

# Linear Manifold Embeddings of Pattern Clusters

Robert Haralick, Rave Harpaz

Pattern Recognition Laboratory, CS Dept.  
The Graduate Center, City University of New York  
365 Fifth Avenue New York, N.Y. 10016  
haralick@ptah.gc.cuny.edu, rbharpaz@sci.brooklyn.cuny.edu

## Abstract

In this paper we propose a new unifying approach for clustering gene expression data which has both theoretical and practical merit, and which is based on the concept of linear manifolds. The power of linear manifold models lies in their ability to capture various types of linear dependencies and correlations among different attributes of the data. We abstract the problem of clustering gene expression data by showing how various types of pattern clusters, including the ones based on the “bicluster” model can be embedded in arbitrarily oriented linear manifolds of various dimensions. On the theoretical side this abstraction can shed light on the important aspects of the problem. In contrast to existing methods, our approach does not require any assumptions such as the underlying distribution of the data, nor does it require to be posed in advance the type of patterns or number of clusters that are present.

Based on this insight we present an efficient and scalable algorithm which exploits the power of randomness and the geometric properties of linear manifolds, to identify pattern clusters that may be hidden in subspaces of the data. As a consequence of our unifying approach this algorithm is capable of detecting patterns that are not necessarily visible in only a subset of the original measurement features, but also patterns that may be induced by linear combinations of the original measurement features. In addition the algorithm does not require the number of clusters to be predetermined, nor does it require data transformations as preprocessing steps to support any apriori assumptions, which in turn enables it to detect different types of patterns that may co-exist in the same data set. Another advantage of our method is that it can easily be extended to support a probabilistic description of the clusters, which besides aiding in the interpretation of the results, supports the concept of overlapping clusters. The algorithm requires two parameters, a sampling level and a goodness of fit or mean squared error threshold similar to the “biclustering” threshold, and operates by repeatedly sampling minimal subsets of points to create trial manifolds from which the one that best fits a subset of data is selected as a candidate linear manifold embedding of a pattern cluster.