

Projet 3A - Extracting Insights From NBA Data Using Topology

Gabriel Azevedo Ferreira

Gustavo Castro

Henrique Gasparini Fiuza do Nascimento

March 16, 2017

École Polytechnique

Table of contents

1. Introduction
2. Extracting the data
3. Visualizing the data via dimensionality reduction
4. Feature selection and clustering analysis
5. Mapper

Introduction

Contextualizing

- Paper published by M. Alagappan in 2013



Extracting insights from the shape of complex data using topology

P. Y. Lum¹, G. Singh¹, A. Lehman¹, T. Ishkanov¹, M. Vejdemo-Johansson², M. Alagappan¹, J. Carlsson³ & G. Gorinson^{1,4}

¹Agendo Inc., Palo Alto, CA, Telcordia Research Center, Bell Core Building, North Haugh, St. Andrews KY16 9SE, Scotland, United Kingdom; ²Automation and System Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN55455, USA,

³Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

SUBJECT AREAS:

APPLIED MATHEMATICS

COMPUTATIONAL SCIENCE

SCIENTIFIC DATA

SOFTWARE

Received

13 September 2012

Accepted

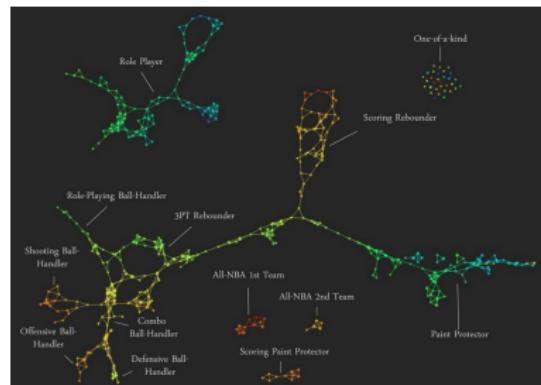
6 December 2012

Published

7 February 2013

Correspondence and
requests for materials

This paper applies topological methods to study complex high dimensional data sets by extracting shapes (patterns) and obtaining insights about them. Our method combines the best features of existing standard methodologies in order to analyze complex high dimensional data sets. This hybrid method is able to analyze complex data sets through this hybrid method, we often find subgroups in data sets that traditional methodologies fail to find. Our method also permits the analysis of individual data sets as well as analysis related to multiple data sets. We demonstrate the use of our method by applying it to three very different kinds of data, namely gene expression from breast tumors, voting data from the United States House of Representatives and player performance data from the NBA, in each case finding stratifications of the data which are more refined than those produced by standard methods.



[M. Alagappan, 2012]

Contextualizing

- Paper published by M. Alagappan in 2013
 - Method Mapper applied to NBA players of 2010-2011 season



Extracting insights from the shape of complex data using topology



SUBJECT AREAS:
APPLIED MATHEMATICS
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
SOFTWARE

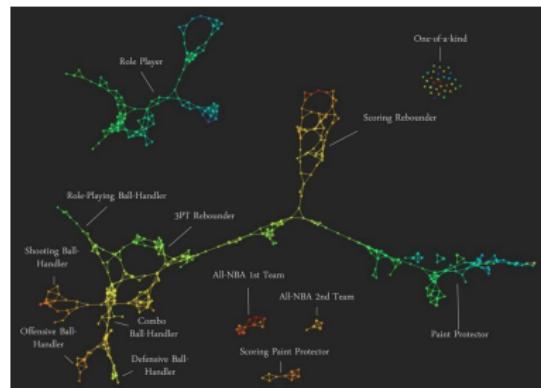
P. Y. Lum¹, G. Singh¹, A. Lehman¹, T. Ishkanov¹, M. Vejdemo-Johansson², M. Alagappan¹, J. Carlsson³& G. Carlsson^{1,4}

¹Maple Inc., Palo Alto, CA, ²School of Computer Science, Bell Chair Building, North Haugh, St Andrews KY16 9SS, Scotland, United Kingdom, ³Computer and System Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN55455, USA, ⁴Department of Mathematics, Stanford University, Stanford, CA, 94305, USA

Received
13 September 2012
Accepted
6 December 2012
Published
7 February 2013

Correspondence and
requests for materials

This paper applies topological methods to study complex high dimensional data sets by extracting shapes (patterns) and obtaining insights about them. Our method combines the best features of existing standard methodologies to find. Our method also permits the analysis of individual data sets as well as analysis related to multiple data sets. We demonstrate the use of our method by applying it to three very different kinds of data, namely gene expression from breast tumors, voting data from the United States House of Representatives and player performance data from the NBA, in each case finding stratifications of the data which are more refined than those produced by standard methods.



[M. Alagappan, 2012]

Contextualizing

- Paper published by M. Alagappan in 2013
 - Method Mapper applied to NBA players of 2010-2011 season
 - Lead to a different classification of playing styles



Extracting insights from the shape of complex data using topology

P. Y. Lum¹, G. Singh¹, A. Lehman¹, T. Ishkhanov², M. Vejdemo-Johansson², M. Alagappan¹, J. Carlsson² & G. Carlsson^{1,2*}

¹Mapd Inc., Palo Alto, CA, ²School of Computer Science, Bell Chair Building, North Haugh, St Andrews KY16 9SS, Scotland, United Kingdom and ²Computer and Systems Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN55455, USA.

*Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

Received

13 September 2012

Accepted

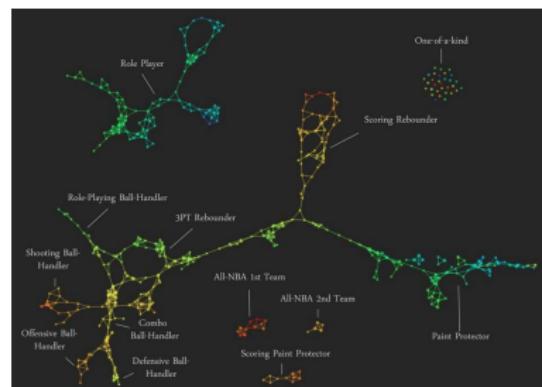
6 December 2012

Published

7 February 2013

Correspondence and
requests for materials

This paper applies topological methods to study complex high dimensional data sets by extracting shapes (patterns) and obtaining insights about them. Our method combines the best features of existing standard methodologies to find clusters in complex high dimensional data sets. This hybrid method also permits the analysis of individual data points and their relationships. We also present the use of our method by applying it to three very different kinds of data, namely gene expression from breast tumors, voting data from the United States House of Representatives and player performance data from the NBA, in each case finding stratifications of the data which are more refined than those produced by standard methods.



[M. Alagappan, 2012]

Our goals

- Analyse data using classical methods of clustering



Extracting insights from the shape of complex data using topology

SUBJECT AREAS:
APPLIED MATHEMATICS
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
SOFTWARE

P. Y. Lum¹, G. Singh¹, A. Leinhardt¹, T. Mekhora¹, H. Vejdemo-Johansson¹, M. Aylagappa¹, J. Carlsson²
& G. Carlsson^{1*}

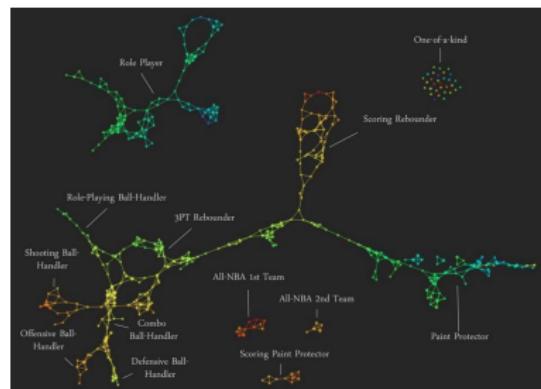
¹Apogee Inc., Palo Alto, CA, ²School of Computer Science, Jack Cole Building, North Haugh, St Andrews KY10 9SX, Scotland, United Kingdom. ¹School of Computer Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN55455, USA.

*Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

Received 13 September 2012
Accepted 6 December 2012
Published 7 February 2013

Correspondence and
requests for materials

This paper applies topological methods to study complex high-dimensional data sets by extracting shapes (patterns) and obtaining insights about them. Our method combines the best features of existing standard methodologies such as principal component and cluster analysis to provide a geometric representation of data sets. In addition, our method is able to handle complex data sets where standard statistical and methodologies fail or find. Our method also permits the analysis of individual data sets as well as the analysis of relationships between related data sets. We illustrate the use of our method by applying it to three very different data sets: namely, network data from the US House of Representatives, data from the United States House of Representatives, and player performance data from the NBA, in each case finding classifications of the data which are more refined than those produced by standard methods.



Our goals

- Analyse data using classical methods of clustering
- Use Mapper algorithm to NBA's last season



Extracting insights from the shape of complex data using topology

SUBJECT AREAS:
APPLIED MATHEMATICS
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
SOFTWARE

P. Y. Lum¹, G. Singh¹, A. Leinhardt¹, T. Mekhora¹, H. Vejdemo-Johansson¹, M. Aikopyan², J. Carlsson³ & G. Carlsson^{4*}

¹Apollon Inc., Palo Alto, CA, ²School of Computer Science, Jack Cole Building, North Haugh, St Andrews KY10 9SX, Scotland, United Kingdom, ³Industrial and Systems Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN55455, USA, ⁴Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

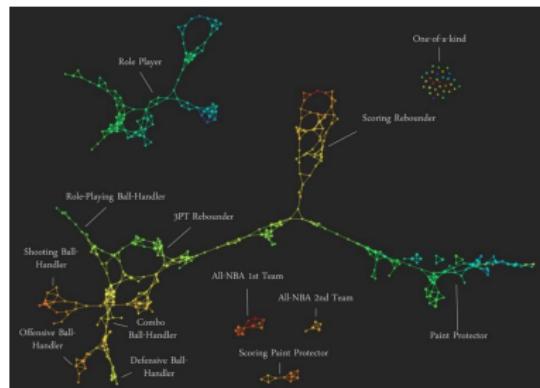
*Correspondence and requests for materials

Received 13 September 2012

Accepted 6 December 2012

Published 7 February 2013

This paper applies topological methods to study complex high-dimensional data sets by extracting shapes (patterns) and obtaining insights about them. Our method combines the best features of existing standard methodologies such as principal component and cluster analysis to provide a geometric representation of data sets that is both meaningful and useful. In addition, our method is able to find structures that standard methodologies fail to find. Our method also permits the analysis of individual data sets as well as the analysis of relationships between related data sets. We illustrate the use of our method by applying it to three very different data sets: namely, the congressional roll call data from the US House of Representatives and player performance data from the NBA, in each case finding stratifications of the data which are more refined than those produced by standard methods.



Our goals

- Analyse data using classical methods of clustering
- Use Mapper algorithm to NBA's last season
- Perform a similar analysis that Alagapan did



Extracting insights from the shape of complex data using topology

SUBJECT AREAS:

APPLIED MATHEMATICS

COMPUTATIONAL SCIENCE

SCIENTIFIC DATA

SOFTWARE

P. Y. Lum¹, G. Singh¹, A. Leinhardt¹, T. Mekhora¹, H. Vejdemo-Johansson¹, M. Alagappan¹, J. Carlsson²

& G. Carlsson^{1*}

¹Apollon Inc., Palo Alto, CA, ²School of Computer Science, Jack Cole Building, North Haig, St Andrews KY10 9SX, Scotland, United Kingdom, ¹Department of Computer Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN55455, USA

*Department of Mathematics, Stanford University, Stanford, CA, 94305, USA.

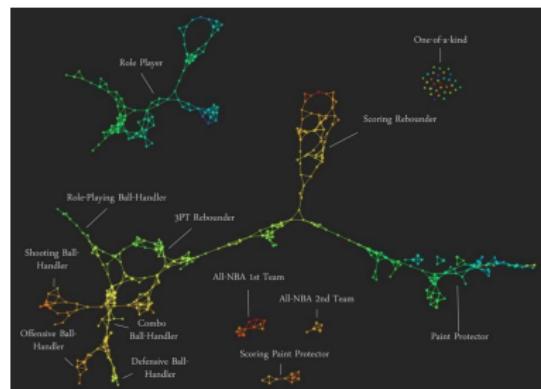
Received 13 September 2012

Accepted 6 December 2012

Published 7 February 2013

Correspondence and
requests for materials

This paper applies topological methods to study complex high-dimensional data sets by extracting shapes (patterns) and obtaining insights about them. Our method combines the best features of existing standard methodologies such as principal component and cluster analyses to provide a geometric representation of data sets that is both meaningful and useful. In addition, our method is able to find structures that other methodologies fail to find. Our method also permits the analysis of individual data sets as well as the analysis of relationships between related data sets. We illustrate the use of our method by applying it to three very different data sets: namely, the 2012 US House of Representatives election data from the United States House of Representatives and player performance data from the NBA, in each case finding stratifications of the data which are more refined than those produced by standard methods.



Extracting the data

Extracting the data

- NBA official website
 - Data per Game (points per game, Blocks per game...)
 - 2015-2016 season
- ESPN website (obtain player's positions)

The screenshot shows the NBA official website with the URL www.nba.com. The page title is "2015-16 Player Official Leaders". It features a search bar and navigation links for Home, Statistics, Players, Teams, Scores, Schedule, and Standings. The main content area displays a table of player statistics for the Regular Season, including columns for MP, FG%, FT%, 3P%, TRB, AST, STL, BLK, and PF. The top five players listed are James Harden, DeMar DeRozan, Kawhi Leonard, Kyle Lowry, and Chris Paul.

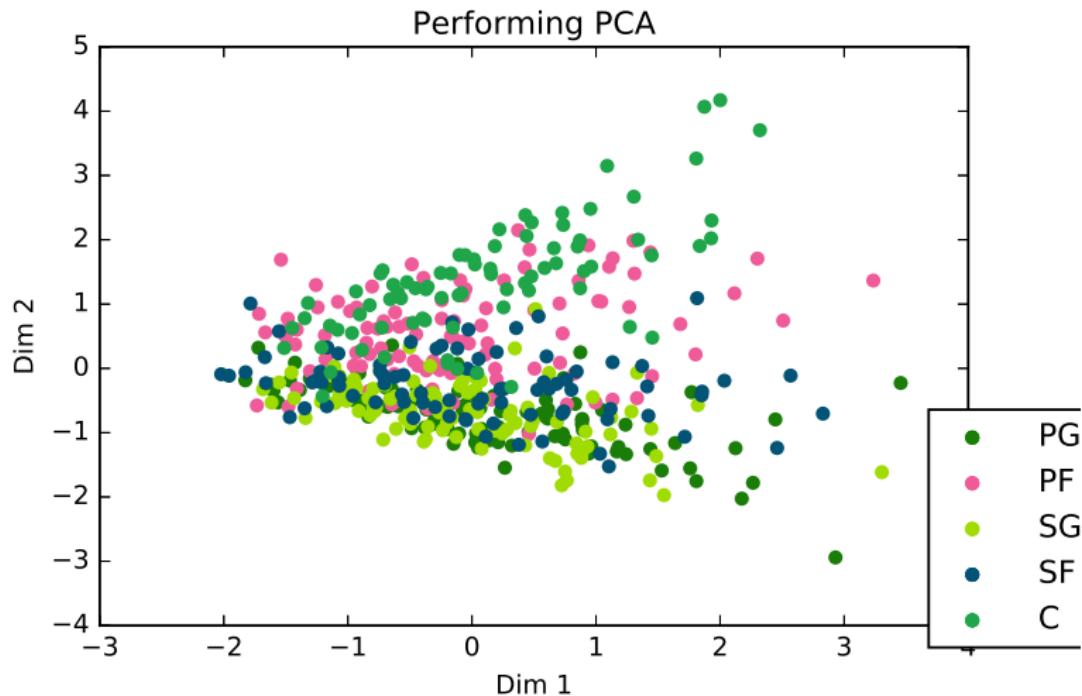
#	PLAYER	MP	FG%	FT%	3P%	TRB	AST	STL	BLK	PF	GP	GS
1	James Harden	80	47.0	77.6	4.2	9.0	4.8	1.5	2.0	1.0	70	70
2	DeMar DeRozan	76	47.0	77.3	4.0	10.0	8.2	1.5	2.0	1.0	69	69
3	Kawhi Leonard	76	47.6	80.0	4.0	9.4	8.2	1.5	1.7	1.0	69	69
4	Kyle Lowry	76	47.0	77.0	4.0	10.2	7.5	1.5	2.0	1.0	69	69
5	Chris Paul	76	47.0	77.0	4.0	10.2	7.5	1.5	2.0	1.0	69	69

The screenshot shows the ESPN website with the URL espn.go.com/nba/statistics/minutes. The page title is "NBA Player Minutes Statistics - 2015-16". It includes filters for Season (2015-16 Regular Season), League (NBA), Splits (Total), Position (All), and Qualification (All Players). The main content area displays a table titled "Minutes Per Game Leaders - Qualified" with columns for RANK, PLAYER, TEAM, GP, MIN, and MGS. The top four leaders are James Harden (SG, HOU), Kyle Lowry (PG, TOR), Jimmy Butler (SF, CHI), and Kentavious Caldwell-Pope (SG, DET).

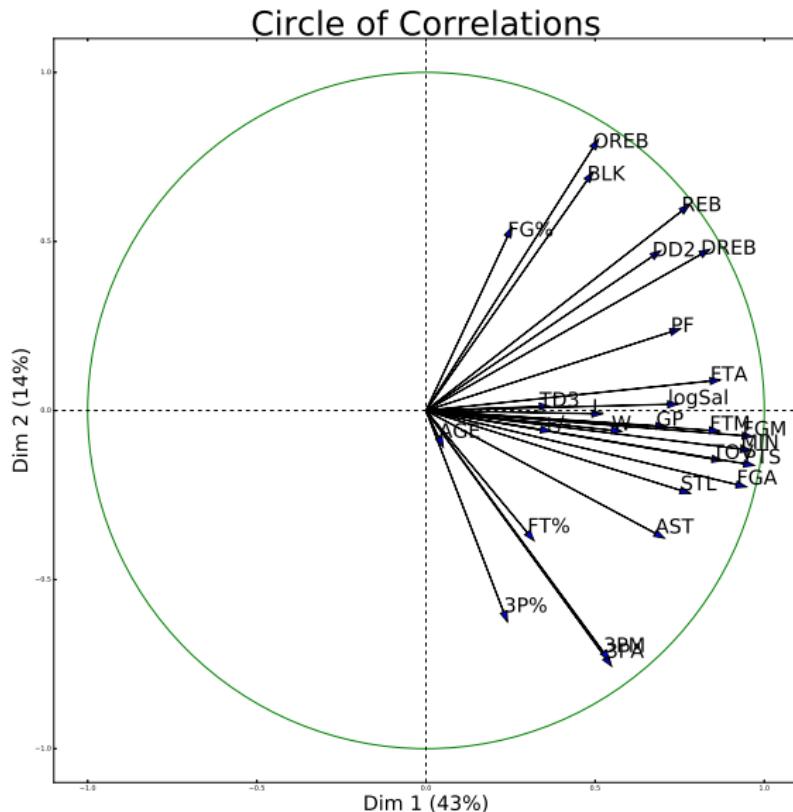
RANK	PLAYER	TEAM	GP	MIN	MGS
1	James Harden, SG	HOU	82	3225	38.1
2	Kyle Lowry, PG	TOR	77	2851	37.0
3	Jimmy Butler, SF	CHI	67	2474	36.9
4	Kentavious Caldwell-Pope, SG	DET	76	2709	36.7

Visualizing the data via dimensionality reduction

Complete Dataset - PCA

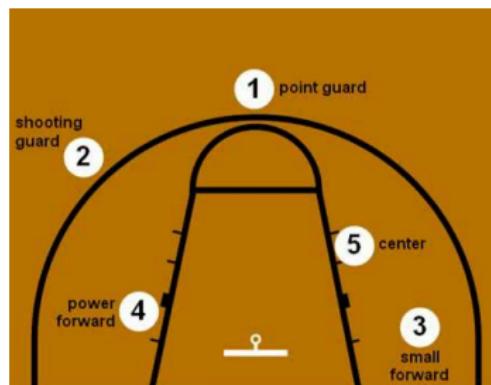


Complete Dataset - Correlation Graph



Complete Dataset

- 1st Dimension: Player ability
 - Salary (log. scale)
 - Points
- 2nd Dimension: Position on the Field
 - (offensive) rebounds,
 - Blocks,
 - 3 point field made/attempts



Feature selection and clustering analysis

Unsupervised methods for feature selection

- Laplacian score

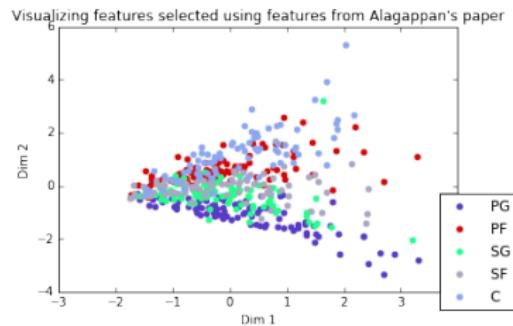
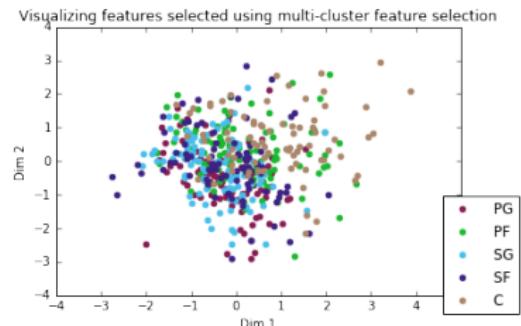
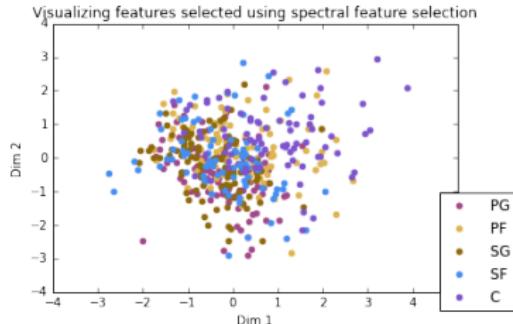
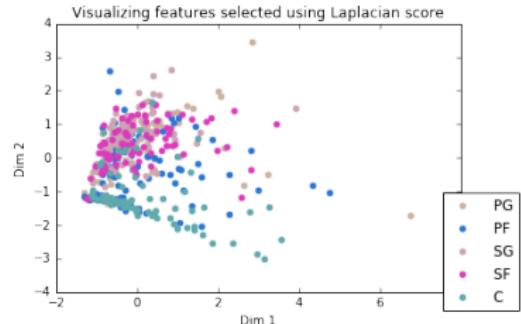
Unsupervised methods for feature selection

- Laplacian score
- Spectral Feature Selection

Unsupervised methods for feature selection

- Laplacian score
- Spectral Feature Selection
- Multi-Cluster Feature Selection

Visualizing the selected features



Evaluating a cluster

- Permutation-wise Accuracy

Evaluating a cluster

- Permutation-wise Accuracy
- Adjusted Rand's Index

Evaluating a cluster

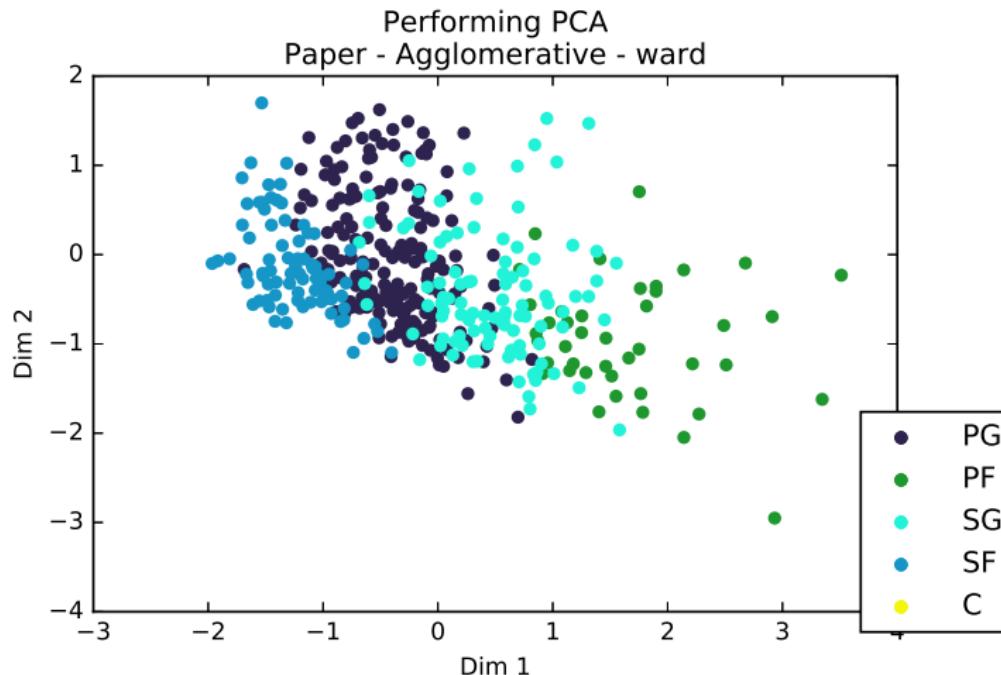
- Permutation-wise Accuracy
- Adjusted Rand's Index
- Silhouette score

Agglomerative Clustering

Table 1: Best performances for each evaluation metrics

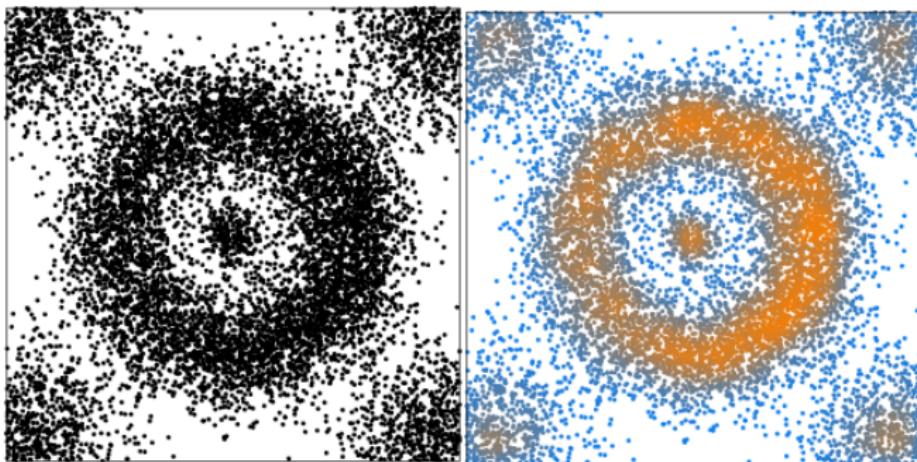
Metrics	Best score	Features	Linkage
Accuracy	37.8%	Alagappan's features	Ward
Rand's Index	0.093	Alagappan's features	Complete
Silhouette score	0.364	All the features	Average

Agglomerative Clustering



Mode-seeking

- Greater importance to peaks of density
- A Hill-Climbing scheme

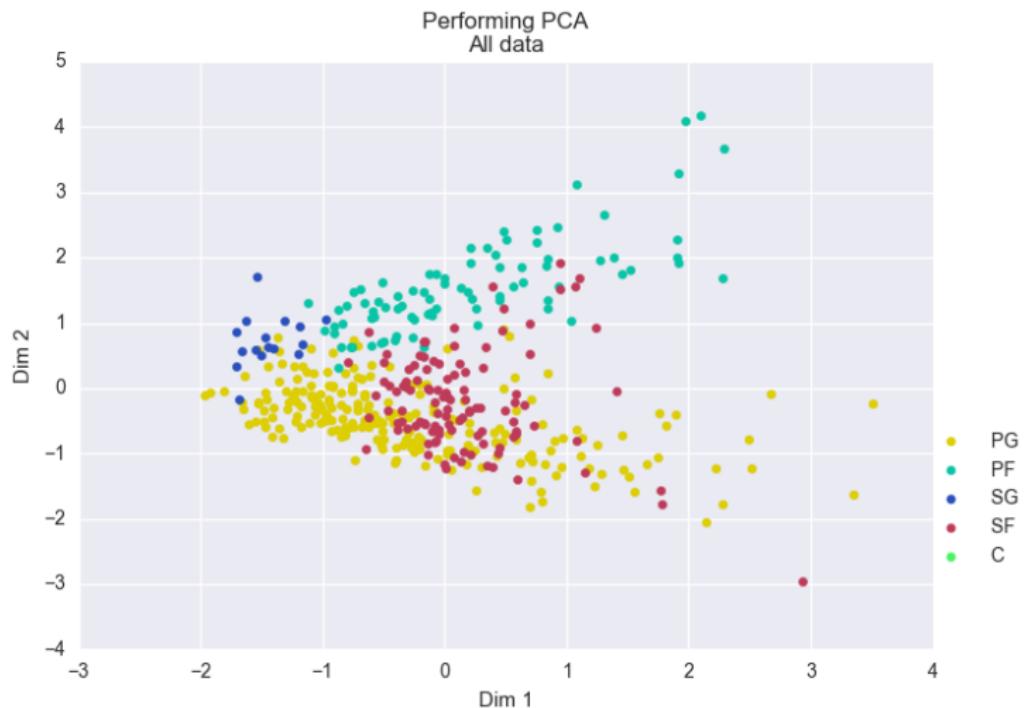


Mode-seeking

Table 2: Best performances for each evaluation metrics

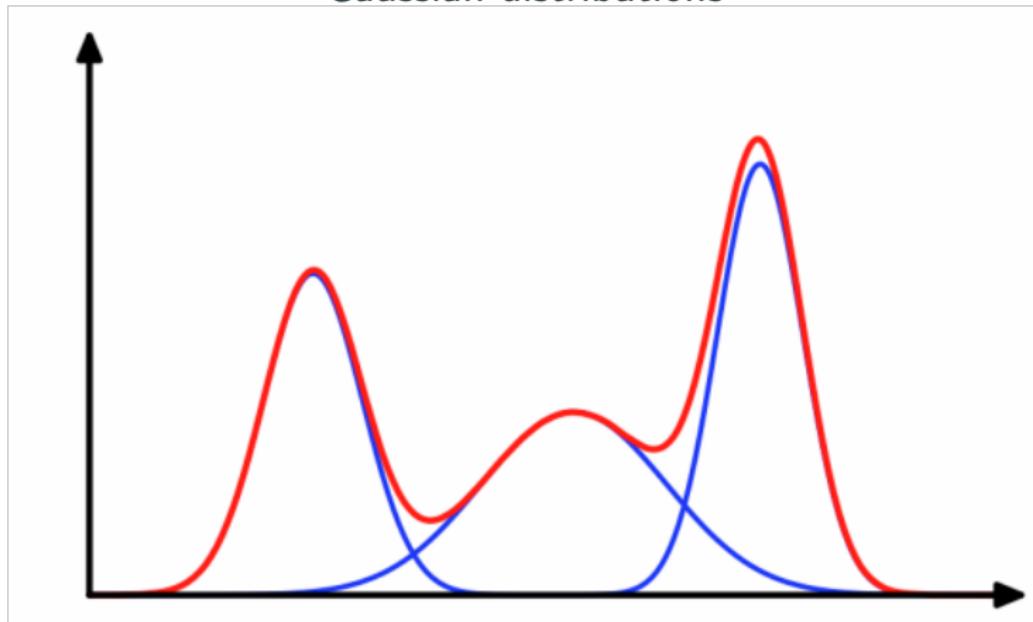
Metrics	Best score	Score range	Features	kDensity
Accuracy	34.9%	[0,100]	All the features	12
Rand's Index	0.137	[0,1]	All the features	16
Silhouette score	0.050	[-1,1]	Laplacian score	24

Mode-seeking



Gaussian Mixture Models

- Assumes data is sampled as a mixture of unknown Gaussian distributions



Gaussian Mixture Models

Table 3: Best performances for each evaluation metrics

Metrics	Best score	Features	Covariance type
Accuracy	39.67%	Alagappan's features	diagonal
Rand's Index	0.126	Alagappan's features	diagonal
Silhouette score	0.287	All the features	diagonal

Comparing with supervised algorithms

Table 4: Best performances on supervised classification

Classifier	Best accuracy	Features
Gaussian Process	55.37%	Alagappan's features
Decision Tree	51.24%	Alagappan's features
QDA	55.10%	All the features
Naive Bayes	43.80%	Alagappan's features
Linear SVM	48.98%	All the features
Neural Net	63.27%	All the features
RBF SVM	49.59%	Alagappan's features
Adaboost	54.55%	Alagappan's features
Random Forest	53.06%	All the features
Nearest Neighbors	52.07%	Alagappan's features

Analyzing other sets of field positions

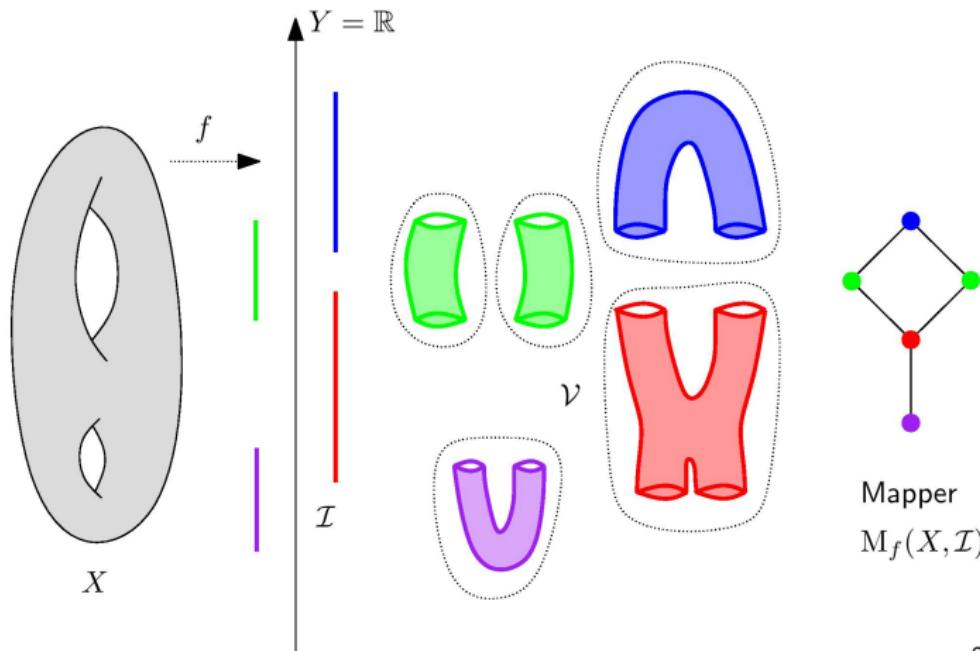
Table 5: Best performances for each method

Method	Accuracy	Selected features	More
Common unsupervised	48.95%	Laplacian score	complete linkage
Mode-seeking	48.7%	All the features	ratio 3:1
Gaussian Mixture Models	67.5%	All the features	tied covariance
Common Supervised	79.59%	All the features	Neural net

Mapper

Theoretical Background

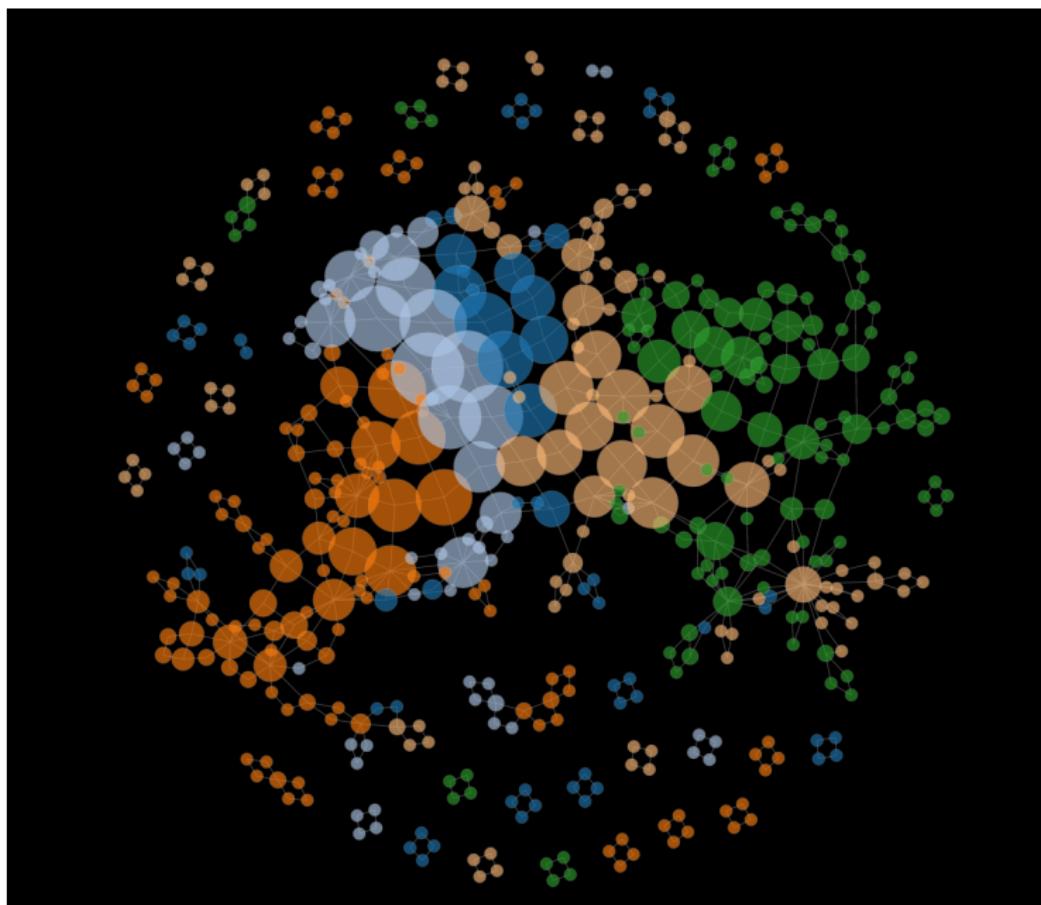
Mapper in the continuous setting



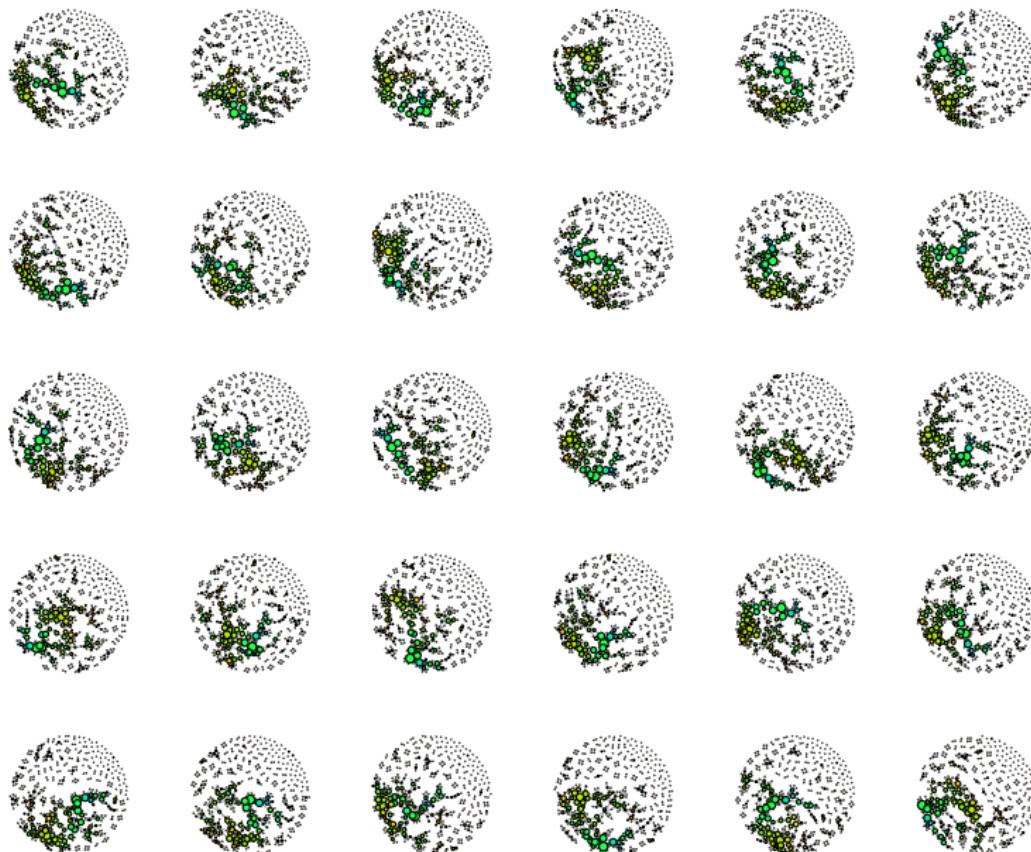
2

KeplerMapper

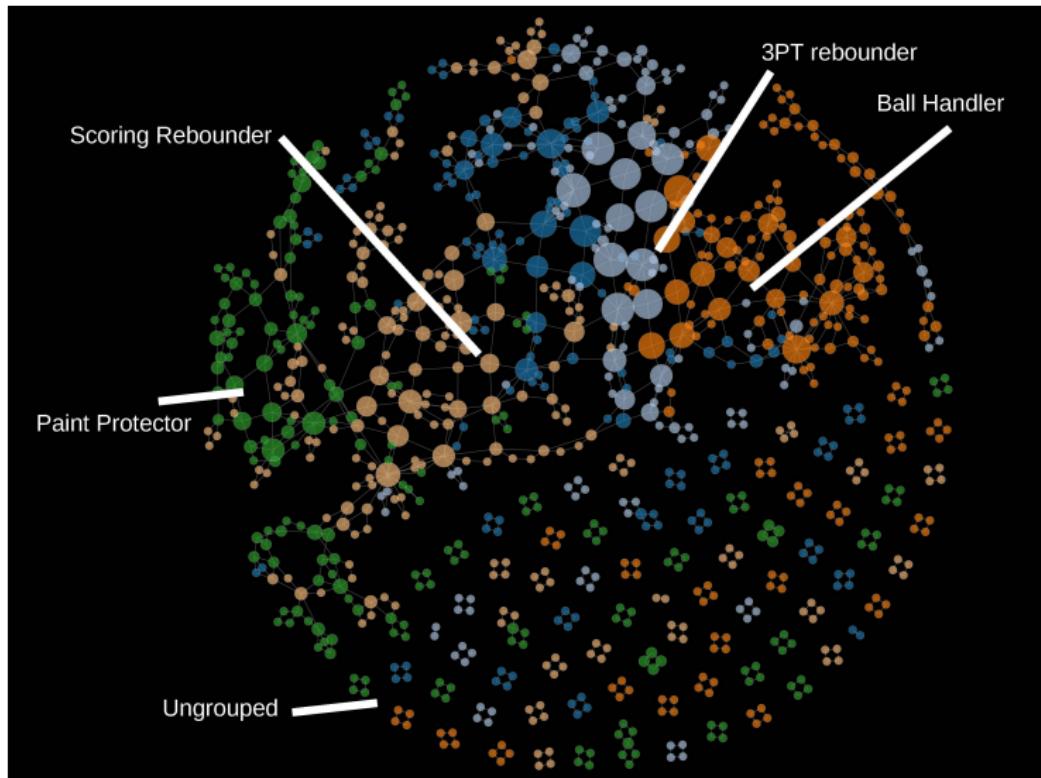




Tuning parameters



Method's result



References I

-  M. Alagappan, P. Y. Lum, G. S. A. L. T. I. M. V.-J. J. C. . G. C. (2012).
Extracting insights from the shape of the data using topology.
SCIENTIFIC REPORTS.

Thank you

Questions?