Inexact MDL for Linear Manifold Clustering

Robert M. Haralick

Computer Science, Graduate Center
City University of New York

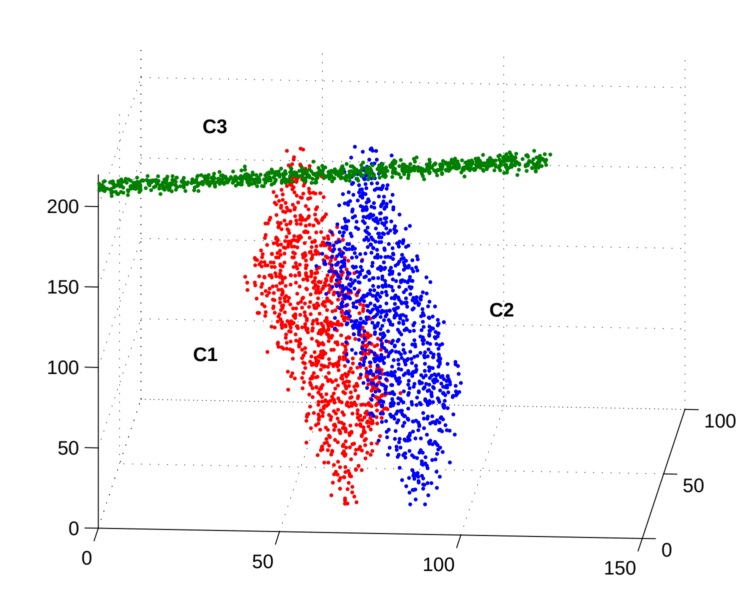
Clustering

- Population has multiple subpopulations
- Observe a Sample of Measurements from the Population
- Identify the subpopulations (Structure)
- For each subpopulation
  - Determine its Center which Serves as the Origin
  - Characterize the relationships among the variables (Dependency)
  - Characterize the Region in Measurement Space Associated with Measurements from each Subpopulation (Structure)
- For any new measurement tuple
  - Determine the Subpopulation to Which It Belongs
  - Adapt the cluster description to new data points

K-Means Clustering

- The cluster mean is the center
- There are no dependencies between variables
- A new measurement is associated with the cluster having the closest center
- Its fitted value is the cluster center

Linear Manifold Clusters



Linear Manifold Clustering

- The cluster mean is the center
- K-Dimensional Linear Manifold
  - The columns of $B^{N \times K}$ are the basis of the linear manifold
  - $x^{N \times 1} = \mu^{N \times 1} + B^{N \times K} \alpha^{K \times 1}$ for some α
  - α is a random vector with large covariance
- Dependencies
  - The columns of $\bar{B}^{N \times N-K}$ are the basis of the orthogonal complement space
  - $\bar{B}'(x - \mu) = \beta$
  - β is a random vector with small covariance
- A new measurement is associated with cluster C_m, having the closest manifold
  - $\rho(x, C_m) = \| \bar{B}'_m(x - \mu) \|^2$ is minimum over all clusters
  - $\rho(x, C_m) \leq \theta$

1-D Manifold

λ is the coordinate relative to the Linear Manifold

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{pmatrix} + \lambda \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{pmatrix}$$

If x_1 is given, then

$$\begin{aligned} \lambda &= \frac{x_1 - \mu_1}{b_1} \\ x_n &= \mu_n + \lambda b_n \\ &= \mu_n + \frac{x_1 - \mu_1}{b_1} b_n \\ &= \mu_n - \mu_1 \frac{b_n}{b_1} + x_1 \frac{b_n}{b_1} \end{aligned}$$

Description Length Model

- Represent the Manifold
- Represent the relative coordinates of a vector in the Manifold
  - Project the vector onto the Manifold
- Represent the distribution of projections on each basis vector of the orthogonal complement space
- Alternative
  - Approximately represent the Manifold orthogonal complement of a vector
    - Project the vector onto the orthogonal complement
    - Represent it approximately

The Linear Manifold

- Linear Manifold has dimension M
- Let the columns of $B^{N \times M}$ be the orthonormal basis of the Manifold
- Relative coordinate projection of x onto the manifold
  - $B'(x - \mu)$
- N-space coordinate projection of x onto the manifold
- $\mu + BB'(x - \mu)$

The Manifold Orthogonal Complement

- Let the columns of $\bar{B}^{N \times N-M}$ be the orthonormal basis of the Manifold Orthogonal Complement
- Relative coordinates of projection of x onto the manifold orthogonal complement space
  - $\bar{B}'(x - \mu)$
- N-space coordinate of projection of x onto the manifold orthogonal complement space
  - $\bar{B}\bar{B}'(x - \mu)$

Model Encoding

- N-dimensional space
- μ
  - N numbers
- B and $\bar{B}$
  - N^2 numbers
  - Constraints
    - Norm 1 N constraints
    - Orthogonal $N(N-1)/2$ constraints
  - $N^2 - N - N(N-1)/2 = N(N-1)/2$
- Total $N + N(N-1)/2 = N(N+1)/2$
- Precision is number of bits
- Precision P_M for all numbers of manifold basis
- Total Bits $P_M N(N+1)/2$

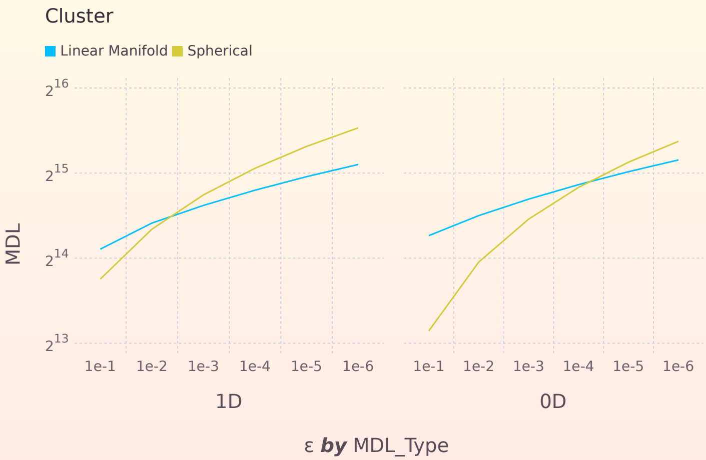
Data Encoding

- J data points
- Precision P_D for coordinates on the Manifold
- $P_D < P_M$
- Manifold has dimension M
  - JMP_D Bits
- All squared errors will be less than ϵ^2
- $S(\epsilon)$ bits on the Manifold Orthogonal Complement
  - $JS(\epsilon)$ bits
- Total is $J(MP_D + S(\epsilon))$

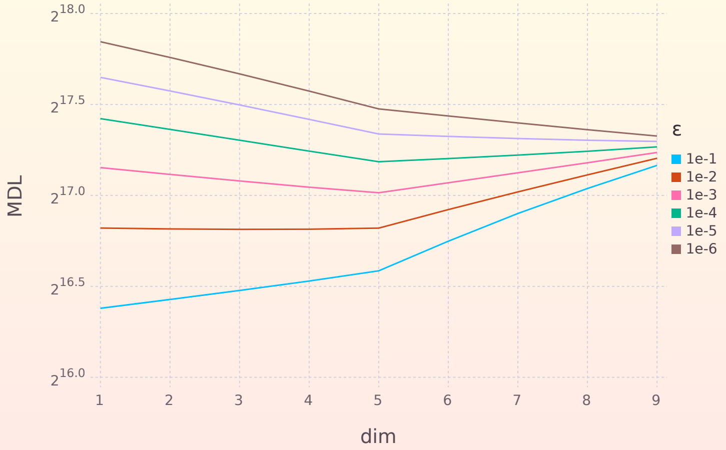
Approximate Orthogonal Complement Encoding

- $K = N - M$ components need to be encoded
- $[-A_k/2, A_k/2]$ range of values for the kth component
- Equal interval quantize with N_k bins
- Each interval has width A_k/N_k
- Assume uniform distribution in each of the N_k intervals
- Variance of the distribution in each interval is $\frac{1}{12}\left(\frac{A_k}{N_k}\right)^2$
- Require $N_k, k = 1, \ldots, K$ to satisfy $\frac{1}{12}\sum_{k=1}^{K}\left(\frac{A_k}{N_k}\right)^2 \leq \epsilon^2$
- Probability p_{nk} for interval n of component k
- $S(\epsilon) = -\sum_{k=1}^{K}\left[\sum_{n=1}^{N_k} p_{nk} \log p_{nk}\right]$

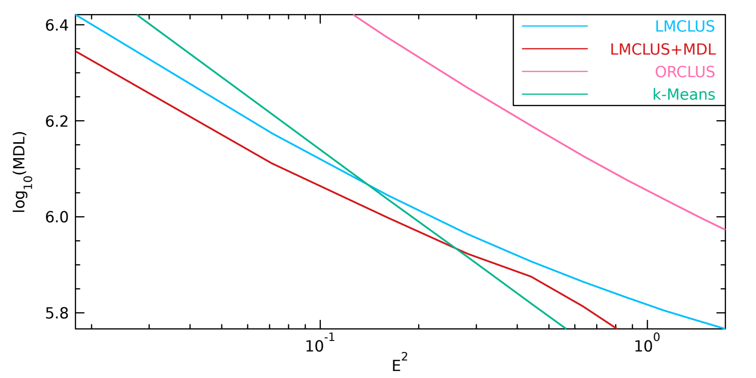
K-means vs MDL



5D Cluster in 10D space



Comparison



LMCLUS+MDL is better than K-Means for smaller errors

Data Normalization

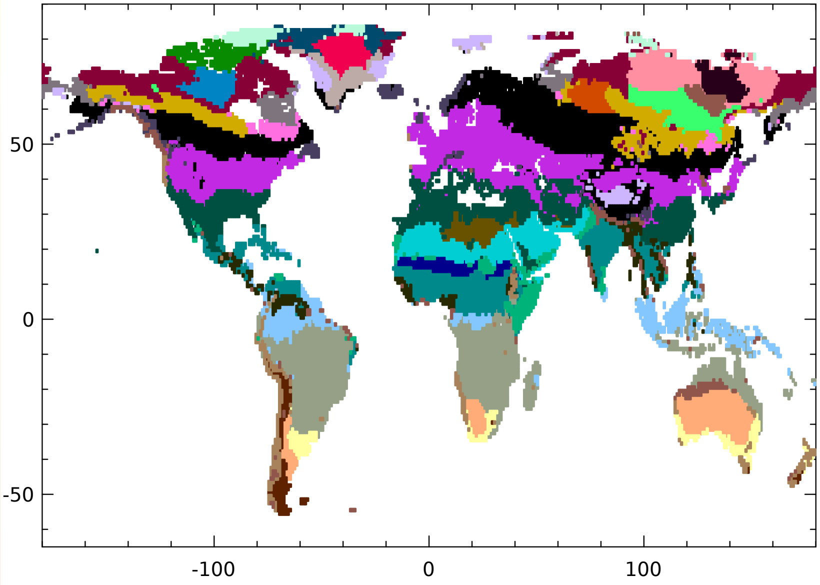
- Temperature and Precipitation have different scales
- Each field is normalized

$$\frac{x - x_{mean}}{x_{max} - x_{min}}$$

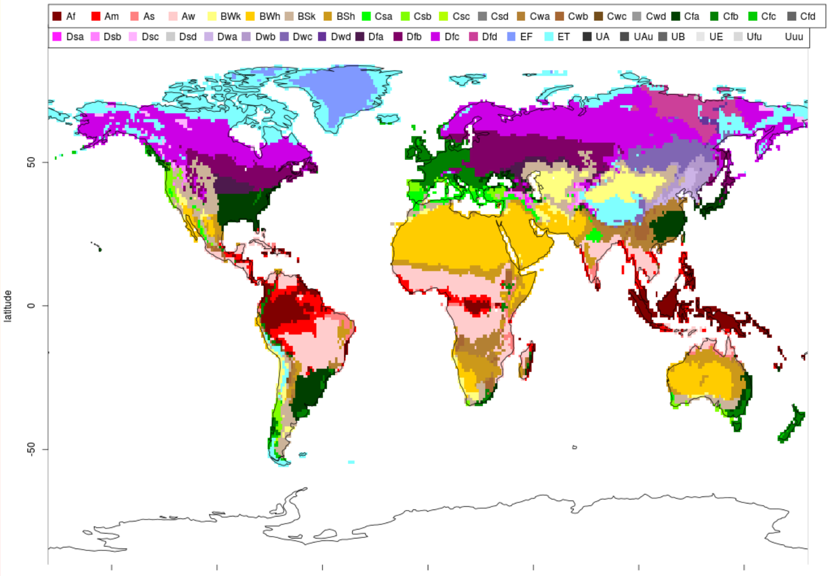
24D Climate Data

- $\epsilon = .001$
- 1951-1980
- Resolution $1^\circ \times 1^\circ$
- CRU 3.22 data set
  - Monthly Surface Temperature Averages
- Global Precipitation Climatology Centre
  - Monthly Precipitation Averages
- 12 Monthly Temperatures Averaged over 30 years
- 12 Monthly Precipitation Averaged over 30 years

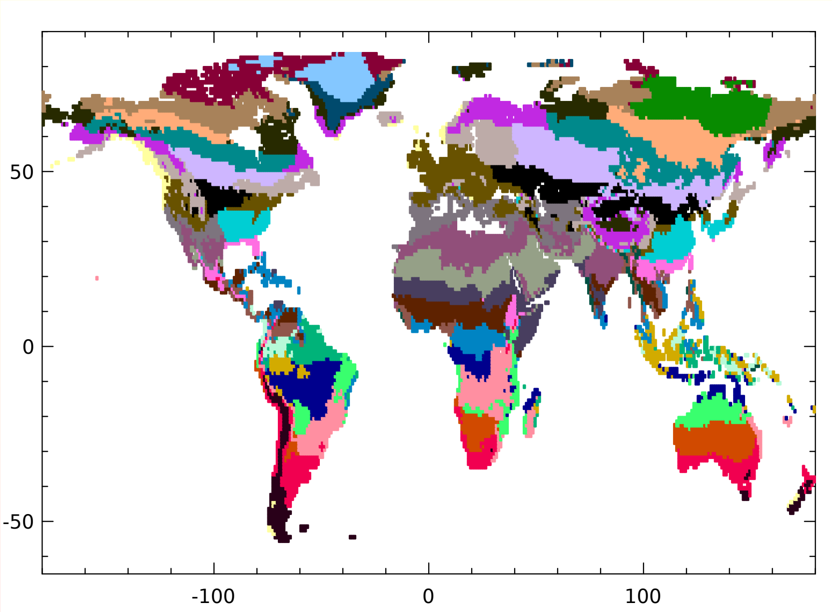
MDL Linear Manifold Clusters



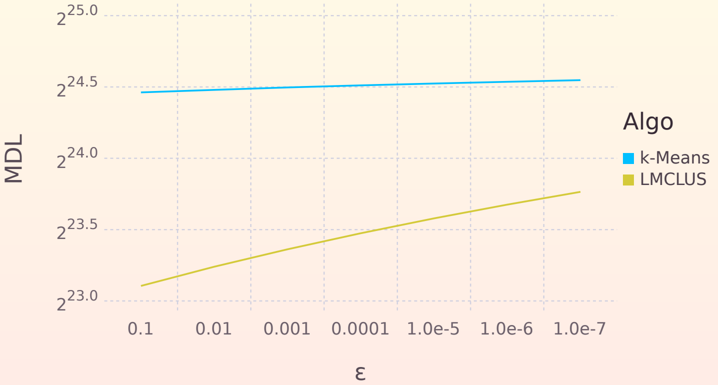
Koeppen-Geiger Map



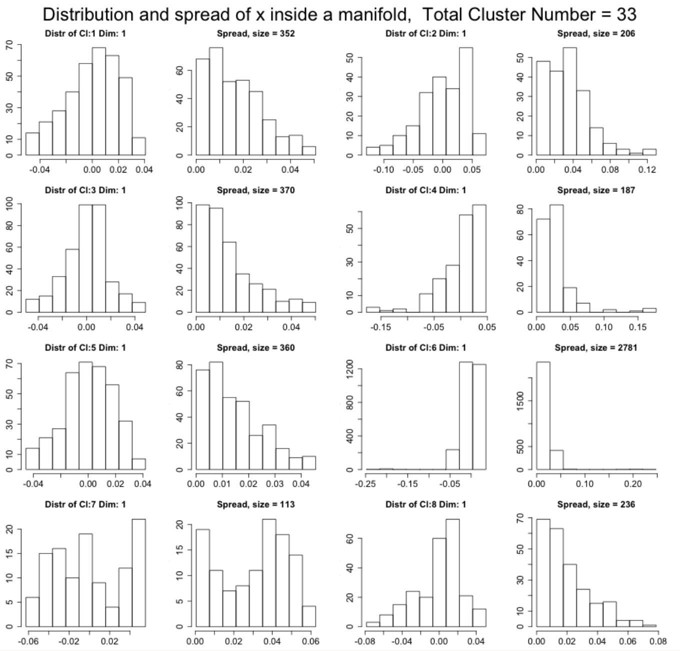
K-Means Clusters



Description Length Comparison



Within Cluster Distribution



Within Cluster Distribution