The Curse of Dimensionality
(formal definition)

- Stone (1982); Optimal rate of convergence for non-parametric regression:
  \[ n^{-2p/(2p+d)} \]

- E.g. for dimensionality \( d = 8 \), and smoothness \( p = 2 \) (p bounded derivatives of the unknown regression function), a sample of size \( n \geq 10^6 \) is needed to make
  \[ n^{-2p/(2p+d)} < .01 \]

- The success of some recent methods suggests that the bound is not optimal for real world problems

- Motivates search for lower dimensional structure
What Are Interesting Directions/How to Reduce Dimensionality


- Therefore, measure some deviation from Normality

- Concentrate on the center of the distribution

- Seek distinguishing features between clusters (Discriminant Analysis)
Second Order Statistics

Principal Components can not find good projections
Exploratory Projection Pursuit (EPP)


- Seeks *interesting* low dimensional projections of a high dimensional point cloud, by numerically maximizing a projection index.

- For review see Huber (1985), Jones and Sibson (1987).
BCM Neuron and Projection Pursuit

- A recent variant of the BCM neuron (Bienenstock Cooper and Munro, 1982) yields synaptic modification equations that maximize a projection index (Intrator 1990; Intrator and Cooper, 1992).

- The projection index emphasizes deviation from normality of a multi-modality type.

- This formulation naturally extends to a lateral inhibition network, which can find several projections at once.
How Does The BCM Neuron Reduce Dimensionality

- The solution to a 2 dimensional problem with 2 inputs is either $m^1$ which is orthogonal to $d^2$ or $m^2$ which is orthogonal to the input $d^1$.
BCM in clustered data

- In a two cluster input problem:

- The distribution of the projections say through \( m^1 \) is bi-modal with one mode centered at zero

- This is the general behavior in high dimensional space as well
BCM Modification Equations

\[ \frac{dm_i}{dt} = \mu \, \phi(c, \Theta_m) d_i, \]

for \( \Theta_m = E[(m \cdot d)^2] \) and \( \phi(c, \Theta_m) = c(c - \Theta_m) \).

In the lateral inhibition network \( c_k = m_k \cdot d; \), where

\[ \tilde{c}_k = \sigma(c_k - \eta \sum_{j \neq k} c_j), \]

\[ \tilde{\Theta}_m = E[\tilde{c}_k^2], \]

\[ R(w_k) = -\left\{ \frac{1}{3} E[\tilde{c}_k^3] - \frac{1}{4} E^2[\tilde{c}_k^2] \right\}. \]

\[ \dot{m}_k = \mu [\phi(\tilde{c}_k, \tilde{\Theta}_m) \sigma'(\tilde{c}_k) - \eta \sum_{j \neq k} \phi(\tilde{c}_j, \tilde{\Theta}_m) \sigma'(\tilde{c}_j)] d. \]
Related Computational Issues

- Use of low order polynomial moments – computationally efficient

- Unsensitive to outliers

- Naturally extends to multi-dimensional projection pursuit

- Number of calculations of the gradient grows \textit{linearly} with the number of projections sought

- The projection index has a stochastic gradient descent version
Related Statistical Issues

- Less biased to the class labels, in contrast to discriminant analysis

- Seeks cluster discrimination not faithful representation of the data (principal components analysis, factor analysis – combines features with high correlation)

- Unlike cluster analysis or multi-dim scaling, the searches is done in the low dimensional projection space

- The search is constrained by seeking projections orthogonal to all but one of the clusters (have a mode at zero). Thus, at most $K$ optimal projections not $\binom{K}{2}$ separating hyperplanes.