Non-parametric Regression

Tan-Jianbin

School of Mathematics
Sun Yat-sen University

Seminar on Statistics 105c

Table of Contents

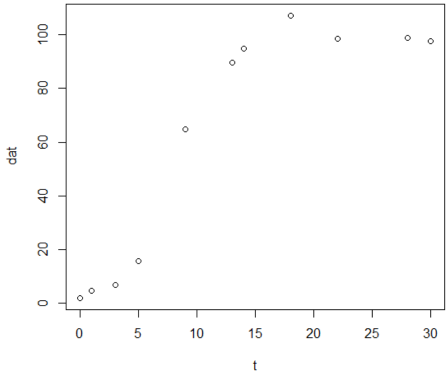
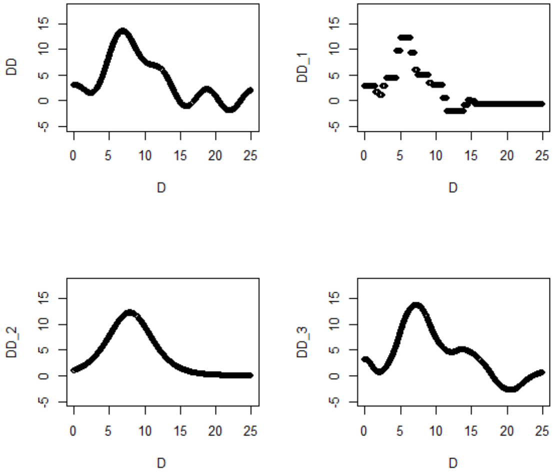


Figure: $f(t) = \frac{KN_0}{N_0+(K-N_0)e^{-rt}} + 5\sin(t) + N(0,5)$


DD
DD_1
DD_2
DD_3
D
D
D
D
-5
0
5
10
15
0
5
10
15
20
25

Definition

$(Y_t)_{t\in E}$ is a random process and $|E|$ is finite. H_1 is a Hilbert space and $T_t \in H_1^*$. $m \in H_1$ is a fix object and ε_t is iid mean zero noise:

$$Y_t = T_t m + \varepsilon_t$$

Statistical model above is called functional linear regression model.

This is only one of type of functional linear regression model with scaler respond and functional predictors.

Example

H_i is Euclidean space equipped with two norm: $Y = X\beta + \varepsilon$.
H_1 is $L^2(E_1)$: $Y_t = \int X_t(s)\beta(s)\mu(ds) + \varepsilon_t$.
If $E_1 = E_2$: $Y_t = \int R(t,s)\beta(s)\mu(ds) + \varepsilon_t$.
H_1 is $H(K)$ on E_1: $Y_n = \beta(t_n) + \varepsilon_n$.

We view Y_t and ε_t are functions of t and let $Y_t, \varepsilon_t \in H_2$, which is a Hilbert space of functions on E. Let $\mathcal{Y}(\omega) = Y_t(\omega)$, $\varepsilon(\omega) = \varepsilon_s(\omega)$ and $T \in B(H_1, H_2)$:

$$\mathcal{Y} = Tm + \varepsilon$$

Table of Contents

Definition

If V_i two Banach spaces, $U \subset V_1$ is open and $f: U \to V_2$. We call f is Gateaux differentiable at $x \in U$ if $\exists\ T \in B(V_1, V_2)$, $\forall\ v \in V_1$ s.t.

$$\lim_{t\to 0} \frac{||f(x+tv)-f(x)-t(Tv)||_2}{t} = 0$$

We mark $T = f^{'}(x)$. If $f^{'}(x)$ exists, it's necessarily unique.
If $\lim_{v\to 0} \frac{||f(x+v)-f(x)-Tv||_2}{||v||_1} = 0$, we call f is Frechet differentiable at $x \in U$.

Theorem

If f is Gateaux differentiable at an open U, and $f^{'}: V_1 \to B(V_1, V_2)$ is continuous at U, then $f^{'}$ is the Frechet derivative of $x \in U$.

Example

Let $V_1 = H$, $V_2 = \mathcal{R}$, $f(\beta) = \langle \alpha, \beta \rangle$, $\alpha \in H$. *Then* $f'(\beta) = \langle \cdot, \alpha \rangle$ *since* $\lim_{v \to 0} \frac{|f(\beta+v)-f(\beta)-\langle \alpha,v \rangle|}{||v||_1} = \lim_{v \to 0} \frac{|\langle \alpha,v \rangle - \langle \alpha,v \rangle|}{||v||_1} = 0$.

Let $f(\beta) = \langle \beta, T\beta \rangle$, $T \in B(H)$ *self-adjoint, then* $f'(\beta) = 2\langle \cdot, T\beta \rangle$.

Theorem

If $f : V \to \mathcal{R}$ *is Gateaux differentiable over* V *and* $x \in V$ *is a local extreme value, then* $f'(x) \in V^*$ *is zero.*

Definition

A collection $\mathcal{F}$ of Ω is said to be a σ-algebra:
(a) $\Omega, \emptyset \in \mathcal{F}$.
(b) If $A \in \mathcal{F}$, $A^c \in \mathcal{F}$. *(Algebra)*
(c) If $\{B_n\} \subset \mathcal{F}$, *then* $\cup_n B_n \in \mathcal{F}$.

A collection $\mathcal{G}$ of Ω can generate many σ-algebra, and $\exists$ a smallest σ-algebra of Ω that contains $\mathcal{G}$, we denote that $\sigma(\mathcal{G})$.

Property

If $\mathcal{G}$ *is* π *system, then* $\sigma(\mathcal{G}) = \lambda(\mathcal{G})$.
Moreover, $\mathcal{G}$ *is an algebra, then* $\sigma(\mathcal{G}) = \mathcal{M}(\mathcal{G})$.

Definition

$(\Omega, \mathcal{F}, \mu)$ is said to be a measurable space, if $\mathcal{F}$ is a σ-algebra, and the function $\mu : \mathcal{F} \to [0, \infty]$:

(a) $\mu(A) \geq \mu(\emptyset) = 0, \ \forall \ A \in \mathcal{F}$.

(b) $\{A_n\}$ disjoint, then $\mu(\cup_n A_n) = \sum_n \mu(A_n)$.

If $\mu(\Omega) = 1$, we call μ a probability measure. If $\exists \ A_n \uparrow \Omega$ s.t. $\mu(A_n) < \infty$, we call μ a σ-finite measure.

Theorem

If $\mathcal{G}$ is an algebra, a σ-finite measure defined on $\mathcal{G}$ can be uniquely extended to $\sigma(\mathcal{G})$.

For a metric space M, we denote $\mathcal{B}(M)$: $\sigma(\mathcal{G})$, $\mathcal{G}$ is the collection of all the open sets of M.

Theorem

Let $\mathcal{G}_1$ be the collection that contains the all sets like $(a_1, b_1] \times ... \times (a_q, b_q]$, and $\mathcal{G}_2 = \{\cup_n A_n$; some finite disjoint $A_n \in \mathcal{G}_1$ $\} \Rightarrow \mathcal{G}_2$ is an algebra and $\sigma(\mathcal{G}_2) = \mathcal{B}(R^p)$.

This theory implies we can define measure μ in $\mathcal{G}_1$: $\mu(A) = \sum_k (F(b_k) - F(a_k))$, then $\forall\ B \in \mathcal{G}_2$, $\mu(B) = \sum_n \mu(A_n)$, then we can extend μ to $\mathcal{B}(R^q)$. F is called Stieltjes measure function and if $F(x) = x$, then μ is Lebesgue measure.

Definition

$f : (\Omega_1, \mathcal{F}_1) \to (\Omega_2, \mathcal{F}_2)$ *is called a measurable map if* $\forall\ A \in \mathcal{F}_2,\ f^{-1}(A) \in \mathcal{F}_1$.

If $\mathcal{F}_2 = \sigma(\mathcal{G})$, then we just need to check all the sets in $\mathcal{G}$. Moreover, if $\Omega_2 = M$, and $\mathcal{F}_2 = \mathcal{B}(M)$, then we just need to check all the open sets in M.

If $(\Omega_1, \mathcal{F}_1, P)$ is probability space, f is measurable, we call f a random element. If $\Omega_2 = \mathcal{R}$, we call f random variable.

The distribution of f: $F_f = P \circ f^{-1}$, which is a probability measure: $\mathcal{F}_2 = \sigma(\mathcal{G})$ $\to [0, 1]$. And it's uniquely determined by F_f which domain is $\mathcal{G}$, if $\mathcal{G}$ is π-system.

Definition

V Banach space, $(\Omega, \mathcal{F}, \mu)$ measure space, $g_i \in V$, $E_k \in \mathcal{F}$, $f : \Omega \to V$ is called a simple function if $f(\omega) = \sum_k I_{E_k}(\omega) g_i$.

If $\mu(E_k) < \infty$, we can easily define the integration of f:

$$\int_\Omega f d\mu = \sum_k \mu(E_k) g_i.$$

We should check the ambiguity of the definition. More, $|| \int f d\mu|| \leq \int ||f|| d\mu$, for simple f.

Definition

A measurable map $f\colon (\Omega, \mathcal{F}) \to (V, \mathcal{B}(V))$ *is integrable if* $\exists$ *a sequence of simple functions* $\{f_n\}$ *s.t.* $\lim_n \int_\Omega ||f_n - f|| d\mu = 0$.

We define the Bochner integral of f*:* $\int_\Omega f d\mu = \lim_n \int_\Omega f_n d\mu$.

The existence of $\lim_n \int_\Omega f_n d\mu$ since $\{\int_\Omega f_n d\mu\}$ is Cauchy:

$$|| \int f_n d\mu - \int f_m d\mu || \leq \int ||f_n - f_m|| d\mu \leq \int ||f - f_m|| d\mu + \int ||f - f_n|| d\mu.$$

Theorem

$f_n \to f$, f_n is integrable, and $\exists\ g > 0$ is Lebesgue integrable and $||f_n|| \leq g$. Then f integrable and $\int f d\mu = \lim_n \int f_n d\mu$.

Proof.

$||f - f_n|| \leq ||f|| + ||f_n|| \leq g + ||f|| \Rightarrow \int ||f - f_n|| d\mu \to 0$.
Take simple $\{g_{n_k}\}$ s.t. $\int ||f_n - g_{n_k}|| d\mu \to 0$, then $\int ||f - g_{n_k}|| d\mu \to 0$. □

One can use this to prove $|| \int f d\mu|| \leq \int ||f|| d\mu$, $\forall$ integrable f.

Lemma

f is a measurable, $\int ||f|| d\mu < \infty$. If $\exists$ measurable $g_n : (\Omega, \mathcal{F}) \to (V_n, \mathcal{B}(V_n))$, $dim(V_n) < \infty$ s.t. $\lim_n \int ||f - g_n|| d\mu = 0$, then f is integrable.

Proof.

Let $K_n = \{x \in V_n; ||x|| \in [1/n, n]\}$, $\Omega_n = g_n^{-1}(K_n)$, $\mu(\Omega_n) < \infty$ since $\int_{\Omega_n} d\mu \leq \int_{\Omega_n} n||g_n|| d\mu < \infty$.

K_n compact, $\exists\ K_n \subset \cup_k B_k$, the diameter of B_k less than $(n\mu(\Omega_n))^{-1}$. Let $f_n = \sum_k g_n(x_k) I_{g_n^{-1}(B_k)}$ and $x_k \in g_n^{-1}(B_k) \Rightarrow \forall\ \omega \in \Omega_n$, $||f_n(\omega) - g_n(\omega)|| \leq (n\mu(\Omega_n))^{-1}$.

$\int ||f_n - f|| d\mu \leq \int ||f_n - g_n|| d\mu + \int ||g_n - f|| d\mu$ and $\int_{\Omega_n} ||f_n - g_n|| d\mu \leq 1/n$. □

Proof.

$\int_{\Omega_n^c} ||f_n - g_n|| d\mu = \int_{||g_n||>n} ||g_n|| d\mu + \int_{||g_n||<1/n} ||g_n|| d\mu.$
$\mu(||g_n|| > n) \leq n^{-1} \int ||g_n|| d\mu \approx n^{-1} \int ||f|| d\mu \to 0 \Rightarrow \int_{||g_n||>n} ||f|| d\mu \to 0.$
Since $\mu(||g_n|| < 1/n, ||f|| \geq \varepsilon) = (\varepsilon - 1/n) \int ||f - g_n|| \mu \to 0$, $\int_{||g_n||<1/n} ||f|| d\mu$ $\leq \int_{||g_n||<1/n} ||f|| I(||f|| \geq \varepsilon) d\mu + \int_{||f||<\varepsilon} ||f|| d\mu, \Rightarrow 0.$ □

Theorem

H separable Hilbert space, if $\int ||f|| d\mu < \infty$, then f integrable.

Proof.

Take a COB of H: $\{e_n\}$, $M_k = span\{e_n\}_{n \leq k}$. Define $g_k : g_k(\omega) = P_{M_k} f(\omega) \Rightarrow$ $||g_k|| \leq ||f||$. $\lim_k \int ||f - g_k|| d\mu = \int \lim_k ||f - g_k|| d\mu = 0.$ □

Property

V_i Banach space, $f : \Omega \to V_1$ integrable. $\forall T \in B(V_1, V_2)$, $\int T(f) d\mu = T(\int f d\mu)$. If $\mathcal{K} : \Omega \to B(V_1, V_2)$ and $\mathcal{K}$ is integrable, $\int \mathcal{K} x d\mu = (\int \mathcal{K} d\mu) x$.

Proof.

$T(f)$ is measurable, let $f = \sum_n I_{A_n} g_n$, $\int T(\sum_n I_{A_n} g_n) d\mu = \sum_n \int I_{A_n} T(g_n) d\mu$ $= \sum_n \mu(A_n) T(g_n) = T(\sum_n \mu(A_n) g_n) = T(\int f d\mu)$.

Let $J_x : B(V_1, V_2) \to V_2$, $J_x(\mathcal{K}) = \mathcal{K} x$. $||J_x(\mathcal{K})||_2 \le ||\mathcal{K}||\ ||x||_1 \Rightarrow J_x$ bounded $\Rightarrow \int \mathcal{K} x d\mu = \int J_x(\mathcal{K}) d\mu = J_x(\int \mathcal{K} d\mu) = (\int \mathcal{K} d\mu) x$. $\square$

Table of Contents

We firstly focus on Hilbert space H_i and bounded T. The model is $\mathcal{Y} = Tm + \varepsilon$. To find a approximation of m, we need to solve the optimization problem:

$$arg\min_{m \in H_1} ||y - Tm||_2$$

We have already shown that $\hat{m} = T^{\dagger}y + Ker(T) = (T^*T)^{\dagger}T^*y + Ker(T)$.

More practical and theoretical valid choice is that $m \in H_1$ s.t. $\langle m, Wm \rangle_1 \leq K$, $W \in B(H_1)$ and $W \gg 0$. It's equivalent to solve m for the loss function:

$$arg\min_{m \in H_1} L(y, m) = ||y - Tm||_2 + \lambda \langle m, Wm \rangle_1$$

Theorem

Assume $T^*T + \lambda W$ is invertible, then $\hat{m} = (T^*T + \lambda W)^{-1} T^* y$, $y \in Dom(T^\dagger)$.

Proof.

$\frac{dL(y,m)}{dm} = \frac{d(\langle m, (\lambda W + T^*T)m \rangle_1 - 2\langle m, T^*y \rangle_1)}{dm} = 2\langle \cdot, (\lambda W + T^*T)m \rangle_1 - 2\langle \cdot, T^*y \rangle_1$, then if $(\lambda W + T^*T)m = T^*y$, m minimizes the loss. $\square$

Example

If T compact then $Ker(T)^{\perp}$ is separable, let $A \subset Ker(T)^{\perp}$ and $W = P_A$. Then $T^*T + \lambda P_A$ compact and self-adjoint.

Supposed $A = span\{e_n\}$ and $T^*T = \sum_n \mu_n e_n \otimes e_n + \sum_m \theta_m g_m \otimes g_m$.

$$\begin{aligned} T^*T + \lambda P_A &= \sum_n \mu_n e_n \otimes e_n + \sum_m \theta_m g_m \otimes g_m + \lambda_n \sum_n e_n \otimes e_n \\ &= \sum_n (\mu_n + \lambda) e_n \otimes e_n + \sum_m \theta_m g_m \otimes g_m \end{aligned}$$

$$\begin{aligned} m_\lambda = (T^*T + \lambda W)^{-1} T^* y &= (\sum_n 1/(\mu_n + \lambda) e_n \otimes e_n + \sum_m (1/\theta_m)\ g_m \otimes g_m) T^* y \\ &= (\sum_n 1/(\mu_n + \lambda) e_n \otimes e_n) T^* y + P_{A^\perp} (T^*T)^\dagger T^* y. \end{aligned}$$

$$|| \sum_n 1/(\mu_n + \lambda) e_n \otimes e_n) || = 1/(\mu_1 + \lambda) \Rightarrow \lim_{\lambda \to \infty} m_\lambda = P_{A^\perp} T^\dagger y.$$
$$|| \sum_n \lambda(\mu_n(\mu_n + \lambda))^{-1} e_n \otimes e_n) || = \lambda(\mu_1(\mu_1 + \lambda))^{-1} \Rightarrow \lim_{\lambda \to 0} m_\lambda = T^\dagger y.$$

$(T^*T+\lambda W)^{-1}T^*y = (T^*T+\lambda W)^{-1}T^*(y - TP_{Ker(W)}T^\dagger y) + P_{Ker(W)}T^\dagger y$ since $P_{Ker(W)}T^\dagger y$ minimizes $L(TP_{Ker(W)}T^\dagger y, m)$. Let $L = TW^{-1/2}$:

$$(T^*T+\lambda W)^{-1}T^*(y - TP_{Ker(W)}T^\dagger y) = W^{-1/2}(L^*L+\lambda I)^{-1}L^*(y - TP_{Ker(W)}T^\dagger y)$$

Let $m_\lambda = (L^*L+\lambda I)^{-1}L^*y$, $L(y,m) = ||y - Lm||_2^2 + \lambda ||m||_1^2$.

Lemma

$m_\lambda \in Ker(L)^\perp$, $||m_\lambda||_1 \le ||L^\dagger y||_1$.

Proof.

$||y - Lm||_2^2 = ||y - LP_{Ker(L)^\perp}m||_2^2$ and $||m||_1 \ge ||P_{Ker(L)^\perp}m||_1$.
$L(y, m_\lambda) \le L(y, L^\dagger y) \Rightarrow ||m_\lambda||_1 \le ||L^\dagger y||_1$. □

Theorem

$y \in Dom(L^\dagger)$, $\lim_{\lambda \to 0} m_\lambda = L^\dagger y$.

Proof.

Let $A = L^*L + \lambda I$, $B = L^*L$, $b = L^*y$, $(L^*L + \lambda I)^{-1}L^*y - (L^*L)^\dagger L^*y = A^{-1}b - B^\dagger b = B^\dagger(B - A)A^{-1}b + (I - B^\dagger B)A^{-1}b$.

$I - (L^*L)^\dagger L^*L = P_{Ker(L^*L)} = P_{Ker(L)} \Rightarrow (I - (L^*L)^\dagger L^*L)m_\lambda = 0$.

$||B^\dagger(B - A)A^{-1}b||_1 = ||(L^*L)^\dagger(L^*L - L^*L - \lambda I)(L^*L + \lambda I)^{-1}L^*y||_1 = \lambda||(L^*L)^\dagger m_\lambda||_1 \le C\lambda$. □

Then $W^{-1/2}(L^*L + \lambda I)^{-1}L^*(y - TP_{Ker(W)}T^\dagger y) + P_{Ker(W)}T^\dagger y \to W^{-1/2}L^\dagger (y - TP_{Ker(W)}T^\dagger y) + P_{Ker(W)}T^\dagger y = T^\dagger(y - TP_{Ker(W)}T^\dagger y) + P_{Ker(W)}T^\dagger y = T^\dagger y + P_{Ker(T)}P_{Ker(W)}T^\dagger y = T^\dagger y$ since $T^*T + \lambda W$ is invertible.

Theorem

$y \in Dom(L^{\dagger})$, $\lim_{\lambda \to \infty} m_{\lambda} = 0$.

Proof.

$||y - Lm_{\lambda}||_2^2 + \lambda ||m_{\lambda}||_1 \leq ||y||_2^2 \Rightarrow m_{\lambda} \to 0.$ □

$$W^{-1/2} \left(L^* L + \lambda I\right)^{-1} L^*(y - TP_{Ker(W)}T^{\dagger}y) + P_{Ker(W)}T^{\dagger}y \to P_{Ker(W)}T^{\dagger}y.$$

Theorem

$$E||Tm - Tm_\lambda||_2^2 = ||ETm_\lambda - Tm||_2^2 + E||Tm_\lambda - ETm_\lambda||_2^2.$$

Proof.

$E||Tm - Tm_\lambda||_2^2 = E||Tm - TEm_\lambda + TEm_\lambda - Tm_\lambda||_2^2$ and $E\langle Tm - TEm_\lambda, TEm_\lambda - Tm_\lambda\rangle_2 = \langle Tm - TEm_\lambda, TEm_\lambda - E(Tm_\lambda)\rangle.$ □

$$Bias^2(\lambda) = ||T(Em_\lambda - m))||_2^2,\ Var(\lambda) = E||T(m_\lambda - Em_\lambda)||_2^2.$$

Theorem

$$Bias^2(\lambda) \leq \lambda \langle m, Wm \rangle_1$$

Proof.

$$Em_\lambda = E(T^*T + \lambda W)^{-1} T^* y = (T^*T + \lambda W)^{-1} T^* T m.$$

$Bias^2(\lambda) = ||TEm_\lambda - Tm||_2^2 \leq L(Tm, g)$, $\forall g \in H_1$, let $g = m$. □

Let $||y||_2 = y^T y/n$, then T is compact.

Theorem

$Var(\lambda) = \frac{\sigma^2}{n} \sum_k (\frac{\mu_k}{\mu_k + \lambda})^2$, μ_k *is the eigenvalue of* $W^{-1/2}T^*TW^{-1/2}$.

Proof.

$TW^{-1/2} = \sum_k \sqrt{\mu_k} e_{2k} \otimes_2 e_{1k}$, let $A = T(T^*T + \lambda W)^{-1}T^* = TW^{-1/2}(W^{-1/2}T^*TW^{-1/2} + \lambda I)^{-1}W^{-1/2}T^* = \sum_k \frac{\mu_k}{\mu_k + \lambda} e_{2k}e_{2k}^T$.

$E||T(m_\lambda - Em_\lambda)||_2^2 = E||A(y - Tm)||_2^2 = E||A\varepsilon||_2^2 = E||\sum_k \frac{\mu_k}{\mu_k + \lambda} e_{2k}e_{2k}^T \varepsilon||_2^2$
$= \sum_k (\frac{\mu_k}{\mu_k + \lambda})^2 E||e_{2k}e_{2k}^T\varepsilon||_2^2 = \sum_k (\frac{\mu_k}{\mu_k + \lambda})^2 E(e_{2k}^T\varepsilon)^2/n^2 = \frac{\sigma^2}{n}\sum_k (\frac{\mu_k}{\mu_k + \lambda})^2$. □

Let $m_\lambda^{[k]}$ be is the minimizer of loss function: $L_1(y_k, m) = \sum_{i \neq k}(y_i - T_i m)^2)/n$ $+\lambda\langle m, Wm\rangle_1$. $m_\lambda[k, z]$ is the minimizer of loss function:

$$L_2(c(z, y_{-k}), m) = ((z - T_k m)^2 + \sum_{i \neq k}(y_i - T_i m)^2)/n + \lambda\langle m, Wm\rangle_1$$

Lemma

$m_\lambda[k, y_k] = m_\lambda$ *and* $m_\lambda[k, T_k m_\lambda^{[k]}] = m_\lambda^{[k]}$.

Proof.

$L_2(c(T_k m_\lambda^{[k]}, y_{-k}), m_\lambda^{[k]}) = \sum_{i \neq k}(y_i - T_i m_\lambda^{[k]})^2)/n + \lambda\langle m_\lambda^{[k]}, P m_\lambda^{[k]}\rangle_1 \leq$ $L_1(y_{-k}, m) \leq L_2(c(T_k m_\lambda^{[k]}, y_{-k}), m)$. □

$CV(\lambda) = \sum_n (y_k - T_k m_\lambda^{[k]})^2/n,\ H = T(T^*T + \lambda P)^{-1}T^*$ is called hat matrix.

Theorem

$$CV(\lambda) = \frac{1}{n}\sum_n \left(\frac{y_k - T_k m_\lambda}{1 - H_{kk}}\right)^2.$$

Proof.

$$1 - a_{kk} = (y_k - T_k m_\lambda)/(y_k - T_k m_\lambda^{[k]}) \Rightarrow a_{kk} = (T_k m_\lambda - T_k m_\lambda^{[k]})\ /(y_k - T_k m_\lambda^{[k]})$$
$$= (T_k m_\lambda[k, y_k] - T_k m_\lambda[k, T_k m_\lambda^{[k]}])/(y_k - T_k m_\lambda^{[k]}) \Rightarrow a_{kk} = \partial T_k m_\lambda[k, y_k]/\partial y_k = H_{kk}.$$
$\square$

We are interesting in the estimation: $E||Y_t - T_t m_\lambda||_2^2 = \sigma^2 + Var(\lambda) + Bias^2(\lambda)$. Let $Risk(\lambda) = Var(\lambda) + Bias^2(\lambda)$ and $MSE(\lambda) = ||(I - H)y||_2^2$.

Lemma

$E(MSE(\lambda)) = Risk(\lambda) + \sigma^2 - 2\sigma^2 tr(H)/n.$

Proof.

$Var(\lambda) = Var||HY||_2 = tr(Var(HY))/n = \sigma^2 tr(H^2)/n.$
$E(MSE(\lambda)) = E||(I - H)Y||_2^2 = Bias^2(\lambda) + E||(I - H)\varepsilon||_2^2 = Bias^2(\lambda) + \sigma^2 tr((I - H)^2)/n = Bias^2(\lambda) + \sigma^2 - 2\sigma^2 tr(H)/n + Var(\lambda).$ □

$C_p(\lambda) = MSE(\lambda) + 2\sigma^2 tr(H)/n$ is an unbiased estimator of $Risk(\lambda) + \sigma^2$.

Theorem

$GCV(\lambda) = MSE(\lambda)/(1 - tr(H)/n)^2$. *Let* $a = tr(H)/n$, $b = tr(H^2)/n$ *and* $c = \frac{a(2-a)}{(1-a)^2} + \frac{a^2}{(1-a)^2 b}$, *then* $|E(GCV(\lambda)) - \sigma^2 - Risk(\lambda)| \leq c\ Risk(\lambda)$.

Proof.

$|E(GCV(\lambda)) - \sigma^2 - Risk(\lambda)| = |\frac{Risk(\lambda)+\sigma^2-2\sigma^2 a}{(1-a)^2} - \sigma^2 - Risk(\lambda)| = |\frac{a(2-a)}{(1-a)^2} Risk(\lambda) - \frac{a^2}{(1-a)^2}\sigma^2| \leq |\frac{a(2-a)}{(1-a)^2} + \frac{a^2}{(1-a)^2}\frac{\sigma^2}{Risk(\lambda)}|Risk(\lambda) \leq c\ Risk(\lambda)$. $\square$

GCV has an advantage that is no need to estimate σ^2.

Table of Contents

We are interesting in the non-parametric Regression questions: $Y_i = f(t_i) + \varepsilon_i$. The optimization question is that $arg\min_f L(y, f)$.

If $f \in H(K)$ and $H(K) = H_0 \oplus H_1$, we confine f: $||P_{H_1} f|| < K$ to ensure the uniqueness of f. Then the loss function is: $L_1(y, f) = L(y, f) + \lambda ||P_{H_1} m||_1$.

Define $T_t f = f(t) = \langle f, \tau_t \rangle_1$, $span\{\phi_i\} = H_0$ and $\xi_j = P_{H_1} \tau_{t_j}$.

Theorem

$\hat{f} = \sum_j a_j \phi_j + \sum_i b_i \xi_i$.

Proof.

$h = \hat{f} + g$, $g \in span\{\phi_j, \xi_i\}^\perp$. $h(t) = \langle h, \tau_t \rangle_1 = \langle \hat{f}, \tau_t \rangle_1 + \langle g, \xi_t \rangle_1 + \langle g, \tau_t - \xi_t \rangle_1$
$= \hat{f}(t)$, but $||P_{H_1} h||_1 \geq ||P_{H_1} \hat{f}||_1$. □

Example

If $H_1 = H(K)$, *then* $\hat{f} = \sum_j a_j K(\cdot, t_j)$ *and* $L_1(y, f) = L(y, \mathcal{K}\alpha) + \lambda\alpha\mathcal{K}\alpha$, $\mathcal{K}_{ij} = K(t_i, t_j)$.

We can view $\alpha \sim N(0, \mathcal{K}^{-1})$, *then we have a prior mean zero Gaussian process* $\{\hat{f}(t)\}_{t \in E}$ *which covariance function* K.

If $L(y, f) = \sum_i (y_i - f(t_i))^2$, $\hat{\alpha} = (\mathcal{K} + \lambda I)^{-1} y$ *and* $\hat{f} = \mathcal{K}(\mathcal{K} + \lambda I)^{-1} y$. *This is equivalence to solve* $\omega(s)$ *from statistical model:* $Y(s) = \omega(s) + \varepsilon(s)$, $\varepsilon(s) \sim N(0, \lambda)$.

Generally, $L_1(y, f) = L(y, K\theta) + \lambda\beta^T\Sigma\beta$. Among $\Phi_{ij} = \phi_i(t_j)$, $\Sigma_{ij} = \xi_i(t_j)$, $\theta = (a_1, \ldots, a_q, \beta^T)^T$, $\beta = (b_1, \ldots, b_n)^T$, $K = (\Phi^T, \Sigma^T)^T$.

Example

Recall $W_q[0,1] = H_0 \oplus H_1$, $H_0 = P_q[t]$, $H_1 = \{G_q \circ g; g \in L^2([0,1])\}$ *is a RKHS with rk* $K(s,t) = \sum_n \Phi_n(s)\Phi_n(t) + \int_0^1 G_q(s-u)G_q(t-u)du$. *We want to find the best approximation function* $\hat{m}$ *on* $W_q[0,1]$*:*

$$\hat{m} = arg\min_{m \in W_q[0,1]} \frac{\sum_i (y_i - m(t_i))^2}{n} + \lambda ||P_{H_1} m||_{W_q[0,1]}$$

Noticed that $||P_{H_1} m||_{W_q[0,1]} = ||m^{(q)}||_2$. $\hat{m} = \sum_j a_j \Phi_j + \sum_i b_i P_{H_1}(K(\cdot, t_i))$ and $P_{H_1}(K(\cdot,t)) = K_1(\cdot, t_i)$.

Definition

$S^q(t_1,...,t_n)$ is a functions space contained all the piecewise polynomials satisfied: $\forall\ f \in S^q$:
(a) $f|_{[t_i,t_{i+1})} \in P_q[t]$;
(b) $f^{(q-1)}$ exists and it's a step function with jump at t_i.

$$S^q(t_1,...,t_n) = \{\sum_i^{q-1} a_i t^i + \sum_{j=1}^n b_j (t-t_j)_+^{q-1}; a_i, b_j \in \mathcal{R}\}.$$

Definition

A natural splines g of order $2q$ is a $2q$-order spline satisfied: $g^{(j)}(0) = g^{(j)}(1)$, $j = q, ..., 2q-1$. We mark that $g \in N^{2q}(t_1,...,t_n)$.

Theorem

$\forall\ f \in W_q[0,1]$ *which interpolates* (t_i, z_i), $\exists!\ g \in N^{2q}(t_1, ..., t_n)$ *satisfied that* $||g^{(q)}||_2 \leq ||f^{(q)}||_2$.

Proof.

Let $h(t) = f(t) - g(t)$, $h(t_i) = 0$. $\int h^{(q)}(t) g^{(q)}(t) dt = - \int g^{(q+1)}(t) h^{(q-1)}(t) dt$
$= (-1)^{q-1} \int g^{(2q-1)}(t) h^{(1)}(t) dt = (-1)^{q-1} \sum_i g^{(2q-1)}(t_i)(h(t_{i+1}) - h(t_i)) = 0$.
$\int |f^{(q)}(t)|^2 dt = \int |g^{(q)}(t)|^2 dt + \int |h^{(q)}(t)|^2 dt \geq \int |g^{(q)}(t)|^2 dt$.
If $\int |h^{(q)}(t)|^2 dt = 0 \Rightarrow h \in P_q[t] \Rightarrow h = 0$. $\square$

We connect the $H(K)$ with $L^2(E)$ by integral operator. Let T be the integral operator: $Tf(s) = \langle f, K(\cdot, s)\rangle_2$, K is a continuous symmetric, non-negative kernel.

Theorem

$K(s,t) = \sum_n \lambda_n e_n(s) e_n(t)$, $\mathcal{G}(K) = \{\sum_n a_n e_n; \sum_n \frac{a_n^2}{\lambda_n} < \infty\}$, which inner product is $\langle \sum_n a_n e_n, \sum_n b_n e_n \rangle_G = \sum_n \frac{a_n b_n}{\lambda_n}$. Then $H(K) = \mathcal{G}(K)$.

Proof.

"$\subset$": $K(\cdot, t) = \sum_n \lambda_n e_n(t) e_n$, $\sum_n \lambda_n e_n^2(t) = K(t,t) < \infty$.
"$\supset$": $f(t) = \sum_n a_n e_n(t) = \sum_n a_n \langle e_n, K(\cdot, t)\rangle_K = \langle f, K(\cdot, t)\rangle_K$. □

Noticed that $\{\sqrt{\lambda_n} e_n\}$ is COB of $H(K)$, we say $H(K)$ is also a function space that is generated by some basis functions.

The loss: $L(y, f) + \lambda ||P_{H_1} m||_1 = L(y, \sum_n a_n e_n) + \mu \sum_{\{m; e_m \in H_1\}} \frac{a_m^2}{\lambda_m}$.

Example

(Kernel trick) Let $K(s,t) = \sum_n \lambda_n e_n(s) e_n(t)$, $h(x) = (\sqrt{\lambda_n} e_n(x))_n \in l^2$. *Define loss:* $L(y, h^T \beta) + \mu ||\beta||_2^2 = L(y, \sum_j \sqrt{\lambda_n} e_j \beta_j) + \mu \sum_j \beta_j^2 = L(y, \sum_j \alpha_j e_j) + \mu \sum_j \frac{\alpha_j^2}{\lambda_j}$.