Ludwig Maximilians Universität München

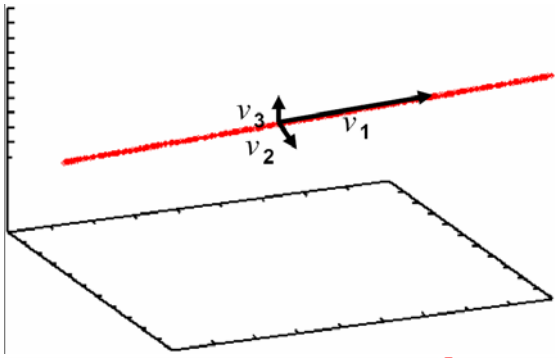
COPAC

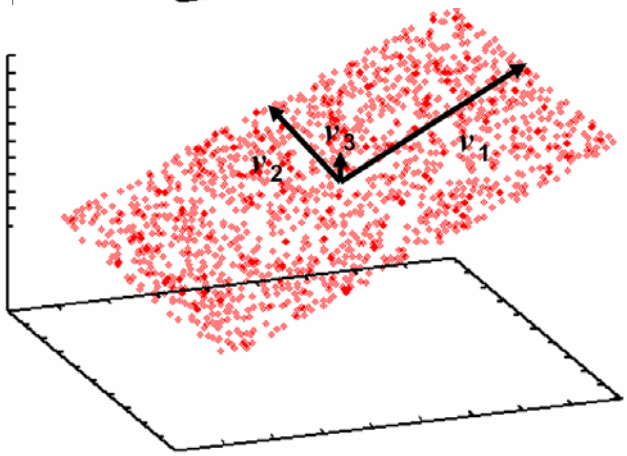
Institut für Informatik

Robust, Complete, and Efficient Correlation Clustering

Motivation and Definitions

Correlation Clusters





Formalization of Correlation Clusters

- Covariance matrix

$$\mathbf{\Sigma}_{\mathcal{C}} = \frac{1}{|\mathcal{C}|} \cdot \sum_{X \in \mathcal{C}} (X - \bar{X}) \cdot (X - \bar{X})^{\mathbf{T}}$$

- Correlation dimensionality

$$\lambda_{\mathcal{C}} = \min_{r \in \{1,\ldots,d\}} \left\{ r \;\middle|\; \frac{\sum_{i=1}^{r} e_i}{\sum_{i=1}^{d} e_i} \geq \alpha \right\}$$

Correlation Dimensionality of a Point *p*

- Correlation dimensionality

$$\lambda_P = \min_{r \in \{1,\ldots,d\}} \left\{ r \;\middle|\; \frac{\sum_{i=1}^{r} e_i}{\sum_{i=1}^{d} e_i} \geq \alpha \right\}$$

Correlation Distance between two Points *P* and *Q*

- Correlation distance matrix

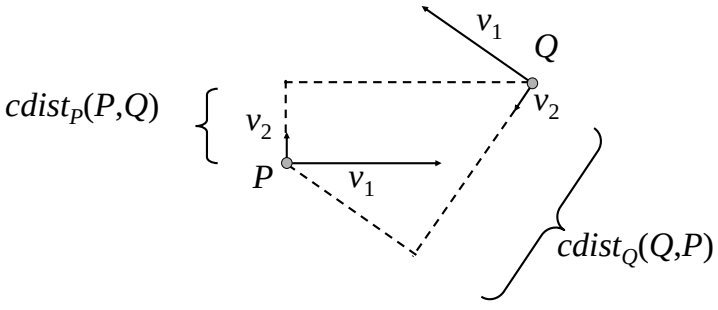
Let $P \in \mathcal{D}$, λ_P the local correlation dimensionality of P, and $\mathbf{V}_P$, $\mathbf{E}_P$ the corresponding Eigenvectors and Eigenvalues of the point P based on the local neighborhood of P, i.e. $\mathcal{N}_P$. An adapted Eigenvalue matrix $\hat{\mathbf{E}}_P$ with entries $\hat{e}_i \in \{0,1\}$, $(i = 1,\ldots,d)$ is derived according to the following rule:

$$\hat{e}_i = \begin{cases} 0 & \text{if } i \leq \lambda_P \\ 1 & \text{if } i > \lambda_P \end{cases}$$

The matrix $\hat{\mathbf{M}}_P = \mathbf{V}_P \hat{\mathbf{E}}_P \mathbf{V}_P^{\mathbf{T}}$ is called the correlation distance matrix *of P.*

- Correlation distance

$$cdist_P(P,Q) = \sqrt{(P-Q) \cdot \hat{\mathbf{M}}_P \cdot (P-Q)^{\mathbf{T}}}$$



E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, A. Zimek

Database Systems LMU

www.dbs.ifi.lmu.de

Ludwig Maximilians Universität München
LMU
Institut für Informatik

COPAC

Robust, Complete, and Efficient Correlation Clustering

Algorithm and Results

Algorithm HiSC

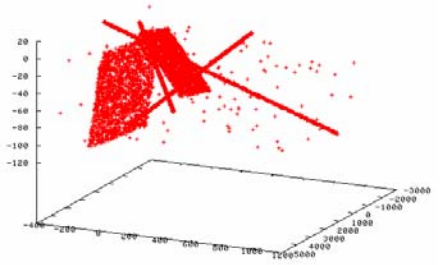
```
algorithm COPAC
  // STEP 1: Partitioning
  for each P in DB do
     compute λ(P);
     DB(λ(P)) = DB(λ(P)) ∪ {P};
  enddo

  // STEP 2: Clustering
  for each DB(λ) do
     compute λ-dimensional
          correlation clusters;
  enddo
```

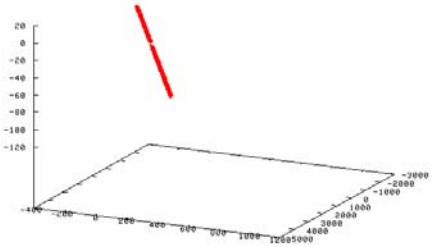
Result on Gene Expression Data

cID	sample gene names	description
1	NDUFB10, MTRF1, TIMM17A, CPS1, NM44, COX10, FIBP, TRAP1, MTERF, HK1, HADHA, ASAH2, CPS1, CA5A, BNI3PL, TOM34, ME2	proteins located in and/or related to mitochondrial membran
2	TNFRSF6, TNFRSF11A, TNFRSF7, TNFRSF1B, TNFRSF5, TNFRSF1A, TRAF5, TRAF2, TNFSF12	proteins related to tumor necrosis factor (TNF)
	IL1A, IL1B, IL2, IL6, IL10, IL18, IL24, IL1RN, IL2RG, IL4R, IL6R, IL7R, IL10RA, IL10RB, IL12A, IL12RB2, IL15RA, IL22R	interleukins or their receptors activating immune response

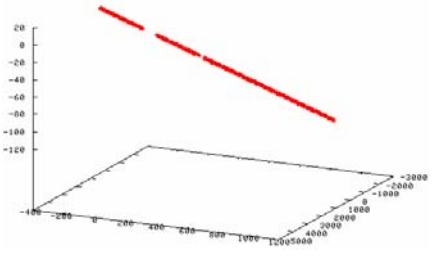
Result on Synthetic Dataset



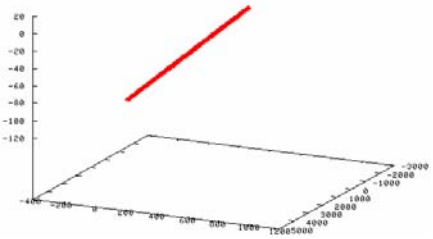
Data set



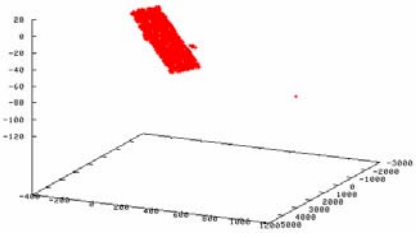
Partition 1 – Cluster 1



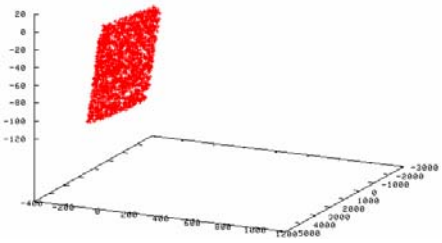
Partition 1 – Cluster 2



Partition 1 – Cluster 3



Partition 2 – Cluster 1



Partition 1 – Cluster 2

Database Systems LMU

E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, A. Zimek

www.dbs.ifi.lmu.de