Greenacre, Michael

Variable selection in compositional data analysis using pairwise logratios. (English)

Zbl 1421.86020
Math. Geosci. 51, No. 5, 649-682 (2019).

Summary: In the approach to compositional data analysis originated by John Aitchison, a set of linearly independent logratios (i.e., ratios of compositional parts, logarithmically transformed) explains all the variability in a compositional data set. Such a set of ratios can be represented by an acyclic connected graph of all the parts, with edges one less than the number of parts. There are many such candidate sets of ratios, each of which explains 100% of the compositional logratio variance. A simple choice consists in using additive logratios, and it is demonstrated how to identify one set that can serve as a substitute for the original data set in the sense of best approximating the essential multivariate structure. When all pairwise ratios of parts are candidates for selection, a smaller set of ratios can be determined by automatic selection, but preferably assisted by expert knowledge, which explains as much variability as required to reveal the underlying structure of the data. Conventional univariate statistical summary measures as well as multivariate methods can be applied to these ratios. Such a selection of a small set of ratios also implies the choice of a subset of parts, that is, a subcomposition, which explains a maximum percentage of variance. This approach of ratio selection, designed to simplify the task of the practitioner, is illustrated on an archaeometric data set as well as three further data sets in an "Appendix". Comparisons are also made with existing proposals for selecting variables in compositional data analysis.

MSC:

86A32 Geostatistics
62H11 Directional data; spatial statistics

Keywords:

compositional data; logratio transformation; logratio analysis; logratio distance; multivariate analysis; ratios; subcompositional coherence; univariate statistics; variable selection

Software:

R; vegan

Full Text: DOI

References:

[1] Aitchison, J., The statistical analysis of compositional data (with discussion), J R Stat Soc B, 44, 139-177, (1982) · Zbl 0491.62017

[2] Aitchison, J., Principal component analysis of compositional data, Biometrika, 70, 57-65, (1983) · Zbl 0515.62057

[3] Aitchison J (1986) The statistical analysis of compositional data. Chapman \& Hall, London. Reprinted in 2003 with additional material by Blackburn Press · Zbl 0688.62004

[4] Aitchison, J., Relative variation diagrams for describing patterns of compositional variability, Math Geol, 22, 487-511, (1990)

[5] Aitchison, J., On criteria for measures of compositional difference, Math Geol, 24, 365-379, (1992) · Zbl 0970.86531

[6] Aitchison, J.; Anderson, TW (ed.); Olkin, I. (ed.); Fang, KT (ed.), Principles of compositional data analysis, 73-81, (1994), Hayward

[7] Aitchison J (2003) Compositional data analysis: where are we and where should we be heading? In: Proceedings of the compositional data analysis workshop, CoDaWork'03, Girona, Spain. CD-format, ISBN 84-8458-111-X · Zbl 1177.86017

[8] Aitchison J (2005) A concise guide to compositional data analysis. http://ima.udg.edu/Activitats/CoDaWork05/A_concise_guide_to_compositio Accessed 29 May 2018

[9] Aitchison, J.; Egozcue, JJ, The statistical analysis of compositional data: where are we and where should we be heading?, Math Geol, 37, 829-850, (2005) · Zbl 1177.86017

[10] Aitchison, J.; Greenacre, MJ, Biplots for compositional data, J R Stat Soc Ser C (Appl Stat), 51, 375-392, (2002) · Zbl 1111.62300

[11] Aitchison, J.; Barceló-Vidal, C.; Martín-Fernández, JA; Pawlowsky-Glahn, V., Logratio analysis and compositional distance, Math Geol, 32, 271-275, (2000) · Zbl 1101.86309

[12] Bacon-Shone, J.; Pawlowsky, V. (ed.); Buccianti, A. (ed.), A short history of compositional data analysis, 3-11, (2011), Chichester

[13] Baxter, MJ; Cool, HEM; Heyworth, MP, Principal component and correspondence analysis of compositional data: some similarities, J Appl Stat, 17, 229-235, (1990)

[14] Baxter, MJ; Beardah, CC; Cool, HEM; Jackson, CM, Compositional data analysis of some alkaline glasses, Math Geol, 37, 183-196, (2005)

[15] Benzécri J-P (1973) Analyse des Données. Tôme II, Analyses des Correspondances. Dunod, Paris · Zbl 0297.62038

[16] Bóna M (2006) A walk through combinatorics: an introduction to enumeration and graph theory, 2nd edn. World Scientific Publishing, Singapore · Zbl 1127.05001

[17] Box, GEP; Cox, DR, An analysis of transformations, J Roy Stat Soc Ser B, 26, 211-252, (1964)

[18] Cortés, J., On the Harker variation diagrams; a comment on "The statistical analysis of compositional data. Where are we and where should we be heading?" by Aitchison and Egozcue (2005), Math Geosc, 41, 817-828, (2009) · Zbl 1178.86017

[19] Dijksterhuis, G.; Frøst, MB; Byrne, DV, Selection of a subset of variables: minimisation of Procrustes loss between a subset and the full set, Food Qual Prefer, 13, 89-97, (2002)

[20] Filzmoser, P.; Hron, K.; Reimann, C., Univariate statistical analysis of environmental (compositional) data: problems and possibilities, Sci Total Environ, 407, 6100-6108, (2009)

[21] Gittins R (1985) Canonical analysis: a review with applications in ecology. Springer, New York · Zbl 0576.62069

[22] Gower JC, Dijksterhuis GB (2004) Procrustes problems. Oxford University Press, Oxford · Zbl 1057.62044

[23] Greenacre, MJ, Power transformations in correspondence analysis, Comput Stat Data Anal, 53, 3107-3116, (2009) · Zbl 1453.62099

[24] Greenacre, MJ, Logratio analysis is a limiting case of correspondence analysis, Math Geosci, 42, 129-134, (2010)

[25] Greenacre MJ (2010b) Biplots in practice. BBVA Foundation, Bilbao. www.multivariatestatistics.org. Accessed 29 May 2018

[26] Greenacre, MJ, Measuring subcompositional incoherence, Math Geosc, 43, 681-693, (2011)

[27] Greenacre, MJ; Pawlowski-Glahn, V. (ed.); Buccianti, A. (ed.), Compositional data and correspondence analysis, 104-113, (2011), Chichester

[28] Greenacre, MJ, Contribution biplots, J Comput Graph Stat, 22, 107-122, (2013)

[29] Greenacre MJ (2016) Correspondence analysis in practice, 3rd edn. Chapman \& Hall/CRC, Boca Raton

[30] Greenacre, MJ; Lewi, PJ, Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements, J Classif, 26, 29-64, (2009) · Zbl 1276.62037

[31] Harary F, Palmer EM (1973) Graphical enumeration. Academic Press, New York · Zbl 0266.05108

[32] Harker A (1909) Natural history of the igneous rocks. Methuen, London

[33] Hron, K.; Filzmoser, P.; Donevska, S.; Fišerová, E., Covariance-based variable selection for compositional data, Math Geosci, 45, 487-498, (2013) · Zbl 1321.86023

[34] Hron, K.; Filzmoser, P.; Caritat, P.; Fišerová, E.; Gardlo, A., Weighted pivot coordinates for compositional data and their application to geochemical mapping, Math Geosci, 49, 777-796, (2017) · Zbl 1369.86019

[35] Kraft, A.; Graeve, M.; Janssen, D.; Greenacre, MJ; Falk-Petersen, S., Arctic pelagic amphipods: lipid dynamics and life strategy, J Plank Res, 37, 790-807, (2015)

[36] Krzanowski, WJ, Selection of variables to preserve multivariate data structure, using principal components, Appl Stat, 36, 22-33, (1987)

[37] Krzanowski WJ (2000) Principles of multivariate analysis: a user's perspective. Oxford University Press, Oxford

[38] Legendre P, Legendre L (2012) Numerical ecology, 3rd edn. Elsevier, Amsterdam

[39] Lewi, PJ, Spectral mapping, a technique for classifying biological activity profiles of chemical compounds, Arzneim Forsch (Drug Res), 26, 1295-1300, (1976)

[40] Lewi, PJ; Linden, GA (ed.), Multivariate data analysis in APL, 267-271, (1980), Amsterdam

[41] Lewi, PJ, Spectral map analysis. Factorial analysis of contrasts, especially from log ratios, Chemometr Intell Lab, 5, 105-116, (1989)

[42] Lewi, PJ, Spectral mapping, a personal and historical account of an adventure in multivariate data analysis, Chemometr Intell Lab, 77, 215-223, (2005)

[43] Lovell, D.; Müller, W.; Taylor, J.; Zwart, A.; Helliwell, C.; Pawlowski-Glahn, V. (ed.); Buccianti, A. (ed.), Proportions, percentges, ppm: do the molecular biosciences treat compositional data right?, 193-207, (2011), Chichester UK

[44] Martín-Fernández, JA; Pawlowsky-Glahn, V.; Egozcue, JJ; Tolosana-Delgado, R., Advances in principal balances for compositional data, Math Geosci, 50, 273-298, (2018) · Zbl 1407.62219

[45] Mert, MC; Filzmoser, P.; Hron, K., Sparse principal balances, Stat Model, 15, 159-174, (2015)

[46] Murtagh, F., Counting dendrograms: a survey, Discrete Appl Math, 7, 191-199, (1984) · Zbl 0528.62055

[47] Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H

(2015) vegan: community ecology package. R package version 2.3-2. https://CRAN.R-project.org/package=vegan. Accessed 11 June 2018

[48] Pawlowski-Glahn V, Buccianti A (eds) (2011) Compositional data analysis. Wiley, Chichester

[49] Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2007) Lecture notes on compositional data analysis. http://dugi-doc.udg.edu/bitstream/handle/10256/297/CoDa-book.pdf?sequence=1. Accessed 11 June 2018

[50] Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. Wiley, Chichester

[51] Rao, CR, The use and interpretation of principal component analysis in applied research, Sankhya A, 26, 329-358, (1964) · Zbl 0137.37207

[52] Tanimoto, S.; Rehren, T., Interactions between silicate and salt melts in LBA glassmaking, J Archaeol Sci, 35, 2566-2573, (2008)

[53] R core team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

[54] van den Boogaart KG, Tolosana-Delgado R (2013) Analyzing compositional data with R. Springer, Berlin · Zbl 1276.62011

[55] Wollenberg, AL, Redundancy analysis—an alternative for canonical analysis, Psychometrika, 42, 207-219, (1977) · Zbl 0354.92050

[56] Wouters, L.; Göhlmann, HW; Bijnens, L.; Kass, SU; Molenberghs, G.; Lewi, PJ, Graphical exploration of gene expression data: a comparative study of three multivariate methods, Biometrics, 59, 1131-1139, (2003) · Zbl 1274.62904