High-dimensional single-index models: link estimation, marginal inference, and efficiency improvement

Kazuma Sawaya

Graduate School of Economics, The University of Tokyo

Model

Single-index model: Suppose that we observe n iid pairs $\{(\boldsymbol{X}_i, y_i)\}_{i=1}^n$ following

$$\mathbb{E}[y_i \mid \boldsymbol{X}_i = \boldsymbol{x}] = g(\boldsymbol{x}^\top \boldsymbol{\beta}),$$

or equivalently $y_i = g(\boldsymbol{X}_i^\top \boldsymbol{\beta}) + \varepsilon_i, \mathbb{E}[\varepsilon_i \mid \boldsymbol{X}_i] = 0$ where

- $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ is a p-dimensional feature vector,
- $y_i \in \mathcal{Y} \subset \mathbb{R}$ is a scalar response,
- $g : \mathbb{R} \to \mathbb{R}$ is an unknown link function,
- $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown coefficient vector.

Typical examples of the model encompass

- *Linear regression*: $y_i \mid \boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{X}_i^\top \boldsymbol{\beta}, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon > 0$ by setting $g(t) = t$.
- *Poisson regression*: $y_i \mid \boldsymbol{X}_i \sim \mathrm{Pois}(\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}))$ by setting $g(t) = \exp(t)$.
- *Binary choice models*: $y_i \mid \boldsymbol{X}_i \sim \mathrm{Bern}(g(\boldsymbol{X}_i^\top \boldsymbol{\beta}))$ with $g : \mathbb{R} \to [0, 1]$. This includes *logistic regression* for $g(t) = 1/(1 + \exp(-t))$ and *the probit model* for $g(t) = \Phi(t)$.

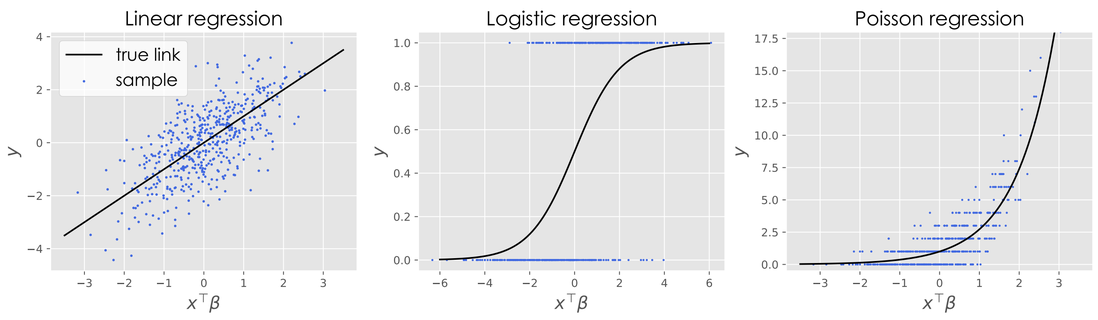

Figure 1. Examples of the single-index model

We consider a prevalent setting where the sample size n and the dimensionality p are large and compatible. To approximate such situations, this study explores *proportional asymptotics* where

$$n, p \to \infty, \qquad p/n =: \kappa \to \bar{\kappa} > 0.$$

Previous Results

Let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top \in \mathbb{R}^{n \times p}$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. An innovative work by P. Bellec [2] demonstrates the average asymptotic normality of M-estimators regardless of the link violation. For instance, when $p < n$, the least squares estimator $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ obeys

$$\frac{1}{p} \sum_{j=1}^{p} \mathbb{E}\left[\left| \frac{\sqrt{p}(\tilde{\beta}_j - \tilde{\mu}\beta_j)}{\tilde{\sigma}} - Z_j \right|^2 \right] \to 0,$$

where $Z_j \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$, $\tilde{\mu}^2 = (\|\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{y}\|^2 - (1 - \kappa)\tilde{\sigma}^2)/n$ and $\tilde{\sigma}^2 = \kappa \|\boldsymbol{P}_{\boldsymbol{X}}^\perp \boldsymbol{y}\|^2 / (n(1 - \kappa)^2)$. However, the marginal asymptotic normality and the construction of the link estimation remain unclear.

Our Goals

1. **Link Estimation**: We propose a uniformly consistent estimator of $g(\cdot)$.
2. **Marginal Inference**: We establish the marginal asymptotic normality of estimators for $\boldsymbol{\beta}$. This facilitates hypothesis testing of $\beta_j = 0$ and variable selection.
3. **Efficiency Enhancement**: Leveraging the information of the estimated link function, we propose a novel estimator of $\boldsymbol{\beta}$ with smaller asymptotic variance.

Assumptions

We impose the following assumptions to establish the theory.

A1 (moderately high dimensions) $0 < \kappa = p/n < \infty$.

A2 (Gaussian feature and identification) $0 < C^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq C$ and $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} = 1$.

A3 (smooth and monotonic link) $g(\cdot) \in C^1(\mathbb{R})$ and $0 < \inf_t g'(t)$.

A4 (moment conditions) $\mathbb{E}[y_1] < \infty$. $m_2(x) = \mathbb{E}[y_1^2 \mid \boldsymbol{X}_1^\top \boldsymbol{\beta} = x]$ is continuous.

Estimation

The estimation procedure comprises three steps. We assume $p \leq n$ for brevity.

Step1: Pilot estimator. Consider the least squares estimator $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$. Applying [2]'s result, we have an approximation

$$W_i := \tilde{\mu}^{-1} \left(\boldsymbol{X}_i^\top \tilde{\boldsymbol{\beta}} + \tilde{\eta}(y_i - \boldsymbol{X}_i^\top \tilde{\boldsymbol{\beta}}) \right) \approx \boldsymbol{X}_i^\top \boldsymbol{\beta} + \mathcal{N}(0, \tilde{\tau}^2), \tag{1}$$

where $\tilde{\eta} = \kappa/(1 - \kappa)$ and $\tilde{\tau}^2 = \tilde{\sigma}^2/\tilde{\mu}^2$. Remind that W_i and $\tilde{\tau}^2$ are computable from the dataset.

Step2: Link estimation. The nonparametric regression between $\boldsymbol{X}_i^\top \boldsymbol{\beta}$ and y_i is infeasible since $\boldsymbol{\beta}$ is unknown. Alternatively, we can use W_i in (1) instead of $\boldsymbol{X}_i^\top \boldsymbol{\beta}$. Here, the problem of link estimation is approximately reduced to the nonparametric regression involving errors-in-variable (EIV) (i.e., noise appears in input variables).

To address the EIV, the deconvolution technique [4] plays an essential role. Since convolution in the frequency domain is equivalent to multiplication, and the density of W_i is the convolution of $\boldsymbol{X}_i^\top \boldsymbol{\beta}$'s and $\mathcal{N}(0, \tau^2)$'s densities, we have

$$f_{\boldsymbol{X}_1^\top \boldsymbol{\beta}}(t) \approx \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\hat{f}_{W_1}]}{\mathcal{F}[f_{\mathcal{N}(0, \tilde{\tau}^2)}]} \right](t),$$

where $f_*(\cdot)$ is a density of $*$, $\mathcal{F}[\cdot]$ is the Fourier transform, and $\hat{f}_W(\cdot)$ is a kernel density estimator (KDE) from $\boldsymbol{W} = (W_1, \ldots, W_n)$. By using the right term as a KDE for $\boldsymbol{X}_1^\top \boldsymbol{\beta}$, we obtain

$$\hat{g}(x) := \sum_{i=1}^{n} y_i K_n\left((x - W_i)/h_n\right) \Big/ \sum_{i=1}^{n} K_n\left((x - W_i)/h_n\right), \tag{2}$$

$$K_n(x) = (2\pi)^{-1} \int_{-\infty}^{\infty} \exp(-\mathrm{i}tx) \frac{\phi_K(t)}{\phi_{\tilde{\tau}}(t/h_n)} dt,$$

where $\phi_K(\cdot), \phi_{\tilde{\tau}}(\cdot)$ are Fourier transforms of a kernel $K : \mathbb{R} \to \mathbb{R}$ and $\mathcal{N}(0, \tilde{\tau}^2)$ density, $h_n > 0$ is a bandwidth, and $\mathrm{i} = \sqrt{-1}$.

Step3: Estimation of $\boldsymbol{\beta}$. We employ the matching/surrogate loss [1] which is strictly convex for any strictly monotonically increasing $\bar{g}(\cdot)$:

$$\sum_{i=1}^{n} \ell(\boldsymbol{b}; \boldsymbol{X}_i, y_i, \bar{g}) := \sum_{i=1}^{n} \left(\bar{G}(\boldsymbol{X}_i^\top \boldsymbol{b}) - y_i \boldsymbol{X}_i^\top \boldsymbol{b} \right),$$

where $\bar{G}(\cdot)$ is any functions satisfying $\bar{G}'(t) = \bar{g}(t)$ for $t \in \mathbb{R}$. Define an M-estimator

$$\hat{\boldsymbol{\beta}}(\hat{g}) = \operatorname*{argmin}_{\boldsymbol{b} \in \mathbb{R}^p} \sum_{i=1}^{n} \ell(\boldsymbol{b}; \boldsymbol{X}_i, y_i, \mathcal{R}[\hat{g}]),$$

where $\mathcal{R}[\cdot]$ is the rearrangement functional [3] which monotonizes the input function,

$$\mathcal{R}[\bar{g}](x) = \inf\left\{ t \in \mathbb{R} : \int_{\mathcal{X}} 1\{\bar{g}(u) \leq t\} du \geq x \right\}.$$

The estimator is a generalization of MLEs for logistic regression, Poisson regression, etc. Also, it has a reasonable property that $\boldsymbol{\beta} = \operatorname{argmin}_{\boldsymbol{b} \in \mathbb{R}^p} \mathbb{E}[\ell(\boldsymbol{b}; \boldsymbol{X}_1, y_1, g)]$.

Theory

Theorem 1 (Consistency of $\hat{g}(\cdot)$)

Under the assumptions, for any $a < b$ and $h_n = c(\log n)^{1/2}$, as $n \to \infty$,

$$\sup_{a \leq x \leq b} |\hat{g}(x) - g(x)| = O_{\mathrm{p}}\left(\frac{1}{\sqrt{\log n}} \right).$$

Theorem 2 (Marginal asymptotic normality of $\hat{\boldsymbol{\beta}}(\hat{g})$)

Consider a finite collection of indices $\mathcal{S} \subset [p]$ such that $\sqrt{p}\lambda_{\max}(\boldsymbol{\Theta}_{\mathcal{S}}^{-1})^{1/2}(\boldsymbol{\beta}_{\mathcal{S}}^\top \boldsymbol{\Theta}_{\mathcal{S}}^{-1} \boldsymbol{\beta}_{\mathcal{S}})^{1/2} = O(1)$ where $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. Then, under the assumptions,

$$\frac{\sqrt{p}\boldsymbol{\Theta}_{\mathcal{S}}^{-1/2}\left(\hat{\boldsymbol{\beta}}_{\mathcal{S}}(\hat{g}) - \hat{\mu}(\hat{g})\boldsymbol{\beta}_{\mathcal{S}} \right)}{\hat{\sigma}(\hat{g})} \xrightarrow{\mathrm{d}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{|\mathcal{S}|}),$$

where

$$\hat{\mu}^2(\hat{g}) = \frac{\|\boldsymbol{X}\hat{\boldsymbol{\beta}}(\hat{g})\|^2}{n} - \kappa(1 - \kappa)\hat{\sigma}^2(\hat{g}), \quad \hat{\sigma}^2(\hat{g}) = \frac{n^{-1}\|\boldsymbol{y} - \hat{g}(\boldsymbol{X}\hat{\boldsymbol{\beta}}(\hat{g}))\|^2}{(n^{-1}\mathrm{tr}(\boldsymbol{V}(\hat{g})))^2},$$

with $\boldsymbol{V}(\hat{g}) = \boldsymbol{D}(\hat{g}) - \boldsymbol{D}(\hat{g})\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{D}(\hat{g})\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{D}(\hat{g})$, $\boldsymbol{D}(\hat{g}) = \mathrm{diag}(\hat{g}'(\boldsymbol{X}\hat{\boldsymbol{\beta}}(\hat{g})))$.

Hence, we can define confidence intervals for each β_j with a preassigned confidence level $(1 - \alpha)$:

$$\mathrm{CI}_{1-\alpha}^j := \frac{1}{\hat{\mu}(\hat{g})} \left[\hat{\beta}_j(\hat{g}) - z_{(1-\alpha/2)} \frac{\hat{\sigma}(\hat{g})}{\sqrt{p}\Theta_{jj}^{-1/2}},\ \hat{\beta}_j(\hat{g}) + z_{(1-\alpha/2)} \frac{\hat{\sigma}(\hat{g})}{\sqrt{p}\Theta_{jj}^{-1/2}} \right],$$

for $j = 1, \ldots, p$ with $z_{(1-\alpha/2)}$ the $(1 - \alpha/2)$-quantile of the standard normal distribution. This exactly regulates the asymptotic coverage proportion at a given confidence level $(1 - \alpha)$.

Numerical Illustrations

The simulations below demonstrate that the proposed estimator exhibits marginal asymptotic normality and indeed enhances estimation efficiency compared to the least squares pilot estimator.

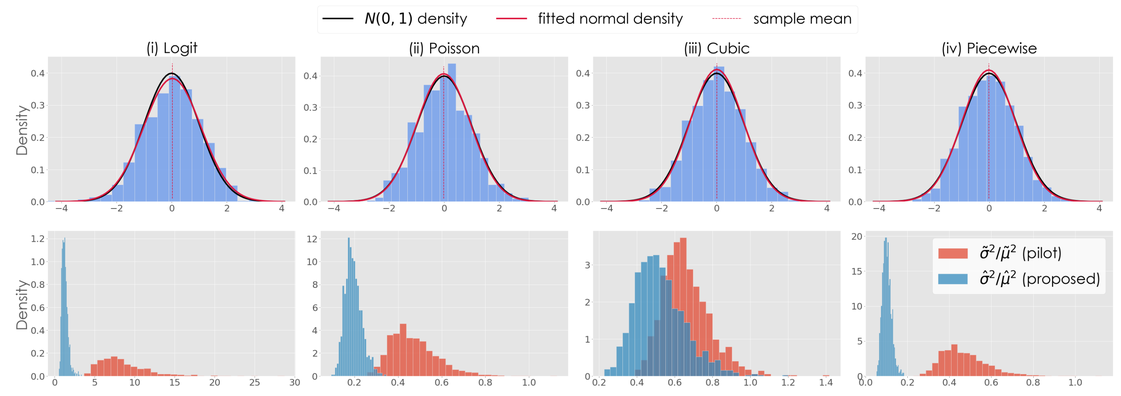

Figure 2. The first row illustrates histograms of the statistics $\sqrt{p}(\hat{\beta}_1(\hat{g}) - \hat{\mu}(\hat{g})\beta_1)/\hat{\sigma}(\hat{g})$ for $n = 500, p = 200$ over $1,000$ replications, which are expected to resemble $\mathcal{N}(0, 1)$ density. The second row shows histograms of estimates for the effective asymptotic variance of $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}(\hat{g})$. We use $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ and the signals $\boldsymbol{\beta} \sim \mathrm{Unif}(\mathbb{S}^{p-1})$. The columns correspond to each model: (i) logistic regression; (ii) Poisson regression; (iii) cubic regression $y_i = (\boldsymbol{X}_i^\top \boldsymbol{\beta})^3/3 + \mathcal{N}(0, 1)$; (iv) piecewise regression $y_i = g(\boldsymbol{X}_i^\top \boldsymbol{\beta}) + \mathcal{N}(0, 1)$ where $g(t) = (0.2t - 2.3)1_{(-\infty, -1]} + 2.5t1_{(-1,1)} + (0.2t + 2.3)1_{[1,\infty)}$ for every $i \in \{1, \ldots, n\}$.

References

[1] Peter Auer, Mark Herbster, and Manfred KK Warmuth.
Exponentially many local minima for single neurons.
Advances in neural information processing systems, 8, 1995.

[2] Pierre C Bellec.
Observable adjustments in single-index models for regularized m-estimators.
arXiv preprint arXiv:2204.06990, 2022.

[3] Victor Chernozhukov, Ivan Fernandez-Val, and Alfred Galichon.
Improving point and interval estimators of monotone functions by rearrangement.
Biometrika, 96(3):559–575, 2009.

[4] Leonard A Stefanski and Raymond J Carroll.
Deconvolving kernel density estimators.
Statistics, 21(2):169–184, 1990.