Detection Problems

A Statistical Viewpoint with Multiscale Insight

Ery Arias-Castro
Department of Statistics, Stanford University
http://www-stat.stanford.edu/ acery/Papers/

Non-Destructive Testing (NDT)

- Also
  - Non-Destructive Evaluation (NDE)
  - Non-Destructive Inspection (NDI)
- Uses
  - Quality control
  - Safety
- Applications
  - Aircraft, Motor Vehicles, Trains
  - Pipelines, Oil Platforms, Refineries
  - Bridges, Buildings

- Techniques
  - Visual (optical)
  - Liquid Penetration
  - Acoustic Emission, Ultrasonic
  - Ultra Magnetic Particle, Eddy Current
  - Radiography

Medical Imaging

- Uses
  - Medical Diagnosis
- Techniques
  - Computed Tomography (CT)
  - Magnetic Resonance Imaging (MRI)
  - Ultrasound
  - Positron Emission Tomography (PET)
  - Single Photon Emission Computed Tomography (SPECT)

Satellite Imagery

- Uses
  - Strategic
  - Planning
  - Monitoring
- Applications
  - Road Detection
  - Ship Wake Detection
  - Forest Fires Early Detection
  - Weather Forecast
- Techniques
  - Synthetic Aperture Radar (SAR)
  - Infrared

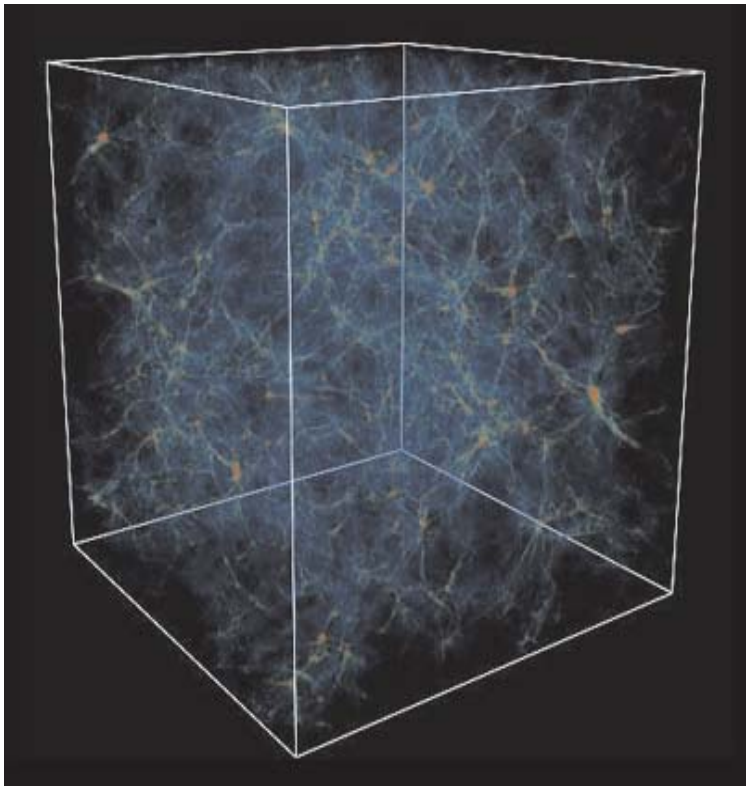
The Challenges of Detection

- Large images (e.g. 3D) – computational burden
- Complex objects – Recognition involved
- Noise – varies with measuring device
- Measure(s) of efficiency

▶ Common to most Image Processing tasks!

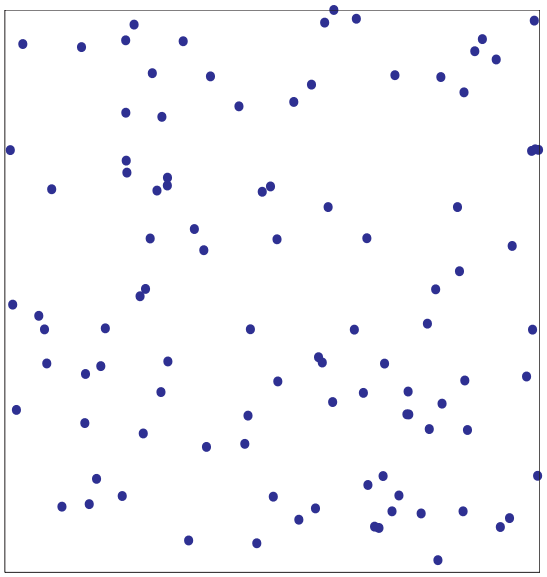
Galaxy Distributions

Finding structures (filaments, sheets) in (large) 3D galaxy catalogs.

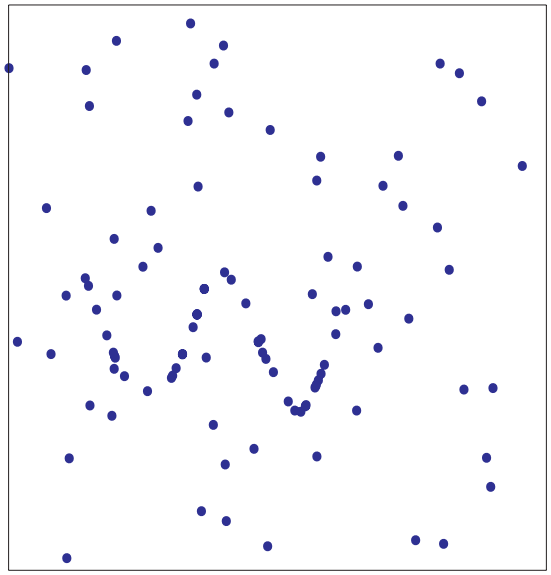


Detection in 2D Point Cloud

Finding a curve in 2D point cloud data.



No curve is present



A curve is present

Hypothesis Testing Problem

H_0 (null hypothesis)

n points are independent, uniformly distributed in the unit square

H_1 (alternative hypothesis)

$n - n_1$ points are independent, uniformly distributed in the unit square

n_1 points are on an unknown curve $\gamma \in \mathcal{C}$

Model for Objects

Objects of interest

$$\mathcal{C} = \{\gamma = f([0,1]) : f = (f_1, f_2), \quad f_1, f_2 \in H(\alpha, \beta)\}$$

where

$$H(\alpha, \beta) = \{f : [0,1] \to [0,1]\},$$

such that

- $|f^{(s)}(x)| \leq \beta$ for all $x \in [0,1]$ and $s = 0, \ldots, \lfloor \alpha \rfloor$
- $|f^{(\lfloor \alpha \rfloor)}(x) - f^{(\lfloor \alpha \rfloor)}(y)| \leq \beta |x - y|^{\alpha - \lfloor \alpha \rfloor}$ for all $x, y \in [0,1]$

Measure of Performance

Detection procedure (test) T:

- $T = 1$ detects an object
- $T = 0$ detects no object

False alarm rate (level): $\mathcal{L}(T) = \mathbf{P}\{T = 1|H_0\}$

Worst-case risk (minimax risk): $\mathcal{R}(T) = \inf_{\gamma \in \mathcal{C}} \mathbf{P}\{T = 1|H_1\}$

Measure of performance: $\mathcal{M}(T) = \mathcal{L}(T) + \mathcal{R}(T)$

Generalized Likelihood Ratio Test

For $S \subset [0,1]^2$, define $N(S)$ number of (data) points in S.

The GLRT rejects for "large" values of

$$N(\mathcal{C}) = \max_{\gamma \in \mathcal{C}} N(\gamma)$$

In other words

- GLRT $= 1$ if $N(\mathcal{C}) > \tau$
- GLRT $= 0$ otherwise

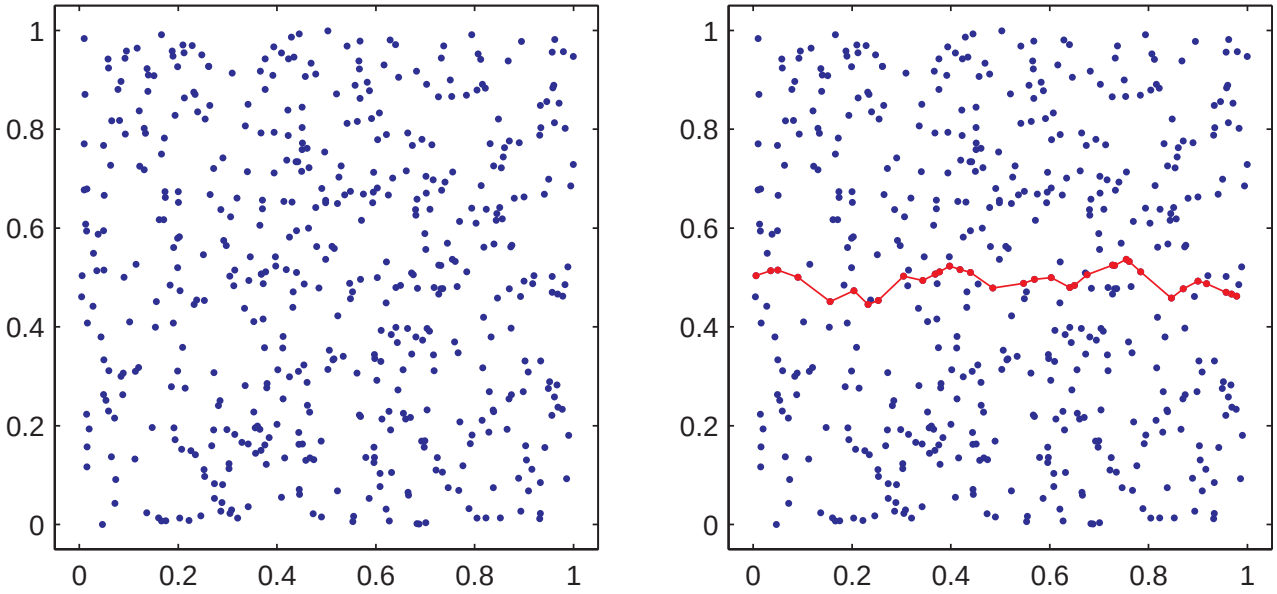
Also known as *Scan Statistics*.

Detection Performance of GLRT

When $n \to \infty$ and a good choice of τ:

- $\mathcal{M}(\text{GLRT}) \to 0$ if $n_1 > Bn^{1/(1+\alpha)}$, B large enough
- $\mathcal{M}(\text{GLRT}) \to 0$ if $n_1 < An^{1/(1+\alpha)}$, A small enough

GLRT in Action



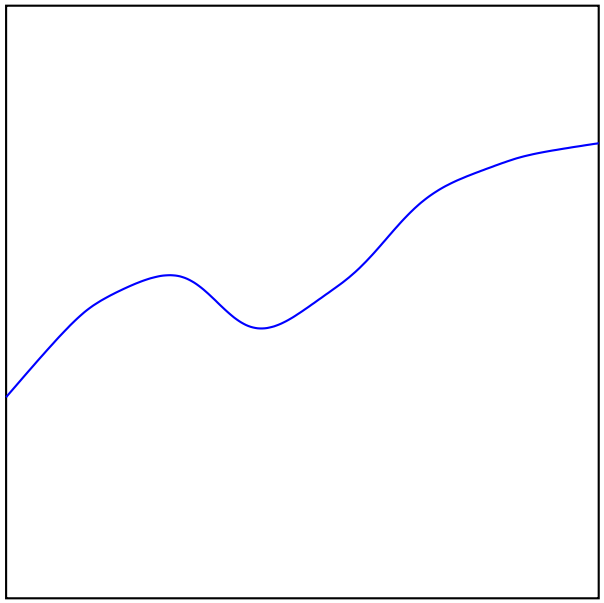
- Here $\mathcal{C}$ is a bounded subset of Lipschitz graphs ($\alpha = 1$)
- In this example, GLRT $= 1$ – GLRT detects the present of a curve in $\mathcal{C}$
- The curve in red maximizes $\{N(\gamma) : \gamma \in \mathcal{C}\}$

Computational Issues

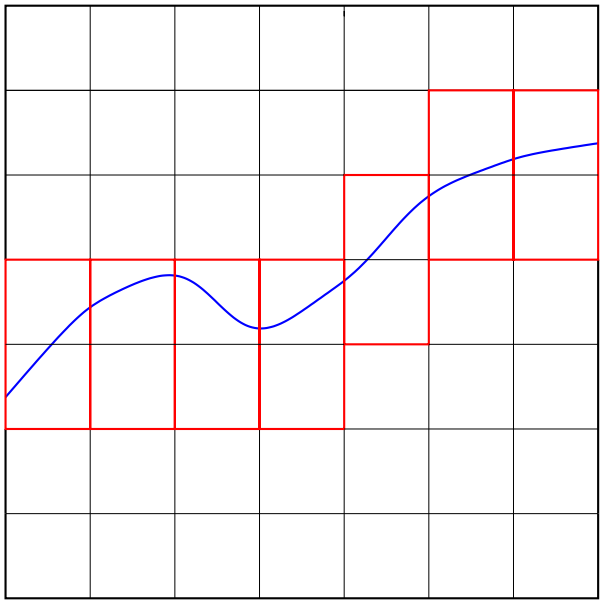
- Only able to compute for curves that are graphs and of specific smoothness ($\alpha \in \{1, 2\}$) – using *Dynamic Programming* (DP)
- DP fails when points are not *exactly* on the curve
- DP fails for detecting higher dimensional objects (e.g. surfaces)

Coverings (Lipschitz, $\alpha = 1$)

Suppose $\mathcal{C}$ is a bounded subset of Lipschitz graphs



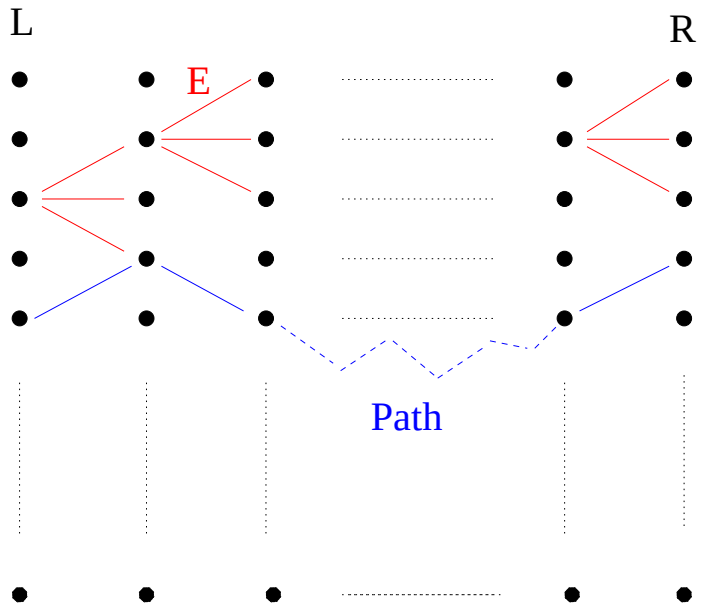
Graph of a Lipschitz function



Corresponding covering

Associated Graphical Structure

To each $\gamma \in \mathcal{C}$, associate a path in graphical structure $\mathcal{G}$ ($\pi(\gamma)$)



Approximation to GLRT

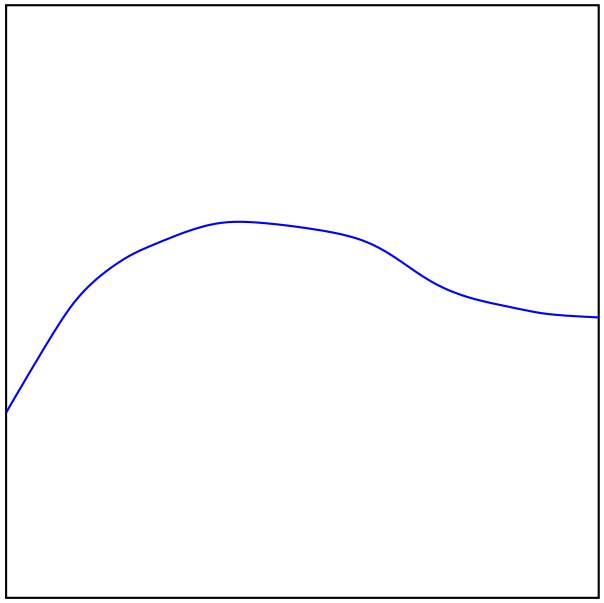
- Because $\gamma \subset \pi(\gamma)$, $N(\gamma) \leq N(\pi(\gamma))$
- Let

$$M(\mathcal{C}) = \max_{\pi \text{path in } \mathcal{G}} N(\pi)$$

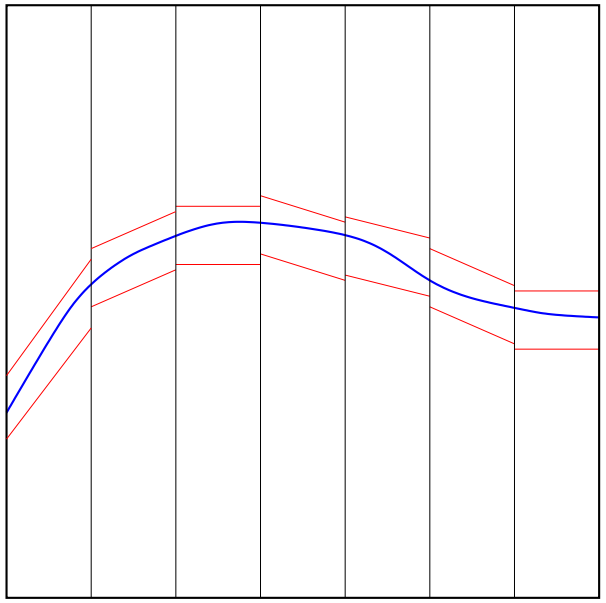
  We have $N(\mathcal{C}) \leq M(\mathcal{C})$
- With good choice of cells, $M(\mathcal{C}) \asymp N(\mathcal{C})$. Use $M(\mathcal{C})$ to approximate $N(\mathcal{C})$. Let GLRT* be the test that rejects when $M(\mathcal{C})$ is large.

Coverings (Hölder, $1 < \alpha \leq 2$)

Suppose $\mathcal{C}$ is a bounded subset of Hölder graphs, with $\alpha \in (1, 2]$

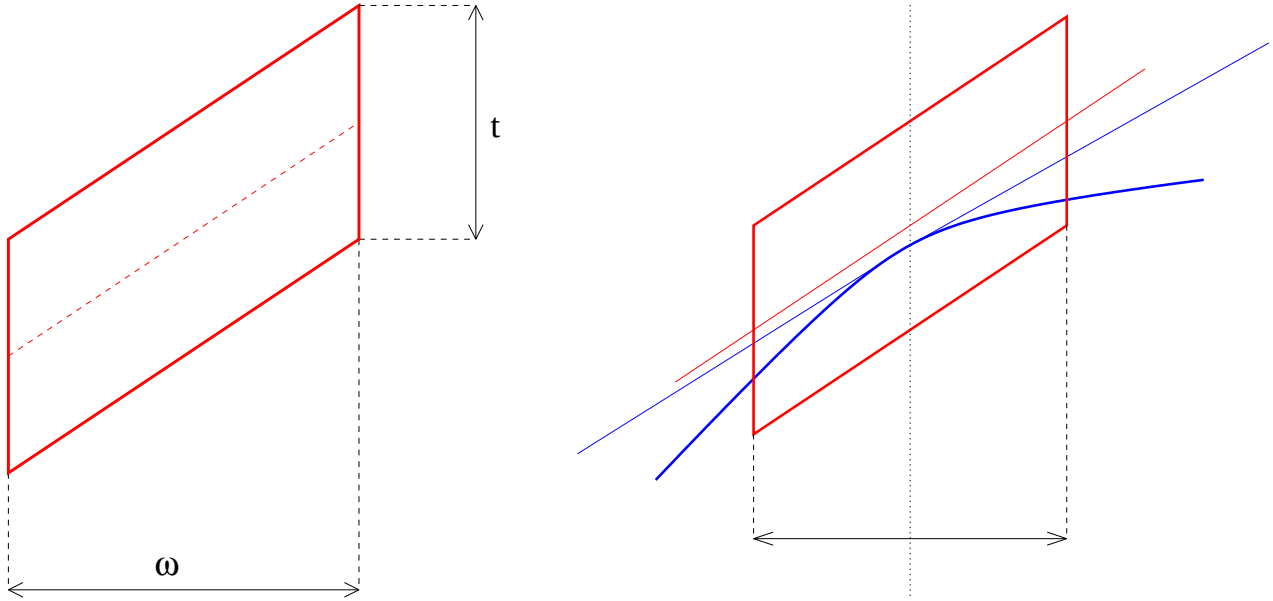


Graph of a Hölder function



Corresponding covering

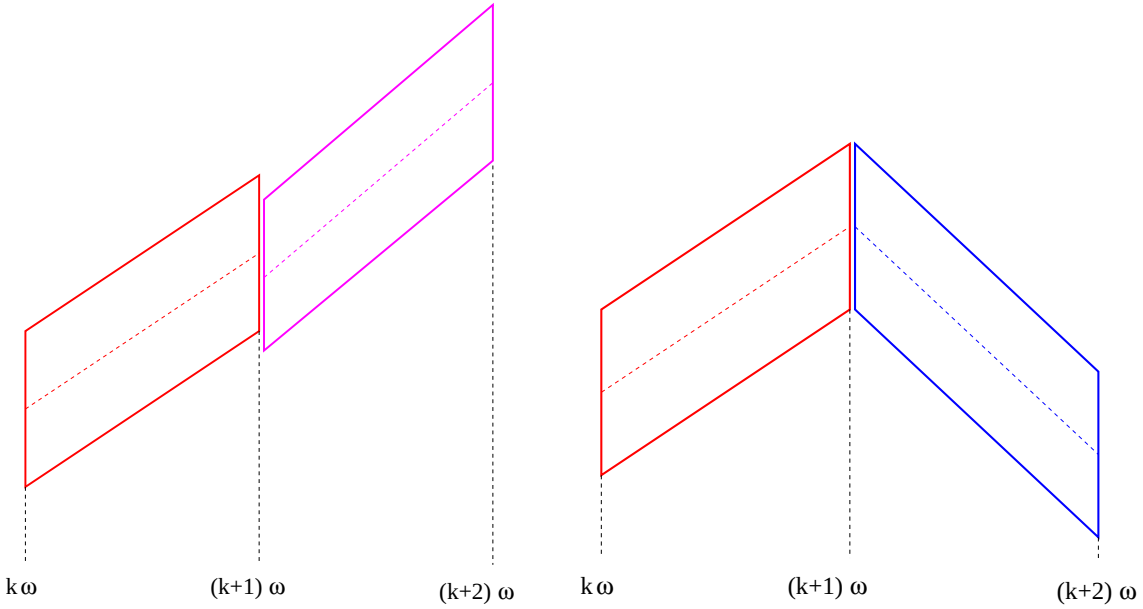
Scale Depends on Smoothness



If the curve is of smoothness α, then $t \asymp w^{\alpha}$

Good Continuation

Determines which cells are connected in the graphical structure



In good continuation In bad continuation

General Coverings

- For general planar curves
- For any $\alpha \geq 1$ (and $\beta > 0$)
- For higher-dimensional objects in higher dimension

Curse of dimensionality!?

- Number of vertices and degree of the graphical structure increase very fast with dimension.

Computational Issues

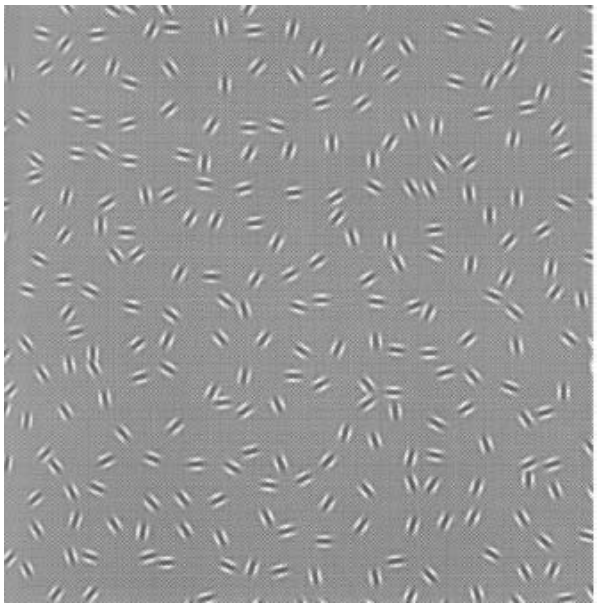
- For graphs of functions, GLRT* can be computed (relatively) fast
- For general curves and other objects, computing GLRT* is NP-har; related to
  - *The Bank Robber Problem*
  - *The Orienteering Problem*
  - *The Prize-Collecting Traveling Salesman Problem*
  - *The Reward Budget Problem*

Other Issues

- How to combine scales efficiently?

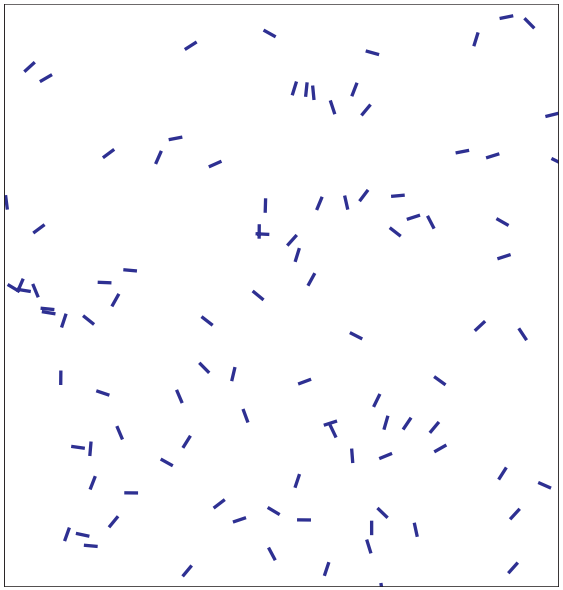
Human Visual System

Understanding the Human Visual System's ability to detect.

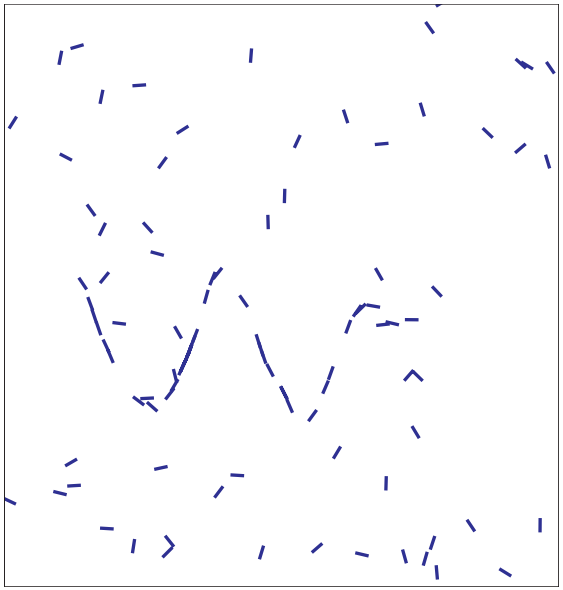


Detection in 2D Vector Field

Finding a curve in 2D vector field data.



No curve is present



A curve is present

Hypothesis Testing Problem

H_0 (null hypothesis)

n points are independent, uniformly distributed in the unit square; associated vectors are independent, uniformly distributed in the unit circle

H_1 (alternative hypothesis)

$n - n_1$ points are independent, uniformly distributed in the unit square; associated vectors are independent, uniformly distributed in the unit circle

n_1 points are on an unknown curve $\gamma \in \mathcal{C}$; associated vectors are tangent to the curve

Generalized Likelihood Ratio Test

For $S \subset [0,1]^2 \times \mathbb{S}^1$, define $\vec{N}(S)$ number of (data) pointed vectors in S.

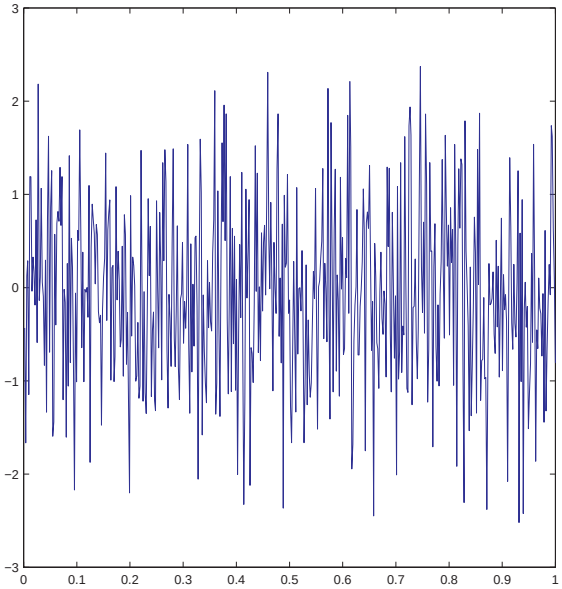
The GLRT rejects for "large" values of

$$\vec{N}(\mathcal{C}) = \max_{\gamma \in \mathcal{C}} \vec{N}(\gamma)$$

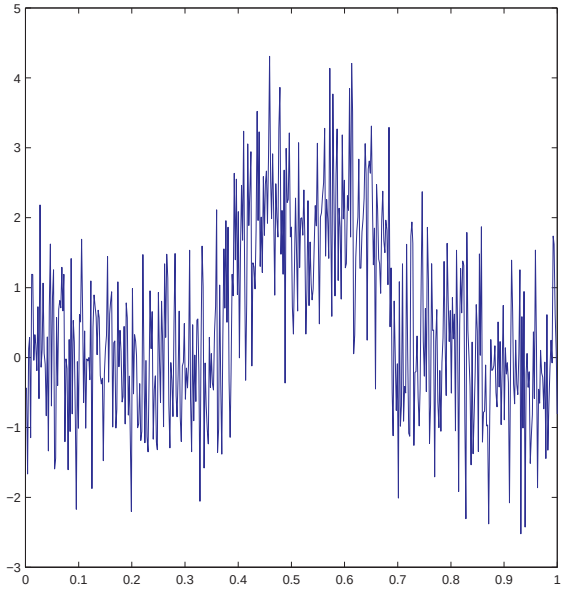
where $\gamma \leftrightarrow (\gamma, \gamma')$

[Parallel results to the previous setting]

Detecting a Boxcar



No boxcar is present



A boxcar is present

Formulation

$$\begin{aligned} X(i) &= \mu 1_{a_0 \le i < b_0}(i) + Z(i), \qquad i = 0, 1, \dots, n-1 \\ &= A \cdot \xi_{[a_0, b_0)}(i) + Z(i), \end{aligned}$$

where

$$\xi_{[a,b)}(i) = 1_{a \le i < b}(i) / \sqrt{b-a}.$$

For $I = [a, b) \in \mathbb{I}_n$, define

$$X[I] = \langle \xi_{[a,b)}, X \rangle = \frac{1}{\sqrt{b-a}} \sum_{i=a}^{b-1} X(i).$$

Naive Approach

- The *Generalized Likelihood Ratio Test* rejects for large

$$X_n^* = \max_{I \in \mathbb{I}_n} X[I].$$

- The GLRT is asymptotically near-optimal ($n \to \infty$) – detects $A = (1+\varepsilon)\sqrt{2 \log n}$ for any $\varepsilon > 0$
- Computing X_n^* requires complexity $O(n^2)$.

Key Remarks

- The $X[I]$ are not independent!
- The index space ($\mathbb{I}_n$) has special structure.

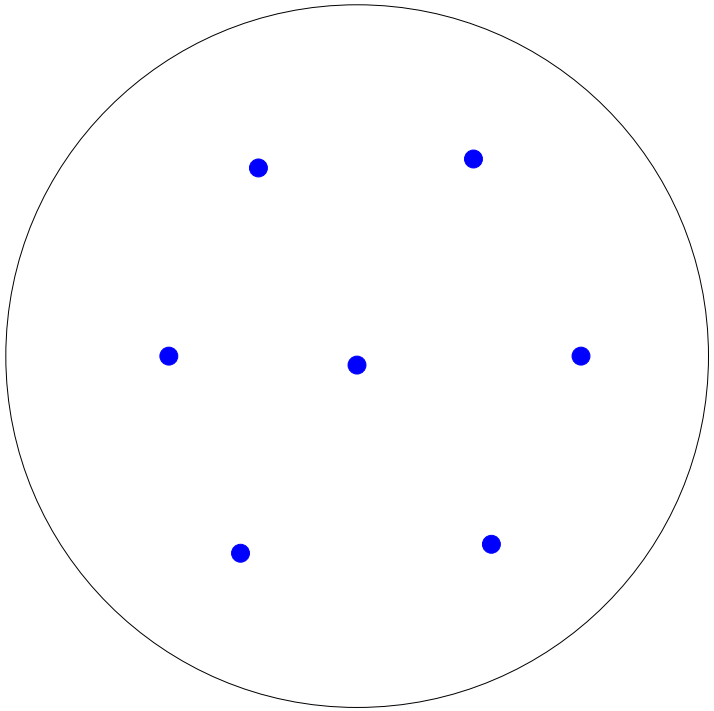
Lemma (Regularity) *Under Null Hypothesis,*

$$\mathrm{var}(X[I] - X[J]) = 2\delta^2(I, J),$$

with

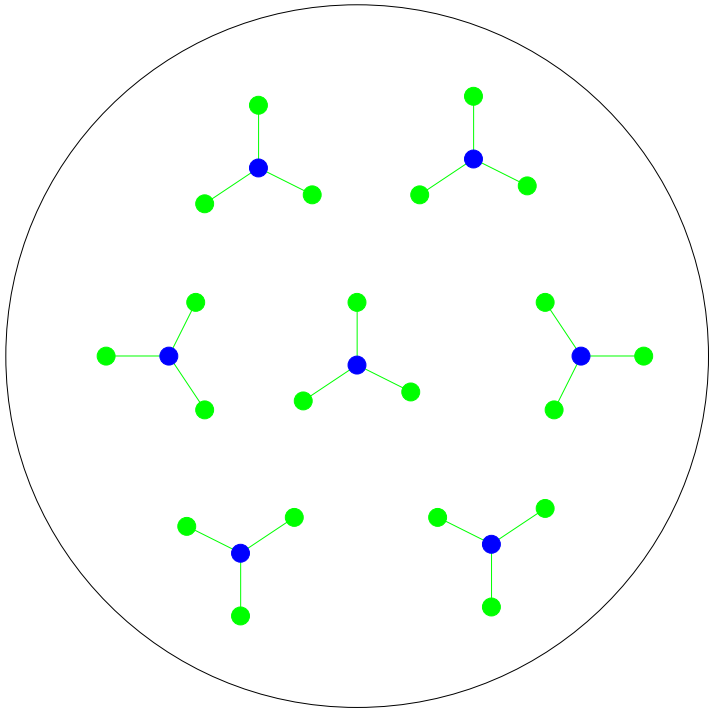
$$\delta^2(I, J) = 1 - \frac{|I \cap J|^2}{|I||J|}.$$

Constructing a δ-Net: Extended Dyadic Intervals



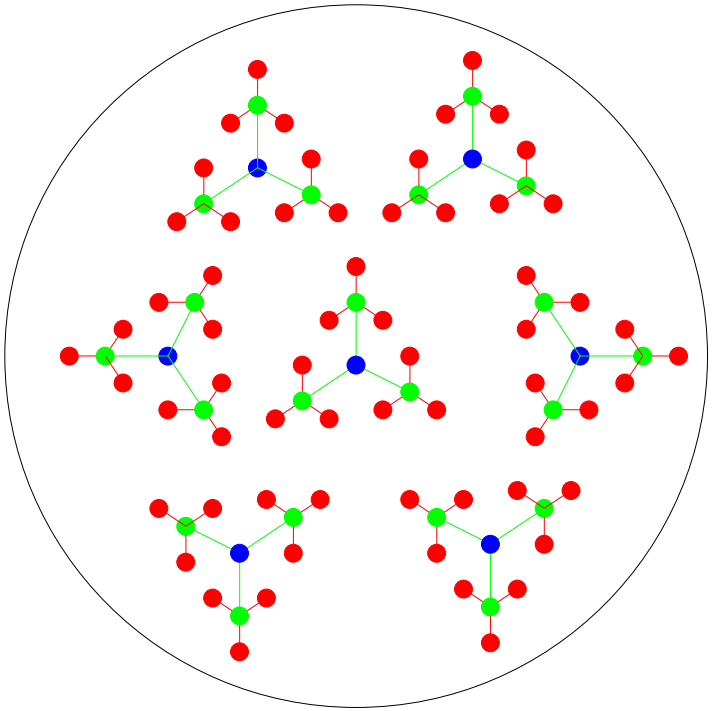
Space of sets

Constructing a δ-Net: Extended Dyadic Intervals



Space of sets

Constructing a δ-Net: Extended Dyadic Intervals



Space of sets

Fast Multiscale Detection Method (FMDM)

(i) *Find promising dyadic intervals.* Identify dyadic intervals I with

$$X[I] > 1/3\sqrt{\log(n)}.$$

If there are more than $n^{19/20}$ such intervals among the $2n$ dyadic ones, reject $H_{0,n}$.

(ii) *Extend promising intervals.* Choose $\ell = \ell_n = \log\log(n)$. For I found at Stage (i), compute

$$X^*_\ell[I] = \max\{X[I'] : I' \in \mathbb{E}_\ell[I]\}.$$

(iii) *Decide.* If the maximum $X[I]$ at Stage 1 exceeds $\sqrt{2\log(n)}$ or if the maximum $X^*_\ell[I]$ at Stage 2 exceeds $\sqrt{2\log(n) + 4\log\log(n)}$, reject $H_{0,n}$.

Performance

Theorem (Detection) *For $\eta > 0$, and $A_n \geq \sqrt{2(1+\eta)\log(n)}$, FMDM has type I and type II errors tending to zero as n increases.*

Theorem (Complexity) *FMDM has complexity of order n.*

Generalizations

Similar results for

- Line-segments (using extended beamlets)
- (Hyper-)rectangles
- Disks
- Rotund Sets

▶ Can this be used to speed-up algorithms presented before?

Collaborators

- David Donoho (Stanford);
- Xiaoming Huo (Georgia Tech);
- Craig Tovey (Georgia Tech).

Websites

- Ery Arias-Castro `http://www-stat.stanford.edu/ acery`
- David Donoho `http://www-stat.stanford.edu/ donoho/`
- Xiaoming Huo `http://www.isye.gatech.edu/ xiaoming/`