Research Report

April 22nd - May 12th

Han Chen

Tanwei Colledge, Tsinghua University

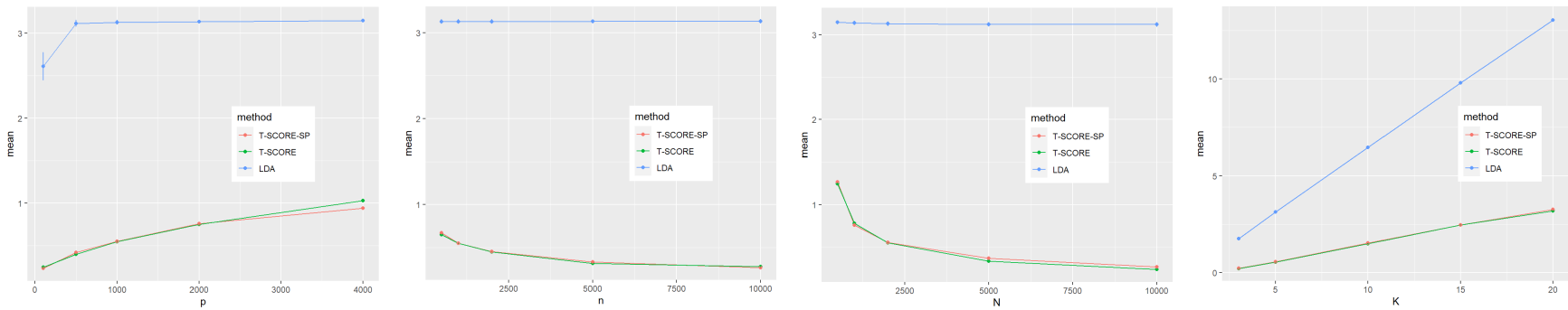
May 12, 2023

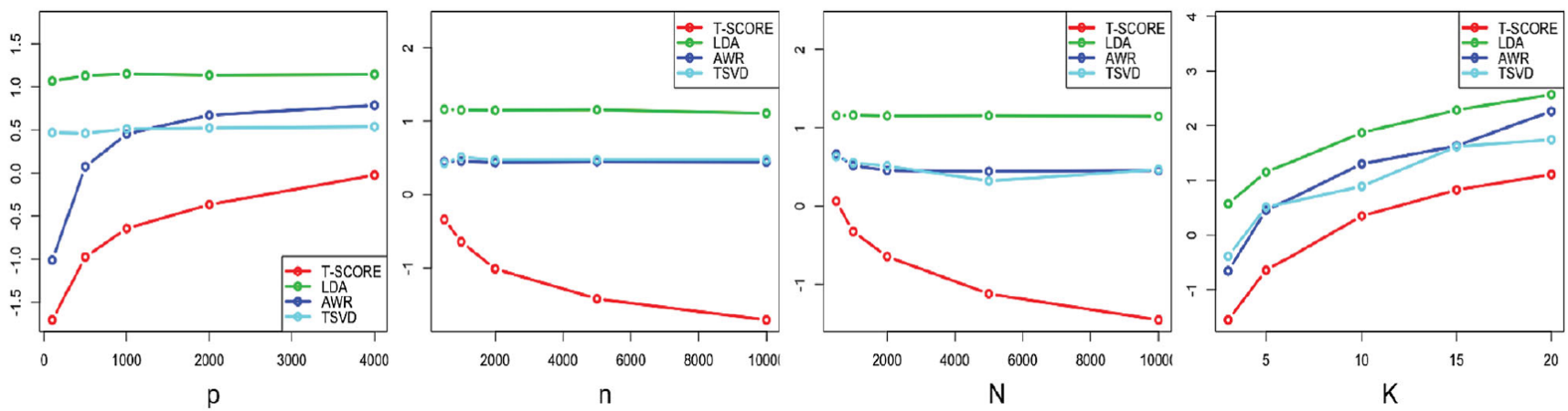
Overview

- Reproduced some simulation results
- Read an article on LDA (Blei, Ng, and Jordan 2003)

Reproducing the simulation results

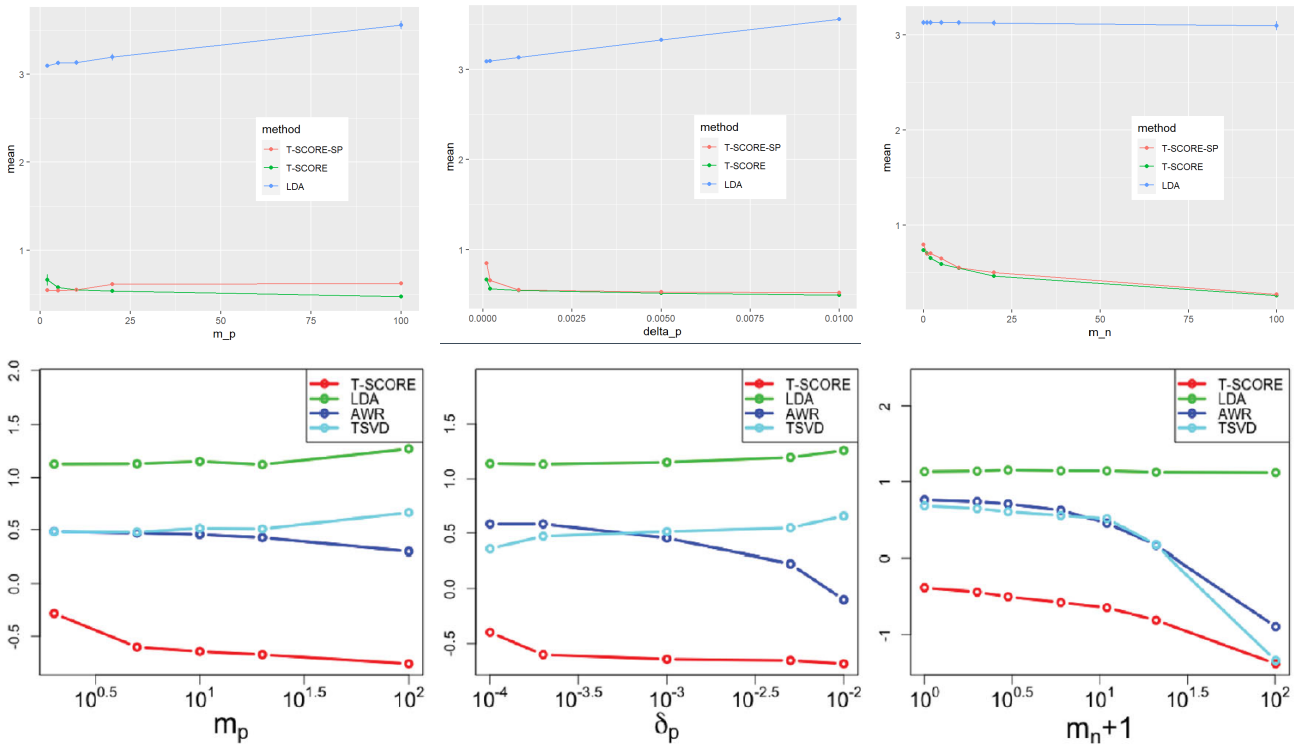
- Above: reproducing results (Topic-SCORE & LDA)
- Below: original results
- TSVD: still debugging
- AWR: not accessible





Reproducing the simulation results (Cont'd)

- Above: reproducing results (Topic-SCORE & LDA)
- Below: original results
- TSVD: still debugging
- AWR: not accessible



Application on real-world data

- Reproduction on AP data
  - **Class I:** reagans, truly, shelter, rotation, forecasters, mice, dingell, suing, costume, brig, liquidation, cxcbt, mess, bands, guardians, easter, wittman, fried, eyewitnesses, renounce
  - **Class II:** testified, reagans, truly, shelter, rotation, forecaster, intensified, mice, dingell, suing, costume, liquidation, cxcbt, bands, connie, wittman, deserts, fried, eyewitnesses, renounce
  - **Class III:** testified, vehicles, match, truly, enacted, proud, forecasters, intensified, mice, retaliation, fernando, dingell, nida, okla, silicon, surrendered, verne, quaker, eyewitnesses, renounce
- However, the original results are

Table 1. Top 15 representative words for each estimated topic in the AP data ($K = 3$). In the word list for the "Finance" topic, *rose* is the past tense of *rise*.

"Crime"	*shootings, injury, mafia, detective, bangladesh, dog, hindus, gunfire, aftershocks, bears, accidentally, handgun, unfortunate, dhaka, police*
"Politics"	*eventual, gorbachevs, openly, soviet, primaries, sununu, yeltsin, cambodia, torture, soviets, herbert, gephardt, afghanistan, citizenship, popov*
"Finance"	*trading, stock, edged, dow, rose, traders, stocks, indicators, exchange, share, guilders, bullion, lire, christies, unleaded*

Personal comments

- The simulation results may be biased (?), since the generation of data favors pLSI model, rather than LDA.
- Application on EHR: the SVD method can only learn patterns from given data, rather than make prediction (?).

Future work

- Topic modeling: complete the reproduction; more reading.
- Pseudo time series?

References

Using SVD for topic modeling.

Zheng Tracy Ke & Minzhe Wang.

Journal of the American Statistical Association, 2022.

Latent dirichlet allocation.

Blei, D. M., Ng, A. Y., & Jordan, M. I.

Journal of machine Learning research 3(Jan), 2003.