NONPARAMETRIC REGRESSION AND WHITE NOISE MODEL

YAN LIU

1. Reference

Nussbaum (1996), AS.
Brown and Low (1996), AS.
Markus Reiss (2008), AS.
Stone (1980), AS
Stone (1982), AS.
Yu and Jones (1998), JASA

2. Notations

1. $(X,Y) \in \mathbb{R}^d \times \mathbb{R}$	a pair of random variables
2. θ	$E(Y\|X) = \theta(X)$
3. k	a nonnegative integer
4. Θ	the collection of k-times continuously differentiable function θ
5. U	an open interval containing $[0,1]$
6. $p > 0$	p-times continuously differentiable on U
7. $\hat{\theta}_n$	estimator
8. $\|\cdot\|$	the Euclidean norm of $\cdot$
9. $\sharp(\cdot)$	the number of elements in $\cdot$
10. δ_n	a seq of positive constants which tend to 0
11. $\mathcal{N}_n(x)$	$= \{i : 1 \leq i \leq n \text{ and } \|X_i - x\| \leq \delta_n\}$
12. $N_n(x)$	$= \sharp(\mathcal{N}_n(x))$
13. $C \subset \mathbb{R}^d$	a compact subset having a nonempty interior
14. q	$\in (0, \infty]$
15. $\\|\cdot\\|_q$	the L^q norm
16. b_n	a seq of eventually positive constants
17. J	an open interval
18. $t \in J$	an unknown real-valued mean parameter of the distribution
19. $\underline{\alpha}$	$= (\alpha_1, \dots, \alpha_d)$
20. $[\alpha]$	$= \alpha_1 + \cdots + \alpha_d$
21. D^α	the differential operator

Date: December 2, 2014.

22. $T(\theta)$	a linear combination with constant coefficients of $D^\alpha\theta$, $[\alpha] \le k$
23. q_α	real constants for $[\alpha] \le k$
24. Q	$= \sum_{[\alpha]\le k} q_\alpha D^\alpha$
25. m	the order of Q, $m = \max([\alpha] : [\alpha] \le k \,\text{and}\, q_\alpha \ne 0)$
26. r	$= (p-m)/(2p+d)$

3. Concepts and definitions

3.1. **Parameter space.** Let $0 < \beta \le 1$ and $0 < K_2 < \infty$. For $x \in \mathbb{R}^d$ and $\theta(x) \in J$,

$$|D^\alpha\theta(x) - D^\alpha\theta(x_0)| \le K_2|x - x_0|^\beta \quad \text{for } x_0, x \in U \text{ and } [\alpha] = k.$$

Note that

$$p = k + \beta$$

is a measure of the smoothness of the functions in Θ.

3.2. **Estimators.**

3.2.1. *response and variable.* θ is the regression function of the response Y on the measurable variable X. $E(Y|X) = \theta(X)$.

3.2.2. *parametric.* If $\hat\theta_n \in \Theta$ for all $n \ge 1$, where Θ is a collection of functions which are defined in terms of a finite number of unknown parameters.

3.2.3. *nonparametric.* Otherwise.

3.3. **Rate of convergece.**

3.3.1. *lower rate of convergence.* If there is a $c > 0$ such that

$$\lim_n \inf_{\hat T_n} \sup_\Theta P_\theta(\|\hat T_n - T(\theta)\|_q \ge cb_n) = 1,$$

where the infimum is over all possible estimator $\hat T_n$.

3.3.2. *achievable rate of convergence.* If there is a sequence $\{\hat T_n\}$ of estimators and a c_0 such that

$$\boxed{\texttt{eq:2.2.1}} \qquad \lim_n \sup_\Theta P_\theta(\|\hat T_n - T(\theta)\|_q \ge cb_n) = 0. \tag{3.1}$$

3.3.3. *optimal rate of convergence.* If it is both a lower and an achievable rate of convergence.

3.3.4. *asymptotically optimal.* If $\{b_n\}$ is the optimal rate of convergence and $\{\hat T_n\}$ satisfies (3.1), the estimators $\hat T_n$, $n \ge 1$, is said to be asymptotically optimal.

3.4. **Assumptions.**

⟨asp:2.2.1⟩ **Assumption 3.1.** Suppose $l(y|x,t) = \log h(y|x,t)$.

(1) As a function of t, h is strictly positive and continuously differentiable.

(2) The equation

$$\int h(y|x,t)\phi(dy) = 1$$

can be differentiated with respect to t to yield

$$\int h'(y|x,t)\phi(dy) = 0$$

and

$$\int h''(y|x,t)\phi(dy) = 0.$$

(3) There are positive constants ϵ_0 and K_1 and there is a function $M(y|x,t)$ such that on the indicated domain

$$|l''(y|x,t+\epsilon)| \leq M(y|x,t) \quad \text{for } |ep| \leq ep_0$$

and

$$\int M(y|x,t)h(y|x,t)\phi(dy) \leq K_1.$$

Remark 3.2. The condition is needed to verify that $\{b_n\}$ is a lower convergence sequence with the assumption that C have a nonempty interior.

⟨asp:2.2.3⟩ **Assumption 3.3.** For some $s > 0$,

$$\int e^{s|y-t|}h(y|x,t)\phi(dy)$$

is bounded for $x \in U$ and $t \in J$.

Remark 3.4. The condition is required to verify that with compactness of C certain rates of convergence are achievable and certain estimators are asymptotically optimal.

⟨asp:2.2.5⟩ **Assumption 3.5.** For every $\lambda \in (0, 1/d)$ and $c > 0$, there is a $c' > 0$ such that

$$\lim_n P(\sharp\{i : 1 \leq i \leq n \, \text{and} \, |X_i - x| \leq cn^{-\lambda}\} \geq c'n^{1-\lambda d} \, \text{for all} \, x \in U) = 1.$$

Remark 3.6. The condition on the asymptotic distribution of $X_1, \dots, X_n$ is required to guarantee achievability and asymptotic optimality.

If U is a polyhedron (polytope?), this condition is implied by the following one.

⟨asp:2.2.7⟩ **Assumption 3.7.** The random variables $X_1, \dots, X_n$ are the first n terms of an i.i.d. sequence of random variables each having distribution F, the density of whose absolutely continuous component is bounded away from 0 on U.

3.5. **Results.**

⟨thm:2.2.8⟩ **Theorem 3.8.** *Under Assumptions 3.1, 3.3 and 3.5,*

- *If* $0 < q < \infty$*, then* $\{n^{-r}\}$ *is the optimal rate of convergence;*
- *If* $q = \infty$*, then* $\{(n^{-1}\log n)^r\}$ *is the optimal rate of convergence.*

Corollary 3.9. Suppose $d = 1$. The estimators $\hat{\theta}_n$, $n \geq 1$, are said to be *asymptotically optimal* if

$$n^{-2p/(2p+1)} \int_0^1 \{\hat{\theta}_n(x) - \theta(x)\}^2 dx$$

is bounded in probability as $n \to \infty$.

Proof. Note that $q = 2$. As the statement of Theorem 3.8, we see that the optimal rate is n^{-r} where

$$r = \frac{p - m}{2p + d}.$$

Here, $m = 0$ and $d = 1$ with L^2 norm, the optimal rate is

$$n^{-2p/(2p+1)}.$$

□

Example 1 (Spiegelman and Sacks (1980))**.** Let $\hat{\theta}_n$ be the kernel estimator.

$$\hat{\theta}_n = \frac{1}{N_n(x)} \sum_{\mathcal{N}_n(x)} Y_i.$$

- If $p = 1$ and $\delta_n = n^{-1/3}$, then $\hat{\theta}_n$ is asymptotically optimal.
- If $p = 2$, $\delta_n = n^{-1/5}$ and f is absolutely continuous and f' is square integrable on U, then $\hat{\theta}_n$ is asymptotically optimal.

3.5.1. *Without any smoothness assumption of f.*

- $\hat{P}_n(\cdot; x)$ the polynomial of degree $p - 1$ which minimizes

  $$\sum_{\mathcal{N}_n(x)} \{Y_i - \hat{P}_n(X_i; x)\}^2,$$

  and then for bounded from zero and infinity f,

  $$\hat{\theta}_n(x) = \hat{P}_n(x; x).$$

- Or without the assumption of boundedness of f if $\hat{P}_n(\cdot; x)$ minimizes

  $$\sum_{\mathcal{N}_n(x)} \frac{\{Y_i - \hat{P}_n(X_i; x)\}}{W_n(X_i; x)},$$

  where $W_n(\cdot; x)$ is an appropriate positive weight function.

Remark 3.10. $\{n^{-r}\}$ is also the optimal rate of convergence if $\|\hat{T}_n - T(\theta)\|_q$ (as function) is replaced by $|\hat{T}_n(x_0) - T(x_0; \theta)|$ (pointwise).

4. QUESTIONS

Suppose $q = 2$.

- Let Assumption 3.7 holds. Under which additional conditions on F, is $\{n^{-r}\}$ an achievable rate of convergence?
- Let $\mathcal{A}$ denote either the collection of functions θ on $\mathbb{R}^d$ which are additive
$$\theta(x_1, \cdots, x_d) = \theta_1(x_1) + \cdots + \theta_d(x_d)$$
or the collection of the form
$$\theta(x_1, \cdots, x_d) = \psi(\beta_1 x_1 + \cdots + \beta_d x_d).$$
Suppose $\Theta = \mathcal{A} \cap \Theta$ and set $r_1 = (p - m)/(2p + 1)$. Is $\{n^{-r_1}\}$ an achievable rate of convergence?
- Suppose that t is the median of the distribution instead of its mean. Is $\{n^{-r}\}$ still an achievable rate of convergence?

5. LOCAL LINEAR QUANTILE REGRESSION

a regression quantile as the minimizer of
$$E\{\rho_p(Y - \theta)|X = x\}$$

- The check function
$$\rho_p(z) = pz\mathbb{1}_{[0,\infty)}(z) - (1 - p)z\mathbb{1}_{(-\infty,0)}(z).$$
- Linear fitting
$$S_{n,l} = \sum_{i=1}^{n} K(\frac{x - X_j}{h_1})(x - X_i)^l, \quad l = 1, 2$$
- the wight function
$$\omega_j(x; h_1) = K(\frac{x - X_j}{h_1})[S_{n,2} - (x - X_j)S_{n,1}]$$
- the double-kernel quantile estimator $\tilde{q}_p$ to solve
$$p = \frac{1}{\sum_j \omega_j(x; h_1)} \sum_j \omega_j(x; h_1)\Omega(\frac{\tilde{q}_p(x) - Y_j}{h_2})$$

6. NOTATIONS

6.1. nonparametric regression.

1. $I \subseteq \mathbb{R}$	a possibly infinite interval
2. $f(\cdot) : I \to \mathbb{R}$	
3. $\sigma^2(\cdot) : I \to (0, \infty)$	
4. $H : \mathbb{R} \to [0, 1]$	an increasing c.d.f.
5. (X_{ni}, Y_{ni}), $i = 1, \ldots, n$	observation

(6) deterministic

$$x_{ni} = H^{-1}(i/(n+1)), \quad i = 1, \dots, n$$

(7) random

$$X_{ni} \sim \text{i.i.d.}\, H, \quad i = 1, \dots, n$$

(8) The conditional distribution of Y given X

$$Y_{ni} = f(x_{ni}) + \sigma(x_{ni})\epsilon_{ni}, \quad \epsilon \sim \text{i.i.d.}\mathcal{N}(0,1), i = 1, \dots, n$$

9. Θ the parameter space consists of f

6.2. **white noise.**

(1) $0 \in I$

2. $\{B_t : t \in I\}$ Brownian motion

(3) the Gaussian process with

$$\begin{aligned} dZ_t^{(n)} &= \mu(t)dt + \lambda(t)dB_t/\sqrt{n} \\ dY_t^{(n)} &= \nu(t)dt + \lambda(t)dB_t/\sqrt{n} \end{aligned}$$

(4) $\mu \in \Theta$

5. $g_X^{(n)}$ the probability density for X with respect any dominating measure ξ

(6) the disparity L_1

$$L_1 = \int |g_Y^{(n)}(\omega) - g_Z^{(n)}(\omega)|\xi(d\omega)$$

Remark 6.1. Under the following additional conditions,

(1) H absolutely continuous with t.
(2) $h = dH/dt$ is $h > 0$ a.e. on I.
(3) Define

$$V_\tau^{(n)} = Z_{H^{-1}(\tau)}^{(n)}.$$

Then,

$$dV_\tau^{(n)} = \mu^*(\tau)d\tau + \lambda^*(\tau)dB_\tau/\sqrt{n},$$

where

$$\mu^*(\tau) = \frac{\mu(H^{-1}(\tau))}{h(H^{-1}(\tau))}, \quad \lambda^{*2}(\tau) = \frac{\lambda^2(H^{-1}(\tau))}{h(H^{-1}(\tau))}.$$

As a result, from Remark 2.1 and 2.1.1. and 2.1.3., we can without loss of generality assume

$$I = [0,1].$$

6.3. statistical equivalence.

1. $\mathscr{P}^{(1)}$, $\mathscr{P}^{(2)}$	two statistical problems
2. $\mathscr{X}^{(1)}$, $\mathscr{X}^{(2)}$	two sample spaces
3. $\{G_\Theta^{(i)} : \theta \in \Theta\}$	the respective families of distributions
4. $\mathcal{A}$	action space
5. $L : \Theta \times \mathcal{A} \to [0, \infty)$	loss function
6. $\delta^{(i)}$	generic symbol for a decision procedure in ith problem
7. $R^{(i)}(\delta^{(i)}, L, \theta)$	The risk from using procedure

$$\|L\| = \sup\{L(\Theta, \alpha) : \theta \in \Theta, \alpha \in \mathcal{A}\}. \tag{7}$$

Definition 6.2 (Le Cam's metric)**.**

$$\Delta(\mathscr{P}^{(1)}, \mathscr{P}^{(2)}) = \max\Big[\inf_{\delta^{(1)}} \sup_{\delta^{(2)}} \sup_{\theta} \sup_{L:\|L\|=1} |R^{(1)}(\delta^{(1)}, L, \theta) - R^{(2)}(\delta^{(2)}, L, \theta)|,$$
$$\inf_{\delta^{(2)}} \sup_{\delta^{(1)}} \sup_{\theta} \sup_{L:\|L\|=1} |R^{(1)}(\delta^{(1)}, L, \theta) - R^{(2)}(\delta^{(2)}, L, \theta)|, \Big].$$

Definition 6.3 (Asymptotically equivalent)**.**

$$\lim_{n\to\infty} \sup_{\theta} \sup_{L:\|L\|=1} |R^{(1)}(\delta_n^{(1)}, L, \theta) - R^{(2)}(\delta_n^{(2)}, L, \theta)| = 0.$$

Lemma 6.4. *Let $\mathscr{X}$ and its σ-field be a Polish space with its associated Borel field. Let $\mathscr{P}^{(1)}$ denote an experiment with sample space $\mathscr{X}$. Let $S : \mathscr{X} \to \mathscr{Y}$ be a sufficient statistic and let $\mathscr{P}^{(2)}$ denote the experiment in which $Y = S(X)$ is observed. Then $\Delta(\mathscr{P}^{(1)}, \mathscr{P}^{(2)}) = 0$.*

Theorem 6.5.

$$|R^{(1)}(\delta^{(1)}, L, \theta) - R^{(2)}(\delta^{(2)}, L, \theta)| \le L_1(\mathscr{P}^{(1)}, \mathscr{P}^{(2)})\|L\|$$

Define the Hellinger metric $H(G^{(1)}, G^{(2)})$ is defined by

$$H^2(G^{(1)}, G^{(2)}) = \int (g^{(1)}(x)^{1/2} - g^{(2)}(x)^{1/2})^2 \xi(dx).$$

Lemma 6.6.

$$L_1(G^{(1)}, G^{(2)}) \le 2H(G^{(1)}, G^{(2)}).$$

Also, for product measures,

$$H^2(G^{(1)}, G^{(2)}) = 2\Big[1 - \prod_{j=1}^{m} \Big[1 - \frac{H^2(G^{(1)}, G^{(2)})}{2}\Big]\Big]$$

For two normal distributions,

$$H^2(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = 2\Big[1 - \Big[\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}\Big]^{1/2} \exp\Big[-\frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)}\Big]\Big].$$

For two multivariate normal distributions with the same mean satisfies

$$H^2(\mathcal{N}(\mu), \alpha\, Q), N(\mu, \alpha\, \mathrm{Id}_{\mathbb{R}^\mathrm{n}}) \leq 2\|\mathrm{Q} - \mathrm{Id}_{\mathbb{R}^\mathrm{n}}\|_{\mathrm{HS}}^2, \quad \mathrm{Q} \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}, \alpha > 0.$$

Remark 6.7.

$$\|f - \mathscr{P}_n f\|_{L^2}^2 = \|f - P_n f\|_{L^2}^2 + \|P_n f - \mathscr{P}_n f\|_{L^2}^2,$$

which shows the distance by classical distance of the Le Cam and the distance caused by the relation between the dimension and wavelets. Or equivalently, canceling the distance can be regarded as

$$\lim_{n\to\infty} \frac{n}{\sigma^2}\Big(\sum_{j=1}^{n}(Ey_j - Ez_j)^2 + \sum_{j=n+1}^{\infty}(Ez_j)^2\Big) = 0.$$

7. WAVELETS

7.1. classification.

Isometric approximation: the Fourier approximation, the Haar wavelet
Isomorphic approximation: B-spline

7.2. Isometric approximation.

Define

$$\mathcal{F}_{\mathrm{S,per}}^d(s, R) = \Big\{f \in L^2([0,1]^d)\Big|\sum_{l\in\mathbb{Z}^2}|l|_\infty^{2s}|\langle f, \varphi_l\rangle|_{L^2}^2 \leq R^2\Big\}.$$

Suppose

$$Z := (D_n|s_n)^{-1}Y.$$

Theorem 7.1 (Reiss (2008))**.** *For d-dimensional periodic Sobolev classes $\mathcal{F}_{\mathrm{S,per}}^d(s, R)$ with regularity $s > d/2$ and equidistant design on the cube $[0,1]^d$, the nonparametric regression experiment $\mathbb{E}_n^d$ and the Gaussian shift experiment $\mathbb{G}_n^d$ are asymptotically equivalent as $n \to \infty$. The Le Cam distance satisfies*

$$\Delta_{\mathcal{F}_{\mathrm{S,per}}^d(s,R)}(\mathbb{E}_n^d, \mathbb{G}_n^d) \lesssim \sigma^{-1}Rn^{1/2-s/d}.$$

7.3. Isomorphic approximation.

Define

$$\mathcal{F}_{\mathrm{S}}^d(s, R) = \Big\{f \in H^s([0,1]^d)\Big|\|f\|_{H^s} \leq R\Big\},$$

where $\|\cdot\|_{H^s}$ denotes the standard L^2-Sobolev norm of regularity s on $[0,1]^d$.

Theorem 7.2 (Reiss (2008))**.** *For general d-dimensional Sobolev classes $\mathcal{F}_{\mathrm{S}}^d(s, R)$ with regularity $s > d/2$ and equidistant design on the cube $[0,1]^d$, the nonparametric regression experiment $\mathbb{E}_n^d$ and the Gaussian white noise experiment $\mathbb{G}_n^d$ are asymptotically equivalent as $n \to \infty$. The Le Cam distance satisfies*

$$\Delta_{\mathcal{F}_{\mathrm{S}}^d(s,R)}(\mathbb{E}_n^d, \mathbb{G}_n^d) \lesssim \sigma^{-1}Rn^{1/2-s/d}.$$

7.4. **Random design.** Suppose

$$Z_r := \sum_{j=1}^{n_0} \langle Y, \varphi_j^n \rangle_n \varphi_j^n + \sum_{j=n_0+1}^{n} \langle Y, \varphi_j^n \rangle_n \varphi_j.$$

Theorem 7.3 (Reiss (2008))**.** *For d-dimensional periodic Sobolev classes $\mathcal{F}^d_{\mathrm{S,per}}(s,R)$ with regularity $s > d/2$, the nonparametric regression experiment $\mathbb{E}^d_{n,r}$ with random design and the Gaussian shift experiment $\mathbb{G}^d_n$ are asymptotically equivalent as $n_0, n \to \infty$ and $n_0 = o(n^{1/2})$. The Le Cam distance satisfies*

$$\Delta_{\mathcal{F}^d_{\mathrm{S,per}}(s,R)}(\mathbb{E}^d_{n,r}, \mathbb{G}^d_n) \lesssim n^{-1/2} n_0 + \sigma^{-1} R n_0^{1/2-s/d}.$$

8. words

1. a possibly infinite interval	無限を含んでもよい区間
2. implicit	暗黙の
3. felicitous	適切な
4. alleviate	軽減する
5. conspicuous	顕著な
6. allude	言及する
7. trait	特徴
8. proximity	近接
9. an intriguing example	興味をそそる例
10. anthropometric	人体測定の
11. tricep	三頭筋
12. skinfold measurement	皮下脂肪測定
13. Gambia villages	ガンビアの村
14. steady	一定の、不変の
15. seminal	先駆的で重要な
16. ad hoc	特別な、臨時の
17. reside	備わっている
18. interest center on	

9. Further Reading

9.1. **nonparametric regression.** For estimating the whole function, see Naussbaum (1985), Speckman (1985), Donoho, Liu and MacGibbon (1990) and Golubev and Nussbaum (1990).

For estimating a point functional such as $f(x_0)$, see Ibragimov and Hasminskii (1982), Donoho and Liu (1991) and Donoho and Low (1992).

For estimating a nonlinear functional in nonparametric regression (and white noise model), see Donoho and Nussbaum (1990).

The minimax sequence based on the regression model is found in Golubev (1987, 1991). For the problem of estimating a linear functional in regression (and white noise), see Lepskii (1991).

The examples of estimating optimal, possibly random, bandwidths in the regression (and also white-noise) are given in Hall and Johnstone (1992).

9.2. **white noise model.** For estimating the whole function, see Pinsker (1980).

For estimating a point functional such as $f(x_0)$, see Ibragimov and Hasminskii (1984).

For estimating a nonlinear functional in white noise model, see Fan (1991b).

For estimating the whole function based on indirect observations, see Fan (1991a).

The minimax sequence based on the white noise model is found in Efroimovich and Pinsker (1984).