Paper Number: 1449

Interpretable principal balances for compositional data in geochemistry

Martín-Fernández, J.A.[1], Egozcue, J.J.[2] and Pawlowsky-Glahn, V.[1]

[1]Universitat de Girona, Campus Montilivi, Edif. P4, E- 17071 Girona, Spain, josepantoni.martin@udg.edu
[2]Universitat Politècnica de Catalunya, Campus Nord, Edif. B2, E-08034 Barcelona, Spain

In applied statistics, reduction of dimensionality of a data set is one of the most common tasks performed by analysts. Principal component analysis (PCA) based methods are well-known and efficient techniques for this goal [1]. A new set of uncorrelated variables, the principal components (PCs), is constructed as linear combinations of original variables with the aim to capture the most variation. However, the interpretation of PCs is not always easy because they are linear combinations of all the original variables. Among the alternatives to PCA made to provide simpler components, the most common techniques are: rotation [1], LASSO [2, 3] and clustering variables [4].

PCA and related methods cannot directly be applied to raw compositional data (CoDa) because its sample space, the simplex, has a particular geometry [5]. The ideas introduced in the early 1980s to deal with CoDa using logratios were further developed, leading to particular isometric coordinates [6], known as balances, obtained through a sequential binary partition (SPB) . Following this methodology, principal balances (PBs) were defined as a compromise between the concepts of PCs and balances [7]. The basic idea in this approach is that, for interpretative purposes, it would be better to have each PC as a balance of a group of components against another group, leaving irrelevant components with a zero weight. That is, PBs are defined as a sequence of orthonormal balances which maximize successively the explained variance in a CoDa set. To compute the PBs three algorithms were proposed: minimize angular proximity to PCs, hierarchical clustering of compositional parts, and maximize explained variance via hierarchical balances. The drawbacks of these algorithms when the data are high-dimensional motivates the introduction of sparse principal balances using the algorithm proposed in [3]. However, this algorithm has other difficulties: it does not use the log-contrast structure of balances, it only constructs non-overlapping groups of parts, and it depends on two arbitrary tuning parameters.

I this work we revise the above methods and propose a new method based on the approaches introduced by Chipman and Gu [9] and Enki et al. [4]. This approach is illustrated using the data set from the groundwater geochemical analysis conducted in the late 1990s in Bangladesh [10].

References:

[1] Jolliffe IT (2002) *Principal Component Analysis*: Springer-Verlag, New York

[2] Jolliffe IT, Trendafilov NT and Uddin M (2003) A modified principal component technique based on the LASSO J. Comput. Graph. Statist. 12: 531–547

[3] Witten DM, Tibshirani R, and Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10(3): 515-534

[4] Enki DG, Trendafilov NT, Jolliffe IT (2013) A clustering approach to interpretable principal components J. of Applied Statistics 40(3): 583-599

[5] Pawlowsky-Glahn V and Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex SERRA 15(5): 384-398

[6] Egozcue J and Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis Math Geo 37: 795–820

[7] Pawlowsky-Glahn V, Egozcue J and Tolosana-Delgado R (2011) Principal balances In: *Egozcue JJ, Tolosana-Delgado R and Ortego M (eds), Proceedings of the 4th International Workshop on Compositional Data Analysis, Girona (Spain)*: 1–10
[8] Mert MC, Filzmoser P and Hron K (2015) Sparse principal balances Stat Model 15(2): 159–174
[9] Chipman HA and Gu H (2005) Interpretable dimension reduction J. Appl. Stat. 32: 969–987
[10] Pawlowsky-Glahn V, Egozcue JJ, Olea RA and Pardo-Igúzquiza E (2015) Cokriging of compositional balances including a dimension reduction and retrieval of original units The Journal of The Southern African Institute of Mining and Metallurgy 115: 59-72