CORD-19 DATASET

AN APPLICATION OF CLUSTERING METHODS

Quynh Anh Nguyen

University of Milan

TABLE OF CONTENTS

INTRODUCTION

- Dataset The data set is CORD-19, resource of over 300,000 scholarly articles about COVID19. A subset of 1000 articles will be used in this project.

- Goal Comparing the results of two methods, K-Means and GMM method on clustering 1000 articles and extracting the topic of each cluster.

METHOD

- Data processing from text into 966×4096 matrix.
- PCA The number of predictors in the data is much larger than the number of data-points. We apply PCA to reduce the dimension of the data.
- Finding the best K by Elbow method and Gap Statistics
- Applying K-Means and GMM for the 2-dimensional projected data.
- Finding the optimal number of clusters for GMM method by AIC and BIC indicator.
- Evaluating the performance of both methods by extracting keywords of each cluster and check whether they are describing for a topic.

Tuning the optimal number of cluster

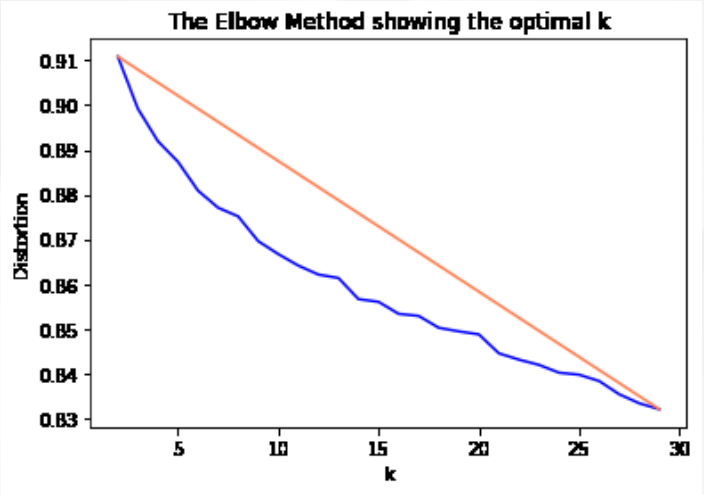

Figure: Elbow method.

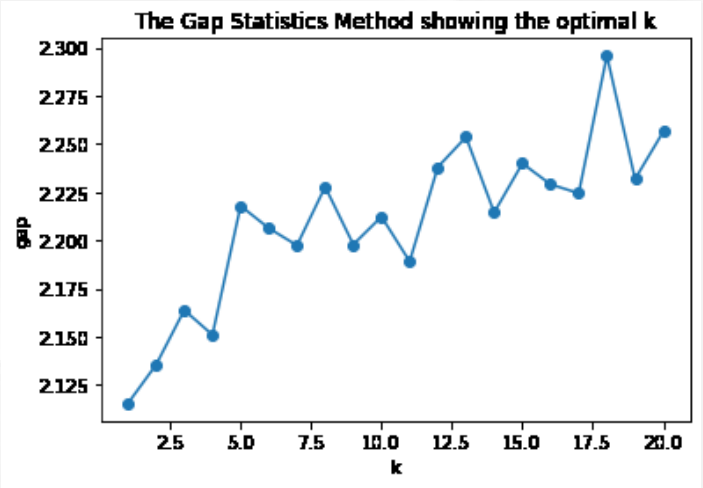

Figure: Gap Statistics method.

Dataset: 966×708 matrix

Distortion: The sum of squared distances from each point to its centroid is computed and plotted against k.

Gap Statistics:

$$Gap(K) = \log W(K) - \log W_{uniform}(K) \tag{1}$$

Keywords of 17 clusters by K-Means

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11
0	[water, chamber, inactivation, dose, power, produce, nozzle, person, spray, radiation, distance, size, solution, provide, device, uv-c, sanitization, lamp, box, cough, procedure, disinfectant, time, sars-cov-, hand, patient, droplet, line, method, air, large, study, sterilization, particle, position, worker, value, world	[patient, infection, case, medical, life, school, social, risk, respondent, survey, mental, future, increase, surgical, support, factor, procedure, service, provider, physical, country, guideline, delivery, home, time, death, datum, anxiety, practice, impact, pediatric, report, pandemic, family, stress, experience, study, woman, present, use, article	[day, kit, concentration, control, laboratory, diagnostic, distribution, cdc, numb, sensitivity, diagnosis, specimen, viral, patient, study, sars-cov-, temperature, rt-pcr, amplification, swab, disease, infection, process, datum, symptom, negative, antibody, report, virus, high]	[protein, vaccine, cell, human, sars-cov, detection, mutation, test, sample, approach, strain, genome, assay, asymptomatic, neutralization, igg, titer, subject, infectious, day, experiment, rate, rbd, ace, wuhan, outbreak, symptom, library, gene, nsp, bind, immune, case, sequence, model, respiratory, rna, host, receptor, antibody, patient]	[treatment, severe, test, surgery, level, need, symptom, swab, viral, report, sars-cov-, study, preoperative, depression, health, pandemic, mental, ocular, breast, manifestation, group, stroke, risk, contact, particle, adverse, hfnc, healthcare, image, score, worker, information, covid, fracture, mask, hip, exposure, tongue, mucus, eye, bronchoscopy, concentration, systemic, positivity, aerosol, medicine, radiologist, hcws, radiology, center	[patient, people, care, individual, police, refugee, human, social, technology, train, node, remote, prison, right, video, trainee, speed, cluster, release, road, vehicle, violence, waste, article, reader, smart, design, model, lockdown, mental, participant, effect, body, report, leave, office, hug, worker, support, belief, professional, mayor, bag, extract, conspiratorial, military, uncertainty, room, woman, ambulance, parliament,	[test, day, cluster, infect, air, infection, epidemic, state, concentration, period, lockdown, transmission, country, disease, analysis, ventilator, prediction, method, traffic, flow, government, passenger, india, tree, population, present, sequence, connect, local, network, coronavirus, search, provide, google, generation, type, viral, frequency, model, equation, individual, differential, secondary, uncertain, emission	[case, people, infection, test, country, anxiety, participant, variable, social, numb, disease, physical, rate, period, symptom, depression, vaccine, person, special, return, respondent, sars-cov-, age, mental, need, knowledge, survey, service, model, city, report, surveillance, optimal, transmission, hazard, example, individual, household, question, protocol, vulnerability, region, epidemic, patient, conduct, collect,	[print, design, mask, food, export, household, income, trade, bank, international, https, china, producer, article, manufacture, www, firm, policy, liability, resilience, disruption, industry, india, sector, company, produce, consumer, ventilator, technology, fresh, impact, home, global, risk, medical]	[institution, virus, recommendation, increase, icu, guideline, recommend, guidance, on-site, procedure, support, control, infection, available, measure, supply, pharmacist, resource, service, team, medication, treatment, work, study, nurse, staff, department, document, time, april, plan, unit, case, information, report, event, previously, therapy, suggest, rely, entire, exposure, similar, set]	[sign, pediatric, pulmonary, score, ggo, rsna, radiograph, initial, male, x-ray, discharge, function, adult, sensitivity, hrct, woman, man, severe, child, decision, involvement, extent, scan, dose, radiation, mass, visible, vessel, abnormality, distribution, typical, negative, asymptomatic, lobe, air, right, high, -ncov, rt-pcr, pattern, severity, positive, day,

Figure: Keywords of clusters by K-Means

WordCloud of clusters by K-Means

Figure: Word-Cloud from Keywords of Clusters by K-Means Method

Plot cluster by K-Means and GMM method

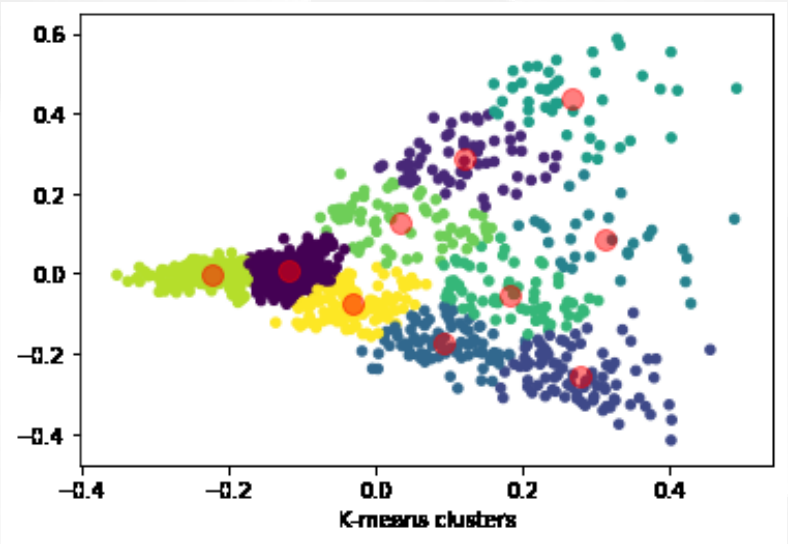

Figure: K-Means Method.

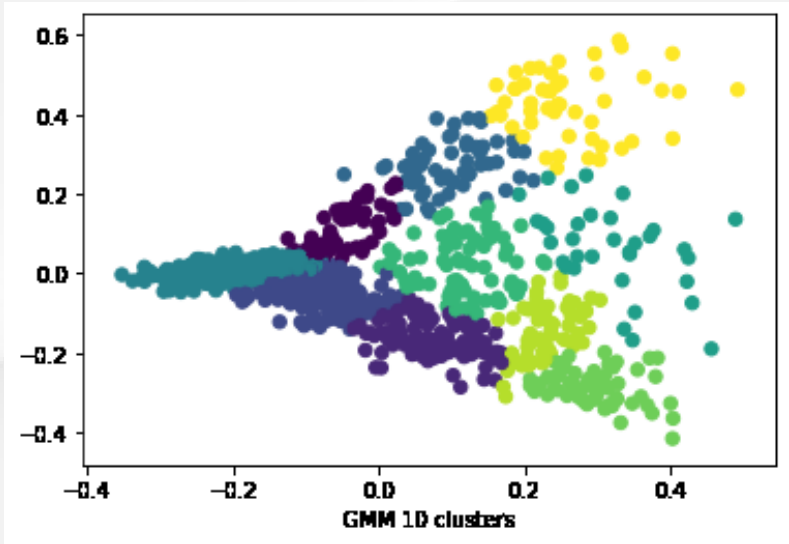

Figure: Gaussian Mixture Method.

Dataset: 966×2 matrix

K-Means method lacks of flexibility in cluster shape and lack of probabilistic cluster assignment. This weakness is improved in GMM method.

Finding optimal component with BIC and AIC indicators

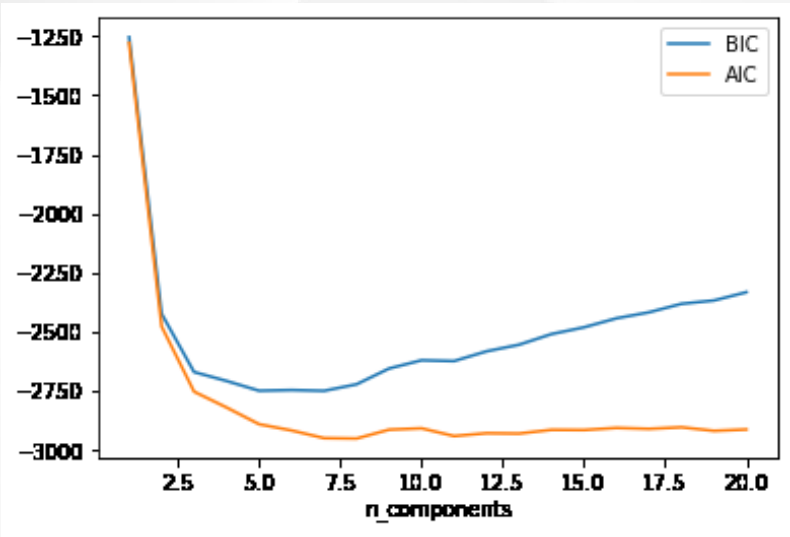

Figure: AIC and BIC index.

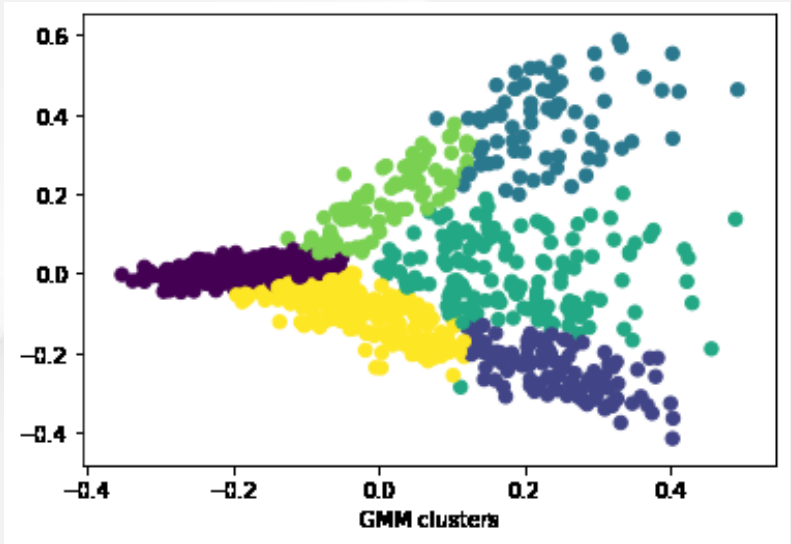

Figure: n_component is 6.

Dataset: 966 $\times$ 2 matrix

The optimal numberof clusters is the value that minimizes the AIC or BIC.

WordCloud of clusters by GMM method

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
0	[vaccine, sequence, antibody, clinical, immune, interaction, genome, mutation, site, transmission, spread, bat, analysis, drug, ace, sars-cov, response, patient]	[social, student, model, learn, people, health, datum, country, firm, public, online, government, case, supply, market, education, numb, food, return, work, economic, period, policy, mask, value]	[model, image, datum, case, study, patient, care, anxiety, country, test, state, population, mental, numb, staff, report, hospital, people, telehealth, survey, group, pharmacist, therapist]	[protein, pro, cell, surface, structure, assay, rna, human, bind, sars-cov-, target, nsp, filter, model, datum, vaccine, cluster, transmission, setting, method, covid-, problem]	[image, chest, lung, opacity, finding, cancer, symptom, care, pulmonary, hospital, treatment, day, risk, surgery, mortality, use, contact, room, transmission, need, embolism, covid, cohort, strain]	[symptom, viral, fever, risk, test, sars-cov-, cell, immune, use, treatment, lung, virus, child, transmission, case, vitamin, aki, medical, bcg, lesion, adverse, country]

Figure: Keywords of 6 clusters by GMM method.

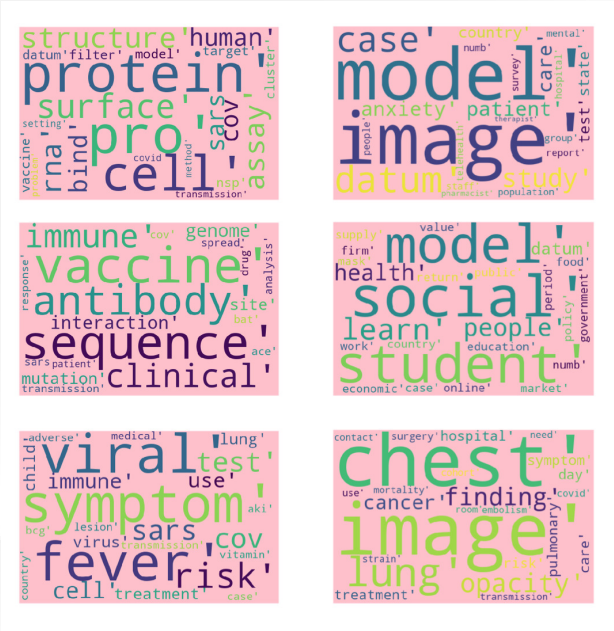

Figure: WordCloud of 6 clusters by GMM method

CONCLUSION

- The optimal number of clusters of K-Means method are 10; whereas it is 6 for GMM model.
- The GMM model shows the better result on two-dimensional data compared to K-Means method.
- In reality, the extracted keywords proved that K-Means is the better method.
- The different results of two methods can be explained by the dimension of the dataset, by the method determine the optimal cluster or components as well.

REFERENCES

Manzi, G. (2020), Lecture 20: Course Recap, lecture notes. *Advanced Multivariate Statistics B7416*, University of Milan, delivered December 2020.

Boehmke, B. & Greenwell, B. (2020). Chapter 22 Model-based Clustering. *Hands-On Machine Learning with R*, https://bradleyboehmke.github.io/HOML/model-clustering.html.

VanderPlas, J. (2016). In Depth: Gaussian Mixture Models. *Python Data Science Handbook*, https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html.

MaksimEkin (2020). Loading data. *COVID-19 Literature Clustering*, https://www.kaggle.com/maksimeren/covid-19-literature-clustering/notebook.

Palafox, L. (2019). A visual introduction to the Gap Statistics. *The Glowing Python*, https://glowingpython.blogspot.com/2019/01/a-visual-introduction-to-gap-statistics.html.