I N S T I T U T D E S T A T I S T I Q U E
B I O S T A T I S T I Q U E E T
S C I E N C E S A C T U A R I E L L E S
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

2013/52

Non-Parametric Approach to Dynamic Time Series Discrete Choice Models

PARK, B.U., SIMAR, L. and V. ZELENYUK

Non-Parametric Approach to Dynamic Time Series Discrete Choice Models

Byeong U. Park
Department of Statistics
Seoul National University, Korea

Léopold Simar
Institut de Statistique, Biostatistique et Sciences Actuarielles
Université Catholique de Louvain, Belgium

Valentin Zelenyuk
School of Economics and Centre for Efficiency and Productivity Analysis
University of Queensland, Australia.

December 10, 2013

Abstract

Dynamic discrete choice models are very important in applied research, frequently used in parametric contexts, and here we develop a non-parametric approach to such models. The main contribution of our work is generalization of the non-parametric quasi-likelihood method to the context that allows for time series models, and in particular with lags of the (discrete) dependent variable appearing among regressors. We show consistency and asymptotic normality of the estimator for such models and illustrate it with simulated and real data examples.

Key words : Nonparametric quasi-likelihood, dynamic discrete choice.
JEL : C14, C22, C25.

1 The Background

The goal of this paper is to develop methodology for the non-parametric estimation of the dynamic discrete response models, i.e., time series models with discrete dependent variables related to its own lagged values and other regressors. In the last few decades, discrete response models received substantial interest in many areas of research. Since the influential work of McFadden (1974), they became very popular in economics, especially applied microeconomics, and since the seminal work of Estrella and Mishkin (1998) they also became popular in macroeconomic analysis, where such models obtained a time series flavor. Common applications of such time series discrete choice models focus on forecasting of economic recessions, decisions of central banks on the interest rate, movements of the stock market or particular stocks, and estimation of the probability of credit defaults or an insurance event; see Estrella and Mishkin (1998), Dueker (1997, 2005), Park and Phillips (2000), Hu and Phillips (2004), Chauvet and Potter (2005), Kauppi and Saikkonen (2008), Harding and Pagan (2011), Kauppi (2012) to mention just a few.

While the approach we consider here can also be used in many other areas of research (e.g., for estimation of the probability of getting a disease, or for estimation of the probability of an earthquake, or the probability of rain, etc.), in this section we will use an economic context to motivate and to describe our approach in a less formal way, before going into the general theory presented in the next two sections. We will also come back to this economic context for the real data illustration of our method.

To facilitate further discussion, let Y^i be a discrete variable for observation (or period) i, representing different (mutually exclusive) states, e.g., such as "economy is in recession" (in a particular period i), or whether a stock market index "is in a bear territory" (i.e., falls), or whether the central bank decides to change the interest rate, etc. Let $\mathbf{X}^i$ be a d-dimensional vector of continuous regressors (which may include lagged variables) and $\mathbf{Z}^i$ be a k-dimensional vector of discrete regressors that impact Y^i through some functional relationship. Importantly, we allow $\mathbf{Z}^i$ to include lags of the dependent variable. The primary goal is then to estimate the probability of an event $Y^i = 1$ (e.g., a stock market going down), given some realizations of the explanatory variables $(\mathbf{X}^i, \mathbf{Z}^i)$, which we denote by $P(Y^i = 1|\mathbf{X}^i = \mathbf{x}, \mathbf{Z}^i = \mathbf{z})$. In the case of binary responses Y^i taking values of 0 and 1, the interest is then to estimate the conditional mean of Y^i, since

$$P(Y^i = 1|\mathbf{X}^i = \mathbf{x}, \mathbf{Z}^i = \mathbf{z}) \quad = \quad E(Y^i|\mathbf{X}^i = \mathbf{x}, \mathbf{Z}^i = \mathbf{z}). \tag{1.1}$$

The secondary goal in our approach is to estimate the marginal effects of the explanatory variables onto the probability of the event $Y^i = 1$, ceteris paribus, i.e., to estimate

$\partial E(Y^i|\mathbf{X}^i = \mathbf{x}, \mathbf{Z}^i = \mathbf{z})/\partial x_j$, for an element j of the vector $\mathbf{x}$ and $E(Y^i|\mathbf{X}^i = \mathbf{x}, \mathbf{Z}^i = \mathbf{z}) - E(Y^i|\mathbf{X}^i = \mathbf{x}, \mathbf{Z}^i = \mathbf{z}^o)$, for some particular values $\mathbf{z}$ and $\mathbf{z}^o$.

To estimate such models, empirical researchers usually impose some structure on the relationship between Y^i and $(\mathbf{X}^i, \mathbf{Z}^i)$, and the most common specification in practice appears to be

$$P(Y^i = 1|\mathbf{X}^i, \mathbf{Z}^i) \quad = \quad \Psi(\xi^i). \tag{1.2}$$

where Ψ is a cumulative distribution function (cdf), usually a standard normal or logistic distribution, and ξ^i is a random variable that aggregates the impacts of all the explanatory variables. A structure for ξ^i commonly used in practice is given by

$$\xi^i \quad = \quad (\mathbf{Z}^i)^\top \boldsymbol{\alpha} + (\mathbf{X}^i)^\top \boldsymbol{\beta} \tag{1.3}$$

where $\mathbf{Z}^i$ includes the lagged values of the discrete response variable, making it a dynamic binary choice model.[1] Such a model was extensively analyzed and compared to other approaches by Dueker (1997), and more recently by Kauppi and Saikkonen (2008). Both works confirmed the empirical importance of the dynamic component in the time-series binary choice models, finding that it substantially improves the accuracy of forecasts of US recession. Similar findings were also obtained in empirical applications of dynamic trichotomous choice models by Kauppi (2012), in forecasting the FED's interest rate decision, where the lagged discrete response variable (decision to increase, decrease or hold the rate) was also found to be very influential in improving the accuracy of forecasts. Similar examples can be mentioned in many other areas where probability models are frequently used. For example, in weather forecasting, one would also naturally expect that the lagged dependent variable, describing whether there was rain in a previous period or not, could play a very important role in explaining the probability of a rainy day coming. All in all, dynamic discrete choice models in general, and with specification (1.3) in particular, are very important in various areas of applied research, frequently implemented in the parametric context, and the goal of this work is to develop a non-parametric approach to such models: to provide the asymptotic results for it and to illustrate it with simulated and real data examples.

The reason for going non-parametric, at least as a complementary approach, is very simple, yet profound: It is well known that parametric maximum likelihood in general, and probit or logit in particular, yield inconsistent estimators if the parametric assumptions are misspecified. Not surprisingly, many important works tried to address this issue in

[1] The asymptotic theory for this model was recently derived by de Jong and Woutersen (2011).

different ways, e.g., see Cosslett (1983, 1987), Klein and Spady (1989), Horowitz (1992), Fan, Heckman and Wand (1995) and more recently Harding and Pagan (2011), to mention just a few. In a nutshell, the main contribution of our work to this literature is generalization of the non-parametric quasi-likelihood method of Fan et al. (1995) to the context that allows for time series, and in particular with lags of the (discrete) dependent variable appearing among the regressors, the practical importance of which we described above. Specifically, we derive the asymptotic theory for the local linear fit, which make it particularly convenient for estimation of derivatives or marginal effects of the regressors onto (the expected value or the probability of) the response variable. In our work we assume stationarity with a strong mixing condition (in the spirit of Masry (1996)).[2]

Our paper is structured as following: Section 2 outlines the methodology, Section 3 outlines theoretical properties of the proposed estimator, and Section 4 provides some illustrative examples for simulated and real data. Appendix provides further theoretical details and proofs.

2 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 is given in the Appendix. The response variable Y^i is of discrete type. For example, it may be binary taking the values 0 and 1. The covariate vector $\mathbf{X}^i$ 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 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 $\sigma^2(\mathbf{x}, \mathbf{z}) \equiv \text{var}(Y | \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = V(m(\mathbf{x}, \mathbf{z}))$. The other 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 parametric assumptions. Thus, a wrong choice would jeopardize estimation of

[2] 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.

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, $K_h(u, v) = h^{-1}K(h^{-1}(u - v))$ for a symmetric kernel function K and a bandwidth h, 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 the spirit of Aitchison and Aitken (1976).

It turns out 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, $0 \leq j \leq d$. 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 much depend 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 the 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, then one maximizes, instead of (2.2),

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

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

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

When Y is binary, our local likelihood approach is related to the binary choice model

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

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 $G(u) = e^u(1+e^u)^{-1}$, 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 choice model.

3 Theoretical Properties

In this section we give the asymptotic distribution of $\hat{f}(\mathbf{x}, \mathbf{z})$. Throughout the paper we assume we take $h_j \sim n^{-1/(d+4)}$ 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 kernel function K is nonnegative, symmetric and assumed to have a compact support, say $[-1, 1]$. Without loss of generality we also assume it integrates to one, $\int K(u)\, du = 1$. 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 tuples $(\hat{\alpha}_j : 0 \le j \le d)$ 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 supress $\mathbf{x}$ 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 in the Appendix. Then, for any compact set* $\mathcal{C} \subset \mathbb{R}^d$

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

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

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

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

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

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

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 $\|\hat{\boldsymbol{\alpha}}^*\| \le \|\hat{\boldsymbol{\alpha}}\|$. This with $\hat{F}(\hat{\boldsymbol{\alpha}}) = \mathbf{0}$ and Lemma 3.2 implies

$$\hat{\boldsymbol{\alpha}} = -\tilde{\mathbf{F}}(\mathbf{0})^{-1}\hat{\mathbf{F}}(\mathbf{0}) + o_p(n^{-2/(d+4)}). \tag{3.1}$$

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 in the Appendix. Then,*

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

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* $(\int K^2(u)\,du)^{d-1}\int u^2K^2(u)\,du$ *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* $\int u^2K(u)\,du$*.*

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

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

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

The above theorem tells that the asymptotic distribution of the estimator $\hat{f}$ is 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 $\sum_{j=1}^{d} f_{jj}(\mathbf{x}, \mathbf{z}) h_j^2 \int u^2 K(u)\, du/2$ 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.

While the theory above is derived for any bandwidths satisfying the mentioned convergence rates, in practice one may use a data driven procedure to optimally select the bandwidth, e.g., by using a cross-validation (CV) leave-one-out likelihood criterion. Specifically, we may select the bandwidths h, and λ that maximizes the following likelihood-based cross-validation criterion

$$CV(h, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \ell \left(g^{-1} \left(\hat{f}_{h,\lambda}^{(-i)}(\mathbf{X}^i, \mathbf{Z}^i) \right), Y^i \right), \tag{3.2}$$

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 ith observation-out" sample with the value (h, λ) for the bandwidths. It is also worth noting here that if there are no continuous regressors, the CV choice of λ will converge to zero at the rate n^{-1}. Finally, note that if $f(\mathbf{u}, \mathbf{v}) = f(\mathbf{x}, \mathbf{z}) + \sum_{j=1}^{d} f_j(\mathbf{x}, \mathbf{z})(u_j - x_j)$ exactly, i.e. the "working parametric model" is true, then one may get the parametric rate of convergence by letting $h \to \infty$.

4 Simulated and Real Data Illustrations

In this section we illustrate how the procedure behaves in finite samples in two simulated situations. In the first, the parametric probit with linear index (hereafter linear probit) is the true model, so we expect that the nonparametric estimator will be less accurate than the correctly specified parametric model, but the idea is to see how our estimator behaves when the "world" is linear. Then, for the second example, we have a model where the linear probit is wrong (we add a quadratic term) and we expect that here our estimator will bring more accurate information on the DGP than the wrong parametric one. Finally we illustrate our methodology through a real data set on recession periods in the US economy, analyzed in several papers in the literature. Of course here we do not know what is the true model so we cannot say that our nonparametric model is more appropriate, but we will see how the flexible nonparametric approach offers more insights in the DGP than the usual linear probit often suggested in the literature to model this data.

4.1 Simulated Example 1

Here we generate the time series according the simple dynamic linear probit model

$$P(Y^i = 1|x_i, y_{i-1}) = \Phi(\beta_0 + \beta_1 x_i + \beta_2 y_{i-1}),\ i = 1, \ldots, 100, \tag{4.3}$$

where $X^i \sim U(-3, 3)$ and $\beta_0 = -0.2$, $\beta_1 = -0.75$, $\beta_2 = 2$ and we initialize the series with $y_0 = 0$. The parametric linear probit estimation gives estimates of the 3 parameters of the linear index, $\widehat{\beta} = (-0.8741,\ -1.0264,\ 3.0431)$. Note that this particular random sample of size $n = 100$ is not particularly favorable to the ML parametric fit, so the resulting estimates are not so close to the true values.

The results of the estimates of the linear index and of the normal probabilities appear, with the true values in Figure 1. The nonparametric estimates seem to be quite good for this particular sample and very similar to the parametric ones. The bandwidths given by cross-validation are $h = 814.34$ (a large value indicating that the linear model is relevant) and $\lambda = 0.04467$. The two models achieve the likelihood values -0.2314 for the nonparametric fit and a slightly better value, as expected, -0.2228 for the parametric fit. The quality of the fit, when using simulated data, can be measured by the MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\Big(P(Y^i = 1|x_i, y_{i-1}) - \widehat{P}(Y^i = 1|x_i, y_{i-1})\Big)^2. \tag{4.4}$$

We obtain here a slightly better fit for the nonparametric estimator $MSE = 0.0020$ against 0.0072 for the parametric one.

We know that parametric estimators have a better rate of convergence (in our case, $\sqrt{n}$ in place of $n^{2/5}$) but this sample was, by chance, a bit more favorable to the nonparametric estimates for fitting the probabilities. Figure 2 illustrates the behavior of the 100 "in sample" forecasts of the two models. It is indeed unclear how to distinguish between the two approaches for this particular sample. Other simulated samples gave qualitatively similar results (sometimes slightly better results for the parametric fit, sometimes on the contrary, as in this particular sample). Overall, the example here illustrates that the nonparametric estimator behaves well relative to the parametric estimator, although the latter relies on a correctly specified model while the former does not.

We also investigate how the two models behave for "out of sample" prediction. Here we investigate 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 (we can imagine X is an exogenous variable $X^{*,i-\ell}$ with lag $\ell \geq 2$). Of course for the one period ahead forecast, the value of y_i is available for forecasting Y^{i+1}, and we can compute $P(Y^{i+1} = 1|x_{i+1}, y_i)$. As suggested in the iterated

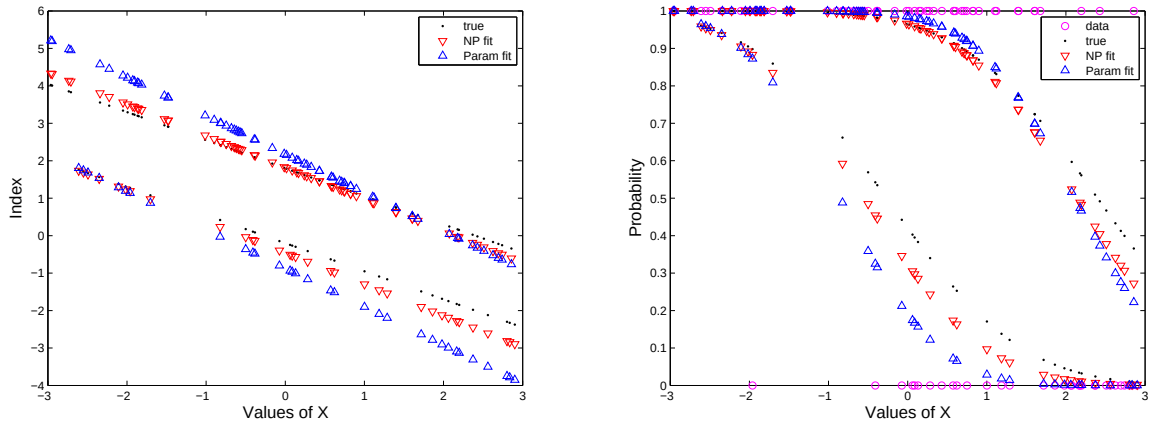

Figure 1: *Example 1, linear probit case. Left panel, true values and estimates of the index function as a function of x, and right panel, the true values and estimates of the probabilities as function of x, evaluated at observed points. We can also see in the right panel the realized y_i. The two levels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

approach of Kauppi and Saikkonen (2008), for two periods ahead we can 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

$$\begin{aligned} P(Y^{i+2} = 1|x_{i+1}, x_{i+2}, y_i) &= P(Y^{i+1} = 1|x_{i+1}, y_i)P(Y^{i+2} = 1|x_{i+2}, y_{i+1} = 1) \\ &+ P(Y^{i+1} = 0|x_{i+1}, y_i)P(Y^{i+2} = 1|x_{i+2}, y_{i+1} = 0), \quad (4.5) \end{aligned}$$

where the true values of all the probabilities on the right hand side are given by our probit model (4.3). We then plug in our estimates to obtain our two-periods ahead forecasts. Here, in this simulated situation, we can also evaluate the MSE of the forecasts. The results are given in Table 1 and Figure 3 displays the predictions and the true probabilities. The forecasts seem particularly good for both the (correctly specified) parametric estimator and the nonparametric estimators. The MSE of the forecasts confirm the slightly better performance of the nonparametric estimates for this particular sample.

Table 1: MSE of forecast values for the probabilities in Example 1.

	Nonparametric	Parametric
One period ahead	0.0020	0.0043
Two periods ahead	0.0028	0.0048

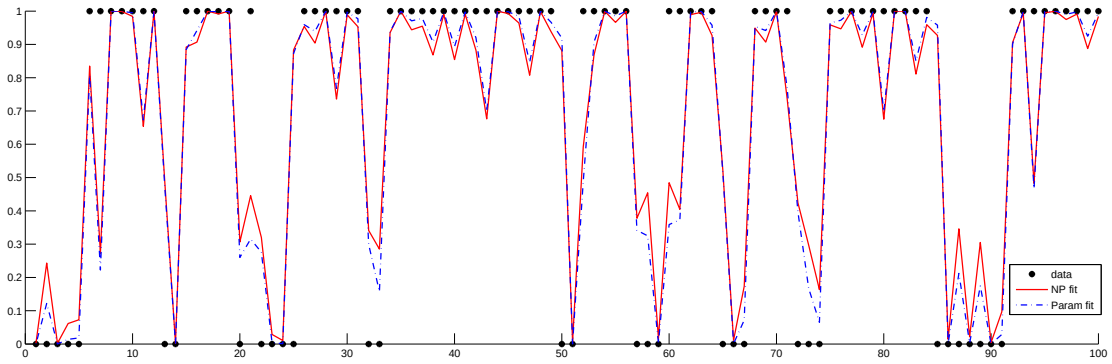

Figure 2: *Example 1: In sample forecasts of the 100 data points of the simulated series, with linear probit and nonparametric estimates. The •'s are the realizations Y_i.*

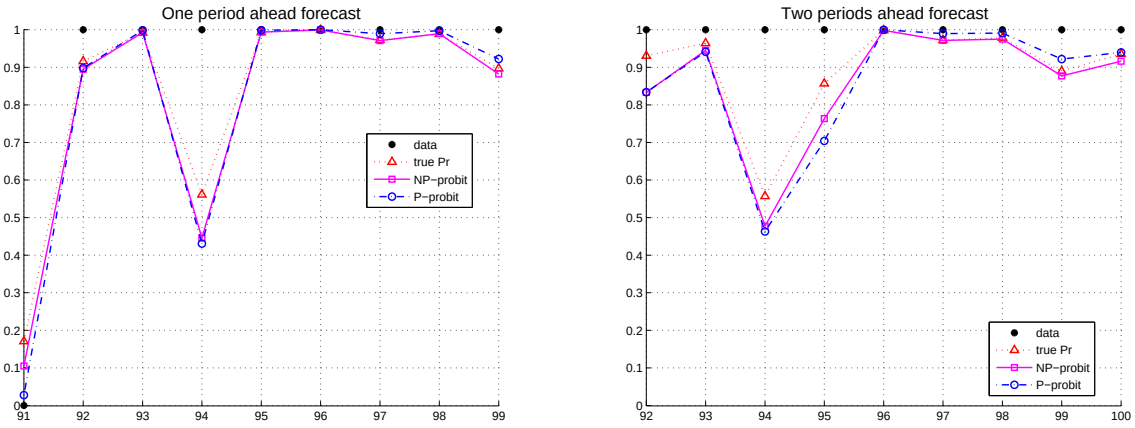

Figure 3: *Example 1: Out of sample forecasts of the 10 last observations of the series, starting with the observation 1 to 90. Left panel, one period ahead forecasts and right panel two periods ahead forecasts. The •'s are the realizations Y_i.*

4.2 Simulated Example 2

Here we simulate the same model as in Example 1, except that we add a quadratic term in x. The true index is now $\beta_0 + \beta_1 x_i + \beta_2 y_{i-1} + \gamma x_i^2$ where the vector β is the same as above and $\gamma = -0.5$. As expected we will observe a bad behavior of the incorrectly specified linear probit model and we will see that the nonparametric model handle this model without any problem.

The results for the estimation are shown in Figures 4 for the index function and the probabilities and in Figure 5 for the 100 in sample forecasts. These figures do not require much comments. The achieved bandwidths for the nonparametric model are $h = 0.6180$ and $\lambda = 0.05498$. The likelihood of the estimated models are -0.4401 for the nonparametric

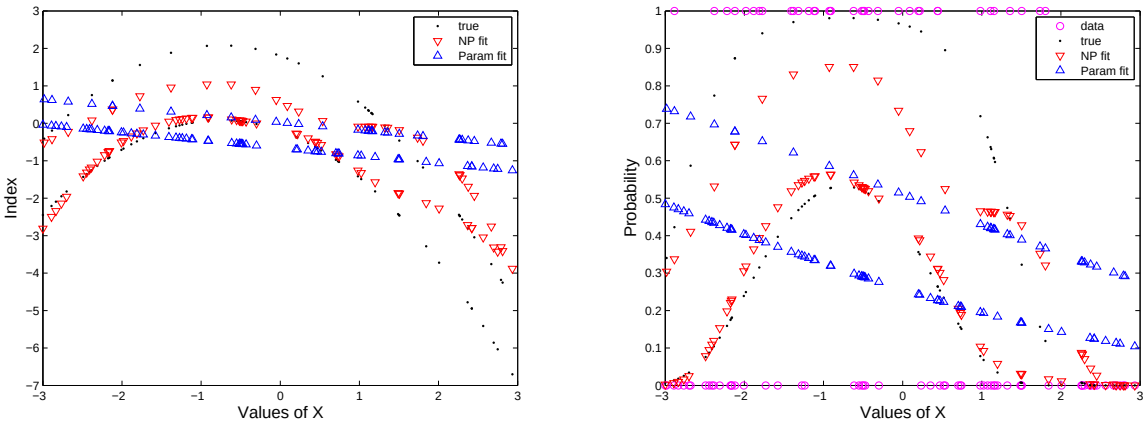

Figure 4: *Example 2, quadratic probit case. Left panel, true values and estimates of the index function as a function of x, and right panel, the true values and estimates of the probabilities, as function of x, evaluated at observed points. The two levels corresponds to the realizations of either $y_{i-1}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

model and -0.6058 for the linear probit. The MSE is 0.0112 for the nonparametric and 0.0618 for the parametric model.

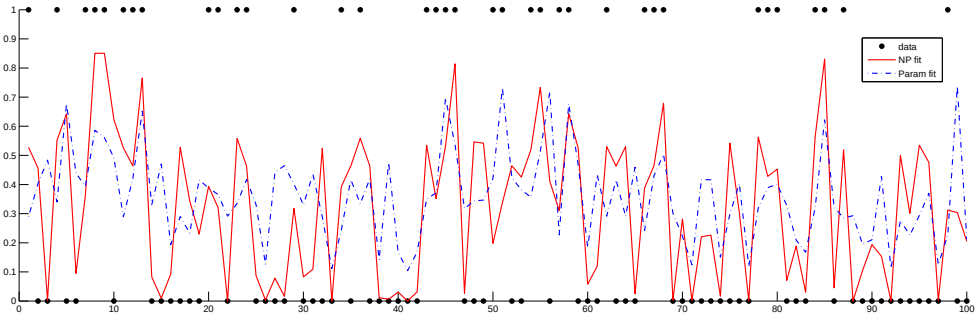

Figure 5: *Example 2: In sample forecasts of the 100 data points of the simulated series, with linear probit and nonparametric estimates. The •'s are the realizations Y_i.*

This bad expected behavior of the linear probit, compared with the nonparametric estimates is confirmed when doing the out of sample forecasts, similarly as was done for Example 1. The results are shown in Figure 6. Here the MSE of the out of sample forecast are given in Table 2. The loss in MSE for the parametric model is, as expected, huge in this case. On the contrary, Figure 6 illustrates how the nonparametric out of sample forecasts follow rather well the true probabilities; and this in both cases, the one and the two periods ahead forecasts.

Table 2: MSE of forecast values for the probabilities, Example 2.

	Nonparametric	Parametric
One period ahead	0.0043	0.0494
Two periods ahead	0.0033	0.0581

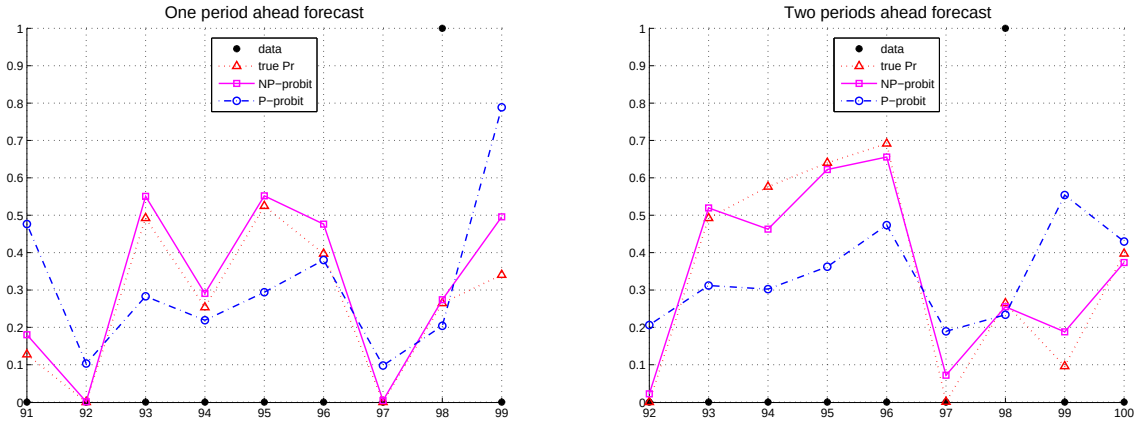

Figure 6: *Example 2: Out of sample forecasts of the 10 last observations of the series, starting with the observation 1 to 90. Left panel, one period ahead forecasts and right panel two periods ahead forecasts. The •'s are the realizations Y_i.*

4.3 Real Data Example

For the real data illustration, we have chosen a context and data set often analyzed in the literature. It is related to the prediction of recessions in an economy—a very popular application of models we consider in this paper (see e.g., Estrella and Mishkin, 1998, Dueker, 1997, Kauppi and Saikkonen, 2008, Harding and Pagan, 2011 and references cited therein).[3]

Many previous works employing parametric binary choice models suggest that a good model for the prediction of US recessions is in fact a parsimonious model with only one continuous predictor, the interest rate spread (hereafter "the spread"), and one discrete variable, the lagged dependent variable. Indeed, in their seminal work, Estrella and Mishkin (1995, 1998) using a static probit approach thoroughly investigated various models with many variables and concluded that the best forecasts resulted from a parsimonious model involving only one explanatory variable—the lagged spread. Dueker (1997) confirmed this discovery, but also found that including the lagged dependent variable among regressors substantially improved the predicting power of the model, especially for the recession of the 1970s and 1990s that were missed by various other forecasting methods. Overall, comparisons in Estrella and Mishkin (1995, 1998) and Dueker (1997) clearly show that their parsimonious

[3]The work of Harding and Pagan (2011) is the closest to ours, as it also uses a non-parametric kernel-based approach, although a different paradigm, based on the Nadaraya-Watson estimator.

model outperformed many alternative models that included many variables to gain a high in-sample fit, yet happened to be poorly forecasting the future. An important advantage of a parsimonious model is that it helps minimizing a danger of overfitting the model by improving the fit via adding more and more variables. We note that the parsimonious modelling is especially useful for our approach because it helps circumventing the curse of dimensionality problem, pertinent to virtually any non-parametric estimator, including ours. It is also more convenient for illustration of the method, which is the main goal of this empirical illustration.

Our data consists of 242 quarterly observations on US recessions and on the spread, starting from quarter 1953:Q1 to 2013:Q1. The variables are constructed following the literature on the subject, and Kauppi and Saikkonen (2008) in particular.[4] To illustrate how our approach behaves with this data set, we will also follow the analysis done in Kauppi and Saikkonen (2008) who did a careful parametric analysis for the prediction of US recessions using a parametric dynamic probit model. Specifically, we will use a model where the explanatory variable x is the spread with 4 lags and the dynamics is introduced by the one lagged value of y. The parametric probit model is

$$P(Y^t = 1|x_{t-4}, y_{t-1}) = \Phi(\beta_0 + \beta_1 x_{t-4} + \beta_2 y_{t-1}).$$

Using our data, we obtain similar parametric linear probit estimates to the ones obtained by Kauppi and Saikkonen (2008): $\widehat{\beta}_0 = -1.1818$, $\widehat{\beta}_1 = -0.4494$ and $\widehat{\beta}_2 = 2.0430$.

Table 3: One-period (1) and Two-periods (2) ahead Forecast values for Real Data example: evaluation of the performances

	Av Log Lik (1)	Estimated MSE (1)	Av Log Lik (2)	Estimated MSE (2)
Nonparametric	-0.1992	0.0653	-0.2843	0.1014
Linear Probit	-0.2062	0.0655	-0.2969	0.1026

[4]Specifically, we constructed the spread variable as the difference between the 10-year US Treasury bond rate and the 3-month US Treasury bill rate, where the information about these rates was sourced from http://www.federalreserve.gov/releases/h15/data.htm. The dependent variable is constructed by setting $y_t = 1$ if "US economy is considered as in recession" in the quarter t and 0 otherwise. It appears that there is some variation in the literature on what is meant by the statement "US economy is considered as in recession". In this illustration example, we follow Kauppi and Saikkonen (2008). Specifically, a given quarter is defined as the first quarter of a recession period if its first month or the preceding quarter's second or third month is classified by the NBER as the "business cycle peak"; A given quarter is defined as the last quarter of a recession period if its second or third month or the subsequent quarter's first month is classified by the NBER as the "business cycle trough". The first and the last quarters define a recession period, during which all quarters are recession quarters (where $y_t = 1$). All the quarters that are not included in a recession period are called expansion quarters (where $y_t = 0$). The dates of the peaks and troughs are obtained from http://www.nber.org/cycles/.

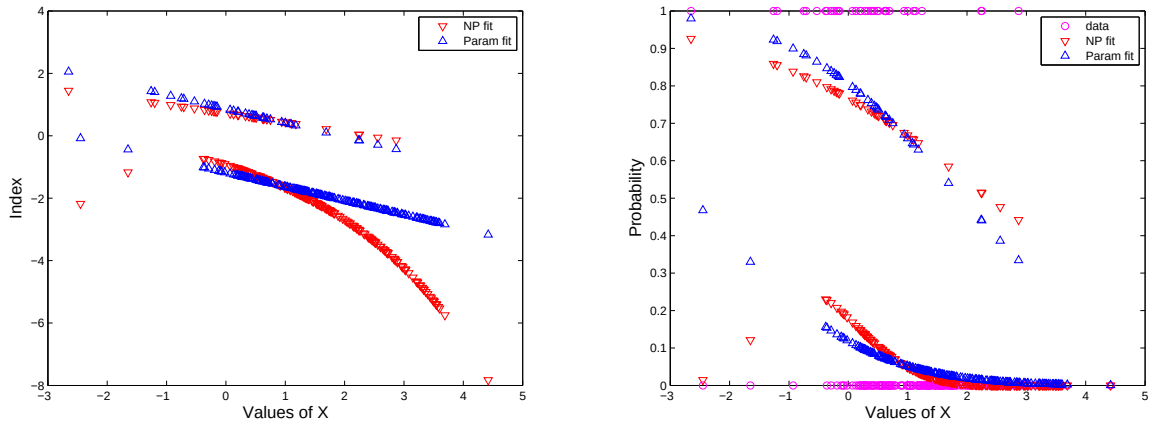

Figure 7: *Recessions in US Example. Left panel, estimates of the index function as a function of x_{t-4}, and right panel, estimates of the probabilities evaluated at observed points, as function of x_{t-4}. The two levels corresponds to the realizations of either $y_{t-1} = 1$ (recession years, higher level) or $y_{t-1} = 0$ (no recessions years, lower level).*

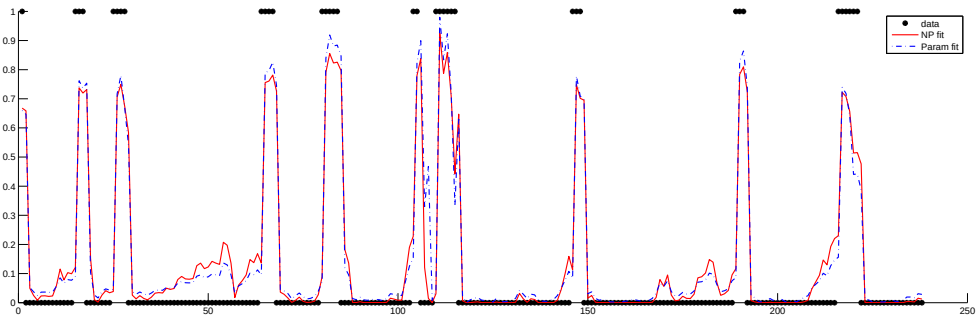

Figure 8: *Recessions in US Example. In sample forecasts, with linear probit and nonparametric estimates. The $\bullet$'s are the realizations Y_i.*

The nonparametric estimators of the index function $f(x_{t-4}, y_{t-1})$ and of the probabilities $m(x_{t-4}, y_{t-1})$ are compared to the fits from the linear probit in Figure 7. Note that for this data set we use an adaptive bandwidth for the continuous x variable, allowing the bandwidth to be different according the value of y_{t-1}. Using the cross-validation approach, we obtained the values $h_0 = 0.9879$ when $y_{t-1} = 0$, $h_1 = 369.78$ when $y_{t-1} = 1$. The resulting CV value for the discrete regressor is $\lambda = 2.2e-05$, confirming the importance of the lagged values y_{t-1} among explanatory variables in the index function. We see that for the group of data where there was no recession in the year $t-1$ ($y_{t-1} = 0$), the nonparametric estimator of the index function displays a clear curvature. For the other case, the effect of the spread is roughly linear. The curvature appears again in the nonparametric probit probabilities (right panel of the figure). Since we do not know what is the true model, we

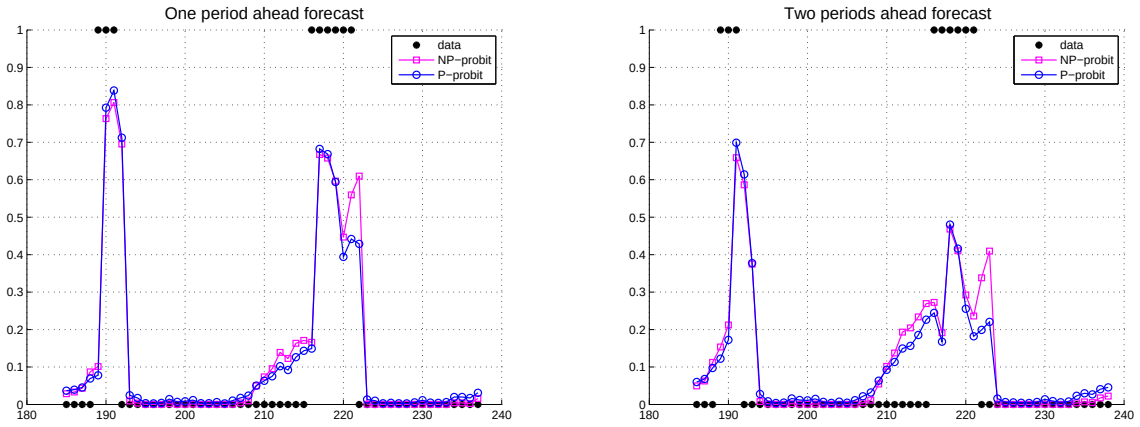

Figure 9: *Recessions in US Example. Out of sample forecasts starting the forecast the first quarter of the year 2000. Top panel, one period ahead forecasts and bottom panel two periods ahead forecasts. The* $\bullet$*'s are the realizations* Y_i*.*

cannot compute the true MSE as in (4.4) above. Usually researchers compare the achieved likelihood of the models and we obtain -0.2082 for the nonparametric approach and -0.2194 for the linear probit approach; the Pseudo-R^2, a monotone transformation of the likelihood facilitating the comparison of models, due to Estrella (1998), and also used by Kauppi and Saikkonen, are 0.4679 and 0.4431 respectively. This indicates a slight improvement for the nonparametric model, not a drastic one though. In a sense, this example indicates that Kauppi and Saikkonen (2008) did a great job when finally selecting this particular model (they tried several models and concluded that the one used here provided the best fit and the best forecasts).[5] The nice thing about the nonparametric approach is that it does not need to start with much *a priori* choices for the model because the bandwidths are selected by data driven methods. The results from the non-parametric approach may also help guiding the parametric specification, by hinting that it might benefit from adding a quadratic or an interaction term to the parametric model to fit the data better. For example, note that our results (see Figure 7) suggest that a dummy variable allowing for interaction between x_{t-4} and y_{t-1}, and including a quadratic term might be useful to make the parametric model more flexible.

Figure 8 confirms the global quality of the in sample fit (as already shown in the original study of Kauppi and Saikkonen). Here too it is difficult to see a clear difference between the two approaches.

Finally, as for the simulated examples we proceed to the out of sample forecasts to see if

[5]In fact, when we applied our estimator to the shorter period, as that used in Kauppi and Saikkonen (2008), the nonparametric approach yielded almost linear fit and so very similar results to those from the linear probit.

using the data from the beginning till 2000:Q1 we can predict (one period and two periods ahead) the recession. The linear probit and the nonparametric approach give very similar results for predictions, as shown in Table 3. For the measure of the quality of the fit we report the likelihood of the observed realizations of Y_t starting the second quarter of 2000, till the end of the series. Also we provide the estimate of the MSE. We observe again slightly better values for the nonparametric estimator, both for the one and for the two periods ahead forecasts. Interestingly, the differences are almost negligible here. The forecast values themselves are displayed in Figure 9. We see that the recession of the 90's and of the 2000's are warned by both models, but better with the one period ahead forecasts.

Acknowledgments

All authors acknowledge the financial support from ARC Discovery Grant (DP130101022), support from the "Interuniversity Attraction Pole", Phase VII (No. P7/06) of the Belgian Science Policy, support by the NRF Grant funded by the Korea government (MEST) (No. 20100017437) and support from their institutions, Seoul National University, Université Catholique de Louvain and The University of Queensland. We also thank for valuable comments our colleagues and participants of conferences and workshops where this paper was presented. We also thank Ms. Ailin Leng for her assistance with data collection for our empirical example.

Appendix: Technical Details

A.1 Assumptions

Here, we collect the assumptions for our theoretical results. The joint distribution of the response variable Y and discrete covariate $\mathbf{Z}$ has a discrete measure with a finite support. For the bandwidths and the discrete weights, we take $h_j \sim \lambda_j^{1/2} \sim n^{-1/(d+4)}$. For the kernel K, we assume that it is bounded, symmetric, nonnegative, compactly supported, say $[-1, 1]$, and $\int K(u)\, du = 1$. 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. The conditional variance $\sigma^2(\mathbf{x}, \mathbf{z}) = \mathrm{var}(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ is continuous in $\mathbf{x}$. 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 discrete covariate $\mathbf{Z}$.

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

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

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

$$\alpha(j) \le (\mathrm{const})(j \log j)^{-(d+2)(2d+5)/4} \tag{A.1}$$

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 (A.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.

A.2 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

$$\sup_{\boldsymbol{\alpha}\in\mathcal{C}} |S_1(\boldsymbol{\alpha})| = O_p(\tau_n^{-1}), \tag{A.2}$$

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

$$\sup_{\boldsymbol{\alpha}\in\mathcal{C}} |S_4(\boldsymbol{\alpha}) - ES_4(\boldsymbol{\alpha})| = O_p(\tau_n^{-1}). \tag{A.4}$$

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. 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.2)–(A.4).

For $S_1(\boldsymbol{\alpha})$, we decomposes 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 do the blocks in the second sum. By using the strong mixing condition (A.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 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 (A.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

$$P\left(\sup_{\boldsymbol{\alpha}\in\mathcal{C}} |S_1(\boldsymbol{\alpha})| > A_1\tau_n^{-1}\right) \le n^{C-\eta_1(A_1)} + \eta_2(A_1)\frac{nL_n}{\tau_n}\left(\frac{n}{h^d\log n}\right)^{1/4}\alpha(\tau_n), \tag{A.5}$$

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.5), we have also used $\max_{1\le i\le n} w_d^i \le 1$. By the strong mixing condition (A.1), we can show that the second term at (A.5) tends to zero at a speed of $(\log n)^{-C}$ for some constant $C > 0$ as n increases. This proves the first assertion (A.2).

To prove (A.3), we claim that

$$\sup_{\boldsymbol{\alpha}\in\mathcal{C}} |S_j(\boldsymbol{\alpha}) - E(S_j(\boldsymbol{\alpha}))| = O_p(\tau_n^{-1} n^{-2/(d+4)}), \quad j = 2, 3. \tag{A.6}$$

This establishes (A.3) 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.6) is similar to that of (A.2). Using the same choices of L_n and q_n, we can obtain the same upper bound as in (A.5) 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.4) and Lemma 3.2 are also similar to that of (A.2).

A.3 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,

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

The second part can be shown to be negligible using the condition (A.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

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

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

$$E(w_c^i w_d^i U_j^i) = E[w_c^i U_j^i I(\mathbf{Z}^i = \mathbf{z})] + \sum_{l=1}^k \lambda_l E[w_c^i U_j^i I(Z_l \neq z_l, \mathbf{Z}_{-l} = \mathbf{z}_{-l})] \tag{A.7}$$

Under the condition that $h_j \sim \lambda_j^{1/2} \sim n^{-1/(d+4)}$, both terms in (A.7) have contributions to the bias that are of magnitude $n^{-2/(d+4)}$. The leading tems of the two parts can be obtained by the standard kernel smoothing theory.

References

[1] Aitchison, J. and Aitken, C. G. G. (1976) . Multivariate Binary Discrimination by the Kernel Method . *Biometrika*, 63(3), 413-420. doi: 10.2307/2335719

[2] Chauvet, Marcelle and Potter, Simon. (2005). Forecasting recessions using the yield curve. *Journal of Forecasting*, 24(2), 77.

[3] Cosslett, Stephen R. (1983). Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model. *Econometrica*, 51(3), 765-782. doi: 10.2307/1912157

[4] Cosslett, Stephen R. (1987). Efficiency Bounds for Distribution-Free Estimators of the Binary Choice and the Censored Regression Models. *Econometrica*, 55(3), 559-585. doi: 10.2307/1913600

[5] de Jong, R. M. and T. Woutersen. (2011). Dynamic times series binary choice. *Econometric Theory*, 27, 673–702. doi: doi:10.1017/S0266466610000472

[6] Dueker, Michael. (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.

[7] Dueker, Michael. (2005). Dynamic Forecasts of Qualitative Variables: A Qual VAR Model of U.S. Recessions. *Journal of Business and Economic Statistics*, 23(1), 96-104. doi: 10.2307/27638797

[8] Estrella, Arturo. (1998). A New Measure of Fit for Equations with Dichotomous Dependent Variables. *Journal of Business and Economic Statistics*, 16(2), 198-205. doi: 10.2307/1392575

[9] Estrella, Arturo and Mishkin, Frederic S. Predicting U.S. Recessions: Financial Variables as Leading Indicators. Working Paper 5379, *National Bureau of Economic Research (Dec. 1995)*. doi: 10.2307/2646728

[10] Estrella, Arturo and Mishkin, Frederic S. (1998). Predicting U.S. Recessions: Financial Variables as Leading Indicators. *The Review of Economics and Statistics*, 80(1), 45-61. doi: 10.2307/2646728

[11] 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. doi: 10.2307/2291137

[12] Harding, Don and Pagan, Adrian. (2011). An Econometric Analysis of Some Models for Constructed Binary Time Series. *Journal of Business and Economic Statistics*, 29(1), 86-95. doi: 10.1198/jbes.2009.08005

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

[14] Hu, Ling and Phillips, Peter C. B. (2004). Dynamics of the federal funds target rate: a nonstationary discrete choice approach. *Journal of Applied Econometrics*, 19(7), 851-867. doi: 10.1002/jae.747

[15] Joon, Y. Park and Phillips, Peter C. B. (2000). Nonstationary Binary Choice. *Econometrica*, 68(5), 1249-1280. doi: 10.2307/2999449

[16] Kauppi, Heikki. (2012). Predicting the Direction of the Fed's Target Rate. *Journal of Forecasting*, 31(1), 47-67. doi: 10.1002/for.1201

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

[18] Klein, Roger W. and Spady, Richard H. (1993). An Efficient Semiparametric Estimator For Binary Response Models. *Econometrica (1986-1998)*, 61(2), 387

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

[20] McFadden, Daniel, Mas-Colell, Andreu, Mantel, Rolf and Richter, Marcel K. (1974). A characterization of community excess demand functions. *Journal of Economic Theory*, 9(4), 361-374.

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