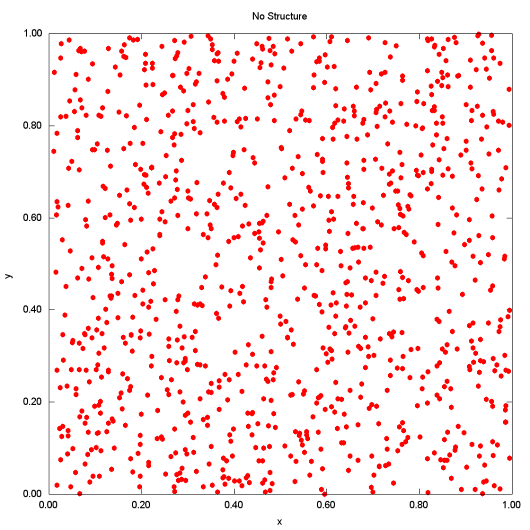
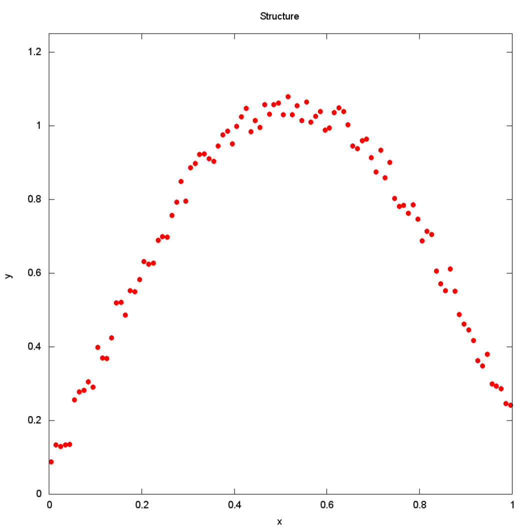
What is Structure?

- Structure is a description of the dependencies
- Dependencies mean constraints
- Un-structured means no constraints
- Constraint means subset

No Structure



Structure



Description of Structure

- Language by which the structure can be described
- Observed data is a sampled perturbed ideal
- Description is inexact
  - Closeness of the description to the observed data
  - Length of the description

Truth and Lies

- Truth
  - Language is able to describe some of the underlying data structure
- Lies
  - What the language cannot describe is a lie by omission
  - Description is an estimate
  - Estimated structures have a random component
  - The difference between the true underlying structure and the estimated structure

Linear Regression Language

- Data: $x_1, \ldots, x_K$
- Dimension: $x_k = (x_k^1, \ldots, x_k^N) \in R^N$
- Dependency: $x_k^N = \sum_{n=1}^{N-1} \alpha_n x_k^n$
- Error: $\epsilon_k^2 = \left| x_K^N - \sum_{n=1}^{N-1} \alpha_n x_k^n \right|^2$
- Assumption: All points arise from the same process
    - All observations have the same dependency

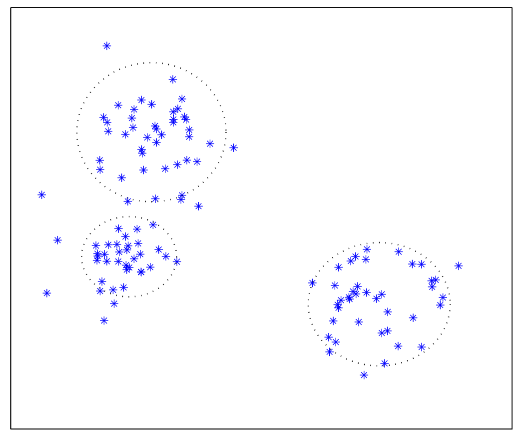
Example

- Population
  - Healthy
  - Illnesses $A_1, \dots, A_K$
- It is not known how many illnesses there are
- Each person is measured with N lab tests
- The structure of the data is the inter-relationship(s) between the values of the lab tests
- Linear Regression is the wrong Language

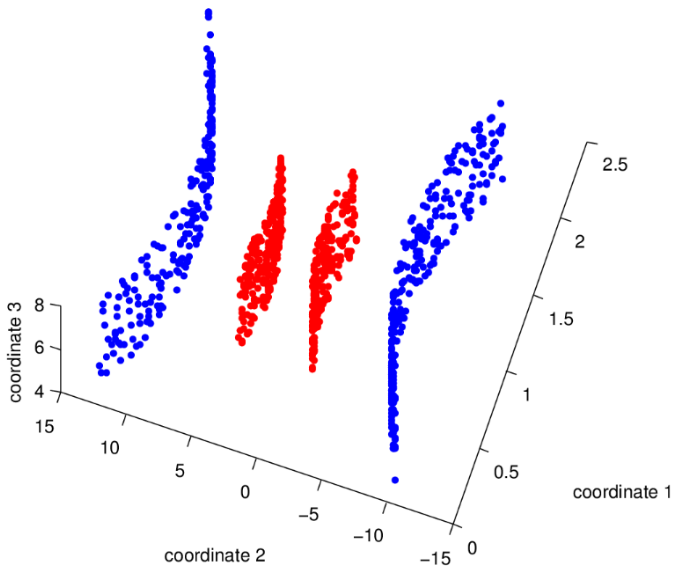
Example Test Report

Alkaline Phosphatase	58	IU/L	25-150
Bilirubin Total	0.4	mg/dL	0.0-1.2
A/G Ratio	1.7		1.1-2.5
Globulin Total	2.5	g/dL	1.5-4.5
Albumin, Serum	4.3	g/dL	3.5-5.5
Protein, Total Serum	6.8	g/dL	6.0-8.5
Phosphorus, Serum	3.6	mg/dL	2.5-4.5
Calcium, Serum	9.3	mg/dL	8.7-10.2
Carbon Dioxide, Total	21	mmol/L	20-32
Chloride, Serum	105	mmol/L	97-108
Potassium, Serum	4.1	mmol/L	3.5-5.2
Sodium, Serum	140	mmol/L	134-144
BUN/Creatinine Ratio	19		9-20
eGFR If Africn Am	126	mL/min/1.73	>59
eGFR If NonAfricn Am	109	mL/min/1.73	>59
Creatinine, Serum	0.81	mg/dL	0.76-1.27
BUN	15	mg/dL	6-24

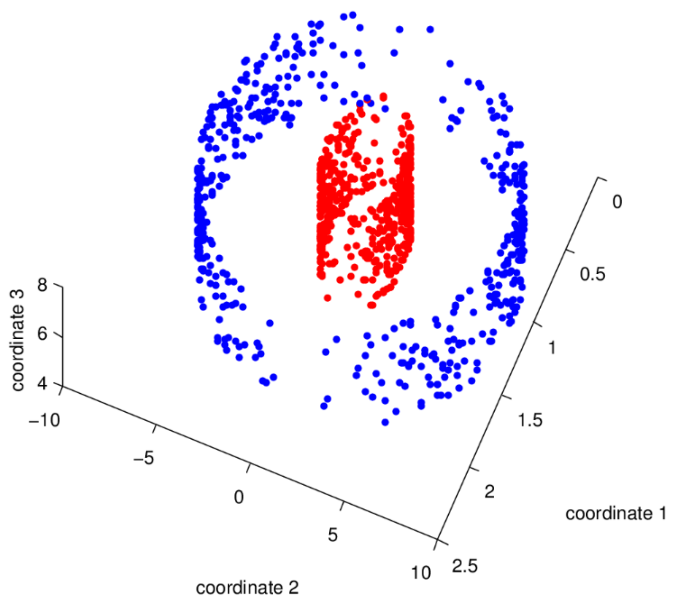
Point Clusters



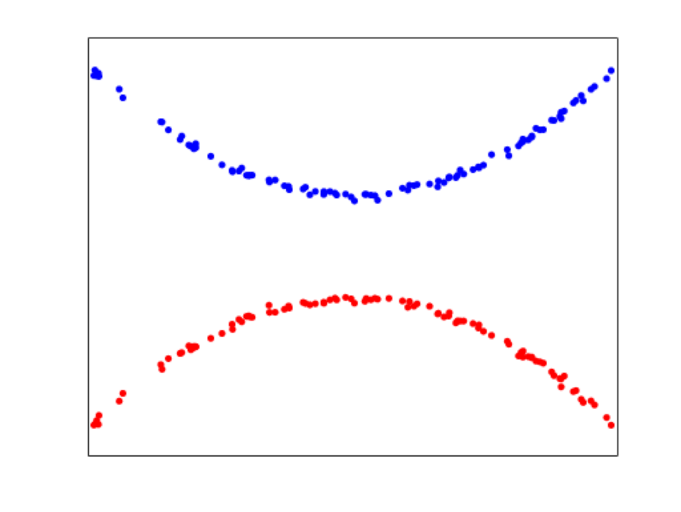
Hyperbolic Clusters



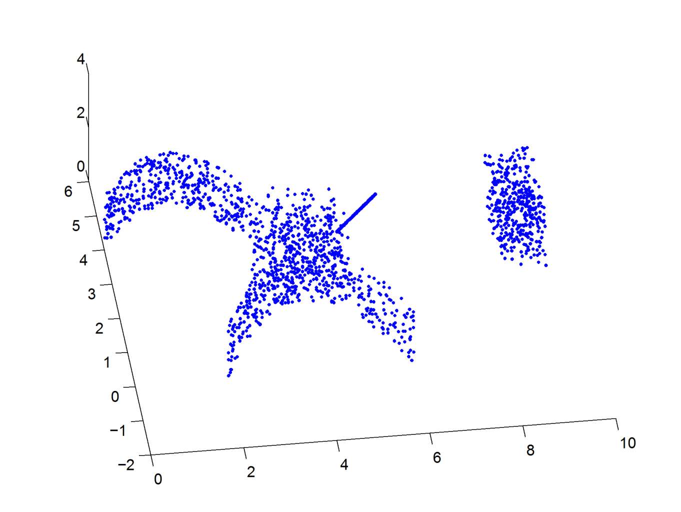
Elliptic Clusters



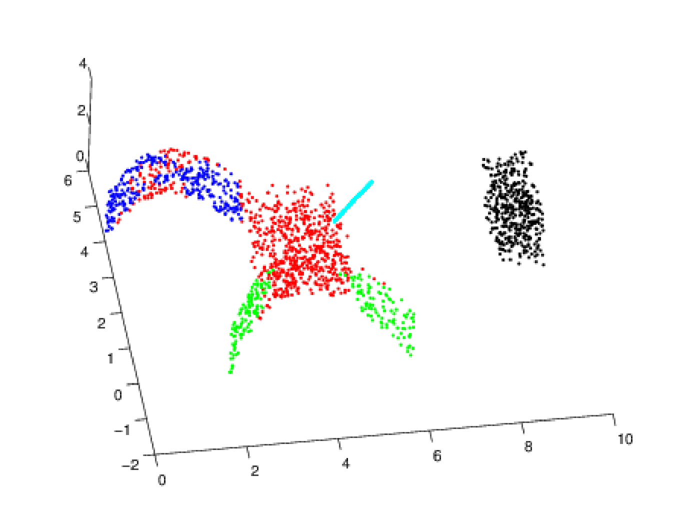
Manifold Clusters



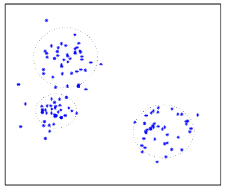
Manifold Clusters



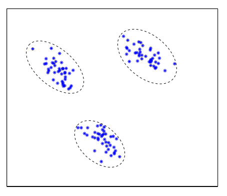
Manifold Clusters



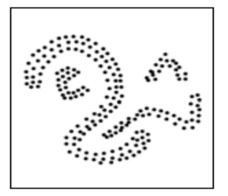
Cluster Models



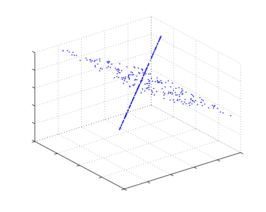
hyper-spherical



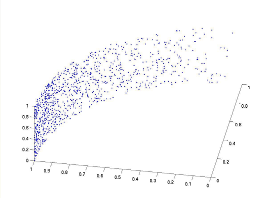
hyper-ellipsoidal



arbitrary shaped



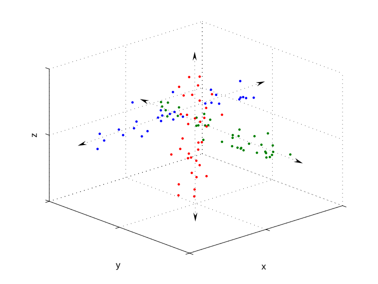
linear

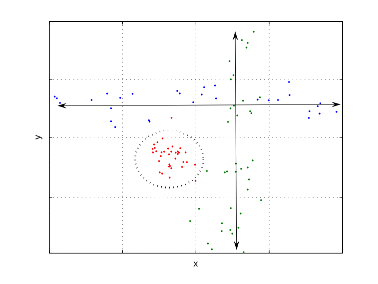


nonlinear

Subspace Clusters

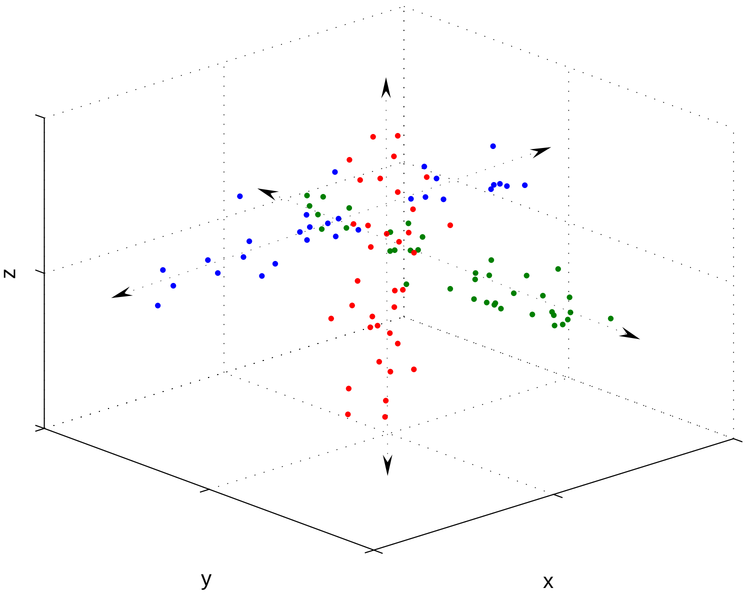
- Consists of a subset of points and a corresponding subset of variables, such that these points form a dense region in a subspace defined by the set of corresponding variables.



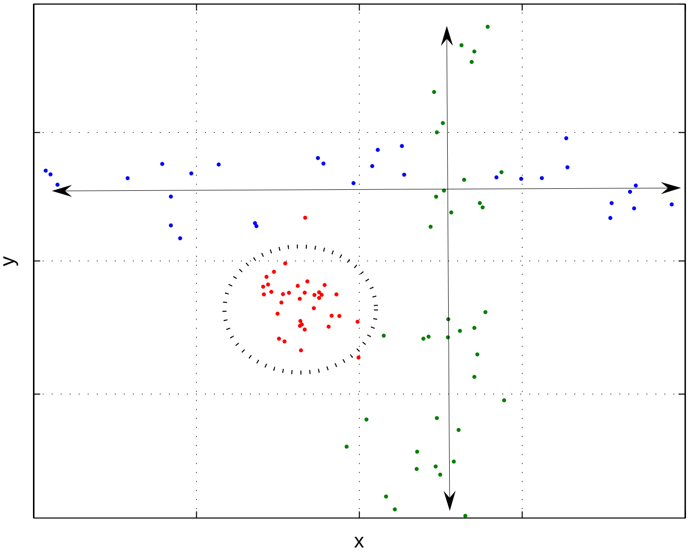


CLIQUE (Agrawal 98), MAFIA (Nagesh 99),PROCLUS (Aggarwal 99)

Subspace Clusters

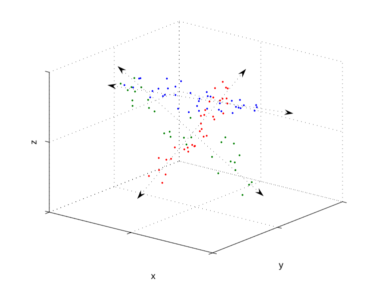


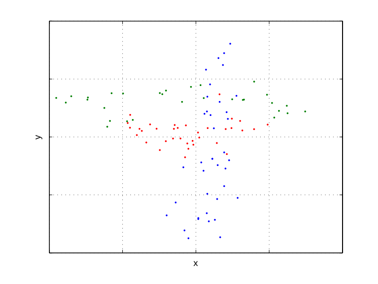
Subspace Clusters



Arbitrary Oriented Subspace Clusters

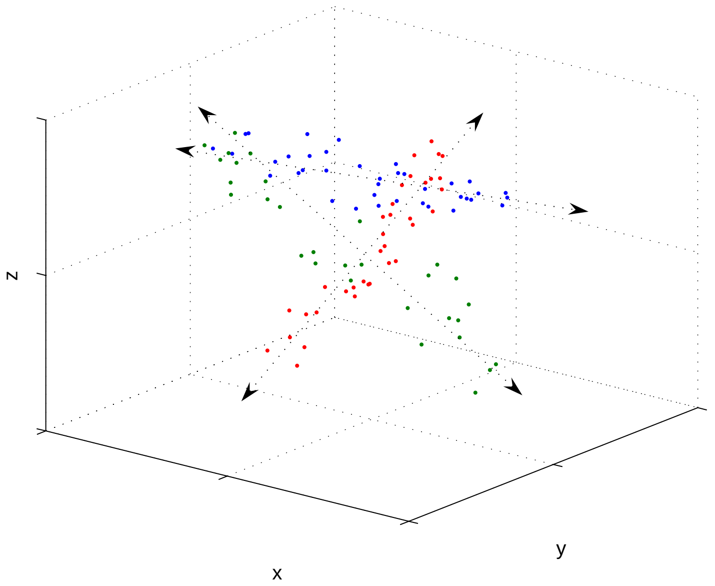
- Consists of a subset of points and a corresponding linear combination of a subset of variables, such that these points form a dense region in a subspace defined by the set of corresponding linear combinations of variables.



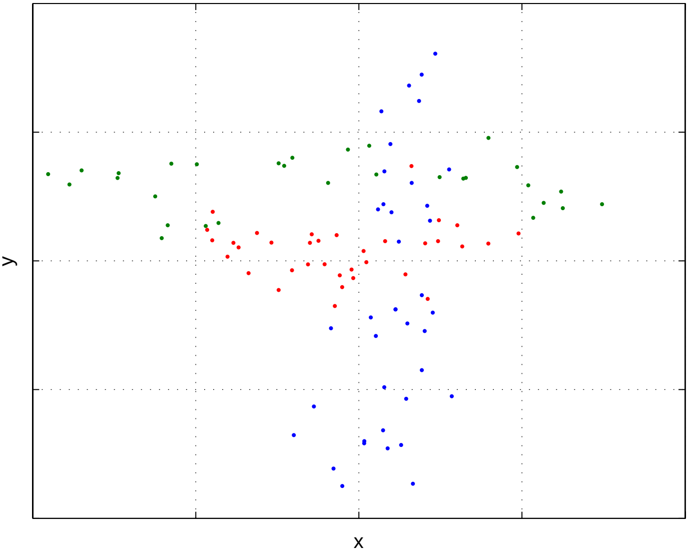


ORCLUS (Aggarwal 00)

Arbitrary Oriented Subspace Clusters

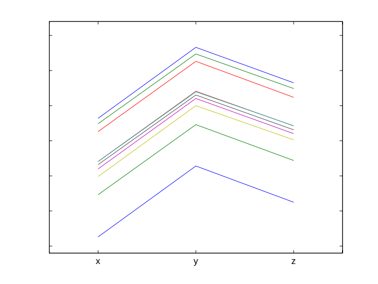


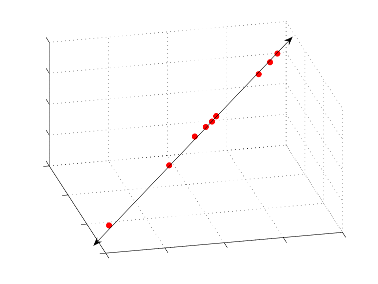
Arbitrary Oriented Subspace Clusters



Pattern (Correlation) Clusters

- Consists of as a subset of objects and variables for which the participating objects show a similar trend rather than being close to each other.





Bicluster (Cheng 00),Floc (Yang 02),pCluster (Wang 02)

Linear Manifold Clusters

Definition

L is a linear manifold of vector space V if and only if for some subspace S of V and translation $t \in V$,

$$L = \{x \in V | \textit{for some } s \in S, x = t + s\}$$

The dimension of L is the dimension of S, and if the dimension of L is one less than the dimension of V then L is called a hyperplane.

Linear Manifold Clusters

Definition

L is a linear manifold of vector space V if and only if for some subspace S of V and translation $t \in V$,

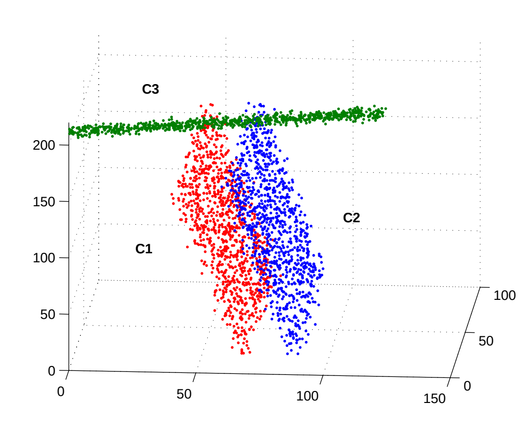
$$L = \{x \in V | \textit{for some } s \in S, x = t + s\}$$

The dimension of L is the dimension of S, and if the dimension of L is one less than the dimension of V then L is called a hyperplane.

A linear manifold is, in other words, a subspace that may have been shifted away from the origin.

A subspace is a linear manifold that contains the origin.

Dense Linear Manifold Clusters



The Linear Manifold Cluster Model

The cluster model has the following properties:

- The points in each cluster lie close to a low dimensional linear manifold.

The Linear Manifold Cluster Model

The cluster model has the following properties:

- The points in each cluster lie close to a low dimensional linear manifold.
- The intrinsic dimensionality of the cluster is the dimensionality of the linear manifold.

The Linear Manifold Cluster Model

The cluster model has the following properties:

- The points in each cluster lie close to a low dimensional linear manifold.
- The intrinsic dimensionality of the cluster is the dimensionality of the linear manifold.
- The manifold is arbitrarily oriented.

The Linear Manifold Cluster Model

The cluster model has the following properties:

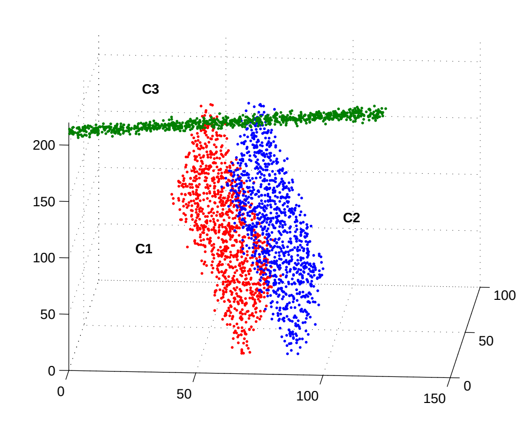
- The points in each cluster lie close to a low dimensional linear manifold.
- The intrinsic dimensionality of the cluster is the dimensionality of the linear manifold.
- The manifold is arbitrarily oriented.
- The points in the cluster induce a correlation among two or more attributes (or linear combinations of attributes) of the data.

The Linear Manifold Cluster Model

The cluster model has the following properties:

- The points in each cluster lie close to a low dimensional linear manifold.
- The intrinsic dimensionality of the cluster is the dimensionality of the linear manifold.
- The manifold is arbitrarily oriented.
- The points in the cluster induce a correlation among two or more attributes (or linear combinations of attributes) of the data.
- In the orthogonal complement space to the manifold the points form a compact densely populated region, which can be used to cluster the data.

Dense Linear Manifold Clusters



The Linear Manifold Cluster Model

Comment

Classical clustering algorithms such as K-means assume that each cluster is associated with a zero dimensional manifold (the center) and therefore omit the possibility that a cluster may have non-zero dimensional linear manifold associated with it.

The Linear Manifold Cluster Model

Definition

- Let D be a set of N-dimensional points in $\mathbb{R}^N$
- $C \subseteq D$ a subset of points that belong to a cluster
- x some point in C
- $b_1, \ldots, b_N$ an orthonormal set of vectors that span $\mathbb{R}^N$
- $(b_i, \ldots, b_j)$ a matrix whose columns are the vectors $b_i, \ldots, b_j$
- μ some point in $\mathbb{R}^N$

Then each $x \in C$ can be modeled by,

$$x = \mu + (b_1, \ldots, b_m)\lambda + (b_{m+1}, \ldots, b_N)\psi$$

The Linear Manifold Cluster Model

$$\begin{aligned} x &= \mu + (b_1, \ldots, b_m)\lambda^{m\times 1} + (b_{m+1}, \ldots, b_N)\psi^{N-m\times 1} \\ x &= \mu + B^{N\times m}\lambda^{m\times 1} + B_c^{N\times N-m}\psi^{N-m\times 1} \end{aligned}$$

- The idea is that each point in a cluster lies close to a m-dimensional linear manifold, defined by $\mu + span\{b_1, \ldots, b_m\}$.
- $\lambda^{m\times 1}$ models the spread of the points in the manifold
  - Each entry of the $m \times 1$ random vector λ is i.i.d. $U(-R/2, +R/2)$
  - In the manifold points are uniformly distributed in each direction

The Linear Manifold Cluster Model

$$
\begin{aligned}
x &= \mu + (b_1, \ldots, b_m)\lambda^{m\times 1} + (b_{m+1}, \ldots, b_N)\psi^{N-m\times 1} \\
x &= \mu + B^{N\times m}\lambda^{m\times 1} + B_c^{N\times N-m}\psi^{N-m\times 1}
\end{aligned}
$$

- $\psi^{N-m\times 1}$ a small perturbation associated with each point in the cluster. The idea is that each point may be perturbed in directions that are orthogonal to the manifold, i.e., the vectors $b_{m+1}, \ldots, b_N$.
- This is modeled by requiring that the $(N-m)\times 1$ random vector $\psi \sim N(\mathbf{0}, \Sigma)$, where the largest eigenvalue of Σ is much smaller than R.
- Since the variance along each of these directions is much smaller than the range R of the embedding, the points are likely to form a compact and densely populated region.

The Algorithm

Main Idea

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

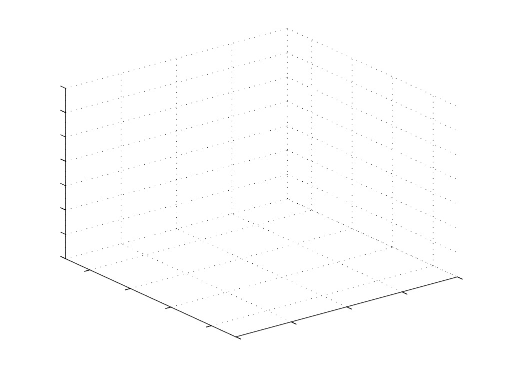
Main Idea

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.

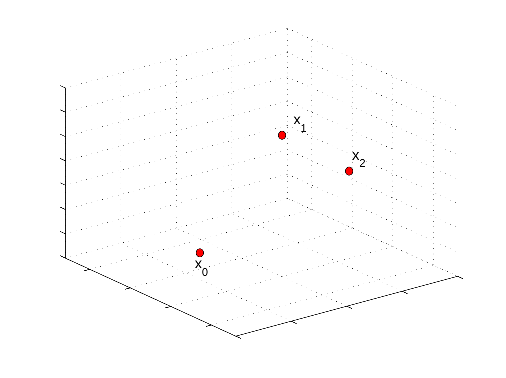
Example- constructing a 2D manifold



How are trial manifolds sampled?

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

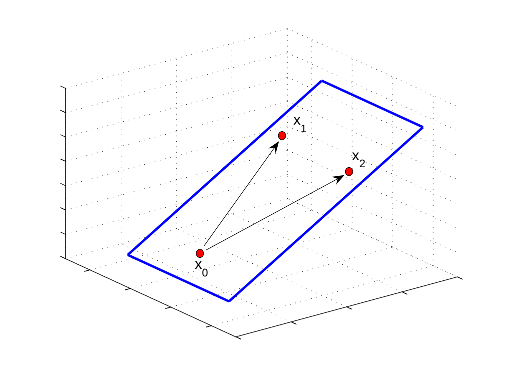
Example- constructing a 2D manifold



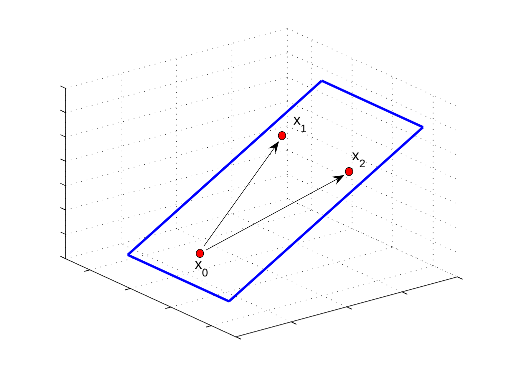
How are trial manifolds sampled?

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

Example- constructing a 2D manifold

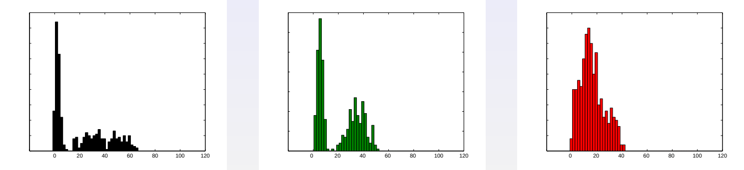


How are the trial manifolds sampled?

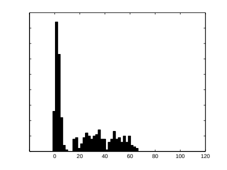


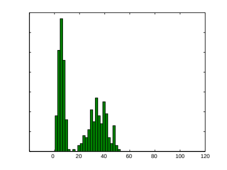
$$
\begin{aligned}
\mu &= x_0 \\
\hat{B} &= (\hat{b}_1, \hat{b}_2) = (x_1 - x_0, x_2 - x_0) \\
B &= \textit{orthonormal } \hat{B} \\
dist(x) &= \|(I - BB')(x - \mu)\|
\end{aligned}
$$

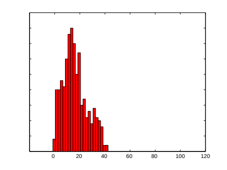
Selecting the best trial manifold/best separation



Selecting the best trial manifold/best separation

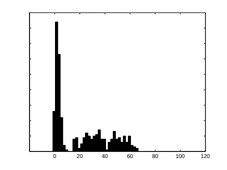


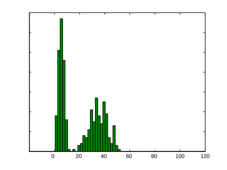


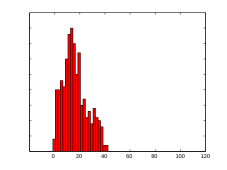


- To compute a separation score we first need to find the two classes or distributions involved.

Selecting the best trial manifold/best separation

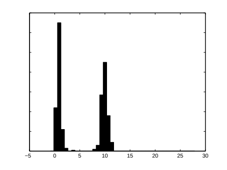


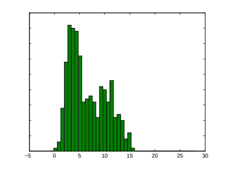


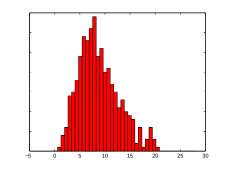


- To compute a separation score we first need to find the two classes or distributions involved.
- This problem is cast into histogram thresholding problem.

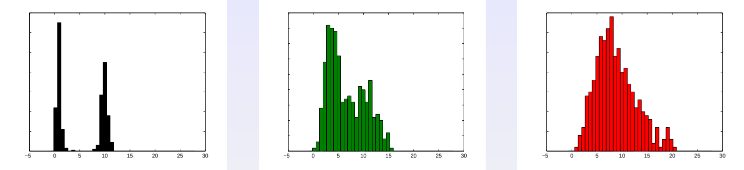
Selecting the best trial manifold

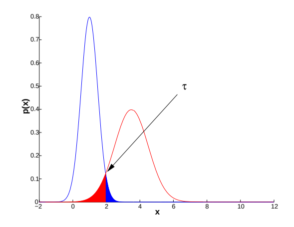






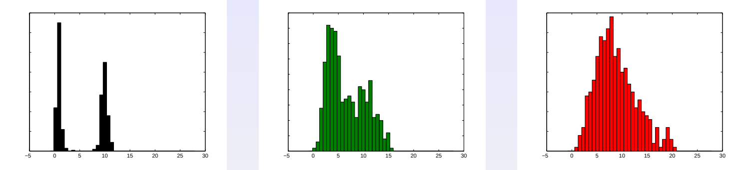
Selecting the best trial manifold

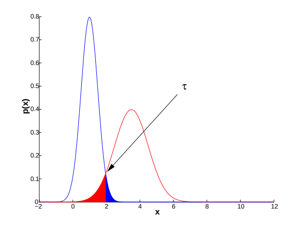




$$\tau = \arg\min_t P(error|t) = \int_{x>t} p(x|C_1)P(C_1)dx + \int_{x\leq t} p(x|C_2)P(C_2)dx$$

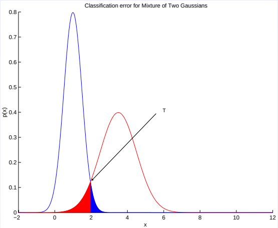
Selecting the best trial manifold





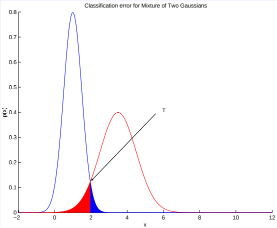
$$\tau = \arg\min_t P(error|t) = \int_{x>t} p(x|C_1)P(C_1)dx + \int_{x\leq t} p(x|C_2)P(C_2)dx$$

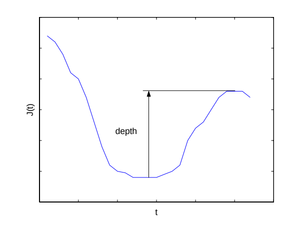
$$\text{Goodness of separation} = \frac{(\mu_1(\tau) - \mu_2(\tau))^2}{\sigma_1^2(\tau) + \sigma_2^2(\tau)}$$



Minimize:

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

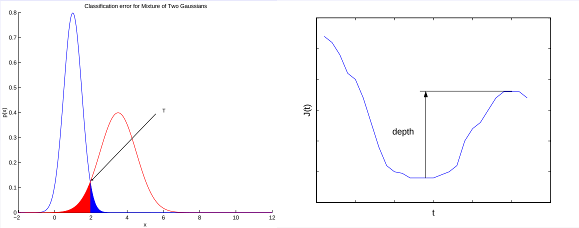




Minimize:

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

Kittler and Illingworth Minimum Error Thresholding (86)

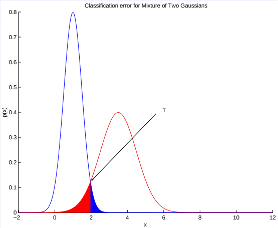


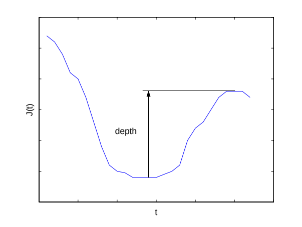
Minimize:

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

KI86:

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





Minimize:

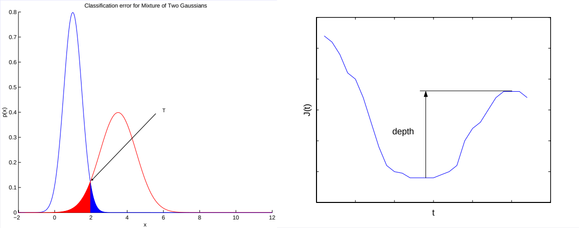
$$P(error) = \int_{x>T} p(x|c_1)P(c_1)dx + \int_{x \le T} p(x|c_2)P(c_2)dx$$

KI86:

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

$$Depth = J(T') - J(T)$$

Kittler and Illingworth Minimum Error Thresholding (86)



Minimize:

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

KI86:

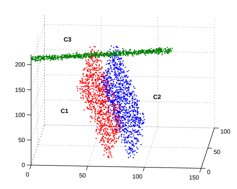
$J(T) = 1 + 2\,(P_1(T)\log\sigma_1(T) + P_2(T)\log\sigma_2(T)) - 2\,(P_1(T)\log P_1(T) + P_2(T)\log P_2(T))$

$Depth = J(T') - J(T)$

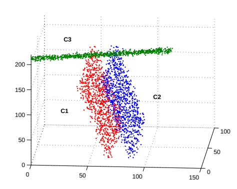
Goodness of separation:

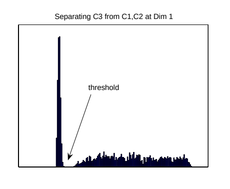
$$Discriminability = \frac{(\mu_1(T) - \mu_2(T))^2}{\sigma_1^2(T) + \sigma_2^2(T)} \quad \times \quad Depth$$

A Run of the Algorithm

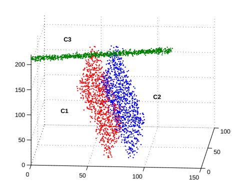


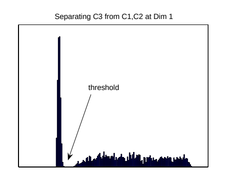
A Run of the Algorithm

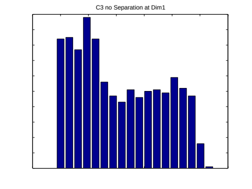




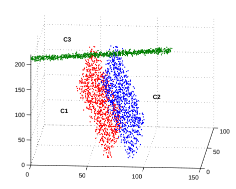
A Run of the Algorithm

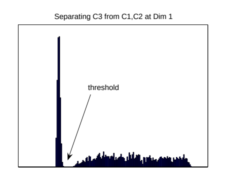


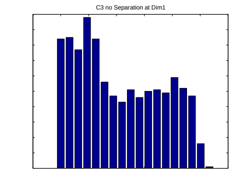


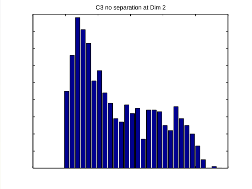


A Run of the Algorithm

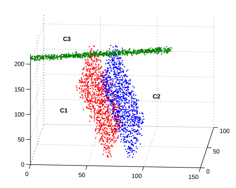


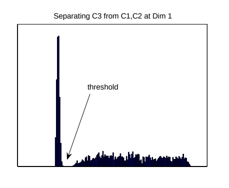


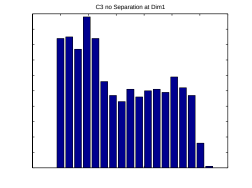


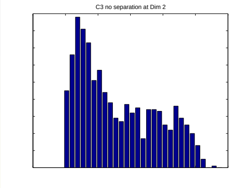


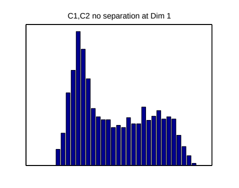
A Run of the Algorithm



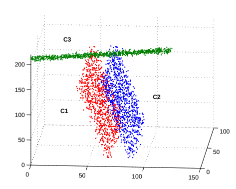


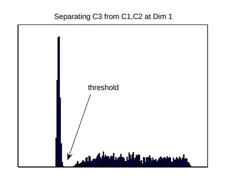


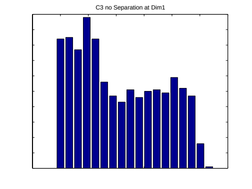


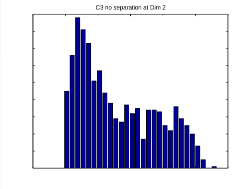


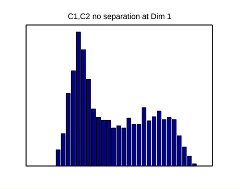
A Run of the Algorithm

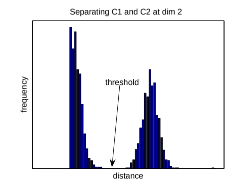




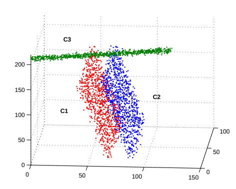


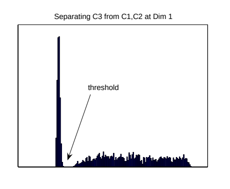


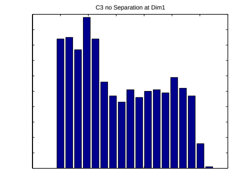


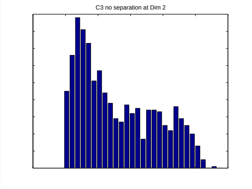


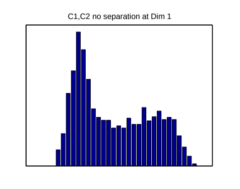
A Run of the Algorithm

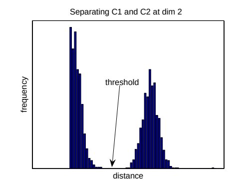


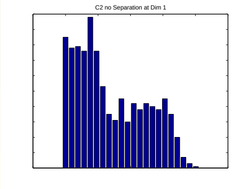




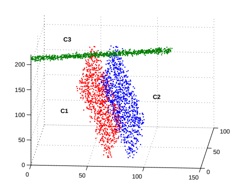


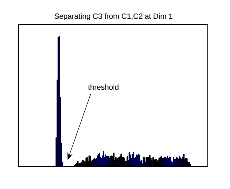


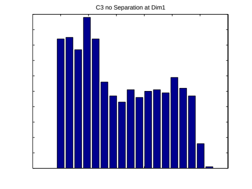


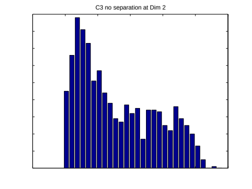


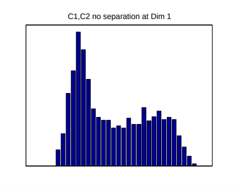
A Run of the Algorithm

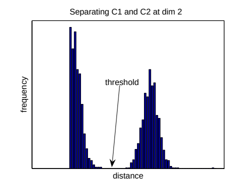


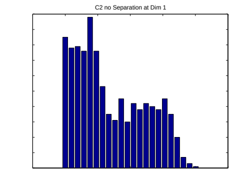


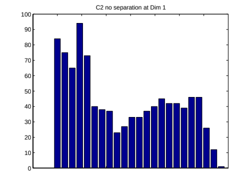




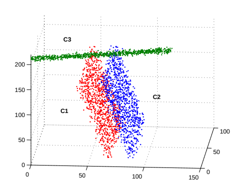


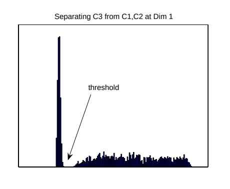


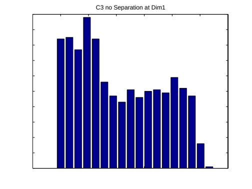


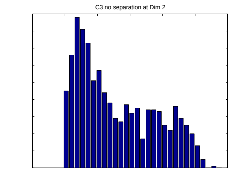


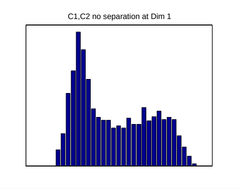
A Run of the Algorithm

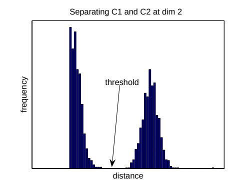


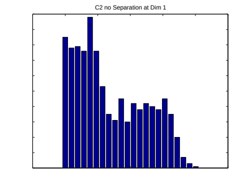


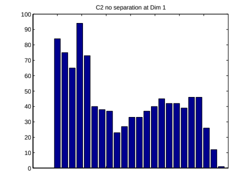


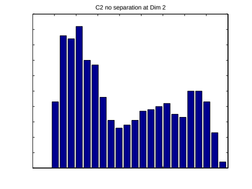












Empirical Evaluation

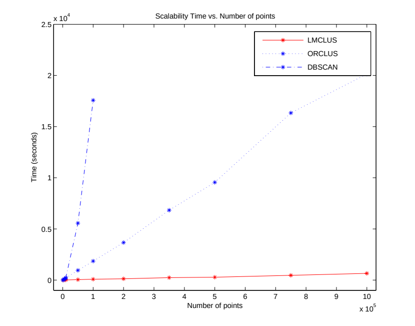
- Evaluation Criteria:
  - Accuracy
  - Efficiency and Scalability
  - Stability
- Data Sets:
  - Synthetic
  - Real
- Algorithms Compared Against:
  - DBSCAN $O(N^2 d)$ (Ester at el. 96)
  - ORCLUS $O(K^3 + kNd + K^2 d^3)$ (Aggarwal 00)
  - HPCluster (Haralick at el. 04)

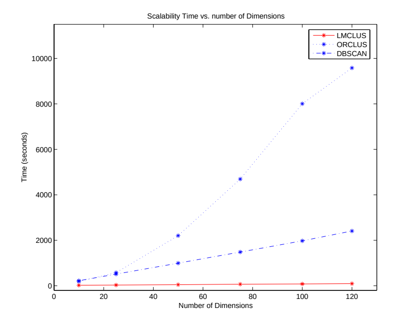
Empirical Evaluation- Accuracy

	size	clusters	dim	LM dim	LMCLUS	ORCLUS	DBSCAN	HPCluster
D_1	3000	3	4	2-3	95% / 0:0:08	80% / 0:0:22	34.6% / 0:0:9	72% / 0:0:51
D_2	3000	3	20	13-17	98.4% / 0:0:33	58.8% / 0:2:18	65.5% / 0:0:36	97.4% / 0:1:39
D_3	30000	4	30	1-4	100% / 0:15:38	64.9% / 1:5:30	100% / 1:31:52	99.3% / 0:1:32
D_4	6000	3	30	4-12	99.9% / 0:9:22	98.3% / 0:8:20	66.5% / 0:3:49	97.1% / 0:0:12
D_5	4000	3	100	2-3	100% / 0:0:20	87.9% / 0:54:30	65.3% / 0:5:24	99% / 0:3:54
D_6	90000	3	10	1-2	99.99% / 0:0:29	100% / 0:29:02	66.7% / 4:58:49	100% / 0:1:23
D_7	5000	4	10	2-6	99.24% / 0:2:05	99.3% / 0:2:41	74.1% / 0:0:54	96% / 0:0:35
D_8	10000	5	50	1-4	99.9% / 0:1:42	63.64% / 1:33:52	100% / 0:17:00	99.2% / 0:3:43
D_9	80000	8	30	2-7	99.9% / 3:12:46	96.9% / 13:30:30	100% / 10:51:15	99.9% / 0:4:57
D_{10}	5000	5	3	1-2	86.5% / 0:0:48	68.2% / 0:0:45	59.6% / 0:0:5	78% / 0:0:33
$*D_{11}$	1500	3	3	1	98.5% / 0:0:01	99.6% / 0:0:10	42.6% / 0:0:02	33.3% / 0:0:52
$*D_{12}$	1500	3	3	2	97% / 0:0:02	99% / 0:0:11	33.8% / 0:0:02	33.3% / 0:0:26
$*D_{13}$	1500	3	7	3	97.7% / 0:0:05	99.1% / 0:0:17	33.9% / 0:0:04	33.3% / 0:0:34
$*D_{14}$	5000	5	20	4	99.9% / 0:5:46	100% / 0:10:42	21.1% / 0:1:39	20% / 0:1:30
$*D_{15}$	4000	4	50	3	99% / 0:9:14	100% / 0:25:52	25% / 0:2:34	25% / 0:3:20

Empirical Evaluation- Efficiency and Scalability

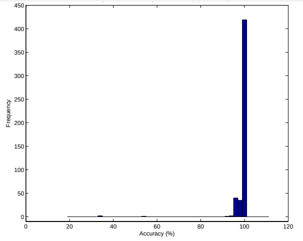
$$O(N^2 K^2 L^3 d)$$

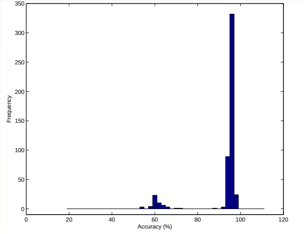




Empirical Evaluation- Stability

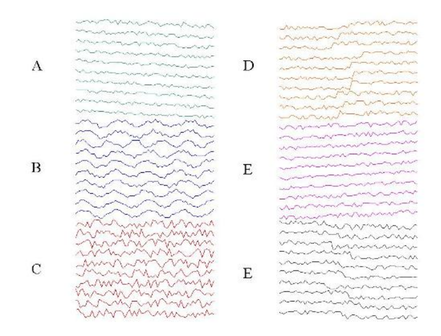
		LMCLUS	ORCLUS
1st data set	mean	99.1	92.1
	median	99.9	95.5
	std	4.7	10.56
2nd data set	mean	97.36	99.26
	median	97.4	99.47
	std	0.0053	0.0049





Time Series Clustering (UCI KDD Archive)

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

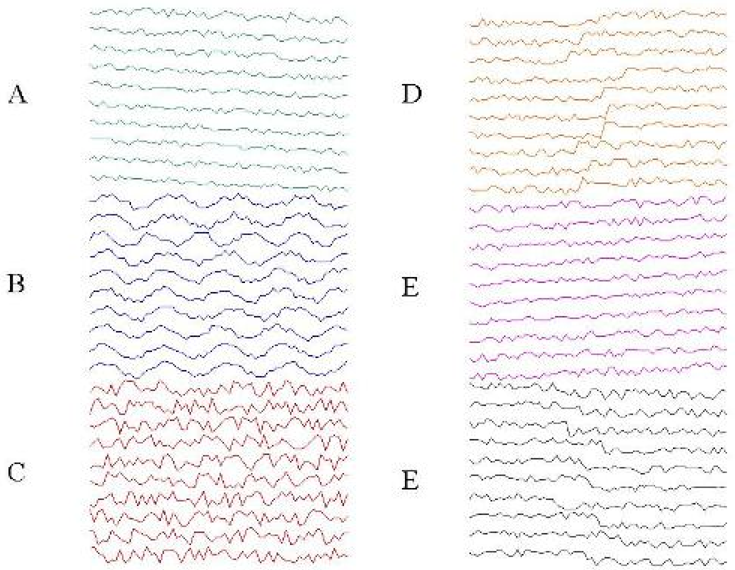


	in1	in2	in3	in4	in5	in6	total
out1	0	0	0	**57**	0	0	57
out2	0	0	**80**	0	1	0	81
out3	0	0	0	43	0	**99**	142
out4	0	0	20	0	**98**	0	118
out5	**99**	0	0	0	0	0	99
out6	0	**41**	0	0	0	0	41
out7	0	**23**	0	0	0	0	23
out8	1	**36**	0	0	1	1	39
total	100	100	100	100	100	100	600

Total Correct=533 Accuracy=88.8333

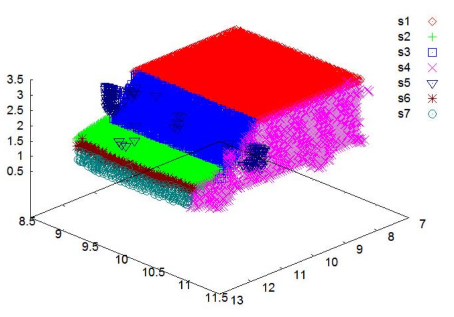
model: $x = \mu + \mathbf{1}\phi + \psi$

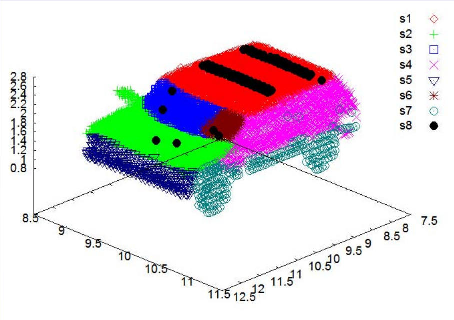
Time Series Clustering (UCI KDD Archive)



out
out
out
out
out
out
out
out
tota

E3D Point Cloud Segmentation (ALPHATECH Inc.)





When to Stop

Clustering Techniques Have A Variety of Ways

- Specify Number of Clusters
- Specify Minimum Cluster Size
- Specify a Minimum Quality Score for a Cluster

Minimum Description Length

- Clustering details the Structure of the data
- The Structure of the data should be more compact that a list of coordinates of each data point
- Good Clustering
  - The Description Length needed for describing the structure of the data is less

Manifold Cluster Description Length

- Description Length for Manifold
- Description Length for points projected to the manifold
- Error Toleration Parameter
- Description Length for point perturbation off the manifold
  - To within Error Tolerance

Manifold Description Length

- 1: Dimension K of Cluster
- N: Offset vector from origin
- Orthonormal Manifold basis set
  - Basis Vectors KN
  - Norm 1: K constraints
  - Orthogonality: $\frac{K(K-1)}{2}$ constraints
  - $KN - \frac{K(K+1)}{2}$ numbers
  - Each number B bits
- Total: $B[1 + N + K(N - \frac{K+1}{2})]$

Manifold Cluster Description Length

- M Data points on a manifold are described by their manifold coordinates
- A Data Point in a K-dimensional manifold has K coordinates
- The K coordinates are the coefficients of its basis vector representation $x^{N\times 1} = \mu^{N\times 1} + B^{N\times K}\lambda^{K\times 1}$
- An observed data point is near but not on the manifold
  - Determine the number of bits that it would take to encode the perturbation that brings a point from its coordinates on the manifold to its associated observation off the manifold to within the Error Tolerance
  - Entropy E of the $N-K$ perturbation distribution
  - Total: $MK + E$

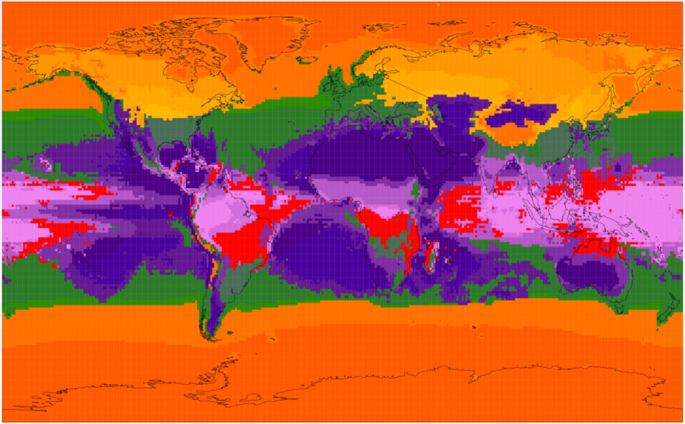
Quality Score

- $X = BMN$: Number of bits to represent the M data points of a cluster in its original representation
- $Y = B[1 + N + K(N - \frac{(K+1)}{2}) + BKM + E$: Number of bits to represent the M data points in the manifold cluster
- If $Y << X$ keep the cluster

Climate Data

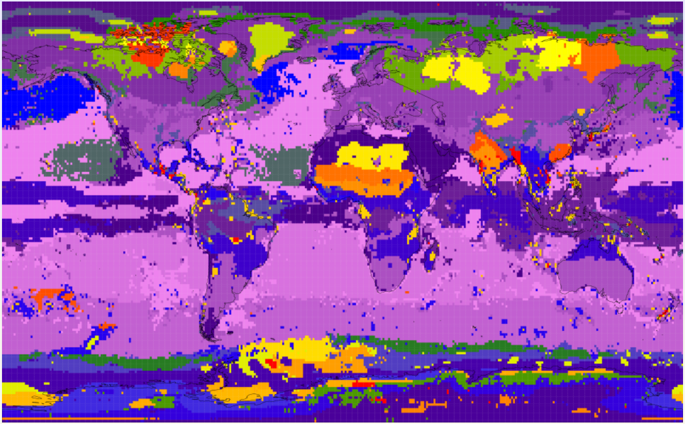
- $\frac{1}{2}^\circ$, $1^\circ, 2^\circ$
- A few decades
- By Month
  - Average Temperature
  - Average Precipitation
- 24-Dimensional

Climate Zones Ground Truth



- Done manually
- Ground Truth Data is known to be faulty

Linear Manifold Clusters



- Done automatically
- Temperature and Precipitation