IMPLIZITE GEWICHTUNGEN AM BEISPIEL DER NONMETRISCHEN MULTIDIMENSIONALEN SKALIERUNG

Claus Möbus
Psychologisches Institut
der
Universität Heidelberg

1. Einleitung

In der nonmetrischen multidimensionalen Skalierung (NMDS) wird mit zwei unterschiedlichen Distanzbegriffen operiert. Die erste Distanzfunktion definiert die Metrik des subjektiven Raumes (z.B. MINKOWSKY - Metriken). Über ihre impliziten „Gewichtungseigenschaften" ist schon relativ intensiv gearbeitet worden (SHEPARD,1964;CROSS,1965a,b;WENDER, 1969;COOMBS,1970;FISCHER & MICKO,1972;AHRENS,1974), obwohl ihre heuristische Brauchbarkeit in letzter Zeit in Zweifel gezogen zu sein scheint (BORTZ, 1974). Der andere Distanzbegriff verbirgt sich hinter dem statistischen Anpassungsmaß ("stress","coefficient of alienation"), das die "Distanz" zwischen Daten (bzw. monoton transfomierten Daten) und Modell mißt und das im Laufe der Parameterschätzung minimiert werden soll. Auch hier kommen implizite Gewichtungen (=Fehlergewichtungen) vor, die, jedoch im Gegensatz zur Wahl einer bestimmten Metrik, ungewollt sind. Obwohl sich das Problem impliziter Fehlergewichtungen auch bei anderen statistischen Schätzproblemen stellt, erscheint es bei der NMDS gravierender, weil die Zahl der unbekannten Parameter (Metrikparameter, Koordinaten der Reizpunkte) im Verhältnis zur Zahl der Daten und Datenrelationen relativ hoch ist, zumal die Dimensionalität ebenfalls unbekannt ist.

Aus diesen Überlegungen heraus wollen wir ein neutrales Anpassungsmaß vorschlagen, das neben einer Reihe weiterer Vorteile frei von impliziten Gewichtungen und damit ungewollten Verzerrungen ist. Es bietet ferner durch seine Einfachheit didaktische und konzeptuelle Vorteile gegenüber den bisher gebräuchlichen Maßen.

2. NMDS und bisher verwendete Anpassungsmaße

Die bisher klarste und einfachste Definition der NMDS gab SHEPARD (1972, S.7f.): "We seek, simply, that configuration of n points in the (Euclidean) space of smallest possible dimension such that, to an acceptable degree of approximation, the resulting interpoint distances d_{ij} are monotonically related to the given proximity data in the sense that:

$$d_{ij} < d_{kl} \quad \text{whenever} \quad s_{ij} > s_{kl} \text{."}$$

Zur Erreichung dieses auf den ersten Blick einfach erscheinenden Zieles sind von einer Reihe von Autoren (KRUSKAL, 1964a,b;McGEE,1966;GUTTMAN, 1968; ROSKAM & LINGOES,1970) zweiphasige und von JOHNSON (1973) ein einphasiger Algorithmus entwickelt worden. Wesentlich bei der NMDS ist dabei die numerische Minimierung des "stress". Bei zweiphasigen Algorithmen erfolgt die Minimierung iterativ unter Verwendung monoton transformierter Daten (=Disparitäten). Die Disparitäten werden über KRUSKAL's 'monotone Regression' oder über GUTTMAN's 'Rank-Image-Prinzip' gewonnen. Das einfachere einphasige Verfahren von JOHNSON verzichtet auf Disparitäten. Die Lossfunktion L ist in allen zweiphasigen Algorithmen eine quadratische Fehlerfunktion (Methode der kleinsten Quadrate), obwohl sich die statistischen Gütemerkmale von least-squares-Schätzungen nicht ohne weiteres in den nonmetrischen Bereich übertragen lassen. (1/N) ist ein algorithmenspezifischer Normierungsfaktor.

$$L^2 = (1/n)\cdot\Sigma(\tilde{d}_{ij} - d_{ij})^2 = (1/N)\cdot\Sigma|\tilde{d}_{ij} - d_{ij}|\,|\tilde{d}_{ij} - d_{ij}|$$

L kann als euklidischer Abstand der gerade vorliegenden Modellkonfiguration von der idealen datenkonformen Konfiguration angesehen werden. Ähnliches gilt für JOHNSON's stress:

$$L^2 = \frac{\sum\limits_{IJ<KL}\ \sum\limits_{\substack{i>j\\k>l}} \delta_{ij,kl}(d_{ij}^2-d_{kl}^2)^2}{\sum\limits_{IJ<KL}\ \sum\limits_{\substack{i>j\\k>l}} (d_{ij}^2-d_{kl}^2)^2} = \frac{\Sigma\delta|d_{ij}-d_{kl}|\,|(d_{ij}-d_{kl})(d_{ij}+d_{kl})^2|}{\Sigma(d_{ij}^2-d_{kl}^2)^2}$$

wobei: $\delta_{ij,kl} = \begin{cases} 1, \text{ wenn signum}(s_{ij}-s_{kl}) = \text{signum}(d_{ij}-d_{kl}) \\ 0, \text{ sonst} \end{cases}$

und: $IJ = (i-1)i/2 + j$ $\quad KL = (k-1)k/2 + l$

s_{ij} = Ähnlichkeit zwischen i und j

Die Ungleichung IJ<KL wurde von JOHNSON noch nicht eingeführt, obwohl sie die Rechenzeit zur Berechnung von L^2 halbiert.

3. Implizite Gewichtung durch die verwendete Metrik

Die MINKOWSKY-r-Metrik, die in einigen NMDS-Programmen implementiert ist, weist folgende Gewichtungseigenschaften auf (vgl. AHRENS,1972):

$$d_r(i,j) = \left[\sum_k |x_{ik}-x_{jk}|^r\right]^{1/r} = \sum_k \left[\frac{|x_{ik}-x_{jk}|}{d_r(i,j)}\right]^{r-1} \cdot |x_{ik}-x_{jk}|$$

$$= \sum \left\{ \begin{matrix}\text{Gewicht der}\\ \text{Aspektdistanz}\end{matrix} \right\} \left\{ \begin{matrix}\text{Aspektdis-}\\ \text{tanz}\end{matrix} \right\}$$

r = Metrikparameter

<u>Bei steigendem r</u> gewinnen die größten spezifischen Reizunterschiede immer mehr an Gewicht bei der Kombination aspektspezifischer Distanzen zu einer globalen Reizdistanz. Im Sinne von WENDER (1969) könnte man bei steigendem r auch von <u>Urteilsvergröberung</u> sprechen. Stärkste Sensibilität liegt dann vor, wenn sich das Urteilsverhalten mittels city-block-Metrik optimal beschreiben läßt.

4. Implizite Gewichtung durch die verwendete Lossfunktion L ("Stress")

Wie man in 2. sehen kann, weisen die L^2 ähnliche Gewichtungen auf, da sie euklidischen Abständen entsprechen. Bei steigendem r dominieren auch hier wieder wie bei 3. die großen Diskrepanzen (=Abweichungen von der Monotonizität). Soll L_r sensibler werden und ist die implizite Gewichtung statistisch nicht zu begründen, sollte L_1 vorgezogen werden.

$$L_r(\tilde{d},d) = \sum_{ij} \left[\frac{|\tilde{d}_{ij} - d_{ij}|}{NL_r(\tilde{d},d)} \right]^{r-1} \cdot |\tilde{d}_{ij} - d_{ij}| \qquad r = 2 \text{ Methode der kleinsten Quadrate}$$

$$\Sigma\{\text{Fehlergewicht}\}\cdot\{\text{'Fehler'}\}$$

Bei JOHNSON's stress wird die den Daten widersprechende Diskrepanz $\delta|d_{ij}-d_{kl}|$ massiv gewichtet. Eine Minimierung dieses Maßes kann auf Kosten der wichtigen Proportion $\Sigma\delta/\Sigma 1$, die Hinweise auf die Rangisomorphie zwischen Daten und Distanzen gibt, geschehen. Ein sensibleres Maß müßte auch hier größere Ähnlichkeit mit der city-block-Metrik aufweisen (s.a.5.).

5. Neues gewichtungsfreies Stressmaß

Aus den oben und noch an anderer Stelle (MÖBUS, 1974) dargelegten Gründen schlagen wir ein L vor, das SHEPARD's verbale Definition (s.a.2.) einfach und klar mathematisch umsetzt. Das Maß ist normiert ($0\leq L\leq 1$), läßt sich als Proportion "datenwidriger Modellanteile" interpretieren, verzichtet auf monoton transformierte Daten und weist keine ungewollten Diskrepanzgewichtungen auf. Auf die schwierigere mathematische Behandlung wird an anderer Stelle (MÖBUS, 1974) eingegangen.

$$L = \frac{\sum\limits_{IJ<KL} \sum\limits_{\substack{i>j\\k>l}} \delta_{ij,kl}|d_{ij} - d_{kl}|}{\sum\limits_{IJ<KL} \sum\limits_{\substack{i>j\\k>l}} |d_{ij} - d_{kl}|}$$

Proportion datenwidriger Distanzen (= Modellanteile)

6. Literaturverzeichnis

AHRENS,H.J.Zur Verwendung des Metrik-Parameters multidimensionaler Skalierungen bei der Analyse von Wahrnehmungsstrukturen.Z.exp.angew.Psych., 1972,19,173-195; AHRENS, H.J. Multidimensionale Skalierung.Weinheim:Beltz 1974; BORTZ,J.Kritische Bemerkungen zur Verwendung nicht-euklidischer Metriken in der multidimensionalen Skalierung,Vortrag auf dem 29.Kongress der DGfP, 1974 in Salzburg; COOMBS, C.H. Scaling and Data Theory.in: C.H. COOMBS, R.M. DAWES & A.TVERSKY. Mathematical Psychology. Englewood Cliffs N.J: Prentice-Hall, 1970,31-77; CROSS, D.V. Metric properties of multidimensional stimulus generalization.in: D.J. MOSTOFSKY(ed).Stimulus Generalisation. Stanford: University Press. 1965(a), 72-93; CROSS, D.V. Metric properties of multidimensional stimulus control. Unpublished doctoral dissertation. University of Michigan, 1965(b); FISCHER & MICKO, H.C. More about metrics of subjective spaces and attention distributions. Journal of mathematical psychology, 1972, 9, 36-54; GUTTMAN, L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. Psychometrika, 1968,33,469-5o6; JOHNSON, R.M. Pairwise nonmetric multidimensional scaling. Psychometrika, 38, 1973, 11-18; KRUSKAL, J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis.Psychometrika,1964(a),29,1-27; KRUSKAL, J.B. Multidimensional scaling: a numerical method.Psychometrika,1964(b),29,115-129; McGEE,V. Multidimensional scaling of 'elastic' distances. The British J. Math.Statist.Psychology,1966,19,181-196; MÖBUS, C. Ein Beitrag zur metrischen und nonmetrischen Analyse von nichtsymmetrischen Proximitätsmatrizen mit multidimensionalen Skalierungsmodellen.Unveröffentlichte Dissertation, Universität Heidelberg, 1974; ROSKAM, E.E. & LINGOES, J.C. Minissa-I: A Fortran-IV (G) program for the smallest space analysis of square symmetric matrices. Behavioral Sciences, 1970,15, 2o4-2o5; SHEPARD R.N. Attention and the metric structure of stimulus space. Journal of Mathematical Psychology, 1964,1,54-87; WENDER,K. Die psychologische Interpretation nichteuklidischer Metriken in der multidimensionalen Skalierung. Dissertation an der TH Darmstadt, 1969.