Multivariate Verfahren 2

cluster analysis

Helmut Waldl

June 11th and 12th 2012

cluster analysis

preface

Methods for grouping, objects in the same group should be "as similar as possible", objects in different groups "as unlike as possible".

Data: the set of objects $I = \{1, \ldots, N\}$, p variates $\mathbf{x} = (x_1 \ldots x_p)^T$ are observed for each object. The data are summarized in the data matrix

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix}$$

There are grouping methods that use the values of the variates directly for grouping.
However at first we will study methods that primarily compute similarities and distances for the $\binom{N}{2}$ pairs of objects and then use these similarities and dissimilarities for grouping.

cluster analysis

preface

Methods for grouping, objects in the same group should be "as similar as possible", objects in different groups "as unlike as possible".

Data: the set of objects $I = \{1, \ldots, N\}$, p variates $\mathbf{x} = (x_1 \ldots x_p)^T$ are observed for each object. The data are summarized in the data matrix

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix}$$

There are grouping methods that use the values of the variates directly for grouping.
However at first we will study methods that primarily compute similarities and distances for the $\binom{N}{2}$ pairs of objects and then use these similarities and dissimilarities for grouping.

cluster analysis

preface

Methods for grouping, objects in the same group should be "as similar as possible", objects in different groups "as unlike as possible".

Data: the set of objects $I = \{1, \ldots, N\}$, p variates $\mathbf{x} = (x_1 \ldots x_p)^T$ are observed for each object. The data are summarized in the data matrix

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix}$$

There are grouping methods that use the values of the variates directly for grouping.
However at first we will study methods that primarily compute similarities and distances for the $\binom{N}{2}$ pairs of objects and then use these similarities and dissimilarities for grouping.

cluster analysis

similarity and distance measures

Definition: **similarity measure** $s : I \times I \longrightarrow \mathbb{R}$
and the following properties:

$$\left.\begin{array}{l} s_{nm} = s_{mn} \\ s_{nm} \leq s_{nn} \end{array}\right\} n, m = 1, \ldots, N$$

sometimes additional $s_{nm} \in [0; 1]$
the bigger s_{nm} the more similar are the objects

Definition: **distance measure** $d : I \times I \longrightarrow \mathbb{R}$
and the following properties:

$$\forall n, m \in I \qquad d_{nn} = 0 \qquad (d_{nm} = 0 \Leftrightarrow n = m) \qquad \text{definiteness}$$

$$\forall n, m \in I \qquad d_{nm} = d_{mn} \qquad \text{symmetry}$$

cluster analysis

similarity and distance measures

Definition: **similarity measure** $s: I \times I \longrightarrow \mathbb{R}$
and the following properties:

$$\left.\begin{array}{l} s_{nm} = s_{mn} \\ s_{nm} \leq s_{nn} \end{array}\right\} n, m = 1, \ldots, N$$

sometimes additional $s_{nm} \in [0; 1]$
the bigger s_{nm} the more similar are the objects

Definition: **distance measure** $d: I \times I \longrightarrow \mathbb{R}$
and the following properties:

$$\forall n, m \in I \qquad d_{nn} = 0 \qquad (d_{nm} = 0 \Leftrightarrow n = m) \qquad \text{definiteness}$$

$$\forall n, m \in I \qquad d_{nm} = d_{mn} \qquad \text{symmetry}$$

cluster analysis

similarity and distance measures

Definition: **metric** is a distance measure with the following additional property:

$$\forall l, m, n \in I: \; d_{nm} \leq d_{nl} + d_{lm} \qquad \text{triangle inequality}$$

counterexample for distance measure without triangle inequality:

$$\mathbb{R}^2; \quad d = (1-\varepsilon) \cdot \min\{d_x, d_y\} + \varepsilon \cdot \max\{d_x, d_y\}; \quad \varepsilon > 0$$

choose $\varepsilon = 0.1$; $A = (0/0)$; $B = (1/1)$; $C = (1/0)$, then $\overline{AB} = 1$; $\overline{AC} = 0.1$; $\overline{CB} = 0.1$ and $\overline{AC} + \overline{CB} < \overline{AB}$

In practice we often use metrics because they correspond to our spacial perception.

cluster analysis

similarity and distance measures

Definition: **metric** is a distance measure with
the following additional property:

$$\forall l, m, n \in I: \; d_{nm} \leq d_{nl} + d_{lm} \qquad \text{triangle inequality}$$

counterexample for distance measure without triangle inequality:

$$\mathbb{R}^2; \quad d = (1-\varepsilon) \cdot \min\{d_x, d_y\} + \varepsilon \cdot \max\{d_x, d_y\}; \quad \varepsilon > 0$$

choose $\varepsilon = 0.1$; $A = (0/0)$; $B = (1/1)$; $C = (1/0)$, then
$\overline{AB} = 1$; $\overline{AC} = 0.1$; $\overline{CB} = 0.1$ and $\overline{AC} + \overline{CB} < \overline{AB}$

In practice we often use metrics because they correspond to our spacial perception.

cluster analysis

similarity and distance measures

Definition: **metric** is a distance measure with
the following additional property:

$$\forall l, m, n \in I: \; d_{nm} \leq d_{nl} + d_{lm} \qquad \text{triangle inequality}$$

counterexample for distance measure without triangle inequality:

$$\mathbb{R}^2; \quad d = (1-\varepsilon) \cdot \min\{d_x, d_y\} + \varepsilon \cdot \max\{d_x, d_y\}; \quad \varepsilon > 0$$

choose $\varepsilon = 0.1$; $A = (0/0)$; $B = (1/1)$; $C = (1/0)$, then
$\overline{AB} = 1$; $\overline{AC} = 0.1$; $\overline{CB} = 0.1$ and $\overline{AC} + \overline{CB} < \overline{AB}$

In practice we often use metrics because they correspond to our spacial perception.

cluster analysis

similarity and distance measures

Similarity and distance measures for subsets of I (classes, groups) are defined analogously: Definition: Let $C_i \neq \{\} \subseteq I$
($C_i \in \mathcal{P}_*(I) \ldots$ power set of I without $\{\}$)
$S(C_i, C_j)$, $D(C_i, C_j)$ are similarity and distance measures for classes if they have the following properties:

$$S,\ D:\ \mathcal{P}_* \times \mathcal{P}_* \longrightarrow \mathbb{R}$$

$$S(C_i, C_j) = S(C_j, C_i) \leq S(C_i, C_i)$$

$$D(C_i, C_j) = D(C_j, C_i) \geq 0$$

mostly also the following monotony condition is satisfied:

$$C_i \subseteq C_j \implies S(C_i, C_k) \geq S(C_j, C_k) \qquad \forall i, j, k$$

$$C_i \subseteq C_j \implies D(C_i, C_k) \leq D(C_j, C_k) \qquad \forall i, j, k$$

We might however use measures without this property (centroid method).

cluster analysis

similarity and distance measures

Similarity and distance measures for subsets of I (classes, groups) are defined analogously: Definition: Let $C_i \neq \{\} \subseteq I$
($C_i \in \mathcal{P}_*(I) \ldots$ power set of I without $\{\}$)
$S(C_i, C_j)$, $D(C_i, C_j)$ are similarity and distance measures for classes if they have the following properties:

$$S,\ D:\ \mathcal{P}_* \times \mathcal{P}_* \longrightarrow \mathbb{R}$$

$$S(C_i, C_j) = S(C_j, C_i) \leq S(C_i, C_i)$$

$$D(C_i, C_j) = D(C_j, C_i) \geq 0$$

mostly also the following monotony condition is satisfied:

$$C_i \subseteq C_j \implies S(C_i, C_k) \geq S(C_j, C_k) \qquad \forall i, j, k$$

$$C_i \subseteq C_j \implies D(C_i, C_k) \leq D(C_j, C_k) \qquad \forall i, j, k$$

We might however use measures without this property (centroid method).

cluster analysis

similarity and distance measures

Similarity and distance measures for subsets of I (classes, groups) are defined analogously: Definition: Let $C_i \neq \{\} \subseteq I$
($C_i \in \mathcal{P}_*(I) \ldots$ power set of I without $\{\}$)
$S(C_i, C_j)$, $D(C_i, C_j)$ are similarity and distance measures for classes if they have the following properties:

$$S,\ D:\ \mathcal{P}_* \times \mathcal{P}_* \longrightarrow \mathbb{R}$$

$$S(C_i, C_j) = S(C_j, C_i) \leq S(C_i, C_i)$$

$$D(C_i, C_j) = D(C_j, C_i) \geq 0$$

mostly also the following monotony condition is satisfied:

$$C_i \subseteq C_j \implies S(C_i, C_k) \geq S(C_j, C_k) \qquad \forall i, j, k$$

$$C_i \subseteq C_j \implies D(C_i, C_k) \leq D(C_j, C_k) \qquad \forall i, j, k$$

We might however use measures without this property (centroid method).

cluster analysis

examples for similarity and distance measures, binary data

binary data: (with values 0 and 1)

For p binary variates we get the contingency table of I_n and I_m:

		I_m		
		1	0	
I_n	1	a	c	$a+c$
	0	b	e	$b+e$
		$a+b$	$c+e$	p

How many times do the values of p variates match? $(a+e)$

How many times do they not match? $(b+c)$

We get the **matching coefficient**: $\boxed{s_{nm} = \frac{a+e}{p}}$

cluster analysis

examples for similarity and distance measures, binary data

binary data: (with values 0 and 1)

For p binary variates we get the contingency table of I_n and I_m:

		I_m		
		1	0	
I_n	1	a	c	$a+c$
	0	b	e	$b+e$
		$a+b$	$c+e$	p

How many times do the values of p variates match? $(a+e)$

How many times do they not match? $(b+c)$

We get the **matching coefficient**: $\boxed{s_{nm} = \frac{a+e}{p}}$

cluster analysis

examples for similarity and distance measures, binary data

If we want to weight matching and mismatching distinctly we have to choose $u \in]0;1[$ (weight for matching). Then we get the similarity measure

$$s_{nm} = \frac{u \cdot (a+e)}{u \cdot (a+e) + (1-u) \cdot (b+c)}$$

For each u we have $s_{nm} \in [0;1]$ and with each u we get the same similarity ranking of pairs of objects.

Thus: If we use a clustering method that uses just the similarity ranking for grouping (e.g. single linkage, complete linkage), then the choice of the weights for the above s_{nm} does not matter.

Another important property of the above coefficient s_{nm} is its invariancy with respect to bijective mappings of one or several variates x.

cluster analysis

examples for similarity and distance measures, binary data

If we want to weight matching and mismatching distinctly we have to choose $u \in]0;1[$ (weight for matching). Then we get the similarity measure

$$s_{nm} = \frac{u \cdot (a+e)}{u \cdot (a+e) + (1-u) \cdot (b+c)}$$

For each u we have $s_{nm} \in [0;1]$ and with each u we get the same similarity ranking of pairs of objects.

Thus: If we use a clustering method that uses just the similarity ranking for grouping (e.g. single linkage, complete linkage), then the choice of the weights for the above s_{nm} does not matter.

Another important property of the above coefficient s_{nm} is its invariancy with respect to bijective mappings of one or several variates x.

cluster analysis

examples for similarity and distance measures, binary data

If we want to weight matching and mismatching distinctly we have to choose $u \in]0;1[$ (weight for matching). Then we get the similarity measure

$$\boxed{s_{nm} = \frac{u \cdot (a+e)}{u \cdot (a+e) + (1-u) \cdot (b+c)}}$$

For each u we have $s_{nm} \in [0;1]$ and with each u we get the same similarity ranking of pairs of objects.

Thus: If we use a clustering method that uses just the similarity ranking for grouping (e.g. single linkage, complete linkage), then the choice of the weights for the above s_{nm} does not matter.

Another important property of the above coefficient s_{nm} is its invariancy with respect to bijective mappings of one or several variates x.

cluster analysis

examples for similarity and distance measures, binary data

similarity coefficient: negative matchings (both objects have value 0 for some variate $\hat{=}$ absence of a property) do not count.

$$s_{nm} = \frac{a}{a+b+c}$$

weighted:

$$s_{nm} = \frac{u \cdot a}{u \cdot a + (1-u) \cdot (b+c)}$$

With each u we get the same similarity ranking of pairs of objects (which does not coincide with the similarity ranking of the matching coefficient).

The similarity coefficient is **not** invariant with respect to bijective transformations of one or several variates x.

cluster analysis

examples for similarity and distance measures, binary data

similarity coefficient: negative matchings (both objects have value 0 for some variate $\hat{=}$ absence of a property) do not count.

$$s_{nm} = \frac{a}{a+b+c}$$

weighted:

$$s_{nm} = \frac{u \cdot a}{u \cdot a + (1-u) \cdot (b+c)}$$

With each u we get the same similarity ranking of pairs of objects (which does not coincide with the similarity ranking of the matching coefficient).

The similarity coefficient is **not** invariant with respect to bijective transformations of one or several variates x.

cluster analysis

examples for similarity and distance measures, binary data

correlation coefficient: we interchange objects and variates and compute the Pearson correlation for binary variates:

$$s_{nm} = \frac{a \cdot e - b \cdot c}{\sqrt{(a+c)(b+e)(a+b)(c+e)}}$$

problematic if the denominator is zero!

The correlation coefficient is invariant with respect to bijective transformations of one or several variates x.

cluster analysis

examples for similarity and distance measures, binary data

correlation coefficient: we interchange objects and variates and compute the Pearson correlation for binary variates:

$$\boxed{s_{nm} = \frac{a \cdot e - b \cdot c}{\sqrt{(a+c)(b+e)(a+b)(c+e)}}}$$

problematic if the denominator is zero!

The correlation coefficient is invariant with respect to bijective transformations of one or several variates x.

cluster analysis

examples for similarity and distance measures, multilevel nominal scaled data

multilevel nominal scaled data: **generalized matching coefficient**:

$$s_{nm} = \frac{u_{nm}}{p}$$

u_{nm} is the number of matching components of x_n and x_m

If we want to consider the number of values of the variates, i.e. matching in a variate with many values carries more weight than matching of a variate with few values, we choose:

$$s_{nm} = \frac{1}{m^*} \sum_{i=1}^{p} m_i \cdot \sigma(x_{ni}, x_{mi})$$

with m_i the number of values of x_i, $m^* = \sum_{i=1}^{p} m_i$ and

$$\sigma(x_{ni}, x_{mi}) = \begin{cases} 1 & \text{for } x_{ni} = x_{mi} \\ 0 & \text{else} \end{cases}$$

The coefficient is again invariant with respect to bijective transformations of one or several variates x.

cluster analysis

examples for similarity and distance measures, multilevel nominal scaled data

multilevel nominal scaled data: **generalized matching coefficient**:

$$s_{nm} = \frac{u_{nm}}{p}$$

u_{nm} is the number of matching components of x_n and x_m

If we want to consider the number of values of the variates, i.e. matching in a variate with many values carries more weight than matching of a variate with few values, we choose:

$$s_{nm} = \frac{1}{m^*} \sum_{i=1}^{p} m_i \cdot \sigma(x_{ni}, x_{mi})$$

with m_i the number of values of x_i, $m^* = \sum_{i=1}^{p} m_i$ and

$$\sigma(x_{ni}, x_{mi}) = \begin{cases} 1 & \text{for } x_{ni} = x_{mi} \\ 0 & \text{else} \end{cases}$$

The coefficient is again invariant with respect to bijective transformations of one or several variates x.

cluster analysis

examples for similarity and distance measures, multilevel nominal scaled data

multilevel nominal scaled data: **generalized matching coefficient**:

$$s_{nm} = \frac{u_{nm}}{p}$$

u_{nm} is the number of matching components of x_n and x_m

If we want to consider the number of values of the variates, i.e. matching in a variate with many values carries more weight than matching of a variate with few values, we choose:

$$s_{nm} = \frac{1}{m^*} \sum_{i=1}^{p} m_i \cdot \sigma(x_{ni}, x_{mi})$$

with m_i the number of values of x_i, $m^* = \sum_{i=1}^{p} m_i$ and

$$\sigma(x_{ni}, x_{mi}) = \begin{cases} 1 & \text{for } x_{ni} = x_{mi} \\ 0 & \text{else} \end{cases}$$

The coefficient is again invariant with respect to bijective transformations of one or several variates x.

cluster analysis

examples for similarity and distance measures, ordinal data

ordinal data: For the m_i values of x_i we have the order relation

$$x_{1i} \prec x_{2i} \prec \cdots \prec x_{m_i i} \qquad i = 1, \ldots, p$$

The closer the values x_{ni} and x_{mi} are with respect to the above order relation, the more similar the objects are assumed. But how?

For each ordinal variate we introduce as many auxiliary binary variates as the ordinal variate has levels. If the value x_{ni} is in position j in the ranking, then the first j binary variates are set to 1, the remaining variates are set to 0.
Finally we use similarity measures for binary variates.

cluster analysis

examples for similarity and distance measures, ordinal data

ordinal data: For the m_i values of x_i we have the order relation

$$x_{1i} \prec x_{2i} \prec \cdots \prec x_{m_i i} \qquad i = 1, \ldots, p$$

The closer the values x_{ni} and x_{mi} are with respect to the above order relation, the more similar the objects are assumed. But how?

For each ordinal variate we introduce as many auxiliary binary variates as the ordinal variate has levels. If the value x_{ni} is in position j in the ranking, then the first j binary variates are set to 1, the remaining variates are set to 0.
Finally we use similarity measures for binary variates.

cluster analysis

examples for similarity and distance measures, ordinal data

ordinal data: For the m_i values of x_i we have the order relation

$$x_{1i} \prec x_{2i} \prec \cdots \prec x_{m_i i} \qquad i = 1, \ldots, p$$

The closer the values x_{ni} and x_{mi} are with respect to the above order relation, the more similar the objects are assumed. But how?

For each ordinal variate we introduce as many auxiliary binary variates as the ordinal variate has levels. If the value x_{ni} is in position j in the ranking, then the first j binary variates are set to 1, the remaining variates are set to 0.
Finally we use similarity measures for binary variates.

cluster analysis

examples for similarity and distance measures, quantitative data

quantitative data: Both measuring scale and point of origin may be selected. Hence it is important to know whether similarity and distance measures are scale invariant (measure does not depend on measuring scale) or translation invariant (measure does not depend on the chosen point of origin).

scale invariance: Let $C = \text{diag}(c_1, \ldots, c_p)$; $c_i > 0$ and $\tilde{x}_n = C \cdot x_n$.
d is scale invariant if $d(x_n, x_m) = d(\tilde{x}_{,n}, \tilde{x}_m)$

translation invariance: Let $b \in \mathbb{R}^p$ and $\tilde{x}_n = x_n + b$.
d is translation invariant if $d(x_n, x_m) = d(\tilde{x}_n, \tilde{x}_m)$

cluster analysis

examples for similarity and distance measures, quantitative data

quantitative data: Both measuring scale and point of origin may be selected. Hence it is important to know whether similarity and distance measures are scale invariant (measure does not depend on measuring scale) or translation invariant (measure does not depend on the chosen point of origin).

scale invariance: Let $C = \text{diag}(c_1, \ldots, c_p)$; $c_i > 0$ and $\tilde{x}_n = C \cdot x_n$.
d is scale invariant if $d(x_n, x_m) = d(\tilde{x},_n, \tilde{x}_m)$

translation invariance: Let $b \in \mathbb{R}^p$ and $\tilde{x}_n = x_n + b$.
d is translation invariant if $d(x_n, x_m) = d(\tilde{x}_n, \tilde{x}_m)$

cluster analysis

examples for similarity and distance measures, quantitative data

quantitative data: Both measuring scale and point of origin may be selected. Hence it is important to know whether similarity and distance measures are scale invariant (measure does not depend on measuring scale) or translation invariant (measure does not depend on the chosen point of origin).

scale invariance: Let $C = \text{diag}(c_1, \ldots, c_p)$; $c_i > 0$ and $\tilde{x}_n = C \cdot x_n$.
d is scale invariant if $d(x_n, x_m) = d(\tilde{x}_{,n}, \tilde{x}_m)$

translation invariance: Let $b \in \mathbb{R}^p$ and $\tilde{x}_n = x_n + b$.
d is translation invariant if $d(x_n, x_m) = d(\tilde{x}_n, \tilde{x}_m)$

cluster analysis

examples for similarity and distance measures, quantitative data

L_q-distances (metrics)

$$d_q(n,m) = \left(\sum_{i=1}^{p} |x_{ni} - x_{mi}|^q \right)^{\frac{1}{q}} \qquad q \geq 1$$

L_q-distances are translation invariant but not scale invariant

Prior to the computation of the L_q-distance the variates must be transformed into the same measuring unit (standardization). Mostly the following standardization is used:

$$\tilde{x}_{ni} = \frac{x_{ni} - \bar{x}_i}{s_i^q} \qquad n = 1, \ldots, N \qquad i = 1, \ldots, p$$

with $s_i^q = \left(\frac{1}{N} \sum |x_{ni} - \bar{x}_i|^q \right)^{\frac{1}{q}}$

cluster analysis

examples for similarity and distance measures, quantitative data

L_q-distances (metrics)

$$d_q(n,m) = \left(\sum_{i=1}^{p} |x_{ni} - x_{mi}|^q \right)^{\frac{1}{q}} \qquad q \geq 1$$

L_q-distances are translation invariant but not scale invariant

Prior to the computation of the L_q-distance the variates must be transformed into the same measuring unit (standardization). Mostly the following standardization is used:

$$\tilde{x}_{ni} = \frac{x_{ni} - \bar{x}_i}{s_i^q} \qquad n = 1, \ldots, N \qquad i = 1, \ldots, p$$

with $s_i^q = \left(\frac{1}{N} \sum |x_{ni} - \bar{x}_i|^q \right)^{\frac{1}{q}}$

L_q-distances (metrics)

$$d_q(n,m) = \left(\sum_{i=1}^{p} |x_{ni} - x_{mi}|^q \right)^{\frac{1}{q}} \qquad q \geq 1$$

L_q-distances are translation invariant but not scale invariant

Prior to the computation of the L_q-distance the variates must be transformed into the same measuring unit (standardization). Mostly the following standardization is used:

$$\tilde{x}_{ni} = \frac{x_{ni} - \bar{x}_i}{s_i^q} \qquad n = 1, \ldots, N \qquad i = 1, \ldots, p$$

with $s_i^q = \left(\frac{1}{N} \sum |x_{ni} - \bar{x}_i|^q \right)^{\frac{1}{q}}$

cluster analysis

examples for similarity and distance measures, quantitative data

Frequently used distances:

$d_1(n,m) = \sum_{i=1}^{p} |x_{ni} - x_{mi}|$ city block metric

$d_2(n,m) = \sqrt{\sum_{i=1}^{p} (x_{ni} - x_{mi})^2} = \|x_n - x_m\|_2$ Euclidean norm, corresponds to our geometric distance perception.

The Euclidean distance is furthermore invariant with respect to orthogonal transformations of x, i.e. if C is an orthogonal matrix, then

$$d_2(x_n, x_m) = d_2(C \cdot x_n, C \cdot x_m)$$

The Euclidean distance does not change with rotation and reflection of the coordinate system.

cluster analysis

examples for similarity and distance measures, quantitative data

Frequently used distances:

$d_1(n,m) = \sum_{i=1}^{p} |x_{ni} - x_{mi}|$ city block metric

$d_2(n,m) = \sqrt{\sum_{i=1}^{p}(x_{ni} - x_{mi})^2} = \|x_n - x_m\|_2$ Euclidean norm, corresponds to our geometric distance perception.

The Euclidean distance is furthermore invariant with respect to orthogonal transformations of x, i.e. if C is an orthogonal matrix, then

$$d_2(x_n, x_m) = d_2(C \cdot x_n, C \cdot x_m)$$

The Euclidean distance does not change with rotation and reflection of the coordinate system.

cluster analysis

examples for similarity and distance measures, quantitative data

Frequently used distances:

$d_1(n,m) = \sum_{i=1}^{p} |x_{ni} - x_{mi}|$ city block metric

$d_2(n,m) = \sqrt{\sum_{i=1}^{p}(x_{ni} - x_{mi})^2} = \|x_n - x_m\|_2$ Euclidean norm, corresponds to our geometric distance perception.

The Euclidean distance is furthermore invariant with respect to orthogonal transformations of x, i.e. if C is an orthogonal matrix, then

$$d_2(x_n, x_m) = d_2(C \cdot x_n, C \cdot x_m)$$

The Euclidean distance does not change with rotation and reflection of the coordinate system.

Mahalanobis distance

$$d_M(n,m) = \sqrt{(x_n - x_m)^T K^{-1} (x_n - x_m)}$$

where $K = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T$ is the empirical covariance matrix.

$d_M(n,m)$ is invariant with respect to arbitrary non-singular linear transformations, i.e. if C is an arbitrary regular $(p \times p)$-matrix and $b \in \mathbb{R}^p$ arbitrary and $\tilde{x}_n = C \cdot x + b$, then

$$d_M(\tilde{x}_n, \tilde{x}_m) = d_M(x_n, x_m)$$

Mahalanobis distance

$$d_M(n,m) = \sqrt{(x_n - x_m)^T K^{-1} (x_n - x_m)}$$

where $K = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T$ is the empirical covariance matrix.

$d_M(n,m)$ is invariant with respect to arbitrary non-singular linear transformations, i.e. if C is an arbitrary regular $(p \times p)$-matrix and $b \in \mathbb{R}^p$ arbitrary and $\tilde{x}_n = C \cdot x + b$, then

$$d_M(\tilde{x}_n, \tilde{x}_m) = d_M(x_n, x_m)$$

cluster analysis

examples for similarity and distance measures, quantitative data

Further important properties of the Mahalanobis distance:

Let $\tilde{x}_n = K^{-\frac{1}{2}} x_n \quad \Longrightarrow \quad \tilde{x}_n$ is uncorrelated. Then

$$d_M(n,m) = \sqrt{(x_n - x_m)^T \underbrace{K^{-\frac{T}{2}} K^{-\frac{1}{2}}}_{K^{-1}} (x_n - x_m)} =$$

$$= \sqrt{(\tilde{x}_n - \tilde{x}_m)^T (\tilde{x}_n - \tilde{x}_m)} = \|\tilde{x}_n - \tilde{x}_m\| = d_2(\tilde{x}_n, \tilde{x}_m)$$

The computation of the Mahalanobis distance is equivalent to

- transform the variates to uncorrelated variates
- then compute the Euclidean distances of the uncorrelated variates

cluster analysis

examples for similarity and distance measures, quantitative data

Further important properties of the Mahalanobis distance:
Let $\tilde{x}_n = K^{-\frac{1}{2}} x_n \quad \Longrightarrow \quad \tilde{x}_n$ is uncorrelated. Then

$$d_M(n,m) = \sqrt{(x_n - x_m)^T \underbrace{K^{-\frac{T}{2}} K^{-\frac{1}{2}}}_{K^{-1}} (x_n - x_m)} =$$

$$= \sqrt{(\tilde{x}_n - \tilde{x}_m)^T (\tilde{x}_n - \tilde{x}_m)} = \|\tilde{x}_n - \tilde{x}_m\| = d_2(\tilde{x}_n, \tilde{x}_m)$$

The computation of the Mahalanobis distance is equivalent to

- transform the variates to uncorrelated variates
- then compute the Euclidean distances of the uncorrelated variates

cluster analysis

examples for similarity and distance measures, quantitative data

Further important properties of the Mahalanobis distance:
Let $\tilde{x}_n = K^{-\frac{1}{2}} x_n \quad \Longrightarrow \quad \tilde{x}_n$ is uncorrelated. Then

$$d_M(n,m) = \sqrt{(x_n - x_m)^T \underbrace{K^{-\frac{T}{2}} K^{-\frac{1}{2}}}_{K^{-1}} (x_n - x_m)} =$$

$$= \sqrt{(\tilde{x}_n - \tilde{x}_m)^T (\tilde{x}_n - \tilde{x}_m)} = \|\tilde{x}_n - \tilde{x}_m\| = d_2(\tilde{x}_n, \tilde{x}_m)$$

The computation of the Mahalanobis distance is equivalent to

- transform the variates to uncorrelated variates
- then compute the Euclidean distances of the uncorrelated variates