Servy, Elsa
Cuesta, Cristina
Marí, Gonzalo
Instituto de Investigaciones Teóricas y Aplicadas en Estadística, Escuela de Estadística

COMPARACIÓN DEL COMPORTAMIENTO DE DIVERSOS ESTIMADORES BASADOS EN NÚCLEOS[*]

1. INTRODUCCIÓN

Este trabajo está dirigido a profundizar y difundir métodos no paramétricos para la estimación de funciones de densidad (métodos de suavizado). Estos métodos hacen pre-supuestos mínimos sobre las densidades que gobiernan las frecuencias observadas de las variables estadísticas.

En los análisis paramétricos se comienza haciendo supuestos rígidos sobre la estructura básica de los datos. Luego se estiman de la manera más eficiente posible los parámetros que definen la estructura. A posteriori se decide si los supuestos iniciales son aceptables. Esta lógica de pensamiento conlleva, muchas veces, círculos viciosos que oscurecen la objetividad del análisis.

Los métodos de suavizado, en cambio, comienzan aceptando su subjetividad y buscan desprenderse de ella a través de métodos de prueba y error tomando como base resultados matemáticos asintóticos. Los fundamentos de los métodos de suavizado son antiguos pero sólo lograron el estado actual de desarrollo gracias a los avances de la ciencia de la computación y los estudios por simulación han permitido evaluar sus comportamientos.

Los métodos paramétricos y no paramétricos, en principio antagónicos, suelen ser usados en forma simultánea en el análisis de conjuntos de datos. Los métodos no paramétricos pueden ayudar en el inicio de la investigación a descubrir la estructura probabilística que gobierna los datos de modo que los supuestos del análisis paramétrico estén bien fundamentados. Después de realizados los análisis, suelen ser utilizados nuevamente para el estudio de los residuos, buscando validar la elección del modelo.

Entre los métodos de estimación de funciones de densidad de probabilidad se encuentran aquellos estimadores basados en núcleos. Estos estimadores logran funciones de densidad suavizadas que se construyen en cada punto del eje real de acuerdo con los valores muestrales más cercanos al mismo que constituyen un entorno denominado “ventana”. Estos valores son ponderados de modo que, por ejemplo, los vecinos más cercanos tengan mayor peso que los más alejados dentro de una ventana de datos. Se pueden utilizar diversas funciones de ponderación (llamadas K o “Kernel”) que son justamente los núcleos en que se basan los estimadores. Las propiedades de las curvas de estimación dependen de la elección del núcleo y del

[*] Trabajo realizado en el marco del proyecto de investigación “Métodos semiparamétricos y no paramétricos para el análisis de regresión con datos univariados y multivariados”. ECO 025. Secretaría de Ciencia y Tecnología. Universidad Nacional de Rosario.

ancho de la ventana. La combinación de la función de ponderación, el ancho de la ventana, el tamaño de muestra y la forma de la densidad verdadera (más o menos "rugosa", con más o menos modos, etc) hacen a la bondad de la estimación resultante.

Lo que se ensaya en este trabajo es una evaluación de la bondad de la estimación de dos funciones de densidad, una unimodal y otro bimodal, cuando se utilizan distintos núcleos para los estimadores y diferentes tamaños de muestra. El ancho de ventana utilizado es el que asintóticamente se considera óptimo. Justamente el motivo de realizar un estudio de simulación es verificar hasta qué punto los resultados asintóticos tienen vigencia cuando la muestra es de tamaño chico o moderado.

2. METODOLOGÍA

Al momento de estimar una función de densidad, el histograma y el polígono de frecuencias son estimadores válidos y de fácil construcción e interpretación. Sin embargo, tienen la desventaja de depender del ancho de la "barra" (o del número de "barras") y del rango de variación del histograma que pueden influir en gran medida en su forma. Por otro lado su forma no es suave y no es sensible a las propiedades locales de la función que se desea estimar. Surge, entonces, la necesidad de construir otro estimador de la densidad que intente diluir dichas desventajas.

Considerando la función de densidad:

$$f(x) \equiv \frac{\partial F(x)}{\partial x} \equiv \lim_{\to\infty} \frac{F(x+h) - F(x-h)}{2h}$$

(donde h es el ancho de la ventana) y estimando la función de distribución acumulada teórica por la empírica se obtiene,

$$\hat{f}(x) = \frac{\#\{x_i \in (x-h, x+h)\}}{2nh}$$

ó, equivalentemente,

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad \text{donde} \quad K(u) = \begin{cases} 1/2, & \text{si } |u| < 1 \\ 0, & \text{en otro caso} \end{cases} \qquad (1)$$

Este estimador puede modificarse considerando otras funciones de núcleo K, dando origen al estimador de densidad basado en núcleos (en inglés Kernel Density Estimator).

Este estimador es más "local" en su naturaleza que el histograma o el polígono de frecuencias. La forma aditiva de (1) traslada implícitamente las propiedades de continuidad y diferenciabilidad de K a $\hat{f}$. Si la función K es discontinua (como la presentada en (1)), también lo será la estimación de la densidad basada en esa función Kernel. Por otro lado, el parámetro de suavizado "h" tiene un fuerte efecto sobre el grado de "suavidad" en la estimación de la densidad.

La función K debe definirse de manera que satisfaga las siguientes condiciones:

$$\int K(u)du = 1, \quad \int uK(u)du = 0, \quad \int u^2 K(u)du = \sigma_K^2 > 0$$

Algunas medidas para evaluar a $\hat{f}(x)$ como estimador de $f(x)$ son:

1. Error cuadrático: $SE(x) = \left[\hat{f}(x) - f(x)\right]^2$

2. Error cuadrático esperado (error cuadrático medio): $\mathrm{MSE}(x) = \mathrm{E}_f\left[\hat{f}(x) - f(x)\right]^2$.
3. Error cuadrático integrado: $\mathrm{ISE} = \int_{-\infty}^{+\infty}\left[\hat{f}(u) - f(u)\right]^2 du$
4. Error cuadrático integrado esperado (o promedio) (MISE)
5. Error cuadrático integrado esperado (o promedio) asintótico (AMISE)

El ajuste más adecuado lo logrará aquella $\hat{f}(x)$ que logre el mejor balance entre sesgo y variancia y puede cuantificarse a través del MISE.

Si K cumple las condiciones recién indicadas y suponiendo que la densidad subyacente es suficientemente suave (f'' absolutamente continua y f''' integrable al cuadrado), desarrollando por series de Taylor y combinando sesgo al cuadrado y variancia puede demostrarse que

$$\mathrm{MSE}\left[\hat{f}(x)\right] = \mathrm{Var}\left[\hat{f}(x)\right] + \mathrm{sesgo}^2\left[\hat{f}(x)\right] = \frac{f(x)}{nh} + \mathbf{o}(n^{-1}) + \frac{f'(x)^2}{4}\left[h - 2(x - b_j)\right]^2 + \mathbf{o}(h^3)$$

Integrando y pasando al límite cuando h y n tienen a infinito se tiene que,

$$\mathrm{AMISE} = \frac{R(K)}{nh} + \frac{h^4 \boldsymbol{\sigma}_K^2 R(f'')}{4}$$

donde $R(\boldsymbol{\phi})$ representa $\int \boldsymbol{\phi}(u)^2 du$. El ancho de ventana óptimo asintótico es,

$$h_0 = \left[\frac{R(K)}{\boldsymbol{\sigma}_K^4 R(f'')}\right]^{1/5} n^{-1/5} \qquad (2)$$

y es el que produce el AMISE mínimo: $\mathrm{AMISE}_0 = \frac{5}{4}\left[\boldsymbol{\sigma}_K R(K)\right]^{4/5} R(f'')^{1/5} n^{-4/5}$

$R(f'')$ da una idea de la suavidad de la densidad subyacente pero es desconocida por el analista. En general, las densidades más rugosas son más difíciles de estimar (tienen mayor AMISE) y requieren anchos de ventanas más angostos.

De modo que para que el error sea mínimo deben tenerse en cuenta: la elección de la función K, la función de densidad subyacente y el ancho de ventana utilizado.

Determinación de la función de núcleo K

La función kernel (K) está bajo el control del analista. La pregunta es qué función K es la "mejor" para minimizar el AMISE. Si K se restringe a ser una función de densidad (lo que significa que $\hat{f}$ también lo será), el mínimo de obtiene a través de una función escalonada de la densidad cuadrática (llamada Epanechnikov):

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2), & \text{si} \quad |u| < 1 \\ 0, & \text{en otro caso} \end{cases}$$

El valor de $\sigma_k R(K)$ para el kernel Epanechnikov es $3/(5\sqrt{5})$, entonces, el cociente $\frac{\sigma_k R(K)}{3/(5\sqrt{5})}$ provee una medida de la eficiencia relativa cometida por usar otra función kernel distinta de la óptima (este cociente es el factor multiplicativo para obtener los tamaños de muestra necesarios para igualar el AMISE en el caso Epanechnikov, al usar otras funciones kernel). La siguiente tabla muestra los valores de estos cocientes para las funciones de núcleo más usuales:

Función de Núcleo	Forma	Ineficiencia
Epanechnikov	$\frac{3}{4}(1-u^2)$	1
Biweight	$\frac{15}{16}(1-u^2)^2$	1.0061
Triweight	$\frac{35}{32}(1-u^2)^3$	1.0135
Normal (Gaussian)	$(2\pi)^{-1/2} e^{-u^2/2}$	1.0513
Uniform	$1/2$	1.0758

Es decir, el AMISE es casi insensible a la elección de la función kernel, de modo que K puede elegirse por otras razones, como por ejemplo su sencillez desde un punto de vista computacional.

Un argumento en contra de la función Epanechnikov es que no es siempre diferenciable y tampoco lo será $\hat{f}$.

Elección del ancho de ventana:

La manera más simple para elegir el ancho de ventana es elegir una función de densidad (f) de referencia y sustituirla en (2). Por ejemplo, si la función de núcleo es Gaussiana y se elige la distribución Normal como distribución de referencia, $h_0 = 1.059\,\sigma\, n^{-1/5}$. Reemplazando σ por alguna estimación, se obtiene un valor de h_0 basado en los datos.

La distribución de referencia puede ser Normal, pero el kernel puede no serlo. Sin embargo, se puede obtener un valor de h_0 basado en otra función kernel a través de un multiplicador:

Kernel	Multiplicador
Epanechnikov	2.214
Biweight	2.623
Triweight	2.978
Uniform	1.74

Así por ejemplo, el ancho de ventana asintóticamente óptimo para una función kernel "biweight" será $h_0 = 2.623\,(1.059\,\sigma\,n^{-1/5})$.

Esta metodología para la determinación del ancho de ventana depende del supuesto de que la verdadera función de densidad es normal. Lo cual es justamente desconocido.

Otro método, que no requiere realizar ningún supuesto con respecto a la densidad a priori, es el de "validación cruzada". Se basa en el ISE. El objetivo es elegir un valor de h que minimice $CV = R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(x_i)$, donde $\hat{f}_{-i}(x_i)$ es la estimación de la densidad sin considerar el caso i-ésimo de la muestra $\{x_1,...,x_n\}$. Minimizar CV es equivalente a minimizar ISE.

El método de validación cruzada también tiene sus problemas. Según la literatura sobre el tema, puede conducir a detectar más modos de los verdaderos, a menudo produce estimaciones muy rugosas (con saltos espurios). Por otro lado, el "nivel" de suavizado depende de los datos (de la muestra elegida). Finalmente la elección de "cuánto" suavizar siempre termina siendo subjetiva.

Otro método para determinar el ancho de ventana es el de Sheather-Jones donde para estimar h_0 se reemplaza $R(f'')$ por una estimación del mismo. Ellos proponen usar $R(\hat{f}'')$ en (2), es decir, $\hat{R}(f'') = R(\hat{f}'')$. La estimación $\hat{f}$ utilizada aquí está basada en un ancho diferente al apropiado para estimar la función f y es estimada a partir de los datos.

3. SIMULACIONES

La bondad del estimador basado en núcleos para estimar una función de densidad depende de la función K elegida así como del ancho de ventana utilizado. Con estos componentes se puede evaluar el error cometido en la estimación. Sin embargo, para el cálculo de los errores es necesario el conocimiento de la función de densidad subyacente, que es desconocida.

El trabajo de simulación propuesto consiste en el cálculo del error empírico sumado para todos los valores muestrales a partir de dos distribuciones prefijadas:

- Unimodal: Normal (5,1)
- Bimodal: 0.20 N(5,1)+ 0.8 N(15,3)

Como la bondad del ajuste también depende del tamaño de muestra utilizado, se generaron tres tamaños de muestra posibles para cada distribución:

- n=30
- n=50
- n=100

Las densidades estimadas se basaron en las siguientes funciones de núcleo:

- Epanechnikov
- Normal
- Biweight
- Triweight
- Uniforme

Los criterios utilizados de selección de la ventana óptima:

- Validación cruzada (CV)
- Criterio basado en una distribución normal (Normal)
- Método de Sheather- Jones (SJ)

Para cada una de las distribuciones teóricas elegidas, las simulaciones siguieron los mismos pasos. Dichos pasos se describen a continuación

3.1. Se seleccionaba una muestra de la distribución teórica con un tamaño dado.

3.2. Se escogía la primera función de núcleo (Epanechnicov) y para ella se calculaba el ancho del intervalo asintóticamente óptimo según el primer método de estimación (Validación Cruzada). Los elementos de la muestra se usaban para estimar la función de densidad teórica con dicha función de núcleo y ancho de intervalo estimado. Luego se computaba la diferencia $\left[\hat{f}(x)-f(x)\right]^2$ en 400 puntos del eje real y se retenía su suma $\sum\left[\hat{f}(x)-f(x)\right]^2$. Este valor mide empíricamente la bondad del ajuste realizado.

3.3. El paso 2 se repetía para cada una de las 5 funciones de núcleo y los 3 métodos de estimación del ancho óptimo. En total, 15 valores empíricos de bondad de ajuste eran retenidos a partir de una sola muestra.

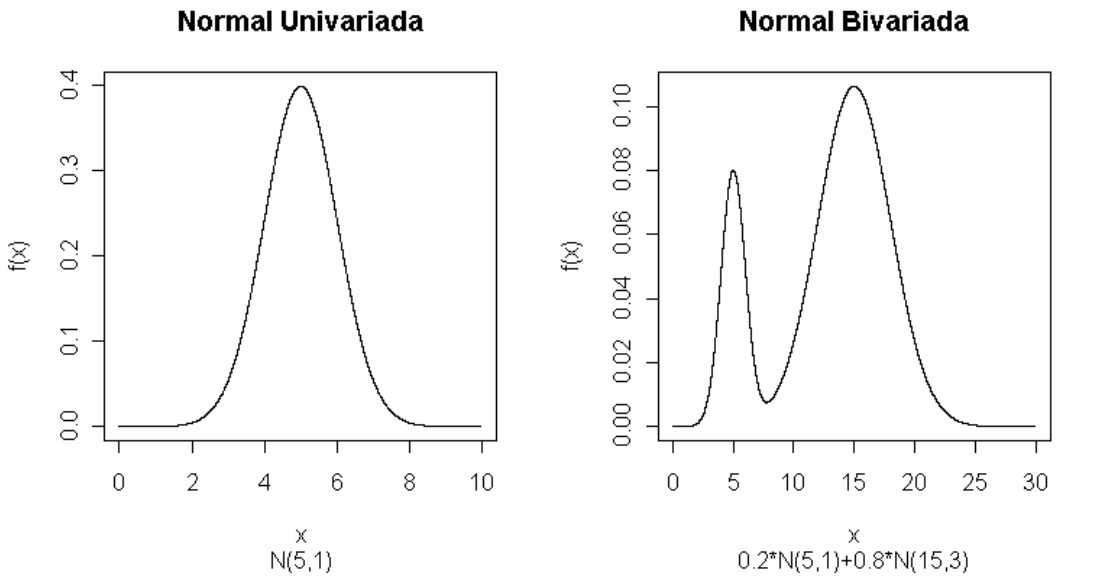
3.4. Una nueva muestra era seleccionada como en el paso 1, y los pasos 2 y 3 se llevaban a cabo sobre ella, arrojando como resultado otros 15 indicadores de bondad de ajuste. El procedimiento continuaba hasta procesar 10000 muestras. Los valores presentados en las tablas son los promedios de esos 10000 valores.

Todos los cálculos y gráficos fueron realizados con el programa R v1.9.1

4. RESULTADOS:

En el Gráfico 1 se muestran las dos distribuciones teóricas que son simuladas.

Gráfico 1. Distribuciones Teóricas



4.1. RESULTADOS DE LAS SIMULACIONES

4.1.1. DISTRIBUCIÓN TEÓRICA UNIMODAL Y SIMÉTRICA

En la Tabla 1 se pueden observar los resultados de las simulaciones para la distribución unimodal: N(5,1). Claramente se observa que utilizar la función de ponderación (K) Normal lleva a tener los menores errores cuando la distribución que se está estimando es Normal. La función triweight es la peor. El estimador del ancho de ventana óptimo que conduce a los menores errores es el obtenido suponiendo que la distribución de origen es normal. Los errores mas altos se encuentran a través del criterio de validación cruzada. Estos resultados se mantienen para todos los tamaños de muestra, aunque lógicamente los errores disminuyen a medida que el tamaño de muestra aumenta.

Por lo dicho anteriormente la mejor opción al momento de elegir un método de obtención del ancho de ventana óptimo es el que supone la distribución teórica normal y con una función de ponderación normal ya que esta opción minimiza el error de estimación de la función de densidad. Esta conclusión es muy lógica ya que efectivamente la distribución de base es normal. Por otro lado la peor opción resulta de la combinación: función de ponderación Triweight y criterio CV para determinar h. En el gráfico 2 se ejemplifica el mejor y peor ajuste para una muestra de tamaño 30.

Tabla 1. Resultado de las simulaciones para la distribución N(5,1). Promedios del error empírico en 10000 repeticiones

n=30

Función de Núcleo	Método de estimación del ancho óptimo de intervalo		
	Validación Cruzada	Normal	Sheather - Jones
Epanechnikov	4.138	2.318	3.354
Normal	1.477	0.656	1.065
Biweight	4.831	2.889	4.027
Triweight	5.432	3.402	4.633
Uniforme	3.449	1.841	2.778

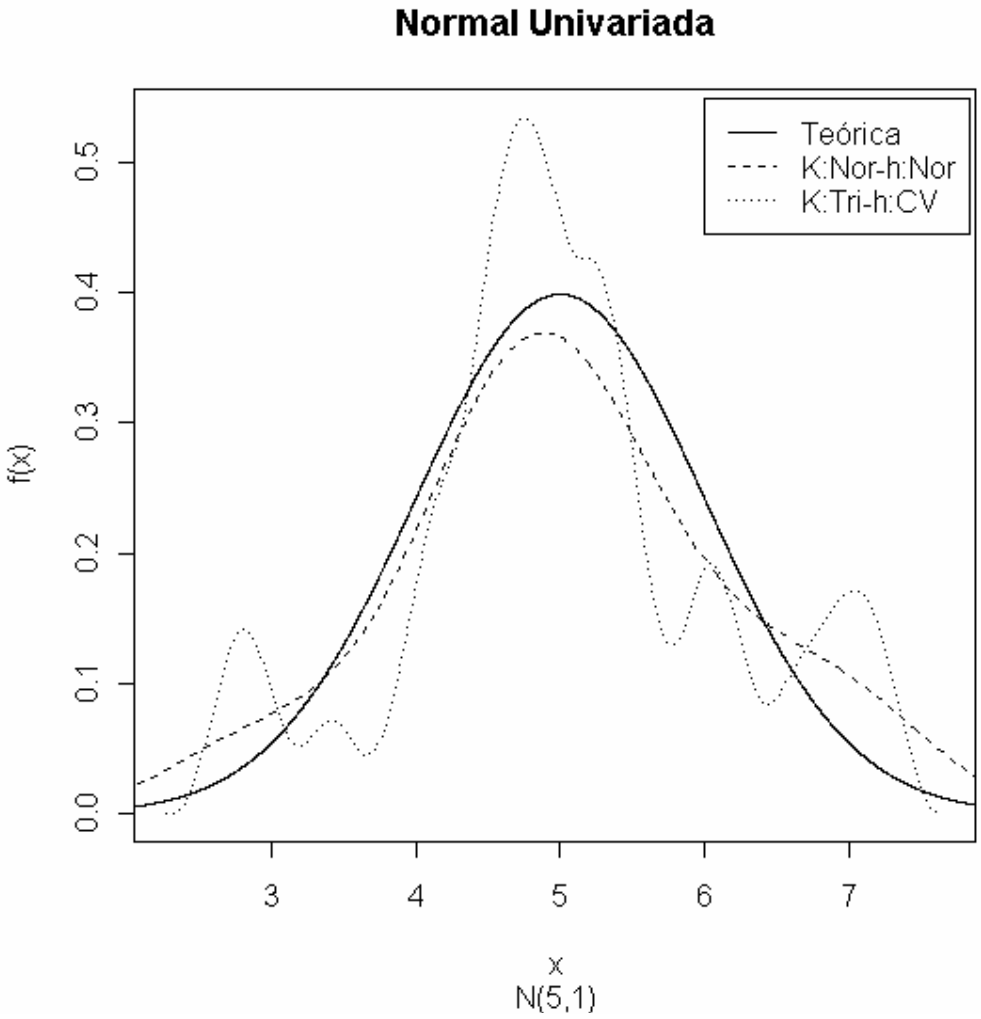
n=50

Función de Núcleo	Método de estimación del ancho óptimo de intervalo		
	Validación Cruzada	Normal	Sheather - Jones
Epanechnikov	2.509	1.512	1.978
Normal	0.927	0.460	0.660
Biweight	2.932	1.866	2.370
Triweight	3.299	2.182	2.723
Uniforme	2.096	1.210	1.640

n=100

Función de Núcleo	Método de estimación del ancho óptimo de intervalo		
	Validación Cruzada	Normal	Sheather - Jones
Epanechnikov	1.265	0.810	0.955
Normal	0.493	0.272	0.337
Biweight	1.481	0.997	1.153
Triweight	1.669	1.164	1.331
Uniforme	1.059	0.652	0.789

Gráfico 2. Distribución Teórica unimodal, Distribución Empírica bajo la mejor y peor situación de estimación



4.1.2. DISTRIBUCIÓN TEÓRICA BIMODAL

En la Tabla 2 se pueden observar los resultados de las simulaciones para la distribución teórica: 0.20*N(5,1)+0.80*N(15,3). Dado que la distribución teórica es bimodal y asimétrica, los resultados no coinciden con los vistos anteriormente. Cuando el tamaño de muestra es chico, es conveniente elegir el tamaño de la ventana a partir del criterio de la distribución normal, excepto si se usa una función de núcleo normal. En tal caso es preferible determinar la ventana óptima por el método de Sheather-Jones. De hecho esta es la combinación con menor error promedio para todos los tamaños de muestra. Siempre resulta preferible utilizar la función de ponderación normal. Mientras que la función más desfavorable es la Triweight. En cuanto a los métodos para estimar el ancho óptimo de ventana, el criterio CV es el más desfavorable. De modo que la peor opción es utilizarlo con una función de ponderación Triweight. En el gráfico 3 se ejemplifica el mejor y peor ajuste para una muestra de tamaño 30.

Tabla 2. Resultado de las simulaciones para la distribución 0.20*N(5,1)+0.80*N(15,3). Promedios del error empírico en 10000 repeticiones

n=30

Función de Núcleo	Método de estimación del ancho óptimo de intervalo		
	Validación Cruzada	Normal	Sheather - Jones
Epanechnikov	0.320	0.148	0.221
Normal	0.135	0.193	0.123
Biweight	0.374	0.167	0.267
Triweight	0.422	0.189	0.310
Uniforme	0.273	0.147	0.189

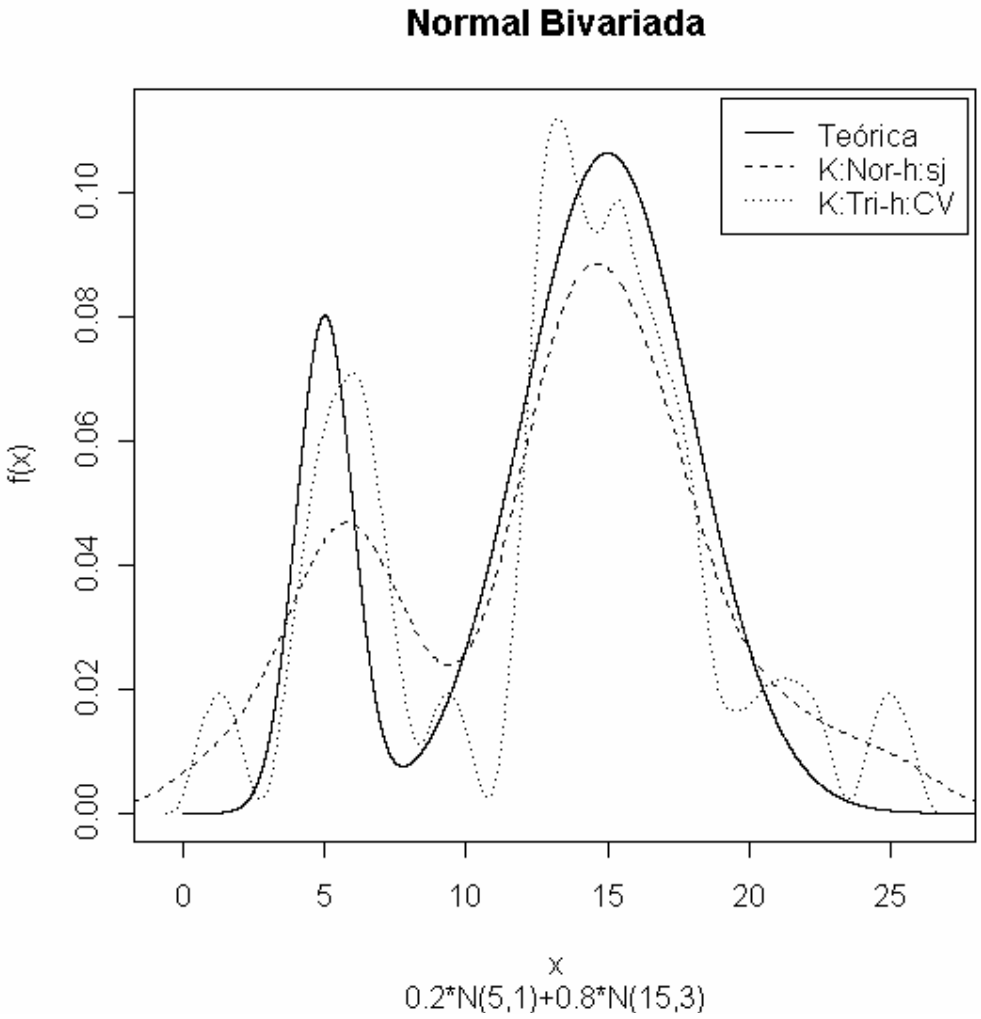
n=50

Función de Núcleo	Método de estimación del ancho óptimo de intervalo		
	Validación Cruzada	Normal	Sheather - Jones
Epanechnikov	0.202	0.116	0.144
Normal	0.086	0.181	0.086
Biweight	0.240	0.131	0.174
Triweight	0.273	0.145	0.202
Uniforme	0.169	0.116	0.123

n=100

Función de Núcleo	Método de estimación del ancho óptimo de intervalo		
	Validación Cruzada	Normal	Sheather - Jones
Epanechnikov	0.121	0.088	0.087
Normal	0.052	0.176	0.060
Biweight	0.144	0.097	0.105
Triweight	0.165	0.106	0.121
Uniforme	0.100	0.087	0.073

Gráfico 3. Distribución Teórica bimodal, Distribución Empírica bajo la mejor y peor situación de estimación



Los gráficos corroboran comentarios realizados por varios autores. Ellos son:

1. El sesgo del estimador basado en núcleos alisa los picos de las distribuciones y "rellena" los valles.
2. El estimador del ancho de ventana frecuentemente sub-suaviza la densidad provocando espúreos saltos en la estimación de la densidad.

5. DISCUSIÓN

No existe un método de suavizado que sea el mejor bajo todas las circunstancias. La simulación permite trabajar con casos "paradigmáticos", que pueden iluminar sobre la utilidad o no de determinados criterios en problemas específicos.

Los casos escogidos han permitido mostrar las ventajas y desventajas de diversos criterios en determinadas situaciones. Como paso inmediato se considerará el efecto de usar anchos de ventanas variables para mejorar los ajustes.

Luego se continuará realizando pruebas con otras distribuciones que suelen presentarse en la práctica, como por ejemplo, las que presentan rugosidades por causa de las correlaciones entre las observaciones, o la que presentan "colas" provocadas por outliers.

6. BIBLIOGRAFÍA

Browman, A.W., Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-plus illustrations*. Oxford University Press.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Simonoff, J. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Venables, W.N., Ripley, B.D. (1999). *Modern Applied Statistics with S-plus*. New York: Springer-Verlag.