

Stéphan Cléménçon

LTCI UMR Telecom ParisTech/CNRS No. 5141 - Institut Telecom

Motivation

Pairwise dissimilarity-based clustering techniques are widely used to segment a dataset into groups, such that data points in the same group are more similar to each other than to those in other groups. The empirical criteria these algorithms seek to optimize are of the form of U -statistics of degree two. We propose to analyze their performance, using recent advances in the theory of **U-processes**. The statistical framework considered permits to establish **learning rates for the excess of clustering risk** and to **design model selection tools** as well.

Statistical Framework

We observe an i.i.d. sample $\mathcal{D}_n = \{(X_i)_{i \leq n}\}$ of $n \geq 1$ observations in a space \mathcal{X} , drawn from a probability distribution $\mu(dx)$. Here, we assume that the space \mathcal{X} is equipped with a **dissimilarity measure** $D : \mathcal{X}^2 \rightarrow \mathbb{R}_+$ that fulfills the properties:

- (SYMMETRY) $D(x, x') = D(x', x)$,
- (SEPARATION) $D(x, x') = 0 \Leftrightarrow x = x'$.

The task is to partition the space \mathcal{X} in a finite number of groups, $K \geq 1$ say, so as to minimize the quantity (*i.e.* the within cluster point scatter):

$$\widehat{W}_n(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{k=1}^K \sum_{1 \leq i < j \leq n} D(X_i, X_j) \cdot \mathbb{I}\{(X_i, X_j) \in \mathcal{C}_k^2\},$$

over all possible partitions $\mathcal{P} = \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$. The **clustering risk** is:

$$W(\mathcal{P}) = \sum_{k=1}^K \mathbb{E} [D(X, X') \cdot \mathbb{I}\{(X, X') \in \mathcal{C}_k^2\}].$$

Optimal partitions are those that minimize $W(\mathcal{P})$.

Generalization Ability

Pairwise-based clustering can be cast in terms of **minimization of a U -statistic** over a class Π of partition candidates.

Analysis of the Empirical Clustering Risk Minimizers requires to study the fluctuations of the **U-process** $\left\{ \widehat{W}_n(\mathcal{P}) - W(\mathcal{P}) : \mathcal{P} \in \Pi \right\}$. Key ingredients:

- COMPLEXITY ASSUMPTION

$$\sup_{\mathcal{C}, \mathcal{P}} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \epsilon_i D(X_i, X_{i+\lfloor n/2 \rfloor}) \cdot \mathbb{I}\{(X_i, X_{i+\lfloor n/2 \rfloor}) \in \mathcal{C}^2\} \right|,$$

- PROJECTION TECHNIQUES

$$U\text{-process} = \text{Empirical Process} + O(1/n)$$

- **Learning rates** of the order $O(1/\sqrt{n})$
- **Tight probability bounds** for the excess of clustering risk
- **Fast rates**
- **Model selection** tools, computing additive complexity penalization terms:
 - ▶ Automatic selection of the **geometry** of the cells
 - ▶ Choosing the **number** of cells