GENERAL

analyzing compositional data with r use r

Published: September 19, 2026 | Index ID: #-

Introduction

In many scientific disciplines—geology, ecology, nutrition, and economics—researchers collect data that are naturally expressed as parts of a whole. Whether it is the percentage of minerals in a rock sample, the proportion of nutrients in a diet, or the market share of companies within an industry, these are examples of *compositional data*. Traditional statistical methods often fail to capture the unique structure of such data, leading to misleading conclusions. This is where the field of **compositional data analysis (CoDA)** comes in.

When it comes to performing CoDA, the programming language R has become the go-to tool for statisticians and data scientists alike. Its rich ecosystem of packages, flexible syntax, and strong community support make it ideal for **analyzing compositional data with R**. In this article, we will walk through the fundamentals of CoDA, explore key R packages, and provide a detailed, step by step tutorial that takes you from raw data to insightful visualizations and robust statistical models.

What Is Compositional Data?

Definition and Examples

Compositional data are multivariate observations where each component represents a part of a whole and the components are constrained to sum to a constant value, typically 1 or 100%. Because of this constant sum constraint, the components are inherently dependent: increasing one component necessarily decreases at least one other.

Common examples include:

- Geochemical compositions of rocks (e.g., percentages of SiO2, Al2O3, FeO, etc.)
- Proportional nutrient intake in dietary studies (e.g., protein, fat, carbohydrate percentages)
- Market share data in economics (e.g., percentage of sales for each company in an industry)
- Microbiome relative abundances (e.g., proportions of bacterial taxa in a sample)

Challenges in Analysis

Because compositional components are not independent, applying conventional multivariate techniques—such as Pearson correlation, linear regression, or PCA directly on raw proportions—can produce spurious results. For instance, a high correlation between two components might simply reflect the constant sum constraint rather than a true association. Moreover, many statistical models assume variables are measured on an unconstrained scale, which is violated by compositional data.

To overcome these issues, CoDA transforms compositional vectors into an unconstrained space where standard statistical tools can be applied safely. The most common transformation is the log ratio approach, which preserves the relative information inherent in the data.

Why Use R for Compositional Data?

R Ecosystem Advantages

R offers several key strengths that make it particularly well suited for compositional data analysis:

- Specialized Packages: Packages like compositions, robCompositions, CoDaPack, and zCompositions provide functions specifically designed for log ratio transformations, imputation, and diagnostics.
- Reproducibility: R scripts can be version controlled with Git, embedded in R Markdown documents for literate programming, and shared on platforms such as GitHub or CRAN.
- Community Support: The R community frequently publishes tutorials, vignettes, and forum discussions on CoDA topics.
- Visualization: Packages such as ggplot2, plotly, and ggtern allow for advanced graphical representations of compositional data, including ternary plots and biplots.

These features make R an attractive choice for scientists who need to analyze compositional data with rigor and transparency.

Core Concepts in Compositional Data Analysis

Log Ratio Transformations

Transforming compositional data into an unconstrained space is essential before applying standard statistical methods. Three main log ratio transformations are widely used:

The CLR transformation divides each component by the geometric mean of all components and then takes the natural logarithm:

CLR(xi) = ln(xi / g(x)) where g(x) is the geometric mean of the components.

This transformation is symmetric and preserves the relative information among all components.

ALR selects one component as a reference and computes log ratios of the other components to this reference:

ALR(xi) = ln(xi / xref) for i "` ref.

While ALR reduces dimensionality, the choice of reference component can influence the results.

ILR constructs orthonormal contrasts between groups of components using a sequential binary partition. The resulting coordinates are orthogonal and isometric, making them suitable for multivariate analyses such as PCA or discriminant analysis.

Choosing between CLR, ALR, and ILR depends on the specific analysis goals and data characteristics.

Distances and Divergences

CoDA also defines specialized distance measures that respect the compositional structure:

- Aitchison distance: The Euclidean distance in CLR space, often used as a general measure of dissimilarity.
- Kullback–Leibler divergence: Useful for comparing probability distributions.
- Euclidean distance in ILR space: Equivalent to Aitchison distance due to orthonormality.

Statistical Models

Once transformed, compositional data can be analyzed using a variety of statistical models:

- Linear regression: Predicting a response variable from log ratio predictors.
- Generalized linear models (GLMs): Extending linear models to non Gaussian responses.
- Multivariate techniques: PCA, PLS DA, and cluster analysis on CLR or ILR coordinates.
- Bayesian approaches: Hierarchical models that incorporate uncertainty in compositional proportions.

Essential R Packages for Compositional Data

compositions

The compositions package provides a comprehensive set of tools for log ratio transformations, diagnostics, and statistical modeling. Functions include clr(), ilr(), alr(), acomp(), and isocomp(). It also offers visualizations such as biplots and ternary plots.

robCompositions

Robust statistical methods are crucial when dealing with outliers or heavy tailed data. robCompositions implements robust versions of CLR, ILR, and ALR, as well as robust distance metrics and clustering algorithms.

CoDaPack

CoDaPack is an integrated suite that combines data manipulation, visualization, and analysis. It includes functions for handling zeros, performing imputation, and conducting hypothesis tests specifically tailored for compositional data.

zCompositions

Zero handling is a common issue in compositional data, especially in microbiome studies. zCompositions offers multiple strategies for replacing zeros before applying log ratio transformations, such as multiplicative replacement and Bayesian-multiplicative approaches.

Others

- ggtern – Ternary plots for three part compositions.
- CoDa – Provides functions for Aitchison geometry and distance calculations.
- tidyverse – For data wrangling and visualization pipelines.

Step by Step Guide to Analyzing Compositional Data with R

Data Preparation

Baca Juga (Artikel Terkait)

- **advanced engineering electromagnetics solution manual edition hardcover by balanis constantine a published by wiley**
  URL: http://denniswildberger.com/advanced-engineering-electromagnetics-solution-manual-edition-hardcover-by-balanis-constantine-a-published-by-wiley.pdf
- **advanced engineering electromagnetics solutions manual**
  URL: http://denniswildberger.com/advanced-engineering-electromagnetics-solutions-manual.pdf
- **advanced engineering electromagnetics wiley 1989 grading**
  URL: http://denniswildberger.com/advanced-engineering-electromagnetics-wiley-1989-grading.pdf
- **advanced engineering fluid mechanics by biswas**
  URL: http://denniswildberger.com/advanced-engineering-fluid-mechanics-by-biswas.pdf

Tags: #analyzing #compositional #data