Centre for Efficiency and Productivity Analysis

Working Paper Series
No. WP06/2016

Nonparametric Estimation of Dynamic Discrete Choice Models for Time Series Data

Byeong U. Park, Leopold Simar, Valentin Zelenyuk

Date: October 2016

School of Economics
University of Queensland
St. Lucia, Qld. 4072
Australia

ISSN No. 1932 - 4398

Nonparametric Estimation of Dynamic Discrete Choice Models for Time Series Data[*]

Byeong U. Park[†], Léopold Simar[‡], Valentin Zelenyuk[§]

7 October 2016

Abstract

The non-parametric quasi-likelihood method is generalized to the context of discrete choice models for time series data where dynamics is modeled via lags of the discrete dependent variable appearing among regressors. Consistency and asymptotic normality of the estimator for such models in the general case is derived under the assumption of stationarity with strong mixing condition. Monte Carlo examples are used to illustrate performance of the proposed estimator relative to the fully parametric approach. Possible applications for the proposed estimator may include modeling and forecasting of probabilities of whether a subject would get a positive response to a treatment, whether in the next period an economy would enter a recession, or whether a stock market will go down or up, etc.

JEL: C14, C22,C25, C44

Keywords: Nonparametric, Dynamic Discrete Choice, Probit

[*] this is a substantial update on the earlier working paper WP09/2013

[†] Department of Statistics, Seoul National University, Korea

[‡] Inst. de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Belgium.

[§] School of Economics and Centre for Efficiency and Productivity Analysis (CEPA) at The University of Queensland, Australia

1 Introduction

Discrete response (or choice) models have received substantial interest in many areas of research. Since the influential works of McFadden (1973, 1974) and Manski (1975), these models have become very popular in economics, especially microeconomics, where they were elaborated on and generalized in many respects. Some very interesting applications of such models are also found in macroeconomic studies where one needs to take into account time series aspects of data. Typical applications of the time series discrete response models deal with forecasting of economic recessions, the decisions of central banks on interest rate, movements of the stock market indices, etc. (See Estrella and Mishkin (1995, 1998), Dueker (1997, 2005), Russell and Engle (1998, 2005), Park and Phillips (2000), Hu and Phillips (2004), Chauvet and Potter (2005), Kauppi and Saikkonen (2008), de Jong and Woutersen (2011), Harding and Pagan (2011), Kauppi (2012) and Moysiadis and Fokianos (2014) to mention just a few.)

The primary goal of this work is to develop a methodology for non-parametric estimation of dynamic time series discrete response models, where the discrete dependent variable is related to its own lagged values as well as other regressors. The theory we develop in the next two sections is fairly general and can be used in many areas of research.

The reason for going non-parametric, at least as a complementary approach, is very simple, yet profound: The parametric maximum likelihood in general, and probit or logit approaches in particular, yield inconsistent estimates if the parametric assumptions are misspecified. Many important works addressed this issue in different ways, e.g., see Cosslett (1983, 1987), Manski (1985), Klein and Spady (1989), Horowitz (1992), Matzkin (1992, 1993), Fan, Heckman and Wand (1995), Lewbel (2000), Honore and Lewbel (2002), Frölich (2006), Dong and Lewbel (2011), Harding and Pagan (2011), to mention just a few.

The main contribution of our work to existing literature is that we generalize the method of Fan et al. (1995) to the context that embraces time series aspects and in particular the case with lags of the (discrete) dependent variable appearing among the regressors. Such a dynamic feature of the model is very important in practice. For example, in weather forecasting, one would also naturally expect that the lagged dependent variable, describing whether the previous day was rainy or not, may play a very important role in explaining the probability that the next day will also be rainy. Another example of the importance of the dynamic component among the explanatory factors in discrete response models can be found in the area of forecasting economic recessions (Dueker (1997) and Kauppi and Saikkonen (2008)).

We derive the asymptotic theory for our estimator under the assumption of stationarity with strong mixing condition (in the spirit of Masry (1996)). Our approach is different from and compliments to another powerful non-parametric approach based on the Nadaraya-Watson estimator (e.g., see Harding and Pagan (2011)). Specifically, we use an alternative estimation paradigm—the one based on the non-parametric quasi-likelihood and the local likelihood concepts—which have well-known advantages over the least squares approach for the context of discrete response models. Furthermore, we consider and derive the theory for the local linear fit, which is known to provide more accurate estimation of a model than the local constant approach and is more convenient for estimation of derivatives or marginal effects of the regressors on (the expected value of) the response variable.

It is also worth noting here that a related approach (a special case of ours) was used by Frölich (2006) who considered the local likelihood method in the case of a binary logit-type regression model with both continuous and discrete explanatory variables, in a cross section set up. Specifically, Frölich (2006) provided very useful and convincing Monte Carlo evidence about superior performance of the local likelihood logit relative to parametric logit for his set up (cross-section), but without deriving asymptotic properties of the resulting estimators. Our work encompasses the work of Frölich (2006) as a special case, and, importantly, allows for time series nature of the data, including the dynamic aspect, and provides key asymptotic results for this set up that appears to be missing in the literature. A natural extension to our work would be to also allow for non-stationarity (e.g., as in Park and Phillips (2000)), which is a subject in itself and so we leave it for future research.

Our paper is structured as following: Section 2 outlines the general methodology, Section 3 outlines theoretical properties of the proposed estimator, Section 4 discusses the choice of bandwidths, Section 5 provides some Monte Carlo evidence, while the Appendix provides further details.

2 General Methodology

Suppose we observe $(\mathbf{X}^i, \mathbf{Z}^i, Y^i)$, $1 \leq i \leq n$, where $\{(\mathbf{X}^i, \mathbf{Z}^i, Y^i)\}_{i=-\infty}^{\infty}$ is a stationary random process. We assume that the process satisfies a strong mixing condition, as described in detail in the next section. The response variable Y^i is of discrete type. For example, it may be binary taking the values 0 and 1. The vector of covariates $\mathbf{X}^i$ is of d-dimension and of continuous type, while $\mathbf{Z}^i$ is of k-dimension and of discrete type. The components of the vector Z^i are allowed to be lagged values of the response variable. For example,

$\mathbf{Z}^i = (Y^{i-1}, \ldots, Y^{i-k})$. Our main interest is to estimate the mean function

$$m(\mathbf{x}, \mathbf{z}) = E(Y^i | \mathbf{X}^i = \mathbf{x}, \mathbf{Z}^i = \mathbf{z}).$$

We employ the quasi-likelihood approach of Fan, Heckman and Wand (1995) to estimate the mean function. It requires two ingredients. One ingredient is the specification of a quasi-likelihood $Q(\cdot, y)$, which is understood to take the role of the likelihood of the mean when $Y = y$ is observed. It is defined by $\partial Q(\mu, y)/\partial\mu = (y - \mu)/V(\mu)$, where V is a chosen function for the working conditional variance model $\sigma^2(\mathbf{x}, \mathbf{z}) \equiv \mathrm{var}(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = V(m(\mathbf{x}, \mathbf{z}))$, where here and below $(\mathbf{X}, \mathbf{Z}, Y)$ denotes the triple that has the same distribution as $(\mathbf{X}^i, \mathbf{Z}^i, Y^i)$. The other ingredient is the specification of a link function g. The link function should be strictly increasing. In a parametric model where it is assumed that $g(m(\mathbf{x}, \mathbf{z}))$ takes a parametric form, its choice is a part of the parametric assumptions. Thus, a wrong choice would jeopardize the estimation of m. In nonparametric settings, its choice is less important. One may take simply the identity function as a link, but one often needs to use a different one. One case is where the target function m has a restricted range, such as the one where Y is binary so that m has the range $[0, 1]$. A proper use of a link function guarantees the correct range.

With a link function g and based on the observations $\{(\mathbf{X}^i, \mathbf{Z}^i, Y^i)\}_{i=1}^n$, the quasi-likelihood of the function f defined by $f(\mathbf{x}, \mathbf{z}) = g(m(\mathbf{x}, \mathbf{z}))$ is given by $\sum_{i=1}^n Q(g^{-1}(f(\mathbf{X}^i, \mathbf{Z}^i)), Y^i)$. Let $(\mathbf{x}, \mathbf{z})$ be a fixed point of interest at which we want to estimate the value of the mean function m or the transformed function f. We apply a local smoothing technique to the observations $(\mathbf{X}^i, \mathbf{Z}^i)$ near $(\mathbf{x}, \mathbf{z})$. In the space of the continuous covariates the weights applied to the data points change smoothly on the scale of the distance to the point $(\mathbf{x}, \mathbf{z})$, while in the space of discrete covariates they take some discrete values, one for the case $\mathbf{Z}^i = \mathbf{z}$ and the others for $\mathbf{Z}^i \neq \mathbf{z}$. Specifically, we use a product kernel $w_c^i(\mathbf{x}) \times w_d^i(\mathbf{z})$ for the weights of $(\mathbf{X}^i, \mathbf{Z}^i)$ around $(\mathbf{x}, \mathbf{z})$, where

$$w_c^i(\mathbf{x}) = \prod_{j=1}^{d} K_{h_j}(x_j, X_j^i), \quad w_d^i(\mathbf{z}) = \prod_{j=1}^{k} \lambda_j^{I(Z_j^i \neq z_j)}.$$

Here, $I(A)$ denotes the indicator such that $I(A) = 1$ if A holds, and zero otherwise, $K_h(u, v) = h^{-1}K(h^{-1}(u - v))$ is a symmetric kernel function K, while bandwidths h_j and λ_j are real numbers such that $0 \leq \lambda_j \leq 1$. The above kernel scheme for the discrete covariates $\mathbf{Z}^i$ is due to Racine and Li (2004) and is in the spirit of Aitchison and Aitken (1976). Note that this approach is different from a special case where λ is set to be zero (e.g., see Harding and Pagan (2011) for the case of Nadaraya-Watson estimator). Indeed, as pointed out by Racine and Li (2004), by setting bandwidths of the discrete variables to zero, the "estimator reverts back to the conventional approach whereby one uses a frequency estimator to deal with the discrete variables", i.e., one performs separate estimation for each group identified by the discrete variable (but with the same bandwidths for the continuous variables). In fact, an important particular case is when $\lambda_j = 1$, implying that the discrete regressor j is irrelevant (see Hall, Li and Racine (2007)). Thus, allowing for the flexibility that each λ_j can be anywhere between 0 and 1 is important both for generalizing the asymptotic theory as well as for the applied work. One can also generalize further by allowing more adaptive bandwidths, e.g., allowing for bandwidths of some or all continuous variables to vary across groups defined by some or all discrete variables, as was discussed in Li, Simar and Zelenyuk (2016). Here we proceed with Aitchison-Aitken/Racine-Li type kernel for the sake of simplicity.

Furthermore, note that approximating $f(\mathbf{X}^i, \mathbf{Z}^i)$ locally by $f(\mathbf{x}, \mathbf{z})$ does not make use of the link function and the quasi-likelihood since it gives an estimator that results from using the local least squares criterion. We take the following local approximation which is linear in the direction of the continuous covariates and constant in the direction of the discrete covariates.

$$f(\mathbf{u}, \mathbf{v}) \simeq \tilde{f}(\mathbf{u}, \mathbf{v}) \equiv f(\mathbf{x}, \mathbf{z}) + \sum_{j=1}^{d} f_j(\mathbf{x}, \mathbf{z})(u_j - x_j), \tag{2.1}$$

where $f_j(\mathbf{x}, \mathbf{z}) = \partial f(\mathbf{x}, \mathbf{z})/\partial x_j$. To estimate $f(\mathbf{x}, \mathbf{z})$ and its partial derivatives $f_j(\mathbf{x}, \mathbf{z})$, we maximize

$$n^{-1} \sum_{i=1}^{n} w_c^i(\mathbf{x}) w_d^i(\mathbf{z}) Q\left(g^{-1}\left(\beta_0 + \sum_{j=1}^{d} \beta_j (X_j^i - x_j)\right), Y^i\right), \tag{2.2}$$

with respect to β_j, $0 \le j \le d$. The maximizer $\hat{\beta}_0$ is the estimator of $f(\mathbf{x}, \mathbf{z})$ and $\hat{\beta}_j$ are the estimators of $f_j(\mathbf{x}, \mathbf{z})$, respectively. Then, one can estimate the mean function $m(\mathbf{x}, \mathbf{z})$ by inverting the link function, $g^{-1}(\hat{\beta}_0)$.

Our theory given in the next section tells us that the asymptotic properties of the estimators do not depend largely on the choice of link function g as long as it is sufficiently smooth and strictly increasing. This is mainly because the estimation is performed locally. Approximating locally the function $g_1(m(\mathbf{x}, \mathbf{z}))$ or $g_2(m(\mathbf{x}, \mathbf{z}))$ for two different links g_1 and g_2 does not make much difference. However, it is suggested to use the canonical link when it is available since its use guarantees the objective function (2.2) to be convex so that the optimization procedure is numerically stable.

When the likelihood of the conditional mean function is available, one may use it in place of the quasi-likelihood Q in the description of our method. This is particularly the case when the response Y is binary. In the latter case $P(Y = y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = m(\mathbf{x}, \mathbf{z})^y[1 - m(\mathbf{x}, \mathbf{z})]^{1-y}$, $y = 0, 1$. Thus, one may replace $Q(\mu, y)$ by

$$\ell(\mu, y) = y \log\left(\frac{\mu}{1-\mu}\right) + \log(1 - \mu).$$

The canonical link g in this case is the logit function defined by $g(t) = \log(t/(1 - t))$. If one uses the *logit* link (e.g., see Frölich (2006)), then one maximizes, instead of (2.2),

$$\begin{aligned} & n^{-1} \sum_{i=1}^{n} w_c^i(\mathbf{x}) w_d^i(\mathbf{z}) \ell\left(g^{-1}\left(\beta_0 + \sum_{j=1}^{d} \beta_j (X_j^i - x_j)\right), Y^i\right) \\ & \quad = n^{-1} \sum_{i=1}^{n} w_c^i(\mathbf{x}) w_d^i(\mathbf{z}) \left[Y^i \left(\beta_0 + \sum_{j=1}^{d} \beta_j (X_j^i - x_j)\right) - \log\left(1 + e^{\beta_0 + \sum_{j=1}^{d} \beta_j (X_j^i - x_j)}\right)\right]. \end{aligned} \tag{2.3}$$

If one uses the *probit* link $g(t) = \Phi^{-1}(t)$ where Φ denotes the cumulative distribution function of the standard normal distribution, then one maximizes

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} w_c^i(\mathbf{x}) w_d^i(\mathbf{z}) & \left[Y^i \log\left(\frac{\Phi\left(\beta_0 + \sum_{j=1}^{d} \beta_j (X_j^i - x_j)\right)}{1 - \Phi\left(\beta_0 + \sum_{j=1}^{d} \beta_j (X_j^i - x_j)\right)}\right)\right. \\ & \left. + \log\left(1 - \Phi\left(\beta_0 + \sum_{j=1}^{d} \beta_j (X_j^i - x_j)\right)\right)\right]. \end{aligned} \tag{2.4}$$

Also note that when Y is binary, our local likelihood approach is related to the binary choice model formulated as

$$Y^i = I\left(f(\mathbf{X}^i, \mathbf{Z}^i) - \varepsilon^i \geq 0\right). \tag{2.5}$$

Thus, the model is a non-parametric extension of the parametric model considered by de Jong and Woutersen (2011) where it is assumed that f is a linear function and ε^i is independent of $(\mathbf{X}^i, \mathbf{Z}^i)$. When ε^i has a distribution function G, then $m(\mathbf{x}, \mathbf{z}) = P(\varepsilon^i \leq f(\mathbf{x}, \mathbf{z})) = G(f(\mathbf{x}, \mathbf{z}))$. Thus, the non-parametric binary choice model (2.5) leads to our local likelihood with link $g = G^{-1}$. For example, the local likelihood (2.3) is obtained when ε^i has the standard logistic distribution with distribution function of the form $G(u) = e^u(1 + e^u)^{-1}$, while the one at (2.4) corresponds to the case where ε^i has the standard normal distribution. In this respect, the choice of a link function amounts to choosing an error distribution in the binary response model.

3 Theoretical Properties

3.1 Assumptions

Here, we collect the assumptions for our theoretical results. Throughout the paper we assume $h_j \sim n^{-1/(d+4)}$, which is known to be the optimal rate for the bandwidths h_j. For the weights λ_j we assume $\lambda_j \sim n^{-2/(d+4)}$. This assumption is mainly for simplicity in the presentation of the theory. Basically, it makes the smoothing bias in the space of the continuous covariates and the one in the space of the discrete covariates be of the same order of magnitudes.

The joint distribution of the response variable Y and the vector of discrete covariates $\mathbf{Z}$ has a discrete measure with a finite support. For the kernel K, we assume that it is bounded, symmetric, nonnegative, compactly supported, say $[-1, 1]$. Without a loss of generality we also assume it integrates to one, i.e., $\int K(u)\, du = 1$.

We also assume the marginal density function of $\mathbf{X}$ is supported on $[0, 1]^d$, and the joint density $p(\mathbf{x}, \mathbf{z})$ of $(\mathbf{X}, \mathbf{Z})$ is continuous in $\mathbf{x}$ for all $\mathbf{z}$, and is bounded away from zero on its support, while the conditional variance $\sigma^2(\mathbf{x}, \mathbf{z}) = \text{var}(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ is continuous in $\mathbf{x}$. We also assume the mean function $m(\mathbf{x}, \mathbf{z})$ is twice continuously differentiable in $\mathbf{x}$ for each $\mathbf{z}$. These are standard conditions for kernel smoothing that are modified for the inclusion of the vector of discrete covariates $\mathbf{Z}$.

Now, we state the conditions on the stationary process $\{(\mathbf{X}^i, \mathbf{Z}^i, Y^i)\}$. The conditional density of $(\mathbf{X}^i, \mathbf{Z}^i)$ given Y^i exists and is bounded. The conditional density of $(\mathbf{X}^i, \mathbf{Z}^i, \mathbf{X}^{i+l}, \mathbf{Z}^{i+l})$ given (Y^i, Y^{i+l}) exists and is bounded. For the mixing coefficients

$$\alpha(j) \equiv \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_j^\infty} \big| P(A \cap B) - P(A)P(B) \big|,$$

where $\mathcal{F}_a^b$ denotes the σ-field generated by $\{(\mathbf{X}^i, \mathbf{Z}^i, Y^i) : a \le i \le b\}$, we assume

$$\alpha(j) \le (\text{const})(j \log j)^{-(d+2)(2d+5)/4}, \tag{3.1}$$

for all sufficiently large j. The assumptions on the conditional densities are also made in Masry (1996) where some uniform consistency results are established for local polynomial regression with strongly mixing processes. Our condition (3.1) on the mixing coefficients is a modification of those assumed in Masry (1996) that fits for our setting.

We also assume typical conditions that are needed for the theory of the quasi-likelihood approach. Specifically, we assume that the quasi-likelihood $Q(\mu, y)$ is three times continuously differentiable with respect to μ for each y in the support of Y, $\partial^2 Q(g^{-1}(u), y)/\partial u^2 < 0$ for all u in the range of the mean regression function and for all y in the support of Y, the link function g is three times continuously differentiable, V is twice continuously differentiable, V and g' are bounded away from zero on the range of the mean regression function, and the second and the third derivatives of g are bounded.

3.2 Main Theoretical Results

In this section we give the asymptotic distribution of $\hat{f}(\mathbf{x}, \mathbf{z})$. Let p denote the density function of $(\mathbf{X}, \mathbf{Z})$ and $f_{jk}(\mathbf{x}, \mathbf{z}) = \partial^2 f(\mathbf{x}, \mathbf{z})/(\partial x_j \partial x_k)$. In the discussion below, we fix $(\mathbf{x}, \mathbf{z})$ at which we estimate the mean function f. For the vector $\mathbf{z}$, we let $\mathbf{z}_{-j}$ denote the $(k-1)$-vector which is obtained by deleting the jth entry of $\mathbf{z}$.

Define $\hat{\alpha}_0 = \hat{f}(\mathbf{x}, \mathbf{z}) - f(\mathbf{x}, \mathbf{z})$ and $\hat{\alpha}_j = h_j(\hat{f}_j(\mathbf{x}, \mathbf{z}) - f_j(\mathbf{x}, \mathbf{z}))$ for $1 \le j \le d$. By the definition of $\tilde{f}$ at (2.1), it follows that the tuple $(\hat{\alpha}_j : 0 \le j \le d)$ is the solution of the equation $\hat{\mathbf{F}}(\boldsymbol{\alpha}) = \mathbf{0}$, where $\hat{\mathbf{F}}(\boldsymbol{\alpha}) = (\hat{F}_0(\boldsymbol{\alpha}), \hat{F}_1(\boldsymbol{\alpha}), \ldots, \hat{F}_d(\boldsymbol{\alpha}))^\top$,

$$\begin{aligned}\hat{F}_0(\boldsymbol{\alpha}) &= n^{-1} \sum_{i=1}^n w_c^i w_d^i \frac{Y^i - m^i(\tilde{f}, \boldsymbol{\alpha})}{V(m^i(\tilde{f}, \boldsymbol{\alpha}))g'(m^i(\tilde{f}, \boldsymbol{\alpha}))}, \\ \hat{F}_j(\boldsymbol{\alpha}) &= n^{-1} \sum_{i=1}^n w_c^i w_d^i \left(\frac{X_j^i - x_j}{h_j}\right) \frac{Y^i - m^i(\tilde{f}, \boldsymbol{\alpha})}{V(m^i(\tilde{f}, \boldsymbol{\alpha}))g'(m^i(\tilde{f}, \boldsymbol{\alpha}))}, \quad 1 \le j \le d,\end{aligned}$$

and g' is the first derivative of the link function g. Here, we suppress $\mathbf{x}$ and $\mathbf{z}$ in w_c^i and w_d^i, and also write for simplicity

$$m^i(\theta, \boldsymbol{\alpha}) = g^{-1}\left(\theta(\mathbf{X}^i, \mathbf{Z}^i) + \alpha_0 + \sum_{j=1}^d \alpha_j \left(\frac{X_j^i - x_j}{h_j}\right)\right),$$

for a function θ defined on $\mathbb{R}^d \times \mathbb{R}^k$. As approximations of $\hat{F}_j$ for $0 \le j \le d$, let

$$\begin{aligned}F_0^*(\boldsymbol{\alpha}) &= E\left[w_c^i w_d^i \frac{m^i(f, \mathbf{0}) - m^i(f, \boldsymbol{\alpha})}{V(m^i(f, \boldsymbol{\alpha}))g'(m^i(f, \boldsymbol{\alpha}))}\right], \\ F_j^*(\boldsymbol{\alpha}) &= E\left[w_c^i w_d^i \left(\frac{X_j^i - x_j}{h_j}\right) \frac{m^i(f, \mathbf{0}) - m^i(f, \boldsymbol{\alpha})}{V(m^i(f, \boldsymbol{\alpha}))g'(m^i(f, \boldsymbol{\alpha}))}\right], \quad 1 \le j \le d.\end{aligned}$$

Note that $m^i(f, \mathbf{0}) = E(Y^i|\mathbf{X}^i, \mathbf{Z}^i)$. The following lemma demonstrates that $\hat{F}_j(\boldsymbol{\alpha})$ are uniformly approximated by $F_j^*(\boldsymbol{\alpha})$ for $\boldsymbol{\alpha}$ in any compact set.

Lemma 3.1. *Assume the conditions stated in subsection 3.1. Then, for any compact set* $\mathcal{C} \subset \mathbb{R}^d$

$$\sup\{|\hat{F}_j(\boldsymbol{\alpha}) - F_j^*(\boldsymbol{\alpha})| : \boldsymbol{\alpha} \in \mathcal{C}\} = O_p\left(n^{-2/(d+4)}(\log n)^{1/2}\right), \quad 0 \leq j \leq d.$$

Under the condition that $Q(g^{-1}(u), y)$ is strictly convex as a function of u, the vector $\mathbf{F}^*(\boldsymbol{\alpha}) \equiv (F_0^*(\boldsymbol{\alpha}), F_1^*(\boldsymbol{\alpha}), \ldots, F_d^*(\boldsymbol{\alpha}))^\top$ is strictly monotone as a function of $\boldsymbol{\alpha}$. Thus, the equation $\mathbf{F}^*(\boldsymbol{\alpha}) = \mathbf{0}$ has a unique solution $\boldsymbol{\alpha} = \mathbf{0}$. This and Lemma 3.1 entail $\hat{\boldsymbol{\alpha}} \to \mathbf{0}$ in probability. The convergence of $\hat{\boldsymbol{\alpha}}$ and the following lemma justify a stochastic expansion of $\hat{\boldsymbol{\alpha}}$. To state the lemma, we define some terms that approximate the partial derivatives $\hat{F}_{jj'}(\boldsymbol{\alpha}) \equiv \partial F_j(\boldsymbol{\alpha})/\partial \alpha_{j'}$. Let

$$\begin{aligned}
\tilde{F}_{00}(\boldsymbol{\alpha}) &= E\left[\frac{w_c^i w_d^i}{V(m^i(\tilde{f}, \boldsymbol{\alpha}))g'(m^i(\tilde{f}, \boldsymbol{\alpha}))^2}\right], \\
\tilde{F}_{0j}(\boldsymbol{\alpha}) &= E\left[\left(\frac{X_j^i - x_j}{h_j}\right)\frac{w_c^i w_d^i}{V(m^i(\tilde{f}, \boldsymbol{\alpha}))g'(m^i(\tilde{f}, \boldsymbol{\alpha}))^2}\right], \quad 1 \leq j \leq d, \\
\tilde{F}_{jj'}(\boldsymbol{\alpha}) &= E\left[\left(\frac{X_j^i - x_j}{h_j}\right)\left(\frac{X_{j'}^i - x_{j'}}{h_{j'}}\right)\frac{w_c^i w_d^i}{V(m^i(\tilde{f}, \boldsymbol{\alpha}))g'(m^i(\tilde{f}, \boldsymbol{\alpha}))^2}\right], \quad 1 \leq j, j' \leq d,
\end{aligned}$$

and form a $(d+1) \times (d+1)$ matrix $\tilde{\mathbf{F}}(\boldsymbol{\alpha})$ with these terms.

Lemma 3.2. *Assume the conditions stated in subsection 3.1. Then, for any compact set* $\mathcal{C} \subset \mathbb{R}^d$

$$\sup\{|\hat{F}_{jj'}(\boldsymbol{\alpha}) - \tilde{F}_{jj'}(\boldsymbol{\alpha})| : \boldsymbol{\alpha} \in \mathcal{C}\} = O_p\left(n^{-2/(d+4)}(\log n)^{1/2}\right), \quad 0 \leq j, j' \leq d.$$

We note that $\tilde{F}_{jj'}(\boldsymbol{\alpha})$ are continuous functions of $\boldsymbol{\alpha}$. Thus, it follows that $\tilde{F}_{jj'}(\hat{\boldsymbol{\alpha}}^*) = \tilde{F}_{jj'}(\mathbf{0}) + o_p(1)$ for any stochastic $\hat{\boldsymbol{\alpha}}^*$ such that $\|\hat{\boldsymbol{\alpha}}^*\| \leq \|\hat{\boldsymbol{\alpha}}\|$. This with $\hat{F}(\hat{\boldsymbol{\alpha}}) = \mathbf{0}$ and Lemma 3.2 implies

$$\hat{\boldsymbol{\alpha}} = -\tilde{\mathbf{F}}(\mathbf{0})^{-1}\hat{\mathbf{F}}(\mathbf{0}) + o_p(n^{-2/(d+4)}). \tag{3.2}$$

In the above approximation we have also used the fact $\hat{\mathbf{F}}(\mathbf{0}) = O_p(n^{-2/(d+4)})$ which is a direct consequence of the following lemma.

Lemma 3.3. *Assume the conditions stated in subsection 3.1. Then,*

$$
\begin{aligned}
&(nh_1 \times \cdots \times h_d)^{1/2} \left[\frac{\sigma^2(\mathbf{x},\mathbf{z})p(\mathbf{x},\mathbf{z})}{V(m(\mathbf{x},\mathbf{z}))^2 g'(m(\mathbf{x},\mathbf{z}))^2} \right]^{-1/2} \mathbf{D}_1^{-1/2} \\
&\times \left[\hat{\mathbf{F}}(\mathbf{0}) - \frac{p(\mathbf{x},\mathbf{z})}{V(m(\mathbf{x},\mathbf{z}))g'(m(\mathbf{x},\mathbf{z}))^2} \left(\frac{1}{2}\mathbf{e}_0 \sum_{j=1}^{d} f_{jj}(\mathbf{x},\mathbf{z}) h_j^2 \int u^2 K(u)\, du + \mathbf{e}_0 b(\mathbf{x},\mathbf{z}) \right) \right] \\
&\qquad \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_{d+1}),
\end{aligned}
$$

where $\mathbf{I}_{d+1}$ denotes the $(d+1)$-dimensional identity matrix, $\mathbf{e}_0$ is the $(d+1)$-dimensional unit vector $(1, 0, \ldots, 0)^\top$, $\mathbf{D}_1$ is a $(d+1) \times (d+1)$ diagonal matrix with the first entry being $(\int K^2(u)\, du)^d$ and the rest $(\int K^2(u)\, du)^{d-1} \int u^2 K^2(u)\, du$ and

$$
b(\mathbf{x},\mathbf{z}) = g'(m(\mathbf{x},\mathbf{z})) \sum_{j=1}^{k} \lambda_j \sum_{z'_j \neq z_j, z'_j \in \mathcal{D}_j} \frac{p(\mathbf{x},\mathbf{z}_{-j}, z'_j)}{p(\mathbf{x},\mathbf{z})} [m(\mathbf{x},\mathbf{z}_{-j}, z'_j) - m(\mathbf{x},\mathbf{z})].
$$

Also, it follows that $\tilde{\mathbf{F}}(\mathbf{0}) = -\mathbf{D}_2 \cdot V(m(\mathbf{x},\mathbf{z}))^{-1} g'(m(\mathbf{x},\mathbf{z}))^{-2} p(\mathbf{x},\mathbf{z}) + o(1)$, where $\mathbf{D}_2$ is a $(d+1) \times (d+1)$ diagonal matrix with the first entry being 1 and the rest $\int u^2 K(u)\, du$.

In Lemma 3.3, we see that the asymptotic variance does not involve the discrete weights λ_j. This is because the contributions to the variance by the terms in $\hat{F}_j(\mathbf{0})$ with $w_d^i < 1$ are negligible in comparison to those by the terms with $w_d^i = 1$ which corresponds to the case where $\mathbf{Z}^i = \mathbf{z}$. This is not the case for the asymptotic bias. Note that the conditional mean of the ith term in $\hat{F}_j(\mathbf{0})$ given $(\mathbf{X}^i, \mathbf{Z}^i)$ contains the factor $m^i(f, \mathbf{0}) - m^i(\tilde{f}, \mathbf{0}) = g^{-1}(f(\mathbf{X}^i, \mathbf{Z}^i)) - g^{-1}(\tilde{f}(\mathbf{X}^i, \mathbf{Z}^i))$. For $\mathbf{Z}^i = \mathbf{z}$, it equals $g^{-1}(f(\mathbf{X}^i, \mathbf{z})) - g^{-1}(f(\mathbf{x},\mathbf{z}) + \sum_{j=1}^d f_j(\mathbf{x},\mathbf{z})(X_j^i - x_j))$, so that the leading terms come from the approximation of f along the direction of $\mathbf{X}^i$. However, $\mathbf{Z}^i$ with $\mathbf{Z}^i \neq \mathbf{z}$ also contribute nonnegligible bias. Note that in this case we have

$$
\begin{aligned}
&m^i(f, \mathbf{0}) - m^i(\tilde{f}, \mathbf{0}) \\
&\quad \simeq g^{-1}\left(f(\mathbf{x}, \mathbf{Z}^i) + \sum_{j=1}^{d} f_j(\mathbf{x}, \mathbf{Z}^i)(X_j^i - x_j) \right) - g^{-1}\left(f(\mathbf{x}, \mathbf{z}) + \sum_{j=1}^{d} f_j(\mathbf{x}, \mathbf{z})(X_j^i - x_j) \right) \\
&\quad \simeq g^{-1}(f(\mathbf{x}, \mathbf{Z}^i)) - g^{-1}(f(\mathbf{x}, \mathbf{z})),
\end{aligned}
$$

where the error of the first approximation is of order $n^{-2/(d+4)}$ and the second one of order $n^{-1/(d+4)}$ for $\mathbf{X}^i$ in the bandwidth range, i.e., for $\mathbf{X}^i$ with $w_c^i > 0$. When the discrete kernel weights w_d^i are applied to the differences, the leading contributions of the differences are made by $\mathbf{Z}^i$ with $\sum_{j=1}^k I(Z_j^i \neq z_j) = 1$ and they are of the magnitude $\lambda_j \sim n^{-2/(d+4)}$.

From (3.2) and Lemma 3.3, we have the following theorem.

Theorem 3.1. *Assume the conditions stated in subsection 3.1. Then, we have*

$$(nh_1 \times \cdots \times h_d)^{1/2} \left[\frac{g'(m(\mathbf{x},\mathbf{z}))^2\sigma^2(\mathbf{x},\mathbf{z})}{p(\mathbf{x},\mathbf{z})}\right]^{-1/2} \left(\int K^2(u)\,du\right)^{-d/2}$$
$$\times \left[\hat{f}(\mathbf{x},\mathbf{z}) - f(\mathbf{x},\mathbf{z}) - \frac{1}{2}\sum_{j=1}^{d} f_{jj}(\mathbf{x},\mathbf{z})h_j^2 \int u^2K(u)\,du - b(\mathbf{x},\mathbf{z})\right] \xrightarrow{d} N(0,1).$$

The theorem stated above tells that the asymptotic distribution of the estimator $\hat{f}$ is normal and invariant under the misspecification of the conditional variance $\sigma^2(\mathbf{x},\mathbf{z})$ in terms of the mean function $m(\mathbf{x},\mathbf{z})$, that is, the asymptotic distribution does not change even if $\sigma^2(\mathbf{x},\mathbf{z})) \neq V(m(\mathbf{x},\mathbf{z}))$. A close investigation into the term $\tilde{\mathbf{F}}(\mathbf{0})$ and Lemma 3.3 reveals that the term $V(m(\mathbf{x},\mathbf{z}))$ cancels out in the asymptotic variance of $\hat{f}(\mathbf{x},\mathbf{z})$. As for the asymptotic bias of the estimator, the term $\sum_{j=1}^d f_{jj}(\mathbf{x},\mathbf{z})h_j^2 \int u^2K(u)\,du/2$ typically appears in nonparametric smoothing over a continuous multivariate regressor, while the term $b(\mathbf{x},\mathbf{z})$ is due to the discrete kernel smoothing.

4 Bandwidths

The asymptotic theory summarized in previous section is derived for any bandwidths satisfying the mentioned convergence rates, namely $h_j \propto n^{-1/(d+4)}$ and $\lambda_j \propto n^{-2/(d+4)}$, and so, theoretically, they are not influenced when they are scaled by a constant. In practice, however, the selection of bandwidths is an important matter. Usually, a small variation of the bandwidths do not lead to dramatic changes in estimation results (as is also confirmed in simulations below), but big changes in the bandwidths may be influential. Indeed, very large values for bandwidths can lead to oversmoothing of the data. On the other hand, choosing very small values may result in overfitting. For a discrete variable, taking very large bandwidth (1 in the limit) would be equivalent to ignoring or omitting the discrete variable. On the other hand, taking very small bandwidth for a discrete variable would be equivalent to treating the different categories as completely different groups (only related through the common continuous bandwidths). It is thus possible that the choice of the bandwidths in practice may influence not only quantitative but also qualitative conclusions implied by the regression estimates and so this choice must be made carefully.

To implement our estimator one may use various approaches already suggested in the literature. Investigating which one of them is the best is by large a subject in itself that is beyond the scope of this paper and so we limit our discussion here to a few practical tips and explanations of what we used in the simulation sections that follow.

The simplest and very fast way, which is quite commonly used in the field of kernel-based estimation, is to start with some types of rules-of-thumb. For example, at various instances below, we make use of the so-called Silverman-type rule-of-thumb adapted to the regression context, which for a continuous variable X is given by

$$h_0(X) = 1.06 \times n^{-1/(4+d)} \hat{\sigma}_X, \tag{4.1}$$

where $\hat{\sigma}_X$ is the empirical standard deviation of observations on variable X. Similarly, for a discrete variable, we make a use of the following rule-of-thumb bandwidth value

$$\lambda_0 = n^{-2/(d+4)}. \tag{4.2}$$

Another, more sophisticated and much more computer intensive, approach is to use a data driven procedure to select the bandwidths *optimally* with respect to some desirable criterion. One of the most popular of such approaches, for example, is based on the so-called leave-one-out cross-validation (CV) criterion, e.g., where one selects such bandwidths h_{cv} and λ_{cv} that jointly maximize the following likelihood-based cross-validation criterion

$$CV(h, \lambda) \;\; = \;\; \frac{1}{n} \sum_{i=1}^{n} \ell \left(g^{-1} \left(\hat{f}_{h,\lambda}^{(-i)}(\mathbf{X}^i, \mathbf{Z}^i) \right), Y^i \right), \tag{4.3}$$

where $\hat{f}_{h,\lambda}^{(-i)}(\mathbf{X}^i, \mathbf{Z}^i)$ is the estimate of the function f at the point $(\mathbf{X}^i, \mathbf{Z}^i)$ computed from the 'leave-the i^{th} observation-out' sample with the value (h, λ) for the bandwidths. (Here, it might be worth noting that if there are no continuous regressors, the CV choice of λ will converge to zero at the rate n^{-1}.)

The statistical properties of CV bandwidth selectors in kernel regression with only continuous type covariates were first studied by Härdle, Hall and Marron (1988). See also Hall and Johnstone (1992) for smoothing parameter selection based on various empirical functionals. It is widely known that a CV bandwidth $\hat{h}$ converges to its optimum, say h_{opt}, in the sense that $(\hat{h} - h_{opt})/h_{opt} = o_p(1)$, which means $\hat{h} = O_p(n^{-1/(d+4)})$ in our context. Racine and Li (2004) extended this result to the case where there is a discrete covariate. They proved that both the bandwidth selectors $\hat{h}$ and $\hat{\lambda}$ based on a CV criterion have the properties that $(\hat{h} - h_{opt})/h_{opt} = o_p(1)$ and $(\hat{\lambda} - \lambda_{opt})/\lambda_{opt} = o_p(1)$, where h_{opt} and λ_{opt} are the corresponding theoretically optimal bandwidths such that $h_{opt} \asymp n^{-1/(d+4)}$ and

$\lambda_{opt} \asymp n^{-2/(d+4)}$. One may prove the same results in our context so that the bandwidth selectors $\hat{h}$ and $\hat{\lambda}$ that minimize the CV criterion (4.3) have asymptotically the magnitudes $n^{-1/(d+4)}$ and $n^{-2/(d+4)}$, respectively.

Besides the rule-of-thumb and the CV bandwidths, many other approaches suggested in the literature can also be used for our estimator (as long as they satisfy the theoretical rates). For example, additional flexibility can be added by allowing more adaptive bandwidths, e.g., some or all bandwidths for continuous variables may be allowed to vary with some or all of the discrete variables (e.g., see Li, Simar and Zelenyuk (2016)).

It should be also noted that maximization of CV function or its variations often is a relatively challenging task, especially for high dimensions and large samples and typically requires numerical optimization. The rule-of-thumb estimates of the bandwidths are often used for the starting values to initiate the iterations of the numerical optimization. Depending on a sample, $CV(h, \lambda)$ may have multiple local minima, some of which are 'spurious' in the sense described by Hall and Marron (1991) in the context of density estimation (also see Park and Marron (1990) for related discussion). Here, this could lead to too small h, leading to overfitting or an even worse value of h such that the local linear estimator is not defined; so, imposing lower bounds on h could prevent to such degenerate solutions.

It is also worth noting that selecting the bandwidth via CV is the most computer intensive part of all the estimation procedure here. For example, an estimation of a model of the type described in Example 2 and 3 below with a given bandwidth is taking about 0.1 and 1 minutes for a sample of $n = 100$ and a sample of $n = 1000$, respectively (on machine with 1.3 GHz Intel Core i5 with 8GB 1600 MHz DDR3). Meanwhile, for the same models, obtaining CV bandwidths by minimizing (4.3) took about 30 minutes and about 2 hours for a sample of $n = 100$ and a sample of $n = 1000$ respectively. While such timing seems not excessive for one or even several estimations, it is prohibitively expensive to do for each of the many replications in the Monte Carlo (MC) simulations. Therefore, some simplified strategy for bandwidths estimation in the MC study is needed.

To expedite the computations in the simulations below, we use the following strategy: for each scenario and different types of sample sizes, we estimate CV-optimal bandwidths by minimizing (4.3) (using several starting values including (4.1) and (4.2)) and compared the performance of the results to those obtained by using the rule-of-thumb bandwidths. We did this only for the in-sample forecasts. Our experiments generally suggested that the performance of our nonparametric approach with CV-bandwidths are very similar to those for rule-of-thumb bandwidths and so we only use the rule-of-thumb bandwidths (because they are much faster) for the MC evaluation of the out-of-sample forecasts. Even with such not optimal but appropriate and very fast bandwidths, the results from the nonparametric model are much better than from the parametric model when the latter is misspecified and very similar when the latter happens to be correctly specified.

5 Simulations

In this section we illustrate how the procedure behaves in finite samples in terms of in-sample and out-of-sample forecasts, considering three simulated situations. In the first scenario, the parametric probit with linear index (hereafter linear probit) is the true model, i.e., the idea is to see how our estimator behaves when the "world" is linear. We expect that the nonparametric estimator will be less accurate than the *correctly* specified parametric model, but it is interesting to see whether the loss is substantial.

For the second scenario we have a model where the linear probit is wrong (we add a quadratic term) and we expect that our estimator brings more accurate information on the data generating process than the linear probit. In the third example, we intensify the nonlinearity by considering a periodic index, to see if the nonparametric estimator is able to capture the essence of the true model and how much it improves upon the linear probit.

In all the examples below we generate the time series according to the following simple binary dynamic probit model

$$Y^i \sim \text{binomial}\left(P(Y^i = 1|x_i, y_{i-1})\right),\ i = 1, \ldots, n, \tag{5.1}$$

where

$$P(Y^i = 1|x_i, y_{i-1}) = \Phi(\psi(x_i, y_{i-1})),\ i = 1, \ldots, n, \tag{5.2}$$

with $X^i \sim U(lb, ub)$ and we initialize the series with $y_0 = 0$. The three examples presented below involve different specifications of $\psi(x_i, y_{i-1})$.

For each replication, we estimated several measures of quality or 'goodness of fit' to get an understanding of relative performance of the parametric linear probit and our nonparametric approach. Specifically, to measure the quality of the in-sample forecasts, the first basic measure that we used is the approximate mean squared error between the true and the estimated probabilities, i.e.,

$$AMSE_P = \frac{1}{n}\sum_{i=1}^{n}\left(P(Y^i = 1|x_i, y_{i-1}) - \widehat{P}(Y^i = 1|x_i, y_{i-1})\right)^2. \tag{5.3}$$

This measure is very useful but limited by the fact that $P(Y^i = 1|x_i, y_{i-1})$ is available only in simulated data and so many other practical alternatives were proposed and used in the literature (e.g., see Estrella (1998) and references cited there in) and we use some of them here.

Specifically, the second measure of fit we use is in the spirit of Efron (1978), defined as

$$PseudoR^2_{Efron} = 1 - AMSE_o/AMSE_c, \tag{5.4}$$

where $AMSE_o$ is the approximate mean squared error between the observation Y^i and the estimated probability, i.e.,

$$AMSE_o = \frac{1}{n}\sum_{i=1}^{n}\Big(Y^i - \widehat{P}(Y^i = 1|x_i, y_{i-1})\Big)^2, \tag{5.5}$$

while $AMSE_c$ is the approximate mean squared error of the naive estimator given by unconditional mean $\bar{Y}$, i.e.,

$$AMSE_c = \frac{1}{n}\sum_{i=1}^{n}\Big(Y^i - \bar{Y}\Big)^2, \tag{5.6}$$

and so, in a sense, this measure indicates about performance of an estimator relative to the naive approach of just looking at the unconditional mean of the sample, $\bar{Y}$ (i.e., proportions of observations where $Y^i = 1$).

Finally, we also use the Pseudo-R^2 proposed (in parametric context) by Estrella (1998), defined as

$$PseudoR^2_{Estrella} = 1 - \left(\frac{\log(L^*_u)}{\log(L^*_c)}\right)^{-2\log(L^*_c)}, \tag{5.7}$$

where $\log(L^*_u)$ is the value of the maximized (parametric or nonparametric) log-likelihood of the full (unconstrained) model and $\log(L^*_c)$ is the value of maximized log-likelihood of the constrained (or naive) model with only the intercept.

In the tables below we present the averages of these measures over $M = 100$ replications. The relatively small number of replications is dictated by the high cost of computations (due to optimization of CV criterion), but to sense the variability of these measures across all the replications ($b = 1, ..., M$), we also present the Monte Carlo standard deviations, i.e.,

$$std_{MC} \;=\; \sqrt{\frac{1}{M(M-1)}\sum_{b=1}^{M}\big(gof_b - \overline{gof}\big)^2},$$

where gof_b is a goodness of fit measure among those presented above and $\overline{gof}$ is its average over M replications.

For the same computational reasons, we mainly focus on MC results for $n \in \{25, 50, 100, 200\}$, where we used both the CV and the rule-of-thumb bandwidths. We also experimented with larger samples but using only the fast-to-compute bandwidths based on the rule-of-thumb described above and slight deviations from it (e.g., changing them by about 10%) and the conclusions were generally the same. We provide such evidence in Appendix B (for $n \in \{400, 800, 1600\}$). In this Appendix B we also provide typical plots of histograms of the bandwidths over 100 MC replications, which helps sensing the variability of CV and rule-of-thumb bandwidths across MC replications.

To investigate how the two models behave for the 'out-of-sample' forecasts, we use $AMSE_P$ as described above except that the averaging is made not over n observations, but over the 'out-of-sample' observations (which were not used in the model estimation) and their forecasts. Specifically, here we investigated the results for the forecasts one-period ahead and two-periods ahead, by starting the forecasting 10 periods before the end of the series, supposing that the value of X^i is known at least two periods in advance (i.e., we can imagine X is an exogenous variable $X^{*,i-\ell}$ with lag $\ell \geq 2$), and then rolling forward. This gave 9 out-of-sample forecasts for each type of (one-period ahead and two-periods ahead) forecasts of probabilities, which were then compared to the true probabilities. Also note that for the one-period ahead forecast, the value of y_i is available for forecasting Y^{i+1}, and so we can compute $P(Y^{i+1} = 1|x_{i+1}, y_i)$ directly. Meanwhile, for the two-periods ahead forecasts, we use the iterated approach of Kauppi and Saikkonen (2008)–we decompose the forecast according to the conditional probabilities, considering the two possible paths for the unobserved y_{i+1}, which is either 0 or 1. Specifically, we have

$$\begin{aligned} P(Y^{i+2} = 1|x_{i+1}, x_{i+2}, y_i) &= P(Y^{i+1} = 1|x_{i+1}, y_i)P(Y^{i+2} = 1|x_{i+2}, y_{i+1} = 1) \\ &+ P(Y^{i+1} = 0|x_{i+1}, y_i)P(Y^{i+2} = 1|x_{i+2}, y_{i+1} = 0), \quad (5.8) \end{aligned}$$

where the true values of all the probabilities on the right hand side are given by our probit model (5.9). We then plug in our estimates to obtain our two-periods ahead forecasts. This strategy was used in all the examples presented below.

Some remarks on the bandwidths selection are in order. Ideally, one may want to compute optimal bandwidths in each replication. Due to computational burden, however, researchers often choose to use a simple way to select bandwidths in each replication of an MC scenario, e.g., using some rule-of thumb bandwidths or even the same bandwidths over all replications (within the same scenario and the same sample size), e.g., median bandwidths obtained for a pilot of 20 or so replications. We tried all these approaches and noticed that selecting optimal bandwidths for each replication may actually yield less favorable results (e.g., higher average AMSE) than using a median of optimal bandwidths for a sub-set of replications or even relative to the rule-of-thumb bandwidths. This is due to the fact that CV sometimes gave too small or too large bandwidths, thus overfitting or oversmoothing the true models relative to the case when the same median bandwidths were used for the same scenario and sample size.

Finally, note that for the out-of-sample forecasts, the new bandwidths must be estimated with each rolling forward–because (i) new information is added and (ii) the sample size changes. Doing so with CV is too computationally intensive, and so we had to resort to a simplified strategy, where we just used the rule-of-thumb bandwidths described in the previous section, recomputing them for any changes in the sample.

5.1 Simulated Example 1

In this first example we generate the time series according to the simple dynamic *linear* index function given by

$$\psi(x_i, y_{i-1}) = \beta_0 + \beta_1 x_i + \beta_2 y_{i-1},\ i = 1, \ldots, n, \tag{5.9}$$

For the results summarized in the tables and figures below, we set $\beta_0 = -0.2$, $\beta_1 = -0.75$, $\beta_2 = 2$, $lb = -3$, $ub = 3$, while also noting that qualitatively similar conclusions were also obtained with other values of these parameters. We use figures below to visually illustrate performance of parametric and nonparametric approaches for a more or less typical MC replication (with $n = 100$), while tables below summarize results from 100 MC replications for different sample sizes.

Of course, *a priori*, we expect that, on average, the parametric linear probit approach must perform better than the nonparametric approach (although in some replications we also observed the opposite), because the latter does not use the information about the (correct) linearity of the true model while the former does. In particular, this is reflected in the faster convergence rates of the parametric approach relative to the nonparametric one (in our case, it is $\sqrt{n}$ for parametric vs. $n^{2/5}$ for the nonparametric with one continuous variable). This expectation is confirmed in most of MC replications, summarized in Table 1 and Table 2. Moreover, Figure 2 displays the time series view of the behavior of the 100 in-sample forecasts of the two approaches, where we plot the realizations of Y in the simulation (0 or 1, depicted with dots) against the respective time-series of probabilities estimated via the

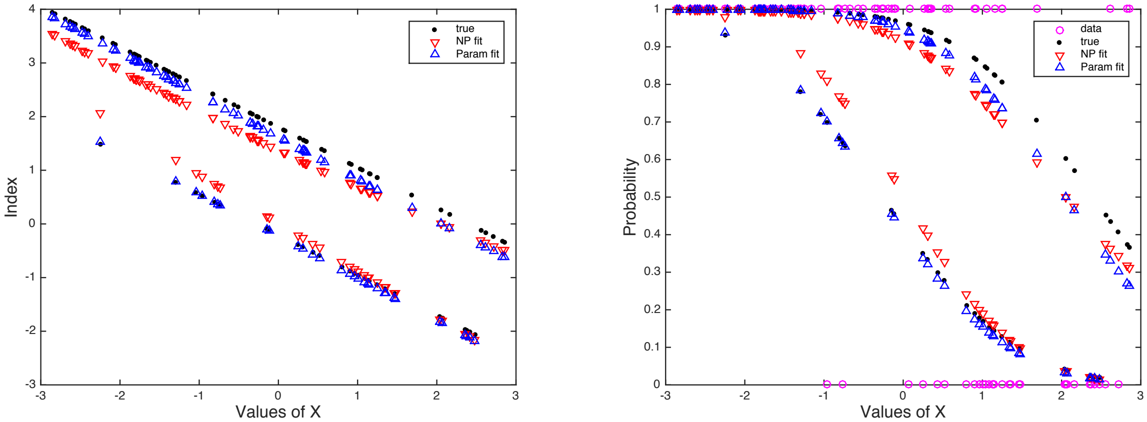

Figure 1: *Example 1, the linear probit case. Left panel, true values and estimates of the index function as a function of x, and right panel, the true values and estimates of the probabilities as function of x, evaluated at observed points. We can also see in the right panel the realized y_i. The two levels correspond to the realizations of either $y_{i-1} = 1$ (higher level) or $y_{i-1} = 0$ (lower level).*

parametric and nonparametric approaches using CV bandwidths (depicted with broken and solid curves).

Table 1 presents the averages of measures of goodness-of-fit over 100 replications and one can draw several conclusions from this scenario. First, note that the parametric approach by using correct parametric information is performing substantially better than the nonparametric method in terms of the in-sample forecasting of the true probabilities, with $AMSE_P$ for both converging to zero as sample size increases. Despite this, however, the nonparametric method is doing similarly well, and sometimes slightly better than the correctly specified parametric approach in terms of both of the $Pseudo - R^2$ measures. Also note that for $n = 25$ and $n = 50$, the nonparametric approach with the rule-of-thumb bandwidths performed better than the same approach with CV bandwidths in terms of $AMSE_P$, but the latter outperformed for the larger samples ($n = 100$ and $n = 200$), although both were already fairly close to converging to zero. The nonparametric approach with the rule-of-thumb bandwidths and with CV bandwidths showed very similar performance in terms of $Pseudo - R^2$ measures, except for $n = 25$ where the estimator with CV bandwidths outperformed. Also note that in all cases except when $n = 25$, the median of the CV bandwidths are very large, suggesting that in most cases the CV approach to bandwidths selection was able to recognize that the true model is linear by yielding CV bandwidths that is well beyond the range of simulated x implying the linear model. Overall, the Table 1 confirms that for the in-sample forecasts, the nonparametric estimator behaves well in the in-sample forecasts

Table 1: Monte-Carlo Results for the In-sample Forecasts, Example 1.

	$n = 25$			$n = 50$			$n = 100$			$n = 200$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Parametric	NP_0	NP_{cv}	Parametric	NP_0	NP_{cv}	Parametric	NP_0	NP_{cv}	Parametric	NP_0	NP_{cv}
$\overline{AMSE_P}$	0.0227	0.0276	0.0424	0.0102	0.0182	0.0211	0.0045	0.0121	0.0093	0.0023	0.0078	0.0040
std_{MC}	0.0019	0.0014	0.0036	0.0009	0.0009	0.0017	0.0004	0.0005	0.0007	0.0002	0.0003	0.0003
$\overline{PseudoR^2_{Efron}}$	0.5842	0.5724	0.6816	0.5037	0.5095	0.5085	0.5073	0.5043	0.5072	0.4806	0.4796	0.4846
std_{MC}	0.0252	0.0176	0.0270	0.0145	0.0117	0.0178	0.0096	0.0093	0.0108	0.0065	0.0063	0.0068
$\overline{PseudoR^2_{Estrella}}$	0.6109	0.5816	0.6978	0.5244	0.5234	0.5271	0.5226	0.5195	0.5253	0.4982	0.4966	0.5050
std_{MC}	0.0252	0.0179	0.0262	0.0153	0.0124	0.0184	0.0106	0.0099	0.0119	0.0075	0.0069	0.0080
median $\hat{h}$	-	0.9634	1.3361	-	0.8448	89.8606	-	0.7330	595.8548	-	0.6375	622.1360
median $\hat{\lambda}$	-	0.2759	0.1188	-	0.2091	0.1257	-	0.1585	0.0532	-	0.1201	0.0233

Notes: (i) ‘Parametric’ stands for the parametric dynamic linear *probit*; (ii) NP_0 stands for our non-parametric approach where the bandwidths were selected via the rules of thumb ($h_0 = 1.06 \times n^{-1/(4+d)} std(X)$; $\lambda_0 = n^{-2/(d+4)}$); (iii) NP_{cv} is the same method as in NP_0 but with bandwidths selected by optimizing the (leave-one-out) CV in each replication;

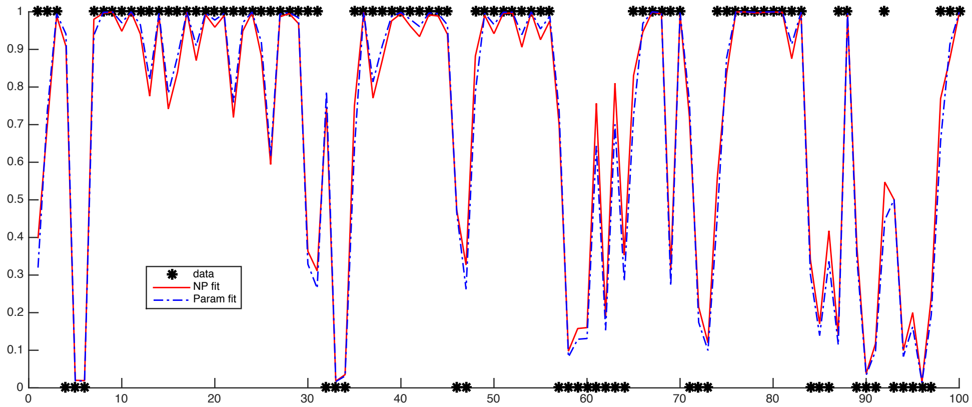

Figure 2: *Example 1: In sample forecasts of the 100 data points of the simulated series, with the linear probit and nonparametric estimates. The* $\bullet$*'s are the realizations* Y^i *(*0 *or* 1*).*

although it is not using the information about linearity, which happened to be correct in this example.

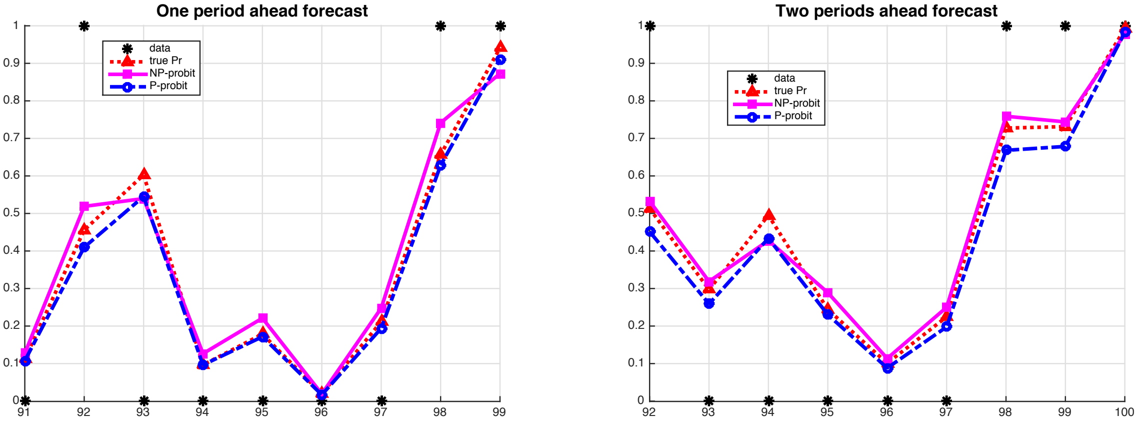

Figure 3: *Example 1: Out of sample forecasts of the 10 last observations of the series, starting with the observation 1 to 90. Left panel, one-period ahead forecasts and right panel two-periods ahead forecasts. The* $\bullet$*'s are the realizations* Y^i *(*0 *or* 1*).*

Let us now look at the performance in terms of out-of-sample forecasts. The results for one replication with $n = 100$ are presented in Figure 3, which displays the true probabilities and their out-of-sample forecasts, for one-period and two-periods ahead. The forecasts seem particularly good for both the (correctly specified) parametric estimator and the nonparametric estimators. Table 2 presents the averages over 100 replications and confirms that the parametric approach is performing substantially better than the nonparametric one. Importantly, $AMSE_P$ for both approaches tend to zero as the sample size increases. Also note

Table 2: Monte-Carlo Results for the Out-of-sample Forecasts, Example 1.

	$n = 25$			$n = 50$			$n = 100$			$n = 200$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Parametric	NP_0	$NP_0/1.1$	Parametric	NP_0	$NP_0/1.1$	Parametric	NP_0	$NP_0/1.1$	Parametric	NP_0	$NP_0/1.1$
$\overline{AMSE_P}$ (1-ahead)	0.0327	0.0764	0.0771	0.0129	0.0381	0.0381	0.0048	0.0211	0.0205	0.0023	0.0089	0.0087
std_{MC}	0.0029	0.0090	0.0092	0.0014	0.0056	0.0056	0.0007	0.0032	0.0033	0.0003	0.0016	0.0016
$\overline{AMSE_P}$ (2-ahead)	0.0368	0.0697	0.0720	0.0144	0.0327	0.0326	0.0047	0.0166	0.0174	0.0025	0.0081	0.0081
std_{MC}	0.0037	0.0085	0.0085	0.0017	0.0049	0.0048	0.0005	0.0028	0.0030	0.0004	0.0012	0.0012
median $\hat{h}$	-	0.9634	0.8759	-	0.8448	0.7680	-	0.7330	0.6664	-	0.6375	0.5796
median $\hat{\lambda}$	-	0.2759	0.2509	-	0.2091	0.1901	-	0.1585	0.1441	-	0.1201	0.1092

Notes: (i) 'Parametric' stands for the parametric dynamic linear *probit*; (ii) NP_0 stands for our non-parametric approach where the bandwidths were selected via the rule-of-thumb ($h_0 = 1.06 \times n^{-1/(4+d)} std(X)$; $\lambda_0 = n^{-2/(d+4)}$); (iii) $NP_0/1.1$ is the same method as in NP_0 but with the rule-of-thumb bandwidths divided by 1.1.

that the nonparametric approach gave similar results whether using the rule-of-thumb bandwidths or their smaller (divided by 1.1) versions, which was the case for both the one-period ahead and the two-periods ahead forecasts.

5.2 Simulated Example 2

Here we simulate the same model as in Example 1, except that we also add a quadratic term in x. The true index is now given by

$$\psi(x_i, y_{i-1}) = \beta_0 + \beta_1 x_i + \beta_2 y_{i-1} + \gamma x_i^2,$$

where β_0 and β_1 are the same as above and $\gamma = -0.5$. As expected we will observe a poor performance of the incorrectly specified linear probit approach and we will see that the nonparametric approach approximates this model fairly well.

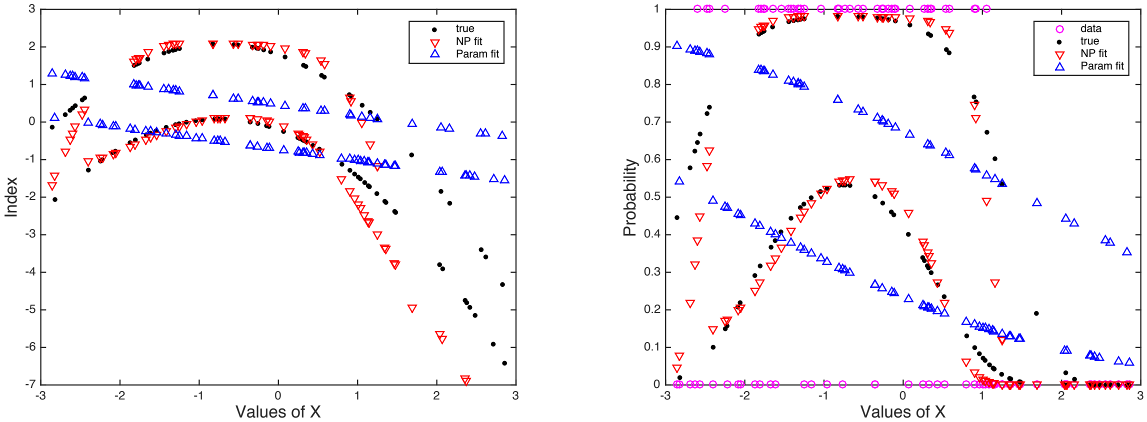

Figure 4: *Example 2, quadratic index case. Left panel, true values and estimates of the index function as a function of x, and right panel, the true values and estimates of the probabilities, as function of x, evaluated at observed points. The two levels correspond to the realizations of either $y_{i-1} = 1$ (higher level) or $y_{i-1} = 0$ (lower level).*

The results for the estimation in one typical replication with $n = 100$ observations are shown in Figure 4 for the index function (left panel) and the probabilities (right panel) and in Figure 5 for the 100 in-sample forecasts over the time series. These figures do not require much comments: the fit of the nonparametric approach here is clearly much better than that for the parametric linear probit.

Table 3 confirms the conclusions from the figures, presenting a summary of the MC results for the in-sample forecasts over 100 replications. One can see that the nonparametric

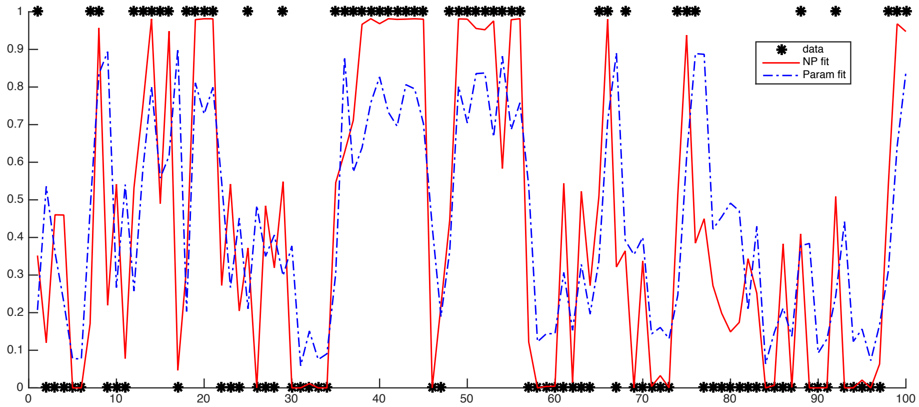

Figure 5: *Example 2, quadratic index case. In sample forecasts of the 100 data points of the simulated series, with the linear probit and nonparametric estimates. The ●'s are the realizations Y^i (0 or 1).*

approach outperforms parametric in all the in-sample goodness-of-fit measures–having substantially lower $AMSE_P$ and substantially higher $PseudoR^2_{Estrella}$ and $PseudoR^2_{Efron}$, even for such small samples as $n = 25$ and $n = 50$. Also note that in terms of $AMSE_P$, the difference in performance increases with sample size, as for the nonparametric approach it tends to zero while for the parametric approach it seems to rather quickly converge to a positive value near 0.05 rather than zero. In a sense, it is an illustration of the so-called 'root-n inconsistency'. It is also worth noting that the nonparametric approach with rule-of-thumb bandwidths showed significantly better performance in terms of $AMSE_P$ for smaller samples ($n = 25$ and $n = 50$) and very similar performance in the larger samples, relative to the nonparametric approach with CV bandwidths estimated in every replication. On the other hand, their performance was almost identical in terms of the $PseudoR^2_{Estrella}$ and $PseudoR^2_{Efron}$ for all the sample considered.

We now turn to the out-of-sample forecasts of this example. The superior performance of the nonparametric approach relative to the parametric linear probit approach is also confirmed here. The results for one typical replication are shown in Figure 6, which illustrates how the nonparametric out-of-sample forecasts (with CV bandwidths) follow rather well the true probabilities, both in the one-period and the two-periods ahead forecasts. Meanwhile, Table 4 presents the summary over 100 replications confirming the same conclusions as drawn from the figure. Specifically, one can see that the nonparametric approach, as expected, is performing substantially better than the parametric approach in terms of both (one-period and two-periods) out-of-sample forecasts of the true probabilities: An exception is for the smallest sample case ($n = 25$) where performance is somewhat similar, while already for

Table 3: Monte-Carlo Results for the In-sample Forecasts, Example 2.

	$n = 25$			$n = 50$			$n = 100$			$n = 200$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Parametric	NP_0	NP_{cv}	Parametric	NP_0	NP_{cv}	Parametric	NP_0	NP_{cv}	Parametric	NP_0	NP_{cv}
$\overline{AMSE_P}$	0.0572	0.0336	0.0447	0.0561	0.0214	0.0267	0.0552	0.0141	0.0138	0.0539	0.0090	0.0078
std_{MC}	0.0017	0.0020	0.0030	0.0009	0.0009	0.0015	0.0007	0.0006	0.0007	0.0004	0.0004	0.0005
$\overline{PseudoR^2_{Efron}}$	0.2950	0.5345	0.5364	0.3117	0.5336	0.5397	0.3025	0.5340	0.5696	0.2825	0.5063	0.5286
std_{MC}	0.0176	0.0155	0.0238	0.0119	0.0094	0.0145	0.0093	0.0085	0.0097	0.0061	0.0053	0.0066
$\overline{PseudoR^2_{Estrella}}$	0.3175	0.5704	0.5771	0.3284	0.5761	0.5877	0.3189	0.5820	0.6276	0.2963	0.5561	0.5889
std_{MC}	0.0195	0.0155	0.0237	0.0127	0.0097	0.0157	0.0101	0.0087	0.0102	0.0065	0.0059	0.0075
median $\hat{h}$	-	0.9634	1.1000	-	0.8448	0.8885	-	0.7330	0.6854	-	0.6375	0.5635
median $\hat{\lambda}$	-	0.2759	0.2490	-	0.2091	0.1110	-	0.1585	0.0669	-	0.1201	0.0385

Notes: (i) ‘Parametric’ stands for the parametric dynamic linear *probit*; (ii) NP_0 stands for our non-parametric approach where the bandwidths were selected via the rules of thumb ($h_0 = 1.06 \times n^{-1/(4+d)} std(X)$; $\lambda_0 = n^{-2/(d+4)}$); (iii) NP_{cv} is the same method as in NP_0 but with bandwidths selected by optimizing the (leave-one-out) CV in each replication;

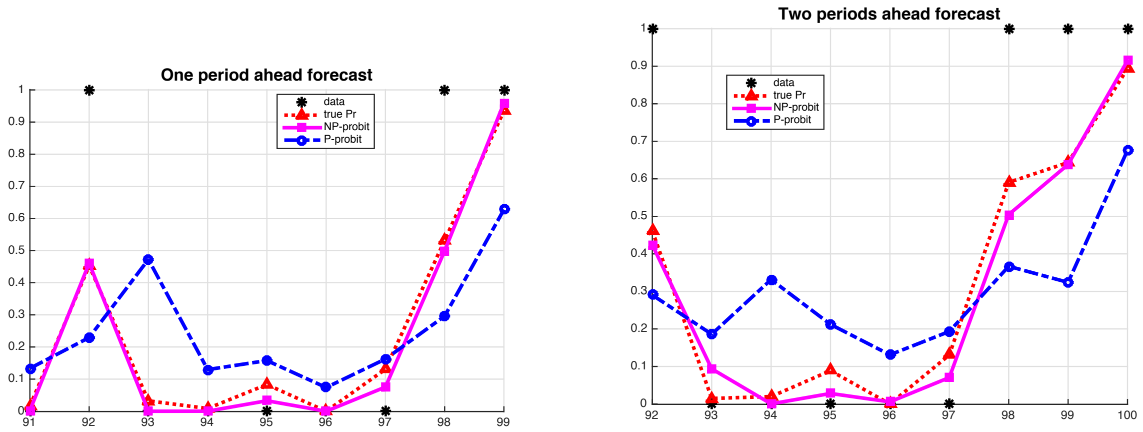

Figure 6: *Example 2, quadratic index case. Out of sample forecasts of the 10 last observations of the series, starting with the observation 1 to 90. Left panel, one-period ahead forecasts and right panel two-periods ahead forecasts. The* $\bullet$*'s are the realizations* Y^i *(*0 *or* 1*).*

$n = 50$ the difference in $AMSE_P$ became about two-fold, while about 4-times and 6-times for $n = 100$ and $n = 200$. This is because $AMSE_P$ for the nonparametric approach tends to zero while for the linear parametric model it appears to be converging to a positive (misspecification) bias around 0.05. Also note that reducing the bandwidths by about 10% from the rule-of-thumb values almost did not change the results and certainly did not change the conclusions.

5.3 Simulated Example 3

In this last example we generate the time series according to the following dynamic index function

$$\psi(x_i, y_{i-1}) = \beta_0 + \sin(\beta_1 x_i + \beta_2 y_{i-1}),\ i = 1, \ldots, n. \tag{5.10}$$

Note that the index function is also nonlinear in parameters and so it is not so simple to approximate it just by adding a quadratic term in the linear index as would have been possible in the preceding example. Also note that the discrete variable is also inside the sin-function and so its impact on the dependent variable is more complicated than just a vertical parallel shift. For the results presented below we had $\beta_0 = -0.2$, $\beta_1 = -1.75$, $\beta_2 = 2$.

Figure 7 illustrates the results for the estimation for one of replications, with $n = 100$ observations, for the index function (left panel) and the probabilities (right panel), meanwhile Figure 8 displays results in time series perspective for the 100 in-sample forecasts for this

Table 4: Monte-Carlo Results for the Out-of-sample Forecasts, Example 2.

	$n = 25$			$n = 50$			$n = 100$			$n = 200$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Parametric	NP_0	$NP_0/1.1$	Parametric	NP_0	$NP_0/1.1$	Parametric	NP_0	$NP_0/1.1$	Parametric	NP_0	$NP_0/1.1$
$\overline{AMSE_P}$ (1-ahead)	0.0935	0.0868	0.0880	0.0659	0.0321	0.0310	0.0639	0.0165	0.0156	0.0571	0.0106	0.0099
std_{MC}	0.0050	0.0094	0.0094	0.0039	0.0031	0.0031	0.0031	0.0017	0.0016	0.0024	0.0010	0.0010
$\overline{AMSE_P}$ (2-ahead)	0.0959	0.0767	0.0770	0.0703	0.0327	0.0322	0.0736	0.0178	0.0167	0.0653	0.0122	0.0113
std_{MC}	0.0049	0.0093	0.0093	0.0045	0.0035	0.0035	0.0037	0.0022	0.0020	0.0026	0.0012	0.0011
median $\hat{h}$	-	0.9634	0.8759	-	0.8448	0.7680	-	0.7330	0.6664	-	0.6375	0.5796
median $\hat{\lambda}$	-	0.2759	0.2509	-	0.2091	0.1901	-	0.1585	0.1441	-	0.1201	0.1092

Notes: (i) 'Parametric' stands for the parametric dynamic linear *probit*; (ii) NP_0 stands for our non-parametric approach where the bandwidths were selected via the rule-of- thumb ($h_0 = 1.06 \times n^{-1/(4+d)} std(X)$; $\lambda_0 = n^{-2/(d+4)}$); (iii) $NP_0/1.1$ is the same method as in NP_0 but with the rule-of-thumb bandwidths divided by 1.1.

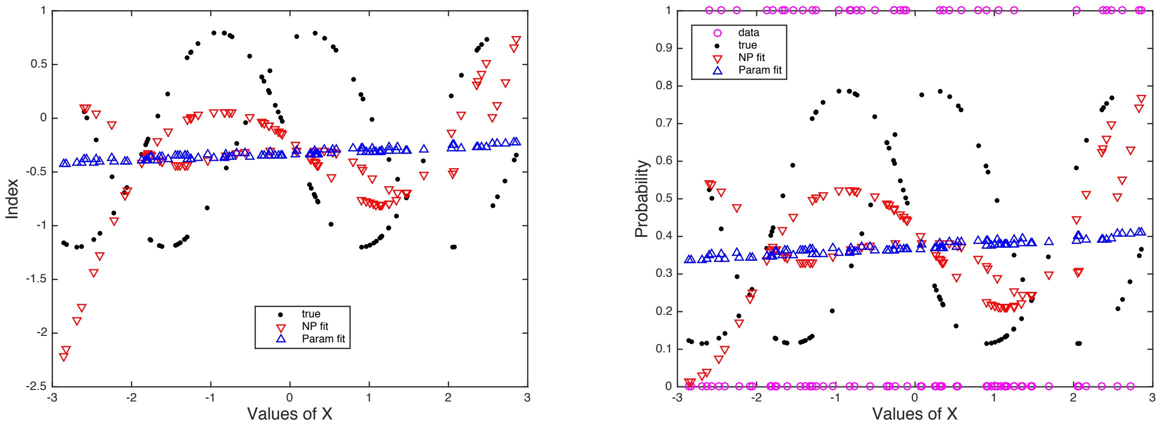

Figure 7: *Example 3, periodic index case. Left panel, true values and estimates of the index function as a function of x, and right panel, the true values and estimates of the probabilities, as function of x, evaluated at observed points. The two levels correspond to the realizations of either $y_{i-1} = 1$ (higher level) or $y_{i-1} = 0$ (lower level).*

same replication. As should be expected, all figures show a rather poor behavior of the parametric linear probit approach and much better (although not perfect) performance of the nonparametric approach, which captures the periodic nature of the true model. Indeed, the parametric linear probit approach here suggests that both the index function and the probabilities are almost flat with respect to x and forecasts probabilities that are fluctuating around 0.4, which is very different from the true model that exhibits a periodic relationship with respect to x. (In Appendix B we also provide typical plots for $n = 1000$, to illustrate the improvement of the fit by the nonparametric approach.)

These conclusions are also confirmed by averages over 100 MC replications. Specifically, Table 5 that summarizes performance in the in-sample forecasts, suggests that the nonparametric approach generally outperforms the parametric linear probit approach in all the goodness-of-fit measures. Indeed, note that as was also the case in the previous example, the difference in performance in terms of $AMSE_P$ was increasing with an increase of the sample size–because for the nonparametric approach $AMSE_P$ tends to zero while for the parametric approach it appears to be converging to a positive value around 0.06. Also note that, as in the previous example, the nonparametric approach with the rule-of-thumb bandwidths showed significantly better performance for smaller samples ($n = 25$ and $n = 50$) and similar performance in the larger samples relative to the nonparametric approach where bandwidths were obtained by optimizing CV in every replication.

Turning to the out-of-sample forecasts, one can also see that the superior performance of the nonparametric approach relative to the parametric linear probit approach is also

Table 5: Monte-Carlo Results for the In-sample Forecasts, Example 3.

	$n = 25$			$n = 50$			$n = 100$			$n = 200$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Parametric	NP_0	NP_{cv}	Parametric	NP_0	NP_{cv}	Parametric	NP_0	NP_{cv}	Parametric	NP_0	NP_{cv}
$\overline{AMSE_P}$	0.0747	0.0540	0.0661	0.0673	0.0379	0.0488	0.0642	0.0276	0.0282	0.0623	0.0194	0.0167
std_{MC}	0.0024	0.0024	0.0026	0.0011	0.0012	0.0017	0.0006	0.0007	0.0010	0.0004	0.0004	0.0006
$\overline{PseudoR^2_{Efron}}$	0.0865	0.3315	0.2211	0.0527	0.2926	0.2782	0.0230	0.2555	0.3033	0.0152	0.2379	0.2883
std_{MC}	0.0085	0.0111	0.0225	0.0045	0.0084	0.0191	0.0020	0.0065	0.0114	0.0013	0.0049	0.0072
$\overline{PseudoR^2_{Estrella}}$	0.0940	0.3474	0.2367	0.0555	0.3060	0.2963	0.0232	0.2632	0.3183	0.0153	0.2431	0.3015
std_{MC}	0.0096	0.0115	0.0236	0.0049	0.0088	0.0204	0.0021	0.0067	0.0123	0.0013	0.0052	0.0079
median $\hat{h}$	-	0.9634	676.1182	-	0.8448	0.9358	-	0.7330	0.5680	-	0.6375	0.4691
median $\hat{\lambda}$	-	0.2759	0.4995	-	0.2091	0.2251	-	0.1585	0.1579	-	0.1201	0.0816

Notes: (i) 'Parametric' stands for the parametric dynamic linear *probit*; (ii) NP_0 stands for our non-parametric approach where the bandwidths were selected via the rules of thumb ($h_0 = 1.06 \times n^{-1/(4+d)} std(X)$; $\lambda_0 = n^{-2/(d+4)}$); (iii) NP_{cv} is the same method as in NP_0 but with bandwidths selected by optimizing the (leave-one-out) CV in each replication;

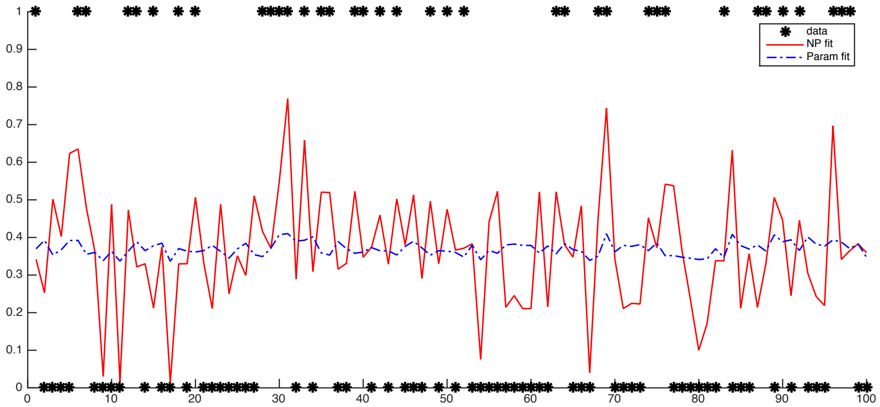

Figure 8: *Example 3, periodic index case: In sample forecasts of the 100 data points of the simulated series, with the linear probit and nonparametric estimates. The* ●*'s are the realizations Y^i (0 or 1).*

quite evident, both for the one-period and the two-periods ahead forecasts. The results are shown in Figure 9, which illustrates a typical replication, while Table 6 presents the averages over 100 replications confirming the general conclusions that we drew also in the previous example. Indeed, one can see that the nonparametric approach, as expected, is performing substantially better than the parametric approach in terms of out-of-sample forecasting of the true probabilities, except perhaps for the smallest sample case ($n = 25$) where their performance is more similar, but the difference in $AMSE_P$ reaches about 1.5 times already for $n = 50$ and about 2-times and 3-times for $n = 100$ and $n = 200$. Again, this is because $AMSE_P$ for the nonparametric approach tend to zero while for the linear parametric model it seems to be converging to a positive value around 0.06. As before, also note that reducing the bandwidths by about 10% from the rule-of-thumb values almost had no impact on results and did not change any conclusions.

6 Concluding Remarks

In this work we generalized the non-parametric quasi-likelihood method to the context of discrete response models for time series data, allowing for lags of the discrete dependent variable to appear among regressors. We derived the consistency and asymptotic normality of the estimator for such models. The theory we presented is fairly general and can be used in many areas of research. The Monte Carlo study confirmed a good performance of our nonparametric approach in finite samples, substantially improving upon the linear paramet-

Table 6: Monte-Carlo Results for the Out-of-sample Forecasts, Example 3.

	$n = 25$			$n = 50$			$n = 100$			$n = 200$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Parametric	NP_0	$NP_0/1.1$	Parametric	NP_0	$NP_0/1.1$	Parametric	NP_0	$NP_0/1.1$	Parametric	NP_0	$NP_0/1.1$
$\overline{AMSE_P}$ (1-ahead)	0.0999	0.0877	0.0895	0.0752	0.0485	0.0470	0.0672	0.0315	0.0293	0.0621	0.0202	0.0183
std_{MC}	0.0047	0.0054	0.0057	0.0027	0.0028	0.0029	0.0020	0.0015	0.0014	0.0018	0.0009	0.0008
$\overline{AMSE_P}$ (2-ahead)	0.0554	0.0547	0.0561	0.0384	0.0293	0.0289	0.0338	0.0204	0.0196	0.0302	0.0134	0.0124
std_{MC}	0.0036	0.0049	0.0050	0.0015	0.0015	0.0016	0.0012	0.0011	0.0011	0.0010	0.0007	0.0006
median $\hat{h}$	-	0.9634	0.8759	-	0.8448	0.7680	-	0.7330	0.6664	-	0.6375	0.5796
median $\hat{\lambda}$	-	0.2759	0.2509	-	0.2091	0.1901	-	0.1585	0.1441	-	0.1201	0.1092

Notes: (i) 'Parametric' stands for the parametric dynamic linear *probit*; (ii) NP_0 stands for our non-parametric approach where the bandwidths were selected via the rule-of- thumb ($h_0 = 1.06 \times n^{-1/(4+d)} std(X)$; $\lambda_0 = n^{-2/(d+4)}$); (iii) $NP_0/1.1$ is the same method as in NP_0 but with the rule-of-thumb bandwidths divided by 1.1;

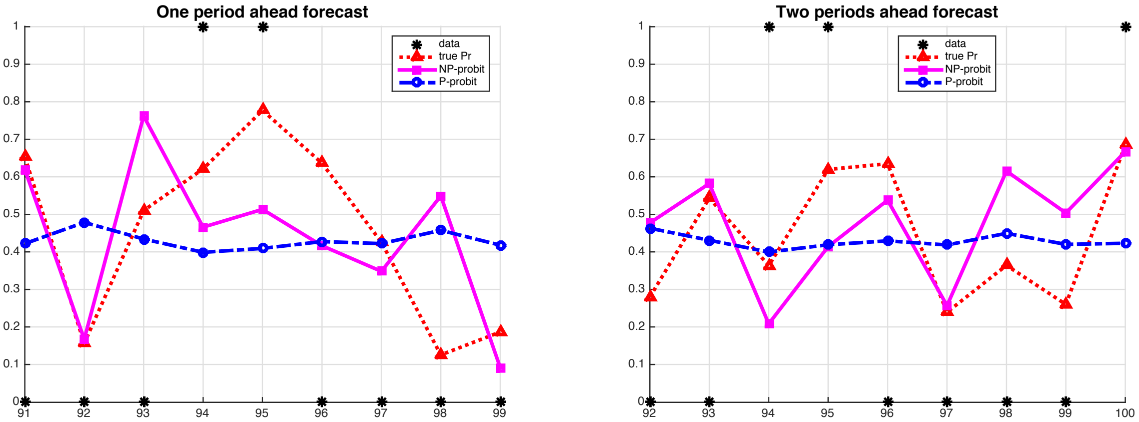

Figure 9: *Example 3, periodic index case: Out of sample forecasts of the 10 last observations of the series, starting with the observation 1 to 90. Left panel, one-period ahead forecasts and right panel two-periods ahead forecasts. The* ●*'s are the realizations* Y^i (0 *or* 1)*.*

ric probit (when the latter is misspecified), and whether using cross-validation bandwidths or the rule-of-thumb bandwidths.

Possible extensions of our work would be to extend our estimator to the case of ordered discrete choice models, the case non-stationary variables, the case of panel data, etc., which we leave for future endeavors.

Acknowledgments

The authors acknowledge the financial support provided by the ARC Discovery Grant (DP130101022), and from the "Interuniversity Attraction Pole", Phase VII (No. P7/06) of the Belgian Science Policy, as well as from the NRF Grant funded by the Korea government (MEST) (No. 20100017437) and from their respective universities: Seoul National University, Université Catholique de Louvain and The University of Queensland. We alWe are also thankful for the feedback provided by our colleagues and participants of conferences and workshops where this paper was presented. We also thank the Editor and two anonymous referees for the fruitful feedback that helped in substantially improving our earlier version of the paper.

References

[1] Aitchison, J. and Aitken, C. G. G. (1976). Multivariate Binary Discrimination by the Kernel Method. *Biometrika*, 63:3, 413-420.

[2] Bierens, H. J. (1983). Uniform Consistency of Kernel Estimators of a Regression Function Under Generalized Conditions. *Journal of the American Statistical Association* 77, 699–707.

[3] Chauvet, M. and Potter, S. (2005). Forecasting recessions using the yield curve. *Journal of Forecasting* 24:2, 77-103.

[4] Cosslett, S. R. (1983). Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model. *Econometrica*, 51(3), 765-782.

[5] Cosslett, S.R. (1987). Efficiency Bounds for Distribution-Free Estimators of the Binary Choice and the Censored Regression Models. *Econometrica*, 55(3), 559-585.

[6] Dong, Y. and Lewbel. A. (2011). Nonparametric identification of a binary random factor in cross section data. *Journal of Econometrics* 163:2, 163-171.

[7] de Jong, R. M. and T. Woutersen. (2011). Dynamic time series binary choice. *Econometric Theory*, 27, 673-702.

[8] Dueker, M. (1997). Strengthening the case for the yield curve as a predictor of U.S. recessions. *Review - Federal Reserve Bank of St. Louis*, 79(2), 41-51.

[9] Dueker, M. (2005). Dynamic Forecasts of Qualitative Variables: A Qual VAR Model of U.S. Recessions. *Journal of Business and Economic Statistics*, 23(1), 96-104.

[10] Estrella, A. (1998). A New Measure of Fit for Equations with Dichotomous Dependent Variables. *Journal of Business and Economic Statistics*, 16(2), 198-205.

[11] Estrella, A. and Mishkin, F. S. (1995). Predicting U.S. Recessions: Financial Variables as Leading Indicators. Working Paper 5379, *National Bureau of Economic Research*

[12] Estrella, A. and Mishkin, F.S. (1998). Predicting U.S. Recessions: Financial Variables as Leading Indicators. *The Review of Economics and Statistics*, 80(1), 45-61.

[13] Fan, J., N. E. Heckman and M. P. Wand. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90, 141-150.

[14] Frölich, M. (2006), Non-parametric regression for binary dependent variables, *Econometrics Journal* 9, 511–540.

[15] Hall, P., Q. Li and J. Racine (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors, *The Review of Economics and Statistics*, 89(4), 784–789.

[16] Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *Journal of American Statistical Association 83, 86*-101.

[17] Hall, P. and Marron, J. S., (1991). Local minima in crossvalidation functions. *Journal of Royal Statistical Society*, Series B, 53, 245–252.

[18] Hall, P. and Johnstone, I. (1992). Empirical functionals and efficient smoothing parameter selection (with discussion). *Journal of Royal Statistical Society*, Series B 54, 475-530

[19] Harding, D. and Pagan, A. (2011). An Econometric Analysis of Some Models for Constructed Binary Time Series. *Journal of Business and Economic Statistics*, 29(1), 86-95.

[20] Honore, B.E. and Lewbel, A. (2002). Semiparametric Binary Choice Panel Data Models Without Strictly Exogeneous Regressors. *Econometrica* 70:5, 2053-2063.

[21] Horowitz, J. L. (1992). A Smoothed Maximum Score Estimator for the Binary Response Model. *Econometrica*, 60(3), 505-531. doi: 10.2307/2951582

[22] Hu, L. and Phillips, P. C. B. (2004). Dynamics of the federal funds target rate: a nonstationary discrete choice approach. *Journal of Applied Econometrics*, 19(7), 851-867.

[23] Joon, Y. Park and Phillips, P. C. B. (2000). Nonstationary Binary Choice. *Econometrica*, 68(5), 1249-1280.

[24] Kauppi, H. (2012). Predicting the Direction of the Fed's Target Rate. *Journal of Forecasting*, 31(1), 47-67.

[25] Kauppi, H. and Saikkonen, P. (2008). Predicting U.S. Recessions with Dynamic Binary Response Models. *Review of Economics and Statistics*, 90(4), 777-791.

[26] Klein, R. W. and Spady, R. H. (1993). An Efficient Semiparametric Estimator For Binary Response Models. *Econometrica* 61:2, 387-421.

[27] Lewbel, A. (2000). Semiparametric Qualitative Response Model Estimation With Unknown Heteroscedasticity or Instrumental Variables, *Journal of Econometrics* 97, 145-177.

[28] Li, Q. and Racine, J. (2007). *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

[29] Li, D. Simar, L., and Zelenyuk, V. (2016). Generalized nonparametric smoothing with mixed discrete and continuous data. *Computational Statistics & Data Analysis* 100, 424-444.

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

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

[32] Masry, E. (1996). Multivariate local polynomial regression for times series: uniform strong consistency and rates. *Journal of Times Series Analysis* 17, 571-599.

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

[34] Matzkin, R. L. (1993). Nonparametric identification and estimation of polychotomous choice models. *Journal of Econometrics* 58, 137-168.

[35] McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. in: P. Zarembka, ed., *Frontiers of Econometrics*. New York: Academic Press.

[36] McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics* 3:4, 303-328.

[37] Moysiadis, T. and Fokianos, K. (2014). On binary and categorical time series models with feedback. *Journal of Multivariate Analysis* 131, 209-228.

[38] Park, B.U., and Marron, J.S. (1990). Comparison of Data-Driven Bandwidth Selectors. *Journal of the American Statistical Association* 85, 66-72.

[39] Racine, J. S. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* 119, 99-130.

[40] Robinson, P. M. (1983). Nonparametric Estimators for Time Series. *Journal of Time Series Analysis* 4:3, 185–207.

[41] Russell, J. R. and Engle, R. F. (1998). Econometric analysis of discrete-valued irregularly-spaced financial transactions data using a new autoregressive conditional multinomial model. SSRN eLibrary, 1998.

[42] Russell, J. R. and Engle, R. F. (2005). A discrete-state continuous-time model of financial transactions prices and times. *Journal of Business and Economic Statistics* 23, 166–180.

Appendix A: Further Theoretical Details

A.1 Proofs of Lemmas 3.1 and 3.2

We prove Lemma 3.1 for $\hat{\mathbf{F}}_0$ only. First, we observe $\hat{\mathbf{F}}_0(\boldsymbol{\alpha}) = \sum_{j=1}^4 S_j(\boldsymbol{\alpha})$, where

$$
\begin{aligned}
S_1(\boldsymbol{\alpha}) &= n^{-1}\sum_{i=1}^n w_c^i w_d^i \frac{Y^i - m^i(f,\mathbf{0})}{V(m^i(\tilde{f},\boldsymbol{\alpha}))g'(m^i(\tilde{f},\boldsymbol{\alpha}))},\\
S_2(\boldsymbol{\alpha}) &= n^{-1}\sum_{i=1}^n w_c^i w_d^i \left[\frac{m^i(f,\boldsymbol{\alpha})}{V(m^i(f,\boldsymbol{\alpha}))g'(m^i(f,\boldsymbol{\alpha}))} - \frac{m^i(\tilde{f},\boldsymbol{\alpha})}{V(m^i(\tilde{f},\boldsymbol{\alpha}))g'(m^i(\tilde{f},\boldsymbol{\alpha}))}\right],\\
S_3(\boldsymbol{\alpha}) &= n^{-1}\sum_{i=1}^n w_c^i w_d^i \left[\frac{m^i(f,\mathbf{0})}{V(m^i(\tilde{f},\boldsymbol{\alpha}))g'(m^i(\tilde{f},\boldsymbol{\alpha}))} - \frac{m^i(f,\mathbf{0})}{V(m^i(f,\boldsymbol{\alpha}))g'(m^i(f,\boldsymbol{\alpha}))}\right],\\
S_4(\boldsymbol{\alpha}) &= n^{-1}\sum_{i=1}^n w_c^i w_d^i \left[\frac{m^i(f,\mathbf{0}) - m^i(f,\boldsymbol{\alpha})}{V(m^i(f,\boldsymbol{\alpha}))g'(m^i(f,\boldsymbol{\alpha}))}\right].
\end{aligned}
$$

Let $\tau_n = n^{2/(d+4)}(\log n)^{-1/2}$. We prove

$$
\begin{aligned}
\sup_{\boldsymbol{\alpha}\in\mathcal{C}} |S_1(\boldsymbol{\alpha})| &= O_p(\tau_n^{-1}), && \text{(A.1)}\\
\sup_{\boldsymbol{\alpha}\in\mathcal{C}} |S_j(\boldsymbol{\alpha})| &= O_p(n^{-2/(d+4)}), \quad j = 2,3, && \text{(A.2)}\\
\sup_{\boldsymbol{\alpha}\in\mathcal{C}} |S_4(\boldsymbol{\alpha}) - ES_4(\boldsymbol{\alpha})| &= O_p(\tau_n^{-1}). && \text{(A.3)}
\end{aligned}
$$

The proofs of these results can be done along the lines of the proofs of Theorems 2 and 5 in Masry (1996) with some modifications (also see Bierens (1983) and Robinson (1983) and Li and Racine (2007)). Specifically, we take a finite number L_n of points in $\mathcal{C}$, denoted by $\mathcal{D}_n$, in such a way that any point in $\mathcal{C}$ has at least one point in $\mathcal{D}_n$ within a distance $L_n^{-1/(d+1)}$. We can bound $|S_1(\boldsymbol{\alpha}) - S_1(\boldsymbol{\alpha}')|$ for all $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ with $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\| \le L_n^{-1/(d+1)}$ by a constant which we can make as small as we want by choosing L_n sufficiently large. This enables us to take care of only $\max_{\boldsymbol{\alpha}\in\mathcal{D}_n} |S_j(\boldsymbol{\alpha})|$ or $\max_{\boldsymbol{\alpha}\in\mathcal{D}_n} |S_j(\boldsymbol{\alpha}) - ES_j(\boldsymbol{\alpha})|$ for (A.1)–(A.3).

For $S_1(\boldsymbol{\alpha})$, we decompose the sum into $2q_n$ equal-sized blocks $V_1, \ldots, V_{2q_n}$, so that we have $S_1(\boldsymbol{\alpha}) = \sum_{j=1}^{q_n} V_{2j-1} + \sum_{j=1}^{q_n} V_{2j}$. Here, we assume $n/(2q_n)$ is an integer without loss of generality. The blocks V_{2j-1} in the first sum are away from each other at the distance $n/(2q_n)$, and so are the blocks in the second sum. By using the strong mixing condition (3.1) we can then approximate these blocks sufficiently well by independent copies V_{2j-1}^* of V_{2j-1} and V_{2j}^* of V_{2j}. Using the independence of V_{2j-1}^* and of V_{2j}^* for different values of j we can derive an exponential inequality for $\sum_{j=1}^{q_n} V_{2j-1}^*$ and for $\sum_{j=1}^{q_n} V_{2j}^*$. For this, we need to use the strong mixing condition (3.1) again to make the covariances within each block V_{2j-1}^* or V_{2j}^*. Let $h = n^{-1/(d+4)}$. By choosing $L_n = (\tau_n/h^d)^{d+1}$ and $q_n = n/\tau_n$, we derive

$$P\left(\sup_{\boldsymbol{\alpha}\in\mathcal{C}} |S_1(\boldsymbol{\alpha})| > A_1\tau_n^{-1}\right) \le n^{C-\eta_1(A_1)} + \eta_2(A_1)\frac{nL_n}{\tau_n}\left(\frac{n}{h^d\log n}\right)^{1/4}\alpha(\tau_n), \tag{A.4}$$

where C is an absolute constant that depends on the dimension d only, η_1 is a function such that $\eta(A_1) \to \infty$ as $A_1 \to \infty$, and η_2 decreases to zero as A_1 increases. In the proof of (A.4), we have also used $\max_{1\le i\le n} w_d^i \le 1$. By the strong mixing condition (3.1), we can show that the second term at (A.4) tends to zero at a speed of $(\log n)^{-C}$ for some constant $C > 0$ as n increases. This proves the first assertion (A.1).

To prove (A.2), we claim that

$$\sup_{\boldsymbol{\alpha}\in\mathcal{C}} |S_j(\boldsymbol{\alpha}) - E(S_j(\boldsymbol{\alpha}))| = O_p(\tau_n^{-1}n^{-2/(d+4)}), \quad j = 2, 3. \tag{A.5}$$

This establishes (A.2) since $\sup_{\boldsymbol{\alpha}\in\mathcal{C}} |E(S_j(\boldsymbol{\alpha}))| = O(n^{-2/(d+4)})$ for $j = 2, 3$. The latter follows from the observation that those with $\mathbf{Z}^i = \mathbf{z}$ in the sum $S_j(\boldsymbol{\alpha})$ contribute $\sum_{j=1}^d h_j^2$, while those with $\mathbf{Z}^i \neq \mathbf{z}$ contribute $\sum_{j=1}^k \lambda_j$. Now, the proof of (A.5) is similar to that of (A.1). Using the same choices of L_n and q_n, we can obtain the same upper bound as in (A.4) for $P\left(\sup_{\boldsymbol{\alpha}\in\mathcal{C}} |S_j(\boldsymbol{\alpha}) - ES_j(\boldsymbol{\alpha})| > A_1\tau_n^{-1}n^{-2/(d+4)}\right)$. The proofs of (A.3) and Lemma 3.2 are also similar to that of (A.1).

A.2 Proof of Lemma 3.3

We write $\hat{F}_j(\mathbf{0}) = n^{-1}\sum_{i=1}^n w_c^i w_d^i U_j^i$ with appropriate definitions of U_j^i. Then,

$$\operatorname{var}(\hat{F}_j(\mathbf{0})) = n^{-2}\sum_{i=1}^n \operatorname{var}(w_c^i w_d^i U_j^i) + n^{-2}\sum_{i\neq i'} w_c^i w_c^{i'} w_d^i w_d^{i'} \operatorname{cov}(U_j^i, U_j^{i'}).$$

The second part can be shown to be negligible using the condition (3.1) on the strong mixing coefficients. The calculation of the first part can be done by the standard kernel smoothing theory. We simply note that

$$\operatorname{var}(w_c^i w_d^i U_j^i) = \operatorname{var}\left(w_c^i U_j^i I(\mathbf{Z}^i = \mathbf{z})\right) + o(n^{-d/(d+4)}).$$

Also, we can prove that $\text{cov}(\hat{F}_j(\mathbf{0}), \hat{F}_l(\mathbf{0})) = o(n^{-4/(d+4)})$ for $1 \leq j \neq l \leq d$. Furthermore, for the bias expansion, we observe that

$$E(w_c^i w_d^i U_j^i) = E[w_c^i U_j^i I(\mathbf{Z}^i = \mathbf{z})] + \sum_{l=1}^{k} \lambda_l E[w_c^i U_j^i I(Z_l \neq z_l, \mathbf{Z}_{-l} = \mathbf{z}_{-l})]. \tag{A.6}$$

Note that under the condition that $h_j \sim \lambda_j^{1/2} \sim n^{-1/(d+4)}$, both terms in (A.6) have contributions to the bias that are of magnitude $n^{-2/(d+4)}$. Finally, note that the leading terms of the two parts can be obtained by the standard kernel smoothing theory.

Appendix B: Further Results from Simulations

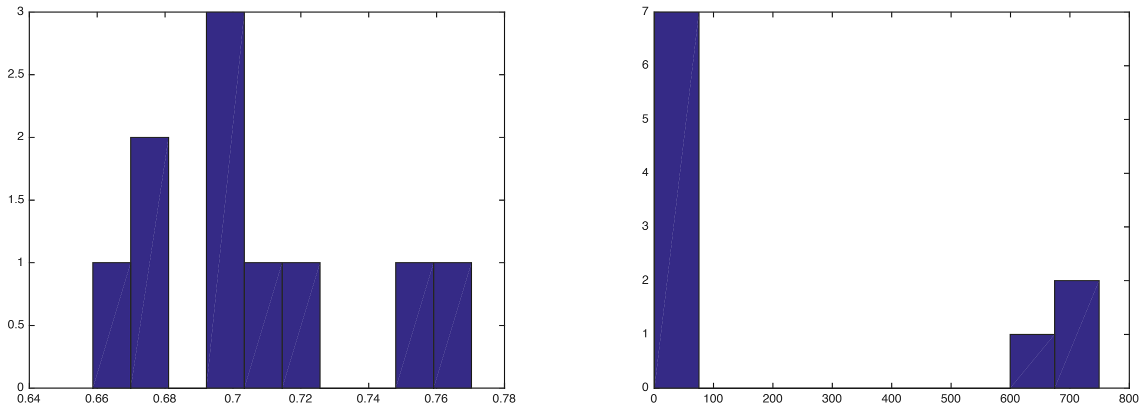

Figure 10: *Example 1 (linear index) with* $n = 100$*. Histograms for estimates of* h *in* 100 *MC replications using the rule-of thumb bandwidth (left panel) and minimization of CV criterion (right panel).*

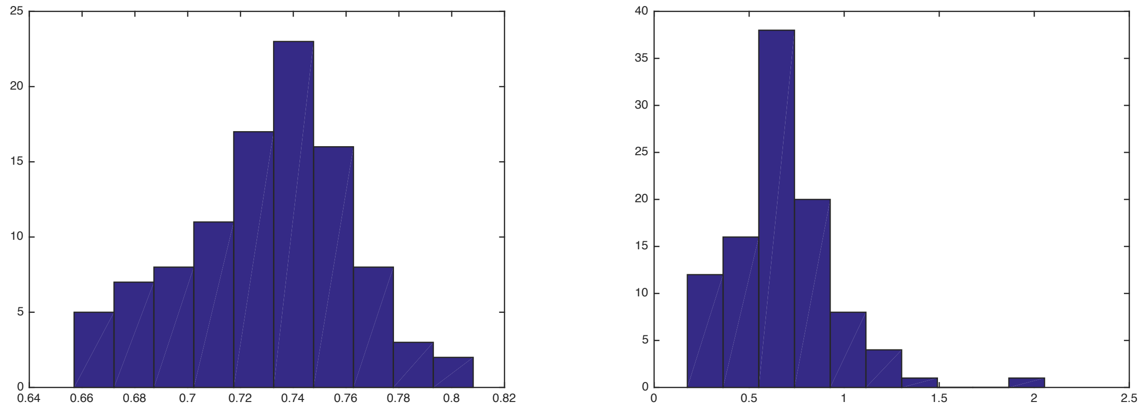

Figure 11: *Example 2 (quadratic index) with* $n = 100$*. Histograms for estimates of* h *in* 100 *MC replications using the rule-of thumb bandwidth (left panel) and minimization of CV criterion (right panel).*

Table 7: Monte-Carlo Results for In-sample Forecasts, Example 1 using rule-of-thumb bandwidths.

	$n = 400$			$n = 800$			1600		
col#	1	2	3	4	5	6	7	8	9
	Parametric	NP_0	$NP_0/1.1$	Parametric	NP_0	$NP_0/1.1$	Parametric	NP_0	$NP_0/1.1$
$\overline{AMSE_P}$	0.0011	0.0049	0.0047	0.0006	0.0028	0.0026	0.0003	0.0017	0.0016
std_{MC}	0.0001	0.0002	0.0002	0.0001	0.0001	0.0001	0.0000	0.0001	0.0001
$\overline{PseudoR^2_{Efron}}$	0.4821	0.4797	0.4836	0.4856	0.4838	0.4861	0.4814	0.4797	0.4812
std_{MC}	0.0052	0.0052	0.0052	0.0035	0.0035	0.0035	0.0023	0.0023	0.0023
$\overline{PseudoR^2_{Estrella}}$	0.4963	0.4924	0.4967	0.5046	0.5013	0.5042	0.5008	0.4980	0.4998
std_{MC}	0.0058	0.0056	0.0056	0.0040	0.0040	0.0040	0.0025	0.0025	0.0025
median $\hat{h}$	-	0.5538	0.5035	-	0.4814	0.4377	-	0.4199	0.3817
median $\hat{\lambda}$	-	0.0910	0.0828	-	0.0690	0.0627	-	0.0523	0.0475

Notes: (i) 'Parametric' stands for the parametric dynamic linear *probit*; (ii) NP_0 stands for our non-parametric approach where the bandwidths were selected via the rule-of-thumb ($h_0 = 1.06 \times n^{-1/(4+d)} std(X)$; $\lambda_0 = n^{-2/(d+4)}$); (iii) $NP_0/1.1$ is the same method as in NP_0 but with the rule-of-thumb bandwidths divided by 1.1.

Table 8: Monte-Carlo Results for In-sample Forecasts, Example 2 using rule-of-thumb bandwidths.

	$n = 400$			$n = 800$			1600		
col#	1	2	3	4	5	6	7	8	9
	Parametric	NP_0	NP_{cv}	Parametric	NP_0	NP_{cv}	Parametric	NP_0	NP_{cv}
$\overline{AMSE_P}$	0.0608	0.0120	0.0105	0.0607	0.0076	0.0065	0.0604	0.0048	0.0040
std_{MC}	0.0003	0.0003	0.0003	0.0001	0.0002	0.0002	0.0001	0.0001	0.0001
$\overline{PseudoR^2_{Efron}}$	0.0113	0.2447	0.2550	0.0087	0.2475	0.2544	0.0083	0.2460	0.2503
std_{MC}	0.0009	0.0035	0.0036	0.0006	0.0028	0.0029	0.0005	0.0020	0.0020
$\overline{PseudoR^2_{Estrella}}$	0.0114	0.2488	0.2602	0.0086	0.2516	0.2595	0.0082	0.2504	0.2555
std_{MC}	0.0009	0.0036	0.0037	0.0006	0.0029	0.0030	0.0005	0.0021	0.0021
median $\hat{h}$	NaN	0.5538	0.5035	NaN	0.4814	0.4377	NaN	0.4199	0.3817
median $\hat{\lambda}$	NaN	0.0910	0.0828	NaN	0.0690	0.0627	NaN	0.0523	0.0475

	$n = 400$			$n = 800$			1600		
col#	1	2	3	4	5	6	7	8	9
	Parametric	NP_0	$NP_0/1.1$	Parametric	NP_0	$NP_0/1.1$	Parametric	NP_0	$NP_0/1.1$
$\overline{AMSE_P}$	0.0538	0.0051	0.0047	0.0536	0.0033	0.0029	0.0535	0.0019	0.0017
std_{MC}	0.0003	0.0002	0.0002	0.0002	0.0001	0.0001	0.0002	0.0001	0.0001
$\overline{PseudoR^2_{Efron}}$	0.2831	0.5186	0.5226	0.2923	0.5150	0.5174	0.2883	0.5152	0.5167
std_{MC}	0.0042	0.0036	0.0036	0.0028	0.0028	0.0028	0.0021	0.0019	0.0019
$\overline{PseudoR^2_{Estrella}}$	0.2963	0.5731	0.5784	0.3054	0.5691	0.5726	0.3018	0.5720	0.5742
std_{MC}	0.0045	0.0041	0.0041	0.0030	0.0031	0.0031	0.0022	0.0022	0.0022
median $\hat{h}$	-	0.5538	0.5035	-	0.4814	0.4377	-	0.4199	0.3817
median $\hat{\lambda}$	-	0.0910	0.0828	-	0.0690	0.0627	-	0.0523	0.0475

Notes: (i) ‘Parametric’ stands for the parametric dynamic linear *probit*; (ii) NP_0 stands for our non-parametric approach where the bandwidths were selected via the rule-of-thumb ($h_0 = 1.06 \times n^{-1/(4+d)} std(X)$; $\lambda_0 = n^{-2/(d+4)}$); (iii) $NP_0/1.1$ is the same method as in NP_0 but with the rule-of-thumb bandwidths divided by

Table 9: Monte-Carlo Results for In-sample Forecasts, Example 3 using rule-of-thumb bandwidths.

	$n = 400$			$n = 800$			1600		
col#	1 Parametric	2 NP_0	3 NP_{cv}	4 Parametric	5 NP_0	6 NP_{cv}	7 Parametric	8 NP_0	9 NP_{cv}
$\overline{AMSE_P}$	0.0608	0.0120	0.0105	0.0607	0.0076	0.0065	0.0604	0.0048	0.0040
std_{MC}	0.0003	0.0003	0.0003	0.0001	0.0002	0.0002	0.0001	0.0001	0.0001
$\overline{PseudoR^2_{Efron}}$	0.0113	0.2447	0.2550	0.0087	0.2475	0.2544	0.0083	0.2460	0.2503
std_{MC}	0.0009	0.0035	0.0036	0.0006	0.0028	0.0029	0.0005	0.0020	0.0020
$\overline{PseudoR^2_{Estrella}}$	0.0114	0.2488	0.2602	0.0086	0.2516	0.2595	0.0082	0.2504	0.2555
std_{MC}	0.0009	0.0036	0.0037	0.0006	0.0029	0.0030	0.0005	0.0021	0.0021
median $\hat{h}$	NaN	0.5538	0.5035	NaN	0.4814	0.4377	NaN	0.4199	0.3817
median $\hat{\lambda}$	NaN	0.0910	0.0828	NaN	0.0690	0.0627	NaN	0.0523	0.0475

Notes: (i) ‘Parametric’ stands for the parametric dynamic linear *probit*; (ii) NP_0 stands for our non-parametric approach where the bandwidths were selected via the rule-of-thumb ($h_0 = 1.06 \times n^{-1/(4+d)} std(X)$; $\lambda_0 = n^{-2/(d+4)}$); (iii) $NP_0/1.1$ is the same method as in NP_0 but with the rule-of-thumb bandwidths divided by

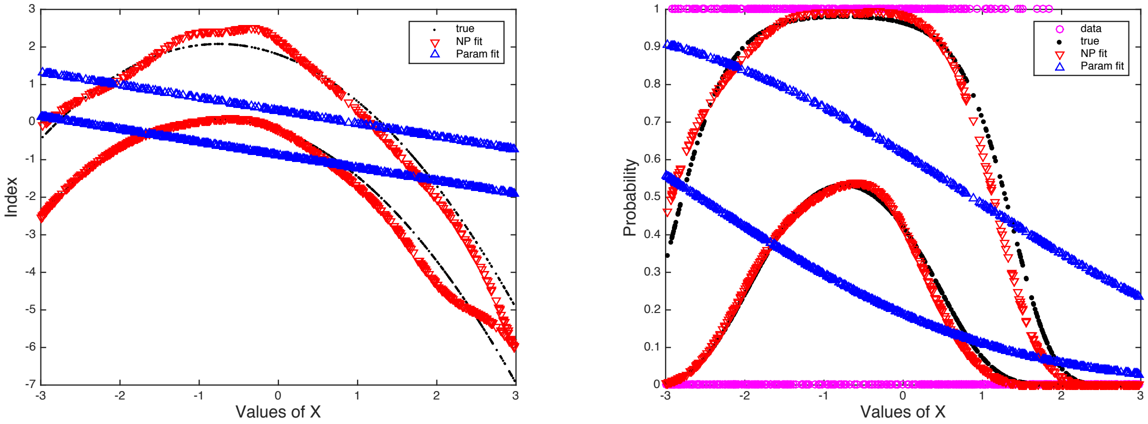

Figure 12: *Example 2 (quadratic index) with* $n = 1000$*. Left panel, true values and estimates of the index function as a function of* x*, and right panel, the true values and estimates of the probabilities, as function of* x*, evaluated at observed points. The two levels correspond to the realizations of either* $y_{i-1} = 1$ *(higher level) or* $y_{i-1} = 0$ *(lower level).*

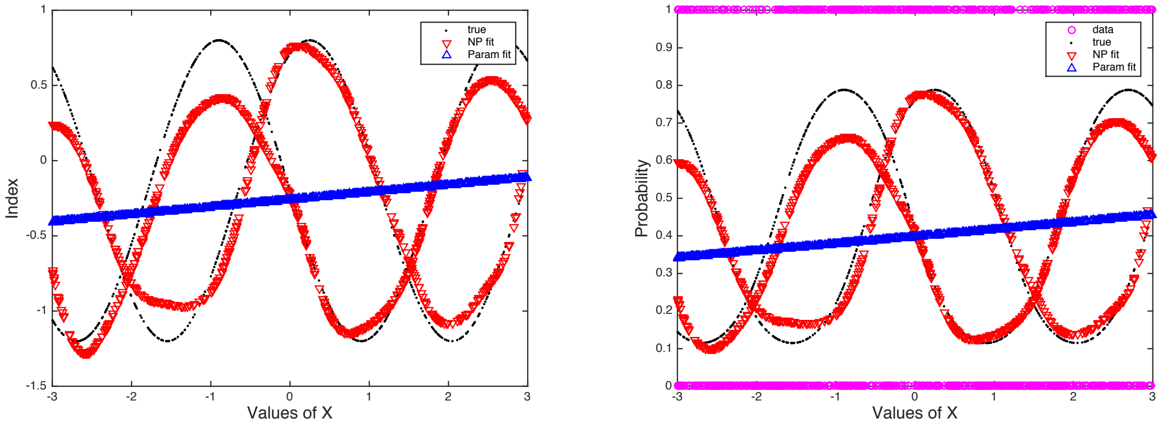

Figure 13: *Example 3 (periodic index) with* $n = 1000$*. Left panel, true values and estimates of the index function as a function of* x*, and right panel, the true values and estimates of the probabilities, as function of* x*, evaluated at observed points. The two levels correspond to the realizations of either* $y_{i-1} = 1$ *(higher level) or* $y_{i-1} = 0$ *(lower level).*