mlEDA: an interactive tool for guided standardized exploratory analysis workflow of multivariate longitudinal data

Ying Jin Shanshan Zhao David Umbach Mandy Goldberg

2026-02-10

Introduction

Motivation

Measures of health are often routinely collected over time in both research and clinical settings. For exmaple, individuals are likely to go through the same set of procedures during annual physical exams, such as vitals and blood/urine tests. The health indicators collected during these procedures are closely monitored for preventive health and can also be important basis for personalized health. In laboratory environment, multiple biomarkers maybe followed over time to understand the evolution of biological pathways. In community settings, large-scale surveys and censuses are conducted regularly to collect environmental, lifestyle, and socioeconomic information as basis of community health improvement. All of the above examples involve the collection of multiple measures temporally, namely the *multivariate temporal data*. Because of the scale and complexity of such information, exploratory analysis is necessary. Investigators need effective and comprehensive representation of the data structure, so that potential problem can be identified, which further facilitates hypothesis generation and resource allocation for further research work.

For a multivariate dataset collected over time, the correlation between the collected measures, especially how the correlation changes temporally, often reveal vital information. For example, changes in correlation between Earth system variables can provide valuable information for the assessment of the climate engineering deployment scenarios(**?**). An effective exploration of the evolving structure of variables is an essential step of Exploratory Data Analysis (EDA) that will have fundamental impact on the following analysis, including hypothesis generating, model selection and scientific discoveries. Unfortunately, exploratory analysis methods and tools for multivariate temporal datasets are not well-developed, especially on the analysis of correlation. While exploration of data largely depend on graphic visualization, making such graphs demands high technical skills (e.g. programming, usage of specific software). As a result, domain experts without easy access to these techniques, are often discouraged from these tools that could significantly improve analysis efficiency. In interdisciplinary collaborations and/or public communication, these technicalities often cause communication barriers. The lack of intuitive, interpretable representation of information causes difficult in comprehension, and as a consequence discourages collaborators and audience from raising questions and feedback. Moreover, existing exploratory analysis are perceived as isolated summary tables and graphs, instead of a complete workflow. Different visualization schemes or summary statistics that can reflect different aspects of the same dataset are presented separately for the sole purpose of supporting more advanced analysis, and are rarely cross-reference. There also lacks a standardize EDA workflow, leaving the choice of methods largely up to discretion. The partiality and inconsistency does not help investigator to view the evidence integrally or objectively.

The challenge above has inspired the development of TempNet-an accessible toolkit with a interactive, standardized workflow for exploratory analysis of multivariate temporal data. It comprehensively presents various aspects of data structure, visualizes the time-dependent intercorrelation, and is easy to operate even for users without statistical or coding background.

Previous work

% exisiting tools for multivariate temporal data EDA: % pairwise

The structure of multivariate temporal datasets prohibits both mutli-dimensionality and temporal evolution. However, existing EDA methods are largely based on single variable or pairwise examination. Summary statistics (e.g., mean, standard deviation) and plots (e.g. boxplots, histograms) can summarize the distribution of variables one by one and time point by time point. The association between variables is often presented by pairwise matrix-based representation, such as correlation matrix, scatterplot matrix or heatmaps. These methods work well on small- or medium-dimensional datasets. However, the increasing dimensionality rapidly expands the information to summarize. The naive method above quickly become infeasible, because the density of information becomes too overwhelming for human eyes. In addition, these tools are also static and cannot reflect temporal changes. With the extra complexity of temporal evolution, they have to be reproduced at every time point. Investigators are required to flip back and forth across slices for examination and comparison. Such operation is not a good fit for information processing in human brain, as its time-consuming and repetitive, which increase the likelihood or error or negligence.

Current literature has proposed methods for the visualization of high-dimensional or large-scale data. Since the challenge essential stems from scale, it is natural to use dimension reduction techniques. When the goal of analysis is to explore variable relationships, the subject of analysis becomes the correlation matrix instead of the data itself. Intuitively, one can also think of the correlation matrix as a dataset where each variable (row) is a subject, and its features are its correlation with other variables (columns). In this case, the correlation matrix essential becomes a time-dependent dataset with sample size equal to number of features. This naturally leads to the consideration of dimension reduction methods. However, one should be alert that this data violates the assumptions of many dimension reduction techniques designed for regular data matrices, render these methods inapplicable. For example, the rows (variables) are unlikely to be independently distributed like subjects. Moreover, the samples used to calculate correlation are not necessary the same across time, which is a very common case for regular large-scale population-level surveys like the NHANES data(**?**). These studies consistently collect the same measure, but participants who took the survey change every time. With correlated variables and inconsistent sample, the correlation matrices, though time-dependent and even with the same set of variables, cannot be classified as "longitudinal" dataset. This makes many EDA methods based on longitudinal model inapplicable, such as Group-Based Model Trajectory (GBMT) and Latent Class Mixed Effect Model(**?**). To decompose such a dataset, we need a dimension reduction method that imposes little assumptions on the data. One such example is Multidimensional Scaling (MDS) (also known as Principal Coordinates Analysis(**?**). It represents a high-dimensional datasets on a lower-dimensional space, similar to Principal Component Analysis (PCA), but with less rigorous restrictions. MDS does not require linearity nor approximate Gaussian distribution. It is also designed to preserve the high-dimensional structure as much as possible, which PCA is not necessarily able to achieve. Considering the purpose of EDA is to examine the correlation matrix, it is ideal to preserve its structure, making MDS a better reduction for this specific case.

The R package *corrr* uses this method visualize a correlation matrices(**?**). The MDS process is embedded into the *network_plot* function. The correlation matrix is represented on a 2D space in a network plot format, where each variable is represented by a node and their coordinates is the output of MDS algorithms. The original value of correlation is also plotted as the edge between nodes, where the magnitude and sign of correlation coefficient are reflected in the thickness, transparency and color of edges. This function offers a clean, comprehensive visualization where variable correlation can be intuitively perceived by their relative positioning-the stronger the correlation is, the closer the variables are positioned to each other. However, it is a static plot and thus is subject to the same repetition procedure mentioned above when visualizing temporal evolution.

Other R packages, although not explicitly for correlation, have been developed to combine MDS reduction with network visualization, such as *igraph*(**?**). One can also implement the two steps separately-generate coordinates with MDS first and then plot variables in a network. R built-in function *cmdscale*(**?**), packages *MASS*(**?**) and *smacof*(**?**) are good tools for implementing MDS and its many variations. Then we could make network plots with packages like *igraph*(**?**) and *ggplot*(**?**). However, these packages also have limitations in