Asymptotic Statistics

Lecture 8

Nonparametric estimation

Distribution function estimation

Empirical distribution function

Observation: $X_1, \ldots, X_n$ i.i.d., unknown dist. function F.

Goal: estimate F without assuming some parametric model.

Idea: $F(x) = P(X_1 \leq x) = E1_{X_1 \leq x}$. Consider the sample counterpart of this expectation.

Definition.
The Empirical distribution function of the sample is the random function $\mathbb{F}_n$ defined by

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x}, \qquad x \in \mathbb{R}.$$

Pointwise properties of the empirical distribution function

For every fixed $x \in \mathbb{R}$:

- $E\mathbb{F}_n(x) = F(x)$ (unbiasedness)
- $\mathbb{F}_n(x) \xrightarrow{P} F(x)$ (consistency)
- $\sqrt{n}(\mathbb{F}_n(x) - F(x)) \rightsquigarrow N(0, F(x)(1 - F(x))$ (asymptotic normality)

Extensions to finitely many points

For fixed $x_1, \ldots, x_k \in \mathbb{R}$:

- $(\mathbb{F}_n(x_1), \ldots, \mathbb{F}_n(x_k)) \xrightarrow{P} (F(x_1), \ldots, F(x_k))$.
- $$\sqrt{n} \begin{pmatrix} \mathbb{F}_n(x_1) - F(x_1) \\ \vdots \\ \mathbb{F}_n(x_k) - F(x_k) \end{pmatrix} \rightsquigarrow N_k(0, \Sigma),$$

  where

  $$\Sigma_{ij} = Cov(1_{X_1 \leq x_i}, 1_{X_1 \leq x_j}) = F(x_i \wedge x_j) - F(x_i)F(x_j).$$

Uniform extensions

View $\mathbb{F}_n - F$ as a random function, for instance as a random element of the function space $\ell^\infty(\mathbb{R})$ of bounded functions on $\mathbb{R}$.

Theorem. (Glivenko-Cantelli)

We have

$$\|\mathbb{F}_n - F\|_\infty \stackrel{\text{as}}{\to} 0$$

Theorem. (Donsker)

We have

$$\sqrt{n}(\mathbb{F}_n - F) \rightsquigarrow \mathbb{G}$$

in $\ell^\infty(\mathbb{R})$, where $\mathbb{G} = (\mathbb{G}(x) : x \in \mathbb{R})$ is a Gaussian random function in $\ell^\infty(\mathbb{R})$, with $E\mathbb{G}(x) = 0$ and $Cov(\mathbb{G}(x), \mathbb{G}(y)) = F(x \wedge y) - F(x)F(y)$ for all $x, y \in \mathbb{R}$.

Density estimation

Kernel estimator - idea

Observation: $X_1, \ldots, X_n$ i.i.d., unknown density f.

Goal: estimate f without assuming some parametric model.

Idea: Put a "mountain" of mass $1/n$ at each observation X_i. Add all these mountains together to get an idea about the density.

Kernel estimator - construction

Choose:

- Kernel: a probability density K with mean 0 and variance 1. Determines the general shape of the mountains.
- Bandwidth: a number $h > 0$ that determines how we rescale each mountain. Small h gives very narrow and spiky mountains, large h gives wide and low mountains.

Define the kernel density estimator $\hat{f}$ by

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

Kernel estimator - (asymptotic) performance

Measure using (pointwise) mean square error $E_f(\hat{f}(x) - f(x))^2$ or the mean integrated square error

$$\int E_f(\hat{f}(x) - f(x))^2 \, dx.$$

Bias-variance decomposition:

$$E_f(\hat{f}(x) - f(x))^2 = (\underbrace{E_f \hat{f}(x) - f(x)}_{\text{bias}})^2 + \underbrace{Var_f \hat{f}(x)}_{\text{variance}}.$$

Kernel estimator - (asymptotic) performance

Theorem.
Suppose $f \in C^2$ and $\int |f''(x)|^2\,dx < \infty$. Suppose $\int |y|K^2(y)\,dy < \infty$. Then $\exists C_f > 0$ s.t. for small h,

$$\int E_f(\hat{f}(x) - f(x))^2\,dx \leq C_f\Big(\underbrace{h^4}_{\text{bias}^2} + \underbrace{\frac{1}{nh}}_{\text{variance}}\Big).$$

For $h \sim n^{-1/5}$, this gives

$$\int E_f(\hat{f}(x) - f(x))^2\,dx \leq C_f n^{-4/5}.$$

Remarks

- The rate of convergence is worse than the rate in parametric models.
- If we assume more smoothness on f, better rates can be attained: if $f \in C^\beta$ and $f^{(\beta)}$ is in L^2, can bound the MISE by $n^{-2\beta/(1+2\beta)}$. But then we need higher order kernels (see p. 67).
- This rate is essentially the best possible. It is possible to prove lower bounds that assert that no estimator can do better, uniformly for $f \in C^\beta$ such that the L^2-norm of $f^{(\beta)}$ is bounded by a constant (see Theorem 5.4).

Remarks

This situation is typical for nonparametric, or high-dimensional statistics:

- Parametric, smooth models: optimal rate is $1/\sqrt{n}$. Asymptotic variance of order $1/n$, asymptotic squared bias of lower order.
- Nonparametric models: optimal rate typically slower, depending on the "complexity" of the unknown quantity of interest. Need to balance squared bias and variance to obtain optimal rates.