Department of Economics
Universidad Carlos III de Madrid

PhD Thesis

Essays on Discrete Choice Models with Fixed Effects

Mian Huang

Supervised by: Jesús M. Carro

November 17, 2015

Essays on Discrete Choice Model with Fixed Effects

Mian Huang

Abstract

Fixed effects in discrete choice models has been a challenge to econometricians from its existence. These unobservable heterogeneities are so important since their impacts can be seen clearly from the behavior of agents being studied. This has been consolidated by lots of studies and simulations including mine. However their existence prevent us from identifying models without restrictive assumptions about them. It is also hard to get rid of fixed effects since they enter the model not in a linear additive way and the outcomes are not continuous, therefore extant difference methods do not apply to discrete choice models with fixed effects. To have flexible specification on the fixed effects, it seems that partial identification is more practicable. There do exists some idea about set identification for discrete choice models and even some estimation methods were proposed for logistic-alike discrete choice models, whose key feature is that all model deduced conditional choice probabilities are well formulated in closed form expressions.

For reasons people may want to have discrete choice models with disturbance other than extreme type I distributed one to overcome some of its implications, e.g the property of independence with irrelevant alternatives among others. The challenge to meet such requirement is that the key feature of closed form expressions does not hold anymore, and techniques like simulation should be used. My PhD thesis provides the foundation and framework on how to practice the simulation based estimation for discrete choice models with rather flexible fixed effects. This framework is both theoretical and practical, I show how to construct the simulation based estimation and study conditions about both the property of model and practice of simulation under which the estimator is consistent. This object is achieved in two steps. I first develop the theory for static discrete choice models where outcomes of behavior does not depend on previous outcomes. In this case specification of disturbance could be rather free and even serial correlation could be included. Later on, I extend the framework to dynamic discrete choice models, where current behavior depends on some state variables which depend on previous behavior in turn. In dynamic models, specification for disturbance is still free except that serial correlation could not be allowed. These two steps consist of the first and second chapters, in both chapter a numeric example is given which shows how well the simulation based estimator works. In the last chapter

I turn to the real data and apply my method to the problem of career decision of young men. Essentially this is a typical application of dynamic programming discrete choice model, which means individual's object function is the lifetime utility and it depends on both previous behavior and future states and what individual should decide is not only the current behavior but also future actions. By introducing a reduced form of the future utility I succeed in fitting this problem into the framework of dynamic discrete choice model with fixed effects.

Acknowledgments

I am glad I could finally touch this part of my PhD thesis, there are many people I should mention. First of all, I want to thank my advisor Jesús M. Carro. It is him who showed me this important field with lots of challenges and give me proper guidance. In this area I learned skills of structural modeling and thinking of counterfactual behavior. I start to comprehend what he told it is difficult to model heterogeneity in microeconometric models. What is more, his patience with me saved me a lot from panic. My thanks also go to Carlos Velasco. He is so informative in statistics and probability. Whenever I went to his office with my questions he could always show me the essential of mathematics and let me know where to go and find the solution. Laura Hospido, one of my previous thesis committee member, who read my first chapter and gave me many precious feedbacks. From her I got the idea of the application to young men's career decision which constitutes my last chapter. I also thank Ricardo Mora, who encouraged me a lot and shared with me his own academic experience. Last but not least, I want to thank all of my family. My wife Jing Yang spent three years far away from China with me and thanks a lot for your accompany. My little daughter Yousi and little son Youping, I am so sorry for not being with you during your very early period. But I can alway feel your love, this time thanks to the Mum again. It is all because of your understanding and unconditional support, this wonderful thing could finally happen. I love you all.

Contents

Chapter 1

Simulation Based Estimation of Multinomial Discrete Choice Model with Fixed Effects

Abstract

Multinomial discrete choice model, including binary choice model, is a class of widely used nonlinear models. Including unobservable heterogeneity in such models is necessary in many applications due to the unobserved variations in individual's preference, attributes or technology in use. Unfortunately this will cause problem for their point identification. I study the estimation of a multinomial discrete choice model in panel data with potentially multidimensional fixed effects that can be set identified on its parameters and conditional average partial effects for the outcome and choice probability. The model I study in this paper is general in the sense that it allows components without closed form, and its conditional probabilities for each alternatives and partial effects can be gotten through Monte Carlo simulation. For this model I propose a simulation based estimator for all the set identified quantities and I show that this estimator is consistent under general conditions and a perturbed bootstrap method can be used to implement its inference. A numeric example with simulated data is given to show the behavior of the estimator and I find that the estimated bounds of partial effects contain their true effects.

1.1 Introduction

Discrete choice model is a powerful tool for econometricians to quantify individual's choice behavior, where individual can be a single consumer, a household, a company or even an organization or government. It can be generally any entity that needs to make a choice out of its available options. Technically, discrete choice models are generally non-separable in the sense that unobserved variations or error terms in its utility expression do not enter the choice function in an additive way as a simple linear regression do. Specifically, since the seminal paper on qualitative choice behavior by D. McFadden 1973, the random utility model (RUM) has become the cornerstone of discrete choice models. Following this practice, even if it is customary to set a linear utility for individuals under investigation, their discrete choice behavior is given by indicator functions instead of linear functions. The nonlinearity of indicator functions makes discrete choice model intrinsically non-additive separable in its error term or other unobserved variations of the utility. In this case if any part of the error term is correlated with one of the other explanatory variables in the utilities, a leading example is the fixed effect which I am working with in this paper, it will be hard to get rid of the endogeneity. This property makes the identification and estimation of discrete choice models with fixed effects a big challenge.

Maximum likelihood method can not give a consistent estimation if we just ignore the fixed effects and treat all the explanatory variables as exogenous. Alternatively, if we take the fixed effects as individual specific parameters to be estimated together with other parameters, Neyman and Scott 1948 showed that the maximum likelihood estimator does not consistently estimate the true parameters because the number of "nuisance parameters[1]" grows with the sample size and this fact violates one of the conditions for the consistency of MLE.

A further alternative method instead of fixed effects is the random effects approach. This method asks you to specify the conditional distribution of the unobservable individual specific attributes given other endogenous explanatory variables and the most prevalent practice is to assume the independence. But for dynamic models[2] random effects approach encounters the initial conditions problem pointed out by Heckman 1981 if your observations can not cover the process from its inception. Even if you use the likelihood conditional on the initial observed dependent variables, random effects approach may still be problematic in internal

[1] Many literatures are devoted to the nuisance or incidental parameter problem, Lancaster 2000 gives a review on this topic.

[2] In the context of discrete choice model with unobservable individual effects, dynamic models indicate models with lagged dependent variables as their explanatory variables. We do not consider those forward looking dynamic discrete choice models in this paper.

consistency across different numbers of periods[3].

Still another well known solution is the BLP method[4], it is named after a series of papers by S. T. Berry 1994, S. Berry, Levinsohn, and Pakes 1995. BLP method decomposes the unobserved utility variation into two parts, of which one is the fixed effects and the other part is independent of all other parts of the model. To avoid the nuisance parameter problem, they do not allow the fixed effects to vary over individuals but only over different choices and markets. The numbers of choices and markets are fixed while the number of individuals increases, therefore the number of fixed effects doesn't increase with the sample size such that they can treat the fixed effects as parameters to be estimated. BLP method actually avoids the nuisance parameter problem by assuming some homogeneity across individuals and limiting the number of heterogeneities.

Browning and Carro 2013 and Browning and Carro 2014 studied the number of heterogeneous types that can be identified in a first order dynamic binary outcome model with maximal heterogeneity. Since the complete heterogeneity causes the nuisance parameter problem and prevents the identification, a restricted maximal number of heterogeneous types can help to point identify the model. They found that as the number of panels increases, the identified number of heterogeneous types also increases.

All the methods aforementioned have to impose some restrictions on the conditional distribution of individual specific heterogeneities in terms of either the functional form or a maximal types of heterogeneity. For BLP method this conditional distribution even degenerates to market or choice specific nonrandom parameters. If we would like to have more flexibility in its conditional distribution, Honoré and Tamer 2006 showed that even for simple models point identification often fails. They also provided an idea of set identification as well as three principles to characterize the identified set. In this paper I study a ready to extend static model[5], where their result of a lack of point identification still apply, and their idea of set identification has been developed by Chernozhukov et al. 2013a in a similar framework.

In this paper I further develop the method by Chernozhukov et al. 2013a. I notice that their method of set identification is general for multinomial discrete choice models, but their estimation can be only applied to situations where all the model components have closed forms. Note that there are situations where some model components have no closed forms, i.e. the choice probability and partial effects in some multinomial discrete choice models with more than two

[3] See Wooldridge 2005 and Honoré 2002 for a discussion.

[4] Nevo 2000 gives a precise but still pellucid explanation of the BLP model.

[5] This idea can be applied to dynamic models, but to keep the notation and assumption consistent and simple, I consider only static models in this paper.

alternatives and even binary choice models allowing serial correlation in error terms. The main object of this paper is to complement this method by developing solution to deal with models with non-closed-form components and to generalize the analysis in average treatment effect for binary choice models to the average partial effect on the choice probability for multinomial choice models.

To be specific, I first introduce models that fit Chernozhukov et al. 2013a and henceforth CFHN with closed-form components. For example, the GEV (generalized extreme value[6]) family discrete choice models with fixed effects. We can derive closed-form choice probabilities from them because of their mathematic tractability. This feature makes them immediate expansions of the binary logit model from CFHN (2013). Then I will discuss specifications other than GEV models which usually do not have closed-form choice probability.

Now let us start our journey with a simple mixed logit model for only cross-sectional data. Consider a general decision problem, the utility for a specific alternative a out of a choice set $\mathcal{F}$ is specified as

$$U_{ia} = \beta^{\in}X_{ia} + \alpha_i Z_{ia} + \epsilon_{ia}, \tag{1.1.1}$$

where X_{ia} and Z_{ia} are vectors of observed variables relating to alternative a and individual i, and β is a vector of coefficients which is constant over i which measures the homogeneous marginal utility with respect to X while α_i is the random marginal utility with respect to Z that is iid over individuals. The preference heterogeneity make the product term $\alpha_i Z_{ia}$ an error component. The mixed logit method assumes that ϵ_{ia} is iid extreme value over both a and i. Given α_i, the probability for individual i to take alternative a is

$$P^a_{i\|\alpha} = \frac{exp(\beta^{\in}X_{ia} + \alpha_i Z_{ia})}{\sum_{b/\mathcal{F}} exp(\beta^{\in}X_{ib} + \alpha_i Z_{ib})}. \tag{1.1.2}$$

The heterogeneous preference parameter α_i is assumed to be independent of explanatory variables X and Z and have a distribution $f(\alpha\|\theta)$. Then the unconditional choice probability is the integration of (1.1.2) over α with respect to the distribution $f(\alpha\|\theta)$ as follows

$$P^a_i = \Big[\; \frac{exp(\beta^{\in}X_{ia} + \alpha_i Z_{ia})}{\sum_{b/\mathcal{F}} exp(\beta^{\in}X_{ib} + \alpha_i Z_{ib})} f(\alpha\|\theta) d\alpha. \tag{1.1.3}$$

It can be seen clearly that in (1.1.3), the mixed logit model actually follows the

[6] The unifying attribute of these models is that the unobserved portions of utility for all alternatives are jointly distributed as a generalized extreme value such that the assumption of independence from irrelevant alternatives (IIA) can be relaxed. See Chapter 4 in Kenneth E. Train 2009 for further information.

random effect method since α_i is supposed to be independent of X and Z. In case there is no information for this independence, CFNH's method and my extension developed later are more useful, where α_i can be freely correlated with them. Actually I allow nonparametric dependence between α_i and explanatory variables X and Z. To complement this nonparametric specification I assume that all the explanatory variables take discrete values or can be convincingly recoded into category variables without seriously loss of information[7]. With this complement the number of α_i's conditional distributions becomes finite and I can consider them one by one. Furthermore, CFNH (2013) has proven that it is only those conditional distributions of α_i with finite support that are relevant if we are concerning the choice probability and average treatment effect[8]. Therefore the arbitrary uncertainty in α_i's conditional distributions can be handled in a finite dimension space $\mathcal{X}_M^K$, where M is the cardinality of α_i's finite support and K is the number of values taken by discrete explanatory variables, and $\mathcal{X}_M^K$ is the Cartesian product of K unit simplexes of dimension M, known as $\mathcal{X}_M$, for all conditional values. The discrete explanatory variable assumption together with the finite dimension property of multinomial choice model help us to reduce the dimension of the original problem from infinity to a finite number.

As in the mixed logit model, we can integrate out the fixed effects α_i using conditional distributions drawn from $\mathcal{X}_M^K$ and get the unconditional choice probability for individuals from each subgroup[9], wherein individuals have the same observed attributes. The idea of set identification is to match the choice probabilities derived from the model with the "real" probabilities. It turns out that only a limited combinations of the parameters and α's conditional distributions make the match holds. This idea is general enough to be applied to any multinomial choice models with fixed effects, however CFHN only showed how to estimate the identified sets for binary logit and probit models, which have closed forms for every components.

For models with non-closed form components, I propose the method of simulation. This also helps us to consider more realistic models than what is mathematical convenient. I show that a simple Accept-Reject (AR) simulator gives a consistent estimation under some general assumptions.

[7] Of course whether this assumption is feasible depends on your purpose of research.

[8] See lemma 7 in Chernozhukov et al. 2013a. They stated this property as "A useful feature of multinomial panel models is that they are finite dimensional, in spite of the presence of distributions." Thanks to the clarification of Jesús M. Carro, this idea can be actually dated back to Ghosal and Van Der Vaart 2001.

[9] After integrating out α_i actually we get a choice probability conditional on observed values of the explanatory variables. Here we say it is unconditional is to emphasize that the probability is not conditional on the unobservable fixed effects α_i. In other words the probability is only conditional on observable explanatory variables.

The rest of the paper is organized as follows. I derive the semiparametric multinomial discrete choice model with fixed effects in panel data under the framework of random utility model in section 1.2. Then I present how to set identify this model by restrictions in choice probability and study how to estimate models with non-closed form components by simulation in section 1.3. In section 1.4, I show that the simulation based estimators of average treatment effect on the outcome variable and treatment effect on choice probability as well as the identified set of parameters are consistent. I also refer to the perturbed bootstrap for valid inference. In section 1.5 a numeric example is given to show the behavior of the simulation based estimator. The last section concludes the paper.

1.2 The Semiparametric Multinomial Discrete Choice Model with Fixed Effects

In this section I showcase the details of the model I am going to study. Briefly individual's choice behavior is modeled by an augmented random utility model with fixed effects. That is to say an utility is assumed for each individual and this utility contains some unobserved variations including fixed effects. The unobserved variations and other observed attributions of individuals in terms of various alternatives reflect individuals' varied features and their idiosyncratic preferences over different alternatives. All other components of the unobserved utility except for the fixed effects are encapsulated into the error term ϵ_{ita}, and it is assumed that ϵ_{ita} is independent from the rest of the utility, including both the fixed effects and other observed variables. What is more, different assumptions about the error term's marginal distribution can be used for different purposes, and the fixed effects are freely correlated with the observed attributions in the utility since there are no specifications about how they are dependent with each other and their relationship can only be revealed by the field data. The fixed effects assumption composes the nonparametric aspect of the model.

The rest of the model is parametrically specified as follows. Individual's decision behavior is studied over a specific period and the number of periods is denoted as T. Individuals are assumed to choose their favorite choice out of a finite set of options $\mathcal{F}$ in each period t, where $\dim(\mathcal{F}) = A$. In each period t, individual i can have a utility (or any other objective functions that serve the purpose to be maximized or minimized) U_{ita} from choosing alternative $a \ / \mathcal{F}$. For decision makers, they know their utilities for all the alternatives such that they can opt for the best alternative a, i.e. $U_{ita} \geqslant U_{ita^{\in}}$for all $a^{\in} / \mathcal{F}$ and $a^{\in} \forall\!\!\!\neq a$.

Of course, decision maker's utility is hardly observed by researchers, instead they observe some attributes of the alternatives and the decision makers. Label these attributes in period t as x_{ita} for any alternative a and individual i, then a utility function V_{ita} that relates these observed attributes to the decision maker's utility can be specified as $V_{ita} = V(x_{ita})$, and this is called the *representative utility* by Kenneth E. Train 2009 or *strict utility* by D. McFadden 1977. This definition for representative utility doesn't consider the fixed effects, α_i, and it can be generalized as a function that also includes the fixed effects, i.e. $V_{ita} = V(x_{ita}, \alpha_{ia})$. In this paper I focus on parametric specifications on V_{ita}, and it generally has the form $V_{ita} = V(x_{ita}, \alpha_{ia}, \beta)$. There are some comments on its specification before I continue.

First of all, should it be possible allowing observable attributes of alternatives entering the representative utility is useful. It turns out an efficient way to include alternatives in the utility function. Without x_{ita} being varied over alternatives a, an alternative method that allows a flexible substitution pattern between alternatives is to use dummy variables for each alternative and let part of the parameters β to be alternative specific. One drawback of this method is that it introduces a lot of alternative specific parameters into the model if there are a host of alternatives for each decision maker[10]. D. McFadden 1973 solved this problem by projecting alternatives onto their characteristics. In other words, we can use a finite common vector of characteristics to distinguish different alternatives and it is possible to represent alternative specific parameters as functions of common parameters and alternative's attributes such that the number of β can be suppressed as the number of alternatives increases. This method has been followed by many papers that analyze demands for differentiated products[11] and my paper is not an exception.

Secondly, letting x_{ita} to change over i is a similar way to capture idiosyncratic preference of individuals without letting β to be specific to i. A borrowed example of car market from Kenneth E. Train 2009 is helpful to clarify this idea. Suppose that the only two attributes of cars that can be observed by the researchers are the purchase price, PR_{ta}, for car a in period t and inches of shoulder room, SR_a, which is a measure of the interior size of the car a. The value that households place on these two attributes varies over households, therefore the representative utility is

$$V_{ita} = \beta_{it1} SR_a + \beta_{it2} PR_{ta}, \tag{1.2.1}$$

where β_{it} are parameters specific to household i and period t. This variation in taste can be modeled in the following way. Suppose the value on shoulder room

[10] See Nevo 2000 for more discussions.

[11] Some examples include: Bresnahan 1987, S. T. Berry 1994, S. Berry, Levinsohn, and Pakes 1995, S. Berry, Levinsohn, and Pakes 2004 and S. T. Berry and Haile 2010.

varies only with the number of members in the household, M_{it}, as

$$\beta_{it1} = \rho M_{it},$$

where ρ is positive such that the value on shoulder room, β_{it1}, increase as M_{it} increases. Similarly, suppose the importance of purchase price is inversely related to income, I_{it}, so that low-income households place more importance on purchase price as

$$\beta_{it2} = \theta / I_{it}.$$

Substituting these relations into (1.2.1) produces

$$V_{ita} = \rho(M_{it}SR_a) + \theta(PR_{ta}/I_{it}),$$

where the product $M_{it}SR_a$ and the quotient PR_{ta}/I_{it} vary over i, t and a and this is exactly covered by the model of general specification.

Furthermore, fixed effects can be included naturally. Suppose that the value of shoulder room varies with household size plus some other factors[12] that are unobserved by the researcher. For example

$$\beta_{it1} = \rho M_{it} + \mu_i,$$

where μ_i is random but constant over t. Similarly, the importance of purchase price consists of its observed and unobserved components as

$$\beta_{it2} = \theta / I_{it} + \eta_i.$$

Substituting into (1.2.1) produces

$$V_{ita} = \rho(M_{it}SR_a) + \theta(PR_{ta}/I_{it}) + \mu_i SR_a + \eta_i PR_{ta},$$

where the term $\mu_i SR_a + \eta_i PR_{ta}$ is unobservable because μ_i and η_i are unobservable. This example shows how the fixed effects come out and what is more, if you would like to keep the structure of $\mu_i SR_a + \eta_i PR_{ta}$, this is also a model with fixed effects of dimension two, where $\alpha_{ia} = (\mu_i SR_a, \eta_i)$ and $V(x_{ita}, \alpha_{ia}, \beta) = \rho(M_{it}SR_a) + \theta(PR_{ta}/I_{it}) + \mu_i SR_a + \eta_i PR_{ta}$.

Last but not least, x_{ita} can contain element that only varies with t and keeps constant over i and a, therefore the model has the capacity of dealing with time effect. Furthermore this model can even deal with dynamic choice over periods, e.g. you can let one element of x_{ita} to be the dummy variable $\mathbb{1}(y_{i(t\quad 1)} = a)$, where

[12] e.g., sizes of the family members, or frequency with which the household travel together.

$y_{i(t\quad 1)}$ is the choice from last period. Parameter for this dummy captures the type of consumption inertia or variety seeking. Since the dynamic choice model needs a different assumption about ϵ_{ita}[13] and causes confusions in the notation, I study dynamic cases in a separate project and focus on static models right now.

After giving comments about specifications on the representative utility, let's continue with other parts of the utility. Since there are still some aspects of the utilities that can not be observed by researchers, $U_{ita} \not\forall V_{ita}$. Let's define $\epsilon_{ita} = U_{ita} \quad V_{ita}$ thus a general utility can be decomposed as

$$U_{ita} = V(x_{ita}, \alpha_{ia}, \beta) + \epsilon_{ita}. \tag{1.2.2}$$

Random effects method assumes that x_{ita} and α_i are independent. The car market example has shown this assumption is too restrictive, since μ_i may be correlated with I_{it} or η_i may be correlated with M_{it}. Therefore in this paper fixed effects method is used such that more flexibility in the dependence between the time invariable heterogeneity, α_{ia}, and other observable covariates, x_{ita} is allowed. However, ϵ_{ita} is assumed to be independent of V_{ita}. For fixed effects in the augmented random utility model, even if an additive functional form is utilized in the utility, i.e. $V(x_{ita}, \alpha_{ia}, \beta) = x_{ita}{}^{\Subset}\beta + \alpha_{ia}$[14], the observed choices in each period are not functions that are additive in the fixed effects. This is because individuals are assumed to behave according to the following rule

$$y_{it} = \arg\max_{a/\mathcal{F}} \} V(x_{ita}, \alpha_{ia}, \beta) + \epsilon_{ita} | \,. \tag{1.2.3}$$

The function above is not additive in the fixed effects and the fixed effects can not be canceled out by simply taking the difference over time. This is the reason as such some authors call this model *nonseparable*.

To simplify notations, let's denote the behavior function (1.2.3) as

$$y_{it} = g_0(x_{it}, \alpha_i, \epsilon_{it}, \beta), \ (i = 1, \times\!\times\!\times, n; t = 1, \times\!\times\!\times, T), \tag{1.2.4}$$

where $x_{it} = \}x_{ita}|_{\,a/\mathcal{F}}$, $\alpha_i = \}\alpha_{ia}|_{\,a/\mathcal{F}}$, and $\epsilon_{it} = \}\epsilon_{ita}|_{\,a/\mathcal{F}}$. For simplicity the only assumption that is going to be imposed on the fixed effects, α_i, is the compact support assumption. Denote the support of α_i as $\mathrm{Y} \to \mathbb{R}^A$ and Y is assumed a compact subset of $\mathbb{R}^A$, where A is the cardinality of the choice set $\mathcal{F}$.

Behavior function (1.2.4) gives individuals' choice at each period t. To use all the information in the panel and to improve the efficiency, individuals' choices over all the T periods should be considered simultaneously. The idea is to treat

[13] We are discussing its definition in the next paragraph.

[14] Note that the results do not depend on this linear additive separable specification.

every complete observation of choices for all the periods as a single alternative. Let's consider the choice set over all the T periods, $\mathcal{F}^T$, and $J = A^T$ is the cardinality of $\mathcal{F}^T$[15]. For example, $\mathcal{Z}^j$ is to denote a general alternative in $\mathcal{F}^T$ with $\mathcal{Z}^j = \}a_1, \ldots, a_T|$, where $a_t \ / \ \mathcal{F}$ for any t. Let's label decision maker's history of complete choices during all the T periods as $Y_i = \}y_{it}|_{\ t/\}1,\ldots,T|}$, then $Y_i = \mathcal{Z}^j$ if and only if $y_{it} = a_t$ for all t.

Remember that the attributes x_{ita} is assumed to be discrete, and the complete history of attributes is denoted as $X_i = \}x_{it}|_{\ t/\}1,\ggg,T|}$, where $x_{it} = \}x_{ita}|_{\ a/\mathcal{F}}$. Suppose x_{ita} has p attributes, thus X_i is actually a discrete vector of dimension $p * A * T$. Denote X_i's support as $\mathcal{Y}$, and K is the cardinality of $\mathcal{Y}$. Thus $\mathcal{Y}^k$ is to indicate a general element in $\mathcal{Y}$, and it could be used to distinguish all the individuals from the perspective of researchers. If ϵ_i follows a specific distribution which is independent from X_i and α_i, where $\epsilon_i = \}\epsilon_{it}|_{\ t/\}1,\ldots,T|}$ and $\epsilon_{it} = \}\epsilon_{ita}|_{\ a/\mathcal{F}}$, the probability of choosing $\mathcal{Z}^j$ out of $\mathcal{F}^T$ for a decision maker with the observable attributes $\mathcal{Y}^k$ and unobservable fixed effects α can be deduced. For example, if

$$\epsilon_i \subset H(\epsilon), \tag{1.2.5}$$

where $\subset$ stands for *distributed as*. Denote the choice probability aforementioned as $\mathcal{P}_j^k(\alpha, \beta)$, where β is the parameter from the representative utility, then $\mathcal{P}_j^k(\alpha, \beta)$ is given by

$$\begin{aligned}
\mathcal{P}_j^k(\alpha, \beta) &= P_r(Y_i = \mathcal{Z}^j \| X_i = \mathcal{Y}^k, \alpha_i = \alpha) \\
&= P_r \Big) g_0(\mathcal{Y}_t^k, \alpha, \epsilon_{it}, \beta) = a_t, \exists t \Big\| X_i = \mathcal{Y}^k, \alpha_i = \alpha \Big(\\
&= \Big[\ \mathbb{1} \Big) g_0(\mathcal{Y}_t^k, \alpha, \epsilon_{it}, \beta) = a_t, \exists t \Big(dH(\epsilon) \\
&= \Big[\ \prod_{t=1}^{T} \mathbb{1} \Big) g_0(\mathcal{Y}_t^k, \alpha, \epsilon_{it}, \beta) = a_t \Big(dH(\epsilon).
\end{aligned} \tag{1.2.6}$$

The third equation holds because we assume ϵ_i are independent of X_i and α_i. Note that the occurrence of the event

$$g_0(\mathcal{Y}_t^k, \alpha, \epsilon_{it}, \beta) = a_t$$

is equivalent to the occurrences of all the following events simultaneously

$$\Big\} V(\mathcal{Y}_{ta_t}^k, \alpha_{a_t}, \beta) + \epsilon_{ita_t} \geqslant V(\mathcal{Y}_{ta_t^{\in}}^k, \alpha_{a_t^{\in}}, \beta) + \epsilon_{ita_t^{\in}} \ {}_{\exists a_t^{\in}/\mathcal{F} \text{ and } a_t^{\in} \not\forall a_t} .$$

[15] In applications, J is actually the number of realized history of choices in the observation. It could be strictly less than A^T for various reasons.

Therefore the indicator function in the last equation of (1.2.6) can be expressed as

$$\prod_{a_t^{\in}\neq a_t} \mathbb{1}\Big) V(\mathcal{Y}^k_{ta_t}, \alpha_{a_t}, \beta) + \epsilon_{ita_t} \geqslant V(\mathcal{Y}^k_{ta_t^{\in}}, \alpha_{a_t^{\in}}, \beta) + \epsilon_{ita_t^{\in}}\Big(. \tag{1.2.7}$$

Substituting (1.2.7) into (1.2.6) produces

(1.2.8)

$$\mathcal{P}^k_j(\alpha,\beta) = \int \prod_{t=1}^{T} \prod_{a_t^{\in}\neq a_t} \mathbb{1}\Big) V(\mathcal{Y}^k_{ta_t}, \alpha_{a_t}, \beta) + \epsilon_{ita_t} \geqslant V(\mathcal{Y}^k_{ta_t^{\in}}, \alpha_{a_t^{\in}}, \beta) + \epsilon_{ita_t^{\in}}\Big(dH(\epsilon).$$

As you can see later in section 1.3, $\mathcal{P}^k_j(\alpha,\beta)$ serves the purpose of identification. In practice, people may not be so interested in the probability of choosing a history of choices per se. For example, to know the potential probability of choosing one specific alternative a_t in period t may be more interesting. For this purpose, let's define the corresponding choice probability as

$$\mathcal{P}^{x_t}_{a_t}(\alpha,\beta) = P_r(y_{it} = a_t \| (x_{it} = x_t), X_i = \mathcal{Y}^k, \alpha_i = \alpha), \tag{1.2.9}$$

where $x_{it} = x_t$ inside the parenthesis of the conditional part is to emphasis that this probability is a potential probability for individuals with $X_i = \mathcal{Y}^k$ and $\alpha_i = \alpha$ behaving as the current attributes is x_t instead of $\mathcal{Y}^k_t$. This is useful for the following counterfactual analysis and it could be easily rewritten as follows

(1.2.10)

$$\begin{aligned}
\mathcal{P}^{x_t}_{a_t}(\alpha,\beta) &= P_r \Big) g_0(x_t, \alpha, \epsilon_{it}, \beta) = a_t \Big\| X_i = \mathcal{Y}^k, \alpha_i = \alpha \Big(\\
&= \Big[\ \mathbb{1}\left(g_0(x_t, \alpha, \epsilon_{it}, \beta) = a_t\right) dH(\epsilon) \\
&= \Big[\prod_{a_t^{\in}\neq a_t} \mathbb{1}\Big) V(x_{ta_t}, \alpha_{a_t}, \beta) + \epsilon_{ita_t} \geqslant V(x_{ta_t^{\in}}, \alpha_{a_t^{\in}}, \beta) + \epsilon_{ita_t^{\in}}\Big(dH(\epsilon).
\end{aligned}$$

Note that this potential probability doesn't depend on the conditional information but depends on the counterfactual value x_t since ϵ_i is assumed to be independent of X_i and α_i.

Now it is time to discuss the specifications about $H(\epsilon)$. As mentioned in the introduction, let's first consider the GEV model where $H(\epsilon)$ is the extreme value type I distribution. This model is a direct extension to the binary logit model and it has closed form expressions for (1.2.8) and (1.2.10).

1.2.1 Generalized Extreme Value Model

Although specifications on $H(\epsilon)$ should reveal as more the economic realities as possible, the freedom of specifying $H(\epsilon)$ as you wish doesn't come without costs. The major problem is that for many specifications of $H(\epsilon)$, the integrations in equation (1.2.8) and (1.2.10) do not have closed-form expressions and a numerical calculation by simulation is needed. This was a challenge thirty years ago when the computation was extremely costly. Econometricians at that time found some models where components like the choice probabilities have closed-form expressions, and out of those models which had been developed in the early 1980's the *logit* model and the *nested logit* model are prominent workhorses. Both models assume a joint extreme value distribution for ϵ_i, hence they are members of the *generalized extreme value models*. Here I first introduce these two models as immediate extensions of the binary logit model proposed by CFHN (2013) without extra efforts of simulation.

The simplest assumption about ϵ_i's joint distribution is made by the logit model, which assumes each term in ϵ_i are iid over both t and a and follows the type I extreme value distribution. That is to assume ϵ_{ita}'s cumulative distribution is

$$F(\epsilon_{ita}) = e^{\ e^{\ \epsilon_{ita}}} \text{ for all } i, t \text{ and } a. \tag{1.2.11}$$

This is a strong assumption under which researchers believe that all the correlations over alternatives and periods have been captured by the representative utility. If this assumption holds, equation (1.2.8) has the following simple expression

$$\mathcal{P}_j^k(\alpha, \beta) = \prod_{t=1}^{T} \frac{e^{V_{ita_t}}}{\sum\limits_{b_t / \mathcal{F}} e^{V_{itb_t}}} = \prod_{t=1}^{T} \frac{e^{V(\mathcal{Y}_{ta_t}^k, \alpha_{a_t}, \beta)}}{\sum\limits_{b_t / \mathcal{F}} e^{V(\mathcal{Y}_{tb_t}^k, \alpha_{b_t}, \beta)}}, \tag{1.2.12}$$

which is just the product of these probabilities of choosing a_t, the alternative of period t in the choice history $\mathcal{Z}^j$, in each time period t. Equation (1.2.10) also has a closed-form expression as follows

$$\mathcal{P}_{a_t}^{x_t}(\alpha, \beta) = \frac{e^{V(x_{ta_t}, \alpha_{a_t}, \beta)}}{\sum\limits_{b_t / \mathcal{F}} e^{V(x_{tb_t}, \alpha_{b_t}, \beta)}}. \tag{1.2.13}$$

This is a immediate extension of the binary logit model from CFHN (2013) by considering more than two choices.

Another workhorse of discrete choice analysis is the nested logit model. It further generalizes the multinomial logit model aforementioned by allowing correlations in between errors corresponding to alternatives with similarity. For example, alternatives can be organized into different groups according to the similarity in their observed attributes. It is believed that alternatives in the same group enjoy some similar unobservable attributes due to their similarity in the observed attributes, thus there should be correlation or dependence in between errors in the same group.

To delineate this idea, let's divide the choice set $\mathcal{F}$ into exclusive subsets $\}\mathcal{F}_n|\,_{n=1}^{N_g}$ according to the similarity of alternatives, where N_g is the number of subsets, and $\mathcal{F} = \cap_{n=1}^{N_g}\mathcal{F}_n$.[16] There exists correlation for errors, $\}\epsilon_{ita}|\,_{a\,/\,\mathcal{F}_n}$, in each group $\mathcal{F}_n$, and σ_n is used to measure the correlation for each group. Errors from different groups are independent and they are also independent over periods. That is to say ϵ_{it_1a} and ϵ_{it_2b} are correlated if and only if $t_1 = t_2$ and alternatives a and b are from the same subset $\mathcal{F}_n$ for some n, otherwise they are independent of each other. For simplicity it is assumed that there is no change in the partitioning $\}\mathcal{F}_n|\,_{n=1}^{N_g}$ over periods and $\}\sigma_n|\,_{n=1}^{N_g}$ are also constant over time. According to D. McFadden 1977's interpretation, $\sigma_n\ /\ (0,1)$ and a larger σ_n indicate a greater similarity or dependence in between $\}\epsilon_{ita}|\,_{a\,/\,\mathcal{F}_n}$. As a result, the conditional choice probability of choosing $a_t\ /\ \mathcal{F}_n$ for individual i in each period t is given by

$$\mathcal{P}_{a_t}^{\mathcal{Y}_t^k}(\alpha,\beta) = \frac{e^{V_{ita_t}/(1\quad\sigma_n)}\Big)\sum_{a\,/\,\mathcal{F}_n} e^{V_{ita}/(1\quad\sigma_n)}\Big(^{\sigma_n}}{\sum_{n=1}^{N_g}\quad\sum_{a\,/\,\mathcal{F}_n} e^{V_{ita}/(1\quad\sigma_n)}\Big[^{1\quad\sigma_n}} = \frac{e^{V(\mathcal{Y}_{ta_t}^k,\alpha_{a_t},\beta)/(1\quad\sigma_n)}\Big)\sum_{a\,/\,\mathcal{F}_n} e^{V(\mathcal{Y}_{ta}^k,\alpha_a,\beta)/(1\quad\sigma_n)}\Big(^{\sigma_n}}{\sum_{n=1}^{N_g}\Big)\sum_{a\,/\,\mathcal{F}_n} e^{V(\mathcal{Y}_{ta}^k,\alpha_a,\beta)/(1\quad\sigma_n)}\Big(^{1\quad\sigma_n}}, \tag{1.2.14}$$

therefore the conditional choice probability of choosing $\mathcal{Z}^j$ for individual i is the product of the formula above over t, and the equation (1.2.8) has the following expression

$$\mathcal{P}_j^k(\alpha,\beta) = \prod_{t=1}^{T}\frac{e^{V_{ita_t}/(1\quad\sigma_n)}\Big)\sum_{a\,/\,\mathcal{F}_n} e^{V_{ita}/(1\quad\sigma_n)}\Big(^{\sigma_n}}{\sum_{n=1}^{N_g}\quad\sum_{a\,/\,\mathcal{F}_n} e^{V_{ita}/(1\quad\sigma_n)}\Big[^{1\quad\sigma_n}} = \prod_{t=1}^{T}\frac{e^{V(\mathcal{Y}_{ta_t}^k,\alpha_{a_t},\beta)/(1\quad\sigma_n)}\Big)\sum_{a\,/\,\mathcal{F}_n} e^{V(\mathcal{Y}_{ta}^k,\alpha_a,\beta)/(1\quad\sigma_n)}\Big(^{\sigma_n}}{\sum_{n=1}^{N_g}\Big)\sum_{a\,/\,\mathcal{F}_n} e^{V(\mathcal{Y}_{ta}^k,\alpha_a,\beta)/(1\quad\sigma_n)}\Big(^{1\quad\sigma_n}}. \tag{1.2.15}$$

Equations (1.2.12) and (1.2.15) are both examples of equation (1.2.8) when the joint distribution in (1.2.5), $H(\epsilon)$, takes the two joint extreme value distributions aforementioned respectively.

These extensions in the family of GEV enjoy the closed form expressions for

[16] A more general assumption is to allow alternatives to be included in more than one subsets. This is because similarity between alternatives can be found in different aspects of their observed attributes. Different partition can be used to tell the similarity in different aspects. That is to say subsets can be overlapped. Models based on overlapped subsets are called *generalized nested logit* models.

equations (1.2.8) and (1.2.10), and their advantage is that we could adopt the method of CFHN (2013) for estimation and inference without extra efforts. Nevertheless, we can not always have the closed-form expressions under other specifications. There are reasons for using other specifications rather than sticking to the multinomial logit and nested logit models. First of all, it is well known that the multinomial logit model implicitly imposes the independent irrelevant alternative assumption (IIA), which states the relative ratio of choice probabilities between any two alternatives doesn't depend on attributes of other alternatives. IIA assumption has been shown too restrictive by the famous example of taking car and bus of different colors as one's means of transport. Secondly, the difference of two type I extreme variables follows the logit distribution which is symmetric. Obviously it can not be used in cases where we have to model some skewness in the difference of error terms. Moreover allowing serial correlation in error terms is not practicable in GEV models. Last but no least, the way to capture alternatives specific dependences between error terms is not natural as we saw in the example of nested logit model, and further sophisticated pattern of correlation will make the closed form probability more and more complicated.

To have a more general result I would like to use the assumption on error ϵ_i in (1.2.5), which includes multinomial logit model and nested multinomial logit model as its special cases.

1.2.2 General Models

For specifications other than a joint GEV distribution for ϵ_{it}, even for the integration of a single period, i.e. (1.2.10), it is hard to find a closed-form expression.[17] Generally, when researchers are completely free to make specifications on ϵ_i's distribution to uncover the economy reality which they believe, to expect closed-form expressions for $\mathcal{P}_j^k(\alpha, \beta)$ and $\mathcal{P}_{a_t}^{x_t}(\alpha, \beta)$ is unrealistic. Furthermore you will see other components in the model would have the same problem very soon. Fortunately simulation methods can be used to approximate them. The rest of the paper will focus on these cases where simulation is used for any components having non-closed-form expressions under the general assumption (1.2.5) of ϵ_i's distribution.

Primarily we are interested in parameters β in the multinomial discrete choice model or some functions of them. In CFHN (2013), they studied the average treatment effect on the outcome for binary choice models, where the discrete outcome took either 0 or 1 such that the average treatment effect has a natural interpre-

[17] For example, although you can have the closed-form probability for binary probit model, you can not do it for a multinomial probit model.

tation as the change of the conditional probability of choosing 1 instead of 0 at period t. For multinomial discrete choice model, the direct expansion of average treatment effect on the outcome does not have a parallel interpretation. However, it is found that in some applications the discrete choice set $\mathcal{F}$ can be mapped into an ordinal scale, for example the choice of different contracts from your mobile carrier indicates different expenditures, minutes of phone calls and so on. In such cases researchers may be interested in the average treatment effect on the mapped outcome. For example Kenneth E Train, D. L. McFadden, and Ben-Akiva 1987 analyzed the number and duration of phone calls made by households, using a discrete choice model instead of a regression because the discrete choice model allows for greater flexibility in handling the nonlinear price schedules. For such potential applications, Blundell and Powell 2006's method of average structural function (ASF) is useful and the conditional average partial effect at period t for individuals with attributes $\mathcal{Y}^k$ can be defined as follows

$$\begin{aligned} \Delta^k &= E \Big] g_0(x_t^a, \alpha_i, \epsilon_{it}, \beta) \quad g_0(x_t^b, \alpha_i, \epsilon_{it}, \beta) \Big\| X_i \ / \ \mathcal{Y}^k \Big\{ \\ &= E_{\alpha_i} \Big] E_{\epsilon_{it}} \Big] g_0(x_t^a, \alpha_i, \epsilon_{it}, \beta) \quad g_0(x_t^b, \alpha_i, \epsilon_{it}, \beta) \Big\{ \Big\| X_i \ / \ \mathcal{Y}^k \Big\{, \end{aligned} \tag{1.2.16}$$

where x_t^a and x_t^b are individual attributes after and before the treatment respectively. The second equation holds for the reason ϵ_i is independent of X_i and α_i. This partial effect depends on k because α_i could be correlated with X_i. To make the estimation for (1.2.16) more precisely, we assume that the distribution for ϵ_{it} doesn't change over t such that in the previous definition Δ^k doesn't depend on t[18].

The average treatment effect with the interpretation in terms of the change of choice probability also has its equivalent in the set up of multinomial discrete choice model. Let's consider the change in the conditional choice probability directly and define the change at period t also conditional on the fixed effects α_i as follows

$$\Delta\mathcal{P}(\alpha, \beta) = \mathcal{P}_{a_t}^{x_t^a}(\alpha, \beta) \quad \mathcal{P}_{a_t}^{x_t^b}(\alpha, \beta), \tag{1.2.17}$$

where x_t^a and x_t^b denote the attributes after the treatment and before the treatment respectively. Therefore the change of choice probability only conditional on the observable attributes, $\mathcal{Y}^k$, can be defined as

$$\Delta^k P = \Big[\ \Delta\mathcal{P}(\alpha, \beta) dF_k(\alpha) = \Big[\ \mathcal{P}_{a_t}^{x_t^a}(\alpha, \beta) \quad \mathcal{P}_{a_t}^{x_t^b}(\alpha, \beta) dF_k(\alpha), \tag{1.2.18}$$

[18] This is not an essential assumption. However, in static models this assumption makes treatment effects over periods homogeneous and improve the efficiency of its estimation. While in dynamic models this assumption per se is not enough for getting homogeneous treatment effects.

where $x_t^b = \mathcal{Y}_t^k$ generally, but you can choose different value for x_t^b.

In this section more new assumptions are introduced, it is better to summarize all the assumptions as follows before I continue

Assumption 1.2.1. *Individuals' potential utility function is of the form of equation* (1.2.2)*, and thus they behave according to function* (1.2.3) *or* (1.2.4)*. And the error term ϵ_i is independent of X_i and α_i and follows a distribution as* (1.2.5)*, and its marginal distributions for each period t are identical. α_i's support* Υ *is a compact subset of Euclidean space.*[19]

Since the assumption of a completely known distribution for ϵ_i, the normalization problem for the general multinomial discrete choice model has been done simultaneously. In the next subsection, I will focus on the individual behavior conditional on observable attributes X_i.

1.2.3 Model the Behavior Conditional on X_i

So far information on the level of individual decision makers has been studied. Since there exists fixed effects, α_i, without knowing its conditional distribution it prevents researchers from using maximum likelihood method as D. McFadden 1973 did. Some literatures use the aggregated information based on the individual behavior rule to match its pertaining observable higher level data.[20]

To see how to aggregate individual decision makers' behavior, it is better to restate the discrete value assumption about X_i's support, such that individuals' behavior can be aggregated conditional on its value.

Assumption 1.2.2. *X_i is a discrete variable or can be convincingly translated into a discrete variable, and the support of X_i is a finite set which can be written as $\mathcal{Y} = \}\mathcal{Y}^1, \times\times\times, \mathcal{Y}^K|$, where K is the cardinality of $\mathcal{Y}$.*

Given the assumption above it is easy to study the aggregate behavior for individuals with the same attributes $X_i = \mathcal{Y}^k$. Especially it is crucial to get model choice probability for different types of individuals pertaining to X_i, i.e. the conditional probability $P_r(Y_i = \mathcal{Z}^j \| X_i = \mathcal{Y}^k)$, where $\mathcal{Z}^j$ is one of the possible choice history out of $\}\mathcal{Z}^1, \times\times\times, \mathcal{Z}^J|$, where J is the number of different choice histories.

In this paper arbitrary dependence between the time invariant heterogeneity, α_i, and the explanatory variables in the representative utility, X_i, is allowed and α_i's possible distribution conditional on $X_i = \mathcal{Y}^k$ is denoted as $F_k(\alpha)$ without imposing any assumptions on it. Since the choice probability conditional on the fixed effects α_i is given by (1.2.8), $P_r(Y_i = \mathcal{Z}^j \| X_i = \mathcal{Y}^k)$ is supposed to be obtained

[19] The dimension of the Euclidean space is determined by both A and the dimension of a_{ia}, e.g., if we allows 2 heterogeneities for each alternative, this will be $\mathbb{R}^{2A}$.

[20] For example S. Berry, Levinsohn, and Pakes 1995 and Chernozhukov et al. 2013a.

by integrating out α_i with respect to $F_k(\alpha)$. This intuition gives the assumption as follows

Assumption 1.2.3. $P_r(Y_i = \mathcal{Z}^j \| X_i = \mathcal{Y}^k) = \int \mathcal{P}_j^k(\alpha, \beta) dF_k(\alpha)$, $(j = 1, \times\times\times, J;$ $k = 1, \times\times\times, K)$.

For the analysis of treatment effect, part of the idea has been discussed in the last subsection. Here I give more details. First of all, let's denote the average treatment effect on the outcome given both X_i and α_i as

$$\Delta(\alpha, \beta) = E_{\epsilon_{it}} \Big] g_0(x_t^a, \alpha, \epsilon_{it}, \beta) \quad g_0(x_t^b, \alpha, \epsilon_{it}, \beta) \Big\{, \tag{1.2.19}$$

where this definition makes sense because ϵ_i is assumed to be independent of X_i and α_i. Therefore equation (1.2.16) can be summarized as the following assumption

Assumption 1.2.4. $\Delta^k = \int \Delta(\alpha, \beta) dF_k(\alpha)$.

Since it is assumed in assumption 2.4.1 that ϵ_{it} is identically distributed over t, $\Delta(\alpha, \beta)$ does not depend on t. Furthermore, x_t^a and x_t^b are controlled values of x_{it} after and before the treatment, as a result $\Delta(\alpha, \beta)$ doesn't depend on X_i either.

To study treatment effect on the choice probability in period t, I introduce the conception of *potential choice probability*. It is defined as the imaginary probability of choosing a_t in period t by individual with attributes $\mathcal{Y}^k$ while acting as other individual with attributes x_t at period t. I denote the potential choice probability by $P_r(y_{it} = a_t \| (x_{it} = x_t), X_i = \mathcal{Y}^k)$. It can be calculated by integrating out α from $\mathcal{P}_{a_t}^{x_t}(\alpha, \beta)$ with respect to $F_k(\alpha)$. This intuition is given as the following assumption

Assumption 1.2.5. $P_r(y_{it} = a_t \| (x_{it} = x_t), X_i = \mathcal{Y}^k) = \int \mathcal{P}_{a_t}^{x_t}(\alpha, \beta) dF_k(\alpha)$.

Thereafter the treatment effect on the choice probability at period t can be defined as:

$$\Delta^k P = P_r(y_{it} = a_t \| (x_{it} = x_t^a), X_i = \mathcal{Y}^k) \quad P_r(y_{it} = a_t \| (x_{it} = x_t^b), X_i = \mathcal{Y}^k), \tag{1.2.20}$$

where x_t^a and x_t^b are attributes after and before the treatment. If the treatment effect on the choice probability conditional on the fixed effects is defined as

$$\Delta \mathcal{P}(\alpha, \beta) = \mathcal{P}_{a_t}^{x_t^a}(\alpha, \beta) \quad \mathcal{P}_{a_t}^{x_t^b}(\alpha, \beta), \tag{1.2.21}$$

then the treatment effect on the choice probability (1.2.20) can be rewritten as

$$\Delta^k P = \Big[\ \mathcal{P}_{a_t}^{x_t^a}(\alpha, \beta) \quad \mathcal{P}_{a_t}^{x_t^b}(\alpha, \beta) dF_k(\alpha) = \Big[\ \Delta \mathcal{P}(\alpha, \beta) dF_k(\alpha). \tag{1.2.22}$$

1.3 Model Identification and Estimation

The aforementioned aggregate behavior can be used to achieve set identification. This idea of set identification was proposed by Honoré and Tamer 2006 and further developed by CFHN (2013). In this section, I follow the schedule of section 7 in CFHN (2013) and give the set identification for both the structure parameter β and the conditional average treatment effects Δ^k and $\Delta^k P$.

First of all I denote the true probability of choosing $\mathcal{Z}^j$ conditional on $X_i = \mathcal{Y}^k$ as

$$\mathcal{S}_j^k = P(Y_i = \mathcal{Z}^j \| X_i = \mathcal{Y}^k), \tag{1.3.1}$$

and let $\mathcal{S} = (\mathcal{S}_1^1, \times\times\times, \mathcal{S}_J^1, \times\times\times, \mathcal{S}_1^K, \times\times\times, \mathcal{S}_J^K)$. Those true conditional probabilities should equal to their model implied counterparts defined in assumption 1.2.3 if our model specifications are correct.

Because there are no restrictions on the conditional distribution functions $F_k(\alpha)$ for all k, the parameter β can only be set identified as a set which consists of all the β, for which we can find a conditional distribution function $F_k(\alpha)$ for each k that makes the model implied conditional probabilities equal the true conditional probabilities. To be concrete, for any given β let's first define the set of all the $F_k(\alpha)$ that are consistent with $(\beta, \mathcal{S})$ as

$$\mathcal{G}_k(\beta, \mathcal{S}) = \Big\} F_k(\alpha) : \mathcal{S}_j^k = \Big[\ \mathcal{P}_j^k(\alpha, \beta) dF_k(\alpha), j = 1, \times\times\times, J \Big(. \tag{1.3.2}$$

Thereafter the identified set of β can be defined as

$$B = \}\beta : \mathcal{G}_k(\beta, \mathcal{S}) \not\forall \emptyset, \exists k = 1, \times\times\times, K| \ , \tag{1.3.3}$$

where B consists of all the parameters that can generate, together with some possible conditional distribution of α_i, $F_k(\alpha)$, the same aggregate choice probabilities as the true DGP does. Given the definition of the identified set of β, the sharp upper and lower bounds Δ_u^k and Δ_l^k for the average treatment effect on the outcome for individuals with attributes $\mathcal{Y}^k$, a.k.a. Δ^k, can be defined as follows

$$\begin{aligned} \Delta_u^k &= \sup_{\beta / B, F_k / \mathcal{G}_k(\beta, \mathcal{S})} \Big[\ \Delta(\alpha, \beta) dF_k(\alpha) \quad \text{and} \\ \Delta_l^k &= \inf_{\beta / B, F_k / \mathcal{G}_k(\beta, \mathcal{S})} \Big[\ \Delta(\alpha, \beta) dF_k(\alpha). \end{aligned} \tag{1.3.4}$$

Similarly, the upper and lower bounds $\Delta_u^k P$ and $\Delta_l^k P$ for the treatment effect on the choice probability for individuals with attributes $\mathcal{Y}^k$, a.k.a. $\Delta^k P$, can be defined as

$$(1.3.5)\qquad \begin{aligned} \Delta_u^k P &= \sup_{\beta / B, F_k / \mathcal{G}_k(\beta,\mathcal{S})} \Big[\ \Delta\mathcal{P}(\alpha,\beta) dF_k(\alpha) \quad \text{and} \\ \Delta_l^k P &= \inf_{\beta / B, F_k / \mathcal{G}_k(\beta,\mathcal{S})} \Big[\ \Delta\mathcal{P}(\alpha,\beta) dF_k(\alpha). \end{aligned}$$

Before starting the estimation for B, Δ_u^k, Δ_l^k, $\Delta_u^k P$ and $\Delta_l^k P$, it is better to first discuss how to calculate them. As is shown by Lemma 7 in CFHN (2013), the conditional distributions of the fixed effects, $F_k(\alpha)$, couldn't generally be identified in multinomial discrete choice models for the reason that every $F_k(\alpha)$, no matter it is continuous or discrete, has a discrete equivalent with no more than J mass points in its support Υ that gives exactly the same aggregate choice probability. Furthermore, restricting our attentions on those discrete equivalents doesn't affect the analysis of the upper and lower bounds for Δ^k.

To include the analysis of bounds for treatment effect on choice probability $\Delta^k P$, I give a slightly augmented variant of Lemma 7 in CFHN (2013) as follows

Lemma 1.3.1. *If Assumptions 1.2.2 and 1.2.3 are satisfied and $\mathcal{P}_j^k(\alpha,\beta)$ is a measurable function of α for each β / $\mathbb{B}$, where $\mathbb{B}$ is the parameter space, then for each β and every CDF F_k on Υ, there is a discrete distribution F_k^J with no more than J support points such that $\int \mathcal{P}_j^k(\alpha,\beta) dF_k^J(\alpha) = \int \mathcal{P}_j^k(\alpha,\beta) dF_k(\alpha)$ $(j = 1, \times\times\times, J)$. If, in addition, $\Delta(\alpha,\beta)$ is bounded for each β, then Δ_u^k and Δ_l^k are not affected by restricting attention to F_k / $\mathcal{G}_k(\beta)$ that are discrete with no more than J support points. Similarly if $\Delta\mathcal{P}(\alpha,\beta)$ is bounded for each β, then $\Delta_u^k P$ and $\Delta_l^k P$ are not affected by restricting attention to F_k / $\mathcal{G}_k(\beta)$ that are discrete with no more than J support points either.*

Consequently, we can consider discrete candidates of F_k exclusively. But there is still a problem that lemma 1.3.1 doesn't tell where are the mass points for each k? CFHN (2013) proposed a refining fixed grids approaching method.

Let's construct a fixed grid for the fixed effects α_i over its support Υ and denote the grid as $\Upsilon_M = (\bar{\alpha}_{1M}, \times\times\times, \bar{\alpha}_{MM})^{\in}$, where M is the number of grids and the fixed grid can be refined by increasing M, and the upper bar indicates something that is fixed or corresponding to a fixed grid. Let $\bar{\pi}^k = (\bar{\pi}_1^k, \times\times\times, \bar{\pi}_M^k)^{\in}$denote a distribution over the fixed grids. Then $\bar{\pi} = (\bar{\pi}^{1\in}, \times\times\times, \bar{\pi}^{K\in})^{\in}$denotes a $MK * 1$ vector of α_i's conditional distributions for all k with $\bar{\pi}^k$ being an element of the M dimensional unit simplex $\mathcal{X}_M$. After all $(\Upsilon_M, \bar{\pi})$ is used to approximate any $\}F_k^J|_{k=1}^K$ with no more than J support points.

Thereafter the model implied conditional choice probabilities can be approximated by

$$P_j^k(\beta, \bar{\pi}^k, M) = \sum_{m=1}^{M} \bar{\pi}_m^k \mathcal{P}_j^k(\bar{\alpha}_{mM}, \beta), \text{ for all } j \text{ and } k. \tag{1.3.6}$$

Afterwards the following quadratic object function is used to serve the purpose of matching the model implied choice probabilities with the true probabilities.

$$T_\lambda(\beta, \bar{\pi}, M) = \sum_{j,k} \omega_j^k \Big] \mathcal{S}_j^k \quad P_j^k(\beta, \bar{\pi}^k, M) \Big\{^2 + \lambda_n \bar{\pi}^{\in} \bar{\pi}, \tag{1.3.7}$$

where ω_j^k are positive weights, and CFHN (2013) proposed a chi-square weight $\omega_j^k = \mathcal{S}^k / \mathcal{S}_j^k$, for $\mathcal{S}^k = P(X_i = \mathcal{Y}^k)$. And λ_n is a penalty multiplier that controls the impact of the penalty term $\lambda_n \bar{\pi}^{\in} \bar{\pi}$.

Then the identified set for β, a.k.a. B, can be approximated by the following set

$$B(M) = \} \beta : \mathcal{A} \bar{\pi} \text{ s.t. } T_\lambda(\beta, \bar{\pi}, M) \leqslant \xi_n | \text{ , } \xi_n > 0, \tag{1.3.8}$$

where the positive threshold ξ_n ensures that the set sequence $\} B(M) |$ is lower hemi-continuous and that $B(M)$ need not be smaller than the identified set.[21]

For the bounds of conditional average treatment effect on the outcome at period t, i.e. Δ^k, note that

$$D^k(M) = \Bigg\} \sum_{m=1}^{M} \bar{\pi}_m^k \Delta(\bar{\alpha}_{mM}, \beta,) : T_\lambda(\beta, \bar{\pi}, M) \leqslant \xi_n \Bigg\lceil \tag{1.3.9}$$

is the approximate set of all the possible conditional average treatment effects. Approximate lower and upper bounds of Δ^k are respectively defined as

$$\Delta_l^k(M) = \min_{\beta, \bar{\pi}} D^k(M) \text{ and } \Delta_u^k(M) = \max_{\beta, \bar{\pi}} D^k(M). \tag{1.3.10}$$

Similarly, for $\Delta^k P$, let's define

$$D^k P(M) = \Bigg\} \sum_{m=1}^{M} \bar{\pi}_m^k \Delta \mathcal{P}(\bar{\alpha}_{mM}, \beta,) : T_\lambda(\beta, \bar{\pi}, M) \leqslant \xi_n \Bigg\lceil \tag{1.3.11}$$

as the approximate set of all the possible treatment effects on the choice probability

[21] See Section 8 of Chernozhukov et al. 2013a for more discussion.

at period t, and its upper and lower bounds can be defined as follows

$$\Delta_l^k P(M) = \min_{\beta,\bar{\pi}} D^k P(M) \text{ and } \Delta_u^k P(M) = \max_{\beta,\bar{\pi}} D^k P(M). \tag{1.3.12}$$

Discussion above gives a calculation method of the identified set of β, i.e. $B(M)$, and the bounds for the conditional average treatment effect on the outcome and treatment effect on the choice probability if the following components are known to researchers: the positive weights ω_j^k, the true choice probability $\mathcal{S}_j^k$, the conditional probability $\mathcal{P}_j^k(\alpha,\beta)$, the conditional average partial effect $\Delta(\alpha,\beta)$ and the treatment effect on the choice probability $\Delta\mathcal{P}(\alpha,\beta)$. Generally these quantities and functions are unknown, and either estimation or simulation is needed to calculate them. The positive weight ω_j^k and true choice probability $\mathcal{S}_j^k$ can be easily estimated by their corresponding sample frequencies. For $\mathcal{P}_j^k(\alpha,\beta)$, $\Delta(\alpha,\beta)$ and $\Delta\mathcal{P}(\alpha,\beta)$, they generally don't have closed-form expressions. To solve this problem, I propose an uniform convergence estimator for each of them using the simulated errors from the distribution of ϵ_i, $H(\epsilon)$.

Suppose there is a generator which gives random vectors $\}\tilde{\epsilon}_i|_{i=1}^{r}$ from the distribution $H(\epsilon)$, where the number of replications $r \quad / \quad \infty$ as the sample size $n \quad / \quad \infty$ and the tilde over ϵ_i emphasizes the fact that they are generated by simulation. For a given triplet $(\mathcal{Y}^k,\alpha,\beta)$, the simulated data $\}\tilde{Y}_i = (\tilde{y}_{i1}, \times\!\times\!\times, \tilde{y}_{iT})|_{i=1}^{r}$ are given by using the behavior function (1.2.4). Thereafter with those generated data the estimator for $\mathcal{P}_j^k(\alpha,\beta)$ can be defined as

$$\tilde{\mathcal{P}}_j^k(\alpha,\beta) = \frac{1}{r}\sum_{i=1}^{r} \mathbb{1}(\tilde{Y}_i = \mathcal{Z}^j). \tag{1.3.13}$$

Following the way of restating (1.2.6) as (1.2.8), (1.3.13) can be restated in terms of indicator functions with simulated data as follows

(1.3.14)

$$\begin{aligned}\tilde{\mathcal{P}}_j^k(\alpha,\beta) &= \frac{1}{r}\sum_{i=1}^{r}\prod_{t=1}^{T}\prod_{a_t^{\in}\neq a_t} \mathbb{1}\Big) V(\mathcal{Y}_{ta_t}^k,\alpha_{a_t},\beta) + \tilde{\epsilon}_{ita_t} \geqslant V(\mathcal{Y}_{ta_t^{\in}}^k, \alpha_{a_t^{\in}}, \beta) + \tilde{\epsilon}_{ita_t^{\in}}\Big(\\ &= \frac{1}{r}\sum_{i=1}^{r}\prod_{t=1}^{T}\prod_{a_t^{\in}\neq a_t} \mathbb{1}\Big) \tilde{\epsilon}_{ita_t^{\in}} \quad \tilde{\epsilon}_{ita_t} \leqslant V(\mathcal{Y}_{ta_t}^k,\alpha_{a_t},\beta) \quad V(\mathcal{Y}_{ta_t^{\in}}^k, \alpha_{a_t^{\in}}, \beta)\Big(.\end{aligned}$$

For given $(x^a, x^b, \alpha, \beta)$, where x^a and x^b are attributes after and before the treatment and the same value for each t is used wherein, i.e. $x_t = x_{t^{\in}}$ another simulated data under treatment $\}\tilde{Y}_i(x^a), \tilde{Y}_i(x^b)|_{i=1}^{r}$ is given by using behavior function (1.2.4) again, where $\tilde{Y}_i(x^a) = \ \tilde{y}_{i1}(x_1^a), \times\!\times\!\times, \tilde{y}_{iT}(x_T^a)\Big[^{\in}$ and $\tilde{y}_{it}(x_t^a) = g_0(x_t^a, \alpha, \tilde{\epsilon}_{it}, \beta)$, similarly $\tilde{Y}_i(x^b) = \ \tilde{y}_{i1}(x_1^b), \times\!\times\!\times, \tilde{y}_{iT}(x_T^b)\Big[^{\in}$ and $\tilde{y}_{it}(x_t^b) = g_0(x_t^b, \alpha, \tilde{\epsilon}_{it}, \beta)$. Afterwards let's define

$$\tilde{\Delta}_t(\alpha,\beta) = \frac{1}{r}\sum_{i=1}^{r}\Big]\tilde{y}_{it}(x_t^a) \quad \tilde{y}_{it}(x_t^b)\Big\{, \tag{1.3.15}$$

which is the estimator of $\Delta(\alpha,\beta)$ using only information from period t. Since it is assumed that ϵ_{it}'s distribution doesn't change over t, a more efficient estimator for $\Delta(\alpha,\beta)$ using information from all periods can be proposed as follows

$$\tilde{\Delta}(\alpha,\beta) = \frac{1}{T}\sum_{t=1}^{T}\tilde{\Delta}_t(\alpha,\beta). \tag{1.3.16}$$

Quite similarly, for $\Delta\mathcal{P}(\alpha,\beta)$ the estimator using only information from period t can be defined as

$$\tilde{\Delta}_t\mathcal{P}(\alpha,\beta) = \frac{1}{r}\sum_{i=1}^{r}\Big]\mathbb{1}\left(\tilde{y}_{it}(x_t^a) = a_t\right) \quad \mathbb{1}\Big)\tilde{y}_{it}(x_t^b) = a_t\Big(\Big\{, \tag{1.3.17}$$

and a more efficient estimator using all information can be defined as its average over periods

$$\tilde{\Delta}\mathcal{P}(\alpha,\beta) = \frac{1}{T}\sum_{t=1}^{T}\tilde{\Delta}_t\mathcal{P}(\alpha,\beta). \tag{1.3.18}$$

Since (1.2.9) can be expressed in terms of indicator functions as in (1.2.10), (1.3.18) can be rewritten in terms of indicator functions of $\tilde{\epsilon}_i$ as follows

(1.3.19)

$$\begin{aligned}
\tilde{\Delta}\mathcal{P}(\alpha,\beta) &= \frac{1}{T}\sum_{t=1}^{T}\Bigg\}\frac{1}{r}\sum_{i=1}^{r}\Big]\mathbb{1}\left(g_0(x_t^a,\alpha,\tilde{\epsilon}_{it},\beta) = a_t\right) \quad \mathbb{1}\Big)g_0(x_t^b,\alpha,\tilde{\epsilon}_{it},\beta) = a_t\Big(\Big\{\Big\lceil \\
&= \frac{1}{T}\sum_{t=1}^{T}\Bigg\}\frac{1}{r}\sum_{i=1}^{r}\Bigg]\prod_{a_t^{\in}\forall a_t}\mathbb{1}\Big]V(x_{ta_t}^a,\alpha_{a_t},\beta) + \tilde{\epsilon}_{ita_t} \geqslant V(x_{ta_t^{\in}}^a,\alpha_{a_t^{\in}},\beta) + \tilde{\epsilon}_{ita_t^{\in}}\Big\{ \\
&\qquad\qquad \prod_{a_t^{\in}\forall a_t}\mathbb{1}\Big]V(x_{ta_t}^b,\alpha_{a_t},\beta) + \tilde{\epsilon}_{ita_t} \geqslant V(x_{ta_t^{\in}}^b,\alpha_{a_t^{\in}},\beta) + \tilde{\epsilon}_{ita_t^{\in}}\Big\{\Big\lfloor\Big\rfloor\Bigg\} .
\end{aligned}$$

Given all the building blocks, here comes the simulation based estimation method. Define the simulated choice probability with fixed grids for (1.3.6) as follows

$$\tilde{P}_j^k(\beta,\bar{\pi}^k,M) = \sum_{m=1}^{M}\bar{\pi}_m^k\tilde{\mathcal{P}}_j^k(\bar{\alpha}_{mM},\beta), \tag{1.3.20}$$

and the estimated quadratic objective function for (1.3.7) is

$$\hat{T}_\lambda(\beta, \bar{\pi}) = \sum_{j,k} \hat{\omega}_j^k \Big] \hat{\mathcal{S}}_j^k \quad \tilde{P}_j^k(\beta, \bar{\pi}^k, M)\Big\{^2 + \lambda_n \bar{\pi}^{\in}\bar{\pi}, \tag{1.3.21}$$

where the true choice probability $\mathcal{S}_j^k$ has been replaced by the data cell probability[22] $\hat{\mathcal{S}}_j^k$ and the model implied choice probability has been replaced by its equivalent of simulation. Afterwards the identified set of β, i.e. B, is estimated by the following random set

$$\hat{B} = \big\}\beta \ / \ \mathbb{B} : \mathcal{A}\bar{\pi} \text{ s.t. } \hat{T}_\lambda(\beta, \bar{\pi}) \leqslant \xi_n \big| \ , \tag{1.3.22}$$

and the lower and upper bounds for the conditional average treatment effect Δ^k are estimated by

$$\hat{\Delta}_l^k = \min \hat{D}^k \text{ and } \hat{\Delta}_u^k = \max \hat{D}^k, \tag{1.3.23}$$

where

$$\hat{D}^k = \Big\}\tilde{\Delta}^k(\beta, \bar{\pi}^k, M) : \hat{T}_\lambda(\beta, \bar{\pi}) \leqslant \xi_n \quad , \tag{1.3.24}$$

and

$$\tilde{\Delta}^k(\beta, \bar{\pi}^k, M) = \sum_{m=1}^{M} \bar{\pi}_m^k \tilde{\Delta}(\bar{\alpha}_{mM}, \beta). \tag{1.3.25}$$

For the bounds of treatment effect on the choice probability, the estimators of the lower and upper bounds are given by

$$\hat{\Delta}_l^k P = \min \hat{D}^k P \text{ and } \hat{\Delta}_u^k P = \max \hat{D}^k P, \tag{1.3.26}$$

where

$$\hat{D}^k P = \Big\}\tilde{\Delta}^k P(\beta, \bar{\pi}^k, M) : \hat{T}_\lambda(\beta, \bar{\pi}) \leqslant \xi_n \quad , \tag{1.3.27}$$

and

$$\tilde{\Delta}^k P(\beta, \bar{\pi}^k, M) = \sum_{m=1}^{M} \bar{\pi}_m^k \tilde{\Delta}\mathcal{P}(\bar{\alpha}_{mM}, \beta). \tag{1.3.28}$$

In the next section, I study asymptotic behavior of my simulation based estimators and I show that under some regulatory conditions they are consistent, and

[22] Aka the data frequency that individuals with attributes $\mathcal{Y}^k$ choose alternative $\mathcal{Z}^j$.

the perturbed bootstrap can be used to learn their asymptotic distribution.

1.4 Asymptotic Behavior of the Simulation Based Estimator

Since simulation is introduced to the estimation, people are interested in its effect on the asymptotic behavior of the new estimator, especially the property of consistency. In this section I do two jobs. First I study the conditions under which the consistency of the new estimator still holds. These conditions are given by a series of assumptions and several lemmas and theorems are given in order such that the profile of the proof can be seen clearly. Some comments are given for these results, and all the details are reserved in the appendix section. Second, I refer to the method of perturbed bootstrap by CFHN (2013) and show how it can be used to learn the new estimator's asymptotic distribution.

1.4.1 Consistency

Besides the notations for the fixed grid Υ_M in α_i's support Υ and its corresponding conditional probability vector $\bar{\pi}$ for α_i's conditional distributions, denote the unknown variable grid for α_i with only $J+1$[23] support points as $\}\alpha_1^k, \times\times\times, \alpha_{J+1}^k|$ and denote its corresponding conditional probabilities as π^k, so that $\pi = (\pi^{1\in}, \times\times\times, \pi^{K⋹})^{\in}$. Let $\alpha^k = (\alpha_1^k, \times\times\times, \alpha_{J+1}^k)^{⋹}$, $\alpha = (\alpha^{1\in}, \times\times\times, \alpha^{K⋹})^{\in}$and $\gamma = (\alpha^{\in}, \pi^{⋹})^{⋹}$ so that all the parameters needed in the model can be denoted as $\phi = (\beta^{⋹}\gamma^{⋹})^{\in} / \Phi = \mathbb{B} * \Upsilon^{(J+1)K} * \mathcal{X}_{J+1}^K$. Also denote $\tilde{P}_j^k(\phi) = \sum_{l=1}^{J+1} \tilde{\mathcal{P}}_l^k(\alpha_l^k, \beta)\pi_l^k$ and $P_j^k(\phi) = \sum_{l=1}^{J+1} \mathcal{P}_l^k(\alpha_l^k, \beta)\pi_l^k$ and $\tilde{\Delta}^k(\phi) = \sum_{l=1}^{J+1} \tilde{\Delta}(\alpha_l^k, \beta)\pi_l^k$ and $\tilde{\Delta}^k P(\phi) = \sum_{l=1}^{J+1} \tilde{\Delta}\mathcal{P}(\alpha_l^k, \beta)\pi_l^k$.

First of all, I show that $\tilde{\mathcal{P}}_j^k(\alpha, \beta)$ uniformly converges to $\mathcal{P}_j^k(\alpha, \beta)$ over $\Upsilon * \ \mathbb{B}$ in probability and the rate of convergence is $\quad \bar{r}$.

Lemma 1.4.1. *Estimator $\tilde{\mathcal{P}}_j^k(\alpha, \beta)$ defined in (1.3.13) uniformly converges to $\mathcal{P}_j^k(\alpha, \beta)$ over $\Upsilon * \ \mathbb{B}$ in probability if assumption (2.4.1) holds. Furthermore,*

$$\bar{r}(\tilde{\mathcal{P}}_j^k(\alpha, \beta) \quad \mathcal{P}_j^k(\alpha, \beta)) \rightsquigarrow G(\alpha, \beta), \tag{1.4.1}$$

where $G(\alpha, \beta)$ is a mean zero Gaussian process, and its finite dimensional distribution is controlled by the distribution of $\tilde{\epsilon}_i$, thus by $H(\epsilon)$.

For other two simulation based random functions $\tilde{\Delta}(\alpha, \beta)$ and $\tilde{\Delta}\mathcal{P}(\alpha, \beta)$, they

[23] Lemma 1.3.1 tells us it is enough to consider unknown grids with no more than J support points, here to consider one more support point is to facilitate the proofs afterwards and this practice doesn't break the lemma.

need also uniformly converge to their expectations $\Delta(\alpha,\beta)$ and $\Delta\mathcal{P}(\alpha,\beta)$ respectively. This is given by the following lemmas.

Lemma 1.4.2. *Estimator $\tilde{\Delta}(\alpha,\beta)$ defined in equation* (1.3.16) *uniformly converges to $\Delta(\alpha,\beta)$ over $\Upsilon * \mathbb{B}$ in probability if assumptions* (2.4.1) *holds.*

Lemma 1.4.3. *Estimator $\tilde{\Delta}\mathcal{P}(\alpha,\beta)$ defined in equation* (1.3.18) *uniformly converges to $\Delta\mathcal{P}(\alpha,\beta)$ over $\Upsilon * \mathbb{B}$ in probability if assumptions* (2.4.1) *holds.*

Since the proof of consistency for bounds of Δ^k and $\Delta^k P$ are similar, I first give the steps of proof for Δ^k in detail and then give the parallel steps for $\Delta^k P$ without details. First of all, I show that for large grid M, here $M \geqslant J+1$ is needed, for every combination of K probability mass functions on fixed grids, i.e. $\bar{\pi} \ / \ \mathcal{X}_M^K$, there are distribution functions $F_k^{J+1}(\alpha), k = 1, \times\times\times, K$, with their supports are subsets of Υ_M having no more than $J+1$ elements, which can generate the same choice probability and partial effect as $\bar{\pi}$ can. This result can be stated as the following Lemma.

Lemma 1.4.4. *For every $\bar{\pi} \ / \ \mathcal{X}_M^K$, where $M > J$, there exists*

$$\phi(\beta,\bar{\pi}) = \Big)\beta^{\in}, \gamma^1(\beta,\bar{\pi}^1)^{\in}, \times\times\times, \gamma^K(\beta,\bar{\pi}^K)^{\in}\Big(^{\in}$$

such that

$$\begin{aligned} \tilde{P}_j^k(\beta,\bar{\pi}^k,M) &= \tilde{P}_j^k \Big)\beta, \gamma^k(\beta,\bar{\pi}^k)\Big(\\ \tilde{\Delta}^k(\beta,\bar{\pi}^k,M) &= \tilde{\Delta}^k \Big)\beta, \gamma^k(\beta,\bar{\pi}^k)\Big(, \end{aligned}$$

for all $j = 1, \times\times\times, J$ and $k = 1, \times\times\times, K$, and where

$$\begin{aligned} \tilde{P}_j^k \Big)\beta, \gamma^k(\beta,\bar{\pi}^k)\Big(&= \sum_{l=1}^{J+1} \pi_l^k \tilde{\mathcal{P}}_j^k(\bar{\alpha}_{m_l^k M}, \beta) \\ \tilde{\Delta}^k \Big)\beta, \gamma^k(\beta,\bar{\pi}^k)\Big(&= \sum_{l=1}^{J+1} \pi_l^k \tilde{\Delta}(\bar{\alpha}_{m_l^k M}, \beta), \end{aligned}$$

and $\}m_l^k|_{l=1}^{J+1}$ is a subset of $\}1, \times\times\times, M|$, which can be different for different k.

One implication of Lemma 1.4.4 is that whenever the approximated model choice probability with fixed grid and simulation, i.e. $\tilde{P}_j^k(\beta,\bar{\pi}^k,M)$, is used, a pertaining choice probability with a variable sub-grid, $\tilde{P}_j^k \ \ \beta, \gamma^k(\beta,\bar{\pi}^k)[$, can be used instead. This fact is very useful for the rest of proof. Consider the following two functions

(1.4.2)

$$\hat{Q}\left(\phi(\beta,\bar{\pi})\right) = \Sigma_{j,k}\, \hat{\omega}_j^k \Big] \hat{\mathcal{S}}_j^k \quad \tilde{P}_j^k \ \ \beta, \gamma^k(\beta,\bar{\pi}^k) \Big[\Big\{^2 = \Sigma_{j,k}\, \hat{\omega}_j^k \Big] \hat{\mathcal{S}}_j^k \quad \tilde{P}_j^k(\beta,\bar{\pi}^k,M) \Big\{^2$$

and

(1.4.3)
$$Q(\phi) = \sum_{j,k} \omega_j^k \Big] \mathcal{S}_j^k \quad P_j^k(\phi) \Big\{^2 .$$

Define three subsets of the parameter space Φ, namely

(1.4.4)
$$\Phi_I = \}\phi \ / \ \Phi : Q(\phi) = 0|$$

and

(1.4.5)
$$\hat{\Phi} = \}\phi(\beta, \bar{\pi}) : \hat{Q}\left(\phi(\beta, \bar{\pi})\right) + \lambda_n \bar{\pi}^{\in} \bar{\pi} \leqslant \xi_n{}^{|}$$

and

(1.4.6)
$$\Phi_M = \Big\}\phi(\beta, \bar{\pi}) : \bar{\pi} \ / \ \mathcal{X}_M^K, \beta \ / \ \mathbb{B} \quad .$$

By construction the projection of $\hat{\Phi}$ onto $\mathbb{B}$ coincides with $\hat{B}$, thus I call $\hat{\Phi}$ the estimation set, and the projection of Φ_I onto $\mathbb{B}$ coincides with B, thus I call Φ_I the identified set. Φ_M contains all ϕ that is equivalent to some distribution on the fixed grid Υ_M, thus I call Φ_M the equivalent set.

In terms of the new notations, the identified set of conditional average treatment effects on the outcome can be expressed again as

(1.4.7)
$$D^k = \Big\}\Delta^k(\phi) : \phi \ / \ \Phi_I \quad ,$$

and its pertaining estimated set can be expressed as

(1.4.8)
$$\hat{D}^k = \Big\}\tilde{\Delta}^k \Big)\beta, \gamma^k(\beta, \bar{\pi}^k)\Big(: \phi(\beta, \bar{\pi}) \ / \ \hat{\Phi} \quad ,$$

where $\Delta^k(\phi) = \sum_{l=1}^{J+1} \Delta(\alpha_l^k, \beta)\pi_l^k$ and $\tilde{\Delta}^k \ \ \beta, \gamma^k(\beta, \bar{\pi}^k)\big[= \sum_{l=1}^{J+1} \tilde{\Delta}(\bar{\alpha}_{m_l^k M}, \beta)\pi_l^k$. In order to continue, the following assumption is also needed.

Assumption 1.4.1. *(i)* Υ *is a compact metric space endowed with the metric* $d(\alpha, \alpha^{⋹})$*; (ii)* $\mathbb{B}$ *is a compact subset of* $\mathbb{R}^b$*, where* b *is the number of elements in* β*; (iii) There is* $L < \infty$ *such that for all* $(\alpha, \beta), (\alpha^{\in}, \beta^{⋹}) \ / \ \Upsilon * \ \mathbb{B}$, $\|\Delta(\alpha, \beta) \quad \Delta(\alpha^{\in}, \beta^{⋹})\| \leqslant L\,[d(\alpha, \alpha^{⋹}) + \backslash\beta \quad \beta^{⋹}\backslash]$*; (iv)* $\Delta^k(\phi)$ *is continuous in* ϕ*.*

Here I introduce the same metric as CFHN (2013) did in the parametric space Φ, this metric is defined as

(1.4.9)
$$d(\phi, \phi^{⋹}) = \max_{j,k} \max \Big\}d(\alpha_j^k, \alpha_j^{⋹k}), \Big\|\pi_j^k \quad \pi_j^{⋹k}\Big\|, \Big(\beta \quad \beta^{\in}\Big(\quad .$$

Thereafter the following lemma holds.

Lemma 1.4.5. *There exists a constant C such that for all ϕ, $\phi^{\in}$ / Φ, the following inequality holds*

$$\left\|\tilde{\Delta}^k(\phi) \quad \tilde{\Delta}^k(\phi^{\in})\right\| \leqslant Cd(\phi,\phi^{\in}) + o_P(1),$$

where $o_P(1)$ doesn't depend on $(\phi,\phi^{\in})$.

It can be shown that the Lipschitz-like condition result in Lemma 1.4.5 implies a similar result in terms of Hausdorff metric, where the Hausdorff metric is defined as

$$d_H(\Phi,\Phi^{\in}) = \max \Big\} \sup_{\phi / \Phi} \inf_{\phi^{\in} / \Phi^{\in}} d(\phi,\phi^{\in}), \sup_{\phi^{\in} / \Phi^{\in}} \inf_{\phi / \Phi} d(\phi^{\in},\phi) \Big\lceil . \tag{1.4.10}$$

I claim this property as the following lemma

Lemma 1.4.6. *Denote the Hausdorff metric as d_H, Lemma 1.4.5 implies*

$$d_H \Big) \tilde{\Delta}^k(\Phi_s), \tilde{\Delta}^k(\Phi_s^{\in}) \Big(\leqslant Cd_H(\Phi_s,\Phi_s^{\in}) + o_P(1),$$

where Φ_s and $\Phi_s^{\in}$are two subsets of Φ and $o_P(1)$ doesn't depend on $(\Phi_s,\Phi_s^{\in})$.

Thereafter it can be proved that if $d_H \quad \hat{\Phi}, \Phi_I [\quad \overset{p}{/} \quad 0$, the simulation based estimators of upper and lower bounds for the conditional average treatment effect, i.e. Δ^k, are consistent. I state this result as the following lemma

Lemma 1.4.7. *If $d_H \quad \hat{\Phi}, \Phi_I [\quad \overset{p}{/} \quad 0$, we have*

$$\hat{\Delta}_l^k \quad \overset{p}{/} \quad \Delta_l^k \text{ and } \hat{\Delta}_u^k \quad \overset{p}{/} \quad \Delta_u^k.$$

Now it's time to consider the case for $\Delta^k P$. First of all, I have a parallel result of Lemma 1.4.4 as stated bellow

Lemma 1.4.8. *For every $\bar{\pi}$ / $\mathcal{X}_M^K$, where $M > J$, there exists*

$$\phi(\beta,\bar{\pi}) = \Big) \beta^{\in}, \gamma^1(\beta,\bar{\pi}^1)^{\in}, \times\times\times, \gamma^K(\beta,\bar{\pi}^K)^{\in} \Big(^{\in}$$

such that

$$\tilde{P}_j^k(\beta,\bar{\pi}^k,M) = \tilde{P}_j^k \Big) \beta, \gamma^k(\beta,\bar{\pi}^k) \Big($$
$$\tilde{\Delta}^k P(\beta,\bar{\pi}^k,M) = \tilde{\Delta}^k P \Big) \beta, \gamma^k(\beta,\bar{\pi}^k) \Big(,$$

for all $j = 1, \times\times\times, J$ and $k = 1, \times\times\times, K$, and where

$$\tilde{P}_j^k \Big) \beta, \gamma^k(\beta, \bar{\pi}^k) \Big(= \sum_{l=1}^{J+1} \pi_l^k \tilde{\mathcal{P}}_j^k(\bar{\alpha}_{m_l^k M}, \beta)$$
$$\tilde{\Delta}^k P \Big) \beta, \gamma^k(\beta, \bar{\pi}^k) \Big(= \sum_{l=1}^{J+1} \pi_l^k \tilde{\Delta}\mathcal{P}(\bar{\alpha}_{m_l^k M}, \beta),$$

and $\}m_l^k|_{l=1}^{J+1}$ is a subset of $\}1, \times\times\times, M|$, which can be different for different k.

Secondly, the identified set of the treatment effect on the choice probability can be rewritten as

(1.4.11) $$D^k P = \Big\}\Delta^k P(\phi) : \phi \ / \ \Phi_I \quad ,$$

and its pertaining estimated set can be rewritten as

(1.4.12) $$\hat{D}^k P = \Big\}\tilde{\Delta}^k P \Big)\beta, \gamma^k(\beta, \bar{\pi}^k)\Big(: \phi(\beta, \bar{\pi}) \ / \ \hat{\Phi} \quad ,$$

where $\Delta^k P(\phi) = \sum_{l=1}^{J+1} \Delta\mathcal{P}(\alpha_l^k, \beta)\pi_l^k$ and $\tilde{\Delta}^k P \quad \beta, \gamma^k(\beta, \bar{\pi}^k)[= \sum_{l=1}^{J+1} \tilde{\Delta}\mathcal{P}(\bar{\alpha}_{m_l^k M}, \beta)\pi_l^k$. Except for Assumption 1.4.1, a new assumption is needed

Assumption 1.4.2. *(i) There is $L < \infty$ such that for all $(\alpha, \beta), (\alpha^{\in}, \beta^{\ni}) \ / \ \Upsilon * \mathbb{B}$, $\|\Delta\mathcal{P}(\alpha, \beta) \quad \Delta\mathcal{P}(\alpha^{\in}, \beta^{\ni})\| \leqslant L\,[d(\alpha, \alpha^{\ni}) + \backslash\beta \quad \beta^{\ni}\backslash]$; (ii) $\Delta^k P(\phi)$ is continuous in ϕ.*

Similarly, there are parallel results of Lemma 1.4.5 and 1.4.6, they are stated as follows

Lemma 1.4.9. *There exists a constant C such that for all ϕ, $\phi^{\in} \ / \ \Phi$, the following inequality holds*

$$\Big\|\tilde{\Delta}^k P(\phi) \quad \tilde{\Delta}^k P(\phi^{\ni})\Big\| \leqslant C d(\phi, \phi^{\ni}) + o_P(1),$$

where $o_P(1)$ doesn't depend on $(\phi, \phi^{\ni})$.

Lemma 1.4.10. *Denote the Hausdorff metric as d_H, Lemma 1.4.9 implies*

$$d_H \Big)\tilde{\Delta}^k P(\Phi_s), \tilde{\Delta}^k P(\Phi_s^{\ni})\Big(\leqslant C d_H(\Phi_s, \Phi_s^{\ni}) + o_P(1),$$

where Φ_s and $\Phi_s^{\in}$are two subsets of Φ and $o_P(1)$ doesn't depend on $(\Phi_s, \Phi_s^{\ni})$.

Finally, here comes the key result

Lemma 1.4.11. *If $d_H \quad \hat{\Phi}, \Phi_I[\quad \overset{p}{/} \quad 0$, we have*

$$\hat{\Delta}_l^k P \quad \overset{p}{/} \quad \Delta_l^k P \text{ and } \hat{\Delta}_u^k P \quad \overset{p}{/} \quad \Delta_u^k P.$$

After all, the rest of the job is to check whether the condition $d_H \quad \hat{\Phi}, \Phi_I [\quad \overset{p}{\prime}$ 0 holds or not. Of course the condition itself has its own merit. To check this condition, I would like to show that $\hat{Q}(\phi)$ uniformly converges to $Q(\phi)$ first, and this property holds as is stated by the following lemma

Lemma 1.4.12. $\sup_{\phi / \Phi} \left\|\hat{Q}(\phi) \quad Q(\phi)\right\| \overset{P}{\prime} \; 0.$

The last assumption I need is as follows

Assumption 1.4.3. *(i)* $\eta(M) = \sup_{\alpha / \Upsilon} \min_{\alpha^{\in} / \Upsilon_M} d(\alpha, \alpha^{\ni}) \prime \quad 0$ *as* $M \prime \quad \infty$*; (ii) there is a constant C such that for all* $(\alpha, \beta), (\alpha^{\in}, \beta^{\ni}) \; / \; \Upsilon * \mathbb{B}$, $\left\|\mathcal{P}_j^k(\alpha^{\in}, \beta^{\ni}) \quad \mathcal{P}_j^k(\alpha, \beta)\right\| \leqslant C \, [d(\alpha, \alpha^{\ni}) + \backslash \beta^{\in} \quad \beta \backslash]$*; (iii)* $Q(\phi)$ *is continuous in* Φ*; and (iv) Let* $\xi_n = n^{\kappa_1}$, $\eta(M) = n^{\kappa_2}$, $\lambda_n = n^{\kappa_3}$, $r = n^{\kappa_4}$ *and* $\}\kappa_2 < 0, \; 0 > \kappa_1 > \max\} \quad 1, \kappa_2|, \; \kappa_4 \geqslant \quad 2\,\kappa_2, \; \kappa_3 < \kappa_1| \, .$

Theorem 1.4.1. *Under all the assumptions,* $d_H(\hat{\Phi}, \Phi_I) \overset{P}{\prime} \; 0.$

Finally, I have the convergence of the bounds for conditional average treatment effect on the outcome and the treatment effect of choice probability. I give this result as follows

Theorem 1.4.2. *By all the lemmas and theorems above, we have*

$$\hat{\Delta}_l^k \quad \overset{p}{\prime} \quad \Delta_l^k \text{ and } \hat{\Delta}_u^k \quad \overset{p}{\prime} \quad \Delta_u^k,$$

$$\hat{\Delta}_l^k P \quad \overset{p}{\prime} \quad \Delta_l^k P \text{ and } \hat{\Delta}_u^k P \quad \overset{p}{\prime} \quad \Delta_u^k P.$$

So far I have shown that for any given joint distribution $H(\epsilon)$, the simulation based set estimator for B is consistent, and so are the simulation based estimators for the lower and upper bounds of the conditional treatment partial effect on the outcome and the treatment effect of the choice probability.

1.4.2 Inference on Simulation Based Estimator

As you can see, estimator in this paper is highly nonlinear with simulation. It is impractical to derive its asymptotic distribution analytically. Even if this is possible I conjecture that the asymptotic behavior is not pivotal. The common practice is to use bootstrap to approximate its asymptotic behavior. But the standard bootstrap is not competent in this case as CFHN (2013) claimed. This idea can be briefly described as follows

1. The Data Generating Process, thus DGP, of this model can be completely described by $\mathcal{S} = \Big\}\mathcal{S}_j^k \quad {}_{j=1,\ggg,J \text{ and } k=1,\ggg,K}$, and any parameter can be written as a function of the DGP: $\theta^{\subseteq} = \theta^{\subseteq}(\mathcal{S})$.[24]

[24] Here $\theta^{\subseteq}$ is the generic form of our interested parameters, i.e. the upper and lower bounds for ATE and treatment effect on choice probability.

2. The estimator of $\theta^{\subsetneq}$ is $\hat{\theta} = \theta^{\subsetneq}(\hat{\mathcal{S}})$, and the inference statistics is $S_n = \hat{\theta} \quad \theta^{\subsetneq} = \theta^{\subsetneq}(\hat{\mathcal{S}}) \quad \theta^{\subsetneq}(\mathcal{S})$. It is supposed to estimate S_n's distribution $G_n(s,\mathcal{S})$ by bootstrap, but standard bootstrap actually gets $G_n(s,\hat{\mathcal{S}})$. It works if only $G_n(s,\hat{\mathcal{S}}) \,\text{ı}\quad G_n(s,\mathcal{S})$ as $n\,\text{ı}\quad \infty$ and $\hat{\mathcal{S}} \overset{P}{\text{ı}}\ \mathcal{S}$.

3. Due to the highly-complexity of the estimator in this paper, the limit version of $G_n(s,\mathcal{S})$ is not continuous in $\mathcal{S}$, the standard bootstrap fails to estimate critical value consistently. Thus CHFN (2013) proposed a variation of bootstrap called the *perturbed bootstrap* to give a consistent confidence region for $\theta^{\subsetneq}$.

In this section I explain the perturbed bootstrap method and show how to use it step by step. But before the introduction of the perturbed bootstrap, I deviate to talk about the DGP projection problem under misspecification or sampling error.[25] All the ideas about set identification I presented above work if the model specification is correct. Note that under misspecification it may turn out an empty set of identified β, and this is also possible because of the sampling error even if the model specification is correct. To overcome this problem it is needed to project $\hat{\mathcal{S}}$ into the model space, where the model space is defined as all the choice probabilities that can be generated by the model specification

$$\text{(1.4.13)} \qquad \text{MODEL SPACE } \Xi = \}P : \mathcal{A}\beta \;/\; \mathbb{B} \text{ s.t. } \mathcal{G}_k(\beta,P) \not\vDash \emptyset, \exists k = 1, \times\!\!\times\!\!\times, K| \;.$$

Define the projection of $\hat{\mathcal{S}}$ as

$$\text{(1.4.14)} \qquad P^{\subsetneq}(\hat{\mathcal{S}}) = \arg\min_{P/\Xi} W(P,\hat{\mathcal{S}}),\ \text{where } W(P,\hat{\mathcal{S}}) = n\sum_{j,k}\hat{\mathcal{S}}^k\frac{\Big)\hat{\mathcal{S}}_j^k \quad P_j^k\Big(^2}{P_j^k}.$$

In practice, the projection is done first and $P^{\subsetneq}(\hat{\mathcal{S}})$ is used afterwards instead of $\hat{\mathcal{S}}$, therefore the identified set can be guaranteed nonempty and the projected DGP can be seen as the best approximation of the true DGP in the model space.

The main idea of the perturbed bootstrap is to construct a confidence region of $\mathcal{S}$ and use each P in the confidence region as a perturbation of $\mathcal{S}$. For every such P, a standard bootstrap is used to construct a confidence region, and out of all the confidence regions their convex hull is constructed as the conservative confidence region and used for consistent inference.

First, the 1 $\quad\gamma$ confidence region for $\mathcal{S}$ is

$$\text{(1.4.15)} \qquad CR_{1\quad\gamma}(\mathcal{S}) = \Big\}P \;/\; S_J^K : W(P,\hat{\mathcal{S}}) \leqslant c_{1\quad\gamma}\Big)\chi^2_{K(J\quad 1)}\Big(\;,$$

[25] See section 9 of Chernozhukov et al. 2013a for more details and proofs, this part is a representation of their idea in the way that makes the notations consistent with the rest of our paper.

where $c_{1\ \ \gamma}\big)\chi^2_{K(J\ \ 1)}\big($ is the $(1\quad\gamma)$-quantile of the $\chi^2_{K(J\ \ 1)}$ distribution and W is the goodness-of-fit statistic

$$W(P,\hat{\mathcal{S}}) = n\sum_{j,k}\hat{\mathcal{S}}^k\frac{\Big)\hat{\mathcal{S}}^k_j\quad P^k_j\Big(^2}{P^k_j}. \tag{1.4.16}$$

Secondly, define the estimates of the lower and upper bounds on the quantiles of $G_n(s,\mathcal{S})$ as follows

$$\underline{G}_n{}^{1}(\alpha,\mathcal{S})/\overline{G}_n{}^{1}(\alpha,\mathcal{S}) = \underset{P/CR_{1\ \ \gamma}(\mathcal{S})}{\inf / \sup}\ G_n{}^{1}(\alpha,P), \tag{1.4.17}$$

where $G_n{}^{1}(\alpha,P) = \inf\}s : G_n(s,P) \geqslant \alpha|$ is the α-quantile of the distribution $G_n(s,P)$. Finally construct a $(1\quad\alpha\quad\gamma)\%$ confidence region for $\theta^{\subseteq}$as follows

$$CR_{1\ \ \alpha\ \ \gamma}(\theta^{\subseteq}) =]\underline{\theta},\overline{\theta}\{, \tag{1.4.18}$$

where for $\alpha = \alpha_1 + \alpha_2$, $\underline{\theta} = \hat{\theta}\quad\overline{G}_n{}^{1}(1\quad\alpha_2,\mathcal{S})$, $\overline{\theta} = \hat{\theta}\quad\underline{G}_n{}^{1}(\alpha_1,\mathcal{S})$.

Since the idea of set identification only depends on information $\mathcal{S}^k_j$ but not $\mathcal{S}^k$, so it is acceptable to treat $\mathcal{S}^k = \hat{\mathcal{S}}^k$ in the analysis and practice. The following assumption about the data generating process from CFNH (2013) is needed.

Assumption 1.4.4. $\Pi\ /\ \mathbb{P} = \Big\}(\mathcal{S}^k,\mathcal{S}^k_j) : \mathcal{S}^k > \varepsilon, \mathcal{S}^k_j > \varepsilon; j = 1, \times\times\times, J, k = 1, \times\times\times, K$ *for some* $\varepsilon > 0$.

Theorem 11 in CFHN (2013) showed that perturbed bootstrap delivers (uniformly) valid inference on the parameter of interest and they also gave an algorithm for practicing the perturbed bootstrap.

1.5 A Numeric Example

To study the behavior of the simulation based estimation method let's consider a binary choice example in this section. The reason to choose a binary choice model instead is that the algorithm in the practice is not efficient enough for too many choices. Further more since I develop the binary choice example exactly in terms of the multinomial discrete choice framework, it won't hamper the demonstration of the idea and method of general models.

Specifically, I consider an artificial car market. In this market, there are only two models of cars. Different models are characterized by the number of seats, i.e. $NS\ /\ \}2,4|$ where consumers have cars with two seats and cars with four seats. Different cars have different prices, but for simplicity I assume that price is completely determined by car's model in our data generating process[26], where consumers have two levels of prices, $PR \ / \ \}10, 15|$. The price for cars with 2 seats is 10 thousand dollars and price for cars with 4 seats is 15 thousand dollars. For cars, there are two observable attributes, NS and PR, where NS is the major common attributes of cars such that I use NS to define the two alternatives in the car market, while PR is the minor common attributes that are not used to define alternatives. Denote cars with 2 seats as a_1 and cars with 4 seats as a_2. On the other hand, consumers in the car market are families. They have two observable attributes: the number of family members, NF, and family income, FI. For values of NF, I assume that there are only single member families with $NF = 1$, couples without child $NF = 2$ and families with one child $NF = 3$. For family income FI, there exist only two levels, they are $FI = 10$ and $FI = 20$. By the combination of NF and FI there are 6 types of families each year, and for a panel with T years, there are 6^T types of families. If $T = 2$, there are only 36 types. To reduce the calculation burden, we draw 4 out of all the 36 types and make replications[27] for each type. By chance the four types of families in the generated sample are

1. $NF_{t_1} = 1, FI_{t_1} = 10, NF_{t_2} = 2, FI_{t_2} = 20,$
2. $NF_{t_1} = 2, FI_{t_1} = 10, NF_{t_2} = 3, FI_{t_2} = 20,$
3. $NF_{t_1} = 2, FI_{t_1} = 10, NF_{t_2} = 1, FI_{t_2} = 20,$
4. $NF_{t_1} = 2, FI_{t_1} = 10, NF_{t_2} = 2, FI_{t_2} = 20.$

Out of the four types above, only the treatment effect of a family income change from 10 to 20 for the 4th type is identified since the family income 10 and 20 appear in different periods while other attributes keep constant.

Before I introduce all the treatment effects I am going to study, I continue to finish the data generate process. For each family the utility of different choice of car is given by

$$U_{ita} = \beta_{1it} NS_a + \beta_{2it} PR_{ta} + \epsilon_{ita}, \tag{1.5.1}$$

where β_{1it} and β_{2it} are random coefficients that change over i and t, they are modeled by

$$\begin{aligned} \beta_{1it} &= \beta_1 NF_{it} \\ \beta_{2it} &= \beta_2 \alpha_i / FI_{it}, \end{aligned} \tag{1.5.2}$$

[26] Generally we may allow price to change over periods since we do not use PR as a major common attribute to define alternatives. See the memo from Huang 2014 for details.

[27] In the simulation application, I have 2000 families for each type such that $n = 8000$.

where α_i is the fixed effects that measure consumer's sensitivity to the price in terms of the family income. For the true DGP I set $\beta_1 = 1$ and $\beta_2 = \quad 10$. Substitute (1.5.2) into (1.5.1) generates

$$U_{ita} = \beta_1 NF_{it} NS_a + \beta_2 \alpha_i PR_{ta} / FI_{it} + \epsilon_{ita}, \tag{1.5.3}$$

where the error term $\epsilon_{it} = \} \epsilon_{ita} |_{\ a=a_1,a_2}$ allows dependence between alternatives and serial dependence. Specifically,

$$\epsilon_{it} = \rho \epsilon_{it \quad 1} + \mu_{it}, \tag{1.5.4}$$

where μ_{it} follows a given two dimensional normal distribution and μ_{ita_1} and μ_{ita_2} are correlated with each other. ρ has a absolute value smaller than unity therefore (1.5.4) gives a stationary DGP for ϵ_{it} that allows dependence between alternatives and serial correlation. Note that giving (1.5.4) and its related parameters, distribution of ϵ_i is completely determined thus the normalization of the utility (1.5.3) has been done simultaneously.

For the true DGP of the fixed effects α_i, I consider two cases. In the first case, it is generated by a uniform distribution over the interval $[0.6, 1]$ which is independent of all the RHS variables in (1.5.3). For the second case, it is dependent with households' income per capital in the first period, it is uniformly distributed over interval $[0.8, 1]$ for households with income per capital less than 10 in the first period and uniformly distributed over interval $[0.6, 0.8]$ for other households. For each individual at any period the generated binary choice outcome is either $y_{it} = a_1$ or $y_{it} = a_2$ which gives the highest utility.

Given the generated data at hand, a series of interesting questions about treatment effects on the probability of choosing a larger car can be asked. For example, it can be asked what are the changes in the probabilities of choosing a larger car if the number of family members increases by one person from 1 to 2 or from 2 to 3 or even from 3 to 4 for each subgroups. Note that we can ask questions where the treatment has never happened in the data over periods, but we can't ask questions for subgroups that hasn't been observed in the data. For example, partial effect of increasing one family member for subgroup with $NF_{t_1} = 1, FI_{t_1} = 10, NF_{t_2} = 2, FI_{t_2} = 20$ can be answered while partial effect of increasing one family member for subgroup with $NF_{t_1} = 1, FI_{t_1} = 20, NF_{t_2} = 2, FI_{t_2} = 20$ can not be answered since the later subgroup doesn't exist in the data.

We may also be interested in treatment effects caused by changes in family income or car's retail price or any combined effects caused by changes in family members, family income and car price. In this numeric example, I won't cover all of them but name a few. Specifically, I will study all the following treatment effects:

1. For the first subgroup
    (a) Partial effect on choosing a larger car if only the number of family members increases from 2 to 3 while the family income is 20.
    (b) Partial effect on choosing a larger car if only the family income increases from 10 to 20 while the number of family members is 1.
    (c) Partial effect on choosing a larger car if only the price for two seats cars increases from 10 to 12 while number of family member is 1 and family income is 10.
2. For the second subgroup
    (a) Partial effect on choosing a larger car if only the number of family members increases from 2 to 3 while the family income is 20.
    (b) Partial effect on choosing a larger car if only the family income increases from 10 to 20 while the number of family members is 2.
    (c) Partial effect on choosing a larger car if only the price for two seats cars increases from 10 to 12 while number of family member is 2 and family income is 10.
3. For the third subgroup
    (a) Partial effect on choosing a larger car if only the number of family members decreases from 2 to 1 while the family income is 10.
    (b) Partial effect on choosing a larger car if only the family income increases from 10 to 20 while the number of family members is 2.
    (c) Partial effect on choosing a larger car if only the price for two seats cars increases from 10 to 12 while number of family member is 2 and family income is 10.
4. For the fourth subgroup
    (a) Partial effect on choosing a larger car if only the number of family members increases from 2 to 3 while the family income is 20.
    (b) Partial effect on choosing a larger car if only the family income increases from 10 to 20 while the number of family members is 2.
    (c) Partial effect on choosing a larger car if only the price for two seats cars increases from 10 to 12 while number of family member is 2 and family income is 10.

Out of all the treatment effects listed above, only the second treatment effect for the fourth subgroup can be identified, while all other partial effects are only set identified. Since I know the true DGP in this example, I can simulate all the partial effects and compare them with these set identification estimations. As the estimator is simulation based, it contains variations from simulation as well as variations from the randomness of the data. Without considering estimator's variance, the true effect are not guaranteed to be inside the estimated intervals. Instead of reporting the result from only one estimation, I run the data generation and estimation 100 times and report the averages. This also gives the empirical distribution of them. I illuminate these distributions by a set of boxplots starting from figure 1.1 to figure 1.4. In those boxplots, three horizontal lines are added to facilitate the comparison of upper bound, lower bound and the true effects, where a purple line indicates the mean of upper bounds, a red line indicates the mean of true effects and a blue line indicates the mean of lower bounds.

I also summarize findings under different DGP in from table 1.1 to table 1.4. Out of all the bounds for 12 effects, most of the effects are included in the bounds while others are quite close to the bounds. Except for the finite sample properties, another reason for the non-inclusion is because of the challenge of searching for all the conditional distribution of the fixed effects and β in the identified set. In practice I use a simpler search that can be implemented by linear programmings which were also used by Chernozhukov et al. 2013a. Theoretically these bounds I report here are actually subsets of their identified sets respectively. Of course because of the simulations, my algorithm costs more time for the computation[28] than CFHN's algorithm without simulation.

1.6 Conclusion

This paper generalizes the estimation approach for a set identified semiparametric discrete choice model with fixed effect proposed by Chernozhukov et al. 2013a. The idea of this estimation method is to use distribution assumption of the error term together with the simulation of individual's behavior to overcome the disadvantage of CFHN's, which is incapable of handling models with components having no closed form expressions, such that this simulation based method can be applied to a wider range of applications, i.e. a general multi-choice model allowing error term's serial dependence. One important feature of this method inherited from CFHN is that it allows multi-dimensional fixed effects in the rep-

[28] It takes more than one hour to finish the data generation and treatment effects bounds estimation on a 2011 intel I5 desktop computer, and my code is written in R and C++. C++ code is to implement the simulation and improve the efficiency.

resentative utility, and it could be non-separable with other covariates. But I do assume that the error term should be additively separable from other components in the representative utility and also be independent of them.

I also note that the advantage comes with extra cost. For example, we need more computation resource to simulate the choice probability as well as any other components which have no closed form expressions either. For quantities like ξ_n, η, λ_n and r, their rates of convergence should depend on each others' and they are controlled by assumption 1.4.3 (iv). Perturbed bootstrap can be used to give valid inference on parameters of interest. In the last section a numeric example based on the simulated data is given, and I give estimator's distribution by simulation instead of perturbed bootstrap[29]. It can be seen the estimators for bounds include the real treatment effects and for this specific example these bounds are quite narrow.

All the models I mentioned in this paper, including binary logit and probit with fixed effects, assume that the distribution of ϵ_i is known completely. I conjecture it is possible to joint identify parameters β and some parameters from $H(\epsilon)$ under some extra assumptions. Of course challenges come after this idea, I will study this case in another paper.

[29] Since this is a numeric example, data are generated such that the simulation is possible and preferred than bootstrap. For application with field data, perturbed bootstrap is your friend.

1.7 Tables and Figures

Table 1.1: Partial effects on choosing a larger car (independendt α_i with serial correlation)

Group 1	Effect a		Effect b		Effect c	
Upper bound	0.05927		0.42861		0.30907	
Lower bound	0.05808		0.41071		0.28743	
True effect	0.05909*	(95)	0.42553*	(97)	0.30485*	(95)
Group 2	**Effect a**		**Effect b**		**Effect c**	
Upper bound	0.05896		0.44602		0.39802	
Lower bound	0.05296		0.43922		0.38660	
True effect	0.05909	(95)	0.44037*	(97)	0.39088*	(97)
Group 3	**Effect a**		**Effect b**		**Effect c**	
Upper bound	-0.42127		0.44811		0.39946	
Lower bound	-0.43307		0.44035		0.38926	
True effect	-0.42556*	(100)	0.44037*	(97)	0.39088*	(97)
Group 4	**Effect a**		**Effect b**		**Effect c**	
Upper bound	0.05851		0.44233		0.39137	
Lower bound	0.05768		0.43997		0.38701	
True effect	0.05909	(95)	0.44037*	(97)	0.39088*	(96)

[1] * indicates the value is included by our estimated upper and lower bounds, where all values are the averages over the 100 simulations.

[2] α_i is independent with households' attributes and follows an uniform distribution over $[0.6, 1]$.

[3] $\rho = 0.7$ and $cov(\mu_{ita_1}, \mu_{ita_2}) = 0.6$, $var(\mu_{ita_1}) = var(\mu_{ita_2}) = 1$.

[4] Effects a, b and c are defined in previous paragraphs, and they may be different for different groups.

[5] Numbers in parentheses are frequencies of containing the true effect in the 95% confidence intervals out of 100 simulations.

Table 1.2: Partial effects on choosing a larger car (independent α_i without serial correlation)

Group 1	Effect a		Effect b		Effect c	
Upper bound	0.01686		0.48451		0.32556	
Lower bound	0.01632		0.45558		0.29300	
True effect	0.01658*	(100)	0.46981*	(99)	0.30879*	(98)
Group 2	**Effect a**		**Effect b**		**Effect c**	
Upper bound	0.02293		0.48710		0.46373	
Lower bound	0.01142		0.47454		0.43240	
True effect	0.01658*	(99)	0.48355*	(97)	0.45235*	(100)
Group 3	**Effect a**		**Effect b**		**Effect c**	
Upper bound	-0.46560		0.48502		0.45540	
Lower bound	-0.47298		0.48086		0.44870	
True effect	-0.46988*	(94)	0.48355*	(96)	0.45235*	(97)
Group 4	**Effect a**		**Effect b**		**Effect c**	
Upper bound	0.01636		0.48441		0.45449	
Lower bound	0.01557		0.48149		0.45087	
True effect	0.01658	(95)	0.48355*	(96)	0.45235*	(97)

[1] * indicates the value is included by our estimated upper and lower bounds, where all values are the averages over the 100 simulations.

[2] α_i is independent with households' attributes and follows an uniform distribution over $[0.6, 1]$.

[3] $\rho = 0.0$ and $cov(\mu_{ita_1}, \mu_{ita_2}) = 0.6$, $var(\mu_{ita_1}) = var(\mu_{ita_2}) = 1$.

[4] Effects a, b and c are defined in previous paragraphs, and they may be different for different groups.

[5] Numbers in parentheses are frequencies of containing the true effect in the 95% confidence intervals out of 100 simulations.

Table 1.3: Partial effects on choosing a larger car (dependent α_i with serial correlation)

Group 1	Effect a		Effect b		Effect c	
Upper bound	0.03783		0.45568		0.34519	
Lower bound	0.03706		0.44772		0.33519	
True effect	0.03693	(94)	0.45668	(97)	0.34660	(96)
Group 2	Effect a		Effect b		Effect c	
Upper bound	0.08800		0.56930		0.49611	
Lower bound	0.08092		0.56168		0.48143	
True effect	0.08120*	(96)	0.56851*	(95)	0.49906	(94)
Group 3	Effect a		Effect b		Effect c	
Upper bound	-0.31739		0.57670		0.50822	
Lower bound	-0.32351		0.56853		0.49753	
True effect	-0.32283*	(96)	0.56851	(98)	0.49906*	(97)
Group 4	Effect a		Effect b		Effect c	
Upper bound	0.08187		0.56685		0.49406	
Lower bound	0.08103		0.56481		0.48916	
True effect	0.08120*	(96)	0.56851	(95)	0.49906	(93)

[1] * indicates the value is included by our estimated upper and lower bounds, where all values are the averages over the 100 simulations.

[2] For households with income per capital in the first period less than 10, α_i follows an uniform distribution over $[0.8, 1]$, and For households with income per capital in the first period more than or equal to 10, α_i follows an uniform distribution over $[0.6, 0.8]$.

[3] $\rho = 0.7$ and $cov(\mu_{ita_1}, \mu_{ita_2}) = 0.6$, $var(\mu_{ita_1}) = var(\mu_{ita_2}) = 1$.

[4] Effects a, b and c are defined in previous paragraphs, and they may be different for different groups.

[5] Numbers in parentheses are frequencies of containing the true effect in the 95% confidence intervals out of 100 simulations.

Table 1.4: Partial effects on choosing a larger car (dependent α_i without serial correlation)

Group 1	Effect a		Effect b		Effect c	
Upper bound	0.00673		0.56155		0.40990	
Lower bound	0.00649		0.55005		0.39632	
True effect	0.00649*	(98)	0.55308*	(97)	0.40084*	(98)
Group 2	**Effect a**		**Effect b**		**Effect c**	
Upper bound	0.03722		0.67936		0.63342	
Lower bound	0.02331		0.66478		0.59605	
True effect	0.02678*	(98)	0.67551*	(97)	0.62532*	(99)
Group 3	**Effect a**		**Effect b**		**Effect c**	
Upper bound	-0.29060		0.67953		0.63156	
Lower bound	-0.29249		0.67687		0.62691	
True effect	-0.29381	(96)	0.67551	(95)	0.62532	(95)
Group 4	**Effect a**		**Effect b**		**Effect c**	
Upper bound	0.02694		0.67717		0.62637	
Lower bound	0.02632		0.67521		0.62380	
True effect	0.02677*	(98)	0.67551*	(96)	0.62532*	(97)

[1] * indicates the value is included by our estimated upper and lower bounds, where all values are the averages over the 100 simulations.

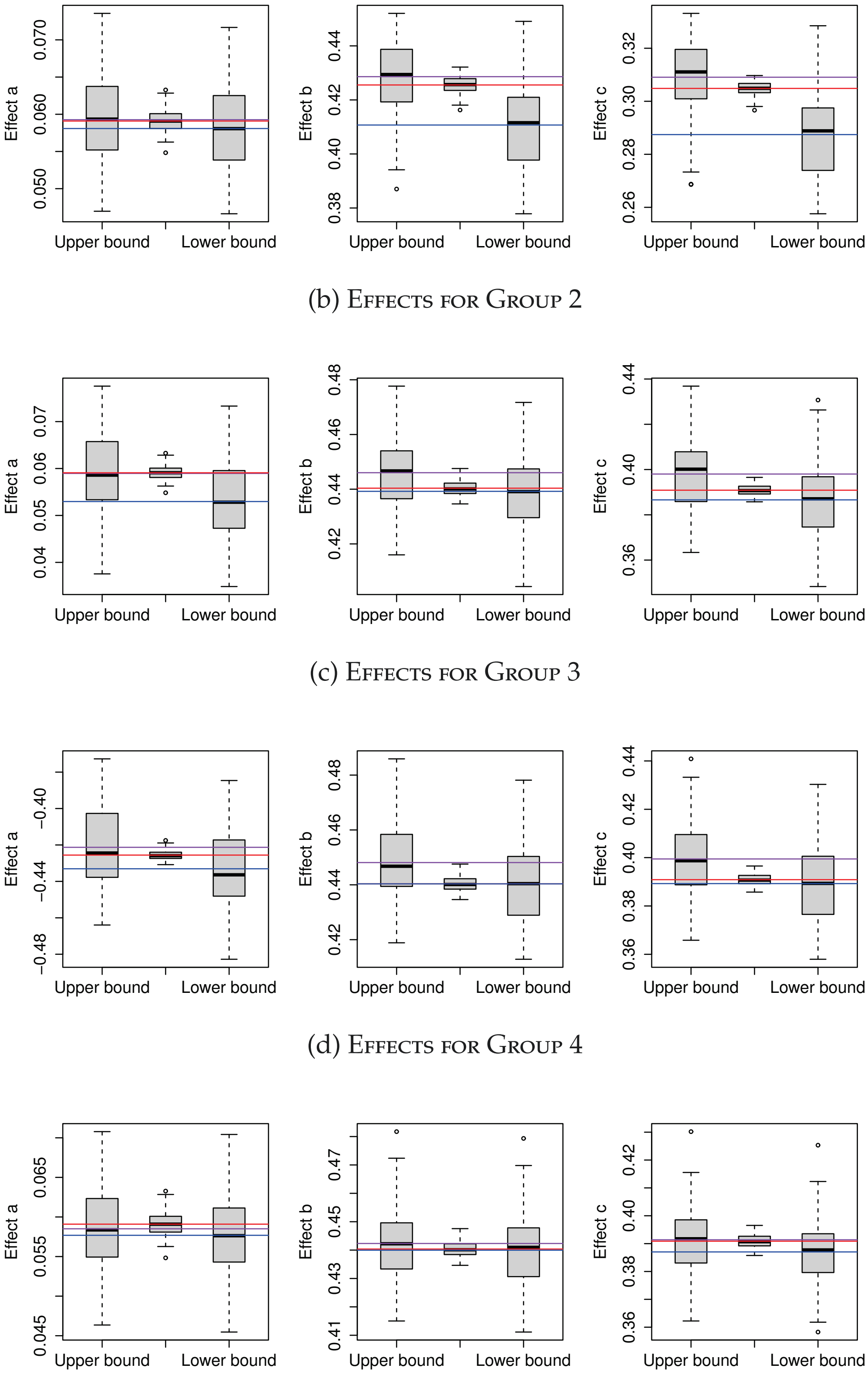
[2] For households with income per capital in the first period less than 10, α_i follows an uniform distribution over $[0.8, 1]$, and For households with income per capital in the first period more than or equal to 10, α_i follows an uniform distribution over $[0.6, 0.8]$.

[3] $\rho = 0.0$ and $cov(\mu_{ita_1}, \mu_{ita_2}) = 0.6$, $var(\mu_{ita_1}) = var(\mu_{ita_2}) = 1$.

[4] Effects a, b and c are defined in previous paragraphs, and they may be different for different groups.

[5] Numbers in parentheses are frequencies of containing the true effect in the 95% confidence intervals out of 100 simulations.

Figure 1.1: Effects on Choosing a Larger Car (indepentent α_i with serial correlation)



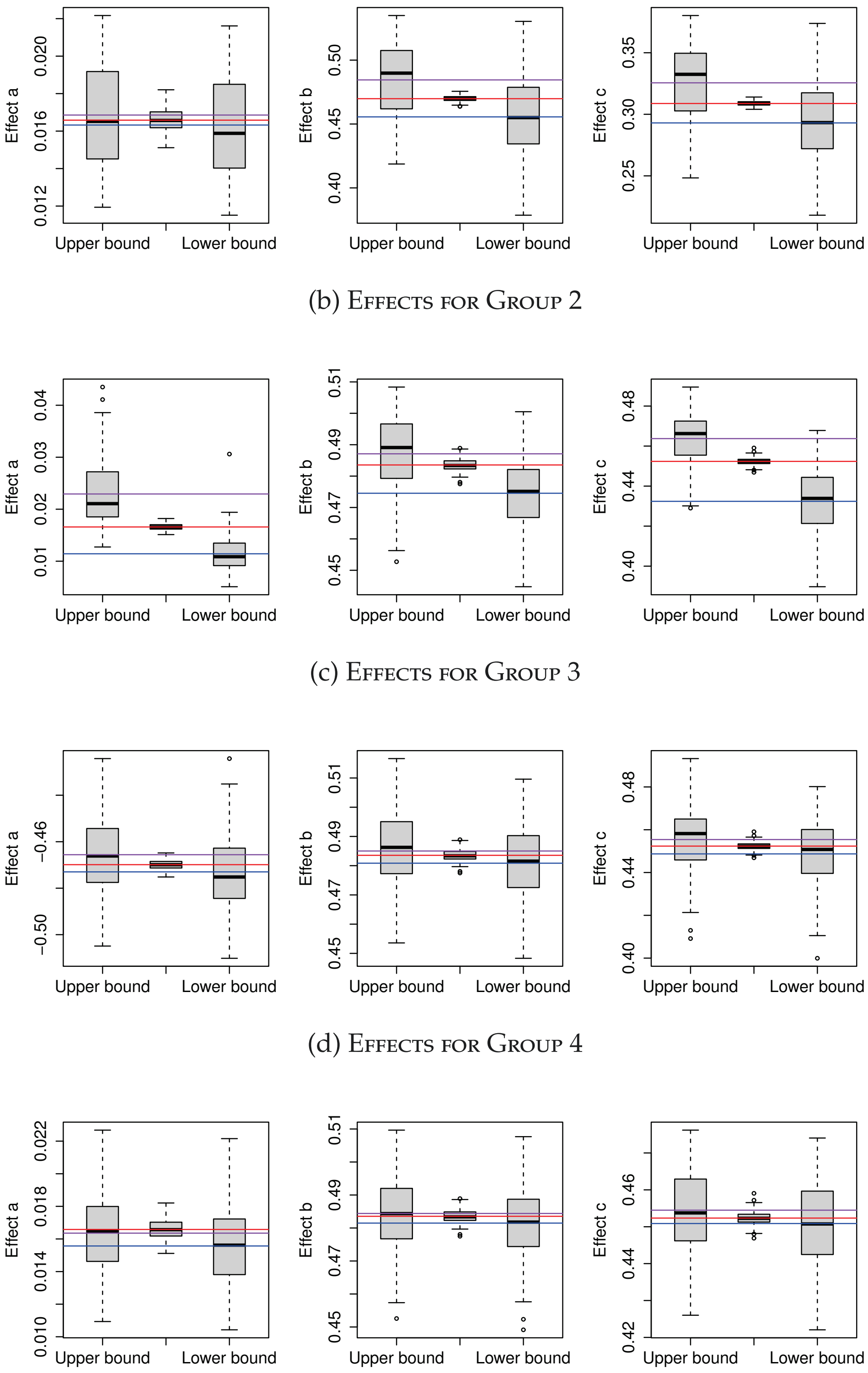
(a) Effects for Group 1

(b) Effects for Group 2

(c) Effects for Group 3

(d) Effects for Group 4

Figure 1.2: Effects on Choosing a Larger Car (independent α_i without serial correlation)



(a) Effects for Group 1

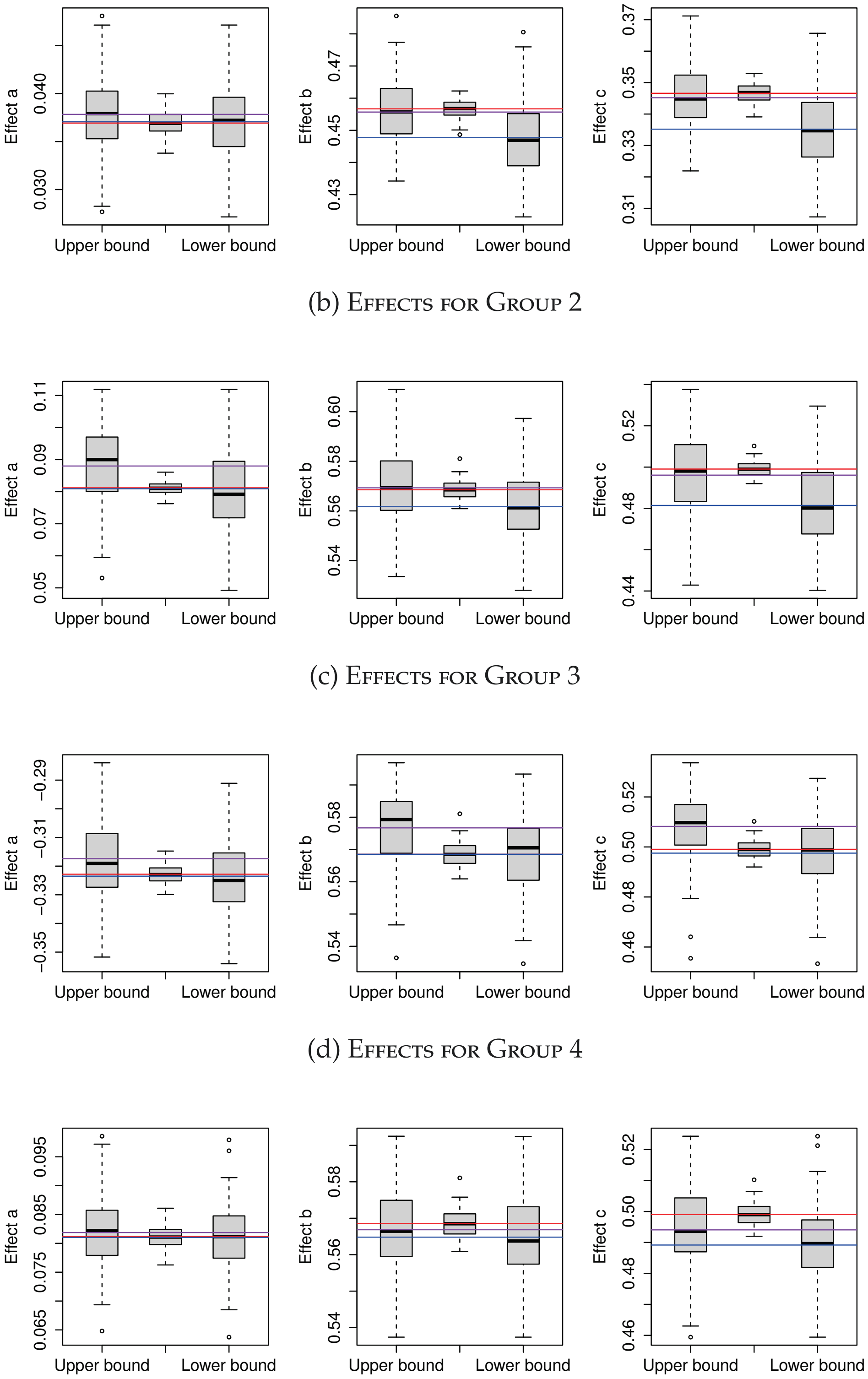
(b) Effects for Group 2

(c) Effects for Group 3

(d) Effects for Group 4

Figure 1.3: Effects on Choosing a Larger Car (depentent α_i with serial correlation)

(a) Effects for Group 1

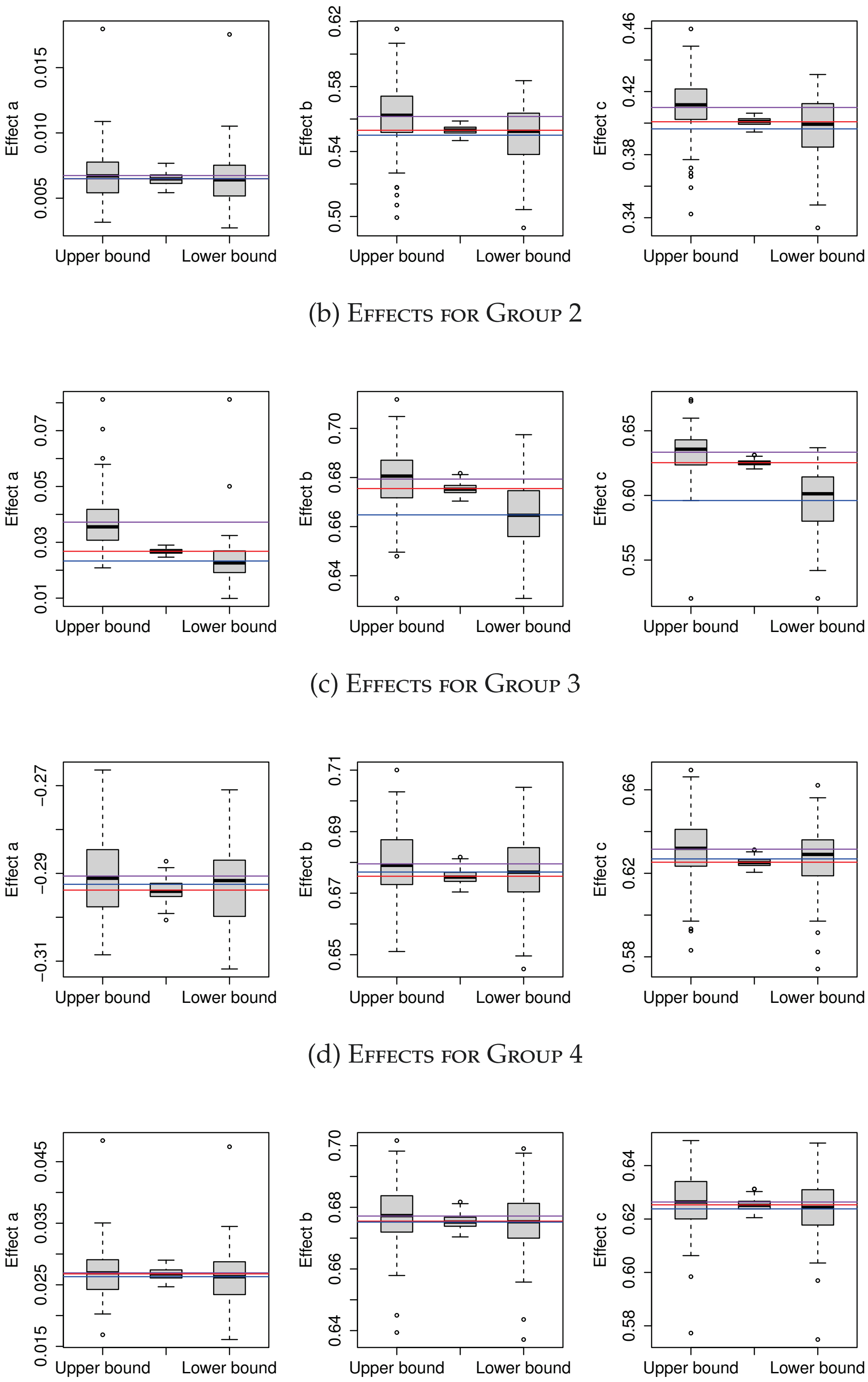


(b) Effects for Group 2

(c) Effects for Group 3

(d) Effects for Group 4

Figure 1.4: Effects on Choosing a Larger Car (depentent α_i without serial correlation)



(a) Effects for Group 1

(b) Effects for Group 2

(c) Effects for Group 3

(d) Effects for Group 4

Chapter 2

Set Identified Dynamic Multinomial Discrete Choice Model with Fixed Effects and Simulation

Abstract

The idea of set identification by matching choice probabilities is general for discrete choice models with fixed effects in the sense that it can be easily applied to both static and dynamic models if all the model components have closed form expressions. Huang 2015b considered the set identification and estimation of fixed effects models without close form probabilities, such as multinomial discrete choice models and nonlinear models with serially correlated errors. However, it was assumed that all the explanatory variables are strictly exogenous, ruling out dynamic models. In this paper I prove that Huang 2015b's simulated estimator is still consistent in dynamic models should we give up the serial correlation in between error terms.

2.1 Dynamic multinomial discrete choice model with fixed effects

In this paper I consider the following dynamic discrete choice model. In each period t, there is an alternative a chosen by individual i out of an alternative set $\mathcal{F}$ which is common over t. Assume that there are some strictly exogenous variables[1] which present individual's characteristics and alternative's attributes of each period, and they are denoted as x_{ita}. There are other dynamic state variables w_{it}, whose values are determined in a deterministic way as $w_{it+1} = f(w_{it}, y_{it})$. The observed data including choice behavior y_{it} may start after the inception of the dynamic process, as such choices in earlier periods is missing but their accumulated result can be summarized by the dynamic state variables in the first period of observation w_{i1}. I denote w_{i1} and those strictly exogenous variables x_{ita} together as z_{ita}. The observed choices prior to t, i.e. $\}y_{i1}, \times\times\times, y_{it\ \ 1}|$, is denoted as $y_{i\ \ t}$, where $t = 2, \times\times\times, T$. There exists fixed effects $\alpha_i = \}\alpha_{ia}|_{\ a/\mathcal{F}}$[2] in the representative utility

$$V_{ita} = V(z_{ita}, y_{i\ \ t}, \alpha_{ia}, \beta), \text{ for } t = 2, \times\times\times, T$$
$$V_{ita} = V(z_{ita}, \alpha_{ia}, \beta), \text{ for } t = 1, \tag{2.1.1}$$

where β are parameters and function V is fully parameterized. Note that z_{ita} includes w_{i1}, and w_{it} evolves according to f, therefore z_{ita} together with $y_{i\ \ t}$ implies w_{it}. Which means (2.1.1) is equivalent to

$$V_{ita} = V(x_{ita}, w_{it}, \alpha_{ia}, \beta), \text{ for } t = 1, \times\times\times, T. \tag{2.1.2}$$

In each period there exists individual and alternative specific shocks ϵ_{ita} to form the utility

$$U_{ita} = V(z_{ita}, y_{i\ \ t}, \alpha_{ia}, \beta) + \epsilon_{ita}, \text{ for } t = 2, \times\times\times, T$$
$$U_{ita} = V(z_{ita}, \alpha_{ia}, \beta) + \epsilon_{ita}, \text{ for } t = 1 \tag{2.1.3}$$

Individuals are assumed to choose the best alternative in each period accord-

[1] These variables are strictly exogenous in the sense that they are independent of ϵ_{it} for every period t.

[2] Notice that the situation where the fixed effects vary across individuals but not across alternatives is a particular case of this.

ing to the utility function above

$$\begin{aligned} y_{it} &= \arg\max_{a/\mathcal{F}} \}V(z_{ita}, y_{i\quad t}, \alpha_{ia}, \beta) + \epsilon_{ita}| \\ &= g_0(z_{it}, y_{i\quad t}, \alpha_i, \epsilon_{it}, \beta),\ (i = 1, \times\times\times, n),\ \text{for } t = 2, \times\times\times, T \\ y_{it} &= \arg\max_{a/\mathcal{F}} \}V(z_{ita}, \alpha_{ia}, \beta) + \epsilon_{ita}| \\ &= g_0(z_{it}, \alpha_i, \epsilon_{it}, \beta),\ (i = 1, \times\times\times, n),\ \text{for } t = 1. \end{aligned} \tag{2.1.4}$$

$\epsilon_{it} = \}\epsilon_{ita}|_{a/\mathcal{F}}$ is assumed to be independent of the fixed effects α_i, $z_i = \}z_{ita}|_{\substack{a/\mathcal{F} \\ t=1,\times\times\times,T}}$ and $\epsilon_{i\quad t}$, where $\epsilon_{i\quad t}$ means all the errors prior to t, and its distribution is known as

$$\epsilon_{it}\|\alpha_i, z_i, \epsilon_{i\quad t} \subset H_t(\epsilon). \tag{2.1.5}$$

This assumption also implies ϵ_{it} is independent of $y_{i\quad t}$. Note ϵ_{it} is independent of z_i, this is because $z_i = \}x_{ita}, w_{i1}|_{\substack{a/\mathcal{F} \\ t=1,\times\times\times,T}}$, where x_{ita} are strictly exogenous variables and w_{i1} are dynamic state variables of the very first period of observation[3].

The fixed effects α_i is allowed to be freely correlated with z_i, and I do not impose any restrictions about their dependence. α_i is also correlated with y_{it}, and their correlation has been modeled by the rule of dynamic behavior as shown in (2.1.4), but with the fixed effects conditional on z_i there is no initial conditions problem.

2.1.1 Interested objects

In this paper I am interested in parameters β and some partial effects of the choice probability, i.e. $\Delta^k P$ defined later in the following paragraphs.

In each period t, let's denote the state as $s_{it} = \}z_{it}, y_{i\quad t}|$. For analysis of partial effects on choice probability, it is useful to consider the following potential conditional probabilities

$$\mathcal{P}^{s_t}_{a_t}(\alpha, \beta) = Pr(y_{it} = a_t\|(s_{it} = s_t), z_i = \cup, \alpha_i = \alpha). \tag{2.1.6}$$

The expression inside the parenthesis in the conditional part is to indicate this is a potential or counterfactual choice probability for individuals with $z_i = \cup$ and unobservable fixed effects $\alpha_i = \alpha$ behaving as if they were with state s_t. It is helpful

[3] $\epsilon_{it^{\in}}$ could not be independent of w_{it} whenever $t^{\in} < t$, since $w_{it} = f(w_{it\quad 1}, y_{it\quad 1})$.

to have it formulated as follows

(2.1.7)

$$\begin{aligned}\mathcal{P}^{s_t}_{a_t}(\alpha,\beta) &= [\ \mathbb{1}\,(g_0(s_t,\alpha,\epsilon_{it},\beta) = a_t)\,dH_t(\epsilon)\\ &= [\ \prod_{a_t^{\in}\not\forall a_t} \mathbb{1}\,\Big)V(s_{ta_t},\alpha_{a_t},\beta)+\epsilon_{ita_t} \geqslant V(s_{ta_t^{\in}},\alpha_{a_t^{\in}},\beta)+\epsilon_{ita_t^{\in}}\Big(\,dH_t(\epsilon).\end{aligned}$$

The first equation holds because ϵ_{it} is independent of z_i and α_i.

Under further assumption such as z_i are finite discrete variables, as such $\cup$ takes finite values out of the set $\}\cup^1, \times\times\times, \cup^K|$. Then I have the following potential probabilities conditional on z_i only,

$$Pr(y_{it} = a_t\|(s_{it} = s_t), z_i = \cup^k) = [\ \mathcal{P}^{s_t}_{a_t}(\alpha,\beta)dF_k(\alpha), \tag{2.1.8}$$

where F_k is α_i's distribution conditional on $z_i = \cup^k$.

Finally it turns out the partial effects on choice probability could be written as

$$\Delta^k P_t = Pr(y_{it} = a_t\|(s_{it} = s_t^a), z_i = \cup^k) \quad Pr(y_{it} = a_t\|(s_{it} = s_t^b), z_i = \cup^k), \tag{2.1.9}$$

where s_t^a and s_t^b are states after and before the treatment. I also define

$$\Delta\mathcal{P}_t(\alpha,\beta) = \mathcal{P}^{s_t^a}_{a_t}(\alpha,\beta) \quad \mathcal{P}^{s_t^b}_{a_t}(\alpha,\beta), \tag{2.1.10}$$

therefore (2.1.9) can be also expressed as

$$\Delta^k P_t = [\ \Delta\mathcal{P}_t(\alpha,\beta)dF_k(\alpha). \tag{2.1.11}$$

2.2 Model Identification

I use the same idea of set identification as Chernozhukov et al. 2013a. This is to use the matching of model implied choice probabilities with true choice probabilities to get all the possible β and F_k.

Denote $y_i = \}y_{it}|_{t=1,\times\times\times,T}$. Since the alternative set $\mathcal{F}$ is finite, all the possible values of y_i can be listed as a set $\}\mathcal{Z}^1, \times\times\times, \mathcal{Z}^J|$. I first introduce a conditional probability as follows

$$\mathcal{P}^k_j(\alpha,\beta) = Pr(y_i = \mathcal{Z}^j\|z_i = \cup^k, \alpha_i = \alpha), \tag{2.2.1}$$

where $\mathcal{Z}^j$ is a general value of y_i and it looks like $\mathcal{Z}^j = \}a_t|_{t=1,\times\times\times,T}$. (2.2.1) can be expressed as

$$
\begin{aligned}
\mathcal{P}_j^k(\alpha,\beta) &= Pr\Big) y_{iT} = a_T, \times\times\times, y_{i1} = a_1 \| z_i = \cup^k, \alpha_i = \alpha\Big(\\
&= Pr\Big) y_{iT} = a_T \| y_{i\quad T} = a\quad_T, z_i = \cup^k, \alpha_i = \alpha\Big(* \\
&\quad \times\times\times \\
&\quad Pr\Big) y_{it} = a_t \| y_{i\quad t} = a\quad_t, z_i = \cup^k, \alpha_i = \alpha\Big(* \\
&\quad \times\times\times \\
&\quad Pr\Big) y_{i1} = a_1 \| z_i = \cup^k, \alpha_i = \alpha\Big(\\
&= \Big[\ \mathbb{1}\Big) g_0(\cup_T^k, y_{i\quad T} = a\quad_T, \alpha, \epsilon_{iT}, \beta) = a_T\Big(dH_T(\epsilon) * \\
&\quad \times\times\times \\
&\quad \Big[\ \mathbb{1}\Big) g_0(\cup_t^k, y_{i\quad t} = a\quad_t, \alpha, \epsilon_{it}, \beta) = a_t\Big(dH_t(\epsilon) * \\
&\quad \times\times\times \\
&\quad \Big[\ \mathbb{1}\Big) g_0(\cup_1^k, \alpha, \epsilon_{i1}, \beta) = a_1\Big(dH_1(\epsilon),
\end{aligned}
\tag{2.2.2}
$$

where integrations hold because ϵ_{it} is assumed to be independent of z_i, $y_{i\quad t}$ and α_i. Also note that

(2.2.3)

$$
\begin{aligned}
&\mathbb{1}\Big) g_0(\cup_t^k, y_{i\quad t} = a\quad_t, \alpha, \epsilon_{it}, \beta) = a_t\Big(= \\
&\prod_{a_t^{\in}\forall\!\!= a_t} \mathbb{1}\Big) V(\cup_{ta_t}^k, y_{i\quad t} = a\quad_t, \alpha_{a_t}, \beta) + \epsilon_{ita_t} \geqslant V(\cup_{ta_t^{\in}}^k, y_{i\quad t} = a\quad_t, \alpha_{a_t^{\in}}, \beta) + \epsilon_{ita_t^{\in}}\Big(\cdot
\end{aligned}
$$

Therefore,(2.2.1) can finally be

(2.2.4)

$$
\begin{aligned}
\mathcal{P}_j^k(\alpha,\beta) = \Big[&\prod_{a_T^{\in}\forall\!\!= a_T} \mathbb{1}\Big) V(\cup_{Ta_T}^k, a\quad_T, \alpha_{a_T}, \beta) + \epsilon_{iTa_T} \geqslant V(\cup_{Ta_T^{\in}}^k, a\quad_T, \alpha_{a_T^{\in}}, \beta) + \epsilon_{iTa_T^{\in}}\Big(dH_T(\epsilon) * \\
&\times\times\times \\
\Big[&\prod_{a_t^{\in}\forall\!\!= a_t} \mathbb{1}\Big) V(\cup_{ta_t}^k, a\quad_t, \alpha_{a_t}, \beta) + \epsilon_{ita_t} \geqslant V(\cup_{ta_t^{\in}}^k, a\quad_t, \alpha_{a_t^{\in}}, \beta) + \epsilon_{ita_t^{\in}}\Big(dH_t(\epsilon) * \\
&\times\times\times \\
\Big[&\prod_{a_1^{\in}\forall\!\!= a_1} \mathbb{1}\Big) V(\cup_{1a_1}^k, \alpha_{a_1}, \beta) + \epsilon_{i1a_1} \geqslant V(\cup_{1a_1^{\in}}^k, \alpha_{a_1^{\in}}, \beta) + \epsilon_{i1a_1^{\in}}\Big(dH_1(\epsilon).
\end{aligned}
$$

Therefore model implied choice probability conditional on $z_i = \cup^k$ is

$$
P_j^k(\beta, F_k) = \Big[\ \mathcal{P}_j^k(\alpha,\beta) dF_k(\alpha).
\tag{2.2.5}
$$

Denote the true choice probability as

$$\mathcal{S}_j^k = Pr(y_i = \mathcal{Z}^j \| z_i = \cup^k), \tag{2.2.6}$$

and $\mathcal{S} = (\mathcal{S}_1^1, \times\times\times, \mathcal{S}_J^1, \times\times\times, \mathcal{S}_1^K, \times\times\times, \mathcal{S}_J^K)$. Define the set of conditional distributions F_k that are consistent with $(\beta, \mathcal{S})$ as

$$\mathcal{G}_k(\beta, \mathcal{S}) = \Big\} F_k(\alpha) : \mathcal{S}_j^k = P_j^k(\beta, F_k), j = 1, \times\times\times, J \quad . \tag{2.2.7}$$

The identified set of β is actually

$$B = \}\beta : \mathcal{G}_k(\beta, \mathcal{S}) \mathrel{\forall\!\!=} \emptyset, k = 1, \times\times\times, K| \ . \tag{2.2.8}$$

As a result, the upper and lower bounds for $\Delta^k P_t$ can be defined as follows
(2.2.9)

$$\Delta_u^k P_t = \sup_{\beta / B, F_k / \mathcal{G}_k(\beta, \mathcal{S})} \int \Delta \mathcal{P}_t(\alpha, \beta) dF_k(\alpha), \ \Delta_l^k P_t = \inf_{\beta / B, F_k / \mathcal{G}_k(\beta, \mathcal{S})} \int \Delta \mathcal{P}_t(\alpha, \beta) dF_k(\alpha).$$

2.3 Estimation

By Lemma 3.1 in Huang 2015b, I can use a discrete distribution with no more than J support point, F_k^J, to replace any real distribution F_k, and this replacement won't undermine the analysis for partial effects bounds. Since F_k^J's supporting points vary with k and completely unknown, I use a fine grid over α_i's support Y and its pertaining distribution $\bar{\pi}^k$ instead in the hope of approximating F_k^J. In other words, the following combination is used to approximate F_k^J, and where M is larger than J.

$$\Big\} \mathrm{Y}_M = (\bar{\alpha}_{1M}, \times\times\times, \bar{\alpha}_{MM}), \bar{\pi}^k = (\bar{\pi}_1^k, \times\times\times, \bar{\pi}_M^k) \quad . \tag{2.3.1}$$

Therefore the approximated choice probability for (2.2.5) is

$$P_j^k(\beta, \bar{\pi}^k, M) = \sum_{m=1}^{M} \bar{\pi}_m^k \mathcal{P}_j^k(\bar{\alpha}_{mM}, \beta). \tag{2.3.2}$$

and the approximated partial effect on choice probability (2.1.11) is

$$\Delta^k P_t(\beta, \bar{\pi}^k, M) = \sum_{m=1}^{M} \bar{\pi}_m^k \Delta \mathcal{P}_t(\bar{\alpha}_{mM}, \beta). \tag{2.3.3}$$

All the approximations aforementioned are summations instead of integrations.

Also note that $\mathcal{P}_j^k(\alpha, \beta)$ and $\Delta \mathcal{P}_t(\alpha, \beta)$ do not usually have closed-form expressions. Their simulated versions are derived according to the following steps. ϵ_{it} follows distribution $H_t(\epsilon)$, and a specific joint distribution for ϵ_i can be constructed. Denote the joint distribution as $H(\epsilon)$, a serial of random errors can be generated according to $H(\epsilon)$. Denote them as $\}\tilde{\epsilon}_i|_{i=1}^{r}$. For a given triplet $(\cup^k, \alpha, \beta)$, simulated data $\}\tilde{y}_i|_{i=1}^{r}$ are given by (2.1.4). Thus simulated $\mathcal{P}_j^k(\alpha, \beta)$ is

$$\tilde{\mathcal{P}}_j^k(\alpha, \beta) = \frac{1}{r}\sum_{i=1}^{r} \mathbb{1}(\tilde{y}_i = \mathcal{Z}^j), \tag{2.3.4}$$

and its expanded expression is

(2.3.5)

$$\begin{aligned}\tilde{\mathcal{P}}_j^k(\alpha, \beta) = \frac{1}{r}\sum_{i=1}^{r} \Bigg\} &\prod_{a_1^{\in} \forall\!\!= a_1} \mathbb{1}\Big) V(\cup^k_{1a_1}, \alpha_{a_1}, \beta) + \tilde{\epsilon}_{i1a_1} \geqslant V(\cup^k_{1a_1^{\in}}, \alpha_{a_1^{\in}}, \beta) + \tilde{\epsilon}_{i1a_1^{\in}}\Big(* \\ &\times\times\times \\ &* \prod_{a_t^{\in} \forall\!\!= a_t} \mathbb{1}\Big) V(\cup^k_{ta_t}, a \quad t, \alpha_{a_t}, \beta) + \tilde{\epsilon}_{ita_t} \geqslant V(\cup^k_{ta_t^{\in}}, a \quad t, \alpha_{a_t^{\in}}, \beta) + \tilde{\epsilon}_{ita_t^{\in}}\Big(* \\ &\times\times\times \\ &* \prod_{a_T^{\in} \forall\!\!= a_T} \mathbb{1}\Big) V(\cup^k_{Ta_T}, a \quad T, \alpha_{a_T}, \beta) + \tilde{\epsilon}_{iTa_T} \geqslant V(\cup^k_{Ta_T^{\in}}, a \quad T, \alpha_{a_T^{\in}}, \beta) + \tilde{\epsilon}_{iTa_T^{\in}}\Big(\int \Bigg\{ .\end{aligned}$$

For given $(s_t^a, s_t^b, \alpha, \beta)$, where s_t^a and s_t^b are states after and before treatment, the simulated counterfactual data $\}\tilde{y}_{it}(s_t^a), \tilde{y}_{it}(s_t^b)|_{i=1}^{r}$ are given by (2.1.4), where $\tilde{y}_{it}(s_t) = g_0(s_t, \alpha, \tilde{\epsilon}_{it}, \beta)$. Thus the simulated $\Delta\mathcal{P}_t(\alpha, \beta)$ is

$$\tilde{\Delta}\mathcal{P}_t(\alpha, \beta) = \frac{1}{r}\sum_{i=1}^{r} \Big] \mathbb{1}(\tilde{y}_{it}(s_t^a) = a_t) \quad \mathbb{1}(\tilde{y}_{it}(s_t^b) = a_t)\Big\{, \tag{2.3.6}$$

and its expansion is

(2.3.7)

$$\begin{aligned}\tilde{\Delta}\mathcal{P}_t(\alpha, \beta) = \frac{1}{r}\sum_{i=1}^{r} \Bigg] &\prod_{a_t^{\in} \forall\!\!= a_t} \mathbb{1}\Big) V(s^a_{ta_t}, \alpha_{a_t}, \beta) + \tilde{\epsilon}_{ita_t} \geqslant V(s^a_{ta_t^{\in}}, \alpha_{a_t^{\in}}, \beta) + \tilde{\epsilon}_{ita_t^{\in}}\Big(\\ &\prod_{a_t^{\in} \forall\!\!= a_t} \mathbb{1}\Big) V(s^b_{ta_t}, \alpha_{a_t}, \beta) + \tilde{\epsilon}_{ita_t} \geqslant V(s^b_{ta_t^{\in}}, \alpha_{a_t^{\in}}, \beta) + \tilde{\epsilon}_{ita_t^{\in}}\Big(\Bigg\{ .\end{aligned}$$

The simulated choice probability and partial effect on choice probability in period t are respectively

$$\tilde{P}_j^k(\beta, \bar{\pi}^k, M) = \sum_{m=1}^{M} \bar{\pi}_m^k \tilde{\mathcal{P}}_j^k(\bar{\alpha}_{mM}, \beta), \tag{2.3.8}$$

and

$$\tilde{\Delta}^k P_t(\beta, \bar{\pi}^k, M) = \sum_{m=1}^{M} \bar{\pi}_m^k \tilde{\Delta}\mathcal{P}_t(\bar{\alpha}_{mM}, \beta). \tag{2.3.9}$$

To practice the estimation, the following quadratic objective function is used.

$$\hat{T}_\lambda(\beta, \bar{\pi}) = \sum_{j,k} \hat{\omega}_j^k \Big] \hat{\mathcal{S}}_j^k \quad \tilde{P}_j^k(\beta, \bar{\pi}^k, M) \Big\{^2 + \lambda_n \bar{\pi}^{\in} \bar{\pi}, \tag{2.3.10}$$

where the weighting $\hat{\omega}_j^k$ is a consistent estimator of $\omega_j^k = \mathcal{S}^k / \mathcal{S}_j^k$ and $\mathcal{S}^k$ is the real probability for $z_i = \cup^k$, $\hat{\mathcal{S}}_j^k$ is a consistent estimator for $\mathcal{S}_j^k$, $\bar{\pi} = \}\bar{\pi}^k|_{\, k=1, \times\!\times\!\times, K}$.

Estimator for the set identified parameters β is

$$\hat{B} = \}\beta \ / \ \mathbb{B} : \mathcal{A}\bar{\pi}, s.t. \hat{T}_\lambda(\beta, \bar{\pi}) \leqslant \xi_n{}^{|} \tag{2.3.11}$$

for a threshold value ξ_n, and where $\mathbb{B}$ is the parameter space of β. The estimated set of partial effects on choice probability in period t is

$$\hat{D}^k P_t = \Big\} \tilde{\Delta}^k P_t(\beta, \bar{\pi}^k, M) : \hat{T}_\lambda(\beta, \bar{\pi}) \leqslant \xi_n \quad . \tag{2.3.12}$$

Consequently, $\Delta_l^k P_t$ and $\Delta_u^k P_t$ are estimated by

$$\hat{\Delta}_l^k P_t = \min \hat{D}^k P_t \text{ and } \hat{\Delta}_u^k P_t = \max \hat{D}^k P_t. \tag{2.3.13}$$

2.4 Consistency

In this section, I prove that those estimators provided in section 2.3 are consistent, under a set of similar assumptions as in Huang 2015b. The sketch of proof follows Huang 2015b and I mention here only differences that arise for the reason of dynamic setting. Before I give lemmas and theories, I summarize all the assumptions which scattered in previous sections as follows

Assumption 2.4.1. *ϵ_{it} is independent of α_i, z_i and $\epsilon_{i \quad t}$, and its distribution is known as $H_t(\epsilon)$. What is more, ϵ_i's distribution is jointly known as $H(\epsilon)$.*

Note that here I do not assume that $H_t(\epsilon)$ is time homogeneous, since in dynamic model partial effects of choice probability at different period t are almost time heterogeneous even if $H_t(\epsilon)$ is time homogeneous. Thus this additional assumption won't help to improve estimation of partial effects of the choice probability as it did in static models and it is never a necessary condition for only consistent estimation.

Assumption 2.4.1 also excludes serially correlated errors while in static models this is allowed. The reason can easily be seen through expression (2.2.2), where the conditional choice probability of any history of choices can only be expressed as products of a sequence of iterative conditional probabilities because of the backwards dependence on state. Thereafter a serial independence of errors are needed to have an expression of overall integration over $H(\epsilon)$ as you can see in appendix **??**. This distinction can be rephrased as an assumption that all the serial dependences in errors are now completely captured by the explicit dynamics of the model.

Assumption 2.4.2. *z_i are discrete variables and the support of z_i is a finite set which can be written as $\cup = \}\cup^1, \times\times\times, \cup^{K|}$.*

z_i includes all the strictly exogenous variables x_i and the accumulated result of choices before the initial round of observations, w_{i1}. For example, individual's previous choices on schooling results in an accumulated variable, the *initial level of education*.

Instead of considering the unknown distribution F_k^J, I consider F_k^{J+1} and denote α_i's support points as $\alpha^k = \Big\}\alpha_1^k, \times\times\times, \alpha_{J+1}^k$, its distribution is π^k, so that let $\pi = \}\pi^1, \times\times\times, \pi^{K|}$ and $'\alpha = \}\alpha^1, \times\times\times, \alpha^K|$ and $\gamma = \}'\alpha, \pi|$. All the parameters in the model can be denoted as $\phi = \}\beta, \gamma| \ / \ \Phi$ and $\Phi = \mathbb{B} * \Upsilon^{(J+1)K} * \mathcal{X}_{J+1}^K$. By lemma 3.1 in Huang 2015b, partial effect on choice probability (2.1.11) can be expressed as

$$\Delta^k P_t = \Delta^k P_t(\phi) = \sum_{l=1}^{J+1} \pi_l^k \Delta \mathcal{P}_t(\alpha_l^k, \beta). \tag{2.4.1}$$

Assumption 2.4.3. *(i) α_i's support Υ is a compact subset of Euclidean space and it is endowed with a metric $d(\alpha, \alpha^{\ni})$; (ii) $\mathbb{B}$ is a compact subset of $\mathbb{R}^b$, where b is the number of elements in β; (iii) There is $L < \infty$ such that for all $(\alpha, \beta), (\alpha^{\in}, \beta^{\ni}) \ / \ \Upsilon * \mathbb{B}$,*

$$\Big\|\Delta \mathcal{P}_t(\alpha, \beta) \quad \Delta \mathcal{P}_t(\alpha^{\in}, \beta^{\ni})\Big\| \leqslant L \Big] d(\alpha, \alpha^{\ni}) + \Big(\beta \quad \beta^{\in}\Big(\{;$$

(iv) $\Delta^k P_t(\phi)$ is continuous in ϕ.

Lemma 2.4.1. *Simulated choice probability $\tilde{\mathcal{P}}_j^k(\alpha, \beta)$ defined in (2.3.5) uniformly converges to $\mathcal{P}_j^k(\alpha, \beta)$ over $\Upsilon * \mathbb{B}$ in probability under assumption (2.4.1). Furthermore,*

$$\bar{r}\Big)\tilde{\mathcal{P}}_j^k(\alpha, \beta) \quad \mathcal{P}_j^k(\alpha, \beta)\Big(\rightsquigarrow G(\alpha, \beta), \tag{2.4.2}$$

where $G(\alpha, \beta)$ is a mean zero Gaussian process, and its finite dimensional distribution is determined by $H(\epsilon)$.

Lemma 2.4.2. *Estimator $\tilde{\Delta}\mathcal{P}_t(\alpha,\beta)$ defined in (2.3.7) uniformly converges to $\Delta\mathcal{P}_t(\alpha,\beta)$ over $\Upsilon * \mathbb{B}$ in probability under assumption (2.4.1).*

Follow the same argument as Huang 2015b, I introduce the following two functions $\hat{Q}$ and Q

$$\hat{Q}\left(\phi(\beta,\bar{\pi})\right) = \sum_{j,k} \hat{\omega}_j^k \Big] \hat{\mathcal{S}}_j^k \quad \tilde{P}_j^k \Big)\beta, \gamma^k(\beta,\bar{\pi}^k)\Big(\Big\{^2, \tag{2.4.3}$$

and in terms of new notations, (2.2.5) can be written as $P_j^k(\phi)$. I define

$$Q(\phi) = \sum_{j,k} \omega_j^k \Big] \mathcal{S}_j^k \quad P_j^k(\phi)\Big\{^2. \tag{2.4.4}$$

Define $\Phi_I = \}\phi \ / \ \Phi : Q(\phi) = 0|$ and $\hat{\Phi} = \}\phi(\beta,\bar{\pi}) : \hat{Q}(\phi(\beta,\bar{\pi})) + \lambda_n \bar{\pi}^{\in}\bar{\pi} \leqslant \xi_n|$. The last assumption is

Assumption 2.4.4. *(i) $\eta(M) = \sup_{\alpha\,/\Upsilon} \min_{\alpha^{\in}/\Upsilon_M} d(\alpha,\alpha^{\ni}) \prime \quad 0$ as $M \prime \quad \infty$; (ii) there is a constant C such that for all $(\alpha,\beta),(\alpha^{\in}\!,\beta^{\ni}) \ / \ \Upsilon * \mathbb{B}$, $\left\|\mathcal{P}_j^k(\alpha^{\in}\!,\beta^{\ni}) \quad \mathcal{P}_j^k(\alpha,\beta)\right\| \leqslant C[d(\alpha,\alpha^{\ni}) + \backslash\beta^{\in} \quad \beta\backslash]$; (iii) $Q(\phi)$ is continuous in Φ; and (iv) let $\xi_n = n^{\kappa_1}$, $\eta(M) = n^{\kappa_2}$, $\lambda_n = n^{\kappa_3}$, $r = n^{\kappa_4}$ and $\}\kappa_2 < 0, 0 > \kappa_1 > \max\} \quad 1, \kappa_2| , \kappa_4 \geqslant \quad 2\kappa_2, \kappa_3 < \kappa_1|$.*

This assumption is exactly the same as assumption 4.3 in Huang 2015b, and also follow the same method I can prove that

Theorem 2.4.1. *Under all the assumptions, $d_H(\hat{\Phi},\Phi_I) \overset{P}{\prime} \quad 0$, where d_H is the Hausdorff metric for two sets.*

Theorem 2.4.2. *If $d_H(\hat{\Phi},\Phi_I) \overset{P}{\prime} \quad 0$, we have*

$$\hat{\Delta}_l^k P_t \overset{P}{\prime} \quad \Delta_l^k P_t \text{ and } \hat{\Delta}_u^k P_t \overset{P}{\prime} \quad \Delta_u^k P_t.$$

2.5 A Numeric Example

In this section, I consider a dynamic extension of the numeric example from Huang 2015b. Note that cars are actually durable goods, family's car purchasing decision should also depends on the number of cars the family has owned. Therefore besides all the variables included in the static car market model, another state variable NS_{it} should be introduced, where NS_{it} stands for the number of seats of all the cars owned by family i at time t after the purchase decision. Also note that it is reasonable not to buy any new cars, thus in this extended numeric example I consider three alternatives, i.e. $a = 0$ (no new cars), $a = 1$ (buy a two-seat car) and $a = 2$ (by a four-seat car). NS_{it} has its evolution governed by the following deterministic function

$$NS_{it} = NS_{it \quad 1} + 2\mathbb{1}(y_{it} = 1) + 4\mathbb{1}(y_{it} = 2), \ t = 1, 2, \tag{2.5.1}$$

where NS_{i0} is the initial state and is observed for each family.

Utilities of choosing different cars are described by the following function

$$U_{ita} = \beta_{1it} NS_a + \beta_{2it} PR_{ta} + \epsilon_{ita}, \ a = 0, 1, 2, \tag{2.5.2}$$

where NS_a is the number of seats of car a, PR_{ta} is the price of car a at time t. Especially when $a = 0$, no car is purchased, and NS_a and PR_{ta} are simply zeros. β_{1it} and β_{2it} are random coefficients that change over i and t, and they are modeled dynamically as follows

$$\begin{aligned} \beta_{1it} &= \left\} \begin{array}{l} \beta_1 NF_{it}/NS_{it}, \text{ if } NS_{it} \text{ is nonzero} \\ \text{any real number, otherwise} \end{array} \right. , \\ \beta_{2it} &= \beta_2 \alpha_i / FI_{it}, \end{aligned} \tag{2.5.3}$$

where the difference appears in the first equation and note that the value of the number of seats of any new cars also depends on the total number of car seats owned by the family after the purchase. The second equation has not been changed. Substitute (2.5.3) into (2.5.2) gives the utility without random coefficients

$$U_{ita} = \left\} \begin{array}{l} \beta_1 NF_{it} NS_a / NS_{it} + \beta_2 \alpha_i PR_{ta} / FI_{it} + \epsilon_{ita}, \text{ if } NS_{it} \text{ is nonzero,} \\ \epsilon_{ita}, \text{ otherwise,} \end{array} \right. \tag{2.5.4}$$

where the error terms $\epsilon_{it} = \}\epsilon_{ita}|_{a=0,1,2}$ are assumed iid joint normal over periods.

Suppose the true parameters are $\beta_1 = 1$ and $\beta_2 = \quad 2$, and the prices of cars are 10 and 15 respectively in all two periods. For the true DGP of the fixed effects α_i, I consider the same two cases as I did in Huang 2015b.

2.5.1 Interested Objects

Suppose the population consists of four types of families and they are listed as follows

1. $NS_{t_0} = 0, NF_{t_1} = 1, FI_{t_1} = 10, NF_{t_2} = 2, FI_{t_2} = 20,$
2. $NS_{t_0} = 2, NF_{t_1} = 2, FI_{t_1} = 10, NF_{t_2} = 3, FI_{t_2} = 20,$
3. $NS_{t_0} = 0, NF_{t_1} = 2, FI_{t_1} = 10, NF_{t_2} = 1, FI_{t_2} = 20,$

4. $NS_{t_0} = 2, NF_{t_1} = 2, FI_{t_1} = 10, NF_{t_2} = 2, FI_{t_2} = 20.$

They are evenly distributed in the population, and my simulated data has a sample of 8000 individuals, that is 2000 for each type. There are many partial effects that can be studied for each period, and the same partial effects can also be studied in all periods. It is impossible to study all of them, therefore I make a short list of them and show my study results. In the following list, there are 16 partial effects on choosing a four-seats car and 4 effects for each type of family.

1. For the first group:
   - (a) Partial effect on choosing a larger car if only the number of family members increases from 1 to 2 in the first period.
   - (b) Partial effect on choosing a larger car if only the family income increases from 20 to 30 in the second period given that a two-seats car was purchased in the first period[4].
   - (c) Partial effect on choosing a larger car if only the price of 4-seats car increases from 15 to 18 in the first period.
   - (d) Partial effect on choosing a larger car in the second period only because of buying a 2-seats car instead of buying no car in the first period.
2. For the second group:
   - (a) Partial effect on choosing a larger car if only the number of family members increases from 2 to 3 in the first period.
   - (b) Partial effect on choosing a larger car if only the family income increases from 20 to 30 in the second period given that no car was purchased in the first period.
   - (c) Partial effect on choosing a larger car if only the price of 4-seats car increases from 15 to 18 in the first period.
   - (d) Partial effect on choosing a larger car in the second period only because of buying a 2-seats car instead of buying no car in the first period.
3. For the third group:
   - (a) Partial effect on choosing a larger car if only the number of family members increases from 2 to 3 in the first period.

[4] Especially note that this partial effect is for a general family from the first group. It is not specific to families of the first group who purchased a two-seats car in the first period.

(b) Partial effect on choosing a larger car if only the family income increases from 20 to 30 in the second period given that no car was purchased in the first period.

(c) Partial effect on choosing a larger car if only the price of 4-seats car increases from 15 to 18 in the first period.

(d) Partial effect on choosing a larger car in the second period only because of buying a 2-seats car instead of buying no car in the first period.

4. For the fourth group:

   (a) Partial effect on choosing a larger car if only the number of family members increases from 2 to 3 in the first period.

   (b) Partial effect on choosing a larger car if only the family income increases from 20 to 30 in the second period given that no car was purchased in the first period.

   (c) Partial effect on choosing a larger car if only the price of 4-seats car increases from 15 to 18 in the first period.

   (d) Partial effect on choosing a larger car in the second period only because of buying a 2-seats car instead of buying no car in the first period.

A monte carlo study of 100 replications has been undertaken, and the average information about true values of these 16 partial effects and their estimated lower and upper bounds are reported in tables 2.1 and 2.2. Further information about their distribution can be seen from these boxplots in figures 2.1 and 2.2, where the 16 partial effects for all the four types of families are orderly named as from effect1 to effect16. Since the bounds of identified partial effects ask to find extreme values out of a set whose structure is complicated and less known, in practice approximations and compromise are made in finding them. Thus I do not report the true bounds of partial effects, instead the true partial effects are reported since they are easy to calculate by simulation. You can not see how close this estimator is to the true bounds in the monte carlo study, but the good news is that the estimated bounds cover true values quite well.

2.6 Conclusion

This paper gives parallel results of Huang 2015b's under a set of assumptions with minor distinctions for a class of dynamic discrete choice models. In both papers, the complete knowledge of the distribution of errors play a key role in the simulation method. In the static case, since all the covariates are strictly exogenous

Table 2.1: Partial effects on choosing a larger car (independent α_i)

Group 1	Effect a		Effect b		Effect c		Effect d	
Upper bound	0.05600		0.09897		-0.03209		0.05412	
Lower bound	0.04928		0.07271		-0.04280		0.01593	
True effect	0.05201*	(100)	0.08598*	(100)	-0.03710*	(100)	0.03270*	(100)
Group 2	Effect a		Effect b		Effect c		Effect d	
Upper bound	0.09978		0.09743		-0.06079		-0.05229	
Lower bound	0.07271		0.07318		-0.06822		-0.07341	
True effect	0.08614*	(100)	0.08556*	(100)	-0.06377*	(100)	-0.06386*	(100)
Group 3	Effect a		Effect b		Effect c		Effect d	
Upper bound	0.04850		0.07681		-0.06238		-0.00183	
Lower bound	0.03816		0.05954		-0.07238		-0.00628	
True effect	0.04535*	(100)	0.06927*	(100)	-0.06757*	(100)	-0.00417*	(100)
Group 4	Effect a		Effect b		Effect c		Effect d	
Upper bound	0.10207		0.09903		-0.06076		-0.03480	
Lower bound	0.07063		0.07280		-0.06834		-0.05264	
True effect	0.08624*	(100)	0.08596*	(100)	-0.06383*	(100)	-0.04414*	(100)

[1] * indicates the value is included by our estimated upper and lower bounds, where all values are the averages over the 100 simulations.

[2] α_i is independent with households' attributes and follows an uniform distribution over $[0.6, 1]$.

[3] $cov(\epsilon_{ita_1}, \epsilon_{ita_2}) = 0.6$, $var(\epsilon_{ita_0}) = var(\epsilon_{ita_1}) = var(\epsilon_{ita_2}) = 1$.

[4] Effects a, b, c and d are defined in previous paragraphs, and they may be different for different groups.

[5] Numbers in parentheses are frequencies of containing the true effect in the 95% confidence intervals out of 100 simulations.

it is possible to allow serial correlation in between them. While in dynamic discrete choice model, serial correlation in errors cause dependence between ϵ_{it} and $y_{i \quad t}$ and it makes the key step (B.0.1) in the following proof a challenge. To avoid this complication, serial independence is assumed in dynamic models. Another prevalent assumption in literatures of dynamic models is $\epsilon_{it}|\alpha_i, z_i, y_{i \quad t} \subset H_t(\epsilon)$, this assumption is weaker than what I use in this paper. This is because the weaker assumption is not adequate to support the equation in (B.0.1), where I need to apply the Fubini theorem. Using the little bit stronger assumption about ϵ_i, it turns out that the framework of proof for static models in Huang 2015b can be used to prove the results for dynamic models with a few alterations which I give in section 2.4.

A monte carlo study suggests this is a satisfactory method since the estimated bounds contain the true effect quite well, although I can not show how close they

Table 2.2: Partial effects on choosing a larger car (dependent α_i)

Group 1	Effect a		Effect b		Effect c		Effect d	
Upper bound	0.06247		0.09086		-0.03897		0.06100	
Lower bound	0.05490		0.06766		-0.04846		0.02249	
True effect	0.06259	(100)	0.07762*	(100)	-0.04686*	(99)	0.03918*	(100)
Group 2	**Effect a**		**Effect b**		**Effect c**		**Effect d**	
Upper bound	0.08236		0.10991		-0.04701		-0.05492	
Lower bound	0.06206		0.08585		-0.05610		-0.07528	
True effect	0.07187*	(100)	0.09746*	(100)	-0.05067*	(100)	-0.06605*	(100)
Group 3	**Effect a**		**Effect b**		**Effect c**		**Effect d**	
Upper bound	0.04499		0.08094		-0.04957		-0.00340	
Lower bound	0.03542		0.06635		-0.06193		-0.00737	
True effect	0.04248*	(100)	0.07449*	(100)	-0.05693*	(100)	-0.00610*	(100)
Group 4	**Effect a**		**Effect b**		**Effect c**		**Effect d**	
Upper bound	0.08453		0.10798		-0.04897		-0.03506	
Lower bound	0.06004		0.08322		-0.05589		-0.05135	
True effect	0.07189*	(100)	0.09537*	(100)	-0.05067*	(99)	-0.04355*	(100)

[1] * indicates the value is included by our estimated upper and lower bounds, where all values are the averages over the 100 simulations.

[2] α_i is dependent with households' attributes, it is uniformly distributed over $[0.6, 0.8]$ if family's first period income per capita is more than or equal to 10 or uniformly distributed over $[0.8, 1]$ if family's first period income per capita is less than 10.

[3] $cov(\epsilon_{ita_1}, \epsilon_{ita_2}) = 0.6$, $var(\epsilon_{ita_0}) = var(\epsilon_{ita_1}) = var(\epsilon_{ita_2}) = 1$.

[4] Effects a, b, c and d are defined in previous paragraphs, and they may be different for different groups.

[5] Numbers in parentheses are frequencies of containing the true effect in the 95% confidence intervals out of 100 simulations.

are to the true bounds.

2.7 Figures

Figure 2.1: Effects on Choosing A Larger Car (Independent α_i)

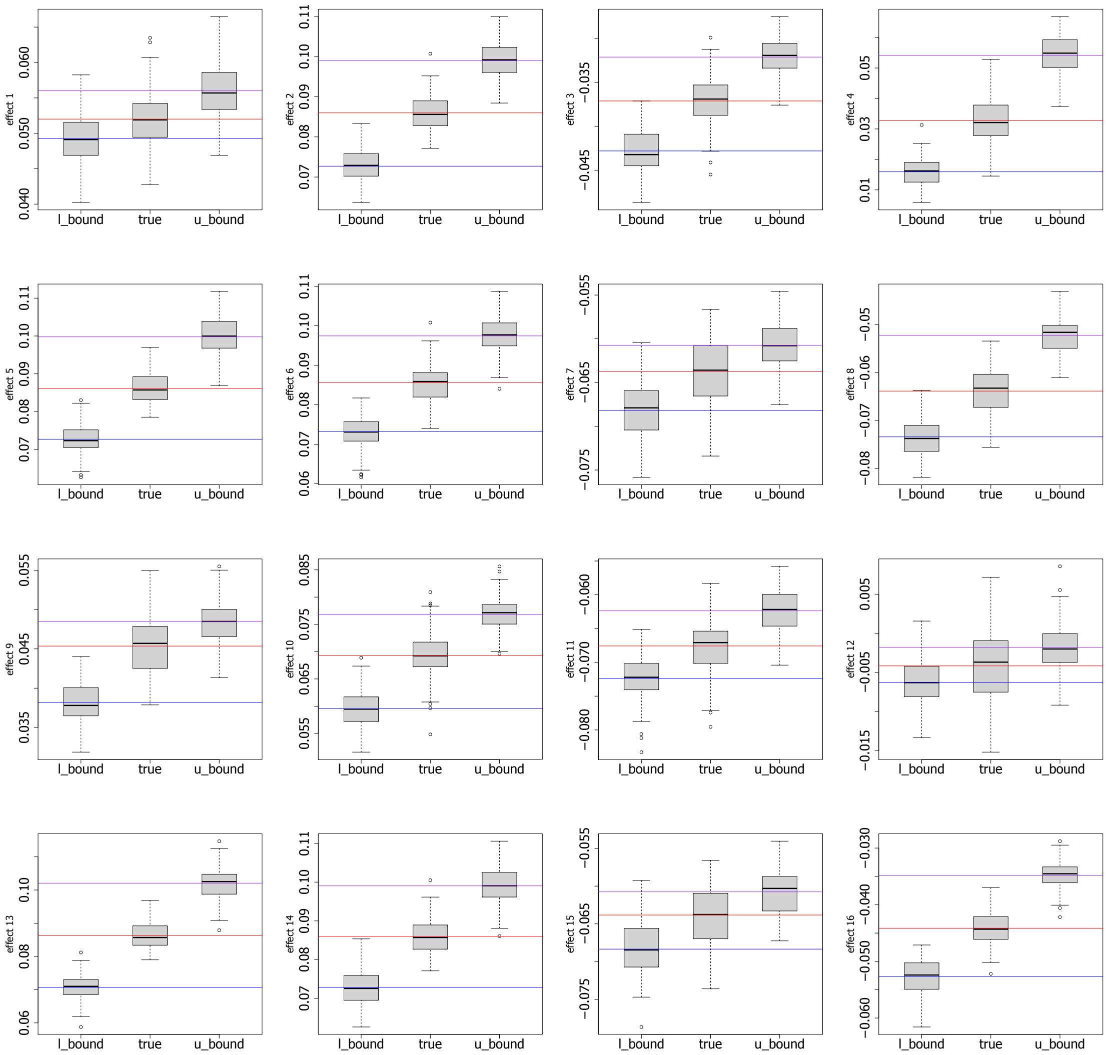
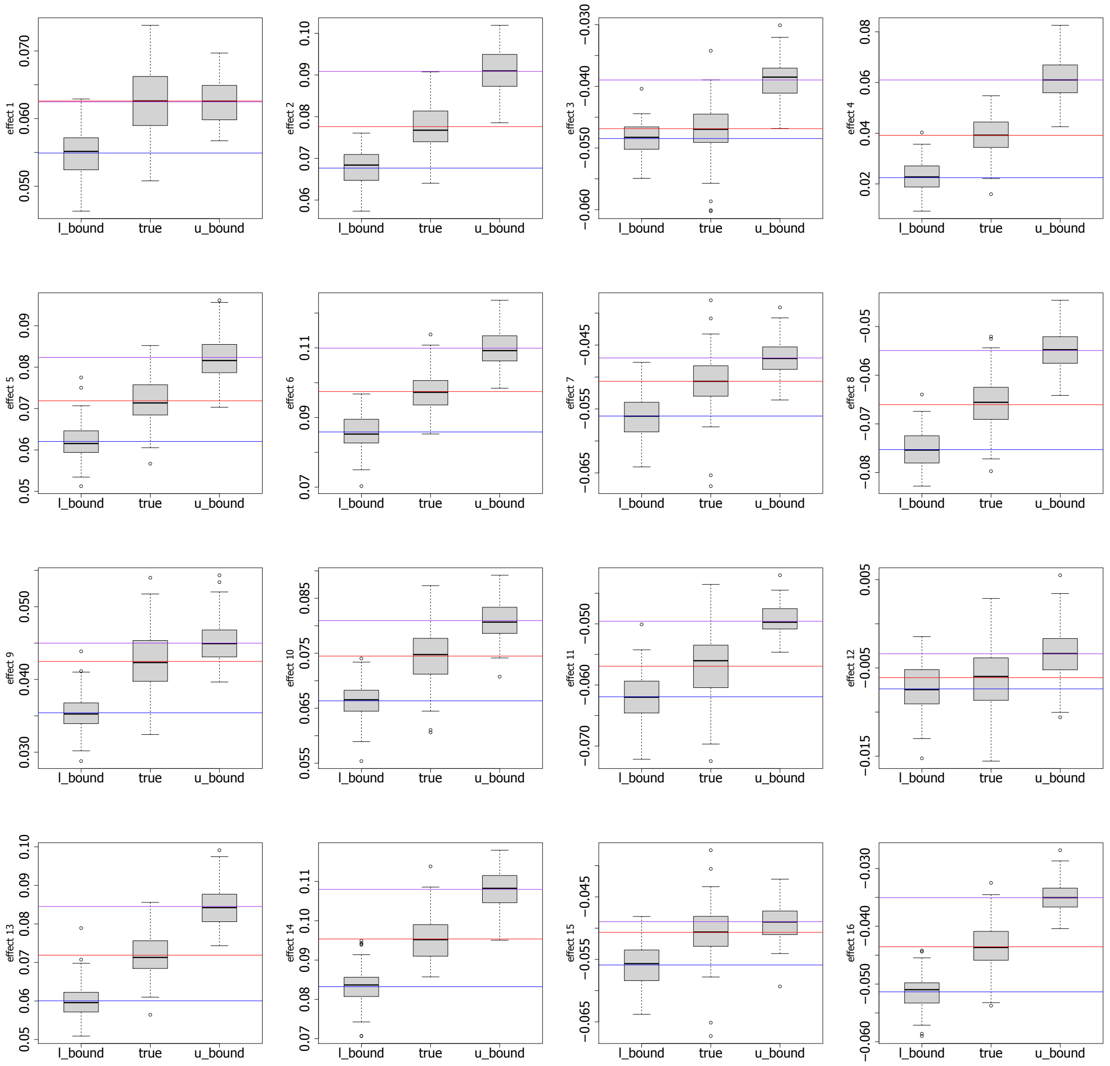


Figure 2.2: Effects on Choosing A Larger Car (Dependent α_i)



Chapter 3

The Career Decisions of Young Men Revised with Fixed Effects Approach

Abstract

Keane and Wolpin 1997 studied the career decisions of young men using a single-agent dynamic programming discrete choice model. Their model is a breakthrough in the development of dynamic programming discrete choice models since their model specification deviates from *Rust's model* in several aspects. One of these improvements is to allow unobservable individual effects that is time invariant. A well-known issue relating to individual effects is the model identification. Their practice is assuming that unobservable individual effects can be captured by a finite number of types and the number of types is known by researchers. And most applications have considered a small number of types due to the reason of identification. Actually, should we give up the point identification a more flexible specification of the individual effects is possible (see Chernozhukov et al. 2013a; Huang 2015b and Huang 2015a). In this paper I restudy the career decision model of young men using a more flexible specification of the individual effects. I am interested in the set identified structural parameters of the model and bounds of partial effects of various choice probabilities.

3.1 Introduction - Fixed Effects and Dynamic Programming Discrete Choice Model

There are various reasons to include unobservable individual effects in an econometric model. For most of the structure model this practice leads to a challenge of model identification since structure model tends to make individual effects non-separable and one can not get rid of them by primitive transformation like the first difference. Bonhomme 2012 provides an idea of *functional form difference* dealing with non-separable individual effects, however this idea does not apply to discrete choice models. It seems that giving up the point identification is an alternative for discrete choice models. Following Honoré and Tamer 2006 and Chernozhukov et al. 2013a, Huang 2015b and Huang 2015a developed the estimation methods for set identified static and dynamic discrete choice models respectively. This development allows non-closed-form components of model by using Monte Carlo simulation, thus it makes the idea of set identification more applicable at the cost of some extra computation and it may become an impediment for some complicated applications.

One such example is the class of dynamic programming discrete choice models[1]. This class of dynamic discrete choice models are forward looking models wherein individual's behavior is not only determined by the current period utility but also the discounted expectation of all future utilities. To practice the identification and calculate choice probabilities, it involves the solution of a dynamic programming problem for each trail of parameters. This is also computation intensive and maybe it is because of this problem, most of the dynamic discrete choice models focusing on the fixed effects do not consider individual's forward looking behavior explicitly or structurally.

To study young men's career decisions as shown by Keane and Wolpin 1997, both individual effects and forward looking behavior are important. Combining both challenges together is not a wise practice. Keane and Wolpin 1997 avoid the fixed effects and identification issue by assuming a known small number of types of the unobservable individual effects and using a mixture likelihood. In this paper I relax this assumption and follow the fixed effects approach making no restrictions on conditional distributions of individual effects[2]. As a cost of this practice one important compromise I make is to use an reduced form expression for the expectation of all future utilities. Due to this compromise, the model I propose is not a fully structural model but partial structural. Although this is a compromise in technique, there are some interpretations that consolidate its practice.

[1] See Aguirregabiria and Mira 2010 for a great survey of this class of models.

[2] One exception is that individual effects' support should be compact.

First, the reduced form of the future utility expectation can be considered as an approximation of the discounted utility from the future. This is because in many applications value functions are smooth and approximation techniques work well for smooth functions. Second, reduced form can be considered as a representation of individual's limited rationality for forward looking reasoning. It acts as a rule of thumb in many complicated situations. By the second interpretation, it is even possible to model idiosyncratic rules of thumb adopted by individuals by introducing further more individual effects. For simplicity I suppress this possibility in later discussion of this paper.

3.2 A Modified Basic Human Capital Model

The model I am studying relies on the basic human capital model of Keane and Wolpin 1997 with some modifications which I declare later. For each individuals the observation starts at their age 16 and ends at a maximum age T. Each year, an individual chooses between staying at home ($a = 5$), attending school ($a = 4$), or working at one of three occupations: blue collar ($a = 1$), white collar ($a = 2$), or the military ($a = 3$). So the set of alternatives is $\mathcal{F} = \}1,2,3,4,5|$. In this application, these five alternatives are hard to be described by their features or attributes[3], as such alternative specific parameters are used to model their effects on individual utility[4]. Assume that there is a vector $\alpha_i = \}\alpha_{ia}|_{a / \mathcal{F}}$ containing occupation and individual-specific endowments which is fixed from age 16 on. The vector of observable state variables is $x_{ita} = \}h_{it}, k_{ita} : a = 1,2,3|$, where h_{it} is schooling (in years), t presents age and k_{ita} is cumulated work experience (in years) in occupation a. Other unobservable state variables are ϵ_{ita} which is known by individual before he/she makes his/her decision. In this application I assume that ϵ_{ita} are independent over i, t and a. There could not be serial correlations due to my finding in Huang 2015a. What is more, for $a = 1,2,3$, ϵ_{ita} distribute as $\exp(N(0,1))$, and for $a = 4,5$, ϵ_{ita} distribute as $N(0,1)$, where $N(0,1)$ is the standard normal distribution. The lower and upper bounds for fixed effects in each dimension are assumed to be 0.7 and 0.7 respectively, which covers almost 52% of the range of variations in a standard normal distribution.

[3] These alternatives are not concrete objects as products or service, therefore there are hardly measures to tell their attributes and there are neither qualified measures in the available data set.

[4] If attributes of alternatives are available, it could be used in conjunction with a shorter common parameter vector to reduce the demension of parameters. See also discussions in Huang 2015b.

The specification of the current period utility functions of each alternative is

(3.2.1)

$$\begin{aligned}
&U(x_{it5}, \alpha_{i5}, \epsilon_{it5}) = \alpha_{i5} + \epsilon_{it5}\\
&U(x_{it4}, \alpha_{i4}, \epsilon_{it4}) = \alpha_{i4} \quad \beta_1 I(h_{it} \geqslant 12) \quad \beta_2 I(h_{it} \geqslant 16) + \epsilon_{it4}\\
&U(x_{ita}, \alpha_{ia}, \epsilon_{ita}) = W_{ita} = r_a \exp\}\alpha_{ia} + \beta_3 h_{it} + \beta_4 k_{ita} \quad \beta_5 k_{ita}^2| + \epsilon_{ita}, \text{ for } a = 1, 2, 3.
\end{aligned}$$

Compared with the original specification in Keane and Wolpin 1997, I take the error term ϵ_{ita} out of the curly bracket in the last equation of (3.2.1). This is a requirement by the technique of simulation which I am going to use. It is assumed working experience for everyone at age 16 is zero, i.e. $k_{i16a} = 0$ for all i and a. Age t evolves naturally every year and schooling and work experience evolves according to the following rules

(3.2.2)
$$\begin{aligned}
h_{it} &= h_{it \quad 1} + \mathbb{1}(y_{it \quad 1} = 4)\\
k_{ita} &= k_{it \quad 1a} + \mathbb{1}(y_{it \quad 1} = a) \text{ for } a = 1, 2, 3,
\end{aligned}$$

where y_{it} is the observed choice of individual i at age t.

Let the discount parameter be ρ, then every period t individual observes x_{it}, α_i and ϵ_{it} and choose his/her action $(a_{it}, \times\times\times, a_{iT})$ to maximize his/her expected utility

(3.2.3)
$$E \Big) \sum_{j=0}^{T \quad t} \rho^j U(x_{it+ja}, \alpha_{ia}, \epsilon_{it+ja} \| a_{it}, x_{it}, \alpha_i, \epsilon_{it}) \Big[.$$

Let $g(x_{it}, \alpha_i, \epsilon_{it})$ and $V(x_{it}, \alpha_i, \epsilon_{it})$ are the policy function and value function of the dynamic programming (DP) problem. By Bellman's principle of optimality the value function can be presented in the recursive expression

(3.2.4)
$$V(x_{it}, \alpha_i, \epsilon_{it}) = \max_{a/\mathcal{F}} \}U(x_{ita}, \alpha_{ia}, \epsilon_{ita}) + \rho \int V(x_{it+1}, \alpha_i, \epsilon_{it+1}) dF(x_{it+1}, \epsilon_{it+1} \| a, x_{it}, \alpha_i, \epsilon_{it})| .$$

Let's assume that $\rho \int V(x_{it+1}, \alpha_i, \epsilon_{it+1}) dF(x_{it+1}, \epsilon_{it+1} \| a, x_{it}, \alpha_i, \epsilon_{it})$ is a smooth function of x_{it}, α_i and ϵ_{it} for any given a. I need to find a strategy to approximate this function. For simplicity, suppose the approximation is given by a function $u(a, x_{it}, \alpha_i)$ and ϵ_{it} is ignored from this approximation. Note U is additive separable in ϵ_{ita} so that $U(x_{ita}, \alpha_{ia}, \epsilon_{ita})$ can be written as $\bar{U}(x_{ita}, \alpha_{ia}, \beta) + \epsilon_{ita}$. Therefore, individual's decision rule can be presented by

(3.2.5)
$$\begin{aligned}
y_{it} &= \arg\max_{a/\mathcal{F}} \}\bar{U}(x_{ita}, \alpha_{ia}, \beta) + u(a, x_{it}, \alpha_i) + \epsilon_{ita}|\\
&= g(x_{it}, \alpha_i, \epsilon_{it}).
\end{aligned}$$

Complement (3.2.5) with the previous assumption about ϵ_{it}

$$\epsilon_{it} \| \alpha_i, x_{i16}, \epsilon_{i\ \ t} \subset H_t(\epsilon), \tag{3.2.6}$$

this application fit the framework proposed by Huang 2015a, where $\epsilon_{i\ \ t} = \}\epsilon_{i16}, \times\times\times, \epsilon_{it\ \ 1}|$, and (3.2.6) is actually the same as

$$\epsilon_{it} \| \alpha_i, h_{i16}, \epsilon_{i\ \ t} \subset H_t(\epsilon), \tag{3.2.7}$$

since $k_{i16a} = 0$ for all occupations $a = 1, 2, 3$ by assumption.

3.2.1 Approximation Specification

How to specify the approximation of the future utility expectation is a challenge in practice. For the reason of accuracy higher order polynomial with richer cross terms is required. But it introduces too many parameters simultaneously and hinder the practice of estimation. Thus I use a first order polynomial to do the approximation. Denote the column vector $z_{it} = (1, h_{it}, k_{it1}, k_{it2}, k_{it3}, \alpha_{i1}, \times\times\times, \alpha_{i5})$, and let

$$u(a, x_{it}, \alpha_i) = z_{it}^{\in}\theta_a, \tag{3.2.8}$$

where θ_a is a choice specific vector of parameters. Therefore (3.2.5) can be rewritten as

$$\begin{aligned} y_{it} &= \underset{a/\mathcal{F}}{\arg\max} \}\bar{U}(x_{ita}, \alpha_{ia}, \beta) + z_{it}^{\in}\theta_a + \epsilon_{ita}| \\ &= g(x_{it}, \alpha_i, \epsilon_{it}, \beta, \theta) \\ &= g(h_{i16}, y_{i\ \ t}, \alpha_i, \epsilon_{it}, \beta, \theta), \end{aligned} \tag{3.2.9}$$

where β are parameters from the structural part of current utility and θ are parameters from the reduced-form expected utility of the future, and $y_{i\ \ t} = \}y_{i16}, \times\times\times, y_{it\ \ 1}|$. The last equation holds because $k_{i16a} = 0$ for all $a = 1, 2, 3$, and h_{it} and k_{ita} evolve according to (3.2.2) such that $(h_{i16}, y_{i\ \ t})$ can give the same information as x_{it} do.

3.2.2 Interested Objects In This Application

Basically parameters such as β and θ are interested, but partial effects are more interesting in set identified models. Denote $s_{it} = \}h_{i16}, y_{i\ \ t}| = \}h_{it^{\in}}, k_{it^{\in}1}, k_{it^{\in}2}, k_{it^{\in}3}|_{16 \leqslant t^{\in} \leqslant t}$, all the partial effects on the choice probability of choosing a_t in period t by individuals with initial schooling $h_{i16} = \mathcal{L}^{k}$ can be generally denoted as follows

(3.2.10)
$$\Delta^k P_t = Pr(y_{it} = a_t \| (s_{it} = s_t^a), h_{i16} = \mathcal{L}^{k}) \quad Pr(y_{it} = a_t \| (s_{it} = s_t^b), h_{i16} = \mathcal{L}^{k}),$$

where s_t^b and s_t^a are different states before and after the treatment, and the parenthesis around it is to emphasize that this probability is the potential probability for the individual with $h_{i16} = \mathcal{L}^{k}$ behaving as one with state s_t^a or s_t^b in period t no matter whatever real state he/she may have.

To understand this probability, it is better to consider another potential probability

(3.2.11)
$$\begin{aligned}\mathcal{P}_{a_t}^{s_t}(\alpha, \beta, \theta) &= Pr(y_{it} = a_t \| (s_{it} = s_t), h_{i16} = \mathcal{L}, \alpha_i = \alpha) \\ &= [\ \mathbb{1}\left(g(s_t, \alpha, \epsilon_{it}, \beta, \theta) = a_t\right) dH_t(\epsilon) \\ &= [\ \prod_{a_t^{\in} \neq a_t} \mathbb{1} \Big) \bar{U}(x_{ita_t^{\in}}, \alpha_{a_t^{\in}}, \beta) + z_{it}{}^{\in}\theta_{a_t^{\in}} + \epsilon_{ita_t^{\in}} \leqslant \bar{U}(x_{ita_t}, \alpha_{a_t}, \beta) + z_{it}{}^{\in}\theta_{a_t} + \epsilon_{ita_t} \Big(dH_t(\epsilon),\end{aligned}$$

where s_t completely determines $x_{it} = \} h_{it}, k_{ita} : a = 1, 2, 3 |$. For any distribution of α_i conditional on $h_{i16} = \mathcal{L}^{k}$, i.e. $F_k(\alpha)$, it holds that

$$Pr(y_{it} = a_t \| (s_{it} = s_t), h_{i16} = \mathcal{L}^{k}) = [\ \mathcal{P}_{a_t}^{s_t}(\alpha, \beta, \theta) dF_k(\alpha). \tag{3.2.12}$$

Or if we define

$$\Delta \mathcal{P}_t(\alpha, \beta, \theta) = \mathcal{P}_{a_t}^{s_t^a}(\alpha, \beta, \theta) \quad \mathcal{P}_{a_t}^{s_t^b}(\alpha, \beta, \theta), \tag{3.2.13}$$

(3.2.10) can be written as

$$\Delta^k P_t = [\ \Delta \mathcal{P}_t(\alpha, \beta, \theta) dF_k(\alpha). \tag{3.2.14}$$

Partial effects as (3.2.10) covers a lot of examples, I name a few later in the section of empirical study.

3.3 Model Identification and Estimation

3.3.1 Identification

In this application, observable strictly exogenous variable is only h_{i16}. Suppose there are K levels of initial schooling in years and they are $\} \mathcal{L}^{1}, \times\times\times, \mathcal{L}^{K} |$. For the observations of individuals' choices during the $T \quad 15$ years, there are J combinations of choices over years and they are $\} \mathcal{Z}^1, \times\times\times, \mathcal{Z}^J |$. Any $\mathcal{Z}^j$ can be represented as $(a_{16}, \times\times\times, a_T)$, where $a_t = 1, \times\times\times, 5$ for any t.

For identification, the method depends on the model implied choice probabilities $\mathcal{P}_j^k$ and P_j^k, and they are defined as follows

$$\begin{aligned}
\mathcal{P}_j^k(\alpha,\beta,\theta) &= Pr(y_i = \mathcal{Z}^j \| h_{i16} = \mathcal{L}^k, \alpha_i = \alpha) \\
&= Pr(y_{iT} = a_T, \times\times\times, y_{i16} = a_{16} \| h_{i16} = \mathcal{L}^k, \alpha_i = \alpha) \\
&= Pr(y_{iT} = a_T \| y_{i \quad T} = a \quad {}_T, h_{i16} = \mathcal{L}^k, \alpha_i = \alpha) * \\
&\quad \times\times\times \\
&\quad Pr(y_{it} = a_t \| y_{i \quad t} = a \quad {}_t, h_{i16} = \mathcal{L}^k, \alpha_i = \alpha) * \\
&\quad \times\times\times \\
&\quad Pr(y_{i16} = a_{16} \| h_{i16} = \mathcal{L}^k, \alpha_i = \alpha).
\end{aligned} \tag{3.3.1}$$

And it could also be expressed as

$$\begin{aligned}
\mathcal{P}_j^k(\alpha,\beta,\theta) = \Big[\;\mathbb{1}\Big) & g(\mathcal{L}^k, y_{i \quad T} = a \quad {}_T, \alpha, \epsilon_{iT}, \beta, \theta) = a_T \Big(dH_T(\epsilon) * \\
& \times\times\times \\
& \Big[\;\mathbb{1}\Big) g(\mathcal{L}^k, y_{i \quad t} = a \quad {}_t, \alpha, \epsilon_{it}, \beta, \theta) = a_t \Big(dH_t(\epsilon) * \\
& \times\times\times \\
& \Big[\;\mathbb{1}\Big) g(\mathcal{L}^k, \alpha, \epsilon_{i16}, \beta, \theta) = a_{16} \Big(dH_{16}(\epsilon).
\end{aligned} \tag{3.3.2}$$

Note the assumption of ϵ_{it} in (3.2.6), and

$$\begin{aligned}
&\mathbb{1}\Big) g(\mathcal{L}^k, y_{i \quad t} = a \quad {}_t, \alpha, \epsilon_{it}, \beta, \theta) = a_t \Big(= \\
&\prod_{a_t^{\in} \forall\!\!\!= a_t} \mathbb{1}\Big) \bar{U}(h_{i16}, y_{i \quad t} = a \quad {}_t, \alpha_{a_t}, \beta) + z_{it}^{\in}\theta_{a_t} + \epsilon_{ita_t} \geqslant \bar{U}(h_{i16}, y_{i \quad t} = a \quad {}_t, \alpha_{a_t^{\in}}, \beta) + z_{it}^{\in}\theta_{a_t^{\in}} + \epsilon_{ita_t^{\in}} \Big(,
\end{aligned} \tag{3.3.3}$$

$\mathcal{P}_j^k(\alpha,\beta,\theta)$ can be further written as

$$\begin{aligned}
\mathcal{P}_j^k(\alpha,\beta,\theta) = \Big[\;\mathbb{1}\Big) & g(\mathcal{L}^k, y_{i \quad T} = a \quad {}_T, \alpha, \epsilon_{iT}, \beta, \theta) = a_T \Big(* \\
& \times\times\times \\
& \mathbb{1}\Big) g(\mathcal{L}^k, y_{i \quad t} = a \quad {}_t, \alpha, \epsilon_{it}, \beta, \theta) = a_t \Big(* \\
& \times\times\times \\
& \mathbb{1}\Big) g(\mathcal{L}^k, \alpha, \epsilon_{i16}, \beta, \theta) = a_{16} \Big(dH(\epsilon) \\
= \Big[\; & \prod_{a_T^{\in} \forall\!\!\!= a_T} \mathbb{1}\Big) \bar{U}(h_{i16}, a \quad {}_T, \alpha_{a_T}, \beta) + z_{iT}^{\in}\theta_{a_T} + \epsilon_{iTa_T} \geqslant \bar{U}(h_{i16}, a \quad {}_T, \alpha_{a_T^{\in}}, \beta) + z_{iT}^{\in}\theta_{a_T^{\in}} + \epsilon_{iTa_T^{\in}} \Big(* \\
& \times\times\times \\
& \prod_{a_t^{\in} \forall\!\!\!= a_t} \mathbb{1}\Big) \bar{U}(h_{i16}, a \quad {}_t, \alpha_{a_t}, \beta) + z_{it}^{\in}\theta_{a_t} + \epsilon_{ita_t} \geqslant \bar{U}(h_{i16}, a \quad {}_t, \alpha_{a_t^{\in}}, \beta) + z_{it}^{\in}\theta_{a_t^{\in}} + \epsilon_{ita_t^{\in}} \Big(* \\
& \times\times\times \\
& \prod_{a_{16}^{\in} \forall\!\!\!= a_{16}} \mathbb{1}\Big) \bar{U}(h_{i16}, \alpha_{a_{16}}, \beta) + z_{i16}^{\in}\theta_{a_{16}} + \epsilon_{i16a_{16}} \geqslant \bar{U}(h_{i16}, \alpha_{a_{16}^{\in}}, \beta) + z_{i16}^{\in}\theta_{a_{16}^{\in}} + \epsilon_{i16a_{16}^{\in}} \Big(dH(\epsilon).
\end{aligned} \tag{3.3.4}$$

Model implied choice probability conditional on $h_{i16} = \mathcal{L}^k$ is thereafter

$$P_j^k(\beta, \theta, F_k) = \Big[\ \mathcal{P}_j^k(\alpha, \beta, \theta) dF_k(\alpha). \tag{3.3.5}$$

Denote the true choice probability as $\mathcal{S}_j^k$, the set of conditional distributions F_k which are consistent with $(\beta, \theta, \mathcal{S})$, where $\mathcal{S} = (\mathcal{S}_1^1, \times\times\times, \mathcal{S}_J^1, \times\times\times, \mathcal{S}_1^K, \times\times\times, \mathcal{S}_J^K)$, are

$$\mathcal{G}_k(\beta, \theta, \mathcal{S}) = \Big\} F_k(\alpha) : \mathcal{S}_j^k = P_j^k(\beta, \theta, F_k), j = 1, \times\times\times, J \quad , \text{ for all } k = 1, \times\times\times, K. \tag{3.3.6}$$

After all, the identified set of (β, θ) can be actually defined as

$$B = \}(\beta, \theta) : \mathcal{G}_k(\beta, \theta, \mathcal{S}) \nvDash \emptyset, k = 1, \times\times\times, K| \ . \tag{3.3.7}$$

Consequently the upper and lower bounds for $\Delta^k P_t$ can be defined as

$$\begin{aligned} \Delta_u^k P_t &= \sup_{(\beta,\theta)/B, F_k/\mathcal{G}_k(\beta,\theta,\mathcal{S})} \Big[\ \Delta\mathcal{P}_t(\alpha, \beta, \theta) dF_k(\alpha), \\ \Delta_l^k P_t &= \inf_{(\beta,\theta)/B, F_k/\mathcal{G}_k(\beta,\theta,\mathcal{S})} \Big[\ \Delta\mathcal{P}_t(\alpha, \beta, \theta) dF_k(\alpha). \end{aligned} \tag{3.3.8}$$

3.3.2 Estimation and Further Details

Following the estimation method proposed by Huang 2015b and Huang 2015a, a serial of random error terms, i.e. $\tilde{\epsilon}_i$, should be first generated according to distribution $H(\epsilon)$. In this application the fixed effects is of five dimension and in each dimension its range is from   0.7 to 0.7 as I mentioned previously. I use a fixed grid of m points to approximate the fixed effects in each dimension, and there are $M = m^5$ points for the fixed grid overall. Therefor the pertaining distribution of the fixed effects conditional on $h_{i16} = \mathcal{L}^k$ is discretized as a positive vector $\bar{\pi}^k$ which sums to unit.

Given simulated error terms, simulated data $\}\tilde{y}_i|_{i=1}^r$ is generated by (3.2.9), and $\mathcal{P}_j^k$ and $\Delta\mathcal{P}_t$ has their versions in simulation as follows

$$\tilde{\mathcal{P}}_j^k(\alpha, \beta, \theta) = \frac{1}{r} \sum_{i=1}^r \mathbb{1}(\tilde{y}_i = \mathcal{Z}^j), \tag{3.3.9}$$

$$\tilde{\Delta}\mathcal{P}_t(\alpha, \beta, \theta) = \frac{1}{r} \sum_{i=1}^r \Big] \mathbb{1}(\tilde{y}_{it}(s_t^a) = a_t) \quad \mathbb{1}(\tilde{y}_{it}(s_t^b) = a_t) \Big\{. \tag{3.3.10}$$

Therefore approximated choice probability and partial effects for (3.3.5) and (3.2.14) are

$$\tilde{P}_j^k(\beta,\theta,\bar{\pi}^k,M) = \sum_{m=1}^{M} \bar{\pi}_m^k \tilde{\mathcal{P}}_j^k(\bar{\alpha}_m,\beta,\theta), \tag{3.3.11}$$

$$\tilde{\Delta}^k P_t(\beta,\theta,\bar{\pi}^k,M) = \sum_{m=1}^{M} \bar{\pi}_m^k \tilde{\Delta}\mathcal{P}_t(\bar{\alpha}_m,\beta,\theta). \tag{3.3.12}$$

The estimation of the identified set of β and θ is given by

$$\hat{B} = \Big\}(\beta,\theta) : \mathcal{A}\bar{\pi},\ \text{s.t.} \sum_{j,k}\omega_j^k \Big]\hat{\mathcal{S}}_j^k \quad \tilde{P}_j^k(\beta,\theta,\bar{\pi}^k,M)\Big\{^2 + \lambda_n \bar{\pi}^{\in}\bar{\pi} \leqslant \xi_n \Big\lceil . \tag{3.3.13}$$

A similar set for partial effects is given by

$$\hat{D}^k P_t = \Big\}\tilde{\Delta}^k P_t(\beta,\theta,\bar{\pi}^k,M) : \sum_{j,k}\omega_j^k \Big]\hat{\mathcal{S}}_j^k \quad \tilde{P}_j^k(\beta,\theta,\bar{\pi}^k,M)\Big\{^2 + \lambda_n \bar{\pi}^{\in}\bar{\pi} \leqslant \xi_n \Big\lceil , \tag{3.3.14}$$

thereafter its lower and upper bounds are given by

$$\begin{aligned} \hat{\Delta}_l^k P_t &= \min \hat{D}^k P_t, \\ \hat{\Delta}_u^k P_t &= \max \hat{D}^k P_t. \end{aligned} \tag{3.3.15}$$

Actually to practice $\hat{B}$ and find the lower and upper bounds for partial effects is really a challenge. I find no perfect algorithm to do that job. Alternatively I adopt the following method as a compromise.

First, let me explain how to find the set of $\hat{B}$. I define an objective function as follows

$$\begin{aligned} \hat{T}_\lambda(\beta,\theta,\bar{\pi}) &= \sum_{j,k}\omega_j^k \Big]\hat{\mathcal{S}}_j^k \quad \tilde{P}_j^k(\beta,\theta,\bar{\pi}^k,M)\Big\{^2 + \lambda_n \bar{\pi}^{\in}\bar{\pi} \\ &= \sum_{j,k}\omega_j^k \Bigg]\hat{\mathcal{S}}_j^k \quad \sum_{m=1}^{M} \bar{\pi}_m^k \tilde{\mathcal{P}}_j^k(\bar{\alpha}_m,\beta,\theta)\Bigg\{^2 + \lambda_n \bar{\pi}^{\in}\bar{\pi}. \end{aligned} \tag{3.3.16}$$

Note for given (β,θ) this objective function is only a function of $\bar{\pi}$. Minimizing $\hat{T}_\lambda(\beta,\theta,\bar{\pi})$ over $\bar{\pi}$ is actually a quadratic programming problem[5] with restrictions that $\bar{\pi}^k$ should be positive and sum to unit for all k. Denote

$$\hat{T}_\lambda(\beta,\theta) = \min_{\bar{\pi}} \hat{T}_\lambda(\beta,\theta,\bar{\pi}). \tag{3.3.17}$$

[5] This step is practiced with the help of R package *quadprog* by Turlach and Weingessel 2013.

$\hat{T}_\lambda(\beta, \theta)$ is a simulated function of β and θ, then I use several algorithms[6] to minimize this function over β and θ. Finally I choose the method "nmkb", which is a Nelder-Mead simplex method enhanced with box constraint, to minimize $\hat{T}_\lambda(\beta, \theta)$. During its process of searching for smaller values of this function, a large number of β and θ have been tested. I collect all the tested β and θ which give smaller function value than the threshold ξ_n. Given enough time and trying different starting points, more parameters β and θ can be found. All these β and θ are used as an approximation of $\hat{B}$ in the practice.

Second, take finding the lower bound of partial effects as an example, let me talk about how to find the bounds of partial effects. Scrutinize the definition of lower bound in (3.3.15), it could actually be re-expressed as

(3.3.18)

$$\hat{\Delta}_l^k P_t = \min_{(\beta,\theta)\,/\,\hat{B}} \min_{\bar{\pi}} \Big\} \tilde{\Delta}^k P_t(\beta, \theta, \bar{\pi}^k, M) : \textstyle\sum_{j,k} \omega_j^k \Big] \hat{\mathcal{S}}_j^k \quad \tilde{P}_j^k(\beta, \theta, \bar{\pi}^k, M) \Big\{^2 + \lambda_n \bar{\pi}^{\in} \bar{\pi} \leqslant \xi_n \Big(.$$

After consider the bound searching problems as such a two-step problem. It can be realized that another challenge comes from the inner searching. Note that for each $(\beta, \theta) \ / \ \hat{B}$, a minimization of $\hat{T}_\lambda(\beta, \theta, \bar{\pi})$ over $\bar{\pi}$ can generate a set of projected choice probabilities $\tilde{P}_j^k$ defined by the given β, θ and optimum $\bar{\pi}$. I use the following linear programming problem to approximate the inner bound searching.

$$\begin{aligned}
&\min_{\bar{\pi}^k} \sum_{m=1}^{M} \bar{\pi}_m^k \tilde{\Delta} \mathcal{P}_t(\bar{\alpha}_m, \beta, \theta) \\
&\text{s.t.} \ \sum_{m=1}^{M} \bar{\pi}_m^k \tilde{\mathcal{P}}_j^k(\bar{\alpha}_m, \beta, \theta) = \tilde{P}_j^k \text{ for all } j \\
&\qquad \bar{\pi}_m^k \geqslant 0 \text{ for all } m \\
&\qquad \sum_{m=1}^{M} \bar{\pi}_m^k = 1,
\end{aligned} \tag{3.3.19}$$

where $\tilde{P}_j^k$ stands for the projected choice probability defined by the given β, θ and optimum $\bar{\pi}$.

3.4 Data

I take the same sauce of data as Keane and Wolpin 1997 did. The NLSY79 data consists of 12,686 individuals, and this re-analysis is based on white males who

[6] Nash and R. Varadhan 2011 provided an R package *optimx* which unifided the interface for several popular algorithms available in R. It makes the comparision of algorithms for your specific problem at hand easy and straightforward. Nash 2014, R. Varadhan, Borchers, and M. R. Varadhan 2011, Bates et al. 2014 and Wright 2010 explain the best practice optimization methods in R and help to understand several algorithms implemented in *optimx*.

were age 16 or less as of the round of 1979 survey.

Schooling data is the highest grade attended and completed at the end of June. I got the data by using two revised main variables recording the enrollment status as of May 1 survey year, and highest grade completed as of May 1 survey year. Only respondents who reported as enrolled and got a higher grade in the next survey year are coded as attending school in period from October 1 to June 30. This period is also used to code the choices between the three occupations, blue-collar, white-collar and military. NLSY79 has the starting and ending date information for at most 5 civilian jobs for each respondent for each survey years. I count the total working dates of these jobs and consider the one as an active worker if she/he worked more than half of the 9 months, and his/her main occupation is the most worked one. While there is no created NLSY79 variable that identifies members of the active military forces, a simple method of identifying these individuals through 1993 is to check whether they valid skipped the first CPS question[7]. Only if respondents are of neither of the above categories, they are classified as being home.

Since the gauge used in my data cleaning is not exactly the same as Keane and Wolpin 1997's. I use table 3.1 to check whether my data cleaning release the same style as theirs did. The comparison shows that although the levels of numbers and percentages are not exactly the same, they are quite close for the most part. What is important is that this table shares the same style of distribution as theirs. For example, as age increases the number of school decrease gradually and the numbers of white-collar and blue-collar increase. The number of military has a single peak at age 21 while it is at age 20 in Keane and Wolpin 1997's.

Since the identification idea depends on choice probabilities conditional on the initial level of schooling, h_{16}. Its marginal distribution is given by table 3.2. Initial schooling concentrates on three values, which are 8, 9 and 10 years and takes up 96.58% of the observations. Therefore I classify the initial level of schooling into only three groups, where group 1 is $4 \leqslant h_{i16} \leqslant 8$, group 2 is $h_{i16} = 9$ and groups 3 is $10 \leqslant h_{i16} \leqslant 12$, and assume that the conditional distribution of α_i does not change within each group such that I explicitly consider only three conditional distributions of fixed effects instead of nine distributions.

3.4.1 Simplify the Output of Choice Combinations

As for the number of combinations of choices over the 10 years of those young men's early career, there are 817 distinct combinations and most of them has only

[7] This is the recommend identification method in *http://nlsinfo.org/content/cohorts/nlsy79/topical-guide/employment/military*

Table 3.1: Choice Distribution: White Males Aged 16-25

	Choice					
Age	School	Home	White-Collar	Blue-Collar	Military	Total
16	1,009	90	4	38	0	1141
	(88.43)	(7.89)	(0.35)	(3.33)	(0.00)	(100%)
17	937	127	5	69	3	1141
	(82.12)	(11.13)	(0.44)	(6.05)	(0.26)	(100%)
18	608	205	52	241	35	1141
	(53.29)	(17.97)	(4.56)	(21.12)	(3.07)	(100%)
19	360	253	96	363	69	1141
	(31.55)	(22.17)	(8.41)	(31.81)	(6.05)	(100%)
20	283	235	124	419	80	1141
	(24.80)	(20.60)	(10.87)	(36.72)	(7.01)	(100%)
21	246	234	137	440	84	1141
	(21.56)	(20.51)	(12.01)	(38.56)	(7.36)	(100%)
22	185	203	172	504	77	1141
	(16.21)	(17.79)	(15.07)	(44.17)	(6.75)	(100%)
23	117	201	251	509	63	1141
	(10.25)	(17.62)	(22.00)	(44.61)	(5.52)	(100%)
24	82	179	316	509	55	1141
	(7.19)	(15.69)	(27.70)	(44.61)	(4.82)	(100%)
25	54	154	338	541	54	1141
	(4.73)	(13.50)	(29.62)	(47.41)	(4.73)	(100%)

[1] Note. - Number of observations and percentages.

Table 3.2: Initial Distribution of Schooling

	School in years									
Age	4	5	6	7	8	9	10	11	12	Total
16	2	1	8	16	112	435	555	9	3	1,141
	0.18	0.09	0.70	1.40	9.82	38.12	48.64	0.79	0.26	100%

[1] Note. - Number of observations and percentages.

one realization. This is a huge impediment to the methodology I proposed previously. Fortunately many combinations differ only in few periods and actually share some common trends. Push this idea further, I classify all the combinations into much fewer types. These types are constructed as follows[8]

1. Divide the ten years into two periods, the early five years and the late five years.

2. Choose a representative action for each periods. i.e $a = 4$ for more than 3

[8] The following rules only give a general guidance, for more details of the classification please check my code for this application.

year implies 4 is a representative action. In case $a = 1, 2, 3$ for more than 3 years, then any $1, 2, 3$ dominating in the 3 or more years is a representative action.

3. Use the simplified choice outcomes in the two five years periods as the final outcomes, then there are at most 25 outcomes and $J \leqslant 25$. Actually in the data there are 22 outcomes.

3.5 Empirical Study Result

There are many partial effects can be studied by applying (3.2.10), As a pedagogic example, in this section I am interested in partial effects of a complete high school education on the choice probability of being a white collar at the year of high school graduation. For group 1 with $4 \leqslant h_{i16} \leqslant 8$, the mode of initial schooling is 8 years. Therefore I choose a representative individual from this group the one with initial schooling of 8 years at age 16. For group 2, a reasonable representative individual is the one with initial schooling of 9 years at age 16. For group 3 with $10 \leqslant h_{i16} \leqslant 12$, the mode of initial schooling is 10 years, thus a representative individual is the one with initial schooling of 10 years. I exhibit the treatment of a complete high school for them in table 3.3.

Table 3.3: Typical Individuals and Their Treatments[1]

	Age									
Init. Sch	16	17	18	19	20	21	22	23	24	25
8 years	.	.	.	.	.	.	.	.	.	.
(8 years	5	5	5	5	.	.	.	.	.	.)
(8 years	4	4	4	4	.	.	.	.	.	.)
9 years	.	.	.	.	.	.	.	.	.	.
(8 years	5	5	5	.	.	.	.	.	.	.)
(9 years	4	4	4	.	.	.	.	.	.	.)
10 years	.	.	.	.	.	.	.	.	.	.
(8 years	5	5	.	.	.	.	.	.	.	.)
(10 years	4	4	.	.	.	.	.	.	.	.)

[1] Note. - The first line in the parenthesis presents a potential state of having not any high school education for the typical individual of each types. The second line in the parenthesis presents another potential state of having a complete high school education for the same typical individual of each types.

For the first group, the age of high school graduation is 20, and for the second group it is age 19 and for the last group it is age 18. I name the complete high school effects on white collar jobs for these three representative individuals as *Effect One*, *Effect Two* and *Effect Three*, and exhibit my findings on their bounds in table 3.4.

Table 3.4: Bounds of Treatment Effects [1] for Typical Individuals

Effects	Lower Bound	Upper Bound
One	-2.776%	4.480%
Two	-3.606%	17.144%
Three	-5.040%	15.109%

[1] Note. - The treatment effect is the change of probability of choosing a white collar job due to a complete high school education and the alternative choices are staying home for 4 years.

The findings are interesting. For average individuals who had finished 9 years of schooling at their age 16, a complete high school education seems help most. Contrarily, it helps only a little for those who had less than 9 years schooling at the same age. For people who exceeded the average initial schooling high school education helps but not as much as the average. Due to the reason of set identification, all the three effects could be negative in their worst cases.

One may consider another possibility of choosing a blue collar job instead of staying home while not taking the high school education. Since four year of blue collar job experience accumulates job-specific human capital, economics theory predicts even larger partial effects since a more experienced blue collar worker is less likely hunting a white collar occupation. To check the intuition, I consider the new treatments as listed in table 3.5, and table 3.6 shows the findings.

To my surprise, only the optimal effect of the first individual conforms the intuition, and larger negative effects in the worst situations may suggest early blue collar career experience have a more important positive effect on becoming a white collar later on than the complete high school education. To make this idea more clear, I further study the pure effects of early blue collar effect compared with staying home on becoming white collar as declared in table 3.7.

Findings on these new treatments are listed in table 3.8. It consolidates the previous finding. Generally blue collar experience has a larger positive effect on the probability of becoming a withe collar than a complete high school education. As a result a direct comparison of high school education and a blue collar experience at the same period as described in table 3.5 and 3.6 shows generally negative effects. To understand this unusual finding, it is helpful to review the utility form

Table 3.5: Typical Individuals and Their Treatments 2

	Age									
Init. Sch	16	17	18	19	20	21	22	23	24	25
8 years	.	.	.	.	.	.	.	.	.	.
(8 years	1	1	1	1	.	.	.	.	.	.)
(8 years	4	4	4	4	.	.	.	.	.	.)
9 years	.	.	.	.	.	.	.	.	.	.
(8 years	1	1	1	.	.	.	.	.	.	.)
(9 years	4	4	4	.	.	.	.	.	.	.)
10 years	.	.	.	.	.	.	.	.	.	.
(8 years	1	1	.	.	.	.	.	.	.	.)
(10 years	4	4	.	.	.	.	.	.	.	.)

[1] Note. - The first line in the parenthesis presents a potential state (as blue collar worker) of having not any high school education for the typical individual of each types. The second line in the parenthesis presents another potential state of having a complete high school education for the same typical individual of each types.

Table 3.6: Bounds of Treatment Effects 2 for Typical Individuals

Effects	Lower Bound	Upper Bound
One	-29.609%	8.032%
Two	-34.765%	11.901%
Three	-23.360%	12.593%

[1] Note. - The treatment effect is the change of probability of choosing a white collar job due to a complete high school education and the alternative choices are working as blue collar workers.

for $a = 1, 2, 3$ in (3.2.1). Note that a quadratic term of occupation-specific working experience following Mincer 1958 is included in the skill production function, therefore it is possible that the quadratic term makes the effect of four years blue collar experience negative for some parameters. It is clear that the data I use could not exclude there parameters according to the idea of set identification. Therefore it is superficial to say that blue collar experience helps to increase the possibility of being a white collar. Instead it is more precise to say that in some cases individual's blue collar experience fails to build up his skill as expected. Unless it is strongly believed that more experience can never reduce its output of skill, the

Table 3.7: Typical Individuals and Their Treatments 3

	Age									
Init. Sch	16	17	18	19	20	21	22	23	24	25
8 years	.	.	.	.	.	.	.	.	.	.
(8 years	5	5	5	5	.	.	.	.	.	.)
(8 years	1	1	1	1	.	.	.	.	.	.)
9 years	.	.	.	.	.	.	.	.	.	.
(8 years	5	5	5	.	.	.	.	.	.	.)
(8 years	1	1	1	.	.	.	.	.	.	.)
10 years	.	.	.	.	.	.	.	.	.	.
(8 years	5	5	.	.	.	.	.	.	.	.)
(8 years	1	1	.	.	.	.	.	.	.	.)

[1] Note. - In this table the states before treatments are from the pre-treatment states of table 3.3, and the states after treatments are from the pre-treatment states of table 3.5.

quadratic term should be included in our model specification[9].

Another interesting finding is that while blue collar experience helps[10] individual with average initial schooling the most, it also help people with less initial schooling a lot.

Table 3.8: Bounds of Treatment Effects 3 for Typical Individuals

Effects	Lower Bound	Upper Bound
One	-10.234%	27.994%
Two	-6.821%	32.727%
Three	-4.549%	21.367%

[1] Note. - The treatment effect is the change of probability of choosing a white collar job due to a blue collar experience instead of staying home.

Another finding is that the ranges of effects are quite large. They cover some negative effects while the majorities are positive. This is implied by the set identification, but there are some other reasons. For example, assumption 4.3 in Huang 2015b guarantees the consistency of the estimator but there are still a lot of freedom in choosing the values of ξ_n, λ_n and M given the sample size, it is still unknown what is the most efficient practice. Another reason is I use a method to

[9] Another result without the quadratic term will be reported soon.

[10] This is only a convenient phrase. Keep in mind the real reason why blue collar experience sometimes increase the probability of being a white collar.

simplify the output of choice combinations as I declare in section 3.4.1. This simplification is a double-edged sword. While it decrease the variation of frequencies and reduce the burden of calculation largely, it also throws away some data information. An improvement is to find a better balance in between data information and burden of calculation. For example it could be tried to have a division of three periods instead of two in the process of output simplification, and in this case there are no more than 125 possible combination of outcomes instead of only 25. Given more computation resource and time, a better estimated range could be achieved.

Although the estimation could be improved somehow, my result suggests that allowing for heterogeneity in a flexible way is important. For example, in this application the partial effect of being a white collar by blue collar experience could be positive largely because of the partial identification of parameters, which in turn invokes us to think about whether a quadratic term of experience should be included.

3.6 Conclusion

This paper is the first application of the simulation based method proposed by Huang 2015b and Huang 2015a. It bridges the single agent dynamic programming discrete choice model with discrete choice model with fixed effects. Instead of point identification, the simulation based method follows the idea of set identification. To handle the complexity caused by fixed effects I use a reduced form for agent's expectation over future utility, and this practice avoids the solution of the dynamic programming problem for every trail of parameters. I also gave reasons on why this reduced form practice is reasonable except for the reason of being practical. This application bases itself on the work of Keane and Wolpin 1997. To study the ten years of white young men's career choices with the set identified idea, a two five-years periods simplification is introduced. It reduces the number of distinct histories of choice thus avoids imprecisions of a lot of small sample frequencies and it also reduces the burden of calculation asked by the method. It also has its own drawbacks and the most serious one is the data information throwing away, which is a reason for the wide bounds of estimation. One needs to find a good balance between its advantages and drawbacks.

Besides the demonstration of the new method, the main finding are also very interesting. I use the change of probability of having a white collar position at the age of high school graduation as the measure of those young men's effects of previous behaviors. I find that both high school education and blue collar experience help them to increase the probability of becoming white collars. Generally the effect of blue collar experience is larger than high school education, and their distributions over initial schooling levels are quite different. While both high school education and blue collar experience help individual with average initial schooling level the most, high school education seems fail to help individuals with less than average initial schooling and blue collar experience helps them a lot. High school education does help individuals with higher initial schooling but not as much as it does for the averages. Blue collar experience also helps individuals with higher initial schooling, but can not be compared with it effects on other two types.

The same method can be used to study effects on other sex or racial groups, such that comparison across sexes and races can be made.

Appendix A

Mathematics for Chapter 1

Lemma 1.3.1: If Assumptions 1.2.2 and 1.2.3 are satisfied and $\mathcal{P}_j^k(\alpha,\beta)$ is a measurable function of α for each $\beta \; / \; \mathbb{B}$, where $\mathbb{B}$ is the parameter space, then for each β and every CDF F_k on Υ, there is a discrete distribution F_k^J with no more than J support points such that $\int \mathcal{P}_j^k(\alpha,\beta)dF_k^J(\alpha) = \int \mathcal{P}_j^k(\alpha,\beta)dF_k(\alpha)$ $(j = 1, \times\times\times, J)$. If, in addition, $\Delta(\alpha,\beta)$ is bounded for each β, then Δ_u^k and Δ_l^k are not affected by restricting attention to $F_k \; / \; \mathcal{G}_k(\beta)$ that are discrete with no more than J support points. Similarly if $\Delta\mathcal{P}(\alpha,\beta)$ is bounded for each β, then $\Delta_u^k P$ and $\Delta_l^k P$ are not affected by restricting attention to $F_k \; / \; \mathcal{G}_k(\beta)$ that are discrete with no more than J support points either.

Proof: See *Proof of Lemma 7* in Chernozhukov et al. 2013b, one important thing is that when we are considering the bounds of Δ^k and $\Delta^k P$, we can not use the same $F_{k\beta}^{J+1}$ for both purposes since this is not implied by the proof. ■

Lemma 2.4.1: Estimator $\tilde{\mathcal{P}}_j^k(\alpha,\beta)$ defined in (1.3.13) uniformly converges to $\mathcal{P}_j^k(\alpha,\beta)$ over $\Upsilon * \; \mathbb{B}$ in probability if assumption (2.4.1) holds. And

$$\bar{r}(\tilde{\mathcal{P}}_j^k(\alpha,\beta) \quad \mathcal{P}_j^k(\alpha,\beta)) \rightsquigarrow G(\alpha,\beta),$$

where $G(\alpha,\beta)$ is a mean zero Gaussian process, and its finite dimensional distribution is controlled by the distribution of $\tilde{\epsilon}_i$, thus by $H(\epsilon)$.

Proof: We only need to prove the second statement. That is to prove $\}\mathbb{1}(\tilde{Y}_i = \mathcal{Z}^j)|_{\alpha \,/\Upsilon,\beta\,/\mathbb{B}}$ is Donsker. To see this more clear, we rewrite the indicator function $\mathbb{1}(\tilde{Y}_i = \mathcal{Z}^j)$ as:

$$f_{\alpha,\beta}(\tilde{\epsilon}_i) = \prod_{t=1}^{T} \prod_{a_t^{\in} \neq a_t} \mathbb{1}\Big)\tilde{\epsilon}_{ita_t^{\in}} \quad \tilde{\epsilon}_{ita_t} \leqslant V(\mathcal{Y}_{ta_t}^k, \alpha_{a_t}, \beta) \quad V(\mathcal{Y}_{ta_t^{\in}}^k, \alpha_{a_t^{\in}}, \beta)\Big(.$$

This is what we did in (1.3.14), and now it is clear that indicator function $\mathbb{1}(\tilde{Y}_i = \mathcal{Z}^j)$ is actually a function of $\tilde{\epsilon}_i$ indexed by $(\alpha,\beta) \; / \; \Upsilon * \; \mathbb{B}$. So we need to prove

$\}f_{\alpha,\beta}|\ _{(\alpha,\beta)\,/\Upsilon *\ \mathbb{B}}$ is Donsker. Define a new class of functions with its element as:

$$h_{\alpha,\beta}(\tilde{\epsilon}_i) = \mathbb{1}\Big)\tilde{\epsilon}_{ita_t^{\in}} \quad \tilde{\epsilon}_{ita_t} \leqslant V(\mathcal{Y}^k_{ta_t}, \alpha_{a_t}, \beta) \quad V(\mathcal{Y}^k_{ta_t^{\in}}, \alpha_{a_t^{\in}}, \beta)\Big(.$$

Note that $\}\mathbb{1}(x_i \leqslant t)|\ _{t\,/\mathbb{R}}$ is P-Donsker for any P, and the newly defined class is actually a subset of $(\tilde{\epsilon}_{ita_t^{\in}} \quad \tilde{\epsilon}_{ita_t})$'s indicator functions. So $\}h_{\alpha,\beta}|\ _{(\alpha,\beta)\,/\Upsilon *\ \mathbb{B}}$ is Donsker and also a bounded class since this is a indicator function class. Furthermore, $\}f_{\alpha,\beta}|\ _{(\alpha,\beta)\,/\Upsilon *\ \mathbb{B}}$ is a subset of the product set which consists of production of functions from Donsker classes, $\}h_{\alpha,\beta}|$. By Kosorok 2007, $\}f_{\alpha,\beta}|\ _{(\alpha,\beta)\,/\Upsilon *\ \mathbb{B}}$ is Donsker. ■

Lemma 2.4.2: Estimator $\tilde{\Delta}(\alpha, \beta)$ defined in equation (1.3.16) uniformly converges to $\Delta(\alpha, \beta)$ over $\Upsilon *\ \mathbb{B}$ in probability if assumptions (2.4.1) holds.

Proof: First we prove that for any t,

$$\text{(A.0.1)} \qquad \sup_{(\alpha,\beta)\,/\Upsilon *\ \mathbb{B}} \left\| \frac{1}{r}\sum_{i=1}^{r} g_0(x^a_t, \alpha, \tilde{\epsilon}_{it}, \beta) \quad E\left[g_0(x^a_t, \alpha, \tilde{\epsilon}_{it}, \beta)\right] \right\| /^{P}\ 0.$$

We treat $g_0(x^a_t, \alpha, \tilde{\epsilon}_{it}, \beta)$ as a function of $\tilde{\epsilon}_{it}$ for the given α and β, and define $\mathcal{H}= \}g_0(x^a_t, \alpha, \tilde{\epsilon}_{it}, \beta) : (\alpha, \beta)\ /\ \Upsilon *\ \mathbb{B}|$. (A.0.1) is actually to claim that the class of functions $\mathcal{H}$is θ-Glivenko-Cantelli.

Define $\mathcal{H}$s envelope as $G(\epsilon) = \sup_{g\,/\mathcal{H}} \|g(\epsilon)\|, \epsilon\ /\ \mathbb{R}^A$, and we have $\|G(\epsilon)\| \leqslant Q$, where Q is the bound for the choice set since our choice set is finite such that we can exclude the unbounded choices. Then $\int G(\epsilon)dH \leqslant \int QdH = Q < \infty$, where H is the distribution of ϵ_{it}. By corollary 4.4.3 of Geer 2006 we only need to prove that $\mathcal{H}$is Vapnik-Chervonenkis. That is to say we need to prove $\}$subgraph$(g) : g\ /\ \mathcal{H}|$ forms a VC class.

Note that any function $g\ /\ \mathcal{H}$takes the form

$$\text{(A.0.2)} \qquad \begin{aligned} g_0(x^a_t, \alpha, \tilde{\epsilon}_{it}, \beta) &= \sum_{c\,/\mathcal{F}} c\mathbb{1}\left(g_0(x^a_t, \alpha, \tilde{\epsilon}_{it}, \beta) = c\right) \\ &= \sum_{c\,/\mathcal{F}} c \prod_{c^{\in}\forall\!\!\!= c} \mathbb{1}\left(\tilde{\epsilon}_{itc^{\in}} \quad \tilde{\epsilon}_{itc} \leqslant V(x^a_{tc}, \alpha_c, \beta) \quad V(x^a_{tc^{\in}}, \alpha_{c^{\in}}, \beta)\right). \end{aligned}$$

Define a set for each $c\ /\ \mathcal{F}$ as follows:

$$\text{(A.0.3)} \quad \mathbb{R}^A_c = \Big\}\tilde{\epsilon}_{it}\ /\ \mathbb{R}^A : \tilde{\epsilon}_{itc^{\in}} \quad \tilde{\epsilon}_{itc} \leqslant V(x^a_{tc}, \alpha_c, \beta) \quad V(x^a_{tc^{\in}}, \alpha_{c^{\in}}, \beta), \exists\ c^{\in}\forall\!\!\!= c \quad ,$$

and a set for each pair $(c, c^{\in})$ as follows:

$$\text{(A.0.4)} \qquad \mathbb{R}^A_{cc^{\in}} = \Big\}\tilde{\epsilon}_{it}\ /\ \mathbb{R}^A : \tilde{\epsilon}_{itc^{\in}} \quad \tilde{\epsilon}_{itc} \leqslant V(x^a_{tc}, \alpha_c, \beta) \quad V(x^a_{tc^{\in}}, \alpha_{c^{\in}}, \beta) \quad .$$

It is easy to see we have the following relation: $\mathbb{R}_c^A = \underset{c\in\not\forall c}{\{} \mathbb{R}_{cc^{\in}}^A$

We can write the subgraph of function g as

$$\text{(A.0.5)} \qquad \text{subgraph}(g) = \bigcap_{c\,/\,\mathcal{F}} (\quad \infty, c] * \mathbb{R}_c^A.$$

Therefore we should prove that $\Big\} \bigcap_{c\,/\,\mathcal{F}} (\quad \infty, c] * \mathbb{R}_c^A \Big(_{(\alpha,\beta)\,/\,\Upsilon *\, \mathbb{B}}$ is a VC class.

It can be easily proved that $\}\mathbb{R}_{cc^{\in}}^{A\,|}{}_{(\alpha,\beta)\,/\,\Upsilon *\, \mathbb{B}}$ is VC class because its VC dimension is bounded by a finite integer. Then by lemma 4.3.2 of Geer 2006, $\}\mathbb{R}_c^{A\,|}{}_{(\alpha,\beta)\,/\,\Upsilon *\, \mathbb{B}}$ is also VC. If $\}(\quad \infty, c] * \mathbb{R}_c^{A\,|}{}_{(\alpha,\beta)\,/\,\Upsilon *\, \mathbb{B}}$ is a VC class for any $c \,/\, \mathcal{F}$, then $\Big\} \bigcap_{c\,/\,\mathcal{F}} (\quad \infty, c] * \mathbb{R}_c^A \Big(_{(\alpha,\beta)\,/\,\Upsilon *\, \mathbb{B}}$ is a VC class by using lemma 4.3.2 of Geer 2006 again, and it is obvious that the VC dimension of $\}(\quad \infty, c] * \mathbb{R}_c^{A\,|}{}_{(\alpha,\beta)\,/\,\Upsilon *\, \mathbb{B}}$ is finite as long as the VC dimension of $\}\mathbb{R}_c^{A\,|}{}_{(\alpha,\beta)\,/\,\Upsilon *\, \mathbb{B}}$ is finite. So we have proved $\mathcal{H}$is VC and hence a Glivenko-Cantelli class. That is to say (A.0.1) holds.

Similarly we can show that

$$\text{(A.0.6)} \qquad \sup_{(\alpha,\beta)\,/\,\Upsilon *\, \mathbb{B}} \left\| \frac{1}{r} \sum_{i=1}^{r} g_0(x_t^b, \alpha, \tilde{\epsilon}_{it}, \beta) \quad E\Big] g_0(x_t^b, \alpha, \tilde{\epsilon}_{it}, \beta) \Big\{ \right\| \prime^{P} \; 0$$

holds for any t.

We have

$$\text{(A.0.7)} \qquad \begin{aligned} & \left\| \tilde{\Delta}(\alpha, \beta) \quad \Delta(\alpha, \beta) \right\| \\ = & \left\| \frac{1}{T} \sum_{t=1}^{T} \tilde{\Delta}_t(\alpha, \beta) \quad E\Big] g_0(x_t^a, \alpha, \epsilon_{it}, \beta) \quad g_0(x_t^b, \alpha, \epsilon_{it}, \beta) \Big\{ \right\| \\ = & \left\| \frac{1}{T} \sum_{t=1}^{T} \Big\} \tilde{\Delta}_t(\alpha, \beta) \quad E\Big] g_0(x_t^a, \alpha, \epsilon_{it}, \beta) \quad g_0(x_t^b, \alpha, \epsilon_{it}, \beta) \Big\{ \quad \right\| \\ \leqslant & \frac{1}{T} \sum_{t=1}^{T} \Bigg\} \left\| \frac{1}{r} \sum_{i=1}^{r} g_0(x_t^a, \alpha, \tilde{\epsilon}_{it}, \beta) \quad E\left[g_0(x_t^a, \alpha, \epsilon_{it}, \beta) \right] \right\| \Bigg\lceil \\ + & \frac{1}{T} \sum_{t=1}^{T} \Bigg\} \left\| \frac{1}{r} \sum_{i=1}^{r} g_0(x_t^b, \alpha, \tilde{\epsilon}_{it}, \beta) \quad E\Big] g_0(x_t^b, \alpha, \epsilon_{it}, \beta) \Big\{ \right\| \Bigg\lceil, \end{aligned}$$

where the second equation follows definition (1.2.19).

Combine (A.0.1), (A.0.6) and (A.0.7) together and note the fact $\tilde{\epsilon}_i$ is generated from ϵ_i's distribution, we have

$$\text{(A.0.8)} \qquad \sup_{(\alpha,\beta)\,/\,\Upsilon *\, \mathbb{B}} \left\| \tilde{\Delta}(\alpha, \beta) \quad \Delta(\alpha, \beta) \right\| \prime^{P} \; 0. \quad \blacksquare$$

Lemma 1.4.3: Estimator $\tilde{\Delta}\mathcal{P}(\alpha,\beta)$ defined in equation (1.3.18) uniformly converges to $\Delta\mathcal{P}(\alpha,\beta)$ over $\Upsilon * \mathbb{B}$ in probability if assumptions (2.4.1) holds.

Proof: This proof follows the same idea as the proof of lemma 2.4.2 with some minor differences. First of all, note that

$$\begin{aligned}&\left\|\tilde{\Delta}\mathcal{P}(\alpha,\beta) \quad \Delta\mathcal{P}(\alpha,\beta)\right\| \\ &\leqslant \frac{1}{T}\sum_{t=1}^{T}\Big\}\Big\|\frac{1}{r}\sum_{i=1}^{r}\mathbb{1}\left(g_0(x_t^a,\alpha,\tilde{\epsilon}_{it},\beta)=a_t\right) \quad \mathcal{P}_{a_t}^{x_t^a}(\alpha,\beta)\Big\|\Big\lceil \\ &+\frac{1}{T}\sum_{t=1}^{T}\Big\}\Big\|\frac{1}{r}\sum_{i=1}^{r}\mathbb{1}\Big)g_0(x_t^b,\alpha,\tilde{\epsilon}_{it},\beta)=a_t\Big(\quad \mathcal{P}_{a_t}^{x_t^b}(\alpha,\beta)\Big\|\Big\lceil,\end{aligned} \tag{A.0.9}$$

and we need to check whether

$$\sup_{(\alpha,\beta)\,/\Upsilon*\,\mathbb{B}}\left\|\frac{1}{r}\sum_{i=1}^{r}\mathbb{1}\left(g_0(x_t,\alpha,\tilde{\epsilon}_{it},\beta)=a_t\right) \quad \mathcal{P}_{a_t}^{x_t}(\alpha,\beta)\right\| \prime^{P} \ 0 \tag{A.0.10}$$

holds for any t. Note $\mathcal{P}_{a_t}^{x_t}(\alpha,\beta)=E_{\epsilon_{it}}\left[\mathbb{1}\left(g_0(x_t,\alpha,\epsilon_{it},\beta)=a_t\right)\right]$, (A.0.10) is to ask

$\}\mathbb{1}\left(g_0(x_t,\alpha,\tilde{\epsilon}_{it},\beta)=a_t\right)|_{\ (\alpha,\beta)\,/\Upsilon*\,\mathbb{B}}$ is Glivenko-Cantelli. Note that

$$\begin{aligned}&\mathbb{1}\left(g_0(x_t,\alpha,\tilde{\epsilon}_{it},\beta)=a_t\right) \\ &=\prod_{a_t^{\in}\not=a_t}\mathbb{1}\Big)V(x_{ta_t^{\in}},\alpha_{a_t^{\in}},\beta)+\tilde{\epsilon}_{ita_t^{\in}}\leqslant V(x_{ta_t},\alpha_{a_t},\beta)+\tilde{\epsilon}_{ita_t}\Big(\\ &=\prod_{a_t^{\in}\not=a_t}\mathbb{1}\Big)\tilde{\epsilon}_{ita_t^{\in}} \quad \tilde{\epsilon}_{ita_t}\leqslant V(x_{ta_t},\alpha_{a_t},\beta) \quad V(x_{ta_t^{\in}},\alpha_{a_t^{\in}},\beta)\Big(\end{aligned} \tag{A.0.11}$$

and we can use the same idea of subgraph to prove (A.0.10) holds for any t. Then by (A.0.9), we have

$$\sup_{(\alpha,\beta)\,/\Upsilon*\,\mathbb{B}}\left\|\tilde{\Delta}\mathcal{P}(\alpha,\beta) \quad \Delta\mathcal{P}(\alpha,\beta)\right\| \prime^{P} \ 0. \quad \blacksquare \tag{A.0.12}$$

Lemma 1.4.4: For every $\bar{\pi} \ / \ \mathcal{X}_M^K$, where $M > J$, there exists

$$\phi(\beta,\bar{\pi})=\Big)\beta^{\in},\gamma^1(\beta,\bar{\pi}^1)^{\in},\times\times\times,\gamma^K(\beta,\bar{\pi}^K)^{\in}\Big(^{\in}$$

such that

$$\begin{aligned}\tilde{P}_j^k(\beta,\bar{\pi}^k,M)&=\tilde{P}_j^k\Big)\beta,\gamma^k(\beta,\bar{\pi}^k)\Big(\\ \tilde{\Delta}^k(\beta,\bar{\pi}^k,M)&=\tilde{\Delta}^k\Big)\beta,\gamma^k(\beta,\bar{\pi}^k)\Big(,\end{aligned}$$

for all $j = 1, \times\times\times, J$ and $k = 1, \times\times\times, K$, and where

$$\tilde{P}^k_j \Big)\beta, \gamma^k(\beta, \bar{\pi}^k)\Big(= \sum_{l=1}^{J+1} \pi^k_l \tilde{\mathcal{P}}^k_j(\bar{\alpha}_{m^k_l M}, \beta)$$
$$\tilde{\Delta}^k \Big)\beta, \gamma^k(\beta, \bar{\pi}^k)\Big(= \sum_{l=1}^{J+1} \pi^k_l \tilde{\Delta}(\bar{\alpha}_{m^k_l M}, \beta),$$

and $\}m^k_l|_{l=1}^{J+1}$ is a subset of $\}1, \times\times\times, M|$, which can be different for different k.

Proof: Denote $\Gamma^{\Delta}_{kM}(\beta) = \Big\}\Big)\tilde{\mathcal{P}}^k(\bar{\alpha}, \beta)^{\in}, \tilde{\Delta}(\bar{\alpha}, \beta)\Big(^{\in} : \bar{\alpha} \;/\; \Upsilon_M\Big($, where $\tilde{\mathcal{P}}^k(\bar{\alpha}, \beta)^{\in} = \Big\}\tilde{\mathcal{P}}^k_j(\bar{\alpha}, \beta) \quad {}_{j=1}^{J}$, and $\check{\Gamma}^{\Delta}_{kM}(\beta)$ is the convex hull of $\Gamma^{\Delta}_{kM}(\beta)$. Then by lemma 3 of Chamberlain 1987, we have:

$$\check{\Gamma}^{\Delta}_{kM}(\beta) = \Bigg\}\Bigg)\Bigg)\Big[\ \ \tilde{\mathcal{P}}^k(\alpha, \beta)dF(\alpha)\Bigg[^{\in}, \Big[\ \ \tilde{\Delta}(\alpha, \beta)dF(\alpha)\Bigg[^{\in} : F \text{ is a cdf on } \Upsilon_M \Bigg\lceil .$$

Therefore for any $\bar{\pi} \;/\; \mathcal{X}^K_M$,

$$\Bigg)\Bigg)\sum_{m=1}^{M} \tilde{\mathcal{P}}^k(\bar{\alpha}_{mM}, \beta)\bar{\pi}^k_m \Bigg[^{\in}, \sum_{m=1}^{M} \tilde{\Delta}(\bar{\alpha}_{mM}, \beta)\bar{\pi}^k_m \Bigg[^{\in} \;/\; \check{\Gamma}^{\Delta}_{kM}(\beta).$$

Note that $\tilde{\mathcal{P}}^k(\bar{\alpha}, \beta) \;/\; \mathcal{X}_J$, then by Caratheodory Theorem there exists a discrete distribution F^{J+1}_k with $J+1$ support points $\Big)\bar{\alpha}_{m^k_1 M}, \times\times\times, \bar{\alpha}_{m^k_{J+1} M}\Big(^{\in} \to \Upsilon_M$ and probability $(\pi^k_1, \times\times\times, \pi^k_{J+1})$ such that

$$\begin{aligned}
&\Big)\Big)\tilde{P}^k(\beta, \bar{\pi}^k, M)\Big(^{\in}, \tilde{\Delta}^k(\beta, \bar{\pi}^k, M)\Big[^{\in} \\
&= \Bigg)\Bigg)\sum_{m=1}^{M} \tilde{\mathcal{P}}^k(\bar{\alpha}_{mM}, \beta)\bar{\pi}^k_m \Bigg[^{\in}, \sum_{m=1}^{M} \tilde{\Delta}(\bar{\alpha}_{mM}, \beta)\bar{\pi}^k_m \Bigg[^{\in} \\
&= \Bigg)\Bigg)\sum_{l=1}^{J+1} \tilde{\mathcal{P}}^k(\bar{\alpha}_{m^k_l M}, \beta)\pi^k_l \Bigg[^{\in}, \sum_{l=1}^{J+1} \tilde{\Delta}(\bar{\alpha}_{m^k_l M}, \beta)\pi^k_l \Bigg[^{\in} \\
&= \Big)\Big)\tilde{P}^k\Big)\beta, \gamma^k(\beta, \bar{\pi}^k)\Big(\Big(^{\in}, \tilde{\Delta}^k\Big)\beta, \gamma^k(\beta, \bar{\pi}^k)\Big(\Big[^{\in},
\end{aligned}$$

where the last equation is the definition. ■

Lemma 1.4.5: There exists a constant C such that for all $\phi, \phi^{\in} \;/\; \Phi$, the following inequality holds: $\|\tilde{\Delta}^k(\phi) \quad \tilde{\Delta}^k(\phi^{\ni})\| \leqslant Cd(\phi, \phi^{\ni}) + o_P(1)$, where $o_P(1)$ doesn't depend on $(\phi, \phi^{\ni})$.

Proof:

$$\begin{aligned}
\left\|\tilde{\Delta}^k(\phi) \quad \tilde{\Delta}^k(\phi^{\ni})\right\| &= \left\|\sum_{l=1}^{J+1}\tilde{\Delta}(\alpha_l^k,\beta)\pi_l^k \quad \sum_{l=1}^{J+1}\tilde{\Delta}(\alpha^{\notin k}_{\ l},\beta^{\ni})\pi^{\notin k}_{\ l}\right\| \\
&\leqslant \sum_{l=1}^{J+1}\left\|\tilde{\Delta}(\alpha_l^k,\beta) \quad \Delta(\alpha_l^k,\beta)\right\|\pi_l^k \\
&+ \sum_{l=1}^{J+1}\left\|\tilde{\Delta}(\alpha^{\notin k}_{\ l},\beta^{\ni}) \quad \Delta(\alpha^{\notin k}_{\ l},\beta^{\ni})\right\|\pi^{\notin k}_{\ l} \\
&+ \sum_{l=1}^{J+1}\left\|\Delta(\alpha_l^k,\beta)\pi_l^k \quad \Delta(\alpha^{\notin k}_{\ l},\beta^{\ni})\pi^{\notin k}_{\ l}\right\| \\
&\leqslant \sum_{l=1}^{J+1}\sup_{\alpha,\beta}\left\|\tilde{\Delta}(\alpha_l^k,\beta) \quad \Delta(\alpha_l^k,\beta)\right\|\pi_l^k \\
&+ \sum_{l=1}^{J+1}\sup_{\alpha,\beta}\left\|\tilde{\Delta}(\alpha^{\notin k}_{\ l},\beta^{\ni}) \quad \Delta(\alpha^{\notin k}_{\ l},\beta^{\ni})\right\|\pi^{\notin k}_{\ l} \\
&+ \sum_{l=1}^{J+1}\left\|\Delta(\alpha_l^k,\beta)\pi_l^k \quad \Delta(\alpha^{\notin k}_{\ l},\beta^{\ni})\pi^{\notin k}_{\ l}\right\| \\
&= o_P(1) + \sum_{l=1}^{J+1}\left\|\Delta(\alpha_l^k,\beta)\pi_l^k \quad \Delta(\alpha^{\notin k}_{\ l},\beta^{\ni})\pi^{\notin k}_{\ l}\right\| \\
&\leqslant \sum_{l=1}^{J+1}\left\|\Delta(\alpha_l^k,\beta)\right\|\left\|\pi_l^k \quad \pi^{\notin k}_{\ l}\right\| \\
&+ \sum_{l=1}^{J+1}\left\|\Delta(\alpha_l^k,\beta) \quad \Delta(\alpha^{\notin k}_{\ l},\beta^{\ni})\right\|\pi^{\notin k}_{\ l} + o_P(1) \\
&\leqslant Cd(\phi,\phi^{\ni}) + \sum_{l=1}^{J+1}L[d(\alpha_l^k,\alpha^{\notin k}_{\ l}) \\
&+ (\beta \quad \beta^{\in}(]\pi^{\notin k}_{\ l} + o_P(1) \\
&= Cd(\phi,\phi^{\ni}) + o_P(1),
\end{aligned}$$

where the last inequality follows by Assumption 1.4.1. ■

Lemma 1.4.6: Denote Hausdorff metric as d_H, Lemma 1.4.5 implies

$$d_H\Big)\tilde{\Delta}^k(\Phi_s),\tilde{\Delta}^k(\Phi_s^{\in})\Big(\leqslant Cd_H(\Phi_s,\Phi_s^{\ni}) + o_P(1),$$

where Φ_s and $\Phi_s^{\in}$ are two subsets of Φ and $o_P(1)$ doesn't depend on $(\Phi_s,\Phi_s^{\ni})$.

Proof: By the definition of Hausdorff metric,

$$
\begin{aligned}
& d_H(\tilde{\Delta}^k(\Phi_s), \tilde{\Delta}^k(\Phi_s^{\in})) \\
& = \max \Big\} \sup_{\phi / \Phi_s} \inf_{\phi^{\in} / \Phi_s^{\in}} \left\| \tilde{\Delta}^k(\phi) \quad \tilde{\Delta}^k(\phi^{\in}) \right\|, \sup_{\phi^{\in} / \Phi_s^{\in}} \inf_{\phi / \Phi_s} \left\| \tilde{\Delta}^k(\phi^{\in}) \quad \tilde{\Delta}^k(\phi) \right\| \Big\lceil .
\end{aligned}
$$

If $\|\tilde{\Delta}^k(\phi) \quad \tilde{\Delta}^k(\phi^{\in})\| \leqslant Cd(\phi, \phi^{\in}) + o_P(1)$ holds for all $\phi, \phi^{\in} / \Phi$, and $o_P(1)$ doesn't depend on $(\phi, \phi^{\in})$, it can be shown that

$$
\sup_{\phi / \Phi_s} \inf_{\phi^{\in} / \Phi_s^{\in}} \left\| \tilde{\Delta}^k(\phi) \quad \tilde{\Delta}^k(\phi^{\in}) \right\| \leqslant \sup_{\phi / \Phi_s} \inf_{\phi^{\in} / \Phi_s^{\in}} Cd(\phi, \phi^{\in}) + o_P(1),
$$

and

$$
\sup_{\phi^{\in} / \Phi_s^{\in}} \inf_{\phi / \Phi_s} \left\| \tilde{\Delta}^k(\phi^{\in}) \quad \tilde{\Delta}^k(\phi) \right\| \leqslant \sup_{\phi^{\in} / \Phi_s^{\in}} \inf_{\phi / \Phi_s} Cd(\phi^{\in}, \phi) + o_P(1).
$$

So that

$$
\begin{aligned}
& d_H \Big) \tilde{\Delta}^k(\Phi), \tilde{\Delta}^k(\Phi^{\in}) \Big(\\
& \leqslant \max \Big\} \sup_{\phi / \Phi_s} \inf_{\phi^{\in} / \Phi_s^{\in}} Cd(\phi, \phi^{\in}) + o_P(1), \sup_{\phi^{\in} / \Phi_s^{\in}} \inf_{\phi / \Phi_s} Cd(\phi^{\in}, \phi) + o_P(1) \Big\lceil \\
& = C \max \Big\} \sup_{\phi / \Phi_s} \inf_{\phi^{\in} / \Phi_s^{\in}} d(\phi, \phi^{\in}), \sup_{\phi^{\in} / \Phi_s^{\in}} \inf_{\phi / \Phi_s} d(\phi^{\in}, \phi) \Big\lceil + o_P(1) \\
& = C d_H(\Phi_s, \Phi_s^{\in}) + o_P(1) \quad \blacksquare
\end{aligned}
$$

Lemma 1.4.7: If $d_H \;\; \hat{\Phi}, \Phi_I \big[\;\; \overset{p}{/} \;\; 0$, we have

$$
\hat{\Delta}_l^k \;\; \overset{p}{/} \;\; \Delta_l^k \text{ and } \hat{\Delta}_u^k \;\; \overset{p}{/} \;\; \Delta_u^k.
$$

Proof: Since the minimum and maximum of a set are continuous in the Hausdorff metric, it is sufficient to show

$$
d_H \Big) \hat{D}^k, D^k \Big(\;\; \overset{p}{/} \;\; 0. \tag{A.0.13}
$$

Note that

$$
\begin{aligned}
d_H \Big) \hat{D}^k, D^k \Big(&= d_H \Big) \tilde{\Delta}^k(\hat{\Phi}), \Delta^k(\Phi_I) \Big(\\
&\leqslant d_H \Big) \tilde{\Delta}^k(\hat{\Phi}), \tilde{\Delta}^k(\Phi_I) \Big(+ d_H \Big) \tilde{\Delta}^k(\Phi_I), \Delta^k(\Phi_I) \Big(\\
&\leqslant C d_H \;\; \hat{\Phi}, \Phi_I \big[+ o_P(1) + d_H \Big) \tilde{\Delta}^k(\Phi_I), \Delta^k(\Phi_I) \Big(.
\end{aligned} \tag{A.0.14}
$$

The first inequality follows by triangular inequality and the second inequality follows by Lemma 1.4.5 and Lemma 1.4.6. By Lemma 2.4.2 we have

$$\begin{aligned}\sup_{\phi / \Phi}\left\|\tilde{\Delta}^k(\phi) \quad \Delta^k(\phi)\right\| &= \sup_{\alpha,\beta,\pi}\left\|\sum_{l=1}^{J+1}\tilde{\Delta}(\alpha_l^k,\beta)\pi_l^k \quad \sum_{l=1}^{J+1}\Delta(\alpha_l^k,\beta)\pi_l^k\right\| \\ &\leqslant \sup_{\pi}\sum_{l=1}^{J+1}\sup_{\alpha,\beta}\left\|\tilde{\Delta}(\alpha_l^k,\beta) \quad \Delta(\alpha_l^k,\beta)\right\|\pi_l^k \\ &= \sup_{\alpha,\beta}\left\|\tilde{\Delta}(\alpha_l^k,\beta) \quad \Delta(\alpha_l^k,\beta)\right\| \overset{p}{\prime} \quad 0.\end{aligned} \tag{A.0.15}$$

Also note that

$$\inf_{\phi^{\in} / \Phi_I}\left\|\tilde{\Delta}^k(\phi) \quad \Delta^k(\phi^{\ni})\right\| \leqslant \left\|\tilde{\Delta}^k(\phi) \quad \Delta^k(\phi)\right\| \quad \text{for any } \phi \ / \ \Phi_I,$$

we have

$$\sup_{\phi / \Phi_I}\inf_{\phi^{\in} / \Phi_I}\left\|\tilde{\Delta}^k(\phi) \quad \Delta^k(\phi^{\ni})\right\| \leqslant \sup_{\phi / \Phi_I}\left\|\tilde{\Delta}^k(\phi) \quad \Delta^k(\phi)\right\|. \tag{A.0.16}$$

Similarly, we have

$$\sup_{\phi / \Phi_I}\inf_{\phi^{\in} / \Phi_I}\left\|\Delta^k(\phi) \quad \tilde{\Delta}^k(\phi^{\ni})\right\| \leqslant \sup_{\phi / \Phi_I}\left\|\Delta^k(\phi) \quad \tilde{\Delta}^k(\phi)\right\|. \tag{A.0.17}$$

It follows by (A.0.15) (A.0.16) and (A.0.17) that

$$d_H \Big) \tilde{\Delta}^k(\Phi_I), \Delta^k(\Phi_I) \Big(\leqslant \sup_{\phi / \Phi_I}\left\|\tilde{\Delta}^k(\phi) \quad \Delta^k(\phi)\right\| \overset{p}{\prime} \quad 0. \tag{A.0.18}$$

So if we have $d_H \quad \hat{\Phi}, \Phi_I [\overset{p}{\prime} \quad 0$, by (A.0.14) and (A.0.18) we have:

$$d_H \Big) \hat{D}^k, D^k \Big(\overset{p}{\prime} \quad 0. \quad ■$$

Lemma 1.4.12: $\sup_{\phi / \Phi}\left\|\hat{Q}(\phi) \quad Q(\phi)\right\| \prime^{P} \ 0$

Proof: First we notice that

$$\sup_{\phi / \Phi}\left\|\tilde{P}_j^k(\phi) \quad P_j^k(\phi)\right\| \leqslant \sup_{\alpha,\beta}\left\|\tilde{\mathcal{P}}_j^k(\alpha_l^k,\beta) \quad \mathcal{P}_j^k(\alpha_l^k,\beta)\right\|,$$

and lemma 2.4.1 tells us

$$\sup_{\alpha,\beta}\left\|\tilde{\mathcal{P}}_j^k(\alpha_l^k,\beta) \quad \mathcal{P}_j^k(\alpha_l^k,\beta)\right\| \prime^{P} \ 0.$$

Follow this result, we have

$$\sup_{\phi / \Phi} \left\| \tilde{P}_j^k(\phi) \quad P_j^k(\phi) \right\| \overset{P}{/} \ 0.$$

Furthermore we notice that $\left\| \tilde{P}_j^k(\phi) + P_j^k(\phi) \right\| \leqslant 2$, and we have

$$\sup_{\phi / \Phi} \left\| \Big) \tilde{P}_j^k(\phi) \Big(^2 \quad \Big) P_j^k(\phi) \Big(^2 \right\| \overset{P}{/} \ 0.$$

By some tedious calculation, we have

$$\begin{aligned} \sup_{\phi / \Phi} \left\| \hat{Q}(\phi) \quad Q(\phi) \right\| \leqslant & \sum_{j,k} \left\| \hat{\omega}_j^k \Big) \hat{\mathcal{S}}_j^k \Big(^2 \quad \omega_j^k \Big) \mathcal{S}_j^k \Big(^2 \right\| \\ & + \sum_{j,k} \sup_{\phi / \Phi} \left\| \hat{\omega}_j^k \Big) \tilde{P}_j^k(\phi) \Big(^2 \quad \omega_j^k \Big) P_j^k(\phi) \Big(^2 \right\| \\ & + 2 \sum_{j,k} \sup_{\phi / \Phi} \left\| \omega_j^k \mathcal{S}_j^k P_j^k(\phi) \quad \hat{\omega}_j^k \hat{\mathcal{S}}_j^k \tilde{P}_j^k(\phi) \right\|. \end{aligned}$$

Of all the right-hand-side items, the first one converges to 0 in probability by law of large numbers and continuous mapping theorem.

For the second part,

$$\begin{aligned} \sum_{j,k} \sup_{\phi / \Phi} \left\| \hat{\omega}_j^k \Big) \tilde{P}_j^k(\phi) \Big(^2 \quad \omega_j^k \Big) P_j^k(\phi) \Big(^2 \right\| \leqslant & \sum_{j,k} \left\| \hat{\omega}_j^k \right\| \sup_{\phi / \Phi} \left\| \Big) \tilde{P}_j^k(\phi) \Big(^2 \quad \Big) P_j^k(\phi) \Big(^2 \right\| \\ & + \sum_{j,k} \left\| \hat{\omega}_j^k \quad \omega_j^k \right\| \overset{P}{/} \ 0. \end{aligned}$$

For the third part,

$$\begin{aligned} 2 \sum_{j,k} \sup_{\phi / \Phi} \left\| \omega_j^k \mathcal{S}_j^k P_j^k(\phi) \quad \hat{\omega}_j^k \hat{\mathcal{S}}_j^k \tilde{P}_j^k(\phi) \right\| \leqslant & \, 2 \sum_{j,k} \left\| \omega_j^k \mathcal{S}_j^k \quad \hat{\omega}_j^k \hat{\mathcal{S}}_j^k \right\| \\ & + 2 \sum_{j,k} \sup_{\phi / \Phi} \left\| \tilde{P}_j^k(\phi) \quad P_j^k(\phi) \right\| \left\| \hat{\omega}_j^k \hat{\mathcal{S}}_j^k \right\| \\ & \overset{P}{/} \ 0. \end{aligned}$$

Since all the right-hand-side terms converge to 0 in probability, we prove that:

$$\sup_{\phi / \Phi} \left\| \hat{Q}(\phi) \quad Q(\phi) \right\| \overset{P}{/} \ 0. \quad \blacksquare$$

Theorem 1.4.1: Under all the assumptions, $d_H(\hat{\Phi}, \Phi_I) \overset{P}{/} \ 0$.

Proof:[1] First step, we prove that:

$$\sup_{\phi^{\in}/\hat{\Phi}} \inf_{\phi/\Phi_I} d(\phi,\phi^{\ni}) < \delta \quad \text{w.p.a.1.} \quad \text{for any } \delta > 0.$$

Define $\Phi_I^\delta = \Big\}\phi : \inf_{\phi^{\in}/\Phi_I} d(\phi,\phi^{\ni}) < \delta\Big($, and Φ_I^δ is open by its definition. So $\Phi \surd \Phi_I^\delta$ is compact. By continuity of $Q(\phi)$, $Q(\phi)$ takes its minimal value on $\Phi \surd \Phi_I^\delta$, i.e. $\inf_{\phi/\Phi \surd \Phi_I^\delta} Q(\phi) = \rho > 0$. We know that $\hat{Q}(\phi)$ uniformly converges to $Q(\phi)$ in probability by Lemma 1.4.12. So $\inf_{\phi/\Phi \surd \Phi_I^\delta} \hat{Q}(\phi) > \frac{\rho}{2}$ w.p.a.1.

By $\xi_n \quad \prime \quad 0$,

$$\sup_{\phi/\hat{\Phi}} \hat{Q}(\phi) \leqslant \sup_{\bar{\pi}} \Big\}\hat{Q}\left(\phi(\beta,\bar{\pi})\right) : \hat{Q}\left(\phi(\beta,\bar{\pi})\right) + \lambda_n \bar{\pi}^{\in}\bar{\pi} \leqslant \xi_n \Big| < \frac{\rho}{2},$$

so that $\hat{\Phi} \sim \Phi_I^\delta$. Therefore w.p.a.1 for all $\phi^{\in} / \hat{\Phi}$ there exists ϕ / Φ_I such that $d(\phi^{\in}, \phi) < \delta$, i.e. $\sup_{\phi^{\in}/\hat{\Phi}} \inf_{\phi/\Phi_I} d(\phi,\phi^{\ni}) < \delta$ w.p.a.1.

The second step is to show that:

$$\sup_{\phi/\Phi_I} \inf_{\phi^{\in}/\hat{\Phi}} d(\phi,\phi^{\ni}) < \delta \quad \text{w.p.a.1.} \quad \text{for any } \delta > 0.$$

First, by definition of $\hat{Q}(\phi)$ in (1.4.2),

$$\begin{aligned}
\hat{Q}(\phi) &= \sum_{j,k} \hat{\omega}_j^k \Big] \hat{\mathcal{S}}_j^k \quad \tilde{P}_j^k(\phi) \Big\{^2 \\
&= \sum_{j,k} \hat{\omega}_j^k \Big] \hat{\mathcal{S}}_j^k \quad P_j^k(\phi) \Big\{^2 + \sum_{j,k} \hat{\omega}_j^k \Big] P_j^k(\phi) \quad \tilde{P}_j^k(\phi) \Big\{^2 \\
&\quad + 2 \sum_{j,k} \hat{\omega}_j^k \Big] \hat{\mathcal{S}}_j^k \quad P_j^k(\phi) \Big\{ \Big] P_j^k(\phi) \quad \tilde{P}_j^k(\phi) \Big\{,
\end{aligned}$$

and by definition of Φ_I in (1.4.4),

$$\begin{aligned}
\sup_{\phi/\Phi_I} \hat{Q}(\phi) \leqslant{}& \sum_{j,k} \hat{\omega}_j^k \Big) \hat{\mathcal{S}}_j^k \quad \mathcal{S}_j^k \Big(^2 + \sup_{\phi/\Phi_I} \sum_{j,k} \hat{\omega}_j^k \Big] P_j^k(\phi) \quad \tilde{P}_j^k(\phi) \Big\{^2 \\
&+ 2 \sup_{\phi/\Phi_I} \sum_{j,k} \hat{\omega}_j^k \Big) \hat{\mathcal{S}}_j^k \quad \mathcal{S}_j^k \Big(\Big] P_j^k(\phi) \quad \tilde{P}_j^k(\phi) \Big\{.
\end{aligned}$$

It can be easily seen that the first part on the right hand side of the last inequal-

[1] The proof here is a adjustment of Chernozhukov et al. 2013a, the most important difference here is that we need to find the rate of convergence of the estimated choice probability to the true choice probability and adjust other terms' convergence rates accordingly. You can see A9.2 in Chernozhukov et al. 2013b for comparison.

ity is $O_P(n\quad^{1})$ since $\hat{\omega}_j^k = \hat{\mathcal{S}}^k/\hat{\mathcal{S}}_j^k$ and both the numerator and denominator are $O_P(n\quad^{\frac{1}{2}})$. We need to find the rate of convergence for the other two parts.

For the second part,

$$\sup_{\phi/\Phi_I}\sum_{j,k}\hat{\omega}_j^k\Big]P_j^k(\phi)\quad\tilde{P}_j^k(\phi)\Big\{^2 \leqslant \sum_{j,k}\hat{\omega}_j^k\Bigg]\sup_{\phi/\Phi_I}\Big\|\tilde{P}_j^k(\phi)\quad P_j^k(\phi)\Big\|\Bigg\{^2,$$

and $\tilde{P}_j^k(\phi)\quad P_j^k(\phi)=\sum_{l=1}^{J+1}\pi_l^k\Big]\tilde{\mathcal{P}}_j^k(\alpha_l^k,\beta)\quad\mathcal{P}_j^k(\alpha_l^k,\beta)\Big\{$ by definition.

So that

$$\sup_{\phi/\Phi_I}\sum_{j,k}\hat{\omega}_j^k\Big]P_j^k(\phi)\quad\tilde{P}_j^k(\phi)\Big\{^2 \leqslant \sum_{j,k}\hat{\omega}_j^k\Bigg]\sum_{l=1}^{J+1}\sup_{\phi/\Phi_I}\Big\|\tilde{\mathcal{P}}_j^k(\alpha_l^k,\beta)\quad\mathcal{P}_j^k(\alpha_l^k,\beta)\Big\|\Bigg\{^2.$$

By Lemma 2.4.1, we know the following process

$$\Big\}\quad\bar{r}\Big]\tilde{\mathcal{P}}_j^k(\alpha_l^k,\beta)\quad\mathcal{P}_j^k(\alpha_l^k,\beta)\Big\{:(\alpha_l^k,\beta)\ /\ \Upsilon*\mathbb{B}$$

converges in distribution to a Gaussian process. Note that taking absolute value and sup are continuous mappings between their proper spaces with proper metrics respectively.We have

$$\sup_{\phi/\Phi_I}\Big\|\quad\bar{r}\Big]\tilde{\mathcal{P}}_j^k(\alpha_l^k,\beta)\quad\mathcal{P}_j^k(\alpha_l^k,\beta)\Big\{\Big\|\rightsquigarrow\sup_{\phi/\Phi_I}\Big\|\text{Gaussian process }\mathbf{G}(\alpha_l^k,\beta)\Big\|=O_P(1).$$

Then we have

$$\sup_{\phi/\Phi_I}\sum_{j,k}\hat{\omega}_j^k\Big]P_j^k(\phi)\quad\tilde{P}_j^k(\phi)\Big\{^2\leqslant O_P(r\quad^{1}).$$

For the third part,

$$2\sup_{\phi/\Phi_I}\sum_{j,k}\hat{\omega}_j^k\Big)\hat{\mathcal{S}}_j^k\quad\mathcal{S}_j^k\Big(\Big]P_j^k(\phi)\quad\tilde{P}_j^k(\phi)\Big\{$$

$$\leqslant 2\Bigg]\sum_{j,k}\Big)\hat{\omega}_j^k\Big(^2\Big)\hat{\mathcal{S}}_j^k\quad\mathcal{S}_j^k\Big(^2\Bigg\{^{\frac{1}{2}}\Bigg\}\sum_{j,k}\Bigg]\sup_{\phi/\Phi_I}\Big\|\tilde{P}_j^k(\phi)\quad P_j^k(\phi)\Big\|\Bigg\{^2\Bigg\}^{\frac{1}{2}},$$

wherein $2\Big]\sum_{j,k}\Big)\hat{\omega}_j^k\Big(^2\Big)\hat{\mathcal{S}}_j^k\quad\mathcal{S}_j^k\Big(^2\Big\{^{\frac{1}{2}}$ is $O_P(n\quad^{\frac{1}{2}})$, and

$$\Bigg\}\sum_{j,k}\Bigg]\sup_{\phi/\Phi_I}\Big\|\tilde{P}_j^k(\phi)\quad P_j^k(\phi)\Big\|\Bigg\{^2\Bigg\}^{\frac{1}{2}}\leqslant O_P(r\quad^{\frac{1}{2}}),$$

since $\sup_{\phi / \Phi_I} \left\| \tilde{P}^k_j(\phi) \quad P^k_j(\phi) \right\| \leqslant \sup_{\alpha,\beta} \left\| \tilde{\mathcal{P}}^k_j(\alpha^k_l, \beta) \quad \mathcal{P}^k_j(\alpha^k_l, \beta) \right\| = O_P(r \;\; ^{\frac{1}{2}})$.

So we have

$$2 \sup_{\phi / \Phi_I} \sum_{j,k} \hat{\omega}^k_j \Big) \hat{\mathcal{S}}^k_j \quad \mathcal{S}^k_j \Big(\Big] P^k_j(\phi) \quad \tilde{P}^k_j(\phi) \Big\{ \leqslant O_P(n \;\; ^{\frac{1}{2}}) O_P(r \;\; ^{\frac{1}{2}}) = O_P \Big) (nr) \;\; ^{\frac{1}{2}} \Big(.$$

Finally we know

$$\sup_{\phi / \Phi_I} \hat{Q}(\phi) \leqslant O_P(n \;\; ^1) + O_P(r \;\; ^1) + O_P(n \;\; ^{\frac{1}{2}}) O_P(r \;\; ^{\frac{1}{2}}). \tag{A.0.19}$$

For any ϕ and $\phi^{\in} / \Phi$, we have

$$\begin{aligned} \left\| \tilde{\mathcal{P}}^k_j(\alpha^k_l, \beta) \quad \tilde{\mathcal{P}}^k_j(\alpha^{\notin k}_l, \beta^{\ni}) \right\| &\leqslant \left\| \tilde{\mathcal{P}}^k_j(\alpha^k_l, \beta) \quad \mathcal{P}^k_j(\alpha^k_l, \beta) \right\| + \left\| \tilde{\mathcal{P}}^k_j(\alpha^{\notin k}_l, \beta^{\ni}) \quad \mathcal{P}^k_j(\alpha^{\notin k}_l, \beta^{\ni}) \right\| \\ &\quad + \left\| \mathcal{P}^k_j(\alpha^k_l, \beta) \quad \mathcal{P}^k_j(\alpha^{\notin k}_l, \beta^{\ni}) \right\| \\ &\leqslant 2 \sup_{\phi / \Phi} \left\| \tilde{\mathcal{P}}^k_j(\alpha^k_l, \beta) \quad \mathcal{P}^k_j(\alpha^k_l, \beta) \right\| + Cd(\phi, \phi^{\ni}). \end{aligned}$$

So that

$$\begin{aligned} &\left\| \tilde{P}^k_j(\phi) \quad \tilde{P}^k_j(\phi^{\ni}) \right\| \\ &\leqslant \sum_{l=1}^{J+1} \left\| \tilde{\mathcal{P}}^k_j(\alpha^k_l, \beta) \pi^k_l \quad \tilde{\mathcal{P}}^k_j(\alpha^{\notin k}_l, \beta^{\ni}) \pi^{\notin k}_l \right\| \\ &= \sum_{l=1}^{J+1} \left\| \tilde{\mathcal{P}}^k_j(\alpha^k_l, \beta) \pi^k_l + \tilde{\mathcal{P}}^k_j(\alpha^k_l, \beta) \pi^{\notin k}_l \quad \tilde{\mathcal{P}}^k_j(\alpha^k_l, \beta) \pi^{\notin k}_l \quad \tilde{\mathcal{P}}^k_j(\alpha^{\notin k}_l, \beta^{\ni}) \pi^{\notin k}_l \right\| \\ &\leqslant \sum_{l=1}^{J+1} \Big] \left\| \tilde{\mathcal{P}}^k_j(\alpha^k_l, \beta) \right\| \left\| \pi^k_l \quad \pi^{\notin k}_l \right\| + \left\| \tilde{\mathcal{P}}^k_j(\alpha^k_l, \beta) \quad \tilde{\mathcal{P}}^k_j(\alpha^{\notin k}_l, \beta^{\ni}) \right\| \pi^{\notin k}_l \Big\{ \\ &\leqslant \sum_{l=1}^{J+1} \Bigg] d(\phi, \phi^{\ni}) + Cd(\phi, \phi^{\ni}) \pi^{\notin k}_l + 2 \Bigg) \sup_{\phi / \Phi} \left\| \tilde{\mathcal{P}}^k_j(\alpha^k_l, \beta) \quad \mathcal{P}^k_j(\alpha^k_l, \beta) \right\| \Bigg[\;\; \pi^{\notin k}_l \Bigg\{ \\ &= Cd(\phi, \phi^{\ni}) + 2 \sup_{\phi / \Phi} \left\| \tilde{\mathcal{P}}^k_j(\alpha^k_l, \beta) \quad \mathcal{P}^k_j(\alpha^k_l, \beta) \right\|, \end{aligned}$$

and

(A.0.20)

$$
\begin{aligned}
&\|\hat{Q}(\phi) \quad \hat{Q}(\phi^{\Subset})\| \\
&= \left\| \sum_{j,k} \hat{\omega}_j^k \Big] \hat{\mathcal{S}}_j^k \quad \tilde{P}_j^k(\phi) \Big\{^2 \quad \sum_{j,k} \hat{\omega}_j^k \Big] \hat{\mathcal{S}}_j^k \quad \tilde{P}_j^k(\phi^{\Subset}) \Big\{^2 \right\| \\
&= \left\| \sum_{j,k} \Big] 2\, \hat{\omega}_j^k \hat{\mathcal{S}}_j^k \Big) \tilde{P}_j^k(\phi^{\Subset}) \quad \tilde{P}_j^k(\phi) \Big(+ \hat{\omega}_j^k \Big) \Big) \tilde{P}_j^k(\phi) \Big(^2 \quad \Big) \tilde{P}_j^k(\phi^{\Subset}) \Big(^2 \Big[\Big\{ \right\| \\
&\leqslant \sum_{j,k} \Big\} 2 \left\| \hat{\omega}_j^k \hat{\mathcal{S}}_j^k \right\| \left\| \tilde{P}_j^k(\phi^{\Subset}) \quad \tilde{P}_j^k(\phi) \right\| + \left\| \hat{\omega}_j^k \right\| \left\| \tilde{P}_j^k(\phi^{\Subset}) \quad \tilde{P}_j^k(\phi) \right\| \left\| \tilde{P}_j^k(\phi^{\Subset}) + \tilde{P}_j^k(\phi) \right\| \\
&= \sum_{j,k} \Big\} \left\| \hat{\omega}_j^k \right\| \left\| \tilde{P}_j^k(\phi^{\Subset}) \quad \tilde{P}_j^k(\phi) \right\| \Big] 2 \left\| \hat{\mathcal{S}}_j^k \right\| + \left\| \tilde{P}_j^k(\phi^{\Subset}) + \tilde{P}_j^k(\phi) \right\| \Big\{ \\
&\leqslant 4 \sum_{j,k} \left\| \tilde{P}_j^k(\phi^{\Subset}) \quad \tilde{P}_j^k(\phi) \right\| \\
&\leqslant C d(\phi, \phi^{\Subset}) + 8 \sum_{j,k} \sup_{\phi / \Phi} \left\| \tilde{\mathcal{P}}_j^k(\alpha_l^k, \beta) \quad \mathcal{P}_j^k(\alpha_l^k, \beta) \right\| \\
&= C d(\phi, \phi^{\Subset}) + O_P(r \quad ^{\frac{1}{2}}).
\end{aligned}
$$

By Assumption (1.4.3) and $M \quad \overset{p}{\prime} \quad \infty$

$$\sup_{\alpha / \Upsilon} \min_{\alpha^{\in} / \Upsilon_M} d(\alpha, \alpha^{\Subset}) = \eta(M) \quad \overset{p}{\prime} \quad 0,$$

therefore for any $\alpha \ / \ \Upsilon$, there is a $\bar{\alpha}_{m(\alpha)M} \ / \ \Upsilon_M$ such that $d \Big) \alpha, \bar{\alpha}_{m(\alpha)M} \Big(\leqslant \eta(M)$. So that for any $\phi \ / \ \Phi$, there is $\bar{\alpha}_{m(\alpha_l^k)M}$ with $\max_{1 \leqslant l \leqslant J+1, k} \Big\} d \Big) \alpha_l^k, \bar{\alpha}_{m(\alpha_l^k)M} \Big(\quad \leqslant \eta(M)$. Let $\alpha^k(\phi) = \Big) \bar{\alpha}_{m(\alpha_1^k)M}, \times\!\!\times\!\!\times, \bar{\alpha}_{m(\alpha_{J+1}^k)M} \Big(^{\in}$, and $\alpha(\phi) = \quad \alpha^1(\phi)^{\Subset}, \times\!\!\times\!\!\times, \alpha^K(\phi)^{\Subset} \Big[^{\in}$ and $\bar{\phi}(\phi) = (\beta^{\Subset}, \alpha(\phi)^{\Subset}, \pi^{\Subset})^{\in}$. By construction $\bar{\phi}(\phi) \ / \ \Phi_M$ and $d\left(\bar{\phi}(\phi), \phi\right) \leqslant \eta(M)$. Thus we have

$$\sup_{\phi / \Phi} \inf_{\phi^{\in} / \Phi_M} d(\phi, \phi^{\Subset}) \leqslant \eta(M). \tag{A.0.21}$$

Given (A.0.20) and (A.0.21), we have

$$\sup_{\phi / \Phi} \inf_{\phi^{\in} / \Phi_M} \|\hat{Q}(\phi) \quad \hat{Q}(\phi^{\Subset})\| \leqslant C\eta(M) + O_P(r \quad ^{\frac{1}{2}}). \tag{A.0.22}$$

That is to say, for any $\phi \ / \ \Phi$, there exist a $\bar{\phi}(\phi) \ / \ \Phi_M$ such that $d\left(\bar{\phi}(\phi), \phi\right) \leqslant \eta(M)$ and $\|\hat{Q}(\phi) \quad \hat{Q}(\bar{\phi})\| \leqslant C\eta(M) + O_P(r \quad ^{\frac{1}{2}})$.

Next let $\delta > 0$ be any positive constant and define four events:

$$\begin{aligned}
\mathbb{E}_1 &= \}\eta(M) < \delta|\ , \\
\mathbb{E}_2 &= \Big\}C\eta(M) + O_P(r^{\ \frac{1}{2}}) < \frac{\xi_n}{3}\Big(\ , \\
\mathbb{E}_3 &= \Bigg\}\sup_{\phi\ /\ \Phi_I} \hat{Q}(\phi) < \frac{\xi_n}{3}\Big\lceil\ , \\
\mathbb{E}_4 &= \Bigg\}\sup_{\bar{\pi}\ /\ \mathcal{X}_M^K} \lambda_n \bar{\pi}^{\in}\bar{\pi} < \frac{\xi_n}{3}\Bigg\}\ .
\end{aligned}$$

By $\xi_n = n^{\kappa_1}, \eta(M) = n^{\kappa_2}, \lambda_n = n^{\kappa_3}, r = n^{\kappa_4}$ and the assumption of their order:

$$\}\kappa_2 < 0,\ 0 > \kappa_1 > \max\}\ \ 1, \kappa_2|\ ,\ \kappa_4 \leqslant\ \ 2\kappa_2,\ \kappa_3 < \kappa_1|\ ,$$

it follows that

$$\begin{aligned}
&P_r(\mathbb{E}_1)\ \ /\ \ 1, \\
&P_r(\mathbb{E}_2) = P_r\Bigg)C + \frac{O_P(r^{\ \frac{1}{2}})}{\eta(M)} < \frac{\eta(M)^{\ \ 1}\xi_n}{3}\Bigg[\ \ \ /\ \ 1, \\
&P_r(\mathbb{E}_3) \geqslant P_r\Bigg)O_P(n^{\ \ 1}) + O_P(r^{\ \ 1}) + O_P\Big)(nr)^{\ \frac{1}{2}}\Big(< \frac{\xi_n}{3}\Bigg[\ \ \ /\ \ 1, \\
&P_r(\mathbb{E}_4) \geqslant P_r\Bigg)\lambda_n K \leqslant \frac{\xi_n}{3}\Bigg[\ \ \ /\ \ 1.
\end{aligned}$$

It follows that $P_r\ \ \{_{r=1}^{4}\mathbb{E}_r[\ \ \ /\ \ \ 1$. When the event $\}\{_{r=1}^{4}\mathbb{E}_r|$ occurs, for every $\phi\ /\ \Phi_I$, there is $\bar{\phi}(\phi)\ /\ \Phi_M$ such that $d(\phi, \bar{\phi}) < \delta$, and

$$\begin{aligned}
\hat{Q}(\bar{\phi}) + \lambda_n \bar{\pi}^{\in}\bar{\pi} &\leqslant \hat{Q}(\bar{\phi}) + \frac{\xi_n}{3} \\
&\leqslant \hat{Q}(\phi) + \hat{Q}(\bar{\phi})\quad \hat{Q}(\phi) + \frac{\xi_n}{3} \\
&\leqslant \sup_{\phi\ /\ \Phi_I} \hat{Q}(\phi) + C\eta(M) + O_P(r^{\ \frac{1}{2}}) + \frac{\xi_n}{3} \leqslant \xi_n,
\end{aligned}$$

which implies $\bar{\phi}\ /\ \hat{\Phi}$.

Thus, w.p.a.1,

$$\sup_{\phi\ /\ \Phi_I}\inf_{\phi^{\in}/\hat{\Phi}} d(\phi, \phi^{\in}) < \delta.$$

By both $\sup_{\phi\ /\ \Phi_I}\inf_{\phi^{\in}/\hat{\Phi}} d(\phi, \phi^{\in}) < \delta$ and $\sup_{\phi^{\in}/\hat{\Phi}}\inf_{\phi\ /\ \Phi_I} d(\phi, \phi^{\in}) < \delta$, it follows that w.p.a.1 $d_H\ \ \hat{\Phi}, \Phi_I[\ < \delta$. Since $\delta > 0$ is arbitrary, it follows that $d_H\ \ \hat{\Phi}, \Phi_I[\ \ \overset{p}{/}\ \ 0$. ■

Appendix B

Mathematics for Chapter 2

Lemma 2.4.1 Proof: First of all, it is useful to realize that the product of integrations in (2.2.2) can be expressed as a single integration with respect to $H(\epsilon)$ as

(B.0.1)
$$\begin{aligned}\mathcal{P}_j^k(\alpha,\beta) = \Big[\ \mathbb{1}\Big) & g_0(\cup_T^k, y_i \quad _T = a \quad _T, \alpha, \epsilon_{iT}, \beta) = a_T\Big(* \\ & \times\times\times \\ \mathbb{1}\Big) & g_0(\cup_t^k, y_i \quad _t = a \quad _t, \alpha, \epsilon_{it}, \beta) = a_t\Big(* \\ & \times\times\times \\ \mathbb{1}\Big) & g_0(\cup_1^k, \alpha, \epsilon_{i1}, \beta) = a_1\Big(dH(\epsilon).\end{aligned}$$

This equation holds since ϵ_{it} is serial independent. With indicator functions of utilities, its expansion is

(B.0.2)
$$\begin{aligned}\mathcal{P}_j^k(\alpha,\beta) = \Big[& \prod_{a_T^{\in}\neq a_T} \mathbb{1}\Big) V(\cup_{Ta_T}^k, a \quad _T, \alpha_{a_T}, \beta) + \epsilon_{iTa_T} \geqslant V(\cup_{Ta_T^{\in}}^k, a \quad _T, \alpha_{a_T^{\in}}, \beta) + \epsilon_{iTa_T^{\in}}\Big(* \\ & \times\times\times \\ & \prod_{a_t^{\in}\neq a_t} \mathbb{1}\Big) V(\cup_{ta_t}^k, a \quad _t, \alpha_{a_t}, \beta) + \epsilon_{ita_t} \geqslant V(\cup_{ta_t^{\in}}^k a \quad _t, \alpha_{a_t^{\in}}, \beta) + \epsilon_{ita_t^{\in}}\Big(* \\ & \times\times\times \\ & \prod_{a_1^{\in}\neq a_1} \mathbb{1}\Big) V(\cup_{1a_1}^k, \alpha_{a_1}, \beta) + \epsilon_{i1a_1} \geqslant V(\cup_{1a_1^{\in}}^k, \alpha_{a_1^{\in}}, \beta) + \epsilon_{i1a_1^{\in}}\Big(dH(\epsilon).\end{aligned}$$

It is needed to prove that these functions under integration above indexed by (α, β) form a Donsker class. These functions are product of different indicators, and by the same way of Huang 2015b each class formed by indicators is Donsker, so do their product. This complete the proof. ■

Lemma 2.4.2 Proof: Note that

$$\left\|\tilde{\Delta}\mathcal{P}_t(\alpha,\beta) \quad \Delta\mathcal{P}_t(\alpha,\beta)\right\| \leqslant \left\|\frac{1}{r}\sum_{i=1}^{r}\mathbb{1}(\tilde{y}_{it}(s_t^a) = a_t) \quad \mathcal{P}_{a_t}^{s_t^a}(\alpha,\beta)\right\| + \left\|\frac{1}{r}\sum_{i=1}^{r}\mathbb{1}(\tilde{y}_{it}(s_t^b) = a_t) \quad \mathcal{P}_{a_t}^{s_t^b}(\alpha,\beta)\right\|,$$

and I need to check whether

$$\sup_{(\alpha,\beta)\,/\mathrm{Y}*\ \mathbb{B}} \left\|\frac{1}{r}\sum_{i=1}^{r}\mathbb{1}(\tilde{y}_{it}(s_t) = a_t) \quad \mathcal{P}_{a_t}^{s_t}(\alpha,\beta)\right\| /^{P} \ 0.$$

This is actually to check whether the indicator functions $\}\mathbb{1}\,(\tilde{y}_{it}(s_t) = a_t)|_{\ (\alpha,\beta)}$ form a Glivenko-Cantelli class. This can be seen by writing it as follows

$$\mathbb{1}\,(\tilde{y}_{it}(s_t) = a_t) = \prod_{a_t^{\in}\forall\!\!= a_t} \mathbb{1}\Big)\tilde{\epsilon}_{ita_t^{\in}} \quad \tilde{\epsilon}_{ita_t} \leqslant V(s_{ta_t}, \alpha_{a_t}, \beta) \quad V(s_{ta_t^{\in}}, \alpha_{a_t^{\in}}, \beta)\Big(.$$

This could be proved by use the same idea of subgraph argument as in Huang 2015b. Then by the first inequality, I have

$$\sup_{(\alpha,\beta)\,/\mathrm{Y}*\ \mathbb{B}} \left\|\tilde{\Delta}\mathcal{P}_t(\alpha,\beta) \quad \Delta\mathcal{P}_t(\alpha,\beta)\right\| /^{P} \ 0. \blacksquare$$

Bibliography

Aguirregabiria, Victor and Pedro Mira (2010). "Dynamic discrete choice structural models: A survey". In: *Journal of Econometrics* 156.1, pp. 38–67.

Bates, D et al. (2014). *minqa: Derivative-free optimization algorithms by quadratic approximation*.

Berry, Steven T. (1994). "Estimating Discrete-Choice Models of Product Differentiation". In: *The RAND Journal of Economics* 25.2, pp. 242–262. ISSN: 07416261. URL: http://www.jstor.org/stable/2555829.

Berry, Steven T. and Philip A. Haile (2010). "Identification in Differentiated Products Markets Using Market Level Data". In: *working paper*.

Berry, Steven, James Levinsohn, and Ariel Pakes (1995). "Automobile Prices in Market Equilibrium". In: *Econometrica* 63.4, pp. 841–890. ISSN: 00129682. URL: http://www.jstor.org/stable/2171802.

— (2004). "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market". In: *Journal of Political Economy* 112.1, pp. 68–105.

Blundell, Richard and James Powell (2006). "Endogeneity in nonparametric and semiparametric regression models". In: CWP09/01. URL: http://ideas.repec.org/p/ifs/cemmap/09-01.html.

Bonhomme, Stéphane (2012). "Functional Differencing". In: *Econometrica* 80.4, pp. 1337–1385. ISSN: 1468-0262. DOI: 10.3982/ECTA9311. URL: http://dx.doi.org/10.3982/ECTA9311.

Bresnahan, Timothy (1987). "Competition and Collusion in the American Automobile Industry: The 1955 Price War". In: *Journal of Industrial Economics* 35.4, pp. 457–482.

Browning, Martin and Jesus M Carro (2013). "The Identification of a Mixture of First-Order Binary Markov Chains". In: *Oxford Bulletin of Economics and Statistics* 75.3, pp. 455–459.

— (2014). "Dynamic binary outcome models with maximal heterogeneity". In: *Journal of Econometrics* 178.2, pp. 805–823.

Chamberlain, Gary (1987). "Asymptotic efficiency in estimation with conditional moment restrictions". In: *Journal of Econometrics* 34.3, pp. 305–334. URL: http://ideas.repec.org/a/eee/econom/v34y1987i3p305-334.html.

Chernozhukov, Victor et al. (2013a). "Average and Quantile Effects in Nonseparable Panel Models". In: *Econometrica* 81.2, pp. 535–580. ISSN: 1468-0262. DOI: 10.3982/ECTA8405. URL: http://dx.doi.org/10.3982/ECTA8405.

— (2013b). "Supplement to Average and Quantile Effects in Nonseparable Panel Models". In: *Econometrica* 81.2, pp. 535–580.

Eddelbuettel, Dirk (2013). *Seamless R and C++ Integration with Rcpp*. Springer.

Geer, Sara van de (2006). *Empirical Process Theory and Applications*.

Ghosal, Subhashis and Aad W Van Der Vaart (2001). "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities". In: *Annals of Statistics*, pp. 1233–1263.

Heckman, James J (1981). "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process." In:

Honoré, Bo E. (2002). "Nonlinear models with panel data". In: *Portuguese Economic Journal* 1, pp. 163–179.

Honoré, Bo E. and Elie Tamer (2006). "Bounds on Parameters in Panel Dynamic Discrete Choice Models". In: *Econometrica* 74.3, pp. 611–629. ISSN: 1468-0262. DOI: 10.1111/j.1468-0262.2006.00676.x. URL: http://dx.doi.org/10.1111/j.1468-0262.2006.00676.x.

Huang, Mian (2014). "Memo: A Framework for Multinomial Discrete Choice Model with Fixed Effects and MC Simulation". In:

— (2015a). "Set Identified Dynamic Multinomial Discrete Choice Model with Fixed Effects and Simulation." In:

— (2015b). "Simulation Based Estimation of Multinomial Discrete Choice Model with Fixed Effects". In:

Keane, Michael P. and Kenneth I. Wolpin (1997). "The Career Decisions of Young Men". In: *Journal of Political Economy* 105.3, pp. 473–522. ISSN: 00223808. URL: http://www.jstor.org/stable/10.1086/262080.

Kosorok, Michael R (2007). *Introduction to empirical processes and semiparametric inference*. Springer.

Lancaster, Tony (2000). "The incidental parameter problem since 1948". In: *Journal of Econometrics* 95.2, pp. 391–413. ISSN: 0304-4076. DOI: 10.1016/S0304-4076(99)00044-5. URL: http://www.sciencedirect.com/science/article/pii/S0304407699000445.

Matloff, Norman (2011). *The art of R programming: a tour of statistical software design*. No Starch Press.

McFadden, Daniel (1973). “Conditional logit analysis of qualitative choice behavior”. In: pp. 105–142.

— (1977). *Modelling the Choice of Residential Location*. Cowles Foundation Discussion Papers 477. Cowles Foundation for Research in Economics, Yale University. URL: http://ideas.repec.org/p/cwl/cwldpp/477.html.

Mincer, Jacob (1958). “Investment in human capital and personal income distribution”. In: *The journal of political economy*, pp. 281–302.

Nash, John C. (2014). “On Best Practice Optimization Methods in R”. In: *Journal of Statistical Software* 60.2.

Nash, John C. and Ravi Varadhan (2011). “Unifying Optimization Algorithms to Aid Software System Users: optimx for R”. In: *Journal of Statistical Software* 43.9.

Nevo, Aviv (2000). “a practitioner’s guide to estimation of random-coefficients logit models of demand”. In: *Journal of Economics & Management Strategy*, pp. 513–548.

Neyman, J. and Elizabeth L. Scott (1948). “Consistent Estimates Based on Partially Consistent Observations”. English. In: *Econometrica* 16.1, ISSN: 00129682. URL: http://www.jstor.org/stable/1914288.

Pitt-Francis, Joe and Jonathan Whiteley (2012). *Guide to scientific computing in C++*. Springer.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Train, Kenneth E. (2009). *Discrete Choice Methods with Simulation 2nd ed*. Cambridge University Press.

Train, Kenneth E, Daniel L McFadden, and Moshe Ben-Akiva (1987). “The demand for local telephone service: a fully discrete model of residential calling patterns and service choices”. In: *The RAND Journal of Economics*, pp. 109–123.

Turlach, BA and A Weingessel (2013). *quadprog: Functions to solve quadratic programming problems. R package version 1.5-5*.

Varadhan, Ravi, Hans W Borchers, and Maintainer Ravi Varadhan (2011). *Package ‘dfoptim’*.

Wooldridge, Jeffrey M. (2005). “Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity”. In: *Journal of Applied Econometrics* 20.1, pp. 39–54. ISSN: 1099-1255. DOI: 10.1002/jae.770. URL: http://dx.doi.org/10.1002/jae.770.

Wright, Margaret H (2010). “Nelder, Mead, and the other simplex method”. In: *Documenta Mathematica* 7.