RESSH 2015, Rennes

University of Antwerp

Clustering of authors through the analysis of their publication patterns

Frederik Verleysen*, Arie Weeren** & Tim Engels*

*ECOOM -University of Antwerp, Belgium

**StatUA - University of Antwerp

University of Antwerp

Introduction

Diversity in SSH beyond disciplinary boundaries?

- beyond fields/disciplines
- beyond the aggregation of SS vs. H
- beyond research units

-> bibliometric analysis at the author level**: *can we show a diversity of SSH publication patterns cutting across disciplinary boundaries?***

-> novel application of distance measures & algorithmic clustering

Publication patterns at higher aggregation levels

Social Sciences

- journal-oriented
- WoS and/or Scopus
- English
- frequent co-authorship

Humanities

- importance of book publications
- national/regional journals
- Dutch
- less frequent co-authorship

Dataset

Data from VABB-SHW (Flanders, Belgium)

- 2000-2011
- 5 publication types
  - Journals (weight 1)
  - Proceedings (weight 0,5)
  - Monographs (weight 4)
  - Edited books (weight 1)
  - Book chapters (weight 1)
- 16 SSH disciplines and 2 general categories
- affiliation with research unit/department -> discipline
- **senior authors (>9 publications, >4 years) (N=2008)**

Method (1)

Premise:

-> Publication patterns of individual authors can be used to explore diversity across disciplines in SSH research and publishing

Technique*: Clustering*

-computer looks for patterns in 'big data'

-based on similarity of observations -> #variables

-explorative, visual method

Method (2)

Step 1: Distance Measure

- How similar/dissimilar are senior authors' publication patterns?
- Calculation of dissimilarity matrix (2008x2008)

Step 2: Clustering Algorithm

- Computer forms clusters of authors w/ similar publication profile
- Visual representation in two dimensions

Method (final)

Total of 11 variables for 2008 authors:

-5 publication types

-publication languages

-share of co-authored publications

Dissimilarity matrix based on *Mahalabonis distance* (correlations between variables)

K-medoids clustering algorithm: clustering around optimal cluster centers + MDS

All implemented in MATLAB

Method (3)

Data input for clustering: weighted fractions of author output (fictitious example)

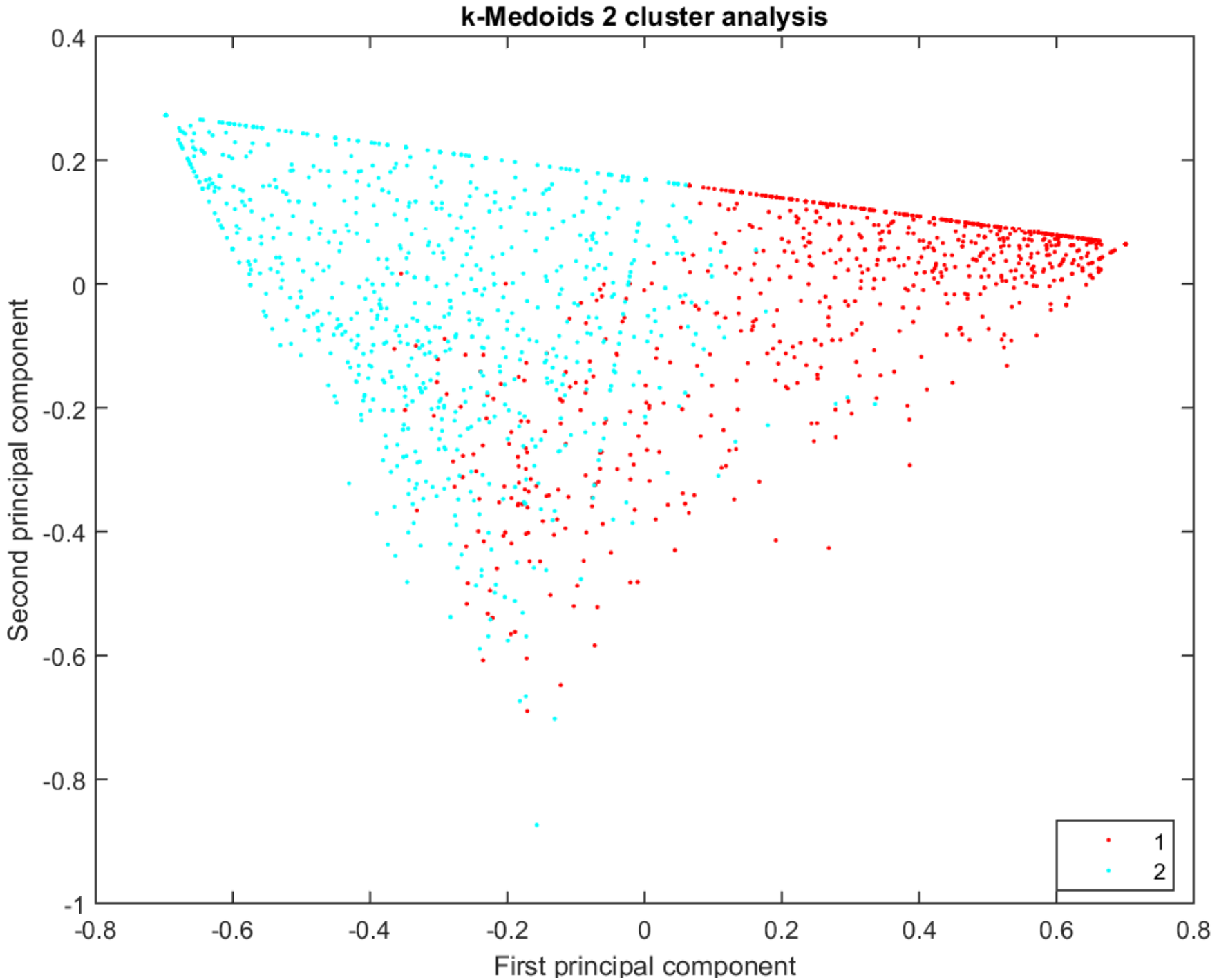
Author	discipline	# Publication type x	# Publication language y	# Co-authorships	Total weighted publications
author1	History	#/25	#/25	#/25	25
author2	Psychology	#/60	#/60	#/60	60
author3	Linguistics	#/40	#/40	#/40	40
author4	Sociology	#/30	#/30	#/30	30
…	…	…	…	…	…

Results

1. Clustering of all authors (N=2008)

2. Location of authors within their discipline

3. Locations of disciplines relative to each other


k-Medoids 2 cluster analysis
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2

Exploring the two clusters

Cluster 1

- journal-oriented (WoS)
- English
- frequent co-authorships

-> = Social Scientists?

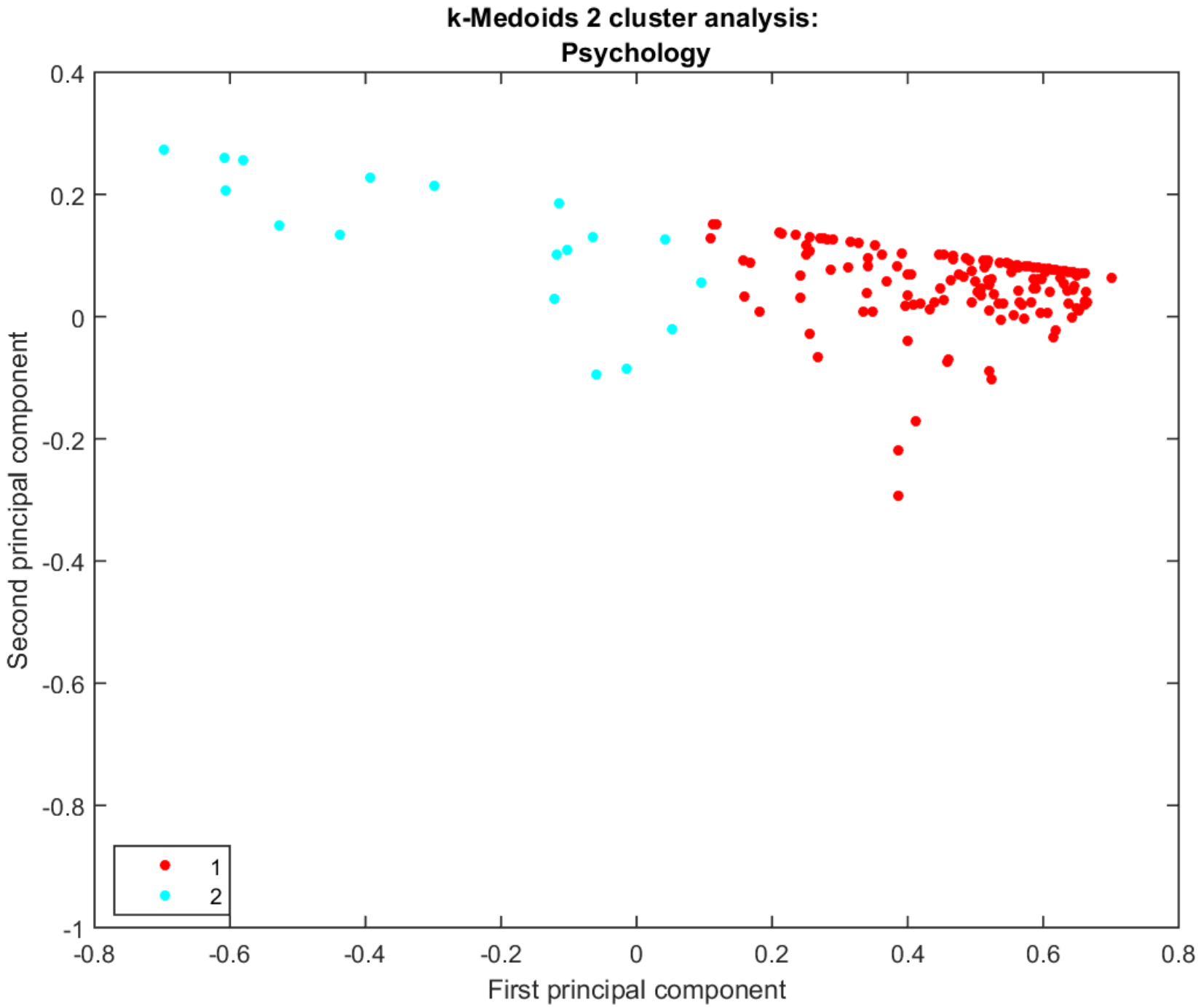
Cluster 2

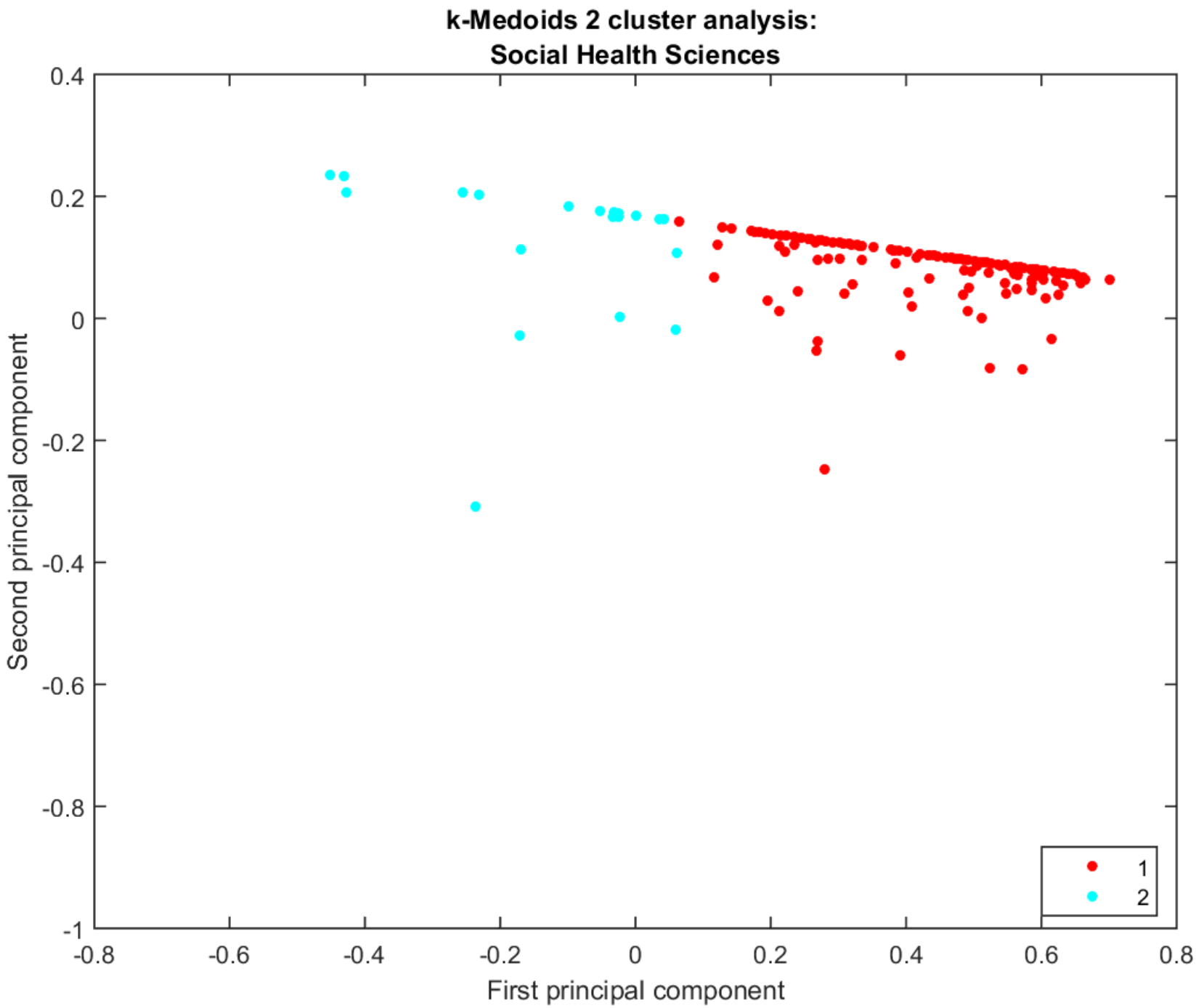
- importance of book publications
- national/regional journals
- Dutch
- sole authorship

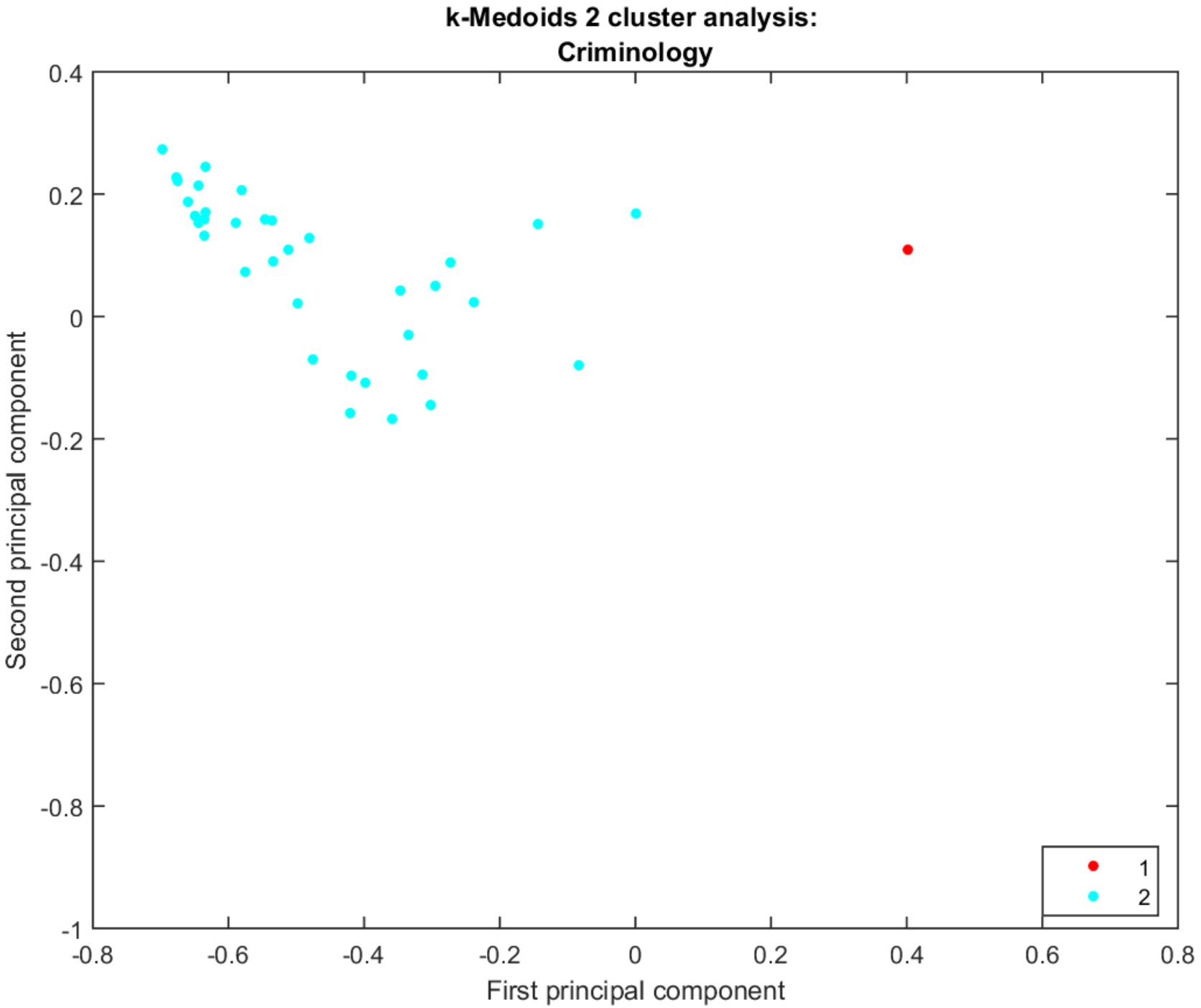
-> = Humanities scholars?

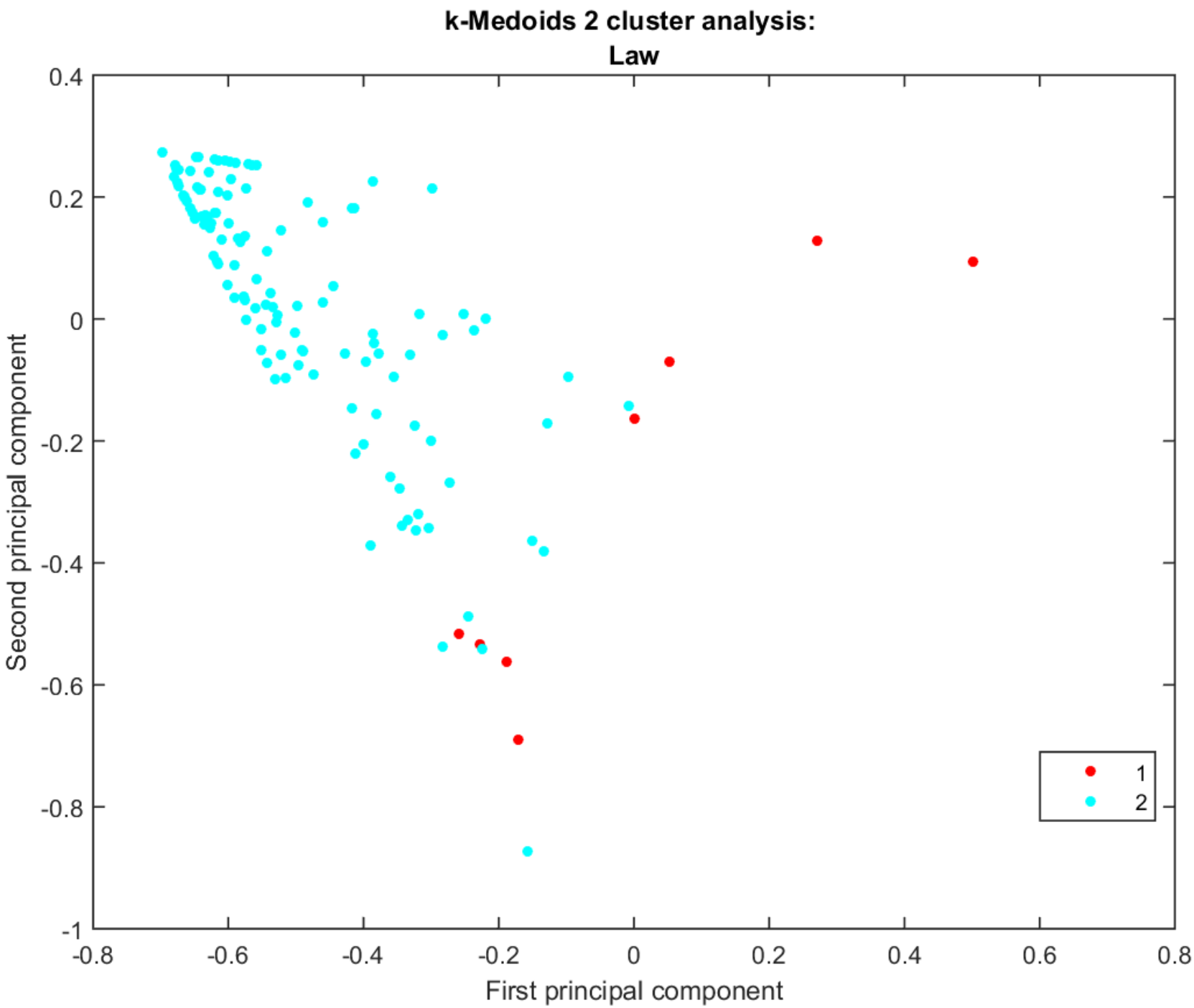
Results

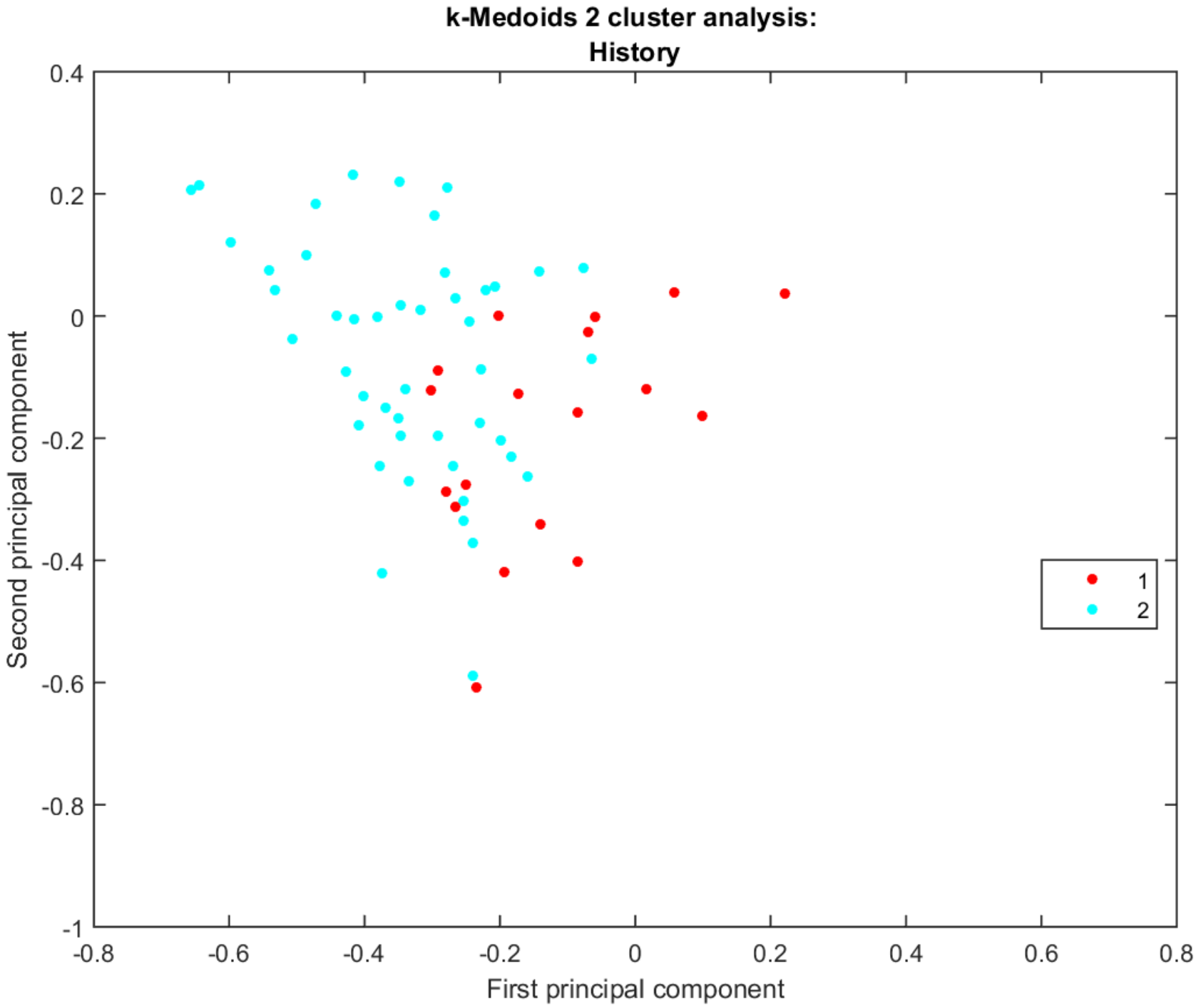
Location of authors within their discipline

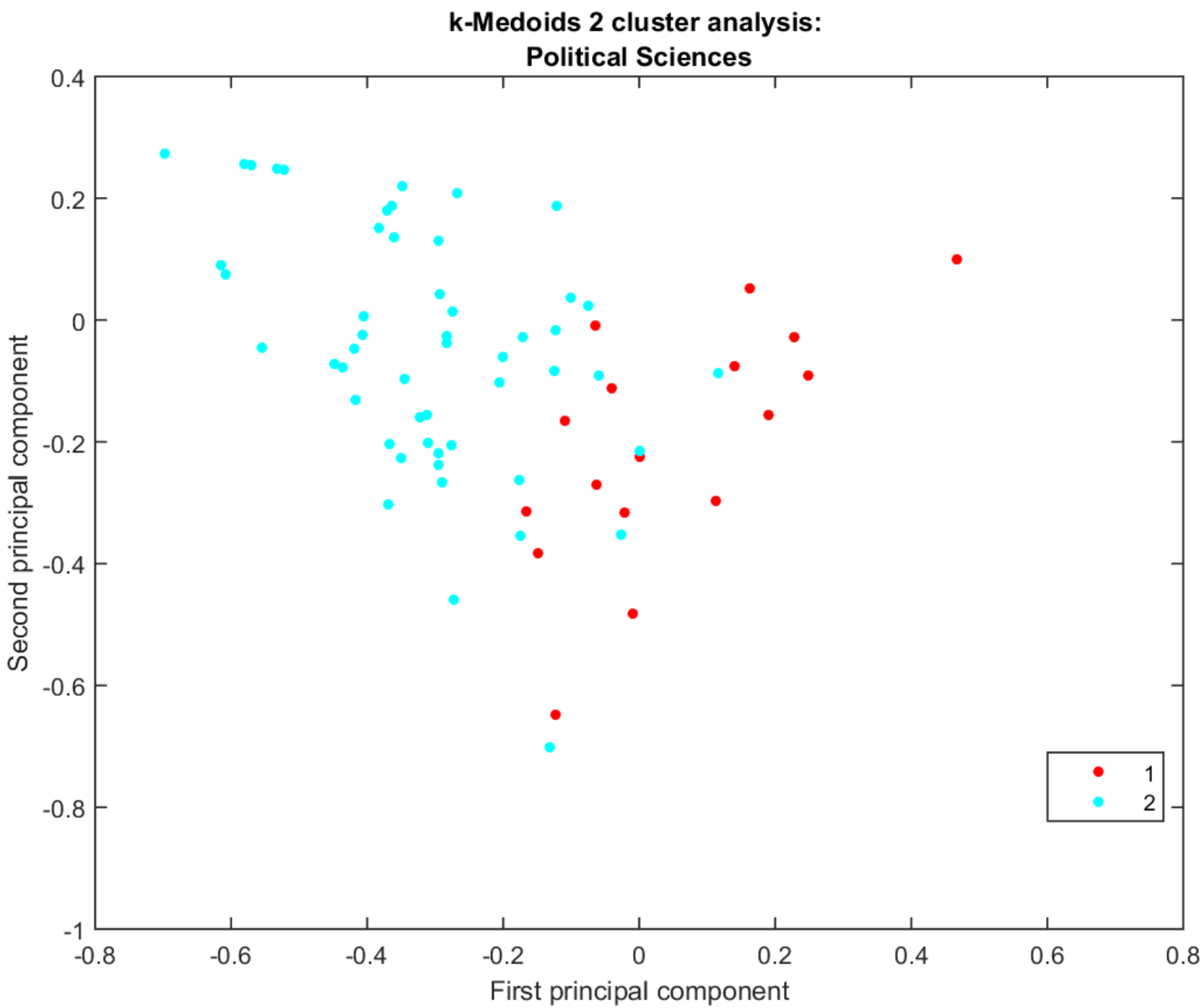

k-Medoids 2 cluster analysis:
Psychology
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2

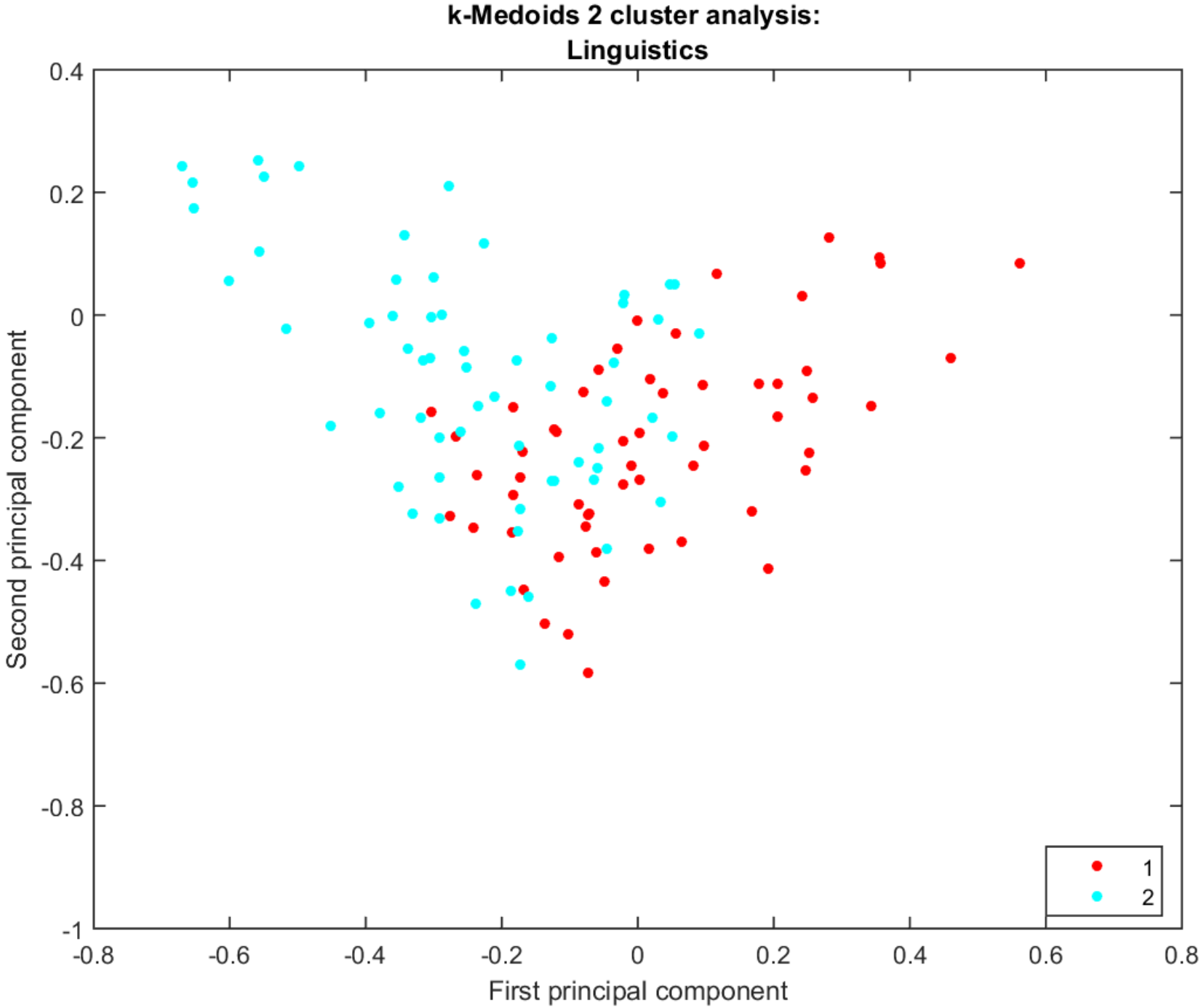

k-Medoids 2 cluster analysis:
Social Health Sciences
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2

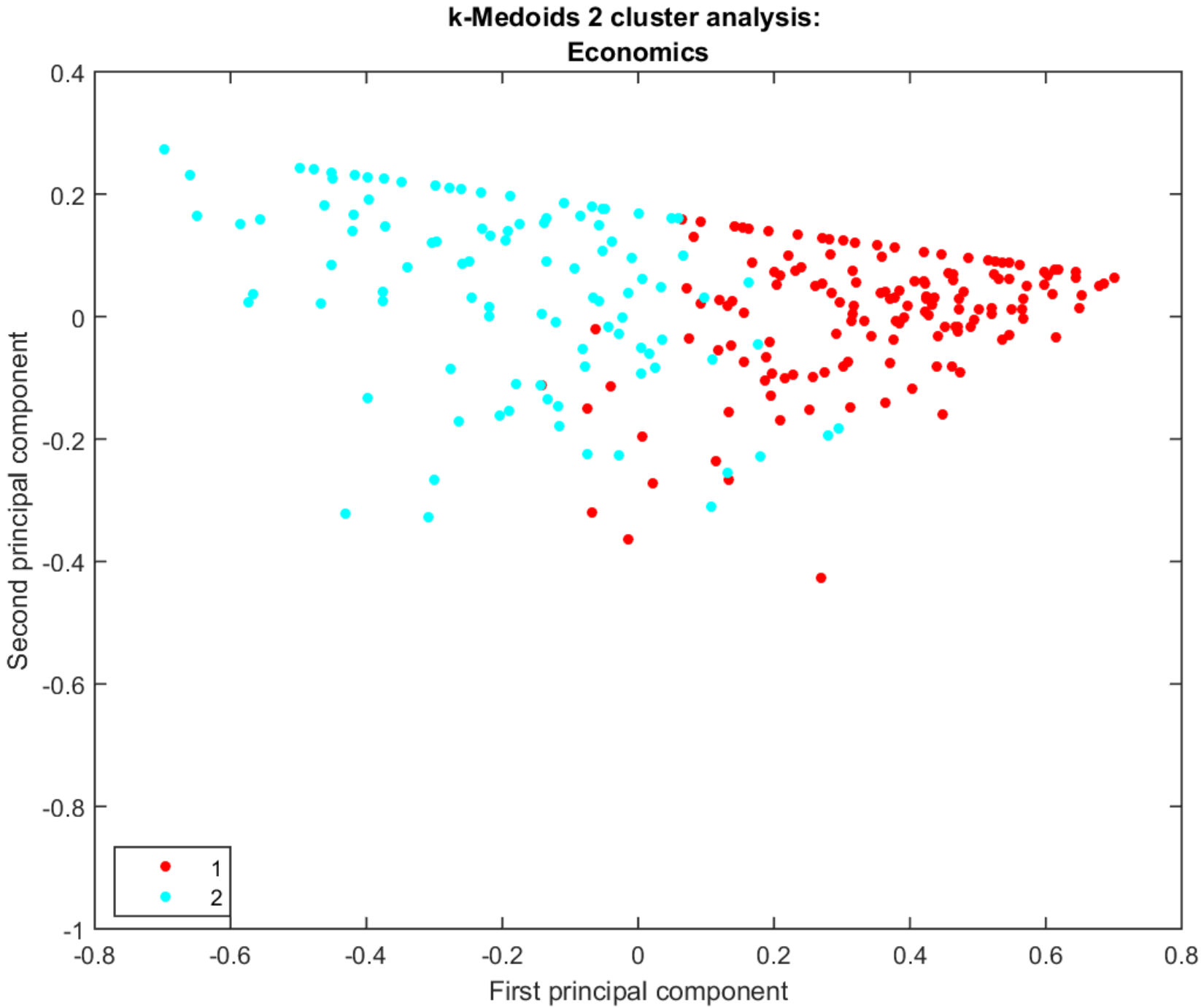

k-Medoids 2 cluster analysis:
Criminology
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2


k-Medoids 2 cluster analysis:
Law
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2


k-Medoids 2 cluster analysis:
History
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2


k-Medoids 2 cluster analysis:
Political Sciences
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2


k-Medoids 2 cluster analysis:
Linguistics
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2


k-Medoids 2 cluster analysis:
Economics
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2

SS Disciplines in 2 clusters in %

Discipline	% Cluster 1	% Cluster 2
Criminology	2.5	97.5
Economics	60.5	39.5
Educational Sciences	44.8	55.2
Social Health Sciences	88.1	11.9
Sociology	47.7	52.3
Political Sciences	22.9	77.1
Psychology	90.9	9.1

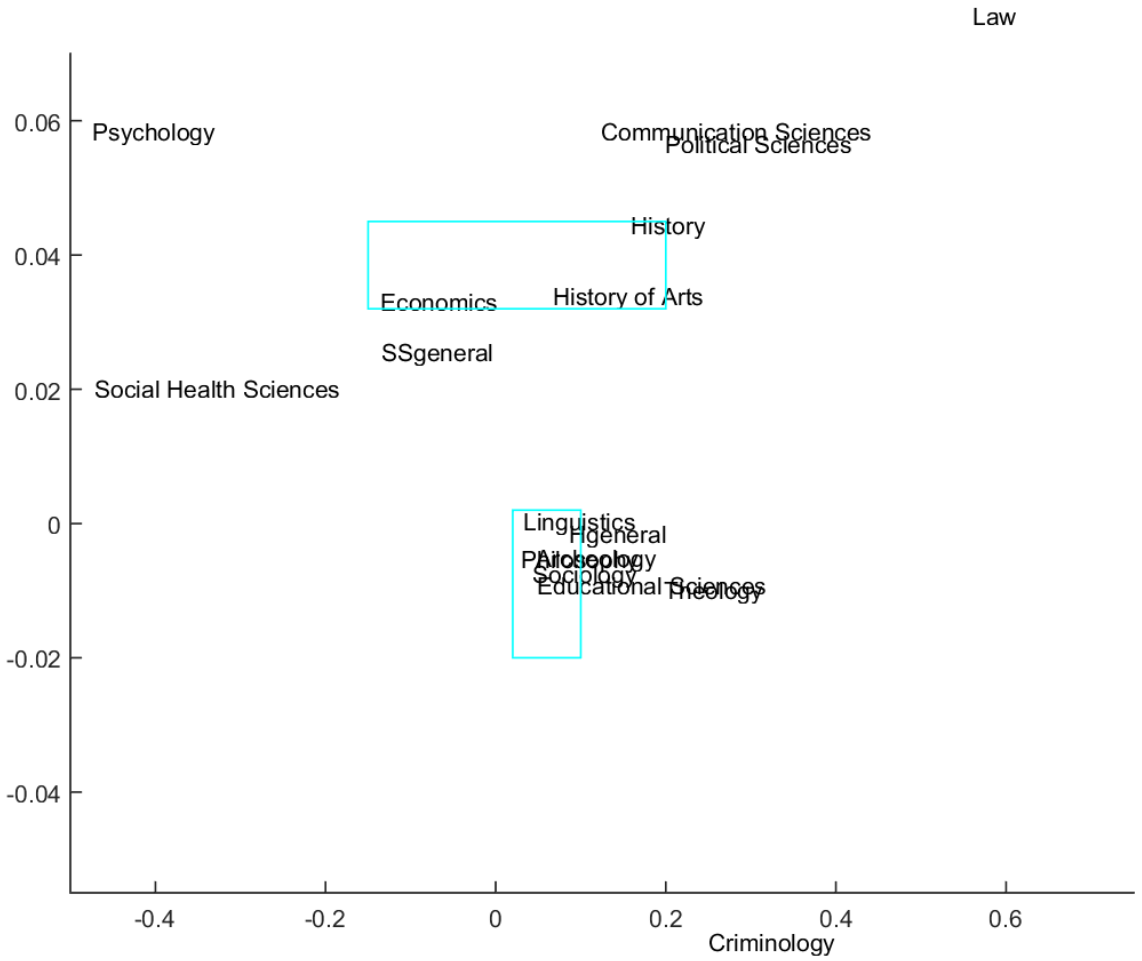
H Disciplines in 2 clusters in %

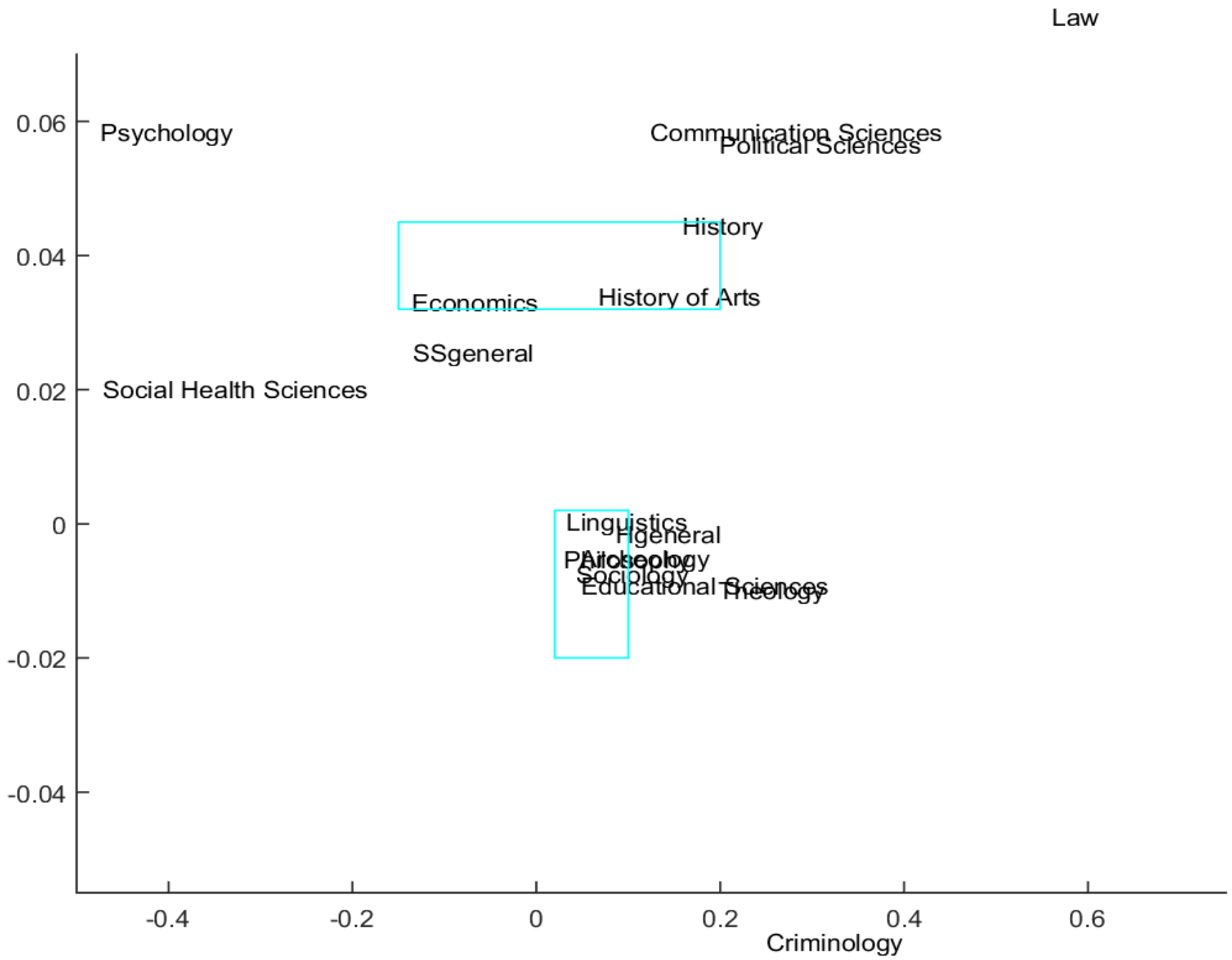
Discipline	% Cluster 1	% Cluster 2
Archaeology	40.6	59.4
Communication Sciences	25.7	74.3
History	27.2	72.8
History of Arts	37.1	62.9
Law	3.8	96.4
Linguistics	47.1	52.9
Literature	51.6	48.4
Philosophy	47.2	52.8
Theology	18.5	81.5

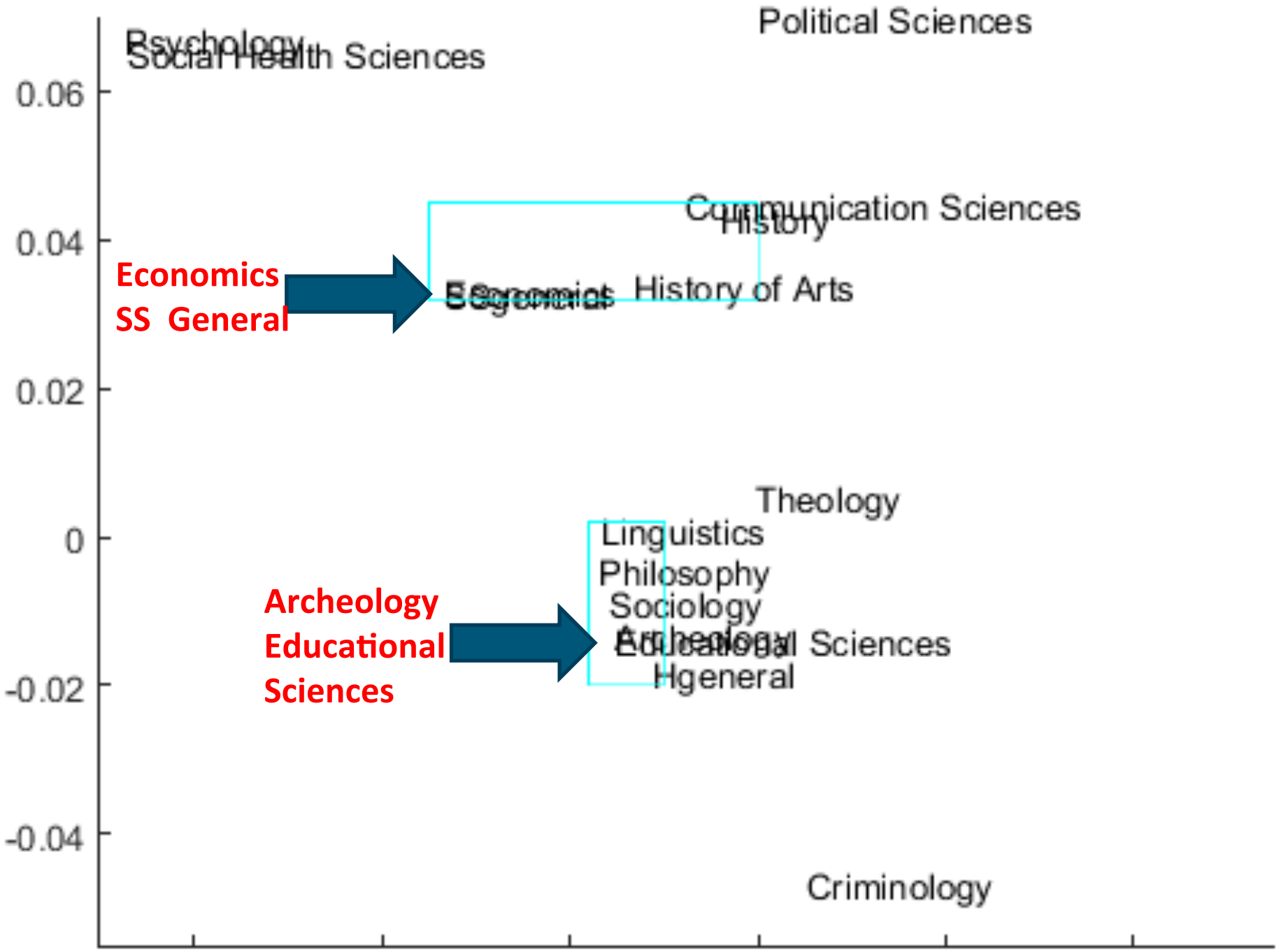
Results

Locations of disciplines relative to each other

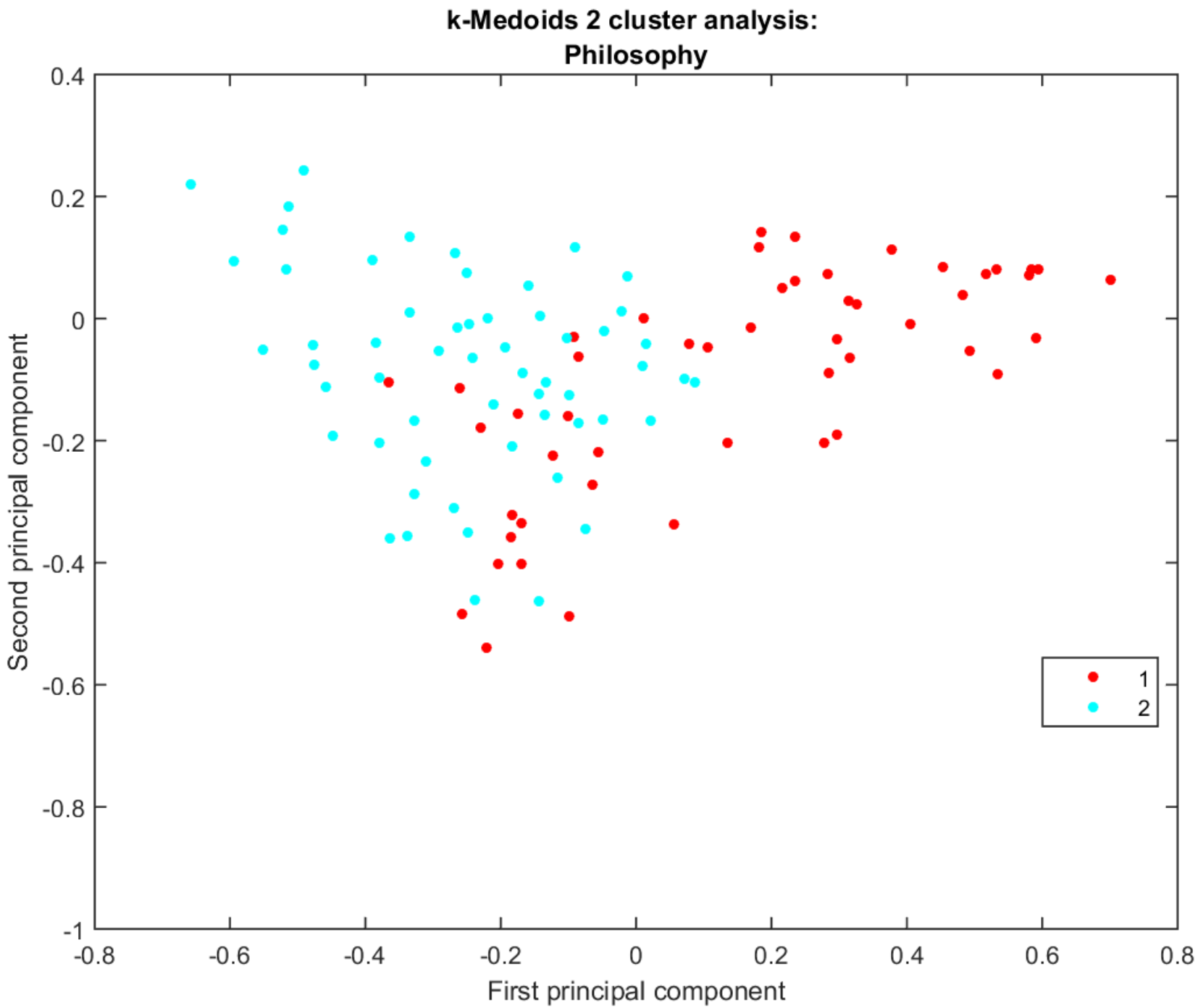
Relative position of Disciplines

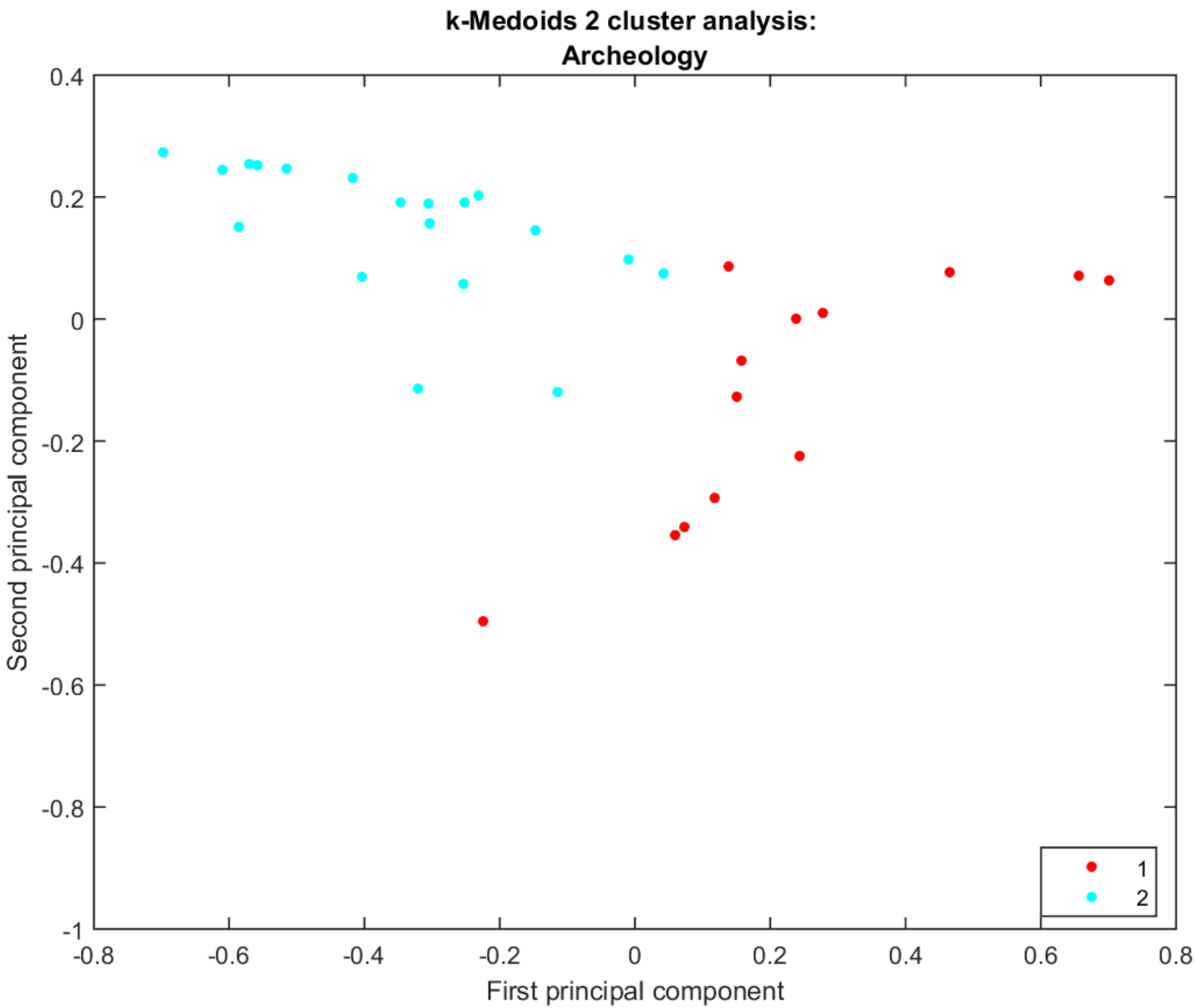


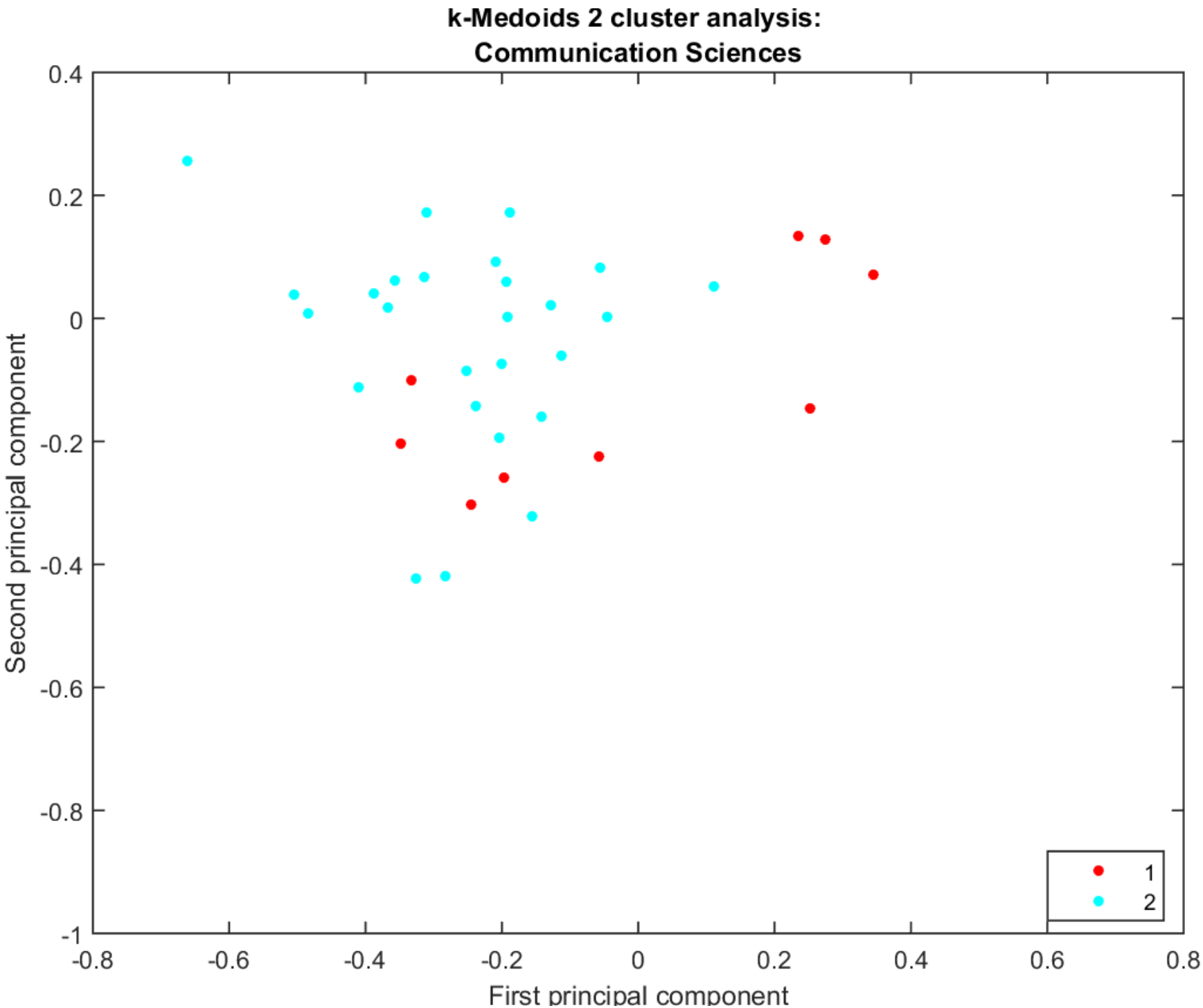

Law
Psychology
Communication Sciences
Political Sciences
History
Economics
History of Arts
SSgeneral
Social Health Sciences
Linguistics
Figeneral
Sociology
Educational Sciences
Theology
0.06
0.04
0.02
0
-0.02
-0.04
-0.4
-0.2
0
0.2
0.4
0.6
Criminology

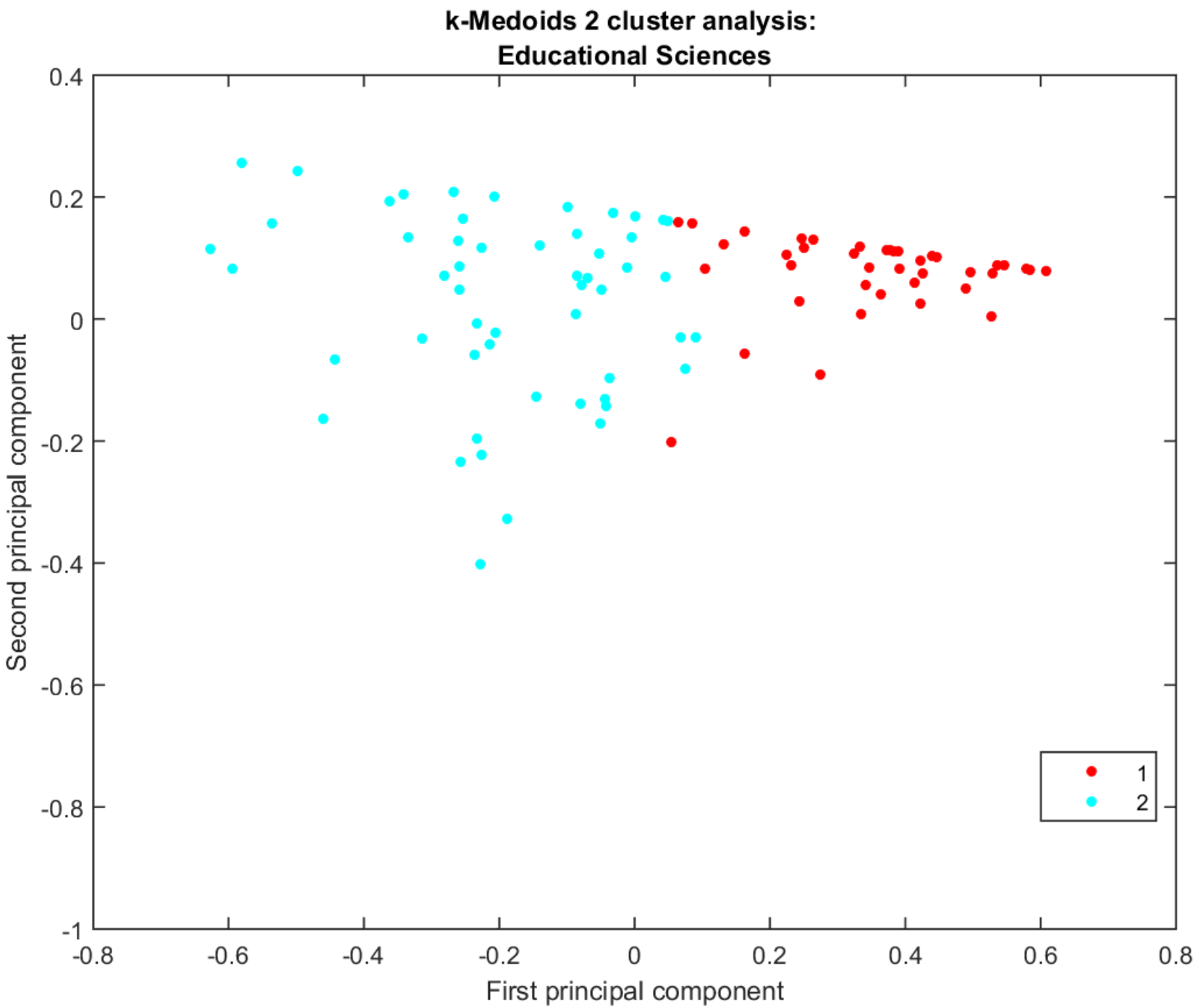

Political Sciences
Psychology
Social Health Sciences
0.06
0.04
0.02
0
-0.02
-0.04
Communication Sciences
History
History of Arts
Economics
SS General
Theology
Linguistics
Philosophy
Sociology
Archeology
Educational Sciences
Hgeneral
Archeology
Educational
Sciences
Criminology

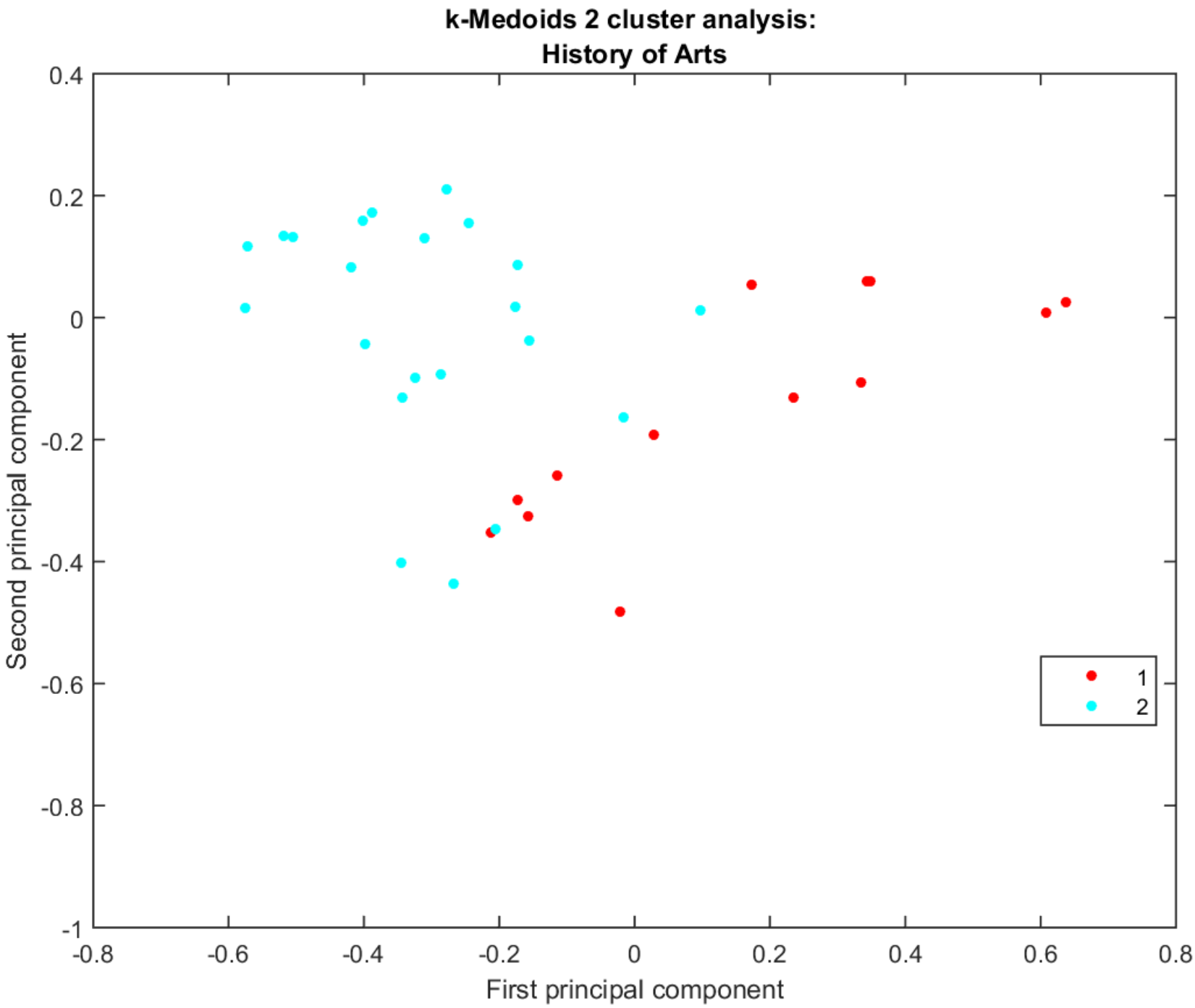
University
of Antwerp

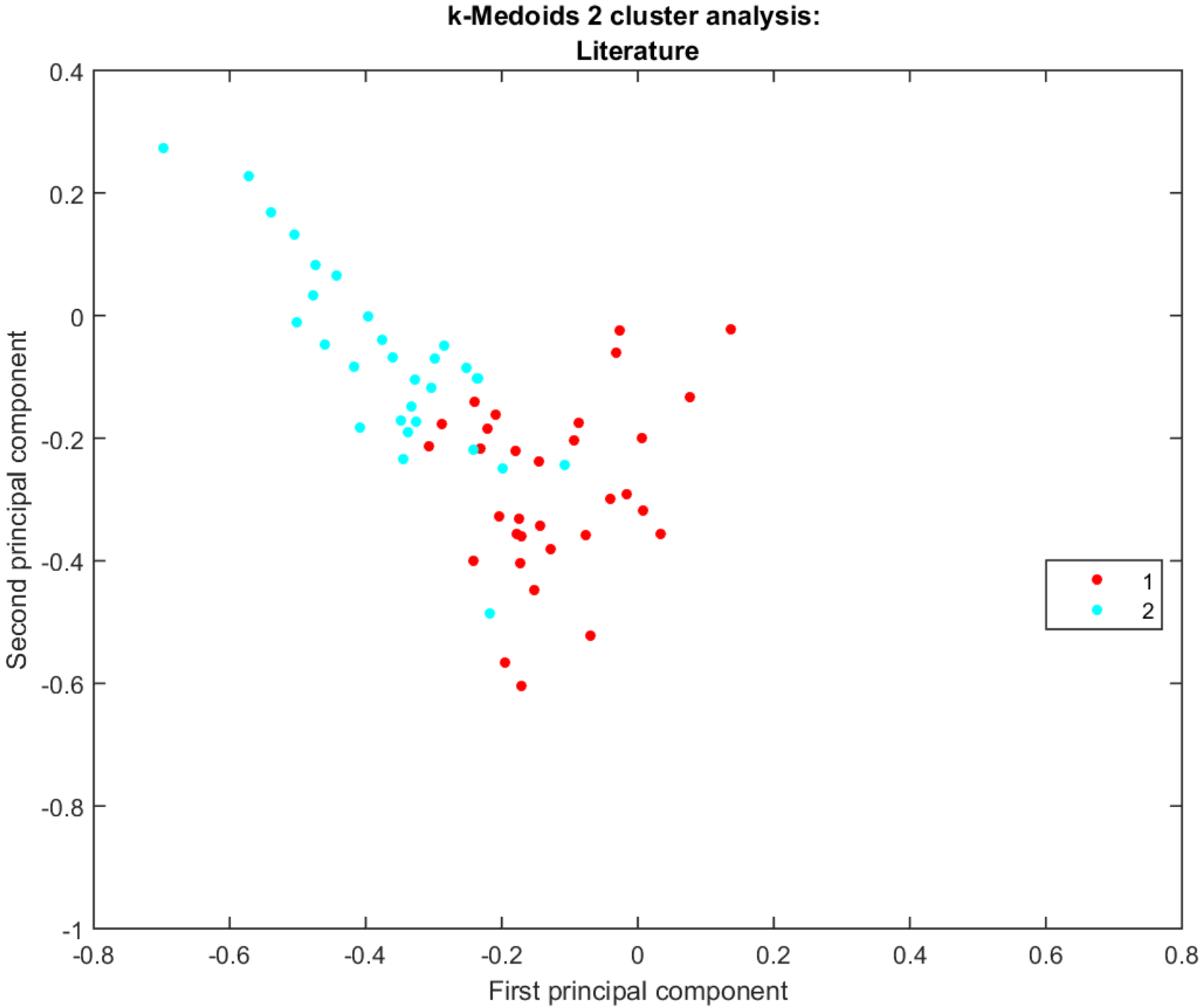

k-Medoids 2 cluster analysis:
Philosophy
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2

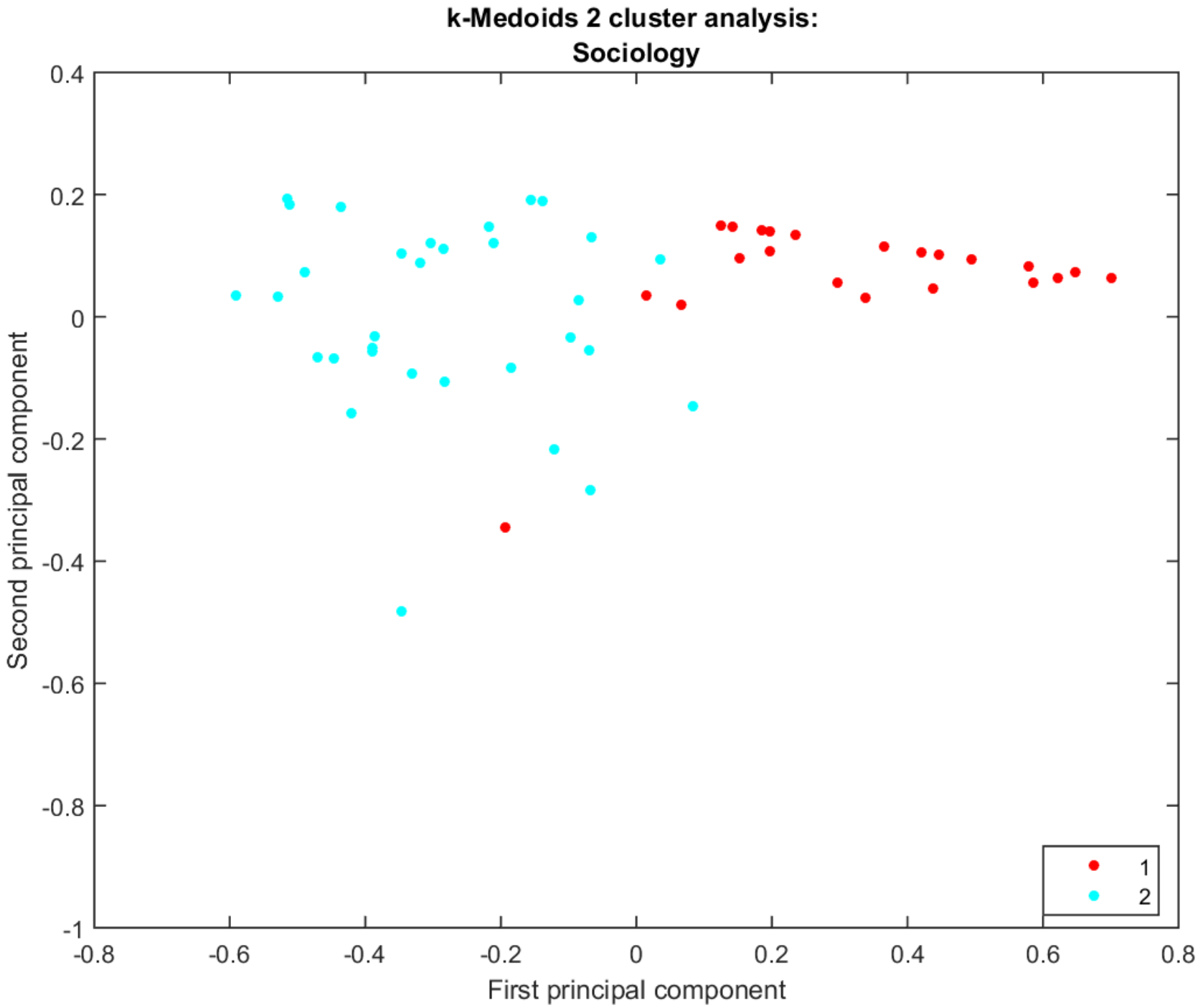

k-Medoids 2 cluster analysis:
Archeology
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2

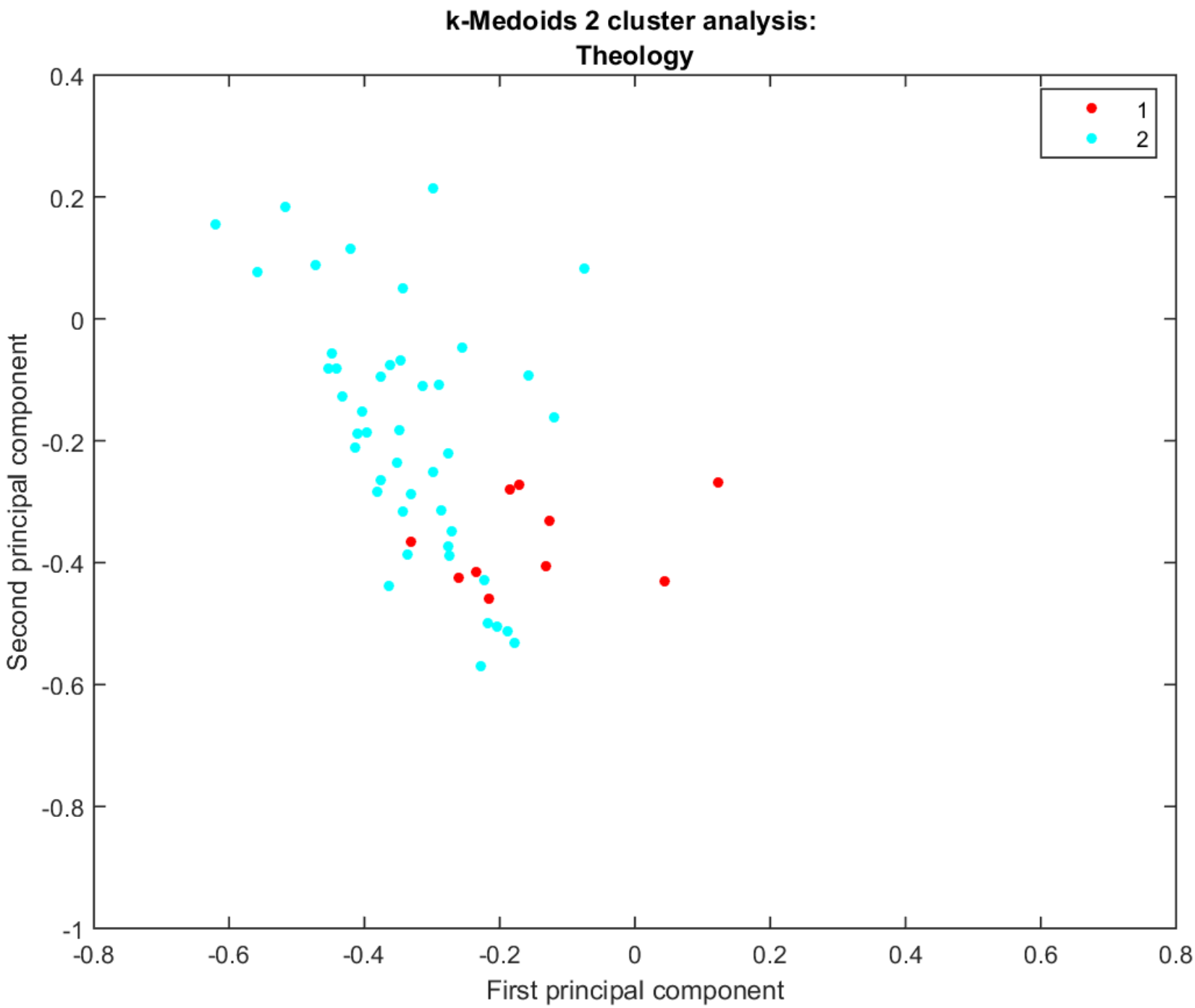

k-Medoids 2 cluster analysis:
Communication Sciences
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2


k-Medoids 2 cluster analysis:
Educational Sciences
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2


k-Medoids 2 cluster analysis:
History of Arts
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2


k-Medoids 2 cluster analysis:
Literature
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2


k-Medoids 2 cluster analysis:
Sociology
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
2


k-Medoids 2 cluster analysis:
Theology
1
2
Second principal component
First principal component
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8