Literature review

Profundidad de datos funcionales

Diego F. Fonseca V. & Rodolfo A. Quintero *

Resumen En el presente documento expondremos algunas de los diferentes versiones de profundidad para datos funcionales, se presentan sus definiciones y su relación con la noción de orden en datos funcionales, lo que permite definir aspectos como la media poblacional y algunos estimadores de ésta. Además, incluimos algunos teoremas referentes a la consistencia de dichos estimadores.

Palabras clave: Profundidad, datos funcionales, α-trimmed mean, semi-región.

1. Introducción

Un dato funcional se entiende como una curva continua que describe la realización de un evento en el tiempo, esta es una descripción escueta, siendo formales un conjunto de datos funcionales se puede entender como un conjunto $X_1(t), X_2(t), \ldots, X_n(t)$ de procesos estocásticos con trayectorias continuas en un intervalo compacto I, es decir, cerrado y acotado, lo que lleva a considerarlo de la forma $[a, b]$. Si graficamos sus trayectorias estas determinan un conjunto de curvas, el objetivo es ver cuál de estas curvas permanece en el medio de todas las demás por el mayor tiempo posible. Dicha curva representa una especie de media pero en este caso para datos funcionales, otra interpretación es que si logramos distinguir dicha curva respecto a las otras se observa que esta es la más profunda, entendiendo profunda como estar más hacia al çentro"(dependerá mucho de la definición que se use de centro). Además, a partir del concepto de profundidad se puede inducir una noción de orden en los datos funcionales: estos se ordenan desde el mas profundo al menos profundo. Dicha noción de orden es importante ya que permite definir estadísticos basados en el orden de los datos como ocurre con los datos unidimensionales.

El presente documento se divide en secciones, cada sección aborda un tipo de profundidad para datos funcionales, estas profundidades fueron propuestas en [1], [2] y [4]. En la Sección 2 se define el concepto de profundidad para datos funcionales como una integral y a partir de esta profundidad se establece un análogo de α-trimmed mean (media truncada) pero en este caso para datos funcionales. En la Sección 3 se aborda un nuevo concepto de profundidad basado en las bandas que generan subconjuntos de la muestra funcional, estas regiones permiten establecer una noción de profundidad basada en el tiempo que permanece una trayectoria dentro de la banda. Por ultimo, la Sección 4 el último concepto de profundidad se basa en el concepto de semi-región generado por una curva.

2. Profundidad integral de datos funcionales

La profundidad para datos funcionales que definiremos se basa en cualquier profundidad de datos unidimensionales que se quiera emplear, en ese sentido, antes de introducir dicha profundidad es importante conocer algunas profundidades de datos no funcionales y sus versiones unidimensionales. En concreto abordaremos tres tipos de profundidad no funcionales que son la *Simplicial*, la de *Tukey* y *estándar unidimensional*:

* Universidad de los Andes,Bogotá, Colombia.

Profundidad de Tukey: Considerando $Y_1, Y_2, \ldots, Y_n$ variables aleatorias independientes e idénticamente distribuidas con distribución F donde $Y_i \in \mathbb{R}^k$. La profundidad de Tukey en x es definida por

$$TD(x) = \text{ínf}\left\{F(H) \mid H \text{ semiespacio en } \mathbb{R}^k \text{ tal que } x \in H\right\}.$$

Si la distribución es la empírica, es decir, F_n entonces la profundidad de Tukey se define de la misma manera y se denota por TD_n. En el caso unidimensional, cuando $k = 1$, se puede inferir que

$$TD(x) = \text{mín}\left\{F(x), 1 - F(x^-)\right\}.$$

Profundidad simplicial: En este caso se consideran $Y_1, Y_2, \ldots, Y_{k+1}$ variables aleatorias independientes e idénticamente distribuidas con distribución F donde $Y_i \in \mathbb{R}^k$. La profundidad simplicial en x es definida por

$$SD(x) = \mathbb{P}_F\left(x \in S[Y_1, \ldots, Y_{k+1}]\right)$$

donde $S[Y_1, \ldots, Y_{k+1}]$ es simplejo cerrado con vértices en $Y_1, \ldots, Y_{k+1}$, es decir

$$S[Y_1, \ldots, Y_{k+1}] := \left\{ \sum_{i=1}^{k+1} t_i y_i \;\middle|\; \sum_{i=1}^{k+1} t_i = 1 \text{ y } t_i \geq 0 \ \forall i = 1, \ldots, k+1 \right\}$$

Si la distribución es la empírica, es decir, F_n entonces la profundidad Simplicial se define de la misma manera y se denota por SD_n. En el caso unidimensional, cuando $k = 1$, se puede inferir que

$$SD(x) = 2F(x)\left(1 - F(x^-)\right).$$

Profundidad estándar unidimensional: Dados $Y_1, \ldots, Y_n$ variables aleatorias unidimensionales independientes e idénticamente distribuidas con distribución F, se define la profundidad estándar unidimensional en x como

$$UD(x) = 1 - \left|\frac{1}{2} - F(x)\right|.$$

Si la distribución es la empírica, es decir, F_n entonces la profundidad estándar unidimensional se define de la misma manera y se denota por UD_n.

De las anteriores profundidades nos interesa sus versiones unidimensionales, por tanto, consideramos nuestro conjunto de datos funcionales como $X_1(t), X_2(t), \ldots, X_n(t)$ independientes e idénticamente distribuidos procesos estocásticos con distribución F_t y con trayectorias continuas en un intervalo compacto I, es decir, cerrado y acotado, lo que lleva a considerarlo de la forma $[a, b]$. Recordemos que un proceso estocástico se define sobre un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ donde $X_i(t) : \Omega \to \mathbb{R}$ es una variable aleatoria, de modo que cuando nos referimos a una trayectoria estamos haciendo referencia a la curva $\gamma_i(t) = X_i(t)(\omega)$ para algún $\omega \in \Omega$. En adelante, $X_i(t)$ se interpretará como curvas sobreentendiendo que esta es la variable aleatoria $X_i(t)$ evaluada en algún $\omega \in \Omega$ (el mismo ω para toda la muestra funcional). La versión empírica de F_t es notada por $F_{n,t}$. Sea D_t una profundidad unidimensional, que puede ser una de las versiones unidimensionales de TD, SD ó UD definidas para la muestra unidimensional $X_1(t), X_2(t), \ldots, X_n(t)$ (es unidimensional pues t esta fijo y son idénticamente distribuidas con distribución F_t). Definimos para cualquier curva x continua en $[a, b]$ lo siguiente

$$I(x) := \int_a^b D_t(x(t))dt$$

A este último valor es a lo que llamamos *profundidad integral* en x, originalmente fue presentada en [1] con el nombre de *functional depth* pero creemos conveniente llamarla profundidad integral.

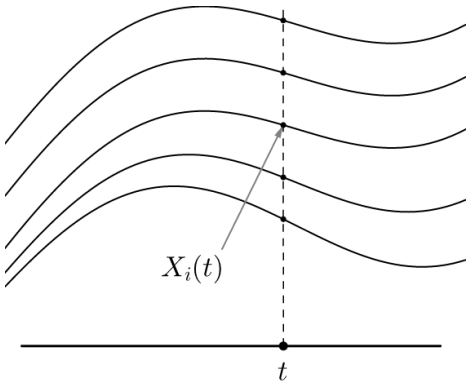


Figura 2.1. Para t fijo los procesos estocásticos forman una muestra unidimensional.

Al igual que con las profundidades no funcionales, la profundidad integral también tiene su versión empírica que se define a partir de la versión empírica de D ($D_{n,t}$), de modo que la versión empírica de I se denota por I_n. Dado que las curvas de interés son las trayectorias $X_i(t)$, entonces para un tiempo t se acostumbra a notar

$$Z_i(t) := D_t(X_i(t)).$$

Si $Z_k(t)$ es el mas grande entre $Z_1(t), \ldots, Z_n(t)$, entonces $X_k(t)$ es el dato más profundo en el tiempo t. Como las trayectorias X_i son continuas entonces Z_i es continua en $[a, b]$, por lo tanto, existe su integral $I(X_i)$. Para efectos de notación se acostumbra a escribir

$$I_i := I(X_i) = \int_a^b D_t(X_i(t))dt = \int_a^b Z_i(t)dt.$$

Dicha profundidad permite inducir un orden en los procesos estocásticos $X_1(t), X_2(t), \ldots, X_n(t)$ desde el más profundo (valores grandes de I_i) hasta el más externo (valores pequeños de I_i). Cuando la distribución no es conocida y no se quieren hacer suposiciones sobre ella se acostumbra a usar la versión empírica de I, I_n, que se fundamenta en las formas empíricas de F_t.

2.1. α-trimmed mean

En el contexto unidimensional es común que dada una muestra existan datos que se resaltan por su lejanía respecto a los demás. En el caso funcional dicha situación también se presenta y estos datos pueden influir en cualquier inferencia que se haga sobre la población. En cualquier caso este aspecto se trata por medio de la profundidad, centrando la atención en un porcentaje de los datos mas profundos. Así nace el concepto de α-trimmed mean (media truncada), un concepto conocido en el caso unidimensional como un promedio sobre los datos que resultan al excluir un porcentaje de los datos mas pequeños y el mismo porcentaje de los datos mas grandes. Siendo formales, si $X_1, \ldots, X_n$ es una muestra unidimensional y su orden es $X_{(1)} < \cdots < X_{(n)}$, entonces se define su α-*trimmed mean* en el caso unidimensional como

$$\alpha - \text{trimmed mean} := \frac{1}{n - 2\lfloor n\alpha \rfloor} \sum_{j=\lfloor n\alpha \rfloor + 1}^{n - \lfloor n\alpha \rfloor} X_{(j)}.$$

Ahora, si consideramos datos funcionales $X_1(t), X_2(t), \ldots, X_n(t)$, la *versión funcional* de α-trimmed mean, notada por μ_n, es dada por

$$\mu_n := \frac{\sum_{i=1}^{n} \mathbb{1}_{[\beta,\infty)} I(X_i) X_i}{\sum_{i=1}^{n} \mathbb{1}_{[\beta,\infty)} I(X_i)}.$$

Como no siempre se tiene acceso a la distribución original de los datos, una buena aproximación de μ_n es un *avance*: dicha aproximación se basa en el hecho que las distribuciones empíricas convergen a la distribución de la muestra de las cuales fueron generadas, a partir de este hecho se define el estimador de la media podada (α-trimmed mean) como

$$\widehat{\mu}_n := \frac{\sum_{i=1}^{n} \mathbb{1}_{[\beta,\infty)} I_n(X_i) X_i}{\sum_{i=1}^{n} \mathbb{1}_{[\beta,\infty)} I_n(X_i)}.$$

Y la *trimmed mean* poblacional es dada por

$$\mu := \frac{\mathbb{E}\left[X_1 \mathbb{1}_{[\beta,\infty)}(X_1)\right]}{\mathbb{E}\left[\mathbb{1}_{[\beta,\infty)}(X_1)\right]}.$$

Esta última se define en X_1, pero se podría definir en cualquier X_i pues estamos asumiendo que los procesos estocásticos son i.i.d. Pensando no-paramétricamente frecuentemente no se conoce μ y tampoco se conoce la distribución que rige cada proceso estocástico, lo que permitiría exhibir explicitamente μ, de modo que un objetivo menos ambicioso pero alcanzable es estimar μ.

2.2. Aspectos de convergencia y consistencia

Bajo ciertas condiciones I_n y $\widehat{\mu}_n$ convergen a I y μ respectivamente, esto es consignado en los siguientes teoremas:

Teorema 1. *Asumiendo las siguientes condiciones:*

(i) Si las trayectorias del proceso estocástico $X_1(t)$ pertenecen a $\mathsf{Lip}_A[a,b]$ para una constante A (suficientemente grande) para todo $i = 1, \ldots, n$ donde

$$\mathsf{Lip}_A[a,b] := \{x : [a,b] \to \mathbb{R} \mid x \text{ función Lipschitz con constante menor o igual a } A\}.$$

(ii) Existe una constante $c > 0$ tal que

$$\mathbb{E}\left[\lambda\left(t \mid X_1 \in [u(t), u(t) + \varepsilon c]\right)\right] \leq \frac{\varepsilon}{2}$$

para todo $u \in \mathsf{Lip}_A[a,b]$ donde λ es la medida de Lebesgue en $\mathbb{R}$.

Si definimos

$$J_n(x) := \int_a^b F_{n,t}(x(t))dt \qquad y \qquad J(x) := \int_a^b F_t(x(t))dt.$$

Entonces tenemos

$$\lim_{n\to\infty} \sup_{x \in \mathsf{Lip}_A[a,b]} |J_n(x) - J(x)| = 0 \quad \mathbb{P} - a.s.$$

y

$$\lim_{n\to\infty} \sup_{x \in \mathsf{Lip}_A[a,b]} |I_n(x) - I(x)| = 0 \quad \mathbb{P} - a.s.$$

El Teorema 1 junto con el siguiente teorema demuestran que $\widehat{\mu}_n$ estima μ.

Teorema 2. *Si las trayectorias del proceso estocástico $X_1(t)$ pertenecen a un espacio arbitrario $\mathcal{E}[a,b]$ de curvas en $[a,b]$, y en dicho espacio se satisface*

$$\lim_{n\to\infty} \sup_{x\in\mathcal{E}[a,b]} |I_n(x) - I(x)| = 0 \quad \mathbb{P}-a.s.$$

entonces

$$\widehat{\mu}_n \longrightarrow \mu \quad \text{como } n \to \infty.$$

En particular, bajo las condiciones del Teorema 1 se obtiene $\widehat{\mu}_n \longrightarrow \mu$.

3. Profundidad por medio de bandas

Una desventaja del concepto de profundidad integral es que no tiene en cuenta la forma de los datos funcionales. En una muestra de datos funcionales pueden existir datos, en este caso curvas, que exhiben una forma que pareciera no seguir la tendencia de los demás datos funcionales, por ejemplo, datos que a priori son curvas diferenciables pero existe una curva que no lo es y exhibe picos pronunciados, lo que se puede llamar un dato contaminado, el inconveniente radica en que dicha curva excepcional podría ser la mas profunda lo cual puede no ser una buena conclusión, es decir, podría ser la media maestral respecto a dicha profundidad. Esto hace muy importante contar con un concepto de profundidad de datos funcionales que contemple la forma de las curvas y no le de demasiado peso a las curvas excepcionales, ese es el objetivo de esta sección.

Consideremos $\mathcal{C}(I)$ el conjunto de las curvas continuas en un intervalo compacto I, para $x \in \mathcal{C}(I)$ definimos el *grafo* de x como

$$\Gamma(x) := \{(t, x(t)) \mid t \in I\}.$$

Lo que permite definir el conjunto de grafos

$$\Gamma(\mathcal{C}(I)) := \{\Gamma(x) \mid x \in \mathcal{C}(I)\}.$$

En dicho conjunto definimos la función indicadora para un conjunto $A \subset I \times \mathbb{R}$ como $\mathbb{1}_A : \Gamma(\mathcal{C}(I)) \to \mathbb{R}$ dada por

$$\mathbb{1}_A(\Gamma(x)) = \begin{cases} 1 & \text{Si } \Gamma(x) \subseteq A \\ 0 & \text{Si } \Gamma(x) \nsubseteq A \end{cases}.$$

Los últimos tres conceptos expuestos junto con el concepto de banda que presentaremos a continuación son indispensables para definir el nuevo concepto de profundidad que queremos presentar.
Dadas las curvas $x_1, x_2, \ldots, x_n$ en $\mathcal{C}(I)$ para $k \leq n$ se define la banda generada por las curvas $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ como

$$\begin{aligned}\mathcal{B}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) &:= \left\{ (t,y) \,\middle|\, t \in I \text{ y } \min_{r=1,\ldots,k} x_{i_k} \leq y \leq \max_{r=1,\ldots,k} x_{i_k} \right\} \\ &= \left\{ (t,y) \,\middle|\, \exists \alpha \in [0,1] \ \forall t \in I, \ y(t) = \alpha \left(\min_{r=1,\ldots,k} x_{i_k} \right) + (1-\alpha) \left(\max_{r=1,\ldots,k} x_{i_k} \right) \right\}\end{aligned}$$

Por lo tanto, para $2 \leq j \leq n$ y $x \in \mathcal{C}(I)$ definimos

$$\mathcal{B}D_n^{(j)}(x) := \frac{1}{\binom{n}{j}} \sum_{1\leq i_1 < \cdots < i_k \leq n} \mathbb{1}_{\mathcal{B}(x_{i_1},\ldots,x_{i_k})}(\Gamma(x)). \tag{1}$$

Este último termino promedia el numero de bandas que contienen al grafo de x, las cuales son generadas por subconjuntos de j curvas de las n curvas iniciales.

Definición 1. *Para $J \in \mathbb{N}$ tal que $2 \leq J \leq n$ y n curvas $x_1, \ldots, x_n$ en $\mathcal{C}(I)$ se define la* ***profundidad por bandas muestral*** *en la curva x (en $\mathcal{C}(I)$) como*

$$\mathcal{BD}_{n,J}(x) := \sum_{j=2}^{J} \mathcal{BD}_n^{(j)}(x).$$

El término *muestral* en esta última definición tiene mucho sentido cuando nos ubicamos en el contexto estocástico, es decir, cuando consideramos $X_1(t), X_2(t), \ldots, X_n(t)$ procesos estocásticos i.i.d cuyas trayectorias pertenecen a $\mathcal{C}(I)$. Una muestra de este conjunto de procesos estocásticos son realizaciones de los mismos, estas realizaciones son trayectorias o curvas, así pues, la *profundidad por bandas* $\mathcal{BD}_{n,J}$ de una curva x respecto al conjunto de procesos estocásticos $X_1(t), X_2(t), \ldots, X_n(t)$ es la dada en la definición 1, pero, haciendo cada x_i como la trayectoria asociada al proceso $X_i(t)$. En ese orden de ideas, se observa que en la definición $\mathcal{BD}_{n,J}(x)$ no se emplea la distribución de los procesos estocásticos, lo que la hace un término eminentemente muestral.

Es natural que exista una versión de $\mathcal{BD}_{n,J}$ poblacional, es decir, que dependa de la distribución de los procesos estocásticos $X_1(t), X_2(t), \ldots, X_n(t)$. En efecto, la expresión (1) también tiene una versión poblacional dada por

$$\begin{aligned} \mathcal{BD}^{(j)}(x) &:= \mathbb{P}\left(\Gamma(x) \subset \mathcal{B}\left(X_1, \ldots, X_j\right)\right) \\ &= \mathbb{P}\left(\{\omega \in \Omega \mid \Gamma(x) \subset \mathcal{B}\left(X_1(\cdot)(\omega), \ldots, X_j(\cdot)(\omega)\right)\}\right) \quad \leftarrow \text{ siendo estrictos.} \end{aligned}$$

Esto permite definir lo siguiente:

Definición 2 (versión poblacional). *Para $J \in \mathbb{N}$ tal que $2 \leq J \leq n$ y n curvas $x_1, \ldots, x_n$ en $\mathcal{C}(I)$ se define para la* ***profundidad por bandas poblacional*** *en la curva x (en $\mathcal{C}(I)$) como*

$$\mathcal{BD}_J(x) := \sum_{j=2}^{J} \mathcal{BD}^{(j)}(x) = \sum_{j=2}^{J} \mathbb{P}\left(\Gamma(x) \subset \mathcal{B}\left(X_1, \ldots, X_j\right)\right).$$

Es importante tener en cuenta que J tiene una influencia importante en el comportamiento de esta profundidad, no todos los J traen buenos resultados, por ello hacemos la siguiente observación:

Observación 1. Se recomienda $J = 3$ (en ambos casos, muestral y poblacional), las razones de dicha elección son las siguientes:

1) Cuando $J > 3$ se tiene que el cálculo de $\mathcal{BD}_{n,J}$ es computacionalmente costoso.
2) Cuando $J > 3$ las bandas formadas en ese caso no se parecerán a la forma de las curvas que conforman la muestra, es común que se pierda la forma.
3) El orden inducido por las profundidades por bandas es muy estable respecto a J.
4) Las profundidades por bandas para $J = 2$ son computacionalmente menos costosas pero las curvas frecuentemente se cruzarán, y con probabilidad uno, ninguna otra curva estará dentro de dicha banda.

Otro concepto importante que emerge gracias a esta definición de profundidad es el de *media muestral*, notada por $\widehat{m}_n$. Esta es una curva que pertenece a la muestra y que satisface

$$\widehat{m}_n = \underset{x \in \{X_1(), \ldots, X_n()\}}{\operatorname{argmin}} \mathcal{BD}_{n,J}(x).$$

Y notamos por m a la *media poblacional* que es la curva que maximiza $\mathcal{BD}_J$ (esta es la versión poblacional).

3.1. Aspectos de convergencia y consistencia

Es de esperar que lo muestral converja a lo poblacional a medida que n se hace grande, dicha pretensión es alcanzada en el siguiente teorema.

Teorema 3. *Sea $\mathbb{P}$ una distribución de probabilidad en $\mathcal{C}(I)$ con marginales absolutamente continuas. Entonces*

1. *En cualquier conjunto $\mathcal{E}(I)$ de funciones equicontinuas en I, se tiene que cuando $n \to \infty$*

$$\sup_{x \in \mathcal{I}} |\mathcal{BD}_{n,J} - \mathcal{BD}_J| \longrightarrow 0 \qquad \mathbb{P}a.s.$$

2. *Si existe $m \in \mathcal{E}(I)$ tal que maximiza $\mathcal{BD}_J$ y $\widehat{m}_n \in \mathcal{E}(I)$ es una sucesión tal que*

$$\mathcal{BD}_{n,J}(\widehat{m}_n) = \sup_{x \in \mathcal{E}(I)} \mathcal{BD}_{n,J}(x).$$

Entonces cuando $n \to \infty$

$$\widehat{m}_n \longrightarrow m \qquad \mathbb{P}.a.s.$$

Note que en particular para $\mathsf{Lip}_A[a,b]$ el conjunto definido en el Teorema 1 se tienen las consecuencias de este teorema.

3.2. Versión finito dimensional de la profundidad por bandas

Un vector $\mathbf{x} \in \mathbb{R}^d$ es de la forma $\mathbf{x} = (x_1, \ldots, x_d)$, pero este vector se puede ver como la función

$$\begin{aligned} \mathbf{x} : \{1, 2, \ldots, d\} &\to \mathbb{R} \\ i &\mapsto \mathbf{x}(i) = x_i. \end{aligned}$$

Entonces, una *banda* generada por los vectores $\mathbf{x}_1, \ldots, \mathbf{x}_j$ en $\mathbb{R}^d$ vistos como funciones es dada por

$$\mathcal{B}(\mathbf{x}_1, \ldots, \mathbf{x}_j) := \left\{ \mathbf{x} \in \mathbb{R}^d \;\middle|\; \min_{i=1,\ldots,j} \mathbf{x}_i(k) \leq \mathbf{x}(k) \leq \max_{i=1,\ldots,j} \mathbf{x}_i(k) \;\; \forall k = 1, \ldots, d \right\}.$$

Pero si volvemos a su caracterización como vectores dicha banda puede ser vista como el rectángulo (ver Figura 3.1)

$$\mathcal{R}(\mathbf{x}_1, \ldots, \mathbf{x}_j) = \left\{ \mathbf{x} \in \mathbb{R}^d \;\middle|\; \min_{i=1,\ldots,j} (\mathbf{x}_i)_k \leq \mathbf{x}_k \leq \max_{i=1,\ldots,j} (\mathbf{x}_i)_k \;\; \forall k = 1, \ldots, d \right\}.$$

Ahora, asumiendo $\mathbf{x}_1, \ldots, \mathbf{x}_n$ en $\mathbb{R}^d$ como la realización de n variables aleatorias i.i.d, entonces para cualquier $\mathbf{x} \in \mathbb{R}^d$ se entiende a $\mathcal{BD}_n^j(\mathbf{x})$ como la proporción de rectángulos $\mathcal{R}(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_j})$ que contienen a $\mathbf{x}$ donde los rectángulos son definidos por todos los posibles j diferentes puntos $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_j}$ de la muestra $\mathbf{x}_1, \ldots, \mathbf{x}_n$, es decir

$$\mathcal{BD}_n^{(j)}(\mathbf{x}) = \frac{1}{\binom{n}{j}} \sum_{1 \leq i_1 z \cdots < i_j \leq n} \mathbb{1}_{\mathcal{R}(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_j})}(\mathbf{x}).$$

Entonces para $2 \leq J \leq n$ la *profundidad por bandas finito dimensional muestral* es dada por

$$\mathcal{BD}_{n,J}(\mathbf{x}) = \sum_{j=2}^{J} \mathcal{BD}_n^{(j)}(\mathbf{x}).$$

La versión poblacional de esta profundidad para $X_1, \ldots, X_n$ variables aleatorias i.i.d. es dada por

$$\mathcal{BD}_J(\mathbf{x}) = \sum_{j=2}^{J} \mathcal{BD}^{(j)}(\mathbf{x})$$

donde $\mathcal{BD}^{(j)}(\mathbf{x}) = \mathbb{P}\left(\mathbf{x} \in \mathcal{R}\left(X_1, \ldots, X_n\right)\right)$.

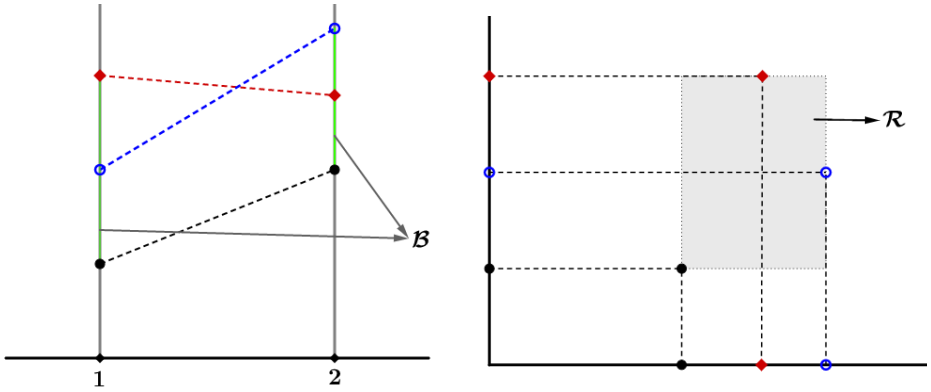


Figura 3.1. Ejemplo de una banda y su respectivo rectángulo para el caso de tres puntos en $\mathbb{R}^2$

3.3. Profundidad por bandas generalizada

Exigir que el grafo de una curva esté completamente contenido en una banda puede ser muy restrictivo, que es justamente exige la profundidad por bandas. Es menos restrictivo exigir que la curva permanezca dentro de la banda la mayor parte del tiempo y dicho enfoque motiva una modificación de la profundidad por bandas. Abordaremos dicha modificación para los dos casos vistos, el funcional y el finito dimensional.

Caso funcional: Sean $x_1, \ldots, x_n$ curvas en $\mathcal{C}(I)$ y $2 \leq j \leq n$, entonces para cualquier subconjunto $x_{i_1}, \ldots, x_{i_j}$ y x una curva en $\mathcal{C}(I)$ se define el término

$$\mathcal{A}(x; x_{i_1}, \ldots, x_{i_j}) := \left\{ t \in I \;\middle|\; \min_{r=i_1,\ldots,i_j} x_r(t) \leq x(t) \leq \max_{r=i_1,\ldots,i_j} x_r(t) \right\}.$$

Ahora, considerando λ como la medida de Lebesgue en I se define

$$\widetilde{\lambda}(\mathcal{A}(x; x_{i_1}, \ldots, x_{i_j})) := \frac{\lambda\left(\mathcal{A}(x; x_{i_1}, \ldots, x_{i_j})\right)}{\lambda(I)}.$$

Esta última expresión mide la proporción de tiempo que la función x permanece en la banda generada por $x_{i_1}, \ldots, x_{i_j}$. Con esto en mente, para $2 \leq j \leq n$ se define una versión más flexible y general de $\mathcal{BD}_n^{(j)}$ como sigue:

$$\mathcal{GBD}_n^{(j)}(x) := \frac{1}{\binom{n}{j}} \sum_{1 \leq i_1 z \cdots < i_j \leq n} \widetilde{\lambda}(\mathcal{A}(x; x_{i_1}, \ldots, x_{i_j})).$$

Note que si la curva x siempre permanece dentro de la banda $\mathcal{B}(x_{i_1}, \ldots, x_{i_j})$, entonces $\mathcal{BD}_n^{(j)} = \mathcal{GBD}_n^{(j)}$.
Por lo tanto, para $2 \leq J \leq n$ se define la *profundidad por bandas generalizada muestral* en x como

$$\mathcal{GBD}_{n,J}(x) := \sum_{j=2}^{J} \mathcal{GBD}_n^j(x).$$

Para $X_1(t), \ldots, X_n(t)$ procesos estocásticos i.i.d, con trayectorias en $\mathcal{C}(I)$ y x una curva en $\mathcal{C}(I)$ la *versión poblacional* de esta profundidad en x es definida como

$$\mathcal{GBD}_J(x) = \sum_{j=2}^{J} \mathcal{GBD}^{(j)}(x)$$

donde

$$\mathcal{GBD}^{(j)}(x) := \mathbb{E}\left[\widetilde{\lambda}\left(\mathcal{A}\left(x; X_1, \ldots, X_j\right)\right)\right].$$

Caso finito dimensional: Considerando $\mathbf{x}_1, \ldots, \mathbf{x}_n$ en $\mathbb{R}^d$, para un subconjunto $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_j}$ con $2 \leq j \leq n$ y $\mathbf{x}$ en $\mathbb{R}^d$ se define

$$\mathcal{GBD}_n^{(j)}(\mathbf{x}) = \frac{1}{\binom{n}{j}} \sum_{1 \leq i_1 z \cdots < i_j \leq n} \frac{1}{d} \sum_{k=1}^{d} \mathbb{1}\left\{\mathbf{y} \;\middle|\; \min_{r=1,\ldots,j} \mathbf{x}_{i_r}(k) \leq \mathbf{y}(k) \leq \max_{r=1,\ldots,j} \mathbf{x}_{i_r}(k)\right\}(\mathbf{x}).$$

Esta ultima expresión mide la proporción de coordenadas de $\mathbf{x}$ que estan dentro del intervalo dado por j diferentes puntos de la muestra. Así, tenemos que la *profundidad por bandas generalizada muestral* es

$$\mathcal{GBD}_{n,J}(\mathbf{x}) = \sum_{i=2}^{J} \mathcal{GBD}_n^{(j)}(\mathbf{x}).$$

4. Profundidad por medio de semi-regiones

Consideremos un proceso estocástico X con caminos en $\mathcal{C}(I)$ con densidad de probabilidad $\mathbb{P}$. Sea $\chi_1(t), \ldots, \chi_n(t)$ una muestra de curvas de $\mathcal{C}(I)$.

Definición 3. *Definimos el* ***hipografo*** *y el* ***epigrafo*** *de una función $\chi \in \mathcal{C}(I)$ como:*

$$\begin{aligned} \text{hyp}(\chi) &= \{(t, y) \in I \times \mathbb{R} : y \leq \chi(t)\}, \\ \text{epi}(\chi) &= \{(t, y) \in I \times \mathbb{R} : y \geq \chi(t)\} \end{aligned}$$

Con esto en mente tenemos lo siguiente:

Definición 4. *La* ***profundidad por semi-región*** *en x con respecto a un conjunto de funciones $\chi_1(t), \ldots, \chi_n(t)$ es*

$$S_{n,H} = \text{mín}\{G_{1n}(\chi), G_{2n}(\chi)\},$$

donde

$$\begin{aligned} G_{1n}(\chi) &= \frac{\sum_{i=1}^{n} \mathbb{1}_{\text{hyp}(\chi)}(\Gamma(\chi_i))}{n} \\ G_{2n}(\chi) &= \frac{\sum_{i=1}^{n} \mathbb{1}_{\text{epi}(\chi)}(\Gamma(\chi_i))}{n} \end{aligned}$$

Por tanto, la profundidad por semi-regiones en χ es el mínimo entre la proporción de funciones de la muestra cuyo grafo está en el ***hipografo*** *de χ y la proporción correspondiente de funciones para el* ***epigrafo*** *de χ.*

La versión **poblacional** de $S_{n,H}$ es

$$S_H(\chi) = \text{mín}\{G_1(\chi), G_2(\chi)\},$$

donde

$$G_1(\chi) = \mathbb{P}(\Gamma(X) \subseteq \text{hyp}(\chi)) = \mathbb{P}(X(t) \leq \chi(t), t \in I)$$
$$G_2(\chi) = \mathbb{P}(\Gamma(X) \subseteq \text{epi}(\chi)) = \mathbb{P}(X(t) \geq \chi(t), t \in I)$$

Una **curva profunda** o **mediana muestral**-S_H es una curva que maximiza la profundidad por semi-regiones y es denotada por

$$\hat{\tau}_n = \text{argmax}_{\chi \in \{\chi_1, \ldots, \chi_n\}} S_{n,H}(\chi)$$

y la mediana S_H de la población se define como la curva en $\mathcal{C}(I)$ que maximiza S_H. Si la muestra de curvas $\chi_1, \chi_2, \ldots, \chi_n$ se ordena de acuerdo a valores decrecientes de $S_{n,H}(x_i)$, obtenemos estadísticos de orden $x_{(1)}, \ldots, x_{(n)}$ donde $x_{(1)}$ es la observación más profunda y x_n es la más superficial. Recordemos que siguiendo este orden podemos definir la *media podada* α como el $(1-\alpha) * 100$ por ciento de las curvas más profundas.

4.1. Versión finito-dimensional

Los conceptos de hipografo y epigrafo se pueden generalizar a datos de dimensión finita. Denotemos por $x(k)$ la componente k-ésima del vector x y si consideremos cada punto en $\mathbb{R}^d$ como una función definida sobre $\{1, 2, \ldots, d\}$ y tenemos:

Definición 5. *El* ***hipografo*** *de x es*

$$\text{hyp}(x) = \{(k, y) \in \{1, 2, \ldots, d\} \times \mathbb{R} : y \leq x(k)\},$$

y el ***epigrafo*** *de x es*

$$\text{epi}(x) = \{(k, y) \in \{1, 2, \ldots, d\} \times \mathbb{R} : y \geq x(k)\}.$$

Sea X una variable aleatoria d-dimensional con función de distribución F. Denotamos con $X \leq x$ al conjunto $\{X(k) \leq x(k), k = 1, \ldots, d\}$ y análogamente hacemos para $X \geq x$. Por lo tanto, podemos dar la siguiente noción de profundidad para el caso de dimensión finita:

$$\begin{aligned} S_H(x, F) := S_H(x) &= \text{mín}\{\mathbb{P}(X \leq x), \mathbb{P}(X \geq x)\} \\ &= \text{mín}\{F_X(x), F_{-X}(-x)\} \\ &= \text{mín}\{F_X(x), F_Y(y)\}, \end{aligned}$$

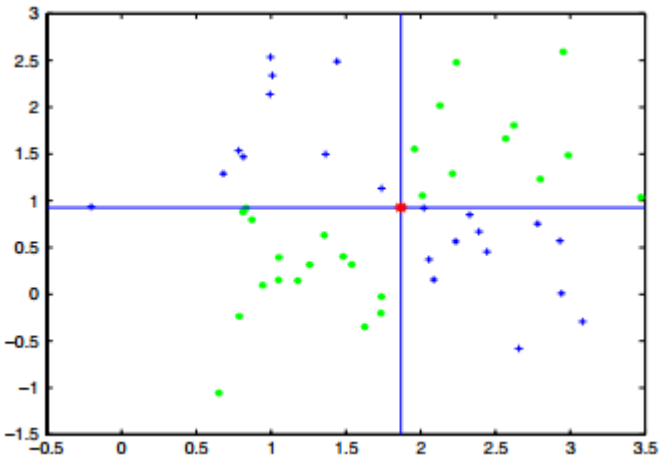
Donde $Y = -X$ y $y = -x$. Sea $x_1, \ldots, x_n$ una muestra aleatoria de la variable X. La versión muestral de la profundidad por semi-regiones es:

$$\begin{aligned} S_{n,H}(x) &= \text{mín}\left\{ \frac{\sum_{i=1}^n \mathbb{1}_{(x_i \leq x)}}{n}, \frac{\sum_{i=1}^n \mathbb{1}_{(x_i \geq x)}}{n} \right\} \\ &= \text{mín}\{F_{X_n}(x), F_{Y_n}(y)\} \end{aligned}$$

La principal ventaja de la profundidad por semi-regiones sobre otras profundidades es la facilidad de cómputo y su aplicabilidad a datos de altas dimensiones ($n << d$). De hecho, se puede mostrar que el costo computacional de la profundidad por semi-regiones de un punto en $\mathbb{R}^d$ con respecto a una muestra de n de datos es $O(nd)$.

La figura 4.1 muestra cómo calcular la profundidad por semi-regiones $S_{n,H}$ del punto más profundo de la muestra (marcado con un asterisco). La proporción de puntos en el rectángulo superior derecho es 12/50 y en el izquierdo inferior es de 17/50. Por lo tanto, la profundad por semi-regiones del punto es 12/50.

Figura 4.1. Punto más profundo en una muestra de 50 puntos normales



4.2. Algunas propiedades de S_H

Proposición 1. *S_H es invariante bajo traslación y algunos tipos de dilaciones. Sea A una matriz diagonal definida positiva o negativa y $b \in \mathbb{R}^d$, entonces*

$$S_H(Ax + b, F_{Ax+b}) = S_H(x, F)$$

Proposición 2. *Para $d = 1$ la profundidad por medio de regiones $s_H(x)$ se puede expresar como*

$$\begin{aligned} S_H(x) &= \text{mín}\{\mathbb{P}(X \leq x), 1 - \mathbb{P}(X < x)\} \\ &= \text{mín}\{F(x), 1 - F(x^-)\}, \end{aligned}$$

Y es equivalente a la profundidad de Tukey por semi-espacios. Además, el valor que maximiza S_H es la mediana usual en $\mathbb{R}$.

Cabe notar que $S_H(x)$ se va a cero conforme x tiende a infinito. La generalización para dimensión finita se resume en la siguiente proposición:

Proposición 3. *Sea $x \in \mathbb{R}^d$, entonces*

$$\sup_{||x|| \geq M} S_H(x) \longrightarrow 0, \text{ cuando } M \longrightarrow \infty$$

y

$$\sup_{||x|| \geq M} S_H(x) \xrightarrow{a.s} 0, \text{ cuando } M \longrightarrow \infty$$

Proposición 4. *$S_{n,H}$ es uniformemente consistente en el siguiente sentido:*

$$\sup_{x \in \mathbb{R}^d} |S_{n,H}(x) - S_H(x)| \xrightarrow{a.s} 0, \text{ cuando } n \longrightarrow \infty$$

Además, si $S_H(x)$ se maximiza unívocamente en τ y $(\tau_n)_n$ es una sucesión de variables aleatorias con $S_{n,H} = \sup_{x \in \mathbb{R}^d} S_{n,H}(x)$, entonces

$$\tau_n \xrightarrow{a.s} \tau, \text{ cuando } n \longrightarrow \infty$$

Propiedades de la profundidad por semi-regiones funcional. Sean $x_1, \ldots, x_n$ copias independientes de un proceso estocástico X en $\mathcal{C}(I)$ con función de distribución P. Supongamos que el proceso estocástico es ajustado, es decir:

$$P(||x||_\infty \geq M) \longrightarrow 0 \text{ cuando } M \to \infty$$

Para el caso funcional tenemos las siguientes propiedades:

Proposición 5 (Invarianza). *Sean a y b funciones en $\mathcal{C}(I)$ con $a(t) > 0$ para todo $t \in I$ o $a(t) < 0$ para todo $t \in I$. Entonces*

$$S_H(ax + b, P_{aX+b}) = S_H(x, P_X)$$

Proposición 6 (Convergencia). *Las profundidades S_H y $S_{n,H}$ satisfacen:*

1. $\sup_{||x||_\infty \geq M} S_H(x) \longrightarrow 0$ *cuando* $M \to \infty$
2. $\sup_{||x||_\infty \geq M} S_{n,H}(x) \xrightarrow{c.s} 0$ *cuando* $M \to \infty$

Proposición 7 (Consistencia). $S_{n,H}$ *es* ***fuertemente consistente*** *en el siguiente sentido:*

$$S_{n,H}(x) \xrightarrow{c.s} S_H(x)$$

Proposición 8. $S_H(\cdot)$ *es un funcional semicontinuo superiormente.*

Para fines de completitud agregamos una versión modificada de la profundidad por semi-regiones.

4.3. Versión modificada de la profundidad por semi-regiones

El gran beneficio de esta versión se encuentra al poder analizar curvas no suaves que se entrecruzan mucho entre sí [4] al ser menos restrictiva en las hipótesis para su definición. Como antes, sea $x \in \mathcal{C}(I)$ (I un intervalo cerrado y acotado) y X un proceso estocástico, definimos

$$SL(x) = \frac{1}{\lambda(I)} \mathbb{E}[\lambda\{t \in I : x(t) \leq X(t)\}],$$
$$IL(x) = \frac{1}{\lambda(I)} \mathbb{E}[\lambda\{t \in I : x(t) \geq X(t)\}]$$

donde λ es la medida de Lebesgue en $\mathbb{R}$. En otras palabras, la forma de pensar $SL(x)$ es como la proporción del tiempo en donde el proceso X está por encima de la curva x. De igual manera, $IL(x)$ es la proporción de tiempo en que el proceso X está por debajo de la curva x. Con estos preliminares podemos hacer la siguiente definición:

Definición 6. *La* ***profundidad por regiones*** *en x* ***modificada*** *es:*

$$MS_H(x) = \text{mín}\{SL(x), IL(x)\}.$$

Ahora, sea $x_1, \ldots, x_n$ un conjunto de curvas con distribución P. La versión muestral de la noción de profundidad por regiones modificada, como sabemos, se obtiene al sustituir P por la distribución empírica P_n:

$$MS_{n,H}(x) = \text{mín}\{SL_n(x), IL_n(x)\},$$

donde

$$SL_n(x) = \frac{1}{n\lambda(I)} \sum_{i=1}^{n} \lambda\{t \in I : x(t) \leq x_i(t)\}$$

y

$$IL_n(x) = \frac{1}{n\lambda(I)} \sum_{i=1}^{n} \lambda\{t \in I : x(t) \geq x_i(t)\}$$

5. Comentarios finales

En [4] se hizo una comparación de los rendimientos de varias de las profundidades expuestas. La profundidad por bandas fue la que tuvo el mejor rendimiento de todas a nuestro parecer: en todas las simulaciones el 100 % de las veces detectó un .ºutlier". La segunda fue la profundidad por semi-regiones, que sólo falló en un caso y de tercera y cuarta tenemos la profundidad por bandas modificada y la profundidad por semi-regiones respectivamente.

Figura 5.1. Porcentaje de veces que se detecta un outlier con 100 repeticiones y 50 curvas

	$\mu_2 = 0.1$ $k_2 = 1$	$\mu_2 = 0.2$ $k_2 = 1$	$\mu_2 = 0.3$ $k_2 = 1$	$\mu_2 = 0.1$ $k_2 = 2$	$\mu_2 = 0.2$ $k_2 = 2$	$\mu_2 = 0.3$ $k_2 = 2$
S_H	100	100	97	100	100	100
MS_H	6	19	4	18	16	18
BD	100	100	100	100	100	100
MBD	15	23	23	52	52	48

La tabla anterior la tomamos del análisis hecho en [4]. BD denota profundidad por bandas y MBD profundidad por bandas modificada. Por otro lado, μ_2 y k_2 son parámetros de una de las funciones de covarianza para generar las cincuenta curvas (contaminadas).

Referencias

1. R. Fraiman & G. Muniz, *Trimmed means for functional data*, In: Sociedad de Estadistica e Investigacidn Operativa Test, Vol. 10, No. 2, Springer, pp. 419-440 (2001)
2. S. López-Pintado & J. Romo, *Depth-based inference for functional data*, In: Computational Statistics & Data Analysis, vol 51, Elsevier, pp. 4957–4968 (2007)
3. S. López-Pintado & J. Romo, *On the Concept of Depth for Functional Data*, In: Journal of the American Statistical Association, vol 104, No 486, pp. 718-734 (2009)
4. S. López-Pintado & J. Romo, *A half-region depth for functional data* , In: Computational Statistics & Data Analysis, vol 55, Elsevier, pp. 1679–1695 (2011)
5. R. Randles & D. Wolfe, *Introduction to Theory of Nonparametric Statistic*, Krieger Publishing Company, pp. 3-114 (1979)