APROXIMACIÓN DE FUNCIONES MULTIVARIANTES MEDIANTE REGRESIÓN LINEAL JERÁRQUICA

VÍCTOR DE BUEN REMIRO

RESUMEN. 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.

1. ANTECEDENTES

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

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

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

$$\mathcal{X} = \{x_j \,|x_j \in \Omega \subset \mathbb{R}^n \forall j = 1 \dots J\} \tag{1.2}$$

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.

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

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

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

$$\begin{gathered} \mathcal{Y} = \{y_h \,|y_h \in \Omega_h \subset \Omega \subset \mathbb{R}^n \forall h = 1 \dots H\} \\ \Omega \subset \bigcup_{h=1}^{H} \Omega_h \end{gathered} \tag{2.1}$$

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

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

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

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

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

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

mediante unas funciónes 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

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

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

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

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

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

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

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

donde

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

Si definimos

$$\begin{gathered} \beta_i = \alpha_{h,k} \\ Q_{i,j} = P_{h,k,j} \end{gathered} \tag{2.10}$$

es posible escribir la regresión en forma matricial

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

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

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

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 stress en los que el valor de la función objetivo es puede ser conocido o no.

$$\begin{gathered} \mathcal{Z}_{h,s} = \{z_{h,s,t} \,|\, z_{h,s,t} \in \Omega_h \cap \Omega_s \subset \Omega \subset \mathbb{R}^n \forall t = 1 \dots T_{h,s}\} \\ T = \sum_{h,s}^{\Omega_h \cap \Omega_s \neq \emptyset} T_{h,s} \end{gathered} \tag{3.1}$$

entonces podemos añadir las T ecuaciones latentes al modelo definido anteriormente

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

4. Diseño del modelo

4.1. Ortonormalización de la geometría del problema. La matriz nodal X es de rango completo por lo que existe un cambio de variable C tal que la nueva matriz nodal es ortonormal

$$\begin{gathered} \dot{X} = X \cdot C \\ \dot{X}^T \cdot \dot{X} = I \end{gathered} \tag{4.1}$$

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

$$\begin{gathered} X^T \cdot X = L \cdot L^T \\ C = L^{-T} \\ \dot{X}^T \cdot \dot{X} = L^{-1} \cdot X^T \cdot X \cdot L^{-T} = L^{-1} \cdot L \cdot L^T \cdot L^{-T} = I \end{gathered} \tag{4.2}$$

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

4.2. 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
$\cdots$	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	$1, x_1, x_2, x_1^2, x_1x_2, x_2^2$	6
3	1	$1, x_1, x_2, x_3$	4
3	2	$1, x_1, x_2, x_3, x_1^2, x_1x_2, x_2^2, x_2x_3, x_3^2$	9

Cuadro 1. Dimensión de la base polinómica

Si se puede presumir algun 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 adecuada K de 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

$$X_{\mathcal{P}} = \begin{pmatrix} p_1(x_1) & \cdots & p_K(x_1) \\ \vdots & \vdots & \vdots \\ p_1(x_J) & \cdots & p_K(x_J) \end{pmatrix} \in \mathbb{R}^{J\times K} \tag{4.3}$$

4.3. Función de ponderación y geometría de los entornos locales. Si para cada punto de referencia se define una función de cercanía local, continua y doblemente diferenciable, no negativa en el entorno local y nula fuera de él y con un único máximo en el punto de referencia

$$\begin{gathered} \lambda_h(x) \geq 0 \forall x \in \Omega_h \\ \lambda_h(x) = 0 \forall x \notin \Omega_h \\ \lambda_h(x) < \lambda_h(y_h) \, \forall x \neq y_h \end{gathered} \tag{4.4}$$

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

$$w_h(x) = \begin{cases} \frac{\lambda_h(x)}{\sum\limits_{x\in\Omega_s} \lambda_s(x)} & \forall x \in \Omega_h \\ 0 & \forall x \notin \Omega_h \end{cases} \tag{4.5}$$

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

$$\lambda(\tau) = \begin{cases} \left(\frac{1}{2} - \frac{1}{2}\tau\right)^4 \left(\frac{1}{2} + \frac{1}{2}\tau\right)^4 & \forall \tau \leq 1 \\ 0 & \forall \tau > 1 \end{cases} \tag{4.6}$$

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 $\mathbb{R}$

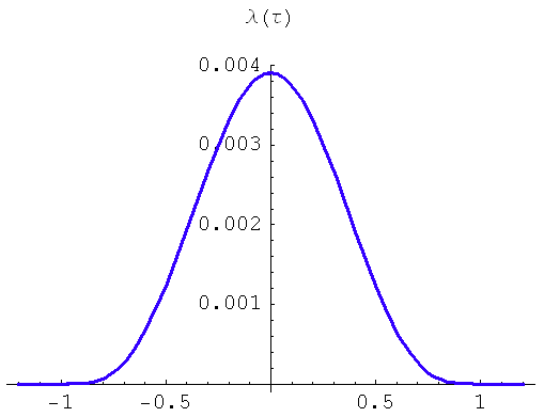

FIGURA 4.1.

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

$$\lambda' = \begin{cases} \frac{1}{32}\tau\left(\tau^2 - 1\right)^3 & \forall \tau \leq 1 \\ 0 & \forall x \notin \Omega_h \end{cases} \tag{4.7}$$

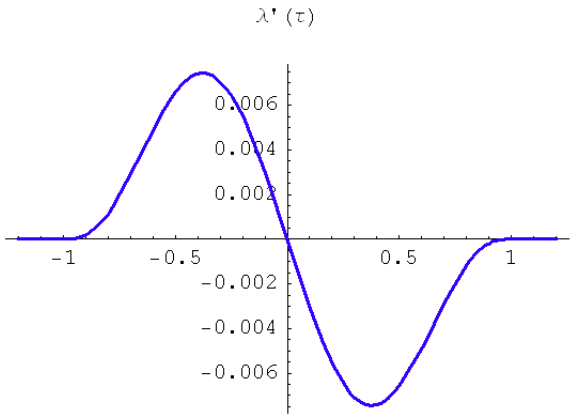

FIGURA 4.2.

$$\lambda'' = \begin{cases} \frac{1}{32}\left(\tau^2-1\right)^2\left(7\tau^2-1\right) & \forall \tau \leq 1 \\ 0 & \forall \tau > 1 \end{cases} \tag{4.8}$$

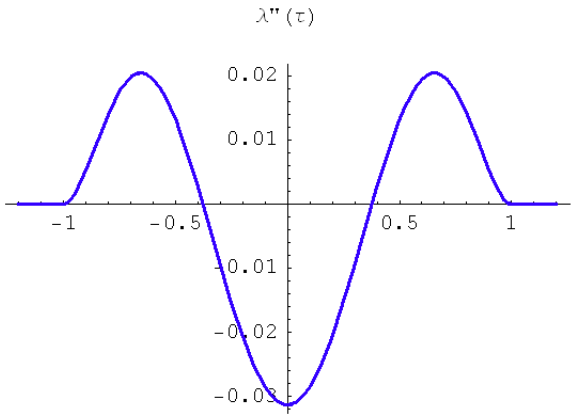

FIGURA 4.3.

Si los entornos locales se definen como hiperesferas de radio δ_h centrados en el punto de referencia y_h

$$x \in \Omega_h \Longleftrightarrow \tau_h(x) = \tfrac{1}{\delta_h}\left\|x - y_h\right\|_2 \leq 1 \tag{4.9}$$

entonces la función

$$\lambda_h(x) = \lambda\left(\tau_h(x)\right) = \begin{cases} \left(\frac{1}{2} - \frac{1}{2}\tau_h(x)\right)^4\left(\frac{1}{2} + \frac{1}{2}\tau_h(x)\right)^4 & \forall x \in \Omega_h \\ 0 & \forall x \notin \Omega_h \end{cases} \tag{4.10}$$

cumple las condiciones de función de cercanía local.

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.

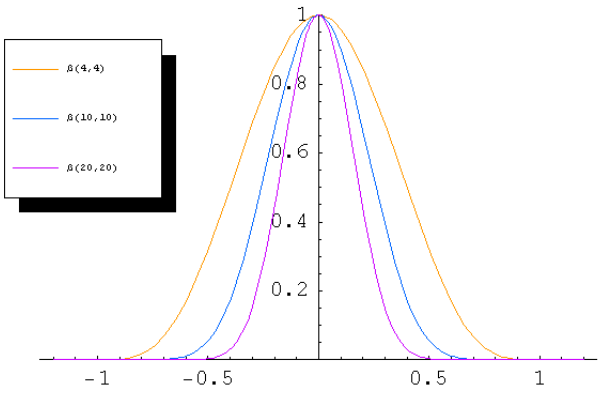

FIGURA 4.4.

Obsérvese por útimo que al estar ortonormalizado el problema la geometría de los entornos hiperesféricos no penaliza ni favorece ninguna dirección en particular.

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

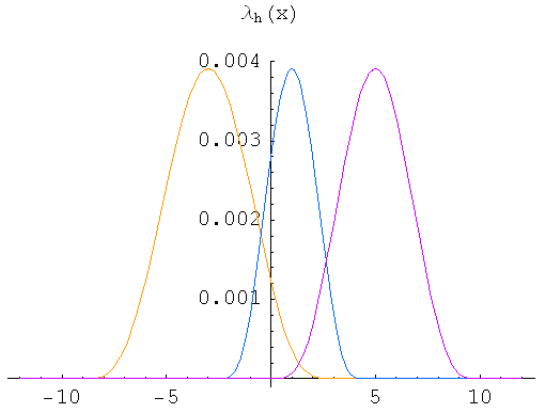

FIGURA 4.5.
Funciones de cercanía local

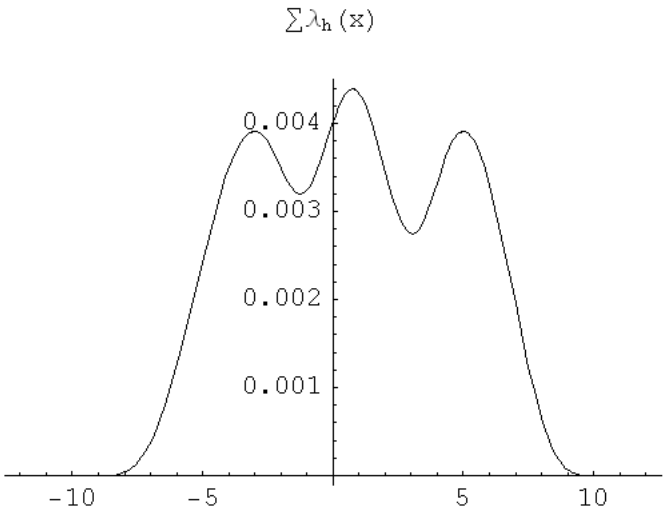

FIGURA 4.6. Sumatorio global de las funciones de cercanía

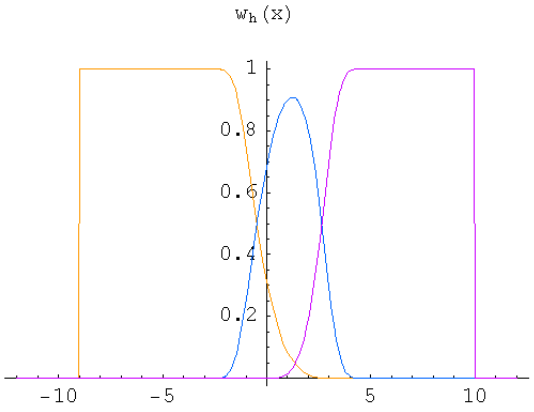

FIGURA 4.7. Funciones de ponderación

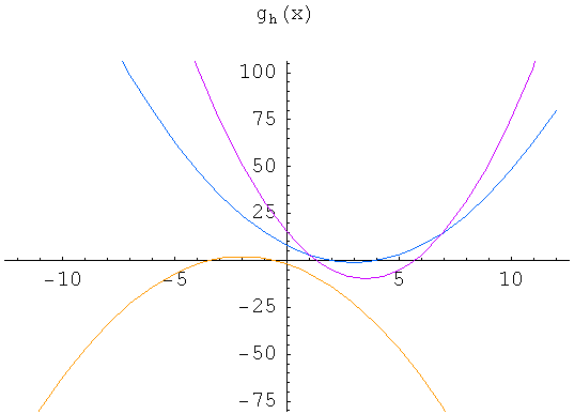

FIGURA 4.8. Funciones de aproximación local

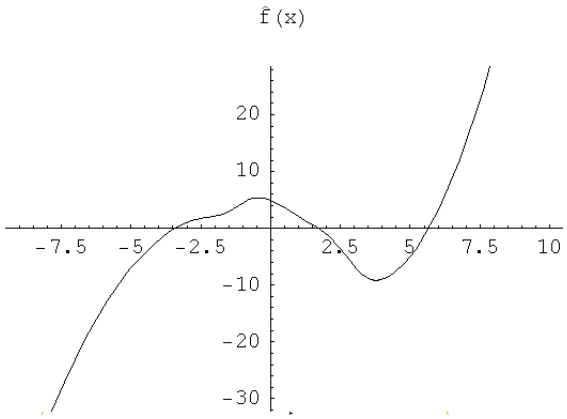

FIGURA 4.9. Función de aproximación global

4.4. Diseño del recubrimiento. Para que la regresión lineal no sea degenerada debe tener más ecuaciones que variables

$$J + T \geq H \cdot K \tag{4.11}$$

Como la base no puede ser vacía es $K \geq 1$, y el número de puntos de referencia no puede exceder al de nodos

$$H \leq J + T \tag{4.12}$$

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

$$SC = \frac{J+T}{H \cdot K} \gg 1 \tag{4.13}$$

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

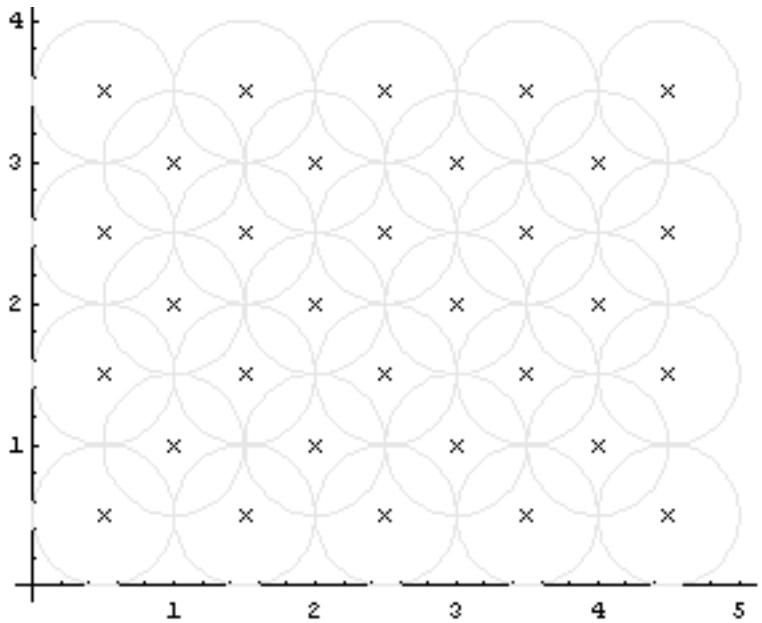

FIGURA 4.10. Recubrimiento bidimensional regular

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.

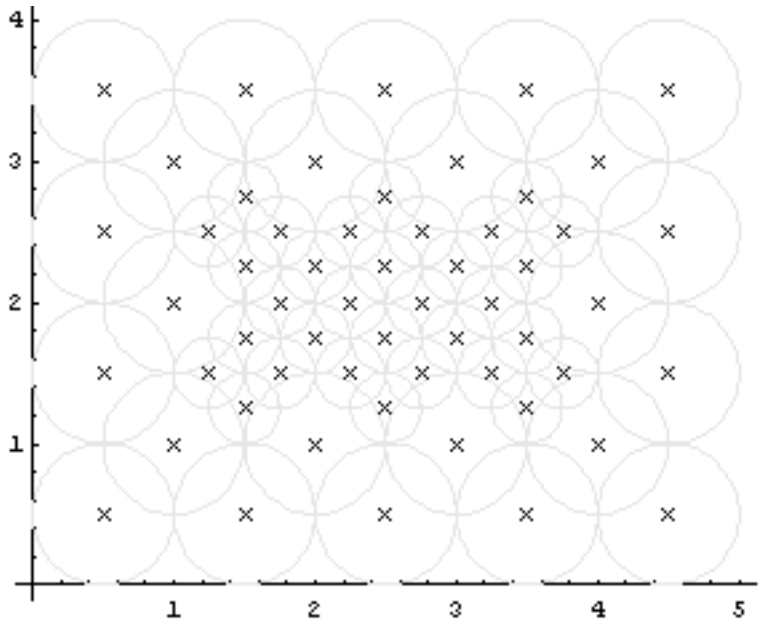

FIGURA 4.11. Recubrimiento bidimensional irregular

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

$$\eta(x) : \Omega \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n \tag{4.14}$$

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