

Linear Manifold Embeddings of Pattern Clusters

Rave Harpaz Robert Haralick

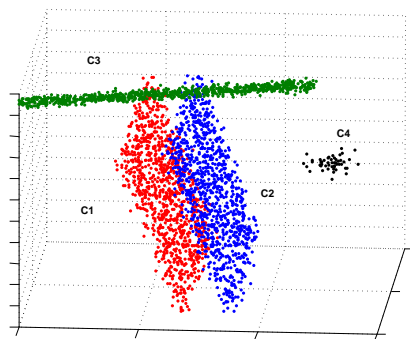
Pattern Recognition Laboratory
The Graduate Center,
City University of New York

DIMACS 2005

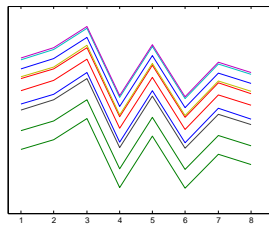


Linear Manifolds

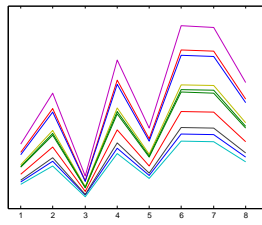
- Informally, a linear manifold is a subspace that may have been shifted away from the origin.
- A subspace is an instance of a linear manifold that contains the origin.



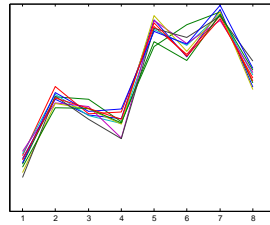
Patterns



shift



scale



0-d lm



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:

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



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:

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

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

$$x_i = p + \mathbf{1}L_i$$

e.g.

$$x_1 = \begin{pmatrix} 2 \\ 6 \\ 4 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} 2 = \begin{pmatrix} 4 \\ 8 \\ 6 \end{pmatrix}$$



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:

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

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

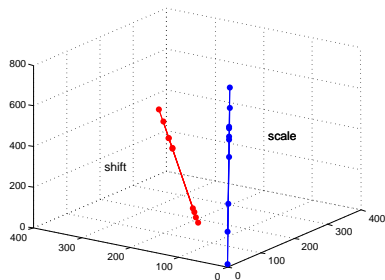
$$x_i = pL_i$$

e.g.

$$x_1 = \begin{pmatrix} 2 \\ 6 \\ 4 \end{pmatrix} 2 = \begin{pmatrix} 4 \\ 12 \\ 8 \end{pmatrix}$$



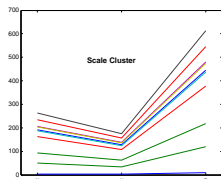
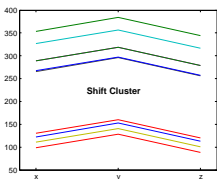
Shift and Scale Patterns as Linear Manifolds



$$PC1_{shift} = (0.5774, 0.5774, 0.5774)'$$

$$PC1_{scale} = (0.3810, 0.2540, 0.8890)'$$

$$R = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$



$$PearsonR = 1$$

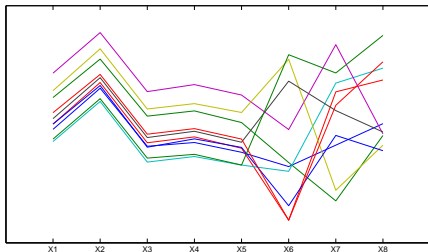
$$MSR_{shift} = 0$$

$$MSR_{scale} = 3236.3$$



Linear Manifolds - Patterns in Subspaces

- Shift pattern that exists only in a subspace:

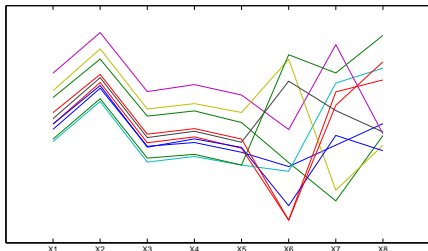


Linear Manifolds - Patterns in Subspaces

- Shift pattern that exists only in a subspace:

$$x_i = B_r(\mu_r + \mathbf{1}_r \phi_i) + B_c(\mu_c + \lambda_i) = B_r \mu_r + B_r \mathbf{1}_r \phi_i + B_c \mu_c + B_c \lambda_i$$

$$(B_r | B_c) = I_8$$

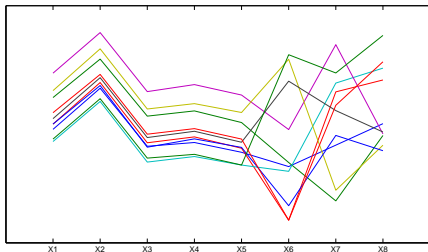


Linear Manifolds - Patterns in Subspaces

- Shift pattern that exists only in a subspace:

$$x_i = B_r(\mu_r + \mathbf{1}_r\phi_i) + B_c(\mu_c + \lambda_i) = B_r\mu_r + B_r\mathbf{1}_r\phi_i + B_c\mu_c + B_c\lambda_i$$

$$(B_r|B_c) = I_8$$

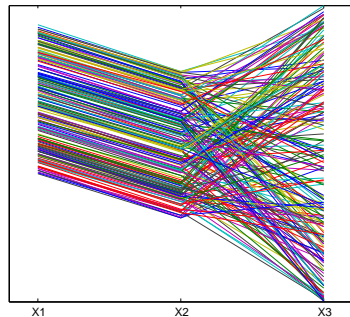
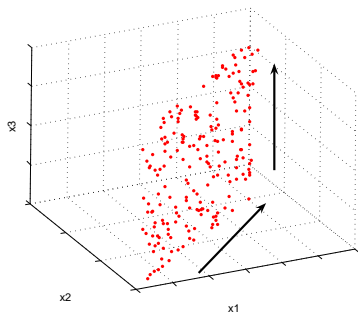


- The linear manifold embedding:

$$x_i = (B_r|B_c) \begin{pmatrix} \mu_r \\ \mu_c \end{pmatrix} + \left(B_r \frac{\mathbf{1}_r}{\sqrt{r}} | B_c \right) \begin{pmatrix} \sqrt{r}\phi_i \\ \lambda_i \end{pmatrix}$$



Linear Manifolds - Patterns in Subspaces



Linear Manifolds - Adding an Error Term

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, \dots, b_m an orthonormal set of vectors that span \mathbb{R}^d , (b_1, \dots, b_m) a matrix whose columns are the vectors b_1, \dots, b_m , and μ some point in \mathbb{R}^d . Then each $x_i \in C$ is modeled by,

$$x_i = \mu + \begin{pmatrix} \vdots & & \vdots \\ b_1 & \cdots & b_m \\ \vdots & & \vdots \end{pmatrix} \lambda_i + \begin{pmatrix} \vdots & & \vdots \\ b_{m+1} & \cdots & b_d \\ \vdots & & \vdots \end{pmatrix} \psi_i$$



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,

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



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,

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

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:

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



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,

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



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:

$$x_i = B_c \mu_c + \left(B_r \frac{\mu_r}{\|\mu_r\|} \mid B_c \right) \begin{pmatrix} \|\mu_r\| \phi_i + \frac{\mu_r'}{\|\mu_r\|} \psi_i \\ \lambda_i \end{pmatrix} + B_r \left(I_r - \frac{\mu_r \mu_r'}{\|\mu_r\|^2} \right) \psi_i$$



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

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



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:

$$x_i = \mathbf{1}\mu + \mathbf{1}\phi_i + \psi + \epsilon_i$$

where ϕ_i is a scalar denoting the residual effect of the i -th gene, $\psi = (\psi_1, \dots, \psi_d)'$ a vector containing the residual effects of the conditions, and $\epsilon_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$



The Bicluster Model (Cheng et al. 00)

Proposition

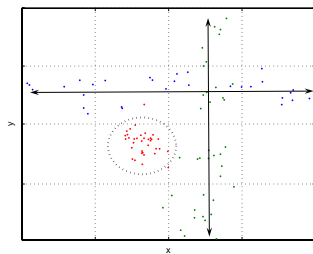
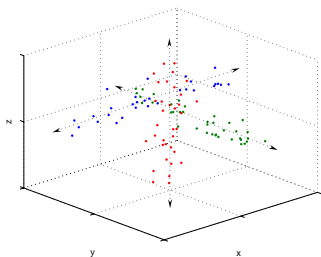
Every point x_i in a d -dim space that fits a bicluster model embedded in an r -dim subspace, also fits the linear manifold cluster model, where the dimension of the linear manifold is $d - r + 1$, and the model is given by:

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



Subspace Clusters

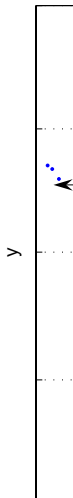
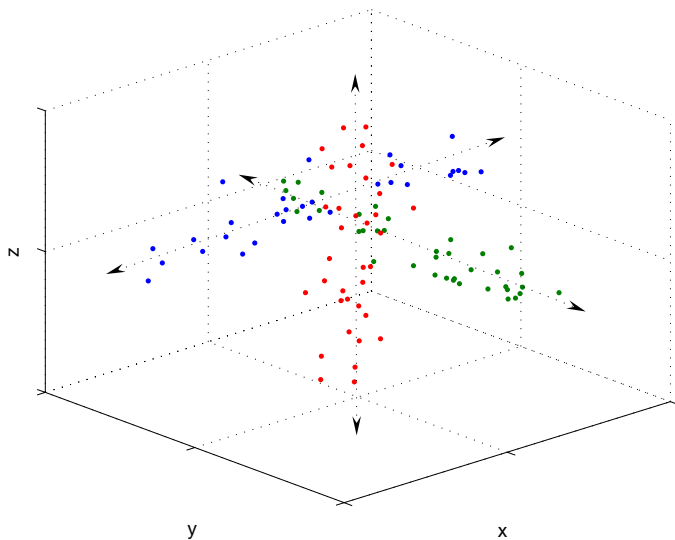
- Consist of a subset of points and a corresponding subset of attributes, such that these points form a dense region in a subspace defined by the set of corresponding attributes.



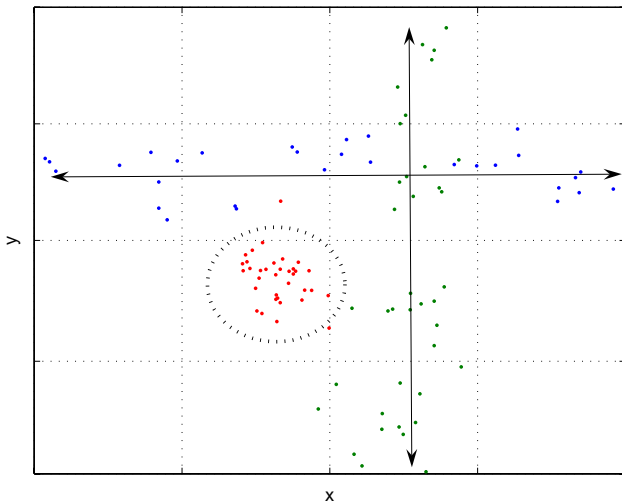
CLIQUE (Agrawal 98), **MAFIA** (Nagesh 99), **PROCLUS** (Aggarwal 99),
ORCLUS (Aggarwal 00)



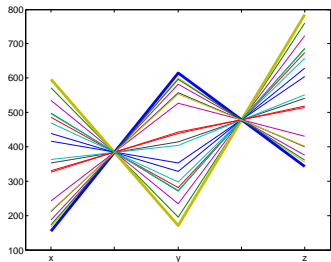
Subspace Clusters



Subspace Clusters



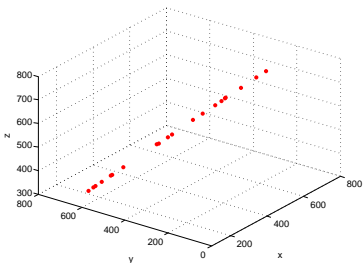
Other Instances of Linear Manifolds - Negative Correlations



$$R = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix}$$

$$\text{Pearson}R = 0.3181$$

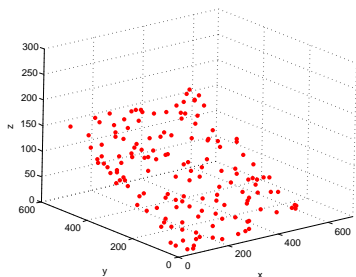
$$\text{MSR} = 18280$$



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



$$z = b_0 + b_1x + b_2y$$

$$\text{Pearson}R = 0.4509$$

$$\text{MSR} = 8975$$

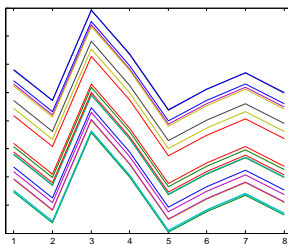
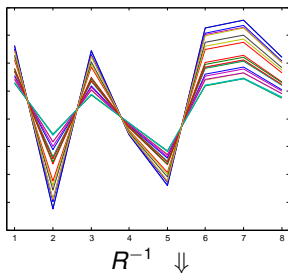
Coefficient of multiple determination:

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

4C, Böhm et al. (2004)



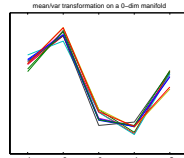
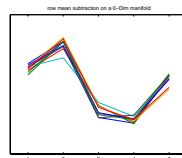
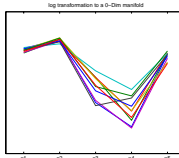
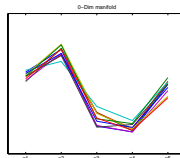
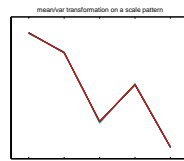
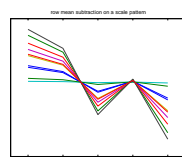
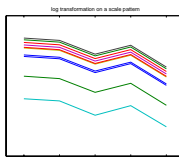
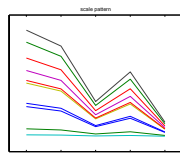
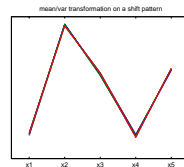
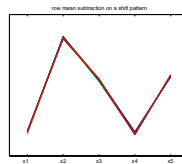
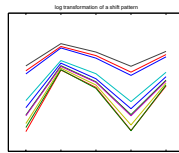
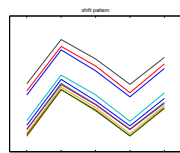
Other Instances of Linear Manifolds - Latent Variables



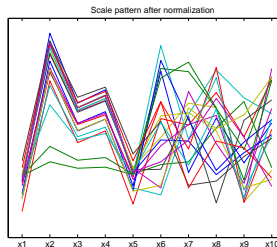
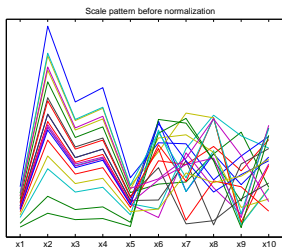
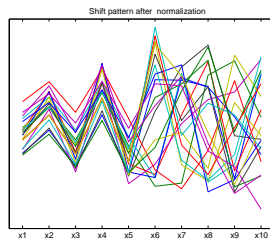
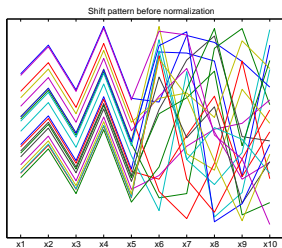
- 1 $x_i = R(\mu + \mathbf{1}_d \phi_i)$
- 2 $y_i = x_i - R\mu = \mathbf{1}_d \phi_i$
- 3 $\phi_i = \sqrt{\frac{x_i' x_i}{d}}$
- 4 $C = \frac{1}{n} \sum_{i=1}^n y_i (\mathbf{1}_d \phi_i)'$
- 5 $[u, s, v] = \text{svd}(C)$
- 6 $R = uv'$



Data Transformations



Data Transformations



The Algorithm

Main Idea



The Algorithm

Main Idea

- 1 Sample minimal subsets of points to construct trial linear manifolds of various dimensions.



The Algorithm

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.**



The Algorithm

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.



The Algorithm

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.**



The Algorithm

Main Idea

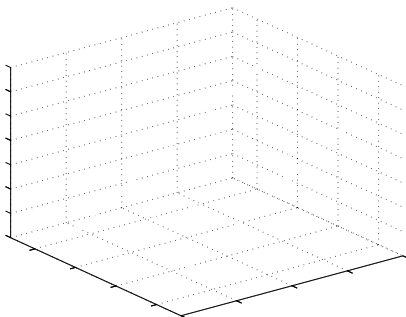
- 1 Sample minimal subsets of points to construct trial linear manifolds of various dimensions.
- 2 Compute distance histograms of the data to each trial manifold.
- 3 Of all the manifolds constructed, select the one whose associated histogram shows the best separation between a mode near zero and the 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.

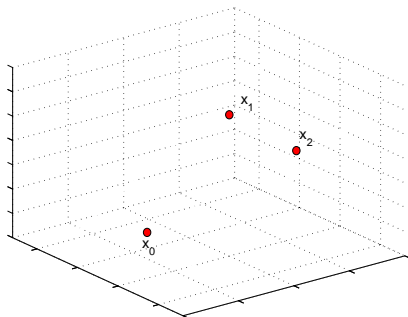
Example- constructing a 2D manifold



How are trial manifolds sampled?

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

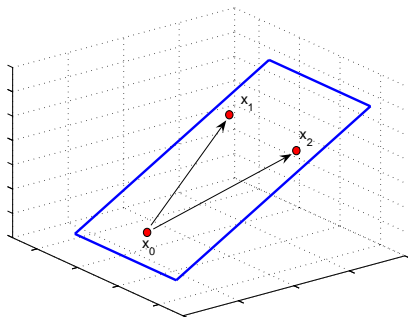
Example- constructing a 2D manifold



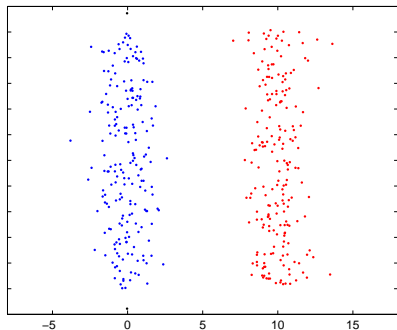
How are trial manifolds sampled?

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

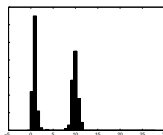
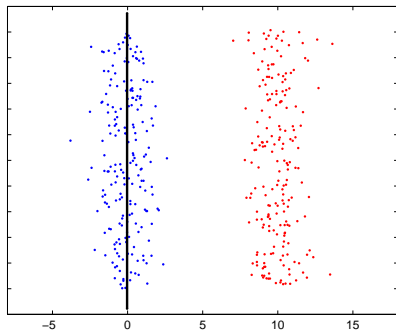
Example- constructing a 2D manifold



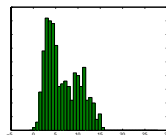
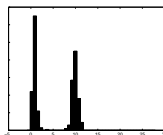
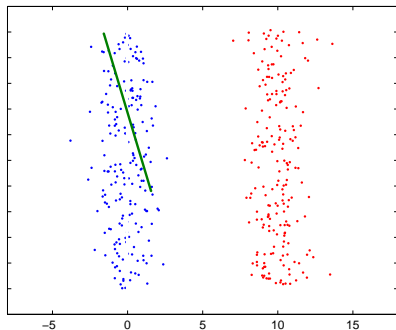
How many trial manifolds need to be examined?



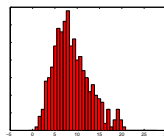
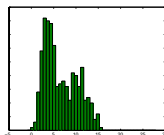
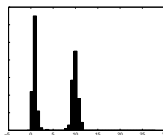
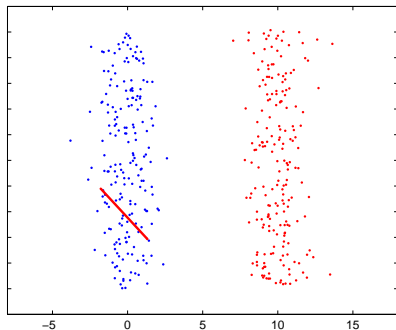
How many trial manifolds need to be examined?



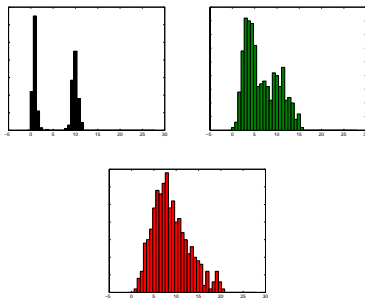
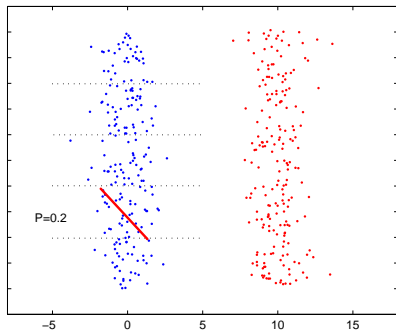
How many trial manifolds need to be examined?



How many trial manifolds need to be examined?



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.



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 $\left(\frac{1}{\hat{K}}\right)^m$.



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 $(1 - (1/\hat{K})^m)^n$



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.



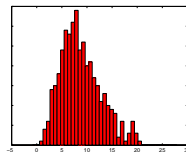
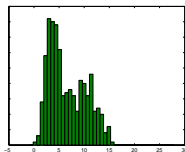
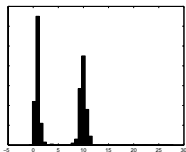
How many trial manifolds need to be examined?

- 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

$$n \geq \frac{\log \epsilon}{\log(1 - (1/\hat{K})^m)}$$



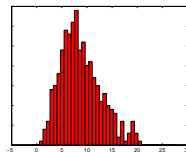
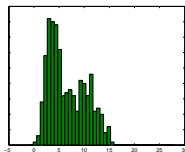
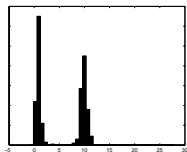
Selecting the best trial manifold/best separation



- To select the best manifold we first need to find the two classes or distributions involved.



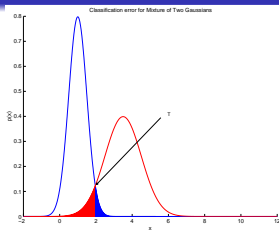
Selecting the best trial manifold/best separation



- To select the best manifold we first need to find the two classes or distributions involved.
- This problem is cast into histogram thresholding problem.



Kittler and Illingworth Minimum Error Thresholding (86)

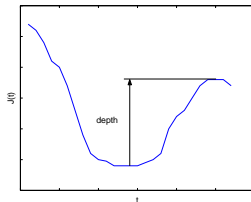
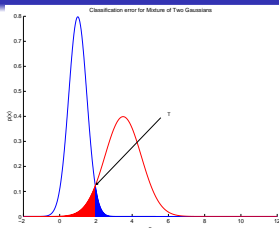


Minimize:

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



Kittler and Illingworth Minimum Error Thresholding (86)



Minimize:

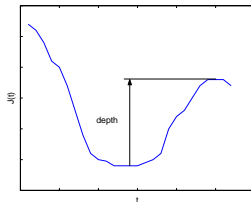
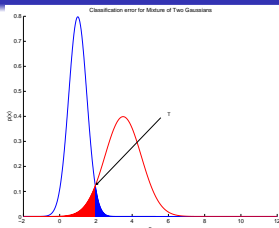
$$P(\text{error}) = \int_{x>T} p(x|c_1)P(c_1)dx + \int_{x\leq T} p(x|c_2)P(c_2)dx$$

KI86:

$$J(T) = 1 + 2(P_1(T) \log \sigma_1(T) + P_2(T) \log \sigma_2(T)) - 2(P_1(T) \log P_1(T) + P_2(T) \log P_2(T))$$



Kittler and Illingworth Minimum 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:

$$J(T) = 1 + 2(P_1(T) \log \sigma_1(T) + P_2(T) \log \sigma_2(T)) - 2(P_1(T) \log P_1(T) + P_2(T) \log P_2(T))$$

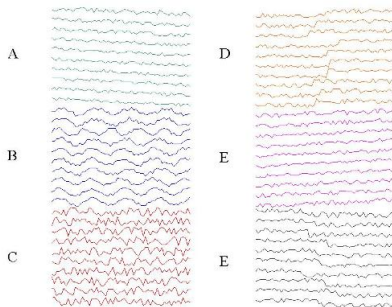
Goodness of separation:

$$\text{discriminability} = \frac{(\mu_1(T) - \mu_2(T))^2}{\sigma_1^2(T) + \sigma_2^2(T)} \quad \times \quad \text{depth} = J(T') - J(T)$$



Time Series Clustering (UCI KDD Archive)

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



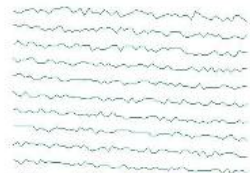
	in1	in2	in3	in4	in5	in6	total
out1	0	0	0	57	0	0	57
out2	0	0	80	0	1	0	81
out3	0	0	0	43	0	99	142
out4	0	0	20	0	98	0	118
out5	99	0	0	0	0	0	99
out6	0	41	0	0	0	0	41
out7	0	23	0	0	0	0	23
out8	1	36	0	0	1	1	39
total	100	100	100	100	100	100	600

Total Correct=533 Accuracy=88.8333

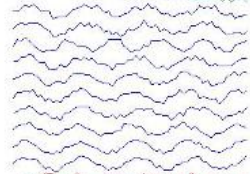


Time Series Clustering (UCI KDD Archive)

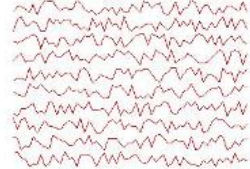
A



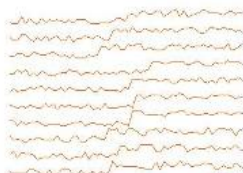
B



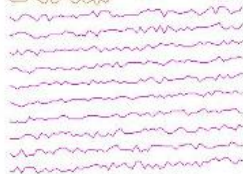
C



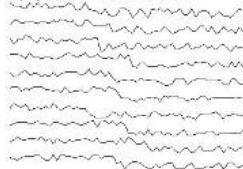
D



E



E



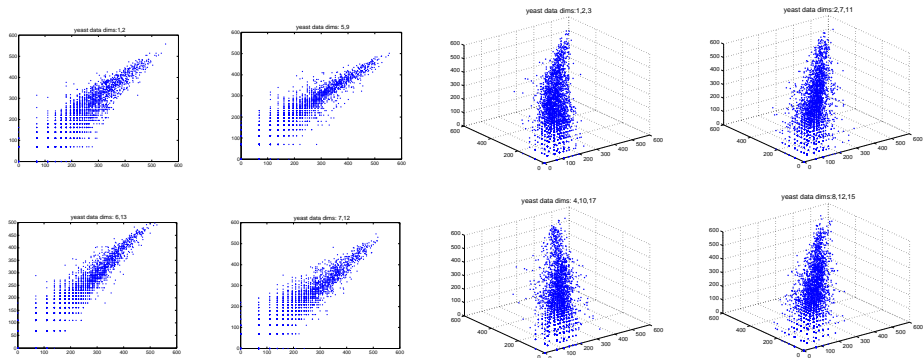
	in1
out1	0
out2	0
out3	0
out4	0
out5	99
out6	0
out7	0
out8	1
total	100



Total C

Yeast Data - mitotic cell cycle 2884×17

(Cho 1998, Tavazoie 1999, <http://arep.med.harvard.edu/biclustering/>)

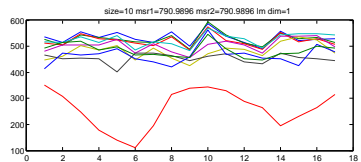
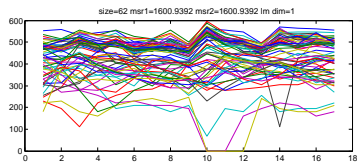
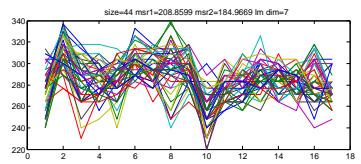
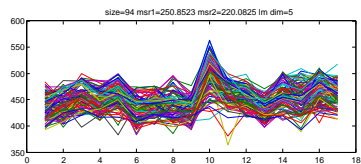
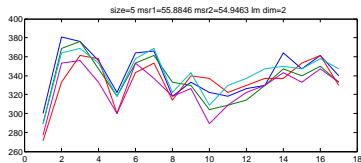
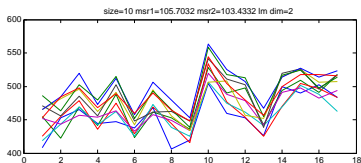


$PC1 = (0.82, 0.95, 1.02, 0.95, 1.02, 0.93, 0.99, 0.97, 0.92, 1.17, 1.05, 1.02, 0.92, 1.04, 1.03, 1.09, 1.04)'$

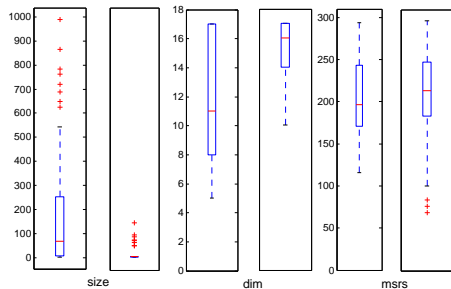
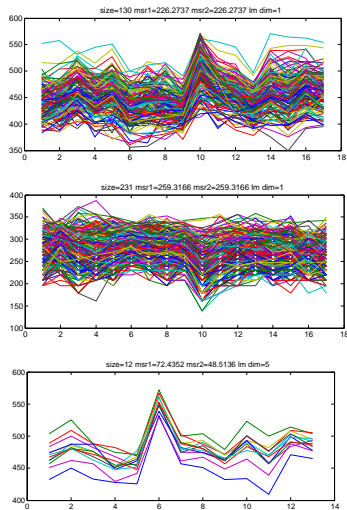
93%



Yeast Data Results (regular manifolds)



Yeast Data Results (MSR manifolds)



Biclustering/Linear Manifold Clustering

