SFB 649 Discussion Paper 2017-013

Adaptive weights clustering of research papers

Larisa Adamyan*
Kirill Efimov*
Cathy*
Yi-Hsuan Chen*
Wolfgang K. Härdle*

*Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

SFB 649 ECONOMIC RISK BERLIN

Noname manuscript No.
(will be inserted by the editor)

Adaptive weights clustering of research papers

Larisa Adamyan · Kirill Efimov · Cathy Yi-Hsuan Chen · Wolfgang K. Härdle

Received: date / Accepted: date

Abstract The JEL classification system is a standard way of assigning key topics to economic articles in order to make them more easily retrievable in the bulk of nowadays massive literature. Usually the JEL (Journal of Economic Literature) is picked by the author(s) bearing the risk of suboptimal assignment. Using the database of a Collaborative Research Center from Humboldt-Universität zu Berlin and Xiamen University, China we employ a new adaptive clustering technique to identify interpretable JEL (sub)clusters. The proposed Adaptive Weights Clustering (AWC) is available on `www.quantlet.de` and is based on the idea of locally weighting each point (document, abstract) in terms of cluster membership. Comparison with k-means or CLUTO reveals excellent performance of AWC.

Keywords Clustering · JEL system · Adaptive algorithm · Economic articles · Nonparametric

1 Introduction

"Words are the new numbers". This quote [1] expresses the insight into the power of the spoken, written or tweeted message in a plethora of applications, social networks and academic discourse. The academic publication industry offers us a rich portfolio

L. Adamyan[†]
E-mail: ladamyan@hu-berlin.de
K. Efimov[†]
E-mail: kirillefimovs@hu-berlin.de
C. Chen[†]
E-mail: chencath@hu-berlin.de
W. K. Härdle[†§]
E-mail: haerdle@hu-berlin.de

[†] Humboldt-Universität zu Berlin, C.A.S.E.-Center of Applied Statistics and Economics, Unter den Linden 6, 10099 Berlin, Germany
[§] Sim Kee Boon Institute for Financial Economics, Singapore Management University, 90 Stamford Road, 6th Level, School of Economics, Singapore 178903.

of research work in variety of outlets like journals, books or epub platforms. The mass of textual data requires pre-structuring in order to avoid the "needle in a haystack" problem that everybody seems to have when one looks for specific information e.g. in a particular domain of a scientific discipline. This is one of the major reasons why abstracts as condensed information of a full research documents are required and that is why e.g. economic papers are classified according to JEL codes. The JEL classification system originated with the Journal of Economic Literature and is a standard method of classifying scholarly literature in the field of Economics [2].

The assignment of such a classification code is done by authors manually and submitted together with a publication. This procedure bears risks. First, author(s) may not be aware of the "best fitting" JEL code in the sense of fast retrieval properties. Second, the spectrum of submitted codes may be too rich or too narrow. For the reasons described above we propose a clustering procedure that automatically assigns the JEL codes to submitted papers.

We analyze papers abstracts from the School of Business and Economics in Humbold-Universität zu Berlin. Papers from year 2005 to year 2017 are stored on the SFB web page [9] and have an open access. Besides the main information such as title, authors and date of issue, this web page stores for every paper its abstract and JEL codes given by the authors. By clustering this collection of documents in an unsupervised learning context we also identify the research directions and activity of economic research on certain topics. Comparing cluster sizes of certain topics will allow us to see whether research groups have biased activity relative to mainstream economic research.

Clustering is a well known data science technique which has a long tradition in statistical learning. Recently a non-parametric technique called Adaptive Weights Clustering (AWC) has emerged that showed excellent performance on various artificial and real world examples [10]. How can we cluster texts? first of all one needs to convert words into numbers. Examples of such numerization of texts into numbers are abound, see e.g. [11] or [12].

We apply Adaptive Weights Clustering algorithm introduced in [10] to cluster the abstracts of the papers and try to find a correlation between the resulting cluster structure and the JEL codes of the papers. All clustering methods will consider finding an accurate clustering structure as a demanding task since all the documents belong to economic domain and share topic areas. During evaluation we compare the performance of AWC with the well known standard *k*-means clustering algorithm [3] and the graph partitioning based algorithm from the CLUTO toolkit [4], which is considered as the state-of-the-art approach to document clustering.

Clustering via *k*-means is one of the mostly used partitional clustering algorithm. These algorithms try to group points by optimizing some specific objective function over the data. The aim of the *k*-means algorithm is to divide M points in N dimensions into K clusters so that the sum of squares within clusters is minimized. It seeks a "local" optimal solution that no movement of a point from one cluster to another will reduce the within-cluster sum of squares[8].

CLUTO [4] is a package for clustering low and high dimensional datasets. It provides different classes of clustering algorithms based on the partitional, agglomerative and graph-partitioning patterns. Agglomerative clustering is a bottom-up hierar-

chical clustering method which proceeds by starting with the individual instances and grouping the ones that have most similarities. It produces a sequence of partitions in which each partition is nested into the next partition in the sequence[7]. As a result it constructs from the data a tree called dendrogram, which displays the intermediate clustering assignments and the merging process. As a contrast to agglomerative paradigm, graph partitioning algorithms perform a sequence of recursive splits until the desired number of clusters are found. CLUTOs Metis graph partitioning based algorithms has been shown to produce high quality clustering results in high dimensional datasets with low computational cost [5].

The comparison with these mentioned cluster techniques we are able to show superior performance for AWC. Our AWC method identifies 5 clusters that can be matched with 'Market Risk', 'Labor Economics', 'Monetary Policy', 'Game Theory', 'Green Energy'.

This paper is organized as follows: Section 2 describes in details the Adaptive Weights Clustering algorithm and a heuristic for tuning of its parameter. In section 3 the process of collecting documents and further preprocessing steps are carried out to prepare the data collection for cluster analysis. In section 4 we choose a clustering performance measure and define a true clustering structure for our data collection. The comparison of clustering methods and experimental results are shown in section 5. Finally, a conclusion is given about the results of the papers main ideas.

2 Adaptive Weights Clustering

An alternative non-parametric clustering technique based on the separation approach via a homogeneity detection test is proposed by [10]. A cluster is defined as a homogeneous region without gaps. A direct advantage of this definition is that it does not require specifying number of clusters. It applies equally well to clusters of convex structure and different density.

The clustering structure of the data is described in terms of binary weights w_{ij}, where $w_{ij} = 1$ indicates being points X_i and X_j in the same cluster, whereas $w_{ij} = 0$ means that these points belong to different clusters. For each point X_i, the associated cluster C_i is given by the collection of positive weights (w_{ij}) over all j. The resulting symmetric matrix of weights W consists of blocks of ones, where each block of ones describes one cluster.

The proposed procedure attempts to iteratively recover the weights w_{ij} from the data. It starts with very local clustering structure $C_i^{(0)}$, that is, the starting positive weights $w_{ij}^{(0)}$ are limited to the closest neighbors X_j of the point X_i in terms of a distance $d(X_i, X_j)$. At each step $k \geq 1$, the weights $w_{ij}^{(k)}$ are recomputed by means of statistical "no gap" tests between $C_i^{(k-1)}$ and $C_j^{(k-1)}$, the local clusters on step $k-1$ for points X_i and X_j correspondingly. Only the neighbor pairs X_i, X_j with $d(X_i, X_j) \leq h_k$ are checked, however the locality parameter h_k and the number of neighbors X_j for each fixed point X_i grow in each step. The resulting matrix of weights W is used for the final clustering. The core element of this adaptive weights clustering (AWC) is the way how the weights $w_{ij}^{(k)}$ are recomputed.

2.1 Clustering by Adaptive Weights

Let $\{X_1, \ldots, X_n\} \subset \mathbb{R}^p$ be the set of all samples X_i, where the dimension p can be very large or even growing. The proposed technique operates with a known distance or similarity matrix $\big(d(X_i, X_j)\big)_{i,j=1}^n$ only. In the experiments we use Euclidean norm: $d(X_i, X_j) = \|X_i - X_j\|$,for every $i, j = 1, \ldots, n$.

The procedure starts from a small scale and considers only points close to each other, then slowly increases the scale and finally considers all pairs of points. For each point X_i, weights $w_{ij}^{(k)}$ are computed using only points from the neighborhood of radius h_k around X_i and X_j. As radius h_k increases with k, weights become more and more data driven during iterations.

A sequence of radii: A growing sequence of radii $h_1 \le h_2 \le \ldots \le h_K$ is fixed which determines how fast the algorithm will accelerate from very local structures to large scale objects. Each value h_k can be viewed as a resolution (scale) of the method at step k. The average number of screened neighbors for each X_i at step k grows at most exponentially with $k \ge 1$.

Initialization of weights: On initialization step each point is connected with its n_0 closest neighbors, where the default choice of $n_0 = 2p + 2$.

Updates at step k: Suppose that the first $k-1$ steps of AWC have been carried out. This results in collection of weights $\big\{w_{ij}^{(k-1)},\ j = 1, \ldots, n\big\}$ for each point X_i. These weights describe a local "cluster" associated with X_i. By construction, only those weights $w_{ij}^{(k-1)}$ can be positive for which X_j belongs to the ball $B(X_i, h_{k-1}) = \{x\colon d(X_i, x) \le h_{k-1}\}$. At the next step k a larger radius h_k is picked and the weights $w_{ij}^{(k)}$ are recomputed using the previous results.

The basic idea behind the definition of $w_{ij}^{(k)}$ is to check for each pair i, j with $d(X_i, X_j) \le h_k$ whether the related clusters are well separated or they can be aggregated into one homogeneous region. A test statistic $T_{ij}^{(k)}$ is computed to compare the data density in the union and overlap of two clusters for points X_i and X_j using the weights $w_{ij}^{(k-1)}$ from the preceding step. The formal definition involves the weighted empirical mass of the overlap and the weighted empirical mass of the union of two balls $B(X_i, h_{k-1})$ and $B(X_j, h_{k-1})$ shown on Figure 1.

The empirical mass of the overlap $N_{i\wedge j}^{(k)}$ is defined as

$$N_{i\wedge j}^{(k)} = \sum_{l \ne i,j} w_{il}^{(k-1)} w_{jl}^{(k-1)}, \tag{1}$$

which is the number of points in the overlap of $B(X_i, h_{k-1})$ and $B(X_j, h_{k-1})$ except points X_i, X_j. Similarly, the *mass of the complement* is defined as

$$N_{i\triangle j}^{(k)} = \sum_{l \ne i,j} \big\{ w_{il}^{(k-1)}\, \mathbb{1}\big(X_l \notin B(X_j, h_{k-1})\big) + w_{jl}^{(k-1)}\, \mathbb{1}\big(X_l \notin B(X_i, h_{k-1})\big) \big\} \tag{2}$$

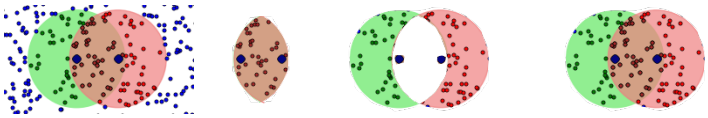

Fig. 1 Test of "no gap between local clusters". From left: Homogeneous case; $N^{(k)}_{i\wedge j}$; $N^{(k)}_{i\triangle j}$; $N^{(k)}_{i\vee j}$

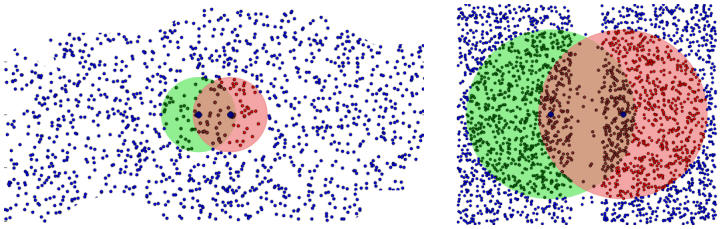

Fig. 2 Left: Homogeneous case. Right: "Gap" case.

$N^{(k)}_{i\triangle j}$ can be seen as the number of points in $C_i^{(k-1)}$ and $C_j^{(k-1)}$ which do not belong to the overlap $B(X_i, h_{k-1})\cap B(X_j, h_{k-1})$. Finally, *mass of the union* $N^{(k)}_{i\vee j}$ is defined via (1), (2) as the sum of the mass of the overlap and the mass of the complement:

$$N^{(k)}_{i\vee j} \;=\; N^{(k)}_{i\wedge j} + N^{(k)}_{i\triangle j}\,. \tag{3}$$

The gap between two regions is measured considering the ratio of these two masses (1), (3):

$$\tilde\theta^{(k)}_{ij} \;=\; N^{(k)}_{i\wedge j} \,/\, N^{(k)}_{i\vee j}\,. \tag{4}$$

The value (4) can be viewed as an estimate of θ_{ij} which measures the ratio of the averaged density in the overlap of two local regions C_i and C_j relative to the average density. In fact (4) should be close to the ratio of the corresponding volumes given local homogeneity:

$$\tilde\theta^{(k)}_{ij} \;\approx\; q^{(k)}_{ij} \;=\; \frac{V_\cap(d_{ij}, h_{k-1})}{2V(h_{k-1}) - V_\cap(d_{ij}, h_{k-1})}.$$

Where $V(h)$ is the volume of a ball with radius h and $V_\cap(d_{ij}, h)$ is the volume of the intersection of two balls with radius h and the distance between centers $d_{ij} = d(X_i, X_j)$.

The new value $w^{(k)}_{ij}$ can be viewed as a randomized test of the null hypothesis H_{ij} of no gap between X_i and X_j against the alternative of a significant gap. The gap is significant if $\tilde\theta^{(k)}_{ij}$ is significantly smaller than $q^{(k)}_{ij}$. The construction is illustrated in Figure 2 for the homogeneous situation (left) and for a situation with a gap (right).

To quantify the notion of significance, the statistical likelihood ratio test of "no gap" between two local clusters is considered, that is $\tilde{\theta}_{ij}^{(k)} > q_{ij}^{(k)}$ vs $\tilde{\theta}_{ij}^{(k)} \leq q_{ij}^{(k)}$:

$$T_{ij}^{(k)} \,=\, N_{i\vee j}^{(k)}\; KL\big(\tilde{\theta}_{ij}^{(k)}, q_{ij}^{(k)}\big)\,\big\{\mathbb{1}(\tilde{\theta}_{ij}^{(k)} \leq q_{ij}^{(k)}) - \mathbb{1}(\tilde{\theta}_{ij}^{(k)} > q_{ij}^{(k)})\big\}. \tag{5}$$

Here $KL(\theta, \eta)$ is the Kullback-Leibler (KL) divergence between two Bernoulli laws with parameters θ and η:

$$KL(\theta, \eta) \,=\, \theta \log \frac{\theta}{\eta} + (1-\theta) \log \frac{1-\theta}{1-\eta}\ . \tag{6}$$

Finally, the weights $w_{ij}^{(k)}$ are updated for all pairs of points X_i and X_j with distance $d_{ij} \leq h_k$:

$$w_{ij}^{(k)} \,=\, \mathbb{1}\left(d_{ij} \leq h_k\right)\, \mathbb{1}\left(T_{ij}^{(k)} \leq \lambda\right)$$

where λ is some hyperparameter controlling the size of the test (5).

Note that the first indicator function in (6) allows to recompute the $n \times n_k$ weights, where n_k is the average number of neighbors in the h_k neighborhood.

The tests $T_{ij}^{(k)}$ are scaled by a global constant λ which is the only tuning parameter of the method. The parameter λ has an important influence on the performance of AWC. Large λ-values will lead to aggregation of in-homogeneous regions. On the contrary, small λ increases the sensitivity of the methods to in-homogeneity but may lead to artificial segmentation.

2.1.1 Parameter tuning

In [10] a heuristic choice of λ based on the effective cluster size is proposed:

Let $w_{ij}^{K}(\lambda)$ be the collection of final AWC weights. Define

$$S(\lambda) \,=\, \sum_{i,j=1}^{n} w_{ij}^{K}(\lambda). \tag{7}$$

Note that given our thoughts above an increase of λ yields larger homogeneous blocks and thus, a larger value $S(\lambda)$. A natural proposal is therefore to pick up the λ-value corresponding to a point right before observing a huge jump in graph of $S(\lambda)$ from (7). This in fact resembles the elbow criterion that we all know from PCA. In the case of a complex cluster structure, several jump points can be observed with the corresponding λ-value for each jump.In this case all those λ-values should be checked and compared the obtained clustering results afterwards.

3 Document Collection and Preprocessing

The SFB web page [9] provides an open access to the Discussion Papers from year 2005 to year 2017 from the department of School of Business and Economics in Humboldt-Universität zu Berlin. We scrape this SFB webpage and extract abstracts of the papers, which form our dataset. For evaluation purposes we also scrape from the website the JEL codes of each paper. There are overall 784 papers from the Economics domain. Further the standard text preprocessing steps are performed to transform the collection of raw data to the vector space. First we split documents into words and transfer all the letters to small ones. Then we perform stemming, remove all punctuation, numbers, special characters, stopwords and words which occurred only once in the dataset. At this step we have a collection of preprocessed documents and the research areas of each document. For details about this information extraction we refer to [11]. The most frequent terms in the collection are plotted on Figure 3. One clearly sees that the documents/abstracts operate in a quantitative economic field, since besides clearly economic terms like "credit" one finds "density" at almost identical frequency.

The basic model for document clustering is the vector space model, therefore we convert the preprocessed documents into tf-idf vector space, [13]. Here each document, X_i is first presented as a term-frequency vector in the term-space: $X_{itf} = \{tf_{ij}\}_{j=1}^{d}$, where tf_{ij} is the frequency of the j-th term in the document i and d is the dimension of the term-space.

Then, each document is weighted via its inverse document frequency (IDF). This weighting factor ensures the frequent term across all documents in a dataset being discounted and considering as a non informative term. Hence, for each i-th document, we obtain the following vector representation:

$X_i = \{x_{ij}\}_{j=1}^{d}$, where

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

Here idf_j is the inverse document frequency , n is the number of documents in a collection and n_j is the number of documents which contain the term j. Hence, tf-idf of a word gives a product of how frequent this word is in the document multiplied by how unique the word is w.r.t. the entire corpus of documents. Words in the document with a high tf-idf score appear frequently in the document and are informative within specific document. The resulting matrix is used further for cluster analysis.

4 Evaluation criteria

In the experiments as a measure of a clustering performance we use the Adjusted Rand Index (ARI) [6]. ARI is considered as a popular measure for cluster validation particularly for textual data. It measures the similarity between the defined true clusters and the estimated clusters.

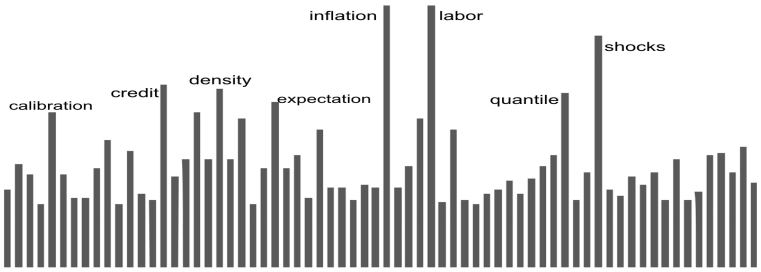

Fig. 3 Most frequent terms in the collection

Suppose that the true clustering structure is $C^* = \{C^*_m\}_{m=1}^M$ and the estimated clustering structure is $C = \{C_l\}_{l=1}^L$, then ARI is defined in the following way:

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

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

When defining the true clustering structure for the economic literature data set, we assign to each document first its JEL code. This choice is based on the idea, that the first JEL code represents the primary topic of the document. for this note, the true partitioning and the corresponding matrix of weights are shown on the Figure 4. Here, on the left panel of the Figure 4 different colors serve as different clusters, while on the right panel white and black colors represent weights being equal to 1 and 0 respectively. There are overall 17 clusters.

The biggest cluster consists of 399 documents and appears as the "C" JEL code which stands for *mathematical and quantitative methods*. In fact 65% of the documents contain the JEL code "C". This may be explained by the fact that the majority of papers from this dataset includes ideas based on statistical methods, particularly econometrics. There are two singleton clusters about *economics teaching* and *history of economic thought*.

5 Experiments

In this section we cluster our dataset using different methods and compare the produced clustering structures with the true C^*. The clustering algorithms used in evaluation are AWC, standard k-means and the *vcluster* algorithm from the CLUTO toolkit [4]. k-means and CLUTO both require as a parameter the number of true clusters K. While AWC has only parameter λ, which is tuned using the heuristic described in Section 2.1.1. CLUTO's *vcluster* algorithm is a bisecting graph partitioning-based algorithm which is greedy in nature and therefore depends on the order of the input documents. The k-means algorithm also includes randomness in the clustering process. Thus we run both CLUTO and k-means 50 times with different random states and choose the best result with the maximal ARI for each $k : 2 \leq k \leq 26$. The results

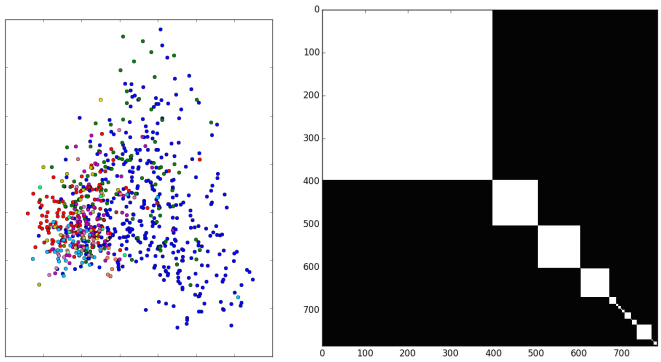

Fig. 4 Left: True clustering structure C^*. Right: Corresponding matrix of weights

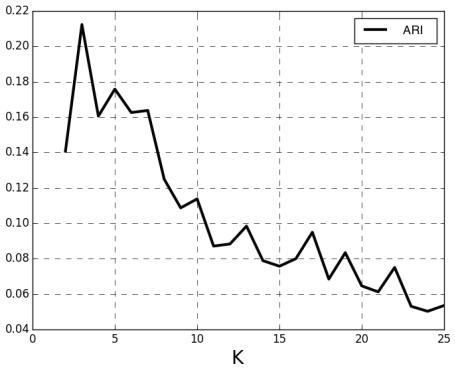

Fig. 5 The ARI values for k-means: 50 runs for each K. Best result for $K = 3$

are shown in Figures 5, 6. As one can see from the plots, both CLUTO and k-means show best performance for $k = 3$.

For AWC we run the algorithm with different λ-values from $[0, 1]$ and compute the sum of weights for each λ-value as proposed in the heuristic for tuning this parameter. Figure 7 demonstrates the dependence between λ and the sum of weights. There are three points before jump which are potential candidates for the parameter λ. As one can see from the plot all three points guarantee ARI measure being higher than 0.2, thus can be considered as good choices for parameter λ. Moreover, ARI measure reaches it's maximum value exactly in the second candidate point. The potential choice of λ is indicated via 3 vectors pointing to the relevant λ value.

To analyze the cluster structure found by AWC we construct for every cluster its word cloud. See Figures 8 - 13. Each word cloud contains the most frequent words in the cluster. Words color varies with words idf. The higher is idf of the word, the darker is its color. Therefore the darkness of a word indicates its importance in the cluster.

We may interpret the detected clusters as follows.

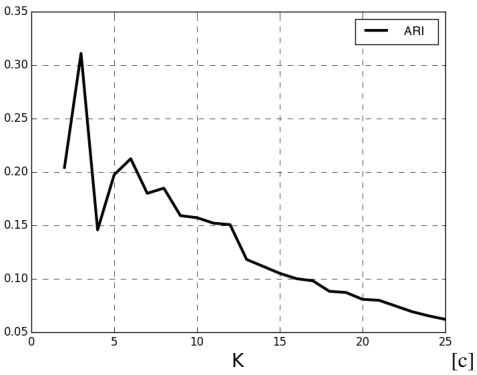

Fig. 6 The ARI values for Cluto: 50 runs for each K. Best result for $K = 3$

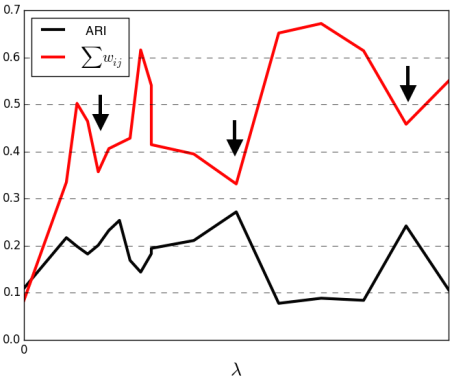

Fig. 7 AWC heuristic choice of the parameter λ

Fig. 8 Word cloud of the first AWC cluster 'Quantitative Methods'

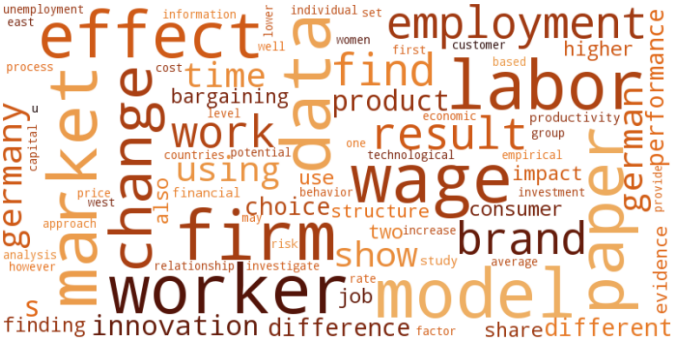

Fig. 9 Word cloud of the second AWC cluster 'Labor Market Analysis'

Fig. 10 Word cloud of the third AWC cluster 'Behavioral Economics'

Fig. 11 Word cloud of the fourth AWC cluster 'Monetary Economics'

Cluster 1 found by AWC contains 46% the JEL code *G*: 'Financial economics' and 81% are assigned to *C*: 'Mathematical and quantitative methods'. In total 86% of the pairs {C, G} are presented in this cluster.

Cluster 2 found by AWC could be described as 'Labor Economics', since 77% contain the JEL code *J*.

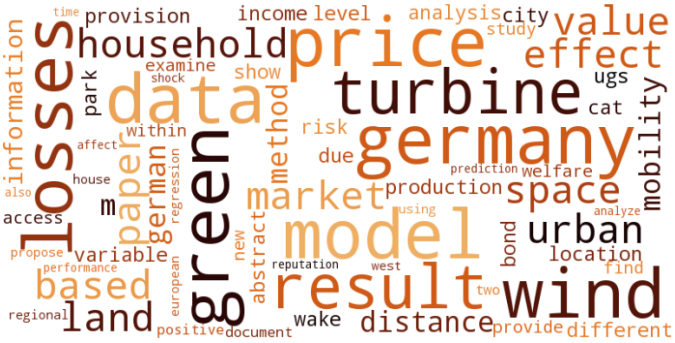

Fig. 12 Word cloud of the fifth AWC cluster 'Green Energy'

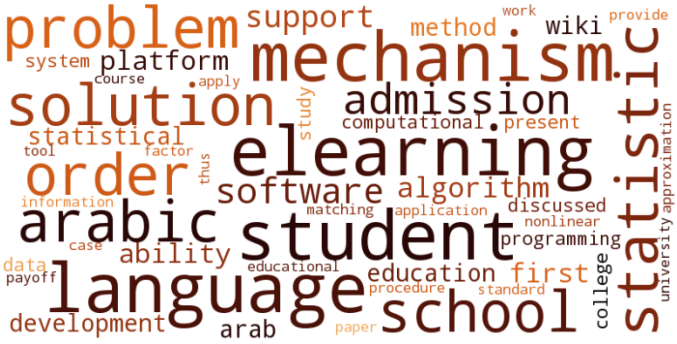

Fig. 13 Word cloud of the sixth AWC cluster

Cluster 3 found by AWC shows a somewhat mixed appearance since 51% contain *D*: 'Microeconomics' and 54% contain *C*: 'Mathematical and quantitative methods'. In total both fields may be combined into the header 'Behavioral Economics'.

Let us look now at the fourth AWC cluster. It contains 73% of the JEL code *E*: 'Macroeconomics and monetary economics'. It also indicates macroeconomics and interest rate as subfields.

Fifth AWC cluster splits up into 32% containing *R*: 'Urban, rural, and regional economic', 24% containing *Q*: 'natural resource economics' and 40% containing *C*: 'Mathematical and quantitative methods'. Thus, this cluster can be related to 'Green Energy Economics'.

Finally sixth cluster has 54% *I*: 'Health, education, and welfare', 80% *C*: 'Mathematical and quantitative methods', also 50% pairs $\{I, C\}$. A somewhat disputable composition.

In order to test the proposed AWC on a different data source we took abstracts from XMU, Xiamen University. In total we obtained 98 paper abstracts from XMU.

Word clouds of the clusters found by AWC are shown on Figures 14 - 17. In the first cluster, see Figure 14, 46% of the abstracts contain the JEL code *G* 'Financial Economics', and 77% contain the JEL code *C* 'Mathematical and quantitative methods'. In fact, 92% of the abstracts contain either *G* or *C*.

Fig. 14 Word cloud of the first AWC cluster

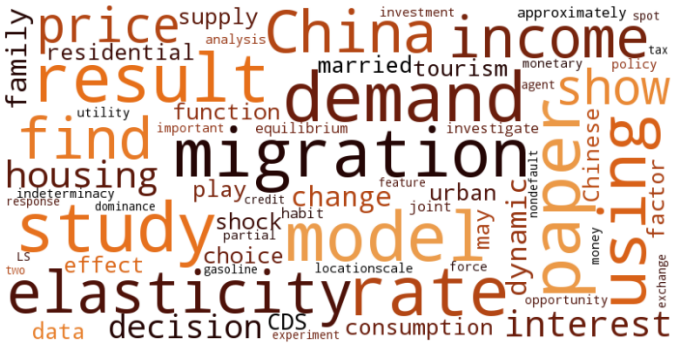

Fig. 15 Word cloud of the second AWC cluster

In the second cluster, see Figure 15, 42% of the abstracts contain the JEL code *E* 'Macroeconomics and Monetary Economics', and 78% contain the JEL code *C* 'Mathematical and quantitative methods'.

In the third cluster, see Figure 16, 69% of the abstracts contain the JEL code *R* 'Urban, Rural, and Regional Economics'.

In the final cluster, see Figure 17, 62% of the abstracts contain the JEL code *L* 'Industrial Organization'. And 75% contain *O* 'Economic Development, Technological Change, and Growth'. And 62% contain joint *L* and*O*.

6 Conclusion

The JEL classification system is a fast way to retrieve corresponding research papers in economics. Based on the CRC HU data we present innovative clustering. It is fully automatic, adaptive and leads to interpretable JEL clusters. The basic idea is based on locally weighting each document or abstract in terms of its cluster membership. The numerical implementation of AWC in Python is available at `www.quantlet.de`. Simulation studies and empirical performance reveal an excellent performance of this new clustering technique.

Fig. 16 Word cloud of the third AWC cluster

Fig. 17 Word cloud of the fourth AWC cluster

Acknowledgements The financial support from the Deutsche Forschungsgemeinschaft, Humboldt-Universität zu Berlin and IRTG 1972 'High Dimensional Non Stationary Time Series' is gratefully acknowledged.

References

1. Thorsrud, Leif Anders. "Words are the new numbers: A newsy coincident index of business cycles." No. 0044. 2016.
2. JEL Classification System. Journal of Economic Literature, vol. 49, no. 4, 2011, pp. 14111425., www.jstor.org/stable/23071718.
3. Lloyd, Stuart. "Least squares quantization in PCM." IEEE transactions on information theory 28.2 (1982): 129-137.
4. Karypis, George. CLUTO-a clustering toolkit. No. TR-02-017. MINNESOTA UNIV MINNEAPOLIS DEPT OF COMPUTER SCIENCE, 2002.
5. Karypis, George, and Vipin Kumar. "A fast and high quality multilevel scheme for partitioning irregular graphs." SIAM Journal on scientific Computing 20.1 (1998): 359-392.
6. Hubert, Lawrence, and Phipps Arabie. "Comparing partitions." Journal of classification 2.1 (1985): 193-218.
7. Abbott, Russ. "What is text processing?." ACM SIGDOC Asterisk Journal of Computer Documentation 4.5 (1977): 20-23.
8. Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108.
9. http://sfb649.wiwi.hu-berlin.de/projects/
10. Kirill Efimov, Larisa Adamyan, Vladimir Spokoiny. "Adaptive Weights Clustering." 2017.
11. Zhang, Junni L., et al. "Distillation of news flow into analysis of stock reactions." Journal of Business & Economic Statistics 34.4 (2016): 547-563.
12. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.
13. Borke, Lukas, and Wolfgang K. Härdle. "Q3-D3-LSA." (2016).

SFB 649 Discussion Paper Series 2017

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Fake Alpha" by Marcel Müller, Tobias Rosenberger and Marliese Uhrig-Homburg, January 2017.
002 "Estimating location values of agricultural land" by Georg Helbing, Zhiwei Shen, Martin Odening and Matthias Ritter, January 2017.
003 "FRM: a Financial Risk Meter based on penalizing tail events occurrence" by Lining Yu, Wolfgang Karl Härdle, Lukas Borke and Thijs Benschop, January 2017.
004 "Tail event driven networks of SIFIs" by Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle and Yarema Okhrin, January 2017.
005 "Dynamic Valuation of Weather Derivatives under Default Risk" by Wolfgang Karl Härdle and Maria Osipenko, February 2017.
006 "RiskAnalytics: an R package for real time processing of Nasdaq and Yahoo finance data and parallelized quantile lasso regression methods" by Lukas Borke, February 2017.
007 "Testing Missing at Random using Instrumental Variables" by Christoph Breunig, February 2017.
008 "GitHub API based QuantNet Mining infrastructure in R" by Lukas Borke and Wolfgang K. Härdle, February 2017.
009 "The Economics of German Unification after Twenty-five Years: Lessons for Korea" by Michael C. Burda and Mark Weder, April 2017.
010 "DATA SCIENCE & DIGITAL SOCIETY" by Cathy Yi-Hsuan Chen and Wolfgang Karl Härdle, May 2017.
011 "The impact of news on US household inflation expectations" by Shih-Kang Chao, Wolfgang Karl Härdle, Jeffrey Sheen, Stefan Trück and Ben Zhe Wang, May 2017.
012 "Industry Interdependency Dynamics in a Network Context" by Ya Qian, Wolfgang Karl Härdle and Cathy Yi-Hsuan Chen, May 2017.
013 "Adaptive weights clustering of research papers" by Larisa Adamyan, Kirill Efimov, Cathy Yi-Hsuan Chen, Wolfgang K. Härdle, July 2017.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".