PANEL DATA DISCRETE CHOICE MODELS WITH LAGGED DEPENDENT VARIABLES*

Bo E. Honoré† Ekaterini Kyriazidou‡

May 1998

Abstract

In this paper, we consider identification and estimation in panel data discrete choice models when the explanatory variable set includes strictly exogenous variables, lags of the endogenous dependent variable as well as unobservable individual-specific effects. For the binary logit model with the dependent variable lagged only once, Chamberlain (1993) gave conditions under which the model is not identified. We present a stronger set of conditions under which the parameters of the model are identified. The identification result suggests estimators of the model, and we show that these are consistent and asymptotically normal, although their rate of convergence is slower than the inverse of the square root of the sample size. We also consider identification in the semiparametric case where the logit assumption is relaxed. We propose an estimator in the spirit of the conditional maximum score estimator (Manski (1987)), and we show that it is consistent. In addition, we discuss an extension of the identification result to multinomial discrete choice models, and to the case where the dependent variable is lagged twice. Finally, we present some Monte Carlo evidence on the small sample performance of the proposed estimators for the binary response model.

1 Introduction.

In many situations, such as in the study of labor force and union participation, accident occurrence, unemployment, purchase decisions, etc., it is observed that an individual who has experienced an event in the past

*Both authors gratefully acknowledge financial support from the National Science Foundation. Honoré in addition aknowledges support from the Alfred P. Sloan Foundation. We thank a co-editor, three anonymous referees and seminar participants at many institutions for helpful comments. All errors are our responsibility. Updated versions of the paper may be found at the world wide web site: http://www.princeton.edu/~honore/papers.

†Department of Economics, Princeton University, Princeton, New Jersey 08544.

‡Department of Economics, University of Chicago, Chicago, Illinois 60637.

is more likely to experience the event in the future than an individual who has not experienced the event. Heckman (1981a, b) discusses two explanations for this phenomenon. The first explanation is the presence of "true state dependence", in the sense that the lagged choice/decision enters the model in a structural way as an explanatory variable. The second is the presence of serial correlation in the unobserved transitory errors that underlie the threshold-crossing econometric specification of the model. Of particular interest is the case where this serial correlation is due to the presence of unobservable permanent individual/choice-specific heterogeneity, i.e. to different propensities across individuals to experience the event. Heckman calls the latter source of serial correlation "spurious state dependence". Distinguishing between these two explanations is important, for example, in evaluating the effect of economic policies that aim to alleviate short-term unemployment (see Phelps (1972)), or the effect of training programs on the future employment of trainees (see Card and Sullivan (1988)). As pointed out by Heckman, longitudinal data on individual histories are required in order to discriminate between true and spurious state dependence. This paper presents methods for discrete choice models with structural state dependence which allow for the presence of unobservable individual heterogeneity in panels with a large number of individuals observed through a small number of time periods.

It is well–known[1] that nonlinear panel data models with individual-specific effects, such as discrete choice, censored and truncated, sample selection models, etc., may be estimated by the "random effects" approach. See Arellano and Carrasco (1996) for an example. This approach requires the specification of the statistical relationship between the observed covariates with the unobservable permanent individual effect. Furthermore, it requires distributional assumptions on the initial conditions of the process, if there is serial correlation in the unobserved transitory error components and/or if lags of the dependent variable are used as explanatory variables. The problems associated with misspecification of these distributions are partly overcome in the "fixed effects" approach. Below we discuss existing estimators for panel data discrete choice models. Estimators for the fixed effects censored and truncated regression models with strictly exogenous regressors have been proposed by Honoré (1992). Honoré (1993) considered the case where the explanatory variable set also includes lags of the dependent variable. Kyriazidou (1997a) proposed estimators for the panel data sample selection model assuming strictly exogenous regressors in both the main equation and the binary sample selection equation. In Kyriazidou (1997b), the main equation is also allowed to contain lags of the continuous dependent variable, while the selection equation may have lags of the endogenous selection indicator.

In the absence of state dependence (that is, with only strictly exogenous regressors), the parametric "fixed effects" approach for the discrete choice model assumes that the time-varying errors are independent of all other covariates and that they are i.i.d. over time with a logistic distribution. No assumptions are made on

[1]For this and other results concerning panel data models, see the survey articles by Chamberlain (1984, 1985), Hsiao (1986) and Maddala (1983).

the distribution of the individual effects conditional on the observed explanatory variables. As Rasch (1960) and Andersen (1970) have shown, the model may then be estimated by conditional maximum likelihood. In the case of binary choice, the model has the form:

$$P(y_{it} = 1|x_i, \alpha_i, y_{i0}, \ldots, y_{i,t-1}) = \frac{\exp(x_{it}\beta + \alpha_i)}{1 + \exp(x_{it}\beta + \alpha_i)} \qquad t = 1, ...T; T \geq 2, \tag{1}$$

where β is the parameter of interest, α_i is an individual–specific effect which may depend on the exogenous explanatory variables $x_i \equiv (x_{i1}, ..., x_{iT})$ in an arbitrary way, and where y_{i0} may or may not be observed. (Throughout, $i = 1, ..., n$ indicates the identity of the individual.) In the case where $T = 2$, inference concerning β is based on the observation that $P(y_{it} = 1|\alpha_i, x_i, y_{i0}, y_{i1} + y_{i2} = 1)$ is independent of α_i.

As Manski (1987) has shown, it is possible to relax the logistic assumption as well as to allow for certain forms of serial correlation in the underlying time-varying errors in (1) above. In the special case where the errors are independent, the model takes the form:

$$P(y_{it} = 1|x_i, \alpha_i, y_{i0}, \ldots, y_{i,t-1}) = F_i\left(x_{it}\beta + \alpha_i\right) \qquad t = 1, ...T; T \geq 2 \tag{1$'$}$$

where F_i is a strictly increasing distribution function with full support on $\Re$ that is allowed to differ across individuals, but not across time for a given individual. In the case where $T = 2$, identification of β is based on the fact that, under certain regularity conditions on the distribution of the exogenous variables, $\operatorname{sgn}\left(P(y_{i2} = 1|x_{i1}, x_{i2}, \alpha_i) - P(y_{i1} = 1|x_{i1}, x_{i2}, \alpha_i)\right) = \operatorname{sgn}\left((x_{i2} - x_{i1})\beta\right)$. This implies that Manski's (1975, 1985) maximum score estimator can be applied to the first differences of the data in the sub–sample for which $y_{i1} \neq y_{i2}$.

The parametric "fixed effects" approach may be also used to estimate panel data logit models with individual effects and lags of the dependent variable, provided that there are no other explanatory variables and that there are at least four observations per individual (see Chamberlain (1985), and Magnac (1997)). In the binary choice case with the dependent variable lagged once, the model is

$$\begin{aligned} P(y_{i0} &= 1|\alpha_i) = p_0(\alpha_i) \\ P(y_{it} &= 1|\alpha_i, y_{i0}, \ldots, y_{i,t-1}) = \frac{\exp(\gamma y_{i,t-1} + \alpha_i)}{1 + \exp(\gamma y_{i,t-1} + \alpha_i)} \qquad t = 1, ...T; T \geq 3 \end{aligned} \tag{2}$$

where y_{i0} is assumed to be observed, although the model is not specified in the initial period. When $T = 3$, inference on γ is based on the observation that $P(y_{i0} = d_0, y_{i1} = 0, y_{i2} = 1, y_{i3} = d_3|y_{i1} + y_{i2} = 1, \alpha_i)$ is independent of α_i. (Here $d_0, d_3 \in \{0, 1\}$).

In this paper, we consider identification and estimation in panel data discrete choice models when the explanatory variable set includes strictly exogenous variables, lags of the endogenous dependent variable, as well as unobservable individual-specific effects that may be correlated with the observed covariates in an unspecified way. For the binary logit model with the dependent variable lagged only once, Chamberlain (1993) has shown that, if individuals are observed in three time periods, then the parameters of the model are not identified. In this paper, we demonstrate that β and γ are both identified (subject to regularity conditions) if the econometrician has access to four or more observations per individual. The identification result suggests estimators of the model. We show that these are consistent and asymptotically normal, although their rate of convergence is not the inverse of the square root of the sample size. This result is in line with recent findings by Hahn (1997) that suggest that the model cannot be estimated at the standard $n^{-1/2}$ rate.

We also consider identification in the semiparametric case where the logit assumption is relaxed. We propose an estimator in the spirit of Manski's (1987) conditional maximum score estimator. For this estimator, we only show consistency. The results by Kim and Pollard (1990) suggest that the estimator will not have a limiting normal distribution and that its rate of convergence will be slower than $n^{-1/3}$.

The paper is organized as follows. Section 2 presents our identification and estimation methods for the case where the panel contains only four observations per individual. Section 3 states the assumptions and derives the asymptotic properties for the estimators proposed in Section 2. Section 4 discusses generalizations and extensions of the estimators to the case of longer panels, to the case where the dependent variable is lagged twice, and to the multinomial choice case. Section 5 presents the results of a small Monte Carlo study investigating the small sample properties of the estimators proposed in Section 2. Section 6 concludes the paper. The proofs of the theorems are in the Appendix.

2 Identification and Motivation of the Estimators.

2.1 The Logit Case

We consider the following fixed effects logit model which combines the features of (1) and (2):

$$\begin{aligned} P(y_{i0} &= 1|x_i,\alpha_i) = p_0(x_i,\alpha_i) \\ P(y_{it} &= 1|x_i,\alpha_i,y_{i0},\ldots,y_{i,t-1}) = \frac{\exp(x_{it}\beta+\gamma y_{i,t-1}+\alpha_i)}{1+\exp(x_{it}\beta+\gamma y_{i,t-1}+\alpha_i)} \qquad t=1,...,T. \end{aligned} \tag{3}$$

where $x_i \equiv (x_{i1},...,x_{iT})$. Throughout this section $T=3$. Here, the logit specification is imposed for periods one to three. The model is left unspecified in the initial period, since the value of the dependent variable is not assumed to be known in periods prior to the sample. We assume that y_{i0} is observed, but it is not necessary to assume that the explanatory variables are observed in the initial period. It is important to note the implicit assumption that the errors in a threshold–crossing model leading to (3) are *i.i.d.* over time with logistic distributions and independent of (x_i,α_i,y_{i0}) in all time periods. While the independence over time assumption is fairly standard (it is also implicitly assumed in (1) and (2)), it is certainly a weakness of the approach taken here (as well as those taken to derive (1) and (2)).

Our identification scheme follows the intuition of the conditional logit approach.[2] The aim is to derive a

[2]Jones and Landwehr (1988) attempt to use the conditional logit approach to estimate the model considered in this paper.

set of probabilities that do not depend on the individual effect. Following Chamberlain (1985), we consider the events:

$$\begin{aligned} A &= \{y_{i0} = d_0, y_{i1} = 0, y_{i2} = 1, y_{i3} = d_3\} \\ B &= \{y_{i0} = d_0, y_{i1} = 1, y_{i2} = 0, y_{i3} = d_3\} \end{aligned}$$

where d_0 and d_3 are either 0 or 1. A straightforward calculation yields:

$$\begin{aligned} P(A|x_i,\alpha_i) &= p_0(x_i,\alpha_i)^{d_0}\left(1-p_0(x_i,\alpha_i)\right)^{1-d_0} \times \frac{1}{1+\exp(x_{i1}\beta+\gamma d_0+\alpha_i)} \\ & \quad \times \frac{\exp(x_{i2}\beta+\alpha_i)}{1+\exp(x_{i2}\beta+\alpha_i)} \times \frac{\exp(d_3 x_{i3}\beta+d_3\gamma+d_3\alpha_i)}{1+\exp(x_{i3}\beta+\gamma+\alpha_i)} \end{aligned}$$

and

$$\begin{aligned} P(B|x_i,\alpha_i) &= p_0(x_i,\alpha_i)^{d_0}\left(1-p_0(x_i,\alpha_i)\right)^{1-d_0} \times \frac{\exp(x_{i1}\beta+\gamma d_0+\alpha_i)}{1+\exp(x_{i1}\beta+\gamma d_0+\alpha_i)} \\ & \quad \times \frac{1}{1+\exp(x_{i2}\beta+\alpha_i+\gamma)} \times \frac{\exp(d_3 x_{i3}\beta+d_3\alpha_i)}{1+\exp(x_{i3}\beta+\alpha_i)} \end{aligned}$$

In general, the probabilities $P(A|x_i,\alpha_i,A\cup B)$ and $P(B|x_i,\alpha_i,A\cup B)$, which condition on the event that the dependent variable changes sign between periods one and two, will depend on α_i. This is the reason why a conditional likelihood approach will not eliminate the fixed effect. Our identification scheme rests on the observation that, if $x_{i2} = x_{i3}$, then the conditional probabilities

$$P\left(A|\, x_i,\alpha_i,A\cup B, x_{i2}=x_{i3}\right) = \frac{1}{1+\exp\left(\left(x_{i1}-x_{i2}\right)\beta+\gamma\left(d_0-d_3\right)\right)} \tag{4}$$

and

$$P\left(B|\, x_i,\alpha_i,A\cup B, x_{i2}=x_{i3}\right) = \frac{\exp\left(\left(x_{i1}-x_{i2}\right)\beta+\gamma\left(d_0-d_3\right)\right)}{1+\exp\left(\left(x_{i1}-x_{i2}\right)\beta+\gamma\left(d_0-d_3\right)\right)} \tag{5}$$

do *not* depend on α_i. This observation is the key to all the results presented in this paper. In the special case where all the explanatory variables are discrete and the x_{it} process satisfies $P(x_{i2}=x_{i3})>0$, one can use (4) to make inference about β and γ. In particular, one may estimate β and γ by maximizing the weighted likelihood function:

$$\sum_{i=1}^{n} 1\{y_{i1}+y_{i2}=1\}1\left\{x_{i2}-x_{i3}=0\right\}\ln\left(\frac{\exp((x_{i1}-x_{i2})b+g(y_{i0}-y_{i3}))^{y_{i1}}}{1+\exp((x_{i1}-x_{i2})b+g(y_{i0}-y_{i3}))}\right)$$

The resulting estimator will have all the usual properties (consistency and root-n asymptotic normality).

While inference based only on observations for which $x_{i2}=x_{i3}$ may be reasonable in some cases (in particular, experimental cases where the distribution of x_i is in the control of the researcher), there are

However, their calculation does not allow for exogenous explanatory variables.

many economic applications where it is not useful. The idea then is to replace the indicator functions $1\{x_{i2} - x_{i3} = 0\}$ in the objective function above with weights that depend inversely on the magnitude of the difference $x_{i2} - x_{i3}$, giving more weight on observations for which x_{i2} is "close" to x_{i3}. Specifically, we propose estimating β and γ by maximizing

$$\sum_{i=1}^{n} 1\{y_{i1} + y_{i2} = 1\} K\left(\frac{x_{i2} - x_{i3}}{\sigma_n}\right) \ln\left(\frac{\exp((x_{i1} - x_{i2})b + g(y_{i0} - y_{i3}))^{y_{i1}}}{1 + \exp((x_{i1} - x_{i2})b + g(y_{i0} - y_{i3}))}\right) \tag{6}$$

with respect to b and g over some compact set. Here $K(\cdot)$ is a kernel density function which gives the appropriate weight to observation i, while σ_n is a bandwidth which shrinks as n increases. The asymptotic theory will require that $K(\cdot)$ be chosen so that a number of regularity conditions, such as $K(\nu) \to 0$ as $|\nu| \to \infty$, are satisfied.

Note that the proposed estimators are maximum-likelihood-type (or extremum or M–) estimators. The key idea behind the estimation is that the limit of the objective function above (as well as of the objective function in the semiparametric case, discussed below), and which may be readily seen to be a conditional expectation given the event that $x_{i2} - x_{i3} = 0$, is uniquely maximized at the true parameter values, under appropriate assumptions. It is clear that identification of the model will require that $x_{i2} - x_{i3}$ be continuously distributed with support in a neighborhood of 0, and that $x_{i1} - x_{i2}$ have sufficient variation conditional on the event that $x_{i2} - x_{i3} = 0$.

The asymptotic properties of the estimators may be derived in a manner similar to that underlying local likelihood estimation (see, for example, Staniswalis (1989), and Tibshirani and Hastie (1987)) and robust regression function estimation (see, for example, Härdle (1984), and Härdle and Tsybakov (1988)). Section 4 states conditions under which the estimators maximizing (6) are consistent and asymptotically normal, although their rate of convergence will be slower than $n^{-1/2}$ and will depend on the number of covariates in x_{it}.

2.2 The Semiparametric Case

In this section, we will use Manski's (1987) insight to relax the logit assumption on the distribution of the time-varying errors underlying a threshold-crossing specification of the model in (3). The independence over time assumption of the previous section will be maintained. Suppose[3] that

$$\begin{aligned} P(y_{i0} &= 1|x_i, \alpha_i) = p_0(x_i, \alpha_i) \\ P(y_{it} &= 1|x_i, \alpha_i, y_{i0}, \ldots, y_{i,t-1}) = F(x_{it}\beta + \gamma y_{i,t-1} + \alpha_i), \quad t = 1, ..., T \end{aligned} \tag{7}$$

where y_{i0} is assumed to be observed and F is a strictly increasing function that has full support on $\Re$. As before, we will focus on the case where there are four observations per individual, i.e. $T = 3$. With A and B

[3]It is possible to generalize the results in the remainder of this section to allow the distribution function, F, to differ across individuals, provided that it does not differ over time for a given individual.

defined as in the previous section, we have:

$$
\begin{aligned}
P(A|x_i, \alpha_i, x_{i2} &= x_{i3}) = p_0(x_i, \alpha_i)^{d_0} \left(1 - p_0(x_i, \alpha_i)\right)^{1-d_0} \times (1 - F(x_{i1}\beta + \gamma d_0 + \alpha_i)) \\
&\times F(x_{i2}\beta + \alpha_i) \times (1 - F(x_{i2}\beta + \gamma + \alpha_i))^{(1-d_3)} \times F(x_{i2}\beta + \gamma + \alpha_i)^{d_3}
\end{aligned}
$$

and

$$
\begin{aligned}
P(B|x_i, \alpha_i, x_{i2} &= x_{i3}) = p_0(x_i, \alpha_i)^{d_0} \left(1 - p_0(x_i, \alpha_i)\right)^{1-d_0} \times F(x_{i1}\beta + \gamma d_0 + \alpha_i) \\
&\times (1 - F(x_{i2}\beta + \gamma + \alpha_i) \times (1 - F(x_{i2}\beta + \alpha_i))^{(1-d_3)} \times F(x_{i2}\beta + \alpha_i)^{d_3}
\end{aligned}
$$

If $d_3 = 0$, then

$$
\begin{aligned}
\frac{P(A|x_i, \alpha_i, x_{i2} = x_{i3})}{P(B|x_i, \alpha_i, x_{i2} = x_{i3})} &= \frac{(1 - F(x_{i1}\beta + \gamma d_0 + \alpha_i))}{(1 - F(x_{i2}\beta + \alpha_i))} \times \frac{F(x_{i2}\beta + \alpha_i)}{F(x_{i1}\beta + \gamma d_0 + \alpha_i)} \\
&= \frac{(1 - F(x_{i1}\beta + \gamma d_0 + \alpha_i))}{(1 - F(x_{i2}\beta + \gamma d_3 + \alpha_i))} \times \frac{F(x_{i2}\beta + \gamma d_3 + \alpha_i)}{F(x_{i1}\beta + \gamma d_0 + \alpha_i)}
\end{aligned}
$$

where the second equality follows from the fact that d_3 is zero. If $d_3 = 1$, then

$$
\begin{aligned}
\frac{P(A|x_i, \alpha_i, x_{i2} = x_{i3})}{P(B|x_i, \alpha_i, x_{i2} = x_{i3})} &= \frac{(1 - F(x_{i1}\beta + \gamma d_0 + \alpha_i))}{(1 - F(x_{i2}\beta + \gamma + \alpha_i))} \times \frac{F(x_{i2}\beta + \gamma + \alpha_i)}{F(x_{i1}\beta + \gamma d_0 + \alpha_i)} \\
&= \frac{(1 - F(x_{i1}\beta + \gamma d_0 + \alpha_i))}{(1 - F(x_{i2}\beta + \gamma d_3 + \alpha_i))} \times \frac{F(x_{i2}\beta + \gamma d_3 + \alpha_i)}{F(x_{i1}\beta + \gamma d_0 + \alpha_i)}
\end{aligned}
$$

where the second equality follows from the fact that $d_3 = 1$, so that $\gamma d_3 = \gamma$. In either case, the monotonicity of F implies that

$$
\operatorname{sgn}\left(P(A|x_i, \alpha_i, x_{i2} = x_{i3}) - P(B|x_i, \alpha_i, x_{i2} = x_{i3})\right) = \operatorname{sgn}\left((x_{i2} - x_{i1})\beta + \gamma(d_3 - d_0)\right) \tag{8}
$$

If $P(x_{i2} = x_{i3}) > 0$, a maximum score estimator may therefore be applied to the observations satisfying $A \cup B$ and $x_{i2} = x_{i3}$. That is, β and γ may be estimated by maximizing

$$
\sum_{i=1}^{n} 1\left\{x_{i2} - x_{i3} = 0\right\}(y_{i2} - y_{i1}) \operatorname{sgn}\left((x_{i2} - x_{i1})\, b + g\,(y_{i3} - y_{i0})\right)
$$

with respect to b and g over some compact set. It is obvious from the expression above that only observations which satisfy $y_{i1} + y_{i2} = 1$ are used in the estimation.

Similarly to the logistic case, when $x_{i2} - x_{i3}$ is continuously distributed with support in a neighborhood of 0, we propose to estimate β and γ (up to scale) by maximizing the score function

$$
\sum_{i=1}^{n} K\left(\frac{x_{i2} - x_{i3}}{\sigma_n}\right)(y_{i2} - y_{i1}) \operatorname{sgn}\left((x_{i2} - x_{i1})\, b + g\,(y_{i3} - y_{i0})\right) \tag{9}
$$

with respect to b and g over some compact set. In Section 4, we show consistency of this estimator. We do not derive its asymptotic distribution, but in view of existing results on the maximum score estimator for the cross-sectional binary response model, we expect the limiting distribution to be non–normal and the rate of convergence to be slower than $n^{-1/3}$.

In both the logistic and the semiparametric case, the main limitations of our approach are (i) the assumption that the errors in the underlying threshold–crossing model are independent over time, and (ii) the assumption that $x_{i2} - x_{i3}$ has support in a neighborhood of 0. The latter restriction rules out time–dummies. In the analysis above we have assumed that all explanatory variables are continuous. The estimators may be modified in a straightforward manner to account for discreteness of some variables in x_{it}, namely multiply the objective functions (6) and (9) by indicators that restrict the discrete regressors to be equal in periods 2 and 3.

3 Asymptotic Properties of Estimators.

In this section, we discuss the asymptotic properties of the estimators proposed in the previous section. The following notation will be useful:

$$\begin{aligned}
\theta &= (b,g)', \quad \theta_0 = (\beta,\gamma)', \\
x_{23} &= x_2 - x_3, \quad x_{12} = x_1 - x_2, \quad x_{21} = x_2 - x_1, \\
y_{03} &= y_0 - y_3, \quad y_{30} = y_3 - y_0, \quad \text{and} \quad y_{21} = y_2 - y_1.
\end{aligned}$$

3.1 The Logit Case.

We first consider the estimator defined by minimizing (6). It will be useful to define the (random) functions:

$$\begin{aligned}
h(\theta) &= 1\{y_1 \neq y_2\} \ln\left(\frac{\exp(z\theta)^{y_1}}{1+\exp(z\theta)}\right) \\
h^{(1)}(\theta) &= \frac{\partial h}{\partial \theta} = 1\{y_1 \neq y_2\}\left(y_1 - \frac{\exp(z\theta)}{1+\exp(z\theta)}\right) z' \\
h^{(2)}(\theta) &= \frac{\partial^2 h}{\partial\theta\partial\theta'} = -1\{y_1 \neq y_2\}\frac{\exp(z\theta)}{(1+\exp(z\theta))^2} z'z
\end{aligned}$$

where $z \equiv (x_{12}, y_{03})$. The following theorem gives sufficient conditions for consistency of the estimator proposed in (6). For simplicity, we will focus on the case where all the exogenous variables are continuously distributed.

Theorem 1 (Consistency) *Let the following assumptions hold:*

(C1) $\{(y_{i0}, y_{i1}, y_{i2}, y_{i3}, x_{i1}, x_{i2}, x_{i3})\}_{i=1}^n$ *is a random sample of n observations from a distribution satisfying (3).*

(C2) $\theta_0 \in \Theta$, *a compact subset of* $\Re^{k+1}$.

(C3) The random vector $x_{23} \in X \subseteq \Re^k$ is absolutely continuously distributed with density $f(\cdot)$ that is bounded from above on its support, and strictly positive and continuous in a neighborhood of zero.

(C4) The function $E[\|x_{12}\| | x_{23} = \cdot]$ is bounded on its support.

(C5) The function $E[h(\theta)| x_{23} = \cdot]$ is continuous in a neighborhood of zero for all $\theta \in \Theta$.

(C6) The function $E[x_{12}' x_{12} | x_{23} = \cdot]$ has full column rank k in a neighborhood of zero.

(C7) $K : \Re^k \to \Re$ is a function of bounded variation that satisfies: (i) $\sup_{\nu \in \Re} |K(\nu)| < \infty$, (ii) $\int |K(\nu)| d\nu < \infty$, and (iii) $\int K(\nu) d\nu = 1$.

(C8) σ_n is a sequence of positive numbers that satisfies: $\sigma_n \to 0$ as $n \to \infty$.

Let $\hat{\theta}_n \equiv \left(\hat{\beta}_n, \hat{\gamma}_n\right)$ be a sequence of solutions to the problem

$$\max_{\theta \in \Theta} \sum_i K\left(\frac{x_{i23}}{\sigma_n}\right) h_i(\theta) \tag{10}$$

Then, $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Assumptions (C7) and (C8) are standard in kernel density and regression function estimation. The strict positiveness of f and the full rank condition of Assumption (C6) are required for identification of θ_0. The rest of the assumptions are regularity conditions that permit the application of a uniform law of large numbers to show convergence of the objective function to a nonstochastic limit that is uniquely maximized at θ_0. Note that, in some cases, the boundedness condition of Assumption (C4) may be restrictive. However, it is clear from the proof of the theorem that it may be relaxed. In particular, we only need to assume that the product $f(\cdot) E[\|x_{12}\| | x_{23} = \cdot]$ is bounded on its support. The same comment applies to similar conditions in the theorems that follow. Finally, the compactness of the parameter space (Assumption (C2)) may be also relaxed if $K(\cdot) > 0$, in which case the objective function is strictly concave (see e.g., Newey and Powell (1987)).

We next present conditions that are sufficient for asymptotic normality of the proposed estimators. Apart from the usual strengthening of regularity conditions on the existence and finiteness of moments higher than those required for consistency, additional smoothness is imposed on the model which allows convergence at a faster rate.

Theorem 2 (Asymptotic Normality) *Let Assumptions (C1)-(C8) hold and $\hat{\theta}_n$ be a solution to (10). In addition assume:*

(N1) $\theta_0 \in int(\Theta)$.

(N2) $f(\cdot)$ is s $(s \geq 1)$ times continuously differentiable on its support and has bounded derivatives.

(N3) The function $E\left[h^{(1)}(\theta_0)\middle| x_{23}=\cdot\right]$ *is s times continuously differentiable on its support and has bounded derivatives.*

(N4) The function $E\left[h^{(2)}(\theta)\middle| x_{23}=\cdot\right]$ *is continuous in a neighborhood of zero for all* $\theta \in \Theta$.

(N5) The function $E\left[\|x_{12}\|^6\middle| x_{23}=\cdot\right]$ *is bounded on its support.*

(N6) The function $E\left[h^{(1)}(\theta_0)h^{(1)}(\theta_0)'\middle| x_{23}=\cdot\right]$ *is continuous in a neighborhood of zero.*

(N7) $K:\Re^k \to \Re$ *is an s'th order bias-reducing kernel that satisfies:*

(i) $\int |\nu|^i |K(\nu)|\, d\nu < \infty$ *for* $i=0$ *and* $i=s$, *where* $s \geq 1$

(ii) $\int \nu_1^{i_1}\nu_2^{i_2}...\nu_k^{i_k}K(\nu_1,\nu_2,...,\nu_k)d\nu_1 d\nu_2...d\nu_k = \begin{cases} 1 & \text{if } i_1=i_2=...=i_k=0 \\ 0 & \text{if } 0<i_1+i_2+...+i_k<s \end{cases}$

Let $\sqrt{n\sigma_n^k}\sigma_n^s \to 0$. *Then*

$$\sqrt{n\sigma_n^k}\left(\hat{\theta}_n-\theta_0\right)\tilde{\to}N\left(0,J^{-1}VJ^{-1}\right)$$

where

$$\begin{aligned} J &\equiv J(\theta_0) \equiv -f(0)\cdot E\left[h^{(2)}(\theta_0)\,|x_{23}=0\right] \\ V &\equiv V(\theta_0) \equiv f(0)\cdot E\left[h^{(1)}(\theta_0)h^{(1)}(\theta_0)'\,|x_{23}=0\right]\cdot\int K^2(\nu)\,d\nu \end{aligned}$$

Remark: Note that the rate of convergence of the proposed estimator is maximized for $\sqrt{n\sigma_n^k}\sigma_n^s \to \sigma \neq 0$. However, in this case the estimator is asymptotically biased. It may be easily shown that in this case, $\sqrt{n\sigma_n^k}\left(\hat{\theta}_n-\theta_0\right)\tilde{\to}N\left(\sigma J^{-1}B, J^{-1}VJ^{-1}\right)$, where B is a function of the s'th order derivative of $f(\cdot)\times E\left[h^{(1)}(\theta_0)\middle| x_{23}=\cdot\right]$ Although it is possible to eliminate the asymptotic bias in the manner suggested, for example, by Horowitz (1992), we have found that the asymptotic bias correction is not effective in reducing the small sample bias in our Monte Carlo experiments.

The next theorem provides consistent estimators of the two components of the asymptotic variance-covariance matrix.

Theorem 3 (Asymptotic Variance Estimation) *Let Assumptions (C1)-(C8) and (N1)-(N7) hold and* $\hat{\theta}_n$ *be a consistent estimator of* θ_0.

(i) *Define*

$$J_n(\theta) \equiv -\frac{1}{n\sigma_n^k}\sum_i K\left(\frac{x_{i23}}{\sigma_n}\right)h_i^{(2)}(\theta)$$

Then, $J_n\left(\hat{\theta}_n\right) \xrightarrow{p} J(\theta_0)$, *where* $J(\theta_0)$ *is defined in Theorem 2.*

(ii) *Define*

$$V_n(\theta) \equiv \frac{1}{n\sigma_n^k}\sum_i K^2\left(\frac{x_{i23}}{\sigma_n}\right)h_i^{(1)}(\theta)h_i^{(1)}(\theta)'$$

If $E\left[h^{(1)}(\theta)\,h^{(1)}(\theta)'\,|x_{23}=\cdot\right]$ *is continuous in a neighborhood of zero as a function of* x_{23} *for all* $\theta\in\Theta$, *then* $V_n\left(\hat{\theta}_n\right)\xrightarrow{p}V(\theta_0)$, *where* $V(\theta_0)$ *is defined in Theorem 2.*

3.2 The Semiparametric Case

In this section, we present conditions that are sufficient to identify and consistently estimate θ_0 when the logit assumption on the distribution of the underlying time-varying errors is relaxed. We define the function

$$h(\theta)=y_{21}\operatorname{sgn}(z\theta)$$

where now $z\equiv(x_{21},y_{30})$.

Theorem 4 (Identification and Consistency) *Let the following conditions hold:*

(CS1) $\{(y_{i0},y_{i1},y_{i2},y_{i3},x_{i1},x_{i2},x_{i3})\}_{i=1}^{n}$ *is a random sample of* n *observations from a distribution satisfying (7).*

(CS2) F *is strictly increasing on* $\Re$ *for almost all* (x_i,α_i).

(CS3) There exists at least one $\kappa\in\{1,...,k\}$, *such that* $\beta_\kappa\neq 0$, *and such that, for almost all* $\tilde{x}_{21}\equiv(x_{21,1},...,x_{21,\kappa-1},x_{21,\kappa+1},...,x_{21,k})$, *the random variable* $x_{21,\kappa}$ *has everywhere positive Lebesgue density conditional on* $\tilde{x}_{21}$ *and conditional on* x_{23} *in a neighborhood of* x_{23} *near zero.*

(CS4) The support of x_{21} *conditional on* x_{23} *in a neighborhood of* x_{23} *near zero is not contained in any proper linear subspace of* $\Re^k$.

(CS5) The random vector $x_{23}\in X\subseteq\Re^k$ *is absolutely continuously distributed with density* $f(\cdot)$ *that is bounded from above on its support and strictly positive in a neighborhood of zero.*

(CS6) For all $\theta\in\Theta$, $f(\cdot)$ *and* $E[h(\theta)|\,x_{23}=\cdot]$ *are continuously differentiable on their support with bounded first-order derivatives.*

(CS7) $K:\Re^k\to\Re$ *is a function of bounded variation that satisfies: (i)* $\sup_{\nu\in\Re}|K(\nu)|<\infty$, *(ii)* $\int|K(\nu)|\,d\nu<\infty$, *and (iii)* $\int K(\nu)\,d\nu=1$.

(CS8) σ_n *is a sequence of positive numbers that satisfies: (i)* $\sigma_n\to 0$ *as* $n\to\infty$, *and (ii)* $n\sigma_n^k/\ln n\to\infty$ *as* $n\to\infty$.

Then:

(i) θ_0 *is identified (up to scale) relative to all* $\theta\in\Re^{k+1}$ *such that* $\theta/\|\theta\|\neq\theta_0/\|\theta_0\|\equiv\theta_0^*$.

(ii) *Let* $\hat{\theta}_n \equiv \left(\hat{\beta}_n, \hat{\gamma}_n\right)$ *be a sequence of solutions to the problem*

$$\max_{\theta\in\Theta} \sum_i K\left(\frac{x_{i23}}{\sigma_n}\right) h_i(\theta) \tag{11}$$

where $\Theta \equiv \left\{\theta \in \Re^{k+1} : \|\theta\| = 1 \cap |b_\kappa| \geq \eta\right\}$ *and* η *is a known positive constant such that* $|\beta_\kappa| / \|\theta_0\| \geq \eta$*. Then,* $\hat{\theta}_n \xrightarrow{p} \theta_0^*$*.*

Assumptions (CS1)-(CS4) are analogous to Manski's (1987) Assumptions 1 and 2. The rest of the assumptions of the theorem are similar to the assumptions in the logit case. Here, however, we strengthen the continuity assumption on $f(\cdot)$ and $E[h(\theta)|\, x_{23} = \cdot]$ to first order differentiability, in order to show uniform convergence of the objective function to its population analog. This is a consequence of the fact that, in the semiparametric case, the summands in the objective function are not continuous in the parameter, although they are uniformly bounded.

4 Extensions.

4.1 Identification with more than four observations per individual

The identification and estimation approach described in Section 2 extends to the case of longer panels. Note that in the case where $T = 4$, the idea behind the identification for the logit model (3) follows Chamberlain's (1985) intuition, namely that conditional on a switch between periods 1 and 2 (and in our case, conditional also on $x_{i2} = x_{i3}$), the probability of a sequence of choices (a "string") does not depend on the individual effect. For general T, identification in the dynamic logit model (2), which has only one lag of the dependent variable and no other explanatory variables, relies on the fact that conditional on the initial and the last observation, and conditional on $\sum_t y_{it}$, the probability of a string is independent of the individual effect. To investigate whether the same intuition holds in the general T case for the model that also contains exogenous variables, we consider $T = 5$. In this case, it is possible to show that the same statement holds, namely that conditional on the initial and the last observation, and conditional on $\sum_t y_{it}$, the probability of a string is independent of the individual effect, provided that we also condition on $x_{i2} = x_{i3} = x_{i4}$. This implies that the rate of convergence would be $\sqrt{n\sigma_n^{2k}}$, i.e. slower than the rate we obtain for $T = 4$. However, as will become clear below, it is possible to retain the same rate of convergence as in the $T = 4$ case, namely $\sqrt{n\sigma_n^k}$, if we instead use a pairwise approach that is based on considering all possible pairs of observations in a string that display a switch in the sign of the dependent variable.

Suppose that individuals are observed for $T + 1$ periods, where $T \geq 3$. In either the logistic or the semiparametric case, identification is based on sequences for which $y_{it} + y_{is} = 1$ for some $1 \leq t < s \leq T - 1$. Consider the event

$$A = \{y_{i0} = d_0, ..., y_{it-1} = d_{t-1}, y_{it} = 0, y_{it+1} = d_{t+1}, ..., y_{is-1} = d_{s-1}, y_{is} = 1, y_{is+1} = d_{s+1}, ..., y_{iT} = d_T\}$$

and its counterpart,

$$B = \{y_{i0} = d_0, ..., y_{it-1} = d_{t-1}, y_{it} = 1, y_{it+1} = d_{t+1}, ..., y_{is-1} = d_{s-1}, y_{is} = 0, y_{is+1} = d_{s+1}, ..., y_{iT} = d_T\}$$

It is not difficult to show that for the logit model (3),

$$\begin{aligned} \Pr(B|\, x_i, \alpha_i, A \cup B, x_{it+1} &= x_{is+1}) \\ &= \frac{\exp\left((x_{it} - x_{is})\,\beta + \gamma\,(d_{t-1} - d_{s+1}) + \gamma\,(d_{t+1} - d_{s-1})\,1\,\{s - t \geq 3\}\right)}{1 + \exp\left((x_{it} - x_{is})\,\beta + \gamma\,(d_{t-1} - d_{s+1}) + \gamma\,(d_{t+1} - d_{s-1})\,1\,\{s - t \geq 3\}\right)} \end{aligned}$$

which does *not* depend on α_i. This suggests estimating β and γ by maximizing

$$\sum_{i=1}^{n}\left(\left(\sum_{1\leq t<s\leq T-1} 1\,\{y_{it} + y_{is} = 1\}\,K\left(\frac{x_{it+1} - x_{is+1}}{\sigma_n}\right) \times \right.\right. \\ \left.\left. \ln\left(\frac{\exp\left((x_{it} - x_{is})\,b + g\,(y_{it-1} - y_{is+1}) + g\,(y_{it+1} - y_{is-1})\,1\,\{s - t \geq 3\}\right)^{y_{it}}}{1 + \exp\left((x_{it} - x_{is})\,b + g\,(y_{it-1} - y_{is+1}) + g\,(y_{it+1} - y_{is-1})\,1\,\{s - t \geq 3\}\right)}\right)\right)\right)$$

For the semiparametric model (7), we consider two cases depending on whether periods t and s are adjacent or not. For $s = t + 1$, it is easily verified that whether $d_{t+2} = 0$ or $d_{t+2} = 1$,

$$\frac{\Pr\left(A|\, x_i, \alpha_i, x_{it+1} = x_{it+2}\right)}{\Pr\left(B|\, x_i, \alpha_i, x_{it+1} = x_{it+2}\right)} = \frac{1 - F\left(x_{it}\beta + \alpha_i + \gamma d_{t-1}\right)}{1 - F\left(x_{it+1}\beta + \alpha_i + \gamma d_{t+2}\right)}\frac{F\left(x_{it+1}\beta + \alpha_i + \gamma d_{t+2}\right)}{F\left(x_{it}\beta + \alpha_i + \gamma d_{t-1}\right)}$$

and therefore,

$$\begin{aligned} & \operatorname{sgn}\left\{\Pr\left(A|\, x_i, \alpha_i, x_{it+1} = x_{it+2}\right) - \Pr\left(B|\, x_i, \alpha_i, x_{it+1} = x_{it+2}\right)\right\} \\ = \; & \operatorname{sgn}\left\{(x_{it+1} - x_{it})\,\beta + \gamma\,(d_{t+2} - d_{t-1})\right\} \end{aligned}$$

If t and s are not adjacent (so $s > t + 1$), then

$$\frac{\Pr\left(A|\, x_i, \alpha_i, x_{it+1} = x_{is+1}, y_{it+1} = y_{is+1}\right)}{\Pr\left(B|\, x_i, \alpha_i, x_{it+1} = x_{is+1}, y_{it+1} = y_{is+1}\right)} = \frac{1 - F\left(x_{it}\beta + \alpha_i + \gamma d_{t-1}\right)}{1 - F\left(x_{is}\beta + \alpha_i + \gamma d_{s-1}\right)}\frac{F\left(x_{is}\beta + \alpha_i + \gamma d_{s-1}\right)}{F\left(x_{it}\beta + \alpha_i + \gamma d_{t-1}\right)}$$

which implies that

$$\begin{aligned} & \operatorname{sgn}\left\{\Pr\left(A|\, x_i, \alpha_i, x_{it+1} = x_{is+1}, y_{it+1} = y_{is+1}\right) - \Pr\left(B|\, x_i, \alpha_i, x_{it+1} = x_{is+1}, y_{it+1} = y_{is+1}\right)\right\} \\ = \; & \operatorname{sgn}\left\{(x_{is} - x_{it})\,\beta + \gamma\,(d_{s-1} - d_{t-1})\right\} \end{aligned}$$

This suggests estimating β and γ by maximizing:

$$\sum_{i=1}^{n}\left(\sum_{t=1}^{T-3} 1\,\{y_{it} + y_{it+1} = 1\}\,K\left(\frac{x_{it+1} - x_{it+2}}{\sigma_n}\right)\operatorname{sgn}\left(y_{it+1} - y_{it}\right)\operatorname{sgn}\left((x_{it+1} - x_{it})\,b + g\,(y_{it+2} - y_{it-1})\right)\right. \\ \left. + \sum_{t=1}^{T-2}\sum_{s=t+2}^{T-1} 1\,\{y_{it} + y_{is} = 1\}\,1\,\{y_{it+1} = y_{is+1}\}\,K\left(\frac{x_{it+1} - x_{is+1}}{\sigma_n}\right)\operatorname{sgn}\left(y_{is} - y_{it}\right)\operatorname{sgn}\left((x_{is} - x_{it})\,b + g\,(y_{is-1} - y_{it-1})\right)\right)$$

It is interesting that although in general time–dummies are ruled out in eiter the logisitic or the semiparametric case, it is possible to allow for seasonal effects in the case of quarterly data and at least seven observations per individual (i.e. $T = 6$).

4.2 Identification with more than one lag of the dependent variable

As noted by Chamberlain (1985), for the logit model without exogenous regressors, it is possible to test for the presence of the endogenous dependent variable lagged twice, when there are at least six observations per individual.[4] The same is true in the presence of exogenous variables for the model:

$$
\begin{aligned}
P(y_{i0} &= 1|x_i, \alpha_i) = p_0(x_i, \alpha_i) \\
P(y_{i1} &= 1|x_i, \alpha_i, y_{i0}) = p_1(x_i, \alpha_i, y_{i0}) \\
P(y_{it} &= 1|x_i, \alpha_i, y_{i0}, \ldots, y_{i,t-1}) = \frac{\exp\left(x_{it}\beta + \gamma_1 y_{i,t-1} + \gamma_2 y_{i,t-2} + \alpha_i\right)}{1 + \exp\left(x_{it}\beta + \gamma_1 y_{i,t-1} + \gamma_2 y_{i,t-2} + \alpha_i\right)} \qquad t = 2, \ldots T; T \geq 5
\end{aligned}
$$

where $x_i \equiv (x_{i2}, ..., x_{iT})$. It is assumed that (y_{i0}, y_{i1}) is observed although the model is not specified for these time periods. For $T = 5$, inference on β and γ_2 can be based on pairs of sequences (strings) that satisfy $\{x_{i3} = x_{i4} = x_{i5}, y_{i2} + y_{i3} = 1\}$ and either $\{y_{i0} \neq y_{i1}, y_{i1} = y_{i4} = y_{i5}\}$ or $\{y_{i4} \neq y_{i5}, y_{i0} = y_{i1} = y_{i4}\}$ (there are four such pairs of strings). Consider, for example, the pair $A = (1, 0, 0, 1, 0, 0)$ and $B = (1, 0, 1, 0, 0, 0)$. It is straightforward to establish that

$$
\Pr\left(A|\, x_i, \alpha_i, A \cup B, x_{i3} = x_{i4} = x_{i5}\right) = \frac{1}{1 + \exp\left(\gamma_2 + (x_{i2} - x_{i3})\,\beta\right)}
$$

and

$$
\Pr\left(B|\, x_i, \alpha_i, A \cup B, x_{i3} = x_{i4} = x_{i5}\right) = \frac{\exp\left(\gamma_2 + (x_{i2} - x_{i3})\,\beta\right)}{1 + \exp\left(\gamma_2 + (x_{i2} - x_{i3})\,\beta\right)}
$$

which does not depend on α_i.

For the semiparametric case, where

$$
P\left(y_{it} = 1|\alpha_i, x_i, y_{i0}, ..., y_{i,t-1}\right) = F\left(x_{it}\beta + \gamma_1 y_{i,t-1} + \gamma_2 y_{i,t-2} + \alpha_i\right) \qquad t = 2, ...T; T \geq 5
$$

and where y_{i0} and y_{i1} are given as above, it is easy to establish that for $T = 5$,

$$
\frac{\Pr\left(A|\, x_i, \alpha_i, x_{i3} = x_{i4} = x_{i5}\right)}{\Pr\left(B|\, x_i, \alpha_i, x_{i3} = x_{i4} = x_{i5}\right)} = \frac{F\left(x_{i2}\beta + \gamma_2 + \alpha_i\right)}{F\left(x_{i3}\beta + \alpha_i\right)} \frac{1 - F\left(x_{i3}\beta + \alpha_i\right)}{1 - F\left(x_{i2}\beta + \gamma_2 + \alpha_i\right)}
$$

Thus,

$$
\operatorname{sgn}\left(\Pr\left(A|\, x_i, \alpha_i, x_{i3} = x_{i4} = x_{i5}\right) - \Pr\left(A|\, x_i, \alpha_i, x_{i3} = x_{i4} = x_{i5}\right)\right) = \operatorname{sgn}\left((x_{i2} - x_{i3})\,\beta + \gamma_2\right)
$$

The analysis above suggests estimators of β and γ_2 analogous to (6) and (9) provided that $(x_{i3} - x_{i4},\ x_{i4} - x_{i5})$ has support in a neighborhood of $(0, 0)$. It may be possible to generalize the preceding identification results for general T and for an arbitrary (but finite) number of included lags. Magnac (1997) provides such results for the dynamic logit model when no exogenous covariates are present.

[4] As noted by Chamberlain, it is possible in the dynamic logit model with two lags of the dependent variable and without exogenous regressors to allow the coefficient of the first lag to be individual-varying. The same observation applies to the model considered in this section, in both the logistic and in the semiparametric case. We are grateful to one of the referees for pointing this out to us.

4.3 Identification in multinomial logit models

We next consider the case where the individual chooses among M alternatives. The model is:

$$\begin{aligned}\Pr\left(y_{i0}=m|\,x_i,\alpha_i\right) &= p_{mi0}\left(x_i,\alpha_i\right)\\ \Pr\left(y_{it}=m|\,x_i,\alpha_i,y_{it-1}=j\right) &= \frac{\exp\left(x_{mit}\beta_m+\alpha_{mi}+\gamma_{jm}\right)}{\sum\limits_{h=1}^{M}\exp\left(x_{hit}\beta_h+\alpha_{hi}+\gamma_{jh}\right)} \qquad t=1,...,T;T\geq 3\end{aligned}$$

where $\alpha_i\equiv\{\alpha_{mi}\}_{m=1}^{M}$ and $x_i\equiv\left\{\{x_{mit}\}_{m=1}^{M}\right\}_{t=1}^{T}$. The model above is obtained if we assume that the underlying errors in the well known random utility maximization framework are independent across alternatives and over time conditional on (x_i,α_i,y_{i0}), and identically distributed according to the Type I extreme value distribution, and hence independent of (x_i,α_i,y_{i0}). Note that we now model individual heterogeneity as depending also on the choice, i.e. each individual has a specific attitude toward each alternative, α_{ji}, where $j\in\{1,...,M\}$. Furthermore, the coefficient γ on the lagged endogenous variable is now allowed to depend upon both the past choice and the current choice, so that there are in total M^2 feedback parameters. Thus, γ_{jm} is the feedback effect when a choice of alternative j at $t-1$ is followed by choice m at time t, where $j,m\in\{1,...,M\}$.

For the model above, identification of $\{\beta_m\}_{m=1}^{M}$ and $\left\{\gamma_{jm}\right\}_{j,m=1}^{M}$ is based on sequences of choices where the individual switches between alternatives at least once during the periods 1 through $T-1$. Out of all possible M^{T+1} sequences of choices among the M alternatives in the $T+1$ periods, there are $\left(M^{T+1}-M^3\right)$ such sequences. However, similar to the dynamic multinomial logit model without time-varying exogenous regressors (see Magnac, 1997), only $\left(M^2-(2M-1)\right)$ feedback parameters γ are identified.

Consider the events:

$$A=\{y_{i0}=d_0,...,y_{it-1}=j,y_{it}=m,y_{it+1}=q,...,y_{is-1}=p,y_{is}=\ell,y_{is+1}=r,...,y_{iT}=d_T\}$$

and

$$B=\{y_{i0}=d_0,...,y_{it-1}=j,y_{it}=\ell,y_{it+1}=q,...,y_{is-1}=p,y_{is}=m,y_{is+1}=r,...,y_{iT}=d_T\}$$

where $1\leq t<s\leq T-1$, and $j,m,q,p,\ell,r,d_0,d_T\in\{1,...,M\}$ with $m\neq\ell$. It is possible to verify that, if $x_{mit+1}=x_{mis+1}$ for all $m\in\{1,...,M\}$, then,

$$\begin{aligned}&\Pr\left(B|\,x_i,\alpha_i,A\cup B,\{x_{mit+1}=x_{mis+1}\}_{m=1}^{M}\right)\\ =\;&\frac{\exp\left((x_{mit}-x_{mis})\beta_m+(x_{\ell is}-x_{\ell it})\beta_\ell+\left(\gamma_{jm}+\gamma_{mq}+\gamma_{p\ell}+\gamma_{\ell r}\right)-\left(\gamma_{j\ell}+\gamma_{\ell q}+\gamma_{pm}+\gamma_{mr}\right)\right)}{1+\exp\left((x_{mit}-x_{mis})\beta_m+(x_{\ell is}-x_{\ell it})\beta_\ell+\left(\gamma_{jm}+\gamma_{mq}+\gamma_{p\ell}+\gamma_{\ell r}\right)-\left(\gamma_{j\ell}+\gamma_{\ell q}+\gamma_{pm}+\gamma_{mr}\right)\right)}\end{aligned}$$

Defining the binary variables $y_{hit}=1$ if alternative $h\in\{1,...,M\}$ is chosen in period t and 0 otherwise, estimation may be based on maximization of

$$\sum_{i=1}^{n}\sum_{1\leq t<s\leq T-1}\sum_{m\neq\ell}1\left\{y_{mit}+y_{\ell is}=1\right\}K\left(\frac{x_{it+1}-x_{is+1}}{\sigma_n}\right)\times$$

$$\ln \frac{\exp\Big((x_{mit}-x_{mis})\beta_m+(x_{\ell is}-x_{\ell it})\beta_\ell+\gamma_{y_{it-1},m}+\gamma_{m,y_{it+1}}+\gamma_{y_{is-1},\ell}+\gamma_{\ell,y_{is+1}}-\gamma_{y_{it-1},\ell}-\gamma_{\ell y_{it+1}}-\gamma_{y_{is-1},m}-\gamma_{my_{is+1}}\Big)^{y_{mit}}}{1+\exp\Big((x_{mit}-x_{mis})\beta_m+(x_{\ell is}-x_{\ell it})\beta_\ell+\gamma_{y_{it-1},m}+\gamma_{m,y_{it+1}}+\gamma_{y_{is-1},\ell}+\gamma_{\ell,y_{is+1}}-\gamma_{y_{it-1},\ell}-\gamma_{\ell y_{it+1}}-\gamma_{y_{is-1},m}-\gamma_{my_{is+1}}\Big)}$$

where $x_{it} \equiv (x_{1it}, ..., x_{Mit})$ and where the necessary $2M-1$ restrictions on the γ's have been imposed (for example $\gamma_{jj} = \gamma_{1j} = 0$ for all $j = 1, ..., M$).

5 Some Monte Carlo Evidence.

In this section, we summarize the main results from a small Monte Carlo experiment designed to illustrate the finite sample properties of the estimators defined in section 2. All the results presented in this section are based on 1000 replications of the model:

$$\begin{aligned} y_{i0} &= 1\{x_{it}\beta + \alpha_i + \varepsilon_{i0} \geq 0\} \\ y_{it} &= 1\{x_{it}\beta + \gamma y_{i,t-1} + \alpha_i + \varepsilon_{it} \geq 0\} \quad t = 1, ..., T-1. \end{aligned}$$

We consider several different designs that differ on the length of the panel, the relative magnitude of β and γ, on the number of variables in x_{it} and on the data generating process for x_{it}. In all designs ε_{it} is i.i.d. logistically distributed over time. The benchmark design has $T = 4$, $\gamma = 0.5$, $\beta = 1$, and only one exogenous variable x_{it} which is i.i.d. over time with distribution $N(0, \pi^2/3)$. This makes the variance of $x_{it}\beta$ equal to the variance of ε_{it}. The fixed effect is generated as $\alpha_i = (x_{i0} + x_{i1} + x_{i2} + x_{i3})/4$. Recall that the estimators proposed in Section 2 use only the observations for which $y_{i1} \neq y_{i2}$. For this design, the "effective" sample size is reduced to about 37%. We consider sample sizes of 250, 500, 1000, 2000 and 4000.

We first focus on the finite sample performance of the estimator proposed for the logit case, which maximizes (6). Given that the distribution of the underlying errors is correctly specified, this estimator is consistent and asymptotically normal. To implement the estimator, one must choose a kernel as well as a bandwidth. All the results presented in this section use a normal kernel. This means that s in Theorem 2 is 2. Since there is only one regressor in the benchmark design, the rate of convergence is maximized by setting $\sigma_n = c \cdot n^{-1/5}$, for some constant c. However, in this case the estiamator is asymptotically biased (see the Remark following Theorem 2). In Table 1, we present results for c = 1, 2, 4, 8, 16, 32 and 64. For each bandwidth and for each sample size, we present the mean bias and the root mean squared error (RMSE) of the estimator. Since these measures can be sensitive to outliers, we also present the median bias and the median absolute error (MAE) of the estimator. In what follows, we focus on these robust measures of bias and precision.

By the Remark following Theorem 2, the estimator converges at rate $n^{-2/5}$. The results in Table 1 suggest that for this design, the theoretical rate of convergence gives a fairly good approximation to the finite sample behavior of the estimator. For example, when we regress the logarithm of the median absolute error of $\widehat{\gamma}_n$ on the logarithm of sample size (and allowing for a bandwidth–specific intercept), we get a coefficient of -0.47, which is fairly close to the predicted value of -0.4. If we exclude the samples of size 250, then the coefficient is -0.41.

When $\gamma = 0$ or $\beta = 0$, the estimator defined by minimization of (6) is consistent and root–n asymptotically normal if the bandwidth is fixed, i.e., does not shrink to 0. The reason for this is that, when $\gamma = 0$ or $\beta = 0$, the terms in (6) are the same terms that would enter into a correctly specified conditional likelihood function.[5] It therefore seems likely that the small–sample performance of the estimator depends on the relative magnitudes of γ and β. In particular, if γ is small, one would expect the estimator to perform well even if the bandwidth is large. In order to investigate the sensitivity of our results to the value of γ, we consider designs with γ equal to 0.25, 1.0 and 2.0. For these designs, the effect of the lagged dependent variable on the first order serial correlation of y_{it} ranges from being smaller, to being greater than the effect of the permanent error component.[6] The results for the alternative values of γ are given in Tables 2 (to conserve space, we present only the results for $\sigma_n = 8 \cdot n^{-1/5}$). Our findings confirm that as γ increases, the bias increases dramatically.[7] Intuitively, we would expect that longer panels would dramatically improve the performance of the estimator for γ. Table 2 therefore also presents results for $T = 8$. Since the same bandwidth is used for the two sample sizes, it is not surprising that the bias is of approximately the same magnitude, but the median absolute errors do decrease dramatically. Interestingly, the gains seem to be about the same for both $\hat{\beta}$ and $\hat{\gamma}$.

It is difficult to interpret Monte Carlo results like the ones presented here without a comparison to competing estimators. Because there is no other consistent (at $n \to \infty$) estimator for the dynamic panel data logit model considered here, we will compare our estimator to the maximum likelihood estimator that estimates all the fixed effects. This estimator is consistent as $T \to \infty$ with n fixed, which suggests that its behavior will depend crucially on the number of time–periods. We therefore report results for $T = 4$, 8, and 16 (to conserve space, we consider only $n = 250$) . As this estimator will be inconsistent (as $n \to \infty$) but converge at rate $n^{-1/2}$ to its probability limit, we expect this estimator to have larger bias but less variability than the estimator proposed here. Columns three through six of Table 3 give the results for this estimator

[5]This observation also suggests that it is possible to test for true state dependence by considering (6) with a fixed bandwidth. Investigation of such a test, which would be in the same spirit as the test proposed by Heckman (1978), is left for future research.

[6]Specifically, if there is no fixed effect and no exogenous explanatory variables in the model, then the first order serial correlation of y_{it}, due entirely to the presence of the lagged dependent variable, is approximately 0.06, 0.12, 0.23 and 0.41, for γ equal to 0.25, 0.5, 1.0 and 2.0, respectively. If only the fixed effect is present in the model, then the first order serial correlation is 0.15. Including the term $x_{it}\beta$, and when there are no individual effects, the corresponding calculations yield first order serial correlation of 0.04, 0.07, 0.15 and 0.29 for the four values of γ. In the absence of the lagged dependent variable, and with the inclusion of the exogenous variable, the first order serial correlation resulting from the fixed effect is 0.16.

[7]In simulations not reported here, we found that this is especially true for large bandwidths. This suggests that it might be important to let the bandwidth be data-dependent.

for the four values of γ considered earlier, whereas the results for the estimator proposed in this paper (with $\sigma_n = 8 \cdot n^{-1/5}$) are presented in columns seven through ten. The most striking feature of these results is that the maximum likelihood estimator that estimates all the fixed effects is inferior in terms of median absolute error to the estimator proposed here for all the values of γ and T, although the results for $\widehat{\gamma}$ are close when $\gamma = 2$ and $T = 16$.

It is not too surprising that our estimator performs better than an estimator which is inconsistent. We therefore also compare the estimator to the infeasible maximum likelihood that uses the fixed effect as one of the explanatory variables (treating its coefficient as an unknown parameter to be estimated). As expected, this estimator performs better than the one proposed here, with the relative performance of our estimator being worse when γ is larger and when T is smaller.

It is well understood that the design of the regressors in Monte Carlo studies may have a large effect on the results. For example, normally distributed regressors often make estimators look better than they are for other distributions of the regressors. In order to investigate this issue, we modify the benchmark design by changing the distribution of x_{it} to a $\chi^2(1)$ random variable, normalized to have the same mean and variance as the regressor in the benchmark design. See the upper left hand corner of Table 4. The results are quite similar to those presented in Table 1. To conserve space, we present only the results for $\sigma_n = 8 \cdot n^{-1/5}$. Another problem with the choice of design is that the rate of convergence of the proposed estimators decreases as the number of regressors increases. In order to obtain information about the finite sample behavior of the estimator when there are more explanatory variables, we add three regressors to the benchmark design. All three are generated as $N\left(0, \pi^2/3\right)$ independently of each other and of all other variables. The true values of the coefficients on the additional regressors are all zero, i.e. $\beta_2 = \beta_3 = \beta_4 = 0$, so the data–generating process is the same as for the benchmark case and the only difference is that three additional regressors are used in the estimation. The results from this experiment are given in the upper right hand corner of Table 4. To conserve space we report only the results for β_1 and for γ and for one of the bandwidth sequences (note that the rate at which the bandwidths decrease is slower as a result of the additional regressors). The results suggest that the cost of adding the additional parameters is not high in terms of the median absolute error of the estimator of the two original parameters, β_1 and γ.

Since γ is essentially identified from the time series behavior of y_{it}, one might worry that the i.i.d. design of the regressors biases the results in favor of our estimator. To investigate this, we change the benchmark design to allow for serial correlation and a time–trend in x. Specifically we generate x_{it} as

$$x_{it} = c \cdot (\zeta_{it} + a + 0.1 \cdot t)$$

where ζ_{it} is an AR(1) with standard normal innovations and autoregressive coefficient equal to 0.5. a is chosen so that $a + 0.1 \cdot t$ averages 0 over the sample, and c is chosen so that the variance of the marginal distribution of x_{it} is the same as in the benchmark design. This design was chosen because it is very close to that used in Heckman (1981b). The fixed effect is constructed as the average of the first four realizations of $c \cdot \zeta_{it}$. The results of this experiment for $T = 4$ and $T = 8$ are given in the lower part of Table 4. A comparison between the results in Table 4 and the corresponding results in Table 2 suggests that the i.i.d. property of the explanatory variable in the benchmark design is favorable to the estimator proposed here, although the difference is surprisingly small. As one might expect, the difference is more pronounced when T is larger, presumably because the time–trend in the regressor is more important with many time–periods.

The asymptotic normality of the estimator may be used to conduct asymptotically valid inference. To investigate how well the asymptotic results approximate the finite sample properties of the resulting confidence intervals, we calculate 80% and 95% confidence intervals for each of the two parameters of the benchmark design. In Table 5, we present the percentage of times that these confidence intervals cover the true parameter values. The results suggest that for the benchmark design, the asymptotics provide a fairly good approximation of the finite sample distribution of the estimator (and the estimator of its variance), although it seems that the confidence intervals have coverage probabilities that are slightly smaller than the asymptotic theory would predict. This is consistent with the fact that the confidence intervals are not centered correctly (due to the asymptotic bias). We expect this problem to be more severe if γ is big, and the sample size and bandwidth are both large, because in this case the bias will be more important relative to the variance of the estimator. Indeed, in additional experiments (not reported in Table 5), we found that, with 4000 observations, $\gamma = 1$, and $\sigma_n = 64 \cdot n^{-1/5}$, the coverage probabilities for the 80% and 95% confidence intervals were approximately 47% and 74% for β and 57 and 80% for γ, respectively.

We finally turn to examine the small sample properties of the maximum score estimator obtained by maximizing (9). In Table 6, we present the results for the benchmark design. Because the assumptions maintained for the maximum score estimator do not identify the scale of the parameter vector, we consider γ/β as the parameter of interest.[8] A priori, we expect the maximum score estimator to perform worse than the estimator that imposes the logit assumption. We also expect the rate of convergence to be slower than that of the logit estimator, because, for the panel data binary response model with only exogenous regressors, the maximum score estimator is known to converge at rate $n^{1/3}$, as opposed to the $n^{-1/2}$ rate of convergence of the logit maximum likelihood estimator. The results in Table 6 confirm this. The median absolute error of the maximum score estimator is larger than that of the logit estimator in all cases, and the relative difference is larger for larger sample sizes.

---

[8] The objective function for the maximum score estimator is not differentiable, and is therefore potentially difficult to maximize. We calculate the estimator by performing a grid search over 600 equally spaced points on the unit circle. If more that one point achieves the maximum value of the objective function, then the estimate is calculated as the average value of those different point estimates of γ/β.

6 Concluding Remarks.

In this paper, we consider discrete choice models that allow for both unobserved individual heterogeneity (Heckman's "spurius" state dependence) and "true" state dependence. We show that it is possible to identify such models within the logit framework, and we propose an estimator which is consistent and asymptotically normal, although the rate of convergence is slower than the usual inverse of the square root of the sample size. The results of a small Monte Carlo study suggest that the estimator performs well (for the very simple designs considered), and that the asymptotics provide a reasonable approximation to the finite sample behavior of the estimator. The paper also proposes an estimator of the semiparametric version of the model. That estimator, which is an adaptation of Manski's maximum score estimator, is consistent, but we do not derive its asymptotic distribution. In future research, we plan to investigate whether it is possible to obtain an asymptotically normal distribution (under a stronger set of regularity conditions), by applying Horowitz's (1992) approach to smoothen the objective function (9) above. This suggests estimating β and γ by maximizing

$$\sum_{i=1}^{n} K\left(\frac{x_{i2}-x_{i3}}{\sigma_n}\right)(y_{i2}-y_{i1})\,L\left(\frac{(x_{i2}-x_{i1})\,b+g\,(y_{i3}-y_{i0})}{h_n}\right)$$

where L is a kernel function that satisfies: $L(\nu) \overset{v\to-\infty}{\longrightarrow} 0$ and $L(\nu) \overset{v\to\infty}{\longrightarrow} 1$, and h_n is another bandwidth sequence that tends to 0 as n increases.

7 References.

Amemiya, T. (1985): *Advanced Econometrics*. Harvard University Press.

Andersen, E. (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society*, Series B, 32, pp. 283-301.

Arrellano, M. and R. Carrasco (1996): "Binary Choice Panel Data Models with Predetermined Variables," unpublished manuscript, CEMFI, Madrid.

Card, D. and D. Sullivan (1988): "Measuring the Effects of Subsidized Training Programs on Movements In and Out of Employment," *Econometrica*, 56, pp. 497-530.

Chamberlain, G. (1984): "Panel Data," in *Handbook of Econometrics*, Vol. II, edited by Z. Griliches and M. Intriligator. Amsterdam: North Holland.

Chamberlain, G. (1985): "Heterogeneity, Omitted Variable Bias, and Duration Dependence," in *Longitudinal Analysis of Labor Market Data*, edited by J.J. Heckman and B. Singer. Cambridge University Press.

Chamberlain, G. (1993): "Feedback in Panel Data Models," unpublished manuscript, Department of Economics, Harvard University. (April 1993)

Chung, K. L. (1974): *A Course in Probability Theory*. Academic Press.

Hahn, J. (1997): "Information Bound of the Dynamic Panel Logit Model with Fixed Effects," unpublished manuscript, Department of Economics, University of Pennsylvania.

Härdle, W. (1984): "Robust Regression Function Estimation," *Journal of Multivariate Analysis*, pp. 169-180.

Härdle, W. (1990): *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Härdle, W. and A. B. Tsybakov (1988): "Robust Nonparametric Regression with Simultaneous Scale Curve Estimation," *The Annals of Statistics*, pp. 120-135.

Heckman, J.J. (1978): "Simple Statistical Models for Discrete Panel Data Developed and Applied to Tests of the Hypothesis of True State Dependence against the Hypothesis of Spurious State Dependence," *Annales de l'INSEE*, 30–31, pp. 227–269.

Heckman, J. J. (1981a): "Statistical Models for Discrete Panel Data," in *Structural Analysis of Discrete Panel Data with Econometric Applications*, edited by C. F. Manski and D. McFadden.

Heckman, J.J. (1981b): "Heterogeneity and State Dependence," in *Studies of Labor Markets*, edited by S. Rosen. The National Bureau of Economic Research. The University of Chicago Press.

Honoré, B. E. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 60, pp. 533-565.

Honoré, B. E. (1993): "Orthogonality Conditions for Tobit Models with Fixed Effects and Lagged Dependent Variables," *Journal of Econometrics*, 59, pp. 35-61.

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

Jones, J. M. and T. J. Landwehr (1988): "Removing Heterogeneity Bias from Logit Model Estimation," *Marketing Science*, 7, pp. 41-59.

Kim, J. and D. Pollard (1990): "Cube Root Asymptotics," *Annals of Statistics*, 18, pp. 191-219.

Kyriazidou, E. (1997a): "Estimation of a Panel Data Sample Selection Model," forthcoming in *Econometrica*, November 1997.

Kyriazidou, E. (1997b): "Estimation of Dynamic Panel Data Sample Selection Models," unpublished manuscript, Department of Economics, University of Chicago.

Maddala, G. S. (1983): "*Limited Dependent and Qualitative Variables in Econometrics*," Cambridge: Cambridge University Press.

Magnac, T. (1997): "State Dependence and heterogeneity in Youth Employment Histories," Working Paper, INRA and CREST, Paris.

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

Manski, C. (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, pp. 313-333.

Manski, C. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55, pp. 357-362.

Newey, W. K., and J. L. Powell (1987): "Asymmetric Least Squares Estimation and Testing," *Econometrica,* 55, pp. 819–847.

Nolan, D. and D. Pollard (1987): "U–Processes: Rates of Convergence," *The Annals of Statistics*, 16, No. 2, pp. 780-799.

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, pp. 1027-1057.

Phelps, E. (1972): *Inflation Policy and Unemployment Theory.* New York: Norton.

Pollard, D. (1984): Convergence of Stochastic Processes. New York: Springer-Verlag.

Pollard, D. (1993): "Uniform Ratio Limit Theorems for Empirical Processes," unpublished manuscript, Department of Statistics, Yale University.

Rasch, G. (1960): *Probabilistic Models for Some Intelligence and Attainment Tests*, Denmarks Pædagogiske Institut, Copenhagen.

Royden, H. L. (1988): *Real Analysis.* New York: Macmillan Publishing Company.

Staniswalis, J. G. (1989): "The Kernel Estimate of a Regression Function in Likelihood-Based Models," *Journal of the American Statistical Association*, pp. 276-283.

Tibshirani, R. and T. Hastie (1987): "Local Likelihood Estimation," *Journal of the American Statistical Association*, pp. 559-567.

8 Appendix

1. LOGIT CASE

The ideas behind the proofs of Theorems 1 and 2 are very closely related to those underlying local likelihood estimation (Staniswalis (1989), and Tibshirani and Hastie (1987)) and robust regression function estimation (Härdle (1984), and Härdle and Tsybakov (1988)). The difference is that the object of interest here is a finite dimensional vector, whereas in those papers it is an unknown function.

Proof of Theorem 1

After re-scaling by $1/n\sigma_n^k$, the objective function (6), can be written as:

$$Q_n(\theta) = \frac{1}{n\sigma_n^k}\sum_i K\left(\frac{x_{i23}}{\sigma_n}\right) h_i(\theta)$$

To show consistency of the maximizer of $Q_n(\theta)$, we will use Theorem 4.1.1 in Amemiya (1985). The theorem requires that the following conditions hold: (A1) Θ is compact, (A2) $Q_n(\theta)$ is continuous in $\theta \in \Theta$, (A3) $Q_n(\theta)$ is measurable for all $\theta \in \Theta$, (A4) $Q_n(\theta)$ converges to a nonstochastic function $Q(\theta)$ in probability uniformly in $\theta \in \Theta$ and (A5) $Q(\theta)$ is uniquely maximized at θ_0.

Notice that (A1) is satisfied by Assumption C2, while (A2) and (A3) are trivially satisfied.

To verify (A4) we will use Lemma 2.9 in Newey and McFadden (1994), which requires that Θ be compact (satisfied by Assumption C2), that $Q_n(\theta)$ converge to $Q(\theta)$ in probability for all θ, where $Q(\theta)$ is continuous in θ, and that there exists $\alpha > 0$ and $Z_n = O_p(1)$ such that for all $\theta, \tilde{\theta} \in \Theta$, $\left|Q_n(\theta) - Q_n\left(\tilde{\theta}\right)\right| \leq Z_n \left\|\theta - \tilde{\theta}\right\|^{\alpha}$.

Note that standard arguments (bounded convergence) and under our assumptions,

$$E\left[Q_n(\theta)\right] = E\left[\frac{1}{\sigma_n^k} K\left(\frac{x_{23}}{\sigma_n}\right) h(\theta)\right] \to f(0)\, E\left[h(\theta)\right| x_{23} = 0] = Q(\theta)$$

and

$$Var\left[Q_n(\theta)\right] = O\left(\frac{1}{n\sigma_n^k}\right) = o(1)$$

which by Chebyshev's theorem imply that $Q_n(\theta) \xrightarrow{p} Q(\theta)$ for all $\theta \in \Theta$. Continuity of $Q(\theta)$ may be easily established by dominated convergence arguments using the fact that $|h(\theta)| \leq \ln 2 + 2\|z\|\|\theta\|$, and that under our assumptions $E\left[\|z\|\right| x_{23}] < \infty$. Next, note that by the multivariate mean value theorem and by the triangular inequality,

$$\begin{aligned}
\left|Q_n(\theta) - Q_n\left(\tilde{\theta}\right)\right| &\leq \left\|\theta - \tilde{\theta}\right\| \frac{1}{n\sigma_n^k}\sum_i \left|K\left(\frac{x_{i23}}{\sigma_n}\right)\right| \left\|h_i^{(1)}(\theta^*)\right\| \\
&\leq 2\left\|\theta - \tilde{\theta}\right\| \frac{1}{n\sigma_n^k}\sum_i \left|K\left(\frac{x_{i23}}{\sigma_n}\right)\right| \|z_i\|
\end{aligned}$$

where θ^* lies between θ and $\tilde{\theta}$, and where the second inequality follows from the fact that $\left\|h_i^{(1)}(\theta^*)\right\| \leq \left(1 + \frac{\exp(z\theta)}{1+\exp(z\theta)}\right)\|z_i\| \leq 2\|z_i\|$. Let $Z_n = \frac{1}{n\sigma_n^k}\sum_i \left|K\left(\frac{x_{i23}}{\sigma_n}\right)\right| \|z_i\|$. It is straightforward to show that under our assumptions, $E\left[Z_n\right] = O(1)$ and $Var\left[Z_n\right] = O\left(\frac{1}{n\sigma_n^k}\right)$, which imply that $Z_n = O_p(1)$ as required.

Finally, to complete the proof of the theorem we need to show (A5), i.e. that $Q(\theta)$ is uniquely maximized at θ_0. Recall that $\theta = (b, g)'$. By assumption, $Q(b, g)$ is well–defined and finite for all b and g, and

$$\begin{aligned} Q(b,g) &= f(0)E\left[h(b,g)|x_{23}=0\right] \\ &= f(0)E\left[P\left(y_1 \neq y_2|x_{23}=0,y_0,y_3\right)E\left[\tilde{h}(b,g)|x_{23}=0,y_0,y_3,y_1\neq y_2\right]\Big|x_{23}=0\right] \end{aligned}$$

where $\tilde{h}(b,g) \equiv \ln\left(\frac{\exp(x_{12}b+y_{03}g)^{y_1}}{1+\exp(x_{12}b+y_{03}g)}\right)$. The logit assumption implies that $P\left(y_1 \neq y_2|x_{23}=0,y_0,y_3\right) > 0$. Next note that conditional on $(x_{23}=0,y_0,y_3,y_1\neq y_2)$, $\tilde{h}\left(b,g\right)$ is the log–likelihood of a logit model with explanatory variables x_{12} and y_{03}. If $y_3 = y_0$, then $E\left[\tilde{h}(\theta)|x_{23}=0,y_0,y_3,y_1\neq y_2\right]$ does not depend on g, but as a function of b, it is uniquely maximized at $b=\beta$ provided that x_{12} is not contained in a proper linear subspace of $\mathcal{R}^k$ with probability 1, conditional on $(x_{23}=0,y_0,y_3,y_1\neq y_2)$ (this follows from standard proofs of consistency of the maximum likelihood estimator of a logit model). By the maintained logit assumption, this "full–rank" condition follows from (C6). If $y_3 \neq y_0$ then $E\left[\tilde{h}(\theta)|x_{23}=0,y_0,y_3,y_1\neq y_2\right]$ depends on both g and b, and it is uniquely maximized at $b=\beta$ and $g=\gamma$ provided that (x_{12},y_{03}) is not contained in a proper linear subspace of $\mathcal{R}^{k+1}$ with probability 1 conditional on $(x_{23}=0,y_0,y_3,y_1\neq y_2)$. Again, the latter will be true by the logit assumption under assumption (C6). This implies that $Q(b,g)$ is uniquely maximized at $b=\beta$ and $g=\gamma$.

Proof of Theorem 2 – Part (i)

The asymptotic normality of the proposed estimator is derived in a standard way. Since the objective function is differentiable, $\theta_0 \in int\left(\Theta\right)$, and $\hat{\theta}_n$ is a consistent estimator of θ_0, then for n large enough the estimator satisfies the first order conditions:

$$\frac{1}{\sqrt{n\sigma_n^k}}\sum_i K\left(\frac{x_{i23}}{\sigma_n}\right)h_i^{(1)}\left(\hat{\theta}_n\right)=0$$

An expansion of these around θ_0 yields:

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n\sigma_n^k}}\sum_i\left(K\left(\frac{x_{i23}}{\sigma_n}\right)h_i^{(1)}\left(\theta_0\right)-E\left[K\left(\frac{x_{i23}}{\sigma_n}\right)h_i^{(1)}\left(\theta_0\right)\right]\right) \\ &\quad +\sqrt{n\sigma_n^k}\frac{1}{n\sigma_n^k}\sum_i E\left[K\left(\frac{x_{i23}}{\sigma_n}\right)h_i^{(1)}\left(\theta_0\right)\right] \\ &\quad +\frac{1}{n\sigma_n^k}\sum_i K\left(\frac{x_{i23}}{\sigma_n}\right)h^{(2)}\left(\theta_n^*\right)\cdot\sqrt{n\sigma_n^k}\left(\hat{\theta}_n-\theta_0\right) \\ &= Z_n\left(\theta_0\right)+\sqrt{n\sigma_n^k}E\left[B_n\left(\theta_0\right)\right]-J_n\left(\theta_n^*\right)\cdot\sqrt{n\sigma_n^k}\left(\hat{\theta}_n-\theta_0\right) \end{aligned}$$

where θ_n^* (which may be different for different rows of $J_n\left(\cdot\right)$) lies between $\hat{\theta}_n$ and θ_0 and therefore converges in probability to θ_0. The asymptotic normality of $\sqrt{n\sigma_n^k}\left(\hat{\theta}_n-\theta_0\right)$ will follow from $Z_n\left(\theta_0\right)\overset{d}{\to}N\left(0,V\right)$, $\sqrt{n\sigma_n^k}EB_n\left(\theta_0\right)\to 0$, and $J_n\left(\theta\right)\overset{p}{\to}J\left(\theta\right)$ uniformly in $\theta\in\Theta$ where $J\left(\theta\right)$ is a nonstochastic function that is continuous at θ_0. This last result will imply (by Theorem 4.1.5 in Amemiya (1985)) that $J_n(\theta_n^*) \xrightarrow{p} J(\theta_0) \equiv J$.

We will first show that $Z_n(\theta_0) \xrightarrow{d} N(0, V)$.

Let c be a $(k+1) \times 1$ vector of finite constants such that $c'c = 1$. To show the claim it suffices to show that $c' Z_n(\theta_0) \xrightarrow{d} N(0, c'Vc)$. Write:

$$c' Z_n(\theta_0) \equiv \frac{1}{\sqrt{n\sigma_n^k}} \sum_i c' \left(q_i^{(1)}(\theta_0) - E\left[q_i^{(1)}(\theta_0)\right]\right) = \frac{1}{\sqrt{n}} \sum_i \xi_{in}$$

where $\{\xi_{in}\}_{i=1}^n$ is an independent sequence of scalar random variables. We will verify that $\{\xi_{in}\}_{i=1}^n$ satisfies the conditions of the Lyapounov CLT for double arrays (see Theorem 7.1.2 in Chung (1974) and comment on page 209). We need $E[\xi_{in}] = 0$, $Var[\xi_{in}] < \infty$, $V \equiv \lim_{n\to\infty} Var[\xi_{in}] < \infty$, and $\sum_{i=1}^n E\left[\left|\frac{\xi_{in}}{\sqrt{n}}\right|^{2+\delta}\right] \to 0$ for some $\delta \in (0,1)$. Indeed:

$$\begin{aligned}
E[\xi_{in}] &= c' E\left[\frac{1}{\sqrt{\sigma_n^k}} q_i^{(1)}(\theta_0) - E\left[\frac{1}{\sqrt{\sigma_n^k}} q_i^{(1)}(\theta_0)\right]\right] = 0 \\
Var[\xi_{in}] &= c' E[\xi_{in}\xi_{in}'] c = c' E\left[\frac{1}{\sigma_n^k} q_i^{(1)}(\theta_0) q_i^{(1)}(\theta_0)'\right] c - \frac{1}{\sigma_n^k} c' E\left[q^{(1)}(\theta_0)\right] E\left[q^{(1)}(\theta_0)\right]' c
\end{aligned}$$

Note that under our assumptions (see Assumptions (N5) and (C7)) both terms of $Var[\xi_{in}]$ are finite. Now, for the first term of the variance, bounded convergence yields:

$$\begin{aligned}
E\left[\frac{1}{\sigma_n^k} q^{(1)}(\theta_0) q^{(1)}(\theta_0)'\right] &= E\left[\frac{1}{\sigma_n^k} K\left(\frac{x_{23}}{\sigma_n}\right)^2 E\left[h^{(1)}(\theta_0) h^{(1)}(\theta_0)' \middle| x_{23}\right]\right] \\
&\to f(0) E\left[h^{(1)}(\theta_0) h^{(1)}(\theta_0)' \middle| x_{23} = 0\right] \int K(\nu)^2 d\nu
\end{aligned}$$

since $f(\cdot)$ and $E\left[h^{(1)}(\theta_0) h^{(1)}(\theta_0)' \middle| x_{23} = \cdot\right]$ are continuous in a neighborhood of zero (Assumptions (C3) and (N6)). The second component of $Var[\xi_{in}]$ goes to 0 since $E\left[q^{(1)}(\theta_0)\right] = O\left(\sigma_n^k\right)$. Therefore,

$$\lim_{n\to\infty} Var[\xi_{in}] = c' \left\{ f(0) E\left[h^{(1)}(\theta_0) h^{(1)}(\theta_0)' \middle| x_{23} = 0\right] \int K^2(\nu) d\nu \right\} c = c'Vc$$

which under our assumptions is bounded away from infinity. Finally, for any $\delta \in (0,1)$,

$$\begin{aligned}
\sum_{i=1}^n E\left[\left|\frac{\xi_{in}}{\sqrt{n}}\right|^{2+\delta}\right] &\leq 2^{2+\delta} \left(\frac{1}{\sqrt{n\sigma_n^k}}\right)^\delta E\left[\frac{1}{\sigma_n^k} \left|K\left(\frac{x_{23}}{\sigma_n}\right)\right|^{2+\delta} \left\|h^{(1)}(\theta_0)\right\|^{2+\delta}\right] \\
&= 2^{2+\delta} \left(\frac{1}{\sqrt{n\sigma_n^k}}\right)^\delta \int |K(\nu)|^{2+\delta} E\left[\left\|h^{(1)}(\theta_0)\right\|^{2+\delta} \middle| x_{23} = \nu\sigma_n\right] f(\nu\sigma_n) d\nu \\
&\leq 2^{2+\delta} \left(\frac{1}{\sqrt{n\sigma_n^k}}\right)^\delta \int |K(\nu)|^{2+\delta} E\left[\|z\|^{2+\delta} \middle| x_{23} = \nu\sigma_n\right] f(\nu\sigma_n) d\nu \\
&= O\left(\frac{1}{\sqrt{n\sigma_n^k}}\right)^\delta = o(1)
\end{aligned}$$

since $\int |K(\nu)|^{2+\delta} d\nu$ is finite by the finiteness and absolute integrability of the kernel, and $E\left[\|z\|^{2+\delta} |x_{23} = \cdot\right] f(\cdot)$ is bounded for all x_{23} under (C3) and (N5).

We will next show that $\sqrt{n\sigma_n^k} E B_n(\theta_0) \to 0$.

By Assumptions (N2) and (N3), $E\left[h^{(1)}(\theta_0) \middle| x_{23} = \cdot\right] f(\cdot) \equiv \varphi(\cdot)$ is s times continuously differentiable. A Taylor expansion around $x_{23} = 0$ yields:

$$
\begin{aligned}
E[B_n(\theta_0)] &= E\left[\frac{1}{\sigma_n^k} K\left(\frac{x_{23}}{\sigma_n}\right) h^{(1)}(\theta_0)\right] \\
&= \int \frac{1}{\sigma_n^k} K\left(\frac{x_{23}}{\sigma_n}\right) E\left[h^{(1)}(\theta_0) \middle| x_{23}\right] f(x_{23})\, dx_{23} \\
&= \int \frac{1}{\sigma_n^k} K\left(\frac{x_{23}}{\sigma_n}\right) \varphi(x_{23})\, dx_{23} \\
&= \int \frac{1}{\sigma_n^k} K\left(\frac{x_{23}}{\sigma_n}\right) \left[\varphi(0) + \varphi^{(1)}(0)\, x_{23} + \ldots + \frac{1}{s!}\varphi^{(s)}(\tilde{x}_{23})\, x_{23}^s\right] dx_{23} \\
&= \varphi(0) \int K(\nu)\, d\nu + \varphi^{(1)}(0)\, \sigma_n \int \nu K(\nu)\, d\nu + \ldots \\
&\quad + \frac{1}{(s-1)!}\varphi^{(s-1)}(0)\, \sigma_n^{(s-1)} \int \nu^{s-1} K(\nu)\, d\nu + \sigma_n^s \frac{1}{s!} \int \nu^s K(\nu)\, \varphi^{(s)}(c_n)\, d\nu \\
&= 0 + \ldots + 0 + \sigma_n^s \frac{1}{s!} \int \nu^s K(\nu)\, \varphi^{(s)}(c_n)\, d\nu
\end{aligned}
$$

since $\varphi(0) \equiv f(0) \cdot E\left[h^{(1)}(\theta_0) \middle| x_{23} = 0\right] = f(0) \cdot 0 = 0$ by the first order conditions of the limiting maximization problem, and K is an s'th order kernel (Assumption N7). Here, c_n is a $(k \times 1)$ vector whose elements lie between 0 and $\nu\sigma_n$. Now, under our assumptions $\frac{1}{s!}\int \nu^s K(\nu)\, \varphi^{(s)}(c_n)\, d\nu = O(1)$. Therefore, since $\sqrt{n\sigma_n^k}\sigma_n^s \to 0$, we obtain,

$$
\sqrt{n\sigma_n^k} E[B_n(\theta_0)] = \sqrt{n\sigma_n^k}\sigma_n^s \frac{1}{s!} \int \nu^s K(\nu)\, \varphi^{(s)}(c_n)\, d\nu \to 0
$$

Finally, we will show that $\sup_{\theta\in\Theta} \|J_n(\theta) - J(\theta)\| = o_p(1)$ where $J(\theta) \equiv -f(0) \cdot E\left[h^{(2)}(\theta) \middle| x_{23} = 0\right]$ is continuous at θ_0.

First note that $J(\theta)$ is continuous in θ for all θ (and hence at θ_0). This may be easily verified using dominated convergence arguments and the proof is omitted. To verify the uniform convergence, we will use Lemma 2.9 of Newey and McFadden (1994).

Note that standard arguments (bounded convergence) and under our assumptions,

$$
E[J_n(\theta)] = -E\left[\frac{1}{\sigma_n^k} K\left(\frac{x_{23}}{\sigma_n}\right) h^{(2)}(\theta)\right] \to -f(0)\, E\left[h^{(2)}(\theta) \middle| x_{23} = 0\right] \equiv J(\theta)
$$

since $f(\cdot)$ and $E\left[h^{(2)}(\theta) |x_{23} = \cdot\right]$ are continuous in a neighborhood of zero, and for the jl'th component of $J_n(\theta)$,

$$
Var\left[J_n(\theta)_{jl}\right] \leq \frac{1}{n\sigma_n^k} E\left[\frac{1}{\sigma_n^k} K\left(\frac{x_{23}}{\sigma_n}\right)^2 h^{(2)}(\theta)_{jl}^2\right] = O\left(\frac{1}{n\sigma_n^k}\right) = o(1)
$$

which by Chebyshev's theorem imply that $J_n(\theta) \xrightarrow{p} J(\theta)$ for all $\theta \in \Theta$. Next, note that by the multivariate mean value theorem and by the triangular inequality,

$$\begin{aligned}\left\|J_n(\theta) - J_n\left(\tilde{\theta}\right)\right\| &\leq \left\|\theta - \tilde{\theta}\right\| \frac{1}{n\sigma_n^k} \sum_i \left|K\left(\frac{x_{i23}}{\sigma_n}\right)\right| \left\|h_i^{(3)}\left(\theta^*\right)\right\| \\ &\leq \left\|\theta - \tilde{\theta}\right\| \frac{1}{n\sigma_n^k} \sum_i \left|K\left(\frac{x_{i23}}{\sigma_n}\right)\right| \left\|z_i\right\|^3\end{aligned}$$

where θ^* lies between θ and $\tilde{\theta}$, and where the second inequality follows from the fact that $\left\|h_i^{(3)}\left(\theta^*\right)\right\| \leq \left|\frac{\exp(z_i\theta^*)(1-\exp(z_i\theta^*))}{(1+\exp(z_i\theta^*))^3}\right| \left\|z_i\right\|^3 \leq \left\|z_i\right\|^3$. Let $Z_n = \frac{1}{n\sigma_n^k}\sum_i \left|K\left(\frac{x_{i23}}{\sigma_n}\right)\right| \left\|z_i\right\|^3$. It is straightforward to show that $E\left[Z_n\right] = O\left(1\right)$ and $Var\left[Z_n\right] \leq O\left(\frac{1}{n\sigma_n^k}\right)$, since by Assumption (N5) $E\left[\left\|z\right\|^6 \middle| x_{23} = \cdot\right] < \infty$ which imply that $Z_n = O_p\left(1\right)$ as required.

Proof of Theorem 3

The proofs of the two parts of the theorem rely on the fact that, under the assumptions, $\hat{\theta}_n \xrightarrow{p} \theta_0$ and on the fact that the proposed estimators of the asymptotic variance components converge in probability uniformly in $\theta \in \Theta$ to nonstochastic limit functions which are continuous at θ_0. Uniform convergence in probability of $J_n\left(\theta\right)$ to $J\left(\theta\right)$ was established above. The proof that $V_n\left(\theta\right)$ converges in probability to $V\left(\theta\right)$ uniformly in $\theta \in \Theta$ follows exactly the same arguments and it is omitted.

2. SEMIPARAMETRIC CASE

The following result on uniform rates of convergence, adapted from Pollard (1993) to the multivariate case, will be useful (see also Theorem 37, Chapter 2 in Pollard, 1984). We will denote by P_n the empirical measure generated by independent sampling from a probability distribution P. Following the empirical process literature, we will also use P to denote the expectation operator.

Lemma 5 (Uniform Rates of Convergence) *Let $\mathcal{F}_n$ be a subclass of a fixed Euclidean class of functions $\mathcal{F}$ that has envelope F. Suppose there exist constants $\sigma_n \to 0$ such that $\sup_{\mathcal{F}_n} P\left|\chi\right| = O\left(\sigma_n^k\right)$. If F is constant and $n\sigma_n^k/\ln n \to \infty$, then:*

$$\sup_{\mathcal{F}_n}\left|P_n\chi - P\chi\right| = O_p\left(\sqrt{\frac{\sigma_n^k \ln n}{n}}\right) = o_p(\sigma_n^k)$$

Note that $n\sigma_n^k/\ln n \to \infty$ implies that $\ln n/n = o\left(\sigma_n^k\right)$. Hence, if $a_n = O_p\left(\sqrt{\frac{\sigma_n^k \ln n}{n}}\right)$ then $a_n = o_p\left(\sqrt{\left(\sigma_n^k\right)^2}\right) = o_p\left(\sigma_n^k\right)$.

Define:

$$Q_n(\theta) = \frac{1}{\sigma_n^k}\cdot\frac{1}{n}\sum_i K\left(\frac{x_{i23}}{\sigma_n}\right)h_i\left(\theta\right) = \frac{1}{\sigma_n^k}\cdot\frac{1}{n}\sum_i q_i\left(\theta\right) = \frac{1}{\sigma_n^k}P_n q$$

where now $h(\theta) = y_{21}\,\mathrm{sgn}(z\theta)$ with $z \equiv (x_{21}, y_{30})$. As we show below, $Q_n(\theta)$ converges in probability uniformly in $\theta \in \Theta$ to the nonstochastic function:

$$Q(\theta) = f(0)E[h(\theta)|x_{23} = 0]$$

Proof of Theorem 4

Note that in the semiparametric case the objective function is no longer continuous in θ. To show consistency we will use Theorem 9.6.1 in Amemiya (1985) which requires that the following conditions hold: (A1) Θ is a compact set, (A2) $Q_n(\theta)$ is a measurable function for all $\theta \in \Theta$, (A3) $Q_n(\theta)$ converges in probability to a nonstochastic function $Q(\theta)$ uniformly in $\theta \in \Theta$, and (A4) $Q(\theta)$ is continuous in θ and is uniquely maximized at θ_0.

The first condition is satisfied by construction of Θ. Condition (A2) is trivially satisfied. For condition (A3) we need to verify that $\sup_{\theta\in\Theta}|Q_n(\theta) - Q(\theta)| = o_p(1)$, which follows from $\sup_{\theta\in\Theta}|Q_n(\theta) - EQ_n(\theta)| = o_p(1)$ and $\sup_{\theta\in\Theta}|EQ_n(\theta) - Q(\theta)| = o(1)$.

Note that $\sup_{\theta\in\Theta}|Q_n(\theta) - EQ_n(\theta)| = o_p(1) \Longleftrightarrow \sup_{\mathcal{F}_n}|P_n q - Pq| = o_p(\sigma_n^k)$ where

$$\mathcal{F}_n \equiv \left\{K\left(\frac{x_{23}}{\sigma_n}\right) h(\theta) : \theta \in \Theta\right\}$$

and $\sigma_n > 0$ and $\sigma_n \to 0$. $\mathcal{F}_n$ is a subclass of the fixed class $\mathcal{F} = \left\{K\left(\frac{x_{23}}{\sigma}\right) h(\theta) : \theta \in \Theta, \sigma > 0\right\} = \mathcal{F}_\sigma \mathcal{F}_\theta$, with $\mathcal{F}_\sigma \equiv \left\{K\left(\frac{x_{23}}{\sigma}\right) : \sigma > 0\right\}$, which is Euclidean for the constant envelope $\sup|K|$ (see Lemma 22(ii) in Nolan and Pollard, 1987) and $\mathcal{F}_\theta \equiv \{h(\theta) = y_{21}\,\mathrm{sgn}(z\theta) : \theta \in \Theta\}$. Note that $h(\theta)$ is uniformly bounded by the constant envelope $F_\theta = 1$. Furthermore, it is easy to see that there is a partition of $\Re$ into $k+1$ intervals on each of which $h(\theta)$ is linear. This implies that $\mathcal{F}$ is Euclidean for the constant envelope $F = \sup|K|$ (see example 2.11 in Pakes and Pollard (1989)). Next, note that

$$\begin{aligned}
\sup_{\mathcal{F}_n} P|q| &\leq \sup_{\mathcal{F}_n} \sigma_n^k \int |K(\nu)|\, E\left[|y_{21}\,\mathrm{sgn}(z\theta)|\,|\, x_{23} = \nu\sigma_n\right] f(\nu\sigma_n)\, d\nu \\
&\leq \sup_{\mathcal{F}_n} \sigma_n^k \int |K(\nu)|\, f(\nu\sigma_n)\, d\nu = O\left(\sigma_n^k\right)
\end{aligned}$$

Applying Lemma 5 we obtain that

$$\sup_{\mathcal{F}_n}|P_n q - Pq| = O_p\left(\sqrt{\frac{\sigma_n^k \ln n}{n}}\right) = o_p(\sigma_n^k)$$

by assumption (CS8).

We next show that $\sup_{\theta\in\Theta}|EQ_n(\theta) - Q(\theta)| = o(1)$. Let $\varphi(\cdot) \equiv E[h(\theta)|\, x_{23} = \cdot] f(\cdot)$. Notice that by assumption (CS6) we can write:

$$\begin{aligned}
\sup_{\theta\in\Theta}|EQ_n(\theta) - Q(\theta)| &= \sup_{\theta\in\Theta}\left|E\left[\frac{1}{\sigma_n^k} K\left(\frac{x_{23}}{\sigma_n}\right) h(\theta)\right] - f(0)\, E[h(\theta)|\, x_{23} = 0]\right| \\
&= \sup_{\theta\in\Theta}\left|\int \frac{1}{\sigma_n^k} K\left(\frac{x_{23}}{\sigma_n}\right) \varphi(x_{23})\, dx_{23} - \varphi(0)\right|
\end{aligned}$$

$$
\begin{aligned}
&= \sup_{\theta\in\Theta}\left|\int \frac{1}{\sigma_n^k} K\left(\frac{x_{23}}{\sigma_n}\right)\left[\varphi(0)+\varphi^{(1)}(\tilde{x}_{23})\,x_{23}\right]dx_{23}-\varphi(0)\right| \\
&= \sup_{\theta\in\Theta}\left|\varphi(0)\int K(\nu)\,d\nu+\sigma_n^k\int \nu K(\nu)\,\varphi^{(1)}(c_n)\,d\nu-\varphi(0)\right| \\
&= \sup_{\theta\in\Theta}\left|\sigma_n^k\int \nu K(\nu)\,\varphi^{(1)}(c_n)\,d\nu\right| \\
&\le \sigma_n^k\int |\nu K(\nu)| \sup_{\theta\in\Theta}\left|\varphi^{(1)}(c_n)\right| d\nu \\
&= O\left(\sigma_n^k\right) \\
&= o(1)
\end{aligned}
$$

and the desired result follows.

To show that the identification condition (A4) holds we will follow Manski (1985, 1987). Let

$$
Z_\theta \equiv \{z : \operatorname{sgn}(z\theta) \neq \operatorname{sgn}(z\theta_0)\}
$$

and

$$
R(\theta) \equiv \int_{Z_\theta} dF_{z|x_{23}=0}
$$

Lemma 6 establishes that $R(\theta) > 0$ for all $\theta \in \Re^{k+1}$ such that $\theta/\|\theta\| \neq \theta_0^*$. Lemma 7 uses this result to establish the desired result, namely that $Q(\theta_0^*) > Q(\theta)$. We will assume throughout that all distributions that condition on the event that $x_{23} = 0$ are well defined.

Lemma 6 *Let assumptions (CS2)-(CS4) hold. Then $R(\theta) > 0$ for all $\theta \in \Re^{k+1}$ such that $\theta/\|\theta\| \neq \theta_0^*$.*

Proof: The claim follows from Lemma 2 of Manski (1985) provided that we can show that (i) $\Pr\left(x_{21,\kappa} \in N \middle| \tilde{x}_{21}, y_0 = d_0, y_3\right.$ 0 for any subset N of $\Re$, and (ii) $\Pr\left(\tilde{x}_{21}\tilde{b} + y_{30}g = 0 \middle| x_{23} = 0\right) < 1$ for all $\left(\tilde{b}, g\right)$.

(i) Note that for any subset N of $\Re$:

$$
\begin{aligned}
&\Pr\left(x_{21,\kappa} \in N \middle| \tilde{x}_{21}, y_0 = d_0, y_3 = d_3, x_{23} = 0\right) \\
&= \frac{\Pr\left(x_{21,\kappa} \in N, y_0 = d_0, y_3 = d_3 \middle| \tilde{x}_{21}, x_{23} = 0\right)}{\Pr\left(y_0 = d_0, y_3 = d_3 \middle| \tilde{x}_{21}, x_{23} = 0\right)} \\
&= \frac{\Pr\left(x_{21,\kappa} \in N \middle| \tilde{x}_{21}, x_{23} = 0\right)\cdot \Pr\left(y_0 = d_0, y_3 = d_3 \middle| x_{21,\kappa} \in N, \tilde{x}_{21}, x_{23} = 0\right)}{\Pr\left(y_0 = d_0, y_3 = d_3 \middle| \tilde{x}_{21}, x_{23} = 0\right)} \\
&= \frac{\Pr\left(x_{21,\kappa} \in N \middle| \tilde{x}_{21}, x_{23} = 0\right)\cdot \int \Pr\left(y_0 = d_0, y_3 = d_3 \middle| x, \alpha, x_{23} = 0\right) dF_{x,\alpha|x_{23}=0,\tilde{x}_{21},x_{21,\kappa}\in N}}{\Pr\left(y_0 = d_0, y_3 = d_3 \middle| \tilde{x}_{21}, x_{23} = 0\right)}
\end{aligned}
$$

By Assumption (CS3) $x_{21,\kappa}$ has positive density on $\Re$ for almost all $\tilde{x}_{21}$ and conditional on $x_{23} = 0$, so that $\Pr\left(x_{21,\kappa} \in N \middle| \tilde{x}_{21}, x_{23} = 0\right) > 0$. Next, note that

$$\int \Pr\left(y_0 = d_0, y_3 = d_3 | \, x, \alpha, x_{23} = 0\right) dF_{x,\alpha | x_{23}=0, \tilde{x}_{21}, x_{21,\kappa} \in N}$$

$$= \int \Pr\left(y_3 = d_3 | \, x, \alpha, y_0 = d_0, x_{23} = 0\right) \Pr\left(y_0 = d_0 | \, x, \alpha, x_{23} = 0\right) dF_{x,\alpha | x_{23}=0, \tilde{x}_{21}, x_{21,\kappa} \in N}$$

$$= \int \sum_{d_1, d_2 \in \{0,1\}} \Pr\left(y_3 = d_3 | \, x, \alpha, y_0 = d_0, y_2 = d_2, y_1 = d_1, x_{23} = 0\right) \Pr\left(y_2 = d_2 | \, x, \alpha, y_0 = d_0, y_1 = d_1, x_{23} = 0\right)$$

$$\Pr\left(y_1 = d_1 | \, x, \alpha, y_0 = d_0, x_{23} = 0\right) \Pr\left(y_0 = d_0 | \, x, \alpha, x_{23} = 0\right) dF_{x,\alpha | x_{23}=0, \tilde{x}_{21}, x_{21,\kappa} \in N}$$

$$= \int \sum_{d_1, d_2 \in \{0,1\}} F\left(x_3 \beta + \alpha + \gamma d_2\right) F\left(x_2 \beta + \alpha + \gamma d_1\right) F\left(x_1 \beta + \alpha + \gamma d_0\right)$$

$$\cdot \Pr\left(y_0 = d_0 | \, x, \alpha, x_{23} = 0\right) dF_{x,\alpha | x_{23}=0, \tilde{x}_{21}, x_{21,\kappa} \in N}$$

where the integrand is strictly positive by Assumption (CS2), and from the fact that $\Pr\left(y_0 = d_0 | \, x, \alpha, x_{23} = 0\right)$ will be always positive even if $y_0 = 1$ (or 0) with probability one conditional on (x, α).

Hence, $\Pr\left(y_0 = d_0, y_3 = d_3 | \, x_{21,\kappa} \in N, \tilde{x}_{21}, x_{23} = 0\right) > 0$ and therefore $\Pr\left(y_0 = d_0, y_3 = d_3 | \, \tilde{x}_{21}, x_{23} = 0\right) > 0$.

(ii) Note that for $g = 0$, $\Pr\left(\tilde{x}_{21} \tilde{b} + y_{30} g = 0 \middle| \, x_{23} = 0\right) = \Pr\left(\tilde{x}_{21} \tilde{b} = 0 \middle| \, x_{23} = 0\right) < 1$ by the full rank condition in Assumption (CS4). Now, for $g \neq 0$

$$\begin{aligned} \Pr\left(\tilde{x}_{21} \tilde{b} + y_{30} g = 0 \middle| \, x_{23} = 0\right) &= \int \Pr\left(y_{30} g = -\tilde{x}_{21} \tilde{b} \middle| \, \tilde{x}_{21}, x_{23} = 0\right) dF_{\tilde{x}_{21} | x_{23}=0} \\ &= \int \sum_{d_0 \in \{0,1\}} \Pr\left(y_3 g = d_0 g - \tilde{x}_{21} \tilde{b} \middle| \, y_0 = d_0, \tilde{x}_{21}, x_{23} = 0\right) \\ &\quad \cdot \Pr\left(y_0 = d_0 | \, \tilde{x}_{21}, x_{23} = 0\right) dF_{\tilde{x}_{21} | x_{23}=0} \end{aligned}$$

By Assumption (CS2), $\Pr\left(y_3 g = d_0 g - \tilde{x}_{21} \tilde{b} \middle| \, y_0 = d_0, \tilde{x}_{21}, x_{23} = 0\right) < 1$, and the desired result follows.

Lemma 7 *Let assumptions (CS2)-(CS5) hold. Then $Q(\theta_0^*) > Q(\theta)$ for all $\theta \in \Re^{k+1}$ such that $\theta / \|\theta\| \neq \theta_0^*$.*

Proof: For all $\theta \in \Re^{k+1}$,

$$\begin{aligned} & Q(\theta_0^*) - Q(\theta) \\ = \; & f(0) E\left[y_{21} \left(\operatorname{sgn}\left(z \theta_0^*\right) - \operatorname{sgn}\left(z \theta\right)\right) | \, x_{23} = 0\right] \end{aligned}$$

$$= \quad 2f(0)\int_{Z_\theta} \operatorname{sgn}(z\theta_0)\, E\left[y_2 - y_1 \mid z, x_{23}=0\right] dF_{z|x_{23}=0}$$

$$= \quad 2f(0)\int_{Z_\theta} \operatorname{sgn}(z\theta_0)\, E\left[E\left[y_2 - y_1 \mid x, \alpha, y_0 = d_0, y_3 = d_3, x_{23}=0\right] \mid z, x_{23}=0\right] dF_{z|x_{23}=0}$$

$$\begin{aligned} = \quad & 2f(0)\int_{Z_\theta} \operatorname{sgn}(z\theta_0)\, E\left[\Pr\left(y_1 = 0, y_2 = 1 \mid x, \alpha, y_0 = d_0, y_3 = d_3, x_{23}=0\right)\right. \\ & \left. - \Pr\left(y_1 = 1, y_2 = 0 \mid x, \alpha, y_0 = d_0, y_3 = d_3, x_{23}=0\right) \mid z, x_{23}=0\right] dF_{z|x_{23}=0} \end{aligned}$$

$$\begin{aligned} = \quad & 2f(0)\int_{Z_\theta} \operatorname{sgn}(z\theta_0)\, E\left[\Pr\left(A \mid x, \alpha, y_0 = d_0, y_3 = d_3, x_{23}=0\right)\right. \\ & \left. - \Pr\left(B \mid x, \alpha, y_0 = d_0, y_3 = d_3, x_{23}=0\right) \mid z, x_{23}=0\right] dF_{z|x_{23}=0} \end{aligned}$$

$$= \quad 2f(0)\int_{Z_\theta} E\left[\operatorname{sgn}(z\theta_0)\left(\frac{\Pr(A|x,\alpha,x_{23}=0) - \Pr(B|x,\alpha,x_{23}=0)}{\Pr\left(y_0 = d_0, y_3 = d_3 \mid x, \alpha, x_{23}=0\right)}\right)\right| \left. z, x_{23}=0\right] dF_{z|x_{23}=0}$$

As we have shown in Section 2, $\operatorname{sgn}\left(\Pr(A|x,\alpha,x_{23}=0) - \Pr(B|x,\alpha,x_{23}=0)\right) = \operatorname{sgn}(z\theta_0)$, so that the integrand above is non-negative which implies that $\operatorname{sgn}(z\theta_0)\, E\left[y_2 - y_1 \mid z, x_{23}=0\right] = \left|E\left[y_2 - y_1 \mid z, x_{23}=0\right]\right|$. Therefore,

$$Q(\theta_0^*) - Q(\theta) = 2f(0)\int_{Z_\theta} \left|E\left[y_2 - y_1 \mid z, x_{23}=0\right]\right| dF_{z|x_{23}=0} \geq 0$$

Now, $E\left[y_2 - y_1 \mid z, x_{23}=0\right] \neq 0$ for almost all z since $E\left[y_2 - y_1 \mid z, x_{23}=0\right] = 0$ if and only if $\operatorname{sgn}(z\theta_0) = 0$, or equivalently $z\theta_0 = 0$, an event that has zero probability measure under our assumptions. It then follows from Lemma 6 and from Assumption (CS5) that $Q(\theta_0^*) - Q(\theta) > 0$ whenever $\theta / \|\theta\| \neq \theta_0^*$.

Finally, to show continuituy of the limiting objective function with respect to θ, we follow Lemma 5 of Manski (1985). From the proof of that Lemma, it is clear that we need to establish continuity of terms of the form:

$$\begin{aligned} & \Pr\left(y_1 = d_1, y_2 = d_2, x_{21,\kappa} b_\kappa > -g\left(y_3 - y_0\right) - \tilde{x}_{21}\tilde{b} \,\middle|\, x_{23} = 0\right) \\ = \quad & \sum_{d_0, d_3 \in \{0,1\}} \Pr\left(y_1 = d_1, y_2 = d_2, x_{21,\kappa} b_\kappa > -g\left(d_3 - d_0\right) - \tilde{x}_{21}\tilde{b} \,\middle|\, x_{23} = 0, y_0 = d_0, y_3 = d_3\right) \times \\ & \Pr\left(y_0 = d_0, y_3 = d_3 \mid x_{23} = 0\right) \\ = \quad & \sum_{d_0, d_3 \in \{0,1\}} \int_{\tilde{x}_{21}} \left[\int_{-g(d_3 - d_0) - \tilde{x}_{21}\tilde{b}} \Pr\left(y_1 = d_1, y_2 = d_2 \mid x, x_{23} = 0, y_0 = d_0, y_3 = d_3\right)\right. \\ & \left. f_{x_{21,\kappa}|x_{23}=0, y_0=d_0, y_3=d_3, \tilde{x}_{21}}\left(x_{21,\kappa}\right) dx_{21,\kappa}\right] dF_{\tilde{x}_{21}|x_{23}=0, y_0=d_0, y_3=d_3} \times \\ & \Pr\left(y_0 = d_0, y_3 = d_3 \mid x_{23} = 0\right) \end{aligned}$$

Note that in order to establish continuity of the last expression above with respect to $\theta = \left(\tilde{b}, g\right)$ it is sufficient that $f_{x_{21,\kappa}|x_{23}=0, y_0=d_0, y_3=d_3}\left(x_{21,\kappa}\right)$ does not have any mass points. This will be true given our assumption

(CS3) on $f_{x_{21,\kappa}|x_{23}=0,\tilde{x}_{21}}\left(x_{21,\kappa}\right).$

Table 1. Benchmark design. ($\gamma = 0.5, \beta = 1, x_{it} \sim N(0, \pi^2/3), \alpha_i = \sum_{t=1}^{4} x_{it}/4$)

		Results for $\widehat{\beta}_n$				Results for $\widehat{\gamma}_n$			
	Sample Size	Mean Bias	RMSE	Median Bias	MAE	Mean Bias	RMSE	Median Bias	MAE
$\sigma_n = 1 \cdot n^{-1/5}$	250	1.393	5.631	0.284	0.450	0.890	6.642	0.103	1.106
	500	0.421	1.479	0.140	0.296	0.054	1.371	0.017	0.670
	1000	0.148	0.445	0.081	0.207	0.047	0.829	0.029	0.448
	2000	0.065	0.259	0.035	0.147	0.025	0.524	-0.007	0.350
	4000	0.037	0.182	0.017	0.109	0.015	0.405	0.006	0.264
$\sigma_n = 2 \cdot n^{-1/5}$	250	0.323	0.891	0.127	0.269	0.166	1.331	0.059	0.688
	500	0.144	0.431	0.075	0.205	-0.005	0.686	-0.007	0.412
	1000	0.073	0.255	0.044	0.148	0.022	0.506	-0.009	0.327
	2000	0.033	0.175	0.019	0.109	0.003	0.356	-0.007	0.242
	4000	0.019	0.120	0.010	0.079	0.011	0.277	0.013	0.192
$\sigma_n = 4 \cdot n^{-1/5}$	250	0.165	0.423	0.080	0.183	0.036	0.787	0.018	0.487
	500	0.085	0.264	0.049	0.144	-0.019	0.467	-0.013	0.295
	1000	0.049	0.173	0.035	0.108	-0.006	0.351	-0.022	0.221
	2000	0.024	0.123	0.019	0.076	-0.015	0.256	-0.024	0.170
	4000	0.015	0.086	0.010	0.053	-0.003	0.196	0.001	0.134
$\sigma_n = 8 \cdot n^{-1/5}$	250	0.137	0.335	0.076	0.154	-0.015	0.645	-0.039	0.403
	500	0.075	0.207	0.044	0.113	-0.044	0.382	-0.052	0.256
	1000	0.049	0.137	0.038	0.086	-0.038	0.282	-0.035	0.178
	2000	0.032	0.098	0.028	0.063	-0.041	0.207	-0.042	0.143
	4000	0.022	0.069	0.019	0.044	-0.031	0.151	-0.035	0.102
$\sigma_n = 16 \cdot n^{-1/5}$	250	0.138	0.320	0.088	0.152	-0.035	0.611	-0.069	0.367
	500	0.078	0.192	0.051	0.102	-0.063	0.361	-0.070	0.255
	1000	0.058	0.130	0.049	0.077	-0.062	0.264	-0.064	0.168
	2000	0.044	0.092	0.040	0.058	-0.065	0.196	-0.065	0.135
	4000	0.035	0.066	0.031	0.043	-0.058	0.141	-0.059	0.098
$\sigma_n = 32 \cdot n^{-1/5}$	250	0.141	0.320	0.090	0.152	-0.042	0.607	-0.068	0.363
	500	0.081	0.190	0.055	0.102	-0.070	0.359	-0.074	0.244
	1000	0.063	0.130	0.052	0.077	-0.071	0.263	-0.069	0.173
	2000	0.050	0.093	0.046	0.058	-0.075	0.197	-0.069	0.134
	4000	0.043	0.069	0.039	0.046	-0.072	0.144	-0.076	0.100
$\sigma_n = 64 \cdot n^{-1/5}$	250	0.142	0.320	0.092	0.156	-0.043	0.607	-0.072	0.359
	500	0.082	0.190	0.056	0.102	-0.072	0.359	-0.077	0.243
	1000	0.064	0.131	0.054	0.077	-0.074	0.264	-0.071	0.176
	2000	0.051	0.094	0.048	0.058	-0.079	0.199	-0.073	0.134
	4000	0.046	0.071	0.042	0.047	-0.076	0.146	-0.079	0.103

Table 2. Benchmark Design with different values of γ

		$T=4$				$T=8$			
		Results for $\widehat{\beta}_n$		Results for $\widehat{\gamma}_n$		Results for $\widehat{\beta}_n$		Results for $\widehat{\gamma}_n$	
	n	Bias	MAE	Bias	MAE	Bias	MAE	Bias	MAE
$\gamma = 0.25$	250	0.060	0.155	-0.027	0.354	0.010	0.049	-0.026	0.124
	500	0.029	0.106	-0.033	0.239	0.005	0.038	-0.025	0.081
	1000	0.028	0.082	-0.019	0.174	0.007	0.027	-0.019	0.062
	2000	0.018	0.060	-0.027	0.133	0.004	0.020	-0.018	0.045
	4000	0.008	0.039	-0.011	0.091	0.003	0.015	-0.018	0.034
$\gamma = 0.5$	250	0.076	0.154	-0.039	0.403	0.014	0.050	-0.053	0.131
	500	0.044	0.113	-0.052	0.256	0.007	0.037	-0.054	0.098
	1000	0.038	0.086	-0.035	0.178	0.009	0.027	-0.041	0.075
	2000	0.028	0.063	-0.042	0.143	0.007	0.020	-0.036	0.051
	4000	0.019	0.044	-0.035	0.102	0.005	0.015	-0.033	0.039
$\gamma = 1$	250	0.121	0.187	-0.028	0.454	0.021	0.054	-0.095	0.152
	500	0.076	0.126	-0.072	0.312	0.014	0.041	-0.092	0.113
	1000	0.062	0.095	-0.085	0.225	0.012	0.029	-0.080	0.098
	2000	0.045	0.072	-0.075	0.167	0.010	0.021	-0.074	0.077
	4000	0.036	0.051	-0.058	0.119	0.007	0.016	-0.064	0.065
$\gamma = 2.0$	250	0.196	0.251	-0.056	0.620	0.016	0.064	-0.195	0.227
	500	0.139	0.181	-0.117	0.417	0.014	0.044	-0.179	0.197
	1000	0.113	0.136	-0.148	0.321	0.016	0.034	-0.160	0.164
	2000	0.087	0.102	-0.142	0.241	0.009	0.023	-0.133	0.134
	4000	0.063	0.074	-0.118	0.163	0.006	0.017	-0.116	0.116

Table 3. Alternative Estimators for Benchmark Design ($n = 250$)

		Estimating the Fixed Effects				Estimator Proposed Here				Infeasible MLE			
		$\widehat{\beta}_n$		$\widehat{\gamma}_n$		$\widehat{\beta}_n$		$\widehat{\gamma}_n$		$\widehat{\beta}_n$		$\widehat{\gamma}_n$	
γ	T	Bias	MAE	Bias	MAE	Bias	MAE	Bias	MAE	Bias	MAE	Bias	MAE
0.25	4	0.796	0.796	-2.663	2.663	0.066	0.161	-0.022	0.341	0.008	0.060	-0.006	0.099
	8	0.263	0.263	-0.760	0.760	0.011	0.051	-0.024	0.131	0.005	0.036	0.003	0.065
	16	0.010	0.010	-0.307	0.307	0.003	0.028	-0.022	0.067	0.001	0.024	-0.003	0.041
0.5	4	0.787	0.787	-2.577	2.577	0.086	0.170	-0.039	0.360	0.009	0.058	-0.006	0.100
	8	0.257	0.257	-0.745	0.745	0.014	0.050	-0.053	0.131	0.003	0.035	0.005	0.062
	16	0.101	0.101	-0.298	0.298	0.005	0.029	-0.053	0.074	0.003	0.024	-0.000	0.040
1.0	4	0.772	0.772	-2.355	2.355	0.107	0.191	-0.057	0.402	-0.001	0.063	-0.011	0.101
	8	0.259	0.259	-0.700	0.700	0.021	0.054	-0.093	0.152	0.001	0.036	-0.000	0.062
	16	0.104	0.104	-0.291	0.291	0.005	0.028	-0.105	0.109	0.003	0.024	0.003	0.040
2.0	4	0.740	0.740	-2.054	2.054	0.195	0.260	-0.074	0.579	0.003	0.068	0.005	0.120
	8	0.268	0.268	-0.636	0.636	0.018	0.062	-0.190	0.222	0.006	0.039	0.008	0.079
	16	0.109	0.109	-0.289	0.289	-0.003	0.034	-0.200	0.201	0.001	0.028	-0.004	0.051

Table 4. Benchmark Design with various designs for x_{it}

	$\chi^2(1)$–Regressors. $\sigma_n = 8 \cdot n^{-1/5}$				Additional Regressors. $\sigma_n = 8 \cdot n^{-1/8}$			
	Results for $\widehat{\beta}_n$		Results for $\widehat{\gamma}_n$		Results for $\widehat{\beta}_{1n}$		Results for $\widehat{\gamma}_n$	
n	Bias	MAE	Bias	MAE	Bias	MAE	Bias	MAE
250	0.090	0.171	-0.011	0.309	0.143	0.186	-0.041	0.431
500	0.052	0.128	-0.010	0.222	0.094	0.141	-0.037	0.273
1000	0.038	0.085	-0.034	0.161	0.070	0.099	-0.026	0.197
2000	0.037	0.071	-0.032	0.117	0.051	0.068	-0.063	0.155
4000	0.032	0.053	-0.021	0.081	0.041	0.056	-0.050	0.116

	Trending Regressor. $T = 4$. $\sigma_n = 8 \cdot n^{-1/5}$				Trending Regressor. $T = 8$. $\sigma_n = 8 \cdot n^{-1/5}$			
	Results for $\widehat{\beta}_n$		Results for $\widehat{\gamma}_n$		Results for $\widehat{\beta}_n$		Results for $\widehat{\gamma}_n$	
n	Bias	MAE	Bias	MAE	Bias	MAE	Bias	MAE
250	0.089	0.184	-0.013	0.385	0.008	0.060	-0.049	0.136
500	0.036	0.127	-0.035	0.260	0.004	0.045	-0.055	0.099
1000	0.040	0.093	-0.046	0.196	-0.002	0.031	-0.040	0.071
2000	0.020	0.066	-0.055	0.141	0.001	0.021	-0.034	0.056
4000	0.015	0.050	-0.039	0.092	-0.001	0.015	-0.030	0.043

Table 5. Coverage Probabilities of CI's for Benchmark Design.

		Results for $\widehat{\beta}_n$		Results for $\widehat{\gamma}_n$	
	n	95% CI	80% CI	95% CI	80% CI
$\sigma_n = 2 \cdot n^{-1/5}$	250	0.911	0.720	0.874	0.679
	500	0.911	0.749	0.927	0.772
	1000	0.929	0.754	0.937	0.773
	2000	0.934	0.778	0.949	0.794
	4000	0.956	0.809	0.940	0.774
$\sigma_n = 8 \cdot n^{-1/5}$	250	0.943	0.786	0.926	0.751
	500	0.933	0.776	0.952	0.795
	1000	0.944	0.765	0.944	0.784
	2000	0.945	0.790	0.939	0.784
	4000	0.941	0.788	0.938	0.785
$\sigma_n = 32 \cdot n^{-1/5}$	250	0.949	0.766	0.935	0.759
	500	0.941	0.763	0.948	0.802
	1000	0.934	0.753	0.941	0.791
	2000	0.928	0.753	0.923	0.748
	4000	0.900	0.695	0.900	0.721

Table 6. Maximum Score Estimation for Benchmark Design.

		$\widehat{\beta}_n/\widehat{\gamma}_n$ (MS)		$\widehat{\beta}_n/\widehat{\gamma}_n$ (logit)	
	n	Bias	MAE	Bias	MAE
$\sigma_n = 2 \cdot n^{-1/5}$	250	-0.059	0.674	0.015	0.591
	500	-0.138	0.593	-0.041	0.397
	1000	-0.044	0.483	-0.025	0.334
	2000	-0.015	0.421	-0.025	0.239
	4000	-0.022	0.330	0.002	0.187
$\sigma_n = 8 \cdot n^{-1/5}$	250	-0.126	0.539	-0.078	0.365
	500	-0.113	0.451	-0.062	0.254
	1000	-0.066	0.354	-0.046	0.177
	2000	-0.072	0.303	-0.060	0.133
	4000	-0.054	0.212	-0.040	0.105
$\sigma_n = 32 \cdot n^{-1/5}$	250	-0.125	0.525	-0.091	0.347
	500	-0.109	0.431	-0.104	0.237
	1000	-0.084	0.339	-0.096	0.175
	2000	-0.109	0.277	-0.090	0.132
	4000	-0.103	0.214	-0.093	0.107