Estimación de la función de densidad

Estadística Computacional I

Departamento de Estadística e Investigación Operativa (Univ. Sevilla)

Introducción

Introducción

- Dada $X_1, \ldots, X_n$ i.i.d. $\rightsquigarrow F(x)$, se trata de obtener una estimación de la función de densidad f(x).
- Una opción consiste en identificar una ley de distribución a la que se pueda ajustar la muestra y posteriormente estimar sus parámetros.
- Sin embargo, la propia identificación de la ley puede requerir un procedimiento previo que proporcione una aproximación a la forma de la función de densidad de la que proviene la muestra.
- Por tanto, se requieren técnicas que proporcionen una estimación no paramétrica de la función de densidad.
- Un método clásico es el histograma (algunos autores lo datan en 1840). Su principal ventaja es la sencillez y facilidad de construcción.
- Sin embargo, presenta deficiencias importantes.

- La construcción de un histograma requiere elegir un número M de intervalos que además deben ser configurados:

$$I_1, \ldots, I_M, \qquad I_j = (e_j - 1, e_j]$$

De este modo, el estimador se define de forma local en cada intervalo I_j:

$$\hat{f}(x) = \frac{card(X_i \in I_j)}{n \times (e_j - e_{j-1})}$$

- Inconvenientes que presenta el histograma:
  - Dependencia del número y configuración de los intervalos.
  - La estimación resultante es de tipo discontinuo.
  - En la estimación en cada punto solo intervienen las observaciones muestrales situadas en el mismo intervalo que el punto.
  - En datos multivariantes, el número de regiones se eleva rápidamente y se requieren muestras muy grandes (Maldición de la Dimensionalidad).

Método ASH

- Una mejora simple consiste en construir B histogramas y obtener un histograma medio. En inglés se conoce con el nombre de **"Average Shifted Histogram" (ASH)**.
- La idea es que los B histogramas se basen en el mismo número M de intervalos y amplitud h, pero desplazando el extremo inferior del primer intervalo. Así, estos orígenes son, por ejemplo, $e_1, e_1 + h/B, e_1 + 2h/B, \ldots, e_1 + (B-1)h/B$.

$$\hat{f}(x) = \frac{1}{B} \sum_{i=1}^{B} \hat{f}_b(x)$$

- Este método reduce la sensibilidad al origen y computacionalmente se puede implementar de manera eficiente teniendo en cuenta que hay BM intervalos en total, cada uno es $(kh/B, (k+1)h/B]$, con centro $(k + 0.5)h/B,\ k = 1, 2, \ldots, BM - 1$.

Mejora

- Otra posible mejora es permitir que todos los intervalos sean tenidos en cuenta en la estimación en cualquier punto, definiendo para ello una suma ponderada, asignando más peso cuanto más cercano a x sea el intervalo:

$$\hat{f}(x) = \frac{1}{B} \sum_{k=1}^{BM} w\,(l_k - x) c_k$$

  - w es una función peso,
  - l_k es el centro de cada intervalo, y
  - c_k es la proporción de observaciones muestrales en el intervalo k.

Método del núcleo

Método del núcleo

- El estimador histograma se define de forma local en cada intervalo.
- Sea S un intervalo de los considerados en el histograma $S = (x - h, x + h]$.
- La **longitud "volumen"** de ese intervalo de amplitud h centrado en x es $V = 2h$.
- Por tanto, dado $X_1, \ldots, X_n$, **el estimador histograma de la función de densidad** es la frecuencia relativa de observaciones muestrales que pertenecen al intervalo S dividido por su amplitud.
- Denotando por N_S el número de observaciones en S y definiendo la función indicador $I(u) = 1$ si u es cierto, 0 en otro caso:

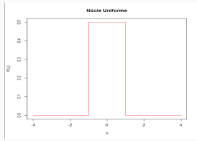
$$\hat{f}(x) = \frac{N_S}{n\,V} = \frac{N_S}{2\,n\,h} = \frac{1}{2\,n\,h}\sum_{i=1}^{n} I_{\{|x-X_i|<h\}} \qquad (1)$$

- El estimador (1) da el mismo peso a todas las observaciones dentro del intervalo S.
- Se puede escribir también de la siguiente forma

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

definiendo para ello:

$$K(u) = \frac{1}{2}\, I_{\{|u|<1\}}$$



- Esta función K es una **función de densidad simétrica definida en el intervalo (-1,1)**, en concreto la densidad de una **ley Uniforme Continua**.
- El **método núcleo utiliza funciones K más generales**, que comparten estas propiedades, pero de forma que los puntos más cercanos a x reciban un peso más elevado.

- Se tiene así el **estimador núcleo de la función de densidad**:

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

- Donde K es una función núcleo y h es el parámetro amplitud.
- En general K es una función simétrica y positiva, y permite asignar a cada X_i un peso que depende de la distancia a x.
- La función **`density`** de R ofrece varias **funciones núcleo**, entre las cuales se pueden destacar las siguientes, para las que se muestra también

$$R(K) = \int K^2(u)\, du$$

- **Normal (Gaussiano)**:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \qquad R(K) = \frac{1}{2\sqrt{\pi}}$$

- **Epanechnikov**:

$$K(u) = \frac{3}{4}(1 - u^2)\, I_{\{|u|<1\}}, \qquad R(K) = 3/5$$

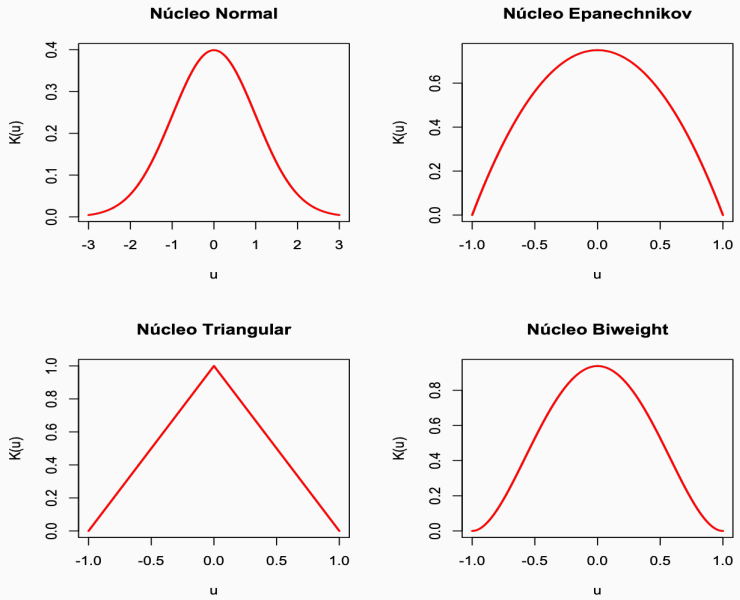
- **Triangular**:

$$K(u) = (1 - |u|)\, I_{\{|u|<1\}}, \qquad R(K) = 2/3$$

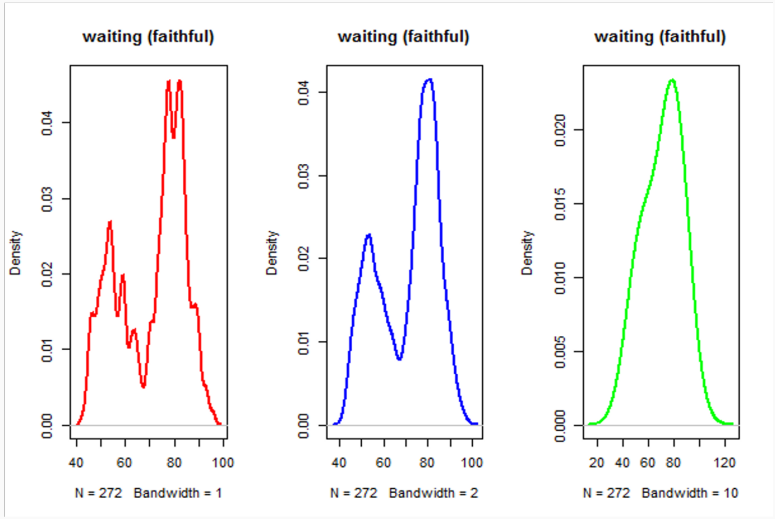
- **Biweight**:

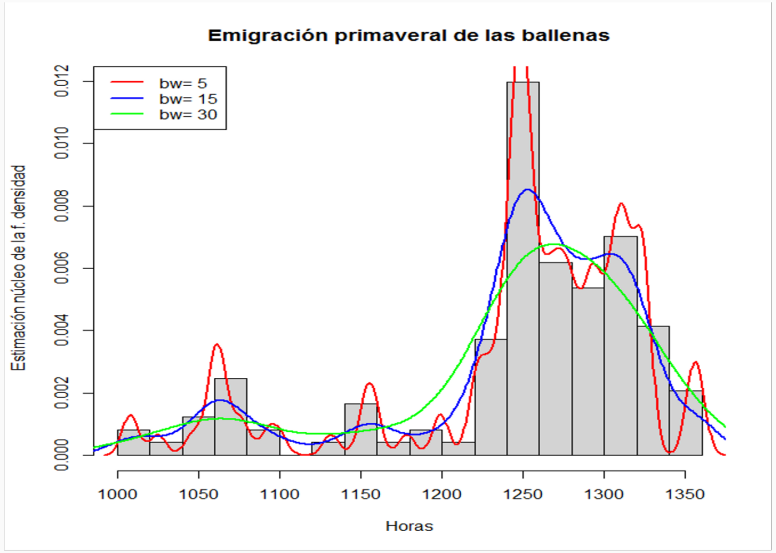
$$K(u) = \frac{15}{16}(1 - u^2)^2\, I_{\{|u|<1\}}, \qquad R(K) = 5/7$$

- El **núcleo Epanechnikov minimiza el criterio ECMIA** que se define más adelante, pero en general el rendimiento de la estimación depende poco de la función núcleo elegida.



Efecto del ancho de ventana (núcleo gaussiano)





Algunas propiedades de los estimadores núcleo

- Teorema.

Si se verifica

i) K está acotado.

ii) $\int |K(x)|\, dx < \infty$

iii) $|x|\ |K(x)| \to 0$, cuando $|x| \to \infty$.

iv) $\int K(x)\, dx = 1$

y f es continua en x, entonces

$$E\left\{\hat{f}(x)\right\} \to f(x)$$

Algunas propiedades de los estimadores núcleo

- **Teorema**.

Si se verifica

i) K está acotado.

ii) $\int |K(x)|\, dx < \infty$

iii) $|x|\, |K(x)| \to 0$, cuando $|x| \to \infty$.

y f es continua en x, entonces

$$ECM\left(\hat{f}(x)\right) \to 0, \qquad \hat{f}(x) \to^P f(x)$$

- Es evidente que el parámetro fundamental es la amplitud h.
- **h pequeño**: se ajusta demasiado a las particularidades de la muestra, sesgo pequeño y alta varianza.
- **h grande**: el estimador es demasiado suave, sesgo alto y varianza pequeña. Se pierden detalles como múltiples modas.
- Se define el **criterio Error Cuadrático Medio** como

$$ECM_h(\hat{f}(x)) = E\left[\left\{\hat{f}(x) - f(x)\right\}^2\right]$$

- Se verifica que

$$ECM_h(\hat{f}(x)) = \underbrace{\left\{E(\hat{f}(x)) - f(x)\right\}^2}_{sesgo^2(\hat{f}(x))} + \underbrace{E\left[\hat{f}(x) - E(\hat{f}(x))\right]^2}_{varianza(\hat{f}(x))}$$

- Integrando ECM se tiene el criterio Error Cuadrático Medio Integrado (ECMI)

$$ECMI(h) = \int ECM_h(\hat{f}(x))\, dx = \int \left\{sesgo(\hat{f}(x))^2 + var(\hat{f}(x))\right\}\, dx$$

- Algunos cálculos conducen a:

$$\int sesgo(\hat{f}(x))^2\, dx = \frac{1}{4} h^4\, \sigma_K^4\, R(f'') + o(h^4)^*, \qquad \int var(\hat{f}(x))\, dx = \frac{R(K)}{n\, h} + o\left(\frac{1}{n\, h}\right)$$

donde (*) es: $o(h^4)/h^4 \xrightarrow[h \to 0]{} 0$.

- Por tanto se tiene la siguiente expresión, donde **ECMIA es el ECMI asintótico**:

$$ECMI(h) = ECMIA(h) + o\left(h^4 + \frac{1}{n\, h}\right), \qquad ECMIA(h) = \frac{h^4\, \sigma_K^4\, R(f'')}{4} + \frac{R(K)}{n\, h}$$

- Si se minimiza $ECMIA(h)$, se obtiene

$$h = \left(\frac{R(K)}{n\, \sigma_K^4\, R(f'')}\right)^{1/5}$$

- Dado que f es desconocida, se han propuesto métodos para estimar $R(f'')$. **Silverman** propuso reemplazar f por la densidad de una normal cuya varianza sea la estimada a partir de la muestra, en tal caso, denotando por ϕ la densidad de la $N(0, 1)$

$$R(f'') = \frac{R(\phi'')}{\hat{\sigma}^5} \Rightarrow h = \left(\frac{4}{3\, n}\right)^{1/5} \hat{\sigma}$$

- El propio Silverman propuso utilizar una estimación robusta de la desviación típica:

$$h = \left(\frac{4}{3n}\right)^{1/5} \cdot Min\left(\hat{\sigma},\ \frac{R_I}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)}\right) \approx \left(\frac{4}{3n}\right)^{1/5} \cdot Min\left(\hat{\sigma},\ \frac{R_I}{1.35}\right)$$

- Este es el método **bw.nrd** disponible en la función **density** de R, mientras que **bw.nrd0** difiere ligeramente.
  - En cualquier caso este método es poco recomendable, ya que tiene tendencia a suavizar demasiado, pero puede proporcionar estimaciones iniciales para métodos más sofisticados.
- Suele ser preferible el método **bw.SJ**, que implementa la propuesta de **Sheather y Jones**, basado en la estimación núcleo de f''.
  - Se puede considerar como un proceso en dos etapas, en la primera se utiliza una regla simple como la de **bw.nrd** para calcular un valor h_0 que se utiliza en la estimación de $R(f'')$, y posteriormente se calcula h mediante la fórmula para minimizar el **criterio ECMIA**.

Proceso de validación cruzada

- Otra estrategia es seguir un proceso de validación cruzada. La siguiente expresión define el estimador núcleo en cada punto muestral calculado con el resto de la muestra:

$$\hat{f}_{-i}(X_i) = \frac{1}{h(n-1)} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right)$$

- En general, el método de validación cruzada se utiliza para ajustar los parámetros de un modelo intentando optimizar algún criterio de ajuste. En este caso se considera el **Error Cuadrático Integrado**

$$ECI(h) = \int \left(\hat{f}(x) - f(x)\right)^2 dx = \int \hat{f}(x)^2 \, dx - 2\, E[\hat{f}(x)] + \int f(x)^2 \, dx =$$
$$= R(\hat{f}) - 2\, E[\hat{f}(x)] + R(f)$$

Proceso de validación cruzada

- $R(f)$ es constante, mientras que el segundo término puede estimarse empleando la media de los estimadores arriba definidos. Se obtiene así el criterio de validación cruzada insesgado, utilizado en la opción **bw.ucv**, donde se minimiza el siguiente criterio:

$$UCV(h) = R(\hat{f}) - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i)$$

- Se le llama insesgado porque

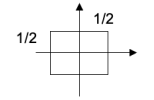
$$E\left\{UCV(h) + R(f)\right\} = ECMI(h)$$

- Si se utiliza el núcleo gaussiano, $UCV(h)$ admite una expresión analítica que facilita los cálculos.
- Algunos autores destacan la gran variabilidad de este método, es decir, ligeros cambios en la muestra pueden producir resultados muy diferentes.
- El **criterio de validación cruzada sesgado** (**bw.bcv**) intenta reducir esta variabilidad **minimizando una estimación de** $ECMIA(h)$. El texto de *Givens y Hoeting (2005)* recomienda el **método de Sheather y Jones**.

Estimación núcleo sobre datos multivariantes

- La estimación núcleo de la función de densidad puede extenderse a datos multivariantes.
- La estimación histograma se basa en una región p-variante S que es un hipercubo de lado h centrado en x: $V = h^p$.
- Si se considera una función núcleo uniforme,

$$K(u) = \begin{cases} 1 & \text{si } |u_j| < \frac{1}{2}, \forall j = 1, \ldots, p \\ 0 & \text{c.c.} \end{cases}$$



- Dada $X_1, \ldots, X_n$

$$K\left(\frac{x - X_i}{h}\right) = \begin{cases} 1 & \text{si } |x_j - X_{ij}| < \frac{h}{2}, \forall j = 1, \ldots, p \\ 0 & \text{c.c.} \end{cases}$$

- Se deduce pues que:

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

- Obteniéndose con una función núcleo general K:

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

- Se definen funciones núcleo como la **gaussiana**:

$$K(u) = \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}||u||^2}$$

Método de los k vecinos más próximos

- El método del núcleo fija h para todo punto.
  - Si h es muy grande, en algunas regiones la suavización es excesiva.
  - Si h es pequeño, en zonas de baja densidad el estimador es muy variable.
- Una alternativa es fijar NS permitiendo que varíe el volumen V.
- Dada la muestra, se calculan las distancias

$$d_i(x) = ||x - X_i||$$

- Supongamos que se ordenan los puntos de la muestra de modo que
- Dado x y el entero k, se considera la región $S = (x - d_k(x), x + d_k(x))$.
- Se define entonces

$$\boxed{\hat{f}(x) = \frac{k-1}{2\, n\, d_k(x)}}$$

- **Justificación**: en un intervalo $(x - r, x + r)$ cabe esperar $2rnp(x)$ elementos de la muestra, y por definición cabe esperar $k - 1$ observaciones en $(x - d_k(x), x + d_k(x))$.
- El estimador no es continuo, como los histogramas. Una alternativa es el **estimador generalizado de los k vecinos más próximos**.

$$\boxed{\hat{f}(x) = \frac{1}{n\, d_k(x)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{d_k(x)}\right)}$$

Es un estimador núcleo con h igual a $d_k(x)$.

Método de los logsplines

- Un **spline cúbico** es una función polinomial cúbica a trozos que es dos veces continuamente diferenciable pero cuya tercera derivada puede ser discontinua en un número finito de puntos, llamados nudos. Por tanto, está construido a partir de distintos polinomios de tercer orden, cada uno definido en un intervalo.
- El **método de logsplines** estima el logaritmo de la función de densidad mediante un **spline cúbico**.
- Para estimar una función de densidad en el intervalo, posiblemente no finito, (L, U), supongamos $M \geq 3$ nudos, $t_j,\ j = 1, 2, \ldots, M$: $L < t_1 < t_2 < \ldots < t_M < U$.
- Sea S el espacio M-dimensional de splines cúbicos con nodos en t_j y que son lineales en $(L, t_1]$ y $[t_M, U)$. Consideremos una base $\{1, B_1, \ldots, B_{M-1}\}$ para S.

- Existen diversas propuestas para elegir la base según las posibles ventajas computacionales. En particular se pueden elegir las funciones base de modo que en el primer intervalo B_1 es lineal con pendiente negativa y las demás B_i son constantes, mientras que en el último intervalo B_{M-1} es lineal con pendiente positiva y las demás B_i son constantes.
- Consideremos la siguiente modelización para la función de densidad,

$$\log f_{X|\theta}(x\ /\theta) = \theta_1 B_1(x) + \ldots + \theta_{M-1} B_{M-1}(x) - c(\theta)$$

donde

$$exp\{c(\theta)\} = \int_L^U exp\{\theta_1 B_1(x) + \ldots + \theta_{M-1} B_{M-1}(x)\}\, dx$$

- Para que se pueda modelizar una densidad, $c(\theta)$ debe ser finito, que se asegura si ocurre

$$i)\ L > -\infty \text{ o } \theta_1 < 0, \text{ y } ii)\ U < \infty \text{ o } \theta_{M-1} < 0$$

- Se tiene la siguiente log-verosimilitud:

$$l(\theta\ /\ X_1, \ldots, X_n) = \sum_{i=1}^{n} \log f_{X|\theta}(X_i\ /\ \theta)$$

- Su forma cóncava garantiza un e.m.v. único sujeto a la restricción de que $c(\theta)$ debe ser finito.
- El **e.m.v.** depende del número y localización de los nodos, por lo que se han desarrollado procedimientos para seleccionar la ubicación y el número de nodos, como el que se describe seguidamente, implementado en la **librería "logsplines"** de R.
- La idea es colocar nudos en el mínimo y máximo $(X(1), \ldots, X(n))$ y otras posiciones $r_2, \ldots, r_{M-1}$ de la muestra ordenada, de forma que las posiciones mantengan cierta simetría en torno a la mediana, si bien no tienen porqué estar igualmente espaciadas.

- En cuanto al número M de nudos, se sigue el siguiente proceso:
  - Se comienza con un número inicial de nodos que depende de n.
  - A continuación, se añade en cada iteración un nuevo nodo elegido en aquella posición donde se maximiza el estadístico de Rao, hasta que se alcanza un valor máximo admisible para el número de nudos.
  - A partir del último modelo de la fase anterior, se va eliminando un nudo en cada iteración, el que corresponde al menor valor del *estadístico de Wald*, hasta llegar a un modelo con un número prefijado de nudos.
  - Para cada uno de estos modelos, se calcula un criterio de bondad de ajuste, por ejemplo el Criterio de Información de Akaike (AIC) o el Criterio de Información Bayesiano (BIC) que dependen de la log-verosimilitud y el número de nudos, y **se elige el modelo con menor criterio**.
  - Se define así un procedimiento secuencial para determinar el número de nudos, que a su vez determina una base, proporcionando por tanto una **estimación de** $\log f(x)$ y finalmente la propia $f(x)$.