Mathematical Foundations of Infinite-Dimensional Statistical Models

Chap.1

Seonghyeon Kim

2018.06.22

Introduction

Introduction

$$Y \sim P_f\,,\ \{P_f : f \in \mathcal{F}\}$$

The problem of statistical inference on f, can be divided into three intimately connected problems.

- Estimate the parameter f by an estimator $T(Y)$.
- Test hypotheses on f based on test functions $\Psi(Y)$.
- Construct confidence sets $C(Y)$ that contain f with high probability.

1.1 Statistical Sampling Models

1.1 Statistical Sampling Models

X : a random experiment with associated sample space $\mathcal{X}$.
$\mathcal{A}$: a $\sigma -$ *field* of subsets of $\mathcal{X}$.
$(\mathcal{X}, \mathcal{A})$: measurable space
P : probability measure on $\mathcal{A}$.
$X_1, \ldots, X_n$: i.i.d. copies from X
$P^n = \otimes_{i=1}^n P$: joint distribution of the $X_1, \ldots, X_n$

- The goal is to recover P from the n observations.
- Classical statistics has been concerned mostly with models where P is explicitly parameterised by a finite-dimensional parameter.
- In this book, we will follow the often more realistic assumption that no such parametric assumptions are made on P.

1.1.1 Nonparametric Models for Probability Measures

Total variation metric

$$||P - Q||_{TV} = \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$$

Bounded Lipschitz metric

$\mathcal{X}$ is endowed with a metric d

$$\beta_{(\mathcal{X},d)}(P,Q) = \sup_{f \in BL(1)} \left| \int_{\mathcal{X}} f(dP - dQ) \right|, \textit{ where}$$

$$BL(M) = \left\{ f : \mathcal{X} \to \mathbb{R}, \sup_{x \in \mathcal{X}} |f(x)| + \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x - y|} \leq M \right\}, \ 0 < M < \infty$$

1.1.1 Nonparametric Models for Probability Measures

Supremum-norm metric (Kolmogorov distance)

$$||F_P - F_Q||_\infty = \sup_{x \in \mathbb{R}} |F_P(x) - F_Q(x)|$$

L^1-distance

$$||f_P - f_Q||_1 = \int_{\mathbb{R}} |f_P(x) - f_Q(x)| dx$$

1.1.1 Nonparametric Models for Probability Measures

- Class of probability densities is more complex than the class of probability-distribution functions.
- we can anticipate that estimating a probability density is harder than estimating the distribution function.
- Instead of P, a particular functional $\Phi(P)$ may be the parameter of statistical interest
- Proving closeness of T to P in some strong loss function then gives access to 'many' continuous functionals Φ for which $\Phi(T)$ will be close to $\Phi(P)$.

1.1.2 Indirect Observations

Indirect Observations

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} P_X$$

$$\epsilon_1, \dots, \epsilon_n \overset{i.i.d.}{\sim} P_\epsilon$$

$$Y_i = X_i + \epsilon_i, \ i = 1, \dots, n$$

$$P_Y = P_X * P_\epsilon$$

- The observer may have very concrete knowledge of the source of the error.
- It is also known as the deconvolution model because one wishes to deconvolve P_ϵ.

1.2 Gaussian Models

1.2.1 Basic Ideas of Regression

Regression model

$$Y_i = f(x_i) + \epsilon_i,\ i = 1, \ldots, n$$

Standard Gaussian linear model

$$f(x) = x_1\theta_i + \cdots + x_p\theta_p$$

$$Y_i = f(x_i) + \epsilon_i \equiv \sum_{j=1}^{p} x_{ij}\theta_j + \epsilon_i,\ i = 1, \ldots, n$$

$$\epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

1.2.1 Basic Ideas of Regression

- If $E(\epsilon_i) \neq 0$, this could be accommodated in the functional model by adding a constant $x_{10} = \cdots = x_{n0} = 1$
- By the CLT, $\epsilon_i = \sum_k \epsilon_{ik}$ should be approximately normally distributed, regardless of the actual distribution of the ϵ_{ik}.
- The assumption that the function f is linear is in principle quite arbitrary.

1.2.2 Some Nonparametric Gaussian Models

Nonparametric regression model with equally spaced design on $[0, 1]$

$$Y_i = f(x_i) + \epsilon_i,\ x_i = \frac{i}{n}, \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2),\ i = 1, \ldots, n$$

- The assumption that the x_i are equally spaced is important for the theory that will follow.
- It may not be reasonable to assume that f has any specific properties other than that it is a continuous or a differentiable function.
- Even if we would assume that f has infinitely many continuous derivatives the set of all such f would be infinite dimensional.

1.2.2 Some Nonparametric Gaussian Models

Nonparametric regression model with equally spaced design on $[0, 1]$

$$Y_i = f(x_i) + \epsilon_i,\ x_i = \frac{i}{n}, \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2),\ i = 1, \ldots, n$$

- The assumption that the x_i are equally spaced is important for the theory that will follow.
- It may not be reasonable to assume that f has any specific properties other than that it is a continuous or a differentiable function.
- Even if we would assume that f has infinitely many continuous derivatives the set of all such f would be infinite dimensional.

1.2.2 Some Nonparametric Gaussian Models

Gaussian White Noise Model

$$dY(t) \equiv dY_f^{(n)}(t) = f(t)dt + \frac{\sigma}{\sqrt{n}} dW(t), t \in [0,1], n \in \mathbb{N},$$

dW is a standard Gaussian white noise process.

$$g \mapsto \int_0^1 g(t)dY^{(n)}(t) \equiv \mathbb{Y}_f^{(n)}(g) \sim N\left(\langle f, g\rangle, \frac{||g||_2^2}{n}\right)$$

$$g \mapsto \int_0^1 g(t)dW(t) \equiv \mathbb{W}(g) \sim N\left(0, ||g||_2^2\right), g \in L^2([0,1])$$

- $\mathbb{W}$ and $\mathbb{Y}^{(n)}$ define Gaussian processes on L^2.
- For any finite set of orthonormal vectors $\{e_k\} \subset L^2$, $\{\mathbb{W}(e_k)\}$ is a multivariate standard normal variable.

1.2.2 Some Nonparametric Gaussian Models

Gaussian Sequence Space Model

$$Y_k \equiv Y_{f,k}^{(n)} = \langle f, e_k \rangle + \frac{\sigma}{\sqrt{n}} g_k, k \in \mathbb{Z}, n \in \mathbb{N},$$

where $\{e_k : k \in \mathbb{Z}\}$ is orthonormal basis of L^2 and g_k are i.i.d. of law $\mathbb{W}(e_k) \sim N(0, ||e_k||_2^2) = N(0, 1)$

- $\{e_k\}$ realise an isometry between L^2 and l^2 through the mapping $f \mapsto \{\langle f, e_k \rangle\}$
- Gaussian White Noise Model and Gaussian Sequence Space Model are equivalent to each other.

1.2.3 Equivalence of Statistical Experiments

The Le Cam Distance of Statistical Experiments

$\mathcal{E}^{(i)} = (\mathcal{Y}_i, P_f^{(i)}),\ i = 1, 2$

$\mathcal{Y}_i$: sample space

$P_f^{(i)}$: probability measure defined on $\mathcal{Y}_i$

$\mathcal{T}$: measurable space of decision rules. $T^{(i)}(Y^{(i)}) \in \mathcal{T}$

$L : \mathcal{F} \times \mathcal{T} \mapsto [0, \infty)$: loss function measuring the performance

$|L| = \sup\{L(f, T) : f \in \mathcal{F}, T \in \mathcal{T}\}$

$R^{(i)}(f, T^{(i)}, L) = \int_{\mathcal{Y}_i} L(f, T^{(i)}(Y^{(i)}))dP_f^{(i)}$

1.2.3 Equivalence of Statistical Experiments

The Le Cam Distance of Statistical Experiments(Conti.)

$$\Delta_{\mathcal{F}}(\mathcal{E}^{(1)}, \mathcal{E}^{(2)}) \equiv \max\Big[\sup_{T^{(2)}} \inf_{T^{(1)}} \sup_{f, L: |L|=1} |R^{(1)}(f, T^{(1)}, L) - R^{(2)}(f, T^{(2)}, L)|,$$
$$\sup_{T^{(1)}} \inf_{T^{(2)}} \sup_{f, L: |L|=1} |R^{(1)}(f, T^{(1)}, L) - R^{(2)}(f, T^{(2)}, L)|\Big]$$

1.2.3 Equivalence of Statistical Experiments

Proposition 1

If $\mathcal{Y}^{(1)} = \mathcal{Y}^{(2)} = \mathcal{Y}$ and $P_f^{(1)}, P_f^{(2)} \ll \mu$,

$$\Delta_{\mathcal{F}}(\mathcal{E}^{(1)}, \mathcal{E}^{(2)}) \leq \sup_{f \in \mathcal{F}} \int_{\mathcal{Y}} \left| \frac{dP_f^{(1)}}{d\mu} - \frac{dP_f^{(2)}}{d\mu} \right| d\mu \equiv ||P^{(1)} - P^{(2)}||_{1,\mu,\mathcal{F}}$$

pf)

$$\inf_{T^{(1)}} \sup_{f,L:|L|=1} |R^{(1)}(f, T^{(1)}, L) - R^{(2)}(f, T^{(2)}, L)| \leq$$

$$\sup_{f,L:|L|=1} |R^{(1)}(f, T^{(2)}, L) - R^{(2)}(f, T^{(2)}, L)|$$

$$|R^{(1)}(f, T, L) - R^{(2)}(f, T, L)| \leq \int_{\mathcal{Y}} |L(f, T(Y))||dP_f^{(1)} - dP_f^{(2)}| \leq$$

$$|L|||P^{(1)} - P^{(2)}||_{1,\mu,\mathcal{F}}$$

1.2.3 Equivalence of Statistical Experiments

Proposition 2

If we can find a bi-measurable isomorphism B of $Y^{(1)}$ with $Y^{(2)}$, independent of f, such that

$$P_f^{(2)} = P_f^{(1)} \circ B^{-1}, P_f^{(1)} = P_f^{(2)} \circ B,$$

then

$$\Delta_{\mathcal{F}}(\mathcal{E}^{(1)}, \mathcal{E}^{(2)}) = 0.$$

1.2.3 Equivalence of Statistical Experiments

Proof of Proposition 2

Let $T^{(2)}(Y^{(2)}) \equiv T^{(1)}(B^{-1}(Y^{(2)}))$

$$R^{(2)}(f, T^{(2)}, L) = \int_{\mathcal{Y}_2} L(f, T^{(1)}(B^{-1}(Y^{(2)})))dP_f^{(2)} = \int_{\mathcal{Y}_1} L(f, T^{(1)}(Y^{(1)}))dP_f^{(1)}$$
$$= R^{(1)}(f, T^{(1)}, L).$$

1.2.3 Equivalence of Statistical Experiments

Proposition 3

If there exists a mapping $S : \mathcal{Y}^{(1)} \to \mathcal{Y}^{(2)}$ independent of f such that

$$Y^{(2)} = S(Y^{(1)}), \; Y^{(2)} \sim P_f^{(2)}$$

and $S(Y^{(1)})$ is a sufficient statistic for $Y^{(1)}$, then

$$\Delta_{\mathcal{F}}(\mathcal{E}^{(1)}, \mathcal{E}^{(2)}) = 0.$$

1.2.3 Equivalence of Statistical Experiments

α-*Hölderian* function

$$\mathcal{F}(\alpha, M) = \left\{ f : [0,1] \to \mathbb{R}, \sup_{x \in [0,1]} |f(x)| + \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x - y|^{\alpha}} \leq M \right\}$$

$$0 < \alpha \leq 1, 0 < M < \infty$$

1.2.3 Equivalence of Statistical Experiments

Theorem 1.2.1

Let $(\mathcal{E}_n^{(i)} : n \in \mathbb{N}), i = 1, 2, 3$, equal the sequence of statistical experiments given by

$$(i = 1)\ Y_i = f(x_i) + \epsilon_i,\ x_i = \frac{i}{n}, \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2),\ i = 1, \dots, n$$

$$(i = 2)\ dY(t) \equiv dY_f^{(n)}(t) = f(t)dt + \frac{\sigma}{\sqrt{n}}dW(t), t \in [0, 1], n \in \mathbb{N}$$

$$(i = 3)\ Y_k \equiv Y_{f,k}^{(n)} = \langle f, e_k \rangle + \frac{\sigma}{\sqrt{n}}g_k, k \in \mathbb{Z}, n \in \mathbb{N}$$

and $\pi_n(f)$ be the function that interpolates f at the x_i and that is piecewise constant on each interval $(x_{i-1}, x_i] \subset [0, 1]$,

1.2.3 Equivalence of Statistical Experiments

Theorem 1.2.1 (Conti.)

Then, for F any family of bounded functions on $[0, 1]$,

$$\Delta_{\mathcal{F}}(\mathcal{E}_n^{(2)}, \mathcal{E}_n^{(3)}) = 0, \Delta_{\mathcal{F}}(\mathcal{E}_n^{(1)}, \mathcal{E}_n^{(2)}) \leq \sqrt{\frac{n\sigma^2}{2}} \sup_{f \in \mathcal{F}} ||f - \pi_n(f)||_2.$$

If $\mathcal{F} = \mathcal{F}(\alpha, M)$ for any $\alpha > 1/2, M > 0$, then

$$\Delta_{\mathcal{F}}(\mathcal{E}_n^{(1)}, \mathcal{E}_n^{(2)}) \to 0 \quad as \quad n \to \infty.$$

1.2.3 Equivalence of Statistical Experiments

Proof of Theorem 1.2.1

$\Delta_{\mathcal{F}}(\mathcal{E}_n^{(2)}, \mathcal{E}_n^{(3)}) = 0$ follows from Proposition 2.
$\phi_{in} := 1_{(x_{i-1}, x_i]}$, $\mathcal{V}_n := span\{\phi_{in} : i = 1, \dots, n\}$, $\langle f, g\rangle_n := \sum_i f(x_i)g(x_i)$
Since $\langle f, \phi_{in}\rangle_n = f(x_i)$, $\pi_n(f)(t) = \sum f(x_i)\phi_{in}(t)$ is $\langle \cdot, \cdot \rangle_n$- projection of f onto $\mathcal{V}_n$.
$Y_i = f(x_i) + \epsilon_i, i = 1, \dots, n$ is equivalent to

$$\sum_{i=1}^n Y_i \phi_{in}(t) = \sum_{i=1}^n f(x_i)\phi_{in}(t) + \sum_{i=1}^n \epsilon_i \phi_{in}(t) = \pi_n(f)(t) + \sum_{i=1}^n \epsilon_i \phi_{in}(t) \cdots (1)$$

Let Π_n be $L^2([0,1])$ projector onto $\mathcal{V}_n$.

$$\int_0^1 h(t) \sum_{i=1}^n \epsilon_i \phi_{in}(t) dt = \int_0^1 \Pi_n(h)(t) \sum_{i=1}^n \epsilon_i \phi_{in}(t) dt \sim N(0, \frac{\sigma^2}{n} ||\Pi_n(h)||_2^2)$$

1.2.3 Equivalence of Statistical Experiments

Proof of Theorem 1.2.1

$$\int_0^1 h(t) \sum_{i=1}^{n} \epsilon_i \phi_{in}(t) dt = \mathbb{W}(\Pi_n(h))$$

It equals the L^2-projection of dW onto $\mathcal{V}_n$, justifying the notation

$$\frac{\sigma}{\sqrt{n}} dW_n(t) \equiv \sum_{i=1}^{n} \epsilon_i \phi_{in}(t) dt, dW_n = \Pi_n(dW)$$

(1) can be rewritten as

$$d\tilde{Y} = \pi_n(f)(t) + \frac{\sigma}{\sqrt{n}} dW_n(t) \cdots (2)$$

1.2.3 Equivalence of Statistical Experiments

Proof of Theorem 1.2.1

Next, consider the model

$$d\bar{Y} = \pi_n(f)(t) + \frac{\sigma}{\sqrt{n}} dW(t) \cdots (3),$$

then $d\tilde{Y} = \Pi_n(d\bar{Y})$, and $\Pi_n(d\bar{Y})$ is sufficient for $d\tilde{Y}$. So, (2) and (3) are equivalent by Proposition 3.
In view of Proposition 1 and using Proposition 6.1.7a) combined with (6.16),

$$\sup_{f \in \mathcal{F}} ||P_f^Y - P_{\pi_n(f)}^Y||_{1,\mu,\mathcal{F}}^2 \leq \frac{n}{\sigma^2} \sup_{f \in \mathcal{F}} ||f - \pi_n(f)||_2^2$$

which gives second claim.

1.2.3 Equivalence of Statistical Experiments

Proof of Theorem 1.2.1

Finally, uniformly in $\mathcal{F} = \mathcal{F}(\alpha, M)$,

$$
\begin{aligned}
||f - \pi_n(f)||_2^2 = & \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} (f(x) - f(x_i))^2 dx \leq M^2 \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} |x - x_i|^{2\alpha} \\
\leq & M^2 n^{-2\alpha} \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} dx = O(n^{-2\alpha})
\end{aligned}
$$

so for $\alpha > 1/2$, the bound of Le Cam distance converges to zero.