Dimension reduction of high dimensional data

May 26, 2006

1 Background

The 21st century has been called the "century of data" [1]. As technology progresses our ability to collect more and more data increases rapidly. Scientists and engineers are now collecting data using hyper-spectral imaging, DNA micro-arrays, biomedical spectroscopy, high definition video, as well as many other new data modalities. All these sources generate heretofore unheard of masses of data. The task then becomes how to understand this sea of data in order to make intelligent decisions based on it. This is becoming a pervasive problem in modern science, and a true challenge for the data analyst.

Traditional statistical science used to revolve around data with a large number of samples (instances, individual records), and a relatively small number of well chosen features (attributes) for each sample. For example, a statistical survey may track the heights (a single feature - height) of 100 school age children (samples) in different locations to see how they vary. With a few features and many samples robust and accurate statistical conclusions can be reached. On the other hand, a cDNA micro-array experiment (a modern biomedical assay that takes a "snapshot" of the activity of every gene in an organism at a given time) can collect numbers representing the expression level of 20 000 different genes (features) for each individual in the experiment, but because of cost and availability of experimental subjects, the experiment consists of only a handful, or at best a hundred distinct samples.

It is useful to think of samples as points in a space whose dimensionality is the number of features in the problem. Thus, the DNA micro-array example above can be expressed as a few dozen or a hundred points lying in a 20 000 dimensional space. In such a high dimensional space, Bellman's curse of dimensionality applies. The few samples that are available cover the space in an extremely sparse manner, making traditional statistical analysis difficult or impossible, without additional structure. Fortunately, this additional structure is often available. High dimensional experimental data often lie on relatively low dimensional sub manifolds of the ambient space, for a variety of reasons.

2 The Problem

For the sake of specificity we restrict ourselves to one type of data for this proposal, cDNA microarrays. Microarrays are a relatively new technology that allow a scientist to determine the expression, or level of activity of many, or even all genes in the cells of a tissue sample. For a primer on this technology see, for example, http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html. A typical microarray experiment will measure the expression levels of tens of thousands of genes (and often other DNA sequences) over a few dozen or perhaps a hundred samples. This technology can be applied to determine the relative expression levels of, for example, genes involved in the progression of cancer [2], or central nervous system genes involved in the host response to infection with scrapie [3] or a host of other diseases.

Genes exist in regulatory networks where the product of one gene serves to promote or inhibit the expression levels of others. Thus one expects there to be hidden patterns in gene expression data that allow some information about these networks to be determined. Given the immense amount of data involved with each sample, however, it is difficult for an experimentalist to make sense of the data. The hope is to be able to use mathematical techniques to reduce the dimensionality and complexity of the microarray data, while at the same time distorting the data as little as possible.

A microarray dataset can be thought of as a number of data points n (the number of experimental samples) contained in a space of dimension m (the number of genes whose expression levels are being measured, where $m >> n$). Is it possible to find a low dimensional manifold of dimension k (where $k < n$) that approximates these points well? How does one estimate the intrinsic dimension of this manifold? How is it identified? How is it parameterized? How does one know when such a manifold exists? Given the existence of such a manifold, how does one project the data onto this manifold, and analyze it? How can one validate that the manifold identified is in fact a good representation to the data? How does the uncertainty in the data influence identifiability of all the things we want to know about the manifold?

Work has been done on dimensionality reduction for quite a while. Principal component analysis (PCA) allow data to be reduced to a low dimensional set capturing the most variance, and many more recent algorithms build and extend on this framework. More recently, several researchers have proposed methods for nonlinear reduction to a low dimensional spanning manifold [4] [5], but these methods are limited in their reconstruction ability, and don't seem to perform well on the massive reduction (from hundreds of dimensions down to a handful) and sparse data associated with microarray applications.

The industrial applications of an algorithm to identify such a low dimensional spanning manifold are immense. Algorithms that help researchers make sense of gene expression data will give deeper understanding of the working of the cell, and drive drug discovery and medical technology for many years.

References

[1] "High dimensional data analysis - the curses and blessings of dimsionality" Donoho, D. at AMS conference "Math Challenges of the 21st Century" Los Angeles, August 6-11, 2000.

[2] "Use of a cDNA microarray to analyse gene expression patterns in human cancer.", DeRisi, J. et. al., Nat Genet. 1996 Dec;14(4):367-70.

[3] "Identification of central nervous system genes involved in the host response to the scrapie agent during preclinical and clinical infection.", Booth S, et. al. J Gen Virol. 2004;85(Pt 11):3459-71.

[4] "A Global Geometric Framework for Nonlinear Dimensionality Reduction" Joshua B. Tenenbaum, Vin de Silva, and John C. Langford Science 22 December 2000: 2319-2323.

[5] "Nonlinear Dimensionality Reduction by Locally Linear Embedding" Sam T. Roweis and Lawrence K. Saul, Science 22 December 2000: 2323-2326.