Stage Master 1
Single cell clustering with sparse feature selection

M. Barlaud, A. Malapert and J.C. Regin
University cote d'azur
Lieu du stage I3S, 2000 Route des Lucioles, F-06903 Sophia-Antipolis
Contacts: barlaud@unice.fr , jean-charles.regin@unice.fr

January 30, 2018

1 An Atlas of Airways at a single cell level (AAAsc)

This project is proposed in the context of a pilot project supported by the Chan Zuckerberg Initiative, entitled 'An Atlas of Airways at a single cell level (AAAsc)', in the context of an international consortium to establish a Human Cell Atlas. The idea is to apply RNA sequencing technologies to the profiling of single cells. Typical experiments currently generate complex profiles of expression for thousands of cells, and future projects will rapidly reach millions of cells. Data analysis methods and software specifically designed to account and model these types of data are now required for proper analysis and interpretation of the biological results. The current project is one of the Master Environns of the Life Science Academy of UCA. It is articulated around 2 themes: the first one aims at using these new methods to develop a gene expression compendium of human airway cells collected by bronchoscopy, and use these reference sets to elucidate differences between pathological and normal airway cells (Pascal Barbry and Agnes Paquet, IPMC), the second one, presented here, aims at developing new mathematical methods for clustering with feature (gene) selection tailored to large datasets.

2 Clustering algorithm

This project deals with machine learning in computational biology. The objective is clustering with feature selection on single-cell RNA sequencing datasets. The main issue is high dimensional features (20,000 genes) since clustering in high

dimension suffers from the curse of dimensionality: As dimensions increase, vectors become indiscernible and the predictive power of the aforementioned methods is drastically reduced. In order to overcome this issue, a popular approach for high-dimensional data is to perform Principal Component Analysis (PCA) prior to clustering using kmeans. This approach is however difficult to justify in general **?**. To address this combinatorial non-convex problem maintaining a strict control on the sparsity of W, we follow an alternating minimization of the Frobenius norm criterion. We provide a new efficient algorithm K-sparse algorithm **?** which alternates a kmeans step and a projection-gradient step.

3 Objectives of the project

- Optimization of the K-sparse algorithm (Matlab) **?**
- Comparison (in term of accuracy, scalability) of K-sparse with state of the art clustering methods **?**
- Develop an R package of K-sparse