Complexity, Information and Geometry

Peyresque July 2008

Alfred Hero
Digiteo and University of Michigan

1. Overview
2. Motivation: Topological inference

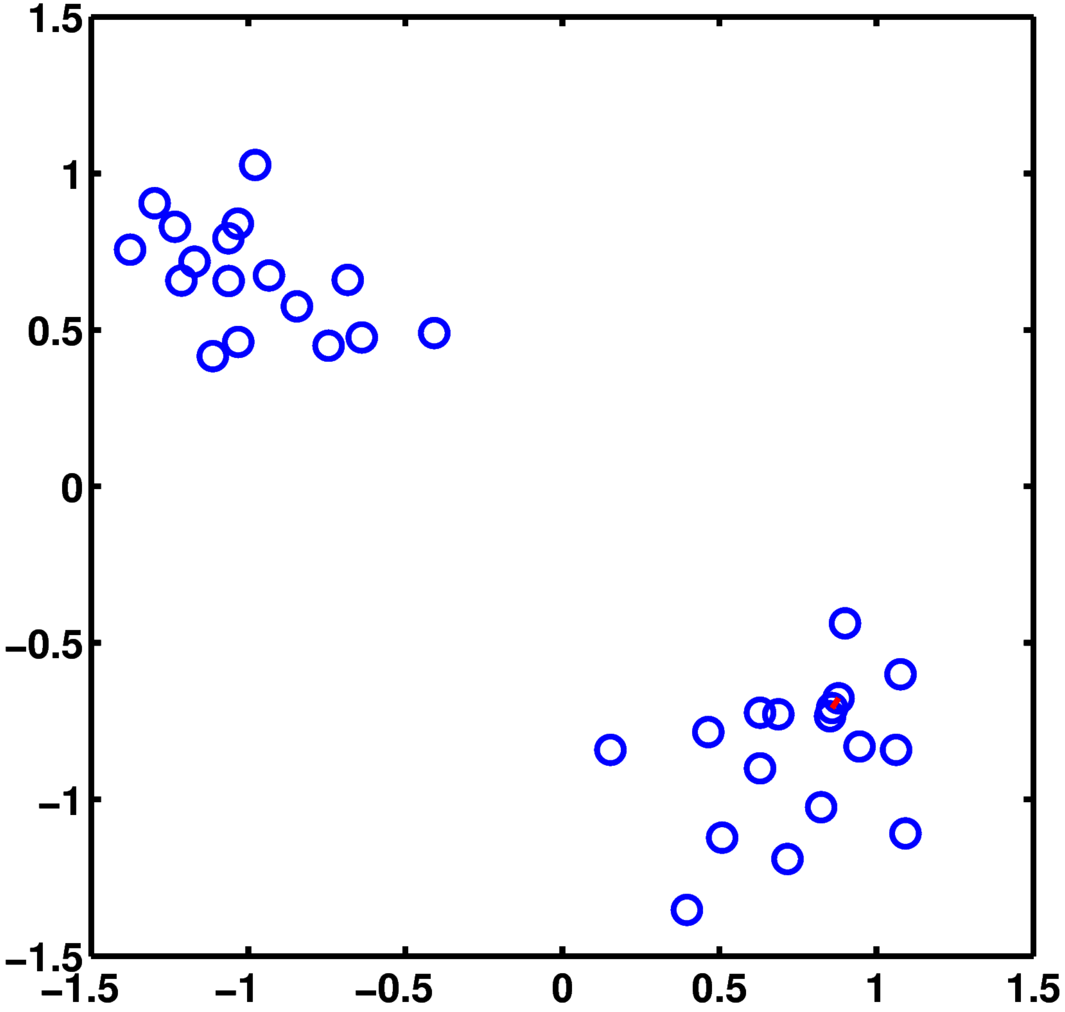
1. Overview of cursus

- Overview and motivation (Now)
- Complexity and Information
- Entropy estimation
- Random geometric graphs
- Applications
    - Dimension estimation
    - Anomaly detection
    - Pattern matching and image registration

2. Motivation: Topological inference

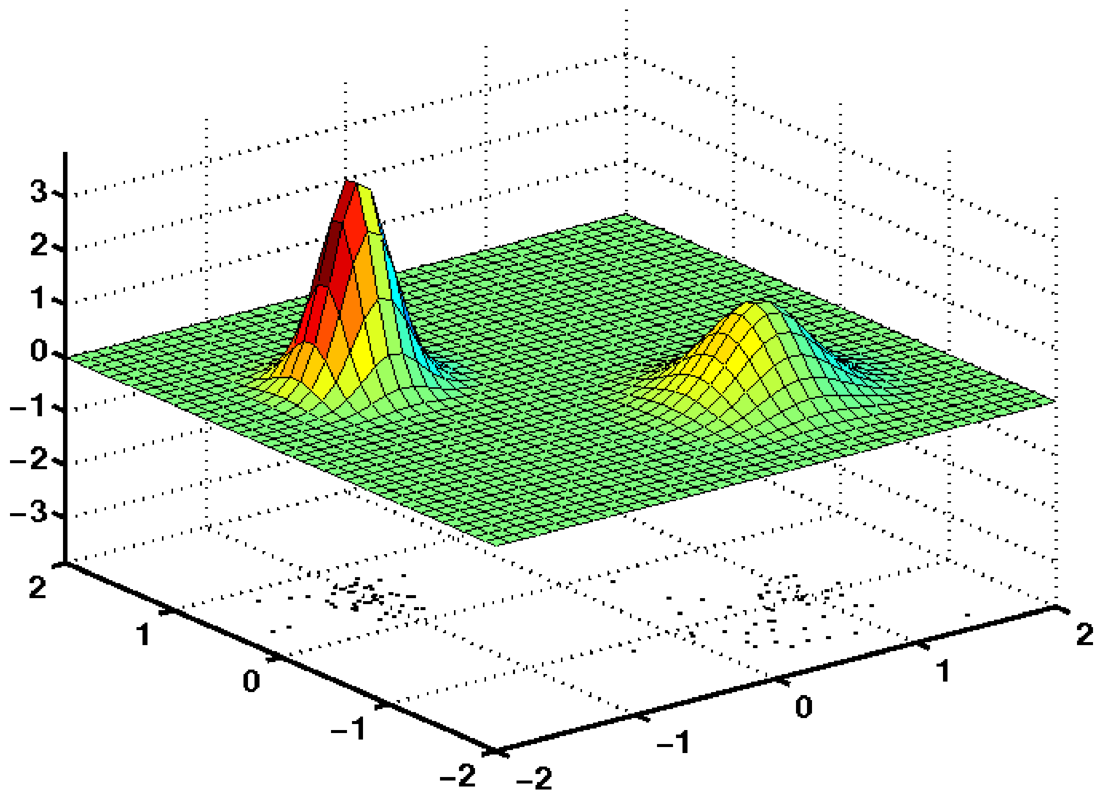
- Three examples
- Structured vs unstructured inference
- Benefit of integrated approaches

A 2D dataset



A 2D dataset

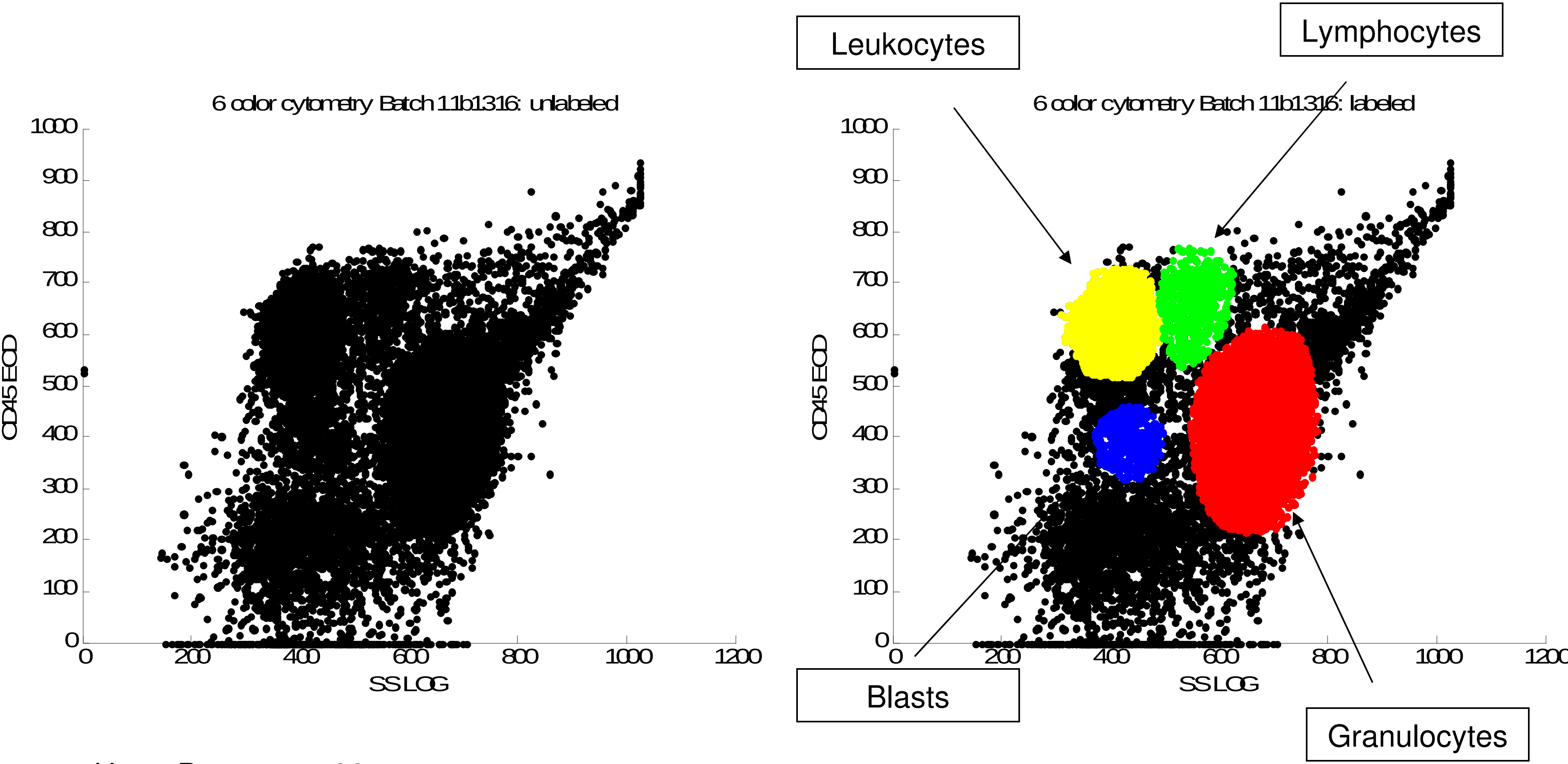
- A simple fit
  - 5 parameter 2 component Gaussian mixture model



Hero. Peyres

I. Fairly low dimensional data: Flow cytometry

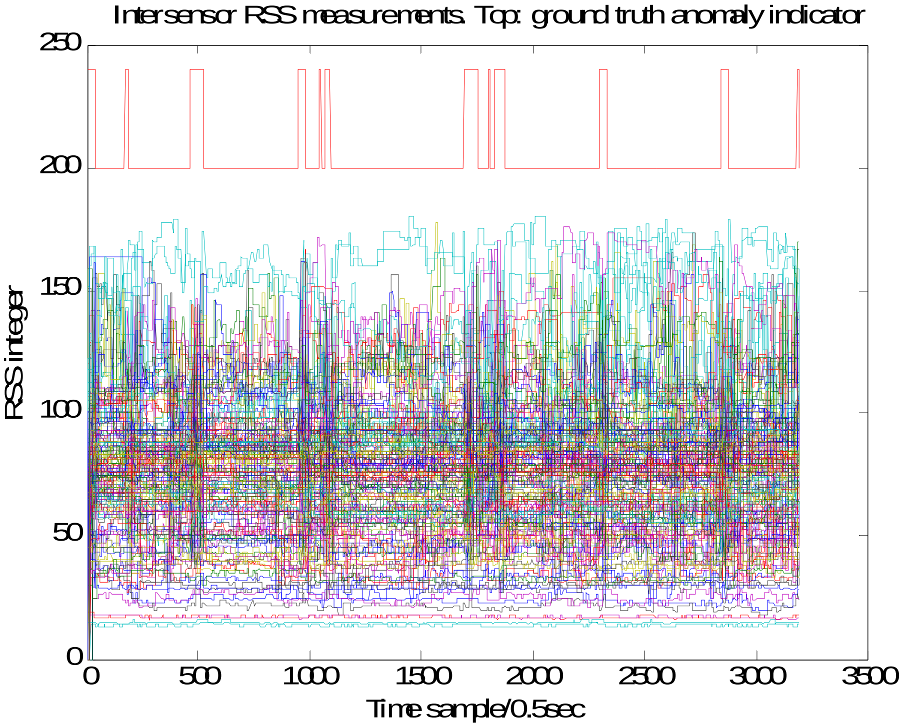
- One 2D projection of 6 color flow cytometry data – N = 30,000 (UM Hemopathology Lab – Dr. W. Finn)



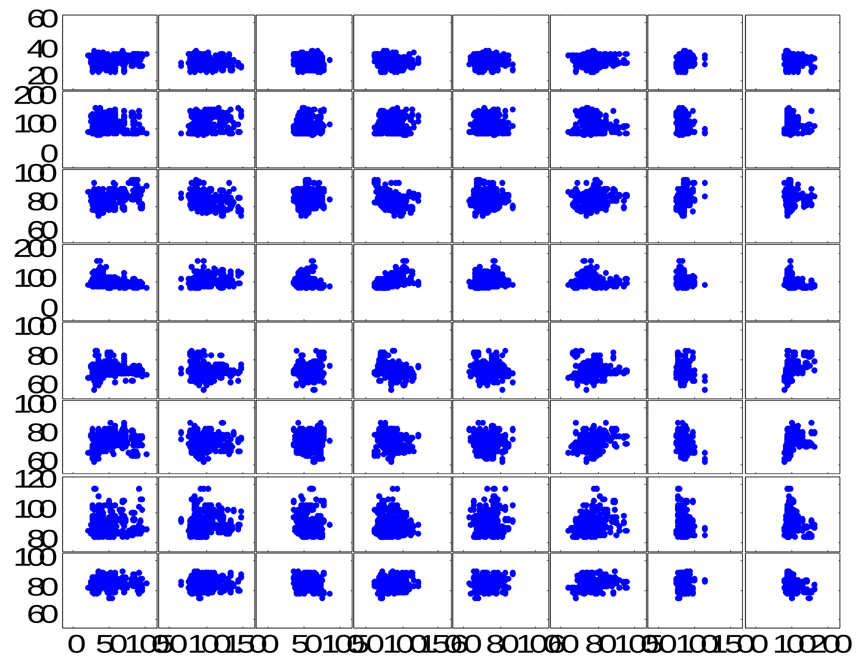
Hero. Peyresque 08

I. Higher dimensional data: Wireless sensor networks

- 13x14=182 dimensional RSS data sample collected from 14 node UM wireless sensor network – N=3500



Sample trajectories over time



64 2D projections of 182 dimensional data

NIPS-07

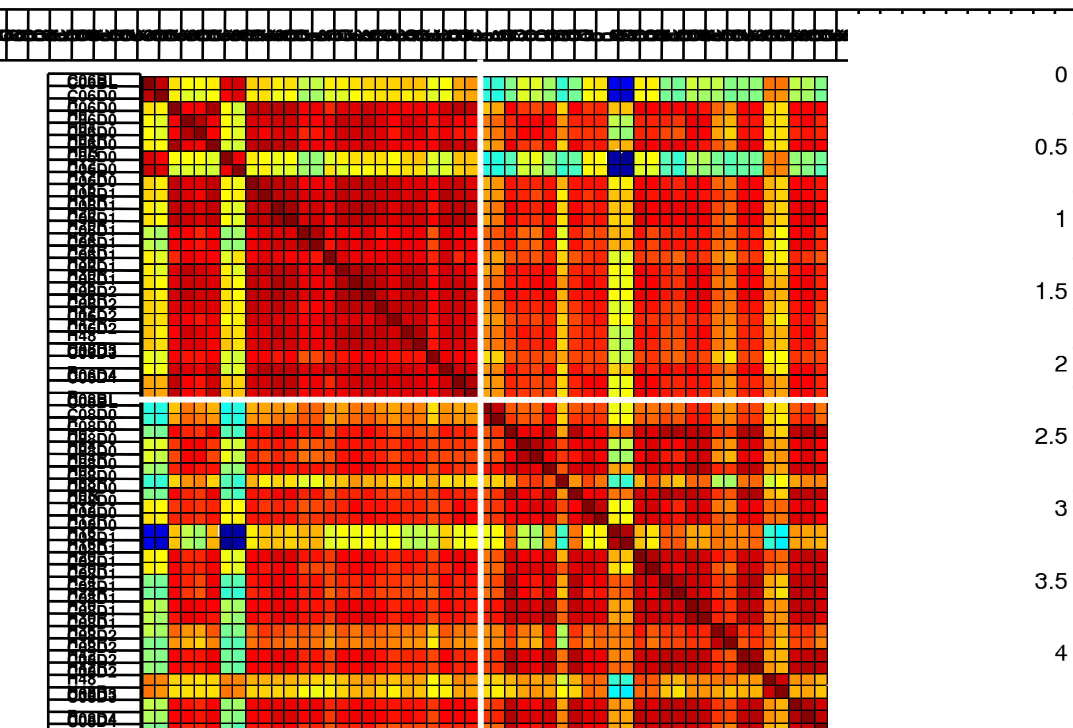
Geometric entropy minimization (GEM) for anomaly detection and localization

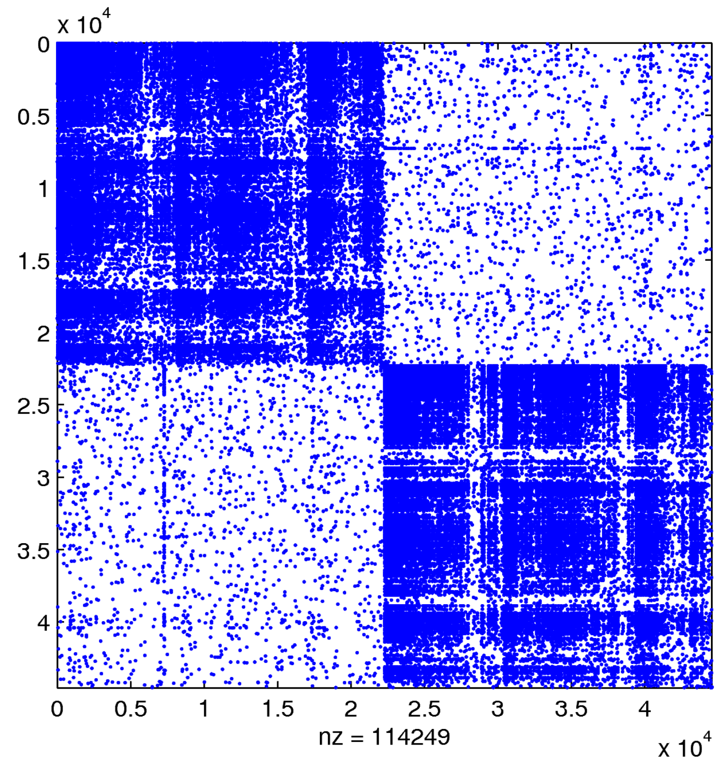
Alfred O Hero, III
University of Michigan, Ann Arbor, MI 48109-2122

NIPS 2006

I. Even Higher Dimensional Data: High throughput genomic time series

- 280 peripheral blood samples of 20 individuals at 14 timepoints.
- mRNA, metabolite, protein, and antibody assays at each time point (24,000 probe dimensions)

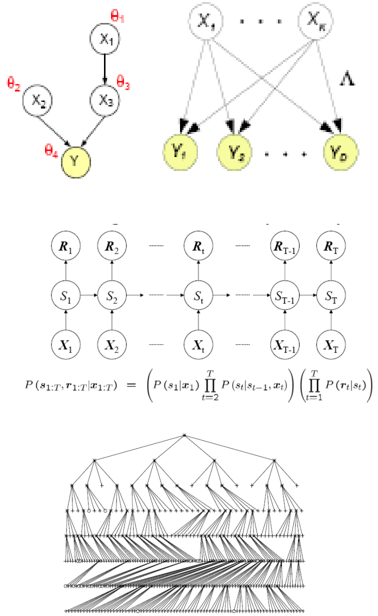




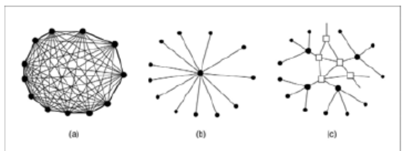
II. Structured vs Unstructured Modeling

- Structured modeling
  - Estimation involves fitting parametric model to a data sample
    - Frequentist parametric models and Fisher's ML principle (Fisher25)
    - Bayesian parametric models and minimum risk estimation (Jeffreys39)
    - Minimum distance parameter estimation (Beran77)
  - (Likelihood and prior) models include
    - Exponential families of densities (Lehman57)
    - Finite mixtures of densities (MclachlanBasford88)
    - Graphical models (Lauritzen96)
- Unstructured modeling
  - Estimation is performed directly on the density in data space
    - Nearest neighbor density estimators (FixHodges51)
    - Partitioning density estimators (NobelLugosi96)
  - Models include
    - Multiscale density representations (WadaSato90)
    - Topological density representations (Wishart69)
    - Cluster tree density representations (Hartigan75)

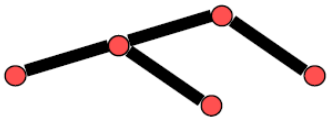
II. Structured graphical model



Structural Priors

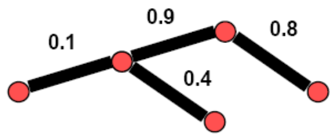


Topologies



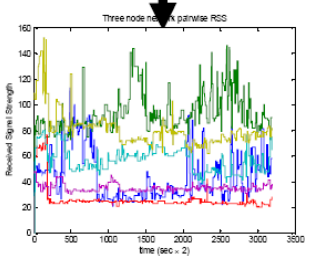
Link priors → **Link jpdf**

Link parameters



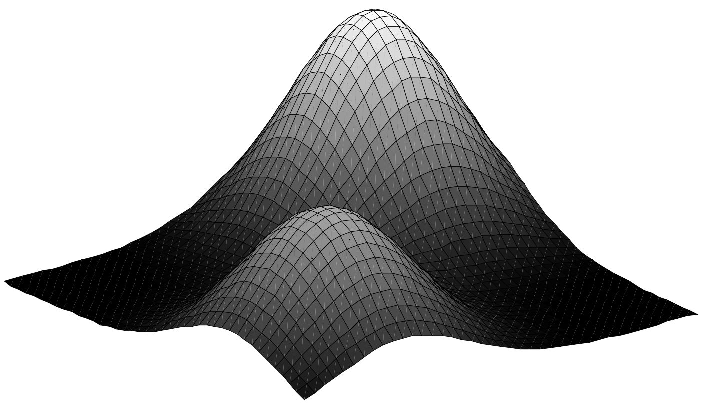
Temporal model → **Observation jpdf**

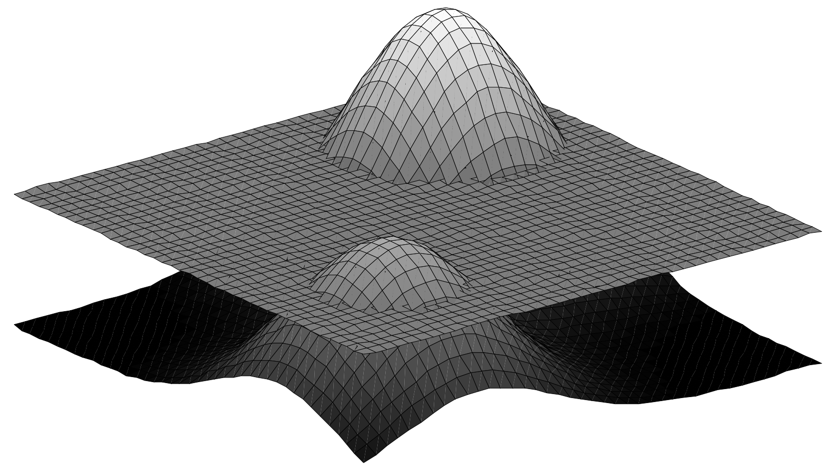
Observations



- Structural priors obtained from context (small world, star, etc)
- Link weight distribution conditioned on
  - Structural parameters
  - Link weight prior

II. Unstructured topological model

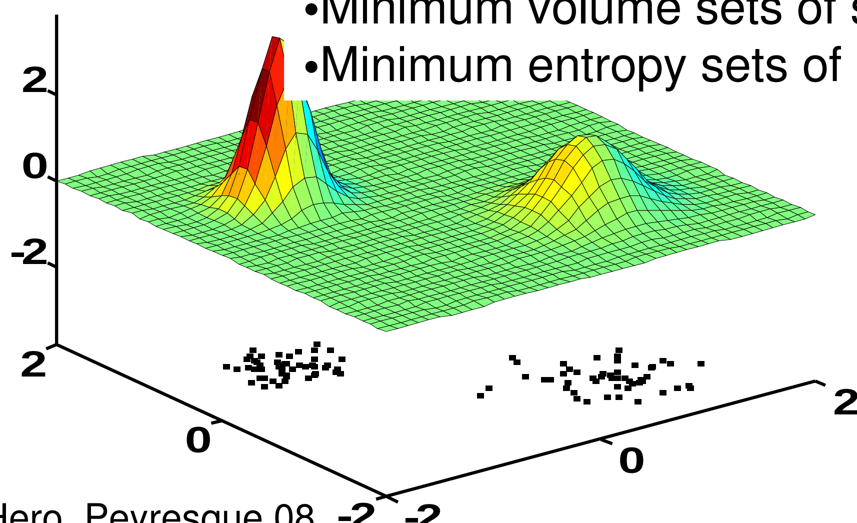


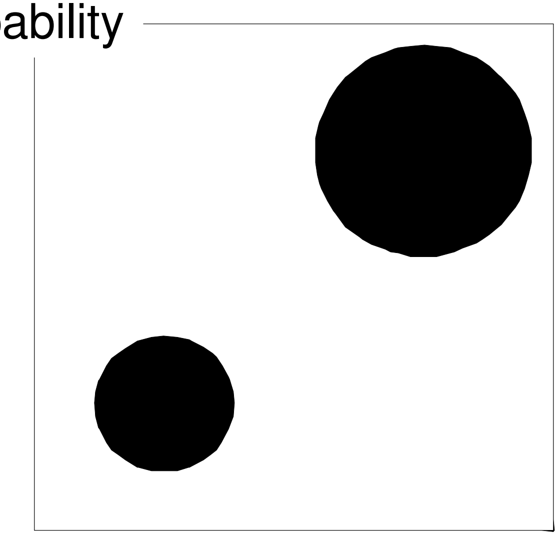


- Density function f(x)

- Cutting plane
- Epigraph sets

These level sets are
•Minimum volume sets of specified probability $= \{x : f(x) \geq l\}$
•Minimum entropy sets of specified probability





II. Toolkit for graphical modeling

- Factored representation of density (factor graphs)

$$f(y_1, y_2, y_3) = f_a(y_1, y_3) f_b(y_1, y_3) f_c(y_1, y_2)$$

- Mixture representation of density (Hidden variable models)

$$f(y_1, y_2, y_3) = \sum_{i=1}^{3} \theta_i \phi_i(y_1, y_2, y_3)$$

- Parameter and structure estimation with EM, variational bayes, MCMC, dependency tests

II. Toolkit for topological modeling

- Density cluster tree representations

Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample

Werner Stuetzle *
Department of Statistics
University of Washington
wxs@stat.washington.edu

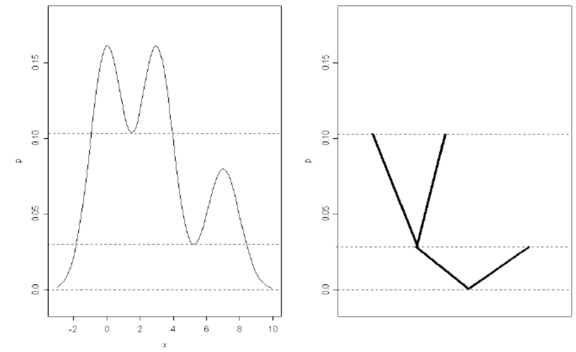


Figure 2: Density and corresponding tree of high density clusters.

- Morse-Smale represention

Maximizing Adaptivity in Hierarchical Topological Models Using Extrema Trees

Peer-Timo Bremer[1], Valerio Pascucci[2], and Bernd Hamann[3]

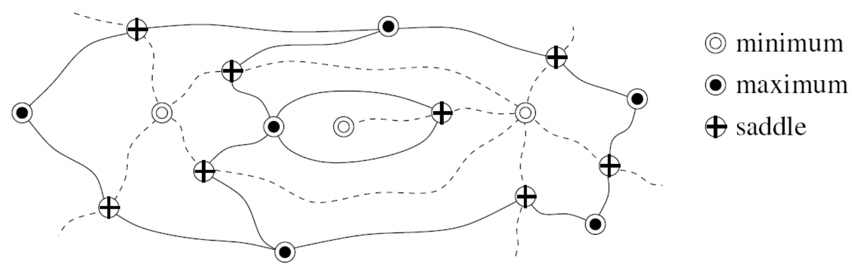


Fig. 1. Morse-Smale complex.

- Non-parametric support set and level set estimation, manifold learning, dimension estimation, entropic graphs, modal analysis

II. Modal analysis

- Wishart (69): one level mode analysis
  - High density clusters extracted from a high level set: single linkage clustering on a level set of kernel density estimator
- Hartigan ('75): multi-level hierarchical mode analysis
  - Concept of density cluster tree introduced
- Software developed: dBScan ('90), OPTICS ('99)
- Stuetzle ('03): mode analysis by runt size pruning
  - NN density estimator and pruned MST are equivalent

III. Benefit of an integrated approach

- Structured model describes broad class of densities with relatively few parameters but suffers from mismodeling errors and bias, especially when there are dimension degeneracies.
- Unstructured model describes general properties of the class of densities without explicit parameterization but suffers from high variance.
- We can bound the theoretical performance gains

III. Theory predicts significant gains

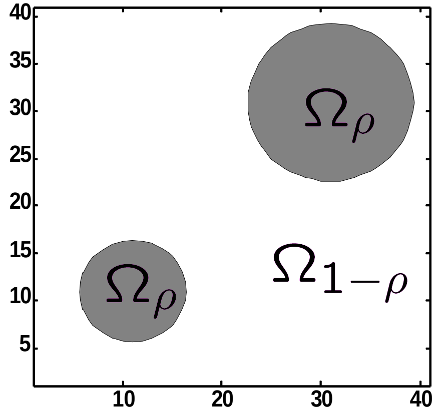
- Achievable minimax density estimation rates over smooth Holder class of densities over $\mathcal{R}^d$ (Tsybakov 1991)

$$\min_{\hat{f}} \max_{f \in \mathcal{F}(\beta,K)} \|f - \hat{f}\| = O\left(n^{-\beta/(2\beta+d)}\right)$$

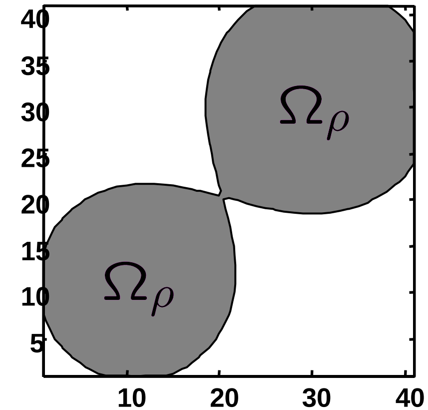
$$\mathcal{F}(\beta, K) = \left\{ f : \|f(z) - p_x^{\lfloor \beta \rfloor}(z)\| \leq K \|x - z\| \right\}$$

- Exponential convergence rate advantage if density were restricted to known domain of dimension m<d

III. Theory predicts significant gains



$\rho = 0.4$



$\rho = 0.8$

- Relative rates inside and outside level set $\Omega_\rho,\ \rho > 0.5$

$$\min_{\hat f} \max_{f \in \mathcal{F}(\beta,K)} \|f - \hat f\| \;=\; n^{-\beta/(2\beta+d)}$$

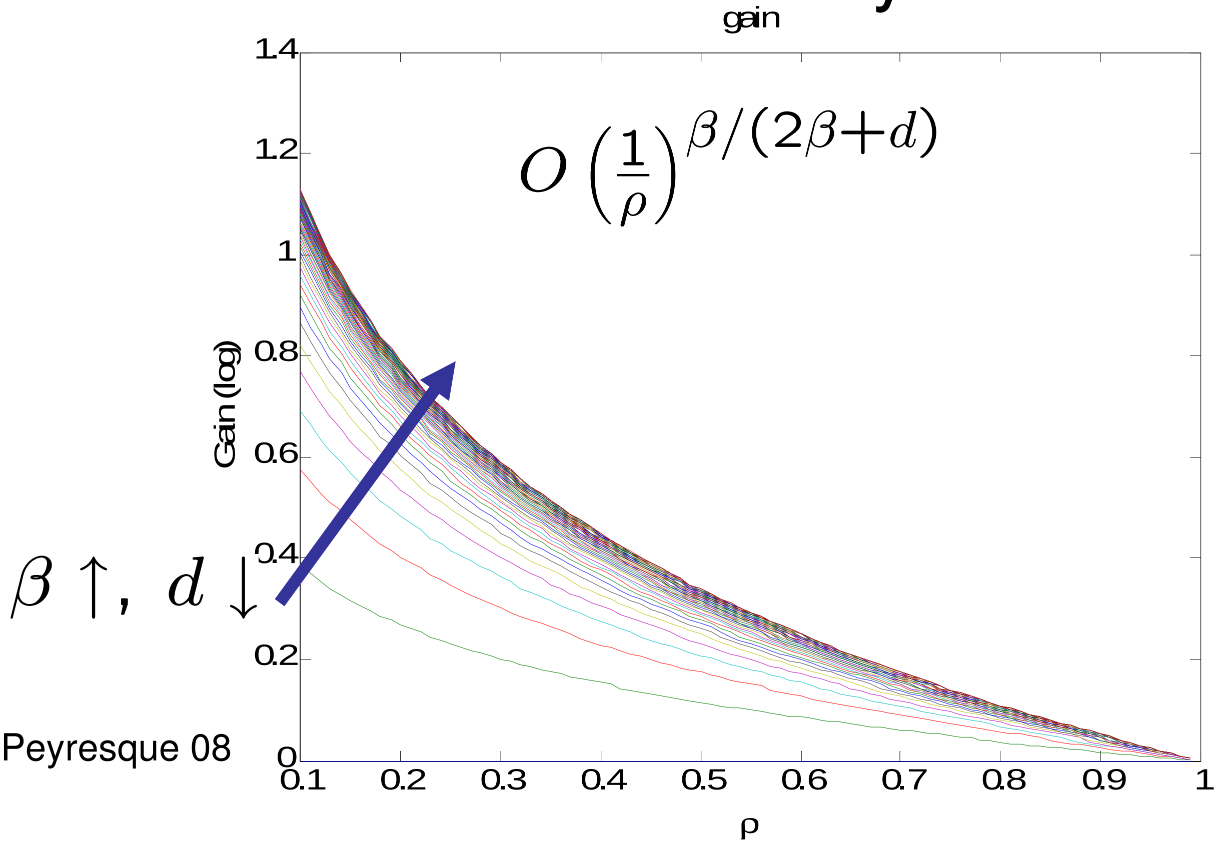
$$\min_{\hat f} \max_{f \in \mathcal{F}(\beta,K)} \|f_{\Omega_\rho} - \hat f\| \;=\; (\rho n)^{-\beta/(2\beta+d)}$$

$$\min_{\hat f} \max_{f \in \mathcal{F}(\beta,K)} \|f_{\Omega_{1-\rho}} - \hat f\| \;=\; ((1-\rho)n)^{-\beta/(2\beta+d)}$$

- Faster convergence occurs inside level set
- Density level set can be estimated with rate no worse than $n^{-\beta/(2\beta+d-2)}$ (Willett&Nowak:2006)

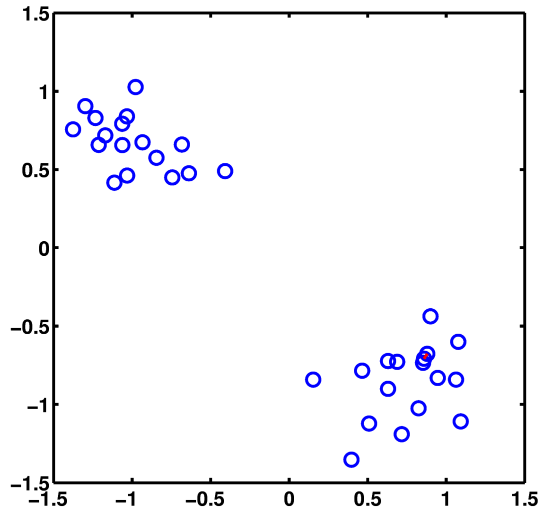
III. Level set screening advantage

- Remark: restricting inference to level set avoids poorest regions of the data sample
- Known level set boundary: $\text{Gain} = O(n^{d/(2\beta+d)})$
- Unknown level set boundary:



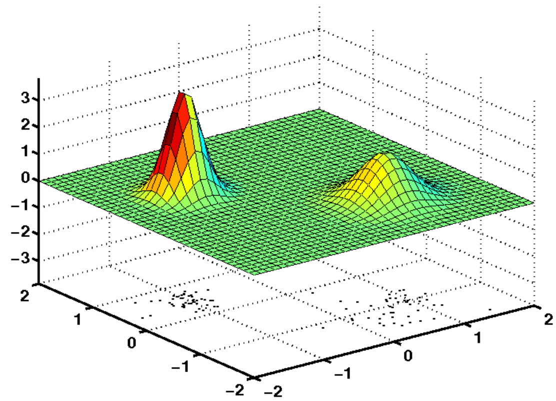
Hero. Peyresque 08

A 2D dataset



A 2D dataset

- A simple fit
  - –5 parameter 2 component Gaussian mixture model



Outlier detection

How to find the “critical region” for testing whether a test point is an outlier

Classical parametric and non-parametric models try to

ii) Fit a smooth function to data

- interpolate the data
- do model checking via statistical tests

WS – Wada & Sato 1990 (ICCPR)

Require upper-semicontinuous function for level sets to exist and epigraph to be bounded