

Semiparametric Single Index Models

Ichimura's and Klein and Spady's methods

Isa Marques Xi Sun Xueying Liu *

University of Bonn
Project Module in Econometrics and Statistics

Prof. Dr. Kneip
Prof. Dr. Liebl

March 5, 2017

Abstract

This study reviews semiparametric single index models discussed in Ichimura (1993) and Klein and Spady (1993). We elaborate on the theoretical justifications, as well as differences in the identification method. Both estimators are root-n consistent, asymptotically normal, as well as computationally expensive. While the maximum likelihood estimator proposed by Klein and Spady (1993) is fully efficient it lacks general applicability. In a simulation study we find that the bias of the estimator shows a decreasing nature for Ichimura's (1993), Klein and Spady's (1993) and the standard parametric logistic estimator. However, Klein and Spady's (1993) model exhibits a smaller finite sample bias. We apply the three estimators to a Voice Recognition data set to test for empirical differences. We find little to no differences in in-sample and out-of-sample accuracy, however, the semiparametric models require a much greater computational effort.

*All parties contributed to the presentation and the writing of the paper. Isa Marques led the theoretical part. Xi Sun and Xueying Liu led the applied part including R coding, simulations and real dataset application.

1 Introduction

Semiparametric single index models (SSIM) are widely applied in economic research. Applications range from finance to labor economics. Masten and Masten (2015) [[30]] use Klein and Spady's (1993) [[12]] SSIM to predict bankruptcies. Coelho et al. (2005) [[29]] use the same SSIM to analyze female labor market in Portugal. Moreover, McMillen and Thorsnes (2012) [[31]] adopt a semiparametric single index average derivative estimator to study the effect of copper smelter on house prices in Tacoma.

This paper reviews Ichimura's (1993) [[6]] and Klein and Spady's (KS) (1993) [[12]] models. Both models are root-n consistent and asymptotically normal. A natural attraction of the KS (1993) [[12]] model is that it reaches the semiparametric efficiency bound. Thus, it is fully efficient (Cameron and Trivedi, 2005 [[5]]). However, it lacks generality compared to Ichimura's (1993) [[6]] model, as the KS (1993) [[12]] model is restricted to binary outcomes.

We thus focus on a comparison between Ichimura's (1993) [[6]] model and the KS (1993) [[12]] model with regards to their theoretical justification, simulation performance, and empirical applicability. We find that Ichimura's (1993) [[6]] model and the KS (1993) [[12]] model share a very similar set of assumptions, specially the estimation of the link function $g(\cdot)$. The most noticeable difference is the use of a weighted nonlinear least squares method for Ichimura's (1993) [[6]] model, while the KS (1993) [[12]] model uses a maximum likelihood method. As to what concerns the Monte Carlo simulations, we found that the proposed estimators are expensive to compute. Moreover, as the number of grids increases due to larger number of independent variables as well as when the length of each grid increases, the number of function evaluations grows exponentially. Also when the true estimator is unknown, the performance of estimation relies heavily on the way grids are preselected. For an increasing sample size, the bias of the estimator shows a decreasing nature for Ichimura's (1993) [[6]], the KS (1993) and the classic logit model. The KS (1993) [[12]] model exhibits a smaller finite sample bias. Finally, we apply the three estimators to a Voice Recognition data set to test for empirical differences. We find little to no differences in in-sample and out-of-sample accuracy, however, the semiparametric models require a much greater computational effort.

This study builds on three blocks: theoretical outline; simulations, where finite sample properties are analyzed; and a real data set example, where one can illustrate the behavior of the models on an applied framework. The paper continues in the following structure. Section 2 of this paper provides a framework for SSIMs and a general comparison with parametric and nonparametric models. Section 3 outlines the identification conditions necessary in SSIMs. Ichimura's (1993) [[6]] and the KS (1993) [[12]] solutions will be analyzed in detail in section 4 and 5, respectively. The former section briefly explains the weight function. In section 6 the two models are briefly compared from a theoretical perspective. The simulation results from the Monte Carlo simulations are explained in section 7. Finally, in section 8 a real data set example is presented using the pre-processed Gender Recognition by Voice and Speech Analysis dataset.

2 Context

In this section, we elaborate on SSIMs' main features and contributions to the class of semiparametric models. Close at hand, a comparison with classical parametric models and nonparametric models is provided. In essence, the risk of misspecification is reduced relative to the overly restrictive but interpretable parametric models. Additionally, SSIMs avoid the inconveniences of fully nonparametric models such as the curse of dimensionality, difficulty of interpretation, and the lack of extrapolation capability. However, this comes at a cost, as computation for SSIMs is often difficult.

The model

$$Y_i = g(X_i' \beta_0) + \varepsilon_i, \quad (1)$$

where

- (i) $\{x_i, y_i\}$ for $i = 1, \dots, n$ is an i.i.d. sample;
- (ii) Y_i is the dependent variable, $X_i \in \mathbb{R}^q$ is a vector of explanatory variables, β_0 is the $q \times 1$ vector of unknown parameters;
- (iii) $X_i' \beta_0$ is a single index because it is a scalar;
- (iv) $E(\varepsilon_i | x_i) = 0$;
- (v) $g : \mathbb{R} \rightarrow \mathbb{R}$ is a smooth unknown link function;

is a SSIM.

Three points become relevant when explaining why the present model is of semiparametric nature. First, unlike fully nonparametric models, the functional form of the linear index is stated. However, as opposed to classical parametric models, the probability of ε conditioned on X is not specified except $E(\varepsilon | X) = 0$. Alongside, $g(\cdot)$ is left fully unspecified.

For illustrative purposes and as a mean of comparison to classical parametric models, we now analyze binary choice models in the setting proposed by Li and Racine (2007) [[1]]. The relationship between a binary dependent variable Y and covariates X is modelled by

$$Y_i = \begin{cases} 1, & \text{if } Y_i^* \stackrel{def}{=} \alpha + X_i' \beta + u_i > 0 \\ 0, & \text{if } Y_i^* = \alpha + X_i' \beta + u_i \leq 0 \end{cases}$$

where Y^* is a latent variable. Assuming a linear relationship between Y and X , the empirical analysis focuses on the estimation of β . Parametric methods to estimate β require assumptions on the distribution of the error term u . A common assumption in the parametric framework is $u \sim N(0, 1)$.¹ Let $F(\cdot)$ denote the true cumulative distribution function (CDF) of u . Then, the conditional expectation of Y has the form

¹With the identification condition $\sigma = 1$, β can be jointly identified (Madalla (1986) [[2]]) and we can use maximum likelihood.

$$\begin{aligned}
E(Y|x) &= \sum_{y=0,1} yP(y|x) = P(Y=1|x) = P(\alpha + x'\beta + u_i > 0) \\
&= P(u_i > -(\alpha + x'\beta)) = 1 - P(u_i \leq -(\alpha + x'\beta)) \\
&= 1 - F(-(\alpha + x'\beta)).
\end{aligned}$$

Moreover, if u has a symmetric distribution, $1 - F(-(\alpha + x'\beta)) = F(\alpha + x'\beta)$. Different functional forms for u lead to different functional forms for the conditional probability of $Y = 1$. If $F(\cdot)$ is the CDF of a standard normal variable, then a Probit model is obtained. Alternatively, for u following a symmetric logistic distribution, a Logit model is obtained. Moreover, consistent estimates of $E(Y|x) = P(Y=1|x)$ require the correct specification of the distribution of u . Hence, while model (1) still clings to many of the parametric model's desirable features, it is a more flexible version, as $g(\cdot)$ is left fully undefined. Distinctively, nonparametric models can be defined as

$$Y_i = g(X_i) + \varepsilon_i$$

with smooth g , assuming additivity of the error.² Thus, while the assumptions of a SSIM are in general weaker than those of a parametric model, they are stronger than those of a nonparametric model.³

Nonparametric models typically suffer from the curse of dimensionality, a term usually attributed to Bellman (1961) [[3]]. This term is defined by Gery Geenens (2011) [[4]] as being caused by the sparsity of data in high-dimensional spaces, which results in a decrease in fastest achievable rates of convergence of regression function estimators toward their target curve as the dimension of the vector of independent variables increases.

The rate of convergence of an estimator quantifies how fast the estimation error decreases as the sample size increases. The single index model avoids the curse of dimensionality by reducing the p -dimensional predictor to a univariate single-index. Thus, the estimator achieves the same convergence rate $n^{-\frac{1}{2}}$ that is optimal for most parametric models. For nonparametric models this rate is only $n^{-\frac{2}{5}}$, if the underlying function is twice continuously differentiable (Cameron and Trivedi (2005) [[5]]). Thus, in general, SSIMs reach greater estimation precision than fully nonparametric estimators with multidimensional vector of independent variables.

Nonetheless, semiparametric models in general have two important disadvantages. First, they are hard to compute. Ichimura's (1993) [[6]] model provides a good example of these computational issues, as it requires nonlinear iteration procedures. Second, semiparametric models often have multiple local optima, as they might require optimization of objective functions that are not unimodal. These problems seem to be exacerbated by increasing sample sizes or number of explanatory variables (Manski, 1975 [[7]], 1985 [[8]]; Manski and Thompson, 1989 [[9]]; Cosslett, 1983 [[10]]; Ichimura, 1993 [[6]]; Horowitz, 1992 [[11]]; and Klein and Spady, 1993 [[12]]).

²One can suggest an even more general model by dropping the additivity of the error term assumption. That is, $Y_i = g(X_i, \varepsilon_i)$.

³The semiparametric single index model might have weaker assumptions than a nonparametric model for the estimation of structural economic models [[13]].

3 Identification conditions

This section provides identification conditions for SSIMs, summarized in proposition 3.1. Brief intuitive explanations follow each of these conditions. Moreover, it is under these conditions that β_0 and $g(\cdot)$ are estimated in Ichimura's (1993) [[6]] model and the KS (1993) [[12]] model, which are analyzed in sections 4 and 5 respectively.

Model (1) implies

$$E(Y|x) = g(x'\beta_0). \quad (2)$$

Thus Y depends on x only through the linear combination $x'\beta_0$, and this relationship is characterized by the link function $g(\cdot)$.

Proposition 3.1. *Identification of β_0 and $g(\cdot)$ in model (2) requires that*

- (i) *The support of $x'\beta_0$ is a bounded convex set with at least one interior point.*
- (ii) *The vector of independent variables x should not contain a constant and it must contain at least one continuous variable with nonzero coefficient. Furthermore, one component of β_0 is set to 1.*
- (iii) *Function g is differentiable and it is not a constant function on the support of $x'\beta_0$;*
- (iv) *For the discrete components of x , changing the values of the discrete variables will not divide the support of $x'\beta_0$ into disjoint subsets.*

Condition (i) is plainly fundamental for the analysis. Thus, we won't dwell much on it. Imposing that the support of $x'\beta_0$ is a bounded convex set can for example prevent that it gets separated into disjoint subsets. This problem is analyzed in more detail in point (iv).

Apart from the identification restrictions on x outlined in condition (ii), we assume x cannot suffer from perfect multicollinearity. That is, there cannot be a perfect linear relationship between the independent variables. Otherwise, β_0 cannot be identified.

Some intuition can be provided to the specific requirements of condition (ii), following quite intertwined arguments. First, requiring x to contain at least one continuous variable (with nonzero coefficient) prevents x from having a finite support. Otherwise, $E(Y|X = x) = g(x'\beta_0)$ would impose only a finite number of restrictions on $g(\cdot)$, leading to an infinite number of different choices for $g(\cdot)$ and β_0 that satisfy those restrictions.⁴ Following similar reasoning, identification requires location and scale normalization. Define the function g^* such that $g^*(\gamma + v\delta) = g(v)$, for all v in the support of $x'\beta_0$. Then

$$E(Y|X = x) = g(x'\beta_0) \quad (3)$$

and

$$E(Y|X = x) = g^*(\gamma + x'\beta_0\delta). \quad (4)$$

⁴Note that if all x components are discrete we can still identify bounds on the components of β_0 , if g is assumed to be increasing. See Horowitz (1998) [[13]] for concrete examples.

Models (3) and (4) are observationally equivalent. Thus, β_0 and g are not identified unless restrictions are imposed to uniquely specify γ and δ . By restricting γ , one provides location normalization conditions. For example, by requiring that x does not include a constant. On the other hand, scale normalization conditions restrict δ . Here, it is assumed that β_0 has one of its components set to 1.⁵

Condition (iii) imposes restrictions on $g(\cdot)$, even though some of these can be weakened. To start with, $g(\cdot)$ cannot be a constant function. Otherwise, β_0 is not identified. Furthermore, what makes the identification of $E(Y|X = x)$ possible is that it remains constant if x changes in such a way that $x'\beta_0$ stays constant. However, $P(x'\beta_0 = c)$ is equal to zero, for $x_0'\beta_0$ continuously distributed and for some constant c . This renders identification impossible. By adding the assumption that $g(\cdot)$ is differentiable, $g(x'\beta_0)$ is close to $g(c)$ whenever $x'\beta_0$ is close enough to c . Then, the set of x for which $x'\beta_0$ is within any specified nonzero distance of c has nonzero probability for c in the interior of the support of $x'\beta_0$. Therefore, we identify β_0 by the approximate constancy of $x'\beta_0$. In fact, Wei Lin and Kulasekera (2007) [[14]] show that the weaker assumption that $g(\cdot)$ is continuous is sufficient for identification. Yet, differentiability is assumed on the remainder of the paper as it will become useful when analyzing Ichimura's (1993) [[6]] model and the KS (1993) [[12]] model.

The need for condition (iv), that is, the need to prevent $x'\beta_0$ from being divided into disjoint subsets, can be explained with an example inspired by Horowitz's (1998) [[13]]. Consider a SSIM in which the vector of independent variables, X , has a continuous component X_1 with support $[0, 1]$, and one discrete component X_2 , with support $\{0, 1\}$. Assume X_1 and X_2 are independent, $g(\cdot)$ is strictly increasing and nonperiodic and set $\beta_1 = 1$ as a *scale normalization* restriction. Consider in particular the case

$$\begin{aligned} E[Y|X = (x_1, 0)] &= g(x_1); \text{support of } g(\cdot) : [0, 1]; \\ E[Y|X = (x_1, 1)] &= g(x_1 + \beta_2); \text{support of } g(\cdot) : [\beta_2, 1 + \beta_2]. \end{aligned}$$

For $X_2 = 0$ the function $g(\cdot)$ is identified on $[0, 1]$. However, for $|\beta_2| > 1$ the support of $X_1 + \beta_2$ is disjoint from $[0, 1]$ and β_2 is an intercept in the model for $E(Y|X = (x_1, 1))$. Therefore, it is not possible to identify β_2 , as otherwise condition (ii) would be contradicted. However, for $0 < \beta_2 < 1$ the support of X_1 and $X_1 + \beta_2$ overlap. The interval of overlap is $[\beta_2, 1]$. Thus, $g(x_1 + \beta_2) = g(v)$ for some $v \in [0, 1]$. Then, $g(v)$ can be identified for $v \in [\beta_2, 1]$ by observations of X_1 for which $X_2 = 0$ and β_2 can be identified by solving

$$E[Y|X = (x_1, 1)] = g(x_1 + \beta_2), \quad (5)$$

on the set of x_1 in which the ranges of $g(x_1 + \beta_2)$ and $E[Y|X = (x_1, 1)]$ overlap.⁶

⁵This implies that X must have at least 2 dimensions. Otherwise β_0 is simply normalized to 1 and a one-dimensional nonparametric model $E(y|x) = g(x)$ with no semiparametric part is obtained instead.

⁶Note that if g was periodic on this set, (5) would have at least two solutions and β_2 would not be identified.

4 Ichimura's estimation model

In this section Ichimura's (1993) [[6]] estimation method for semiparametric models is analyzed. This method exhibits root-n consistency and asymptotic normality. A weight function that maximizes asymptotic efficiency is investigated. Nonetheless, multiple local minima may result.

Let β_0 denote the true value of β . For known g , β_0 is estimated by minimizing the nonlinear least squares (NLS) problem

$$S(\beta) = E[Y - g(x'\beta)]^2.$$

Exclusively as a means to understand the latter statement, assume $E[\varepsilon^2|X] = \sigma^2$. Then,

$$\begin{aligned} E[(Y - g(x'\beta))^2] &= E[\{Y - g(x'\beta_0) + (g(x'\beta_0) - g(x'\beta))\}^2] \\ &= E[\varepsilon^2] + 2E[\varepsilon(g(x'\beta_0) - g(x'\beta))] + E[(g(x'\beta_0) - g(x'\beta))^2] \\ &= \sigma^2 + 2E[E[\varepsilon(g(x'\beta_0) - g(x'\beta))|X] + E[(g(x'\beta_0) - g(x'\beta))^2]] \\ &= \sigma^2 + 2E[(g(x'\beta_0) - g(x'\beta))E[\varepsilon|X]] + E[(g(x'\beta_0) - g(x'\beta))^2] \\ &= \sigma^2 + E[(g(x'\beta_0) - g(x'\beta))^2] \end{aligned}$$

Where the last equality follows from the assumption $E[\varepsilon|X] = 0$. Hence, given the scale and location normalization conditions from Proposition 3.1., $E[(Y - g(x'\beta))^2]$ is minimal at $\beta_0 = \beta$.

Replace $S(\beta)$ by the empirical counterpart

$$S_n(\beta) = \frac{1}{n} \sum_{i=1}^n [Y_i - g(X_i'\beta)]^2. \quad (6)$$

Hence, one minimizes $S_n(\beta)$ instead of $S(\beta)$.

In the present case, $g(\cdot)$ is in fact unknown and it must be estimated. However, this cannot be done directly by kernel estimation as β_0 is also unknown. Still, for a given β we can estimate

$$G(X_i'\beta) \stackrel{\text{def}}{=} E(Y_i|X_i'\beta) = E[g(X_i'\beta_0)|X_i'\beta] \quad (7)$$

by a kernel method.

The Nadaraya-Watson (NW) kernel density estimator commonly takes the form

$$\hat{G}(X_i'\beta) = \frac{1}{nh_n\hat{p}(X_i'\beta)} \sum_{i=1}^n Y_i K\left(\frac{x'\beta - X_i'\beta}{h_n}\right)$$

where $\hat{p}(X_i'\beta) = (nh_n)^{-1} \sum_{i=1}^n K\left(\frac{x'\beta - X_i'\beta}{h_n}\right)$.

Ichimura (1993) [[6]] proposes modifications of the usual kernel estimation to estimate $G(X_i'\beta)$. Markedly, observation i is excluded from the calculation of $G(X_i'\beta)$. Otherwise, for a relatively small bandwidth, $S_n(\beta)$ is trivially minimized when $\hat{G}(X_i'\beta) = Y_i$. By leaving one observation out, this problem is resolved. Also, it validates the ability to predict the i th observation using the remaining observations in the sample. Therefore, outside the sample prediction is improved. Moreover, the denominator $\hat{p}(X_i'\beta)$ is random and it becomes

necessary to trim small values, particularly at the tails of the distribution. Otherwise, the value for the NW kernel estimator grows out of bound. Let $p(x'\beta)$ denote the probability density function of $X'_i\beta$ and A_δ and A_n be the sets

$$A_\delta = \{x : p(x'\beta) \geq \delta, \text{ for all } \beta \in \mathcal{B}\}$$

where $\delta > 0$ is a constant, $\mathcal{B} \in \mathbb{R}^q$, and

$$A_n = \{x : \|x - x^*\| \leq 2h_n \text{ for some } x^* \in A_\delta\}.$$

Then, for $x \in A_\delta$ the denominator does not get too close to zero. The set A_n where $\|\cdot\|$ is a Euclidean norm, is larger than A_δ but as $n \rightarrow \infty$, $h_n \rightarrow 0$ and A_n shrinks to A_δ .

With all of this in mind, a leave-one-out Nadaraya-Watson (NW) kernel estimator is obtained

$$\hat{G}_{-i}(X'_i\beta) = \frac{1}{nh_n\hat{p}_{-i}(X'_i\beta)} \sum_{j=1, j \neq i}^n w(x_j) \mathbf{1}(X_j \in A_n) Y_j K\left(\frac{X'_i\beta - X'_j\beta}{h_n}\right), \quad (8)$$

where $\hat{p}_{-i}(X'_i\beta) = (nh_n)^{-1} \sum_{j=1, j \neq i}^n w(x_j) \mathbf{1}(X_j \in A_n) K\left(\frac{X'_i\beta - X'_j\beta}{h_n}\right)$, $\mathbf{1}(X_i \in A_n)$ is a trimming function and $w(\cdot)$ is a nonnegative weight function chosen to maximize the asymptotic efficiency of the estimator. The latter is explained in detail in section 4.2.

Furthermore, without going into lengthy technical details, additional restrictions must be imposed on the model. A bounded second order kernel $K(u)$ with compact support is used. A kernel is of second order if its second moment is the first nonzero moment.⁷ This kind of kernel satisfies $0 \leq K(u) < \infty$, $K(u) = K(-u)$, $\int_{-\infty}^{\infty} K(u) du = 1$ and $\sigma_k^2 = \int_{-\infty}^{\infty} u^2 K(u) du < \infty$. Additionally, $g(\cdot)$ is twice continuously differentiable, which is a classical assumption for consistency of the kernel estimator. Moreover, an approximately optimal bandwidth balancing bias and variance is required. That is, $h_n = O(n^{-\frac{1}{5}})$.

With all of this in mind, it is possible to choose β by using a weighted NLS (WNLS) method

$$S_n(\beta) = \frac{1}{n} \sum_{i=1}^n w(x_i) \mathbf{1}(X_i \in A_\delta) [Y_i - \hat{G}_{-i}(X'_i\beta)]^2 \quad (9)$$

where $\mathbf{1}(X_i \in A_\delta)$ is a trimming function, and $w(x_i)$ is an appropriate nonnegative weighting function to maximize asymptotic efficiency.

Theorem 4.1. *According to Ichimura (1993) [6],*

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, \Omega_I),$$

with $\Omega_I = V^{-1} \Sigma V^{-1}$, where I stands for Ichimura and

$$\Sigma = E\{\mathbf{1}(X_i \in A_\delta) w(X_i)^2 \sigma^2(X_i) (g_i^{(1)})^2 (X_i - E(X_i | X'_i\beta_0)) \times (X_i - E(X_i | X'_i\beta_0))'\},$$

with $g_i^{(1)} = [\partial g(v) / \partial v]_{v=X'_i\beta_0}$, and

⁷The assumption that the support of the kernel is compact is used in order to simplify arguments.

$$V = E[\mathbf{1}(X_i \in A_\delta) w(X_i) (g_i^{(1)})^2 (X_i - E(X_i|X_i'\beta_0))(X_i - E(X_i|X_i'\beta_0))'].$$

It follows from the theorem 4.1. that $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_p(1)$, and thus $(\hat{\beta}_n - \beta_0) = O_p\left(\frac{1}{\sqrt{n}}\right)$. Consequently, root-n consistency is attained, which is the optimal convergence rate for most parametric methods. Furthermore, $\sqrt{n}(\hat{\beta}_n - \beta_0)$ is asymptotically normally distributed and its asymptotic distribution is centered at zero. The latter fact contrasts with the case of nonparametric density estimation, whose asymptotic distributions are in general not centered at zero when the estimators have their fastest possible rates of convergence (Stone (1980) [[15]] and Goldstein and Messer (1992) [[16]]).

A consistent estimator of Ω_I is given by

$$\hat{\Omega}_I = \hat{V}^{-1} \hat{\Sigma} \hat{V}^{-1},$$

where $\hat{V} = n^{-1} \sum_i w(X_i) (\hat{g}^{(1)})^2 (X_i' \hat{\beta}_n) (X_i - \hat{E}(X_i|X_i' \hat{\beta}_n)) (X_i - \hat{E}(X_i|X_i' \hat{\beta}_n))'$, $\hat{\Sigma} = n^{-1} \sum_i w(X_i)^2 \hat{\varepsilon}_i^2 (\hat{g}^{(1)})^2 (X_i' \hat{\beta}_n) (X_i - \hat{E}(X_i|X_i' \hat{\beta}_n))'$, $\hat{\varepsilon}_i = Y_i - \hat{g}(X_i' \hat{\beta}_n)$, $\hat{g}^{(1)}(X_i' \hat{\beta}_n) = [\partial \hat{g}_{-i} / \partial v]_{v=X_i' \hat{\beta}_n}$, $\hat{g}_{-i}(X_i' \hat{\beta}_n)$ is defined in (8), $\hat{E}(X_i|X_i' \hat{\beta}_n)' = \sum_j X_j K((X_i - X_j)' \hat{\beta}) / \sum_j K((X_i - X_j)' \hat{\beta}_n)$.

Heuristics for Theorem 4.1 are now provided, under a rather strong assumption. That is, assume $\beta_n - \beta_0 = O(n^{-\frac{1}{2}})$. Moreover, for what follows the trimming function $\mathbf{1}(X_i \in A_\delta)$ is ignored and $w(\cdot)$ is set to 1. Then,

$$\begin{aligned} S_n(\beta_n) &= \frac{1}{n} \sum_i \{Y_i - \hat{G}_{-i}(X_i' \beta_n)\}^2 = \frac{1}{n} \sum_i \{Y_i - \hat{G}_{-i}(X_i' \beta_n) + \hat{G}_{-i}(X_i' \beta_0) \\ &\quad - \hat{G}_{-i}(X_i' \beta_0)\}^2 = \frac{1}{n} \sum_i \{Y_i - G(X_i' \beta_n) + o_p(1) + \hat{G}_{-i}(X_i' \beta_0) \\ &\quad - \hat{G}_{-i}(X_i' \beta_0)\}^2 = \frac{1}{n} \sum_i \{Y_i - G(X_i' \beta_n) + \hat{G}_{-i}(X_i' \beta_0) - g(X_i' \beta_0) \\ &\quad + o_p(1)\}^2 = \frac{1}{n} \sum_i \{g(X_i' \beta_0) + \varepsilon_i - G(X_i' \beta_n) + \hat{G}_{-i}(X_i' \beta_0) - g(X_i' \beta_0) \\ &\quad + o_p(1)\}^2 = \frac{1}{n} \sum_i \{\varepsilon_i + \hat{G}_{-i}(X_i' \beta_0) - E[g(X_i' \beta_0)|X_i' \beta_n] + o_p(1)\}^2 \\ &= \frac{1}{n} \sum_i \{g(X_i' \beta_0) - E[g(X_i' \beta_0)|X_i' \beta_n] + \varepsilon_i + o_p(1)\}^2 \\ &= \frac{1}{n} \sum_i \{g(X_i' \beta_0) - E[g(X_i' \beta_0)|X_i' \beta_n] + \varepsilon_i\}^2 + o_p(1) \end{aligned}$$

Using Taylor expansions:

$$\begin{aligned} g(X_i' \beta_0) - E[g(X_i' \beta_0)|X_i' \beta_n] &= g(X_i' \beta_0) - g(X_i' \beta_n) \\ &\quad - g^{(1)}(X_i' \beta_n) E[(\beta_0 - \beta_n) X_i' | X_i' \beta_n] + o_p(1) \\ &= g^{(1)}(X_i' \beta_n) (X_i - E[X_i' | X_i' \beta_n]) (\beta_0 - \beta_n) + o_p(1) \\ &= g^{(1)}(X_i' \beta_0) (X_i - E[X_i' | X_i' \beta_0]) (\beta_0 - \beta_n) + o_p(1) \end{aligned}$$

Hence, for $g_{i0}^{(1)} = g^{(1)}(X_i' \beta_0)$ and $v_{i0} = X_i - E[X_i' | X_i' \beta_0]$

$$\begin{aligned} S_n(\beta_n) &= (\beta_0 - \beta_n)' \left[\frac{1}{n} \sum_i (g_{i0}^{(1)})^2 v_{i0} v_{i0}' \right] (\beta_0 - \beta_n) \\ &\quad + 2 \frac{1}{n} \sum_i \varepsilon_i g_{i0}^{(1)} v_{i0}' (\beta_0 - \beta_n) + \frac{1}{n} \sum_i \varepsilon_i^2 + o_p(1) \end{aligned}$$

Now following closely Li and Racine (2007) [[1]], the objective function is minimized with respect to β , ignoring terms independent of β_n and keeping the term $o_p(1)$ so as to make the previous approximations evident.⁸ Then, replace β_n by the minimizing value $\hat{\beta}_n$. Ultimately, one obtains

$$2 \frac{1}{n} (\hat{\beta}_n - \beta_0) \sum_i (g_{i0}^{(1)})^2 v_{i0} v_{i0}' - 2 \frac{1}{n} \sum_i \varepsilon_i g_{i0}^{(1)} v_{i0}' + o_p(1) = 0.$$

Hence,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \left(\frac{1}{n} \sum_i (g_{i0}^{(1)})^2 v_{i0} v_{i0}' \right)^{-1} \frac{1}{\sqrt{n}} \sum_i \varepsilon_i g_{i0}^{(1)} v_{i0}' + o_p(1).$$

With this in mind, as well as the Lindeberg-Levy Central Limit Theorem and the Law of Large Numbers, we obtain the result from Theorem 4.1 if we set $w(X_i) = 1$ and ignore the trimming function $\mathbf{1}(X_i \in A_\delta)$.

4.1 Bandwidth Selection

Ichimura (1993) [[6]] suggests the use of the optimal smoothing parameter h_n , which balances bias and variance. That is, $h_n = O(n^{-\frac{1}{5}})$.⁹ However, even though Ichimura (1993) [[6]] gives a range of bandwidth choices which enables the construction of a root- n consistent $\hat{\beta}_n$, it excludes the size of bandwidth that is optimal for estimating $g(\cdot)$.¹⁰

With this in mind, Härdle et al. (1993) [[18]] suggest an empirical way of selecting the bandwidth for optimal smoothing of both $g(\cdot)$ and β . This can be attained by selecting h_n and β simultaneously by minimizing

$$M(\beta, h_n) = \sum_i \left[Y_i - \hat{G}_{-i}(X_i' \beta, h_n) \right]^2 \mathbf{1}(X_i \in A_\delta), \quad (10)$$

where $\hat{G}_{-i}(X_i' \beta, h_n) = \hat{G}_{-i}(X_i' \beta)$ follows equation (8), and $\mathbf{1}(X_i \in A_\delta)$ is the trimming function, defined in the beginning of the present section.

⁸The sketch for the proof can be found on pages 256-277.

⁹The original paper (1993) [[6]] suggests conditions that, according to Ichimura, satisfy optimal smoothing. However, no explicit explanation is provided, and thus we refrain from further details.

¹⁰Hall (1989) [[17]] shows that two very different bandwidths may be necessary to construct good estimators of both $g(\cdot)$ and β .

4.2 Weight Function

A weight function is introduced for efficiency reasons. Efficiency is desirable because the more efficient an estimator is, the smaller the amount of dispersion it has around its expected value and the more precise it is as an estimator of the corresponding parameter. Ichimura's (1993) [[6]] model does not preclude heteroskedasticity, which is a source of inefficiency. Heteroskedasticity occurs when the variance of the unobservable error ε , conditional on the vector of independent variables, is not constant. That is, $\text{Var}(\varepsilon_i|X_i) = \sigma_i^2$.

To take heteroskedasticity into account, an analogue of Generalized Least Squares (GLS) is used. This method transforms the model in order to maximize asymptotic efficiency. Also analogously to GLS, if the transformation depends upon unknown parameters, these need to be estimated and an analogue of Feasible GLS (FGLS) is obtained instead. On the latter case, a weight function is used that assumes a general form of heteroskedasticity. If $\text{Var}(\varepsilon_i|X_i) = \sigma^2$ is constant, it can be shown that the optimal choice of $w(x_i)$ is $w(x_i) = 1$. Thus, by simply setting $w(x_i) = 1$, Ω_I is the semiparametric efficiency bound. However, if $\text{Var}(\varepsilon_i|X_i) = \sigma_i^2$, asymptotic efficiency is less easily achieved.

The problem of efficient estimation of β_0 in a single index model with unknown $g(\cdot)$ is analyzed by Härdle et al. (1993) [[18]] and Newey and Stocker (1993) [[19]]. Under certain regularity conditions, the efficiency bound for the single index model, with unknown $g(\cdot)$ and using only data for which $X \in A_\delta$, is Ω_I from Theorem 4.1., for $w(x) = \frac{1}{\sigma^2(x)}$.¹¹ Thus, the weight function allows us to weight each observation by a factor proportional to the error variance. Moreover, observations with higher variance get a smaller weight.

The efficiency bound is then

$$\Omega_{SI} = \left\{ E \left[\frac{\mathbf{1}(X_i \in A_\delta)}{\sigma^2(x)} \frac{\partial}{\partial \beta} G(X' \beta) \frac{\partial}{\partial \beta} G(X' \beta)' \right] \right\}^{-1} \quad (11)$$

where SI stands for single index. Thus, the efficiency bound from equation (11) is achieved by the semiparametric WNLS estimator if $\sigma^2(X)$ is known.

To make the equality between equation (11) and Ω_I clear, for known $\sigma^2(X)$, consider the result shown in the present section

$$E[g(X'_i \beta_0)|X'_i \beta] = g(X'_i \beta_0) - g^{(1)}(X'_i \beta_0)(X_i - E[X'_i|X'_i \beta_0])(\beta_0 - \beta) + o_p(1).$$

Minimizing the previous result in order to β , ignoring terms independent of β and keeping the term $o_p(1)$ so as to make the previous approximations evident, one obtains

$$\frac{\partial}{\partial \beta} G(X'_i \beta) = g^{(1)}(X'_i \beta_0)(X_i - E[X'_i|X'_i \beta_0]) + o_p(1).$$

Even when $\sigma^2(X)$ is unknown, an asymptotic efficient estimator of β_0 can be obtained by using an analogue of FGLS. Consider a given consistent estimator of $\sigma^2(X)$, say $\hat{\sigma}_n^2(x)$, that adopts a two-step procedure. On the first step, minimize function (9) with respect to β for $w(x) = 1$. The resulting estimator $\hat{\beta}_n$ is root- n consistent and asymptotically normal but inefficient. This estimator is used on

¹¹The assumption that only data where $X \in A_\delta$ is used can be relaxed by letting A_δ grow very slowly as n increases.

the second step to calculate the weight function $\hat{w}_i(x) = \frac{1}{\hat{\sigma}_i^2}$. Robinson (1987) [[20]] estimates $\hat{\sigma}_i^2$ using a nearest-neighbor nonparametric regression estimator where $\hat{\varepsilon}_i = Y_i - \hat{G}_{-i}(X_i' \hat{\beta}_n)$ and $\hat{\beta}_n$ is the estimator from the first step. This method is used to avoid technical problems if X has unbounded support or a density that can be arbitrarily close to zero. However, in practice a kernel estimator will suffice as A_δ and A_n are chosen so as to keep the estimated density of X away from zero.

4.3 Model's disadvantages

It is important to realize that the minimization of a nonlinear objective function such as equation (9) might be computationally costly. The WNLS estimator is computed by iterative methods. Start with an initial guess for the estimator $\hat{\beta}_n^1$ such as $\hat{\beta}_n^1 = -\frac{1}{n} \sum_i y_i \hat{f}'(x_i)$, where \hat{f}' can be obtained by calculation of the first derivative of the kernel estimator of the density of x_i .¹² Moreover, $\hat{\beta}_n^1$ follows the restrictions from section 3. First, one reaches the kernel estimate $\hat{G}_{-i}(X_i' \hat{\beta}_n^1)$ and thus $S_n(\hat{\beta}_n^1)$. Afterwards, $\hat{\beta}_n^1$ must be perturbed so as to obtain $\frac{\partial S_n(\beta)}{\partial \beta} \big|_{\hat{\beta}_n^1}$. Then, one updates $\hat{\beta}_n^2 = \hat{\beta}_n^1 + A_n \frac{\partial S_n(\beta)}{\partial \beta} \big|_{\hat{\beta}_n^1}$ where A_n is the size of the random disturbance. This process should be repeated until convergence. Yet, this is computationally difficult, specially because there might be multiple local minima, in case, for example, the objective function is multimodal or nonconvex. One may consider as an alternative a direct estimation method that does not require optimization of problems involving iterative solutions. Härdle and Stoker (1989) [[22]] provide such a method. However, according to Racine and Li (2007) [[1]], for small sample settings, Härdle and Stoker's (1989) [[22]] direct method may still be less appealing than Ichimura's (1993) [[6]] iterative method.

5 Klein and Spady's binary estimation model

In this section we analyze the KS (1993) [[12]] SSIM model, which is used to estimate equation (2) when $Y \in \{0, 1\}$. This method exhibits root-n consistency, asymptotic normality and asymptotic efficiency.

The model is defined as

$$Y_i = \mathbf{1}(X_i' \beta_0 \geq \varepsilon_i), \quad (12)$$

where ε is a random disturbance. Furthermore, equation (7) holds and for known $g(\cdot)$, the asymptotically efficient estimator of β_0 is a maximum likelihood estimator (MLE).

The *log-likelihood* function is

$$\mathcal{L}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \{(1 - Y_i) \ln[1 - g(X_i' \beta)] + Y_i \ln[g(X_i' \beta)]\}. \quad (13)$$

It is clear from equation (13) that restrictions must be imposed such that any estimate of $g(\cdot)$ is kept sufficiently far away from 0 and 1. As in Ichimura's (1993) [[6]] model, this can be achieved by using a simplified trimming function

¹²This estimator is a density weighted average derivative estimator such as in Stoker (1986) [[21]].

$\mathbf{1}(X_i \in A_\delta)$ that restricts X to a fixed set A_δ on which $g(\cdot)$ is bounded away from 0 and 1. The set A_n is defined as in section 4.

For the same reason as for Ichimura's method, $g(\cdot)$ cannot be estimated directly by kernel estimation. As equation (7) holds, KS (1993) [[12]] suggest using an unweighted version of the leave-one-out NW estimator from equation (8). Again, one observation is left out as otherwise, for a relatively small bandwidth, the objective function would be maximized by setting $\hat{G}(X'_i\beta) = 0$ when $Y_i = 0$ and $\hat{G}(X'_i\beta) = 1$ when $Y_i = 1$. That is, \mathcal{L}_n would be zero. By leaving one observation out, outside the sample prediction is improved.

Furthermore, without spanning much on technical details, further restrictions must be imposed on the model. KS (1993) [[12]] assume a higher order kernel density estimator $K(u)$ with compact support is used. Higher order kernels have order bigger than two and, in general, allow for faster asymptotic rates of convergence by relaxing the restriction that the kernel must be a density function. However, these are seldomly used in applied work as they involve local averaging with negative weights (Marron (1994) [[23]]).¹³ The bandwidth must satisfy the rate $n^{-\frac{1}{6}} < h_n < n^{-\frac{1}{8}}$. Moreover, $P(Y = 1|X = x)$ need not be monotonic in x .

With all of this in mind, the maximization problem takes the form

$$\mathcal{L}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_i \in A_\delta) \{ (1 - Y_i) \ln[1 - \hat{G}_{-i}(X'_i\beta)] + Y_i \ln[\hat{G}_{-i}(X'_i\beta)] \}, \quad (14)$$

where \hat{G}_{-i} follows equation (8) and $\mathbf{1}(X_i \in A_\delta)$ is a trimming function.

Theorem 5.1. *According to KS (1993) [[12]],*

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, \Omega_{KS}),$$

with

$$\Omega_{KS} = \left\{ E \left[\frac{\mathbf{1}(X_i \in A_\delta)}{g(X'_i\beta_0)(1 - g(X'_i\beta_0))} \frac{\partial}{\partial \beta} G(X'_i\beta) \frac{\partial}{\partial \beta} G(X'_i\beta)' \right] \right\}^{-1},$$

where KS stands for Klein and Spady and $\Omega_{KS} = \Omega_{SI}$, i.e., the estimator is asymptotically efficient.

It follows from theorem 5.1. that $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_p(1)$, and thus $(\hat{\beta}_n - \beta_0) = O_p\left(\frac{1}{\sqrt{n}}\right)$. Consequently, root-n consistency is attained, which is the optimal convergence rate for most parametric methods. Furthermore, $\sqrt{n}(\hat{\beta}_n - \beta_0)$ is asymptotically normally distributed and its asymptotic distribution is centered at zero. Once again, the latter fact contrasts with the case of nonparametric density estimation, whose asymptotic distributions are in general not centered at zero when the estimators have their fastest possible rates of convergence (Stone (1980) [[15]] and Goldstein and Messer (1992) [[16]]).

However, to have a good grasp of Theorem 5.1., it is fundamental to understand the expression for Ω_{KS} . In particular, if one is to show that $\Omega_{KS} = \Omega_{SI}$, it is

¹³In fact, for its desirable properties to “kick in”, a very large number of observations might be necessary. Marron and Wand (1992) [[24]] conclude that, in most situations, the benefits in terms of performance of higher order kernels are not sufficient to outweigh the loss of interpretability.

necessary to understand that $\frac{1}{g(X'_i\beta_0)(1-g(X'_i\beta_0))}$ indeed corresponds to $\frac{1}{\sigma^2(x)}$ in equation (11).

The variance of the present binary choice model is

$$\begin{aligned} \text{Var}(Y|X=x) &= P(Y=1|X=x)[1-P(Y=1|X=x)] \\ &= P((x'\beta_0 \geq \varepsilon)|X=x)[1-P((x'\beta_0 \geq \varepsilon)|X=x)] \\ &= g(x'\beta_0)[1-g(x'\beta_0)]. \end{aligned}$$

Hence, the model can handle heteroskedasticity only if the “error” distribution depends on the same index, $x'\beta_0$, that determines the “mean response” (Klein and Vella (2006) [[25]]).

If one differentiates the right-hand side of equation (14) with respect to β and then replace β by the minimizing value $\hat{\beta}_n$, the first order condition is

$$\frac{1}{n} \sum_{i=1}^n \frac{\hat{G}'_{-i}(X'_i\hat{\beta}_n)}{\hat{G}_{-i}(X'_i\hat{\beta}_n)(1-\hat{G}_{-i}(X'_i\hat{\beta}_n))} (Y_i - \hat{G}_{-i}(X'_i\hat{\beta}_n)) \mathbf{1}(X_i \in A_\delta) X_i = 0 \quad (15)$$

with probability approaching 1 as $n \rightarrow \infty$ and where $\hat{G}'_{-i}(z) = \partial \hat{G}_{-i}(z) / \partial z$. Consider now a semiparametric WNLS estimation of β_0 such as in Ichimura’s (1993) [[6]] model. The first order condition in equation (15) is the same as the first-order condition for the semiparametric WNLS estimation of β_0 with the estimated weight function

$$\begin{aligned} w_i &= \{\hat{G}_{-i}(X'_i\hat{\beta}_n)[1-\hat{G}_{-i}(X'_i\hat{\beta}_n)]\}^{-1} = \{G(X'_i\beta_0)[1-G(X'_i\beta_0)]\}^{-1} + o_p(1) \\ &= \{g(X'_i\beta_0)[1-g(X'_i\beta_0)]\}^{-1} + o_p(1) = \text{Var}(Y|X=x)^{-1} + o_p(1) \end{aligned}$$

Based on section 4.2., it follows that the KS (1993) [[12]] model is asymptotically efficient.

5.1 Bandwidth Selection

KS (1993) [[12]] does not discuss the choice of bandwidth, apart from the requirement that $n^{-\frac{1}{6}} < h_n < n^{-\frac{1}{8}}$. However, analogously to Ichimura’s (1993) [[6]] method, one can conjecture that the result from Härdle et al. (1993) [[18]] applies here. An empirical way of selecting the bandwidth for optimal smoothing of both $g(\cdot)$ and β is attained by choosing h_n jointly with β when maximizing

$$M_n(\beta, h_n) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_i \in A_\delta) \{(1-Y_i) \ln[1-\hat{G}_{-i}(X'_i\beta, h_n)] + Y_i \ln[\hat{G}_{-i}(X'_i\beta, h_n)]\} \quad (16)$$

where $\hat{G}_{-i}(X'_i\beta, h_n) = \hat{G}_{-i}(X'_i\beta)$ follows equation (8), and $\mathbf{1}(X_i \in A_\delta)$ is the trimming function, supposedly defined as at the beginning of section 5.

5.2 Model’s disadvantages

Analogously to Ichimura’s (1993) [[6]] model, optimization of the maximum likelihood function, using the iteration between the computation of $\hat{\beta}_n$ given \hat{G}_{-i} and the computation of \hat{G}_{-i} given $\hat{\beta}_n$, is often difficult. Especially so when

the optimization leads to multiple local maxima. Such an event could be caused by an objective function that is either multimodal or nonconcave.

Zhou and Lang (1995) [[26]], for example, provide an alternative using an “easy to compute” semiparametric estimator for binary choice models. It preserves root-n consistency and asymptotically normally, but has the novelty of being fast and easy to compute.¹⁴

6 Theoretical comparison between Ichimura’s and Klein and Spady’s models

If one restricts outcomes to be binary, the KS (1993) [[12]] model seems more adequate than Ichimura’s (1993) [[6]] model. The KS (1993) [[12]] model is efficient in the sense that it reaches the semiparametric efficiency bound. Maximum-likelihood functions have “natural” weighting. On the other hand, Ichimura’s (1993) [[6]] model requires a weight function, and potentially the two-step procedure described in section 4.2., to reach the asymptotic efficiency bound. Thus, much is saved in terms of simplicity and asymptotic efficiency is in general improved by using the KS (1993) [[12]] model for the binary case.

However, restricting analysis to the case of a binary outcomes is quite stringent. Indeed, this fact gives a clear advantage to Ichimura’s (1993) [[6]] model, as his WNLS model allows for continuous outcomes. Thus, it can be applied to a much broader variety of settings than the KS (1993) [[12]] model.

7 Simulation results

In this section, datasets are simulated from Ichimura’s (1993) [[6]] and the KS (1993) [[12]] theoretical models. Estimators are then estimated using this dataset. We reflect on the properties of these estimates, generated via a Monte Carlo experiment with sample size of 250 and 1000 trials. Moreover, we analyze finite sample bias for sample sizes of 50, 150 and 250.

Several points are worth noting in the design of the experiment. Firstly, two scenarios are simulated based on different error distributions: standard normal $N(0, 1)$ and joint normal $0.75 \cdot N(-0.5, 1) + 0.25 \cdot N(1.5, 2.5)$. For each scenario, a logit estimator is included as an additional means of comparison with the two SSIMs. Furthermore, both specifications contain two independent/exogenous variables, each independently generated from a standard normal distribution. Lastly, we restrict our model to the binary case, as this satisfies the restrictions of the KS (1993) [[12]] model. Following section 3, the coefficient of the first exogenous variable is normalized to 1. Moreover, the true value for the second is arbitrarily set to -2. The dependent variable Y is thus generated based on the following equation

$$Y_i = I(x_{1i} - 2x_{2i} > \epsilon_i).$$

Our simulation characterizes Ichimura’s (1993) [[6]] nonlinear least squares estimator and the KS (1993) [[12]] maximum likelihood estimator. Again for convenience of comparison, a fourth order Gaussian kernel function is selected

¹⁴The method is based on a semiparametric interpretation of the Expectation and Maximization principle (Dempster et al (1977) [[27]]) and the least squares approach.

for both cases, in order to achieve the requirement in the KS (1993) [[12]] model. Additionally, a trimming function is defined, following the theoretical requirements from section 4 and 5. This guarantees sensible calculation of estimates over each randomly drawn data set. Although Ichimura’s (1993) [[6]] and the KS (1993) [[12]] estimators of one individual simulation are extracted from the same trimmed data set, a “floor” is executed on the latter. This is due to the stronger restriction that $g(\cdot)$ is between 0 and 1 for the KS (1993) [[12]] model. As for bandwidth selection, we preselect two bandwidths respectively for the two models, through a likelihood-based cross-validation approach, which is used in the function *npindexbw* of the empirical np-package created by Hayfield and Racine (2008) [[28]].

In accordance with Ichimura (1993) [[6]], grid search was chosen as the method of optimization nested in our estimation process. This method performs sufficiently well and is stable in finding “good” estimates, as supported by histograms of estimates shown in *Figure 1*. However, it may have two considerable deficiencies: first, as the number of grids increases due to larger number of independent variables as well as when the length of each grid increases, the number of function evaluations grow exponentially; second, when the true β is unknown, the performance of estimation relies heavily on the way grids are preselected. An extreme example of the latter fact would be when the true beta lies outside the pre-specified grid, in which case we would never obtain an estimator close to the true value. In our experiment grid search is set to be $(-4, 0)$ with width 0.05. This leads to 81 function evaluations for each data generation, which is already computationally costly, and even more so with 1000 data generation trials.

Estimator ¹	Standard Normal		Joint Normal	
	Bias	RMSE	Bias	RMSE
Ichi	-0.045	0.3672	-0.057	0.3665
KS	-0.0095	0.3231	-0.015	0.3125
Logit	-0.030	0.2760	-0.034	0.2813

Table 1: Bias and Root Mean Squared Error Comparison

¹Ichi = Ichimura’s method, KS = Klein and Spady’s method.

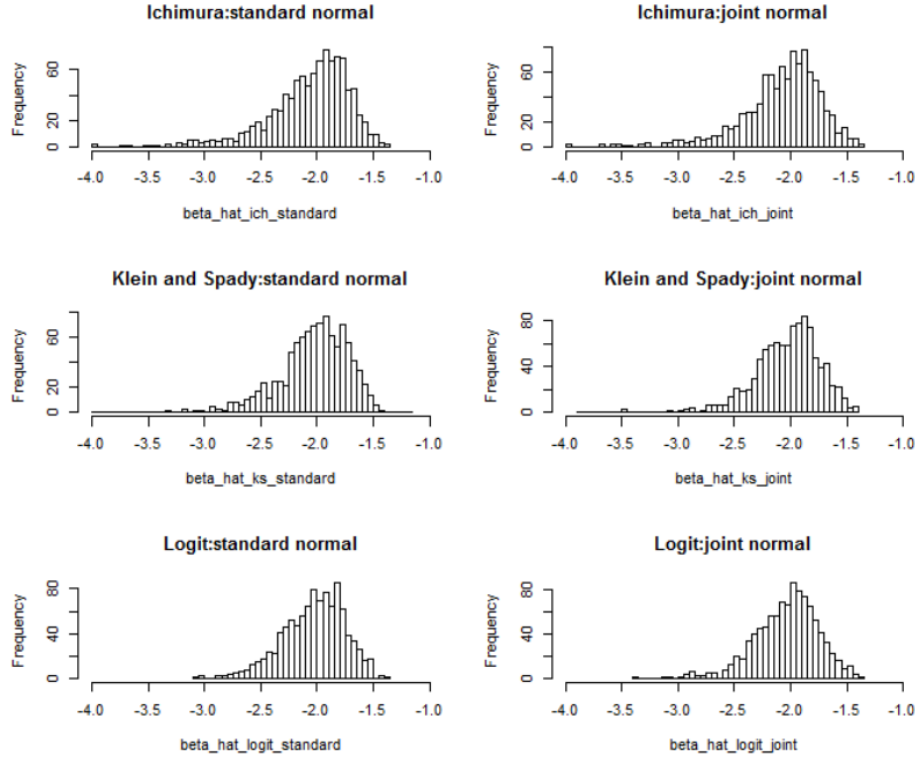


Figure 1: Plot of the estimates

First, note that for a given numbers of simulations m

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{\beta}_i - \beta_0)^2}.$$

Combining simulation results from *Table 1* and *Figure 1*, it can be observed that the distributions of all six estimators are skewed to the left, which can be due to finite sample properties. Moreover, the biases are bounded below by - 0.6. As to what concerns the mean squared error, there is no significant difference among estimators in terms of magnitude. Since binary choice models are inherently heteroskedasticity, Ichimura's (1993) [[6]] model in particular would require the use of a weight function. However, this is out of the scope of our work and it is also excluded from the np-package created by Hayfield and Racine (2008) [[28]].

Estimator	n=50	n=150	n=250
Ichi	-0.1477	-0.06515	-0.04485
KS	0.0299	-0.03015	-0.00945
Logit	-0.240235	-0.05246	-0.030321

Table 2: Bias of the estimators for different sample sizes

To fully understand the finite sample properties of the estimators, we need to study consistency. Indeed, observing *Table 2*, the bias of the estimators presents a decreasing nature for all cases, despite small increases. The KS (1993) [[12]] model exhibits a smaller finite sample bias (ranging in absolute terms from 0.0095 to 0.0302) than Ichimura’s (1993) [[6]] model (ranging from 0.0449 to 0.1477) and the logit model (ranging from 0.0303 to 0.2402). Further simulations with increasing number of observations would be desirable, for a more indepth and reliable study of consistency. However, computational issues limit our analysis. More details on consistency and asymptotic normality can be found in the appendix (section 11.3).

Ultimately, a comparison between our simulation results and Hayfield and Racine’s (2008) np-package [[28]] encouraged us to rely on our own implementation. This study is included in the appendix (section 11.2). The weight function can however be employed in further simulation studies.

8 Empirical Application

This section presents an empirical application on gender recognition by voice using the aforementioned statistical methods. A comparison with simple logistic regression is also made.

The dataset we used is the pre-processed Gender Recognition by Voice and Speech Analysis dataset, a publicly available dataset from online resources containing 3,168 recorded voice samples by 1584 male and 1584 female speakers. 21 independent variables of acoustic properties of voice and 1 binary dependent indicator variable for gender are included in the dataset. More information regarding the dataset can be found in the appendix.

As a first step, we removed three variables that have shown to cause multicollinearity problem, namely; IQR, centroid, and dfrange. Next, the dataset is split into training sample for estimation of parameters, and test sample for prediction using estimated parameters. 70% of the data are randomly selected into the training sample, with the remaining 30% left for the test sample.

For estimation, we use the pre-implemented np-package in R (Racine and Hayfield, 2008 [[28]]) for computational simplicity. In our self-implemented Ichimura (1993) [[6]] and KS (1993) [[12]] methods we employed grid search in the minimization process for equation (9) and (14), which is also the method proposed by Ichimura (1993) [[6]]. The minimization results are highly dependent on the specification of the grid for each independent variables and an appropriate resolution of the grid is necessary for precise results. As such, the grid search method suffers from the curse of dimensionality. Numerical approximation becomes infeasible in our case as the data contains 18 independent variables. We

therefore use the np-package which employs the *npksum()* and *nlm()* minimization procedures in R with multiple starting values to best avoid local minima. A comparison with logistic regression is also provided using the pre-implemented logistic regression routine in R.

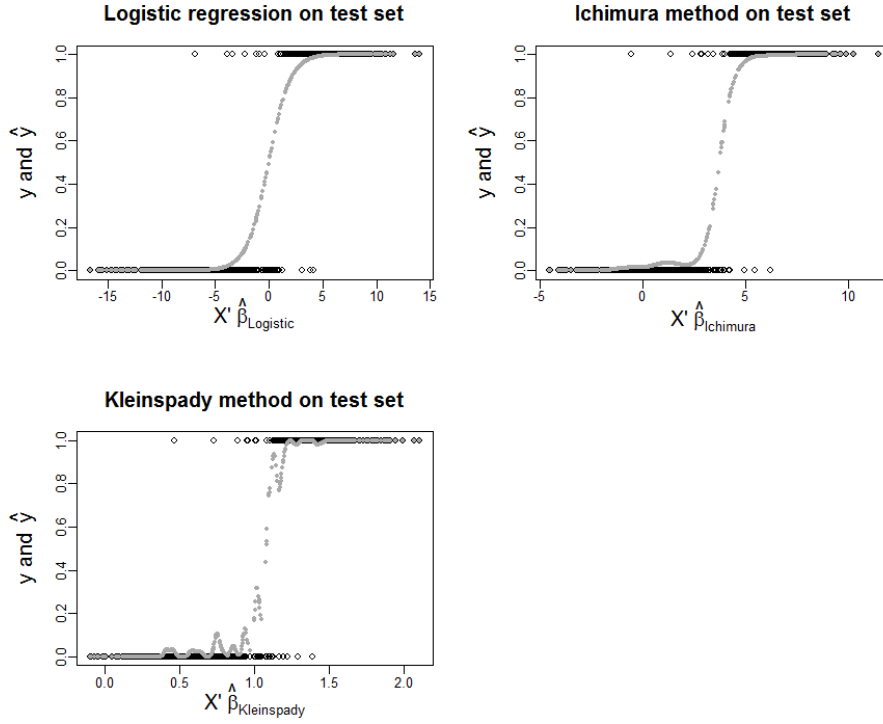


Figure 2: Plot of Estimates on Test Sample

The prediction results and the true values are plotted in *Figure 2*, with grey points representing predicted values and black points true values. A comparison of in-sample and prediction accuracy rate and approximate calculation time are also presented in *Table 3*. The accuracy rate is defined as ratio of correct gender prediction to the total number of data points.

From *Table 3* and *Figure 2*, we could conclude that the logistic regression would be the preferred method in this case since the improvement in accuracy rate by using semiparametric methods are minimum while the calculation time increases significantly. However, this could be due to the fact that the underlying data fit the logistic distribution better, and thus this comparison should not be generalized to other datasets.

We also take note of the two peculiar features of the results. First, the accuracy rate is higher for test sample than for training sample. As the difference is relatively small, we postulate that this could be random and of no particular significance. Moreover, the fact that the training sample contains 3.5 times more data points than the test sample could exacerbate the random effect. The second feature is that the resulting graph for the KS (1993) [\[\[12\]\]](#) method is not

	Logistic Regression	Ichimura Method	Klein-Spady Method
Approximate calculation time	< 1 second	> 3 hours	> 3 hours
In-sample Accuracy Rate (training sample)	0.971145	0.974752	0.976104
Prediction Accuracy Rate (test sample)	0.978947	0.977894	0.981052

Table 3: Estimation time and accuracy rate

very smooth. This could be due to a problem with bandwidth selection included in the np-package (Racine and Hayfield, 2008) [[28]] or the minimization process. However, as the result is not significantly worse than the other estimators, actually slightly better, we do not think that this should be of great concern.

9 Conclusion

This study investigates the class of semiparametric single index models as discussed in Ichimura (1993) [[6]] and Klein and Spady (1993) [[12]]. These models allow to leave the link function and the error term distribution unspecified, while achieving root-n consistency. We present key properties, such as root-n consistency, asymptotic normality and asymptotic efficiency.

We implement a brief simulation study to analyze finite sample properties of the proposed estimators. The findings suggest that the finite sample bias decreases for increasing sample size. Klein and Spady’s (1993) [[12]] model exhibits a smaller finite sample bias. The semiparametric estimators weakly outperform a standard parametric logistic estimator, yet require significantly more computational effort.

This result is confirmed in a brief empirical study using a gender voice recognition data set. While the semiparametric estimators achieve weakly higher in-sample and out-of-sample accuracy, computational issues remain.

10 Bibliography

11 Appendix

11.1 Empirical Definition of Trimming Function

Section 7 develops an empirical method to implement the trimming functions for Ichimura's (1993) [[6]] and the KS (1993) [[12]] models.

Referring to theoretical definitions, the aim of a trimming function in Ichimura's (1993) [[6]] model is to guarantee $\hat{p}_{-i}(X'_i\beta)$ is not too close to zero.

The estimate $\hat{p}_{-i}(X'_i\beta)$ is calculated over a given bandwidth h . There are two ways to keep such an estimate far away enough from zero: either we employ A_δ directly, which can be quite stringent, or we use observations that are close enough to the independent variables that satisfy $x \in A_\delta$. The latter also leads $\hat{p}_{-i}(X'_i\beta)$, the kernel density estimator for $X'_i\beta$, to be large enough. We use a similar reasoning to A_n for the primary trimming function.

Before starting the estimation, simulated sample data is sieved by examining whether the single index's NW leave-one-out kernel estimates are large enough. A lower bound is defined, such that $\hat{p}_{-i}(x'\beta)$ is not too close to zero and also so that enough observations are left for the estimation to be reliable. If $\hat{p}_{-i}(x'\beta)$ fails to achieve the chosen lower bound, this data point will be removed from the dataset. Furthermore, as the grid search method is used to carry out the estimation, the same set of grids is used to define \mathcal{B} , as defined in section 4. Data sieving is carried out over this set such that the estimation procedure is sensible for each single grid.

Firstly, the same trimmed dataset is used for Ichimura's (1993) [[6]] and the KS (1993) [[12]] estimation methods. However, this is not enough for the latter model to work, as it only prevents $g(\cdot)$ from growing out of bounds. The KS (1993) [[12]] model is more restrictive than Ichimura's (1993) [[6]] model on this point, as it requires any estimate of \hat{G}_i not to be too close to 0 or 1. Recall that the numerator follows the sum of dependent binary variables y_i weighted by its corresponding kernel evaluation. This issue becomes apparent for relatively large evaluations. Therefore, we further introduce a lower bound equal to square root of machine double epsilon and set any estimate smaller than the lower bound to this value.

11.2 A Comparison in Results of self implemented functions and pre-implemented np-package

Model design of Monte Carlo simulation following Ichimura (1993):

- sample size is 250;
- the SSIM is specified as binary choice model;
- the model includes two exogenous variables, both are independently generated from standard normal distribution
- when constructing the single index, coefficient of the first exogenous variable is normalized to 1, and the true value for the coefficient of the second exogenous variable is -2;
- standard normal distribution for error distribution;

- the amount of Monte carlo simulation is 1000 times.

To summarize, the simulation is based on the following model setup

$$y_i = I(x_{1i} - 2x_{2i} > \epsilon_i).$$

The *Table 4* presents a comparison between the pre-implemented np-package and our own implementation using grid search.

Method	np-package	self-implementation
(standard normal distribution for x_1, x_2, ϵ)		
Ichimura	0.00363	0.00201
Klein and Spady	0.00416	0.0000893

Table 4: Mean Squared Error Comparison

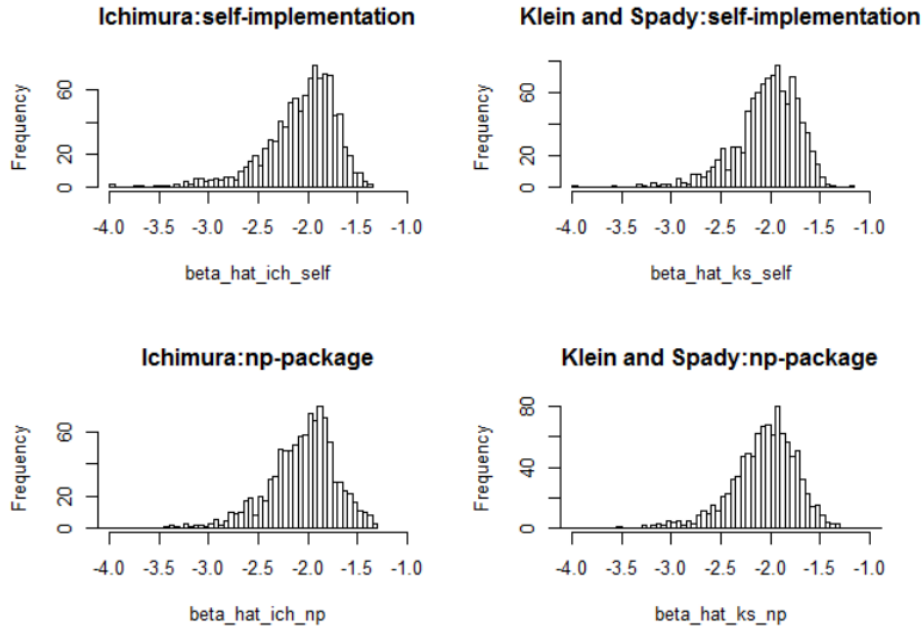


Figure 3: Comparison of Estimates on the Same Scale

From *Table 4* and *Figure 3*, we see that our self-implemented code does not perform significantly worse than the np-package [[28]]. In fact it is even better for the KS (1993) [[12]] method. This could be due to the use of our *a-priori* knowledge of the true value of the coefficients when specifying the grid. We thus conclude that it is an useful and successful practice in constructing our own code for single index model functions.

11.3 Analysis of consistency

These are the graphs that follow the analysis of finite sample bias on section 7. For the three models, the variance of the estimates decreases with the sample

size. Moreover, the histograms resemble more and more closely the distribution of normal variables.

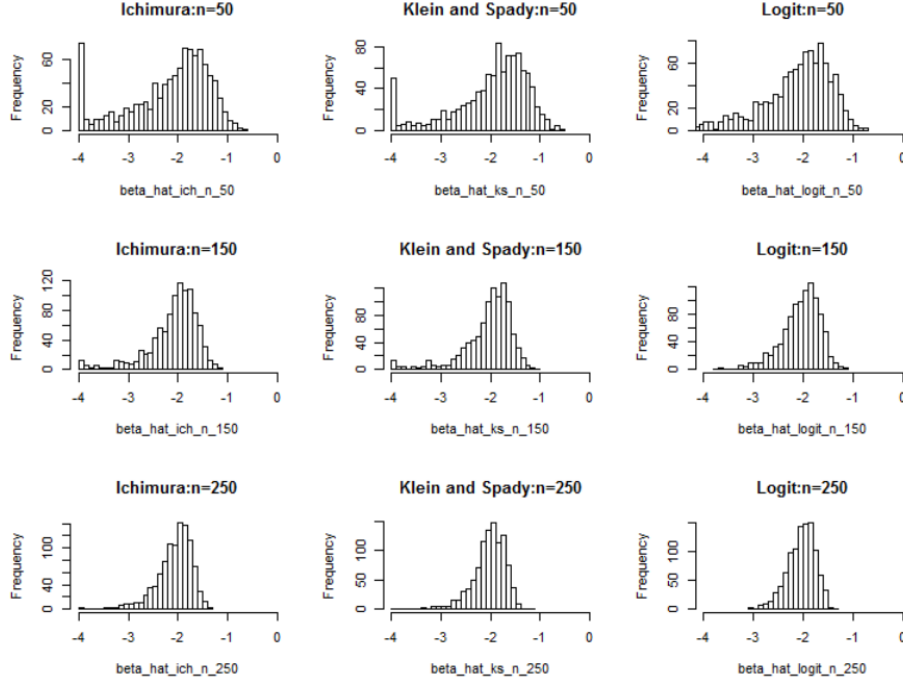


Figure 4: Plot of estimates for different sample sizes

11.4 Real Dataset Variable Explanation

The following list shows the 21 independent variables and 1 binary gender indicator contained in the dataset.

Variable Name	Description
meanfreq	mean frequency (in kHz)
sd	standard deviation of frequency
median	median frequency (in kHz)
Q25	first quantile (in kHz)
Q75	third quantile (in kHz)
IQR	interquartile range (in kHz)
skew	skewness
kurt	kurtosis
sp.ent	spectral entropy
sfm	spectral flatness
mode	mode frequency
centroid	frequency centroid
peakf	peak frequency (frequency with highest energy)
meanfun	average of fundamental frequency measured across acoustic signal
minfun	minimum fundamental frequency measured across acoustic signal
maxfun	maximum fundamental frequency measured across acoustic signal
meandom	average of dominant frequency measured across acoustic signal
mindom	minimum of dominant frequency measured across acoustic signal
maxdom	maximum of dominant frequency measured across acoustic signal
dfrange	range of dominant frequency measured across acoustic signal
modindx	modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range
label	male or female