XVII Congreso Galego de Estatística e Investigación de Operacións
Ourense, 23 - 25 de outubro de 2025

Robust Nonparametric Regression for Compositional Data

A. M. Bianco[1], G. Boente[1], W. González-Manteiga[2]
Francisco Gude Sampedro[2] and A. Pérez-González[31]

[1]Universidad de Buenos Aires & CONICET, Argentina
[2]Universidad de Santiago de Compostela, Santiago de Compostela, Spain
[3]Universidad de Vigo, Campus Ourense, Ourense, Spain

ABSTRACT

Statistical analysis in other contexts than the Euclidean space $\mathbb{R}^D$ has become an important area of research in modern statistics. Traditional methods, which are designed to handle Euclidean spaces equipped with the usual inner product, often fail to leverage the underlying geometric structures inherent in non-standard domains. In fact, these procedures typically assume that data points are drawn from a space where the classical notions of distance and operations in $\mathbb{R}^D$, such as averaging or summation, are applicable as in $\mathbb{R}^D$. However, when dealing with data that lie in more complex spaces, this assumption may not be still valid.

A relevant class among non-standard settings is that of compositional data, which has gained a lot of attention due to their great potential in applications. A feature of these data is that they are multivariate vectors that lie in the simplex, that is, the components of each vector are positive and sum up a constant value. This fact poses a challenge to the analyst due to the internal dependency of the components which exhibit a spurious negative correlation. Since classical multivariate techniques are not appropriate in this scenario, it is necessary to endow the simplex of a suitable algebraic-geometrical structure, which is a starting point to develop adequate methodology and strategies to handle compositions.

A way to analyse compositional data relies on isomorphisms that transform from the simplex to the real space. Egozcue et al.(2003) introduced the isometric log-ratio (ilr) transformation of D-parts compositions that gives coordinates in $\mathbb{R}^{D-1}$ ($ilr-$coordinates) improving previously proposed transformations.

Motivated by a real dataset, our goal is to deepen our understanding of the relationship between diet composition and glycaemic control. We adopt a nonparametric approach to address regression problems involving real-valued responses and compositional covariates due to its flexibility and ability to handle complex relationships. Aware of the potential damage that outliers may produce on traditional statistical techniques, we introduce two families of robust estimators within the framework of nonparametric regression for compositional data.

Through a numerical experiment we compare the performance of the robust estimators with those proposed in Di Marzio et al.(2015) under different contamination schemes. The analysis of the motivating real dataset further reveals the advantages of using the proposed robust procedure.

Keywords: Aitchison geometry, Compositional data, Glycaemia measurement, Nonparametric regression, Robust estimators

REFERENCES

Di Marzio, M., Panzera, A. and Venieri, C. (2015). Non-parametric regression for compositional data. *Statistical Modelling*, 15:113–133.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35:279–300.