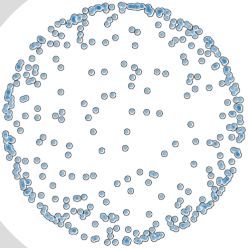


Lecture 4a | **Anomalies**

Outline

Motivations

'Classical' Anomaly Detection

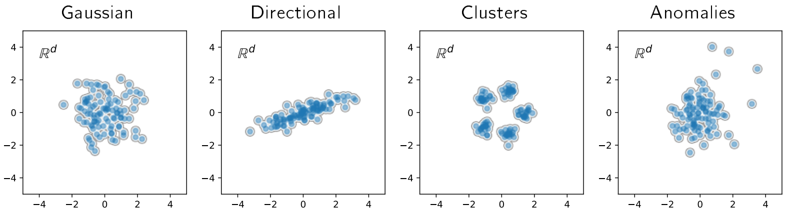
- Univariate Data and the Z-Score
- Multivariate Data and the Mahalanobis Distance

Boundary-Based Anomaly Detection

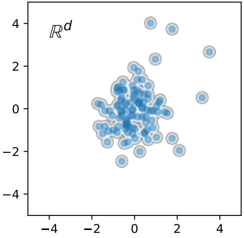
- Formulation as a Constrained Optimization Problem
- KKT Conditions and the Lagrange Dual
- Example: Adjusting a Threshold

Modeling Data Dispersion

Various types of Dispersion:



Anomalies:



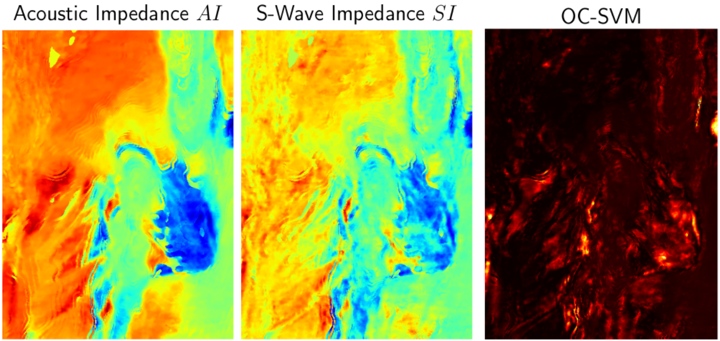
- Anomalies are by nature points that escape the models prediction capability.
- Hence, a model for anomaly prediction should focus and what is normal, and predict anomaly in opposition to what has been predicted to be normal.

Part 1

Motivating Examples

Motivating Examples

Analysis of geological data:



- Discovering anomalous/rare geological properties.
- Application to resources monitoring / resources extraction.

Motivating Examples

Removal of eye blink artifacts:

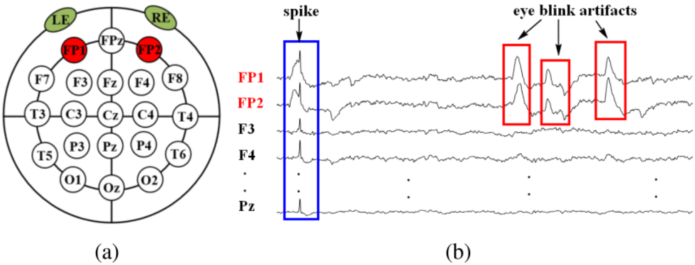


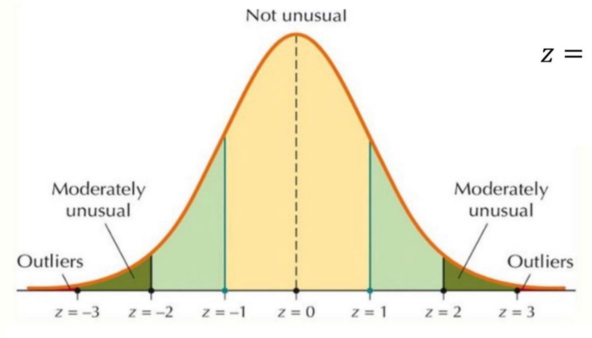
Image source: doi: 10.1109/JBHI.2021.3057891

- Eye blink anomalies are not interesting per se, but they can perturb the model and therefore need to be detected/removed.

Part 2

‘Classical’ Anomaly Detection

The Z-Score



Source: Analytics Vidhya

$$z = \frac{(x - \mu)}{\sigma}$$

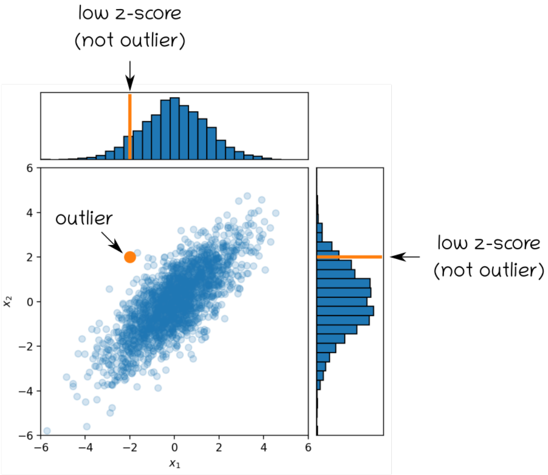
Data point

Mean

Standard deviation

- Very common measure of anomaly in statistical studies (e.g. mortality, temperatures, etc.).
- Points can be considered as outliers if their z-score is above a certain threshold, typically $|z| > 3$.

The Z-Score and Multivariate Data



Problem:

- When input features are correlated, z-score of individual features cannot properly detect outlierness.

The Mahalanobis Distance

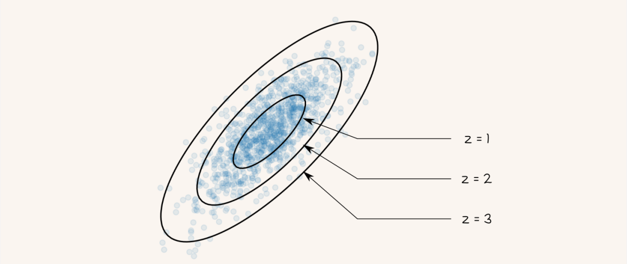
$x \in \mathbb{R}^d$ $\quad \Sigma \in \mathbb{R}^{d \times d}$

$\mu \in \mathbb{R}^d$

Definition:

- Generalization of the concept of z-score to multiple dimensions, i.e. how many standard deviations the model is away from the center of the data. It takes into account correlations in the data.
- The Mahalanobis distance of a point $\boldsymbol{x}$ to a reference distribution of mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is defined as:

$$z = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}$$

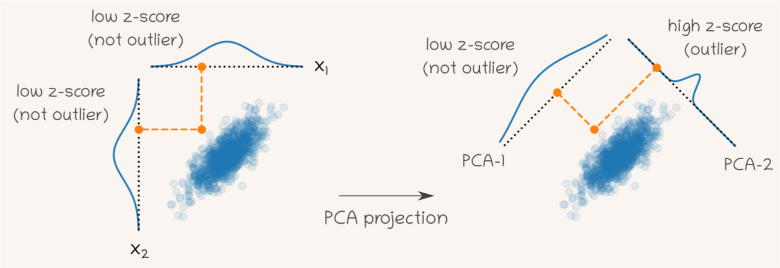


Relating Mahalanobis Distance and Z-Scores

Observation:

- The Mahalanobis distance can be related to the z-scores computed *in PCA space* by the formula:

$$z = \sqrt{\sum_{j=1}^{d} (z_{\text{PCA-}j})^2}$$



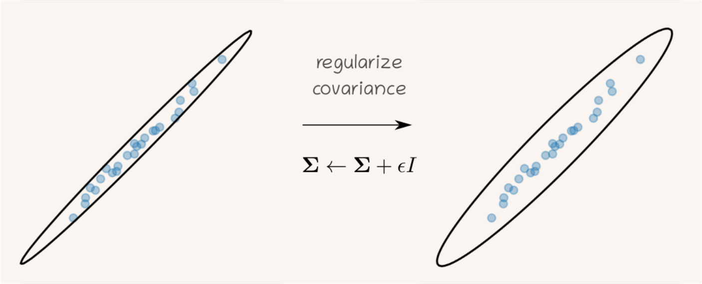
Limits of Mahalanobis Distance

Problem:

- In high dimensions, the matrix $\boldsymbol{\Sigma}^{-1}$ tends to become uncontrollably large, due to correlations that arise from the limited data. The Mahalanobis distance typically produces inflated anomaly scores.

Remedy:

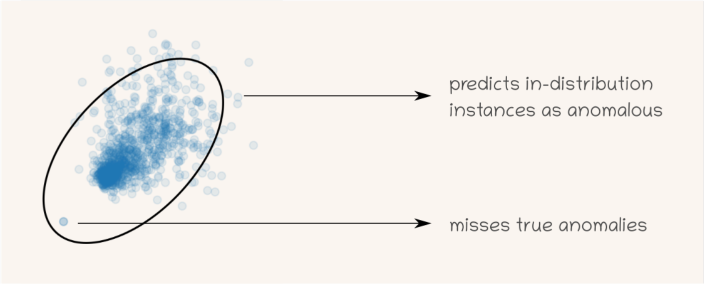
- This instability can be overcome by adding a small diagonal term to the covariance matrix. This modification of the covariance matrix makes the anomaly score more robust.



Limits of Mahalanobis Distance

Observation:

- In presence of skewed, non-Gaussian, distributions, the Malahanobis distance does not describe well the data.



Solution:

- No easy remedy. We need a different principle for anomaly detection that places the focus on the actual boundary between anomalous and non-anomalous data.

Part 3

Boundary-Based Anomaly Detection

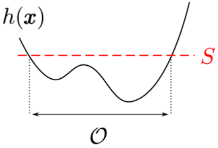
Boundary-Based Anomaly Detection

Idea:

- Learn a geometrical object

$$\mathcal{O}(S, \boldsymbol{c}) = \{\boldsymbol{x} \in \mathbb{R}^d : h(\boldsymbol{x}; \boldsymbol{c}) \leq S\}$$

  e.g. a hypersphere, a polytope, etc., that encloses most the data (except outliers).



Observation:

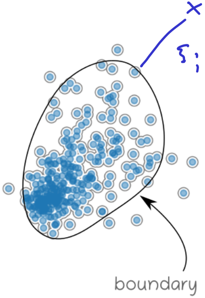
- The problem of finding a minimum enclosing object can be stated as the constrained optimization problem:

$$\min_{\underbrace{S, \boldsymbol{c}, \boldsymbol{\xi}}_{\boldsymbol{\theta}}} \; S + C \sum_{i=1}^{N} \xi_i$$

s.t.

$$\forall_{i=1}^{N} : h(\boldsymbol{x}_i; \boldsymbol{c}) \leq S + \xi_i$$

$$\forall_{i=1}^{N} : 0 \leq \xi_i$$



- Optimization problems with inequality constraints can be studied within the framework of KKT conditions.

Some Theory: KKT Conditions

Consider the constrained optimization problem

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \quad \text{s.t.} \quad \forall_{i=1}^{M} : \; g_i(\boldsymbol{\theta}) \leq 0 \tag{1}$$

and define the Lagrange function

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = f(\boldsymbol{\theta}) + \sum_{i=1}^{M} \lambda_i g_i(\boldsymbol{\theta}).$$

KKT Conditions

If $\boldsymbol{\theta}^\star$ is a solution of (1) and the optimization problem satisfies some *regularity conditions (e.g. objective convex, all constraints convex, and existence a point $\boldsymbol{\theta}$ that satisfies them with strict inequalities)*, then there exists a constant vector $\boldsymbol{\lambda}$ such that the solution necessarily satisfies

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^\star, \boldsymbol{\lambda}) &= 0 && \text{(stationarity)} \\ \forall_{i=1}^{M} : g_i(\boldsymbol{\theta}^\star) &\leq 0 && \text{(primal feasibility)} \\ \forall_{i=1}^{M} : \lambda_i &\geq 0 && \text{(dual feasibility)} \\ \forall_{i=1}^{M} : \lambda_i g_i(\boldsymbol{\theta}^\star) &= 0 && \text{(complementary slackness)} \end{aligned}$$

Some Theory: Primal vs. Dual

KKT conditions provide a set of equations which a solution $\boldsymbol{\theta}^\star$ needs to satisfy. However, these equations can typically not be solved analytically and one needs an optimization procedure. There are two approaches:

Primal

- Solve the optimization problem

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \quad \text{s.t.} \quad \forall_{i=1}^M : \; g_i(\boldsymbol{\theta}) \leq 0.$$

Lagrange Dual

- Solve the optimization problem

$$\max_{\boldsymbol{\lambda} \succeq \mathbf{0}} \inf_{\boldsymbol{\theta}} \Big\{ f(\boldsymbol{\theta}) + \sum_{i=1}^M \lambda_i \, g_i(\boldsymbol{\theta}) \Big\}$$

- Intuition: penalize constraint violations directly into the objective.
- The primal parameters $\boldsymbol{\theta}^\star$ can be recoved from the dual solution $\boldsymbol{\lambda}^\star$ using KKT conditions.

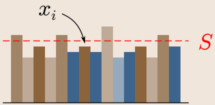
Example: Adjusting a Threshold

$$\min_{S,\boldsymbol{\xi}} \quad S + C\sum_{i=1}^{N} \xi_i \qquad \text{s.t.} \quad \forall_{i=1}^{N} : x_i \leq S + \xi_i$$

$$x_i - S - \xi_i \leq 0$$

$$\forall_{i=1}^{N} : \xi_i \geq 0 \qquad -\xi_i \leq 0$$

$$S \geq 0 \qquad -S \leq 0$$



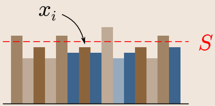
Step 1: Apply the KKT Conditions (and simplify them)

$$\mathcal{L}(S,\xi) = S + C\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i (x_i - S - \xi_i) + \sum_{i=1}^{N} \beta_i (-\xi_i) + \gamma \cdot (-S)$$

$$\frac{\partial \mathcal{L}}{\partial S} = 1 - \sum_{i=1}^{N} \alpha_i - \gamma \overset{\text{def}}{=} 0 \quad \Rightarrow \quad \boxed{\sum_{i=1}^{N} \alpha_i = (1-\gamma)}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \beta_i \overset{\text{def}}{=} 0 \quad \Rightarrow \quad \boxed{\alpha_i \leq C}$$

Example: Adjusting a Threshold



Step 2: Derive the Lagrange Dual

$$\mathcal{L}(S,\xi) = \cancel{S} + C\sum_{i=1}^{N}\cancel{\xi_i} + \sum_{i=1}^{N}\alpha_i(x_i - \cancel{S} - \cancel{\xi_i}) + \sum_{i=1}^{N}\beta_i(-\cancel{\xi_i}) + \gamma\cdot(-\cancel{S})$$

$$= \sum_{i=1}^{N}\alpha_i x_i$$

$$\sum_{i=1}^{N}\alpha_i = (1-\gamma)$$

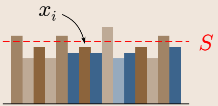
$$C - \alpha_i - \beta_i = 0$$

$$\max_{\alpha,\gamma} \sum_{i=1}^{N}\alpha_i x_i$$

$$0 \leq \alpha_i \leq C$$

$$\gamma \geq 0$$

Example: Adjusting a Threshold



Step 3: Recover Original Parameters from the Dual

Dual gives a solution α, γ

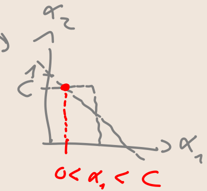
Recall the complementary slackness constraints in KKT:

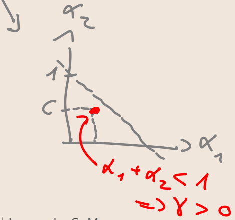
$$\forall i: \quad \alpha_i(x_i - S - \xi_i) = 0 \qquad (\alpha_i - C)(-\xi_i) = 0 \qquad \gamma \cdot (-S) = 0$$

The following cases enable to recover a solution of the primal:

if $\exists k: \; 0 < \alpha_k < C$: then $S = x_k$

if $\gamma > 0$: then $S = 0$





Summary

Summary

- Anomalies are a common property of data distributions. Anomalies can be interesting on their own (e.g. rare events), or some disturbance (e.g. an erroneous sensor measurement).
- Simple approaches (e.g. **z-score**, and its multi-dimensional generalization, the **Mahalanobis distance**) are effective and easy to put to practice (no machine learning involved except estimating parameters). However, they do not work well e.g. when the data distribution exhibit skewness.
- For more effective modeling of anomalies, it would be better to have an algorithm that focuses on the boundary of the data distribution, such as learning a **minimum enclosing** object of the data. This can be formulated as a constrained optimization problem.
- Constrained optimization problems, in particular, convex ones, come with a lot of theory (**KKT conditions**, **Lagrange Primal/Dual**) that allows us to gain a better understanding of the problem.