Master Degree in Artificial Intelligence for Science and Technology

Anomaly Detection:

Clustering Based, Statistical Approaches and Reconstruction Based

Fabio Stella

Department of Informatics, Systems and Communication

University of Milan-Bicocca

fabio.stella@unimib.it

OUTLOOK

- Clustering Based
- Statistical Approaches
- Reconstruction Based

CLUSTERING BASED

- **ADVANTAGE**
  - unsupervised algorithm
  - existing clustering algorithms can be plugged in
- **DRAWBACKS**
  - if the data object does not have a natural clustering or the clustering algorithm is not able to detect the natural clusters, the techniques may fail
  - computationally expensive
    - using indexing structures (k-d tree, R* tree) may alleviate this problem
  - in high dimensional spaces, data is sparse and distances between any two data objects may become quite similar
  - can be difficult to decide on a clustering technique
  - can be difficult to decide on number of clusters
  - outliers can distort the clusters

CLUSTERING BASED

- **KEY ASSUMPTION**: normal data instances belong to large and dense clusters, while
  anomalies do not belong to any significant cluster.

- **GENERAL APPROACH**:
  - cluster data objects into a finite number of clusters
  - analyze each data object with respect to its closest cluster
  - anomalous data objects
    - do not fit into any cluster (residuals from clustering)
    - belong to small clusters
    - are located in low density clusters
    - are far from other data objects within the same cluster

CLUSTERING BASED: BASIC ALGORITHM

- Fixed-width clustering is first applied
  - the first data object is the center of first cluster
  - two data objects p_1 and p_2 are "near" if $d(p_1, p_2) < \mathcal{E}$ ($\mathcal{E}$ is a user specified parameter)
  - if every subsequent data objects is "near", add to the current cluster
    - otherwise create a new cluster
- Data objects in small clusters are anomalies

CLUSTERING BASED: CLUSTER BASED LOCAL OUTLIER FACTOR (CBLOF)

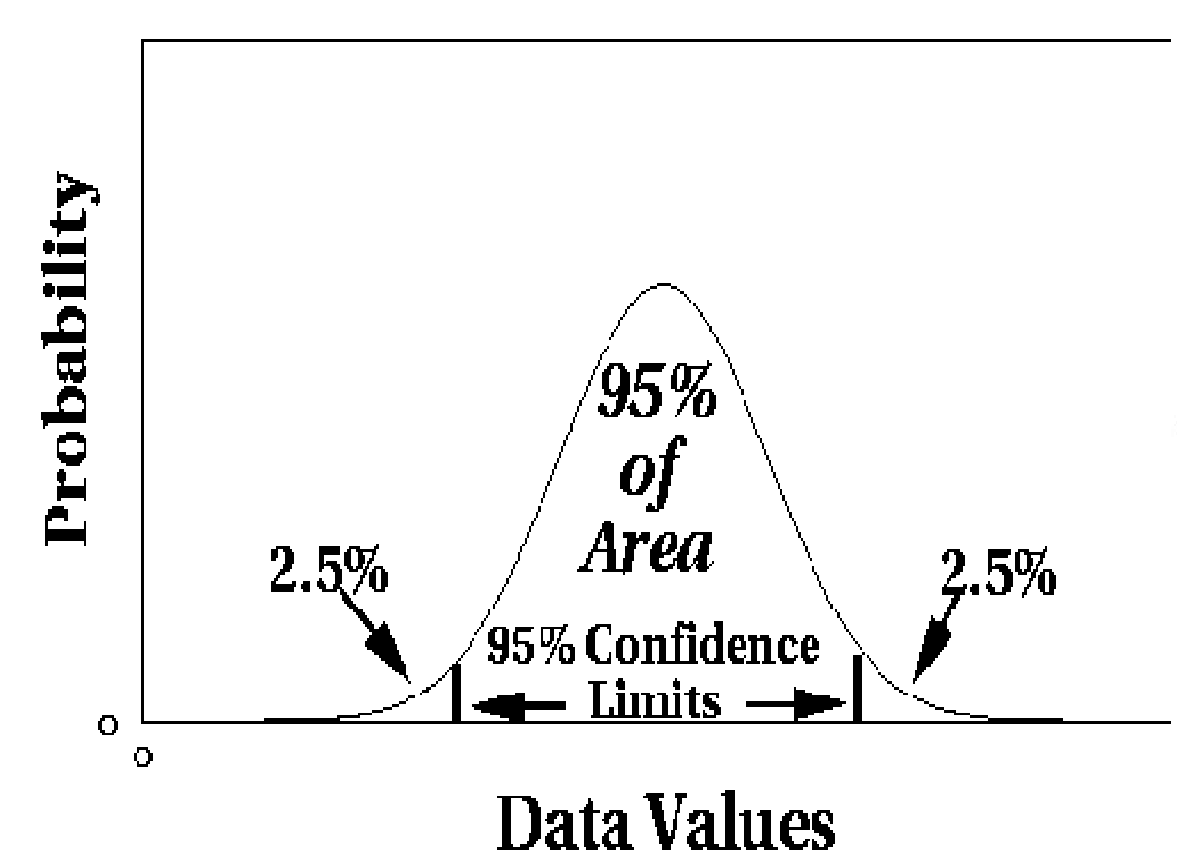
- An data object is a cluster-based outlier if it does not strongly belong to any cluster
  - for prototype-based clusters, an data object is an outlier if it is not close enough to a cluster center
    - outliers can impact the clustering produced
  - for density-based clusters, an data object is an outlier if its density is too low
    - can't distinguish between noise and outliers
  - for graph-based clusters, an data object is an outlier if it is not well connected

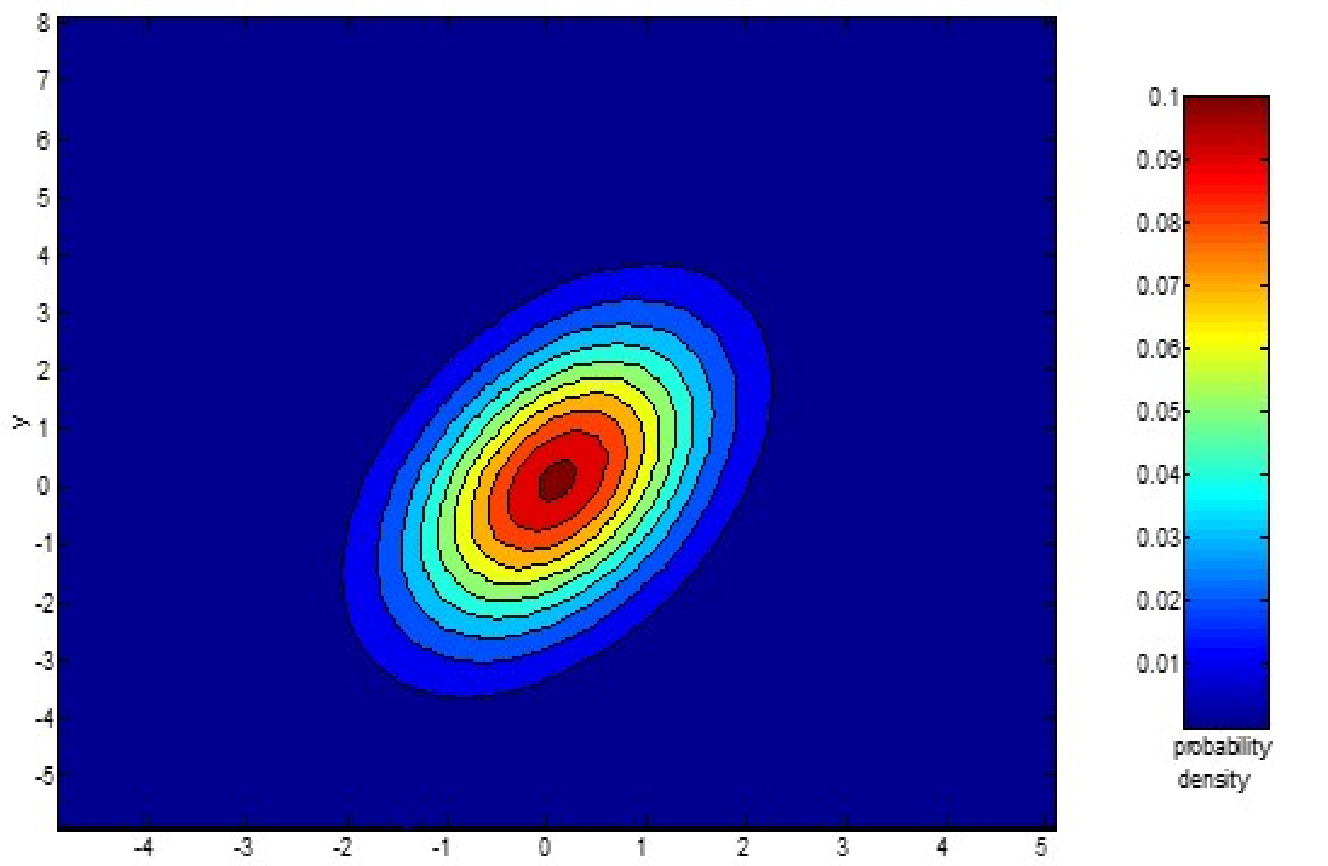
STATISTICAL APPROACHES

Probabilistic definition of an outlier: an outlier is an data object that has a low probability with respect to a probability distribution model of the data.

- Usually assume a parametric model describing the distribution of the data (e.g., normal distribution)
- Apply a statistical test that depends on
  - data distribution
  - parameters of distribution (e.g., mean, variance)
  - number of expected outliers (confidence limit)
- Issues
  - identifying the distribution of a data set
    - heavy tailed distribution
  - number of attributes
  - is the data a mixture of distributions?

STATISTICAL APPROACHES: NORMAL DISTRIBUTION





one-dimensional Gaussian

two-dimensional Gaussian

STATISTICAL APPROACHES: GRUBBS'S TEST

- Detects outliers in univariate data
- Assumes data comes from normal distribution
- Detects one outlier at a time, remove the outlier, and repeat
  - H_0: there is no outlier in data
  - H_1: there is at least one outlier
- Grubbs's test statistic: $$G = \frac{\max\left|X - \overline{X}\right|}{s}$$
- Reject H_0 if: $$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/(2N),N-2)}}{N - 2 + t^2_{(\alpha/(2N),N-2)}}}$$

STATISTICAL APPROACHES: LIKELIHOOD APPROACH

- Assumes the data set D contains samples from a mixture of two probability distributions:
  - M (majority/non-anomalous distribution)
  - A (anomalous distribution)
- General Approach:
  - initially, assumes all the data objects belong to M
  - let $LL_t(D)$ be the log likelihood of D at time t
  - for each data object x_t that belongs to M, move it to A
- Let $LL_{t+1}(D)$ be the new log likelihood
- Computes the difference, $\Delta = LL_t(D) - LL_{t+1}(D)$
- If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

STATISTICAL APPROACHES: LIKELIHOOD APPROACH

- Data distribution, $D = (1-\lambda)M + \lambda A$
- M is a probability distribution estimated from data
  - can be based on any modeling method (naïve Bayes, maximum entropy, etc.)
- A is initially assumed to be uniform distribution
- Likelihood at time t:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i)\right)\left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i)\right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

STATISTICAL APPROACHES: STRENGTHS AND WEAKNESSES

- Firm mathematical foundation
- Can be very efficient
- Good results if distribution is known
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution
- Anomalies can distort the parameters of the distribution

RECONSTRUCTION BASED

- Based on assumptions there are patterns in the distribution of the normal class that can be captured using lower-dimensional representations

- Reduce data to lower dimensional data
  - e.g. use Principal Components Analysis (PCA) or auto-encoders

- Measure the reconstruction error for each object
  - the difference between original and reduced dimensionality version

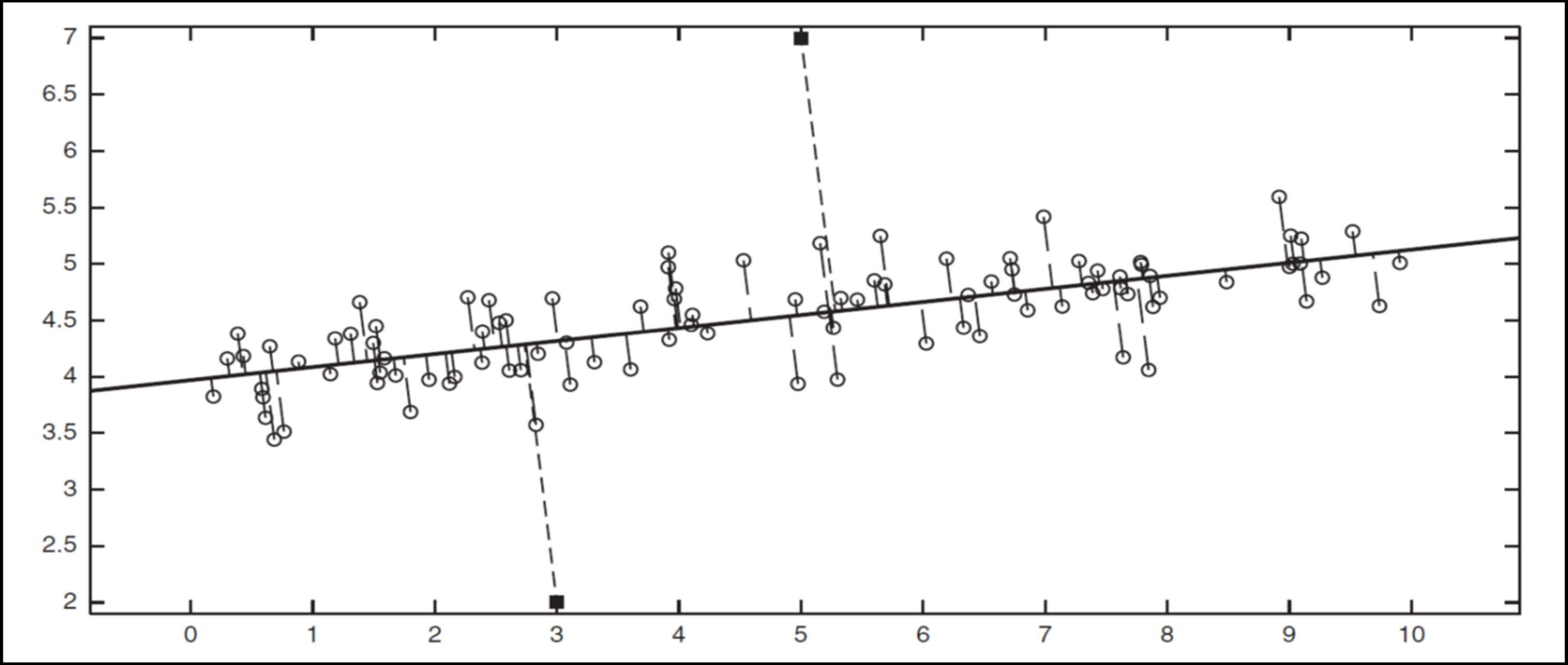
RECONSTRUCTION BASED: RECONSTRUCTION ERROR

- Let x be the original data object
- Find the representation of the data object in a lower dimensional space
- Project the object back to the original space
- Call this object $\hat{x}$

$$\text{Reconstruction Error} = ||x - \hat{x}||$$

- Objects with large reconstruction error are anomalies

RECONSTRUCTION BASED: RECONSTRUCTION OF TWO DIMENSIONAL DATA



RECONSTRUCTION BASED: PRINCIPAL COMPONENTS ANALYSIS

- Compute the principal components of the dataset
- For each test data object, compute its projection on these components
- If y_i denotes the i^th component, then the following has a chi-squared distribution

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \cdots + \frac{y_q^2}{\lambda_q} \qquad q < n$$

— an data object is anomalous, if for a given significance level α

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} > \chi_q^2(\alpha)$$

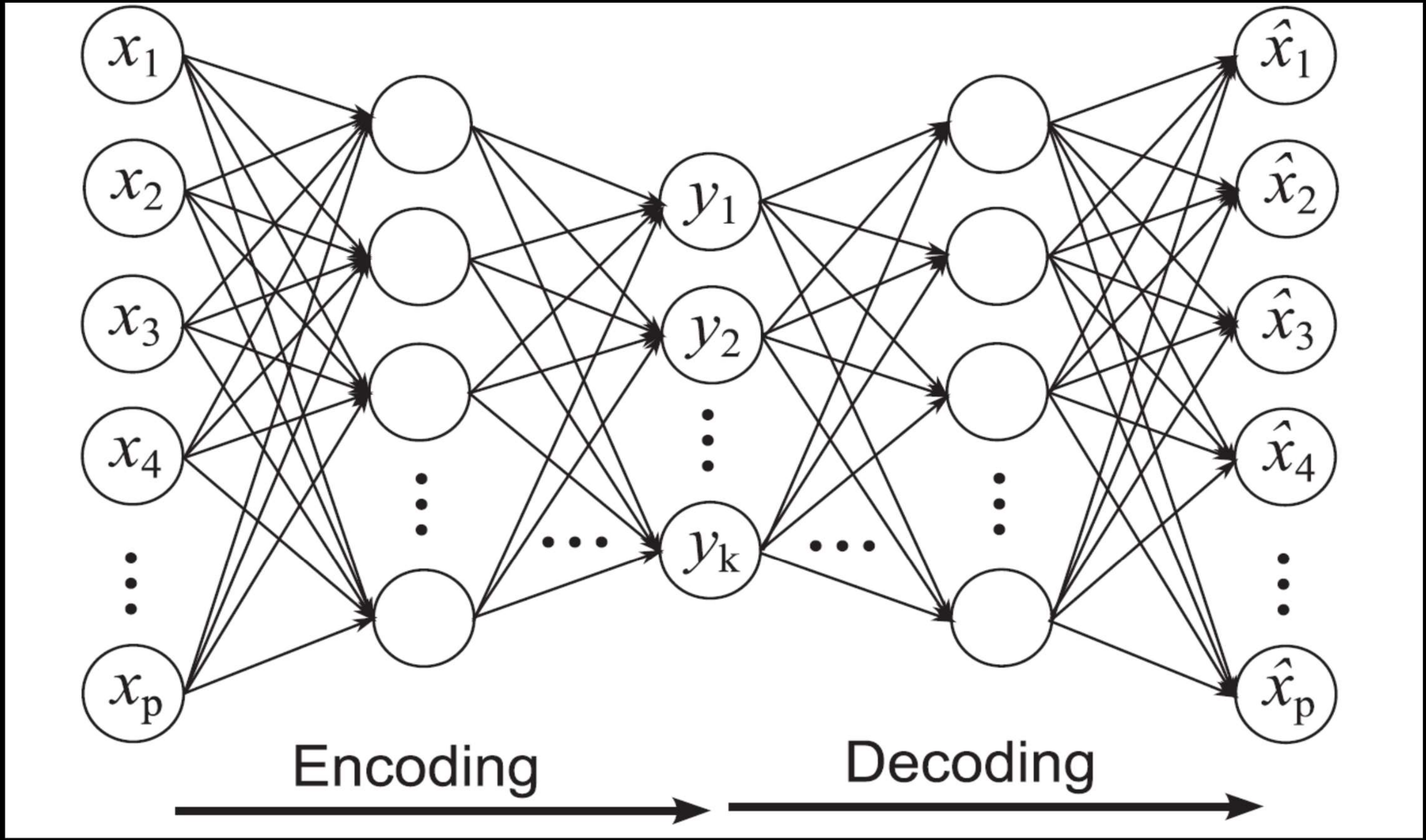
- Another measure is to observe last few principal components

$$\sum_{i=p-r+1}^{p} \frac{y_i^2}{\lambda_i}$$

— anomalies have high value for the above quantity

RECONSTRUCTION BASED: AUTO-ENCODER

- An auto-encoder is a multi-layer neural network
- The number of input and output neurons is equal to the number of original attributes



RECONSTRUCTION BASED

- **STRENGHTS**
  - — does not require assumptions about distribution of normal class
  - — can use many dimensionality reduction approaches

- **WEAKNESSES**
  - — the reconstruction error is computed in the original space
    - this can be a problem if dimensionality is high

RECAP

- Clustering Based
- Statistical Approaches
- Reconstruction Based