<table>
<thead>
<tr>
<th><strong>Statistica Sinica Preprint No:</strong> SS-2017-0451</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
<tr>
<td><strong>Complete List of Authors</strong></td>
</tr>
<tr>
<td><strong>Corresponding Author</strong></td>
</tr>
<tr>
<td><strong>E-mail</strong></td>
</tr>
<tr>
<td><strong>Notice:</strong> Accepted version subject to English editing.</td>
</tr>
</tbody>
</table>
ESTIMATION OF SINGLE-INDEX MODELS WITH FIXED CENSORED RESPONSES

Hailin Huang\textsuperscript{1*}, Yuanzhang Li\textsuperscript{1}, Hua Liang\textsuperscript{1}, Yanlin Tang\textsuperscript{2*},

\textsuperscript{1}George Washington University and \textsuperscript{2}Tongji University

Abstract: We propose a new procedure to estimate the index parameter and link function of single-index models, where the response variable is subject to fixed censoring. Under some regularity conditions, we show that the estimated index parameter is root-$n$ consistent and asymptotically normal, and the estimated nonparametric link function achieves the optimal convergence rate and is asymptotically normal. In addition, we propose a linearity testing method for the nonparametric link function. The simulation study shows that the proposed procedures perform well in finite sample experiments. An application to an HIV dataset is presented for illustration.

Key words and phrases: Nonparametric censored regression, single-index model, semi-parametric least-squares.

1. Introduction

Because of non-negativity or detection limit, data with fixed censored

*The two authors contribute equally to the paper.
responses are common in econometrics and biometrics studies (Maddala, 1986; Adesina and Zinnah, 1993; Nizar Al-Malkawi, 2007; Haab, Dunham and Brown, 2001; Van der Pouw Kraan et al., 1995). For instance, in our motivating HIV data set, the viral load in blood serum can only be observed if it is above 50 units (Kobie et al., 2012).

To explore the relationship between the fixed censored response variable and the covariates, several models and associated estimation methods have been proposed. Earlier work focused on parametric regression models, and representative work includes the Tobit model (Tobin, 1958) and its variants (Amemiya, 1984, 1979; Blundell and Meghir, 1987), which assume a linear relationship with normal errors. However, both linearity and normality assumptions can be violated in practice (Maddala and Nelson, 1975; Gawande, 1995; Chen, Dahl and Khan, 2005). To make it more flexible, several researchers studied nonparametric regression models with fixed censored data. For example, Lewbel and Linton (2002) proposed a two-stage moment based method to estimate the nonparametric conditional mean function; Chen, Dahl and Khan (2005) studied the identification and estimation problems of the conditional median function in nonparametric location-scale models. These nonparametric methods achieve more flexibility and generally do not require distributional assumptions, but they suffer
from “curse of dimensionality”, and their performance can be poor even when the dimension of the covariates is moderate.

To amend the limitations of the existing methods aforementioned, we consider the single-index models along with the fixed censored responses. Single-index models have been widely studied in the literature (Powell, Stock and Stoker, 1989; Duan and Li, 1991; Härdle, Hall and Ichimura, 1993; Ichimura, 1993; Horowitz and Härdle, 1996; Carroll et al., 1997; Xia and Härdle, 2006; Liang et al., 2010). However, the majority of the existing literature is devoted to the case where the response $Y$ is fully observed, though some researchers studied the estimation when $Y$ is randomly censored (Lopez, 2009; Bücher, El Ghouch and Van Keilegom, 2014; Chiang, Wang and Huang, 2017; Kong and Xia, 2017). It is worth pointing out that, the methods for single-index models with randomly censored responses, implicitly assume that we can always observe uncensored observations below any given value of the censoring point (Lopez, 2009; Bücher, El Ghouch and Van Keilegom, 2014; Chiang, Wang and Huang, 2017; Kong and Xia, 2017), which is not true for fixed censoring case since the probability of observing uncensored observations below the given fixed censored point is zero, and thus the associated methods cannot be applied. To the best of our knowledge, no estimation methods for single-index models with fixed censored responses is available.
In this paper, by establishing a relationship between the fixed censored single-index models and uncensored single-index models, we propose a new procedure to estimate the index parameter. Under certain regularity conditions, the proposed estimator is root-$n$ consistent and asymptotically normal. After plugging in the index parameter, the single-index model is simplified to a univariate fixed-censored nonparametric model, and we apply the method of Lewbel and Linton (2002) to estimate the nonparametric link function. The estimated nonparametric link function achieves the optimal convergence rate and is asymptotically normal. Finally, a hypothesis testing procedure is proposed to check the linearity of the nonparametric link function.

The rest of the paper is organized as follows. Section 2 presents the model, and gives the estimation and testing procedures, and Section 3 presents the asymptotic properties. Section 4 explores the finite sample performance through simulation studies, and an HIV dataset is analyzed in Section 5 for illustration. Technical proofs are referred to the online Supplementary Materials.
2. Model and Methods

2.1 Model

Consider the following single-index model for the latent responses,

\[ Y_i^* = m(X_i^\top \beta) - \epsilon_i, \quad i = 1, \ldots, n, \quad (2.1) \]

where \( X_i = (X_{i,1}, \ldots, X_{i,d})^\top \) is a \( d \)-dimensional covariate vector, \( \beta = (\beta_1, \ldots, \beta_d)^\top \) is an unknown index parameter vector, \( m(u) = E(Y_i^*|X_i^\top \beta = u) \) is an unknown smooth function, and \( \epsilon_i \) is the random error. Due to fixed censoring, \( Y_i^* \) cannot be fully observed, instead we can only observe \((Y_i, \delta_i)\), where \( Y_i = \max(Y_i^*, c) \), \( \delta_i = I(Y_i^* > c) \), \( c \) is the known lower detection limit, \( I(\cdot) \) is the indicator function. Without loss of generality, we assume \( c = 0 \). Instead of making parametric distribution assumptions such as normality, here we only assume that \( \epsilon_i \) are independently and identically distributed (i.i.d.), from an unknown distribution symmetric around 0, with finite variance. Furthermore, we assume that no intercept is included in the index function \( X_i^\top \beta \), \( \|\beta\| = 1 \) and the first element of \( \beta \) is positive, to ensure identification, where \( \|\cdot\| \) denotes the \( L_2 \) norm. In addition, we assume that \( \beta \in \Theta \subset \mathbb{R}^d \) for some compact set \( \Theta \), and \( X \in D_X \subset \mathbb{R}^d \) for some compact set \( D_X \).

Remark 1. To facilitate theoretical derivations, we consider our model with
error term “− ϵ_i” instead of “ϵ_i”, and similar model setting can be found in [Lewbel and Linton (2002)]. It is worthy noting that with the symmetry assumption on ϵ_i around 0, ϵ_i and −ϵ_i have the same distribution.

2.2 Profile least-squares estimator of β

Under model (2.1), the proposed estimation procedure for β is inspired by considering a connection between fixed censored single-index models and uncensored single-index models. Under mild assumptions, this connection transfers the estimation of single-index Tobit models to standard single-index models, and well-developed estimation procedures can be applied.

Assumption A.1 (i) The latent response Y* has first ν(≥ 3) absolute moments. (ii) The common density function of ϵ_i, denoted as f(·), is symmetric around zero and its derivative is continuous.

Proposition 1. Let F(·) be the distribution function of ϵ. Under Assumption A.1, if \( \lim_{\epsilon \to -\infty} \epsilon F(\epsilon) = 0 \), then \( E(Y_i | X_i^\top \beta) = \int_{-\infty}^{m(X_i^\top \beta)} F(\epsilon) d\epsilon \). 

Assumption A.1 and the assumption \( \lim_{\epsilon \to -\infty} \epsilon F(\epsilon) = 0 \) are mild, and most commonly used symmetric distributions, such as normal distribution, Student t_v distribution (v ≥ 4), uniform distribution on a symmetric interval, satisfy these assumptions [Lewbel and Linton 2002]. Proposition 1 implies that \( E(Y_i | X_i^\top \beta) \) can be represented as a new uncensored single-index
2.2 Profile least-squares estimator of $\beta$

model with the same index parameter $\beta$, but a new link function. More specifically, $E(Y|X^T_i \beta = u) = r(u) = w \circ m(u)$, where $w(t) = \int_{-\infty}^{t} F(\epsilon)d\epsilon$, and ‘$\circ$’ means composition of two functions; similar derivation can be found in [Lewbel and Linton (2002)].

According to Proposition 1, we can assign a new single-index model for the observed responses as

$$Y_i = r(X^T_i \beta) - \epsilon'_i, \ i = 1, 2, \ldots, n,$$ \hspace{1cm} (2.2)

where $\epsilon'_i = \epsilon_i + (Y^*_i - Y_i) + r(X^T_i \beta) - m(X^T_i \beta)$. By Proposition 1, $E(Y_i|X^T_i \beta) = r(X^T_i \beta)$, thus we have $E(\epsilon'_i|X^T_i \beta) = 0$. Therefore, existing estimation methods for single-index models could be applied to estimate $\beta$. Here we adopt the profile least-squares method in [Liang et al. (2010)] as follows. Given $\beta$, we employ the local linear regression technique to estimate $r(\cdot)$, i.e.,

$$\min_{a, b} \sum_{i=1}^{n} \{a + b(X^T_i \beta - u) - Y_i\}^2 K_h(X^T_i \beta - u)$$ \hspace{1cm} (2.3)

with respect to $a$ and $b$, where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot) \geq 0$ is a kernel function and $h > 0$ is the bandwidth. Let $(\hat{a}, \hat{b})$ be the minimizer of (2.3), then $\hat{r}(u) = \hat{a}$. As discussed in [Jennrich (1969)], there exists a profile least-squares estimator $\hat{\beta}$ by minimizing

$$Q(\beta) = \sum_{i=1}^{n} \{Y_i - \hat{r}(X^T_i \beta)\}^2$$
2.3 Nonparametric estimation of $m(\cdot)$

with respect to $\beta$, where the minimization problem can be solved by standard optimization algorithms such as Newton-Raphson algorithm, and convergence is guaranteed.

Remark 2. The estimation procedure above treats all the covariates as important ones. In practice, especially when the dimension of $X$ is high, it is quite possible that irrelevant covariates are included. This may motivate us to consider variable selection. Given the expression (2.2), any variable selection method for single-index models is applicable to achieve the purpose of variable selection, for example the penalized profile least-squares method in [Liang et al. (2010)]. A detailed discussion can be found in [Huang (2017)].

2.3 Nonparametric estimation of $m(\cdot)$

Given $\hat{\beta}$, we can estimate the unknown link function $m(\cdot)$. For notational ease, we rewrite model (2.1) as

$$Y_i^* = m(U_i) - \epsilon_i, \quad i = 1, \ldots, n,$$

(2.4)

where $U_i = X_i^\top \beta$. Recall the definition of $r(\cdot)$ in (2.2), $r(u) = E(Y|U = u) = E(Y|X^\top \beta = u)$, and define $s = r(u)$, $q(s) = q(r(u)) = P\{Y > 0|U = u\} = P\{Y > 0|U = u\}$ and $\hat{U}_i = X_i^\top \hat{\beta}$. We propose to estimate $m(\cdot)$ similarly as [Lewbel and Linton (2002)].
2.4 Testing the linearity of the link function

Step 1. Smooth the observed response $Y_i$ over $\hat{U}_i$ to estimate $r(\hat{U}_i)$ by the local linear smoother [Fan and Gijbels 1996]; i.e.,

$$(\hat{a}_{i,0}, \hat{a}_{i,1}) = \arg \min_{(a_0, a_1) \in \mathbb{R}^2} \sum_{j=1}^{n} \{Y_j - a_0 - a_1(\hat{U}_j - \hat{U}_i)\}^2 K_{h_1}(\hat{U}_j - \hat{U}_i), \quad (2.5)$$

and then $r(\hat{U}_i)$ is estimated as $\hat{r}(\hat{U}_i) = \hat{a}_{i,0}$, where $h_1 > 0$ is a bandwidth.

Step 2. Smooth $I(Y_i > 0)$ over $\hat{r}(\hat{U}_i)$ to estimate $q(\cdot)$ by local linear smoother; i.e.,

$$(\hat{b}_0, \hat{b}_1) = \arg \min_{(b_0, b_1) \in \mathbb{R}^2} \sum_{i=1}^{n} [I(Y_i > 0) - b_0 - b_1\{\hat{r}(\hat{U}_i) - \hat{r}(u)\}]^2 K_{h_2}(\hat{r}(\hat{U}_i) - \hat{r}(u)),$$

and then $q(\hat{r}(u))$ is estimated as $\hat{q}(\hat{r}(u)) = \hat{b}_0$, where $\hat{r}(u)$ is estimated by replacing $\hat{U}_i$ with $u$ in (2.5), and $h_2 > 0$ is a bandwidth.

Step 3. Estimate $m(u)$ by $\hat{m}(u) = \hat{\lambda}_r - \int_{\hat{r}(u)}^{\hat{\lambda}_r} \frac{1}{\hat{q}(s)} ds$, where $\hat{\lambda}_r = \max_{i=1,\ldots,n} \hat{r}(X_i^\top \hat{\beta})$. For the integration part, any one-dimensional numerical integration approach, such as Trapezoid rule, can be employed.

2.4 Testing the linearity of the link function

In practice, one may be interested in whether $m(\cdot)$ is a linear function so that the single-index model can be simplified to a linear model. In this
2.4 Testing the linearity of the link function

In this section, we study the hypothesis

\[ H_0 : m(u) = \zeta_0 + \zeta_1 u \text{ versus } H_1 : H_0 \text{ is not true.} \]

To test the linearity of \( m(\cdot) \), we further assume \( \epsilon_i \sim N(0, \sigma^2) \), where \( \sigma \) is an unknown scale parameter.

Recalling Proposition 1, we have

\[ w'(u) = \frac{\partial r(u)}{\partial m(u)} = F(m(u)) = \Phi(m(u)/\sigma) > 0, \]

which indicates that \( r(\cdot) \) is a strictly increasing function of \( m(\cdot) \). As a result, testing \( H_0 \) against \( H_1 \) is equivalent to

\[ K_0 : r_0(u) = \int_{-\infty}^{\zeta_0 + \zeta_1 u} \Phi(\epsilon/\sigma) d\epsilon \text{ versus } K_1 : K_0 \text{ is not true.} \]

We adopt the idea of Koul, Song and Liu (2014) to test \( K_0 \) versus \( K_1 \).

Given a root-\( n \) consistent estimator of \( \beta_0 \), say \( \hat{\beta} \), we define

\[ \tilde{\epsilon}_i = Y_i - \int_{-\infty}^{\zeta_0 + \zeta_1 X_i^\top \hat{\beta}} \Phi(\epsilon/\sigma) d\epsilon, \]

where \( \zeta_0, \zeta_1 \) and \( \sigma \) are estimated by the maximum likelihood method in the Tobit model (Tobin, 1958; Amemiya, 1984). Define

\[
V_n = \frac{1}{n(n-1)h} \sum_{i \neq j} K \left( \frac{X_i^\top \hat{\beta} - X_j^\top \hat{\beta}}{h} \right) \tilde{\epsilon}_i \tilde{\epsilon}_j,
\]

\[
\hat{\gamma}^2 = \frac{2}{n(n-1)h} \sum_{i \neq j} K^2 \left( \frac{X_i^\top \hat{\beta} - X_j^\top \hat{\beta}}{h} \right) \tilde{\epsilon}_i^2 \tilde{\epsilon}_j^2,
\]
where $h > 0$ is a bandwidth the same as that for the profile least squares estimator of the index parameter, specified in equation (2.3). The test statistic is then defined as

$$T_n = n h^{1/2} V_n / \hat{\gamma}.$$  

Under certain regularity conditions, we can prove that $T_n$ is asymptotically normal under the null hypothesis. Thus, a large value of $T_n$ indicates deviation from a Tobit model.

3. Asymptotic Properties

In this section, we present the asymptotic properties of the proposed estimators for the index parameter and the link function, and also property of the test statistic. The true index parameter and unknown link function are denoted as $\beta_0 = (\beta_{1,0}, \ldots, \beta_{d,0})^\top$ and $m(\cdot)$. Besides Assumption A.1, the following assumptions are needed for the asymptotic results.

**Assumption A.2.** (i) $r(\cdot)$ and $m(\cdot)$ are not constant on the support $\Omega = \{ u | u = x^\top \beta, x \in D_X, \beta \in \Theta \}$, and their third derivatives are uniformly Lipschitz continuous for all $u \in \Omega$. (ii) Let $f_X(x)$ be the density function of $X$, and the third derivative of $f_X(x)$ is continuous. (iii) The second derivative of the function $q(\cdot)$ is continuous, and $\inf_{u \in \Omega} q(r(u)) > 0$, and $q(\lambda_r) = 1$ where $\lambda_r = \sup_{u \in \Omega} r(u)$, where the supremum is taken over
\[ u = x^\top \beta_0, \ x \in D_X. \] (iv) \( r^2(u) = E[\{Y - r(X^\top \beta_0)\}^2|X^\top \beta_0 = u] \) and \( v^4(u) = E[\{Y - r(X^\top \beta_0)\}^4|X^\top \beta_0 = u] \) are bounded functions, and they have continuous derivatives.

**Assumption A.3.** (i) \( nh^8 \to 0 \) and \( nh^{3+3/(\nu - 1)}/\log n \to \infty \), as \( n \to \infty \), where \( \nu \geq 3 \) is specified in A.1. (ii) \( nh_2^2/\log^2(n) \to \infty \), \( nh_2^3/\log^2(n) \to \infty \) and \( h_1/h_2 \leq C_1, nh_1^5 \leq C_2, nh_2^5 \leq C_3 \) for some positive constants \( C_1, C_2, C_3 \).

**Assumption A.4.** The support of the kernel function \( K(\cdot) \) is \([-1, 1]\), and its second derivative is Lipschitz continuous. Moreover, \( \int_{-1}^{1} K(s)ds = 1; \int_{-1}^{1} sK(s)ds = 0; \int_{-1}^{1} s^2K(s)ds > 0. \)

Assumptions A.2 (i)-(ii) are similar to the regularity conditions in Carroll et al. (1997); Liang et al. (2010) for uncensored data. A.2 (iii) is adopted from Assumption 2 in Lewbel and Linton (2002), which is necessary to ensure that the estimated nonparametric link function achieves the optimal convergence rate. Assumption A.2 (iv) is adopted from Assumption (C2) in Koul, Song and Liu (2014), which is a necessary condition for the asymptotic normality of the test statistic. Assumption A.3 provides us the guideline to select appropriate bandwidths, and as pointed out by Liang et al. (2010), Assumption A.3 (i) implies that the estimation performance remains stable in a reasonable range of bandwidth, especially when the sample size is large.

In practice, the bandwidth can be chosen by cross-validation. Assumptions
A.4 is standard for nonparametric regression.

The following Theorems 1-2 present the asymptotic properties of the estimated index parameter and nonparametric link function.

**Theorem 1.** Under Assumptions A.1–A.4, we have

\[
\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, W_0^+),
\]

where \( W_0 = E\left[\sigma^2(X^\top \beta_0)\{X - E(X|X^\top \beta_0)\}\{X - E(X|X^\top \beta_0)\}^\top \sigma^2(X^\top \beta_0)\right], \)
and \( W_0^+ \) denotes its Moore-Penrose inverse.

**Theorem 2.** Under Assumptions A.1–A.4, for an interior point \( u = x^\top \beta, \)
where \( x \in D_X \) and \( \beta \in \Theta_{c_0} = \{\beta : ||\beta - \beta_0|| \leq c_0 n^{-1/2}\} \) for some positive constant \( c_0, \) we have

\[
\sqrt{nh_1}\left\{\hat{m}(u) - m(u) - k_0 - b_m(u)h_1^2\right\} \xrightarrow{D} N\left\{0, \frac{1}{s_0^2(u)}\sigma_u^2\right\}.
\]

Here, \( \sigma_u^2 = \tau^2(u)f_U^{-1}(u)\int_{-1}^1 K^2(t)dt, \) with \( f_U(\cdot) \) being the density function of \( U = X^\top \beta_0; \) \( k_0 = \lambda_r - F_1^{-1}(\lambda_r), \) with \( F_1(\lambda_r) = \int_{-\infty}^{\lambda_r} F(\epsilon)d\epsilon; \) \( s_0(u) = q(r(u)); \) \( b_m(\cdot) \) is a bounded continuous function that is determined by terms \( T_2 \) and \( T_6 \) in the online Supplementary Materials. If we further assume that \( \sup_{\epsilon \in \Omega_\epsilon} \epsilon \leq \lambda_r, \) where \( \Omega_\epsilon \) is the domain of \( \epsilon, \) then the term \( k_0 \) disappears and we have

\[
\sqrt{nh_1}\left\{\hat{m}(u) - m(u) - \frac{1}{2} b_m(u)h_1^2\right\} \xrightarrow{D} N\left\{0, \frac{1}{s_0^2(u)}\sigma_u^2\right\}.
\]
Theorem 1 shows that the estimator \( \hat{\beta} \) is \( \sqrt{n} \)-consistent and asymptotically normal. Theorem 2 indicates that, up to a location constant, the proposed nonparametric estimator achieves the optimal convergence rate. Furthermore, it is worth pointing out that although \( k_0 \) is theoretically nonzero, it is numerically negligible in many situations based on our numerical experiences. Theorem 2 theoretically further justifies that the location shift \( k_0 \) disappears with slightly stronger assumptions.

**Remark 3.** If we want to construct confidence intervals for \( \beta_0 \) and \( m(\cdot) \), it may be required to estimate the asymptotic variances involved in Theorems 1-2. The weighting function \( \tau(\cdot) \) and asymptotic covariance matrix \( W_0^{+} \) of \( \hat{\beta} \) can be estimated by typical variance estimation methods for heterogeneous single-index models (Ichimura, 1993; Härdle, Hall and Ichimura, 1993; Chiou and Müller, 1998, 1999). The asymptotic variance \( \sigma^2 / s_0^2(u) \) of the link function estimator can be obtained by replacing \( f_U^{-1}(\cdot) \) and \( s_0(\cdot) \) by their consistent estimators (Lewbel and Linton, 2002). Considering the potential complexity in estimation of the variances, the bootstrap method is a good alternative to construct confidence intervals for \( \beta_0 \) and \( m(\cdot) \).

Lastly, we state the asymptotic properties of the proposed test. We need two additional assumptions.

**Assumption A.5.** The random noise \( \epsilon_i \sim N(0, \sigma^2) \), where \( \sigma \in \Omega_\sigma \) is an
unknown parameter.

**Assumption A.6.** For any given $\beta \in \Theta$, and any root-$n$ consistent estimator $\hat{\sigma}$ of $\sigma$, $\sup_{(x,\sigma) \in D_X \times \Omega, \sigma} |r(x^T \beta, \hat{\sigma}) - r(x^T \beta, \sigma) - (\hat{\sigma} - \sigma)r'(x^T \beta, \sigma)| = O_p(1/n)$, where $r(x^T \beta, \sigma) = \int_{-\infty}^{m(x^T \beta)} \Phi(\epsilon/\sigma) d\epsilon$.

Assumption A.6 is adapted from Assumption (C.4) of Koul, Song and Liu (2014). We have the following result for $T_n$.

**Theorem 3.** Assume Assumptions A.1–A.6 hold, then under $H_0$,

$$T_n = nh^{1/2}V_n/\hat{\gamma} \xrightarrow{D} N(0,1).$$

### 4. Simulation Studies

In this section, we investigate the finite sample performance of the proposed estimation and testing methods by Monte Carlo simulations. Examples 4.1 and 4.2 focus on the estimation of $\beta_0$ and $m(\cdot)$, respectively, and Example 4.3 studies the performance of $T_n$.

**Example 4.1.** In this example, we focus on the estimation of $\beta_0$. We generate 100 replicates from the following two models,

$$Y_{i}^* = e^{(X_{i1}+X_{i2})/\sqrt{2}} - \epsilon_i, \ i = 1, \ldots, n, \quad (4.4)$$

and

$$Y_{i}^* = \sin \left( \pi \{(X_{i1} + X_{i2})/\sqrt{2}\}/(b - a) \right) - \epsilon_i, \ i = 1, \ldots, n, \quad (4.5)$$
where $X_{i1}, X_{i2}$ are i.i.d. from $Uniform(0, 1)$, $\epsilon_i$ is from either $N(0, 0.1^2)$ or Laplace distribution $\mathcal{L}(0, 0.1^2)$, and $a = \sqrt{2}/2$, $b = \sqrt{3}/2 + 1.645/\sqrt{12}$. In both (4.4) and (4.5), the true index parameter is $\beta_0 = (\beta_{01}, \beta_{02})^T = (0.701, 0.701)^T$. The observed responses $Y_i$ are set as $Y_i = \max(Y_i^*, c)$, where $c$ is properly chosen to result two censoring proportions (Cen), Cen=20% and Cen=40%. We consider two sample sizes, $n = 200$ and 400.

Since no other existing estimation method is available for such models, we compare our estimator to the omniscient profile least-squares estimator based on the latent data $Y_i^*$ (corresponding to Cen=0). The performance is evaluated by the $L_2$ difference $||\beta_0 - \hat{\beta}||_2$ across replicates. We select the bandwidth $h$ by grid search to minimize simulation based estimates of the $L_2$ differences, which follows the same manner as in [Liang et al. (2010)].

The average CPU time for each replicate is 28 seconds for $n = 200$ and 101 seconds for $n = 400$, running on an Intel(R) Core(TM) i7-6700HQ CPU with 2.60GHz. Table 1 summarizes the averaged estimates (AVE) of $\beta_0$ and the corresponding MSE. From Table 1, we find that the biases based on $Y_i$ are comparable to those from $Y_i^*$, while the MSE based on $Y_i$ are larger, but are still within reasonable ranges.

**Example 4.2.** In this example, we focus on the estimation of $m(\cdot)$. We generate 200 replicates, and each replicate consists of $n = 400$ observations,
Table 1: Example 4.1, average estimates (AVE) and MSE×10^4 of the index parameter.

<table>
<thead>
<tr>
<th>n</th>
<th>Cen</th>
<th>Model (4.4)</th>
<th>Model (4.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AVE</td>
<td>MSE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>β₁</td>
<td>β₂</td>
</tr>
</tbody>
</table>

$\epsilon_i \sim N(0, 0.1^2)$

<table>
<thead>
<tr>
<th>n</th>
<th>Cen</th>
<th>AVE</th>
<th>MSE</th>
<th>AVE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0%</td>
<td>0.7067</td>
<td>0.7080</td>
<td>4.84</td>
<td>5.07</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>0.7065</td>
<td>0.7076</td>
<td>5.16</td>
<td>4.51</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>0.7080</td>
<td>0.7061</td>
<td>7.91</td>
<td>7.43</td>
</tr>
<tr>
<td>400</td>
<td>0%</td>
<td>0.7078</td>
<td>0.7064</td>
<td>1.16</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>0.7067</td>
<td>0.7060</td>
<td>1.63</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>0.7056</td>
<td>0.7070</td>
<td>1.64</td>
<td>1.67</td>
</tr>
</tbody>
</table>

$\epsilon_i \sim \mathcal{L}(0, 0.1^2)$

<table>
<thead>
<tr>
<th>n</th>
<th>Cen</th>
<th>AVE</th>
<th>MSE</th>
<th>AVE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0%</td>
<td>0.7067</td>
<td>0.7074</td>
<td>4.12</td>
<td>3.67</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>0.7075</td>
<td>0.7068</td>
<td>5.19</td>
<td>4.47</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>0.7071</td>
<td>0.7061</td>
<td>6.27</td>
<td>6.34</td>
</tr>
<tr>
<td>400</td>
<td>0%</td>
<td>0.7070</td>
<td>0.7072</td>
<td>1.27</td>
<td>1.30</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>0.7069</td>
<td>0.7730</td>
<td>1.28</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>0.7071</td>
<td>0.7070</td>
<td>1.76</td>
<td>1.72</td>
</tr>
</tbody>
</table>
from models (4.4) and (4.5). We estimate $m(\cdot)$ at 400 grid points uniformly spaced within the range of $X^\top \beta_0$. The censoring point $c$ is set to result Cen=20%, which mimics our real HIV data in Section 5. To alleviate computational burden, the bandwidths for estimating the link function are chosen by the rule of thumb (Silverman, 1986), i.e., $h_1 = 1.06 s(X^\top \hat{\beta}) n^{-1/5}$ and $h_2 = 1.06 s(\hat{r}(X^\top \hat{\beta})) n^{-1/5}$, where $s(\cdot)$ denotes the sample standard deviation.

Figure 1 presents the point-wise median curve (solid line) of the estimated function $\hat{m}(u)$ on the selected grid, point-wise 5% and 95% quantiles (dotted line) of $\hat{m}(u)$, and the true $m(u)$ (dashed lines). The difference between the median curve and the true curve provides a measure of bias, while the 5% and 95% lines provide a measure of spread, which can be interpreted as simulation based point-wise confidence band. In general, regardless of normal errors or Laplace errors, the fitted curves are close to the true one, and the confidence bands covers the true curve except for a small region. Finally, as pointed out by Lewbel and Linton (2002) that the assumption that $\sup_{\epsilon \in \Omega} \epsilon \leq \sup_u r(u) = \lambda_r$ in Theorem 2 is not satisfied, a location shift may be expected. However, for these scenarios, $\int_{-\infty}^{\hat{\lambda}_r} \epsilon f(\epsilon) d\epsilon$ is almost zero and $F(\hat{\lambda}_r) = 1$ numerically, which implies that $\int_{-\infty}^{\hat{\lambda}_r} \epsilon f(\epsilon) d\epsilon = \hat{\lambda}_r - \int_{-\infty}^{\hat{\lambda}_r} F(\epsilon) d\epsilon = 0$, i.e. $\hat{\lambda}_r = F_1^{-1}(\hat{\lambda}_r)$, and therefore the
location bias can be ignored.

**Example 4.3.** In this example, we focus on the linearity test. We generate 200 replicates from the model

\[ Y^*_i = m \left( \frac{(X_{i1} + X_{i2})}{\sqrt{2}} \right) - \sigma \epsilon_i, \quad Y_i = \max(Y^*_i, c), \quad i = 1, \ldots, n, \]

where \( X_{i1}, X_{i2} \) are i.i.d. from \( N(0, 1) \), \( \epsilon_i \sim N(0, 1) \), \( \sigma \) equals either 0.1 or 0.25, \( n \) equals either 200 or 400, and \( c \) is chosen to lead to Cen=20%. The true \( m(\cdot) \) function is

\[ m(u) = u + c_2 \exp(u), \]

where \( c_2 \) ranges from 0 to 0.16 with increment 0.04, and \( c_2 = 0 \) corresponding the null hypothesis.

Table 2 summarizes the rejection rates for all cases given the nominal level 0.05. We see that, (i) under the null hypothesis, the empirical sizes are less than the nominal level, hence the proposed tests are conservative, which is common for nonparametric smoothing based tests (Zheng, 1996; Koul, Song and Liu, 2014); (ii) when the alternative is true, the power approaches to 1 quickly.
Figure 1: Simulation results for Model (8) and (9) with the normal error: fitted curves (dashed lines) and true curves (solid lines) with 90% confidence bands (dotted lines)

Figure 2: Simulation results for Model (8) and (9) with the Laplace error: fitted curves (dashed lines) and true curves (solid lines) with 90% confidence bands (dotted lines)
Table 2: Rejection rates for the linearity test of the link function when 
\( n = 200 \) or \( 400 \), and \( \sigma = 0.1 \) or \( \sigma = 0.25 \):

<table>
<thead>
<tr>
<th>( c_2 )</th>
<th>( \sigma = 0.1 )</th>
<th>( \sigma = 0.25 )</th>
<th>( \sigma = 0.1 )</th>
<th>( \sigma = 0.25 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>0.04</td>
<td>0.58</td>
<td>0.04</td>
<td>0.96</td>
<td>0.13</td>
</tr>
<tr>
<td>0.08</td>
<td>0.98</td>
<td>0.35</td>
<td>1</td>
<td>0.75</td>
</tr>
<tr>
<td>0.12</td>
<td>1</td>
<td>0.78</td>
<td>1</td>
<td>0.98</td>
</tr>
<tr>
<td>0.16</td>
<td>1</td>
<td>0.93</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

5. Analysis of an HIV Study

A primary goal of vaccine strategies to prevent HIV infection is the induction of a protective humoral response. Some HIV infected patients develop potent serum antibodies that are able to neutralize a broad range of HIV isolates. By studying the characteristics of the T cells in such HIV-infected patients, mechanisms for the induction of potent neutralizing antibodies may be revealed.

In this section, we apply the proposed methods to analyze a dataset from a study focusing on measuring T-cell-related parameters in such HIV
infected patients with varying degrees of HIV viral load. The dataset consists of observations of 414 patients with 4 variables: CD4, CD8, the difference of CD4 (diffcd4), and difference of CD8 (diffcd8). Due to detection limit, 20% of viral load values are left censored at 50 units. All the covariates are standardized to $[0,1]$, and log-transformation is applied to the response variable.

We first apply the linearity test for the link function, and the resulting p-value is 0.002, which suggests a strong evidence that the link function is nonlinear. So the proposed model would be more appropriate for this dataset. We then estimate the index parameter and the link function. The bandwidth for estimating $\beta$ is selected by 10-fold cross-validation, which yields $h_{\text{real}} = 0.14$, and the bandwidths for estimating the unknown link function are selected by rule of thumb as in the simulation study.

The estimated coefficients are 0.3970 (CD4), 0.0002 (CD8), 0.5919 (diffcd4), and $-0.7015$ (diffcd8). Figure 2 presents the estimated curve of the link function and the 90% point-wise confidence band at 50 grid points uniformly spaced between $[0,0.3]$. The figure indicates that the viral load shows a logarithmically descending trend with the composite single-index. Combining the index parameter signs and the descending trend of the link function, we find that CD4, CD8 and diffcd4 have negative effects, while
Figure 3: Fitted link function (solid line) and 90% confidence band (shaded area)

diffcd8 has a positive effect, on viral load, though the effect of CD8 is very small. These results are largely consistent with the conclusions in the scientific literature. For example, [Jiao et al. (2006)] discovered that there is a negative relation between CD4 and viral load.

Supplementary Materials

The online Supplementary Materials include the proofs of Proposition 1 and Theorems 1-3.

Acknowledgements

The authors would like to thank two reviewers, an associate editor
and the editor for constructive comments and helpful suggestions. This re-
search was partially supported by National Science Foundation grant DMS-
1418042 and DMS-1620898, National Natural Science Foundation of China 
grant, Award Number 11529101 (Liang), and the OSR-2015-CRG4-2582 
grant from KAUST, Shanghai Pujiang Program 18PJ1409800, and Key 
Laboratory for Applied Statistics of MOE, Northeast Normal University 
130028849 (Tang).

References

Adesina, A. A. and Zinnah, M. M. (1993). Technology characteristics, farmers’ perceptions and 
adoption decisions: A Tobit model application in sierra leone. *Agricultural Econom.* 9, 
297–311.


34, 179–200.


Kong, E. and Xia, Y. (2017). Uniform bahadur representation for nonparametric censored quan-


REFERENCES


Department of Statistics, George Washington University, Washington, D.C. 20052, USA

E-mail: hhl1988@email.gwu.edu

Department of Statistics, George Washington University, Washington, D.C. 20052, USA

E-mail: hliang@gwu.edu

Department of Statistics, George Washington University, Washington, D.C. 20052, USA

E-mail: yuanzhang.li@yahoo.com

School of Mathematical Sciences, Tongji University, Shanghai, 200092, China

E-mail: tang_y_l@hotmail.com