

CONSTRUCCIÓN DE KERNELS Y FUNCIONES DE DENSIDAD DE PROBABILIDAD

Luis Rodríguez Ojeda¹

Resumen. En este artículo se describen los detalles del conocido método kernel para construir una función de densidad de probabilidad para una muestra univariada. Se proponen nuevas formas para los kernels y criterios adicionales para su elección y para estimar el ancho de banda óptimo. Como soporte para esta investigación se instrumentó un software para experimentación y obtención de resultados gráficos y numéricos.

Palabras clave: Kernel. Densidad de Probabilidad. Ancho de banda. Estimación no Paramétrica

Abstract. This paper describes the details of the well known kernel method used to construct a probability density function for a univariate sample. New forms for the kernels are proposed and additional criteria for their election and for estimating the optimal bandwidth. As support for this research a software was implemented for experimentation and obtaining of graphical and numerical results.

Keywords: Kernel. Probability Density. Bandwidth. Nonparametric Estimation

Recibido:

Aprobado

¹Luis Rodríguez Ojeda, M.Sc., Departamento de Matemáticas, ESPOL. (e-mail: lrodrig@espol.edu.ec)

1. INTRODUCCIÓN

La distribución de probabilidad de una variable aleatoria continua X se describe mediante una función denominada función de densidad $f(x)$ con la cual se pueden determinar valores de probabilidad con la definición:

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

Si $f(x)$ no es un modelo conocido, es de interés para la investigación poder estimar $f(x)$ a partir de una muestra de observaciones x_1, x_2, \dots, x_n que suponemos son resultados independientes y tienen la misma distribución de probabilidad. Como ocurre frecuentemente en problemas de ingeniería, la obtención de estos datos se basa en ensayos que involucran tiempo y costo, por lo tanto su cantidad es limitada.

El método kernel utiliza un conjunto de datos que provienen de una distribución continua, univariada y desconocida para aproximar esta función. Los fundamentos matemáticos son conocidos pero la investigación aún continúa en la selección de los parámetros de ajuste adecuados para su aplicación. En este documento se describe en detalle la formulación y se proponen algunos criterios para la aplicación de este método.

Los kernels son funciones que se asocian a cada uno de los datos. Entonces, la suma ponderada de estas funciones es un estimador para aproximar la función de densidad desconocida.

Estas funciones son objetos matemáticos conocidos, pero en esta contribución se desarrollan nuevas formas basadas en consideraciones geométricas y se establecen otros criterios para compararlos.

Como ocurre en muchas áreas del conocimiento, la formulación desarrollada no es adecuada para el tratamiento manual, por lo que se ha construido un programa computacional para su aplicación. Existen programas para usar el método Kernel, pero no incluyen los modelos para experimentar como se describe en este trabajo.

El programa usa como soporte el lenguaje MATLAB y constituye un pequeño laboratorio con el que se pueden probar diferentes modelos para construir el estimador de la función $f(x)$ realizando pruebas con los parámetros de ajuste. Los resultados que se muestran son gráficos, simbólicos y numéricos.

2. PROPIEDADES DE KERNELS

2.1 Definición de kernel

Un kernel es una función de variable real: $K: \mathbf{R} \rightarrow \mathbf{R}$ con las siguientes propiedades

- $K(x) \in [0, \infty)$, $x \in [-1, 1]$
- $K(x) = 0$, $x \notin [-1, 1]$
- $K(x) = K(-x)$
- $\int_{-1}^1 K(x) dx = 1$
- $\int_{-1}^1 xK(x) dx = 0$
- $\int_{-1}^1 x^2 K(x) dx \in \mathbf{R}^+$

De aquí en adelante, nos referiremos solamente al intervalo en el cual K no es negativo. En la definición este intervalo es $[-1, 1]$ pero puede modificarse mediante un parámetro. Además, es deseable que K sea diferenciable

2.2 Parametrización de kernels

Sea $h \in \mathbf{R}^+$, el kernel parametrizado en h es

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right), \quad x \in [-h, h]$$

Esta modificación mantiene las propiedades anteriores, pero referidas al intervalo $[-h, h]$.

h se denomina el **ancho de banda** de K y es la semi-amplitud del kernel en el intervalo de interés.

La propiedad **d)** se prueba mediante la sustitución: $u = x/h$: $x=-h \Rightarrow u = -1$, $x=h \Rightarrow u=1$
 $dx=hu$, $dx=hd u$

$$\int_{-h}^h \frac{1}{h} K\left(\frac{x}{h}\right) dx = \int_{-1}^1 \frac{1}{h} K(u) h du \\ = \int_{-1}^1 K(u) du = 1$$

Es importante interpretar el rol de h en $K_h(x)$

Si h se incrementa, $h > 1$, la amplitud de K aumenta, pero el factor $1/h$, reduce el rango de K para mantener el área igual a 1

Si h se reduce, $h < 1$, la amplitud de K se reduce, pero el factor $1/h$ incrementa el rango de K para mantener el área igual a 1

2.3 Traslación de kernels

El kernel se puede centrar en cualquier punto $x_i \in \mathbf{R}$. El kernel parametrizado en h y centrado en x_i es:

$$K_h(x) = \frac{1}{h} K\left(\frac{x-x_i}{h}\right), \quad x \in [x_i - h, x_i + h]$$

Esta modificación mantiene las propiedades anteriores en el nuevo intervalo de interés:

$[x_i - h, x_i + h]$

Para verificar la propiedad **d)** se realiza la sustitución: $u = \frac{x-x_i}{h}$: $x = x_i - h \Rightarrow u = -1$,

$x = x_i + h \Rightarrow u = 1$, $dx=hd u$

$$\int_{x_i-h}^{x_i+h} \frac{1}{h} K\left(\frac{x-x_i}{h}\right) dx = \int_{-1}^1 \frac{1}{h} K(u) h du \\ = \int_{-1}^1 K(u) du = 1$$

Se concluye que la función $K_h(x) = \frac{1}{h} K\left(\frac{x-x_i}{h}\right)$, $x \in [x_i - h, x_i + h]$ es una función de densidad de probabilidad centrada en el punto x_i .

3. MODELOS DE KERNELS

Se describen a continuación algunos kernels que son bien conocidos [1].

3.1 Kernel Rectangular o Uniforme

Es un rectángulo que se coloca sobre cada punto. Al interactuar con los kernels de los otros puntos, el efecto en la suma es un cambio abrupto.

$$K(x) = 0.5, \quad x \in [-1, 1]$$

3.2 Kernel Triangular

Es un triángulo que se coloca sobre cada punto. Al interactuar con los otros kernels el efecto combinado es lineal pero más liso que los rectángulos

$$K(x) = 1 - |x|, \quad x \in [-1, 1]$$

3.3 Kernel de Epanechnikov

Es el kernel más estudiado. Es un segmento del perfil de un arco de parábola que se coloca sobre cada punto.

$$K(x) = \frac{3}{4}(1-x^2), \quad x \in [-1, 1]$$

3.4 Kernel Normal o Gaussiano

Este kernel es un caso especial. Para este kernel se define como intervalo el conjunto \mathbf{R} , por lo que cada kernel influye en todos los otros kernels colocados en los puntos de la muestra. La suma resultante es continua y suave.

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in (-\infty, +\infty)$$

3.5 Kernel Biweight o Cuártico

$$K(x) = \frac{15}{16}(1-x^2)^2, \quad x \in [-1, 1]$$

3.6 Kernel Arco Coseno

$$K(x) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}x\right), \quad x \in [-1, 1]$$

4. CONSTRUCCIÓN DE KERNELS

El diseño de nuevas formas para los kernels es un ejercicio matemático. El diseño se basa en consideraciones geométricas. Estos objetos matemáticos son los componentes con los que se construye el estimador de la función de densidad.

Las formas propuestas difieren en algún aspecto las formas conocidas que normalmente son funciones de potencia par. Las funciones propuestas usan polinomios cúbicos entre otros.

4.1 Kernel Cúbico Natural

El perfil son dos segmentos de un polinomio cúbico que se conectan manteniendo continuidad hasta la primera derivada y con los extremos libres, sin curvatura. Se ilustra su obtención. Similarmente se obtienen los otros.

Polinomio cúbico en el tramo derecho:

$$K(x) = ax^3 + bx^2 + cx + d, \quad x \in [0, 1] \\ K'(x) = 3ax^2 + 2bx + c, \quad K''(x) = 6ax + 2b$$

Condiciones geométricas:

$$a) K(1) = 0 \Rightarrow a + b + c + d = 0$$

$$b) K'(0) = 0 \Rightarrow c = 0$$

$$c) K''(1) = 0 \Rightarrow 3a + b = 0$$

La altura t del kernel es el parámetro para convertirlo en función de densidad:

$$d) K(0) = t \Rightarrow d = t$$

Resolviendo y sustituyendo se obtiene

$$K(x) = \frac{t}{2}x^3 - \frac{3}{2}tx^2 + t$$

Para que K sea función de densidad:

$$\int_0^1 K(x) dx = \frac{1}{2} \Rightarrow t = \frac{4}{5}$$

Entonces

$$K(x) = \frac{2}{5}(x^3 - 3x^2 + 2), \quad x \in [0, 1]$$

$$K(-x) = \frac{2}{5}(-x^3 - 3x^2 + 2), \quad x \in [-1, 0]$$

Se pueden escribir con una regla:

$$K(x) = \frac{2}{5}(|x|^3 - 3x^2 + 2), \quad x \in [-1, 1]$$

4.2 Kernel Cúbico Plano

El perfil son dos segmentos de un polinomio cúbico que se conectan manteniendo continuidad hasta la segunda derivada

$$K(x) = \frac{2}{3}(1 - |x|^3), \quad x \in [-1, 1]$$

4.3 Kernel Cúbico Sujeto

El perfil son dos segmentos de un polinomio cúbico que se conectan manteniendo continuidad hasta la primera derivada, pero con los extremos horizontales (primera derivada nula), con el objetivo de que al combinarse con otros kernels, la interacción sea lisa.

$$K(x) = 2|x|^3 - 3x^2 + 1, \quad x \in [-1, 1]$$

4.4 Kernel Coseno

El perfil es un segmento modulado de la función coseno.

$$K(x) = \frac{1}{2}(\cos(\pi x) + 1), \quad x \in [-1, 1]$$

4.5 Kernel Arco Circular

El perfil es un arco de una circunferencia:

$$K(x) = \sqrt{1.145834 - x^2} - 0.381882, \quad x \in [-1, 1]$$

En la **Figura 18** al final de este artículo se muestra el perfil de algunos kernels formulados.

5. CONSTRUCCIÓN DE FUNCIONES DE DENSIDAD DE PROBABILIDAD

Sea \mathbf{X} una variable aleatoria con distribución de probabilidad continua, univariada y desconocida $f(x)$ de la cual se dispone de una muestra aleatoria de n observaciones independientes: x_1, x_2, \dots, x_n . El objetivo es usar estos datos para obtener un estimador $\hat{f}(x)$ de la función de densidad de probabilidad $f(x)$

El método clásico para construir $\hat{f}(x)$ es el histograma que agrupa los datos en clases con amplitud que debe elegirse. La representación gráfica son rectángulos excluyentes cuya altura es el conteo de observaciones en cada clase y se denomina frecuencia de clase.

El histograma es una función continua pero cambia con saltos entre clases. La altura de cada rectángulo puede asociarse a valores de probabilidad. Esta altura solo depende de la cantidad de datos incluidos en cada clase ignorando la influencia de los datos adyacentes aunque estén muy próximos.

5.1 El Método Kernel

Un kernel es una función de densidad. Si se coloca un kernel en cada uno de los datos de la muestra, la suma ponderada de estas funciones también será una función de densidad de probabilidad. Esta suma es una función continua que suaviza el perfil de la distribución captando la influencia de los datos cercanos y constituye el estimador $\hat{f}(x)$ del modelo teórico del cual provienen los datos, permitiendo observar diferencias que los rectángulos del histograma no puede mostrar [2].

Sea $K_h(x) = \frac{1}{h} K\left(\frac{x-x_i}{h}\right)$, $x \in [x_i - h, x_i + h]$

kernel parametrizado y centrado en cada punto $x_i, i=1, 2, 3, \dots, n$

K_h es una función de densidad de probabilidad. Si cada kernel se multiplica por $1/n$, entonces la suma de los n kernels también será una función de densidad de probabilidad.

Definición: Estimador por kernels $\hat{f}(x)$:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_h(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x-x_i}{h}\right)$$

$x \in [x_i - h, x_i + h]$ en cada kernel i

Intervalo de $\hat{f}(x)$: $[x_1 - h, x_n + h]$

Se supondrá que $x_1 \leq x_2 \leq x_3 \leq \dots \leq x_n$

El ancho de banda h es el parámetro de ajuste o suavizado de $\hat{f}(x)$ su elección es crítica para el modelo.

Mientras más pequeño es h , más concentrada está la contribución del kernel en cada punto x_i . Mientras más grande es h , mayor es la influencia e interacción del kernel hacia los puntos vecinos.

En el límite, cuando $h \rightarrow 0$, la contribución de cada kernel estará concentrada en cada punto x_i así el estimador $\hat{f}(x)$ tendrá una distribución puntual concentrada en cada dato. Por otra parte, cuando $h \rightarrow \infty$, la distribución de $\hat{f}(x)$ se aplanará, con un solo cúmulo y con mayor dispersión. Es necesario buscar un ancho de banda adecuado para construir el estimador.

Ejemplo. *Dados los siguientes datos de una muestra aleatoria (variables independientes y con la misma distribución de probabilidad), analizar un modelo de densidad de probabilidad:*

$X: 1.1, 2.1, 2.3, 2.7, 3.8$

Todos los resultados gráficos y numéricos que se muestran a continuación fueron obtenidos con el software **KDEN** desarrollado para esta investigación. El programa puede mostrar también el modelo matemático del estimador $\hat{f}(x)$ y algunas medidas estadísticas de interés.

Figura 1
Diagrama de puntos

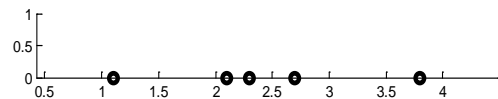
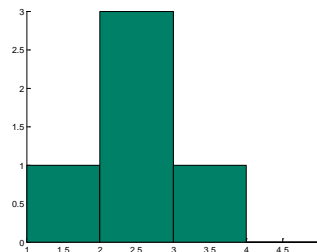


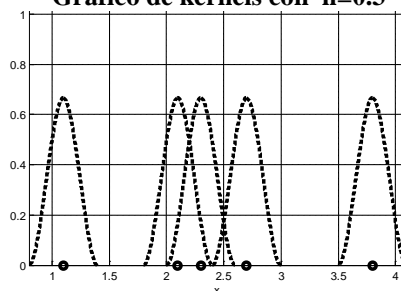
Figura 2
Histograma, amplitud de clase = 1



La representación del histograma muestra solamente parte de la información de los datos.

Kernel elegido: Cúbico Sujeto

Figura 3
Gráfico de kernels con $h=0.3$



Si h es muy pequeño, los kernels están concentrados en cada punto y no interactúan con los otros. La suma es la función de densidad $\hat{f}(x)$ y su perfil se muestra superpuesto al gráfico de los kernels en la siguiente figura.

Figura 4
Kernels y la función de densidad $\hat{f}(x)$ con $h=0.3$

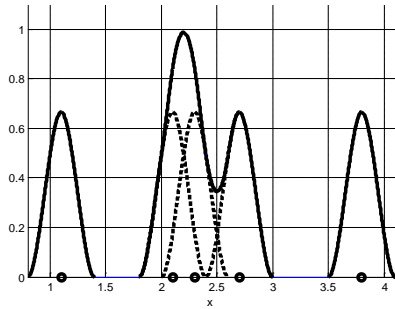
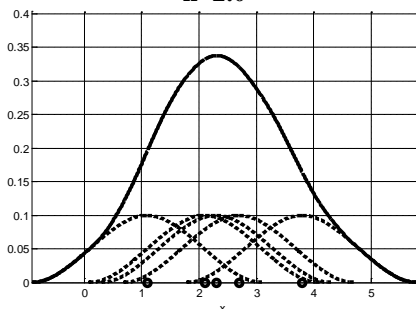
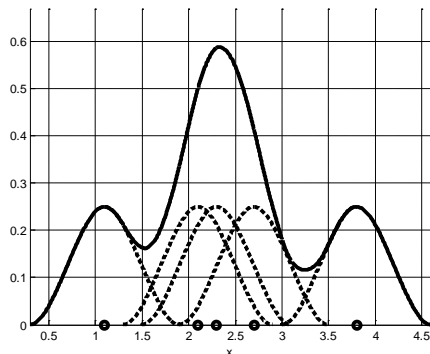


Figura 5
Kernels y la función de densidad $\hat{f}(x)$ con $h=2.0$



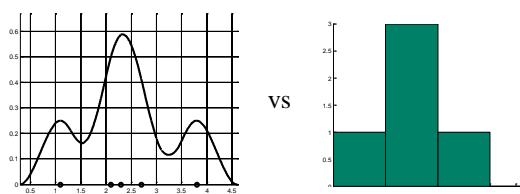
Si h es muy grande hay un sobre ajuste y la distribución sumada será más plana, con un solo cúmulo y con mayor dispersión

Figura 6
Kernels y la función de densidad $\hat{f}(x)$ con $h=0.8$



Si el valor de h es adecuado, la distribución se suaviza y permite observar más detalles de la distribución de probabilidad.

Figura 7
Función de densidad $\hat{f}(x)$ con $h=0.8$ y el histograma



Cálculo de probabilidad

Calcular la probabilidad que la variable X tome un valor entre 1.5 y 2.5

$$P(1.5 \leq X \leq 2.5) = \int_{1.5}^{2.5} \hat{f}(x) dx = 0.3973$$

Medidas estadísticas de $\hat{f}(x)$

Para el ejemplo anterior:

Media	2.4000
Varianza	0.8532
Sesgo	0.1092
Rango	[0.3, 4.6]
Amplitud	4.3000
Mediana	2.3578
Primer Cuartil	1.8302
Tercer Cuartil	2.9226

En resumen, el método kernel proporciona un estimador continuo $\hat{f}(x)$. Si se elige adecuadamente el ancho de banda h los rectángulos del histograma son reemplazados por cúmulos suavizados que se solapan e interactúan de tal manera que al sumarlos producen una función que presenta detalles que el histograma no puede mostrar.

5.2 Propiedades de la variable aleatoria con densidad $\hat{f}(x)$

Sea X : variable aleatoria con densidad $\hat{f}(x)$

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_h(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x-x_i}{h}\right),$$

$K_h(x)$, $x \in [x_i - h, x_i + h]$, Kernel
 h : Ancho de banda
 $\hat{f}(x)$, $x \in [x_1 - h, x_n + h]$
 $x_1 \leq x_2 \leq x_3 \leq \dots \leq x_n$

5.2.1 $\hat{f}(x)$ es una función de densidad

Demostración

$$\int_{x_1-h}^{x_n+h} \hat{f}(x) dx = \int_{x_1-h}^{x_n+h} \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x-x_i}{h}\right) dx$$

$$= \frac{1}{nh} \sum_{i=1}^n \int_{x_i-h}^{x_i+h} K\left(\frac{x-x_i}{h}\right) dx$$

Con la sustitución: $u = \frac{x-x_i}{h}$: $x = x_i - h \Rightarrow u = -1$

$$x = x_i + h \Rightarrow u = 1, \quad dx = h du$$

$$\frac{1}{nh} \sum_{i=1}^n \int_{x_i-h}^{x_i+h} K\left(\frac{x-x_i}{h}\right) dx = \frac{1}{n} \sum_{i=1}^n \int_{-1}^1 K(u) du$$

$$= \frac{1}{n} (n) = 1$$

5.2.2 Valor esperado de la variable aleatoria

Sea $x \in [x_1-h, x_n+h]$ una variable aleatoria con distribución $\hat{f}(x)$. Su valor esperado:

$E(x) =$

$$\int_{x_1-h}^{x_n+h} x \hat{f}(x) dx = \frac{1}{nh} \sum_{i=1}^n \int_{x_i-h}^{x_i+h} x K\left(\frac{x-x_i}{h}\right) dx$$

Mediante la sustitución

$$u = \frac{x-x_i}{h}: x = x_i - h \Rightarrow u = -1, x = x_i + h \Rightarrow u = 1,$$

$$x = hu + x_i, \quad dx = h du$$

$$\begin{aligned}
E(x) &= \frac{1}{n} \sum_{i=1}^n \int_{-1}^1 (\mathbf{h}u + \mathbf{x}_i) \mathbf{K}(u) du = \\
&= \frac{h}{n} \sum_{i=1}^n \int_{-1}^1 u \mathbf{K}(u) du + \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \int_{-1}^1 \mathbf{K}(u) du \\
&= \frac{h}{n} (\mathbf{0}) + \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i (\mathbf{1}) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \bar{\mathbf{X}} \\
E(\mathbf{x}) &= \bar{\mathbf{X}}
\end{aligned}$$

El valor esperado de la variable aleatoria \mathbf{x} con densidad $\hat{\mathbf{f}}(\mathbf{x})$ coincide con la media muestral, independientemente del kernel \mathbf{K} .

5.2.3 Varianza de la variable aleatoria

Sea $\mathbf{x} \in [\mathbf{x}_1 - \mathbf{h}, \mathbf{x}_n + \mathbf{h}]$ una variable aleatoria con densidad $\hat{\mathbf{f}}(\mathbf{x})$ [7].

Su varianza:

$$\begin{aligned}
\sigma_x^2 &= E(\mathbf{x}^2) - E^2(\mathbf{x}) \\
E(\mathbf{x}^2) &= \\
\int_{\mathbf{x}_1 - \mathbf{h}}^{\mathbf{x}_n + \mathbf{h}} \mathbf{x}^2 \hat{\mathbf{f}}(\mathbf{x}) d\mathbf{x} &= \frac{1}{nh} \sum_{i=1}^n \int_{\mathbf{x}_i - \mathbf{h}}^{\mathbf{x}_i + \mathbf{h}} \mathbf{x}^2 \mathbf{K}\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) d\mathbf{x}
\end{aligned}$$

Mediante la sustitución:

$$\begin{aligned}
\mathbf{u} &= \frac{\mathbf{x} - \mathbf{x}_i}{h} : \mathbf{x} = \mathbf{x}_i - \mathbf{h} \Rightarrow \mathbf{u} = -1, \\
\mathbf{x} &= \mathbf{x}_i + \mathbf{h} \Rightarrow \mathbf{u} = 1, \quad \mathbf{x} = \mathbf{h}\mathbf{u} + \mathbf{x}_i, \quad d\mathbf{x} = \mathbf{h}d\mathbf{u}
\end{aligned}$$

$$\begin{aligned}
E(\mathbf{x}^2) &= \frac{1}{n} \sum_{i=1}^n \int_{-1}^1 (\mathbf{h}u + \mathbf{x}_i)^2 \mathbf{K}(u) du \\
&= \frac{1}{n} \sum_{i=1}^n \int_{-1}^1 \mathbf{u}^2 \mathbf{h}^2 \mathbf{K}(u) du + \\
&= \frac{1}{n} \sum_{i=1}^n \int_{-1}^1 2\mathbf{u}\mathbf{h}\mathbf{x}_i \mathbf{K}(u) du + \\
&= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^2 \int_{-1}^1 \mathbf{K}(u) du \\
&= \frac{\mathbf{h}^2}{n} \sum_{i=1}^n \int_{-1}^1 \mathbf{u}^2 \mathbf{K}(u) du + \\
&= \frac{2\mathbf{h}}{n} \sum_{i=1}^n \mathbf{x}_i \int_{-1}^1 \mathbf{u} \mathbf{K}(u) du + \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^2 \int_{-1}^1 \mathbf{K}(u) du \\
&= \frac{\mathbf{h}^2}{n} \sum_{i=1}^n \sigma_K^2 + \frac{2\mathbf{h}}{n} \sum_{i=1}^n \mathbf{x}_i (\mathbf{0}) + \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^2 (\mathbf{1}) \\
&= \mathbf{h}^2 \sigma_K^2 + \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^2 \\
\sigma_x^2 &= E(\mathbf{x}^2) - E^2(\mathbf{x}) \\
&= \mathbf{h}^2 \sigma_K^2 + \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^2 - \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \right)^2
\end{aligned}$$

$$\sigma_x^2 = \mathbf{h}^2 \sigma_K^2 + \hat{\mathbf{S}}_x^2$$

En donde

σ_K^2 es la varianza del Kernel original

$\hat{\mathbf{S}}_x^2$ es un valor asociado a la muestra (varianza muestral)

La varianza σ_x^2 de la variable aleatoria depende linealmente de la varianza del kernel, pero cuadráticamente del ancho de banda \mathbf{h} , por ello este es el factor crítico.

5.3 Criterios para elegir el kernel

5.3.1 Varianza

El primer criterio para elegir al kernel más eficiente es seleccionar el de menor varianza. Este valor se suma al construir el estimador $\hat{\mathbf{f}}(\mathbf{x})$ y aumentará su dispersión. Los valores

calculados están en la primera columna de la **Tabla 1** al final de este artículo

5.3.2 Enlace

Definimos el coeficiente $\mathbf{c} = e^{-|\mathbf{d}|}$, en donde \mathbf{d} es el valor de la tangente en el borde.

El valor más alto es **1** como en el modelo normal, que se conecta con suavidad a los otros kernels. El menor valor es **0**, como en el kernel rectangular. Los valores calculados están en la segunda columna de la **Tabla 1**.

5.3.3 Efecto del factor enlace

Al sumar kernels el perfil resultante puede ser liso y continuo como el caso del kernel gaussiano, o cambiar abruptamente como el caso del kernel rectangular. Se debe seleccionar el kernel y el ancho de banda que permitan detectar detalles y una apariencia aceptable. A esto contribuye también el factor enlace. El siguiente ejemplo muestra dos kernels con diferente factor de enlace. Se observa la diferencia significativa alrededor de $\mathbf{x} = 3$.

Suponer una muestra $\mathbf{X}: 2, 4, 5$

Figura 8

Kernel Epanechnikov, $\mathbf{h} = 1.5$, enlace=0.2231

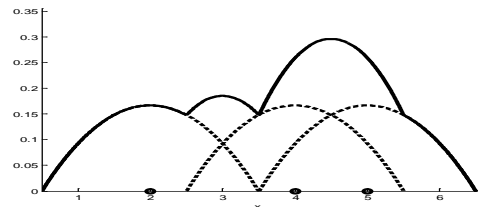
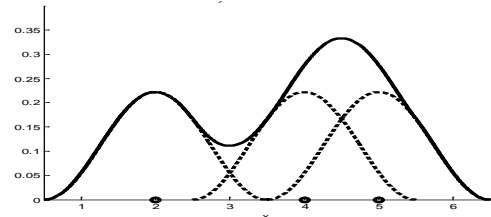


Figura 9

Kernel Coseno, $\mathbf{h} = 1.5$, enlace = 1



Como se verá más adelante, la valoración de los kernels cambia al analizarlos mediante otros criterios. Para esto se considerarán las restantes columnas de la **Tabla 1**.

En esta primera valoración, es mejor el kernel cuya varianza sea menor, adicionalmente el coeficiente de enlace debe ser alto para que el perfil del estimador sea liso.

6. EFICIENCIA DEL ESTIMADOR $\hat{f}(x)$ CON RESPECTO A $f(x)$

Sean

$f(x)$: Función de densidad de probabilidad teórica (desconocida)

$\hat{f}(x)$: Estimador de $f(x)$ basado en los datos y el kernel elegido

Definiciones

$B(\hat{f}(x)) = E[\hat{f}(x)] - f(x)$: Sesgo del estimador $\hat{f}(x)$ con respecto a $f(x)$

$V(\hat{f}(x)) = E[\hat{f}(x) - E(\hat{f}(x))]^2$: Varianza del estimador $\hat{f}(x)$ con respecto a $E(\hat{f}(x))$

$ECM(\hat{f}(x)) = E[\hat{f}(x) - f(x)]^2$: Error cuadrático medio. (Medición de la diferencia puntual)

Si se desarrolla el cuadrado y se sustituyen las definiciones se obtiene [7].

$$ECM(\hat{f}(x)) = B^2(\hat{f}(x)) + V(\hat{f}(x))$$

La siguiente definición determina la exactitud global del estimador, integrando ECM

$$ECMI(\hat{f}(x)) = \int_{\mathbb{R}} ECM(\hat{f}(x)) dx = \int_{\mathbb{R}} B^2(\hat{f}(x)) dx + \int_{\mathbb{R}} V(\hat{f}(x)) dx$$

Desarrollo de los componentes

$$E(\hat{f}(x)) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} E[K\left(\frac{x-x_i}{h}\right)] = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{x-t}{h}\right) f(t) dt = \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{x-t}{h}\right) f(t) dt$$

Con las sustituciones: $z = \frac{x-t}{h}$, $t = x-hz$, $dt = h dz$

$$E(\hat{f}(x)) = \int_{\mathbb{R}} K(z) f(x-hz) dz$$

Esto muestra que $E(\hat{f}(x)) \neq f(x)$ pero $E(\hat{f}(x)) \rightarrow f(x)$ cuando $h \rightarrow 0$ siempre que f sea continua y acotada. Entonces $\hat{f}(x)$ es asintóticamente insesgado. Este resultado parece extraño pues cuando $h \rightarrow 0$, la distribución de $\hat{f}(x)$ se hace puntual.

Para obtener alguna aproximación, y suponiendo que f es diferenciable alrededor de x se desarrolla $f(x-hz)$ con la serie de Taylor. x es la variable aleatoria del estimador $\hat{f}(x)$, t , z son variables aleatorias del modelo teórico $f(x)$ pero x también depende de t , z

$$f(x-hz) = f(x) - hzf'(x) + \frac{1}{2}(hz)^2 f''(x) + O(h^3)$$

$$E(\hat{f}(x)) = \int_{\mathbb{R}} K(z) f(x) dz -$$

$$\int_{\mathbb{R}} K(z) hzf'(x) dz + \int_{\mathbb{R}} \frac{1}{2} K(z) (hz)^2 f''(x) dz + O(h^3)$$

$$E(\hat{f}(x)) = f(x)(1) - hf'(x)(0) +$$

$$\frac{h^2}{2} f''(x) \int_{\mathbb{R}} z^2 K(z) dz + O(h^3)$$

$$= f(x) + \frac{h^2}{2} f''(x) \sigma_K^2 + O(h^3)$$

6.1 Sesgo del estimador $\hat{f}(x)$

Sustituyendo en $B(\hat{f}(x))$ y siendo σ_K^2 la varianza del kernel original se obtiene

$$B(\hat{f}(x)) \approx \frac{h^2}{2} f''(x) \sigma_K^2$$

Sesgo del estimador $\hat{f}(x)$ con respecto a $f(x)$

El sesgo de $\hat{f}(x)$ con respecto a $f(x)$ depende linealmente de la varianza del kernel K y cuadráticamente del ancho de banda h . Adicionalmente, aparece un nuevo factor, la derivada de la densidad teórica desconocida f . Se puede notar que el sesgo de $\hat{f}(x)$ no depende del tamaño muestral. También se observa que $\lim(B(\hat{f}(x))) = 0$, cuando $h \rightarrow 0$. Este resultado parece contradecir el hecho que si $h \rightarrow 0$, la distribución de $\hat{f}(x)$ se hace puntual.

6.2 Varianza del estimador $\hat{f}(x)$ [3]

$$V(\hat{f}(x)) = V\left(\frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x-x_i}{h}\right)\right) = \frac{1}{n^2 h^2} \sum_{i=1}^n V\left(K\left(\frac{x-x_i}{h}\right)\right)$$

x_i independientes, con igual distribución

$$V\left(K\left(\frac{x-x_i}{h}\right)\right) = E\left(\left(K^2\left(\frac{x-x_i}{h}\right)\right)\right) - [E\left(K\left(\frac{x-x_i}{h}\right)\right)]^2 = \int_{\mathbb{R}} K^2\left(\frac{x-t}{h}\right) f(t) dt - \left[\int_{\mathbb{R}} K\left(\frac{x-t}{h}\right) f(t) dt\right]^2$$

$$V(\hat{f}(x)) = \frac{1}{n^2 h^2} \sum_{i=1}^n \int_{\mathbb{R}} K^2\left(\frac{x-x_i}{h}\right) f(t) dt - \frac{1}{n^2 h^2} \sum_{i=1}^n \left[\int_{\mathbb{R}} K\left(\frac{x-t}{h}\right) f(t) dt\right]^2 = \frac{1}{n} \int_{\mathbb{R}} \frac{1}{h^2} K^2\left(\frac{x-x_i}{h}\right) f(t) dt - \frac{1}{n} \left(\int_{\mathbb{R}} K\left(\frac{x-t}{h}\right) f(t) dt\right)^2$$

Mediante la sustitución:

$$z = \frac{x-t}{h} \Rightarrow t = x-hz, dt = -h dz$$

$$V(\hat{f}(x)) = \frac{1}{nh} \int_{\mathbb{R}} K^2(z) f(x-hz) dz - \frac{1}{n} \left(\int_{\mathbb{R}} K(z) f(x-hz) dz\right)^2$$

Desarrollo de los componentes usando dos términos de la Serie de Taylor

$$\int_{\mathbb{R}} K^2(z) f(x-hz) dz = \int_{\mathbb{R}} K^2(z) [f(x) - hzf'(x) + O(h^2)] dz = \int_{\mathbb{R}} K^2(z) [f(x) - hzf'(x)] dz + O(h^2)$$

$$\int_{\mathbb{R}} K(z) f(x-hz) dz =$$

$$\int_{\mathbb{R}} K(z) [f(x) - hzf'(x) + O(h^2)] dz$$

$$= \int_{\mathbb{R}} K(z) f(x) dz - \int_{\mathbb{R}} K(z) hzf'(x) dz + O(h^2)$$

$$= f(x) \int_{\mathbb{R}} K(z) dz - hf'(x) \int_{\mathbb{R}} zK(z) dz + O(h^2)$$

$$= f(x)(1) - hf'(x)(0) + O(h^2) = f(x) + O(h^2)$$

Sustituyendo en la definición de varianza:

$$\begin{aligned} \mathbf{V}(\hat{\mathbf{f}}(\mathbf{x})) &= \frac{1}{nh} \int_{\mathbf{R}} \mathbf{K}^2(\mathbf{z}) \mathbf{f}(\mathbf{x} - h\mathbf{z}) d\mathbf{z} \\ &- \frac{1}{n} \left(\int_{\mathbf{R}} \mathbf{K}(\mathbf{z}) \mathbf{f}(\mathbf{x} - h\mathbf{z}) d\mathbf{z} \right)^2 \\ &= \frac{1}{nh} \left[\int_{\mathbf{R}} \mathbf{K}^2(\mathbf{z}) [\mathbf{f}(\mathbf{x}) - h\mathbf{z}\mathbf{f}'(\mathbf{x})] d\mathbf{z} + \right. \\ &\quad \left. \mathbf{O}(h^2) \right] - \frac{1}{n} [\mathbf{f}(\mathbf{x}) + \mathbf{O}(h^2)]^2 \end{aligned}$$

Si se supone que n es grande y h pequeño se llega a la siguiente aproximación

$$\mathbf{V}(\hat{\mathbf{f}}(\mathbf{x})) \approx \frac{1}{nh} \mathbf{f}(\mathbf{x}) \int_{\mathbf{R}} \mathbf{K}^2(\mathbf{z}) d\mathbf{z}$$

Según este resultado $\mathbf{V}(\hat{\mathbf{f}}(\mathbf{x}))$ aumenta si h se reduce.

6.3 Medición de la exactitud de $\hat{\mathbf{f}}(\mathbf{x})$

Sustituyendo $\mathbf{V}(\hat{\mathbf{f}}(\mathbf{x}))$ en la definición de $\mathbf{ECM}(\hat{\mathbf{f}}(\mathbf{x}))$ [3]

$$\begin{aligned} \mathbf{ECM}(\hat{\mathbf{f}}(\mathbf{x})) &= \mathbf{B}^2(\hat{\mathbf{f}}(\mathbf{x})) + \mathbf{V}(\hat{\mathbf{f}}(\mathbf{x})) \approx \\ &\left(\frac{h^2}{2} \mathbf{f}''(\mathbf{x}) \sigma_{\mathbf{K}}^2 \right)^2 + \frac{1}{nh} \mathbf{f}(\mathbf{x}) \int_{\mathbf{R}} \mathbf{K}(\mathbf{z})^2 d\mathbf{z} \end{aligned}$$

Finalmente, integrando sobre \mathbf{x}

$$\begin{aligned} \mathbf{ECMI}(\hat{\mathbf{f}}(\mathbf{x})) &\approx \frac{1}{4} h^4 (\sigma_{\mathbf{K}}^2)^2 \int_{\mathbf{R}} (\mathbf{f}''(\mathbf{x}))^2 d\mathbf{x} + \\ &\frac{1}{nh} \int_{\mathbf{R}} \mathbf{K}^2(\mathbf{z}) d\mathbf{z} \int_{\mathbf{R}} \mathbf{f}(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Pero $\int_{\mathbf{R}} \mathbf{f}(\mathbf{x}) d\mathbf{x} = \mathbf{1}$, y se llega a la siguiente expresión

$$\begin{aligned} \mathbf{ECMI}(\hat{\mathbf{f}}(\mathbf{x})) &\approx \frac{1}{4} h^4 (\sigma_{\mathbf{K}}^2)^2 \int_{\mathbf{R}} (\mathbf{f}''(\mathbf{x}))^2 d\mathbf{x} + \\ &\frac{1}{nh} \int_{\mathbf{R}} \mathbf{K}^2(\mathbf{z}) d\mathbf{z} \end{aligned}$$

Se puede ver que ambos componentes actúan en forma inversa al variar h .

Para determinar el ancho de banda h , tal que $\mathbf{ECMI}(\hat{\mathbf{f}}(\mathbf{x}))$ sea mínimo [1]:

$$\frac{d\mathbf{ECMI}(\hat{\mathbf{f}}(\mathbf{x}))}{dh} = 0$$

$$h^3 (\sigma_{\mathbf{K}}^2)^2 \int_{\mathbf{R}} (\mathbf{f}''(\mathbf{x}))^2 d\mathbf{x} - \frac{1}{nh^2} \int_{\mathbf{R}} \mathbf{K}^2(\mathbf{z}) d\mathbf{z} = 0$$

De donde se obtiene la siguiente fórmula para estimar h que minimiza a $\mathbf{ECMI}(\hat{\mathbf{f}}(\mathbf{x}))$

$$h^* = \left[\frac{1}{n} \frac{(\sigma_{\mathbf{K}}^2)^{-2} \int_{\mathbf{R}} \mathbf{K}^2(\mathbf{z}) d\mathbf{z}}{\int_{\mathbf{R}} (\mathbf{f}''(\mathbf{x}))^2 d\mathbf{x}} \right]^{1/5}$$

Separando en sus tres componentes:

$$h^* = n^{-\frac{1}{5}} \left[\int_{\mathbf{R}} (\mathbf{f}''(\mathbf{x}))^2 d\mathbf{x} \right]^{-\frac{1}{5}} \left[\frac{\int_{\mathbf{R}} \mathbf{K}^2(\mathbf{z}) d\mathbf{z}}{(\sigma_{\mathbf{K}}^2)^2} \right]^{1/5}$$

Se observa que el valor óptimo h^* que minimiza el valor de $\mathbf{ECMI}(\hat{\mathbf{f}}(\mathbf{x}))$ depende de tres componentes independientes: el tamaño de la muestra n , el modelo de densidad teórico desconocido $\mathbf{f}''(\mathbf{x})$, y el kernel elegido: \mathbf{K} y $\sigma_{\mathbf{K}}^2$.

Los métodos usados en la actualidad para obtener h^* se basan en métodos para estimar $\mathbf{f}''(\mathbf{x})$ mediante una aproximación basada en los mismos datos muestrales y en suposiciones acerca del modelo $\mathbf{f}(\mathbf{x})$. Es importante anotar que la validez de esta fórmula requiere n grande y h pequeño.

En resumen, existe bastante incertidumbre en la estimación de h^* pero si se dispone de software se puede experimentar directamente con los modelos, los datos y el valor de h .

Sustituyendo h^* en $\mathbf{ECMI}(\hat{\mathbf{f}}(\mathbf{x}))$ se obtiene la expresión con el valor mínimo para el error global del estimador $\hat{\mathbf{f}}(\mathbf{x})$

$$\begin{aligned} \mathbf{ECMI}^* &\approx \frac{5}{4} n^{-\frac{4}{5}} \left(\int_{\mathbf{R}} \mathbf{K}^2(\mathbf{z}) d\mathbf{z} \right)^{\frac{4}{5}} (\sigma_{\mathbf{K}}^2)^{\frac{2}{5}} \left(\int_{\mathbf{R}} (\mathbf{f}''(\mathbf{x}))^2 d\mathbf{x} \right)^{\frac{1}{5}} \end{aligned}$$

La exactitud global del estimador depende del tamaño de la muestra en el orden $n^{-\frac{4}{5}}$ y del kernel elegido \mathbf{K} , pero también de la densidad desconocida que se desea estimar $\mathbf{f}(\mathbf{x})$, por lo que no se puede calcular directamente. La columna 4 de la **Tabla 1** muestra la contribución de cada kernel para el valor de $\mathbf{ECMI}(\hat{\mathbf{f}}(\mathbf{x}))$ y se puede constatar que el kernel que más reduce este valor es el de **Epanechnikov**. La diferencia con respecto a los otros kernels no es muy significativa, por lo tanto, si n es un valor fijo, la medida de $\mathbf{ECMI}(\hat{\mathbf{f}}(\mathbf{x}))$ depende principalmente del factor desconocido, la densidad $\mathbf{f}(\mathbf{x})$.

6.4 Cálculo del ancho de banda óptimo

Para evaluar $\mathbf{f}''(\mathbf{x})$ se supondrá que $\mathbf{f}(\mathbf{x})$ tiene distribución normal. Este es el caso más común.

Sea $\mathbf{f}(\mathbf{x}) = \mathbf{N}(\mathbf{0}, \sigma^2)$ densidad normal con media cero y varianza σ^2

$$\mathbf{f}(\mathbf{x}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

Entonces

$$\begin{aligned} (\mathbf{f}''(\mathbf{x}))^2 &= \frac{(\sigma^2 - x^2)^2}{2\pi\sigma^{10}} e^{-\frac{x^2}{\sigma^2}} = \\ &\frac{1}{2\pi\sigma^{10}} \left(\sigma^4 e^{-\frac{x^2}{\sigma^2}} - 2\sigma^2 x^2 e^{-\frac{x^2}{\sigma^2}} + x^4 e^{-\frac{x^2}{\sigma^2}} \right) \end{aligned}$$

Con la sustitución $x = \sigma u$, $dx = \sigma du$

$$\begin{aligned} \int_{\mathbb{R}} (f''(x))^2 dx &= \int_{\mathbb{R}} (f''(\sigma u))^2 \sigma du, \\ &= \frac{\sigma}{2\pi\sigma^{10}} (\sigma^4 \int_{\mathbb{R}} e^{-u^2} du - 2\sigma^4 \int_{\mathbb{R}} u^2 e^{-u^2} du \\ &\quad + \sigma^4 \int_{\mathbb{R}} u^4 e^{-u^2} du) \\ &= \frac{\sigma^5}{2\pi\sigma^{10}} (1.7724 - 2(0.8862) + 1.3293) \\ &= 0.2115 \sigma^{-5} \end{aligned}$$

6.5 Ancho de banda óptimo para el kernel Gaussiano

Para obtenerlo se elige el kernel Gaussiano. Los valores se toman de las columnas 1 y 3 de la **Tabla 1** y se reemplazan en la fórmula de h^*

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in (-\infty, +\infty)$$

$$\int_{\mathbb{R}} K^2(z) dz = 0.2821, \quad \sigma_K^2 = 1$$

Sustituyendo en h^* ,

$$\begin{aligned} h_G^* &= n^{-\frac{1}{5}} (0.2115 \sigma^{-5})^{-\frac{1}{5}} \left(\frac{0.2821}{1^2}\right)^{\frac{1}{5}} \\ &= 1.0592 \sigma n^{-\frac{1}{5}} \end{aligned}$$

Este resultado es bien conocido. Es adecuado si $f(x)$ se parece a la distribución normal. En este trabajo se usaron métodos numéricos para calcularlo.

Se han desarrollado modificaciones a esta fórmula. La siguiente se debe a Silverman y funciona bien para diferentes tipos de densidades [6].

$$h_G^* = 0.9 n^{-\frac{1}{5}} \min\left(\sigma, \frac{\text{rango intercuartil}}{1.349}\right)$$

En donde σ puede sustituirse con una estimación tomada de la muestra.

Ejemplo. Para entender la aplicación de la fórmula usamos el micro ejemplo anterior:

$$X: 1.1, 2.1, 2.3, 2.7, 3.8$$

$$S^2 = 0.96 \quad (\text{varianza muestral})$$

$$\sigma \approx S = \sqrt{S^2} = \sqrt{0.96} = 0.9798$$

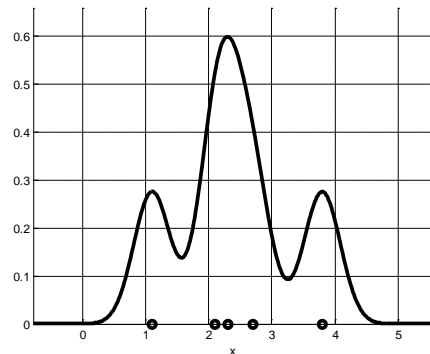
$$\text{Rango intercuartil} = 2.7 - 2.1 = 0.6$$

$$\begin{aligned} h_G^* &= 0.9 (5^{-\frac{1}{5}}) \min(0.9798, 0.6/1.349) \\ &= 0.2901 \end{aligned}$$

Estos datos suministrados al programa **KDEN** produjeron el siguiente gráfico con el estimador kernel Gaussiano. El resultado se muestra en la **Figura 10**, muy similar en forma al que se

obtuvo en la **Figura 6**, con otro kernel y con un ancho de banda elegido intuitivamente.

Figura 10
Kernel Gaussiano, $h = 0.2901$



6.6 Ancho de banda óptimo para otro kernel

Si se requiere usar un kernel diferente, se puede usar el valor óptimo h_G^* del kernel Gaussiano para convertirlo en un valor óptimo h^* para el kernel seleccionado.

Sean

$K_G(x)$: Kernel Gaussiano con $h = 1$

$K(x)$: Kernel elegido

$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$: Kernel parametrizado con h

Proponemos la siguiente expresión para encontrar el valor h_K^* para el kernel seleccionado K_h . Este es el valor de h que minimiza la diferencia global con respecto al kernel Gaussiano K_G :

$$h_K^* = \arg \min_h \left(\int_{-h}^h |K_G(x) - K_h(x)| dx + 2 \int_h^\infty K_G(x) dx \right)$$

Su interpretación gráfica es el área sombreada en la **Figura 19** en la que se muestra el caso del kernel Coseno respecto al kernel Gaussiano. Los resultados h_K^* calculados para los kernels requirieron usar métodos numéricos y están en la antepenúltima columna de la **Tabla 1**.

En la penúltima columna de la **Tabla 1** está la varianza para el kernel y en la última columna está un criterio adicional para comparación de kernels según el cual es preferible el que tiene la mayor amplitud con menor varianza. Según este criterio, el mejor sería el kernel Coseno.

Se puede estimar el valor óptimo h^* para un kernel específico con la siguiente fórmula:

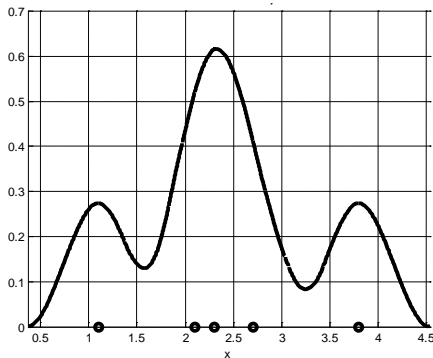
$$h^* = h_G^* h_K^*$$

Ejemplo. Para los datos del ejemplo anterior, determine el ancho de banda óptimo si se desea usar el kernel Cúbico Sujeto.

$$h^* = h_G^* h_K^* = 0.2901(2.5150) = 0.7296$$

El gráfico obtenido con el programa KDEN para el kernel Cúbico Sujeto y $h = 0.7296$ es muy parecido en el nivel de detalle al que muestra el gráfico del kernel Gaussiano con $h = 0.2901$

Figura 11
Kernel Cúbico Sujeto, $h = 0.7296$

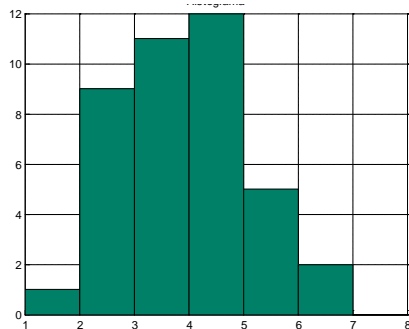


Ejemplo. La siguiente es una muestra ordenada del tiempo (minutos) que se utilizó para atender a 40 personas en una estación de servicio:

1.80 2.10 2.20 2.50 2.50 2.70 2.80 2.80
2.90 2.90 3.10 3.10 3.50 3.50 3.60 3.60
3.60 3.70 3.70 3.80 3.90 4.10 4.10 4.20
4.20 4.30 4.40 4.50 4.60 4.70 4.80 4.90
4.90 5.10 5.10 5.10 5.60 5.70 6.10 6.20

Analizar un modelo de densidad con KDEN

Figura 12
Histograma, amplitud de clase = 1



Método kernel (modelo Coseno)

Determinar el ancho de banda óptimo

$$S^2 = 1.2259 \quad (\text{Varianza})$$

$$\sigma \approx S = \sqrt{S^2} = \sqrt{1.2259} = 1.1072$$

$$Q_1 = 0.5(x_{10} + x_{11}) = 3 \quad (\text{Cuartiles})$$

$$Q_3 = 0.5(x_{30} + x_{31}) = 4.75$$

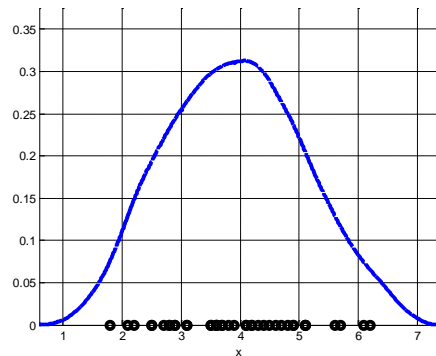
$$\text{Rango intercuartil} = 4.75 - 3 = 1.75$$

$$h_G^* = 0.9(40^{-\frac{1}{5}}) \min(1.1072, 1.75/1.349) = 0.4765$$

Para el modelo Coseno

$$h^* = h_G^* h_K^* = 0.4765(2.5260) = 1.2036$$

Figura 13
Kernel Coseno, $h = 1.2036$



Calcule la probabilidad que el tiempo de atención sea mayor a 5

Con el **histograma**

$$7/40 = 17.5\%$$

Con el **modelo kernel** y el programa **KDEN**

$$19.01\%$$

El programa **KDEN** calcula y sugiere el ancho de banda óptimo, pero si se desea observar más detalles de la función de densidad se puede experimentar con el programa, y así como se puede afinar el histograma reduciendo la amplitud de clase, también se puede ensayar con el estimador kernel cambiando el ancho de banda y probando otros kernels. En los siguientes gráficos se muestran algunos resultados con los mismos 40 datos del ejemplo anterior. En estos casos el ancho de banda ya no es el valor óptimo.

Figura 14
Histograma, amplitud de clase = 0.5

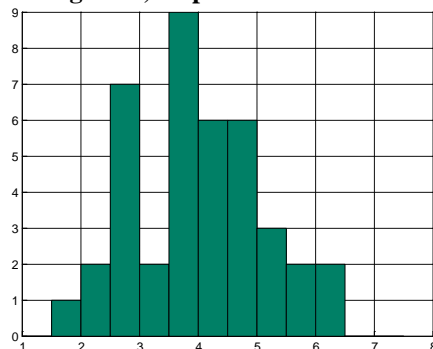


Figura 15
Kernel Coseno, $h=0.8$

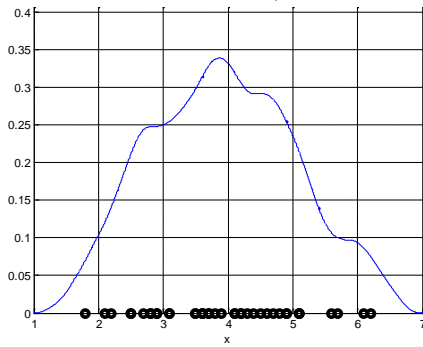


Figura 16
Kernel Coseno, $h=0.6$

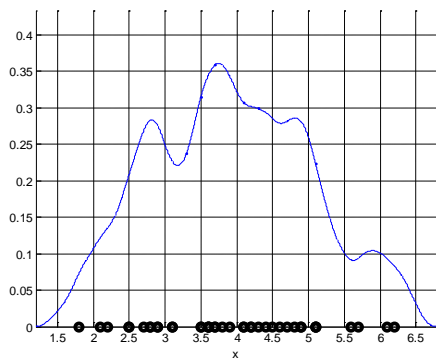
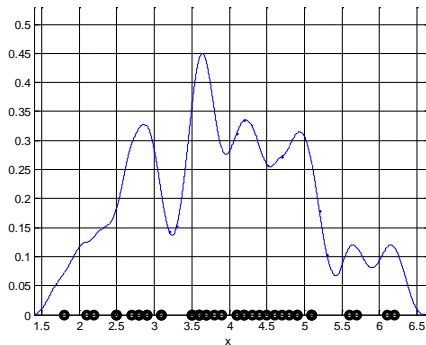


Figura 17
Kernel Coseno, $h=0.4$



7. CONCLUSIONES

Se realizaron ensayos con muestras de diferente tamaño y se obtuvieron resultados coherentes. Sin embargo siempre será conveniente realizar pruebas con varios kernels y sus parámetros y constatar si el modelo de probabilidad muestra los detalles que uno desea. Por ello la necesidad de tener un programa para experimentar hasta llegar al modelo que nuestra intuición nos dice que es adecuado. Esto ocurre especialmente cuando los datos tienen un patrón multimodal.

El programa KDEN calcula y sugiere el valor óptimo del ancho de banda usando el criterio desarrollado en este artículo. La literatura contiene otros métodos más complejos, para estimar este valor crítico del método kernel.

Es importante anotar que el desarrollo de este trabajo tuvo como soporte la aplicación de métodos numéricos y de un lenguaje computacional para diseñar e instrumentar el software para facilitar la investigación.

El código fuente MATLAB del programa KDEN está disponible en el Repositorio de la ESPOL para que usuarios interesados puedan descargarlo y mejorarlo.

Figura 18
Gráfico de algunos kernels

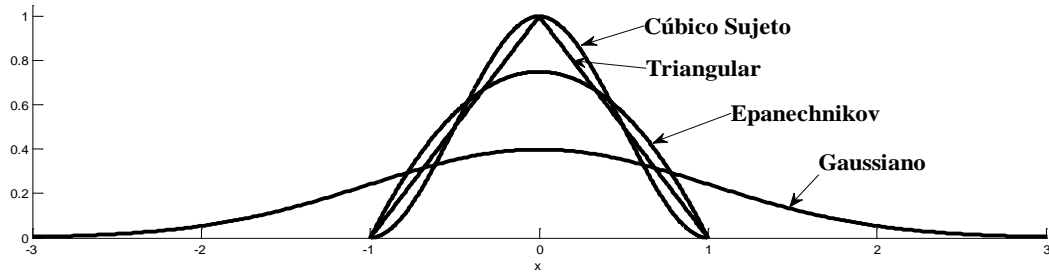


Figura 19
Gráfico del Kernel Coseno parametrizado sobre el kernel Gaussiano

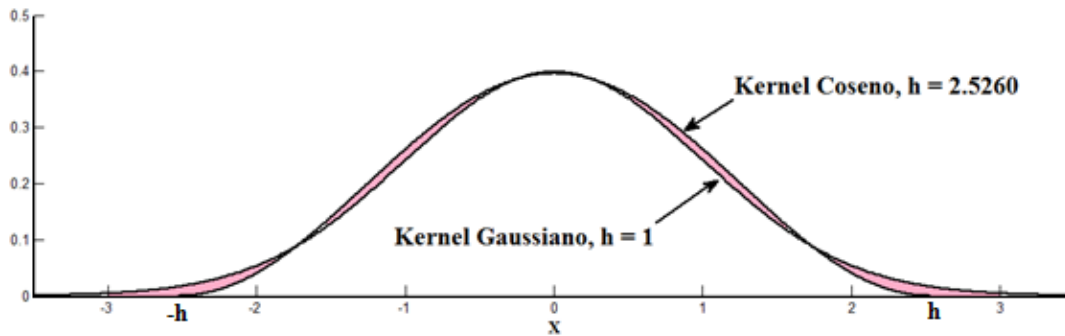


Tabla 1
Algunas medidas desarrolladas para comparar kernels

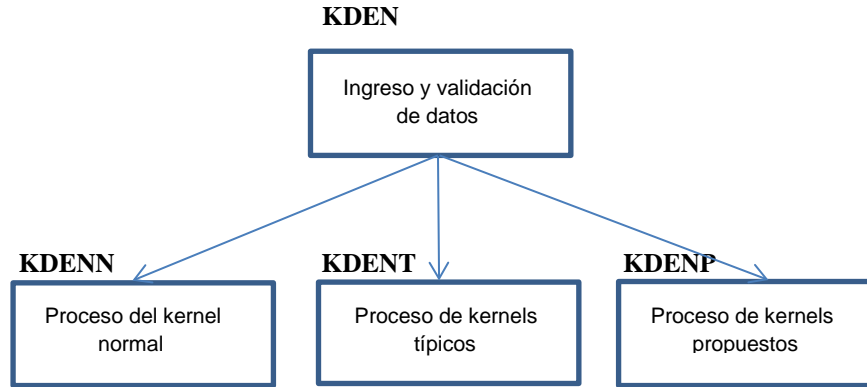
Kernel	σ_K^2	Enlace	$\int K^2(z)dz$	K_{ECMI}^*	h_K^*	σ_{Kh}^2	h_K^*/σ_{Kh}^2
Normal	1.0000	1.0000	0.2821	0.3633	1.0000	1.0000	1.0000
Rectangular	0.3333	0.0000	0.5000	0.3701	1.3800	0.6348	2.1739
Triangular	0.1666	0.3679	0.6666	0.3531	2.2120	0.8155	2.7125
Epanechnikov	0.2000	0.2231	0.6000	0.3491	1.9470	0.7582	2.5681
Arco coseno	0.1894	0.2912	0.6169	0.3492	1.9850	0.7464	2.6594
Biweight	0.1429	1.0000	0.7143	0.3508	2.3700	0.8024	2.9536
Arco circular	0.2228	0.0729	0.5709	0.3503	1.7810	0.7049	2.5267
Coseno	0.1307	1.0000	0.7500	0.3520	2.5260	0.8339	3.0292
Cúbico plano	0.2222	0.1353	0.5714	0.3502	1.9530	0.8476	2.3042
Cúbico natural	0.1867	0.3012	0.6217	0.3494	2.0180	0.7602	2.6547
Cúbico sujeto	0.1333	1.0000	0.7429	0.3521	2.5150	0.8434	2.9821

APÉNDICE A

DISEÑO CONCEPTUAL DEL PROGRAMA KDEN

Desarrollado modularmente en el lenguaje MATLAB con el soporte de su capacidad simbólica, numérica y gráfica.

A.1 Estructura de KDEN



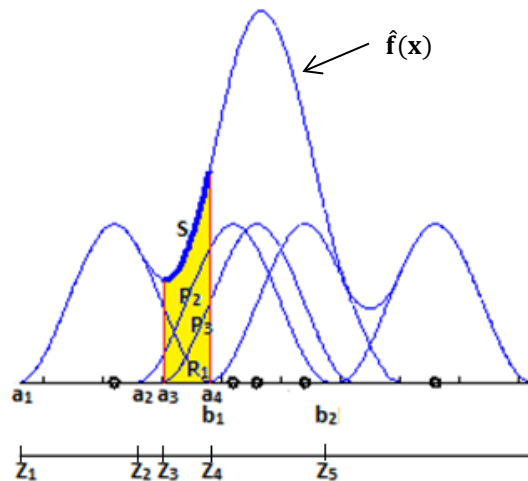
A.2 Módulos subyacentes

- Graficación de puntos
- Graficación de histograma
- Graficación de kernels
- Graficación de la función de densidad
- Graficación de kernels y la función de densidad
- Cálculos estadísticos muestrales
- Cálculos estadísticos de densidad
- Cálculo del ancho de banda óptimo
- Cálculo de probabilidad
- Descripción simbólica matemática de funciones
- Función para integración numérica
- Funciones para manejo de vectores y texto

A.3 Estructuras de datos

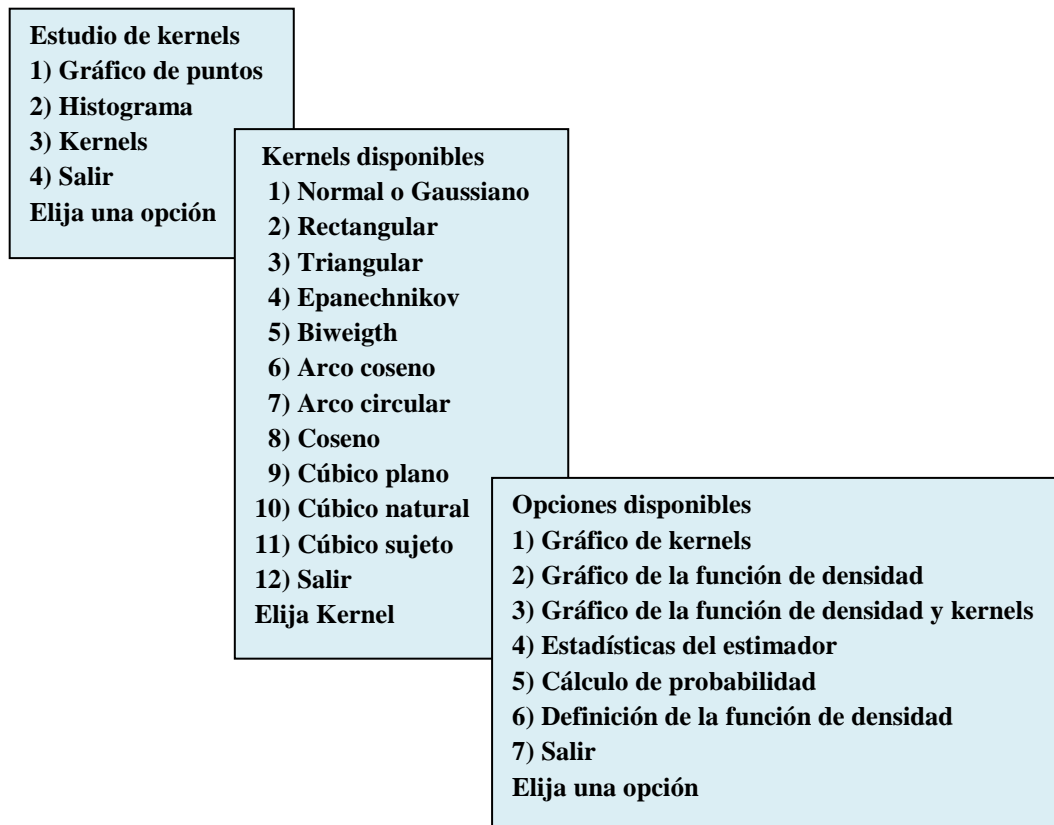
- Vector de celdas para almacenar kernels
- Vectores para almacenar bordes de kernels
- Vector para almacenar puntos de cambio de intervalo para la función de densidad
- Vector de celdas para almacenar sumas de kernels

A.4 Algoritmo para construir la función de densidad $\hat{f}(x)$



- 1) Seleccionar el kernel
- 2) Ingresar el vector con las observaciones
- 3) Ingresar el ancho de banda h
- 4) Aplicar la ponderación y colocar el kernel en cada punto muestral
- 5) Crear el vector con los bordes izquierdos y derechos a distancia h alrededor de cada punto muestral
- 6) Combinar los vectores de bordes, en un solo vector Z con los puntos de cambio de intervalo
- 7) Recorrer cada intervalo del vector Z y sumar los kernels en ese intervalo
- 8) Almacenar la suma en el vector de sumas de kernels
- 9) El vector de sumas de kernels es el estimador $\hat{f}(x)$

A.5 Interacción con el programa KDEN



REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

- (1) **BERTIN, KARINE. (2012)** “Estimación no-paramétrica de funciones”. Universidad de Valparaiso, Chile, pp26-pp63
- (2) **CORTES DE LA FUENTE, JORGE. (2011).** “La información mutua como medida de asociación y su utilidad en análisis genéticos”. Universidad Politécnica de Catalunya, pp29-pp32
- (3) **ZUCCHINI, WALTER. (2003).** “Kernel Density Estimation”, pp2-pp19
- (4) **HANSEN, BRUCE (2009).** “Lecture Notes on Nonparametrics”. University of Wisconsin, pp3-pp16
- (5) **MURAT, KAIRY. (2009).** “Kernel Smoothing Function and choosing Bandwidth for nonparametrics Regression Methods”. Ozean Journal of Applied Sciences, Ozean Publication, pp2-pp6
- (6) **BRUFMAN, JUANA, URBISAIA, HERIBERTO. (2006).** “Distribución del Ingreso Según Género: Un enfoque no paramétrico”. Cuadernos del CIMBAGE N0. 8, pp9-pp16
- (7) **PONOMAREVA, MARIA. (2010).** “Nonparametrics: regresión smoothing with kernels”. Northern Illinois University, pp2-pp13
- (8) **HANSEN, BRUCE (2004).** “Bandwidth Selection for Nonparametrics Distribution Estimation”. University of Wisconsin, pp3-pp9
- (9) **WAN, BING. (2007).** “Bandwidth Selection for Eweighted Kernel Density Estimation”. Electronic Journal of Statistics. ISSN 1935-7524, pp1-pp9

Escuela Superior Politécnica del Litoral
Facultad de Ciencias Naturales y Matemáticas
Departamento de Matemáticas

Manual de uso del programa KDEN

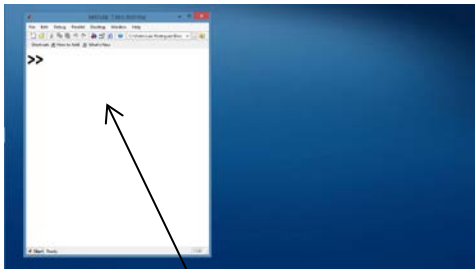
El programa **KDEN** permite construir un estimador de la función densidad de probabilidad para una muestra de datos independientes.

KDEN usa como soporte el lenguaje **MATLAB** y constituye un dispositivo para obtener la función de densidad probando diferentes modelos y variando un parámetro de ajuste. Los resultados que se muestran son gráficos, numéricos y simbólicos.

Procedimiento para usar KDEN

1) Ingrese a **MATLAB**

2) Reduzca el tamaño de la ventana de comandos. Se sugiere ubicarla en la mitad izquierda



Ventana de comandos de MATLAB

3) Seleccione la carpeta en la cual está el programa **KDEN**, si no está en la carpeta inicial de MATLAB

4) En la ventana de comandos defina un vector con los datos muestrales

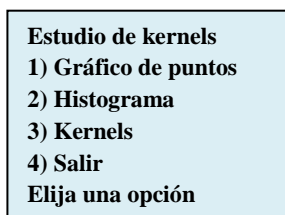
Ejm. Escriba

`>> x=[1.1, 2.1, 2.3, 2.7, 3.8];`

5) Ingrese al programa **KDEN**. Escriba

`>> kden`

6) El programa muestra el menú inicial. Elija una de las opciones.



7) Si elige la opción **1** ingrese el nombre del vector con los datos muestrales.

Ejm. Escriba **x**

Se muestra el gráfico de puntos y las medidas muestrales

8) Si elige la opción **2** ingrese el nombre del vector con los datos muestrales.

Ejm. Escriba **x**

Ingrese los bordes de las clases.

Ejm. Si los bordes de clase son 1, 2, 3, 4 escriba **1:4**

Ejm. Si los bordes de clase son 1, 1.5, 2, 2.5, 3, 3.5, 4 escriba **1:0.5:4**

Se muestra el histograma y las medidas muestrales

9) Si elige la opción **3**, se muestra el menú de kernels disponibles

- Kernels disponibles**
- 1) Normal o Gaussiano
 - 2) Rectangular
 - 3) Triangular
 - 4) Epanechnikov
 - 5) Biweight
 - 6) Arco coseno
 - 7) Arco circular
 - 8) Coseno
 - 9) Cúbico plano
 - 10) Cúbico natural
 - 11) Cúbico sujeto
 - 12) Salir
- Elija Kernel

Ingrese el número del kernel.

Ejm. Si quiere el kernel Coseno, escriba **8**

Ingrese el nombre del vector con los datos muestrales.

Ejm. Escriba **x**

El programa muestra el ancho de banda óptimo sugerido. Escriba este valor o el que desea probar

El programa muestra las opciones disponibles

- Opciones disponibles**
- 1) Gráfico de kernels
 - 2) Gráfico de la función de densidad
 - 3) Gráfico de la función de densidad y kernels
 - 4) Estadísticas del estimador
 - 5) Cálculo de probabilidad
 - 6) Definición de la función de densidad
 - 7) Salir
- Elija una opción

Las opciones **1, 2, 3** producen resultados gráficos. Para copiar un gráfico a algún documento marque **Edit** en el menú del gráfico y seleccione la opción **Copy Figure**

La opción **4** muestra las medidas estadísticas básicas de la función de densidad

La opción **5** permite calcular el valor de probabilidad en un rango especificado. Si no se escribe alguno de los dos extremos del rango, se calcula la probabilidad desde el inicio o hasta el final del dominio de la densidad de probabilidad, según corresponda.

La opción **6** muestra la forma algebraica de la función de densidad

Si la cantidad de datos es grande, el programa demorará algunos segundos para responder

10) Elija la opción **Salir** para retroceder al nivel anterior y probar otras opciones de **KDEN**

