Weekly Report

WU Zihan

September 21, 2023

Outline

Method for Noisy Data

S_{th} denotes the thereshold for finding a co-cluster in a submatrix; T_p denotes the times of reparitioning.

- Step 1: Decide S_{th} with confidence level $1 - \beta_1$
- Step 2: Decide T_p with confidence level $1 - \beta_2$
- $p = (1 - \beta_1)(1 - \beta_2) \geq 1 - (\beta_1 + \beta_2)$
- Select $\beta_1 = \beta_2 = 0.005$, then $p \geq 0.99$

Results on Noisy Data

Settings

- $M = N = 10^4$
- $M_k = N_k \in \{50, 60, \ldots, 190\}$
- $\kappa = \dfrac{\sigma^2}{B_{\max}} \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$

Results

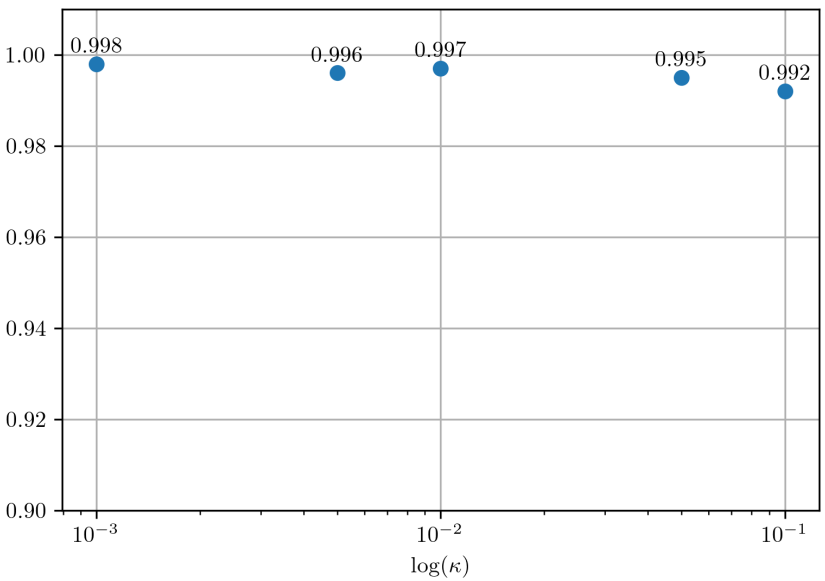
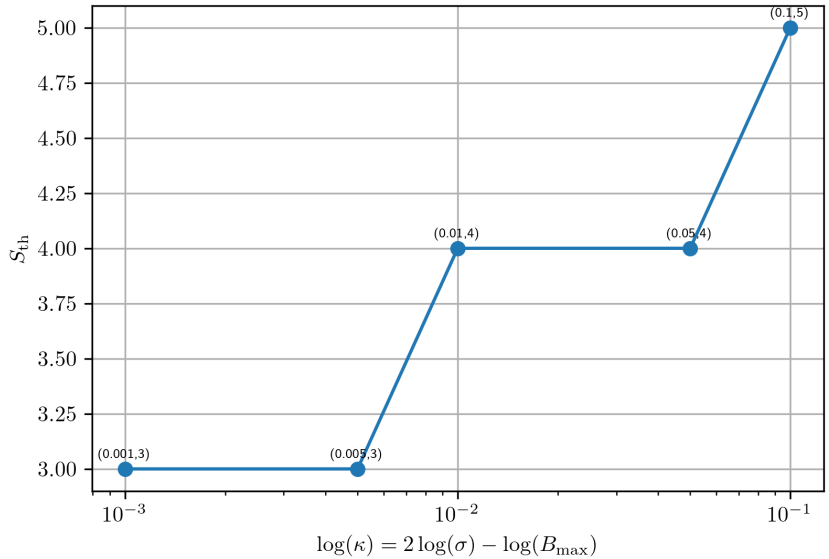
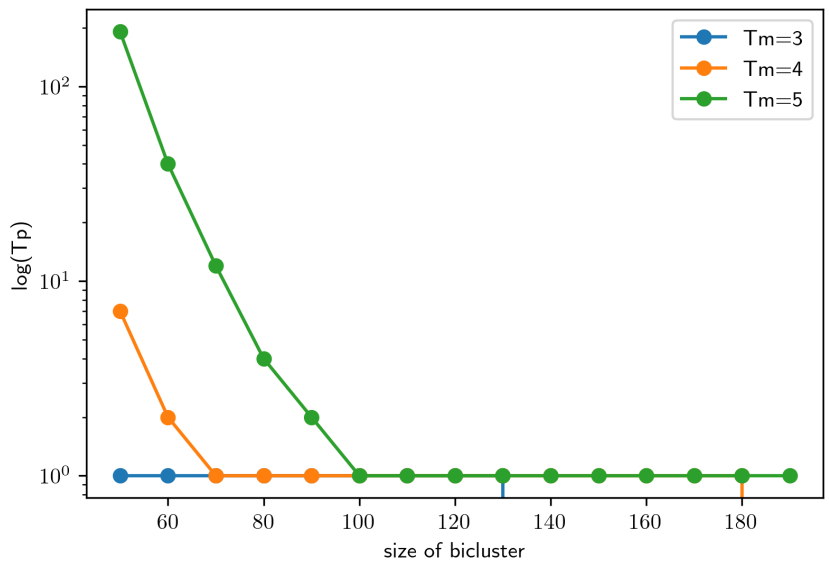


Figure: Results on Noisy Data



(a) Decide S_{th}



(b) Decide T_p

Application

- **Big Data and NLP:** Partitioning allows for smaller memory requirements, making it highly efficient for handling big data. This is particularly useful in parallel computing environments. Natural Language Processing (NLP) can greatly benefit from this, as it often involves processing large datasets like text corpora or social media feeds [1]
- **Real-Time Data Processing:** Partitioning is also conducive for batch stream processing. This is advantageous for real-time data processing applications. For instance, stream data such as recommendation systems can be processed in real-time, allowing for quicker decision-making and analytics [2]


Séverine Affeldt, Lazhar Labiod, and Mohamed Nadif.
Ensemble block co-clustering: A unified framework for text data.
In *International Conference on Information and Knowledge Management*, pages 5–14, Virtual Event Ireland, France, October 2020. ACM.


Laiwen Zheng, Hong Huo, Yiyou Guo, and Tao Fang.
Supervised Adaptive Incremental Clustering for data stream of chunks.
Neurocomputing, 219:502–517, January 2017.