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

ESTIMACIÓN DE FUNCIONES DE DENSIDAD DE DATOS CORRELACIONADOS Y/O CON VALORES ATÍPICOS

1. INTRODUCCIÓN

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. Por ejemplo no considera si la distribución corresponde a datos correlacionados o si en ella se presentan valores extremos o atípicos. Surge, entonces, la necesidad de construir otro estimador de la densidad que intente diluir dichas desventajas.

Un método de estimación de funciones de densidad de probabilidad más adecuado para estos escenarios, es el basado 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 funciones de núcleos (K) en las que se basan los estimadores. Las propiedades de las curvas de estimación dependen de la elección de la función de núcleo y del 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.

En este trabajo se presenta la estimación de funciones de densidad mediante núcleos teniendo en cuenta dos situaciones que suelen presentarse en la práctica y que resultan difíciles de modelar mediante las técnicas tradicionales, en particular con los métodos paramétricos. La primera situación se presenta cuando la variable en estudio está medida sobre unidades que se agrupan en conglomerados, y sin embargo esta estructura no se tiene en cuenta al momento de estimar la función de densidad (un ejemplo de esto lo constituyen los datos provenientes de muestras complejas que suelen analizarse como si se tratase de una muestra simple al azar). La segunda, cuando se presentan valores extremos o atípicos en los datos, que provocan "colas" en las distribuciones.

2. METODOLOGÍA

Considerando la definición de 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,

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

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

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

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

Entre las medidas para evaluar a $\hat{f}(x)$ como estimador de $f(x)$ se encuentran

1. Error cuadrático integrado: $ISE = \int_{-\infty}^{+\infty} \left[\hat{f}(u) - f(u)\right]^2 du$. El ISE es una variable aleatoria construida sobre el espacio de todas las muestras posibles.
2. Error cuadrático integrado esperado (o promedio) (MISE) es el valor esperado del ISE sobre todas las muestras posibles que el método de muestreo es capaz de generar. El MISE depende del tamaño muestral. Para $n \to \infty$, el MISE tiene por límite al AMISE (error cuadrático integrado esperado asintótico)

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.

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.

La elección de la función kernel (K) está bajo el control del analista, y debe ser aquella que logre minimizar el error. Sin embargo su elección no modifica sustancialmente el error, por tanto K puede elegirse por otras razones, como por ejemplo su sencillez desde un punto de vista computacional. Una función que computacionalmente es sencilla y se adapta a la mayoría de las situaciones en la práctica es la Normal

Por otro lado, la selección del ancho de ventana es independiente de la función Kernel utilizada. Sin embargo , 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. Esta metodología para la determinación del ancho de ventana depende del supuesto de que la verdadera función de densidad es normal.

3. SIMULACIONES

El trabajo de simulación propuesto consiste en el cálculo del error cuadrático integrado (ISE) y del ancho de ventana óptimo en dos situaciones generales.

1. Datos correlacionados

Se generan muestras de conglomerados a partir del modelo $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, $i = 1,...,a;\ j = 1,...,b$, siendo $\alpha_i \sim N(0,\sigma_\alpha^2)$ $\varepsilon_{ij} \sim N(0,\sigma_\varepsilon^2)$, con α_i y ε_{ij} independientes. De esta forma se obtienen datos correlacionados dentro de cada uno de los *a* conglomerados, pero independientes entre conglomerados. Es decir,

$$Corr(y_{ij}, y_{ij'}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2} \qquad i = 1,...,a;\ j \wedge j' = 1,...,b;\ j \neq j'$$

$$Corr(y_{ij}, y_{i'j}) = 0 \qquad i \wedge i' = 1,...,a;\ i \neq i';\ j = 1,...,b$$

El procedimiento de generación es el siguiente:

i) Generar un valor de α_1 proveniente de $N(0,\sigma_\alpha^2)$. Generar *b* valores de la variable y_{ij} , *j=1,…,b,* proveniente de $N(\mu + \alpha_1, \sigma_\varepsilon^2)$

ii) Repetir el paso i) *a* veces, cambiando α_1 por α_i con *i=2,…,a*

iii) Se calcula el ISE y el ancho de ventana óptimo

Se realizan 1000 repeticiones y se calculan el promedio de los anchos de ventana óptimos, y de los ISE. Éste último promedio estima el MISE. Los escenarios planteados corresponden a las 18 combinaciones que surgen de fijar $\sigma_\varepsilon^2 = 1$ y $\mu = 3$, y hacer variables los parámetros:

- Número de conglomerados (a): 3, 5, 10
- Tamaño del conglomerado (b): 5, 10, 30
- σ_α^2: 0.5, 10

2. Datos con outliers

Se realizan 1000 repeticiones. Cada una de estas repeticiones genera datos de una muestra donde el 90% de los mismos corresponden a una distribución Normal con promedio 7 y desvío estándar 2, y el 10% restante corresponde a una Normal con promedio 15 y desvío estándar 1. Se generan muestras de tamaño 30, 50, y 100. En cada caso se calcula el promedio de los anchos de ventana óptimos, y de los ISE.

Todos los cálculos y gráficos se realizaron con el programa R v2.0.1

4. RESULTADOS:

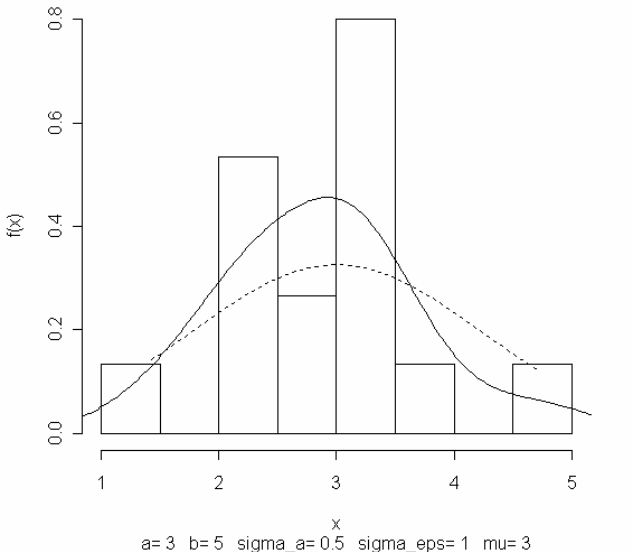
4.1. DATOS CORRELACIONADOS

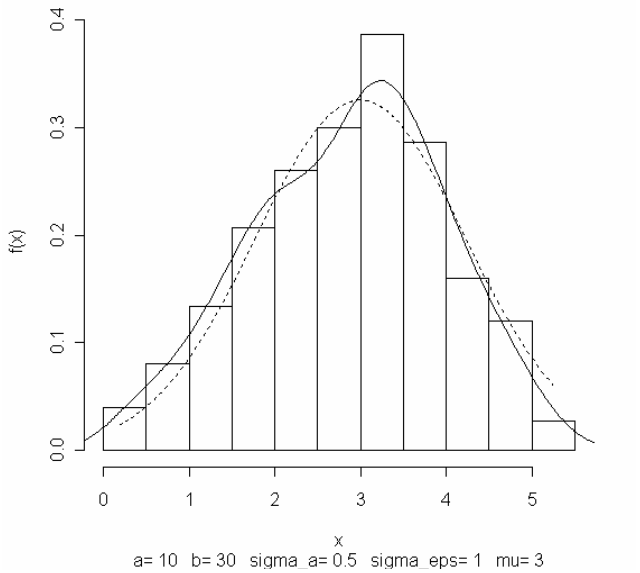
		a (N° de Conglomerados)								
		3			5			10		
σ_α^2	b (tamaño de los cong.)	5	10	30	5	10	30	5	10	30
0.5	h	0.659	0.577	0.465	0.603	0.528	0.423	0.532	0.467	0.375
	MISE	0.005	0.004	0.002	0.004	0.003	0.002	0.003	0.002	0.001
10	h	4.747	4.012	3.196	4.782	4.122	3.35	4.477	3.994	3.141
	MISE	0.018	0.031	0.058	0.019	0.03	0.05	0.016	0.021	0.033

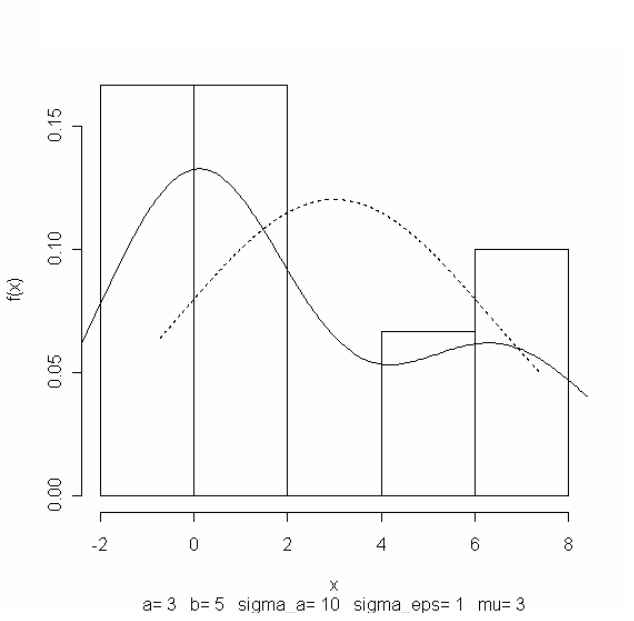
- El ancho de ventana óptimo (*h*) es menor a medida que el tamaño de muestra total (*a*b*) aumenta, lo cual permite una mejor estimación de la curva de densidad en aquellas funciones dificiles de modelar con modelos paramétricos. Esto también se ve reflejado en los errores.
- A mayor variancia entre conglomerados, la correlación entre datos de un mismo conglomerado aumenta. En esta situación los errores que se producen en la estimación de la función de densidad son mayores, lo cual indica que el modelo subyacente para la variable en estudio, $N(\mu, \sigma_\alpha^2 + \sigma_\varepsilon^2)$, no es apropiado. De aquí la importancia de utilizar modelos no paramétricos
- Considerando los casos con tamaños de muestra totales igual a 50 (a=5 y b=10, ó a=10 y b=5), no existen diferencias entre los resultados si la variabilidad entre los conglomerados es pequeña, sin embargo si la variabilidad entre conglomerados es grande, los errores en las estimaciones de la función de densidad son considerablemente mayores cuando hay menos conglomerados (a=5 y b=10)

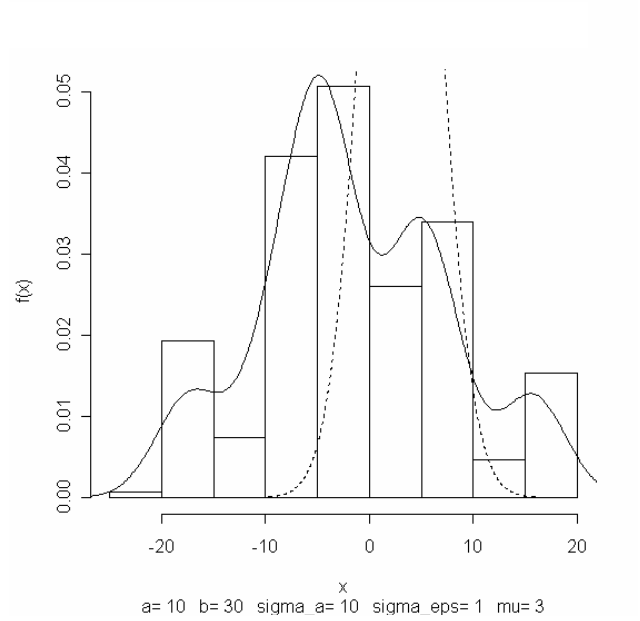
Los gráficos que se muestran a continuación son casos particulares de generación de datos correlacionados correspondiente a una de las 1000 repeticiones para algunos de los escenarios considerados.

DATOS PROVENIENTES DE UNIDADES QUE SE AGRUPAN EN CONGLOMERADOS









------ Función de Densidad $N(\mu, \sigma_\alpha^2 + \sigma_\varepsilon^2)$ ignorando la correlación de los datos

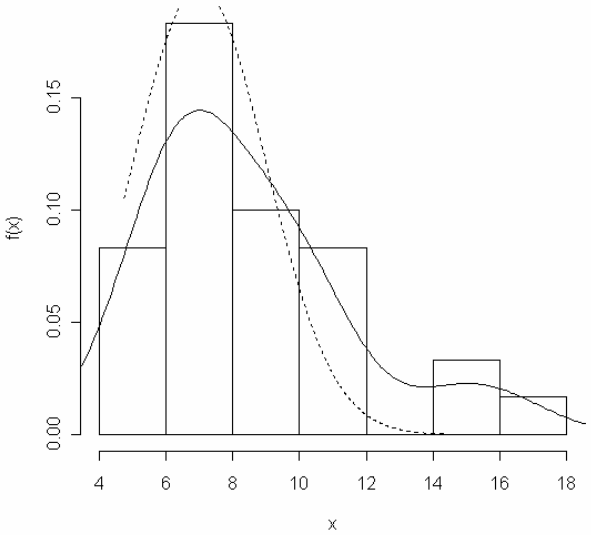
—— Estimación No Paramétrica de la función de densidad

4.2. DATOS CON OUTLIERS

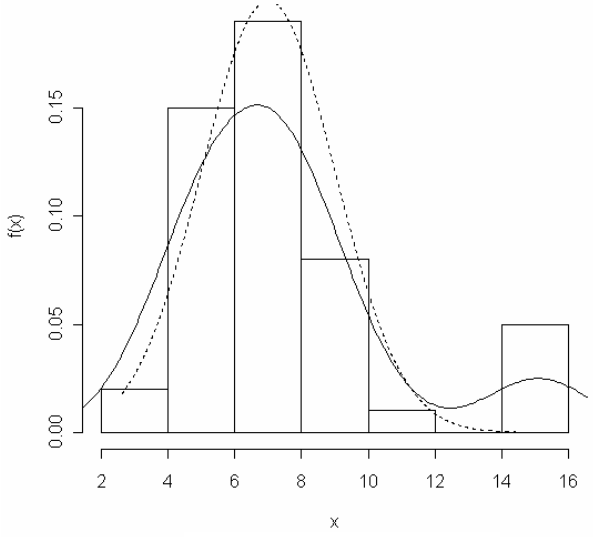
Tamaño de muestra	Promedio de	
	h	MISE
30	1.664	0.020
50	1.494	0.022
100	1.300	0.025

- Tal como en los datos correlacionados, el ancho de ventana disminuye a medida que el tamaño de muestra aumenta.
- El MISE aumenta a medida que el tamaño de muestra aumenta debido a que la cantidad de outliers es mayor y por lo tanto la diferencia entre la densidad teórica y la ajustada no paramétricamente se hace mayor.

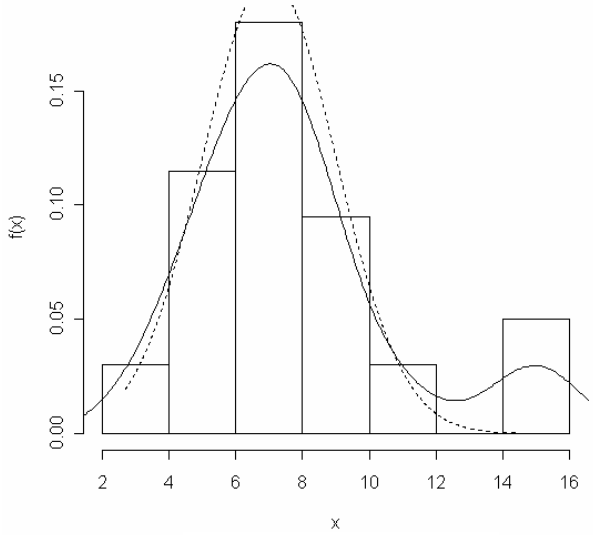
DATOS CON VALORES EXTREMOS Ó ATÍPICOS



n=30



n=50



n=100

------ Función de Densidad N(7,2), ignorando valores extremos
—— Estimación No Paramétrica de la función de densidad

5. DISCUSIÓN

Los estimadores de densidad simple como histogramas o polígonos de frecuencia son sencillos y muy informativos, pero tienen dos contras importantes: no son suaves y no son sensibles a las propiedades locales de f(x), más aún, considera que los datos son independientes y con igual distribución.

Cuando se cuenta con datos provenientes de unidades que se agrupan en conglomerados, o con datos con valores extremos o atípicos, los histogramas no son suaves y los métodos clásicos de estimación de función de densidad no son aplicables debido a que no logran captar la estructura interna de los datos.

En la práctica se presentan muchas situaciones donde los datos provienen de unidades agrupadas en conglomerados. Por ejemplo, si se quiere analizar la distribución del ingreso, a partir de datos provenientes de la Encuesta Permanente de Hogares, y no se tiene en cuenta el diseño muestral utilizado, difícilmente se puede estimar la función de densidad del ingreso con un modelo de función de densidad conocido, por ejemplo, Distribución Gamma, ya que el mismo no captará la estructura de correlación de los datos, especialmente si la variabilidad entre conglomerados es muy grande.

Una alternativa es utilizar el estimador de densidad no paramétrico de kernel que es "más local" en su naturaleza y que provee una estimación más suave, capturando localmente las formas rugosas que generan los datos en las situaciones antes mencionadas.

En este trabajo se muestra que cuando los datos están correlacionados, y especialmente cuando la variancia entre conglomerados es grande, la distribución presenta muchos picos, y los mismos son captados con precisión por el estimador de kernel. En esta situación sería difícil encontrar un modelo paramétrico que describa adecuadamente los datos. Igual situación se presenta con datos con valores extremos, cuando el tamaño de muestra es grande.

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.