AGENCIA NACIONAL DE INVESTIGACIÓN E INNOVACIÓN

Informe final publicable de proyecto
Datos funcionales y en alta dimension.

Código de proyecto ANII: FCE_1_2019_1_156054

Fecha de cierre de proyecto: 01/05/2023

CHOLAQUIDIS NOBLÍA, Alejandro (Responsable Técnico - Científico)
CUEVAS, Antonio (Investigador)
FEBRERO BANDE, Manuel (Investigador)
FRAIMAN, Ricardo (Investigador)
GAMBOA, Fabrice (Investigador)
GHATTAS, Badih (Investigador)
MORENO ROMERO, Leonardo Fabian (Investigador)
PATEIRO LOPEZ, Beatriz (Investigador)

UNIVERSIDAD DE LA REPÚBLICA. FACULTAD DE CIENCIAS (Institución Proponente) \\
UNIVERSIDAD DE LA REPÚBLICA. FACULTAD DE CIENCIAS

Resumen del proyecto

La estadística de datos en el alta dimensión así como la de datos funcionales requiere de nuevas técnicas ya que los métodos tradicionales en estadística clásica resultan inadecuados para abordarlos. Esto se debe a que en alta dimensión a menudo el tamaño de la muestra es menor a la dimensión de los datos. En el caso de datos funcionales requiere del manejo de procesos estocásticos.
En este proyecto abordamos el estudio de tres problemas estadísticos importantes cuando los datos se encuentran en espacios no necesariamente euclideanos utilizando su estructura métrica.
Consideramos los siguientes problemas,
1) el modelo lineal funcional de respuesta escalar, mediante el uso de RKHS (Reproducing Kernel Hilbert spaces)
2) test de hipótesis para datos binarios en alta dimensión
3) estimación de conjuntos bi-convexos
Analizamos las propiedades asintóticas de los procedimientos, generamos nuevos algoritmos para resolverlos, estudiamos su comportamiento para tamaños de muestra moderada por simulaciones y presentamos ejemplos de aplicación en datos reales: en biología, en particular en genética, text mining, datos nutricionales, datos electorales, target marketing, reconocimiento de patrones y de imágenes.

Ciencias Naturales y Exactas / Matemáticas / Estadística y Probabilidad / Estadística matemática
Palabras clave: Datos Funcionales / Estadística en espacios métricos / Datos binarios en alta dimensión /

Antecedentes, problema de investigación, objetivos y justificación.

A nivel nacional somos el único grupo que trabaja en estadística en espacios abstractos, en particular en los problemas mencionados. Tenemos potenciales interacciones con el grupo de imágenes del Instituto de Ingeniería Eléctrica de la Facultad de Ingeniería, y con el grupo de Bioinformática del Instituto Pasteur de Montevideo. También integramos el Centro Académico de Análisis de Big Data (CABIDA) . Interactuaremos con el ICT4V de la ANNI. Algunos problemas y áreas que desarrollamos fueron:

1) Datos Funcionales.
A nivel internacional hay un importante desarrollo de la estadística de datos funcionales (FDA). Nuestro grupo ha trabajado desde hace años en esta temática (ver por ejemplo The Oxford Handbook of Functional Data (2011)). En esta área tenemos una importante interacción con el Prof. Antonio Cuevas de la Universidad Autónoma de Madrid, los profesores Manuel Febrero y Beatriz Pateiro de la Universidad de Santiago de Compostela.
Entre otros problemas, abordaremos:

2) Estudio del modelo lineal funcional de respuesta escalar mediante el uso de "Reproducing Kernel Hilbert Spaces" (RKHS por sus siglas en inglés), que se ha popularizado, recientemente, en la comunidad estadística.
2) Test de hipótesis para datos binarios en alta dimensión. Caso no independiente.
En muchas aplicaciones, particularmente en genética nos encontramos con un conjunto de datos binarios que generalmente no son independientes, o pensamos que no lo son. En ese caso, podemos considerar que tenemos un conjunto de variables aleatorias Bernoulli, Z1,..., Zn, cada una con un parámetro diferente, que en principio no son independientes. Si lo fueran, y todos los parámetros fueran iguales, tendríamos una distribución Binomial(n,p). Si los parámetros fueran distintos, pero las variables aún independientes, tendríamos una distribución Poisson-Binomial, pero eso ocurriría en ambos casos si las variables fueran independientes.
Si asumimos que no lo son, el problema a abordar es testear si la distribución conjunta de las variables Z1, ...Zn tiene cierta estructura o no.

3) Estimación de Conjuntos. La idea de convexidad es transversal a toda la matemática, y en particular ha sido ampliamente estudiada en estadística, en estimación de conjuntos. Existen algoritmos que dan el cierre convexo de una muestra, y resultados asintóticos para su convergencia al conjunto, no obstante desde un punto de vista práctico resulta muy restrictiva en muchas aplicaciones (por ejemplo en el caso en que el conjunto es el hábitat de especies). Para eso

se han propuesto diversas generalizaciones: r-convexidad, rho-cono convexidad, biconvexidad.
Varias de ellas dependen de parámetros desconocidos, por ejemplo la r-convexidad. Su estimación, o tener una idea de si se esta sobre estimando o no, es clave para obtener estimadores consistentes, y será uno de los problemas que abordaremos. Otro de los problemas es el estudio de la biconvexidad, la cual es una restricción de forma que no depende de parámetros.

Metodología/Diseño del estudio

Aunque la estadística teórica puede ser un campo muy amplio, usamos la siguiente metodología.

Estimación de parámetros: En estadística, un parámetro es una característica de una población. Por ejemplo, la media y la desviación estándar son parámetros que describen la localización y la dispersión de una distribución, respectivamente. La estimación de parámetros es el proceso de usar datos para estimar estos parámetros. Por ejemplo, la estimación de máxima verosimilitud es un método que busca el valor del parámetro que maximiza la probabilidad de observar los datos dados. La estimación de momentos, por otro lado, implica igualar los momentos (como la media y la varianza) de una muestra a los de la población.

Modelado estadístico: En el modelado estadístico, los datos se representan a través de un modelo matemático que se ajusta a estos datos. Estos modelos pueden ser tan simples como una línea recta (como en la regresión lineal) o tan complejos como se necesite. Los modelos estadísticos se usan para describir las relaciones entre variables, predecir futuros resultados, o probar hipótesis científicas. Algunos de los tipos de modelado más comunes incluyen la regresión lineal, el análisis de varianza (ANOVA), la regresión logística y los modelos de series temporales.

Teoremas límite: Los teoremas límite son resultados que describen lo que sucede a medida que el tamaño de la muestra se acerca al infinito. La Ley de los Grandes Números, por ejemplo, establece que el promedio de una muestra de valores independientes e idénticamente distribuidos converge hacia su media real a medida que el tamaño de la muestra se acerca al infinito. El Teorema del Límite Central, otro teorema límite fundamental, dice que la suma de una gran cantidad de variables aleatorias independientes e idénticamente distribuidas tiende a seguir una distribución normal, independientemente de la forma de la distribución original. Estos teoremas son fundamentales para muchas técnicas en estadística e inferencia.

Resultados, análisis y discusión

En nuestro estudio, se obtuvieron resultados de consistencia asintótica casi segura de los estimadores que se propusieron. En términos sencillos, esta consistencia asintótica sugiere que nuestros estimadores se vuelven más precisos a medida que aumenta la cantidad de datos o el tamaño de la muestra. Esto indica que nuestros estimadores son altamente confiables y pueden ser robustos en una variedad de aplicaciones, incluyendo la estimación de parámetros y conjuntos de nivel en varios casos de problemas de estimación de conjuntos.

Además, fuimos capaces de obtener tasas de convergencia para nuestros estimadores dentro de un marco muy general. Las tasas de convergencia son una característica crucial de cualquier estimador, ya que indican la velocidad a la que las estimaciones se acercan a los verdaderos valores a medida que aumenta el tamaño de la muestra. Nuestros resultados demostraron que estas tasas de convergencia son aplicables en una amplia variedad de contextos, incluyendo espacios métricos y variedades, lo que proporciona una base sólida para la utilidad de nuestros estimadores en diversos problemas de estimación.

Enfocándonos en los modelos lineales funcionales, hemos examinado de cerca el modelo lineal funcional del Espacio Hilbert de Kernel Reproductor (RKHS). A través de nuestro análisis, pudimos demostrar que el modelo RKHS puede igualar y en muchos casos superar el rendimiento del modelo L2 clásico. Una de las ventajas más significativas del modelo RKHS es su interpretabilidad, lo que facilita la comprensión de los resultados y sus implicaciones en el contexto de la investigación.

Finalmente, es importante destacar que nuestras propuestas se basan en gran medida en los métodos plug-in, que son ampliamente reconocidos por su versatilidad y robustez en el campo de la estadística. Un ejemplo específico de esto es nuestra propuesta basada en la estimación universalmente consistente del alcance (reach). Esta estrategia nos permitió proporcionar un estimador muy general del alcance de un conjunto, ampliando aún más la utilidad y aplicabilidad de nuestros estimadores en un amplio rango de problemas.

Para mas detalle, ver las publicaciones obtenidas:

• Level sets of depth measures in abstract spaces (2023) – Cholaquidis, Fraiman, Moreno. TEST

• A counter example on a Borsuk conjecture. (2023) - Cholaquidis. - Appl. Gen. Topol. Vol 24, Nro. 1

• Universally consistent estimation of the reach – (2023) - Cholaquidis, Fraiman, Moreno – Journal of Statistical planning and Inference.

• Estimation of surface area – (2022) – Aaron, Cholaquidis, Fraiman. Electronic Journal of Statistics. Vol. 16, No. 2, 3751-3788

• Weighted lens depth: Some applications to supervised classification. (2022) Cholaquidis, Fraiman, Gamboa, Moreno. The Canadian Journal of Statistics.

• Level set and density estimation on manifolds. (2022) Cholaquidis, Fraiman, Moreno. Journal of Multivariate Analysis. Vol. 89

• Set estimation under biconvexity restrictions. (2020) Cholaquidis, Cuevas. ESAIM.

• On boundary detection. (2020) Aaron, Cholaquidis. Annales De L'institut Henri Poincaré, Probabilité et Statistique.

• A quantitative Heppes theorem and multivariate Bernoulli distributions Ricardo F., L. Moreno , T. Ransford Journal of the Royal Statistical Society Series B: Statistical Methodology, (2023)

Conclusiones y recomendaciones

Recomendamos seguir investigando en estadística en datos complejos ya que en particular quedan muchos problemas abiertos y conjeturas, algunas que mencionamos a continuación.
Detallamos algunas conclusiones de algunos de los trabajos (se recomienda ver las versiones publicadas para mas detalles):

En el trabajo On the Functional Regression Model and its Finite-Dimensional Approximations, exploramos un marco matemático, diferente al enfoque clásico L2 para el problema de la regresión lineal con la variable explicativa funcional X y la respuesta escalar Y.
Esta formulación matemática incluye, como casos particulares, los modelos de dimensión finita obtenidos considerando como variables explicativas un conjunto finito X(t1), ..., X(tp) de marginales. Esto permitiría, por ejemplo, comparar dichos modelos para la selección de variables o considerar, dentro de un marco unificado, el estudio del comportamiento asintótico de los modelos a medida que el número p de covariables crece hasta el infinito para un análisis reciente en el modelo de regresión logística). También cabe destacar que en el caso funcional, el análisis asintótico como p ? ? aparece de manera más natural que en el caso de los estudios generales de regresión, ya que las nuevas covariables entrantes son homogéneas en naturaleza ya que provienen del único depósito predefinido de las marginales unidimensionales del proceso X(t). Queda por estudiar el comportamiento de los autovalores de la matriz de covarianza, para incluir mas casos al modelo estudiado.

En el trabajo Universally consistent estimation of the reach nos enfrentamos a un problema importante en la estimación de conjuntos relacionado con la teoría de la medida geométrica: la estimación del alcance, r0, de un conjunto M ? R^d, que incluye el caso en el que M es una variedad. Se propone un estimador universalmente consistente de r0, en el sentido de que no se requieren suposiciones excepto que r0 sea positivo. Demostramos que basándonos en una muestra finita con densidad f con respecto a la medida de Lebesgue, no es posible determinar si el alcance del soporte de f es cero o no. El resultado de consistencia está relacionado con la convergencia a cero de dH(M, Xn). Conjeturamos que, dado cualquier estimador consistente del alcance, rn, y dada cualquier secuencia ?n ? 0, es posible encontrar un conjunto (dependiendo de ?n), con alcance positivo, para el cual, rn converge al alcance a una tasa más lenta que ?n.

Sin embargo, bajo la suposición adicional débil de estandarización, proporcionamos tasas de convergencia para el estimador propuesto. Para el caso en que M es una variedad, adaptamos nuestro procedimiento añadiendo un método de corrección de sesgo. Una forma alternativa de abordar este problema es considerar el estimador rn = (1 - ??n)rˆn, pero la elección óptima del parámetro ? debe ser abordada.

En relación a "Weighted lens depth: Some applications to supervised classification": Estudiar profundidades en espacios métricos generales permite abordar problemas en los que la estructura de los datos no es ni de dimensión finita ni funcional. Algunos ejemplos importantes, entre otros, son los árboles filogenéticos dados por la evolución de la historia de los genes, o datos pertenecientes a una variedad desconocida. Para espacios métricos separables y completos, hemos estudiado las principales propiedades de la profundidad de lente (LD, por sus siglas en inglés), hemos demostrado la consistencia de la versión empírica con la poblacional y hemos proporcionado tasas de convergencia casi paramétricas. Para variedades riemannianas, hemos proporcionado una extensión de LD, llamada WLDp, que es más flexible que LD y es capaz de capturar también información sobre la geometría de la distribución subyacente. También hemos demostrado resultados de consistencia para el estimador plug-in de WLDp. Para la mayoría de los problemas mencionados anteriormente, demostramos, utilizando WLDp, que es crucial que la profundidad tenga en cuenta no solo la estructura geométrica de los datos sino también la distribución subyacente. Utilizamos LD y WLDp en el problema de clasificación supervisada, poniendo en acción el método de profundidad-profundidad introducido en Liu (1990). Estas clasificaciones se realizan en ejemplos de datos simulados y reales. Obtenemos un rendimiento mejor que algunos competidores y resultados interesantes en los ejemplos de datos de la vida real.

Referencias bibliográficas

Berrendero, J.R., Bueno-Larraz, B. and Cuevas, A. (2019). An RKHS model for variable selection in functional linear regression. Journal of Multivariate Analysis, 170, 22-45.
Berrendero, J.R., Cuevas, A. and Torrecilla, J.L. (2018). On the use of reproducing kernel Hilbert spaces in functional classification. Journal of the American Statistical Association, 113, 1210-1218.
Yuan, Ming, and T. Tony Cai. "A reproducing kernel Hilbert space approach to functional linear regression." The Annals of Statistics 38.6 (2010): 3412-3444.
Estimación de Conjuntos:
Cholaquidis, A., Cuevas, A. Fraiman, R. (2014) On Poincaré cone property. Ann. Statist.
Cholaquidis, A., Fraiman, R., Lugosi, G. and Pateiro-López, B. (2016) - Set estimation from reflected Brownian motion. J. R. Stat. Soc. Ser. B. Stat. Methodol.
Cuevas, A., Fraiman, R. and Pateiro-López, B. (2012). On statistical properties of sets fulfilling rolling-type conditions. Adv. in Appl. Probab.
Aaron, C. Cholaquidis, A., and Fraiman, R. 2017). A Generalization of the maximal-spacings in several dimensions and a convexity test. Extremes.
Aumann, R. and Hart, S. (1986). Bi-convexity and bi-martingales. Isr. J. Math., 54, 159–180.
Bae, S. W., Lee, C., Ahn, H. K., Choi, S., and Chwa, K. Y. (2009). Computing minimum-area rectilinear convex hull and L-shape. Computational Geometry,
Aumann, R. and Hart, S. (1986). Bi-convexity and bi-martingales. Isr. J. Math.,
Alegría-Galicia, C., Orden, D., Seara, C., and Urrutia, J. (2018). On the beta -hull of a planar point set. Computational Geometry.

Licenciamiento