Robustez en el Simplex

Marina Fragalá
Instituto de Ciencias, Universidad Nacional de General Sarmiento, Argentina
mfragala@campus.ungs.edu.ar

Los datos composicionales son vectores de coordenadas positivas con suma fija (sin pérdida de generalidad se puede suponer 1). Tienen su representación geométrica en el Simplex. En 1897 Pearson K. [1] puso de manifiesto en su artículo "On a form of spurious correlation", la inadecuación de los métodos estadísticos clásicos para el estudio de estos datos. En 1982 Aitchison J. [2] definió en el Simplex una geometría que le otorga las propiedades de un espacio de Hilbert. Existen isomorfismo que transforman las composiciones del Simplex de d partes en $\mathbb{R}^{d-1}$ (con la geometría euclídea usual) y de esta manera es posible hacer ajustes clásicos. Una de las primeras y más conocidas distribuciones en el Simplex es la Dirichlet. Una generalización de esta distribución se debe a Graf M. [3] y la llamó Beta Generalizada Simplicial (SGB) que permite una gama más amplia de ajustes. La misma autora desarrolló, a partir de una muestra de covariables $x_1, \ldots, x_n$, un modelo lineal generalizado para estimar por máxima verosimilitud los parámetros de dicha ditribución. En nuestro trabajo proponemos un método robusto y eficiente para resolver el problema de la estimación de los parámetros de la SGB. Para esta propuesta utlizamos la estimación robusta τ de Ben M. y Yohai V. [4] y adaptamos el método de detección de outliers de Gervini D. y Yohai V. [5]. A partir de estos procedimientos construimos un estimador robusto inicial. Luego entre el estimador robusto inicial y el de máxima verosimilitud de Graf M., planteamos una combinación convexa parecida a la propuesta por Marazzi A. [6]. El parámetro de ajuste lineal de la combinación convexa se busca de tal modo que haya un balanceo entre robustez y eficiencia. Si no hay outliers, la combinación convexa devuelve el estimador de máxima verosimilitud. Si hay outliers, prevalece el robusto. Mediante estudios de simulación de Monte Carlo, nuestro estimador demostró ser suficientemente robusto para muestras contaminadas y al mismo tiempo altamente eficiente en muestras limpias. Mostraremos sus propiedades asintóticas y lo compararemos con otros estimadores en un ejemplo con datos reales.

Trabajo en conjunto con Alfio Marazzi, Université de Lausanne, Suiza y Marina Valdora, Instituto de Ciencias, Universidad de Buenos Aires y CONICET, Argentina.

Referencias

[1] Pearson K., .ºn a form of spurious correlation", Proc. Roy. Soc (1897), vol 60, pag 489-498.

[2] Aitchison J., "The statistical analysis of compositional data", Journal of the Royal Statistical Society, Series B (1982), 44, 139-160.

[3] Graf M., Regression for compositions based on a generalization of the Dirichlet distribution", Statistical Methods and Applications (2020), 913-936.

[4] Ben M., Matinez E., Yohai V., Robust estimation for the multivariate linear model based on a Tau-scale", Journal of Multivariate Analysis, Elsevier (2006), Vol 97, Nº7, pag 166-1622.

[5] Gervini D., Yohai V., .ª class of robust and fully efficient regression estimators", The Annals of Statistics, (2002), 30(2):583–616.

[6] Marazzi A., ¨Improving the efficiency of robust estimators for the generalized linear model", Stats, MDP (2021), Vol 4, Nº1, pág 88-107.