Anomaly Detection based on Markov Data: A Statistical Depth Approach

Carlos Fernández Stephan Clémençon

LTCI, Télécom Paris, Institut Polytechnique de Paris

July 2nd, 2025

Outline

1. Introduction
2. Statistical Depth
3. Markovian depth functions
4. Theoretical Results
5. Applications
6. Conclusion

Background and Notations

- P_X: distribution of r.v. X in $\mathbb{R}^d$
- $\mathcal{P}(\mathbb{R}^d)$: set of all probability distributions on $\mathbb{R}^d$
- $\mathbb{T}$: set of finite length sequences of elements in $\mathbb{R}^d$
- $\mathbf{X} = (X_0, X_1, \ldots)$: Harris recurrent Markov chain in $E \subseteq \mathbb{R}^d$ with initial distribution ν, kernel Π, stationary distribution μ.

Background and Notations

- P_X: distribution of r.v. X in $\mathbb{R}^d$
- $\mathcal{P}(\mathbb{R}^d)$: set of all probability distributions on $\mathbb{R}^d$
- $\mathbb{T}$: set of finite length sequences of elements in $\mathbb{R}^d$
- $\mathbf{X} = (X_0, X_1, \ldots)$: Harris recurrent Markov chain in $E \subseteq \mathbb{R}^d$ with initial distribution ν, kernel Π, stationary distribution μ.

Kernel

The kernel Π rules the one step transitions of the chain, that is

$$\forall x \in \mathbb{R}^d, A \in \mathcal{B}(E), \quad \mathbb{P}(X_{n+1} \in A | X_n = x) = \Pi(x, A) = \Pi_x(A),$$

i.e.

$$P_{(X_{n+1}|X_n)} = \Pi_{X_n}.$$

Depth Functions on $\mathbb{R}^d$

Definition

A depth function in $\mathbb{R}^d$ with respect to a probability distribution $P \in \mathcal{P}(\mathbb{R}^d)$ is a function $D_P : \mathbb{R}^d \to [0, 1]$ that satisfies (most of) the following properties:

- P1 (Affine invariance) $D_{P_{AX+b}}(Ax + b) = D_P(x)$ for all $x \in \mathbb{R}^d$
- P2 (Vanishing at infinity) $D_P(x) \to 0$ as $||x||$ tends to infinity
- P3 (Maximality at center) D_P is maximized at the center of symmetry
- P4 (Monotonicity) D_P decreases along rays from the deepest point

Utility of Depth Functions

Main Applications

- **Generalizations** - Ex: multivariate medians
- **Robust statistics** - Less sensitive to outliers
- **Outlier detection** - Points with low depth are potential outliers
- **Data ordering** - Center-outward ranking
- **Data visualization** - Through depth contours

Examples of Depth functions

Halfspace depth

For any $x \in \mathbb{R}^d$ denote by $\mathcal{H}_x$ the collection of all halfspaces that contain x, then, the Halfspace depth is defined as

$$D_h(x, P) = \inf_{H \in \mathcal{H}_x} P(H)$$

Other examples include Simplicial depth, Mahalanobis depth, Leans depth, IRW depth ...

Motivation

Main Objectives

Develop computationally feasible tools for assessing the *centrality* (or *outlyingness*) of finite length trajectories w.r.t. the distribution of a Markov chain $\mathbf{X}$.

Motivation

Main Objectives

Develop computationally feasible tools for assessing the *centrality* (or *outlyingness*) of finite length trajectories w.r.t. the distribution of a Markov chain $\mathbf{X}$.

Main Challenges

- The set of finite length trajectories ($\mathbb{T}$) is infinite dimensional
- Its topology ($\mathcal{T}$) is not metrizable
- Even with fixed-length trajectories, standard methods suffer from:
  - Failure to capture the Markovian structure
  - Curse of dimensionality

Motivation - A Concrete Example

Reflected Random Walk Model

$X_0 = x_0 \geq 0$ and
$X_{n+1} = \max(0, X_n + W_n)$ for
$n \in \mathbb{N}$ where:

$$W_n = -1.1 \times Y_n + 1$$

The Y_n's are i.i.d. exponential r.v.'s with mean 1.

Notice that

If $X_{n+1} > X_n$, then $X_{n+1} - X_n \leq 1$.

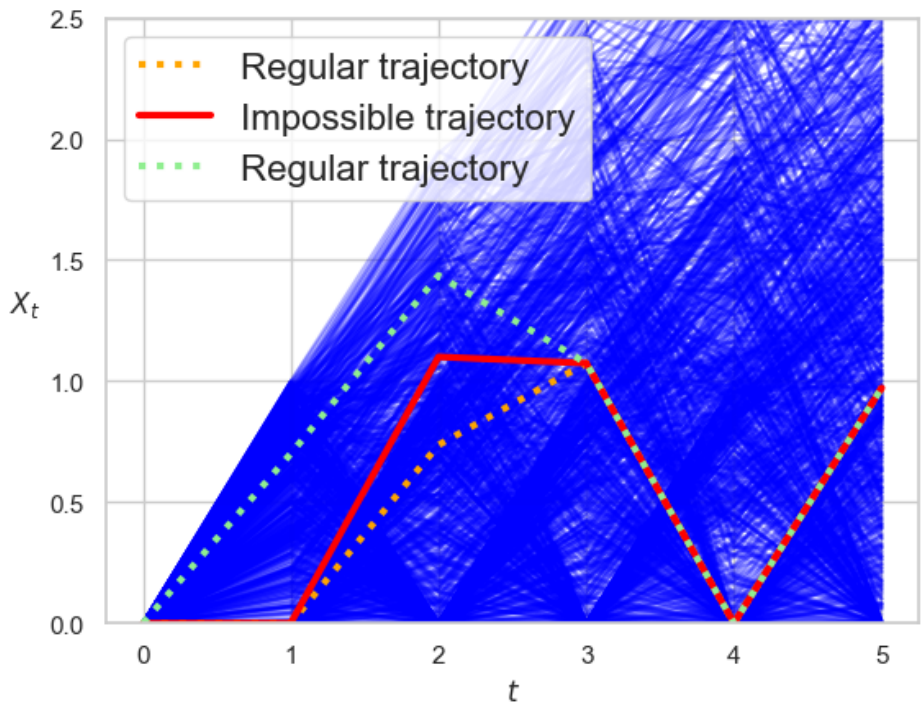


Figure: Unfeasible trajectory (red) vs. regular ones (green/orange)

Limitations of Standard Approaches

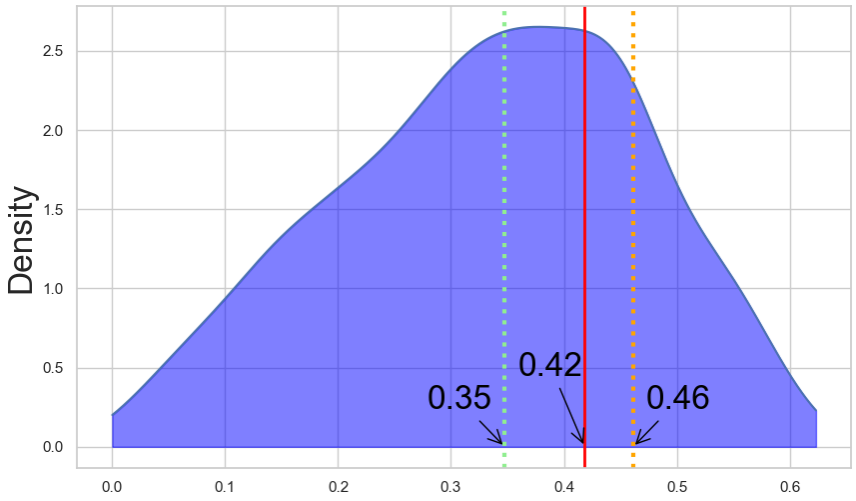


Figure: Lens depth density

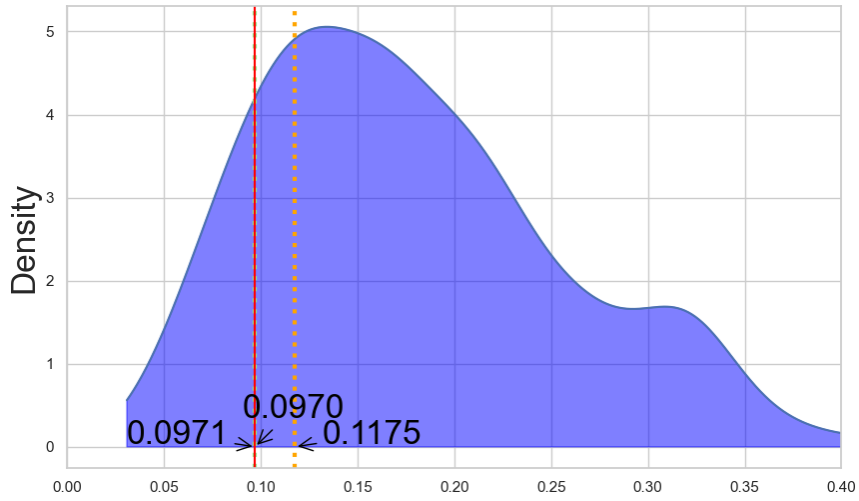


Figure: Mahalanobis depth density

Key Limitations

- They fail to account for the sequential structure and temporal dynamics of Markov data.
- The computational complexity of the best methods generally increases exponentially with n.

Markovian Depth Functions

Definition

A statistical depth function w.r.t. the distribution of a Markov chain $\mathbf{X}$ is a function $\mathbf{D}_{\mathbf{X}} : \mathbb{T} \to [0, 1]$ that may satisfy:

M0 (Initial Law Independence) If $\mathbf{X}$ and $\mathbf{X}'$ have the same transition kernel:

$$\mathbf{D}_{\mathbf{X}}(\mathbf{x}) = \mathbf{D}_{\mathbf{X}'}(\mathbf{x})$$

M1 (Affine Invariance) For any non-singular A and $b \in \mathbb{R}^d$:

$$\mathbf{D}_{A\mathbf{X}+b}(A\mathbf{x} + b) = \mathbf{D}_{\mathbf{X}}(\mathbf{x})$$

M2 (Vanishing at Infinity) $\mathbf{D}_{\mathbf{X}}(\mathbf{x}) \to 0$ as $\mathbf{x}$ "tends to infinity" in $\mathbb{T}$'s topology

Markovian Sample Path Depth

Key Idea

Given a depth function D on $\mathbb{R}^d$, we define the Markovian sample path depth as:

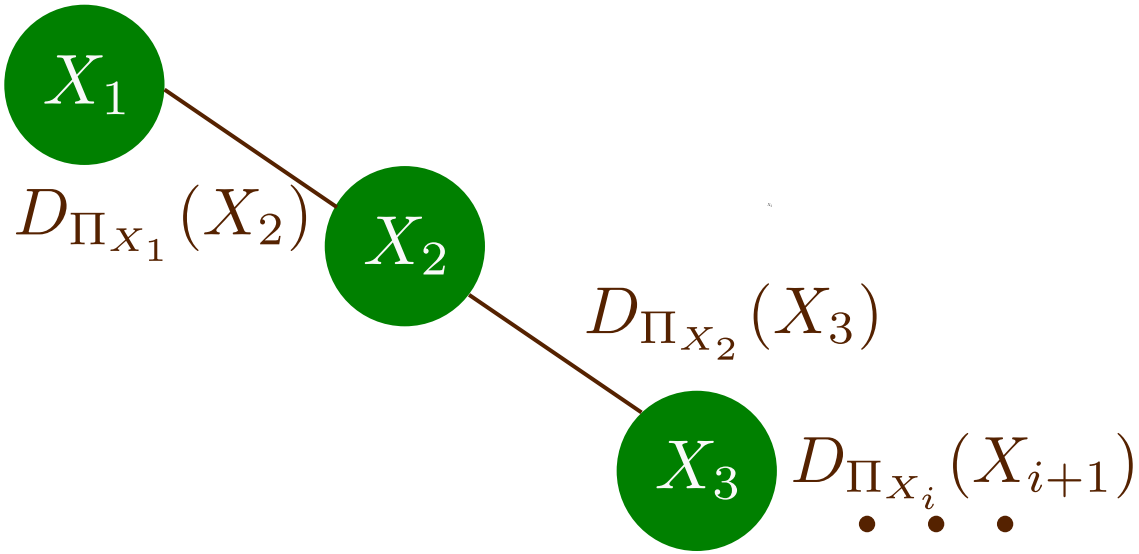
$$\forall \mathbf{x} = (x_0, \ldots, x_n) \in \mathbb{T},\ D_{\Pi}(\mathbf{x}) = \sqrt[n]{\prod_{i=1}^{n} D_{\Pi_{x_{i-1}}}(x_i)}$$

Markovian Sample Path Depth

Key Idea

Given a depth function D on $\mathbb{R}^d$, we define the Markovian sample path depth as:

$$\forall \mathbf{x} = (x_0, \ldots, x_n) \in \mathbb{T},\ D_\Pi(\mathbf{x}) = \sqrt[n]{\prod_{i=1}^{n} D_{\Pi_{x_{i-1}}}(x_i)}$$



Markovian Sample Path Depth

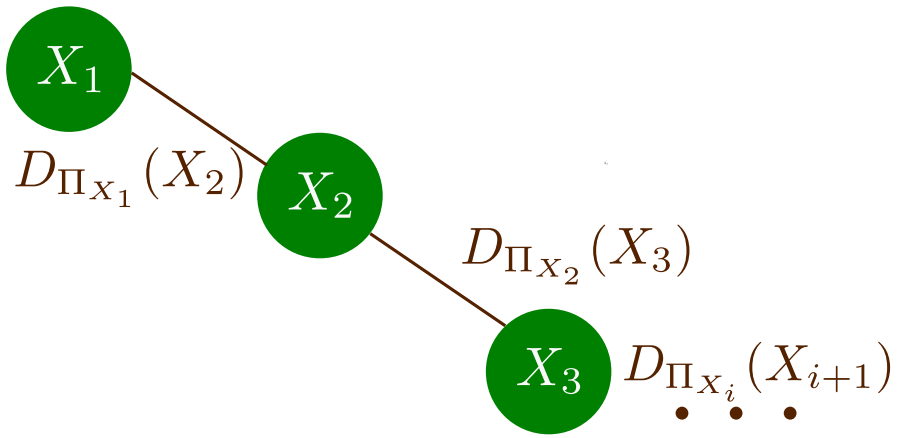
Key Idea

Given a depth function D on $\mathbb{R}^d$, we define the Markovian sample path depth as:

$$\forall \mathbf{x} = (x_0, \ldots, x_n) \in \mathbb{T},\ D_\Pi(\mathbf{x}) = \sqrt[n]{\prod_{i=1}^{n} D_{\Pi_{x_{i-1}}}(x_i)}$$

Interpretation

- Geometric mean of depths of one-step transitions
- Accounts for Markov dependency structure



Key Properties of Markovian Sample Path Depth

Main properties

- (M0) Independence from initial law.
- (M1) Affine invariance.
- (M2) Vanishing at infinity.
- Continuity as a function of $\mathbf{x}$.
- Continuity as a function of Π

Asymptotic Results

Theorem (Consistency)

If $\log(D_{\Pi_x}(y))$ *is integrable w.r.t.* $\mu d(x)\Pi_x(dy)$ *in* E^2*, then, for any initial distribution* ν*,*

$$D_{\Pi}(X_0,\, X_1,\, \ldots,\, X_n) \to D_{\infty}(\Pi)$$

$\mathbb{P}_\nu$*-almost surely as* $n \to \infty$*, where*

$$D_{\infty}(\Pi) = \exp\left(\int_{E^2} \log\left(D_{\Pi_x}(y)\right) \Pi_x(dy)\mu(dx)\right).$$

Asymptotic Results

Theorem (Consistency)

If $\log(D_{\Pi_x}(y))$ *is integrable w.r.t.* $\mu d(x)\Pi_x(dy)$ *in* E^2*, then, for any initial distribution* ν*,*

$$D_{\Pi}(X_0,\ X_1,\ \ldots,\ X_n) \to D_{\infty}(\Pi)$$

$\mathbb{P}_\nu$*-almost surely as* $n \to \infty$*, where*

$$D_{\infty}(\Pi) = \exp\left(\int_{E^2} \log\left(D_{\Pi_x}(y)\right) \Pi_x(dy)\mu(dx)\right).$$

Theorem (Asymptotic Normality)

If, in addition, $\log(D_{\Pi_x}(y))$ *is square integrable w.r.t.* $\mu d(x)\Pi_x(dy)$*, then* $D_{\Pi}(X_0, X_1, \ldots, X_n)$ *is asymptotically normal.*

Non-Asymptotic Bounds

Theorem (Finite sample inequalities)

Let $n \geq 1$ and $\mathbf{x} \in E^{n+1}$. Consider two transition probabilities Π and $\hat{\Pi}$ on E. Suppose:

1. $\exists \epsilon > 0$ *s.t.* $\min\{D_{\hat{\Pi}_i}(x_{i+1}), D_{\Pi_i}(x_{i+1})\} > \epsilon$ *for* $i < n$
2. $\exists \mathcal{A} \subset \mathcal{B}(\mathbb{R}^d)$ *and a finite constant* C_d *such that:*

$$\sup_{x \in \mathbb{R}^d} |D_P(x) - D_Q(x)| \leq C_d ||P - Q||_{\mathcal{A}}$$

Then,

$$|D_\Pi(\mathbf{x}) - D_{\hat{\Pi}}(\mathbf{x})| \leq \frac{7}{4} C_d \frac{D_\Pi(\mathbf{x})}{\epsilon} \max_{i<n} ||\Pi_{x_i} - \hat{\Pi}_{x_i}||_{\mathcal{A}}$$

Computational Aspects

Algorithm: Estimation of $D_{\hat{\Pi}}(\mathbf{x})$

Input:

- Path $\mathbf{x} = (x_0, \ldots, x_n)$
- Transition probability $\hat{\Pi}$ (an estimator of the true Π)
- Precision control integer $M \geq 1$

Steps:

For $i = 0$ to $n - 1$:

1. Generate M samples $x_{1,i}, \ldots, x_{M,i}$ from $\hat{\Pi}_{x_i}$
2. Compute $\widehat{D}_i$ of $D_{\hat{\Pi}_{x_i}}(x_{i+1})$

Output: $D_{\hat{\Pi}}(\mathbf{x}) = \sqrt[n]{\prod_{i=1}^{n} \widehat{D}_i}$

Computational Aspects

Main advantages

- We can use efficient algorithms to obtain $\hat{\Pi}$. Ex. Nadaraya-Watson.
- The loop can be executed in parallel.
- The complexity is linear on n.

Applications: Anomaly detection

GI/G/1 Queuing System

- Consider a GI/G/1 queuing system with:
  - Interarrival times $\{T_n\}_{n\geq 0}$ i.i.d. exponential with mean 0.5
  - Service times $\{V_n\}_{n\geq 0}$ i.i.d. exponential with mean 0.45
  - $\{T_n\}$ and $\{V_n\}$ are independent sequences
- Waiting time process: $X_{n+1} = \max(0, X_n + W_n)$ where $W_n = V_n - T_n$
- This corresponds to the reflected random walk model introduced earlier

Applications: Experimental Setup

Dataset construction

- **Training data**:
  - One long trajectory ($n = 1000$) from normal system behavior
  - Used to learn the transition kernel of the Markov chain
- **Testing data**:
  - Four contaminated datasets, each focused on one anomaly type
  - Each dataset contains 200 paths of random length (50-200)
  - 50% of paths contain the specific anomaly

Anomaly detection

1. Obtain an estimator $\hat{\Pi}$ of the kernel Π using the long trajectory.
2. Apply the algorithm to estimate D_{Π} using the halfspace depth.
3. Use this estimation as the scoring function.

Anomaly 1: Shock Anomaly

Service Time Shock

- Generated by increasing the service time distribution mean
- Normal system: $V_n \sim$ Exponential(mean $= 0.45$)
- Anomaly segment: $V_n \sim$ Exponential(mean $= 2.25$)
- The anomaly starts at a random time and affects 10% of the trajectory.
- Results in sudden spikes in waiting times

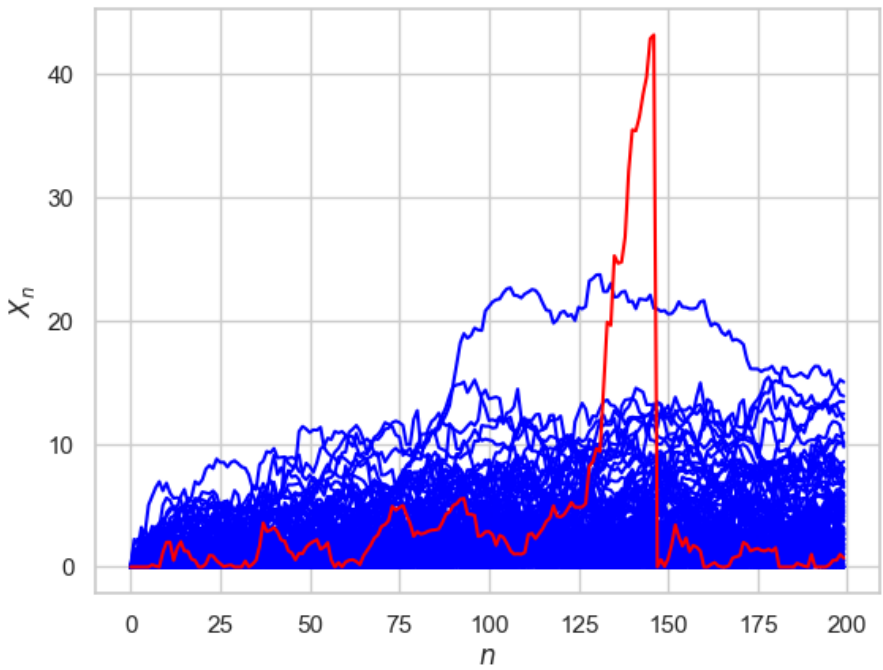


Figure: Service time shock

Anomaly 2: Dynamic anomaly

Faster Customer Arrivals

- Changes the interarrival time distribution
- Normal:
  $T_n \sim \text{Exp}(\text{mean} = 0.5)$
- Anomaly:
  $T_n \sim \text{Exp}(\text{mean} = 0.1)$
- The anomaly starts at a random time and affects 20% of the trajectory.
- Simulates period of increased customer arrivals

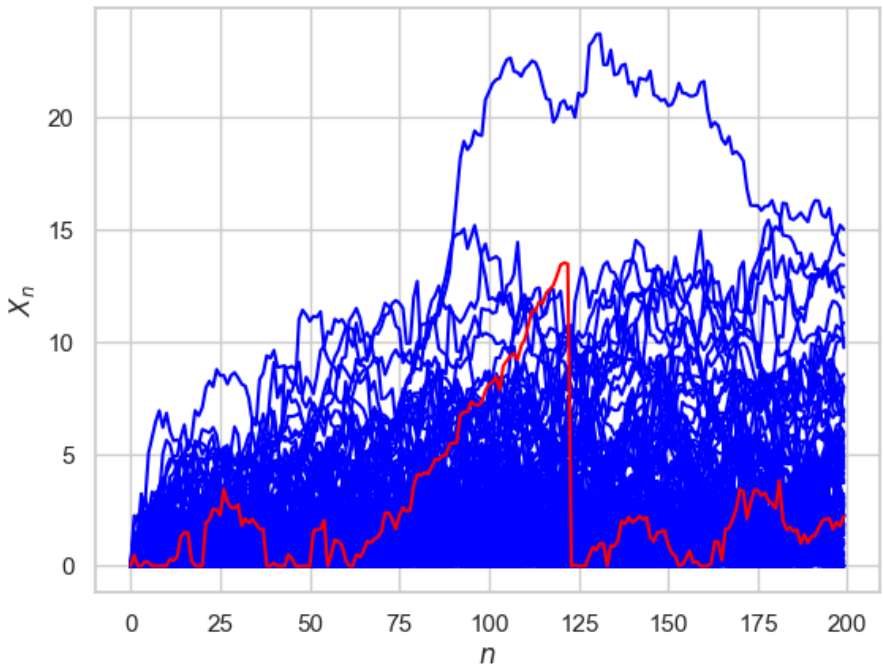


Figure: Faster arrivals

Anomaly 3: Dynamic anomaly

Slower service times

- Modifies service time distribution
- Normal: $V_n \sim \text{Exp}(\text{mean} = 0.45)$
- Anomaly: $V_n \sim 0.55 \cdot \mathcal{U}(0, 2)$
- The anomaly starts at a random time and affects 30% of the trajectory
- Represents period of degraded service

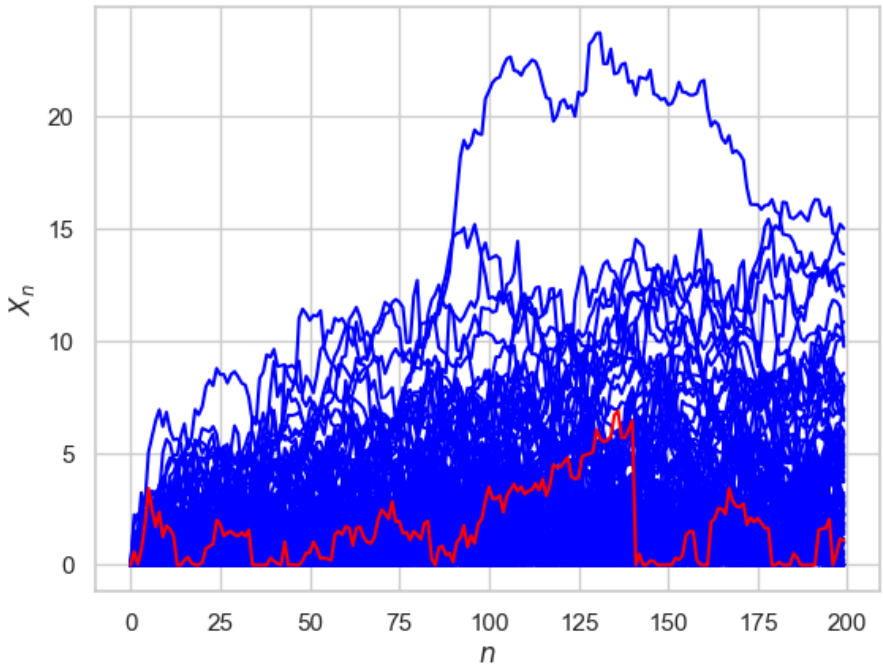


Figure: Slower service

Anomaly 4: Shift anomaly

Deterministic arrivals

- Changes arrival process from stochastic to deterministic
- Normal: $T_n \sim \text{Exponential}(\text{mean} = 0.5)$
- Anomaly: $T_n = 2^{-n}$ (deterministic function)
- The anomaly starts at a random time and affects 25 consecutive steps.
- Represents systematic rather than random arrivals

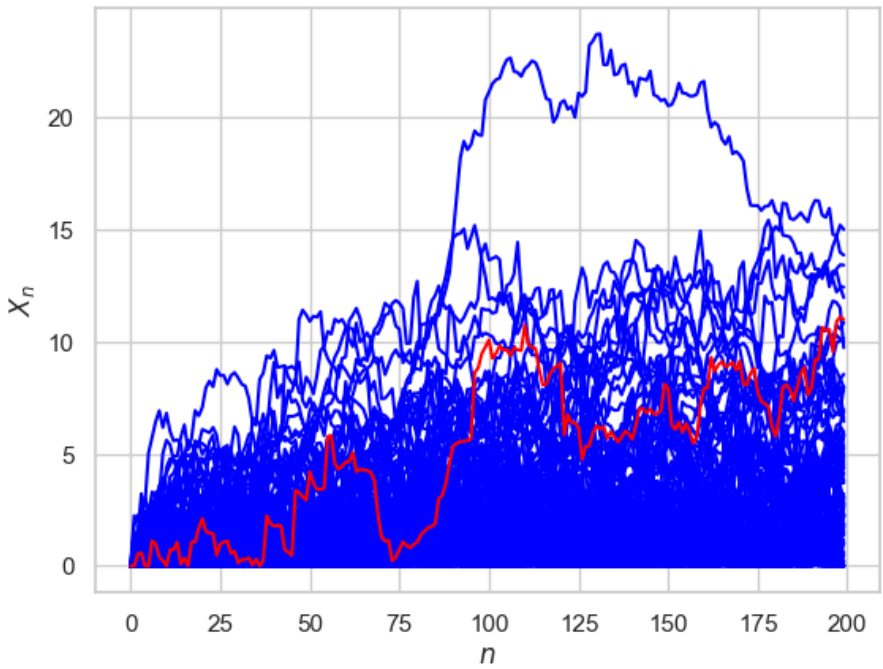


Figure: Deterministic arrivals

Detection Performance - ROC Curves

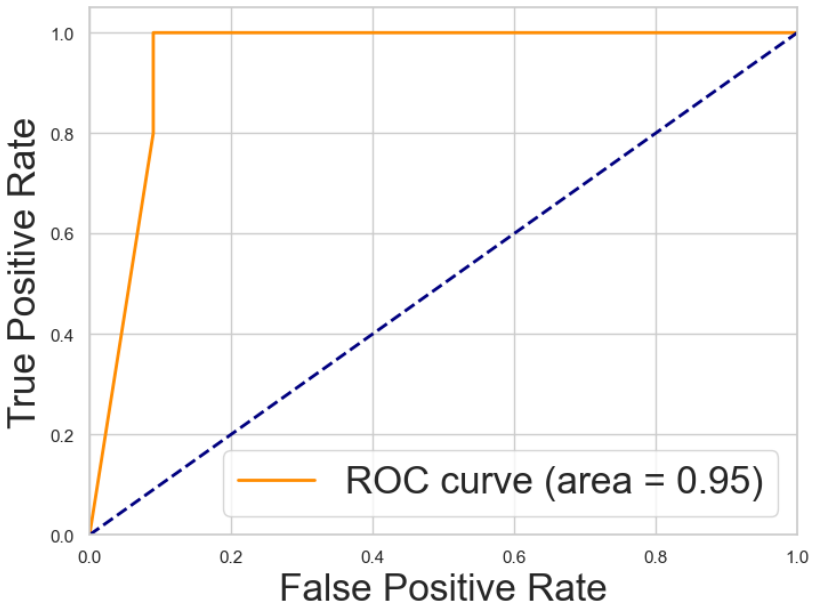


Figure: Service time shock

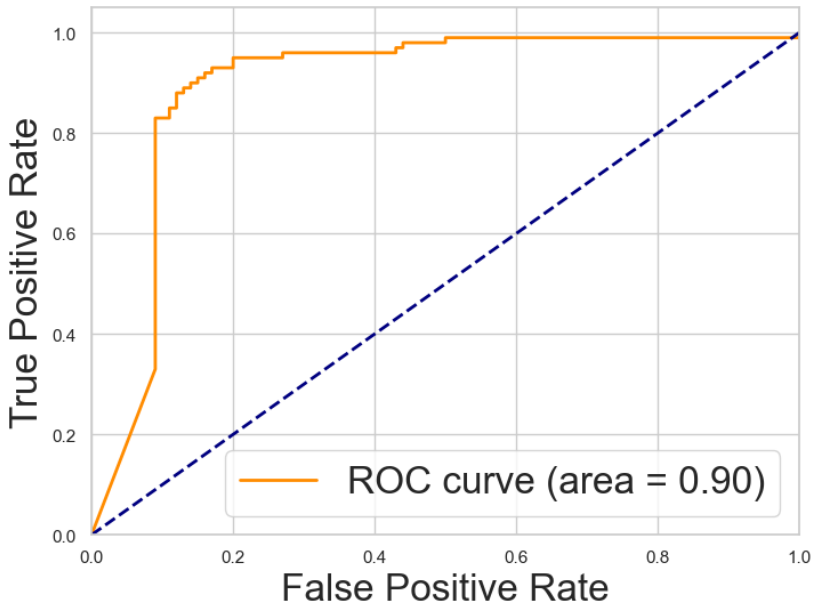


Figure: Faster arrivals

Detection Performance (cont.)

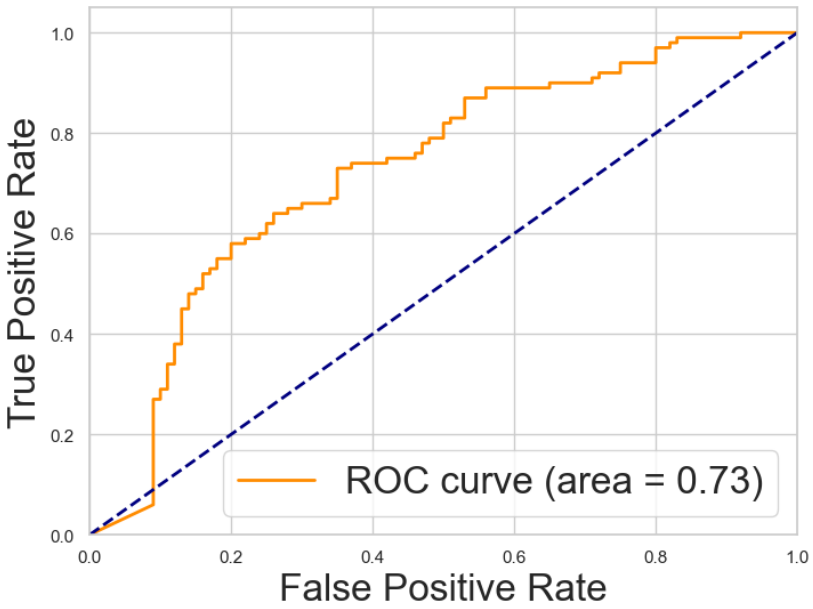


Figure: Slower service times

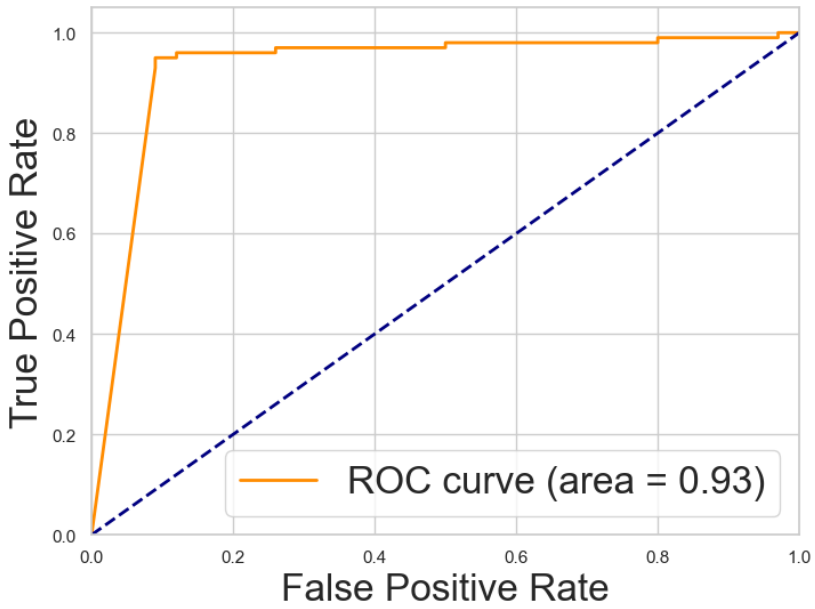


Figure: Deterministic arrivals

Method comparison

Comparative Analysis Setup

- We compared our Markovian depth (D_Π) against standard methods:
  - Isolation Forest (IF)
  - Local Outlier Factor (LOF)
  - Mahalanobis Depth (MD)
- For fair comparison, all trajectories must have equal length
- Created 4 datasets (one per anomaly type):
  - 100 trajectories of fixed length (200 points each)
  - 5% contamination rate
- Applied all four methods to each dataset

Method Comparison

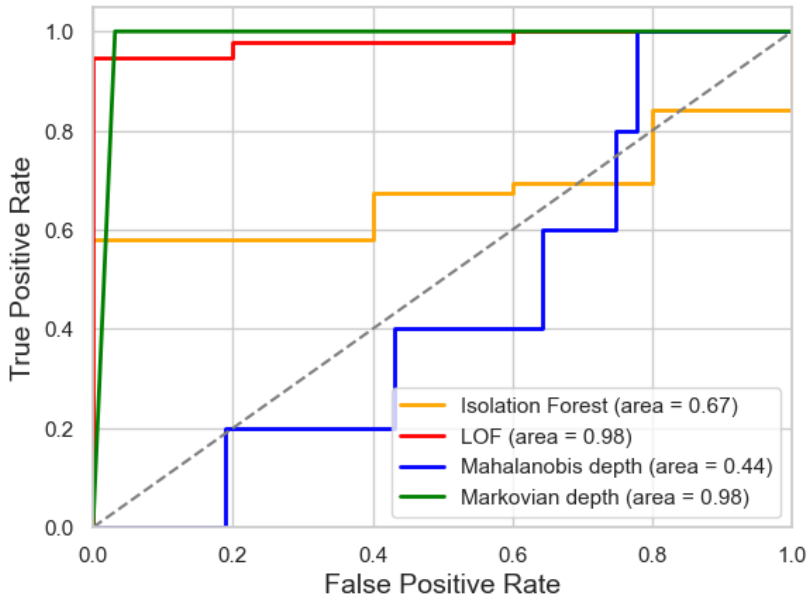


Figure: Service time shock

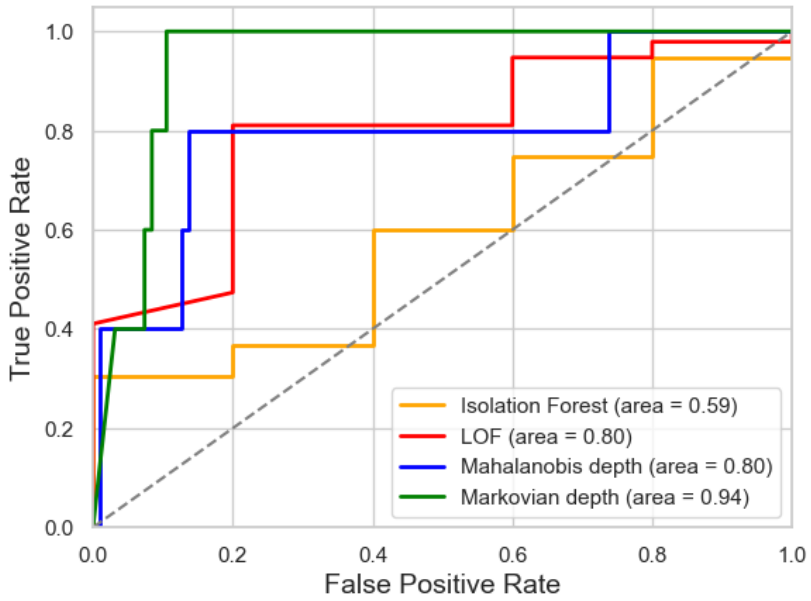


Figure: Faster arrivals

Method Comparison

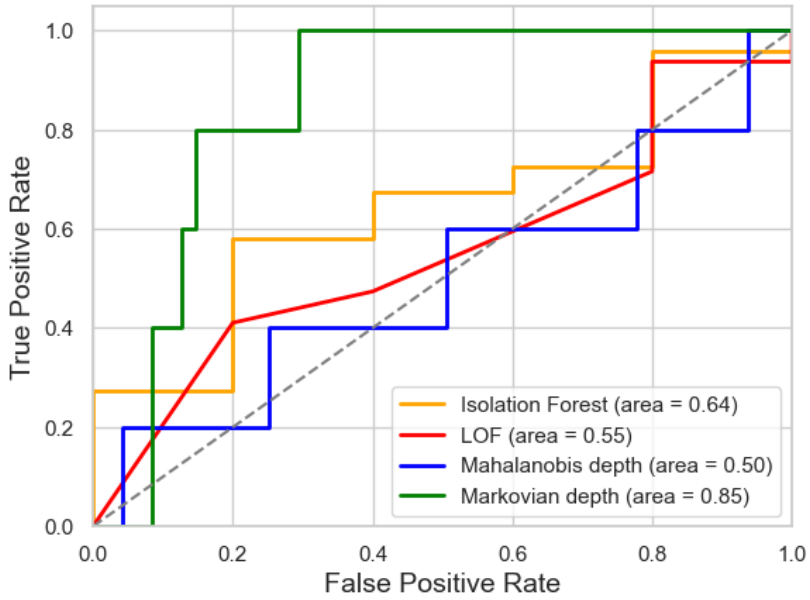


Figure: Slower service times

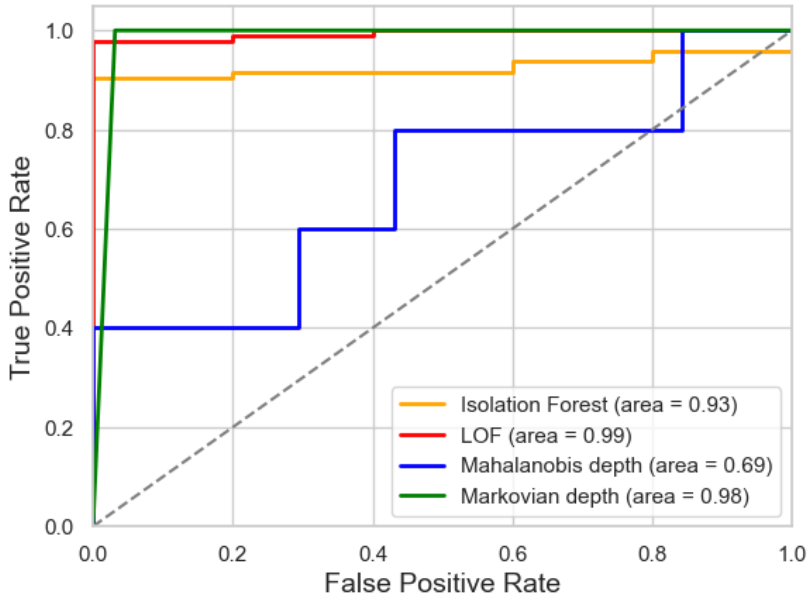


Figure: Deterministic arrivals

Method Comparison

Table: Comparison of the AUC for different classifiers.

	ARCH(1) model				Queuing model			
Anomaly type	IF	LOF	MD	D_Π	IF	LOF	MD	D_Π
Shock	0.75	0.82	0.68	0.85	0.67	0.98	0.44	0.98
Dynamic anomaly I	0.42	0.63	0.52	0.84	0.59	0.80	0.80	0.94
Dynamic anomaly II	0.68	0.90	0.91	0.99	0.64	0.55	0.50	0.85
Shift	1	1	0.6	1	0.93	0.99	0.69	0.98

Application: Clustering

DD-plots

For two sets of Markov trajectories $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}\}$ and $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}\}$ with corresponding kernel estimators $\hat{\Phi}$ and $\hat{\Psi}$, the Markovian DD-plot is obtained by plotting in the Euclidean plane the points $\left\{\left(D_{\hat{\Phi}}(\mathbf{x}), D_{\hat{\Psi}}(\mathbf{x})\right) : \mathbf{x} \in \mathcal{X} \cup \mathcal{Y}\right\}$.

Applications: Clustering

Data generation

5 data sets, each one containing 50 trajectories of random lengths (between 50 and 200 steps). These datasets, labeled $\mathcal{X}, \mathcal{Y}_a, \mathcal{Y}_b, \mathcal{Y}_c, \mathcal{Y}_d$, are constructed following an ARCH(1) model: $X_{n+1} = m(X_n) + \sigma(X_n)\epsilon_n$.

Table: Parameters of the ARCH(1) model used

Dataset	$m(x)$	$\sigma(x)$
$\mathcal{X}$	$(1+\exp(-x))^{-1}$	$\psi(x+1.2)+1.5\psi(x-1.2)$
$\mathcal{Y}_a$	$(1+\exp(-x))^{-1}$	$\psi(x+1.2)+1.5\psi(x-1.2)$
$\mathcal{Y}_b$	$(2+\exp(-x))^{-1}$	$\psi(x+1.2)+1.5\psi(x-1.2)$
$\mathcal{Y}_c$	$(4+\exp(-x))^{-1}$	

Applications: Clustering

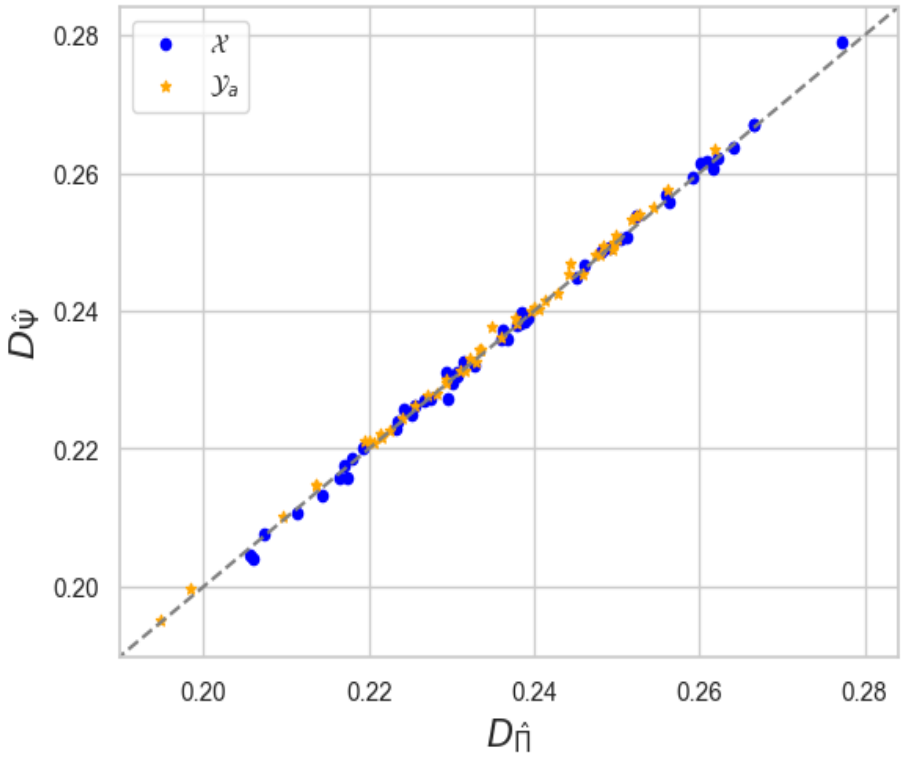


Figure: Markovian DD-plot for trajectories $\mathcal{X}$ and $\mathcal{Y}_a$ (same kernel).

Applications: Clustering

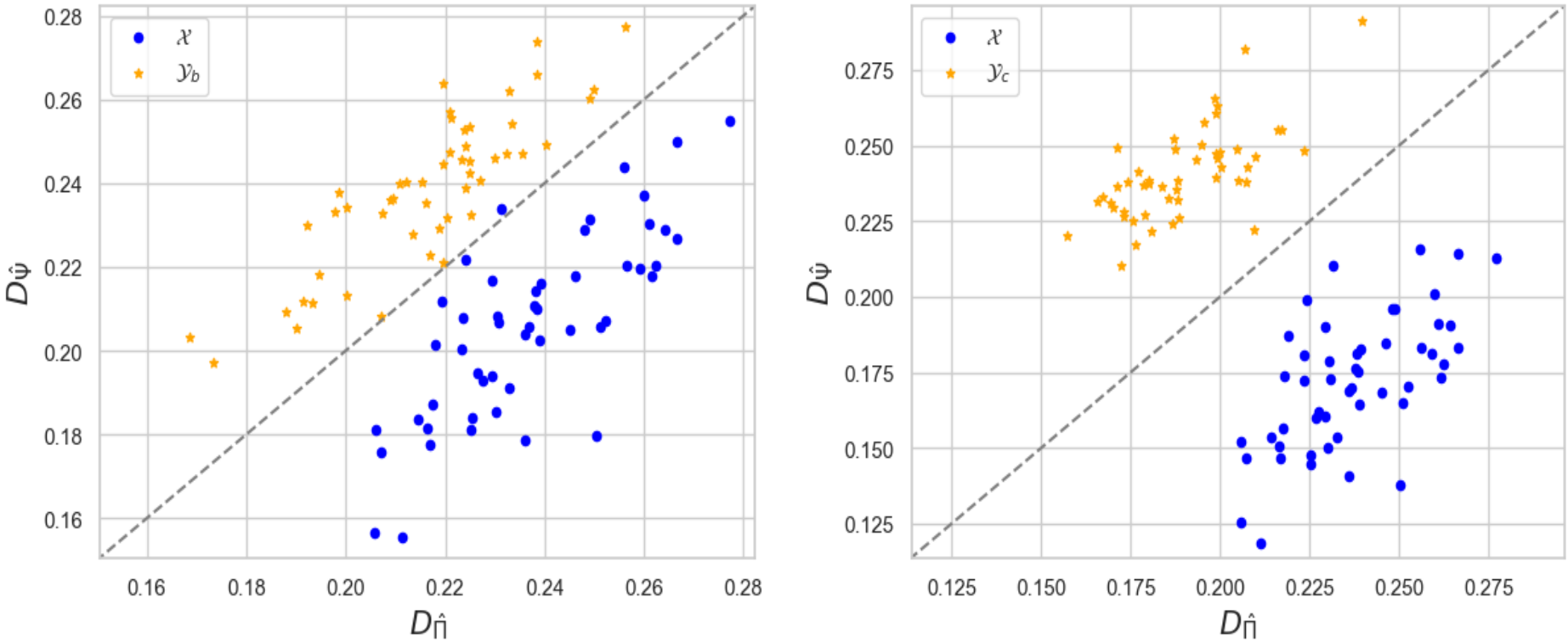


Figure: Markovian DD-plot for trajectories $\mathcal{X}$, and $\mathcal{Y}_b$, and $\mathcal{Y}_c$.

Conclusion

Summary

- Novel framework for depth-based anomaly detection in Markov data
- Overcomes limitations of traditional methods for sequential data
- Strong theoretical guarantees with practical implementation
- Linear scaling with trajectory length
- Superior performance across different types of anomalies

Thank you for your attention!

Code:
github.com/carlosds731/depth_markov

Paper:
arxiv.org/abs/2406.16759