Resumen Detallado del Análisis de Datos Funcionales (FDA)

Introducción

El Análisis de Datos Funcionales (FDA) constituye una metodología estadística robusta dedicada al análisis y la teoría de datos que inherentemente se manifiestan como funciones, imágenes, formas u objetos de naturaleza más general. La prevalencia de estos datos ha experimentado un crecimiento significativo, impulsada por los avances tecnológicos que facilitan su registro continuo en intervalos temporales o de manera intermitente en múltiples puntos discretos.

Naturaleza y Desafíos de los Datos Funcionales

Una característica fundamental de los datos funcionales reside en su dimensionalidad infinita. Esta alta dimensionalidad intrínseca presenta desafíos considerables, tanto desde una perspectiva teórica como computacional, cuya índole se modula según la estrategia de muestreo empleada. Sin embargo, esta misma estructura compleja alberga una abundante fuente de información, abriendo vastas oportunidades para la investigación y el análisis de datos sofisticado.

Tipos de Datos Funcionales

El documento centra su atención primordialmente en los datos funcionales de "primera generación", concebidos como realizaciones concretas de un proceso estocástico subyacente. Se alude brevemente a los datos de "próxima generación", los cuales se integran en objetos de datos más intrincados, potencialmente multivariados, correlacionados o que abarcan modalidades como imágenes o formas, ejemplificados por los datos de neuroimagen.

La adquisición de datos funcionales puede llevarse a cabo mediante diversas estrategias:

- **Muestreo Continuo (o en malla densa y regular):** La observación ininterrumpida de datos, exenta de errores, representa el escenario más tratable desde los puntos de vista teórico y metodológico.
- **Muestreo Discreto en Puntos Temporales:** Los datos muestreados discretamente en instantes temporales que pueden ser fijos o aleatorios, y que pueden variar entre los sujetos, se clasifican en:
    - **Densos:** Cuando el número de puntos (p_n) converge a infinito con la suficiente rapidez para permitir que los estimadores, como la función media, alcancen la tasa de convergencia paramétrica ($\sqrt{n}$) para métricas estándar. A pesar de lograr esta tasa, pueden exhibir un sesgo asintótico no despreciable.
    - **Dispersos (o Longitudinales):** Si el número de mediciones por sujeto (n_i) se encuentra limitado o acotado por una constante finita. Estos datos suelen demandar un mayor rigor teórico y metodológico, constituyendo un caso particular de datos no densos. Un ejemplo ilustrativo son los datos de recuento de CD4 visualizados en un gráfico de tipo "spaghetti".

En términos generales, los diseños de muestreo se categorizan en no densos (sin alcanzar la tasa $\sqrt{n}$), densos (alcanzando la tasa $\sqrt{n}$ con sesgo asintótico) y ultra-densos (alcanzando la tasa $\sqrt{n}$ sin sesgo asintótico). Los esquemas de muestreo disperso conllevan las tasas de convergencia más lentas.

Robustez ante el Error de Medición

Una ventaja distintiva del FDA radica en su capacidad para integrar de manera natural los errores de medición (ruido aleatorio) que inevitablemente contaminan las observaciones. Estos errores se modelan como fluctuaciones aleatorias alrededor de una trayectoria subyacente suave, manifestándose exclusivamente en los instantes temporales en los que se efectúan las mediciones (e_{ij}).

Conceptos Estadísticos Fundamentales

Los pilares conceptuales del FDA incluyen la función media $\mu(t) = E(X(t))$ y la función de covarianza $\Sigma(s,t) = \text{cov}(X(s), X(t))$.

La estimación de estos parámetros se adapta al plan de muestreo específico. Para mallas temporales variables (datos dispersos), se recurre a suavizadores no paramétricos aplicados al diagrama de dispersión de datos agrupados $\{(t_{ij}, Y_{ij})\}$ para la estimación de la media. La covarianza se estima mediante el suavizado de un diagrama de dispersión bidimensional de covarianzas "en bruto" $(Y_{ik} - \hat{\mu}(t_{ik}))(Y_{il} - \hat{\mu}(t_{il}))$, excluyendo los términos diagonales que incorporan la varianza del error de medición. La varianza del error de medición $\sigma^2(t)$ se puede inferir a partir de la distancia entre el pico de la diagonal de la covarianza "en bruto" y la superficie de covarianza suavizada.

Se exploran diversos esquemas de ponderación para la estimación, contrastando la ponderación equitativa de cada observación (donde sujetos con mayor cantidad de datos ejercen una influencia proporcionalmente mayor) con la ponderación uniforme por sujeto. La elección del esquema óptimo depende intrínsecamente del diseño del estudio, analizándose en un marco unificado que considera funciones de ponderación generales y sus propiedades asintóticas (convergencia L_2, L_∞, normalidad asintótica). El esquema de igual ponderación por sujeto demuestra ser más eficiente para datos ultra-densos o ciertos tipos de datos densos, mientras que la ponderación uniforme por observación resulta superior para una gama más amplia de planes, incluyendo los datos dispersos.

Las tasas de convergencia de los estimadores están intrínsecamente ligadas al diseño de muestreo. Se observa una "transición de fase" entre tasas no paramétricas y paramétricas $(\sqrt{n})$ en función de la densidad del muestreo. La estimación en datos dispersos típicamente conduce a tasas de convergencia no paramétricas más lentas.

Inferencia Estadística

La inferencia estadística en el contexto del FDA, incluyendo pruebas de hipótesis para la comparación de funciones medias (pruebas de dos muestras y ANOVA funcional), ha sido objeto de considerable atención. Asimismo, se han propuesto pruebas para distribuciones y funciones de covarianza. La construcción de bandas de confianza simultáneas para datos densos y dispersos también ha sido investigada, aunque el problema aún no se considera completamente resuelto. Los principales desafíos emanan de la dimensionalidad infinita y la naturaleza no paramétrica de la función objetivo. El fenómeno de "transición de fase" influye en la distribución límite del estimador de la media, que converge a un proceso gaussiano. Para datos ultra-densos, el proceso límite posee media cero; para datos densos, se encuentra descentrado debido al sesgo asintótico, análogo al sesgo en la estimación de funciones de regresión escalares no paramétricas. Esta distinción implica la necesidad de emplear metodologías diferenciadas para la construcción de bandas de confianza en datos ultra-densos, densos y dispersos.

Análisis de Componentes Principales Funcionales (FPCA)

Una herramienta fundamental y ampliamente utilizada en el FDA es el Análisis de Componentes Principales Funcionales (FPCA). El FPCA desempeña un papel crucial en la reducción de dimensionalidad, transformando datos funcionales de dimensionalidad infinita en un vector finito de puntuaciones aleatorias no correlacionadas, denominadas componentes principales funcionales (FPCs). Los conceptos fundacionales se remontan a los trabajos de Grenander, Karhunen, Loève y Rao, con un marco más comprehensivo desarrollado por Dauxois y Pousse.

La expansión de Karhunen-Loève expresa una función aleatoria $X_i(t)$ como una suma (infinita en teoría, truncada a K términos en la práctica) de la función media $\mu(t)$ más combinaciones lineales de eigenfunciones (ϕ_k) ponderadas por las puntuaciones FPC (A_{ik}):

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} A_{ik}\phi_k(t)$$

Los A_{ik} son incorrelacionados, con media cero y varianza λ_k.

La reducción de dimensionalidad se logra mediante la aproximación de la función utilizando los primeros K términos de esta expansión, concentrando esencialmente la información en el vector K-dimensional de puntuaciones $A_i = (A_{i1}, ..., A_{iK})$.

El FPCA se distingue de otras expansiones de base (splines, Fourier) en su capacidad para explicar la mayor parte de la variabilidad en los datos en el sentido L_2 para un número fijo K de componentes.

La estimación de los eigen-componentes (eigenfunciones y eigenvalores) se realiza aproximando la superficie de covarianza estimada en una malla y aplicando una descomposición espectral matricial. Para datos dispersos, la convergencia de los eigen-componentes estimados se ve influenciada por el método de suavizado empleado. Una cuestión abierta es el desarrollo de un enfoque que estime directamente la eigenfunción sin la necesidad de estimar explícitamente la covarianza.

La selección óptima del número de componentes (K) y de los parámetros de suavizado sigue siendo un desafío. Se han propuesto procedimientos ad hoc (gráfico de sedimentación, fracción de varianza explicada), criterios pseudo-AIC/BIC y validación cruzada, aunque esta última tiende a sobreajustar los modelos.

El FPCA ha sido extensamente estudiado para datos completamente observados, densos y, con mayor complejidad, dispersos. También se ha extendido para incorporar covariables, tanto vectoriales como funcionales.

El FPCA es sensible a la presencia de valores atípicos (outliers), que pueden manifestarse como mediciones inusuales en puntos específicos o como formas funcionales atípicas. Se requiere el desarrollo de métodos robustos y una mayor investigación en esta área. Herramientas exploratorias como los box plots funcionales también resultan útiles.

Aplicaciones del FPCA: El FPCA motiva el concepto de modos de variación para la visualización y descripción de la variabilidad inherente en los datos. Estos modos a menudo poseen interpretaciones significativas. Se aplica en la regresión principal de componentes funcionales (proyectando predictores funcionales en los FPCs), clasificación y clustering funcional. Además, facilita la identificación de modelos paramétricos más parsimoniosos para datos longitudinales, como los modelos de efectos mixtos basados en los componentes dominantes.

Problemas Inversos y Correlación Funcional

Un desafío significativo en el FDA reside en los problemas inversos, particularmente en la regresión funcional y en ciertas medidas de correlación funcional. Esto se debe a la compacidad del operador de covarianza Σ, lo que conlleva operadores inversos no acotados. La regularización se erige como una necesidad rutinaria para abordar estos problemas mal planteados.

La correlación funcional ha sido objeto de considerable investigación.

El Análisis de Correlación Canónica Funcional (FCCA) busca identificar funciones peso (u, v) que maximicen la correlación entre combinaciones lineales de dos procesos funcionales (X, Y). Representa una extensión del CCA multivariado. Sin embargo, sufre de ser un problema mal planteado debido a la naturaleza de dimensionalidad infinita de las direcciones de proyección, lo que puede conducir a sobreajuste si el tamaño de la muestra es insuficiente. La formulación del FCCA implica el análisis de los eigenvalores de un operador $R = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}$, y las inversas de los operadores de covarianza son no acotadas. En la práctica, se requiere regularización, ya que de lo contrario la primera correlación canónica siempre sería igual a uno. El FCCA exige datos densamente registrados, y la búsqueda de un FCCA efectivo para datos dispersos sigue siendo un problema abierto. A pesar de estos desafíos, el FCCA puede emplearse para implementar la regresión funcional utilizando las funciones peso canónicas como bases.

Se han propuesto medidas alternativas de correlación funcional que eluden el problema inverso. Una de ellas es la correlación singular, definida a partir de la base singular del operador de covarianza conjunta. Otra es la correlación dinámica, basada en el coseno del ángulo en el espacio L_2 entre las partes "dinámicas" de las funciones (obtenidas sustrayendo la integral de la función, que representa la parte "estática").

Regresión Funcional

La regresión funcional constituye un área activa de investigación. Las técnicas empleadas varían en función del tipo de respuesta y de las covariables (funcionales o vectoriales).

Un caso común es la Regresión Lineal Funcional (FLM) con respuesta escalar y covariables funcionales: $Y = \beta_0 + \int_{\mathcal{I}} X(t)\beta(t)dt + \epsilon$. Este modelo representa una extensión del modelo lineal multivariado. Los enfoques de estimación incluyen la expansión en bases (como el FPCA) o métodos penalizados (utilizando splines). El FLM con respuesta escalar a menudo se presenta como un problema mal planteado que requiere regularización, similar al FCCA, debido al operador de covarianza. La tasa de convergencia para la estimación de β depende de la velocidad de decaimiento de los eigenvalores y de la regularidad de β. Bajo ciertas condiciones, la predicción puede alcanzar la tasa $\sqrt{n}$.

El FLM puede extenderse para acomodar múltiples covariables funcionales y vectoriales.

Modelos de Coeficientes Variables: $Y(s) = \beta(s)X(s) + \epsilon(s)$. En este modelo, el coeficiente $\beta(s)$ varía con el tiempo y no constituye un problema mal planteado.

FLM con Respuesta Funcional: $Y(s) = \int_{\mathcal{I}_X} \beta(s,t)X(t)dt + \epsilon(s)$, donde $\beta(s,t)$ representa una superficie de coeficientes. Este modelo sí es un problema mal planteado que requiere regularización. El "modelo lineal funcional histórico" es un caso especial donde $\beta(s,t) = 0$ para $s < t$, implicando que solo el pasado influye en el presente.

Modelos Lineales Funcionales Generalizados (GFLM): Estos modelos incorporan una función de enlace no lineal g: $Y = g(\beta_0 + \int_{\mathcal{I}} X(t)\beta(t)dt) + \epsilon$. Se han estudiado tanto con g conocida como desconocida. La presencia de la función de enlace no lineal complica el análisis, requiriendo descomposiciones específicas. Los modelos de índice único múltiple extienden la idea del predictor lineal único.

Se han explorado generalizaciones de estos modelos para coeficientes variables y para el FLM con respuesta funcional.

El FPCA ajustado por covariables puede emplearse para respuestas funcionales y covariables vectoriales, modelando los efectos de las covariables en las puntuaciones FPC, aunque requiere datos densos.

Clustering y Clasificación Funcional

El clustering y la clasificación de datos funcionales son herramientas esenciales en el FDA, ilustradas con el ejemplo de datos de velocidad vehicular. El clustering es un aprendizaje no supervisado, mientras que la clasificación es supervisada.

Clustering Funcional: Este proceso de aprendizaje no supervisado a menudo extiende métodos tradicionales como k-means o el clustering jerárquico al dominio funcional. Las funciones medias de los clústeres actúan como centros. Un enfoque común consiste en proyectar los datos en un espacio de baja dimensión (como el espacio de los FPCs) antes de aplicar algoritmos de clustering multivariado. Sin embargo, si la estructura de covarianza es relevante, el uso exclusivo de las medias resulta insuficiente.

El clustering con subespacios como centros aprovecha la estructura estocástica de las funciones aleatorias (representación de Karhunen-Loève: media + eigenfunciones). Los clústeres se identifican minimizando la discrepancia entre las funciones y su proyección en los subespacios definidos por cada clúster.

La determinación del número óptimo de clústeres es un problema abierto, con procedimientos propuestos para identificar el número significativo de clústeres (como el procedimiento de prueba funcional hacia adelante) o enfoques bayesianos que incorporan la incertidumbre.

Clasificación Funcional: Este procedimiento de aprendizaje supervisado tiene como objetivo asignar una etiqueta de grupo a un nuevo objeto funcional. Los enfoques populares incluyen métodos basados en la regresión (como la regresión logística funcional, una variante del GFLM) y el análisis discriminante funcional (extensión del LDA clásico, a menudo utilizando el FPCA para la reducción de dimensionalidad). La clasificación bayesiana asigna la etiqueta con la mayor probabilidad posterior. Se ha introducido

la noción de "clasificación perfecta" para probabilidades de clasificación errónea que desaparecen asintóticamente.

Métodos No Lineales

Los métodos no lineales se vuelven necesarios cuando los datos funcionales exhiben características intrínsecamente no lineales que disminuyen la efectividad de los métodos lineales.

Las extensiones directas a modelos no lineales, como los GFLM o los modelos de índice único, incorporan funciones de enlace no lineales con un predictor lineal.

La extensión directa del suavizado de núcleo a predictores funcionales se enfrenta a la "maldición de la dimensionalidad" en espacios funcionales. Esto subraya la importancia de seleccionar una métrica apropiada o de imponer restricciones estructurales.

Los modelos aditivos ofrecen otra herramienta poderosa para la reducción de dimensionalidad. Pueden ser aditivos en el tiempo (asumiendo un efecto aditivo del tiempo, a veces restrictivo) o aditivos en las FPCs (frecuencia-aditivos), donde la media condicional es una suma de funciones de las FPCs. Los modelos aditivos en FPCs son más fáciles de implementar y analizar bajo supuestos de independencia.

El aprendizaje de dinámicas a partir de datos funcionales implica el uso de derivadas. Modelos como $X(1)(t) = f(X(t), t) + Z(t)$ pueden aprender la función de deriva f a partir de los datos. El comportamiento de la función de deriva (como $\beta(t)$ en $X(1)(t) = \beta(t)X(t) + Z(t)$) puede caracterizar dinámicas como la regresión a la media o el comportamiento explosivo.

Cuando los datos funcionales presentan variación temporal (diferencias en el timing de las características), esto se conoce como el problema de registro de curvas o time warping. Ignorar estas diferencias distorsiona las funciones medias. Métodos como el alineamiento por landmarks o el dynamic time warping buscan alinear las curvas. Identificar y separar la variación de amplitud y tiempo es un desafío si ambas están presentes.

El **aprendizaje de variedades funcionales (Functional Manifold Learning)** es un enfoque para el time warping y otras características no lineales. Asume que los datos funcionales se encuentran en una variedad no lineal de baja dimensión incrustada en el espacio funcional. Métodos como Isomap pueden descubrir esta estructura y proporcionar representaciones de baja dimensión. Esto permite definir una media de variedad y componentes de variedad funcionales análogos a la media y componentes de FPCA, que pueden ser más eficientes para datos en variedades. La visualización de los primeros componentes de FPCA puede sugerir la presencia de una variedad (por ejemplo, formas de "herradura"). La distancia entre funciones es un input esencial para Isomap, y debe ajustarse para datos dispersos.

El campo de FDA ha ampliado su alcance, incluyendo el análisis de datos longitudinales y datos de alta dimensión. Los desarrollos futuros incluyen el análisis de datos funcionales dependientes (series de tiempo funcionales), datos funcionales multivariados y datos funcionales indexados espacialmente. El surgimiento de nuevos tipos de datos ("segunda generación") está impulsando los desarrollos recientes en FDA. Estos nuevos datos provienen de seguimiento continuo (movimiento, salud), sensores (tráfico, clima), genómica y finanzas. El documento menciona herramientas útiles para el FDA como métodos de suavización (kernel, splines) y áreas teóricas como el análisis funcional y los procesos estocásticos. Señala que esta revisión es una selección subjetiva de temas de interés para los autores.