Oracle-efficient M-estimation for single-index models with a smooth simultaneous confidence band

Jiuzhou Miao

School of Statistics and Mathematics, Zhejiang Gongshang University

The 16th National Symposium on Survival Analysis and Applied Statistics

Joint work with Li Cai, Lei Jin and Suojin Wang

November 17, 2024

Single-index model

- Consider the following semiparametric model,

$$Y = g\left(\mathbf{X}^T \boldsymbol{\theta}_0\right) + \varepsilon, \tag{1}$$

  where $g(\cdot)$ is an unknown link function whose domain is (a, b) and $\boldsymbol{\theta}_0$ an unknown d-vector.
- Identifiability: Assume that $\|\boldsymbol{\theta}_0\| = 1$ and the first element of $\boldsymbol{\theta}_0$ is positive.
- Model (1) is called single-index model.

Simultaneous confidence band

- Simultaneous confidence band (SCB) is a powerful tool to investigate the global shape of unknown curves.
- It investigates the asymptotic distribution of the maximal deviation between an unknown function and its estimation.
- Recent works: Wang and Yang (2009), Cao et al. (2012), Ma et al. (2012), Wang et al. (2014), Cai and Yang (2015), Gu and Yang (2015), Wang et al. (2016), Cai et al. (2019), Cai et al. (2020), Cai et al. (2021), Li and Yang (2023), etc.

Estimation

- Let $(\mathbf{X}_i, Y_i)_{i=1}^n$ and $(\varepsilon_i)_{i=1}^n$ be i.i.d. copies of $(\mathbf{X}, Y, \varepsilon)$ from the single-index model (1) with $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d})^T$.
- If $\boldsymbol{\theta}_0$ is known by oracle, one can easily obtain a local M-type kernel estimator $\tilde{g}(x)$ for the link function $g(x)$ through minimizing

$$n^{-1} \sum_{i=1}^{n} K_h \left(\mathbf{X}_i^T \boldsymbol{\theta}_0 - x \right) \rho \left(Y_i - \lambda \right).$$

- In other words, $\tilde{g}(\cdot)$ is obtained by solving

$$n^{-1} \sum_{i=1}^{n} K_h \left(\mathbf{X}_i^T \boldsymbol{\theta}_0 - x \right) \psi \left(Y_i - \lambda \right) = 0. \tag{2}$$

Estimation

- However, one often has limited knowledge about $\boldsymbol{\theta}_0$. Thus, the estimator $\tilde{g}(x)$ becomes infeasible.
- In such cases, Replacing the unknown $\boldsymbol{\theta}_0$ in $\tilde{g}(x)$ in (2) with its $\sqrt{n}$-consistent estimator $\hat{\boldsymbol{\theta}}$, one can immediately derive the feasible M-estimator $\hat{g}(x)$ for $g(x)$.
- i.e., $\hat{g}(x)$ is obtained by solving

$$n^{-1} \sum_{i=1}^{n} K_h \left(\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - x \right) \psi \left(Y_i - \lambda \right) = 0. \tag{3}$$

Asymptotic properties

- Under general conditions, according to the asymptotic distribution of the maximal deviation in Härdle (1989), one can obtain that, for any closed subinterval $[a_0, b_0] \subseteq (a, b)$,

$$\gamma_h \left[\sqrt{nh} \sup_{x \in [a_0, b_0]} \left| \{\tilde{g}(x) - g(x)\} / v(x) \right| - d_h \right] \xrightarrow{d} Z. \quad (4)$$

- Z is a random variable with its cumulative distribution function being $P(Z \leq z) = \exp\{-2\exp(-z)\}$ for any $z \in \mathbf{R}$, and

$$v(x) = \Lambda_K^{1/2} f^{-1/2}(x) p_{0,2}^{1/2}(x) p_{1,1}^{-1}(x).$$

- Therefore, for any $\alpha \in (0,1)$, an asymptotic $100(1-\alpha)\%$ infeasible SCB for the link function $g(x)$, $x \in [a_0, b_0]$ is

$$\tilde{g}(x) \pm (nh)^{-1/2} v(x) \left[d_h - \gamma_h^{-1} \log\left\{-2^{-1}\log(1-\alpha)\right\}\right].$$

Oracle efficiency and asymptotic distribution of $\hat{g}(x)$

Theorem 1 (Oracle efficiency)

Under some mild assumptions, as $n \to \infty$, one has that

$$\sup_{x \in [a_0, b_0]} |\hat{g}(x) - \tilde{g}(x)| = O_p(n^{-1/2}).$$

Theorem 2 (Uniform asymptotic distribution of $\hat{g}(x)$)

Under some mild assumptions, as $n \to \infty$, one has that

$$\gamma_h \left[\sqrt{nh} \sup_{x \in [a_0, b_0]} |\{\hat{g}(x) - g(x)\} / v(x)| - d_h \right] \xrightarrow{d} Z,$$

where γ_h, d_h and $v(x)$ are same as that in (4).

A M-type feasible SCB for $g(x)$

- By Theorem 2, for any $\alpha \in (0,1)$, an asymptotic $100(1-\alpha)\%$ theoretical SCB for $g(x)$, $x \in [a_0, b_0]$ is

$$\hat{g}(x) \pm (nh)^{-1/2} v(x) \left[d_h - \gamma_h^{-1} \log \left\{ -2^{-1} \log(1-\alpha) \right\} \right].$$

- Unfortunately, the theoretical SCB above still relies on the unknown asymptotic standard error function $v(x)$.
- The pilot kernel density estimator $\hat{f}(x)$ is employed to estimate the density function $f(x)$:

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} K_{h_f} \left(\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - x \right),$$

where h_f is chosen using the rule-of-thumb provided in Silverman (1986) equation (3.31).

A M-type feasible SCB for $g(x)$

- Denote the residuals as $\hat{\varepsilon}_i = Y_i - \hat{g}(\mathbf{X}_i^T \hat{\boldsymbol{\theta}})$, where $1 \le i \le n$. For $k = 0, 1$ and $s = 1, 2$, we estimate the mean functions $p_{k,s}(x)$ using kernel regression:

$$\hat{p}_{k,s}(x) = \frac{\sum_{i=1}^{n} K_{h_p}(\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - x) \left\{ \psi^{(k)} \left(\hat{\varepsilon}_i \right) \right\}^s}{\sum_{i=1}^{n} K_{h_p}(\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - x)}.$$

- Here, h_p is an appropriate smoothing parameter. We then obtain the following estimator for $v(x)$:

$$\hat{v}(x) = \Lambda_K^{1/2} \hat{f}^{-1/2}(x) \hat{p}_{0,2}^{1/2}(x) \hat{p}_{1,1}^{-1}(x).$$

- We recommend using the optimal rule-of-thumb bandwidth given in Fan and Gijbels (1996) equation (4.3) to select h_p.

A M-type feasible SCB for $g(x)$

Theorem 3 (Uniform convergence of $\hat{v}(x)$)

Under some mild assumptions, as $n \to \infty$, one has that

$$\sup_{x \in [a_0, b_0]} \left| \hat{v}^2(x) - v^2(x) \right| = O_p(n^{-1/2} h^{-1/2} h_p^{-1} \log^{1/2} n).$$

Corollary 4 (A M-type feasible SCB for $g(x)$)

Under the assumptions of Theorem 3, for any $\alpha \in (0, 1)$, an asymptotic $100(1-\alpha)\%$ SCB for $g(x)$, $x \in [a_0, b_0]$, is given by

$$\hat{g}(x) \pm (nh)^{-1/2} \hat{v}(x) \left[d_h - \gamma_h^{-1} \log \left\{ -2^{-1} \log(1-\alpha) \right\} \right]. \quad (5)$$

Implementation

- The Huber-type loss function with its derivative

$$\psi(u) = \max\{-\kappa, \min\{u, \kappa\}\}$$

in the M-estimation (3) is employed.

- The cutoff constant κ in $\psi(u)$ is set as the absolute minimum value of the 10% and the 90% quantiles of the residuals

$$\hat{\varepsilon}_i = Y_i - \hat{g}^*(\mathbf{X}_i^T \hat{\boldsymbol{\theta}}), 1 \leq i \leq n,$$

where $\hat{g}^*(\cdot)$ is the Nadaraya-Watson kernel estimator of $g(\cdot)$.

Implementation

- $\sqrt{n}$-consistent estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$: Robust cubic B-spline method studied in Zou and Zhu (2014) with absolute loss.
- The domain of the link function $g(x)$:

$$(\hat{a}, \hat{b}) = (\min_{1 \leq i \leq n} \mathbf{X}_i^T \hat{\boldsymbol{\theta}}, \max_{1 \leq i \leq n} \mathbf{X}_i^T \hat{\boldsymbol{\theta}}).$$

  The sub-interval $[\hat{a}_0, \hat{b}_0]$ is determined by excluding 5% on both the left and right sides of the interval $(\hat{a}, \hat{b})$.
- Kernel function: Biweight kernel function

$$K(u) = 15(1 - u^2)^2 \mathrm{I}(|u| \leq 1)/16.$$

- Undersmoothing bandwidth h for $\hat{g}(x)$: $h = h_{\text{rot}} \log^{-1/2} n$.

Simulation

- M-SCB: The proposed M-type SCB in (5).
- L-SCB: The least square type SCB studied in Li et al. (2014) and Gu and Yang (2015).
- INF-SCB: The infeasible SCB.
- $1-\alpha = 0.95, 0.99$; $n = 200, 400, 600, 800$.
- The distributions of the error ε:

(i) Standard normal: $\varepsilon \sim N(0,1)$;

(ii) t distribution with degrees of freedom of 4: $\varepsilon \sim t(4)$;

(iii) Contaminated normal: $\varepsilon \sim 0.8N(0,1) + 0.2N(0,4^2)$;

(iv) Contaminated normal: $\varepsilon \sim 0.9N(0,1) + 0.1N(0,7^2)$.

Simulation

- Scenario 1:

$$g(x) = 0.5 \exp(x),\ \boldsymbol{\theta}_0 = (1,1,1,1)^T/2,$$

$\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})^T$ are i.i.d. from the 4-dimensional standard normal distribution with a mean vector of zeros and an identity covariance matrix.

- Scenario 2-4 are omitted here.

Empirical coverage percentages and average widths (in parentheses) of the SCBs for Scenario 1

n	$1-\alpha$	(i): $\varepsilon \sim N(0,1)$		
		M-SCB	L-SCB	INF-SCB
200	0.95	0.902 (1.474)	0.882 (1.414)	0.918 (1.488)
	0.99	0.974 (1.805)	0.972 (1.730)	0.982 (1.822)
400	0.95	0.910 (1.193)	0.930 (1.149)	0.946 (1.200)
	0.99	0.980 (1.447)	0.994 (1.392)	0.992 (1.454)
600	0.95	0.932 (1.051)	0.932 (1.014)	0.928 (1.054)
	0.99	0.988 (1.268)	0.990 (1.222)	0.988 (1.271)
800	0.95	0.944 (0.955)	0.944 (0.923)	0.956 (0.958)
	0.99	0.994 (1.148)	0.992 (1.109)	0.994 (1.151)

Empirical coverage percentages and average widths (in parentheses) of the SCBs for Scenario 1

n	$1-\alpha$	(ii): $\varepsilon \sim t(4)$		
		M-SCB	L-SCB	INF-SCB
200	0.95	0.920 (1.653)	0.902 (1.870)	0.926 (1.649)
	0.99	0.976 (2.035)	0.980 (2.300)	0.982 (2.031)
400	0.95	0.934 (1.298)	0.940 (1.494)	0.938 (1.301)
	0.99	0.982 (1.584)	0.988 (1.822)	0.990 (1.587)
600	0.95	0.966 (1.141)	0.962 (1.315)	0.958 (1.144)
	0.99	0.994 (1.384)	0.996 (1.594)	0.996 (1.388)
800	0.95	0.968 (1.034)	0.968 (1.200)	0.962 (1.037)
	0.99	0.998 (1.250)	1.000 (1.451)	0.996 (1.253)

Empirical coverage percentages and average widths (in parentheses) of the SCBs for Scenario 1

n	$1-\alpha$	(iii): $\varepsilon \sim 0.8N(0,1)+0.2N(0,4^2)$		
		M-SCB	L-SCB	INF-SCB
200	0.95	0.928 (1.809)	0.902 (2.514)	0.948 (1.795)
	0.99	0.974 (2.238)	0.984 (3.106)	0.990 (2.223)
400	0.95	0.944 (1.383)	0.948 (1.994)	0.966 (1.383)
	0.99	0.996 (1.698)	0.994 (2.446)	0.994 (1.698)
600	0.95	0.960 (1.200)	0.958 (1.735)	0.970 (1.199)
	0.99	0.996 (1.465)	0.986 (2.117)	0.996 (1.464)
800	0.95	0.962 (1.085)	0.954 (1.579)	0.974 (1.085)
	0.99	1.000 (1.320)	0.996 (1.920)	0.996 (1.320)

Empirical coverage percentages and average widths (in parentheses) of the SCBs for Scenario 1

n	$1-\alpha$	(iv): $\varepsilon \sim 0.9N(0,1)+0.1N(0,7^2)$		
		M-SCB	L-SCB	INF-SCB
200	0.95	0.904 (1.622)	0.928 (2.909)	0.926 (1.614)
	0.99	0.938 (2.013)	0.980 (3.601)	0.948 (2.005)
400	0.95	0.960 (1.221)	0.962 (2.313)	0.964 (1.223)
	0.99	0.982 (1.503)	0.996 (2.846)	0.990 (1.507)
600	0.95	0.962 (1.048)	0.966 (2.004)	0.974 (1.051)
	0.99	0.992 (1.284)	0.998 (2.453)	0.994 (1.287)
800	0.95	0.974 (0.944)	0.980 (1.816)	0.972 (0.945)
	0.99	0.998 (1.152)	0.998 (2.214)	0.994 (1.152)

Empirical power of two SCBs test

- Considered the following setup:

$$Y_i = X_{i1} + X_{i2} + \gamma \cdot \exp\{-2(X_{i1} + X_{i2})^2\} + \varepsilon_i$$

  with $\mathbf{X}_i = (X_{i1}, X_{i2})$ generated from the i.i.d. uniform distribution $U^2(-1, 1)$.
- The random errors ε_i generated from $N(0, 1)$ and $0.8N(0, 1) + 0.2N(0, 4^2)$.
- $\gamma = 0, 0.1, \ldots, 0.9, 1$; $n = 200, 400$; $\alpha = 0.05$.
- $$H_0 : g(x) = \sqrt{2}x \quad \text{vs} \quad H_1 : g(x) = \sqrt{2}x + \gamma \cdot \exp\left(-2x^2\right).$$

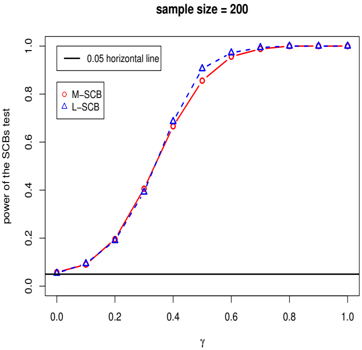

Figure 1: Plots of the empirical power functions out of 500 replications by the M-SCB and the L-SCB tests with the N(0, 1) error at the significance level $\alpha = 0.05$.

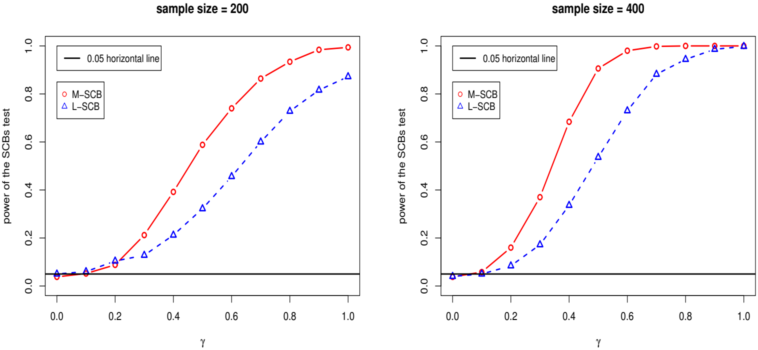

Figure 2: Plots of the empirical power functions out of 500 replications by the M-SCB and the L-SCB tests with the contaminated normal error $\varepsilon \sim 0.8N(0,1) + 0.2N(0,4^2)$ at the significance level $\alpha = 0.05$.

Real data analysis

- The proposed method is applied to analyze a car purchasing dataset.
- This dataset comprises 500 records.
  - X_1: The customer's annual salary
  - X_2: The customer's credit card debt
  - X_3: The customer's net worth
  - Y: The car purchase amount.
- Single-index model

$$\begin{aligned}\log Y &= g\left(\theta_{01}\log X_1 + \theta_{02}\log X_2 + \theta_{03}\log X_3\right) + \varepsilon \\ &= g\left(\mathbf{X}^T\boldsymbol{\theta}_0\right) + \varepsilon,\end{aligned}$$

was utilized to fit the data, where $\boldsymbol{\theta}_0 = (\theta_{01}, \theta_{02}, \theta_{03})^T$ and $\mathbf{X} = (\log X_1, \log X_2, \log X_3)^T$.

Real data analysis

- The M-type B-spline estimation method studied in Zou and Zhu (2014)

$$\hat{\boldsymbol{\theta}}_{\mathsf{R}} = (0.970, 0.241, 0.005)^T$$

- The regular B-spline regression in Wang and Yang (2009)

$$\hat{\boldsymbol{\theta}}_{\mathsf{NR}} = (0.970, 0.241, 0.009)^T .$$

- The proposed M-type SCB for the link function in Corollary 4 was utilized to test the linear trend:

$$H_0 : g(x) = \beta_0 + \beta_1 x \quad \text{vs.} \quad H_1 : g(x) \neq \beta_0 + \beta_1 x \tag{6}$$

- β_0, β_1 were computed using robust linear regression with Huber's loss function applied to fit $(Y_i, \mathbf{X}_i^T \hat{\boldsymbol{\theta}}_{\mathsf{R}})_{i=1}^n$.

$$(\hat{\beta}_0, \hat{\beta}_1)^T = (-1.042, 0.846)^T$$

.

Real data analysis

- The asymptotic p-value for proposed SCB is 0.088.
- The asymptotic p-value for the SCB in Li et al. (2014), Gu and Yang (2015) is 0.037.
- Additionally, the linear trend hypothesis in (6) is also tested by computing β_0 and β_1 using ordinary linear regression on the observed $(Y_i, \mathbf{X}_i^T \hat{\boldsymbol{\theta}}_{\mathsf{NR}})_{i=1}^n$.
- p-values are 0.105 and 0.043 for the proposed method (M-type SCB test) and the least square type SCB test, respectively.
- Annual salary, credit card debt, and net worth have positive effects on the car purchase amount.

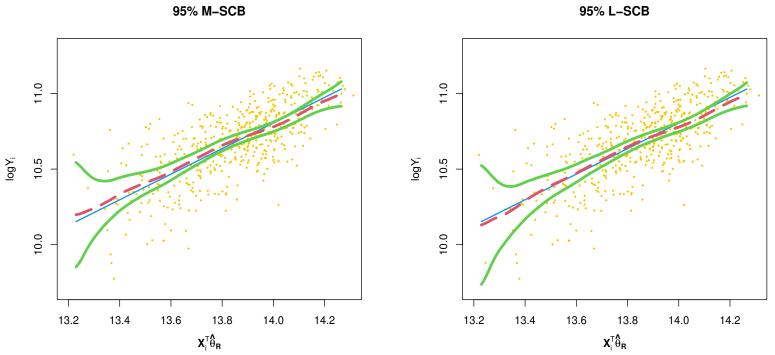

Figure 3: Plots of the M-type kernel estimate (dashed line in left panel) in (7) and ordinary least square kernel estimate (dashed line in right panel) for the link function, the estimated curve under the null hypothesis (solid line) and the 95% M-type SCB (thick solid line in left panel) and the 95% least square type SCB (thick solid line in right panel).

Conclusions

- It effectively reduces the influence of outliers and addresses the sensitivity of least squares estimation to heavy-tailed errors.
- It is a robust approach achieved by treating the coefficients as nuisance parameters. Any $\sqrt{n}$-consistent index coefficient estimator can be employed in the procedures for making global inferences for the link function.
- The method significantly extends the desirable uniform convergence properties of ordinary least squares kernel regression to M-type estimation in single-index models.

Thank You