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

Ying Jin Shanshan Zhao David Umbach Mandy Goldberg

2026-02-11

Contents

1 Introduction

1.1 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(Mengis et al. 2019). 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.

1.2 Previous work

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(Centers for Disease Control and Prevention and National Center for Health Statistics 2025). 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(Lu, n.d.). 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(Gower 1966). 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(Kuhn, Jackson, and Cimentada 2022). 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*(Csardi and Nepusz 2005). 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*(R Core Team 2025), packages *MASS*(Venables and Ripley 2002) and *smacof*(Mair, Groenen, and de Leeuw 2022) are good tools for implementing MDS and its many variations. Then we could make network plots with packages like *igraph*(Csardi and Nepusz 2005) and *ggplot*(Wickham 2016). However, these packages also have limitations in representing temporal evolution-they can only represent one time point at a time. Integration over time requires additional tool to combine slices into graphic interchange format or animation. There have been a few packages developed for interactive or dynamic visualization, such as *ndtv*(Bender-deMoll 2024) and *networkDynamic*(Butts et al. 2024). These packages do not require repetition of the entire process, but the data has to be formatted carefully such that time-varying features are explicitly specified for each time point. All the packages mentioned require high technical ability, such as fluency in R programming and understanding of data structure. They are not friendly to users without much technical background or coding experience, such as domain scientists or public user.

Another shortcoming of existing EDA tools is their lack of coherence and interactivity. Each analysis methods are often implemented in isolation, producing single figure or summary statistics reflecting only a small component of the entire dataset. In practice, we often desire an integrated view of the entire dataset, especially with multiple variables across multiple time. Cross-reference across different components of data is frequently needed but not sufficiently supported by the existing exploratory analysis scheme. A more fluid, interactive workflow could release these operation burdens and thus improve efficiency of analysis. It can also overcome interdisciplinary communication barrier when accompanied with intuitive representation. Existing literature has demonstrated the preference toward interactive tools (Lakkaraju et al. 2022). Unfortunately, development in this areas has been sparse, with existing work largely restricted to non-exploratory modeling or univariate and bivariate summary (Saxena, n.d.), or generative AI applications based on black-box deep learning.

2 The TempNet App

To address the challenges in Section 1, we have created TempNet, a web-based application designed for interactive, standardized exploratory analysis workflow of multivariate temporal dataset. This application is an integration of classical EDA components and novel multidimensional visualization scheme. It also has the ability to visualize temporal evolution dynamically and intuitively to human vision. Moreover, it guides users through an comprehensive analytical workflow with interactive operations. Its usage requires minimal technical background or coding experience, thus desirable for the interdisciplinary collaboration settings. This section will demonstrate the functionalities of TempNet in details, using the the Infant Feeding and Early Development (IFED) study (Adgent et al. 2018) as a motivational example. The same dataset will also serve as a case study example in Section {#sec:case}.

2.1 Motivational dataset

The Infant Feeding and Early Development (IFED) study (Adgent et al. 2018) is an NIH-funded longitudinal observational study of infant development after birth. Although it recruited both boys and girls, different biological sexes are often analyzed separately due to the essential difference in early development process. For this paper, we use the subset of 136 girls followed from birth to 36 weeks. The participants are followed up through routine clinical visits every 2-4 weeks. In each visit, the participants went through one or more of the following procedures: 1) physical examination; 2) specimen collection; and 3) ultrasounds. Each procedures collected multiple measures. Physical examination collected basic growth information such as height, weight, head circumference, as well as manually measured reproductive development, such as anogenital and bud bead size. Blood specimen collected hormones measures including FSH, estradiol and testosterone. Ultrasound collected organ size, including thyroid, uterus, ovary and bud bead.

The IFED data is a typical example of the multivariate temporal data structure, where multiple variables are repeatedly collected over time. Although the repeated measures can be collected from the same individual, the timing of collection procedures are not aligned, and at each visit there are significant proportion of missing values. Therefore, the actual effective sample changes across time and at each visit. The physical exam is the most consistent procedure, implemented at every scheduled visit with small number of missing values. Ultrasound, on the other hand, is scheduled every other visit, with over 50% missing values at most scheduled visits. Depending on how the time variable is coded, the inconsistency can impose challenge to data exploration. For example, when the "time" information is coded as "age in days after birth", mostly time points have only a few observations, causing correlation or association measures at this specific time points unreliable. If there are only two complete pairs between two variables, the correlation will always be 1, which is clearly not an accurate representation.

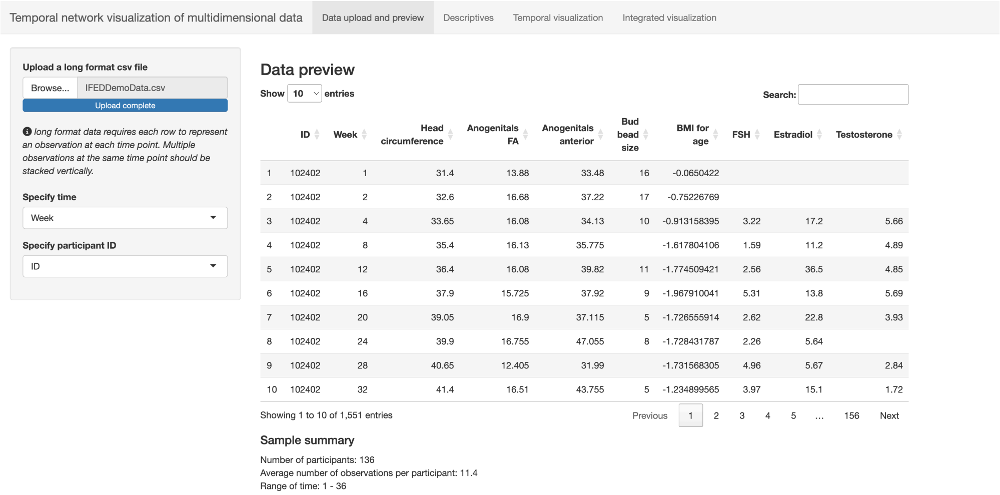
2.2 Functionalities

The mlEDA Shiny App guides user through a series of operations to explore a full dataset. The entire work flow is briefly summarized in Figure **??**. Each step corresponds to a tab in the shiny app with funcionalities focusing on certain aspects of the dataset, such as descriptive graphs for univariate distribution and bivariate correlation, multivariate structure and time evolution. All tabs in mlEDA consists of a input panel on the left and an output panel on the right. The former for user inputs, and the later for analytical output following user operation.

2.2.1 Data upload and preview

The first tab of mlEDA allows user to upload a dataset from their own local device. Currently, the dataset is required to be a csv file with long-format, meaning each row is an observation from one subject at a single time point, and repeated measures at different times should be vertically stacked. The uploaded dataset will then be displayed on the output panel, together with a brief summary of sample size and follow-up times.

It is also important that user should specify the time and subject identifier in this tab, as those are the defining factors of repeated measures over time and will affect future analysis. In some cases a dataset can contain multiple variables for time or subject ID. The IFED dataset, as an example, recorded time both by the week a procedure is scheduled, or the number of days after birth. These variables are often highly correlated as they contains the same piece of information in different forms. Therefore, we recommend user to keep only one of them in the following analysis to avoid redundancy, unless they must be included for specific purposes.



Upload a long format csv file

IFEDDemoData.csv

Upload complete

long format data requires each row to represent an observation at each time point. Multiple observations at the same time point should be stacked vertically.

Specify time: Week

Specify participant ID: ID

Data preview

Show 10 entries

Search:

	ID	Week	Head circumference	Anogenitals FA	Anogenitals anterior	Bud bead size	BMI for age	FSH	Estradiol	Testosterone
1	102402	1	31.4	13.88	33.48	16	-0.0650422			
2	102402	2	32.6	16.68	37.22	17	-0.75226769			
3	102402	4	33.65	16.08	34.13	10	-0.913158395	3.22	17.2	5.66
4	102402	8	35.4	16.13	35.775		-1.617804106	1.59	11.2	4.89
5	102402	12	36.4	16.08	39.82	11	-1.774509421	2.56	36.5	4.85
6	102402	16	37.9	15.725	37.92	9	-1.967910041	5.31	13.8	5.69
7	102402	20	39.05	16.9	37.115	5	-1.726555914	2.62	22.8	3.93
8	102402	24	39.9	16.755	47.055	8	-1.728431787	2.26	5.64	
9	102402	28	40.65	12.405	31.99		-1.731568305	4.96	5.67	2.84
10	102402	32	41.4	16.51	43.755	5	-1.234899565	3.97	15.1	1.72

Showing 1 to 10 of 1,551 entries

1 2 3 4 5 … 156

Sample summary

Number of participants: 136
Average number of observations per participant: 11.4
Range of time: 1 - 36

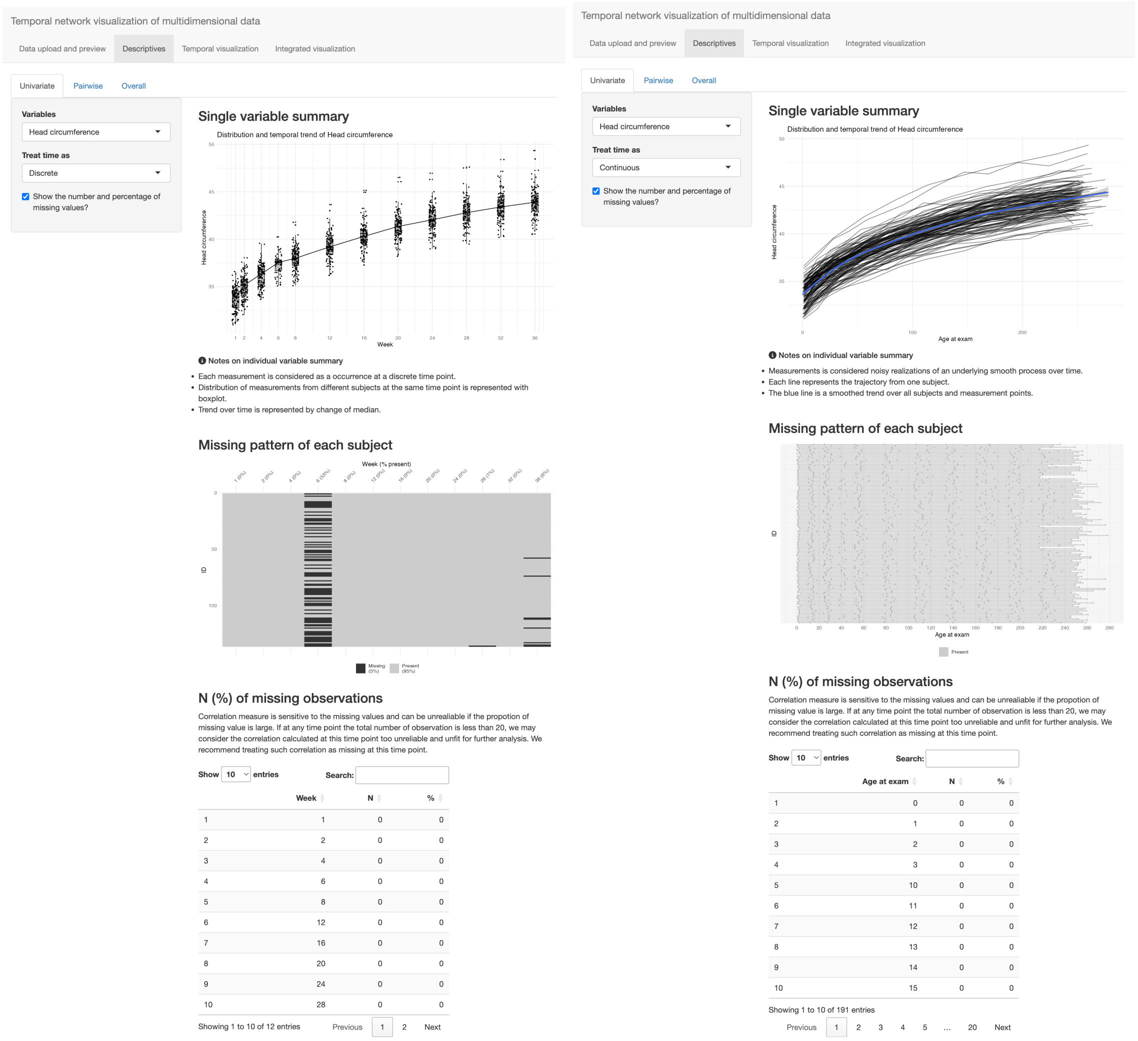
Figure 1: Temporal correlation structure

2.2.2 Descriptives

The second tab of mlEDA provides graphs and statistics for detailed exploration of subsets of the data. If offers three aspects of examination: univariate distribution, pairwise relationship and overall structure, implemented in three subtabs respectively.

2.2.2.1 Univaraiate On the first "univariate" subtab, users can choose any single variable from the dataset and examine their distribution at each time point, as will as the change in distribution over time. However, please note that the appropriate methods to visualize these features largely depend on how the "time" variable is distributed. If time grid is relatively sparse (e.g. a handful of time points), users are more likely to think of time as discrete points and focus on each point specifically. On the other hand, a dense grid (e.g. a few hundreds of time points) is more likely to be interpreted as a continuous variable, where the trajectories over time become more important than distribution at each time point. The choice of perceiving time as categorical or continuous depends on the data structure as well as analysis context, and thus is often up to user discretion. Taking the IFED data as example, "week" will be treated as discrete since there are only 12 unique points over the entire dataset. However, infant age coded as days after birth ranges from 0 to 278, with 191 unique values present in between. It should be considered continuous due to this high-density.

The mlEDA app accomodates this preference by giving user the option to specify time as discrete or continuous, and adjusts the visualization method accordingly. A variable over discrete time will be visualized as a series of boxplots, one at each measurement point (as Figure **??**) Temporal trend is represented by the change in sample median, while the discrepancy across sample is still highlighted at each time point (as Figure **??**). If the user with the perceive time time as continuous, the app generates a spaghetti plot of individual trajectories, accompanied by as summary trajectory that is smoothed over time. Additional, it provides time-dependent summary of missing pattern across both sample and time. This is often of interest because high-proportion of missing values can severely affect the reliability of correlation measure. User could chose to display a summary table of proportion of missing at each time point, where high-missing will be highlighted. User may wish to remove high-missing variables or time points from down stream analysis to improve robustness.



Variables: Head circumference

Treat time as: Discrete

☑ Show the number and percentage of missing values?

Single variable summary

Notes on individual variable summary

- Each measurement is considered as a occurrence at a discrete time point.
- Distribution of measurements from different subjects at the same time point is represented with boxplot.
- Trend over time is represented by change of median.

Missing pattern of each subject

N (%) of missing observations

Correlation measure is sensitive to the missing values and can be unrealiable if the propotion of missing value is large. If at any time point the total number of observation is less than 20, we may consider the correlation calculated at this time point too unreliable and unfit for further analysis. We recommend treating such correlation as missing at this time point.

Show 10 entries Search:

	Week	N	%
1	1	0	0
2	2	0	0
3	4	0	0
4	6	0	0
5	8	0	0
6	12	0	0
7	16	0	0
8	20	0	0
9	24	0	0
10	28	0	0

Showing 1 to 10 of 12 entries

(a) Discrete time

Variables: Head circumference

Treat time as: Continuous

☑ Show the number and percentage of missing values?

Single variable summary

Notes on individual variable summary

- Measurements is considered noisy realizations of an underlying smooth process over time.
- Each line represents the trajectory from one subject.
- The blue line is a smoothed trend over all subjects and measurement points.

Missing pattern of each subject

N (%) of missing observations

Correlation measure is sensitive to the missing values and can be unrealiable if the propotion of missing value is large. If at any time point the total number of observation is less than 20, we may consider the correlation calculated at this time point too unreliable and unfit for further analysis. We recommend treating such correlation as missing at this time point.

Show 10 entries Search:

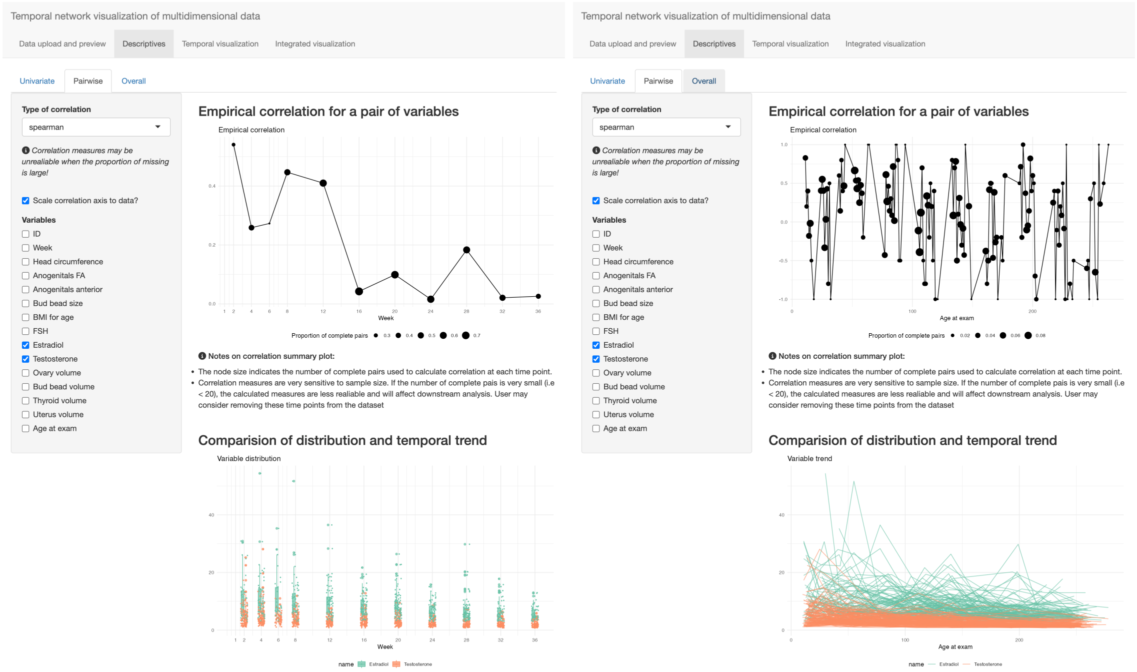
	Age at exam	N	%
1	0	0	0
2	1	0	0
3	2	0	0
4	3	0	0
5	10	0	0
6	11	0	0
7	12	0	0
8	13	0	0
9	14	0	0
10	15	0	0

Showing 1 to 10 of 191 entries

(b) Continuous time

Figure 2: Descriptives of individual variable distribution and change over time. Figure (a) uses week of scheduled procedure as time. Figure (b) uses infant age as time, coded as days after birth.

2.2.2.2 Paiwise The second subtab examines the time-varying correlation between any pair of variables. User can choose any two variable from the dataset and their preferred correlation measure (pearson or spearman). A plot of correlation trend is then generated on the output panel showing the empirical correlation value at each measurement point. However, please note that the scale of missing is also visualized on the same plot. The points are not only representing the value of correlation coefficient. Their size also indicate how many pairs of observations are used for calculating this correlation. This information is very important because it is relevant to the reliability of correlation. Large point size indicates a large sample, and thus more reliable correlation, and vice versa. In some cases, a pair of variables may have only 2 complete pairs at a single time point, and the correlation measure will always be 1 or -1, which clearly is meaningless. This in fact happened in the IFED dataset, especially when age is considered as time. The 136 participants are spread over 278 days, making sample size at each time point very small. Therefore empirical correlation frequently hits extreme values.



Temporal network visualization of multidimensional data

Data upload and preview | Descriptives | Temporal visualization | Integrated visualization

Univariate | Pairwise | Overall

Type of correlation

spearman

Correlation measures may be unrealiable when the proportion of missing is large!

☑ Scale correlation axis to data?

Variables

☐ ID
☐ Week
☐ Head circumference
☐ Anogenitals FA
☐ Anogenitals anterior
☐ Bud bead size
☐ BMI for age
☐ FSH
☑ Estradiol
☑ Testosterone
☐ Ovary volume
☐ Bud bead volume
☐ Thyroid volume
☐ Uterus volume
☐ Age at exam

Empirical correlation for a pair of variables

Notes on correlation summary plot:

- The node size indicates the number of complete pairs used to calculate correlation at each time point.
- Correlation measures are very sensitive to sample size. If the number of complete pais is very small (i.e < 20), the calculated measures are less realiable and will affect downstream analysis. User may consider removing these time points from the dataset

Comparision of distribution and temporal trend

Temporal network visualization of multidimensional data

Data upload and preview | Descriptives | Temporal visualization | Integrated visualization

Univariate | Pairwise | Overall

Type of correlation

spearman

Correlation measures may be unrealiable when the proportion of missing is large!

☑ Scale correlation axis to data?

Variables

☐ ID
☐ Week
☐ Head circumference
☐ Anogenitals FA
☐ Anogenitals anterior
☐ Bud bead size
☐ BMI for age
☐ FSH
☑ Estradiol
☑ Testosterone
☐ Ovary volume
☐ Bud bead volume
☐ Thyroid volume
☐ Uterus volume
☐ Age at exam

Empirical correlation for a pair of variables

Notes on correlation summary plot:

- The node size indicates the number of complete pairs used to calculate correlation at each time point.
- Correlation measures are very sensitive to sample size. If the number of complete pais is very small (i.e < 20), the calculated measures are less realiable and will affect downstream analysis. User may consider removing these time points from the dataset

Comparision of distribution and temporal trend

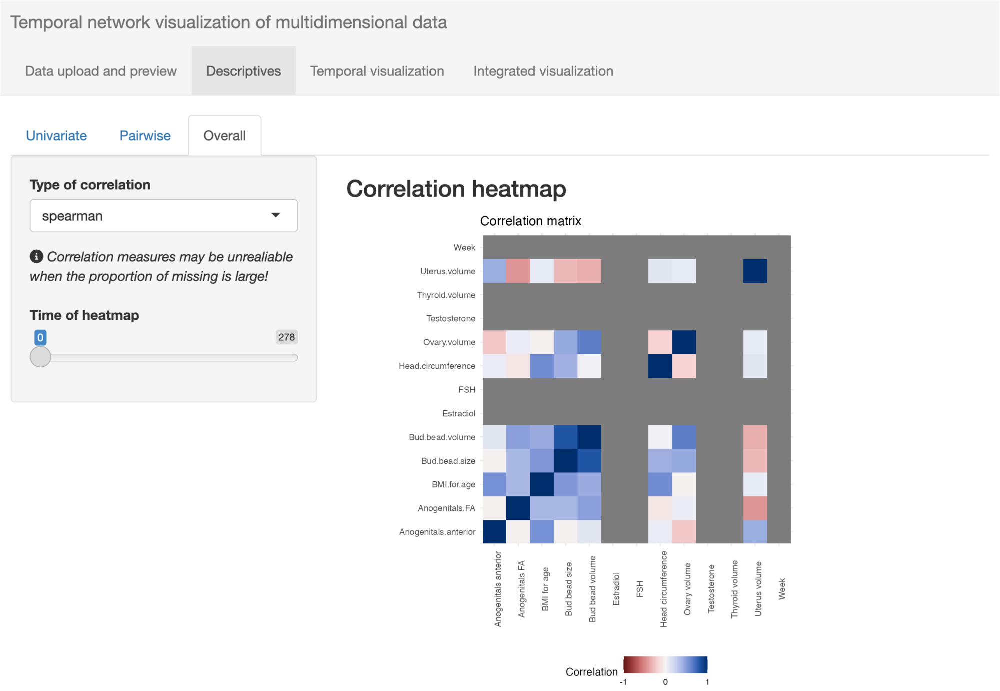
(a) Discrete time

(b) Continuous time

Figure 3: Descriptives of pairwise correlation and its change over time. Figure (a) uses week of scheduled procedure as discrete time. Figure (b) uses infant age as continuous time, coded as days after birth.

Distribution and/or trajectories of the two variables are displayed together for comparison. The visualization scheme is similar to the previous tab, and thus not displayed in the manuscript for the conciseness. User can also chose to fix the vertical axis range correlation plot between -1 and 1 if they prefer by unchecking the "scale correlation axis to data" box. If user selects less than or more than two variables by mistake, the app will display a warning message.

2.2.2.3 Overall Tab 3 display the empirical correlation matrix in a heatmap, using different colors for direction and hue for magnitude. User can scroll over the time axis on the input panel to examine its change over time. The structure is simple and straightforward, thus are not displayed in the manuscript for simplicity.



Type of correlation
spearman

Correlation measures may be unrealiable when the proportion of missing is large!

Time of heatmap
0
278

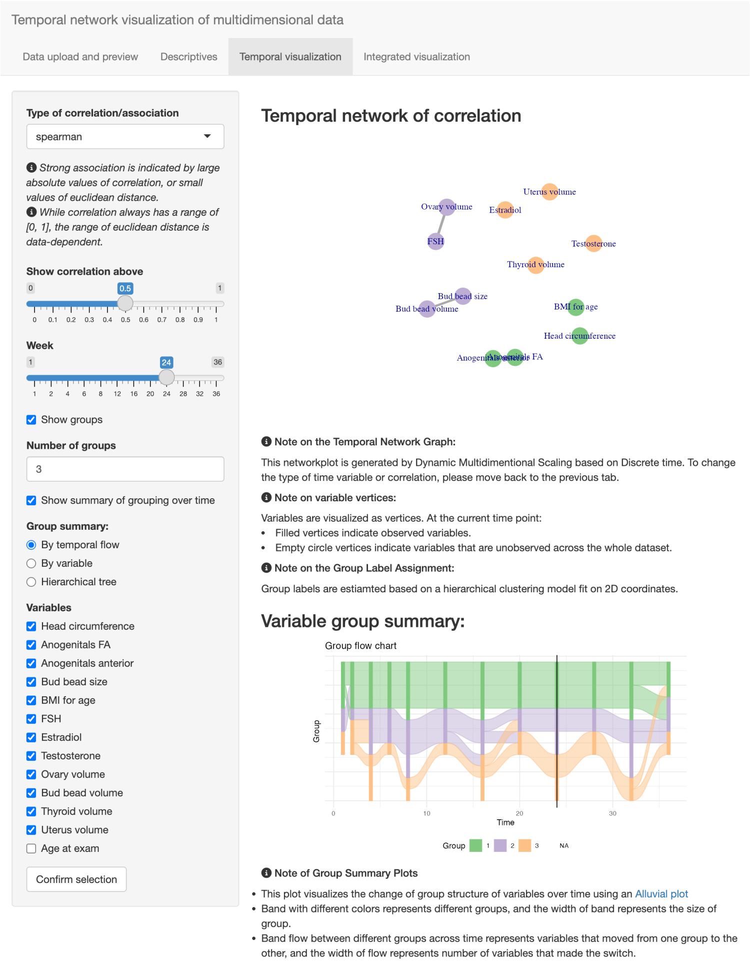
Correlation heatmap

Figure 4: Descriptives of the correlation structure of the entire dataset

2.2.3 Temporal visualization

This tab hosts the core functionality of the mlEDA app - a dynamic network graph of multivariate correlation and/or association (DCAN). We proposed this novel visualization scheme to offer a comprehensive, concise and interpretable representation of the complicated and evolving relationship between multiple measures. The layout reflects the matrix of correlation or association measures on one 2D surface, with nodes indicates variables, and their relative position indicating the strength of relationship. Strongly associated variables are positioned closer to each other while weakly associated variables are positioned further away from each other. By scroll over the time axis on the input panel, user will see the DCAN graph layout adjusts in a visually stable way, using minimal movement to adapt to the data structure at the current time point. As Figure (**fig-tab3?**) shows, the plot is generated by different two different methods, based on user with to perceive time. The following Section 3 will explain in detail the how the layout is generated and visual stability is preserved. Here we focus on the visualization scheme. For example, in Figure **??**, the two anogenital measures are very close to each other, indicating very strong correlation between them. This is intuitive considering they measure the same subject from different aspects. On the other hand, uterus volume and head circumference are very far in position, indicating weak correlation. As is shown, the plot maps the strength of correlation onto the distance between nodes on a 2D surface, thus lead to an intuitive visualization that can be efficiently interpreted.

More interactive features are incorporated in this tab to accommodate users' needs and preferences. Users can choose to display correlation over certain thresholds. These large correlations are visualized as an edge connecting the corresponding two variables, and the thickness and color hue of the edges reflects the magnitude of the correlation. In Figure 5, correlation threshold is set to 0.51, and edges appeared between FSH and ovary volume, and bud bead size and bud bead volume, meaning they are the only variable pairs whose correlation is greater than 0.5. In practice, user may with to identity groups of variables with stronger intercorrelation (i.e., "clusters" of variables). TempNet has a corresponding complementary function on this tab to group variables by performing a time-dependent hierarchical clustering on the correlation/association measures. The clustering is implmented on the lower-dimensional coordinates instead of the empirical higher dimensional structure, since the lower dimensional structure is more stable over time (see section 3 for



Type of correlation/association

spearman

Strong association is indicated by large absolute values of correlation, or small values of euclidean distance.
While correlation always has a range of [0, 1], the range of euclidean distance is data-dependent.

Show correlation above

Week

☑ Show groups

Number of groups

3

☑ Show summary of grouping over time

Group summary:

● By temporal flow
○ By variable
○ Hierarchical tree

Variables

☑ Head circumference
☑ Anogenitals FA
☑ Anogenitals anterior
☑ Bud bead size
☑ BMI for age
☑ FSH
☑ Estradiol
☑ Testosterone
☑ Ovary volume
☑ Bud bead volume
☑ Thyroid volume
☑ Uterus volume
☐ Age at exam

Temporal network of correlation

Note on the Temporal Network Graph:

This networkplot is generated by Dynamic Multidimentional Scaling based on Discrete time. To change the type of time variable or correlation, please move back to the previous tab.

Note on variable vertices:

Variables are visualized as vertices. At the current time point:

- Filled vertices indicate observed variables.
- Empty circle vertices indicate variables that are unobserved across the whole dataset.

Note on the Group Label Assignment:

Group labels are estiamted based on a hierarchical clustering model fit on 2D coordinates.

Variable group summary:

Note of Group Summary Plots

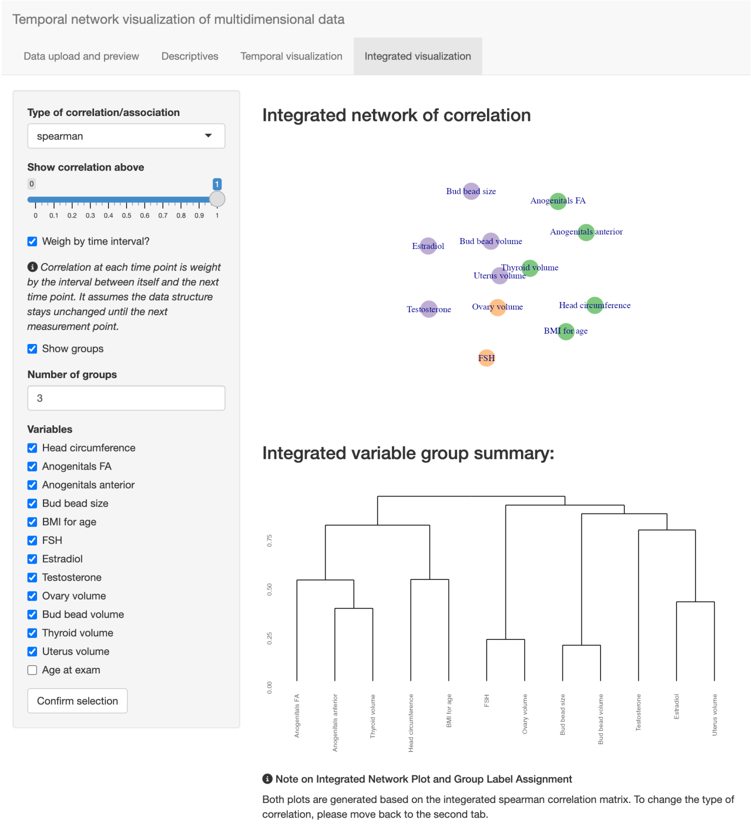
- This plot visualizes the change of group structure of variables over time using an Alluvial plot
- Band with different colors represents different groups, and the width of band represents the size of group.
- Band flow between different groups across time represents variables that moved from one group to the other, and the width of flow represents number of variables that made the switch.

Figure 5: Dynamic Correlation and Association Network plot.

details). User can choose the number of groups, and the variable vertices will be colored differently by group assignment. A few summary of group labels over time can be displayed below the network graph, where user can examine the change of grouping results over time from different perspectives. If a group of interest in identified, user can also regenerated the plot only for these subset of variables for more detailed analysis.

2.2.4 Integrated visualization

The fourth tab offers an time-independent overall summary of the entire follow up period. It averages the correlation or associatia matrix from all time points during the follow up time, and then visualize this "average" matrix using the same mechanism as DCAN in tab 3. User can either treat all time points equally, or weigh them by the time interval in between. The "weighted" option takes a "last observation carried over" approach, assuming correlation structure does not change until the next time point. It therefore weighs each time point by the interval following. This time-independent summary could be good complementary information to user inspection. As Figure 6 shows, the structure of this tab is very similar to tab 3, except for the time-varying components. They also share similar interactive functionalities, such as specifying measure threshold, grouping, and chooing subsets of data.



Type of correlation/association: spearman

Show correlation above: 0 – 1

☑ Weigh by time interval?

Correlation at each time point is weight by the interval between itself and the next time point. It assumes the data structure stays unchanged until the next measurement point.

☑ Show groups

Number of groups: 3

Variables

☑ Head circumference
☑ Anogenitals FA
☑ Anogenitals anterior
☑ Bud bead size
☑ BMI for age
☑ FSH
☑ Estradiol
☑ Testosterone
☑ Ovary volume
☑ Bud bead volume
☑ Thyroid volume
☑ Uterus volume
☐ Age at exam

Note on Integrated Network Plot and Group Label Assignment

Both plots are generated based on the integerated spearman correlation matrix. To change the type of correlation, please move back to the second tab.

Figure 6: Integrated netowrk visualization of correlation or association matrix

3 Dynamic Correlation and Association Netowrk

In this section, we explain in more details about the core functionality of mlEDA: the Dynamic Correlation and Association Network (DCAN). We explain how the plot is generated to reflect the correlation structure of the dataset, while also maintain a visual conciseness and stability.

3.1 Multidimensional Scaling

As in Section 1, exploratory analysis methods for multivariate longitudinal datasets are not well-developed. Existing literature of EDA largely focus on studying the relationship between subjects but not variables. Dimension reduction or clustering, which are arguably the two most widely used EDA methods, both set up the data with raw as subjects and columns as variables (Hastie, Tibshirani, and Friedman 2009). However, exploration of correlation is the study of relationship between *variables*, or in other words, the matrix of correlation, association or distance measures. This distinction makes a lot of existing methods inapplicable, because such matrices do not satisfy may common assumptions these methods make, such as i.i.d observations. However, multidimensional scaling (MDS) is appropriate in this case, because it is highly flexible with few assumptions on the dataset.

Essentially, MDS generates variable layout by estimating a set of low-diensional coordinates that mimics the high-dimensional matrix (i.e. correlation) as much as possible. Specifically, for a dataset $\mathbf{X}$ with P variables X_1, ..., X_P at a specific time point, the relationship between variables will be represented by a P-dimensional matrix: $\Sigma(\mathbf{X}) = \{\rho_{lp}\}$, where ρ_{lp} is the relationship between $\mathbf{X_l}$ and $\mathbf{X_p}$. There are many options for this relationship measure. The mlEDA app temporarily supports three: Pearson correlation, Spearman correlation and euclidean distance. MDS will transform this higher-dimensional into a "dissimilarity" matrix: $\mathbf{\Delta} = \{\delta_{lp}\}$, with the method of transformation dependent on the chosen measure. For correlation, $\delta_{lp} = 1 - |\rho_{lp}|$. Euclidean distance on the other hand can be a dissimilarity measure itself and needs no more transformation.

The next step is to project the higher-dimensional dissimilarity $\mathbf{\Delta}$ onto a 2D surface. A variable X_p is represented by two coordinates (c_{p1}, c_{p2}) indicating horizontal and vertical positions respectively. The distance between any pair of variables, given their coordinates, can be easily calculated: $d_{lp} = \sqrt{(c_{l1} - c_{p1})^2 + (c_{l2} - c_{p2})^2}$. The matrix of pairwise Euclidean distance is a P by P matrix: $\mathbf{D} = \{d_{lp}\}$. Since we would like the geometrical distance to reflect higher-dimensional measures, this problems comes down to minimizing the difference between $\mathbf{\Delta}$ and $\mathbf{D}$. This leads to the stress loss function**?**:

$$Stress(\mathbf{C}) = \frac{||\mathbf{\Delta} - \mathbf{D}||^2}{||\mathbf{\Delta}||^2}. \tag{1}$$

By minimizing the stress loss 1, we will be able to get the 2D coordinates that is most reflective of the correlation structure.

3.2 Temporal stability

Note that the MDS algorithm alone does not consider time changes. The classic algorithm proposed by (Gower 1966), as well as many its extensions, are able to visualize one matrix at a time. For a series of matrices at different time point, there needs to be one separate MDS procedure done at each time point. We can consider the evolution process of data structure as a "movie", then each representation generated by MDS at a specific time point would be a "screenshot". However, it is not ideal to simply regenerate a separate plot at each time point. Regeneration will induce random initialization of vertices positions. When implemented repeatedly over time, this causes variable nodes to bounce around the surface, causing high visual instability and disorientation. This goes against the purpose of mlEDA, which is to introduce intuitive visualization that is easy for human comprehension. A temporally stable extension of classic MDS is needed to solve this issue. In this section, we will introduce two different approaches for this purpose: Dynamic MDS and Splines MDS. The major difference between these two methods is how they perceive the time variable. While Dynamic MDS treats time as discrete points, Splines MDS treats time as the observations of a continuous variable.

3.2.1 Dynamic MDS

One way to achieve visual stability over time is to minimize the change of layout between adjacent time points. This leads us to the Dynamic MDS methods by (Xu, Kliger, and Hero 2012), which impose penalization on the coordinate change between all adjacent slices. The loss function in this case is penalized stress below:

$$Loss(\mathbf{C}) = \sum_t \frac{||\boldsymbol{\Delta}_t - \mathbf{D}_t||^2}{||\boldsymbol{\Delta}_t||^2} + \lambda||\mathbf{C}_t - \mathbf{C}_{t-1}||^2. \tag{2}$$

This method is very straightforward and works well on datasets with low measurement density. As an example, when we use "week" as time in the IFED dataset, there are 12 unique time points, leading to 12 sets of coordinates to estimated. However, the computation time increases exponentially with measurement grid density and quickly becomes unfeasible. Our simulation shows a dataset with 100 unique repeated measures can take up to 2.5 hours to generate a compete graph.

In addition to computational burden, this method also works poorly with the presence of missing values, or when the measurement grids are different across variables. Unfortunately, both issues are very common in practice. In the IFED datasets, the three procedures are not scheduled in conjunction. Office visits and sample collection are scheduled twice as frequent as ultrasounds. From Equation 2 we can see that only vertices that are present in both adjacent time points can be counted into the penalization. At time points when growth measures and hormones are present but ultrasounds are complete missing, the relationship between ultrasounds and other measures cannot be calculated, and these variables cannot be visualized. The corresponding vertices, as a consequence, will appear and disappear from the graph repeatedly, causing visual disorientation. During each procedure, there are also sometimes large proportion of missing values. For example, Estradiol has consistently high proportion of missing, with over 20% missing at all time points. When implemented on such datasets, Dynamic MDS often fail to generate stable visualization even with heavy penalization. This is because missing values cannot contribute to the penalization and thus cannot influence the generated layout.

3.2.2 Splines MDS

While the relationship between variables changes with time, we can also assume this evolution to be a smooth process over time. It follows that its lower-dimensional representation will also be a smooth process, represented by the smooth change over time of vertex coordinates $\mathbf{C}_t = (c_{p1}(t), c_{p2}(t))$, where $p = 1....P$. A natural choise is to model the coordinates using spline basis functions:

$$\begin{aligned} c_{p1}(t) &= \sum_{k=1}^{K} \xi_{pk}\phi_k(t) = \boldsymbol{\xi}_p\ (t) \\ c_{p2}(t) &= \sum_{k=1}^{K} \zeta_{pk}\phi_k(t) = \boldsymbol{\zeta}_p\ (t) \end{aligned} \tag{3}$$

The position of the vertex indicating variable X_p will then change smoothly with t, determined by parameters $\boldsymbol{\xi}_p$ and $\boldsymbol{\zeta}_p$.It follows that the lower dimensional representation $\mathbf{D}$ will be a collection of the smooth coordinates: $\mathbf{D}(t, \boldsymbol{\xi}, \boldsymbol{\zeta}) = \{\mathbf{c}_p(t), p = 1...P\}$. And the stress function can be written as $Stress(\mathbf{C}(t, \boldsymbol{\xi}, \boldsymbol{\zeta})) = \frac{||\boldsymbol{\Delta}_t - \mathbf{D}(t)||^2}{||\boldsymbol{\Delta}_t||^2}$.

Following these assumptions, the stability of layout can be controlled by penalizing the "smoothness" of coordinate functions, such as their second derivatives:

$$Loss(\boldsymbol{\xi}, \boldsymbol{\zeta}) = \int_t Stress(\mathbf{C}(t, \boldsymbol{\xi}, \boldsymbol{\zeta})) + \sum_{p=1}^{P} \lambda(\boldsymbol{\xi}_p + \boldsymbol{\zeta}_p)\boldsymbol{\Phi}'' \tag{4}$$

Splines MDS layout estimated by minimizing Equation 4 has a few advantages over the Dynamic MDS approach. First of all, the Spliens basis functions are shared across variables. Threrfore, the vertex coordinates computed here is determined by spline basis coefficients $\boldsymbol{\xi}_p, \boldsymbol{\zeta}_p$, which are time-independent. With irregular measurement grid, although some variables can be completely missing at some time points, their layout could still be computed in a consistent manner. The mlEDA app support this function and annotates the completely missing variables with unfilled circles instead of solid nodes. This keeps all variables in the plot and thus makes the plot more visually stable, revealing smooth, gradual structural evolution over time.

In addition, with the measurement grid is dense, the loss function 4 is a lot more efficient to minimize, because it estimates a lot less parameters and also can be parallzied oever time. In the IFED dataset with age as time, the different of computation time between Dynamic and Splines MDS is a few hours versus a few minutes. When the density of data makes Dynamic MDS infeasible, Splines MDS will be the only choice for visualization.

4 Case Study

In this case study, we apply the mlEDA app to our motivational dataset and demonstrate its use and efficiency in a practical setting. Recall Section **??** has discussed the different ways of perceiving time-discrete versus continuous. The appropriate choice depends on the data structure and analytically context, and is often up to user discretion. In the mlEDA app, this choice is made on the first descriptives tab (Tab 2.1) and will affect downstream analysis, especially the method used to generate the Dynamic Correlation and Association graph. Therefore, we use two different time variables in IFED as time to demonstrate the two different workflows. Visit scheduled by weeks after birth will be the discrete time variables, with 12 unique values range from 1 to 36 with uneven distribution. The age of infants, recorded as days after birth, will be the continuous time variable, with 191 unique values also unevenly spaced from 0 to 278. In fact, visit week can be considered as binned age in days.

We use a subset of all the measures collected and removed other measures due to duplication or high-colinearity. The dataset of interest includes basic growth measures, such as age-adjusted BMI, head circumference and anogentital measures, hormones and volume measures of reproductive organs derived from ultrasound. The investigators are particularly interested in variables relationship regarding hormones and hoping to make inferences of the underlying biological process by examining the change of correlation between hormones and other meausres. Through out this section, Spearman correlation will be used for the relationship measure between variables. It is nonparametric with straightforward interpretation and also popularly used for practice.

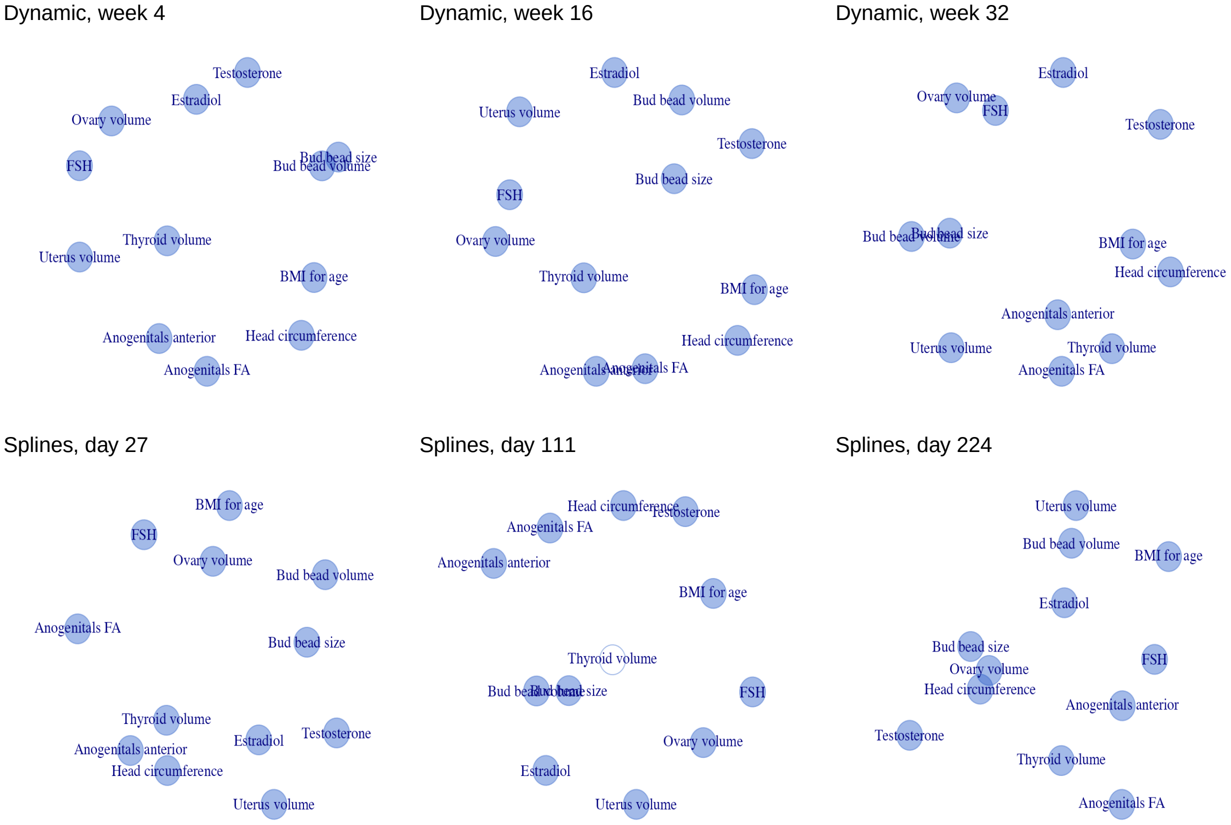
Figure **??** shows a few selected slices during the entire following up period, The animation of the full dynamic graphs over all time points can be found in the supplemetary materials. The first row are network graphs generated by Dynamic MDS with week as discrete time. Weeks 4, 16 and 32 are chosen because they have relatively complete data and also represents the start, middle and end of the study. The second row are Splines MDS graphs with age in days as time. The 27th, 111st and 224th days are the sample median age corresponding to week 4, 16 and 32, therefore chosen for display to offer a comparision between two methods.

References

Adgent, Margaret A, David M Umbach, Babette S Zemel, Andrea Kelly, Joan I Schall, Eileen G Ford, Kerry James, et al. 2018. "A Longitudinal Study of Estrogen-Responsive Tissues and Hormone Concentrations in Infants Fed Soy Formula." *The Journal of Clinical Endocrinology & Metabolism* 103 (5): 1899–909. https://doi.org/10.1210/jc.2017-02249.

Bender-deMoll, Skye. 2024. *Ndtv: Network Dynamic Temporal Visualizations.* https://doi.org/10.32614/CRAN.package.ndtv.

Butts, Carter T., Ayn Leslie-Cook, Pavel N. Krivitsky, and Skye Bender-deMoll. 2024. *networkDynamic: Dynamic Extensions for Network Objects.* https://doi.org/10.32614/CRAN.package.networkDynamic.


Dynamic, week 4

Dynamic, week 16

Dynamic, week 32

Splines, day 27

Splines, day 111

Splines, day 224

Figure 7: (#fig:ifed_dcan)Selected correlation network graph of the IFED data.

Centers for Disease Control and Prevention, and National Center for Health Statistics. 2025. “National Health and Nutrition Examination Survey Data.” Hyattsville, MD: https://www.cdc.gov/nchs/nhanes/index.htm; U.S. Department of Health; Human Services.

Csardi, Gabor, and Tamas Nepusz. 2005. “The Igraph Software Package for Complex Network Research.” *InterJournal* Complex Systems (November): 1695.

Gower, J. C. 1966. “Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis.” *Biometrika* 53 (3/4): 325–38. http://www.jstor.org/stable/2333639.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning.* Springer New York, NY. https://doi.org/https://doi.org/10.1007/978-0-387-84858-7.

Kuhn, Max, Simon Jackson, and Jorge Cimentada. 2022. *Corrr: Correlations in r.* https://doi.org/10.32614/CRAN.package.corrr.

Lakkaraju, Himabindu, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. “Rethinking Explainability as a Dialogue: A Practitioner’s Perspective.” *NeurIPS Workshop on Human Centered AI.* https://par.nsf.gov/biblio/10462932.

Lu, Zihang. n.d. “Clustering Longitudinal Data: A Review of Methods and Software Packages.” *International Statistical Review* n/a (n/a). https://doi.org/https://doi.org/10.1111/insr.12588.

Mair, Patrick, Patrick Groenen, and Jan de Leeuw. 2022. “Multidimensional Scaling Using Majorization: SMACOF in r.” *Journal of Statistical Software* 102. https://doi.org/10.18637/jss.v102.i10.

Mengis, Nadine, Dennis P. Keller, Wilfried Rickels, and et al. 2019. “Climate Engineering–Induced Changes in Correlations Between Earth System Variables–Implications for Appropriate Indicator Selection.” *Climatic Change* 153: 305–22. https://doi.org/10.1007/s10584-019-02389-7.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Saxena, Pulkit. n.d. “Advanced Biostatistics and Epidemiology Toolkit for Clinical Monitoring and Research.” https://epistatkit.vercel.app/.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s.* Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Xu, Kevin S., Mark Kliger, and Alfred O. Hero. 2012. “A Regularized Graph Layout Framework for Dynamic Network Visualization.” *Data Mining and Knowledge Discovery* 27 (1): 84–116. https://doi.org/10.1007/s10618-012-0286-6.