Comparative Analysis of M-Estimators and Trimmed Means for Fuzzy Set-Valued Data

Beatriz Sinova Fernández
With the collaboration of Pedro Terán

Universidad de Oviedo

Resumen del trabajo candidato para el Premio Ramiro Melendreras
XXXVI Congreso Nacional de Estadística e Investigación Operativa

seio 2016

Toledo, 2016

Resumen del trabajo

Los datos imprecisos constituyen uno de los nuevos tipos de datos que van surgiendo al ampliarse aún más el campo de aplicaciones de la Estadística. Esta contribución se ha centrado en el caso de los datos difusos (o *fuzzy*) de cualquier dimensión (es decir, conjuntos difusos), con el fin de generalizar al máximo los resultados. La formalización matemática de numerosos conceptos estadísticos referentes a los mecanismos aleatorios que generan conjuntos difusos ha dado paso al interés por el desarrollo de técnicas estadísticas que faciliten su análisis. Sin embargo, debido a las peculiaridades del espacio de los conjuntos difusos, su análisis estadístico presenta una serie de retos.

La filosofía perseguida en este trabajo es la de conservar, en la medida de lo posible, las nociones y técnicas clásicas utilizadas para variables y vectores aleatorios, aunque dicho enfoque no carece de obstáculos, como se ha adelantado ya. En este sentido, cabe destacar la falta de linealidad del espacio de valores de los conjuntos difusos con la aritmética usual y, como consecuencia, la inexistencia de una operación diferencia que simultáneamente esté siempre bien definida y conserve las propiedades que posee en el caso real en relación con la operación suma. Por lo tanto, las distancias entre conjuntos difusos serán una herramienta clave para el desarrollo o adaptación de técnicas estadísticas que involucren diferencias.

El objetivo del trabajo candidato al Premio Ramiro Melendreras es la propuesta de medidas de localización robustas. Hay que remarcar la importancia y el interés de este tipo de estudio, ya que en los últimos años el número de métodos estadísticos que se han generalizado a conjuntos difusos aleatorios ha ido creciendo exponencialmente, pero la prática totalidad de ellos están basados en la media tipo Aumann. La media tipo Aumann es una generalización del concepto de media de una variable aleatoria y, como tal, goza de múltiples propiedades, muy convenientes desde el punto de vista tanto estadístico como probabilístico, pero también hereda la excesiva sensibilidad ante la existencia de datos atípicos (*outliers*) o los cambios en los datos.

Con el propósito de reforzar en este aspecto los métodos existentes en la literatura y cimentar adecuadamente los modelos aún por desarrollar, se vuelve acuciante la búsqueda de medidas que resuman la tendencia central de los conjuntos difusos aleatorios y que no se vean tan sumamente influenciadas por errores u observaciones atípicas como la media tipo Aumann. En pocas palabras, que salvaguarden las conclusiones estadísticas obtenidas a través de esos métodos, incluso bajo esos supuestos cambios con respecto a las condiciones iniciales o ideales para el estudio.

Esta descripción de la línea de investigación abordada en el trabajo es bastante general, así que en las siguientes secciones se profundizará en las propuestas previas que se pueden encontrar en la literatura y, finalmente, en una explicación un poco más detallada de las aportaciones de este estudio.

Estado del arte

A continuación, se comentarán sucintamente las ideas que ya se habían recogido en la literatura y están relacionadas con el problema planteado en este trabajo. El planteamiento como tal, de búsqueda de medidas de localización robustas para conjuntos difusos aleatorios de cualquier dimensión, es completamente novedoso. Hay que tener en cuenta que, ya en el caso más sencillo de

conjuntos difusos aleatorios de dimensión uno (es decir, los llamados "números difusos aleatorios"), el estudio robusto completo de la localización, en cuanto a la amplia propuesta de medidas y el uso de herramientas de la Teoría de la Robustez Estadística para fundamentar la mejora que suponen en comparación con la media tipo Aumann, se originó con la tesis doctoral de la autora principal de este trabajo. Algunos artículos de Ban *et al.* [1], Bodjanova [2], Grzegorzewski [8], Kersten [9] y Yamashiro [17, 18] proponen ciertas extensiones del concepto de mediana para números difusos aleatorios, pero sus ideas no se complementan con ningún estudio formal de la robustez. Cabe citar, en este sentido, las publicaciones de Sinova *et al.* [15, 16, 14] acerca de la noción de mediana y la adaptación de la M-estimación a los números difusos aleatorios.

Sin embargo, es complicado, o incluso inviable, adaptar muchas de estas propuestas al caso más general de conjuntos difusos aleatorios de dimensión $p > 1$. Las nociones de mediana, por ejemplo, se basan en procedimientos *ad hoc* cuya validez depende de la posibilidad de caracterizar adecuadamente los números difusos aleatorios. En el caso unidimensional, existen representaciones caracterizadoras sencillas de manejar, pero en el caso p-dimensional se hace necesario operar con la función soporte, lo que entraña una gran dificultad práctica. Por todo ello, se vuelve imprescindible la búsqueda de otras alternativas.

Si bien las medidas de localización que se presentan en este trabajo fueron introducidas en su momento en la tesis doctoral de la autora, lamentablemente su estudio era incompleto. Los resultados expuestos aquí son a la vez básicos y cruciales (como la medibilidad, la consistencia o el punto de ruptura) y se han desarrollado expresamente para este estudio.

Por lo tanto, la originalidad de este trabajo radica en la fundamentación teórica de las medidas de localización robustas para conjuntos difusos aleatorios introducidas en dicha tesis doctoral (*M-estimadores de localización y medias recortadas difusas*), con el consiguiente impacto que esta teoría pueda tener en el análisis estadístico de elementos aleatorios imprecisos.

Contribuciones de este trabajo

En esta sección se ahondará en el contenido, sobre todo la parte novedosa, del trabajo candidato. En primer lugar, nos centraremos en la M-estimación.

La motivación para la extensión de los M-estimadores de localización de variables aleatorias reales al contexto de conjuntos difusos aleatorios proviene, principalmente, de su aceptación general como una herramienta exitosa en el estudio de la tendencia central. Recuérdese que los M-estimadores se definen a través de un problema de minimización de cierta función de pérdida evaluada en las distancias euclídeas entre los valores que toma la variable aleatoria y la clase de números reales, pues la idea subyacente es la de proponer estimadores intermedios entre la media (con función de pérdida la función cuadrado) y la mediana (con función de pérdida la función valor absoluto).

Existen en la literatura algunos trabajos que adaptan los M-estimadores de localización al caso de elementos aleatorios con valores en un espacio de Hilbert, si bien sus autores los enmarcan únicamente en el contexto de la estimación robusta de la función núcleo de densidad (véanse Kim [10] y Kim y Scott [11, 12]). Gracias a la existencia de un encaje isométrico del espacio de conjuntos difusos en un cono convexo de cierto espacio de Hilbert, mediante el uso de ciertas métricas y la función soporte, es posible introducir los M-estimadores de localización

para conjuntos difusos aleatorios.

De hecho, el estudio de la *existencia y unicidad de solución* de los M-estimadores difusos (el Teorema de Representación) es una aplicación del análisis publicado por Kim y Scott. Hay que tener en cuenta que bajo las condiciones de este teorema, es posible expresar los M-estimadores difusos como combinaciones lineales convexas de las observaciones muestrales (que son, evidentemente, conjuntos difusos) y, por lo tanto, garantizar que la M-estimación pertenecerá al espacio paramétrico (cónico) de los valores difusos. Es, por lo tanto, muy interesante asentar una teoría sobre esta propuesta y cimentar la aplicación de esta medida en el futuro.

En este sentido, el trabajo presentado al Premio Ramiro Melendreras *analiza formalmente la medibilidad, consistencia y robustez* de esta alternativa, así como otras propiedades convenientes en la práctica, como la equivarianza por traslaciones y escala o la simetría cuando se analizan conjuntos difusos simétricos. Nótese que muchas demostraciones no se pueden abordar igual que en el caso intervalar, ya que el nuevo espacio topológico no satisface las mismas propiedades que $\mathbb{R}\times[0,\infty)$. Por eso, hay propiedades subyacentes que se han tenido que comprobar explícitamente para este estudio, como, por ejemplo, que el espacio de los conjuntos difusos es localmente compacto con la métrica utilizada, y demostraciones que han tenido que plantearse de manera absolutamente diferente.

En cuanto a la segunda alternativa propuesta para el análisis de la tendencia central de los conjuntos difusos aleatorios, se trata de las medias recortadas. A diferencia de la M-estimación, las medias recortadas han sido estudiadas en los espacios de Hilbert de manera recurrente en la literatura (como muestra, véanse Cuesta-Albertos y Nieto-Reyes [6], Cuesta-Albertos y Fraiman [5], Cuevas y Fraiman [7] o López-Pintado y Romo [13], entre otros). Recientemente, se ha contemplado su particularización al caso difuso a través del encaje isométrico nombrado antes (ver Colubi y González-Rodríguez [3]), pero, en cualquier caso, el tratamiento formal de las propiedades que se han comentado para los M-estimadores difusos tampoco se ha llevado a cabo previamente con las medias recortadas, a excepción de la consistencia (Cuesta-Albertos y Fraiman [5, 4]). Así, también se *analiza formalmente la medibilidad, equivarianza por traslaciones y escala, simetría y robustez* de las medias recortadas difusas (y los resultados son, de hecho, válidos en los espacios más generales de Hilbert).

Finalmente, y con el objetivo de ilustrar el uso de los conjuntos difusos, se presentan un ejemplo con datos reales y varios *estudios de simulación*, en los que se calculan la media tipo Aumann, los M-estimadores difusos con algunas de las funciones de pérdida más habituales y las medias recortadas difusas. En el caso de los estudios de simulación, se ha tratado de representar las situaciones más comunes en la práctica, ya que es preciso hacer notar que no se dispone todavía de modelos de distribuciones para conjuntos difusos aleatorios que sean suficientemente realistas. La dificultad práctica para encontrar ejemplos de conjuntos difusos p-dimensionales en la literatura (por la gran complicación técnica que lleva asociado el manejo de la función soporte), ha limitado la parte práctica al caso unidimensional.

Referencias

[1] Ban A, Coroianu L, Grzegorzewski P (2013) A fixed-shape fuzzy median of a fuzzy sample. *8th Conference of the European Society for Fuzzy Logic and Technology, EUSFLAT 2013, Advances in Intelligent Systems Research* **32**: 215–222

[2] Bodjanova S (2005) Median value and median interval of a fuzzy number. *Inform. Sci.* **172(1)**: 73–89

[3] Colubi A, González-Rodríguez G (2015) Fuzziness in data analysis: Towards accuracy and robustness. *Fuzzy Sets Syst.* **281**: 60–271

[4] Cuesta-Albertos JA, Fraiman R (2007) Impartial trimmed k-means for functional data. *Comp. Stat. Data Anal.* **51(10)**: 4864–4877

[5] Cuesta-Albertos JA, Fraiman R (2006) Impartial trimmed means for functional data. In: *Data Depth: Robust Multivariate Statistical Analysis, Computational Geometry and Applications*, Vol. 72, USA: Amer. Math. Soc. in DIMACS Series: pp. 121–145

[6] Cuesta-Albertos JA, Nieto-Reyes A (2008) The random Tukey depth. *Comp. Stat. Data Anal.* **52(11)**: 4979–4988

[7] Cuevas A, Fraiman R (2009) On depth measures and dual statistics. A methodology for dealing with general data. *J. Multivar. Anal.* **100**: 753–766

[8] Grzegorzewski P (1998) Statistical inference about the median from vague data. *Control and Cybernetics* **27**: 447–464

[9] Kersten PR (1995) The fuzzy median and the fuzzy MAD. In: *Proc. 3rd International Symposium on Uncertainty Modeling and Analysis and Annual Conference of the North American Fuzzy Information Processing Society, (ISUMA - NAFIPS'95)*, IEEE: pp. 85–88

[10] Kim JS (2011) *Kernel Methods for Classification with Irregularly Sampled and Contaminated Data.* PhD Thesis, University of Michigan (http://deepblue.lib.umich.edu/bitstream/handle/2027.42/89858/stannum_1.pdf?sequence=1)

[11] Kim JS, Scott CD (2011) On the robustness of kernel density M-estimators. In: *Proc. 28th Int. Conf. Mach. Lear.*, Bellevue, Washington: pp. 697–704

[12] Kim JS, Scott CD (2012) Robust kernel density estimation. *J. Mach. Learn. Res.* **13**: 2529–2565

[13] López-Pintado S, Romo J (2009) On the concept of depth for functional data. *J. Amer. Stat. Assoc.* **104(486)**: 718–734

[14] Sinova B, Gil MA, Colubi A, Van Aelst S (2012) The median of a random fuzzy number. The 1-norm distance approach. *Fuzzy Sets Syst.* **200**: 99–115

[15] Sinova B, Gil MA, Van Aelst S. M-estimates of location for the robust central tendency of fuzzy data. *IEEE Trans. Fuzzy Syst.*, accepted for publication, DOI:10.1109/TFUZZ.2015.2489245.

[16] Sinova B, Pérez-Fernández S, Montenegro M (2015) The wabl/ldev/rdev median of a random fuzzy number and statistical properties. In: *Strengthening Links Between Data Analysis and Soft Computing. Advances in Intelligent Systems and Computing* (Grzegorzewski P, Gagolewski M, Hryniewicz O, Gil MA, Eds), Adv. Int. Syst. Comp. Vol. 315. Springer, Heidelberg: pp. 143–150

[17] Yamashiro M (1995) The median for a L-R fuzzy number. *Microelectronics Reliability* **35(2)**: 269–271

[18] Yamashiro M (1994) The median for a trapezoidal fuzzy number. *Microelectronics Reliability* **34(9)**: 1509–1511