Clustering SFB Abstracts

Larisa Adamyan

Linxi Wang

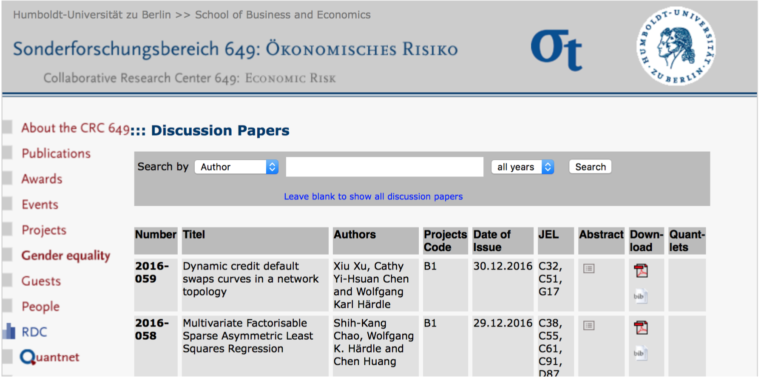
Kirill Efimov

Wolfgang Karl Härdle

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. - Center for Applied Statistics
and Economics
International Research Training Group 1792
Humboldt–Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de
irtg1792.hu-berlin.de

Topic extraction

- Find a **cluster structure** in the abstracts of SFB papers
- Compare it with *JEL* or *project codes*



Humboldt-Universität zu Berlin >> School of Business and Economics

Sonderforschungsbereich 649: Ökonomisches Risiko

Collaborative Research Center 649: Economic Risk

::: Discussion Papers

Leave blank to show all discussion papers

Number	Titel	Authors	Projects Code	Date of Issue	JEL	Abstract	Down-load	Quant-lets
2016-059	Dynamic credit default swaps curves in a network topology	Xiu Xu, Cathy Yi-Hsuan Chen and Wolfgang Karl Härdle	B1	30.12.2016	C32, C51, G17		bib	
2016-058	Multivariate Factorisable Sparse Asymmetric Least Squares Regression	Shih-Kang Chao, Wolfgang K. Härdle and Chen Huang	B1	29.12.2016	C38, C55, C61, C91, D87		bib	

Figure 1: Papers on SFB website

Outline

1. Motivation ✓
2. Data Preparation
3. Adaptive Weights Clustering
4. True clustering structure
5. References

Data Extraction

- Scrape SFB webpage with discussion papers
- For each paper extract:
  - Abstract
  - Project code
  - JEL Codes
- Store all the information in database on HU server

Data Preprocessing

- Tokenize
- Transfer all letters to small ones
- Remove punctuation, numbers, stopwords, special characters
- Lemmatize/stemming
- Remove words which occur only once

Term-Document Matrix (TDM)

- Rows correspond to the documents
- Columns correspond to the terms
- Each cell represents frequency of a word in a document

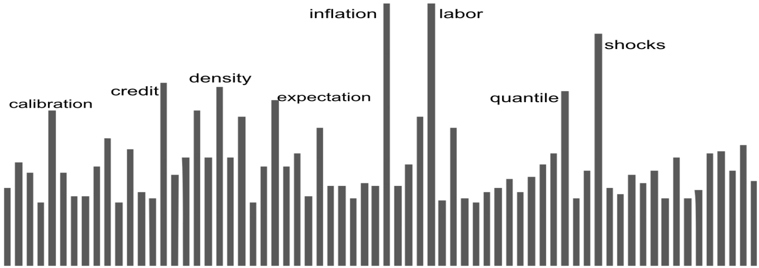


Figure 2: Most frequent terms from abstracts on SFB website

Term frequency- inverse document frequency (TF-IDF)

- A weighting factor
- Reflects how important a word is to a document in a collection
- i-th document is presented as vector $X_i = \{x_{ij}\}_{j=1}^{d}$, where

$$x_{ij} = tf_{ij} \times idf_j, \qquad idf_j = \log \frac{1+n}{1+n_j} + 1.$$

tf_{ij} : frequency of term j in the document i
idf_j : inverse document frequency
n : number of documents
n_j : number of documents which contain the term j.

True clustering structure

What to consider as true clustering structure?

- Project codes
  - represent project areas
  - 5 project area (Individual and contractual answers to risks, Macroeconomic risk, Financial markets, Risk Data Center, Transfer projects)

- JEL codes
  - represent topics
  - 17 JEL (Mathematical and quantitative methods, International economics, Financial economics, Business administration...)
  - paper abstracts can have up to 5 JEL codes

Comparison

- Adaptive Weights Clustering (AWC)
- K-means
  - minimize the objective function over partitions.
  - require to fix the number of clusters
  - produce only spherical clusters
- Cluto
  - a software package for clustering high dimensional datasets
  - hierarchical clustering
  - require to fix the number of clusters
  - produce high quality clustering solutions in text clustering

Normalized Mutual Information NMI

- True clustering structure $C^* = \{C^*_m\}_{m=1}^M$
- Answer clustering structure $C = \{C_l\}_{l=1}^L$

$$NMI(C, C^*) = \frac{\sum_{ml} n_{ml} \log \frac{n n_{ml}}{n_m n_l}}{\sqrt{\sum_m n_m \log \frac{n_m}{n} \cdot \sum_l n_l \log \frac{n_l}{n}}},$$

where $n_{ml} = |C^*_m \cap C_l|$, $n_m = |C^*_m|$, $n_l = |C_l|$.

- Maximize NMI

Misweighting Error used in AWC

- True weights w^*_{ij}
- Answer weights $\hat{w}_{ij}$

$$e = \frac{\sum_{i \neq j} |\hat{w}_{ij}| \mathbb{1}_{(w^*_{ij}=0)} + \sum_{i \neq j} |1 - \hat{w}_{ij}| \mathbb{1}_{(w^*_{ij}=1)}}{\sum_{i \neq j} \mathbb{1}_{(w^*_{ij}=0)} + \sum_{i \neq j} \mathbb{1}_{(w^*_{ij}=1)}}$$

Rand index:

$$R = 1 - e$$

- Minimize e

Adjusted Rand Index AdR

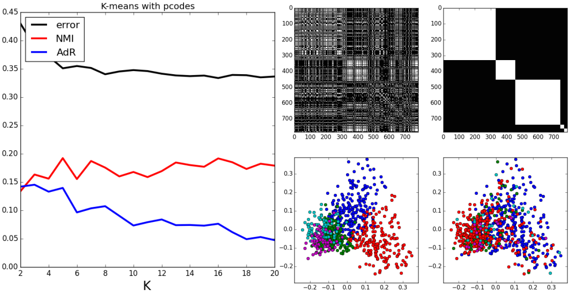
- True clustering structure $C^* = \{C^*_m\}_{m=1}^M$
- Answer clustering structure $C = \{C_l\}_{l=1}^L$

$$AdR(C, C^*) = \frac{\sum_{ml} \binom{n_{ml}}{2} - \sum_m \binom{n_m}{2} \sum_l \binom{n_l}{2} / \binom{n}{2}}{\frac{1}{2}(\sum_m \binom{n_m}{2} + \sum_l \binom{n_l}{2}) - \sum_m \binom{n_m}{2} \sum_l \binom{n_l}{2} / \binom{n}{2}}$$

- Maximize AdR

K-means

- Project codes as true clustering structure
- 50 runs for each K
- Try with PCA (number of components = 2, 5, 10)
- Best result without PCA
- Best result when K = 5



Cluto

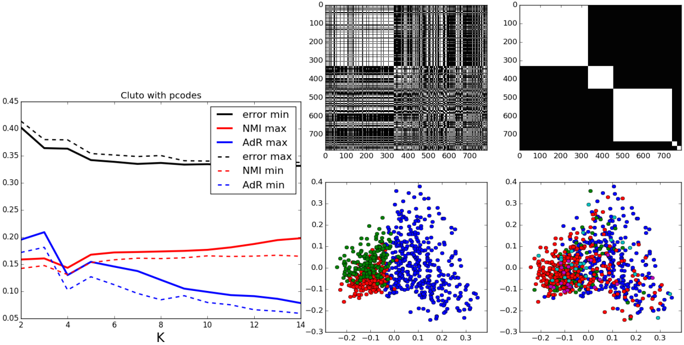


Figure 3: 50 runs for each K. K = 3 best result

AWC

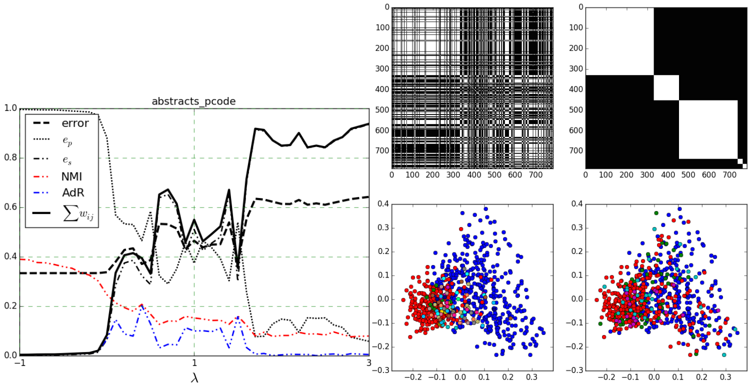
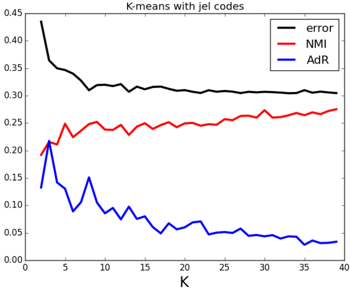
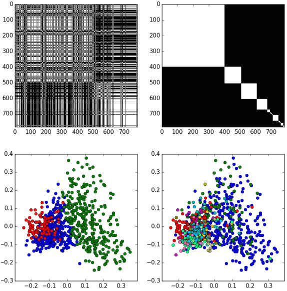


Figure 4: left: plateau heuristics, right: AWC result for $\lambda = 0.4$

K-means

- JEL codes as true clustering structure
- 50 runs for each K
- Try with PCA (number of components = 2, 5, 10)
- Best result without PCA
- Best result when K = 3





Cluto

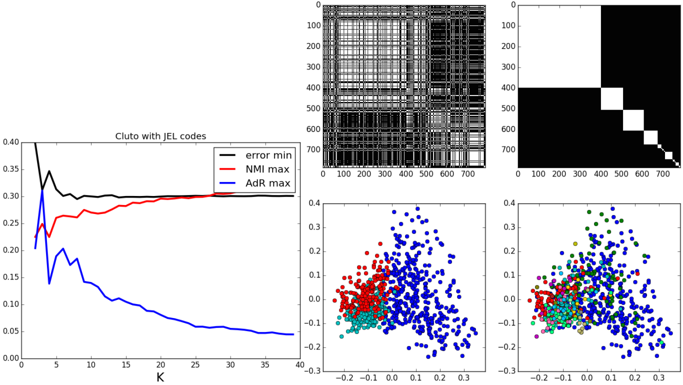


Figure 5: 50 runs for each K. K = 3 best result

AWC

- ⊡ JEL codes as true clustering structure

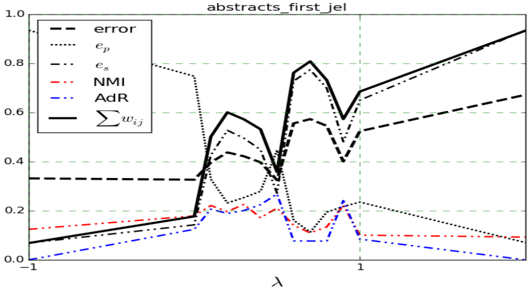


Figure 6: Plateau heuristics

AWC Result

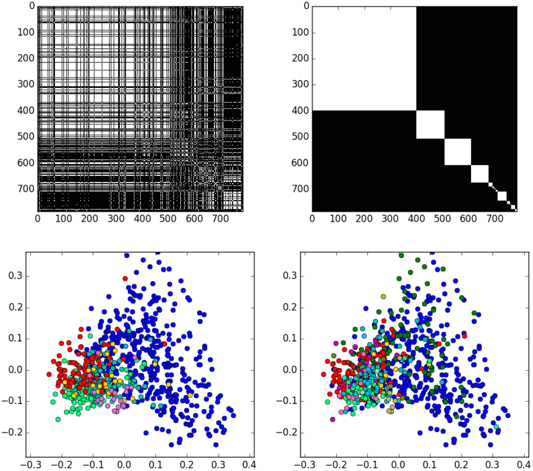


Figure 7: Result for $\lambda = 0.5$ from plateau heuristics

Cluster 1 found by AWC

- 46% contain *G*: 'Financial economics'
- 81% contain *C*: 'Mathematical and quantitative methods'
- Contains 86% of pairs {C, G}

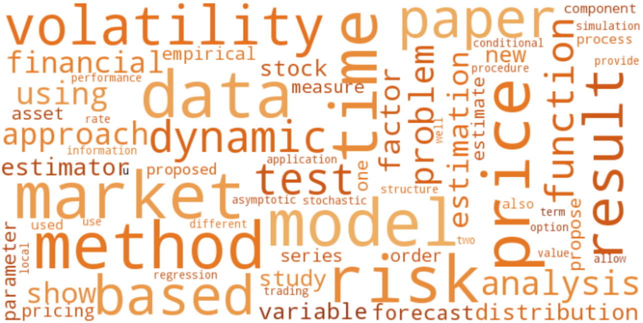


Figure 8: size = word frequency, darker color – higher idf

Cluster 2 found by AWC

- 77% contain *J*: 'Labor economics'

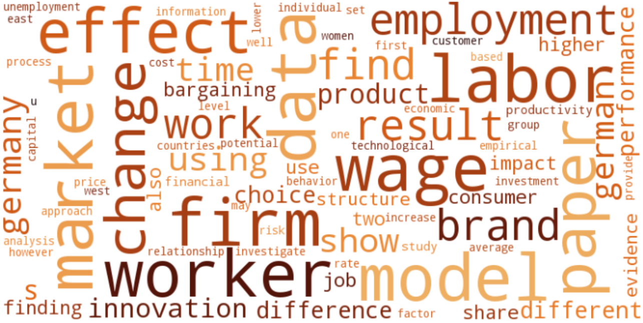


Figure 9: size = word frequency, darker color – higher idf

Cluster 3 found by AWC

- 51% contain *D*: 'Microeconomics'
- 54% contain *C*: 'Mathematical and quantitative methods'

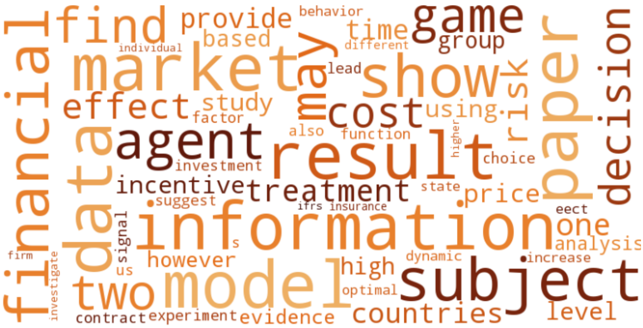


Figure 10: size = word frequency, darker color – higher idf

Cluster 4 found by AWC

- 73% contain *E*: 'Macroeconomics and monetary economics'

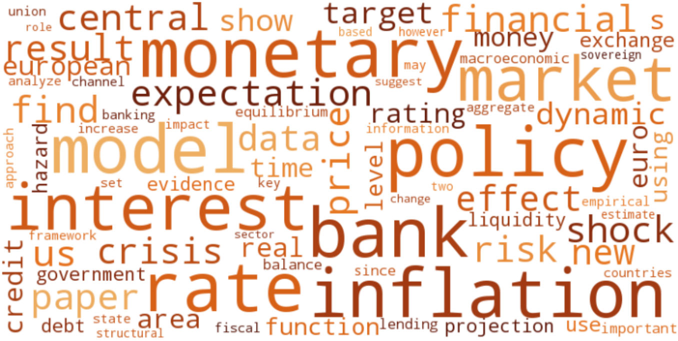


Figure 11: size = word frequency, darker color – higher idf

Cluster 5 found by AWC

- 32% contain *R*: 'Urban, rural, and regional economic'
- 24% contain *Q*: 'natural resource economics'
- 40% contain *C*: 'Mathematical and quantitative methods'

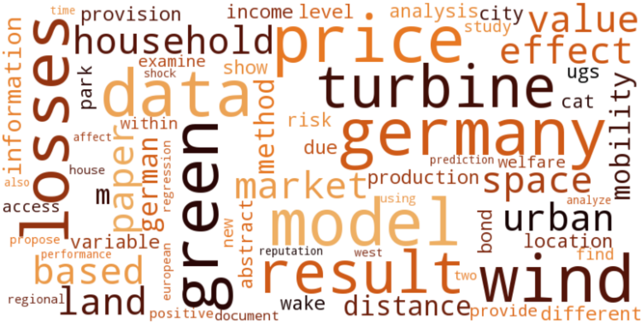


Figure 12: size = word frequency, darker color – higher idf

Cluster 6 found by AWC

- 54% contain I: 'Health, education, and welfare'
- 80% contain C: 'Mathematical and quantitative methods'
- 50% contain pairs $\{I, C\}$

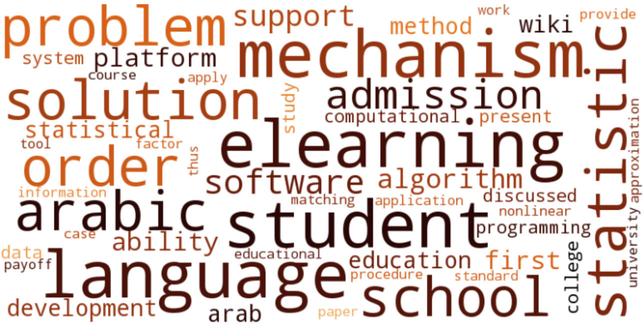


Figure 13: size = word frequency, darker color – higher idf

Conclusion

- The best run of *k-means* among 50 runs for each $2 \leq k \leq 30$ provides best AdR $= 0.22$ when $k = 3$
- CLUTO can provide partitioning with AdR $= 0.32$ (best result among 50 runs for k=3)
- The best result of CLUTO for $k \neq 3$ is AdR $= 0.20$
- **AWC automatically finds meaningful cluster structure with AdR = 0.27**

References

Steinhaus, H.
"Sur la division des corp materiels en parties." Bull. Acad. Polon. Sci 1.804 (1956): 801.

Strehl, A. and Ghosh, J.
"Cluster ensembles a knowledge reuse framework for combining multiple partitions". Journal of machine learning research, 3, 583-617. (2002)

Lawrence, H. and Phipps, A.
"Comparing partitions". Journal of Classification (1985).

Karypis, G.
"CLUTO - a clustering toolkit".No. TR-02-017. MINNESOTA UNIV MINNEAPOLIS DEPT OF COMPUTER SCIENCE, (2002).

Clustering SFB Abstracts

Larisa Adamyan
Linxi Wang
Kirill Efimov
Wolfgang Karl Härdle

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. - Center for Applied Statistics
and Economics
International Research Training Group 1792
Humboldt–Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de
irtg1792.hu-berlin.de