Assignments of JEL codes via Adaptive Weights Clustering

Larisa Adamyan, Kirill Efimov, Linxi Wang, Wolfgang K. Härdle

International Research Training Group 1792
Ladislaus von Bortkiewicz Chair of Statistics
Humboldt–Universität zu Berlin

Outline

- Publication industry offers a rich portfolio of research work
- The mass of textual data requires pre-structuring
  - Abstracts are a condensed information of full documents
  - Economic papers require manually specify the JEL codes (project areas, topics)
- Clustering can be a solution to identify the research directions and activity of economic research on certain topics.

Motivation

- Analyze Discussion paper abstracts from School of Business and Economics at Humbold-Universität zu Berlin
- Find a **cluster structure** using Adaptive Weights Clustering
- Examine its correlation with paper's *JEL codes*

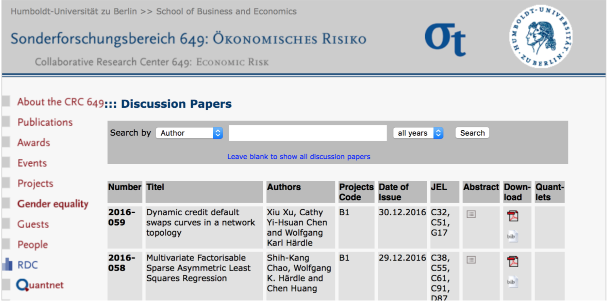


Abbildung: Papers on SFB website

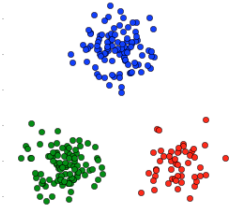
What is clustering?

Data: $X_1, \ldots, X_n \in \mathbb{R}^d$.

Aim: split into homogeneous groups (clusters).

Number and structure/shape of clusters usually unknown.

Ideal picture:



- Scrape SFB webpage with discussion papers and extract:
  - Abstracts
  - JEL Codes
- Preprocess abstracts:
  - 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

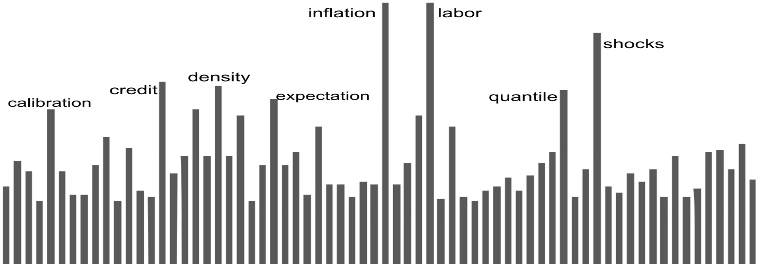


Abbildung: Most frequent terms from abstracts on SFB website

- 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.

Outline

Aim: an efficient procedure which adapts to unknown cluster structure.

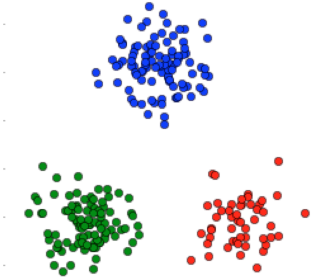
Approach: Describe the cluster structure by an adjacency matrix $W = (w_{ij})$, where

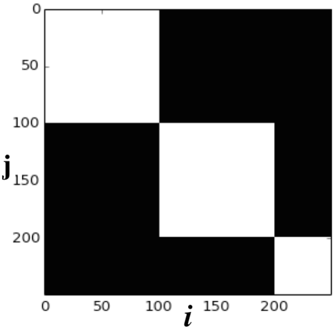
$$w_{ij} = \begin{cases} 1 & i,j \text{ from the same cluster,} \\ 0, & \text{otherwise} \end{cases}$$

The matrix W is recovered from the data by an iterative procedure.

Let $\{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ be the set of all samples X_i.

Example: 250 points, 3 normal clusters (100 + 100 + 50) and the corresponding matrix of weights W.





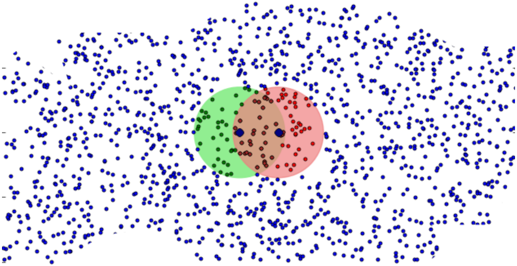
$$W = (w_{ij})_{i,j=1,\ldots,n} \ , \ w_{ij} \in [0,1] \, .$$

AWC Procedure

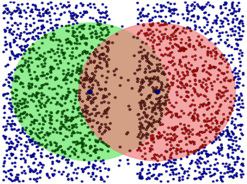
- Initialize with one cluster $\mathscr{C}_i^{(0)}$ per point X_i ;
- At each step, increase the locality parameter h_k and recompute the local weights $w_{ij}^{(k)}$ using a statistical test of no gap between two local clusters $\mathscr{C}_i^{(k-1)}$ and $\mathscr{C}_j^{(k-1)}$.
- Stop when the bandwidth h_k reaches the global value.

Test of "no gap between local clusters"

Homogeneous case:



"Gap" case:



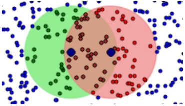
After $k-1$ steps, for each $i \le n$, the cluster $\mathscr{C}_i^{(k-1)}$ is given via weights $w_{ij}^{(k-1)}$, $j \le n$.

At step k, suppose the locality parameter h_k to be fixed and consider any pair (X_i, X_j) with $\|X_i - X_j\| \le h_k$.

Problem: For two local clusters $\mathscr{C}_i^{(k-1)}$ and $\mathscr{C}_j^{(k-1)}$ with $\|X_i - X_j\| \le h_k$, compute the value $w_{ij}^{(k)}$ reflecting the gap between $\mathscr{C}_i^{(k-1)}$ and $\mathscr{C}_j^{(k-1)}$.

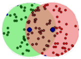
Principal idea: check the data density in the overlap $\mathscr{C}_i^{(k-1)} \cap \mathscr{C}_j^{(k-1)}$.

Test of "no gap between local clusters". Cont

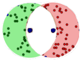


Mass of the overlap $N^{(k)}_{i\cap j}$,  :

$$N^{(k)}_{i\cap j} \stackrel{\text{def}}{=} \sum_{l\neq i,j} w^{(k-1)}_{il} w^{(k-1)}_{jl} \approx \#\text{ points in } \mathcal{B}(X_i, h_{k-1}) \cap \mathcal{B}(X_j, h_{k-1})$$

Mass of the union $N^{(k)}_{i\cup j}$  :

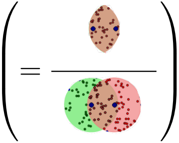
$$N^{(k)}_{i\cup j} \stackrel{\text{def}}{=} N^{(k)}_{i\cap j} + N^{(k)}_{i\triangle j} \approx \#\text{ points in } \mathcal{B}(X_i, h_{k-1}) \cup \mathcal{B}(X_j, h_{k-1})$$

where $N^{(k)}_{i\triangle j}$ is the mass of the complementary parts  :

$$N^{(k)}_{i\triangle j} \stackrel{\text{def}}{=} \sum_{l\neq i,j:\{\|X_i - X_l\| \le h_{k-1}\} \triangle \{\|X_j - X_l\| \le h_{k-1}\}} \left(w^{(k-1)}_{il} + w^{(k-1)}_{jl} \right).$$

Estimated relative density in the overlap:

$$\widetilde{\theta}^{(k)}_{i\cap j} = \frac{N^{(k)}_{i\cap j}}{N^{(k)}_{i\cup j}}$$



Local homogeneous case corresponds to the nearly uniform distribution:

$$\widetilde{\theta}^{(k)}_{i\cap j} \approx q^{(k)}_{ij} \stackrel{\text{def}}{=} \frac{\mathrm{Vol}_{\cap}(\rho_{ij}, h_{k-1})}{2\,\mathrm{Vol}(h_{k-1}) - \mathrm{Vol}_{\cap}(\rho_{ij}, h_{k-1})} = q\left(\frac{\rho_{ij}}{h_{k-1}}\right),$$

where $\mathrm{Vol}(h)$ is the volume of a ball with radius h and $\mathrm{Vol}_{\cap}(\rho, h)$ is the volume of the intersection of two balls with radii h and the distance ρ between centers, $\rho_{ij} = \|X_i - X_j\|$.

Null (no gap): $\theta^{(k)}_{i\cap j} > q^{(k)}_{ij}$ vs alternative (a gap) $\theta^{(k)}_{i\cap j} < q^{(k)}_{ij}$.

Outline

1 Motivation

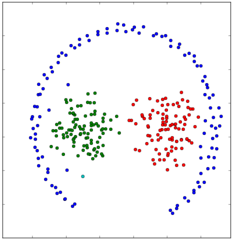
2 Data Preparation

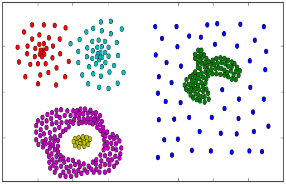
3 Adaptive Weights Clustering

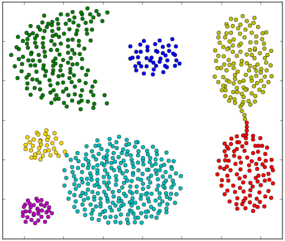
4 **AWC Results**

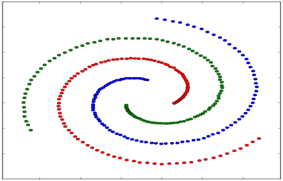
5 AWC on SFB Abstracts

Artificial Examples









AWC

More examples

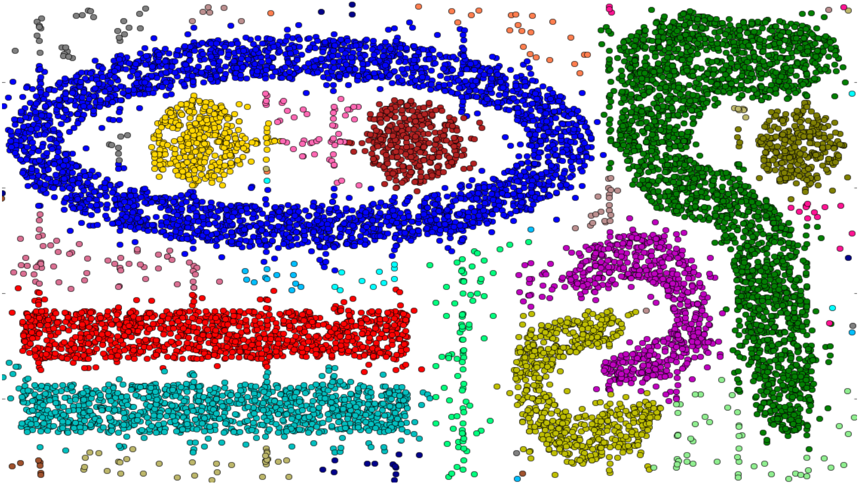
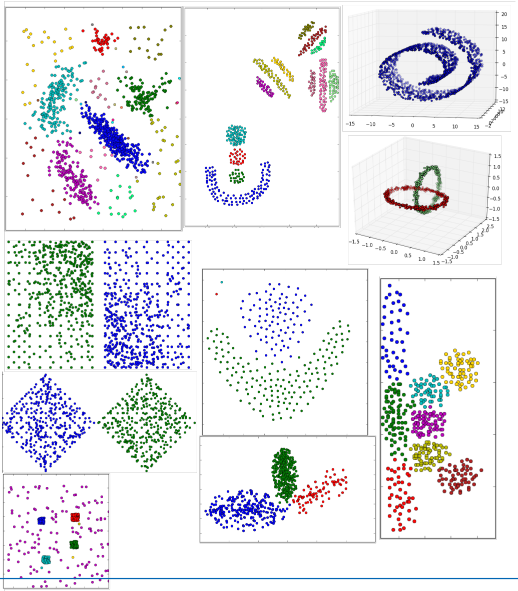


Abbildung: $DS4$, $n = 10000$ points

More examples



Outline

AWC Result

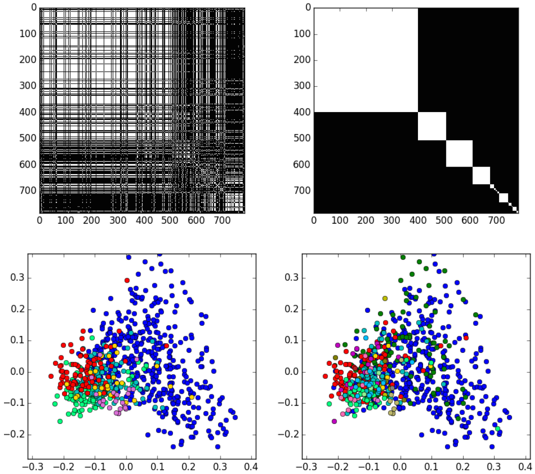


Abbildung: Clustering Structure found by AWC vs Original

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

Abbildung: size = word frequency, darker color – higher idf

Cluster 2 found by AWC

- 77% contain J: 'Labor economics'

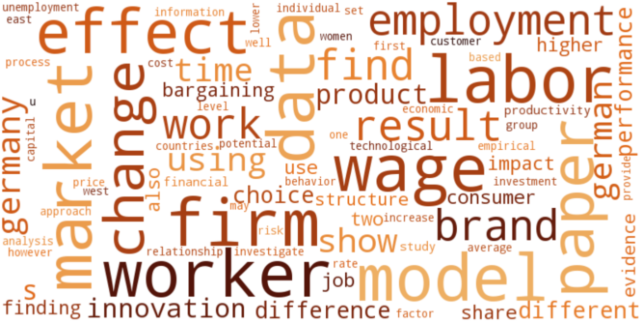


Abbildung: size = word frequency, darker color – higher idf

Cluster 3 found by AWC

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

Abbildung: size = word frequency, darker color – higher idf

Cluster 4 found by AWC

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

Abbildung: 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'

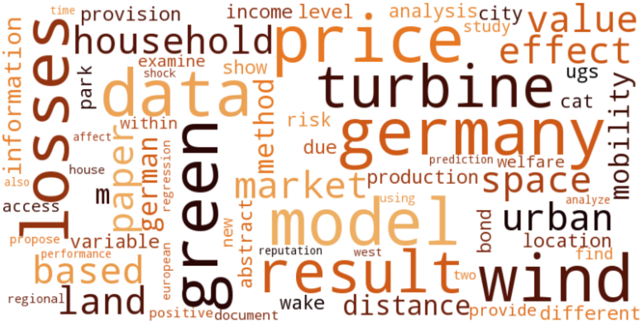


Abbildung: size = word frequency, darker color – higher idf

Cluster 6 found by AWC

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

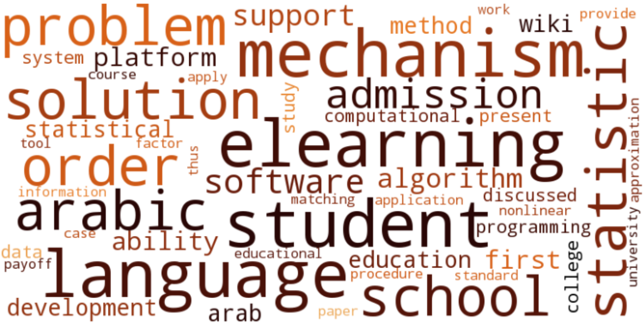


Abbildung: size = word frequency, darker color – higher idf

- Clustering as a specification of research directions of the papers
- Clustering using adaptive weights:
  - numerically feasible and applicable for large data sets
  - fully adaptive to unknown clustering structure including the number and shape of clusters and the separation distance
  - State-of-the-art performance of a wide range of artificial and real life examples
- Clustering procedure automatically assigns the JEL codes to submitted papers.

Thank you!

Test of “no gap between local clusters”. Formal definition

We need to test if $\theta_{i\cap j}^{(k)} > q_{ij}^{(k)}$. Following to Polzehl and Spokoiny (2006), define the test statistic $T_{ij}^{(k)}$

$$T_{ij}^{(k)} = N_{i\cup j}^{(k)} \mathcal{K}(\widetilde{\theta}_{i\cap j}^{(k)}, q_{ij}^{(k)})(-1)^{\mathbf{I}(\widetilde{\theta}_{i\cap j}^{(k)} > q_{ij}^{(k)})},$$

where $\mathcal{K}(\theta, q)$ is the Kullback-Leibler divergence:

$$\mathcal{K}(\theta, q) = \theta \log \frac{\theta}{q} + (1-\theta) \log \frac{1-\theta}{1-q}.$$

Algorithm steps:

- Initialization of weights w_{ij}
  - each point is connected with its n_0 closest neighbors
  - default choice of $n_0 = 2d+1$
- Fix a sequence of radii h_k:
  - The average number of screened neighbors for each X_i at step *k* grows at most exponentially.
- Iterative update of the weights:
  - At step k for all pairs of points X_i and X_j with distance $\|X_i - X_j\| \le h_k$ compute

$$w_{ij}^{(k)} = \mathbb{1}\left(\|X_i - X_j\| \le h_k\right) \mathbb{1}\left(T_{ij}^{(k)} \le \lambda\right)$$

- Cluster extraction from matrix of weights

- Take the point X_i with maximal local cluster $\mathscr{C}(X_i) = (X_j : w_{ij} > 0)$.
- Consider it as a separated cluster if majority of points X_j in $\mathscr{C}(X_i)$ have almost the same number of neighbors as X_i.
- Delete this cluster $\mathscr{C}(X_i)$ and repeat the procedure for remaining points until we delete all points.

Run the procedure with different λ and select one by checking an increase of the sum of weights $\sum_{i,j} w_{ij}^{(K)}$.

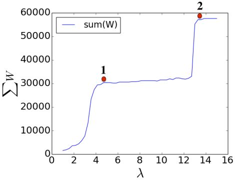


Abbildung: $\sum_{i,j} w_{ij}^{(K)}(\lambda)$

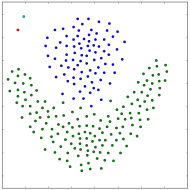


Abbildung: AWC for 1