

APROXIMACIÓN DE FUNCIONES MULTIVARIANTES MEDIANTE REGRESIÓN LINEAL JERÁRQUICA

VÍCTOR DE BUEN REMIRO

RESUMEN. El enfoque usual de aproximación de funciones como un problema de mero cálculo numérico es reemplazado en este artículo por una visión estadística del problema que permite una formulación más robusta y un mecanismo más eficiente.

A semejanza de los métodos clásicos de interpolación y aproximación sin malla, la función de aproximación global se construirá como una combinación lineal convexa de funciones que aproximan localmente la función objetivo en entornos que recubren su dominio. Estas funciones de aproximación local son a su vez combinaciones lineales de unas pocas funciones básicas linealmente independientes. De este modo, para evaluar la función de aproximación global sólo será necesario localizar los entornos locales a los que pertenece y evaluar en ellos las funciones de aproximación local.

Se trabajará bajo la hipótesis de errores de aproximación normales independientes y la existencia de relaciones latentes entre las funciones de aproximación local para construir un modelo de regresión lineal jerárquico con matriz de inputs muy dispersa y por lo tanto muy eficaz en cuanto a su estimación máximo-verosímil, todo ello sin necesidad de tener que aproximar cada zona por separado para tener luego que ensamblarlas ni reajustarlas.

1. ANTECEDENTES

Sea la función objetivo real continua y doblemente diferenciable en su dominio Ω

$$(1.1) \quad f(x) : \Omega \subset \mathbb{R}^n \longrightarrow \mathbb{R}$$

de la cual en principio se desconoce su formulación analítica y no se tiene ninguna otra forma de evaluarla o al menos no con la velocidad que se requiere.

Sea una colección no vacía ordenada de $J > 0$ puntos en una región $\Omega \subset \mathbb{R}^n$ a los que llamaremos puntos nodales

$$(1.2) \quad \mathcal{X} = \{x_j \mid x_j \in \Omega \subset \mathbb{R}^n \forall j = 1 \dots J\}$$

para los cuales se conoce el valor de la función objetivo $f_j = f(x_j)$. La matriz X cuyas filas son los puntos nodales debe ser de rango completo y se denominará matriz nodal del problema. Si no fuera así habría que reducir la dimensión del problema eliminando las dependencias lineales.

Se desea construir mediante un modelo de regresión lineal sparse con residuos normales independientes una función de aproximación global de f que sea continua y doblemente diferenciable

$$(1.3) \quad \hat{f}(x) : \Omega \subset \mathbb{R}^n \longrightarrow \mathbb{R}$$

Como introducción al estado del arte en el ataque clásico de la ingeniería mecánica acerca de los métodos de malla, y en especial de los de mínimos cuadrados móviles, sería interesante la lectura del artículo [4]

2. MODELO LINEAL OBSERVACIONAL PURO

Sea otra colección ordenada no vacía de $H > 0$ puntos dentro de esa misma región convexa $\Omega \subset \mathbb{R}^n$ a los que llamaremos puntos de referencia en los que el valor de la función objetivo puede ser conocido o no, es decir, un punto de referencia puede ser, o bien uno de los puntos nodales o bien otro punto cualquiera. Sean Ω_h regiones de entorno local de cada uno, tales que recubran entre todas la región Ω

$$(2.1) \quad \mathcal{Y} = \{y_h \mid y_h \in \Omega_h \subset \Omega \subset \mathbb{R}^n \forall h = 1 \dots H\}$$

$$\Omega \subset \bigcup_{h=1}^H \Omega_h$$

Sea \mathcal{P} una colección no vacía ordenada de $K > 0$ de funciones linealmente independientes continuas y doblemente diferenciables en todo \mathbb{R}^n a las que llamaremos base de la aproximación local o simplemente base

$$(2.2) \quad \mathcal{P} = \{p_k(x) : \mathbb{R}^n \longrightarrow \mathbb{R}\}$$

Para cada punto de referencia se definirá una función de aproximación local como una combinación lineal de dicha base en su región correspondiente

$$(2.3) \quad g_h(x) = \sum_{k=1}^K \alpha_{h,k} \cdot p_k(x - y_h) \forall x \in \Omega_h \wedge h = 1 \dots H$$

por lo que es obvio que deben ser continuas y doblemente diferenciables en todo \mathbb{R}^n .

La función de aproximación global en un punto se definirá a su vez como una combinación convexa de las funciones locales cuyos entornos contengan dicho punto

$$(2.4) \quad \hat{f}(x) = \sum_{x \in \Omega_h} w_h(x) \cdot g_h(x) \forall x \in \mathbb{R}^n$$

mediante unas funciones de ponderación $w_h(x)$ que no dependen de ningún parámetro y deben ser continuas, doblemente diferenciables en todo \mathbb{R}^n , alcanzar el máximo valor en el punto de referencia, y dar siempre suma 1 al aplicarla a todos los entornos locales a los que pertenece un punto cualquiera

$$(2.5) \quad \begin{aligned} w_h(x) &\geq 0 \forall x \in \Omega_h \\ w_h(x) &= 0 \forall x \notin \Omega_h \\ w_h(x) &\leq w_h(y_h) \forall x \in \mathbb{R}^n \\ \sum_{x \in \Omega_h} w_h(x) &= 1 \forall x \in \mathbb{R}^n \end{aligned}$$

Se supondrá que el error de aproximación global sigue una distribución normal independiente de media nula y varianza desconocida común para todos los puntos de referencia

$$f(x) - \hat{f}(x) = e(x) \sim N(0, \sigma^2) \forall x \in \Omega$$

$$(2.6) \quad f(x) = \hat{f}(x) + e(x) = \sum_{x \in \Omega_h} w_h(x) \cdot g_h(x) + e(x)$$

$$(2.7) \quad f(x) = \sum_{x \in \Omega_h} \sum_{k=1}^K w_h(x) \cdot p_k(x - y_h) \cdot \alpha_{h,k} + e(x)$$

Contrastando con los valores conocidos en los puntos nodales se reduce el problema de la aproximación a estimar la siguiente regresión lineal

$$(2.8) \quad F_j = \sum_{x_j \in \Omega_h} \sum_{k=1}^K P_{h,k,j} \alpha_{h,k} + e_j$$

donde

$$(2.9) \quad \begin{aligned} e_j &= e(x_j) \sim N(0, \sigma^2) \\ F_j &= f(x_j) \\ P_{h,k,j} &= w_h(x_j) p_k(x_j - y_h) \end{aligned}$$

Si definimos

$$(2.10) \quad \begin{aligned} \beta_i &= \alpha_{h,k} \\ Q_{i,j} &= P_{h,k,j} \end{aligned}$$

es posible escribir la regresión en forma matricial

$$(2.11) \quad \begin{aligned} F &= Q \cdot \beta + e \\ e &\sim N(0, \sigma^2 I_J) \\ F &\in \mathbb{R}^J \\ \beta &\in \mathbb{R}^{HK} \\ Q &\in \mathbb{R}^{J \times HK} \end{aligned}$$

Como la matriz de input Q es muy dispersa es relativamente poco costoso calcular su descomposición de Cholesky

$$(2.12) \quad Q^T \cdot Q = L \cdot L^T$$

lo cual nos da un mecanismo de estimación máximo-verosimil bastante asequible de los parámetros del modelo

$$(2.13) \quad \hat{\beta} = (Q^T \cdot Q)^{-1} Q \cdot F = L^{-T} L^{-1} Q \cdot F$$

Por ejemplo, el paquete de software CHOLMOD es capaz de resolver este sistema lineal sin necesidad de efectuar el producto $Q^T \cdot Q$ y con mayor eficacia cuanto más dispersa sea la matriz.

3. MODELO JERÁRQUICO

Para forzar que la función de aproximación global sea más suave es posible añadir más hipótesis al modelo que obliguen a las funciones de aproximación locales parecerse en cierta medida a sus vecinas en las zonas compartidas de sus entornos locales. Una hipótesis que parece bastante razonable es que en los puntos nodales comunes a varios entornos locales las diferencias entre las correspondientes aproximaciones locales siga una distribución normal similar a la de los errores de aproximación global, es decir

$$g_h(x) - g_s(x) = e_{h,s}(x) \sim N(0, \sigma^2) \forall x \in \Omega_h \cap \Omega_s$$

Si seleccionamos una nueva colección ordenada no vacía de puntos en la intersección de dos entornos locales $\Omega_h \cap \Omega_s$ a los que llamaremos puntos de estrés en los que el valor de la función objetivo es puede ser conocido o no.

$$(3.1) \quad \begin{aligned} \mathcal{Z}_{h,s} &= \{z_{h,s,t} \mid z_{h,s,t} \in \Omega_h \cap \Omega_s \subset \Omega \subset \mathbb{R}^n \forall t = 1 \dots T_{h,s}\} \\ T &= \sum_{h,s}^{\Omega_h \cap \Omega_s \neq \emptyset} T_{h,s} \end{aligned}$$

entonces podemos añadir a la matriz Q las T ecuaciones latentes al modelo definido anteriormente

$$(3.2) \quad 0 = \sum_{k=1}^K p_k(z_{h,s,t} - y_h) \alpha_{h,k} - \sum_{k=1}^K p_k(z_{h,s,t} - y_s) \alpha_{s,k} + e_{h,s,j}$$

4. ESTIMACIÓN ADAPTATIVA

El diseño de un modelo puede ser globalmente aceptable y sin embargo acumular en alguna zona concreta demasiados residuos altos. También puede ocurrir que con el tiempo se vaya disponiendo de nuevos datos que produzcan esos residuos inesperados, o que simplemente caigan fuera del alcance del recubrimiento actual. En ambos casos podría ser necesario introducir más puntos de referencia y de estrés por las proximidades. Si lo que ocurre es que la aproximación no consigue adaptarse a la curvatura aparente de los datos podría ser que hubiera que aumentar el tamaño de la base.

Cuando las dimensiones de la regresión son pequeñas se podría readaptar el diseño sin más y volver a estimarla desde cero, pero en los casos de datos masivos eso puede tener un coste difícil de asumir. En estos casos se puede aplicar un conocido mecanismo de ampliación matricial sobre el algoritmo de Cholesky.

Supongamos que tenemos el problema original resuelto para una matriz de input Q_{11} de $J_1 + T_1$ ecuaciones y $H_1 \cdot K_1$ variables y un output F_1

$$(4.1) \quad Q_{11}^T \cdot Q_{11} = L_0 \cdot L_0^T$$

$$(4.2) \quad \hat{\beta}_0 = L_1^{-T} L_1^{-1} Q_1 \cdot F_1$$

Por las razones antedichas tenemos que añadir un número notablemente inferior $J_2 + T_2$ de ecuaciones y de $H_2 \cdot K_2$ variables nuevas para formar las nuevas matrices

$$(4.3) \quad Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}$$

$$(4.4) \quad F = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix}$$

Desarrollando el producto por bloques

$$(4.5) \quad Q^T \cdot Q = \begin{pmatrix} Q_{11}^T & Q_{21}^T \\ Q_{12}^T & Q_{22}^T \end{pmatrix} \cdot \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} = \begin{pmatrix} Q_{11}^T \cdot Q_{11} + Q_{21}^T \cdot Q_{21} & Q_{11}^T \cdot Q_{12} + Q_{21}^T \cdot Q_{22} \\ Q_{12}^T \cdot Q_{11} + Q_{22}^T \cdot Q_{21} & Q_{12}^T \cdot Q_{12} + Q_{22}^T \cdot Q_{22} \end{pmatrix}$$

y haciendo lo propio con la descomposición de Cholesky

$$(4.6) \quad Q^T \cdot Q = LL^T = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \cdot \begin{pmatrix} L_{11}^T & L_{21}^T \\ 0 & L_{22}^T \end{pmatrix} = \begin{pmatrix} L_{11} \cdot L_{11}^T & L_{11} \cdot L_{21}^T \\ L_{21} \cdot L_{11}^T & L_{21} \cdot L_{21}^T + L_{22} \cdot L_{22}^T \end{pmatrix}$$

se obtienen las igualdades bloque a bloque

$$(4.7) \quad \begin{aligned} L_{11} \cdot L_{11}^T &= L_0 \cdot L_0^T + Q_{21}^T \cdot Q_{21} \\ L_{21} \cdot L_{11}^T &= Q_{12}^T \cdot Q_{11} + Q_{22}^T \cdot Q_{21} \\ L_{21} \cdot L_{21}^T + L_{22} \cdot L_{22}^T &= Q_{12}^T \cdot Q_{12} + Q_{22}^T \cdot Q_{22} \end{aligned}$$

La primera ecuación de bloque es la que concentra el mayor peso del problema pero en realidad está ya casi resuelta pues el término $Q_{21}^T \cdot Q_{21}$ es apenas una pequeña alteración de $L_0 \cdot L_0^T$ que es fácilmente soluble tal y como se explica en [1, 2]. Las otras dos ecuaciones de bloque tienen poco coste computacional por ser de pequeño tamaño y su efecto se reduce a añadir nuevas filas a la matriz triangular según se expone en [3].

5. DISEÑO DEL MODELO

5.1. Ortonormalización de la geometría del problema. Si la matriz nodal X es de rango completo entonces existe un cambio de variable $C \in \mathbb{R}^{n \times n}$ tal que la nueva matriz nodal es ortonormal

$$(5.1) \quad \begin{aligned} \dot{X} &= X \cdot C \\ \dot{X}^T \cdot \dot{X} &= I \end{aligned}$$

Existen muchos métodos para calcular C pero en principio la forma más eficaz es mediante la descomposición de Cholesky

$$(5.2) \quad \begin{aligned} X^T \cdot X &= L \cdot L^T \\ C &= L^{-T} \\ \dot{X}^T \cdot \dot{X} &= L^{-1} \cdot X^T \cdot X \cdot L^{-T} = L^{-1} \cdot L \cdot L^T \cdot L^{-T} = I \end{aligned}$$

Si la matriz nodal es de rango $r < n$ entonces habría que buscar un cambio de variable $C \in \mathbb{R}^{n \times r}$ aplicando la descomposición de Jordan

$$(5.3) \quad \begin{aligned} X^T \cdot X &= V \cdot D^2 \cdot V^T \\ C &= V \cdot D^{-1} \\ \dot{X}^T \cdot \dot{X} &= D^{-1} \cdot V^T \cdot X^T \cdot X \cdot V \cdot D^{-1} = I \end{aligned}$$

En adelante supondremos que la matriz nodal ya ha sido ortonormalizada previamente.

5.2. Selección de la base de funciones de aproximación local. La base más habitual será en principio la base polinómica de grado r

n	r	\mathcal{P}	K
\dots	0	1	1
1	1	$1, x_1$	2
1	2	$1, x_1, x_1^2$	3
2	1	$1, x_1, x_2$	3
2	2	$1, x_1, x_2, x_1^2, x_1 x_2, x_2^2$	6
3	1	$1, x_1, x_2, x_3$	4
3	2	$1, x_1, x_2, x_3, x_1^2, x_1 x_2, x_2^2, x_2 x_3, x_3^2$	9

CUADRO 1. Dimensión de la base polinómica

Si se puede presumir algún tipo de periodicidad o pseudo-periodicidad en la función objetivo, entonces sería mejor usar una base de Fourier. Si se observan hay grandes saltos en poco espacio se pueden añadir componentes exponenciales. Por supuesto se pueden combinar cualesquiera de los antedichos o bien otros tipos de función, con la única condición de que sean linealmente independientes.

La dimensión K adecuada para la base de funciones de aproximación local dependerá de la dimensión del espacio y de lo suave o rugosa que sea la función. Si K empieza a no ser muy pequeño puede ser conveniente ortogonalizar la base aplicando la misma técnica del punto anterior a la matriz

$$(5.4) \quad X_{\mathcal{P}} = \begin{pmatrix} p_1(x_1) & \cdots & p_K(x_1) \\ \vdots & \vdots & \vdots \\ p_1(x_J) & \cdots & p_K(x_J) \end{pmatrix} \in \mathbb{R}^{J \times K}$$

5.3. Función de ponderación y geometría de los entornos locales. A continuación se va a construir una función de ponderación con las características especificadas anteriormente. En primer lugar se describirá una función de partida γ en el caso unidimensional que opera sobre entornos locales en forma de intervalo cerrado $[-1, 1]$, la cual se extenderá al caso multivariante por rotación y escalado, lo cual implica entornos locales hipersféricos de radio arbitrario.

Sea la función real escalar continua y doblemente diferencial en todo \mathbb{R}

$$(5.5) \quad \gamma(\tau) = \begin{cases} \left(\frac{1}{2} - \frac{1}{2}\tau\right)^4 \left(\frac{1}{2} + \frac{1}{2}\tau\right)^4 & \forall \tau \leq 1 \\ 0 & \forall \tau > 1 \end{cases}$$

correspondiente a la función de densidad de la distribución beta $\beta(4, 4)$ transportada al intervalo $[-1, 1]$ y extendida con ceros en el resto de \mathbb{R}

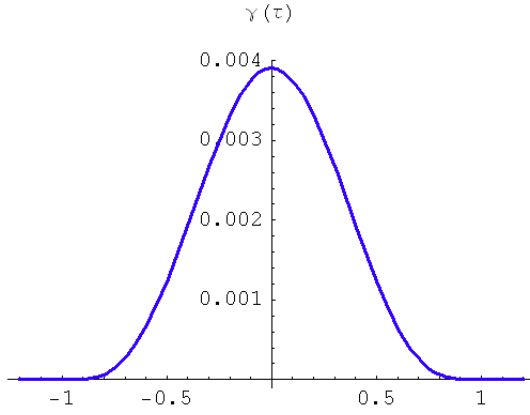


FIGURA 5.1.

Sus derivadas primera y segunda existen y son continuas en todo \mathbb{R}

$$(5.6) \quad \gamma' = \begin{cases} \frac{1}{32} \tau (\tau^2 - 1)^3 & \forall \tau \leq 1 \\ 0 & \forall \tau \notin \Omega_h \end{cases}$$

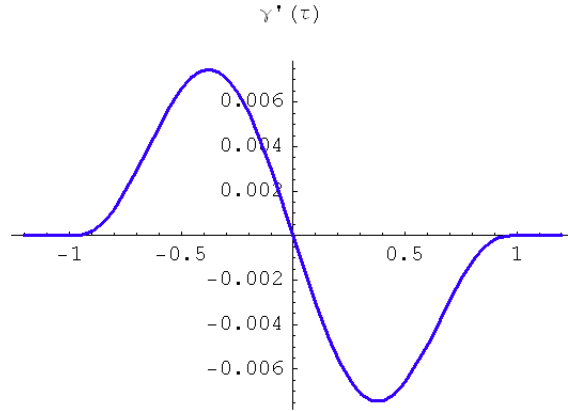


FIGURA 5.2.

$$(5.7) \quad \gamma'' = \begin{cases} \frac{1}{32} (\tau^2 - 1)^2 (7\tau^2 - 1) & \forall \tau \leq 1 \\ 0 & \forall \tau > 1 \end{cases}$$

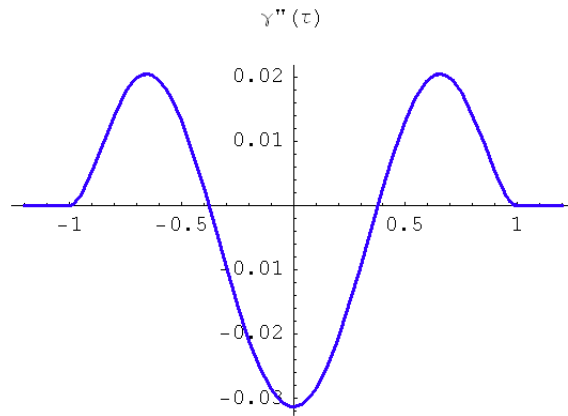


FIGURA 5.3.

A continuación se describirá la extrapolación de estos resultados al caso n -dimensional. Sólo es necesario tener en cuenta que un intervalo en la recta real no deja de ser una hipersfera de dimensión uno y, además, al estar ortonormalizado el problema la geometría de los entornos hipersféricos no penaliza ni favorece ninguna dirección en particular. Si los entornos locales se definen como hipersferas de radio δ_h centrados en el punto de referencia y_h

$$(5.8) \quad x \in \Omega_h \iff \tau_h(x) = \frac{1}{\delta_h} \|x - y_h\|_2 \leq 1$$

entonces es posible definir por rotación la siguiente función de cercanía local

$$(5.9) \quad \lambda_h(x) = \gamma(\tau_h(x)) = \begin{cases} \left(\frac{1}{2} - \frac{1}{2}\tau_h(x)\right)^4 \left(\frac{1}{2} + \frac{1}{2}\tau_h(x)\right)^4 & \forall x \in \Omega_h \\ 0 & \forall x \notin \Omega_h \end{cases}$$

Esta función es continua y doblemente diferenciable, y es no negativa en el entorno local y nula fuera de él y tiene un único máximo en el punto de referencia

$$(5.10) \quad \begin{aligned} \lambda_h(x) &\geq 0 \forall x \in \Omega_h \\ \lambda_h(x) &= 0 \forall x \notin \Omega_h \\ \lambda_h(x) &< \lambda_h(y_h) \forall x \neq y_h \end{aligned}$$

Resulta evidente que las siguientes funciones cumplen todas las condiciones de las funciones de ponderación

$$(5.11) \quad w_h(x) = \begin{cases} \frac{\lambda_h(x)}{\sum_{x \in \Omega_s} \lambda_s(x)} & \forall x \in \Omega_h \\ 0 & \forall x \notin \Omega_h \end{cases}$$

Todo lo anterior seguiría siendo cierto partiendo de cualquier distribución beta simétrica de orden entero mayor o igual que 4 pero no con orden 3 pues en ese caso la segunda derivada no sería continua en la frontera. En la gráfica siguiente se muestra cómo la función de cercanía se hace más concentrada en el punto de referencia cuanto más grande es el orden de la beta simétrica.

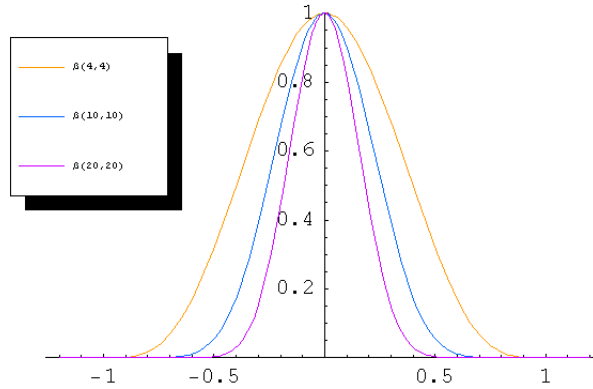


FIGURA 5.4.

A continuación se ilustrará lo anterior con un ejemplo unidimensional con tres puntos de referencia y base polinómica de grado 2 y se supondrá que ya se ha efectuado la estimación de la regresión lineal.

h	y_h	δ_h	$g_h(x)$
1	-3	6	$1 + 2(x - y_1) - (x - y_1)^2$
2	1	3,5	$3 - 4(x - y_2) + (x - y_2)^2$
3	5	5	$-5 + 6(x - y_3) + 2(x - y_3)^2$

CUADRO 2. Datos del ejemplo 1

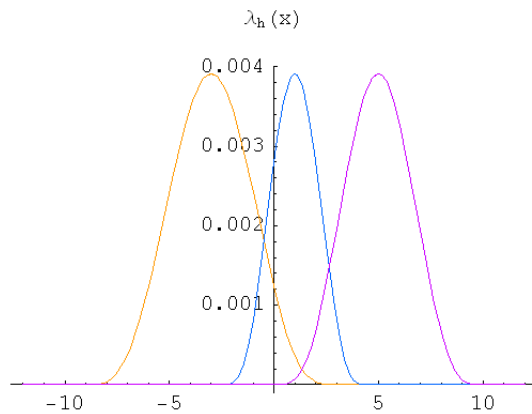


FIGURA 5.5.

Funciones de cercanía local

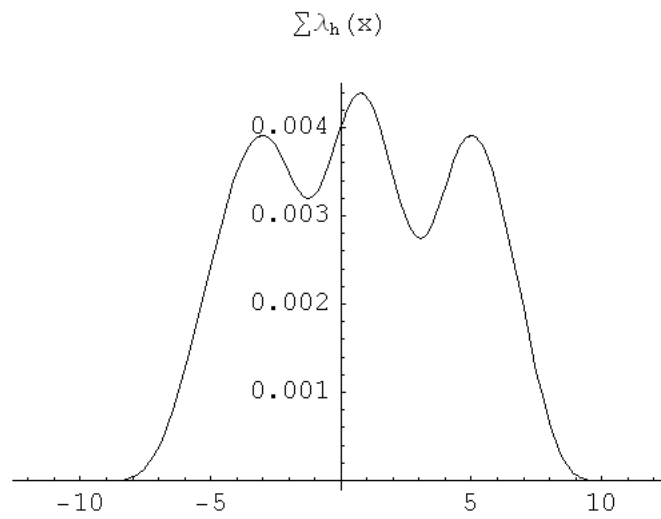


FIGURA 5.6.

Sumatorio global de las funciones de cercanía

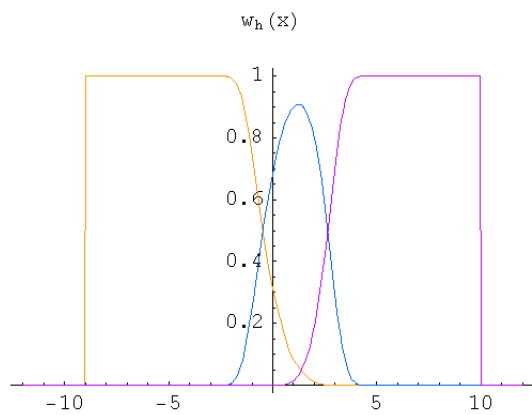


FIGURA 5.7.

Funciones de ponderación

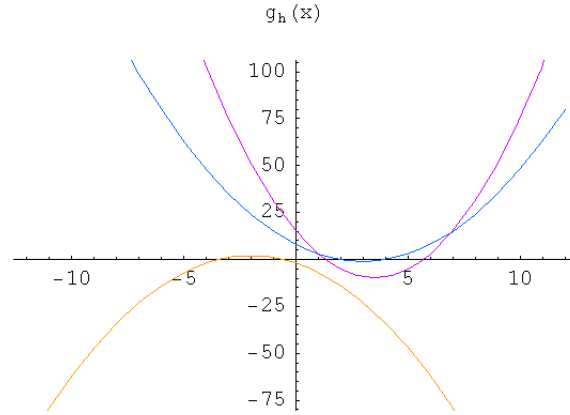


FIGURA 5.8.

Funciones de aproximación local

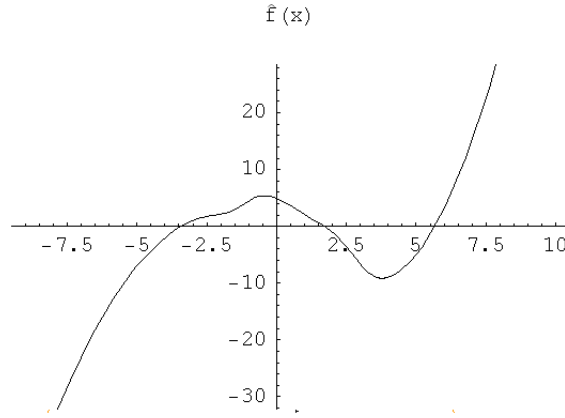


FIGURA 5.9.

Función de aproximación global

5.4. Diseño del recubrimiento. Si se coloca un punto de referencia en cada punto nodal y no se usan puntos de estrés entonces estaríamos ante el caso particular de aproximación de funciones sin malla, aunque dado el sistema de estimación no aportaría ninguna ganancia sino más bien al contrario: de una buena selección de la malla puede depender la calidad y la eficiencia de la aproximación.

Para que la regresión lineal no sea degenerada debe tener más ecuaciones que variables

$$(5.12) \quad J + T \geq H \cdot K$$

De hecho, para que se pueda hacer inferencia la superficie de contraste ha de ser razonablemente grande

$$(5.13) \quad SC = \frac{J+T}{H \cdot K} \gg 1$$

Para aumentar la superficie de contraste sin rebajar el tamaño de la base ni el número de puntos de referencia se deben añadir más puntos de estrés. En las intersecciones entre pares de entornos locales en los que haya algún punto nodal lo más prudente sería que estos ejercieran a su vez de puntos de estrés. En aquellas en las que no los hubiera sería recomendable introducirlos de forma artificial.

Cuanto menos solapamiento haya entre las regiones de entorno local más dispersa será la matriz de regresión y los cálculos podrán ejecutarse en menos tiempo, pero menos suave será la transición entre ellos.

Lo ideal sería que cada entorno de referencia encerrara aproximadamente el mismo número de puntos nodales para que el contraste de sus parámetros fuera del mismo rango para todos ellos. Si los puntos nodales se hallan distribuidos más o menos uniformemente la forma de malla regular es la manera más eficaz de colocar los puntos de referencia y sus entornos para recubrir una región del espacio con el mínimo solpamiento.

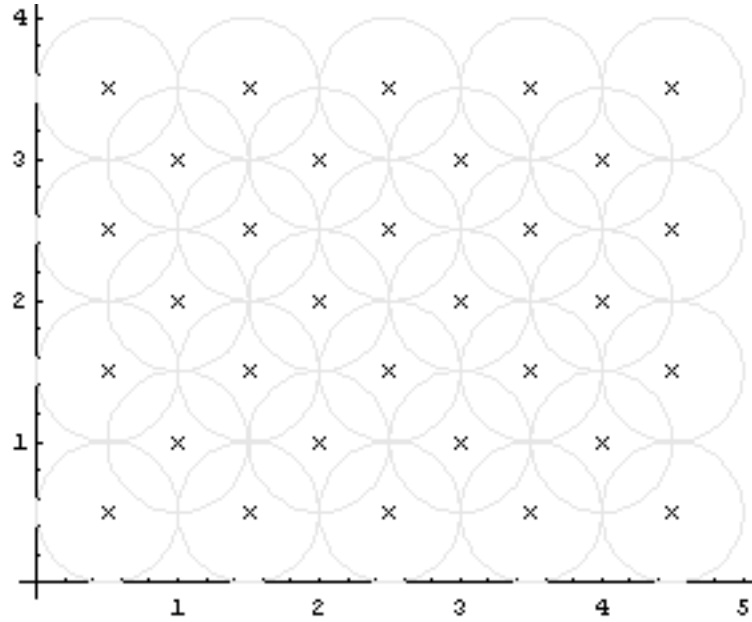


FIGURA 5.10.

Recubrimiento bidimensional regular

Cuando la densidad de puntos nodales no es homogénea lo más prudente es adaptar el recubrimiento a las circunstancias. Resulta relativamente sencillo pasar de un recubrimiento de un radio dado a otro de la mitad de radio de una forma bastante eficiente.

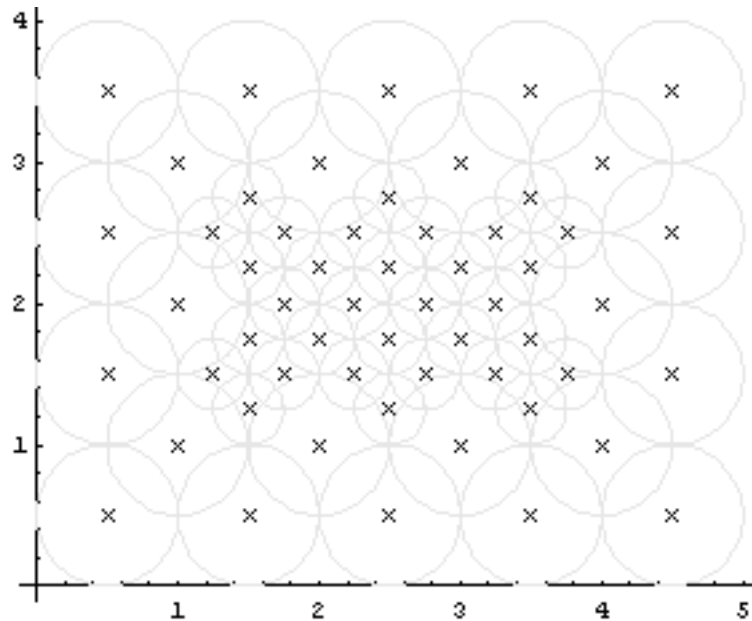


FIGURA 5.11.

Recubrimiento bidimensional irregular

Una forma alternativa y más sencilla de homogeneizar el recubrimiento es aplicar una función continua y doblemente diferenciable de deformación del espacio que convierta los puntos originales en puntos cuya distribución sea aproximadamente uniforme

$$(5.14) \quad \eta(x) : \Omega \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

REFERENCIAS

- [1] Modifying a Sparse Cholesky factorization. T. A. Davis, J. R. Gilbert, S. I. Larimore, and E. G. Ng. SIAM J. Matrix Anal. Applic. 20(3):606–627, 1999.
- [2] Multiple-rank modifications of a sparse Cholesky factorization. T. A. Davis and W. W. Hager. SIAM J. Matrix Anal. Applic., 22(4):997–1013, 2001.
- [3] Row modifications of a sparse Cholesky factorization. SIAM J. Matrix Anal. Applic. 26(3):621–639, 2005.
- [4] Meshless methods: An overview and recent developments. T. Belytschko, Y. Krongauz, D. Organ, M. Fleming and P. Krysl. doi:10.1016/S0045-7825(96)01078-X

E-mail address: `vdebuen@tol-project.org`