cardinalR: Generating interesting high-dimensional data structures

by Jayani P.G. Lakshika, Dianne Cook, Paul Harrison, Michael Lydeamore, and Thiyanga S. Talagala

Abstract A high-dimensional dataset is where each observation is described by many features, or dimensions. Such a dataset might contain various types of structures that have complex geometric properties, such as nonlinear manifolds, clusters, or sparse distributions. We can generate data containing a variety of structures using mathematical functions and statistical distributions. Sampling from a multivariate normal distribution will generate data in an elliptical shape. Using a trigonometric function we can generate a spiral. A torus function can create a donut shape. High-dimensional data structures are useful for testing, validating, and improving algorithms used in dimensionality reduction, clustering, machine learning, and visualization. Their controlled complexity allows researchers to understand challenges posed in data analysis and helps to develop robust analytical methods across diverse scientific fields like bioinformatics, machine learning, and forensic science. Functions to generate a large variety of structures in high dimensions are organized into the R package `cardinalR`, along with some already generated examples.

1 Introduction

The generation of synthetic datasets with well-defined structural properties is crucial for evaluating and benchmarking algorithms across various fields, including machine learning, data mining, and spatial statistics. Researchers often require the ability to create data with specific dimensionalities, noise characteristics, and underlying cluster structures to rigorously test the performance and robustness of their methods. While existing R packages like `geozoo` and `snedata` offer valuable tools for working with geometric objects and spatial network data, respectively, a gap exists in the comprehensive and flexible generation of high-dimensional data structures with integrated geometric underpinnings and controlled noise injection.

Specifically, `geozoo` provides an extensive collection of geometric objects, primarily focused on fixed, often low-dimensional, shapes. While useful for representing and visualizing these forms, it lacks direct functionalities for generating high-dimensional datasets based on these geometries, nor does it inherently incorporate mechanisms for adding controlled noise or creating clustered structures based on geometric primitives with user-defined parameters such as position, scale, and orientation in arbitrary dimensions.

On the other hand, `snedata` focuses on spatial network datasets, which, while inherently spatial, are often represented in lower-dimensional geographic spaces. It does not offer general tools for generating high-dimensional non-networked data structures based on flexible geometric layouts or the systematic addition of noise dimensions and background noise. Furthermore, the generation of clustered data based on basic geometric shapes with fine-grained control over cluster properties is not a primary focus of this package.

Therefore, a need exists for a dedicated R package that provides a versatile framework for generating synthetic data structures in arbitrary high dimensions, starting from basic geometric primitives. Such a package would ideally offer functionalities to: (i) create data structures based on geometric shapes that can be embedded and extended to higher dimensions by adding noise dimensions; (ii) introduce controlled levels of background noise to these structures; and (iii) generate clustered data by leveraging fundamental geometric shapes with user-specified positions, scales, orientations, and sample sizes in any desired dimensionality. This would empower researchers to create more realistic and challenging synthetic datasets for evaluating algorithms in various high-dimensional scenarios and spatial contexts.

This paper introduces the `cardinalR` R package, which aims to address these limitations by providing a comprehensive suite of functions for generating customizable high-dimensional data structures based on geometric primitives, with integrated noise injection and flexible cluster generation capabilities. By offering these functionalities, `cardinalR` seeks to provide a valuable resource for researchers needing to create synthetic datasets tailored to their specific evaluation needs in the realm of high-dimensional data analysis and beyond.

2 Implementation

Installation

The package can be installed from CRAN using

```
install.packages("cardinalR")
```

and from GitHub using

```
remotes::install_github("jayanilakshika/cardinalR")
```

to install the development version.

Web site

More documentation of the package can be found at the web site https://jayanilakshika.github.io/cardinalR/.

Data sets

The `cardinalR` package comes with several data sets that load with the package. These are described in Table **??**.

Functions

Swiss Roll

To generalize the Swiss roll structure to arbitrary dimensions, we introduce a function `generate_swiss_roll(n, p)`, which constructs a high-dimensional version of the classic 3D Swiss roll while preserving its core characteristics.

The function generates `n` points in a `p`-dimensional space, where the first two dimensions (`X_1`, `X_2`) define the primary Swiss roll shape using a parametric equation:

$$X_1 = t\cos(t), \quad X_2 = t\sin(t), \quad \text{where } t \sim U(0, 3\pi)$$

The third dimension (`X_3`) introduces variation perpendicular to the roll, sampled uniformly from $[-1, 1]$. Additional dimensions (`X_4` to `X_p`) extend the data structure by applying a **sinusoidal transformation** of the parameter `t`, ensuring continuity in higher-dimensional spaces:

$$X_i = \frac{\sin(it)}{i}, \quad \text{for } i \geq 4.$$

This transformation ensures a gradual decay in variance across dimensions, mimicking real-world high-dimensional structures where later dimensions often capture subtler variations.

3 Examples

Add one or two datasets and evaluate how it will be useful.... (one with NLDR, one with clustering)

4 Discussion

5 Code

The code is available at https://github.com/JayaniLakshika/cardinalR, and source material for this paper is available at https://github.com/JayaniLakshika/paper-cardinalR.

6 Acknowledgements

This article is created using **knitr** (**?**) and **rmarkdown** (**?**) in R with the `rjtools::rjournal_article` template.

Bibliography

Jayani P.G. Lakshika
Monash University
Department of Econometrics and Business Statistics, VIC 3800 Australia
https://jayanilakshika.netlify.app/
ORCiD: 0000-0002-6265-6481
jayani.piyadigamage@monash.edu

Dianne Cook
Monash University
Department of Econometrics and Business Statistics, VIC 3800 Australia
http://www.dicook.org/
ORCiD: 0000-0002-3813-7155
dicook@monash.edu

Paul Harrison
Monash University
MGBP, BDInstitute, VIC 3800 Australia
ORCiD: 0000-0002-3980-268X
paul.harrison@monash.edu

Michael Lydeamore
Monash University
Department of Econometrics and Business Statistics, VIC 3800 Australia
ORCiD: 0000-0001-6515-827X
michael.lydeamore@monash.edu

Thiyanga S. Talagala
University of Sri Jayewardenepura
Department of Statistics, Gangodawila, Nugegoda 10100 Sri Lanka
https://thiyanga.netlify.app/
ORCiD: 0000-0002-0656-9789
ttalagala@sjp.ac.lk