

# STA 663 Project - Scalable K-means++

Bai Li, Jialiang Mao

April 5, 2016

## Abstract

K-means algorithm is one of the most popular clustering algorithm. A crucial part of k-means algorithm is the choice of initial centers while a poor initial centers may lead to locally optimal solution. To beat this, the k-means++ initialization is proposed to obtain an initial set of centers that is close to the global optimum solution. However, because of its sequential nature, the k-means++ is not scalable. The paper introduces a scalable k-means++ algorithm by reducing the number of passes needed to obtain a good initialization. In this project, we implement this algorithm and test its performance with simulation studies.

## 1 Introduction

## 2 Implementation

## 3 Testing

## 4 Optimization

## 5 High performance computing

## 6 Application and comparison

## Appendix