zbMATH Open
THE FIRST RESOURCE FOR MATHEMATICS

Egozcue, Juan José; **Pawlowsky-Glahn, Vera**

Compositional data: the sample space and its structure. (English) Zbl 1428.62220

Test 28, No. 3, 599-638 (2019).

Summary: The log-ratio approach to compositional data (CoDa) analysis has now entered a mature phase. The principles and statistical tools introduced by *J. Aitchison* [The statistical analysis of compositional data. Boca Raton, FL: CRC Press (1986; Zbl 0688.62004)] have proven successful in solving a number of applied problems. The algebraic-geometric structure of the sample space, tailored to those principles, was developed at the beginning of the millennium. Two main ideas completed the J. Aitchison's seminal work: the conception of compositions as equivalence classes of proportional vectors, and their representation in the simplex endowed with an interpretable Euclidean structure. These achievements allowed the representation of compositions in meaningful coordinates (preferably Cartesian), as well as orthogonal projections compatible with the Aitchison distance introduced two decades before. These ideas and concepts are reviewed up to the normal distribution on the simplex and the associated central limit theorem. Exploratory tools, specifically designed for CoDa, are also reviewed. To illustrate the adequacy and interpretability of the sample space structure, a new inequality index, based on the Aitchison norm, is proposed. Most concepts are illustrated with an example of mean household gross income per capita in Spain.

MSC:

62H12 Estimation in multivariate analysis
62H25 Factor analysis and principal components; correspondence analysis
62P20 Applications of statistics to economics
60F05 Central limit and other weak theorems
62G30 Order statistics; empirical distribution functions

Cited in **4** Reviews
Cited in **1** Document

Keywords:

simplex; equivalence class; isometric log-ratio coordinates; Euclidean space; Aitchison geometry; principal balances; dendrogram; principal components; biplot; household income; normal distribution on the simplex; logistic-normal; compositional data (CoDa); central limit theorem

Software:

coDaPack; R; zCompositions

Full Text: DOI

References:

[1] Äijö, T.; Müller, CL; Bonneau, R., Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing, Bioinformatics, 34, 372-380, (2018)

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

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

[4] Aitchison J (1986) The statistical analysis of compositional data. Monographs on statistics and applied probability. Chapman \& Hall Ltd., London (reprinted in 2003 with additional material by The Blackburn Press)

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

[6] Aitchison J (1994) Multivariate analysis and its applications, volume 24 of lecture notes—monograph series, chapter principles of compositional data analysis. Institute of Mathematical Statistics, Hayward, pp 73-81

[7] Aitchison J (1997) The one-hour course in compositional data analysis or compositional data analysis is simple. In: Pawlowsky-Glahn V (ed) Proceedings of IAMG'97—the III annual conference of the international association for mathematical geology, volume I, II and addendum, Barcelona (E). CIMNE, Barcelona, pp 3-35, ISBN 978-84-87867-76-7

[8] Aitchison, J.; Bacon-Shone, J., Log contrast models for experiments with mixtures, Biometrika, 71, 323-330, (1984)

[9] Aitchison, J.; Egozcue, JJ, Compositional data analysis: Where are we and where should we be heading?, Math Geol, 37,

© 2021 FIZ Karlsruhe GmbH

829-850, (2005) · Zbl 1177.86017

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

[11] Aitchison, J.; Shen, S., Logistic-normal distributions. Some properties and uses, Biometrika, 67, 261-272, (1980) · Zbl 0433.62012

[12] 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

[13] Aitchison, J.; Barceló-Vidal, C.; Martín-Fernández, JA; Pawlowsky-Glahn, V., Reply to letter to the editor by S. Rehder and U. Zier on ""'Logratio analysis and compositional distance"", Math Geol, 33, 849-860, (2001) · Zbl 1101.86310

[14] Aitchison J, Barceló-Vidal C, Egozcue JJ, Pawlowsky-Glahn V (2002) A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. In: Bayer U, Burger H, Skala W (eds) Proceedings of IAMG'02—the VIII annual conference of the international association for mathematical geology, vol I and II. Selbstverlag der Alfred-Wegener-Stiftung, Berlin, pp 387-392

[15] Atkinson, AB, On the measurement of inequality, J Econ Theory, 2, 244-263, (1970)

[16] Bacon-Shone J (2003) Modelling structural zeros in compositional data. In: Thió-Henestrosa S, Martín-Fernández JA (eds) Proceedings of CoDaWork'03, the 1st compositional data analysis workshop, Girona (E). Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork2003/

[17] Barceló-Vidal, C.; Martín-Fernández, JA, The mathematics of compositional analysis, Austrian J Stat, 45, 57-71, (2016)

[18] Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2001) Mathematical foundations of compositional data analysis. In: Ross G (ed) Proceedings of IAMG'01—the VII annual conference of the international association for mathematical geology, Cancun (Mex), p 20

[19] Billheimer, D.; Guttorp, P.; Fagan, W., Statistical interpretation of species composition, J Am Stat Assoc, 96, 1205-1214, (2001) · Zbl 1073.62573

[20] Buccianti, A.; Pawlowsky-Glahn, V., New perspectives on water chemistry and compositional data analysis, Math Geol, 37, 703-727, (2005) · Zbl 1103.62111

[21] Chayes F (1971) Ratio correlation. University of Chicago Press, Chicago, p 99

[22] Chen, J.; Zhang, X.; Li, S., Multiple linear regression with compositional response and covariates, J Appl Stat, 44, 2270-2285, (2017)

[23] Chipman, HA; Gu, H., Interpretable dimension reduction, J Appl Stat, 32, 969-987, (2005) · Zbl 1121.62347

[24] Comas-Cufí M, Thió-Henestrosa S (2011) Codapack 2.0: a stand-alone, multi-platform compositional software. See Egozcue et al. (2011c)

[25] Connor, RJ; Mosimann, JE, Concepts of independence for proportions with a generalization of the Dirichlet distribution, J Am Stat Assoc, 64, 194-206, (1969) · Zbl 0179.24101

[26] Daunis-i Estadella J, Barceló-Vidal J, Buccianti A (2006) Exploratory compositional data analysis. In: Compositional data analysis in the geosciences: from theory to practice, volume 264 of special publications. Geological Society, London, pp 161-174 · Zbl 1158.86333

[27] Finetti, B., Considerazioni matematiche sull'ereditarietà mendeliana, Metron, 6, 3-41, (1926) · Zbl 52.0542.05

[28] Egozcue, JJ, Reply to ""'On the Harker variation diagrams;..."" by J. A. Cortés, Math Geosci, 41, 829-834, (2009) · Zbl 1178.86018

[29] Egozcue, JJ; Jarauta-Bragulat, E., Differential models for evolutionary compositions, Math Geosci, 46, 381-410, (2014) · Zbl 1323.37051

[30] Egozcue, JJ; Pawlowsky-Glahn, V., Groups of parts and their balances in compositional data analysis, Math Geol, 37, 795-828, (2005) · Zbl 1177.86018

[31] Egozcue JJ, Pawlowsky-Glahn V (2011a) Basic concepts and procedures. See Pawlowsky-Glahn and Buccianti (2011), pp 12-28

[32] Egozcue JJ, Pawlowsky-Glahn V (2011b) Evidence information in Bayesian updating. See Egozcue et al. (2011c)

[33] Egozcue, JJ; Pawlowsky-Glahn, V., Evidence functions: a compositional approach to information (invited paper), Stat Oper Res Trans, 42, 1-24, (2018) · Zbl 1403.60008

[34] Egozcue JJ, Pawlowsky-Glahn V (2018b) Modelling compositional data. The sample space approach, Chapter 4, p XXV, 875. Handbook of mathematical geosciences—fifty years of IAMG. Springer, Berlin

[35] Egozcue, JJ; Pawlowsky-Glahn, V.; Mateu-Figueras, G.; Barceló-Vidal, C., Isometric logratio transformations for compositional data analysis, Math Geol, 35, 279-300, (2003) · Zbl 1302.86024

[36] Egozcue, JJ; Díaz-Barrero, JL; Pawlowsky-Glahn, V., Hilbert space of probability density functions based on Aitchison geometry, Acta Math Sin, 22, 1175-1182, (2006) · Zbl 1113.46016

[37] Egozcue JJ, Barceló-Vidal C, Martín-Fernández JA, Jarauta-Bragulat E, Díaz-Barrero JL, Mateu-Figueras G (2011a) Elements of simplicial linear algebra and geometry. See Pawlowsky-Glahn and Buccianti (2011), pp 141-157

[38] Egozcue JJ, Jarauta-Bragulat E, Díaz-Barrero JL (2011b) Calculus of simplex-valued functions. See Pawlowsky-Glahn and Buccianti (2011), pp 158-175

[39] Egozcue JJ, Tolosana-Delgado R, Ortego MI (eds) (2011c) Proceedings of the 4th international workshop on compositional data analysis, Sant Feliu de Guixols, Girona. CIMNE, Barcelona, ISBN 978-84-87867-76-7

© 2021 FIZ Karlsruhe GmbH

[40] Egozcue, JJ; Daunis-i-Estadella, J.; Pawlowsky-Glahn, V.; Hron, K.; Filzmoser, P., Simplicial regression. The normal model, J Appl Probab Stat, 6, 87-108, (2012) · Zbl 06186205

[41] Egozcue, JJ; Pawlowsky-Glahn, V.; Tolosana-Delgado, R.; Ortego, MI; Boogaart, KG, Bayes spaces: use of improper distributions and exponential families, Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales, Serie A Matemáticas, 107, 475-486, (2013) · Zbl 1280.62030

[42] Egozcue, JJ; Pawlowsky-Glahn, V.; Templ, M.; Hron, K., Independence in contingency tables using simplicial geometry, Commun Stat Theory Methods, 44, 3978-3996, (2015) · Zbl 1327.62360

[43] Egozcue, JJ; Pawlowsky-Glahn, V.; Gloor, GB, Linear association in compositional data analysis, Austrian J Stat, 47, 3-31, (2018)

[44] Erb, I.; Notredame, C., How should we measure proportionality on relative gene expression data?, Theory Biosci, 135, 21-36, (2016)

[45] Fernandes, AD; Reid, JN; Macklaim, JM; McMurrough, TA; Edgell, DR; Gloor, GB, Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis, Microbiome, 2, 15.1-15.13, (2014)

[46] Filzmoser, P.; Hron, K.; Templ, M., Discriminant analysis for compositional data and robust parameter estimation, Comput Stat, 27, 585-604, (2012) · Zbl 1304.65033

[47] Filzmoser P, Hron K, Templ M (2018) Applied compositional analysis. With worked examples in R. Springer, Switzerland AG, p 280 · Zbl 1304.65033

[48] Fisher, RA, The analysis of covariance method for the relation between a part and the whole, Biometrics, 3, 65-68, (1947)

[49] Fréchet, M., Les éléments Aléatoires de Nature Quelconque dans une Espace Distancié, Annales de l'Institut Henri Poincaré, 10, 215-308, (1948) · Zbl 0035.20802

[50] Fry, JM; Fry, TRL; McLaren, KR, Compositional data analysis and zeros in micro data, Appl Econ, 32, 953-959, (2000)

[51] Gini, C., Measurement of inequality of incomes, Econ J, 31, 124-126, (1921)

[52] Greenacre, M., Measuring subcompositional incoherence, Math Geosci, 43, 681-693, (2011)

[53] Halmos P (1974) Finite dimensional vector spaces. Springer, Berlin · Zbl 0288.15002

[54] Hijazi, RH; Jernigan, RW, Modelling compositional data using Dirichlet regression models, J Appl Probab Stat, 4, 77-91, (2009) · Zbl 1166.62053

[55] Hron, K.; Filzmoser, P.; Thompson, K., Linear regression with compositional explanatory variables, J Appl Stat, 39, 1115-1128, (2012)

[56] Hrůzová, K.; Todorov, V.; Hron, K.; Filzmoser, P., Classical and robust orthogonal regression between parts of compositional data, Statistics, 50, 1261-1275, (2016) · Zbl 1353.62072

[57] INE (2016) Renta disponible bruta de los hogares (per cápita). Serie 2010-2014. Contabilidad regional de España. Base 2010

[58] Kurtz, ZD; Müller, CL; Miraldi, ER; Littman, DR; Blaser, MJ; Bonneau, RA, Sparse and compositionally robust inference of microbial ecological networks, PLoS Comput Biol, 11, e1004226, (2015)

[59] Kynčlová, P.; Hron, K.; Filzmoser, P., Correlation between compositional parts based on symmetric balances, Math Geosci, 49, 777-796, (2017) · Zbl 1369.86020

[60] Lin, W.; Shi, P.; Feng, R.; Li, H., Variable selection in regression with compositional covariates, Biometrika, 101, 785-797, (2014) · Zbl 1306.62164

[61] Lovell, D.; Pawlowsky-Glahn, V.; Egozcue, JJ; Marguerat, S.; Bähler, J., Proportionality: a valid alternative to correlation for relative data, PLoS Comput Biol, 11, e1004075, (2015)

[62] Martín-Fernández, JA; Barceló-Vidal, C.; Pawlowsky-Glahn, V., Dealing with zeros and missing values in compositional data sets using nonparametric imputation, Math Geol, 35, 253-278, (2003) · Zbl 1302.86027

[63] Martín-Fernández, JA; Hron, K.; Templ, M.; Filzmoser, P.; Palarea-Albaladejo, J., Model-based replacement of rounded zeros in compositional data: classical and robust approaches, Comput Stat Data Anal, 56, 2688-2704, (2012) · Zbl 1255.62116

[64] Martín-Fernández, JA; Hron, K.; Templ, M.; Filzmoser, P.; Palarea-Albaladejo, J., Bayesian-multiplicative treatment of count zeros in compositional data sets, Stat Model, 15, 134-158, (2015) · Zbl 1255.62116

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

[66] Mateu-Figueras G (2003) Models de distribució sobre el símplex. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona

[67] Mateu-Figueras, G.; Pawlowsky-Glahn, V., The skew-normal distribution on the simplex, Commun Stat Theory Methods, 36, 1787-1802, (2007) · Zbl 1315.60023

[68] Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ (2011) The principle of working on coordinates. See Pawlowsky-Glahn and Buccianti (2011), pp 31-42

[69] Mateu-Figueras, G.; Pawlowsky-Glahn, V.; Egozcue, JJ, The normal distribution in some constrained sample spaces, Stat Oper Res Trans, 37, 29-56, (2013) · Zbl 1296.60002

[70] McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, London · Zbl 0744.62098

[71] Menafoglio, A.; Secchi, P.; Dalla Rosa, M., A universal kriging predictor for spatially dependent functional data of a Hilbert space, Electron J Stat, 7, 2209-2240, (2013) · Zbl 1293.62120

[72] Menafoglio, A.; Guadagnini, A.; Secchi, P., Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems

© 2021 FIZ Karlsruhe GmbH

through a bayes space approach, Water Resour Res, 52, 5708-5726, (2016)

[73] Morais, J.; Thomas-Agnan, C.; Simioni, M., Using compositional and Dirichlet models for market share regression, J Appl Stat, 45, 1670-1689, (2018)

[74] Mosimann, JE, On the compound multinomial distribution, the multivariate \(\beta\)-distribution and correlations among proportions, Biometrika, 49, 65-82, (1962) · Zbl 0105.12502

[75] Ortego MI, Egozcue JJ (2013) Spurious copulas. In: Hron PFK MT (eds) Proceedings of the 5th workshop on compositional data analysis, CoDaWork 2013, pp 123-130

[76] Palarea-Albaladejo, J.; Martín-Fernández, J., A modified EM alr-algorithm for replacing rounded zeros in compositional data sets, Comput Geosci, 34, 2233-2251, (2008)

[77] Palarea-Albaladejo, J.; Martín-Fernández, JA, zCompositions—R package for multivariate imputation of left-censored data under a compositional approach, Chemom Intell Lab Syst, 143, 85-96, (2015)

[78] Pawlowsky-Glahn V, Buccianti A (eds) (2011) Compositional data analysis: theory and applications. Wiley, New York, p 378

[79] Pawlowsky-Glahn, V.; Egozcue, JJ, Geometric approach to statistical analysis on the simplex, Stoch Environ Res Risk Assess, 15, 384-398, (2001) · Zbl 0987.62001

[80] Pawlowsky-Glahn, V.; Egozcue, JJ, BLU estimators and compositional data, Math Geol, 34, 259-274, (2002) · Zbl 1031.86007

[81] Pawlowsky-Glahn, V.; Egozcue, J., Exploring compositional data with the coda-dendrogram, Austrian J Stat, 40, 103-113, (2011)

[82] Pawlowsky-Glahn, V.; Egozcue, JJ; Lovell, D., Tools for compositional data with a total, Stat Model, 15, 175-190, (2015)

[83] Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015b) Modeling and analysis of compositional data. Statistics in practice. Wiley, Chichester, p 272

[84] Pearson, K., Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs, Proc R Soc Lond, LX, 489-502, (1897) · Zbl 28.0209.02

[85] Queysanne M (1973) Álgebra Básica. Editorial Vicens Vives, Barcelona (E), p 669

[86] Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, Paredes R, Noguera-Julian M, Calle ML (2018) Balances: a new perspective for microbiome analysis. mSystems 3(4):e00053-18. https://doi.org/10.1128/mSystems.00053-18

[87] Robert CP (1994) The Bayesian choice. A decision-theoretic motivation. Springer, New York · Zbl 0808.62005

[88] Scealy, JL; Welsh, AH, Regression for compositional data by using distributions defined on the hypersphere, J R Stat Soc Ser B Stat Methodol, 73, 351-375, (2011) · Zbl 1411.62179

[89] Shi, P.; Zhang, A.; Li, H., Regression analysis for microbiome compositional data, Ann Appl Stat, 10, 1019-1040, (2016) · Zbl 1398.62346

[90] Shorrocks, AF, The class of additively decomposable inequality measures, Econometrica, 48, 613-625, (1980) · Zbl 0435.90040

[91] Theil H (1967) On the measurement of inequality. North Holland, Amsterdam

[92] Tolosana-Delgado, R.; Eynatten, H., Grain-size control on petrographic composition of sediments: compositional regression and rounded zeros, Math Geosci, 41, 869-886, (2009) · Zbl 1178.86025

[93] Tolosana-Delgado, R.; Eynatten, H., Simplifying compositional multiple regression: application to grain size controls on sediment geochemistry, Comput Geosci, 36, 577-589, (2010)

[94] van den Boogaart KG, Tolosana-Delgado R (2013) Analysing compositional data with R. Springer, Berlin, p 258 · Zbl 1276.62011

[95] Boogaart, KG; Egozcue, JJ; Pawlowsky-Glahn, V., Bayes linear spaces, Stat Oper Res Trans, 34, 201-222, (2010) · Zbl 1208.62003

[96] Boogaart, KG; Egozcue, JJ; Pawlowsky-Glahn, V., Bayes Hilbert spaces, Aust NZ J Stat, 56, 171-194, (2014) · Zbl 1335.62025

[97] Vistelius, AB, The skew frequency distributions and the fundamental law of the geochemical processes, J Geol, 68, 1-22, (1960)

[98] Wang, H.; Shangguan, L.; Wu, J.; Guan, R., Multiple linear regression modeling for compositional data, Neurocomputing, 122, 490-500, (2013)

[99] Wikipedia (2018) Homogeneous function—Wikipedia, The Free Encyclopedia. Accessed 5 Aug 2018

This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.