XI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 24–25–26 de outubro de 2013

ESTUDIO COMPARATIVO DE SELECTORES DE VENTANA. ESTIMACIÓN TIPO NÚCLEO DE LA DESNSIDAD.

Alejandro Saavedra Nieves*

*Máster en Técnicas Estadísticas
Universidade de Santiago de Compostela

RESUMEN

El problema de selección del parámetro ventana en la estimación tipo núcleo de la densidad es un problema clásico de la Estadística no Paramétrica. Mientras la elección de la función tipo núcleo no es determinante para reconstruir la función de densidad, seleccionar adecuadamente el parámetro ventana resulta decisivo. En este trabajo, revisamos de forma breve algunos de los algoritmos clásicos para elegir el parámetro de suavizado para, posteriormente, compararlos a través de un estudio de simulación. Como criterio de error hemos considerado el Error Cuadrático Integrado. Además, el comportamiento de las ventanas obtenidas será analizado.

Palabras y frases clave: Estadística no Paramétrica, estimador tipo núcleo, parámetro ventana o de suavizado.

1. INTRODUCCIÓN

El estimador tipo núcleo de la densidad de una variable aleatoria X se define como

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

donde K es una función de densidad llamada núcleo, h denota el parámetro ventana y $\{X_1, \ldots, X_n\}$ representa una muestra aleatoria simple de X. Este estimador ha sido ampliamente estudiado en la literatura. La elección del núcleo no es demasiado influyente en la estimación. Sin embargo, seleccionar el parámetro de suavizado h resulta determinante. Valores altos para el parámetro h proporcionan estimaciones sobresuavizadas de la densidad. Valores bajos, provocan el efecto contrario.

En la Sección 2, realizaremos una breve descripción de algunos de los selectores de ventana más utilizados. En la Sección 3, a través de un estudio de simulación, serán comparados empleando como densidades de prueba las propuestas en Marron y Wand (1992). Como criterio de error hemos considerado el ISE definido como

$$ISE(h) = \int (\hat{f}_h(x) - f(x))^2 dx.$$

Su valor será aproximado empleando el método de los trapecios compuesto. En la Sección 4, exponemos las conclusiones obtenidas del estudio realizado.

2. ALGUNOS SELECTORES CLÁSICOS DEL PARÁMETRO DE SUAVIZADO

En esta sección, presentamos algunos de los selectores de ventana más utilizados.
El Error Cuadrático Medio Integrado asintótico se puede escribir como

$$AMISE(h) = \frac{1}{nh}R(K) + \frac{1}{4}h^2\mu_2(K)^2 R(f^{''})$$

con $\mu_2(K) = \int u^2 K(u)\ du < \infty$ y $R(K) = \int K(u)^2\ du$.

Selector de ventana normal o regla del pulgar: Este algoritmo reemplaza la parte desconocida de la expresión de la ventana $AMISE(h)$, $R(f'')$, por una estimación basada en una familia parámetrica. Jones, Marron y Sheather (1996) consideran $h_{NS} = 1.06\hat{\sigma}n^{-1/5}$, siendo $\hat{\sigma}$ la estimación de la desviación típica y 1.06 el valor de la parte constante de la ventana $AMISE$ suponiendo que nuestra densidad es normal. Posteriormente, se propone una corrección a ese término, considerando en lugar de $\hat{\sigma}$ el mínimo entre ese valor y $\hat{\sigma}_{IQR}$, donde

$$\hat{\sigma}_{IQR} = \frac{IQR}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)}$$

y Φ denota la función de distribución de una normal estándar. Scott y Terrell (1985) y Terrell (1990) desarollaron un método para la estimación de la ventana que minimiza $R(f'')$. Dado que el parámetro de localización de dicha familia no es importante, debemos estudiar únicamente la escala, de ahí que una elección natural sea la familia $N(0, \sigma^2)$.

Validación cruzada insesgada: Bowman (1984) propone elegir como ventana el valor de h que minimiza la siguiente expresión:

$$\int \hat{f}_h^2(y)dy - 2\int \hat{f}_h(x)f(x)dx.$$

El primer sumando, puede reescribirse como $(1/n)\sum_{i=1}^{n}\int \hat{f}_{-i}^2(y)dy$ mientras que el segundo puede ser estimado según el método de los momentos. Así, podemos escribir

$$\frac{1}{n}\sum_{i=1}^{n}\int \hat{f}_{-i}^2(y)dy - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i),$$

donde $\hat{f}_{-i}$ denota el estimador tipo núcleo eliminando el dato X_i del cálculo. Denoratemos este método por UCV.

Validación cruzada sesgada: El método de validación cruzada sesgada minimiza la expresión del $AMISE(h)$ descrita al principio de la sección. Reemplazando $R(f'')$ por $R(\hat{f}_h'') - 1/(nh^5)R(K'')$ obtenemos la expresión $BCV(h)$ a minimizar, ver Seather (2004). Este método será denotado por BCV.

Selector *plug-in*: De acuerdo con el método del pulgar, los métodos $plug-in$ reemplazan $R(f'')$ por $R(\hat{f}_g'')$, siendo $\hat{f}_g$ un estimador tipo núcleo de la densidad y g una ventana piloto prefijada de antemano. Sheather y Jones (ver Wand y Jones, 1995) y la regla plug-in directa son dos selectores de este grupo. Serán denotados por SJ-ste y SJ-dpi, respectivamente.

3. COMPARACIÓN DE SELECTORES DE VENTANA A TRAVÉS DE UN ESTUDIO DE SIMULACIÓN

En esta sección, exponemos algunos de los resultados del estudio de simulación realizado para comparar los selectores de ventana expuestos en la Sección 2. Hemos tomado como densidades de prueba las propuestas por Marron y Wand (1992). Para cada una de ellas, hemos generado 1000 muestras de tamaño 1000. Para cada muestra, hemos estimado la función de densidad usando los distintos selectores de ventana y calculado el ISE asociado a cada estimación.

En la Tabla 1, mostramos las medias de los errores ISE para cada selector. A partir de los resultados obtenidos, el selector de ventana de validación cruzada insesgada (UCV) es el más competitivo para los modelos 3, 10, 12, 13, 14 y 15. En general, salvo el modelo 3, son densidades con gran número de modas. La regla del pulgar (selector normal) solo ofrece buenos resultados para densidades unimodales, concretamente para las densidades 1, 2 y 5. Los peores resultados para el resto de densidades son los obtenidos con este selector. Para el selector de validación cruzada sesgada (BCV), los mejores resultados son obtenidos para las densidades 1, 2, 4, 5 y 6.

Modelo	Normal	UCV	BCV	SJste	SJdpi
1	0.0011	0.0014	0.0011	0.0011	0.0011
2	0.0016	0.0020	0.0016	0.0016	0.0016
3	0.0912	0.0091	0.0103	0.0118	0.0198
4	0.0435	0.0085	0.0080	0.0080	0.0090
5	0.0106	0.0144	0.0106	0.0108	0.0107
6	0.0020	0.0017	0.0015	0.0015	0.0015
7	0.0141	0.0023	0.0021	0.0020	0.0021
8	0.0033	0.0022	0.0021	0.0020	0.0020
9	0.0034	0.0020	0.0021	0.0019	0.0020
10	0.0430	0.0071	0.0441	0.0091	0.0351
11	0.0034	0.0032	0.0030	0.0029	0.0029
12	0.0182	0.0067	0.0150	0.0090	0.0124
13	0.0065	0.0050	0.0057	0.0054	0.0055
14	0.0580	0.0105	0.0174	0.0164	0.0219
15	0.0770	0.0092	0.0193	0.0182	0.0213

Tabla 1: Valores medios del ISE para cada densidad de Marron y Wand (1992) y cada uno de los selectores propuestos.

	Media						Varianza				
Modelo	Normal	UCV	BCV	SJste	SJdpi	W-J	Normal	UCV	BCV	SJste	SJdpi
1	0.264	0.249	0.277	0.260	0.261	1	0.007	0.056	0.012	0.016	0.015
2	0.198	0.179	0.193	0.180	0.181	2	0.008	0.039	0.012	0.011	0.010
3	0.248	0.043	0.054	0.063	0.088	3	0.014	0.006	0.011	0.004	0.004
4	0.127	0.040	0.047	0.048	0.056	4	0.012	0.007	0.004	0.003	0.004
5	0.030	0.022	0.029	0.026	0.026	5	0.001	0.008	0.002	0.002	0.002
6	0.320	0.219	0.242	0.223	0.232	6	0.005	0.044	0.021	0.015	0.013
7	0.421	0.156	0.165	0.164	0.173	7	0.004	0.026	0.008	0.005	0.004
8	0.292	0.171	0.199	0.182	0.197	8	0.006	0.033	0.023	0.015	0.013
9	0.340	0.183	0.232	0.206	0.224	9	0.005	0.037	0.030	0.015	0.012
10	0.231	0.051	0.248	0.074	0.172	10	0.005	0.007	0.006	0.010	0.012
11	0.319	0.210	0.241	0.222	0.231	11	0.005	0.051	0.019	0.015	0.013
12	0.296	0.070	0.245	0.140	0.199	12	0.006	0.016	0.089	0.014	0.014
13	0.317	0.122	0.238	0.209	0.220	13	0.005	0.051	0.017	0.013	0.011
14	0.438	0.052	0.121	0.115	0.159	14	0.006	0.005	0.021	0.005	0.004
15	0.450	0.051	0.124	0.114	0.147	15	0.006	0.001	0.004	0.003	0.003

Tabla 2: Valores medios de ventana y varianza para cada densidad de Marron y Wand (1992) y cada uno de los selectores propuestos.

Los métodos SJ proporcionan resultados similares salvo para los modelos 3, 10, 12, 14 y 15. Este selector aparece como la mejor opción para gran parte de las densidades.

En la Tabla 2, exponemos las medias de las ventanas seleccionadas y su variabilidad para los distintos selectores de ventana. Podemos observar que el método de validación cruzada insesgada proporciona, en media, ventanas de menor tamaño. Sin embargo, con el selector de ventana normal, dicho valor es el mayor de todos los selectores para casi todas las densidades estudiadas, a excepción de las densidades 1 y 10 en las que los mayores tamaños de ventana se alcanzan con el selector BCV.

Al estudiar la variabilidad de las ventanas, vemos que sus valores mínimos se alcanzan para el selector de ventana normal para la mayor parte de las densidades. Para las densidades 1, 3 y 4, la menor varianza se consigue con el selector $plug - in$. Para la densidad 14, el selector UCV minimiza dicho valor.

4. CONCLUSIONES

Como conclusión, no existe un selector de ventana que sea universalmente el más competitivo. El comportamiento de cada método depende, en gran medida, del modelo de prueba considerado y de las propiedades del mismo.

REFERENCIAS

A. Bowman (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, **71**, 353–360.

M.C. Jones, J.S. Marron y S.J. Sheather (1992). Progress in Data- Based Bandwidth Selection for Kernel Density Estimation *(unpublished)*.

M.C. Jones, J.S. Marron y S.J. Sheather (1996). A brief survey of bandwidth selection for density estimation, *J. Amer. Statist. Assoc.*, **91**, 401–407.

S. Marron y M. Wand (1992). Exact Mean Integrated Squared Error, *Annals of Statistcs*, **20**, 712–736.

D.W. Scott y G.R. Terrell (1985). Oversmoothed nonparametric density estimates, *J. Amer. Statist. Assoc.*, **80**, 209–214.

S.J. Sheather (2004). Density Estimation, *Statistical Science*, **1009**, 19, No. 4, 588–597 34–45.

G.R. Terrell (1990). The maximal smoothing principle in density estimation. *J. Amer. Statist. Assoc.*, **85**, 470–477.

M.P. Wand y C. Jones (1995). *Kernel Smoothing*, Chapman and Hall, London.