Inferring Time Varying Processes from Cross-Sectional Data

Thomas Nealon[1]; Supervisors: David Weston[2], Dr. Allan Tucker[3]
[1]Birkbeck, University of London, [2]Birkbeck, University of London, [3]Brunel University London

The aim of this research is to allow cross sectional data to by analysed as longitudinal data.

Introduction

The goal of this research is to develop methods to perform longitudinal or panel data analysis on cross sectional data. Longitudinal data is particularly important as biological systems, such as disease development, are predominantly developmental and dynamic. However, some observations in biomedical research are often made only at crucial stages, such as during initial diagnosis. By developing methods to give cross sectional data characteristics that enable it to be analysed as longitudinal data.

Trajectories represent longitudinal sequences.

Pseudotemporal Bootstrap

The Pseudotemporal Bootstrap (PTB) method is described in `The Pseudotemporal Bootstrap for Predicting Glaucoma From Cross-Sectional Visual Field Data'[1]. It is an algorithm for generating a Pseudotemporal trajectory through a data set that is then used to generate a predictive Hidden Markov Model[2] (HMM) from trajectories between two classes within the data set being examined. The trajectories are 'Pseudotemporal' because they have a start, end and an order but do not assign a specific time at which each point occurs.

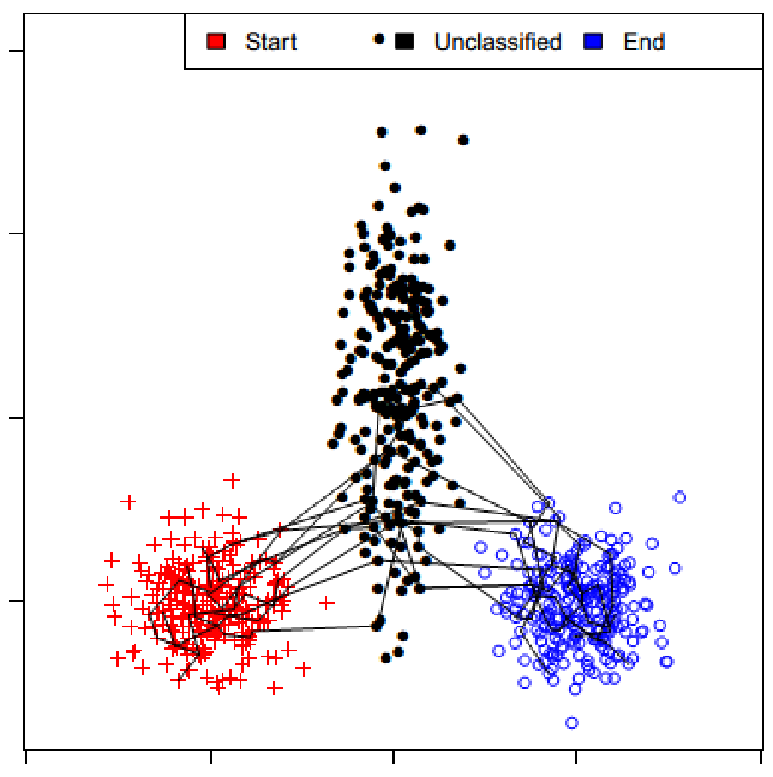

Figure 1

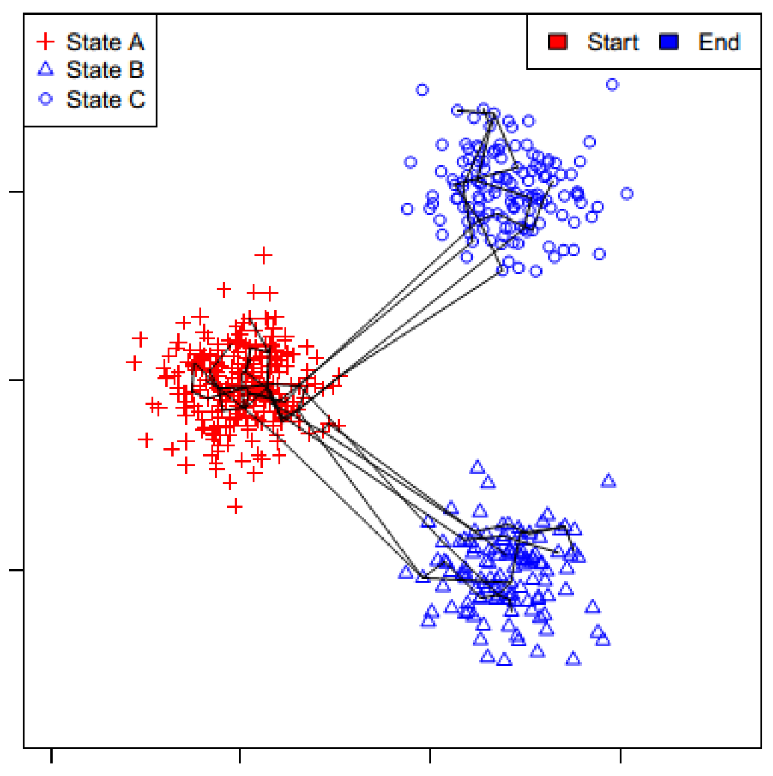

Figure 2

The Pseudotemporal Bootstrap

1. A sample with replacement is taken from the data set that includes at least one member from each of the two classes.
2. A distance matrix is calculated from the sampled points. This matrix represents a graph with edges connecting each point to every other with weights generated by finding the Euclidean distance between them, in terms of the attributes of the point.
3. A minimum spanning tree is constructed from distance matrix.
4. The start and end point of the trajectory is randomly selected from the two classes representing possible start and end points in the sample.
5. The shortest path, within the minimum spanning tree, between the start and end point is calculated.

Pseudotemporal Bootstrap Evaluation

When trajectories are created from real world data sets it is difficult to evaluate how well they represent theoretical longitudinal data. To evaluate and test trajectory methods in a controlled fashion it was decided to create synthetic data that would be made up of sequences made from a known HMM. Figure 1 shows some sample trajectories drawn between a model where the middle cluster is stretched on one axis and its centre of mass is moved away from the other clusters in the other axis. In this example the PTB method does not estimate the sequences that produced the data very well. In Figure 2 the HMM generates sequences that lead to one of two possible clusters. The sequences that produced the data are much better estimated using the PTB method.

Contact: **thomas@dcs.bbk.ac.uk**

Hamiltonian Bootstrap

This method of producing trajectories is a development of the Christofides Algorithm[3]. Which is an algorithm for finding an approximate solution to the Travelling Salesman Problem. The Hamiltonian Bootstrap method makes use of the same sampling method seen in the Pseudotemporal Bootstrap. The main difference between this method and the Pseudotemporal Bootstrap is that it uses every sampled point to create a trajectory. This method generates a Hamiltonian Path between the start and end points using all the points in a sample taken from the Hamiltonian Path. This is a path between two vertices that visits each vertex in the graph exactly once. In order to ensure that the start and end points are at the beginning and end of the path the distance matrix used to produce the MST has the distance between the start and end points forced to zero.

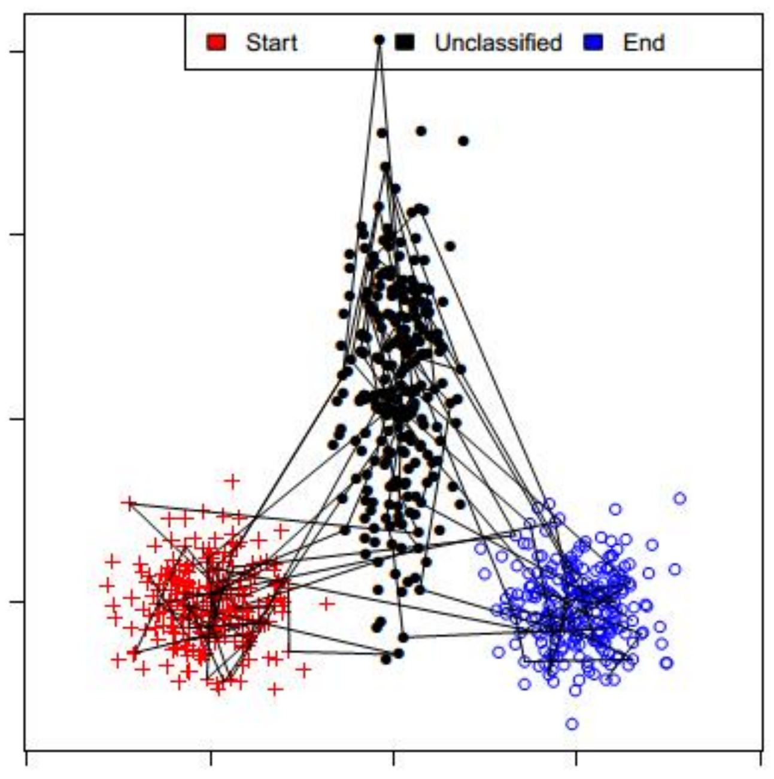

Figure 3

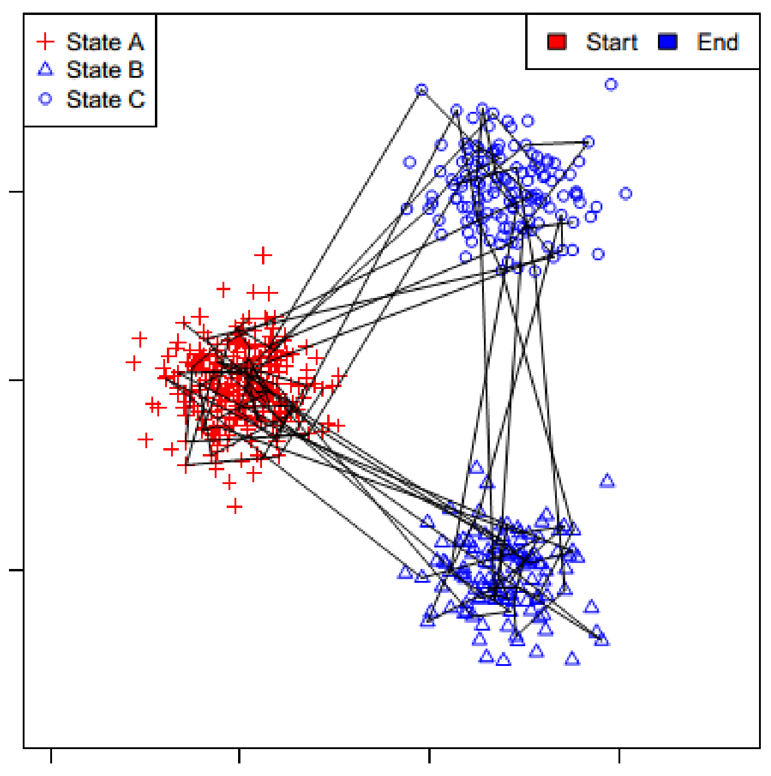

Figure 4

The Hamiltonian Bootstrap

1. A sample with replacement is taken from the data set that includes at least one member from each of the two classes.
2. A distance matrix is calculated from the sampled points. This matrix represents a graph with edges connecting each point to every other with weights generated by finding the Euclidean distance between them, in terms of the attributes of the point.
3. The distance between the start and end points in the distance matrix is forced to be zero.
4. A minimum spanning tree is constructed from distance matrix.
5. An edge list is created from the MST and each edge is duplicated..
6. An Euler Tour is created from the duplicated edge list.
7. Non-unique vertices are removed to create the shortest possible trajectory.

Hamiltonian Bootstrap Evaluation

Figure 3 and 4 show the same HMM clusters from Figures 1 and 2 with trajectories produced by the Hamiltonian Bootstrap method. In Figure 3 the new method is more successful than the PTB in estimating the sequences that created the data. In Figure 4 the trajectories created by the new method often travel between end states and are much less successful than the PTB.

Different trajectory methods better model different kinds of data.

Summary

This research has investigated the use of an existing trajectory method and developed a new method. It has found that they each have different strength and weaknesses when it comes to estimating the underlying structure of data. In the future this research will investigate selecting appropriate trajectory methods for different data sets and combining trajectory methods to model more complicated data sets.

References
[1]Tucker, A., & Garway-Heath, D. (2010). The Pseudotemporal Bootstrap for Predicting Glaucoma From Cross-Sectional Visual Field Data. IEEE Transactions on Information Technology in Biomedicine, pp.79-85.
[2]Baum, L. E., & Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. The Annals of Mathematical Statistics, 37(6), pp. 1554-1563
[3]Christofides, N. (1976). Worst-case analysis of a new heuristic for the travelling salesman problem. Report 388, Graduate School of Industrial Administration, CMU, 1976