Exploiting the Geometry of Gene Expression Patterns for Unsupervised Learning

Rave Harpaz, Robert Haralick
Pattern Recognition Laboratory
The Graduate Center, City University of New York,
365 Fifth Avenue New York, NY 10016, USA
rbharpaz@sci.brooklyn.cuny.edu, haralick@ptah.gc.cuny.edu

Abstract

Typical gene expression clustering algorithms are restricted to a specific underlying pattern model while overlooking the possibility that other information carrying patterns may co-exist in the data. This may potentially lead to a large bias in the results. In this paper we discuss a new method that is able to cluster simultaneously various types of patterns. Our method is based on the observation that many of the patterns that are considered significant to infer gene function and regulatory mechanisms all share the geometry of linear manifolds.

1 Introduction

The emergence of DNA microarray technology now offers researchers the ability to monitor the behavior patterns of thousands of genes simultaneously. Elucidating these patterns offers potential insight into gene function and regulatory mechanisms. Unsupervised learning techniques such as clustering can be used to address this challenge, as they would cluster genes that behave similarly under a set of conditions. Unlike traditional clustering methods that focus on grouping objects with similar values, in gene expression analysis the emphasis is on the similarity of expression patterns genes exhibit under some subset of conditions. That is, genes that exhibit coherent rise and fall patterns in subspaces of the data. These types of patterns are often discussed in terms of correlation, hence the term pattern/correlation clustering. The most widely studied patterns are the *shift* and *scaling* patterns, which induce only positive linear correlations and are typically referred to as *biclusters* [1, 2, 3, 4]. In the case of a shift pattern the expression pattern of one gene under a set of conditions is offset from another by some constant, whereas in the case of scaling the expression pattern of one gene is a scaler multiple of another. The second pattern is often reduced to the first by various transformations (e.g. log transform), none of which is pattern preserving when more than one type of pattern co-exist in the same experiment [5]. Fig. 1 illustrates the concept of a shift pattern cluster embedded in a 5D subspace, consisting of the first 5 dimensions of an 8D space. In recent studies other types of information carrying patterns have been suggested which are completely overlooked by most gene expression clustering methods [6, 7]. While there is no consensus on what type of patterns should be considered meaningful to infer genetic regulatory structure, in practice each pattern based clustering algorithm postulates a unique underlying "globally expressed" pattern or cluster model, while overlooking or rejecting the possibility that other types of information carrying patterns may exit in the data. This in turn potentially leads to a large bias in the results. In addition many of these methods require the number of clusters or the dimensionality of the subspaces in which they are embedded as input parameters, along with some sort of data transformation to support their underlying model, both which may dramatically effect study conclusions. In a recent study [8] we proposed a new clustering technique which is based on the concept of linear manifolds that are embedded in lower dimensionality subspaces of data, and demonstrated its successful application to various clustering problems. In this paper we extend this concept and propose a new gene expression clustering technique that is based on the observation that many pattern models share the geometry of linear manifolds or equiva-

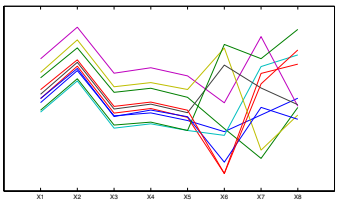

Figure 1. A shift pattern embedded in a 5-dimensional subspace

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00

lently are instances of linear manifolds. By focusing on linear manifolds rather than on particular patterns our method allows the data to "speak for itself", offering a researcher a data exploratory and clustering tool that is not limited to one type of pattern but that takes into account several types of patterns simultaneously and does not require any data transformations prior to the clustering. Moreover our method does not require the number of clusters nor their dimensionality to be posed in advance.

2 Formal Models

A linear manifold is a subspace that may have been translated away from the origin. A subspace is a special case of a linear manifold that contains the origin. For example, a 1D manifold can be geometrically visualized as a line embedded in the space, a 2D manifold as a plane, and a 0D manifold as a point. In the following we denote by D a set of d-dimensional points, $C \subseteq D$ the subset of points that belong to a cluster, x some point in C, $b_1, \ldots, b_d$ a set orthonormal vectors that span a d-dimensional space, B a $d \times k$ matrix whose k columns are a subset of the vectors $b_1, \ldots, b_d$, and $\overline{B}$ a $d \times d-k$ matrix whose columns are the remaining vectors.

Definition 1 (Linear Manifold Model) *Let μ be some point in $\mathbb{R}^d$, λ a zero mean $k \times 1$ random vector whose entries are i.i.d. $U(-R/2, +R/2)$ where R is the range of the data, and ψ is a zero mean $d-k \times 1$ random vector with small variance independent of λ. Then each $x \in C$, a linear manifold cluster is modeled by,*

$$x = \mu + B\lambda + \overline{B}\psi. \tag{1}$$

The idea is that each point in a cluster lies close to a k-dimensional linear manifold with mean μ, which is defined by μ plus the space spanned by the columns of B. Classical clustering algorithms such as K-means assume a 0D linear manifold ($k = 0$) and therefore omit the possibility that a cluster has a non-zero dimensional linear manifold associated with it. On the manifold the points are assumed to be uniformly distributed in each direction according to $U(-R/2, +R/2)$. However this assumption is not binding, and the uniform distribution can be replaced by any other distribution. It is in this manifold that the cluster is embedded, and therefore the intrinsic dimensionality of the cluster will be k. What characterizes this type of cluster is the third component that models a small error associated with each point on the manifold. The idea is that each point may be perturbed in directions that are orthogonal to the subspace spanned by the columns of B. We model this behavior by requiring that ψ be a $(d-k) \times 1$ random vector, normally distributed according to $N(\mathbf{0}, \Sigma)$, where the largest eigenvalue of Σ is much smaller than R the range of the data, otherwise the signal cannot be distinguished from the noise. Thus each point has a random component in the orthogonal subspace spanned by the columns of $\overline{B}$. For example adding an error term to a 1D manifold transforms it into an elongated thin cylinder.

To define pattern clusters we extend the notation used to define a manifold by letting C now be a set points manifesting some pattern such as the shift or scaling pattern in some k-dimensional subspace, $\mu \in \mathbb{R}^k$, $\overline{\mu} \in \mathbb{R}^{d-k}$, ϕ a scaler uniformly distributed within some range, λ a zero mean $d-k \times 1$ random vector whose entries are uniformly i.i.d. within some range, and ψ a $k \times 1$ normal random vector with zero mean and small variance.

Definition 2 (Shift Pattern Model) *Each $x \in C$, a shift pattern cluster can be modeled by,*

$$x = B\mu + B\mathbf{1}_k\phi + B\psi + \overline{B}\overline{\mu} + \overline{B}\lambda \tag{2}$$

In this case $B\mu$ models the mean or template of the pattern in the k dimensional subspace of relevant features, $B\mathbf{1}_k\phi$ a possible shift from the mean of the pattern, $B\psi$ a small error associated with the pattern, while $\overline{B}\overline{\mu}$ and $\overline{B}\lambda$ model the random behavior of points in the subspace of irrelevant features. For example in Fig. 1 the matrix $(B|B_c) = I_8$ and $k = 5$.

Proposition 1 *Every point x in a d-dimensional space that fits the shift pattern model, also fits the linear manifold model, where the dimension of the linear manifold is $d - k + 1$, and the linear manifold model is given by,*

$$x = (B|\overline{B}) \begin{pmatrix} \mu \\ \overline{\mu} \end{pmatrix} + \left(B\frac{\mathbf{1}_k}{\sqrt{k}}|\overline{B}\right) \begin{pmatrix} \sqrt{k}\phi + \frac{\mathbf{1}'_k}{\sqrt{k}}\psi \\ \lambda \end{pmatrix} + B\left(I_k - \frac{\mathbf{1}_k\mathbf{1}'_k}{k}\right)\psi \tag{3}$$

The first term in eq. (3) models the mean of the manifold/pattern cluster. In the subspace of relevant features the dimension of the manifold is one and is spanned by the vector $B\frac{\mathbf{1}_k}{\sqrt{k}}$. This vector together with the vectors of $\overline{B}$ in the second term define the manifold in the full space whose dimension is the rank of $\left(B\frac{\mathbf{1}_k}{\sqrt{k}}|\overline{B}\right) = 1 + d - k$. Finally, the third component includes the vectors spanning the space orthogonal to the manifold, modeling only a portion of the error associated with the pattern, since the other portion is absorbed in the manifold. The proof of prop. 1 is straight forward, multiplying out the three terms in eq. (3) gives eq. (2).

Similarly for the scaling pattern we have the following.

Definition 3 (Scaling Pattern Model) *Each $x \in C$, a scaling pattern cluster can be modeled by,*

$$x = \phi B\mu + B\psi + \overline{B}\overline{\mu} + \overline{B}\lambda \tag{4}$$

The idea is similar to the shift pattern cluster model, however in this case the scaling in the subspace of relevant features is induced by $\phi B\mu$, where μ can be thought of as the pattern template and ϕ its scaling.

Proposition 2 *Every point x in a d-dimensional space that fits the scaling pattern model, also fits the linear manifold model, where the dimension of the linear manifold is $d - k + 1$, and the linear manifold model is given by,*

$$x = \overline{B}\overline{\mu} + \left(B\frac{\mu}{\parallel \mu \parallel} | \overline{B} \right) \begin{pmatrix} \parallel \mu \parallel \phi + \frac{\mu'}{\|\mu\|}\psi \\ \lambda \end{pmatrix}$$

$$+B\left(I_k - \frac{\mu\mu'}{\parallel \mu \parallel^2} \right)\psi \qquad (5)$$

Note that for the scaling pattern the 1D manifold (line) embedded in the space of relevant dimensions in now spanned by $B\frac{\mu}{\|\mu\|}$.Thus, the geometrical difference between the two patterns in the space of relevant dimensions is that for the shift pattern the orientation of the manifold is in the direction of the vector $\mathbf{1}$ and the manifold does not necessarily have to pass through the origin, whereas for the scaling pattern the direction is some vector μ and the manifold passes through the origin. Fig. 2 illustrats the linear manifold embedding of the two patterns in a 3D space, each consisting of 10 points, where all dimensions are relevant, i.e. $k = d = 3$, whereas Fig. 3 illustrates a linear manifold embedding of a shift pattern where only two of the dimensions (x_1 and x_2) manifest the pattern, i.e. $k = 2$.

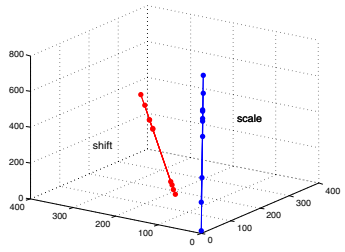

Figure 2. A linear manifold embedding of shift and scaling patterns in the full space.

The bicluster concept introduced by Cheng et al. [1] is based on a two-way *analysis of variance* (ANOVA) which attempts to decompose the total variation in the data into contributions from two different sources, which in the context of gene expression analysis are the genes and conditions. If we let Y_{ij} denote the value of the ith gene under the jth condition, and μ the overall mean of Y, then a two-way ANOVA says that

$$Y_{ij} = \mu + \phi_i + \psi_j + \epsilon_{ij} \qquad (6)$$

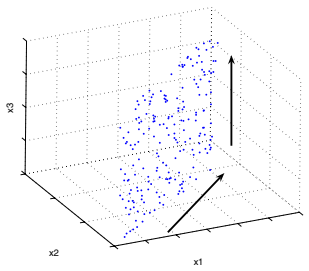

Figure 3. A linear manifold embedding of a shift pattern manifested in only 2 (x_1 and x_2) of 3 dimensions.

where ϕ_i and ψ_j are the residual effects of the ith gene and jth condition on the overall mean of expression levels, and $\epsilon_{ij} \sim N(0, \sigma)$ an error term. Since this model assumes an additive effect of the two sources causing the variation, it essentially models a shift pattern. Putting eq. (6) in parametric form we can model each each k-dimensional point x_i $i = 1, \ldots, n$ belonging to a bicluster as follows,

$$x_i = \mathbf{1}_k\mu + \mathbf{1}_k\phi_i + \psi + \epsilon \qquad (7)$$

where $\mu \in \mathbb{R}^k$, ϕ_i a scalar denoting the residual effect of the ith gene, $\psi = (\psi_1, \ldots \psi_k)'$ a vector containing the residual effects of the conditions, and ϵ a realization of a multivariate normal $N(\mathbf{0}, \sigma^2 I_k)$ containing the error terms. In the case that the bicluster is embedded in some k-dimensional subspace of the data we prefix each term in eq. (7) by B and add the terms $\overline{B}\overline{\mu}$ and $\overline{B}\lambda$ as before.

Proposition 3 *Every point x_i in a d-dimensional space that fits the bicluster model, where the bicluster is embedded in some k-dimensional space, also fits the linear manifold cluster model, where the dimension of the linear manifold is $d - k + 1$, and the linear manifold model is given by,*

$$x_i = (B|\overline{B}) \begin{pmatrix} \mathbf{1}_k\mu + \psi \\ \overline{\mu} \end{pmatrix} + \left(B\frac{\mathbf{1}_k}{\sqrt{k}} | \overline{B} \right) \begin{pmatrix} \sqrt{k}\phi_i + \frac{\mathbf{1}'_k}{\sqrt{k}}\epsilon \\ \lambda \end{pmatrix}$$

$$+B\left(I_k - \frac{\mathbf{1}_k\mathbf{1}'_k}{k} \right)\epsilon \qquad (8)$$

Note that shift pattern clusters and biclusters are embedded in the same type of manifold, i.e., the vectors spanning the manifold in both cases are the columns of $(B\frac{\mathbf{1}_k}{\sqrt{k}}|\overline{B})$. This coincides with the previous statement about the assumption of additivity.

As mentioned earlier most methods are not able to detect negative correlations, and only consider patterns which are visible in some subset of the original measurement features. They assume that the matrices B and $\overline{B}$ consist of

different columns of a d-dimensional identity matrix. However, it is possible that patterns may be visible in some linear combination of the original measurement features, or that expressions levels are determined by some linear combination of other expression levels and other factors such as column/row means. To model these cases all that is necessary is to replace the columns of B and $\overline{B}$ by any set of orthonormal vectors that span $\mathbb{R}^d$. Note that by changing these matrices the geometry of the resulting pattern clusters does not change, i.e. they are still characterized by linear manifolds. Fig. 4 for example shows what a shift pattern that is only visible in some linear combination of the original measurement features and embedded in a 1D linear manifold, would look like when viewed through the original measurement features. In addition the figure reveals that the features are still highly correlated, but rather than being only positively correlated some are also negatively correlated (features 1, 2, and 3 are negatively correlated with 4 and 5).

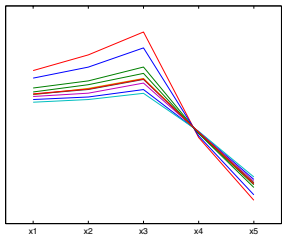

Figure 4. A Shift pattern induced by a linear combination of the original measurement features.

3 A Linear Manifold Clustering Algorithm

Since the main characteristic of linear manifold clusters is that their constituting points are located on or close to a lower dimensional linear manifold, the problem of clustering data which is embedded in linear manifolds can be restated as the problem of fitting different linear manifolds of different dimensions to different subsets of the data. This is similar to the regression problem of fitting a hyperplane (a linear manifold of dimension one less than the dimension of the space) to the data. However, since the linear manifolds localize to different subsets of the data, a global approach such as in the regression case can not be taken. Instead, by using random sampling and by taking into account the geometry of linear manifolds we can find sufficiently accurate estimates of linear manifolds in which clusters are embedded. Our method operates as follows. Minimal subsets of points are repeatedly sampled to construct trial linear manifolds of various dimensions. A k-dimensional trial manifold is constructed by sampling $k+1$ points from the data. By choosing one of these points as the origin and taking the difference of each of the other k points from the origin we obtain the k basis vectors that span the manifold. Histograms of the distances of the data points to each trial manifold are computed. These histograms will either be unimodel or reveal a mixture of two populations, one with a mode near zero which corresponds to the points located in the vicinity of a trial manifold, i.e. the points that are potentially embedded in the manifold, and those that are not. The sampling corresponding to the histogram having the best separation between a mode near zero and the rest of the data is selected, and the data points are partitioned on the basis of the best separation. The best separation can be thought of as the best fitting of a subset of points to a linear manifold. The separation criteria is computed by first thresholding the histogram using Kittler and Illingworth [9] technique, and then using the threshold to compute a *discriminability* measure between the two populations of the mixture. After partitioning the data, the sampling is then repeated on each block of the partitioned data until no further partitioning is possible, in which case a cluster is assumed to be found. The points belonging to this cluster are then removed from the data set and the algorithm is applied to the remaining points, detecting one cluster at a time. A more detailed description of the algorithm is presented in [8].

4 Experiments

We conducted extensive tests on two typical data sets that have become standard benchmarking data sets for clustering gene expression data–the yeast *Saccharomyces Cerevisiae* cell cycle expression data [1] [10], and the *Colon Cancer* data [2] [11]. The yeast data contained 2884 genes and 17 conditions, and the cancer data contained 2000 genes and 62 tissue samples, of which 40 were colon tumor and 22 normal colon samples. Our method was applied on the yeast data in two different modes. One requiring it to detect all types of patterns that may be embedded in linear manifolds, and the other forcing it to search only for shift patterns so that the results can be compared with other reported results. Forcing the algorithm to search only for shift patterns was achieved by replacing our separation/discriminability criteria with the *mean squared residue score* (MSRS) used in [1] to assess the quality of biclusters. Applied in this mode our algorithm detected 98 clusters. We compared these clusters with the 100 clusters reported by the biclustering algorithm [1]. Fig. 5 shows boxplots highlighting the differences between the two results in terms of the size, dimension, and MSRS of the clusters detected. Our method detected smaller clusters (maximum of about 150 genes per cluster as opposed to 1000), which biologically makes sense,

[1] obtained from http://arep.med.harvard.edu/biclustering/
[2] obtained from http://microarray.princeton.edu/oncology/

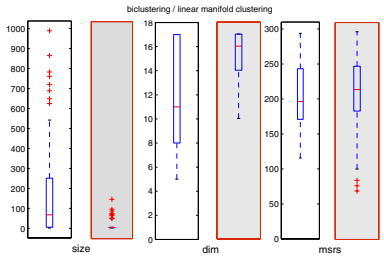

Figure 5. Boxplots comparing biclusters and linear manifold clusters (gray boxs) of the yeast genome.

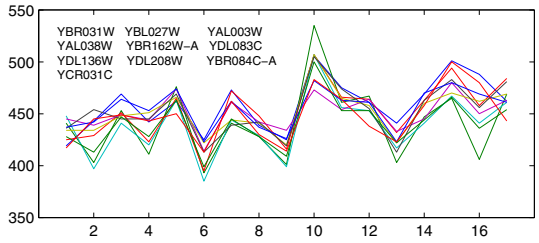

Figure 6. A yeast cluster manifesting a scaling pattern and negative correlations.

Table 1. MIPS gene function enrichment.

Genes in Cluster	MIPS Functional Category	Genes in Category	Clustered Genes	P-value
68	ribosome biogenesis	215	49	<1e-14
	protein synthesis	359	52	<1e-14
	cytoplasm	554	54	<1e-14
7	endoplasmic reticulum	157	4	1.251e-05
12	DNA processing	251	6	2.934e-06
	cell cycle and DNA processing	628	8	3.345e-06

since the whole yeast genome contains roughly only 6000 genes, and typical functional categories of the yeast genome contain dozens rather than hundreds of genes. Our method also found clusters in higher dimensions (as low as 10). Finally, the median MSRS of the our clusters is slightly larger, but our method was able to find many clusters with much smaller MSRS than the ones found by the biclustering algorithm. We also evaluated the biological significance of the clusters our algorithm produced by means of *function enrichment* [10]–the degree to which the clusters grouped genes of common function. This was done by computing for each cluster P-values (using the hypergeometric distribution) of observing a certain number of genes within a cluster from a particular MIPS [3] functional category. Some of the clusters demonstrated significant grouping (very small P-values) of genes within the same functional class. Table 1 shows three of them. Fig. 6 shows a cluster of ten genes embedded in a 3-dimensional linear manifold, which was discovered using the original criterion function, and which manifests a scaling pattern and negative correlations. Among these genes, six (YBL027W, YBR031W, YBR084C-A, YCR031C, YDL083C, YDL136W) enriched the ribosome biogenesis functional category with a P-value of 2.848e-07.

Applied on the cancer data our method detected 14 clusters. The goal was to find gene clusters that differentiate the cancerous tissues from the normal ones. These genes can then be used to construct a classifier for diagnosis purposes. We found one such cluster containing 229 genes expressed in 12 cancerous tissues and no normal tissues. We also found one cluster containing 44 genes which were expressed in all 62 tissues, implying that these genes cannot be used to differentiate the normal from the cancerous tissues. The rest of the clusters contained a portion of the samples but none with an overwhelmingly majority of normal or cancerous tissues.

[3] Munich Information Center for Protein Sequences, http://mips.gsf.de/

References

[1] Y. Cheng and G. Church. Biclustering of expression data. In *International Conference on Intelligent Systems for Molecular Biology*, pages 93–103, 2000.

[2] Yang et al. δ-clusters: Capturing subspace correlation in a large data set. In *ICDE*, pages 517–528, 2002.

[3] Wang et al. A fast algorithm for subspace clustering by pattern similarity. In *SSDBM*, 2004.

[4] Sra et al. Minimum sum-squared residue co-clustering of gene expression data. In *SDM*, 2004.

[5] Jess S. Aguilar-Ruiz. Shifting and scaling patterns from gene expression data. *Bioinformatics*, 21(10):3840–3845, 2005.

[6] Erdal et al. A time series analysis of microarray data. In *BIBE*, page 366, 2004.

[7] Böhm et al. Computing clusters of correlation connected objects. In *ACM SIGMOD*, pages 455–466, 2004.

[8] R. Haralick and R. Harpaz. Linear manifold clustering. In *4th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2005), Springer Verlag, LNAI 3587*, pages 132–141, 2005.

[9] J. Kittler and J. Illingworth. Minimum error thresholding. *Pattern Recogn.*, 19(1):41–47, 1986.

[10] Tavazoie et al. Systematic determination of genetic network architecture. *Nature Genetics*, 22(3):281 – 285, 1999.

[11] Alon et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In *PNAS*, volume 96, pages 6745–6750, 1999.