



HAL
open science

Leveraging the properties of the Christoffel function for anomaly detection in data streams

Louise Travé-Massuyès

► **To cite this version:**

Louise Travé-Massuyès. Leveraging the properties of the Christoffel function for anomaly detection in data streams. POP23 - Future Trends in Polynomial OPTimization, LAAS-CNRS, Toulouse, Nov 2023, Toulouse, France. hal-04795310

HAL Id: hal-04795310

<https://hal.science/hal-04795310v1>

Submitted on 21 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Leveraging the properties of the Christoffel function for anomaly detection in data streams

Louise Travé-Massuyès

Kévin Ducharlet, Jean-Bernard Lasserre

POP23 - Future Trends in Polynomial OPTimization

13-17 November 2023

LAAS-CNRS, Toulouse



DATA QUALITY



FAULT DETECTION



Are there intruders ?

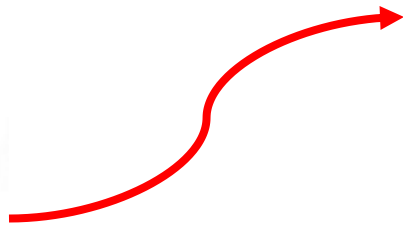


Anomaly

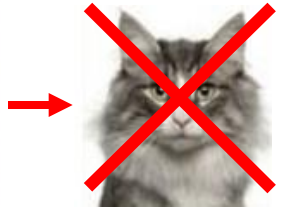
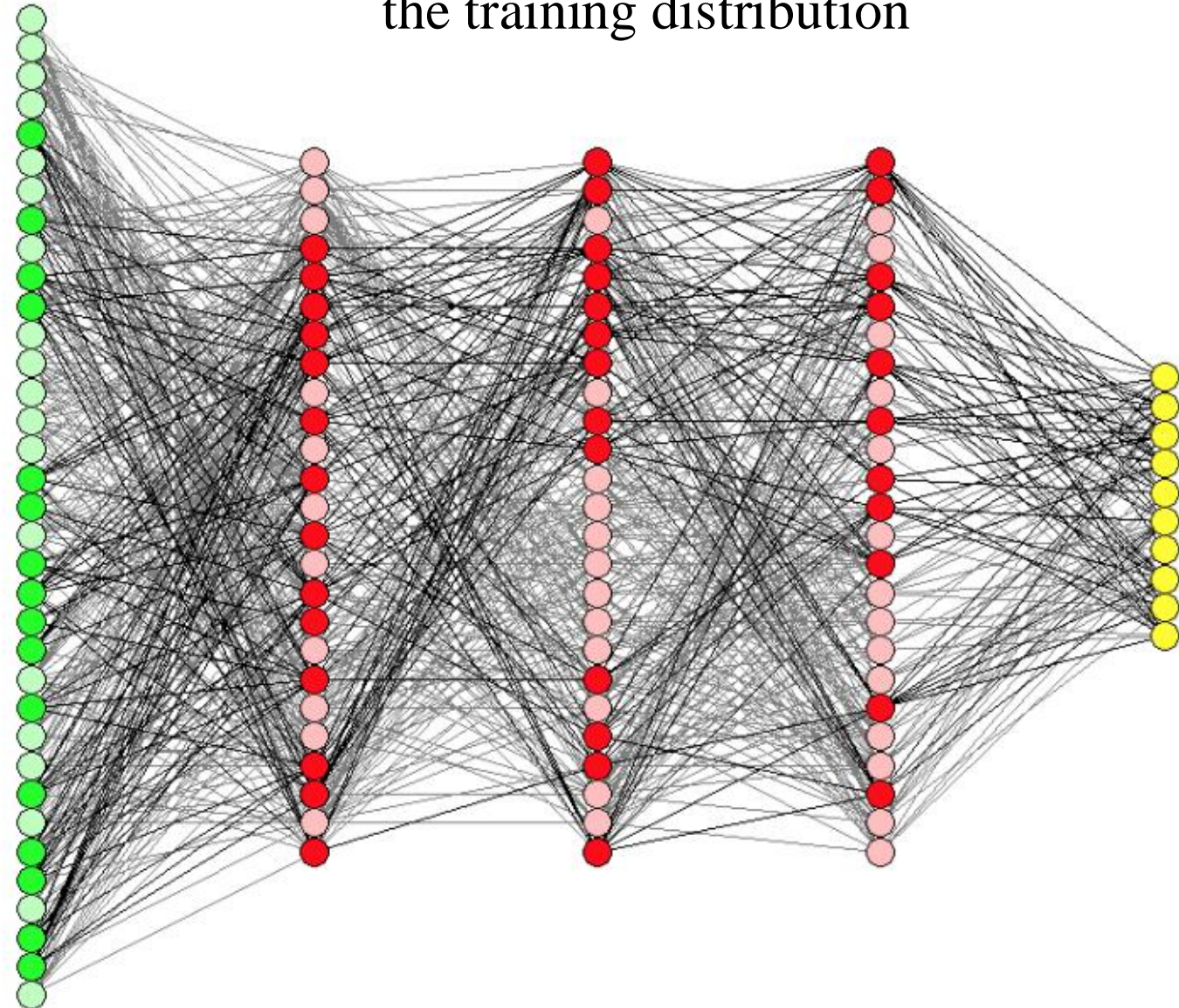
- an instance from a distinct distribution
- a rare or low-probability instance

Data quality

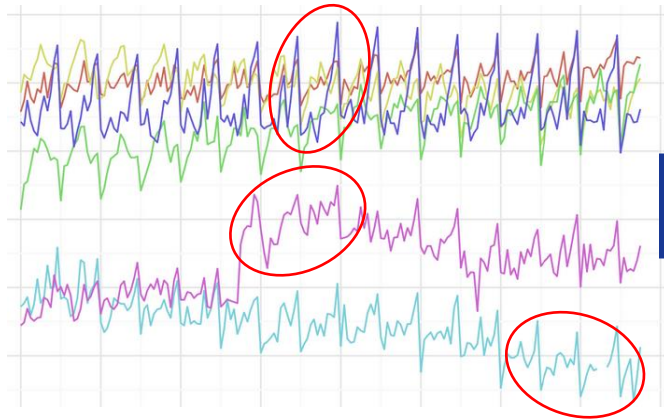
Training data



AI systems reliability is based on inputs lying in the training distribution



Anomalies



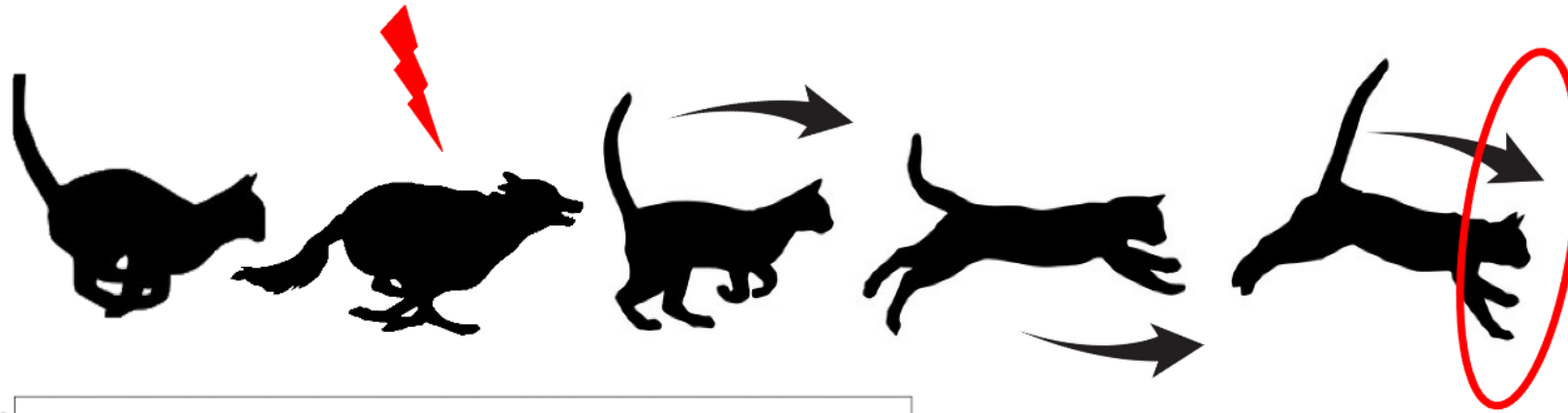
Symptoms

DIAGNOSIS



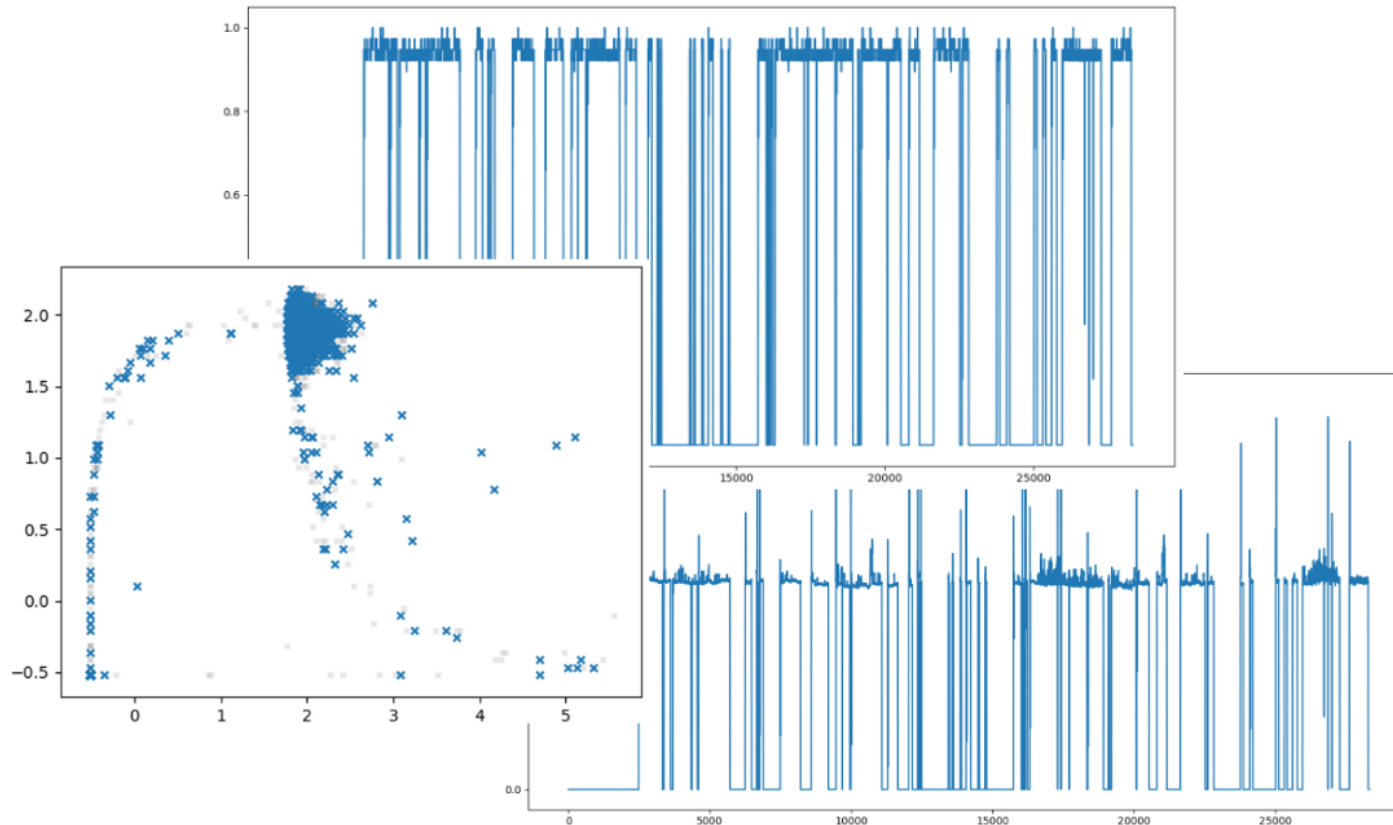
Root cause

Anomaly detection in data streams



Data stream requirements

- detection on the fly
- incremental update
- low complexity
- complex distributions



▶ Adaptation of time series methods

- Prediction models, e.g. based on exponential smoothing or LSTM
- Do not account for concept drift

▶ Dynamic clustering

- Outliers do not belong to clusters or are in low density clusters

▶ Methods relying on KNN

- Based on number of neighbours, e.g., MCODE
- Based on local density based (LOF and variants)

▶ Statistical methods

- Parametric methods, e.g., based on GMM (Samrtisifter)
- Non parametric methods, e.g., on line Multiple Kernel Density Estimation (MKDE)

Many methods deal with *transiency*, *concept drift*, *infinity* and *time dependency*,
mostly through the **use of windows**

But **no rapid model update** and tend **no memory** of previously acquired
knowledge, **window size dependency**

A **hybrid AI** anomaly detection method for data streams that:

- ▶ leverages the **Christoffel function**
 - ▶ related to the **Christoffel-Darboux kernel** borrowed from the **theory of approximation and orthogonal polynomials**
 - ▶ advocated for data mining by J.-B. Lasserre and E. Pauwels (2019)
- ▶ benefits from a clean algebraic framework
- ▶ fulfils all data stream requirements
- ▶ needs **little tuning or no tuning at all**

A **hybrid AI** anomaly detection method for data streams that:

- ▶ leverages the **Christoffel function**
 - ▶ related to the **Christoffel-Darboux kernel** borrowed from the **theory of approximation** and **orthogonal polynomials**
 - ▶ advocated for data mining by J.-B. Lasserre and E. Pauwels (2019)
- ▶ benefits from a clean algebraic framework
- ▶ fulfils all data stream requirements
- ▶ needs **little tuning**

A collaboration between **two ANITI chairs**:

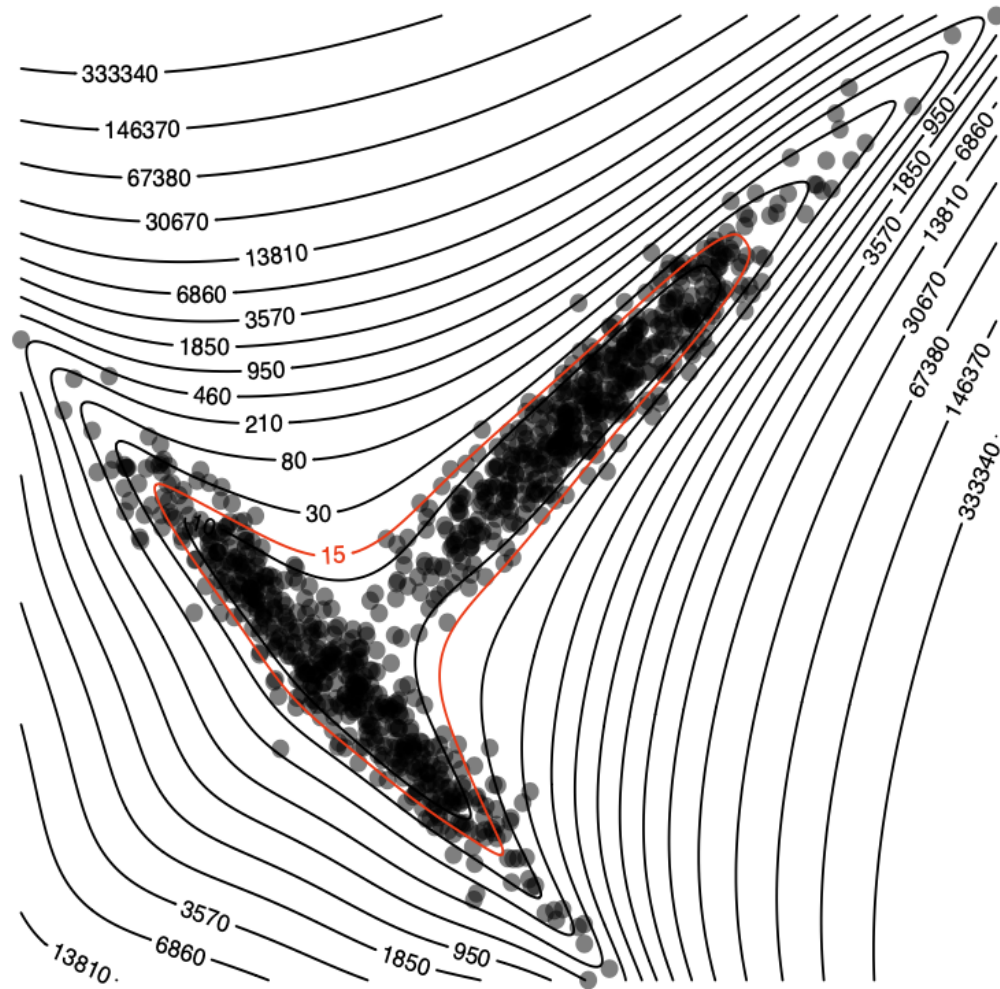
- ▶ **Polynomial Optimization** for Machine Learning and Data Analysis
(*Jean-Bernard Lasserre*)
- ▶ Synergistic Transformations in Model Based and Data Based **Diagnosis**
(*Louise Travé-Massuyès*)



PhD thesis of Kévin Ducharlet

Détection d'anomalies dans les flux de données pour une application dans les réseaux de capteurs (in french), PhD thesis, Computer science & Control, INSA, defended on September 28, 2023.

Capturing the shape of a cloud of points



Consider a cloud of data points
 $(x(i))_{i \in \mathbb{N}} \subset \mathbb{R}^p$

The red curve is the level set:
 $\mathcal{L}_\gamma = \{x : Q_d(x) \leq \gamma\}, \gamma \in \mathbb{R}_+$

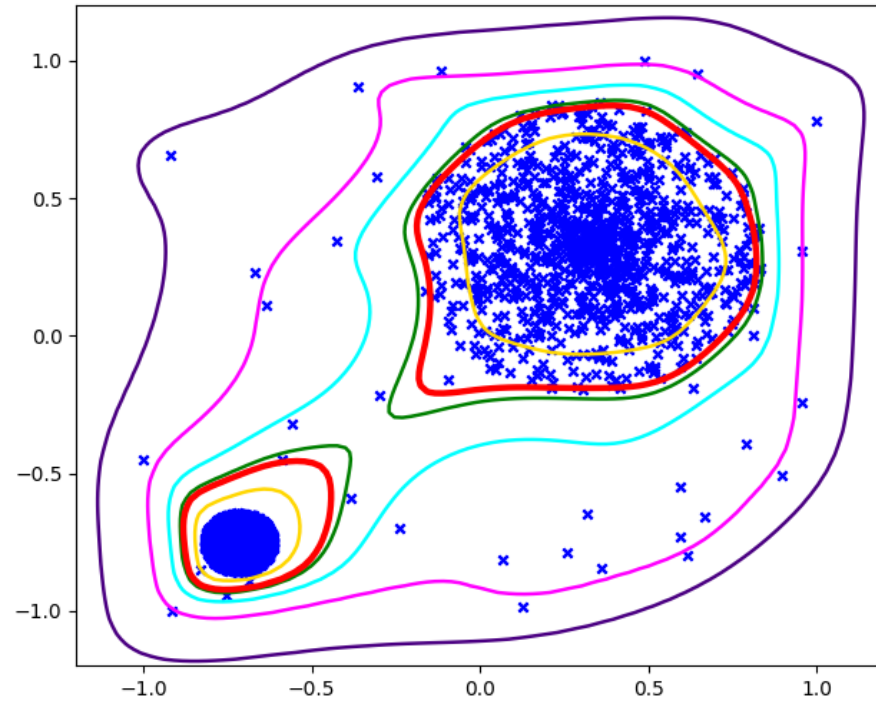
of a certain polynomial $Q_d \in \mathbb{R}[x_1, x_2]$ of
degree $2d$.

Notice that \mathcal{L}_γ captures the shape of the
cloud.

Capturing the shape of a cloud of points (2)

Level sets obtained for a multi-density two disks dataset

CF level sets (d=6)



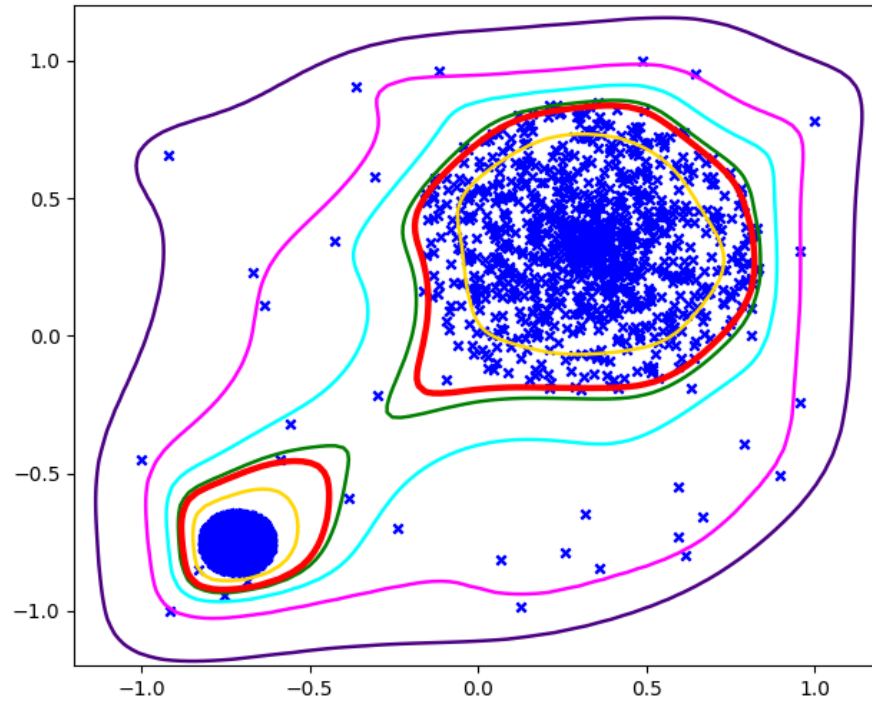
- ✓ The red level set nicely captures the two clusters



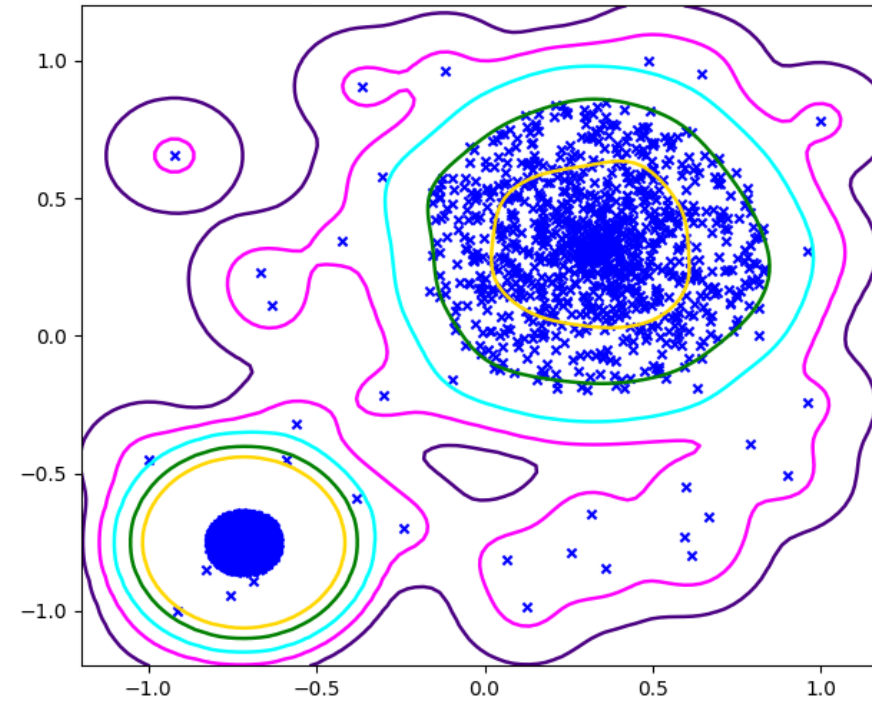
Capturing the shape of a cloud of points (2)

Level sets obtained for a multi-density two disks dataset with CF and MKDE gaussian kernel

CF level sets (d=6)



MKDE level sets



Method	AUROC	AP
CF	0.9745	0.7050
MKDE	0.9645	0.6648



- ▶ Let μ be a Borel measure on a compact set $\Omega \subset \mathbb{R}^p$ with nonempty interior,
- ▶ Form the vector $\mathbf{v}_d(\mathbf{x})$ from a basis of p -variate polynomials of degree at most d :

$$\mathbf{v}_d(\mathbf{x}) = (P_1(\mathbf{x}), \dots, P_{s(d)}(\mathbf{x}))^T \quad \text{of size } s(d) = \binom{p+d}{p}.$$

$$\text{👉 } \mathbf{Q}_d^\mu(\mathbf{x}) = \mathbf{v}_d(\mathbf{x})^T \underbrace{\mathbf{M}_d(\mu)}^{-1} \mathbf{v}_d(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^p$$

Moment matrix of μ

The **Christoffel function** $\Lambda_d^\mu : \mathbb{R}^p \rightarrow \mathbb{R}_+$ is defined by:

$$\Lambda_d^\mu(\mathbf{x})^{-1} = \mathbf{Q}_d^\mu(\mathbf{x})$$

Λ_d^μ encodes properties of the underlying measure μ .

In our case

$$\mu_n := \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}(i)}$$

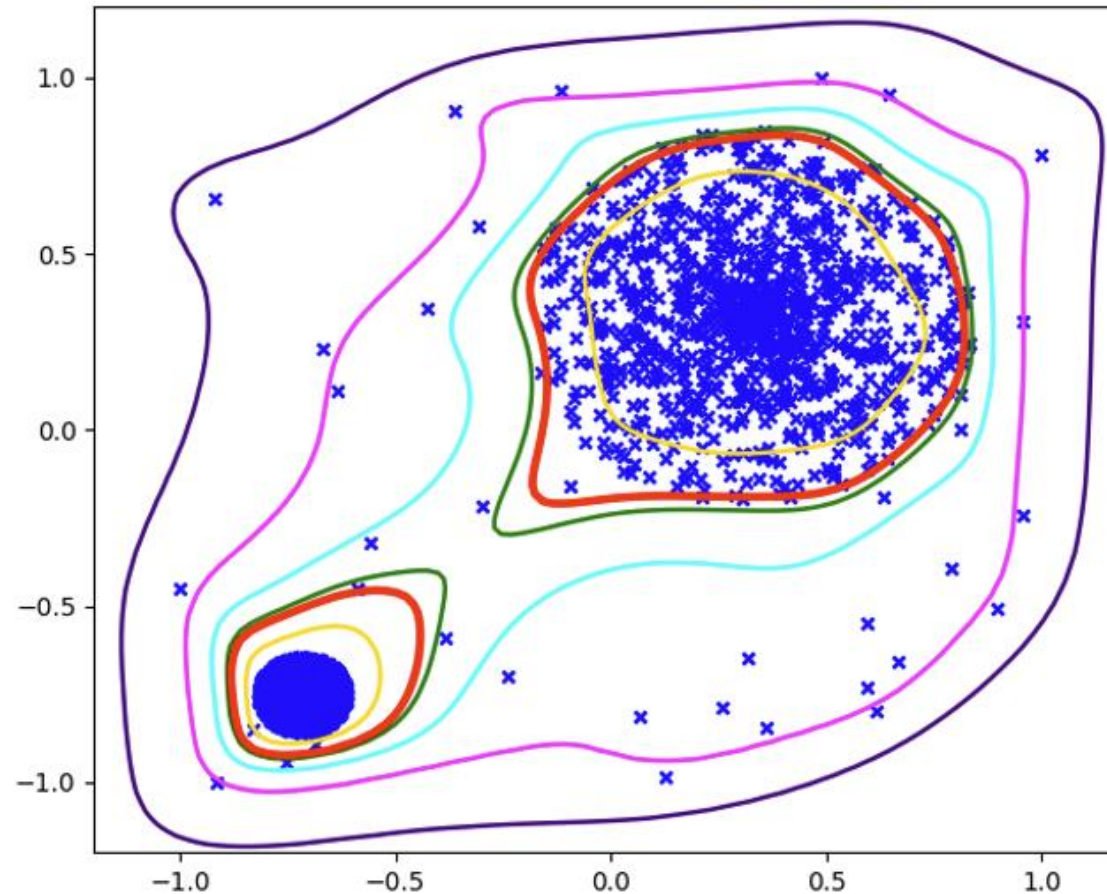
is the EMPIRICAL measure associated with the **cloud** of **data points** $(\mathbf{x}(i))_{i \leq n}$ sampled from an unknown measure μ on Ω .

Empirical moment matrix of μ_n :

$$\mathbf{M}_d(\mu) = \int_{\mathbb{R}^p} \mathbf{v}_d(\mathbf{x}) \mathbf{v}_d(\mathbf{x})^T d\mu(\mathbf{x}) \quad \longrightarrow \quad \mathbf{M}_d(\mu_n) = \sum_{\mathbf{x} \in \mathcal{X}} \mathbf{v}_d(\mathbf{x}) \mathbf{v}_d(\mathbf{x})^T$$

Property: The samples belonging to the support Ω of the empirical measure μ_n are confined by a specific level set $\Omega_{\gamma_{d,p}}$, where $\gamma_{d,p} = C d^{3p/2}$ and C a problem-related constant (cf. (Lasserre *et al.* 2022), Theorem 7.3.3).

CF level sets (d=6)



The red level set corresponds to the set $\Omega_{\gamma_{d,p}}$ with the threshold $\gamma_{d,p} = d^{3p/2}$ as dictated by the CF theory (C=1)

In our case

$$\mu_n := \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}(i)}$$

is the EMPIRICAL measure associated with the **cloud** of **data points** $(\mathbf{x}(i))_{i \leq n}$ sampled from an unknown measure μ on Ω .

 ... and quite remarkably

The **level sets** of $\Lambda_d^{\mu_n}(\mathbf{x})^{-1}$ match the density variations of the **cloud of points** $(\mathbf{x}(i))_{i \leq n}$

→ $\Lambda_d^{\mu_n}(\mathbf{x})^{-1}$ is a **good scoring function for anomaly detection**

In particular, the level set

$$\{\mathbf{x} \in \mathbb{R}^p : \Lambda_d^{\mu_n}(\mathbf{x})^{-1} \leq \gamma_{d,p} = \mathbf{C}d^{3p/2}\}$$

identifies the support Ω of μ , even for moderate values of d .

Dealing with data streams: DyCF method

- 1) Low memory: $\mathbf{M}_d(\mu_n)^{-1}$ can be seen as an encoding of the whole data set
- 2) Low computation: incremental update of $\Lambda_d^{\mu_n}(\mathbf{x})^{-1}$ with rank-one update of the inverse $\mathbf{M}_d(\mu_n)^{-1}$

When a point ξ is added to the cloud of r points, i.e.,

$$\mu_n \rightarrow \frac{1}{n+1} (n \mu_n + \delta_\xi)$$

→ a new cloud with $n+1$ points

👉 The Sherman-Morrison-Woodbury formula allows for a simple **RANK-ONE UPDATE** of the inverse $\mathbf{M}_d(\mu_n)^{-1}$

Incremental update

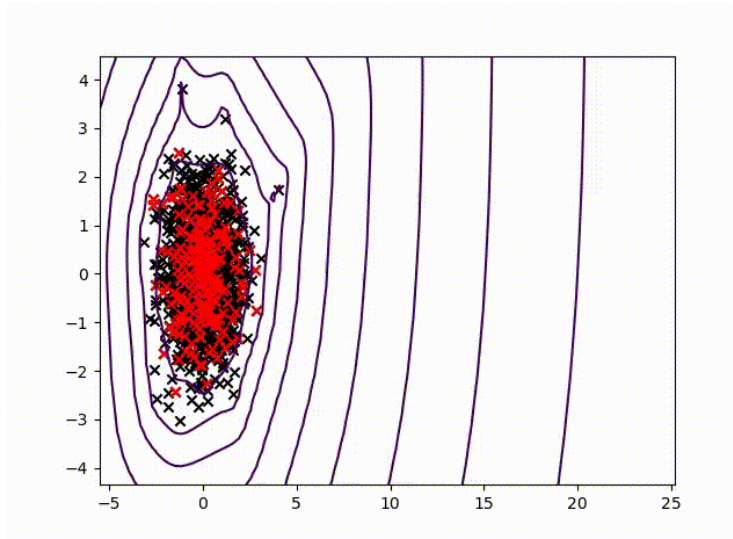
The moment matrix $\mathbf{M}_d(\mu_n)^{-1}$ can be rewritten with the incremental formula:

$$\mathbf{M}_d(\mu_{n+1}) = \frac{1}{n+1} [n\mathbf{M}_d(\mu_n) + \mathbf{v}_d(\mathbf{x}_{n+1})^T \mathbf{v}_d(\mathbf{x}_{n+1})]$$

Incremental inversion of a matrix A with the Sherman-Morrison formula:

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1}u}$$

$$\mapsto ((n+1)\mathbf{M}_d(\mu_{n+1}))^{-1} = (n\mathbf{M}_d(\mu_n))^{-1} - \frac{(n\mathbf{M}_d(\mu_n))^{-1} \mathbf{v}_d(\mathbf{x}_{n+1}) \mathbf{v}_d(\mathbf{x}_{n+1})^T (n\mathbf{M}_d(\mu_n))^{-1}}{1 + \mathbf{v}_d(\mathbf{x}_{n+1})^T (n\mathbf{M}_d(\mu_n))^{-1} \mathbf{v}_d(\mathbf{x}_{n+1})}$$



DyCF requires **only one parameter** to be fixed: d

The theory dictates to use the level set defined by $\Omega_{\gamma_{d,p}}$, where $\gamma_{d,p} = Cd^{3p/2}$.

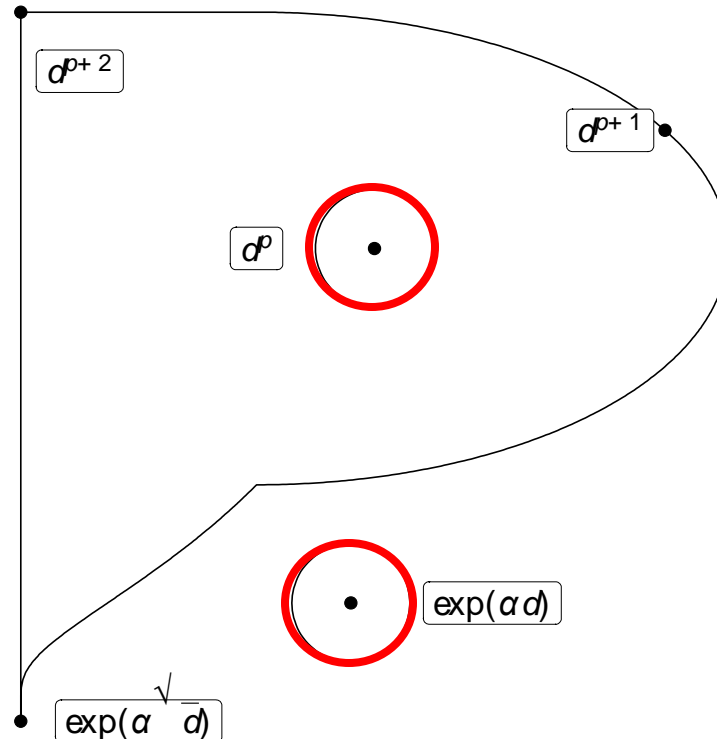
$$\text{Normalized scoring function : } S_{d,p}(\mathbf{x}) = \frac{\Lambda_d^\mu(\mathbf{x})^{-1}}{\gamma_{d,p}}.$$

If $C=1$, a point \mathbf{x} is defined as an **outlier** if $S_{d,p}(\mathbf{x}) \geq 1$.

Leveraging the growth properties of CF

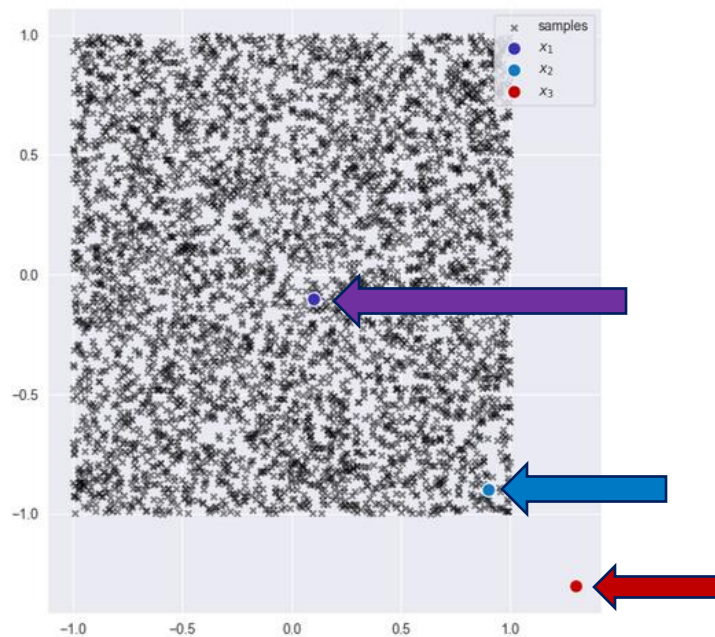
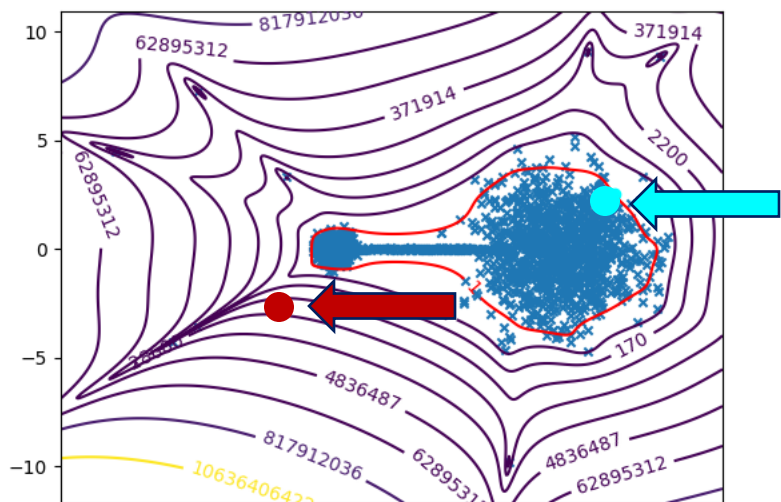
As d grows, $\Lambda_d^\mu(\mathbf{x})^{-1}$ has:

{	POLYNOMIAL growth	INSIDE Ω
	EXPONENTIAL growth	OUTSIDE Ω



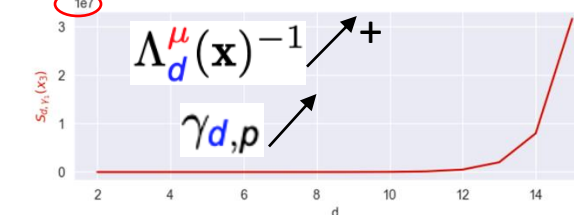
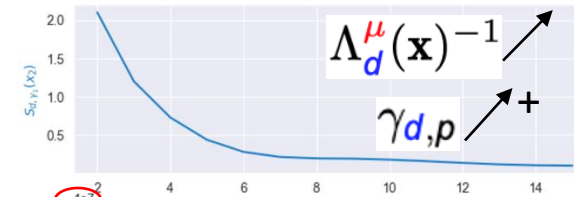
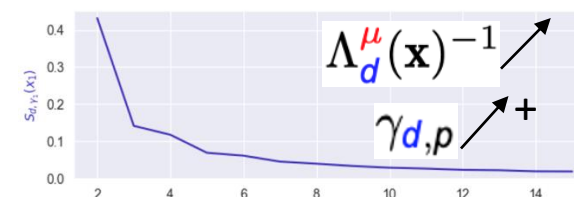
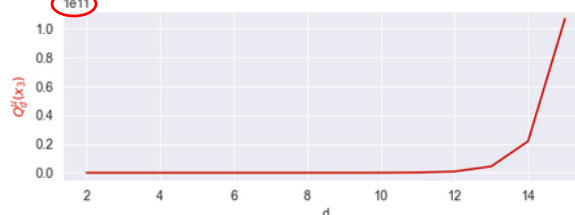
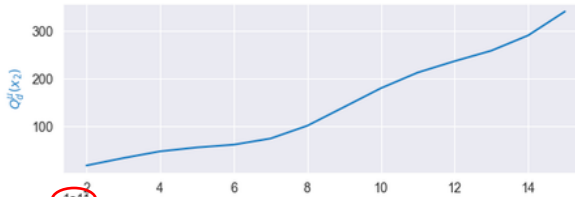
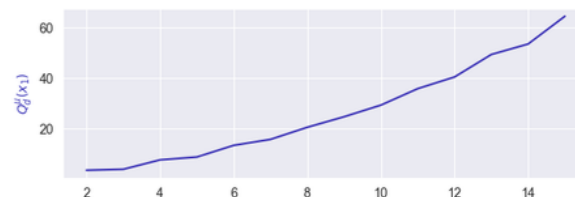
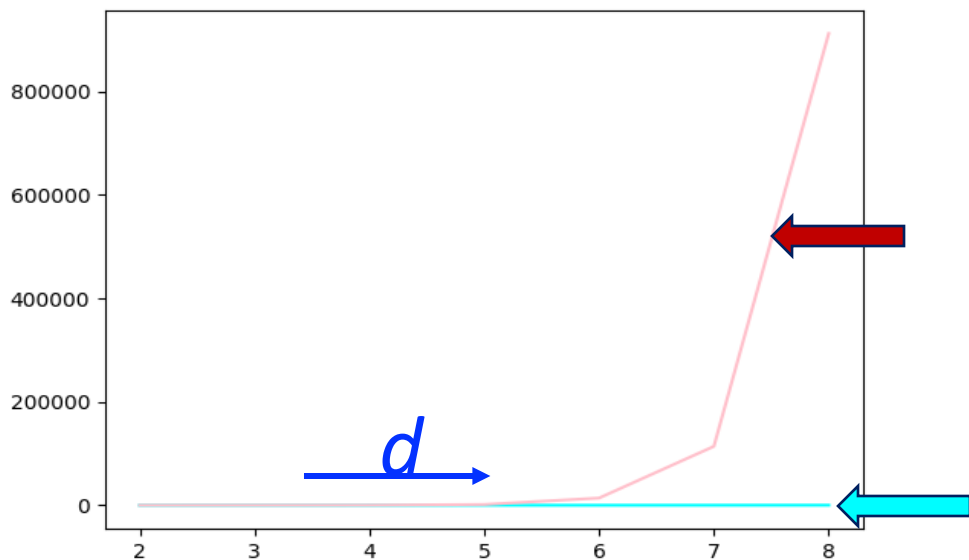
Cf. (Lasserre *et al.* 2022), Lemmas 4.3.1 and 4.3.2)

CF growth property



Normalized score:

$$S_{d,p}(\mathbf{x}) = \frac{\Lambda_d^\mu(\mathbf{x})^{-1}}{\gamma_{d,p}}$$



DyCG : two DyCF models of degrees d_{min} and d_{max}

$$\text{DyCG scoring function: } S'_{d_{max}, d_{min}, p}(\mathbf{x}) = \frac{S_{d_{max}, p}(\mathbf{x}) - S_{d_{min}, p}(\mathbf{x})}{d_{max} - d_{min}}$$

Outlierness threshold is 0:

$$\text{Inliers} \longrightarrow S_{d_{max}, p}(\mathbf{x}) < S_{d_{min}, p}(\mathbf{x}) \longrightarrow S'_{d_{max}, d_{min}, p}(\mathbf{x}) < 0$$

$$\text{Outliers} \longrightarrow S_{d_{max}, p}(\mathbf{x}) \geq S_{d_{min}, p}(\mathbf{x}) \longrightarrow S'_{d_{min}, d_{max}, p}(\mathbf{x}) \geq 0$$

d_{min} and d_{max} are fixed at 2 and 8 once and for all: **DyCG is tuning free.**

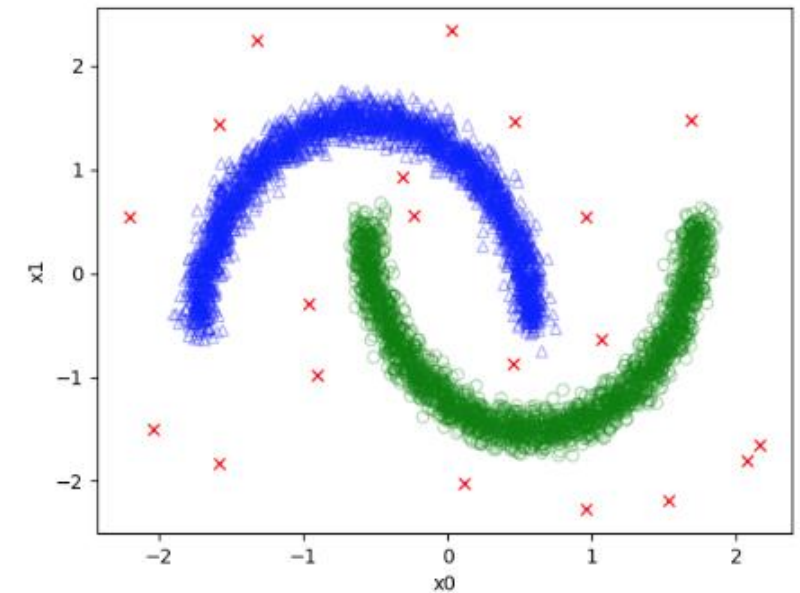
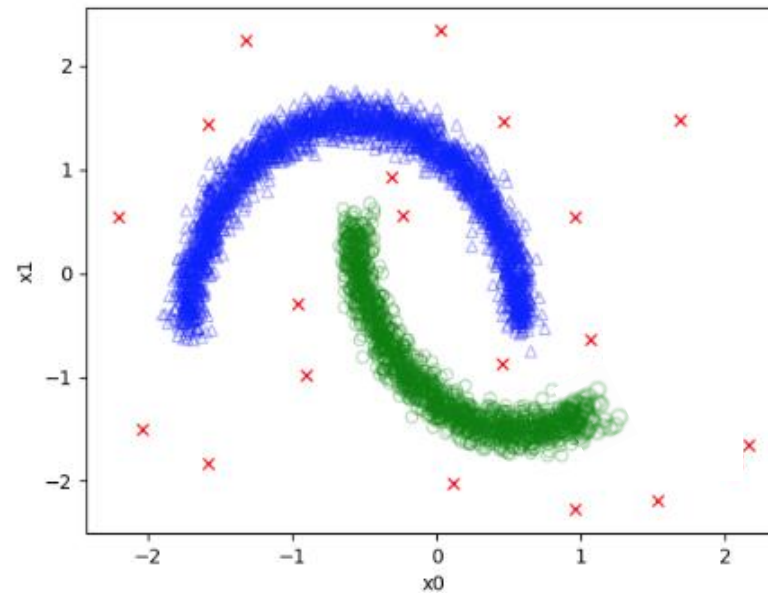
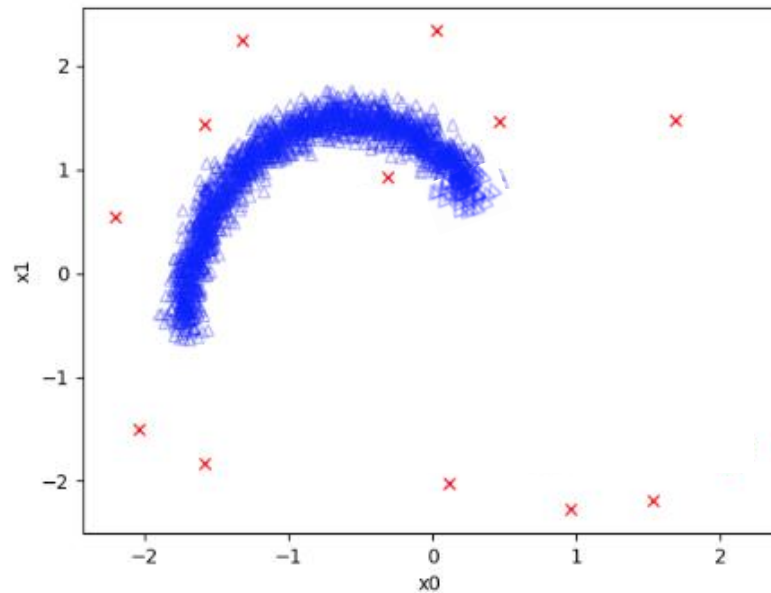
Evaluation

Two moons data stream (tm)

2 x 2500 samples building two moons
(500 for initial training)

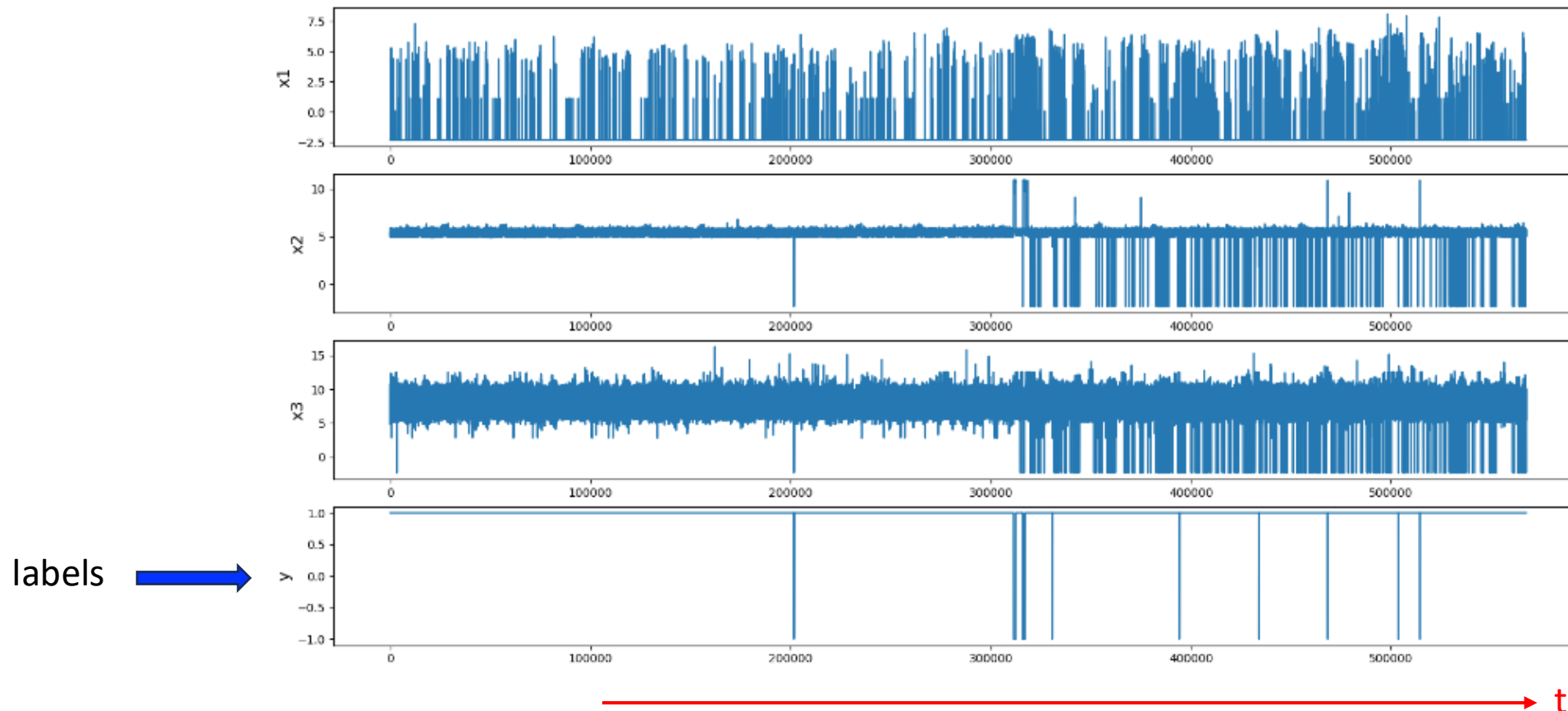
+

20 outliers uniformly distributed



Http data stream (http)

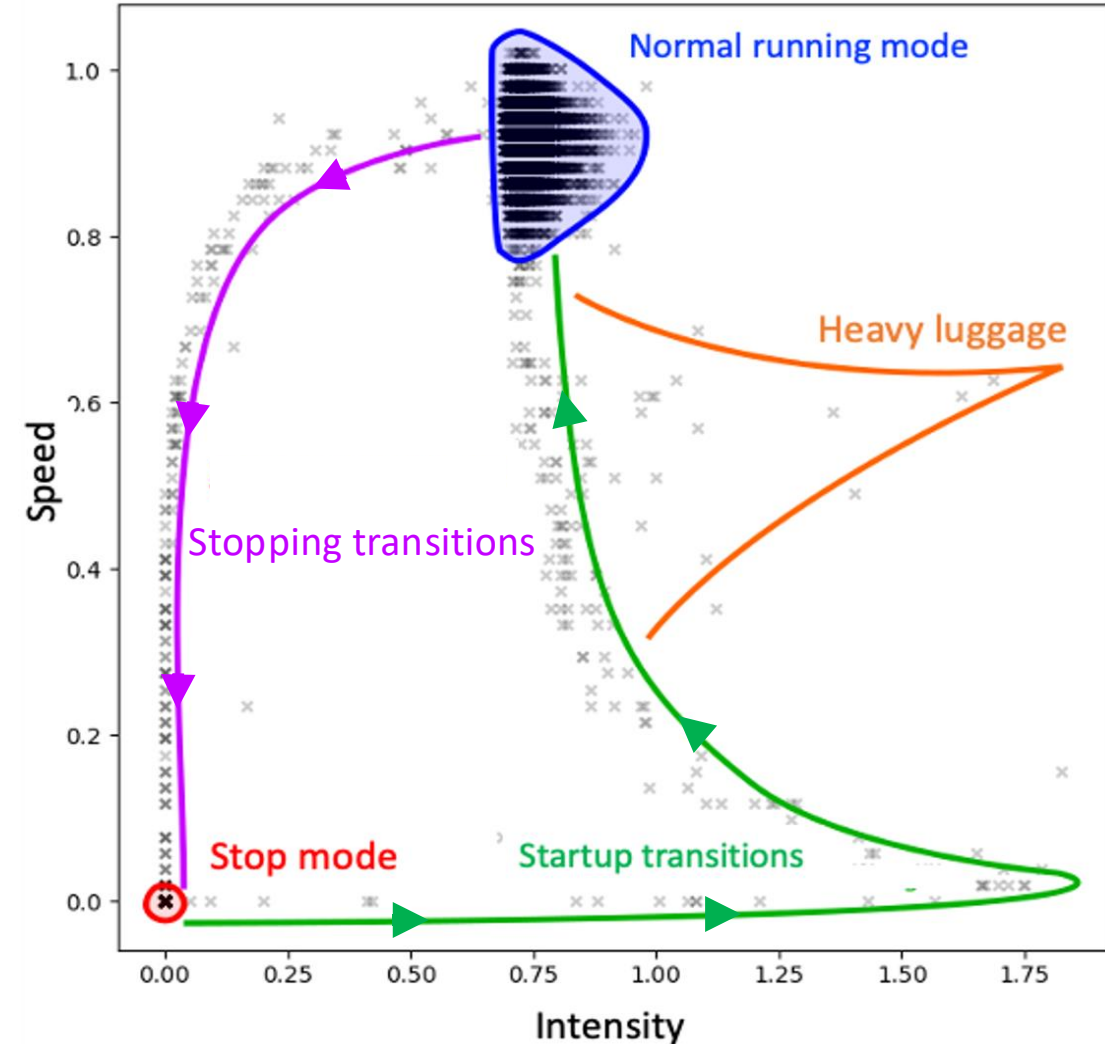
From the Outlier Detection Data Streams ODDS library (<https://odds.cs.stonybrook.edu/>)
3 variables, 567498 observations (50000 for initial training), 2211 outliers ($\approx 0,4\%$)



Industrial luggage conveyor data stream (Ic)

Carl Berger-Levrault project

Multimode system, 2 variables, 166926 observations (15000 for initial training), 17 introduced outliers



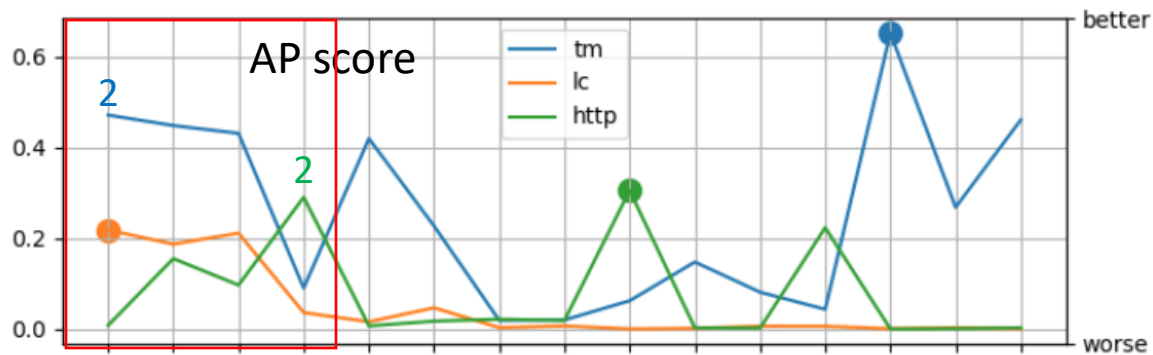
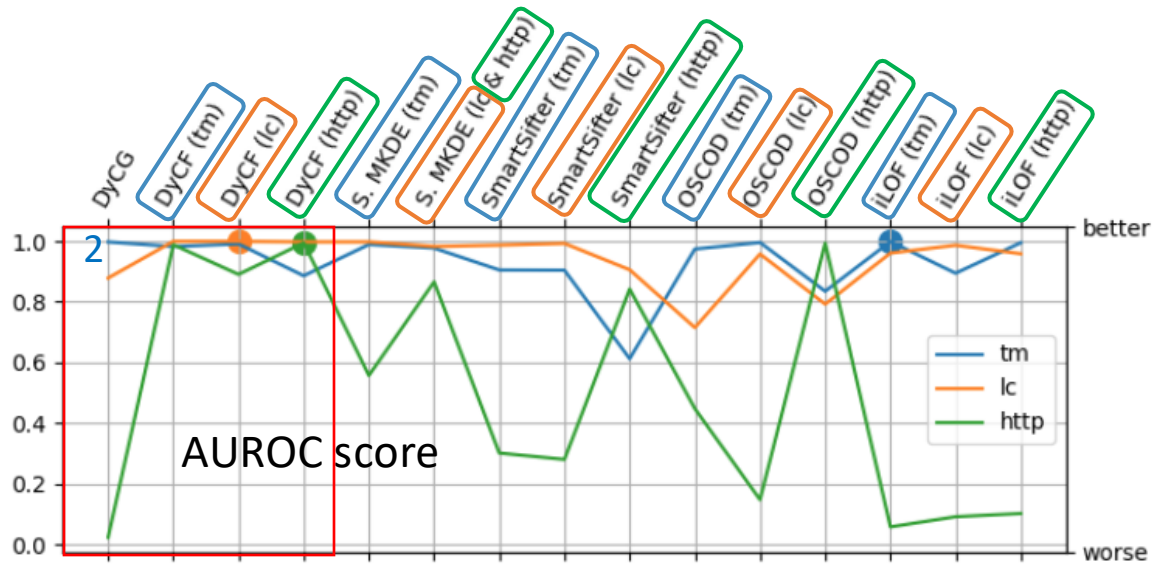
Methods

Method	Parameters	Range	ω_{tm}	ω_{lc}	ω_{http}
Sliding window Multivariable Kernel Density Estimation	threshold (not informed)	-	-	-	-
	window size W	{100, 200, 500, 1000, 2000, 5000}	200	5000	5000
	kernel (fixed)	gaussian	-	-	-
	bandwidth (fixed)	Scott's rule	-	-	-
Based on discounting learning of a GMM model	threshold (not informed)	-	-	-	-
	nb of gaussians k	{2, 5, 10, 15, 20}	15	10	2
	discounting param r	{0.001, 0.005, 0.01}	0.01	0.01	0.001
	stability param α	{1, 1.5, 2}	1.5	1	2
Based on the number (parameter k) of neighboring points laying at a given distance (parameter R)	nb of neighbors k (not informed)	-	-	-	-
	radius R	{0.1, 0.2, 0.5, 1, 1.2, 1.5}	1.2	0.2	0.5
	window size W	{100, 200, 500, 1000, 2000, 5000}	200	100	5000
KNN based method that contrasts sample local density with that of its neighbors	threshold (not informed)	-	-	-	-
	nb of neighbors k	{5, 10, 15, 20, 25, 30}	5	25	30
	window size W	{100, 200, 500}	200	100	500
DyCF	degree d	{2, 4, 6, 8}	6	8	2
DyCG	no need	-	-	-	-

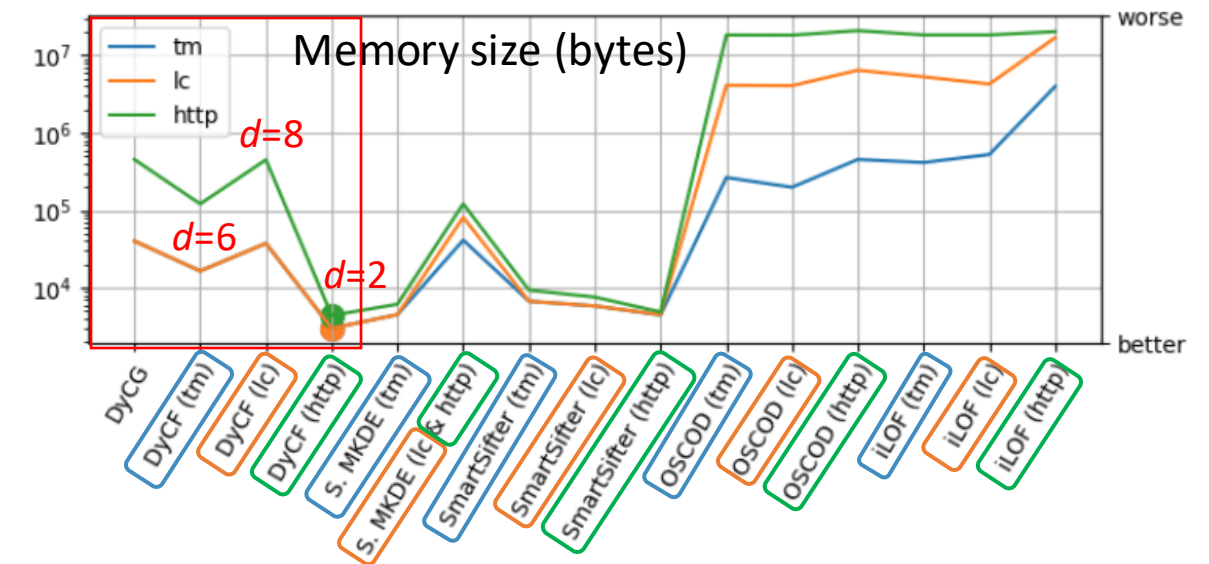
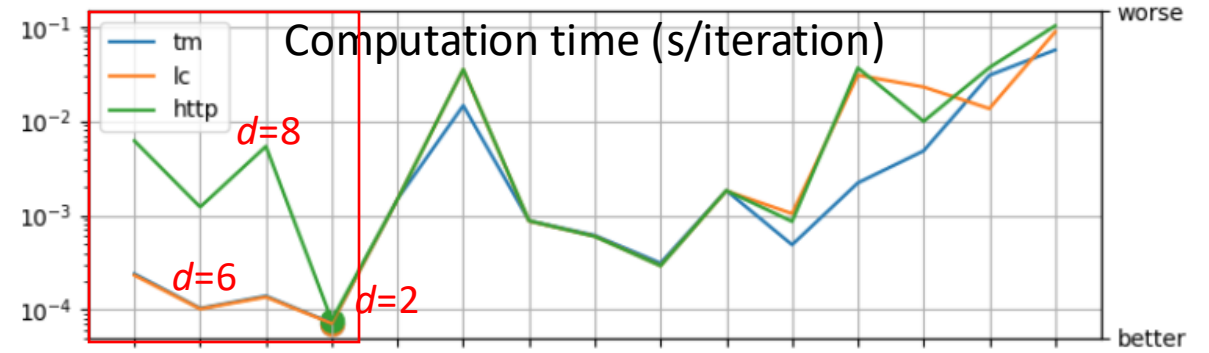
No deep learning method because no frugality, no fast update, no low tuning.

- ▶ AUROC (Area Under the ROC Sensitivity-Specificity curve)
 - The higher the better (a value of 0.5 is not better than a random classifier)
- ▶ AUPRC (Area Under the Precision-Recall Curve) estimated as AP (Average Precision)
 - Higher value indicates better precision-recall performance
 - Relevant for imbalanced data sets
- ▶ Computation time for one iteration
- ▶ Memory size

Results

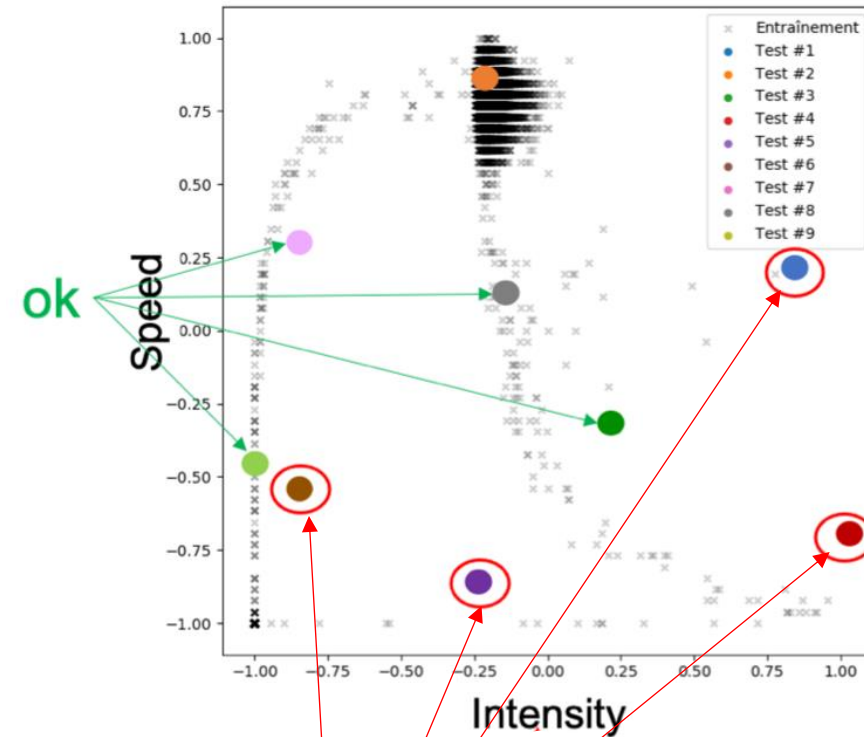


- DyCF and DyCG are globally positioned at the top
- Best scores for the industrial data stream
- DyCG and ILOF are bad for the http data stream ($p=3, d=8$)



Industrial luggage conveyor data stream (Ic)

Carl Berger-Levrault project



Anomalies

- ▶ *Dy-CF* and *Dy-CG* are simple and easy-to-use methods with **little or no tuning at all**
- ▶ They achieve excellent results compared to other more tricky anomaly detection methods
- ▶ The Christoffel function provides interesting **theoretical foundations**
- ▶ It nicely deals with data streams thanks to the moment matrix encoding and its **incremental update**
- ▶ Future work:
 - ▶ adding forgetting ability
 - ▶ scaling up to high dimensions
 - ▶ extend to abnormal trajectory detection
 - ▶ integrate with other tasks.