A General Framework for Increasing the Robustness of PCA-based Correlation Clustering Algorithms

Hans-Peter Kriegel, Peer Kröger, Erich Schubert, Arthur Zimek

Ludwig-Maximilians-Universität München
Munich, Germany
www.dbs.ifi.lmu.de

Presenter: Matthias Renz

Outline

- Motivation
- Increasing the Robustness of PCA for Correlation Clustering
- Evaluation
- Summary

Outline

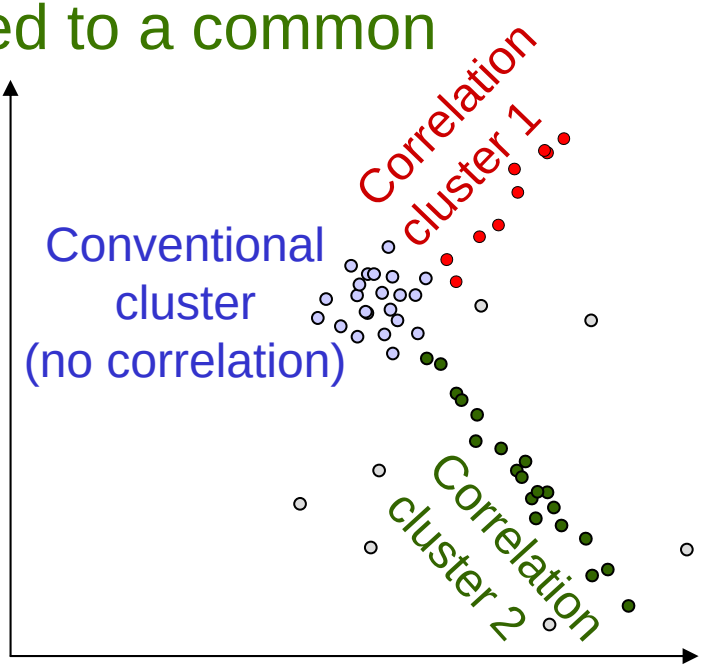
- Motivation
- Increasing the Robustness of PCA for Correlation Clustering
- Evaluation
- Summary

Motivation

- Correlation Clustering aims at finding groups of *d*-dimensional points that are associated to a common
  - lower-dimensional
  - arbitrarily oriented

  hyperplane

=> points exhibit a common correlation among a subset of attributes



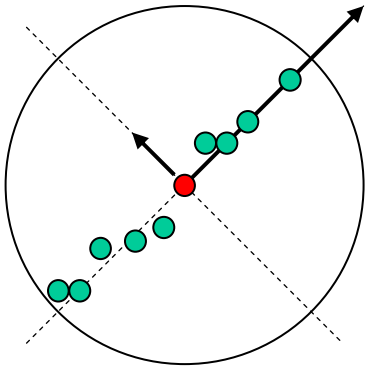
- Extends the problem of subspace/projected clustering from axis-parallel to arbitrarily oriented subspace clusters
- Prominent sample methods: ORCLUS, 4C, ERiC, …

Motivation

- Challenge of correlation clustering algorithms
  - To find the subspace, a set of cluster members need to be known

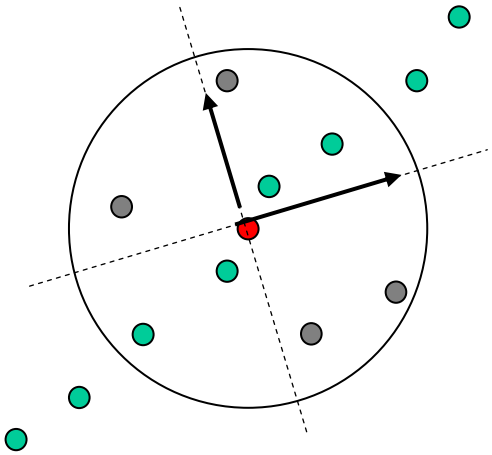
    => apply PCA on these cluster members
  - To assign cluster memberships and determine noise/outliers, the subspaces of the clusters need to be known

    => assignment based on the eigensystem of the cluster
- Most correlation clustering algorithm use the following ***locality assumption***
  - The local neighbors of cluster members or cluster representatives (e.g. centers) reflect the subspace of the cluster



Motivation

- Challenge of correlation clustering algorithms
  - To find the subspace, a set of cluster members need to be known
    => apply PCA on these cluster members
  - To assign cluster memberships and determine noise/outliers, the subspaces of the clusters need to be known
    => assignment based on the eigensystem of the clusters
- Most correlation clustering algorithm use the following ***locality assumption***
  - The local neighbors of cluster members or cluster representatives (e.g. centers) reflect the subspace of the cluster
  - But in high-dimensional spaces, it is likely that the local neighborhood contains noise points!!!!

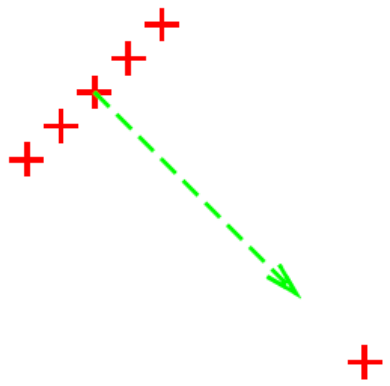


Outline

- Motivation
- Increasing the Robustness of PCA for Correlation Clustering
- Evaluation
- Summary

Problem Analysis

- Impacts of the locality assumption on PCA
  1. PCA is very sensitive to noise/outliers
     - Eigenvector with the largest eigenvalue of 6 points

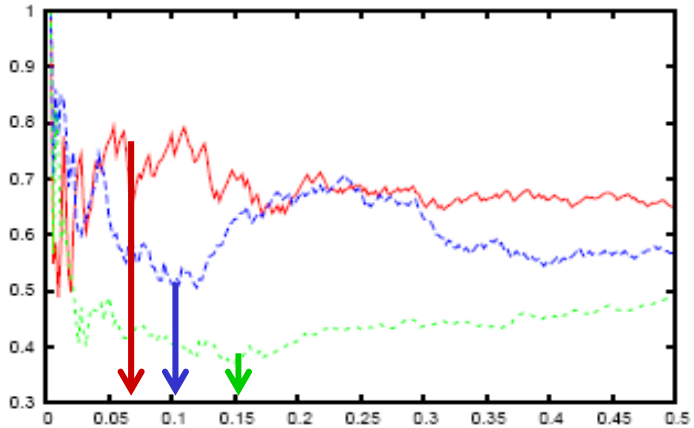


=> only one point flips the eigenvector into a wrong direction

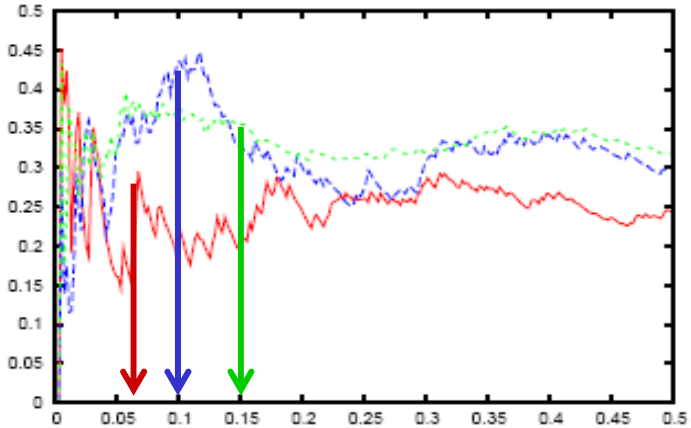
     - Consequence: learning the correct subspace from local neighbors is misled by noise
     - NOTE: noise/outliers cannot be eliminated because we need to know the correct subspaces of the cluster before

Problem Analysis

2. A different number of neighbors (of cluster members) are sufficient to represent the correct subspace for different clusters
   - Relative strength of the first two eigenvectors of sample points in a 3D toy data set w.r.t. *k* = number of neighbors considered for PCA



k/data set size



k/data set size

   - Red: point of a 1D cluster; *k* = 7% of data set size is the perfect choice (80% of the variance is explained by the first eigenvector)
   - Blue: point of a 2D cluster; *k* = 10% of data set size is the perfect choice (45% of the variance is explained by the first two eigenvectors)
   - Green: point of noise; k = 15% of the data set size is the perfect choice

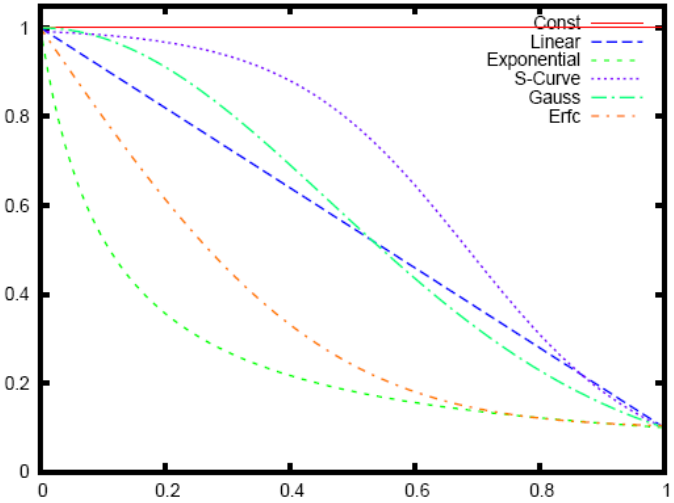
Aim of this work

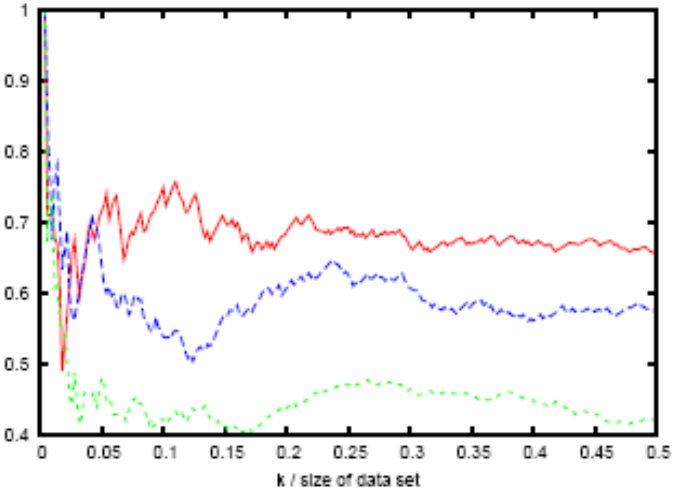
- Problem summary
  - PCA is very sensitive to noise/outliers
  - A different number of neighbors should be taken for different points/cluster representatives

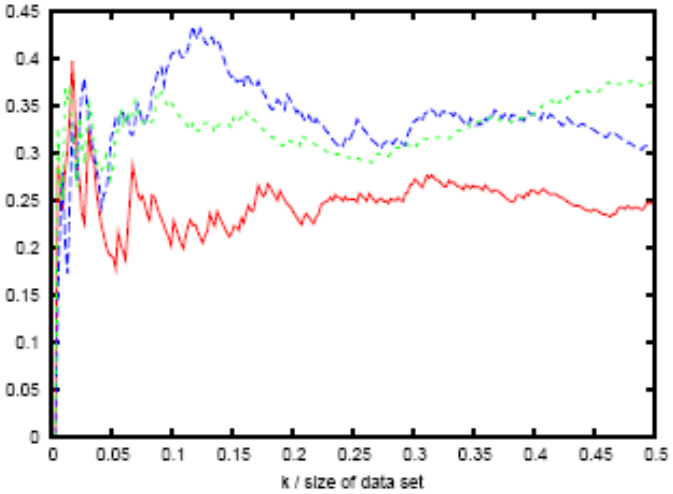
- In this work
  - We do not overcome the locality assumption
  - But we try to ease the effects of the locality assumption on the PCA and, thus, on the quality of the clustering results
  - Tackle the above described problems in a general way
  - Show how these concepts can be integrated into correlation clustering algorithms (partitioning-based ORCLUS and density-based ERiC)

Solution 1: Weighting

- Ease the impact of outliers
  - Give weights to all neighbors
    - Close neighbors are more likely to be cluster members than far neighbors
    - Use any distance-based weighting function
  - Compute weighted covariance matrix and apply PCA on that weighted matrix
  - Effect on the relative strength of the first two eigenvectors







Observations:

Sudden drops have disappeared

Generally, choosing k "high enough" works fine

Solution 2: Auto-tuning

- Choose an individual number of neighbors to be considered for determining the subspace by PCA
  - Use a globally fixed number of k_{max} neighbors and choose individually for each point/cluster representative the best k neighbors
  - Do ***NOT*** test all possible $O(2^{k_{\max}})$ subsets
  - Rather test for all $k \leq k_{max}$ the $O(k_{\max})$ subsets containing the k nearest neighbors
  - Evaluate the goodness of the particular selections of k by
    - Using a sliding window for dimensionality filtering
    - Averaging the explained variance of the largest eigenvalues

  (see details in the paper)

Outline

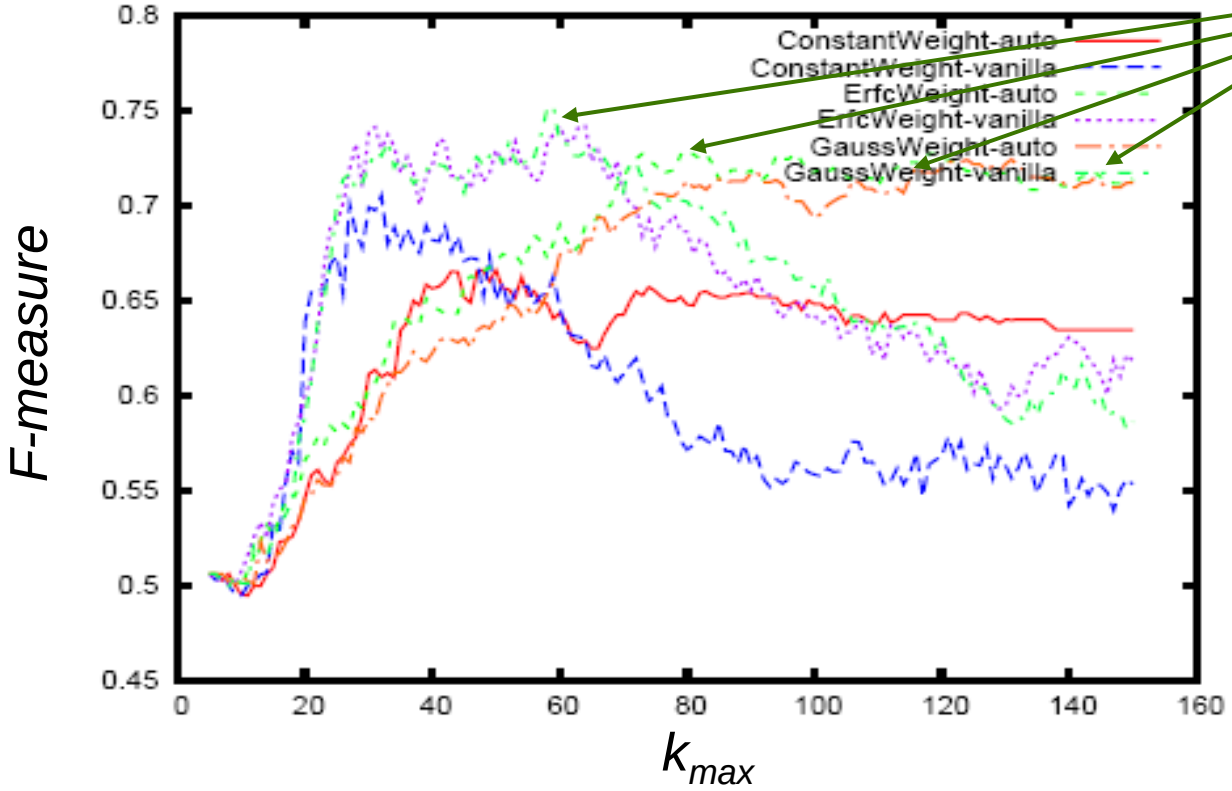
- Motivation
- Increasing the Robustness of PCA for Correlation Clustering
- Evaluation
- Summary

Evaluation

- Test Bed
  - Quality measurement
    - Pair-counting F-Measure
  - Clustering methods
    - ERiC (density-based)
    - ORCLUS (partitioning-based)
  - Competitors
    1. Original method
    2. Weighting function
    3. Auto-tuning of numbers of neighbors
    4. All combinations of 1. 2. and 3.

Evaluation

- 10D Synthetic data (ERiC)



Observation:

Erfc weighting + auto-tuning provides highest F-measure scores and is rather robust for k_{max} choosen "high enough"

- (Erfc-)weighting seems to be very important
- Auto-tuning still gives some benefit

Evaluation

- 10D Synthetic data (ORCLUS, 100 randomly initialized runs)

Variant	Average F-measure	Standard Deviation
ORCLUS	0.667	0.046
ORCLUS + Gauss weight	0.684	0.055
ORCLUS + Exponential weight	0.676	0.054
ORCLUS + Erfc weight	0.683	0.061
ORCLUS + Linear weight	0.686	0.056
ORCLUS + Auto	0.751	0.070
ORCLUS + Auto + Gauss	0.763	0.069
ORCLUS + Auto + Exponential	0.754	0.075
ORCLUS + Auto + Erfc	0.754	0.075
ORCLUS + Auto + Linear	0.771	0.078

  - Auto-tuning seems to be very important
  - Weighting still gives some benefit

Evaluation

- Results of ERiC using autotuning and Erfc weighting on NBA player stats data (15 dimensions)
  - Four meaningful clusters
  - Several players classified as noise

cluster ID	dim	Description
1	4	"go-to-guys"
2	4	guards
3	4	reserves
4	5	small forwards

- Results of ERiC using auto-tuning and Erfc weighting on Metabolic Screening data (containing metabolite concentrations – 43 dimensions)
  - Newborns are labeled with metabolic disorders
  - Five pure clusters of newborns suffering PKU and healthy newborns
  - Several newborns classified as noise

cluster ID	dim	Description
1	10	PKU
2	10	controll
3	11	PKU
4	12	PKU
5	13	PKU

Outline

- Motivation
- Increasing the Robustness of PCA for Correlation Clustering
- Evaluation
- Summary

Summary

- A general framework for increasing the robustness of PCA-based correlation clustering algorithms
  - Locality assumption is still there
  - But it's negative effects on PCA and, thus, final result is decreased
- Main ideas
  - ***Weighting***: Impact of neighbors are weighted such that close neighbors have higher impact than far neighbors
  - ***Auto-tuning***: Number of neighbors on which PCA is applied is locally optimized (i.e. for each point/cluster representative separately)
- Integration into any existing PCA-based correlation clustering algorithm (see paper for details)
- Experimental results show the benefit of these concepts; none of the two concepts is the clear winner

The last slide

Any questions

Just please mail me

kroegerp@dbs.ifi.lmu.de