Math/Csc 870: Computational-Geometry
Final Project Proposial

Inho Choi

San Francisco State University,
Department of Mathematics

e-mail: hasbegun@sfsu.edu

Oct 20, 2005

1 Topic

The topic of the project that I will fulfill is "Geometric Approaches for Reconstructing Time-Series Data", which is one of the topics that have been proposed.

2 Motivations

We can find a lot of processes that make time sequence data. A biological processes, for instance, are difficult to estimate due to the various factors such as various sampling rate and synchronization. But we need mothods that can utilize multi-dimensional data in order to reconstruct the series of unorder or poorly ordered set of data.

3 Research Method

1. Articles: There are a couple of articles, but the main article that I will focus on is Reconstructing the temporal ordering of biological samples using microarray data by Paul M. Magwene, Paul Lizardi, and Junhyong Kim at Yale University.
   Besides the above article, I will also pay attention to following articles:
   - T. Hastie and W. Stuetzle, "Principal Curves", J of Am. Stat. Assoc., Vol 84, pp. 502-516.
   - R. Singh, V. Cherkassky, and N. P. Papanikolopoulous, "Self-Organizing Maps for the Skeletonization of Sparse Shapes", IEEE Transactions on Neural Networks, Vol. 11, No. 1, pp. 241-248, 2000.
   - B. Kegl et.al., "Learning and Design of Principal Curves" IEEE Trans. on PAMI, Vol 22, 2000.
2. Algorithm Implemention: I will analyze the MST(Minimum Spanning Tree)-based algorithms in the article and implementate them.
3. Test Algorithms: Collect data sets and test out the robustness of the proposed algorithms in the article.
4. Improve Algorithms: Research similar works that have been done and improve the existing and proposed algorithms in the paper.
5. Test New Algorithms: Use the same data sets and test out the new algorithms. Compare and contrast the old and new algorithms.

4 Introduction

Natural biologocal systems are dynamical. Usually we collect samples, tissues, and other relevant units at intervals in order to study the behavior of dynamic biological phenomena.

Definition: ***A time series*** is an ordered collection of samples.

Mnay problems with sampling associated with dynamic biological systems are difficult to obtain exact or accurate the time-series since (1) they are usually drawn from population without synchronization, hence it doesn't contain information of temporal process. (2) Rate heterogeneity among the sample make even more complicated problems because sample ordering are correct with respect to the absolute time but incorrect with respect to the dynamics of biological process. The fundamential problem that this research will investigate is the problem that there no clear phenotypic or genotypic markers in that it may be impossible to explicitly time index collected samples. So we will use microarray data [1] with an assumption that the temporal changes in the transcriptome are relatively smooth and continuous.

5 Background

Assume that the biologocal study can be treated as a continuous funcion then the problem of ordering can be solved by reconstructing vector valued functions.

Definition: **Vactor valued function** $\overrightarrow{f}(t) = [x_1(t), x_2(t), ..., x_d(t)]$ where each d-dimentional point on the function represents the state of a system at a particular point in time.

$\overrightarrow{\Psi}(s) = \overrightarrow{f}(s) + \overrightarrow{\delta}(s)$ where $\overrightarrow{\delta}(s)$ is a noise vector from some distribution. More precisily the noise vector $\overrightarrow{\delta}(s) = \overrightarrow{\delta}$ meaning that it is time homogeneous. An observed data set consists of a finite data set, $V = \overrightarrow{\Psi}(s_0), \overrightarrow{\Psi}(s_2), ..., \overrightarrow{\Psi}(s_n)$ where we assume that the sampled time,s, is unknown excpet the starting and ending points.

The problem of*estimating the geometry* of $\overrightarrow{f}(s)$ from the finite points has been refered to as*the curve reconstruction.*
There are two major techniques in order to solve the curve reconstruction problems.

1. Polygonal reconstruction.
   Definition: Polygonal Reconstruction of $\overrightarrow{f}(s)$ *from* V is the procedure that connects every pair of samples that are adjacent of $\overrightarrow{f}(s)$ and ot others. The polygonal reconstruction assumes following:

   - The samples are collected from a smooth embeded in R^d dimensions.
   - Samples are collected without errors.
   - Sampling intensity is sufficient to achive reconstruction.

2. Principal curves.
   Definition: ***Principal Curves*** self-consistent smooth curve that passes though the middle of a d-dimentional data cloud.

- The samples are collected from a smooth embeded in R^d dimensions.
- Samples are collected without errors.
- Sampling intensity is sufficient to achive reconstruction.

[1]The microarray data such as the gene expression data may be publicly available at http://genome-www.stanford.edu/cellcycle and http://caulobacter.stanford.edu/CellCycle

6 Expected Results

At the end of the semester I will present following:

1. Well working algorithms that have been proposed by the article and improved algorithms.
2. Compared and contrasted results of old and new algorithms.