Linear Manifold Embeddings of Pattern Clusters

Robert Haralick    Rave Harpaz

Pattern Recognition Laboratory
The Graduate Center,
City University of New York

DIMACS 2005
Informally, a linear manifold is a subspace that may have been shifted away from the origin.

A subspace is an instance of a linear manifold that contains the origin.
Linear Manifolds

- Each point $x_i$ in a set of a $d$-dim points that all lie on an $m$-dim linear manifold can be modeled as:

$$x_i = \mu + \begin{pmatrix} \vdots & \cdots & \vdots \\ b_1 & \cdots & b_m \end{pmatrix} \lambda_i$$

- Each point $x_i$ in a set of points that all manifest a shift pattern in the full space can be modeled as:

$$x_i = p + 1L_i$$

e.g.

$$x_1 = \begin{pmatrix} 2 \\ 6 \\ 4 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \cdot 2 = \begin{pmatrix} 4 \\ 8 \\ 6 \end{pmatrix}$$
Each point $x_i$ in a set of a $d$-dim points that all lie on an $m$-dim linear manifold can be modeled as:

$$x_i = \mu + \begin{pmatrix} \vdots & \cdots & \vdots \\ b_1 & \cdots & b_m \\ \vdots & \cdots & \vdots \end{pmatrix} \lambda_i$$

Each point $x_i$ in a set of points that all manifest a scale pattern in the full space can be modeled as:

$$x_i = pL_i$$

e.g.

$$x_1 = \begin{pmatrix} 2 \\ 6 \\ 4 \end{pmatrix} 2 = \begin{pmatrix} 4 \\ 12 \\ 8 \end{pmatrix}$$
Shift and Scale Patterns as Linear Manifolds

\[ PC_{1\text{shift}} = (0.5774, 0.5774, 0.5774)' \]
\[ PC_{1\text{scale}} = (0.3810, 0.2540, 0.8890)' \]

\[ R = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \]

\[ \text{PearsonR} = 1 \]
\[ \text{MSR}_{\text{shift}} = 0 \]
\[ \text{MSR}_{\text{scale}} = 3236.3 \]
Linear Manifolds - Patterns in Subspaces

- Shift pattern that exists only in a subspace:

\[
x_i = B_r(\mu_r + 1_r \phi_i) + B_c(\mu_c + \lambda_i) = B_r \mu_r + B_r 1_r \phi_i + B_c \mu_c + B_c \lambda_i
\]

\[
(B_r | B_c) = I_8
\]

- The linear manifold embedding:

\[
x_i = (B_r | B_c) \begin{pmatrix} \mu_r \\ \mu_c \end{pmatrix} + \left( B_r \frac{1_r}{\sqrt{r}} | B_c \right) \begin{pmatrix} \sqrt{r} \phi_i \\ \lambda_i \end{pmatrix}
\]
Linear Manifolds - Patterns in Subspaces

Linear Manifold Embeddings of Pattern Clusters

Haralick, Harpaz
Definition (The Linear Manifold Cluster Model)

Let $D$ be a set of $d$-dimensional points, $C \subseteq D$ a subset of points that constitute a cluster, $x_i$ some point in $C$, $b_1, \ldots, b_d$ an orthonormal set of vectors that span $\mathbb{R}^d$, $(b_i, \ldots, b_j)$ a matrix whose columns are the vectors $b_i, \ldots, b_j$, and $\mu$ some point in $\mathbb{R}^d$. Then each $x_i \in C$ is modeled by,

$$x_i = \mu + \left( \begin{array}{ccc} \vdots & \cdots & \vdots \\ b_1 & \cdots & b_m \\ \vdots & \cdots & \vdots \end{array} \right) \lambda_i + \left( \begin{array}{ccc} \vdots & \cdots & \vdots \\ b_{m+1} & \cdots & b_d \\ \vdots & \cdots & \vdots \end{array} \right) \psi_i$$
**Shift Pattern** - Bicluster (Cheng 00), Floc (Yang 02), pCluster (Wang 02)

**Definition (Shift Pattern Cluster Model)**

Let $D$ be a set of $d$-dimensional points, $C \subseteq D$ the subset of points manifesting a shift pattern in some $r$-dimensional subspace of the data, and $x_i$ some point in $C$. Then each $x_i \in C$ can be modeled by,

$$x_i = B_r \mu_r + B_r \overline{1}_r \phi_i + B_r \psi_i + B_c \mu_c + B_c \lambda$$

**Proposition**

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

$$x_i = (B_r | B_c) \begin{pmatrix} \mu_r \\ \mu_c \end{pmatrix} + \left( B_r \frac{1_r}{\sqrt{r}} | B_c \right) \begin{pmatrix} \sqrt{r} \phi_i + \frac{1_r'}{\sqrt{r}} \psi_i \\ \lambda \end{pmatrix} + B_r \left( I_r - \frac{1_r 1_r'}{r} \right) \psi_i$$
Scale Pattern

Definition (Scale Pattern Cluster Model)

Let $D$ be a set of $d$-dimensional points, $C \subseteq D$ the subset of points manifesting a scale pattern in some $r$-dimensional subspace of the data, and $x_i$ some point in $C$. Then each $x_i \in C$ can be modeled by,

$$x_i = \phi_i B_r \mu_r + B_r \psi_i + B_c \mu_c + B_c \lambda_i$$

Proposition

*Every point $x_i$ in a $d$-dimensional space that fits the scale pattern cluster model, also fits the linear manifold cluster model, where the dimension of the linear manifold is $d - r + 1$, and the model is given by:*

$$x_i = B_c \mu_c + \left( B_r \frac{\mu_r}{\|\mu_r\|} | B_c \right) \left( \|\mu_r\| \phi_i + \frac{\mu_r'}{\|\mu_r\|^2} \psi_i \right) + B_r \left( I_r - \frac{\mu_r \mu_r'}{\|\mu_r\|^2} \right) \psi_i$$
The Bicluster Model (Cheng et al. 00)

- **MSRS** = $H(I, J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (Y_{ij} - \bar{Y}_i - \bar{Y}_j - \bar{Y}_{IJ})^2$

- The Underlying Model - Two Way ANOVA

$$Y_{ij} = \mu + \phi_i + \psi_j + \epsilon_{ij}$$

- Each point in a bicluster can be modeled by:

$$x_i = 1\mu + 1\phi_i + \psi + \epsilon_i$$

where $\phi_i$ is a scalar denoting the residual effect of the $i$-th gene, $\psi = (\psi_1, \ldots, \psi_d)'$ a vector containing the residual effects of the conditions, and $\epsilon_i \sim N(0, \sigma^2I)$
The Bicluster Model (Cheng et al. 00)

Proposition

Every point \( x_i \) in a \( d \)-dim space that fits a bicluster model embedded in an \( r \)-dim subspace, also fits the linear manifold cluster model, where the dimension of the linear manifold is \( d - r + 1 \), and the model is given by:

\[
x_i = (B_r | B_c) \left( \begin{pmatrix} 1_r \mu_r + \psi \\ \mu_c \end{pmatrix} \right) + \left( B_r \frac{1_r}{\sqrt{r}} | B_c \right) \left( \begin{pmatrix} \sqrt{r} \phi_i + \frac{1_r'}{\sqrt{r}} \epsilon_i \\ \lambda_i \end{pmatrix} \right) + B_r \left( I_r - \frac{1_r 1_r'}{r} \right) \epsilon_j
\]
Subspace Clusters

- Consist of a subset of points and a corresponding subset of attributes, such that these points form a dense region in a subspace defined by the set of corresponding attributes.

**CLIQUE** (Agrawal 98), **MAFIA** (Nagesh 99), **PROCLUS** (Aggarwal 99), **ORCLUS** (Aggarwal 00)
Other Instances of Linear Manifolds - Negative Correlations

\[
R = \begin{pmatrix}
1 & -1 & 1 \\
-1 & 1 & -1 \\
1 & -1 & 1 \\
\end{pmatrix}
\]

PearsonR = 0.3181

MSR = 18280

Yip et al. (2004)- HARP, to detect co-regulated genes, create a reflective copy of the data set, cluster and remove the copy.
Other Instances of Linear Manifolds - Linear Combinations of Variables

\[ z = b_0 + b_1 x + b_2 y \]

Pearson \( R = 0.4509 \)

MSR = 8975

Coefficient of multiple determination:

\[ R^2 = \frac{\sum (\hat{z} - \bar{z})^2}{\sum (z - \bar{z})^2} = 1 \]

4C, Böhm et al. (2004)
Other Instances of Linear Manifolds - Latent Variables

1. $x_i = R(\mu + 1_d \phi_i)$
2. $y_i = x_i - R\mu = 1_d \phi_i$
3. $\phi_i = \sqrt{x_i'x_i/d}$
4. $C = \frac{1}{n} \sum_{i=1}^{n} y_i(1_d \phi_i)'$
5. $[u, s, v] = \text{svd}(C)$
6. $R = uv'$
Data Transformations

- Shift pattern
- Log transformation of a shift pattern
- Row mean subtraction on a shift pattern
- Mean/var transformation on a shift pattern

- Scale pattern
- Log transformation on a scale pattern
- Row mean subtraction on a scale pattern
- Mean/var transformation on a scale pattern

- 0–Dim manifold
- Log transformation to a 0–Dim manifold
- Row mean subtraction on a 0–Dim manifold
- Mean/var transformation on a 0–Dim manifold
Main Idea

1. Sample minimal subsets of points to construct trial linear manifolds of various dimensions.
2. Compute distance histograms of the data to each trial manifold.
3. Of all the manifolds constructed, select the one whose associated histogram shows the best separation between a mode near zero and the rest of the data.
4. Partition the data based on the best separation.
5. Repeat the procedure on each block of the partitioned data.
How are trial manifolds sampled?

To construct an $m$-dimensional manifold we need to sample $m + 1$ points.

Example- constructing a 2D manifold
How many trial manifolds need to be examined?
How many trial manifolds need to be examined?

- Assuming there are $\hat{K}$ clusters having approximately the same number of points.
- Then the probability that a sample of $m + 1$ points all come from the same cluster is approximately $\left(\frac{1}{\hat{K}}\right)^m$.
- The probability that out of $n$ samples of $m + 1$ points, none come from the same cluster, is approximately $(1 - (1/\hat{K})^m)^n$.
- $1 - (1 - (1/\hat{K})^m)^n$ will be the probability that at least for one of the samples all of its $m + 1$ points come from the same cluster.
- Therefore the sample size $n$ required such that this probability is greater than some value $1 - \epsilon$ is given by

$$n \geq \frac{\log \epsilon}{\log(1 - (1/\hat{K})^m)}$$
To compute a separation score we first need to find the two classes or distributions involved.

This problem is cast into histogram thresholding problem.
Kittler and Illingworth Minimum Error Error Thresholding (86)

Minimize:
\[
P(\text{error}) = \int_{x > T} p(x|c_1)P(c_1)dx + \int_{x \leq T} p(x|c_2)P(c_2)dx
\]

KI86:
\[
J(T) = 1 + 2 \left( P_1(T) \log \sigma_1(T) + P_2(T) \log \sigma_2(T) \right) - 2 \left( P_1(T) \log P_1(T) + P_2(T) \log P_2(T) \right)
\]

Goodness of separation:
\[
\text{discriminability} = \frac{(\mu_1(T) - \mu_2(T))^2}{\sigma_1^2(T) + \sigma_2^2(T)} \times \text{depth} = J(T') - J(T)
\]
Time Series Clustering (UCI KDD Archive)

600 × 60, A-decreasing trend, B-cyclic, C-normal, D-upward shift, E-increasing trend, F-downward shift.

<table>
<thead>
<tr>
<th></th>
<th>in1</th>
<th>in2</th>
<th>in3</th>
<th>in4</th>
<th>in5</th>
<th>in6</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>out1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>57</td>
<td>0</td>
<td>0</td>
<td>57</td>
</tr>
<tr>
<td>out2</td>
<td>0</td>
<td>0</td>
<td>80</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>81</td>
</tr>
<tr>
<td>out3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>43</td>
<td>0</td>
<td>99</td>
<td>142</td>
</tr>
<tr>
<td>out4</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>98</td>
<td>0</td>
<td>118</td>
</tr>
<tr>
<td>out5</td>
<td>99</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>out6</td>
<td>0</td>
<td>41</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>41</td>
</tr>
<tr>
<td>out7</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>out8</td>
<td>1</td>
<td>36</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>39</td>
</tr>
<tr>
<td>total</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>600</td>
</tr>
</tbody>
</table>

Total Correct=533   Accuracy=88.8333
Yeast Data - mitotic cell cycle $2884 \times 17$


$PC1 = (0.82, 0.95, 1.02, 0.95, 1.02, 0.93, 0.99, 0.97, 0.92, 1.17, 1.05, 1.02, 0.92, 1.04, 1.03, 1.09, 1.04)'$

93%
Yeast Data Results (regular manifolds)

- Size 10: mnr1 = 105.7032, mnr2 = 103.4332, lm dim = 2
- Size 5: mnr1 = 55.8846, mnr2 = 54.9463, lm dim = 2
- Size 44: mnr1 = 208.8599, mnr2 = 184.9669, lm dim = 7
- Size 62: mnr1 = 1600.9392, mnr2 = 1600.9392, lm dim = 1
Yeast Data Results (MSR manifolds)

Biclustering/Linear Manifold Clustering