0932114 CPS: Foundations of Implicit and Explicit Communications

High Dimensional Anomaly Detection

Venkatesh Saligrama, ECE Dept., Boston University

Applications: Comm. Nets, Social Nets, Health, Fraud detection, Imaging

I. Problem Setup

- Training set: S = {x_1, x_2,……, x_n } from f_0, all nominal
- Test Point: a mixture distribution

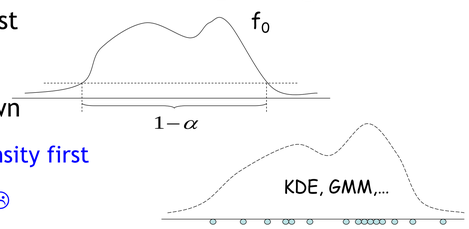
$$f_\times(\eta) = (1-\pi) f_0(\eta) + \pi f_1(\eta)$$

Uniform

- For test point η : $H_0 : \eta \sim f_0$ vs. $H_1 : \eta \sim f_1$
  - ✓ Develop algorithm that controls false alarm & miss detection
- Challenges of the problem
  - ✓ f_0 and mixture weight are unknown
  - ✓ Limited number of training points, high dimensionality

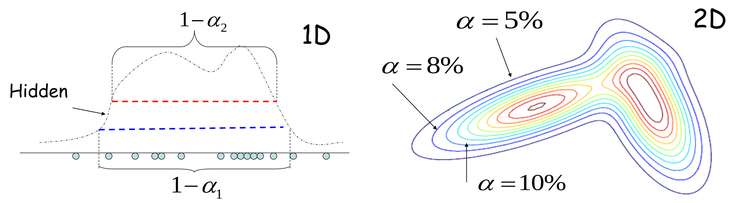
Ideal Anomaly Detector: Everything Known

- If density is known: LR test
  - ✓ Pick up a threshold
- Reality: density is unknown
  - ✓ Naïve idea: estimate density first
  - ✓ Curse of dimensionality ☹



Idea 1: Level Sets

- Estimate level-set [Scott 2006]
  - ✓ OK for 1D or 2D, still hard for high dim. ☹



Idea 2: Level Set Certificate!

- Estimate the measure of the level set
  - ✓ Much easier (scalar estimation) ☺
  - ✓ No information is lost , need not recalculate for different α

II. Point-wise Minimum Volume

- Volume of the smallest level-set containing η:

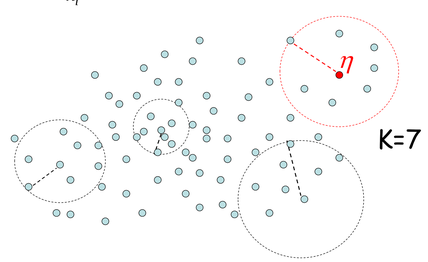
$$\mathrm{vol}(\eta) = P_0(x : f_0(\eta) \geq f_0(x))$$

- Property: the uniformly most powerful test for H_0: $\eta \sim f_0$ vs. H_1: $\eta \sim f_1$ at false alarm rate α is:

Declare anomaly iff. $\mathrm{vol}(\eta) \leq \alpha$

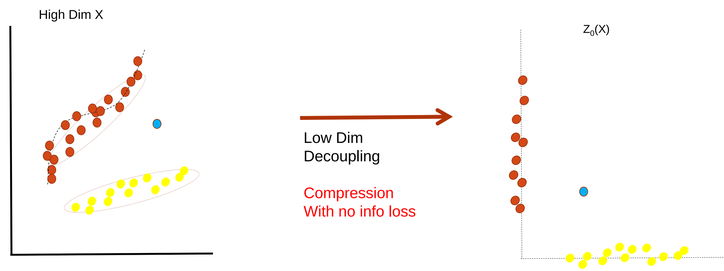
III. KNNG Ranking Algorithm

- Inputs and preconditions:
  - ✓ Desired FA rate, α, neighborhood size K
  - ✓ Training set S = {x_1,x_2,…,x_n}, test point η
  - ✓ Distance metric d(x,y), Euclidean, geodesic,…
- Ranking scheme
  - ✓ $R(x) = \frac{1}{n}\sum_{x_i} 1_{\{KNNG(x) < KNNG(x_i)\}}$



IV. Low-Dim Manifold & Local Structures

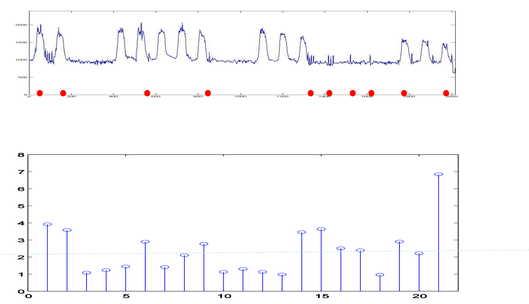
- Kernel Low-rank optimization using nuclear norm relaxations



✓Define KNNG in Z(X) representation

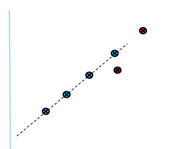
- Local Structures:
  - ✓Markoviantity on Baseline & Anomalous Distribution
  - ✓Marginal PDFs identical outside local window
  - ✓Define KNNG as max or sum over all local KNNG distances

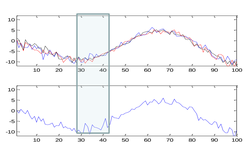
Some results: Power Data Set



V. Issues

- Distance vs. Structure
  - ✓Out-of-Manifold vs. In-Manifold
  - ✓ Different Weights
- Structured Anomalies
  - ✓ Anomaly is Local
  - ✓Space/Time Anomaly





VI. Properties:

- Computationally efficient
- No complicated tuning parameter
- Asymptotically optimal
- Incorporate Manifold Structure
  - ✓Sample complexity ~ intrinsic dim.
- Incorporate Locality Structure
  - ✓ Sample complexity ~ local size