Dynamic time series binary choice*

Robert M. de Jong† Tiemen Woutersen ‡

May 31, 2007

Abstract

This paper considers dynamic time series binary choice models. It proves near epoch dependence and strong mixing for the dynamic binary choice model with correlated errors. Using this result, it shows in a time series setting the validity of the dynamic probit likelihood procedure when lags of the dependent binary variable are used as regressors, and it establishes the asymptotic validity of Horowitz' smoothed maximum score estimation of dynamic binary choice models with lags of the dependent variable as regressors. For the semiparametric model, the latent error is explicitly allowed to be correlated. It turns out that no long-run variance estimator is needed for the validity of the smoothed maximum score procedure in the dynamic time series framework.

1 Introduction

For a dynamic linear time series model

$$y_n = \sum_{j=1}^{p} \rho_j y_{n-j} + \gamma' x_n + u_n, \tag{1}$$

$n = 1, \ldots, N$, it is well-known that a sufficient condition for consistency as $N \to \infty$ of the least squares estimator is that $E(u_n | y_{n-1}, \ldots, y_{n-p}, x_n) = 0$, and that even if u_n is weakly dependent, consistency can be proven as long as this condition holds,

*We thank Stephen Cosslett, James Davidson, Jon Faust, Lung-Fei Lee, Benedikt Pötscher, Jim Stock and Jeff Wooldridge for helpful discussions.

†Department of Economics, Ohio State University, email dejong@econ.ohio-state.edu.

‡Department of Economics, Johns Hopkins University, email woutersen@jhu.edu.

without the assumption of normality on u_n. In this paper, we analyze maximum likelihood estimation of the dynamic probit model of order p, and maximum score estimation of dynamic binary choice models of order p, and we explicitly allow the error to be correlated. We define the dynamic binary choice model of order p as

$$y_n = I(\sum_{j=1}^{p} \rho_j y_{n-j} + \gamma' x_n + u_n > 0), \tag{2}$$

where $I(\cdot)$ denotes the indicator function, x_n is predetermined, and u_n can be correlated and heteroskedastic. We first show near epoch dependence and strong mixing for this model. We then impose identifying assumptions to ensure identification of the probit model and the binary choice model. For conditional maximum likelihood estimation of the dynamic probit model, the key condition that is needed will turn out to be

$$E(y_n|x_n, y_{n-1}, y_{n-2}, \ldots) = \Phi(\sum_{j=1}^{p} \rho_j y_{n-j} + \gamma' x_n), \tag{3}$$

while in the smoothed maximum score setting, we will need the condition

$$\text{Median}(u_n|y_{n-1}, \ldots, y_{n-p}, x_n) = 0. \tag{4}$$

Therefore, this paper analyzes the dynamic time series binary choice model at a level of generality that is comparable to the level of generality at which linear dynamic time series models can be analyzed.

Manski (1975) uses the sign function to develop the first semiparametric estimator for the binary choice model. Cosslett (1983) and Ichimura (1993) derive alternative estimators for the binary choice model. Imbens (1992) and Matzkin (1992) also develop estimators for the semiparametric binary choice model. Finally, in his seminal paper, Horowitz (1992) smooths the sign function of Manski (1975, 1985) and derives an estimator that is asymptotically normally distributed. However, all these estimators assume that one has a random sample. Thus, none of these estimators allows for lagged dependent explanatory variables. Park and Phillips (2000) assume that one of the regressors in a binary choice model is integrated, and they assume that all regressors are exogenous, thereby excluding predetermined variables and lagged y_n as possible regressors. Other recent papers that consider multinomial choice models in the presence of an integrated regressor are Hu and Phillips (2004) and Moon (2004).

In this paper we consider the binary choice model in a time series setting and we allow for lagged dependent variables and predetermined regressors as explanatory variables. For the semiparametric case, we only impose a median assumption. Thus, we allow the variance (and other moments of the error distribution) to

depend on lagged error terms, lagged dependent variables as well as regressors. Moreover, the median assumption allows for heterogeneity that is caused by random coefficients, e.g. a data generating process whose parameters are random and symmetrically distributed around $(\rho', \gamma')'$.
Binary choice models in a time series context have been used in a number of settings. Eichengreen, Watson, and Grossman (1985) considered a dynamic binary choice model

$$y_n = I(y_t^* > 0) \quad \text{where} \quad \phi(L)y_n^* = \beta' x_n + \varepsilon_n.$$

The stationarity of this model simply requires that the roots of $\phi(z)$ are outside the unit circle. This model does not allow for the actual state y_n to impact y_n and we can view y_n as a form of imperfect measurement of y_n^*. In a setting where y_n represents the presence in time period n of an intervention by a monetary authority, or a change in the Federal Funds rate, it seems natural to presume that lagged values of y_n will have a direct impact on y_n, which is ruled out by this specification. Dueker (1997) discusses how US recessions can be predicted using the slope of the yield curve and estimates the dynamic binary choice model that is considered in this paper. Kauppi and Saikkonen (2005) use binary choice models to predict recessions in the US and note that lagged dependent variables improce the predictions. Hahn and Kuersteiner (2005) use the geometric mixing results derived in this paper in a panel data logit model to show asymptotic properties of their estimator. Other work on discrete-valued dependent variables is for example Hamilton and Jorda (2002), who analyzed how the Federal Reserve determines the level of the Federal Funds target, and the Autoregressive Conditional Multinomial (ACM) model of Russell and Engle (1998). Furthermore, Ruud (1981) and Poirier and Ruud (1988) have considered the probit model with correlated errors. Robinson (1982) considered the tobit model with correlated errors.
However, no formal stationarity properties for dynamic probit models are derived in these papers, nor anywhere else in the literature as far as the authors are aware. Other potential applications include finance models concerning the likelihood of a financial transaction in a given time period as well as models concerning labor market participation decisions in which the relative importance of wealth versus welfare effects are studied.
The setup of this paper is as follows. In Section 2, the weak dependence properties of y_n are analyzed. Section 3 of this paper will analyze the dynamic probit procedure when lagged values of y_n have been included among the regressors and normality of u_n is assumed. In Section 4, we consider consistency of the smoothed maximum score estimator of the dynamic time series binary choice model. The smoothed maximum score estimator was first suggested in Horowitz (1992). Section 5 establishes asymptotic normality of the smoothed maximum score estima-

tor[1].

2 Properties of the dynamic time series binary choice model

A key aspect of the analysis below is to show that y_n satisfies the appropriate "fading memory" property when generated through a general dynamic binary choice model with regressors and possibly correlated errors. For the analysis of the smoothed maximum score estimator, this "fading memory" property that is proven for y_n needs to be strong enough to allow a proof of an equivalent of the Hoeffding inequality, and in addition, it needs to allow for a proof of a central limit theorem (CLT) for a function of y_n and x_n that depends on N in a situation where no martingale difference CLT can be applied. For a proof of validity of the dynamic probit model, the "fading memory" property only needs to support laws of large numbers and uniform laws of large numbers.

The "fading memory" property that we will prove for y_n is that of *near epoch dependence*. The idea of the proof is similar to that of proofs for showing fading memory properties of processes y_n of the form

$$y_n = f(y_{n-1}) + \varepsilon_n, \tag{5}$$

where $f(.)$ is such that $|f(x) - f(y)| \leq L|x - y|$ for some $L < 1$. Functions $f(.)$ satisfying this condition are called contraction mappings. Such proofs can be found in Bierens (1981) and Pötscher and Prucha (1997), for example. Pötscher and Prucha (1997, Section 6.4) contains a thorough discussion of these types of results, but the approach in the proof of this paper is different from the techniques discussed there. The differences are that the $f(.)$ function in the dynamic binary choice case is not continuous, depends on ε_n, is not strictly less than 1, and depends on more than one lagged value of y_n. These problems are essentially solved by smoothing the response function by the expectations operator, by using the fact that y_n is a binary random variable, and by the use of the appropriate metric on the arguments of the $f(.)$ function.

Near epoch dependence of random variables y_n on a base process of random variables η_n is defined as follows:

Definition 1 *Random variables y_n are called* near epoch dependent on η_n *if*

$$\sup_{n \in \mathbb{Z}} E|y_n - E(y_n|\eta_{n-m}, \eta_{n-m+1}, \ldots, \eta_n)|^2 = \nu(m)^2 \to 0 \quad \textit{as} \quad m \to \infty. \tag{6}$$

[1] In addition, some corrections to Horowitz' proof of the validity of the smoothed maximum score procedure are provided.

The idea behind the near epoch dependence condition is that given the last m error terms η_n, y_n should be predictable up to arbirary accuracy. The base process η_n needs to satisfy a condition such as strong or uniform mixing or independence. For the definitions of strong (α-) and uniform (ϕ-) mixing see e.g. Gallant and White (1988, p. 23) or Pötscher and Prucha (1997, p. 46). The near epoch dependence condition functions as a device that allows approximation of y_n by a function of finitely many mixing or independent random variables η_n. An intuitive explanation of the NED concept is that for large m, the conditional expectation of y_n given the last m elements of the base process η_n is close to y_n.

Note also that for strictly stationary (y_n, η_n), the "sup" in the above definition can be removed, because in that case

$$E|y_n - E(y_n|\eta_{n-m}, \eta_{n-m+1}, \ldots, \eta_n)|^2 \tag{7}$$

does not depend on n. The reader is referred to Gallant and White (1988) for a detailed account of the near epoch dependence condition. See also Pötscher and Prucha (1997) for a more up-to-date treatment of dependence concepts such as near epoch dependence.

The main results of the paper are the conditions under which y_n is stationary and near epoch dependent (Theorem 1) and the conditions under which y_n is strong mixing (Theorem 2). Unlike the linear model autoregressive model, no restrictions on the parameter space are needed for stationarity, near epoch dependence or strong mixing.

For establishing near epoch dependence of y_n, we have the following result. Define S as the set of all 2^p possible p -vectors s such that its elements s_i are 0 or 1, and define

$$\mathbf{\Phi} = \{\phi : \phi = \sum_{i=1}^{p} \rho_i s_i, s \in S\}. \tag{8}$$

Let ϕ_{min} denote the smallest element of $\mathbf{\Phi}$, and let ϕ_{max} denote the largest element.

Theorem 1 *Consider the model $y_n = I(\sum_{j=1}^{p} \rho_j y_{n-j} + \eta_n > 0)$. Let η_n be strong mixing and strictly stationary. Assume that there is some $\delta > 0$ and a positive integer K such that*

$$P(\phi_{max} + \max_{i=1,\ldots,p} \eta_{n-i} > 0|\eta_{n-p-K}, \eta_{n-p-K-1}, \ldots)$$

$$-P(\phi_{min} + \min_{i=1,\ldots,p} \eta_{n-i} > 0|\eta_{n-p-K}, \eta_{n-p-K-1}, \ldots) < 1 - \delta \qquad \text{almost surely.} \tag{9}$$

Then there exists a strictly stationary solution $y_n = f(\eta_n, \eta_{n-1}, \ldots)$ to the model that is near epoch dependent on η_n, and its near epoch dependence sequence $\nu(.)$ satisfies $\nu(m) \leq C_1 \exp(-C_2 m)$ for positive constants C_1 and C_2. Also, $f(\eta_n, \eta_{n-1}, \ldots)$ is unique in the sense that if $g(\eta_n, \eta_{n-1}, \ldots)$ is also a strictly stationary solution to the model, then $f(\eta_n, \eta_{n-1}, \ldots) = g(\eta_n, \eta_{n-1}, \ldots)$ almost surely.

Note that if $\eta_n = \gamma' x_n + u_n$ for strong mixing and strictly stationary (x_n', u_n), clearly η_n is mixing as well. This observation will be used below.

The formulation of the above theorem can be compared to what can be shown for a simple AR(1) model $y_n = \rho y_{n-1} + \eta_n$. In that case, $\sum_{j=0}^{\infty} \rho^j \eta_{n-j} + C\rho^n$ will solve the model for any value of C, but the strictly stationary solution is obtained for $C = 0$. In addition, for our model our uniqueness statement has to rule out forward-looking solutions that are functions of $\eta_{n+1}, \eta_{n+2}, \ldots$, which are also possible.

The assumption of Equation (9) limits the predictability of y_n given the distant past. If η_n is M-dependent (i.e. η_n and η_{n+M} are independent for some value of M), then the required condition is

$$P(\phi_{max} + \max_{i=1,\ldots,p} \eta_{n-i} > 0) - P(\phi_{min} + \min_{i=1,\ldots,p} \eta_{n-i} > 0) < 1 - \delta. \tag{10}$$

This condition is implied by the assumption of full support on $\mathbb{R}^p$ for $(\eta_{n-1}, \ldots, \eta_{n-p})$. It also follows from the definition of uniform mixing that the assumption of Equation (10) suffices for the verification of the assumption of Equation (9) if η_n is uniform mixing. Furthermore, if $p = 1$ and η_n is an invertible MA(∞) process, i.e. $\eta_n = \sum_{i=0}^{\infty} b_i \zeta_{n-i}$ where b_i is deterministic and the ζ_{n-i} are i.i.d., then we have

$$\begin{aligned}
&P(\phi_{max} + \eta_{n-1} > 0 | \eta_{n-2}, \eta_{n-3}, \ldots) - P(\phi_{min} + \eta_{n-1} > 0 | \eta_{n-2}, \eta_{n-3}, \ldots) \\
&= P(\phi_{max} + b_0 \zeta_{n-1} > \sum_{i=1}^{\infty} b_i \zeta_{n-1-i} | \zeta_{n-2}, \zeta_{n-3}, \ldots) \\
&\qquad - P(\phi_{min} + b_0 \zeta_{n-1} > \sum_{i=1}^{\infty} b_i \zeta_{n-i} | \zeta_{n-2}, \zeta_{n-3}, \ldots) \\
&\leq \sup_{x \in \mathbb{R}} |F((\phi_{max} - x)/b_0) - F((\phi_{min} - x)/b_0)|,
\end{aligned}$$

where $F(\cdot)$ denotes the distribution of ζ_n, and the latter expression is less than 1 if $b_0 \neq 0$, $F(\cdot)$ is strictly increasing, and $\phi_{max} > \phi_{min}$.

The proof of Theorem 1 is substantially easier for the case where η_n is i.i.d., only one lagged y_n is used as regressor, and no other regressors are included. In that case, we can write

$$y_n = y_{n-1} I(\rho_1 + \eta_n > 0) + (1 - y_{n-1}) I(\eta_n > 0), \tag{11}$$

implying that

$$
\begin{aligned}
\nu_m &\equiv \sup_{n\in\mathbb{Z}} E|y_n - E(y_n|\eta_{n-m},\ldots,\eta_n)|^2 \\
&= \sup_{n\in\mathbb{Z}} E|(I(\rho_1+\eta_n>0)-I(\eta_n>0))(y_{n-1}-E(y_{n-1}|\eta_{n-m},\ldots,\eta_{n-1})|^2 \\
&= |P(\rho_1+\eta_n>0)-P(\eta_n>0)|\sup_{n\in\mathbb{Z}} E|y_{n-1}-E(y_{n-1}|\eta_{(n-1)-(m-1)},\ldots,\eta_{n-1})|^2 \\
&= |P(\rho_1+\eta_n>0)-P(\eta_n>0)|\nu_{m-1}, \qquad (12)
\end{aligned}
$$

which implies that the $\nu(m)$ sequence decays geometrically under the condition of Equation (9). The proof of Theorem 1 should be viewed as an extension to the above reasoning.
The fact that y_n is a 0/1-valued near epoch dependent random variable can now be exploited to show that (y_n, x_n') is also strong mixing. Note that this is an observation that apparently has not been made in the literature before. The result is as follows:

Theorem 2 *Suppose that $y_n = f(\eta_n, \eta_{n-1}, \ldots)$ is a sequence of 0/1-valued random variables that is near epoch dependent on (u_n, x_n') with near epoch dependence coefficients $\nu(m)$, where $\eta_n = \gamma' x_n + u_n$ and $(u_n, x_n')'$ is strong mixing with mixing coefficients $\alpha(m)$. Then $(y_n, x_n')'$ is strong mixing with strong mixing coefficients is an invertible linear MA(∞) process and $p = 1$, i.e. the case $N(0,1)$, we have $C(\nu(m)+\alpha(m))$ for some $C > 0$.*

The mixing property of (y_n, x_n') will be used in the proofs for consistency and asymptotic normality of the next sections.

3 The dynamic probit model

This section examines the behavior of the dynamic probit model estimator that results from including lagged y_n among the regressors. Let $\beta = (\rho', \gamma')'$ denote the true parameter and let $b = (r', c')'$, $\rho, r \in \mathbb{R}^p$ and $\gamma, c \in \mathbb{R}^q$, and let R and Γ denote the parameter spaces for r and c respectively, and let $B = R \times \Gamma$. We assume normality of the errors so that the normalized loglikelihood conditional on $y_1, ..., y_p$ has the following form,

$$
L_N(b) = (N-p)^{-1} \sum_{n=p+1}^{N} l_n(b)
$$

$$
= (N-p)^{-1} \sum_{n=p+1}^{N} [y_n \log(\Phi(\sum_{j=1}^{p} r_j y_{n-j} + c' x_n)) + (1-y_n)\log(1-\Phi(\sum_{j=1}^{p} r_j y_{n-j} + c' x_n))]. \quad (13)
$$

Given the result of Theorem 2, it is now straightforward to find standard conditions under which the maximum likelihood estimator b_N^{ML} is consistent.

Assumption A

1. x_n is a sequence of strictly stationary strong mixing random variables with α-mixing numbers $\alpha(m)$, where $x_n \in \mathbb{R}^q$ for $q \geq 0$ and $\gamma \in \mathbb{R}^q$, and the second absolute moment of x_n exits. The distribution of $w_n = (x_n', y_{n-1}, \ldots, y_{n-p})'$ is not contained in any linear subspace of $\mathbb{R}^q$.

2. $u_n | x_n, y_{n-1}, y_{n-2} \ldots, y_{n-p} \sim$ iid $N(0,1)$.

3. $y_n = I(\sum_{i=1}^p \rho_i y_{n-i} + \gamma' x_n + u_n > 0)$.

4. β is an element of the interior of a convex set B.

The assumption that the distribution of w_n is not contained in a linear subspace of $\mathbb{R}^q$ is used in Manski (1975, 1985) and is equivalent to the assumption that $Ew_n w_n'$ has full rank.

Theorem 3 *Under Assumption A, $b_N^{ML} \xrightarrow{p} \beta$. If in addition (i) the strong mixing coefficients satisfy $\alpha(m) \leq Cm^{-\eta}$ for positive constants C and η and (ii) $E|l_n(b)|^{1+\delta} < \infty$ for some $\delta > 0$ and all $b \in B$, and (iii) B is compact, then $b_N^{ML} \xrightarrow{as} \beta$.*

Let $I = -E(\partial/\partial b)(\partial/\partial b')l_n(\beta)$. For asymptotic normality, we need an additional assumption.

Assumption B

1. $u_n | (x_n, y_{n-1}), (x_{n-1}, y_{n-2}) \ldots \sim$ iid $N(0,1)$.

Theorem 4 *Under Assumptions A and B, $N^{1/2}(b_N^{ML} - \beta) \xrightarrow{d} N(0, I^{-1})$.*

Under the above Assumptions A and B, it also follows that the usual estimators of I, using either the outer product or Hessian approach, will both be weakly consistent for I.

Note that given the weak dependence property of Theorem 2, it is also possible to set forth conditions such that for weakly dependent u_n with arbitrary distribution, $N^{1/2}(b_N^{ML} - \beta^*) \xrightarrow{d} N(0, J)$ for some matrix J and a β^* that uniquely minimizes the objective function. Here of course β^* does not necessarily equal the true parameter value β. However, in order to show that the probit objective function is uniquely maximized at β, we need that a first order condition of the type

$$E(y_n - \Phi(\sum_{i=1}^{p} \rho_i y_{n-i} + \gamma' x_n))m(y_{n-1}, \ldots, y_{n-p}, x_n) = 0 \tag{14}$$

holds for some function $m(., \ldots, .)$. This condition is implied by

$$E(y_n | y_{n-1}, \ldots, x_n) = \Phi(\sum_{i=1}^{p} \rho_i y_{n-i} + \gamma' x_n), \tag{15}$$

and the latter condition is equivalent to assuming that u_n is i.i.d. and standard normal if lagged values of y_n are included.

4 Consistency of the smoothed maximum score estimator

The smoothed maximum score estimator is defined as $\text{argmax}_{b \in B} S_N(b, \sigma_N)$, where

$$S_N(b, \sigma_N) = (N-p)^{-1} \sum_{n=p+1}^{N} (2 \cdot I(y_n = 1) - 1) K((\sum_{j=1}^{p} r_j y_{n-j} + c' x_n)/\sigma_N) \tag{16}$$

and σ_N is a bandwidth-type sequence such that $\sigma_N \to 0$ as $N \to \infty$, where $K(.)$ is a function such that $K(-\infty) = 0$ and $K(\infty) = 1$. This objective function is a smoothed version of the maximum score objective function

$$S_N^*(b) = (N-p)^{-1} \sum_{n=p+1}^{N} (2 \cdot I(y_n = 1) - 1) I(\sum_{j=1}^{p} r_j y_{n-j} + c' x_n \geq 0). \tag{17}$$

In addition, let $S(b) = ES_N^*(b)$. This notation is justified because we will use conditions under which (y_n, x_n) will be proven to be strictly stationary. See Manski (1985) and Kim and Pollard (1990) for more information and results regarding the maximum score estimator.

Horowitz' maximum score estimator can reach the optimal rate of convergence (Horowitz (1992, 1993)). Kim and Pollard (1990) showed that the maximum score estimator in general is consistent of order $N^{-1/3}$, the optimal rate for that model.

The following five assumptions are needed for the proof of our consistency result:

Assumption 1 *$(x_n', u_n)'$ is a sequence of strictly stationary strong mixing random variables with α-mixing numbers $\alpha(m)$, where $x_n \in \mathbb{R}^q$ for $q \geq 1$ and $\gamma \in \mathbb{R}^q$, and*

$$y_n = I(\sum_{i=1}^{p} \rho_i y_{n-i} + \gamma' x_n + u_n > 0). \tag{18}$$

The following assumption is simply the assumption of Equation (9) in Theorem 1 for $\eta_n = \gamma' x_n + u_n$.

Assumption 2 *For ϕ_{max} and ϕ_{min} as defined before, for some $\delta > 0$ there exists a positive integer K such that*

$$\begin{aligned} &P(\phi_{max} + \max_{i=1,\ldots,p}(\gamma' x_{n-i} + u_{n-i}) > 0 | y_{n-p-K}, y_{n-p-K-1}, \ldots) \\ &\quad -P(\phi_{min} + \min_{i=1,\ldots,p}(\gamma' x_{n-i} + u_{n-i}) > 0 | y_{n-p-K}, y_{n-p-K-1}, \ldots) < 1 - \delta. \end{aligned} \tag{19}$$

By Theorem 1 and the discussion following that theorem, $(y_n, x_n)'$ is strictly stationary under the above two assumptions. This justifies the formulation of the assumptions below in their current forms[2]. Define x_{nj} as the elements of x_n, i.e. $x_n' = (x_{n1}, x_{n2}, \ldots, x_{nq})'$, and define $\tilde{x}_n = (y_{n-1}, \ldots, y_{n-p}, x_{n2}, \ldots, x_{nq})$.

Assumption 3 *The support of the distribution of $(x_{n1}, \tilde{x}_n')'$ is not contained in any proper linear subspace of $\mathbb{R}^{p+q}$. (b) $0 < P(y_n = 1 | x_{n1}, \tilde{x}_n) < 1$ almost surely. (c) $\gamma_1 \neq 0$, and for almost every $\tilde{x}_n$, the distribution of x_{n1} conditional on $\tilde{x}_n$ has everywhere positive density with respect to Lebesgue measure.*

Assumption 4 $\text{Median}(u_n | x_n, y_{n-1}, \ldots, y_{n-p}) = 0$ almost surely.

Assumption 4 allows for heteroskedasticity of arbitrary form, including heteroskedasticity that depends on lagged values of y_n. If all regressors are exogenous, Assumption 4 allows for correlated errors, e.g. the errors could follow an ARMA process.

Assumption 5 *$|\gamma_1| = 1$, and $\tilde{\beta} = (\rho_1, \ldots, \rho_p, \gamma_2, \ldots, \gamma_q)$ is contained in a compact subset $\tilde{B}$ of $\mathbb{R}^{p+q-1}$.*

We need some form of scale normalization; we set $|b_1| = 1$ here, as in Horowitz (1992). Therefore, the estimator b_N is defined as

$$b_N = \text{argmax}_{b:|b_1|=1} S_N(b, \sigma_N). \tag{20}$$

The following result shows the consistency of b_N:

Theorem 5 *Under Assumptions 1,3,4,5 and 2, $b_N \xrightarrow{p} \beta$. If in addition the strong mixing coefficients satisfy $\alpha(m) \leq Cm^{-\eta}$ for positive constants C and η, then $b_N \xrightarrow{as} \beta$.*

[2]Assumptions 1-5 imply the assumptions of Theorem 1

5 Asymptotic normality of the smoothed maximum score estimator

Define, analogously to Horowitz (1992), $\tilde{b} = (r_1, \dots, r_p, c_2, \dots, c_q)$, and let

$$T_N(b, \sigma_N) = \partial S_N(b, \sigma_N)/\partial \tilde{b}, \tag{21}$$

$$Q_N(b, \sigma_N) = \partial^2 S_N(b, \sigma_N)/\partial \tilde{b} \partial \tilde{b}'. \tag{22}$$

Also, define

$$z_n = \sum_{j=1}^{p} \rho_j y_{n-j} + \gamma' x_n, \tag{23}$$

and let $p(z_n|\tilde{x}_n)$ denote the density of z_n given $\tilde{x}_n$, let $P(.)$ denote the distribution of $\tilde{x}_n$, let $F(.|z_n, \tilde{x}_n)$ denote the cumulative distribution of u_n conditional on z_n and $\tilde{x}_n$. For each positive integer i, define

$$F^{(i)}(-z, x, \tilde{x}) = \partial^i F(-z|z, \tilde{x})/\partial z^i \tag{24}$$

Let h denote a positive integer that satisfies the conditions of Assumptions 8, 9 and 10 below, and let

$$\alpha_A = \int_{-\infty}^{\infty} v^h K'(v) dv \tag{25}$$

$$\alpha_D = \int_{-\infty}^{\infty} K'(v)^2 dv. \tag{26}$$

Also analogously to Horowitz (1992), define

$$A = -2\alpha_A \sum_{i=1}^{h} \{[i!(h-i)!]^{-1} E[F^{(i)}(0, 0, \tilde{x}_n) p^{(h-i)}(0|\tilde{x}_n)\tilde{x}_n]\}, \tag{27}$$

$$D = \alpha_D \cdot E[\tilde{x}_n \tilde{x}_n' p(0|\tilde{x}_n)], \tag{28}$$

$$Q = 2 \cdot E[\tilde{x}_n \tilde{x}_n' F^{(1)}(0|0, \tilde{x}_n) p(0|\tilde{x}_n)]. \tag{29}$$

The following assumption is the analogue of Horowitz' Assumption 5, which is the assumption below for $s = 4$. It appears that Horowitz' truncation argument is in error (see also notes 2, 3, 4 and 5), but that his argument is correct for bounded data. This explains the presence here of a condition that is stronger than that of Horowitz.

Assumption 6 *For all vectors ξ such that $|\xi| = 1$, $E|\xi' \tilde{x}|^s < \infty$ for some $s > 4$.*

We need to strengthen the fading memory conditions of Assumption 1 in order to establish asymptotic normality:

Assumption 1' *(x_n', u_n) is a sequence of strictly stationary strong mixing random variables with α-mixing numbers $\alpha(m)$ such that $\alpha(m) \leq Cm^{-(2s-2)/(s-2)-\eta}$ for some $\eta > 0$, where $x_n \in \mathbb{R}^q$ for $q \geq 1$ and $\gamma \in \mathbb{R}^q$, and*

$$y_n = I(\sum_{i=1}^{p} \rho_i y_{n-i} + \gamma' x_n + u_n > 0). \tag{30}$$

The assumption below is needed in lieu of Horowitz' Assumption 6.

Assumption 7 *For some sequence $m_N \geq 1$,*

$$\sigma_N^{-3(p+q-1)} \sigma_N^{-2} N^{1/s} \alpha(m_N) + \sigma_N^{-2(p+q-1)/\beta} N^{2/s} \alpha(m_N) \\ + |\log(N m_N)| (N^{1-4/s} \sigma_N^4 m_N^{-2})^{-1} \to 0 \quad \text{as} \quad N \to \infty. \tag{31}$$

For the case of independent (x_n, u_n), $\alpha(m) = 0$ for $m \geq 1$, and we can set $m_N = 1$ for that case. The condition of Assumption 7 then becomes

$$(\log(N))(N^{1-4/s} \sigma_N^4)^{-1} \to 0 \quad \text{as} \quad N \to \infty, \tag{32}$$

implying that for bounded data, we can set $s = \infty$ and obtain Horowitz' condition

$$(\log(N))(N \sigma_N^4)^{-1} \to 0 \quad \text{as} \quad N \to \infty. \tag{33}$$

The following assumptions are identical to Horowitz' Assumptions 7-11:

Assumption 8 *(a) $K(.)$ is twice differentiable everywhere, $|K(.)|$ and $K''(.)$ are uniformly bounded, and each of the following integrals over $(-\infty, \infty)$ is finite: $\int [K'(v)]^4 dv$, $\int [K''(v)]^2 dv$, $\int |v^2 K''(v)| dv$. (b) For some integer $h \geq 2$ and each integer i $(1 \leq i \leq h)$, $\int |v^i K'(v)| dv < \infty$, and*

$$\int_{-\infty}^{\infty} v^i K'(v) dv = \begin{cases} 0 & \text{if } i < h, \\ d & \text{(nonzero) if } i = h. \end{cases} \tag{34}$$

(c) For any integer i between 0 and h, any $\eta > 0$, and any sequence $\{\sigma_N\}$ converging to 0,

$$\lim_{N \to \infty} \sigma_N^{i-h} \int_{|\sigma_N v| > \eta} |v^i K'(v)| dv = 0 \tag{35}$$

and

$$\lim_{N \to \infty} \sigma_N \int_{|\sigma_N v| > \eta} |K''(v)| dv = 0. \tag{36}$$

Assumption 9 *For each integer i such that* $1 \leq i \leq h-1$*, all z in a neighborhood of 0, almost every* $\tilde{x}_n$*, and some* $M < \infty$*,* $p^{(i)}(z_n|\tilde{x}_n)$ *exists and is a continuous function of* z_n *satisfying* $|p^{(i)}(z_n|\tilde{x}_n)| < M$*. In addition,* $p(z_n|\tilde{x}_n) < M$ *for all z and almost every* $\tilde{x}$*.*

Assumption 10 *For each integer i such that* $1 \leq i \leq h$*, all* z_n *in a neighborhood of 0, almost every* $\tilde{x}_n$*, and some* $M < \infty$*,* $F^{(i)}(-z_n, z_n, \tilde{x}_n)$ *exists and is a continuous function of* z_n *satisfying* $|F^{(i)}(-z_n, z_n, \tilde{x}_n)| < M$*.*

Assumption 11 $\tilde{\beta}$ *is an interior point of* $\tilde{B}$*.*

Assumption 12 *The matrix Q is negative definite.*

In addition to the above equivalents to Horowitz' assumptions, we will also need the following two assumptions. The first assumption is needed to assure proper behavior of covariance terms.

Assumption 13 *The conditional joint density* $p(z_n, z_{n-j}|x_n, x_{n-j})$ *exists for all* $j \geq 1$ *and is continuous at* $(z_n, z_{n-j}) = (0, 0)$ *for all* $j \geq 1$*.*

The next condition on $K''(.)$ is needed to formally show a uniform law of large numbers for the second derivative of the objective function.

Assumption 14 $K''(.)$ *satisfies, for some* $\mu \in (0, 1]$ *and* $L \in [0, \infty)$ *and all* $x, y \in \mathbb{R}$*,*

$$|K''(x) - K''(y)| \leq L|x - y|^{\mu}. \tag{37}$$

To prove asymptotic normality, we need an inequality in the spirit of Hoeffding's inequality, but for weakly dependent random variables. We derive such an inequality in the Appendix as Lemma 10. The inequality of Lemma 10 also allows for martingale difference sequences so that it covers both the random sample case of Horowitz (1992) as well as the dynamic case.
Our asymptotic normality result now is the following. This result, of course, is nearly identical to Horowitz' in the non-dynamic cross-section case.

Theorem 6 *Let Assumptions 1' and Assumptions 3-14 hold for some* $h \geq 2$*. Then*

1. *If* $N\sigma_N^{2h+1} \to \infty$ *as* $N \to \infty$*,* $\sigma_N^{-h}(\tilde{b}_N - \tilde{\beta}) \xrightarrow{p} -Q^{-1}A$*.*

2. If $N\sigma_N^{2h+1}$ has a finite limit λ as $N \to \infty$,

$$(N\sigma_N)^{1/2}(\tilde{b}_N - \tilde{\beta}) \xrightarrow{d} N(-\lambda^{1/2}Q^{-1}A, Q^{-1}DQ^{-1}). \tag{38}$$

In order to estimate the matrices A, D and Q, we need an additional result, the analogue of Horowitz' (1992) Theorem 3.

Theorem 7 *Let b_N be a consistent smoothed maximum score estimator based on σ_N such that $\sigma_N = O(n^{-1/(2h+1)})$. For $b \in \{-1, 1\} \times \tilde{B}$, define*

$$t_n(b, \sigma) = (2 \cdot I(y_n = 1) - 1)(\tilde{x}_n/\sigma_N)K'((\sum_{j=1}^{p} r_j y_{n-j} + c'x_n)/\sigma). \tag{39}$$

Let σ_N^* be such that $\sigma_N^* = O(N^{-\delta/(2h+1)})$, where $0 < \delta < 1$. Then: (a) $\hat{A}_N \equiv (\sigma_N^*)^{-h}T_N(b_N, \sigma_N^*)$ converges in probability to A; (b) the matrix

$$\hat{D}_N \equiv \sigma_N(N-p)^{-1} \sum_{n=p+1}^{N} t_n(b_N, \sigma_N)t_n(b_N, \sigma_N)' \tag{40}$$

converges in probability to D; (c) $Q_N(b_N, \sigma_N)$ converges in probability to Q.

6 Simulations

In this section, we conduct a limited Monte Carlo simulation experiment in order to evaluate the performance of the maximum likelihood estimator and the smooth maximum score estimator for the dynamic binary choice model. In order to achieve an empirically relavant setup for conducting our simulations experiment we set our parameters as the estimates that were obtained from a probit regression as suggested in Kauppi and Saikkonen (2005). Kauppi and Saikkonen (2005) estimated the dynamic binary choice model and then draw several conclusions and make several predictions. We use the estimates to generate 2000 datasets and then study the performance of the maximum likelihood estimator and the smooth maximum score estimator. The dependent variable in Kauppi and Saikkonen (2005)'s setup is a dummy that indicates whether the US economy was in recession during a particular quarter. The explanatory variables are a dummy that indicates whether a recession occurred during the previous quarter, a dummy that indicates whether a recession occurred in the quarter before that, and the slope of the yield curve.

The slope of the yield curve is approximated by the difference between the yield of the ten year treasury bond and the yield of the three month treasury bill[3].

The first model that we consider corresponds to Table 1, column 2 of Kauppi and Saikkonen (2005) and assumes that the probability of a recession in the US depends on the slope of the yield curve and on whether there was a recession in the previous quarter. In particular,

$$P\left(y_t = 1 | y_{t-1}, x_{t-1}\right) = \Phi(\rho y_{t-1} + \beta_1 x_{t-1} + c),$$

where $\Phi(\cdot)$ is the standardnormal distribution function, y_{t-1} denotes a dummy that indicates whether there was a recession in the previous period, and x_{t-1} denotes the slope of the yield curve in the previous period. Estimating this model yields

Probit Model	Estimate	Std.Error
ρ	2.899	0.663
β_1	-0.496	0.253
c	-1.256	0.343

We now take the above estimates as parameters for our data-generating process, generate 2000 datasets, and estimate the parameters repeatedly. We find the following:

Probit Model	Mean estimate	Bias	RMSE
ρ	3.052	0.153	0.712
β_1	-0.549	-0.053	0.178
c	-1.309	-0.053	0.389

Also, the simulated recession probabilities closely matched the recession probabilities that were estimated in the sample. For the smoothed maximum score estimator, we find

SMSE	Estimate	Bias	RMSE
ρ	2.887	-0.0123	0.719
c	-1.235	0.020	0.513

We also calculated the median of the estimate, the median bias and the median

[3]The sample is 1981:01 - 2005:04 (100 time observations). The dummy variable for recessions and expansions has been constructed using the official NBER dating of US recessions (1 for periods of recession; 0 for periods of expansions). The variable y_t denotes the dummy variable constructed in this way; x_t is the interest rate spread and is constructed as the difference between the 10 year treasury bond yield and the 3 month treasury bill yield and averaged over three months.

absolute error.

Probit Model	Median estimate	Median bias	Median absolute error
ρ	2.947	0.048	0.349
β_1	-0.524	-0.028	0.092
c	-1.280	-0.024	0.219

Similarly, for the smoothed maximum score estimator, we find[4]

SMSE	Median estimate	Median bias	Median absolute error
ρ	2.860	-0.039	0.479
c	-1.220	0.036	0.350

The second model that we consider corresponds to Table 2, column 3 of Kauppi and Saikkonen (2005) and uses a lag of the yield curve; this specification assumes that the yield curve can help predicting recessions 6 months ahead. The model now is

$$P\left(y_t = 1|y_{t-1}, x_{t-2}\right) = \Phi\left(\rho y_{t-1} + \beta_1 x_{t-2} + c\right),$$

where x_{t-2} denotes the slope of the yield curve in period $t-2$. The parameter estimates now are

Probit Model	Estimate	Std.Error
ρ	2.987	0.711
β_1	-0.791	0.292
c	-1.140	0.344

Again taking these values as the parameters for our data-generating process, we generate 2000 datasets and find

Probit Model	Mean estimate	Bias	RMSE
ρ	3.24	0.253	0.862
β_1	-0.891	-0.100	0.273
c	-1.218	-0.078	0.437

and

Probit Model	Median estimate	Median bias	Median absolute error
ρ	3.097	0.110	0.438
β_1	-0.850	-0.059	0.131
c	-1.181	-0.041	0.250

[4]Note that the parameter vector is normalized such that $|\beta_1| = 1$.

For the smoothed maximum score estimator, we find

SMSE	Estimate	Bias	RMSE
ρ	2.970	-0.017	0.723

and

SMSE	Median estimate	Median bias	Median absolute error
ρ	2.960	-0.027	0.467

The simulations show that the estimators performs reasonably well in a realistic setting. The standard errors are also reasonably close to the RMSE that is derived using simulations. Given Theorem 1 and 2, one may have expected that the performance of the estimators is reasonably close to the performance of these estimators in an i.i.d. setting; after all, the data generating process has exponentially decreasing near epoch dependence sequence, which suggests that the dependence properties of the process are not severe. The simulations support this view.

7 Conclusions

This paper proves near epoch dependence and strong mixing for the dynamic binary choice model with correlated errors. Using this result, it shows in a time series setting the validity of the dynamic probit likelihood procedure when lags of the dependent binary variable are used as regressors, and it establishes the asymptotic validity of Horowitz' (1992) smoothed maximum score estimation of dynamic binary choice models with lags of the dependent variable as regressors. For the semiparametric model, the latent error is allowed to be correlated. Unlike the linear autoregressive model, no restrictions on the parameter space are needed for the stationarity, near epoch dependence or strong mixing properties of the data.

References

Andrews, D.W.K. (1987), Consistency in nonlinear econometric models: a generic uniform law of large numbers, *Econometrica* 55, 1465-1471.

Andrews, D.W.K. (1988), Laws of large numbers for dependent non-identically distributed random variables, *Econometric Theory* 4, 458-467.

Azuma, K. (1967), Weighted sums of certain dependent random variables, *Tokohu*

Mathematical Journal 19, 357-367.

Bierens, H.J. (1981) *Robust methods and asymptotic theory in nonlinear econometrics.* New York: Springer-Verlag.

Bierens, H. J. (2004) *Introduction to the mathematical and statistical foundations of Econometrics.* Cambridge University Press, forthcoming, available at `http://econ.la.psu.edu/~hbierens/CHAPTER7.PDF`.

Cosslett, S. R. (1983), Distribution-free maximum likelihood estimator of the binary choice model, *Econometrica* 51, 765-782.

Davidson, J. (1994) *Stochastic limit theory.* Oxford: Oxford University Press.

de Jong, R.M. (1995), Laws of large numbers for dependent heterogeneous processes, *Econometric Theory* 11, 347-358.

de Jong, R.M. (1997), Central limit theorems for dependent heterogeneous random variables, *Econometric Theory* 13, 353-367.

Dueker, M. J. (1997), Strengthening the case for the yield curve as a predictor of U.S. recessions, *Federal Reserve Bank of St. Louis Review in Business & Finance*, vol. 2, 41-51.

Eichengreen, B., Watson, M. and R. Grossman (1985), Bank rate policy under the interwar gold standard: a dynamic probit model, *Economic Journal* 95, 725-745.

Gallant, A.R. and H. White (1988) *A unified theory of estimation and inference for nonlinear dynamic models.* New York: Basil Blackwell.

Hahn, J. and G. Kuersteiner (2005), Bias Reduction for dynamic nonlinear panel models with fixed effects, Boston University working paper.

Horowitz, J. (1992), A smoothed maximum score estimator for the binary response model, *Econometrica* 60, 505-531.

Hu, L. and P.C.B. Phillips (2004), Nonstationary discrete choice, *Journal of Econometrics* 120, 103-138.

Ichimura, I. (1993), Semiparametric least squares (SLS) and weighted SLS estimation of single-index models, *Journal of Econometrics* 58, 71-120.

Imbens, G. W. (1992), An efficient method of moment estimator for discrete choice models with choice-based sampling, *Econometrica* 60, 1187-1214.

Kauppi, H. and P. Saikkonen (2005), Predicting U.S. recession with dynamic response models, unpublished manuscript, University of Helsinki.

Kim, J., and D. Pollard (1990), Cube root asymptotics, *Annals of Statistics* 18, 191-219.

Manski, C.F. (1975), Maximum score estimation of the stochastic utility model of choice, *Journal of Econometrics* 3, 205-228.

Manski, C.F. (1985), Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator, *Journal of Econometrics* 27, 313-333.

Matzkin, R.L. (1992), Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models, *Econometrica* 60, 239-270.

McLeish, D. L. (1974), Dependent central limit theorems and invariance principles, *Annals of Probability* 2, 620-628.

Moon, H.R. (2004), Maximum score estimation of a nonstationary binary choice model, *Journal of Econometrics* 120, 385-403.

Newey, W. K., and D. McFadden (1994), Large sample estimation and hypothesis testing, in *Handbook of Econometrics,* Vol. 4, ed. by R. F. Engle and D. MacFadden. Amsterdam: North-Holland.

Park, Y., and P.C.B. Phillips (2000), Nonstationary binary choice, *Econometrica* 68, 1249-1280.

Poirier and Ruud (1988), Probit with dependent observations, *Review of Economic Studies* 55, 593-614.

Pötscher, B.M. and I.R. Prucha (1997) *Dynamic nonlinear econometric models.* Berlin: Springer-Verlag.

Robinson, P.M. (1982), On the asymptotic properties of estimators of models containing LDV, *Econometrica* 50, 27-41.

Ruud, P. (1981), Conditional minimum distance estimation and autocorrelation in limited dependent variable models, Chapter 3 of Ph.D. thesis, Department of Economics, MIT.

White, H. (2001) *Asymptotic theory for Econometricians.* New York: Academic Press.

Wooldridge, J. (1994), Estimation and inference for dependent processes, in *Handbook of Econometrics*, volume 4, ed. by R. F. Engle and D. MacFadden. Amsterdam: North-Holland.

Proofs

Proof of Theorem 1:

The dynamic binary choice model of order p can be written as

$$y_n = I(\sum_{i=1}^{p} \rho_i y_{n-i} + \eta_n > 0) = g(y_{n-1}, y_{n-2}, \ldots, y_{n-p}, \eta_n).$$

This $g(., \ldots, .)$ satisfies, for all 0-1 valued $y_1, y_2, \ldots, y_{n-p}$ and $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{n-p}$,

$$|g(y_1, y_2, \ldots, y_p, \eta_n) - g(\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_p, \eta_n)| \leq L(\eta_n) \max_{j=1,\ldots,p} |y_j - \tilde{y}_j|,$$

where

$$L(\eta_n) = \sup_{\phi, \phi' \in \mathbf{\Phi}} |I(\phi + \eta_n > 0) - I(\phi' + \eta_n > 0)|$$

and $\mathbf{\Phi}$ was defined in Equation (8). The idea of the proof is to show that the process y_n can be approximated arbitrarily well by using a function of a finite number of η_n - this is the content of the near epoch dependence concept. We do this by using for our approximation $\hat{y}_n^m$ the y that would have resulted if the process had been started up using 0 values for the y_n and η_n that occurred m periods or longer ago. Formally, for all n define $\hat{y}_n^m = 0$ for $m \leq 0$. Then for all $m \geq 1$ recursively define

$$\hat{y}_n^m = g(\hat{y}_{n-1}^{m-1}, \hat{y}_{n-2}^{m-2}, \ldots, \hat{y}_{n-p}^{m-p}, \eta_n).$$

Note that by construction, $\hat{y}_n^m = f_m(\eta_n, \eta_{n-1}, \ldots, \eta_{n-m})$. Define $\max_{j \in A} c_j = 0$ if A is empty. Next, note that by assumption there exists positive integer K and $\delta > 0$, such that for ϕ_{max} and ϕ_{min} as defined below Equation (8),

$$|E(\max_{j=1,\ldots,p} L(\eta_{n-j+1}) | \eta_{n-pK}, \eta_{n-pK-1}, \ldots)|$$

$$\leq |P(\phi_{max} + \max_{j=1,\ldots,p} \eta_{n-j+1} > 0 | \eta_{n-pK}, \eta_{n-pK-1}, \ldots)$$

$$-P(\phi_{min} + \min_{j=1,\ldots,p} \eta_{n-j+1} > 0 | \eta_{n-pK}, \eta_{n-pK-1}, \ldots)| < 1 - \delta$$

for some $\delta > 0$. Then for the approximators $\hat{y}_n^m$ we have, using $0 \leq L(\cdot) \leq 1$, for any $k \geq 0$,

$$\max_{j=1,\ldots,p} |\hat{y}_{n-j+1}^{m+k-j-1} - \hat{y}_{n-j+1}^{m-j+1}|$$

$$= \max(|g(\hat{y}_{n-1}^{m+k-1}, \hat{y}_{n-2}^{m+k-2}, \ldots, \hat{y}_{n-p}^{m+k-p}, \eta_n) - g(\hat{y}_{n-1}^{m-1}, \hat{y}_{n-2}^{m-2}, \ldots, \hat{y}_{n-p}^{m-p}, \eta_n)|, \max_{j=2,\ldots,p} |\hat{y}_{n-j+1}^{m+k-j+1} - \hat{y}_{n-j+1}^{m-j+1}|)$$

$$\leq \max(L(\eta_n) \max_{j=1,\ldots,p} |\hat{y}_{n-j}^{m-j+k} - \hat{y}_{n-j}^{m-j}|, \max_{j=2,\ldots,p} |\hat{y}_{n-j+1}^{m-j+1+k} - \hat{y}_{n-j+1}^{m-j+1}|)$$

$$\leq \max(L(\eta_n)|\hat{y}_{n-p}^{m-p+k} - \hat{y}_{n-p}^{m-p}|, \max_{j=2,\ldots,p} |\hat{y}_{n-j+1}^{m-j+1+k} - \hat{y}_{n-j+1}^{m-j+1}|)$$

$$\leq \max(L(\eta_n)|\hat{y}_{n-p}^{m-p+k} - \hat{y}_{n-p}^{m-p}|, L(\eta_{n-1}) \max_{j=1,\ldots,p} |y_{n-j-1} - \hat{y}_{n-j-1}^{m-j-1}|, \max_{j=3,\ldots,p} |y_{n-j+1} - \hat{y}_{n-j+1}^{m-j+1}|)$$

$$\leq \max(L(\eta_n)|\hat{y}_{n-p}^{m-p+k} - \hat{y}_{n-p}^{m-p}|, L(\eta_{n-1})|\hat{y}_{n-p-1}^{m-p-1+k} - \hat{y}_{n-p-1}^{m-p-1}|,$$
$$L(\eta_{n-1})|\hat{y}_{n-p}^{m-p+k} - \hat{y}_{n-p}^{m-p}|, \max_{j=3,\ldots,p} |\hat{y}_{n-j+1}^{m-j+1+k} - \hat{y}_{n-j+1}^{m-j+1}|)$$

$$\leq \max_{j=1,\ldots,p} L(\eta_{n-j+1}) \max_{j=1,\ldots,p} |\hat{y}_{n-p-j+1}^{m-p-j+1+k} - \hat{y}_{n-p-j+1}^{m-p-j+1}|,$$

and again using $0 \leq L(\cdot) \leq 1$, we also have by repeating this reasoning K times, for all $K \geq 1$,

$$a_{n,m,k} \equiv \max_{j=1,\ldots,p} |\hat{y}_{n-j+1}^{m+k-j+1} - \hat{y}_{n-j+1}^{m-j+1}|$$

$$\leq \max_{j=1,\ldots,p} L(\eta_{n-j+1}) \max_{j=1,\ldots,p} |\hat{y}_{n-pK-j+1}^{m+k-pK-j+1} - \hat{y}_{n-pK-j+1}^{m-pK-j+1}| \equiv L_n a_{n-pK,m-pK,k}.$$

Therefore,

$$a_{n,m,k} \leq L_n L_{n-pK} L_{n-2pK} \ldots L_{n-[m/(pK)]pK} a_{n-[m/(pK)]pK,m-[m/(pK)]pK,k}$$

$$\leq 2 \prod_{i=0}^{[m/(pK)]} L_{n-ipK} \quad \text{a.s.}$$

where $[x]$ denotes the integer part of x. Next, note that by assumption there exists positive integer K and $\delta > 0$, such that for ϕ_{max} and ϕ_{min} as defined below Equation (8),

$$|E(\max_{j=1,\ldots,p} L(\eta_{n-j+1})|\eta_{n-pK}, \eta_{n-pK-1}, \ldots)|$$

$$= |P(\phi_{max} + \max_{j=1,\ldots,p} \eta_{n-j+1} > 0|\eta_{n-pK}, \eta_{n-pK-1}, \ldots)$$
$$-P(\phi_{min} + \min_{j=1,\ldots,p} \eta_{n-j+1} > 0|y_{n-pK}, y_{n-pK-1}, \ldots)| < 1 - \delta$$

for some $\delta > 0$. Therefore, by successive conditioning it follows that

$$E \max_{k\geq 0} |\hat{y}_n^{m+k} - \hat{y}_n^m|^2 \leq 2(1-\delta)^{[m/(pK)]}.$$

By the Cauchy criterion it now follows that $\hat{y}_n^m$ converges a.s. as $m \to \infty$. Also, $\lim_{m\to\infty} \hat{y}_n^m$ satisfies

$$\psi(r,n) = E\exp(i\sum_{j=1}^{h} r_j \lim_{m\to\infty} \hat{y}_{n-j}^m) = \lim_{m\to\infty} E\exp(i\sum_{j=1}^{h} r_j \hat{y}_{n-j}^m)$$

by the dominated convergence theorem. Because $\hat{y}_n^m = f_m(\eta_n, \ldots, \eta_{n-m})$ and by the strict stationarity of η_n, it follows that $E\exp(i\sum_{j=1}^h r_j \hat{y}_{n-j}^m)$ does not depend on n, implying that $\lim_{m\to\infty} \hat{y}_n^m$ is strictly stationary. A similar argument shows that the pair (y_n, x_n) is strictly stationary. In addition,

$$E|\lim_{m\to\infty} \hat{y}_n^m - \hat{y}_n^m|^2 \leq E\sup_{k\geq 0} |\hat{y}_n^{m+k} - \hat{y}_n^m|^2 \leq 2(1-\delta)^{[m/(pK)]},$$

which shows the asserted property of the near epoch dependence numbers. Finally, if $g(\eta_n, \eta_{n-1}, \ldots)$ would be an alternative solution to the model, then for any $m > 0$

$$E|g(\eta_n, \eta_{n-1}, \ldots) - \hat{y}_t^m|^2 \leq 2(1-\delta)^{[m/(pK)]}$$

by the same argument as before, implying that $f(., \ldots) = g(., \ldots)$ a.s.. $\square$

Proof of Theorem 2:

Let $\mathcal{X}_{a,b}$ denote the σ-algebra generated by $((x_a, y_a), \ldots, (x_b, y_b))$. The definition of the strong mixing coefficients is

$$\sup_{n\in\mathbb{Z}} \sup_{F\in\mathcal{X}_{-\infty,t}, G\in\mathcal{X}_{t+m,\infty}} \{|P((x_n, y_n) \in F, (x_{n+m}, y_{n+m}) \in G) - P((x_n, y_n) \in F)P((x_{n+m}, y_{n+m}) \in G)|\},$$

see for example White (2001, page 47). Because y_n is a 0/1-valued random variable, there are only four possibilities for the possible values of the (y_n, y_{n-m}) pair. Therefore,

$$\begin{aligned}
&P((x_n, y_n) \in F, (x_{n+m}, y_{n+m}) \in G) \\
&= E\sum_{i=0}^{1}\sum_{j=0}^{1} I((x_n, y_n) \in F, (x_{n+m}, y_{n+m}) \in G)I(y_n = i)I(y_{n+m} = j) \\
&= E\sum_{i=0}^{1}\sum_{j=0}^{1} I((x_n, i) \in F, (x_{n+m}, j) \in G)I(y_n = i)I(y_{n+m} = j)
\end{aligned}$$

and

$$P((x_n, y_n) \in F) = EI((x_n, y_n) \in F)\sum_{i=0}^{1} I(y_n = i),$$

implying that

$$|P((x_n, y_n) \in F, (x_{n+m}, y_{n+m}) \in G) - P((x_n, y_n) \in F)P((x_{n+m}, y_{n+m}) \in G)|$$

$$\leq \sum_{i=0}^{1}\sum_{j=0}^{1} |EI((x_n, i) \in F, (x_{n+m}, j) \in G)I(y_n = i)I(y_{n+m} = j)$$

$$-EI((x_n, i) \in F)I(y_n = i)EI((x_{n+m}, j) \in G)I(y_{n+m} = j)|.$$

For the case $y_n = 1$, $y_{n+m} = 1$, we now have, defining $\mathcal{F}_n = \sigma(v_n, v_{n-1}, \ldots)$ for $v_n = (u_n, x_n')'$,

$$|EI((x_n, 1) \in F)y_n I((x_{n+m}, 1) \in G)y_{n+m} - EI((x_n, 1) \in F)y_n EI((x_{n+m}, 1) \in G)y_{n+m}|$$

$$= |EI((x_n, 1) \in F)y_n[E(I((x_{n+m}, 1) \in G)y_{n+m}|\mathcal{F}_n) - EI((x_{n+m}, 1) \in G)y_{n+m}]|$$
$$\leq E|E((I(x_{n+m}, 1) \in G)y_{n+m}|\mathcal{F}_n) - E(I((x_{n+m}, 1) \in G)y_{n+m})|,$$

and convergence to zero with m of the last expression constitutes the L_1-mixingale condition for $I((x_n, 1) \in G)y_n$ with respect to $\mathcal{F}_n$; see for example Pötscher and Prucha (1997) for a definition on an L_1-mixingale. Now $I((x_n, 1) \in G)y_n$ is a sequence that is bounded and near epoch dependent on v_n, implying that it is an L_1-mixingale, which in turn implies that

$$E|E(I((x_n, 1) \in G)y_n|\mathcal{F}_{n-m}) - E(I((x_n, 1) \in G)y_n)|$$

$$\leq C(\nu(m) + \alpha(m)).$$

The cases $y_n = 1$, $y_{n+m} = 0$; $y_n = 0$, $y_{n+m} = 1$; and $y_n = 0$, $y_{n-m} = 0$ are analogous, which then proves the result. $\square$

For the proof of Theorem 3, we need the following two lemmas. Let $w_n = (y_{n-1}, ..., y_{n-p}, x_n')'$.

Lemma 1 *Under the conditions of Theorem 3, and B being compact, $E \sup_{b \in B} |l_n(b)| < \infty$.*

Proof of Lemma 1:

Note that $Ew_n w_n'$ exists by Assumption A1. Existence of $Ew_n w_n'$ and the probit specification imply the result. The reasoning is similar to the result for cross-section probit, see Newey and McFadden (1994, page 2125, Example 1.2). $\square$

Lemma 2 *Under the conditions of Theorem 3, (i) $Ew_n w_n'$ is positive definite and (ii) $El_n(b)$ is uniquely maximized at $b = \beta$.*

Proof of Lemma 2:

Note that Ew_nw_n' exists by Assumption A1. The assumptions of Theorem 1 are satisfied so that $(x_n', y_n)'$ is strongly stationary. The assumption that distribution of w_n is not contained in any linear subspace of $\mathbb{R}^{p+q}$ implies that Ew_nw_n' is nonsingular so that Ew_nw_n' is positive definite. Let $b \neq \beta$ so that $E[(w_n'(b-\beta))^2] = (b-\beta)'Ew_nw_n'(b-\beta) > 0$, implying that $w_n'(b-\beta) \neq 0$ on a set with positive probability, implying that $w_n'b \neq w_n'\beta$ on a set with positive probability. Both $\Phi(z)$ and $\bar{\Phi}(z) = 1-\Phi(z)$ are strictly monotonic, and therefore $w_n'b \neq w_n'\beta$ implies that both $\Phi(w_n'b) \neq \Phi(w_n'\beta)$ and $\bar{\Phi}(w_n'b) \neq \bar{\Phi}(w_n'\beta)$. Thus, the density

$$p(y_n|w_n, b) = \Phi(w_n'b)^{y_n}\bar{\Phi}(w_n'b)^{1-y_n} \neq p(y_n|w_n, \beta)$$

on a set with positive probability. Note that $El_n(b)$ is concave so that it is uniquely minimized at $b = \beta$.

□

Proof of Theorem 3:

For convergence in probability, we check the conditions of Theorem 2.7 of Newey and McFadden (1994). The objective function $L_n(b)$ is concave. The stationarity and strong mixing assumptions imply ergodicity, see White (2001, theorem 3.44). This implies pointwise convergence, $L_n(b) \xrightarrow{p} El_n(b)$ for all b. Lemma 1 proves that $El_n(b)$ is uniquely maximized at β. Therefore, all conditions of Theorem 2.7 of Newey and McFadden (1994) are satisfied and consistency follows. For almost sure convergence, note that it is easily seen from Lemma 1 and Lemma 2 that all the conditions of Theorem A1 of Wooldridge (1994) are satisfied, except for the condition of uniform convergence in probability of $L_N(b)$. Note that Wooldridge's Theorem A1 can be extended to include a strong convergence result if instead of uniform convergence in probability of $L_N(b)$, uniform almost sure convergence $L_N(b)$ is assumed. To show this uniform convergence, we use the generic uniform law of large numbers of Andrews (1987). To show strong uniform law of large numbers, this theorem requires compactness of the parameter space, and in addition it needs to be verified that the summands $q_n(w_n, b)$ are such that $q_n(w_n, b)$, $q_n^*(w_n, b) = \sup\{q_n(w_n, \tilde{b}) : \tilde{b} \in B, |\tilde{b} - b| < \rho\}$ and $q_{n*}(w_n, b) = \inf\{q_n(w_n, \tilde{b}) : b \in B, |\tilde{b} - b| < \rho\}$ are well-defined and satisfy a strong law of large numbers, and that for all $b \in B$,

$$\lim_{\rho \to 0} \sup_{n \in \mathbb{Z}} |N^{-1} \sum_{n=1}^{N} Eq_n^*(w_n, b) - Eq_{*n}(w_n, b)| = 0.$$

The latter condition follows from stationarity of (y_n, x_n), continuity, and the envelope condition of Assumption A. In addition, $q_n(w_n, b)$, $q_n^*(w_n, b)$ and $q_{n*}(w_n, b)$ are well-defined and strong mixing random variables, so that we can apply the strong law of large numbers of Theorem 4 of de Jong (1995), from which it follows that if $\alpha(m) + \nu(m) \leq Cm^{-\eta}$ for some positive constants C and η, these variables will satisfy a strong law of large numbers. This is because under the condition that $E|l_n(b)|^{1+\delta} < \infty$, the summands will be an $L_{1+\delta/2}$-mixingale. $\square$

Lemma 3 *Under the conditions of Theorem 4,*

$$(N-p)^{1/2}(\partial L_N(b)/\partial b)|_{b=\beta} \xrightarrow{d} N(0, I).$$

Proof:

Note that by assumption, $E((\partial L_n(b)/\partial b)|_{b=\beta}|w_n) = 0$ so that $E(\partial L_n(b)/\partial b)|_{b=\beta} = 0$. Moreover, $(\partial L_n(b)/\partial b)|_{b=\beta}$ is a martingale difference sequence that is strong mixing and strictly stationary. In particular, the version of Bierens (2004, Theorem 7.11) of a central limit theorem of McLeish (1974) yields asymptotic normality. Applying the information matrix equality yields the result. $\square$

Proof of Theorem 4:

We prove Theorem 4 by checking the conditions of Newey and McFadden (1994, theorem 3.1). Consistency was shown in Theorem 3. Condition (i) was assumed. Condition (ii), twice differentiability of the log likelihood, follows from the probit specification. Condition (iii) was shown in Lemma 3. Note that stationarity and strong mixing imply ergodicity, see White (2001, theorem 3.44). Condition (iv) then follows from the probit specification and reasoning similar to Newey and McFadden, page 2147, example 1.2. Nonsingularity follows from the probit specification and $Ew_n w_n'$ being positive definite so that condition (iv) is satisfied. $\square$

For the proof of Theorem 5, we need the following lemmas.

Lemma 4 *For all $a \in \mathbb{R}$, if $0 \leq z_n \leq 1$ and (z_n, x_n) is strictly stationary and strong mixing, then*

$$\sup_{b \in B} |N^{-1} \sum_{n=1}^{N} (z_n I(b' x_n \leq a) - E z_n I(b' x_n \leq a))| \xrightarrow{p} 0.$$

In addition, if $\alpha(m) \leq C m^{-\eta}$ for positive constants C and η, the convergence is almost surely.

Proof of Lemma 4:

We will apply the generic uniform law of large numbers of the Theorem of Andrews (1987). It requires compactness of the parameter space B (which is assumed), and in addition it needs to be verified that the summands $q_n(w_n, b)$ are such that $q_n(w_n, b)$, $q_n^*(w_n, b) = \sup\{q_n(w_n, \tilde{b}) : \tilde{b} \in B, |\tilde{b} - b| < \rho\}$ and $q_{n*}(w_n, b) = \inf\{q_n(w_n, \tilde{b}) : b \in B, |\tilde{b} - b| < \rho\}$ are well-defined and satisfy a (respectively weak or strong) law of large numbers, and for all $b \in B$,

$$\lim_{\rho \to 0} \sup_{n \in \mathbb{Z}} |N^{-1} \sum_{n=1}^{N} E q_n^*(w_n, b) - E q_{*n}(w_n, b)| = 0.$$

To show the last result, note that (z_n, x_n) is strictly stationary under the conditions of the theorem, and therefore

$$\begin{aligned}
&\lim_{\rho \to 0} \sup_{n \in \mathbb{Z}} |N^{-1} \sum_{n=1}^{N} E q_n^*(w_n, b) - E q_{*n}(w_n, b)| \\
&= \lim_{\rho \to 0} \sup_{n \in \mathbb{Z}} |E z_n I(\sup_{\tilde{b}:|b-\tilde{b}|<\rho} b' x_n < a) - E z_n I(\inf_{\tilde{b}:|b-\tilde{b}|<\rho} b' x_n < a)| \\
&\leq \limsup_{K \to \infty} \lim_{\rho \to 0} \sup_{n \in \mathbb{Z}} |E z_n (I(b' x_n < a + \rho |x_n|) - I(b' x_n < a - \rho |x_n|)) I(|x_n| \leq K) \\
&\quad + \limsup_{K \to \infty} \lim_{\rho \to 0} \sup_{n \in \mathbb{Z}} |E z_n (I(b' x_n < a + \rho |x_n|) - I(b' x_n < a - \rho |x_n|)) I(|x_n| > K))| \\
&\leq \limsup_{K \to \infty} \lim_{\rho \to 0} |P(b' x_n < a + \rho K) - P(b' x_n < a - \rho K)| + \limsup_{K \to \infty} P(|x_n| > K) = 0,
\end{aligned}$$

because x_{1n} has a continuous distribution. Furthermore, note that $q_n(z_n, b)$,

$$q_n^*(w_n, b) = z_n I(\sup_{\tilde{b}:|b-\tilde{b}|<\rho} b' x_n < a)$$

and

$$q_{*n}(w_n, b) = z_n I(\inf_{\tilde{b}:|b-\tilde{b}|<\rho} b' x_n < a)$$

are well-defined and strong mixing random variables, implying that weak law of large numbers for mixingales of Andrews (1988) applies; or alternatively we can apply the strong law of large numbers of Theorem 4 of de Jong (1995), from which it follows that if $\alpha(m)+\nu(m) \leq Cm^{-\eta}$ for some positive constants C and η, these variables will satisfy a strong law of large numbers (note that because of boundedness of the summands, the summands are L_2-mixingales). $\square$

Lemma 5 *Under Assumptions 1,3, 4, 5 and 2,*

$$\sup_{b\in B} |S_N(b,\sigma_N) - ES_N(b,\sigma_N)| \xrightarrow{p} 0.$$

In addition, if $\alpha(m) \leq Cm^{-\eta}$ for positive constants C and η, the convergence is almost surely.

Proof of Lemma 5:

First note that Horowitz' proof of his Lemma 4 (i.e. $\sup_{b\in B} |S_N(b,\sigma_N) - S_N^*(b)| \xrightarrow{as} 0$) goes through as it stands, except for the proof of uniform convergence of the term in his Equation (A4), which uses a uniform law of large numbers for i.i.d. random variables. To show that

$$\sup_{b\in B} |N^{-1}\sum_{n=1}^{N}(I(|\sum_{j=1}^{p} r_j y_{n-j} + c'x_n| < \alpha) - EI(|\sum_{j=1}^{p} r_j y_{n-j} + c'x_n| < \alpha))|$$

satisfies a strong or weak law of large numbers, we can use Lemma 4. To do so, note that

$$N^{-1}\sum_{n=1}^{N} I(|\sum_{j=1}^{p} r_j y_{n-j} + c'x_n| < \alpha)$$

$$= \sum_{j_1=0}^{1} \ldots \sum_{j_p=0}^{1} N^{-1}\sum_{n=1}^{N} I(y_{n-1}=j_1)\ldots I(y_{n-p}=j_p) I(|\sum_{i=1}^{p} r_i j_i + c'x_n| < \alpha)$$

and note that $I(y_{n-1}=j_1)\ldots I(y_{n-p}=j_p)$ is strong mixing, because it is the product of strong mixing random variables. It now only remains to be proven that

$$\sup_{b\in B} |S_N^*(b) - S(b)| \xrightarrow{p} 0 \quad \text{or} \quad \xrightarrow{as} 0,$$

which Horowitz shows by referring to Manski (1985). This can be shown by noting that

$$S_N^*(b) = N^{-1}\sum_{n=1}^{N}(2 \cdot I(y_n = 1) - 1)I(b'x_n \geq 0)$$

$$= 2N^{-1}\sum_{n=1}^{N} y_n I(b'x_n \geq 0) - N^{-1}\sum_{n=1}^{N} I(b'x_n \geq 0),$$

and by Lemma 4, both terms satisfy a (weak or strong) uniform law of large numbers. $\square$

Lemma 6 *Under Assumptions 1,3, 4 and 5, $S(b) \leq S(\beta)$ with equality holding only if $b = \beta$.*

Proof of Lemma 6:

This result follows by noting that all conditions from Lemma 3 of Manski (1985) are satisfied. $\square$

Proof of Theorem 5 :

The proof of the theorem now follows from Theorem A1 of Wooldridge (1994) and the results of Lemma 5 and Lemma 6. $\square$

Let $z_n = \sum_{j=1}^{p} \rho_j y_{n-j} + \gamma' x_n$. The following lemma shows that Horowitz' Lemma 5 holds as it stands in our setting:

Lemma 7 *Under Assumptions 1' and Assumptions 3-14,*

$$\lim_{N\to\infty} E[\sigma_N^{-h} T_N(\beta, \sigma_N)] = A;$$

$$\lim_{N\to\infty} \mathrm{Var}[(N\sigma_N)^{1/2} T_N(\beta, \sigma_N)] = D.$$

Proof of Lemma 7:

The only adjustment to Horowitz' Lemma 5 that needs to be made is to show that the covariance terms in $\mathrm{Var}[(N\sigma_N)^{1/2}T_N(\beta,\sigma_N)]$ are asymptotically negligible. To prove this, we show that for all vectors ξ such that $|\xi|=1$,

$$\lim_{N\to\infty}\sigma_N\sum_{m=1}^{\infty}|\mathrm{cov}(\xi'(\tilde{x}_n/\sigma_N)K'(z_n/\sigma_N),\xi'(\tilde{x}_{n-m}/\sigma_N)K'(z_{n-m}/\sigma_N))|=0.$$

By the covariance inequality for mixingales, for the same s as in Assumption 6, (see Davidson (1994, p. 212, Corollary 14.3)),

$$\begin{aligned}
&\sigma_N\mathrm{cov}(\xi'(\tilde{x}_n/\sigma_N)K'(z_n/\sigma_N),\xi'(\tilde{x}_{n-m}/\sigma_N)K'(z_{n-m}/\sigma_N))\\
&\le\sigma_N C\alpha(m)^{1-2/s}(E|\xi'(\tilde{x}_n/\sigma_N)K'(z_n/\sigma_N)|^s)^{1/s}(E|\xi'(\tilde{x}_{n-m}/\sigma_N)K'(z_{n-m}/\sigma_N)|^s)^{1/s}\\
&=\sigma_N^{-1}C\alpha(m)^{1-2/s}(\int|\xi'\tilde{x}|^s|K'(z/\sigma_N)|^s p(z|\tilde{x})dzdP(\tilde{x}))^{2/s}\\
&=C\alpha(m)^{1-2/s}\sigma_N^{2/s-1}(\int|\xi'\tilde{x}|^s|K'(\zeta)|^s p(\sigma_N\zeta|\tilde{x})d\zeta dP(\tilde{x}))^{2/s}
\end{aligned}$$

by substituting $\zeta=z/\sigma_N$. The last term is smaller than $C'\sigma_N^{2/s-1}\alpha(m)^{1-2/s}$ for some constant C'. In view of the fact that summing the latter expression over m will give a term that diverges as $N\to\infty$, we also need to use a second bound. To obtain this second bound, note that by Horowitz' arguments, under the conditions of the theorem,

$$\sigma_N E\xi'(\tilde{x}_{n-m}/\sigma_N)K'(z_{n-m}/\sigma_N)=O(\sigma_N),$$

implying that

$$\begin{aligned}
&\sigma_N\mathrm{cov}(\xi'(\tilde{x}_n/\sigma_N)K'(z_n/\sigma_N),\xi'(\tilde{x}_{n-m}/\sigma_N)K'(z_{n-m}/\sigma_N))\\
&=O(\sigma_N)+\sigma_N E(\sigma_N\xi'(\tilde{x}_n/\sigma_N)K'(z_n/\sigma_N)\xi'(\tilde{x}_{n-m}/\sigma_N)K'(z_{n-m}/\sigma_N))\\
&=O(\sigma_N)+\sigma_N^{-1}\int\xi'\tilde{x}_nK'(z_n/\sigma_N)\xi'\tilde{x}_{n-m}K'(z_{n-m}/\sigma_N)dP(x_n,x_{n-m},z_n,z_{n-m})\\
&=O(\sigma_N)+\sigma_N^{-1}\int\xi'\tilde{x}_nK'(z_n/\sigma_N)\xi'\tilde{x}_{n-m}K'(z_{n-m}/\sigma_N)dp(z_n,z_{n-m}|x_n,x_{n-m})dz_ndz_{n-m}dP(x_n,x_{n-m})\\
&=O(\sigma_N)+\sigma_N\int\int K'(\zeta_n)K'(\zeta_{n-m})p(\sigma_N\zeta_n,\sigma_N\zeta_{n-m}|x_n,x_{n-m})d\zeta_nd\zeta_{n-m}\xi'\tilde{x}_{n-m}\xi'\tilde{x}_ndP(x_n,x_{n-m})\\
&=O(\sigma_N)
\end{aligned}$$

under the assumptions of the theorem. Therefore for any $\kappa \in (0,1)$,

$$\sum_{m=1}^{\infty} |\mathrm{cov}(\xi'(\tilde{x}_n/\sigma_N)K'(z_n/\sigma_N), \xi'(\tilde{x}_{n-m}/\sigma_N)K'(z_{n-m}/\sigma_N))|$$

$$\leq C \sum_{m=1}^{\infty} (\sigma_N)^{\kappa}(\alpha(m)^{(1-2/s)}\sigma_N^{2/s-1})^{1-\kappa},$$

and by choosing $\kappa = (s-2)/(2s-2) + \eta$ and $\eta > 0$ small enough, the last term can be bounded by

$$C(\sum_{m=1}^{\infty} \alpha(m)^{(s-2)/(2s-2)-\eta(s-2)/s})\sigma_N^{(2s-2)\eta/s} = O(\sigma_N^{(2s-2)\eta/s}) = o(1),$$

where the finiteness of the summation follows from the assumptions. □

Horowitz' Lemma 6 now holds as follows:

Lemma 8 *Under Assumptions 1' and Assumptions 3-14, (a) If $N\sigma_N^{2h+1} \to \infty$ as $N \to \infty$, $\sigma_N^{-h}T_N(\beta, \sigma_N) \xrightarrow{p} A$. (b) If $N\sigma_N^{2h+1}$ has a finite limit λ as $N \to \infty$, $(N\sigma_N)^{1/2}T_N(\beta, \sigma_N) \xrightarrow{d} N(\lambda^{1/2}A, D)$.*

Proof of Lemma 8:

The modification of Horowitz (1992) that is needed is to show that for all vectors ξ such that $|\xi| = 1$,

$$(\sigma_N/N)^{1/2}\xi' \sum_{n=1}^{N}(t_{Nn} - Et_{Nn}) \xrightarrow{d} N(0, \xi' D\xi),$$

where

$$t_{Nn} = (2y_n - 1)(\tilde{x}_n/\sigma_N)K'(z_n/\sigma_N).$$

Since t_{Nn} is strong mixing, Theorem 2 of de Jong (1997) for strong mixing arrays can now be applied to show this result under the conditions of the lemma. Note that the condition $\alpha(m) \leq Cm^{-s/(s-2)-\eta}$ from that theorem follows from the assumptions of the lemma. □

For reproving Horowitz' Lemma 7 for the case of strong mixing data, we need the following lemmas:

Lemma 9 (Azuma(1967)) *If η_n is a martingale difference sequence with respect to*
$\mathcal{F}_n$ and $|\eta_n| \leq C_N$, then

$$P(|N^{-1}\sum_{n=1}^{N}\eta_n| > \delta) \leq 2\exp(-N\delta^2/C_N^2).$$

Proof of Lemma 9:

See Azuma (1967). $\square$

An m_N-fold application of the above lemma now gives the following result:

Lemma 10 *If $\mathcal{F}_n$ is a sequence of sigma-fields such that $\eta_n - E(\eta_n|\mathcal{F}_{n-1})$ is a martingale difference sequence with respect to $\mathcal{F}_n$ and $|\eta_n| \leq C_N$, then for any integer-valued sequence m_N such that $m_N \geq 1$,*

$$P(|N^{-1}\sum_{n=1}^{N}(\eta_n - E(\eta_n|\mathcal{F}_{n-m_N}))| > \delta) \leq 2m_N\exp(-\delta^2/(m_N^2C_N^2)).$$

Proof of Lemma 10:

Obviously

$$N^{-1}\sum_{n=1}^{N}(\eta_n - E(\eta_n|\mathcal{F}_{n-m_N})) = \sum_{j=0}^{m_N-1}N^{-1}\sum_{n=1}^{N}(E(\eta_n|\mathcal{F}_{n-j}) - E(\eta_n|\mathcal{F}_{n-j-1})),$$

and therefore

$$P(|N^{-1}\sum_{n=1}^{N}(\eta_n - E(\eta_n|\mathcal{F}_{n-m_N}))| > \delta)$$

$$\leq \sum_{j=0}^{m_N-1}P(|N^{-1}\sum_{n=1}^{N}(E(\eta_n|\mathcal{F}_{n-j}) - E(\eta_n|\mathcal{F}_{n-j-1}))| > \delta/m_N)$$

$$\leq 2m_N\exp(-\delta^2/(m_N^2C_N^2))$$

by m_N applications of Lemma 9. $\square$

Analogously to Horowitz (1992), define

$$g_{Nn}(\theta) = (2 \cdot I(y_n = 1) - 1)\tilde{x}_n K'(z_n/\sigma_N + \theta'\tilde{x}_n)$$

The following result is now the analogue[5] of Horowitz' Lemma 7.

Lemma 11 *If (y_n, x_n) is strong mixing with strong mixing sequence $\alpha(m)$, and there exists a sequence $m_N \geq 1$ such that*

$$\sigma_N^{-3(p+q-1)}\sigma_N^{-2}N^{1/s}\alpha(m_N) + (\log(Nm_N))(N^{1-2/s}\sigma_N^4 m_N^{-2})^{-1} \to 0,$$

then

$$\sup_{\theta\in\Theta_N} |(N\sigma_N^2)^{-1}\sum_{n=1}^{N}(g_{Nn}(\theta) - Eg_{Nn}(\theta))| \xrightarrow{p} 0.$$

Note that the second part of Horowitz' Lemma 7 will hold without modification. Also note that the case of i.i.d. (y_n, x_n) is a special case, because then $\alpha(m) = 0$ for $m \geq 1$, and we could set $m_N = 1$ for that case.

Proof of Lemma 11:

Consider

$$g_{Nn}^{C_N}(\theta) = (2 \cdot I(y_n = 1) - 1)\tilde{x}_n K'(z_n/\sigma_N + \theta'\tilde{x}_n)I(|\tilde{x}_n| \leq C_N)$$

and note that obviously,

$$\begin{aligned} g_{Nn}(\theta) - Eg_{Nn}(\theta) &= (g_{Nn}^{C_N}(\theta) - Eg_{Nn}^{C_N}(\theta)) \\ &\quad +(g_{Nn}(\theta) - g_{Nn}^{C_N}(\theta) - Eg_{Nn}(\theta) + Eg_{Nn}^{C_N}(\theta)). \end{aligned} \tag{41}$$

[5]Note that Horowitz' Lemma 7 only holds for bounded regressors, and that the truncation argument at the start of Lemma 8 appears to be in error. Horowitz does not explicitly consider the remainder statistic containing the summation elements for which $|\tilde{x}_n|$ exceeds a. Horowitz' Lemma 9 appears to have a similar problem in its proof. Therefore, Lemma 11 also serves to correct this aspect of Horowitz' proof. This is because the conditioning on the event C_γ does not appear relevant; while Horowitz' $\tilde{x}$ stands for a random variable distributed identically to any $\tilde{x}_n$, the conditioning should be with respect to every $\tilde{x}_n$, $n = 1, \ldots, N$, in order for this argument to work. However, unless $\tilde{x}_n$ is almost surely bounded, such a conditioning set C_γ would depend on N, and will not have the desired property that $\limsup_{\gamma\to\infty}\limsup_{N\to\infty} P(C_\gamma) = 0$.

Now define $C_N = \eta^{-1/s} N^{1/s} (E|\tilde{x}_n|^s)^{1/s}$ for any $\eta > 0$. Then because $C_N \to \infty$ as $N \to \infty$, following the reasoning as in the proof of (A16) of Horowitz (1992, page 525-526), it follows that

$$\sup_{\theta\in\Theta} |E g_{Nn}(\theta) - E g_{Nn}^{C_N}(\theta)| \to 0. \tag{42}$$

In addition,

$$P(\sup_{\theta\in\Theta} |\sum_{n=1}^{N} (g_{Nn}(\theta) - g_{Nn}^{C_N}(\theta))| = 0) \leq P(\exists n : |\tilde{x}_n| > C_N) \leq N E|\tilde{x}_n|^s C_N^{-s} \leq \eta, \tag{43}$$

and we can choose η arbitrarily small. For the case $s = \infty$, it is trivial that these two terms disappear asymptotically for some constant C_N not depending on N. To deal with the first part of Equation (41), note that

$$g_{Nn}(\theta) - E g_{Nn}(\theta) = (g_{Nn}(\theta) - E(g_{Nn}(\theta)|\mathcal{F}_{n-m_N})) + (E(g_{Nn}(\theta)|\mathcal{F}_{n-m_N}) - E g_{Nn}(\theta)). \tag{44}$$

To deal with the first part of the right-hand side of Equation (44), we can copy the argument on page 525 of Horowitz (1992), except that now, by Lemma 10,

$$\begin{aligned}&\sum_{i=1}^{\Gamma_N} P((N\sigma_N^2)^{-1} |\sum_{n=1}^{N} (g_{Nn}(\theta_{Ni}) - E g_{Nn}(\theta_{Ni}))| > \varepsilon/2)\\ &\leq 2\Gamma_N m_N \exp(-\varepsilon^2 4^{-1} N \sigma_N^4 C_N^{-2} m_N^{-2}).\end{aligned}$$

where Γ_N is as defined in Horowitz (1992). Since $\Gamma_N = O(\sigma_N^{-3(p+q-1)})$, this term will converge to zero if

$$(\log(N m_N))(N \sigma_N^4 C_N^{-2} m_N^{-2})^{-1} \to 0, \tag{45}$$

which is assumed. For dealing with the second part of the right-hand side of Equation (44), note since $g_{Nn}(\theta)$ is strong mixing, it is also an L_1-mixingale (see for example Davidson (1994, p. 249, Example 16.3), implying that

$$E|E(g_{Nn}^{C_N}(\theta)|\mathcal{F}_{n-m_N}) - E g_{Nn}^{C_N}(\theta)| \leq 6 C_N \alpha(m_N).$$

Using Horowitz' reasoning of page 525, it now suffices to show that for all $\varepsilon > 0$,

$$\sum_{i=1}^{\Gamma_N} P((N\sigma_N^2)^{-1} |\sum_{n=1}^{N} E(g_{Nn}^{C_N}(\theta_{Ni})|\mathcal{F}_{n-m_N}) - E g_{Nn}^{C_N}(\theta_{Ni})| > \varepsilon) \to 0.$$

By the Markov inequality,

$$\sum_{i=1}^{\Gamma_N} P((N\sigma_N^2)^{-1} |\sum_{n=1}^{N} E(g_{Nn}^{C_N}(\theta_{Ni})|\mathcal{F}_{n-m_N}) - E g_{Nn}^{C_N}(\theta_{Ni})| > \varepsilon)$$

$$\leq \sum_{i=1}^{\Gamma_N} \varepsilon^{-1} \sigma_N^{-2} N^{-1} \sum_{n=1}^{N} E|E(g_{Nn}^{C_N}(\theta)|\mathcal{F}_{n-m_N}) - Eg_{Nn}^{C_N}(\theta)|$$
$$= O(\sigma_N^{-3(p+q-1)} \sigma_N^{-2} C_N \alpha(m_N)) = o(1)$$

by assumption. □

Lemma 12 *Under Assumptions 1' and Assumptions 3-14,* $(\tilde{b}_N - \tilde{\beta})/\sigma_N \xrightarrow{p} 0.$

Proof of Lemma 12:

This follows from Lemma 11 and the reasoning[6] of Horowitz' (1992) Lemma 8. □

The following lemma corresponds[7] to Horowitz' Lemma 9.

Lemma 13 *Let* $\{\beta_N\} = \{\beta_{N1}, \tilde{\beta}_N\}$ *be such that* $(\beta_N - \beta)/\sigma_N \xrightarrow{p} 0$ *as* $N \to \infty$*. Then under Assumptions 1' and Assumptions 3-14,*

$$Q_N(\beta_N, \sigma_N) \xrightarrow{p} Q.$$

Proof of Lemma 13:

Remember that

$$Q_N(\beta_N, \sigma_N) = [\sigma_N^{-2} N^{-1} \sum_{n=1}^{N} (2y_n - 1) \tilde{x}_n \tilde{x}_n' K''((\sum_{j=1}^{p} r_j y_{n-j} + c' x_n)/\sigma_N)]_{b=\beta_N}.$$

Since $P(b_1 = \beta_1) \to 1$ and by the assumption that $(\beta_N - \beta)/\sigma_N \xrightarrow{p} 0$ as $N \to \infty$, it suffices to show that for all $\eta > 0$ and any vector ξ such that $|\xi| = 1$,

$$\sup_{|\tilde{\theta}| \leq \eta} |N^{-1} \sum_{n=1}^{N} r_{nN}(\tilde{\theta}) - E r_{nN}(\tilde{\theta})| \equiv \sup_{|\tilde{\theta}| \leq \eta} |\sigma_N^{-2} N^{-1} \sum_{n=1}^{N} (2y_n - 1)(\xi' \tilde{x}_n)^2 K''(z_n/\sigma_N + \tilde{\theta}' \tilde{x}_n)$$

---

[6] See footnote 2.

[7] We discuss Horowitz' conditioning on X_N in footnote 5. Note that when Horowitz uses his Lemma 8 in the proof of his Theorem 2, a uniform law of large numbers appears to be needed rather than the result of his Lemma 8.

$$-E(2y_n - 1)(\xi'\tilde{x}_n)^2 K''(z_n/\sigma_N + \tilde{\theta}'\tilde{x}_n)| \xrightarrow{p} 0. \tag{46}$$

Note that Horowitz (1992) shows the continuity of $Er_{nN}(\tilde{\theta})$ in $\tilde{\theta}$ uniformly in N. To show the result of Equation (46), note that

$$P(\sup_{|\tilde{\theta}|\leq\eta} |N^{-1}\sum_{n=1}^{N} r_{nN}(\tilde{\theta})I(|r_{nN}(\tilde{\theta})| > C_N)| = 0)$$

$$\leq \sum_{n=1}^{N} P((\xi'\tilde{x}_n)^2 > C_N) \leq NE|\xi'\tilde{x}_n|^s C_N^{-s/2}$$

and the last term can be made smaller than ε by choosing $C_N^{-s/2} = N^{-1}\varepsilon(E|\xi'\tilde{x}_n|^s)^{-1}$. In addition, it is easily verified that

$$\sup_{|\tilde{\theta}|\leq\eta} |N^{-1}\sum_{n=1}^{N} E(r_{nN}(\tilde{\theta})I(|r_{nN}(\tilde{\theta})| > C_N))| \to 0.$$

Because of these two results, it suffices to show uniform convergence to zero in probability of

$$R_N(\tilde{\theta}) = N^{-1}\sum_{n=1}^{N}(r_{nN}(\tilde{\theta})I(|r_{nN}(\tilde{\theta})| \leq C_N) - Er_{nN}(\tilde{\theta})I(|r_{nN}(\tilde{\theta})| \leq C_N)$$

$$+C_N I(|r_{nN}(\tilde{\theta})| > C_N) - EC_N I(|r_{nN}(\tilde{\theta})| > C_N)).$$

Now note that since $\tilde{\theta} \in \mathbb{R}^{p+q-1}$, we can cover the parameter space $\{\tilde{\theta} : |\tilde{\theta}| \leq \eta\}$ with $O(\sigma_N^{-2(p+q-1)/\mu})$ balls of size $\sigma_N^{2/\mu}$ and with centers $\tilde{\theta}_j$. Now note that, by Assumption 14,

$$\sup_{N\geq 1} E \sup_{|\tilde{\theta}-\tilde{\theta}'|<\delta\sigma_N^{2/\mu}} |R_N(\tilde{\theta}) - R_N(\tilde{\theta}')|$$

$$\leq \sup_{N\geq 1} E(\xi'\tilde{x}_n)^2 L \sup_{|\tilde{\theta}-\tilde{\theta}'|<\delta\sigma_N^{2/\mu}} |\tilde{\theta} - \tilde{\theta}'|^\mu \sigma_N^{-2} \to 0 \qquad \delta \to 0.$$

Using Lemma 11 and following the same reasoning as in the proof of that lemma, we can now argue

$$\limsup_{n\to\infty} P(\sup_{|\tilde{\theta}|\leq\eta} |R_N(\tilde{\theta}) - ER_N(\tilde{\theta})| > \varepsilon)$$

$$\leq \limsup_{n\to\infty} P(\max_j |R_N(\tilde{\theta}_j) - ER_N(\tilde{\theta}_j)| > \varepsilon/2)$$

$$\leq \limsup_{n\to\infty} \sum_j P(|R_N(\tilde{\theta}_j) - ER_N(\tilde{\theta}_j)| > \varepsilon/2)$$

$$= O(\sigma_N^{-2(p+q-1)/\mu}[2m_N \exp(-N\varepsilon^2/(4\sigma_N^4 C_N^2 m_N^2)) + \varepsilon^{-1} C_N \alpha(m_N)])$$

and because $C_N = O(N^{2/s})$, the last term converges to 0 if

$$\sigma_N^{-2(p+q-1)/\mu} N^{2/s} \alpha(m_N) + (m_N^{-2} \sigma_N^{-4} N^{1-4/s})^{-1} \log(N m_N) \to 0,$$

which is assumed. □

Proof of Theorem 6:

This proof is identical to the proof of Horowitz' Theorem 2, where we need to use our Lemma 12 and Lemma 13 instead of Horowitz' Lemma 8 and Lemma 9. □

Proof of Theorem 7:

Part (a) now follows exactly[8] as in Horowitz' proof of his Theorem 3, where our Lemma 12 and Lemma 13 replace Horowitz' Lemma 8 and Lemma 9. Part (c) follows from Lemma 13.

$\square$

[8]To show part (b), one can use a uniform law of large numbers result of the type

$$\sup_{|\tilde{\theta}|\leq\eta} |\sigma_N^{-1}N^{-1}\sum_{n=1}^{N}(\xi'\tilde{x}_n)^2K'(z_n/\sigma_N+\tilde{\theta}'\tilde{x}_n)-E((\xi'\tilde{x}_n)^2K'(z_n/\sigma_N+\tilde{\theta}'\tilde{x}_n)| \xrightarrow{p} 0$$

for all ξ such that $|\xi| = 1$. Under the conditions of our theorem, this result can be proven analogously to the proof of Lemma 13, using the same C_N and ball size sequences. Note that $K'(.)$ is Lipschitz-continuous with $\mu = 1$, since $K''(.)$ is assumed to exist and to be uniformly bounded.