Generalized Principal Component Analysis via Lossy Coding and Compression

Yi Ma

Image Formation & Processing Group, Beckman
Decision & Control Group, Coordinated Science Lab.
Electrical & Computer Engineering Department

University of Illinois at Urbana-Champaign

OUTLINE

MOTIVATION

PROBLEM FORMULATION AND EXISTING APPROACHES

SEGMENTATION VIA LOSSY DATA COMPRESSION

SIMULATIONS (AND EXPERIMENTS)

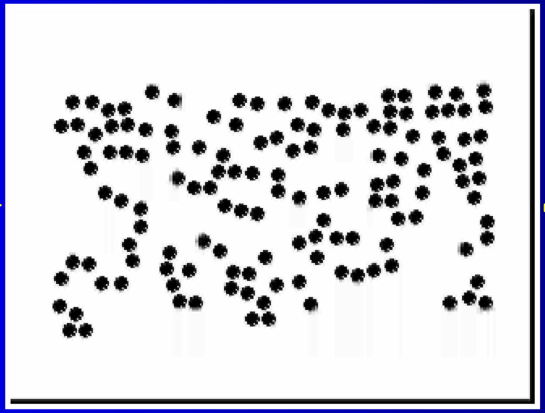
CONCLUSIONS AND FUTURE DIRECTIONS

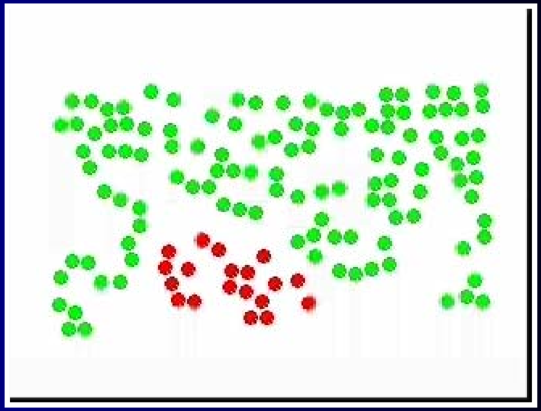
MOTIVATION – Motion Segmentation in Computer Vision

Goal: Given a sequence of images of multiple moving objects, determine:

1. the number and types of motions (rigid-body, affine, linear, etc.)
2. the features that belong to the same motion.

QuickTime™ and a
Cinepak decompressor
are needed to see this picture.



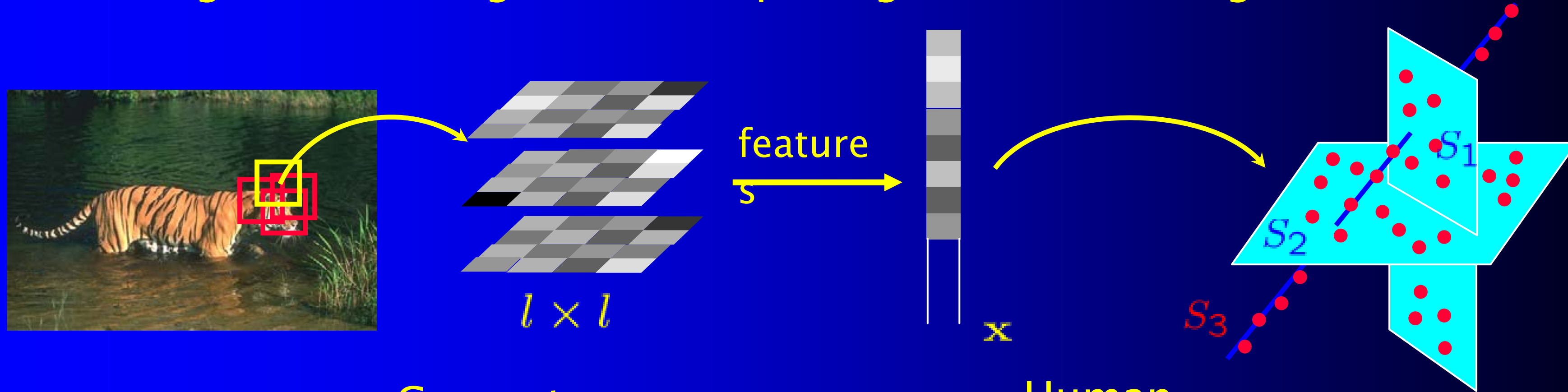


The "chicken-and-egg" difficulty:

- Knowing the segmentation, estimating the motions is easy;
- Knowing the motions, segmenting the features is easy.

MOTIVATION – Image Segmentation

Goal: segment an image into multiple regions with homogeneous texture.



Computer Human

Difficulty: A mixture of models of different dimensions or

Multiscale Hybrid Linear Models for Lossy Image Representation, [Hong-Wright-Ma, TIP

MOTIVATION – Video Segmentation

Goal: segmenting a video sequence into segments with "stationary" dynamics

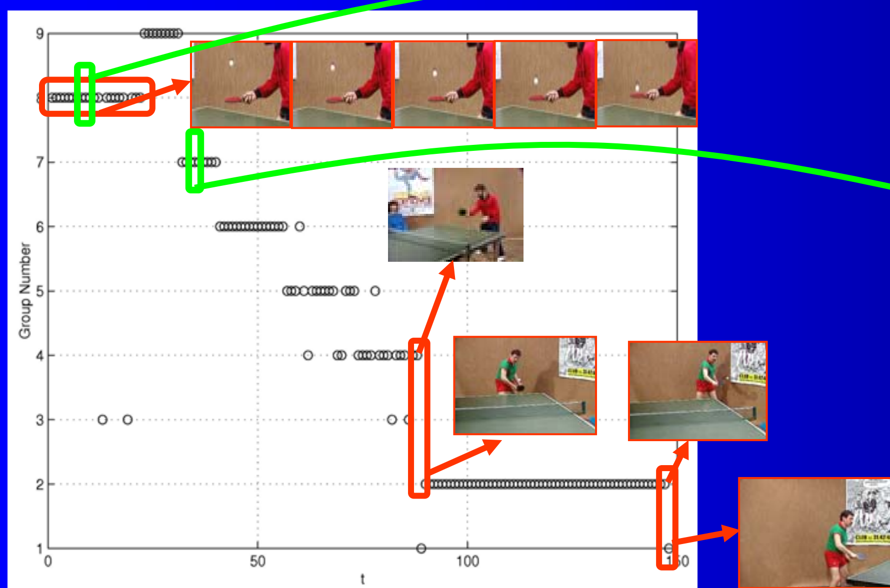
Model: different segments as outputs from different (linear) dynamical systems:

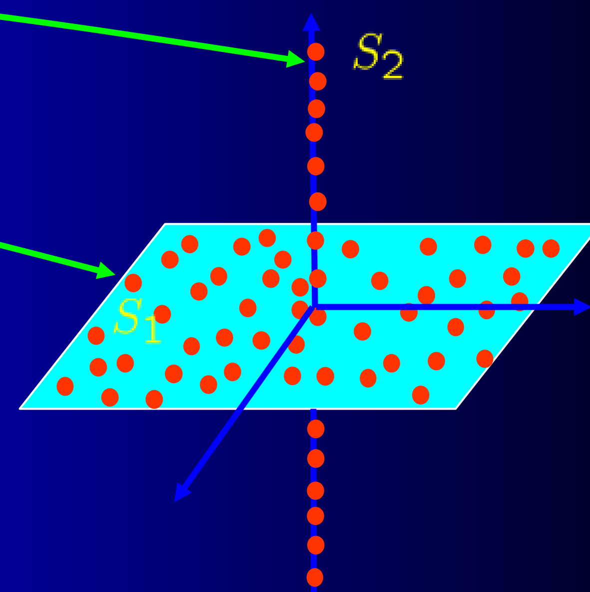
$$x_{t+1} = A_{\lambda(t)} x_t + B_{\lambda(t)} u_t$$

$$y_t = C_{\lambda(t)} x_t + D_{\lambda(t)} u_t$$

$$\lambda(t) \in \{1, 2, \ldots, n\}$$

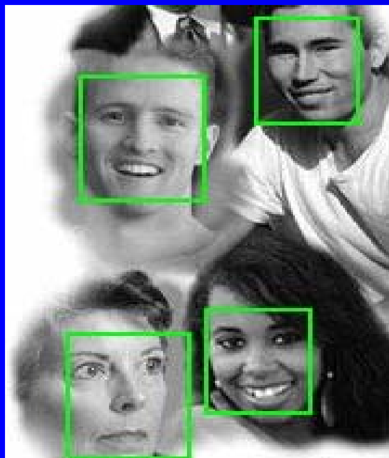
QuickTime™ and a H.264 decompressor are needed to see this picture.



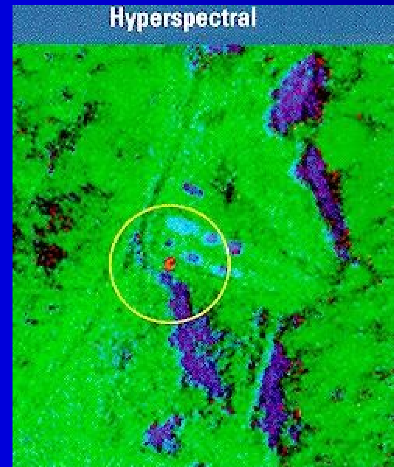


Identification of Hybrid Linear Systems via Subspace Segmentation, [Huang-Wagner-Ma, C

MOTIVATION – Massive Multivariate Mixed Data



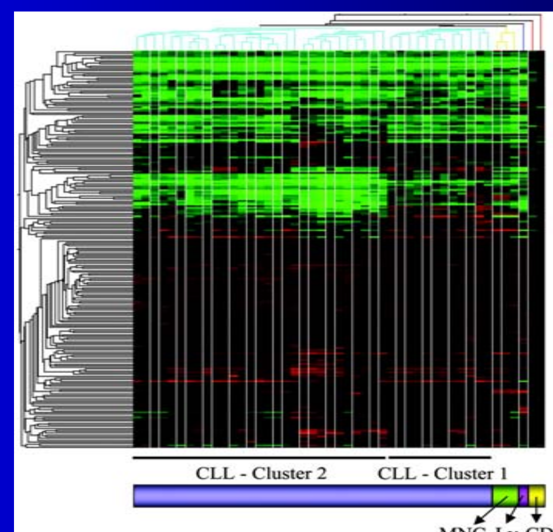
Face database



Hyperspectral images

Articulate motions

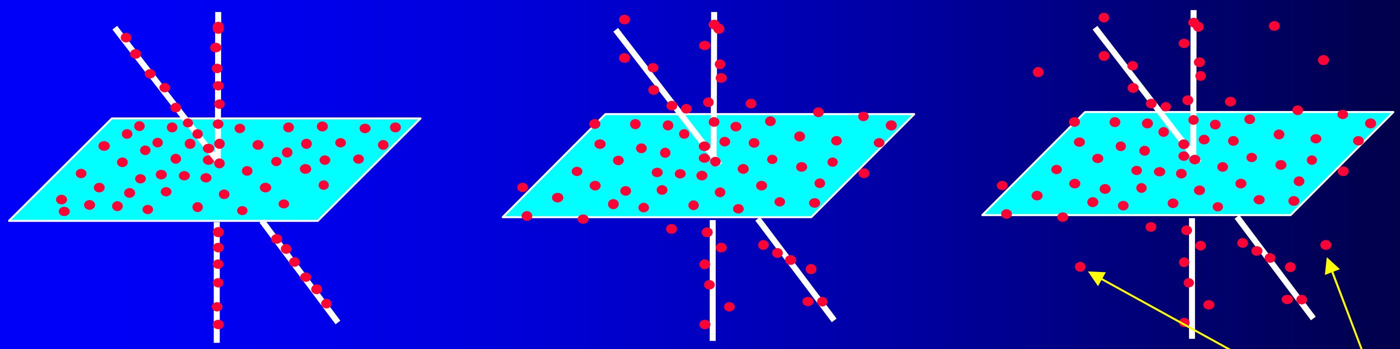
Hand written digits



Microarrays

SUBSPACE SEGMENTATION – Problem Formulation

Assumption: the data $\{x_1, x_2, \ldots, x_N\}$ are noisy samples from an arrangement of linear subspaces $A = S_1 \cup S_2 \cup \cdots \cup S_n$.



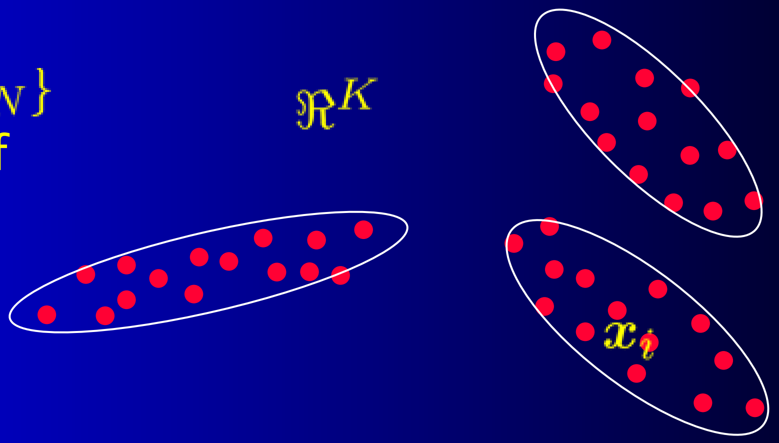
noise-free samples noisy samples samples with outliers

Difficulties:

- the dimensions of the subspaces can be **different**
- the data can be **corrupted** by noise or **contaminated** by outliers
- the number and dimensions of subspaces may be **unknown**

SUBSPACE SEGMENTATION – Statistical Approaches

Assume that the data $\{x_1, x_2, \ldots, x_N\}$ are i.i.d. samples from a mixture of probabilistic distributions:



$$p(x,\theta) = \sum_{i=1}^{n} \pi_i p_i(x,\theta)$$

Solutions:

- Expectation Maximization (EM) for the maximum-likelihood estimate [Dempster et. al.'77], e.g., Probabilistic PCA [Tipping-Bishop'99]:

$$\max_{\theta,\pi} \sum_{j=1}^{N} \log \Big(\sum_{i=1}^{n} \pi_i p_i(x_j,\theta) \Big)$$

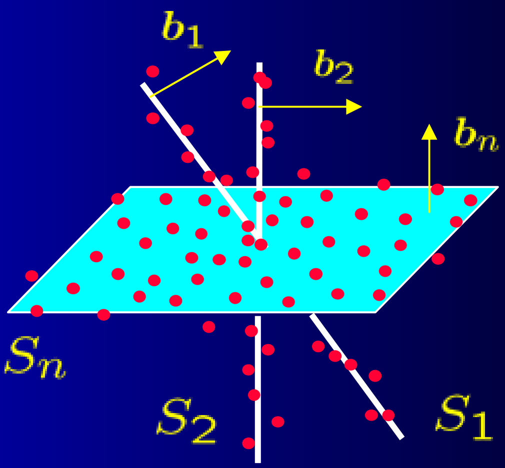
- K-Means for a minimax-like estimate [Forgy'65, Jancey'66, MacQueen'67], e.g., K-Subspaces [Ho and Kriegman'03]:

$$\min_{\theta} \sum_{j=1}^{N} \min_i \Big(-\log p_i(x_j,\theta) \Big)$$

Essentially iterate between data segmentation and model estimation.

SUBSPACE SEGMENTATION – An Algebro-Geometric Approach

Idea: a union of linear subspaces is an algebraic set -- the zero set of a set of (homogeneous) polynomials:



$$\begin{aligned} \mathcal{A} &= S_1 \cup S_2 \cup \cdots \cup S_n \\ &= \{x : p(x) = 0, p \in I(\mathcal{A})\}. \end{aligned}$$

Solution:

- Identify the set of polynomials of degree n that vanish on $\mathcal{A}$

$$\left\{p(x) = (b_1^T x)(b_2^T x)\cdots(b_n^T x) = c^T \nu_n(x),\ b_i \in S_i^\perp\right\}.$$

- Gradients of the vanishing polynomials are normals to the subspaces

$$\left.\frac{\partial p(x)}{\partial x}\right|_{x \in S_i} = b_i \prod_{j \neq i}(b_j^T x) \in S_i^\perp,\ i = 1, 2, \ldots, n.$$

Complexity exponential in the dimension and number of subspaces.

Generalized Principal Component Analysis, [Vidal-Ma-Sastry, IEEE Transactions PAMI'0

SUBSPACE SEGMENTATION – An Information-Theoretic Approach

Problem: If the number/dimension of subspaces not given and data corrupted
by noise and outliers, how to determine the optimal subspaces that fit the data?

Solutions: Model Selection Criteria?

- Minimum message length (MML) [Wallace-Boulton'68]
- Minimum description length (MDL) [Rissanen'78]
- Bayesian information criterion (BIC)
- Akaike information criterion (AIC) [Akaike'77]
- Geometric AIC [Kanatani'03], Robust AIC [Torr'98]

Key idea (MDL):

- a good balance between model complexity and data fidelity.
- minimize the length of codes that describe the model and the data:

$$\min_{\theta \in \Theta} \text{Length}(X, \theta) = \text{Length}(\theta) + \text{Length}(X|\theta).$$

with a quantization error optimal for the model.

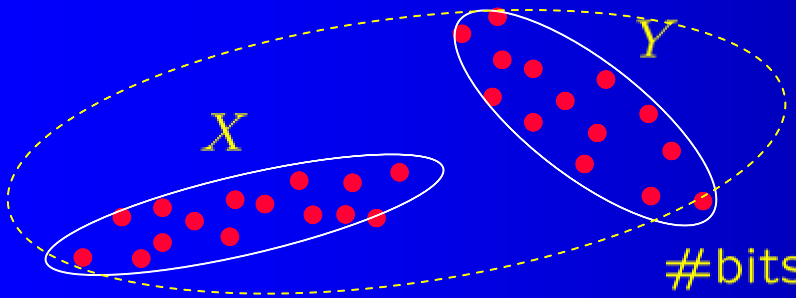
LOSSY DATA COMPRESSION

Questions:

- What is the “gain” or “loss” of segmenting or merging data?

- How does tolerance of error affect segmentation results?

Basic idea: whether the number of bits required to store “the whole is more than the sum of its parts”?



$\#\text{bits}(X \cup Y) \geq \#\text{bits}(X) + \#\text{bits}(Y)$?

LOSSY DATA COMPRESSION – Problem Formulation

– A coding scheme maps a set of vectors $V = [v_1, v_2, \ldots, v_m] \in \Re^{K\times m}$ to a sequence of bits, from which we can decode $\hat{v}_i : \mathbb{E}\|\hat{v}_i - v_i\|^2 \leq \epsilon^2$. The **coding length** is denoted as:

$$\begin{aligned} L : \Re^{K\times m} &\rightarrow Z_+ \\ V &\mapsto L(V) \end{aligned}$$

– Given a set of real-valued mixed data $X = [x_1, x_2, \ldots, x_N] \in \Re^{K\times N}$ the **optimal segmentation** $X = X_1 \cup X_2 \cup \cdots \cup X_n$ minimizes the overall coding length:

$$L^s(X) \doteq L(X_1) + L(X_2) + \cdots + L(X_n) + H(|X_1|, |X_2|, \ldots, |X_n|)$$

where

$$H(|X_1|, |X_2|, \ldots, |X_n|) \doteq \sum_{i=1}^n |X_i| \Big(-\log_2(|X_i|/N)\Big).$$

LOSSY DATA COMPRESSION – Coding Length for Multivariate Data

Theorem.

Given $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \Re^{K \times N}$ with $\mu = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i, \bar{X} = X - \mu$

$$L(X) = \frac{N+K}{2}\log_2 \det\left(I + \frac{K}{\epsilon^2 N}\bar{X}\bar{X}^T\right) + \frac{K}{2}\log_2\left(1 + \frac{\mu^T\mu}{\epsilon^2}\right)$$

is the number of bits needed to encode the data s.t. $\mathbb{E}\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2 \le \epsilon^2$

.

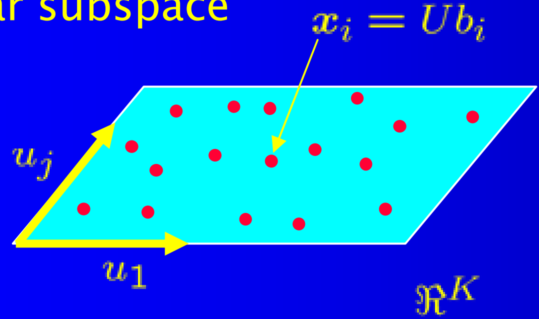
A nearly optimal bound for even a small number of vectors drawn from a subspace or a Gaussian source.

Segmentation of Multivariate Mixed Data, [Ma-Derksen-Hong-Wright, PAMI'

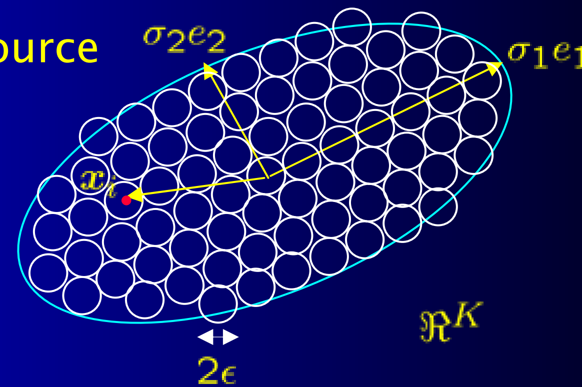
LOSSY DATA COMPRESSION – Two Coding Schemes

Goal: code $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$ s.t. a mean squared error $\mathbb{E}\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2 \leq \epsilon^2$

Linear subspace



Gaussian source



$$X = U \Sigma V^T \doteq U B$$

$$\delta u_{ij} \sim \left[-\frac{\epsilon\sqrt{N}}{\sigma_j K}, \frac{\epsilon\sqrt{N}}{\sigma_j K}\right], \quad \delta b_{ij} \sim \left[-\frac{\epsilon}{\sqrt{K}}, \frac{\epsilon}{\sqrt{K}}\right]$$

$$\#\text{bits}(U) \leq \frac{K}{2}\sum_{i=1}^{K} \log_2\left(1 + \frac{K\sigma_i^2}{N\epsilon^2}\right)$$

$$\#\text{bits}(B) \leq \frac{N}{2}\sum_{i=1}^{K} \log_2\left(1 + \frac{K\sigma_i^2}{N\epsilon^2}\right)$$

$$\#\text{bits} \leq \frac{N+K}{2}\sum_{i=1}^{K} \log_2\left(1 + \frac{K\sigma_i^2}{N\epsilon^2}\right)$$

$$\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i + \boldsymbol{z}_i, \quad \boldsymbol{z}_i \sim N(0, \frac{\epsilon^2}{K} I)$$

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i^T\right] = \frac{\epsilon^2}{K} I + \frac{1}{N} X X^T$$

$$\text{vol}(\hat{X}) \propto \sqrt{\det\left(\frac{\epsilon^2}{K} I + \frac{1}{N} X X^T\right)}$$

$$\text{vol}(\boldsymbol{z}) \propto \sqrt{\det\left(\frac{\epsilon^2}{K} I\right)}$$

$$=$$

$$\#\text{bits} = (N + K)\log_2\left(\text{vol}(\hat{X})/\text{vol}(\boldsymbol{z})\right)$$

LOSSY DATA COMPRESSION – Properties of the Coding Length

$$L(X) = \frac{N+K}{2}\log_2 \det\left(I + \frac{K}{\epsilon^2 N} X X^T\right)$$

1. *Commutative Property:* $\det\left(I + \frac{K}{\epsilon^2 N} X X^T\right) = \det\left(I + \frac{K}{\epsilon^2 N} X^T X\right).$

   For high-dimensional data, computing the coding length only needs the **kernel matrix:** $X^T X$.

2. *Asymptotic Property:* $\lim_{N\to\infty} \frac{1}{N} L(X) \doteq R(\epsilon) = \frac{1}{2}\log_2\left[\det\left(I + \frac{K}{\epsilon^2}\Sigma_X\right)\right].$

   At high SNR, this is the **optimal** rate distortion for a Gaussian source.

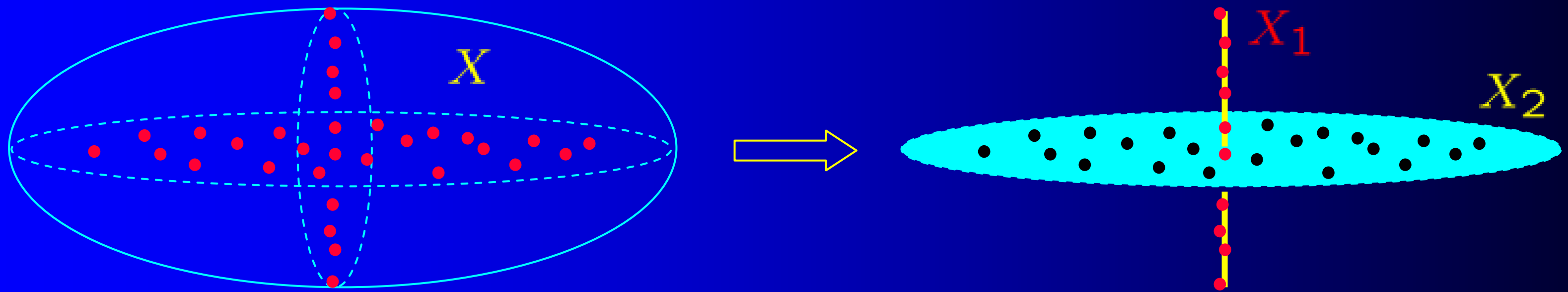
3. *Invariant Property:* $L(X) = L(UX) = L(XV),\ \forall U \in O(K), V \in O(N).$

   Harmonic Analysis is useful for data compression only when the data are **non-Gaussian or nonlinear** ……… **so is segmentation!**

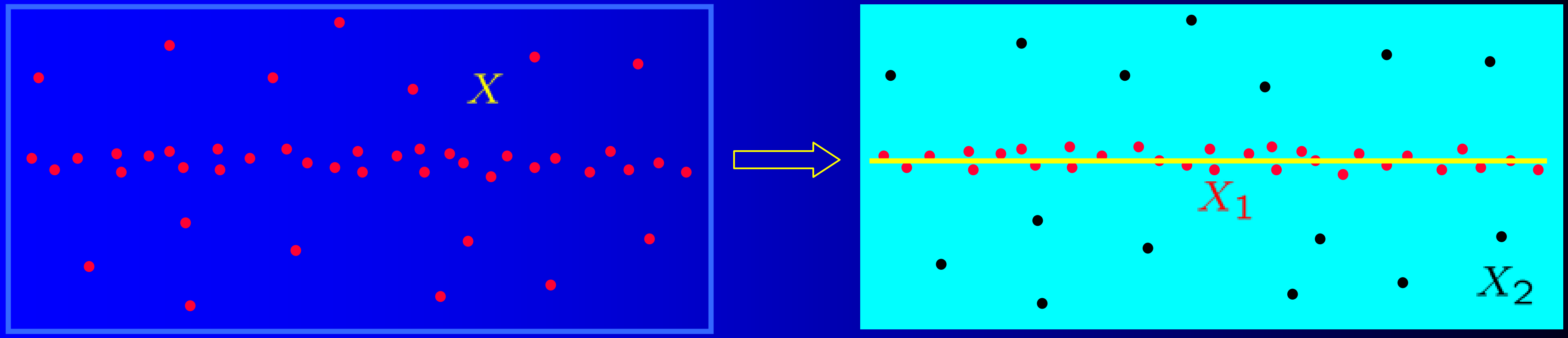
LOSSY DATA COMPRESSION – Why Segment?

$$L(X) > L^s(X) = L(X_1) + L(X_2) + H(|X_1|, |X_2|)$$

partitioning:



sifting:



LOSSY DATA COMPRESSION – Probabilistic Segmentation?

Assign the ith point to the jth group with probability $\pi_{ij} \in [0, 1], j = 1, 2, \ldots, n.$

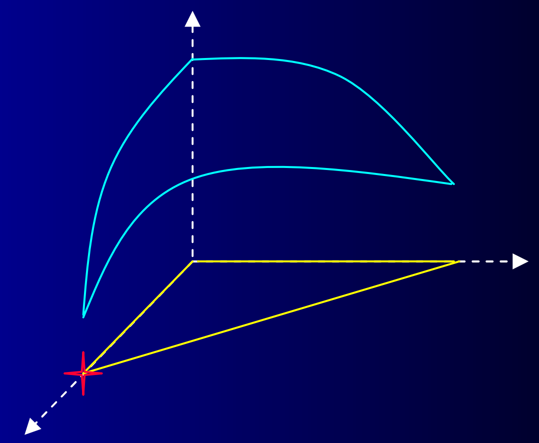
$$\Pi_j \doteq \begin{bmatrix} \pi_{1j} & 0 & \cdots & 0 \\ 0 & \pi_{2j} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \pi_{Nj} \end{bmatrix} \in \Re^{N \times N}, \quad \Pi_j \succeq 0, \sum_{j=1}^{n} \Pi_j = I_{N \times N}.$$

Theorem. The expected coding length of the segmented data

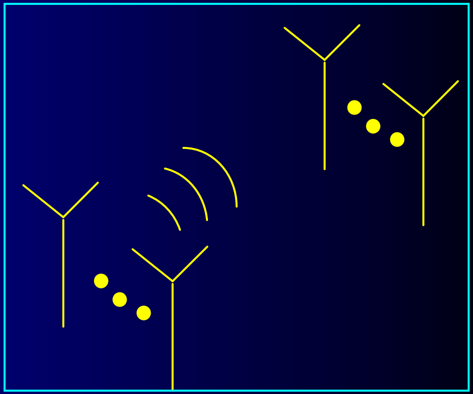
$$L^s(X, \Pi) = \sum_{j=1}^{n} \frac{\mathrm{tr}(\Pi_j)}{2} \log_2 \det\left(I + \frac{K}{\epsilon^2 \mathrm{tr}(\Pi_j)} X \Pi_j X^T\right) - \mathrm{tr}(\Pi_j) \log_2\left(\frac{\mathrm{tr}(\Pi_j)}{N}\right)$$

is a *concave* function in Π over the domain of a convex polytope.

Minima are reached at the vertexes of the polytope -- no probabilistic segmentation!



LOSSY DATA COMPRESSION – Segmentation & Channel Capacity



A MIMO additive white Gaussian noise (AWGN) channel

$$y = Wx + z, \quad W \in \Re^{K\times N},\ z \sim \mathcal{N}(0, \sigma^2 I)$$

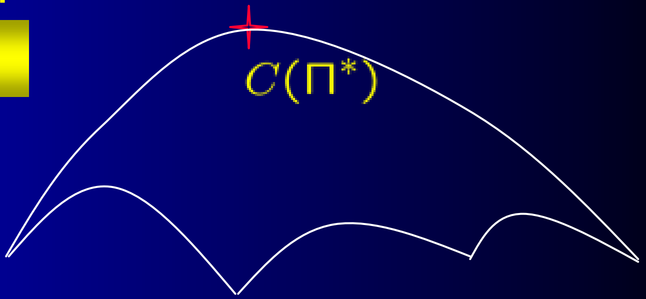
has the capacity: $C(W) \doteq \frac{1}{2}\log_2 \det\left(I + \frac{P}{N\sigma^2} W W^T\right).$

If allowing probabilistic grouping of transmitters, the expected capacity

$$C(W, \Pi) = \sum_{j=1}^{n} \frac{\mathrm{tr}(\Pi_j)}{2N} \log_2 \det\left(I + \frac{P}{\mathrm{tr}(\Pi_j)\sigma^2} W \Pi_j W^T\right)$$

is a *concave* function in Π over a convex polytope.

Maximizing such a capacity is a convex



On Coding and Segmentation of Multivariate Mixed Data, [Ma-Derksen-Hong-Wright, PAMI

LOSSY DATA COMPRESSION – A Greedy (Agglomerative) Algorithm

Objective: minimizing the overall coding length

$$\min L^s(X) = L(X_1) + L(X_2) + \cdots + L(X_n) + H(|X_1|, |X_2|, \ldots, |X_n|).$$

Input: $X = \{x_1, x_2, \ldots, x_N\} \subset \Re^K$, $\epsilon > 0$

$\mathcal{S} = \{S = \{x\} \mid x \in X\}$

while true do

choose two sets $S_1, S_2 \in \mathcal{S}$ such that $L(S_1 \cup S_2) - L^s(S_1, S_2)$ is minimal

if $L(S_1 \cup S_2) - L^s(S_1, S_2) < 0$ **then** $\mathcal{S} = \big(\mathcal{S} \setminus \{S_1, S_2\}\big) \cup \{S_1 \cup S_2\}$

else break

endif

end

Output: $\mathcal{S}$

"Bottom-up" merge

QuickTime™ and a
PNG decompressor
are needed to see this picture.

Segmentation of Multivariate Mixed Data via Lossy Coding and Compression, [Ma-Derksen-Hong-Wright, PAMI'07]

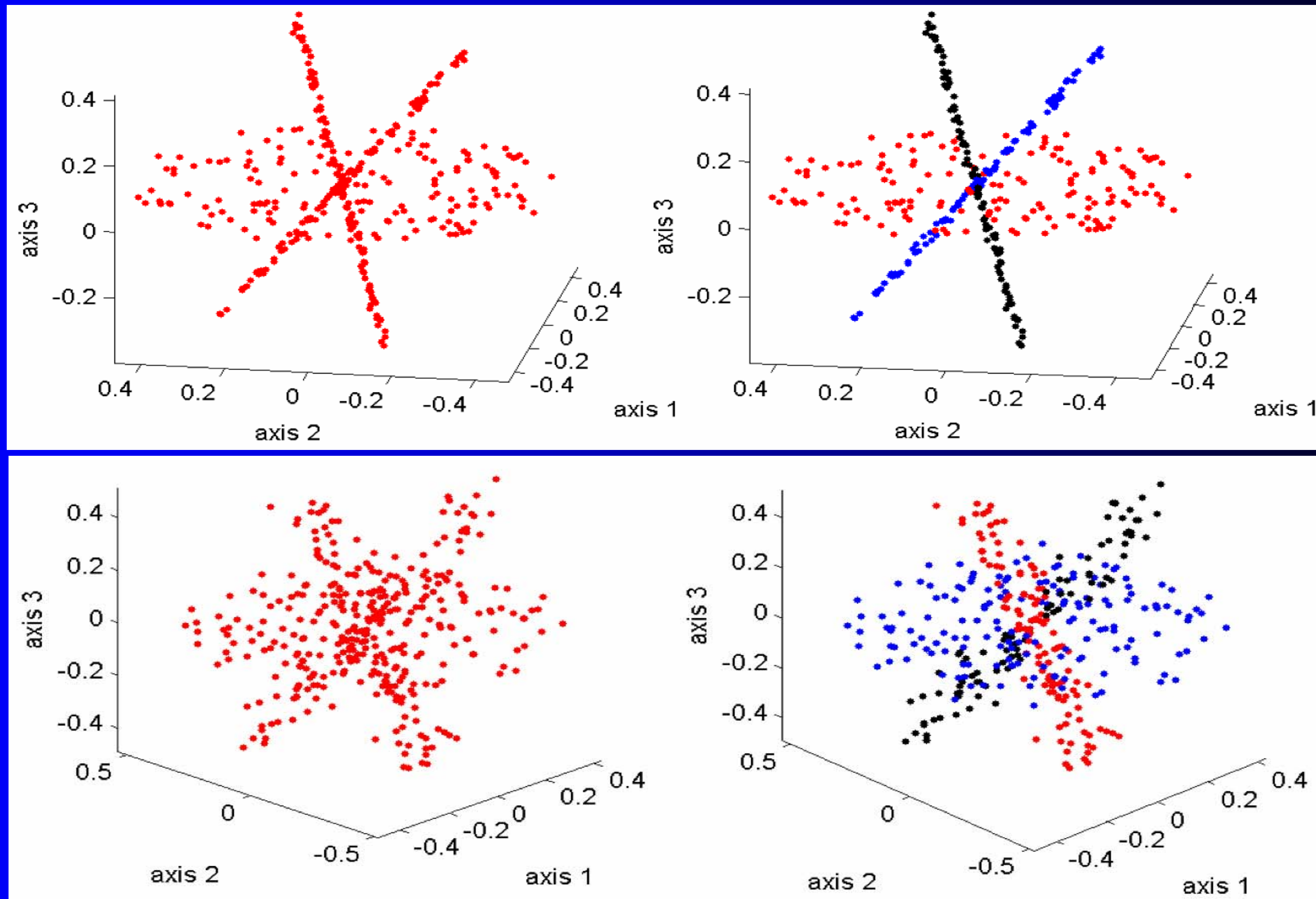
SIMULATIONS – Mixture of Almost Degenerate Gaussians

Noisy samples from two lines and one plane in $\Re^3$

Given Data | Segmentation Results

$\varepsilon_0 = 0.01$

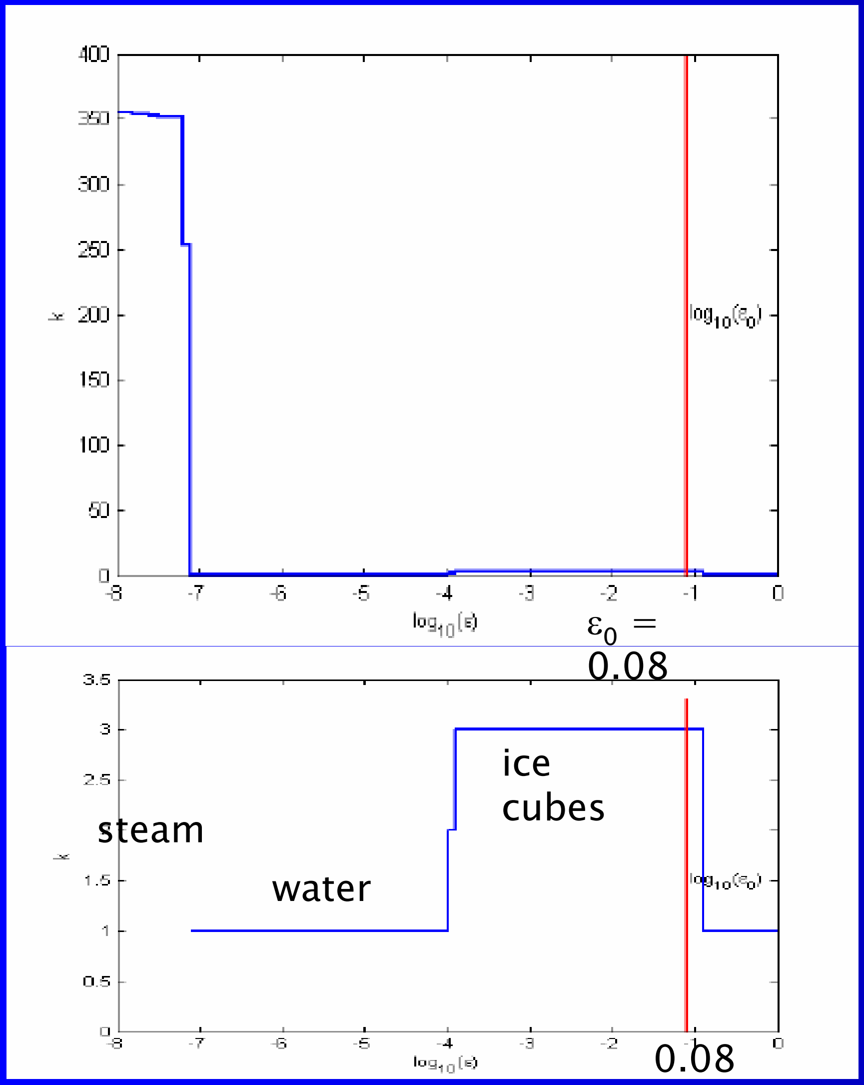
$\varepsilon_0 = 0.08$



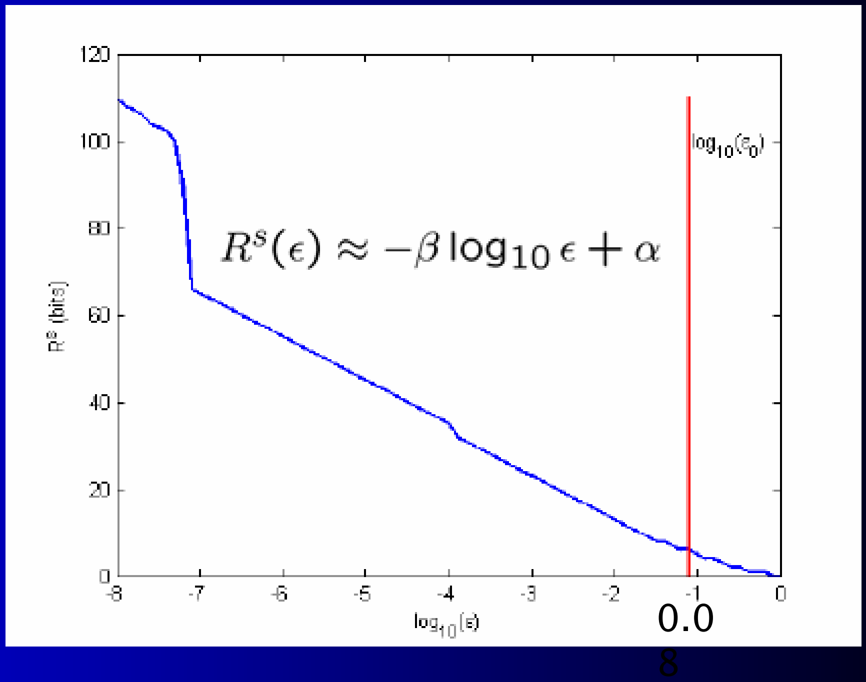
Segmentation of Multivariate Mixed Data via Lossy Coding and Compression, [Ma-Derksen-Hong-Wright, PAMI'07]

SIMULATIONS – "Phase Transition"

#group v.s. distortion



Rate v.s. distortion



Stability: the same segmentation for ε across 3 magnitudes!

Segmentation of Multivariate Mixed Data via Lossy Coding and Compression, [Ma-Derksen-Hong-Wright, PAMI'07]

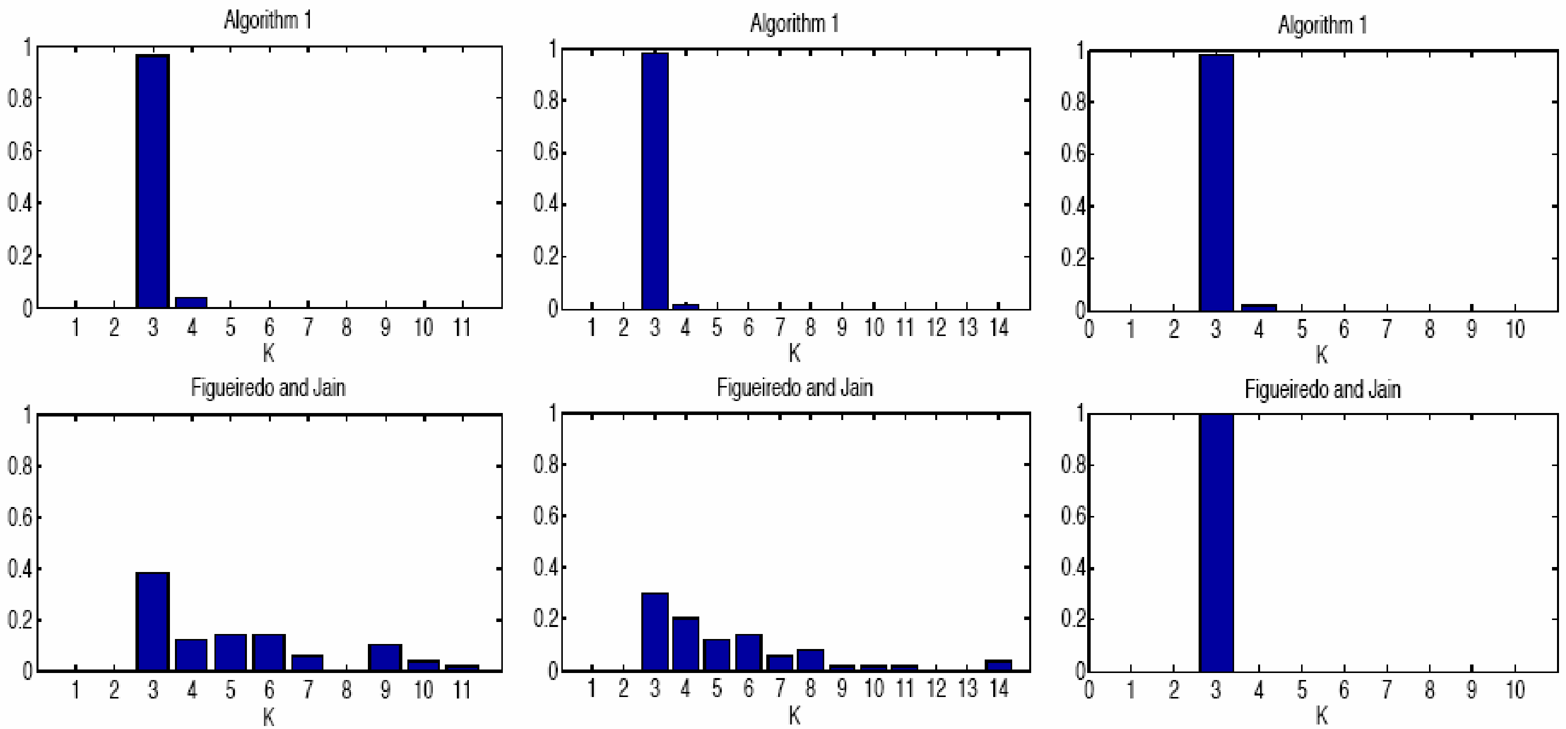
SIMULATIONS – Comparison with EM

100 x d uniformly distributed random samples from each subspace, corrupte
with 4% noise. Classification rate averaged over 25 trials for each case.

Subspace dimensions	Identified dimensions	Classification (%) (Greedy Algorithm)	Classification (%) (E-M)
$(2,1,1)$ in $\Re^3$	$2,1,1$	96.62	39.33
$(2,2,1)$ in $\Re^3$	$2,2,1$	90.00	68.98
$(4,2,2,1)$ in $\Re^5$	$4,2,2,1$	98.53	43.36
$(6,3,1)$ in $\Re^7$	$6,3,1$	99.77	66.16
$(7,5,2,1,1)$ in $\Re^8$	$7,5,2,1,1$	98.04	42.29

SIMULATIONS – Comparison with EM

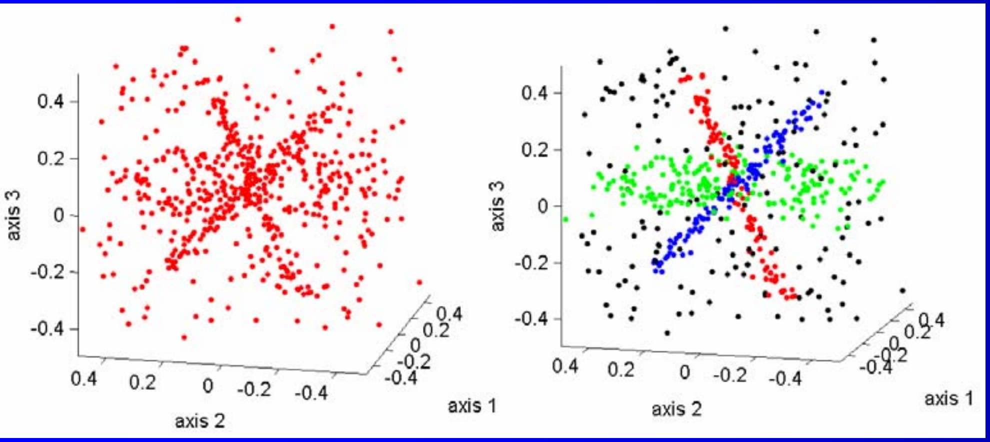
Segmenting three degenerate or non-degenerate Gaussian clusters for 50 tri



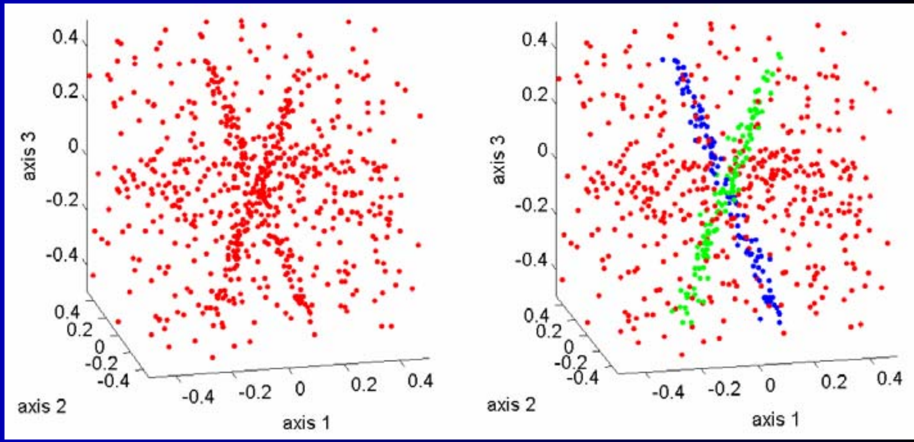
(a) $(2, 1, 1)$ in $\mathbb{R}^3$ (b) $(2, 2, 1)$ in $\mathbb{R}^3$ (c) 3 Gaussians in $\mathbb{R}^2$

SIMULATIONS – Robustness with Outliers

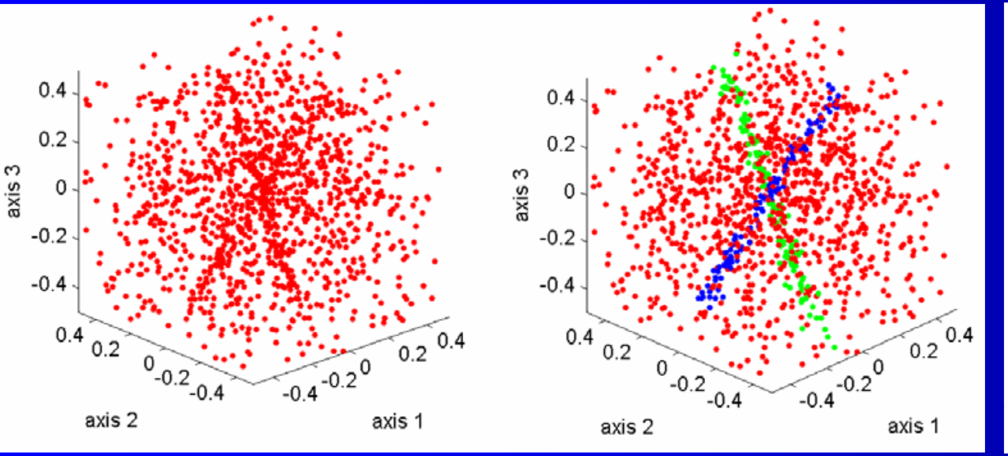
35.8% outliers



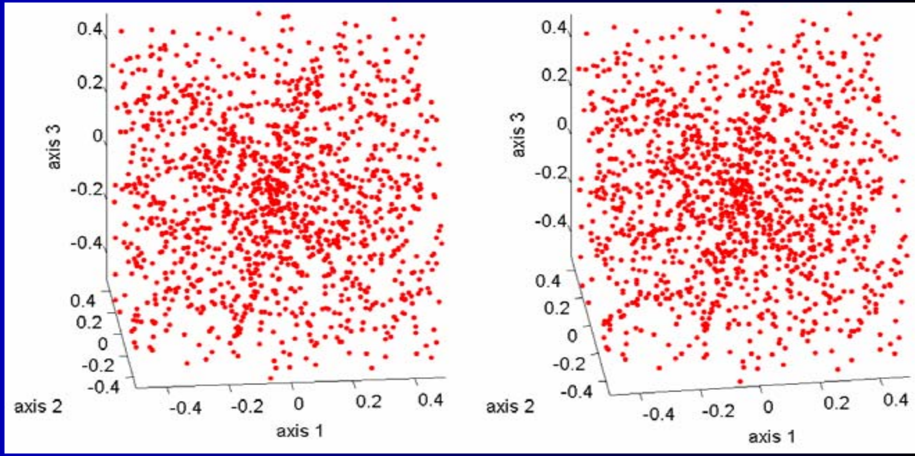
45.6%



71.5%

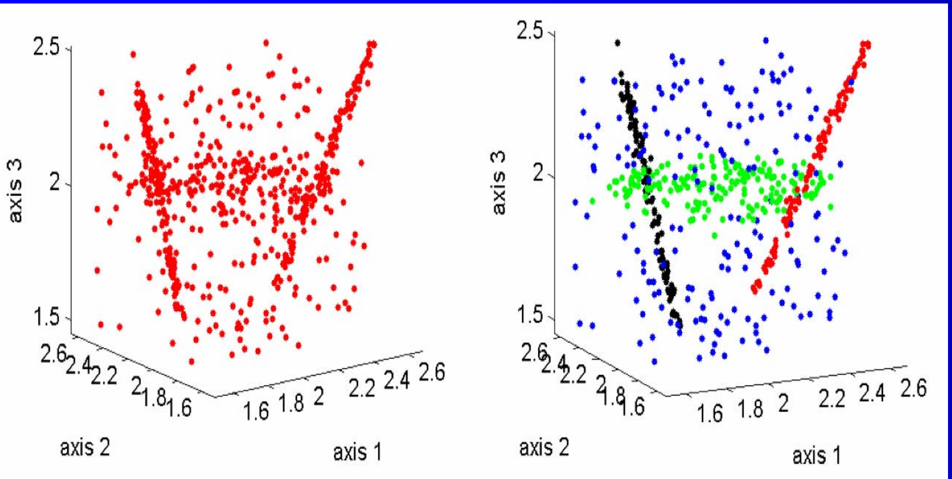


73.6%

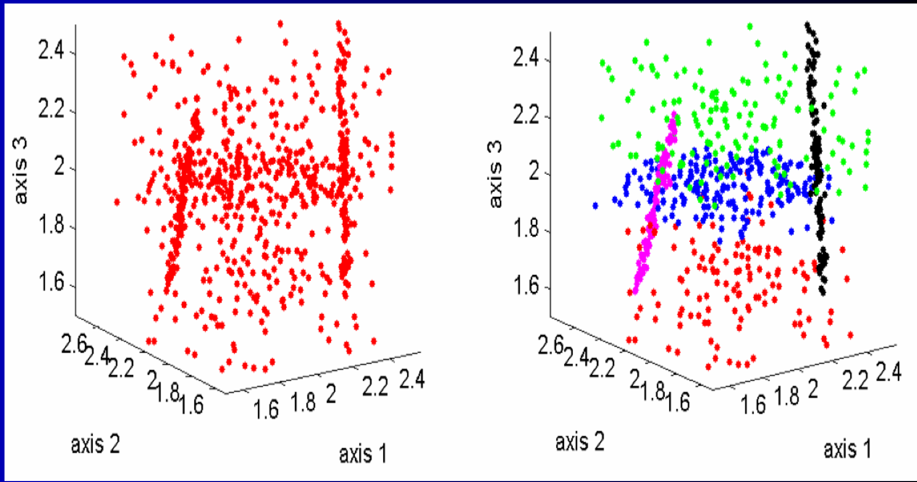


SIMULATIONS – Affine Subspaces with Outliers

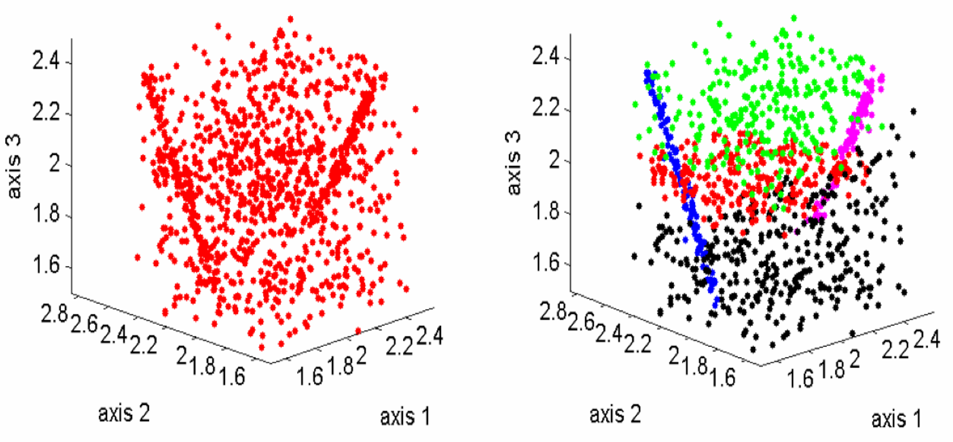
35.8% outliers



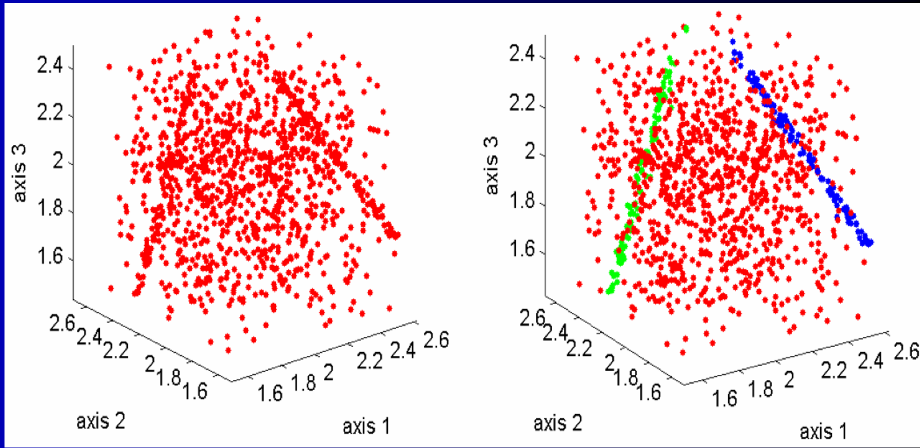
45.6%



66.2%



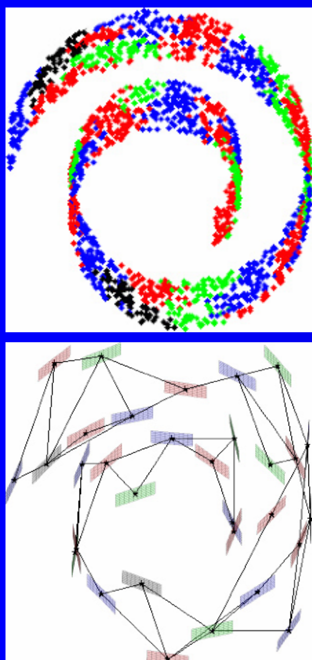
69.1%



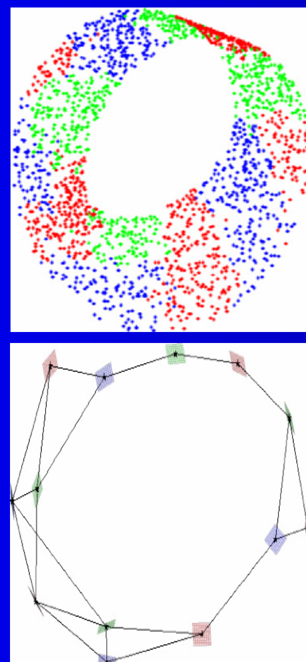
Segmentation of Multivariate Mixed Data via Lossy Coding and Compression, [Ma-Derksen-Hong-Wright, PAMI'07]

SIMULATIONS – Piecewise-Linear Approximation of Manifolds

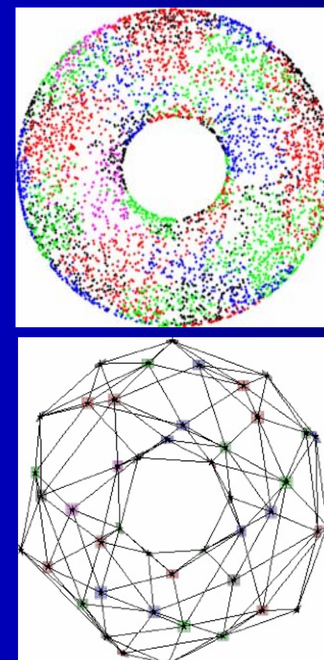
Swiss roll



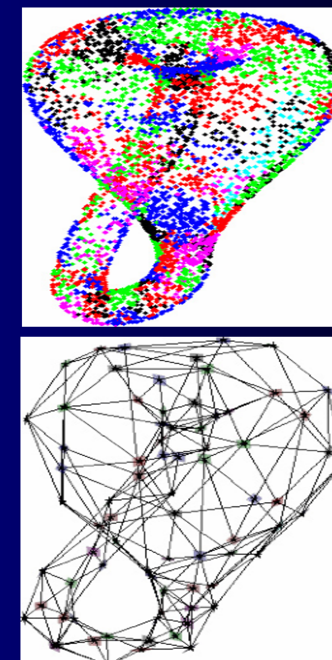
Mobius strip



Torus



Klein bottle



SIMULATIONS – Summary

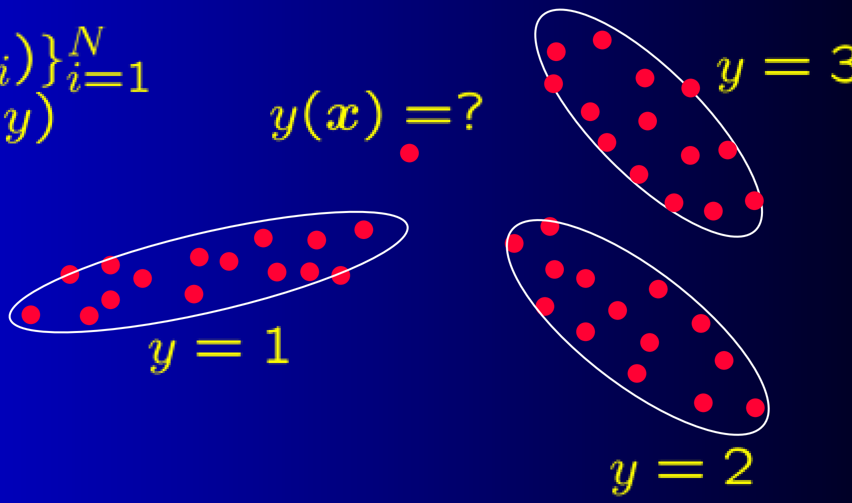
– The minimum coding length objective automatically addresses the
model selection issue: the optimal solution is very **stable and robust.**

– The segmentation/merging is physically meaningful (measured in bits).
The results resemble **phase transition** in statistical physics.

– The greedy algorithm is **scalable** (polynomial in both K and N) and
converges well when ε is not too small w.r.t. the sample density.

Clustering from a Classification Perspective

Assumption: The training data $\{(x_i, y_i)\}_{i=1}^N$ are drawn from a distribution $p_{X,Y}(x, y)$



Goal: Construct a classifier $\hat{y}(x)$ such that the misclassification error

$$E[I_{\hat{y}(X) \neq Y}]$$

reaches minimum.

Solution: Knowing the distributions $p_{X,Y}(x, y)$ and $p_Y(y)$, the optimal classifier is the maximum a posteriori (MAP) classifier:

$$\hat{y}(x) = \arg\max_y \ln p_{X|Y}(x|y) + \ln p_Y(y)$$

Difficulties: How to learn the two distributions $p_{X,Y}, p_Y$ from samples? (parametric, non-parametric, model selection, high-dimension, outliers…)

MINIMUM INCREMENTAL CODING LENGTH – Problem Formulation

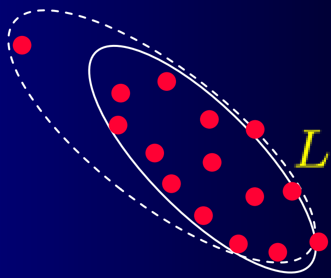
Ideas: Using the lossy coding length

$$L_\epsilon(X) = \frac{N+K}{2}\log_2\det\left(I + \frac{K}{\epsilon^2 N}\bar{X}\bar{X}^T\right) + \frac{K}{2}\log_2\left(1+\frac{\mu^T\mu}{\epsilon^2}\right)$$

as a surrogate for the Shannon lossless coding length w.r.t. true distributions.

Additional bits need to encode the test sample x with the jth training set is

$$\delta L_\epsilon(x,j) = L_\epsilon(\mathcal{X}_j \cup \{x\}) - L_\epsilon(\mathcal{X}_j) + L(j)$$



Classification Criterion: Minimum Incremental Coding Length (MICL)

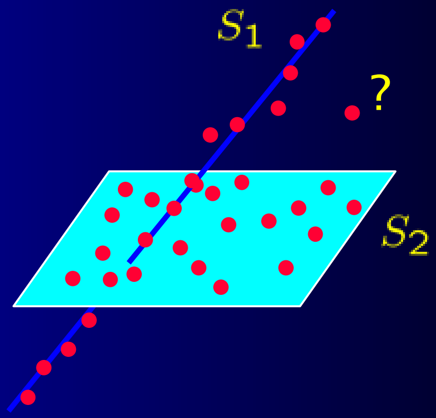
$$\boxed{\hat{y}(x) = \arg\min_j \ \delta L_\epsilon(x,j)}$$

MICL (“Michael”) – Asymptotic Properties

Theorem: As the number of samples N goes to infinity, the MICL criterion converges with probability one to the following criterion:

$$\hat{y}_\epsilon(x) = \arg\max_j \mathcal{L}_G\left(x \,\middle|\, \mu_j, \Sigma_j + \frac{\epsilon^2}{K} I\right) + \ln \pi_j + \frac{1}{2} D_\epsilon(\Sigma_j),$$

where $D_\epsilon(\Sigma_j) \doteq \mathrm{trace}\Sigma_j \left(\Sigma_j + \frac{\epsilon^2}{K} I\right)^{-1}$



$D_\epsilon(\Sigma_j)$ is the “number of effective parameters” of the j-th model (class).

Theorem: The MICL classifier converges to the above asymptotic form at the rate of $c \cdot N^{-\frac{1}{2}}$ for some constant c.

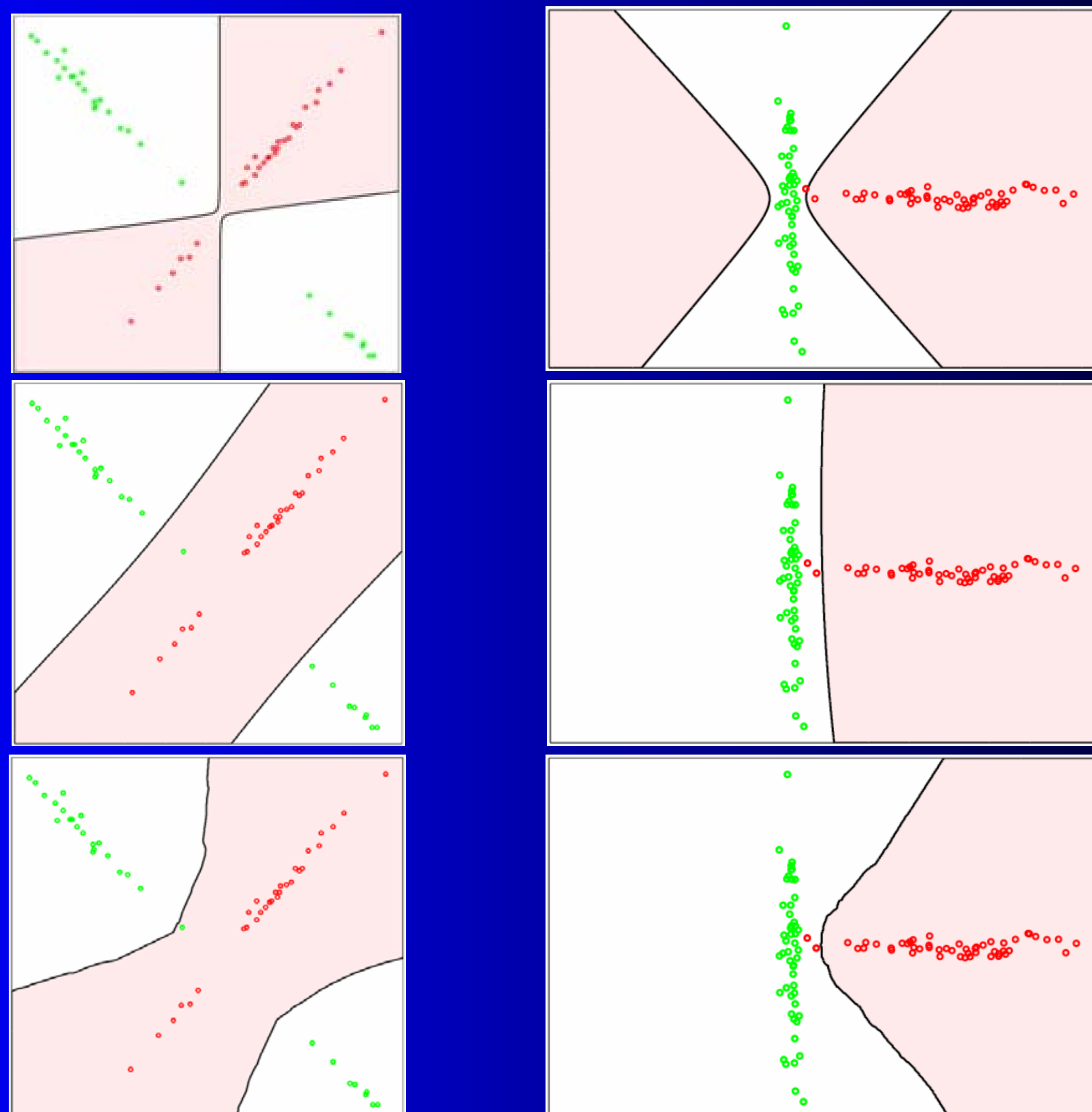
Minimum Incremental Coding Length (MICL), [Wright and Ma et. a.., NIPS’07]

SIMULATIONS – Interpolation and Extrapolation via MICL

MICL

SVM

k-NN



Minimum Incremental Coding Length (MICL), [Wright and Ma et. a.., NIPS’07]

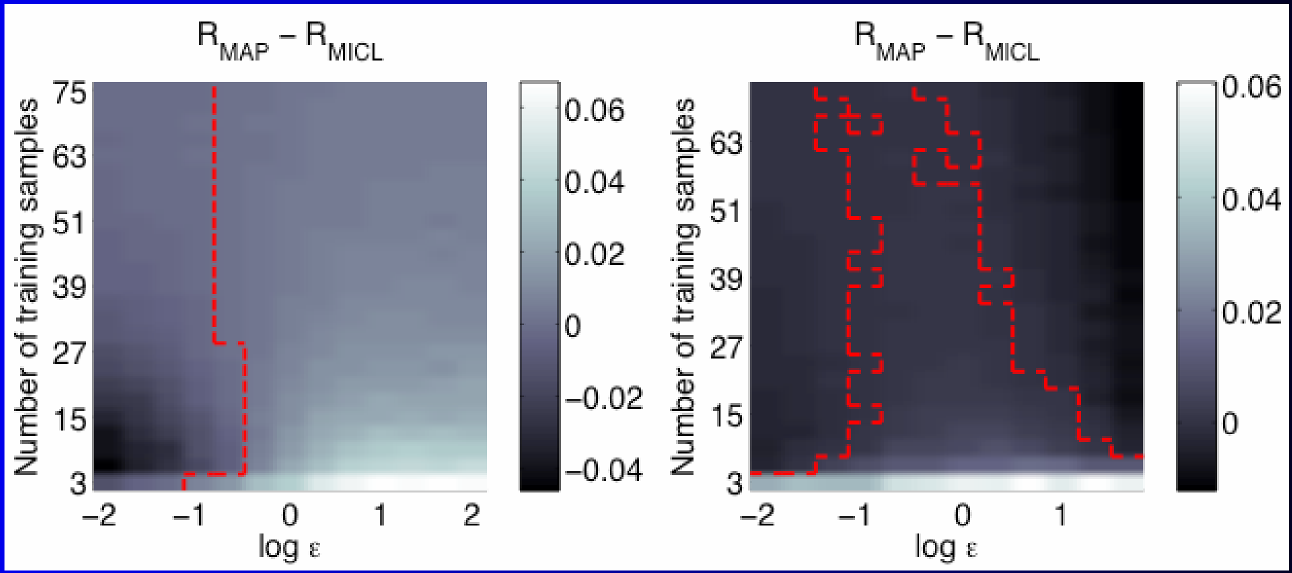
SIMULATIONS – Improvement over MAP and RDA [Friedman1989]

Two Gaussians in $\mathcal{R}^2$

isotropic (left)

anisotropic (right)

(500 trials)



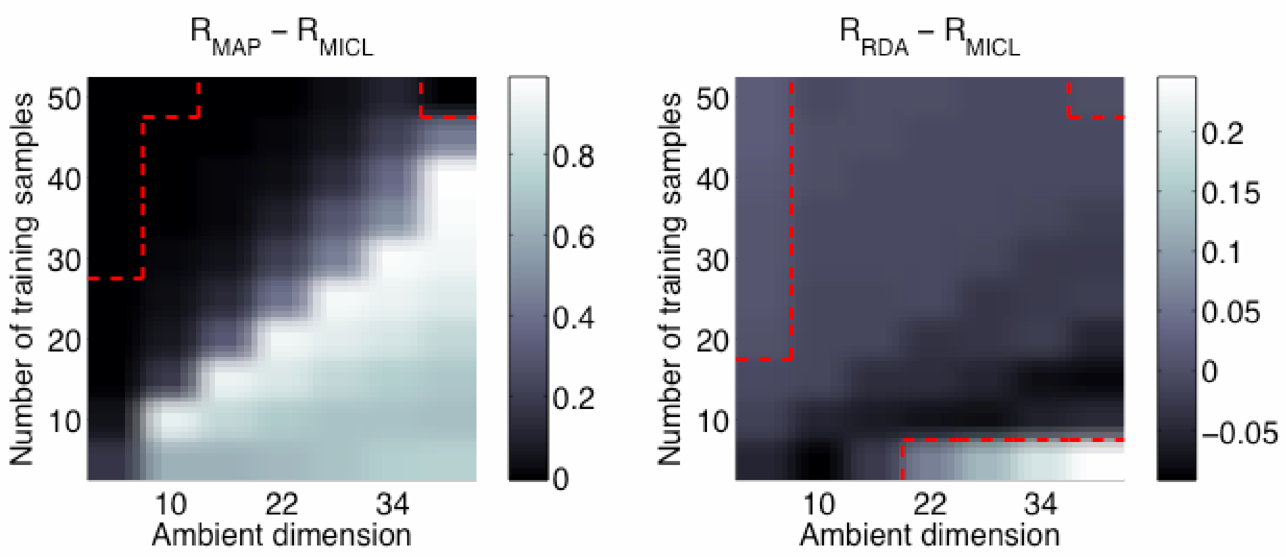
Three Gaussians in $\mathcal{R}^n$

dim = n

dim = n/2

dim = 1

(500 trials)



Minimum Incremental Coding Length (MICL), [Wright and Ma et. a.., NIPS'07]

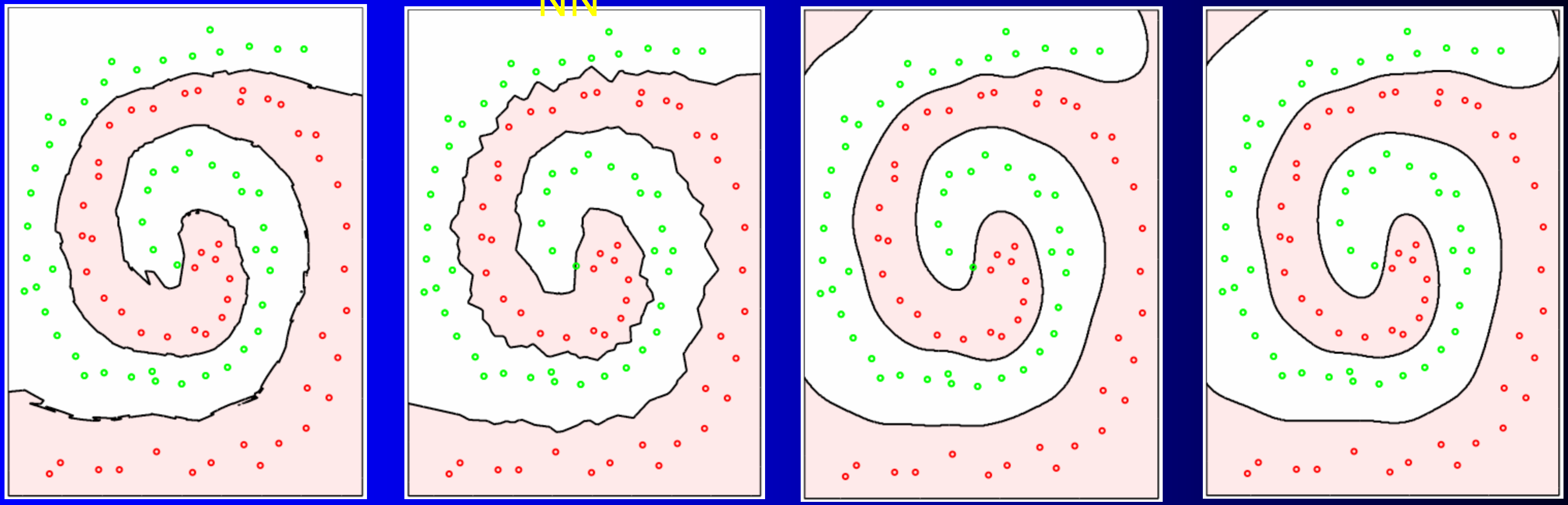
SIMULATIONS – Local and Kernel MICL

Local MICL (LMICL): Applying MICL locally to the k-nearest neighbors of the test sample (frequencylist + Bayesianist).

Kernel MICL (KMICL): Incorporating MICL with a nonlinear kernel naturally through the identity ("kernelized" RDA):

$$\det\left(I + \frac{K}{\epsilon^2 N} X X^T\right) = \det\left(I + \frac{K}{\epsilon^2 N} X^T X\right).$$

LMICL k-NN KMICL-RBF SVM-RBF



Minimum Incremental Coding Length (MICL), [Wright and Ma et. a.., NIPS'07]

CONCLUSIONS

- Assumptions: Data are in a high-dimensional space but have low-dimensional structures (subspaces or submanifolds).
- Compression => Clustering & Classification:
  - Minimum (incremental) coding length subject to distortion.
  - Asymptotically optimal clustering and classification.
  - Greedy clustering algorithm (bottom-up, agglomerative).
  - MICL corroborates MAP, RDA, k-NN, and kernel methods.
- Applications (Next Lectures):
  - Video segmentation, motion segmentation (Vidal)
  - Image representation & segmentation (Ma)
  - Others: microarray clustering, recognition of faces and handwritten digits (Ma)

FUTURE DIRECTIONS

- Theory
  - More complex structures: manifolds, systems, random fields…
  - Regularization (ridge, lasso, banding etc.)
  - Sparse representation and subspace arrangements
- Computation
  - Global optimality (random techniques, convex optimization…)
  - Scalability: random sampling, approximation…
- Future Application Domains
  - Image/video/audio classification, indexing, and retrieval
  - Hyper-spectral images and videos
  - Biomedical images, microarrays
  - Autonomous navigation, surveillance, and 3D mapping
  - Identification of hybrid linear/nonlinear systems

REFERENCES & ACKNOWLEGMENT

- References:
  - *Segmentation of Multivariate Mixed Data via Lossy Data Compression*, Yi Ma, Harm Derksen, Wei Hong, John Wright, PAMI, 2007.
  - *Classification via Minimum Incremental Coding Length (MICL)*, John Wright et. al., NIPS, 2007.
  - Website: http://perception.csl.uiuc.edu/coding/home.htm

- People:
  - **John Wright**, PhD Student, ECE Department, University of Illinois
  - **Prof. Harm Derksen**, Mathematics Department, University of Michigan
  - **Allen Yang** (UC Berkeley) and **Wei Hong** (Texas Instruments R&D)
  - **Zhoucheng Lin and Harry Shum**, Microsoft Research Asia, China

- Funding:
  - ONR YIP N00014-05-1-0633
  - NSF CAREER IIS-0347456, CCF-TF-0514955, CRS-EHS-0509151

"*The whole is more than the sum of its parts*."

-- Aristotle

Questions, please?