Uniform Histograms for Change Detection in Multivariate Data

Giacomo Boracchi, Cristiano Cervellera, Danilo Macciò

DEIB, Politecnico di Milano and,

ISSIA - National Research Council, Genova

giacomo.boracchi@polimi.it

May 16th, 2017

IJCNN 2017,

Anchorage, Alaska, USA

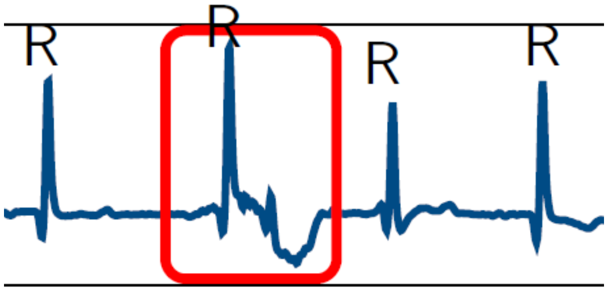
EXAMPLES OF CHANGE-DETECTION APPLICATIONS

Learning in Nonstationary Environments: active classifiers, online classification systems, fraud-detection systems

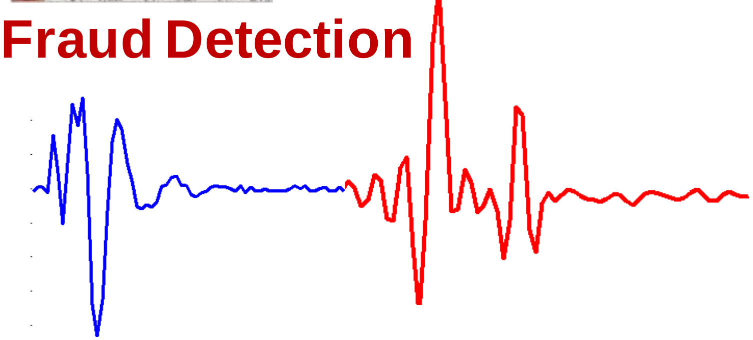
Environmental/Industrial monitoring: quality inspection systems, fault-detection systems

Health monitoring: arrhythmias detection

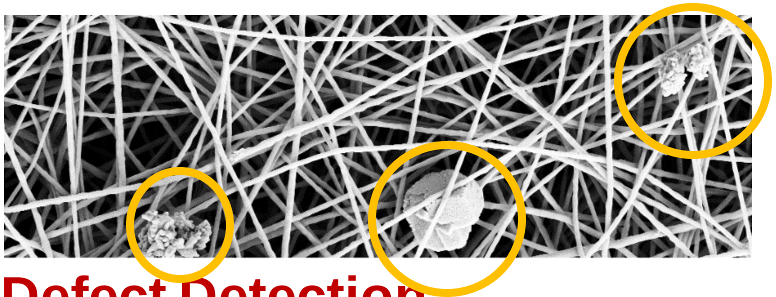
Fraud Detection



ECG Monitoring



Environmental Monitoring



Defect Detection

OFTEN, CHANGE DETECTION BOILS DOWN TO THE FOLLOWING

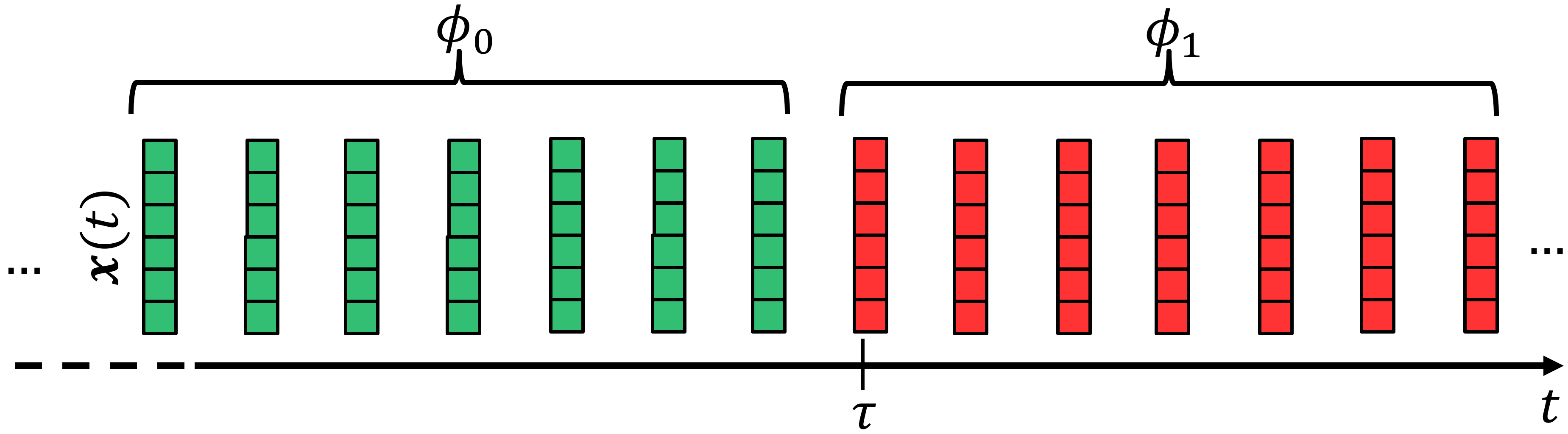
The input **stream** $\{\boldsymbol{x}(t), t = 1, \dots\}$, $\boldsymbol{x}(t) \in \mathbb{R}^d$ contains realizations of a **random variable**, and is modeled as

$$\boldsymbol{x}(t) \sim \begin{cases} \phi_0 & t < \tau \qquad \textcolor{green}{\text{in control state}} \\ \phi_1 & t \geq \tau \qquad \textcolor{red}{\text{out of control state}} \end{cases},$$

where $\{\boldsymbol{x}(t),\ t < \tau\}$ are i.i.d. and $\phi_0 \neq \phi_1$

Goal: detect the change-point τ, by solely analyzing the stream $\{\boldsymbol{x}(t),\ t < \tau\}$

Multivariate data $d > 1$



OFTEN, CHANGE DETECTION BOILS DOWN TO THE FOLLOWING

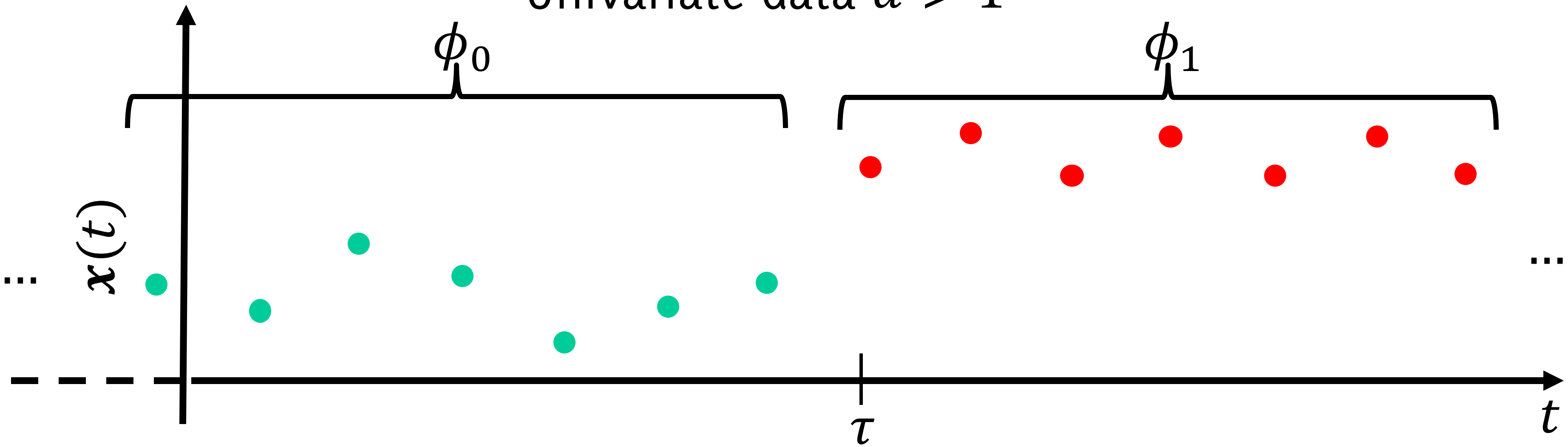
The input **stream** $\{\boldsymbol{x}(t), t = 1, \dots\}$, $\boldsymbol{x}(t) \in \mathbb{R}^d$ contains realizations of a **random variable**, and is modeled as

$$\boldsymbol{x}(t) \sim \begin{cases} \phi_0 & t < \tau & \text{in control state} \\ \phi_1 & t \geq \tau & \text{out of control state} \end{cases},$$

where $\{\boldsymbol{x}(t),\ t < \tau\}$ are i.i.d. and $\phi_0 \neq \phi_1$

Goal: detect the change-point τ, by solely analyzing the stream $\{\boldsymbol{x}(t),\ t < \tau\}$

Univariate data $d > 1$



PRACTICAL ISSUES

Many applications pose the **following constraints/challenges**:

- ϕ_1 the **post-change distribution is unknown**, and it is convenient not to make many assumptions on that
- ϕ_0 the **stationary distribution is unknown**, there is no good parametric model to fit
- An initial training set of stationary data is provided

$$X = \{\boldsymbol{x}_i, i = 1, \dots N, \text{s.t.}, \boldsymbol{x}_i \sim \phi_0\}$$

- Data have to be analyzed batch-wise

$$W = \{\boldsymbol{x}_j, j = 1, \dots, \nu\}$$

determining whether each incoming batch W was from ϕ_0

A PRACTICAL APPROACH

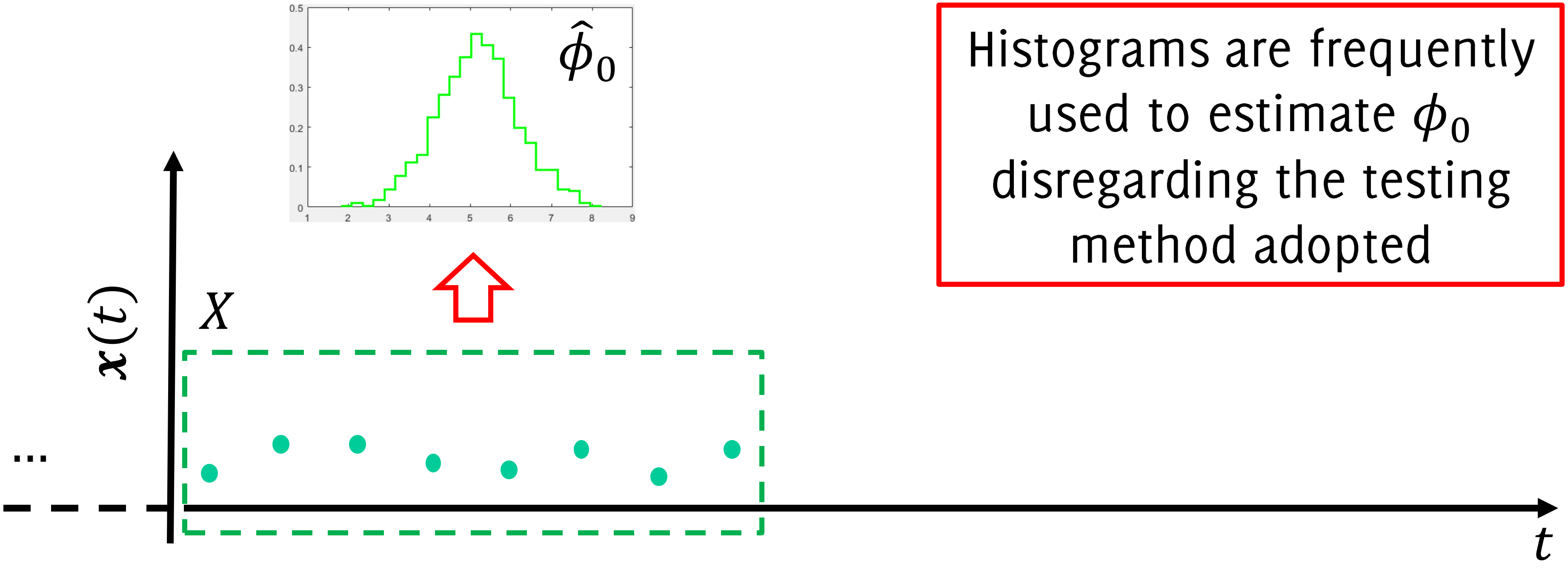
A **convenient approach** is to

- **Train: fit a nonparametric model** $\widehat{\phi}_0$ from X to learn the stationary distribution
- **Test:** compute a suitable statistic to **determine whether incoming batches match** $\widehat{\phi}_0$ or not

A PRACTICAL APPROACH

A **convenient approach** is to

- **Train: fit a nonparametric model** $\hat{\phi}_0$ from X to learn the stationary distribution
- **Test:** compute a suitable statistic to **determine whether incoming batches match** $\hat{\phi}_0$ or not



Histograms are frequently used to estimate ϕ_0 disregarding the testing method adopted

HISTOGRAM DESIGN & CHANGE DETECTION

The **way we build histograms is crucial**, as it determines

- the viable change-detection approach (likelihood-based or distance-based)
- the change-detection performance

This has never been investigated before!

HISTOGRAM DESIGN & CHANGE DETECTION

The **way we build histograms is crucial**, as it determines

- the viable change-detection approach (likelihood-based or distance-based)
- the change-detection performance

This has never been investigated before!

Our contribution: we investigate the use of

- Histograms defined on a **regular grids (uniform volume)**
- Histograms yielding **uniform density**

for **change-detection purposes.**

Our experiments show that **uniform-density histograms** combined with **distance-based methods** are typically **more effective**.

PAPER OUTLINE

- Uniform Histograms
- Efficient construction of Uniform Density Histograms
- Distance-based monitoring
- Likelihood-based monitoring
- Experiments: change-detection performance

Uniform Histograms

HISTOGRAMS

An histogram h^0 defined over the input domain $\mathcal{X} \subset \mathbb{R}^d$ is

$$h^0(\mathcal{X}) = \{(S_k, p_k^0)\}_{k=1,\dots,K}$$

Where $\{S_k\}_k$ is a disjoint covering of $\mathcal{X}$, namely $S_k \subset \mathcal{X}$

$$\bigcup_k S_k = \mathcal{X} \text{ and } S_j \cap S_i = \delta_{i,j}$$

and $p_k^0 \in [0,1]$ is an the probability (estimated from X) for a sample drawn from ϕ_0 to fall inside S_k, i.e.

$$p_k^0 = \frac{m_k}{N}$$

where $N = \#X$

There is quite a lot of freedom in designing $\{S_k\}_k$

HISTOGRAMS BUILT ON A REGULAR GRID

This is the most common way of constructing histograms.

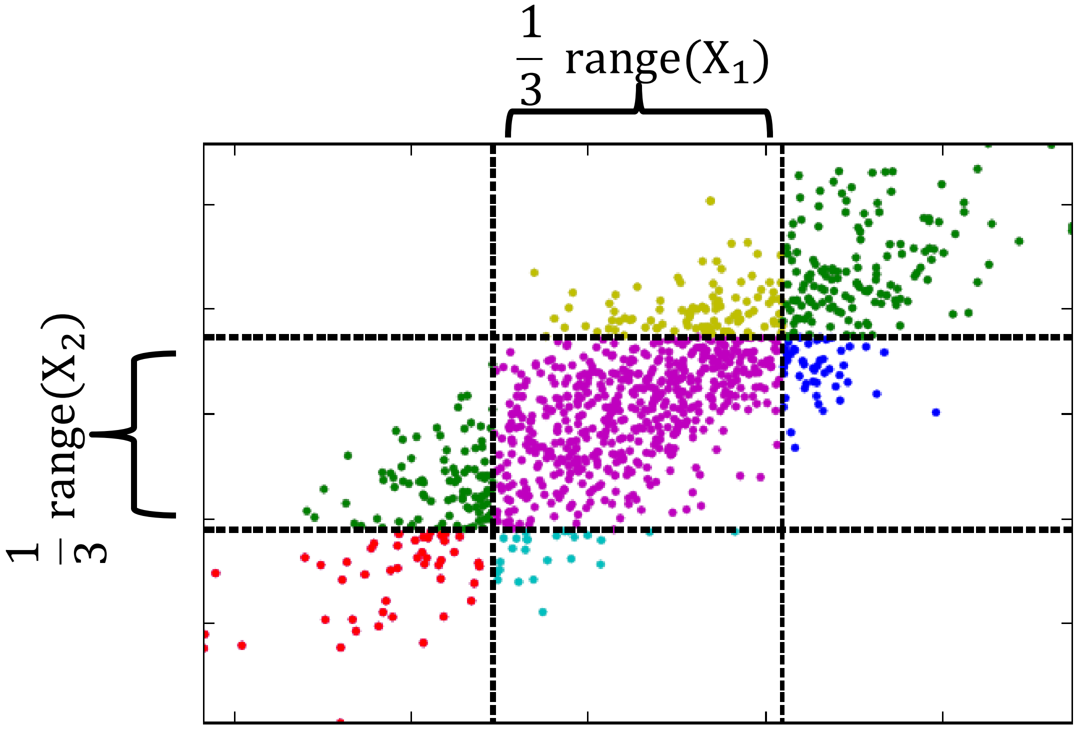
Build a tessellation of supp(X) by splitting each component in q equally sized parts.

This yields q^d hyper-rectangles $\{S_k\}$ having the **same volume**

Add to the histogram a region to gather points that during operation, won’t fall in supp(X)

$$S_K = \bar{X}, p_K^0 = 0$$

being $K = q^d + 1$



An example of 2D histogram $q = 1/3$

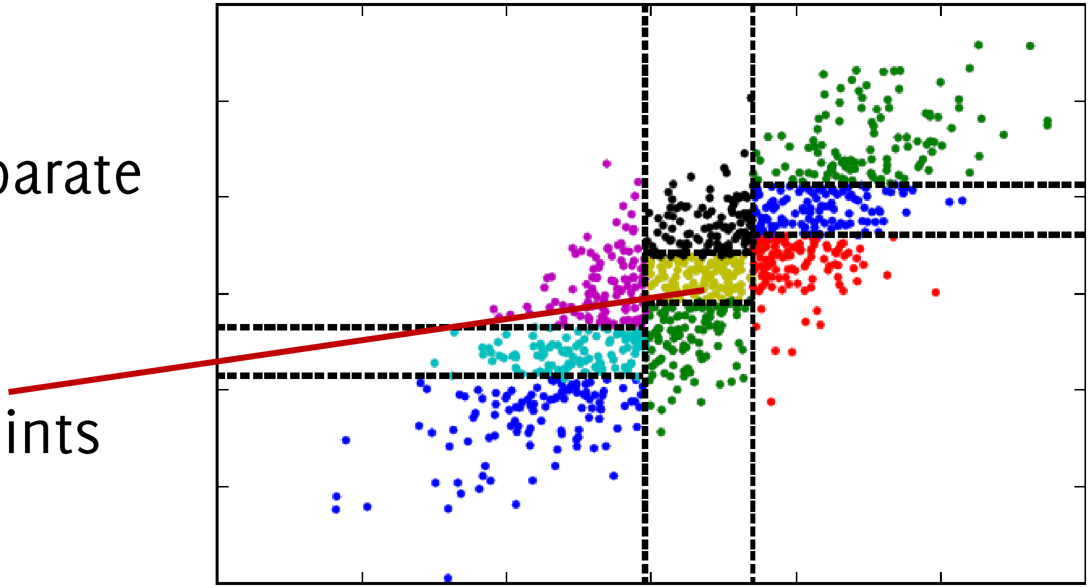
HISTOGRAMS YIELDING UNIFORM DENSITY

Define the partition $\{S_k\}_k$ in such a way that all the subsets have the **uniform density**, i.e.,

$$p_k^0 \approx \frac{1}{K}, k = 1, .., K$$

Such that each of the q^d hyper-rectangles contains the same number of points

No need to consider a separate region for $\bar{X}$



$\frac{N}{9}$ points

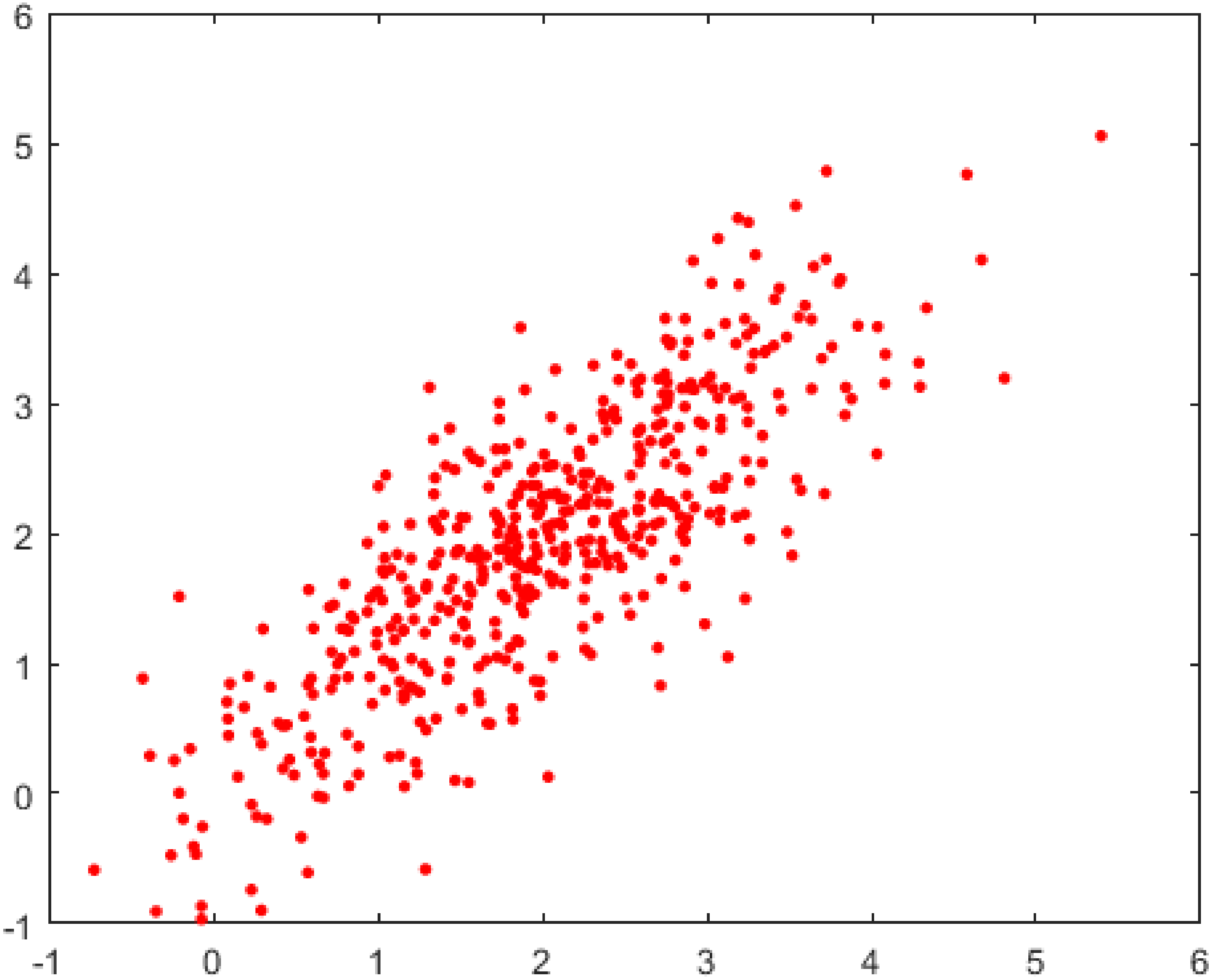
An example of 2D histogram $q = 1/3$

CONSTRUCTION OF UNIFORM DENSITY HISTOGRAMS

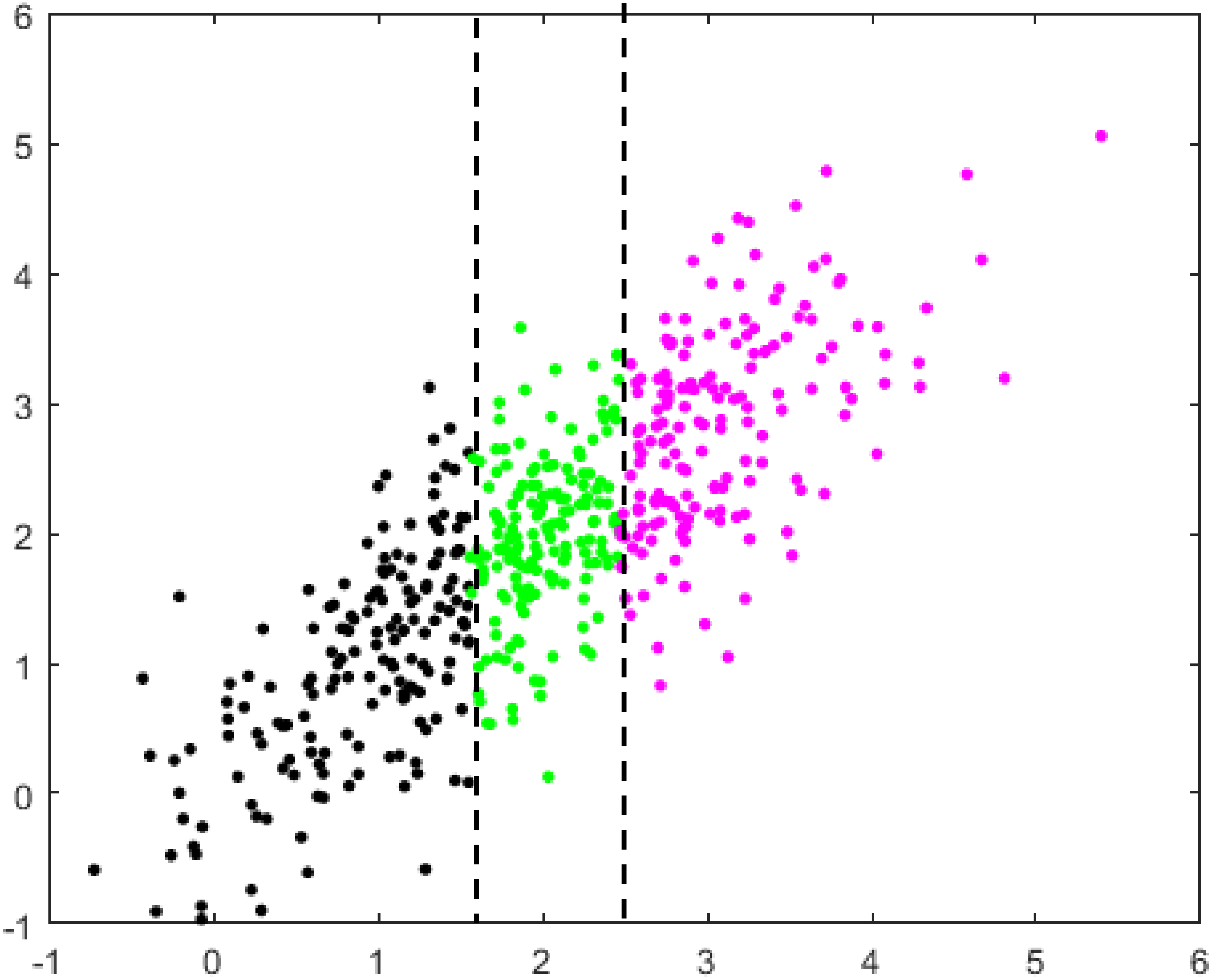
Uniform histograms can be constructed through an iterative procedure that scans all the dimensions

- Initialize $i = 1, v_1 = \mathcal{X}$ (the only subset)
- For each dimension $i = 1, \dots, d$
  - Define K_i as the current number of sets
  - Split each subset $\{S_1, \dots, S_{K_i}\}$ along the i-th component in q subsets containing the same number of points. Subsets are defined by the percentiles of the marginal distribution
- Given the final histogram $\{S_1, \dots, S_K\}, K = q^d$ compute the probabilities

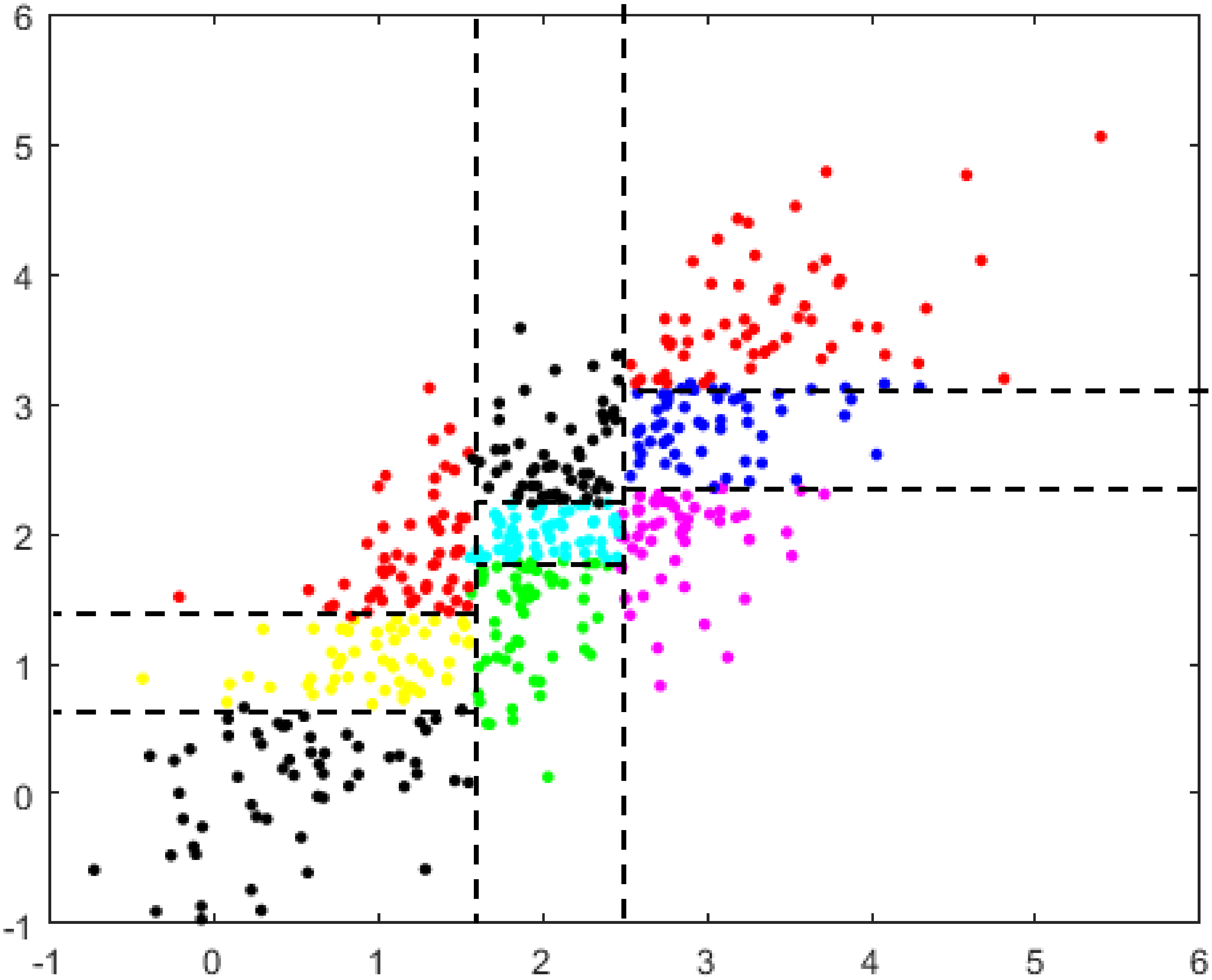
FIRST ITERATION, S_1



SECOND ITERATION, $\{S_1, S_2, S_3\}$



THIRD ITERATION, $\{S_1, S_2, S_3, S_4, \ldots, S_9\}$



REMARKS

Histograms can be considered as trees and it is very fast to determine the cell where each input point belongs to.

The number of subsets is bounded to an exponential growth (q^d)

As d increases,

- many subsets in h^0 uniform volume will be empty
- all the subsets in h^0 uniform density will be equally populated

Different histograms can implement different monitoring schemes!

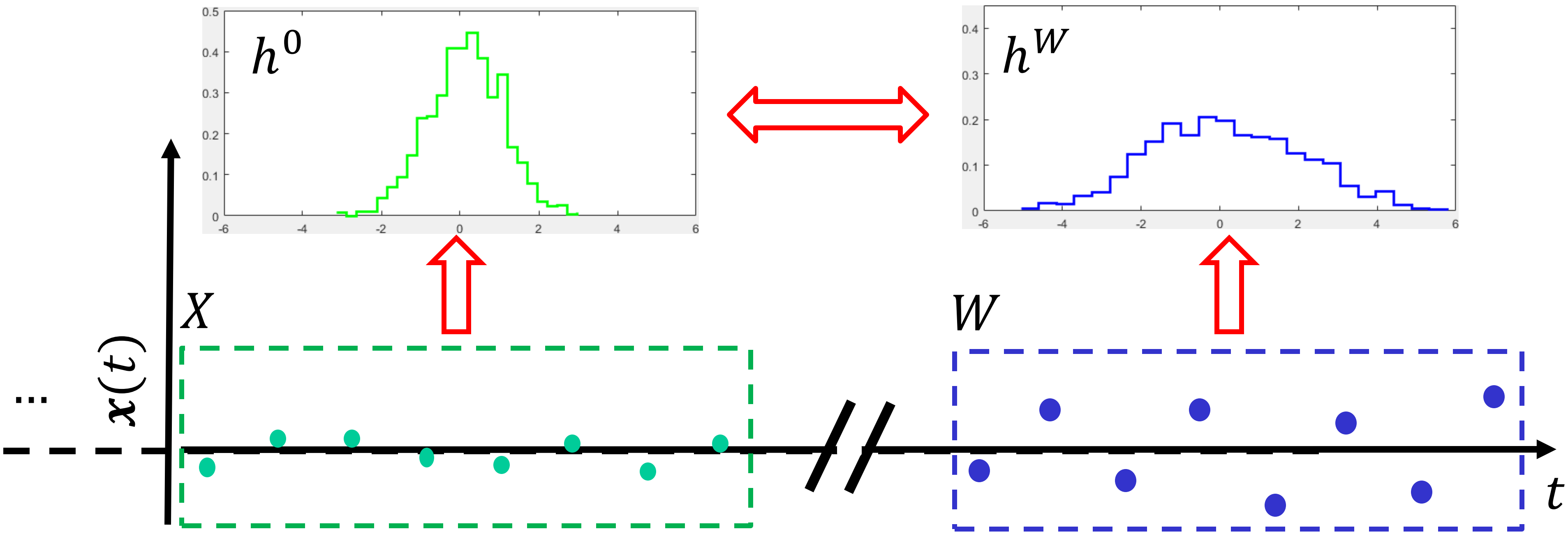
Change Detection Using Histograms

DISTANCE – BASED MONITORING USING HISTOGRAMS

Compute an histogram h^W from an incoming batch and compare it against h^0.

Measure the distance between histograms and run an HT.

$$d(h^0, h^W)$$



DISTANCE – BASED MONITORING USING HISTOGRAMS

1. Compute the probabilities for an incoming batch W over $\{S_k\}$

$$p_k^W = \frac{\#\{\boldsymbol{x}_i \in S_k \cap W\}}{\nu}$$

2. Compare h^0 and h^W by a suitable distance, e.g.

$$d_{TV}(h^0, h^W) = \frac{1}{2}\sum_k |p_k^0 - p_k^W| \quad \text{(total variation)}$$

or

$$d_{PS}(h^0, h^W) = \nu \sum_k \frac{|p_k^0 - p_k^W|}{p_k^0} \quad \text{(Pearson)}$$

3. Run an HT on d_{TV} (having estimated its p-values empirically) or d_P (this follows a χ-square distribution)

LIKELIHOOD – BASED MONITORING SCHEME

Advantages:

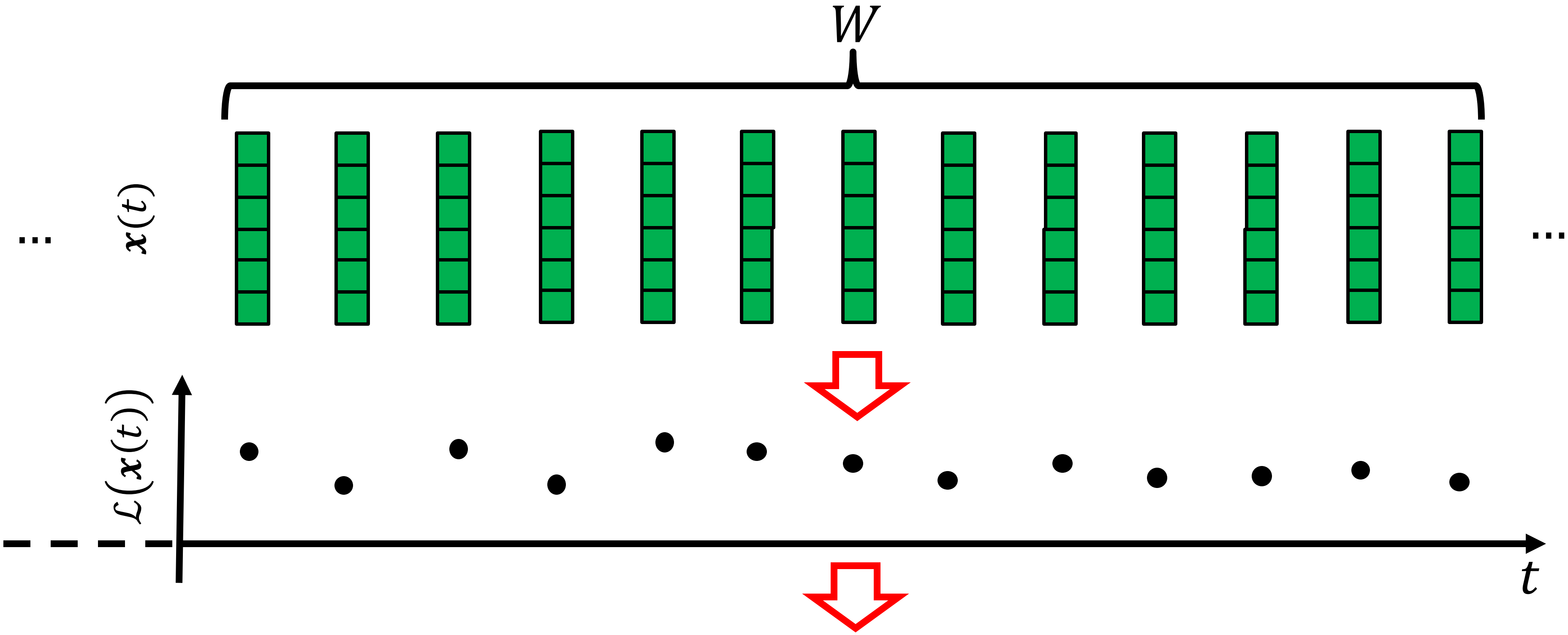
- If we run the Pearson test we do not have to compute empirical distributions. As far as there are enough samples in each S_k we can use it
- Other distances between discrete distributions could be considered, e.g. Kolmogorov-Smirnov

Cons:

- Distance-based methods cannot be applied to perform element-wise monitoring
- Sequential monitoring schemes have to be implemented at batch-level
- Perason test does not admit zero-probabilities,

LIKELIHOOD – BASED MONITORING USING HISTOGRAMS

As in density-based methods, the histogram is used to compute the likelihood as $\hat{\phi}_0$. Monitor likelihood values to see whether it has decreased w.r.t. stationary data



$$\bar{l}_W = \frac{1}{\nu} \sum_{t \in W} p_k^0(t)$$

Test whether $\bar{l}_W = \bar{l}_0$

LIKELIHOOD – BASED MONITORING USING HISTOGRAMS

1. During training, estimate h^0 from X and the distribution of $\overline{l}_0$, the average likelihood over test-batches $W \sim \phi_0$
2. During testing, compute

$$\mathcal{L}\big(\boldsymbol{x}(t)\big) = \hat{\phi}_0\Big(h^0\big(\boldsymbol{x}(t)\big)\Big) = p_k^0 \text{ , s.t. } \boldsymbol{x}(t) \in S_k$$

3. Monitor $\{\mathcal{L}\big(\boldsymbol{x}(t)\big),\ t = 1, \ldots\}$ in a batch-wise manner, i.e.,
   - compute $\overline{l}_W$, the average likelihood on W
   - Run an hypothesis test with null hypothesis $\overline{l}_W = \overline{l}_0$

LIKELIHOOD – BASED MONITORING SCHEME

Advantages:

- Computing the likelihood is seen as an effective way to reduce dimensionality for change-detection purposes
- Likelihood based monitoring schemes enable element-wise monitoring of data-streams. We adopted batch-wise monitoring scheme to enable a fair comparison with
- Many tests can be applied to univariate quantities

Cons:

- Some tests based on ordinal statistics might fail when probabilities $\{p_k^0\}$ assume few distinct values
- Likelihood-based monitoring cannot be applied with uniform-density histograms

Experiments

DATASET AND METHODS CONSIDERED

Gaussian Data: We synthetically generate 10000 pairs of Gaussians $\phi_0 \to \phi_1$ having that $sKL(\phi_0, \phi_1) = 1$, $d = 2, \dots, 5$ and

$$\phi_1 = \phi_0(Q\boldsymbol{x} + \boldsymbol{v})$$

Protein Data: "Physicochemical Properties of Protein Tertiary Structure" dataset from the UCI repository. We have extracted $d = 2, \dots, 5$ components and introduced changes by shifting the observations.

Alippi C., Boracchi G., Carrera D.,"*CCM: Controlling the Change Magnitude in High Dimensional Data*" INNS Conference on Big Data, 2016 https://home.deib.polimi.it/carrerad/projects.html (Matlab Package Available)

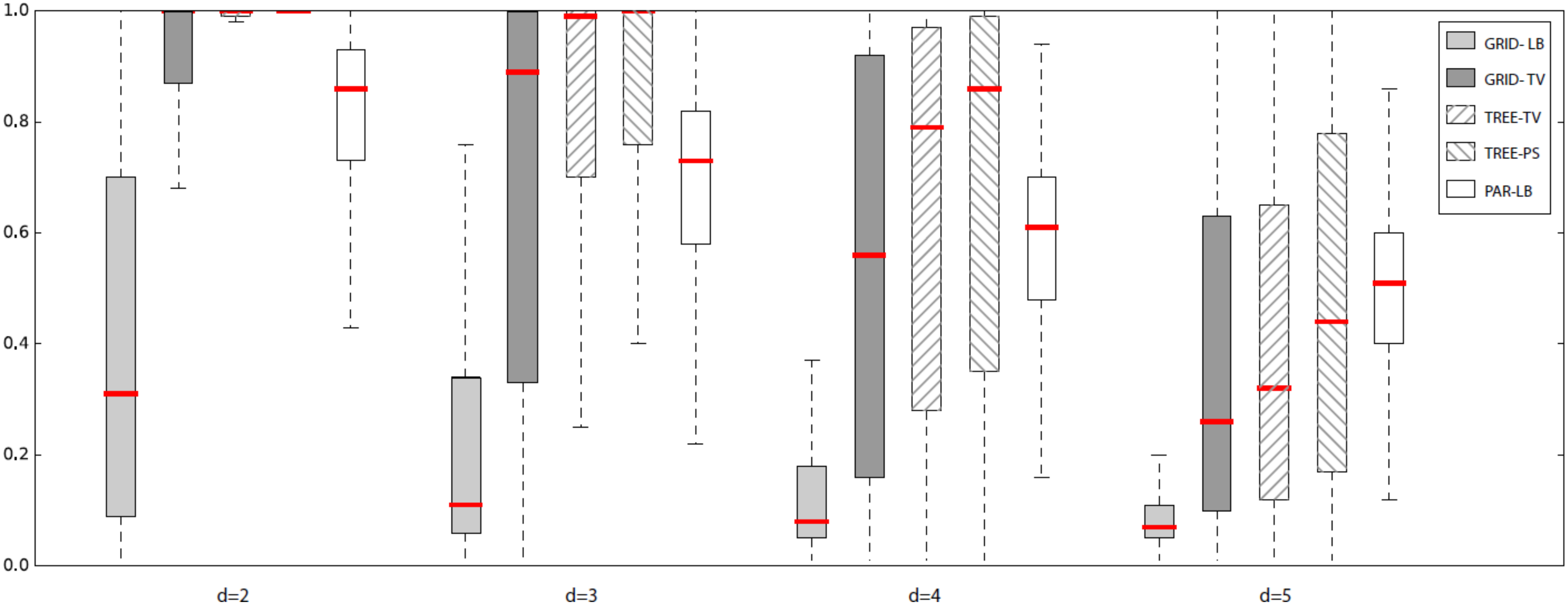
METHODS

- **GRID-LB**: histograms built upon regular grids using the likelihood-based test.
- **GRID-TV**: histograms built upon regular grids using a test on the p-values of d_{TV}.
- **TREE-TV:** histograms yielding uniform density using a test on the p-values of d_{TV}.
- **TREE-PS**: histograms yielding uniform density using the Pearson chi-square test, i.e. d_{PS}
- **PAR-LB:** test on the likelihood computed by assuming ϕ_0 known, it holds only for the Gaussian sequences

All these methods have been configured to yield 5% FPR

TESTS ON GAUSSIAN DATASETS

We report the power of the corresponding datasets in boxplots,



Median of the powers for Gaussian data ($q = 2$)

	$d = 2$	$d = 3$	$d = 4$	$d = 5$
GRID-TV	0.99	0.86	0.62	0.33
TREE-TV	1	0.97	0.81	0.46
TREE-PS	1	0.98	0.85	0.49
GRID-LB	0.3	0.14	0.09	0.08

Median of the powers for Gaussian data ($q = 3$).

	$d = 2$	$d = 3$	$d = 4$	$d = 5$
GRID-TV	1	0.89	0.56	0.26
TREE-TV	1	0.99	0.79	0.32
TREE-PS	1	1	0.86	0.44
GRID-LB	0.31	0.11	0.08	0.07

TESTS ON THE PROTEIN DATASET

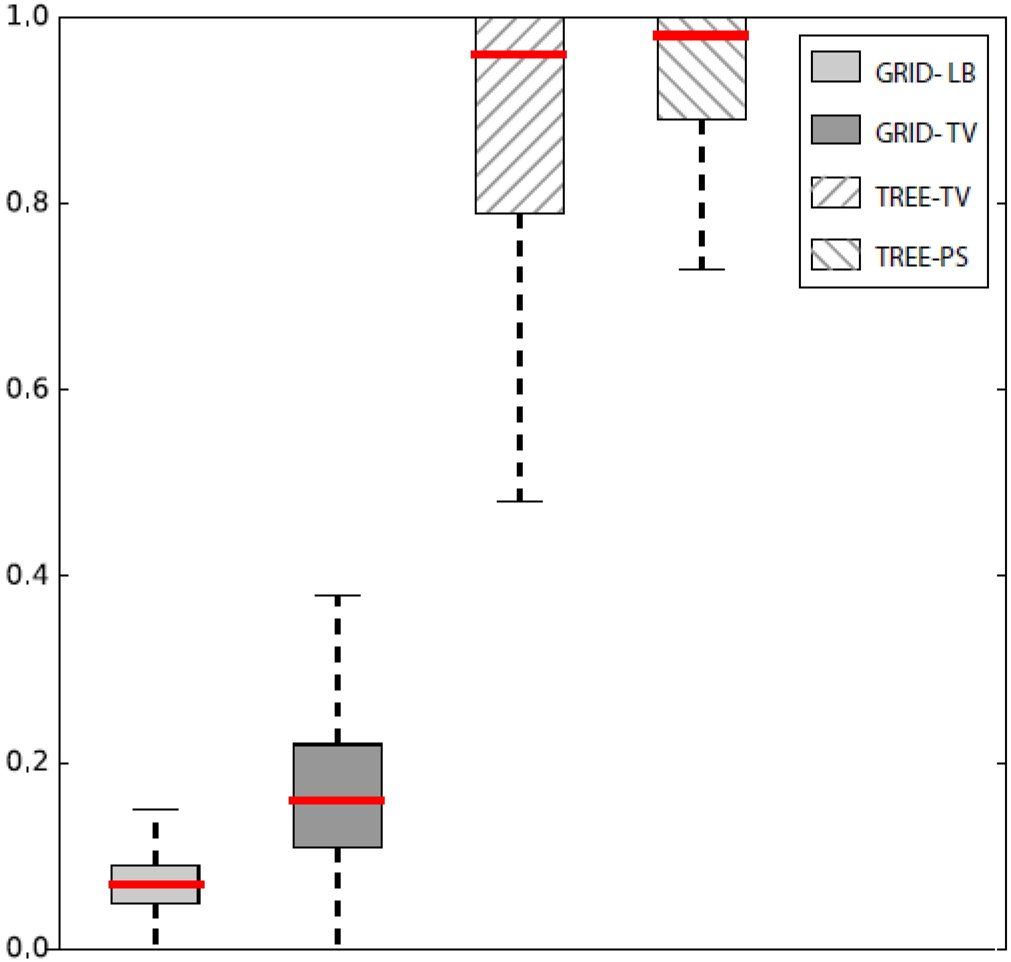


TABLE III: PAR-LB: median of the powers for Gaussian data.

	$d = 2$	$d = 3$	$d = 4$	$d = 5$
PAR-LB	0.86	0.73	0.61	0.51

TABLE IV: Median of the powers for the *Protein* datase

	GRID-TV	TREE-TV	TREE-PS	GRID-LB
q=2	0.11	0.87	0.91	0.08
q=3	0.16	0.96	0.98	0.07

COMMENTS

distance-based distance outperform likelihood-based ones

- In a batch-wise monitoring scheme, it is better to compare the distributions rather than only the average likelihood

The boxplots are very wide

- Differences are significant, thought
- This is mainly due to the fact that ϕ_0 was randomly defined, bad-conditioned covariances are more challenging
- PAR-ML does not suffer of this problem

The power of all the tests is decreasing when d increases

- This is in agreement with the *"detectability loss"* problem we have shown in **[Alippi et al, IJCAI16]**

[Alippi et Al IJCAI 2016] C. Alippi, G. Boracchi, D. Carrera, M. Roveri, "*Change Detection in Multivariate Datastreams: Likelihood and Detectability Loss*" IJCAI 2016, New York, USA, July 9 - 13

COMMENTS

Histograms yielding uniform density are typically better than histograms yielding uniform volumes

- This might sound surprising since these have not been much considered in the literature!
- Grids might provide very unbalanced coverage of the input domain and this might be a problem

Performance on synthetic and real dataset are consistent

Concluding Remarks

CONCLUDING REMARKS

Histograms used for change-detection problems have to be carefully designed depending on the monitoring conditions

Our experiments show that the best option in batch-wise monitoring consists in combining uniform density histograms with a distance-based method.

Future work aims at developing further uniform density histograms for high-dimensional change-detection problems, where it is not feasible to grow the partition size as q^d

Analyze other nonparametric monitoring schemes such as kernel density estimation methods used for LNSE