Profundidad de datos funcionales

Diego Fonseca & Rodolfo Quintero

Universidad de los Andes

23/May/2017

1. **Introduccion**

2. **Profundidad integral**
   - Algunas profundidades finito dimensionales
   - Profundidad integral
   - α-trimmed mean
   - Aspectos de convergencia y consistencia

3. **Profundidad por medio de bandas**
   - Aspectos de convergencia y consistencia
   - Versión finito dimensional de la profundidad por bandas
   - Profundidad por bandas generalizada
     - Caso funcional
     - Caso finito dimensional

4. **Profundidad por medio de semi-regiones**
   - Versión finito-dimensional
   - Propiedades de la profundidad por semi-región

Introduccion

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.

El objetivo es determinar cual de estas curvas permanece en el medio de todas las curvas 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 mas al centro, en ese sentido, a partir del concepto de profundidad se puede inducir una noción de orden en los datos funciones, estos se ordenan desde el mas profundo al menos profundo.

Algunas profundidades finito dimensionales

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

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

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

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

Algunas profundidades finito dimensionales

Profundidad de 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 de 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 $Y_1, \ldots, Y_{k+1}$. 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

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

Algunas profundidades finito dimensionales

Profundidad standard unidimensional

Dados $Y_1, \ldots, Y_n$ variables aleatorias unidimensionales independientes e idénticamente distribuidas con distribución F, se define la Profundidad standard 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 standard unidimensional se define de la misma manera y se denota por UD_n.

Profundidad integral

De las anteriores profundidades nos interesa sus versiones unidimensionales, en ese sentido, consideramos nuestro conjunto de datos funcionales como $X_1(t), X_2(t), \ldots, X_n(t)$ de independientes es idénticamente distribuidos procesos estocásticos con distribución F_t y con trayectorias continuas en un intervalo compacto I. La versión empírica de F_t es notada por $F_{n,t}$.

En ese sentido, considerando D_t una profundidad unidimensional la cual 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 $[a, b]$ su **profundidad integral**

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

Profundidad integral

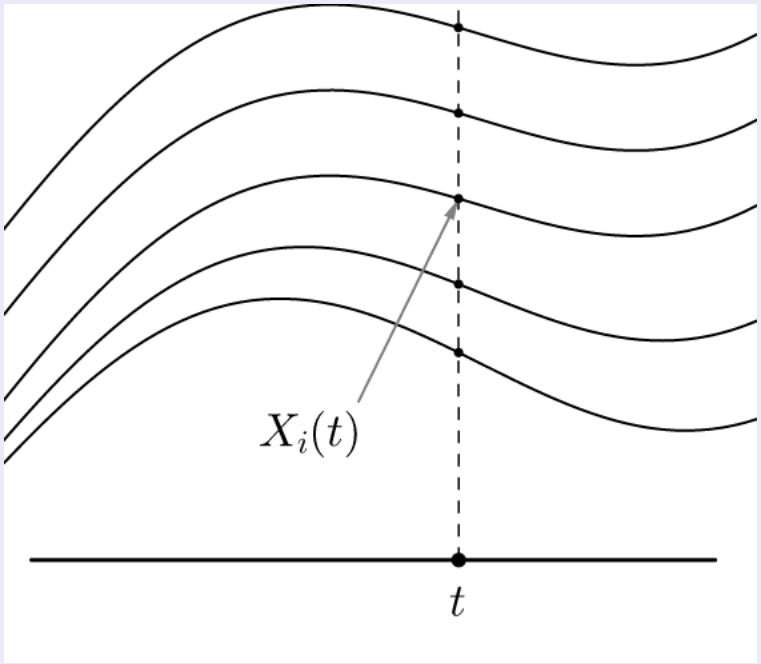


Figure 1: Para t fijo los procesos estocásticos forman una muestra unidimensional.

Profundidad integral

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 que notamos $D_{n,t}$, 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 mas profundo en el tiempo t. Como las trayectorias X_i son continuas entonces Z_i es continua en $[a, b]$, por lo tanto, es posible integrar, de modo que existe $I(X_i)$, para efectos de notación se acostumbra a escribir

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

Profundidad integral

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

α-trimmed mean (media truncada)

El concepto de α-trimmed mean (media truncada) es 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 el α-*trimmed mean* en el caso unidimensional como

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

α-trimmed mean (media truncada)

Ahora, si consideramos datos funcionales $X_1(t), X_2(t), \ldots, X_n(t)$ la versión funcional de α-trimmed mean notada 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 un buena aproximación de μ_n es un avance,en ese sentido se define el α-**trimmed mean estimador** 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

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

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 teorema:

Asumiendo las siguientes condiciones:

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

$$\text{Lip}_A[a,b] := \left\{ x : [a,b] \to \mathbb{R} \;\middle|\; \begin{array}{l} x \text{ función Lipschitz con} \\ \text{constante menor o igual a } A \end{array} \right\}.$$

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

Aspectos de convergencia y consistencia

Teorema 1

Asumiendo las condiciones (I) y (II) y definiendo

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

Entonces tenemos

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

y

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

Aspectos de convergencia y consistencia

Teorema 2

Si las trayectorias del proceso estocástico $X_1(t)$ pertenecen a un espació 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 $\widehat{\mu}_n \longrightarrow \mu$.

Profundidad por 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 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. En ese sentido, es 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.

Profundidad por bandas

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

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

Profundidad por bandas

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

$$\mathcal{B}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) := \left\{ (t, y) \,\middle|\, t \in I \text{ y } \min_{r=1,\ldots,k} x_{i_k} \leq y \leq \max_{r=1,\ldots,k} x_{i_k} \right\}$$

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

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

Este ultimo 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.

Profundidad por bandas

Definición

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 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 termino "*muestral*" en esta ultima 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 $\mathcal{C}(I)$.

Profundidad por bandas

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

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

Esto permite definir lo siguiente:

Definition (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

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

Profundidad por bandas

Observación

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 calculo de $\mathcal{B}D_{n,J}$ 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 cruzaran, y con probabilidad uno, ninguna otra curva estará dentro de dicha banda.

Profundidad por bandas

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

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

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

Aspectos de convergencia y consistencia

Teorema

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 $n \to \infty$

$$\sup_{x \in \mathcal{I}} |\mathcal{B}D_{n,J} - \mathcal{B}D_J| \longrightarrow 0 \qquad \mathbb{P}a.s.$$

2. Si existe $m \in \mathcal{E}(I)$ tal que maximiza $\mathcal{B}D_J$ y $\widehat{m}_n \in \mathcal{E}(I)$ es una sucesión tal que $\mathcal{B}D_{n,J}(\widehat{m}_n) = \sup_{x \in \mathcal{E}(I)} \mathcal{B}D_{n,J}(x)$. Entonces cuando $n \to \infty$ se tiene $\widehat{m}_n \longrightarrow m \qquad \mathbb{P}.a.s.$

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

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

$$\mathbf{x} : \{1, 2, \ldots, d\} \to \mathbb{R} \quad \text{donde } i \mapsto \mathbf{x}(i) = x_i.$$

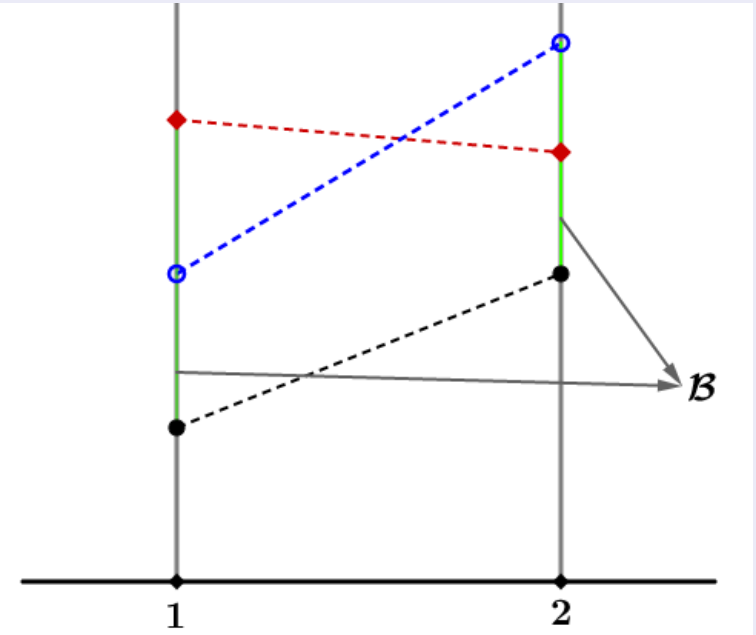
Entonces, una **banda** generada por los vectores $\mathbf{x}_1, \ldots, \mathbf{x}_j$ en $\mathbb{R}^d$ vistos como funciones es dada por

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

Pero si volvemos a su caracterización como vectores dicha banda puede ser vista como el rectángulo (ver Figura 2)

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

Versión finito dimensional de la profundidad por bandas



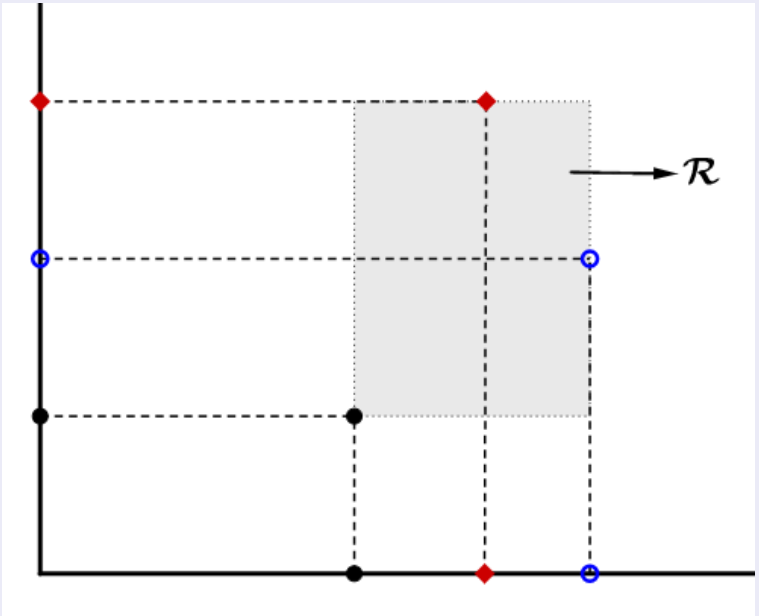


Figure 2: Ejemplo de una banda y su respectivo rectángulo para el caso de tres puntos en $\mathbb{R}^2$

Versión finito dimensional de la profundidad por bandas

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

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

Versión finito dimensional de la profundidad por bandas

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

$$\mathcal{B}D_J(\mathbf{x}) = \sum_{j=2}^{J} \mathcal{B}D^{(j)}(\mathbf{x})$$

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

Profundidad por bandas generalizada

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

Profundidad por bandas generalizada

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 termino

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

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 ultima expresión mide la proporción de tiempo que la función x permanece en la banda generada por $x_{i_1}, \ldots, x_{i_j}$.

Profundidad por bandas generalizada

Caso funcional

Con lo anterior en mente, para $2 \leq j \leq n$ se define una versión mas flexible y general de $\mathcal{B}D_n^{(j)}$ como sigue:

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

Note que si la curva x siempre permanece dentro de la banda $\mathcal{B}(x_{i_1}, \ldots, x_{i_j})$, entonces $\mathcal{B}D_n^{(j)} = \mathcal{GB}D_n^{(j)}$.

Profundidad por bandas generalizada

Caso funcional

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

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

Profundidad por bandas generalizada

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

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

Profundidad por medio de semi-regiones

Definition

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

$$\mathrm{hyp}(\chi) = \{(t, y) \in I \times \mathbb{R} : y \leq \chi(t)\},$$
$$\mathrm{epi}(\chi) = \{(t, y) \in I \times \mathbb{R} : y \geq \chi(t)\}$$

Profundidad por medio de semi-regiones

Definition

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

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

donde

$$G_{1n}(\chi) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\mathrm{hyp}(\chi)}(\Gamma(\chi_i))}{n}$$

$$G_{2n}(\chi) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\mathrm{epi}(\chi)}(\Gamma(\chi_i))}{n}$$

Profundidad por medio de semi-regiones

Versión poblacional

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

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

donde

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

Versión finito-dimensional

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:

Hipografo y epigrafo en dimensión finita

El **hipografo** de x es

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

y el **epigrafo** de x es

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

Versión finito-dimensional

Denotamos con $X \leq x$ al conjunto $\{X(k) \leq x(k), k = 1, \ldots, d\}$ y análogamente hacemos para $X \geq x$.

PSR para dimensión finita

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

Donde $Y = -X$ y $y = -x$.

Versión finito-dimensional

Caso poblacional

Sea $x_1, \ldots, x_n$ una muestra aleatoria de la variable aleatoria X. La versión poblacional de la profundidad por semi-regiones es:

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

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 ($n << d$).

Propiedades de la profundidad por semi-región

Propiedad 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

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

Propiedad 2

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

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

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

Propiedades de la profundidad por semi-región

Propiedad 3

Sea $x \in \mathbb{R}^d$, entonces

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

y

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

Propiedades de la profundidad por semi-región

Propiedad 4

$S_{n,H}$ es uniformemente consistente en el siguiente sentido:

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

Además, si $S_H(x)$ se maximiza únicamente 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

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

Gracias