



Università
degli Studi
di Palermo



Outlier Analysis

CORSO DI BIG DATA – MODULO ANALISI PER I BIG DATA

a.a. 2023/2024

Prof. Roberto Pirrone





Outline

- **Introduction**
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- Summary



Introduction (1)

□ A Quote

“You are unique, and if that is not fulfilled, then something has been lost.”—Martha Graham

□ An Informal Definition

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.”

□ A Complementary Concept to Clustering

- Clustering attempts to determine groups of data points that are **similar**
- Outliers are individual data points that are **different** from the remaining data



Introduction (2)

□ Applications

- Data cleaning

- ✓ Remove noise in data

- Credit card fraud

- ✓ Unusual patterns of credit card activity

- Network intrusion detection

- ✓ Unusual records/changes in network traffic



Introduction (3)

□ The Key Idea

- Create a model of **normal** patterns
- Outliers are data points that **do not naturally fit** within this normal model
- The “outlierness” of a data point is quantified by a **outlier score**

□ Outputs of Outlier Detection Algorithms

- Real-valued outlier score
- Binary label



Outline

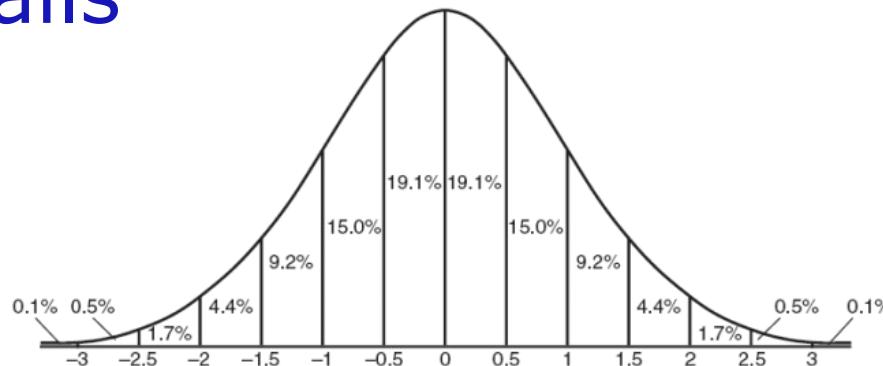
- Introduction
- **Extreme Value Analysis**
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- Summary



Extreme Value Analysis (1)

□ Statistical Tails

<http://www.regentsprep.org/regents/math/algtrig/ats2/normallesson.htm>

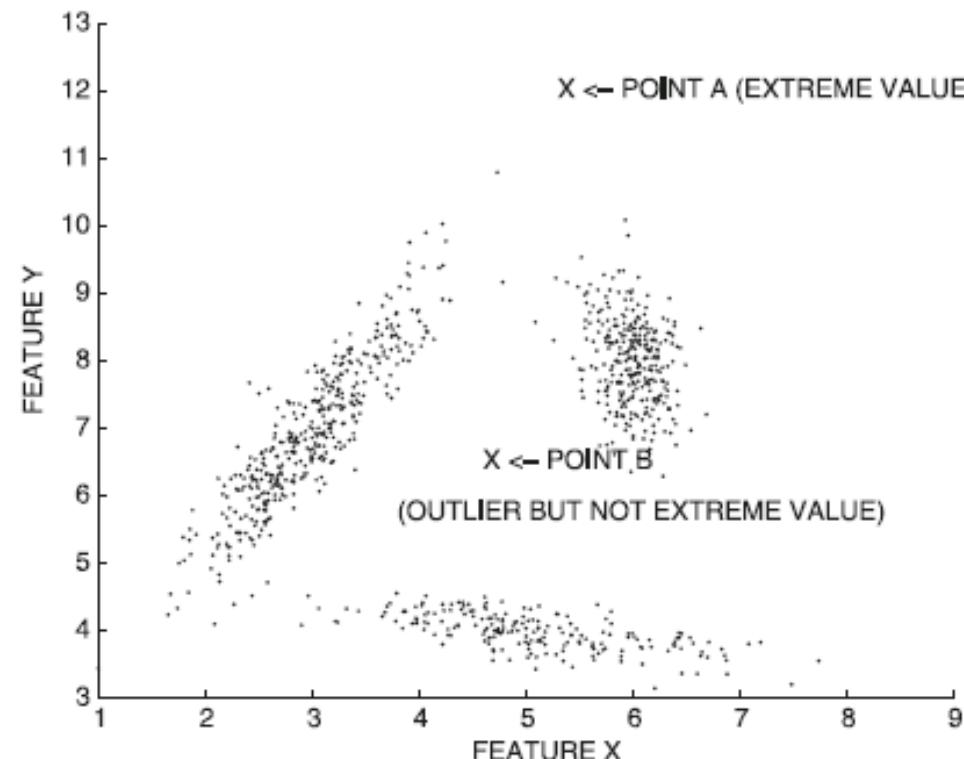


- All extreme values are outliers
- Outliers may not be extreme values
 - {1,3,3,3,50,97,97,97,100}
 - 1 and 100 are extreme values
 - 50 is an outlier but not extreme value



Extreme Value Analysis (2)

- All extreme values are outliers
- Outliers may not be extreme values





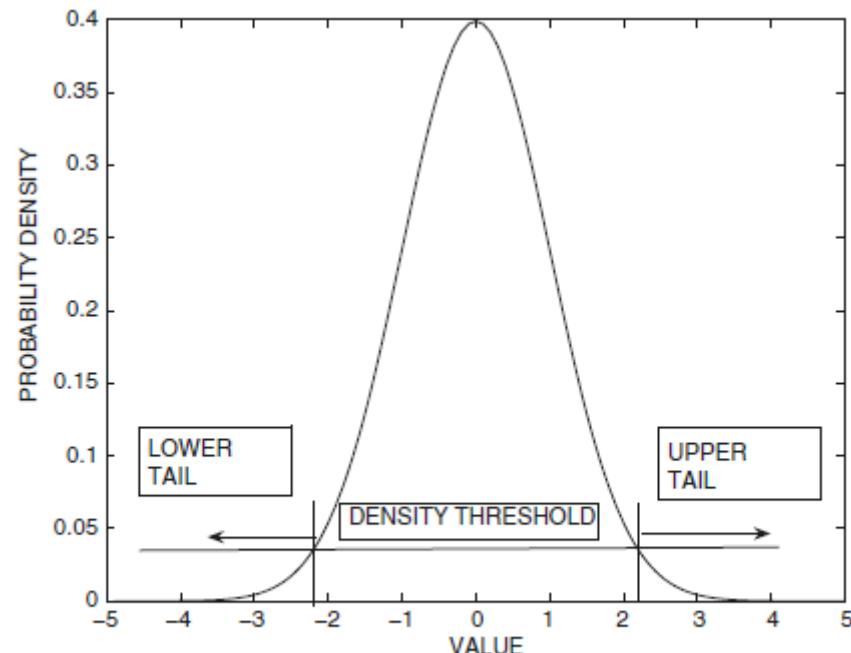
Univariate Extreme Value Analysis (1)

□ Statistical Tail Confidence Tests

- Suppose the density distribution is $f_X(x)$
- Tails are **extreme** regions s.t. $f_X(x) \leq \theta$

□ Symmetric Distribution

- Two symmetric tails
- The areas inside tails represent the cumulative probability



(a) Symmetric distribution



Univariate Extreme Value Analysis (2)

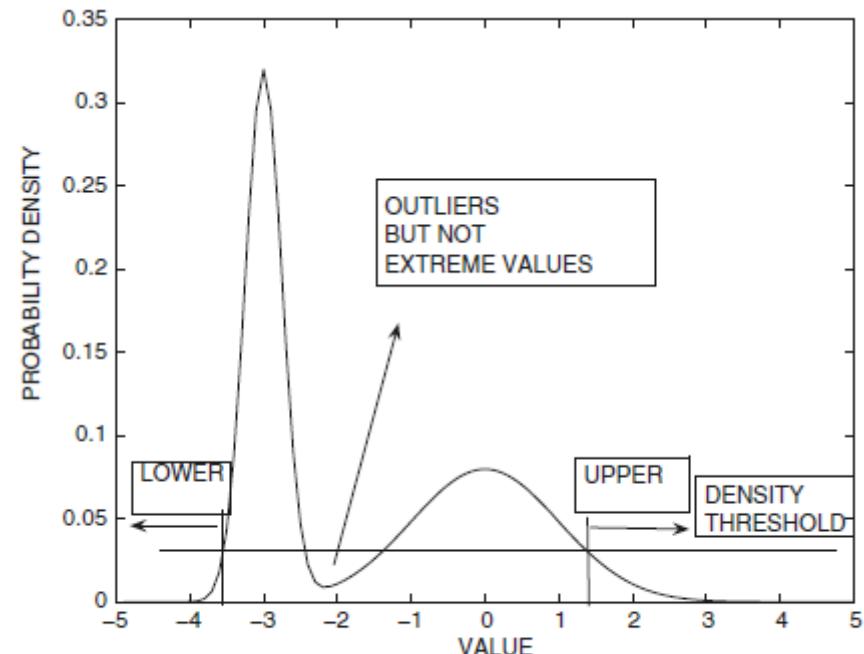
□ Statistical Tail Confidence Tests

- Suppose the density distribution is $f_X(x)$
- Tails are **extreme** regions s.t. $f_X(x) \leq \theta$

□ Asymmetric Distribution

- Areas in two tails are different
- Regions in the interior are not tails

E' necessario trovare un modello per quantificare la probabilità delle code



(b) Asymmetric distribution



The Procedure (1)

- A model distribution is selected
 - Normal Distribution with mean μ and standard deviation σ

$$f_X(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\frac{-(x-\mu)^2}{2 \cdot \sigma^2}}$$

- Parameter Selection
 - Prior domain knowledge
 - Estimate from data

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

The Procedure (2)

□ Z-value of a random variable

$$z_i = \frac{x_i - \mu}{\sigma}$$

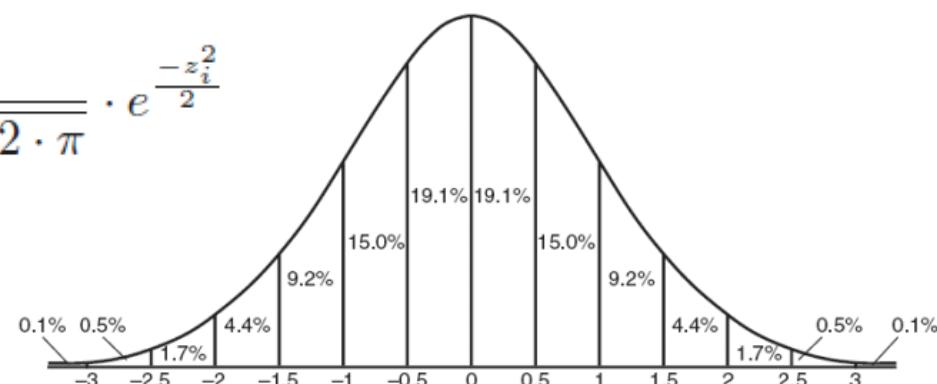
- Large positive values of z_i correspond to the upper tail
- Large negative values of z_i correspond to the lower tail
- z_i follows the **standard normal distribution**

$$f_X(z_i) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\frac{-z_i^2}{2}}$$

□ Extreme values

- $|z_i| \geq \tau$

$|z_i| \geq 3$ garantisce che i punti selezionati siano outliers





Multivariate Extreme Values (1)

- Unimodal probability distributions with a single peak
 - Suppose the density distribution is $f_X(x)$
 - Tails are **extreme** regions s.t. $f_X(x) \leq \theta$
- Multivariate Gaussian Distribution

$$\begin{aligned}f(\bar{X}) &= \frac{1}{\sqrt{|\Sigma|} \cdot (2 \cdot \pi)^{(d/2)}} \cdot e^{-\frac{1}{2} \cdot (\bar{X} - \bar{\mu}) \Sigma^{-1} (\bar{X} - \bar{\mu})^T} \\&= \frac{1}{\sqrt{|\Sigma|} \cdot (2 \cdot \pi)^{(d/2)}} \cdot e^{-\frac{1}{2} \cdot Maha(\bar{X}, \bar{\mu}, \Sigma)^2}\end{aligned}$$

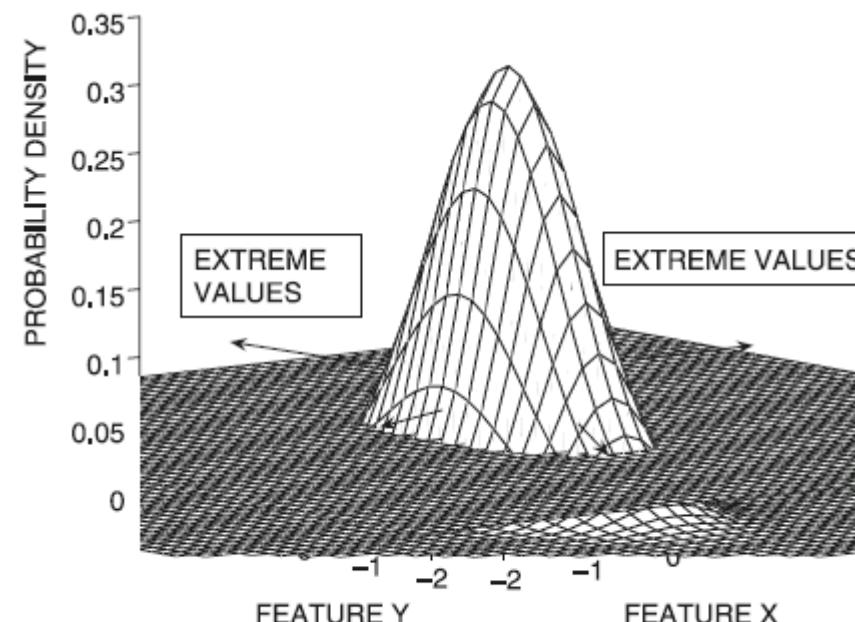
where $Maha(\bar{X}, \bar{\mu}, \Sigma)$ is the Mahalanobis distance between \bar{X} and $\bar{\mu}$



Multivariate Extreme Values (2)

□ Extreme-value Score of \bar{X}

- $Maha(\bar{X}, \bar{\mu}, \Sigma)$
- Larger values imply more extreme behavior



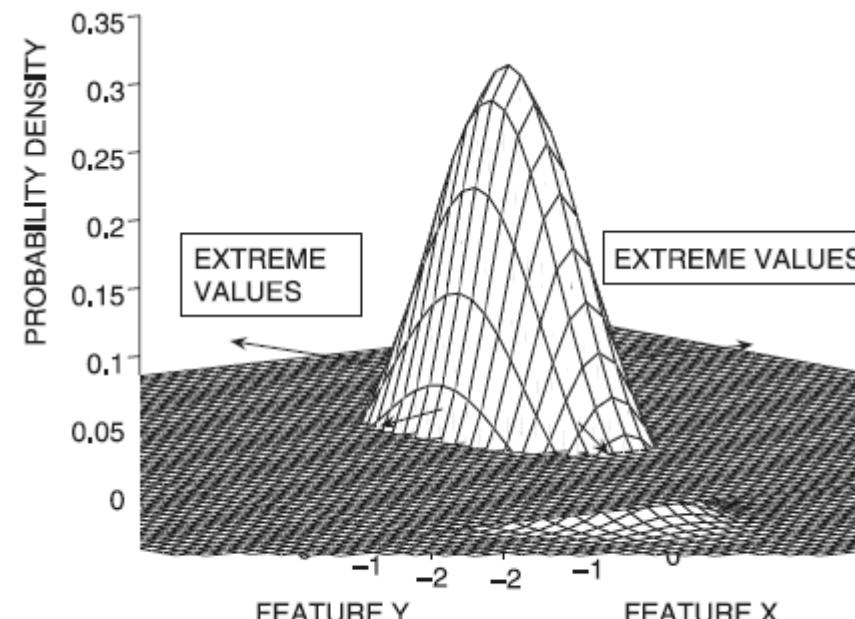
(b) Multivariate extreme values

Multivariate Extreme Values (2)

□ Extreme-value Score of \bar{X}

- $Maha(\bar{X}, \bar{\mu}, \Sigma)$
- Larger values imply more extreme behavior

La distanza di Mahalanobis diventa quella Euclidea se il dataset è riferito alle sue componenti principali



(b) Multivariate extreme values

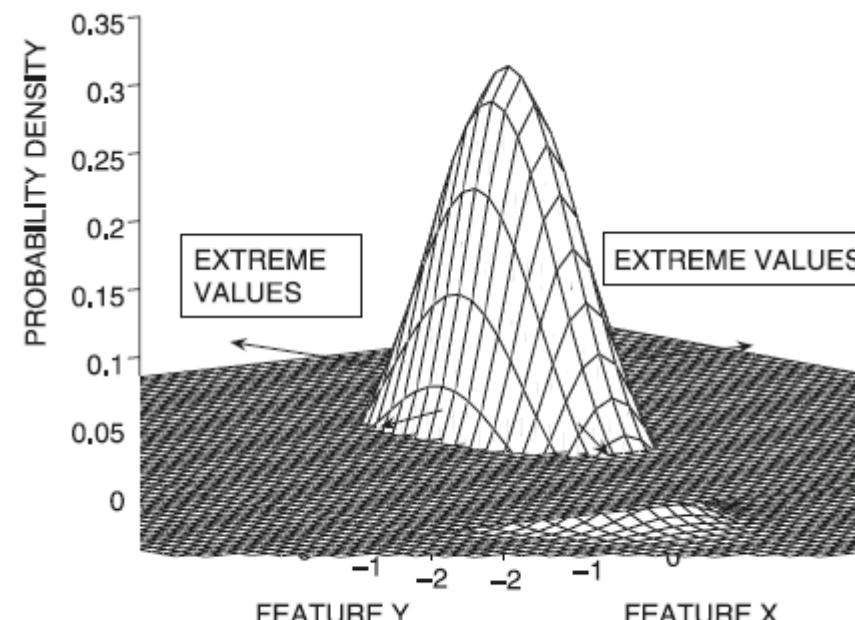
Multivariate Extreme Values (2)

□ Extreme-value Score of \bar{X}

- $Maha(\bar{X}, \bar{\mu}, \Sigma)$
- Larger values imply more extreme behavior

La distanza di Mahalanobis diventa quella Euclidea se il dataset è riferito alle sue componenti principali

Lungo ogni componente principale si può riproporre l'analisi svolta per la gaussiana unidimensionale



(b) Multivariate extreme values



Multivariate Extreme Values (2)

□ Extreme-value Score of \bar{X}

- $Maha(\bar{X}, \bar{\mu}, \Sigma)$
- Larger values imply more extreme behavior

□ Extreme-value Probability of \bar{X}

- Let \mathcal{R} be the region
$$\mathcal{R} = \{\bar{Y} | Maha(\bar{Y}, \bar{\mu}, \Sigma) \geq Maha(\bar{X}, \bar{\mu}, \Sigma)\}$$
- Cumulative probability of \mathcal{R}
- Cumulative Probability of χ^2 distribution for which the value is larger than $Maha(\bar{X}, \bar{\mu}, \Sigma)$



Why χ^2 distribution?

□ The Mahalanobis distance

- Let Σ be the covariance matrix

$$Maha(\bar{Y}, \bar{\mu}, \Sigma) = \sqrt{(\bar{Y} - \bar{\mu})\Sigma^{-1}(\bar{Y} - \bar{\mu})^\top}$$

■ Projection+Normalization

- ✓ Let $\Sigma = U\Lambda U^\top = \sum_{i=1}^d \sigma_i^2 \mathbf{u}_i \mathbf{u}_i^\top$
- ✓ Then, $\Sigma^{-1} = U\Lambda^{-1}U^\top = \sum_{i=1}^d \sigma_i^{-2} \mathbf{u}_i \mathbf{u}_i^\top$

$$Maha(\bar{Y}, \bar{\mu}, \Sigma) = \sqrt{(\bar{Y} - \bar{\mu}) \left(\sum_{i=1}^d \sigma_i^{-2} \mathbf{u}_i \mathbf{u}_i^\top \right) (\bar{Y} - \bar{\mu})^\top} = \sqrt{\sum_{i=1}^d \left(\frac{\mathbf{u}_i (\bar{Y} - \bar{\mu})^\top}{\sigma_i} \right)^2}$$



Why χ^2 distribution?

□ The Mahalanobis distance

- Let Σ be the covariance matrix

$$Maha(\bar{Y}, \bar{\mu}, \Sigma) = \sqrt{(\bar{Y} - \bar{\mu}) \Sigma^{-1} (\bar{Y} - \bar{\mu})^\top}$$

vettore riga

- Projection+Normalization

- ✓ Let $\Sigma = U\Lambda U^\top = \sum_{i=1}^d \sigma_i^2 \mathbf{u}_i \mathbf{u}_i^\top$
- ✓ Then, $\Sigma^{-1} = U\Lambda^{-1}U^\top = \sum_{i=1}^d \sigma_i^{-2} \mathbf{u}_i \mathbf{u}_i^\top$

$$Maha(\bar{Y}, \bar{\mu}, \Sigma) = \sqrt{(\bar{Y} - \bar{\mu}) \left(\sum_{i=1}^d \sigma_i^{-2} \mathbf{u}_i \mathbf{u}_i^\top \right) (\bar{Y} - \bar{\mu})^\top} = \sqrt{\sum_{i=1}^d \left(\frac{\mathbf{u}_i (\bar{Y} - \bar{\mu})^\top}{\sigma_i} \right)^2}$$



Why χ^2 distribution?

□ The Mahalanobis distance

- Let Σ be the covariance matrix

$$Maha(\bar{Y}, \bar{\mu}, \Sigma) = \sqrt{(\bar{Y} - \bar{\mu}) \Sigma^{-1} (\bar{Y} - \bar{\mu})^\top}$$

vettore riga

- Projection+Normalization

✓ Let $\Sigma = U\Lambda U^\top = \sum_{i=1}^d \sigma_i^2 \mathbf{u}_i \mathbf{u}_i^\top$ *Σ è semidefinita positiva e ammette sempre una diagonalizzazione*

✓ Then, $\Sigma^{-1} = U\Lambda^{-1}U^\top = \sum_{i=1}^d \sigma_i^{-2} \mathbf{u}_i \mathbf{u}_i^\top$

$$Maha(\bar{Y}, \bar{\mu}, \Sigma) = \sqrt{(\bar{Y} - \bar{\mu}) \left(\sum_{i=1}^d \sigma_i^{-2} \mathbf{u}_i \mathbf{u}_i^\top \right) (\bar{Y} - \bar{\mu})^\top} = \sqrt{\sum_{i=1}^d \left(\frac{\mathbf{u}_i (\bar{Y} - \bar{\mu})^\top}{\sigma_i} \right)^2}$$



Why χ^2 distribution?

- The Mahalanobis distance
 - Let Σ be the covariance matrix

$$Maha(\bar{Y}, \bar{\mu}, \Sigma) = \sqrt{(\bar{Y} - \bar{\mu}) \Sigma^{-1} (\bar{Y} - \bar{\mu})^\top}$$

vettore riga

- Projection+Normalization

✓ Let $\Sigma = U\Lambda U^\top = \sum_{i=1}^d \sigma_i^2 \mathbf{u}_i \mathbf{u}_i^\top$ *Σ è semidefinita positiva e ammette sempre una diagonalizzazione*

✓ Then, $\Sigma^{-1} = U\Lambda^{-1}U^\top = \sum_{i=1}^d \sigma_i^{-2} \mathbf{u}_i \mathbf{u}_i^\top$
vettore colonna

$$Maha(\bar{Y}, \bar{\mu}, \Sigma) = \sqrt{(\bar{Y} - \bar{\mu}) \left(\sum_{i=1}^d \sigma_i^{-2} \mathbf{u}_i \mathbf{u}_i^\top \right) (\bar{Y} - \bar{\mu})^\top} = \sqrt{\sum_{i=1}^d \left(\frac{\mathbf{u}_i (\bar{Y} - \bar{\mu})^\top}{\sigma_i} \right)^2}$$



Why χ^2 distribution?

□ The Mahalanobis distance

- Let Σ be the covariance matrix

$$Maha(\bar{Y}, \bar{\mu}, \Sigma) = \sqrt{(\bar{Y} - \bar{\mu}) \Sigma^{-1} (\bar{Y} - \bar{\mu})^\top}$$

vettore riga

- Projection+Normalization

✓ Let $\Sigma = U\Lambda U^\top = \sum_{i=1}^d \sigma_i^2 \mathbf{u}_i \mathbf{u}_i^\top$ *Σ è semidefinita positiva e ammette sempre una diagonalizzazione*

✓ Then, $\Sigma^{-1} = U\Lambda^{-1}U^\top = \sum_{i=1}^d \sigma_i^{-2} \mathbf{u}_i \mathbf{u}_i^\top$
vettore colonna

$$Maha(\bar{Y}, \bar{\mu}, \Sigma) = \sqrt{(\bar{Y} - \bar{\mu}) \left(\sum_{i=1}^d \sigma_i^{-2} \mathbf{u}_i \mathbf{u}_i^\top \right) (\bar{Y} - \bar{\mu})^\top} = \sqrt{\sum_{i=1}^d \left(\frac{\mathbf{u}_i (\bar{Y} - \bar{\mu})^\top}{\sigma_i} \right)^2}$$

z-score delle singole componenti di \bar{Y} lungo le direzioni degli autovettori di Σ



Why χ^2 distribution?

□ The Mahalanobis distance

- Let Σ be the covariance matrix

$$Maha(\bar{Y}, \bar{\mu}, \Sigma) = \sqrt{(\bar{Y} - \bar{\mu}) \Sigma^{-1} (\bar{Y} - \bar{\mu})^\top}$$

vettore riga

- Projection+Normalization

✓ Let $\Sigma = U\Lambda U^\top = \sum_{i=1}^d \sigma_i^2 \mathbf{u}_i \mathbf{u}_i^\top$ Σ è semidefinita positiva e ammette sempre una diagonalizzazione

✓ Then, $\Sigma^{-1} = U\Lambda^{-1}U^\top = \sum_{i=1}^d \sigma_i^{-2} \mathbf{u}_i \mathbf{u}_i^\top$
vettore colonna

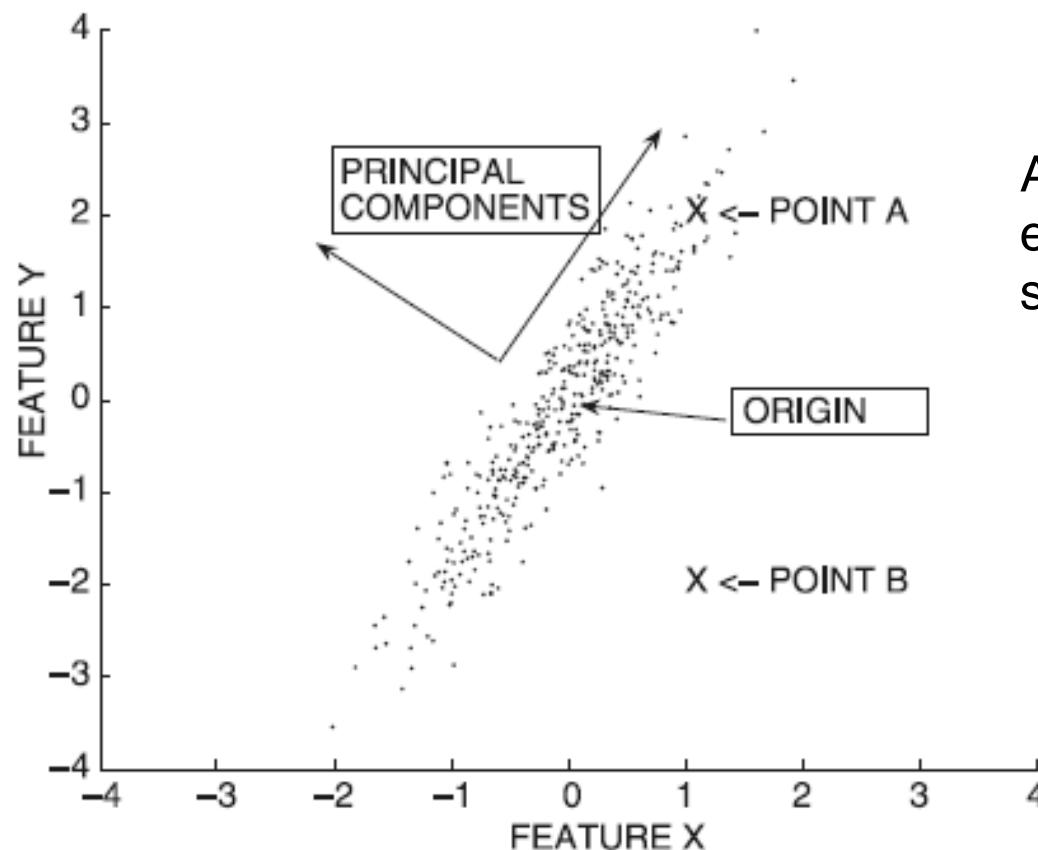
$$Maha(\bar{Y}, \bar{\mu}, \Sigma) = \sqrt{(\bar{Y} - \bar{\mu}) \left(\sum_{i=1}^d \sigma_i^{-2} \mathbf{u}_i \mathbf{u}_i^\top \right) (\bar{Y} - \bar{\mu})^\top} = \sqrt{\sum_{i=1}^d \left(\frac{\mathbf{u}_i (\bar{Y} - \bar{\mu})^\top}{\sigma_i} \right)^2}$$

Per definizione la distribuzione chi-square con d gradi di libertà è quella che assume la somma dei quadrati di d variabili con distribuzione normale standard ($\mu=0, \sigma=1$)

z-score delle singole componenti di \bar{Y} lungo le direzioni degli autovettori di Σ

Adaptive to the Shape

- B is an extreme value



A sembra essere un punto estremo, ma non lo è nel senso di Mahalanobis.



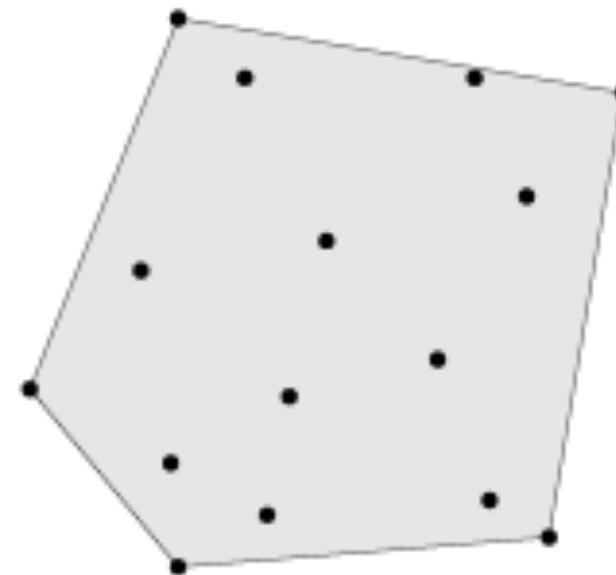
Depth-Based Methods

□ Convex Hull

The *convex hull* of a set C , denoted $\text{conv } C$, is the set of all convex combinations of points in C :

$$\text{conv } C = \{\theta_1 x_1 + \cdots + \theta_k x_k \mid x_i \in C, \theta_i \geq 0, i = 1, \dots, k, \theta_1 + \cdots + \theta_k = 1\}.$$

■ Corners





The Procedure

- The index k is the outlier score
 - Smaller values indicate a grate tendency

Algorithm *FindDepthOutliers*(Data Set: \mathcal{D} , Score Threshold: r)

begin

$k = 1$;

repeat

 Find set S of corners of convex hull of \mathcal{D} ;

 Assign depth k to points in S ;

$\mathcal{D} = \mathcal{D} - S$;

$k = k + 1$;

until(D is empty);

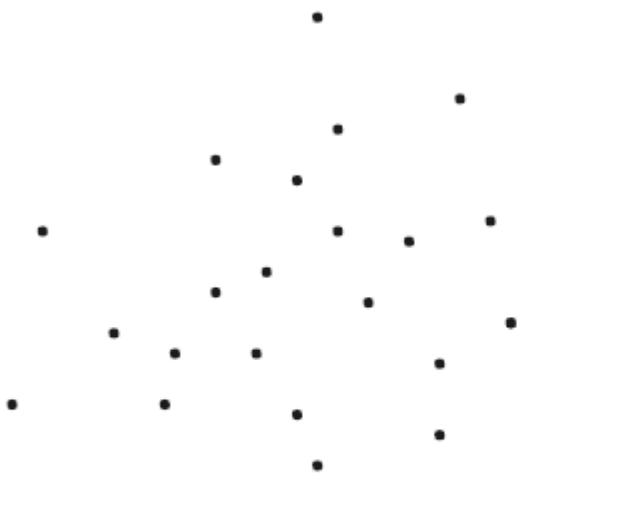
 Report points with depth at most r as outliers;

end



An Example

□ Peeling Layers of an Onion

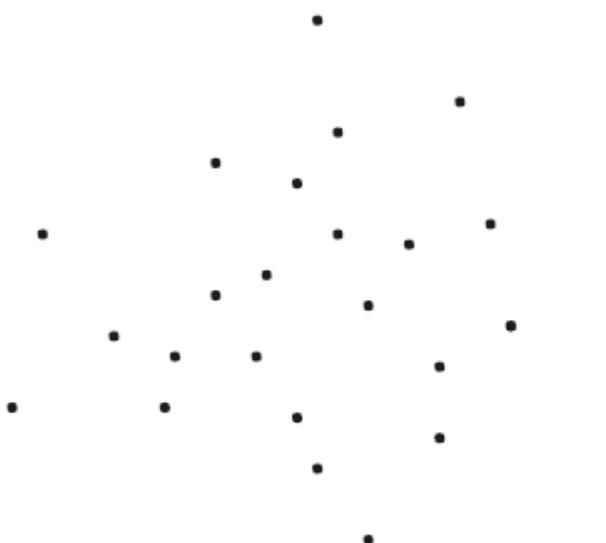


(a)

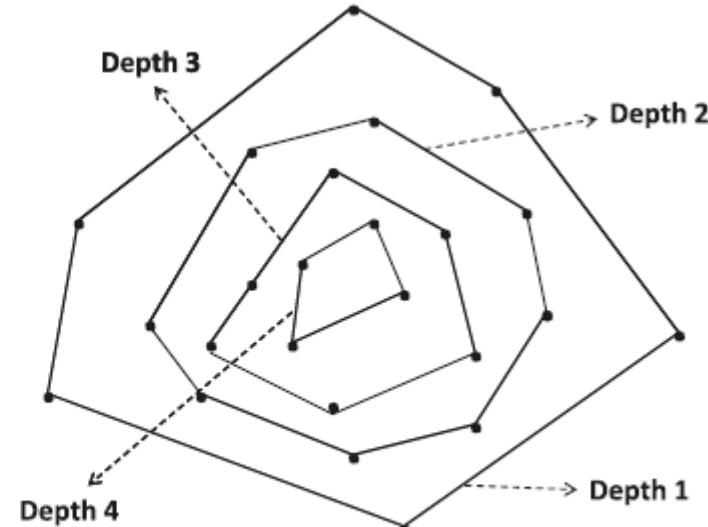


An Example

□ Peeling Layers of an Onion



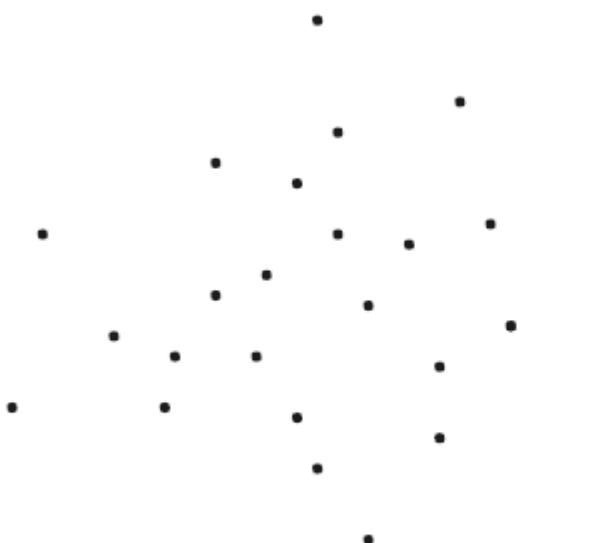
(a)



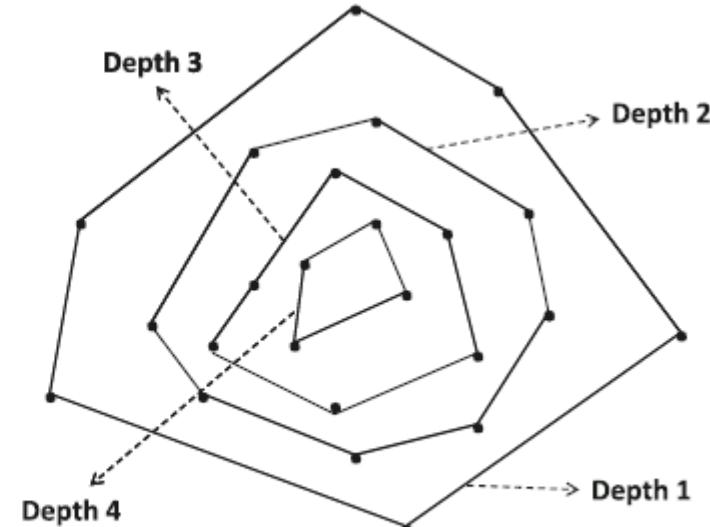
(b)

An Example

□ Peeling Layers of an Onion



(a)



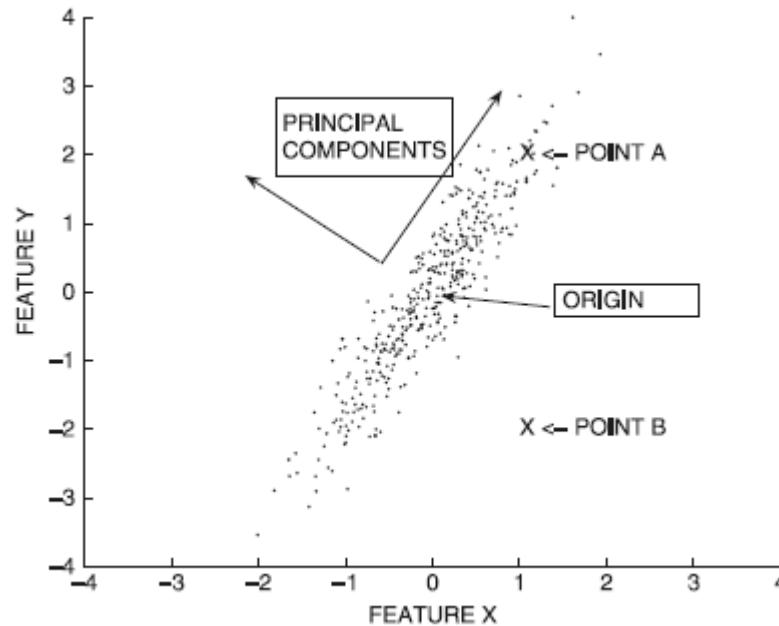
(b)

Saranno davvero tutti outlier? Non c'è distinzione!!!



Limitations

No Normalization



- Many data points are indistinguishable
- The computational complexity increases significantly with dimensionality



Outline

- Introduction
- Extreme Value Analysis
- **Probabilistic Models**
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- Summary



Probabilistic Models

- Related to Probabilistic Model-Based Clustering
- The Key Idea
 - Assume data is generated from a mixture-based generative model
 - Learn the parameter of the model from data
 - ✓ EM algorithm
 - Evaluate the probability of each data point being generated by the model
 - ✓ Points with low values are outliers



Mixture-based Generative Model

- Data was generated from a mixture of k distributions with probability distribution $\mathcal{G}_1, \dots, \mathcal{G}_k$
- \mathcal{G}_i represents a cluster/mixture component
- Each point \bar{X} is generated as follows
 - Select a mixture component with probability $\alpha_i = P(\mathcal{G}_i)$, $i = 1, \dots, k$
 - Assume the r -th component is selected
 - Generate a data point from G_r



Learning Parameter from Data

- The probability that \bar{X}_j generated by the mixture model \mathcal{M} is given by

$$f^{point}(\bar{X}_j | \mathcal{M}) = \sum_{i=1}^k P(\mathcal{G}_i, \bar{X}_j) = \sum_{i=1}^k P(\mathcal{G}_i)P(\bar{X}_j | \mathcal{G}_i) = \sum_{i=1}^k \alpha_i \cdot f^i(\bar{X}_j)$$

- The probability of the data set $\mathcal{D} = \{\bar{X}_1, \dots, \bar{X}_n\}$ generated by \mathcal{M}

$$f^{data}(\mathcal{D} | \mathcal{M}) = \prod_{j=1}^n f^{point}(\bar{X}_j | \mathcal{M}).$$

- Learning parameters that maximize

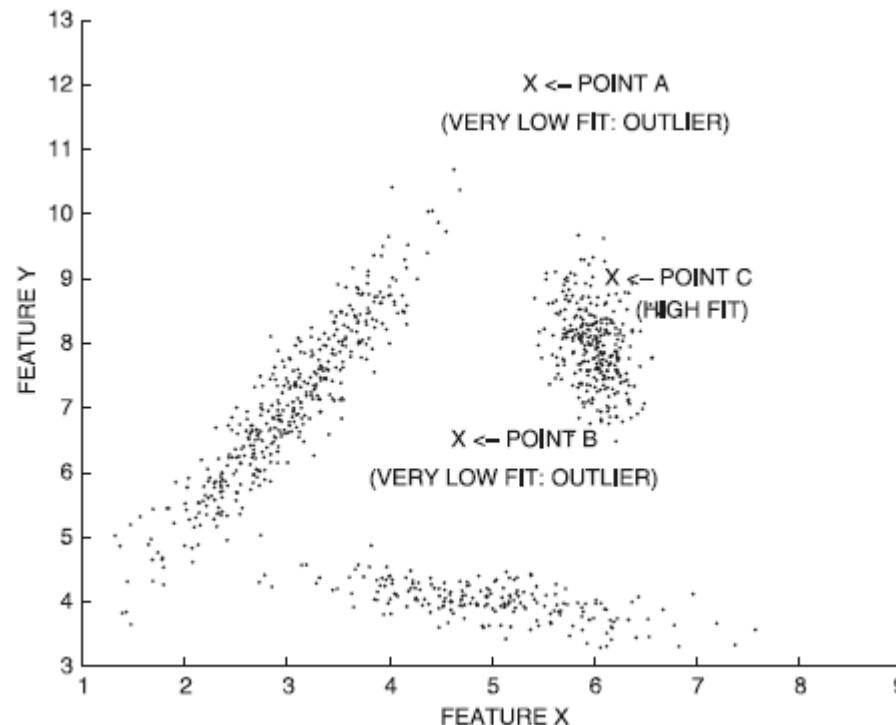
$$\mathcal{L}(\mathcal{D} | \mathcal{M}) = \log\left(\prod_{j=1}^n f^{point}(\bar{X}_j | \mathcal{M})\right) = \sum_{j=1}^n \log\left(\sum_{i=1}^k \alpha_i f^i(\bar{X}_j)\right)$$



Identify Outliers

- Outlier Score is defined as

$$f^{point}(\bar{X}_j | \mathcal{M}) = \sum_{i=1}^k P(\mathcal{G}_i, \bar{X}_j) = \sum_{i=1}^k P(\mathcal{G}_i)P(\bar{X}_j | \mathcal{G}_i) = \sum_{i=1}^k \alpha_i \cdot f^i(\bar{X}_j)$$





Outline

- Introduction
- Extreme Value Analysis
- Probabilistic Models
- **Clustering for Outlier Detection**
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- Summary



Clustering for Outlier Detection

- Outlier Analysis v.s. Clustering
 - Clustering is about finding “crowds” of data points
 - Outlier analysis is about finding data points that are far away from these crowds
- Every data point is
 - Either a member of a cluster
 - Or an outlier
- Some clustering algorithms also detect outliers
 - DBSCAN, DENCLUE



Clustering for Outlier Detection

- Outlier Analysis v.s. Clustering
 - Clustering is about finding “crowds” of data points
 - Outlier analysis is about finding data points that are far away from these crowds
- Every data point is
 - Either a member of a cluster
 - Or an outlier
- Some clustering algorithms also detect outliers
 - DBSCAN, DENCLUE

Algoritmo di clustering basato sul concetto di “Kernel density estimation”



Clustering for Outlier Detection

- Outlier Analysis v.s. Clustering
 - Clustering is about finding “crowds” of data points
 - Outlier analysis is about finding data points that are far away from these crowds
- Every data point is
 - Either a member of a cluster
 - Or an outlier
- Some clustering algorithms also detect outliers
 - DBSCAN, DENCLUE

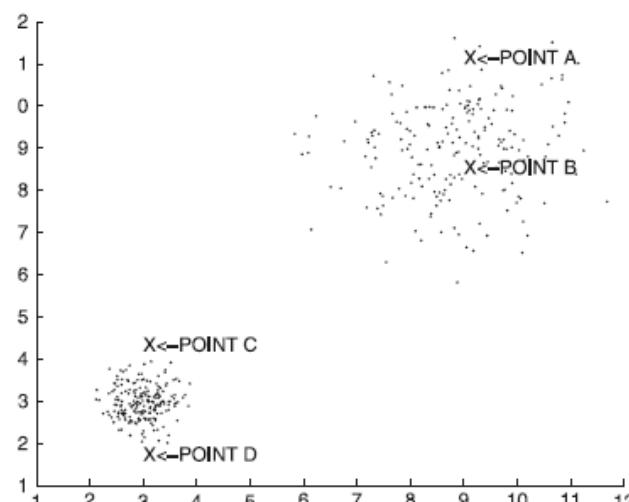
Algoritmo di clustering basato sul concetto di “Kernel density estimation”

Si fa una stima della densità di funzioni kernel $K(\cdot)$ che si sostituiscono ai punti del data set e si ottiene una stima “dolce” dei contorni dei cluster.

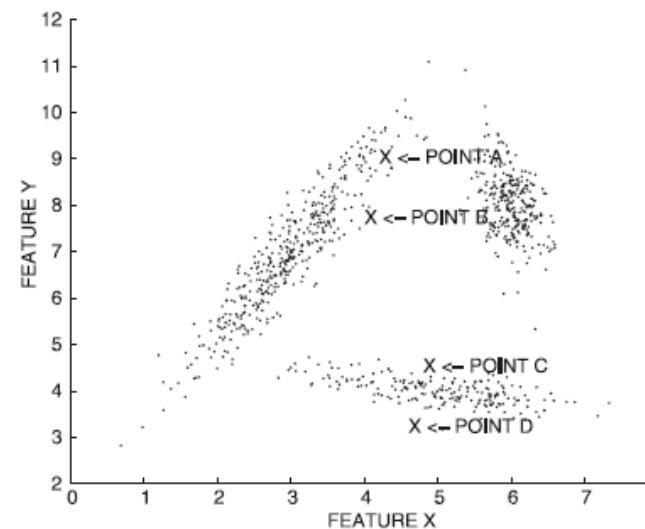
The Procedure (1)

□ A Simple Way

1. Cluster the data
2. Define the outlier score as the distance of the data point to its cluster centroid



(a) local density variation



(b) local orientation variation

*La struttura locale di ogni cluster può inficiare la validità
di uno score basato su distanze dirette dai centroidi*



The Procedure (2)

□ A Better Approach

1. Cluster the data
2. Define the outlier score as the **local Mahalanobis distance**
 - ✓ Suppose \bar{X} belongs to cluster r

$$Maha(\bar{X}, \bar{\mu}_r, \Sigma_r) = \sqrt{(\bar{X} - \bar{\mu}_r) \Sigma_r^{-1} (\bar{X} - \bar{\mu}_r)^T}.$$

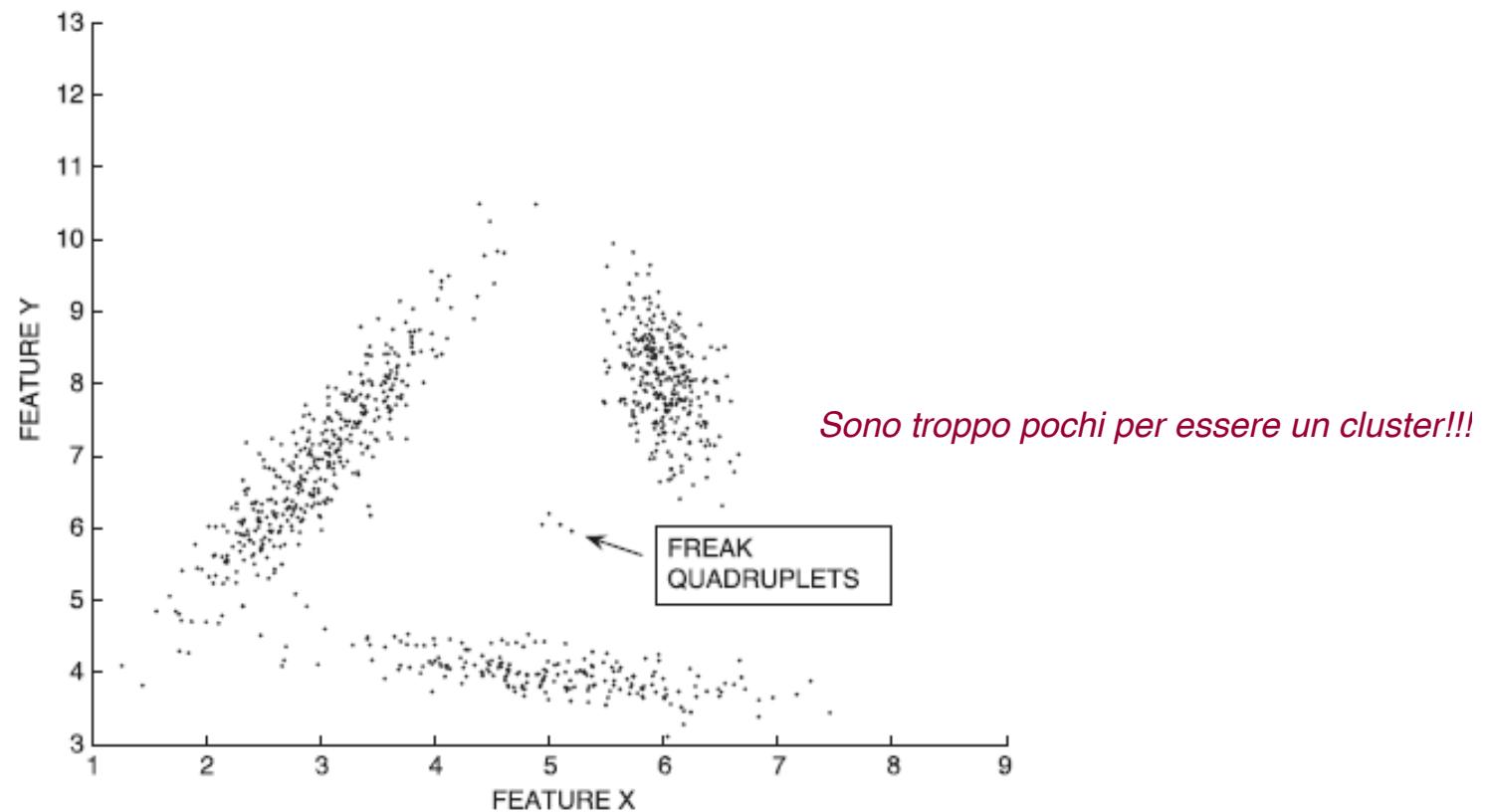
- ✓ $\bar{\mu}_r$ is the mean vector of the r -th cluster
- ✓ Σ_r is the covariance matrix of the r -th cluster

□ Multivariate Extreme Value Analysis

■ Global Mahalanobis distance

A Post-processing Step

Remove Small-Size Clusters





Outline

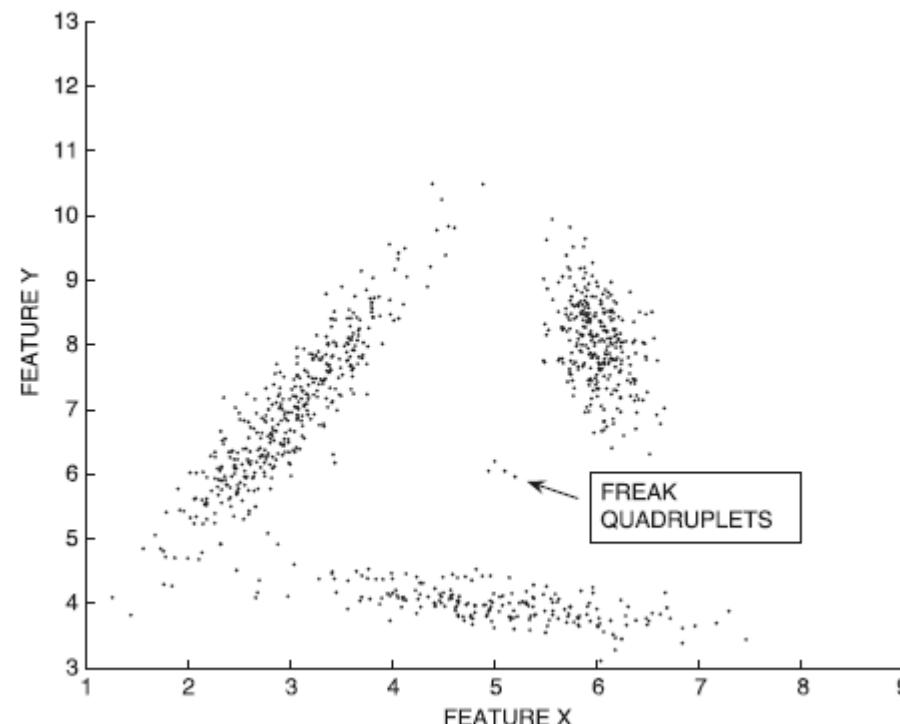
- Introduction
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- **Distance-Based Outlier Detection**
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- Summary



Distance-Based Outlier Detection

- An *Instance-Specific* Definition
 - The distance-based outlier score of an object O is its distance to its k -th nearest neighbor

k è un parametro
scelto dall'utente





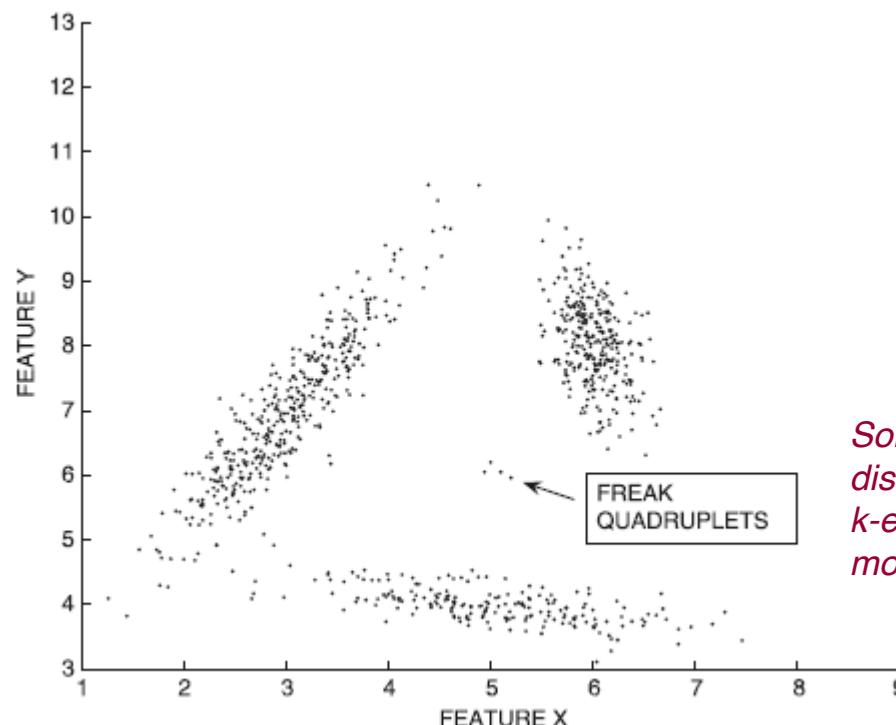
Distance-Based Outlier Detection

□ An *Instance-Specific* Definition

- The distance-based outlier score of an object O is its distance to its k -th nearest neighbor

k è un parametro
scelto dall'utente

$$k > 3$$



Sono tutti outlier perché la distanza $V_k(X)$ dal k -esimo vicino è sempre molto elevata



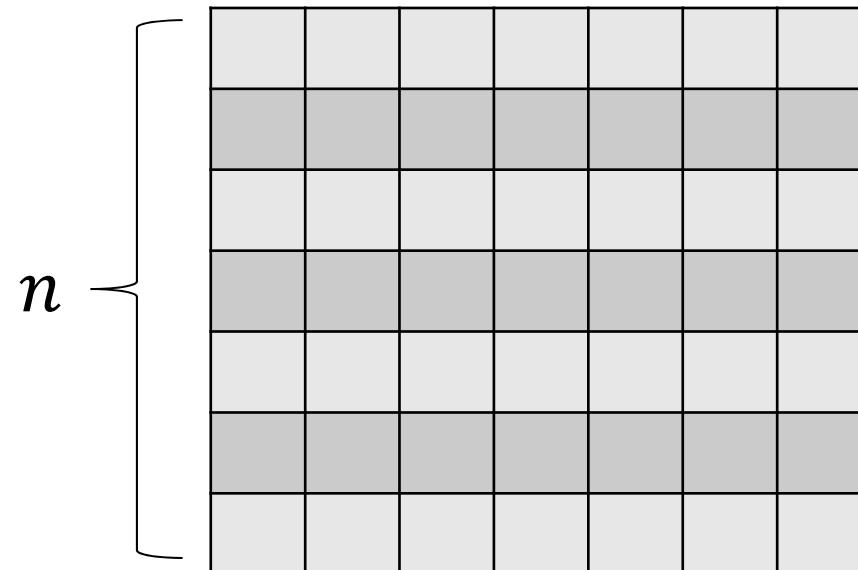
Distance-Based Outlier Detection

- An *Instance-Specific* Definition
 - The distance-based outlier score of an object O is its distance to its k -th nearest neighbor
 - Sometimes, average distance is used
- High-computational Cost $O(n^2)$
 - Index structure
 - ✓ Effective when the dimensionality is low
 - Pruning tricks
 - ✓ Designed for the case that only the top- r outliers are needed

The Naïve Approach for Finding Top r -Outliers



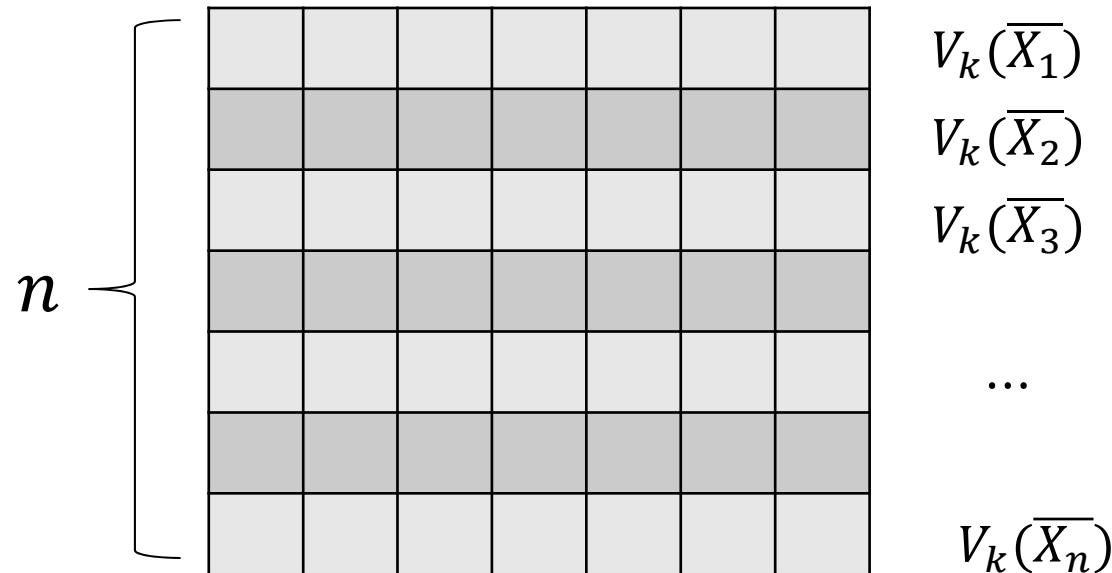
1. Evaluate the $n \times n$ distance matrix



The Naïve Approach for Finding Top r -Outliers



1. Evaluate the $n \times n$ distance matrix

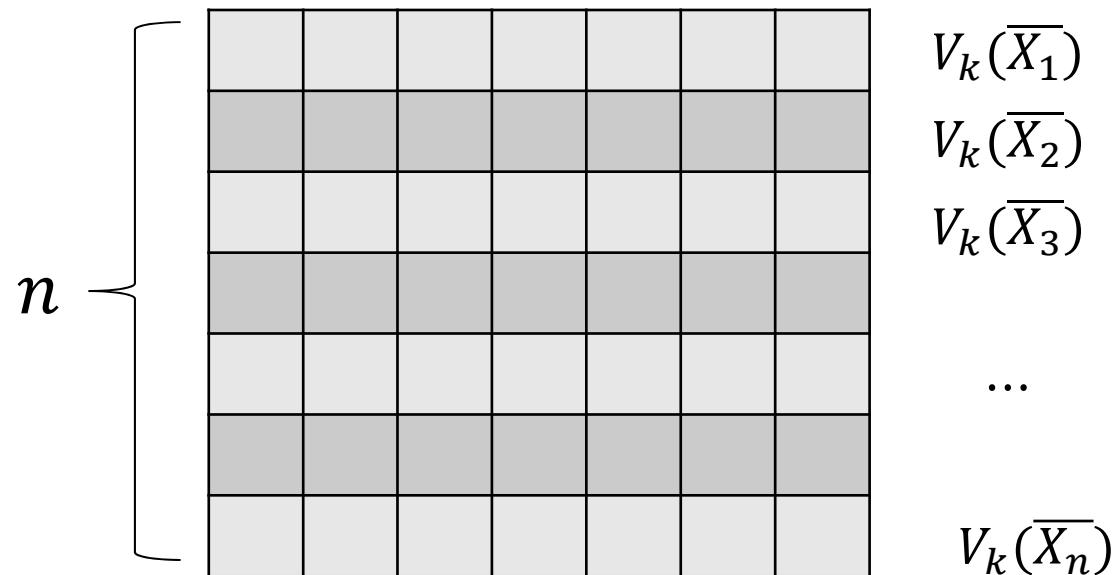


2. Find the k -th **smallest** value in each row



The Naïve Approach for Finding Top r -Outliers

1. Evaluate the $n \times n$ distance matrix

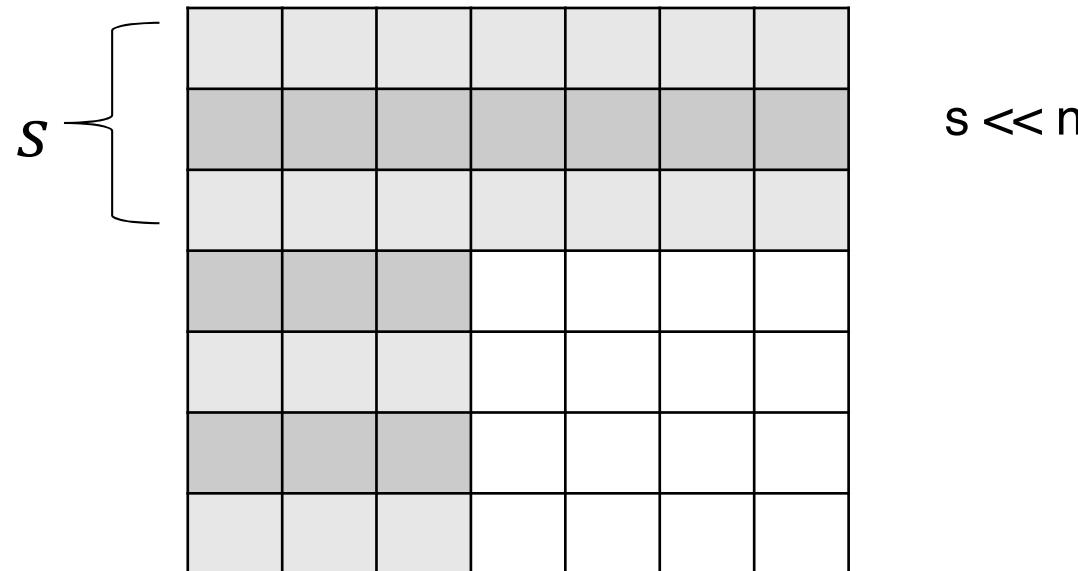


2. Find the k -th **smallest** value in each row
3. Choose r data points with **largest** $V_k(\cdot)$



Pruning Methods—Sampling

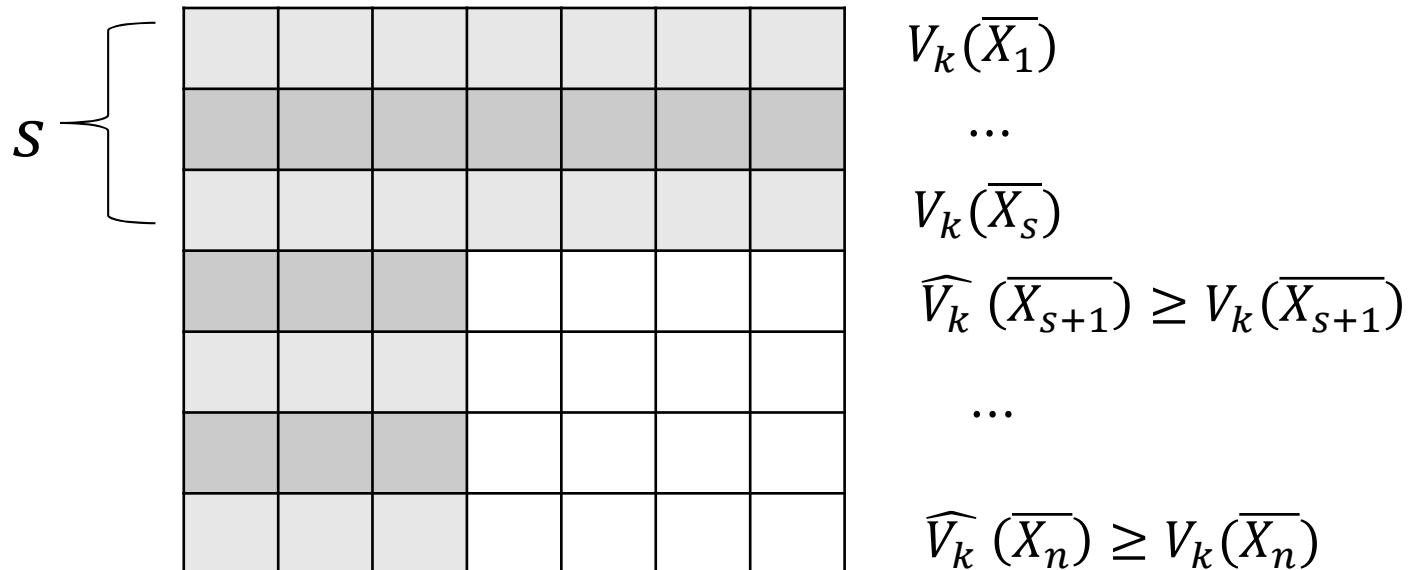
1. Evaluate a $s \times n$ distance matrix





Pruning Methods—Sampling

1. Evaluate a $s \times n$ distance matrix

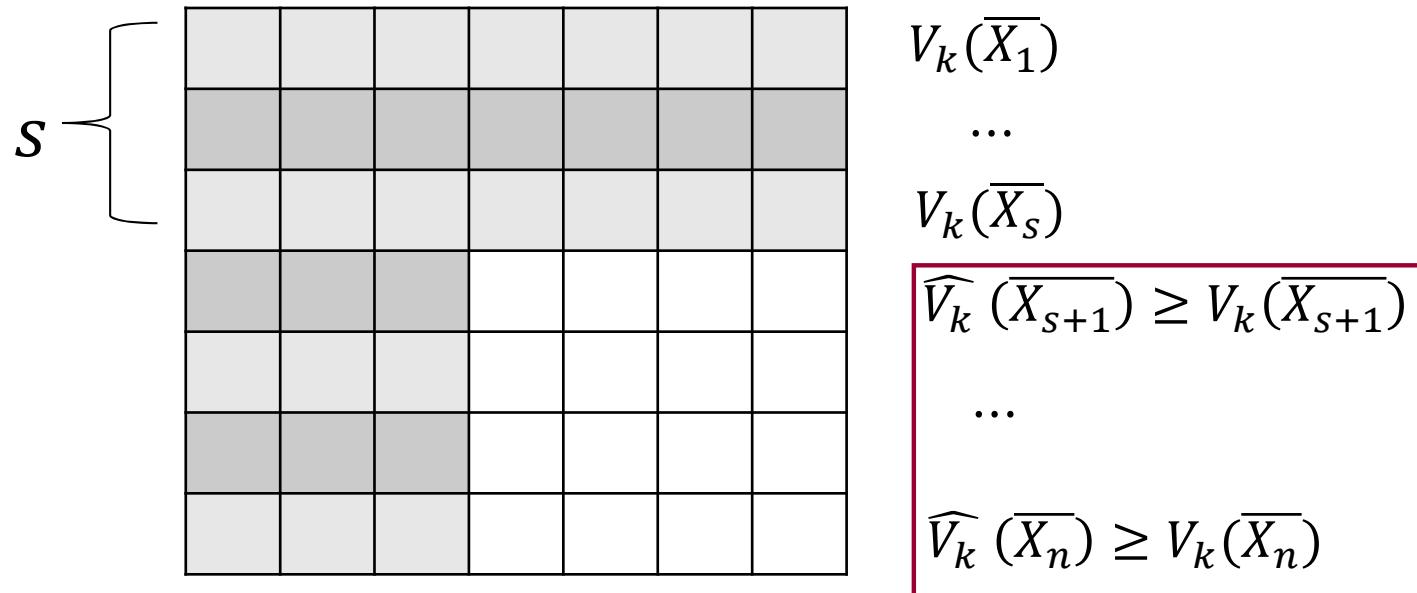


2. Find the k -th smallest value in each row



Pruning Methods—Sampling

1. Evaluate a $s \times n$ distance matrix



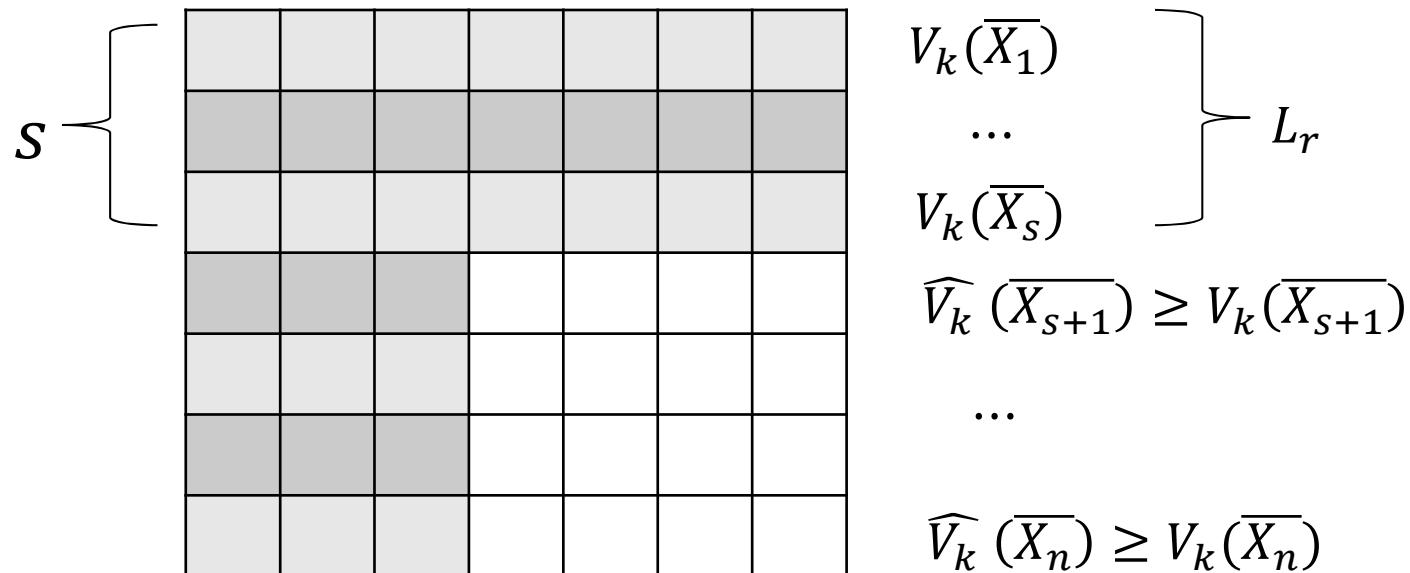
2. Find the k -th smallest value in each row

Il valore stimato della distanza dal k -esimo vicino è certamente un limite superiore del valore vero



Pruning Methods—Sampling

1. Evaluate a $s \times n$ distance matrix

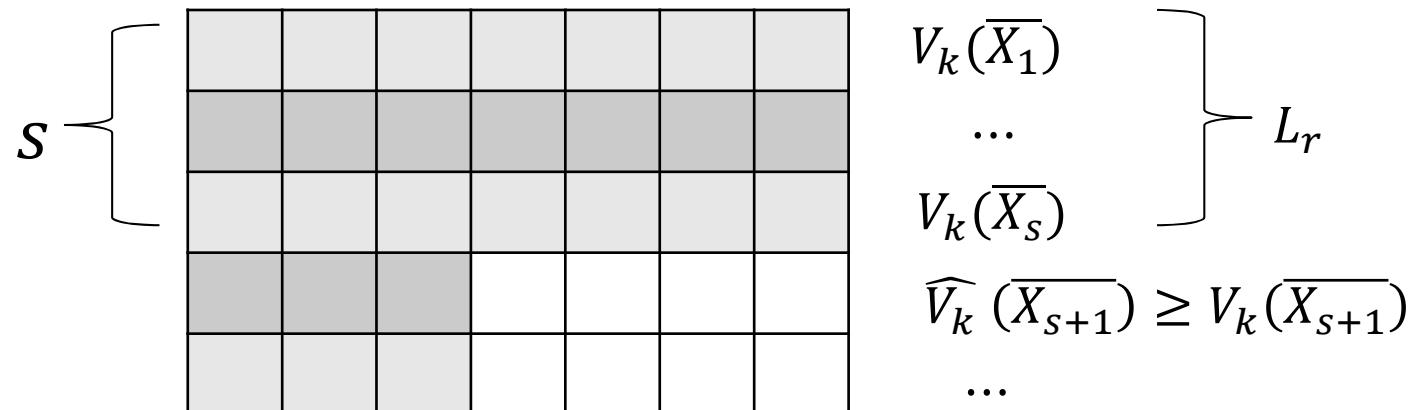


2. Find the k -th smallest value in each row
3. Identify the r -th score in top s -rows



Pruning Methods—Sampling

1. Evaluate a $s \times n$ distance matrix

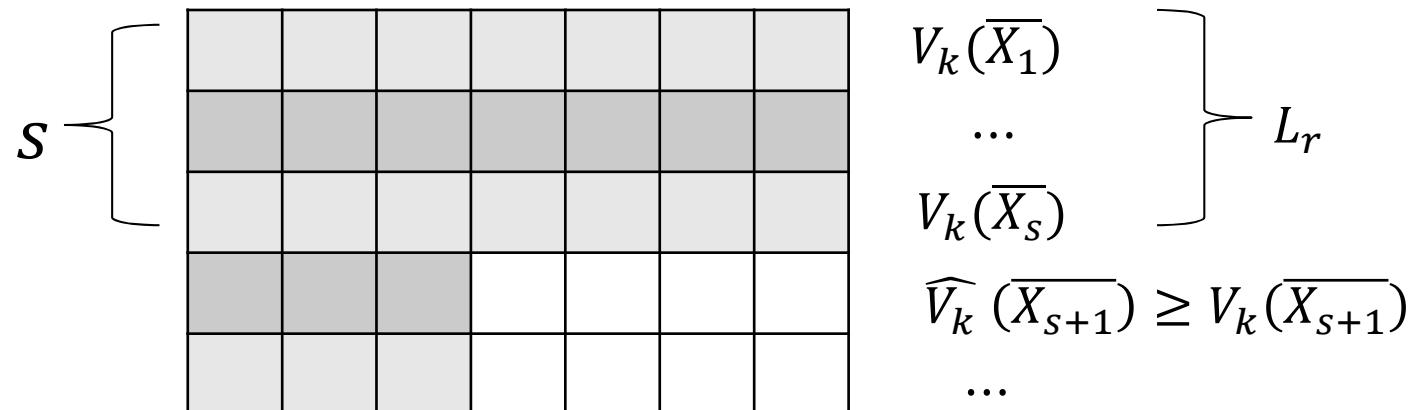


2. Find the k -th smallest value in each row
3. Identify the r -th score in top s -rows
4. Remove points with $\widehat{V}_k(\cdot) \leq L_r$



Pruning Methods—Sampling

1. Evaluate a $s \times n$ distance matrix



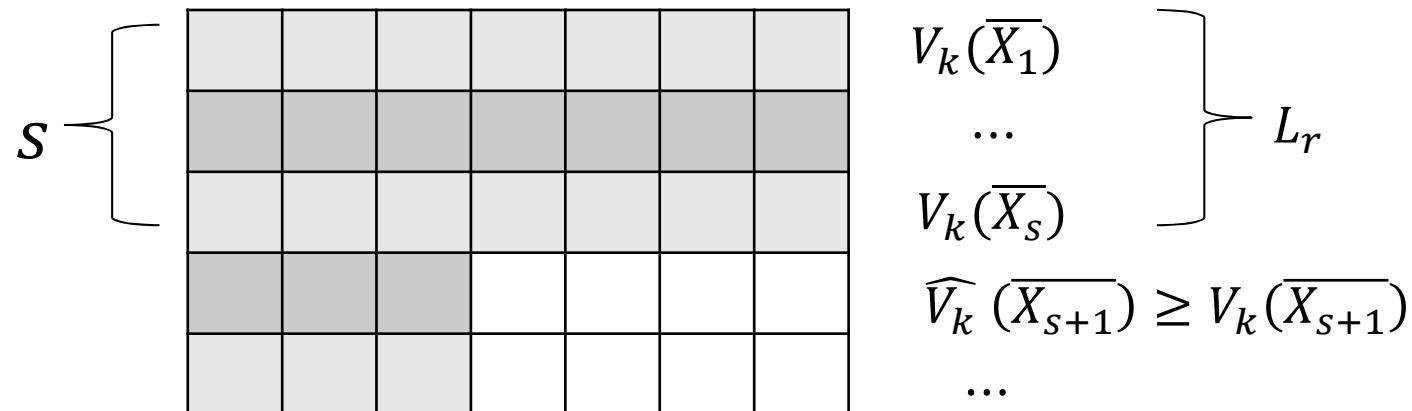
2. Find the k -th smallest value in each row
3. Identify the r -th score in top s -rows
4. Remove points with $\widehat{V}_k(\cdot) \leq L_r$

Il “limite superiore” alla loro distanza dal k -esimo vicino è certamente più piccolo del minimo degli r top outliers



Pruning Methods—Sampling

1. Evaluate a $s \times n$ distance matrix



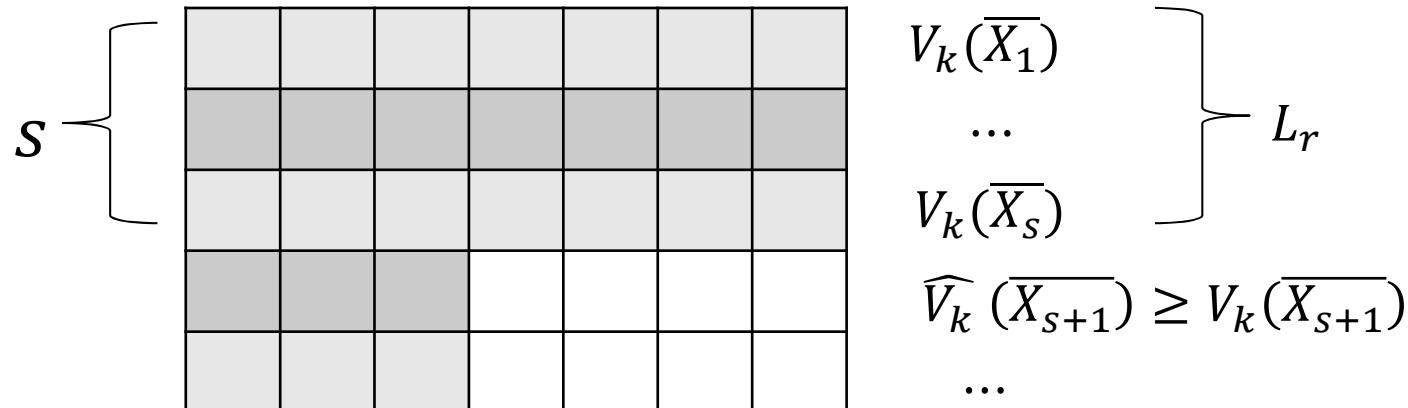
A questo punto è necessario calcolare le distanze rimanenti per individuare i veri r top outliers

2. Find the k -th smallest value in each row
3. Identify the r -th score in top s -rows
4. Remove points with $\widehat{V}_k(\cdot) \leq L_r$



Pruning Methods—Early Termination

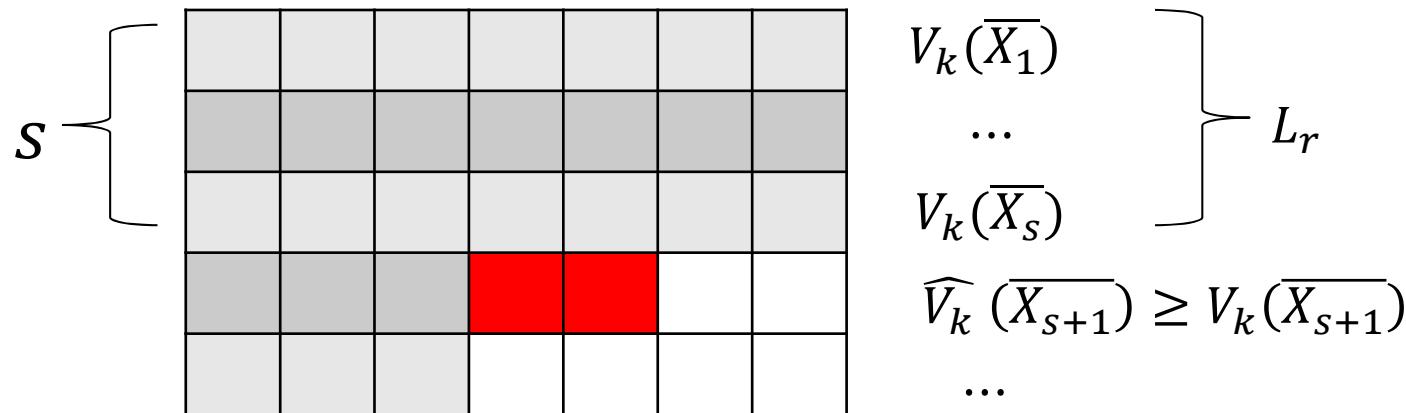
- When completing the empty area





Pruning Methods—Early Termination

- When completing the empty area



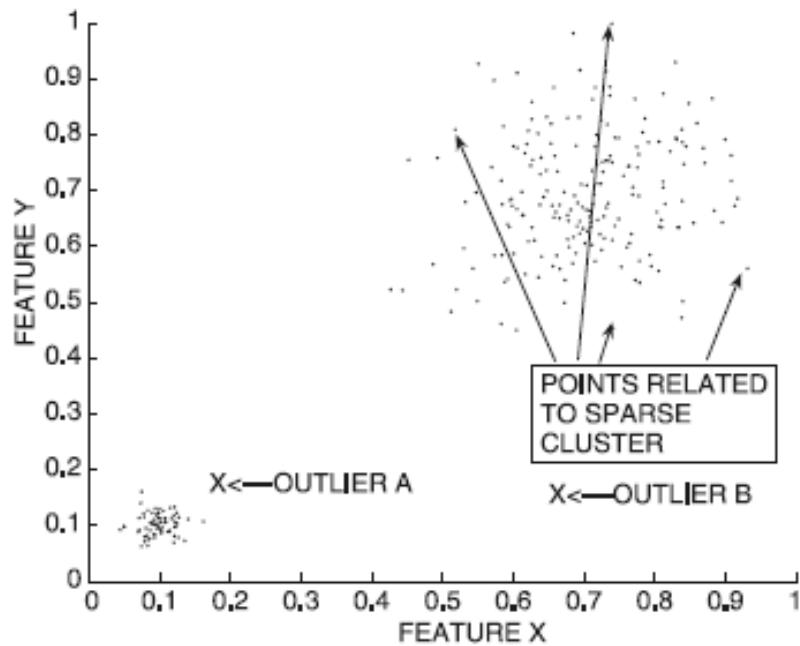
- Update $\widehat{V}_k(\cdot)$ when more distances are known
- Stop if $\widehat{V}_k(\cdot) \leq L_r$
- Update L_r if necessary

Local Distance Correction Methods

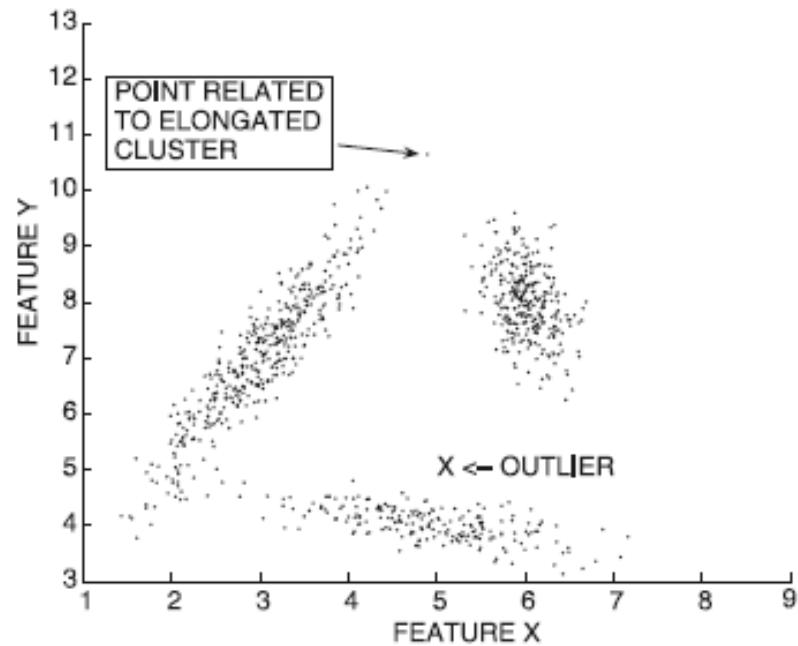


□ Impact of Local Variations

La distanza euclidea non è un rank corretto per gli outlier.



(a) Varying cluster density



(b) Varying cluster shape

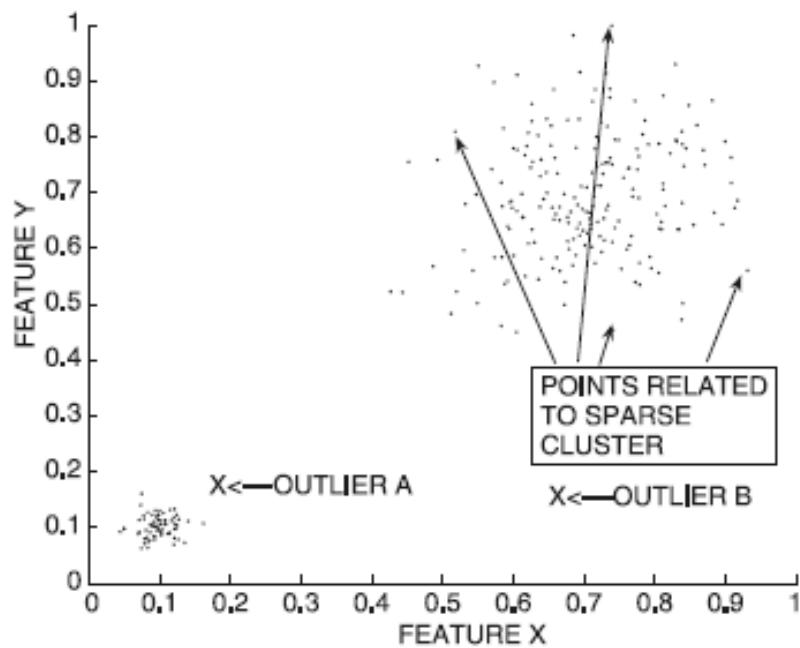
Local Distance Correction Methods



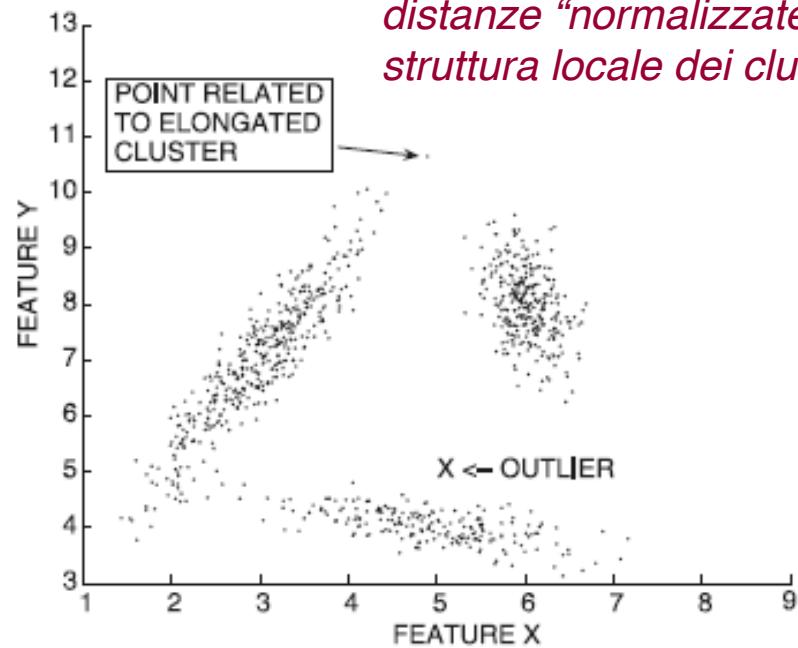
□ Impact of Local Variations

La distanza euclidea non è un rank corretto per gli outlier.

E' necessario calcolare delle distanze "normalizzate" rispetto alla struttura locale dei cluster.



(a) Varying cluster density



(b) Varying cluster shape



Local Outlier Factor (LOF)

- Let $V^k(\bar{X})$ be the distance of \bar{X} to its k -nearest neighbor



Local Outlier Factor (LOF)

- Let $V^k(\bar{X})$ be the distance of \bar{X} to its k -nearest neighbor
- Let $L_k(\bar{X})$ be the set of points within the k -nearest neighbor distance of \bar{X}



Local Outlier Factor (LOF)

- Let $V^k(\bar{X})$ be the distance of \bar{X} to its k -nearest neighbor
- Let $L_k(\bar{X})$ be the set of points within the k -nearest neighbor distance of \bar{X}
- Reachability Distance

$$R_k(\bar{X}, \bar{Y}) = \max\{Dist(\bar{X}, \bar{Y}), V^k(\bar{Y})\}$$



Local Outlier Factor (LOF)

- Let $V^k(\bar{X})$ be the distance of \bar{X} to its k -nearest neighbor
- Let $L_k(\bar{X})$ be the set of points within the k -nearest neighbor distance of \bar{X}
- Reachability Distance

$$R_k(\bar{X}, \bar{Y}) = \max\{Dist(\bar{X}, \bar{Y}), V^k(\bar{Y})\}$$

- Not symmetric between \bar{X} and \bar{Y}
- If $Dist(\bar{X}, \bar{Y})$ is large, $R_k(\bar{X}, \bar{Y}) = Dist(\bar{X}, \bar{Y})$
- Otherwise, $R_k(\bar{X}, \bar{Y}) = V^k(\bar{Y})$
 - ✓ Smoothed out by $V^k(\bar{Y})$, more stable



Local Outlier Factor (LOF)

□ Average Reachability Distance

$$AR_k(\bar{X}) = \text{MEAN}_{\bar{Y} \in L_k(\bar{X})} R_k(\bar{X}, \bar{Y})$$



Local Outlier Factor (LOF)

□ Average Reachability Distance

$$AR_k(\bar{X}) = \text{MEAN}_{\bar{Y} \in L_k(\bar{X})} R_k(\bar{X}, \bar{Y})$$

□ Local Outlier Factor

$$LOF_k(\bar{X}) = \text{MEAN}_{\bar{Y} \in L_k(\bar{X})} \frac{AR_k(\bar{X})}{AR_k(\bar{Y})}$$



Local Outlier Factor (LOF)

□ Average Reachability Distance

$$AR_k(\bar{X}) = \text{MEAN}_{\bar{Y} \in L_k(\bar{X})} R_k(\bar{X}, \bar{Y})$$

I valori di $AR_k(X)$ e $LOF_k(X)$ sono calcolati rispetto alla Reachability dei punti appartenenti all'insieme L_k dei punti all'interno di $V_k(X)$

□ Local Outlier Factor

$$LOF_k(\bar{X}) = \text{MEAN}_{\bar{Y} \in L_k(\bar{X})} \frac{AR_k(\bar{X})}{AR_k(\bar{Y})}$$



Local Outlier Factor (LOF)

□ Average Reachability Distance

$$AR_k(\bar{X}) = \text{MEAN}_{\bar{Y} \in L_k(\bar{X})} R_k(\bar{X}, \bar{Y})$$

□ Local Outlier Factor

$$LOF_k(\bar{X}) = \text{MEAN}_{\bar{Y} \in L_k(\bar{X})} \frac{AR_k(\bar{X})}{AR_k(\bar{Y})}$$

I valori di $AR_k(X)$ e $LOF_k(X)$ sono calcolati rispetto alla Reachability dei punti appartenenti all'insieme L_k dei punti all'interno di $V_k(X)$

In questo modo si introduce il concetto di “normalizzazione” della distanza

Local Outlier Factor (LOF)

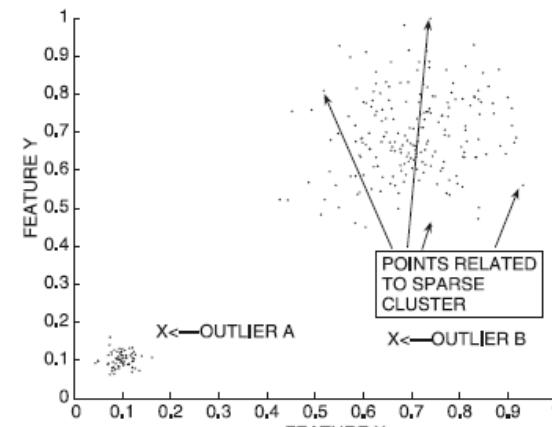
□ Average Reachability Distance

$$AR_k(\bar{X}) = \text{MEAN}_{\bar{Y} \in L_k(\bar{X})} R_k(\bar{X}, \bar{Y})$$

□ Local Outlier Factor

$$LOF_k(\bar{X}) = \text{MEAN}_{\bar{Y} \in L_k(\bar{X})} \frac{AR_k(\bar{X})}{AR_k(\bar{Y})}$$

- Larger for Outliers
- Close to 1 for Others



(a) Varying cluster density

I valori di $AR_k(X)$ e $LOF_k(X)$ sono calcolati rispetto alla Reachability dei punti appartenenti all'insieme L_k dei punti all'interno di $V_k(X)$

In questo modo si introduce il concetto di “normalizzazione” della distanza

Local Outlier Factor (LOF)

□ Average Reachability Distance

$$AR_k(\bar{X}) = \text{MEAN}_{\bar{Y} \in L_k(\bar{X})} R_k(\bar{X}, \bar{Y})$$

□ Local Outlier Factor

$$LOF_k(\bar{X}) = \text{MEAN}_{\bar{Y} \in L_k(\bar{X})} \frac{AR_k(\bar{X})}{AR_k(\bar{Y})}$$

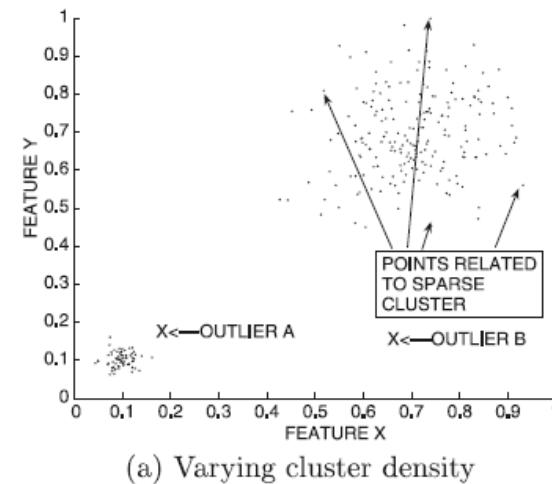
- Larger for Outliers
- Close to 1 for Others

□ Outlier Score

$$\max_k LOF_k(\bar{X})$$

I valori di $AR_k(X)$ e $LOF_k(X)$ sono calcolati rispetto alla Reachability dei punti appartenenti all'insieme L_k dei punti all'interno di $V_k(X)$

In questo modo si introduce il concetto di "normalizzazione" della distanza



Instance-Specific Mahalanobis Distance (1)



- Define a local Mahalanobis distance for each point
 - Based on the covariance structure of the neighborhood of a data point

- The Challenge
 - Neighborhood of a data point is hard to define with the Euclidean distance
 - Euclidean distance is biased toward capturing the circular region around that point

Instance-Specific Mahalanobis Distance (2)



- An agglomerative approach for neighborhood construction
 - Add \bar{X} to $L^k(\bar{X})$
 - Data points are **iteratively** added to $L^k(\bar{X})$ that have the smallest distance to $L^k(\bar{X})$

$$\operatorname{argmin}_{\bar{Y} \in \mathcal{D}} \min_{\bar{Z} \in L^k(\bar{X})} dist(\bar{Y} - \bar{Z})$$

- Instance-specific Mahalanobis score

$$LMaha_k(\bar{X}) = Maha(\bar{X}, \overline{\mu_k(X)}, \Sigma_k(\bar{X}))$$

- Outlier score $\max_k LMaha_k (\bar{X})$

Instance-Specific Mahalanobis Distance (2)



- An agglomerative approach for neighborhood construction
 - Add \bar{X} to $L^k(\bar{X})$

Instance-Specific Mahalanobis Distance (2)



- An agglomerative approach for neighborhood construction
 - Add \bar{X} to $L^k(\bar{X})$
 - Data points are **iteratively** added to $L^k(\bar{X})$ that have the smallest distance to $L^k(\bar{X})$

$$\operatorname{argmin}_{\bar{Y} \in \mathcal{D}} \min_{\bar{Z} \in L^k(\bar{X})} dist(\bar{Y} - \bar{Z})$$

Instance-Specific Mahalanobis Distance (2)



- An agglomerative approach for neighborhood construction
 - Add \bar{X} to $L^k(\bar{X})$
 - Data points are **iteratively** added to $L^k(\bar{X})$ that have the smallest distance to $L^k(\bar{X})$

$$\operatorname{argmin}_{\bar{Y} \in \mathcal{D}} \min_{\bar{Z} \in L^k(\bar{X})} dist(\bar{Y} - \bar{Z})$$

- Instance-specific Mahalanobis score

$$LMaha_k(\bar{X}) = Maha(\bar{X}, \overline{\mu_k(X)}, \Sigma_k(\bar{X}))$$

Instance-Specific Mahalanobis Distance (2)



- An agglomerative approach for neighborhood construction
 - Add \bar{X} to $L^k(\bar{X})$
 - Data points are **iteratively** added to $L^k(\bar{X})$ that have the smallest distance to $L^k(\bar{X})$

$$\operatorname{argmin}_{\bar{Y} \in \mathcal{D}} \min_{\bar{Z} \in L^k(\bar{X})} dist(\bar{Y} - \bar{Z})$$

- Instance-specific Mahalanobis score

$$LMaha_k(\bar{X}) = Maha(\bar{X}, \overline{\mu_k(X)}, \boxed{\Sigma_k(\bar{X})})$$

- Outlier score

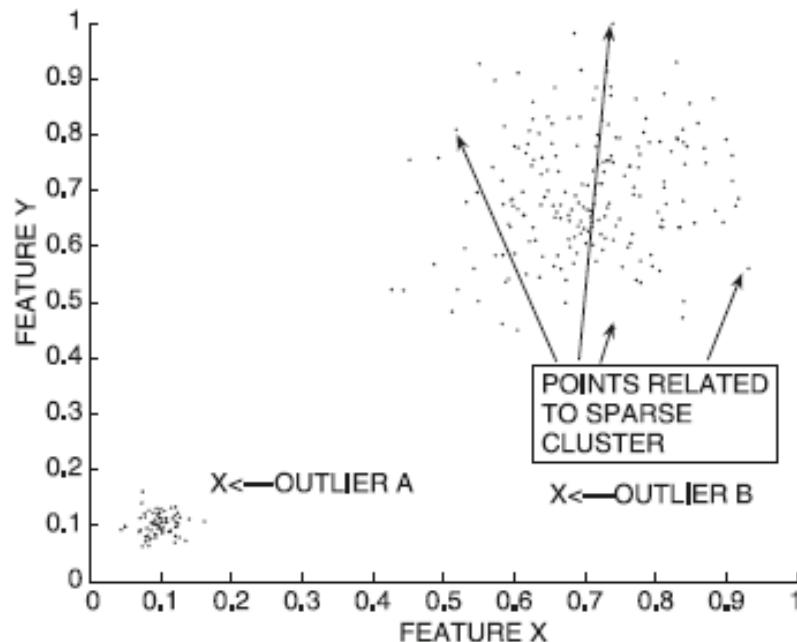
$$\max_k LMaha_k (\bar{X})$$

Il calcolo della matrice di covarianza locale introduce il concetto di “normalizzazione” della distanza

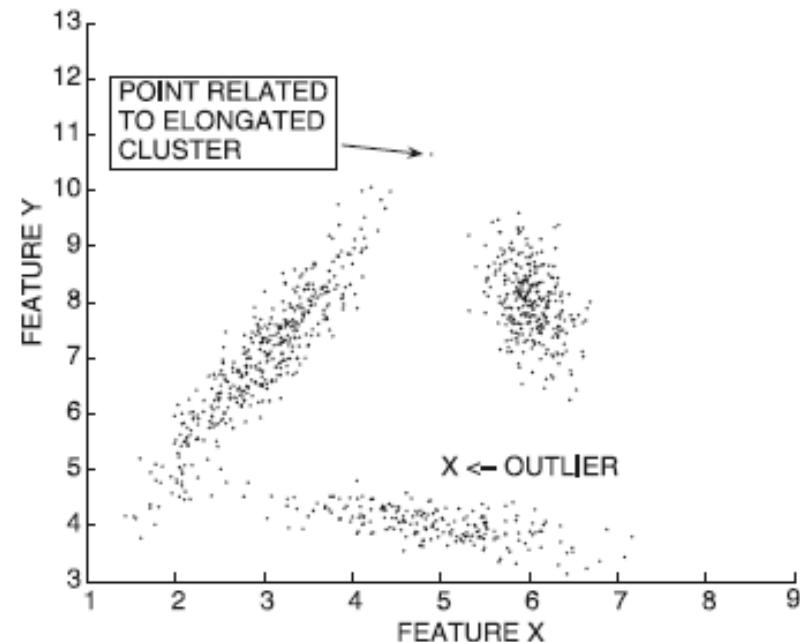
Instance-Specific Mahalanobis Distance (3)



- Can be applied to both cases



(a) Varying cluster density



(b) Varying cluster shape

- Relation to clustering-based approaches



Outline

- Introduction
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- **Density-Based Methods**
- Information-Theoretic Models
- Outlier Validity
- Summary



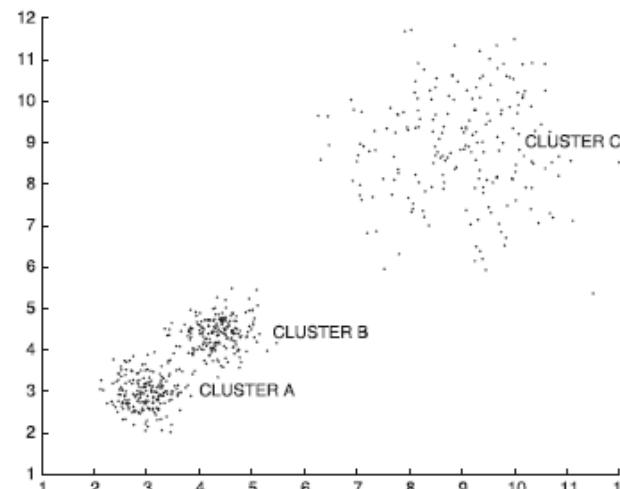
Density-Based Methods

□ The Key Idea

- Determine sparse regions in the underlying data

□ Limitations

- Cannot handle variations of density



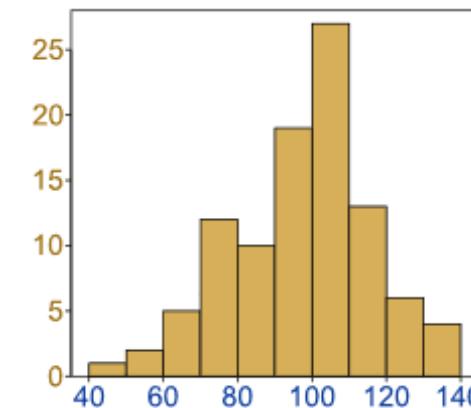


Histogram- and Grid-Based Techniques

□ Histogram for 1-dimensional data

- Data points that lie in bins with very low frequency are reported as outliers

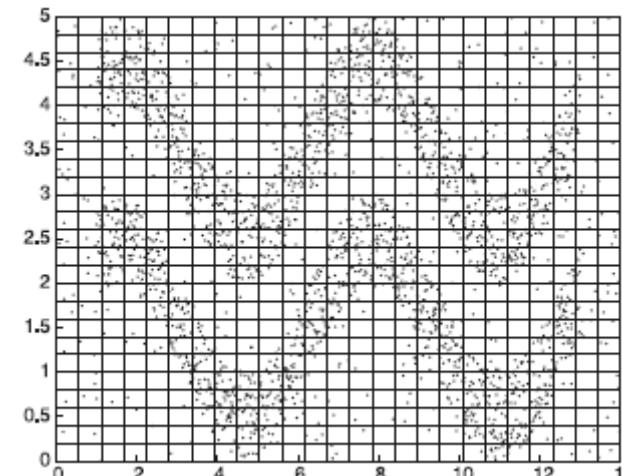
<https://www.mathsisfun.com/data/histograms.html>



□ Grid for high-dimensional data

□ Challenges

- Size of grid
- Too local
- Sparsity



Kernel Density Estimation

- Nella Kernel Density Estimation (KDE) la densità in un punto dello spazio \mathbb{R}^d viene stimata attraverso una *composizione di funzioni kernel* $K(\cdot)$ centrate nei vari punti del dataset \mathcal{D} :

$$f(\bar{X}) = \frac{1}{n} \sum_{i=1}^n K(\bar{X} - \bar{X}_i).$$

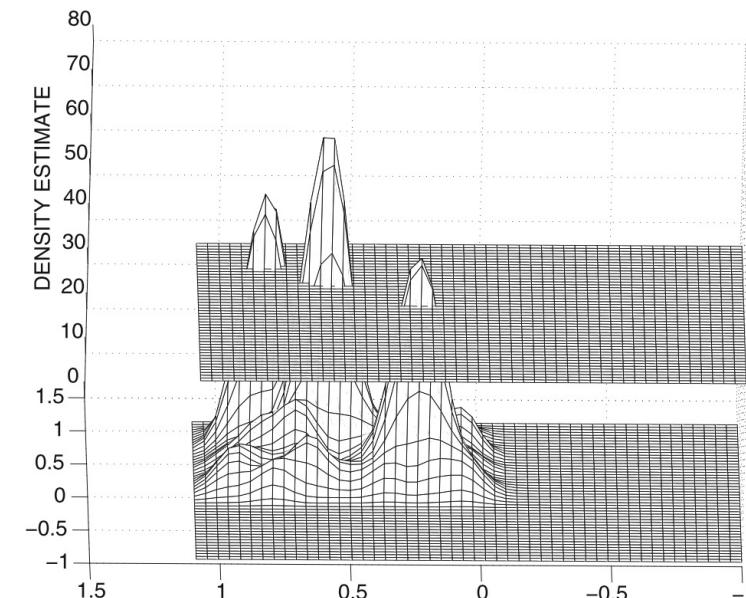
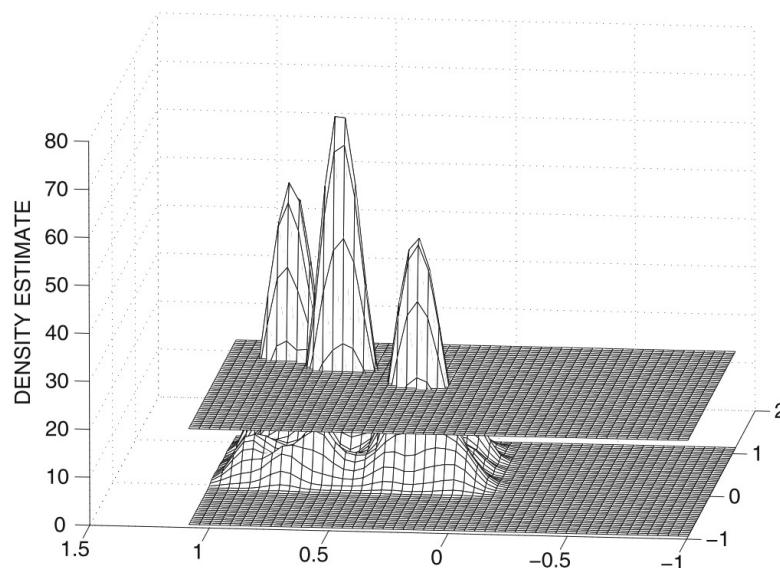
- Una scelta tipica è quella del kernel gaussiano:

$$K(\bar{X} - \bar{X}_i) = \left(\frac{1}{h\sqrt{2\pi}} \right)^d e^{-\frac{||\bar{X} - \bar{X}_i||^2}{2 \cdot h^2}}$$

- Il parametro h definisce una misura di *smoothness* della stima e viene scelto euristicamente dai dati

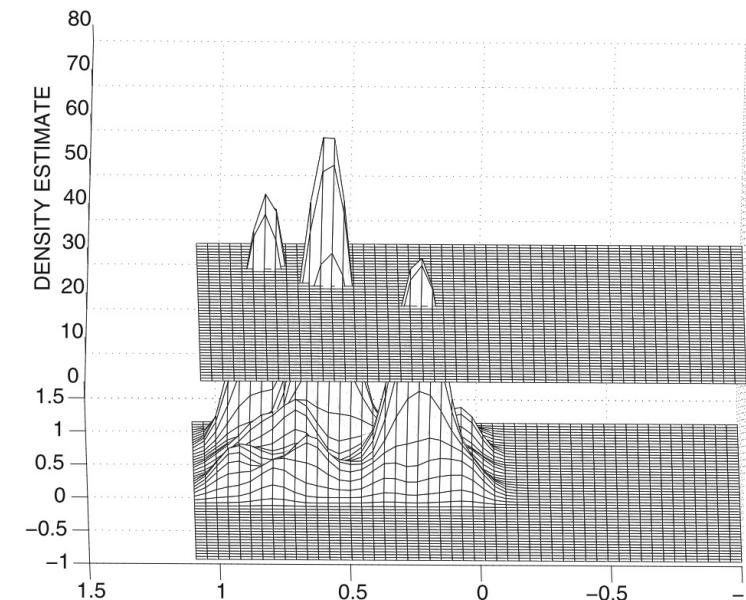
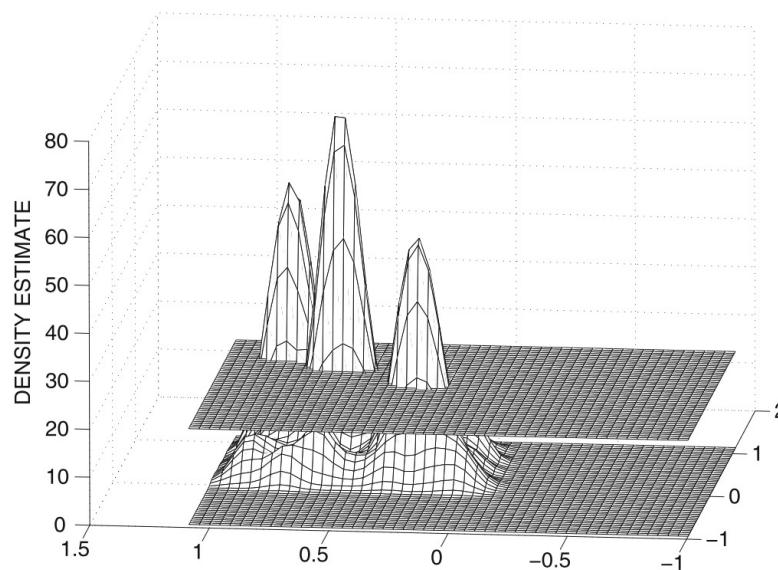
DENCLUE clustering tramite KDE

- DENCLUE cerca i cluster come i profili di intersezione della KDE con una soglia τ di densità
- Tutti i punti per cui $f(\cdot) < \tau$ sono considerati *outlier*
- La ricerca di τ è empirica: dipende dai dati ed è difficile determinarla



DENCLUE clustering tramite KDE

- DENCLUE usa il concetto di *attrattore di densità*:
 - Ogni picco di $f(\cdot)$ sarà un attrattore verso cui tendono i punti appartenenti a quel picco



DENCLUE clustering tramite KDE

- DENCLUE utilizza un approccio di tipo gradient ascent iterativo applicato ai punti di \mathcal{D} fino alla convergenza verso un massimo locale che sarà l'attrattore del cluster i -esimo

$$\overline{X^{(t+1)}} = \overline{X^{(t)}} + \alpha \nabla f(\overline{X^{(t)}})$$

$$\nabla f(\overline{X}) = \frac{1}{n} \sum_{i=1}^n \nabla K(\overline{X} - \overline{X_i})$$

DENCLUE clustering tramite KDE

- La forma del gradiente $\nabla f(\cdot)$ dipende dalla forma di $K(\cdot)$
- Il kernel gaussiano è una buona scelta perché si può mostrare che:

$$\nabla K(\bar{X} - \bar{X}_i) \propto (\bar{X}_i - \bar{X})K(\bar{X} - \bar{X}_i)$$

DENCLUE clustering tramite KDE

- Alternativamente, se ottimizziamo ponendo $\nabla f(\cdot)=0$:

$$\sum_{i=1}^n \overline{X} K(\overline{X} - \overline{X_i}) = \sum_{i=1}^n \overline{X_i} K(\overline{X} - \overline{X_i})$$

- Da cui viene una semplice regola di aggiornamento dei punti che converge più velocemente dell'altra

$$\overline{X^{(t+1)}} = \frac{\sum_{i=1}^n \overline{X_i} K(\overline{X^{(t)}} - \overline{X_i})}{\sum_{i=1}^n K(\overline{X^{(t)}} - \overline{X_i})}$$

DENCLUE clustering tramite KDE

- Complessità computazionale $O(n^2)$ come DBSCAN
 - n computazioni di $f(\cdot)$ ad ogni iterazione
- Si possono trascurare i contributi delle gaussiane centrate in punti lontani oltre $3h$
 - Si usano delle griglie per questa pre-computazione
- DBSCAN caso particolare di DENCLUE in cui:

$$K(\bar{X} - \bar{X}_i) = \begin{cases} 1, & \| \bar{X} - \bar{X}_i \|^2 < Eps \\ 0, & \text{altrimenti} \end{cases}$$



Outline

- Introduction
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- **Information-Theoretic Models**
- Outlier Validity
- Summary



Information-Theoretic Models

An Example

- The 1st One: “AB 17 times”
 - C in 2nd string increases its minimum description length

Conventional Methods

- Fix model, then calculate the deviation

□ Information-Theoretic Models

- Fix the deviation, then learn the model
 - Outlier score of \bar{X} : increase of the model size when \bar{X} is present



Information-Theoretic Models

An Example

AB17

АВАВАСАВАВАВАВАВАВАВАВАВАВАВАВАВА

- The 1st One: “AB 17 times”
 - C in 2nd string increases its minimum description length

Conventional Methods

- Fix model, then calculate the deviation

□ Information-Theoretic Models

- Fix the deviation, then learn the model
 - Outlier score of \bar{X} : increase of the model size when \bar{X} is present



Information-Theoretic Models

An Example

AB17

AB2A1C1AB14

- The 1st One: “AB 17 times”
 - C in 2nd string increases its minimum description length

Conventional Methods

- Fix model, then calculate the deviation

□ Information-Theoretic Models

- Fix the deviation, then learn the model
 - Outlier score of \bar{X} : increase of the model size when \bar{X} is present



Probabilistic Models

□ The Conventional Method

- Learn the parameters of generative model with a fixed size
- Use the fit of each data point as the outlier score

□ Information-Theoretic Method

- Fix a maximum allowed deviation (a minimum value of fit)
- Learn the size and values of parameters
- Increase of size is used as the outlier score



Outline

- Introduction
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- **Outlier Validity**
- Summary



Outlier Validity

□ Methodological Challenges

- Internal criteria are rarely used in outlier analysis
- A particular validity measure will favor an algorithm using a similar objective function criterion
- Magnified because of the small sample solution space

□ External Measures

- The known outlier labels from a synthetic data set
- The rare class labels from a real data set



Receiver Operating Characteristic (ROC) curve

- \mathcal{G} is the set of outliers (ground-truth)
- An algorithm outputs a outlier score
- Given a threshold t , we denote the set of outliers by $\mathcal{S}(t)$
 - True-positive rate (recall)

$$TPR(t) = Recall(t) = 100 * \frac{|\mathcal{S}(t) \cap \mathcal{G}|}{|\mathcal{G}|}$$

- The false positive rate

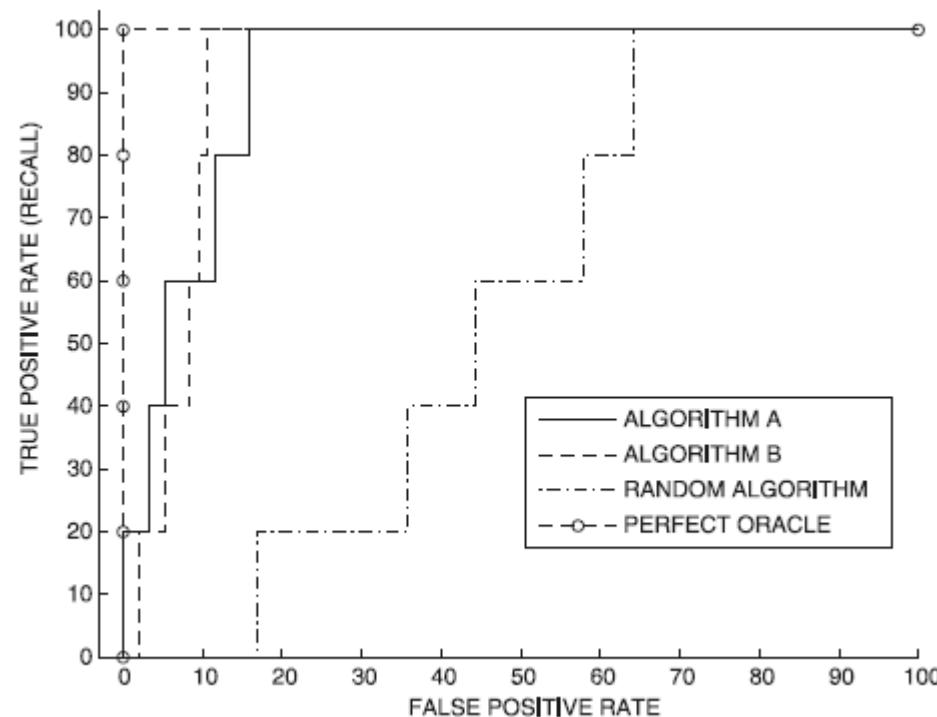
$$FPR(t) = 100 * \frac{|\mathcal{S}(t) - \mathcal{G}|}{|\mathcal{D} - \mathcal{G}|}$$

- ROC Curve
 - Plot $TPR(t)$ versus $FPR(t)$



An Example

Algorithm	Rank of ground-truth outliers
Algorithm A	1, 5, 8, 15, 20
Algorithm B	3, 7, 11, 13, 15
Random Algorithm	17, 36, 45, 59, 66
Perfect Oracle	1, 2, 3, 4, 5





Outline

- Introduction
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- **Summary**



Summary

- Extreme Value Analysis
 - Univariate, Multivariate, Depth-Based
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
 - Pruning, LOF, Instance-Specific
- Density-Based Methods
 - Histogram- and Grid-Based, Kernel Density
- Information-Theoretic Models
- Outlier Validity
 - ROC curve