

# DFG-SNF Research Group FOR916

Statistical Regularization and Qualitative Constraints

Stefan Hoderlein

Enno Mammen

Kyusang Yu

## Nonparametric Models in Binary Choice Fixed Effects Panel Data

Preprint FOR916 10-14

Preprint-Series of the Research Group FOR916

# Nonparametric Models in Binary Choice Fixed Effects Panel Data

Stefan Hoderlein<sup>1</sup>, Enno Mammen<sup>2</sup> and Kyusang Yu<sup>2</sup>

<sup>1</sup>) Department of Economics, Brown University, Providence, RI 02916, USA

<sup>2</sup>) Department of Economics, University of Mannheim, L 7, 3-5, 68131 Mannheim, Germany

stefan\_hoderlein@yahoo.com, emammen@rumms.uni-mannheim.de and

kyusangu@yahoo.co.kr

## SUMMARY

In this paper we extend the fixed effects approach to deal with endogeneity arising from persistent unobserved heterogeneity to nonlinear panel data with nonparametric components. Specifically, we propose a nonparametric procedure that generalizes Chamberlain's (1984) conditional logit approach. We develop an estimator based on nonlinear stochastic integral equations and provide the asymptotic property of the estimator and an iterative algorithm to implement the estimator. We analyze the small sample behavior of the estimator through a Monte Carlo study, and consider the decision to retire as an illustrative application.

**JEL Classification:** C14; C23

**Keywords:** Panel data, Binary Choice, Endogenous Regressors, Stochastic integral equation.

## 1 Introduction

In the linear model, panel data have become very popular for dealing with the issue of endogeneity arising from unobserved heterogeneity, because they allow to eliminate such potentially endogenous terms by taking differences or quasi-differences. This very appealing solution to handle persistent endogeneity can be extended to the binary choice panel data

model which is linear in parameters provided one is willing to accept that the distribution of the errors is logistic. This is the well known conditional ML approach dating back to work by Rasch (1960, 1961), see also Andersen (1970) and Chamberlain (1984).

To illustrate this approach, let  $Y_j^i$  denote the dependent variable for individual  $i$  in period  $j$ , let  $X_j^i$  denote the  $d$ -dimensional vector of regressors, let  $\alpha^i$  denote a time invariant unobservable effect that is idiosyncratic to the individual (potentially correlated with the regressors), and let  $\varepsilon_j^i$  be a time varying orthogonal individual error. The model is then given by:

$$Y_j^i = \mathbb{I} \left\{ X_j^{i'} \beta + \alpha^i + \varepsilon_j^i > 0 \right\}, \quad (1)$$

where  $\mathbb{I} \{ \}$  denotes the indicator function. In this paper, we assume that we have an *iid* sample of two periods,  $(Y_1^i, Y_2^i, X_1^i, X_2^i)$ ,  $i = 1, \dots, n$ . In addition, assume that for  $j = 1, 2$ ,  $\varepsilon_j^i$  is independent of  $(X_1^i, X_2^i, \alpha^i)$  and logistically distributed. Define  $N^i = Y_1^i + Y_2^i$ , so that  $N^i \in \{0, 1, 2\}$ . Consider the conditional distribution of  $Y_1^i$  given  $N^i = 1$ ,

$$p^i \equiv \mathbb{P}(Y_1^i = 1 | X_1^i, X_2^i, \alpha^i, N^i = 1) = \frac{\exp \left( (X_1^i - X_2^i)' \beta \right)}{1 + \exp \left( (X_1^i - X_2^i)' \beta \right)}. \quad (2)$$

This conditional probability is free of  $\alpha^i$  and hence allows consistent estimation of the parameter of interest  $\beta$ .

Obviously, there are two ways in which this model is restrictive. The first is the already mentioned assumption that  $\varepsilon_j^i$  is identically logistically distributed and independent of  $(X_1^i, X_2^i, \alpha^i)$ . The second restrictive assumption is that in model (1) all variables enter through a single parametric index. Since generally no a priori knowledge on the true relationship between the latent variable and the regressors is available, misspecification in this part of the model may have severe consequences.

This paper aims at relaxing the single index assumption by replacing  $X_j^{i'} \beta$  in model (1) by  $\eta(X_j^i)$ , where the function  $\eta$  is smooth but otherwise unrestricted. However, we still retain the spirit of Chamberlain's (1984) conditional logit idea. More formally, the one dimensional

responses  $Y_j^i$  obey the following conditional law:

$$Y_j^i | X_1^i, X_2^i, \alpha^i \stackrel{\text{indep.}}{\sim} \text{Bernoulli}(p(X_j^i, \alpha^i)), \quad i = 1, \dots, n, \quad j = 1, 2, \quad (3)$$

where  $\alpha^i$  are unobserved individual effects and  $p(X_j^i, \alpha^i) = \exp(\eta(X_j^i) + \alpha^i) / \{1 + \exp(\eta(X_j^i) + \alpha^i)\}$ . Since we consider estimation in the subset for which  $N^i = 1$  only, we put  $n$  equal to the numbers of indices  $i$  with  $N^i = 1$  and, in abuse of notation, we write  $(Y_1^i, Y_2^i, X_1^i, X_2^i)$ ,  $i = 1, \dots, n$  for this subsample. We define  $Y^i = \mathbb{I}(Y_1^i = 1)$  and assume in the following that all statements are made conditionally on the event that  $N^i = 1$ ,  $i = 1, \dots, n$ . Using these notations, we get the following regression model

$$\mathbb{E}(Y^i | X_1^i, X_2^i, N^i = 1) = \text{logit}(\eta(X_1^i) - \eta(X_2^i)) \quad (4)$$

where  $\text{logit}(u) = \exp(u) / (1 + \exp(u))$ ,  $(X_1^i, X_2^i, Y^i)$  are i.i.d. data, and  $\eta$  is an unknown smooth function. Note that the unobserved terms  $\alpha^i$  do not appear in model (4).

While our main motivation comes from the binary choice model, we consider a nonparametric procedure for a more general class of models for which

$$\mathbb{E}(Y^i | X_1^i, X_2^i) = G(\eta(X_1^i) - \eta(X_2^i)) \quad (5)$$

where  $G$  is a known link function. Model (5) belongs to the class of generalized additive models (GAMs), which is characterized by the form of  $\mathbb{E}(Y^i | X_1^i, X_2^i) = G(\eta_1(X_1^i) + \eta_2(X_2^i))$ . This class of models dates back to seminal work by Hastie and Tibshirani (1990). Compared to this line of work, the speciality of Model (5) is that the component functions  $\eta_1$  and  $\eta_2$  in the additive regression model are identical (up to sign). To the best of our knowledge, this paper is the first to propose an estimator that directly utilizes this restriction.

As a positive byproduct of our general framework we can also handle the linear nonparametric model with fixed effects,  $Y_j^i = \eta(X_j^i) + \alpha_i + \varepsilon_j^i$  with errors  $\varepsilon_j^i$  of which the conditional mean  $E(\varepsilon_j^i | X_1^i, X_2^i) = 0$ . If we put  $Y^i = Y_1^i - Y_2^i$ , we obtain a regression model

$$\mathbb{E}(Y^i | X_1^i, X_2^i) = \eta(X_1^i) - \eta(X_2^i). \quad (6)$$

While there is no directly comparable work in the binary choice case, Model (6) has been considered in Porter (1996), who proposed marginal integration type estimators. This type

estimator has some drawbacks that it uses a pilot estimator which often has the instabilities arising from fitting a more complex model and the related unnecessary higher order smoothness assumptions. However, our one-step projection type procedure avoids these problems by projecting the data directly onto the space of functions. Hence, our approach is advantageous because it not just allows to consider more general panel data models than have been considered previously in the literature - it also has important advantages in setups where models have already been proposed.

As already mentioned, our approach generalizes the one of Chamberlain (1984) to the nonparametric setting. In local likelihood nonparametric regression related work can be found in Fan, Gijbels and King (1997) and Chen and Zhou (2007). These two papers consider the nonparametric estimation of the risk function in Cox proportional hazard model. They utilize the exponential structure to eliminate the baseline hazard function. In the case of a  $d$ -dimensional covariate, their setup leads to a  $d$ -dimensional nonparametric model in the induced partial likelihood whereas our problem involves a  $2d$ -dimensional nonparametric additive regression model in the reduced form of the panel data. Other related work includes Yu, Mammen and Park (2008) who study a smooth backfitting estimator in unrestricted generalized additive models. In contrast to this work, our model imposes the equality between the two component functions arising from the first differencing step, and results hence in a quite different analysis.

Nonlinear panel data models with parametric coefficients have frequently been analyzed. For an overview of work related to the discrete choice models, see Arellano (2003). Closely related is the work of Manski (1987), who considers semiparametric estimation of a binary choice panel data model. Chamberlain (1992) discusses the identification of the panel data binary choice model, and why the logistic distribution assumption is required for identification of  $\beta$ , unless one is willing to assume unbounded support for one of the regressors (as is the case in Manski (1987)). For other nonlinear fixed effects model see also Hausman, Hall and Grilliches (1984) for panel count data, Honore (1992) for panel censored regression, and Kyriazidou (1997) for a panel sample selection model.

Like all of this work, our approach assumes a fixed number of time periods  $T$ . Recently, reduction of the first order bias has been achieved by letting  $T$  grow as well, see Arellano and Hahn (2007), and Hahn and Newey (2004). An interesting alternative way to treat the bias are bounds as in Honore and Tamer (2004). For a general overview, see Chamberlain (1984) and Arellano and Honore (2001).

In the following section we introduce our estimator for model (5) and establish the asymptotic behavior of our estimator. In the third and fourth section, the focus is on the finite sample behavior. We first provide a simulation study that describes the behavior of our estimator in a realistic finite sample setup. Then we show how our estimator performs in a typical application. As an illustrating example we take an example from a retirement study. Finally, an outlook concludes the paper.

## 2 Description and Asymptotic Properties of the Estimator

We suppose that we observe a random sample  $(Y^1, X_1^1, X_2^2), \dots, (Y^n, X_1^n, X_2^n)$  from model (5) where  $g$  is not necessarily the logit function. However, we assume that the distribution of the response  $Y$  (given  $(X_1, X_2)$ ) belongs to an exponential family with conditional log density

$$\log f_{Y|(X_1, X_2)}(y|x_1, x_2) = \{y[\eta(x_1) - \eta(x_2)] - b[\eta(x_1) - \eta(x_2)]\} + c(y).$$

Note that since the conditional variance of  $Y$  given  $(X_1, X_2) = (x_1, x_2)$  is equal to  $b''(\eta(x_1) - \eta(x_2))$ , we may naturally assume that  $b''(\cdot)$  is positive and hence  $b$  is strictly convex. Furthermore, the function  $b$  is infinitely often differentiable in each compact subset of the interior of the natural parameter space. We assume that the link function  $G$  is the canonical link, i.e.,  $G = b'$ . All these assumptions include our motivating example of matched binary response data with logit link. Our theory can be extended to a more general regression problems of the form (5) where the link function is not the canonical link and even where the conditional

density does not belong to an exponential family. Note, however, that in the general case this may not be the reduced form of a Panel data model any longer.

We make use of the following version of a smoothed log-likelihood

$$SL(\eta) \equiv \int \frac{1}{n} \sum_{i=1}^n \{Y^i(\eta(x_1) - \eta(x_2)) - b(\eta(x_1) - \eta(x_2))\} K_{h_1}(x_1, X_1^i) K_{h_2}(x_2, X_2^i) dx_1 dx_2.$$

Here  $K_{h_1}(x_1, u_1)$  and  $K_{h_2}(x_2, u_2)$  are product kernel weights  $K_{h_{1,1}}(x_{1,1}, u_{1,1}) \times \cdots \times K_{h_{1,d}}(x_{1,d}, u_{1,d})$  and  $K_{h_{2,1}}(x_{2,1}, u_{2,1}) \times \cdots \times K_{h_{2,d}}(x_{2,d}, u_{2,d})$  with bandwidth vectors  $h_1 = (h_{1,1}, \dots, h_{1,d})$  and  $h_2 = (h_{2,1}, \dots, h_{2,d})$ . For  $x_{j,k}$  in the interior of the support of  $X_{j,k}$  the kernel weights are equal to  $K_{h_{j,k}}(x_{j,k} - u_{j,k})$  with  $K_h(v) = h^{-1}K(h^{-1}v)$  for a kernel function  $K$  and a bandwidth  $h$ . At the boundary we will use boundary corrected kernels. For more details, see Assumption (A.3), below. We define the maximum smoothed likelihood estimator as the maximizer of  $SL(\eta)$ . Note that the differential of  $SL(\cdot)$  at  $\eta$  is represented by the following linear operator:

$$dSL(\eta)g = \int \frac{1}{n} \sum_{i=1}^n \{Y^i - b'(\eta(x_1) - \eta(x_2))\} \{g(x_1) - g(x_2)\} K_{h_1}(x_1, X_1^i) K_{h_2}(x_2, X_2^i) dx_1 dx_2.$$

Thus the maximizer  $\hat{\eta}$  satisfies that, for any  $g$ ,

$$\begin{aligned} & \int \left[ \hat{m}_1(u) \hat{p}_1(u) - \int b'(\hat{\eta}(u) - \hat{\eta}(v)) \hat{p}(u, v) dv \right] g(u) du \\ & - \int \left[ \hat{m}_2(u) \hat{p}_2(u) - \int b'(\hat{\eta}(v) - \hat{\eta}(u)) \hat{p}(v, u) dv \right] g(u) du = 0. \end{aligned} \quad (7)$$

Here,  $\hat{p}_j$ ,  $\hat{p}$  and  $\hat{m}_j$  denote marginal kernel density estimators, joint kernel density estimator and marginal Nadaraya-Watson estimators, respectively:  $\hat{p}_j(u) = n^{-1} \sum_{i=1}^n K_{h_j}(u, X_j^i)$ ,  $\hat{p}(u_1, u_2) = n^{-1} \sum_{i=1}^n K_{h_1}(u_1, X_1^i) K_{h_2}(u_2, X_2^i)$ ,  $\hat{m}_j(u) = \hat{p}_j(u)^{-1} n^{-1} \sum_{i=1}^n K_{h_j}(u, X_j^i) Y^i$ .

The maximizer in (7) is given by the solution of the following nonlinear integral equation:

$$\hat{m}_1(u) \hat{p}_1(u) - \hat{m}_2(u) \hat{p}_2(u) - \int \{b'(\eta(u) - \eta(v)) \hat{p}(u, v) - b'(\eta(v) - \eta(u)) \hat{p}(v, u)\} dv = 0. \quad (8)$$

It is clear that if a function  $\eta$  is a solution of the equation (8) then also  $\eta + C$  solves (8) for any constant  $C$ . For technical simplicity we use the following normalization condition for the identification:

$$\int (\eta(u) + \eta(v)) b''(\eta(v) - \eta(u)) \hat{p}(v, u) dudv = 0. \quad (9)$$

The discussion simplifies when  $b'$  is the identity function that is the case of Model (6). Then, the equation (8) can be simplified to

$$\eta(u) = \widehat{f}(u) + \int \eta(v) \widehat{\mathcal{K}}(u, v) dv \quad (10)$$

where  $\widehat{f}(u) = (\widehat{p}_1(u) + \widehat{p}_2(u))^{-1}(\widehat{m}_1(u)\widehat{p}_1(u) - \widehat{m}_2(u)\widehat{p}_2(u))$  and  $\widehat{\mathcal{K}}(u, v) = (\widehat{p}_1(u) + \widehat{p}_2(u))^{-1}(\widehat{p}(u, v) + \widehat{p}(v, u))$ . This integral equation is a Fredholm integral equation of the second type. Under assumptions we make below, the integral operator in (10) is a strict contraction and thus the integral equations can be solved by an iterative algorithm which converges geometrically fast. In the iteration the  $(k - 1)$ th fit  $\widehat{\eta}^{[k-1]}$  is updated by

$$\widehat{\eta}^{[k]}(u) = \widehat{f}(u) + \int \widehat{\eta}^{[k-1]}(v) \widehat{\mathcal{K}}(u, v) dv. \quad (11)$$

The algorithm could start with the initial value  $\widehat{\eta}^{[0]}(\cdot) = \widehat{f}(\cdot)$ .

In general, the estimator is defined by the nonlinear integral equation (8). This equation can be solved by using Newton type algorithms. For this purpose one considers a linear approximation. With an increment  $\xi$  we have the following linearized equation for  $\xi$  at  $\eta^0$  :

$$\xi(u) = \widehat{f}(u; \eta^0) + \int \xi(v) \widehat{\mathcal{K}}(u, v; \eta^0) dv \quad (12)$$

where

$$\widehat{f}(u; \eta^0) = \frac{\widehat{m}_1(u)\widehat{p}_1(u) - \widehat{m}_2(u)\widehat{p}_2(u) - \int \{b'(\eta^0(u) - \eta^0(v))\widehat{p}(u, v) - b'(\eta^0(v) - \eta^0(u))\widehat{p}(v, u)\} dv}{\widehat{w}_1(u; \eta^0) + \widehat{w}_2(u; \eta^0)}$$

and

$$\widehat{\mathcal{K}}(u, v; \eta^0) = \widehat{a}(u; \eta^0) \frac{\widehat{w}(u, v; \eta^0)}{\widehat{w}_1(u; \eta^0)} + (1 - \widehat{a}(u; \eta^0)) \frac{\widehat{w}(v, u; \eta^0)}{\widehat{w}_2(u; \eta^0)}.$$

Here,  $\widehat{w}(u, v; \eta^0) = b''(\eta^0(u) - \eta^0(v))\widehat{p}(u, v)$ ,  $\widehat{w}_1(u; \eta^0) = \int \widehat{w}(u, v; \eta^0) dv$ ,  $\widehat{w}_2(u; \eta^0) = \int \widehat{w}(v, u; \eta^0) dv$  and  $\widehat{a}(u; \eta^0) = \widehat{w}_1(u; \eta^0)[\widehat{w}_1(u; \eta^0) + \widehat{w}_2(u; \eta^0)]^{-1}$ . Note that equation (12) has the same form as (10). For the solution of the integral equation (8) this suggests the following iterative algorithm that uses two nested loops. The outer loop uses a Newton type iteration. The inner loop proceeds similarly as (11). We now give a more detailed description of this algorithm.

*Step 1.* Choose an initial value  $\widehat{\eta}^{[0]}$  and put  $k = 1$ .

*Step 2.* Calculate  $f(u; \hat{\eta}^{[k-1]})$  and  $\hat{w}(u, v; \hat{\eta}^{[k-1]})$ .

*Step 3.* (inner loop) Set  $j = 1$  and  $\xi^{[0,k]} = f(u; \hat{\eta}^{[k-1]})$  and repeat (13) for  $j = 1, 2, \dots$

$$\xi^{[j,k]}(u) = f(u; \hat{\eta}^{[k-1]}) + \int \xi^{[j-1,k]}(v) \mathcal{K}(u, v; \hat{\eta}^{[k-1]}) dv \quad (13)$$

until a convergence criterion is satisfied (or for a fixed number of iterations). The result of the iterations is denoted by  $\xi^{[k]}$ .

*Step 4.* Update  $\hat{\eta}^{[k]}$  by  $\hat{\eta}^{[k]} = \hat{\eta}^{[k-1]} + \xi^{[k]}$  and repeat from Step 2 with  $k$  replaced by  $k + 1$ . This is done until a convergence criterion is fulfilled (or for a fixed number of iterations).

To analyze this estimator, we make the following assumptions:

- (A1) We suppose that  $X_1$  and  $X_2$  have compact supports, saying without loss of generality,  $[0, 1]^d$ . The density  $p$  of  $(X_1, X_2)$  is bounded from zero and from infinity on  $[0, 1]^{2d}$  and is continuously differentiable.
- (A2) The function  $\eta$  is twice continuously differentiable and the interval  $I = [\min\{\eta(x_1) - \eta(x_2) : x_1, x_2 \in [0, 1]^d\}, \max\{\eta(x_1) - \eta(x_2) : x_1, x_2 \in [0, 1]^d\}]$  lies in the interior of the natural parameter space of the exponential family of  $Y$ , i.e.  $\int \exp[\theta y + c(y)] dy < \infty$  for  $d(\theta, I) < \varepsilon$  for  $\varepsilon$  small enough. The exponential family is non degenerate, i.e.  $b''(\theta) \neq 0$  for  $\theta \in I$ .
- (A3) The kernel weights  $K_g(v, w)$  are bounded by  $Cg^{-1}$  for a constant  $C$  and vanish for  $|v - w| > g$ . It holds that  $\int K_g(v, w) dv = 1$  for  $0 \leq w \leq 1$ . For  $g < v < 1 - g$  the weight is equal to  $K_g(v, w) = K_g(v - w)$  with  $K_g(t) = g^{-1}K(g^{-1}t)$  for a kernel function  $K$  with support  $[-1, 1]$ . The kernel  $K$  is a Lipschitz continuous probability density function.
- (A4)  $n^{1/(4+d)}h_{j,k}$  converges to constants  $\delta_{j,k} > 0$  for  $j = 1, 2$  and  $k = 1, \dots, d$ .

These assumptions are standard in the nonparametric literature. The first two are common assumptions about design density and smoothness of the function and the latter two

are also standard assumptions on the kernel function and bandwidth sequences. Related to Assumption (A1), for a series estimator usually an additional assumption is supposed on the minimum singular value of the design matrix obtained from the basis functions. This assumption can be easily violated and a series estimator becomes unstable when the covariates are strongly correlated. Our procedure based on smoothed likelihood avoids this, which is a key advantage.

We now need some notation for the statement of our first theorem. Put

$$\mathcal{K}(u, v) = a(u) \frac{w(u, v)}{w_1(u)} + (1 - a(u)) \frac{w(v, u)}{w_2(u)}.$$

Here,  $w(u, v) = b''(\eta(u) - \eta(v))p(u, v)$ ,  $w_1(u) = \int w(u, v)dv$ ,  $w_2(u) = \int w(v, u)dv$  and  $a(u) = w_1(u)[w_1(u) + w_2(u)]^{-1}$ . Furthermore,  $\delta_{1,prod} = \delta_{1,1} \cdot \dots \cdot \delta_{1,d}$  and  $\delta_{2,prod} = \delta_{2,1} \cdot \dots \cdot \delta_{2,d}$ . We also write  $K$  for the linear integral operator that maps a function  $f$  to the function  $Kf(u) = \int K(u, v)f(v)dv$ .

**THEOREM 1** (*Asymptotic normality*) Under (A1)–(A4) the following convergence holds for  $x \in (0, 1)^d$ :

$$n^{2/(4+d)} (\widehat{\eta}(x) - \eta(x)) \xrightarrow{d} N(\beta(x), v(x)),$$

where

$$\begin{aligned} v(x) &= \frac{w_1(x)/\delta_{1,prod} + w_2(x)/\delta_{2,prod}}{(w_1(x) + w_2(x))^2} \int K^2(w)dw \\ \beta(x) &= (I - \mathcal{K})^{-1} \left\{ \frac{1}{w_1 + w_2} \left[ \lim_{n \rightarrow \infty} n^{2/(4+d)} B_{1,n} - \lim_{n \rightarrow \infty} n^{2/(4+d)} B_{2,n} \right] \right\} (x) \\ B_{1,n}(x) &= \int K_{h_1}(x, x_1) \left\{ b'[\eta(x_1) - \eta(x_2)] - \left( \int b'[\eta(x) - \eta(v)]K_{h_2}(v, x_2)dv \right) p(x_1, x_2) \right\} dx_1 dx_2 \\ B_{2,n}(x) &= \int K_{h_2}(x, x_2) \left\{ b'[\eta(x_1) - \eta(x_2)] - \left( \int b'[\eta(v) - \eta(x)]K_{h_1}(v, x_1)dv \right) p(x_1, x_2) \right\} dx_1 dx_2 \end{aligned}$$

It can be easily checked that  $n^{2/(4+d)}B_{1,n}(x)$  and  $n^{2/(4+d)}B_{2,n}(x)$  have a finite limit. An explicit formula of these expressions is complicated by the calculation of the integrals at boundary regions. A consistent estimate of the bias terms is given if an estimator of  $\eta$  is plugged in with a uniformly consistent first and second derivative.

In our binary choice fixed effects model, Theorem 1 states that

$$\mathcal{L} \left( m^{2/(4+d)} (\widehat{\eta}(x) - \eta(x)) \mid N = m \right) \rightarrow N(\beta(x), v(x)) \text{ as } m \rightarrow \infty.$$

Here,  $N = \sum_{i=1}^n N^i$  and  $\mathcal{L}(Z_n \mid N = m)$  denotes the conditional law of  $Z_n$  given  $N = m$ . This implies that

$$N^{2/(4+d)} (\widehat{\eta}(x) - \eta(x)) \xrightarrow{d} N(\beta(x), v(x)) \text{ as } n \rightarrow \infty.$$

The next theorem states conditions under which our algorithm converges.

**THEOREM 2** (*convergence of algorithm*) *Under conditions (A1)–(A4) there exists a positive number  $r$  such that if  $\widehat{\eta}^{[0]} \in B(\widehat{\eta}, r)$  (with probability tending to one) then the outer loop converges with geometric rate (with probability tending to one), i.e. there exist  $C > 0$  and  $0 < \gamma < 1$  such that*

$$\int [\widehat{\eta}^{[k]}(u) - \widehat{\eta}(u)]^2 [\widehat{w}_1(u, \eta) + \widehat{w}_2(u, \eta)] du \leq C\gamma^k$$

*with probability tending to one.*

As our estimator is based on the smoothed likelihood - which can be regarded as an empirical version of Kullback-Liebler divergence between the model density and the true density - efficiency properties of our estimator are comparable to the maximum likelihood estimator in parametric models. Although we have a  $2d$ -dimensional model, our estimator moreover achieves the optimal rate of the  $d$ -dimensional nonparametric regression under the given smoothness assumption. Thus, our estimator successfully makes use of the additive structure of the model to achieve the best convergence rate for estimating  $d$ -dimensional function  $\eta$ . By econometric theory criteria, this is a favorable estimator. In simulations and the application below we show that these properties translate into equally favorable small sample behavior.

As mentioned in the introduction, there is no directly comparable estimator. However, since model (5) is a subclass of the class of GAMs,  $E(Y^i \mid X_1^i, X_2^i) = G(\eta_1(X_1^i) + \eta_2(X_2^i))$ ,

one could, in principle, use estimators based on unrestricted GAM<sup>1</sup>. A possible estimation procedure is the following: Let  $\tilde{\eta}_1$  and  $\tilde{\eta}_2$  be unrestricted estimators of  $\eta_1$  and  $\eta_2$ . Then, take an (weighted) average of the two estimators,  $\theta(x)\tilde{\eta}_1(x) + (1 - \theta(x))(-\tilde{\eta}_2(x))$  with a weight  $\theta(x) \in [0, 1]$  to obtain an estimator of  $\eta$ . Obviously, the properties of this two step estimator depends on the properties of the first step estimators  $\tilde{\eta}_1$  and  $\tilde{\eta}_2$  and the weight  $\theta$ . An optimal choice of  $\theta(x)$  depends on several unknown quantities including the marginal and joint design densities, and is very difficult. Moreover, the weighting also affects the choice of bandwidths in the first step estimation of  $\tilde{\eta}_1$  and  $\tilde{\eta}_2$ . Finally, the optimization can also lead to a discontinuous weight function and thus to a non-smooth estimator. This happens in particular when the two covariates have different support or when either of the covariates  $X_1$  or  $X_2$  has very low density in a certain area (in that case, the first step estimators  $\tilde{\eta}_1$  and  $\tilde{\eta}_2$  could also be unstable because the asymptotic variances of these estimators are reciprocal to the marginal densities of the covariates). All these disadvantages make estimation based on first stage unrestricted GAMs rather unattractive, and underscore the need for our different one-step estimation strategy. These are not just theoretical issues: In the simulation study below we find that our estimator outperforms the two stage estimator in small samples.

Our approach is based on local constant (Nadaraya-Watson) smoothing. One could also use local linear or local polynomial approximations in the construction of the local model. In that case we will have simpler bias expressions. Our approach can also be generalized to more structured models of  $\eta$ , e.g. additive specifications  $\eta(x) = \eta_1(x_1) + \dots + \eta_d(x_d)$  or semiparametric models like the partial linear model  $\eta(x, z) = \eta(x) + z^T\beta$ . Finally, our procedure can be directly applied to the case of more than two periods by using it for pairs of time points.

---

<sup>1</sup>For instance, one can use the estimator proposed by Yu, Mammen and Park (2008) which is oracle efficient.

### 3 Simulation

We conducted some numerical experiments to assess the finite sample performance of the estimators. These numerical experiments were done by R on windows. The simulation was done under the following model for the conditional distribution: for  $j = 1, 2$  and  $i = 1, \dots, n$ ,

$$Y_j^i | X_j^i, \alpha^i \sim \text{Bernoulli}(m(X_j^i, \alpha^i))$$

where  $\text{logit}(m(X_j^i, \alpha^i)) = \eta(X_j^i) + \alpha^i$ . Here,  $\eta(x) = \sin(\pi x)$  and  $\alpha^i$  are unobserved individual effects i.i.d.  $N(0, 1/9)$ . We considered two models for the covariate vector  $(X_1, X_2)$ . One case (Case I) is the identically distributed case that has uniform distribution on  $[-1, 1]$ . The other case (Case II) is that the covariate  $X_1$  and  $X_2$  are distributed on  $[-1, 1]$  with the marginal densities  $p_1(x) = -0.45x + 0.5$  and  $p_2(x) = 0.45x + 0.5$ , respectively. In both cases, we generated bivariate normal random numbers with zero means, unit variances and 0.9 correlation and then took a transformation with standard normal distribution function to obtain uniform random numbers on  $[0, 1]$  and transformed these random number according to each design. The actual correlations in Case I and II are about 0.89 and 0.84, respectively. Note that in Case II the covariate  $X_1$  has low density around the right end of the support while the covariate  $X_2$  has low density around the left end of the support. We generated 500 pseudo samples of size  $n = 500$  from each model. All the integrals involved in the procedures were calculated by a trapezoidal rule based on 201 equally spaced grid points on  $[-1, 1]$  for each direction. The average numbers of observations of which  $N^i = 1$  are about 225 in Case I and 248 in Case II. We tried 36 different combination of bandwidths  $(h_1, h_2)$ . Each of them is from  $\{0.2, 0.25, \dots, 0.45\}$ .

In Table 1, we report the performance with small bandwidth  $(0.2, 0.2)$  and the best performances among the choices of bandwidths in terms of mean integrated squared error (MISE). The first method (HMY) is our method, the second one (YPM I) is a simple average,  $\theta(x) = 0.5$ , of two step estimators based on Yu, Mammen and Park (2008), and the third one (YPM II) is the two step estimators with  $\theta(x) = \delta_1 w_1(x)(\delta_1 w_1(x) + \delta_2 w_2(x))^{-1}$ . The third one gives the best asymptotic variance among two step estimators for the given bandwidths.

In Case I, three estimators show similar performances but HMY is slightly better than the others with bigger difference in small bandwidth case. In Case II, HMY clearly outperforms the others. Especially, the difference in small bandwidth case is quite interesting. We consider small bandwidth because these estimators can be used for estimating average effects  $\int \eta(x)W(x)dx$  or in semiparametric estimation. In these cases, undersmoothing might be needed to avoid large bias terms. The result shows that our estimator is more robust for small bandwidths. Although we do not report here, YPM I showed the worst performances and HMY produced the best results for other choices of bandwidths.

Figure 1 illustrates the result of HMY under Case I with bandwidths (0.3,0.3). The left-upper panel shows the pointwise confidence band. The other plots shows the quantile plot of 500 estimates. These plots show that the asymptotic distribution approximate the distribution of the estimator quite well. As usual in nonparametric regression, we see that there are downward biases and upward biases at the hills and valleys, respectively.

## 4 Application: The Effect of Life Cycle Income on Retirement

In our application we look at the decision to retire. In particular, we want to focus on the effect of life cycle income on the decision to retire. We start with a brief discussion of the literature, introduce the data, state our formal model, and discuss the empirical results.

**Relationship to the Literature:** The decision to retire, and the associated institutions have always attracted a great deal of interest in applied economics. In an influential paper, Rust and Phelan (1997) considered the decision to retire using a discrete dynamic programming approach. Concentrating on a subset of the population that has relatively low wage in the RHS (Retirement History Survey), they are able to explain the retirement decision in this subpopulation, especially the pronounced peaks at retirement age 62 and 65. Their explanation of retirement behavior rests largely on constraint. Other studies that use the same data obtain in parts different findings: Gustman and Steinmeier (1986) explain the

		Case I			Case II	
	HMY	YPM I	YPM II	HMY	YPM I	YPM II
<i>Best Performances</i>						
MISE	0.208	0.212	0.212	0.083	0.188	0.118
IV	0.111	0.117	0.116	0.059	0.113	0.077
Bandwidths	(0.35,0.35)	(0.35,0.35)	(0.35,0.35)	(0.4,0.4)	(0.4,0.4)	(0.35,0.4)
<i>Small Bandwidth Performances (0.2, 0.2)</i>						
MISE	0.282	0.301	0.298	0.130	0.539	0.222
IV	0.259	0.280	0.275	0.124	0.482	0.211

Table 1: Finite sample performances-mean integrated squared error (MISE) and integrated variance (IV) of the estimators. HMY denotes the proposed estimator and YPM I and II denote two step estimators based on GAM. Case I has identical distribution of the two covariates and Case II has different areas of low densities for two covariates. The bandwidths are given in the parentheses for each case.

retirement behavior by a dynamic structural model without constraints, while Lumsdaine, Stock and Wise (1995) attribute at least the peaks in retirement behavior to social customs.

In contrast to this literature we aim at a much less structural analysis. Our focus is on the ceteris paribus effect of higher life cycle income on the decision to retire. Our analysis is more in line with traditional smooth Euler equation labor supply models (like MaCurdy (1983)), though we acknowledge the problem of corner solutions to the individuals optimization problem. However, compared to the previously mentioned approaches we allow for more smoothness, which in parts is driven by the data. Unlike the data sets used in this literature (as in Rust and Phelan (1998)), from our HRS data set we obtain a less Social Security heavy

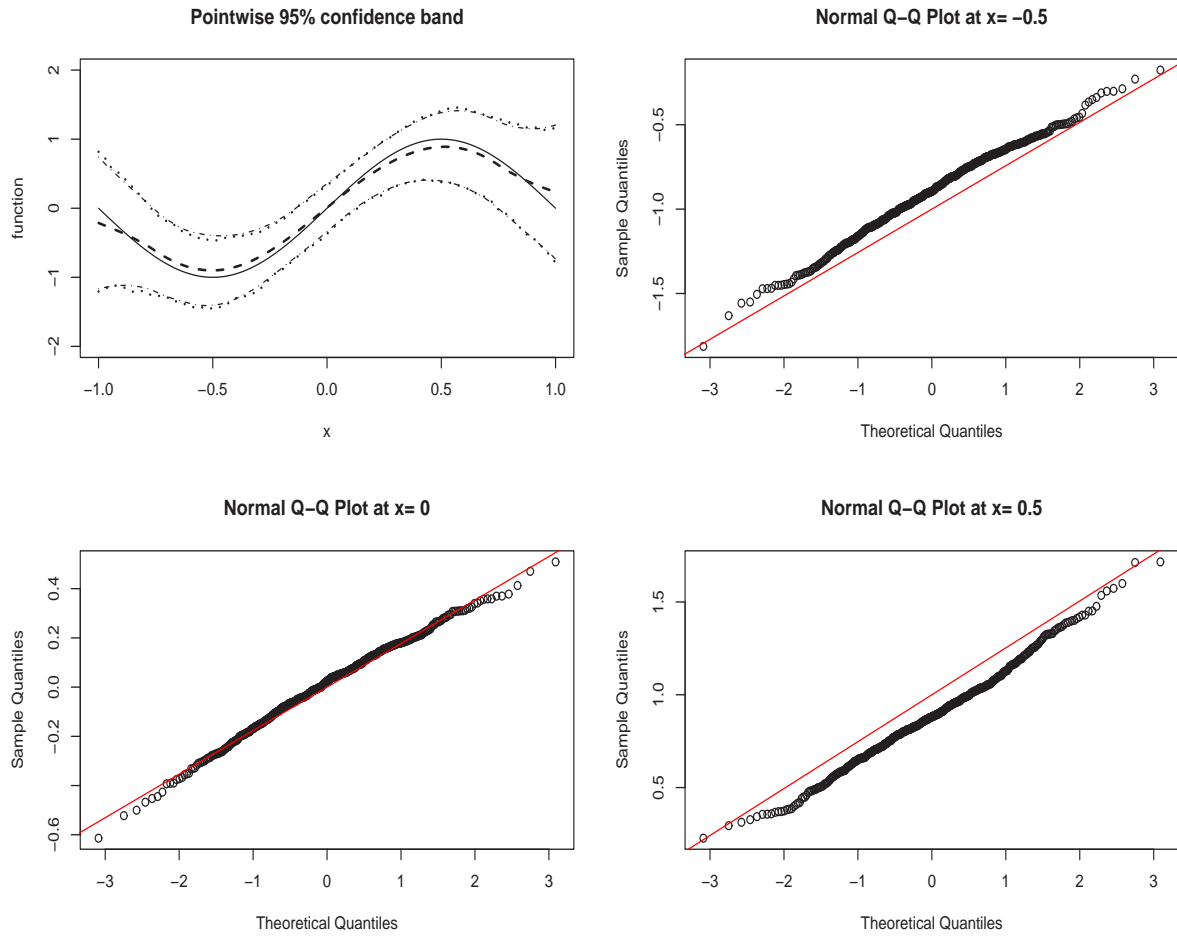


Figure 1: In the left upper figure, the real line in the center represents the true function and the dashed line represents the average of the 500 estimates. The outer lines represents pointwise 95% confidence band. Dot-dashed lines are calculated according to the asymptotic theory and dotted lines are sample quantiles from 500 estimates. The other figures shows the Q-Q plots of 500 estimates at  $x = -0.5, 0$  and  $0.5$ . The reference lines represent theoretical asymptotic distributions without biases.

sample of the population. Indeed, Social Security pensions only account for 50-60% of the population, and individuals with private pension plans (or both) are an important factor, a group which is much less restricted by institutional settings, or at least not affected equally.

**Data Description:** The data come from a supplemental survey to the Health and

Retirement Study (HRS), the Consumption and Activities Mail Survey (CAMS). We use data from the 2000, 2002, and 2004 waves of the HRS main survey and from the 2001 and 2003 waves of CAMS. The HRS is a biennial panel of older Americans; the target population of the HRS was the cohorts born in 1931–1941 but additional cohorts were added later (see Juster and Suzman, 1995). In 2000 the HRS interviewed about 20,000 subjects in 13,100 households. While the HRS main surveys are conducted as computer-assisted telephone or personal interviews, the CAMS supplements to the HRS are self-administered mail questionnaires. For the first wave of CAMS, a random sample of 5,000 HRS households was contacted with a response rate of 77.3 percent, see Hurd and Rohwedder (2005) for more details.

The dependent variable in our analysis is whether one is retired or not. Consumption is a sum of nondurables, and taken from CAMS. All other covariates are mostly taken from the RAND version of the HRS 2000, 2002, and 2004 data, and matched to CAMS where necessary. Life cycle income is a key concept in our analysis. We look at two different measures of life cycle income, namely pension and total nondurable consumption. The first is defined as the sum of the payment from social security and private pension, the second one is a large sum over nondurable expenditures<sup>2</sup>.

Our estimator uses the observations of those individuals who switch from work to retirement. These are 285 observations, out of which 68 actually switch from retirement back to (part time) work, which is in line with previous observations in the literature.

**The Model and Estimation:** As discussed previously, our analysis is of fairly reduced form. We consider the probability to retire given covariates. Our analysis critically exploits the fixed effects assumption. In particular as most household observables are constant over time, as are many unobservables, we really only focus on the effect of two regressors at a time. One of them is a measure of life cycle income, the other reflects a more transitory

---

<sup>2</sup>Essentially, these are the answers to categories: B10-13, 15, 18, 26, 28-29, 31, 34, 97-99, and include typical nondurable categories like clothing, household goods, home repairs, food and beverages, mortgage etc.

component. According to standard labor supply theory, (permanent) income is a substitute for leisure. Hence the probability of being retired should be inversely related to the wage rate (or current income), and positively related to life cycle income (i.e. pension or consumption). The model on individual level is then given by:

$$Y_j^i = \mathbb{I} \{ \eta(X_{1j}^i, X_{2j}^i) + \alpha^i + \varepsilon_j^i > 0 \}, \quad (14)$$

where  $X_{1j}^i$  and  $X_{2j}^i$  are two scalar regressors, the first one denoting the permanent component, the second one the transitory. The disturbance  $\varepsilon_j^i$  are assumed to be iid logistic and independent with covariates. The influence of all time invariant regressors are summarized in  $\alpha^i$ . In this application, we do not consider other time varying covariates. Indeed, in our approach they could be introduced in a variety of forms, e.g., through a partially linear structure, but we leave such a model for future research.

To apply our method, we used rough choices of bandwidths. The bandwidths are proportional to  $n^{-1/6}$  where  $n$  is the sample size. To obtain an estimate for the derivatives we applied smoothed differentiation, see Mammen and Park (2005). This method is based on applying a local linear smoothing to the fitted values of the regression functions. In the illustration of the results we presented one-dimensional graphs of  $y = \eta(x, a)$  for several fixed values of  $a$ . We do not present the (pointwise) confidence intervals in the following figures but we can calculate them from our asymptotic results. In the formula of the asymptotic variance  $v(x)$ , only the functions  $\eta$  and  $p$  are unknown. We can replace  $\eta$  with  $\hat{\eta}$  and  $p$  with a kernel density estimator  $\hat{p}$  to estimate the asymptotic variance  $v(x)$ . Note that, in this model, one has  $b''(t) = \exp(t)/\{1 + \exp(t)\}^2$ .

**Empirical Results:** We start with the first scenario in which we use pension and labor income as regressors. In pension, there are roughly 30% of the population that have private pension, while it is roughly half the population that have both Social Security or pension. Of those who have private pensions, these account for more than one quarter of the pension on average, providing means to smooth consumption. Our results are summarized in the following Figure 2, showing the conditional probability of being retired, evaluated at zero fixed effect:

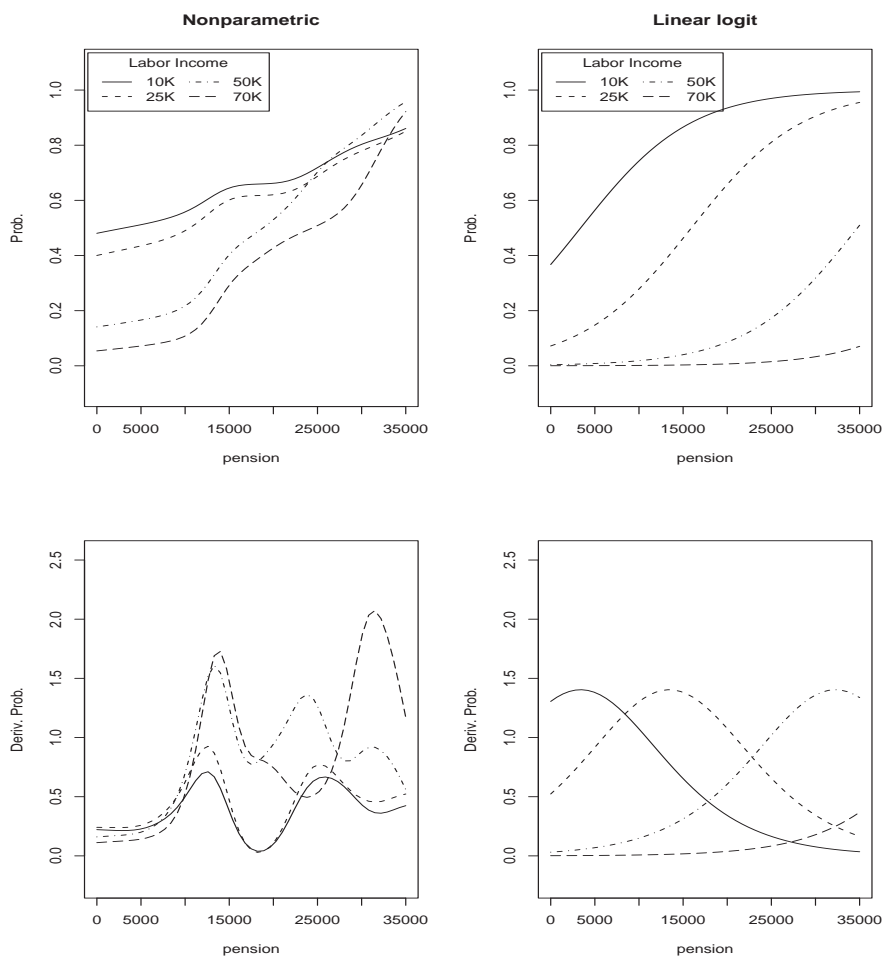


Figure 2: Implied conditional probabilities(levels and derivatives) to retire as a function of pension wealth at zero fixed effect, controlling for transitory income. Comparison Nonparametric vs parametric regression.

Since this application is meant to be an illustration of the superiority of our estimator over binary single index model fixed effects estimators, we plot the parametric conditional probabilities for comparison. Several things are noteworthy: 1. The nonparametric results are quite different, revealing a number of interesting details. 2. Still, both estimators produce the same essential result: First, the probability of retirement increases with increasing pension (life cycle income). Second, the probability decreases with increasing income (wage,

opportunity costs).

When it comes to the details, it is noteworthy that the nonparametric estimator of the probabilities reveals that they coincide at high pensions, while they show a remarkable gap at low income. As was to be expected, this means that transitory components have a higher impact at low LCI compared to high one, reflecting e.g., the relative magnitude of these shocks, or the presence of constraints, while for high LCI this effect is not of great importance. Observe how these plausible effects are obscured by the parametric fit.

The second scenario we consider is where we replace pensions as a measure of LCI with consumption, see Figure. 3. This graph reveals essentially the same picture: Similar sign of effect between parametric and nonparametric, but noticeable differences in the details. As before, the probability of retirement increases with increasing consumption (life cycle income), while the probability decreases with increasing income (wage, opportunity costs). However, this time there is no “convergence” of the nonparametric estimators at high probabilities - the functions remain apart for all of its range.

Finally, it is interesting to consider the effect of various definitions of LCI. Figure. 4 illustrates the effects of including consumption and pension as regressors. Following a simplistic understanding of theory, both definitions should produce the same result, i.e., after controlling for pensions, consumption should not have an effect on the conditional probabilities. This is what we get from the parametric model, too. Indeed, there is virtually no difference in the conditional probabilities for various values of the regressors. The nonparametric estimator, however, produces a much more differentiated picture. At low pensions, for most values of consumption there is not much of a difference. But at high values of pensions, the functions start to “fan out”, with the higher probabilities at higher values of consumption. This is suggestive of consumption being a better measure of LCI at the higher end of the pension scale: Increasing consumption may for instance reflect higher financial assets of the individuals. Consequently, for a given level of pensions, the LCI of such individuals may be higher. And, as theory predicts, a higher LCI yields in the data to a lower probability of being in the workforce, and hence a higher probability of being retired.

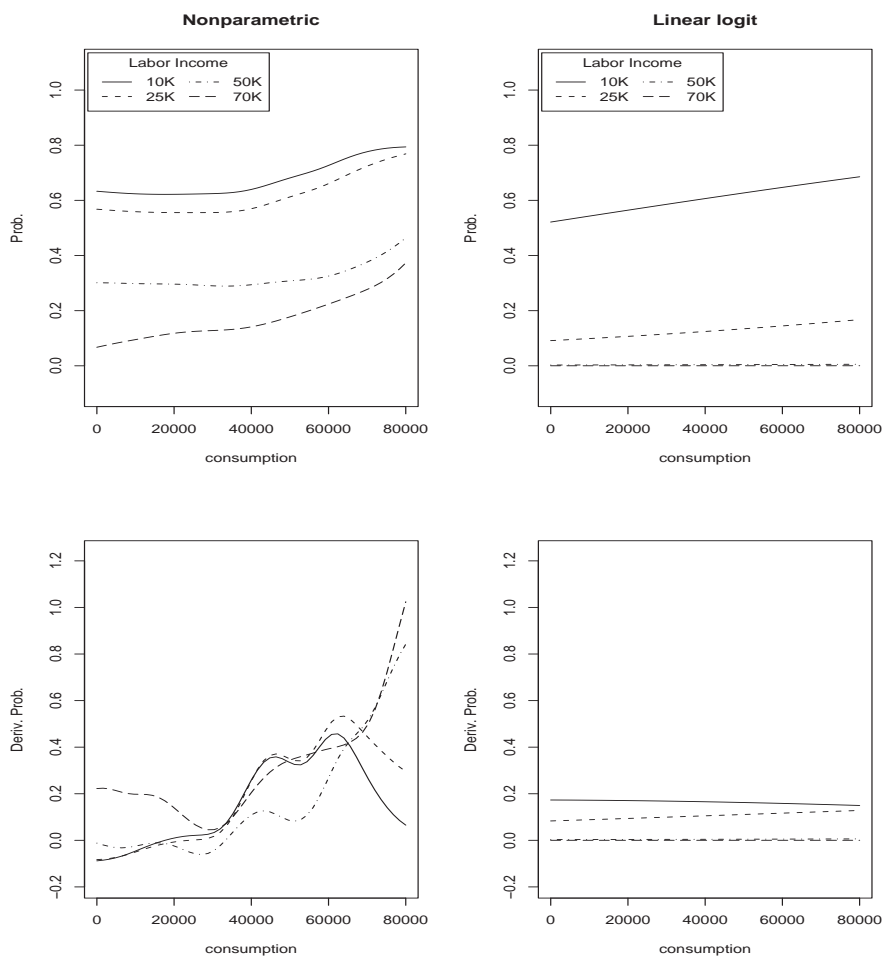


Figure 3: Implied conditional probabilities(levels and derivatives) to retire as a function of consumption at zero fixed effect, controlling for transitory income. Comparison Nonparametric vs parametric regression.

Summarizing, the fact that our nonparametric estimator disentangles the effects of wages on the lower end of the pension scale, and the more precise quantification of consumption as a measure for LCI at the upper end compared to the parametric estimator illustrates the advantages of our nonparametric procedure in a data set of the size that typically is encountered in practise.

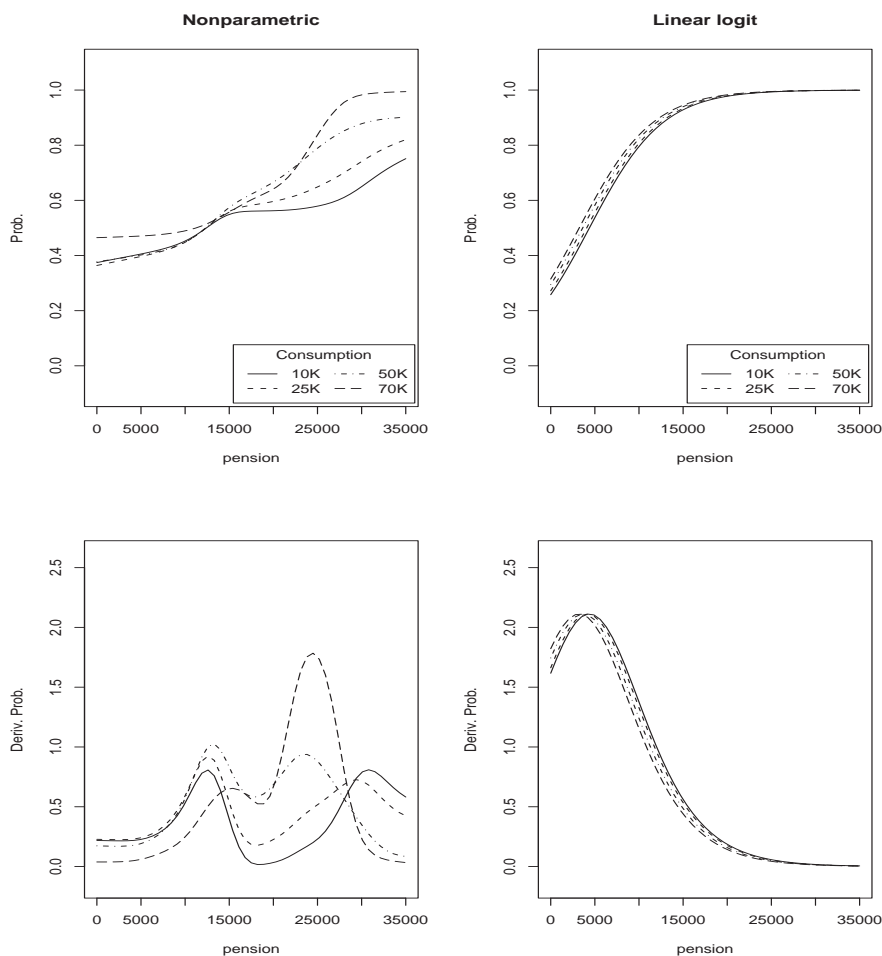


Figure 4: Implied conditional probabilities(levels and derivatives) to retire as a function of pension wealth at zero fixed effect, controlling for consumption. Comparison Nonparametric vs parametric regression.

## 5 Conclusion and Outlook

In this paper we are concerned with relaxing the linear single index assumption commonly encountered in binary choice panel data estimators with fixed effects. We propose a nonparametric estimator for the influence of the regressors of interest, and establish its asymptotic distribution. We show that the estimator works well in an application, where we were able to reveal interesting features that a simple parametric single index model missed.

The key maintained assumption is that the time varying errors are logistically distributed. As it is only possible to remove the fixed effects using observations that switch sign under this assumption, the impact of misspecification in this part is not clear. However, the nonparametric structural function certainly accommodates some of the misspecification. Hence, while acknowledging the limitations of assuming a certain functional form for the errors, our approach may be seen as a more robust way of quantifying the marginal effects of a change in the regressors. Practical implementations with a bandwidth selector and an efficient procedure for the case with more than two repeated observations is postponed to further research.

## Appendix

**Proof of Theorems 1 and 2.** For clarity in this section the true function  $\eta$  is denoted by  $\eta^*$  and the symbol  $\eta$  is used as wildcard character. Let  $(\widehat{F}(\eta))(u) = \widehat{m}_1(u)\widehat{p}_1(u) - \widehat{m}_2(u)\widehat{p}_2(u) - \int \{b'(\eta(u) - \eta(v))\widehat{p}(u, v) - b'(\eta(v) - \eta(u))\widehat{p}(v, u)\} dv$ . Then the integral equation (8) can be written as  $\widehat{F}(\eta) = 0$ . In our proof, we make use of the Newton-Kantorovich theorem which is used in Yu, Mammen and Park (2008) to linearize the nonlinear integral equation. The main step in the proof is to show that the operator used in the inner loop is a contraction, see Lemma 5. This argument makes the main difference to the approach in Yu, Mammen and Park (2008). The operator considered there does not have this property. Thus a more complicated iterative algorithm than ours is required and has to be analyzed.

The Newton-Kantorovich theorem makes the following statement: if a function  $\bar{\eta}$  fulfills

$$\|\widehat{F}(\bar{\eta})\| = O_p(\epsilon_n) \text{ ( or } o_p(\epsilon_n), \text{ respectively),} \quad (15)$$

and the conditions listed below then this implies that

$$\|\widehat{\eta} - \bar{\eta}\| = O_p(\epsilon_n) \text{ ( or } o_p(\epsilon_n), \text{ respectively).} \quad (16)$$

Moreover, the theorem implies that a Newton type algorithm for the nonlinear integral equation  $\widehat{F}(\boldsymbol{\eta}) = 0$  converges geometrically fast.

We will apply the Newton-Kantorovich theorem with the  $L_2(w^*)$  norm defined as  $\|g\|_2 = (\int g^2(u)w^*(u))^{1/2}$  and with the uniform norm  $\|g\|_\infty = \sup_{u \in I_n} |g(u)|$ . Here,  $w^*(u) = (w_1(u; \eta^*) + w_2(u; \eta^*))/2$  and  $I_n$  is the compact set  $\prod_{j=1}^d [2 \max\{h_{1j}, h_{2j}\}, 1 - 2 \max\{h_{1j}, h_{2j}\}]$ . Note that for the identification we restrict the mapping  $\widehat{F}$  to the functions that are orthogonal to the constants, i.e.,  $\widehat{F}$  is restricted on the set of functions  $\{\eta : \|\eta\| < \infty \text{ and } \int \eta(x)w^*(x)dx = 0\}$ .

We will apply the Newton-Kantorovich theorem with  $\eta = \eta^*$  and  $\eta = \check{\eta}$ , defined below. For the validity of the Newton-Kantorovich theorem one has to check the validity of (15) for these choices of  $\eta$ . This is done in Lemmas 1 and 2, below. Furthermore, one has to show that the Fréchet differential of  $\widehat{F}$ ,  $\widehat{F}'$ , has a bounded inverse and is Lipschitz continuous near  $\eta^*$  or  $\check{\eta}$  with stochastically bounded Lipschitz constant. This is shown in Lemma 4. The application of the Newton-Kantorovich theorem with  $\eta = \eta^*$  gives the statement of Theorem 2. The application with  $\eta = \check{\eta}$  shows that  $\widehat{\eta}$  has the same asymptotic limit as  $\check{\eta}$ . The asymptotic distribution of  $\check{\eta}$  is given in Lemma 2. This shows the statement of Theorem 1. For a statement of the Newton-Kantorovich theorem and its application in nonparametric statistics, see Mammen and Nielsen (2003).

We now come to the definition of  $\check{\eta}$ . This is defined as  $\check{\eta} = \eta^* + \check{\xi}$  where  $\check{\xi}$  is the solution of the following linear integral equation:

$$\xi(u) = \check{f}(u; \eta^*) + \int \xi(v)\mathcal{K}^*(u, v)dv, \quad (17)$$

where

$$\check{f}(u; \eta^*) = \frac{\widehat{m}_1(u)\widehat{p}_1(u) - \widehat{m}_2(u)\widehat{p}_2(u) - \int \{b'(\eta^*(u) - \eta^*(v))\widehat{p}(u, v) - b'(\eta^*(v) - \eta^*(u))\widehat{p}(v, u)\} dv}{\widehat{w}_1(u; \eta^*) + \widehat{w}_2(u; \eta^*)}$$

and

$$\mathcal{K}^*(u, v) = \widehat{a}(u; \eta^*) \frac{\widehat{w}(u, v; \eta^*)}{\widehat{w}_1(u; \eta^*)} + (1 - \widehat{a}(u; \eta^*)) \frac{\widehat{w}(v, u; \eta^*)}{\widehat{w}_2(u; \eta^*)}.$$

We now state Lemmas 1–4.

LEMMA 1 *Under the conditions of Theorem 1, we have*

$$\|\widehat{F}(\eta^*)\|_{w^*} = O_p(n^{-2/(4+d)}) \text{ and } \|\widehat{F}(\eta^*)\|_\infty = O_p(n^{-2/(4+d)} \sqrt{\log n}).$$

LEMMA 2 *Under the conditions of Theorem 1, we have the following convergence for  $x \in (0, 1)$ :*

$$n^{2/(4+d)} (\check{\eta}(x) - \eta^*(x)) \xrightarrow{d} N(\beta(x), v(x)).$$

LEMMA 3 *Under the conditions of Theorem 1, we have*

$$\|\widehat{F}(\check{\eta})\|_{w^*} = o_p(n^{-2/(4+d)}) \text{ and } \|\widehat{F}(\check{\eta})\|_{\infty} = o_p(n^{-2/(4+d)}).$$

LEMMA 4 *Under the conditions of Theorem 1, the operator  $\widehat{F}'$  has a bounded inverse and it is Lipschitz continuous with stochastically bounded Lipschitz constant. Both properties hold near  $\eta^*$  and  $\check{\eta}$  with respect to the norms  $\|\cdot\|_{w^*}$  and  $\|\cdot\|_{\infty}$ .*

Lemma 1 and 3 can be shown by standard kernel smoothing arguments. Thus the proof of Theorems 1 and 2 is reduced to the treatment of a linear integral equations (Lemma 2) and the check of smoothness properties of the nonlinear integral operator  $\widehat{F}$  (Lemma 4).

**Proof of Lemma 2.** Because  $\check{\eta}$  is defined as the solution of a linear integral equation, with the operator  $K^*$  defined by  $K^*f(u) = \int K^*(u, v)f(v)dv$  we have that

$$\check{\xi} = (I - \mathcal{K}^*)^{-1} \check{f}(u; \eta^*).$$

Now, we have  $\check{f}(u; \eta^*) = \check{f}^A(u; \eta^*) + \check{f}^B(u; \eta^*)$  with

$$\begin{aligned} \check{f}^A(u; \eta^*) &= \rho^*(u)^{-1} \frac{1}{n} \sum_{i=1}^n K_{h_1}(u, X_1^i) \varepsilon^i - \rho^*(u)^{-1} \frac{1}{n} \sum_{i=1}^n K_{h_2}(u, X_2^i) \varepsilon^i, \\ \check{f}^B(u; \eta^*) &= \rho^*(u)^{-1} \frac{1}{n} \sum_{i=1}^n K_{h_1}(u, X_1^i) E[Y^i | X_1^i, X_2^i] - \rho^*(u)^{-1} \frac{1}{n} \sum_{i=1}^n K_{h_2}(u, X_2^i) E[Y^i | X_1^i, X_2^i] \\ &\quad - \rho^*(u)^{-1} \int b[\eta^*(u) - \eta^*(v)] \widehat{p}(u, v) dv + \rho^*(u)^{-1} \int b[\eta^*(v) - \eta^*(u)] \widehat{p}(v, u) dv \end{aligned}$$

with

$$\begin{aligned} \varepsilon^i &= Y^i - E[Y^i | X_1^i, X_2^i], \\ \rho^*(u) &= \widehat{w}_1(u; \eta^*) + \widehat{w}_2(u; \eta^*). \end{aligned}$$

We now use that uniformly for  $u \in [0, 1]^d$

$$(I - \mathcal{K}^*)^{-1} \check{f}^A(u; \eta^*) = \check{f}^A(u; \eta^*) + o_P(n^{-2/(4+d)}).$$

This can be shown as in the proof of Theorem 3 in Mammen, Støve and Tjøstheim (2008) where another linear integral equation is treated. This representation implies that

$$n^{2/(4+d)}(I - \mathcal{K}^*)^{-1} \check{f}^A(x; \eta^*) \xrightarrow{d} N(0, v(x)).$$

For the statement of the lemma it remains to show that

$$n^{2/(4+d)} [(I - \mathcal{K}^*)^{-1} \check{f}^B(x; \eta^*) - \eta(x)] = \beta(x) + o_P(n^{-2/(4+d)}).$$

For the proof of this claim it suffices to show that uniformly for  $u \in [0, 1]^d$

$$n^{2/(4+d)} [\check{f}^B(u; \eta^*) - (I - \mathcal{K}^*)\eta(u)] = (I - \mathcal{K}^*)\beta(u) + o_P(n^{-2/(4+d)}), \quad (18)$$

compare again the proof of Theorem 3 in Mammen, Støve and Tjøstheim (2007). For a proof of (18) note first that

$$\check{f}^B(u; \eta^*) = \rho^*(u)^{-1} [T_1(u) - T_2(u)]$$

with

$$\begin{aligned} T_1(u) &= \frac{1}{n} \sum_{i=1}^n K_{h_1}(u, X_1^i) \left\{ b'[\eta^*(X_1^i) - \eta^*(X_2^i)] - \int b'[\eta^*(u) - \eta^*(v)] K_{h_2}(v, X_2^i) dv \right\}, \\ T_2(u) &= \frac{1}{n} \sum_{i=1}^n K_{h_2}(u, X_2^i) \left\{ b'[\eta^*(X_1^i) - \eta^*(X_2^i)] - \int b'[\eta^*(v) - \eta^*(u)] K_{h_1}(v, X_1^i) dv \right\}. \end{aligned}$$

The statement of Lemma 2 now follows from the fact that  $T_j(u) - E[T_j(u)] = o_P(n^{-2/(4+d)})$ , uniformly for  $j = 1, 2$  and  $u \in [0, 1]^d$ .  $\square$

For the proof of Lemma 4 we need the following additional notation. The differential of  $\widehat{F}$  at  $\eta$  with increment  $g$ ,  $\widehat{F}'(\eta)g$ , is given by

$$-(\widehat{F}'(\eta)g)(u) = \{\widehat{w}_1^\eta(u) + \widehat{w}_2^\eta(u)\} \left[ g(u) - \left\{ \widehat{a}(u) \int g(v) \frac{\widehat{w}^\eta(u, v)}{\widehat{w}_1^\eta(u)} dv + (1 - \widehat{a}(u)) \int g(v) \frac{\widehat{w}^\eta(v, u)}{\widehat{w}_2^\eta(u)} dv \right\} \right]$$

where  $\widehat{w}^\eta(u, v) = b''(\eta(u) - \eta(v))\widehat{p}(u, v)$ ,  $\widehat{w}_1^\eta(u) = \int \widehat{w}^\eta(u, v) dv$ ,  $\widehat{w}_2^\eta(u) = \int \widehat{w}^\eta(v, u) dv$  and  $\widehat{a}(u) = \widehat{w}_1^\eta(u) \{\widehat{w}_1^\eta(u) + \widehat{w}_2^\eta(u)\}^{-1}$ .

Define  $w^\eta$ ,  $w_1^\eta$ ,  $w_2^\eta$ ,  $a$  and  $F'$  as above by replacing  $\widehat{p}$  with  $p$ . Consider an  $L_2$  space equipped with the norm  $\|g\|_2 = (\int g^2(u, v)w^\eta(u, v)dudv)^{1/2}$ , say  $H$ . Let  $H_j$  be  $L_2(w_j^\eta)$  spaces for  $j = 1, 2$  and denote the projections from  $H$  onto  $H_1$  and  $H_2$  with  $\Pi_1$  and  $\Pi_2$ , respectively. Then we have,

$$-(F'(\eta)g)(u) = \{w_1^\eta(u) + w_2^\eta(u)\} [g(u) - \{a(u)(\Pi_1g)(u) + (1 - a(u))(\Pi_2g)(u)\}].$$

Let the operator  $A^\eta$  be defined by  $(A^\eta g)(u) = a(u)(\Pi_1g)(u) + (1 - a(u))(\Pi_2g)(u)$ . Then we have

$$-(F'(\eta)g)(u) = \{w_1^\eta(u) + w_2^\eta(u)\} (I - A^\eta)g(u).$$

Here,  $I$  is the identity operator. Let  $w_{12}^\eta(u) = (w_1^\eta(u) + w_2^\eta(u))/2$ .

The proof of Lemma 4 makes use of the following lemma.

LEMMA 5 *There exists a positive constant  $\gamma_\eta$  such that*

$$\sup_{\|g\|_{w_{12}^\eta} \leq 1} \frac{\|A^\eta g\|_{w_{12}^\eta}}{\|g\|_{w_{12}^\eta}} < \gamma_\eta < 1.$$

**Proof of Lemma 5.** We will show that

$$\frac{\|A^\eta g\|_{w_{12}^\eta}^2}{\|g\|_{w_{12}^\eta}^2} \leq \frac{\|\Pi_1g\|_{w_1^\eta}^2 + \|\Pi_2g\|_{w_2^\eta}^2}{\|g\|_{w_1^\eta}^2 + \|g\|_{w_2^\eta}^2} \quad (19)$$

and that for  $k = 1, 2$  with a constant  $\rho_\eta < 1$

$$\frac{\|\Pi_k g\|_{w_k^\eta}^2}{\|g\|_{w_k^\eta}^2} < \rho_\eta. \quad (20)$$

Because of

$$\frac{\|\Pi_1g\|_{w_1^\eta}^2 + \|\Pi_2g\|_{w_2^\eta}^2}{\|g\|_{w_1^\eta}^2 + \|g\|_{w_2^\eta}^2} = \alpha \frac{\|\Pi_1g\|_{w_1^\eta}^2}{\|g\|_{w_1^\eta}^2} + (1 - \alpha) \frac{\|\Pi_2g\|_{w_2^\eta}^2}{\|g\|_{w_2^\eta}^2}$$

with  $\alpha = \|g\|_{w_1^\eta}^2 / (\|g\|_{w_1^\eta}^2 + \|g\|_{w_2^\eta}^2)$  equations (19), (20) and  $\|g\|_{w_{12}^\eta} = (\|g\|_{w_1^\eta}^2 + \|g\|_{w_2^\eta}^2)/2$  imply the statement of Lemma 5.

It remains to show (19), (20). The inequality (19) can be shown as follows

$$\begin{aligned}
& \|A^\eta g\|_{w_{12}^\eta}^2 \\
&= \int \{a^2(u)(\Pi_1 g)^2(u) + (1-a(u))^2(\Pi_2 g)^2(u) + 2a(u)(1-a(u))(\Pi_1 g)(u)(\Pi_2 g)(u)\} w_{12}^\eta(u) du \\
&= \frac{1}{2} \left[ \int (\Pi_1 g)^2(u) w_1^\eta(u) du + \int (\Pi_2 g)^2(u) w_2^\eta(u) du \right. \\
&\quad \left. - \int a(u)(1-a(u)) \{(\Pi_1 g)(u) - (\Pi_2 g)(u)\}^2 (w_1^\eta(u) + w_2^\eta(u)) du \right] \\
&\leq \frac{1}{2} \left\{ \|\Pi_1 g\|_{w_1^\eta}^2 + \|\Pi_2 g\|_{w_2^\eta}^2 \right\}.
\end{aligned}$$

It remains to show the inequality (20). Since  $\Pi_j$  for  $j = 1, 2$  are Hilbert-Schmidt operators,  $H_1 + H_2$  is closed. (See Bickel et al., 1993, A4, Proposition 2). Theorem 2 in Appendix 4 of Bickel et al. (1993) states that

$$\rho_\eta = \sup\{\langle h_1, h_2 \rangle_{w^\eta} : h_i \in \mathcal{H}_i \cap (\mathcal{H}_1 \cap \mathcal{H}_2)^\perp, i = 1, 2; \|h_i\|_{w^\eta} \leq 1\} < 1.$$

Now, since  $\|\Pi_k g\|_w \leq \|g\|_w$ , for nonconstant  $g$  and for  $k = 1, 2$ ,

$$\frac{\|\Pi_k g\|_{w_k^\eta}^2}{\|g\|_{w_k^\eta}^2} = \frac{\|\Pi_k g\|_{w^\eta}^2}{\|g\|_{w^\eta}^2} = \left\langle \frac{\Pi_k g}{\|g\|_{w^\eta}}, \frac{\Pi_k g}{\|g\|_{w^\eta}} \right\rangle = \left\langle \frac{\Pi_k g}{\|g\|_{w^\eta}}, \frac{g}{\|g\|_{w^\eta}} \right\rangle \leq \rho_\eta.$$

This completes the proof of Lemma 5. □

**Proof of Lemma 4.** Lipschitz continuity of  $\widehat{F}'$  can be shown by direct arguments. We only give a proof that the operator  $\widehat{F}'$  has a bounded inverse near  $\eta^*$  and  $\check{\eta}$  with respect to the norms  $\|\cdot\|_{w^*}$  and  $\|\cdot\|_\infty$ . Since  $w_1^\eta(u) + w_2^\eta(u)$  is strictly positive it suffices to show that  $I - A^\eta$  has a bounded inverse for the choice of  $\eta = \eta^*$  and  $\check{\eta}$ . We apply a version of the inverse mapping theorem for Banach spaces, see e.g. Chapter III in Conway (1990). The inverse mapping theorem states that a linear operator  $T$  has a bounded inverse if it is one-to-one, self-adjoint and bounded. We apply the inverse mapping theorem with  $T = I - A^\eta$ . We now show that  $I - A^\eta$  is one-to-one, self-adjoint and bounded.

For a function  $g$ ,  $(I - A^\eta)g = 0$  implies  $\langle (I - A^\eta)g, g \rangle_{w^*} = 0$ . Note that

$$\begin{aligned} \langle (I - A^\eta)g, g \rangle_{w^*} &= \langle a\{g - \Pi_1 g\} + (1 - a)\{g - \Pi_2 g\}, g \rangle_{w^*} \\ &= \frac{1}{2} \left\{ \langle g - \Pi_1 g, g \rangle_{w_1^\eta} + \langle g - \Pi_2 g, g \rangle_{w_2^\eta} \right\} \\ &= \frac{1}{2} \left\{ (\|g\|_{w_1^\eta}^2 - \|\Pi_1 g\|_{w_1^\eta}^2) + (\|g\|_{w_2^\eta}^2 - \|\Pi_2 g\|_{w_2^\eta}^2) \right\}. \end{aligned}$$

Thus, for nonconstant functions  $g$ ,  $\langle (I - A^\eta)g, g \rangle_{w^*}$  is positive. Since  $(I - A^\eta)g = 0$  implies  $g = 0$ , the operator  $I - A^\eta$  is one-to-one.

Now we will show that the operator  $I - A^\eta$  is self-adjoint. This can be shown as follows.

$$\begin{aligned} &\langle (I - A^\eta)g, h \rangle_{w^*} \\ &= \int h(u) \left\{ a(u) \int g(v) \frac{w^\eta(u, v)}{w_1^\eta(u)} dv + (1 - a(u)) \int g(v) \frac{w^\eta(v, u)}{w_2^\eta(u)} dv \right\} \frac{w_1^\eta(u) + w_2^\eta(u)}{2} du \\ &= \int g(u) \left\{ a(u) \int h(v) \frac{w^\eta(u, v)}{w_1^\eta(u)} dv + (1 - a(u)) \int h(v) \frac{w^\eta(v, u)}{w_2^\eta(u)} dv \right\} \frac{w_1^\eta(u) + w_2^\eta(u)}{2} du \\ &= \langle g, (I - A^\eta)h \rangle_{w^*}. \end{aligned}$$

Since  $\|I - A\| \leq 1 + \|A\|$ , the boundedness of the operator with respect to the norm  $\|\cdot\|_{w^*}$  follows from Lemma 5. The boundedness of the operator with respect to the norm  $\|\cdot\|_\infty$  follows from the following inequality

$$\begin{aligned} \|Ag\|_\infty &= \sup_u \left| a(u) \frac{\int g(v) w^\eta(u, v) dv}{w_1^\eta(u)} + (1 - a(u)) \frac{\int g(v) w^\eta(v, u) dv}{w_2^\eta(u)} \right| \\ &\leq \sup_u \left| a(u) \frac{\int |g(v)| w^\eta(u, v) dv}{w_1^\eta(u)} + (1 - a(u)) \frac{\int |g(v)| w^\eta(v, u) dv}{w_2^\eta(u)} \right| \\ &\leq \sup_u |a(u)| \|g\|_\infty + (1 - a(u)) \|g\|_\infty \\ &= \|g\|_\infty. \end{aligned}$$

This completes the proof of Lemma 4. □

## References

- [1] Andersen, E. (1970): ‘‘Asymptotic Properties of Conditional Maximum Likelihood Estimators’’, *Journal of the Royal Statistical Society Series B*, 32, 283-301.

- [2] Arellano, M. (2003): “Discrete Choice with Panel Data”, *Investigaciones Economicas* 27, 423-458.
- [3] Arellano, M. and Honore, B. (2001): “Panel Data Models. Some Recent Developments”, in J.Heckman and E. Leamer (eds.), *Handbook of Econometrics*, Vol. 5, Ch. 53.
- [4] Arellano, M. and Hahn, J. (2007): “Understanding Bias in Nonlinear Panel Data Models: Some Recent Developments”, In: R. Blundell, W. Newey, and T. Persson (eds.): *Advances in Economics and Econometrics*, Ninth World Congress, Cambridge University Press.
- [5] Chamberlain, G. (1984): “Panel Data”, in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics*, Vol. 2.
- [6] Chamberlain, G. (1992): “Binary Response Models for Panel Data: Identification and Information”, unpublished manuscript, Harvard University.
- [7] Chen, S. and Zhou, L. (2007): “Local partial likelihood estimation in proportional hazards regression”, *Annals of Statistics*, 35, 888-916.
- [8] Conway, J. (1990): *A course in functional analysis*, Springer, New York.
- [9] Fan, J., Gijbels, I. and King, M. (1997): “Local likelihood and local partial likelihood in hazard regression”, *Annals of Statistics*, 25, 1661-1690.
- [10] Hahn, J. and Newey, W. (2004): “Jackknife and Analytical Bias Reduction for Nonlinear Panel Data Models”, *Econometrica*, 72, 1295-1319.
- [11] Hastie, T. J. and Tibshirani, R. J. (1990): *Generalized Additive Models*. Chapman and Hall, London.
- [12] Hausman, J., Hall, B. and Griliches, Z. (1984): “Econometric Models for Count Data with an Application to the Patents-R&D Relationship”, *Econometrica*, 52, 909-938.

- [13] Honore, B. (1992): “Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects”, *Econometrica*, 60, 533-565.
- [14] Honore, B. and Tamer, E. (2006): “Bounds on Parameters in Dynamic Discrete Choice Models without Strict Exogeneity”, *Econometrica*, 74, 611-629.
- [15] Kyriazidou, E. (1997): “Estimation of a Panel Data Sample Selection Model”, *Econometrica*, 65, 1335-1364.
- [16] Mammen, E. and Nielsen, J. P. (2003): “Generalised structured models”, *Biometrika*, 90, 551-566.
- [17] Mammen, E. and Park, B. U. (2005): “Bandwidth selection for smooth backfitting in additive models”, *Annals of Statistics*, 33, 1260-1294.
- [18] Mammen, E., Støve, B. and Tjøstheim, D. (2008): “Nonparametric additive models for panels of time series ”, *Econometric Theory*, 25, 442-481.
- [19] Porter, J. (1996): *Essays in Econometrics* , PhD Thesis, (MIT).
- [20] Rasch. G. (1960): “Probabilistic Models for some Intelligence and Attainment Tests”, *Copenhagen: Danish Institute for Educational Research*. (Reissued, 1980, Chicago, University of Chicago Press).
- [21] Rasch, G. (1961): “On the General Law and the Meaning of Measurement in Psychology”, *Proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, UC Press, Berkeley and Los Angeles.
- [22] Yu, K., Mammen, E. and Park, B. U. (2008): “Smooth backfitting in generalized additive models”, *Annals of Statistics*, 36, 228-260.