Efficient Estimation in Convex Single Index Models[1]

Rohit Patra
University of Florida

http://arxiv.org/abs/1708.00145

[1] Joint work with Arun K. Kuchibhotla (UPenn) and Bodhisattva Sen (Columbia)

Overview

Introduction

A semiparametric model

Convex single index model

$$Y = m_0(\boldsymbol{\theta}_0^\top \boldsymbol{X}) + \epsilon, \quad \mathbb{E}(\epsilon|\boldsymbol{X}) = 0.$$

- $(Y, \boldsymbol{X}) \in \mathbb{R} \times \mathbb{R}^d \sim P_{\boldsymbol{\theta}_0, m_0}$.
- $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ is the unknown coefficient vector.
- $m_0 : \mathbb{R} \to \mathbb{R}$ is an unknown convex link function with no parametric restriction.

- Offers a balance between flexibility of nonparametric models and interpretability of parametric models.

Goal and Applicability

Model

$$Y = m_0(\boldsymbol{\theta}_0^\top \boldsymbol{X}) + \epsilon, \quad \mathbb{E}(\epsilon|\boldsymbol{X}) = 0.$$

Problem

Estimate $\boldsymbol{\theta}_0$ and m_0 simultaneously, when we have i.i.d. data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from the above model.

Why convex link

- Convex/concave SIMs are widely used in economics, operations research, financial engineering among other fields.
- Production functions, utility functions, and call option prices are known to be concave.

Restrictions

- *Identifiability:* If $m_1(t) = m_0(-t/2)$ and $\theta_1 = -2\boldsymbol{\theta}_0$, then $m_0(\boldsymbol{\theta}_0^\top \boldsymbol{x}) = m_1(\theta_1^\top \boldsymbol{x})$. Thus we need to assume

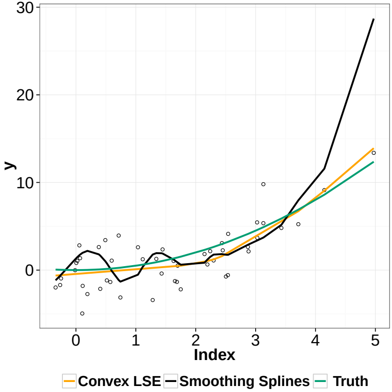
$$\boldsymbol{\theta}_0 \in \Theta := \{\boldsymbol{\beta} \in \mathbb{R}^d : |\boldsymbol{\beta}| = 1, \boldsymbol{\beta}_1 > 0\}.$$

- We need some "regularity" assumptions on the class of link functions.
    - Only Shape constraints
    - Shape and Smoothness constraints

Some relevant works in the shape constrained single index model include: Murphy et al. (1999), Chen and Samworth (2015), Groeneboom and Hendrickx (2016), and Balabdaoui et al. (2016).

- $Y = 2(\boldsymbol{X}^\top \theta_0)^2 + N(0, 5)$,
- $\boldsymbol{X} \sim \mathrm{U}[0,1]^3$, $m(t) = 2t^2$.

- Y: Output for 555 Belgian Firms
- X: Labour, Capital, Wage



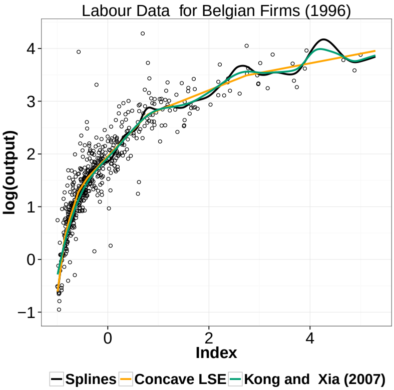


Figure: Estimated link functions: stability due to convex constraint. Index $:= \hat{\boldsymbol{\theta}} x$

.

Estimation

Estimation in Convex SIM [Kuchibhotla, Patra, Sen, 2017]

Lipschitz LSE

$$(\hat{m}, \hat{\theta}) = \underset{m \in \mathcal{C}_L, \boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{y_i - m(\boldsymbol{\theta}^\top \boldsymbol{x}_i)\}^2,$$

where

$$\mathcal{C}_L = \Big\{ m \big| m \text{ is convex and } |m(t_1) - m(t_2)| \leq L|t_1 - t_2| \quad \forall\, t_1, t_2 \in \mathbb{R} \Big\}.$$

Penalized LSE

$$(\check{m}, \check{\boldsymbol{\theta}}) := \underset{(m, \boldsymbol{\theta}) \in \mathcal{R} \times \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{y_i - m(\boldsymbol{\theta}^\top x_i)\}^2 + \check{\lambda}_n^2 \int \{m''(t)\}^2 dt,$$

where $\mathcal{R}$ denotes the class of all convex functions that have absolutely continuous first derivative.

Computation: Alternating Scheme

Recall:

$(\hat{m}, \hat{\boldsymbol{\theta}}) = \text{argmin}_{m \in \mathcal{C}_L, \boldsymbol{\theta} \in \Theta} \, Q_n(m, \theta)$, where

$$Q_n(m, \theta) := \frac{1}{n} \sum_{i=1}^{n} \{y_i - m(\boldsymbol{\theta}^\top \boldsymbol{x}_i)\}^2.$$

Minimization for fixed $\boldsymbol{\theta}$

- For a fixed $\boldsymbol{\theta}$, define

$$\hat{m}_{\boldsymbol{\theta}} := \underset{m \in \mathcal{C}_L}{\text{argmin}} \, Q_n(m, \theta).$$

- The minimization is a convex optimization problem.
- $\hat{m}_\theta$ can computed efficiently using the `nnls` package in R.
- $\check{m}_\theta$ can be computed via a damped newton type algorithm. Our R package `simest` has a implementation of this computation.

Computation contd.

Minimization for fixed $\boldsymbol{\theta}$

- We can now define the profiled loss $\mathbb{Q}_n : \Theta \to \mathbb{R}$,

$$\mathbb{Q}_n(\boldsymbol{\theta}) := Q_n(\hat{m}_\theta, \theta)$$

Minimization over θ

- To find $\hat{\boldsymbol{\theta}}$, we now minimize $\mathbb{Q}_n(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta$.
- The loss function is not convex.
- Simulations suggest a large domain of attraction for moderate d.
- We use a gradient step on the unit sphere based on the right derivative of $\hat{m}_\theta$.
- Some initial work has suggested that the convergence of this alternating scheme is linear, i.e., $|\theta^{(k+1)} - \hat{\theta}| \leq (1-\rho)|\theta^{(k)} - \hat{\theta}|$.

Asymptotics

Theoretical properties of LLSE

Rates of convergence of the estimators [Kuchibhotla, Patra, Sen, 2017]

Under some regularity conditions on m_0 and distribution of $\boldsymbol{X}$ (bounded support), sub-Gaussian errors and $L \geq L_0$ the Lipschitz LSE satisfies

$$\begin{aligned}
\|\hat{m} \circ \hat{\boldsymbol{\theta}} - m_0 \circ \boldsymbol{\theta}_0\| &= O_p(n^{-2/5}), && \text{[Estimation error]} \\
\|\hat{m} \circ \boldsymbol{\theta}_0 - m_0 \circ \boldsymbol{\theta}_0\| &= O_p(n^{-2/5}), && \text{[Estimation error of } \hat{m}\text{]} \\
\|\hat{m}' \circ \boldsymbol{\theta}_0 - m_0' \circ \boldsymbol{\theta}_0\| &= O_p(n^{-2/15}), && \text{[Estimation error of } \hat{m}'\text{]} \\
|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| &= O_p(n^{-2/5}). && \text{[Estimation error of } \hat{\boldsymbol{\theta}}\text{]}
\end{aligned}$$

For a convex Lipschitz function $g : \mathbb{R} \to \mathbb{R}$ let g' define the right derivative of g that satisfies $g(b) = g(a) + \int_a^b g'(t)dt$.

Here for any $f : \mathbb{R} \to \mathbb{R}$, and $\theta \in \mathbb{R}^d$, we define $\|f \circ \theta\|^2 := \int |f(\theta^\top x)|^2 dP_X(x)$.

Asymptotic normality: Homoscedastic model

Recall:

$$(\hat{m}, \hat{\theta}) = \underset{(m,\boldsymbol{\theta})\in\mathcal{C}_L\times\Theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\{y_i - m(\boldsymbol{\theta}^\top \boldsymbol{x}_i)\}^2.$$

Semiparametric efficiency of $\hat{\boldsymbol{\theta}}$ [Kuchibhotla, Patra, and Sen, 2017]

Assume that $\mathbb{E}(\epsilon^2|\boldsymbol{X}) \equiv \sigma^2$. Let $\boldsymbol{\ell}_{\boldsymbol{\theta},m} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d-1}$ be the efficient score and let us define the efficient information matrix

$$\mathcal{I}_{\boldsymbol{\theta}_0,m_0} := \mathbb{E}(\boldsymbol{\ell}_{\boldsymbol{\theta}_0,m_0}\boldsymbol{\ell}_{\boldsymbol{\theta}_0,m_0}^\top) \in \mathbb{R}^{(d-1)\times(d-1)}.$$

If m_0 is twice differentiable, then under some regularity conditions we can conclude

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, H_{\theta_0}\mathcal{I}_{\boldsymbol{\theta}_0,m_0}^{-1}H_{\theta_0}^\top).$$

Inference

- Our result readily yields confidence sets for θ_0.
- We can use the following plug-in estimator for the covariance estimator:

$$\hat{\Sigma} := \hat{\sigma}^4 H_{\hat{\theta}} P_{\hat{\theta},\hat{m}}[\boldsymbol{\ell}_{\hat{\theta},\hat{m}}(Y,X)\boldsymbol{\ell}_{\hat{\theta},\hat{m}}^\top(Y,X)]^{-1} H_{\hat{\theta}}^\top,$$

where

$$\hat{\sigma}^2 := \sum_{i=1}^{n}[y_i - \hat{m}(\hat{\theta}^\top x_i)]^2/n$$

$$\ell_{\theta,m}(y,x) := \left(y - m(\theta^\top x)\right) m'(\theta^\top x) H_\theta^\top \left\{x - h_\theta(\theta^\top x)\right\}.$$

Asymptotic $1-2\alpha$ confidence interval

$$\left[\hat{\theta}_i - \frac{z_\alpha}{\sqrt{n}}\left(\hat{\Sigma}_{i,i}\right)^{1/2},\ \hat{\theta}_i + \frac{z_\alpha}{\sqrt{n}}\left(\hat{\Sigma}_{i,i}\right)^{1/2}\right],$$

Asymptotic properties of the PLSE

If $\check{\lambda}_n^{-1} = O_p(n^{2/5})$ and $\check{\lambda}_n = o_p(n^{-1/4})$, then under some similar regularity assumptions the PLSE satisfies

$$\|\check{m} \circ \check{\boldsymbol{\theta}} - m_0 \circ \boldsymbol{\theta}_0\| = O_p(\check{\lambda}_n), \qquad \text{[Estimation error]}$$

$$\|\check{m} \circ \boldsymbol{\theta}_0 - m_0 \circ \boldsymbol{\theta}_0\| = O_p(\check{\lambda}_n), \qquad \text{[Estimation error for } \check{m}\text{]}$$

$$\|\check{m}' \circ \boldsymbol{\theta}_0 - m_0' \circ \boldsymbol{\theta}_0\| = O_p(\check{\lambda}_n^{1/2}), \qquad \text{[Estimation error for } \check{m}'\text{]}$$

and

$$\sqrt{n}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, H_{\theta_0} \mathcal{I}^{-1}_{\boldsymbol{\theta}_0, m_0} H_{\theta_0}^\top).$$

- An example choice of $\check{\lambda}_n := C\, n^{-2/5}$.
- Proof borrows ideas from the empirical process theory; see e.g., Mammen and van de Geer (1997) and van de Geer (2000).

Difficulty in proving efficiency

The LLSE $\hat{m}$ is a piecewise affine function and lies on the boundary of $\mathcal{C}_L$.

Nuisance tangent space

- Consider the following model:

$$Y = m(\theta^\top X) + \epsilon, \quad \text{where} \quad m \in \mathcal{C}_L, \text{ and } \theta \in \Theta.$$

- A linear perturbation/submodel around m:

$$m_{s,a}(t) = m(t) - s\, a(t), \qquad \text{where } s \in \mathbb{R}.$$

- The score for the single index model along this submodel is proportional to $a(\cdot)$.
- $$\overline{\mathrm{lin}}\{a : D \to \mathbb{R} | m_{s,a} \in \mathcal{R} \text{ for small enough } s\} \subseteq L_2(\Lambda).$$
- The set inclusion is strict when m is not strongly convex.

Efficient score

- Let S_θ denote the parametric score of the model and Λ_m denote the nuisance tangent space.

- When m is strongly convex, the efficient score is known to be

$$\Pi(S_\theta|\Lambda_m^\perp) := \ell_{\theta,m}(y,x) = \big(y - m(\theta^\top x)\big)\, m'(\theta^\top x) H_\theta^\top \left\{x - h_\theta(\theta^\top x)\right\}.$$

- Since $\hat{m}$ is not strongly convex, it is not clear if one can show that

$$\Pi(S_{\hat{\theta}}|\Lambda_{\hat{m}}^\perp) \stackrel{??}{=} \ell_{\hat{\theta},\hat{m}}.$$

Efficient score

- Since $\hat{m}$ lies on the boundary of $\mathcal{C}_L$, the "least favorable path" does not exist, i.e., we can not find (θ_t, m_t) (centered at $(\hat{\theta}, \hat{m})$) such that

$$\ell_{\hat{\theta},\hat{m}} \overset{??}{=} \frac{\partial}{\partial t}\left(y - m_t(\theta_t^\top x)\right)^2\Big|_{t=0}.$$

- Since LLSE is the minimizer of the least squares loss, this would mean

$$\mathbb{P}_n \ell_{\hat{\theta},\hat{m}} = 0.$$

- Since $\hat{m}$ is piecewise affine, it is not clear if one can show that

$$\mathbb{P}_n \ell_{\hat{\theta},\hat{m}} \overset{??}{=} o_p(n^{-1/2}).$$

Approximations

- We next try to construct some paths around $(\hat{\theta}, \hat{m})$ that will have score "close" to $\ell_{\hat{\theta},\hat{m}}$.
- We find a path that has the following score:

$$\mathfrak{S}_{\theta,m} = \{(y - m(\theta^\top x)) H_\theta^\top \Big[m'(\theta^\top x)x + \int_{s_0}^{\theta^\top x} m'(u)k'(u)du - m'(\theta^\top x)k(\theta^\top x)$$

$$+ m_0'(s_0)k(s_0) - m_0'(s_0)h_{\theta_0}(s_0)\Big].$$

- Note $\ell_{\theta_0,m_0} = \mathfrak{S}_{\theta_0,m_0} = (y - m(\theta_0^\top x)) H_{\theta_0}^\top m_0'(\theta_0^\top x)\left\{x - h_{\theta_0}(\theta_0^\top x)\right\}$.
- van der Vaart (2002) calls such a path "approximately least favorable".

Approximation, Part 2

- $\mathfrak{S}_{\theta,m}$ is not very tractable, so we further approximate this by

$$\psi_{\theta,m}(x,y) := (y - m(\theta^\top x))H_\theta^\top[m'(\theta^\top x)x - h_{\theta_0}(\theta^\top x)m_0'(\theta^\top x)].$$

- Compare $\psi_{\theta,m}$ to

$$\ell_{\theta,m} = \left(y - m(\theta^\top x)\right)H_\theta^\top m'(\theta^\top x)\left\{x - h_\theta(\theta^\top x)\right\}.$$

  Also note $\psi_{\theta_0,m_0} = \ell_{\theta_0,m_0}$.
- We show that

$$\mathbb{P}_n\psi_{\hat{\theta},\hat{m}} = o_p(n^{-1/2}).$$

Simulation study

Another Estimator

Convex LSE

$$(\tilde{m}, \tilde{\theta}) := \underset{(m,\theta)\in \mathcal{C}\times\Theta}{\operatorname{argmin}} Q_n(m, \theta).$$

- We can compute the LSE via the alternating minimization algorithm.
- Simulations suggest $\tilde{\theta}$ is $\sqrt{n}-$consistent.
- The behavior of $\tilde{m}'$ at the boundary is not well-understood.
- In fact, in univariate regression, $\tilde{m}'$ is unbounded in probability at the boundary.

Choice of L

$$Y = (\theta_0^\top X)^2 + N(0, .1^2), \quad \text{where } X \sim \text{Uniform}[-1,1]^4 \text{ and } \theta_0 = \mathbf{1}_4/2.$$

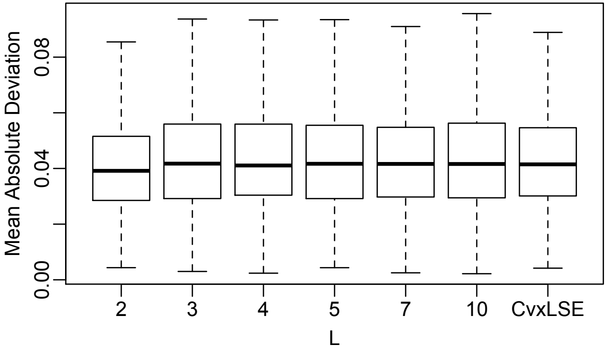


Figure: Box plots of $\frac{1}{4}\sum_{i=1}^{4}|\theta_i - \theta_{0,i}|$ (over 1000 replications, $n = 500$) from the following model as the tuning parameter varies over $\{3, 4, 5, 7, 10\}$ and `CvxLSE`. Here $L_0 = 4$.

Confidence Interval

$$Y = (\theta_0^\top X)^2 + N(0, .3^2), \quad X \sim \text{Uniform}[-1,1]^3 \text{ and } \theta_0 = \mathbf{1}_3/\sqrt{3}. \quad (1)$$

Table: The estimated coverage probabilities and average lengths (obtained from 800 replicates) of nominal 95% confidence intervals for the first coordinate of θ_0 for the model (1).

n	CvxLip		CvxPen	
	Coverage	Avg Length	Coverage	Avg Length
50	0.92	0.30	0.94	0.29
100	0.91	0.18	0.92	0.19
200	0.92	0.13	0.93	0.13
500	0.94	0.08	0.92	0.08
1000	0.93	0.06	0.92	0.06
2000	0.92	0.04	0.93	0.04

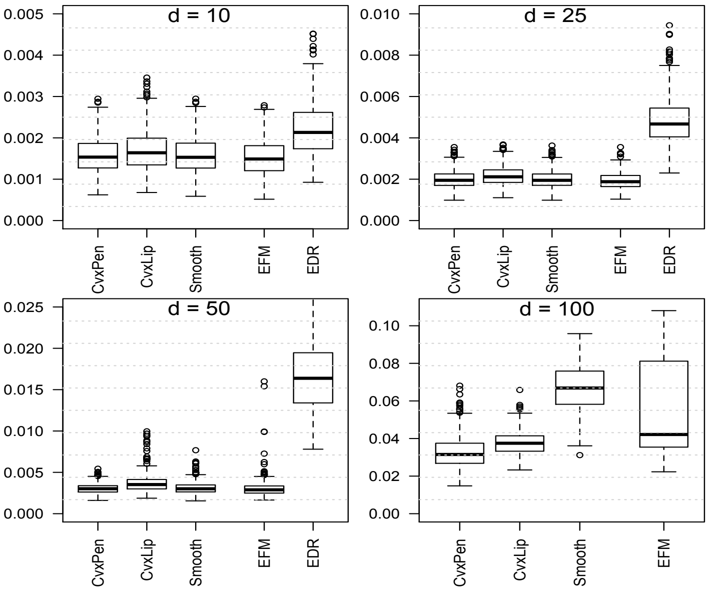


Figure: Boxplots of $\sum_{i=1}^{d} |\hat{\theta}_i - \theta_{0,i}|/d$ (over 500 replications) based on 200 observations for dimensions $10, 25, 50,$ and 100.

Summary

- First work providing efficient estimator in shape-constrained LSE (in a bundled parameter problem) when $\hat{m}$ is piecewise affine.
- Our estimators readily lead to asymptotic confidence sets for θ_0.
- Our methods are robust towards the choice of the tuning parameter.
- The proposed estimators are implemented in the R package `simest`.

References

[1] Kuchibhotla, A. K. and Patra, R. K. (2016).
simest: Single Index Model Estimation with Constraints on Link Function.
R package version 0.2.

[2] Kuchibhotla, A. K., Patra, R. K., and Sen, B. (2017).
Efficient Estimation in Convex Single Index Models.
arxiv.org/abs/1708.00145.

[3] Kuchibhotla, A. K., and Patra, R. K. (2017).
Efficient estimation in single index models through smoothing splines.
arxiv.org/abs/1612.00068.

[4] Balabdaoui, F., Durot, C. and Jankowski, H. (2016).
Least squares estimation in the monotone single index model.
arxiv.org/abs/1610.06026.

[5] Groeneboom, P. and Hendrickx, K. (2016).
Current status linear regression.
Annals of Statistics (Forthcoming).

[6] Chen, Y. and Samworth, R. J. (2014).
Generalised additive and index models with shape constraints.
JRSSB. 78(4), 729–754..

[7] Murphy, S. A., van der Vaart, A. W. and Wellner, J. A. (1999).
Current status regression.
Math. Methods Statist. 8(3), 407425.

Thank You! Questions?