

Análisis de Datos Funcionales y Profundidad Estadística

Este trabajo se centra en el análisis de **datos funcionales** (Functional Data Analysis - FDA) y la **profundidad estadística** (statistical depth).

1. **Datos Funcionales:** A diferencia de los datos tradicionales que son puntos en un espacio multidimensional (\mathbb{R}^p), los datos funcionales se ven como **realizaciones parcialmente observadas de funciones aleatorias**. Ejemplos incluyen trayectorias de crecimiento o datos de imágenes cerebrales. Estos datos tienen características topológicas como la continuidad, suavidad y contigüidad. También tienen una característica inherente de ser **observados solo en un conjunto finito de puntos de discretización**, no completamente.
2. **Profundidad Estadística:** Es una herramienta para el análisis de datos que ayuda a **ordenar puntos en un conjunto de datos de "centro hacia afuera"**. Revela características de la distribución subyacente como la dispersión, forma y simetría. En el caso univariante (una dimensión), el orden es obvio, pero en dimensiones superiores no lo es. Tukey acuñó el término "depth" y propuso el "halfspace depth". El "simplicial depth" de Liu revitalizó la investigación en el área.
3. **El Problema:** Las definiciones existentes de profundidad para datos multivariantes (\mathbb{R}^p) no son adecuadas para datos funcionales porque **ignoran las características topológicas** y a menudo dan resultados sin sentido. Se necesitaba una definición formal y robusta para la profundidad en el espacio de funciones.
4. **La Contribución del Trabajo:** El objetivo principal de este documento es proporcionar una **definición formal de profundidad estadística para datos funcionales**. Esta definición se basa en **seis propiedades** (P-1 a P-6) que reconocen las características topológicas de los datos funcionales, como continuidad, suavidad y contigüidad.
5. **Propiedades Clave de la Definición (Resumidas):**
 - **P-1 (Distance invariance):** La profundidad no cambia si se aplica una transformación a los datos que preserva las distancias relativas (excepto por un factor de escala).
 - **P-2 (Maximality at centre):** La profundidad es máxima en el centro de simetría único de la distribución (si existe). Se considera el caso de procesos Gaussianos con media cero como un caso particular (P-2G).
 - **P-3 (Strictly decreasing w.r.t. the deepest point):** La profundidad disminuye estrictamente a medida que un punto se aleja del punto más profundo (el de máxima profundidad). Esto induce el orden de centro hacia afuera.
 - **P-4 (Upper semi-continuity in x):** La profundidad es semicontinua superiormente en el punto de evaluación (x). Es una propiedad mínima deseable.
 - **P-5 (Receptivity to convex hull width across the domain):** La profundidad debe dar menos peso a las regiones del dominio donde las funciones varían poco o se solapan mucho. Esto es importante debido a la observación parcial.
 - **P-6 (Continuity in P):** La profundidad es continua con respecto a la distribución de probabilidad (P). Esta propiedad es crucial porque **garantiza la convergencia de la profundidad empírica** (calculada a partir de una muestra) a la profundidad poblacional, lo cual es vital para la inferencia estadística. También aborda inherentemente el problema de la **observación parcial** de los datos funcionales, asegurando la convergencia incluso cuando se usan funciones reconstruidas a partir de datos discretos. Además, la satisfacción de P-6 confiere **robustez cualitativa** a la profundidad empírica, haciendo que las conclusiones no se vean afectadas desproporcionadamente por valores atípicos (outliers).
6. **Análisis de Profundidades Existentes:** Se examinaron seis construcciones de profundidad funcional existentes para ver cuáles satisfacían estas seis propiedades. Ninguna de las construcciones existentes satisface las seis propiedades.
7. **Resultados del Análisis:** La **h-depth** fue la que satisfizo el mayor número de propiedades (5 de 6, fallando en P-1). Sin embargo, los autores señalan que esto no significa que la h-depth sea la mejor, ya que la elección de la profundidad debe depender de la aplicación específica, pues cada propiedad tiene diferentes implicaciones. Varias otras profundidades (como band depth, modified band depth, half-region depth, modified half-region depth, random Tukey depth) fallaron en propiedades clave como P-3 (estrictamente decreciente) y P-5 (receptividad al ancho del convex hull).

En resumen, este trabajo propone una **nueva definición formal para la profundidad estadística de datos funcionales** basada en seis propiedades que capturan la naturaleza funcional y abordan desafíos como la observación parcial. Demuestra que las definiciones existentes no cumplen completamente estos criterios y evalúa su adhesión a estas propiedades.

Aplicación a Datos Longitudinales

Según la información proporcionada en las fuentes, los **datos longitudinales son un ejemplo prototípico de datos funcionales**.

A continuación, se explica cómo se aplica la profundidad estadística de datos funcionales a los datos longitudinales, basándose en las fuentes:

- 1. Los Datos Longitudinales como Datos Funcionales:** Los datos longitudinales consisten en **mediciones repetidas de individuos a lo largo del tiempo**. En el marco de este trabajo, estos se conceptualizan como **realizaciones parcialmente observadas de funciones aleatorias**.
 - El "dominio" de la función (V) es el tiempo, que se considera un subconjunto compacto de \mathbb{R} (es decir, un intervalo de tiempo).
 - La "función" representa la trayectoria o el comportamiento de una variable para cada individuo a lo largo de ese intervalo de tiempo.
 - La naturaleza de los datos longitudinales, donde las mediciones se toman solo en momentos específicos, significa que son **inherentemente parcialmente observados en un conjunto finito de puntos de discretización** (los momentos en que se toman las mediciones).
- 2. El Problema con la Profundidad Multivariante para Datos Longitudinales:** Las definiciones existentes de profundidad estadística para datos multivariantes (como si cada conjunto de mediciones longitudinales discretas fuera simplemente un vector en un espacio de alta dimensión) **ignoran las características topológicas** de las trayectorias longitudinales, como la continuidad y la suavidad a lo largo del tiempo. Aplicar ingenuamente la profundidad multivariante a estos datos puede llevar a **cálculos de profundidad absurdos**.
- 3. La Definición de Profundidad Funcional Aborda Esto:** La definición formal de profundidad estadística para datos funcionales propuesta en este trabajo se basa en **seis propiedades (P-1 a P-6)** diseñadas para reconocer las características topológicas como la continuidad, la suavidad y la contigüidad que son cruciales para los datos longitudinales.
 - Una propiedad clave, **P-6 (Continuidad en P)**, aborda específicamente el desafío de la **observabilidad parcial intrínseca** de los datos funcionales (incluidos los longitudinales).
 - P-6 garantiza que la profundidad empírica (calculada a partir de una muestra) **converge casi seguramente a la profundidad poblacional**. Esto es vital porque, en la práctica con datos longitudinales discretos, se suelen utilizar funciones **reconstruidas** (por ejemplo, mediante interpolación o suavizado) a partir de las observaciones discretas. Si la reconstrucción es adecuada (es decir, la medida de probabilidad de los datos reconstruidos converge casi seguramente a la medida de probabilidad poblacional), la propiedad P-6 asegura la **convergencia de la profundidad calculada sobre las funciones reconstruidas** a la profundidad poblacional de las funciones reales.
 - Además, la satisfacción de P-6 implica **robustez cualitativa** para la profundidad empírica, lo que es importante para asegurar que las conclusiones sobre los datos longitudinales **no se vean afectadas desproporcionadamente por valores atípicos** (outliers).
 - Otra propiedad relevante, **P-5 (Receptividad al ancho del convex hull a través del dominio)**, es importante para los datos longitudinales que pueden exhibir poca variabilidad o superponerse significativamente durante ciertos períodos de tiempo. P-5 asegura que las regiones del dominio (tiempo) donde las funciones longitudinales varían poco tengan un peso reducido en la asignación de profundidad.
- 4. Aplicaciones para Datos Longitudinales:** Al aplicar una definición de profundidad funcional que satisface estas propiedades (como la h-depth, que satisface la mayoría según el análisis comparativo), se pueden lograr varios objetivos para el análisis de datos longitudinales:

- **Ordenación:** Se puede **ordenar las trayectorias longitudinales de "centro hacia afuera"**. Esto permite identificar trayectorias típicas (profundidad alta) y atípicas (profundidad baja).
- **Visualización:** La profundidad puede usarse para crear gráficos que revelen características de la distribución de las trayectorias, como la dispersión, la forma y la simetría.
- **Inferencia Estadística:** La convergencia garantizada por P-6 es fundamental para usar la profundidad en problemas de inferencia estadística sobre la población de trayectorias longitudinales.
- **Clasificación:** La ordenación y la información sobre la proximidad relativa al centro (inducidas por P-2, P-3 y P-4) mejoran la capacidad de los clasificadores basados en profundidad para diferenciar grupos de trayectorias longitudinales.

En resumen, la estructura y las propiedades de la definición de profundidad funcional presentada en las fuentes están intrínsecamente diseñadas para manejar la naturaleza de los datos longitudinales, tratándolos como funciones y abordando el desafío práctico de que solo se observan en puntos discretos a lo largo del tiempo.

Aplicación a Datos de Estrés de Reloj Garmin

Basándonos en la información de las fuentes, los **datos de estrés extraídos de un reloj Garmin encajan perfectamente en el marco de los datos funcionales** que estamos discutiendo.

1. **Tus Datos de Estrés son Datos Funcionales (Longitudinales):** Las mediciones de estrés que recoge el reloj Garmin a lo largo del tiempo para una persona son un ejemplo clásico de **datos longitudinales**. Estos datos longitudinales son, a su vez, un **ejemplo prototípico de datos funcionales**. Cada conjunto de mediciones a lo largo del tiempo para un individuo se considera una **realización parcialmente observada de una función aleatoria**. La "función" representa la trayectoria de estrés de esa persona a lo largo del tiempo.
2. **La Necesidad de Profundidad Funcional:** Tu reloj registra mediciones de estrés en **momentos discretos**, no continuamente. Si trataras estas mediciones simplemente como un vector de datos multivariantes (una dimensión por cada momento de medición), y aplicarás una definición de profundidad multivariante, estarías **ignorando por completo la estructura temporal y las características topológicas** de la trayectoria de estrés, como la continuidad o la suavidad a lo largo del tiempo. Esto podría llevar a resultados poco significativos. La **definición formal de profundidad funcional** propuesta en las fuentes está diseñada específicamente para reconocer estas **características topológicas**.
3. **Propiedades Clave que te Ayudan con Datos de Reloj (Observación Parcial):**
 - **P-6 (Continuidad en P):** Esta es una propiedad **fundamental** para tus datos. El reloj Garmin te proporciona datos de forma **discreta**, lo que significa que tus trayectorias de estrés están **inherentemente observadas de forma parcial**. En la práctica, probablemente tendrás que **reconstruir** la función completa de estrés (por ejemplo, usando suavizado o interpolación) a partir de estas mediciones discretas. P-6 garantiza que la **profundidad empírica** (calculada a partir de tus datos muestrales, posiblemente reconstruidos) **convergerá a la profundidad poblacional** (la profundidad de la verdadera trayectoria de estrés subyacente). Esto es **crucial para poder hacer inferencia estadística** sobre los patrones de estrés basados en tus datos limitados.
 - **Robustez (derivada de P-6):** El cumplimiento de P-6 confiere **robustez cualitativa** a la profundidad empírica. Esto significa que las conclusiones que obtengas sobre la profundidad de las trayectorias de estrés **no se verán afectadas de forma desproporcionada por valores atípicos** (outliers) o posibles errores de medición del reloj.
 - **P-5 (Receptividad al ancho del convex hull):** Podría ser relevante si hay momentos del día o situaciones en las que los niveles de estrés de todas las personas son muy similares o se solapan mucho (poca variabilidad en las funciones en esa región del dominio/tiempo). P-5 asegura que estas regiones donde las funciones varían poco tengan un **peso reducido** en el cálculo de la profundidad. Esto es importante porque una pequeña variación o error de medición en esas regiones de baja variabilidad podría distorsionar la profundidad si no se tuviera en cuenta esta propiedad.

4. **Aplicaciones para tus Datos de Estrés:** Aplicar una definición de profundidad funcional que satisfaga estas propiedades te permitiría:
- **Ordenar las trayectorias de estrés:** Podrías **ordenar las funciones de estrés de "centro hacia afuera"**. Esto te ayudaría a identificar qué personas tienen trayectorias de estrés más "típicas" o centrales (alta profundidad) frente a aquellas con patrones de estrés más "atípicos" o extremos (baja profundidad).
 - **Identificar Outliers:** Las trayectorias con baja profundidad serían **candidatos a outliers**, lo que te ayudaría a detectar individuos con patrones de estrés inusualmente altos, bajos o erráticos.
 - **Visualizar:** Podrías crear gráficos basados en la profundidad que te muestren la **dispersión, forma y simetría de las trayectorias de estrés** en tu grupo de estudio.
 - **Comparación de grupos o clasificación:** Si tienes distintos grupos de personas (por ejemplo, bajo diferentes condiciones), la profundidad podría usarse para comparar la distribución de las trayectorias de estrés entre esos grupos o incluso construir clasificadores para diferenciar a las personas basándose en sus patrones de estrés.
5. **Elección de una Profundidad Existente:** Las fuentes analizan seis construcciones de profundidad funcional [43, Table 1]. Aunque ninguna satisface las seis propiedades, la **h-depth es la que cumple más (5 de 6)**, fallando solo en P-1. Sin embargo, para tus datos discretos de un reloj, la **satisfacción de P-6 es particularmente importante**. La h-depth, la Modified Band Depth y la Modified Half-Region Depth cumplen P-6 (con una restricción mencionada para la Band Depth y la Random Tukey Depth). La elección final dependerá de las propiedades que consideres más importantes para tu análisis específico de estrés.

En resumen, el marco de FDA y la definición de profundidad funcional son muy relevantes para analizar tus datos de estrés de Garmin. Las propiedades de la profundidad funcional, especialmente P-6, abordan directamente los desafíos de la observación parcial y la necesidad de robustez, lo que te permitiría realizar análisis significativos sobre las trayectorias de estrés.