Temporal Analysis of Cross Section Data

Han Chen

Tanwei Colledge, Tsinghua University

April 8, 2023

Outline

1. Introduction

2. Pseudo time series construction

3. States identification

4. Discussion

Motivation

- Clinical trials are typically conducted over a population within a defined time period.
- The construction of pseudo time series for clinical data allows for an improved understanding of the nature of disease, therefore we can make more reliable predictions.

Overview

1. **Temporal bootstrap:** resampling data from a crosssectional study.
2. **Pseudo time series construction:** each trajectory begins at a randomly selected datum from a healthy individual and ends at a random datum classified as diseased.
3. **States identification:** unlabelling the healthy/disease states in order to cluster the data into increasingly fine-grain regions using the Expectation Maximisation (EM) algorithm.

Outline

1. Introduction

2. Pseudo time series construction

3. States identification

4. Discussion

Temporal bootstrap

- $\boldsymbol{D} \in \mathbb{R}^{m \times n}$ is a real-valued matrix where m is the number of samples and n the number of variabels.
- $\boldsymbol{c} = (c_1, \dots, c_m)^\top$ represents the defined class.
- $\boldsymbol{P} = (\boldsymbol{p}_1, \dots, \boldsymbol{p}_k)$ is a set of pseudo time indices, where each $\boldsymbol{p}_i$ has length T and each $p_{ij} \in \{1, \dots, m\}$.
- $\boldsymbol{F}(\boldsymbol{p}_i) \in \mathbb{R}^{T \times n}$ where each row of $\boldsymbol{F}(\boldsymbol{p}_i) = \boldsymbol{D}(p_{ij})$.
- The corresponding class vector of $\boldsymbol{F}(\boldsymbol{p}_i)$ is given by $\boldsymbol{G}(\boldsymbol{p}_i) = (c(p_{i1}), \dots, c(p_{iT}))^\top$.

- Let $\boldsymbol{D} = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \\ d_{41} & d_{42} & d_{43} \end{bmatrix}$.
- If $\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2)$, where $\boldsymbol{p}_1 = (1, 3, 1)^\top$ and $\boldsymbol{p}_2 = (2, 3, 1)^\top$, then

$$\boldsymbol{F}(\boldsymbol{p}_1) = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{31} & d_{32} & d_{33} \\ d_{11} & d_{12} & d_{13} \end{bmatrix}, \quad \boldsymbol{G}(\boldsymbol{p}_1) = (c_1, c_3, c_1)^\top.$$

Algorithm 1: Pseudo time series construction

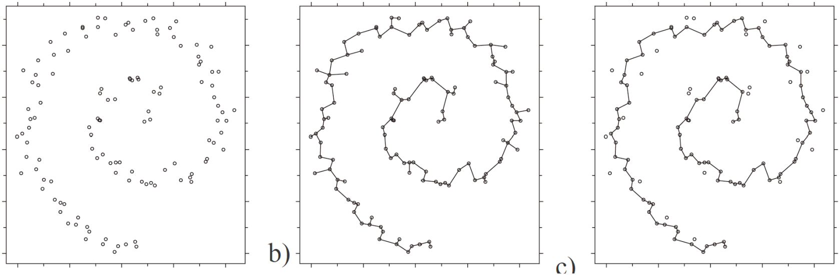
Data: Cross section data $\boldsymbol{D}$, label $\boldsymbol{c}$, sample size T, number of series k

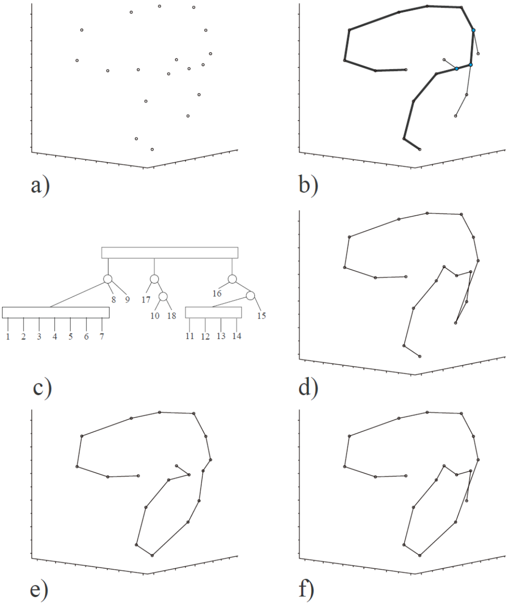
Result: Pseudo time series model

```
standardization;
for $i=1 \to k$ do
    Uniformly randomly sample $T$ row indices from $\boldsymbol{D}$ to create $d_i$ such that there is at least one healthy and one diseased class (in $\boldsymbol{c}$) corresponding to any of the indices in $d_i$;
    Uniformly randomly select a row index from $d_i$, start, from where $1 \leq \text{start} \leq T$ and an endpoint, end, where $1 \leq \text{end} \leq T$ where $c(d_{i,start})$ represents a healthy class and $c(d_{i,end})$ represents a diseased class;
    Calculate distance;
    Order $d_i$ to create $d_i^*$ based upon the shortest path between $\boldsymbol{D}(d_{i,start})$ and $\boldsymbol{D}(d_{i,end})$ given the weighted graph $G_i$ using the Floyd-Warshall algorithm;
end
```

Ordering observations

1. Find the minimum spanning tree, $G_{mst} = (V, E_{mst})$ of the weighted graph G. If G is a path, then we are done.
2. If G_{mst} is not a path, assess the *diameter path noise ratio*, *branch distribution*, and *sampling intensity*.
   - If the sampling appears to be relatively intense and the diameter path branch distribution appears to be relatively uniform, then assigned the same ordering index as the diameter path element to which they connect.
   - Otherwise two additional steps are taken. First a data structure called PQ-tree is used to summarize all the uncertainties of path variations. Next, a secondary criterion of *shortest path ordering* (motivated by the TSP algorithm) is applied to the variations of the paths.


b)
c)


a)
b)
8
9
17
16
10
18
15
1
2
3
4
5
6
7
11
12
13
14
c)
d)
e)
f)

Outline

1. Introduction

2. Pseudo time series construction

3. States identification

4. Discussion

Hidden Markov Model (HMM)

- $\boldsymbol{A} \in \mathbb{R}^{h \times h}$ is a time-independent stochastic transition matrix where h is $\#$ of states.
- $\boldsymbol{B} = \{b_j(\cdot)\}$ is the complete collection of parameters for all observation distributions.
- Assume that $b_j(\cdot)$ follows a mixture of M multivariate Guassians.
- $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_h)$ is the initial state distribution.
- $\boldsymbol{X}$ is the collection of observations, $\boldsymbol{Z}$ is the collection of hidden states.
- $\gamma(\boldsymbol{z}_i)$ is the conditional distribution of $\boldsymbol{z}_i$, $\xi(\boldsymbol{z}_{i-1}, \boldsymbol{z}_i)$ is the joint conditional distribution of $\boldsymbol{z}_{i-1}$ and $\boldsymbol{z}_i$.

- E step.

$$Q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \tilde{\boldsymbol{\theta}}) \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}).$$

- M step.

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{h} \gamma(z_{1j})}, \quad A_{jk} = \frac{\sum_{n=2}^{m} \xi(z_{n-1}, j, z_{nk})}{\sum_{l=1}^{h} \sum_{n=2}^{m} \xi(z_{n-1,j} z_{nl})},$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{m} \gamma(z_{nk}) \boldsymbol{x}_n}{\sum_{n=1}^{m} \gamma(z_{nk})}, \quad \boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{m} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_{n=1}^{m} \gamma(z_{nk})}.$$

- E step.

$$\gamma(\boldsymbol{z}_n) = \frac{\alpha(\boldsymbol{z}_n)\beta(\boldsymbol{z}_n)}{p(\boldsymbol{X})},$$

$$\xi(\boldsymbol{z}_{n-1}, \boldsymbol{z}_n) = \frac{\alpha(\boldsymbol{z}_{n-1}) p(\boldsymbol{x}_n|\boldsymbol{z}_n) p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}) \beta(\boldsymbol{z}_n)}{p(\boldsymbol{X})}.$$

```
Algorithm 2: States identification
Data: Pseudo time series matrix $\boldsymbol{P}$, cross-section data $\boldsymbol{D}$, label $\boldsymbol{c}$.
Result: HMM with new intermediate or end states.
Remove classification labels $\boldsymbol{c}$;
Set $h = \#$of classes $+ 1$;
while no clear end state do
    Train a HMM on the $\boldsymbol{P}$ with $h$ hidden states using the EM algorithm;
    h=h+1;
end
```

- Definition of *clear end state*?

Outline

1. Introduction

2. Pseudo time series construction

3. States identification

4. Discussion

Discussion

- Classical algorithms for MST: Kruscal & Prim. The choice of end points cannot be decided in advance, though.
- The authors didn't explain the procedure of HMM in their work in detail.
- There may be time heterogeneity between $\boldsymbol{p}_i$'s, i.e. the time (phase) of p_{iT} may be far away from that of p_{jT} if $i \neq j$.
- Choice of distance: Euclide v.s. cosine.

References I

Modelling and analysing the dynamics of disease progression from cross-sectional studies.
Yuanxi Li, Stephen Swift and Allan Tucker.
Journal of Biomedical Informatics 46(2), 2013.

Reconstructing the temporal ordering of biological samples using microarray data.
Paul M. Magwene, Paul Lizardi and Junhyong Kim.
Bioinformatics 19, 2003.

A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models.
Jeff A. Bilmes.
International Computer Science Institute, 1998.

Updating Markov models to integrate cross-sectional andlongitudinal studies
Allan Tuckera, Yuanxi Lia and David Garway-Heath.
Artificial Intelligence in Medicine 81, 2017.

References II

A mutagenetic tree hidden Markov model for longitudinal clonal HIV sequence data.
Niko Beerenwinkel, Mathisas Drton.
Biostatistics 8, 2007.