CSE 435: Data Mining

Chapter 11: Outlier Detection

Md Atikuzzaman
Lecturer
Department of Computer Science & Engineering
Green University of Bangladesh
atik@cse.green.edu.bd

GREEN UNIVERSITY

Learning Outcomes

By the end of this chapter, students should be able to:

1. explain global, contextual, and collective outliers;
2. distinguish supervised, semi-supervised, and unsupervised detection;
3. calculate statistical, distance, density, and reconstruction scores;
4. compare clustering- and classification-based approaches;
5. explain why high-dimensional data require subspace-aware methods;
6. select a method, threshold, and evaluation metric for a real problem.

Seven Views of Abnormality

Section	Main question	Typical evidence
11.1	What counts as an outlier?	deviation, context, group behavior
11.2	Is the object improbable?	probability or statistical test
11.3	Is the object far away or locally sparse?	distance and density
11.4	Is the object difficult to reconstruct?	residual error
11.5	Does it fit a cluster or learned class?	membership and boundary
11.6	Is it abnormal only in context or as a group?	conditional and sequence behavior
11.7	Which dimensions reveal the anomaly?	subspaces and ensembles

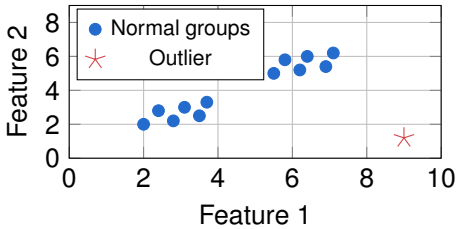
Source: Han, Pei, and Tong, *Data Mining: Concepts and Techniques*, 4th ed., Chapter 11.

11.1 Basic Concepts

Outliers Suggest a Different Data Mechanism

Working definition

An outlier is a data object that deviates significantly from normal objects, as if it were generated by a different mechanism.



Possible meanings

- a valid rare event;
- fraud or intrusion;
- equipment failure;
- measurement or entry error;
- a new population or concept drift.

Detection identifies unusualness. Investigation determines the cause.

Outlier vs. Noise

	Outlier	Noise
Meaning	A meaningful object that strongly deviates from a pattern	Random error or unwanted variation
Example	A real high-value transaction from a compromised account	A sensor spike caused by electrical interference
Goal	Detect, rank, explain, and investigate	Remove, smooth, correct, or model robustly
Danger	Deleting it may remove the event of interest	Keeping it may distort the learned pattern

Data-centric warning

A valid extreme value is not automatically an error, and an in-range value is not automatically normal.

Three Types of Outlier

Global

Is one object unusual compared with the entire data set?

$s(\boldsymbol{x})$ uses $\mathcal{D}$

Example: an amount of 50,000.

Contextual

Is one object unusual under a specific context?

$s(\boldsymbol{x})$ uses $\mathcal{D}_c$

Example: 30°C in winter.

Collective

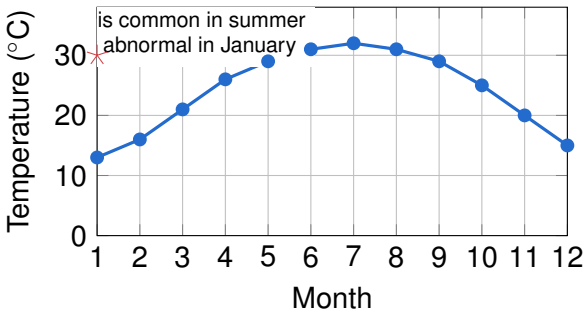
Is a group unusual although its members may look ordinary?

$s(W)$ uses a set or sequence

Example: many rapid login attempts.

Always define the unit of analysis: object, context-specific object, window, sequence, or graph substructure.

Context Changes What Counts as an Outlier



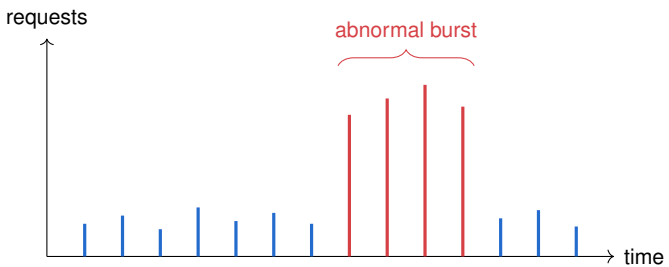
Contextual attributes

- month;
- location;
- user or cohort;
- device type.

Behavioral attribute

- temperature.

Collective Outliers Live in Relationships



Each request may be valid by itself.

- The rate is abnormal.
- The ordering is abnormal.
- The repeated source-target pattern is abnormal.

Examples: denial-of-service traffic, ECG subsequences, coordinated reviews.

Labels Determine the Learning Setting

Setting	Training data	Main strength	Main difficulty
Supervised	labeled normal and outlier objects	learns task-specific differences	anomalies are rare, diverse, and expensive to label
Semi-supervised	mostly verified normal objects	avoids requiring every anomaly type	contaminated normal data can weaken the boundary
Unsupervised	no reliable labels	usable when incidents are unknown	model assumptions define normality

Practical default

Use simple unsupervised scores to rank objects, then collect expert feedback and move toward semi-supervised or supervised learning when labels become reliable.

Four Challenges Shape Every Detector

1. **Rarity:** high class imbalance makes accuracy misleading.
2. **Heterogeneity:** anomalies may be produced by many mechanisms.
3. **Context and drift:** normal behavior changes across groups and time.
4. **Dimensionality:** irrelevant attributes hide informative deviation.

$$\underbrace{\text{data representation}}_{\text{what is compared}} + \underbrace{\text{reference population}}_{\text{what is normal}} + \underbrace{\text{score and threshold}}_{\text{what is flagged}} = \text{detector behavior}$$

The algorithm is only one part of an outlier detection system.

11.2 Statistical Approaches

Statistical Detection Finds Improbable Objects

Assume data are generated by a probability model with parameter θ:

$$\boldsymbol{x} \sim p(\boldsymbol{x} \mid \theta).$$

A simple anomaly score is negative log-likelihood:

$$s(\boldsymbol{x}) = -\log p(\boldsymbol{x} \mid \hat{\theta}).$$

High density

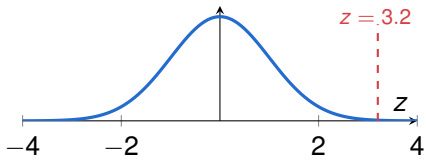
- object agrees with the fitted model;
- score is small.

Low density

- object is improbable under the model;
- score is large.

Assumption check: A low likelihood may reflect model mismatch rather than a true anomaly.

Z-Scores Standardize Gaussian Values



Worked example

$$\mu = 100, \quad \sigma = 5, \quad x = 116$$

$$z = \frac{x - \mu}{\sigma} = \frac{116 - 100}{5} = 3.2$$

The value is 3.2 standard deviations above the mean.

A large $|z|$ is evidence of unusualness, not proof of an error.

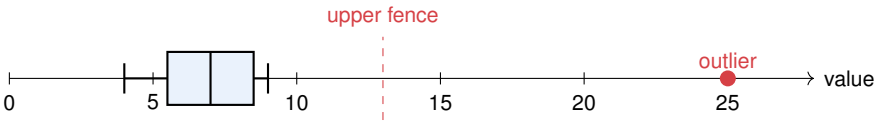
The IQR Rule Is Robust to Extreme Values

For the sorted data

$$4,\ 5,\ 6,\ 6,\ 7,\ 8,\ 8,\ 9,\ 25$$

$$Q_1 = 5.5, \quad Q_3 = 8.5, \quad IQR = Q_3 - Q_1 = 3.$$

$$\text{Lower fence} = 1, \qquad \text{Upper fence} = 13.$$



Since $25 > 13$, the IQR rule flags 25.

Robust Statistics Resist Extreme Values

Classical estimate

$$\bar{x} = \frac{1}{n}\sum_i x_i, \qquad s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$$

Extreme objects pull both $\bar{x}$ and s toward themselves.

Robust estimate

$$\tilde{x} = \text{median}(x_i)$$

$$MAD = \text{median}\,|x_i - \tilde{x}|$$

$$z_i^{(r)} = 0.6745\frac{x_i - \tilde{x}}{MAD}$$

Masking and swamping

Masking: anomalies distort the model and escape detection.

Swamping: valid objects are incorrectly flagged.

Mahalanobis Handles Scale and Correlation

$$D_M^2(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}).$$

Worked example

$$\boldsymbol{\mu} = \begin{bmatrix} 10 \\ 20 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 3 \\ 3 & 9 \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} 14 \\ 26 \end{bmatrix}$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{27} \begin{bmatrix} 9 & -3 \\ -3 & 4 \end{bmatrix}$$

$$D_M^2 = \frac{144}{27} = 5.33, \qquad D_M = 2.31$$

Why not Euclidean alone?

- units may differ;
- features may correlate;
- normal regions may be elliptical.

Chi-Square Turns Distance into a Test

If normal data follow a d-dimensional Gaussian distribution, then approximately

$$D_M^2(\boldsymbol{x}) \sim \chi_d^2.$$

For the previous example, $d = 2$ and $D_M^2 = 5.33$.

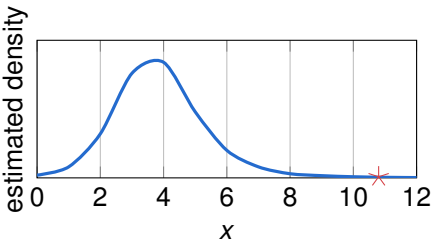
$$\chi_{2,0.95}^2 = 5.99.$$

Because $5.33 < 5.99$, the object is not flagged at the 5% significance level.

Interpretation

A score can be large relative to nearby data but still remain inside the chosen statistical boundary. The threshold encodes how much false-alarm risk is accepted.

KDE Estimates Density Without a Fixed Family



$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- Small h: sensitive, irregular estimate.
- Large h: smooth estimate that may hide local anomalies.
- Score: $-\log \hat{f}_h(x)$.

When Statistical Methods Work Best

Method	Good choice when	Main limitation
Z-score	one symmetric, light-tailed attribute	not robust; univariate
IQR or robust z	one skewed or contaminated attribute	ignores relationships
Mahalanobis	correlated continuous features with an elliptical normal region	covariance is difficult in high dimensions
Gaussian mixture	multiple probabilistic normal groups	model selection and local optima
Histogram or KDE	distribution family is unknown	bin width or bandwidth; dimension grows costly

Start with the simplest statistical model that matches the data-generating process.

11.3 Proximity-Based Approaches

Proximity Uses Neighborhood Evidence

Core assumption

An outlier is far from most objects or has much lower local density than its neighbors.

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{d} |x_j - y_j|^p \right)^{1/p}$$

- $p = 1$: Manhattan distance.
- $p = 2$: Euclidean distance.
- Similarity for text or sparse vectors may require cosine distance.
- Scaling is essential when feature units differ.

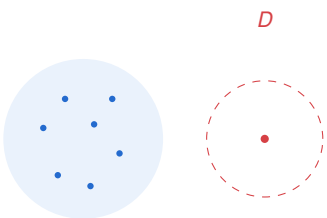
  $d((1000 \text{ taka}, 1 \text{ km}))$ is meaningless until units are made comparable.

Distance-Based Outliers Use a Radius

DB(p, D) definition

An object $\boldsymbol{o}$ is an outlier if at least a fraction p of all objects are farther than distance D from $\boldsymbol{o}$.

$$\frac{|\{\boldsymbol{x} \in \mathcal{D} : d(\boldsymbol{x}, \boldsymbol{o}) > D\}|}{|\mathcal{D}|} \geq p.$$



Question How many objects fall outside the red object's D-neighborhood?

k-NN Scores Rank Objects by Distance

Two common scores are

$$s_{\max}(\boldsymbol{x}) = d\big(\boldsymbol{x}, k\text{-th NN}(\boldsymbol{x})\big),$$

$$s_{\text{avg}}(\boldsymbol{x}) = \frac{1}{k} \sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} d(\boldsymbol{x}, \boldsymbol{o}).$$

Small k

- detects very local deviation;
- a small anomalous group may protect its members.

Large k

- produces a more global score;
- small valid clusters may be over-flagged.

Worked Example: k-NN Distance

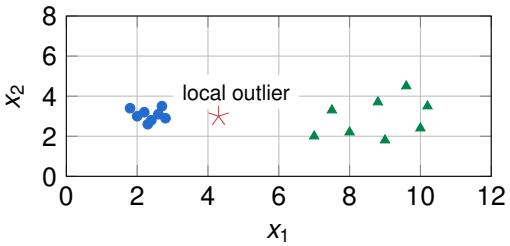
Data: $\{1,\ 2,\ 2.5,\ 3,\ 10\}$, using absolute distance and $k = 2$.

x	Two nearest distances	$s_{\text{avg}}(x)$	Rank
1	1, 1.5	1.25	2
2	0.5, 1	0.75	3
2.5	0.5, 0.5	0.50	5
3	0.5, 1	0.75	3
10	7, 7.5	7.25	1

$$s_{\text{avg}}(10) = \frac{7 + 7.5}{2} = 7.25.$$

Ranking is clear; the alert threshold still requires a decision rule.

Global Distance Can Miss Local Outliers



The green cluster is valid but sparse.

- Global distance may over-score green objects.
- Local density compares an object with its own neighborhood.

LOF Compares Neighbor Densities

$$\text{reach-dist}_k(p, o) = \max\{k\text{-distance}(o), d(p, o)\}$$

$$\text{lrd}_k(p) = \left(\frac{\sum_{o \in N_k(p)} \text{reach-dist}_k(p, o)}{|N_k(p)|} \right)^{-1}$$

$$\text{LOF}_k(p) = \frac{1}{|N_k(p)|} \sum_{o \in N_k(p)} \frac{\text{lrd}_k(o)}{\text{lrd}_k(p)}$$

- LOF ≈ 1: density resembles the neighbors.
- LOF > 1: object is locally sparser.
- A much larger value provides stronger local-outlier evidence.

Worked Example: LOF

Suppose point p has

$$\mathrm{lrd}_k(p) = 0.25,$$

and its three neighbors have local reachability densities

$$0.50, \quad 0.40, \quad 0.60.$$

Then

$$\mathrm{LOF}_k(p) = \frac{1}{3}\left(\frac{0.50}{0.25} + \frac{0.40}{0.25} + \frac{0.60}{0.25}\right) = \frac{2.0 + 1.6 + 2.4}{3} = 2.0.$$

Interpretation

The neighbors are, on average, twice as dense as p's region. Thus p is a strong local-outlier candidate.

Proximity Methods: Strengths and Limits

Strengths

- no explicit distribution;
- intuitive explanations through neighbors;
- distance and density capture different structures.

Limitations

- naive pairwise computation is $O(n^2)$;
- performance depends on metric, scale, and k;
- distances become less informative in high dimensions.

Scalability tools

Spatial indexes, approximate nearest neighbors, grids, sampling, and parallel computation reduce search cost.

11.4 Reconstruction Methods

Reconstruction Learns Compact Normal Structure

Represent the data matrix as

$$\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{V}^{\top},$$

where the rank of $\boldsymbol{U}\boldsymbol{V}^{\top}$ is small.

$$\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^{\top}.$$

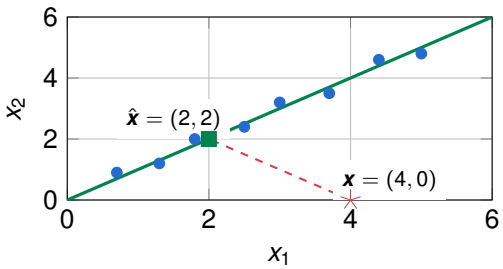
For object i, use row residual

$$s(\boldsymbol{x}_i) = \|\boldsymbol{e}_i\|_2 \quad \text{or} \quad s(\boldsymbol{x}_i) = \|\boldsymbol{e}_i\|_1.$$

Assumption

Normal objects share a compact structure. Outliers do not fit that structure and therefore leave large residuals.

PCA Scores Distance from a Normal Subspace



The green line is the learned one-dimensional normal subspace.

$$\hat{\boldsymbol{x}} = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}$$

$$s(\boldsymbol{x}) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2$$

Worked Example: PCA Residual

Let the normal direction be

$$\boldsymbol{u} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}.$$

Projection:

$$\hat{\boldsymbol{x}} = \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{x} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}.$$

Residual and score:

$$\boldsymbol{e} = \boldsymbol{x} - \hat{\boldsymbol{x}} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \qquad s(\boldsymbol{x}) = \|\boldsymbol{e}\|_2^2 = 2^2 + (-2)^2 = 8.$$

A normal point near the line has a small residual; a point off the line has a large residual.

Robust PCA Separates Structure and Corruption

$$\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S},$$

where $\boldsymbol{L}$ is low-rank normal structure and $\boldsymbol{S}$ contains sparse unusual entries.

$$\min_{\boldsymbol{L},\boldsymbol{S}} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{S}\|_1 \quad \text{subject to} \quad \boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}.$$

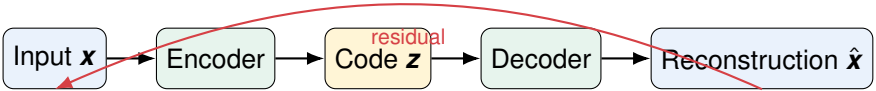
- $\|\boldsymbol{L}\|_*$ encourages low rank.
- $\|\boldsymbol{S}\|_1$ encourages sparse deviations.
- Large row or entry values in $\boldsymbol{S}$ identify anomalies.

Caution

If anomalies are common or form a coherent low-rank pattern, they may be absorbed into the normal component.

Autoencoders Learn Nonlinear Reconstruction

$$\boldsymbol{z} = f_\theta(\boldsymbol{x}), \qquad \hat{\boldsymbol{x}} = g_\phi(\boldsymbol{z}), \qquad s(\boldsymbol{x}) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2.$$



- Useful for images, signals, and nonlinear manifolds.
- A model with excessive capacity may also reconstruct anomalies well.
- Train on a trusted normal reference when possible.

11.5 Clustering vs. Classification

Clustering Reuses Group Structure

Three common assumptions:

1. Normal objects belong to a cluster; nonmembers are outliers.
2. Normal objects lie near a cluster center; distant members are outliers.
3. Normal clusters are large and dense; tiny or sparse clusters are anomalous.

For centroid $\boldsymbol{\mu}_{c(i)}$ of object i:

$$s(\boldsymbol{x}_i) = d(\boldsymbol{x}_i, \boldsymbol{\mu}_{c(i)}).$$

For a cluster-level score:

$$s(C_j) = \alpha \frac{1}{|C_j|} + (1 - \alpha) \frac{1}{\text{density}(C_j)}.$$

Worked Example: Centroid Distance

Suppose a cluster has centroid

$$\boldsymbol{\mu} = (2, 3).$$

For three assigned objects:

$$\boldsymbol{x}_1 = (2, 4), \qquad \boldsymbol{x}_2 = (3, 3), \qquad \boldsymbol{x}_3 = (7, 8).$$

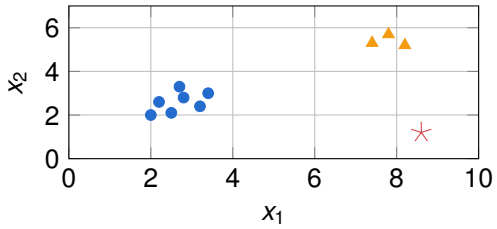
$$s(\boldsymbol{x}_1) = \sqrt{(2-2)^2 + (4-3)^2} = 1,$$

$$s(\boldsymbol{x}_2) = \sqrt{(3-2)^2 + (3-3)^2} = 1,$$

$$s(\boldsymbol{x}_3) = \sqrt{(7-2)^2 + (8-3)^2} = \sqrt{50} = 7.07.$$

Decision: $\boldsymbol{x}_3$ is strongest, provided the cluster correctly represents the population.

Small Clusters Are Not Automatically Abnormal



The orange group may represent:

- a valid minority population;
- a new product or cohort;
- a coordinated anomaly.

Check domain meaning, persistence, and external evidence.

Classification Learns an Outlier Boundary

Two-class classification

Train on labeled normal and outlier examples.

$$s(\boldsymbol{x}) = P(y = \text{outlier} \mid \boldsymbol{x}).$$

Best when: labels cover important anomaly types.

One-class classification

Train a boundary around normal data.

$$s(\boldsymbol{x}) = \max\{0, \rho - \boldsymbol{w}^\top \phi(\boldsymbol{x})\}.$$

Best when: trusted normal examples exist but anomalies are poorly sampled.

Generalization risk

A supervised classifier often detects anomalies similar to its training labels but can miss a new mechanism.

One-Class SVM Learns the Normal Support

$$\min_{\boldsymbol{w},\rho,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n} \xi_i - \rho$$

$$\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) \geq \rho - \xi_i, \qquad \xi_i \geq 0.$$

Decision function:

$$f(\boldsymbol{x}) = \operatorname{sign}\big(\boldsymbol{w}^\top \phi(\boldsymbol{x}) - \rho\big).$$

- ν controls tolerated violations and the support-vector fraction.
- An RBF kernel can learn a nonlinear normal boundary.
- Scaling and kernel width strongly affect the result.

Clustering vs. Classification

	Clustering-based	Classification-based
Evidence	membership, distance, cluster size, density	label probability or boundary violation
Labels	not required	required for two-class; normal labels for one-class
Strength	reveals structure and multiple groups	learns task-specific separation
Failure mode	valid small clusters may be flagged	unseen anomaly types may be missed
Explanation	nearest cluster and centroid distance	features, margin, probability, prototypes

Choose a Method from the Available Evidence

Data condition	Strong starting method	Reason
One interpretable variable	IQR or robust z-score	transparent baseline
Elliptical continuous data	robust Mahalanobis	models covariance
Unequal local density	LOF	compares nearby density
Low-rank numerical structure	PCA or robust factorization	residual identifies poor fit
Several meaningful groups	clustering-based score	uses group structure
Reliable anomaly labels	supervised classifier	learns task-specific distinction
Trusted normal reference	one-class model	estimates the normal support

11.6 Contextual and Collective

Context Separates Circumstances from Behavior

Partition the attributes into

$$\boldsymbol{x} = (\boldsymbol{c}, \boldsymbol{b}),$$

where $\boldsymbol{c}$ contains contextual attributes and $\boldsymbol{b}$ contains behavioral attributes.

$$s(\boldsymbol{x}) = -\log p(\boldsymbol{b} \mid \boldsymbol{c}).$$

Context

- time and season;
- location;
- user, device, or cohort.

Behavior

- amount or rate;
- temperature;
- duration or resource use.

Compare an object with the right peer group, not automatically with the entire data set.

Two Ways to Model Context

1. **Stratify or condition:** build a reference distribution for each context.

$$z_{\text{context}} = \frac{x - \mu_{\text{context}}}{\sigma_{\text{context}}}.$$

2. **Model expected behavior:** predict behavior from context and score the residual.

$$\hat{b} = f(\boldsymbol{c}), \qquad e = b - \hat{b}, \qquad s = |e|.$$

Examples

Compare transaction amount with the customer's history, energy use with the same hour and season, or response time with the same service and traffic load.

Worked Example: Seasonal Residual

Suppose expected January temperature is

$$\mu_{\text{Jan}} = 14^\circ C, \qquad \sigma_{\text{Jan}} = 3^\circ C.$$

Observed temperature:

$$x = 30^\circ C.$$

Contextual z-score:

$$z_{\text{Jan}} = \frac{30 - 14}{3} = 5.33.$$

Yet relative to a summer distribution with

$$\mu_{\text{Jul}} = 31^\circ C, \qquad \sigma_{\text{Jul}} = 2^\circ C,$$

the same value has

$$z_{\text{Jul}} = \frac{30 - 31}{2} = -0.5.$$

The same value can be anomalous or normal depending on context.

Collective Detection Scores Groups and Sequences

For a sequence window

$$W_t = (x_{t-w+1}, \ldots, x_t),$$

a likelihood-based score is

$$s(W_t) = -\sum_{j=t-w+1}^{t} \log p(x_j \mid x_{j-1}, \ldots).$$

Other collective scores may use:

- an unusual subsequence shape;
- a burst in rate or frequency;
- an unexpected transition pattern;
- a dense but suspicious coordinated group;
- an abnormal graph substructure.

Worked Example: An Abnormal Login Burst

Normal login rate is $\lambda = 2$ attempts per minute. Under a Poisson model,

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

For 10 attempts in one minute:

$$P(X \geq 10) = 1 - \sum_{k=0}^{9} e^{-2} \frac{2^k}{k!} \approx 4.65 \times 10^{-5}.$$

Interpretation

Each login attempt looks ordinary, but the group is highly improbable. This is a collective outlier.

Window Length Controls Collective Detection

Window too short

- misses slow campaigns;
- sees fragments instead of a pattern;
- reacts quickly.

Window too long

- dilutes short bursts;
- increases delay;
- captures long-term coordination.

$$\text{granularity} + \text{window length} + \text{overlap} \longrightarrow \text{detectable collective pattern}$$

Practical response

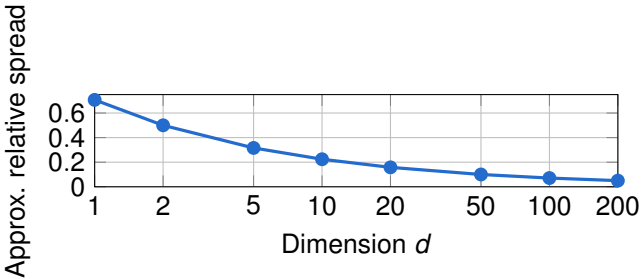
Evaluate multiple window lengths or use multiscale detectors when event duration is unknown.

11.7 High-Dimensional Outliers

Distance Concentrates as Dimension Grows

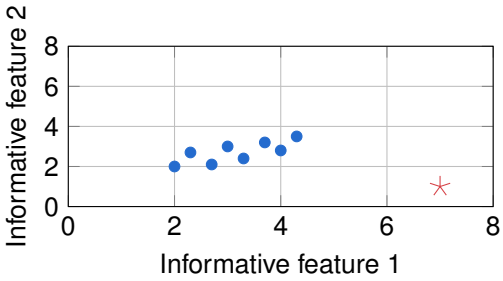
For independent standardized features, pairwise Euclidean distance becomes increasingly concentrated. An approximate relative spread is

$$\mathrm{CV}(d) \approx \frac{1}{\sqrt{2d}}.$$



Meaning: Similar nearest and farthest distances weaken global scores.

Outliers Can Hide in Small Subspaces



If 98 irrelevant features are added:

- full-space distance may hide the anomaly;
- the correct two-feature subspace still separates it.

Three Strategies for High Dimensions

1. **Extend a conventional detector:** use regularization, robust distance, or angle-based evidence.
2. **Find informative subspaces:** search feature subsets where the object is sparse or separated.
3. **Model high-dimensional structure:** use projections, factorization, ensembles, or learned representations.

$$s(\boldsymbol{x}) = \text{aggregate}_{m=1}^{M}\, s_m(\boldsymbol{x}_{S_m}),$$

where S_m is a feature subset.

Feature bagging

Run detectors on many random feature subsets and aggregate their rankings. An anomaly hidden in full space may become visible in some subsets.

Angle-Based Evidence in High Dimensions

For point $\boldsymbol{x}$ and pairs $\boldsymbol{a}, \boldsymbol{b}$, consider the angle

$$\theta_{\boldsymbol{axb}} = \cos^{-1} \frac{(\boldsymbol{a} - \boldsymbol{x})^\top (\boldsymbol{b} - \boldsymbol{x})}{\|\boldsymbol{a} - \boldsymbol{x}\| \, \|\boldsymbol{b} - \boldsymbol{x}\|}.$$

- A central point is viewed from many directions, producing varied angles.
- A far-away point sees most other points in a similar direction, producing low angle variance.

small angle variance $\Rightarrow$ strong outlier evidence.

Cost

Exact angle calculations are expensive, so sampling or approximation is often needed.

Reduction Must Preserve Anomaly Signals

Helpful

- remove verified irrelevant features;
- use robust PCA or task-aware embeddings;
- inspect reconstruction residuals;
- compare several subspaces.

Dangerous

- fit preprocessing on all data;
- retain components only by variance;
- assume a 2-D plot preserves every anomaly;
- use one arbitrary feature subset.

A low-variance direction may contain the exact signal that makes an object anomalous.

Evaluation and Practice

Scores Rank; Thresholds Create Alerts

A detector produces a score $s(\boldsymbol{x})$. An operational rule flags

$$s(\boldsymbol{x}) \geq \tau.$$

Choose τ from

- a clean-score quantile;
- a target false-positive rate;
- a review budget;
- expected cost or utility.

Monitor after deployment

- score distribution;
- alerts per day;
- review outcomes;
- feature and population drift.

Accuracy Fails When Outliers Are Rare

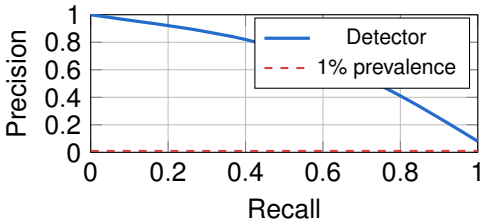
Suppose 10 of 10,000 transactions are fraudulent.

Rule	Accuracy	Recall	Precision
Predict all normal	99.90%	0%	undefined
Flag 20, including 8 frauds	99.88%	80%	40%

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}.$$

The slightly lower-accuracy rule is useful because it finds 8 frauds with only 20 reviews.

Precision-Recall Focuses on Rare Events



- PR curves expose false-alert burden.
- ROC curves can appear optimistic under extreme imbalance.
- Precision@k is useful when only k cases can be reviewed.

Cost-Based Thresholds Make Trade-Offs Explicit

Let a missed anomaly cost C_{FN} and an unnecessary review cost C_{FP}.

$$\text{ExpectedCost}(\tau) = C_{FN}\, FN(\tau) + C_{FP}\, FP(\tau).$$

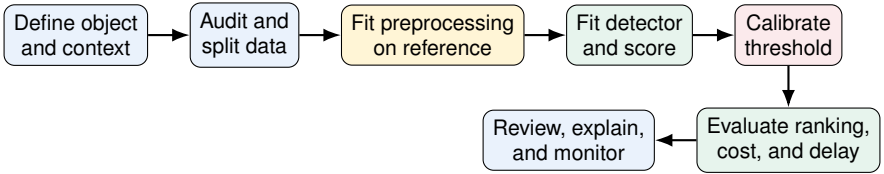
Example:

$$C_{FN} = 1000, \quad C_{FP} = 10.$$

Threshold	FN	FP	Expected cost
τ_1	8	20	$8(1000) + 20(10) = 8200$
τ_2	2	100	$2(1000) + 100(10) = 3000$

Even with more false alerts, τ_2 is better under these costs.

A Leakage-Safe Detection Workflow



Common leakage

Full-data scaling, future statistics, post-incident fields, and the same entity appearing in train and test can produce unrealistic performance.

Card Fraud Detection Case Study

Design choice	Decision
Unit	one transaction plus a recent-customer history window
Context	customer, hour, region, merchant category, device
Behavior	amount ratio, velocity, travel distance, merchant novelty
Baselines	robust contextual z-scores and k-NN distance
Complementary model	LOF for local density or a supervised classifier when labels mature
Split	train on the past; test on later time; prevent customer leakage
Metrics	precision@k, recall, monetary loss prevented, review effort
Explanation	unusual amount ratio, new device, distant location, nearest historical examples

Common Failures Have Direct Corrections

Failure	Better response
Delete every outlier	trace the cause and retain valid rare cases
Use raw Euclidean distance	scale features and choose a meaningful metric
Assume one normal population	model context, cohorts, and local density
Use a default contamination rate	calibrate from evidence, capacity, and cost
Evaluate with accuracy	use PR metrics, precision@k, recall, cost, and delay
Fit transformations on all data	fit only on the training or reference set
Treat a score as an explanation	show contributing features, neighbors, residuals, or cluster distance

Class Activity: Design the Detector

A university platform records login time, device, IP region, course, page, action, and session duration. The goal is to identify compromised accounts.

1. Choose the unit: event, session, day, or student profile.
2. Give one global, one contextual, and one collective outlier.
3. Propose five useful features.
4. Select two complementary detectors and justify them.
5. Describe a time-based split.
6. Choose two evaluation metrics and one threshold policy.

Practice Problems

1. For $\{4, 5, 5, 6, 6, 7, 8, 20\}$, calculate the IQR fences and identify any outlier.
2. For $\mu = (0, 0)$, $\Sigma = \mathrm{diag}(4, 1)$, and $\boldsymbol{x} = (4, 2)$, calculate D_M^2.
3. For $\{1, 2, 3, 4, 12\}$ and $k = 2$, calculate the average k-NN score for every point.
4. A point has lrd $= 0.4$ and neighbor densities $0.8, 0.6, 1.0$. Calculate LOF.
5. A PCA model reconstructs $\boldsymbol{x} = (5, 1)$ as $\hat{\boldsymbol{x}} = (3, 2)$. Calculate squared reconstruction error.
6. Explain why 30°C may be a contextual outlier but not a global outlier.

Answer Check

1. $Q_1 = 5$, $Q_3 = 7.5$, $IQR = 2.5$, fences 1.25 and 11.25; therefore 20 is flagged.
2. $D_M^2 = (4^2/4) + (2^2/1) = 8$.
3. Average 2-NN scores: $1.5,\ 1,\ 1,\ 1.5,\ 8.5$.
4. LOF $= \frac{1}{3}(0.8/0.4 + 0.6/0.4 + 1.0/0.4) = 2.0$.
5. $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 = (5-3)^2 + (1-2)^2 = 5$.
6. Its abnormality depends on season, location, or another contextual attribute.

Decision Map: Object-Level Evidence

Question	Method family
Is it improbable under a model?	statistical likelihood, z-score, Mahalanobis, KDE
Is it far from other objects?	distance-based and k-NN scores
Is its region locally sparse?	density-based detection such as LOF
Does it reconstruct poorly?	PCA, matrix factorization, robust PCA, autoencoder

Match the evidence to the assumption made by the detector.

Decision Map: Beyond Objects

Question	Method family
Does it fit a group or class?	clustering- and classification-based detection
Is it unusual only under context?	conditional model or contextual residual
Is a sequence or set abnormal?	collective window, subsequence, or graph score
Is it hidden by many features?	subspaces, feature bagging, angle or representation methods

Change the unit or subspace when object-level global scores cannot reveal the anomaly.

Key Takeaways

1. Outlierness is relative to a representation, reference population, context, and time.
2. Statistical, proximity, reconstruction, cluster, and classification methods encode different assumptions.
3. Global, contextual, and collective outliers require different units of analysis.
4. High-dimensional detection must identify informative subspaces or structures.
5. Scores rank unusualness; thresholds translate scores into operational decisions.
6. Evaluation must reflect rarity, false-alert cost, review capacity, and deployment time.

References

J. Han, J. Pei, and H. Tong,
Data Mining: Concepts and Techniques, 4th ed.,
Morgan Kaufmann, 2023, Chapter 11.

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander,
LOF: Identifying Density-Based Local Outliers,
Proc. ACM SIGMOD, 2000.

V. Chandola, A. Banerjee, and V. Kumar,
Anomaly Detection: A Survey,
ACM Computing Surveys, vol. 41, no. 3, 2009.

E. J. Candès, X. Li, Y. Ma, and J. Wright,
Robust Principal Component Analysis?
Journal of the ACM, vol. 58, no. 3, 2011.

H.-P. Kriegel, M. Schubert, and A. Zimek,
Angle-Based Outlier Detection in High-Dimensional Data,
Proc. ACM SIGKDD, 2008.

Recommended Reading

- Han, Pei, and Tong, Chapter 11, for the core taxonomy and methods.
- The authors' official Chapter 11 lecture slides for chapter-aligned review.
- LOF paper for local density and reachability distance.
- Chandola et al. for a broader anomaly-detection taxonomy.
- Robust PCA and angle-based detection papers for reconstruction and high-dimensional extensions.

Suggested study order

Definition $\rightarrow$ type of outlier $\rightarrow$ data setting $\rightarrow$ method assumption $\rightarrow$ score calculation $\rightarrow$ threshold and evaluation.