Adaptive Rank Inference in
Semiparametric Multinomial Response Models

S. Khan  F. Ouyang  E. Tamer
Duke University  Duke University  Harvard University

June 20, 2016
Preliminary and Incomplete Version

Abstract

We consider estimation and inference on regression coefficients in semi parametric multinomial response models. Our approach is based on a localized rank objective function, loosely analogous to that used in Abrevaya, Hausman, and Khan (2010), which we show achieves sharp identification of the identified region. This is in contrast to existing procedures in the literature such as Ahn, Powell, Ichimura, and Ruud (2014), which although providing consistent estimators when conditions for point identification are satisfied, do not yield sharp set estimates when they are not. A leading case when point identification is not satisfied is when all covariates are discrete, as happens often in empirical settings. In these settings we show that the aforementioned procedures sometimes only estimate the trivial set. In contrast, our procedure is adaptive in the sense that it provides an estimator of the sharp set when point identification does not hold, and a root-n consistent point estimator when it does. Furthermore, we show our rank procedure readily extends to panel data settings for inference in models with fixed effects. This includes dynamic panel models with lagged dependent variables as covariates. We also propose a new algorithm to accommodate empirically relevant settings when the number of choices available to the agent are large relative to the sample size of cross sectional observations. A simulation study establishes adequate finite sample properties of our new procedures.

Keywords: Multinomial choice, Rank Estimation, Adaptive Inference. Dynamic Panel Data.
1 Introduction

Many important economic decisions involve households’ or firms choice among qualitative or discrete alternatives. Examples are individuals’ choice among transportation alternatives, labor force participation, family size, residential location, type and level of education, brand of automobile, etc. In transportation choices, for example, the commuter typically chooses among a few available transportation modes, each characterized by traveling time, cost, convenience and comfort. If a behavioral model for commuters transportation choices has been established and estimated one can apply this model to assess the changes in the aggregate distribution of commuters that follow from changing prices and traveling times or by introducing a new transportation alternative, or closing down an old one.

The theory of discrete choice is designed to model these kind of choice settings, and to provide the corresponding econometric methodology for empirical analyses. Due to variables that are unobservable to the econometrician the observations from a sample of agents discrete choices can be viewed as outcomes generated by a stochastic model.

Statistically, these observations can be considered as outcomes of multinomial experiments, since the alternatives typically are mutually exclusive. In the context of choice behavior, the probabilities in the multinomial model are to be interpreted as the probability of choosing the respective alternatives (choice probabilities), and the purpose of the theory of discrete choice is to provide a structure of the probabilities that can be justified from behavioral arguments. Specifically, one is interested in expressing the choice probabilities as functions of the agents’ preferences and the choice constraints. The choice constraints are often represented by economic budget constraint and in addition, the choice set, which is the set of alternatives that are feasible to the agent.

Over the last 25 years there has been extensive development in the theoretical and methodological literature within the field of discrete choice. Early developments are discussed in to standard textbooks and surveys in econometric modelling of discrete choice such as Maddala (1983), Train and Mehrez (1994), Amemiya (1985), McFadden (1978) and Ben-Akiva, McFadden, and Train (1987)). The earlier work on estimating multinomial choice models took a parametric approach by assuming known distributions of unobserved components of the model.

The most popular of these were the multinomial logit model and the multinomial probit. The former had the disadvantage of suffering from the well known IIA problem, as pointed out in McFadden (1978), The latter was computationally difficult as the likelihood function lacked an analytic form. Both also suffered from the sensitivity of results to the parametric assumptions.

the IIA encountered in the multinomial logic. But it is worth pointing out they are both "distribution free", and also allow for cross sectional heteroskedasticity. Lee (1995), Powell and Ruud (2008), Ahn, Powell, Ichimura, and Ruud (2014), Shi, Shum, and Song (2015) allow for correlation in unobservables across choices so do not suffer from IIA, but are based on index/independence assumptions.

Even more recently, the literature has considered panel data variants of the multinomial choice model. This is important for applied work as the increased availability of longitudinal panel data sets has presented new opportunities for econometricians to control for individual unobserved heterogeneity across agents. Examples include Shi, Shum, and Song (2015) and Pakes and Porter (2014). The latter takes a partial identification approach by deriving moment inequalities.

In this paper we propose new estimation and inference procedures for cross sectional and panel data multinomial choice response models. The proposed procedure is semiparametric, and the panel data setting, dynamic as it allows for lagged dependent variables as covariates, enabling the distinction between state dependence and unobserved heterogeneity. Compared to the existing work mentioned above the main contributions of our work are two fold. Unlike the previous approaches ours is robust in the sense that it achieves sharp bounds for the regression coefficients when conditions for point identification are not satisfied. Furthermore the procedure is adaptive to point identification in the sense that the same procedure yield a consistent estimator when the point identification conditions are satisfied. Finally, another contribution is the new results for the dynamic multinomial choice model. To our knowledge theory are the first point identification and consistency results for this model.

We structure the paper as follows. In the next section we formally introduce the cross sectional model, and state standard regularity conditions on both observed and unobserved random variables. These conditions yield the inequalities we used for our identification strategy, which in turn motivate a localized rank based objective function. We then show that the optima of the objective function form an estimator that is adaptive, consistent, and under certain DGPs, root-n consistent and asymptotically normal.

Section 3 generalizes the model by assuming the availability of a longitudinal panel data set and introducing unobserved individual specific effects. For this model we propose a localized maximum score Manski (1987) estimator, that we show is adaptive and under certain DGP’s is point consistent. Most interestingly in this paper, we further generalize the model by introducing dynamics. Specifically, we do so by allowing lagged values of dependent variables to be explanatory variables. This approach of remodeling dynamics was taken in the binary choice model- see, e.g. Heckman (1978) and Honore and Kyriazidou (2000). Here again we propose an adaptive procedure that can be point consistent under standard conditions.

Section 4 discusses the case where the number of possible choices can be arbitrarily large, as occurs in many empirical settings- for example in marketing urban economics. Here we propose two novel algorithm that provided a data driven method to simultaneous estimated the number of choices and the regression coefficients.
Section 5 explores finite sample properties of the new procedures/algorithms through a small scale simulation study.

Section 6 concludes by summarizing results and proposing areas for future research. The appendix collects proofs of many of the theorems stated in the paper.

2 Semiparametric Multinomial Choice

For the standard multinomial choice model assume the dependent variable takes one of $J+1$ mutually-exclusive and exhaustive alternatives (numbered from $j = 0$ to $j = J$) is chosen.

Specifically, for individual $i$, alternative $j$ is assumed to have an unobservable indirect utility $y^*_ij$ for that individual. The alternative with the highest indirect utility is assumed chosen.

Thus the observed variable $y_{ij}$ has the form

$$y_{ij} = I[y^*_ij > y^*_ik \text{ for } k = 0, ..., J]$$

with the convention that $y_{ij} = 0$ indicates choice of alternative $j = 0$. As is standard in this literature an assumption of joint continuity of the indirect utilities rules out ties (with probability one);

In our semiparametric model, the indirect utilities are further restricted to have the linear form

$$y^*_ij = x'_{ij} \beta_0 + \epsilon_{ij}$$

for $j = 1, ..., J$, where the vector $\epsilon_i$ of unobserved error terms is assumed to be jointly continuously distributed and independent of the $J \times (p+1)$-dimensional matrix of regressors $X_i$ (whose $j^{th}$ row is $x_{ij}$). (For alternative $j = 0$, the standard normalization $y_{ij} = 0$ is imposed.)

Parametric assumptions on the unobservables that have been traditionally assumed are $\epsilon_i$ include iid Type 1 extreme value (Multinomial Logit) or $\epsilon_i$ multivariate normal (Multinomial Probit)

The multinomial logit model suffers from the well known IIA problem McFadden (1978) The multinomial probit is analytically intractable, requiring simulation methods. These can be computationally difficult, and as is the case with multinomial logit, is inconsistent if the error distribution is misspecified and/or heteroskedastic.

As mention, these have motivated semi parametric approaches. This includes Lee (1995), who imposed a profile likelihood approach, extending the results in Klein and Spady (1993) for the binary choice model. Ahn, Powell, Ichimura, and Ruud (2014), propose a 2-step, estimator, that requires nonparametric methods but shoe second stage is closed form. Shi,
Shum, and Song (2015) also propose a 2 step estimator that exploits a cyclic monotonicity condition, but also requires high dimensional nonparametric first stage, but whose second stage is not closed form as Ahn, Powell, Ichimura, and Ruud (2014) is.

Unlike the parametric approaches these approaches are distribution free and in one sense more robust. But unlike parametric approaches they all require one continuous regressor for point identification. Consequently, they do not nest parametric approaches, and furthermore can perform very poorly when all regressors are discrete.

Furthermore point identification is lost when all regressors are discrete, and these estimators may converge to a point that does not lie in the identified region.

These drawbacks motivate the approach proposed in this paper. Specifically, we propose an approach which: converges to identified region when parameter is not point identified, and furthermore, is indeed point consistent and asymptotically normal when the parameter is point identified.

2.1 Local Rank Procedure

To illustrate our approach consider the following multinomial choice model (3 choices for now) where the latent utilities for choices 0, 1, 2 are:

\[
\begin{align*}
y_{i0}^* &= 0 \\
y_{i1}^* &= x_{i1}^\prime \beta + \epsilon_{i1} \\
y_{i2}^* &= x_{i2}^\prime \beta + \epsilon_{i2}
\end{align*}
\]

with the maintained assumption that

\[(\epsilon_{i1}, \epsilon_{i2}) \perp (x_{i1}, x_{i2})\]

but we allow arbitrary correlation between \(\epsilon_{i1}\) and \(\epsilon_{i2}\).

Assume from a random sample of choices \((y_{i0}, y_{i1}, y_{i2}; x_{i1}; x_{i2})\) such that:

\[y_{i1} = 1[y_{i1}^* \geq y_{i2}^*; y_{i1}^* \geq 0]\]

We are interested in the sharp set for \(\beta\) when all the regressors have discrete support.

From the model and assumptions we have for a given \(\beta\) and joint distribution on the epsilons:

\[
G(x; \beta, F) = \begin{bmatrix}
P(x_{i1} \beta + \epsilon_{i1} \leq 0; x_{i2} \beta + \epsilon_{i2} \leq 0) \\
P(x_{i1} \beta + \epsilon_{i1} \geq 0; x_{i1} \beta + \epsilon_{i1} \geq x_{i2} \beta + \epsilon_{i2}) \\
P(x_{i1} \beta + \epsilon_{i1} \leq x_{i2} \beta + \epsilon_{i2}; x_{i2} \beta + \epsilon_{i2} \geq 0)
\end{bmatrix}
\]
To illustrate how to get bounds for $\beta$ we first fix $x_{i2}$. With $x_{i2}$ fixed, we have what we call a conditional monotone index model.

By this we mean that conditional on $x_{i1}, x_{i2}$ $P(y_{i1}|x_i)$ is increasing in $x_{i1}\beta$.

Thus we have have the relationship

$$P(y_{i1}|x_{i1}, x_{i2} = x_2) \geq P(y_{i1}|x_{j1}, x_{j2} = x_2) \text{ iff } x_{i1} \beta \geq x_{j1}\beta$$

This will be the basis for our identification and inference results. Furthermore, we can show how this easily extends to more then 3 choices by conditioning on all the other regressors except for those in the first choice.

Crucially, the above conditional moment inequalities can be repeated for all values of $x_2$ (finitely many if support of $x_2$ is finite). Furthermore, note for a fixed $x_2$, $P(y_{i0} = 1|x)$ and $P(y_{i2} = 1|x)$ are both decreasing in $x_{i1}\beta$. Of course, this can be exploited also fixing $x_1$ resulting in a slew of other moment inequalities. Collectively, all these moment inequalities would be the basis for our bounds. As we will show here this will translate into an estimation procedure, which will converge to the identified region.

For estimation assume a random sample of $n$ observations. We propose the following weighted rank correlation estimator, analogous to the MRC proposed in Han (1987).

$$G_{1n}(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} I[x_{i2} = x_{j2}]I[y_{i1} > y_{j1}]I[x_{i1}\beta > x'_{j1}\beta]$$

The above function can be used for one set of moment inequalities. But as alluded to from the identification inequalities, can also work with $y_{i0}, y_{i2}$. For the latter, another objective function is

$$G_{2n}(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} I[x_{i2} = x_{j2}]I[y_{i2} > y_{j2}]I[x'_{i1}\beta < x'_{j1}\beta]$$

In addition to these we can fix $x_{i1}$; this can yield an objective function of the form

$$G_{3n}(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} I[x_{i1} = x_{j1}]I[y_{i2} > y_{j2}]I[x'_{i2}\beta > x'_{j2}\beta]$$

Note we can combine all objective functions by, for example, adding them up, resulting in an objective function $G_{4n}$ and maximize with respect to $\beta$. Next, we establish some of the properties of this type of approach.

We first consider the case where all the regressors are discrete. We have the following theorem:
Theorem 2.1 Under Assumptions A1-A3, the optimizer of the sum of the objective function $G_{4n}$ converges to the set $B$. $B$ is to singleton, but the identified region under Assumptions A1-A3.

What this theorem implies is that when all of the regressors are discrete, $\beta$ is not point identified.

In fact our point identification result require the following assumption.

A1: One of the vectors $x_1, x_2, ... x_J$ has one component which is continuously distributed with positive density on the real line.

Such a support condition is analogous to that assumed in Manski (1975) and Han (1987). Also, the approach effectively assumed exclusion restrictions but that is not necessary.

We note the continuity condition is only for point identification, and an advantage of our procedure is one gets meaningful bounds when all the regressors are discrete.

Importantly, note also that under condition A1 we will never have ties when matching regressors, thus the value of the objective function will always be 0. But we can construct kernel weights as follows. To illustrate for the $G_{1n}$ objective function, construct the approximate binary weights:

$$I[x_{i2} = x_{j2}] \approx K_h(x_{i2} - x_{j2}) \equiv w_{ij}$$

where $K$ denotes a kernel density function. $K_h(\cdot) = K(\cdot/h)$ $h$ is a bandwidth sequence that converges to 0 as $n \to \infty$.

We have the following theorem:

Theorem 2.2 Under Assumptions A1-A3,A.1, $\beta_0$ is point identified, the optimizer of $G_{4n}$ is unique with probability approaching 1, and the optimizer converges in probability to $\beta_0$.

Regularity Conditions Here we outline the regularity conditions for the point consistency and asymptotic normality of the proposed estimator. We note that the estimator has a similar structure to that proposed in Abrevaya, Hausman, and Khan (2010). Consequently we impose regularity conditions similar to those used in that paper.

We first note that since $\beta$ is only identified up to scale, we will normalize its last component to 1 and denote its other components by $\theta_0$ and the corresponding estimator by $\hat{\theta}$, where

$$\hat{\theta} = \arg \max_{\theta \in \Theta} G_{1n}(\theta) \equiv \frac{1}{n(n-1)} \sum_{i \neq j} w_{ij} I[y_{i1} > y_{j1}] I[x'_{i1} \theta > x'_{j1} \hat{\theta}]$$

where above the index now reflects the scale normalization, so $x'_{i1} \theta = x'_{i1} \beta$ with the first component of $\beta$ is set to 1.
We impose the following regularity conditions:

**Assumption A0** Letting the vectors $y_i, x_i$ denote the collected choice outcomes and choice regressors, we assume of random sample of them is drawn, $i = 1, 2, \ldots, n$.

**Assumption A1** (Parameter Space) $\theta_0$ lies in the interior of $\Theta$, a compact subset of $\mathbb{R}^{k-1}$.

**Assumption A2** (Matching stage kernel function) The kernel function $k(\cdot)$ used in matching the regressors $x_{2i}$ is assumed to have the following properties:

- **A2.1** $k(\cdot)$ is twice continuously differentiable, has compact support and integrates to $1$.
- **A2.2** $k(\cdot)$ is symmetric about $0$.
- **A2.3** $k(\cdot)$ is a $p^{th}$ order kernel, where $p$ is an even integer:

$$
\int u^l k(u) du = 0 \text{ for } l = 1, 2, \ldots, p - 1
$$

$$
\int u^p k(u) du \neq 0
$$

**Assumption A3** (Matching stages bandwidth sequence) The bandwidth sequence $h_n$ used in the second stages satisfies the conditions, $\sqrt{n}h_n \to 0$, $\sqrt{n}h_n^3 \to \infty$.

**Assumption A4** (First Regressor Properties) $x_{i1}^{(1)}$ is continuously distributed, with positive density on the real line.

**Assumption A5** (Error Distribution) $\epsilon_i$ denotes the vector of disturbance terms for individual $i$ for all choices. We assume it is distributed independently of the regressors $x_i$, and is continuously distributed with positive density on the real plane.

**Assumption A6** (Full Rank Condition) The support of $x_{i1}^{(1)}$ does not lie in a proper linear subspace of $\mathbb{R}^k$, conditional on $x_{i2}^{(2)}$.

We can next strengthen the assumptions to ensure root-$n$ consistency and asymptotic normality of the proposed estimator. In doing so, we introduce the following notation, which is chosen to deliberately be as close as possibly to that used in Sherman (1993). Let $z_i^{(1)} = (y_i^{(1)}, x_i^{(1)})$. Next, define

$$
\tau(z^{(1)}, x^{(2)}, \theta) = P(y^{(1)} > y_i^{(1)}, x^{(1)} \beta > x_i^{(1)} \beta | x_i^{(2)} = x^{(2)}) + P(y^{(1)} < y_i^{(1)}, x^{(1)} \beta < x_i^{(1)} \beta | x_i^{(2)} = x^{(2)})
$$

**Assumption AN1** Let $\mathcal{N}$ denote a neighborhood of $\theta_0$.

1. For each $z^{(1)}$, all mixed second partial derivatives of $\tau(z, x^{(2)}, \cdot)$ exist on $\mathcal{N}$. Let $\nabla_1, \nabla_2$ denote first and second derivative operators, respectively.
2. There is an integrable function $M(z^{(1)}, x^{(2)})$ such that for all $z^{(1)}, x^{(2)},$ and $\theta$ in $\mathcal{N}$,

\[|\nabla_2 \tau(z^{(1)}, x^{(2)}, \theta) - \nabla_2 \tau(z^{(1)}, x^{(2)}, \theta_0)| \leq M(z^{(1)}, x^{(2)}) ||\theta - \theta_0||\]

3. $E[||\nabla_1 \tau(\cdot, \cdot, \theta_0)||^2] < \infty$

4. $E[||\nabla_2 \tau(\cdot, \cdot, \theta_0)||] < \infty$

5. The matrix $E[\nabla_2 \tau(\cdot, \theta_0)]$ is negative definite.

With these conditions, we are able to conclude that our local rank estimator is root-$n$ consistent and asymptotically normal, as stated in the following theorem:

**Theorem 2.3** Assume the density function of $x^{(2)}_i$ is $p$ times continuously distributed and that $p > d_x/2$ with $d_x$ denoting the dimension of $\beta$. Then, under Assumptions A1-A6 and AN1, we have

\[\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, V^{-1} \Delta V^{-1})\]

where $V = \frac{1}{2} E[\nabla_2 \tau(\cdot, \cdot, \theta_0)]$ and $\Delta = E[\nabla_1 \tau(\cdot, \cdot, \theta_0) \nabla_1 \tau(\cdot, \cdot, \theta_0)'$]

3 Panel Data Multinomial Choice

Paralleling the increase in popularity of estimating multinomial regression models is the estimation of panel data models. The increased availability of longitudinal panel data sets has presented new opportunities for econometricians to control for individual unobserved heterogeneity across agents. In linear panel data models, unobserved additive individual specific heterogeneity, if assumed constant over time (i.e. “fixed effects”), can be controlled for when estimating the slope parameters by first differencing the observations.

Discrete panel data models have received a great deal of interest in both the econometrics and statistics literature, beginning with the seminal paper of Andersen (1970). For a review of the early work on this model see Chamberlain (1984), and for a survey of more recent contributions see Arellano and Honoré (2001). More generally speaking there is a vibrant and growing literature on both partial and point identification in nonlinear panel data models. There are a set of recent papers that deal with various nonlinearities in models with (short $T$) panels. See for example the work of Arellano and Bonhomme (2009), Bester and Hansen (2009), Bonhomme (2012), Chernozhukov, Fernandez-Val, Hahn, and Newey (2013), Evdokimov (2010), Graham and Powell (2012) and Hoderlein and White (2012). See also the survey in Arellano and Honoré (2001).
Here we consider a panel data model for multinomomial choice Chamberlain (1984)

\[ y_{ijt}^* = \alpha_{ij} + x'_{ijt}\beta_0 + \epsilon_{ijt} \]

\[ i = 1, 2, \ldots n \]
\[ j = 1, 2, \ldots J \]
\[ t = 1, 2, \ldots T \]
\[ y_{it} = \arg\max_j y_{ijt}^* \]

Here we will consider identification and asymptotics is \( J, T \) fixed, \( n \to \infty \). Note also \( \alpha_{ij} \) denotes the individual and choice specific effect. Existing results for binary choice with fixed effects: Andersen (1970), Manski (1987), Chamberlain (2010).

The literature on multinomial choice for panel data is more limited. Recent results include Shi, Shum, and Song (2015), Pakes and Porter (2014). The latter only attains partial identification. The former attains point identification but under random effects and a strict exogeneity assumption usually not required in the existing literature, such as Manski (1987)).

Here we propose point identification results that are robust to misspecification, yet still impose the weak conditions in Manski (1987) To illustrate our identification results, assume \( T = 2, J = 3 \) w.l.o.g.

And as before impose normalization that \( y_{i0t}^* \equiv 0 \).

Our identification strategy will be analogous to the cross sectional case but now we match and compare individuals over time as oppose to pairs of individuals. As we will show the analogy is not 100% complete as we have to condition on “switchers”, in a way similar to estimation of the conditional Logit model.

So basically we need a subset of the population whose value for a choice changes over time but whose regressors for a different choice do not change over time.

Specifically, we propose working with the objective function:

\[ G^p_n(\beta) = \frac{1}{n} \sum_{i=1}^n I[y_{i11} = 1] \neq I[y_{i12} = 1]I[x_{i21} = x_{i22}]I[x_{i11}'\beta \geq x_{i12}'\beta] \]

Under conditions analogous to Manski (1987) (full rank of \( \Delta x_{i1} \), continuity and large support of \( x_{it} \)), \( \beta_0 \) is identified. Furthermore, we can formally show that

**Theorem 3.1** Under Assumptions P1-P5,

\[ \hat{\beta} = \arg\max_\beta G^p_n(\beta) \overset{p}{\to} \beta_0 \]
However, in contrast to cross sectional case here not "enough" matches for standard asymptotics to hold.

For binary choice, Manski(1987) resulted in “cube root” asymptotics, nongaussian limiting distribution. Here, things may be worse as we have to also match in the regressors.

A smoothed objective function (Horowitz (1992), Charlier, Melenberg, and van Soest (1995)) should result in asymptotic normality under stronger smoothness conditions.

3.1 Dynamic Multinomial Choice

In this section we extend the base model of the previous section by introducing dynamics into the model. The dynamics we consider will be of the form of including lagged discrete dependent variables as one of the explanatory variables. This model is well motivate in empirical settings.

In many situations, such as in the study of labor force and union participation, or transportation choice, it is observed that an individual who has experienced an event, or made some choice in the past is more likely to experience the event or make that same choice in the future than an individual who has not experienced the event or made that choice. Heckman (1981), Heckman (1991) discusses two explanations for this phenomenon. The first explanation is the presence of "true state dependence," in the sense that the lagged choice/decision enters the model as an explanatory variable. The second is the presence of serial correlation in the unobserved transitory errors that are in the model. Here we will be interested in the case where this serial correlation is due to the presence of unobservable permanent individual and choice specific heterogeneity.

This section expands results from the previous one by presenting identification and estimation methods for discrete choice models with structural state dependence that allow for the presence of unobservable individual heterogeneity in panels with a large number of individuals observed over a small number of time periods.

One way to express this model \( J = 3 \) as follows:

\[
y_{it}^* = \alpha_i + x_{it}'\beta_0 + \gamma_0 I[y_{i(t-1)} = 1] + \epsilon_{it}
\]

This version of the dynamic model is narrow in the sense that only the lagged value of a particular choice enters as an explanatory variable as opposed to allowing for lagged values of other possible choices. This simplification is made solely for simplicity in illustrating the identification approach here.

In this model, the parameters of interest are \( \beta_0, \gamma_0 \). Identification is more complicated in dynamic model, even for binary choice. For example, Chamberlain(1985) shows \( \beta_0 \) is not identified when \( T = 3 \). (But \( \gamma_0 \) is identified if \( \beta_0 = 0 \)). Honore and Kyriazidou (2000) show identification of \( \gamma_0 \) and \( \beta_0 \) \( T = 4 \).

Their identification based on conditioning on the subset of the population whose re-
gressors do not change in periods 2,3. Interestingly, this is analogous to what we do in multinomial choice for regressors for a different choice.

Thus our identification strategy for the dynamic multinomial choice model is based on conditioning on the subpopulation whose regressors are time invariant for 2 choices. Specifically, in the three choice setting 4 time period setting we condition on the subpopulation wise regressor values for voice 2 do not change over time in some periods and whose regressor values for choice 1 do not change over time for a different set of time periods.

Therefore, we consider the objective function:

\[ G_n^{DP}(\beta, g) = \sum_{i=1}^{n} I[x_{i21} = x_{i23}]I[x_{i12} = x_{i13}](I[y_{i11} = 1] > I[y_{i12} = 1]) \]

\[ I[x'_{i12} \beta + g \cdot I[y_{i11} = 1] > x'_{i11} \beta + g \cdot I[y_{i10} = 1]] \]

Note that for point identification we require that one of the components of the regressors for the first choice has to be continuously distributed. Consequently, when matching regressors for this choice, we would need to assign kernel weights as illustrated before. Regarding matching the regressors for the other choices, as was the case before we do not need continuity for point identification. As such, all the regressors there can be discrete and we would not need kernel weights.

Under the standard “initial conditions” assumption, as in, e.g. Honore and Kyriazidou (2000) the maximizer of this objective function can be shown to be consistent, although as in the static model, the asymptotic theory is nonstandard.

We outline here the regularity condition, and outline a proof of point identification. We deliberately keep the notation as close as possible to Honore and Kyriazidou (2000).

DP1 Let \( y_{it} \) denote the \( J + 1 \) vector of binary choice indicators in period \( t \) for each of the choices. Let \( x_{it} \) be the set of regressor vectors for each of the choices in period \( t \).

DP2 Let \( \epsilon_{it} \) denote the \( J \) dimensional vector of disturbance terms in the latent utilities. Then for each \( \epsilon_{it} \) is stationary across time and has positive density on \( \mathbb{R}^J \) conditional on \( (x_i, \alpha_i) \), where \( \alpha_i \) denotes the \( J \) dimensional vector of choice effects for individual \( i \).

DP3 For the first choice regressors, the first component has positive density on the real line conditional on all values of the other components and the regressors for the other choices, for all time periods.

DP4 For the regressors for the first choice, let \( x_{ilm} \) denote \( x_{il} - x_{im} \) for time periods \( l, m \). Then the support of \( x_{i12} \) conditional on \( x_{i23} \) lying in a neighborhood of 0 and conditional on the regressors for the other choices not changing over time, does not lie in a proper linear subspace of \( \mathbb{R}^k \).
DP5. The random vector $x_{123}$ is continuously distributed with is density that is strictly positive in a neighborhood of 0.

DP6. Let $h(\theta) = y_{i21} \text{sgn}(z_i \theta)$ where $y_{i21} = y_{i2} - y_{i1}$, with $y_{it}$ here denoting the binary indicator for the first choice. $z_i$ denotes $(x_{i21}, y_{i30})$. Then for for all $\theta \in \Theta$, $E[h(\theta)|x_{123}]$ is continuously differentiable with bounded derivatives.

DP7. Let $K(\cdot)$ denote the kernel function used in matching regressors for the first choice. Then $K(\cdot)$ is of bounded variation, is bounded on its support, and integrates to 1.

DP8. Let $h_n$ denote the bandwidth sequence used in the matching of the first choice regressors. Then $h_n \to 0$ and $nh_n^{k_n} \to \infty$.

The above conditions suffice for point consistency of our proposed estimator as stated in the following theorem:

**Theorem 3.2** Let $\hat{\kappa} \equiv \hat{\beta}, \hat{\gamma}$ be defined as maximizers of $G_n^{DP}(\beta, g)$ and let $\kappa_0 \equiv \beta_0, \gamma_0$. Then under conditions DP1 – DP8,

$$
\hat{\kappa} \overset{p}{\to} \kappa_0
$$

The crucial aspect of the proof of the consistent is the point identification of both the regression coefficients and that on the lagged dependent variable. Consequently, we outline the steps for the point identification result in the appendix.

## 4 Simulation Study

In this section we explore the relative finite sample performances of the new proposed estimation procedures in cross-sectional and panel data (both static and dynamic) designs.

We generated 1000 replications of the following designs: For the cross-sectional model we generated data for 3 choices using a model with 3 latent utilities: $y_0^* = 0; \ y_1^* = x_{11} + x_{12}\beta_0 + \epsilon_{i1}; \ y_2^* = x_{21} + x_{22}\beta_0 + \epsilon_{i2}$ For the covariates in this model for the first latent utility $x_{11}, x_{12}$ each standard normal, independent of each other and all other variables. For the second latent utility $x_{12}, x_{22}$ are each bernoulli, $p = 0.5$. For the disturbance terms in this model

$\epsilon_{i1}, \epsilon_{i2}$ bivariate normal, mean 0, variance 1, correlation 0.5.

The one regression coefficient to be estimated was $\beta_0 = 1.25$ Table 1 below reports mean bias, median bias and RMSE for 50, 100, 200 and 400 observations We also explored the
effect of dimensionality on the estimator by considering a design with an additional discrete regressor. Tables 2 and 3 report the same statistics for each of the two regression coefficients. Table 2 for the one step estimator and Table 3 for the two step estimator.

For panel data with fixed effects, we simulated from a similar design but now two time periods, so for each person the error term is 4 dimensional (2 choices for each period). Here we simulated from a 4 variate normal, mean 0 variances 1; correlation across choices 0.5; correlation for same choice error across time 0.25; correlation for errors differing by choice and time to be 0; So for example, the error for second choice in first time period is uncorrelated with error for first choice in second time period; The individual specific and choice specific effects were set to to the constant 0 for each choice and each individual.

For the dynamic multinomial choice we generated data from 4 periods 0,1,2,3 3 choices like before; normalize $y_{0}^* = 0$.

We set latent utilities to $y_{0}^* = 0; \ y_{1}^* = \alpha_{i1} + x_{11} + x_{12}\beta_0 + \gamma_0[I(y_{i,t-1} = 1) + \epsilon_{i1}; \ y_{2}^* = \alpha_{i2} + x_{21} + x_{22}\beta_0 + \epsilon_{i2}$ parameters of interest are $\beta_0, \gamma_0$. First component of regressors are normal, second bernoulli as before. We assumed errors iid across time but bivariate normal across choices, mean 0, variance 1, correlation 0.5;

Finally, we also considered a discrete design. This design was identical to the initial cross sectional design except for one change. Now make the first component of $x_{1}$ discrete instead of standard normal. Specifically let it take the values -2,-1,1,2 each with probability 0.25. Note that in this design the regression coefficients are not point identified. But an advantage of our new local rank estimator is that it yields sharp bounds in this case. So in the simulation here we plot the objective function for one replication of 5000 observations. This is to demonstrate how our approach can provide meaningful bounds that can be used for inference even though point identification is lost. We compare this plot to the continuous design when the model is point identified.

Results are reported in the tables and graphs below. As demonstrated, the performance is in line with the asymptotic theory. Specifically, the cross sectional estimator is root-$n$ consistent as both the bias and RMSE shrink at the parametric rate. Furthermore, the two step estimator performs better in finite sample. This was also anticipated as it is based on matching indexes instead of regressors.

The static panel data estimator appears also to be consistent, but appears to converge more slowly in terms of bias and RMSE. It takes samples sixes that are larger than 800 before the estimator performs adequately well.

The estimation works as desired in the discrete design as well. Here the objective function appears to have multiple optima, indicating a bound of the coefficient to be in-between 1 and 1.5. This is in contrast to the continuous design when the objective function has a unique maximum at the true value of $\beta_0$. 

14
Table 1: Cross Sectional Estimator

<table>
<thead>
<tr>
<th>$N$</th>
<th>Mean</th>
<th>Median</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.1656</td>
<td>0.1900</td>
<td>0.5366</td>
</tr>
<tr>
<td>100</td>
<td>0.0173</td>
<td>0.0100</td>
<td>0.4656</td>
</tr>
<tr>
<td>200</td>
<td>0.0410</td>
<td>0.0500</td>
<td>0.3389</td>
</tr>
<tr>
<td>400</td>
<td>0.0001</td>
<td>-0.0010</td>
<td>0.2544</td>
</tr>
</tbody>
</table>

Table 2: Cross Sectional Estimator

<table>
<thead>
<tr>
<th>$N$</th>
<th>Mean</th>
<th>Median</th>
<th>RMSE</th>
<th>Mean</th>
<th>Median</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>-0.2022</td>
<td>-0.1949</td>
<td>0.5130</td>
<td>-0.1828</td>
<td>-0.1160</td>
<td>0.4855</td>
</tr>
<tr>
<td>500</td>
<td>-0.0819</td>
<td>-0.0365</td>
<td>0.4213</td>
<td>-0.1393</td>
<td>-0.1271</td>
<td>0.3857</td>
</tr>
<tr>
<td>1000</td>
<td>-0.0652</td>
<td>-0.0413</td>
<td>0.1010</td>
<td>-0.1312</td>
<td>-0.0513</td>
<td>0.3531</td>
</tr>
</tbody>
</table>

Table 3: Cross Sectional 2- Step Estimator

<table>
<thead>
<tr>
<th>$N$</th>
<th>Mean</th>
<th>Median</th>
<th>RMSE</th>
<th>Mean</th>
<th>Median</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>-0.2528</td>
<td>-0.1384</td>
<td>0.5204</td>
<td>-0.0936</td>
<td>0.3586</td>
<td>-0.0999</td>
</tr>
<tr>
<td>500</td>
<td>-0.0717</td>
<td>-0.0104</td>
<td>0.3476</td>
<td>-0.0332</td>
<td>-0.0334</td>
<td>0.2518</td>
</tr>
<tr>
<td>1000</td>
<td>-0.0356</td>
<td>-0.0154</td>
<td>0.2132</td>
<td>-0.0182</td>
<td>-0.0126</td>
<td>0.1503</td>
</tr>
<tr>
<td>N</td>
<td>Mean</td>
<td>Median</td>
<td>RMSE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>-------</td>
<td>--------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.4839</td>
<td>0.7500</td>
<td>0.6465</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>0.3579</td>
<td>0.6070</td>
<td>0.5861</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>800</td>
<td>0.2344</td>
<td>0.2910</td>
<td>0.5073</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td>0.1284</td>
<td>0.1260</td>
<td>0.4255</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Dynamic Panel Data Estimator

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>Mean</th>
<th>RMSE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>-0.2403</td>
<td>-0.2380</td>
<td>0.2441</td>
<td>0.2452</td>
</tr>
<tr>
<td>800</td>
<td>-0.2188</td>
<td>-0.2295</td>
<td>0.2365</td>
<td>0.2445</td>
</tr>
<tr>
<td>1600</td>
<td>-0.1610</td>
<td>-0.1955</td>
<td>0.2226</td>
<td>0.2374</td>
</tr>
<tr>
<td>3200</td>
<td>-0.1029</td>
<td>-0.1874</td>
<td>0.1898</td>
<td>2324</td>
</tr>
</tbody>
</table>

Figure 1: Objective function for cross-sectional model 1

Figure 2: Objective function for cross-sectional model 2
5 Extensions- Large J (incomplete)

Estimation of discrete-choice models in which consumers face high-dimensional choice sets is encountered in many empirical settings, in fields such as marketing, urban economics and consumer choice with many brands. These models are both computationally and theoretically challenging. These models are in contrast to what we considered in the previous sections of the paper where we treated the number of choices, $J$, as fixed where as the sample size $n$ was regarded as large to develop our asymptotic theory. In this section we propose new algorithms that aims to achieve dimension reduction with the purpose of estimating the finite dimensional regression coefficients with adequate finite sample performance. Under our conditions to algorithms will converge to an estimator of the regression coefficients that is consistent and asymptotically normal when $J$ is large. It is important to note that this
will also be the case for the local rank procedure proposed in the previous sections\textsuperscript{1} but the local rank estimator may have poor finite sample properties in these designs.

Estimating multinomial choice models with very large choice sets have been considered in the econometrics literature. The large number of choices is not problematic in the multinomial logit model McFadden (1978). This is because under the IIA condition smaller subsamples of the choices can be selected to estimate the regression coefficients. However as is well known, the IIA assumption is not constant with most behavioral models of utility maximization. Berry, Linton, and Pakes (2004) develop asymptotic theory for GMM estimators in the study of differentiated product markets when the number of products grows large. Recently, Fox (2007) proposed using a subset of choices for Manski (1975) maximum score estimator. But this approach requires an exchangeability condition on the disturbance terms which is not much of a generalization of the original IIA condition. Thus a procedure not based on IIA would be welcome in the literature due to the many empirical papers of models with a large number of choices. We note that our local rank estimator in the previous section does not impose an IIA condition, so our algorithm for the many \( J \) case will be based on that. Our local rank estimator was based on an objective function for a set \( J \). Our algorithm is based on optimizing the the objective function with respect to \( J \) and the regression coefficients \( \beta \).

The approach we take is based on exploring the mean squared estimator of our rank based procedure. One approach to constructing a data driven method for choosing the number of choices from a large "consideration set", is to explore the mean squared error (MSE) of the rank estimator as a function of \( J \), and the sample size. In that sense we would be regarding the choice of \( J \) analogously to selecting a tuning parameter (i.e. bandwidth) in standard nonparametric estimation. Interestingly, the analysis of \( J \) as a function of the MSE was explored in Berry, Linton, and Pakes (2004) for the multinomial probit.

Our approach to doing so will be based on the linear representation of the MRC shown in Sherman (1993). As we explain below, a crucial parameter that ways into our calculations, is the probability associated with the weight function, which recall, is turned on when the regressors associated with the other choices besides the first one, match up. Specifically,

\[
 w_{im} = I[x_i^{(2)} = x_m^{(2)}]I[x_i^{(3)} = x_m^{(3)}]...I[x_i^{(J)} = x_m^{(J)}] 
\]

where here we have assumed discrete support of their regressors, recognizing that in the continuous case, indicator weights could be replaced with kernel weights.

A crucial parameter for our calculations below is \( p(J) = E[w_{im}] \). Generally speaking we can say \( p(J) \) is decreasing in \( J \). (For an illustrative example, assuming the \( J \) indicators are iid, \( p(J) = p^{J} \)). This monotonicity will result in the variance of the estimator being an increasing function of \( J \) yet the bias of the estimator being decreasing in \( J \). This will imply we can choose a \( J \) to minimize the MSE of the rank estimator, analogously to choosing a bandwidth in nonparametric estimation.

\textsuperscript{1}In fact, the new algorithms make use of the local rank procedure for smaller subsets of choices.
To calculate the bias and variance as a function of $J$, we use the results in Sherman (1993). Consequently, we will keep our notation here as close as possible to that used in that paper.

Taking expectations conditional on the regressors, our objective function is of the form

$$Q_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq m} I[p_i > p_m] I[x_i' \beta > x_m' \beta] w_{im}$$

Let $Q(\beta) = E[Q_n(\beta)]$.

Following Sherman (1993), the asymptotic variance of the rank estimator will be related to the second derivative of $Q(\beta)$, which itself is a smooth function of $\beta$ based on smoothness conditions on the distribution of the index function for the first choice. The asymptotic variance has a "Hessian" term, based on the second derivative, and an outer score term based on the first derivative. Note the second derivative of $Q(\beta)$, and hence the Hessian term will be of order $p(J)$ because of the presence of the weight function, as will be the outer score term. Consequently, following Sherman (1993), the asymptotic variance will be

$$V^{-1} \Delta V^{-1} = O(p(J))^{-1} O(p(J)) O(p(J))^{-1} = O(p(J))^{-1}$$

This agrees with our intuition that the variance of the estimator increases in $J$. Furthermore, note this also implies that the finite sample variance is of order $\frac{1}{n} p(J)^{-1}$.

Having said that, our framework here will be different from Sherman (1993) in terms of the bias term. Specifically, for a given $J < \bar{J}$ where $\bar{J}$ denotes the number of choices available in the consideration set, we do not have that $\nabla \beta Q(\beta_0) = 0$

This is because that using a subset of choices will result in a bias, since we are not imposing the IIA type assumption in Manski (1975). The bias will be of order $p(J) - p(\bar{J})$ which by assuming that $\bar{J}$ is of a larger order than $J$, the bias of our estimator is $O(p(J))$.

Collecting these arguments, the MSE of the estimator is of order

$$MSE(J) \approx p(J)^2 + \frac{p(J)^{-1}}{n}$$

Optimizing this with respect to $p(J)$ gives us $p(J) = O(n^{-1/3})$.

This can be useful in selecting $J$ in finite samples, since we can estimate $p(J)$ from the data as it varies with different designs. For example, if the regressors vector is one dimensional for each choice, and independently and identically distributed across choices then $p(J) = p$.

This would imply that, as a function of the sample size, we would set $J$ to be of order $J = -O\left(\frac{\log n}{3 \log p}\right)$. Then use this choice of $J$ for the rank estimator.
Another thing that can be done is to simultaneously estimate \( J \) and \( \beta \). To do this one can add the rank objective function to the negative of the MSE function. Then maximize the composite objective function with respect to \( J, \beta \).

Specifically, for each \( J \), evaluate \( G_n(\beta, J)) - \text{MSE}(J) = S_n(\beta, J) \).

where

\[
G_n(\beta, J) = \frac{1}{n(n-1)} \sum_{i \neq j} I[y_i > y_j]I[x_{1i}^J > x_{1j}^J]w_{ij}^{(J)}
\]

and maximize \( S_n(\beta, J) \) wrt \( \beta, J \).

Thus collecting these results an algorithm would be based based on the following steps:

1. let \( \bar{J} \) denote the number of possible choices the consumer faces.
2. From the data estimate the unconditional probability of each potential choice, using the sample fraction of times that choice was made in the data.
3. Our rank based objective function in the previous sections was for a given \( J \), the number of choices. Denote this objective function by \( G_n(J, \beta) \).
4. Estimate the MSE minimizing \( J, \bar{J} \) using the method described above, and estimate \( \beta \) by optimizing \( G_n(\bar{J}, \beta) \).
5. Alternatively, for each \( J = 1, \ldots, \bar{J} \), optimize \( S_n(J, \beta) \) with respect to \( \beta \), using the \( J \) choices with the highest unconditional probabilities, This results in \( \bar{J} \) objective function values, each evaluated at the optimal \( \beta \) value for each \( J \).
6. The estimator of \( \beta \) is the value of \( \beta \) that maximizes \( S_n(J, \beta) \) across all values of \( J \) from 1 to \( \bar{J} \).

Under strong sparsity conditions where the values of the choice probabilities are strictly separated, standard arguments such as those used in the previous section can be used to establish root-\( n \) consistency and asymptotic normality of this algorithm. As mentioned the original estimator will be consistent but could have poor finite sample properties when evaluating \( G_n(\bar{J}, \beta) \) has too few observations receive positive weight. Since we are not assuming IIA, using a subset of choices can induce a bias of the estimator, which is the price to pay for reducing the variability of the original estimator in the large \( J \) setting.

Interestingly, for the large \( J \) problem, an algorithm based on random sparse projection was proposed by Chiong and Shum (2014). Convergence of their algorithm is proven but further asymptotic properties are not. Further comparison of the two proposed procedures is worth pursuing both theoretically and in finite samples.
6 Conclusions

In this paper we proposed new estimation procedures for semi parametric multinomial choice models. The local rank based procedure was shown to be root $n$ consistent and asymptotically normal, even in designs where no smoothing parameters were required. The pairwise differencing readily extended to time differencing, enabling a consistent estimator for a panel data estimator of a model with choice and individual specific effects. Furthermore we attained anew identification result for a dynamic multinomial choice model with lagged discrete dependent variables. Finally, we considered the empirical settings where there are large number of choices facing the consumer. We propose a new algorithm which simultaneous reduces the number of choices based on features of the data and estimates the regression coefficients.

The work hear leaves many open areas for future research. For example formal asymptotic properties of the algorithm dimension reduction estimator have to be established. This also case with the dynamic model. Finally, in both panel data settings the propose procedure suffers from a curse of dimensionality in the number of choices. It is thus an open question if the procure results in a rate optimal estimator. Rate optimality for dynamic binary choice models was established in Seo and Otsu (2015), but such bounds are lacking in the multinomial case.

References


A Point Identification for Dynamic Model

A.1 Model Description

\[ i \in \{1, \ldots, n\} \equiv \mathcal{N}, j \in \{0, 1, 2\} \equiv \mathcal{J}, t \in \{0, 1, 2, 3\} \equiv \mathcal{T} \]

\[
\begin{align*}
  u_{1t} &= x_{1t, 1} + \beta x_{1t, 2} + \gamma [y_{t-1} = 1] + \alpha_1 + \epsilon_{1t} \\
  u_{2t} &= x_{2t, 1} + x_{2t, 2} \beta + \alpha_2 + \epsilon_{2t} \\
  u_{0t} &= 0
\end{align*}
\]

\[ y_t = j \iff u_{jt} = \max \{u_{0t}, u_{1t}, u_{2t}\} \]

A.2 Identification

Denote \( d_t = \mathbb{I}[y_t = 1] \),

\[ A = \{\mathbb{I}[y_0 = 1] = d_0, y_1 = 1, y_2 \neq 1, \mathbb{I}[y_3 = 1] = d_3\} \]

and

\[ B = \{\mathbb{I}[y_0 = 1] = d_0, y_1 \neq 1, y_2 = 1, \mathbb{I}[y_3 = 1] = d_3\} \]

Let

\[ \Omega = \{x, \alpha, x_{21} = x_{22}, x_{12} = x_{13}, x_{22} = x_{23}\} \]

be a conditioning set. By definition, for all \( t \in \mathcal{T} \setminus \{0\} \),

\[
\Pr (y_t = 1|\Omega) = \Pr \left( \frac{x_{1t, 1} + \beta x_{1t, 2} + \gamma [y_{t-1} = 1] + \alpha_1 + \epsilon_{1t} > 0}{x_{1t, 1} + \beta x_{1t, 2} + \gamma [y_{t-1} = 1] + \alpha_1 + \epsilon_{1t} > x_{2t, 1} + x_{2t, 2} \beta + \alpha_2 + \epsilon_{2t}} | \Omega \right)
\]
Then,

\[
\Pr(A|\Omega) = \frac{d_0}{p_0} (1 - p_0)^{1-d_0} \\
\times \Pr\left( x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > 0 \middle| \Omega \right) \\
\times \left( 1 - \Pr\left( x_{12,1} + \beta x_{12,2} + \gamma + \alpha_1 + \epsilon_{12} > 0 \middle| \Omega \right) \right) \\
\times \Pr\left( x_{13,1} + \beta x_{13,2} + \alpha_1 + \epsilon_{13} > 0 \middle| \Omega \right) \\
\times \left( 1 - \Pr\left( x_{13,1} + \beta x_{13,2} + \alpha_1 + \epsilon_{1t} > x_{23,1} + x_{23,2} + \alpha_2 + \epsilon_{23} \middle| \Omega \right) \right)^{1-d_3}
\]

and

\[
\Pr(B|\Omega) = \frac{d_0}{p_0} (1 - p_0)^{1-d_0} \\
\times \left( 1 - \Pr\left( x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > 0 \middle| \Omega \right) \right) \\
\times \Pr\left( x_{12,1} + \beta x_{12,2} + \alpha_1 + \epsilon_{12} > 0 \middle| \Omega \right) \\
\times \Pr\left( x_{13,1} + \beta x_{13,2} + \gamma + \alpha_1 + \epsilon_{13} > 0 \middle| \Omega \right) \\
\times \left( 1 - \Pr\left( x_{13,1} + \beta x_{13,2} + \gamma + \alpha_1 + \epsilon_{1t} > x_{23,1} + x_{23,2} + \alpha_2 + \epsilon_{23} \middle| \Omega \right) \right)^{1-d_3}
\]
If $d_3 = 0,$

$$\frac{\Pr(A|\Omega)}{\Pr(B|\Omega)} = \Pr \left( \begin{array}{c}
x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > 0 \\
x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > x_{21,1} + x_{21,2} \beta + \alpha_2 + \epsilon_{21} | \Omega \end{array} \right)$$

$$\times \left( 1 - \Pr \left( \begin{array}{c}
x_{12,1} + \beta x_{12,2} + \gamma + \alpha_1 + \epsilon_{12} > 0 \\
x_{12,1} + \beta x_{12,2} + \gamma + \alpha_1 + \epsilon_{12} > x_{22,1} + x_{22,2} \beta + \alpha_2 + \epsilon_{22} | \Omega \end{array} \right) \right)$$

$$\times \left( 1 - \Pr \left( \begin{array}{c}
x_{13,1} + \beta x_{13,2} + \gamma d_3 + \alpha_1 + \epsilon_{13} > 0 \\
x_{13,1} + \beta x_{13,2} + \gamma d_3 + \alpha_1 + \epsilon_{13} > x_{23,1} + x_{23,2} \beta + \alpha_2 + \epsilon_{23} | \Omega \end{array} \right) \right)$$

$$/ \left( \left( 1 - \Pr \left( \begin{array}{c}
x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > 0 \\
x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > x_{21,1} + x_{21,2} \beta + \alpha_2 + \epsilon_{21} | \Omega \end{array} \right) \right) \right)$$

$$\times \Pr \left( \begin{array}{c}
x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{12} > 0 \\
x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{12} > x_{22,1} + x_{22,2} \beta + \alpha_2 + \epsilon_{22} | \Omega \end{array} \right)$$

$$\times \left( 1 - \Pr \left( \begin{array}{c}
x_{13,1} + \beta x_{13,2} + \gamma + \alpha_1 + \epsilon_{13} > 0 \\
x_{13,1} + \beta x_{13,2} + \gamma + \alpha_1 + \epsilon_{13} > x_{23,1} + x_{23,2} \beta + \alpha_2 + \epsilon_{23} | \Omega \end{array} \right) \right) \right)$$

$$= \Pr \left( \begin{array}{c}
x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > 0 \\
x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > x_{21,1} + x_{21,2} \beta + \alpha_2 + \epsilon_{21} | \Omega \end{array} \right)$$

$$\times \left( 1 - \Pr \left( \begin{array}{c}
x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{13} > 0 \\
x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{13} > x_{22,1} + x_{22,2} \beta + \alpha_2 + \epsilon_{23} | \Omega \end{array} \right) \right)$$

$$/ \left( \left( 1 - \Pr \left( \begin{array}{c}
x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{12} > 0 \\
x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{12} > x_{22,1} + x_{22,2} \beta + \alpha_2 + \epsilon_{22} | \Omega \end{array} \right) \right) \right)$$

$$\times \left( 1 - \Pr \left( \begin{array}{c}
x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > 0 \\
x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > x_{21,1} + x_{21,2} \beta + \alpha_2 + \epsilon_{21} | \Omega \end{array} \right) \right) \right)$$
Similarly, if $d_3 = 1$,

$$\frac{\Pr (A|\Omega)}{\Pr (B|\Omega)} = \Pr \left( \begin{array}{c}
  x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > 0 \\
  x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > x_{21,1} + x_{21,2} \beta + \alpha_2 + \epsilon_{21} |\Omega
\end{array} \right) \\
\times \left( 1 - \Pr \left( \begin{array}{c}
  x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{12} > 0 \\
  x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{12} > x_{22,1} + x_{22,2} \beta + \alpha_2 + \epsilon_{22} |\Omega
\end{array} \right) \right) \\
\times \Pr \left( \begin{array}{c}
  x_{12,1} + \beta x_{12,2} + \alpha_1 + \epsilon_{13} > 0 \\
  x_{12,1} + \beta x_{12,2} + \alpha_1 + \epsilon_{13} > x_{22,1} + x_{22,2} \beta + \alpha_2 + \epsilon_{23} |\Omega
\end{array} \right) \\
/ \left\{ \begin{array}{c}
  \left( 1 - \Pr \left( \begin{array}{c}
  x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > 0 \\
  x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > x_{21,1} + x_{21,2} \beta + \alpha_2 + \epsilon_{21} |\Omega
\end{array} \right) \right) \\
\times \Pr \left( \begin{array}{c}
  x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{12} > 0 \\
  x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{12} > x_{22,1} + x_{22,2} \beta + \alpha_2 + \epsilon_{22} |\Omega
\end{array} \right) \\
\times \Pr \left( \begin{array}{c}
  x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{13} > 0 \\
  x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{13} > x_{22,1} + x_{22,2} \beta + \alpha_2 + \epsilon_{23} |\Omega
\end{array} \right) \right\}$$

$$= \Pr \left( \begin{array}{c}
  x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > 0 \\
  x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > x_{21,1} + x_{21,2} \beta + \alpha_2 + \epsilon_{21} |\Omega
\end{array} \right) \\
\times \left( 1 - \Pr \left( \begin{array}{c}
  x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{12} > 0 \\
  x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{12} > x_{22,1} + x_{22,2} \beta + \alpha_2 + \epsilon_{22} |\Omega
\end{array} \right) \right) \\
/ \left\{ \begin{array}{c}
  \Pr \left( \begin{array}{c}
  x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{13} > 0 \\
  x_{12,1} + \beta x_{12,2} + \gamma d_3 + \alpha_1 + \epsilon_{13} > x_{22,1} + x_{22,2} \beta + \alpha_2 + \epsilon_{23} |\Omega
\end{array} \right) \\
\times \left( 1 - \Pr \left( \begin{array}{c}
  x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > 0 \\
  x_{11,1} + \beta x_{11,2} + \gamma d_0 + \alpha_1 + \epsilon_{11} > x_{21,1} + x_{21,2} \beta + \alpha_2 + \epsilon_{21} |\Omega
\end{array} \right) \right) \right\}$$

It can be seen that for all $d_3 \in \{0, 1\}$, we have $\Pr (A|\Omega) > \Pr (B|\Omega)$ if and only if

$$x_{11,1} + \beta x_{11,2} + \gamma d_0 > x_{12,1} + \beta x_{12,2} + \gamma d_3$$

**Objective Function**

The objective function for the corresponding maximum score estimator can be written as

$$Q_n (b, g) = \frac{1}{n} \sum_{i=1}^{n} K_n (x_{i12,1} - x_{i13,1}) \mathbb{I} [x_{i12,2} = x_{i13,2}] \mathbb{I} [x_{i21} = x_{i22}] \mathbb{I} [x_{i22} = x_{i23}] \\
\times (\mathbb{I} [A_i] - \mathbb{I} [B_i]) \times \mathbb{I} [x_{i11,1} + bx_{i11,2} + gd_i0 > x_{i12,1} + bx_{i12,2} + gd_i3]$$