Computational Data Analysis

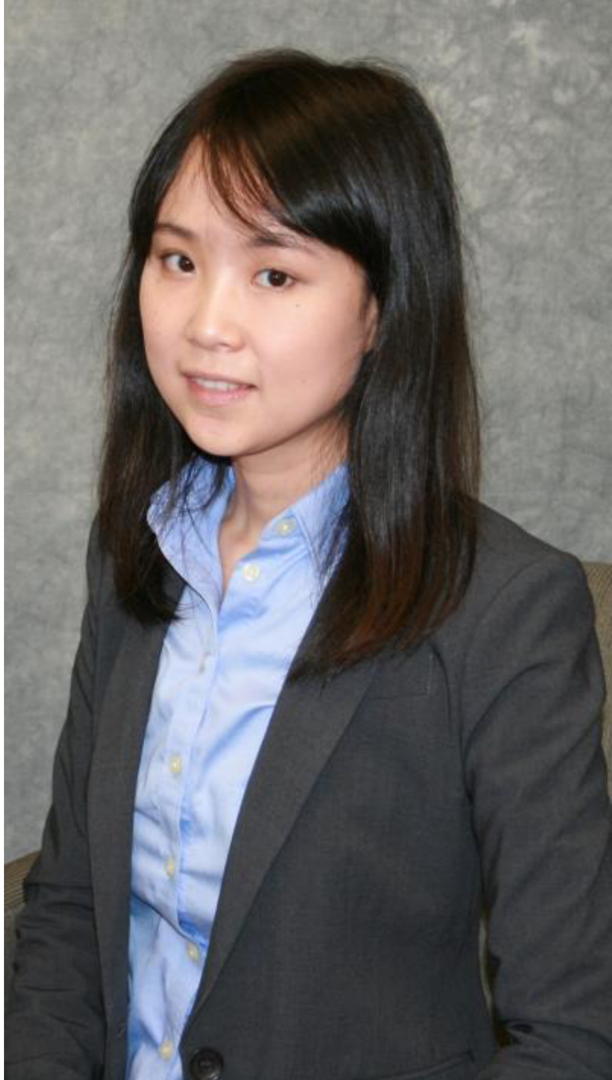
Machine Learning

Yao Xie, Ph.D.

Associate Professor

Harold R. and Mary Anne Nash Early Career Professor

H. Milton Stewart School of Industrial and Systems Engineering



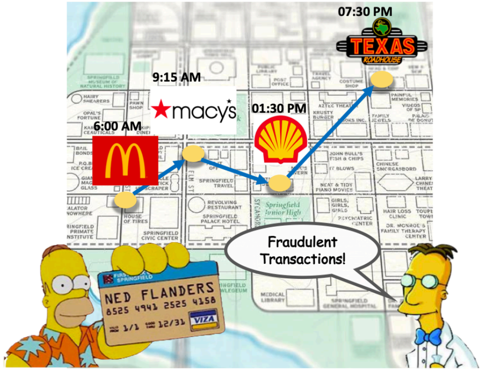
Anomaly detection

Motivation

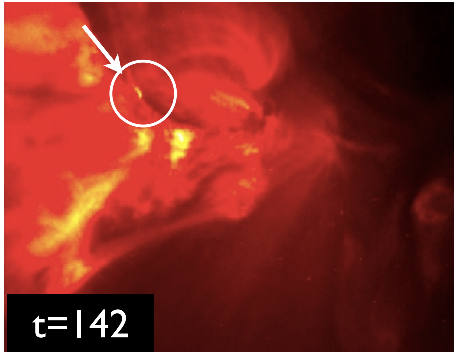
Power network monitoring

Northeast blackout, 2003

Credit card fraud detection

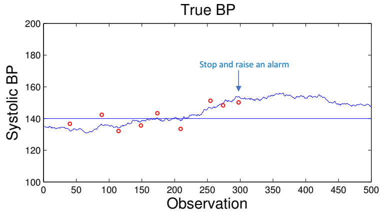


Solar flare detection

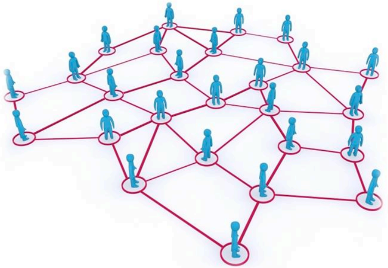


Blood pressure monitoring

Omron Project Zero 2.0.

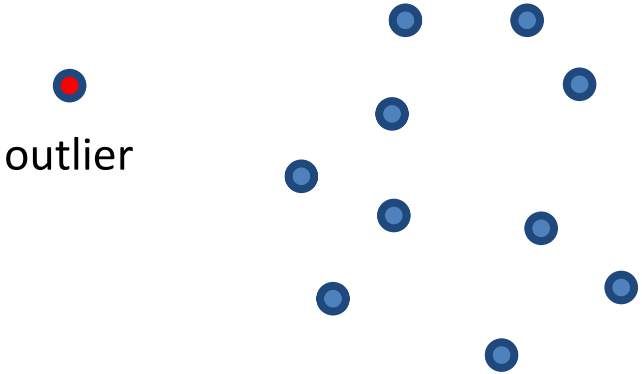


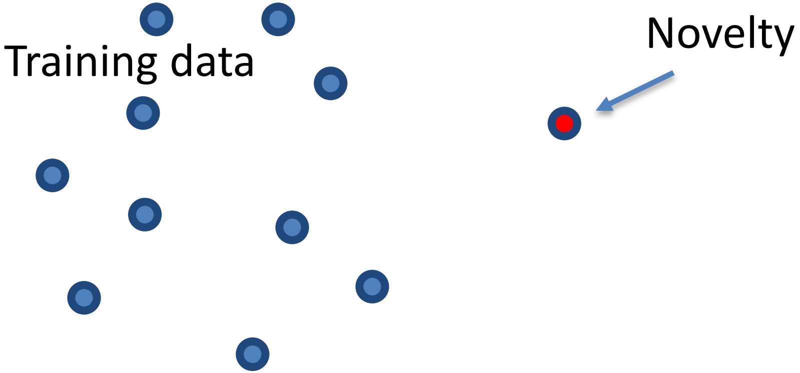
Event detection from social networks



Anomaly detection

- Outlier detection: identify outliers in training data, which are defined as observations that are "far" from the others in a chosen metric. --- As pre-processing step to clean data.
- Novelty detection: Detecting whether a **new** observation is an outlier. Related to change point detection (e.g., detecting change-points in time series).



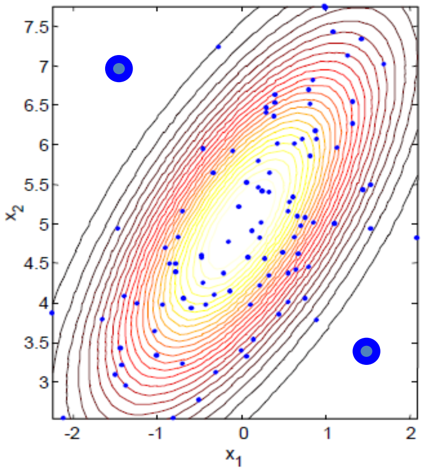


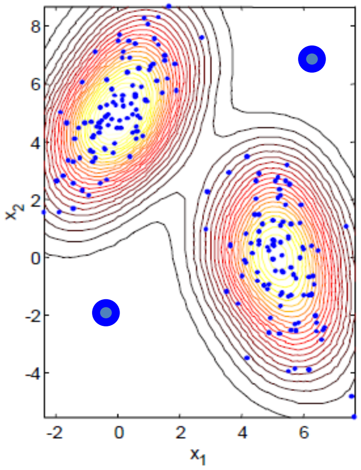
Anomaly detection

- Statistical methods: likelihood based
- Geometric methods: one-class SVM
- Some theory

Outlier detection: density-based

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





Covariance based detector

- Multivariate Gaussian

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

- Find parameter estimation from data

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

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

$$(x_{new} - \hat{\mu})^T \hat{\Sigma}^{-1}(x_{new} - \hat{\mu}) \geq b$$

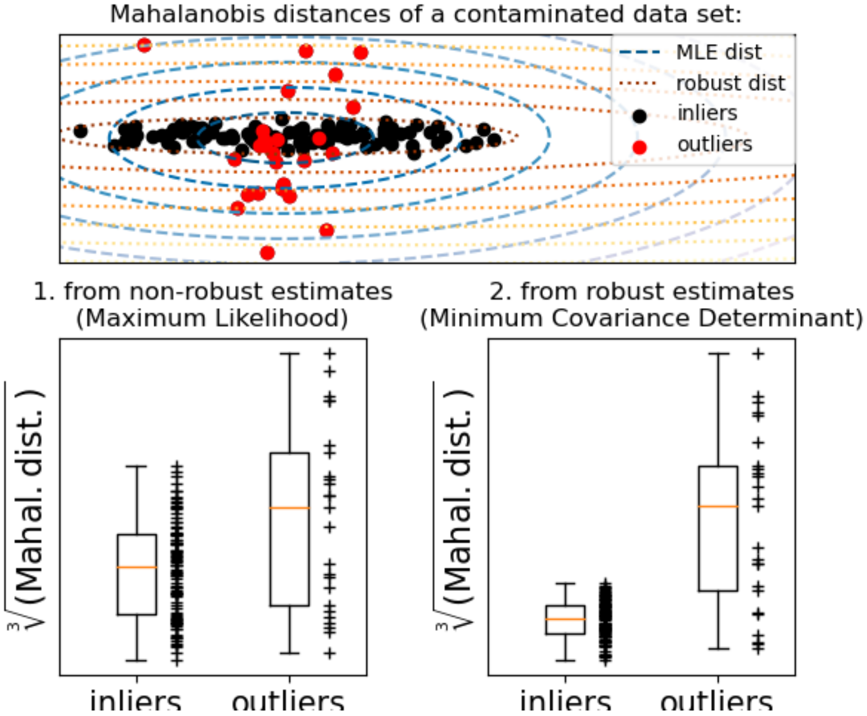
Mahalanobis distance

for some threshold b

Robust covariance estimation

- Mahalanobis distance can be affected by outliers
- Minimum Covariance Determinant estimator (MCD):

find a subset of observations whose empirical covariance has the smallest determinant $|\hat{\Sigma}|$

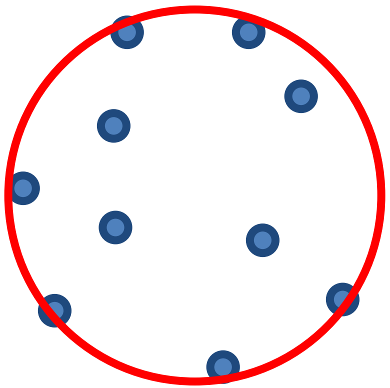


Anomaly detection

- Statistical methods: likelihood based
- Geometric methods: one-class SVM
- Some theory

Geometric method for anomaly detector

Rather than first fit density and then find threshold, directly find the boundary



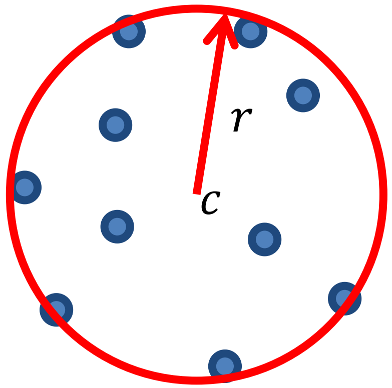
SVM based one-class classification (OCC)

Find the smallest ball such that it include all training data points.

Parameterized the ball with a center c and radius r

Optimization problem

$$\min_{r,c} \quad r^2$$
$$s.t. ||x^i - c||^2 \leq r^2,$$
$$r \geq 0$$
$$\forall i = 1, \ldots, m$$



Estimating the support of a high-dimensional distribution Schölkopf, Bernhard, et al. Neural computation 13.7 (2001): 1443-1471.

Solving one-class SVM

Lagrangian function

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

Set partial derivative of L to 0

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

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

Expressing Lagrangian function L in α_i

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

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

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

Can be solved efficiently

Optimal radius can be

found based on c^*

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

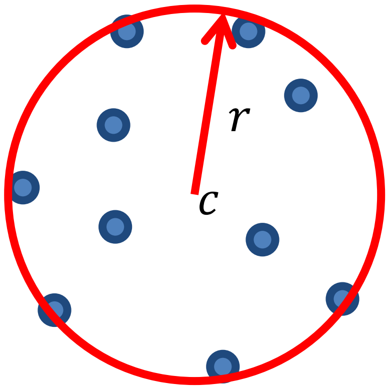
Sparsity in α_i

Complementary slackness (for all i):

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

Interesting observation:

- Many data points will be inside the circle
  - $\left(x^i - c\right)^\top \left(x^i - c\right) - r^2 < 0$
  - $\alpha_i = 0$
- Very few data point will be exactly on the boundary
  - $\left(x^i - c\right)^\top \left(x^i - c\right) - r^2 = 0$
  - α_i can be nonzero



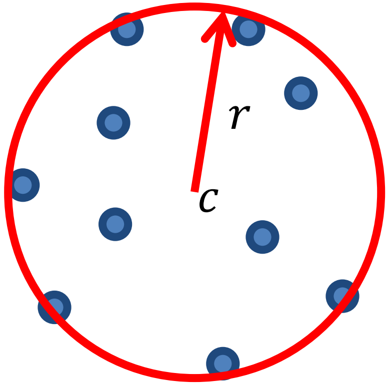
Solution in α is very sparse!

Support vectors

- Each α_i is associated with a one data point

- Data points with nonzero α_i are called support vectors. They determine the center and radius

$$c^* = \sum_{i=1}^{m} \alpha_i x^i$$
$$r = \min_{i:\alpha_i > 0} ||x^i - c^*||$$



- Data points with zero α_i is not affect c and r

Kernalize the algorithm

Dual problem

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

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

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

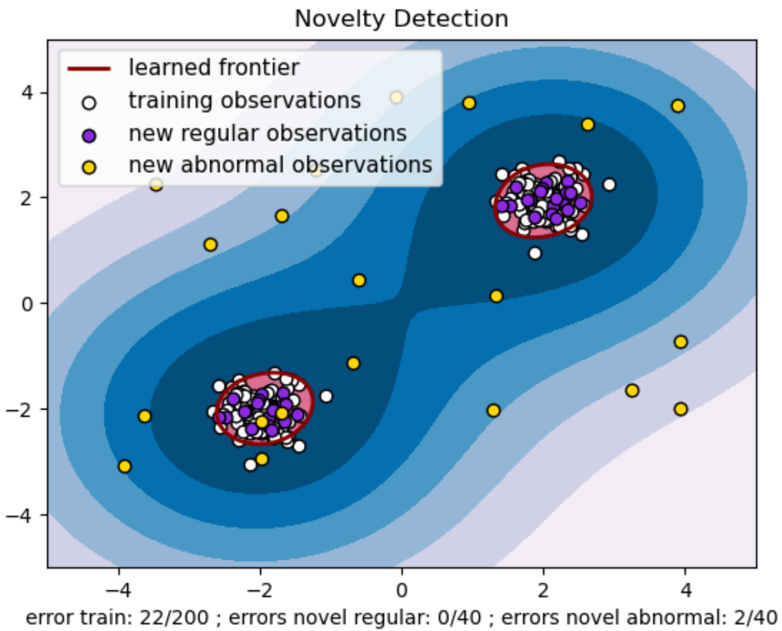
$b_i = {x^i}^\top x^i$

α

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

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

Demo: Novelty detection

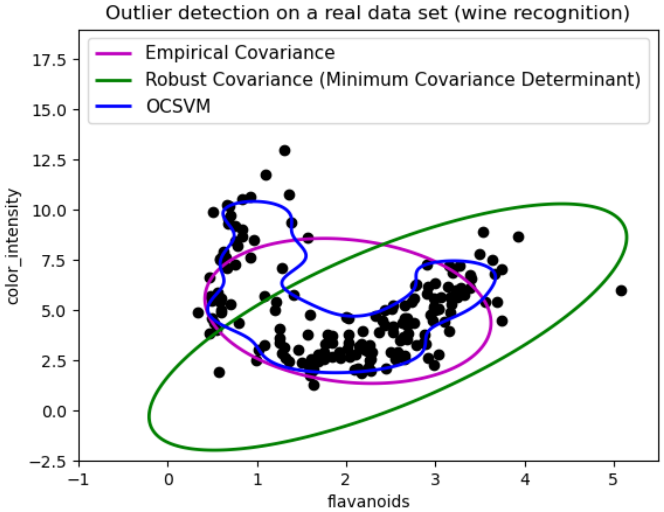


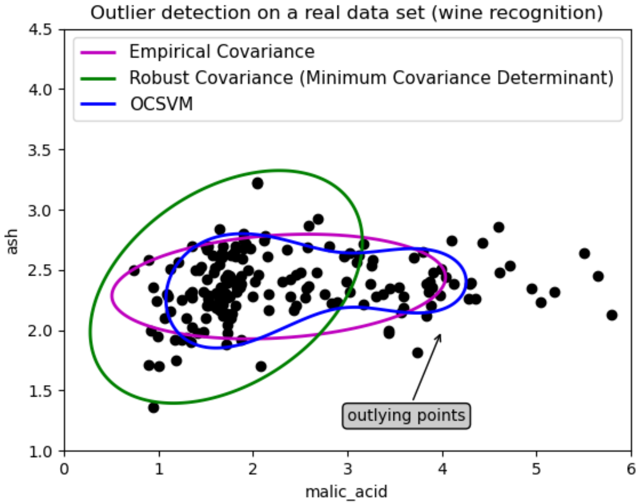
https://scikit-learn.org/stable/auto_examples/svm/plot_oneclass.html#sphx-glr-auto-examples-svm-plot-oneclass-py

Demo: Wine data

- The wine data set was introduced by Forina et al. (1986).
- It originally included the results of 27 chemical measurements on 178 wines made in the same region in Italy but derived from three different cultivars: Barolo, Grignolino and Barbera.

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





Anomaly detection

- Statistical methods: likelihood based
- Geometric methods: one-class SVM
- Some theory

Hypothesis test

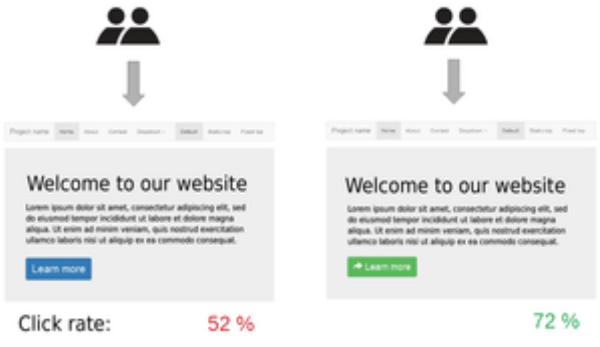
- Common hypothesis test

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

- Anomaly detection

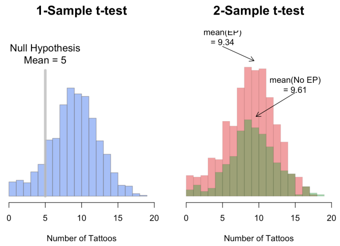
$$H_0: x^i \sim f_0, \qquad i = 1, \dots, m$$
$$H_1: x^i \nsim f_0, \qquad i = 1, \dots, m$$

- A/B testing



- Two-sample test

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

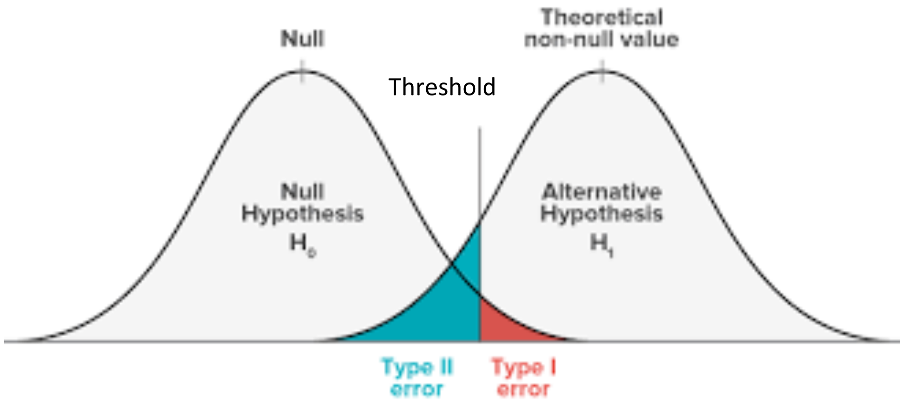


Performance metrics

Decision making: Choose H_0 or H_1

Performance metrics

- Type-I error = $P_{H_0}\{\text{reject } H\_0\}$
- Type-II error = $P_{H_1}\{\text{accept } H\_0\}$



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

Likelihood ratio test

- The likelihood ratio test is optimal when f_0 and f_1 are known exactly (Neyman-Pearson lemma)

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

- In practice, we can estimate the distributions
- Set the threshold b to control false alarms (Type-I error)

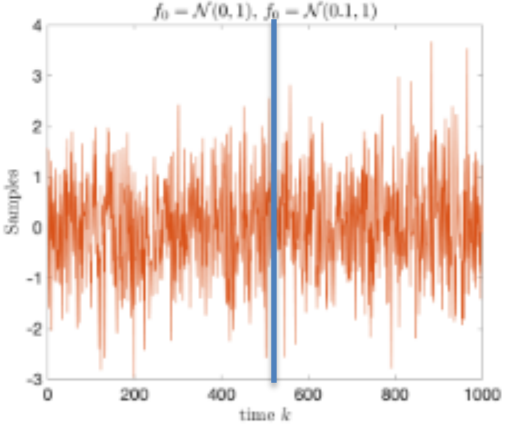
Novelty detection for sequential data

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

CUSUM procedure (cumulative sum over log-likelihood ratios)

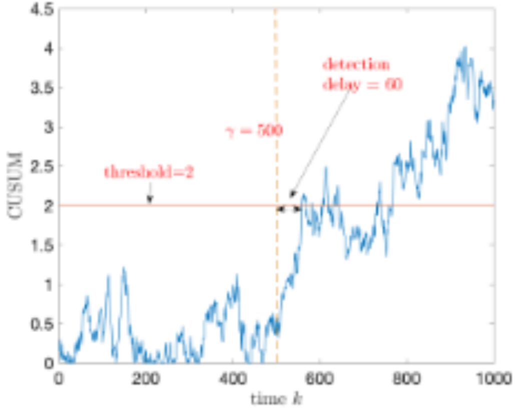
$$W_0 = 0, \qquad W_t = \max\left(W_{t-1} + \log\frac{f_1(X_t)}{f_0(X_t)}, 0\right)$$

Sequential data



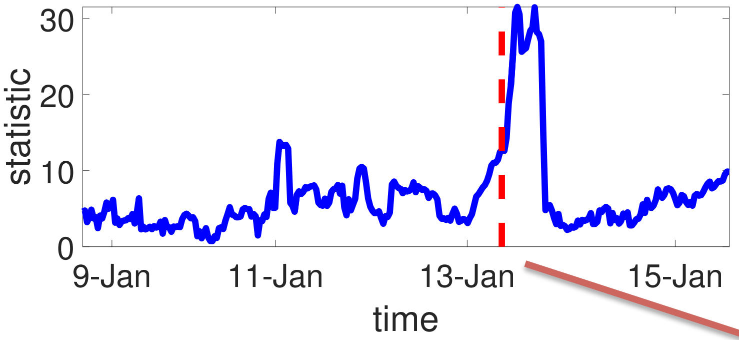
Change-point

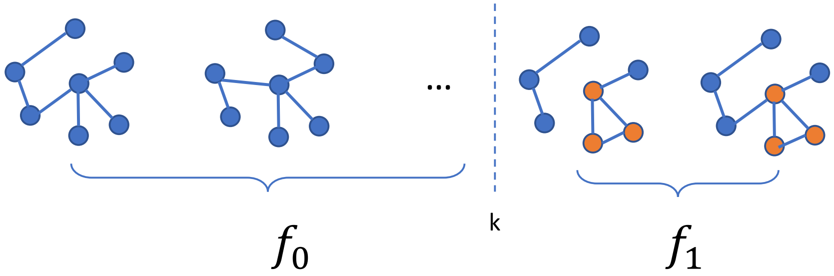
CUSUM statistics

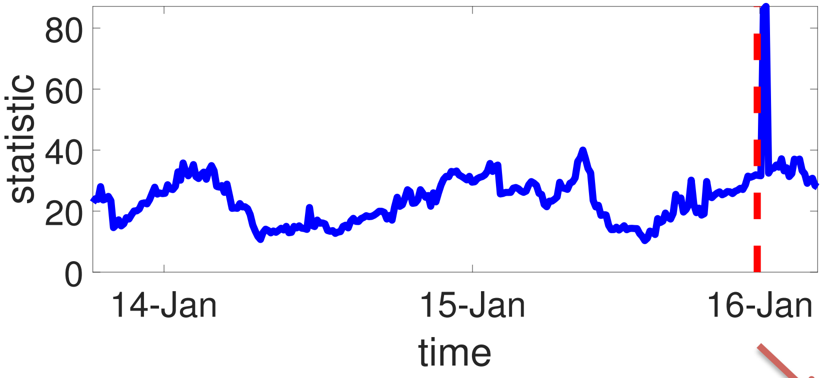


Example: Social network change detection

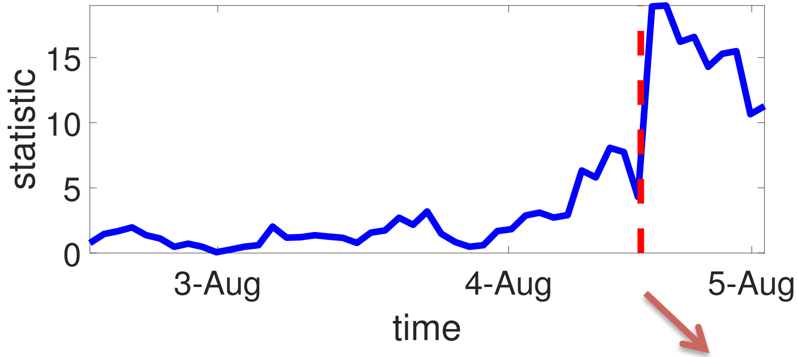
Detecting change in random graph, using twitter data







Israel announces ceasefire in Gaza War in 2009v



Beijing Olympic opening in 2008

Other anomaly detection algorithms

- Local outlier factor (LOF): distance metric based
- Neural network based: e.g., anomalous sequence detection using RNN

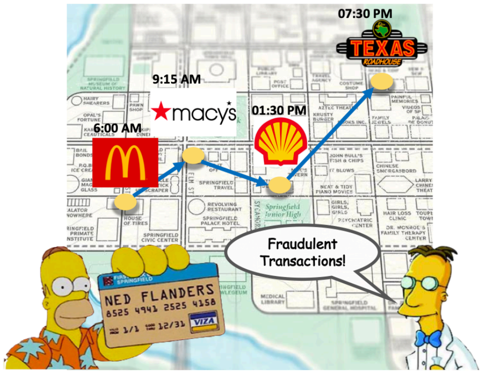
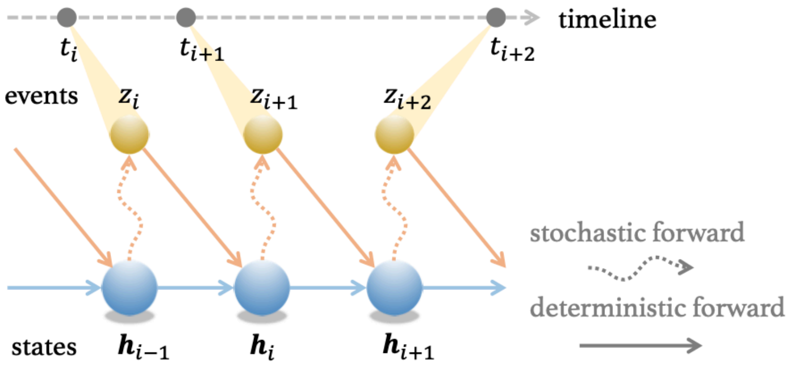


Fig. 1: An example of a sequence of anomalous events that are dependent: one leads to another.

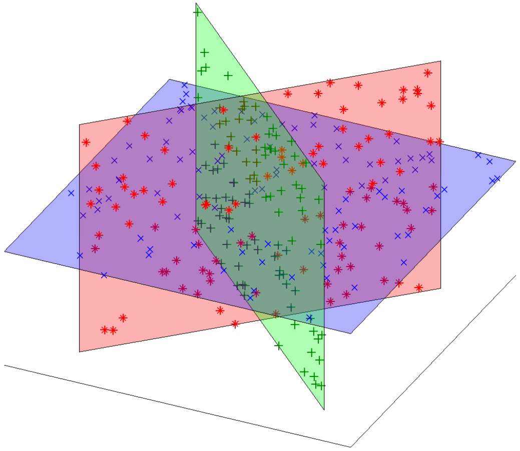


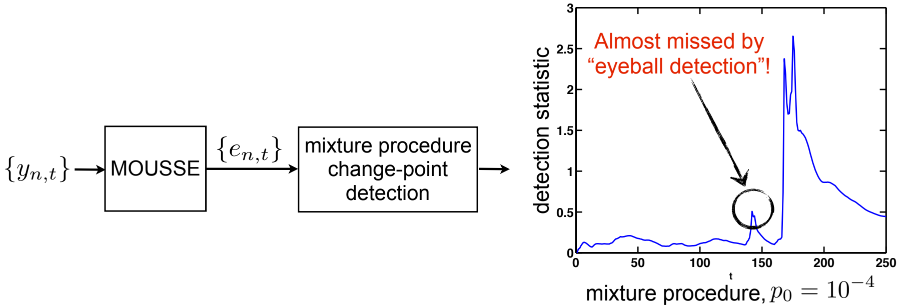
RNN-based anomalous seq. detection

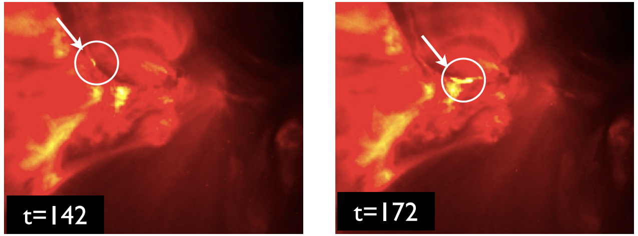
Other anomaly detection algorithms (cont.)

- Subspace based anomaly detection

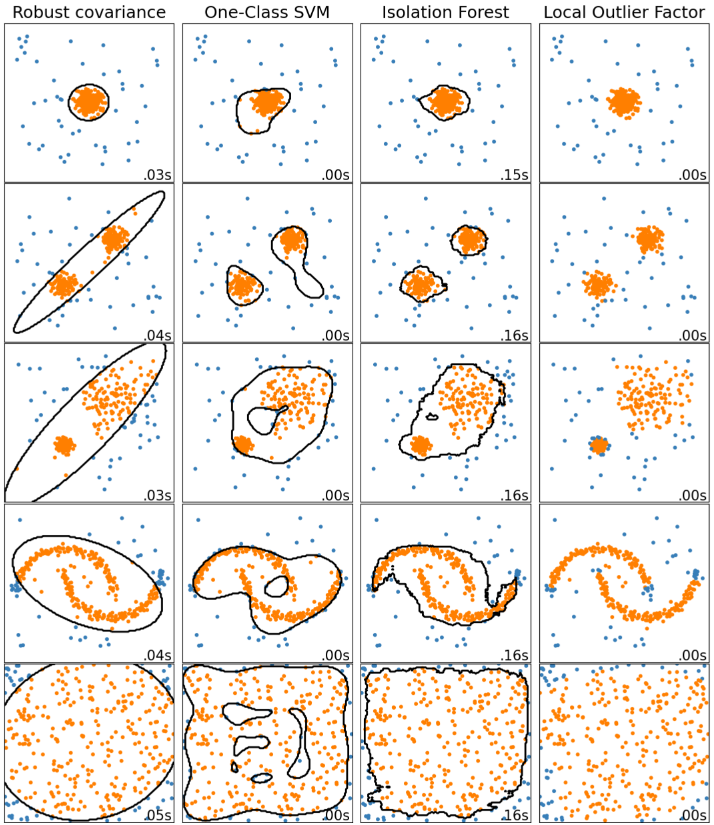
Represent normal data using subspace







Comparison of algorithms



https://scikit-learn.org/stable/auto_examples/miscellaneous/plot_anomaly_comparison.html#sphx-glr-auto-examples-miscellaneous-plot-anomaly-comparison-py


SUMMARY