Unsupervised anomaly detection

Data set: $X_1, \dots, X_n \in \mathbb{R}^d$ (d = number of parameters)

- Unlabeled data set.
- Anomaly = rare

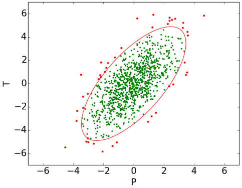
Statistical approach

- $X_1, \dots, X_n$ realizations of an unknown probability distribution P
- λ Lebesgue measure
- Find the normal region: region of minimum volume among all regions with probability greater than $\alpha \in (0, 1)$

Minimum volume set [Polonik, 1997]

$$\Omega^*_\alpha = \underset{\Omega \in \mathcal{B}(\mathbb{R}^d)}{\operatorname{argmin}} \{\lambda(\Omega), P(\Omega) \geq \alpha\}$$

(density level set with regularity assumptions)

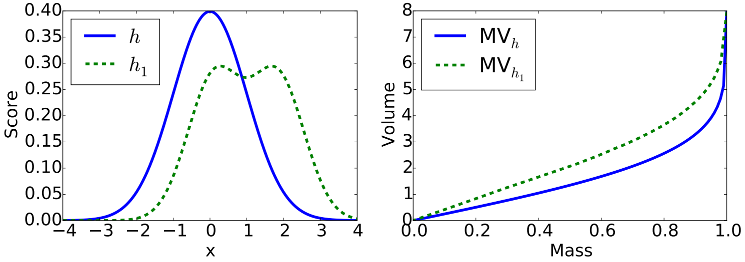


Mass Volume curve

Mass Volume curve MV_s of a scoring function s [Clémençon and Jakubowicz, 2013]:

$$t \in \mathbb{R} \mapsto (\alpha_s(t), \lambda_s(t))$$

- $\alpha_s(t) = \mathbb{P}(s(X) \geq t)$ **mass**
- $\lambda_s(t) = \lambda(\{x, s(x) \geq t\})$ **volume**



(a) Scoring functions

(b) Mass Volume curves

Mass Volume curve

MV_s also defined as the plot of the function

$$\mathrm{MV}_s : \alpha \in (0,1) \mapsto \lambda_s(\alpha_s^{-1}(\alpha)) = \lambda(\{x, s(x) \geq \alpha_s^{-1}(\alpha)\})$$

where α_s^{-1} generalized inverse of α_s.

Property [Clémençon and Jakubowicz, 2013]

Assume that the underlying density h has no flat parts. Let MV^* be the MV curve of h, then for all scoring functions s,

$$\forall \alpha \in (0,1), \quad \mathrm{MV}^*(\alpha) \leq \mathrm{MV}_s(\alpha)$$

The closer is MV_s to MV^* the better is s