Clusteranalyse

Multivariate Analyse

Mathematischer Hintergrund, durchgerechnetes Beispiel, Handhabung des Tools

Deutsch

Wolfgang Böhme

wolfgang.boehme@economic-modeling.de

www.economic-modeling.de

8. November 2026

Dokumentationen dieser Webseite wurden mit Unterstützung der KI Claude (Anthropic) entwickelt.

Inhaltsverzeichnis

1 Einleitung und Anwendungsgebiete **2**

2 Mathematischer Hintergrund **2**
2.1 Distanzmasse . . . 2
2.2 Hierarchische Verfahren . . . 2
2.2.1 Ward-Methode (Varianzminimierung) . . . 2
2.2.2 Complete Linkage (Maximum-Distanz) . . . 3
2.2.3 Average Linkage (Durchschnittsdistanz) . . . 3
2.3 k-Means-Algorithmus . . . 3
2.4 Wahl der Clusteranzahl . . . 3
2.4.1 Elbow-Methode . . . 3
2.4.2 Silhouetten-Koeffizient . . . 3

3 Durchgerechnetes Beispiel: Bilanzkennzahlen **4**
3.1 Datenbasis . . . 4
3.2 Schritt 1: Distanzmatrix (euklidisch, standardisiert) . . . 4
3.3 Schritt 2: Hierarchisches Clustering (Ward) . . . 4
3.4 Schritt 3: Clusterprofile ($k = 3$) . . . 5
3.5 Schritt 4: Silhouetten-Koeffizient und Wahl von k . . . 5

4 Handhabung des interaktiven Clusteranalyse-Tools **6**
4.1 Dateneingabe . . . 6
4.2 Parameterauswahl . . . 6
4.3 Ergebnistabs . . . 7
4.4 Typischer Arbeitsablauf . . . 7

5 Formelübersicht **8**

6 Weiterführende Literatur **8**

Einleitung und Anwendungsgebiete

Die **Clusteranalyse** ist ein unueberwachtes Klassifikationsverfahren, das Beobachtungen anhand ihrer Aehnlichkeit in homogene Gruppen (*Cluster*) einteilt – ohne vorherige Kenntnis der Gruppenstruktur.

Typische Anwendungsgebiete:

- Marktsegmentierung (Kundentypologien)
- Biologische Taxonomien (Artenklassifikation)
- Dokumenten-Clustering (Textmining)
- Anomalieerkennung (Ausreisseridentifikation)
- Bildverarbeitung (Farbquantisierung)

Mathematischer Hintergrund

Distanzmasse

Euklidische Distanz (Standard):

$$d_E(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{1}$$

Manhattan-Distanz (robuster gegenueber Ausreissern):

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}| \tag{2}$$

Kosinus-Distanz (fuer proportionale Muster):

$$d_C(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|} \tag{3}$$

Hierarchische Verfahren

2.2.1 Ward-Methode (Varianzminimierung)

Die Ward-Methode minimiert den Anstieg der **totalen Within-Cluster-Streuung**:

$$d_W(C_i, C_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}} \, d_E(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j) \tag{4}$$

Mit Clustermittelwerten $\bar{\mathbf{x}}_i$, $\bar{\mathbf{x}}_j$ und Clustergroessen n_i, n_j.

2.2.2 Complete Linkage (Maximum-Distanz)

$$d_{CL}(C_i, C_j) = \max_{a \in C_i,\, b \in C_j} d(\mathbf{x}_a, \mathbf{x}_b) \tag{5}$$

2.2.3 Average Linkage (Durchschnittsdistanz)

$$d_{AL}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{a \in C_i} \sum_{b \in C_j} d(\mathbf{x}_a, \mathbf{x}_b) \tag{6}$$

Vergleich der Methoden:

Tabelle 1: Eigenschaften der hierarchischen Verfahren

Methode	Clusterform	Ausreisser	Empfehlung
Ward	kompakt, aehnl. Groesse	empfindlich	allgemein
Complete	runde Cluster	maessig	bei Ausreissern
Average	Kompromiss	robust	allgemein

k-Means-Algorithmus

1. **Initialisierung** (k-Means++): Ersten Centroid zufaellig waehlen; weitere Centroide mit Wahrscheinlichkeit proportional zu $d^2(\mathbf{x}_i, \text{naechster Centroid})$
2. **Zuweisung**: Jede Beobachtung dem naechsten Centroid zuordnen: $c_i = \arg\min_k d(\mathbf{x}_i, \boldsymbol{\mu}_k)$
3. **Update**: Centroide aktualisieren: $\boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{x}_i$
4. Schritte 2–3 wiederholen bis Konvergenz

Zielfunktion (WCSS – Within-Cluster Sum of Squares):

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \tag{7}$$

Wahl der Clusteranzahl

2.4.1 Elbow-Methode

WCSS gegen k auftragen; der Knick (Ellbogen) im Graphen markiert das optimale k.

2.4.2 Silhouetten-Koeffizient

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \quad s_i \in [-1, 1] \tag{8}$$

mit a_i = mittlere Distanz zu Beobachtungen im eigenen Cluster, b_i = minimale mittlere Distanz zum naechsten fremden Cluster.

Tabelle 2: Interpretation des mittleren Silhouetten-Koeffizienten

$\bar{s}$	Interpretation
≥ 0.70	starke Clusterstruktur
$0.50 - 0.69$	vernuenftige Struktur
$0.25 - 0.49$	schwache Struktur
< 0.25	keine Struktur

Durchgerechnetes Beispiel: Bilanzkennzahlen

Datenbasis

Verwendet werden dieselben 24 synthetischen Unternehmen U01–U24 mit sechs Bilanzkennzahlen wie in Handbuch 1 (vollständige Datentabelle dort): X_1 Eigenkapitalquote, X_2 Liquidität 2. Grades, X_3 Anlagendeckung II, X_4 Umsatzrendite, X_5 Gesamtkapitalrendite, X_6 Umsatzwachstum. Einstellungen im Tool: Ward, euklidische Distanz, Z-Transformation, 3 Cluster (Werkseinstellungen).

Ziel der Clusteranalyse: Die Unternehmen sollen allein anhand der Kennzahlen zu homogenen Typen gruppiert werden – *ohne* Kenntnis des Ratings. Erst im Anschluss wird geprüft, wie gut die gefundenen Cluster mit den Ratings A/B/C übereinstimmen.

Schritt 1: Distanzmatrix (euklidisch, standardisiert)

Die Kennzahlen werden z-standardisiert (Mittelwerte und Standardabweichungen siehe Handbuch 1); die euklidische Distanzmatrix $\mathbf{D}$ hat die Dimension 24×24. Einige Werte:

$$d(U11, U14) = 0.726, \quad d(U01, U02) = 0.979, \quad d(U01, U12) = 3.007,$$
$$d(U01, U19) = 3.848, \quad d(U12, U19) = 5.514$$

Unternehmen desselben Typs liegen nahe beieinander; am weitesten voneinander entfernt sind ertragsstarke B-Unternehmen (U12) und gefährdete C-Unternehmen (U19).

Schritt 2: Hierarchisches Clustering (Ward)

Tabelle 3: Erste und letzte Fusionsschritte (Ward)

Schritt	Cluster A	Cluster B	d_W
1	U11	U14	0.513
2	U01	U02	0.692
3	U18	U19	0.717
...	...	...	...
20	{U17, U20}	{U03, U18, U19, U21–U24}	2.468
21	{U13, U16}	{U09–U12, U14, U15}	2.523
22	{U01, U02, U04–U08}	{U09–U16}	5.972
23	{U03, U17–U24}	{U01, U02, U04–U16}	7.652

Beispiel Schritt 1: $d_W(U11, U14) = \sqrt{\frac{1 \cdot 1}{2}}\, d_E(U11, U14) = 0.7071 \cdot 0.726 = 0.513$.

Bis Schritt 21 steigen die Fusionsdistanzen nur langsam (höchstens 2.52). Danach folgt ein großer Sprung auf 5.97 und 7.65: Die Daten zerfallen **natürlich in 3 Cluster**. Der Schnitt bei $k = 3$ (Voreinstellung) liegt zwischen Schritt 21 und 22 und trennt genau diese drei Gruppen.

Schritt 3: Clusterprofile ($k = 3$)

Tabelle 4: Clustermittelwerte (Originaldaten, in %)

Cluster	X_1 EK-Q.	X_2 Liqu.	X_3 AD II	X_4 UR	X_5 GKR	X_6 Wachst.
1 ($n = 7$)	47.01	130.29	128.71	5.33	7.59	5.30
2 ($n = 9$)	28.44	87.11	90.11	2.20	4.64	0.18
3 ($n = 8$)	34.12	99.38	96.75	8.55	11.16	11.71
Gesamt ($n = 24$)	35.75	103.79	103.58	5.23	7.68	5.52

Interpretation:

- Cluster 1 (U01, U02, U04–U08) – **solide Substanzwerte**: deutlich überdurchschnittliche Stabilität, durchschnittliche Ertragskraft.
- Cluster 2 (U03, U17–U24) – **gefährdete Unternehmen**: unterdurchschnittlich in allen sechs Kennzahlen, Umsatz stagniert.
- Cluster 3 (U09–U16) – **ertragsstarke Wachstumsunternehmen**: höchste Renditen und Wachstumsraten bei leicht unterdurchschnittlicher Stabilität.

Tabelle 5: Vergleich Cluster – Rating

	Rating A	Rating B	Rating C
Cluster 1	**7**	0	0
Cluster 2	1	0	**8**
Cluster 3	0	**8**	0

23 von 24 Unternehmen landen im Cluster „ihres“ Ratings, obwohl das Verfahren das Rating nicht kennt. Die Ausnahme ist **U03** (Rating A): Mit einer Eigenkapitalquote von 32.0 %, einer Anlagendeckung von 100 % und einer Gesamtkapitalrendite von nur 5.0 % liegt es bei fast allen Kennzahlen unter dem Durchschnitt. Sein Abstand zum Zentrum von Cluster 2 (1.76 ohne U03) ist kleiner als zum Zentrum von Cluster 1 (2.45). Für die Praxis heißt das: Entweder beruht das A-Rating auf Informationen, die nicht in den sechs Kennzahlen stecken, oder es ist ein Warnsignal für eine mögliche Herabstufung.

Schritt 4: Silhouetten-Koeffizient und Wahl von k

$\bar{s} = 0.441$ bei $k = 3$ – nach Tabelle 2 eine **schwache Struktur** (im Tool als „akzeptabel“ bewertet, Schwellen 0.5/0.25). Die besten Einzelwerte haben U02 ($s_i = 0.593$) und

U22 (0.588); die niedrigsten U07 (0.195) und U03 (0.212) – dieselben Grenzfälle, die auch in der PCA (Handbuch 1) und der Diskriminanzanalyse (Handbuch 4) auffallen.

Tabelle 6: Silhouetten und WCSS für verschiedene k

	$k = 2$	$k = 3$	$k = 4$
$\bar{s}$ (Ward)	0.366	**0.441**	0.381
WCSS (k-Means)	79.45	43.79	37.31

$k = 3$ liefert den höchsten Silhouetten-Koeffizienten, und der Elbow-Plot von k-Means knickt bei $k = 3$ deutlich ab (WCSS sinkt von 79.45 auf 43.79, danach nur noch auf 37.31). k-Means mit $k = 3$ liefert dieselbe Einteilung wie Ward. Alle Kriterien sprechen also übereinstimmend für drei Unternehmenstypen.

Handhabung des interaktiven Clusteranalyse-Tools

Dateneingabe

Zusaetzlich zu den Standard-Tabs (Manuell, CSV, Beispiel) besitzt das Clusteranalyse-Tool eine **Bezeichnungsspalte** (»Name«) links in der manuellen Eingabetabelle. Hier koennen aussagekraeftige Namen fuer Beobachtungen eingetragen werden, die im Dendrogramm und Streudiagramm sichtbar werden.

Vorgeladen: Bilanzkennzahlen (24 Unternehmen U01–U24, 6 Vars.). Weitere Beispiele: Iris (15 Obs., 4 Vars.), Städteprofil (12 Städte, 4 Vars.) und Kundensegmentierung (20 Kunden, 4 Vars.). Zu jedem Beispiel erscheint über der Eingabetabelle ein Erläuterungskasten mit den Daten, dem Ziel der Analyse und dem erwarteten Ergebnis; er verschwindet bei »Leer« oder beim Laden einer CSV-Datei.

Parameterauswahl

Verfahren
: Hierarchisch: Ward (Standard), Complete Linkage, Average Linkage. Nichthierarchisch: k-Means. Bei Wahl von k-Means wird der Dendrogramm-Tab ausgeblendet.

Distanzmass
: Euklidisch (Standard), Manhattan, Kosinus-Aehnlichkeit. Bei Ward wird die gewaehlte Distanz zwischen den Clusterzentren verwendet; theoretisch begruendet ist Ward nur fuer die euklidische Distanz. k-Means arbeitet stets euklidisch (Auswahl dann gesperrt). Der Silhouetten-Koeffizient verwendet das gewaehlte Distanzmass.

Standardisierung
: Z-Transformation dringend empfohlen bei unterschiedlichen Einheiten.

Clusteranzahl (hierarchisch)
: Bestimmt den Schnitt im Dendrogramm. Die goldene gestrichelte Linie liegt zwischen den beiden Fusionsstufen, an denen geschnitten wird; Aeste unterhalb des Schnitts sind gruen, oberhalb rot.

k und kmax (k-Means)
k: gewuenschte Clusteranzahl. $k_{\max}$: obere Grenze fuer Elbow-Plot und Silhouetten-Analyse (bis zu 12).

Ergebnistabs

Uebersicht
Kennzahlen (n, p, Cluster, Silhouette) + vollstaendige Zuordnungstabelle.

Dendrogramm
Dendrogramm mit Schnittlinie; die Beobachtungen sind in Baumreihenfolge angeordnet und nach zugeordnetem Cluster farbig markiert. Nur bei hierarchischen Verfahren verfuegbar.

Grafiken
k-Means: Elbow-Plot (WCSS) und Silhouetten-Balken je k. Beide Verfahren: Streudiagramm der ersten beiden Variablen (standardisiert), Beobachtungen nach Cluster eingefaerbt.

Clusterprofil
Mittelwerttabelle je Cluster (Originaldaten) + Radar-Chart mit standardisierten Clusterprofilen.

Interpretation
Automatische Bewertung: Silhouetten-Koeffizient, Dendrogramm-Hinweise, k-Means-Elbow-Interpretation.

Export
CSV: Clusterzuordnung, Clusterprofil. PNG: Streudiagramm, Radar-Chart, Dendrogramm.

Typischer Arbeitsablauf

1. Daten laden, Standardisierung pruefen
2. Ward-Methode und euklidische Distanz waehlen
3. Clusteranzahl auf 3 setzen (oder nach Dendrogramm-Inspektion anpassen)
4. »**Clusteranalyse berechnen**« klicken
5. Dendrogramm-Tab: grosse Fusionsdistanzen als Hinweis auf natuerliche Grenzen
6. Silhouette-Wert pruefen ($\bar{s} \geq 0.5$ = gut)
7. Bei Bedarf k-Means mit Elbow-Plot zur Bestaetigung
8. Clusterprofil-Tab: inhaltliche Interpretation der Gruppen

Wichtige Hinweise:
- Mindestens $n = 4$ vollstaendige Beobachtungen erforderlich.
- k-Means: 10 Startwiederholungen (k-Means++) verhindern lokale Minima.
- Ward setzt sphaerische, aehnlich grosse Cluster voraus.

- Bei $p > 2$ visualisiert das Streudiagramm nur die ersten zwei Variablen.
- Dendrogramm: Download als PNG fuer Veroeffentlichungen.

Formelübersicht

Tabelle 7: Formeln der Clusteranalyse

Bezeichnung	**Formel**
Euklidische Distanz	$d_E = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$
Manhattan-Distanz	$d_M = \sum_k \lvert x_{ik} - x_{jk} \rvert$
Ward-Fusionsdistanz	$d_W = \sqrt{\frac{n_i n_j}{n_i + n_j}}\, d_E(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j)$
Kosinus-Distanz	$d_C = 1 - \mathbf{x}_i^\top \mathbf{x}_j / (\lVert\mathbf{x}_i\rVert \, \lVert\mathbf{x}_j\rVert)$
WCSS (k-Means)	$\sum_k \sum_{i \in C_k} \lVert\mathbf{x}_i - \boldsymbol{\mu}_k\rVert^2$
Silhouette	$s_i = (b_i - a_i) / \max(a_i, b_i)$

Weiterführende Literatur

- Kaufman, L. & Rousseeuw, P. J. (2005). *Finding Groups in Data*. Wiley.
- Backhaus, K. et al. (2021). *Multivariate Analysemethoden*, 16. Aufl. Springer Gabler.
- Arthur, D. & Vassilvitskii, S. (2007). k-Means++: The Advantages of Careful Seeding. *SODA 2007*, 1027–1035.