TRABAJOS DE ESTADISTICA
Vol. 5, núm. 2, 1990, pp. 23 a 32

ORDENES DE CONVERGENCIA PARA LAS APROXIMACIONES NORMAL Y BOOTSTRAP EN ESTIMACION NO PARAMETRICA DE LA FUNCION DE DENSIDAD*

CAO ABAD, R.**
Dpto. de Estadística e I.O.
Facultad de Matemáticas
Universidad de Santiago Compostela

RESUMEN

Este artículo concierne las distribuciones usadas para construir intervalos de confianza para la función de densidad en una situación no paramétrica. Se comparan los órdenes de convergencia para el límite normal, su aproximación «plug in» y el método bootstrap. Se deduce que el bootstrap se comporta mejor que las otras dos aproximaciones tanto en su forma clásica como con la aproximación bootstrap normal.

Palabras clave: Estimación no Paramétrica de la Ventana, Método Kernel, Bootstrap.

Clasificación AMS: 62G05.

ABSTRACT

Title: Rates of Convergence for the Normal and the Bootstrap Approximations in Nonparametric Density Estimation.

This paper is concerned with the distributions used to construct confidence intervals for the density function in a nonparametric situation. The rates of convergence for the normal limit, its plug in approach and the bootstrap method are compared. It turns out that the bootstrap performs better than the

Recibido noviembre 1989.
* Este trabajo ha obtenido el Premio «Ramiro Melendreras» correspondiente a la XVIII Reunión Nacional de Estadística, Investigación Operativa e Informática.
** Parte de este trabajo ha sido realizada durante una visita del autor en la Rechts- und Staatswissenschaftliche Fakultät, Wirtschaftstheorie Abteilung II, Universidad de Bonn.

two other approximations either with its classical shape or with the normal bootstrap approximation.

Key words: Nonparametric bandwidth estimation, Kernel method, Bootstrap.

AMS classification: 62G05.

1. INTRODUCCION

El presente artículo estudia varias posibilidades para la construcción de intervalos de confianza para la función de densidad f de una variable aleatoria monodimensional X en un punto x. Se consideran estimadores tipo kernel como el introducido por Parzen (1962) dado por:

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) \tag{1.1}$$

con $K_h(u) = h^{-1}K(u/h)$ y siendo K una función kernel.

Tomando como h un cierto estimador de la ventana ISE o MISE se tiene el siguiente resultado asintótico:

$$(nh)^{1/2}(\hat{f}_h(x) - f(x)) \xrightarrow{d} \mathcal{N}(B, V)$$

con

$$B = 1/2c_K^{5/2}d_k f''(x)$$
$$V = c_0 f(x)$$

siendo c_0 la constante que verifica $n^{1/5}h \xrightarrow{p} c_0$ y $c_K = \int K(t)^2\, dt$, $d_K = \int t^2 K(t)\, dt$.

Un posible modo de construcción de intervalos de confianza para $f(x)$ descansa en la aproximación de la distribución de $(nh)^{1/2}(\hat{f}_h(x) - f(x))$ mediante la de una $\mathcal{N}(\hat{B}, \hat{V})$ siendo $\hat{B}$ y $\hat{V}$ ciertos estimadores de B y V apropiados.

Una alternativa a este procedimiento es la utilización del método bootstrap. En este contexto la forma natural de realizar el remuestreo consiste en arrojar muestras $(X_1^*, X_2^*, ..., X_n^*)$ a partir de la densidad estimada $\hat{f}_g$. En efecto más adelante se justificará que la ventana usada

en el remuestreo bootstrap es sustancialmente distinta de la ventana h; de hecho es asintóticamente más grande. Construyendo los estimadores bootstrap $\hat{f}_h^*$ de modo análogo a (1.1), reemplazando la muestra inicial observada por la muestra bootstrap, se intenta aproximar la distribución de $(nh)^{1/2}(\hat{f}_h(x) - f(x))$ por la distribución bootstrap de $(nh)^{1/2}(\hat{f}_h^*(x) - \hat{f}_g(x))$.

Este trabajo constituye un estudio teórico comparativo entre los órdenes para la aproximación normal y para la aproximación bootstrap. En el segundo capítulo se establecen los resultados obtenidos cuyas demostraciones se exponen en la siguiente sección. Por último el apartado cuarto recoge algunas conclusiones obtenidas de todo ello.

2. RESULTADOS

A lo largo de este capítulo la letra h denotará la ventana MISE, es decir aquella que minimiza

$$E\left[\int (\hat{f}_h(x) - f(x))^2 \, dx\right]$$

y la letra g denotará la ventana piloto en el remuestreo bootstrap. A continuación se detallan las hipótesis utilizadas para probar los resultados que se presentan en esta sección:

(1) La función kernel K es simétrica, no negativa y verifica $d_K < \infty$, $c_K < \infty$.

2) La función de densidad f es cuatro veces diferenciable y su cuarta derivada está acotada.

3) La función K es cuatro veces diferenciable y además la función $t^5 K^{IV}(t)$ está acotada (para ello basta pedir que K^{IV} sea continua y que $\lim_{t \to \infty} t^5 K^{IV}(t) = \text{cte} \in \mathbb{R}$).

Se usará frecuentemente la letra Φ para denotar la función de distribución de una $\mathcal{N}(0, 1)$. Así mismo usaremos también las siguientes notaciones:

$$B_n = (nh)^{1/2}(EK_h(x - X_1) - f(x))$$

$$V_n = h \operatorname{Var} K_h(x - X_1)$$

Una vez definidos estos términos podemos establecer el siguiente resultado:

Lema 1: Bajo las condiciones (1) y (2) se verifica

$$\sup_{z\in\mathbb{R}}\left|P\{(nh)^{1/2}(\hat{f}_h(x)-f(x))\leqslant z\}-\Phi\left(\frac{z-B_n}{V_n^{1/2}}\right)\right|=O(n^{-2/5}) \tag{2.1}$$

lo cual implica que

$$\sup_{z\in\mathbb{R}}\left|P\{(nh)^{1/2}(\hat{f}_h(x)-f(x))\leqslant z\}-\Phi\left(\frac{z-B}{V^{1/2}}\right)\right|=O(n^{-2/5}) \tag{2.2}$$

Como ya que se ha indicado en la sección 1 la expresión (2.2) no se puede utilizar directamente para construir intervalos de confianza. En realidad hemos de estimar B y V previamente. El estimador natural de V es $\hat{V}=c_K\hat{f}_h(x)$, obteniéndose $\hat{V}-V=O_p(h^4+n^{-1}h^{-1})^{1/2})=O_p(n^{-2/5})$ sin embargo el problema es distinto para B ya que la estimación de la segunda derivada conduce a la elección de otra ventana g de orden $n^{-1/9}$ que es la que minimiza $\text{Var}\,\hat{f}_g''(x)+E(\hat{f}_g''(x)-f''(x))^2$. A partir de aquí el estimador de B se define como $\hat{B}=1/2(nh^5)^{1/2}d_K\hat{f}_g''(x)$ y se tiene que $\hat{B}-B=O_p((n^{-1}g^{-5}+g^4)^{1/2})=O_p(n^{-2/9})$.

Estos hechos junto con la expresión (2.2) nos permiten enunciar el siguiente teorema:

Teorema 1: Las condiciones (1) y (2) implican

$$\sup_{z\in\mathbb{R}}\left|P\{(nh)^{1/2}(\hat{f}_h(x)-f(x))\leqslant z\}-\Phi\left(\frac{z-B}{\hat{V}^{1/2}}\right)\right|=O_p(n^{-1/5}) \tag{2.3}$$

Definamos ahora el sesgo y la varianza bootstrap

$$B_n^*=(nh)^{1/2}(E^*K_h(x-X_1^*)-\hat{f}_g(x))$$
$$V_n^*=h\,\text{Var}^*\,K_h(x-X_1^*)$$

entonces se verifica el siguiente resultado

Lema 2: Si se cumplen las condiciones (1) y (2) entonces

$$\sup_{z\in\mathbb{R}}\left|P^*\{(nh)^{1/2}(\hat{f}_h^*(x)-\hat{f}_g(x))\leqslant z\}-\Phi\left(\frac{z-B_n^*}{V_n^{*1/2}}\right)\right|=O_p(n^{-2/5}) \tag{2.4}$$

donde g es una ventana de orden $n^{-1/9}$.

Se definen ahora

$$\hat{B}_n^* = 1/2(nh^5)^{1/2} d_K \hat{f}_g''(x)$$

y

$$\hat{V}_n^* = c_K \hat{f}_g(x) - h\hat{f}_g(x)^2.$$

Finalmente enunciamos el siguiente teorema:

Teorema 2: Bajo las condiciones (1), (2) y (3) se verifica

$$\sup_{z \in \mathbb{R}} \Big| P\{(nh)^{1/2}(\hat{f}_h(x) - f(x)) \leqslant z\} - \\ - P^*\{nh)^{1/2}(\hat{f}_h^*(x) - \hat{f}_g(x)) \leqslant z\} \Big| = O_p(n^{-2/9}) \tag{2.5}$$

y además

$$\sup_{z \in \mathbb{R}} \left| P\{(nh)^{1/2}(\hat{f}_h(x) - f(x)) \leqslant z\} - \Phi\left(\frac{z - \hat{B}_n^*}{\hat{V}_n^{*1/2}}\right) \right| = O_p(n^{-2/9}) \tag{2.6}$$

Todos estos resultados que sirven como comparación entre ambos métodos se demuestran en la sección siguiente.

3. DEMOSTRACIONES

Demostración del lema 1:

Se consideran las variables aleatorias

$$Z_i = K_h(x - X_i) - EK_h(x - X_i) \quad , \quad i = 1, 2, ..., n$$

que son independientes e idénticamente distribuidas y además verifican

$$EZ_1 = 0$$
$$\sigma^2 = EZ_1^2 = h^{-1} c_K f(x) - f(x)^2 + O(h)$$
$$E|Z_1|^3 \leqslant O(h^{-2})$$

como se puede comprobar haciendo cambio de variable dentro de las correspondientes integrales y luego utilizando desarrollos de Taylor.

Aplicando la desigualdad de Berry-Esseen, ver Petrov (1975), se obtiene

$$\sup_{z\in\mathbb{R}}\left|P\left\{\frac{1}{\sigma n^{1/2}}\sum_{i=1}^{n}Z_i\leqslant z\right\}-\Phi(z)\right|\leqslant A\frac{E|Z_1|^3}{\sigma^3 n^{1/2}}$$

para una cierta constante absoluta A.

Teniendo en cuenta los cálculos previos se llega a

$$\frac{E|Z_1|^3}{\sigma^3 n^{1/2}}=O((nh)^{-1/2})=O(n^{-2/5}).$$

Por otra parte

$$\hat{f}_h(x)-E\hat{f}_h(x)=n^{-1}\sum_{i=1}^{n}Z_i$$

con lo cual

$$\sup_{z\in\mathbb{R}}\left|P\left\{\frac{(nh)^{1/2}(\hat{f}_h(x)-E\,\hat{f}_h(x))}{(\sigma^2 h)^{1/2}}\leqslant z\right\}-\Phi(z)\right|=O(n^{-2/5})$$

que con la notación introducida puede escribirse como

$$\sup_{z\in\mathbb{R}}\left|P\left\{\frac{(nh)^{1/2}(\hat{f}_h(x)-f(x))-B_n}{V_n^{1/2}}\leqslant z\right\}-\Phi(z)\right|=O(n^{-2/5})$$

de la cual se deduce inmediatamente (2.1).

Obsérvese que

$$V_n=\sigma^2 h=c_K f(x)-hf(x)^2+O(h^2) \tag{3.1}$$

de modo que $V_n-V=O(h)=O(n^{-1/5})$. Además cálculos directos permiten afirmar

$$B_n=1/2(nh^5)^{1/2}d_K f''(x)+O((nh^9)^{1/2}) \tag{3.2}$$

así que $B_n-B=O(n^{-2/5})$. Utilizando además la acotación de las funciones $\Phi'(t)$ y $t\Phi'(t)$ se deduce

$$\sup_{z\in\mathbb{R}}\left|\Phi\left(\frac{z-B_n}{V_n^{1/2}}\right)-\Phi\left(\frac{z-B}{V^{1/2}}\right)\right|=O((B_n-B)+(V_n-V))=O(n^{-1/5}).$$

Este hecho junto con (2.1) implica (2.2).

Demostración del teorema 1:

La expresión (2.3) es evidente a partir de (2.2) y de que $\hat{B} - B = O_p(n^{-2/9})$ y $\hat{V} - V = O_p(n^{-2/5})$.

Demostración del lema 2:

Sigue los mismos pasos que la del lema 1, pero para los análogos bootstrap.

Se definen las variables

$$Z_i^* = K_h(x - X_i^*) - E^*K_h(x - X_i^*) \quad , \quad i = 1, 2, ..., n$$

que son independientes, con distribución bootstrap idéntica y verifican

$$E^*Z_1^* = 0$$
$$\sigma^{*2} = E^*Z_1^{*2} = h^{-1}c_K\hat{f}_g(x) - \hat{f}_g(x)^2 + O_p(h)$$
$$E^*|Z_1^*|^3 \leqslant O_p(h^{-2})$$

De aquí se deduce que

$$\frac{E^*|Z_1^*|^3}{\sigma^{*3}n^{1/2}} = O_p((nh)^{-1/2}) = O_p(n^{-2/5})$$

luego la desigualdad de Berry-Esseen permite concluir

$$\sup_{z\in\mathbb{R}}\left|P^*\left\{\frac{1}{\sigma^* n^{1/2}}\sum_{i=1}^{n} Z_i^* \leqslant z\right\} - \Phi(z)\right| = O_p(n^{-2/5}).$$

Teniendo ahora en cuenta que

$$\hat{f}_h^*(x) - E^*\hat{f}_h^*(x) = n^{-1}\sum_{i=1}^{n} Z_i$$

y la notación introducida en la sección previa se llega a

$$\sup_{z\in\mathbb{R}}\left|P\left\{\frac{(nh)^{1/2}(\hat{f}_h^*(x) - \hat{f}_g(x)) - B_n^*}{V_n^{*1/2}} \leqslant z\right\} - \Phi(z)\right| = O_p(n^{-2/5})$$

que es equivalente a (2.4).

Demostración del teorema 2:

De los cálculos anteriores se deduce que

$$V_n^* = \sigma^2 h = c_K \hat{f}_g(x) - h\hat{f}_g(x)^2 + O_p(h^2). \tag{3.3}$$

Por otra parte, mediante cálculos directos se obtiene

$$B_n^* = 1/2(nh^5)^{1/2} d_K \hat{f}_g''(x) + O_p((nh^9)^{1/2}). \tag{3.4}$$

Teniendo en cuenta (3.1) y (3.2) se concluye

$$V_n^* - V_n = c_K(\hat{f}_g(x) - f(x)) - h(\hat{f}_g(x)^2 - f(x)^2) + O_p(n^{-2/5})$$
$$B_n^* - B_n = 1/2(nh^5)^{1/2} d_K(\hat{f}_g''(x) - f''(x)) + O_p(n^{-2/5}).$$

Calculando el sesgo y la varianza de $\hat{f}_g(x)$ y de $\hat{f}_g''(x)$ se llega a las siguientes expresiones:

$$E(\hat{f}_g(x) - f(x))^2 = O(n^{-1}g^{-1}c_K f(x) + 1/4 d_K^2 f''(x) g^4)$$
$$E(\hat{f}_g''(x) - f''(x))^2 = O(n^{-1}g^{-5}c_K f''(x) + 1/4 d_K^2 f^{IV}(x) g^4)$$

De modo que

$$\sup_{z\in\mathbb{R}} \left| \Phi\left(\frac{Z - B_n^*}{V_n^{*1/2}}\right) - \Phi\left(\frac{z - B_n}{V_n^{1/2}}\right) \right| = O_p((B_n^* - B_n) + (V_n^* - V_n)) =$$
$$= O_p((n^{-1}g^{-5}c_K f''(x) + 1/4 d_K^2 (f''(x)^2 + f^{IV}(x)^2) g^4)^{1/2}) + O_p(n^{-2/5}),$$

pero esta última función de g se minimiza para

$$g = \left(\frac{5 c_K f''(x)}{n d_K^2 (f''(x)^2 + f^{IV}(x)^2)} \right)^{1/9} \tag{3.5}$$

De modo que la ventana óptima en el remuestreo bootstrap es de orden $n^{-1/9}$ y además para esa ventana se tiene

$$\sup_{z\in\mathbb{R}} \left| \Phi\left(\frac{z - B_n^*}{V_n^{*1/2}}\right) - \Phi\left(\frac{z - B_n}{V_n^{1/2}}\right) \right| = O_p(n^{-2/9}). \tag{3.6}$$

Esta última expresión junto con (2.1) y (2.4) permite probar (2.5).

Para probar (2.6) obsérvese la definición de $\hat{B}_n^*$ y $\hat{V}_n^*$ y téngase en cuenta las expresiones (3.3) y (3.4). Entonces

$$V_n^* - \hat{V}_n^* = O_p(n^{-2/5}) \quad \text{y} \quad B_n^* - \hat{B}_n^* = O_p(n^{-2/5}).$$

Estas igualdades junto con (3.6) conducen a

$$\sup_{z\in\mathbb{R}}\left|\Phi\left(\frac{z-\hat{B}_n^*}{\hat{V}_n^{*1/2}}\right)-\Phi\left(\frac{z-B_n}{V_n^{1/2}}\right)\right| = O_p(n^{-2/9})$$

de la cual se sigue (2.6) sin más que utilizar la expresión (2.1).

4. COMENTARIOS

Lo primero que se advierte es que la aproximación bootstrap (B) dada por (2.5) presenta un orden mejor que el de la teoría normal (TN) como puede observarse en (2.2). Del mismo modo la aproximación plug in (PI), consistente en sustituir los valores de la media y de la varianza en (2.2) por ciertos estimadores, proporciona un orden peor que el bootstrap, como se puede apreciar en (2.3). A modo ilustrativo se pueden reflejar los órdenes obtenidos en la siguiente tabla

TN	PI	B
$O(n^{-1/5})$	$O_p(n^{-1/5})$	$O_p(n^{-2/9})$

Los resultados obtenidos aquí para la densidad son análogos a los relativos a la función de regresión como puede verse en Cao Abad (1989). En ese otro contexto el método bootstrap utilizado fue el Wild Bootstrap.

La expresión (2.6), que proporciona el orden $n^{-2/9}$ y que es el mismo que el del método bootstrap, puede ser interpretada como una aproximación bootstrap normal en el sentido de que sustituye la distribución bootstrap por la correspondiente distribución normal que se ajusta a la bootstrap en media y varianza. Otro punto de vista consiste en considerar (2.6) como una aproximación plug in de la expresión (2.1) que no es directamente utilizable para la construcción de intervalos de confianza. Esto es el plug in no se realiza en la distribuión asintótica sino en el paso inmediatamente anterior.

A la vista de esto las expresiones (2.5) y (2.6) resultan las más favorables. No obstante un problema que presentan es el de la elección de la ventana piloto g. Aunque disponemos de una expresión asintótica para la ventana piloto óptima, dada por (3.5), ésta depende de valores

desconocidos. Una posibilidad sería estimar mediante plug in la expresión (3.5). Otra alternativa consistiría en estimar

$$\int (\hat{f}_g(x) - f(x))^2 \, dx + \int (\hat{f}''_g(x) - f''(x))^2 \, dx$$

u otra función de g que presente el mínimo en el mismo punto que ésta y después cross-validar dicha estimación para finalmente minimizar en g. En cualquiera de los dos casos se trata de generalizaciones de métodos de selección de la ventana (plug in o cross validation en los casos mencionados) a la situación que se nos plantea.

Es interesante resaltar que el orden de la ventana piloto es el mismo que el obtenido en el contexto de la función de regresión para el remuestreo «Wild». Esto puede observarse en Härdle y Marron (1989) o bien en Cao Abad (1989).

AGRADECIMIENTO

El autor desea agradecer a Wolfgang Härdle la sugerencia de este estudio así como varias aportaciones de interés.

REFERENCIAS

CAO ABAD, R. (1989): «Rates of Convergence for the Normal and the Bootstrap Approximations in Nonparametric Regression». Manuscrito.

HÄRDLE, W. y MARRON, J. S. (1990): «Bootstrapping Simultaneous Error Bars for Nonparametric Regression», aparecerá próximamente en *The Annals of Statistics.*

PARZEN, E. (1962): «On Estimating Probability Density and Mode», *Annals of Mathematical Statistics,* 35, 1065-1076.

PETROV, V. (1975): *Sums of Independent Random Variables,* Springer Verlag, Berlin-Heidelberg-Nueva York.