

The DynamicNet Package

Version 0.1.0

Yang Zhenyu

The Center for Computational Biolog at Beijing Forestry University, China

Section 1: Introduction
Section 2: Installation
Section 3: Pattern recognition of sequential representation data
Section 4: Network sparsity based on regularization
Section 5: Dynamic Network construction
Section 6: Results
Section 7: Function List
Section 8: References

1. Introduction

This R package, DynamicNet, is based on the construction method of high-dimensional temporal data dynamic network in the paper Dynamic Network Reconstruction and R Package Development. It is used to reconstruct and parse the dynamic characteristic interactive network of biological data. The package characterizes interactions with time-dependent variables by analyzing the direct effects of each target variable on its own and by adjusting other variables. In this software package, the expression pattern recognition function Fclusty divides all variables into different modules based on spatiotemporal similarity. The selection of variables based on regularization realizes the screening of important features of target variables. By constructing and solving the differential equation system based on dynamic evolutionary game theory, A sparse, directed and weighted dynamic interaction network that changes over time between variables can be constructed. The dynamic interactive network inferred by DynamicNet provides a valuable tool for mining the organizational structure hidden behind big data and exploring large biological networks.

We refer to Kim et al. (2010), Wu et al. (2021), Zeng et al. (2017) and Yang et al. (2022) for the theoretical foundation of this package. If you benefited from this software, then please cite the following paper in your work:

1. Kim B R, McMurry T, Zhao W, Wu R, Berg A. Wavelet-based functional clustering for patterns of high-dimensional dynamic gene expression[J]. J Comput Biol, 2010, 17(8): 1067-1080.
2. Wu R, Jiang L. Recovering dynamic networks in big static datasets[J]. Physics Reports, 2021, 912: 1-57.
3. Zeng Y, Yang T, Breheny P. Hybrid safe-strong rules for efficient optimization in lasso-type problems[J]. 2017.
4. Yang Z, Zhu X, Ren X, Cao Y, Ye M, Wu R. Genome-wide epistatic interaction network construction for Plant height traits in Arabidopsis thaliana. Seedlings. Journal of Beijing Forestry University.

2. Installation

The DynamicNet package hides most of the dependent functions, provides the user with functions for cluster dimension reduction, feature selection, network construction, and visualization, and only considers

limited input data validation to minimize the performance overhead. When using the R package, it is the user's responsibility to ensure that the input data is in the desired format and does not contain invalid characters.

The `hfunmap` package depends on the ***ggplot2,igraph*** package (available on CRAN), so these packages should be installed firstly.

2.1 Local Offline Installation

To install DynamicNet, download the package file and type the appropriate command below or click the menu item "Install packages from local zip/tar.gz files".

*Windows/Linux OS:

```
>install.packages("C:/yourpath/DynamicNet.zip", repos=NULL)
```

2.2 Online installation

Enter the following command in R to install the software online:

*Windows/Linux OS:

```
> require("devtools")
> devtools::install_github("https://github.com/ZhenyuYang1996/DynamicNet")
```

3. Pattern recognition of sequential representation data

The function that implements cluster dimension reduction is `Fclust()`. The input data of this function are time series samples, and all samples are divided into different modules according to the similarity of expression patterns among samples.

3.1 Statistical Model

Gaussian mixture model is a finite mixture model, its parameter learning and modeling has become an important direction of statistical learning, and has been widely used in various fields. In the multivariate finite Gaussian mixture model, the data to be clustered follows the mixed probability distribution from multiple multivariate normal distributions, which represent different classes and their related parameters (mean vector and covariance matrix) can be used as class markers. The general form of this model is as follows:

$$p(x|\theta) = \sum_{i=1}^K \pi_i p_i(x|\theta_i)$$

On the basis of multivariate Gaussian mixture clustering, this study embedded the statistical principle of time series analysis into maximum likelihood model for clustering similar expression patterns. In order to better describe the actual biological significance, this study uses legendard orthogonal polynomial family to approximate the specific mean vector under the k class(Kim et al. 2010), and its recursive formula is:

$$(r+1)P_{r+1}(t) = (2r+1)tP_r(t) - rP_{r-1}(t)$$

The covariance matrix approximated by AR1 can be expressed as:

$$\Sigma = \sigma^2 R = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{T-1} \\ \rho & 1 & \cdots & \rho^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \cdots & 1 \end{pmatrix}_{T \times T}$$

3.2 Example

Load the R package:

```
> library("DynamicNet")
```

Read the data:

```
> data(dynamicnet_data)
> X <- dynamicnet_data$data
> times <- dynamicnet_data$times
```

Cluster time series data:

```
> Fclusty(data = X,
+         times = times,
+         startK = 10, endK = 20, r = 6,
+         outfilename = "cluster",
+         verbose = TRUE)
```

4. Network sparsity based on regularization

4.1 Statistical Model

The R package uses regularized variable selection based on an elastic net regression model (Zou and Hastie, 2005) to filter a subset of the most significant variables associated with the focus variables in order to reduce complexity for subsequent calculations. As a method of feature selection, elastic net regression model is defined as the following optimization problem:

$$\begin{aligned} & \arg \min_{\beta \in R^p} \left\{ \frac{1}{2n} \|y - \mathbf{X}\beta\|^2 + \alpha\lambda \|\beta\|_1 + \frac{(1-\alpha)\lambda}{2} \|\beta\|^2 \right\} \\ & = \arg \min_{\beta \in R^p} \left\{ \frac{1}{2n} \sum_i \left(\sum_j x_{ij} \beta_j - y_i \right)^2 + \alpha\lambda \sum_i |\beta_i| + \frac{(1-\alpha)\lambda}{2} \sum_i \beta_i^2 \right\} \end{aligned}$$

At present, many methods have been developed to solve such problems. Among them, Hybrid safe-strong rules (HSSR) is a heuristic rule for discarding features. This rule can not only accurately screen out a large number of features with a zero coefficient, but also compared with SSR rules, HSSR hybrid feature screening rules have higher computational efficiency (Zeng et al. 2017).

4.2 Example

Feature filtering is performed on modules as variables:

```
> model_connect <- Dnelastic(X = clust_20_means_data,
+                             eps = 1e-07,
+                             outfile = /data/elastic_data,
+                             verbose = FALSE)
```

5. Dynamic Network construction

In DynamicNet, the function to build a dynamic network is Dynetwork().

5.1 Statistical Model

In general, ordinary differential equation models using the continuous time dimension can be used to simulate how individuals in a biological system react over time, both through their own strategies and in combination with the strategies of other individuals (Wu and Jiang, 2021). Specifically, the behavior of a variable (observed value) can be expressed as the sum of components generated by its own strategy and components generated by its interacting object (competition or cooperation). A bidirectional, signed and weighted network system consisting of M variables can be represented by a nonparametric sparse additive ordinary differential equation with time derivatives:

$$\frac{dG_i}{dt} = \sum_{t=1}^T \left(F_i(G_i(t)) + \sum_{j \in L_{it}} F_{j \rightarrow i}(G_j(t)) \right) + \varepsilon_i, i = 1, 2, \dots, m$$

For ordinary differential equations which have no analytical solution or are difficult to be solved by analytical solution, approximate analytical solution or numerical solution can be used to approximate them. Among them, the basic idea of numerical solution is to transform differential equations into difference equations by some discretization means. The commonly used numerical methods mainly include Euler method, improved Euler method, Runge-Kutta method and linear multi-step method. For a complex differential equation, The Runge-Kutta method uses Taylor expansion to construct a high-precision numerical solution, which has the advantage of high computational accuracy. The general form of the Runge-Kutta method is:

$$\begin{cases} y_{n+1} = y_n + h \sum_{i=1}^r c_i K_i \\ K_1 = f(x_n, y_n) \\ K_i = f(x_n + \lambda_i h, y_n + h \sum_{j=1}^{i-1} \mu_{ij} K_j) \\ (i = 2, \dots, r) \end{cases}$$

5.2 Example

Build networks between modules:

```
> model_net <- dynetwork(connect_mat = model_connect,
+                          Y = X,
+                          times = times,
```

```
+ order = 6,  
+ outfilename = "data/model",  
+ verbose = FALSE)
```

6. Results

6.1 Simulation results of Fclusty()

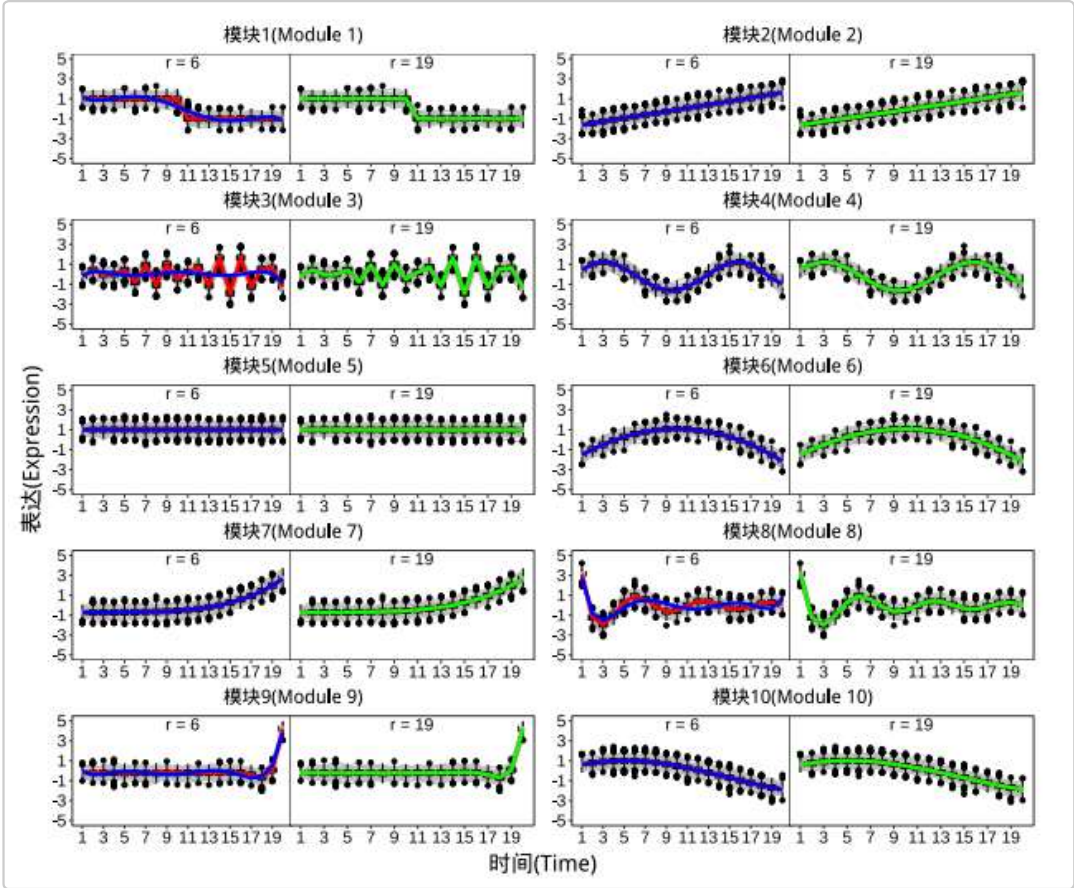


Fig.1 Expression pattern recognition of test data at different Legendre orders.

Note: The red line is the true mean value; The gray line is the expression curve divided into the same category. The blue and green lines are mean curves smoothed using order 6 and 19 Legendre polynomials.

6.2 Simulation results of Dynamic Network Construction

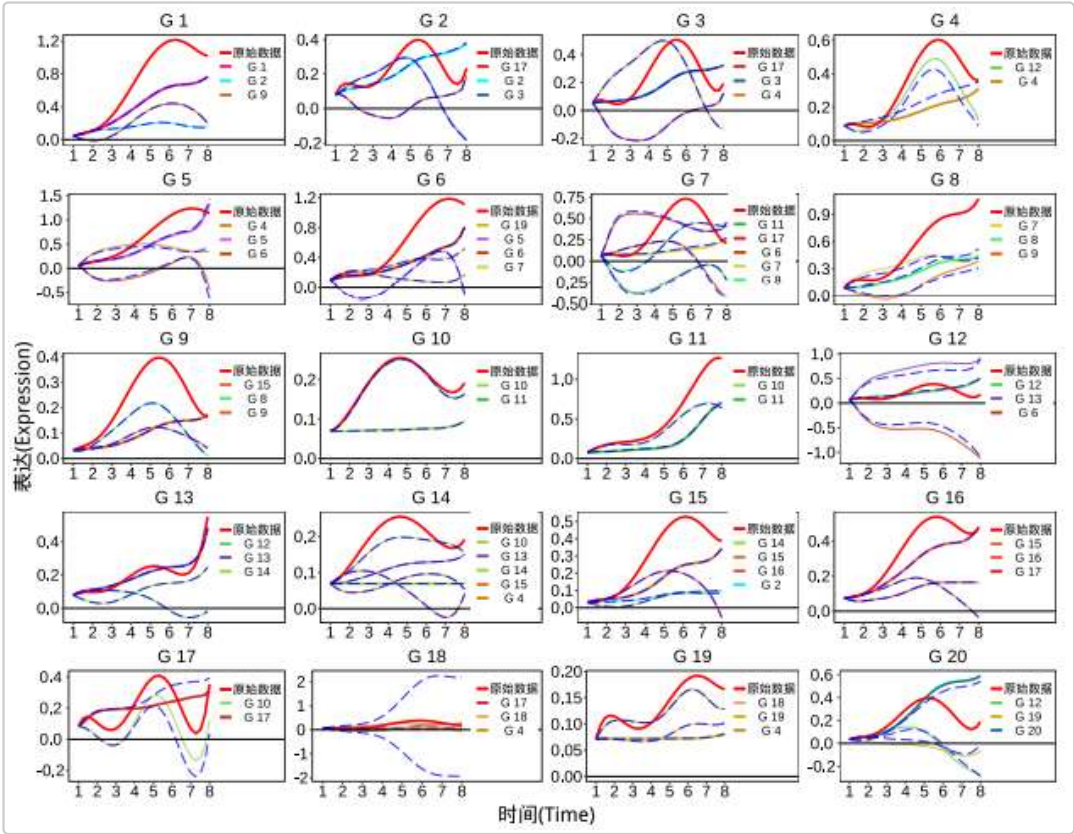


Fig.2 The comparison between the effect curve estimated in simulation experiment and the actual effect curve.

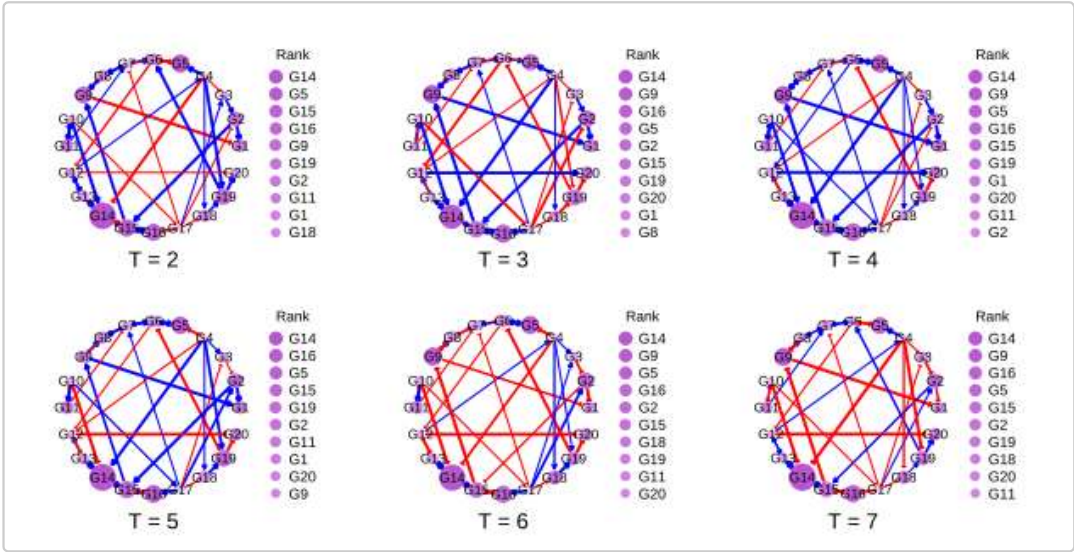


Fig.3 Network interaction and node importance at different time points.

7. Function list

NO		escription
1	Fclusty(data, times, startK, endK, r, outfilename, verbose = FALSE)	
	Pattern recognition of sequential data expression based on Orthogonal polynomial family.	
2	Dnelastic(X, eps, core_n, outfilename, verbose = FALSE)	
	Filter the subset of the most significant variables associated with the focus variable.	

NO escription	
3	Dynetwork(X, connect, times, order, outfile, verbose = FALSE)
	Dynamic network construction and topology analysis across
	time scales.

8. References

1.Kim B R, McMurry T, Zhao W, Wu R, Berg A. Wavelet-based functional clustering for patterns of high-dimensional dynamic gene expression[J]. J Comput Biol, 2010, 17(8): 1067-1080.

2.Wu R, Jiang L. Recovering dynamic networks in big static datasets[J]. Physics Reports, 2021, 912: 1-57.

3.Zeng Y, Yang T, Breheny P. Hybrid safe-strong rules for efficient optimization in lasso-type problems[J]. 2017.

4.Yang Z, Zhu X, Ren X, Cao Y, Ye M, Wu R. Genome-wide epistatic interaction network construction for Plant height traits in Arabidopsis thaliana. Seedlings. Journal of Beijing Forestry University.

5.R package: multi-dimensional functional mapping. Beijing Forestry University, 2022