Profundidad de Datos Funcionales

Nicolas Maldonado Baracaldo

Andrés Felipe Patiño López

Noviembre 2021

Departamento de Matemáticas
Facultad de Ciencias
Universidad de los Andes

Motivación

Motivación

Para una muestra aleatoria real $X_1, X_2, ..., X_n$ poemos definir:

Motivación

Para una muestra aleatoria real $X_1, X_2, ..., X_n$ poemos definir:

- El k-ésimo estadístico de orden $X_{(k)}$.

Motivación

Para una muestra aleatoria real $X_1, X_2, ..., X_n$ poemos definir:

- El k-ésimo estadístico de orden $X_{(k)}$.
- El vector de rangos $\mathcal{R} = (R_1, \ldots, R_n)$.

Motivación

Para una muestra aleatoria real $X_1, X_2, ..., X_n$ poemos definir:

- El k-ésimo estadístico de orden $X_{(k)}$.
- El vector de rangos $\mathcal{R} = (R_1, \ldots, R_n)$.
- L-estadísticos. En partícular, la media truncada.

A lo largo de este curso, ya hemos visto la utilidad del vector de rangos y los estadísticos de orden.

Motivación

Profundidad: Una medida de profundidad en un conjunto de datos proporciona una noción de centralidad de los mismos. Es una función $\mathrm{D} : \mathbb{R}^{\mathrm{k}} \to [0, 1]$ que depende de la distribución de los datos (o de la empírica).

Motivación

Profundidad: Una medida de profundidad en un conjunto de datos proporciona una noción de centralidad de los mismos. Es una función $\mathrm{D} : \mathbb{R}^{\mathrm{k}} \to [0, 1]$ que depende de la distribución de los datos (o de la empírica).

- Para finitos datos sin empates la profundidad induce estadísticos de orden.

Motivación

Profundidad: Una medida de profundidad en un conjunto de datos proporciona una noción de centralidad de los mismos. Es una función $D : \mathbb{R}^k \to [0, 1]$ que depende de la distribución de los datos (o de la empírica).

- Para finitos datos sin empates la profundidad induce estadísticos de orden.
- La mediana será análoga al dato "más profundo" o "más central". Es decir, el estadístico de orden $X_{(1)}$.

Motivación

Profundidad: Una medida de profundidad en un conjunto de datos proporciona una noción de centralidad de los mismos. Es una función $\mathrm{D} : \mathbb{R}^{\mathrm{k}} \to [0, 1]$ que depende de la distribución de los datos (o de la empírica).

- Para finitos datos sin empates la profundidad induce estadísticos de orden.
- La mediana será análoga al dato "más profundo" o "más central". Es decir, el estadístico de orden $\mathrm{X}_{(1)}$.
- Los datos atípicos o los datos extremos serán los "menos profundos" o "menos centrales". Por ejemplo, el estadístico de orden $\mathrm{X}_{(\mathrm{n})}$.

Motivación

Profundidad: Una medida de profundidad en un conjunto de datos proporciona una noción de centralidad de los mismos. Es una función $D : \mathbb{R}^k \to [0, 1]$ que depende de la distribución de los datos (o de la empírica).

- Para finitos datos sin empates la profundidad induce estadísticos de orden.
- La mediana será análoga al dato "más profundo" o "más central". Es decir, el estadístico de orden $X_{(1)}$.
- Los datos atípicos o los datos extremos serán los "menos profundos" o "menos centrales". Por ejemplo, el estadístico de orden $X_{(n)}$.
- Esto implica que las definiciones de profundidad son ligeramente diferentes en dimensiones > 1.

Motivación

Debido a la naturaleza y la libertad de movimiento en dimensiones más altas, existen una gran variedad de formas de extender las profundidades de datos. Algunas profundidades (Datos multivariados):

- **Tukey's Depth**
- **Simplicial Depth**
- **L1 Depth**

Motivación

Motivación

Motivación

Objetivo: Extender estas definiciones a $X_1(t), \cdots, X_n(t)$ datos funcionales i.i.d. F_t con trayectorias continuas.

Profundidad vía integración

Profundidad vía integración

Suponga que tenemos una profundidad $\mathrm{D_t}$ definida sobre $\mathbb{R}$ para cada $\mathrm{t} \in [0,1]$ y $\mathrm{x(t)} \in \mathrm{C}[(0,1)]$. Defina:

$$\mathrm{Z(t)} = \mathrm{D_t(x(t))}, \quad \mathrm{t} \in [0,1]$$

Profundidad vía integración

Suponga que tenemos una profundidad D_t definida sobre $\mathbb{R}$ para cada $t \in [0,1]$ y $x(t) \in C[(0,1)]$. Defina:

$$Z(t) = D_t(x(t)), \quad t \in [0,1]$$

La profundidad funcional (poblacional) de $x = x(t)$ es

$$I(x) = \int_0^1 Z(t)\, dt$$

Profundidad vía integración

Suponga que tenemos una profundidad $\mathrm{D_t}$ definida sobre $\mathbb{R}$ para cada $\mathrm{t} \in [0,1]$ y $\mathrm{x(t)} \in \mathrm{C}[(0,1)]$. Defina:

$$\mathrm{Z(t)} = \mathrm{D_t}(\mathrm{x(t)}), \quad \mathrm{t} \in [0,1]$$

La profundidad funcional (poblacional) de $\mathrm{x} = \mathrm{x(t)}$ es

$$\mathrm{I(x)} = \int_0^1 \mathrm{Z(t)}\,\mathrm{dt}$$

Obs:

- Decimos que x es profundo si $\mathrm{I(x)}$ es grande.

Profundidad vía integración

Suponga que tenemos una profundidad $\mathrm{D_t}$ definida sobre $\mathbb{R}$ para cada $\mathrm{t} \in [0, 1]$ y $\mathrm{x(t)} \in \mathrm{C}[(0, 1)]$. Defina:

$$\mathrm{Z(t) = D_t(x(t)), \quad t \in [0, 1]}$$

La profundidad funcional (poblacional) de $\mathrm{x = x(t)}$ es

$$\mathrm{I(x)} = \int_0^1 \mathrm{Z(t)\, dt}$$

Obs:

- Decimos que x es profundo si $\mathrm{I(x)}$ es grande.
- Si los valores de $\mathrm{D_t(x(t))}$ son pequeños (grandes), los valores de $\mathrm{I(x)}$ serán pequeños (grandes).

EDO's and ranks

Dados $X_1(t), \ldots, X_n(t)$ procesos estocásticos i.i.d. con trayectorias continuas definidas en $[0, 1]$, definimos

$$Z_i(t) = D_t(X_i(t)) \quad \text{e} \quad I_i = \int_0^1 Z_i(t)\, dt.$$

Ordenamos los valores I_k de forma decrecientes en profundidad. Si I_j está en la posición i, entonces $X_{(i)}(t) = X_j(t)$. Esto es $R_j = i$ y

$$X_{(R_j)}(t) = X_j(t).$$

Version muestral: En la definición de la profundidad, reemplazamos F_t por $F_{n,t}$ (la distribución empírica asociada) en la profundidad.

Medias truncadas

Dado $\alpha > 0$ (usualmente no superior 0.25). Podemos definir la media truncada a nivel α como el promedio de los $\mathrm{n} - [\mathrm{n}\alpha]$ datos más profundos.

Más precisamente, elegimos $\beta > 0$ tal que el intervalo $[\beta, \infty)$ contenga $\mathrm{n} - [\mathrm{n}\alpha]$ valores de $\mathrm{I_n}(\mathrm{X_i})$ para $\mathrm{i} = 1, \cdots, \mathrm{n}$. Entonces:

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

Consistencia

Bajo dos condiciones H_1 y H_2 uno puede garantizar que

$$\widehat{\mu}_{\mathrm{n}} \to \mu \text{ c.s.}$$

Donde

$$\mu = \frac{\mathrm{E}(\mathbb{1}_{[\beta,\infty)}(\mathrm{I}(\mathrm{X}_1))\mathrm{X}_1)}{\mathrm{E}(\mathbb{1}_{[\beta,\infty)}(\mathrm{I}(\mathrm{X}_1)))}$$

es la media poblacional recortada.

Profundidad vía bandas

Profundidad vía bandas

El **grafo** de una curva $x \in C[a, b]$ está definido por:

$$G(x) = \{(t, x(t)) : t \in [a, b]\}$$

Y, la **banda** delimitada por j curvas $x_1, \ldots, x_j$ es:

$$B(x_1, \ldots, x_j) = \left\{(t, y) : t \in [a, b], \min_{k=1,\ldots,j} x_k(t) \leq y \leq \max_{k=1,\ldots,j} x_k(t)\right\}$$

Profundidad vía bandas

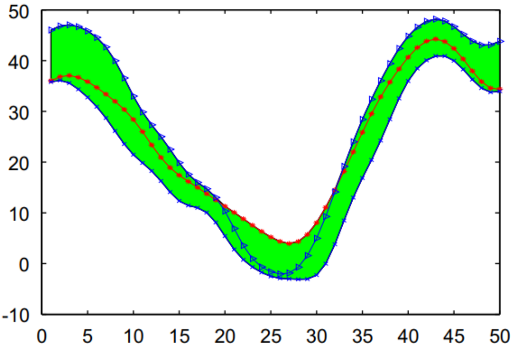

Fig. 1. Band defined by three curves.

La medida de profundidad **BD** para una curva x dado un conjunto de curvas $x_1, \cdots, x_n$ está dada por:

$$S_{n,J}(x) = \sum_{j=2}^{J} S_n^{(j)}(x)$$

donde

$$S_n^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 \leq \cdots \leq i_j \leq n} \mathbb{1}_{\{G(x) \subset B(x_{i_1}, \cdots, x_{i_j})\}}(x)$$

BD

EDOs: Dados $\mathrm{X}_1(\mathrm{t}), \cdots, \mathrm{X}_\mathrm{n}(\mathrm{t})$ procesos estocásticos i.i.d.

EDOs: Dados $X_1(t), \cdots, X_n(t)$ procesos estocásticos i.i.d.

- La mediana se define como

$$X_m = \text{argmax}\, S_{n,J}(X_i)$$

EDOs: Dados $X_1(t), \cdots, X_n(t)$ procesos estocásticos i.i.d.

- La mediana se define como

$$X_m = \text{argmax}\, S_{n,J}(X_i)$$

- Puedo definir estadísticos de orden, rangos y medias truncadas con esta profundidad tal como antes.

Restricciones:

EDOs: Dados $X_1(t), \cdots, X_n(t)$ procesos estocásticos i.i.d.

- La mediana se define como

$$X_m = \operatorname{argmax} S_{n,J}(X_i)$$

- Puedo definir estadísticos de orden, rangos y medias truncadas con esta profundidad tal como antes.

Restricciones:

- Empates cuando $J = 2$.

EDOs: Dados $X_1(t), \cdots, X_n(t)$ procesos estocásticos i.i.d.

- La mediana se define como

$$X_m = \operatorname{argmax} S_{n,J}(X_i)$$

- Puedo definir estadísticos de orden, rangos y medias truncadas con esta profundidad tal como antes.

Restricciones:

- Empates cuando $J = 2$.
- Variabilidad y cálculo compuacional intensivo cuando $J > 3$.

BD

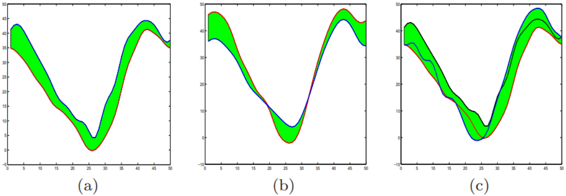

FIG 2. (a) Band determined by two curves, (b) Band determined by two curves, where the curves cross, and (c) Band determined by three curves.

Idea: Debilitar un poco la **BD** cambiando la función indicadora por la proporción de tiempo que una curva x vive "dentro" de un conjunto de curvas dadas.

Considere las curvas $x, x_1, \ldots, x_n$ y para $j \leq n$ defina

$$A(x; x_{i_1}, \ldots, x_{i_j}) = \left\{ t \in [a, b] : \min_{k=1,\ldots,j} x_{i_k}(t) \leq y \leq \max_{k=1,\ldots,j} x_{i_k}(t) \right\}$$

Dados $x(t), x_1(t), \ldots, x_n(t)$ definimos la profundidad generalizada de banda **BGD** por

$$\mathrm{GS}_{n,J}(x) = \sum_{j=2}^{J} \mathrm{GS}_n^{(j)}(x)$$

donde

$$\mathrm{GS}_n^{(j)}(x) := \binom{n}{j}^{-1} \sum_{1 \leq i_2 \leq \cdots \leq i_j \leq n} \frac{A(x; x_{i_1}, \cdots, x_{i_j})}{b - a}$$

Profundidad vía "half-region"

Profundidad vía “half-region”

Grafo, hipografo, epigrafo

$$\mathrm{G}(\mathrm{x}) = \{(\mathrm{t}, \mathrm{x}(\mathrm{t})) : \mathrm{t} \in [\mathrm{a}, \mathrm{b}]\}$$

$$\mathrm{hyp}(\mathrm{x}) = \{(\mathrm{t}, \mathrm{y}) \in [\mathrm{a}, \mathrm{b}] \times \mathbb{R} : \mathrm{y} \leq \mathrm{x}(\mathrm{t})\}$$

$$\mathrm{epi}(\mathrm{x}) = \{(\mathrm{t}, \mathrm{y}) \in [\mathrm{a}, \mathrm{b}] \times \mathbb{R} : \mathrm{y} \geq \mathrm{x}(\mathrm{t})\}$$

Profundidad vía “half-region”

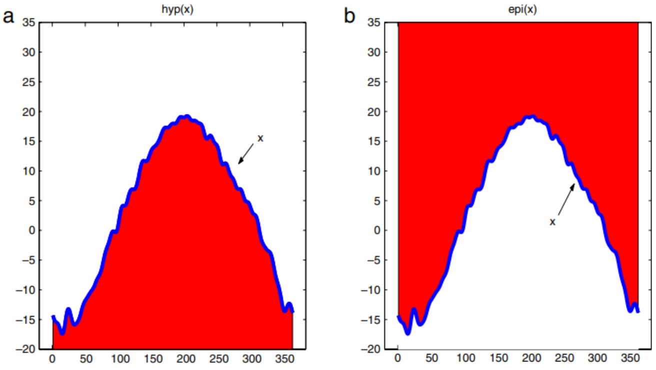

Fig. 1. (a) Hypograph and (b) epigraph of function x.

Profundidad vía "half-region"

Para una curva x defina:

$$\mathrm{DG}_1(\mathrm{x}) = \mathrm{P}(\mathrm{G}(\mathrm{x}) \subset \mathrm{hyp}(\mathrm{X})) = \mathrm{P}(\mathrm{x}(\mathrm{t}) \leq \mathrm{X}(\mathrm{t}), \mathrm{t} \in [\mathrm{a}, \mathrm{b}])$$

$$\mathrm{DG}_2(\mathrm{x}) = \mathrm{P}(\mathrm{G}(\mathrm{x}) \subset \mathrm{epi}(\mathrm{S})) = \mathrm{P}(\mathrm{x}(\mathrm{t}) \geq \mathrm{X}(\mathrm{t}), \mathrm{t} \in [\mathrm{a}, \mathrm{b}])$$

Entonces la profundidad half-region (poblacional) de x con respecto a P es definida por:

$$\mathrm{S}_{\mathrm{H}}(\mathrm{x}) = \mathsf{mín}\{\mathrm{DG}_1(\mathrm{x}), \mathrm{DG}_2(\mathrm{x})\}$$

Profundidad vía "half-region"

La profundidad half-region de x con respecto a $x_1, \ldots, x_n$ es:

$$S_{n,H}(x) = \min\{DG_{1,n}(x), DG_{2,n}(x)\}$$

donde

$$DG_{1,n}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{G(x_i) \subset hyp(x)\}}(x)$$

$$DG_{2,n}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{G(x_i) \subset epi(x)\}}(x).$$

Profundidad vía "half-region"

Cómo en los ejemplos anteriores, podemos definir: mediana, estadísticos de orden, rangos y medias truncadas a nivel α.

- El dato que maximiza la profundidad es la mediana, es decir,

$$\mathrm{m_n} = \mathrm{argmax}\, \mathrm{S_{n,H}(X)}$$

- El dato que tiene la profundidad en la posición k-ésima es el estadístico de orden $\mathrm{X_{(k)}}$

Profundidad vía “half-region” generalizada

De nuevo tenemos un problema similar con los empates. Por tanto, podemos debilitar la definición y considerar la proporción de que la curva está en el epigrafo o en el hipografo. Es decir, para

$$\mathrm{SL}(x) = \frac{1}{b-a} E[\lambda \{t \in [a,b] : x(t) \leq X(t)\}]$$
$$\mathrm{IL}(x) = \frac{1}{b-a} E[\lambda \{t \in [a,b] : x(t) \geq X(t)\}]$$

se define la profundidad:

$$\mathrm{MS_H}(x) = \mathsf{mín}\{\mathrm{SL}(x), \mathrm{IL}(x)\}$$

Profundidad vía “half-region” generalizada

Con respecto a un conjunto de curvas $x_1, \ldots, x_n$ para

$$SL_n(x) = \frac{1}{n(b-a)} \sum_{i=1}^{n} \lambda\{t \in [a,b] : x(t) \leq x_i(t)\}$$

$$IL_n(x) = \frac{1}{n(b-a)} \sum_{i=1}^{n} \lambda\{t \in [a,b] : x(t) \geq x_i(t)\}$$

se define la profundidad

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

Gracias