Closed Form Estimators for a Class of Semiparametric Multinomial Choice Models

Daniel Ackerberg*
Dept. of Economics, University of Texas at Austin

Xiao Meng
Dept. of Finance and Economics, Texas State University

Haiqing Xu
Dept. of Economics, University of Texas at Austin

PRELIMINARY - PLEASE DO NOT DISTRIBUTE

November 13, 2020

Abstract

We consider a semiparametric multinomial choice model that allows for an arbitrary joint distribution of choice specific unobservables that are independent of explanatory variables. This model permits relatively flexible substitution patterns between choices. To minimize computational difficulties, we restrict attention on estimators of the model that can be expressed in closed form. We combine and extend various results from the existing literature to enforce economic restrictions implied by the model and to attain "as efficient estimators as we can" - given the closed form requirement. Some aspects of our estimators achieve the semiparametric efficiency bound, while others do not. In Monte-Carlo experiments, we study how various strategies increase efficiency, and compare the efficiency of our best estimators to computationally more challenging, non-closed form, estimators that are efficient.

*daniel.ackerberg@gmail.com, xm_11@txstate.edu, h.xu@austin.utexas.edu. All errors are our own.

1 Introduction

In this paper we propose and investigate *closed-form* estimators of the following semiparametric multinomial discrete choice model where the utility consumer i obtains from choice j is given by

$$U_{ij} = X_{ij}'\beta_j + \epsilon_{ij} \tag{1}$$

The deterministic component of utility is a linear index of the observables X_{ij} where the coefficients β_j are permitted to vary across choices j. Our model is semiparametric in the sense that we allow for an *arbitrary joint distribution* of $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{iJ})$, but we do assume ϵ_i is independent of X_i.

We restriction attention to closed form estimators of various aspects of this model because in an empirical situation, we believe avoiding numeric optimization (or root solving) can be a large benefit. For example, likelihood functions based on models like (1) are not generally globally concave. Local minima imply that the numeric optimization needed to maximize the likelihood function can be time consuming and error prone. Moreover, if one is trying to be flexible with the joint distribution of the error terms in the context of maximum likelihood estimation, the number of parameters one needs to numerically maximize over can increase very quickly, likely increasing these computational costs and the potential for optimization mistakes. Our closed form estimators avoid this while being completely non-parametric on the joint distribution of the error terms.

However, we do lose something by restricting attention to closed form estimators. In particular, our estimators will not enforce all the economic restrictions implied by the discrete choice model that, e.g., a sieve based maximum likelihood estimator would (e.g. Chen (2007)). However, by combining and extending results across a number of literatures, we are able to propose estimators that we believe can be helpful in practice (and some aspects of the estimated model will in fact achieve the semiparametric efficiency bound). One can think of this paper as an effort to combine different techniques to determine "the best we can do" given our restriction to closed form estimators.[1]

Our model (1) can thought of as a generalization of a multinomial probit model with correlated errors - a model that is already thought to allow fairly flexible substitution patterns. For example, in a three choice case, ϵ_{i1} and ϵ_{i2} might be more highly correlated than ϵ_{i1} and ϵ_{i3}. This means that, all else equal, changes in X_{i1} have a bigger impact on the probability of choosing alternative 2 than on the probability of choosing alternative 3. In other words, alternatives 1 and 2 are stronger substitutes than are alternatives 1 and 3. Like the multinomial probit model with correlated errors, the model we study allows these differential substitution patterns, but in a non-parametric fashion as we do not assume joint normality. That said, the assumption of independence between ϵ_i and X_i is a restriction, ruling out models with random coefficients (e.g. the mixed logit of McFadden and Train (2000) or Berry, Levinsohn, and Pakes (1995)). But as we detail later, our model is neither

[1] Because the set of all possible closed-form estimators would be extremely hard to characterize, we are not able to formally prove that our estimators are the most efficient within such a set. However, we show in Monte-Carlo's that our estimators perform reasonably well compared to non-closed-form, efficient, parametric estimators.

more nor less general than the typical random coefficients model.

In our goal to construct the best $\sqrt{n}$-consistent, closed-form, estimators for the β_j's in (1) we can, we combine and extend results in a number of important papers on semiparametric index models (e.g. Powell, Stock, and Stoker (1989), Klein and Spady (1993), Ahn, Ichimura, Powell, and Ruud (2018 - henceforth AIPR) and Allen and Rehbeck (2019 - henceforth AR)), results on efficiency gains of single Newton-Steps (e.g. Lehmann (1983), Horowitz (1998)), and kernel estimation techniques that can enforce shape restrictions (e.g. Stone 1977, Cleveland, 1979, Fan, 1992, 1993).[2] Focusing at this point on AIPR and AR, AIPR also propose closed form estimators for a class of multi-index models similar to the multinomial choice problem in (1). However, the multi-index estimators proposed by AIPR can only identify each index up to its own scale. As such AIPR only consider a restricted form of (1), i.e.

$$U_{ij} = X_{ij}'\beta + \epsilon_{ij}$$

where the β is assumed to be the same across alternatives. We are able to estimate the more general model (1) because we make use of a symmetry condition that AIPR do not use, and our estimator illustrates how the information in this symmetry condition can be used in a closed form way. On the other hand, AIPR's estimator can directly apply when elements of X_{ij} are discrete, while ours cannot. There are other differences in the estimators - our estimator directly uses derivatives of choice probabilities, while AIPR looks at variation across level sets of choice probabilities - this means that our estimator has less tuning parameters than AIPR. In any case, given the tradeoffs between the two estimators, we advocate a combination of the two that can accomodate both the more general model with β_j's and discrete X_{ij}'s.

AR (2019) have previously utilized the symmetry condition that we use, showing that it holds in a very broad set of models, including choice between bundles, matching, and multinomial choice. However, their focus is on identification and not estimation. We focus on estimation, and in performing that estimation in closed form. Another difference from both AIPR and AR is that we consider not only β but also the joint distribution of ϵ_i. We show how we can also obtain a closed form estimator of a function this joint distribution, and how we can enforce some of the restrictions implied by the choice model on this estimator. This is useful for calculation of counterfactuals, e.g. elasticities of choice probabilities w.r.t. observables.

After proposing our basic estimators, we illustrate how they can be improved in terms of effi-

[2] There are other estimators that also treat multinomial models like (1) semiparametrically. In particular maximum score estimators (see Manski (1985) and Fox (2007)) also do not fully specify the joint distribution of ϵ_i. But maximum score estimators also require numerical optimization, and this is well known to be particularly challenging given that score functions are not continuous. Maximum score estimators are also not $\sqrt{n}$-consistent without smoothing, e.g. Horowitz (1992), Yan (2018). While this smoothing can also help with numeric optimization, these objective functions are still challenging to maximize in practice. Maximum score estimators also require different assumptions on the joint distribution of ϵ_i than we do - maximum score estimators can allow some dependence between ϵ_i and X_i that we cannot (e.g. median independence rather than full independence), but our model allows arbitrary correlations between the elements of ϵ_i while the maximum score estimator needs to restrict this with what are often perceived as fairly high level assumptions

ciency. This is done in various ways - using a single Newton-Step to improve efficiency, leveraging overidentification efficiently, and using Local Linear Kernels (Stone (1977)) to impose the AR cross-derivative restrictions in kernel estimation - all preserve the closed form nature of the estimators. We illustrate the performance of all these estimators in Monte-Carlo experiments, in particular assessing 1) how much precision we lose, relative to non-closed form MLE based estimators, and 2) how much precision our various improvements add to the estimators. This helps inform us about the tradeoff between potentially more reliable closed form solutions and loss of precision.

2 Model specification

Suppose an individual (decision-maker) i makes a choice Y_i among $J+1$ alternatives, i.e. $\mathcal{J}_0 \equiv \{0, 1, \cdots, J\}$, where $J \geq 2$. By convention, alternative 0 refers to the "outside" choice and the utility of such an option is normalized to zero, i.e. $U_{i0} = 0$. Let further $\mathcal{J} \equiv \mathcal{J}_0/\{0\}$. Moreover, individual i's utility for option $j \in \mathcal{J}$ is given by

$$U_{ij} = X_{ij}'\beta_j - \epsilon_{ij},$$

where $X_{ij} \in \mathbb{R}^{d_j}$ $(d_j \geq 2)$ is a vector of covariates, $\beta_j \in \mathbb{R}^{d_j}$ is the coefficient, and $\epsilon_{ij} \in \mathbb{R}$ is an unobserved shock to utility. For notation simplicity, we denote $X \equiv (X_1', \cdots, X_J') \in \mathbb{R}^{\sum_{j\in\mathcal{J}} d_j}$, $\beta \equiv (\beta_1' \cdots, \beta_J')' \in \mathbb{R}^{\sum_{j\in\mathcal{J}} d_j}$, and $\epsilon \equiv (\epsilon_1, \cdots, \epsilon_J)' \in \mathbb{R}^J$. Moreover, we denote the density function of ϵ_i by $f_\epsilon : \mathbb{R}^J \to \mathbb{R}^+$.

Thus, individual i's optimal decision on the discrete choice set can be described as

$$Y_i = \begin{cases} 0, & \text{if} \quad \max_{k\in\mathcal{J}_1} U_{ik} \leq 0; \\ j, & \text{if} \quad U_{ij} \geq 0 \text{ and } \ U_{ij} \geq \max_{k\in\mathcal{J}_1}\{U_{ik}\}. \end{cases} \tag{2}$$

In some empirical applications, some of the regressors may not vary over alternatives. See e.g. Cameron and Trivedi (2005). In such a situation, the utility function for option j can be modeled as

$$U_{ij} = Z_i'\alpha_j + X_{ij}'\beta_j - \epsilon_{ij},$$

where $Z_i \in \mathbb{R}^{d_z}$ is a vector of covariates associated with all the alternatives (e.g. household/individual demographics), and α_j is its coefficient. Another natural source of Z_i are covariates associated with the outside option, if the choice probabilities vary with these outside-choice specific variables.[3] For simplicity of the presentation, we first consider our benchmark estimators without common covari-

[3]Namely, suppose the utility from choosing alternative 0 is given by $U_{i0} = X_{i0}'\beta_0 - u_{i0}$, where $X_{i0} \in \mathbb{R}^{k_0}$ is a vector of covariates, $u_{i0} \in \mathbb{R}$ is the error term, and β_0 is the coefficient. Then if we normalize the utility function at alternative 0 to zero, i.e., $\tilde{U}_{i0} = 0$, we also need to specify the utility for alternative j as the difference between the (original) utilities from alternative j and 0, i.e. $\tilde{U}_{ij} = -X_{i0}'\beta_0 + X_{ij}'\beta_j - \tilde{u}_{ij}$, where $\tilde{u}_{ij} = u_{ij} - u_{i0}$. Note that X_{i0} and β_0 are invariant across $j \in \mathcal{J}_1$.

ates. Later we extend our techniques to the more general situation. Note that X_{ij} can implicitly include components that are constant across i. Such components will simply generate alternative specific constant terms, so can be subsumed into different means of the distribution of the u_{ij} across j.

We do not make parametric distributional assumption on ϵ_{ij} - instead we introduce some weak/non-distributional assumptions on the model specification.

Assumption 1 *(i) The distribution of ϵ be absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^J$. Let f_ϵ denote its density function which has a full support on $\mathbb{R}^J$; (ii) The distribution of X be absolutely continuous with respect to the Lebesgue measure on $S_X \subseteq \mathbb{R}^{\sum_{j\in\mathcal{J}} d_j}$, with density denoted by f_X; (iii) There exists no proper linear subspace of $\mathbb{R}^{\sum_{j\in\mathcal{J}} d_j}$ having probability one under the probability distribution of X; (iv) Let $\beta_{11} = 1$.*

Later we relax Condition (ii) to allow for some (but not all) discrete components in X_j. Condition (iii) is made to exclude perfect multi-collinearity, which is also a standard assumption in the semiparametric binary response model literature. See e.g. Manski (1975, 1985). Condition (iv) is a scale normalization on β. This is necessary with our non parametric treatment of $f_\epsilon(\cdot)$, because $(\beta, f_\epsilon(\cdot))$ and $(c \times \beta, f_\epsilon(\cdot/c)/c)$ are observationally equivalent as long as $c > 0$.

Assumption 2 *The covariates X are independent of the error term ϵ, i.e. $X \perp\!\!\!\perp \epsilon$.*

The independence restriction in assumption4 is strong, but a key aspect of our analysis. This rules out random coefficient models, but since our model allows arbitrary correlation in ϵ_j across j, it permits fairly flexible substitution patterns across the alternatives. Defining $W_j = X_j'\beta_j$ and $W = (W_1, \cdots, W_J)$, we have:

Lemma 1 (Semiparametric Multinomial Model) *Suppose Assumption 1 and Assumption 2 hold. Then, the probability distribution of the multinomial choice is given by: for each $j \in \mathcal{J}_1$,*

$$\mathbb{P}(Y = j|X) = \psi_j(W),$$

for some differentiable function $\psi_j : \mathbb{R}^J \to [0, 1]$. Moreover, ψ_j is monotonically increasing in W_j, decreasing in W_k for $k \in \mathcal{J}_1/\{j\}$, and satisfies

$$\frac{\partial \psi_k}{\partial W_j} = \frac{\partial \psi_j}{\partial W_k} < 0. \tag{3}$$

The first part of Lemma 1 defines ψ_j as a smooth functional of the density function f_ϵ. This will be a useful alternative representation of the structural parameter. Note that we can invert the

density function f_ϵ out from $\{\psi_j : j \in \mathcal{J}_1\}$ under some regularity conditions. To see this, let $J = 2$ for simplicity. Note that

$$\Pr(\epsilon_1 \geq t_1; \epsilon_2 \geq t_2) = 1 - \sum_{j \in \{1,2\}} \psi_j(t_1, t_2).$$

By taking the second derivatives of the above equation, we have

$$f_\epsilon(t) = \sum_{j,k,\ell \in \{1,2\}} \frac{\partial^2 \psi_j(t_1, t_2)}{\partial t_k \partial t_\ell}$$

The second part of Lemma 1 is a Slutsky symmetry type condition implied by the structure of the discrete choice "demand" model. We follow Allen and Rehbeck (2019) in using this condition for identification, and parts of our estimation procedures also rely directly on this condition.

2.1 Discussion

On one hand the assumption of independence between ϵ_i and X_i is a strong restriction. On the other, the model is more flexible than logit and nested logit model, as well as multinomial probit models that allow arbitrary correlations between the errors of the different alternatives. These multinomial probit models are already thought to be quite flexible in representing substitution patterns across alternatives (e.g. McCulloch and Rossi (1994), Imbens and Wooldridge (2007)). What our independence restriction does rule out is random coefficients on the X_i's (e.g. the mixed logit of McFadden and Train (2000) or Berry, Levinsohn, and Pakes (1995)). But it is important to note that the model we study is neither more nor less general than a typical model with random coefficients on product characteristics X plus additive errors ϵ_{ij} that are uncorrelated across j (in such a model ϵ_{ij} can be interpreted as including both the ϵ_{ij} and the random coefficient terms containing X's). In a typical random coefficients model, unobserved correlations in preferences across j are a function of the X's. So, for example, two alternatives with the same value of X_{ij}'s will have the same unobserved correlation in preferences with respect to a third alternative. The model we study is more flexible in that it allows completely arbitrary correlations (and thus substitution patterns) across different alternatives j, even those with the same value of X_{ij}'s.[4] On the other hand, random coefficients generate unobserved correlations in preferences across alternatives that change when X_{ij}'s change. The model we consider does not allow this.

[4] Of course, if X_{ij} contained a set of dummy variables corresponding to every product j and one allowed a fully flexible joint distribution of random coefficients on those dummy variables, it would encompass our model (as essentially, those random coefficients would correspond to our ϵ_i).

3 Identification and Basic Estimation

In this section, we first briefly establish identification of our model. While this model is already known to be identified, e.g. Matzkin (1993), our identification proof is useful to reparameterize and reexpress the model in a way that helps for estimation. It also highlights some overidentified aspects of the model. We then focus on the main point of the paper, i.e. proposing closed form estimation procedures for this model that are computationally simple and relatively efficient.

3.1 Identification

First, reparameterize the model as

$$\beta_{js} = \beta_{j1}\gamma_{js}$$

with the normalizations $\beta_{11} = 1$ and $\gamma_{j1} = 1\ \forall j$.[5] With this parameterization, the γ_{js}'s measure "within-choice" relative scales, relative to the first X for that choice, i.e. $\gamma_{js} = \frac{\beta_{js}}{\beta_{j1}}$. For example, for choice 2, the γ_{23} measures the marginal effect of increasing X_{i23} relative to the marginal effect of increasing X_{i21}. The β_{j1}'s measure "across-choice" relative scales, all with respect to the first X_i for each choice, and relative to the normalized $\beta_{11} = 1$. For example β_{j3} measures the marginal effect of increasing X_{i31} on the utility of choice 3 relative to the marginal effect of increasing X_{i11} on the utility of choice 1.[6]

We now state a compact theorem that proves identification of the entire β (up to one normalization, i.e. $\beta_{11} = 1$). We then show explicitly how this result applies to identify the individual elements of our reparameterization, i.e. the γ_{js}'s and β_{j1}'s. Let $\beta_j = (\beta_{j1}, ..., \beta_{jd_j})$ be the vector of β's for choice j.

Theorem 2 *Under Assumptions 1 and 2, β_j is (over) identified by the following equation system:*

$$\beta_j = -\frac{1}{2}\lambda_{kj} \times \mathbb{E}\Big[\frac{\partial \Pr(Y = k \mid X)}{\partial X_j} f_X(X)\Big] = \lambda_{kj} \times \mathbb{E}\Big[\mathbb{I}(Y = k)\frac{\partial f_X(X)}{\partial X_j}\Big], \quad \forall\ j, k = 1, ..., J,$$

where λ_{kj} is a scalar satisfying $\lambda_{kj} = \lambda_{jk} > 0$.

[5] It is straightforward that the sign (i.e. positive, negative, or zero) of β_{js} is identified under weak conditions. For simplicity, we assume that X_{js} for some j and s is known to have a strictly positive coefficient. We can then reindex to make this X_{11} and then normalize its coefficient.

[6] Again, for expositional simplicity, we assume $\beta_{j1} \neq 0$ for all $j = 1, ..., J$.

Note that the second equality holds because

$$\begin{aligned}
& \mathbb{E}\left[\frac{\partial \mathbb{P}(Y=k|X)}{\partial X} f_X(X)\right] + 2\mathbb{E}\left[\mathbb{I}(Y=k)\frac{\partial f_X(X)}{\partial X}\right] \\
= & \int_{-\infty}^{+\infty}\left[\frac{\partial \mathbb{P}(Y=k|X=x)}{\partial x} f_X^2(x)\right] dx + \int_{-\infty}^{+\infty}\left[\mathbb{P}(Y=k|X=x)\frac{\partial f_X^2(x)}{\partial x}\right] dx \\
= & \int_{-\infty}^{+\infty}\left\{\frac{\partial\left[\mathbb{P}(Y=k|X=x) f_X^2(x)\right]}{\partial x}\right\} dx \\
= & \mathbb{P}(Y=k|X=x) f_X^2(x)\Big|_{-\infty}^{+\infty} \\
= & 0
\end{aligned}$$

Using terms of the form $\mathbb{E}\left[\mathbb{I}(Y=k)\frac{\partial f_X(X)}{\partial X}\right]$ to express the implications of a choice model is common in the related literature (e.g. Powell, Stock, and Stoker (1989)) because it is a straightforward object to estimate using kernel techniques.[7]

Theorem 2 simultaneously encompasses both the index restrictions of our model (which relates the effects of changing X_{js} versus $X_{js'}$ for the same alternative j) and the Slutzky symmetry restrictions in Lemma 1 (which relate the the effects of changing X_j versus $X_{j'}$ for two different alternatives). More specifically, with our reparameterization, the theorem first implies that the γ_{js}'s are (over) identified by

$$\gamma_{js} \equiv \frac{\mathbb{E}\left[\mathbb{I}(Y=k)\frac{\partial f_X(X)}{\partial X_{js}}\right]}{\mathbb{E}\left[\mathbb{I}(Y=k)\frac{\partial f_X(X)}{\partial X_{j1}}\right]} \qquad \forall\ k \in \mathcal{J}_1. \tag{4}$$

Intuitively, this is leveraging the fact that because of the linear index structure, the ratio of the derivatives of a choice probability w.r.t. two elements of X_j must be equal to the ratio of the respective β's, i.e. $\gamma_{js} = \frac{\beta_{js}}{\beta_{j1}}$. Note that the overidentification comes from the fact that this is true for any choice probability, i.e. one could also look at how the probability of choice k varies as the two elements of X_j vary. As J increases, these cross derivatives may get small, so in practice (e.g. in our Monte-Carlos), we only use (4) for $k = j$. Note that this identification result is essentially equivalent to existing results in the single index literature (e.g. Powell, Stock, and Stoker (1989)).

In contrast, identification of the "across-choice" relative scales, i.e. the β_{j1}'s, rests on the utility maximizing structure of the multinomial choice model. This identification argument can be interpreted as a special case of the identification results of Allen and Rehbeck (2019), which apply very generally to index models with restrictions from choice theory. This choice model structure is key to Lemma 1 and its implication that $\lambda_{kj} = \lambda_{jk}$ in Theorem 2. Intutively, the choice model implies that $\frac{\partial \mathbb{P}(Y=j|X)}{\partial X_k' \beta_k} = \frac{\partial \mathbb{P}(Y=k|X)}{\partial X_j' \beta_j}$ since it is differences in utilities that what determine choice,

[7] While expressing identification in terms of this representation is useful for us since our eventual goal is estimation, we should note that identification has been established under weaker assumptions than those necessary for this representation.

and locally, it is the same consumers switching from j to k (or vice-versa) that are on each margin. Formally, Theorem 2 implies

$$\frac{\beta_{j1}}{\beta_{k1}} = \frac{\mathbb{E}\left[\mathbb{I}(Y=k)\frac{\partial f_X(X)}{\partial X_{j1}}\right]}{\mathbb{E}\left[\mathbb{I}(Y=j)\frac{\partial f_X(X)}{\partial X_{k1}}\right]}, \quad \forall\ k \neq j. \tag{5}$$

Setting $k=1$ directly identifies the β_{j1} for $j>1$ with our normalization that $\beta_{11}=1$.

Given that the γ_{js}'s are identified, Theorem 2 implies one could also identify the model with equations similar to ((5)) based on differentiating w.r.t. elements of the X_j's other than the first. To utilize this overidentification and reduce the dimensionality of the kernel we end up using to estimate ((5)), define $\gamma_j \equiv (1, \gamma_{j2}, \cdots, \gamma_{jd_j})'$ and $\bar{W}_j = X_j'\gamma_j$, and let $f_{\bar{W}}(\bar{W})$ represent the joint distribution of these indices. Note that $\bar{W}_j = W_j/\beta_{j1}$ as the "normalized index" of W_j. We then can show

$$\frac{\beta_{j1}}{\beta_{k1}} = \frac{\mathbb{E}\left[\mathbb{I}(Y=k)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_j}\right]}{\mathbb{E}\left[\mathbb{I}(Y=j)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_k}\right]}, \quad \forall\ k \neq j. \tag{6}$$

Even when we restrict attention to variation in the indices W_j, note that there are $\frac{J(J-1)}{2}$ of such equations, so these "across-choice" relative scales β_{j1}'s are still over-identified if $J \geq 3$. We discuss how to leverage this over-identification in our identification in the estimation section.

Given that the γ_{js}'s (and β_{j1}'s) are identified, identification of ψ_j is straightforward. Namely, for any $w \in \mathbb{R}^J$, we have:

$$\psi_j(w) = \Pr\big(Y=j|W=w\big). \tag{7}$$

As we discuss later, estimates of ψ_j are necessary for many counterfactuals, e.g. derivatives of choice probabilities w.r.t. X's. To estimate these derivatives, our estimation method will take into account the restriction on ψ established in Lemma 1 for efficiency gains.

3.2 Basic Estimates

We first present very basic closed form estimators for the parameters of the model - i.e. the γ_{js}'s, β_{j1}'s, and ψ_j's. We then present various improvements on some of these estimators intended to increase efficiency. Let $\{(Y_i, X_{i1}, \cdots, X_{iJ}) : i \leq n\}$ be an i.i.d. random sample of $(Y, X_1, \cdots, X_J)$ with sample size n.

3.2.1 Within-Choice Relative Scales γ_{js}'s

First, natural kernel estimates for $\gamma_j \equiv (1, \gamma_{j2}, \cdots, \gamma_{jd_j})'$ based on ((4)) are

$$\tilde{\gamma}_j = \frac{\sum_{i=1}^N \mathbb{I}(Y_i = j)\frac{\partial \hat{f}_X(X_i)}{\partial \tilde{X}_j}}{\sum_{i=1}^N \mathbb{I}(Y_i = j)\frac{\partial \hat{f}_X(X_i)}{\partial X_{j1}}} = \frac{1}{h_j^{d_j-2}} \frac{\sum_{i=1}^N \sum_{n=1;n\neq i}^N \mathbb{I}(Y_i = j)\frac{\partial K_j(\frac{X_{ij}-X_{nj}}{h_j})}{\partial \tilde{u}_j} \prod_{\ell\neq j} K_\ell(\frac{X_{i\ell}-X_{n\ell}}{h_\ell})}{\sum_{i=1}^N \sum_{n=1;n\neq i}^N \mathbb{I}(Y_i = j)\frac{\partial K_j(\frac{X_{ij}-X_{nj}}{h_j})}{\partial u_{j1}} \prod_{\ell\neq j} K_\ell(\frac{X_{i\ell}-X_{n\ell}}{h_\ell})} \quad (8)$$

where $h_j \in \mathbb{R}_+$ is a bandwidth, and $K_\ell : \mathbb{R}^{d_\ell} \to \mathbb{R}$ is a Parzen-Rosenblatt kernel with compact support, in which $u_{\ell 1}$ denotes its first component and $\tilde{u}_\ell$ is the other $d_\ell - 1$ arguments. In the above expression, note that $f_X(X_i)$ is estimated by the leave-one-out kernel density estimator, i.e.

$$\hat{f}_X(X_i) = \frac{1}{(N-1)\prod_{\ell=1}^J h^{d_\ell}} \sum_{n=1;n\neq i}^N \prod_{\ell=1}^J K_\ell(\frac{X_{i\ell} - X_{n\ell}}{h_\ell}).$$

Following Powell, Stock, and Stoker (1989), under regularity conditions, $\hat{\gamma}_j$ is $\sqrt{N}$ consistent and has a limiting normal distribution. One could think about refining this estimator in a couple of ways, first based on the observation above that γ_j is overidentified (i.e. (8) uses own derivatives, but as noted in the prior section, a similar equation holds using cross derivatives), and the second based on the possibility of trying to reduce the dimension of the kernels K_ℓ using the index restrictions of the model. However, we do not pursue this further, because in a moment we provide an alternative refinement to the estimator of γ_j that will attain the semiparametric efficiency bound.[8]

3.2.2 Across-Choice Relative Scales β_{j1}'s

We next form natural kernel estimates for β_{j1}'s. Theorem 2 and (5) suggests using

$$\bar{\beta}_{j1} = \frac{\sum_{i=1}^N \mathbb{I}(Y_i = 1)\frac{\partial \hat{f}_X(X_i)}{\partial X_{j1}}}{\sum_{i=1}^N \mathbb{I}(Y_i = j)\frac{\partial \hat{f}_X(X_i)}{\partial X_{11}}} = \frac{h_1}{h_j} \times \frac{\sum_{i=1}^N \sum_{n=1;n\neq i}^N \mathbb{I}(Y_i = 1)\frac{\partial K_j(\frac{X_{ij}-X_{nj}}{h_j})}{\partial \tilde{u}_{j1}} \prod_{\ell\neq j} K_\ell(\frac{X_{i\ell}-X_{n\ell}}{h_\ell})}{\sum_{i=1}^N \sum_{n=1;n\neq i}^N \mathbb{I}(Y_i = j)\frac{\partial K_1(\frac{X_{i1}-X_{n1}}{h_1})}{\partial u_{11}} \prod_{\ell\neq 1} K_\ell(\frac{X_{i1}-X_{n1}}{h_1})}. \quad (9)$$

However, given the linear index structure and estimates of $\tilde{\gamma}_j$ (and thus the indices $\widetilde{W}_{ij} = X_{ij}'\tilde{\gamma}_j$ as estimates of $\bar{W}_{ij} = X_{ij}'\gamma_j$ for all observations i), we can reduce the dimensionality of this nonparametric problem, instead following (6) and using

$$\widetilde{\beta}_{j1} = \frac{\sum_{i=1}^N \mathbb{I}(Y_i = 1)\frac{\partial \hat{f}_W(\widetilde{W}_i)}{\partial W_{ij}}}{\sum_{i=1}^N \mathbb{I}(Y_i = j)\frac{\partial \hat{f}_W(\widetilde{W}_i)}{\partial W_{i1}}} = \frac{\bar{h}_1}{\bar{h}_j} \times \frac{\sum_{i=1}^N \sum_{n=1;n\neq i}^N \mathbb{I}(Y_i = 1)\frac{d\bar{K}_j(\frac{\widetilde{W}_{ij}-\widetilde{W}_{nj}}{\bar{h}_j})}{du_j} \prod_{\ell\neq j} \bar{K}_\ell(\frac{\widetilde{W}_{i\ell}-\widetilde{W}_{n\ell}}{\bar{h}_\ell})}{\sum_{i=1}^N \sum_{n=1;n\neq i}^N \mathbb{I}(Y_i = j)\frac{d\bar{K}_1(\frac{\widetilde{W}_{i1}-\widetilde{W}_{n1}}{\bar{h}_1})}{du_1} \prod_{\ell\neq 1} \bar{K}_\ell(\frac{\widetilde{W}_{i\ell}-\widetilde{W}_{n\ell}}{\bar{h}_\ell})} \quad (10)$$

[8] Obviously linearity of the utility indices over the entire support of X is a strong assumption . But note that identification and estimation is not reliant on these linear indices holding globally. For example, one could estimate two different linear indexes - one where, e.g. $X_{j1} > 0$, and another where $X_{j1} < 0$.

where

$$\hat{f}_{\bar{W}}(\widetilde{W}_i) = \frac{1}{(N-1)h^J} \sum_{n=1;n\neq i}^{N} \prod_{\ell=1}^{J} \bar{K}_\ell(\frac{\widetilde{W}_{i\ell} - \widetilde{W}_{n\ell}}{\bar{h}_\ell}),$$

in which $\bar{K}_\ell : \mathbb{R} \to \mathbb{R}$ are Parzen-Rosenblatt kernels with compact support and $\bar{h}_\ell$ are bandwidths. Again, under similar conditions in Powell, Stock and Stocker (1989), both $\bar{\beta}_{j1}$ and $\widetilde{\beta}_{j1}$ are $\sqrt{N}$ consistent and have a limiting normal distribution. However, $\widetilde{\beta}_{j1}$ only requires a J dimensional kernel rather than a $\sum_{j=1}^{J} d_j$ dimensional kernel.

3.2.3 Error Distribution Parameters ψ_j

With estimates of the γ_{js}'s and β_{j1}'s in hand, we can think about estimating the ψ_j functions, which can be thought of as transformations of the joint distribution of the error terms ϵ_{ij}. Estimation of the ψ_j functions allows us to compute additional counterfactuals, in particular, the effect of changes in X on choice probabilities. Here we propose very simple kernel based closed form estimator of ψ_j based on (7)

$$\widetilde{\psi}_j(w) = \frac{\sum_{i=1}^{N} \mathbb{I}(Y_i = j) \prod_{\ell=1}^{J} \bar{K}_\ell\big(\frac{w_\ell - X'_{i\ell}\tilde{\beta}_j}{\bar{h}_\ell}\big)}{\sum_{i=1}^{N} \prod_{\ell=1}^{J} \bar{K}_\ell\big(\frac{w_\ell - X'_{i\ell}\tilde{\beta}_j}{\bar{h}_\ell}\big)}. \tag{11}$$

$\widetilde{\psi}_j$ can be used to compute probabilities at any X by evaluating $\widetilde{\psi}_j$ at $(X'_1\tilde{\beta}_1, ..., X'_J\tilde{\beta}_J)$.[9] Depending on the precise kernel used, (11) might be analytically differentiated to estimate the derivative of a choice probability at any X. As well known, whether a particular counterfactual estimand is $\sqrt{N}$ consistent will depend on its precise form, e.g. average derivatives across the distribution of X in the data, versus derivatives at a single point.

Like with our closed-form estimates of the γ_{js}'s and β_{j1}'s, there is a tradeoff we have to make in obtaining these closed form estimates - i.e. we are not fully imposing restrictions of choice theory. Later we examine alternative estimators of ψ_j that do utilize some of these economic restrictions. Also note that these estimates $\widetilde{\psi}_j$ do not depend on the estimates of the across-choice scale parameters $\widetilde{\beta}_{\ell 1}$. If we multiplied the indices by the $\widetilde{\beta}_{\ell 1}$'s, it would not change anything since we are treating ψ_j non-parametrically. This illustrates just one aspect of how our closed-form estimates $\widetilde{\psi}_j$ are not fully imposing restrictions of choice theory – there are others, like monotonicity implied by the model and the Slutzky symmetry restriction.[10]

[9] Note that the estimation of ψ_j requires the estimates of β_{j1}. Without the latter, one could alternatively estimate $\Pr(Y = j|\bar{W} = \cdot)$, which is closely related but different from our structural function $\psi_j(\cdot) = \Pr(Y = j|W = \cdot)$.

[10] Also note that this does not mean that our estimates of $\widetilde{\beta}_{j1}$ are uninteresting, as they help answer different sorts of counterfactual questions, e.g. how changes in X's relatively affect utilities of individuals forced to make various choices.

4 Improved Estimates

As we have noted, the cost of our closed-form estimators is that we do not fully enforce restrictions on the model implied by choice theory. To the best of our knowledge, fully enforcing those restrictions would generally require a non-closed form estimator and a numeric search over a relatively large dimensional parameter space, e.g. sieve MLE of the multinomial choice model where the non-parametric joint distribution $(\epsilon_{i1}, ..., \epsilon_{iJ})$ is flexibly specified, e.g. a flexible mixture of normals.

Because we are not fully enforcing these restrictions, we want to do the best we can in terms of precision. So, continuing to restrict attention to closed form estimators, we next pursue three distinct directions for providing improved estimators of the γ_{js}'s, β_{j1}'s, and ψ_j's to increase efficiency. First, we improve our estimates of γ_{js}'s using the well known result on taking a single Newton-Step from an initial consistent estimator (Lehmann (1983), Horowitz (1998)). Second, we show how overidentification of the β_{j1}'s can be leveraged in an optimal way. Lastly, we show how we can improve our estimates of the ψ_j's using local linear kernel approach (e.g. Stone 1977, Cleveland, 1979, Fan, 1992, 1993). We show how these kernels can very easily incorporate the Slutsky symmetry restriction implied by choice theory, i.e. (3). Again, these improved estimators continue to be in closed form.

4.1 A Newton-Step on the $\tilde{\gamma}_j$'s

We first consider doing a Newton-Step on our initial consistent estimates of the $\tilde{\gamma}_j$'s. Following Klein and Spady (1993) and Lee (1995), consider the following pseudo-likelihood MLE:

$$\hat{\gamma}^* = \arg\max_{\gamma \in \Gamma} \ \frac{1}{N} \sum_{i=1}^{N} \sum_{j=0}^{J} \mathbb{I}(Y_i = j) \log \hat{\mathbb{P}}(Y_i = j | X_i; \gamma),$$

where $\Gamma \equiv \left\{(\gamma_1, \cdots, \gamma_J) : \gamma_j \in \mathbb{R}^{d_j} \text{ with } \gamma_{j1} = 1 \text{ for } j = 1, \cdots, J\right\}$ and

$$\hat{\mathbb{P}}(Y_i = j | X_i; \gamma) = \frac{\sum_{n=1; n \neq i}^{N} \mathbb{I}(Y_n = j) \prod_{\ell=1}^{J} \bar{K}_\ell(\frac{X'_{\ell i}\gamma_\ell - X'_{\ell n}\gamma_\ell}{\bar{h}_\ell})}{\sum_{n=1; n \neq i}^{N} \prod_{\ell=1}^{J} \bar{K}_\ell(\frac{X'_{\ell i}\gamma_\ell - X'_{\ell n}\gamma_\ell}{\bar{h}_\ell})},$$

is a nonparametric kernel estimator of $\mathbb{P}(Y_i = j | X'_{1i}\gamma_1, \cdots, X'_{Ji}\gamma_J)$.

One could estimate γ by maximizing this pseudo likelihood function. In fact, Klein and Spady (1993) and Lee (1995) show that this produces estimates of the index parameters γ that achieve the semiparametric efficiency bound. However, this is not a closed-form estimator. So what we do instead is perform a Newton-Step using the pseudo likelihood starting from our consistent

estimator $\tilde{\gamma}$, i.e.

$$\hat{\gamma} = \tilde{\gamma} - \left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \tilde{s}(Y_i, X_i; \tilde{\gamma})}{\partial \gamma}\right]^{-1} \times \frac{1}{N}\sum_{i=1}^{N}\tilde{s}(Y_i, X_i; \tilde{\gamma}),$$

where $\tilde{s}(Y_i, X_i; \gamma) = \sum_{k\in\mathcal{J}} \frac{\mathbb{I}(Y_i=k)}{\hat{\mathbb{P}}(Y_i=k|X_i;\gamma)} \frac{\partial \hat{\mathbb{P}}(Y_i=k|X_i;\gamma)}{\partial \gamma}$ is the (pseudo) score function. Note that in the above equation, $-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \tilde{s}(Y_i,X_i;\tilde{\gamma})}{\partial \gamma}$ can be replaced by the (pseudo) Fisher information matrix

$$\frac{1}{N}\sum_{i=1}^{N}\tilde{s}(Y_i, X_i; \tilde{\gamma}) \times \tilde{s}'(Y_i, X_i; \tilde{\gamma}).$$

For each $r_j \in \mathbb{R}^{d_j}$ and $r = (r_1', \cdots, r_J')'$, let $\bar{W}_j(r_j) = X_j' r_j$ and $\bar{W}(r) = \big(\bar{W}_1(r_1), \cdots, \bar{W}_J(r_J)\big)$. Following e.g. Lehmann (1983) and Horowitz (1998), we can show that the proposed one-step-updating estimator $\hat{\gamma}$ is as efficient as the pseudo MLE, i.e. $\hat{\gamma}^*$.

Lemma 3 *(i) Let $\{(Y_i, X_i) : i \leq N\}$ be an i.i.d. random sample. (ii) Let Γ be a compact space and γ is in the interior of Γ. (iii) There exists a $\underline{p}$ such that $0 < \underline{p} < \mathbb{P}(Y = k|\bar{W}(\gamma))$, holding for all $k \in \mathcal{J}$ and a neighborhood of γ; (iv) Let X be continuously distributed with continuously differentiable density function f_X; Moreover, let $\bar{W}(\gamma)$ be continuously distributed with continuously differentiable density function $f_{\bar{W}(\gamma)}$ for any $\gamma \in \Gamma$; (v) Let $\mathbb{P}[Y = k|\bar{W}(\gamma) = \bar{w}]$ be continuously differentiable in $\bar{w}$, with $\left|\frac{\partial \mathbb{P}[Y=k|\bar{W}(\gamma)=\bar{w}]}{\partial \bar{w}}\right| \leq c$ for some constant $c > 0$; Moreover, let $\mathbb{P}[Y = k|\bar{W}(\gamma) = \bar{w}]$ be continuously differentiable in γ; (vi) Suppose $\tilde{\gamma}$ is $\sqrt{N}$-consistent estimator of γ and*

$$\sup_{\tilde{\gamma}\in B_\epsilon(\gamma)} \sup_{i\in\{1,\cdots,N\}} \left|\hat{\mathbb{P}}(Y_i = k|W_i(\tilde{\gamma})) - \mathbb{P}(Y_i = k|W_i(\tilde{\gamma}))\right| = o_p(N^{-1/4}).$$

Then we have $\hat{\gamma} - \hat{\gamma}^* = o_p(N^{-1/2})$ and

$$\sqrt{N}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, I^{-1}(\gamma))$$

where $I(\gamma) = \mathbb{E}[s(Y, X; \gamma)s'(Y, X; \gamma)]$ and $s(Y, X; \gamma) = \sum_{k\in\mathcal{J}} \frac{\mathbb{I}(Y=k)}{\mathbb{P}(Y=k|W(\gamma))} \frac{\partial \mathbb{P}(Y=k|W(\gamma))}{\partial \gamma}$.

In Lemma 3, Condition (vi) is a high-level assumption that requires the choice of kernel function K_w and bandwidth h_w to ensure $\hat{\mathbb{P}}(Y_i = k|X_i; \tilde{\gamma})$ uniformly converges to $\mathbb{P}(Y_i = k|W_i(\gamma))$ at a rate sufficiently fast.[11] Note that $I(\gamma)$ can be estimated by $\hat{I}(\gamma) = \frac{1}{N}\sum_{i=1}^{N}\tilde{s}(Y_i, X_i; \hat{\gamma})\tilde{s}'(Y_i, X_i; \hat{\gamma})$.

Combining this result and the results in Klein and Spady (1993) and Lee (1995), our closed form estimate $\hat{\gamma}$ achieves the semiparametric efficiency bound. Note that we cannot do a Newton-Step on the $\widetilde{\beta}_{j1}$'s as those are not identified with the Klein-Spady pseudo-likelihood (for the same reason that $\widetilde{\beta}_{j1}$ is irrelevant for the estimates of the $\widetilde{\psi}_j$'s). However, after doing the Newton Step to

[11] By a similar argument to (**?**), we may need to modify $\hat{\mathbb{P}}(Y_i = k|X_i; \tilde{\gamma})$ to $\underline{p}$ if it's too small. By Assumption (iii), the probability of such a modification, however, should approach to zero as the sample size increases.

obtain new estimates $\widehat{\gamma}_j$'s, one would likely want to reestimate β_{j1}'s based on newly constructed indices $\widehat{W}_{ij} = X_{ij}\widehat{\gamma}_j$. Similarly, one would likely want to construct new estimates of ψ_j's based on those new $\widehat{W}_{ij}$ indices.

There are some practical issues involved in taking a Newton Step from an initial consistent estimate. For example, in some cases, the Newton-Step might go to a parameter value with a worse psuedo-likelihood. If this were the case, choosing not to take the Newton-Step would be an option that preserves the asymptotic efficiency result[12], but perhaps improves small sample performance. More problematic is a case where at the initial consistent estimate, the pseudo-likelihood function is not concave. If this is the case, one should certainly not take a Newton-Step, as this would tend to minimize the pseudo-likelihood instead of maximizing it. One option would be to try to find a nearby point where the pseudo-likelihood is at least as high as at the initial point but concave, and take a Newton-Step from there. This would also preserve efficiency, but is not closed form (though it is presumably less computationally taxing than try to find the maximum).

4.2 Leveraging Overidentification of β_{j1}'s

Unlike the γ_j's, we cannot take pseudo-likelihood Newton Steps on the β_{j1}'s to obtain (semiparametrically) efficient estimators. This is because the multiple-index pseudo-likelihood function does not enforce the model restriction ((6)), and it is this restriction that is used for the identification of these "across-choice" scale parameters β_{j1}'s. Hence, we feel it is important to leverage usable overidentification information on the β_{j1}'s. As noted earlier, when there are more than 3 choices ($J > 2$), ((6)) implies that the β_{j1}'s are overidentified. More specifically, there are $(J-1)$ parameters of β_{j1} in $J(J-1)/2$ equations (recall that $\beta_{11} = 1$ is the normalization).

To optimally use this information, start by representing these equations in (6) as two set of conditions: First, let $k = 1$ and $j = 2, \cdots, J$, then we have

$$\mathbb{E}(\mathbb{A}) \times \begin{bmatrix} \beta_{21} \\ \vdots \\ \beta_{J1} \end{bmatrix} = \mathbb{E} \begin{bmatrix} \mathbb{I}(Y=1)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_2} \\ \vdots \\ \mathbb{I}(Y=1)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_J} \end{bmatrix}$$

where $\mathbb{A} = \text{diag}\left(\mathbb{I}(Y=2)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_1}, \cdots, \mathbb{I}(Y=J)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_1}\right)$, a $(J-1)\times(J-1)$ matrix. On the RHS, note that $\beta_{11} = 1$ as a normalization. Thus, β_{j1}'s are identified by the above equation.

[12] Since one is choosing as one's estimate that has a higher pseudo-likelihood than the parameter vector from the Newton-Step.

Second, choose $2 \leq k < j \leq J$, then we have

$$\mathbb{E}(\mathbb{B}) \times \begin{bmatrix} \beta_{21} \\ \vdots \\ \beta_{J1} \end{bmatrix} = \mathbf{0}_{\frac{(J-1)(J-2)}{2}}$$

where $\mathbb{B} = [B_2', \cdots, B_{J-1}']'$ is a $\frac{(J-1)(J-2)}{2} \times (J-1)$ matrix, in which B_j is a $(J-j) \times (J-1)$ matrix defined as

$$B_j = \begin{pmatrix} & -\mathbb{I}(Y=j+1)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_j} & \mathbb{I}(Y=j)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_{j+1}} & 0 & \cdots & 0 \\ & -\mathbb{I}(Y=j+2)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_j} & 0 & \mathbb{I}(Y=j)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_{j+2}} & \cdots & 0 \\ \mathbf{0}_{(J-j)\times(j-2)} & \cdots & \cdots & \cdots & \cdots & \cdots \\ & -\mathbb{I}(Y=J)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_j} & 0 & 0 & \cdots & \mathbb{I}(Y=j)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_J} \end{pmatrix}.$$

It should be noted that $\mathbb{B}$ is degenerated to be void when $J = 2$. Moreover, it should also be noted that $\mathbb{E}(B_j)$ and $\mathbb{E}(\mathbb{B})$ has a rank of $J - j$ and $J - 2$, respectively.

Combining these two set of conditions together, we obtain

$$\mathbb{E}\begin{bmatrix} \mathbb{A} \\ \mathbb{B} \end{bmatrix} \times \begin{bmatrix} \beta_{21} \\ \vdots \\ \beta_{J1} \end{bmatrix} = \mathbb{E}\begin{bmatrix} \mathbb{I}(Y=1)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_2} \\ \vdots \\ \mathbb{I}(Y=1)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_J} \\ \mathbf{0}_{\frac{(J-2)(J-1)}{2}} \end{bmatrix}.$$

Therefore, for any $\frac{(J-1)J}{2} \times \frac{(J-1)J}{2}$ positive definite weight matrix Ω, we have

$$\begin{bmatrix} \beta_{21} \\ \vdots \\ \beta_{J1} \end{bmatrix} = \left\{ \mathbb{E}\begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix}' \Omega\, \mathbb{E}\begin{bmatrix} \mathbb{A} \\ \mathbb{B} \end{bmatrix} \right\}^{-1} \times \mathbb{E}\begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix}' \Omega\, \mathbb{E}\begin{bmatrix} \mathbb{I}(Y=1)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_2} \\ \cdots \\ \mathbb{I}(Y=1)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_J} \\ \mathbf{0}_{\frac{(J-2)(J-1)}{2}} \end{bmatrix}.$$

Following the GMM literature, we can choose the following optimal weighting matrix:

$$\Omega^* = \mathbb{E}(\xi\xi'),$$

where $\xi = \left\{ \begin{bmatrix} \mathbb{A} \\ \mathbb{B} \end{bmatrix} \times \begin{bmatrix} \beta_{21} \\ \vdots \\ \beta_{J1} \end{bmatrix} - \begin{bmatrix} \mathbb{I}(Y=1)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_2} \\ \cdots \\ \mathbb{I}(Y=1)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_J} \\ \mathbf{0}_{\frac{(J-2)(J-1)}{2}} \end{bmatrix} \right\}$. Because ξ depends on unknown parameter $\{\beta_{j1} : j \geq 2\}$, therefore we apply the standard two-step approach in which we first choose $\Omega = I$ to estimate $\{\beta_{j1} : j \geq 2\}$ for estimating the optimal weighting matrix Ω^* by $\hat{\Omega}$. In the second

stage we plug $\hat{\Omega}^*$ into the above formula to obtain an estimator of $(\beta_{21}, \cdots, \beta_{J1})'$, i.e.

$$\begin{bmatrix} \widehat{\beta}_{21} \\ \vdots \\ \widehat{\beta}_{J1} \end{bmatrix} = \left\{ \hat{\mathbb{E}} \begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix}' \hat{\Omega}\, \hat{\mathbb{E}} \begin{bmatrix} \mathbb{A} \\ \mathbb{B} \end{bmatrix} \right\}^{-1} \times \hat{\mathbb{E}} \begin{bmatrix} \mathbb{A}' \\ \mathbb{B}' \end{bmatrix}' \hat{\Omega}\, \hat{\mathbb{E}} \begin{bmatrix} \mathbb{I}(Y=1)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_2} \\ \cdots \\ \mathbb{I}(Y=1)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_J} \\ \mathbf{0}_{\frac{(J-2)(J-1)}{2}} \end{bmatrix}$$

in which $\mathbb{E}[\mathbb{I}(Y=k)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_j}]$ is estimated again by (**?**)-type estimator:

$$\hat{\mathbb{E}}\left[\mathbb{I}(Y=k)\frac{\partial f_{\bar{W}}(\bar{W})}{\partial \bar{W}_j}\right] = \frac{1}{(N-1)\bar{h}_j^{J+1}} \sum_{i=1}^{N} \sum_{n=1;n\neq i}^{N} \left[\mathbb{I}(Y_i=k) \prod_{\ell \neq j} \bar{K}_\ell(\frac{\widehat{W}_{i\ell} - \widehat{W}_{n\ell}}{\bar{h}_\ell})\right] \frac{d\bar{K}_j(\frac{\widehat{W}_{ij} - \widehat{W}_{nj}}{\bar{h}_j})}{du_j}.$$

where $\widehat{W}_{ij} = X_{ij}'\hat{\gamma}_j$ are improved estimates of $\bar{W}_{ij} = X_{ij}'\gamma_j$ for all observations i. Note that $\bar{K}$ and $\bar{h}$ have been introduced above.

4.3 Local Linear Kernels

Lastly, we show how we can potentially improve our estimates of the ψ_j's using a local linear kernel approach (e.g. Stone 1977, Cleveland, 1979, Fan, 1992, 1993). In particular, we show how these kernels can very easily incorporate the Slutsky symmetry restriction implied by choice theory – in other words we estimate $\psi_j(W) = \mathbb{P}(Y=j|W)$ and its derivatives while imposing the model restrictions on the cross derivatives of the choice probabilities, i.e.

$$\frac{\partial \mathbb{P}(Y=j|W)}{\partial W_k} = \frac{\partial \mathbb{P}(Y=k|W)}{\partial W_j}.$$

The LLK method fits a linear regression line through the observations in a local neighborhood of W, obtaining a kernel smoothed estimator of the regression function and its (partial) derivatives in the same time. Such an approach allows us to obtain a closed-form local estimator of $\psi_j(W)$ under our model restrictions on the choice probabilities' cross derivatives.

Let $\hat{\beta}_j = \hat{\beta}_{j1} \times \hat{\gamma}_j$. For $w \equiv (w_1, \cdots, w_J) \in Supp^\circ(W)$, let θ_{j0} and θ_{jk} $(k \in \mathcal{J}_1)$ be $\mathbb{P}(Y=j|W=w)$ and its partial derivative w.r.t. w_k, respectively. In other words, at a given point w, the parameter θ_{j0} is equal to ψ_j at that point, and the parameters θ_{jk} are equal to the derivatives of ψ_j w.r.t. the indexes at that point.

Denote $\theta_j = (\theta_{j0}, \theta_{j1}, \cdots, \theta_{jJ})' \in \mathbb{R}^{J+1}$ and $\theta = (\theta_1', \cdots, \theta_J')' \in \mathbb{R}^{J(J+1)}$. Then, at a given point w,

our estimator $\hat{\theta}$ is defined as follows:

$$\hat{\theta}(w) = \underset{\theta \in \mathbb{R}^{J(J+1)}}{\arg\min} \frac{1}{n\prod_{\ell=1}^{J} h_\ell} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} \Big[\mathbb{I}(Y_i = j) - \theta_{j0} - \sum_{k=1}^{J} (X_{ik}' \hat{\beta}_k - w_k)\theta_{jk} \Big]^2 \right\} \prod_{\ell=1}^{J} K_\ell\big(w_\ell - \frac{X_{i\ell}' \hat{\beta}_\ell}{h_\ell}\big)$$
$$s.t. \;\; \theta_{jk} - \theta_{kj} = 0, \;\; \text{for all} \;\; k = \mathcal{J}/\{J\}, \; j \geq k+1.$$

We now use matrix algebra to rewrite this objective function and obtain its first order condition. Let $\mathbb{W}_n(w) = \frac{1}{n}\text{diag}\Big[K\big(\frac{w_1 - X_{11}'\hat{\beta}_1}{h_1}, \cdots, \frac{w_J - X_{1J}'\hat{\beta}_J}{h_J}\big), \cdots, K\big(\frac{w_1 - X_{n1}'\hat{\beta}_1}{h_1}, \cdots, \frac{w_J - X_{nJ}'\hat{\beta}_J}{h_J}\big)\Big]$ and

$$\mathbb{Y}_{jn} = (\mathbb{I}(Y_1 = j), \cdots, \mathbb{I}(Y_n = j))';$$
$$\mathbb{X}_n(w) = \begin{bmatrix} 1 & X_{11}'\hat{\beta}_1 - w_1 & \cdots & X_{1J}'\hat{\beta}_J - w_J \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1}'\beta_1 - w_1 & \cdots & X_{nJ}'\beta_J - w_J \end{bmatrix}.$$

Let Ψ_j be a $(J+1) \times \frac{(J-1)J}{2}$ matrix, with k–th row defined as

$$\Psi_j(k,\ell) = \begin{cases} 0 & \text{for } k = 1, j; \\ -\mathbb{I}(\ell = \frac{(k-1)k}{2} + j - k) & \text{for } 1 < k < j; \\ \mathbb{I}(\ell = \frac{(j-1)j}{2} + k - j) & \text{for } k > j, \end{cases}$$

and $\Psi = (\Psi_1', \cdots, \Psi_J')'$. Let further Δ be a $\frac{(J-1)J}{2} \times J(J+1)$ matrix, with $[\frac{(2J-k)(k-1)}{2} + \ell]$–th row, $k = 1, \cdots, J-1$ and $\ell = 1, \cdots, J-k$, defined as

$$\Delta\Big(\frac{(2J-k)(k-1)}{2} + \ell, q\Big) = \begin{cases} 1 & \text{if } q = (k-1)(J+1) + k + \ell + 1; \\ -1 & \text{if } q = (k+\ell-1)(J+1) + k + 1; \\ 0 & \text{otherwise.} \end{cases}$$

By definition, both Ψ and Δ are known matrices. Thus, we obtain the Lagrange function of the above minimization problem as follows:

$$L_n(\theta; w) = \frac{1}{2}\sum_{j=1}^{J} [\mathbb{Y}_{jn} - \mathbb{X}_n(w)\theta_j]' \mathbb{W}_n(w) [\mathbb{Y}_{jn} - \mathbb{X}_n(w)\theta_j] + \lambda' \Delta\theta.$$

where $\lambda = (\lambda_{12}, \cdots, \lambda_{1J}, \lambda_{23}, \cdots, \lambda_{2J}, \cdots\cdots, \lambda_{J-1,J}) \in \mathbb{R}^{\frac{(J-1)J}{2}}$ is the Lagrange multiplier. Then, we obtain the f.o.c.:

$$\mathbb{X}_n(w)'\mathbb{W}_n(w)\mathbb{X}_n(w) \times \hat{\theta}_j + \Psi_j \hat{\lambda} = \mathbb{X}_n(w)'\mathbb{W}_n(w)\mathbb{Y}_{jn}, \;\; \text{for } j = 1, \cdots, J;$$
$$\Delta\hat{\theta} = 0.$$

Therefore, we express $\hat{\theta}_j(w)$ in terms of $\hat{\lambda}(w)$

$$\hat{\theta}_j(w) = \tilde{\theta}_j(w) - \left[\mathbb{X}_n(w)'\mathbb{W}_n(w)\mathbb{X}_n(w)\right]^{-1}\Psi_j\hat{\lambda}(w),$$

where $\tilde{\theta}_j(w) = \left[\mathbb{X}_n(w)'\mathbb{W}_n(w)\mathbb{X}_n(w)\right]^{-1}\mathbb{X}_n(w)'\mathbb{W}_n(w)\mathbb{Y}_{jn}$. Note that $\tilde{\theta}_j(w)$ is the usual local linear kernel estimator (i.e. without imposing the model restrictions on the cross derivatives). It follows that

$$\hat{\theta}(w) = \tilde{\theta}(w) - \text{Diag}\left\{\left[\mathbb{X}_n(w)'\mathbb{W}_n(w)\mathbb{X}_n(w)\right]^{-1},\cdots,\left[\mathbb{X}_n(w)'\mathbb{W}_n(w)\mathbb{X}_n(w)\right]^{-1}\right\}\Psi\hat{\lambda}(w).$$

Moreover, we plug $\hat{\theta}(w)$ into the f.o.c.:

$$\hat{\lambda}(w) = \left\{\Delta \times \text{Diag}\left[\left(\mathbb{X}_n(w)'\mathbb{W}_n(w)\mathbb{X}_n(w)\right)^{-1},\cdots,\left(\mathbb{X}_n(w)'\mathbb{W}_n(w)\mathbb{X}_n(w)\right)^{-1}\right]\times\Psi\right\}^{-1}\Delta\times\tilde{\theta}(w).$$

Therefore, we obtain $\hat{\theta}(w)$ of closed form:

$$\begin{aligned}\hat{\theta}(w) = \tilde{\theta}(w) &- \text{Diag}\left[\left(\mathbb{X}_n(w)'\mathbb{W}_n(w)\mathbb{X}_n(w)\right)^{-1},\cdots,\left(\mathbb{X}_n(w)'\mathbb{W}_n(w)\mathbb{X}_n(w)\right)^{-1}\right]\Psi\\ &\times\left\{\Delta\times\text{Diag}\left[\left(\mathbb{X}_n(w)'\mathbb{W}_n(w)\mathbb{X}_n(w)\right)^{-1},\cdots,\left(\mathbb{X}_n'(w)\mathbb{W}_n(w)\mathbb{X}_n(w)\right)^{-1}\right]\Psi\right\}^{-1}\Delta\times\tilde{\theta}(w).\end{aligned}$$

Fan (1992) established the asymptotic properties of $\tilde{\theta}(w)$, from which we obtain $\hat{\theta}(w)$'s limiting distribution.

In sum, note that enforcing the cross derivative restrictions on the ψ_j's is quite simple to do in closed form. Of course, this is only a subset of the economic restriction that the choice model places on the ψ_j's. For example, the model also implies that the ψ_j is weakly increasing in W_j, and weakly decreasing in the W_k's. However, it is not clear how one can impose these monotonicity restrictions in closed form estimation.

5 Extensions

5.1 Covariates that are Constant Across Alternatives

The above procedure does not work when an X is constant across alternatives, i.e. $X_{1s} = ... = X_{Js} = Z_s$ for $s = 1,\cdots,d_z$. In this case, the derivatives w.r.t. Z_s are different since they affect the utility index for all (or multiple) choices. However, it is straightforward to do adapt our closed for estimators for this case. Again, the identification result would follow from the general proof of Allen and Rehbeck (2019), but the estimator and $\sqrt{N}$ consistency result are new. Intutively, what is going on here is that given the existence of one X that varies across alternatives, one can use its variation to net out the effect of Z on all but one index.

For notational simplicity, let $Z = (X_{j1}, \cdots, X_{jd_z})'$, $\bar{X}_j = (X_{j,d_z+1}, \cdots, X_{jd_j})'$, and $X_j = (Z', \bar{X}_j')'$ for $j \in \mathcal{J}$. Further, let $\beta_j^z = (\beta_{j1}, \cdots, \beta_{jd_z})'$ and $\beta_j^\circ = (\beta_{j,d_z+1}, \cdots, \beta_{jd_j})'$. Again, let $\beta_{js} = \beta_{j,d_z+1}\gamma_{js}$ for $j = 1, \cdots, d_j$. Denote $\gamma_j^z = (\gamma_{j,1}^\circ, \cdots, \gamma_{jd_z}^\circ)'$ and $\gamma_j^\circ = (\gamma_{j,d_z+1}^\circ, \cdots, \gamma_{jd_j}^\circ)'$. By definition, $\gamma_{j,d_z+1} = 1$. Again, we normalize $\beta_{1,d_z+1} = 1$.

By a similar argument to theorem1, we have

$$\gamma_j^\circ = \frac{1}{2}\lambda_{kj} \times \mathbb{E}\left[\frac{\partial \mathbb{P}(Y=k|X)}{\partial \bar{X}_j} f(X)\right] = \lambda_{kj} \times \mathbb{E}\left[\mathbb{I}(Y=k)\frac{\partial f(X)}{\partial \bar{X}_j}\right]$$

which identifies the γ_j°, i.e. the within-choice relative scales for the X's that do vary across j. Next, let $\bar{W}_j = \bar{X}_j'\gamma_j^\circ$ and $\bar{W} = (\bar{W}_1', \cdots, \bar{W}_J')'$. Then we identify β_{j,d_z+1} as follows

$$\frac{\beta_{j,d_z+1}}{\beta_{k,d_z+1}} = \frac{\mathbb{E}\left[\mathbb{I}(Y=k)\frac{\partial \bar{f}(\bar{W},Z)}{\partial \bar{W}_j}\right]}{\mathbb{E}\left[\mathbb{I}(Y=j)\frac{\partial \bar{f}(\bar{W},Z)}{\partial \bar{W}_k}\right]}$$

where $\bar{f}$ is the density of $(\bar{W}, Z)$.

Now consider the identification and estimation of γ_j^z, i.e. the within-choice relative scales for the X's that are constant across j. Note that

$$\frac{\partial \mathbb{P}(Y=k|X)}{\partial Z} = \sum_{j=1}^{J} \frac{\partial \mathbb{P}(Y=k|\bar{W},Z)}{\partial \bar{W}_j} \times \gamma_j^z, \quad k \in \mathcal{J}.$$

It follows that

$$\mathbb{E}\left[\mathbb{I}(Y=k)\frac{\partial \bar{f}(\bar{W},Z)}{\partial Z}\right] = \sum_{j=1}^{J} \mathbb{E}\left[\mathbb{I}(Y=k)\frac{\bar{f}(\bar{W},Z)}{\partial \bar{W}_j}\right] \times \gamma_j^z, \quad k \in \mathcal{J}.$$

Let $\mathbb{C}_{kj} = \mathbb{E}\left[\mathbb{I}(Y=k)\frac{\partial \bar{f}(\bar{W},Z)}{\partial \bar{W}_j}\right]$ and $\mathbb{C}$ be a $J \times J$ matrix, defined as

$$\mathbb{C} = \begin{bmatrix} \mathbb{C}_{11} & \cdots & \mathbb{C}_{1J} \\ \cdots & \cdots & \cdots \\ \mathbb{C}_{J1} & \cdots & \mathbb{C}_{JJ} \end{bmatrix}.$$

Thus, the above equation can be rewritten as

$$\mathbb{C} \times \begin{pmatrix} \gamma_1^{z\prime} \\ \vdots \\ \gamma_J^{z\prime} \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left[\mathbb{I}(Y=1)\frac{\partial \bar{f}(\bar{W},Z)}{\partial Z'}\right] \\ \vdots \\ \mathbb{E}\left[\mathbb{I}(Y=J)\frac{\partial \bar{f}(\bar{W},Z)}{\partial Z'}\right] \end{pmatrix}$$

So $\gamma^z = (\gamma_1^z, \cdots, \gamma_J^z)$ is identified by

$$\gamma^{z\prime} = \mathbb{C}^{-1} \times \begin{pmatrix} \mathbb{E}\left[\mathbb{I}(Y=1)\frac{\partial \bar{f}(\bar{W},Z)}{\partial Z'}\right] \\ \vdots \\ \mathbb{E}\left[\mathbb{I}(Y=J)\frac{\partial \bar{f}(\bar{W},Z)}{\partial Z'}\right] \end{pmatrix}, \tag{12}$$

provided we have a full rank condition on $\mathbb{C}$.

Applying Powell, Stock and Stocker (1989) just as we did earlier, we can obtain a $\sqrt{N}$-consistent estimator of both $\mathbb{C}$ and $\mathbb{E}\left[\mathbb{I}(Y=k)\frac{\partial \bar{f}(\bar{W},Z)}{\partial Z}\right]$. Since (12) shows that the the parameter of interest γ^z is a function of these two quantities, plugging these estimates provides us a $\sqrt{N}$–consistent estimator. Note that we can use a Newton-Step on this estimator to obtain a more efficient estimator of γ^z (and γ^0), just as is illustrated in Section 4.1.

5.2 Discrete Covariates

Since our method requires estimating derivatives of choice probabilities w.r.t. X, it should not be applied when X is discrete. However, we can combine our estimator with the estimator proposed by AIPR to deal with this case. We do need at least one continuous X for each alternative, as this is what allows us to identify the cross-choice relative scales (the β_{j1}'s) that AIPR do not identify.

Before describing the combination of our estimator and AIPR, we first describe how our basic estimator differs from AIPR. Both estimators, ours based on Powell, Stock and Stocker (1989 - PSS), and AIPR, can be used to estimate γ in a model with continuous X (though as noted, ours has problems with discrete X, AIPR cannot identify β_{j1}). How do they differ? Intuitively, both estimators start by non-parametrically estimating choice probabilities given X. Our PSS based estimator is based directly on derivatives of those estimated choice probabilities. In contrast, AIPR do an additional step where they in essence find level sets of those choice probabilities - i.e. sets of X with the same choice probability (or probabilities). Intutively, e.g. in a binary case with two X_{11} and X_{12}, one can think about regressing X_{11} on X_{12} conditional on being in a given probability level set. The coefficient tells us, given a change in X_{12}, how much X_{11} has to change by to "hold that probability constant". This therefore is an estimate of γ_{12}. While they both can be used to estimate γ in a model with continuous X, AIPR works with discrete X (while our PSS derivative based approach does not), while our approach identifies β_{j1} (while AIPR cannot). This is why we propose a combination of the two.

More formally describing this, let $X_j = (X_{jc}, X_{jd})$ where $X_{jc} \in \mathbb{R}^{c_j}$ and $X_{jd} \in \mathbb{R}^{d_j}$ is a vector of continuous and discrete random variables respectively. Our first step is to apply our PSS based estimator to estimate the γ's on the continuous X_{jc}'s. After this, we can combine those X_{jc}'s into an index for each j. Redefine this index as $X_{j1} = X_{jc}\gamma_{j,1:c}$. The first component of each X_j, i.e. X_{j1}, is now a continuously distributed random variable, with the rest of the components being

discrete.

Denote $p_k(X) \equiv \mathbb{P}(Y = k|X)$ and $p(X) = (p_1(X), \cdots, p_J(X))$. Because for each $j \in \mathcal{J}$,

$$X_j'\gamma_j = \mathbb{E}[X_j|p(X)]'\gamma_j,$$

it follows that

$$\sum_{\ell=2}^{1+d_j} \{X_{j\ell} - \mathbb{E}[X_{j\ell}|p(X)]\}\gamma_{j\ell} = -\{X_{j1} - \mathbb{E}[X_{j1}|p(X)]\}.$$

Therefore,

$$\begin{pmatrix} \gamma_{j2} \\ \vdots \\ \gamma_{j,1+d_j} \end{pmatrix} = -\left\{ \mathbb{E}\begin{bmatrix} X_{j2} - \mathbb{E}[X_{j2}|p(X)] \\ \vdots \\ X_{j,1+d_j} - \mathbb{E}(X_{j,1+d_j}|p(X)) \end{bmatrix} \begin{bmatrix} X_{j2} - \mathbb{E}[X_{j2}|p(X)] \\ \vdots \\ X_{j,1+d_j} - \mathbb{E}(X_{j,1+d_j}|p(X)) \end{bmatrix}' \right\}^{-1} \times$$
$$\mathbb{E}\begin{bmatrix} X_{j2} - \mathbb{E}[X_{j2}|p(X)] \\ \vdots \\ X_{j,1+d_j} - \mathbb{E}(X_{j,1+d_j}|p(X)) \end{bmatrix} [X_{j1} - \mathbb{E}[X_{j1}|p(X)]].$$

Thus, we can estimate $(\gamma_{j2}, \cdots, \gamma_{j,1+d_j})$ by[13]

$$-\left\{ \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} X_{ij2} - \hat{\mathbb{E}}[X_{ij2}|p(X_i)] \\ \vdots \\ X_{ij,c_j+d_j} - \hat{\mathbb{E}}(X_{ij,c_j+d_j}|p(X_i)) \end{bmatrix} \begin{bmatrix} X_{ij2} - \hat{\mathbb{E}}[X_{ij2}|p(X_i)] \\ \vdots \\ X_{ij,c_j+d_j} - \hat{\mathbb{E}}(X_{ij,c_j+d_j}|p(X)) \end{bmatrix}' \right\}^{-1} \times$$
$$\left\{ \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} X_{ij2} - \hat{\mathbb{E}}[X_{ij2}|p(X_i)] \\ \vdots \\ X_{ij,c_j+d_j} - \hat{\mathbb{E}}(X_{ij,c_j+d_j}|p(X_i)) \end{bmatrix} \left[X_{ij1} - \hat{\mathbb{E}}[X_{ij1}|p(X_i)\right]' \right\}$$

where

$$\hat{\mathbb{E}}(X_{ij\ell}|p(X_i)) = \frac{\sum_{n\neq i} X_{nj\ell} K_p(\frac{\hat{p}(X_i) - \hat{p}(X_n)}{h_p})}{\sum_{n\neq i} K_p(\frac{\hat{p}(X_i) - \hat{p}(X_n)}{h_p})} \tag{13}$$

where $h_p \in \mathbb{R}_+$ is a bandwidth, $K_p : \mathbb{R}^J \to \mathbb{R}$ is a Parzen-Rosenblatt kernel with compact support, and $\hat{p}(X_i)$ is some kernel estimator of $p(X_i)$. Note that one could alternative use the AIPR procedure to obtain the entire γ (including the coefficients on continuous X's), then using our approach to obtain β_{j1}. But one advantage of the hybrid approach to estimating γ is that the our PSS approach only requires one tuning parameter, while the AIPR approach requires two - one for the kernel estimates of the probabilities $\hat{p}$, and then one to compute the expections $\hat{\mathbb{E}}$ conditional on the probability in (13).

[13] AIPR suggest trimming in practice.

6 Monte-Carlo

Lastly, we do a simple Monte-Carlo to examine how our estimators work in practice. We are interested in studying two particular questions. First, we want to examine the extent to which some of the "improvements" in Section 4 improve the precision of our basic estimators. Second, we want to compare the precision of our "best" closed-form estimators to alternative, non-closed form, estimators - estimators that do impose all restrictions from choice theory. This can give us a sense of the tradeoff between our more robust (but generally inefficient) closed form estimators vs more efficient (but computationally challenging) non-closed form estimators.

The setup of our experiments are as follows. We consider a model with three choices, $j = 0, 1, 2$, the first being the outside alternative with "mean" utility normalized to 0. The utility from the two inside goods each depend on two X variables.

$$\begin{aligned} U_{i0} &= \epsilon_{i0} \\ U_{i1} &= \beta_{10} + \beta_{11} X_{i11} + \beta_{12} X_{i12} + \epsilon_{i1} \\ U_{i2} &= \beta_{20} + \beta_{21} X_{i21} + \beta_{22} X_{i22} + \epsilon_{i2} \end{aligned}$$

Normalizing $\beta_{11} = 1$ and reparameterizing the model as in Section 3.1, we use

$$\begin{aligned} U_{i0} &= \epsilon_{i0} \\ U_{i1} &= \beta_{10} + X_{i11} + \gamma_{12} X_{i12} + \epsilon_{i1} \\ U_{i2} &= \beta_{20} + \beta_{21} X_{i21} + \beta_{21}\gamma_{22} X_{i22} + \epsilon_{i2} \end{aligned}$$

so the parameters of interest are γ_{12}, γ_{22}, and β_{21}. Note that β_{10} and β_{20} are not directly identified by our methods (though they implicitly enter the model through the ψ_j's).

We consider three basic data generating processes for our experiments. In all of them, $\gamma_{12} = \gamma_{22} = \beta_{21} = 1$, $\beta_{10} = \beta_{20} = 0$, and $X_{i11}, X_{i12}, X_{i21}, X_{i22}$ ~ iid normals with mean 0 and standard deviation 2. In DGP A, we consider a simple logit model, i.e. ϵ_{i0}, ϵ_{i1}, and ϵ_{i2} are iid Type 1 Extreme Value unobservables. In DGP B, we add a large amount of correlation between ϵ_{i1} and ϵ_{i2}, by adding an additional unobservable ν_i, normally distributed with mean 0 and standard deviation 3, to both. In other words, ϵ_{i1} is an iid Type 1 Extreme value plus ν_i, and ϵ_{i2} is an iid Type 1 Extreme value plus that same ν_i. In DGP C, we change the shape of the distribution of ϵ_{i2}. In particular, we assume ϵ_{i2} comes from a 50/50 mixture of two normals, the first centered around -2 with variance 0.5, and the second centered around 2 with variance 0.5. This is a highly bimodal distribution.

For each specification, we estimate the parameters (and counterfactuals) with 4 different procedures. As a benchmark, we first estimate using a simple logit model and maximum likelihood. This produces consistent estimates under DGP A, but inconsistent estimates under DGPs B and C, since in those specifications, the errors are not iid logit. We then construct 3 estimates of γ_{12},

γ_{22},and β_{21} using our proposed semi-parametric closed-form methods. The first, "AMXbasic", is based on (8) and (9). The second, "AMXindex" instead uses (10), utilizing the index restriction to achieve more precise estimates of β_{21}. The estimates of γ_{12} and γ_{22} do not change from "AMXbasic".[14] Lastly, "AMXNewton" takes a single Newton-Step on γ_{12} and γ_{22} starting from the estimates of these parameters in AMXindex. "AMXNewton" also re-estimates β_{21} using an index restriction with those improved estimates of γ_{12} and γ_{22}.

For each of the four estimators, Table 1 reports estimates of the index parameters γ_{12}, γ_{22},and β_{21}. Because one is also often interested in estimates of the distributions ψ_j, we also report two additional parameters related to these distributions. Specifically, we report estimates of the following average derivatives

$$E\left[\frac{\partial\psi_2(w_i)}{\partial X_{i22}}\right] \approx \frac{1}{N}\sum_i \frac{\partial\widetilde{\psi}_2(w_i)}{\partial X_{i22}} \quad \text{and } E\left[\frac{\partial\psi_1(w_i)}{\partial X_{i22}}\right] \approx \frac{1}{N}\sum_i \frac{\partial\widetilde{\psi}_1(w_i)}{\partial X_{i22}} \tag{14}$$

where the estimates of $\widetilde{\psi}_1$ and $\widetilde{\psi}_2$ are given by (11). Note that the first term is an own-characteristic derivative, and the second is a cross-characteristic derivative. Since these are averages, these estimates should be $\sqrt{N}$ consistent[15]. For each of the DGPs we report means, standard deviations, and root-mean squared error across 5000 monte-carlo replications, for datasets of size $N = 3000$ and $N = 10000$.[16]

For DGP A, the Logit estimates are clearly consistent. Comparing the 3 closed form estimators, imposing the index restriction in estimating β_{21} reduces rootMSE considerably - from 0.1299 to 0.0899 with $N = 3000$ and from 0.0696 to 0.0486 with $N = 10000$. The Newton-Step improves estimates of γ_{12} and γ_{22} - the rootMSEs of these parameters drop from about 0.057 to 0.042 when $N = 3000$, and from 0.031 to 0.022 when $N = 10000$. Even though β_{21} is not directly impacted by the Newton-Step, reestimating it based on the Newton-Step estimates of γ_{12} and γ_{22} does seem to lower its rootMSE by about 5%. Note that the rootMSE estimates of the average derivatives do not seem to be affected by the Newton-Step - these "parameters" depend much more on the estimates of the ψ's (which we work to improve momentarily).

Continuing to examine DGP A, it is particularly interesting to compare the rootMSEs of the AMXNewton estimator to that of the Logit Estimator. This tells us how much precision we lose by 1) our closed form estimators not imposing all the theoretical restrictions of the choice model, and 2) our closed form estimators treating ψ non-parametrically. Interestingly, the loss in rootMSE for γ_{12} and γ_{22} is extremely small - less than 5%. This reflects the fact that the Newton-Step makes a huge improvement over our basic estimates of γ_{12} and γ_{22} . The loss in rootMSE for β_{21} is significantly higher - for AMXNewton it is about 80% more than the logit rootMSE. This is

[14] One could utilize the index restriction for γ_{12} and γ_{22} as well, but given we are about to do a Newton-Step on these parameters, we opted not to.

[15] *Appropriate cite?*

[16] Following Hansen (20?), we just use Silverman like optimal bandwidths for estimating derivatives.

presumably because the estimate of β_{21} is not directly benefiting from the Newton-Step. The loss in rootMSE for the average derivatives is even higher - 2-4 times for the average derivatives. This also makes sense because the logit model is imposing a (correct) parametric assumption on ψ. To assess how much of this last difference is due to the parametric assumption, vs how much is due to imposing the theoretical restrictions of the choice model, it would be constructive to compare to another estimator - a sieve MLE estimator of the discrete choice model where the joint distribution of $(\epsilon_{i0}, \epsilon_{i1}, \epsilon_{i2})$ is modelled as a sieve, e.g. a mixture of normals. This sieve estimator would enforce the theoretical restrictions of the choice model, but not benefit from parametric assumptions on ψ. The problem is that this is highly challenging to estimate correctly (e.g. ensure one does not end up at local minima of the likelihood function) - especially when one needs to do that thousands of replications. Returning to the AMXNewton estimates, note that there are clearly some small sample biases in the estimates of the average derivatives - this is not unusual given the non-parametric treatment of ψ.

Moving to DGP B and DGP C, a first observation is that in these models, the logit estimates are not consistent, as they are imposing incorrect parametric assumptions. In DGP B, this inconsistency shows up mainly in the average cross derivative. The logit model underestimates the cross derivative by about 30% - which makes sense as it is assumes away the positive correlation between ϵ_{i1} and ϵ_{i2}, which tends to increase this cross-derivative (e.g. it makes the two choices more substitutable). Interestingly, the estimates of γ_{12}, γ_{22},and β_{21} in the logit model do not appear biased. This is because in this specification, choices 1 and 2 are completely symmetric (and the X's are also distributed symmetrically). In DGP C, which is asymmetric since only ϵ_{i2} has a bimodal distribution, β_{21} is underestimated by about 20% (γ_{12} and γ_{22} are still consistently estimated, but again, this is specific to the case where the X's are symmetric). Turning to our closed-form estimates, the patterns follow DGP A fairly closely. Doing the Newton-Step and imposing the index restriction increase the precision of the estimates of γ_{12}, γ_{22},and β_{21} substantially, but do not do much to the average derivative estimates.

Next we consider the proposal in Section 4.3, to improve the estimates of ψ (and its derivatives, e.g. (14)) using Local Linear Kernels. As illustrated in that Section, Local Linear Kernels provide a closed form way to enforce on ψ the cross-derivative restrictions implied by the choice model. Tables 2, 3, 4, examine various levels and derivatives of ψ, for DGP A, DGP B, and DGP C respectively. In each table we again consider datasets with both N=3000 and N=10000. For each dataset, we compare estimates using a LLK where we don't enforce the cross-derivative restriction to estimates using a LLK where we do enforce the cross-derivative restriction. This allows us to assess the effect of imposing the restriction, all else equal. Note that we estimate many different derivatives. The first two rows are average derivatives with respect to X_{i22}, corresponding to what was reported in Table 1. However, we also report derivatives at specific points in ($W_{i1} = X_{i11} + \gamma_{12}X_{i12}$,$W_{i2} = \beta_{21}X_{i21} + \beta_{21}\gamma_{22}X_{i22}$) space. For example, rows 3-6 report estimated derivatives of choice probabilities w.r.t. X_{i22}, at the point where the indices are ($W_{i1} = 0$,$W_{i2} = 0$),

as well as the two choice probabilies themselves at that point. The lower columns do this at other points in a grid on (W_{i1},W_{i2}). Note that since these are values of a non-parametric object (or derivatives of a non-parametric object) at a point, these estimands are not estimable at the parametric rate, but they are still things that researchers may be interested in measuring.

Looking at the first two rows of Table 2, one can see that the precision of the estimate of the average own derivative does not improve when enforcing the cross derivative restriction. The standard deviation of the estimate of the average cross derivative does decrease, but not by much - from 0.025 to 0.023 when $N = 3000$, and from 0.14 to 0.12 when $N = 10000$. It makes sense that imposing the cross-derivative restrictions would most benefit estimates of the cross-derivative. The findings are similar with point derivatives. Imposing the cross-derivative restriction has little effect on either the estimates of levels of ψ or the own derivatives. It does effect the cross-derivatives, however, and for the point cross-derivatives, the size of the effect is considerably larger. For example, at ($W_{i1} = 0$,$W_{i2} = 0$), the standard deviation of the cross derivative estimate decreases from 0.0127 to 0.0090 with $N = 3000$, and 0.0092 to 0.0064 for $N = 10000$. These are substantial increases in precision. Interestingly, this size of this increase can vary alot across the points - for example, at ($W_{i1} = 2$,$W_{i2} = -2$) the effect is considerably larger (in percentage terms), but at ($W_{i1} = -2$,$W_{i2} = 2$) it is quite small. Tables 3 and 4 show similar patterns for DGP B and DGP C.

7 Conclusion

Numeric optimization of objective functions, especially with many parameters, can be fraught with error. Hence, we study closed form estimators of a general class of semiparametric multinomial choice models. We combine and extend various results from the existing literature to enforce economic restrictions implied by the model and to try to attain estimators that are as efficient as possible. These closed form estimators appear to perform quite well in our Monte-Carlo experiments, though the sacrifice in efficiency relative to non-closed form, more computationally challenging estimators depends on the parameter or counterfactual of interest.

References

[1] Ahn, H., Ichimura, H., Powell, J., and Ruud, P. (2018) "Simple Estimators for Invertible Index Models", *Journal of Business and Economic Statistics,* 36:1, 1-10

[2] Allen, R. and Rehbeck, J. (2019), "Identification With Additively Separable Heterogeneity", *Econometrica*, 87:3, 1021-1054

[3] Berry, S., Levinsohn, J. and Pakes, A. (1995) "Automobile Prices in Market Equilibrium", *Econometrica*, 63:4, 841-890

[4] Cameron, AC and Trivedi, P (2005) *Microeconometrics: Methods and Applications*, Cambridge University Press

[5] Chen, X. (2007) "Large Sample Sieve Estimation of Semi-Nonparametric Models", *Handbook of Econometrics*, Vol 6B, Elsevier

[6] Cleveland, W. (1979) "Robust Locally Weighted Regression and Smoothing Scatterplots", *Journal of the American Statistical Association*, 74:368, 829-836

[7] Fan, J. (1992) "Design-adaptive Nonparametric Regression", *Journal of the American Statistical Association*, 87, 998-1004

[8] Fan, J. (1993) "Local Linear Regression Smoothers and their Minimax Efficiency", *Annals of Statistics*, 21, 196-216

[9] Fox, J. (2007) "Semiparametric estimation of multinomial discrete-choice models using a subset of choices", *The RAND Journal of Economics*, 38:4, 1002-1019

[10] Horowitz, J. (1998) *Semiparametric Methods in Econometrics*, Springer

[11] Horowitz, J. (1992) "A Smoothed Maximum Score Estimator for the Binary Response Model", *Econometrica*, 60:3, 505-531

[12] Imbens, G. and Wooldridge, J. (2007) Lecture Notes for Summer NBER

[13] Klein, and Spady, (1993) "An Efficient Semiparametric Estimator for Binary Response Models" *Econometrica*, 61:2, 387-421

[14] Lee, LF (1995) "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models", *Journal of Econometrics*, 65, 381-428

[15] Lehmann, EL (1983) Theory of Point Estimation, John Willey and Sons Inc.

[16] Manski, C. (1975) "Maximum Score Estimation of the Stochastic Utility Model of Choice", *Journal of Econometrics*, 3, 205-228

[17] Manski, C. (1985) "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator", *Journal of Econometrics*, 27:3, 313-333

[18] Matzkin, R. (1993) "Nonparametric identification and estimation of polychotomous choice models", *Journal of Econometrics*, 58:1-2, 137-168

[19] McCulloch, J. and Rossi, P. (1994) "An exact likelihood analysis of the multinomial probit model", *Journal of Econometrics*, 64:1-2, 207-240

[20] McFadden, D. and Train, K. (2000) "Mixed MNL Models for Discrete Response" *Journal of Applied Econometrics*, 15, 447-470

[21] Powell, J., Stock, J., and Stoker, T. (1989) "Semiparametric Estimation of Index Coefficients" *Econometrica*, 57:6, 1403-1403

[22] Stone, C. (1977) "Consistent Nonparametric Regression" *Annals of Statistics*, 5:4, 595-620

[23] Yan, J. (2017) "A Smoothed Maximum Score Estimator for Multinomial Discrete Choice Models" Mimeo

A Proofs

A.1 Proof of Lemma 1

Proof: For notational simplicity, Let $J = 2$. Note that

$$\mathbb{P}(Y = 1|X) = \int_{-\infty}^{X_1'\beta_1} \int_{u_1 - X_1'\beta_1 + X_2'\beta_2}^{+\infty} f_u(u_1, u_2) du_2 du_1 \equiv \psi_j(X_1'\beta_1, X_2'\beta_2)$$

which is differentiable in $X_j'\beta_j$.

We now show (3). W.l.o.g., let $k = 1$ and $j = 2$. Note that

$$\frac{\partial \mathbb{P}(Y = 1|X)}{\partial X_2'\beta_2} = -\int_{-\infty}^{X_1'\beta_1} f_u(u_1, u_1 - X_1'\beta_1 + X_2'\beta_2) du_1 = -\int_{-\infty}^{X_2'\beta_2} f_u(\tilde{u}_1 - X_2'\beta_2 + X_1'\beta_1, \tilde{u}_1) d\tilde{u}_1,$$

where in the last step we change the variable by letting $\tilde{u}_1 = u_1 - X_1'\beta_1 + X_2'\beta_2$. Moreover, we have

$$\frac{\partial \mathbb{P}(Y = 2|X)}{\partial X_1'\beta_1} = -\int_{-\infty}^{X_2'\beta_2} f_u(u_2 - X_2'\beta_2 + X_1'\beta_1, u_2) du_2 = \frac{\partial \mathbb{P}(Y = 1|X)}{\partial X_2'\beta_2}$$

which is strictly negative under assumption2-(i).

A.2 Proof of Theorem 2

Proof: For notational simplicity, we prove this theorem by letting $J = 2$. Note that

$$0 = \frac{\partial \mathbb{E}\Big[\mathbb{I}(Y = k) f_X(X)\Big]}{\partial X_j} = \frac{\partial \mathbb{E}\Big[\mathbb{P}(Y = k|X) f_X(X)\Big]}{\partial X_j}$$
$$= \mathbb{E}\left[\frac{\partial \psi_k(X_1'\beta_1, X_2'\beta_2)}{\partial X_j} f_X(X)\right] + \mathbb{E}\left[\mathbb{P}(Y = k|X) \times \frac{\partial f_X(X)}{\partial X_j}\right]$$

where the second equality applies the law of iterated expectation and Lemma 1. Thus,

$$\mathbb{E}\left[\mathbb{P}(Y = k|X) \times \frac{\partial f_X(X)}{\partial X_j}\right] = -\mathbb{E}\left[\frac{\partial \psi_k(X_1'\beta_1, X_2'\beta_2)}{\partial X_j'\beta_j} f_X(X)\right] \beta_j.$$

Let $\lambda_{kj} \equiv -\mathbb{E}^{-1}\left\{\left[\frac{\partial \psi_k(X_1'\beta_1, X_2'\beta_2)}{\partial X_j'\beta_j}\right] f_X(X)\right\}$. It follows that

$$\beta_j = \lambda_{kj} \mathbb{E}\left[\mathbb{P}(Y = k|X) \frac{\partial f_X(X)}{\partial X_j}\right],$$

and by Lemma 1, $\lambda_{jk} = \lambda_{kj} > 0$. ∎

A.3 Proof of Lemma 3

Proof: Let $I(\gamma) = \mathbb{E}[s(Y,X;\gamma) \times s'(Y,X;\gamma)]$ and $\hat{I}(\hat{\gamma}) = \frac{1}{N}\sum_{i=1}^{N}[\tilde{s}(Y_i,X_i;\tilde{\gamma}) \times \tilde{s}'(Y_i,X_i;\tilde{\gamma})]$. By a similar argument to Lemma 5 of Klein and Spady (1993), it is straightforward to see

$$\hat{I}(\tilde{\gamma}) - I(\gamma) = o_p(1),$$

as long as $\tilde{\gamma} \xrightarrow{p} \gamma$. Therefore, it suffices to show

$$\frac{1}{N}\sum_{i=1}^{N} s(Y_i,X_i;\tilde{\gamma}) - \frac{1}{N}\sum_{i=1}^{N} \tilde{s}(Y_i,X_i;\tilde{\gamma}) = o_p(N^{-\frac{1}{2}}).$$

Because

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} s(Y_i,X_i;\tilde{\gamma}) - \frac{1}{N}\sum_{i=1}^{N} \tilde{s}(Y_i,X_i;\tilde{\gamma}) \\
&\quad= \sum_{k\in\mathcal{J}_0}\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\mathbb{I}(Y_i=k)}{\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})}\frac{\partial\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})}{\partial\gamma} - \frac{\mathbb{I}(Y_i=k)}{\hat{\mathbb{P}}(Y_i=k|X_i,\hat{\gamma})}\frac{\partial\hat{\mathbb{P}}(Y_i=k|X_i,\hat{\gamma})}{\partial\gamma}\right] \\
&\quad= \sum_{k\in\mathcal{J}}\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\mathbb{I}(Y_i=k)}{\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})}\frac{\partial\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})}{\partial\gamma} - \frac{\mathbb{I}(Y_i=k)}{\hat{\mathbb{P}}(Y_i=k|X_i,\hat{\gamma})}\frac{\partial\hat{\mathbb{P}}(Y_i=k|X_i,\hat{\gamma})}{\partial\gamma}\right] \\
&\quad\quad- \sum_{k\in\mathcal{J}}\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\mathbb{I}(Y_i=0)}{\mathbb{P}(Y_i=0|X_i,\tilde{\gamma})}\frac{\partial\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})}{\partial\gamma} - \frac{\mathbb{I}(Y_i=0)}{\hat{\mathbb{P}}(Y_i=0|X_i,\hat{\gamma})}\frac{\partial\hat{\mathbb{P}}(Y_i=k|X_i,\hat{\gamma})}{\partial\gamma}\right] \\
&\quad= \sum_{k\in\mathcal{J}}\frac{1}{N}\sum_{i=1}^{N}\frac{[\mathbb{I}(Y_i=k)\mathbb{P}(Y_i\in\{0,k\}|X_i,\tilde{\gamma}) - \mathbb{I}(Y_i\in\{0,k\})\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})]}{\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})\mathbb{P}(Y_i=0|X_i,\tilde{\gamma})}\frac{\partial\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})}{\partial\gamma} \\
&\quad\quad- \sum_{k\in\mathcal{J}}\frac{1}{N}\sum_{i=1}^{N}\frac{[\mathbb{I}(Y_i=k)\hat{\mathbb{P}}(Y_i\in\{0,k\}|X_i,\tilde{\gamma}) - \mathbb{I}(Y_i\in\{0,k\})\hat{\mathbb{P}}(Y_i=k|X_i,\tilde{\gamma})]}{\hat{\mathbb{P}}(Y_i=k|X_i,\tilde{\gamma})\hat{\mathbb{P}}(Y_i=0|X_i,\tilde{\gamma})}\frac{\partial\hat{\mathbb{P}}(Y_i=k|X_i,\tilde{\gamma})}{\partial\gamma}.
\end{aligned}$$

Hence, it suffices to show that for any $k \in \mathcal{J}$, there is

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{[\mathbb{I}(Y_i=k)\mathbb{P}(Y_i\in\{0,1\}|X_i,\tilde{\gamma}) - \mathbb{I}(Y_i\in\{0,k\})\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})]}{\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})\mathbb{P}(Y_i=0|X_i,\tilde{\gamma})}\frac{\partial\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})}{\partial\gamma} \\
&\quad- \frac{1}{N}\sum_{i=1}^{N}\frac{[\mathbb{I}(Y_i=k)\hat{\mathbb{P}}(Y_i\in\{0,1\}|X_i,\tilde{\gamma}) - \mathbb{I}(Y_i\in\{0,k\})\hat{\mathbb{P}}(Y_i=k|X_i,\tilde{\gamma})]}{\hat{\mathbb{P}}(Y_i=k|X_i,\tilde{\gamma})\hat{\mathbb{P}}(Y_i=0|X_i,\tilde{\gamma})}\frac{\partial\hat{\mathbb{P}}(Y_i=k|X_i,\tilde{\gamma})}{\partial\gamma} = o_p(N^{-\frac{1}{2}}).
\end{aligned}$$

Denote $f_{\bar{W}}(\cdot\,;\gamma)$ as the density function of $\bar{W}(\gamma) \equiv (X_1'\gamma_1, \cdots, X_J'\gamma_J)$, which clearly depends on the value of γ. Moreover, let

$$\hat{f}_{\bar{W}}(\bar{W}_i(\gamma);\gamma) = \frac{1}{(N-1)\bar{h}^J}\sum_{n=1;n\neq i}^{N}\bar{K}\big(\frac{\bar{W}_i(\gamma) - \bar{W}_n(\gamma)}{\bar{h}_w}\big).$$

be an estimator of $f_{\bar{W}}(\bar{W}_i(\gamma);\gamma)$. Next, we rewrite the above condition as:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{[\mathbb{I}(Y_i=k)\mathbb{P}(Y_i\in\{0,k\}|X_i,\tilde{\gamma})-\mathbb{I}(Y_i\in\{0,k\})\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})]}{f_{\bar{W}}(\bar{W}_i(\tilde{\gamma});\tilde{\gamma})\times\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})\mathbb{P}(Y_i=0|X_i,\tilde{\gamma})}\times f_{\bar{W}}(\bar{W}_i(\tilde{\gamma});\tilde{\gamma})\frac{\partial\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})}{\partial\gamma}$$
$$-\frac{1}{N}\sum_{i=1}^{N}\frac{[\mathbb{I}(Y_i=k)\hat{\mathbb{P}}(Y_i\in\{0,1\}|X_i,\tilde{\gamma})-\mathbb{I}(Y_i\in\{0,k\})\hat{\mathbb{P}}(Y_i=k|X_i,\tilde{\gamma})]}{\hat{f}_{\bar{W}}(\bar{W}_i(\tilde{\gamma});\tilde{\gamma})\times\hat{\mathbb{P}}(Y_i=k|X_i,\tilde{\gamma})\hat{\mathbb{P}}(Y_i=0|X_i,\tilde{\gamma})}\times\hat{f}_{\bar{W}}(\bar{W}_i(\tilde{\gamma});\tilde{\gamma})\frac{\partial\hat{\mathbb{P}}(Y_i=k|X_i,\tilde{\gamma})}{\partial\gamma}$$
$$=o_p(N^{-\frac{1}{2}})$$

holding for all $k\in\mathcal{J}$.

Let $a_{ki}(\gamma)=f_{\bar{W}}(\bar{W}_i(\gamma);\gamma)\times\mathbb{P}(Y_i=k|X_i,\gamma)\mathbb{P}(Y_i=0|X_i,\gamma)$ and $\hat{a}_{ki}(\gamma)=\hat{f}_{\bar{W}}(\bar{W}_i(\gamma);\gamma)\times\hat{\mathbb{P}}(Y_i=k|X_i,\gamma)\hat{\mathbb{P}}(Y_i=0|X_i,\gamma)$. Let further

$$r_{ki}(\gamma)=\frac{[\mathbb{I}(Y_i=k)\mathbb{P}(Y_i\in\{0,1\}|X_i,\gamma)-\mathbb{I}(Y_i\in\{0,k\})\mathbb{P}(Y_i=k|X_i,\gamma)]}{a_{ki}(\gamma)},$$
$$\hat{r}_{ki}(\gamma)=\frac{[\mathbb{I}(Y_i=k)\hat{\mathbb{P}}(Y_i\in\{0,1\}|X_i,\gamma)-\mathbb{I}(Y_i\in\{0,k\})\hat{\mathbb{P}}(Y_i=k|X_i,\gamma)]}{\hat{a}_{ki}(\gamma)},$$
$$w_{ki}(\gamma)=f_{\bar{W}}(\bar{W}_i(\gamma);\gamma)\frac{\partial\mathbb{P}(Y_i=k|X_i,\gamma)}{\partial\gamma},\quad \hat{w}_{ki}(\gamma)=\hat{f}_{\bar{W}}(\bar{W}_i(\gamma);\gamma)\frac{\partial\hat{\mathbb{P}}(Y_i=k|X_i,\gamma)}{\partial\gamma}.$$

Therefore,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{[\mathbb{I}(Y_i=k)\mathbb{P}(Y_i\in\{0,k\}|X_i,\tilde{\gamma})-\mathbb{I}(Y_i\in\{0,k\})\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})]}{f_{\bar{W}}(\bar{W}_i(\tilde{\gamma});\tilde{\gamma})\times\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})\mathbb{P}(Y_i=0|X_i,\tilde{\gamma})}\times f_{\bar{W}}(\bar{W}_i(\tilde{\gamma});\tilde{\gamma})\frac{\partial\mathbb{P}(Y_i=k|X_i,\tilde{\gamma})}{\partial\gamma}$$
$$-\frac{1}{N}\sum_{i=1}^{N}\frac{[\mathbb{I}(Y_i=k)\hat{\mathbb{P}}(Y_i\in\{0,1\}|X_i,\tilde{\gamma})-\mathbb{I}(Y_i\in\{0,k\})\hat{\mathbb{P}}(Y_i=k|X_i,\tilde{\gamma})]}{\hat{f}_{\bar{W}}(\bar{W}_i(\tilde{\gamma});\tilde{\gamma})\times\hat{\mathbb{P}}(Y_i=k|X_i,\tilde{\gamma})\hat{\mathbb{P}}(Y_i=0|X_i,\tilde{\gamma})}\times\hat{f}_{\bar{W}}(\bar{W}_i(\tilde{\gamma});\tilde{\gamma})\frac{\partial\hat{\mathbb{P}}(Y_i=k|X_i,\tilde{\gamma})}{\partial\gamma}$$
$$=\frac{1}{N}\sum_{i=1}^{N}[r_{ki}(\tilde{\gamma})w_{ki}(\tilde{\gamma})-\hat{r}_{ki}(\tilde{\gamma})\hat{w}_{ki}(\tilde{\gamma})]=\frac{1}{N}\sum_{i=1}^{N}[r_{ki}(\tilde{\gamma})-\hat{r}_{ki}(\tilde{\gamma})]w_{ki}(\tilde{\gamma})+\frac{1}{N}\sum_{i=1}^{N}r_{ki}(\tilde{\gamma})[w_{ki}(\tilde{\gamma})-\hat{w}_{ki}(\tilde{\gamma})]$$
$$-\frac{1}{N}\sum_{i=1}^{N}[r_{ki}(\tilde{\gamma})-\hat{r}_{ki}(\tilde{\gamma})][w_{ki}(\tilde{\gamma})-\hat{w}_{ki}(\tilde{\gamma})]\equiv\mathbb{I}_1+\mathbb{I}_2+\mathbb{I}_3.$$

We now apply the Taylor expansion to $\mathbb{I}_1$ and obtain

$$\hat{r}_{ki}(\tilde{\gamma})=\hat{r}^*_{ki}(\tilde{\gamma})+o_p(N^{-1/2})$$

where $\hat{r}^*_{ki}(\gamma)=\frac{[\mathbb{I}(Y_i=k)\hat{\mathbb{P}}(Y_i\in\{0,1\}|X_i,\gamma)-\mathbb{I}(Y_i\in\{0,k\})\hat{\mathbb{P}}(Y_i=k|X_i,\gamma)]}{a_{ki}(\gamma)}\times[1-\frac{\hat{a}_{ki}(\gamma)-a_{ki}(\gamma)}{a_{ki}(\gamma)}]$. Let $\mathbb{I}^*_1$ be obtained from $\mathbb{I}_1$ by replacing $\hat{r}_{ki}(\tilde{\gamma})$ with $\hat{r}^*_{ki}(\tilde{\gamma})$. Thus, to establish $\mathbb{I}_1=o_p(N^{-\frac{1}{2}})$, it suffices to show $N\times\mathbb{E}[(\mathbb{I}^*_1)^2]\to 0$. Note that

$$N\times\mathbb{E}[(\mathbb{I}^*_1)^2]=\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}w^2_{ki}(\tilde{\gamma})[\hat{r}^*_{ki}(\tilde{\gamma})-r_{ki}(\tilde{\gamma})]^2+\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}w_{ki}(\tilde{\gamma})w_{kj}(\tilde{\gamma})[\hat{r}^*_{ki}(\tilde{\gamma})-r_i(\tilde{\gamma})][\hat{r}^*_{kj}(\tilde{\gamma})-r_{kj}(\tilde{\gamma})],$$

in which the expectation treats $\tilde{\gamma}$ as a constant (i.e. non-random). It is straightforward that the first term converges to zero since $\hat{r}^*_{ki}(\gamma)\xrightarrow{p}r_{ki(\gamma)}$. For the second term, let $\mathbb{X}=\{X_i:i\leq N\}$ be all the observed

covariates and $\mathbb{V}(\gamma) = \{(X_{i1}'\gamma_1, \cdots, X_{iJ}'\gamma_J) : i \leq N\}$. Then

$$\begin{aligned}
& \mathbb{E} w_{ki}(\tilde{\gamma}) w_{kj}(\tilde{\gamma}) [\hat{r}_{ki}^*(\tilde{\gamma}) - r_{ki}(\tilde{\gamma})][\hat{r}_{kj}^*(\tilde{\gamma}) - r_{kj}(\tilde{\gamma})] \\
= & \mathbb{E}\left(\mathbb{E}\left\{w_{ki}(\tilde{\gamma}) w_{kj}(\tilde{\gamma}) [\hat{r}_{ki}^*(\tilde{\gamma}) - r_{ki}(\tilde{\gamma})][\hat{r}_{kj}^*(\tilde{\gamma}) - r_{kj}(\tilde{\gamma})] \middle| \mathbb{X}\right\}\right) \\
= & \mathbb{E}\left(\mathbb{E}\left\{w_{ki}(\tilde{\gamma}) w_{kj}(\tilde{\gamma}) [\hat{r}_{ki}^*(\tilde{\gamma}) - r_{ki}(\tilde{\gamma})][\hat{r}_{kj}^*(\tilde{\gamma}) - r_{kj}(\tilde{\gamma})] \middle| \mathbb{V}(\tilde{\gamma})\right\}\right) \\
= & \mathbb{E}\left(\mathbb{E}[w_{ki}(\tilde{\gamma}) w_{kj}(\tilde{\gamma}) | \mathbb{V}(\tilde{\gamma})] \times \mathbb{E}\left\{[\hat{r}_{ki}^*(\tilde{\gamma}) - r_{ki}(\tilde{\gamma})][\hat{r}_{kj}^*(\tilde{\gamma}) - r_{kj}(\tilde{\gamma})] \middle| \mathbb{V}(\tilde{\gamma})\right\}\right)
\end{aligned}$$

Because $\tilde{\gamma} \in \mathcal{B}(\gamma, \frac{\ln N}{\sqrt{N}})$ with probability approaching to one and $\mathbb{E}[w_{ki}(\gamma)|\mathbb{V}(\gamma)] = 0$, therefore,

$$\mathbb{E}[w_{ki}(\tilde{\gamma}) w_{kj}(\tilde{\gamma}) | \mathbb{V}(\tilde{\gamma})] = \mathbb{E}[w_{ki}(\tilde{\gamma}) | \mathbb{V}(\tilde{\gamma})] \times \mathbb{E}[w_{kj}(\tilde{\gamma}) | \mathbb{V}(\tilde{\gamma})] = o_p(\ln^2 N / N).$$

and

$$\begin{aligned}
& \left|\mathbb{E}\left\{[\hat{r}_{ki}^*(\tilde{\gamma}) - r_{ki}(\tilde{\gamma})][\hat{r}_{kj}^*(\tilde{\gamma}) - r_{kj}(\tilde{\gamma})] \middle| \mathbb{V}(\tilde{\gamma})\right\}\right| \\
& \qquad \leq \mathbb{E}\left\{[\hat{r}_{ki}^*(\tilde{\gamma}) - r_{ki}(\tilde{\gamma})]^2 \middle| \mathbb{V}(\tilde{\gamma})\right\} + \mathbb{E}\left\{[\hat{r}_{kj}^*(\tilde{\gamma}) - r_j(\tilde{\gamma})]^2 \middle| \mathbb{V}(\tilde{\gamma})\right\} = 2\mathbb{E}\left\{[\hat{r}_{ki}^*(\tilde{\gamma}) - r_{ki}(\tilde{\gamma})]^2 \middle| \mathbb{V}(\tilde{\gamma})\right\}
\end{aligned}$$

which converges to zero at a nonparametric rate (faster than $\ln^2 N$). Thus,

$$\mathbb{E}\left\{w_{ki}(\tilde{\gamma}) w_{kj}(\tilde{\gamma}) [\hat{r}_{ki}^*(\tilde{\gamma}) - r_{ki}(\tilde{\gamma})][\hat{r}_{kj}^*(\tilde{\gamma}) - r_{kj}(\tilde{\gamma})]\right\} = o_p(N).$$

Regarding $\mathbb{I}_2$, similarly we have

$$\begin{aligned}
N \times \mathbb{E}(\mathbb{I}_2^2) = & \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left\{r_{ki}^2(\tilde{\gamma}) [\hat{w}_{ki}(\tilde{\gamma}) - w_{ki}(\tilde{\gamma})]^2\right\} \\
& + \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq i} \mathbb{E}\left\{r_{ki}(\tilde{\gamma}) r_{kj}(\tilde{\gamma}) [\hat{w}_{ki}(\tilde{\gamma}) - w_{ki}(\tilde{\gamma})][\hat{w}_{kj}(\tilde{\gamma}) - w_{kj}(\tilde{\gamma})]\right\}.
\end{aligned}$$

The first term is $o_p(1)$. For the second term, note that

$$\mathbb{E}[r_{ki}(\tilde{\gamma}) r_{kj}(\tilde{\gamma}) \hat{w}_{ki}(\tilde{\gamma}) w_{kj}(\tilde{\gamma})] = \mathbb{E}\left\{\mathbb{E}[r_{ki}(\tilde{\gamma}) | \mathbb{V}(\tilde{\gamma})] \mathbb{E}[r_{kj}(\tilde{\gamma}) \hat{w}_{ki}(\tilde{\gamma}) w_{kj}(\tilde{\gamma}) | \mathbb{V}(\tilde{\gamma})]\right\}$$

in which $\mathbb{E}[r_{ki}(\tilde{\gamma})|\mathbb{V}(\tilde{\gamma})] = \mathbb{E}[r_{ki}(\tilde{\gamma})|X_i, \tilde{\gamma}] = 0$. Thus, the second term equals to

$$\begin{aligned}
& \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq i} \mathbb{E}\left[r_{ki}(\tilde{\gamma}) r_{kj}(\tilde{\gamma}) \hat{w}_{ki}(\tilde{\gamma}) \hat{w}_{kj}(\tilde{\gamma})\right] \\
& \qquad = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq i} \mathbb{E}\left[r_{ki}(\tilde{\gamma}) r_{kj}(\tilde{\gamma}) [\hat{w}_{ki(j)}(\tilde{\gamma}) + \hat{\delta}_{ki(j)}(\tilde{\gamma})][\hat{w}_{kj(i)}(\tilde{\gamma}) + \hat{\delta}_{kj(i)}(\tilde{\gamma})]\right]
\end{aligned}$$

where $\hat{w}_{ki(j)}(\tilde{\gamma})$ obtains from $\hat{w}_{ki}(\tilde{\gamma})$ by replacing Y_j with 0, and $\hat{\delta}_{ki(j)}(\tilde{\gamma}) = \hat{w}_{ki}(\tilde{\gamma}) - \hat{w}_{ki(j)}(\tilde{\gamma})$. By the law of iterated expectation, we have

$$\begin{aligned}
& \mathbb{E}[r_{ki}(\tilde{\gamma}) r_{kj}(\tilde{\gamma}) \hat{\delta}_{ki(j)}(\tilde{\gamma}) \hat{w}_{kj(i)}(\tilde{\gamma})] = 0, \\
& \mathbb{E}[r_{ki}(\tilde{\gamma}) r_{kj}(\tilde{\gamma}) \hat{w}_{ki(j)}(\tilde{\gamma}) \hat{\delta}_{kj(i)}(\tilde{\gamma})] = 0, \\
& \mathbb{E}[r_{ki}(\tilde{\gamma}) r_{kj}(\tilde{\gamma}) \hat{w}_{ki(j)}(\tilde{\gamma}) \hat{w}_{kj(i)}(\tilde{\gamma})] = 0.
\end{aligned}$$

Thus, the second term in $N \times \mathbb{E}(\mathbb{I}_2^2)$ is also an $o(1)$. Furthermore, it is straightforward that $\mathbb{I}_3 = o_p(N^{-\frac{1}{2}})$.

TABLE 1- DGP A

N=3000

	Truth	Logit			AMXbasic			AMXindex			AMXNewton		
		Mean	SD	rootMSE	Mean	SD	rootMSE	Mean	SD	rootMSE	Mean	SD	rootMSE
gamma1	1.0000	1.0023	0.0396	0.0397	1.0029	0.0572	0.0573				0.9995	0.0410	0.0410
gamma2	1.0000	1.0011	0.0405	0.0405	1.0022	0.0575	0.0575				0.9986	0.0421	0.0421
beta12	1.0000	1.0023	0.0516	0.0517	1.0111	0.1294	0.1299	1.0053	0.0897	0.0899	1.0050	0.0853	0.0854
Ave(dPr(y=2)/dx22)	0.1047	0.1043	0.0030	0.0030	0.0950	0.0032	0.0102				0.0949	0.0026	0.0101
Ave(dPr(y=1)/dx22)	-0.0455	-0.0454	0.0019	0.0019	-0.0416	0.0023	0.0045				-0.0415	0.0021	0.0045
Bad Step %											0		

N=10000

	Truth	Logit			AMXbasic			AMXindex			AMXNewton		
		Mean	SD	rootMSE	Mean	SD	rootMSE	Mean	SD	rootMSE	Mean	SD	rootMSE
gamma1	1.0000	0.9999	0.0219	0.0219	1.0007	0.0313	0.0313				0.9992	0.0224	0.0224
gamma2	1.0000	1.0002	0.0216	0.0216	1.0009	0.0303	0.0303				0.9995	0.0220	0.0220
beta12	1.0000	1.0001	0.0278	0.0278	1.0016	0.0696	0.0696	1.0008	0.0486	0.0486	1.0006	0.0463	0.0463
Ave(dPr(y=2)/dx22)	0.1047	0.1042	0.0016	0.0017	0.0976	0.0018	0.0073				0.0975	0.0014	0.0073
Ave(dPr(y=1)/dx22)	-0.0455	-0.0454	0.0010	0.0010	-0.0426	0.0013	0.0032				-0.0426	0.0012	0.0031
Bad Step %											0		

DGP B

N=3000

	Truth	Logit			AMXbasic			AMXindex			AMXNewton		
		Mean	SD	rootMSE	Mean	SD	rootMSE	Mean	SD	rootMSE	Mean	SD	rootMSE
gamma1	1.0000	1.0021	0.0528	0.0528	1.0033	0.0725	0.0726				0.9975	0.0526	0.0527
gamma2	1.0000	1.0027	0.0523	0.0524	1.0055	0.0729	0.0731				0.9979	0.0521	0.0521
beta12	1.0000	1.0026	0.0627	0.0628	1.0103	0.1375	0.1379	1.0046	0.0975	0.0976	1.0047	0.0903	0.0904
Ave(dPr(y=2)/dx22)	0.0892	0.0894	0.0031	0.0031	0.0809	0.0035	0.0090				0.0807	0.0028	0.0089
Ave(dPr(y=1)/dx22)	-0.0453	-0.0354	0.0016	0.0100	-0.0414	0.0025	0.0046				-0.0413	0.0023	0.0046
Bad Step %											0		

N=10000

	Truth	Logit			AMXbasic			AMXindex			AMXNewton		
		Mean	SD	rootMSE	Mean	SD	rootMSE	Mean	SD	rootMSE	Mean	SD	rootMSE
gamma1	1.0000	1.0002	0.0280	0.0280	1.0003	0.0392	0.0392				0.9988	0.0273	0.0273
gamma2	1.0000	1.0003	0.0284	0.0284	1.0006	0.0389	0.0389				0.9991	0.0272	0.0272
beta12	1.0000	1.0004	0.0329	0.0329	1.0010	0.0731	0.0731	1.0007	0.0524	0.0524	1.0005	0.0484	0.0484
Ave(dPr(y=2)/dx22)	0.0892	0.0894	0.0017	0.0017	0.0831	0.0019	0.0064				0.0831	0.0015	0.0063
Ave(dPr(y=1)/dx22)	-0.0453	-0.0354	0.0009	0.0099	-0.0424	0.0014	0.0032				-0.0424	0.0013	0.0032
Bad Step %											0		

DGP C

N=3000

	Truth	Logit			AMXbasic			AMXindex			AMXNewton		
		Mean	SD	rootMSE	Mean	SD	rootMSE	Mean	SD	rootMSE	Mean	SD	rootMSE
gamma1	1.0000	1.0007	0.0423	0.0423	1.0010	0.0601	0.0601				0.9982	0.0424	0.0424
gamma2	1.0000	1.0022	0.0540	0.0540	1.0049	0.0783	0.0785				0.9961	0.0518	0.0519
beta12	1.0000	0.8163	0.0469	0.1896	1.0099	0.1724	0.1727	1.0041	0.1142	0.1143	1.0067	0.1101	0.1103
Ave(dPr(y=2)/dx22)	0.0897	0.0894	0.0031	0.0031	0.0805	0.0036	0.0099				0.0803	0.0028	0.0098
Ave(dPr(y=1)/dx22)	-0.0391	-0.0368	0.0018	0.0029	-0.0355	0.0025	0.0044				-0.0354	0.0022	0.0043
Bad Step %											0		

N=10000

	Truth	Logit			AMXbasic			AMXindex			AMXNewton		
		Mean	SD	rootMSE	Mean	SD	rootMSE	Mean	SD	rootMSE	Mean	SD	rootMSE
gamma1	1.0000	1.0002	0.0230	0.0230	1.0003	0.0329	0.0329				0.9994	0.0225	0.0225
gamma2	1.0000	0.9997	0.0295	0.0295	0.9996	0.0423	0.0423				0.9979	0.0268	0.0269
beta12	1.0000	0.8162	0.0260	0.1856	1.0043	0.0923	0.0924	1.0026	0.0636	0.0637	1.0028	0.0603	0.0604
Ave(dPr(y=2)/dx22)	0.0897	0.0894	0.0017	0.0017	0.0828	0.0020	0.0072				0.0828	0.0015	0.0071
Ave(dPr(y=1)/dx22)	-0.0391	-0.0368	0.0010	0.0025	-0.0365	0.0014	0.0030				-0.0365	0.0012	0.0029
Bad Step %											0		

Table 2 - Specification A

	N=3000				N=10000			
	LLK no Restriction		LLK with Restriction		LLK no Restriction		LLK with Restriction	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Ave(dPr(y=2)/dX22)	0.1086	0.0032	0.1086	0.0032	0.1076	0.0017	0.1076	0.0017
Ave(dPr(y=1)/dx22)	-0.0475	0.0025	-0.0475	0.0023	-0.047	0.0014	-0.047	0.0012
Pr(y=1| x1B1=0, x2B2=0)	0.3449	0.0173	0.3449	0.0173	0.3430	0.0109	0.3431	0.0109
Pr(y=2| x1B1=0, x2B2=0)	0.3452	0.0173	0.3452	0.0173	0.3428	0.0109	0.3428	0.0109
dPr(y=2| x1B1=0, x2B2=0)/dX22	0.1840	0.0129	0.1840	0.0128	0.1913	0.0090	0.1913	0.0090
dPr(y=1| x1B1=0, x2B2=0)/dX22	-0.0875	0.0127	-0.0874	0.0090	-0.0917	0.0092	-0.0917	0.0064
Pr(y=1| x1B1=1, x2B2=1)	0.4240	0.0197	0.4240	0.0195	0.4244	0.0124	0.4243	0.0122
Pr(y=2| x1B1=1, x2B2=1)	0.4245	0.0198	0.4245	0.0196	0.4240	0.0123	0.4240	0.0122
dPr(y=2| x1B1=1, x2B2=1)/dX22	0.1909	0.0137	0.1909	0.0138	0.1997	0.0097	0.1997	0.0097
dPr(y=1| x1B1=1, x2B2=1)/dX22	-0.1264	0.0135	-0.1264	0.0097	-0.1348	0.0099	-0.1349	0.0070
Pr(y=1| x1B1=-1, x2B2=-1)	0.2382	0.0172	0.2382	0.0170	0.2331	0.0107	0.2331	0.0106
Pr(y=2| x1B1=-1, x2B2=-1)	0.2377	0.0168	0.2377	0.0166	0.2328	0.0105	0.2328	0.0105
dPr(y=2| x1B1=-1, x2B2=-1)/dX22	0.1593	0.0126	0.1593	0.0126	0.1616	0.0088	0.1616	0.0088
dPr(y=1| x1B1=-1, x2B2=-1)/dX22	-0.0477	0.0123	-0.0477	0.0087	-0.0476	0.0087	-0.0476	0.0061
Pr(y=1| x1B1=-1, x2B2=1)	0.1244	0.0122	0.1244	0.0122	0.1174	0.0076	0.1174	0.0076
Pr(y=2| x1B1=-1, x2B2=1)	0.6217	0.0189	0.6217	0.0188	0.6299	0.0118	0.6299	0.0118
dPr(y=2| x1B1=-1, x2B2=1)/dX22	0.1983	0.0137	0.1983	0.0137	0.2044	0.0099	0.2044	0.0099
dPr(y=1| x1B1=-1, x2B2=1)/dX22	-0.0652	0.0095	-0.0651	0.0079	-0.0649	0.0066	-0.0649	0.0055
Pr(y=1| x1B1=1, x2B2=-1)	0.6215	0.0192	0.6215	0.0192	0.6301	0.0120	0.6301	0.0120
Pr(y=2| x1B1=1, x2B2=-1)	0.1246	0.0123	0.1246	0.0123	0.1174	0.0077	0.1173	0.0077
dPr(y=2| x1B1=1, x2B2=-1)/dX22	0.1060	0.0105	0.1060	0.0105	0.1024	0.0071	0.1024	0.0071
dPr(y=1| x1B1=1, x2B2=-1)/dX22	-0.0651	0.0141	-0.0651	0.0082	-0.0649	0.0100	-0.0648	0.0057
Pr(y=1| x1B1=2, x2B2=2)	0.4677	0.0254	0.4678	0.0243	0.4687	0.0159	0.4687	0.0154
Pr(y=2| x1B1=2, x2B2=2)	0.4690	0.0257	0.4689	0.0246	0.4681	0.0156	0.4681	0.0151
dPr(y=2| x1B1=2, x2B2=2)/dX22	0.1895	0.0158	0.1895	0.0158	0.1986	0.0112	0.1986	0.0112
dPr(y=1| x1B1=2, x2B2=2)/dX22	-0.1543	0.0158	-0.1541	0.0111	-0.1648	0.0112	-0.1649	0.0080
Pr(y=1| x1B1=-2, x2B2=-2)	0.1343	0.0168	0.1343	0.0163	0.1284	0.0101	0.1284	0.0099
Pr(y=2| x1B1=-2, x2B2=-2)	0.1339	0.0167	0.1338	0.0163	0.1284	0.0101	0.1283	0.0099
dPr(y=2| x1B1=-2, x2B2=-2)/dX22	0.1169	0.0126	0.1169	0.0126	0.1133	0.0087	0.1133	0.0087
dPr(y=1| x1B1=-2, x2B2=-2)/dX22	-0.0197	0.0120	-0.0197	0.0084	-0.0176	0.0084	-0.0177	0.0059
Pr(y=1| x1B1=-2, x2B2=2)	0.0276	0.0070	0.0276	0.0071	0.0251	0.0043	0.0251	0.0043
Pr(y=2| x1B1=-2, x2B2=2)	0.8266	0.0184	0.8266	0.0178	0.8348	0.0113	0.8348	0.0111
dPr(y=2| x1B1=-2, x2B2=2)/dX22	0.1451	0.0138	0.1451	0.0138	0.1406	0.0093	0.1406	0.0093
dPr(y=1| x1B1=-2, x2B2=2)/dX22	-0.0283	0.0061	-0.0283	0.0057	-0.0249	0.0041	-0.0249	0.0038
Pr(y=1| x1B1=2, x2B2=-2)	0.8267	0.0184	0.8266	0.0179	0.8350	0.0110	0.8350	0.0108
Pr(y=2| x1B1=2, x2B2=-2)	0.0276	0.0070	0.0276	0.0071	0.0251	0.0043	0.0251	0.0043
dPr(y=2| x1B1=2, x2B2=-2)/dX22	0.0385	0.0066	0.0385	0.0066	0.0325	0.0043	0.0325	0.0043
dPr(y=1| x1B1=2, x2B2=-2)/dX22	-0.0285	0.0135	-0.0284	0.0059	-0.0250	0.0094	-0.0249	0.0038

Table 3 - Specification B

	N=3000				N=10000			
	LLK no Restriction		LLK with Restriction		LLK no Restriction		LLK with Restriction	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Ave(dPr(y=2)/dX22)	0.0923	0.0033	0.0923	0.0033	0.0917	0.0017	0.0917	0.0017
Ave(dPr(y=1)/dx22)	-0.0473	0.0027	-0.0472	0.0026	-0.0468	0.0015	-0.0468	0.0013
Pr(y=1| x1B1=0, x2B2=0)	0.2995	0.0171	0.2994	0.0171	0.2977	0.0106	0.2977	0.0106
Pr(y=2| x1B1=0, x2B2=0)	0.2990	0.0170	0.2990	0.0170	0.2974	0.0105	0.2974	0.0105
dPr(y=2| x1B1=0, x2B2=0)/dX22	0.1425	0.0126	0.1424	0.0126	0.1477	0.0089	0.1477	0.0089
dPr(y=1| x1B1=0, x2B2=0)/dX22	-0.0881	0.0126	-0.0880	0.0093	-0.0927	0.0088	-0.0928	0.0063
Pr(y=1| x1B1=1, x2B2=1)	0.3513	0.0194	0.3513	0.0192	0.3503	0.0120	0.3503	0.0120
Pr(y=2| x1B1=1, x2B2=1)	0.3515	0.0197	0.3515	0.0195	0.3499	0.0120	0.3499	0.0119
dPr(y=2| x1B1=1, x2B2=1)/dX22	0.1565	0.0137	0.1565	0.0137	0.1633	0.0096	0.1633	0.0096
dPr(y=1| x1B1=1, x2B2=1)/dX22	-0.1067	0.0136	-0.1068	0.0100	-0.1133	0.0096	-0.1133	0.0069
Pr(y=1| x1B1=-1, x2B2=-1)	0.2438	0.0178	0.2437	0.0176	0.2416	0.0108	0.2416	0.0107
Pr(y=2| x1B1=-1, x2B2=-1)	0.2432	0.0173	0.2432	0.0170	0.2413	0.0109	0.2413	0.0108
dPr(y=2| x1B1=-1, x2B2=-1)/dX22	0.1256	0.0128	0.1256	0.0128	0.1288	0.0087	0.1288	0.0087
dPr(y=1| x1B1=-1, x2B2=-1)/dX22	-0.0696	0.0129	-0.0696	0.0093	-0.0724	0.0091	-0.0724	0.0063
Pr(y=1| x1B1=-1, x2B2=1)	0.1079	0.0118	0.1079	0.0118	0.1013	0.0071	0.1013	0.0071
Pr(y=2| x1B1=-1, x2B2=1)	0.5265	0.0206	0.5265	0.0204	0.5314	0.0127	0.5314	0.0126
dPr(y=2| x1B1=-1, x2B2=1)/dX22	0.1471	0.0146	0.1471	0.0146	0.1487	0.0103	0.1487	0.0103
dPr(y=1| x1B1=-1, x2B2=1)/dX22	-0.0637	0.0093	-0.0637	0.0080	-0.0632	0.0065	-0.0631	0.0056
Pr(y=1| x1B1=1, x2B2=-1)	0.5266	0.0208	0.5266	0.0206	0.5315	0.0127	0.5315	0.0127
Pr(y=2| x1B1=1, x2B2=-1)	0.1077	0.0117	0.1077	0.0118	0.1013	0.0072	0.1013	0.0072
dPr(y=2| x1B1=1, x2B2=-1)/dX22	0.0866	0.0099	0.0866	0.0099	0.0839	0.0067	0.0839	0.0067
dPr(y=1| x1B1=1, x2B2=-1)/dX22	-0.0639	0.0148	-0.0637	0.0084	-0.0629	0.0106	-0.0630	0.0056
Pr(y=1| x1B1=2, x2B2=2)	0.3966	0.0260	0.3965	0.0252	0.3951	0.0159	0.3952	0.0155
Pr(y=2| x1B1=2, x2B2=2)	0.3965	0.0260	0.3965	0.0252	0.3954	0.0159	0.3954	0.0154
dPr(y=2| x1B1=2, x2B2=2)/dX22	0.1669	0.0162	0.1669	0.0162	0.1751	0.0116	0.1751	0.0116
dPr(y=1| x1B1=2, x2B2=2)/dX22	-0.1244	0.0159	-0.1247	0.0115	-0.1330	0.0116	-0.1328	0.0081
Pr(y=1| x1B1=-2, x2B2=-2)	0.1883	0.0203	0.1883	0.0196	0.1857	0.0123	0.1857	0.0119
Pr(y=2| x1B1=-2, x2B2=-2)	0.1880	0.0202	0.1880	0.0194	0.1863	0.0121	0.1863	0.0118
dPr(y=2| x1B1=-2, x2B2=-2)/dX22	0.1068	0.0137	0.1068	0.0137	0.1078	0.0097	0.1078	0.0097
dPr(y=1| x1B1=-2, x2B2=-2)/dX22	-0.0524	0.0145	-0.0523	0.0102	-0.0535	0.0099	-0.0537	0.0070
Pr(y=1| x1B1=-2, x2B2=2)	0.0232	0.0066	0.0232	0.0066	0.0209	0.0039	0.0209	0.0039
Pr(y=2| x1B1=-2, x2B2=2)	0.6916	0.0242	0.6916	0.0234	0.6940	0.0146	0.6940	0.0142
dPr(y=2| x1B1=-2, x2B2=2)/dX22	0.1172	0.0163	0.1172	0.0163	0.1145	0.0115	0.1145	0.0115
dPr(y=1| x1B1=-2, x2B2=2)/dX22	-0.0257	0.0059	-0.0257	0.0057	-0.0221	0.0038	-0.0221	0.0037
Pr(y=1| x1B1=2, x2B2=-2)	0.6912	0.0240	0.6912	0.0232	0.6944	0.0146	0.6945	0.0142
Pr(y=2| x1B1=2, x2B2=-2)	0.0231	0.0065	0.0231	0.0065	0.0210	0.0039	0.0210	0.0039
dPr(y=2| x1B1=2, x2B2=-2)/dX22	0.0319	0.0063	0.0319	0.0063	0.0269	0.0040	0.0269	0.0040
dPr(y=1| x1B1=2, x2B2=-2)/dX22	-0.0259	0.0168	-0.0257	0.0058	-0.0220	0.0119	-0.0221	0.0038

Table 4 - Specification C

	N=3000				N=10000			
	LLK no Restriction		LLK with Restriction		LLK no Restriction		LLK with Restriction	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Ave(dPr(y=2)/dX22)	0.0919	0.0032	0.0919	0.0032	0.0914	0.0017	0.0914	0.0017
Ave(dPr(y=1)/dx22)	-0.0405	0.0026	-0.0405	0.0023	-0.0403	0.0014	-0.0403	0.0012
Pr(y=1| x1B1=0, x2B2=0)	0.3414	0.0174	0.3414	0.0174	0.3369	0.0111	0.3369	0.0111
Pr(y=2| x1B1=0, x2B2=0)	0.3406	0.0180	0.3406	0.0180	0.3454	0.0112	0.3454	0.0112
dPr(y=2| x1B1=0, x2B2=0)/dX22	0.1160	0.0135	0.1160	0.0135	0.1159	0.0095	0.1159	0.0095
dPr(y=1| x1B1=0, x2B2=0)/dX22	-0.0592	0.0131	-0.0591	0.0094	-0.0595	0.0093	-0.0595	0.0067
Pr(y=1| x1B1=1, x2B2=1)	0.4421	0.0203	0.4421	0.0199	0.4411	0.0127	0.4411	0.0125
Pr(y=2| x1B1=1, x2B2=1)	0.3856	0.0201	0.3856	0.0199	0.3888	0.0124	0.3888	0.0123
dPr(y=2| x1B1=1, x2B2=1)/dX22	0.1134	0.0147	0.1134	0.0147	0.1112	0.0103	0.1112	0.0103
dPr(y=1| x1B1=1, x2B2=1)/dX22	-0.0776	0.0147	-0.0775	0.0104	-0.0777	0.0104	-0.0776	0.0073
Pr(y=1| x1B1=-1, x2B2=-1)	0.2275	0.0169	0.2275	0.0168	0.2205	0.0102	0.2205	0.0102
Pr(y=2| x1B1=-1, x2B2=-1)	0.2687	0.0183	0.2687	0.0181	0.2723	0.0115	0.2722	0.0114
dPr(y=2| x1B1=-1, x2B2=-1)/dX22	0.1234	0.0127	0.1234	0.0127	0.1276	0.0093	0.1276	0.0093
dPr(y=1| x1B1=-1, x2B2=-1)/dX22	-0.0399	0.0121	-0.0400	0.0088	-0.0401	0.0089	-0.0401	0.0065
Pr(y=1| x1B1=-1, x2B2=1)	0.1572	0.0144	0.1572	0.0143	0.1528	0.0089	0.1528	0.0089
Pr(y=2| x1B1=-1, x2B2=1)	0.4978	0.0210	0.4979	0.0209	0.4977	0.0127	0.4977	0.0126
dPr(y=2| x1B1=-1, x2B2=1)/dX22	0.1122	0.0147	0.1122	0.0147	0.1094	0.0107	0.1094	0.0107
dPr(y=1| x1B1=-1, x2B2=1)/dX22	-0.0360	0.0105	-0.0359	0.0087	-0.0335	0.0076	-0.0336	0.0062
Pr(y=1| x1B1=1, x2B2=-1)	0.6035	0.0193	0.6035	0.0193	0.6094	0.0121	0.6094	0.0121
Pr(y=2| x1B1=1, x2B2=-1)	0.1526	0.0139	0.1526	0.0139	0.1496	0.0086	0.1496	0.0086
dPr(y=2| x1B1=1, x2B2=-1)/dX22	0.1115	0.0108	0.1115	0.0108	0.1168	0.0075	0.1168	0.0075
dPr(y=1| x1B1=1, x2B2=-1)/dX22	-0.0725	0.0140	-0.0725	0.0089	-0.0776	0.0100	-0.0777	0.0061
Pr(y=1| x1B1=2, x2B2=2)	0.5096	0.0260	0.5096	0.0248	0.5098	0.0164	0.5099	0.0158
Pr(y=2| x1B1=2, x2B2=2)	0.4111	0.0260	0.4111	0.0249	0.4127	0.0162	0.4127	0.0156
dPr(y=2| x1B1=2, x2B2=2)/dX22	0.1144	0.0173	0.1144	0.0173	0.1119	0.0125	0.1119	0.0125
dPr(y=1| x1B1=2, x2B2=2)/dX22	-0.0936	0.0173	-0.0936	0.0122	-0.0934	0.0125	-0.0933	0.0086
Pr(y=1| x1B1=-2, x2B2=-2)	0.1288	0.0166	0.1288	0.0163	0.1225	0.0100	0.1225	0.0098
Pr(y=2| x1B1=-2, x2B2=-2)	0.1708	0.0194	0.1708	0.0188	0.1694	0.0119	0.1694	0.0116
dPr(y=2| x1B1=-2, x2B2=-2)/dX22	0.1239	0.0127	0.1239	0.0127	0.1294	0.0091	0.1294	0.0091
dPr(y=1| x1B1=-2, x2B2=-2)/dX22	-0.0219	0.0117	-0.0219	0.0087	-0.0213	0.0081	-0.0213	0.0062
Pr(y=1| x1B1=-2, x2B2=2)	0.0544	0.0106	0.0543	0.0105	0.0517	0.0065	0.0517	0.0065
Pr(y=2| x1B1=-2, x2B2=2)	0.6527	0.0250	0.6527	0.0245	0.6530	0.0152	0.6530	0.0149
dPr(y=2| x1B1=-2, x2B2=2)/dX22	0.1336	0.0159	0.1336	0.0159	0.1390	0.0114	0.1390	0.0114
dPr(y=1| x1B1=-2, x2B2=2)/dX22	-0.0239	0.0078	-0.0240	0.0071	-0.0229	0.0055	-0.0229	0.0050
Pr(y=1| x1B1=2, x2B2=-2)	0.8286	0.0184	0.8286	0.0179	0.8387	0.0110	0.8387	0.0108
Pr(y=2| x1B1=2, x2B2=-2)	0.0261	0.0060	0.0261	0.0060	0.0214	0.0033	0.0214	0.0033
dPr(y=2| x1B1=2, x2B2=-2)/dX22	0.0528	0.0074	0.0528	0.0074	0.0465	0.0046	0.0465	0.0046
dPr(y=1| x1B1=2, x2B2=-2)/dX22	-0.0392	0.0137	-0.0390	0.0066	-0.0352	0.0092	-0.0354	0.0041