

Proximitätsmaße

Strategieoptionen unter Missing Values

Dipl.-Wirtsch.-Ing. Matthias Runte
Universität Kiel, Lehrstuhl für Marketing
Westring 425, 24098 Kiel
Tel 0431/880-1535
Email: matthias@runte.de
URL: <http://www.runte.de/matthias>

Inhaltsverzeichnis

<i>Tabellenverzeichnis</i>	<i>III</i>
<i>Abbildungsverzeichnis</i>	<i>III</i>
1 Übersicht	1
2 Proximitätsmaße	1
2.1 Eigenschaften von Proximitätsmaßen.....	1
2.2 Einteilung von Proximitätsmaßen	2
3 Proximitätsmaße unter Missing Values	3
3.1 Missing Values.....	3
3.2 Aufbereitung der Datenbasis	4
3.3 Unterschiedliche Überlappungszahlen	5
4 Vorbereitende Strategien zur Proximitätsberechnung	6
4.1 Pairwise Available Case Elimination.....	6
4.2 Pairwise Imputation.....	7
4.3 Complete Imputation	8
5 Korrigierte Proximitätsmaße unter Missing Values	9
5.1 Korrigierte Minkowski-Metrik.....	9
5.2 Korrigierter Q-Korrelationskoeffizient.....	12
5.3 Korrigierte abgeleitete Proximitätsmaße	14
6 Zusammenfassung	15
<i>Quellenverzeichnis</i>	<i>16</i>
<i>Anhang</i>	<i>i</i>
A Eigenschaften von Proximitätsmaßen	i
A.1 Direkte und indirekte Erhebung von Proximitäten	i
A.2 Transformationen von Ähnlichkeiten und Distanzen	i
A.3 Skalenniveau der Merkmale	iii
A.4 Mittelwertzentrierung und Standardisierung	iii
A.5 Skaleninvarianz und Translationsinvarianz.....	iv
A.6 Allgemeine Minkowski-Metrik	v
A.7 Korrelationskoeffizienten.....	vi
A.8 Mahalanobis-Distanz.....	vii
A.9 Zur Verzerrung von Proximitätsfunktionen unter MV	viii
B Abbildungen	xi

Tabellenverzeichnis

<i>Tabelle 1 – Strategieoptionen zur Aufbereitung der Datenbasis: Beispiel</i>	<i>5</i>
<i>Tabelle 2 – Korrekturterme K für L_q-Metriken</i>	<i>12</i>
<i>Tabelle 3 - Spezielle L_q-Distanzmaße</i>	<i>vi</i>

Abbildungsverzeichnis

<i>Abbildung 1 – Korrekturterme für L_q-Metriken ($q=2$).....</i>	<i>xi</i>
<i>Abbildung 2 – Relative Erwartungswerte für korrigierte L_q-Metriken ($q=2$)</i>	<i>xi</i>
<i>Abbildung 3 – Korrekturterme K für ein abgeleitetes Ähnlichkeitsmaß.....</i>	<i>xii</i>
<i>Abbildung 4 - Relative Erwartungswerte für ein korrigiertes Ähnlichkeitsmaß</i>	<i>xii</i>

Abkürzungsverzeichnis

MV Missing Value(s)

Symbolverzeichnis

d_{ij}	<i>Distanz zwischen Objekt i und j</i>
f_m	<i>Partialfunktion</i>
K	<i>Korrekturterm für Distanz- und Ähnlichkeitsmaße</i>
M	<i>Anzahl Merkmale</i>
N	<i>Anzahl Objekte</i>
q	<i>Exponent der Minkowski-q-Metrik</i>
r_{ij}	<i>Korrelation zwischen den Vektoren i und j</i>
s_{ij}	<i>Ähnlichkeit zwischen Objekt i und j</i>
X	<i>Datenmatrix</i>
x_i	<i>Ausprägungsvektor eines Objektes</i>
x_{im}	<i>Ausprägung des Merkmals m in Objekt i</i>

1 Übersicht

In der Betriebswirtschaftslehre und insbesondere im Marketing ist es oft notwendig, die Ähnlichkeit oder Verschiedenheit von Objekten zu bestimmen und zu quantifizieren. Die Bestimmung von Ähnlichkeiten kann unterschiedliche Gründe und Anwendungen haben. Beispielsweise seien die Clusteranalyse und neuerdings das Collaborative Filtering genannt.

Zur Ähnlichkeitsbestimmung kommt eine Reihe unterschiedlicher Proximitätsmaße in Frage. Deren Eigenschaften unterscheiden sich teilweise deutlich voneinander. In Abschnitt 2.1 wird daher zunächst auf einige wichtige Definitionen und Eigenschaften von Proximitätsmaßen eingegangen. Daran anschließend werden in Abschnitt 2.2 einige gebräuchliche Proximitätsmaße und ihre Eigenschaften dargestellt.

Oft liegen zur Bestimmung der Ähnlichkeit zwischen Objekten jedoch keine vollständigen Daten vor. Beim Vorliegen von *Missing Values* sind viele Proximitätsmaße nicht mehr uneingeschränkt anwendbar, ohne daß es zu Verzerrungen oder inhaltlichen Problemen kommt. Auf zu beachtende Besonderheiten unter Missing Values und mögliche Lösungsansätze wird in Abschnitt 3 ff. eingegangen.

2 Proximitätsmaße

2.1 Eigenschaften von Proximitätsmaßen

Für die nachfolgenden Betrachtungen ist es hilfreich, eine Menge $I = \{I_1, \dots, I_N\}$ von **Objekten** zu unterstellen, welche über eine Anzahl M von **Merkmalen** verfügt. Weiterhin unterstellen wir die Existenz einer Matrix X , deren Zeilen x_n die Objektvektoren und deren Spalten die Merkmale $m=1..M$ repräsentiert. X wird als **Datenmatrix** bezeichnet.

Proximitätsmaße messen die Ähnlichkeit oder Unähnlichkeit zwischen zwei Objekten, indem die Unterschiede in den Merkmalen untersucht werden. Man unterscheidet Ähnlichkeits- und Distanzmaße (auch Unähnlichkeitsmaße genannt, s. Bacher 1994, S. 198). Die Maße unterscheiden sich inhaltlich dadurch, daß der Wert des Ähnlichkeits- bzw. Distanzmaßes um so größer ist, je ähnlicher bzw. unähnlicher die Objekte sind. Zur Inhaltsvalidität von Meßinstrumenten s. Schnell/Hill/Esser (1998, S. 152).

Ähnlichkeitsmaße sind definiert als eine Funktion $s = s(x_i, x_j) = s_{ij}$ des Kreuzproduktes einer Menge $I = \{I_1, \dots, I_N\}$ von Objekten auf die Menge der reellen Zahlen, wobei $s_{ij} = s_{ji}$ und $s_{ij} \leq s_{ii}$ für $i, j = 1, \dots, N$ (Fahrmeier/Hamerle 1996, S. 440). Zusätzlich läßt sich fordern, daß $s_{ij} \geq 0$ und $s_{ii} = 1$. Dies ist aber nicht zwingend erforderlich.

Distanzmaße hingegen sind definiert als eine Funktion $d = d(x_i, x_j)$ des Kreuzproduktes einer Menge $I = \{I_1, \dots, I_N\}$ von Objekten auf die Menge der reellen Zahlen \mathfrak{R} , wobei $d_{ii} = 0$, $d_{ij} = d_{ji}$ und $d_{ij} \geq 0$ für jeweils $i, j = 1, \dots, N$ (a.a.O.).

Metrische Distanzmaße bilden eine Untermenge der Distanzmaße. Sie sind definiert als Distanzmaße, welche die Eigenschaft $d_{ij} \leq d_{ik} + d_{jk}$ erfüllen. Sie ermöglichen unter bestimmten Umständen eine räumliche Vorstellung der Distanz.

In dieser Arbeit werden nur *indirekt berechnete* Proximitäten betrachtet, bei denen die Distanzen bzw. Ähnlichkeiten aus der Datenmatrix berechnet werden. Zur *direkten Erhebung* von Proximitäten s. Anhang A.1.

Distanz- und Ähnlichkeitsmaße können ineinander überführt werden. Gängige Transformationen sind im Anhang A.2 aufgeführt.

Diese Arbeit bezieht sich ausschließlich auf quantitative Merkmale. Für Distanzmaße reicht in der Regel die Unterstellung von intervallskalierten Merkmalen aus, für Korrelationskoeffizienten ist jedoch die Verhältnisskalierung erforderlich. Generell können die Merkmale aber auch ein anderes Skalenniveau besitzen (Anhang A.3).

Über weitere grundlegende Begriffe wie Mittelwertzentrierung, Standardisierung sowie Skalen- und Translationsinvarianz geben Anhang A.4 und A.5 Auskunft.

2.2 Einteilung von Proximitätsmaßen

Nach Bacher (1994, S. 198 f.) lassen sich Ähnlichkeits- und Unähnlichkeitsmaße für Objekte mit metrischen Merkmalen in vier Gruppen einteilen:

1. Korrelationskoeffizienten als Ähnlichkeitsmaße, auch als Assoziations- oder Zusammenhangsmaße bezeichnet.
2. Distanzmaße, berechnet auf Basis der Minkowski- L_q -Metrik.
3. Aus 1. oder 2. durch monotone Transformation abgeleitete Proximitätsmaße.
4. Andere Proximitätsmaße für spezifische Fragestellungen und Meßniveaus.

Diese Gruppierung findet sich in ähnlicher Form auch in anderen Quellen (Backhaus 1994, S. 264). Sie ist geprägt durch die Benennung der beiden **Grundmaße** (Korrelationsmaße und L_q -Distanzen) und bietet eine übersichtliche und klare Strukturierung unterschiedlicher Proximitätsmaße. Aus diesem Grunde orientiert sich die Darstellung von Proximitätsmaßen in dieser Arbeit ebenfalls im wesentlichen an dieser Unterteilung.

Die Grundmaße werden im Anhang A.6 und A.7 kurz dargestellt. Um die Grundmaße an spezielle Anforderungen und Situationen anzupassen, lassen sich beliebige Variationen dieser Grundmaße durch monotone Transformationen entwickeln. In der Literatur findet sich eine Vielzahl unterschiedlicher Varianten (Gower 1985). Zu den abgeleiteten Proximitätsmaße gehört z.B. die Klasse der teildifferenzbasierten Ähnlichkeitsmaße (s. Abschnitt 5.3).

Eine weitere Klasse der abgeleiteten metrischen Distanzfunktionen bilden die quadratischen Distanzfunktionen (Steinhauser/Langer 1997, S. 61). Den prominentesten Vertreter dieser Proximitätsmaße bildet die Mahalanobis-Distanz, auf welche an dieser Stelle nicht weiter eingegangen werden kann. Die Definition und Beschreibung einiger Eigenschaften finden sich im Anhang A.6.

3 Proximitätsmaße unter Missing Values

3.1 Missing Values

Wie in den vorangehenden Abschnitten dargestellt, messen Proximitätsmaße die Ähnlichkeit oder Unähnlichkeit zwischen zwei Vektoren der Datenmatrix X . Je nach Anwendung und Datenbasis kann es nun vorkommen, daß X nicht vollständig ist. Die fehlenden Daten werden als *Missing Values* (MV) oder *Lücken* bezeichnet.

Der Anteil fehlender Werte in der Datenmatrix kann je nach Anwendung stark variieren. In den meisten Anwendungen treten Missing Values nur vereinzelt auf (Santos 1981, S. 17, in: Schnell 1986, S. 7). Hingegen ist in einigen Verfahren die Nicht-Existenz von Werten sogar die Regel, wie beispielsweise dem Collaborative Filtering (Balabanovic/Shoham 1997, Konstan et al. 1997) oder vielen Data-Mining-Anwendungen. Da in vielen der von MV betroffenen Verfahren auch Proximitätsmaße eine Rolle spielen, ist es notwendig, sich mit den Auswirkungen von MV auf Proximitätsmaße zu beschäftigen.

Zur allgemeinen Behandlung von Missing Values läßt sich eine Vielzahl von Ansätzen und Verfahren unterscheiden (Runte 1999). Die in dieser Arbeit verfolgte Einteilung der Proxi-

mitätsmessung unter Missing Values basiert auf einem zweistufigen Verfahren. Es berücksichtigt dabei die wesentlichen aus der Literatur bekannten MV-Strategien (Bankhofer 1995). Die hier vorgeschlagenen Verfahren der Proximitätsberechnung unter Missing Values lassen sich jeweils in zwei Schritte unterteilen:

1. Festlegung einer Strategie zur Aufbereitung der Datenmatrix X .
2. Korrektur des Proximitätsmaßes zum Ausgleich unterschiedlicher Überlappungszahlen.

Die Schritte 1 und 2 werden in den beiden nachfolgenden Abschnitten kurz anhand eines Beispiels erläutert.

3.2 Aufbereitung der Datenbasis

Die beiden wichtigsten Verfahrensklassen zur Behandlung von MV bilden die **Eliminierungsverfahren** und **Imputationsverfahren**. Sie bilden die Basis der in dieser Arbeit verfolgten Aufbereitungs-Strategien der Datenmatrix. Aus den verfolgten Strategien leitet sich vor allen Dingen ab, *wieviele* und *welche* Merkmale in die anschließende Proximitätsberechnung einfließen.

Wir unterscheiden drei Aufbereitungsverfahren:

1. Pairwise available case analysis
2. Pairwise imputation analysis
3. Complete imputation analysis

Tabelle 1 verdeutlicht die Auswirkungen der unterschiedlichen Aufbereitungsverfahren auf die Datenmatrix bzw. die Objektvektoren x_i und x_j . Die beiden Ausgangsvektoren umfassen fünf Merkmale mit jeweils zwei Missing Values, welche als Punkt (●) dargestellt sind. Das erste (5;6) und letzte (2;4) Merkmal liegen in beiden Vektoren paarweise vor. Wir bezeichnen dies im folgenden als **Überlappung**.

Bei der **pairwise available case analysis** werden nur die jeweils *paarweise verfügbaren* Werte der Datenmatrix zur Proximitätsberechnung herangezogen. Bei diesem Verfahren handelt es sich also um ein Eliminierungsverfahren.

Bei der **pairwise imputation analysis** werden fehlende Werte vor der Proximitätsberechnung imputiert (dargestellt als „i“), falls die Merkmalsausprägung im anderen Vektor vorliegt (Merkmale 2 in x_1 , Merkmal 4 in x_2). Paarweise fehlende Werte werden bei der Proximität jedoch nicht berücksichtigt (Merkmal 3).

Bei der **complete imputation analysis** werden zunächst vollständige Vektoren erzeugt, bevor die Proximitätsberechnung durchgeführt wird. Damit werden alle Merkmale berücksichtigt, auch wenn beide Merkmalswerte fehlen.

In Abschnitt 4 werden die unterschiedlichen Strategien im Detail behandelt.

Ausgangsvektoren	Strategieoption zur Aufbereitung der Datenbasis					
	Pairwise Available		Pairwise Imputation		Complete Imputation	
$x_1' = \begin{pmatrix} 5 \\ \bullet \\ \bullet \\ 1 \\ 2 \end{pmatrix}$ $x_2' = \begin{pmatrix} 6 \\ 7 \\ \bullet \\ \bullet \\ 4 \end{pmatrix}$	$\begin{pmatrix} 5 \\ \bullet \\ \bullet \\ \bullet \\ 2 \end{pmatrix}$	$\begin{pmatrix} 6 \\ \bullet \\ \bullet \\ \bullet \\ 4 \end{pmatrix}$	$\begin{pmatrix} 5 \\ i \\ \bullet \\ 1 \\ 2 \end{pmatrix}$	$\begin{pmatrix} 6 \\ 7 \\ \bullet \\ i \\ 4 \end{pmatrix}$	$\begin{pmatrix} 5 \\ i \\ i \\ 1 \\ 2 \end{pmatrix}$	$\begin{pmatrix} 6 \\ 7 \\ i \\ i \\ 4 \end{pmatrix}$

Tabelle 1 – Strategieoptionen zur Aufbereitung der Datenbasis: Beispiel

3.3 Unterschiedliche Überlappungszahlen

Prinzipiell lassen sich unter Missing Values ähnliche Proximitätsmaße wie für vollständige Datenmatrizen verwenden. Dabei sind jedoch einige Besonderheiten zu beachten, wie die folgenden Überlegungen zeigen.

In alle Proximitätsfunktionen fließen die Merkmalsvektoren x_i und x_j der Objekte I_i und I_j ein. Die Berechnung der Proximität läßt sich in der Regel durch Kombinationen von **Partialfunktionen** $f_m(x_{im}, x_{jm})$ darstellen. In f_m geht das Ausprägungspaar $(x_{im}; x_{jm})$ des Merkmals m der Vektoren x_i und x_j ein. Distanz- und Ähnlichkeitsfunktionen lassen sich damit in der Regel darstellen als $d = d(f_1, f_2, \dots, f_M)$ bzw. $s = s(f_1, f_2, \dots, f_M)$.

Beispielsweise wäre im Falle der L_q -Metrik f_m definiert als $f_m = |x_i - x_j|$. Für die euklidische Distanz ergäbe sich

$$d_2 = \sqrt{\sum_{m=1}^M f_m^2} \quad \text{mit} \quad f_m = |x_i - x_j|.$$

Betrachten wir nun den Einfluß von Missing Values auf die Partialfunktion. Offensichtlich läßt sich ein Teil der Partialfunktionen f_m nicht mehr berechnen, da die unabhängigen Variablen x_{im} und x_{jm} nicht oder nicht paarweise zur Verfügung stehen. Das Merkmal m muß dann bei der Berechnung der Proximität *eliminiert*, oder die fehlenden Werte müssen *imputiert*

werden. Im Falle der Eliminierung oder der unvollständigen Imputation (d.h. nicht alle MV werden imputiert) verkleinert sich die Datenbasis zur Proximitätsberechnung im Vergleich zum Fall ohne Missing Values.

Da Missing Values in der Regel unregelmäßig über die Datenmatrix verteilt sind, ist die Anzahl fehlender Werte und damit der Überlappungen zwischen unterschiedlichen Objektpaaren meist nicht konstant. Je nach Proximitätsmaß kann die Anzahl der Überlappungen jedoch ein Faktor sein, der die Proximität stark beeinflusst. Die Überlappungsanzahl ist daher bei der Berechnung je nach Proximitätsmaß zu berücksichtigen (s. Abschnitt 5).

4 Vorbereitende Strategien zur Proximitätsberechnung

4.1 Pairwise Available Case Elimination

Eliminierungsverfahren lassen sich in die Verfahrensklassen **complete case analysis** und **available case analysis** unterteilen (Schwab 1991, S. 4). Für beide Klassen existieren die Varianten der **Merkmals-** und **Objekteliminierung** (Bankhofer/Praxmarer 1998, S. 91; Runte 1999).

Proximitätsmaße messen die Ähnlichkeit zwischen Objekten. Im Falle der Objekteliminierung würden alle Objektvektoren mit Missing Values aus der Analyse ausgeschlossen werden. Da dies offensichtlich nicht sinnvoll ist, scheidet demnach die Objekteliminierung prinzipiell aus. Es verbleibt die Merkmalseliminierung, also der Ausschluß einzelner Merkmale aus der Analyse.

Bei der **complete case analysis** mit Merkmalseliminierung wird das gesamte Merkmal aus der Datenmatrix eliminiert, sobald ein oder mehr Werte in diesem Merkmal fehlen. Das Verfahren kann damit erstens in Fällen von vergleichsweise wenigen Missing Values angewendet werden oder zweitens, wenn sich die Missing Values auf einige wenige Merkmale konzentrieren. Der wesentliche Nachteil der complete case analysis ist jedoch der teilweise hohe Informationsverlust durch die Eliminierung. Er vermindert das mögliche Anwendungsspektrum des Verfahrens in starkem Maße. Aus diesem Grunde wird im folgenden auf die complete case analysis nicht weiter eingegangen.

Bei der **available case analysis** mit Merkmalseliminierung werden alle existierenden Merkmalsausprägungen für die Analyse verwendet. Bei der hier untersuchten Proximitätsberechnung bedeutet dies, daß jeweils *paarweise vorhandene* Merkmalswerte bei der Proximitätsberechnung verwendet werden. Liegt für ein Merkmal nur ein Wert oder gar kein Wert vor, wird

dieses Merkmal eliminiert und steht nicht mehr für die Berechnung zur Verfügung (s. Tabelle 1, pairwise available analysis).

Eliminierungsverfahren zeichnen sich insbesondere durch einfache Handhabung aus. Es werden ausschließlich existierende Werte für die Proximitätsberechnung verwendet. Dadurch wird verhindert, daß sich ggf. Fehler von Imputationsschätzern in nachfolgenden Stufen der Proximitätsberechnung fortsetzen. Nachteilig wirkt sich jedoch der durch die Eliminierung eintretende Informationsverlust aus (Buck 1960). Bei einer Missing-Value-Wahrscheinlichkeit von jeweils 50% würden unter MCAR (Rubin 1976) nur bei 25% der Merkmale paarweise existierende Daten vorliegen. Die Hälfte der existierenden Daten würde eliminiert.

Weiterhin ist zu beachten, daß eine Eliminierung von Werten bei korrekter statistischer Betrachtung nur unter ganz bestimmten Voraussetzungen überhaupt zulässig ist. Prinzipiell kann davon ausgegangen werden, daß diese Voraussetzungen vorliegen, wenn für ein Merkmal die Ausprägungen der Merkmale einer gemeinsamen Verteilung entspringen, und das Fehlen der Werte weder von anderen Werten noch von den unbeobachteten Werten selbst abhängt. Die entspricht gerade einem Ausfallmechanismus mit MCAR.

Unter bestimmten statistischen Annahmen liegt ein weiteres Problem in der Verzerrung der Proximitätsfunktion. Diese Verzerrung kann insb. bei Eliminierung von Merkmalen auftreten, bei denen je ein Wert des Merkmals fehlt und existiert (s. Anhang A.9).

Die Überlappungsanzahl bestimmt die Anzahl der in die Berechnung einfließenden Werte. Typisch für Eliminierungsverfahren der available case-Klasse ist, daß die Überlappungsanzahl von Objektpaar zu Objektpaar aufgrund unregelmäßiger MV-Strukturen unterschiedlich hoch ist. Wie noch gezeigt werden wird, erschwert dies Eigenschaft insbesondere die Vergleichbarkeit der Proximitäten zwischen Objektpaaren. Die konkreten Auswirkungen hängen von der Art des gewählten Proximitätsmaßes ab und müssen im Einzelfall untersucht werden. Der Versuch einer solchen Analyse wird für einige spezielle Proximitätsmaße in Kapitel 5 unternommen.

4.2 Pairwise Imputation

Neben Eliminierungsverfahren können auch Imputationsverfahren als vorbereitende Maßnahme zur Proximitätsberechnung eingesetzt werden. Dabei werden fehlende Werte „ersetzt“ und in die Datenmatrix eingefügt. Auf diese Weise wird die Datenbasis zur Proximitätsberechnung erhöht, da unvollständige oder teilweise unvollständige Merkmalsausprägungen nicht eliminiert werden.

Die hier verfolgten Imputationsstrategien lassen sich einteilen in die Verfahren der paarweisen Imputation (*pairwise imputation*) und vollständigen Imputation (*complete imputation*).

Bei der **pairwise imputation** werden alle fehlenden Merkmalsausprägungen imputiert, für die ein Wert im jeweils anderen Vektor vorliegt (s. Tabelle 1). Der Vorteil dieses Vorgehens im Vergleich zu den Eliminierungsverfahren liegt in der Tatsache, daß keine existierenden Werte eliminiert werden. Die Anzahl der zur Berechnung verwendbaren Merkmale steigt.

In Anwendungen mit sehr wenigen Missing Values treten oft keine Merkmalspaare mit paarweise fehlenden Werten auf. Das Vorgehen der *pairwise imputation* führt in diesen Fällen bereits zu vollständig imputierten Vektoren.

Auch in Szenarien mit massivem Missing-Value-Aufkommen (z.B. Collaborative Filtering) hat das Verfahren Vorteile: Bei extremen Missing Value-Anteilen ist das Eliminierungsverfahren oft nicht mehr sinnvoll einsetzbar, weil Vektorenpaare vorliegen können, in denen keine einzige Überlappung zwischen den Vektoren existiert. Die *pairwise imputation* erzeugt hingegen paarweise vorliegende Merkmale bei allen Merkmalen, in denen mindestens eine Ausprägung existiert.

Einschränkend ist zu erwähnen, daß Imputationsverfahren nur unter bestimmten Bedingungen statistisch korrekt einsetzbar sind. So muß die Bedingung MCAR vorliegen oder ein Modell des Ausfallmechanismus bekannt sein.

Als Imputationsverfahren unter MCAR eignet sich beispielsweise die Mittelwertersetzung (Rubin 1987). In Abhängigkeit des Proximitätsmaßes müssen jedoch auch andere Imputationsschätzer in Betracht gezogen werden. Maßgeblich ist dabei, daß das durch die Imputation erzeugte Ergebnis zu keiner Verzerrung der Proximitätsfunktion führt. Hierauf wird weiter unten noch eingegangen.

4.3 Complete Imputation

Das Verfahren der **complete imputation** führt eine vollständige Imputation der beiden Vektoren durch. Dabei ist es im Vergleich zur *pairwise imputation* unerheblich, ob beide oder nur ein Wert eines Merkmals fehlen.

Der Vorteil des Verfahrens liegt insbesondere in einer konstanten Überlappungszahl. Die Berechnung basiert stets auf allen Merkmalen. Dies schafft die Vergleichbarkeit des Proximitätsmaßes zwischen unterschiedlichen Objektpaaren. Ein weiterer Vorteil ist, daß nach erfolgter Imputation alle Proximitätsmaße verwendet werden können, die auf vollständigen Daten basieren.

Das Verfahren ist jedoch problematisch bei hohen Ausfallraten der Daten. Die Proximitätsberechnung basiert in diesen Fällen auf imputierten Daten ohne signifikanten Informationsgehalt. Selbst wenn die notwendigen Imputations-Voraussetzungen (MCAR) vorliegen bzw. der Ausfallmechanismus richtig modelliert wurde, sind die errechneten Proximitäten unter massivem MV-Aufkommen vermutlich kaum noch aussagekräftig.

Folgende Variante könnte dem Verfahren jedoch möglicherweise zur Einsetzbarkeit auch bei hohen MV-Anteilen verhelfen. Bei dieser Variante findet vor oder nach der Imputation eine *objektweise Standardisierung* der Vektoren statt (s. Anhang A.4). Es werden also die Abweichungen der fehlenden Merkmalsausprägungen von den Merkmalsmittelwerten imputiert.

5 Korrigierte Proximitätsmaße unter Missing Values

Nachdem die Aufbereitung der Datenbasis durchgeführt wurde, liegt eine bestimmte Anzahl paarweise existierender Merkmale in je zwei Vektoren vor. Nun muß geprüft werden, ob für das zu berechnende Proximitätsmaß eine geeignete Korrektur durchzuführen ist, um ggf. unterschiedlichen Überlappungszahlen gerecht zu werden.

Diese Korrektur muß je nach Proximitätsmaß unterschiedlich vorgenommen werden. Wir beginnen mit der Korrektur von L_q -Metriken.

5.1 Korrigierte Minkowski-Metrik

Minkowski- L_q -Metriken bilden im Kern eine additive Verknüpfung der Ausprägungsdifferenzen der Merkmale, wobei eine Gewichtung der Differenzen anhand des Exponenten q vorgenommen wird. Bei $q > 1$ ergibt sich eine stärkere Gewichtung größerer Differenzen. Unter Verwendung der oben eingeführten Partialfunktionen f_m ergibt sich d_q als

$$d_q(x_i, x_j) = \left[\sum_{m=1..M} f_m^q \right]^{1/q} \quad \text{mit } f_m = |x_i - x_j|.$$

Wenn unkorrigierte L_q -Metriken auf Vektorenpaare mit unterschiedlicher Überlappungszahl angewendet werden, ergeben sich unerwünschte Effekte. Hierzu ein einfaches Beispiel. Für die Vektoren $x_1 = (2 \quad 2)'$ und $x_2 = (3 \quad 4)'$ ergibt sich eine L_2 -Distanz von

$$d_2(x_1, x_2) = \sqrt{(2-3)^2 + (2-4)^2} = \sqrt{5} \approx 2,236.$$

Wir vergleichen die Proximität dieser Vektoren nun mit der Distanz zweier erweiterter Vektoren $\tilde{x}_1 = (2 \ 2 \ 2 \ 2)'$ und $\tilde{x}_2 = (3 \ 4 \ 2 \ 3)'$. Die beiden ersten Merkmale dieser Vektoren entsprechen den Werten von x_1 und x_2 . Die Partialfunktionen haben die Werte $f_1=1$ und $f_2=2$. Die dritten und vierten Merkmale besitzen Partialfunktionen von $f_3=0$ und $f_4=1$, also im Mittel weniger als f_1 und f_2 . Auf den ersten Blick würde man daher eine kleinere Distanz zwischen den erweiterten Vektoren vermuten.

Die euklidische Distanz wächst jedoch auf

$$d_2(\tilde{x}_1, \tilde{x}_2) = \sqrt{(2-3)^2 + (2-4)^2 + (2-2)^2 + (2-3)^2} = \sqrt{6} \approx 2,449,$$

was inhaltlichen Validitätskriterien eines Distanzmaßes in der Regel nicht standhalten kann: Das Distanzmaß mißt nicht, was es eigentlich messen soll (Schnell/Hill/Esser 1988, S. 152).

Einen ersten Ansatz zur Lösung des Überlappungsproblems bietet die Normierung der Partialfunktionen über die Anzahl der Überlappungen:

$$d_q^m(x_i, x_j) = \left[\sum_{m=1..M} \left(\frac{f_m}{M} \right)^q \right]^{1/q} = \frac{1}{M} \left[\sum_{m=1..M} f_m^q \right]^{1/q} = \frac{1}{M} d_q(x_i, x_j)$$

Ein Kriterium für Proximitätsmaße mit unterschiedlichen Überlappungszahlen ist das der Erweiterbarkeit von Vektoren. Bei diesem Kriterium werden die Vektoren durch identische Werte „verdoppelt“. So wird beispielsweise aus x_1 und x_2 das Vektorenpaar $\tilde{x}_1 = (x_1 \ x_1)' = (2 \ 2 \ 2 \ 2)'$ und $\tilde{x}_2 = (3 \ 4 \ 3 \ 4)'$. Ändert sich die Proximität zwischen den Vektoren nicht, so gilt das **Erweiterbarkeitskriterium**:

$$d(x_i, x_j) = d(\tilde{x}_i, \tilde{x}_j) = d((x_i, x_i), (x_j, x_j))$$

Das Erweiterbarkeitskriterium bedeutet, daß der Erwartungswert der Proximität zwischen Objekten unter konstanten stochastischen Bedingungen *unabhängig von der Überlappungszahl* ist. Einfluß auf die Proximität haben also nur die vorhandenen Merkmalsausprägungen.

Das Erweiterbarkeitskriterium gilt für $d_q(x_i, x_j)$ und $d_q^m(x_i, x_j)$ nicht, wie sich jeweils exemplarisch zeigen läßt. Um dieses Kriterium zu gewährleisten, läßt sich speziell für L_q -Metriken folgende Korrektur durchführen. Die Korrektur fügt eine Mittelwertbildung über die Summe der mit q potenzierten Teildistanzen ein:

$$d_q^e(x_i, x_j) = \left[\frac{1}{M} \sum_{m=1..M} |x_{im} - x_{jm}|^q \right]^{1/q} = \left(\frac{1}{M} \right)^{1/q} \left[\sum_{m=1..M} |x_{im} - x_{jm}|^q \right]^{1/q} = \left(\frac{1}{M} \right)^{1/q} d_q(x_i, x_j)$$

Die Gültigkeit des Erweiterbarkeitskriteriums für d_q^e läßt sich durch Einsetzen erweiterter Vektoren in die Definition von d_q^e zeigen.

In einigen Applikationen kann die Überlappungszahl selbst Anhaltspunkte für die Proximität von Vektoren liefern. Inhaltlich könnte dabei unterstellt werden, daß eine hohe Überlappungszahl der Vektoren bereits auf eine hohe Ähnlichkeit der Vektoren schließen läßt. Beispielsweise läßt sich hier das Merkmal „Durchschnittsalter der Kinder“ in einer demografischen Erhebung anführen. Bei kinderlosen Befragten ist die Ausprägung dieses Merkmals „Missing“. Die Existenz des Merkmalswertes läßt hingegen auf Kinder schließen. Existiert bei zwei Befragten das Merkmal paarweise, so könnte inhaltlich bereits eine bestimmte „Ähnlichkeit“ unterstellt werden, da beide Befragten ein oder mehrere Kinder haben. Gleiches gilt bei kinderlosen Probanden im Falle paarweiser Missing Values.

Analog gilt dies in Recommender-Systemen (Resnick/Varian 1997), bei denen die Befragten Objekte bewerten sollen, falls sie diese kennen. Allein die gleichzeitige Kenntnis eines Objektes kann bereits ein Indiz für „Ähnlichkeit“ sein. Dies gilt in verstärktem Maße, wenn bei einer Vielzahl von Merkmalen nur vergleichsweise wenig existierende Werte vorliegen, wie dies beispielsweise beim Collaborative Filtering der Fall ist.

Um diesem Aspekt Rechnung zu tragen, läßt eine ähnliche Forderung wie beim Erweiterbarkeitskriterium stellen. Allerdings wird hier nicht die Gleichheit zwischen der Proximität der Ausgangsvektoren (x_1, x_2) und der erweiterten Vektoren $(\tilde{x}_1, \tilde{x}_2)$ gefordert, sondern eine geringere Distanz bzw. höhere Ähnlichkeit der erweiterten Vektoren. Wir bezeichnen diese Eigenschaft als **Überlappungskriterium**. Es gilt:

$$d(x_i, x_j) > d(\tilde{x}_i, \tilde{x}_j) \text{ mit } d(x_i, x_j) > 0, \tilde{x}_i = (x_i, x_i) \text{ und } \tilde{x}_j = (x_j, x_j).$$

Das Überlappungskriterium bedeutet statistisch, daß unter gleichen stochastischen Bedingungen der Erwartungswert der Distanz mit zunehmender Überlappungszahl abnimmt.

Für d_q gilt das Kriterium nicht. Für d_q^e gilt wie oben beschrieben das Erweiterbarkeitskriterium, welches offensichtlich die Gültigkeit des Überlappungskriteriums ausschließt. d_q^m hingegen hält für $q > 1$ dem Überlappungskriterium stand.

Wir halten folgendes Ergebnis fest: Die unkorrigierte L_q -Metrik ist für unterschiedlichen Überlappungszahlen meist ungeeignet und entspricht nicht sinnlogischen Forderungen. Im Einzelfall ist zu überprüfen, ob die Überlappungszahl selbst Aussagegehalt über die Proximität von Objekten liefern kann. Ist dies der Fall, so eignet sich beispielsweise d_q^m als korrigiertes Proximitätsmaß, für welches das Überlappungskriterium gilt. Ein vom Einfluß der Überlappungszahl unabhängiges Proximitätsmaß ist d_q^e , für welches das Erweiterbarkeitskriterium gilt.

Die bislang behandelten korrigierten L_q -Metriken lassen sich allgemein als $d_q^k(x_i, x_j) = K \cdot d_q(x_i, x_j)$ darstellen, wobei K die Bedeutung eines Korrekturterms zur Kompensation unterschiedlicher Überlappungsanzahl einnimmt (s. Tabelle 2).

Distanzmaß	Korrekturterm K	Erfülltes Kriterium
d_q	1	-
d_q^m	M^{-1}	Erweiterbarkeitskriterium
d_q^e	$M^{-1/q}$	Überlappungskriterium

Tabelle 2 – Korrekturterme K für L_q -Metriken

Abbildung 1 und Abbildung 2 (s. Anhang B) stellen den Verlauf von K bzw. den Erwartungswert der betrachteten (korrigierten) Distanzmaße d_q , d_q^m und d_q^e in Abhängigkeit der Überlappungszahl M dar. Für den Erwartungswert wird dabei unterstellt, daß alle Merkmale eine identische Wahrscheinlichkeitsverteilung haben und nicht korreliert sind.

5.2 Korrigierter Q-Korrelationskoeffizient

Der Q-Korrelationskoeffizient mißt die lineare Abhängigkeit zwischen zwei Vektoren. Bezeichne (ij) die Menge der paarweise existierenden Merkmale. Im einfachsten Falle werden alle paarweise existierenden Merkmale $m \in (ij)$ zur Berechnung des Korrelationskoeffizienten herangezogen:

$$r_{ij}^{(ij)} = \frac{s_{ij}^{(ij)}}{s_i^{(ij)} \cdot s_j^{(ij)}} = \frac{\sum_{m \in (ij)} (x_{im} - \bar{x}_i^{(ij)}) \cdot (x_{jm} - \bar{x}_j^{(ij)})}{s_i^{(ij)} \cdot s_j^{(ij)}}$$

Dabei sind $s_i^{(ij)}$ und $s_j^{(ij)}$ die Standardabweichung der Vektoren x_i und x_j , jeweils über die Merkmale (ij) , sowie M_{ij} als Anzahl der Merkmale in (ij) .

Der Erwartungswert des Korrelationskoeffizienten ist unter der restriktiven Bedingung MCAR unabhängig von der Überlappungsanzahl. Damit ist er im Vergleich zu den oben behandelten Proximitätsmaßen vergleichsweise einfach unter Missing Values zu behandeln, indem man wie dargestellt die jeweils paarweise verfügbaren Merkmale zur Korrelationsberechnung heranzieht.

Dabei ist allerdings nach den verwendeten Datenaufbereitungsverfahren zu unterscheiden. Im Falle der Verwendung von Imputationsverfahren ist die dabei verwendete Datenbasis frei von Informationsverlust, da einzeln fehlende Werte durch entsprechende Imputationen eingefügt werden. Dies ist im Falle der pairwise available analysis jedoch nicht der Fall. Hierbei gehen die Informationen über die einzeln verfügbaren Merkmale verloren, die nicht in die Korrelationsberechnung miteinbezogen werden (Buck 1960).

Abhilfe kann hierbei die Einbeziehung aller jeweils verfügbaren Werte pro Merkmal bei der Berechnung der Varianzen helfen. Statt $s_i^{(ij)}$ und $s_j^{(ij)}$ werden aus diesem Grunde $s_i^{(i)}$ und $s_j^{(j)}$ zur Berechnung des folgenden Korrelationskoeffizienten verwendet (Little/Rubin 1987, S. 42):

$$r_{ij}^* = \frac{s_{ij}^{(ij)}}{s_i^{(i)} \cdot s_j^{(j)}} \quad \text{mit} \quad s_i^{(i)} = \frac{\sum_{m \in (i)} (x_{im} - \bar{x}_i^{(i)})^2}{M_i} \quad \text{und} \quad s_j^{(j)} \quad \text{analog.}$$

wobei M_i und M_j die Anzahl existierender Merkmale der Objekte i und j sind, und $\bar{x}_i^{(i)}$ bzw. $\bar{x}_j^{(j)}$ die entsprechenden Mittelwerte über die existierenden Merkmale der Vektoren x_i bzw. x_j .

Weitere Varianten sind möglich durch die Verwendung aller verfügbaren Werte (i) bzw. (j) bei der Berechnung der Erwartungswerte $\bar{x}_i^{(i)}$ bzw. $\bar{x}_j^{(j)}$ statt lediglich der paarweise existierenden Werte (ij) mit der Berechnung von $\bar{x}_i^{(ij)}$ bzw. $\bar{x}_j^{(ij)}$.

r_{ij}^* und die nachfolgend skizzierten Varianten versuchen, den durch das Eliminierungsverfahren eintretenden Informationsverlust auszugleichen. Die dabei ermittelten univariaten Varianzen und Kovarianzen sind unter MCAR erwartungstreu (Little/Rubin 1987, S. 43).

Die paarweise Betrachtung verfügt jedoch auch über einige Nachteile, die ihre praktische Einsetzbarkeit nachhaltig beeinträchtigen können. Zu nennen sind hierbei unter anderem, daß bei

r_{ij}^* und den nachfolgenden skizzierten Varianten die Korrelationsmaße den Wertebereich von $[-1;1]$ verlassen können (Holm 1975, S. 162). Dieser Fall tritt vor allen Dingen bei hohen Korrelationen und geringen univariaten Varianzen im Vergleich zu den Varianzen der paarweise existierenden Werte auf.

5.3 Korrigierte abgeleitete Proximitätsmaße

Bei den (korrigierten) L_q -Metriken handelt es sich um Distanzmaße, also um Unähnlichkeitsmaße. Mit Korrelationskoeffizienten betrachtet man die lineare Abhängigkeit zwischen den Vektoren. Zusätzlich interessiert man sich jedoch auch oft für die Ähnlichkeit zwischen Objekten, wobei neben der linearen Abhängigkeit auch die absolute Differenz der Werte eine Rolle spielt. In der Praxis werden Distanzmaße in der Regel in entsprechende Ähnlichkeitsmaße transformiert (s. Anhang A.2).

Die nachträgliche Transformation kann jedoch durch die alleinige Verwendung von Ähnlichkeitsmaßen statt Distanzmaßen überflüssig gemacht werden. Im Unterschied zu den Distanzmaßen beinhalten die Partialfunktionen f_m hierbei nicht Teildistanzen, sondern **Teilähnlichkeiten** s_m . Durch Aggregation der Teilähnlichkeiten s_m erhält man das Ähnlichkeitsmaß.

Die Teilähnlichkeiten lassen sich auf unterschiedliche Art und Weise berechnen. Beispielsweise eignet sich in Anlehnung an die L_q -Metrik

$$s_m = \frac{1}{1 + (x_{im} - x_{jm})^q} \text{ oder allgemeiner } s_m = \frac{p}{p + (x_{im} - x_{jm})^q} \text{ mit } p > 0.$$

In diesen Ansätzen liegt für $x_{im}=x_{jm}$ das Maximum der Partialfunktion ($s_m=1$) vor, für große $(x_{im} - x_{jm})^q$ bzw. $(x_{im} - x_{jm})^q / p$ geht s_m gegen Null.

Die Aggregation der Teilähnlichkeiten kann analog zur L_q -Metrik unterschiedlich durchgeführt werden. Naheliegende Ansätze für Ähnlichkeitsmaße mit unterschiedlichen Überlappungszahlen bezeichnen wir in Analogie zu den oben dargestellten L_q -Metriken mit s^s und s^m .

Das Maß s^s summiert alle Teilähnlichkeiten auf. Es erfüllt das Überlappungskriterium, da die Proximität mit steigender Überlappungszahl monoton zunimmt:

$$s^s = \sum_{m=1}^M s_m$$

Das Maß s^m bildet den arithmetischen Mittelwert über die Teilähnlichkeiten s_m . Es erfüllt das Erweiterbarkeitskriterium, da der Erwartungswert der Proximität mit steigender Überlap-

pungszahl konstant bleibt (Unabhängigkeit der s_m und identische Verteilungen der Merkmale unterstellt):

$$s^m = \frac{1}{M} \sum_{m=1}^M s_m = K_m \cdot \sum_{m=1}^M s_m \quad \text{mit} \quad K_m = \frac{1}{M}$$

Weitere Ähnlichkeitsmaße sind je nach Anwendung und inhaltlichen Überlegungen denkbar. So ist es möglich, Korrekturfaktoren einzuführen, mit denen das lineare Wachstum des Erwartungswertes von s^s mit zunehmender Überlappungszahl gebremst wird. Hierzu würden sich beispielsweise Korrekturterme wie $K_w = \frac{1}{\sqrt{M}}$, $K_{hyp} = \frac{2M-1}{M^2}$ oder $K_{\log} = \frac{\ln(M+1)}{M}$ eignen (s. Abbildung 3, Anhang B). In diesen Fällen würde zwar das Überlappungskriterium gelten, aber das Wachstum des Erwartungswertes der Ähnlichkeitsfunktion degressiv statt linear ausfallen (s. Abbildung 4). Derartig korrigierte Ähnlichkeitsmaße könnten sich generell bei Methoden eignen, in denen die Überlappungszahl selbst insbesondere im Bereich geringer Überlappungszahlen deutliche Auswirkungen auf die Proximität hat, aber im Bereich höherer Überlappungszahlen weniger aussagekräftiger wird. Hierzu zählen möglicherweise Methoden wie das Collaborative Filtering.

6 Zusammenfassung

Missing Values können gravierende Auswirkungen auf die Unverzerrtheit und inhaltliche Bedeutung von Proximitätsmaßen besitzen. In dieser Arbeit wurden neben den aus der gängigen Literatur bekannten Eliminierungsverfahren weitere Verfahrensvorschläge zur Aufbereitung der Datenbasis und Korrektur von Proximitätsmaßen unter verschiedenen inhaltlichen Bedingungen und statistischen Annahmen gemacht. Dazu wurden die Partialfunktion f_m sowie das Erweiterbarkeits- und Überlappungskriterium eingeführt, mit deren Hilfe sich die Auswirkung von unterschiedlichen Überlappungszahlen auf den Erwartungswert von Proximitätsmaßen beschreiben läßt.

Ob die vorgeschlagenen Verfahren und Korrekturen die Proximitätsmessung in konkreten Anwendungen mit Missing Values verbessern können, ist nach Kenntnis des Autors noch nicht untersucht worden. In diesem Bereich besteht noch Forschungsbedarf.

Quellenverzeichnis

- Bacher 1994: Clusteranalyse - Anwendungsorientierte Einführung, München, Wien, Oldenbourg
- Backhaus; Erichson; Plinke; Weiber (1994): Multivariate Analysemethoden, 7. Auflage, Berlin, Heidelberg, New York, Tokio
- Balabanovic, Marko; Shoham, Yoav (1997): Fab - Content-Based, Collaborative Recommendation, Communications of the ACM, Vol. 40, No. 3 (March), S. 66-72.
- Bankhofer, Udo (1995): Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse, Bergisch-Gladbach, Eul
- Bankhofer, Udo; Praxmarer, Sandra (1998): Zur Behandlung fehlender Daten in der Marktforschungspraxis, Marketing ZFP, Heft 2, 2. Quartal 1998
- Buck, S.F. (1960): A Method of Estimation of Missing Values in Multivariate Data suitable for use with an Electronic Computer, Journal of the Royal Statistical Society, Series B, S. 302-306
- Dixon, J.K. (1979): Pattern Recognition with Missing Data, IEEE Transactions on Systems, Man and Cybernetics, SMC9, S. 617-621
- Everitt, Brian S. (1993): Cluster Analysis, third Edition, London, Melbourne, Auckland, Wiley
- Gower, J.C. (1985): Measures of Similarity, Dissimilarity and Distance, in: Kotz, S.; Johnson N.L.; Read, C.B. (Eds.): Encyclopedia of Statistical Sciences, Volume 5, New York, Wiley
- Fahrmeir, Ludwig; Hamerle, Alfred; Tutz, Gerhard (1996): Multivariate statistische Verfahren, 2. Auflage, de Gruyter, Berlin, New York
- Hartung, Joachim; Elpelt, Bärbel (1995): Multivariate Statistik: Lehr- und Handbuch der angewandten Statistik, 5. Auflage, München, Wien, Oldenbourg
- Holm, Kurt (1975): Die Erstellung einer Korrelationsmatrix, in: Holm, Kurt (Hrsg.): Die Befragung, Band 2, München, UTB, S. 155-218
- Konstan, Joseph A.; Miller, Bradley N.; Maltz, David; Herlocker, Jonathan L.; Gordon, Lee R., Riedl, John (1997): Grouplens - Applying Collaborative Filtering to Usenet News, Communications of the ACM, Vol. 40, No. 3 (March), S. 77-87
- Little, Roderick J.A.; Rubin, Donald B. (1987): Statistical Analysis with Missing Data, New York, Wiley
- Resnick, Paul; Varian, Hal R. (1997): Recommender Systems, Communications of the ACM, Vol. 40, No. 3 (March), S. 56-58
- Rubin, Donald B. (1976): Inference and missing data, in: Biometrika, Vol. 63, S. 581-592
- Rubin, Donald B. (1987): Multiple Imputation for Nonresponse in Surveys, New York, Wiley

Runte, Matthias (1999): Missing Values – Konzepte und statistische Literatur, Kiel, <http://www.runte.de/matthias>

Schnell, Rainer (1986): Missing-Data-Probleme in der empirischen Sozialforschung, Bochum

Schnell, Rainer; Hill, Paul B.; Esser, Elker (1988): Methoden der empirischen Sozialforschung, München

Schwab, Georg (1991): Fehlende Werte in der angewandten Statistik, Wiesbaden, DUV

Späth (1977): Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion, 2. Auflage, München, Wien, Oldenbourg

Steinhausen/Langer 1977: Clusteranalyse - Einführung in Methoden und Verfahren der automatischen Klassifikation, Berlin, New York, de Gruyter

Anhang

A Eigenschaften von Proximitätsmaßen

A.1 Direkte und indirekte Erhebung von Proximitäten

Anstatt Ähnlichkeiten und Distanzen *indirekt* durch die Datenmatrix zu berechnen, können sie auch *direkt* erhoben werden. Bei der direkten Erhebung müssen die befragten Personen die Ähnlichkeit oder Distanz von Objektpaaren schätzen.

Die direkte Erhebung von Proximitäten ist in den meisten Fällen nicht oder nur unter Schwierigkeiten durchführbar. Selbst in den Fällen, in denen eine direkte Erhebung prinzipiell möglich wäre, ist die Konformität mit der Definition von Distanzen und Ähnlichkeiten meist nicht gewährleistet (Fahrmeier/Hamerle 1996, S. 441). Zu beachten ist dabei jeweils das Kriterium der Inhaltsvalidität (Schnell/Hill/Esser 1988, S. 152). Die Probleme zeigen sich zum Beispiel in Fällen, bei denen identischen Objekten eine Distanz von $d_{ii} > 0$ zugeordnet wird. Ausgesprochen unähnliche Objekte hingegen könnten eine sehr kleine Distanz erhalten. Außerdem ist die Reliabilität der erhaltenen Daten oft ungenügend.

Inkonsistenzen bezüglich der Definitionen können bei indirekt erhobenen Proximitätsmaßen nicht auftreten. Die Ähnlichkeits- und Distanzmaße werden entsprechend den Definitionen gewählt und gewährleisten damit entsprechend der Definition konsistente Proximitätsmaße. Dies erleichtert insbesondere die Weiterverwendung von Proximitätsmaßen in multivariaten Verfahren.

A.2 Transformationen von Ähnlichkeiten und Distanzen

Ähnlichkeitsmaße und Distanzmaße lassen sich gegenseitig transformieren. Die Transformation basiert auf der Interpretation von Ähnlichkeits- und Distanzmaßen. Je geringer die Distanz zwischen den Objekten ist, desto ähnlicher sind die Objekte.

Für die Transformation von Ähnlichkeitsmaßen in Distanzen eignet sich jede streng monoton fallende Funktion f mit $f(s_{ii}) = d_{ii} = 0$, wobei wegen $d_{ii} = 0$ und der Eindeutigkeit von f die Forderung $s_{ii} = s_{jj}$ für alle $i, j = 1, \dots, N$ zu stellen ist.

Für $s_{ii} = 1$ und $0 \leq s_{ij} \leq 1$ eignet sich insbesondere die Transformation

$$d_{ij} = 1 - s_{ij},$$

welche für d_{ij} einen Wertebereich von $0 \leq d_{ij} \leq 1$ festlegt.

Für Ähnlichkeitsmaße mit $s_{ii}=1$ und einem Wertebereich von $-1 \leq s_{ij} \leq 1$ ist nach Fahrmeier/Hamerle (1996, S. 442) die Transformation

$$d_{ij} = \sqrt{2(1-s_{ij})}$$

üblich.

Analog eignet sich für die Transformation von Distanzmaßen in Ähnlichkeitsmaße ebenfalls jede streng monotone fallende Funktion. Im einfachsten Falle berechnet man die transformierte Ähnlichkeitsfunktion als

$$s_{ij} = -d_{ij}.$$

Da allerdings $d_{ij} \geq 0$ gilt, wäre $s_{ij} \leq 0$. Damit eignet sich die Ähnlichkeitsfunktion für viele Anwendungen nicht, bei denen $s_{ij} \geq 0$ gefordert wird. Vorteilhafter ist aus diesem Grunde folgende Transformation:

$$s_{ij} = t - d_{ij} \text{ mit } t = \max\{d_{ij} \mid i, j = \{1..N\}, i \neq j\}$$

Die Funktion schließt negative Ähnlichkeitswerte aus und definiert für die höchste Distanz eine Ähnlichkeit von Null. Möglich ist auch eine Funktion, welche eine Normierung der Ähnlichkeiten über alle Objekte durchführt, so daß $0 \leq s_{ij} \leq 1$ folgt (Fahrmeier/Hamerle 1996, S. 442):

$$s_{ij} = 1 - \frac{d_{ij}}{t} \text{ mit } t = \max\{d_{ij} \mid i, j = \{1..N\}, i \neq j\}$$

Desweiteren kommt auch eine Transformation

$$s_{ij} = \frac{1}{d_{ij}}$$

in Frage. Hierbei ergibt sich jedoch eine Überbetonung sehr kleiner Distanzmaße. Für Distanzen mit $d_{ij}=0$ ist s_{ij} nicht definiert. Dieses Problem kann umgangen werden, indem man die Transformation anhand der folgenden Gleichung vornimmt:

$$s_{ij} = \frac{1}{1 + d_{ij}} \text{ oder allgemeiner } s_{ij} = \frac{p}{p + d_{ij}} \text{ mit } p > 0.$$

Für $d_{ij}=0$ ergibt sich $s_{ij}=1$, für alle $d_{ij}>0$ gilt $0 < s_{ij} < 1$.

A.3 Skalenniveau der Merkmale

Proximitätsmaße werden anhand der Datenmatrix X berechnet. Die Messung der Ausprägungen kann jedoch unterschiedlicher Natur sein. Proximitätsmaße lassen sich nach dem Skalenniveau der zur Berechnung verwendeten Merkmale klassifizieren. Man unterscheidet:

- Nominalskalierte binäre Merkmale
- Nominalskalierte mehrstufige (kategoriale) Merkmale
- Ordinalskalierte Merkmale
- Intervallskalierte Merkmale
- Verhältnisskalierte Merkmale

Nominalskalierte mehrstufige Merkmale lassen sich allgemein in binäre Merkmale überführen. Intervallskalierte und verhältnisskalierte Merkmale werden auch als quantitative oder kardinalskalierte Merkmale bezeichnet und zusammengefaßt. Die Verhältnisskala wird auch als metrische Skala bezeichnet.

Weiterhin lassen sich unabhängig vom Skalenniveau der Merkmale diskrete und stetige Ausprägungen unterscheiden (Fahrmeier/Hamerle 1996, S. 11):

Ein Merkmal heißt *diskret*, wenn es höchstens abzählbar viele Ausprägungen annehmen kann. Ein Merkmal heißt *stetig*, wenn mit jeweils zwei Ausprägungen auch jeder Zwischenwert zulässig ist. Stetige Merkmale können nicht nominalskaliert sein. Diskrete Merkmale können beliebig skaliert sein.

Viele als metrisch angenommene Merkmale sind in der Realität nur ordinalskaliert. So muß beispielsweise im Zuge einer Erhebung die Testperson oft einen Zahlenwert auf einer diskreten Rating-Skala festlegen. Den Stufen der Rating-Skala werden diskrete numerische Werten zugeordnet. Unterstellt man äquidistante Ausprägungen der Rating-Skala, ergibt sich eine Intervallskala.

A.4 Mittelwertzentrierung und Standardisierung

Der Vektor $x_i = (x_{i1}, x_{i2}, \dots, x_{im})'$ der Merkmalsausprägungen eines Objektes I_i verfügt über einen Mittelwert \bar{x}_i und eine Standardabweichung s_i . Der Mittelwert \bar{x}_i der Ratings wird auch

Profilhöhe, die Standardabweichung s_i wird **Profilstreuung** oder Profilstandardabweichung genannt (Bacher 1994, S. 191). Mit diesen Kennwerten lassen sich folgende Transformationen vornehmen:

Bei der **Mittelwertzentrierung** wird die Profilhöhe eines jeden Vektors normalisiert:

$$x_i' = x_i - \bar{x}_i$$

Inhaltlich bedeutet dies, daß nicht mehr die absolute Höhe der Merkmalsausprägungen untersucht wird, sondern lediglich die relativen Ausprägungen.

Bei der **Standardisierung** wird neben der Profilhöhe auch die Profilstreuung herausgerechnet:

$$x_i'' = \frac{x_i - \bar{x}_i}{s_i} = \frac{x_i'}{s_i}$$

Alle Ratingvektoren erhalten damit eine Profilhöhe von 0 und eine Standardabweichung von 1. Die Standardisierung bedeutet inhaltlich, daß nur noch die *relative Ausprägungshöhe*, bezogen auf alle Merkmale eines Vektors, untersucht wird.

A.5 Skaleninvarianz und Translationsinvarianz

Bei intervall- und verhältnisskalierten Merkmalen ist die Maßeinheit frei wählbar. Es läßt sich nun fordern, daß die Distanz bzw. Ähnlichkeit zweier Objekte von der Maßeinheit unabhängig sein soll. Man spricht in diesem Falle von Skaleninvarianz.

Ein Proximitätsmaß ist skaleninvariant, wenn die Transformation $\tilde{x}_n = C \cdot x_n$ mit $C = \text{diag}(c_1, \dots, c_M)$ und $c_i > 0$ ($i=1..M$) keinen Einfluß auf die Distanz hat, also $d(\tilde{x}_i, \tilde{x}_j) = d(x_i, x_j)$ gilt (Fahrmeir/Hamerle 1996, S. 448).¹

Die Forderung der Skaleninvarianz ist eine sehr strenge Forderung, die nur von wenigen Proximitätsmaßen erfüllt wird. Liegen für alle Merkmale gleiche Maßeinheiten vor, wird man aus diesem Grunde im allgemeinen lediglich fordern, daß $c_1 = \dots = c_M = c$ mit $c > 0$, und daß $d(\tilde{x}_i, \tilde{x}_j) = c \cdot d(x_i, x_j)$. Diese Bedingung wird beispielsweise von der allgemeinen Lq-Metrik erfüllt:

¹ Der Operator $\text{diag}(c_1, \dots, c_M)$ erzeugt eine Diagonalmatrix der Dimension $M \times M$ mit den Elementen c_1 bis c_M auf der Hauptdiagonalen und allen anderen Elementen gleich Null.

$$d_q(\tilde{x}_i, \tilde{x}_j) = \left[\sum_{m=1..M} |c \cdot x_{im} - c \cdot x_{jm}|^q \right]^{1/q} = c \cdot d_q(x_i, x_j)$$

Bei intervallskalierten Merkmalen ist die Wahl des Koordinatenursprungs frei wählbar und darf keinen Einfluß auf das Proximitätsmaß haben. So darf es bei intervallskalierten Merkmalen keine Rolle spielen, ob man eine Skala die Abweichung vom Nullpunkt (-3 = sehr schlecht, 0 = mittel, 3 = sehr gut) oder eine Skala mit ganzzahligen Werten von 1 bis 7 (z.B. 1 = sehr schlecht, 4 = mittel, 7 = sehr gut) verwendet wird. Liegt diese Eigenschaft vor, spricht man von Translationsinvarianz.

Formal bedeutet dies, daß die Transformation $\tilde{x}_1 = x_1 + b$ und $\tilde{x}_2 = x_2 + b$ keinen Einfluß auf die Distanz hat, also $d(\tilde{x}_1, \tilde{x}_2) = d(x_1, x_2)$ gilt (Fahrmeier/Hamerle 1996, S. 448).

Für quantitative Merkmale sind **alle** Proximitätsfunktionen translationsinvariant, in deren Partialfunktionen f_m (s. Abschnitt 3.3) die Differenz der Merkmalsausprägungen $x_1 - x_2$ berechnet wird. Dies ist leicht einzusehen wegen:

$$\tilde{x}_1 - \tilde{x}_2 = (x_1 + b) - (x_2 + b) = x_1 - x_2$$

Der Translationsparameter b wird also durch die Subtraktion der Merkmale eliminiert.

A.6 Allgemeine Minkowski-Metrik

Zu den am häufigsten verwendeten Proximitätsmaßen gehören Distanzmaße, die auf den verallgemeinerten L_q -Distanzen (Minkowski- q -Metriken) basieren (Hartung/Elpelt 1995, S. 72):

$$d_q(x_i, x_j) = \left[\sum_{m=1..M} |x_{im} - x_{jm}|^q \right]^{1/q}$$

Für q ist prinzipiell jeder Parameter mit $q > 0$ geeignet. Die Verwendung bestimmter Werte für q liefert die *speziellen L_q -Distanzmaße* (Tabelle 3). Der Wert des Parameters q führt zu einer mehr oder weniger starken Gewichtung der Merkmalsdifferenzen in Abhängigkeit der Merkmalsdifferenz selbst. Je größer q ist, desto stärker werden höhere Abweichungen gewichtet. Im Extremfall der Chebychev-Distanz ($q = \infty$) ist diese Gewichtung so stark, daß ausschließlich das Maximum der absoluten Merkmalsdifferenzen die Distanz zwischen I_i und I_j bestimmt, also $d_\infty(x_i, x_j) = \max_{m=1..M} |x_{im} - x_{jm}|$.

Bei der City-Block-Metrik ($q=1$) geht der Betrag jeder Einzeldifferenz $|x_{im} - x_{jm}|$ linear additiv in die Distanz ein. Man kann sich für den zweidimensionalen Fall eine Großstadt vorstellen

len, deren Straßen jeweils rechtwinklig zueinander laufen. Die Distanz zwischen zwei Koordinaten entspricht dann dem kürzesten Weg, den man auf dem rechtwinkligen Gitter von einem zum anderen Punkt zurücklegen würde.

Sehr häufig werden euklidische Distanzen ($q=2$) als Distanzmaß verwendet. Sie ermöglichen im Falle von drei Dimensionen eine räumliche Vorstellung der Distanz. Die euklidische Distanz gewichtet durch die Quadrierung der Differenzen die Merkmale mit größeren Abweichungen höher als geringe Abweichungen.

q=1	City-Block-Metrik
q=2	Euklidische Distanz
q= ∞	Chebychev-Distanz

Tabelle 3 - Spezielle L_q -Distanzmaße

Als Variante der L_q -Maße läßt sich vor der Berechnung eine Mittelwertzentrierung oder Standardisierung der Objektvektoren durchführen, so daß nur die relativen und normalisierten Abweichungen der Merkmale von den Mittelwerten zur Berechnung der Distanz herangezogen wird (Anhang A.4).

Zur Transformation der L_q -Distanzen in Ähnlichkeitsmaße können u.a. die im Anhang A.2 dargestellten Transformationen verwendet werden.

A.7 Korrelationskoeffizienten

Als Ähnlichkeitsmaß kommt auch der Pearsonsche oder Q-Korrelationskoeffizient r_{ij} in Betracht (Schwarze 1988, S. 147). r_{ij} mißt die Stärke des linearen Zusammenhangs zwischen zwei Vektoren

$$r_{ij} = \frac{s_{ij}}{s_i \cdot s_j} = \frac{\frac{1}{M} \sum_{m=1}^M (x_{im} \cdot x_{jm}) - \bar{x}_i \cdot \bar{x}_j}{s_i \cdot s_j}$$

mit \bar{x}_i, \bar{x}_j als Profilhöhe und s_i, s_j als Profilstreuung von x_i bzw. x_j (s. Anhang A.4). Der Korrelationskoeffizient führt eine implizite Standardisierung (Elimination der Profilhöhe und -streuung) der Vektoren durch. r_{ij} kann damit Werte zwischen -1 und 1 annehmen.

Sind keine negativen Werte erwünscht, läßt sich durch eine Transformation

$$\tilde{r}_{ij} = \frac{(r_{ij} + 1)}{2}$$

ein Wertebereich von $[0;1]$ erzielen.

Ist nur die *Stärke*, nicht aber die *Richtung* der stochastischen Abhängigkeit von Interesse, läßt sich diese durch eine Transformation

$$\tilde{r}_{ij} = r_{ij}^2 \text{ oder } \tilde{r}_{ij} = |r_{ij}|$$

bestimmen, welche ebenfalls zu einem Wertebereich von $[0;1]$ führt. Der Informationsgehalt über die Richtung der Abhängigkeit wird dabei jedoch eliminiert.

Die implizite Standardisierung des Korrelationskoeffizienten kann in einigen Anwendungen unerwünscht sein (Everitt 1993, S. 43). Dies wäre beispielsweise der Fall, wenn neben der linearen Abhängigkeit ein auf absoluten Größen basierendes Zusammenhangsmaß von Interesse ist. In diesen Fällen läßt sich statt r_{ij} auch die Kovarianz s_{ij} verwenden, welche keine implizite Standardisierung durchführt.

An dieser Stelle ist nochmals deutlich zu betonen, daß r_{ij} lediglich ein Maß für die *lineare Abhängigkeit* von zwei Vektoren ist. So sind die Vektoren $x_1 = (0 \ 0,1 \ 0,2)'$ und $x_2 = (10 \ 10,1 \ 10,2)'$ zwar vollständig linear abhängig ($r_{ij}=1$), könnten aber nach inhaltlichen Validitätskriterien ausgesprochen unähnlich sein. Trotz dieser Einschränkung können wir r_{ij} oder aus r_{ij} abgeleitete Transformationen für die folgenden Betrachtungen als mögliche Ähnlichkeitsmaße ansehen.

A.8 Mahalanobis-Distanz

Eine Klasse der metrischen Distanzfunktionen bilden die *quadratischen Distanzfunktionen* (Steinhauser/Langer 1997, S. 61). Dabei ist

$$d_B^2(x_i, x_j) = \left[(x_i - x_j)' B (x_i - x_j) \right]^{1/2},$$

wobei B eine positiv definite Matrix ist. Einen häufig verwendeten Fall der quadratischen Distanzfunktionen bildet die Mahalanobis-Distanz, bei der für B die Inverse der empirischen Kovarianzmatrix K^{-1} verwendet wird:

$$d_M(x_i, x_j) = \left[(x_i - x_j)' K^{-1} (x_i - x_j) \right]^{1/2} \text{ mit}$$

$$K = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})(x_n - \bar{x})' \text{ und } \bar{x} = \frac{1}{N} \sum_{n=1}^N x_n .$$

Der Vorteil der Mahalanobis-Distanz gegenüber der euklidischen Distanz liegt in der Möglichkeit zur Berücksichtigung von Korrelationen zwischen den Merkmalen (Hartung/Elpelt 1995, S. 74). Die Einbeziehung von stark korrelierende Merkmalen führt bei der euklidischen Distanz zu einer Überbetonung der hinter den korrelierenden Variablen stehenden Dimensionen. Die Mahalanobis-Distanz führt eine *korrelationsfreie* Berechnung der Distanz durch, wie von Steinhausen/Langer (1977, S.59 f.) gezeigt wird. Damit eignet sich die Mahalanobis-Distanz für Anwendungen, bei denen starke Merkmalskorrelationen auftreten, die Korrelationen aber selbst keine inhaltliche Bedeutung für die Proximität haben.

Die Mahalanobis-Distanz verfügt über eine Reihe von weiteren interessanten Eigenschaften. Sie ist invariant gegenüber beliebigen nichtsingulären linearen Transformationen. Damit ist sie das einzige in dieser Arbeit betrachtete Proximitätsmaß, welches neben der *Translationsinvarianz* auch die Eigenschaft der *Skaleninvarianz* besitzt. Der Beweis wird von Fahrmeir/Hamerle (1996, S. 450 f.) geführt.

Der bei der Berechnung der Mahalanobis-Distanz verwendete Mittelwertvektor \bar{x} setzt stillvoraus, daß es sich bei den auf Ähnlichkeit zu untersuchenden Merkmalsvektoren um Realisierungen ein und desselben multivariaten Zufallsvektors handelt. Diese Annahme ist oft nicht adäquat. Steinhausen/Langer (1977, S. 61) geben Lösungsempfehlungen für dieses Problem.

A.9 Zur Verzerrung von Proximitätsfunktionen unter MV

Missing Values können zu Verzerrung in nachfolgenden Analysen führen (Little/Rubin 1987). Dies gilt insbesondere für die Proximitätsberechnung. Zu untersuchen ist insbesondere die Auswirkung verschiedener Vorgehensweisen auf die Verzerrung des Erwartungswertes des Proximitätsmaßes. In der folgenden Betrachtung umfassen dabei die in dieser Arbeit betrachteten Eliminierungs- und Imputationsverfahren. Beim Eliminierungsverfahren werden unvollständige Merkmalspaare entfernt, beim Imputationsverfahren wird der fehlende Wert imputiert.

Zur Vereinfachung unterstellen wir die Gültigkeit des Erweiterbarkeitskriteriums (s. Abschnitt 5.1), Unabhängigkeit und Gleichverteilung der Merkmale sowie die Missing Value-Bedingung MCAR (Rubin 1976). Faßt man die Merkmalswerte als Zufallszahlen auf, folgt aus der Gleichverteilung der Merkmale, daß auch die Partialfunktionen f_m gleichverteilt sind und damit einen identischen Erwartungswert haben, d.h. $E(f_1) = E(f_2) = \dots = E(f_M) = E_f$.

Das Erweiterbarkeitskriteriums besagt, daß der Erwartungswert der Proximität unabhängig von der Überlappungsanzahl ist. Wir betrachten beispielhaft das Erweiterbarkeitskriterium für die korrigierte L_q -Metrik d_q^e :

$$d_q^e(x_1, x_2) = \left[\frac{1}{M} \sum_{m=1..M} f_m^q \right]^{1/q} \quad \text{mit } f_m = |x_{1m} - x_{2m}|.$$

Zur Berechnung von d_q^e unterstellen wir nun unterschiedliche Überlappungszahlen bzw. Merkmalsteilmengen $M_a, M_b \subset \{1, \dots, M\}$. Die sich ergebenden Distanzen werden als $d_q^{(a)}$ und $d_q^{(b)}$ bezeichnet. Dann gilt für die Erwartungswerte

$$E(d_q^{(a)}) = E \left[\frac{1}{|M_a|} \sum_{m \in M_a} f_m^q \right]^{1/q} = \left[\frac{1}{|M_a|} \sum_{m \in M_a} E(f_m^q) \right]^{1/q} = \left[\frac{M_a \cdot E_f^q}{M_a} \right]^{1/q} = E_f = E(d_q^{(b)})$$

und somit das Erweiterbarkeitskriterium.

Kommen wir nun zur Unverzerrtheit des Erwartungswertes unter Missing Values. Man betrachte das Beispiel aus Tabelle 1, S. 5. Zur Vereinfachung unterstellen wir einen reellen Wertebereich der Merkmale von $[1;7]$ mit Gleichverteilung. Missing Values werden als Zufallszahlen X_{im} interpretiert. Für die fehlenden Werte gilt $E(X_{im})=4$. Für den Erwartungswert der Partialfunktion E_f gilt

$$E_f = E(f_m) = E(|X_{1m} - X_{2m}|) = 2,$$

wie sich leicht zeigen läßt.

Betrachten wir nun das Merkmal im Beispiel, welches die Ausprägung $(x_{21}, x_{22})=(\bullet;7)$ besitzt. Mit der Kenntnis des Wertes $x_{22}=7$ ergibt sich für die Partialfunktion f_2 der Erwartungswert

$$E(f_2) = E(|X_{12} - 7|) = 7 - E(X_{12}) = 7 - 4 = 3.$$

Offensichtlich unterscheidet sich dieser Wert von dem oben betrachteten Erwartungswert E_f . E_f ist der Schätzer für den Erwartungswert der Partialfunktion bei vollständiger Unkenntnis der Merkmalsausprägungen, was beispielsweise bei zwei Missing Values der Fall wäre. Das Vernachlässigen der Information über x_{22} kann also zum einen als Informationsverlust gedeutet werden, aber ebenso als Grund für eine Verzerrung der Proximitätsfunktion.

Analog ergibt sich bei Anwendung des Imputationsverfahrens ein verzerrter Schätzer, wenn die Auswirkungen auf f_m nicht beachtet werden. So würde eine Mittelwertersetzung im obigen Beispiel ebenfalls einen verzerrten Schätzer für das Proximitätsmaß liefern.

Zusammenfassend können Eliminierungsverfahren also zu einer Verzerrung der Proximitätsfunktion führen, da Informationen eliminiert werden, die eine erwartungstreue Schätzung ermöglichen würden. Bei Imputationsverfahren gilt dies analog, nur daß statt eines Informationsverlustes ggf. zusätzliche „falsche“ Werte imputiert werden, die zur Verzerrung der Proximität führen.

B Abbildungen

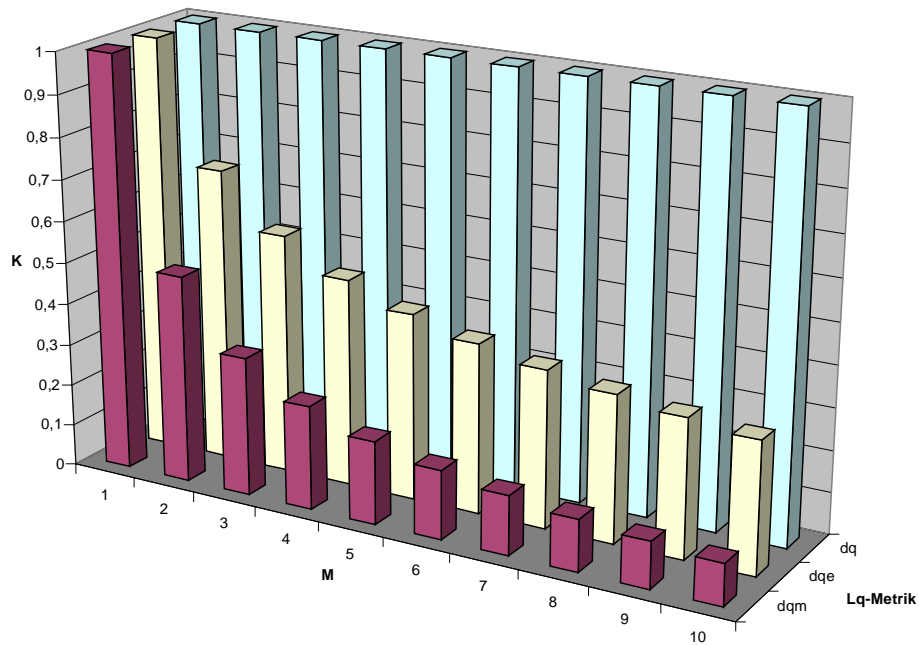


Abbildung 1 – Korrekturterme für Lq-Metriken (q=2)

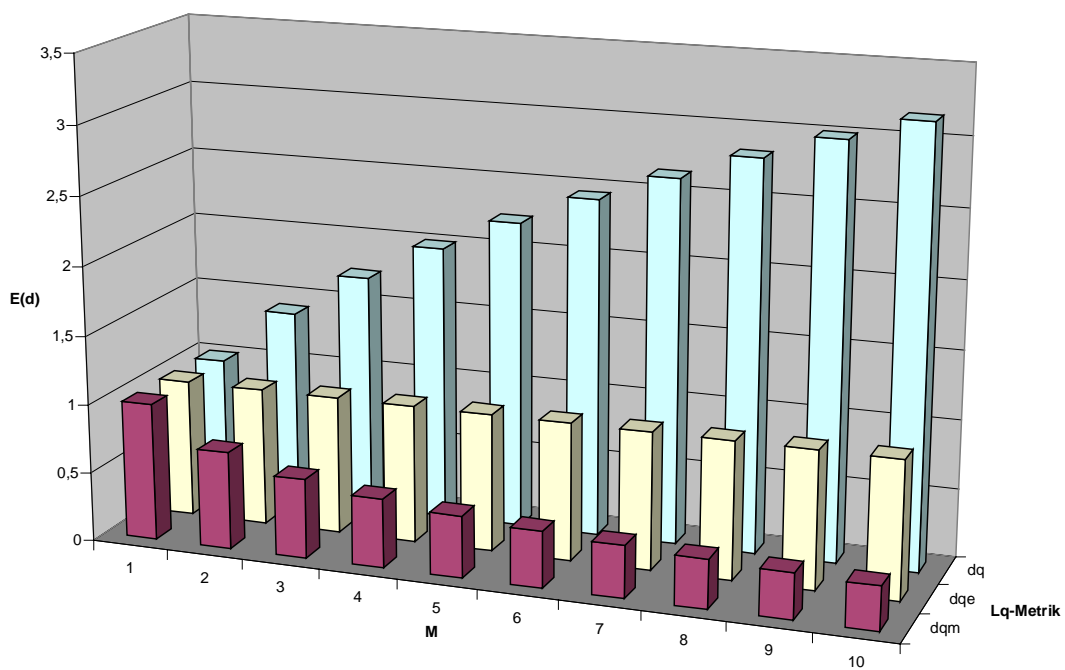


Abbildung 2 – Relative Erwartungswerte für korrigierte Lq-Metriken (q=2)

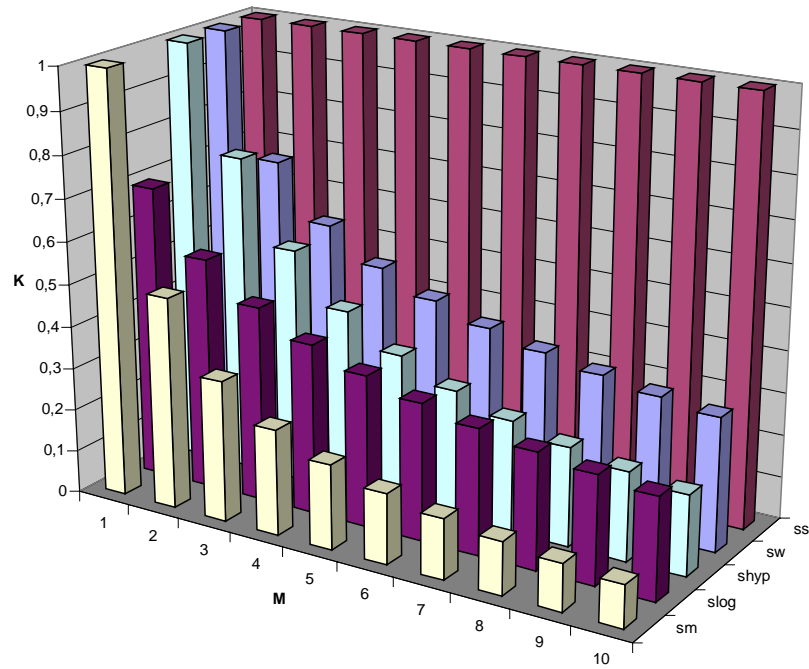


Abbildung 3 – Korrekturterme K für ein abgeleitetes Ähnlichkeitsmaß

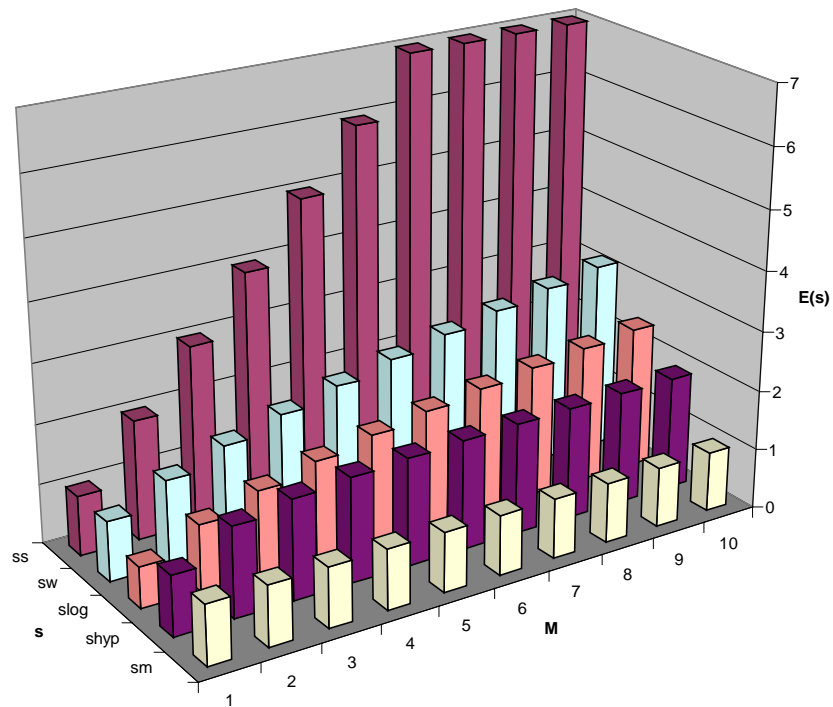


Abbildung 4 - Relative Erwartungswerte für ein korrigiertes Ähnlichkeitsmaß