Selección del parámetro de suavizado en estimación noparamétrica de la densidad y de la regresión

María Leyenda Rodríguez[1] y Silvia Suárez Crespo[1]

[1]Departamento de Estadística e Investigación Operativa
Universidad de Santiago de Compostela
España

- El objetivo de esta sección será comprobar el funcionamiento de la selección del parámetro de suavizado en la estimación núcleo de la densidad univariante utilizando el método de validación cruzada. Para ello se comparará con el método de escala Normal.
- Como modelos de prueba se considerarán las 15 densidades descritas en (Marron and Wand, 1992), denotándolas por #1 hasta #15, extrayéndose 500 muestras de cada densidad, cada una de tamaño 100. Para cada muestra:
  - Se calculará la ventana de validación cruzada $\hat{h}_{\text{CV}}$ y la ventana Normal $\hat{h}_{\text{NS}}$.
  - Se calcularán los estimadores núcleo de la densidad $\hat{f}_{\hat{h}_{\text{CV}}}(\cdot)$ y $\hat{f}_{\hat{h}_{\text{NS}}}(\cdot)$.
  - Se calculará el criterio de error $\text{ISE}(h) = \int (\hat{f}_h(x) - f(x))^2 \, dx$ para los dos estimadores $\hat{f}_{\hat{h}_{\text{CV}}}(\cdot)$ y $\hat{f}_{\hat{h}_{\text{NS}}}(\cdot)$, siendo $f(\cdot)$ la función de desidad teórica de la densidad correspondiente.

- Denotando por $X_1, \ldots, X_n$ a cualquiera de las muestras aleatorias simples consideradas proveniente de una densidad $f(\cdot)$, la ventana de validación cruzada $\hat{h}_{\text{CV}}$ se obtiene minimizando la expresión:

$$\begin{aligned} CV(h) &= \frac{1}{n^2 h} \sum_{i,j} K \star K \left(\frac{X_i - X_j}{h} \right) \\ &- \frac{4}{n(n-1)h} \sum_{i=1}^{n} \sum_{j>i} K \left(\frac{X_i - X_j}{h} \right), \end{aligned}$$

donde $\star$ denota la operación convolución. Se utilizará $K(\cdot)$ un núcleo gaussiano, por los que $K \star K(\cdot)$ se corresponde con una $N(0, 2)$. La minimización puede ser problemática si existen mínimos locales.

- La ventana Normal viene dada por la expresión explícita:

$$\hat{h}_{\text{NS}} = 1{,}06\,\text{mín}\left(S, \frac{RIC}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)}\right) n^{-1/5},$$

donde S es la cuasidesviación típica muestral, RIC el rango intercuartílico estandarizado y Φ^{-1} la función cuantil de la $N(0,1)$.

- Se construirá el estimador núcleo de la densidad univariante $f(\cdot)$:

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

utilizando h igual a h_{CV} y h_{NS}, en cada caso.

- El criterio de error ISE(h) se calculará utilizando integración numérica, en concreto aplicando la Regla de Trapecios Compuesta (R.T.C), que aproxima cualquier integral $I(g) = \int_a^b g(x)\,dx$ por la siguiente expresión:

$$I(g) \sim \frac{h}{2}\left(g(x_0) + g(x_s)\right) + h\sum_{j=1}^{s-1} g(x_j),$$

siendo $x_0, x_1, \ldots, x_s$ una rejilla de puntos equidistantes del intervalo de integración y h la distancia entre dos puntos consecutivos.

El algoritmo principal a implementar es claro:

Algoritmo 1. Algoritmo principal:

1. Cálculo de $\hat{h}_{\text{CV}}$ y $\hat{h}_{\text{NS}}$.
2. Cálculo de $\hat{f}_{h_{\text{CV}}}(\cdot)$ y $\hat{f}_{h_{\text{NS}}}(\cdot)$.
3. Cálculo de $\text{ISE}(h_{\text{CV}})$ y $\text{ISE}(h_{\text{NS}})$.

Es necesaria la minimización de $CV(h)$ para la obtención de la ventana h_{CV}. Dado un intervalo de minimización I, se pueden considerar dos algoritmos posibles:

Algoritmo 2. Algoritmo de minimización de $CV(h)$ (1):

1. Dividir el intervalo I en m subintervalos $I_1, \ldots, I_m$ más pequeños.
2. Efectuar la minimización respecto de h de la expresión $CV(h)$ en cada uno de los subintervalos elaborados en el paso 1 y escoger el mínimo.

Algoritmo 3. Algoritmo de minimización de $CV(h)$ (2):

1. Construír una rejilla "suficientemente fina" que cubra el intervalo I (es decir, una secuencia de ventanas $h_1, \ldots, h_k$) y evaluar la $CV(h)$ en cada uno de las ventanas.
2. Escoger el mínimo de los valores obtenenidos en en apartado 1.

Dado un intervalo de integración $J = [a, b]$, $h = \hat{h}_{\text{CV}}$ o $h = \hat{h}_{\text{NS}}$ según corresponda y utilizando integración numérica con R.T.C:

Algoritmo 4. Algoritmo para la construcción de ISE(h):

1. Dividir el intervalo J en s subintervalos, $J_1 = [a = x_0, x_1]$, $J_2 = [x_1, x_2]$, ... $J_s = [x_{s-1}, x_s = b]$.
2. Evaluar las funciones $\hat{f}_h(\cdot)$ y $f(\cdot)$ en los extremos de los intervalos obtenidos en el paso 1.
3. Calcular los valores $(\hat{f}_h(z) - f(z))^2$ para $z = x_0, x_1, \ldots, x_s$.
4. Aproximar $I(\hat{f}_h - f)$ utilizando los valores del paso 3.

- Se deberá escoger entre el algoritmo 2 y el algoritmo 3. Para ello se obtendrán las curvas CV(h) para las 5 primeras muestras de cada densidad y se escogerá la densidad que presente gráficas más "irregulares". Se le aplicarán a las 30 primeras muestras de dicha densidad los dos algoritmos y se compararán los resultados. Para la representación se utiliza una rejilla de extremos 0 y 4 y paso `0.01`.

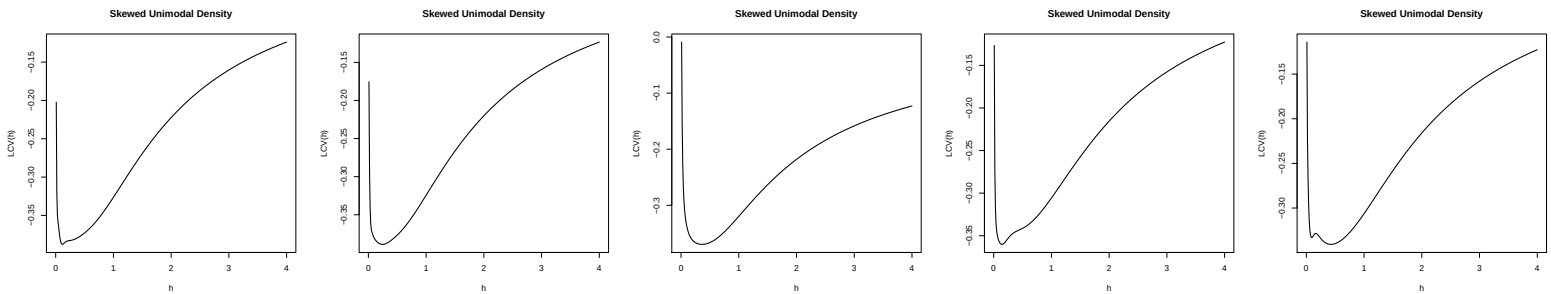

Figura: Representaciones de la curva CV(h) para las cinco primeras simulaciones de la densidad #2.

Muestra	Algoritmo 3	Algoritmo 2 (1)	Algoritmo 2 (2)
1	0.35	0.3511	0.3511
2	0.30	0.2983	0.2983
3	0.39	0.3920	0.3920
4	0.38	0.3839	0.3839
5	0.30	0.2989	0.2989
6	0.29	0.2942	0.2942
7	0.33	0.3251	0.3251
8	0.30	0.2989	0.2989
9	0.31	0.3145	0.3145
10	0.36	0.3567	0.3567

Cuadro: Ventanas de validación cruzada obtenidas al aplicar el algoritmo 3 sobre la rejilla de representación, el algoritmo 2 con 8 subintervalos del intervalo $[0, 4]$ (1) y el algoritmo 2 con 4 subintervalos (2). PARTE I

Muestra	Algoritmo 3	Algoritmo 2 (1)	Algoritmo 2 (2)
11	0.35	0.3506	0.3506
12	0.44	0.4361	0.4361
13	0.33	0.3254	0.3254
14	0.29	0.2927	0.2927
15	0.44	0.4430	0.4430
16	0.32	0.3175	0.3175
17	0.31	0.3131	0.3131
18	0.28	0.2788	0.2788
19	0.37	0.3665	0.3665
20	0.38	0.3798	0.3798

Cuadro: Ventanas de validación cruzada obtenidas al aplicar el algoritmo 3 sobre la rejilla de representación, el algoritmo 2 con 8 subintervalos del intervalo $[0, 4]$ (1) y el algoritmo 2 con 4 subintervalos (2). PARTE II

Muestra	Algoritmo 3	Algoritmo 2 (1)	Algoritmo 2 (2)
21	0.37	0.3653	0.3653
22	0.24	0.2439	0.2439
23	0.41	0.4134	0.4134
24	0.34	0.3407	0.3407
25	0.30	0.3029	0.3029
26	0.27	0.2737	0.2737
27	0.23	0.2292	0.2292
28	0.27	0.2716	0.2716
29	0.31	0.3056	0.3056
30	0.21	0.2145	0.2145

Cuadro: Ventanas de validación cruzada obtenidas al aplicar el algoritmo 3 sobre la rejilla de representación, el algoritmo 2 con 8 subintervalos del intervalo $[0, 4]$ (1) y el algoritmo 2 con 4 subintervalos (2). PARTE III

- Teniendo en cuenta que el algoritmo 2 es más eficiente computacionalmente que el algoritmo 3 y que tan sólo hemos tenido en cuenta las 30 primeras muestras de una densidad en particular, utilizaremos el algoritmo 2 con 8 subintervalos.
- El intervalo escogido para la integración numérica es $J = [-4, 4]$, dado que todos los valores de las muestras generadas para cualquiera de las densidades se encuentran dentro de dicho intervalo.

- **Representación gráfica de las estimaciones núcleo $\hat{f}_h(\cdot)$ de la función de densidad $f(\cdot)$ construídas con las ventanas $\hat{h}_{\mathsf{CV}}$ y $\hat{h}_{\mathsf{NS}}$.** Se realiza para las 5 primeras muestras de cada densidad y se incluirá aquí la gráfica de una de ellas.

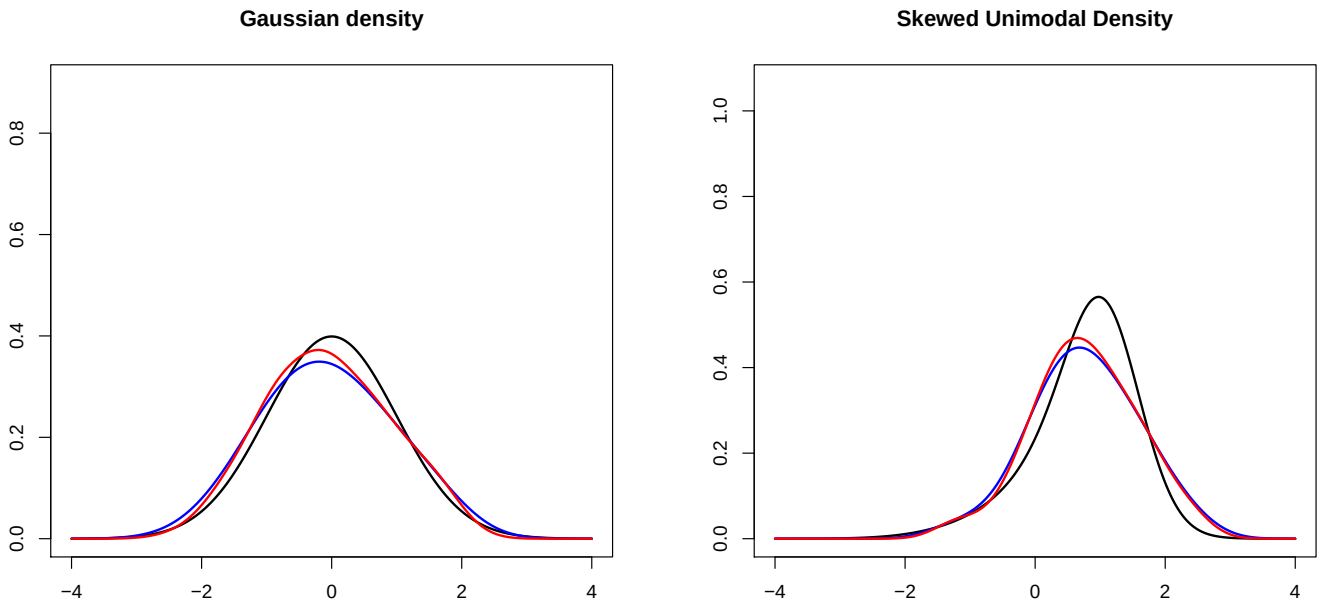

Figura: De izquierda a derecha, curva de densidad teórica (línea negra) acompañada de las estimaciones núcleo con ventanas $\hat{h}_{\mathsf{CV}}$ (línea azul) y $\hat{h}_{\mathsf{NS}}$ (línea roja) para #1 y #2, respectivamente.

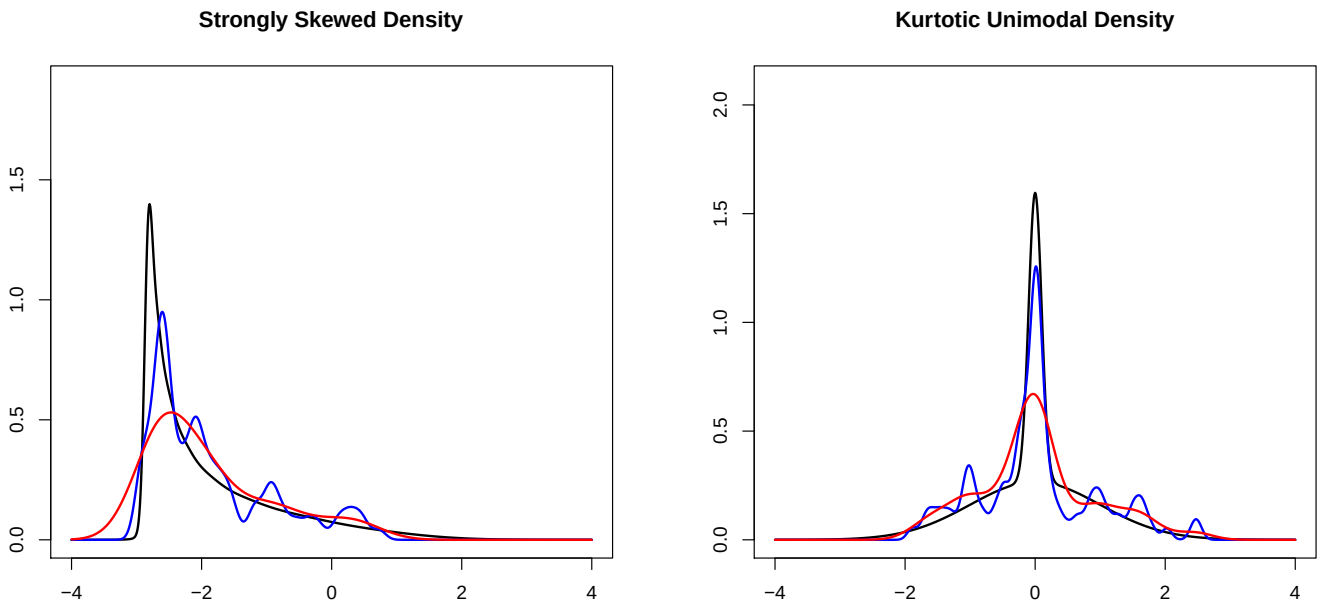

Figura: De izquierda a derecha, curva de densidad teórica (línea negra) acompañada de las estimaciones núcleo con ventanas $\hat{h}_{\text{CV}}$ (línea azul) y $\hat{h}_{\text{NS}}$ (línea roja) para #3 y #4, respectivamente.

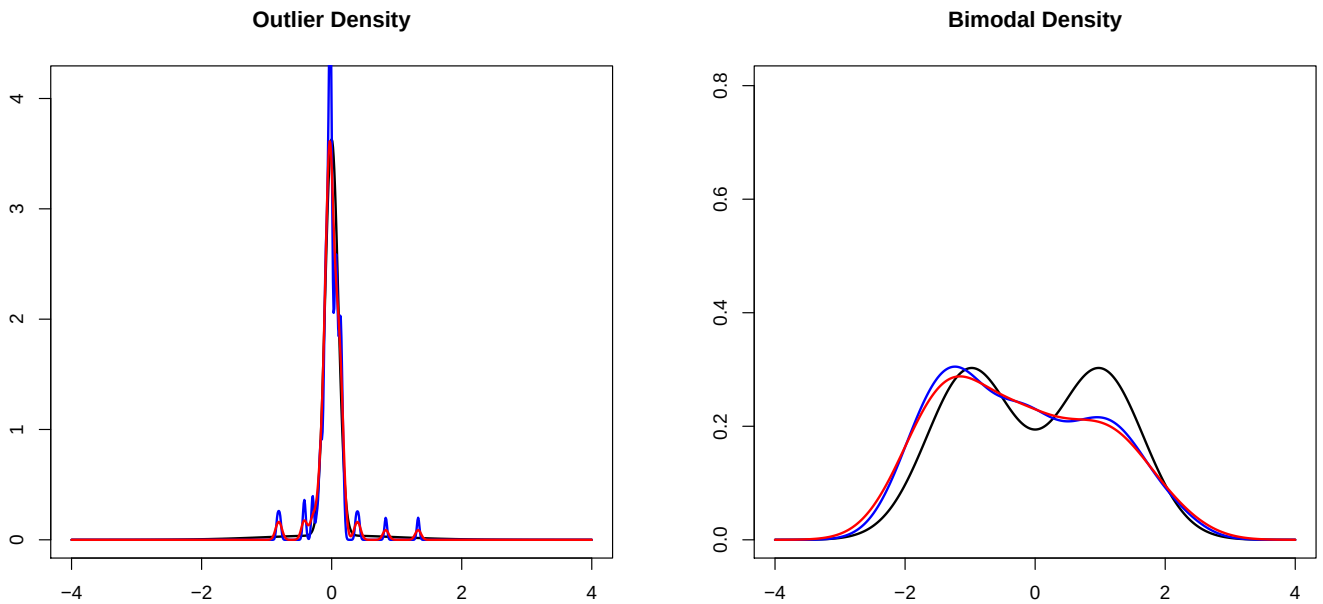

Figura: De izquierda a derecha, curva de densidad teórica (línea negra) acompañada de las estimaciones núcleo con ventanas $\hat{h}_{\text{CV}}$ (línea azul) y $\hat{h}_{\text{NS}}$ (línea roja) para #5 y #6, respectivamente.

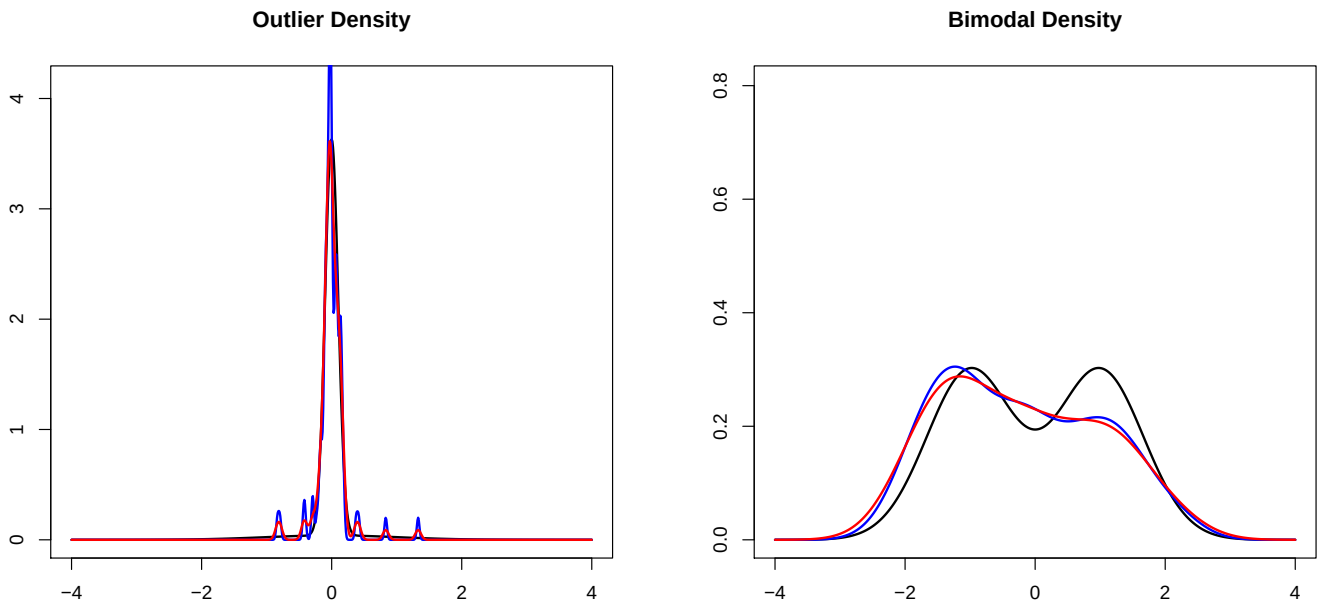

Figura: De izquierda a derecha, curva de densidad teórica (línea negra) acompañada de las estimaciones núcleo con ventanas $\hat{h}_{\text{CV}}$ (línea azul) y $\hat{h}_{\text{NS}}$ (línea roja) para #5 y #6, respectivamente.

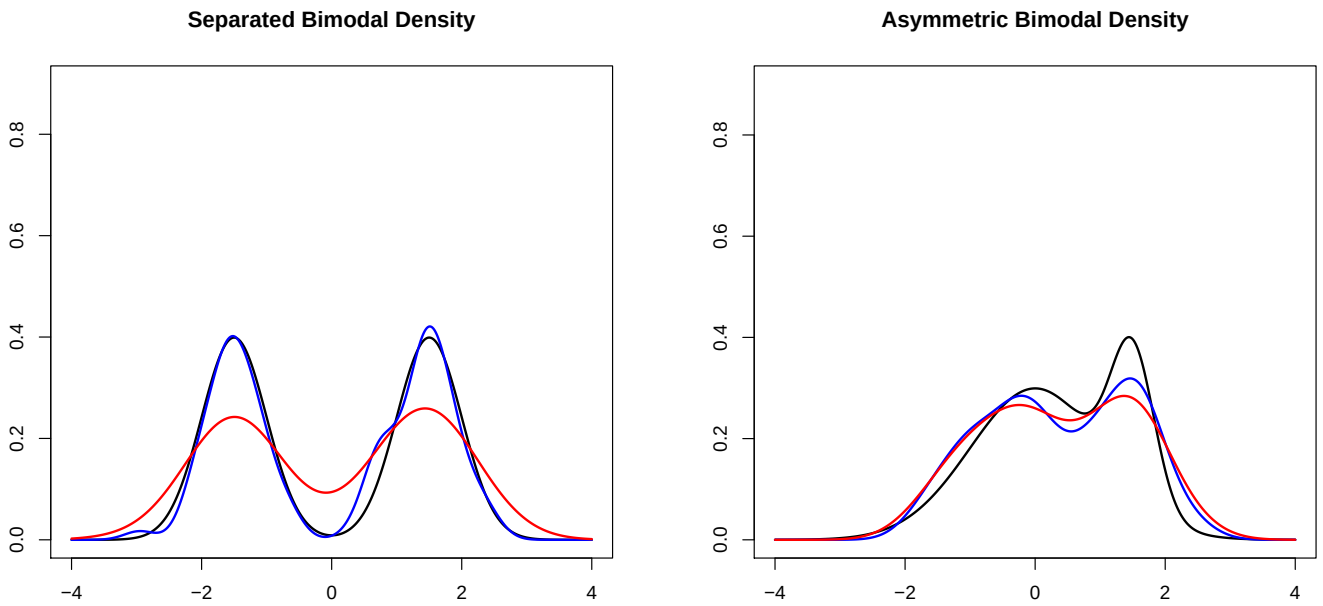

Figura: De izquierda a derecha, curva de densidad teórica (línea negra) acompañada de las estimaciones núcleo con ventanas $\hat{h}_{\mathrm{CV}}$ (línea azul) y $\hat{h}_{\mathrm{NS}}$ (línea roja) para #7 y #8, respectivamente.

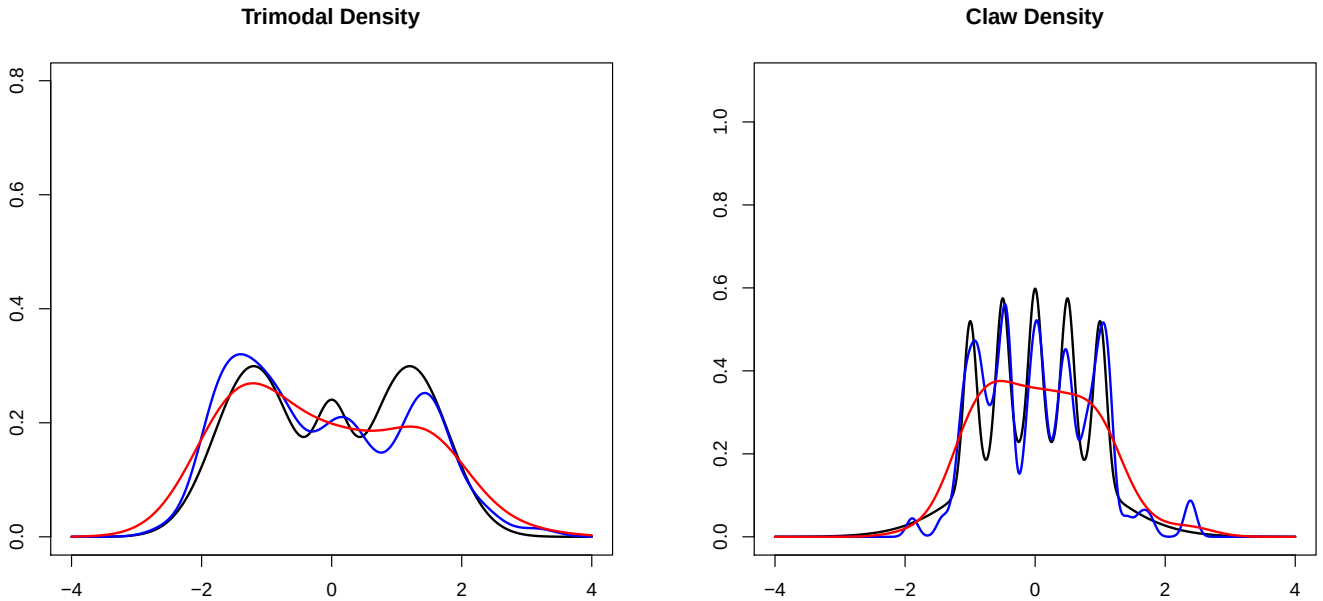

Figura: De izquierda a derecha, curva de densidad teórica (línea negra) acompañada de las estimaciones núcleo con ventanas $\hat{h}_{\text{CV}}$ (línea azul) y $\hat{h}_{\text{NS}}$ (línea roja) para #9 y #10, respectivamente.

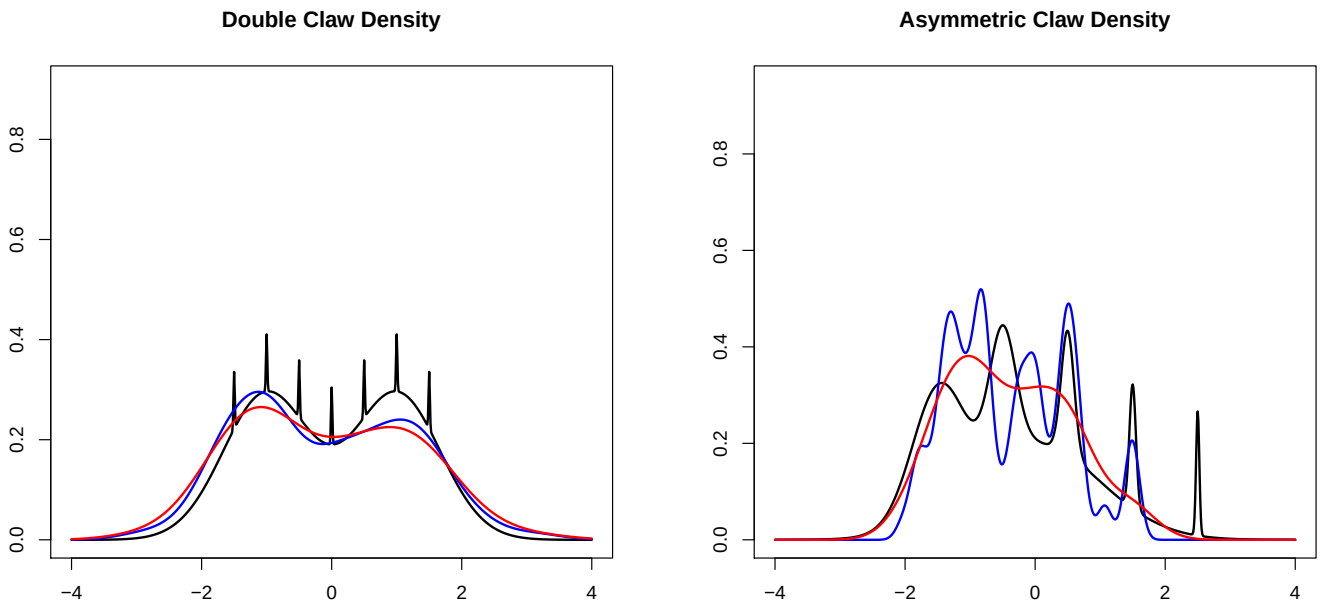

Figura: De izquierda a derecha, curva de densidad teórica (línea negra) acompañada de las estimaciones núcleo con ventanas $\hat{h}_{\mathrm{CV}}$ (línea azul) y $\hat{h}_{\mathrm{NS}}$ (línea roja) para #11 y #12, respectivamente.

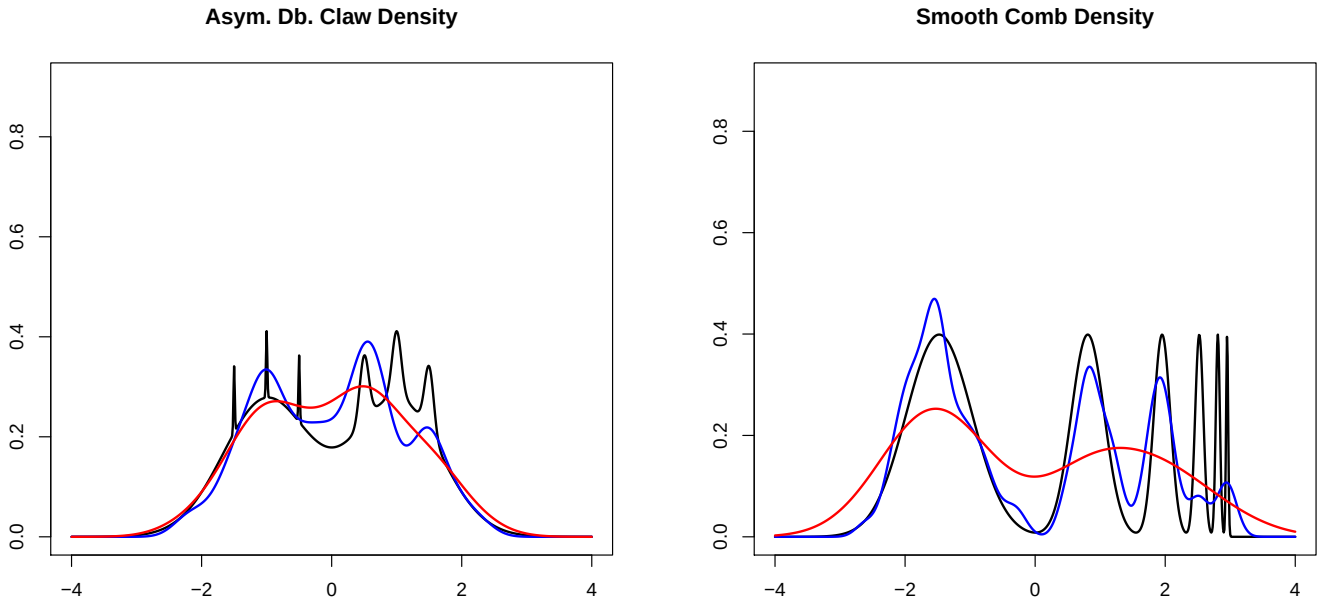

Figura: De izquierda a derecha, curva de densidad teórica (línea negra) acompañada de las estimaciones núcleo con ventanas $\hat{h}_{\mathrm{CV}}$ (línea azul) y $\hat{h}_{\mathrm{NS}}$ (línea roja) para #13 y #14, respectivamente.

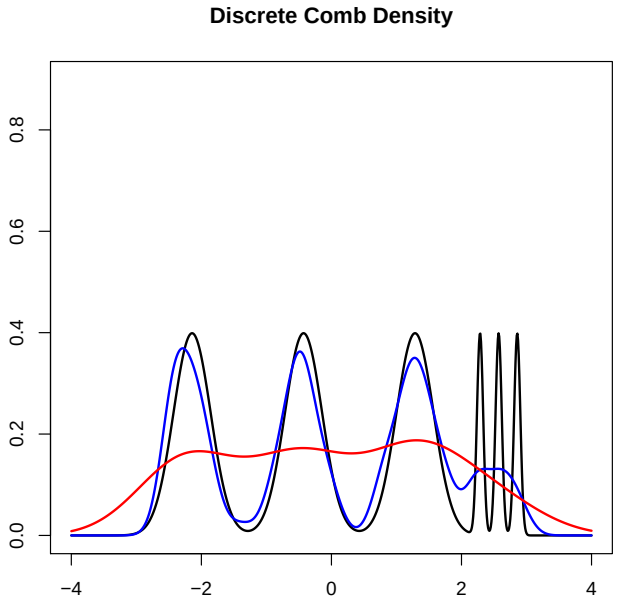

Figura: Curva de densidad teórica (línea negra) acompañada de las estimaciones núcleo con ventanas $\hat{h}_{\mathrm{CV}}$ (línea azul) y $\hat{h}_{\mathrm{NS}}$ (línea roja) para #15.

- **Resumen del porcentaje de veces que la ventana normal $\hat{h}_{\mathsf{NS}}$ ofrece mejores resultados que la ventana de validacin cruzada $\hat{h}_{\mathsf{CV}}$ (es decir, el porcentaje de veces asociado a que $\mathrm{ISE}(\hat{h}_{\mathrm{NS}})$ es menor que $\mathrm{ISE}(\hat{h}_{\mathrm{CV}})$), para cada densidad.**

Densidad	#1	#2	#3	#4	#5	#6
Porcentaje	74.6 %	71.6 %	0.0 %	1.4 %	68.4 %	29.2 %
Densidad	#7	#8	#9	#10	#11	#12
Porcentaje	0.0 %	23.0 %	13.6 %	23.0 %	24.6 %	22.6 %
Densidad	#13	#14	#15			
Porcentaje	19.0 %	0.0 %	0.0 %			

Cuadro: Porcentaje de muestras en las que $\mathrm{ISE}(\hat{h}_{\mathrm{NS}})$ es menor que $\mathrm{ISE}(\hat{h}_{\mathrm{CV}})$, para cada una de las densidades.

- **Resumen de estadísticos descriptivos de las ventanas $\hat{h}_{\mathbf{CV}}$ y $\hat{h}_{\mathbf{NS}}$ y de los valores de las expresiones** $\mathrm{ISE}(\hat{h}_{\mathrm{CV}})$ **y** $\mathrm{ISE}(\hat{h}_{\mathrm{NS}})$ **asociadas, para cada densidad**.
  - Se observan diferencias entre las desviaciones típicas asociadas a las ventanas Normales y las asociadas a las ventanas de validación cruzada: la variabilidad de las primeras es pequeña y prácticamente constante a lo largo de las densidades, mientras que el rango de variabilidad de las segundas no es independiente de las densidades y es en general mayor que el alcanzado para las ventanas Normales.
  - Si nos fijamos en los valores medios de ambas ventanas se tiene que en general, las ventanas Normales son más grandes que las ventanas de validación cruzada (excepto para las densidades #1 y #2).
  - Se puede observar como, en media, el error cuadrático integrado asociado a la ventana de validación cruzada en general es más pequeño que el asociado a la ventana Normal (excepto para las densidades #1, #2 y#5).

	Dens.	Mín.	Media	Med.	Máx.	Sd
h_{CV}	#1	0.0404	0.4394	0.4696	0.6681	0.1308
	#2	0.0241	0.3123	0.3271	0.5160	0.0826
	#3	0.0138	0.0903	0.0888	0.1965	0.0310
	#4	0.0160	0.0834	0.0825	0.2242	0.0255
	#5	0.0071	0.0464	0.0485	0.0729	0.0122
	#6	0.0712	0.4122	0.4149	0.7393	0.1360
	#7	0.0662	0.2625	0.2738	0.3872	0.0607
	#8	0.0528	0.3418	0.3313	0.6521	0.1274
	#9	0.0584	0.3717	0.3681	0.7674	0.1302
	#10	0.0299	0.1970	0.1098	0.5446	0.1453
	#11	0.0840	0.4029	0.3976	0.7780	0.1336
	#12	0.0475	0.2888	0.2144	0.7201	0.1706
	#13	0.0734	0.3827	0.3833	0.6958	0.1310
	#14	0.0319	0.1458	0.1455	0.2758	0.0486
	#15	0.0307	0.1523	0.1663	0.2486	0.0474

Cuadro: Estadísticos descriptivos de $\hat{h}_{CV}$.

	Dens.	Mín.	Media	Med.	Máx.	Sd
h_{NS}	#1	0.2707	0.4049	0.4100	0.5022	0.0393
	#2	0.2156	0.3055	0.3045	0.4109	0.0353
	#3	0.2087	0.3805	0.3797	0.5413	0.0609
	#4	0.0831	0.2016	0.2001	0.3590	0.0513
	#5	0.0286	0.0469	0.0465	0.0644	0.0055
	#6	0.4281	0.5067	0.5059	0.6018	0.0267
	#7	0.6072	0.6672	0.6672	0.7412	0.0214
	#8	0.3598	0.4621	0.4627	0.5432	0.0282
	#9	0.4538	0.5375	0.5372	0.6196	0.0263
	#10	0.2676	0.3591	0.3583	0.4378	0.0275
	#11	0.4296	0.5049	0.5060	0.5892	0.0261
	#12	0.3547	0.4658	0.4662	0.5530	0.0322
	#13	0.4143	0.5011	0.5015	0.5971	0.0254
	#14	0.6199	0.6918	0.6911	0.7697	0.0283
	#15	0.6067	0.7110	0.7126	0.8032	0.0345

Cuadro: Estadísticos descriptivos de $\hat{h}_{NS}$.

	Dens.	Mín.	Media	Med.	Máx.	Sd
$\mathrm{ISE}(\hat{h}_{\mathrm{CV}})$	#1	0.0014	0.0696	0.0573	0.2888	0.0493
	#2	0.0035	0.0693	0.0542	0.3832	0.0550
	#3	0.0269	0.0967	0.0771	0.4619	0.0605
	#4	0.0214	0.0895	0.0702	0.8180	0.0693
	#5	0.0063	0.0667	0.0567	0.3427	0.0458
	#6	0.0083	0.0771	0.0653	0.2723	0.0486
	#7	0.0064	0.0683	0.0607	0.2297	0.0374
	#8	0.0077	0.0851	0.0735	0.3158	0.0516
	#9	0.0090	0.0830	0.0718	0.3141	0.0506
	#10	0.0187	0.1974	0.0989	0.6998	0.1712
	#11	0.0361	0.1172	0.1014	0.4039	0.0604
	#12	0.0297	0.1570	0.1059	0.5471	0.1054
	#13	0.0439	0.1276	0.1047	0.3945	0.0673
	#14	0.0465	0.1217	0.1099	0.3801	0.0495
	#15	0.0510	0.1191	0.1137	0.2827	0.0377

Cuadro: Estadísticos descriptivos de $\mathrm{ISE}(\hat{h}_{\mathrm{CV}})$.

	Dens.	Mín.	Media	Med.	Máx.	Sd
$\mathrm{ISE}(\hat{h}_{\mathrm{NS}})$	#1	0.0032	0.0524	0.0449	0.2699	0.0372
	#2	0.0040	0.0575	0.0454	0.2600	0.0438
	#3	0.3529	1.1133	1.0904	2.1936	0.3626
	#4	0.0472	0.4505	0.3992	1.4928	0.2818
	#5	0.0078	0.0578	0.0461	0.2743	0.0412
	#6	0.0144	0.0849	0.0814	0.2261	0.0373
	#7	0.4253	0.6192	0.6161	0.9169	0.0834
	#8	0.0297	0.1084	0.1022	0.2909	0.0406
	#9	0.0391	0.1180	0.1129	0.2688	0.0387
	#10	0.2933	0.3799	0.3697	0.5613	0.0483
	#11	0.0484	0.1375	0.1321	0.3726	0.0433
	#12	0.1863	0.2594	0.2518	0.4770	0.0413
	#13	0.0716	0.1601	0.1545	0.3948	0.0453
	#14	1.0385	1.2583	1.2525	1.6337	0.1017
	#15	1.3216	1.6354	1.6320	1.9609	0.1077

Cuadro: Estadísticos descriptivos de $\mathrm{ISE}(\hat{h}_{\mathrm{NS}})$.

- **Gráficos boxplot comparativos entre las dos ventanas en estudio ($\hat{h}_{\text{CV}}$ y $\hat{h}_{\text{NS}}$) y de los errores cuadrticos integrados asociados a cada una de ellas (ISE$(\hat{h}_{\text{CV}})$ y ISE$(\hat{h}_{\text{NS}})$), para cada densidad**.
  - Estos gráficos generalizan los resultados que acabamos de comentar en el análisis descriptivo.
  - Cabe destacar el caso de la densidad Normal: aún cuando la ventana Normal proporciona buenas aproximaciones de una forma fácil y sencilla, el proceso de selección de ventana por el método de validación cruzada no acaba de definirse por un intervalo de ventanas concreto, obteniéndose así una gran cantidad de *outliers* en el gráfico.
  - Llama también la atención los box-plots de los criterios de error asociados a las densidades #14 y #15, en las que se ve claramente que el error cuadrático integrado asociado a la ventana de validación cruzada es mucho menor que el asociado a la ventana Normal.

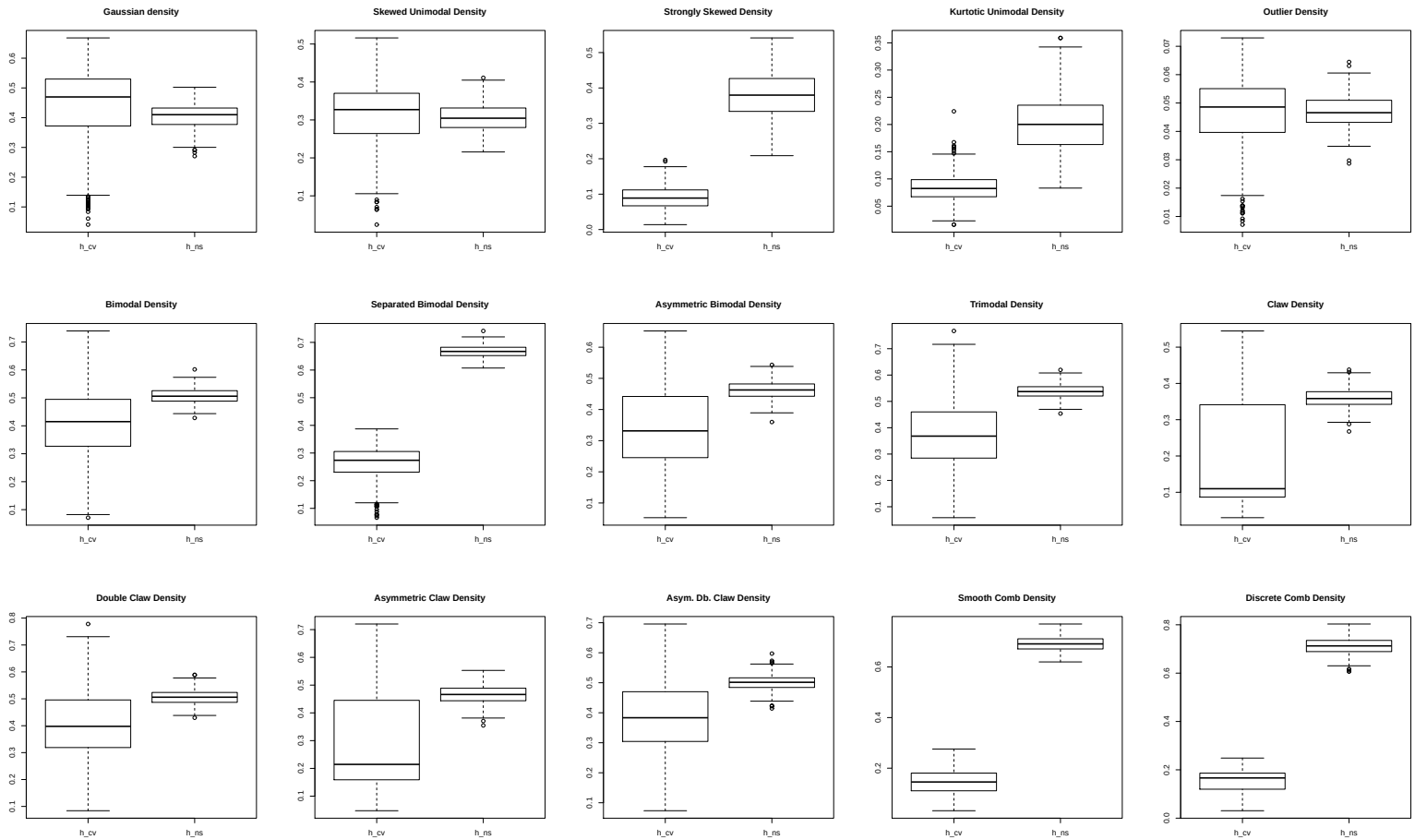

Figura: De izquierda a derecha, y de arriba a abajo, gráficos boxplot comparativos asociados a las ventanas $\hat{h}_{\mathrm{CV}}$ y $\hat{h}_{\mathrm{NS}}$, de la densidad #1 hasta la #15.

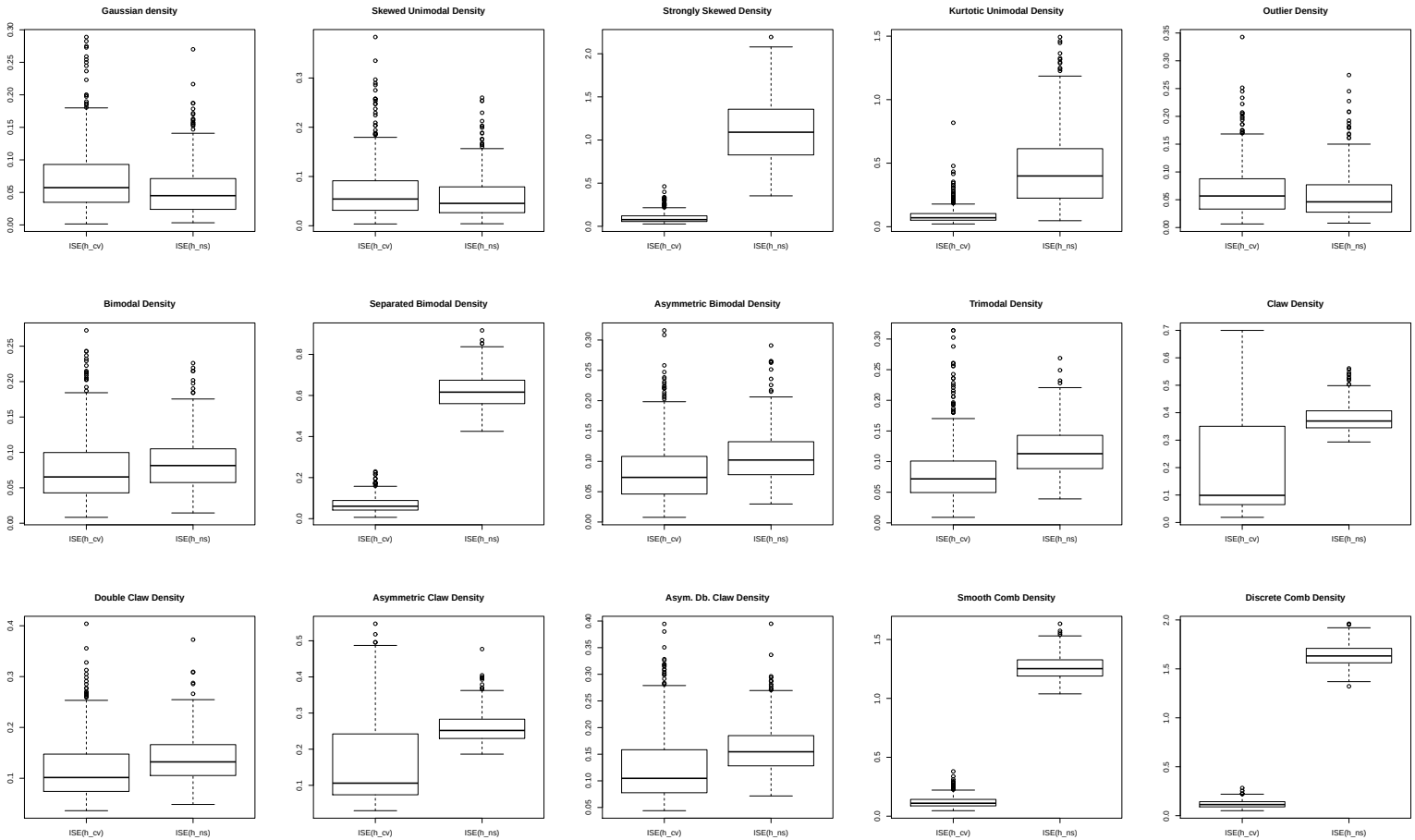

Figura: De izquierda a derecha, y de arriba a abajo, gráficos boxplot comparativos asociados a las ventanas $\mathrm{ISE}(\hat{h}_{\mathrm{CV}})$ y $\mathrm{ISE}(\hat{h}_{\mathrm{NS}})$, de la densidad #1 hasta la #15.

Conclusiones

- Los selectores de ventana de escala Normal proporcionan una primera aproximación de forma fácil y rápida, y cabe esperar resultados razonables cuando la muestra de la que se dispone está próxima a la de una Normal. Para este tipo de densidades funciona mejor la ventana Normal, pues la ventana de validación cruzada tiende a infrasuavizar en algunos casos.
- Para densidades unimodales fuertemente asimétricas, la ventana Normal tiende a sobresuavizar. En estos casos se prefiere la ventana de validación cruzada.

Conclusiones

- Para densidades unimodales leptocúrticas (exceptuando los casos extremos como la densidad #5), se prefiere la ventana de validación cruzada. Sin embargo, para los casos extemos no está claro cuál escoger pues la ventana de validación cruzada en general se aproxima más a la moda pero a base de incluír mucho ruído en las colas, de forma que para estas muestras funciona mejor la ventana Normal.
- Para densidades multimodales no se debe utilizar la ventana Normal, pues tiende a sobresuavizar en exceso y a enmascarar la existencia de modas. En este caso conviene utilizar la ventana de validación cruzada.

- El objetivo de esta sección será comprobar el funcionamiento de la selección del parámetro de suavizado en la estimación núcleo de la regresión, realizando en este caso el análisis sobre un conjunto de datos reales.
- El conjunto de datos a emplear será el *data frame* denominado `airquality`, que contiene diversas medidas de la calidad del aire en Nueva York entre mayo y septiembre de 1973.
- Se desea analizar la relación existe entre la temperatura (en grados F) y la concentración de ozono (en -*ppb*) en la ciudad de Nueva York, siendo la variable independiente la temperatura. Se realizará un análisis descriptivo de ellas.
- Se comparará el estimador lineal local construído con ventana de validación cruzada con el mismo estimador construído con ventana *plug-in* y con el estimador de Nadaraya-Watson construído con la ventana de validación cruzada asociada.
- Finalmente se propondrá un modelo alternativo basado en la regresión lineal simple.

La variable concentración de ozono posee datos faltantes. El análisis descriptivo (y la posterior construcción de los estimadores) se realizará eliminando los índices correspondientes a dichos datos de la variable ozono y también de la temperatura.

- **Estadísticos descriptivos**

Var.	Mín.	Media	Med.	Máx.	Sd
Ozono	1.00	42.13	31.50	168.00	32.98788
Temperatura	57.00	77.87	79.00	97.00	9.485486

Cuadro: Estadísticos descriptivos asociados a las variables ozono y temperatura.

- **Gráficos boxplot**

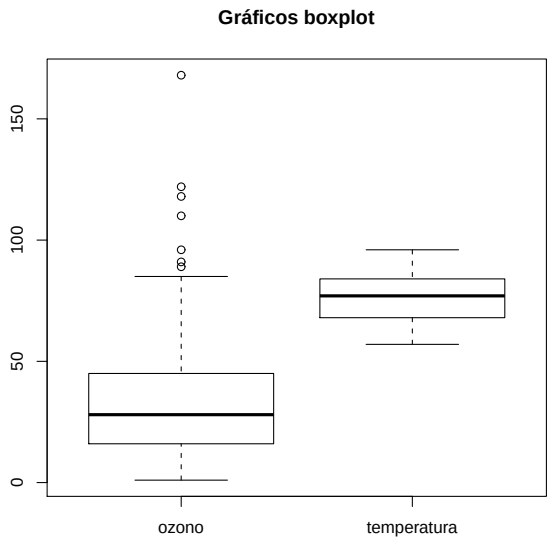

Figura: Diagrama de cajas del ozono y de la temperatura.

Se observa la existencia de datos atípicos en la concentración de ozono.

- **Distribución de las variables**

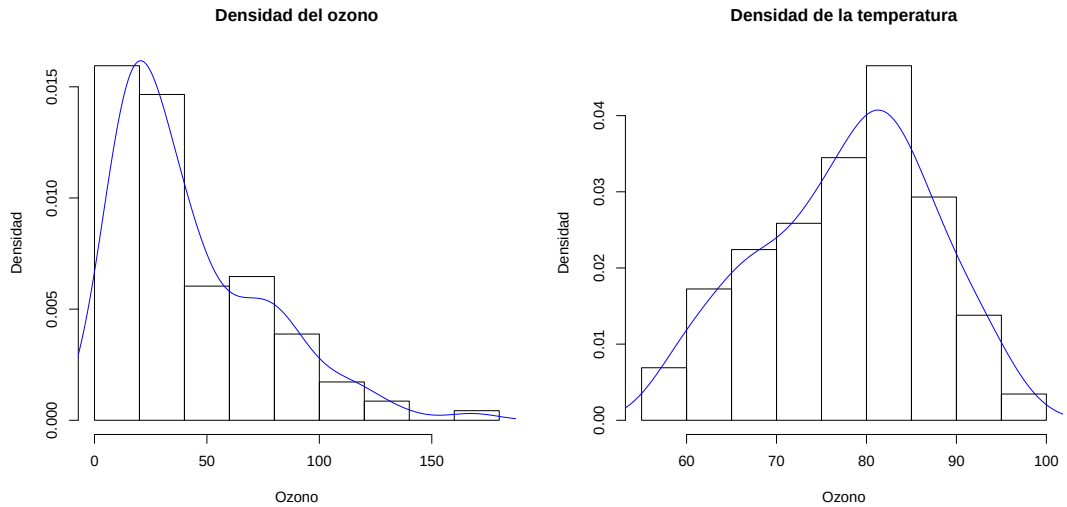

Figura: De izquiera a derecha, histogramas del ozono y de la temperatura, respectivamente, acompañados de las correspondientes estimaciones núcleo realizadas con la ventana de validación cruzada.

Realizando un test de normalidad de Shapiro-Wilks, se obtiene para el ozono un p-valor de 2.790e−08 y para la temperatura un p-valor de 0.0719. Por lo tanto se acepta normalidad para la temperatura con un nivel de confianza de 0.95.

- **Serie de tiempo de la temperatura**

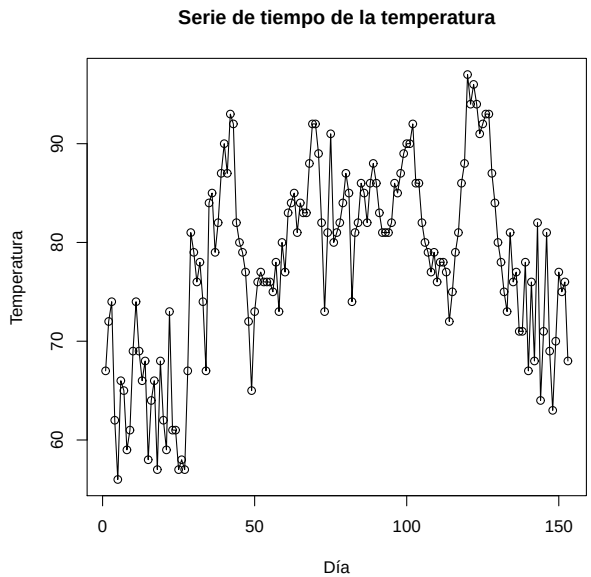

Figura: Serie de tiempo asociada a la variable temperatura, en el período en el que registran se las observaciones.

Notemos que en este caso hemos utilizado todas las observaciones (diarias) registradas para la temperatura.
Como es de esperar, las temperaturas son mayores en los meses centrales (julio y agosto) y más bajas en los restantes meses.

- **Gráfico de dispersión de la variable ozono sobre la temperatura**

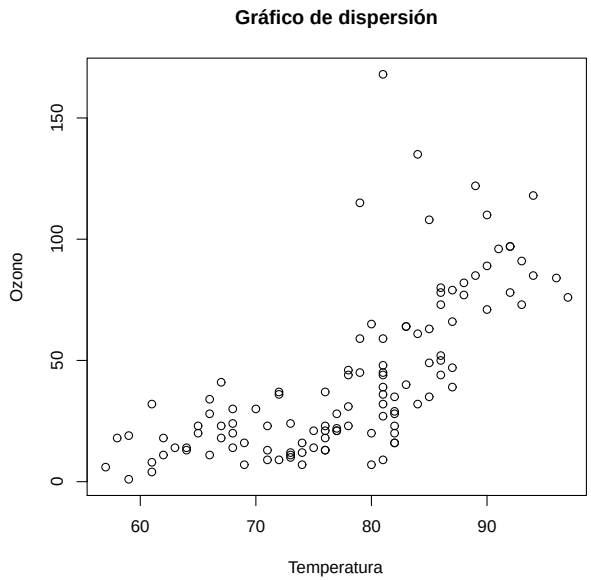

Figura: Gráfico de dispersión de la concentración de ozono con respecto a la temperatura.

La concentración de ozono se mantiene casi constante durante un rango de temperaturas, para luego aumentar de forma lineal con pendiente alta (existe relacíon).

- **Ajuste por regresión lineal simple**

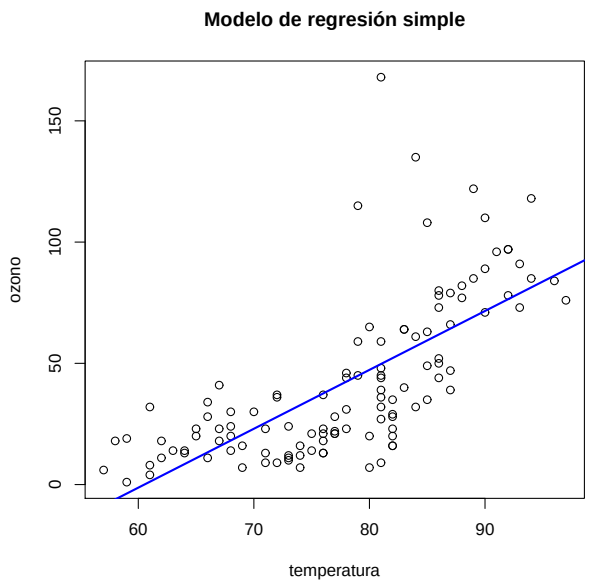

Figura: Ajuste obtenido mediante el método de regresión lineal simple.

Puede observarse que el ajuste no es bueno (además, el estadístico R^2 ajustado es 0.4832)

- **Diagramas de cajas comparativos (temperaturas menores que 75 y mayores que 75)**

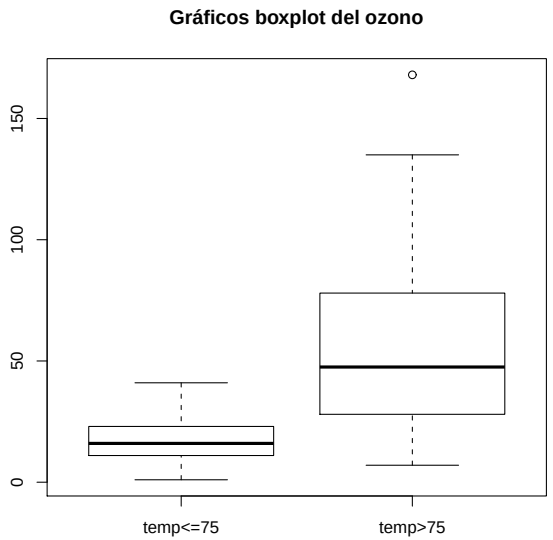

Figura: Diagramas de cajas de la concentración de ozono sobre dos rangos de temperaturas (menores o iguales a 75 y mayores que 75).

Se observa como los rangos entre los que varía la concentración de ozono son muy distintos.

- Considérese una muestra de una variable aleatoria bidimensional (X, Y), es decir, un conjunto de pares $(X_1, Y_1), \ldots, (X_n, Y_n)$ independientes e identicamente distribuídos a (X, Y), y denotemos $\mathbb{X} = (X_1, \ldots, X_n)$.
- Se necesitará escribir el estimador lineal local y el estimador de Nadaraya-Watson como estimadores lineales, es decir, en la forma

$$\hat{m}(x) = \sum_{j=1}^{h} l_j(x) Y_j.$$

- Así, para una rejilla de puntos $t_1, \ldots, t_m$, el valor del estimador en el punto t_i viene dado por:

$$\hat{m}(t_i) = \sum_{j=1}^{n} T_{ij} Y_j,$$

con $T_{ij} = l_j(t_i)$.

- En particular, si la rejilla de puntos coincide con los valores de $\mathbb{X}$, escribiremos $L_{ij} \equiv T_{ij}$.
- Por tanto, si denotamos $\hat{\mathbf{m}} = (\hat{m}(t_1), \ldots, \hat{m}(t_m))$, $\mathbf{Y}$ el vector de las observaciones $(Y_1, \ldots, Y_n)$ y $\mathbf{T} = (T_{ij})_{i=1}^{m},\ j = 1 \ldots, n$ ($\mathbf{L}$ en el caso de $\mathbb{X}$) la matriz de dimensiones $m \times n$ $(n \times n)$ cuyos elementos se definieron anteriormente, entonces

$$\hat{\mathbf{m}} = \mathbf{TY}.$$

- Se verifica además para los estimadores lineales que la función de validación cruzada viene dada por la expresión

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{m}(X_i)}{1 - L_{ii}} \right)^2,$$

donde L_{ii} es la matriz diagonal de la matriz $\mathbf{L}$ definida anteriormente. La idea es tomar aquel h que haga que CV sea mínimo.

- Dada una ventana h, se tiene que el estimador lineal local $\hat{m}_{\mathsf{LL}}(\cdot)$ se puede escribir en la forma

$$\hat{m}_{\mathsf{LL}}(t_i) = \sum_{j=1}^{n} T_{ij} Y_j,$$

donde

$$T_{ij} = \frac{b_j(t_i)}{\sum_{k=1}^{n} b_k(t_i)},$$

con

$$b_j(t_i) = K\left(\frac{X_j - t_i}{h}\right)(S_{n,2}(t_i) - (X_j - t_i)S_{n,1}(t_i))$$

y

$$S_{n,r}(t_i) = \sum_{j=1}^{n} K\left(\frac{X_j - t_i}{h}\right)(X_j - t_i)^r,\ r = 1, 2.$$

- Dada una ventana h, se tiene que el estimador de Nadaraya-Watson $\hat{m}_{\text{NW}}(\cdot)$ se puede escribir en la forma

$$\hat{m}_{\text{NW}}(t_i) = \sum_{j=1}^{n} T_{ij} Y_j,$$

donde

$$T_{ij} = \frac{b_j(t_i)}{\sum_{k=1}^{n} b_k(t_i)},$$

con

$$b_j(t_i) = K\left(\frac{X_j - t_i}{h}\right)$$

- Observamos por tanto que el estimador de Nadaraya-Watson es un caso particular del estimador lineal local.

Supongamos que disponemos de una rejilla $t_1, \ldots, t_m$. Dada una ventana h:

Algoritmo 1. Cálculo de la matriz $\mathbf{T} = (T_{ij})_{i=1}^{m},\ j = 1, \ldots, n$

1. Calcular las evaluaciones $\frac{X_j - t_i}{h}$ y $K\left(\frac{X_j - t_i}{h}\right)$, para cada $t_i,\ i = 1, \ldots, m$.
2. Calcular $\mathbf{T}$ (en el caso del estimador local se calcularán previamente las cantidades $b_j(t_i)$ y $S_{n,r}(t_i)$).

Notemos de que la función que implemente el algoritmo 1 nos servirá igualmente para calcular dicha matriz, pero evaluada en la muestra $\mathbb{X}$, es decir, para calcular la matriz $\mathbf{L} = (L_{ij})_{i,j=1}^{n}$, necesaria para la obtención de la ventana de validación cruzada).

Algoritmo 2. Cálculo del estimador (lineal local o de Nadaraya-Watson)

1. Cálculo de la matriz $\mathbf{T} = (T_{ij})_{i=1}^{m},\ j = 1, \ldots, n$ utilizando el algoritmo 1.
2. Resolución de $\hat{\mathbf{m}} = \mathbf{TY}$.

Dada una rejilla de valores $h = h_1, \ldots, h_s$ razonable:

Algoritmo 3. Cálculo de la ventana de validación cruzada

1. Cálculo de la matriz $\mathbf{L} = (L_{ij})_{i,j=1}^{n}$ utilizando el algoritmo 1, para cada $h_p,\ p = 1, \ldots, s$.
2. Cálculo de $L_{ii},\ i = 1, \ldots, n$ (diagonal de la matriz anterior), para cada $h_p,\ p = 1, \ldots, s$.
3. Cálculo de la función de validación cruzada $CV(h)$, para cada $h_p,\ p = 1, \ldots, s$.
4. Cálculo del mínimo de las expresiones del paso anterior y del valor de h asociado (ventana de validación cruzada).

- Se buscarán las ventanas de validación cruzada en la rejilla de extremos 0 y 5, con paso 0.01. Además, se evaluarán los estimadores tan sólo en los puntos de la muestra.
- La ventana de validación cruzada para el método lineal local el 1.59 mientras que para el estimador de Nadaraya-Watson se obtiene el valor 1.31.

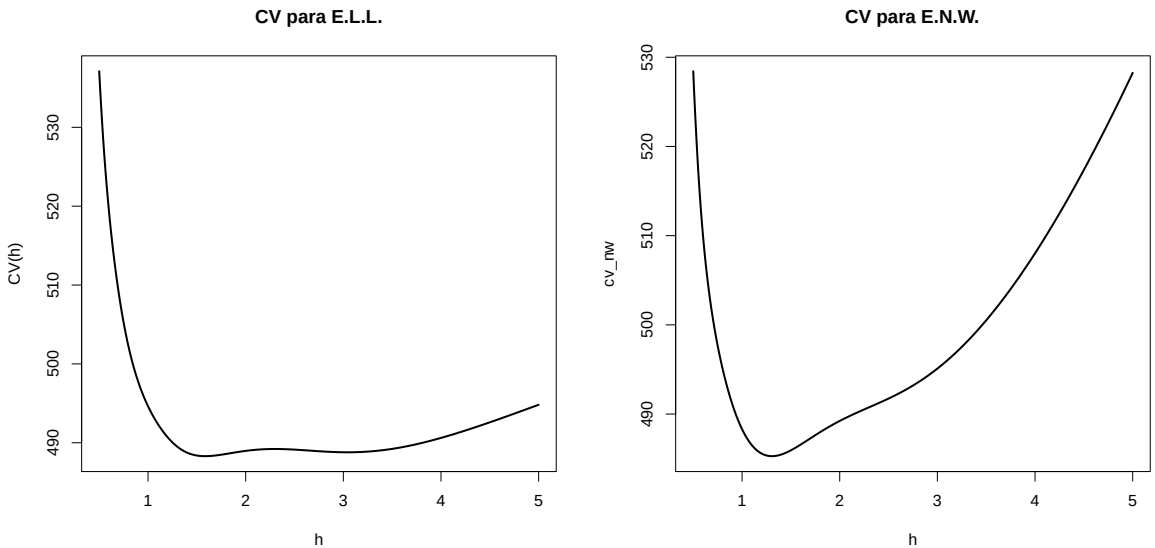

Figura: De izquierda a derecha, representación de la función de validación cruzada para el estimador lineal local y el estimador de Nadaraya - Watson, respectivamente.

- En la siguiente gráfica se presenta la comparación entre las estimaciones realizadas por uno y otro método. En este caso particular, los estimadores se comportan de forma similar, aunque se sabe que el estimador de Nadaraya-Watson no suele aproximar muy bien en la frontera (*edge effect*).

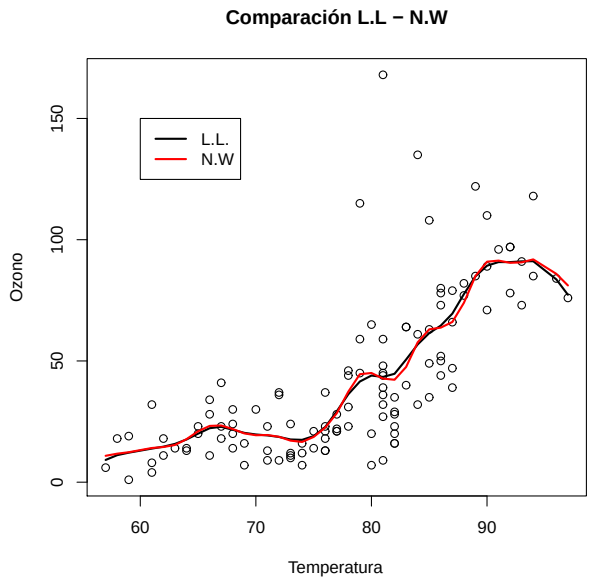

Figura: Estimaciones realizadas por el método lineal local (línea negra) y el método de Nadaraya-Watson (línea roja) utilizando sendas ventanas de validación cruzada.

- La ventana *plug-in* se calcula intrínsecamente en el paquete `R` mediante la función `dpill` del paquete `KernSmooth` (el valor de la ventana es `1.598009`). Ambas estimaciones se superponen en este caso, aunque en general la ventana *plug-in* suele ofrecer estimaciones más *suaves* que la otra.

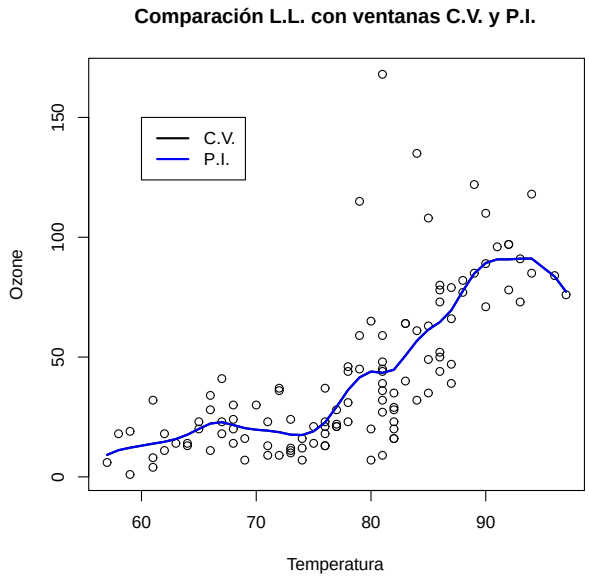

Figura: Estimaciones realizadas por el método lineal local utilizando la ventana de validación cruzada (línea negra) y la ventana *plug-in* (línea azul).

- En el análisis descriptivo habíamos detectado datos atípicos en la concentración de ozono. Podríamos pensar que dichos datos están influyendo en las estimaciones realizadas.
- Si realizamos el estudio eliminando dichos datos, varían evidentemente los valores concretos de las ventanas y de los estimadores, pero las conclusiones generales son exactamente las mismas que las obtenidas para el estudio realizado hasta ahora.

- A la vista de los resultados obtenidos en las anteriormente, se podría considerar un modelo sencillo para la descripción de los datos.
- Se observa que cuando la temperatura se encuentra entre los valores 0 y 75, los estimadores no paramétricos de la regresión se comportan de forma más o menos lineal con poca pendiente; entre los valores de temperatura 75 y 91 los estimadores se comportan de una forma más o menos lineal pero con mucha más pendiente; finalmente, cuando la temperatura toma valores entre 91 y 96 los estimadores presentan un comportamiento constante.
- En este último tramo cabe notar que hemos eliminado el último valor de la temperatura, 97, pues provoca un efecto frontera.

- Por todo esto, hemos probado a estimar la función de regresíon aplicando el modelo de regresíon lineal simple a trozos, cuya expresión viene dada por

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, \ldots, n,$$

donde los errores $\epsilon_i, i = 1, \ldots, n$ son variables aleatorias que siguen una distribución Normal $N(0, \sigma^2)$ y son mutuamente independentes.

- El ajuste obtenido se puede ver en la gráfica siguiente:

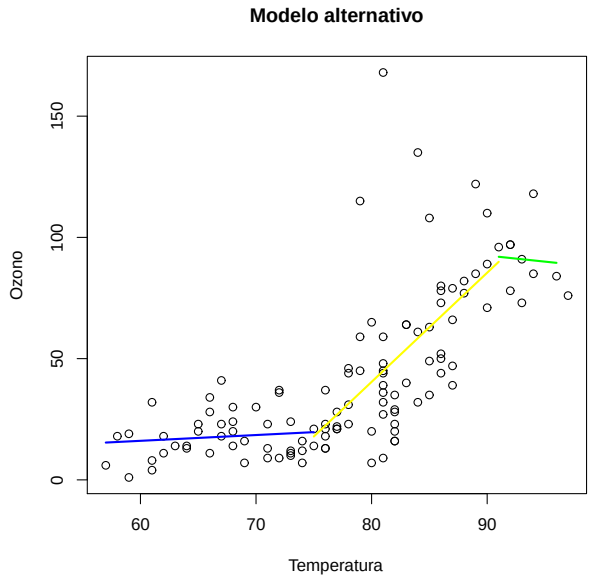

Figura: Estimación de la curva de regresión utilizando el modelo de regresión simple a trozos (en los tramos situados entre el mínimo de las temperaturas y 75, entre 75 y 91 y entre 91 y el máximo de las temperaturas).

- Para realizar un ajuste de mejor calidad, pero siguiendo el modelo secillo que hemos planteado, se podría utilizar el modelo de regresión spline con polinomios de grado 1 (en la figura se incluyen los ajustes también para grado 2 y 3). Se utiliza la función `bs` de la librería `splines`.

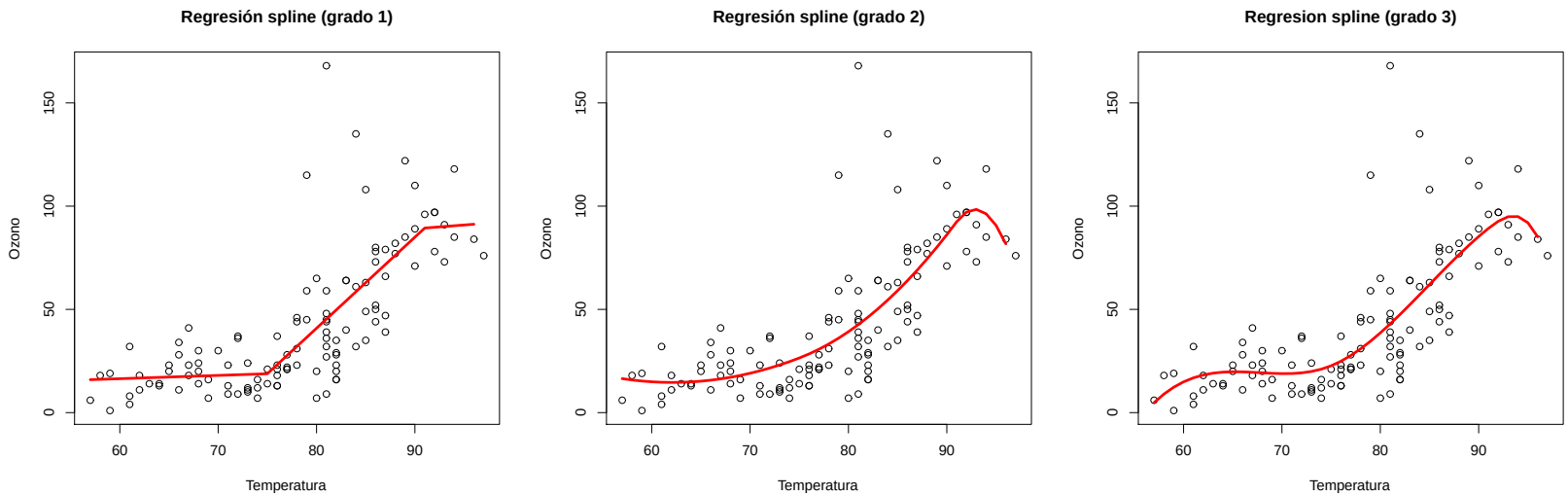

Figura: De izquierda a derecha, estimación de la curva de regresión utilizando el modelo de regresión spline con polinomios de grado uno, dos y tres, respectivamente.

Gracias por vuestra atención

María Leyenda Rodríguez
Silvia Suárez Crespo

 Fan, J. and Gijbels, I. (1996) *Local polynomial modelling and its applications*. Chapman and Hall, London.

 Marron, J. S. and Wand, M. P. (1992) Exact mean integrated squared error. *The Annals of Statistics*, **20**, 712-736.

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