Dimension Adaptive Estimation

by Chencheng Fang

IFS

Brown Bag, December 2023

Motivation

- Goal: estimate the function g in the model

$$Y = g(X) + U, \quad E(U \mid X) = 0$$

  where $X = (X_1, X_2, \ldots, X_d)' \in \mathbb{R}^d$.
- Choose a estimator with a good rate of convergence, such as
  - Linear regression
  - Nonparametric regression
  - Nonparametric additive regression
  - $\cdots$
- Problem: The underlying form of $g(X)$ is unknown, making the choice of estimator difficult.

- Again, assume $X = (X_1, X_2, \ldots, X_d)' \in \mathbb{R}^d$
- Series approximation

$$
\begin{aligned}
Y &= g(X) + U \\
&\approx \sum_{k_1=0}^{K_n} \cdots \sum_{k_d=0}^{K_n} b_{k_1,\ldots,k_d} p_{k_1}(X_1) \cdots p_{k_d}(X_d) + U
\end{aligned}
$$

- Regress Y on the set of basis functions using least squares

- Suppose $g(X)$ has α continuous derivatives in all directions.
- Then:

$$\int \left(\hat{g}(x) - g(x)\right)^2 f_X(x)dx = O_p\big(\underbrace{K_n^d/n}_{\text{variance}} + \underbrace{K_n^{-2\alpha}}_{\text{squared bias}}\big)$$

- To get the fastest convergence rate,

$$K_n^d/n \asymp K_n^{-2\alpha} \;\Rightarrow\; K_n \asymp n^{\frac{1}{d+2\alpha}} \;\Rightarrow\; IMSE = O_p\left(n^{-\frac{2\alpha}{d+2\alpha}}\right)$$

- Curse of dimensionality

- If we impose additivity assumption on $g(X)$, that is

$$Y = g_1(X_1) + \ldots + g_d(X_d) + U, \quad E(U \mid X) = 0$$

- Then, we have to estimate $d(K_n + 1)$ terms, and by following a similar argument,

$$dK_n/n \asymp K_n^{-2\alpha} \Rightarrow K_n \asymp n^{\frac{1}{1+2\alpha}} \Rightarrow IMSE = O_p(n^{-\frac{2\alpha}{1+2\alpha}})$$

which is the same as that in univariate case.

- Suppose $d = 3$ and $g(X)$ has 2 continuous derivatives in all directions, i.e., $\alpha = 2$. X is normalized such that $X_d \in [0, 1]$. Consider three types of underlying form of $g(X)$ and three types of estimator.
- Would each estimator converge to the true $g(X)$? If yes, what is the convergence rate of IMSE?

$g(X)$	OLS Estimator	Additive Series Estimator	Series Estimator
$X_1 + 2X_2 - X_3$	yes; n^{-1}	yes; $n^{-\frac{4}{5}}$	yes; $n^{-\frac{4}{7}}$
$\sin(X_1) + \log(2 + X_2) - X_3^2$	no	yes; $n^{-\frac{4}{5}}$	yes; $n^{-\frac{4}{7}}$
$\sin(X_1 + 2X_2 - X_3)$	no	no	yes; $n^{-\frac{4}{7}}$

- The "best" estimator is different in different case of underlying $g(X)$!

- Could we find out a single estimator which always converges to true $g(X)$ and adapts its convergence rate to different underlying forms?
- The answer is yes! We could do so by

$$\text{Series Estimation} + \text{Lasso}$$

- It is termed as dimension adaptive estimator in this research.

Contribution

- In literature, this type of penalized series estimation is also widely discussed, e.g., Chen(2011); Zhang and Simon (2022); Luo and Sang (2022). But few of them analyzes it from the perspective of convergence rate. The contribution of this research is twofold.
- In theory, we find out that our dimension adaptive estimator could always converge to true function and achieve good convergence rate under all three cases of underlying models.
- In practice, an R-package dimada is developed to help with easy implementation of this estimator.

Dimension Adaptive Estimation

Estimator

- Let $(X, Y) \in (\mathcal{X}, \mathbb{R})$ where $\mathcal{X}$ is a Borel subset of $\mathbb{R}^d$. Here, X is normalized such that $X_d \in [0, 1]$. Again, consider the model

$$Y = g(X) + U, \quad E(U \mid X) = 0$$

- Our dimension adaptive estimator is basically a series estimator with ℓ_1 penalty, as shown below.

$$\hat{\beta} = \arg\min_{b_{k_1,\dots,k_d}} \sum_{i=1}^{n} \left(Y_i - \sum_{k_1=0}^{K_n} \cdots \sum_{k_d=0}^{K_n} b_{k_1,\dots,k_d} p_{k_1}(X_1) \cdots p_{k_d}(X_d) \right)^2 + \lambda \sum_{k_1=0}^{K_n} \cdots \sum_{k_d=0}^{K_n} \omega_{k_1,\dots,k_d} |b_{k_1,\dots,k_d}|$$

- Basis functions for series estimation:
  - Power series
  - Legendre polynomials: orthogonal on $[-1, 1]$
  - Splines: piecewise polynomial
  - B-Splines: basis spline. All possible splines could be built from a combination of B-splines.
  - Trigonometric polynomials (on $[-1, 1]$): $1, \cos(\pi x), \sin(\pi x), 2\cos(\pi x), 2\sin(\pi x), \ldots$
  - $\ldots$
- The choice of basis functions affects the values of some parameters in a theorem later.

- Bunea, Tsybakov and Wegkamp (2007) derive oracle properties of ℓ_1-penalized least squares in non-parametric regression setting with random design.

Sparsity

Let $M(\beta)$ denote the number of non-zero coefficients of β, then

$$M(\beta) = \sum_{j=1}^{M} \mathbb{I}_{\{\beta_j \neq 0\}} = \mathrm{Card} J(\beta)$$

where $\mathbb{I}_{\{\cdot\}}$ is indicator function and $J(\beta) = \{j \in \{1, \ldots, M\} : \beta_j \neq 0\}$.

- The smaller $M(\beta)$, the sparser β.
- However, in series estimation, series of basis functions is not an exact representation of true function g, but an approximation instead.

Dimension Adaptive Estimation

Bunea, Tsybakov and Wegkamp (2007)

Weak Sparsity

Let $C_g > 0$ be a constant depending on g, and denote

$$\mathcal{B} = \{\beta \in \mathbb{R}^M : ||g_\beta - g||^2 \leq C_g r_{n,M}^2 M(\beta)\}$$

as the oracle set. Here, $||\cdot||$ is the $L_2(\mu)$-norm with probability measure μ. $g_\beta = \sum_{j=1}^M \beta_j g_j$ for g_j in the dictionary $\mathcal{G}_M = \{g_1, \ldots, g_M\}$. Then, if $\mathcal{B}$ is non-empty, g is said to have a weak sparsity property relative to $\mathcal{G}_M$.

- $r_{n,M}$ is a tuning parameter that is chosen to ensure the bias-variance balance of $||g_\beta - g||^2$ realized for $M(\beta) \sim n^{\frac{1}{2\alpha+1}}$, so $r_{n,M} \sim \sqrt{\frac{\log M}{n}}$
- In random design, $M = M(n) = \log(n)M(\beta) \geq M(\beta)$ for n large.
- With weak sparsity, one may believe that, for some $\beta^* \in \mathbb{R}^M$, squared approximation error is bounded, up to logarithmic factors, by $M(\beta^*)/n$.

Theorem of Sparsity Oracle Inequalities

Assume some assumptions hold, then for all $\beta \in \mathcal{B}$ we have

$$\mathbb{P}\left\{||\widehat{g} - g||^2 \leq B_1 \kappa_M^{-1} r_{n,M}^2 M(\beta)\right\} \geq 1 - \pi_{n,M}(\beta)$$
$$\mathbb{P}\left\{|\widehat{\beta} - \beta|_1 \leq B_2 \kappa_M^{-1} r_{n,M} M(\beta)\right\} \geq 1 - \pi_{n,M}(\beta)$$

where $B_1 > 0$ and $B_2 > 0$ are constants depending on c_0 and C_g only and

$$\pi_{n,M}(\beta) \leq 10M^2 \exp\left(-c_1 n \min\left\{r_{n,M}^2, \frac{r_{n,M}}{L}, \frac{1}{L^2}, \frac{\kappa_M^2}{L_0 M^2(\beta)}, \frac{\kappa_M}{L^2 M(\beta)}\right\}\right) + \exp\left(-c_2 \frac{M(\beta)}{L^2(\beta)} n r_{n,M}^2\right)$$

for some positive constants c_1, c_2 depending on c_0, C_g and b only and $L(\beta) = ||g - g_\beta||_\infty$.

- κ_M, c_0 and b are some constants defined in assumptions. L and L_0 are parameters, which can be constant or associated with sample size, depending on the basis functions.
- This theorem is very useful in the case of weak sparsity.
- By replacing β and $M(\beta)$ with β^* and $M(\beta^*)$, this theorem indicates that, the squared approximation error of Lasso-type estimator is controlled, up to a logarithmic factor, by a bound that only relates to the number of non-zero components of oracle vector, instead of all components.

- Unrestricted True Model: $\exists\ \alpha > 0$ and $\beta \in \mathbb{R}^{(K_n+1)^d}$ such that

$$\sup_{x\in\mathcal{X}} \left| g(x) - \sum_{k_1=0}^{K_n} \cdots \sum_{k_d=0}^{K_n} \beta_{k_1,\ldots,k_d} \prod_{l=1}^{d} p_{k_l}(x_l) \right| = O(K_n^{-\alpha})$$

- Additive True Model: $\exists\ \alpha > 0$ and $\beta_1, \ldots, \beta_d \in \mathbb{R}^{K_n+1}$ such that

$$\sup_{x\in\mathcal{X}} \left| g(x) - \sum_{l=1}^{d} \sum_{k=0}^{K_n} \beta_{lk} p_k(x_l) \right| = O(K_n^{-\alpha})$$

- Parametric True Model: $\exists\ \beta \in \mathbb{R}^{(K+1)^d}$ such that

$$\sup_{x\in\mathcal{X}} \left| g(x) - \sum_{k_1=0}^{K} \cdots \sum_{k_d=0}^{K} \beta_{k_1,\ldots,k_d} \prod_{l=1}^{d} p_{k_l}(x_l) \right| = 0$$

- For a chosen basis function, we can verify assumptions for the theorem of sparsity oracle inequalities.
- By applying that theorem, we can derive mean squared convergence rates (up to logarithmic factors) of our dimension adaptive estimator.
  - Unrestricted True Model:

$$M(\beta) \asymp n^{\frac{1}{d+2\alpha}} \Rightarrow r_{n,M}^2 M(\beta) \asymp \log(M) n^{-\frac{2\alpha}{2\alpha+d}} \Rightarrow O_p(n^{-\frac{2\alpha}{2\alpha+d}})$$

  - Additive True Model:

$$M(\beta) \asymp n^{\frac{1}{1+2\alpha}} \Rightarrow r_{n,M}^2 M(\beta) \asymp \log(M) n^{-\frac{2\alpha}{2\alpha+1}} \Rightarrow O_p(n^{-\frac{2\alpha}{2\alpha+1}})$$

  - Parametric True Model:

$$M(\beta) = (K+1)^d \Rightarrow r_{n,M}^2 M(\beta) \asymp \log(M) n^{-1} \Rightarrow O_p(n^{-1})$$

- The estimator adapts its convergence rate to different true models!

- One more thing! To achieve the convergence rates above, $\pi_{n,M}(\beta)$ in the theorem needs to be asymptotically zero. As shown,

$$\pi_{n,M}(\beta) \leq 10M^2 \exp\left(-c_1 n \min\left\{r_{n,M}^2, \frac{r_{n,M}}{L}, \frac{1}{L^2}, \frac{\kappa_M^2}{L_0 M^2(\beta)}, \frac{\kappa_M}{L^2 M(\beta)}\right\}\right) + \exp\left(-c_2 \frac{M(\beta)}{L^2(\beta)} n r_{n,M}^2\right)$$

- It is sufficient that both parts go to zero. Then, for the first part, it is sufficient that all terms in $\min\{\cdot\}$ go to infinity; for the second part, it is proved that it goes to zero.
- For some basis functions, L and L_0 are unbounded and are associated to n and smoothness α; For others, they are some constants.

- (a) For normalized Legendre polynomials, $\underline{\alpha} > \frac{(2d-1)+\sqrt{8d^2+1}}{4}$, or a sufficient condition $\underline{\alpha} \geq \frac{3d-1}{2}$.
- (b) For orthonormalized B-splines, $\underline{\alpha} > \frac{d}{2}$, or a sufficient condition $\underline{\alpha} \geq \frac{d+1}{2}$.
- (c) For normalized Haar wavelets, the same as (a).
- (d) For normalized trigonometric polynomials, same as (b).

Simulation
Setup

- Three True Functions: for $x \in [0,1]^5$,

$$
\begin{aligned}
m_1(x) =& 3x_1 + 1.8x_2 + x_3 + 2.5x_4 + x_5 \\
m_2(x) =& \sin(4x_1) + 1.5\log(x_2) + \frac{1}{\cos(x_3)} + \sin(\sqrt{x_4}) + \sin(x_5^2) \\
m_3(x) =& 3\sqrt[4]{x_1 + 4x_2 + x_3x_4x_5} + 2\sin(x_4 + x_5^2 + x_1x_2x_3) \\
& + 3\log(x_3^2 + x_4 + 2x_5)
\end{aligned}
$$

- The n observations of type (X, Y) are generated with the following data generation process:

$$Y = m_i(X) + \sigma_j \cdot \epsilon \qquad (i \in \{1,2,3\}, j \in \{1,2\})$$

where X is uniformly distributed on $[0,1]^5$ and ϵ is standard normally distributed and independent of X. The parameters scaling the noise are $\sigma_1 = 5\%$ and $\sigma_2 = 20\%$.

Simulation
Setup

- We compare our dimension adaptive estimator (*dimada*) with two other estimators.
    - *dimada*: a series estimator with ℓ_1-regularization
    - *addt*: the same as *dimada* but with additive restriction
    - *ols*: OLS (parametric) estimator
- Out-of-sample empirical squared L_2 error is applied in our simulation to examine the performance. The size of train and test datasets are 400 and 1000 respectively.
- To account for the data generation randomness, the empirical squared L_2 errors are computed for 500 repeatedly generated realization of X. The medians are examined.

Median of out-of-sample empirical MSE for m_1 (parametric true model)

		Post LASSO		Post Adaptive LASSO	
Basis Function	Estimator	MSE	# Terms	MSE	# Terms
Legendre	*dimada*	0.00262	17	0.00254	5
	addt	0.00255	7	0.00254	5
	ols	**0.00254**	5	**0.00254**	5
B-Splines	*dimada*	0.01053	191	0.00917	107
	addt	0.00268	28	0.00268	25
	ols	**0.00258**	5	**0.00254**	5
Trigonometric	*dimada*	0.00614	159	0.00436	14
	addt	0.00661	12	0.00656	10
	ols	**0.00254**	5	**0.00254**	5

- *ols* outperforms the other two. But, *dimada* and *addt* perform almost as good as *ols*.

Simulation

Out-of-sample Empirical MSE ($\sigma_1 = 5\%$)

Median of out-of-sample empirical MSE for m_2 (additive true model)

		Post LASSO		Post Adaptive LASSO	
Basis Function	Estimator	MSE	# Terms	MSE	# Terms
Legendre	*dimada*	0.174	41	0.166	22
	addt	**0.144**	12	**0.146**	10
	ols	0.750	5	0.750	5
B-Splines	*dimada*	0.182	72	0.164	44
	addt	**0.084**	27	**0.084**	22
	ols	0.750	5	0.750	5
Trigonometric	*dimada*	0.202	172	0.185	105
	addt	**0.065**	24	**0.067**	19
	ols	0.750	5	0.750	5

- *ols* is quite bad.
- *addt* is the best, and *dimada* is almost as good.

Out-of-sample Empirical MSE ($\sigma_1 = 5\%$)

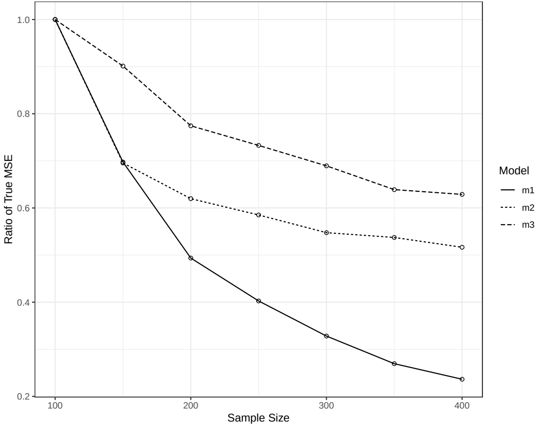
Median of out-of-sample empirical MSE for m_3 (unrestricted true model)

		Post LASSO		Post Adaptive LASSO	
Basis Function	Estimator	MSE	# Terms	MSE	# Terms
Legendre	*dimada*	**0.037**	32.0	**0.041**	24
	addt	0.257	10.0	0.257	9
	ols	0.314	5.0	0.314	5
B-Splines	*dimada*	**0.095**	177.0	**0.084**	121
	addt	0.262	27.0	0.264	24
	ols	0.314	5.0	0.314	5
Trigonometric	*dimada*	**0.032**	204.5	**0.032**	117
	addt	0.265	21.0	0.265	15
	ols	0.314	5.0	0.314	5

- *ols* and *addt* are bad.
- *dimada* outperform the other two.

Simulation

Convergence Rates of Dimension Adaptive Estimator ($\sigma_1 = 5\%$)



- The change of true MSE ratio with sample size.
- In parametric true model (m_1), *dimada* converges at the fastest rate of verifiable n^{-1}.

The End