Deriving Quantitative Models for Correlation Clusters

Elke Achtert   Christian Böhm   Hans-Peter Kriegel   Peer Kröger
Arthur Zimek

Institute for Computer Science
Ludwig-Maximilians-Universität München

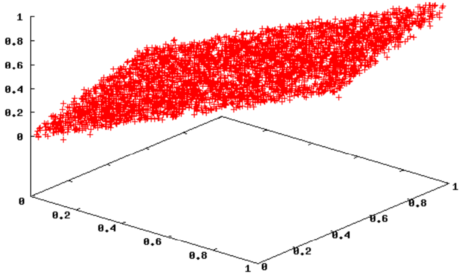
12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, 2006

Overview

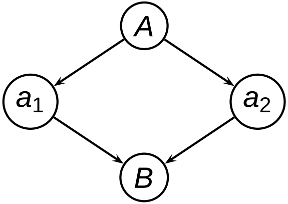
1. Correlation Clustering and Beyond
   - What are Correlation Clusters?
   - Why are Correlation Clusters Interesting?
   - What are Models for?
2. Deriving Quantitative Models for Correlation Clusters
   - Formal Description of Correlation Clusters
   - Quantitative Models for Correlation Clusters
   - Interpretation of Correlation Cluster Models
   - Quantitative Models as Predictive Models
3. Evaluation
   - Identifying Models
   - Predictive Models
4. Conclusions

Correlation Clusters

- hyperplanes exhibiting a high density of data points
- strong correlations between different features
- corresponding linear dependencies

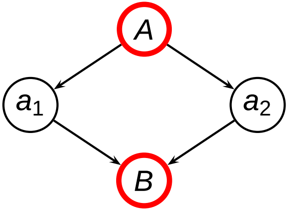


Example: Metabolic Pathways



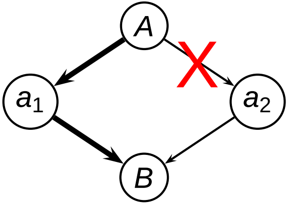
- There are certain pathways for degradation of metabolites.
- Concentrations of input and output metabolites may be correlated, the concentration of alternative intermediate states may vary depending on the environment.
- Genetic disorders may lead to failure of some pathways, other pathways are used more intensely.
- The concentrations of more metabolites are correlated if samples suffer from certain diseases.

Example: Metabolic Pathways



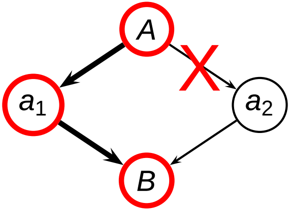
- There are certain pathways for degradation of metabolites.
- Concentrations of input and output metabolites may be correlated, the concentration of alternative intermediate states may vary depending on the environment.
- Genetic disorders may lead to failure of some pathways, other pathways are used more intensely.
- The concentrations of more metabolites are correlated if samples suffer from certain diseases.

Example: Metabolic Pathways



- There are certain pathways for degradation of metabolites.
- Concentrations of input and output metabolites may be correlated, the concentration of alternative intermediate states may vary depending on the environment.
- Genetic disorders may lead to failure of some pathways, other pathways are used more intensely.
- The concentrations of more metabolites are correlated if samples suffer from certain diseases.

Example: Metabolic Pathways



- There are certain pathways for degradation of metabolites.
- Concentrations of input and output metabolites may be correlated, the concentration of alternative intermediate states may vary depending on the environment.
- Genetic disorders may lead to failure of some pathways, other pathways are used more intensely.
- The concentrations of more metabolites are correlated if samples suffer from certain diseases.

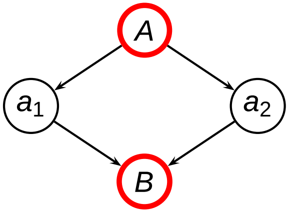
Correlation Clustering

- clustering: find sets of points building a hyperplane of arbitrary dimensionality

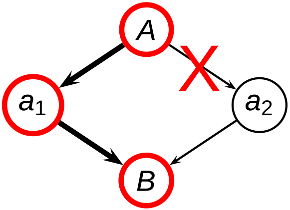
  e.g. metabolic pathways: Which patients are suffering from a disease?

- post-clustering: what are the underlying linear dependencies?

  e.g. metabolic pathways: What is the nature of the disease?



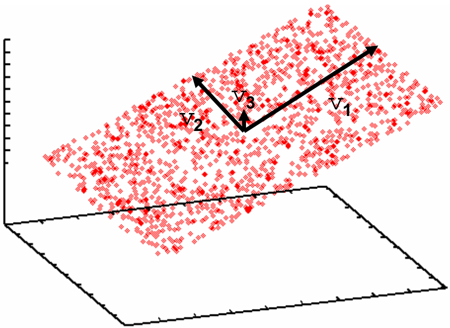
$$\left(\begin{array}{cccc} A & & -B & = & c_1 \end{array} \right)$$



$$\left(\begin{array}{ccccc} A & & -B & = & c_1 \\ & a_1 & -3 \cdot B & = & c_2 \end{array} \right)$$

Correlation Cluster Model

- A model can describe a cluster.
- A model can possibly be interpreted by domain experts, leading to new insights.
- A model can possibly be used to predict the class of a new point.



Formal Description of Correlation Clusters

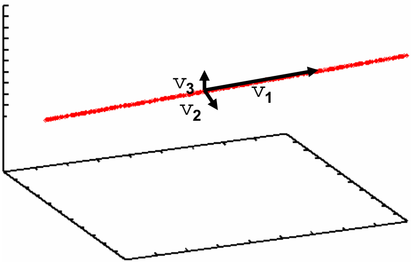
- decompose covariance matrix of correlation cluster $\mathcal{C}$ to eigenvalues and eigenvectors:

$$\Sigma_{\mathcal{C}} = V_{\mathcal{C}} \cdot E_{\mathcal{C}} \cdot V_{\mathcal{C}}^{\mathsf{T}}$$

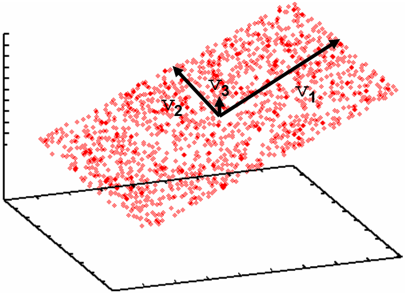
- most of the variance covered by small number of eigenvectors
- number of eigenvectors covering most of the variance is called correlation dimensionality $\lambda_{\mathcal{C}}$ of cluster $\mathcal{C}$
- eigenvectors $\#1, \ldots, \#\lambda_{\mathcal{C}}$: strong eigenvectors $\check{V}_{\mathcal{C}}$
- eigenvectors $\#\lambda_{\mathcal{C}} + 1, \ldots, \#d$: weak eigenvectors $\hat{V}_{\mathcal{C}}$

Strong and Weak Eigenvectors

- Strong eigenvectors span the hyperplane corresponding to a correlation cluster.
- Weak eigenvectors are orthogonal to the hyperplane.

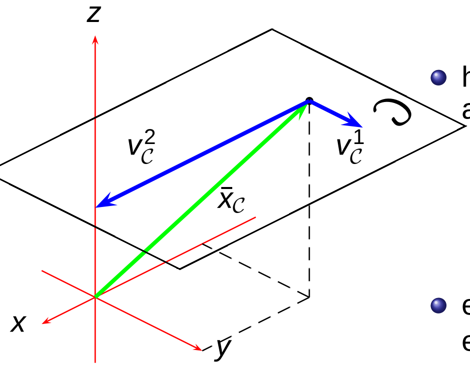


(a) $\lambda = 1$



(b) $\lambda = 2$

Eigenvectors Define a Correlation Cluster



- hyperplane of a correlation cluster $\mathcal{C}$ as affine subspace:
  - strong eigenvectors $\check{V}_{\mathcal{C}}$ as basis
  - an affinity w.r.t. original data space (e.g. centroid $\bar{x}_{\mathcal{C}}$ of the cluster $\mathcal{C}$)

$$x = \bar{x}_{\mathcal{C}} + a_1 v_{\mathcal{C}}^1 + \ldots + a_\lambda v_{\mathcal{C}}^\lambda, a_i \in \mathbb{R}$$

- equivalent: definition by weak eigenvectors $\hat{V}_{\mathcal{C}}$ and affinity $\bar{x}_{\mathcal{C}}$

Quantitative Models

- equation system to describe the hyperplane of correlation cluster $\mathcal{C}$ based on weak eigenvectors $\hat{V}_\mathcal{C}$ and affinity $\bar{x}_\mathcal{C}$:

$$\hat{V}_\mathcal{C}^T \cdot x = \hat{V}_\mathcal{C}^T \cdot \bar{x}_\mathcal{C}.$$

- equivalently:

$$\begin{array}{ccccccccc}
v_{(\lambda+1),1}(x_1-\bar{x}_1) & + & v_{(\lambda+1),2}(x_2-\bar{x}_2) & + & \cdots & + & v_{(\lambda+1),d}(x_d-\bar{x}_d) & = & 0 \\
v_{(\lambda+2),1}(x_1-\bar{x}_1) & + & v_{(\lambda+2),2}(x_2-\bar{x}_2) & + & \cdots & + & v_{(\lambda+2),d}(x_d-\bar{x}_d) & = & 0 \\
 & & & & \vdots & & & & \\
v_{d,1}(x_1-\bar{x}_1) & + & v_{d,2}(x_2-\bar{x}_2) & + & \cdots & + & v_{d,d}(x_d-\bar{x}_d) & = & 0
\end{array}$$

- defect of $\hat{V}_\mathcal{C}^T$: number of free attributes

Algorithm

1. run an arbitrary correlation clustering algorithm (e.g. 4C, ORCLUS) on data set $\mathcal{D} \subset \mathbb{R}^d$
2. for each correlation cluster $\mathcal{C}_i \subset \mathcal{D}$ found in the first step:
   1. derive covariance matrix $\Sigma_{\mathcal{C}_i}$
   2. select weak eigenvectors $\hat{V}_{\mathcal{C}_i}$ of $\Sigma_{\mathcal{C}_i}$
   3. derive equation system describing the correlation hyperplane:

      $$\hat{V}_{\mathcal{C}_i}^T \cdot x = \hat{V}_{\mathcal{C}_i}^T \cdot \bar{x}_{\mathcal{C}_i}$$

   4. apply Gauss-Jordan elimination to the derived equation system to obtain a unique description of quantitative dependencies by means of the reduced row echelon form

What a Correlation Cluster Model Tells Us

Example Model:

$$\left(\begin{array}{l} 1x_1 + 0x_2 + c_1x_3 + 0x_4 + e_1x_5 = f_1 \\ 0x_1 + 1x_2 + c_2x_3 + 0x_4 + e_2x_5 = f_2 \\ 0x_1 + 0x_2 + 0x_3 + 1x_4 + e_3x_5 = f_3 \end{array}\right)$$

- correlation dimensionality: 2 (number of free attributes, number of strong eigenvectors)
- linear dependencies by given factors c_1, e_1, c_2, e_2, and e_3 among:
  $\{x_1, x_3, \text{and } x_5\}$ $\{x_2, x_3, \text{and } x_5\}$ $\{x_4 \text{ and } x_5\}$

Result:

quantitatively and uniquely defined relations between certain attributes

What a Correlation Cluster Model Does Not Tell Us...

What a Correlation Cluster Model Does Not Tell Us...

Trivially, correlations do not allow to directly conclude causalities.

... and how we can make use of it anyway

BUT: domain experts could make use of the model to refine experiments.

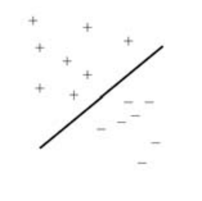
Quantitative Models as Predictive Models

- Refine descriptive model by determining the average distance of cluster members from correlation hyperplane.
- The standard deviation σ of the distances of all cluster members defines a Gaussian model of deviations from the common correlation hyperplane.
- for each model: probability for a new data point to be generated by this specific Gaussian distribution
- set of models: convenient instrument for classification in the perspective of different linear dependencies

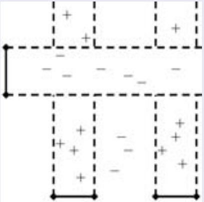
$$P(\mathcal{C}_j|x) = \frac{\frac{1}{\sigma_j\sqrt{2\pi}} e^{-\frac{1}{2\sigma_j^2}\left(d\left(x,\mathcal{C}_j\right)\right)^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2\sigma_i^2}(d(x,\mathcal{C}_i))^2}}$$

A New Type of Classifier

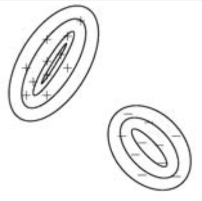
Decision models of different types of classifiers



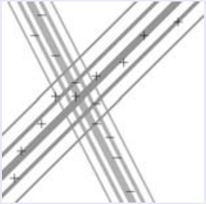
Linear decision boundaries



Axis parallel decision rules



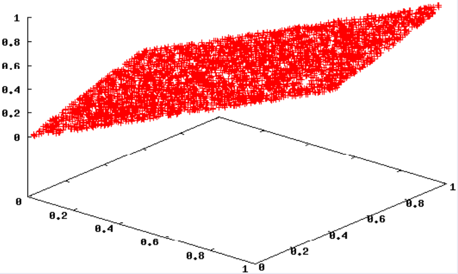
Density functions



Deviations from hyperplanes

Synthetic Data

Dataset with increasing standard deviation
Created with dependency: $x_1 - 0.5x_2 - 0.5x_3 = 0$



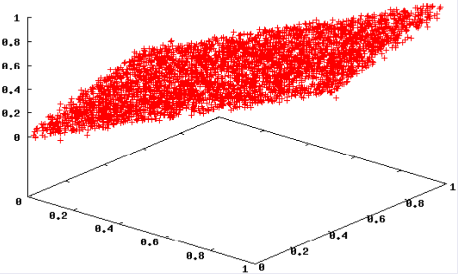
$$\sigma = 0 \text{ (no jitter)}$$

$$x_1 - 0.5000x_2 - 0.5000x_3 = 0.0000$$

Synthetic Data

Dataset with increasing standard deviation

Created with dependency: $x_1 - 0.5x_2 - 0.5x_3 = 0$

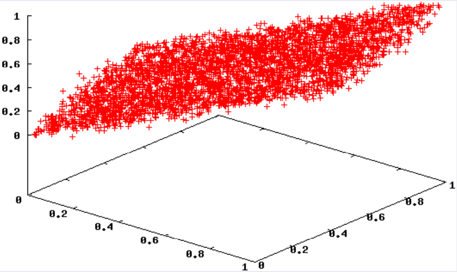


$\sigma = 0.0173$ (jitter of 1% of maximum distance in unit cube)

$$x_1 - 0.4989x_2 - 0.5002x_3 = 0.0000$$

Synthetic Data

Dataset with increasing standard deviation
Created with dependency: $x_1 - 0.5x_2 - 0.5x_3 = 0$

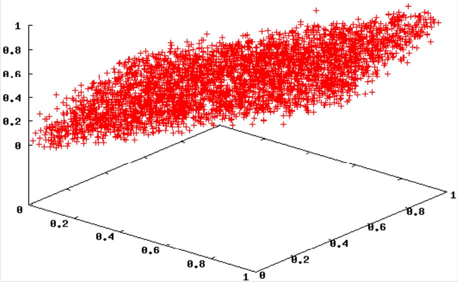


$\sigma = 0.0346$ (jitter of 2% of maximum distance in unit cube)

$$x_1 - 0.5017x_2 - 0.4951x_3 = 0.0016$$

Synthetic Data

Dataset with increasing standard deviation
Created with dependency: $x_1 - 0.5x_2 - 0.5x_3 = 0$

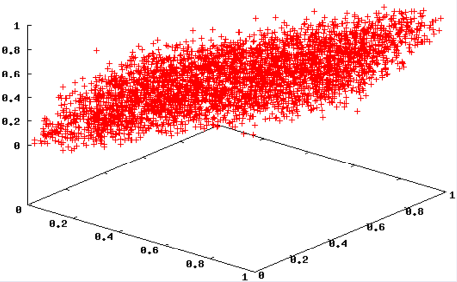


$\sigma = 0.0520$ (jitter of 3% of maximum distance in unit cube)

$$x_1 - 0.5030x_2 - 0.5047x_3 = -0.0059$$

Synthetic Data

Dataset with increasing standard deviation
Created with dependency: $x_1 - 0.5x_2 - 0.5x_3 = 0$

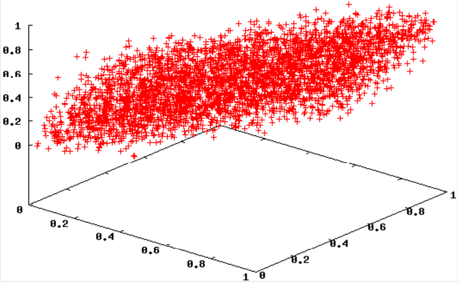


$\sigma = 0.0693$ (jitter of 4% of maximum distance in unit cube)

$$x_1 - 0.4962x_2 - 0.5106x_3 = -0.0040$$

Synthetic Data

Dataset with increasing standard deviation
Created with dependency: $x_1 - 0.5x_2 - 0.5x_3 = 0$



$\sigma = 0.0866$ (jitter of 5% of maximum distance in unit cube)

$$x_1 - 0.4980x_2 - 0.4956x_3 = 0.0064$$

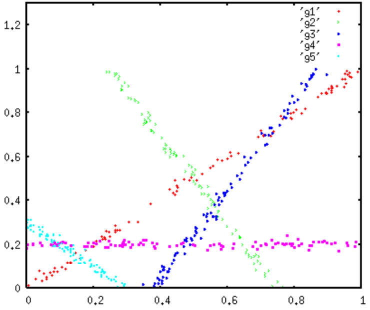
Models for Real World Data

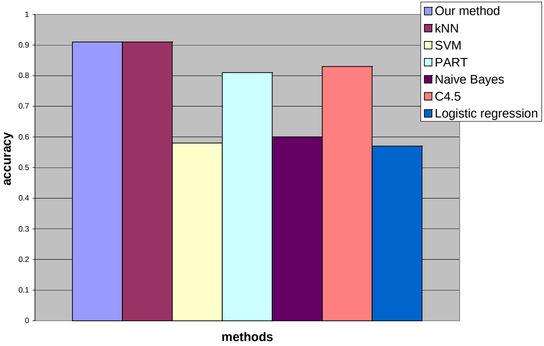
Wages data

Dependencies among age (A), years of education (YE), years of work experience (YW), and wage (W) found for three clusters in the 1985 Current Population Survey:

1. $YE = 12$
   $YW - 1 \cdot A = -18$
   $W - 0.07 \cdot A = 5.14$
2. $YE = 16$
   $YW - 1 \cdot A = -22$
3. $YE + 1 \cdot YW - 1 \cdot A = -6$

Predictive Models





Conclusions

- So far, correlation cluster analysis yields sets of points connected by common correlations.
- Based on such approaches, we propose to derive a model to grasp the underlying possible causalities.
- Our model is interpretable for domain experts and may help them to refine their experimental setting (descriptive model).
- Our model may also be used to classify new data points (predictive model).