ARTU Research Project: First Report
Anthony Grebe under the direction of Professor Victor Wickerhauser

One procedure for studying data in high-dimensional space is to find an approximate embedding of this data in a lower-dimensional space. Currently, an algorithm called Isomap is widely used for this purpose. In this research project, we will explore diffusion mapping, an alternative algorithm.

The diffusion mapping algorithm is based on Markov chains. A Markov chain is a process in which a system moves among various states in a state-space (which, in this project, we assume to be finite, and we number the states 1 to n) but has no memory of where it has been. Thus, if the system is in state i, the probability that it will move to state j is a constant of the system, which we call P_{ij}. These probabilities can be written as a transition matrix P; then P is row-stochastic ($\sum_{j=1}^{n} P_{ij} = 1 \forall i$) since, from state i, the system must move to exactly one state j with probability 1. The system may be allowed to stay in its current state; i.e. P_{ii} need not be 0.

If the exact state of the system is unknown, the probability distribution of states can be written as a row vector σ with $\sigma_i = \Pr(X = i)$, where X is the random variable representing the state of the system, and with $\sum_{i=1}^{n} \sigma_i = 1$. If the system's probability distribution can be described by σ, then after one step, the system can be described by σP, which is also a probability distribution since

$$\sum_{j=1}^{n} (\sigma P)_j = \sum_{j=1}^{n} \sum_{i=1}^{n} \sigma_i P_{ij} = \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{n} P_{ij} = \sum_{i=1}^{n} \sigma_i = 1$$

We often seek a stationary distribution of the system in which $\sigma P = \sigma$. The Perron-Frobenius Theorem states that if P is a positive matrix (i.e. all entries are positive), its largest eigenvalue λ (in absolute value) is simple, positive, strictly larger than all other eigenvalues' absolute values, and corresponds to a positive eigenvector. Furthermore, if any eigenvector is positive, its corresponding eigenvalue must be λ. Thus, for some positive σ, λ, we have $\sigma P = \lambda \sigma$, and since $\lambda \sigma$ must be a probability distribution, we have $\lambda = 1$. The stationary state can be obtained by taking any initial distribution σ_0 and considering $\lim_{n \to \infty} \sigma_0 P^n$; in practice, one can use a finite value of n to approximate the stationary distribution.

Although the matrices with which we will deal are not necessarily positive (they may have some zero entries), they will be irreducible, which means that $\forall i, j \exists k$ such that $(P^k)_{ij} > 0$. In the context of Markov chains, this means that there is a nonzero probability of the system moving from any state i to any other state j in a finite amount of time. The Perron-Frobenius Theorem can be generalized from positive matrices to nonnegative irreducible matrices; most results still apply to the irreducible case, although there may be multiple eigenvalues with absolute value equal to λ.

The next stage in the research project will be to learn how to construct diffusion maps using Markov chains and how to implement diffusion mapping algorithms in R and Octave. Right now, we are examining the effect of different parameters of the diffusion mapping using simple data structures, and we are comparing the results of diffusion mapping with the Isomap algorithm. We will then move on to explore more complex data sets.