

Reinhard Dobbener
Universität Bamberg

Zur Skalen- und Translationsinvarianz von Metriken

(On the Invariance of Metrics on $\mathbb{R}^n$ under Certain Families of Transformation of $\mathbb{R}^n$)

Dobbener, R.: Zur Skalen- und Translationsinvarianz von Metriken. (On the invariance of metrics on $\mathbb{R}^n$ under certain families of transformations of $\mathbb{R}^n$) (In German)

In: Intern. Classificat. 8 (1981) No. 2, p. 64–68, 4 refs.

A problem of special interest in the field of cluster analysis is the characterisation of the metrics on $\mathbb{R}^n$ which are invariant under certain groups of transformation of $\mathbb{R}^n$. If all variables under consideration are measured on an interval scale it is appropriate to use metrics which are invariant under the group of transformations generated by the diagonal $n \times n$ -matrices and the translations of $\mathbb{R}^n$. The characterisation of these metrics leads to a frustrating result: This article proves that any metric invariant under this group of transformations treats the interval variables as classificatory variables. In case of ratio variables one can restrict one's attention to $\mathbb{R}^{n+}$ and the group of positive diagonal-matrix-transformations of $\mathbb{R}^n$. The result is better now: A logarithm-like mapping of $\mathbb{R}^{n+}$ to $\mathbb{R}^n$ makes it possible to identify these metrics with the (well known) translation-invariant metrics on $\mathbb{R}^n$.

Author

0. Problemstellung

Ein Ziel der Numerischen Klassifikation ist es, eine gegebene Objektmenge $\Omega = \{O_1, \dots, O_k\}$ anhand einer Reihe von Merkmalen $X_1, \dots, X_n$ so in disjunkte Klassen zu zerlegen, daß Objekte einer Klasse einander „möglichst ähnlich“, Objekte aus verschiedenen Klassen dagegen „möglichst unähnlich“ sind. Für numerische Verfahren reicht eine intuitive Vorstellung von Ähnlichkeit und Unähnlichkeit nicht aus, sie benötigen eine in (reellen) Zahlen ausgedrückte Quantifizierung dieser Begriffe. Zur Quantifizierung dieser (Un-)Ähnlichkeiten greift man üblicherweise auf die an den Objekten beobachteten Merkmalsausprägungen $x_i = (x_{i1}, \dots, x_{in})^1$ (beim Objekt O_i) zurück und benutzt eine Metrik d auf dem Merkmalsraum: die Zahl $d(x_i, x_j)$ ist dann ein Maß für die Unähnlichkeit der Objekte O_i und O_j . Sind die Merkmale $X_1, \dots, X_n$ metrisch (der Merkmalsraum ist dann $\mathbb{R}^n$ oder ein Teil davon), so sollte man erwarten, daß die Unähnlichkeit zweier Objekte nicht von den mehr oder weniger zufällig gewählten Maßeinheiten der einzelnen Merkmale abhängt (vgl. [3]). Für die zu benutzende Metrik d bedeutet dies, daß sie unter Transformationen D , die eine solche Skalenänderung darstellen, invariant sein sollen:

$$d(Dx_i, Dx_j) = d(x_i, x_j)$$

für alle x_i und x_j ($1 \leq i, j \leq k$).

In dieser Arbeit werden für bestimmte Familien von

Transformationen die invarianten Metriken charakterisiert.

1. Definition und elementare Eigenschaften

Die Überlegungen aus Abschnitt 0 führen zu der allgemeinen Definition (vgl. [1] S. 27):

Definition: Eine Metrik d auf $\mathbb{R}^n$ heißt $\mathcal{H}$ -invariant ($\mathcal{H}$ ist eine Familie von Transformationen des $\mathbb{R}^n$), wenn

$$d(x, y) = d(Hx, Hy)$$

für alle $x, y \in \mathbb{R}^n$ und $H \in \mathcal{H}$ gilt.

Man kann zeigen, daß die Menge

$$J(d) = \{H: \mathbb{R}^n \rightarrow \mathbb{R}^n \text{ bijektiv; } d \text{ ist } H\text{-invariant}\}$$

eine Gruppe im algebraischen Sinne ist (vgl. [2] S. 5). Dies ist einer der Gründe dafür, daß es sinnvoll ist, hier nur solche Familien $\mathcal{H}$ von Transformationen zu studieren, die auch Gruppen sind. (Ein anderer Grund ist in [2], S. 1 zu finden.)

Beispiele für solche Gruppen sind:

$$(1) \quad \mathcal{T} = \{T_t: \mathbb{R}^n \rightarrow \mathbb{R}^n; T_t(x) = x + t; t \in \mathbb{R}^n\}$$

$\mathcal{T}$ -invariante Metriken heißen translationsinvariant.

Alle durch Normen $\|\cdot\|$ auf $\mathbb{R}^n$ definierten Metriken ($d(x, y) = \|x - y\|$) sind translationsinvariant.

$$(2) \quad \mathcal{O} = \{O: \mathbb{R}^n \rightarrow \mathbb{R}^n; O \text{ ist orthogonal}\}$$

Invarianz unter $\mathcal{O}$ ist in weiten Bereichen der Physik wichtig (Unabhängigkeit physikalischer Erscheinungen von der Lage des räumlichen Koordinatensystems, in dem sie beschrieben werden). Die Euklidische Metrik ist das klassische Beispiel einer $\mathcal{O}$ -invarianten Metrik.

$$(3) \quad \mathcal{D} = \{D: \mathbb{R}^n \rightarrow \mathbb{R}^n; D \text{ ist regulär und diagonal}\}$$

$\mathcal{D}$ ist die Gruppe der Transformationen, die Maßstabsänderungen darstellen. Sie ist, wie schon im Abschnitt 0 gezeigt, für die Numerische Klassifikation von besonderer Bedeutung. Die Untersuchung $\mathcal{D}$ -invarianter Metriken ist der wesentliche Gegenstand dieser Arbeit.

(4) Der Vollständigkeit halber sei hier auch die Gruppe $\mathcal{A}$ aller regulären, linearen Transformationen des $\mathbb{R}^n$ erwähnt. Obwohl $\mathcal{A}$ -Invarianz im Rahmen der Numerischen Klassifikation eine kaum begründbare Forderung an Metriken darstellt, lassen sich die in dieser Arbeit benutzten Methoden leicht zur Charakterisierung $\mathcal{A}$ -invarianter Metriken benutzen.

Für das weitere Vorgehen ist es notwendig, zwischen intervall- und verhältnisskalierten (metrischen) Merkmalen zu unterscheiden.

- 1) Ein Merkmal heißt intervallskaliert, wenn es positive und negative Werte annehmen kann und die verwendete Skala keinen ausgezeichneten (natürlichen) Nullpunkt besitzt. Beispiele: Temperatur in $^{\circ}\text{C}$, Erträge im betriebswirtschaftlichen Sinn.
- 2) Ein metrisches Merkmal heißt verhältnisskaliert, wenn seine Ausprägungen nur positive reelle Zahlen sein können und die benutzte Skala einen natürlichen Nullpunkt besitzt. Beispiele: Alter, Länge, Gewicht.

Bei intervallskalierten Merkmalen ist der Merkmalsraum der gesamte $\mathbb{R}^n$. Da sowohl Nullpunkt als auch Skalierung der einzelnen Merkmale zunächst als zufällig anzusehen sind, ist es notwendig, bei intervallskalierten Merkmalen Skalen- und Translations- ($\mathcal{D}$ - und $\mathcal{T}$ -) Invarianz von der verwendeten Metrik zu verlangen. Im Abschnitt 2 werden die $\mathcal{D}$ - und $\mathcal{T}$ -invarianten Metriken auf $\mathbb{R}^n$ charakterisiert.

Sind dagegen alle Merkmale verhältnisskaliert, so ist es möglich, auf Translationsinvarianz zu verzichten, da es ja einen natürlichen Nullpunkt gibt. Da außerdem der Merkmalsraum nur $\mathbb{R}^{n+} = \{x \in \mathbb{R}^n; x_i \geq 0, i = 1, \dots, n\}$ umfaßt, reicht es aus, statt Invarianz unter $\mathcal{D}$ nur Invarianz unter $\mathcal{D}^+ = \{D \in \mathcal{D}; d_{ii} > 0, i = 1, \dots, n\}$ zu verlangen. In Abschnitt 3 werden zunächst die $\mathcal{D}^+$ -invarianten Metriken auf $\mathbb{R}^{n+} = \{x \in \mathbb{R}^n, x_i > 0, i = 1, \dots, n\}$ beschrieben. Eine Methode, mit der solche Metriken auf $\mathbb{R}^{n+}$ fortgesetzt werden können, wird ebenfalls in Abschnitt 3 beschrieben.

2. $\mathcal{D}$ - und $\mathcal{T}$ -invariante Metriken auf $\mathbb{R}^n$

Zunächst werden Eigenschaften $\mathcal{D}$ - und $\mathcal{T}$ -invarianter Metriken auf $\mathbb{R}^n$ abgeleitet, von denen sich dann zeigen läßt, daß sie die $\mathcal{D}$ - und $\mathcal{T}$ -invarianten Metriken vollständig charakterisieren.

Da zu sei d eine solche $\mathcal{D}$ - und $\mathcal{T}$ -invariante Metrik. Weiterhin ist es nützlich, $[x]$ (Klasse von x) einzuführen:

$$[x] = \{D \cdot x; D \in \mathcal{D}\}.$$

Aus der Definition der $\mathcal{D}$ -Invarianz folgt, daß für $x, y \in [x]$

$$d(x, 0) = d(y, 0)$$

gilt; also hängt $d(x, 0)$ nur von $[x]$ ab. $[x]$ ist aber durch $A(x) = \{i, x_i \neq 0\}^2$ vollständig bestimmt: $A(x) = A([x])$. Da mit gilt:

$$d(x, 0) = f_d(A(x))$$

f_d ist dabei eine von d abhängige Funktion auf $\mathcal{P}(\{1, \dots, n\})$, der Potenzmenge von $\{1, \dots, n\}$.

Aus der Translationsinvarianz von d folgt dann für beliebige $x, y \in \mathbb{R}^n$:

$$d(x, y) = d(x - y, 0) = f_d(A(x - y))$$

Damit ist schon das folgende Lemma bewiesen:

Lemma 1: Ist d eine $\mathcal{D}$ - und $\mathcal{T}$ -invariante Metrik auf $\mathbb{R}^n$, so hängt $d(x, y)$ nur von $A(x - y)$ ab:

$$d(x, y) = f_d(A(x - y))$$

Die Umkehrung von Lemma 1 ist auch richtig:

Lemma 2: Gibt es zu der Metrik d auf $\mathbb{R}^n$ eine Funktion $f_d: \mathcal{P}(\{1, \dots, n\}) \rightarrow \mathbb{R}^+$, so daß

$$d(x, y) = f_d(A(x - y))$$

ist, so ist d $\mathcal{D}$ - und $\mathcal{T}$ -invariant.

Beweis: Da A nur von $[x - y]$ abhängt, gilt für $D \in \mathcal{D}$ und $t \in \mathbb{R}^n$

$$\begin{aligned} A(Dx + t - (Dy + t)) &= A(D \cdot (x - y)) \\ &= A(x - y). \end{aligned}$$

Deshalb ist

$$\begin{aligned} d(Dx + t, Dy + t) &= f_d(A(Dx + t - (Dy + t))) \\ &= f_d(A(x - y)) \\ &= d(x, y). \end{aligned}$$

Da dies für alle $x, y \in \mathbb{R}^n$ und beliebige $D \in \mathcal{D}$ und $t \in \mathbb{R}^n$ gilt, ist damit die $\mathcal{D}$ - und $\mathcal{T}$ -Invarianz von d gezeigt.

Diese beiden Lemma stellen auch den wichtigsten Teil des folgenden Satzes dar:

Satz 1: Eine Funktion d auf $\mathbb{R}^n \times \mathbb{R}^n$ ist genau dann eine $\mathcal{D}$ - und $\mathcal{T}$ -invariante Metrik, wenn es eine Funktion f_d auf $\mathcal{P}(\{1, \dots, n\})$ gibt mit den Eigenschaften

- (i) $d(x, y) = f_d(A(x - y))$
- (ii) $f_d(A) \geq 0$ für alle $A \in \mathcal{P}(\{1, \dots, n\})$
- (iii) $f_d(A) = 0$ genau dann, wenn $A = \emptyset$
- (iv) Für $A, B, C \subset \{1, \dots, n\}$ mit $A \subset (B \cup C)$ und $((B \cup C) - (B \cap C)) \subset A$ gilt:
 $f_d(A) \leq f_d(B) + f_d(C)$

Beweis: Der wesentliche Teil des Satzes, daß nämlich bei translations- und skaleninvarianten Metriken $d(x, y)$ nur von $A(x - y)$, der Menge der Indizes i , für die $x_i \neq y_i$ ist, abhängt, wurde in den beiden Lemmata bewiesen. Die Eigenschaften (ii) – (iv) drücken mit Hilfe der Funktion f_d aus, daß d eine Metrik ist. Der vollständige Beweis dieses Satzes ist eine recht einfache, aber umfangreiche Rechnung und wird deshalb hier nicht aufgeführt. Er ist in [2] S. 10ff. zu finden.

Das wichtigste Ergebnis von Satz 1 ist, daß für $\mathcal{D}$ - und $\mathcal{T}$ -invariante Metriken $d(x, y)$ nur von $A(x - y) = \{i, x_i \neq y_i\}$ abhängt. Dies hat zur Folge, daß $\mathcal{D}$ - und $\mathcal{T}$ -invariante Metriken nicht nur unter Transformationen aus der von $\mathcal{D}$ und $\mathcal{T}$ erzeugten Gruppe invariant sind, sondern unter allen Abbildungen von der Form $(x_1, \dots, x_n) \rightarrow (f_1(x_1), \dots, f_n(x_n))$, wobei die f_i bijektive Abbildungen von $\mathbb{R}$ in irgendwelche Mengen M_i sind ($i = 1, \dots, n$). Dies bedeutet aber, daß von der ganzen topologischen und algebraischen Struktur des $\mathbb{R}^n$ nur die Eigenschaft des n-fachen kartesischen Produkts erhalten bleibt. Man kann für Objektpaare (Punktpaare) nur noch komponentenweise entscheiden, ob sie gleich oder verschieden sind. Die n metrischen Merkmale werden also wie n klassifikatorische Merkmale behandelt. Dies ist gegenüber der ursprünglichen Struktur der Merkmale ein so großer Informationsverlust, daß $\mathcal{D}$ - und $\mathcal{T}$ -invariante Metriken für praktische Anwendungen kaum in Frage kommen.

3. Skaleninvarianz bei verhältnisskalierten Merkmalen

Sind die Objekte durch verhältnisskalierte Merkmale charakterisiert, so ist es vernünftig (vgl. Abschn. 1), auf die Forderung der Translationsinvarianz zu verzichten. Da die Merkmalsausprägung außerdem auf $\mathbb{R}^{n+}$ beschränkt sind, reicht es aus, nur solche Skalentransformationen zu studieren, die den $\mathbb{R}^{n+}$ auf sich selbst abbilden. Das ist aber gerade

$$\mathcal{D}^+ = \{D \in \mathcal{D}; d_{ii} > 0\}.$$

Ziel dieses Abschnittes ist es also, die Metriken d auf $\mathbb{R}^{n+}$ zu charakterisieren, für die $d(x, y) = d(Dx, Dy)$ (für alle $x, y \in \mathbb{R}^{n+}$ und $D \in \mathcal{D}^+$) gilt.

$$\text{Setzt man } e = \underbrace{(1, \dots, 1)}_{n\text{-mal}} \text{ und } x/y = \left(\frac{x_1}{y_1}, \dots, \frac{x_n}{y_n}\right)$$

(wegen $x, y \in \mathbb{R}^{n+}$ ist x/y immer definiert), so gilt für $\mathcal{D}^+$ -invariante Metriken d auf $\mathbb{R}^{n+}$:

$$d(x, y) = d(e, x/y)$$

Durch komponentenweises Logarithmieren

$$(\log x = (\log x_1, \dots, \log x_n))$$

kann man diesen Ausdruck folgendermaßen umformen:

$$\begin{aligned} d(x, y) &= d^+(\log x, \log y) = \\ &= d^+(0, \log x - \log y) \end{aligned}$$

d^+ ist dabei eine von d abhängige Funktion auf $\mathbb{R}^n \times \mathbb{R}^n$. Man kann sofort erkennen, daß aus der Skaleninvarianz von d die Translationsinvarianz von d^+ folgt. Bevor Satz 2, der die $\mathcal{D}^+$ -invarianten Metriken d auf $\mathbb{R}^n$ durch Funktionen d^+ auf $\mathbb{R}^n \times \mathbb{R}^n$ charakterisiert, formuliert werden kann, sind noch einige Bezeichnungen zu vereinbaren:

$$\exp x = (\exp x_1, \dots, \exp x_n) \quad x \in \mathbb{R}^n \quad (\exp x \in \mathbb{R}^{n+})$$

$$D_{\exp t} = \begin{pmatrix} \exp t_1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \ddots & \\ 0 & \dots & 0 & \exp t_n \end{pmatrix} \quad t \in \mathbb{R}^n \quad (D_{\exp t} \in \mathcal{D}^+)$$

$$t_{\log D} = (\log d_{1,1}, \dots, \log d_{n,n}) \quad d \in \mathcal{D}^+ \quad (t_{\log D} \in \mathbb{R}^n)$$

Mit diesen Vereinbarungen gilt folgender Satz:

Satz 2: Eine Abbildung d auf $\mathbb{R}^{n+} \times \mathbb{R}^{n+}$ ist genau dann eine $\mathcal{D}^+$ -invariante Metrik, wenn die durch

$$(1) \quad d^+(x, y) = d(\exp x, \exp y) \quad x, y \in \mathbb{R}^n$$

definierte Abbildung eine translationsinvariante Metrik auf $\mathbb{R}^n$ ist.

Beweis:

a) Sei d eine $\mathcal{D}^+$ -invariante Metrik auf $\mathbb{R}^{n+}$, so ist

zu zeigen, daß die durch (1) definierte Abbildung d^+ eine Metrik ist:

- (i) $d^+(x, y) \geq 0$
- (ii) $d^+(x, y) = 0$ genau dann, wenn $x = y$
- (iii) $d^+(x, y) = d^+(y, x)$
- (iv) $d^+(x, y) \leq d^+(x, z) + d^+(z, y)$
- (v) $d^+(x + t, y + t) = d^+(x, y) \quad t \in \mathbb{R}^n$

Die Eigenschaften (i) – (iv) drücken aus, daß d^+ eine Metrik ist und folgen direkt aus den entsprechenden Eigenschaften von d und der Tatsache, daß $\exp: \mathbb{R}^n \rightarrow \mathbb{R}^{n+}$ bijektiv ist.

Die Translationsinvarianz (v) folgt so:

$$\begin{aligned} d^+(x + t, y + t) &= d(\exp(x + t), \exp(y + t)) \\ &= d(D_{\exp t} \cdot \exp x, D_{\exp t} \cdot \exp y) \\ &= d(\exp x, \exp y) \\ &= d^+(x, y) \end{aligned}$$

Damit ist gezeigt, daß d^+ eine translationsinvariante Metrik auf $\mathbb{R}^n$ ist, wenn d eine skaleninvariante Metrik auf $\mathbb{R}^{n+}$ ist.

b) Sei andererseits d^+ eine translationsinvariante Metrik auf $\mathbb{R}^n$; es muß jetzt gezeigt werden, daß durch

$$d(x, y) = d^+(\log x, \log y) \quad x, y \in \mathbb{R}^{n+}$$

eine skaleninvariante Metrik auf $\mathbb{R}^{n+}$ definiert wird. Mit den gleichen Argumenten wie in a) folgt, daß d eine Metrik ist. Zum Nachweis der Skaleninvarianz sei $D \in \mathcal{D}$ beliebig. Dann gilt für $x, y \in \mathbb{R}^{n+}$:

$$\begin{aligned} d(D \cdot x, D \cdot y) &= d^+(\log(D \cdot x), \log(D \cdot y)) \\ &= d^+(t_{\log D} + \log x, t_{\log D} + \log y) \\ &= d^+(\log x, \log y) \\ &= d(x, y) \end{aligned}$$

Satz 2 ist damit vollständig bewiesen.

Ein Beispiel möge die Aussage von Satz 2 verdeutlichen. Definieren wir $d_2(x, y) = \|\log x - \log y\|_2$ (d_2 kann als „Euklidische“ Metrik auf $\mathbb{R}^{n+}$ bezeichnet werden), so gilt für die Punkte $x = (1, 1)$ und $y = (e^{-1}, e)$ ($\in \mathbb{R}^{2+}$) $d_2(x, y) = \|(0, 0) - (-1, 1)\|_2 = \sqrt{2}$. Führt man dann die Skalentransformation

$$D = \begin{pmatrix} e^2 & 0 \\ 0 & e^{-1} \end{pmatrix}, \text{ also } D \cdot x = (e^2, e^{-1}), D \cdot y = (e, 1),$$

durch, so gilt:

$$d_2(Dx, Dy) = \|(2, -1) - (1, 0)\|_2 = \sqrt{2}.$$

Satz 2 gestattet es also, die skaleninvarianten Metriken auf $\mathbb{R}^{n+}$ mit den translationsinvarianten auf $\mathbb{R}^n$ zu identifizieren.

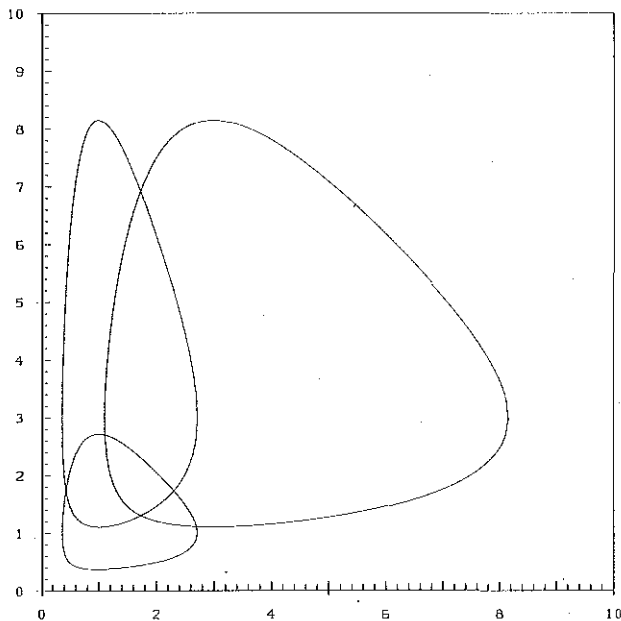


Abb. 1:

So kann man beispielsweise durch jede Norm $\|\cdot\|$ auf $\mathbb{R}^n$ eine $\mathcal{D}^+$ -invariante Metrik d auf $\mathbb{R}^{n+}$ durch

$$d(x, y) = \|\log x - \log y\|$$

definieren. Um einen geometrischen Eindruck von solchen Metriken zu gewinnen, sind in Abb. 1 für die „Euklidische“ Metrik d_2 und für einige Punkte x_0 die Mengen

$$K(x_0; 1) = \{x \in \mathbb{R}^{2+}; d_2(x, x_0) = 1\}$$

gezeichnet.

Da logarithmische Transformationen eine wichtige Rolle bei der Charakterisierung der $\mathcal{D}^+$ -invarianten Metriken spielen, ist es verständlich, daß als Merkmalsraum zunächst $\mathbb{R}^{n+}$ (und nicht $\mathbb{R}^{n+}$) gewählt wurde, da $\log 0$ nicht definiert ist. Eine Möglichkeit, eine Metrik auf $\mathbb{R}^{n+}$ fortzusetzen, ist diese:

Man wählt eine beschränkte, translationsinvariante Metrik

$$\delta \text{ auf } \mathbb{R} \text{ (z.B.: } \delta_1(x_i, y_i) = \frac{|x_i - y_i|}{1 + |x_i - y_i|})$$

und setzt diese Metrik nach $-\infty$ fort durch

$$\delta(-\infty, x_i) = \sup_{x_i, y_i \in \mathbb{R}} \delta(x_i, y_i)$$

(also z.B. $\delta_1(-\infty, x_i) = 1$).

Auf $(-\infty \cup \mathbb{R})^n$ ist dann durch

$$\delta(x, y) = \sum_{i=1}^n \delta(x_i, y_i)$$

eine translationsinvariante Metrik definiert.

Vereinbart man jetzt $\log 0 = -\infty$, so ist durch

$$d(x, y) = \delta(\log x, \log y)$$

eine skaleninvariante Metrik auf $\mathbb{R}^{n+} = \{x \in \mathbb{R}^n; x_i \geq 0\}$ definiert.

4. Schlußbemerkungen

1) Die Ergebnisse aus Abschnitt 1 und Abschnitt 2 lassen sich ohne Schwierigkeiten auf Distanzindizes übertragen. (Distanzindizes sind „Metriken“, für die die Dreiecksungleichung nicht gelten muß). In Satz 1 muß lediglich die Bedingung (iv) gestrichen werden, wenn man Distanzindizes statt Metriken studiert. Satz 2 bleibt richtig, wenn man das Wort „Metrik“ durch „Distanzindex“ ersetzt.

2) Da die Gruppe $\mathcal{R}$ der regulären $n \times n$ Matrizen größer ist als $\mathcal{D}$, gibt es natürlich weniger $\mathcal{R}$ - (bzw. $\mathcal{R}$ - und $\mathcal{F}$ -) als $\mathcal{D}$ - (bzw. $\mathcal{D}$ - und $\mathcal{F}$ -) invariante Metriken. Insbesondere gibt es im wesentlichen nur eine $\mathcal{R}$ - und $\mathcal{F}$ -invariante Metrik auf $\mathbb{R}^n$; das liegt daran, daß

$$\mathcal{R} \cdot x = \{Rx; R \in \mathcal{R}\} = \begin{cases} \mathbb{R}^n - \{0\} & \text{falls } x \neq 0 \\ \{0\} & \text{falls } x = 0 \end{cases}$$

gilt (vgl. Abschnitt 1 und [2] S. 6).

Will man $\mathcal{R}$ -invariante Metriken auf $\mathbb{R}^{n+}$ studieren, so muß man sich auf diejenigen $R \in \mathcal{R}$ beschränken, die $\mathbb{R}^{n+}$ bijektiv auf sich selbst abbilden:

$$\mathcal{R}^+ = \{R \in \mathcal{R}; R(\mathbb{R}^{n+}) = \mathbb{R}^{n+}\}.$$

Überlegungen der linearen Algebra zeigen jedoch, daß $\mathcal{R}^+$ die von $\mathcal{D}^+$ und $\mathcal{D}^3$ erzeugte Gruppe ist.

Die $\mathcal{R}^+$ -invarianten Metriken auf $\mathbb{R}^{n+}$ sind also die $\mathcal{D}^+$ -invarianten, die gegenüber Vertauschungen der Reihenfolge der Merkmale invariant sind.

3) Dem aufmerksamen Leser wird nicht entgangen sein, daß bisher an keiner Stelle dieser Arbeit der Mahalanobis-Abstand erwähnt wurde, obwohl er ja üblicherweise zu den $\mathcal{R}$ - (und $\mathcal{F}$ -) invarianten Metriken gezählt wird. Ich werde deshalb jetzt die Vorgehensweise des Mahalanobis-Abstandes im Hinblick auf seine Invarianzeigenschaften beschreiben.

Dazu ist es zunächst wichtig festzustellen, daß der Mahalanobis-Abstand

$$d(x, y) = (x - y) \cdot \hat{\Sigma}^{-1} \cdot (x - y)^t$$

($\hat{\Sigma}$ ist die Varianz-Kovarianz-Matrix des Datensatzes $\mathcal{X} = \{x_1, \dots, x_k\}$) wesentlich von der Datenmatrix $\mathcal{X}$ abhängt:

$$d(x, y) = d_{\mathcal{X}}(x, y).$$

Der Mahalanobis-Abstand erfüllt dann die Gleichung:

$$d_{\mathcal{X}}(x, y) = d_R \cdot \mathcal{X}(Rx, Ry) \quad R \in \mathcal{R}$$

Dies ist aber nicht die Bedingung, die in Abschn. 1 an $\mathcal{R}$ -invariante Metriken gestellt wurde. Der Mahalanobis-Abstand ist also keine $\mathcal{R}$ -invariante Metrik (im Sinne von Abschnitt 1).

Das Invarianzverhalten des Mahalanobis-Abstandes läßt sich besser auf eine andere Art beschreiben:

In der Äquivalenzklasse $\mathcal{R} \cdot \mathcal{X} = \{R \cdot \mathcal{X}; R \in \mathcal{R}\}$ der Datenmatrix gibt es genau eine Datenmatrix $\mathcal{X}_0$, die durch die Einheitsmatrix als Varianz-Kovarianz-Matrix ausgezeichnet ist. Der Mahalanobis-Abstand kann dann so interpretiert werden, daß er $\mathcal{X}$ in die aufgezeichnete Datenmatrix $\mathcal{X}_0$ ($\in \mathcal{R} \cdot \mathcal{X}$) transformiert und aus diesen Daten die Distanzen zwischen den Objekten berechnet. Ein solches Verfahren ist selbstverständlich $\mathcal{R}$ -invariant, da jede Datenmatrix $R \cdot \mathcal{X}$ ($\in \mathcal{R} \cdot \mathcal{X}$) in $\mathcal{X}_0$ transformiert wird.

Problematisch an diesem Vorgehen ist die Annahme, daß die Datenmatrix $\mathcal{X}_0$ (mit I_n als Varianz-Kovarianz-Matrix) tatsächlich für die Numerische Klassifikation eine „ausgezeichnete“ Rolle spielt: Es gibt kein Klassifikationsverfahren, von dem man annehmen muß, daß die Voraussetzungen für seine Anwendung in $\mathcal{X}_0$ eher erfüllt sind als in irgendeinem anderen Datensatz aus $\mathcal{R} \cdot \mathcal{X}$. Ähnliche Überlegungen für Verfahren, die mit standardisierten Merkmalen arbeiten, führen zu entsprechenden Ergebnissen; auch hier muß man sich fragen, ob der „standardisierte“ Datensatz im Hinblick auf die anzuwendenden Verfahren tatsächlich ausgezeichnet ist.

4) Satz 2 führt auch zu Überlegungen, die über das Invarianzproblem der Numerischen Klassifikation hinausgehen: Genauso wie es möglich ist, Metriken des $\mathbb{R}^n$ auf $\mathbb{R}^{n+}$ zu übertragen, kann es sinnvoll sein, Lage- und Streuungsmaße von $\mathbb{R}$ auf $\mathbb{R}^+$ mittels der exp-Transformation zu übertragen. Beispielsweise kann man mit Hilfe des arithmetischen Mittels für m Beobachtungen $x_1, \dots, x_m$ eines verhältnisskalierten Merkmals das Lokalisationsmaß $\bar{x}$ durch

$$\bar{x} = \exp \left(\frac{1}{m} \sum_{i=1}^m \log x_i \right)$$

definieren.

Es stellt sich dabei heraus, daß

$$\bar{x} = \left(\prod_{i=1}^m x_i \right)^{1/m}$$

das geometrische Mittel ist. Mit der Varianz erhält man als Streuungsmaß $\bar{s}^2$

$$\begin{aligned} \bar{s}^2 &= \exp \left(\frac{1}{m} \sum_{i=1}^m \left(\log x_i - \frac{1}{m} \sum_{j=1}^m \log x_j \right)^2 \right) \\ &= \left(\prod_{i=1}^m \left(\frac{x_i}{\bar{x}} \right) \log \frac{x_i}{\bar{x}} \right)^{1/m} \end{aligned}$$

$\bar{s}^2$ ist ein Streuungsmaß, für das $\bar{x}$ die Minimalitätseigenschaft besitzt (so wie $\bar{x}$ für s^2 ; vgl. [4] S. 20).

Mißt man beispielsweise durchschnittliches Wachstum durch das geometrische Mittel der entsprechenden Wachstumsfaktoren, so bietet es sich an, die „Stetigkeit“

bzw. Ungleichmäßigkeit des Wachstums durch $\bar{s}^2$ zu messen.

Notes

- (Merkmals-) Vektoren aus $\mathbb{R}^n$ werden in dieser Arbeit der Übersichtlichkeit wegen immer als Zeilenvektoren dargestellt; das führt dazu, daß auftretende Matrixprodukte (z.B. $D \cdot x$) nicht ganz korrekt definiert sind.
- Beispielsweise ist $A((0, 1, 4, 0, 3, 1)) = 2, 3, 5, 6$
- $\mathcal{P} = \{P_\Pi: \mathbb{R}^n \rightarrow \mathbb{R}^n; P_\Pi(x_1, \dots, x_n) = (x_{\Pi(1)}, \dots, x_{\Pi(n)}), \Pi \text{ Permutation von } \{1, \dots, n\}\}$

References

- (1) Bock, H. H.: Automatische Klassifikation. Göttingen 1974
- (2) Dobbener, R.: Zur Invarianz von Klassifikationsverfahren. Diskussionspapier des Lehrstuhls für Statistik, Universität Bamberg 1980
- (3) Vogel, F.: Subjektivitäten bei der Klassifikation von Einheiten. In Proceedings in Operations Research, 1978. Tagung in Kiel 1977 p. 105–129
- (4) Vogel, F.: Beschreibende und Schließende Statistik. München-Wien 1979.

Dipl. Math. R. Dobbener
Universität Bamberg, Statistik
Feldkirchenstr. 21
D-8600 Bamberg

Research and Development in Information Retrieval, Berlin 1982

At the Technical University of West-Berlin a conference on the theme mentioned will be held from May 18–20, 1982 sponsored by the Gesellschaft für Informatik in cooperation with ACM and BCS. Its theme will be basic research ideas in information storage and retrieval as well as interesting new applications. Papers covering the following or other relevant topics are invited:

Methodology, Theory and Applications in the fields of System Modelling Evaluation Problems, Natural Language Processing, Automatic Indexing, Query Languages, Complexity Problems, File Organization, Data Models, Integration of IRS into DBMS, New Technology.

Contributors should submit a four page synopsis and an abstract by Oct. 15, 1981. Chairmen of the conference are Prof. H.-J. Schneider, TU Berlin and Prof. G. Salton, Cornell University. Intending contributors or those requiring further details should write to Dr. Peter Bollmann, TU Berlin, Kurfürstendamm 202, V, D-1000 Berlin 15.