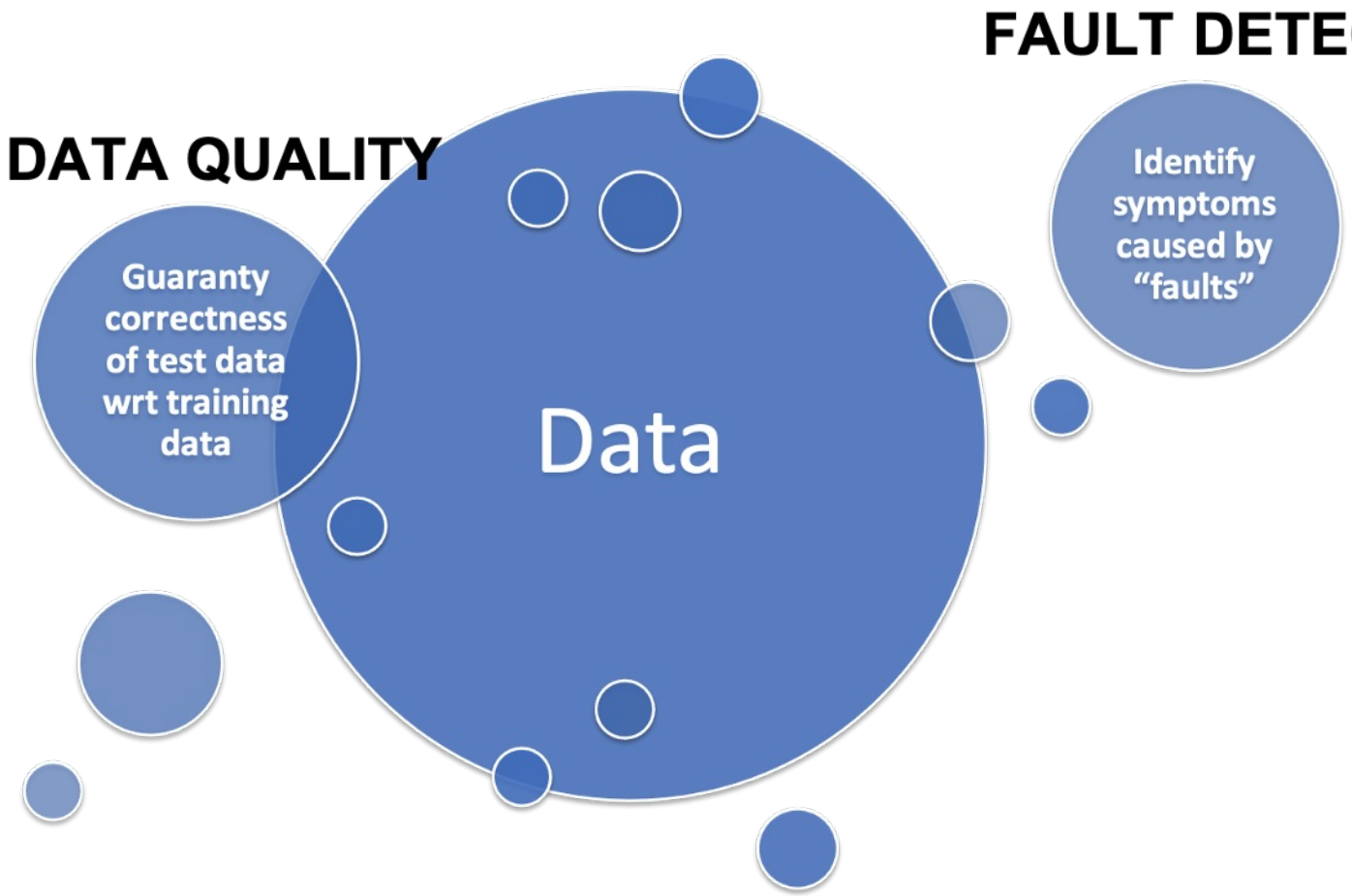
Leveraging the properties of the Christoffel function for anomaly detection in data streams

Louise Travé-Massuyès

Kévin Ducharlet, Jean-Bernard Lasserre

2ème Congrès de la SAGIP
29-31 mai 2024
Villeurbanne

Anomaly detection



Are there intruders ?

Anomaly

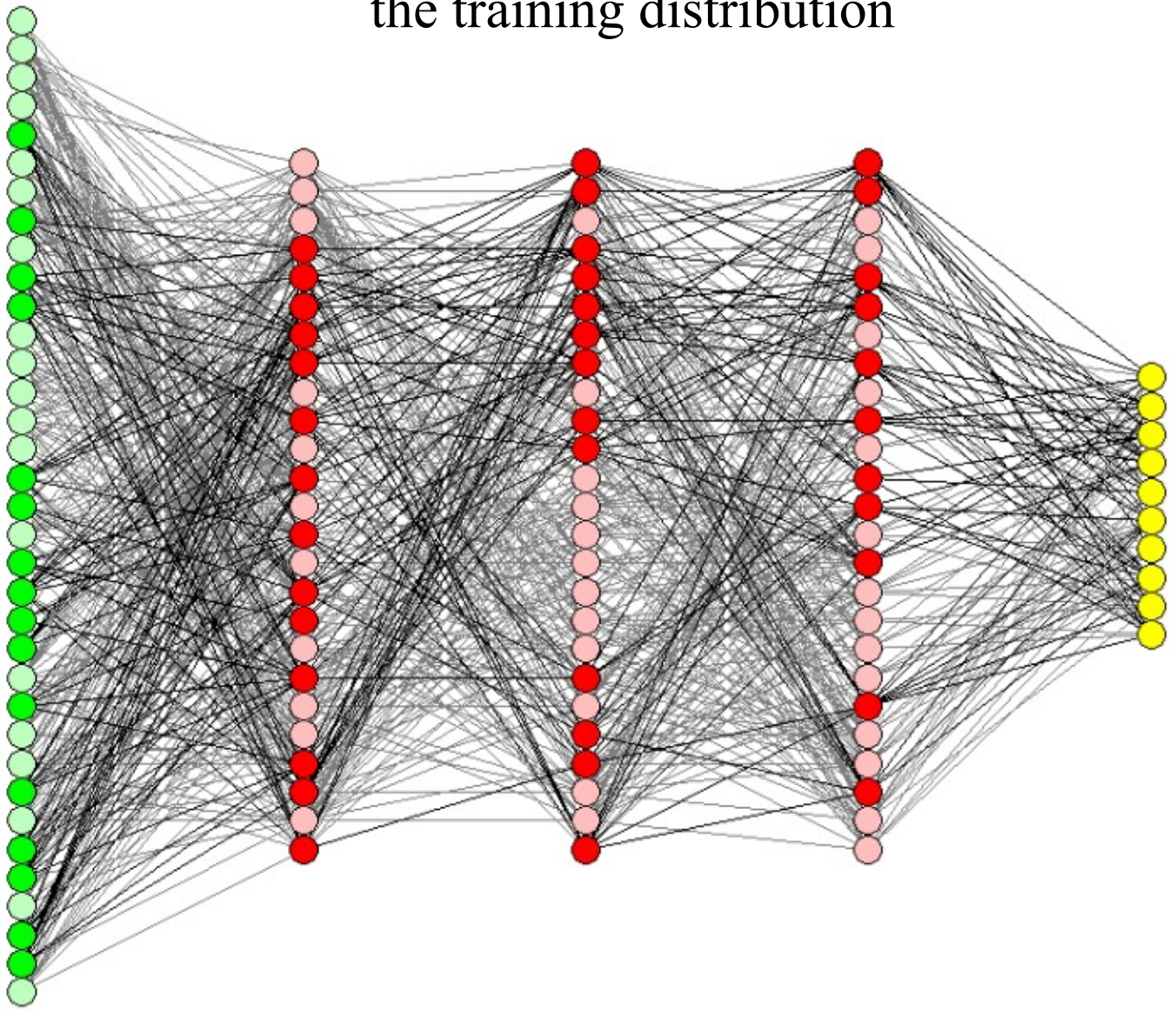
- an instance from a distinct distribution
- a rare or low-probability instance

↳ *outlier, out-of-distribution (OOD) sample, or novelty*

Data quality

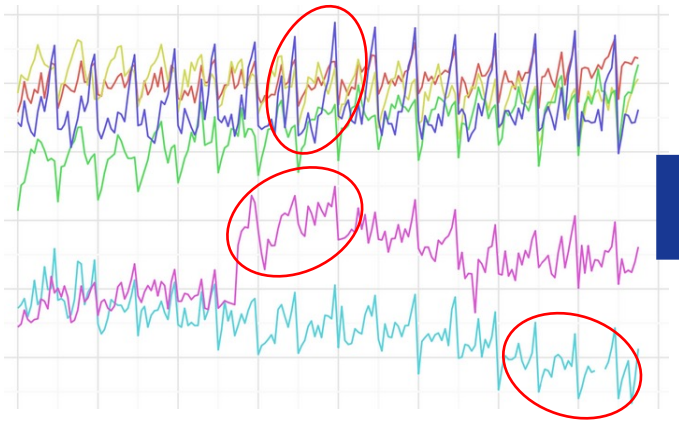
Training data

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



Fault detection

Anomalies

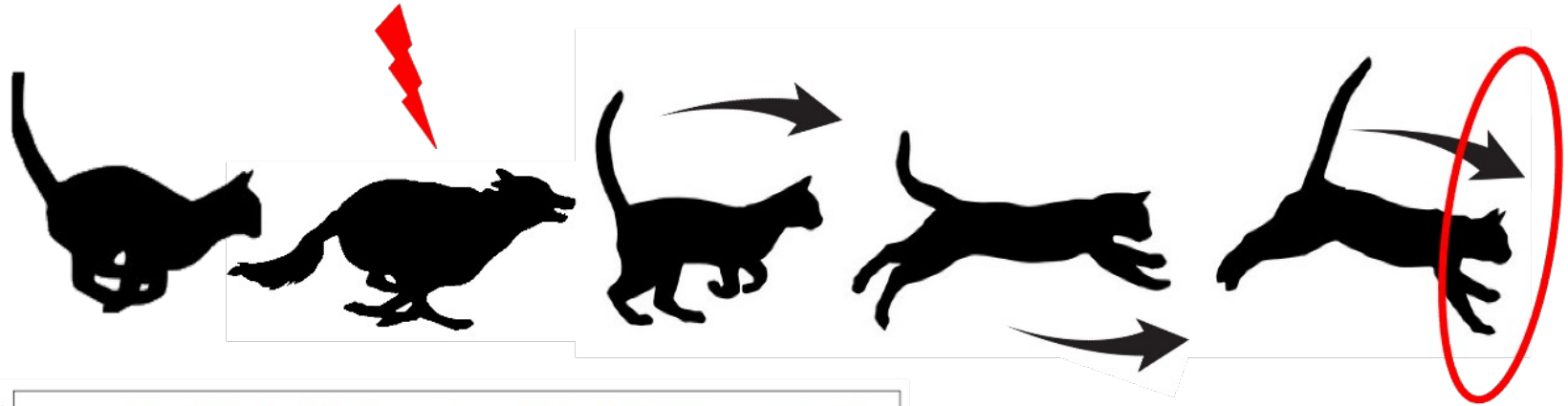


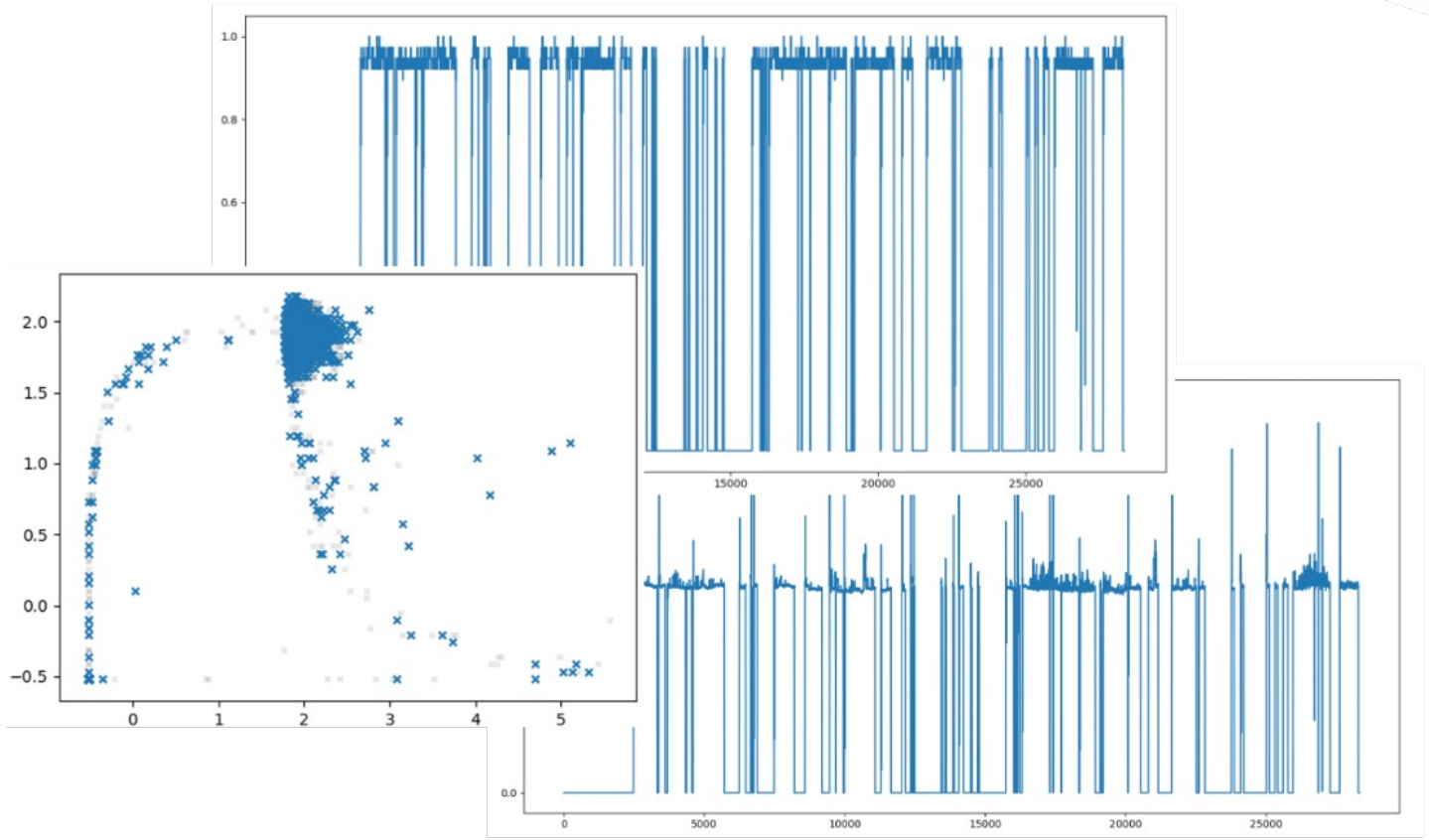
Symptoms

DIAGNOSIS

Root cause

Anomaly detection in data streams





Data stream requirements

- detection on the fly

No knowledge
Just data

Peculiarities of data streams and requirements

- **Transiency**: the significance of each data point diminishes over time
- **Time dependency**: a data point is assessed in a changing temporal context
- **Infinity**: samples cannot be stored in memory entirely
- **Arrival rate**: may vary over time
- **Concept drift**: data distributions are non-stationary
- **Non labelled data**: continuous evolution renders labeling impractical and it can swiftly become outdated

- *fast update* to match incoming measurement frequency
- *Dynamic learning* to deal with non stationarity
- *frugality* allowing to embed outlier detection models in low memory and CPU capacity devices
- *Unsupervised dynamic learning* to handle unlabelled data

+

Little or no fine-tuning to meet automation and generalization needs

State of the Art

Long time research initiated by statisticians:

F. Y. Edgeworth (1887) XLI. *On discordant observations*. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 23(143):364–375. https://doi.org/10.1080/14786448708628471

- *Supervised methods* that rely on the availability of datasets labeled with the outlierness status of samples
- *Semi-supervised methods* that rely on datasets in which only normal samples are labeled
- *Unsupervised methods* that can accept datasets without any information on outlierness

State of the Art

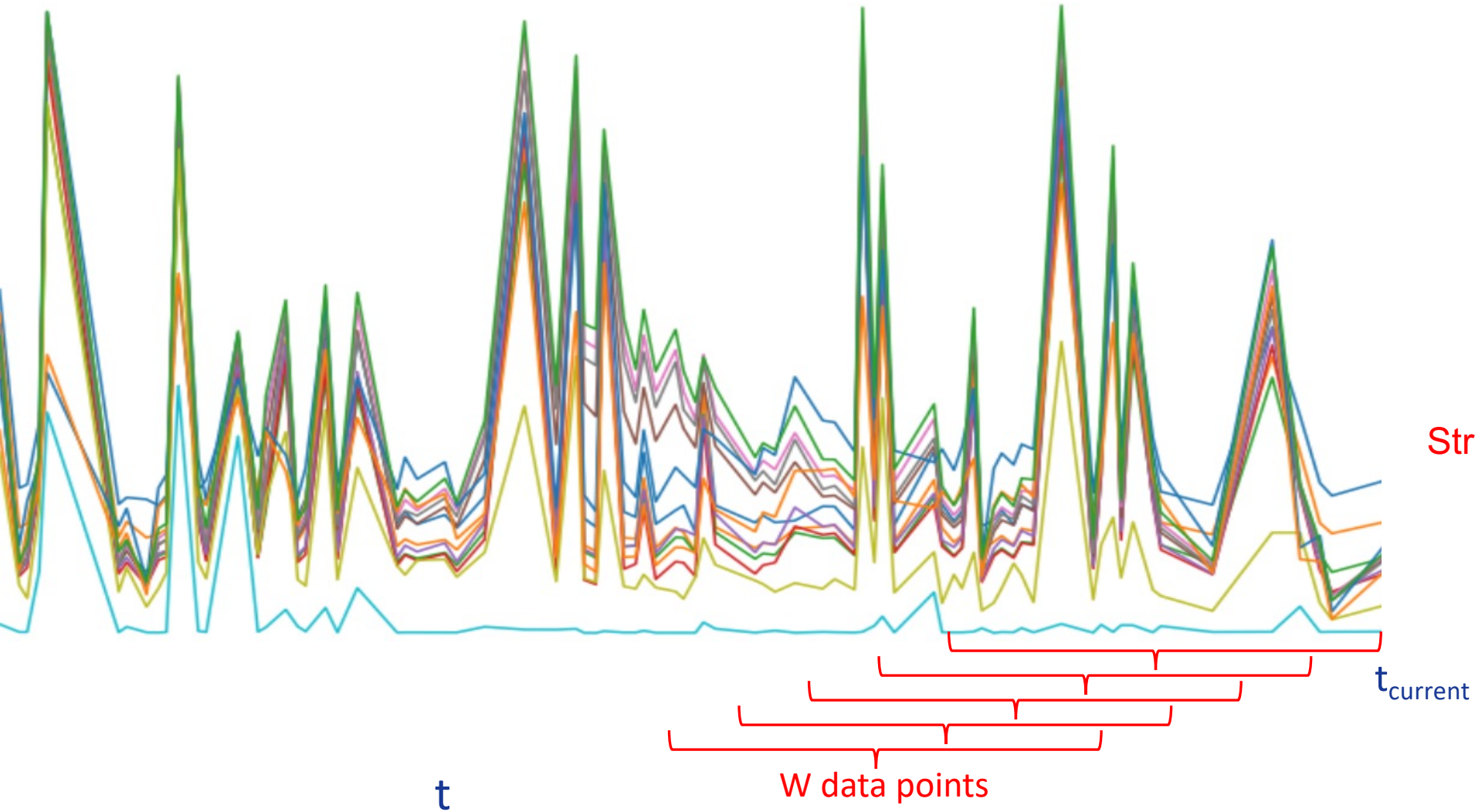
- Adaptation of time series methods
  - Prediction models, e.g. based on exponential smoothing or LSTM
  - Trend and seasonality must follow a regular pattern that is not guaranteed with data stream
  - Do not account for concept drift

State of the Art

- Adaptation of time series methods
  - Prediction models, e.g. based on exponential smoothing or LSTM
  - Trends and seasonalities must follow a regular pattern that is not guaranteed with data stream
  - Do not account for concept drift
- Dynamic clustering
  - Outliers do not belong to clusters or are in low density clusters
- Methods relying on kNN *(k Nearest Neighbours)*
  - Based on number of neighbours, e.g., MCOD
  - Based on local density (LOF and variants)
- Statistical methods
  - Parametric methods, e.g., based on *Gaussian Mixed Models* (Smartsifter)
  - 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 no memory of previously acquired knowledge,

Windowing techniques



- Landmark windows
- Sliding windows
- Damped windows
- Adaptive windows

Strong window size dependency

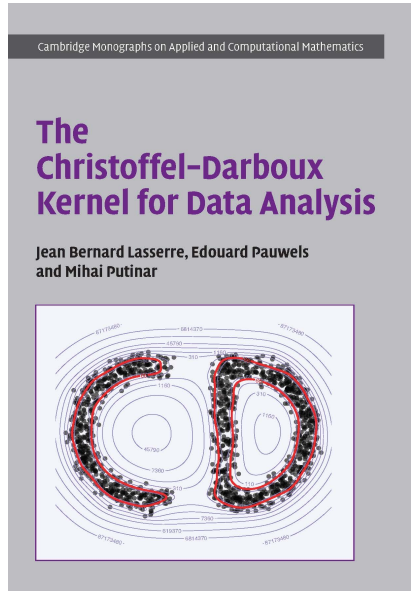
W ↗ : satisfying results

W ↘ : fast model update

This work

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



(Lasserre, Pauwels & Putinar 2022)
https://doi.org/10.1017/ 9781108937078

This work

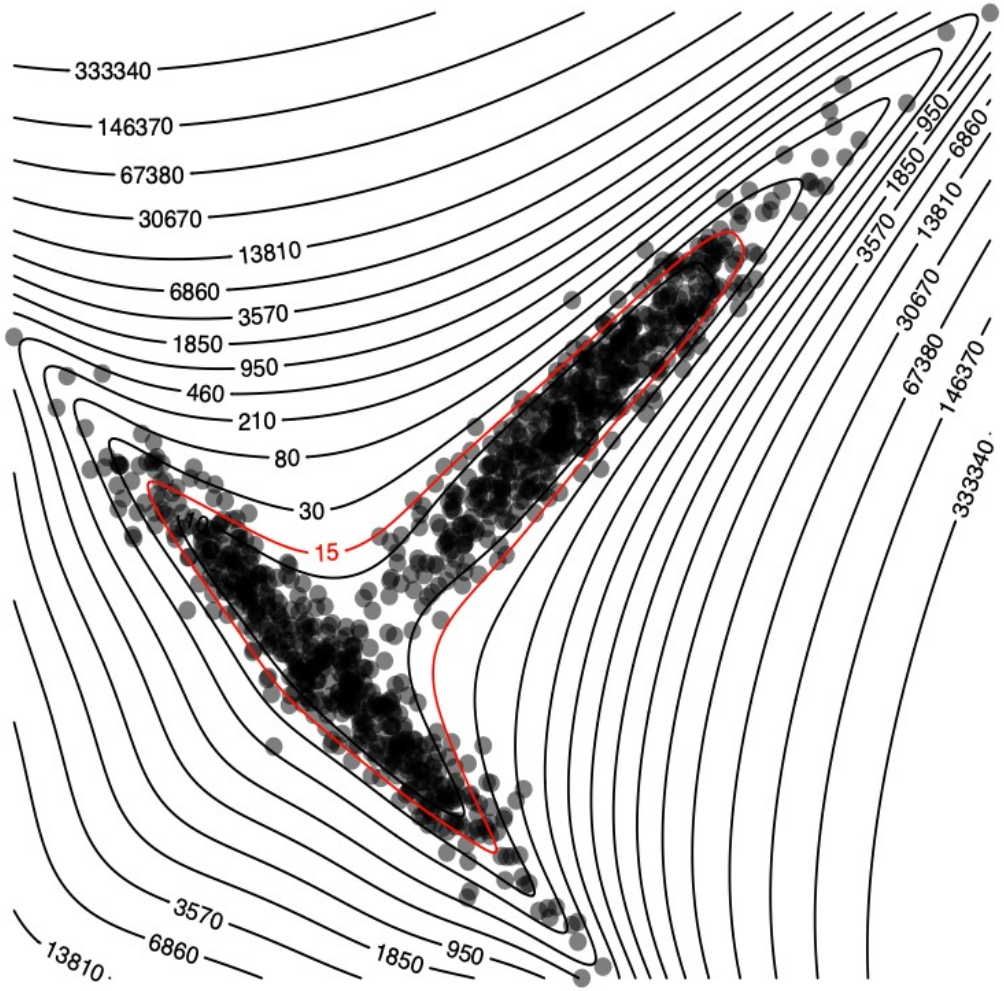
A collaboration between two ANITI chairs:

- **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 Septembre 28, 2023.

Capturing the shape of a cloud of points



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

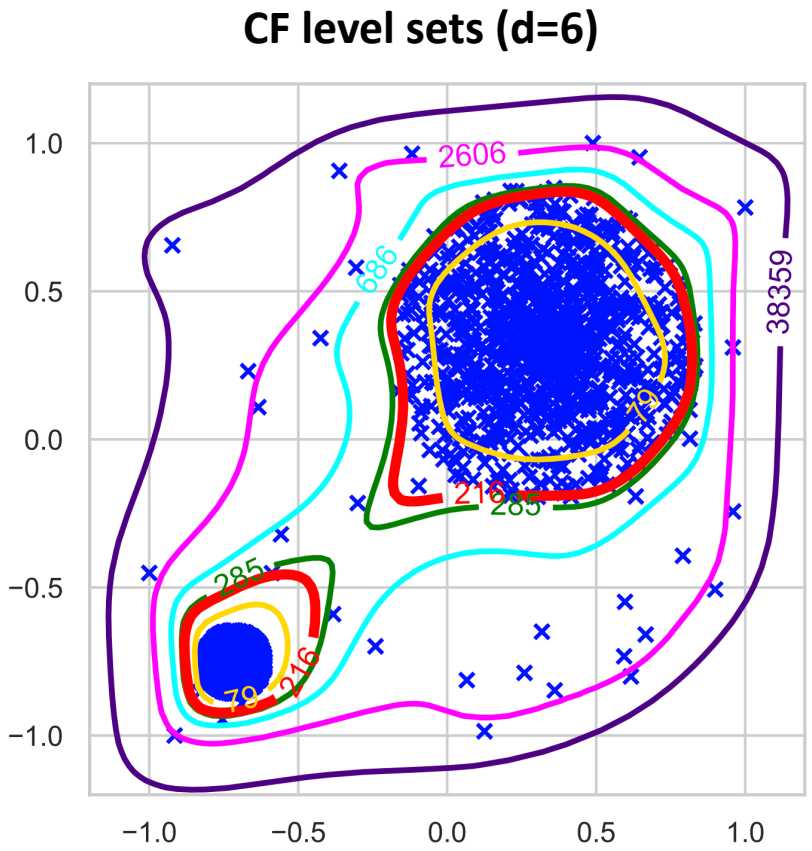
The red curve is the level set:
$\mathcal{L}_\gamma = \{\mathbf{x} : Q_d(\mathbf{x}) \leq \gamma\}, \gamma \in \mathbb{R}_+$

of a certain polynomial $Q_d \in \mathbb{R}[x1, x2]$ of degree $2d$.

Notice that $\mathcal{L}_\gamma$ 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

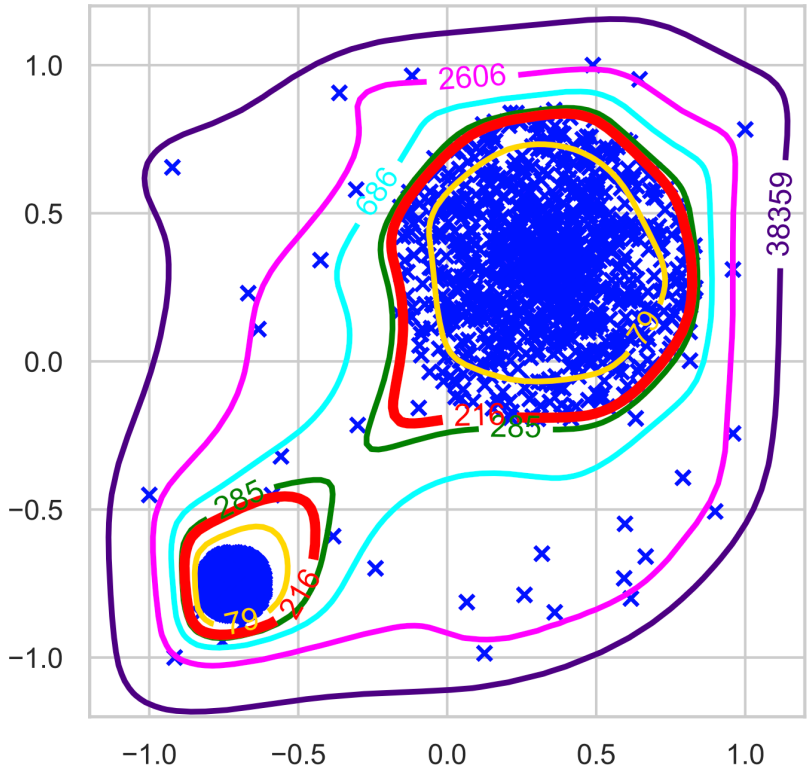


✓ The red level set nicely captures the two clusters 😊

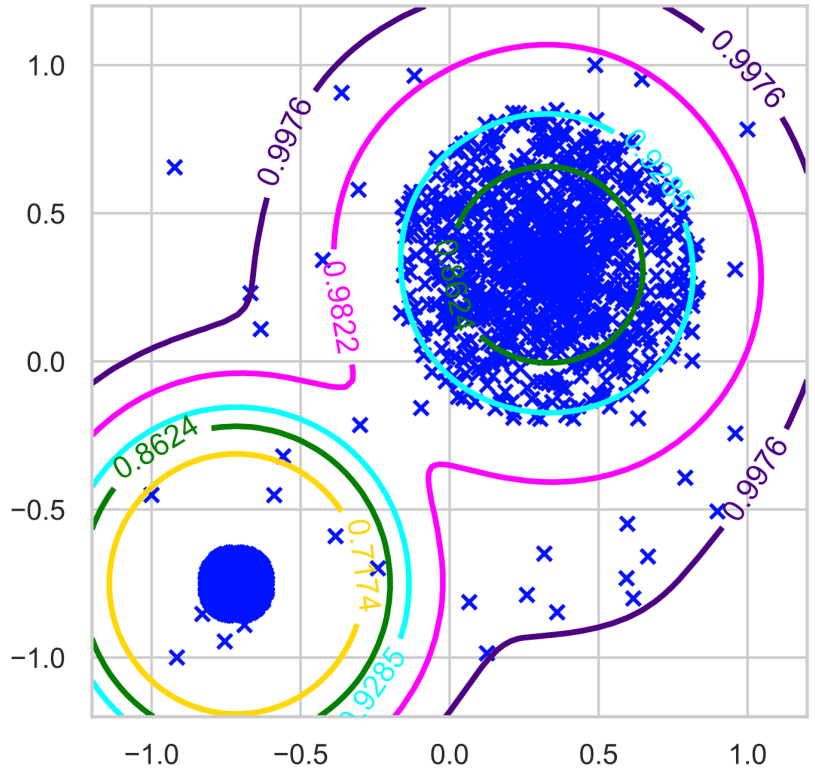
Capturing the shape of a cloud of points (2)

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	0.9644	0.7250
KDE	0.9372	0.6042

The Christoffel function

- 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:

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

$$\text{☞} \quad Q_d^\mu(\mathbf{x}) = \mathbf{v}_d(\mathbf{x})^T \underbrace{\mathbf{M}_d(\mu)^{-1}}_{\text{Moment matrix of } \mu} \mathbf{v}_d(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^p$$

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

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

Λ_d^μ encodes properties of the underlying measure μ.

Empirical measure

In our case

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

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

Empirical moment matrix of μ_n:

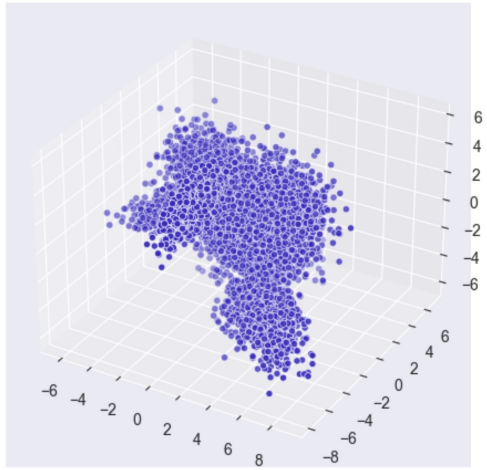
$$\mathbf{M}_d(\mu) = \int_{\mathbb{R}^p} \mathbf{v}_d(\mathbf{x})\mathbf{v}_d(\mathbf{x})^T \, d\mu(\mathbf{x}) \quad \longrightarrow \quad \mathbf{M}_d(\mu_n) = \tfrac{1}{n}\textstyle\sum_{\mathbf{x}\in\mathcal{X}} \mathbf{v}_d(\mathbf{x})\mathbf{v}_d(\mathbf{x})^T$$

If n is large enough compared to d, it has been proved that, by the strong law of large numbers, $\Lambda_d^{\mu_n}$ and Λ_d^{μ} share the same properties.

Λ_d^{μ} encodes properties of the underlying measure μ.

Moments

Moments serve to quantify three essential parameters of distributions: location, shape and scale.



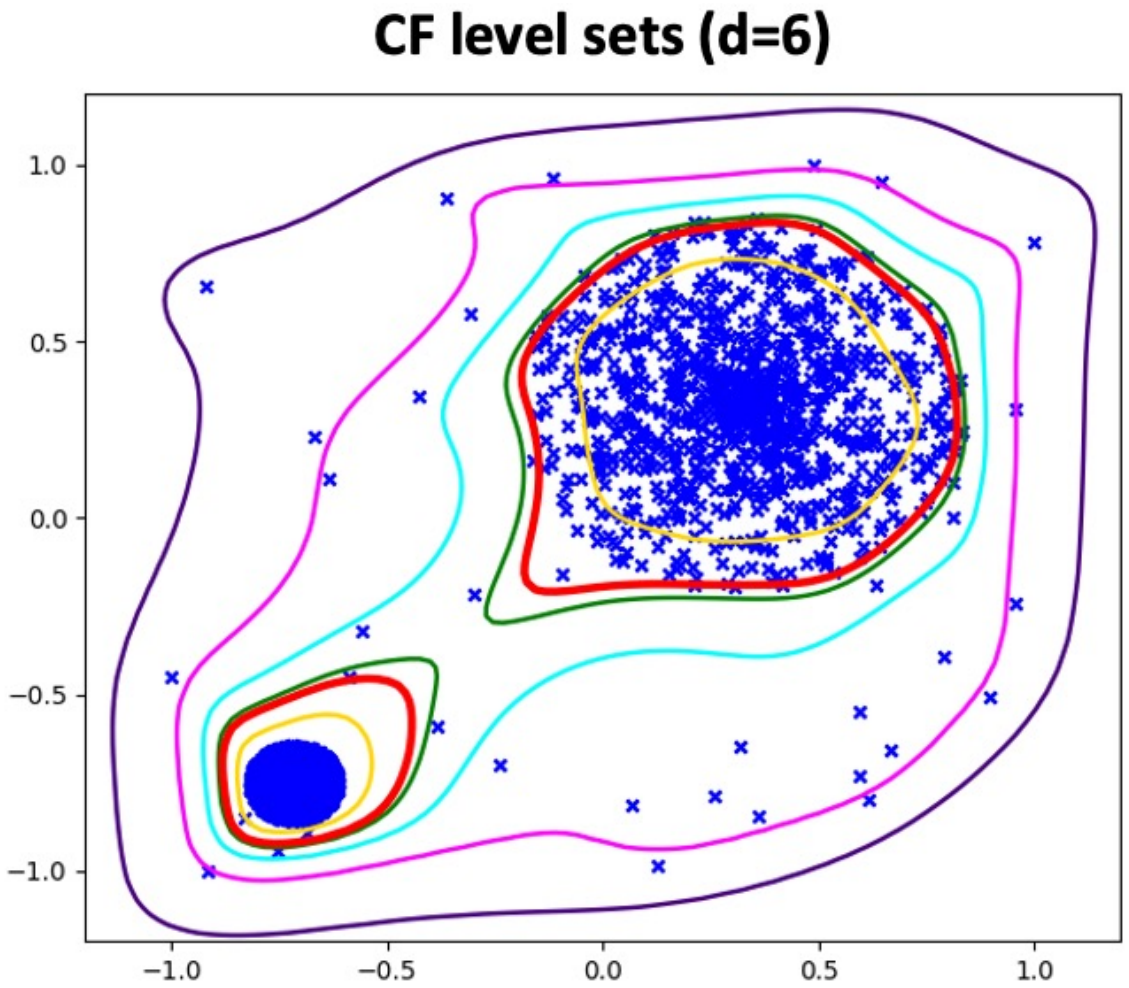
- The location of a distribution is given by the position of its center of mass.
- The scale denotes the extent to which a distribution is spread out
- the shape of a distribution encompasses its overall geometry, including characteristics such as bimodality, asymmetry, and heavy-tailedness.

- The first moment delineates a distribution's location
- The second moment characterizes its scale
- Higher moments collectively elucidate its shape

Capturing the support

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 $\gamma_{d,p} = Cd^{3p/2}$ and C a problem-related constant (cf. (Lasserre *et al.* 2022), Theorem 7.3.3).

Specific level set



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 summary

In our case

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

is the EMPIRICAL measure associated with the cloud of data points $(x(i))_{i\leq n}$ 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 $(x(i))_{i\leq n}$

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

In particular, the level set

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

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 en 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 n points, i.e.,

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

$\rightarrow$ a new cloud with n+1 points

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

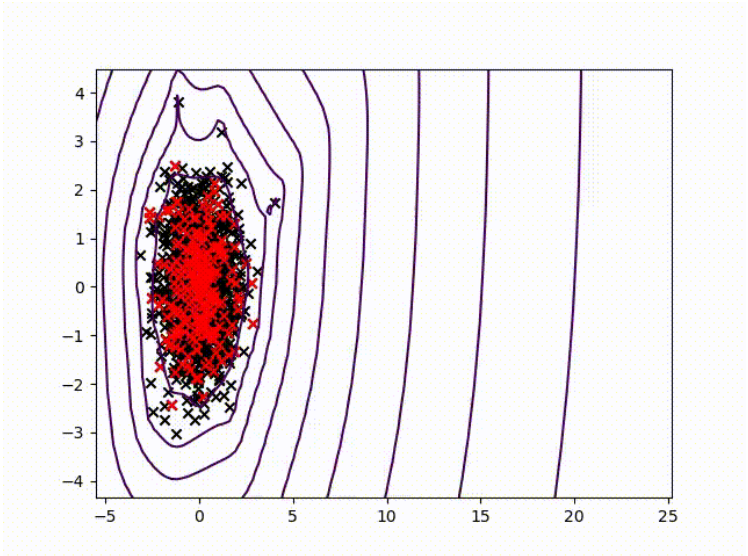
Incremental inversion of a matrix A with the Sherman-Morrison-Woodbury formula:

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

The moment matrix $\mathbf{M}_d(\mu_{n+1})$ can be rewritten with the incremental formula:

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

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



Christoffel function score

DyCF requires only one parameter to be fixed: d

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

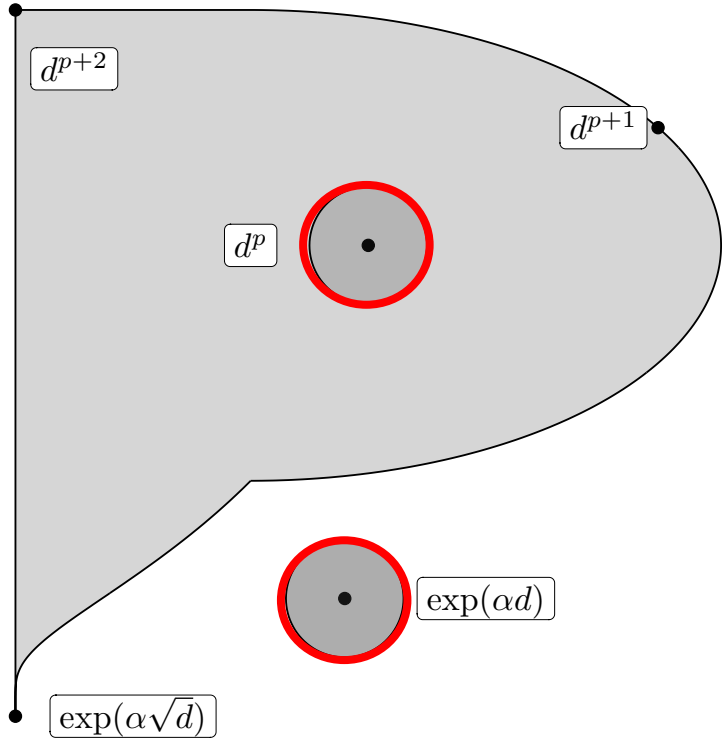
Normalized scoring function : $S_{d,p}(\mathbf{x}) = \frac{\Lambda_d^{\mu}(\mathbf{x})^{-1}}{\gamma_{d,p}}$.

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

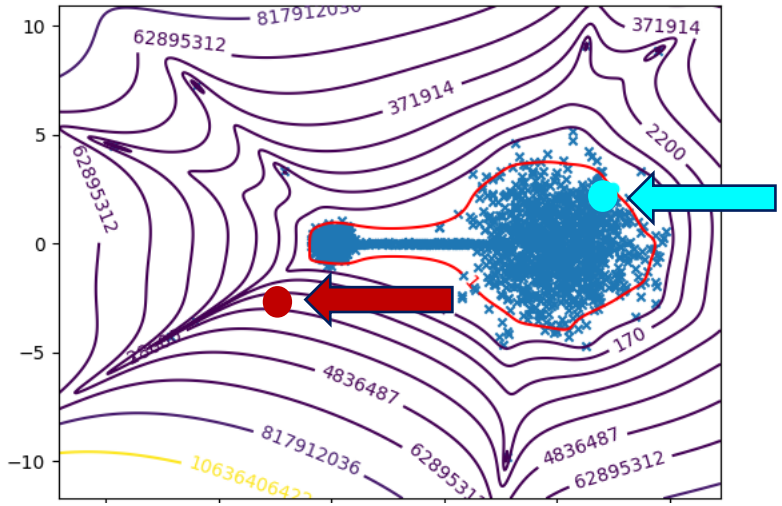
As d grows, $\Lambda_d^{\mu}(\mathbf{x})^{-1}$ has:

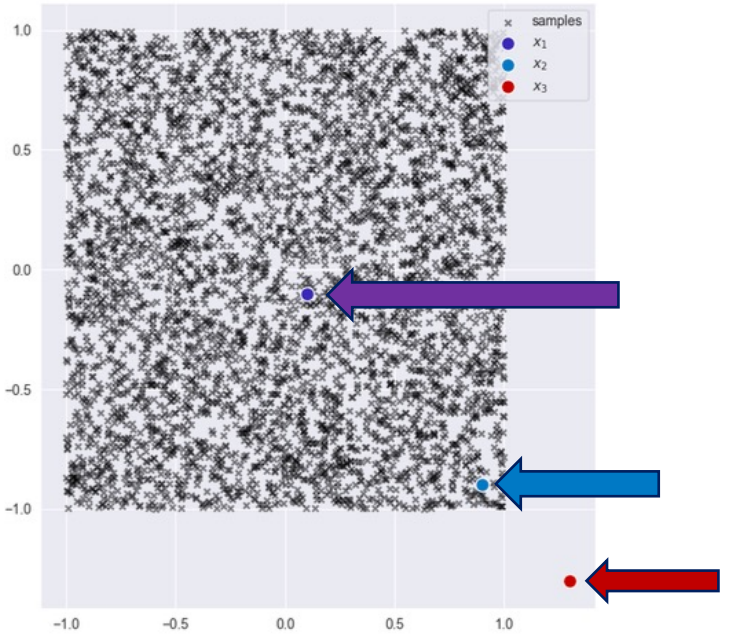
$$\begin{cases} \text{POLYNOMIAL growth} & \text{INSIDE } \Omega \\ \text{EXPONENTIAL growth} & \text{OUTSIDE } \Omega \end{cases}$$

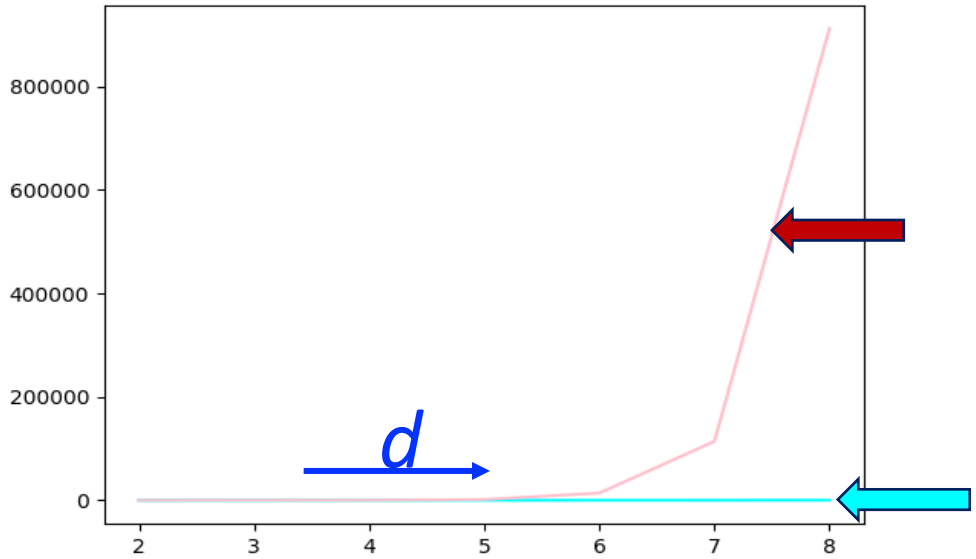


Cf. (Lasserre *et al.* 2022), Lemmas 4.3.1 and 4.3.2

CF growth property

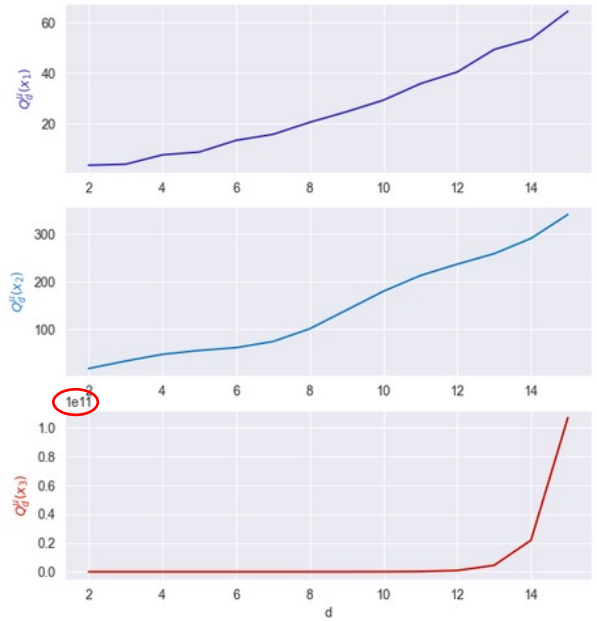


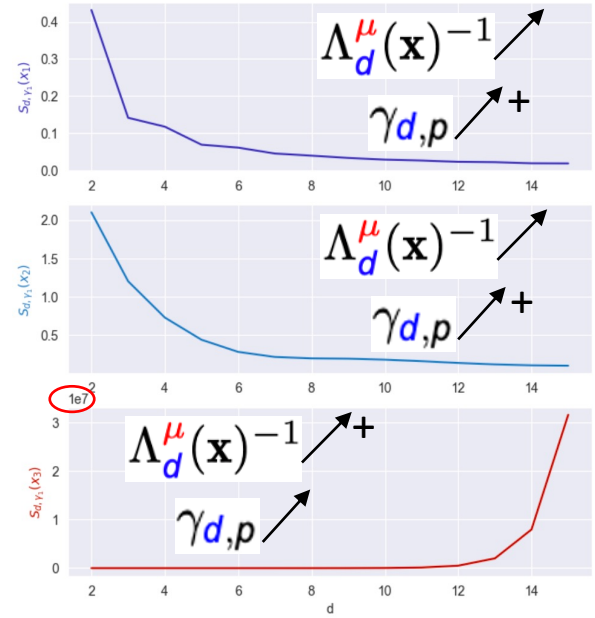




Normalized score:

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





DyCG: Dynamic CF Growth method

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

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.

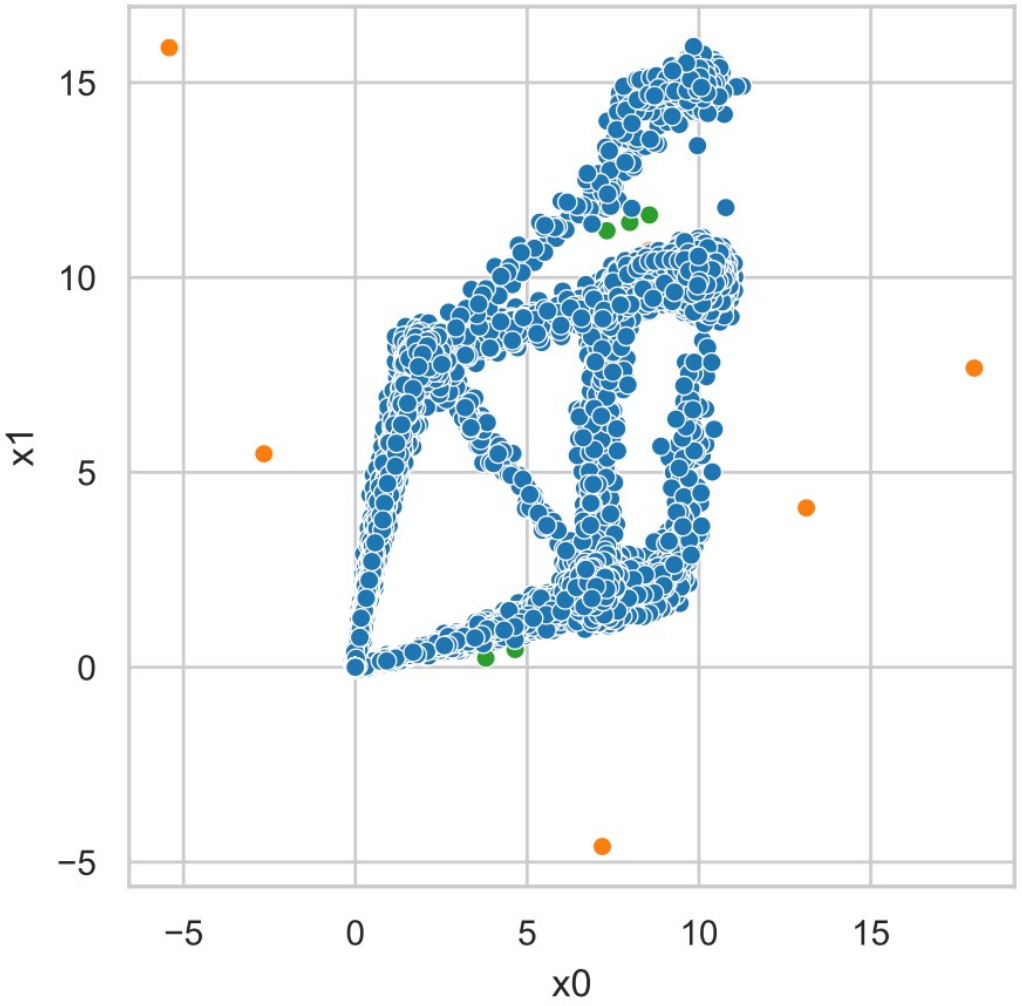
Evaluation

- Labelled synthetic data streams
- Unlabelled real-world data streams

Synthetic data streams

- Multimode distributions
- Transitions according to assigned probabilities

200k point



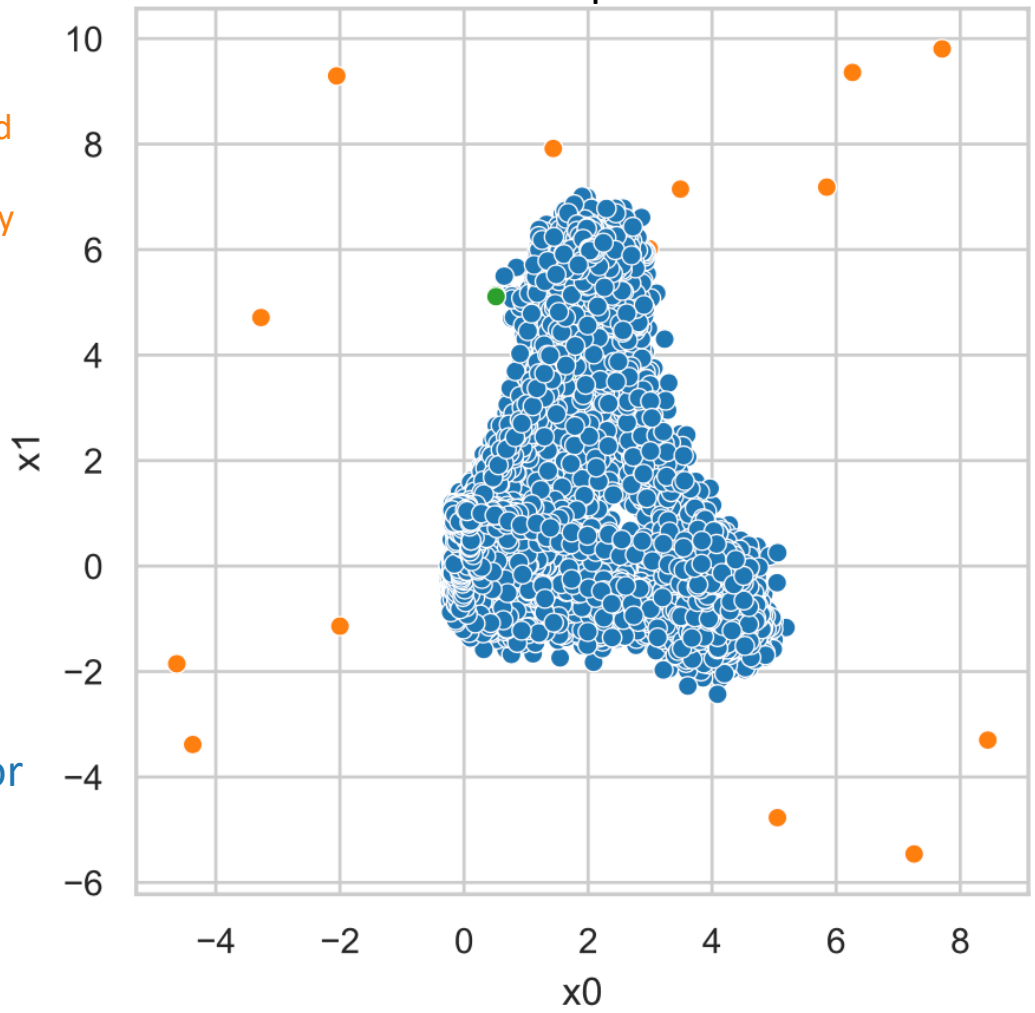
Type I outliers: Uniformly distributed with specified occurence probability

Type II outliers: short off- set from normal behavior with appearing and lasting probabilities

\+

Alteration of normal behavior

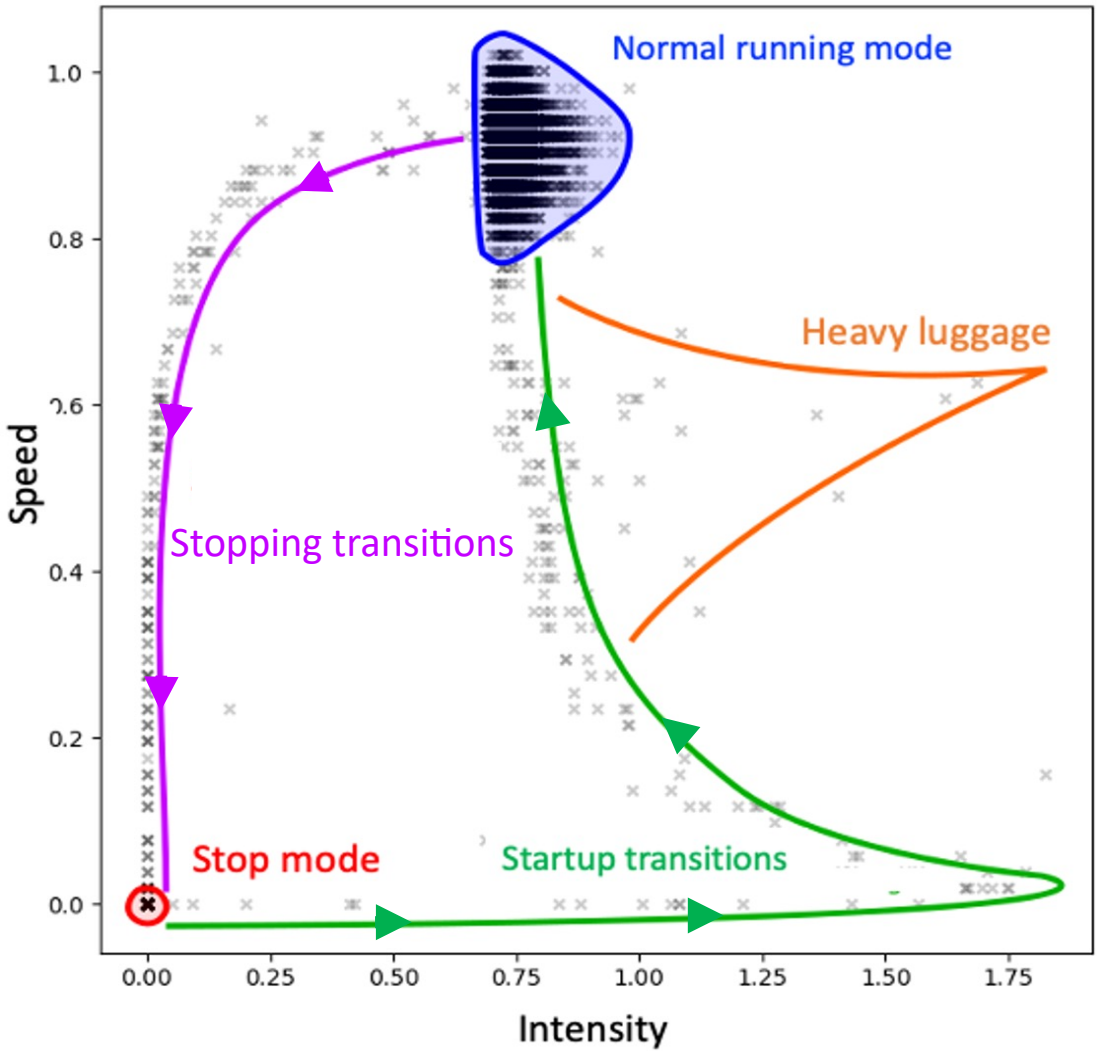
200k point



Industrial luggage conveyor data stream

Carl Berger-Levrault project

Multimode system, 5 conveyors, 2 variables, 7 successive days, measurements every second (86400 samples per day)



Metrics

- Labelled synthetic data streams
  - **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
- Unlabelled real-world data streams
  - **EM** (Excess-Mass curve): the higher, the better
  - **MV** (Mass-Volume curve): the lower, the better
  - EM and MV evaluate the extent to which a scoring function aligns with the statistical distribution of samples
- Computation time for one iteration

Methods

Sliding window Multivariable Kernel Density Estimation

Based on discounting learning of a GMM model

Based on estimating the number of neighbors laying at a given distance (parameter « radius »)

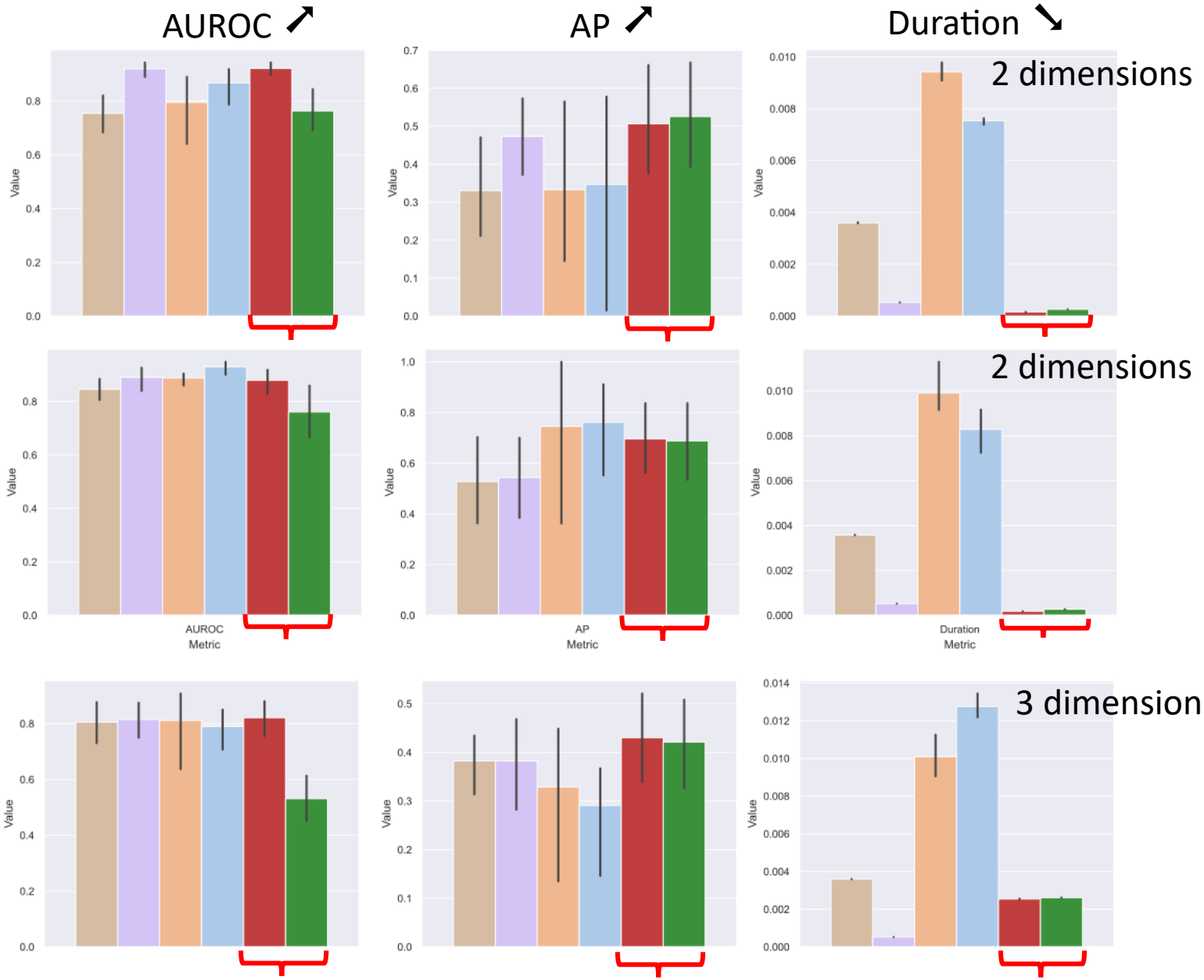
Method that contrasts sample local density with that of its neighbors

Method	Parameters	Values
KDE	Threshold	Meaningless
	Window size	1000
	Kernel	Gaussian
	Bandwidth	Scott's rule
SmartSifter	Threshold	Meaningless
	Nb components	12
	Discounting parameter	1e-3
	Stability parameter	1.5
DBOKDE	Nb neighbors	Meaningless
	Search radius	0.1
	Window size	1000
	Kernel	Epanechnikov
	Bandwidth	Scott's rule
ILOF	Threshold	Meaningless
	Nb neighbors	10
	Window size	1000
	Min children	3
	Max children	12
	Reinsertion strategy	close
	Reinsertion tolerance	4
DyCF	Degree	6
	C (threshold-like)	Meaningless
DyCG	Degrees	(2, 6)

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

Evaluation process

Results for the synthetic data streams



AUROC / AP

Performance of DyCF at least on par (close to the best).

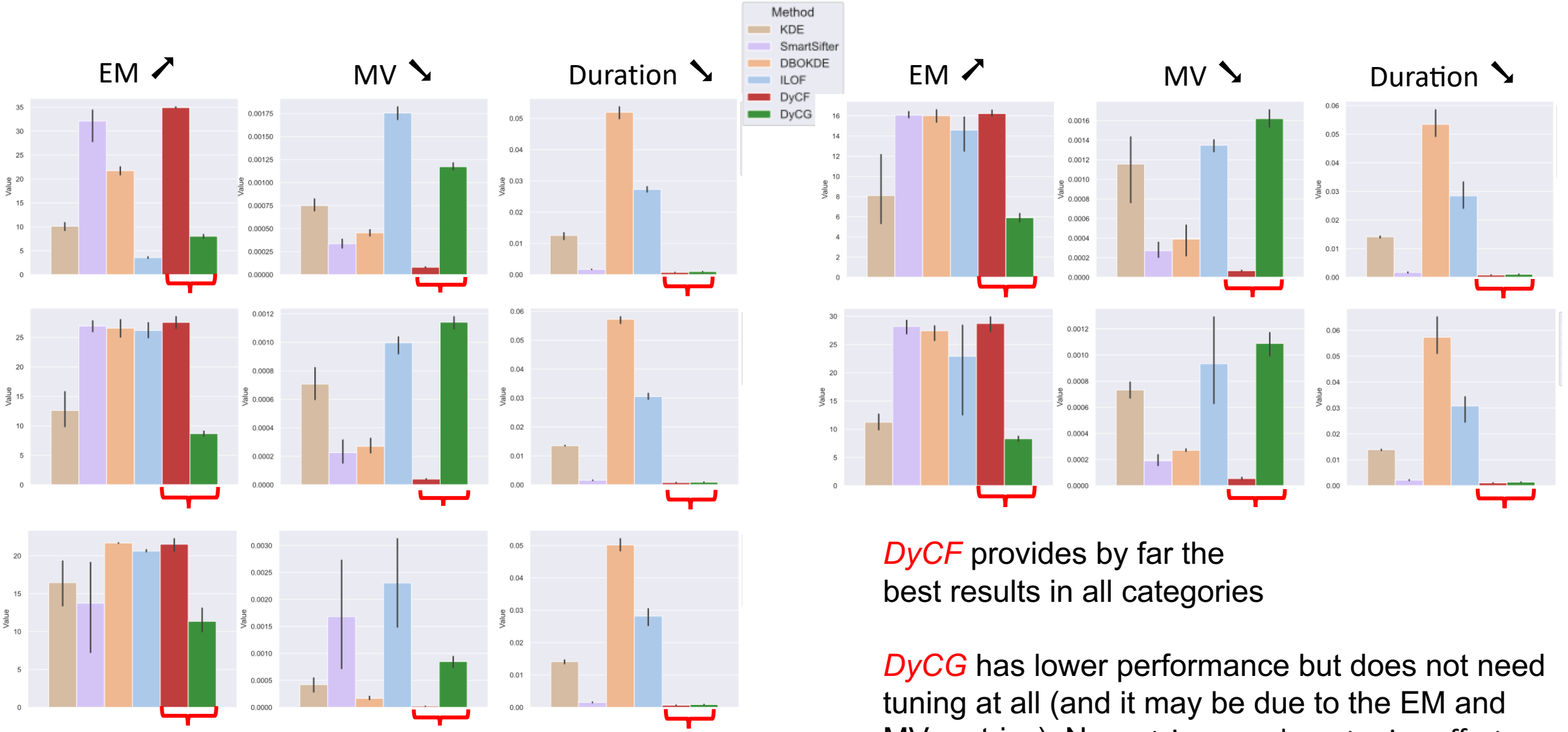
DyCG exhibits slightly lower performance.

Duration

DyCF and DyCG outperform other methods

In 3 dimensions, SmartSifter is slightly better (dependence in *p*)

Results for the 5 industrial conveyors



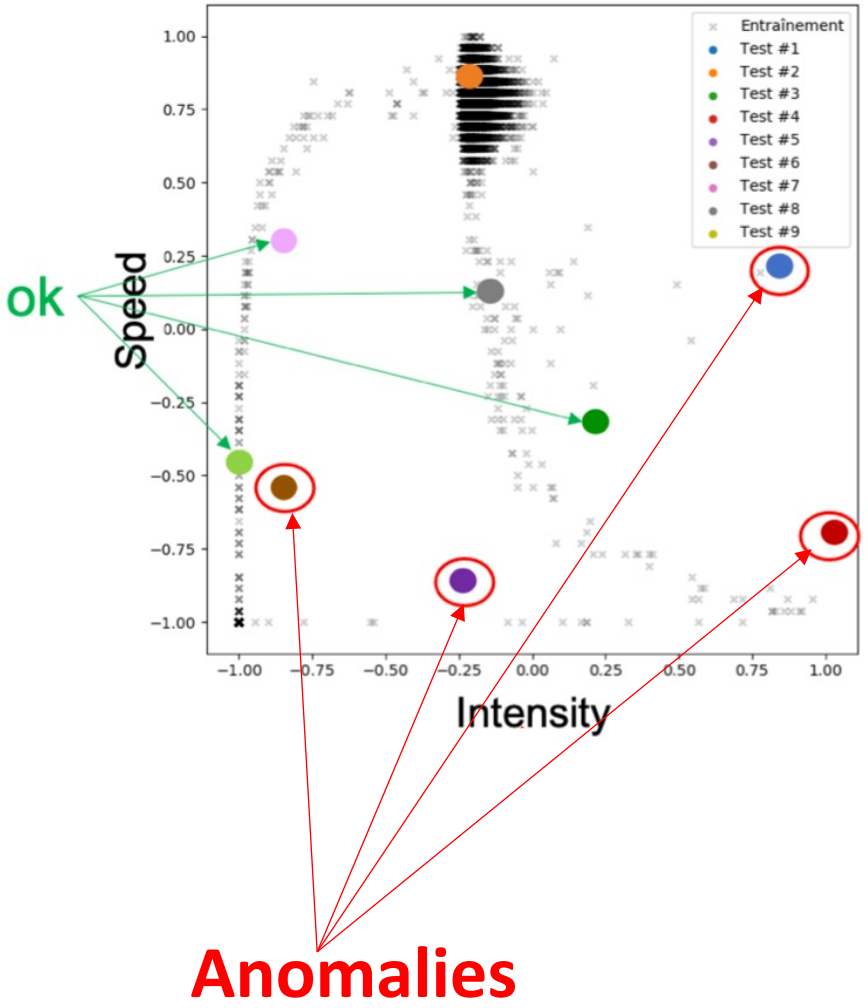
DyCF provides by far the best results in all categories

DyCG has lower performance but does not need tuning at all (and it may be due to the EM and MV metrics). No metric rewards no tuning effort.

Industrial luggage conveyor data stream

Carl Berger-Levrault project

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



Conclusions

- *DyCF* and *DyCG* are simple and easy-to-use methods with very little tuning 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
  - Extending to abnormal trajectory detection

Ducharlet K, Travé-Massuyès L, Lasserre JB, Le Lann MV, Miloudi Y, Leveraging the Christoffel Function for Outlier Detection in Data Streams, submitted to the Int. J. of Data Science and Analytics.

Lasserre JB, Pauwels E, Putinar M (2022) The Christoffel–Darboux Kernel for Data Analysis. Cambridge Monographs on Applied and Computational Mathematics, Cambridge, University Press, https://doi.org/10.1017/9781108937078

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

Louise Travé-Massuyès

Kévin Ducharlet, Jean-Bernard Lasserre

2ème Congrès de la SAGIP
29-31 mai 2024
Villeurbanne