

1 Das Shepard-Diagramm

Das Ergebnis der MDS einer (n, n) -Distanzmatrix $\mathbf{D} = (d_{ij})$ ist eine Konfiguration von Punkten $\mathbf{x}_1, \dots, \mathbf{x}_n$. Die euklidische Distanz zwischen \mathbf{x}_i und \mathbf{x}_j ist

$$\tilde{d}_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}$$

Wir bezeichnen die Matrix der euklidischen Distanzen \tilde{d}_{ij} zwischen den Punkten der Konfiguration mit \tilde{D} . Bei der metrischen MDS sollten die \tilde{d}_{ij} und die d_{ij} so gut wie möglich übereinstimmen. Bei der nichtmetrischen MDS nach Kruskal sollten die \tilde{d}_{ij} wie die d_{ij} geordnet sein. Es liegt nahe, die \tilde{d}_{ij} und die d_{ij} zu vergleichen. Shepard ([3]) hat vorgeschlagen, das Streudiagramm der \tilde{d}_{ij} gegen die d_{ij} zu zeichnen. Man nennt es auch das **Shepard-Diagramm**. Um die Übereinstimmung der \tilde{d}_{ij} und die d_{ij} zu überprüfen, sollte man die Winkelhalbierende einzeichnen.

Zur Überprüfung der Monotonie sollte man die Punkte miteinander verbinden. Man erhält einen Polygonzug, der im Idealfall monoton wachsend ist. Schauen wir uns das Beispiel 3 auf Seite 5 im Buch an.

Beispiel 1

Im Wintersemester 1996/97 wurden an der Fakultät für Wirtschaftswissenschaften der Universität Bielefeld 265 Erstsemesterstudenten in der Statistik I Vorlesung befragt. Neben dem Merkmal **Geschlecht** mit den Ausprägungsmöglichkeiten **w** und **m** wurden die Merkmale **Gewicht**, **Alter** und **Größe** erhoben. Außerdem wurden die Studenten gefragt, ob sie rauchen und ob sie ein Auto besitzen. Diese Merkmale bezeichnen wir mit **Raucher** und **Auto**. Auf einer Notenskala von 1 bis 5 sollten sie angeben, wie ihnen Cola schmeckt. Das Merkmal bezeichnen wir mit **Cola**. Als letztes wurde noch gefragt, ob die Studenten den Leistungskurs Mathematik besucht haben. Dieses Merkmal bezeichnen wir mit **MatheLK**. Tabelle 1 gibt die Ergebnisse von 5 Studenten wieder.

Tabelle 1: Ergebnis der Befragung von 5 Erstsemesterstudenten

Geschlecht	Alter	Größe	Gewicht	Raucher	Auto	Cola	MatheLK
m	23	171	60	n	j	2	j
m	21	187	75	n	j	1	n
w	20	180	65	n	n	3	j
w	20	165	55	j	n	2	j
m	23	193	81	n	n	3	n

Wir bestimmen die Distanzen mit dem Gower-Koeffizienten. Die Distanzmatrix ist:

$$\mathbf{D} = \begin{pmatrix} 0.000 & 0.414 & 0.502 & 0.551 & 0.512 \\ 0.414 & 0.000 & 0.621 & 0.799 & 0.389 \\ 0.502 & 0.621 & 0.000 & 0.303 & 0.510 \\ 0.551 & 0.799 & 0.303 & 0.000 & 0.812 \\ 0.512 & 0.389 & 0.510 & 0.812 & 0.000 \end{pmatrix}$$

Wir führen eine metrische MDS durch. Die Konfiguration ist im Buch in Abbildung 6.1 auf Seite 137 zu finden.

Die Punkte sind

$$\mathbf{x}_1 = \begin{pmatrix} -0.04 \\ -0.23 \end{pmatrix} \quad \mathbf{x}_2 = \begin{pmatrix} -0.33 \\ -0.11 \end{pmatrix} \quad \mathbf{x}_3 = \begin{pmatrix} 0.21 \\ 0.18 \end{pmatrix}$$

$$\mathbf{x}_4 = \begin{pmatrix} 0.46 \\ -0.04 \end{pmatrix} \quad \mathbf{x}_5 = \begin{pmatrix} -0.30 \\ 0.21 \end{pmatrix}$$

Die Matrix der euklidischen Distanzen ist

$$\tilde{\mathbf{D}} = \begin{pmatrix} 0.000 & 0.314 & 0.480 & 0.535 & 0.511 \\ 0.314 & 0.000 & 0.613 & 0.793 & 0.321 \\ 0.480 & 0.613 & 0.000 & 0.333 & 0.511 \\ 0.535 & 0.793 & 0.333 & 0.000 & 0.800 \\ 0.511 & 0.321 & 0.511 & 0.800 & 0.000 \end{pmatrix}$$

Abbildung 1 zeigt das Shepard-Diagramm zur Überprüfung der Übereinstimmung. Die Anpassung ist gut. Es fällt aber auf, dass kleine Distanzen schlechter angepasst werden als große.

Abbildung 1: Shepard-Plot mit Winkelhalbierender

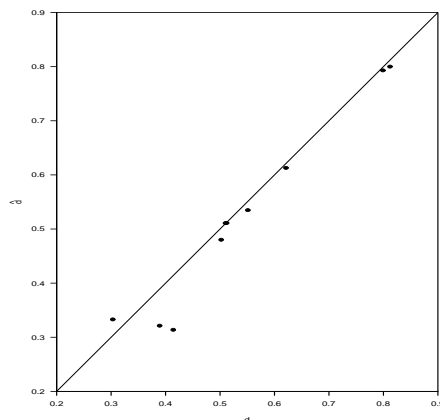
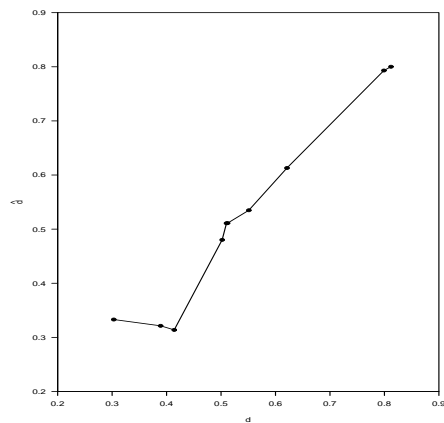


Abbildung 2 zeigt das Shepard-Diagramm zur Überprüfung der Übereinstimmung. Die Monotoniebedingung ist bei großen Distanzen erfüllt, bei kleinen hingegen nicht.

Abbildung 2: Shepard-Plot mit verbundenen Punkten



2 Wie geht man bei einer metrischen MDS vor?

Will man im \mathbb{R}^2 eine Darstellung mit einer metrischen MDS gewinnen, so sollte man folgende Punkte abarbeiten.

1. Liegen die Daten in Form einer Distanzmatrix \mathbf{D} vor?

Falls dies der Fall ist, kann man zu Punkt 2. übergehen.

Ansonsten bestimmt man die Distanzmatrix \mathbf{D} aus der Datenmatrix \mathbf{X} und geht zu Punkt 2..

2. Man führt die metrische MDS durch, d.h. man arbeitet die folgenden Punkte ab:

- (a) Bilde die Matrix $\mathbf{A} = (a_{rs})$ mit

$$a_{rs} = -0.5 d_{rs}^2.$$

- (b) Bilde die Matrix $\mathbf{B} = (b_{rs})$ mit

$$b_{rs} = a_{rs} - \bar{a}_{r.} - \bar{a}_{.s} + \bar{a}_{..}$$

- (c) Führe eine Spektralzerlegung von \mathbf{B} durch:

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'.$$

- (d) Bilde die Diagonalmatrix $\mathbf{\Lambda}_1$ mit den beiden größten Eigenwerten λ_1 und λ_2 von \mathbf{B} und die Matrix \mathbf{U}_1 mit den zu λ_1 und λ_2 gehörenden normierten Eigenvektoren. Die Konfiguration bilden dann die Zeilenvektoren von

$$\mathbf{X}_1 = \mathbf{U}_1 \mathbf{\Lambda}_1^{0.5}.$$

Sind die beiden größten Eigenwerte positiv und alle anderen Eigenwerte gleich 0, so ist die Darstellung im \mathbb{R}^2 exakt.

3. Man überprüft die Güte der Anpassung mit einem Shepard-Plot
4. Man stellt die Konfiguration grafisch dar.
5. Man interpretiert die Konfiguration.

Wir illustrieren die Punkte an einem Beispiel.

Beispiel 2

Bei einer Befragung wurden von Studierenden unter anderem die Ausprägungen folgender Merkmale erfragt:

- x_1 Geschlecht (1 ist weiblich)
- x_2 Wollten Sie in Bielefeld studieren? (1 ist ja)
- x_3 Würden Sie noch einmal in Bielefeld studieren? (1 ist ja)
- x_4 Haben Sie den Leistungskurs Mathematik besucht? (1 ist ja)
- x_5 Durchschnittsnote im Abitur
- x_6 Durchschnittsnote im Vordiplom

In Tabelle 2 sind die Daten zu finden.

Tabelle 2: Ergebnisse einer Befragung

Student	x_1	x_2	x_3	x_4	x_5	x_6
1	1	1	0	1	1.7	1.6
2	0	1	0	0	3.0	3.0
3	0	0	0	1	2.5	3.3
4	0	1	1	1	2.1	2.6
5	0	1	1	1	2.2	2.5

1. Es liegt keine Distanzmatrix vor. Wir bestimmen die Distanzen mit dem Gower-Koeffizienten. Die Distanzmatrix ist:

$$\mathbf{D} = \begin{pmatrix} 0.000 & 3.824 & 3.615 & 2.896 & 2.914 \\ 3.824 & 0.000 & 2.561 & 2.928 & 2.910 \\ 3.615 & 2.561 & 0.000 & 2.719 & 2.701 \\ 2.896 & 2.928 & 2.719 & 0.000 & 0.136 \\ 2.914 & 2.910 & 2.701 & 0.136 & 0.000 \end{pmatrix}$$

2. Wir führen die metrische MDS durch.

(a) Wir bilden die Matrix $\mathbf{A} = (a_{rs})$ mit $a_{rs} = -0.5 d_{rs}^2$. Es gilt

$$\mathbf{A} = \begin{pmatrix} 0.000 & -7.311 & -6.534 & -4.193 & -4.246 \\ -7.311 & 0.000 & -3.279 & -4.287 & -4.234 \\ -6.534 & -3.279 & 0.000 & -3.696 & -3.648 \\ -4.193 & -4.287 & -3.696 & 0.000 & -0.009 \\ -4.246 & -4.234 & -3.648 & -0.009 & 0.000 \end{pmatrix}$$

(b) Wir bilden die Matrix $\mathbf{B} = (b_{rs})$ mit $b_{rs} = a_{rs} - \bar{a}_{r.} - \bar{a}_{.s} + \bar{a}_{..}$. Es gilt

$$\mathbf{B} = \begin{pmatrix} 5.60 & -2.35 & -1.96 & -0.61 & -0.68 \\ -2.35 & 4.33 & 0.66 & -1.34 & -1.30 \\ -1.96 & 0.66 & 3.55 & -1.14 & -1.10 \\ -0.61 & -1.34 & -1.14 & 1.56 & 1.54 \\ -0.68 & -1.30 & -1.10 & 1.54 & 1.54 \end{pmatrix}$$

(c) Wir führen eine Spektralzerlegung von \mathbf{B} durch.

Die Eigenwerte sind

$$\lambda_1 = 8.42 \quad \lambda_2 = 4.93 \quad \lambda_3 = 3.22 \quad \lambda_4 = 0.01 \quad \lambda_5 = 0$$

Die Eigenvektoren, die zu den beiden größten Eigenwerten gehören, sind

$$\mathbf{u}_1 = \begin{pmatrix} 0.69 \\ -0.56 \\ -0.42 \\ 0.15 \\ 0.14 \end{pmatrix} \quad \mathbf{u}_2 = \begin{pmatrix} 0.57 \\ 0.30 \\ 0.19 \\ -0.53 \\ -0.53 \end{pmatrix}$$

(d) Wir bilden die Diagonalmatrix $\mathbf{\Lambda}_1$ mit den beiden größten Eigenwerten λ_1 und λ_2 von \mathbf{B}

$$\mathbf{\Lambda}_1 = \begin{pmatrix} 8.42 & 0.00 \\ 0.00 & 4.93 \end{pmatrix}$$

und die Matrix \mathbf{U}_1 mit den zu λ_1 und λ_2 gehörenden normierten Eigenvektoren

$$\mathbf{U}_1 = \begin{pmatrix} 0.69 & 0.57 \\ -0.56 & 0.30 \\ -0.42 & 0.19 \\ 0.15 & -0.53 \\ 0.14 & -0.53 \end{pmatrix}$$

Die Konfiguration bilden dann die Zeilenvektoren von

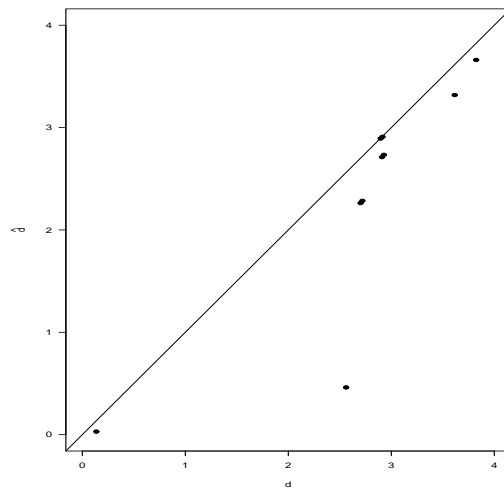
$$\mathbf{X}_1 = \mathbf{U}_1 \mathbf{\Lambda}_1^{0.5} = \begin{pmatrix} 2.00 & 1.26 \\ -1.61 & 0.66 \\ -1.21 & 0.42 \\ 0.43 & -1.17 \\ 0.40 & -1.17 \end{pmatrix}$$

3. Die euklidischen Distanzen der Konfiguration sind:

$$\tilde{\mathbf{D}} = \begin{pmatrix} 0.00 & 3.66 & 3.32 & 2.89 & 2.91 \\ 3.66 & 0.00 & 0.46 & 2.73 & 2.71 \\ 3.32 & 0.46 & 0.00 & 2.28 & 2.26 \\ 2.89 & 2.73 & 2.28 & 0.00 & 0.03 \\ 2.91 & 2.71 & 2.26 & 0.03 & 0.00 \end{pmatrix}$$

Abbildung 3 zeigt das Shepard-Diagramm.

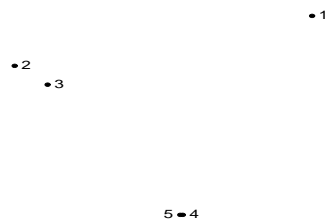
Abbildung 3: Shepard-Plot mit Winkelhalbierender



Wir sehen, dass die Distanz zwischen dem zweiten und dritten Studierenden sehr schlecht angepasst ist.

4. Abbildung 4 zeigt die Konfiguration.

Abbildung 4: Konfiguration der Metrischen MDS



5. Ein Blick in die Daten in Tabelle 2 auf Seite 5 liefert eine einfache Interpretation der ersten Achse der Darstellung. Sie bezieht sich auf die Noten der Studierenden. Die Daten zeigen auch, warum der erste Studierende von allen anderen entfernt ist. Es handelt sich um die einzige Frau, die außerdem die besten Noten hat.

3 Das Verfahren von Sammon

Im Beispiel 2 auf Seite 5 haben wir gesehen, dass bei der metrischen mehrdimensionalen Skalierung kleine Distanzen besser als große angepasst werden. Eine Erklärung ist bei Mardia, Kent und Bibby ([1], S.406) zu finden. Von Sammon ([2]) wurde ein Verfahren vorgeschlagen, bei dem kleine Distanzen besser angepasst werden. Ausgehend von den Distanzen d_{ij} sucht Sammon eine Konfiguration mit Distanzen \tilde{d}_{ij} so, dass

$$\sum_{i \neq j} \frac{(d_{ij} - \tilde{d}_{ij})^2}{d_{ij}}$$

minimal wird. Das Optimierungsverfahren wird von Sammon beschrieben. Es ist in R implementiert. Wir schauen uns hier nur das Ergebnis für ein Beispiel an.

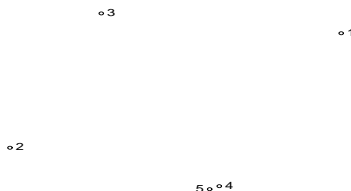
Beispiel 2 (fortgesetzt)

Mit dem Verfahren von Sammon erhalten wir folgende Konfiguration, die in den Zeilen der Matrix \mathbf{X} steht.

$$\mathbf{X} = \begin{pmatrix} 2.15 & 1.15 \\ -2.17 & -0.48 \\ -0.98 & 1.43 \\ 0.56 & -1.03 \\ 0.44 & -1.07 \end{pmatrix}$$

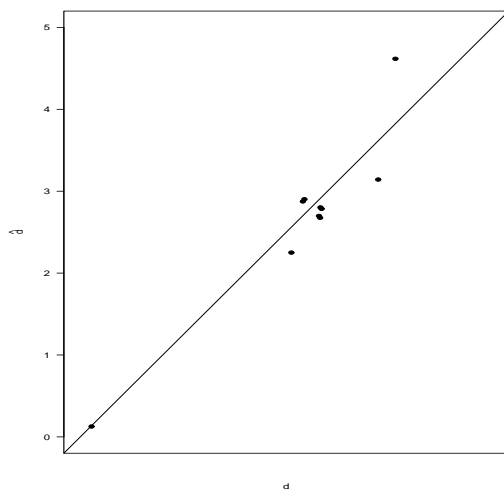
Abbildung 5 zeigt die Konfiguration.

Abbildung 5: Konfiguration der Metrischen MDS



Vergleichen wir Abbildung 5 mit Abbildung 4 auf Seite 8, so sehen wir, dass die Studierenden 2 und 3 in der durch das Verfahren von Sammon gewonnenen Darstellung viel weiter voneinander entfernt sind als in der durch die metrische MDS gewonnenen Darstellung. Bei der metrischen MDS war dies die Distanz, die am schlechtesten angepasst wurde. Dieser Defekt tritt bei Sammon nicht auf, dafür werden die großen Distanzen nicht so gut angepasst, wie das Shepard-Diagramm in Abbildung 6 zeigt. Die Anpassung durch das Verfahren von Sammon ist auf jeden Fall viel besser als durch die metrische MDS.

Abbildung 6: Shepard-Plot mit Winkelhalbierender



4 Wie geht man bei der nichtmetrischen MDS nach Kruskal vor?

Will man im \mathbb{R}^2 eine Darstellung mit einer nichtmetrischen MDS nach Kruskal gewinnen, so sollte man folgende Punkte abarbeiten.

1. Liegen die Daten in Form einer Distanzmatrix \mathbf{D} vor?
Falls dies der Fall ist, kann man zu Punkt 2. übergehen.
Ansonsten bestimmt man die Distanzmatrix \mathbf{D} aus der Datenmatrix \mathbf{X} und geht zu Punkt 2..
2. Man führt die metrische MDS durch und erstellt das Shepard-Diagramm zur Überprüfung der Monotonie.
Ist der Polygonzug monoton wachsend, so hat man die gesuchte Konfiguration bereits mit der metrischen MDS gefunden. Man kann zu Punkt 5. übergehen.
Ist der Polygonzug nicht monoton wachsend, so geht man zu Punkt 3..
3. Man wendet das iterative Verfahren von Kruskal an. Dieses liefert die Konfiguration der Punkte.
4. Man beurteilt die Güte der Konfiguration an Hand des Wertes von STRESS1.
5. Man stellt die Konfiguration grafisch dar.
6. Man interpretiert die Konfiguration.

Schauen wir uns zwei Beispiele an.

Beispiel 3

Ein Student sollte die 10 Paare, die man aus 5 Politikern bilden kann, mit der Rangreihung vergleichen. Er sollte also dem Paar, bei dem sich die Politiker am ähnlichsten sind, den Rang 1 geben, dem Paar, bei dem sich die Politiker am zweitähnlichsten sind, den Rang 2 geben, u.s.w.. Die Daten sind in Tabelle 3 zu finden.

Tabelle 3: Distanzen auf Grund der Rangreihung

	Berlusconi	Blair	Bush	Putin	Schroeder
Berlusconi	0	3	2	4	9
Blair	3	0	1	6	7
Bush	2	1	0	8	10
Putin	4	6	8	0	5
Schroeder	9	7	10	5	0

Wir arbeiten die einzelnen Punkte ab.

1. Die Daten liegen als Distanzmatrix vor.
2. Wir führen eine metrische MDS durch. Die Konfiguration ist:

$$\mathbf{X} = \begin{pmatrix} -2.5 & 2 \\ -1.6 & -1.8 \\ -4.4 & -1.2 \\ 2.7 & 2.8 \\ 5.8 & -1.8 \end{pmatrix}$$

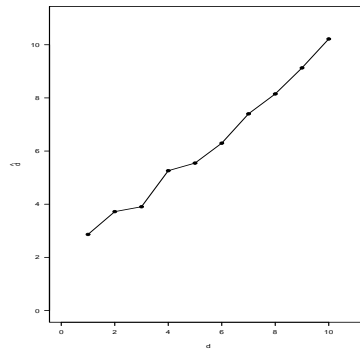
Tabelle 4 zeigt die euklidischen Distanzen der Konfiguration.

Tabelle 4: Die euklidischen Distanzen der Konfiguration

	Berlusconi	Blair	Bush	Putin	Schroeder
Berlusconi	0.0	3.9	3.7	5.3	9.1
Blair	3.9	0.0	2.9	6.3	7.4
Bush	3.7	2.9	0.0	8.1	10.2
Putin	5.3	6.3	8.1	0.0	5.5
Schroeder	9.1	7.4	10.2	5.5	0.0

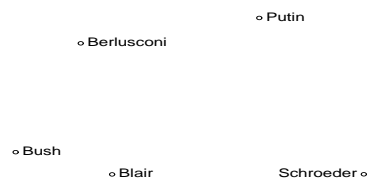
Im Shepard-Plot in Abbildung 7 ist der Polygonzug monoton. Also erfüllt die Konfiguration der metrischen MDS bereits die Bedingung.

Abbildung 7: Shepard-Plot mit verbundenen Punkten



5. Abbildung 8 zeigt die Konfiguration.

Abbildung 8: Konfiguration der Metrischen MDS



Beispiel 4

Ein Student sollte die 10 Paare, die man aus 5 Politikern bilden kann, mit der Rangreihung vergleichen. Er sollte also dem Paar, bei dem sich die Politiker am ähnlichsten sind, den Rang 1 geben, dem Paar, bei dem sich die Politiker am zweitähnlichsten sind, den Rang 2 geben, u.s.w.. Die Daten sind in Tabelle 5 zu finden.

Tabelle 5: Distanzen auf Grund der Rangreihung

	Berlusconi	Blair	Bush	Putin	Schroeder
Berlusconi	0	7	2	4	8
Blair	7	0	6	9	3
Bush	2	6	0	1	10
Putin	4	9	1	0	5
Schroeder	8	3	10	5	0

Wir arbeiten die einzelnen Punkte ab.

1. Die Daten liegen als Distanzmatrix vor.
2. Wir führen eine metrische MDS durch. Die Konfiguration ist:

$$\mathbf{X} = \begin{pmatrix} 2.7 & 0.5 \\ -3.7 & 3.4 \\ 4.0 & 1.8 \\ 2.1 & -3.5 \\ -5.1 & -2.3 \end{pmatrix}$$

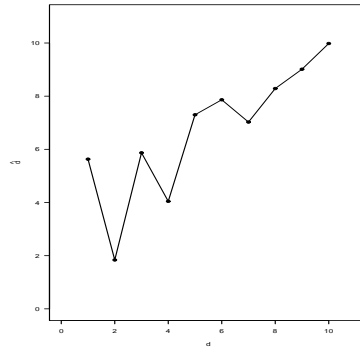
Tabelle 6 zeigt die euklidischen Distanzen der Konfiguration.

Tabelle 6: Die euklidischen Distanzen der Konfiguration

	Berlusconi	Blair	Bush	Putin	Schroeder
Berlusconi	0.0	7.0	1.8	4.0	8.3
Blair	7.0	0.0	7.9	9.0	5.9
Bush	1.8	7.9	0.0	5.6	10.0
Putin	4.0	9.0	5.6	0.0	7.3
Schroeder	8.3	5.9	10.0	7.3	0.0

Im Shepard-Plot in Abbildung 9 ist der Polygonzug nicht monoton.

Abbildung 9: Shepard-Plot mit verbundenen Punkten



3. Das Verfahren von Kruskal liefert folgende Konfiguration:

$$\mathbf{X} = \begin{pmatrix} 4.80 & -0.71 \\ -7.12 & 1.47 \\ 4.76 & -0.93 \\ 4.74 & -1.02 \\ -7.18 & 1.19 \end{pmatrix}$$

4. Tabelle 7 zeigt die euklidischen Distanzen der Konfiguration. Aus diesen können wir den Wert von STRESS berechnen, nachdem wir eine monotone Regression durchgeführt haben.

Tabelle 7: Die euklidischen Distanzen der Konfiguration

	Berlusconi	Blair	Bush	Putin	Schroeder
Berlusconi	0.000	12.118	0.224	0.316	12.130
Blair	12.118	0.000	12.120	12.119	0.286
Bush	0.224	12.120	0.000	0.092	12.127
Putin	0.316	12.119	0.092	0.000	12.123
Schroeder	12.130	0.286	12.127	12.123	0.000

Wir führen die monotone Regression durch. Zunächst bringen wir die Distanzen aus Tabelle 7 in die Reihenfolge der Distanzen aus Tabelle 5 auf Seite 14.

0.092 0.224 0.286 0.316 12.123 12.120 12.118 12.130 12.119 12.127

Die Monotoniebedingung ist beim 5-ten, 6-ten und 7-ten Element verletzt. Wir ersetzen das 5-te, 6-te und 7-te Element durch ihren Mittelwert.

0.092 0.224 0.286 0.316 12.1203 12.1203 12.1203 12.13 12.119 12.127

Jetzt ist die Monotoniebedingung beim 8-ten und 9-ten Element verletzt. Wir ersetzen das 8-te und 9-te Element durch ihren Mittelwert.

0.092 0.224 0.286 0.316 12.1203 12.1203 12.1203 12.1245 12.1245 12.127

Nun ist die Folge monoton. Wir berechnen

$$\text{STRESS1} = \sqrt{\frac{\sum_{i < j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i < j} d_{ij}^2}} = 0.00029.$$

Die Konfiguration ist also hervorragend.

5. Abbildung 10 zeigt die Konfiguration.

Abbildung 10: Konfiguration der Nichtmetrischen MDS nach Kruskal

• Blair

• Schroeder

Berlusconi •

Bush •
Putin •

Literatur

- [1] Mardia, K.V., J.T. Kent, J.M. Bibby (1979): Multivariate analysis. London: Acad. Press.
- [2] Sammon, J.W. (1969): A nonlinear mapping for data structure analysis. IEEE Transactions on Computers, 18, 401-409
- [3] Shepard, R.N. (1962): The analysis of proximities: Multidimensional scaling with unknown distance function. Psychometrika, vol. 27, 219-246