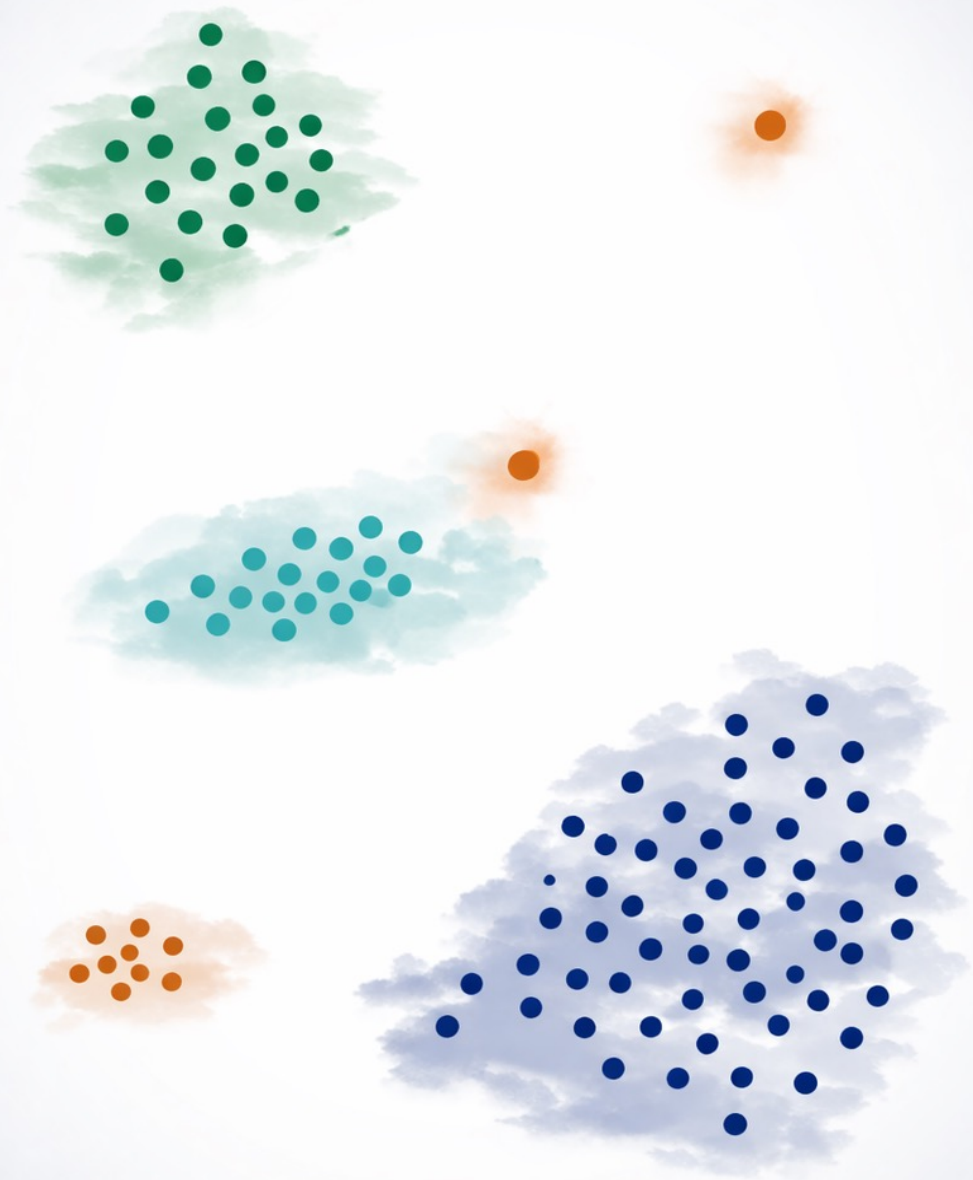
Georgia Tech

Anomaly Detection



Mohsen Moghaddam, Ph.D.

Gary C. Butler Family Associate Professor
H. Milton Stewart School of Industrial and Systems Engineering
George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

SAIL
Symbiotic & Augmented Intelligence Laboratory

Spring 2026

Learning Outcomes

- Distinguish between outlier detection, novelty detection, and out-of-distribution detection
- Explain and apply density-based anomaly detection methods, including Gaussian and robust covariance models
- Formulate and interpret geometric anomaly detectors, particularly one-class support vector machine (SVM)
- Connect anomaly detection to hypothesis testing, including false alarms and detection power
- Apply sequential change-point detection methods such as the cumulative sum over log-likelihood ratios (CUSUM)
- Compare anomaly detection algorithms and select methods based on data characteristics and assumptions

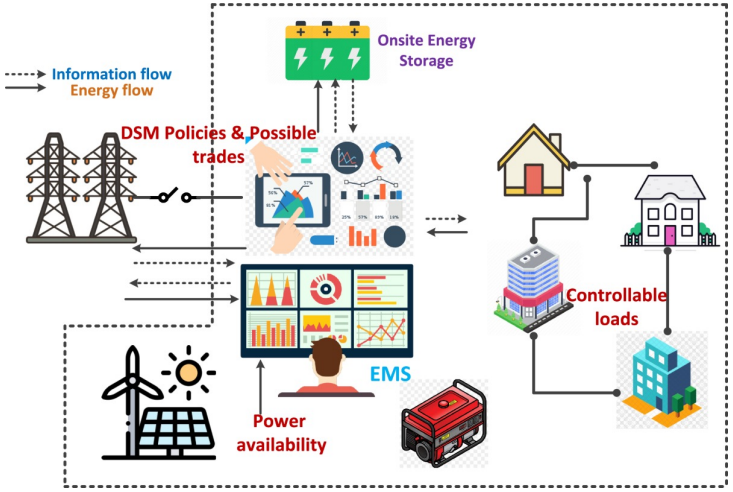
Problem Setting & Motivation

Motivation

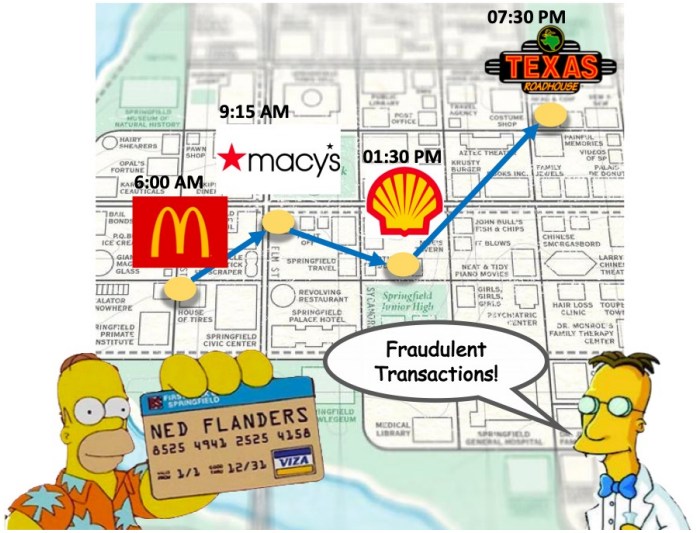
Anomalies correspond to rare but high-impact events

→ **Challenges:** rarity, noise, high-dimensionality, evolving distributions

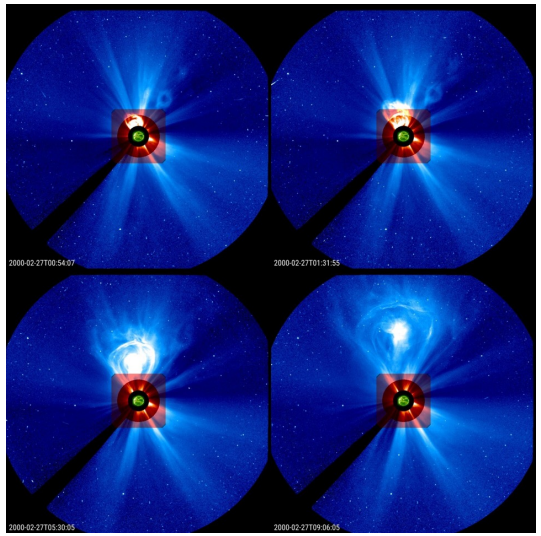
https://www.nature.com/articles/s41598-024-70336-3



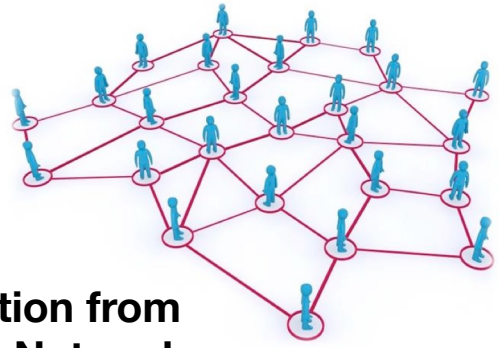
Power Network Monitoring



Credit Card Fraud Detection



Solar Flare Detection



Event Detection from Social Networks

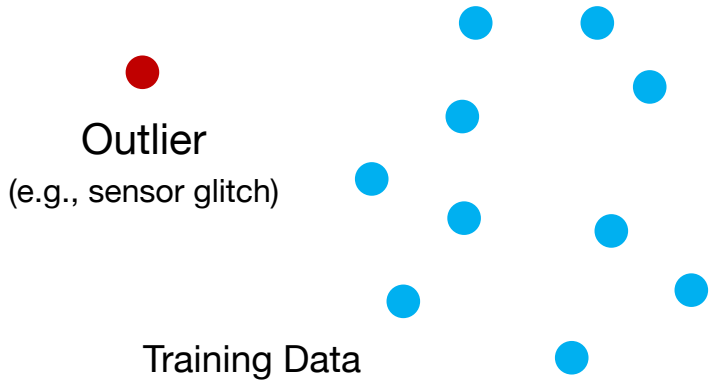
Omron Project Zero 2.0

Blood Pressure Monitoring

Anomaly Detection

Outlier Detection

Identify outliers in training data, defined as observations that are “far” from the others in a chosen metric—as pre-processing step to **clean data**



Novelty Detection

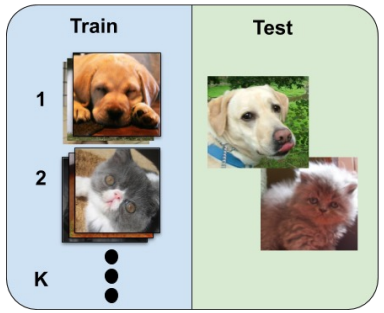
Detect whether a <u>new</u> observation is an outlier—related to **change point detection** (e.g., detecting change-points in time series)

Whether training data is contaminated or clean changes the problem

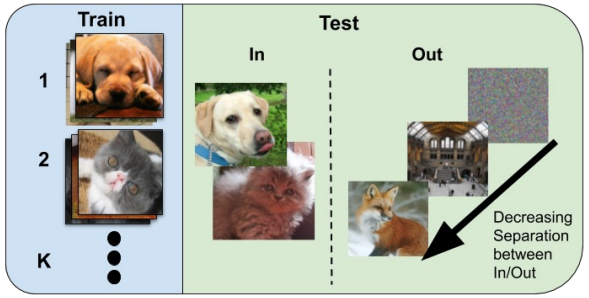
Out-of-Distribution (OOD) Detection

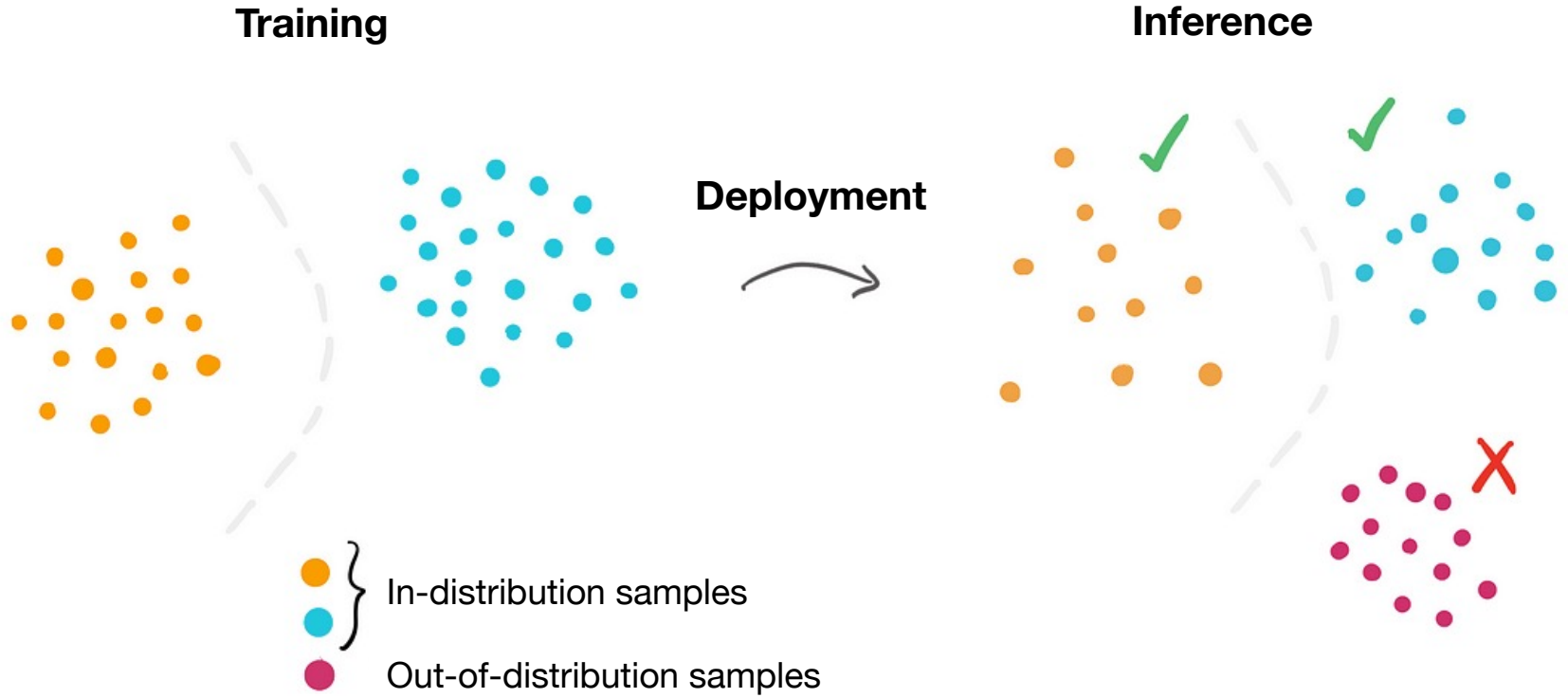
A classifier can be accurate and still unsafe if it cannot detect OOD inputs—OOD detection is a form of **novelty detection** **under distribution shift**

Multi-Class Classification



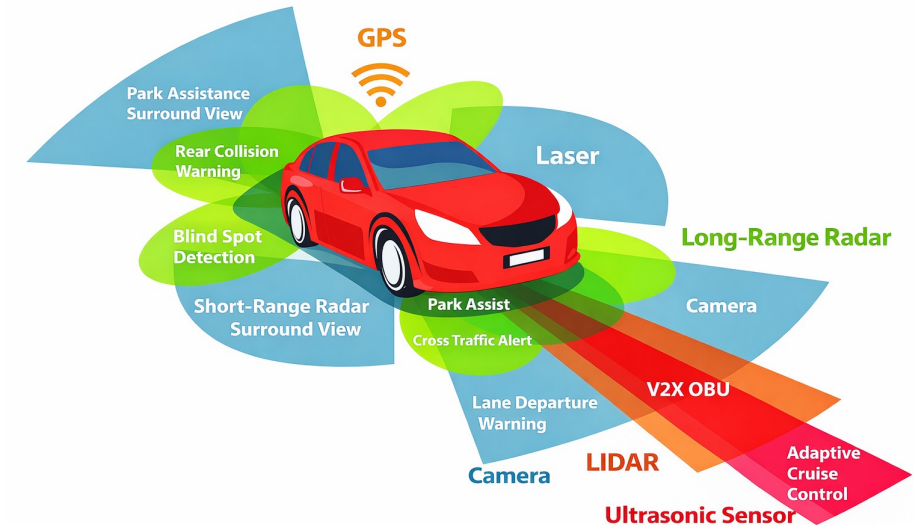
Classification with OOD Detection

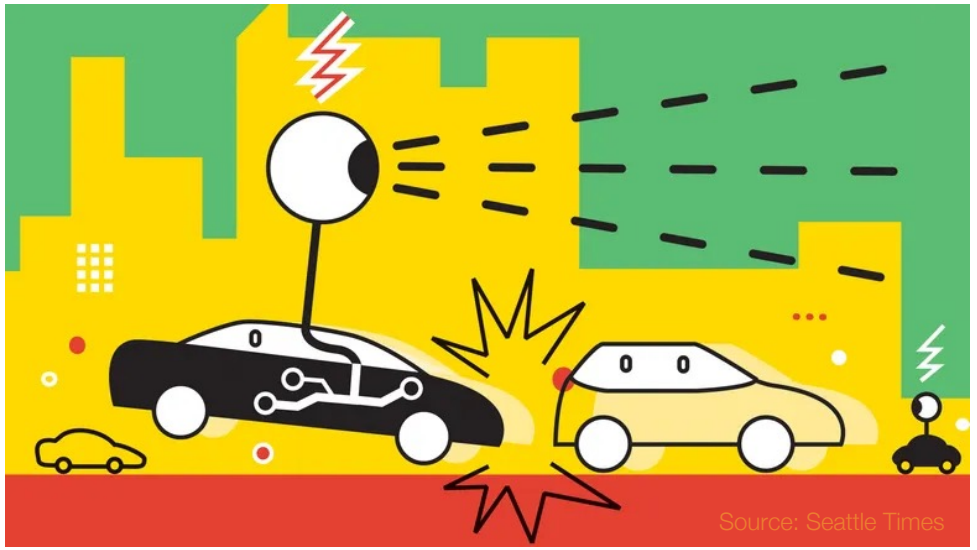




Example: Self-Driving Cars

Anomaly detection is critical for safety when encountering unseen objects or conditions: **out-of-distribution (OOD) detection**



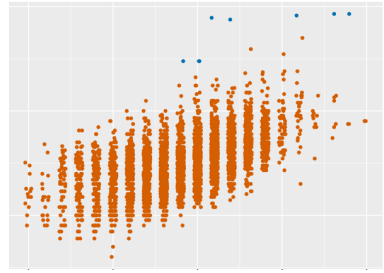


Why isn't high classification accuracy enough for self-driving cars?

Overview of Methods

Statistical Methods

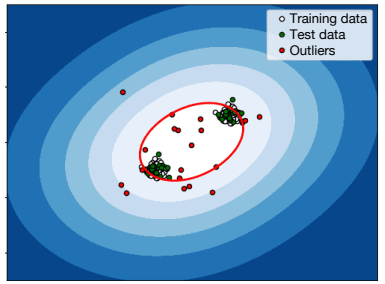
- Model the probability distribution of normal data
- Flag anomalies in low-density/low-likelihood regions
- Parametric (e.g., Gaussian) and nonparametric (e.g., KDE) approaches
- Detection via likelihood or distance thresholding



Statistical

Geometric Methods

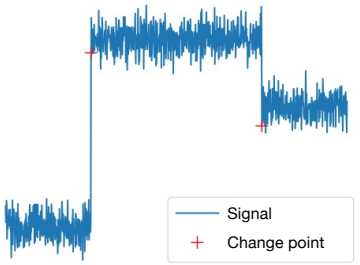
- Learn a decision boundary enclosing normal data
- Points outside the boundary are anomalies
- E.g., one-class SVM (a small set of support vectors)



Geometric

Sequential Methods

- Designed for time-ordered data
- Detect persistent distributional changes, not isolated points
- Accumulate evidence over time to balance fast detection and false alarms



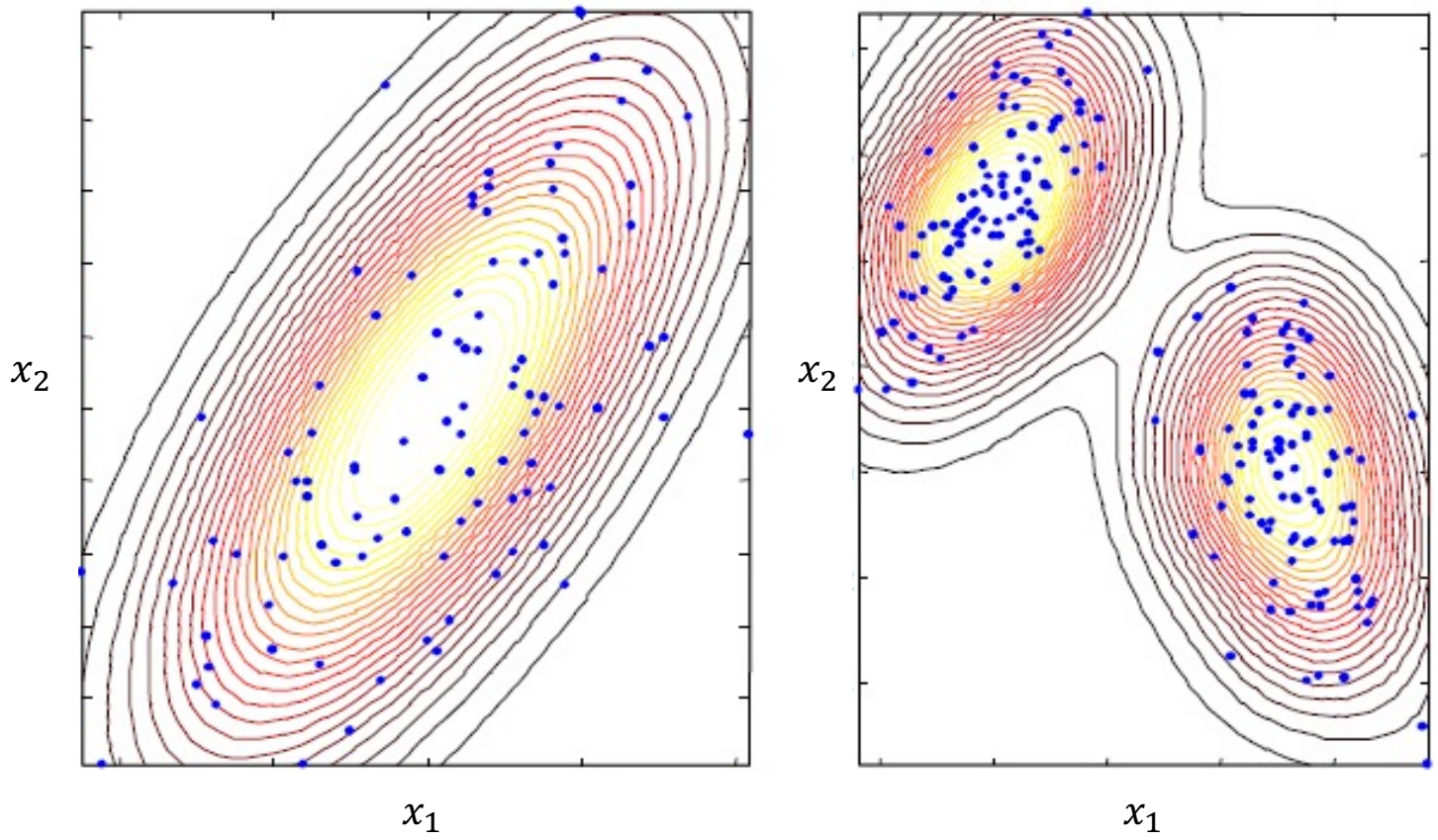
Sequential

Statistical Methods

Density-Based Outlier Detection

Declare anomalies as data points in **low density region**

- First fit density to normal data $\widehat{f_0}(x)$ and then set a threshold ϵ (e.g., using kernel density estimation or **KDE**)
- Data points in region with density small than ϵ, i.e., $\widehat{f_0}(x) \leq \epsilon$ are anomalous



Why might low density correspond to anomalies, but not always?

Covariance-Based Detector

Multivariate Gaussian

$$f_0(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}$$

Find parameter estimation from data

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x^i, \qquad \hat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m}(x^i - \hat{\mu})(x^i - \hat{\mu})^\top$$

For a new data, having $\widehat{f_0}(x_{\text{new}}) \leq \epsilon$ is equivalent to having

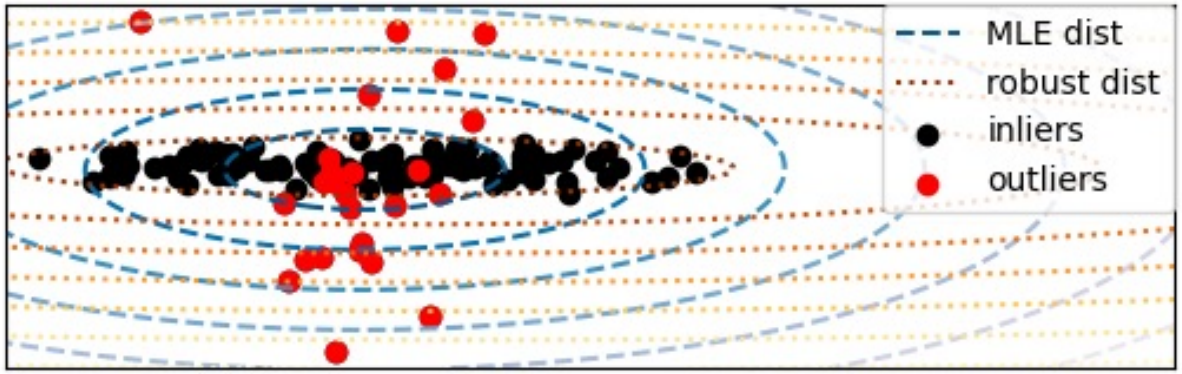
(Mahalanobis distance)^2 ← $(x_{\text{new}} - \hat{\mu})^\top \hat{\Sigma}^{-1}(x_{\text{new}} - \hat{\mu}) \geq b$ → Cutoff threshold

Testing whether a point lies far from the mean under the learned covariance

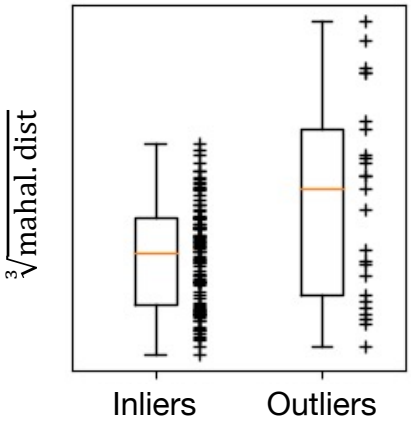
Robust Covariance Estimation

- **Problem:** Mahalanobis distance can be affected by outliers
- **Minimum Covariance Determinant (MCD) estimator:** Find a subset of observations whose empirical covariance has the smallest determinant $|\hat{\Sigma}|$

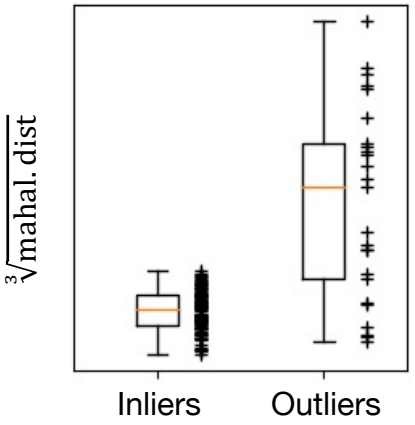
Mahalanobis distances of a contaminated dataset



From non-robust estimates (Maximum Likelihood)



From robust estimates (MCD)

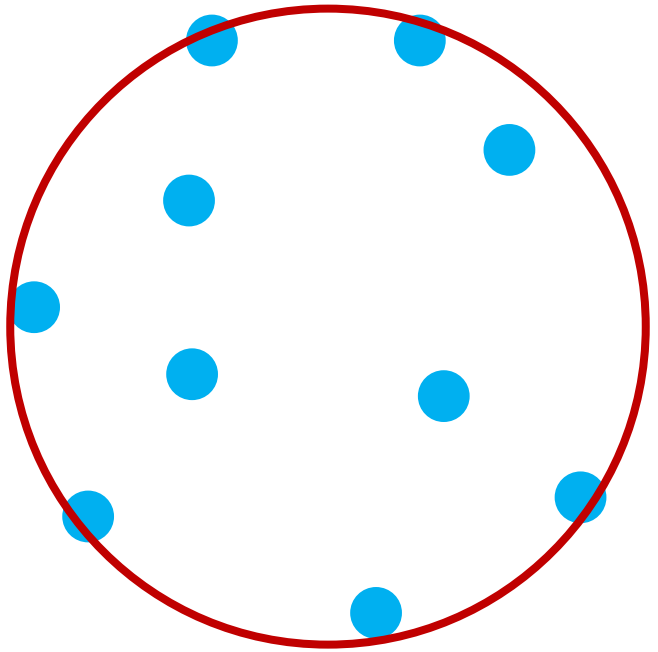


MCD finds the "tightest" subset of points that represent normal behavior

Geometric Methods

Geometric Viewpoint

Instead of fitting density and then finding threshold, **directly find the boundary**; i.e., learning the **support** of normal data



Geometric methods avoid explicit density estimation in high dimensions

One-Class SVM (Primal Formulation)

Find the smallest **ball** (region) with a center c and radius r that contains most **normal** data points

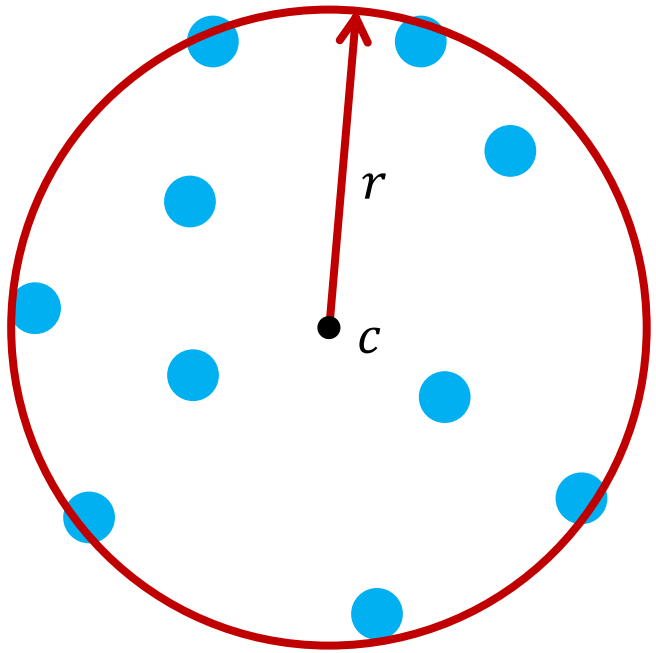
Optimization problem:

$$\min_{r,c} \; r^2$$

$$\text{s.t.} \quad \left\|x^i - c\right\|^2 \leq r^2$$

$$r \geq 0$$

$$\forall i = 1, \ldots, m$$



This is the anomaly detection analogue of margin maximization in SVM

Solving One-Class SVM

Lagrangian function

$$L(r, c, \alpha) = r^2 + \sum_{i=1}^{m} \alpha_i \left((x^i - c)^\top (x^i - c) - r^2 \right), \qquad \alpha_i \geq 0$$

Set partial derivative of L to 0

$$\frac{\partial L}{\partial r} = 2r - 2\sum_{i=1}^{m} \alpha_i r = 0 \quad \Rightarrow \quad \sum_{i=1}^{m} \alpha_i = 1$$

$$\frac{\partial L}{\partial c} = \sum_{i=1}^{m} \alpha_i (-2x^i + 2c) = 0 \quad \Rightarrow \quad c^* = \sum_{i=1}^{m} \alpha_i x^i$$

→ $r^* = \|x_i - c^*\|$ for any **support vector**

What does it mean that the center is a weighted sum of points?

Dual Problem

$$\max_{\alpha} g(\alpha) := L(r^*, c^*, \alpha) = \sum_{i=1}^{m} \alpha_i x^{i\top} x^i - \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j x^{i\top} x^j$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \alpha_i = 1\,, \qquad \alpha_i \geq 0$$

A quadratic programming problem $g(\alpha) = b^\top \alpha - \alpha^\top A \alpha$

$$A_{ij} = x^{i\top} x^j \qquad b_i = x^{i\top} x^i$$

Why is it useful that the dual depends only on inner products?

Sparsity in the Solution

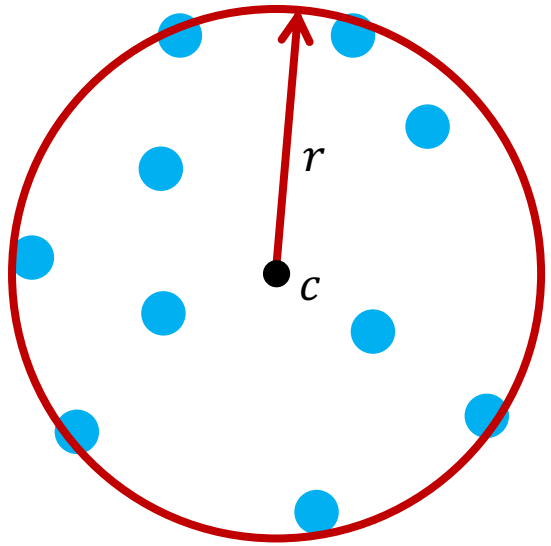
Complementary slackness (for all i)

$$\alpha_i \left((x^i - c)^\top (x^i - c) - r^2 \right) = 0, \qquad \alpha_i \geq 0$$

Many data points will be inside the circle

$$(x^i - c)^\top (x^i - c) - r^2 < 0$$

$$\alpha_i = 0$$



Few data points will be on the boundary (**support vectors**)

$$(x^i - c)^\top (x^i - c) - r^2 = 0$$

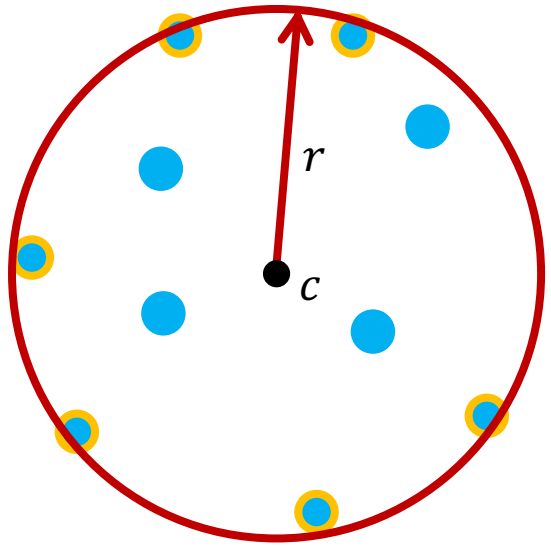
α_i can be nonzero: solution in α is very sparse!

Support Vectors

- Each α_i is associated with **one data point**
- Data points with nonzero α_i are **support vectors**
- They determine the **center** and **radius**

Center ← $c^* = \sum_{i=1}^{m} \alpha_i x^i$

Radius ← $r = \max_{i:\alpha_i>0} \|x^i - c^*\|$



Memory and computation depend on the number of support vectors

Kernelization

Dual problem

$$\max_{a} g(a) = b^{\top}\alpha - \alpha^{\top}A\alpha$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \alpha_i = 1\,, \qquad \alpha_i \geq 0$$

$$A_{ij} = x^{i\top}x^{j} \qquad b_i = x^{i\top}x^{i}$$

$k(x^i, x^j)$ $k(x^i, x^i)$

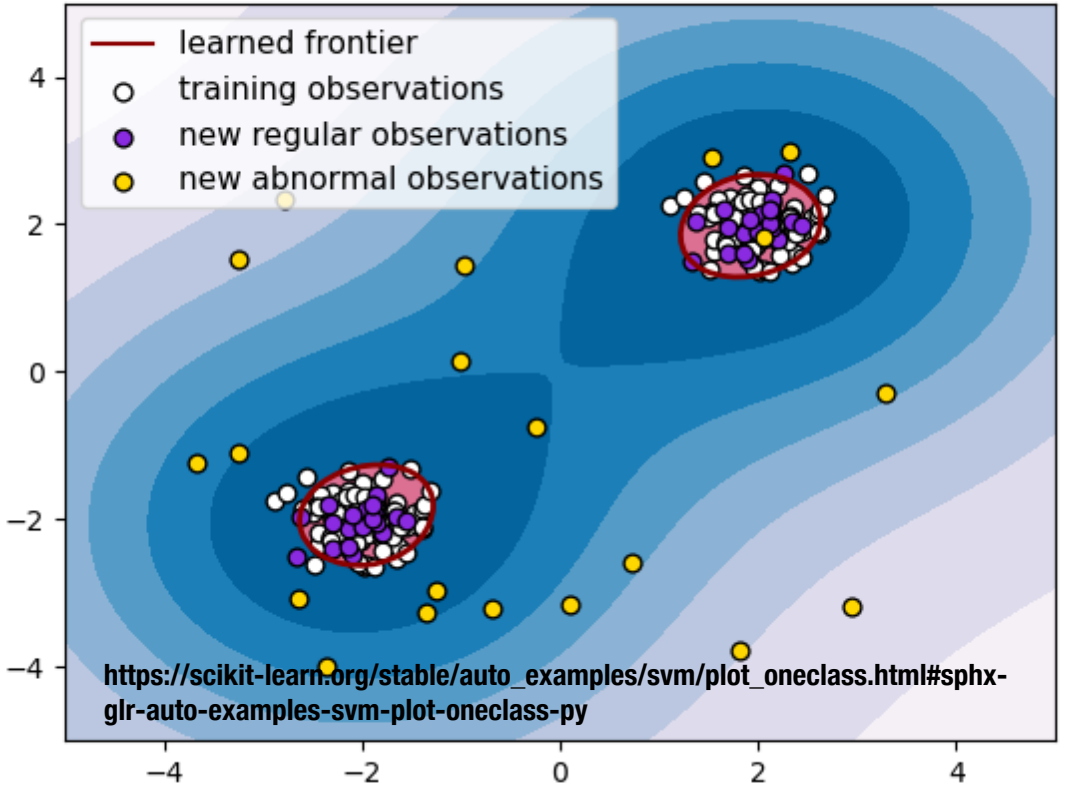
Why might nonlinear boundaries be essential for anomaly detection?

Demo: Novelty Detection

One-class SVM learns a boundary around normal data and flags novel points at test time

- Only normal data are used during training; anomalies appear only at inference
- This is powerful in practice: labeled anomalies are often unavailable or extremely rare

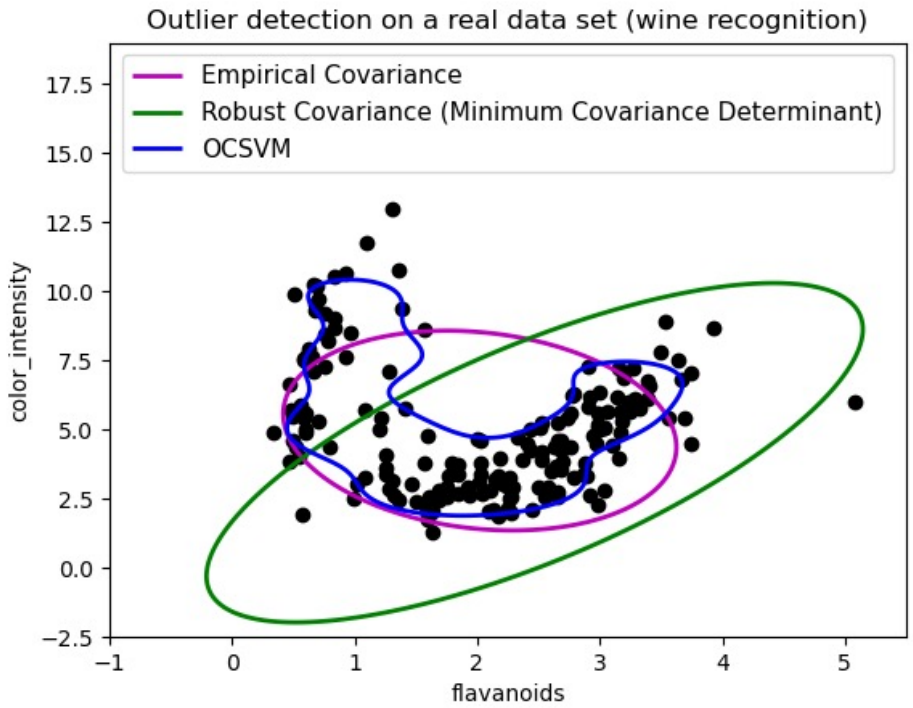
Why don't we need labeled anomalies to train this model?

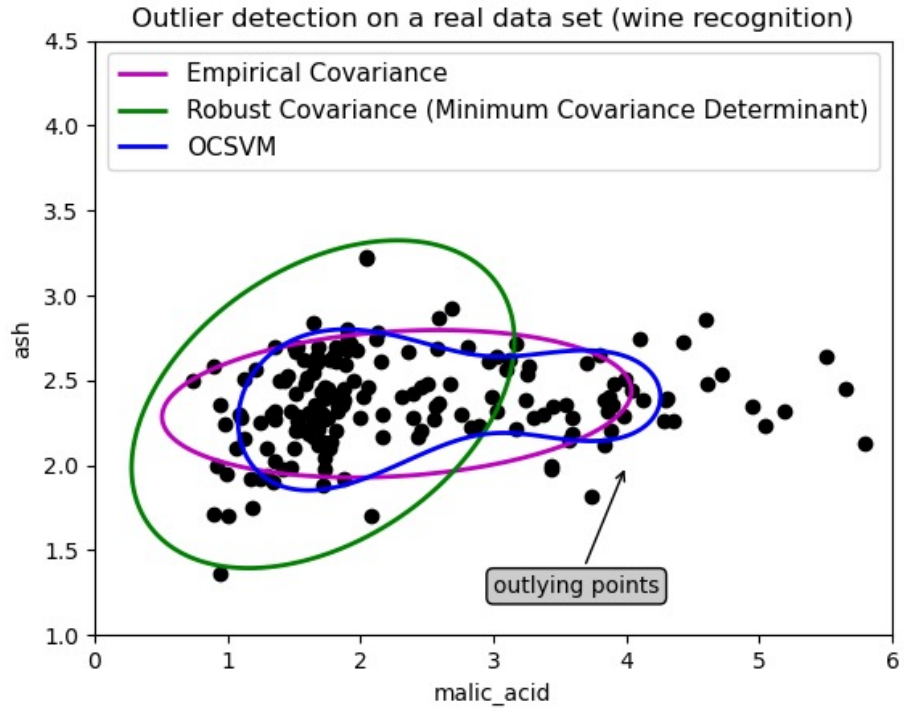


Example: Wine Dataset

Dataset: 27 chemical measurements of 178 wines

Task: Compare how different methods define **“normality”**





https://scikit-learn.org/stable/auto_examples/applications/plot_outlier_detection_wine.html#sphx-glr-auto-examples-applications-plot-outlier-detection-wine-py

Hypothesis Testing

Hypothesis Testing View

Common hypothesis test

$$H_0: x^i \sim f_0, \qquad i = 1, \ldots, m$$
$$H_1: x^i \sim f_1, \qquad i = 1, \ldots, m$$

Two-sample test

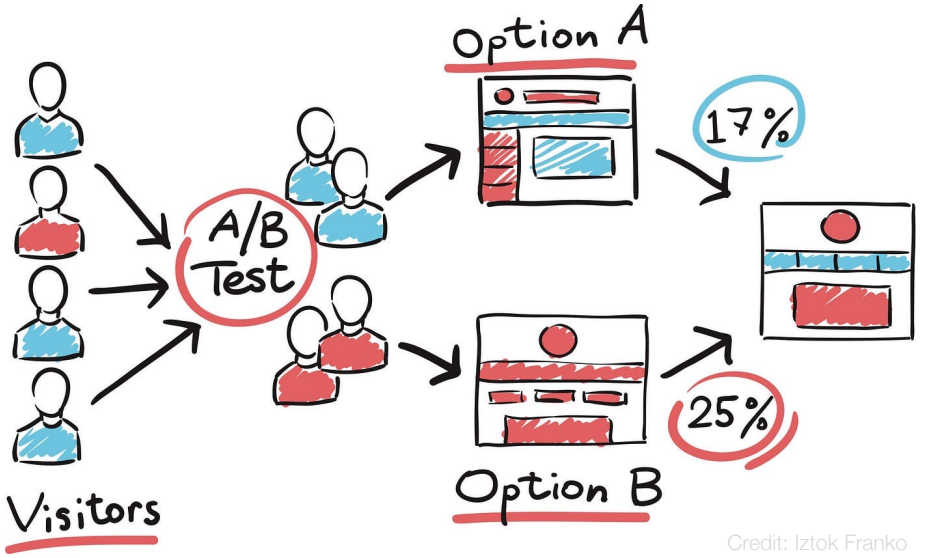
$$H_0: x^i \sim f_0, z^i \sim f_0, \qquad i = 1, \ldots, m$$
$$H_1: x^i \sim f_1, z^i \sim f_1 \qquad i = 1, \ldots, m$$

Anomaly detection

$$H_0: x^i \sim f_0, \qquad i = 1, \ldots, m$$
$$H_1: x^i \not\sim f_0, \qquad i = 1, \ldots, m$$

Hypothesis Testing	Anomaly Detection
Null hypothesis	Normal behavior
Alternative	Anomaly
Test statistic	Anomaly score

A/B Testing



Credit: Iztok Franko

Performance Metrics

Decision

Choose H_0 or H_1

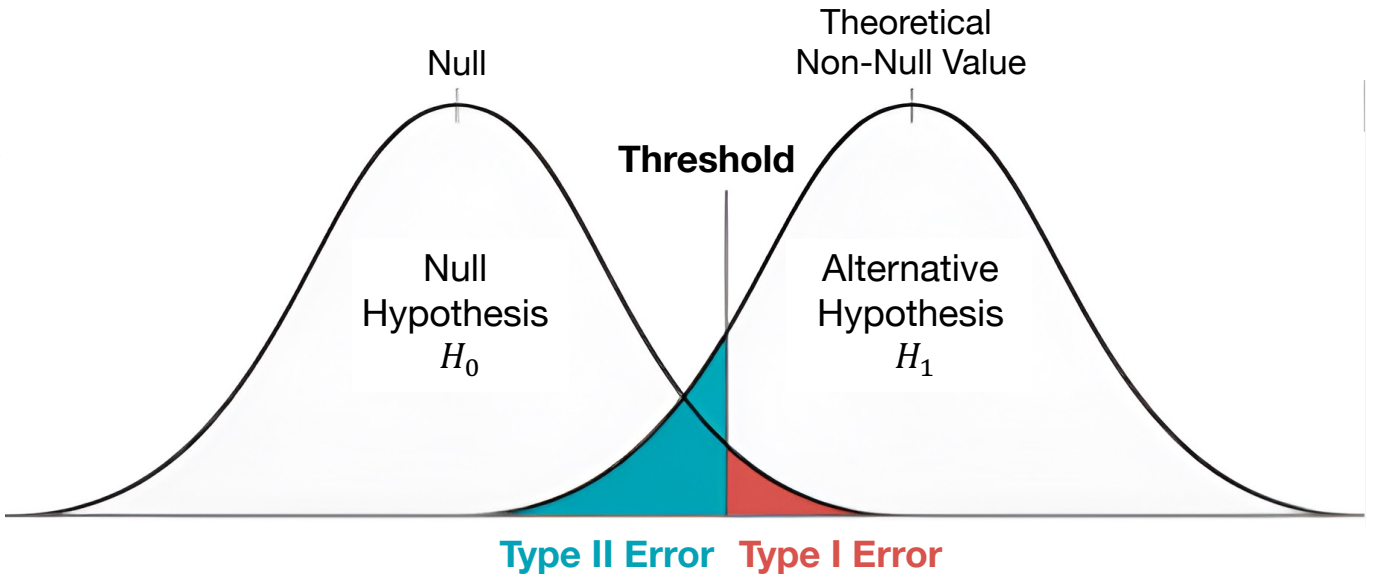
Decision	H_0 Is True	H_0 Is False
Fail to reject H_0	No error	**Type II error**
Reject H_0	**Type I error**	No error

Type-I Error = $p_{(H_0)}$ {reject H_0}

Falsely flag normal data as an anomaly

Type-II Error = $p_{(H_0)}$ {accept H_0}

Fail to detect an actual anomaly



In anomaly detection, false alarms are often more costly than misses

Likelihood Ratio Test

Compare how likely the data is under two distributions f_0 and f_1, and declare anomaly when the likelihood ratio exceeds a threshold:

$$\text{decide } H_1 \text{ when } \frac{f_1(x)}{f_0(x)} > b$$

- **Neyman-Pearson lemma:** it measures relative plausibility under two models
- Set the threshold b to control false alarms (Type-I error)
- **Limitation:** f_1 is rarely known in practice

Optimal when distributions are known—but rarely true in practice

Sequential Methods

Sequential Novelty Detection (CUSUM)

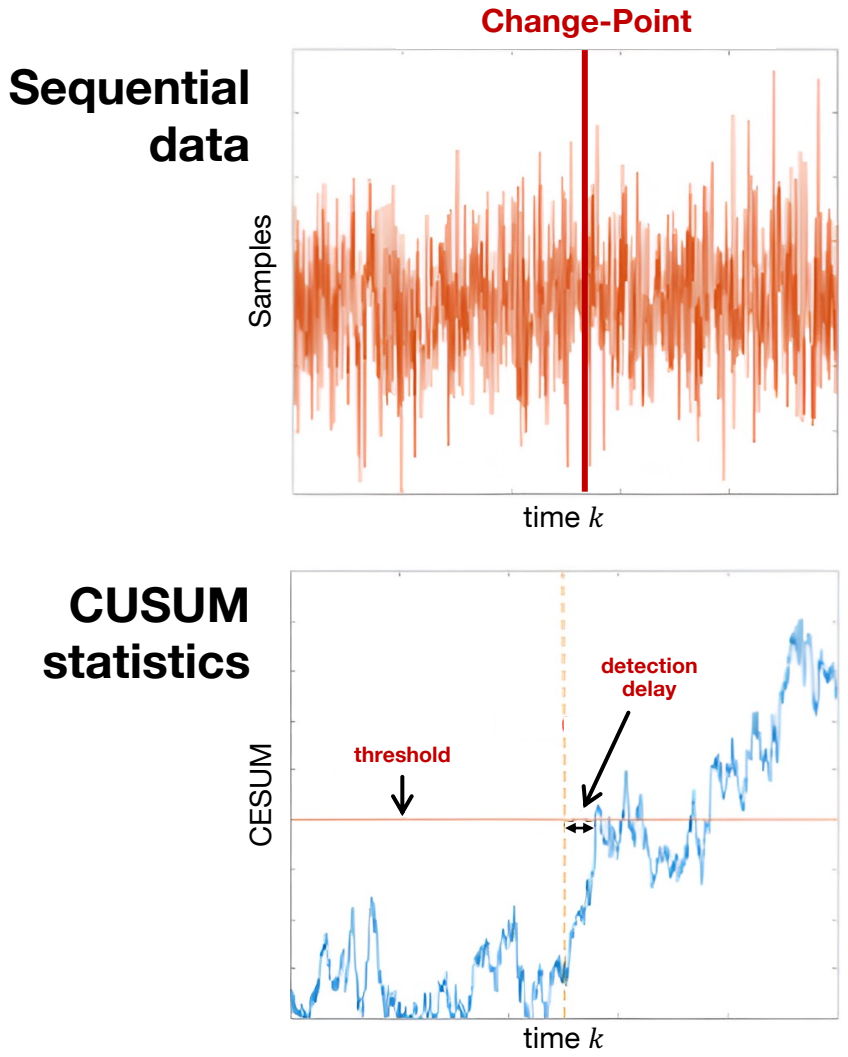
Assume the distribution before the change: $f_0(x)$, after the change: $f_1(x)$

CUMulative **SUM** over log-likelihood ratios:

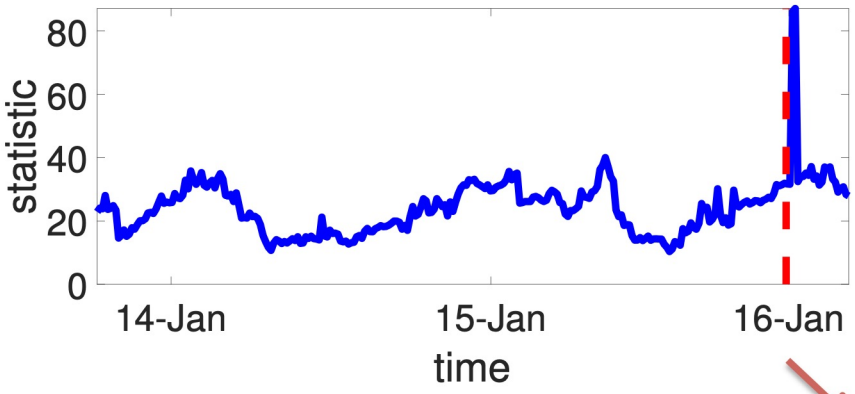
$$W_0 = 0 \qquad W_t = \max\left(W_{t-1} + \log\frac{f_1(X_t)}{f_0(X_t)}, 0\right)$$

- W_t accumulates evidence when data looks more like f_1 than f_0
- Anomaly is flagged if $W_t > h$ (h: threshold)
- W_t resets to 0 if evidence weakens

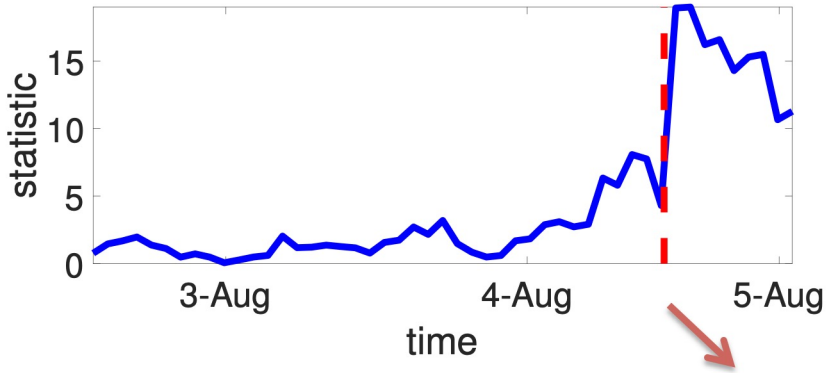
Why not just test each point independently?



Change-Points → Major Events



Israel announces ceasefire in Gaza War in 2009

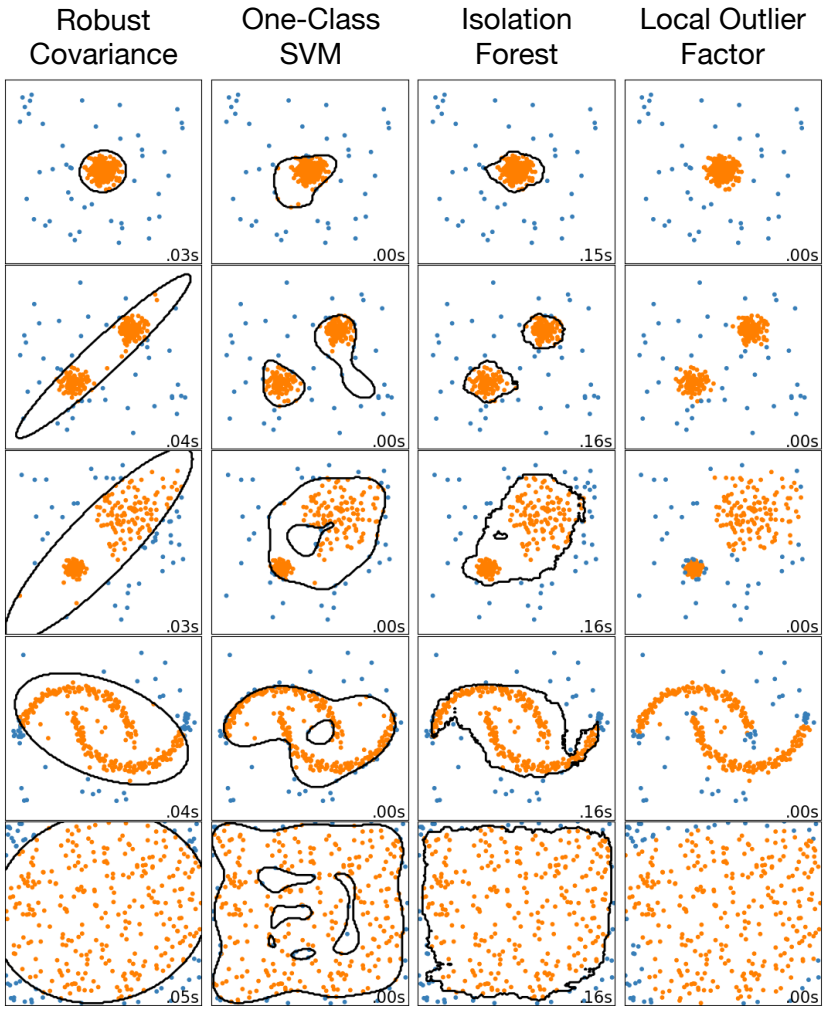


Beijing Olympics opening in 2008

Comparison of Algorithms

Which method works best depends on data geometry, noise, and dimensionality

- **Robust Covariance:** fits a Gaussian to normal data (robustly) and flags points far via Mahalanobis distance
- **One-Class SVM:** learns a boundary enclosing normal data and labels outside points as anomalies
- **Isolation Forest:** isolates anomalies by randomly partitioning data—anomalies are easier to isolate (short paths)
- **Local Outlier Factor (LOF):** detects points whose local density is much lower than their neighbors



Key Takeaways

What We Learned This Week

- Anomaly detection focuses on identifying rare or abnormal patterns and includes outlier detection, novelty detection, and out-of-distribution detection
- Statistical methods flag anomalies as low-likelihood events using models such as Gaussian/Mahalanobis, robust covariance, and kernel density estimation
- Geometric methods (e.g., one-class SVM) learn a boundary around normal data without explicitly modeling probability densities
- Anomaly detection can be framed as a hypothesis testing problem, with thresholds chosen to control false alarm rates (Type-I error)
- Sequential methods like CUSUM detect changes over time by accumulating weak evidence into strong signals
- Many real-world anomalies arise in structured data (time series, networks), where detecting changes in summary statistics enables scalable detection