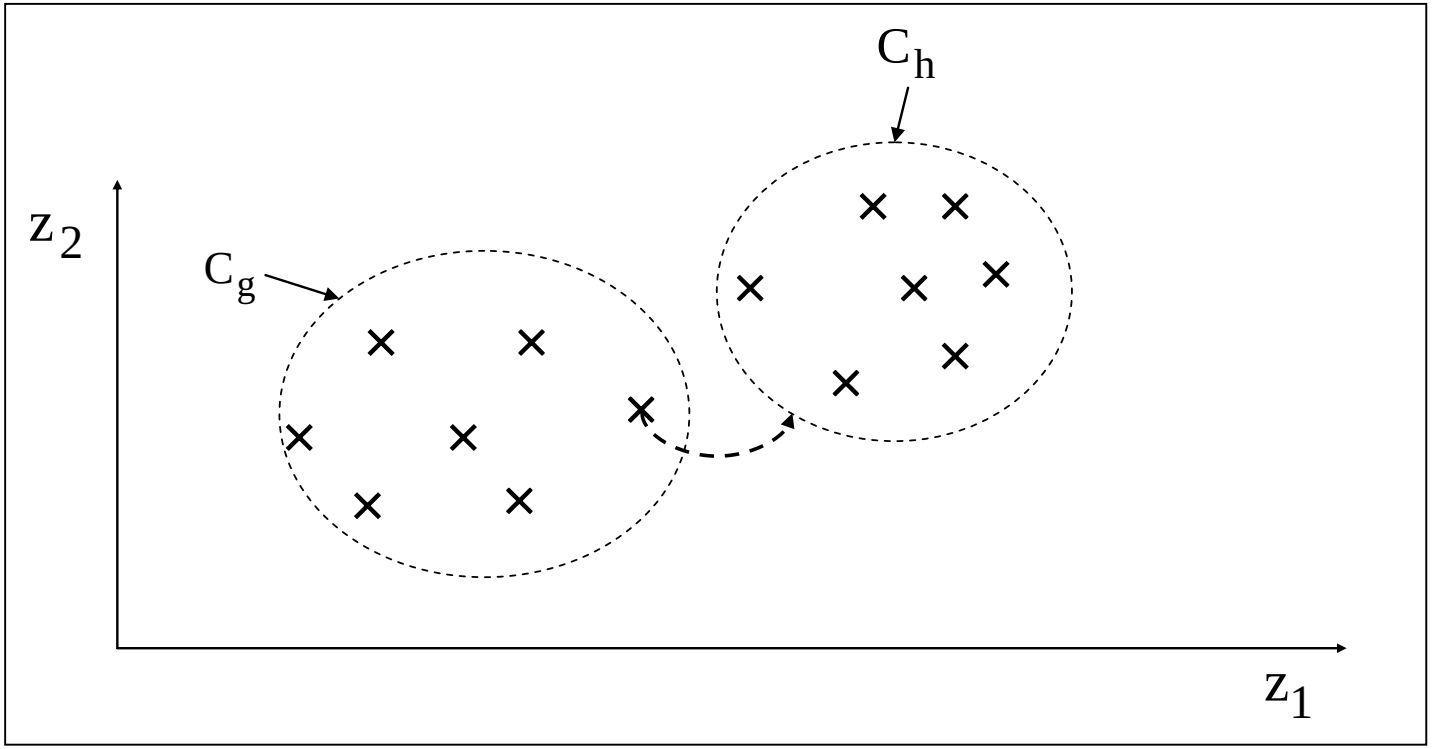
4.4 Partitionierende Klassifikationsverfahren

Partitionierenden Verfahren:

- gegeben: eine **Zerlegung der Objektmenge in G Cluster**, die jedoch nicht als "optimal" angesehen wird;
- Verbesserung der Ausgangspartition – die etwa durch ein hierarchisches Klassifikationsverfahren bestimmt worden ist - durch Umgruppierung von Objekten;
- Anzahl der Cluster bleibt in jeder Stufe konstant (G Cluster).

Abbildung 4.9: Umgruppierung eines Objektes bei Partitionsverfahren



Partitionsverfahren

Optimierende Austauschverfahren
(Hill-climbing methods)

Verbesserung einer Partition aufgrund eines Optimierungskriteriums

- Varianzkriterium
- Spurkriterium
- Determinantenkriterium

Iterative Minimal-Distanz-Verfahren

Verbesserung einer Partition durch Zuordnung eines Objekts zu einem Cluster mit einer geringeren Distanz zum Clusterschwerpunkt

- K-Means-Verfahren

Ablaufschema des Minimal-Distanz-Verfahrens

	Start: Vorgabe einer Ausgangspartition und eines Distanzmaßes
	Bestimmung der Clusterschwerpunkte für die vorliegende Partition.
	Berechnung der Distanz des ersten bzw. des nächsten Objekts zu allen Clusterschwerpunkten
nein ←	Ist eine Distanz zu einem anderen als dem bisherigen Cluster-schwerpunkt geringer?
	ja ↓
	Umgruppierung des Objekts: Zuordnung des Objekts zum Cluster mit der geringsten Distanz
nein ←	Ist eine Umgruppierung für alle Objekte bei gegebenen Clusterschwer-punken überprüft?
	ja ↓
ja ←	Ist mindestens ein Objekt umgruppiert worden?
	nein ↓
	Ende

Mögliche Variation des Ablaufschemas:

Jeweils neue Berechnung der Schwerpunkte der beiden bei einer Verschiebung eines Objektes tangierten Cluster vor Ermittlung der minimalen Distanz des nächsten Objektes

Anmerkung: Nach unserem Ablaufschema erfolgt eine Neuberechnung der Clusterschwerpunkte erst nach einem vollständigen Durchlauf (pass)

K-Means-Verfahren

Eine Konkretisierung erfährt das iterative Minimal-Distanz-Verfahren in Gestalt der K-Means-Methode. Hierbei werden die **Clusterschwerpunkte** durch die **Zentroide** definiert, deren Komponenten die arithmetischen Mittelwerte (means) der Merkmale innerhalb der Gruppen sind:

$$(4.19)\quad \bar{\mathbf{z}}_g = \begin{pmatrix} \bar{z}_{g1} & \bar{z}_{g2} & \dots & \bar{z}_{gm} \end{pmatrix}', \; g = 1,2,\dots,G.$$

Zur Ermittlung der Minimal-Distanz-Partition macht die K-Means-Methode von dem **quadrierten euklidischen Distanzmaß** Gebrauch:

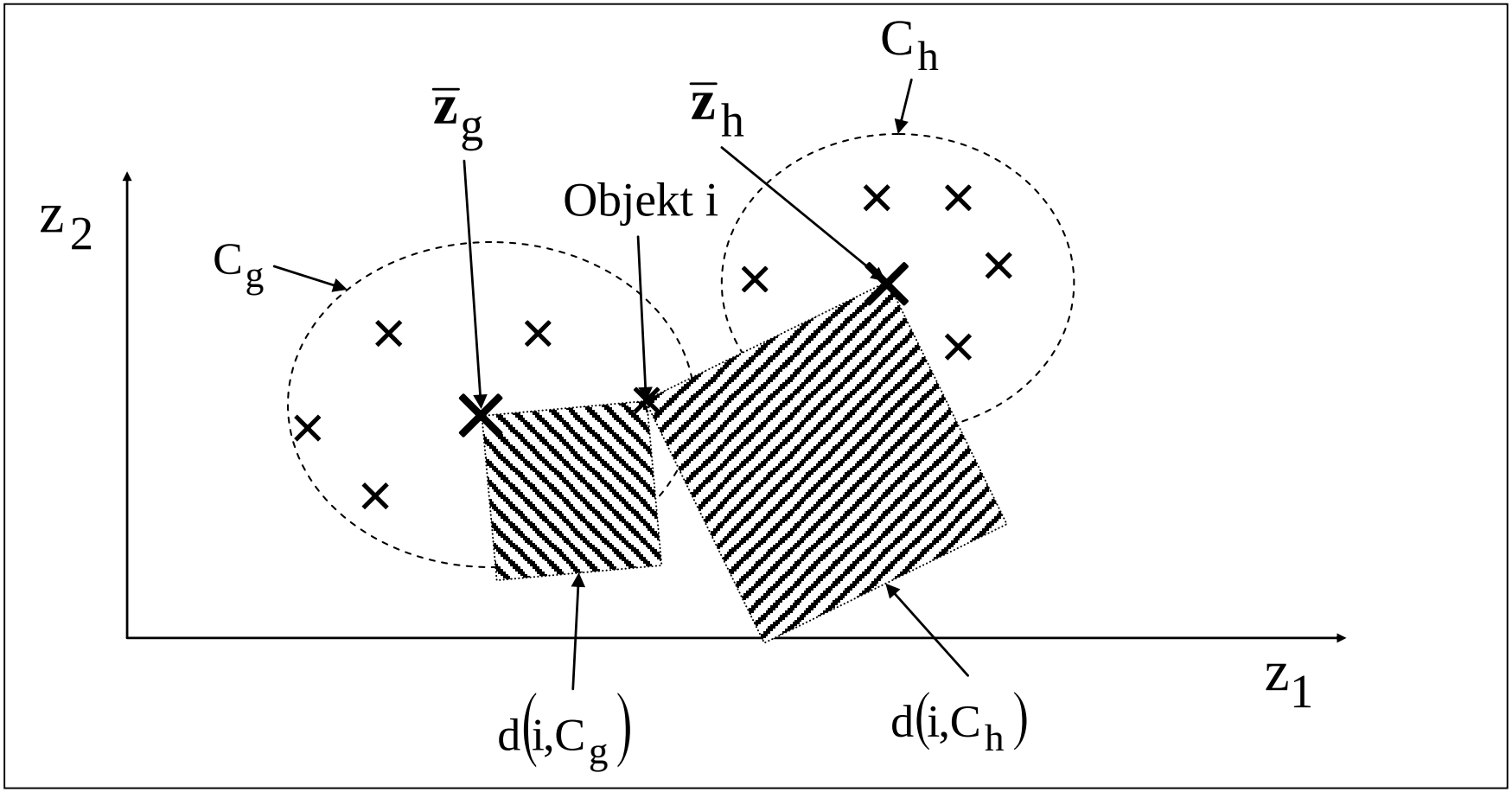
$$(4.20)\quad d(i, C_g) = \sum_{k=1}^{m} (z_{ik} - \bar{z}_{gk})^2.$$

Angenommen, das Objekt i befindet sich in dem Cluster C_g . Dann erfolgt ein Wechsel, wenn es ein Cluster $C_h, h \neq g$, mit der Eigenschaft

$$(4.21)\quad d(i, C_h) = \sum_{k=1}^{m} (z_{ik} - \bar{z}_{hk})^2 < d(i, C_g) = \sum_{k=1}^{m} (z_{ik} - \bar{z}_{gk})^2, h \neq g,$$

gibt.

Abbildung 4.10: Quadrierter Euklidischer Abstand zwischen dem Objekt i und den Schwerpunkten der Cluster C_g und C_h



Speziell wird das Objekt i dem Cluster C_ℓ mit der Minimal-Distanz-Eigenschaft

$$(4.22)\quad d(i,C_\ell) = \sum_{k=1}^{m}(z_{ik} - \bar{z}_{\ell k})^2 = \min_h \{d(i,C_h)\} = \min_h \left\{ \sum_{k=1}^{m}(z_{ik} - \bar{z}_{hk})^2 \right\}$$

zugeordnet.

Beispiel 4.15: Die Arbeitsweise der K-Means-Methode sei anhand des Regionenbeispiels aufgezeigt. Hierzu unterstellen wir eine Drei-Gruppen-Lösung, für die wir die Ausgangspartition

Cluster C_1 Regionen A, C, H

Cluster C_2 Regionen B, E, F

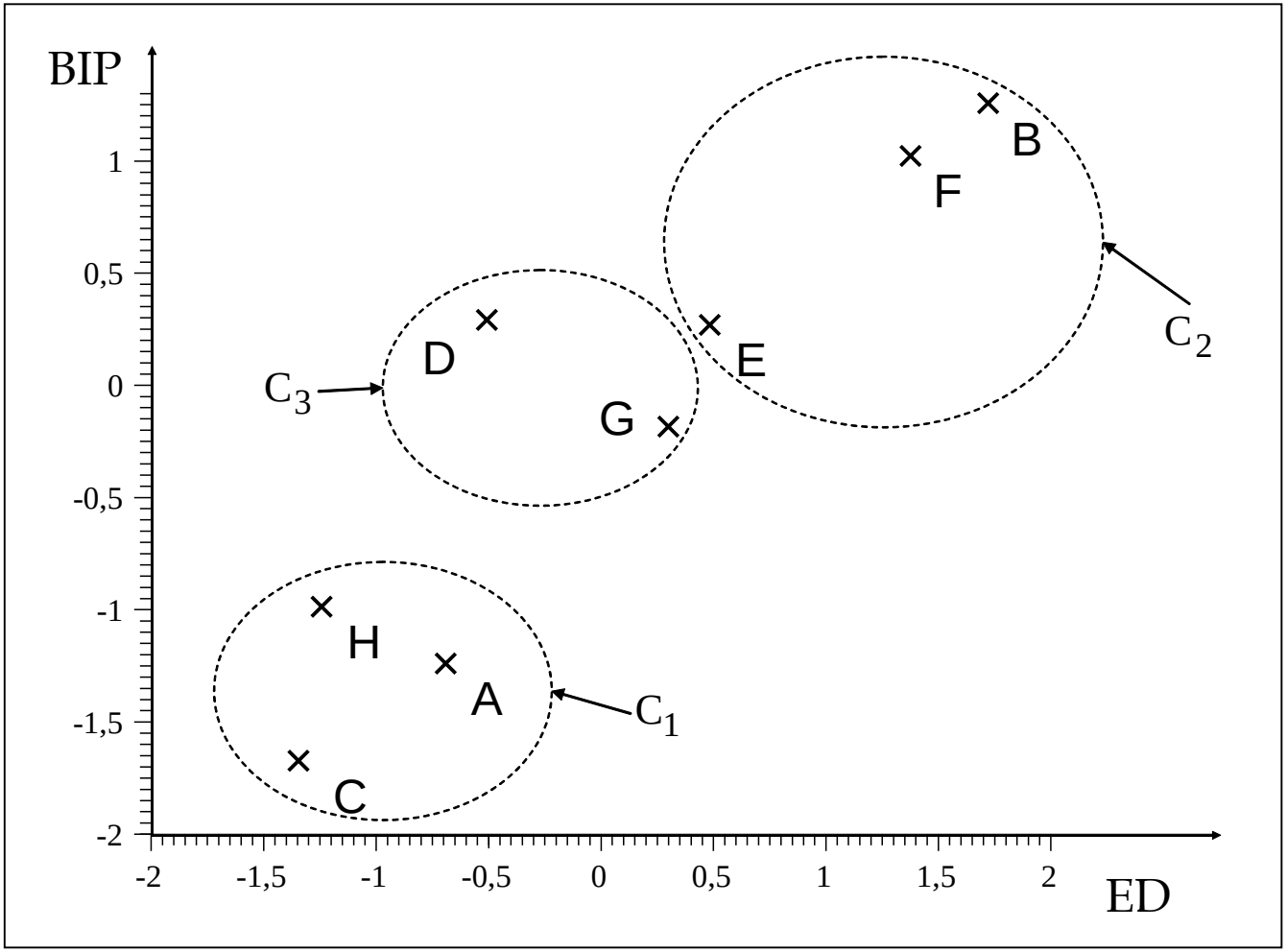
Cluster C_3 Regionen D, G

vorgeben. Der Einfachheit halber berücksichtigen wir bei der Ermittlung der Minimal-Distanz-Partition allein die Merkmale X_1 (ED) und X_2 (BIP). Bei Verwendung der standardisierten Merkmalswerte lassen sich die Cluster der Ausgangspartition wie folgt charakterisieren:

Cluster C_1		
Region	X_1	X_2
A	-0,657	-1,245
C	-1,343	-1,653
H	-1,237	-0,988
	$\bar{z}_{11} =$ -1,079	$\bar{z}_{12} =$ -1,2953

Cluster C_2		
Region	X_1	X_2
B	1,709	1,254
E	0,491	0,280
F	1,381	1,022
	$\bar{z}_{21} =$ 1,1937	$\bar{z}_{22} =$ 0,852

Cluster C_3		
Region	X_1	X_2
D	-0,516	0,299
G	0,031	-0,178
	$\bar{z}_{31} =$ -0,2425	$\bar{z}_{32} =$ 0,0605



Zwei Clustermittelwerte seien beispielhaft berechnet:

$$\bar{z}_{11} = \frac{1}{n_1} \cdot \sum_{i \in C1} z_{i1} = \frac{1}{3}\left[(-0{,}657) + (-1{,}343) + (-1{,}237)\right] = -1{,}0790$$

$$\bar{z}_{32} = \frac{1}{n_3} \cdot \sum_{i \in C3} z_{i2} = \frac{1}{2} \cdot \left[0{,}299 + (-0{,}178)\right] = 0{,}0605.$$

1. Durchlauf (1. Iteration)

Um zu erkennen, ob eine Region vom jetzigen Cluster in eines der beiden anderen Cluster zu verschieben ist, sind ihre quadratischen euklidischen Distanzen zu den Clusterschwerpunkten (Gruppenzentroiden) zu ermitteln. Für die Region A ergibt sich aus (4.19) eine **quadrierte euklidische Distanz** zu den Schwerpunkten der Cluster C_1, C_2 und C_3 von

$$d(A,C_1) = (z_{A1} - \bar{z}_{11})^2 + (z_{A2} - \bar{z}_{12})^2$$
$$= [(-0{,}657) - (-1{,}079)]^2 + [(-1{,}245) - (-1{,}2953)]^2 = 0{,}181.$$

$$d(A,C_2) = (z_{A1} - \bar{z}_{21})^2 + (z_{A2} - \bar{z}_{22})^2$$
$$= [(-0{,}657) - 1{,}1937]^2 + [(-1{,}245) - 0{,}852]^2 = 7{,}823$$

$$d(A,C_3) = (z_{A1} - \bar{z}_{31})^2 + (z_{A2} - \bar{z}_{32})^2$$
$$= [(-0{,}657) - (-0{,}2425)]^2 + [(-1{,}245) - 0{,}0605]^2 = 1{,}876$$

Die Distanz der Region A zu den Clustern C_2 und C_3 ist mithin größer als zum Schwerpunkt (Zentroid) des Clusters C_1, dem sie aktuell zugeordnet ist. Da die Bedingung (4.20) nicht erfüllt ist, unterbleibt eine Umgruppierung der Region A.

Berechnet man auf analoge Weise für die übrigen Regionen die Distanzen zu den Clusterschwerpunkten, dann ergibt sich folgendes Bild:

	Distanz zum Schwerpunkt des Clusters			Gruppe	
Region	C_1	C_2	C_3	alt	neu
A	0,181	7,822	1,876	C_1	C_1
C	0,198	12,710	4,147	C_1	C_1
H	0,119	9,294	2,088	C_1	C_1
B	14,272	0,040	4,169	C_2	C_2
E	4,947	0,821	0,586	C_2	C_3
F	11,422	0,064	3,560	C_2	C_2
D	2,859	3,229	0,132	C_3	C_3
G	2,481	2,413	0,132	C_3	C_3

Für jedes Objekt ist die **minimale Distanz** zu einem Cluster grau unterlegt. Man erkennt unmittelbar, dass allein die Region E "falsch" gruppiert ist. In Bezug auf das Cluster C_3 wird die Wechseleigenschaft (4.21) wirksam, womit gleichzeitig die Minimal-Distanz-Eigenschaft (4.22) erfüllt ist.

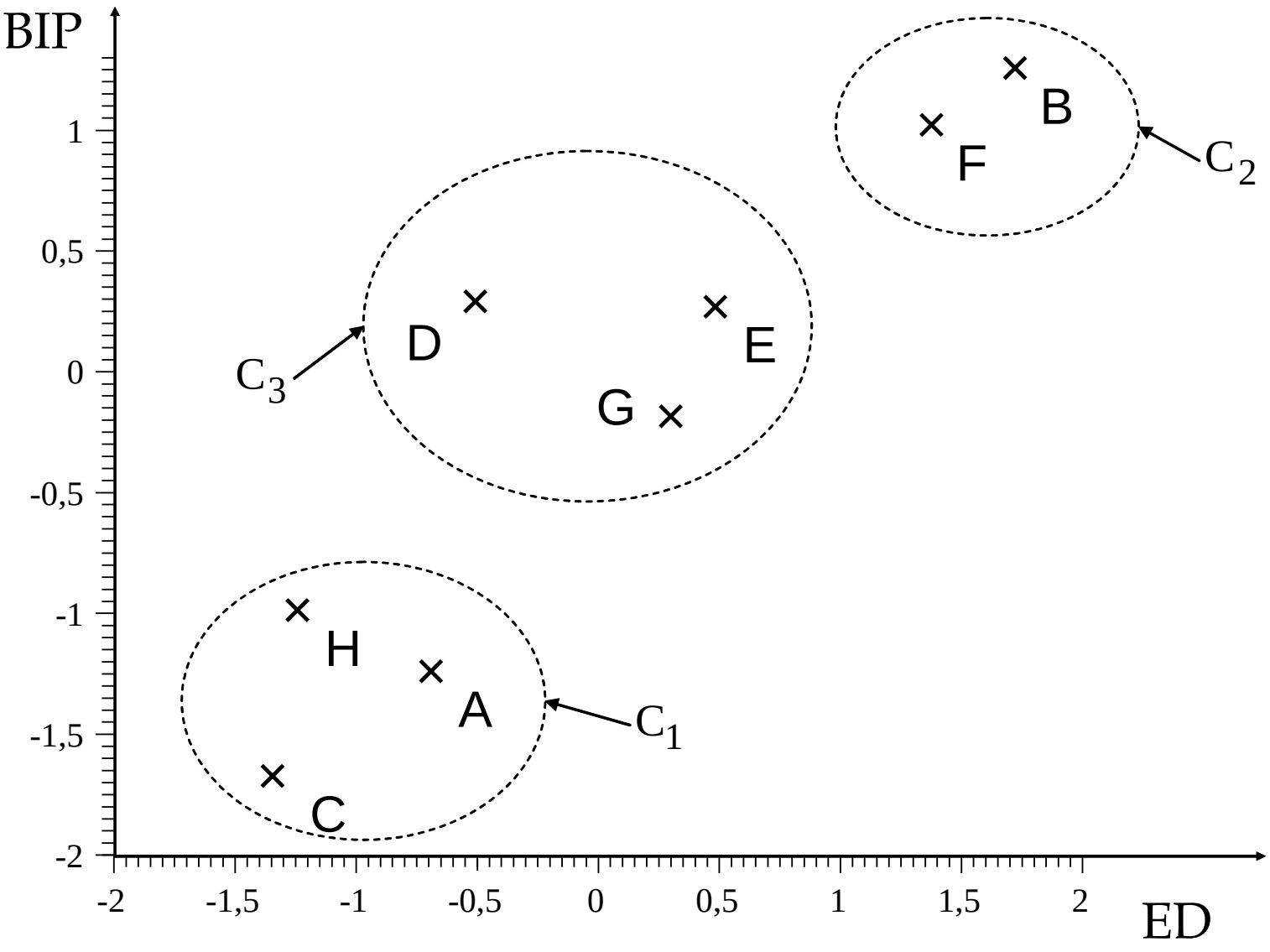
Der Wechsel der Region E von C_2 nach C_3 führt zur Partition

Cluster C_1: Regionen A, C, H
Cluster C_2: Regionen B, F
Cluster C_3: Regionen D, E, G

die durch die Struktur

Cluster C_1			Cluster C_2			Cluster C_3		
Region			Region			Region		
A	-0,657	-1,245	B	1,709	1,254	D	-0,516	0,299
C	-1,343	-1,653	F	1,381	1,022	E	0,491	0,280
H	-1,237	-0,988				G	0,031	-0,178
	$\bar{z}_{11} =$ -1,079	$\bar{z}_{12} =$ -1,2953		$\bar{z}_{21} =$ 1,545	$\bar{z}_{22} =$ 1,138		$\bar{z}_{31} =$ 0,002	$\bar{z}_{32} =$ 0,134

gekennzeichnet ist.


BIP
1
0,5
0
-0,5
-1
-1,5
-2
-2
-1,5
-1
-0,5
0
0,5
1
1,5
2
ED
B
F
C2
D
E
C3
G
H
A
C1
C

2. Durchlauf (2. Iteration)

Im zweiten Durchlauf werden die quadratischen euklidischen Distanzen zwischen den Regionen und den Zentroiden der drei Cluster neu berechnet:

	Distanz zum Schwerpunkt des Clusters			Gruppe	
Region	C_1	C_2	C_3	alt	neu
A	0,181	10,527	2,330	C_1	C_1
C	0,198	16,130	5,001	C_1	C_1
H	0,119	12,259	2,793	C_1	C_1
B	14,272	0,040	4,169	C_2	C_2
F	11,422	0,040	2,691	C_2	C_2
D	2,859	4,952	0,296	C_3	C_3
E	4,947	1,847	0,261	C_3	C_3
G	2,481	4,024	0,098	C_3	C_3

Wie sich zeigt, lassen sich die Abstände der Regionen zu den drei Clustern durch eine Umgruppierung nicht mehr verringern. Die sich am Ende des ersten Durchlaufs herausbildende Partition ist mithin bereits eine **Minimal-Distanz-Partition**, so dass das Iterationsverfahren nach zwei Durchläufen beendet wird. ♦

K-Means-Verfahren mit SPSS

Wir illustrieren die Anwendung des K-Means-Verfahrens mit SPSS anhand des manuell durchgerechneten Beipiels der 8 Regionen (A, B, C, D, E, F, G, H) und 2 Merkmale (ED, BIP). Das K-Means-Verfahren ist in SPSS unter

Analysieren
    Klassifizieren
        Clusterzentrenanalyse …

aufrufbar. Die standardisierten Daten müssen vorab berechnet worden sein. Wir berechnen daher zunächst die standardisierten Werte für alle 12 Regionen und 6 Merkmale durch Wahl der Menüpunkte

Analysieren
    Deskriptive Statistiken
        Deskriptive Statistiken ….

Anschließend speichern dann aber nur die Variablen

- Region: A, B, C, D, E, F, G, H,
- ED (standardisiert),
- BIP (standardisert),
- Cluster: 1 für A, C, H; 2 für B, E, F, 3 für D, G

für die ersten 8 Regionen in eine SPSS-Datendatei Regionen(Z8).sav. Da wir bei der Umgruppierung von Objekten nicht mit einer Zufallspartition starten wollen, son-dern mit der aus unserem durchgerechneten Beispiel vorgegebene Anfangspartition, muss diese SPSS in einer Datendatei zur Verfügung gestellt werden. Diese SPSS-Datendatei - in unserem Fall Clustermittelwerte.sav – muss einen ganz bestimmten Aufbau haben, der aus der folgenden Tabelle hervorgeht:

cluster_	ED	BIP
1	-1.0790	-1.2953
2	1.1937	.8520
3	-.2425	.0605

Die erste Variable, die die Gruppennummern enthält, muss cluster_ heißen. Die Na-men der anderen Variablen müssen aus der zu analysierenden Datendatei übernom-men werden, d.h. in unserem Beispiel ED und BIP. Sie enthalten hier die Mittelwerte der Merkmale in den einzelnen Clustern.

Bei geladener SPSS-Datendatei Regionen(Z8).sav sind im Fenster „Clusterzentren-analyse“ verschiedene Einstellungen vorzunehmen:

- Variablen ED und BIP → Feld „Variablen“,
- Variable Region → Feld “Fallbeschriftung”,
- Anzahl der Cluster: 3,
- im Feld “Clusterzentren”: Anklicken von “Anfangswerte einlesen aus” → Schaltfläche „Datei“ anklicken: Datei Clustermittelwerte.sav übernehmen.

Wir klicken die Schaltfäche „Optionen…“ an und wählen im Feld „Statistik“ das Item „Clusterinformationen für jeden Fall“. Schließlich klicken wir noch die Schaltfläche „Speichern“, wonach sich das Fester „Clusterzentrenanalyse: Neue Variablen“ öffnet, in dem wir beide Einträge („Cluster-Zugehörigkeit, „Distanz vom Clusterzentrum“) wählen.

Als Ergebnis erhalten wir im SPSS-Viewer verschiedene Ausgabetabellen. Als erstes wird eine Tabelle mit den anfänglichen Clustermittelwerten ausgegeben:

Anfängliche Clusterzentren

	Cluster		
	1	2	3
Einwohnerdichte	-1.079	1.194	-.243
BIP je Einwohner	-1.295	.852	.061

Aus Unterbefehl FILE eingeben

Im „Iterationsprotokoll“ wird die Änderung der Clusterzentren in den beiden Iterationen angegeben. Die Änderung wird dabei jeweils durch die euklidische Distanz zwischen dem vorherigen und aktuellen Schwerpunkt (Zentrum, Zentroid) des betreffenden Clusters gemessen.

Iterationsprotokoll[a]

Iteration	Änderung in Clusterzentren		
	1	2	3
1	3.335E-13	.453	.255
2	.000	.000	.000

a. Konvergenz wurde aufgrund geringer oder keiner Änderungen der Clusterzentren erreicht. Die maximale Änderung der absoluten Koordinaten für jedes Zentrum ist .000. Die aktuelle Iteration lautet 2. Der Mindestabstand zwischen den anfänglichen Zentren beträgt 1.593.

Beispiel 4.16: 1. Iteration: Änderung des Clusterzentrums des 3. Clusters = 0,255
Vorheriger Schwerpunkt des Clusters C_3: $(\bar{z}_{31} = -0{,}2425 \quad \bar{z}_{32} = 0{,}0605)$
Aktueller Schwerpunkt des Clusters C_3: $(\bar{z}_{31} = 0{,}002 \quad \bar{z}_{32} = 0{,}134)$

Euklidische Distanz zwischen dem aktuellen und vorherigen Schwerpunkt des Clusters C_3:

$$\sqrt{[0{,}002-(-0{,}2425)]^2+(0{,}134-0{,}0605)^2} = \sqrt{0{,}0598+0{,}0054} = 0{,}255$$

In der zweiten Iteration kann – wie bereits manuell aufgezeigt – keine weitere Verbesserung mehr erreicht werden, so dass die Partition am Ende der ersten Iteration eine Minimal-Distanz-Partition ist.

In der Tabelle „Cluster-Zugehörigkeit“ wird die Zuordnung der Regionen zu den drei Clustern wiedergegeben. Die letzte Spalte „Distanz“ in dieser Tabelle enthält die euklidischen Distanzen der Regionen zu den Clusterschwerpunkten in der endgültigen Partition (=Minimal-Distanz-Partition). Da wir in unserem manuellen Beispiel – wie üblich beim K-Means-Verfahren - die quadrierten euklidischen Distanzen verwendet haben wieder, lassen sich beiden Größen durch Wurzelziehen bzw. Quadrieren einfach ineinander überführen.

Beispielsweise hatten wir für die Region A eine quadrierte euklidische Distanz von 0,181 zum Clusterzentrum C_1 berechnet. Zieht man hieraus die Wurzel, dann erhält man unmittelbar die von SPSS ausgewiesene Distanz von 0,425: $\sqrt{0{,}181} = 0{,}425$. Entsprechend erhält man aus der von uns ausgewiesenen quadrierten euklidischen Distanz der Region C zum Clusterzentrum C_1 von 0,198 durch Wurzelziehen den euklidischen Distanzwert 0,445 von SPSS: $\sqrt{0{,}198} = 0{,}445$.

Cluster-Zugehörigkeit

Fallnummer	Region	Cluster	Distanz
1	A	1	.425
2	C	1	.445
3	H	1	.346
4	B	2	.201
5	E	3	.510
6	F	2	.201
7	D	3	.544
8	G	3	.313

Die Tabelle „Clusterzentren der endgültigen Lösung“ gibt die Clustermittelwerte der Merkmale in der Minimal-Distanz-Partition wieder:

Clusterzentren der endgültigen Lösung

	Cluster		
	1	2	3
Einwohnerdichte	-1.079	1.545	.002
BIP je Einwohner	-1.295	1.138	.134

Zusätzlich lassen sich die Distanzen zwischen den Clusterzentren der endgültigen Lösung berechnen:

Distanz zwischen Clusterzentren der endgültigen Lösung

Cluster	1	2	3
1		3.579	1.792
2	3.579		1.841
3	1.792	1.841	

<u>Übung:</u> Vollziehen Sie die von SPSS berechneten Distanzen zwischen den Clusterzentren der endgültigen Lösung nach! Um welches Distanzmaß handelt es sich hierbei?

Schließlich gibt SPSS noch die Verteilung der Objekte (Fälle) auf die einzelnen Cluster an:

Anzahl der Fälle in jedem Cluster

Cluster	1	3.000
	2	2.000
	3	3.000
Gültig		8.000
Fehlend		.000

Entsprechend unserer Einstellung ist unsere SPSS-Datendatei Regionen(Z8).sav um eine Variable der Clusterzugehörigkeit der Objekte und ihrer Distanz zum zugehörigen Clusterschwerpunkt erweitert worden.

♦