Forest Guided smoothing

이동균

November 25, 2021

목차

Abstract

- Randomforest 출력값, Spatially adaptive bandwidth matrice를 사용하여 local smoother 를 정의.
- Smoother는 forest의 유연성을 가지면서 ,선형 Smoother이므로 해석하기가 쉬워진다.
- bias correction, confidence interval 등도 이용가능.

1. Introduction

Randomforest는 비모수 회귀 분석으로써 정확한 방법이지만 해석이 어렵다는 단점이 있다. 따라서, Standard error, confidence interval 등을 구하는데 어려움이 있다.
이 논문에서는 spatially adaptive local linear smoother를 구성하여 forest 의 값을 근사시킨다.

2. Forest-Guided Smoothers

$$(X_1, Y_1), ..., (X_n, Y_n) \sim P$$

where $Y_i \in R$ and $X_i \in R^d$, $\mu(x) = E(Y|X = x)$, 그리고 $d < n$ 가정할 때,

$$\hat{\mu}_{RF}(x) = \frac{1}{B} \sum_{j=1}^{B} \hat{\mu}_j(x)$$

2. Forest-Guided Smoothers

$\hat{\mu}_{RF}(x)$를 다른 임무에도 사용되는 다루기 쉬운 forest에 대한 근사치로 추정

$$\hat{\mu}_{RF}(x) = \sum_{i=1}^{n} w_i(x) Y_i$$

where $w_i(x) \geq 0$, $\sum_i w_i(x) = 1$

2. Forest-Guided Smoothers

데이터를 D_1 and D_2를 분할하고 각각의 사이즈가 n이라고 가정.
D_1 에서 bandwidth matrix 구성

$$H_x = \left(\frac{1}{n} \sum_i w_i(x)(X_i - x)(X_i - x)^T \right)^{1/2} \tag{1}$$

K는 spherically symmetric kernel 이고 아래와 같이 정의한다.

$$K(x; H_x) = |H_x|^{-1} K(H_x^{-1} x).$$

여기서 bandwidth matrices의 하나의 parameter family를 정의한다.
$\Xi = \{hH_x : h > 0, x \in R^d\}$

그리고 forest guided loce linear smoother (FGS) 정의한다. local linear smoother는 아래식을 최소화하는 $\hat{\mu}_h(x) = \hat{\beta}_0(x)$를 찾는다.

$$\sum_i \left(Y_i - \beta_0(x) - \beta(x)^T(X_i - x)\right)^2 K(X_i - x; hH_x).$$

2. Forest-Guided Smoothers

$$\hat{\mu}_h(x) = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x Y = \sum_i \ell_i(x; hH_x) Y_i$$

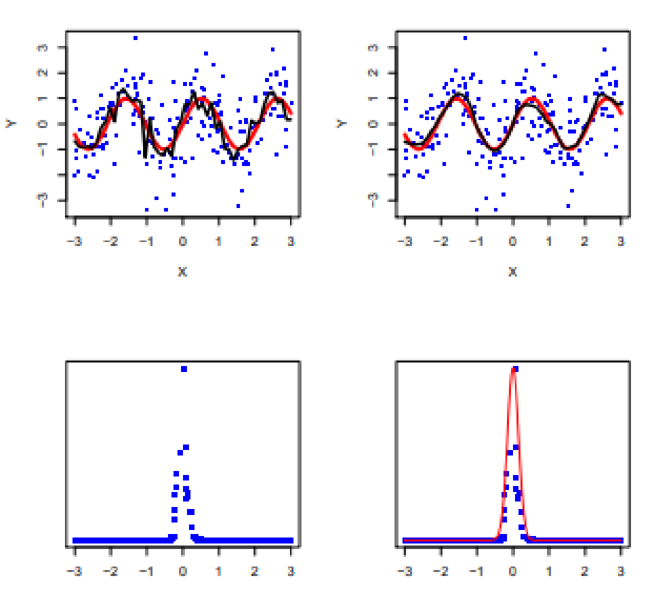
where

$$X_x = \begin{bmatrix} 1 & (X_1 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{bmatrix}$$

W_x 는 $W_x(i,i) = K(X_i - x; hH_x), e_1 = (1, 0, ..., 0)^T$ 인 대각행렬

$$\ell_i(x; hH_x) = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x \tag{2}$$

2. Forest-Guided Smoothers



$\sigma^2(x) = Var(Y|X = x)$ 를 추정해서 standard error를 구할 수 있다.

σ^2을 추정하는 방법으로는 forest 로부터 잔차 $r_i = Y_i - \hat{\mu}_R F(X_i)$ 를 구한 후, r_i^2을 반응변수, X_i를 설명변수로 한 Random forest 모델로 r_i^2을 추정하는 방법을 사용한다.

위에서 추정한 $\hat{\sigma}^2(x)$는 분산을 과소추정하는 경향이 있으므로 $\hat{\sigma}(x)$ 대신 $c\hat{\sigma}(x)$ 를 사용한다.

$H_x \equiv H_{n,x}$, $\mu_2(K)I = \int uu^T K(u)du$, and $R(K) = \int K^2(u)du$.

Assumptions

(A1) K is compactly supported and bounded. All odd moments of K vanish.
(A2) $\sigma^2(x)$ is continuous at x and f is continuously differentiable. Also, the second order derivatives of μ are continuous. Further, $f(x) > 0$ and $\sigma^2(x) > 0$
(A3) $H_{n,x}$ is symmetric and positive definite. As $n \to \infty$ we have $n^{-1}|H_{n,x}| \to 0$ and $H_{n,x}(i,j) \to 0$ for every i, j.
(A4) There exists c_λ such that

$$\frac{\lambda_{max}(H_{n,x})}{\lambda_{min}(H_{n,x})} \leq c_\lambda$$

3. Confidence Intervals - Properties of Smoothers

앞의 (A1)-(A4) 조건 하에서 bias와 vararariance를 아래와 같이 구한다.

$$B(x, H_x) = \frac{1}{2}\mu_2(K) tr(H_x^2 Hess(x)) + o_p(tr(H_x^2)) \tag{3}$$

$$V(x, H_x) = \frac{\sigma^2(x) R(X)}{n|H_x| f(x)}(1 + o_p(1)). \tag{4}$$

bias에 bandwidth matrix를 hH_x를 사용하면 아래의 식을 만족한다.

$$B(x, hH_x) = h^2 c_n(x) + o_p(h^2 tr(H_x^2)).$$

for some $c_n(x)$

3. Confidence Intervals - Properties of Smoothers

(A4) There exist a sequence $\phi_n \to 0$ and positive definite symmetric matrix C_x such that $H_{n,x} \sim \phi_n C_x$ where $\phi_n \asymp (1/n)^a$ for some $0 < a < 1$
(A4) 가정을 통해 $B(x, hH_x) = h^2 c(x)/n^2 + o_p(h^2)$ 으로 표현가능하다. 여기서 bias correction을 위해서는 더 강한 smoothness condition이 필요하다.

Assumptions

(A5) For some t, the t^{th} order derivatives of μ are continuous and there exist function $c_1(t), ...c_t(x)$ such that, for any $h > 0$,

$$B(x, hH_x) = \sum_{j=2}^{t} \frac{c_j(x)h^j}{n^{aj}} + o_p(\frac{1}{n^{at}})$$

Bias를 추정하기위해 아래와 같이 정의를 한다. b개의 bandwidth 선택 $h_1, h_2, ..., h_b$.
$\hat{m} = (\mu_{h_1}(x), ..., \mu_{h_b}(x))$.
$\kappa_n = (\mu(x), \kappa_{2,n}(x), ...\kappa_{t,n}(x))^T$ where $\kappa_{j,n}(x) = c_j(x)/n^{aj}$

$$H = \begin{bmatrix} 1 & h_1^2 & h_1^3 & \cdots & h_1^t \\ 1 & h_2^2 & h_2^3 & \cdots & h_2^t \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & h_b^2 & h_b^3 & \cdots & h_b^t \end{bmatrix}$$

$$\hat{\kappa}_n = argmin_c ||\hat{m} - Hc||^2 = (H^T H)^{-1} H^T \hat{m}.$$

3. Confidence Intervals - Properties of Smoothers

$\hat{m} = LY$ where

$$L = \begin{bmatrix} \ell_1(x;, h_1 H_x) & \ell_1(x;, h_1 H_x) & \cdots & \ell_n(x;, h_1 H_x) \\ \ell_1(x;, h_2 H_x) & \ell_2(x;, h_2 H_x) & \cdots & \ell_n(x;, h_2 H_x) \\ \vdots & \vdots & \vdots & \vdots \\ \ell_1(x;, h_b H_x) & \ell_2(x;, h_b H_x) & \cdots & \ell_n(x;, h_b H_x) \end{bmatrix}$$

$$\hat{\kappa}_n = (H^T H)^{-1} H^T LY.$$

$$\hat{B}(x, h) = \sum_{j=2}^{t} \hat{\kappa}_{j,n}(x) h^j = g^T (H^T H)^{-1} H^T LY$$

where $g = (0, h^2, ..., h^t)^T$.

$\hat{\kappa}_n$ 의 첫번째 요소는 μ 의 de-biased estimator 가 된다.

$$\mu^{\dagger}(x) = e_1^T (H^T H)^{-1} H^T L Y = \sum_i \tilde{\ell}_i(x)$$

where $\tilde{\ell}(x) = e_1^T (H^T H)^{-1} H^T L$.

$\mu^{\dagger}(x)$ variance 와 esimated variance 는 아래와 같이 구한다.

$$Var[\mu^{\dagger}(x)] = \sum_i \tilde{\ell_i^2}(x) \sigma^2(X_i)$$

$$s^2(x) = \sum_i \tilde{\ell_i^2}(x) \hat{\sigma}^2(X_i)$$

여기서 CLT에 적용하기 위해서 bandwidth를 더 구체적으로 줄 필요가 있다. bandwidth를 $h_j = \alpha_j n^{-\gamma}$, for j=1,2,...b, with $0 < \alpha_1 < \cdots < \alpha_b$ 로 정하고, n에 의존하지 않는다고 가정한다.

3. Confidence Intervals - Properties of Smoothers

Assumptions

Theorem 1 Assume that, conditional on D_1, assumptions (A1)-(A5) hold and:

(i) $sup_x|\hat{\sigma}^2(x) - \sigma^2(x)| \xrightarrow{p} 0,$

(ii) $-a < \gamma < \frac{1-ad}{d}$

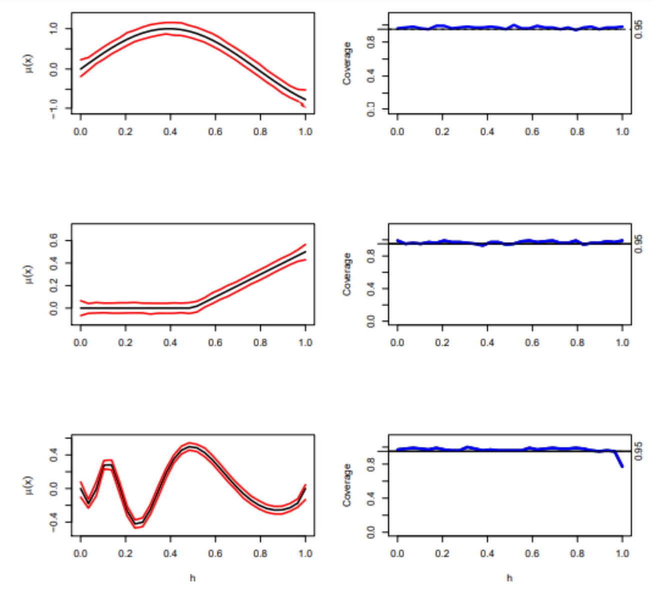
Further, if $t < d/2$ we require $a < 1/(d-2t)$. Also asuume that Y is bounded and that $b > t+1$. Then

$$\frac{\mu^\dagger(x) - \mu(x)}{s(x)} \xrightarrow{d} N(0,1).$$

$$P(\mu(x) \in C_n(x)) \to 1 - \alpha$$

where $C_n(x) = \mu^\dagger(x) \pm z_{\alpha/2}s(x)$

3. Confidence Intervals - Examples of Confidence Intervals



4. Exploring the Forest - Summarizing the Spatial Adaptivity of the Kernels

Wasserstein barycenter 개념을 도입한다.

먼저 두 분포 사이의 Warsserstein distance는 아래와 같이 구한다.

$$W_2^2(P_1, P_2) = \underset{J}{inf} E_J[||X - Y||^2]$$

J는 $X \sim P_1$와 $Y \sim P_2$의 Joint distribution

특이 케이스로 $P_1 = N(\mu_1, \sigma_1)$ and $P_2 = N(\mu_2, \sigma_2)$ 일 때,

$$W_2^2(P_1, P_2) = ||\mu_1 - \mu_2||^2 + tr(\Sigma_1) + tr(\Sigma_2) - 2tr\Big\{(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Big\}$$

4. Exploring the Forest - Summarizing the Spatial Adaptivity of the Kernels

Q_x의 Wasserstein barycenter는 아래의 식을 최소화(minimize)하는 distribution $\overline{Q}$이다.

$$\int W^2(Q_x, \overline{Q}) dP_X(x)$$

예를 들어, $N(\mu_1, 1)$ and $N(\mu_2, 1)$ 의 barycenter 는 $N((\mu_1 + \mu_2)/2, 1)$ 이다. barycenter는 기존의 분포의 모양을 보존하는 분포가 나온다.

4. Exploring the Forest - Summarizing the Spatial Adaptivity of the Kernels

이제 $K(0, H_{x_i})$ 들의 barycenter를 찾는 것이 우리의 목적이다.
barycenter 는 $K(0, \overline{H})$ 로 나오고, $\overline{H}$ 는 고유한 양정치 행렬 (unique positive definite matrix)

$$\overline{H} = \int (\overline{H}^{1/2} H_x \overline{H}^{1/2})^{1/2} dP_X(x) \tag{5}$$

$\overline{H}$를 구하면, Frechet variance 도 아래와 같이 구할 수 있다.

$$V = \int W^2(\overline{H}, H_x) dP_X(x) \tag{6}$$

4. Exploring the Forest - Comparing the Forest and the Smoother

Smoother를 쓰면서 얼마나 많은 Prediction accuracy의 손실을 보는지의 척도

$$\Gamma = E[(Y - \hat{\mu}(X))^2 - (Y - \hat{\mu}_{RF}(X))^2]$$

실제 데이터에서 계산은?

데이터를 각각 사이즈가 $m \approx n/4$ 인 그룹 D_1, D_2, D_3, D_4 로 분할한 후, D_1에서 $\hat{\mu}_{RF}$, D_2에서 $\hat{\mu}$ 를 추정한다.

$$\hat{\Gamma} = \frac{1}{m}\sum_{i \in D_3} r_i - \frac{1}{m}\sum_{i \in D_4} s_i$$

where $r_i = (Y_i - \hat{\mu}_{RF}(X_i))^2$, $s_i = (Y_i - \hat{\mu}(X_i))^2$

4. Exploring the Forest - Comparing the Forest and the Smoother

$$\sqrt{m}(\hat{\Gamma} - \Gamma) \xrightarrow{d} N(0, \tau^2)$$

τ^2 의 일치 추정량 $\hat{\tau} = m^{-1}(\sum_i (r_i - \overline{r})^2 + \sum_i (s_i - \overline{s})^2)$

Γ 의 confidence interval :

$$\hat{\Gamma} \pm z_{\alpha/2}\hat{\tau}/\sqrt{m}$$

4. Exploring the Forest - Multiresolution Local Variable Importance

local variable importance를 평가하는 가장 알려져 있는 방법은 μ의 gradient를 추정하는 방법, 즉 local linear approximation을 사용하는 방법이다.
forest guided local linear smoother를 사용하여 gradient와 standard error를 추정할 수 있었다.

$$\beta_{h,j}(x) = \sum_i Y_i \ell_{ij}(x; hH_x)$$

where $\ell_{ij}(x; hH_x)$는 벡터 $e_{j+1}^T (X_x^T W_x X_x)^{-1} X_x W_x$의 i번째 요소

4. Exploring the Forest - Multiresolution Local Variable Importance

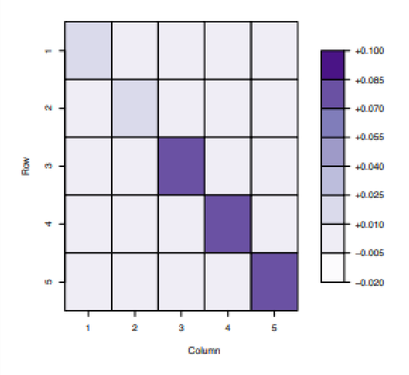
$\hat{\beta}_{h,j}$ 의 standard error

$$se_{h,j}(x) = \sqrt{\sum_i \hat{\sigma}^2(X_i)\ell_{ij}(x; hH_x)}$$

$1 - \alpha$ variability interval:

$$\hat{\beta}_{h,j}(x) \pm z_{\alpha/2} se_{j,h}(x)$$

5. Examples - Synthetic Exmaple



(a) Barycenter of bandwidth matrix

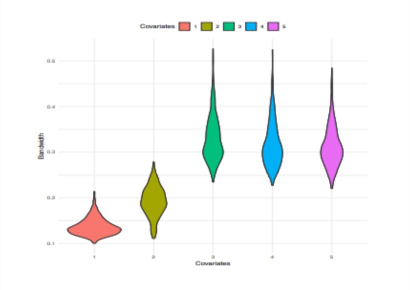


Figure 5: Example (6). Effective bandwidths for each covariate.

(b) Effective bandwidths

5. Examples - Synthetic Exmaple

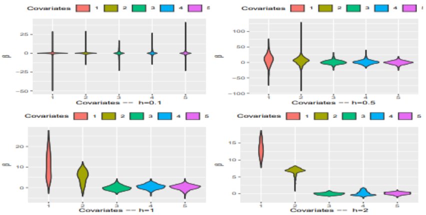


Figure 6: $\{\hat{\beta}_j(X_1), \ldots, \hat{\beta}_j(X_n)\}$ for each covariate at four resolutions, $h = 0.1, 0.5, 1, 2$.

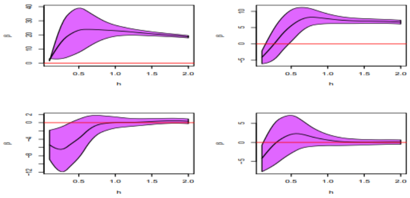


Figure 7: Variability intervals for $\hat{\beta}_{1,h}(x), \cdots, \hat{\beta}_{4,h}(x)$ at $x = (1/2, 1/2, 1/2, 1/2, 1/2)$.

5. Examples - Synthetic Exmaple