THE BOOTSTRAP AND EDGEWORTH EXPANSION

YAN LIU

REFERENCE

- Peter Hall[1992]
  The Bootstrap and Edgeworth Expansion
- 吉田 [2006]
  数理統計学
- 前園 [2001]
  統計的推測の漸近理論

Part 1. preliminaries

1. INEQUALITY

Theorem 1.1 (Lyapunov's inequality)**.** *Let X be a random variable and $0 < s < r$. If $E(|X|^r) < \infty$, then*

$$\{E(|X|^s)\}^{1/s} \leq \{E(|X|^r)\}^{1/r}.$$

Corollary 1.2. *Suppose $p > q > 0$. If $E(|X|^p) < \infty$, then*

$$E(|X|^q) < \infty.$$

We will also give some inequalities for martingales.

Theorem 1.3. *Suppose $\{Y_n\}_{n=0,1,\ldots}$ is a martingale. Let $\{X_n\}$ be the martingale difference for $\{Y_n\}$, i.e, $X_k = Y_k - Y_{k-1}$, then*

- *for $1 \leq r < 2$, if $E(|X_k|^r) < \infty$ for all k, then*

$$E(|Y_n|^r) \leq 2 \sum_{k=1}^{n} E(|X_k|^r).$$

- *for $r \geq 2$, if $E(|X_k|)^r < \infty$ for all k, then*

$$E(|Y_n|^r) \leq C_r n^{r/2-1} \sum_{k=1}^{n} E(|X_k|^r),$$

where $C_r = [\, 8(r-1)\max(1, 2^{r-2})\,]^r$.

Date: September 10, 2016.

2. Edgeworth Expansion

Let $\phi_X(u) = E[\exp(iu'X)]$ be the characteristic function of X. Then the cumulant function is given by

$$\chi_r(u; X) = i^r \kappa_r[u'X] = (\partial_\epsilon)_0^r \log \varphi_X(\epsilon).$$

If we use tensor to rewrite it, then

$$\chi_r(u; X) = i^r \sum_{\alpha_1,\ldots,\alpha_r=1}^{d} u_{\alpha_1} \cdots u_{\alpha_r} \lambda^{\alpha_1 \cdots \alpha_r}[X],$$

where $\lambda^{\alpha_1 \cdots \alpha_r}[X]$ is defined as

$$\lambda^{\alpha_1 \cdots \alpha_r}[X] = (-i)^r (\partial_{u_{\alpha_1}})_0 \cdots (\partial_{u_{\alpha_r}})_0 \log \varphi_X(\epsilon).$$

Part 2. Principles of Bootstap

3. Introduction

If we wish to estimate a functional of a population function F, such as a population mean

$$\mu = \int x dF(x),$$

consider employing the same functional of the sample (or empirical) distribution function $\hat{F}$, which in this instance is the sample mean

$$\bar{X} = \int x d\hat{F}(x).$$

This argument is not always practicable– a probability density is just one example of a functional of F that is not directly amenable to this treatment.

The abbreviation "d.f." stands for "distribution function." The paris $(F, \hat{F})$, representing population d.f., sample d.f., will often be written as (F_0, F_1) during this chapter, hence making it easier to introduce bootstrap iteration, where for general $i \geq 1$, F_i denotes the d.f. of a sample drawn from F_{i-1} conditional on F_{i-1}. For the same reason, the pair $(\mathcal{X}, \mathcal{X}^*)$ will often be written as $(\mathcal{X}_1, \mathcal{X}_2)$.

An estimate $\hat{\theta}$ is a function of the data and may also be regarded as a functional of the sample distribution function $\hat{F}$. This we do by using square brackets for the former and round brackets for the latter:

$$\hat{\theta} = \theta[\mathcal{X}] = \theta(\hat{F}).$$

Formally, given a functional f_t from a class $\{f_t : t \in \tau\}$, we wish to determine that value t_0 of t that solves an equation

$$E\{ f_t(F_0, F_1) \,|\, F_0\} = 0, \tag{3.1}$$

where F_0 denotes the population distribution function and F_1 the distribution function "of the sample". we call (3.1) the population equation because we need properties of the population if we are to solve this equation exactly.

For example, correcting $\hat{\theta}$ for bias is equivalent to finding that value t_0 that solves (3.1) when

$$f_t(F_0, F_1) = \theta(F_1) - \theta(F_0) + t. \tag{3.2}$$

Our bias-corrected estimate would be $\hat{\theta}+t_0$. On the other hand, to construct a symmetric, 95% confidence interval for $\theta_0(\equiv \theta(F_0))$, solve (3.1) when

$$f_t(F_0, F_1) = I\{\theta(F_1) - t \le \theta(F_0) \le \theta(F_1) + t\} - 0.95. \tag{3.3}$$

The confidence interval is $(\hat{\theta} - t_0, \hat{\theta} + t_0)$, where $\hat{\theta} = \theta(F_1)$.

Similarly, replace the pair (F_0, F_1) in (3.1) by (F_1, F_2), thereby transforming (3.1) to

$$E\{\, f_t(F_1, F_2) \,|\, F_1\} = 0.$$

We call this the sample equation because we know everything about it once we know everything about it once we know the sample distribution function F_1. In particular, its solution $\hat{t}_0$ is a function of the sample values.

We shall refer to this iteration as "the bootstrap principle".

We call $\hat{t}_0$ and $E\{\, f_t(F_1, F_2) \,|\, F_1\}$ "the bootstrap estimates" of t_0 and $E\{\, f_t(F_0, F_1) \,|\, F_0\}$, respectively. They are obtained by replacing F_0 by F_1 in formulae for t_0 and $E\{\, f_t(F_0.F_1) \,|\, F_0\}$. In the bias correction problem, where f_t is given by (3.2), the bootstrap version of our bias-corrected estimate is $\hat{\theta}_0+\hat{t}_0$. In the confidence interval problem where (3.3) describes f_t, our bootstrap confidence interval is $(\hat{\theta}-\hat{t}_0, \hat{\theta}+\hat{t}_0)$. The latter is commonly called a (symmetric) percentile-method confidence interval for θ_0.

It is appropriate now to give detailed definitions of F_1 and F_2. There are at least two approaches, suitable for nonparametric and parametric problems respectively. In both, inference is based on a sample $\mathcal{X}$ of n random observations of the population. In the nonparametric case, if we denote the population by $\mathcal{X}_0$ then we have a nest of sampling operations, like our nest of dolls: $\mathcal{X}$ is drawn at random from $\mathcal{X}_0$ and $\mathcal{X}^*$ is drawn at random from $\mathcal{X}^*$. In the parametric case, F_0 is assumed completely known up to a finite vector $\boldsymbol{\lambda}_0$ of unknown parameters. To indicate this dependence we write $F_0 = F_{(\boldsymbol{\lambda}_0)}$, and element of a class $\{\, F_{(\boldsymbol{\lambda})}, \boldsymbol{\lambda} \in \Lambda \,\}$ of possible distributions. Let $\hat{\boldsymbol{\lambda}}$ be an estimate of $\boldsymbol{\lambda}_0$ computed from $\mathcal{X}$, then $F_1 = F_{(\hat{\lambda})}$, the distribution function obtained on replacing "true" parameter values by their sample estimates. Let $\mathcal{X}^*$ denote the sample drawn at random from the distribution with distribution function $F_{(\hat{\boldsymbol{\lambda}})}$ and let $\hat{\boldsymbol{\lambda}}^* = \boldsymbol{\lambda}[\,\mathcal{X}^*\,]$ denote the version of $\hat{\boldsymbol{\lambda}}$ computed for $\mathcal{X}^*$ instead of $\mathcal{X}$. Then $F_2 = F_{(\hat{\boldsymbol{\lambda}}^*)}$.

Part 3. Principles of Edgeworth Expansion

In this chapter we define, develop, and discuss Edgeworth expansions as approximations to distributions of estimates $\hat{\theta}$ os unknown quantities θ_0. We call θ_0 a "parameter", for want of a better term. Briefly, if $\hat{\theta}$ is constructed from a sample of size n, and if $n^{1/2(\hat{\theta}-\theta_0)}$ is asymptotically normally distributed with 0 mean and variance σ^2, then in a great many

cases of practical interest the distribution function of $n^{1/2}(\hat{\theta}-\theta_0)$ may be expanded as a power series in $n^{-1/2}$,

$$P\{n^{1/2}(\hat{\theta}-\theta_0)/\sigma \le x\} = \Phi(x) + n^{-1_2}p_1(x)\phi(x) + \cdots + n^{-j/2}p_j(x)\psi(x) + \ldots, \tag{3.4}$$

where $\phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$ is the standard normal density function and

$$\Phi(x) = \int_{-\infty}^{x} \phi(u)\, du$$

is the standard normal distribution function. Formula (3.4) is termed an *Edgeworth expansion*. The functions p_j are polynomials with coefficients depending on cumulants of $\hat{\theta}-\theta_0$.

For a general random variable Y with characteristic function χ, the jth cumulants, κ_j, of Y is defined to be the coefficient of $\frac{1}{j!}(it)^j$ in a power series expansion of $\log \chi(t)$:

$$\chi(t) = \exp\{\, \kappa_1 it + \frac{1}{2}\kappa_2(it)^2 + \cdots + \frac{1}{j!}\kappa_j(it)^j + \ldots\}.$$

Equivalently, since

$$\chi(t) = 1 + E(Y)it + \frac{1}{2}E(Y^2)(it)^2 + \cdots + \frac{1}{j!}E(Y^j)(it)^j + \ldots,$$

cumulants may be defined in terms via the formal identity

$$\sum_{j\ge 1}\kappa_j(it)^j = \log\left\{\, 1 + \sum_{j\ge 1}\frac{1}{j!}E(Y^j)(it)^j \right\} = \sum_{k\ge 1}(-1)^{k+1}\frac{1}{k}\left\{\sum_{j\ge 1}\frac{1}{j!}E(Y^j)(it)^j\right\}^k.$$

Equating coefficients of $(it)^j$ we deduce that

$$\left.\begin{aligned} \kappa_1 &= E(Y), \\ \kappa_2 &= E(Y^2) - (EY)^2 = \mathrm{Var}(Y), \\ \kappa_3 &= E(Y^3) - 3E(Y^2)E(Y) + 2(EY)^3 = E(Y-EY)^3, \\ \kappa_4 &= E(Y^4) - 4E(Y^3)E(Y) - 3(EY^2)^2 + 12E(Y^2)(EY)^2 - 6(EY)^4 \\ &= E(Y-EY)^4 - 3(\mathrm{Var}Y)^2, \end{aligned}\right\} \tag{3.5}$$

and so on, the formula for $\kappa_1 0$ containing 41 such terms.

Part 4. Bootstrap Curve Estimation

4. Nonparametric Density Estimation

Under the assumption that the density is univariate with at least two bounded derivatives, and using a nonnegative kernel function, the size of bandwidth that optimizes performance of the estimator in any L^p metric ($1 \le p < \infty$) is $n^{-1/5}$. The number of "parameter" needed to model the unknown density within a given interval is approximately equal to the number of bandwidths that can be fitted into that interval, and so is roughly of size

$n^{1/5}$. Thus, nonparametric density estimation (using a second-order kernel) involves the adaptive fitting of approximately $n^{1/5}$ parameters, this number growing with increasing n.

we would argue that the problems of point estimation and interval estimation are distinctly different, with different aims and different ends in mind. There is still a lot to be learned about the interval estimation problem in the context of density estimation, and our theoretical contributions here should be regarded only as a first step towards a deeper, more practically oriented appreciation.

4.1. **Different Bootstrap Methods in Density Estimation.** Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ denote a random sample drawn from a distribution with density f, and let K be a bounded function with the property that for some integer $r \geq 1$

$$\int_{-\infty}^{\infty} y^i K(y)\, dy \begin{cases} = 1 & \text{if } i = 0, \\ = 0 & \text{if } 1 \leq i \leq r-1, \\ \neq 0 & \text{if } i = r. \end{cases} \tag{4.1}$$

If

$$\int (1 + |y|^r)|K(y)|\, dy < \infty$$

then the integrals in (4.1) are well defined. A function K satisfying (4.1) is called an rth order kernel and may be used to construct a kernel-type nonparametric density estimator of $f(x)$,

$$\hat{f}(x) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\},$$

where h is the "bandwidth" or "window size" of the estimator. The bandwidth is a very important ingredient of the definition of $\hat{f}$. It determines the sizes of bias and variance, and so plays a major role in governing the performance of $\hat{f}$. Let think about the expectation of kernel-type nonparametric densities μ_j,

$$\mu_j(x) = h^{-1} E[\, K\{(x - X)/h\}^j \,] = \int K(y)^j f(x - hy)\, dy.$$

Then we can see that

$$\begin{aligned} b(x) = E\{\hat{f}(x)\} - f(x) &= h^r k_1 f^{(r)}(x) + o(h^r); \\ \mathrm{Var}\{\hat{f}(x)\} &= (nh)^{-1} k_2 f(x) + o\{(nh)^{-1}\}. \end{aligned}$$

5. Edgeworth expansion for time series

Let $U_T(u_1,\dots,u_p)'$ be a measurable function of a sample $X_1,\dots,X_T$. Suppose that all order of cumulants of U_T exist and satisfy the following:

$$\begin{aligned} c_i &= \operatorname{cum}(u_i) = T^{-1/2}c_i^{(1)} + T^{-1}c_i^{(2)} + o(T^{-1}),\\ c_{ij} &= \operatorname{cum}(u_i,u_j) = c_{ij}^{(1)} + T^{-1/2}c_{ij}^{(2)} + T^{-1}c_{ij}^{(3)} + o(T^{-1}),\\ c_{ijk} &= \operatorname{cum}(u_i,u_j,u_k) = T^{-1/2}c_{ijk}^{(1)} + T^{-1}c_{ijk}^{(2)} + o(T^{-1}),\\ c_{ijkm} &= \operatorname{cum}(u_i,u_j,u_k,u_m) = T^{-1}c_{ijkm}^{(1)} + o(T^{-1}). \end{aligned}$$

$$\begin{aligned} P(u_1<y_1,\dots,u_p<y_p) = \int_{-\infty}^{y_1}\cdots\int_{-\infty}^{y_p}\mathcal{N}(y;\Omega)\Big[1+\sum_i\Big(\frac{c_i^{(1)}}{\sqrt{T}}+\frac{c_i^{(2)}}{T}\Big)H_i(y)\\ +\frac{1}{2}\sum_{i,j}\Big(\frac{c_{ij}^{(2)}}{\sqrt{T}}+\frac{c_{ij}^{(3)}}{T}+\frac{c_i^{(1)}c_j^{(1)}}{T}\Big)H_{ij}(y)+\sum_{i,j,k}\Big(\frac{c_{ijk}^{(1)}}{6\sqrt{T}}+\frac{c_{ijk}^{(2)}}{6T}+\frac{c_i^{(1)}c_{jk}^{(2)}}{2T}\Big)H_{ijk}(y)\\ +\sum_{i,j,k,m}\Big(\frac{c_{ijkm}^{(1)}}{24T}+\frac{c_{ij}^{(2)}c_{km}^{(2)}}{8T}+\frac{c_i^{(1)}c_{jkm}^{(1)}}{6T}\Big)H_{ijkm}(y)+\sum_{i,j,i',j',k'}\frac{c_{ij}^{(2)}c_{i'j'k'}^{(1)}}{12T}H_{iji'j'k'}(y)\\ +\sum_{i,j,k,i',j',k'}\frac{c_{ijk}^{(1)}c_{i'j'k'}^{(1)}}{72T}H_{ijki'j'k'}(y)\Big]dy+o(T^{-1}), \end{aligned}$$

where $y=(y_1,\dots,y_p)'$, $\mathcal{N}(y;\Omega)=(2\pi)^{-p/2}|\Omega|^{1/2}\exp(-1/2y'\Omega^{-1}y)$, $\Omega=(c_{ij}^{(1)})$

$$H_{j_1,\dots j_s}(y)=\frac{(-1)^s}{\mathcal{N}(y;\Omega)}\frac{\partial^s}{\partial y_{j_1}\cdots\partial y_{j_s}}\mathcal{N}(y;\Omega).$$

The validity of Edgeworth expansion will be verified by the argument on the higher order cumulants and moments and Chebyshev's inequality.

6. words and phrases

(1) matryoshka dolls
each of a set of brightly painted hollow dolls of varying sizes, designed to nest inside one another.
(2) hollow
having a hole or empty space inside
(3) amenable
open and responsive to suggestion
(4) elaboration
(5) awkward causing difficulty; hard to do or deal with

(6) trifle
a thing of little value or importance
treat without seriousness or respect
(7) freckle
a small patch of light brown color on the skin, often becoming more pronounced through exposure to the sun
(8) patch

(9) refine

(10) facility
space of equipment necessary for doing something
(11) allude
suggest or call attention to indirectly; hint at
(12) ingredient
any of the foods or substances that are combined to make a particular dish