+	+	?	?	+
2	−	0	?	+	+	?
3	−	?	0	+	+	?
4	?	−	−	0	?	+
5	?	−	−	?	0	+
6	−	?	?	−	−	0

$$W(D_1, D_5) = V\left(\psi\left(\frac{1}{3}, 1, \alpha_1\right), \psi\left(\frac{2}{3}, \frac{1}{3}, \alpha_2\right), \psi\left(1, \frac{2}{3}, \alpha_3\right)\right)$$

$$W(D_2, D_6) = V\left(\psi\left(\frac{1}{3}, 1, \alpha_1\right), \psi\left(1, \frac{2}{3}, \alpha_2\right), \psi\left(\frac{2}{3}, \frac{1}{3}, \alpha_3\right)\right)$$

$$W(D_2, D_3) = V\left(\psi\left(\frac{1}{3}, \frac{2}{3}, \alpha_1\right), \psi\left(1, \frac{1}{3}, \alpha_2\right), \psi\left(\frac{2}{3}, 1, \alpha_3\right)\right)$$

$$W(D_4, D_5) = V\left(\psi\left(\frac{2}{3}, 1, \alpha_1\right), \psi\left(1, \frac{1}{3}, \alpha_2\right), \psi\left(\frac{1}{3}, \frac{2}{3}, \alpha_3\right)\right)$$

By Lemma 1 and 2, given $\{\alpha_1, \alpha_2\}$, we can find α_3^* such that $W(D_1, D_4)$, $W(D_3, D_6)$, $W(D_1, D_5)$ and $W(D_2, D_6)$ are positive for any $\alpha_3 > \alpha_3^*$. Now we have Matrix 2.

Table B.3: Matrix 2

	1	2	3	4	5	6
1	0	+	+	+	+	+
2	−	0	?	+	+	+
3	−	?	0	+	+	+
4	−	−	−	0	?	+
5	−	−	−	?	0	+
6	−	−	−	−	−	0

By Matrix 2, we have

$$1 \succ (2, 3) \succ (4, 5) \succ 6$$

Note that the elements $(2, 3)$ and $(3, 2)$, and the elements $(4, 5)$ and $(5, 4)$ have opposite signs. Therefore, the preference order is transitive no matter what the signs of $W(D_2, D_3)$ and $W(D_4, D_5)$ are. For example,

1. If $W(D_2, D_3) \geqslant 0,\ W(D_4, D_5) \geqslant 0$, then

$$1 \succ 2 \succsim 3 \succ 4 \succsim 5 \succ 6$$

2. If $W(D_2, D_3) \geqslant 0,\ W(D_4, D_5) < 0$, then

$$1 \succ 2 \gtrsim 3 \succ 5 \succ 4 \succ 6$$

3. If $W(D_2, D_3) < 0,\ W(D_4, D_5) \geqslant 0$, then

$$1 \succ 3 \succ 2 \succ 4 \gtrsim 5 \succ 6$$

4. If $W(D_2, D_3) < 0,\ W(D_4, D_5) < 0$, then

$$1 \succ 3 \succ 2 \succ 5 \succ 4 \succ 6$$

■

B.3 Proof of Theorem 3.1

Let Ω be the set of all possible rankings, and $\mathcal{Y}$ be the set $\Omega \setminus X^*$. Define set $\mathcal{Y}_k$ as all possible rankings where their differences between X^* only involve the first $k \in \{2, \ldots, n\}$ individuals. Note that for each ranking $Y_k^i \in \mathcal{Y}_k$, the k^{th} person's ranking has to be different from X^*. Then $\{\mathcal{Y}_2, \ldots, \mathcal{Y}_n\}$ are independent sets and $\mathcal{Y} = \mathcal{Y}_2 \cup \mathcal{Y}_3 \cup \ldots \cup \mathcal{Y}_n$.

I need to find conditions such that if they are satisfied, then $X^* \succ Y_k^i$ for any $Y_k^i \in \mathcal{Y}_k$ and $k \in \{2, \ldots, n\}$; and if they are not satisfied, there exists at least one ranking preferred to X^*.

1. If $k = 2$, $X^* \succ Y_2^1$ is always true as $\alpha_1 < \alpha_2$ is given.

2. If $k = 3$, we have $Y_3 = \{Y_3^1, \ldots, Y_3^4\}$. To simplify, I use relative definitions, but the proof also applies to absolute definitions.

$$Y_3^1 = \left(\frac{2}{n}, \frac{3}{n}, \frac{1}{n}, \frac{4}{n}, \cdots, 1\right)$$

$$Y_3^2 = \left(\frac{3}{n}, \frac{1}{n}, \frac{2}{n}, \frac{4}{n}, \cdots, 1\right)$$

$$Y_3^3 = \left(\frac{1}{n}, \frac{3}{n}, \frac{2}{n}, \frac{4}{n}, \cdots, 1\right)$$

$$Y_3^4 = \left(\frac{3}{n}, \frac{2}{n}, \frac{1}{n}, \frac{4}{n}, \cdots, 1\right)$$

Then

$$W(X^*, Y_3^1) = V\left(\psi\left(\frac{1}{n}, \frac{2}{n}, \alpha_1\right), \psi\left(\frac{2}{n}, \frac{3}{n}, \alpha_2\right), \psi\left(\frac{3}{n}, \frac{1}{n}, \alpha_3\right), 0, \ldots, 0\right) \tag{B.1}$$

$$W(X^*, Y_3^2) = V\left(\psi\left(\frac{1}{n}, \frac{3}{n}, \alpha_1\right), \psi\left(\frac{2}{n}, \frac{1}{n}, \alpha_2\right), \psi\left(\frac{3}{n}, \frac{2}{n}, \alpha_3\right), 0, \ldots, 0\right) \tag{B.2}$$

$$W(X^*, Y_3^3) = V\left(0, \psi\left(\frac{2}{n}, \frac{3}{n}, \alpha_2\right), \psi\left(\frac{3}{n}, \frac{2}{n}, \alpha_3\right), 0, \ldots, 0\right) \tag{B.3}$$

$$W(X^*, Y_3^4) = V\left(\psi\left(\frac{1}{n}, \frac{3}{n}, \alpha_1\right), 0, \psi\left(\frac{3}{n}, \frac{1}{n}, \alpha_3\right), 0, \ldots, 0\right) \tag{B.4}$$

By Lemma B.1, equations (B.3) and (B.4) are greater than 0. So I only need to discuss equations (B.1) and (B.2). By Lemma B.1 again, we have

$$V\left(\psi\left(\frac{1}{n}, \frac{2}{n}, \alpha_1\right), \psi\left(\frac{2}{n}, \frac{3}{n}, \alpha_2\right), 0, \ldots, 0\right) > 0$$

$$V\left(\psi\left(\frac{1}{n}, \frac{3}{n}, \alpha_1\right), \psi\left(\frac{2}{n}, \frac{1}{n}, \alpha_2\right), 0, \ldots, 0\right) > 0$$

By Lemma B.2, we can always find $\alpha_{3,1}^* \geqslant \alpha_2$ and $\alpha_{3,2}^* \geqslant \alpha_2$, such that if $\alpha_3 > \alpha_3^* = max\{\alpha_{3,1}^*, \alpha_{3,2}^*\}$, then both equations (B.1) and (B.2) are greater than 0. Therefore, if $\alpha_3 > \alpha_3^*$, $W(X^*, Y_3^i) > 0$ for each $Y_3^i \in Y_3$ and X^* is the best; if $\alpha_3 < \alpha_3^*$, $W(X^*, Y_3^i) < 0$ for some $Y_3^i \in Y_3$ and X^* is not the best.

3. If $k > 3$, and the conditions are satisfied such that $X^* \succ Y_j^i$, where $Y_j^i \in \mathcal{Y}_2 \cup \ldots \cup \mathcal{Y}_{k-1}$.

   I want to find conditions such that if they are satisfied, $X^* \succ Y_k^i$ for any $Y_k^i \in \mathcal{Y}_k$; if not, X^* is not the best choice.

   Similarly, I use relative definition. Suppose individual k gets rank $\frac{k'}{n}$, and individual k'' gets rank $\frac{k}{n}$ in Y_k^i. As the k^{th} individual always gets a worse rank in X^* than in Y_k^i, we have $k' < k$ and $\psi\left(\frac{k}{n}, \frac{k'}{n}, \alpha_k\right) < 0$.

$$W(X^*, Y_k^i) = V\left(\psi_1, \ldots, \psi\left(\frac{k''}{n}, \frac{k}{n}, \alpha_{k''}\right), \ldots, \psi_{k-1}, \psi\left(\frac{k}{n}, \frac{k'}{n}, \alpha_k\right), 0, \ldots, 0\right) \tag{B.5}$$

   Consider another ranking $Y_{k-1}^i \in \mathcal{Y}_{k-1}$, the only difference between Y_k^i and Y_{k-1}^i is that individual k gets ranking $\frac{k}{n}$ and individual k'' gets ranking $\frac{k'}{n}$ in Y_{k-1}^i. We have

$$W(X^*, Y^i_{k-1}) = V\left(\psi_1, \ldots, \psi\left(\frac{k''}{n}, \frac{k'}{n}, \alpha_{k''}\right), \ldots, \psi_{k-1}, 0, \ldots, 0\right)$$

By $W(X^*, Y^i_{k-1}) > 0$ and monotonicity, we have

$$V\left(\psi_1, \ldots, \psi\left(\frac{k''}{n}, \frac{k}{n}, \alpha_{k''}\right), \ldots, \psi_{k-1}, 0, \ldots, 0\right) > W(X^*, Y^i_{k-1}) > 0 \quad \text{(B.6)}$$

By lemma B.2, we can find $\alpha^*_k \geqslant \alpha_{k-1}$ such that if $\alpha_k > \alpha^*_k$, we have $X^* \succ Y^i_k$ for any $Y^i_k \in Y_k$ and X^* is the best; if $\alpha_k < \alpha^*_k$, X^* is not the best.

By induction on k, if $k = n$ and $\alpha_n > \alpha^*_n$, then X^* is the best choice. ■

To understand the way to get critical value functions, I discuss more details when the aggregate function is additive. Define Y^j_k as one of the rankings where person k takes position $k - j$ and individuals $\{k + 1, \ldots, n\}$ take the same positions as X^*. Suppose person $k - j$, whose original position has been taken by person k takes position p_2. So on and so forth, finally, a person p_m takes position k. Define such a loop as one circle and write it as

$$k \to k - j \to p_2 \to \ldots \to p_m \to k$$

Define T as the first term of equation (B.6). Given j, to get α^j_k, I want to minimize the value of T. Therefore, it satisfies the followings.

1. There should be only one circle involved in the equation. Denote it circle 1. Suppose not, there is another circle denoted by circle 2. The aggregation of circle 2 must be positive because of the induction process, which increases the value of T.

2. In circle 1, only person k gets a better rank in Y^j_k than X^*, while all the others get lower rank in Y^j_k, which implies that

$$k - j < p_2 < \ldots < p_m < k$$

Suppose not, part of the circle, $(p_l, \ldots, p_t)$, satisfies $p_{l-1} < p_l$, $p_t < p_{t+1}$ and $p_l > \ldots > p_t$. To simplify, I use absolute rankings. We have

$$T' = \ldots + \psi(p_{l-1}, p_l, \alpha_{p_{l-1}}) + \psi(p_l, p_{l+1}, \alpha_{p_l}) + \ldots$$

$$+\psi(p_t, p_{t+1}, \alpha_{p_t}) + \psi(p_{t+1}, p_{t+2}, \alpha_{p_{t+1}}) + \dots$$

By regret aversion,

$$\psi(p_{l-1}, p_l, \alpha_{p_{l-1}}) > \psi(p_{l-1}, p_t, \alpha_{p_{l-1}}) + \psi(p_t, p_{t-1}, \alpha_{p_{l-1}}) + \dots + \psi(p_{l+1}, p_l, \alpha_{p_{l-1}})$$

$$> \psi(p_{l+1}, p_l, \alpha_{p_l}) + \dots + \psi(p_t, p_{t-1}, \alpha_{p_{t-1}}) + \psi(p_{l-1}, p_t, \alpha_{p_{l-1}})$$

Then

$$\psi(p_{l-1}, p_t, \alpha_{p_{l-1}}) < \psi(p_{l-1}, p_l, \alpha_{p_{l-1}}) + \psi(p_l, p_{l+1}, \alpha_{p_l}) + \dots + \psi(p_{t-1}, p_t, \alpha_{p_{t-1}})$$

Therefore, if we remove persons $(p_l, \dots, p_{t-1})$, T will be smaller. The circle changes to

$$k \to k-j \to \dots \to p_{l-1} \to p_t \dots \to p_m \to k$$

and we have

$$T'' = \dots + \psi(g_{l-1}, g_t, \alpha_{p_{l-1}}) + \dots < T'$$

If $p_{l-1} > p_t$, we can keep removing. In the end, we have only one circle and each person except person k gets a lower rank in this circle.

3. Given j, in circle 1, person k moves upward j positions, and the persons from $k-j$ to $k-1$ move downward 1 position.

   Suppose not. At least one person moves down s positions, where $s > 1$, then by regret aversion, we can decrease T by moving this person downward 1 position, and the next person $s-1$ positions. Keep breaking down, it ends up with each person moving downward 1 position.

Therefore, given j, the particular equation to calculate α_k^j is the following

$$\psi(k-j, k, \alpha_k^j) = \psi(k-j, k-j+1, \alpha_{k-j}) + \dots + \psi(k-1, k, \alpha_{k-1})$$

And

$$\alpha_k^* = max(\alpha_k^1, \dots, \alpha_k^{k-1})$$

B.4 Proof of Theorem 3.2

By the proof of Theorem 3.1, we have

$$W(X^*, Y_k^i) = \psi\left(\frac{1}{n}, \alpha_{k-i}\right) + \ldots + \psi\left(\frac{1}{n}, \alpha_{k-1}\right) - \psi\left(\frac{i}{n}, \alpha_k^i\right) = 0$$

$$\Rightarrow \psi(\frac{i}{n}, \alpha_k^i) = \sum_{j=k-i}^{k-1} \psi(\frac{1}{n}, \alpha_j) \tag{B.7}$$

Take derivative with respect to n on both sides,

$$\psi_1(\frac{i}{n}, \alpha_k^i)(-\frac{i}{n^2}) + \psi_2(\frac{i}{n}, \alpha_k^i)\frac{\partial \alpha_k^i}{\partial n} = -\frac{1}{n^2}\sum_{j=k-i}^{k-1} \psi_1(\frac{1}{n}, \alpha_j)$$

$$\Longrightarrow \psi_2(\frac{i}{n}, \alpha_k^i)\frac{\partial \alpha_k^i}{\partial n} = -\frac{1}{n^2}\sum_{j=k-i}^{k-1}\left[\psi_1(\frac{1}{n}, \alpha_j) - \psi_1(\frac{i}{n}, \alpha_k^i)\right]$$

$$\Longrightarrow \frac{\partial \alpha_k^i}{\partial n} = \frac{1}{n^2}\frac{1}{\psi_2(\frac{i}{n}, \alpha_k^i)}\sum_{j=k-i}^{k-1}\left[\psi_1(\frac{i}{n}, \alpha_k^i) - \psi_1(\frac{1}{n}, \alpha_j)\right] \tag{B.8}$$

As $\psi_2(y - x, \alpha_i) < 0$ for any i, I want to show that

$$\psi_1(\frac{i}{n}, \alpha_k^i) - \frac{1}{i}\sum_{j=k-i}^{k-1} \psi_1(\frac{1}{n}, \alpha_j) \geqslant 0$$

for any $i \in [1, k-1]$. Let $\psi_1(\frac{1}{n}, \bar{\alpha}_{k,i}) = \frac{1}{i}\sum_{j=k-i}^{k-1} \psi_1(\frac{1}{n}, \alpha_j)$, then

$$\psi_1(\frac{i}{n}, \alpha_k^i) - \psi_1(\frac{1}{n}, \bar{\alpha}_{k,i}) \geqslant 0$$

which is the assumption.

Moreover, I want to show that there exists k for which α_k^* is not constant. If $\alpha_k^* = \alpha_k^i$, where $i > 1$, by equation (B.8), if $\psi_1(\frac{i}{n}, \alpha_k^i) - \psi_1(\frac{1}{n}, \alpha_j) > 0$ and $\alpha_j \neq \alpha_{k-1}$ for some j, then $\frac{\partial \alpha_k^*}{\partial n} > 0$. So I only need to show that for some k, $\alpha_k^* \neq \alpha_k^1 = \alpha_{k-1}$.

Suppose not, we have $\alpha_{k-1} = \alpha_{k-2} = \ldots = \alpha_2$ and $\alpha_k^i \leqslant \alpha_{k-1}$ for any $i > 1$, because α_k^* is the maximum value among α_k^j, where $j \in [1, k-1]$. By monotonicity,

$\psi(\frac{i}{n}, \alpha_k^i) \geqslant \psi(\frac{i}{n}, \alpha_{k-1})$. By equation (B.7), we have

$$\psi(\frac{i}{n}, \alpha_k^i) = i \cdot \psi(\frac{1}{n}, \alpha_{k-1}) \geqslant \psi(\frac{i}{n}, \alpha_{k-1})$$

A violation of regret aversion.

Regarding the limiting case, I will prove it by induction.

1. $k = 3$.

   Recall that we have

   $$\psi\left(\frac{i}{n}, \alpha_k^i\right) = \sum_{j=k-i}^{k-1} \psi\left(\frac{1}{n}, \alpha_j\right)$$

   If $k = 3$, i could be either 1 or 2. If $i = 1$, we have $\alpha_3^1 = \alpha_2$. If $i = 2$, we have

   $$\psi\left(\frac{2}{n}, \alpha_3^2\right) = \psi\left(\frac{1}{n}, \alpha_1\right) + \psi\left(\frac{1}{n}, \alpha_2\right)$$

   I want to show that,

   $$\lim_{n\to\infty} \frac{\psi\left(\frac{1}{n}, \alpha_1\right) + \psi\left(\frac{1}{n}, \alpha_2\right)}{\psi\left(\frac{2}{n}, \alpha_2\right)} > 1$$

   Because this inequality implies that $\lim_{n\to\infty} \alpha_3^2 < \alpha_2$, so $\lim_{n\to\infty} \alpha_3^* = \alpha_2$. As $n \to \infty$, both the denominator and numerator of the LHS of the inequality go to zero. Use L'Hopital's Rule, we have

   $$LHS = \frac{\psi_1(0, \alpha_1) + \psi_1(0, \alpha_2)}{2\psi_1(0, \alpha_2)} > 1$$

   which is implied by $\alpha_1 < \alpha_2$.

2. $k > 3$, given the claim is true for $\{3, \ldots, k-1\}$.

   (a) If $i \in [1, k-2]$, we have

   $$\psi\left(\frac{i}{n}, \alpha_k^i\right) = \sum_{j=k-i}^{k-1} \psi\left(\frac{1}{n}, \alpha_j\right) = i \cdot \psi\left(\frac{1}{n}, \alpha_2\right)$$

   Again, by L'Hopital's Rule, we have

   $$\lim_{n\to\infty} \frac{\psi\left(\frac{i}{n}, \alpha_2\right)}{\psi\left(\frac{1}{n}, \alpha_2\right)} = \frac{i \cdot \psi_1(0, \alpha_2)}{\psi_1(0, \alpha_2)} = i$$

Therefore, $\lim_{n\to\infty}\psi\left(\frac{i}{n},\alpha_k^i\right)=\psi\left(\frac{i}{n},\alpha_2\right)$, then $\lim_{n\to\infty}\alpha_k^i=\alpha_2$ for $i\in[1,k-2]$.

(b) Consider $i=k-1$, we have

$$\psi\left(\frac{k-1}{n},\alpha_k^{k-1}\right)=\sum_{j=1}^{k-1}\psi\left(\frac{1}{n},\alpha_j\right)$$

$$=(k-2)\psi\left(\frac{1}{n},\alpha_2\right)+\psi\left(\frac{1}{n},\alpha_1\right)$$

By L'Hopital's Rule, if $n\to\infty$, we have

$$\lim_{n\to\infty}\frac{(k-2)\psi\left(\frac{1}{n},\alpha_2\right)+\psi\left(\frac{1}{n},\alpha_1\right)}{\psi\left(\frac{k-1}{n},\alpha_2\right)}=\frac{(k-2)\psi_1\left(0,\alpha_2\right)+\psi_1\left(0,\alpha_1\right)}{(k-1)\psi_1\left(0,\alpha_2\right)}>1$$

as $\psi_{12}(y-x,\alpha)<0$, which implies $\psi_1\left(0,\alpha_1\right)>\psi_1\left(0,\alpha_2\right)$. We have $\alpha_k^{k-1}<\alpha_2$.

Therefore, $\lim_{n\to\infty}\alpha_k^*\to\alpha_2$ for $k>3$. ■

B.5 Proof of Example 3.3

Given $x<y$, we have

$$\psi_1(y-x,\alpha)=\alpha(y-x)^{\alpha-1}$$

I want to show that

$$\psi_1\left(\frac{j}{n},\alpha_k^j\right)=\alpha_k^j\left(\frac{j}{n}\right)^{\alpha_k^j-1}$$

$$\geqslant\psi_1\left(\frac{1}{n},\bar{\alpha}_{k,j}\right)=\frac{1}{j}\sum_{i=k-j}^{k-1}\alpha_i\left(\frac{1}{n}\right)^{\alpha_i-1}$$

which is

$$\alpha_k^j\left(\frac{j}{n}\right)^{\alpha_k^j}\geqslant\sum_{i=k-j}^{k-1}\alpha_i\left(\frac{1}{n}\right)^{\alpha_i}$$

By the definition of α_k^j, we have

$$\left(\frac{j}{n}\right)^{\alpha_k^j}=\sum_{i=k-j}^{k-1}\left(\frac{1}{n}\right)^{\alpha_i}$$

By $\alpha_k^j \geqslant \alpha_i$,

$$\alpha_k^j \left(\frac{j}{n}\right)^{\alpha_k^j} = \sum_{i=k-j}^{k-1} \alpha_k^j \left(\frac{1}{n}\right)^{\alpha_i} \geqslant \sum_{i=k-j}^{k-1} \alpha_i \left(\frac{1}{n}\right)^{\alpha_i}$$

for $i = k-j, \ldots, k-1$. ■

B.6 Proof of Theorem 3.1*

Let Ω_G be the set of all possible rankings, and $\mathcal{Y}$ be the set $\Omega_G \setminus X_G$. Define set $\mathcal{Y}_k$ as all possible rankings where their differences between X_G only involve the first $k \in \{2, \ldots, I\}$ groups. Note that for each ranking $Y_k^j \in \mathcal{Y}_k$, at least one individual in the k^{th} group has a rank different from her rank in X_G. Then $\{\mathcal{Y}_2, \ldots, \mathcal{Y}_I\}$ are independent sets and $\mathcal{Y} = \mathcal{Y}_2 \cup \mathcal{Y}_3 \cup \ldots \cup \mathcal{Y}_I$.

We need to find conditions such that if they are satisfied, then $X_G \succ Y_k^j$ for any $Y_k^j \in \mathcal{Y}_k$ and for any $k \in \{2, \ldots, I\}$; if not, X_G is not the best choice.

1. $k = 2$. Let $\alpha_2^* = \alpha_{N_1}$. Suppose a set of individuals A_2 in group 2 of X_G are not in group 2 of Y_2^j, where $Y_2^j \in \mathcal{Y}_2$. And a set of individuals A_2' in group 2 of Y_2^j are not in group 2 of X_G. Sets A_2 and A_2' have the same number of individuals. As $\alpha_{n_I+1} > \alpha_2^*$, the individuals in set A_2' are more sensitive than the individuals in set A_2. By lemma 1, we have $X_G \succ Y_2^j$, where Y_2^j is any ranking in $\mathcal{Y}_2$.

2. If $k > 2$, and the conditions are satisfied such that $X_G \succ Y_{k-1}^j$, where $Y_{k-1}^j \in \mathcal{Y}_2 \cup \ldots \cup \mathcal{Y}_{k-1}$.

   Suppose a set of individuals A_k in the k^{th} group of X_G has different rank in Y_k^j, and a set of individuals A_k' in the k^{th} group of Y_k^j is not in the k^{th} group of X_G. Define the set of other individuals, who are not in either A_k or A_k' but have different rank in X_G and Y_k^j, as A_k''. We have

   $$W(X_G, Y_k^j) = V(\underbrace{\psi_{m1}, \ldots}_{m_1 \in A_k'}, \underbrace{\psi_{m_2}, \ldots}_{m_2 \in A_k''}, \underbrace{\psi_{m_3}, \ldots}_{m_3 \in A_k}, 0) \tag{B.9}$$

   As each individual in A_k gets a lower rank in X_G than in Y_k^j, we have $\psi_{m_3} < 0$ where $m_3 \in A_k$.

   Consider another ranking $Y_{k-1}^j \in \mathcal{Y}_{k-1}$, the only difference between Y_k^j and Y_{k-1}^j is that individuals A_k stay in the k^{th} group and individuals A_k' take the positions of individuals A_k in Y_k^j. We have $W(X_G, Y_{k-1}^j) > 0$. By monotonicity, we have

$$V(\underbrace{\psi'_{m1},\ldots}_{m_1\in A'_k},\underbrace{\psi_{m_2},\ldots}_{m_2\in A''_k},\underbrace{0,\ldots}_{m_3\in A_k},0)>0$$

By lemma B.2, we can find $\alpha^{k*}\geqslant\alpha_{N_{k-1}}$ such that if $\alpha_{N_{k-1}+1}>\alpha^{k*}$, we have $X_G\succ Y_k^i$ for any $Y_k^i\in Y_k$ and X_G is the best; if $\alpha_{N_{k-1}+1}<\alpha^{k*}$, X_G is not the best.

By induction on k, if $k=I$ and $\alpha_{N_{I-1}+1}>\alpha^{I*}$, then X_G is the best choice. ■

To understand the way to get critical value functions, I discuss more details when the aggregate function is additive. Given j, to get α_k^j, we want the value of equation (B.10) be as small as possible.

$$V'=\underbrace{\psi_{m1}+\ldots}_{m_1\in A'_k}+\underbrace{\psi_{m_2}+\ldots}_{m_2\in A''_k}+\underbrace{0,\ldots}_{m_3\in A_k}+0 \tag{B.10}$$

From X_G to Y_k^j, observe that one person p_0 in group k moves to group g_1. The person p_1 in group g_1, whose original position has been taken by person p_0, moves to group g_2. So on and so forth, finally, a person in group g_m moves to group k and takes the original position of person p_0. Note that $g_i\leqslant k$ for any $i\in\{1,\ldots,m\}$. Define such a loop as one circle and write it as

$$p_0\to p_1\to p_2\to\ldots\to p_m\to p_0$$

or

$$k\to g_1\to g_2\to\ldots\to g_m\to k$$

1. I claim that to get α_k^j, there should be only one circle involved in equation (B.10). Denote it circle 1. Suppose not, there is another circle denoted circle 2. If no persons in group k were involved in circle 2, then the aggregation of circle 2 must be positive because of the induction, which increases the value of V'. If circle 2 involves at least one person in group k, we can calculate α_k^j separately. If the critical values calculated from the two circles are different, then we should only keep the circle leading to a greater α_k^j; otherwise, the extra circle also increases the value of V'. Because the α values of group k have to be strictly greater than the larger critical value. Thus, there is only one circle.

2. I claim that it is impossible to have more than one persons from the same group in one circle as we can always break it into several circles. Consider the following

circle

$$p_0 \to p_1 \to \ldots \to p_{u-1} \to p_u \to \ldots \to p_{v-1} \to p_v \to \ldots \to p_m \to p_0$$

where p_u and p_v are from the same group. Then we can break this circle into two circles as

$$p_0 \to p_1 \to \ldots \to p_{u-1} \to p_v \to \ldots \to p_m \to p_0$$

and

$$p_{v-1} \to p_u \to \ldots \to p_{v-1}$$

Therefore, we have only one person from group k in circle 1. As person $N_{k-1}+1$ has the highest level of sensitivity in group k, if this person has a smaller level of sensitivity than the critical value, then everyone in group k has. So we only need to consider the circles including person $N_{k-1}+1$ and calculate the critical value of $\alpha^j_{N_{k-1}}$ as $\alpha^{k,j}$.

3. I claim that in circle 1, only person p_0 gets a higher rank in Y^j_k, while all the others get lower rank in Y^j_k, which implies that

$$p_1 < p_2 < \ldots < p_m < p_0$$

or

$$g_1 < g_2 < \ldots < g_m < k$$

Suppose not, part of the circle, $(p_l, \ldots, p_t)$ in groups $(g_l, \ldots, g_t)$, satisfies $g_{l-1} < g_l > \ldots > g_t < g_{t+1}$. To simplify, let the group number stands for the rank. We have

$$V_1' = \ldots + \psi(g_{l-1}, g_l, \alpha_{p_{l-1}}) + \psi(g_l, g_{l+1}, \alpha_{p_l}) + \ldots$$
$$+\psi(g_t, g_{t+1}, \alpha_{p_t}) + \psi(g_{t+1}, g_{t+2}, \alpha_{p_{t+1}}) + \ldots$$

By regret aversion,

$$\psi(g_{l-1}, g_l, \alpha_{p_{l-1}}) > \psi(g_{l-1}, g_t, \alpha_{p_{l-1}}) + \psi(g_t, g_{t-1}, \alpha_{p_{l-1}}) + \ldots + \psi(g_{l+1}, g_l, \alpha_{p_{l-1}})$$

$$> \psi(g_{l+1}, g_l, \alpha_{p_l}) + \ldots + \psi(g_t, g_{t-1}, \alpha_{p_{t-1}}) + \psi(g_{l-1}, g_t, \alpha_{p_{l-1}})$$

Then

$$\psi(g_{l-1}, g_t, \alpha_{p_{l-1}}) < \psi(g_{l-1}, g_l, \alpha_{p_{l-1}}) + \psi(g_l, g_{l+1}, \alpha_{p_l}) + \ldots + \psi(g_{t-1}, g_t, \alpha_{p_{t-1}})$$

Therefore, if remove persons $(p_l, \ldots, p_{t-1})$, the value of V' will decrease. The circle changes to

$$p_0 \to p_1 \to \ldots \to p_{l-1} \to p_t \ldots \to p_m \to p_0$$

or

$$k \to g_1 \to \ldots \to g_{l-1} \to g_t \ldots \to g_m \to k$$

and we have

$$V_2' = \ldots + \psi(g_{l-1}, g_t, \alpha_{p_{l-1}}) + \ldots < V_1'$$

If $g_{l-1} > g_t$, we can keep removing. In the end, we have only one circle and each person except person p_0 gets a lower rank in this circle.

4. I claim that given j, in circle 1, the most sensitive person in group k moves upward j groups, and the least sensitive persons in group $k-j$ to group $k-1$ moves downward 1 group. Suppose not. First, if at least one person is not the least sensitive person in the group, then we can replace this person by the least sensitive person and it decreases the value of V'. Second, if at least one person moves downward s groups, where $s > 1$, then by regret aversion, we can decrease the value of V' by moving this person down by 1 group, and the least sensitive person in the next group by $s-1$ groups. Keep breaking down, it ends up with each person moving downward 1 group.

Therefore, given any $j = 1, \ldots, k-1$, the particular equation to calculate $\alpha^{k,j}$ is the following

$$\psi(k-j, k, \alpha^{k,j}) = \psi(k-j, k-j+1, \alpha_{N_{k-j}}) + \ldots + \psi(k-1, k, \alpha_{N_{k-1}})$$

and

$$\alpha^{k*} = max_j\{\alpha^{k,j}\}$$

X_G is the best among $\mathcal{Y}_2 \cup \ldots \cup \mathcal{Y}_k$ if and only if $\alpha_{N_{k-1}+1} \geqslant \alpha^{k*}$.

By induction, we have X_G is the best if and only if $\alpha_{N_{k-1}+1} \geqslant \alpha^{k*}$ for any $k = \{2, \ldots, I\}$. ■

B.7 Proof of Claim 3.2

Suppose X_G is not the best. As ν is stable, there is a best ranking X'_G for ν, and $X'_G \succ X_G$. Suppose X'_G and X_G differ only in (some of) the first k groups. Suppose a set of individuals G_k are in the k^{th} group of X_G but not in the k^{th} group of X'_G; and a set of individuals G'_k are in the k^{th} group of X'_G but not in the k^{th} group of X_G. Sets G_k and G'_k are disjoint and have the same number of individuals (recall that X_G and X'_G have the same structure ν). As X_G puts people in a descending order of sensitivity, the individuals in G_k are in the k^{th} group and they are the least sensitive people before the $(k+1)^{th}$ group. Therefore, all individuals in G'_k are more sensitive than the individuals in G_k.

Suppose both G_k and G'_k have m individuals. Starting from X'_G, switch the positions of individuals in G_k and G'_k in the following way. Put the individuals in each set in a descending order of sensitivity. Switch the positions of individual j in G_k and individual j in G'_k, where $j = 1, \ldots, m$. Define the new ranking as X^k_G.

Observe that all individuals in G'_k get a higher rank in X^k_G than in X'_G. As all of them are more sensitive than the individuals in G_k, by Lemma 1, we have $X^k_G \succ X'_G$. A contradiction. ■

B.8 Proof of Claim 3.3

Consider a structure with one group. It has a best choice as there is only one ranking in the choice set.

Consider structures with two groups. There are $n-1$ of them. The structures are $[i, n-i]$ for any $i = 1, \ldots, n-1$. By the proof of Theorem 1*, we only need to compare the sensitivities of person i and $i+1$. As person i is more sensitive than person $i+1$, where $\alpha_i < \alpha_{i+1}$, X^*_G is always preferred to the alternative rankings. Therefore, the $n-1$ structures with two groups are stable. ■

B.9 Proof of Claim 3.4

Consider a case where each group has the same level of sensitivity. The sequence of sensitivities is simplified as $(\alpha_1, \ldots, \alpha_I)$. Create a sequence

$$\alpha^* = (\alpha_1, \alpha_2, \alpha^*_3, \ldots, \alpha^*_I)$$

where $\alpha^*_3 > \alpha_2$ and α^*_i is the critical value of α_i for $i = 3, \ldots, I$.

Consider any alternative structures $\nu' = [n_1, n_2, n_3', \ldots, n_I']$. Suppose the first different group between ν and ν' is group $k \geqslant 3$. There are two cases: one is $n_k' > n_k$ where ν' merges group k with some other individuals; the other one is $n_k' < n_k$ where ν' splits group k. I can show that the critical value of group k in ν' is larger than α_k in the first case, and the critical value of group $k+1$ in ν' is larger than α_k which is the actual value of group $k+1$ in the second case. Then ν' is not stable in either case and ν is the only stable structure and therefore the optimal stable structure.

1. If $n_k' > n_k$.

   First, I want to show that $\alpha_k^* = \alpha_k^{k-1}$ given $\alpha_3^* > \alpha_2$ and $\alpha_i = \alpha_i^*$ for $i = 3, \ldots, I$. I prove it by induction on k. For $k = 3$, we have $\alpha_3^* = \max\{\alpha_3^1, \alpha_3^2\}$. As $\alpha_3^1 = \alpha_2$ and $\alpha_3^* > \alpha_2$, we have $\alpha_3^* = \alpha_3^2$. If it is true for $k-1$, then I need to show that it is also true for k. Recall that α_k^j is implied by the equation

$$\psi\left(\frac{n_{k-j} + 2(n_{k-j+1} + \ldots + n_{k-1}) + n_k}{2n}, \alpha_k^j\right)$$

$$= \psi\left(\frac{n_{k-j} + n_{k-j+1}}{2n}, \alpha_{k-j}\right) + \ldots + \psi\left(\frac{n_{k-2} + n_{k-1}}{2n}, \alpha_{k-2}\right)$$
$$+\psi\left(\frac{n_{k-1} + n_k}{2n}, \alpha_{k-1}\right)$$

$$= \psi\left(\frac{n_{k-j} + 2(n_{k-j+1} + \ldots + n_{k-2}) + n_{k-1}}{2n}, \alpha_{k-1}^{j-1}\right)$$
$$+\psi\left(\frac{n_{k-1} + n_k}{2n}, \alpha_{k-1}\right)$$

   By the assumption, α_k^j increases with both terms $\frac{n_{k-j}+2(n_{k-j+1}+\ldots+n_{k-2})+n_{k-1}}{2n}$ and α_{k-1}^{j-1}. As both of them are maximized at $j = k-1$, we have $j = k-1$. Therefore, $\alpha_k^* = \alpha_k^{k-1}$ for each $k \geqslant 3$.

   Second, I want to show that $\alpha_k^{*\prime}$ increases with n_k'. By the first part, $\alpha_k^{*\prime} = \alpha_k^{k-1\prime}$, then we have

$$\psi\left(\frac{n_1 + 2(n_2 + \ldots + n_{k-1}) + n_k}{2n}, \alpha_k^{*\prime}\right)$$

$$= \psi\left(\frac{n_1 + 2(n_2 + \ldots + n_{k-2}) + n_{k-1}}{2n}, \alpha_{k-1}\right) + \psi\left(\frac{n_{k-1} + n_k'}{2n}, \alpha_{k-1}\right)$$

   As $\alpha_k^{*\prime} = \alpha_k$ when $n_k' = n_k$, again by the assumption, $\alpha_k^{*\prime} > \alpha_k$ when $n_k' > n_k$. Therefore, α_k is smaller than the critical value and ν' is not stable.

2. If $n'_k < n_k$.

   First, I want to show that if $n'_k + n'_{k+1} < n_k$, then the structure is not stable. Consider ${\alpha^2_{k+2}}'$, we have

$$\psi\left(\frac{n'_k + 2n'_{k+1} + n'_{k+2}}{2n}, \alpha^2_{k+2}\right)$$

$$= \psi\left(\frac{n'_k + n'_{k+1}}{2n}, \alpha_k\right) + \psi\left(\frac{n'_{k+1} + n'_{k+2}}{2n}, \alpha_k\right)$$

$$< \psi\left(\frac{n'_k + 2n'_{k+1} + n'_{k+2}}{2n}, \alpha_k\right)$$

   Then $\alpha_k < \alpha^2_{k+2} \leqslant {\alpha^*_{k+2}}'$ and it is not stable. Therefore, $n_k \leqslant n'_k + n'_{k+1}$.

   Second, I want to show that ${\alpha^*_{k+1}}' > \alpha_k$. Note that

$$\psi\left(\frac{n_1 + 2(n_2 + \ldots + n'_k) + n'_{k+1}}{2n}, {\alpha^k_{k+1}}'\right)$$

$$= \psi\left(\frac{n_1 + 2(n_2 + \ldots + n_{k-2}) + n_{k-1}}{2n}, \alpha_{k-1}\right) + \psi\left(\frac{n_{k-1} + n'_k}{2n}, \alpha_{k-1}\right)$$

$$+\psi\left(\frac{n'_k + n'_{k+1}}{2n}, \alpha_k\right)$$

   and

$$\psi\left(\frac{n_1 + 2(n_2 + \ldots + n_{k-1}) + n_k}{2n}, \alpha_k\right)$$

$$= \psi\left(\frac{n_1 + 2(n_2 + \ldots + n_{k-2}) + n_{k-1}}{2n}, \alpha_{k-1}\right) + \psi\left(\frac{n_{k-1} + n_k}{2n}, \alpha_{k-1}\right)$$

   We have

$$\psi\left(\frac{n_1 + 2(n_2 + \ldots + n'_k) + n'_{k+1}}{2n}, {\alpha^k_{k+1}}'\right)$$

$$= \psi\left(\frac{n_1 + 2(n_2 + \ldots + n_{k-1}) + n_k}{2n}, \alpha_k\right) - \psi\left(\frac{n_{k-1} + n_k}{2n}, \alpha_{k-1}\right)$$

$$+\psi\left(\frac{n_{k-1} + n'_k}{2n}, \alpha_{k-1}\right) + \psi\left(\frac{n'_k + n'_{k+1}}{2n}, \alpha_k\right)$$

   As $n'_k < n_k$, we have

$$\psi\left(\frac{n_{k-1} + n_k}{2n}, \alpha_{k-1}\right) > \psi\left(\frac{n_{k-1} + n'_k}{2n}, \alpha_{k-1}\right)$$

$$+\psi\left(\frac{n_k - n'_k}{2n}, \alpha_{k-1}\right)$$

then

$$-\psi\left(\frac{n_{k-1} + n_k}{2n}, \alpha_{k-1}\right) + \psi\left(\frac{n_{k-1} + n'_k}{2n}, \alpha_{k-1}\right)$$
$$< -\psi\left(\frac{n_k - n'_k}{2n}, \alpha_{k-1}\right)$$

We have

$$\psi\left(\frac{n_1 + 2(n_2 + \ldots + n'_k) + n'_{k+1}}{2n}, {\alpha^k_{k+1}}'\right)$$
$$< \psi\left(\frac{n_1 + 2(n_2 + \ldots + n_{k-1}) + n_k}{2n}, \alpha_k\right) - \psi\left(\frac{n_k - n'_k}{2n}, \alpha_{k-1}\right)$$
$$+\psi\left(\frac{n'_k + n'_{k+1}}{2n}, \alpha_k\right)$$

The first part indicates that $n_k < 2n'_k + n'_{k+1}$. By strict convexity, we have $\psi(x, \alpha) + \psi(z - x, \alpha) < \psi(y, \alpha) + \psi(z - y, \alpha)$, where $y < x < z - x < z - y$.[22] So

$$\psi\left(\frac{n_1 + 2(n_2 + \ldots + n_{k-1}) + n_k}{2n}, \alpha_k\right) + \psi\left(\frac{n'_k + n'_{k+1}}{2n}, \alpha_k\right)$$
$$< \psi\left(\frac{n_1 + 2(n_2 + \ldots + n'_k) + n'_{k+1}}{2n}, \alpha_k\right) + \psi\left(\frac{n_k - n'_k}{2n}, \alpha_{k-1}\right)$$

Then

$$\psi\left(\frac{n_1 + 2(n_2 + \ldots + n'_k) + n'_{k+1}}{2n}, {\alpha^k_{k+1}}'\right)$$
$$< \psi\left(\frac{n_1 + 2(n_2 + \ldots + n'_k) + n'_{k+1}}{2n}, \alpha_k\right)$$

Therefore, $\alpha_k < {\alpha^k_{k+1}}' \leqslant {\alpha^*_{k+1}}'$ and the structure is not stable.

By strict convexity, we have $\psi(x + y, \alpha) > \psi(x, \alpha) + \psi(y, \alpha)$. Given $\frac{\partial\psi(x,\alpha)}{\partial\alpha} < 0$, for some very small $0 \leqslant \varepsilon_1 \leqslant \varepsilon_2$, we can still have $\psi(x+y, \alpha+\varepsilon_2) > \psi(x, \alpha)+\psi(y, \alpha+\varepsilon_1)$. Therefore, if the sensitivities of each group are slightly different, the claim still holds. ■

[22] Convexity indicates that $f''(x) < 0$. Then $\frac{\psi(x,\alpha)-\psi(y,\alpha)}{x-y} < \frac{\psi(z-y,\alpha)-\psi(z-x,\alpha)}{(z-y)-(z-x)}$ as the LHS can be seen as the slope of some point between y and x and the RHS can be seen as the slope of some point between $z - x$ and $z - y$. Therefore, $\psi(x, \alpha) + \psi(z - x, \alpha) < \psi(y, \alpha) + \psi(z - y, \alpha)$.

C Proofs of Chapter 4

C.1 Appendix I

Recall $\widehat{\alpha} = \hat{W}_f \hat{W}_g \widehat{\alpha}_h + \hat{W}_f \left(1 - \hat{W}_g\right) \widehat{\alpha}_g + (1 - \hat{W}_f)\widehat{\alpha}_f$. To avoid confusion, we collect our notation here. The sample and population moments are

$$\widehat{g}(\alpha,\beta) \equiv \frac{1}{n}\sum_{i=1}^{n} G(Z_i,\alpha,\beta), \quad \widehat{h}(\alpha,\gamma) \equiv \frac{1}{n}\sum_{i=1}^{n} H(Z_i,\alpha,\gamma), \quad \hat{f}(\alpha,\beta,\gamma) \equiv \frac{1}{n}\sum_{i=1}^{n} F(Z_i,\alpha,\beta,\gamma),$$
$$g_0(\alpha,\beta) \equiv E\{G(Z,\alpha,\beta)\}, \qquad h_0(\alpha,\gamma) \equiv E\{H(Z,\alpha,\gamma)\}, \qquad f_0(\alpha,\beta,\gamma) \equiv E\{F(Z,\alpha,\beta,\gamma)\}.$$

The true and pseudo-true parameters are ($\theta^j = \theta_0^j$ if the model is correct, $j = g, h, f$)

$$\theta_0^g \equiv \{\alpha_0,\beta_0\},\ \theta_0^h \equiv \{\alpha_0,\gamma_0\},\ \theta_0^f \equiv \{\alpha_0,\beta_0,\gamma_0\}, \qquad \theta^g \equiv \{\alpha_g,\beta_g\},\ \theta^h \equiv \{\alpha_h,\gamma_h\},\ \theta^f \equiv \{\alpha_f,\beta_f,\gamma_f\},$$
$$c_g \equiv g_0(\theta^g) \neq 0 \text{ if } \theta^g \neq \theta_0^g, \qquad c_h \equiv h_0(\theta^h) \neq 0 \text{ if } \theta^h \neq \theta_0^h, \qquad c_f \equiv f_0(\theta^f) \neq 0 \text{ if } \theta^f \neq \theta_0^f.$$

With $\hat{\Omega}_g \to^p \Omega_g$ and $\hat{\Omega}_h \to^p \Omega_h$,

$$\{\widehat{\alpha}_g, \hat{\beta}_g\} \text{ minimizes } \widetilde{Q}^g(\alpha,\beta) \equiv \widehat{g}(\alpha,\beta)'\hat{\Omega}_g\widehat{g}(\alpha,\beta), \quad \{\widehat{\alpha}_h, \hat{\gamma}_h\} \text{ minimizes } \widetilde{Q}^h(\alpha,\gamma) \equiv \widehat{h}(\alpha,\gamma)'\hat{\Omega}_h\widehat{h}(\alpha,\gamma),$$
$$\{\alpha_g,\beta_g\} \text{ minimizes } \widetilde{Q}_0^g(\alpha,\beta) \equiv g_0(\alpha,\beta)'\Omega_g g_0(\alpha,\beta), \quad \{\alpha_h,\gamma_h\} \text{ minimizes } \widetilde{Q}_0^h(\alpha,\gamma) \equiv h_0(\alpha,\gamma)'\Omega_h h_0(\alpha,\gamma)$$
$$\hat{Q}^g(\alpha,\beta) \equiv \frac{\widetilde{Q}^g(\alpha,\beta)}{k_g}, \hat{Q}^h(\alpha,\gamma) \equiv \frac{\widetilde{Q}^h(\alpha,\gamma)}{k_h},\ \hat{Q}^g \equiv \hat{Q}^g(\widehat{\alpha}_g,\widehat{\beta}_g), \hat{Q}^h \equiv \hat{Q}^h(\widehat{\alpha}_h,\widehat{\gamma}_h),\ Q_0^g \equiv \frac{c_g'\Omega_g c_g}{k_g}, Q_0^h \equiv \frac{c_h'\Omega}{k}$$
$$\hat{W}_g \equiv \frac{\hat{Q}^g(\widehat{\alpha}_g,\widehat{\beta}_g)}{\hat{Q}^g(\widehat{\alpha}_g,\widehat{\beta}_g) + \hat{Q}^h(\widehat{\alpha}_h,\widehat{\gamma}_h)} \qquad \text{and} \qquad \hat{W}_f \equiv 1 - \frac{1}{n^\tau \hat{Q}^f(\widehat{\alpha}_f,\widehat{\beta}_f,\widehat{\gamma}_f) + 1}.$$

C.1.1 Proof of Lemma 4.1

To obtain the probability limits of $\hat{W}_g$ and $\hat{W}_f$, first we consider without loss of generality the probability limit of $\hat{Q}^g$ when model G is correctly specified, and when it's misspecified. The asymptotics for $\hat{Q}^h$ and $\hat{Q}^f$ are obtained following the same logic. After these derivations, we then obtain the probability limits of $\hat{W}_g$ and $\hat{W}_f$ based on $\hat{Q}^g$, $\hat{Q}^h$ and $\hat{Q}^f$. First we have

$$n\hat{Q}^g = \{\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)\}'\{\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)\}\frac{1}{k_g}. \tag{C.1}$$

From the first order condition for $\hat{\theta}^g$ minimizing $\widetilde{Q}^g(\theta)$, we have

$$\sqrt{n}\nabla_\theta \widehat{g}(\hat{\theta}^g)\cdot\hat{\Omega}_g\widehat{g}(\hat{\theta}^g)=0.$$

Taylor-expanding the last term $\widehat{g}(\hat{\theta}^g)$ around θ^g gives

$$\begin{aligned}0&=\sqrt{n}\nabla_\theta \widehat{g}(\hat{\theta}^g)\cdot\hat{\Omega}_g\{\widehat{g}(\theta^g)+\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)(\hat{\theta}^g-\theta^g)\}\\&=\nabla_\theta \widehat{g}(\hat{\theta}^g)\cdot\hat{\Omega}_g\sqrt{n}\widehat{g}(\theta^g)+\nabla_\theta \widehat{g}(\hat{\theta}^g)\cdot\hat{\Omega}_g\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)\sqrt{n}(\hat{\theta}^g-\theta^g)\end{aligned}$$

where $\overline{\theta}^g$ is a mean value between θ^g and $\hat{\theta}^g$. If the model is correctly specified, $\theta^g=\theta_0^g$. This gives

$$\sqrt{n}(\hat{\theta}^g-\theta^g)=-(\hat{H}^g)^{-1}\nabla_\theta \widehat{g}(\hat{\theta}^g)\cdot\hat{\Omega}_g\sqrt{n}\widehat{g}(\theta^g)\quad\text{where}\quad \hat{H}^g\equiv\nabla_\theta \widehat{g}(\widehat{\theta}^g)\hat{\Omega}_g\nabla_{\theta'}\widehat{g}(\overline{\theta}^g). \tag{C.2}$$

Case i). Suppose that G is correctly specified. By Assumption A1, A2, A3, A5, and A6, the conditions of Theorem 2.1 of in Newey and McFadden (1994) (uniqueness, compactness, continuity, and uniform convergence) hold for GMM estimation of model G, so that $\hat{\theta}^g\to^p\theta_0^g$. For $\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)$ in equation (C.1), expanding $\widehat{g}$ around θ_0^g, we have

$$\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)=\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)+\hat{\Omega}_g^{1/2}\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)\sqrt{n}(\hat{\theta}^g-\theta_0^g)$$

where $\overline{\theta}^g$ is a mean value between θ_0^g and $\hat{\theta}^g$. Plug equation (C.2) with θ^g replaced by θ_0^g into this equation to get

$$\begin{aligned}&\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)=\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)-\hat{\Omega}_g^{1/2}\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)(\hat{H}^g)^{-1}\nabla_\theta \widehat{g}(\hat{\theta}^g)\hat{\Omega}_g\sqrt{n}\widehat{g}(\theta_0^g)\\&=\{I_{\widetilde{k}_g}-\hat{\Omega}_g^{1/2}\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)(\hat{H}^g)^{-1}\nabla_\theta \widehat{g}(\hat{\theta}^g)\hat{\Omega}_g^{1/2}\}\cdot\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)=\hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)\quad\text{(C.3)}\\&\text{where}\quad\hat{\Pi}_g^*\equiv I_{\widetilde{k}_g}-\hat{\Omega}_g^{1/2}\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)(\hat{H}^g)^{-1}\nabla_\theta \widehat{g}(\hat{\theta}^g)\hat{\Omega}_g^{1/2}\end{aligned}$$

and $I_{\widetilde{k}_g}$ is the $\widetilde{k}_g\times\widetilde{k}_g$ identity matrix and $\widetilde{k}_g$ is the number of moments in the model G.

Under Assumption A7, A9, A10, A11 and A12, $\sqrt{n}\widehat{g}(\theta_0^g)\to^d N(0,\Sigma_g)$ where $\Sigma_g=E\{G(Z,\theta_0^g)G(Z,\theta_0^g)'\}$, and with $\Omega_g^{-1}=\Sigma_g$, $\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)\to^d N(0,I_{\widetilde{k}_g})$. By Assumption A11, $\nabla_\theta \widehat{g}(\overline{\theta}^g)\to^p\nabla_\theta g_0(\theta_0^g)$, $\nabla_\theta \widehat{g}(\hat{\theta}^g)\to^p\nabla_\theta g_0(\theta_0^g)$, and $\hat{H}^g\to^p H^g$ which is non-singular by Assumption A8. Then, we have

$$\hat{\Pi}_g^*\to^p\Pi_g^*\equiv I_{\widetilde{k}_g}-\Omega_g^{1/2}\nabla_{\theta'}g_0(\theta_0^g)(H^g)^{-1}\nabla_\theta g_0(\theta_0^g)\Omega_g^{1/2}$$

where $\hat{\Pi}_g^*$ is a $\widetilde{k}_g \times \widetilde{k}_g$ symmetric matrix that is idempotent with $trace(\Pi_g) = k_g$, $k_g \equiv \widetilde{k}_g - k_g^*$ where k_g^* is the number of parameters in the model G. Therefore,

$$n\hat{Q}^g = \{\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)\}'\hat{\Pi}_g^*\{\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta_0^g)\}/k_g \to^d \chi^2_{k_g}/k_g.$$

Case ii). Suppose that G is misspecified. Under Assumption A1, A3, A4, A5, and A6, $\hat{\theta}^g \to^p \theta^g \neq \theta_0^g$ by Lemma 1 of Hall (2000). For $\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)$ in (C.1), Taylor-expand $\widehat{g}$ around θ^g to get, with $c_g \equiv g_0(\alpha_g, \beta_g)$

$$\begin{aligned}\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g) &= \hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta^g) + \hat{\Omega}_g^{1/2}\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)\sqrt{n}(\hat{\theta}^g - \theta^g) \\ &= \hat{\Omega}_g^{1/2}\sqrt{n}\{\widehat{g}(\theta^g) - c_g\} + \hat{\Omega}_g^{1/2}\nabla_{\theta'}\widehat{g}(\overline{\theta}^g)\sqrt{n}(\hat{\theta}^g - \theta^g) + \hat{\Omega}_g^{1/2}\sqrt{n}c_g. \qquad \text{(C.4)}\end{aligned}$$

Under Assumption A7, A9, A10, A11 and $\hat{\Omega}_g^{1/2} \to^p \Omega_g^{1/2}$, the first term is asymptotically normal. Also, by Assumptions A12 to A15, using Theorem 2 of Hall and Inoue (2003), $\sqrt{n}(\hat{\theta}^g - \theta^g)$ is asymptotically normal with mean zero. Thus, the sum of first two terms in (C.4) are bounded in probability. However, the third term in (C.4) diverges at the rate $\sqrt{n}$ $(= O_p(n^{1/2}))$, and consequently, $n\hat{Q}^g$ diverges at the rate n as $n \to \infty$.

In short, the asymptotics of $n\hat{Q}^g$ is summarized as follows:

$$\begin{aligned}&\text{Case i) } G \text{ is correctly specified} &&\implies n\hat{Q}^g \to^d \chi^2_{k_g}/k_g \text{ as } n \to \infty \\ &\text{Case ii) } G \text{ is misspecified} &&\implies n\hat{Q}^g \text{ diverges as } n \to \infty.\end{aligned}$$

In the following, we investigate the probability limits of $\hat{W}_g$ and $\hat{W}_f$, using these results.

Case 1). Suppose both $g_0(\alpha_0, \beta_0) = 0$ and $h_0(\alpha_0, \gamma_0) = 0$. Then, $f_0(\alpha_0, \beta_0, \gamma_0) = 0$. By A1, A2, A3, A5, and A6, $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_0, \beta_0\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_0, \gamma_0\}$, and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_0, \beta_0, \gamma_0\}$, so $\hat{Q}^g \to^p 0$, $\hat{Q}^h \to^p 0$, and $\hat{Q}^f \to^p 0$. For $n^\tau\hat{Q}^f$, following the same derivation in (C.3), we have

$$\begin{aligned}n^\tau\hat{Q}^f &= n^\tau\hat{f}(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)'\hat{\Omega}_f\hat{f}(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)\frac{1}{k_f} \\ &= n^{\tau-1}\left\{\hat{\Pi}_f\hat{\Omega}_f^{1/2}\sqrt{n}\widehat{f}(\theta_0^f)\right\}'\left\{\hat{\Pi}_f\hat{\Omega}_f^{1/2}\sqrt{n}\widehat{f}(\theta_0^f)\right\}\frac{1}{k_f} = n^{\tau-1}\hat{\chi}^2_{k_f}\frac{1}{k_f}\end{aligned}$$

where $\hat{\Pi}_f \equiv I_{\widetilde{k}_f} - \hat{\Omega}_f^{1/2}\nabla_{\theta'}\widehat{f}(\overline{\theta}^f)(\hat{H}^f)^{-1}\nabla_\theta\widehat{f}(\hat{\theta}^f)\hat{\Omega}_f^{1/2}$ and $\hat{\chi}^2_{k_f} \equiv \{\hat{\Pi}_f\hat{\Omega}_f^{1/2}\sqrt{n}\widehat{f}(\theta_0^f)\}'\{\hat{\Pi}_f\hat{\Omega}_f^{1/2}\sqrt{n}\widehat{f}(\theta_0^f)\}$.

Following the same steps as in Case i) of Lemma 1, we have $\hat{\chi}^2_{k_f} \to^d \chi^2_{k_f}$ which is bounded in probability, and consequently, $n^\tau \hat{Q}^f \to^p 0$ for $\tau < 1$, and

$$\hat{W}_f = 1 - \frac{1}{n^\tau \hat{Q}^f + 1} = 1 - \frac{1}{n^{\tau-1} n\hat{Q}^f + 1} \to^p 0.$$

As for $\hat{W}_g$, due to $n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) \to_d \chi^2_{k_g}/k_g$ and $n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h) \to_d \chi^2_{k_h}/k_h$, $\hat{W}_g = n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)/\{n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) + n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\}$ converges to a ratio of possibly dependent random variables, which lies between zero and one with probability one. We do not need the limiting distribution of $\hat{W}_g$ [23], as it is enough to have $\hat{W}_g$ bounded in probability to ensure $\hat{W}_f \hat{W}_g \to^p 0$ when $\hat{W}_f \to^p 0$.

Case 2). Suppose $g_0(\alpha_0, \beta_0) = 0$ but $h_0(\alpha_0, \gamma_0) \neq 0$. Then $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_0, \beta_0\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_h, \gamma_h\}$, and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_f, \beta_f, \gamma_f\}$. By the continuous mapping theorem and uniform convergence of $\hat{Q}^g$ and $\hat{Q}^h$, we have $\hat{Q}^g \to^p Q_0^g = c_g' \Omega_g c_g/k_g = 0$, $\hat{Q}^h \to^p Q_0^h = c_h' \Omega_h c_h/k_h > 0$, and $\hat{Q}^f \to^p Q_0^f = c_f' \Omega_f c_f/k_f > 0$. From Case ii), $n\hat{Q}^h$ diverges as $n \to \infty$ while $n\hat{Q}^g$ is bounded in probability, and thus $\hat{W}_g = n\hat{Q}^g/(n\hat{Q}^g + n\hat{Q}^h) \to^p 0$. As for $\hat{W}_f$, due to $\hat{Q}^f \to^p Q_0^f = c_f' \Omega_f c_f/k_f > 0$, we have

$$\hat{W}_f = 1 - \frac{1}{n^\tau \hat{Q}^f + 1} = 1 - \frac{1}{n^{\tau-1} n\hat{Q}^f + 1} \to^p 1$$

and $\hat{W}_f \hat{W}_g \to^p 0$.

Case 3). Suppose now $g_0(\alpha_0, \beta_0) \neq 0$ but $h_0(\alpha_0, \gamma_0) = 0$. Then $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_g, \beta_g\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_0, \gamma_0\}$, and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_f, \beta_f, \gamma_f\}$. So $\hat{Q}^g \to^p Q_0^g = c_g' \Omega_g c_g/k_g > 0$, $\hat{Q}^h \to^p Q_0^h = c_h' \Omega_h c_h/k_h = 0$, and $\hat{Q}^f \to^p Q_0^f = c_f' \Omega_f c_f/k_f > 0$. Following the same argument as in Case 2), $\hat{W}_g \to^p 1$ and $\hat{W}_f \to^p 1$. In short, the probability limits of $\hat{W}_f$ and $\hat{W}_g \hat{W}_f$ are categorized as follows:

Case 1) Both G and H are correctly specified $\implies \hat{W}_f \to^p 0$ and $\hat{W}_f \hat{W}_g \to^p 0$
Case 2) G is correctly specified, but H is not $\implies \hat{W}_f \to^p 1$ and $\hat{W}_f \hat{W}_g \to^p 0$
Case 3) H is correctly specified, but G is not $\implies \hat{W}_f \to^p 1$ and $\hat{W}_f \hat{W}_{fg} \to^p 1$.

Q.E.D.

[23] If $\hat{Q}^g$ and $\hat{Q}^f$ happen to be independent, then $\hat{W}_g$ would be a ratio of independent Chi-squareds and so converges to a beta distribution with shape parameters $k_g/2$ and $k_k/2$. But there is no reason to impose that these distributions be independent.

C.1.2 Proof of Theorem 4.2

Recall equation (4.1) and rewrite it as

$$\widehat{\alpha} = \alpha_0 + \hat{W}_f\hat{W}_g(\widehat{\alpha}_h - \alpha_0) + \hat{W}_f\left(1 - \hat{W}_g\right)(\widehat{\alpha}_g - \alpha_0) + (1 - \hat{W}_f)(\widehat{\alpha}_f - \alpha_0)$$

$$\begin{aligned}
\implies \sqrt{n}(\widehat{\alpha} - \alpha_0) &= \hat{W}_f\hat{W}_g\sqrt{n}(\widehat{\alpha}_h - \alpha_0) + \hat{W}_f\left(1 - \hat{W}_g\right)\sqrt{n}(\widehat{\alpha}_g - \alpha_0) + (1 - \hat{W}_f)\sqrt{n}(\widehat{\alpha}_f - \alpha_0) \\
&= \hat{W}_f\hat{W}_g\sqrt{n}(\widehat{\alpha}_h - \alpha_h) + \hat{W}_f\left(1 - \hat{W}_g\right)\sqrt{n}(\widehat{\alpha}_g - \alpha_g) + (1 - \hat{W}_f)\sqrt{n}(\widehat{\alpha}_f - \alpha_f) \quad \text{(C.5)} \\
&\quad + \hat{W}_f\hat{W}_g\sqrt{n}(\alpha_h - \alpha_0) + \hat{W}_f\left(1 - \hat{W}_g\right)\sqrt{n}(\alpha_g - \alpha_0) + (1 - \hat{W}_f)\sqrt{n}(\alpha_f - \alpha_0)
\end{aligned}$$

Now we show the asymptotic normality of $\widehat{\alpha}$ and the form of $\widetilde{V}$ depending on which model is correct.

Case 1). Suppose G and H are both correct. Then, because of $\alpha_g = \alpha_h = \alpha_f = \alpha_0$, α_g, α_h, α_f in the first line of (C.5) are replaced by α_0, and the second line disappears. Following the same argument as in Theorem 3.4 of Newey and McFadden, under Assumption A7, A9, A10 and A11, the central limit theorem yields $n^{-1/2}\sum_i G(Z_i, \alpha_0, \beta_0) \to^d N(0, \Sigma_g)$ where $\Sigma_g = E\{G(Z, \alpha_0, \beta_0)G(Z, \alpha_0, \beta_0)'\}$. Along with $\widehat{g}(\widehat{\alpha}, \widehat{\beta}_g) \to^p g_0(\theta_0^g) = 0$ and $\nabla_\alpha \widehat{g}(\widehat{\alpha}, \widehat{\beta}_g) \to^p \nabla_\alpha g_0(\theta_0^g)$, we can establish asymptotic normality of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$. Following the same argument, along with the consistency of $(\widehat{\alpha}_h, \widehat{\beta}_h)$ and $(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f)$, the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_h - \alpha_0)$ and $\sqrt{n}(\widehat{\alpha}_f - \alpha_0)$ are established. Therefore, by Lemma 1 on $\hat{W}_f \to^p 0$ and $\hat{W}_g\hat{W}_f \to^p 0$, and boundedness of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$ and $\sqrt{n}(\widehat{\alpha}_h - \alpha_0)$ in probability, the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_f - \alpha_0)$, and the continuous mapping theorem, we have

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(0, \widetilde{V}^f).$$

By Assumption A8

$$\frac{1}{n}\sum_i \widehat{\eta}_i^f \widehat{\eta}_i^{f\prime} \to^p \widetilde{V}^f \equiv E(\eta^f \eta^{f\prime}) \qquad \text{where} \quad \sqrt{n}(\widehat{\alpha}_f - \alpha_0) = \frac{1}{\sqrt{n}}\sum_i \widehat{\eta}_i^f;$$

$\widehat{\eta}_i^f$ is the influence function of $\widehat{\alpha}_f$ (the details of $\widehat{\eta}_i^f$ are given in (C.14) of the Appendix III), making $\widetilde{V}^f = \widetilde{V}$.

Case 2). Suppose G is correct, but H is not ($\alpha_h - \alpha_0 \equiv \delta_h \neq 0$). In this case, F

is also misspecified ($\alpha_f - \alpha_0 \equiv \delta_f \neq 0$). Then, (C.5) can be rewritten as

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) = \hat{W}_f\hat{W}_g\sqrt{n}(\widehat{\alpha}_h - \alpha_h) + \hat{W}_f\left(1 - \hat{W}_g\right)\sqrt{n}(\widehat{\alpha}_g - \alpha_0) + (1 - \hat{W}_f)\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$$
$$+ \hat{W}_f\hat{W}_g\sqrt{n}\delta_h + (1 - \hat{W}_f)\sqrt{n}\delta_f \qquad \text{(C.6)}$$

By Theorem 1, $(\widehat{\alpha}_g, \widehat{\beta}_g) \to^p (\alpha_0, \beta_0)$, while $(\widehat{\alpha}_h, \widehat{\gamma}_h) \to^p (\alpha_h, \gamma_h)$ and $(\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f) \to^p (\alpha_f, \beta_f, \gamma_f)$. Following the same argument as above, we have the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$. Under Assumption A7, A9, A10, A11, A12, A13, A14, and A15, by Theorem 2 of Hall and Inoue (2003), $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$ is asymptotically normal with mean zero and a complex form of the variance. The same argument holds for $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ too. In the second line of (C.6), $\hat{W}_f\hat{W}_g\sqrt{n} = \left(1 - \frac{1}{O_p(n^\tau)+1}\right)\frac{O_p(1)}{O_p(1)+O_p(n)}O(\sqrt{n})$ and $(1 - \hat{W}_f)\sqrt{n} = \frac{1}{O_p(n^\tau)+1}O(\sqrt{n})$, and thus for $\tau > 1/2$, the second line disappears as $n \to \infty$. By Lemma 1 on $\hat{W}_f \to^p 1$ and $\hat{W}_g\hat{W}_f \to^p 0$, boundedness of $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$ and $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ in probability, the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$ and the continuous mapping theorem, we have $\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(0, \widetilde{V}^g)$. By Assumption A8, we get

$$\frac{1}{n}\sum_i \widehat{\eta}_i^g\widehat{\eta}_i^{g\prime} \to^p \widetilde{V}^g \equiv E(\eta^g\eta^{g\prime}) \qquad \text{where} \quad \sqrt{n}(\widehat{\alpha}_g - \alpha_0) = \frac{1}{\sqrt{n}}\sum_i \widehat{\eta}_i^g;$$

$\widehat{\eta}_i^g$ is the influence function of $\widehat{\alpha}_g$ (the details of $\widehat{\eta}_i^g$ are given in (C.16) of the Appendix III), making $\widetilde{V}^g = \widetilde{V}$.

Case 3). Suppose H is correct, but G is not ($\Longrightarrow \alpha_g - \alpha_0 \equiv \delta_g \neq 0$). Then the same argument as in Case 2) applies, replacing $\hat{W}_g$ with $1 - \hat{W}_g$, and switching the roles of β and γ and the roles of g and h.

C.2 Appendix II

Let the model G be "locally misspecified" when the parameter in the data generating process takes the form $\theta^g = \theta_0^g + \delta_g n^{-s}$ for a constant δ_g and $s > 0$, while θ_0^g satisfies $E\{G(Z, \theta_0^g)\} = 0$ due to Assumption A3. Analogously, let the model H be "locally misspecified" when the parameter in the data generating process is $\theta^h = \theta_0^h + \delta_h n^{-s}$ with $E\{H(Z, \theta_0^h)\} = 0$. When $s = 1/2$, $\delta_g n^{-s}$ is 'Pitman drift' as in Pitman (1949), Newey and West (1987), Bera and Yoon (1993) and Newey and McFadden (1994).

When model G or H is locally misspecified, we have, respectively,

$$g_0(\theta^g) \equiv g_0(\theta_0^g) + \nabla_{\theta'} g_0(\widetilde{\theta}^g)\delta_g n^{-s} = \nabla_{\theta'} g_0(\widetilde{\theta}^g)\delta_g n^{-s} \quad \text{with} \quad \omega_g \equiv \nabla_{\theta'} g_0(\widetilde{\theta}^g)\delta_g,$$
$$h_0(\theta^h) \equiv h_0(\theta_0^h) + \nabla_{\theta'} h_0(\widetilde{\theta}^h)\delta_h n^{-s} = \nabla_{\theta'} h_0(\widetilde{\theta}^h)\delta_h n^{-s} \quad \text{with} \quad \omega_h \equiv \nabla_{\theta'} h_0(\widetilde{\theta}^g)\delta_h,$$

$\widetilde{\theta}^g$ is a mean value between θ^g and θ_0^g, and $\widetilde{\theta}^h$ is a mean value between θ^h and θ_0^h.

Before presenting the detailed proofs, we summarize here our main findings when one of the models is locally misspecified but another is correctly specified. Suppose that model H is correctly specified and model G is locally misspecified, with $\theta^g = \theta_0^g + \delta_g n^{-s}$. This local misspecification does not affect the consistency of our estimator $\hat{\alpha}$, because the local misspecification reduces to the correct specification as $n \to \infty$ and the weights $\hat{W}_g$ and $\hat{W}_g\hat{W}_f$ still have finite probability limits under the local misspecification. As for asymptotic distribution, when $s > 0.5$, the limiting distribution of $\sqrt{n}(\hat{\alpha} - \alpha_0)$ is the same as when both models are correct, because the drift approaches 0 sufficiently quickly. Second, when $s = 0.5$, $\hat{\alpha}$ is consistent but not $\sqrt{n}$-consistent. Third, when $s < 0.5$, if $s + 0.5 < \tau$, then the asymptotic distribution of $\sqrt{n}(\hat{\alpha} - \alpha_0)$ is the same as if model G was globally misspecified (and is still $\sqrt{n}$-consistent, because asymptotically all weight goes on model H) .

Assumption A16: Either 1) model G is correct but model H is locally misspecified, or 2) model H is correct but model G is locally misspecified.

C.2.1 Lemma App.1

Lemma App.1: Let Assumption A1 and Assumptions A3 to A16 hold. For any τ with $0 < \tau < 1$, $\hat{W}_f$ and $\hat{W}_g\hat{W}_f$ have finite probability limits.

Proof for Lemma App.1.

Analogously to the proof for Lemma 1, first we consider without loss of generality the probability limit of $\hat{Q}^g$ when the model is locally misspecified. Then, the probability limits of $\hat{Q}^h$ and $\hat{Q}^f$ can be found following the same logic. Next, we find the probability limits of $\hat{W}_g$ and $\hat{W}_f$, based on those of $\hat{Q}^g$, $\hat{Q}^h$, and $\hat{Q}^f$.

Case iii). Suppose that G is locally misspecified ($\theta^g = \theta_0^g + \delta_g n^{-s}$). Replacing θ_0^g with θ^g in (C.3) gives

$$\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g) = \hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\theta^g) = \hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\sqrt{n}\{\widehat{g}(\theta^g) - \omega_g n^{-s}\} + \hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\omega_g n^{1/2-s} \quad \text{(C.7)}$$

note $E\{\widehat{g}(\theta^g)\} = g_0(\theta^g) = \nabla_{\theta'} g_0(\widetilde{\theta}^g)\delta_g n^{-s} = \omega_g n^{-s}$. Under Assumption A1, A2, A3, A5, and A6, the corresponding GMM estimator is still consistent $\hat{\theta}^g \to^p \theta_0^g$ by Theorem 9.1 of Newey and McFadden (1994). By Assumption A8, A11 and A12, $\nabla_\theta \widehat{g}(\bar{\theta}^g) \to^p \nabla_\theta g(\theta_0^g)$ for $\bar{\theta}^g$ in $\hat{\Pi}_g^*$ and $\hat{\Omega}_g^{-1} \to^p \Omega_g^{-1} = E\{G(Z,\theta_0^g)G(Z,\theta_0^g)'\}$, and thus, $\hat{\Pi}_g^* \to^p \Pi_g^*$, which is a $\widetilde{k}_g \times \widetilde{k}_g$ symmetric and idempotent matrix with $trace(\Pi_g) = k_g$. Therefore, applying the same the argument in Case i) of Lemma 1, along with the consistency of $\hat{\theta}^g$, the first term in the right-hand side of (C.7) is asymptotically standard normal, and thus bounded in probability. Consequently, we can characterize the asymptotics of $\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)$ depending on s using the last term $\hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\omega_g n^{1/2-s}$ in (C.7).

If $s = 1/2$, then $\hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\omega_g n^{1/2-s} \to^p \Pi_g^*\Omega_g^{1/2}\omega_g$ and $\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)$ is asymptotically normal with mean $\Pi_g^*\Omega_g^{1/2}\omega_g$ and unit variance. Hence,

$$n\hat{Q}^g = \{\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g)\}'\hat{\Omega}_g^{1/2}\sqrt{n}\widehat{g}(\hat{\theta}^g) \to^d \chi^2_{k_g}(\omega_g'\Omega_g^{1/2}\Pi_g^*\Omega_g^{1/2}\omega_g)/k_g;$$

$\chi^2_{k_g}(\omega_g'\Omega_g^{1/2}\Pi_g^*\Omega_g^{1/2}\omega_g)$ is the noncentral chi-squared distribution with noncentrality parameter $\omega_g'\Omega_g^{1/2}\Pi_g^*\Omega_g^{1/2}\omega_g$. If $s > 1/2$ in (C.7), the noncentrality parameter shrinks to zero, so that $n\hat{Q}^g \to^d \chi^2_{k_g}(0)/k_g$ as $n \to \infty$, analogously to Case i) of Lemma 1. If $s < 1/2$ in (C.7), then $\hat{\Pi}_g^*\hat{\Omega}_g^{1/2}\omega_g n^{1/2-s} = O_p(n^{1/2-s})$ diverges as $n \to \infty$, analogously to Case ii) of Lemma 1. In short,

$$\begin{aligned}
&\text{Case iii) with } s < 1/2 \quad \Longrightarrow n\hat{Q}^g \text{ diverges as } n \to \infty;\\
&\text{Case iii) with } s = 1/2 \quad \Longrightarrow n\hat{Q}^g \to^d \chi^2_{k_g}(\omega_g'\Omega_g^{1/2}\Pi_g^*\Omega_g^{1/2}\omega_g)/k_g \text{ as } n \to \infty;\\
&\text{Case iii) with } s > 1/2 \quad \Longrightarrow n\hat{Q}^g \to^d \chi^2_{k_g}(0)/k_g \text{ as } n \to \infty.
\end{aligned}$$

Next, we investigate the probability limits of $\hat{W}_g$ and $\hat{W}_f$ based on those of $\hat{Q}^g$, $\hat{Q}^h$, and $\hat{Q}^f$, doing analogously to what was done for Case iii).

Case 4). Suppose that model G is correct, but H is locally misspecified with $\theta^h = \theta_0^h + \delta_h n^{-s}$. In this case, F is also locally misspecified with $\theta^f = \theta_0^f + \delta_f n^{-s}$ for some δ_f.

Case 4-1). If $s = 1/2$, as shown in Case iii), $n\hat{Q}^h \to_d \chi^2_{k_h}(\omega_h'\Omega_h^{1/2}\Pi_h^*\Omega_h^{1/2}\omega_h)/k_h$ and $n\hat{Q}^f \to_d \chi^2_{k_f}(\omega_f'\Omega_f^{1/2}\Pi_f^*\Omega_f^{1/2}\omega_f)/k_f$ for some ω_f as $n \to \infty$. Thus $\hat{W}_g = n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g)/\{n\hat{Q}^g(\widehat{\alpha}_g, \widehat{\beta}_g) + n\hat{Q}^h(\widehat{\alpha}_h, \widehat{\gamma}_h)\}$ converges to a distribution on $(0,1)$. For $\hat{W}_f$, we have

$$\hat{W}_f = 1 - \frac{1}{n^\tau \hat{Q}^f + 1} = 1 - \frac{1}{n^{\tau-1} n\hat{Q}^f + 1} \to^p 0,$$

because $n\hat{Q}^f$ is bounded in probability, and $n^{\tau-1} \to^p 0$. Thus, $\hat{W}_g\hat{W}_f \to^p 0$.

Case 4-2). If $s > 1/2$, as shown in Case iii), $n\hat{Q}^h \to_d \chi^2_{k_h}/k_h$, and $n\hat{Q}^f \to_d \chi^2_{k_f}/k_f$. Therefore, it is asymptotically the same as Case 1) of Lemma 1.

Case 4-3). If $s < 1/2$, as shown in Case iii), $n\hat{Q}^h$ and $n\hat{Q}^f$ are $O_p(n^{2(1/2-s)})$, as each is a squared version of a term analogous to (C.7). In this case, whereas $\hat{W}_g \to^p 0$, convergence of $\hat{W}_f$ depends on the relationship between τ and s. Because $n^\tau\hat{Q}^f = O(n^{\tau-1})O_p(n^{2(1/2-s)}) = O_p(n^{\tau-2s})$, when $\tau > 2s$, $n^\tau\hat{Q}^f$ diverges to result in $\hat{W}_f \to^p 1$ and $\hat{W}_g\hat{W}_f \to^p 0$. When $\tau < 2s$, $n^\tau\hat{Q}^f \to^p 0$, and consequently $\hat{W}_f \to^p 0$ and $\hat{W}_f\hat{W}_g \to^p 0$. When $\tau = 2s$, however, (C.7) shows that $n^\tau\hat{Q}^f \to^p \omega_f'\Omega_f^{1/2}\Pi_f^*\Omega_f^{1/2}\omega_f$ because only the last term of (C.7) matters, so that $\hat{W}_f \to^p W_f^* \equiv 1 - (\omega_f'\Omega_f^{1/2}\Pi_f^*\Omega_f^{1/2}\omega_f + 1)^{-1}$ and $\hat{W}_g\hat{W}_f \to^p 0$.

Case 5). Suppose that model G is locally misspecified with $\theta^g = \theta_0^g + \delta_g n^{-s}$, but model H is correct. Then essentially the same arguments as in Case 4) apply.

Case 5-1). If $s = 1/2$, then $n\hat{Q}^g \to_d \chi^2_{k_g}(\omega_g'\Omega_g^{1/2}\Pi_g^*\Omega_g^{1/2}\omega_g)/k_g$ and $n\hat{Q}^f \to_d \chi^2_{k_f}(\omega_f'\Omega_f^{1/2}\Pi_f^*\Omega_f^{1/2}\omega_f)/k_f$. Thus, $\hat{W}_f \to^p 0$ and $\hat{W}_g\hat{W}_f \to^p 0$.

Case 5-2). If $s > 1/2$, then $n\hat{Q}^g \to_d \chi^2_{k_g}/k_g$, and $n\hat{Q}^f \to_d \chi^2_{k_f}/k_f$ as $n \to \infty$, which is asymptotically the same as Case 1) of Lemma 1.

Case 5-3). If $s < 1/2$, then since $n\hat{Q}^g$ and $n\hat{Q}^f$ diverge, $\hat{W}_g \to^p 1$ but the asymptotics of $\hat{W}_f$ depends on the relationship between τ and s. For $\tau > 2s$, $n^{\tau-1}n\hat{Q}^f$ diverges, and thus, $\hat{W}_f \to^p 1$ and $\hat{W}_g\hat{W}_f \to^p 1$; for $\tau < 2s$, $\hat{W}_f \to^p 0$ and $\hat{W}_g\hat{W}_f \to^p 0$. When $\tau = 2s$, $\hat{W}_f \to^p W_f^*$ and $\hat{W}_g\hat{W}_f \to^p W_f^*$ because $\hat{W}_g \to^p 1$.

In sum, the probability limits of $\hat{W}_f$ and $\hat{W}_g\hat{W}_f$ are categorized as follows:

Case 4-1) and 4-2) with $s \geq 1/2$,	$\Longrightarrow \hat{W}_f \to^p 0$ and $\hat{W}_g\hat{W}_f \to^p 0$
Case 4-3) with $s < 1/2$ and $2s < \tau$	$\Longrightarrow \hat{W}_f \to^p 1$ and $\hat{W}_g\hat{W}_f \to^p 0$
Case 4-3) with $s < 1/2$ and $\tau = 2s$	$\Longrightarrow \hat{W}_f \to^p W_f^*$ and $\hat{W}_g\hat{W}_f \to^p 0$
Case 4-3) with $s < 1/2$ and $\tau < 2s$	$\Longrightarrow \hat{W}_f \to^p 0$ and $\hat{W}_g\hat{W}_f \to^p 0$
Case 5-1) and 5-2) with $s \geq 1/2$,	$\Longrightarrow \hat{W}_f \to^p 0$ and $\hat{W}_g\hat{W}_f \to^p 0$
Case 5-3) with $s < 1/2$ and $2s < \tau$	$\Longrightarrow \hat{W}_f \to^p 1$ and $\hat{W}_g\hat{W}_f \to^p 1$
Case 5-3) with $s < 1/2$ and $\tau = 2s$	$\Longrightarrow \hat{W}_f \to^p W_f^*$ and $\hat{W}_g\hat{W}_f \to^p W_f^*$
Case 5-3) with $s < 1/2$ and $\tau < 2s$	$\Longrightarrow \hat{W}_f \to^p 0$ and $\hat{W}_g\hat{W}_f \to^p 0$.

Q.E.D.

C.2.2 Theorem App.1

Theorem App.1: Under Assumptions A1 and A3 to A16, for $\widehat{\alpha}$ given by equation (4.1), $\widehat{\alpha} \to^p \alpha_0$.

Proof for Theorem App.1.

Case 4). Suppose that G is correct, but H is the locally misspecified with $\theta^h = \theta_0^h + \delta_h n^{-s}$. By Theorem 9.1 of in Newey and McFadden (1994), still $\{\widehat{\alpha}_g, \widehat{\beta}_g\} \to^p \{\alpha_0, \beta_0\}$, $\{\widehat{\alpha}_h, \widehat{\gamma}_h\} \to^p \{\alpha_0, \gamma_0\}$ and $\{\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}_f\} \to^p \{\alpha_0, \beta_0, \gamma_0\}$. By Lemma App.1, if $s \geq 1/2$, then $\hat{W}_f \to^p 0$ and $\hat{W}_g \hat{W}_f \to^p 0$, and the consistency of $\widehat{\alpha}$ in (4.1) follows from consistency of $\widehat{\alpha}_f$. If $s < 1/2$, the probability limits of $\hat{W}_f$ and $\hat{W}_g \hat{W}_f$ depend on the relationship between τ and s. If $s < 1/2$ and $\tau < 2s$, the limits are the same as in the case with $s \geq 1/2$ by Lemma App.1, and thus, the same argument holds for $\widehat{\alpha}$. If $s < 1/2$ and $\tau > 2s$, by Lemma App.1 $\hat{W}_f \to^p 1$ and $\hat{W}_g \hat{W}_f \to^p 0$ and the consistency of $\widehat{\alpha}$ follows from the consistency of $\widehat{\alpha}_g$. If $s < 1/2$ and $\tau = 2s$, then by Lemma App.1 $\hat{W}_f \to^p W_f^*$ and $\hat{W}_g \hat{W}_f \to^p 0$ and the consistency of $\widehat{\alpha}$ follows from the consistency of $\widehat{\alpha}_g$ and $\widehat{\alpha}_f$, and $\widehat{\alpha}_g - \widehat{\alpha}_f \to^p 0$.

Case 5). Suppose that H is correct, but G is locally misspecified. Then, essentially the same arguments as in Case 4 apply. Q.E.D.

C.2.3 Theorem App.2

Theorem App.2: Under Assumptions A1 and A3 to A16, for $1/2 < \tau < 1$, when $s > 1/2$ or $s + 1/2 < \tau$, there exists a matrix $\widetilde{V}$ such that

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(0, \widetilde{V}),$$

and

$$\frac{1}{n} \sum_i \widehat{\eta}_i \widehat{\eta}_i' \to^p \widetilde{V}$$

$$\text{where } \widehat{\eta}_i \equiv \hat{W}_f \hat{W}_g \widehat{\eta}_i^h + \hat{W}_f (1 - \hat{W}_g) \widehat{\eta}_i^g + (1 - \hat{W}_f) \widehat{\eta}_i^f.$$

Proof of Theorem App.2.

To ease referencing, recall (C.5):

$$\sqrt{n}(\widehat{\alpha}-\alpha_0) = \hat{W}_f\hat{W}_g\sqrt{n}(\widehat{\alpha}_h-\alpha_h) + \hat{W}_f\left(1-\hat{W}_g\right)\sqrt{n}(\widehat{\alpha}_g-\alpha_g) + (1-\hat{W}_f)\sqrt{n}(\widehat{\alpha}_f-\alpha_f)$$
$$+ \hat{W}_f\hat{W}_g\sqrt{n}(\alpha_h-\alpha_0) + \hat{W}_f\left(1-\hat{W}_g\right)\sqrt{n}(\alpha_g-\alpha_0) + (1-\hat{W}_f)\sqrt{n}(\alpha_f-\alpha_0).$$

Case 4). Suppose model G is correct ($\alpha_g = \alpha_0$), but H is locally misspecified with $\alpha_h = \alpha_0 + \delta_h n^{-s}$; then F is also locally misspecified with $\alpha_f = \alpha_0 + \delta_f n^{-s}$. Rewrite (C.5) as

$$\sqrt{n}(\widehat{\alpha}-\alpha_0) = \hat{W}_f\hat{W}_g\sqrt{n}(\widehat{\alpha}_h-\alpha_h) + \hat{W}_f\left(1-\hat{W}_g\right)\sqrt{n}(\widehat{\alpha}_g-\alpha_0) + (1-\hat{W}_f)\sqrt{n}(\widehat{\alpha}_f-\alpha_f)$$
$$+ \hat{W}_f\hat{W}_g\delta_h n^{1/2-s} + (1-\hat{W}_f)\delta_f n^{1/2-s}. \tag{C.8}$$

Following the same argument in Case 1) of Theorem 2, we can establish asymptotic normality of $\sqrt{n}(\widehat{\alpha}_g-\alpha_0)$. Call (C.15) in the Appendix III below replacing g with h to have

$$\sqrt{n}(\widehat{\alpha}_h-\alpha_h) = \widehat{A}_h^{-1}\nabla_\alpha\widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^*\sqrt{n}\widehat{h}(\theta^h)$$

where

$$\widehat{A}_h \equiv \nabla_\alpha\widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^*\nabla_{\alpha'}\widehat{h}(\overline{\theta}^h), \quad \hat{\Omega}_h^* \equiv \hat{\Omega}_h^{1/2\prime}\hat{\Pi}_h\hat{\Omega}_h^{1/2},$$
$$\hat{\Pi}_h \equiv [I_{\tilde{k}_h} - \hat{\Omega}_h^{1/2}\nabla_{\gamma'}\widehat{h}(\overline{\theta}^h)\{\nabla_\gamma\widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h\nabla_{\gamma'}\widehat{h}(\overline{\theta}^h)\}^{-1}\nabla_\gamma\widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^{1/2\prime}].$$

With $h_0(\theta^h) \equiv \omega_h n^{-s}$, add and subtract $E\{\widehat{h}(\theta^h)\} = \omega_h n^{-s}$ to get

$$\sqrt{n}(\widehat{\alpha}_h-\alpha_h) = \widehat{A}_h^{-1}\nabla_\alpha\widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^*\sqrt{n}\{\widehat{h}(\theta^h)-\omega_h n^{-s}\} + \widehat{A}_h^{-1}\nabla_\alpha\widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^*\sqrt{n}\omega_h n^{-s}. \tag{C.9}$$

Applying the vectorization part in Hall and Inoue (2003, p.367) and using the population first-order condition $\nabla_\alpha h_0(\theta^h)\Omega_h^* h_0(\theta^h) = 0$, rewrite $\nabla_\alpha\widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^*\sqrt{n}\omega_h n^{-s}$ in the last term other than $\widehat{A}_h^{-1}$ as

$$\sqrt{n}\nabla_\alpha\widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^*\omega_h n^{-s} =$$
$$\sqrt{n}\{\nabla_\alpha\widehat{h}(\hat{\theta}^h) - \nabla_\alpha\widehat{h}(\theta^h)\}\hat{\Omega}_h^*\omega_h n^{-s} + \sqrt{n}\{\nabla_\alpha\widehat{h}(\theta^h) - \nabla_\alpha h_0(\theta^h)\}\hat{\Omega}_h^*\omega_h n^{-s} + \nabla_\alpha h_0(\theta^h)\sqrt{n}(\hat{\Omega}_h^* - \Omega_h^*)\omega_h n^{-}$$
$$= \omega_h n^{-s}\hat{M}^h\sqrt{n}(\widehat{\alpha}_h-\alpha_h) + \sqrt{n}\{\nabla_\alpha\widehat{h}(\theta^h) - \nabla_\alpha h_0(\theta^h)\}\hat{\Omega}_h^*\omega_h n^{-s} + \nabla_\alpha h_0(\theta^h)\sqrt{n}(\hat{\Omega}_h^* - \Omega_h^*)\omega_h n^{-s} \tag{C.10}$$

for some symmetric $k_h^* \times k_h^*$ matrix $\hat{M}^h$ involving the second-order derivative of $h(\cdot)$ that is bounded in probability. Plugging (C.10) into (C.9) and solving for $\sqrt{n}(\widehat{\alpha}_h-\alpha_h)$

gives

$$\sqrt{n}(\widehat{\alpha}_h - \alpha_h) = [I_{k_h^*} - \widehat{A}_h^{-1}\omega_h n^{-s}\hat{M}^h]^{-1}\widehat{A}_h^{-1}\widehat{\Gamma}_h, \tag{C.11}$$

$$\begin{aligned}\widehat{\Gamma}_h \equiv{}& \nabla_\alpha \widehat{h}(\widehat{\theta}^h)\hat{\Omega}_h^*\sqrt{n}\{\widehat{h}(\theta^h) - \omega_h n^{-s}\} \\ &+ \sqrt{n}\{\nabla_\alpha \widehat{h}(\theta^h) - \nabla_\alpha h_0(\theta^h)\}\hat{\Omega}_h^*\omega_h n^{-s} + \nabla_\alpha h_0(\theta^h)\sqrt{n}(\hat{\Omega}_h^* - \Omega_h^*)\omega_h n^{-s}.\end{aligned}$$

Under Assumptions A12 to A16, $\sqrt{n}\{\nabla_\alpha \widehat{h}(\theta^h) - \nabla_\alpha h_0(\theta^h)\}$ and $\sqrt{n}(\hat{\Omega}_h^* - \Omega_h^*)$ are bounded in probability, so that the last two terms of $\widehat{\Gamma}_h$ converge to zero in probability. Under Assumption A7, A9, A10 and A11, $\sqrt{n}\{\widehat{h}(\theta^h) - \omega_h n^{-s}\}$ is asymptotically normal with mean zero. Therefore, due to $\widehat{\theta}_h \to^p \theta_0$ (from Theorem 9.1 of Newey and McFadden, 1994), $\omega_h n^{-s}\hat{M}^h \to^p 0$, $\nabla_\theta \widehat{h}(\widehat{\theta}^h) \to^p \nabla_\theta h_0(\theta_0^h)$, $\hat{\Omega}_h^* \to^p \Omega_h^*$, $\hat{\Omega}_h^{-1} \to^p \Omega_h^{-1} = E\{H(Z,\alpha_0,\gamma_0)H(Z,\alpha_0,\gamma_0)'\}$, and the continuous mapping theorem, we get

$$\sqrt{n}(\widehat{\alpha}_h - \alpha_h) \to^d N(0, \widetilde{V}^h)$$

where $\widetilde{V}^h$ is the same asymptotic variance as in Case 3) as if model H were correct. Analogously, the same argument holds for $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$, so that we have $\sqrt{n}(\widehat{\alpha}_f - \alpha_f) \to^d N(0, \widetilde{V}^f)$ as if model F were correct. Hence, all of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$, $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$ and $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ in the first line of (C.8) are asymptotically normal with mean zero and variance being that of the corresponding GMM estimator under correct specification.

Recall (C.8):

$$\begin{aligned}\sqrt{n}(\widehat{\alpha} - \alpha_0) = {}& \hat{W}_f\hat{W}_g\sqrt{n}(\widehat{\alpha}_h - \alpha_h) + \hat{W}_f\left(1 - \hat{W}_g\right)\sqrt{n}(\widehat{\alpha}_g - \alpha_0) + (1 - \hat{W}_f)\sqrt{n}(\widehat{\alpha}_f - \alpha_f) \\ &+ \hat{W}_f\hat{W}_g\delta_h n^{1/2-s} + (1 - \hat{W}_f)\delta_f n^{1/2-s}.\end{aligned}$$

Recalling (C.7) and its "squared version", we have

$$n\hat{Q}^h = O_p(n^{2(1/2-s)}) \quad \text{and} \quad n\hat{Q}^f = O_p(n^{2(1/2-s)}) \Longrightarrow n^\tau \hat{Q}^f = n^{\tau-1}n\hat{Q}^f = O_p(n^{\tau-1+2(1/2-s)}) = O_p(n^{\tau-2s})$$

Consequently, for the last two terms in (C.8), we get

$$\begin{aligned}&\hat{W}_f\hat{W}_g\delta_h n^{1/2-s} + (1 - \hat{W}_f)\delta_f n^{1/2-s} = \left(1 - \frac{1}{n^\tau \hat{Q}^f + 1}\right)\left(\frac{n\hat{Q}^g \cdot \delta_h n^{1/2-s}}{n\hat{Q}^g + n\hat{Q}^h}\right) + \left(\frac{1}{n^\tau \hat{Q}^f + 1}\right)\delta_f n^{1/2-s} \\ &= \left(1 - \frac{1}{O_p(n^{\tau-2s}) + 1}\right)\left(\frac{O_p(1)O(n^{1/2-s})}{O_p(1) + O_p(n^{2(1/2-s)})}\right) + \left(\frac{1}{O_p(n^{\tau-2s}) + 1}\right)O(n^{1/2-s}).\end{aligned} \tag{C.12}$$

For the first term in the left-hand side in (C.12), if $s = 1/2$, its probability limit is zero because $\hat{W}_f \to^p 0$ and $\hat{W}_g n^{1/2-s}$ is bounded in probability. If $s > 1/2$, the probability limit is zero because $\hat{W}_f \to^p 0$ and $\hat{W}_g n^{1/2-s} \to^p 0$. If $s < 1/2$, the probability limit is zero because $\hat{W}_f$ is bounded between zero and one in probability and $\hat{W}_g n^{1/2-s} \to^p 0$. Therefore, the first term in the left-hand side in (C.12) disappears as $n \to \infty$, regardless of s. However, the probability limit of the second term $(1 - \hat{W}_f)\delta_f n^{1/2-s}$ in the left-hand side in (C.12) varies, depending on the relationship between s and τ. So, the asymptotic behavior of $\sqrt{n}(\widehat{\alpha} - \alpha_0)$ depends on the values of s and τ as follows.

Case 4-1). If $s = 1/2$, $\hat{W}_f \to^p 0$ and $(1-\hat{W}_f)\delta_f n^{1/2-s} \to^p \delta_f$ as $n \to \infty$. By Lemma App.1, $\hat{W}_f \to^p 0$ and $\hat{W}_g\hat{W}_f \to^p 0$, boundedness of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$ and $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$ in probability, the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ and the continuous mapping theorem, only $(1 - \hat{W}_f)\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ survives in (C.8) and we get

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(\delta_f, \widetilde{V}^f).$$

Case 4-2). If $s > 1/2$, $\hat{W}_f \to^p 0$ and $(1 - \hat{W}_f)\delta_f n^{1/2-s} \to^p 0$ as $n \to \infty$. Therefore, we get $\hat{W}_g\hat{W}_f \to^p 0$ by Lemma App.1. Hence,

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(0, \widetilde{V}^f),$$

which is asymptotically equivalent to Case 1).

Case 4-3). If $s < 1/2$, the probability limit of the second term in (C.12) depends on s and τ:

$$(1 - \hat{W}_f)\delta_f n^{1/2-s} = \left(\frac{1}{O_p(n^{\tau-2s}) + 1}\right) O(n^{1/2-s}) = O_p(n^{s+1/2-\tau}).$$

When $s + 1/2 < \tau$, $(1 - \hat{W}_f)\delta_f n^{1/2-s}$ disappears as $n \to \infty$, which implies that the second line of (C.8) disappears. Due to $s < 1/2$,

$s + 1/2 < \tau \Longrightarrow 2s < 1/2 + s < \tau \Longrightarrow \hat{W}_f \to^p 1$ because of $2s < \tau$, and $\hat{W}_g\hat{W}_f \to^p 0$

by Lemma App.1. Therefore, by the boundedness of $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$ and $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ in probability, the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$ and the continuous mapping theorem, we have

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(0, \widetilde{V}^g),$$

which is asymptotically equivalent to Case 2) of Theorem 2.

When $s + 1/2 > \tau$, $(1 - \hat{W}_f)\delta_f n^{1/2-s}$ diverges as $n \to \infty$. Therefore, $\sqrt{n}(\widehat{\alpha} - \alpha_0)$ is not bounded in probability.

When $s + 1/2 = \tau$, $(1 - \hat{W}_f)\delta_f n^{1/2-s}$ converges to a constant, say ν, times δ_f, as $n \to \infty$. Also, we have

$$s + 1/2 = \tau \Longrightarrow 2s < 1/2 + s = \tau \Longrightarrow \hat{W}_f \to^p 1 \text{ because of } 2s < \tau, \text{ and } \hat{W}_g \hat{W}_f \to^p 0.$$

Therefore, by the boundedness of $\sqrt{n}(\widehat{\alpha}_h - \alpha_h)$ and $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$ in probability, the asymptotic normality of $\sqrt{n}(\widehat{\alpha}_g - \alpha_0)$ and the continuous mapping theorem, we get

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(\nu\delta_f, \widetilde{V}^g).$$

In sum, when G is correct but H is locally misspecified, $\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(0, \widetilde{V}^f)$ if $s > 1/2$, or $\sqrt{n}(\widehat{\alpha} - \alpha_0) \to^d N(0, \widetilde{V}^g)$ if $s + 1/2 < \tau$.

Case 5). Suppose H is correct specified, but G is locally misspecified with $\alpha^g = \alpha_0^g + \delta_g n^{-s}$. Then essentially the same arguments as in Case 4) apply, replacing $\hat{W}_g$ with $1 - \hat{W}_g$, and switching the roles of β and γ and the roles of g and h. Q.E.D.

C.3 Appendix III

Derivation of $\widehat{\eta}_i^f$, $\widehat{\eta}_i^g$, and $\widehat{\eta}_i^h$.

To find the influence functions $\widehat{\eta}_i^f$, let $\widehat{\theta}^f$ denote the first-stage estimator

$$\widehat{\theta}^f \equiv (\widehat{\alpha}_f, \widehat{\beta}_f, \widehat{\gamma}^f) = \arg \min_{\{\alpha,\beta,\gamma\} \in \Theta_\alpha \times \Theta_\beta \times \Theta_\gamma} \widetilde{Q}^f(\alpha, \beta, \gamma) = \widehat{f}(\alpha, \beta, \gamma)\hat{\Omega}_f \widehat{f}(\alpha, \beta, \gamma).$$

Under Assumption A7 and A10-12, the following first-order conditions in the first-stage for $\widehat{\theta}^f$ hold:

$$FD_\alpha^f = \frac{\partial \widetilde{Q}^f(\widehat{\theta}^f)}{\partial \alpha} = \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \widehat{f}(\widehat{\theta}^f) = 0, \qquad FD_\beta^f = \frac{\partial \widetilde{Q}^f(\widehat{\theta}^f)}{\partial \beta} = \nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \widehat{f}(\widehat{\theta}^f) = 0,$$
$$FD_\gamma^f = \frac{\partial \widetilde{Q}^f(\widehat{\theta}^f)}{\partial \gamma} = \nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f \widehat{f}(\widehat{\theta}^f) = 0.$$

Expend $\widehat{f}$ around the unique minimizer $\theta^f \equiv \{\alpha_f, \beta_f, \gamma_f\}$ to get

$$\widehat{f}(\widehat{\theta}^f) = \widehat{f}(\theta^f) + \nabla_{\alpha'} \widehat{f}(\overline{\theta}^f)(\widehat{\alpha}_f - \alpha_f) + \nabla_{\beta'} \widehat{f}(\overline{\theta}^f)(\widehat{\beta} - \beta_f) + \nabla_{\gamma'} \widehat{f}(\overline{\theta}^f)(\widehat{\gamma} - \gamma_f)$$

where $\overline{\theta}^f$ is the mean value to apply the mean value theorem. Substitute these into each FD^f to get

$$
\begin{aligned}
FD^f_\alpha &= \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\{\widehat{f}(\theta^f) + \nabla_{\alpha'}\widehat{f}(\overline{\theta}^f)(\widehat{\alpha}_f - \alpha_f) + \nabla_{\beta'}\widehat{f}(\overline{\theta}^f)(\widehat{\beta}_f - \beta_f) + \nabla_{\gamma'}\widehat{f}(\overline{\theta}^f)(\widehat{\gamma} - \gamma_f)\},\\
FD^f_\beta &= \nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\{\widehat{f}(\theta^f) + \nabla_{\alpha'}\widehat{f}(\overline{\theta}^f)(\widehat{\alpha}_f - \alpha_f) + \nabla_{\beta'}\widehat{f}(\overline{\theta}^f)(\widehat{\beta}_f - \beta_f) + \nabla_{\gamma'}\widehat{f}(\overline{\theta}^f)(\widehat{\gamma} - \gamma_f)\}.\\
FD^f_\gamma &= \nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\{\widehat{f}(\theta^f) + \nabla_{\alpha'}\widehat{f}(\overline{\theta}^f)(\widehat{\alpha}_f - \alpha_f) + \nabla_{\beta'}\widehat{f}(\overline{\theta}^f)(\widehat{\beta}_f - \beta_f) + \nabla_{\gamma'}\widehat{f}(\overline{\theta}^f)(\widehat{\gamma} - \gamma_f)\},\\
FD^f &= \{FD^f_\alpha, FD^f_\beta, FD^f_\gamma\} = \widehat{I}^f + \widehat{H}^f(\widehat{\theta}^f - \theta^f), \text{ and from these, } \sqrt{n}(\widehat{\theta}^f - \theta^f) = \widehat{H}^{f-1}\sqrt{n}\widehat{I}^f,
\end{aligned}
$$

$$
\widehat{I}^f \equiv \begin{bmatrix} \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\widehat{f}(\theta^f) \\ \nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\widehat{f}(\theta^f) \\ \nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\widehat{f}(\theta^f) \end{bmatrix}, \widehat{H}^f \equiv \begin{bmatrix} \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\nabla_{\alpha'}\widehat{f}(\overline{\theta}^f) & \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\nabla_{\beta'}\widehat{f}(\overline{\theta}^f) & \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\nabla_{\gamma'}\widehat{f}(\overline{\theta}^f) \\ \nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\nabla_{\alpha'}\widehat{f}(\overline{\theta}^f) & \nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\nabla_{\beta'}\widehat{f}(\overline{\theta}^f) & \nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\nabla_{\gamma'}\widehat{f}(\overline{\theta}^f) \\ \nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\nabla_{\alpha'}\widehat{f}(\overline{\theta}^f) & \nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\nabla_{\beta'}\widehat{f}(\overline{\theta}^f) & \nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\nabla_{\gamma'}\widehat{f}(\overline{\theta}^f) \end{bmatrix}
$$

In this expression for$\sqrt{n}(\widehat{\theta}^f - \theta^f)$, examine the part for $\sqrt{n}(\widehat{\alpha}_f - \alpha_f)$, i.e., the first $k_\alpha \times 1$ components:

$$
\sqrt{n}(\widehat{\alpha}_f - \alpha_f) = \widehat{A}_f^{-1}\nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}^*_f\sqrt{n}\widehat{f}(\theta^f), \qquad \widehat{A}_f \equiv \nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}^*_f\nabla_{\alpha'}\widehat{f}(\overline{\theta}^f), \quad \hat{\Omega}^*_f \equiv \hat{\Omega}_f^{1/2}\hat{\Pi}_f\hat{\Omega}_f^{1/2}, \tag{C.13}
$$

$$
\begin{aligned}
\hat{\Pi}_f \equiv I_{\widetilde{k}_f} &- \hat{\Omega}_f^{1/2}\nabla_\beta \widehat{f}(\overline{\theta}^f)\{\nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\nabla_{\beta'}\widehat{f}(\overline{\theta}^f)\}^{-1}\nabla_\beta \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f^{1/2} \\
&- \hat{\Omega}_f^{1/2}\nabla_\gamma \widehat{f}(\overline{\theta}^f)\{\nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f\nabla_{\gamma'}\widehat{f}(\overline{\theta}^f)\}^{-1}\nabla_\gamma \widehat{f}(\widehat{\theta}^f)\hat{\Omega}_f^{1/2}.
\end{aligned}
$$

Then we have

$$
\sqrt{n}(\widehat{\alpha}_f - \alpha_f) = \frac{1}{\sqrt{n}}\sum_i \widehat{\eta}^f_i, \qquad \widehat{\eta}^f_i \equiv \widehat{A}_f^{-1}\nabla_\alpha \widehat{f}(\widehat{\theta}^f)\hat{\Omega}^*_f F(Z_i, \theta^f), \tag{C.14}
$$

and $\widehat{\eta}^f_i$ is the influence function of the first-stage estimate $\widehat{\alpha}_f$. If F is correct, θ^f is replaced by θ^f_0.

To find the influence functions $\widehat{\eta}^g_i$, let $\widehat{\theta}^g$ denote the first-stage estimator

$$
\widehat{\theta}^g \equiv (\widehat{\alpha}_g, \widehat{\beta}_g) = \arg \min_{\{\alpha,\beta\}\in\Theta_\alpha\times\Theta_\beta} \widetilde{Q}^g(\alpha, \beta) = \widehat{g}(\alpha, \beta)\hat{\Omega}_g\widehat{g}(\alpha, \beta).
$$

Under Assumption A7 and A10-12, with probability approaching one, the following first-order conditions in the first-stage for $\widehat{\theta}^g$ hold:

$$
FD^g_\alpha = \frac{\partial \widetilde{Q}^g(\widehat{\theta}^g)}{\partial \alpha} = \nabla_\alpha \widehat{g}(\widehat{\theta}^g)\hat{\Omega}_g\widehat{g}(\widehat{\theta}^g) = 0, \quad FD^g_\beta = \frac{\partial \widetilde{Q}^g(\widehat{\theta}^g)}{\partial \beta} = \nabla_\beta \widehat{g}(\widehat{\theta}^g)\hat{\Omega}_g\widehat{g}(\widehat{\theta}^g) = 0.
$$

Expend $\widehat{g}$ around the unique minimizer $\theta^g \equiv \{\alpha_g, \beta_g\}$ to get

$$\widehat{g}(\widehat{\theta^g}) = \widehat{g}(\theta^g) + \{\nabla_{\alpha'}\widehat{g}(\overline{\theta}^g)\}(\widehat{\alpha}_g - \alpha_g) + \{\nabla_{\beta'}\widehat{g}(\overline{\theta}^g)\}(\widehat{\beta} - \beta_g)$$

where $\overline{\theta}^g$ is the value for the mean value theorem. Substitute these into each FD^g to get

$$\begin{aligned}
FD^g_\alpha &= \nabla_\alpha \widehat{g}(\widehat{\theta^g})\hat{\Omega}_g[\widehat{g}(\theta^g) + \nabla_{\alpha'}\widehat{g}(\overline{\theta}^g)(\widehat{\alpha}_g - \alpha_g) + \nabla_{\beta'}\widehat{g}(\overline{\theta}^g)(\widehat{\beta}_g - \beta_g)] \\
FD^g_\beta &= \nabla_\beta \widehat{g}(\widehat{\theta^g})\hat{\Omega}_g[\widehat{g}(\theta^g) + \nabla_{\alpha'}\widehat{g}(\overline{\theta}^g)(\widehat{\alpha}_g - \alpha_g) + \nabla_{\beta'}\widehat{g}(\overline{\theta}^g)(\widehat{\beta}_g - \beta_g)] \\
FD^g &= \{FD^g_\alpha, FD^g_\beta\} = \widehat{I}^g + \widehat{H}^g(\widehat{\theta^g} - \theta^g), \quad \text{and from these, } \sqrt{n}(\widehat{\theta^g} - \theta^g) = \widehat{H}^{g-1}\sqrt{n}\widehat{I}^g, \\
\widehat{I}^g &\equiv \begin{bmatrix} \nabla_\alpha \widehat{g}(\widehat{\theta^g})\hat{\Omega}_g\widehat{g}(\theta^g) \\ \nabla_\beta \widehat{g}(\widehat{\theta^g})\hat{\Omega}_g\widehat{g}(\theta^g) \end{bmatrix}, \ \widehat{H}^g \equiv \begin{bmatrix} \nabla_\alpha \widehat{g}(\widehat{\theta^g})\hat{\Omega}_g\nabla_{\alpha'}\widehat{g}(\overline{\theta}^g) & \nabla_\alpha \widehat{g}(\widehat{\theta^g})\hat{\Omega}_g\nabla_{\beta'}\widehat{g}(\overline{\theta}^g) \\ \nabla_\beta \widehat{g}(\widehat{\theta^g})\hat{\Omega}_g\nabla_{\alpha'}\widehat{g}(\overline{\theta}^g) & \nabla_\beta \widehat{g}(\widehat{\theta^g})\hat{\Omega}_g\nabla_{\beta'}\widehat{g}(\overline{\theta}^g) \end{bmatrix}.
\end{aligned}$$

In this expression for$\sqrt{n}(\widehat{\theta^g} - \theta^g)$, examine the part for $\sqrt{n}(\widehat{\alpha}_g - \alpha_g)$, i.e., the first $k_\alpha \times 1$ components:

$$\sqrt{n}(\widehat{\alpha}_g - \alpha_g) = \widehat{A}_g^{-1}\nabla_\alpha \widehat{g}(\widehat{\theta^g})\hat{\Omega}^*_g\sqrt{n}\widehat{g}(\theta^g), \qquad \widehat{A}_g \equiv \nabla_\alpha \widehat{g}(\widehat{\theta^g})\hat{\Omega}^*_g\nabla_\alpha \widehat{g}(\overline{\theta}^g), \quad \hat{\Omega}^*_g \equiv \hat{\Omega}^{1/2}_g\hat{\Pi}_g\hat{\Omega}^{1/2}_g, \tag{C.15}$$

$$\hat{\Pi}_g \equiv I_{\tilde{k}_g} - \hat{\Omega}^{1/2}_g\nabla_{g'}\widehat{g}(\overline{\theta}^g)\{\nabla_\beta \widehat{g}(\widehat{\theta^g})\hat{\Omega}_g\nabla_{\beta'}\widehat{g}(\overline{\theta}^g)\}^{-1}\nabla_\beta \widehat{g}(\widehat{\theta^g})\hat{\Omega}^{1/2}_g.$$

Then, we have

$$\sqrt{n}(\widehat{\alpha}_g - \alpha_g) = \frac{1}{\sqrt{n}}\sum_i \widehat{\eta}^g_i, \qquad \widehat{\eta}^g_i \equiv \widehat{A}_g^{-1}\nabla_\alpha \widehat{g}(\widehat{\theta^g})\hat{\Omega}^*_g G(Z_i, \theta^g), \tag{C.16}$$

and $\widehat{\eta}^g_i$ is the influence function of the first-stage estimate $\widehat{\alpha}_g$. If G is correct, θ^g is replaced by θ^g_0.

Analogously, we can obtain the influence function $\widehat{\eta}^h_i$ switching the roles of β and γ, and switching the roles of g and h. If H is correct, θ^h is replaced by θ^h_0.

C.4 Appendix IV

Over-Identified Doubly Robust Identification and Estimation by Arthur Lewbel, Jin-Young Choi, and Zhuzhu Zhou

Original 2018, Revised February 2021

Supplemental Online Appendix

In this Supplemental Appendix, we provide two additional examples of applying our ODR estimator. For both examples, DR estimators already exist, so we can comparing the requirements of our ODR estimator to existing DR applications. The first example is average treatment effect estimation, while the second concerns additive regression models.

Average Treatment Effect Estimation

Going back to the earliest DR estimators like Robins, Rotnitzky, and van der Laan (2000), Scharfstein, Rotnitzky, and Robins (1999), and Robins, Rotnitzky, and Zhao (1994), here we describe the construction of DR estimates of average treatment effects, as in, e.g., Bang and Robins (2005), Funk, Westreich, Wiesen, Stürmer, Brookhart, and Davidian (2011), Rose and van der Laan (2014), Lunceford and Davidian (2004), Słoczyński and Wooldridge (2018) and Wooldridge (2007). We then show how this model could alternatively be estimated using our ODR construction. Note that other DR estimators of treatment effects also exist, e.g., Lee and Lee (2018).

The assumption in this application is that either the conditional mean of the outcome or the propensity score of treatment is correctly parametrically specified. Let $Z = \{Y, T, X\}$ where Y is an outcome, T is a binary treatment indicator, and X is a J vector of other covariates (including a constant). The average treatment effect we wish to estimate is

$$\alpha = E\{E(Y|T=1,X) - E(Y|T=0,X)\}. \tag{C.17}$$

As is well known, an alternative propensity score weighted expression for the same average treatment effect is

$$\alpha = E\left\{\frac{YT}{E(T|X)} - \frac{Y(1-T)}{1-E(T|X)}\right\}. \tag{C.18}$$

Let $\widetilde{G}(T, X, \beta)$ be the proposed functional form of the conditional mean of the outcome, for some K vector of parameters β. So if $\widetilde{G}$ is correctly specified, then $\widetilde{G}(T, X, \beta) = E(Y|T, X)$. Similarly, let $\widetilde{H}(X, \gamma)$ be the proposed functional form of the propensity score for some J vector of parameters γ, so if $\widetilde{H}$ is correctly specified, then $\widetilde{H}(X, \gamma) = E(T|X)$.

One standard estimator of α, based on equation (C.17), consists of first estimating β by least squares, minimizing the sample average of $E[\{Y - \widetilde{G}(T, X, \beta)\}^2]$, and then estimating α as the sample average of $\widetilde{G}(1, X, \beta) - \widetilde{G}(0, X, \beta)$. This estimator is equivalent to GMM estimation of α and β, using the vector of moments

$$E\left[\begin{array}{c} \{Y - \widetilde{G}(T, X, \beta)\} r_1(T, X) \\ \alpha - \{\widetilde{G}(1, X, \beta) - \widetilde{G}(0, X, \beta)\} \end{array}\right] = 0 \tag{C.19}$$

for some vector valued function $r_1(T, X)$. Least squares estimation of β specifically chooses $r_1(T, X)$ to equal $\partial \widetilde{G}(T, X, \beta)/\partial\beta$, but alternative functions could be used, corresponding to, e.g., weighted least squares estimation, or to the score functions associated with a maximum likelihood based estimator of β, given a parameterization for the error terms $Y - \widetilde{G}(T, X, \beta)$. Note that to identify the K vector β, the function $r_1(T, X)$ needs to be a $\widetilde{K}$ vector for some $\widetilde{K} \geq K$. The problem with this estimator is that in general α will not be consistently estimated if the functional form of $\widetilde{G}(T, X, \beta)$ is not the correct specification of $E(Y|T, X)$.

An alternative common estimator of α, based on equation (C.18), consists of first estimating γ by least squares, minimizing the sample average of $E[\{T - \widetilde{H}(X, \gamma)\}^2]$, and then estimating α as the sample average of $\frac{YT}{\widetilde{H}(X,\gamma)} - \frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)}$. This estimator is equivalent to GMM estimation of α and γ, using the vector of moments

$$E\left[\begin{array}{c} \{T - \widetilde{H}(X, \gamma)\} r_2(X) \\ \alpha - \left\{\frac{YT}{\widetilde{H}(X,\gamma)} - \frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)}\right\} \end{array}\right] = 0 \tag{C.20}$$

for some $\widetilde{J}$ vector valued function $r_2(X)$. As above, least squares estimation of γ sets $r_2(X)$ equal to $\partial \widetilde{H}(X, \gamma)/\partial\gamma$, but as above alternative functions could be chosen for $r_2(X)$. To identify the J vector γ, the function $r_2(X)$ needs to be a $\widetilde{J}$ vector for some $\widetilde{J} \geq J$. With this estimator, in general α will not be consistently estimated if the functional form of $\widetilde{H}(X, \gamma)$ is not the correct specification of $E(T|X)$.

A doubly robust estimator like that of Bang and Robins (2005) and other authors

assumes α can be expressed as

$$\alpha = E\left\{ \frac{YT}{\widetilde{H}(X,\gamma)} - \frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)} + \frac{T-\widetilde{H}(X,\gamma)}{\widetilde{H}(X,\gamma)}\widetilde{G}(1,X,\beta) - \frac{T-\widetilde{H}(X,\gamma)}{1-\widetilde{H}(X,\gamma)}\widetilde{G}(0,X,\beta) \right\}. \tag{C.21}$$

Observe that if $\widetilde{H}(X,\gamma) = E(T|X)$, then the first two terms in the above expectation equal equation (C.18) and the second two terms have mean zero. By rearranging terms, equation (C.21) can be rewritten as

$$\alpha = E\left[\widetilde{G}(1,X,\beta) - \widetilde{G}(0,X,\beta) + \frac{T}{\widetilde{H}(X,\gamma)}\{Y - \widetilde{G}(1,X,\beta)\} - \frac{1-T}{1-\widetilde{H}(X,\gamma)}\{Y - \widetilde{G}(0,X,\beta)\} \right]. \tag{C.22}$$

Rewriting the equation this way, it can be seen that if $\widetilde{G}(T,X,\beta) = E(Y|T,X)$, then the first two terms in equation (C.22) equal equation (C.17), and the second two terms have mean zero. This shows that equation (C.21) or equivalently (C.22) is doubly robust, in that it equals the average treatment effect α if either $\widetilde{G}(T,X,\beta)$ or $\widetilde{H}(X,\gamma)$ is correctly specified. The GMM estimator associated with this doubly robust estimator estimates α, β, and γ, using the moments

$$E\left[\begin{array}{c} \{Y - \widetilde{G}(T,X,\beta)\} r_1(T,X) \\ \{T - \widetilde{H}(X,\gamma)\} r_2(X) \\ \alpha - \left\{ \frac{YT}{\widetilde{H}(X,\gamma)} - \frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)} + \frac{T-\widetilde{H}(X,\gamma)}{\widetilde{H}(X,\gamma)}\widetilde{G}(1,X,\beta) - \frac{T-\widetilde{H}(X,\gamma)}{1-\widetilde{H}(X,\gamma)}\widetilde{G}(0,X,\beta) \right\} \end{array} \right] = 0. \tag{C.23}$$

Construction of this doubly robust estimator required finding equation (C.21) which is special to the problem at hand and possesses the DR property. In general, finding such expressions for any particular problem may be difficult or impossible.

In contrast, our proposed ODR estimator does not require any such creativity. All that is required for constructing our ODR for this problem is to know the two alternative standard estimators, based on equations (C.17) and (C.18), expressed in GMM form, i.e., equation (C.19) and equation (C.20). Just define $G(Z,\alpha,\beta)$ to be the vector of functions given in equation (C.19) and define $H(Z,\alpha,\gamma)$ to be the vector of functions given in equation (C.20). That is,

$$G(Z,\alpha,\beta) = \left[\begin{array}{c} \{Y - \widetilde{G}(T,X,\beta)\} r_1(T,X) \\ \alpha - \{\widetilde{G}(1,X,\beta) - \widetilde{G}(0,X,\beta)\} \end{array} \right] \tag{C.24}$$

and

$$H(Z,\alpha,\gamma) = \begin{bmatrix} \{T - \widetilde{H}(X,\gamma)\} r_2(X) \\ \alpha - \left\{ \frac{YT}{\widetilde{H}(X,\gamma)} - \frac{Y(1-T)}{1-\widetilde{H}(X,\gamma)} \right\} \end{bmatrix}. \tag{C.25}$$

These functions can then be plugged into the expressions in the previous section to obtain our ODR estimator, equation (4.1), without having to find an expression like equation (C.21) with its difficult to satisfy properties.

The vector $r_2(X)$ can include any functions of X as long as the corresponding moments $E\{H(Z,\alpha,\gamma)\}$ exist. To satisfy the required overidentification (discussed earlier, and formally given later in Assumption A3), we will want to choose $r_2(X)$ to include $\widetilde{J}$ elements where $\widetilde{J}$ is strictly greater than J. What we require is that, if the propensity score is incorrectly specified, then there is no α, γ (in the set of permitted values) that satisfies the moments $E\{H(Z,\alpha,\gamma)\} = 0$, while, if the propensity score is correctly specified, then the only α, γ that satisfies $E\{H(Z,\alpha,\gamma)\} = 0$ is α_0, γ_0. By the same logic, we will want to choose the $\widetilde{K}$ vector $r_1(T,X)$ to include strictly more than K elements. For efficiency, it could be sensible to let $r_2(X)$ and $r_1(T,X)$ include $\partial\widetilde{H}(X,\gamma)/\partial\gamma$ and $\partial\widetilde{G}(T,X,\beta)/\partial\beta$, respectively.

An Instrumental Variables Additive Regression Model

Okui, Small, Tan, and Robins (2012) propose a DR estimator for an instrumental variables (IV) additive regression model. The model is the additive regression

$$Y = M(W,\alpha) + \widetilde{G}(X) + U, \tag{C.26}$$

$$E(Q \mid X) = \widetilde{H}(X),$$

$$E(U \mid X, Q) = 0, \tag{C.27}$$

where Y is an observed outcome variable, W is a S vector of observed exogenous covariates, X is a J vector of observed confounders, and Q is a $K \geq S$ vector of observed instruments. Note that this model has features that are unusual for instrumental variables estimation, in particular, the assumption that $E(U \mid X, Q) = 0$ is stronger than the usual $E(U \mid Q) = 0$ assumption. The function $M(W,\alpha)$ is assumed to be correctly parameterized, and the goal is estimation of α.

Okui, Small, Tan, and Robins (2012) construct a DR estimator assuming that, in addition to the above, either $\widetilde{G}(X) = \widetilde{G}(X,\beta)$ is correctly parameterized, or that $\widetilde{H}(X) = \widetilde{H}(X,\gamma)$ is correctly parameterized. Let $Z = \{Y, W, X, Q\}$, and let $r_1(X)$ and $r_2(X)$ be vectors of functions chosen by the user. Define $G(\alpha,\beta,Z)$ and $H(\alpha,\gamma,Z)$

by

$$G(Z,\alpha,\beta) = \begin{bmatrix} \{Y - M(W,\alpha) - \widetilde{G}(X,\beta)\}r_1(X) \\ \{Y - M(W,\alpha) - \widetilde{G}(X,\beta)\}Q \end{bmatrix} \tag{C.28}$$

and

$$H(Z,\alpha,\gamma) = \begin{bmatrix} \{Q - \widetilde{H}(X,\gamma)\}r_2(X) \\ \{Y - M(W,\alpha)\}\{Q - \widetilde{H}(X,\gamma)\} \end{bmatrix}. \tag{C.29}$$

Okui, Small, Tan, and Robins (2012) take $r_1(X) = \partial\widetilde{G}(X,\beta)/\partial\beta$ and $r_2(X) = \partial\widetilde{H}(X,\gamma)/\partial\gamma$. If $\widetilde{G}(X,\beta)$ is correctly specified, then $E\{G(Z,\alpha,\beta)\} = 0$, while if $\widetilde{H}(X,\gamma)$ is correctly specified then $E\{H(Z,\alpha,\gamma)\} = 0$.

To get their doubly robust estimator, Okui, Small, Tan, and Robins (2012) first specify $\widetilde{G}(X_i,\beta)$ and $\widetilde{H}(X_i,\gamma)$, then estimate $\hat{\gamma}$ by the moment:

$$E(Q|X_i) = \widetilde{H}(X_i,\gamma)$$

and then estimate α and β by minimizing a quadratic form of $\hat{B}(\alpha,\beta;\hat{\gamma})$, where

$$\hat{B}(\alpha,\beta;\hat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} \{Y_i - M(W_i,\alpha) - \widetilde{G}(X_i,\beta)\}\{Q_i - \widetilde{H}(X_i,\hat{\gamma})\} \\ \{Y_i - M(W_i,\alpha) - \widetilde{G}(X_i,\beta)\}r_1(X_i) \end{bmatrix}.$$

In place of the Okui, Small, Tan, and Robins (2012) DR construction, we could estimate this model using the ODR estimator, equation (4.1), with G and H given by equations (C.28) and (C.29). To satisfy the required overidentification (Assumption A3), $r_1(X)$ and $r_2(X)$ need to include more than J elements. So, e.g., we would want to include at least one more function of X into $r_1(X)$ and $r_2(X)$, in addition to the functions $\partial\widetilde{G}(X,\beta)/\partial\beta$ and $\partial\widetilde{H}(X,\gamma)/\partial\gamma$ used by Okui, Small, Tan, and Robins (2012).