

4.3 Hierarchische Klassifikationsverfahren

Hierarchische Klassifikationsverfahren:

Einsatz zum Zwecke einer Aufdeckung von Clusterstrukturen, wenn keine Kenntnisse über die Gruppenzahl verfügbar sind


Agglomerativen Verfahren:

- Ausgehend von der feinsten Gruppierung einelementiger Cluster werden sukzessive die "ähnlichsten" Klassen bis hin zur größten Gruppierung eines n-elementigen Clusters zusammengefasst,
- Gruppierung wird von Stufe zu Stufe heterogener, da zu den Clustern immer „entferntere“ Objekte hinzukommen

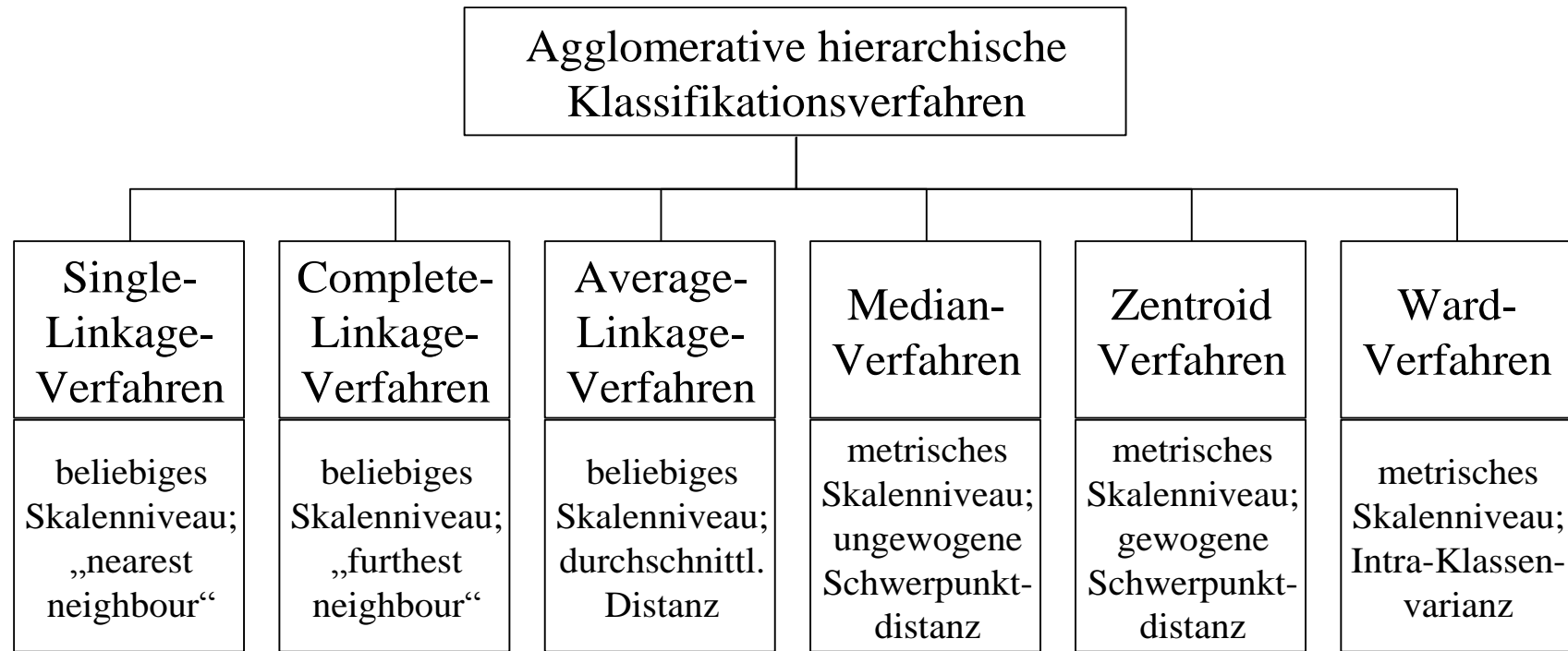
Divisive Verfahren:

Genau umgekehrte Vorgehensweise, d.h. von einem n-elementigen Cluster zu n einelementigen Cluster (keine praktische Bedeutung)

Ablauf einer hierarchischen Klassifikation (Agglomerationsverfahren)

<u>Start</u> : Feinste Partition (n einelementige Cluster)		
Berechnung der Ausgangsdistanz- (Ähnlichkeits-)Matrix		
Ermittlung der beiden Cluster mit der geringsten Distanz (größten Ähnlichkeit)		
Vereinigung der beiden Cluster mit der geringsten Distanz (größten Ähnlichkeit)		
Gibt es nur noch eine Gruppe (= n- elementiges Cluster)?	nein	Neuberechnung der Distanzmatrix (Ähnlichkeitsmatrix)
↓ ja		
<u>Ende</u>		

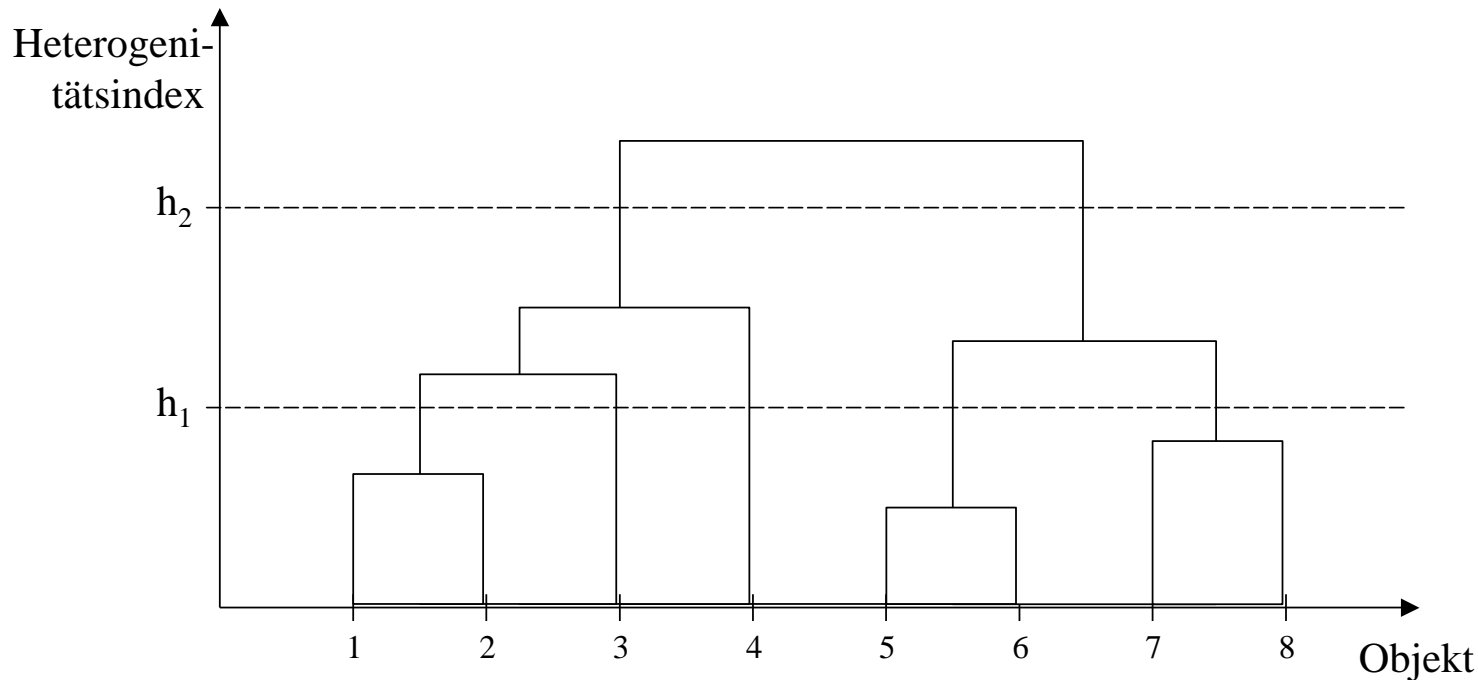
Hierarchische Klassifikationsverfahren



Die Verfahren **Single-Linkage**, **Complete-Linkage** und **Average-Linkage**, die bereits bei nominalskalierten Klassifikationsmerkmalen anwendbar sind, könnten gleichwertig auf der Basis von **Distanz- und Ähnlichkeitsmaßen** eingesetzt werden. Das **Median-** und das **Zentroid-Verfahren** setzen dagegen **metrisch skalierte Merkmale** voraus, da der Homogenitätsverlust im Falle einer Fusion zweier Klassen hierbei anhand des Abstandes der beiden Clusterschwerpunkte gemessen wird. Ein **metrisches Skalenniveau** der Klassifikationsmerkmale setzt auch das **Ward-Verfahren** voraus. Hier erfolgt die Fusion zweier Klassen abweichend zu dem obigen Ablaufschema jedoch auf der Grundlage eines **globalen Heterogenitätskriteriums**. Auf jeder Stufe werden die beiden Cluster fusioniert, deren Zusammenlegung die Streuung innerhalb der Klassen am geringsten erhöht.

Die Ergebnisse einer hierarchischen Klassifikation lassen sich anschaulich in Form eines Baumdiagramms visualisieren, das als **Dendrogramm** bezeichnet wird.

Dendrogramm



Dendrogramm:

- Stufen der hierarchischen Klassifikation anschaulich nachvollziehbar,
- Erkennbar, bei welchem Heterogenitätsgrad eine Fusion zweier Gruppen erfolgt,
- Heterogenitätsgrad z.B. durch die Distanz der beiden zuletzt fusionierten Gruppen oder die Intra-Klassen-Varianz gemessen

Aus der obigen Abbildung geht hervor, dass bei einem **Heterogenitätsindex** h_1 drei Zweiergruppen mit den Objekten 5 und 6, 1 und 2 sowie 7 und 8 neben zwei ein-₄elementigen Gruppen der Objekte 3 und 4 bestehen.

Dagegen verringert sich die Clusterzahl bei einem Heterogenitätsindex h_2 auf zwei: Das erste Cluster setzt sich aus den Objekten 1, 2, 3 und 4 zusammen, während die Objekte 5, 6, 7 und 8 das zweite Cluster bilden.

Ein **sprunghafter Anstieg des Heterogenitätsindex** spiegelt eine starke Abnahme der Ähnlichkeit der Objekte einer Klassifikation wider. In der Abbildung zeigt sich ein starker Anstieg des Heterogenitätsindex nach Bildung der beiden Vierergruppen. Das Dendrogramm würde daher hier eine Clusterzahl von zwei nahe legen.

Single-Linkage-Verfahren

Beim **Single-Linkage-Verfahren** ist die Distanz D zwischen zwei Clustern C_g und C_h durch die kleinste Distanz zwischen zwei Objekten i und j der beiden Cluster definiert:

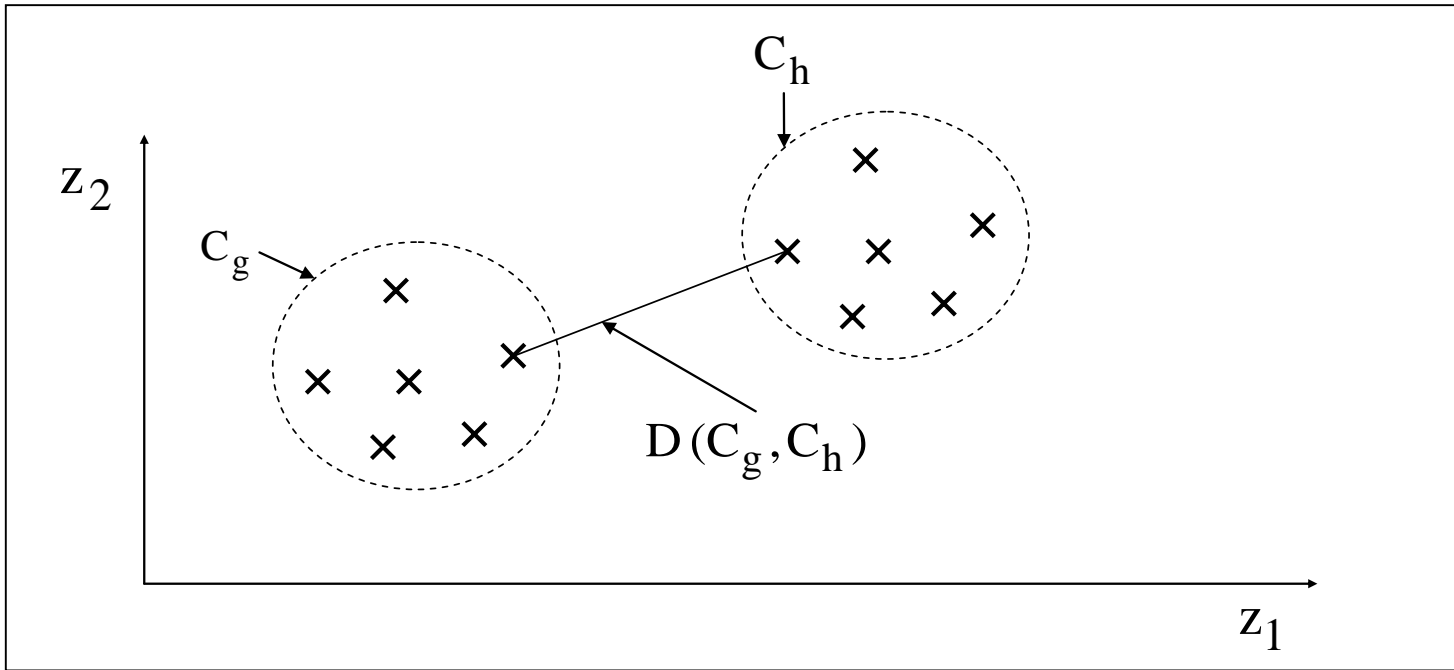
$$(4.10) \quad D(C_g, C_h) = \min \{d(i, j)\}, i \in C_g, j \in C_h$$

Aufgrund dieser Art der Festlegung der Clusterdistanzen spricht man von einer **Nearest-Neighbour-Methode**. Auf jeder Stufe werden die Clusterdistanzen aufgrund von Gleichung (4.10) bestimmt. Es werden dann stets die beiden Cluster r und s fusioniert, für die die Clusterdistanz minimal ist:

$$(4.11) \quad D(C_r, C_s) = \min \{D(C_g, C_h)\}, g \neq h$$

\Rightarrow Fusion der Cluster C_r und C_s .

Abbildung 4.4: Single-Linkage-Verfahren im Zwei-Variablen-Fall



Beispiel 4.11: Um die hierarchische Klassifikation unter Anwendung des Single-Linkage-Verfahrens aufzuzeigen, gehen wir von den Regionen A, B, C und D aus, deren Ähnlichkeiten durch die euklidische Distanz gemessen werden. Aufgrund der Symmetrie geben wir nur die untere Dreiecksmatrix wieder:

$$(4.12) \quad \mathbf{D} = \begin{bmatrix} 0 & & & \\ 4,438 & 0 & & \\ 3,084 & 6,777 & 0 & \\ 2,259 & 2,887 & 4,339 & 0 \end{bmatrix} \begin{matrix} A \\ B \\ C \\ D \end{matrix}$$

Die Ausgangspartition besteht aus vier Clustern, die die einzelnen Regionen A, B, C und D enthalten:

Ausgangspartition: $C_1 = \{A\}$, $C_2 = \{B\}$, $C_3 = \{C\}$, $C_4 = \{D\}$.

Stufe 1

In Stufe 1 entspricht die niedrigste Distanz zwischen den Clustern exakt der geringsten Objektdistanz. Aus der Distanzmatrix **D** ist erkennbar, dass dies die Distanz zwischen den Objekten A ($\hat{=}$ Cluster C1) und D ($\hat{=}$ Cluster C4) ist:

$$D(C_1, C_4) = d(A, D) = 2,259$$

Aus diesem Grund werden die beiden Objekte A und D zu einem Cluster vereinigt, so dass sich die

Partition (1. Stufe): $C_1 = \{A, D\}$, $C_2 = \{B\}$, $C_3 = \{C\}$,

ergibt.

Zu dieser Partition geben wir die Distanzmatrix an, wozu wir die neuen Clusterdistanzen ermitteln:

$$D(C_1, C_2) = \min\{d(A, B) = 4,438, d(D, B) = 2,887\} = d(D, B) = 2,887$$

$$D(C_1, C_3) = \min\{d(A, C) = 3,084, d(D, C) = 4,339\} = d(A, C) = 3,084$$

$$D(C_2, C_3) = d(B, C) = 6,777$$

Die Distanzmatrix für die Partition der 1. Stufe lautet somit

$$\mathbf{D} = \begin{array}{ccc} & \begin{array}{c} C_1 \quad C_2 \quad C_3 \end{array} \\ \begin{array}{c} C_1 : A, D \\ C_2 : B \\ C_3 : C \end{array} & \begin{bmatrix} 0 & & \\ 2,887 & 0 & \\ 3,084 & 6,777 & 0 \end{bmatrix} \end{array}.$$

2. Stufe:

Erneut sind die beiden Cluster mit der geringsten Distanz zu bestimmen. Man erkennt anhand der Distanzmatrix der 1. Stufe, dass die Distanz zwischen den Clustern C_1 und C_2 mit einem Distanzwert von 2,887 minimal ist, weshalb diese beiden Cluster vereinigt werden. Damit ergibt sich die

Partition (2. Stufe): $C_1 = \{A, B, D\}$, $C_2 = \{C\}$

als Ergebnis des Fusionsprozesses der zweiten Stufe.

Wir berechnen die Distanz zwischen den beiden Clustern C_1 und C_2 :

$$D(C_1, C_2) = \min\{d(A, C) = 3,084, d(B, C) = 6,777, d(D, C) = 4,339\} = d(A, C) = 3,084$$

und erhalten die Distanzmatrix

$$\mathbf{D} = \begin{array}{cc} & \begin{array}{c} C_1 \quad C_2 \end{array} \\ \begin{array}{c} C_1 : A, B, D \\ C_2 : C \end{array} & \begin{bmatrix} 0 & \\ 3,084 & 0 \end{bmatrix} \end{array}.$$

3. Stufe:

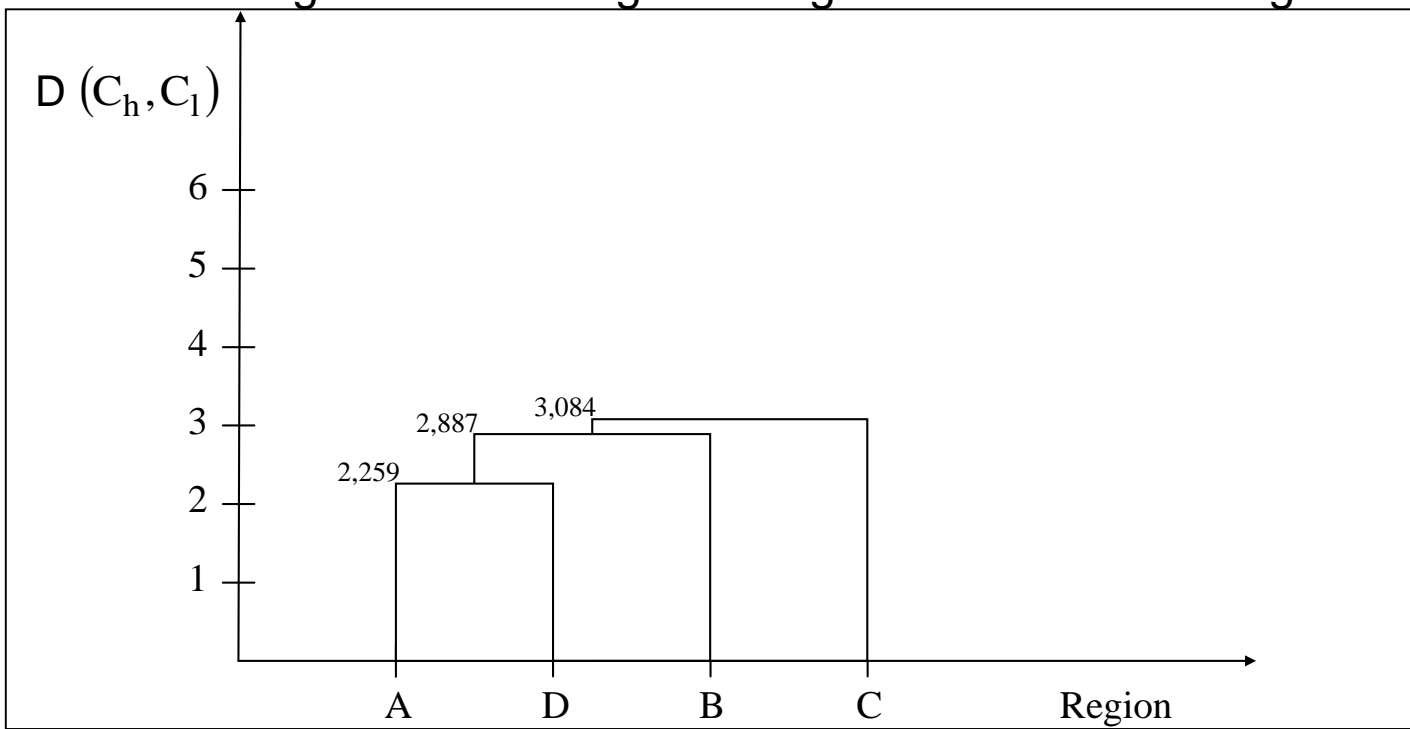
In der 3. Stufe werden schließlich noch die beiden verbliebenen Cluster C_1 und C_2 bei einer Distanz von $D(C_1, C_2) = 3,084$ zu einem Cluster vereinigt:

Partition (3. Stufe): $C_1 = \{A, B, C, D\}$

Da es nur noch ein Cluster mit allen Objekten gibt, wird der Gruppierungsprozess beendet.

Die Ergebnisse der hierarchischen Klassifikation auf der Basis des Single-Linkage-Verfahrens lassen sich auch durch das in der folgenden Abbildung wiedergegebene **Dendrogramm** transparent machen.

Abb.: Dendrogramm des Single-Linkage-Verfahrens der Regionen A, B, C, D



Single-Linkage-Verfahren mit SPSS

Wir wollen die manuell mit dem Single-Linkage-Verfahren durchgeführte hierarchische Klassifikation der 4 Regionen A, B, C und D nun mit SPSS ausführen. Hierzu legen wir eine verkleinerte SPSS-Datendatei für die 4 Regionen an, wobei wir unsere Daten vorher durch Wahl der Menüpunkte

Analysieren

Deskriptive Statistiken

Deskriptive Statistiken...

standardisieren. Wir speichern dann ausschließlich die standardisierten Merkmalswerte für die 4 Regionen A, B, C und D zusammen mit der Variablen Region (A, B, C, D) in der Datendatei Regionen(Z4).sav. Dort haben die standardisierten Variablen das Präfix Z, also Zed, Zbip, etc.

Hinweis: Um eine hierarchische Klassifikation durchführen zu können, wird in SPSS keine standardisierte Datendatei benötigt. SPSS kann die Variablen auch innerhalb der hierarchischen Klassifikationsprozedur standardisieren. Wir verwenden eine standardisierte Datendatei für eine Teilmenge von Objekten (Regionen), um die Objekte mit den für den gesamten Datensatz gültigen standardisierten Werten zu klassifizieren.

Der Aufruf der hierarchischen Klassifikation erfolgt in SPSS über die Menüpunkte

Analysieren

Klassifizieren

Hierarchische Cluster....

Im Fenster „Hierarchische Clusteranalyse“ bringen wir die z-Werte der Variablen in das Feld „Variablen“ und die Variable Region in das Feld „Fallbeschriftung“. Wir betätigen die Schaltfläche „Statistik“ und versehen das Item „Distanzmatrix“ mit einem Haken. Mit der Schaltfläche „Diagramm“ gelangen wir in das Fenster „Hierarchische Clusteranalyse: Diagramme“. Dort geben wir im Feld „Eiszapfendiagramm“ „keine“ an und wählen das Item „Dendrogramm“. Im Methoden-Fenster wählen wir mit dem Pull-down-Menü die „Cluster-Methode“ „Nächstgelegener Nachbar“ (=Single-Linkage-Verfahren). Im Feld „Maß“ wählen wir aus dem Pull-down-Menü „Euklidische Distanz“.

Als Ausgabe erhalten wir im SPSS-Viewer verschiedene Tabellen. Nach einer Tabelle über die Anzahl der verarbeiteten Fälle wird die von SPSS bezeichnete „Näherungsmatrix“ (=Distanzmatrix) ausgegeben:

Näherungsmatrix

Fall	Euklidisches Distanzmaß			
	1:A	2:B	3:C	4:D
1:A	.000	4.438	3.084	2.259
2:B	4.438	.000	6.777	2.887
3:C	3.084	6.777	.000	4.339
4:D	2.259	2.887	4.339	.000

Dies ist eine Unähnlichkeitsmatrix

Sieht man einmal davon ab, dass wir stets nur die untere Dreiecksmatrix verwendet haben, stimmen beide Matrizen überein.

Anschließend gibt SPSS die Tabelle „Zuordnungsübersicht“ aus. Hierbei ist zu beachten, dass SPSS die Cluster stets mit der kleinsten Nummer des Objekts kennzeichnet, das ihm angehört. In der Spalte „Koeffizienten“ wird diejenige Distanz ausgewiesen, zu der eine Verschmelzung der beiden links daneben stehenden Cluster stattfindet.

Zuordnungsübersicht

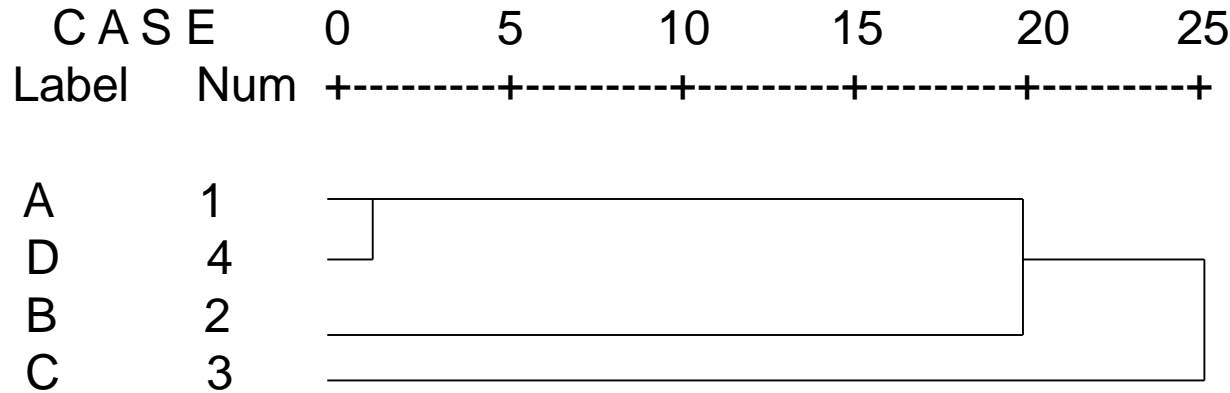
Schritt	Zusammengeführte Cluster		Koeffizienten	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	4	2.259	0	0	2
2	1	2	2.887	1	0	3
3	1	3	3.084	2	0	0

Während in der „Zuordnungsübersicht“ die tatsächlichen Distanzwerte ausgewiesen werden, normiert SPSS die Clusterdistanzen beim Dendrogramm auf den Wertebereich [0; 25].

***** HIERARCHICAL CLUSTER ANALYSIS *****

Dendrogram using Single Linkage

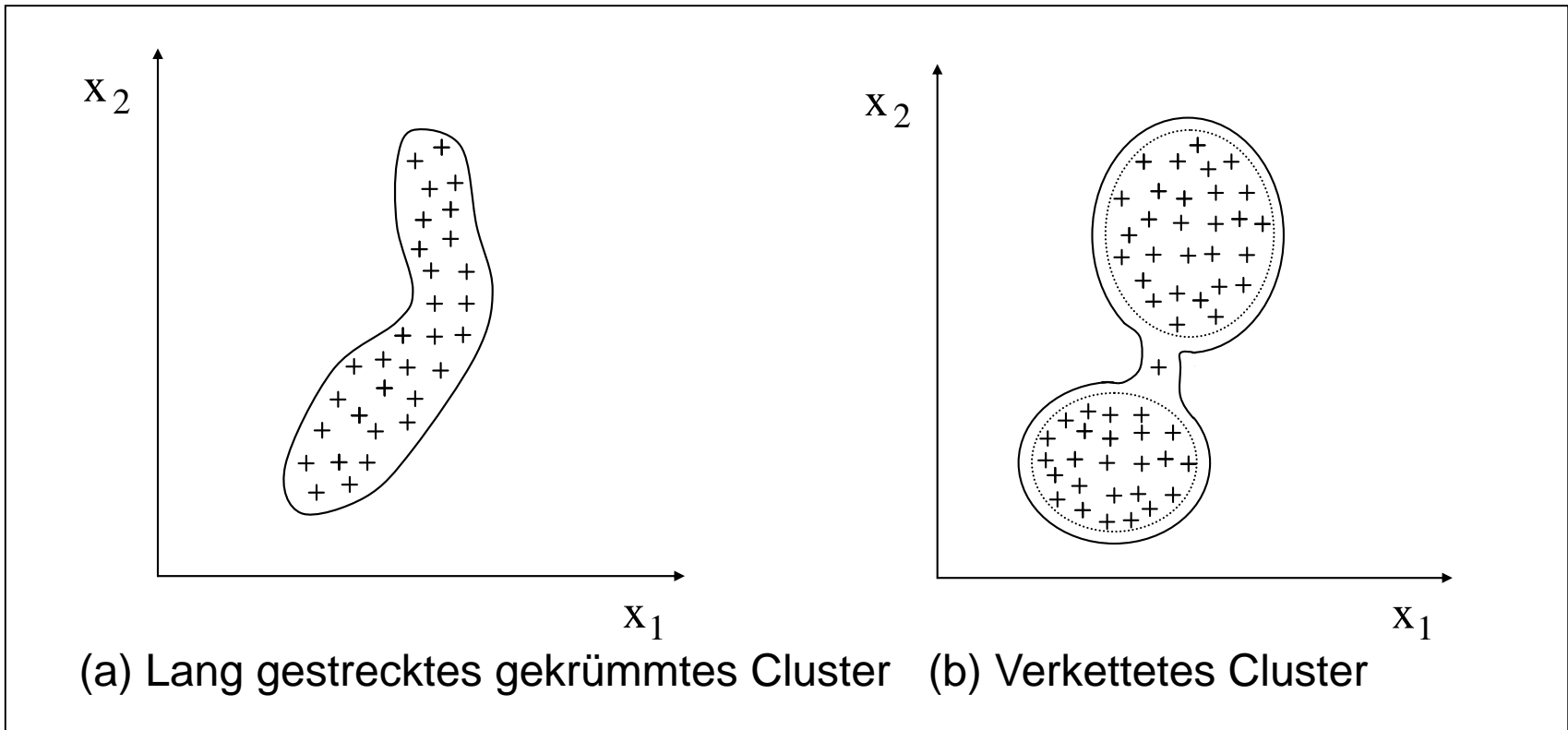
Rescaled Distance Cluster Combine



Eigenschaften des Single-Linkage-Verfahrens:

- Geeignet verzweigte, gekrümmte oder lang gestreckte Cluster zu "erkennen", da es genügt, dass ein Objekt einer Klasse nahe bei einem Objekt einer anderen Klasse liegt
- Gruppen werden zusammengefasst, die nur durch eine "Brücke" miteinander verbunden sind, ansonsten aber deutlich separiert voneinander im Raum liegen (**kontrahierend**) → **Verkettungseffekt (chaining effect)**, der zu außerordentlich **heterogenen Clustern** führen kann
- Monotonieeigenschaft (Clusterdistanz nimmt von Stufe zu Stufe zu)

Abbildung 4.5: Identifikation von Clustern mittels des Single-Linkage-Verfahrens



Hauptsächliche Anwendung des Single-Linkage-Verfahrens:
Aufdeckung von Ausreißern [Objekte, die auf einer höheren Stufe des Klassifikationsprozesses trotz der Neigung des Verfahrens zur Bildung weniger großer Cluster (kontrahierendes Verfahren) noch unklassiert geblieben sind]

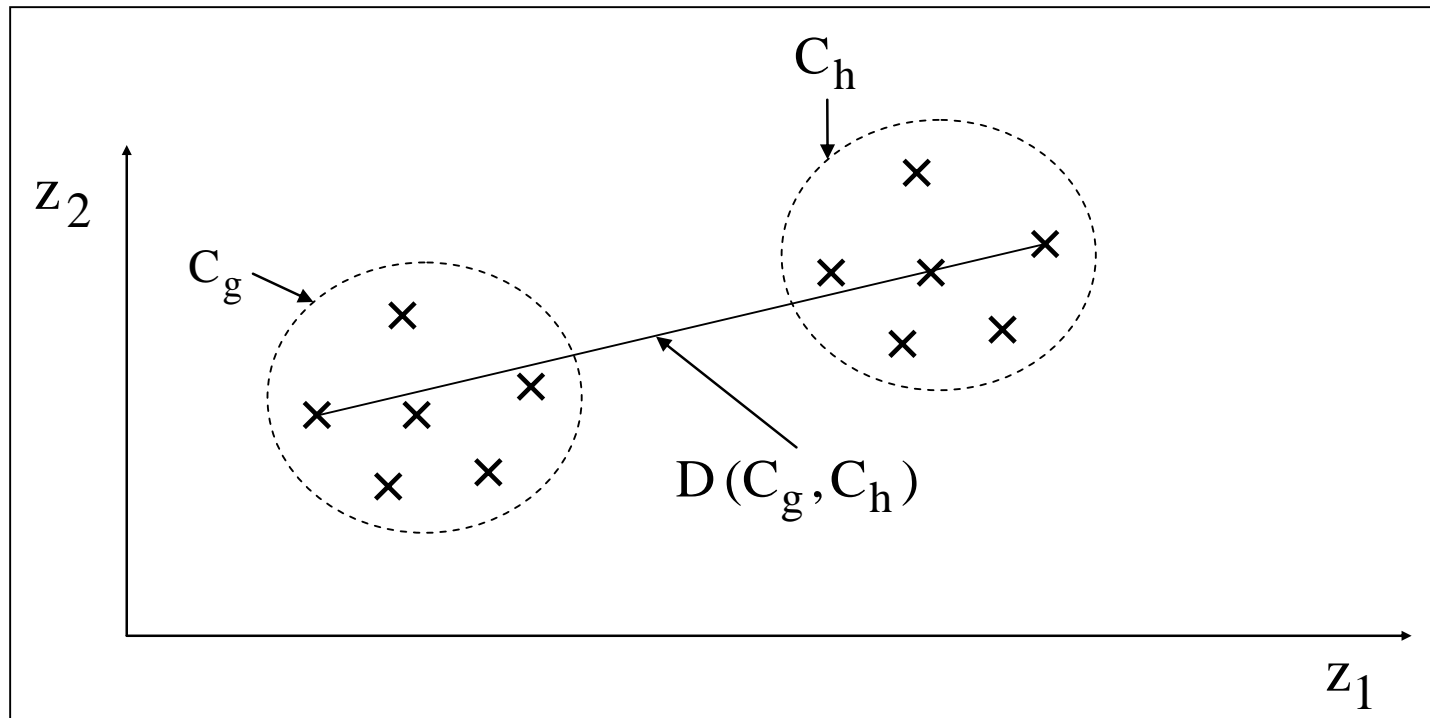
Complete-Linkage-Verfahren

Das **Complete-Linkage-Verfahren** geht bei der Messung der Clusterdistanzen von den beiden entferntesten Objekten (**Furthest-Neighbour-Methode**) aus. Die Distanz zwischen den beiden Clustern C_g und C_h ist hierin demzufolge durch

$$(4.13) \quad D(C_g, C_h) = \max \{d(i, j)\}, i \in C_g, j \in C_h$$

definiert. Sofern die Clusterdistanzen auf einer Stufe des Klassifikationsprozesses durch (4.13) ermittelt worden sind, erfolgt eine Fusion der beiden Cluster mit der minimalen Distanz gemäß der Regel (4.11).

Abbildung 4.6: Complete-Linkage-Verfahren im Zwei-Variablen



Beispiel 4.12: Die Arbeitsweise des Complete-Linkage-Verfahrens lässt sich wiederum anhand des vereinfachten Regionenbeispiels unter Verwendung der Distanzmatrix aufzeigen:

$$D = \begin{array}{c} \begin{array}{cccc} & A & B & C & D \\ \begin{bmatrix} 0 & 4,438 & 3,084 & 2,259 \\ & 0 & 6,777 & 2,887 \\ & & 0 & 4,339 \\ & & & 0 \end{bmatrix} & \begin{array}{l} A \\ B \\ C \\ D \end{array} \end{array} \end{array}$$

Ausgangspartition: $C_1 = \{A\}$, $C_2 = \{B\}$, $C_3 = \{C\}$, $C_4 = \{D\}$.

Stufe 1

Da die Ausgangspartition aus einelementigen Clustern besteht, sind die Clusterdistanzen stets mit den in der Distanzmatrix wiedergegebenen Objektdistanzen identisch, so dass der Regel (4.11) zufolge das Cluster 1 (Region A) mit dem Cluster 4 (Region D) bei einem Distanzwert von 2,259 zu verschmelzen ist:

Partition (1. Stufe): $C_1 = \{A, D\}$, $C_2 = \{B\}$, $C_3 = \{C\}$

Wir berechnen die Distanzen zwischen den drei Clustern nach dem Complete-Linkage-Verfahren,

$$D(C_1, C_2) = \max \{d(A, B) = 4,438, d(D, B) = 2,887\} = d(A, B) = 4,438,$$

$$D(C_1, C_3) = \max \{d(A, C) = 3,084, d(D, C) = 4,339\} = d(D, C) = 4,339,$$

$$D(C_2, C_3) = d(B, C) = 6,777$$

und erhalten die Distanzmatrix

$$\mathbf{D} = \begin{array}{ccc} & \begin{matrix} C_1 & C_2 & C_3 \end{matrix} \\ \begin{bmatrix} 0 & & \\ 4,438 & 0 & \\ 4,339 & 6,777 & 0 \end{bmatrix} & \begin{matrix} C_1 : A, D \\ C_2 : B \\ C_3 : C \end{matrix} \end{array}.$$

2. Stufe:

Das Minimum der Furthest-Neighbour-Distanzen liegt bei einem Wert von 4,339, der die Distanz zwischen den Clustern C_1 und C_3 wiedergibt. Eine Fusion dieser beiden Cluster führt zu der

Partition (2. Stufe): $C_1 = \{A, C, D\}$, $C_2 = \{B\}$

die sich von der durch das Single-Linkage-Verfahren erzeugten Partition der zweiten Stufe unterscheidet.

Mit der Distanz zwischen den beiden Clustern C_1 und C_2 von

$$D(C_1, C_2) = \max \{d(A, B) = 4,438, d(C, B) = 6,777, d(D, B) = 2,887\} = d(C, B) = 6,777$$

erhalten wir die Distanzmatrix

$$\mathbf{D} = \begin{array}{cc} & \begin{matrix} C_1 & C_2 \end{matrix} \\ \begin{bmatrix} 0 & \\ 6,777 & 0 \end{bmatrix} & \begin{matrix} C_1 : A, C, D \\ C_2 : B \end{matrix} \end{array}.$$

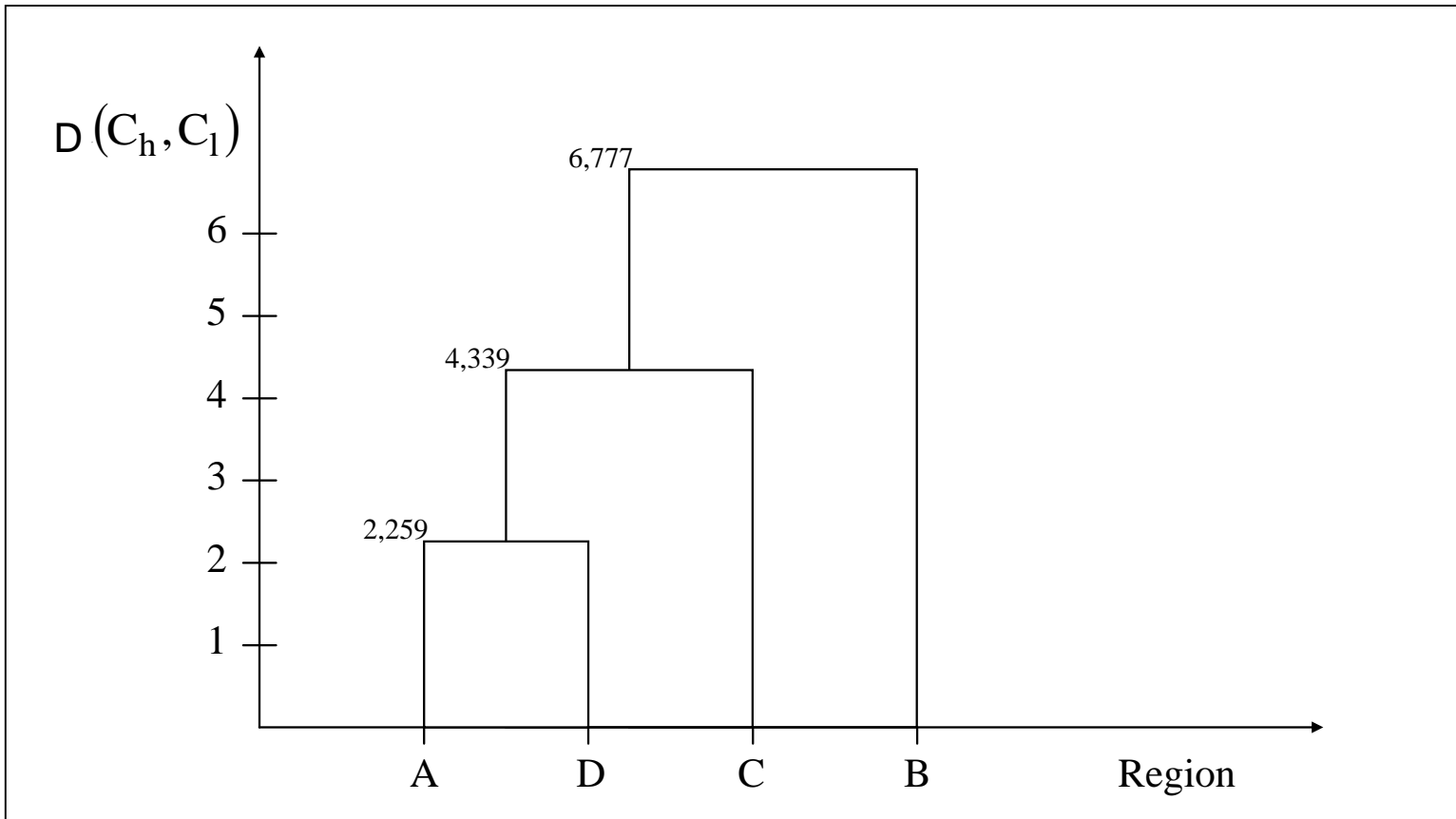
3. Stufe:

In der 3. Stufe werden wiederum die beiden noch verbliebenen Cluster $C_1 = \{A, C, D\}$ und $C_2 = \{B\}$ zu einem Cluster vereinigt:

Partition (3. Stufe): $C_1 = \{A, B, C, D\}$

Da es nur noch ein Cluster mit allen Objekten gibt, wird der Gruppierungsprozess beendet.

Abb.: Dendrogramm des Complete-Linkage-Verfahrens der Regionen A, B, C und D



Complete-Linkage-Verfahren mit SPSS

Wir führen dieselben Einstellungen wie beim Single-Linkage-Verfahren, wählen jedoch im Methoden-Fenster die „Cluster-Methode“ „Entferntester Nachbar“ (=Complete-Linkage-Verfahren).

Die im SPSS-Viewer ausgegebenen Tabellen und Grafiken lassen sich analog zu denen des Single-Linkage-Verfahrens interpretieren.

Näherungsmatrix

Fall	Euklidisches Distanzmaß			
	1:A	2:B	3:C	4:D
1:A	.000	4.438	3.084	2.259
2:B	4.438	.000	6.777	2.887
3:C	3.084	6.777	.000	4.339
4:D	2.259	2.887	4.339	.000

Dies ist eine Unähnlichkeitsmatrix

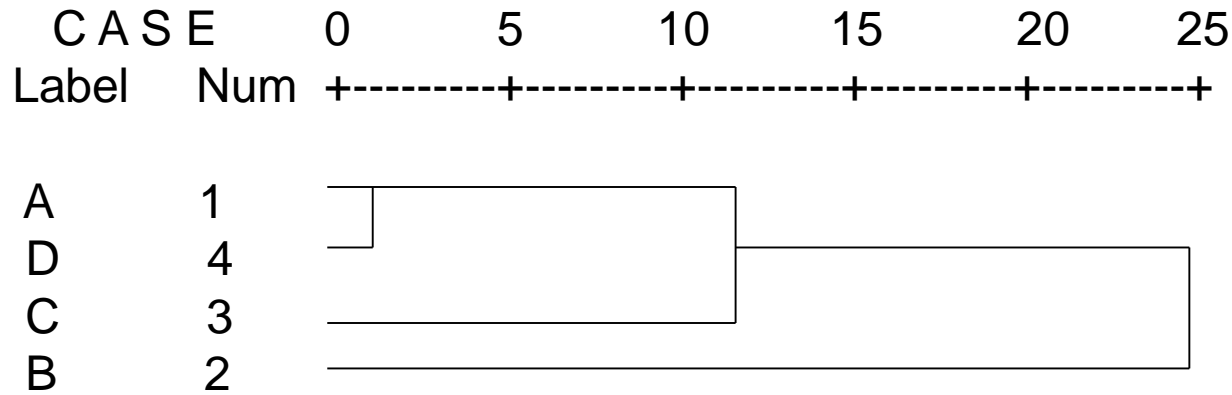
Zuordnungsübersicht

Schritt	Zusammengeführte Cluster		Koef fizienten	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	4	2.259	0	0	2
2	1	3	4.339	1	0	3
3	1	2	6.777	2	0	0

***** HIERARCHICAL CLUSTER ANALYSIS ***

Dendrogram using Complete Linkage

Rescaled Distance Cluster Combine



Eigenschaften des Complete-Linkage-Verfahrens:

- Tendenz zur Bildung kleiner, kompakter Gruppen (**dilatierendes Verfahren**), die häufig in sich erheblich homogener sein werden
- Die Orientierung an den beiden maximal unähnlichsten Objekten kann dazu führen, dass eine Fusion zweier Cluster unterbleibt, auch wenn die mittlere Distanz – zwischen den Objekten nicht notwendig eine merkliche Erhöhung der Heterogenität anzeigen würde.
- Monotonieeigenschaft

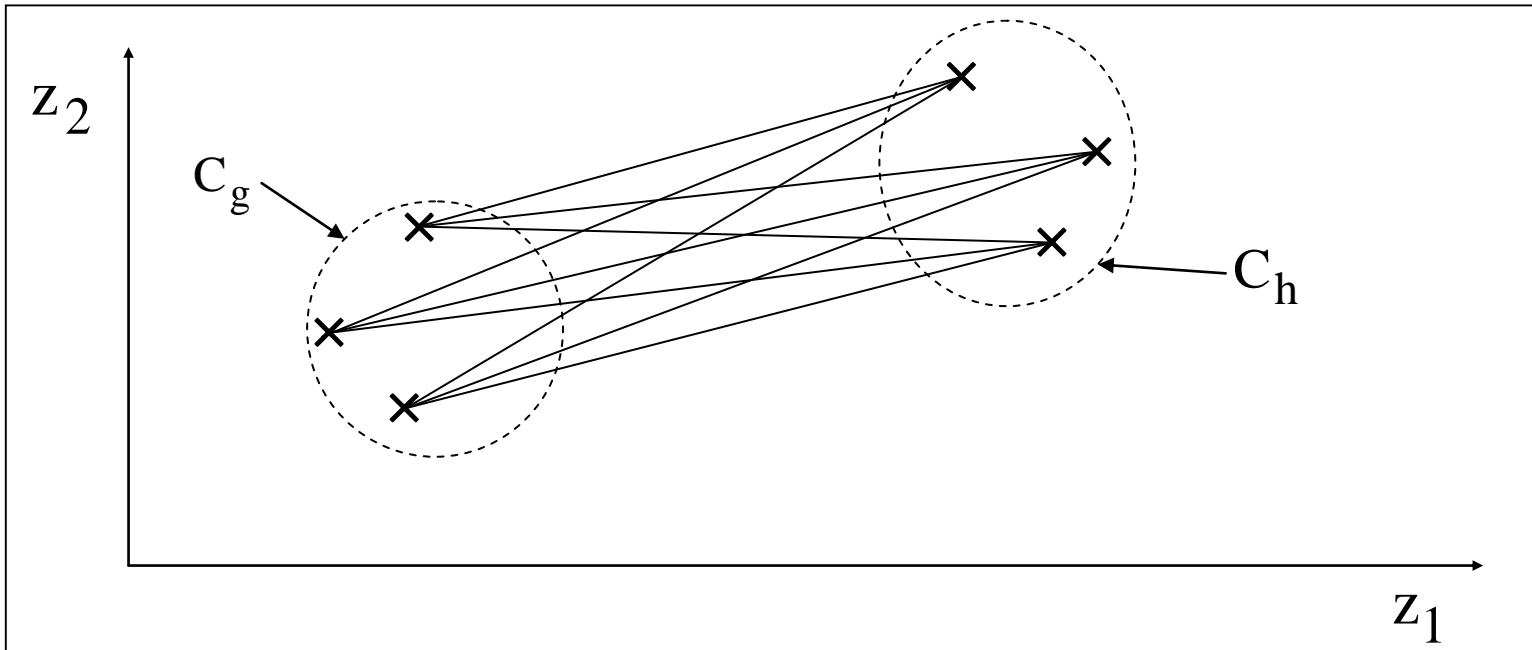
Average-Linkage-Verfahren

Die Distanz zwischen zwei Clustern C_g und C_h entspricht beim **Average-Linkage-Verfahren** dem arithmetischen Mittel der Distanzen zwischen den Objekten der Cluster C_g und C_h :

$$(4.14) \quad D(C_g, C_h) = \frac{1}{n_g \cdot n_h} \sum_{i \in C_g} \sum_{j \in C_h} d(i, j)$$

Hierbei geben n_g und n_h die Anzahl der in den Clustern C_g und C_h enthaltenen Objekte wieder.

Abbildung 4.7: Alle Objektdistanzen im Zwei-Cluster-Fall



Beispiel 4.13: Ausgegangen wird wiederum von der Distanzmatrix:

$$\mathbf{D} = \begin{array}{c} \begin{array}{cccc} & A & B & C & D \end{array} \\ \begin{array}{l} \left[\begin{array}{cccc} 0 & & & \\ 4,438 & 0 & & \\ 3,084 & 6,777 & 0 & \\ 2,259 & 2,887 & 4,339 & 0 \end{array} \right] \end{array} \begin{array}{l} A \\ B \\ C \\ D \end{array} \end{array}$$

mit der Ausgangspartition:

Ausgangspartition: $C_1 = \{A\}$, $C_2 = \{B\}$, $C_3 = \{C\}$, $C_4 = \{D\}$.

1. Stufe:

Bei gleicher Vorgehensweise wie beim Single-Linkage- bzw. Complete-Linkage-Verfahren erhalten wir die

Partition (1. Stufe): $C_1 = \{A, D\}$, $C_2 = \{B\}$, $C_3 = \{C\}$

Während die Distanz zwischen den Clustern C_2 und C_3 auf der zweiten Stufe unverändert 6,777 bleibt, verändern sich die Distanzen zwischen C_1 und C_2 sowie C_1 und C_3 ($n_1 = 2$, $n_2 = 1$, $n_3 = 1$):

$$D(C_1, C_2) = \frac{1}{n_1 \cdot n_2} [d(A, B) + d(D, B)] = \frac{1}{2} (4,438 + 2,887) = 3,663$$

und

$$D(C_1, C_3) = \frac{1}{n_1 \cdot n_3} [d(A, C) + d(D, C)] = \frac{1}{2} (3,084 + 4,339) = 3,712$$

Die zur Partition der 1. Stufe gehörende Distanzmatrix ist daher von der Form

$$\mathbf{D} = \begin{array}{ccc} & \begin{matrix} C_1 & C_2 & C_3 \end{matrix} \\ \begin{matrix} C_1 : A, D \\ C_2 : B \\ C_3 : C \end{matrix} & \begin{bmatrix} 0 & & \\ 3,663 & 0 & \\ 3,712 & 6,777 & 0 \end{bmatrix} & \end{array}.$$

2. Stufe:

Aufgrund der minimalen Clusterdistanz von 3,663 sind die Cluster C_1 und C_2 zu fusionieren:

Partition (2. Stufe): $C_1 = \{A, B, D\}, C_2 = \{C\}$

Die beiden verbleibenden Cluster weisen mit $n_1 = 3$ und $n_2 = 1$ eine mittlere Distanz von $D(C_1, C_2) = \frac{1}{n_1 \cdot n_2} [d(A, C) + d(B, C) + d(D, C)] = \frac{1}{3} (3,084 + 6,777 + 4,339) = 4,733$

auf, womit sich die Distanzmatrix

$$\mathbf{D} = \begin{array}{cc} & \begin{matrix} C_1 & C_2 \end{matrix} \\ \begin{matrix} C_1 : A, B, D \\ C_2 : C \end{matrix} & \begin{bmatrix} 0 & \\ 4,733 & 0 \end{bmatrix} \end{array}$$

ergibt.

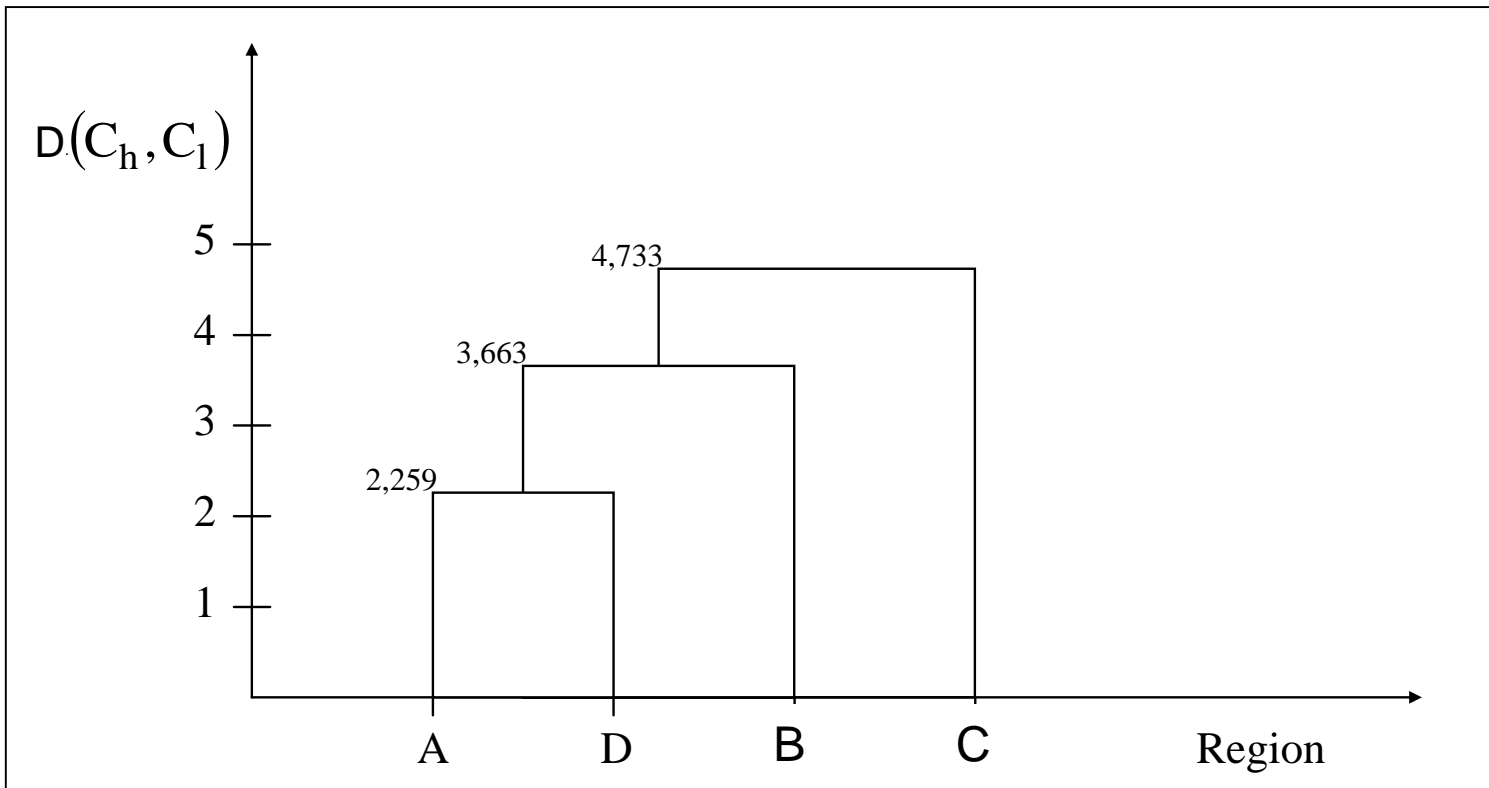
3. Stufe:

In der 3. Stufe werden erneut die beiden noch verbliebenen Cluster $C_1 = \{A, B, D\}$ und $C_2 = \{C\}$ zu einem Cluster vereinigt:

Partition (3. Stufe): $C_1 = \{A, B, C, D\}$

Da es nur noch ein Cluster mit allen Objekten gibt, wird der Gruppierungsprozess beendet.

Abb.: Dendrogramm des Average-Linkage- Verfahrens der Regionen A, B, C und D



Average-Linkage-Verfahren mit SPSS

Hierzu wählen wir im Methoden-Fenster die „Cluster-Methode“ „Linkage zwischen den Gruppen“ (=Average-Linkage-Verfahren).

Näherungsmatrix

Fall	Euklidisches Distanzmaß			
	1:A	2:B	3:C	4:D
1:A	.000	4.438	3.084	2.259
2:B	4.438	.000	6.777	2.887
3:C	3.084	6.777	.000	4.339
4:D	2.259	2.887	4.339	.000

Dies ist eine Unähnlichkeitsmatrix

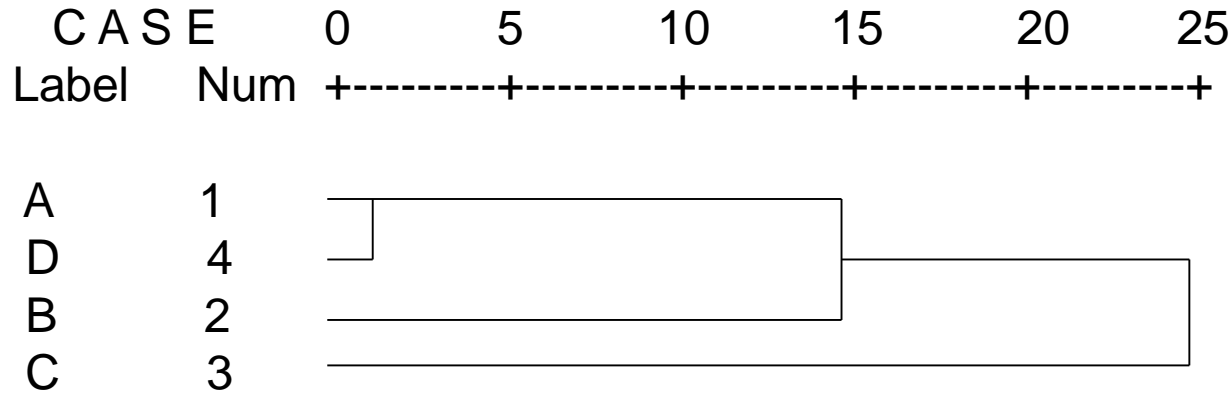
Zuordnungsübersicht

Schritt	Zusammengeführte Cluster		Koef fizienten	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	4	2.259	0	0	2
2	1	2	3.662	1	0	3
3	1	3	4.733	2	0	0

***** HIERARCHICAL CLUSTER ANALYSIS *****

Dendrogram using Average Linkage (Between Groups)

Rescaled Distance Cluster Combine



Eigenschaft des Average-Linkage-Verfahrens:

- Konservatives Verfahren, das zwischen dem kontrahierenden Single-Linkage- Verfahren und dem dilatierenden Complete-Linkage-Verfahren eingeordnet werden kann,
- Objekte zweier Gruppen müssen "im Mittel" ähnlich sein müssen, damit es zu einer Fusion kommt. Größere Distanzen zwischen Objekten können hierbei durch geringere Distanzen nahe beieinander liegender Objekte kompensiert werden.
- Monotonieeigenschaft.

Ward-Verfahren

Beim Ward-Verfahren werden nicht wie bei den bisher behandelten hierarchischen Verfahren die Cluster mit der geringsten Distanz zueinander vereinigt. Vielmehr erfolgt die Fusion von Clustern auf der Grundlage eines **Varianzkriteriums**. Hierbei werden stets metrisch skalierte Merkmale vorausgesetzt.

Die Summe der Abweichungsquadrate der (standardisierten) Beobachtungswerte z_{ik} des Clusters C_g von den Merkmalsmittelwerten \bar{z}_{gk} ,

$$(4.15) \quad V_g = \sum_{k=1}^m \sum_{i \in C_g} (z_{ik} - \bar{z}_{gk})^2$$

gibt die Streuung innerhalb des g-ten Clusters wieder. Hierbei werden die Merkmalsmittelwerte \bar{z}_{gk} aus den (standardisierten) Beobachtungen berechnet, die zum Cluster C_g gehören:

$$(4.16) \quad \bar{z}_{gk} = \frac{1}{n_g} \sum_{i \in C_g} z_{ik}$$

Die Gesamtstreuung innerhalb der G Cluster einer vorliegenden Partition ist dann durch

$$(4.17) \quad V = \sum_{g=1}^G \sum_{k=1}^m \sum_{i \in C_g} (z_{ik} - \bar{z}_{gk})^2$$

gegeben. Mit jeder Fusion geht ein Homogenitätsverlust der Klassifikation in Form einer Steigerung der Streuung innerhalb der Klassen (within-groups sum of squares) einher.

Bei Anwendung des **Varianzkriteriums der Ward-Methode** werden in jeder Stufe des Fusionsprozesses stets die **beiden Cluster fusioniert**, die zu einer **minimalen Erhöhung der Gesamtstreuung V** führen.

Wie sich zeigt, kann die Erhöhung der Kriteriumsgröße V im Falle einer Fusion des Cluster C_g und C_h mittels des Ausdrucks

$$(4.18) \quad \Delta V(C_g \cup C_h) = \frac{n_g \cdot n_h}{n_g + n_h} \sum_{k=1}^m (\bar{z}_{gk} - \bar{z}_{hk})^2$$

bestimmt werden.

Auf jeder Stufe des Klassifikationsprozesses sind für alle Clusterpaare die Zuwächse ΔV zu berechnen. **Vereinigt** wird auf einer bestimmten Stufe jeweils das **Clusterpaar mit dem geringsten ΔV -Wert**.

Beispiel 4.14: In unserem Beispiel liegen für die Regionen A, B, C und D folgende standardisierte Merkmalswerte für die Variablen Einwohnerdichte (X_1) und BIP (X_2) vor:

$$\mathbf{Z} = \begin{array}{cc} & \begin{array}{cc} \text{ED} & \text{BIP} \end{array} \\ \left[\begin{array}{cc} -0,657 & -1,245 \\ 1,709 & 1,254 \\ -1,343 & -1,653 \\ -0,516 & 0,299 \end{array} \right] & \begin{array}{l} \text{A} \\ \text{B} \\ \text{C} \\ \text{D.} \end{array} \end{array}$$

Die Ausgangspartition lautet wieder:

Ausgangspartition: $C_1 = \{A\}$, $C_2 = \{B\}$, $C_3 = \{C\}$, $C_4 = \{D\}$.

Wir berechnen für die Ausgangspartition die Clustermittelwerte:

Cluster $C_1 = \{A\}$: $\bar{z}_{11} = -0,657$, $\bar{z}_{12} = -1,245$

Cluster $C_2 = \{B\}$: $\bar{z}_{21} = 1,709$, $\bar{z}_{22} = 1,254$

Cluster $C_3 = \{C\}$: $\bar{z}_{31} = -1,343$, $\bar{z}_{32} = -1,653$

Cluster $C_4 = \{D\}$: $\bar{z}_{41} = -0,516$, $\bar{z}_{42} = 0,299$

1. Stufe:

Nach Gleichung (4.18) würde der Zuwachs der Kriteriumsgröße V im Falle einer Fusion der beiden Cluster C_1 und C_2

$$\Delta V(C_1 \cup C_2) = \frac{1 \cdot 1}{1 + 1} \left\{ [(-0,657) - 1,709]^2 + [(-1,245) - 1,254]^2 \right\} = 5,921$$

betragen. Dagegen würde sich das Varianzkriterium bei einer Fusion der Cluster C_1 und C_3 nur um 0,319 erhöhen:

$$\Delta V(C_1 \cup C_3) = \frac{1 \cdot 1}{1 + 1} \left\{ [(-0,657) - (-1,343)]^2 + [(-1,245) - (-1,653)]^2 \right\} = 0,319$$

Entsprechend erhält man für die übrigen Clusterpaare die ΔV -Werte

$$\Delta V(C_1 \cup C_4) = 1,202 \text{ ,}$$

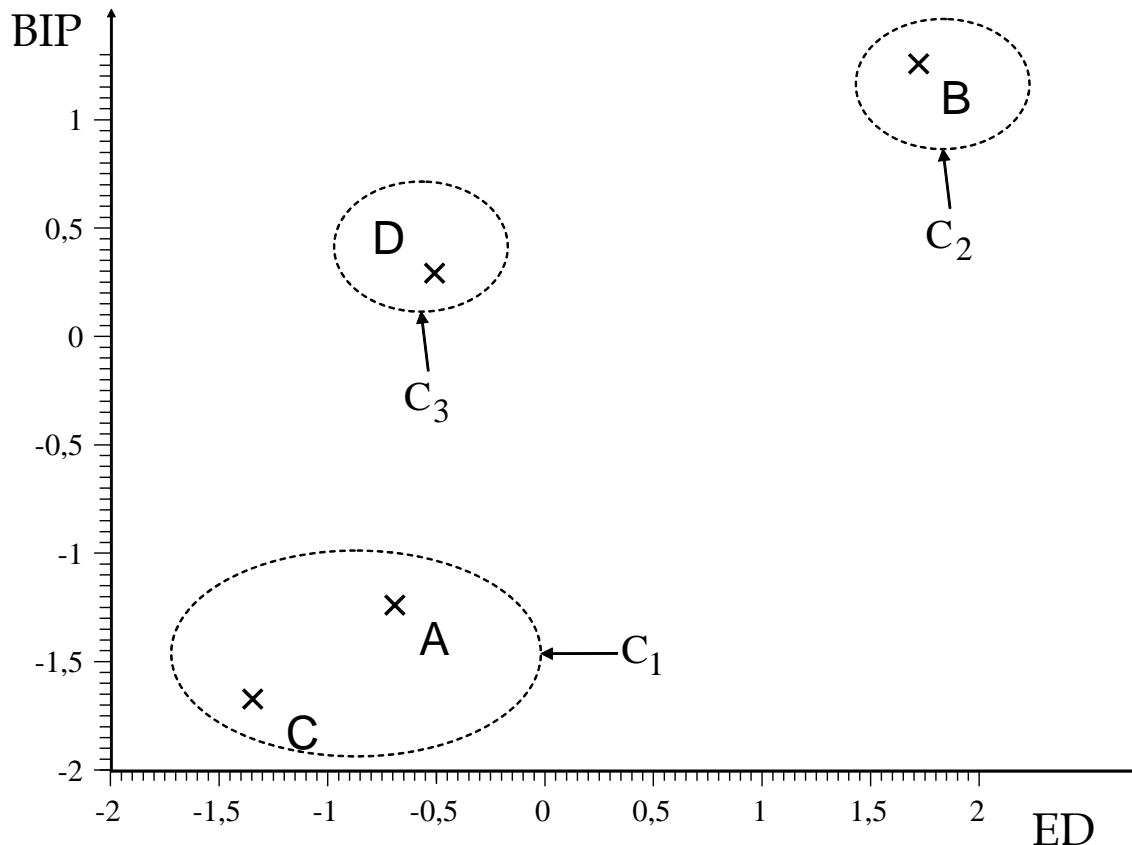
$$\Delta V(C_2 \cup C_3) = 8,883 \text{ ,}$$

$$\Delta V(C_2 \cup C_4) = 2,931 \text{ ,}$$

$$\Delta V(C_3 \cup C_4) = 2,247 \text{ ,}$$

so dass auf der ersten Stufe eine Vereinigung der Cluster C_1 und C_3 erfolgt:

Partition (1. Stufe): $C_1 = \{A, C\}$, $C_2 = \{B\}$, $C_3 = \{D\}$



Die Clustermittelwerte dieser Partition lauten

$$\text{Cluster } C_1 = \{A, C\}: \bar{z}_{11} = \frac{1}{2}[(-0,657) + (-1,343)] = -1,000,$$

$$\bar{z}_{12} = \frac{1}{2}[(-1,245) + (-1,653)] = -1,449$$

$$\text{Cluster } C_2 = \{B\}: \bar{z}_{21} = 1,709, \bar{z}_{22} = 1,254$$

$$\text{Cluster } C_3 = \{D\}: \bar{z}_{31} = -0,516, \bar{z}_{32} = 0,299$$

2. Stufe:

Die zweite Stufe beginnt erneut mit der Berechnung der aus der potenziellen Fusion hervorgehenden Erhöhung des Varianzkriteriums. Bei einer Fusion der Cluster C_1 und C_2 würde z.B. eine ΔV -Erhöhung von

$$\Delta V(C_1 \cup C_2) = \frac{2 \cdot 1}{2 + 1} \left\{ [(-1,000) - 1,709]^2 + [(-1,449) - 1,254]^2 \right\} = 9,763$$

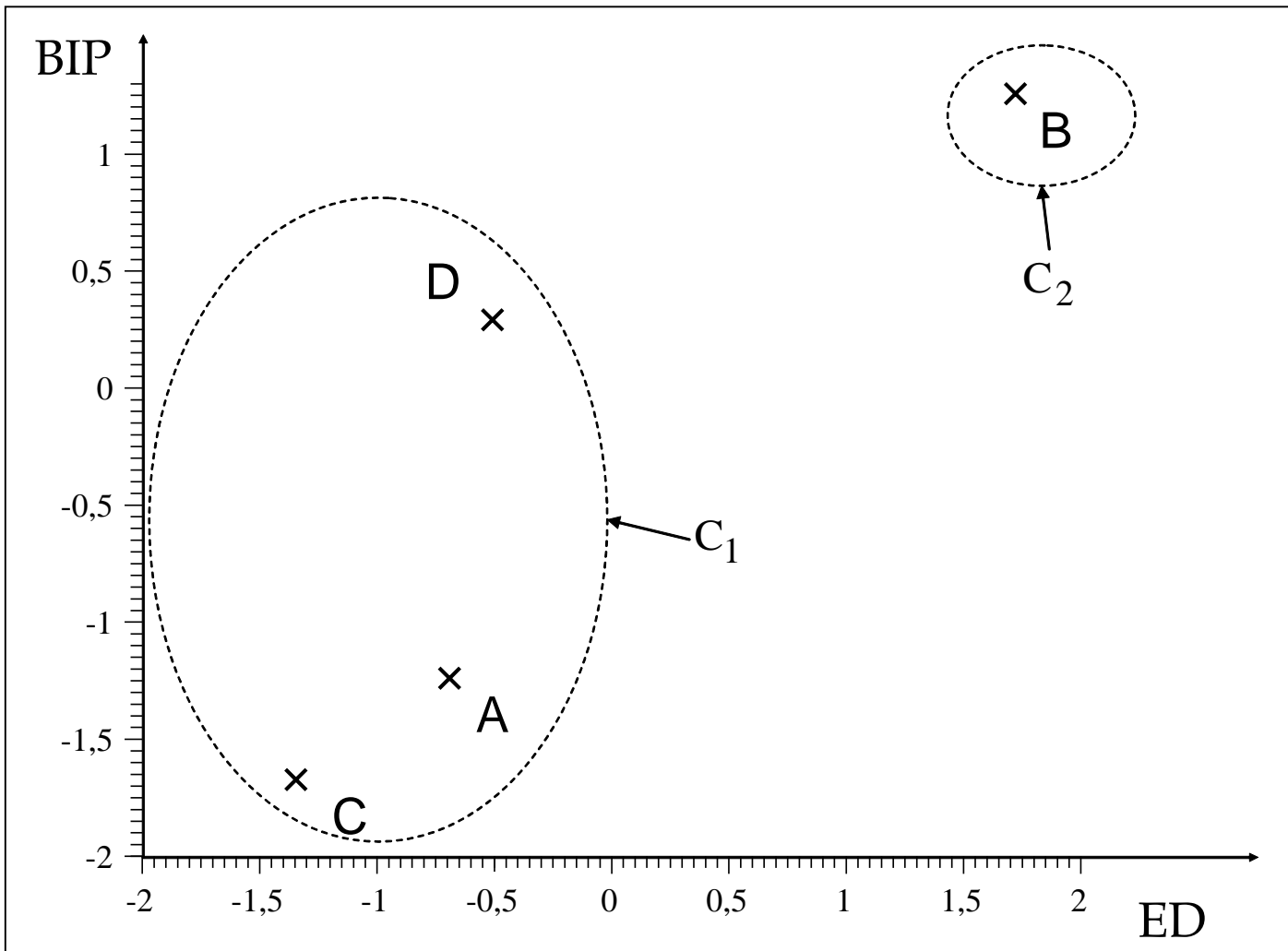
erfolgen. Für die beiden anderen Clusterpaare erhält man

$$\Delta V(C_1 \cup C_3) = 2,193,$$

$$\Delta V(C_2 \cup C_3) = 2,931,$$

was eine Fusion von C_1 und C_3 indiziert.

Partition (2. Stufe): $C_1 = \{A, C, D\}, C_2 = \{B\}$



Hierfür ergeben sich die Clustermittelwerte

$$\text{Cluster } C_1 = \{A, C, D\}: \quad \bar{z}_{11} = \frac{1}{3}[(-0,657) + (-1,343) + (-0,516)] = -0,839,$$

$$\bar{z}_{12} = \frac{1}{3}[(-1,245) + (-1,653) + 0,299] = -0,866,$$

Cluster $C_2 = \{B\}$: $\bar{z}_{21} = 1,709$, $\bar{z}_{22} = 1,254$

3. Stufe:

Aus einer Fusion der beiden verbleibenden Cluster auf der dritten Stufe resultiert schließlich ein Streuungszuwachs in Höhe von

$$\Delta V(C_1 \cup C_2) = \frac{3 \cdot 1}{3 + 1} \left\{ [-(0,839) - 1,709]^2 + [(-0,866) - 1,254]^2 \right\} = 8,240$$

bei Gesamtmittelwerten von

Cluster $C_1 = \{A, B, C, D\}$:

$$\bar{z}_{11} = \frac{1}{4} [(-0,657) + 1,709 + (-1,343) + (-0,516)] = -0,202,$$

$$\bar{z}_{12} = \frac{1}{4} [(-1,245) + 1,254 + (-1,653) + 0,299] = -1,345$$

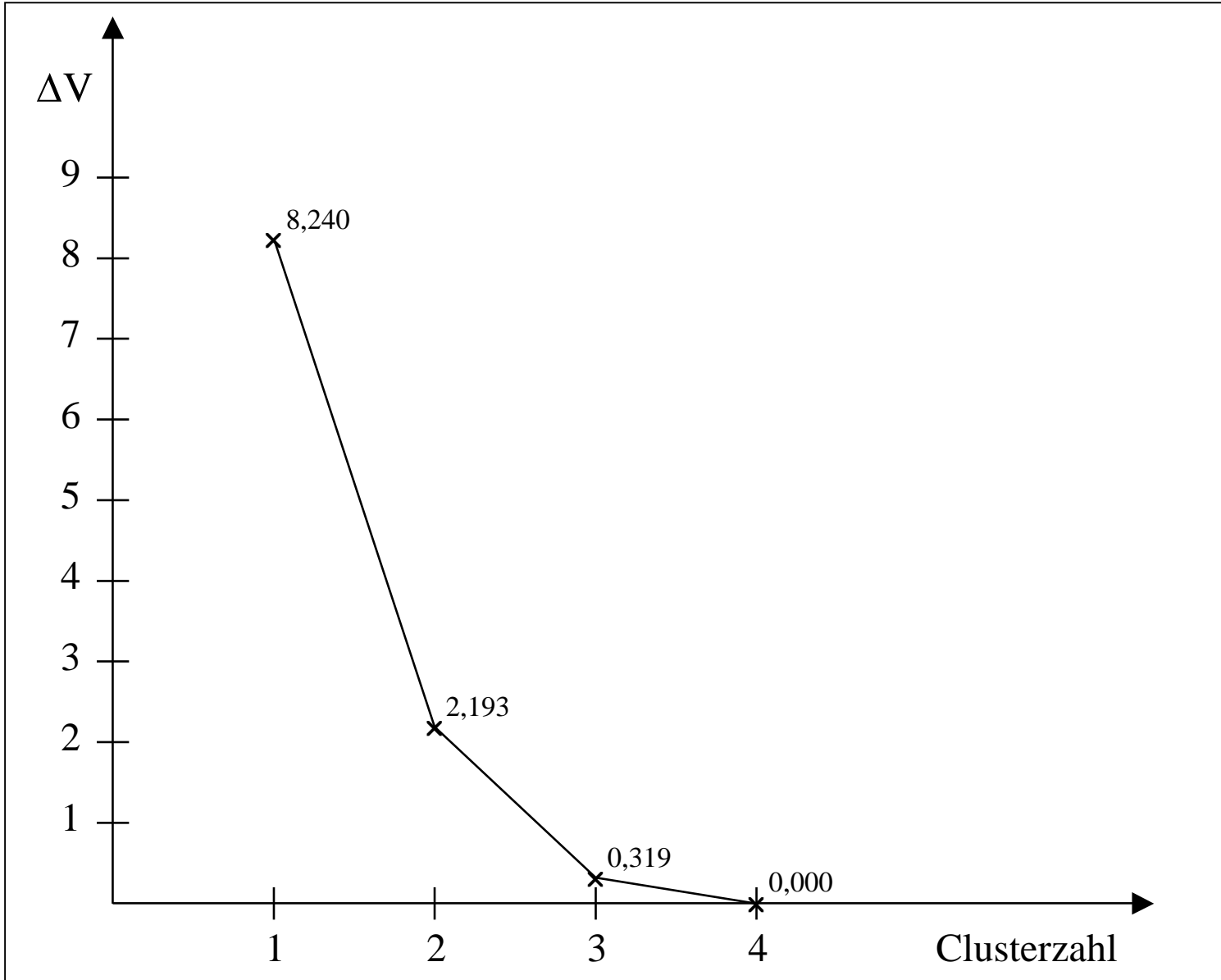


Struktogramm

Der Klassifikationsprozess könnte hier ebenfalls anhand eines Dendrogramms transparent gemacht werden. Zusätzlich lässt sich die Anzahl der Cluster mit Hilfe eines **Struktogramms** bestimmen, in dem der Streuungszuwachs ΔV gegen die Clusterzahl abgetragen wird.

Das Struktogramm ist vergleichbar mit dem Scree-Test in der Faktorenanalyse. Ein starker "Knick" spiegelt eine beträchtliche Abnahme der Streuung zwischen den Klassen wieder. Umgekehrt würden die Cluster erheblich heterogener werden, wenn man von rechts nach links im Struktogramm zu einer niedrigeren Clusterzahl überginge. Zur Bestimmung der Clusterzahl bietet sich daher die Lokalisation eines steilen "Knicks" vor einem flacheren Verlauf der Kurve in dem zugehörigen Struktogramm an. In unserem Beispiel (siehe Abbildung) ist er auffällig beim Übergang der Zwei-Klassen-Partition auf eine Ein-Klassen-Partition vorzufinden, so dass aufgrund dieses Kriteriums zwei Klassen zu bilden wären.

Abbildung 4.8: Struktogramm der Regionen A, B, C und D



Ward-Verfahren mit SPSS

Illustration des Ward-Verfahrens haben wir die Anzahl der Variablen auf 2 reduziert (ED, BIP). Im Methoden-Fenster wählen wir als „Cluster-Methode“ die „Ward-Methode“ und als „Maß“ den „Quadrierten Euklidischen Abstand“.

Die von SPSS ausgegebene „Näherungsmatrix“ enthält jetzt die quadrierten euklidischen Distanzen:

Näherungsmatrix

Fall	Quadriertes euklidisches Distanzmaß			
	1:A	2:B	3:C	4:D
1:A	.000	11.840	.637	2.405
2:B	11.840	.000	17.763	5.862
3:C	.637	17.763	.000	4.495
4:D	2.405	5.862	4.495	.000

Dies ist eine Unähnlichkeitsmatrix

Bei der Interpretation der „Zuordnungsübersicht“ ist zu beachten, dass die Spalte „Koeffizienten“ beim Ward-Verfahren keine Distanzen, sondern die V-Werte, d.h. die Streuung der Gruppierung der einzelnen Stufen (=Intra-Klassen-Streuung), ausweist.

Zuordnungsübersicht

Schritt	Zusammengeführte Cluster		Koeffizienten	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	3	.319	0	0	2
2	1	4	2.512	1	0	3
3	1	2	10.751	2	0	0

Wir haben bei unserer manuell durchgeführten Berechnung dagegen jeweils die ΔV -Werte, d.h. die Veränderungen der Streuung der Gruppierung der einzelnen Stufen, ausgewiesen. Man kann jedoch aufzeigen, dass beide Vorgehensweisen aufeinander abgestimmt sind:

1. Stufe:

Vor der ersten Stufe sind alle Cluster einelementig, damit ist keine Streuung innerhalb der Cluster vorhanden. Durch die Fusion von C_1 und C_3 erhöht sich V damit von 0 um $\Delta V = 0,319$ auf:

$$V = 0 + \underbrace{0,319}_{=\Delta V} = 0,319 .$$

2. Stufe:

Auf der zweiten Stufe tritt ein Zuwachs von V um $\Delta V = 2,193$ ein:

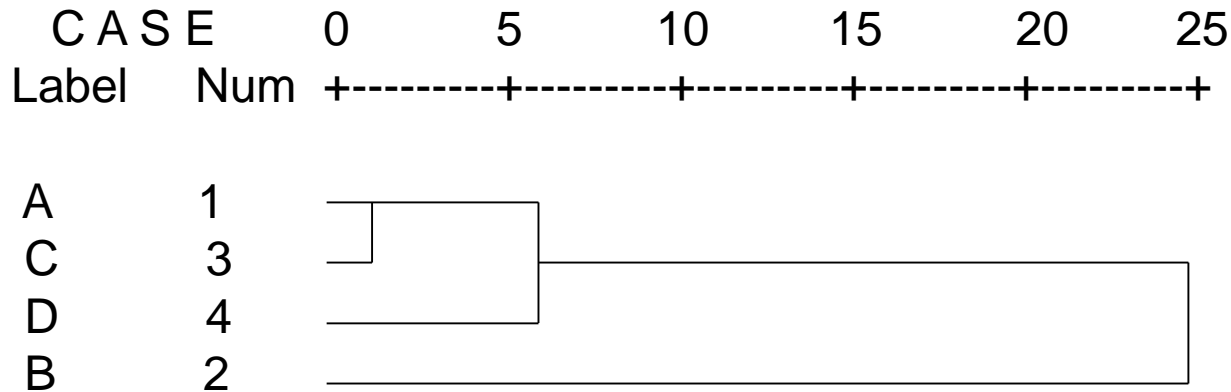
$$V = 0,319 + \underbrace{2,193}_{=\Delta V} = 2,512$$

3. Stufe: $V = 2,512 + \underbrace{8,240}_{=\Delta V} = 10,752$

***** HIERARCHICAL CLUSTER ANALYSIS *****

Dendrogram using Ward Method

Rescaled Distance Cluster Combine



Eigenschaften des Ward-Verfahrens:

- Konservatives Klassifikationsverfahren (nicht kontrahierend und nicht dilatierend),
- Tendenz des Verfahrens kompakte kugelförmige Cluster mit etwa gleichen Besetzungszahlen zu bilden,
- Hohe Anforderungen an das Skalenniveau (metrisch skalierten Merkmale),
- Unzureichende Eignung zu einer "Entdeckung" von ellipsoiden Clustern,
- Monotonieeigenschaft.