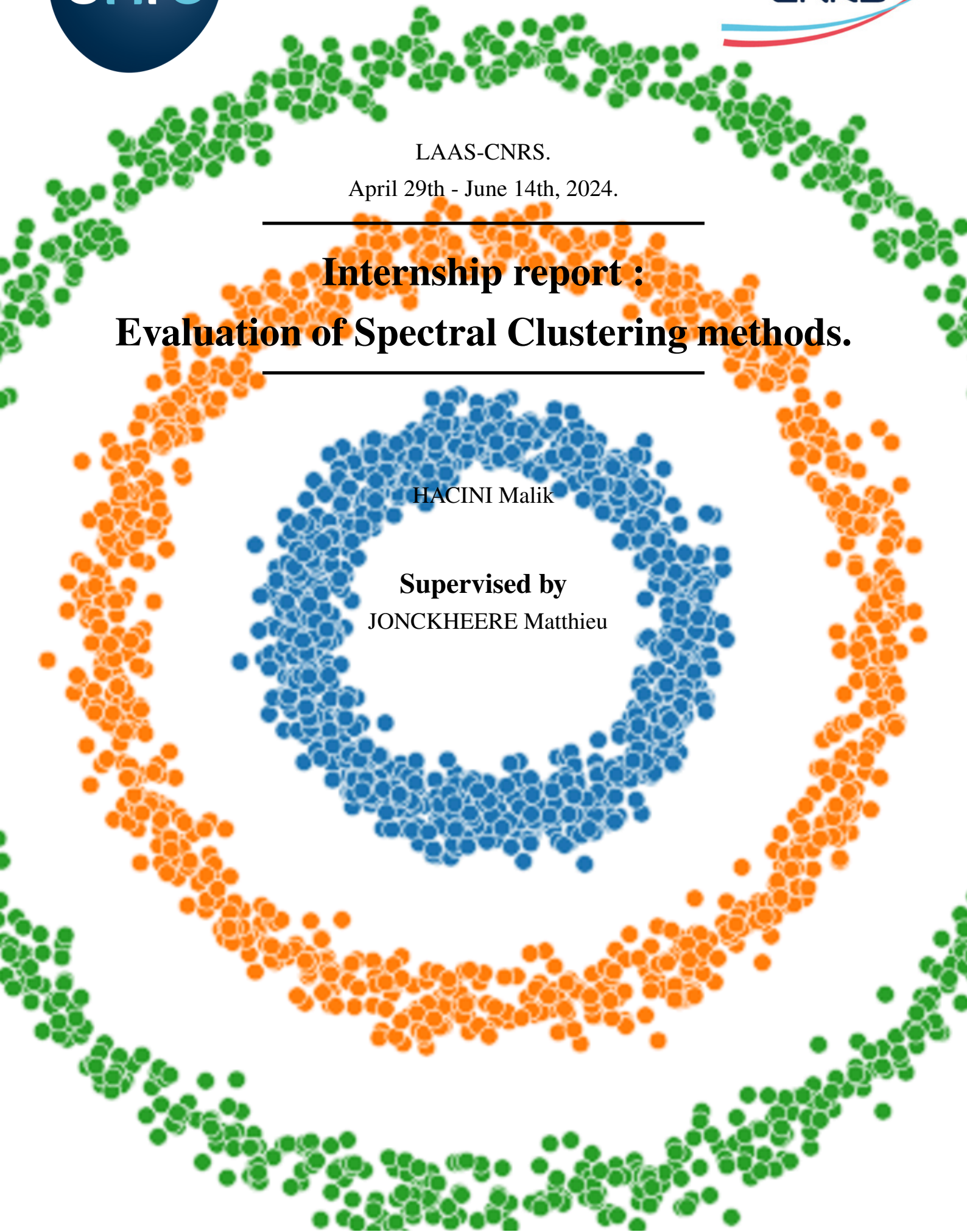

LAAS-CNRS.

April 29th - June 14th, 2024.

Internship report :
Evaluation of Spectral Clustering methods.

HACINI Malik

Supervised by

JONCKHEERE Matthieu

Abstract

Placeholder

Contents

1 Introduction

Clustering data is crucial in various fields as it allows us to identify patterns, group similar entities together and derive meaningful insights from large datasets. In virtually every scientific field dealing with empirical data, people attempt to get a first impression on their data by trying to identify groups of "similar behavior" in their data. These fields include machine learning where it is referred to as unsupervised learning, healthcare where clustering techniques are used to identify groups of patients with similar medical characteristics facilitating the conducting of studies and marketing where clustering allows businesses to segment customers based on arbitrary chosen criteria, often used to personalize advertisement. To perform clustering, algorithms are built using mathematical tools from probability theory, statistics, linear algebra and functional analysis.

A classic approach is the k-means algorithm, which aims at partitioning the data into k predefined clusters centered at specific points and minimizing within-cluster variances. However, this problem is computationally NP-Hard and although efficient heuristics exist, k-means performs poorly on datasets with inadequate geometry or very high dimensional ones. Moreover, the number k of clusters need to be known before performing the algorithm, which can be a big problem. A new theoretical framework for clustering has developed in the past 25 years with the goal of overcoming these limitations. It aims to identify clusters based on the similarity between data points. It views the data points as nodes in a graph and analyzes the graph's structure to partition the data. This analysis uses spectral properties of the graph, leading to the name of **spectral clustering**. In the last 3 decades, spectral clustering has become one of the most widely used clustering methods due to its simplicity, efficiency, and strong theoretical background.

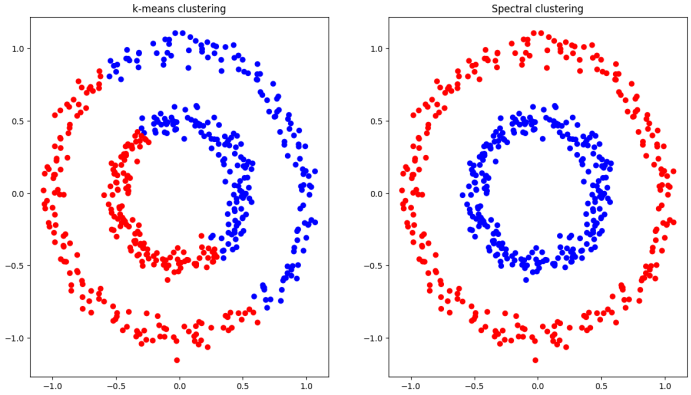

Figure 1: Comparison between k-means clustering and spectral clustering on a toy dataset

However, due to it's relative recency and important practical applications, spectral clustering is still heavily studied today. It still has theoretical limitations that hinders it's performance. Improving the theory behind these algorithms is the goal of a team of researchers at LAAS-CNRS in Toulouse. This report presents the internship I have done in their team over the course of 6 weeks.

The goal of the internship was understanding classical spectral clustering (SC) in theory, then dive into **generalized spectral clustering** (GSC), the renewed theory of SC that the team is working on. I could then try helping the team by implementing SC and GSC on synthetic and real datasets, aiming to conduct experiments on the performance of these algorithms. This would hopefully lead to guiding theoretical research towards methods with the best experimental results, as good experimental results may be the sign of the existence of good theoretical results. This internship allowed me to reinvest the mathematical tools learned during La Prépa in an Applied Mathematics context. To deeply understand the theory of SC and GSC, I also had to learn new mathematics, mostly in probability theory and functional analysis. I was also able to practice my Python programming skills and learn the art of presenting experimental results in a scientific way. Most importantly, I was fully involved in the team, working at the lab and sharing everyday with it's researchers from all around the globe, discovering the world of academics.

2 An overview of LAAS-CNRS

LAAS-CNRS is a french research lab of the *Centre National de la Recherche Scientifique* (CNRS), the biggest public research organism in France. LAAS stands for *Laboratoire d'analyse et d'architecture des systèmes* (Laboratory of system analysis and architecture). Behind this rather complex acronym lies 4 historical disciplinary fields : computer science, robotics, automatics and micro and nano systems. The 'systems' considered in LAAS' research activites are of diffrent kinds :integrated systems, robotic systems, biological systems...

They fall in various application domains such as aeronautics and space, telecommunications,

transports, production, services, security and defense, energy management, healthcare, environment and sustainable development.

2.1 Foundation

LAAS was created in 1968 under the name *Laboratoire d'automatique et de ses applications spatiales* (Laboratory of automatics and it's spatial applications). Indeed, it is located in Toulouse, a leading city in spatial technology, near other important academic entities such as ENAC or CNES (National Centre for Space Studies).

Figure 2: LAAS-CNRS Facility in Toulouse

2.2 Lab organization and philosophy

LAAS is the home of 6 research departments made up of 26 teams dedicated to their 4 disciplinary fields. All departments combined, over 800 people work at LAAS, including 400+ permanent researchers.

Figure 3: LAAS' 6 research departments.

The lab has a history of strong relationships with industry and works in a large number of collaborative projects with international, national and regional industries of all size. LAAS was one of the 20 first "Carnot Institutes" labeled in 2006, a label given to labs putting an emphasis on industry partnership. LAAS also takes great advantage of it's pluridisciplinary nature : teams from different departments often collaborate to build projects.LAAS promotes transdisciplinary research through 4 strategic axes: Ambient intelligence, Living (biology, environment, medicine), Space and Energy. Examples include LAAS' mathematicians teaming up with robotics researchers to provide efficient Machine Learning (ML) algorithms for their projects.

2.3 Daily life