LA FUNCIÓN DE VEROSIMILITUD SUAVIZADA EN MODELOS DE REGRESIÓN

Lilian Muñiz Álvarez[1], Facultad de Matemática y Computación, Universidad de La Habana
Rolando J. Biscay Lirio[2], Instituto de Cibernética, Matemática y Física (ICIMAF)

RESUMEN
En el presente trabajo se introduce una modificación de la función de verosimilitud, llamada verosimilitud suavizada. Esto se logra mediante una estimación por núcleo de la distribución empírica de los datos. Además, se estudia la aplicación de esta función en la estimación de modelos de regresión, y se brindan resultados teóricos acerca de la consistencia de los estimadores basados en ella. También se ilustra, a través de simulaciones, el comportamiento de esta función en la regresión polinomial.

Palabras clave: función de verosimilitud suavizada, modelos de regresión, consistencia de estimadores.

ABSTRACT
In the present work a modification of the likelihood function is introduced. It is based on what we call the smoothed likelihood function, which is obtained by substituting a smoothed (kernel) estimate for the empirical measure into the standard likelihood function. Also, applications of the smoothed likelihood-based inference in regression models estimation are shown, and theoretical results about the consistency of the estimators based on this function are given. It is also shown, by means of simulations, the behavior of this function in polynomial regression models.

1. INTRODUCCIÓN

La función de verosimilitud, salvo una constante aditiva, es una versión empírica de la divergencia de Kullback-Leibler (KL); más precisamente, es la divergencia KL con respecto a la medida de probabilidad empírica de los datos. Esta función es un instrumento clave de la inferencia estadística para modelos paramétricos (ver e.g. Cox y Hinkley, 1974). Sobre su base se ha desarrollado la llamada inferencia basada en verosimilitud, que comprende las regiones de verosimilitud, los estimadores máximo verosímiles (que, como es sabido, son asintóticamente eficientes), las regiones de verosimilitud-confianza (que son asintóticamente de precisión óptima) y las dócimas de hipótesis basadas en cocientes de verosimilitud.

Sin embargo, el enfoque basado en verosimilitud no es directamente aplicable a situaciones en las que no se conoce un modelo paramétrico regular para la distribución de los datos. Por ejemplo, cuando el modelo especificado consiste en una familia de distintos modelos paramétricos regulares. Esto incluye en particular el caso de modelos anidados de diferentes dimensiones. También este enfoque presenta serias limitaciones en modelos regulares en los que el número de parámetros es grande en comparación con la cantidad de datos disponibles (modelos “grandes”). Una de estas limitaciones es que no evita el llamado fenómeno de "sobreajuste" de parámetros.

En tales situaciones se utilizan enfoques alternativos. En el caso de modelos que comprenden submodelos de varias dimensiones, usualmente se aplican primero técnicas de selección de modelos, y posteriormente se realiza la inferencia basada en verosimilitud para el modelo seleccionado. La selección suele hacerse mediante criterios informacionales (como AIC; ver Burbham(2002), y bibliografía citada allí) o criterios de remuestreo (ver Efron(1982)). En caso de un modelo regular de alta dimensión en comparación con el tamaño de muestra, frecuentemente se utiliza el enfoque de verosimilitud penalizada para la construcción de estimadores, lo cual incluye el uso de estimadores bayesianos MAP como en Carlin(1996). La penalización es típicamente ponderada mediante un hiper-parámetro no negativo, que suele seleccionarse mediante criterios de remuestreo o informacionales.

Como consecuencia, en tales situaciones la función de verosimilitud no ofrece una base unificadora de la inferencia. La inferencia se desintegra en una fase de selección de modelos e hiper-parámetros y otra posterior fase de estimación y prueba de hipótesis (clásicas) dentro del modelo seleccionado, utilizando en la

E-mail: [1]lilian@matcom.uh.cu
[2]biscay@icmf.inf.cu

fase de selección criterios distintos a la divergencia KL. Además, las penalizaciones son en la práctica especificadas por el investigador siguiendo criterios más o menos arbitrarios ajenos a dicha divergencia (e.g., medidas de complejidad o suavidad del modelo).

El objetivo de este trabajo es introducir una modificación de la función de verosimilitud (llamada verosimilitud suavizada, VS) tal que: i) no requiera de la especificación por el investigador de un término de penalización; ii) conserve la propiedad de ser una divergencia KL, y iii) pueda ser utilizada como base de la inferencia en modelos constituidos por conjuntos arbitrarios de distribuciones sin conducir a sobreajustes. Específicamente, la VS es definida como la divergencia KL con respecto a un suavizamiento (por núcleo) de la distribución empírica de los datos. De este modo contiene a la función de verosimilitud clásica como caso particular cuando la cantidad de suavizamiento es cero.

Centraremos la atención del presente trabajo en la aplicación de la verosimilitud suavizada para la estimación de modelos de tipo regresión. En la Sección 2 definimos la VS en este contexto. En la Sección 3 se brindan resultados teóricos acerca de la consistencia de los estimadores basados en ella. En la Sección 4 se ilustra, a través de simulaciones, el comportamiento de la VS en la regresión polinomial. Finalmente, la Sección 5 contiene algunos problemas abiertos y generalizaciones, y también las conclusiones del trabajo.

2. LA FUNCIÓN DE VEROSIMILITUD SUAVIZADA

2.1. Formulación general

Consideremos un problema de análisis de regresión entre una variable respuesta $Y \in \Re$ y un vector de variables predictoras $X = (X_1,\ldots,X_p) \in \Re^p$. Más específicamente, supongamos que se tienen datos $\{(Y_i, x_i)\}_{i=1}^n$, con $x_i = (x_{i1},\ldots,x_{ip})$, que satisfacen la relación funcional:

$$Y_i = m(x_i) + \varepsilon_i, \qquad (1)$$

donde ε_i son variables aleatorias iid. con distribuciones Gaussianas, $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$, y μ es una función con dominio en $\Re^p$ (llamada función de regresión). Supongamos además que los puntos de diseño x_i son controlados, i.e., interesa estudiar el problema "condicionado" a la variable x. Obviamente, $E(Y_i / x_i) = \mu(x_i)$ y $V(Y_i / x_i) = \sigma^2$.

Existen diversos tipos de modelos para la función de regresión. Dos grandes clases son las siguientes.

a) En un modelo de regresión paramétrico (clásico o regular) se supone que la función de regresión μ tiene una expresión funcional especificada en dependencia de un número fijo y conocido q de parámetros reales desconocidos $\beta = (\beta_1,\ldots,\beta_q) \in B \subset \Re^q$. Se suponen además condiciones de suavidad convenientes acerca de la dependencia de μ con respecto a β. Un ejemplo es la regresión polinomial con grado q conocido, donde los parámetros son los coeficientes del polinomio de regresión.

b) En un modelo de regresión no paramétrico, sólo se supone que μ pertenece a cierta clase infinito-dimensional M de funciones "suaves", no indizada por un parámetro finito-dimensional. Por ejemplo, M puede ser la clase de las funciones con derivadas continuas hasta cierto orden sobre cierto dominio de $\Re^p$.

En la práctica, a veces el modelo paramétrico considerado por el investigador no ajusta bien los datos, y se carece de suficiente información previa para proponer un único modelo paramétrico alternativo que sea adecuado. Entonces la modelación no paramétrica constituye una atrayente opción, y las técnicas de suavizamiento no paramétrico ofrecen una herramienta flexible para estudiar la función de regresión desconocida.

Uno de los métodos de suavizamiento más simples es el de estimación por núcleos. En particular, el estimador de Nadaraya-Watson para la media $\mu(x)$ tiene la forma:

$$\tilde{\mu}_\lambda(x) = \frac{\sum_{i=1}^n Y_i K_\lambda(x - x_i)}{\sum_{i=1}^n K_\lambda(x - x_i)}, \qquad (2)$$

donde $K_\lambda(u) = \frac{1}{\lambda} K(u/\lambda)$ y el núcleo K es una función real continua, acotada, simétrica alrededor del cero y cuya integral es uno. El parámetro no negativo λ es llamado ancho del núcleo (Nadaraya (1964), Watson (1964)).

El estimador $\tilde{\mu}_\lambda(x)$ no es insesgado para muestras finitas. Pero bajo condiciones de regularidad convenientes es consistente (de acuerdo a varias métricas) y tiene distribución asintótica Normal (ver Hardle (1994)). Además, es sabido que $\tilde{\mu}_\lambda(x)$ no es muy sensible a la selección del núcleo K, sino sólo a la selección del ancho λ que controla el grado de suavizamiento (menor para valores pequeños de λ).

De la estimación (2) para la media se obtiene el siguiente estimador no paramétrico $\tilde{\sigma}^2_\lambda$ para la varianza σ^2:

$$\tilde{\sigma}^2_\lambda = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{\mu}_\lambda(x_i))^2. \qquad (3)$$

Luego una estimación natural de la densidad condicional f(y/x) de Y dado x es $f_\lambda(y/x) = N(\tilde{\mu}_\lambda(x), \tilde{\sigma}^2_\lambda)(y)$, donde $N(\tilde{\mu}_\lambda(x), \tilde{\sigma}^2_\lambda)(y)$, denota la función de densidad Normal de media $\tilde{\mu}_\lambda(x)$ y varianza $\tilde{\sigma}^2_\lambda$ evaluada en y.

c) Una situación intermedia entre las (a) y (b) anteriores es cuando se supone que la función de regresión pertenece a un subconjunto especificado M_0 de M que no es un modelo paramétrico. Por ejemplo, M_0 puede consistir en todos los polinomios de grados arbitrarios en la variable x. En este caso, μ pertenece a un conjunto de funciones que no puede describirse mediante un parámetro de dimensión finita.

Nuestro interés es extender el enfoque de verosimilitud de modo que sea aplicable no sólo a modelos paramétricos clásicos (a) sino también a los modelos de tipo (c) evitando el llamado fenómeno de sobreajuste. Para esto, a continuación definiremos, en el contexto de la regresión, lo que llamamos la función de VS (la Sección 5 discute su definición para modelos asociados a muestras *iid*).

Definición: Sea B un conjunto arbitrario y sea un modelo de regresión (1), donde la función de regresión $\mu(x) = \mu_\beta(x)$ está indizada por un parámetro $\beta \in B$. Asociada a este modelo, la función de log-verosimilitud suavizada (VS) $l_\lambda(\theta)$, con $\theta = (\beta, \sigma^2) \in \Theta = B \times \Re^*_+$, se define como:

$$l_\lambda(\theta) = \sum_{i=1}^{n}\int f_\lambda(y/x_i)\ln f(y/x_i;\theta)dy, \qquad (4)$$

donde

$$f_\lambda(y/x_i) = N(\tilde{\mu}_\lambda(x_i), \tilde{\sigma}^2_\lambda)(y), \qquad (5)$$

con $\tilde{\mu}_\lambda(x)$ y $\tilde{\sigma}^2_\lambda$ definidos por (2) y (3) respectivamente, $f(y / x_i; \theta) = N(\mu_\beta(x_i), \sigma^2)(y)$ y la integración es sobre todo $\Re$.

Salvo una constante aditiva, l_λ es la divergencia KL de $f(./x_i; \theta)$ con respecto a un suavizamiento $f_\lambda(./x_i)$ por núcleo de la distribución empírica de los datos. Nótese que ella contiene a la función de log-verosimilitud clásica

$$l(\theta) = \sum_{i=1}^{n}\ln f(Y_i / x_i;\theta) \qquad (6)$$

como caso particular cuando el ancho λ tiende a cero. Nótese que el hecho de que el conjunto B sea arbitrario permite incluir modelos de regresión lineales y no lineales con números de parámetros fijos, y también familias de modelos de regresión lineales o no lineales con diferentes números de parámetros.

Mediante la maximización con respecto a θ de la función de log-verosimilitud (6) se obtiene, como es sabido, el estimador máximo verosímil $\hat{\theta}$ de θ. Maximizando la función de log-verosimilitud suavizada l_λ definimos el estimador máximo verosímil suavizado $\hat{\theta}_\lambda$ de θ.

La selección del ancho λ es un paso crucial en los métodos no paramétricos de estimación. Un método de selección de λ por validación cruzada que utiliza la forma de $\mu(x)$ igual a $\mu_\beta(x)$ utilizada en la definición anterior es propuesto a continuación:

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{i=1}^{n} \ln f(Y_i / x_i; \hat{\theta}_\lambda^{(i)}), \tag{7}$$

donde

$$\hat{\theta}_\lambda^{(i)} = (\hat{\beta}_\lambda^{(i)}, \hat{\sigma}_\lambda^{2(i)}) = \arg\max_{\theta=(\beta,\sigma^2)} \sum_{j\neq i}^{n} \int f_\lambda^{(i)}(y / x_j) \ln f(y / x_j; \theta) dy. \tag{8}$$

Aquí $f_\lambda^{(i)}(y / x_j)$ denota al estimador por núcleo (5) basado en la muestra sin el dato (Y_i, x_i). Luego este criterio de selección consiste en hallar λ de modo que se maximice la log-verosimilitud de los datos en un sentido predictivo.

En la siguiente sección se estudia el caso particular de la regresión lineal, que tiene la ventaja computacional de que en él se obtiene de forma explícita el estimador máximo verosímil suavizado de $\theta = (\beta, \sigma^2)$.

2.2. Caso de la regresión lineal

En la sección anterior la función de regresión $\mu_\beta(x)$ podía ser una función cualquiera, tanto lineal como no lineal. En el caso de la regresión lineal, la variable respuesta Y depende de forma lineal del parámetro β. Más específicamente, supongamos que se tienen datos $\{(Y_i, x_i)\}_{i=1}^{n}$ que satisfacen la relación funcional:

$$Y_i = \mu_\beta(x_i) + \varepsilon_i, \tag{9}$$

donde las variables aleatorias ε_i son como en la sección anterior, $\beta = (\beta_1,\ldots,\beta_p)'$, $x_i = (x_{i1},\ldots,x_{ip})'$ y

$$\mu_\beta = (x_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\ x_i = x_i'\beta. \tag{10}$$

La ecuación de regresión (9) que satisface (10) puede escribirse de forma matricial como $Y = X\beta + \varepsilon$, donde $Y\ (Y_1,\ldots,Y_p)'$, $\varepsilon(\varepsilon_1,\ldots,e_n)'$ y $X = (x_{ij})$, con i = 1,...,n y j = 1,...,p.

En este caso la función de log-verosimilitud suavizada $l_\lambda(\theta)$ con $\theta = (\beta, \sigma^2)$ según (4) es:

$$l_\lambda(\theta) = \sum_{i=1}^{n} \int f_\lambda(y / x_i) \ln f(y / x_i; \theta) dy,$$

donde $f_\lambda(y/x_i) = N(\tilde{\mu}_\lambda(x_i), \tilde{\sigma}_\lambda^2(y)\, f(y)$ como anteriormente y, a diferencia de la sección anterior, $f(y / x_i; \theta) = N(x_i', \beta, \sigma^2)(y)$.

Teniendo en cuenta que tanto $f_\lambda(y/x_j)$ como $f(y/x_i; \theta)$ son densidades normales, el problema de maximización (8) toma la forma:

$$\theta_\lambda^{(i)} = \arg\max_{\theta=(\beta,\sigma^2)} \sum_{j\neq 1} \left[\ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \left(\tilde{\sigma}_\lambda^{2(i)} + \left(\tilde{\mu}_\lambda^{(i)}(x_j) - x_j'\beta \right)^2 \right) \right],$$

por lo que se obtiene que

$$\hat{\beta}_\lambda^{(i)} = (X^{(i)\prime}X^{(i)})^{-1}X^{(i)\prime}\tilde{\mu}_\lambda^{(i)} \tag{11}$$

y

$$\hat{\sigma}_\lambda^{2(i)} = \frac{1}{n-1}\sum_{j\neq 1}\left(\tilde{\sigma}_\lambda^{2(i)} + \left(\tilde{\mu}_\lambda^{(i)}(x_j) - x_j'\beta_\lambda^{(i)}\right)^2\right)'. \tag{12}$$

Aquí $X^{(i)}$ denota la matriz X sin la fila i-ésima y los valores de $\tilde{\mu}_\lambda^{(i)}$ y $\tilde{\sigma}_\lambda^{2(i)}$ son calculados sin usar el dato (Y_i, x_i). De (11) y (12) se deduce que los estimadores máximos verosímiles suavizados $\hat{\beta}_\lambda$ y $\tilde{\sigma}_\lambda^2$ de β y σ^2 respectivamente tienen la forma:

$$\hat{\beta}_\lambda = (X'X)^{-1}X'\tilde{\mu}_\lambda, \tag{13}$$

donde

$$\tilde{\mu}_\lambda = (\tilde{\mu}_\lambda(x_1),\ldots,\tilde{\mu}_\lambda(x_n))' \text{ y } \tilde{\sigma}_\lambda^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\sigma}_\lambda^2 + \left(\tilde{\mu}_\lambda(x_i) - x_j'\hat{\beta}_\lambda\right)^2\right)'.$$

Un ejemplo de regresión lineal es el caso de la regresión polinomial. En la Sección 4 se estudiará el comportamiento de la VS en este contexto.

3. CONSISTENCIA DEL ESTIMADOR MÁXIMO VEROSÍMIL SUAVIZADO

La consistencia (según convergencia en probabilidad) del estimador máximo verosímil suavizado de β en el caso de la regresión lineal es consecuencia de la consistencia del estimador por núcleo de Nadaraya-Watson $\tilde{\mu}_\lambda(x)$, resultado este último que aparece por ejemplo en Hardle(1994). Este asegura que si:

(C1) La verdadera función de regresión μ es una función de Lipschitz.

(C2) El conjunto X donde toma valores la variable x es compacto.

(C3) Los errores ε_i están acotados.

(C4) $|K(u)| \leq 1$.

Entonces se tiene que:

$$\sup_{x\in X}|\tilde{\mu}_\lambda(x) - \mu(x)| = O_p\left(\max\left\{\left(\frac{n\lambda}{\log n}\right)^{-1/2}, \lambda\right\}\right).$$

Por tanto si $n \underset{n\to\infty}{\to} 0$ y $\left(\frac{n\lambda}{\log n}\right)^{-1/2} \underset{n\to\infty}{\to} 0$ y entonces el resultado anterior implica que el estimador de Nadaraya-Watson $\tilde{\mu}_\lambda(x)$ converge en probabilidad a la verdadera función de regresión $\mu(x)$, que en el caso de la regresión lineal es igual a $m_\beta(x) = x'\beta$. Luego puede plantearse el siguiente teorema.

Teorema: Sea un problema de regresión lineal como en (9) y (10). Supongamos además que se satisfacen las condiciones (C1)-(C4) y que se cumplen los siguientes supuestos:

(A1) $\lambda \underset{n\to\infty}{\to} 0$ y $\left(\frac{n\lambda}{\log n}\right)^{-1/2} \underset{n\to\infty}{\to} 0$.

(A2) $\left\|\sqrt{n}(X'X)^{-1}X'\right\|_{n\to\infty}^2 = O(1)..$

Entonces el estimador máximo verosímil suavizado (13) de β es consistente en probabilidad.

Demostración:

Utilizando la forma explícita del estimador máximo verosímil suavizado de $\hat{\beta}_\lambda = (X'X)^{-1}X'\tilde{\mu}_\lambda$, y algunas propiedades de normas se tiene que:

$$\begin{aligned}\left\|\hat{\beta}_\lambda - \beta\right\|^2 &= \left\|(X'X)^{-1}X'\tilde{\mu}_\lambda - (X'X)^{-1}X'X\beta\right\|^2 \\ &= \left\|(X'X)^{-1}X'(\tilde{\mu}_\lambda - X\beta)\right\|^2 \\ &\le \left\|(X'X)^{-1}X'\right\|^2 \left\|\tilde{\mu}_\lambda - X\beta\right\|^2 \\ &= n\left\|(X'X)^{-1}X'\right\|^2 \frac{1}{2}\left\|\tilde{\mu}_\lambda - X\beta\right\|^2.\end{aligned}$$

Según (A2),

$$n\left\|(X'X)^{-1}X'\right\|^2 = \left\|\sqrt{n}(X'X)^{-1}X'\right\|^2 = O(1).$$

Además, teniendo en cuenta las condiciones (C1)-(C4) y la hipótesis (A1), el término $\frac{1}{n}\left\|\tilde{\mu}_\lambda - X\beta\right\|^2$ tiende a cero en probabilidad. En efecto,

$$\frac{1}{n}\left\|\tilde{\mu}_\lambda - XB\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}(\tilde{\mu}_\lambda(x_i) - x_i'\beta)^2 \le \frac{1}{n} n \sup_{x\in X}\left|(\tilde{\mu}_\lambda(x) - x'\beta)\right|^2 = \sup_{x\in X}\left|(\tilde{\mu}_\lambda(x) - x'\beta)\right|^2 \xrightarrow[n\to\infty]{P} 0.$$

Luego queda demostrado que $\left\|\hat{\beta}_\lambda - \beta\right\| \xrightarrow[n\to\infty]{P} 0$, por lo que el estimador máximo verosímil suavizado $\hat{\beta}_\lambda$ de β es consistente.

4. ESTUDIO POR SIMULACIONES DEL COMPORTAMIENTO DE LA FUNCIÓN DE VS EN LA REGRESIÓN POLINOMIAL

Consideremos una regresión polinomial entre una variable respuesta Y y una variable predictora escalar t. Sean n datos $\{(Y_i, t_i)\}_{i=1}^n$, se tiene la ecuación:

$$Y_i = \mu_\beta(t_i) + \varepsilon_i, \qquad (14)$$

donde $\beta = (\beta_0, \beta_1,\ldots,\beta_p)'$ y $\mu_\beta(t_i) = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \ldots + \beta_p t_i^p$. De forma matricial la ecuación de regresión (14) puede escribirse como $Y = X\beta + \varepsilon$, donde $Y = (Y_1,\ldots,Y_n)'$, $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_n)'$ y $X = (t_i^{j-1})$, con i = 1,...,n y j = 1,...,p+1. O sea, $\mu_\beta(t_i) = x_i'\beta$, donde $x_i = (1, t_i, t_i^2,\ldots t_i^p)' \in \Re^{p+1}$ denota la i-ésima fila de la matriz X.

Nos interesa la situación, frecuente en la práctica, en que los datos satisfacen un modelo del tipo (14) pero con un verdadero grado $p = p_0$ desconocido por el investigador. Consideremos pues el modelo (14) con un grado suficientemente grande $p \ge p_0$. Nótese que debido a que típicamente p se toma mucho mayor que p_0, el enfoque de verosimilitud clásica no es aplicable pues conduciría a sobreajuste de los parámetros.

Para el estudio por simulaciones se tomaron n = 20 observaciones (Y_i, t_i), donde los n valores t_i de la variable predictora escalar t son equidistantes en el intervalo [0, 1]. La verdadera densidad se tomó como $f(y/x_i; \theta_0) = N(N(x_i'\beta^0, \sigma_0^2)(y)$, donde $\beta^0 = (1.6913, 8.4207, -9.2430, 3.5334, 0,0,0)'$ y $\sigma_0^2 = 0.7$. Notar que aquí el verdadero grado es $p_0 = 3$ mientras que el modelo se tomó con polinomios ("candidatos") de grados hasta p = 5.

Para el estudio, se calcularon varias estimaciones de la verdadera función de regresión m_{β^0}. Estas son:

a) Las estimaciones por el método de los mínimos cuadrados de las funciones de regresión polinomiales de grados p = 0,1,…,5.

b) El polinomio óptimo por el método de Validación Cruzada Generalizada (VCG), propuesto por Wahba (1977). Este consiste en tomar el grado óptimo como sigue:

$$p_{VCG} = \underset{k\in\{0,1,2,\ldots,5\}}{\arg\min} \frac{\frac{1}{n}\left\| Y - H_k Y \right\|^2}{(1 - \mathrm{tr} H_k)^2},$$

donde $H_k = X_k (X_k' X_k)^{-1} X_k'$ denota la matriz "hat" calculada a partir de la matriz X_k que contiene las primeras k+1 columnas de X. Se calculó entonces el estimador por mínimos cuadrados $\hat{\mu}_{VCG}$ de la función de regresión dada por el polinomio de grado p_{VCG}.

c) La estimación por núcleo $\hat{\mu}_{\hat{\lambda}_K} = \tilde{\mu}_{\hat{\lambda}_K}$ según (2) tomando $\lambda = \hat{\lambda}_K$ determinado según

$$\hat{\lambda}_K = \underset{\lambda}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \tilde{\mu}_\lambda^{(i)}(t_i) \right)^2,$$

donde $\tilde{\mu}_\lambda^{(i)}(t_i)$ denota que el estimador de Nadaraya-Watson (2) es hallado sin usar el dato (Y_i, t_i) y está evaluado en t_i. Esta estimación por validación cruzada del ancho del núcleo fue propuesta por Clark (1980).

d) La estimación por el método de verosimilitud suavizada: $\hat{\mu}_{\hat{\beta}_{\hat{\lambda}_S}} = X\hat{\beta}_{\hat{\lambda}_S}$, donde según (7),

$$\hat{\lambda}_S = \underset{\lambda}{\arg\min} \sum_{i=1}^{n} \ln f \left(Y_i - x_i; \hat{\theta}_\lambda^{(i)} \right)^2,$$

$\theta_\lambda^{(i)} = (\hat{\beta}_\lambda^{(i)}, \hat{\sigma}_\lambda^{2(i)})$ es el vector de los estimadores por verosimilitud suavizada según (11)-(12) utilizando dimensión p = 5, y $\hat{\beta}_{\hat{\lambda}_S}$ se calcula por (13) con $\lambda = \hat{\lambda}_S$.

e) La estimación por núcleo $\hat{\mu}_{\hat{\lambda}_S} = \tilde{\mu}_{\hat{\lambda}_S}$ según (2) tomando $\lambda = \hat{\lambda}_S$ como en (d).

f) La estimación $\hat{\mu}_{\hat{\beta}_{\hat{\lambda}_K}} = X\hat{\beta}_{\hat{\lambda}_K}$, donde $\hat{\lambda}_K$ definido como en (c) y $\hat{\beta}_{\hat{\lambda}_S}$ se calcula por (13) con $\lambda = \hat{\lambda}_K$.

En todos los ajustes mencionados en (c)-(f) el estimador de Nadaraya-Watson (2) se halló tomando el núcleo gaussiano K(u) = N(0, 1)(u).

Para ilustrar, las dos figuras siguientes muestran todos los ajustes en una muestra simulada. En la Figura 1 se muestran los ajustes por el método de mínimos cuadrados de las distintas regresiones polinomiales. Se observa que los polinomios de grados 0 y 1 no tienen suficiente flexibilidad para aproximar la verdadera función de regresión μ_{β^0}, mientras que los polinomios de grados 3, 4 y 5 presentan demasiadas oscilaciones. El polinomio ajustado más próximo a μ_{β^0} es el de grado 2, grado que es inferior al 3 del verdadero polinomio. Esto está en concordancia con la común experiencia acerca de que, cuando la cantidad de datos es moderada, modelos "parsimoniosos", i.e. con pocos parámetros, resultan más convenientes para el ajuste que modelos "grandes".

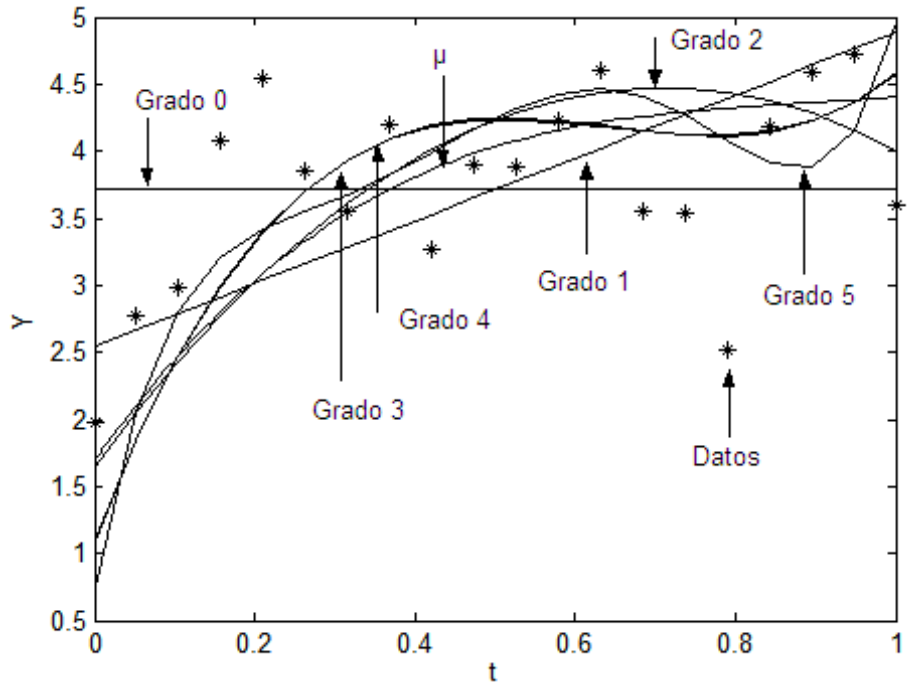

Figura 1. Ajustes por mínimos cuadrados (a) con $\mu = \mu_{\beta^0}$.

En la Figura 2 se muestran los ajustes (b)-(f). Se observa que los mismos aproximan la verdadera media evitando el gran sesgo de los polinomios de grados 0 y 1 a la vez que la extrema variabilidad de los polinomios de alto grado 3, 4 y 5 cuando estos se ajustan por mínimos cuadrados.

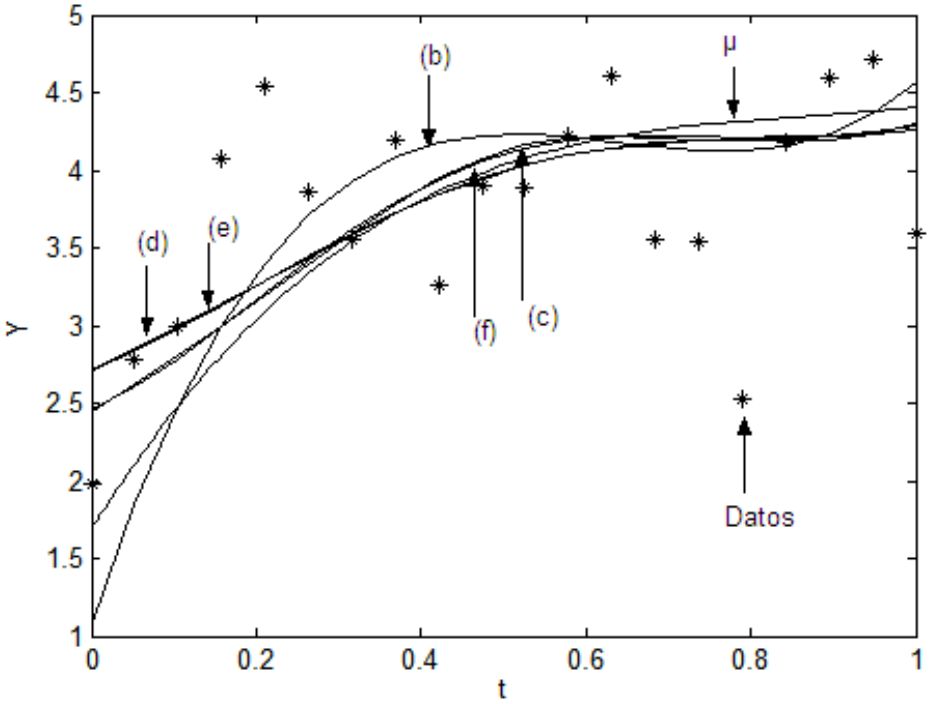

Figura 2. Ajustes (b): $\hat{\mu}_{VCG}$, (c): $\hat{\mu}_{\hat{\lambda}_K}$, (d): $\hat{\mu}_{\hat{\beta}_{\hat{\lambda}_S}}$, , (e): $\hat{\mu}_{\hat{\lambda}_S}$ y (f): $\hat{\mu}_{\hat{\beta}_{\hat{\lambda}_K}}$, con $\mu = \mu_{\beta^0}$.

Para estudiar el comportamiento promedio de estos estimadores a través de réplicas se realizaron las siguientes simulaciones. Se generaron un número B = 300 de muestras independientes de tamaño n = 20; para cada una de estas réplicas se calcularon las estimaciones (a) - (f); y finalmente se calcularon los errores cuadráticos medios de cada uno de ellos con respecto a la verdadera media mediante

$$\text{ECM} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}^{b}(t_i) - \mu_{\beta 0}(t_i)\right)^2,$$

donde $\hat{\mu}$ denota el estimador que corresponda de los mencionados en (a)-(f) obtenido en la b-ésima réplica.

Análogamente, también se calcularon el sesgo y varianza de cada estimador. Como es sabido, el ECM se descompone como suma de la varianza y el cuadrado del sesgo (Sesgo2).

Los resultados obtenidos para los estimadores por mínimos cuadrados de la función de regresión basados en polinomios de distintos grados se muestran en la Tabla I.

Tabla I.

	p = 0	**p = 1**	**p = 2**	**p = 3**	**p = 4**	**p = 5**
ECM	0.6729	0.1643	0.0858	0.1065	0.1309	0.1537
Sesgo2	0.6462	0.1106	0.0059	0.0002	0.0002	0.0002
Varianza	0.0267	0.0537	0.0799	0.1063	0.1307	0.1535

Para los estimadores (b)-(f), los resultados obtenidos se presentan en la Tabla II.

Tabla II.

	$\hat{\mu}_{VCG}$	$\hat{\mu}_{\hat{\lambda}_K}$	$\hat{\mu}_{\hat{\beta}_{\hat{\lambda}_S}}$	$\hat{\mu}_{\hat{\lambda}_S}$	$\hat{\mu}_{\hat{\beta}_{\hat{\lambda}_K}}$
ECM	0.1259	0.1379	0.1326	0.1333	0.1208
Sesgo2	0.0015	0.0168	0.0565	0.0565	0.0165
Varianza	0.1244	0.1211	0.0761	0.0768	0.1043

El análisis de estas tablas revela los siguientes hechos que merecen destacarse:

Los ajustes por mínimos cuadrados con polinomios de grados 2, 3 y 4 brindan los menores valores del error cuadrático medio. Pero estos tienen la desventaja de que en la práctica no se conoce el verdadero grado del polinomio.

En general los estimadores $\hat{\mu}_{VCG}$, $\hat{\mu}_{\hat{\beta}_{\hat{\lambda}_K}}$ y $\hat{\mu}_{\hat{\beta}_{\hat{\lambda}_K}}$ tienen un ECM comparable con el de los mejores polinomios (grados 2 al 4). Los estimadores $\hat{\mu}_{\hat{\lambda}_K}$ y $\hat{\mu}_{\hat{\lambda}_S}$ presentan peor ECM. Esto pudiera deberse a que los estimadores por núcleo no usan la información adicional de que la media es polinomial.

De los estimadores (b)-(f), $\hat{\mu}_{\hat{\lambda}_S}$ y $\hat{\mu}_{\hat{\beta}_{\hat{\lambda}_K}}$ son los que muestran mayor balance entre las componentes de sesgo y varianza de sus errores cuadráticos medios.

Curiosamente, $\hat{\mu}_{\hat{\beta}_{\hat{\lambda}_K}}$ muestra el menor error cuadrático medio entre los estimadores (b)-(f), el cual resulta comparable con el de los dos mejores polinomios (grados 2 y 3). Nótese que tal estimador se obtiene simplemente sustituyendo en (13) a λ por una estimación no paramétrica estándar del ancho del núcleo, $\hat{\lambda}_K$.

Debe advertirse que, teniendo en cuenta la cantidad limitada de simulaciones realizadas, las pequeñas diferencias observadas entre $\hat{\mu}_{VCG}$, $\hat{\mu}_{\hat{\beta}_{\hat{\lambda}_K}}$ y $\hat{\mu}_{\hat{\beta}_{\hat{\lambda}_K}}$ deben interpretarse cautelosamente.

5. CONCLUSIONES Y PROBLEMAS ABIERTOS

Los resultados obtenidos en las secciones anteriores nos permiten concluir que:

1. La función de VS introducida en este trabajo permite en problemas de regresión definir la verosimilitud sobre espacios de funciones de regresión mucho más generales que los modelos paramétricos clásicos, evitando no obstante el sobreajuste de la estimación basada en ella. Tales espacios pueden ser arbitrarios con la única condición de estar contenidos en el conjunto de las funciones de regresión estimables por núcleo. De este modo se incluyen modelos consistentes en familias de distintos modelos de regresión paramétricos, quizás de distintas dimensiones.

2. A diferencia de la verosimilitud penalizada, la VS tiene una interpretación directa como divergencia KL y no requiere de la especificación de un término de penalización.

3. En modelos de regresión lineales con número desconocido de variables predictoras, los estimadores basados en la VS son consistentes.

4. Aplicada a modelos de regresión polinomiales, los resultados de simulaciones muestran que los estimadores basados en la VS son factibles y muestran un comportamiento comparable a los estimadores según enfoque VCG.

No obstante, trabajos posteriores son necesarios para profundizar más sobre las propiedades asintóticas y no asintóticas de la estimación basada en la VS. Estudios con mayor número de simulaciones se requieren para arribar a conclusiones más precisas. Por otra parte, el criterio de ECM debe complementarse con otros criterios de calidad de la estimación, en especial el criterio de divergencia KL con respecto a la verdadera distribución de la muestra es de interés en este contexto. Además, si bien una notable ventaja potencial del enfoque de VS es en su posibilidad de tratar modelos complejos que comprendan submodelos paramétricos lineales y no lineales -situación para la cual el enfoque VCG no está diseñado-, la exploración del comportamiento práctico de la VS para tales situaciones está abierta a trabajos futuros.

Por otra parte, la inferencia basada en la VS es susceptible de extenderse en varias direcciones.

En particular, pudiera formularse para modelos de regresión heterocedásticos, descritos por funciones de regresión para la media y la varianza.

También puede formularse el enfoque VS para el caso de datos iid de la manera siguiente. Sea dado un modelo estadístico $(\aleph, \{f(.;\theta):\theta \in \Theta\})$ consistente en un conjunto de funciones de densidad sobre un mismo espacio muestral $\aleph$ y espacio de parámetros Θ arbitrario, al cual pertenece la verdadera densidad $f(.;\theta)$. Dadas observaciones iid x_i, con i = 1,...,n con densidad $f(.;\theta_0)$, puede definirse la función de log-verosimilitud suavizada como:

$$l_\lambda(\theta) = \int f_\lambda(z) \ln f(z:\theta) dz,$$

donde $f_\lambda(.)$ es una estimación no paramétrica por núcleo de $f(.;\theta_0)$:

$$f_\lambda(z) = \frac{1}{n\lambda} \sum_{i=1}^{n} K\left(\frac{z - x_i}{\lambda} \right).$$

Aquí el núcleo K es una función real continua, acotada, simétrica alrededor del cero y cuya integral es uno (ver e.g. Van der Vaart, 1998). El estimador máximo verosímil suavizado de θ se definiría por maximización de $l_\lambda(\theta)$. En esta situación, un criterio predictivo para la determinación de un valor $\hat{\lambda}$ para el ancho λ puede ser el siguiente:

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{i=1}^{n} \ln f(x_i; \hat{\theta}_\lambda^{(i)}),$$

donde

$$\hat{\theta}_\lambda^{(i)} = \arg\max_{\theta} \int f_\lambda^{(i)}(z) \ln f(z;\theta) dz,$$

y $\hat{\theta}_{\lambda}^{(i)}$ y $f_{\lambda}^{(i)}(z)$ son, respectivamente, la estimación máximo verosímil suavizada de θ y la estimación por núcleo de $f(.; \theta_0)$ basadas en todos los datos menos x_i.

REFERENCIAS

BURBHAM, K. P. and D.R. ANDERSON (2002): **Model Selection and Multimodel Inference**. Springer: N.Y.

CARLIN, B.P. and T.A. LOUIS (1996): **Bayes and empírical Bayes methods for data analysis**. London: Chapman and Hall.

CLARK, R. M. (1980): "Calibration, cross-validation and carbon 14 ii". **Journal of the Royal Statistical Society**, Series A 143: 177-194.

COX, D. R. and D.V. HINKLEY (1974): **Theoretical Statistics**. Chapman and Hall: London.

EFRON, B. (1982): "The Jacknife, the Bootstrap and Other Resampling Plans", Regional Conference Series in **Applied Mathematics**, 38. Philadelphia: SIAM.

HARDLE, W. (1994): **Applied Non-parametric Regression**. Cambridge Univ. Press: Cambridge.

NADARAYA, E. A. (1964): "On estimating regresión". **Theory Prob. Appl**. 10: 186-190.

VAN der VAART, A.W. (1998): **Asymptotic Statistics**. Cambridge Univ. Press: Cambridge.

WAHBA, G. (1977): "Applications of statistics", in P. Krishnaiah (ed.), A survey of some smoothing problems and the method of generalized cross-validation for solving them, North Holland, Amsterdam.

WATSON, G. S. (1964): **Smooth regression analysis, Sankhya**, Series A 26: 359-372.