Gruppierungsanalysen

5. Gruppierungsanalysen

Bei den Gruppierungsanalysen geht es darum, Stichproben, also z.B. eine Reihe von Vegetationsaufnahmen, in Gruppen möglichst ähnlicher Individuen zu unterteilen. Ist eine solche Analyse erfolgreich, so kann in der Folge anstatt mit den Eigenschaften der Individuen mit denjenigen der viel weniger zahlreichen Gruppen weitergearbeitet werden. Dies erleichtert das Erkennen von Zusammenhängen und Gesetzmässigkeiten innerhalb eines Datensatzes. In der Vegetationskunde gibt es Theorien, welche von der Existenz organismusähnlicher Pflanzengesellschaften und mithin natürlicher Gruppen ausgehen. Analysen dienen dabei der Erkennung "echter" Pflanzengesellschaften. Andere Auffassungen gehen dahin, die Vegetationsdecke der Erde als "Kontinuum" zu betrachten (GLEASON 1926, 1939). Logisch gefolgert müssten Gruppierungsanalysen bevorzugte Methoden der Anhänger diskreter Pflanzengesellschaften sein. Wie nun aber im folgenden zu zeigen ist, können mit geeigneten Methoden auch gradientenhafte Datenstrukturen analysiert werden. Die Resultate sind in jedem Falle vom Datensatz einerseits, vom Gruppierungsalgorithmus andererseits abhängig.

5.1 Gruppenstruktur

Zunächst sollen verschiedene Möglichkeiten von Gruppenstrukturen erörtert werden. Der einfachste Fall ist in Abb. 5.1 dargestellt. Diese zweidimensionale Struktur ist kontinuierlich und einigermassen linear. Die gesamte Wolke von Punkten kann sinnvollerweise nur als eine einzige natürliche Gruppe aufgefasst werden. Während gewisse Gruppierungsmethoden tatsächlich zu diesem Ergebnis kommen, lassen andere eine weitergehende Unterteilung zu (gestrichelte Trennlinien in Abb. 5.1). Dabei kann von natürlichen Gruppen nicht die Rede sein. Eine willkürliche Unterteilung einer an sich kompakten Punktewolke kann aber sinnvoll und notwendig sein, wenn eine gradientenhafte Aehnlichkeitsstruktur vorliegt.

Ein wichtiger Spezialfall kontinuierlicher Strukturen ist die mehrdimensionale Normalverteilung. Um eine solche erken-

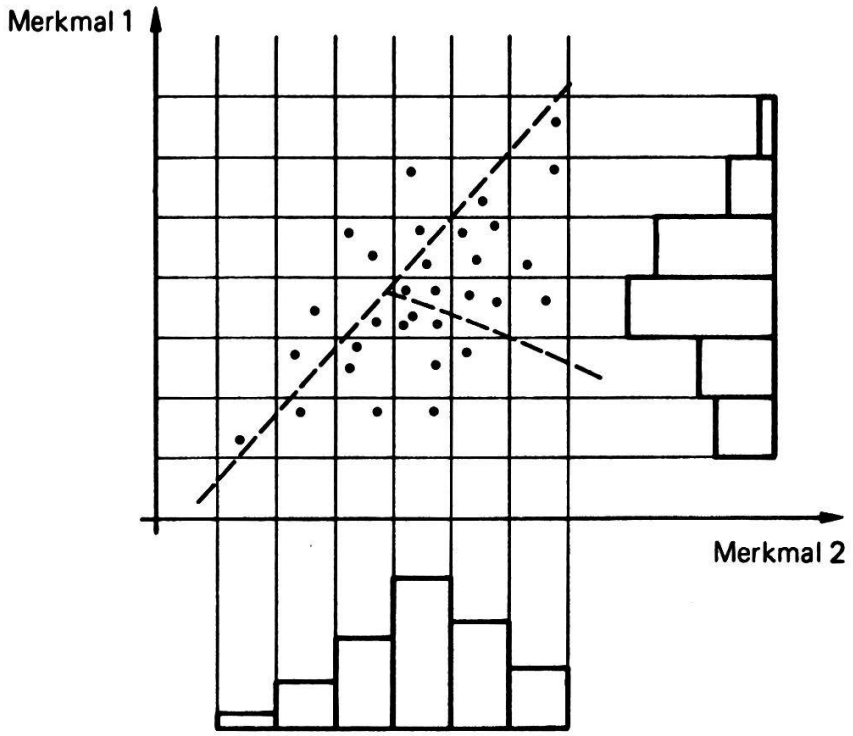

Abb. 5.1 Angenähert multivariat-normalverteilte Stichprobe. Die Balkendiagramme zeigen die Häufigkeitsverteilungen bezüglich der Merkmale 1 und 2. Gestrichelt ist eine Möglichkeit zur Unterteilung angedeutet.

nen zu können, werden alle Achsen in Klassen unterteilt und die Anzahl der Individuen pro Klasse in Balkendiagrammen dargestellt (Abb. 5.1). Die so entstehenden Häufigkeitsverteilungen müssen bezüglich jeder Dimension durch eine Glockenkurve angenähert werden können. In Vegetationstabellen tritt eine solche Struktur fast ausschliesslich bei Artvergleichen auf, beim Vergleich von Aufnahmen dagegen fast nie. Aus diesem Grunde sind zur Gruppierung von Arten und Aufnahmen gelegentlich verschiedene Methoden zu verwenden.

Recht häufig sind Strukturen, wie sie in Abb. 5.2, A, dargestellt sind (vgl. auch BARTEL 1974, S. 83). Sie sind kontinuierlich, oft ausgesprochen länglich und gekrümmt. Bei deren Analyse sind all jene Methoden ungeeignet, welche speziell mehrdimensional normalverteilte Gruppen auseinanderhalten können. Wirkungsvoller sind hier Methoden, die mehr oder weniger willkürliche, eher kompakte Gruppierungen herbeiführen.

Abb. 5.2, B, zeigt den Fall disjunkter "natürlicher" Gruppen. Methoden, die solche Strukturen aufzudecken vermögen, gibt es zahlreiche, doch können viele davon durch intermediäre Individuen (dünner Pfeil in Abb. 5.2, B) gestört werden. Aberrante Individuen (dicker Pfeil) lassen sich dagegen leicht aufspüren und werden in der Regel als eigene Gruppe betrachtet.

Damit ist das Spektrum in der Pflanzensoziologie auftretender Strukturen noch nicht erschöpft. Fast immer zeigt es sich, dass irgendwelche Kombinationen der in Abb. 5.2, A und B, gezeigten Fälle vorliegen. Abb. 5.2, C, ist ein Beispiel. In der Tat gibt es Methoden, die auch diese Konfiguration als aus drei Gruppen bestehend erkennt.

Die hier gezeigten Beispiele lassen die Vermutung aufkommen, dass die visuelle Gruppierung einer Stichprobe viel rascher und sicherer sein könnte, als eine numerische. SPäTH (1977) erwähnt, dass dies für ein- und zweidimensionale Strukturen zutrifft. Liegen jedoch drei und mehr Dimensionen vor, so werden numerische Verfahren rasch überlegen. Bei

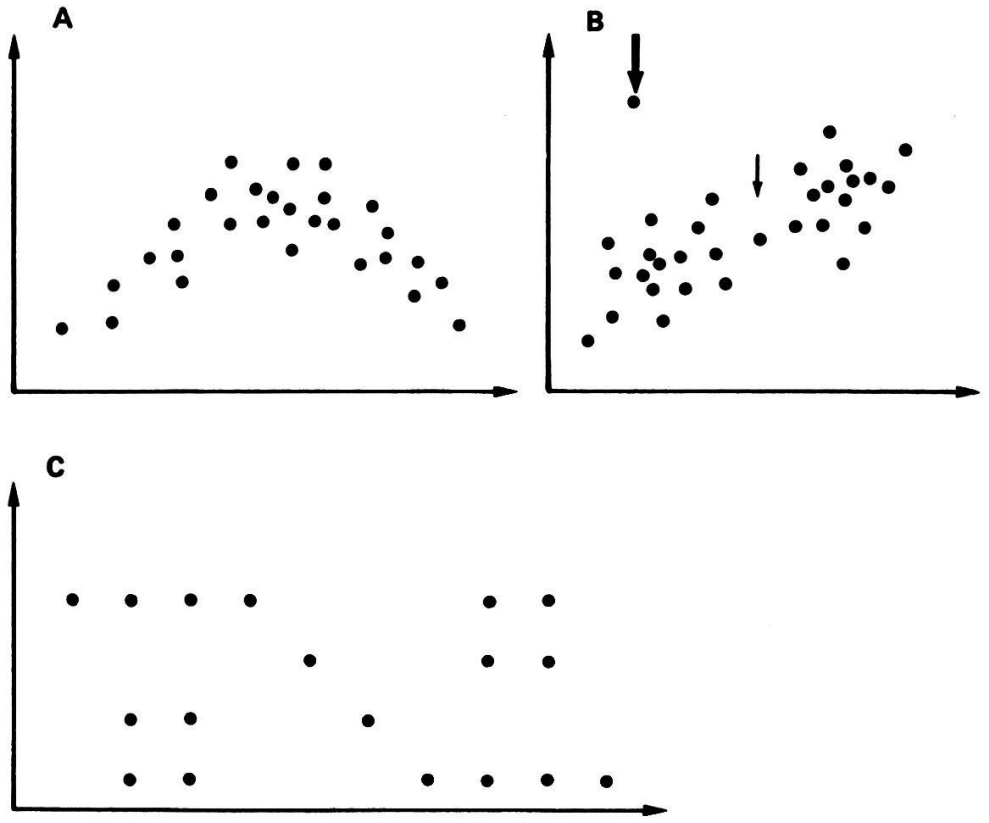

Abb. 5.2 Verschiedene Gruppenstrukturen. A: Langgezogene, gekrümmte Gruppe. B: Zwei getrennte Gruppen mit intermediärem (dünner Pfeil) und aberrantem Individuum (fetter Pfeil). C: Zwei quadratische und eine langgezogene Gruppe.

vegetationskundlichen Anwendungen mit oft hundert und mehr Dimensionen - nämlich Arten - bilden sie den einzigen Weg zu nachvollziehbaren Ergebnissen.

5.2 Heuristische Verfahren

Heuristische Verfahren erfüllen die Qualitätsanforderungen nicht, welche normalerweise an multivariate Verfahren gestellt werden. In der Regel basieren sie nicht auf einer vollständigen Aehnlichkeitsmatrix. Damit kann, streng genommen, die wirkliche Gruppenstruktur einer Stichprobe gar nicht vollständig erfasst werden. Vorteile sind dagegen ein geringer Rechen- und Speicheraufwand. Tausende von Vegetationsaufnahmen lassen sich rasch und leicht provisorisch gruppieren. Beispiel eines iterativen Verfahrens dieser Art ist TABORD (VAN DER MAAREL et al. 1978), welches Vegetationstabellen grösseren Ausmasses zu strukturieren erlaubt.

Einen typischen, ausgesprochen einfachen Vertreter heuristischer Verfahren finden wir bei ANDERBERG (1973) und SPäTH (1977). Er wird als LEADER-Algorithmus bezeichnet, lehnt sich stark an ein intuitives Vorgehen an und dient bei VAN DER MAAREL et al. (1978) als erstes grobes Ordnungsverfahren. Als Aehnlichkeitsmass wird oft die Euklidsche Distanz verwendet. Es gelte folgende Notation:

i i = 1,...,n ist die momentan zu verarbeitende Aufnahme bei einem Total von n;

rho ist eine vom Benützer zu definierende Distanz. Sie begrenzt die in einer Gruppe auftretenden Unterschiede zwischen den Aufnahmen;

NMAX ist die maximal erlaubte Anzahl Gruppen.

Die LEADER-Methode verläuft sodann wie folgt:

1. Die erste Aufnahme des Datensatzes (i=1) wird erste Aufnahme (Leitaufnahme) der ersten Gruppe.

2. i wird um 1 erhöht. Die nächstfolgende Aufnahme i wird verarbeitet (im ersten Durchgang ist i = 2). Dazu wird die Euklidsche Distanz zu den ersten Aufnahmen (Leitaufnahmen) der bereits bestehenden Gruppen berechnet.

3. Aufnahme i wird der ersten Gruppe zugeordnet, zu deren Leitaufnahme die Distanz kleiner ist als rho. Falls dies gelingt, wird mit 2. weitergefahren.

4. Ist die Distanz von i zu allen Leitaufnahmen grösser als rho, so wird i zur Leitaufnahme einer neuen Gruppe. Das Verfahren wird mit 2. fortgesetzt.

5. Uebersteigt die Zahl der Gruppen NMAX, so werden die noch nicht verarbeiteten Aufnahmen nicht klassifiziert. Rho sollte etwas vergrössert werden. Der ganze Gruppierungsprozess ist zu wiederholen (1.), bis alle Aufnahmen klassifiziert sind.

Der Ablauf lässt sich noch vereinfachen, indem die Zahl der Gruppen nicht festgelegt wird. Schwerwiegendster Nachteil der Methode ist, dass das Resultat von der Reihenfolge der Eingabe der Aufnahmen abhängt. Das illustrieren die Beispiele in Abb. 5.3. Rho sei gleich 2. Im Falle A sei die Reihenfolge der Aufnahmen (1,2,3,4,5). Die erste in den Prozess eingeschleuste Aufnahme wird zur Leitaufnahme der ersten Gruppe. Aufnahme 2 besitzt Distanz d = 1 zu dieser und kommt damit ebenfalls in Gruppe 1. Aufnahme 3 besitzt d $= 5^{1/2} = 2.24$ zu Aufnahme 1 und wird Leitaufnahme der Gruppe 2. Aufnahme 4 wird letzterer zugeordnet und Aufnahme 5 schliesslich bildet eine selbständige Gruppe 3. Im Falle B soll versuchsweise die Reihenfolge der Aufnahmen so geändert werden, dass Nummer 3 an erster Stelle steht. Es gilt also 3,1,2,4,5. Leitaufnahmen werden zunächst 3 und 1. Aufnahme 2 wird der Leitaufnahme der Gruppe 1 zugeordnet, also Aufnahme 3. Auch Aufnahme 4 gehört zu Gruppe 1, während Aufnahme 5 eine neue Gruppe 3 bildet.

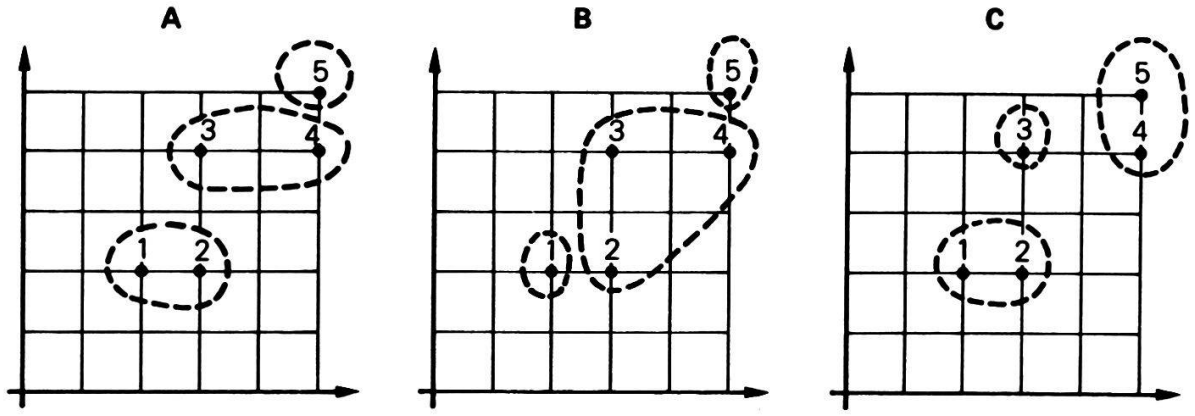

Abb. 5.3 Verschiedene Lösungen des Leader-Algorithmus. A und B ergeben sich durch unterschiedliche Reihenfolge in der Verarbeitung der Aufnahmen. Lösung C resultiert, wenn der maximal zulässige Durchmesser der Gruppen (rho) reduziert wird.

Tabelle 5.1

Beispiel für die Durchführung der Assoziationsanalyse (A). Aufbau der Kontingenztafel zur Berechnung von Chiquadrat zwischen den Arten 1 und 2 (B).

A

Art	Aufnahme 1	2	3	4	5
1	1	1	0	0	0
2	1	1	1	0	0
3	1	1	1	1	0
4	0	0	0	1	1

B

Art 1 / Art2	+	-	
+	a=2	b=1	a+b=3
-	c=0	d=2	c+d=2
	a+c=2	b+d=3	N = 5

Das Resultat der Analyse ist stark von der Wahl von rho abhängig. Setzt man z.B. rho = 1.5, so erhält man die visuell leicht als optimal erkennbare Lösung mit den Gruppen (1,2), (3), (4,5). Für das LEADER-Verfahren gibt es zahlreiche Verbesserungsvorschläge und Alternativen (ANDERBERG 1973), doch sind die Resultate immer von der ursprünglichen Reihenfolge der Aufnahmen abhängig. Heuristische Verfahren werden ihrer Einfachheit wegen eingesetzt, um sehr grosse vegetationskundliche Datensätze zu strukturieren. Das Beispiel in Abb. 5.3 zeigt deren Grenzen auf. Eine provisorische Gliederung ist aber doch zu erzielen. Sie kann in einem anschliessenden Schritt korrigiert und optimiert werden (z.B. im Programm CLUSLA, LOUPPEN und VAN DER MAAREL (1979)).

5.3 Teilungsverfahren

5.3.1 Assoziationsanalyse

Bei Teilungsverfahren wird versucht, für die Gesamtstichprobe von Individuen eine möglichst sinnvolle Unterteilung zu finden. Als typischer Vertreter soll zuerst die Assoziationsanalyse dargestellt werden, die von WILLIAMS und LAMBERT (1959) publiziert und später mehrfach variiert wurde. Es handelt sich um ein monothetisches Verfahren, d.h. jede Unterteilung wird nur aufgrund der Gegenwart oder des Fehlens einer einzelnen Art durchgeführt. Immerhin wird zur Auswahl dieser Art die Gesamtähnlichkeitsstruktur der Vegetationstabelle berücksichtigt. Ausgangspunkt bildet eine Aehnlichkeitsmatrix S der Arten mit den Elementen

$$S_{ij} = \chi^2_{ij}/N.$$

Das Chiquadrat berechnet sich wie in Kapitel 4.5 gezeigt. N ist die Anzahl Aufnahmen der Vegetationstabelle, die Division durch N hat daher auf das Ergebnis keinen Einfluss. Anhand von Tabelle 5.1, A, soll die Berechnung gezeigt werden. Für die Arten 1 und 2 konstruiert man die Kontingenztafel der Tabelle 5.1, B. Daraus ergibt sich

$$\chi^2(1,2) = \frac{(ad - bc)^2 N}{(a+b)(a+c)(c+d)(b+d)} = \frac{(4 - 0)^2 5}{3*2*2*3} = \frac{80}{36} = 2.22$$

In der Originalversion der Assoziationsanalyse wird die Unabhängigkeit der Arten - unbesehen von Voraussetzungen der schliessenden Statistik - anhand einer χ^2-Tabelle auf Signifikanz getestet (Anzahl Freiheitsgrade df = 1). Da N in unserem Beispiel sehr klein ist, müsste χ^2 dazu erst korrigiert werden. Entsprechende Formeln finden sich bei MÜLLER-DOMBOIS und ELLENBERG (1974) und PIELOU (1977). Nicht signifikante Werte werden meist durch null ersetzt. Als Element der S - Matrix erhalten wir zum Beispiel

$$S(1,2) = \chi^2(1,2)/N = 2.22/5 = 0.44$$

In gleicher Weise werden nun die andern Elemente berechnet. Man setzt die Werte der Diagonalen gleich null und erhält

$$S = \begin{matrix} 0 & 0.44 & 0.166 & 0.44 \\ 0.44 & 0 & 0.375 & 1.00 \\ 0.166 & 0.375 & 0 & 0.375 \\ 0.44 & 1.00 & 0.375 & 0 \end{matrix}$$

Wie in Kapitel 4.5 erwähnt, lässt sich χ^2 in relativierter Form als Korrelationskoeffizient für die Kontingenztafel verstehen (V - Wert). Um nun diejenige Art zu finden, die den grössten gemeinsamen Zusammenhang mit allen andern Arten aufweist, müssen nur die Kolonnen (oder Zeilen) in S aufaddiert zu werden. Wir erhalten

$$S_{.j} = 1.046,\ 1.815,\ 0.916,\ 1.815$$

Das maximale Chiquadrat weisen Art 2 und 4 auf. Anhand ihrer Gegenwart oder Abwesenheit in den Aufnahmen wird Tabelle 5.1, A, unterteilt in die Gruppen (1,2,3) und (4,5). Das Verfahren geht weiter, indem beide der neu gebildeten Gruppen in derselben Weise analysiert werden. Da Art 2 sicher zu keiner weiteren Unterteilung zu verwenden ist,

kann sie (und im vorliegenden Fall auch Art 4) weggelassen werden. Die beiden neuen S-Matrizen, welche zur Suche der nächsten zwei Trennarten dienen, haben deshalb immer eine um 1 verminderte Dimension. Beim vorliegenden Beispiel ist es jedoch sinnvoll, die Analyse hier abzubrechen.

Die Gruppen der Assoziationsanalyse sind streng hierarchisch gegliedert. Weil sie auf eindeutigen Trennarten beruhen, lassen sie sich sehr leicht identifizieren. Für eine einfache Unterteilung wird eine einzige Art benötigt, für vier Gruppen mindestens zwei Arten, für acht Gruppen mindestens drei Arten usw. Da aber eine einzige Art über die Gruppenzugehörigkeit entscheidet, ist das Verfahren stark von Zufälligkeiten abhängig. Wie zu erwarten ist, befriedigt die so entstehende Struktur pflanzensoziologisch kaum, krasse Fehlklassifikationen wegen unerwartet auftretender oder abwesender Arten sind die Regel. Günstiger liegen die Verhältnisse, wenn statt der Artmächtigkeiten Koordinaten aus Hauptkomponentenanalysen verwendet werden (NOY MEIR 1973). Auf diese Weise entfällt jedoch die Möglichkeit einfacher Identifikation. Um pflanzensoziologische Einheiten trotzdem mit Hilfe weniger Arten charakterisieren zu können, bieten sich Rangierungsmethoden an (vgl. dazu Kap. 6 sowie HILL (1979b)).

Die Assoziationsanalyse entspricht ihrer Idee nach dem Konzept der Charakter- und Differentialarten des Systems Braun-Blanquet. Wohl aus diesem Grunde hat sie einen bemerkenswerten Bekanntheitsgrad erlangt.

5.3.2 Gridanalyse

Die Gridanalyse versucht, die hauptsächlichsten Nachteile der Assoziationsanalyse zu vermeiden (WILDI 1979). Aus diesem Grunde soll nicht eine, sondern es sollen mehrere Dimensionen gleichzeitig zur Abgrenzung von Gruppen beigezogen werden. Gesucht werden echte Gruppen, welche einer lokalen Anhäufung ähnlicher Individuen entsprechen (Noda im Sinne von POORE 1955). Wir verfolgen den Algorithmus anhand von Abb. 5.4. Der Uebersichtlichkeit halber wird nur ein

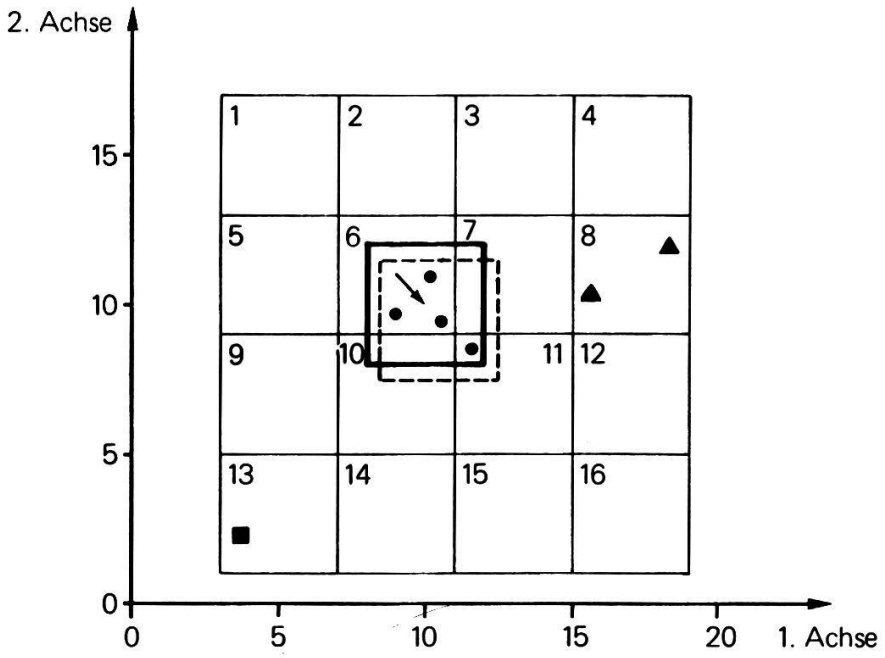

Abb. 5.4 GRID-Analyse im zweidimensionalen Fall.

zweidimensionaler Fall dargestellt.

x_{ij} sei die Koordinate bezüglich der i-ten Achse (i = 1,...,p ; Achsen können Arten, Standortsmessungen, Ordinationsachsen usw. sein) der Aufnahme j, j = 1,...,n. Es sind folgende Operationen durchzuführen:

1. Für jede Dimension ist der Bereich δ_i festzulegen.

   $$\delta_1 = \max(x_{1j}) - \min(x_{1j}),\ j = 1,...,n$$

   $$\delta_2 = \max(x_{2j}) - \min(x_{2j}),\ j = 1,...,n$$

   In Abb. 5.4 ist $\delta_1 = 15$, $\delta_2 = 10$.

2. Bestimme $\delta_{max} = \max(\delta_1,...,\delta_p)$. In Abb. 5.4 ist $\delta_{max} = \delta_1 = 15$.

3. Lege ein p-dimensionales Netz mit der Seitenlänge von mindestens δ_{max} über den Aufnahmeraum, so dass alle Individuen innerhalb des Netzes liegen. Die Auflösungskraft wird durch die Anzahl Unterteilungen jeder Dimension bestimmt (m = 4 in Abb. 5.4).

4. In den entstandenen $m^p = 16$ Zellen werden die Individuen gezählt. Wir erhalten

Zelle	Individuen
6	3
8	2
11	1
13	1

5. Die individuenreichste, noch nicht verarbeitete Zelle (Nr. 6) wird geprüft, ob sie ein lokales Dichtemaximum enthält, also Zentrum einer echten Gruppe darstellt. Dazu wird kontrolliert, ob nicht eine der Nachbarzellen (2,5,7,10) bereits ein Zentrum enthält. Im Beispiel bleibt Zelle Nr. 6 Zentrum der ersten Gruppe.

6. Weil das p-dimensionale Netz willkürlich in den Raum gelegt wurde, wird nun die genauere Lage der Gruppe 1 gesucht. Dazu ist die aktuelle Zelle in Richtung des Schwerpunkts der in ihr enthaltenen Individuen zu verschieben (Pfeil, fett ausgezogenes Quadrat). Enthält die verschobene Zelle mehr Individuen, so wird weiter geschoben. In Abb. 5.4 kommt ein Individuum in Zelle 11 dazu, so dass das erste Gruppenzentrum durch 4 Individuen repräsentiert wird.

7. Nun wird die Zelle mit der nächstniedrigeren Anzahl Individuen verarbeitet (Schritte 5 und 6). Dies wäre nach dem ersten Durchgang Zelle 8.

Der Prozess wird so lange fortgesetzt, bis alle Zellen abgearbeitet sind. Einzelindividuen können den Status einer unabhängigen Gruppe erhalten, oder aber dem nächstgelegenen Gruppenzentrum zugeordnet werden. In Abb. 5.4 erhalten wir somit 3 Gruppen, erkennbar an den unterschiedlichen Symbolen.

Die GRID-Analyse liefert natürliche, nicht hierarchische Gruppen. Dies allerdings nur, wenn die Auflösung zweckmässig gewählt wird. Man sollte daher zuerst mit einer niedrigen Zellenzahl (d.h. niedriger Auflösung) beginnen. Steigt die Gruppenzahl mit stetig wachsender Zellenzahl sprunghaft an, so sind vermutlich echte Gruppen halbiert worden. Im weitern ist zu bemerken, dass die Zahl der Dimensionen aus Gründen der Uebersichtlichkeit nicht zu gross gewählt werden sollte. Das Verfahren hat sich bewährt, wenn statt einfacher Artmächtigkeiten aus Hauptkomponentenanalysen stammende Koordinaten verwendet werden (WILDI 1979).

Die GRID-Analyse eignet sich zum Auffinden diskreter Gruppen. Sie hat gegenüber anderen Verfahren den Vorteil, dass lokale Verdichtungen von Punkten als Gruppenzentren interpretiert werden, sodass intermediäre Punkte (vgl. Abb. 5.2) nicht störend wirken. Ist der zu untersuchende Datensatz genügend gross, so können Gruppen fast beliebiger Form entdeckt werden. Beschränkungen ergeben sich aus der prak-

tisch begrenzten Anzahl von Dimensionen, die sich noch sinnvoll bearbeiten lassen. Die GRID-Analyse ist deshalb nicht geeignet, um strukturelle Details zu untersuchen.

5.4 Agglomerative Verfahren

Bei dieser Kategorie von Gruppierungsverfahren werden schrittweise Individuen - später Gruppen von Individuen - zu neuen Gruppen zusammengeschlossen. Dabei können Dendrogramme gebildet werden. Diese dienen der übersichtlichen Darstellung von Resultaten hierarchischer Gruppierungsmethoden divisiver oder agglomerativer Art. Die Sachverhalte lassen sich anhand der einfachsten agglomerativen Methode demonstrieren, nämlich der Single Linkage Analysis (ANDERBERG 1973).

5.4.1 Single Linkage Analysis

Das Prinzip der Single Linkage Analysis lässt sich am univariaten Fall verfolgen. Abb. 5.5, A zeigt ein Beispiel. Jede der 4 Aufnahmen wird charakterisiert durch ein einziges Merkmal, entsprechend den folgenden Werten:

Merkmal	Aufn.	1	2	3	4
1		2	4	7	8

Im ersten Schritt muss nun eine Aehnlichkeitsmatrix berechnet werden. Die Euklidsche Distanz hat gegenüber andern Massen den Vorteil, dass sie direkt aus Abb. 5.5, A, herausgelesen werden kann. Für $d_{1,2}$ erhält man 2, für $d_{2,3}$ = 3, $d_{3,4}$ = 1 usw. Die Gesamtähnlichkeitsstruktur ergibt:

$$D = \begin{matrix} 0 & 2 & 5 & 6 \\ 2 & 0 & 3 & 4 \\ 5 & 3 & 0 & 1 \\ 6 & 4 & 1 & 0 \end{matrix}$$

Die Gruppierung erfolgt nach einer einzigen Vorschrift: Es sind stets jene 2 Gruppen zu einer neuen Gruppe zusammen-

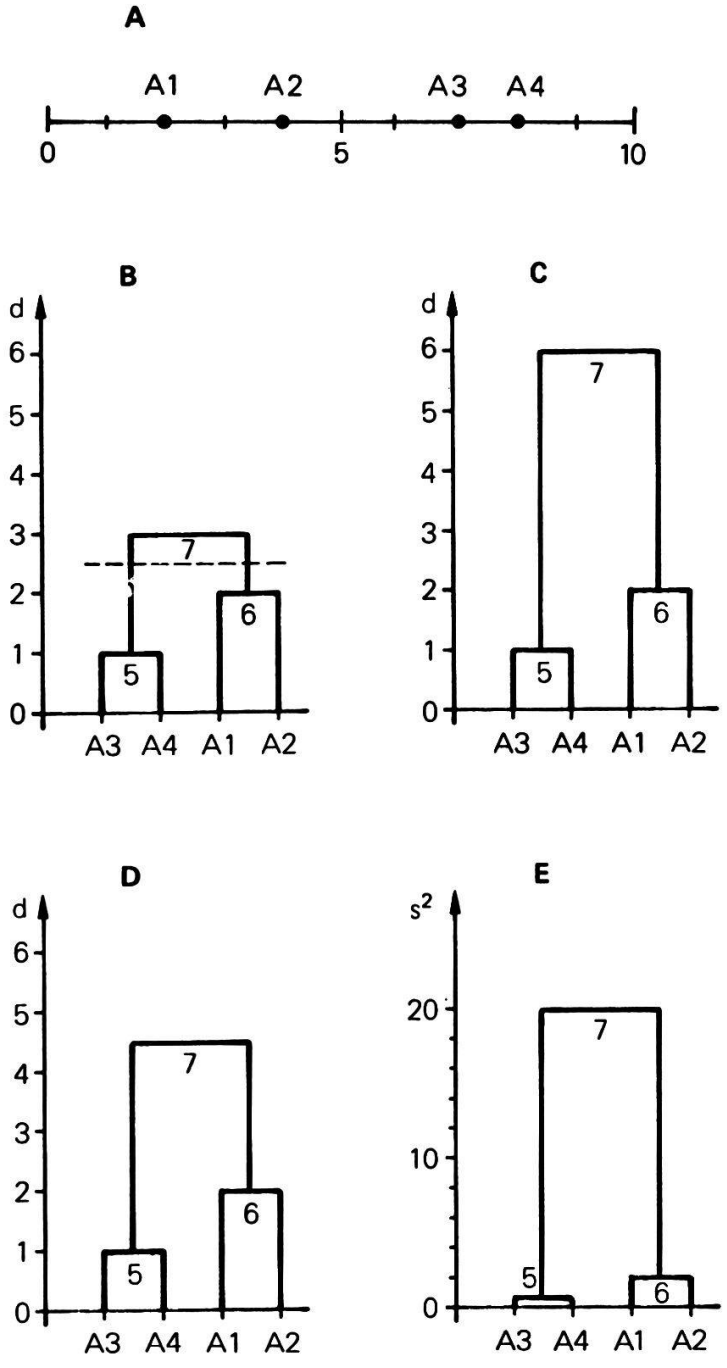

Abb. 5.5 Vier univariat charakterisierte Punkte (A) mit verschiedenen Methoden gruppiert: Single Linkage Analysis (B), Complete Linkage Analysis (C), Average Linkage Analysis (D) und Minimalvarianzanalyse (E).

zufassen, welche sich bezüglich ihrer **ähnlichsten** Individuen am nächsten stehen. Zu Beginn der Analyse finden wir in Abb. 5.5 4 Gruppen, die aus je einem einzigen Individuum bestehen. Der grafischen Darstellung, aber auch der Distanzmatrix entnehmen wir, dass A3 und A4 am ähnlichsten sind und mit $d_{3,4} = 1.0$ zu einer neuen Gruppe zusammengeschlossen werden können. Diese soll zweckmässigerweise die Nummer 5 erhalten. Im Dendrogramm (Abb. 5.5, B) sind die beiden Aufnahmen als erste aufgezeichnet und durch einen Bügel der Höhe d = 1 miteinander verbunden. Der zweite Zusammenschluss gestaltet sich komplizierter. Zu prüfen sind nun noch 3 Distanzen, nämlich $d_{1,2}$, $d_{2,5}$ und $d_{1,5}$. $d_{1,2}$ kann sofort der Distanzmatrix entnommen werden. Für $d_{1,5}$ gilt bei der Single Linkage Analysis, dass der Wert von $d_{1,3} = 5$ gewählt werden muss, da A3 der Gruppe 1 näher steht als A4. Entsprechend gilt $d_{2,5} = d_{2,3} = 3$. Zusammengefasst erhält man:

$$d_{1,2} = 2$$
$$d_{1,5} = 5$$
$$d_{2,5} = 3.$$

Damit werden A1 und A2 als nächste Gruppe, Nr. 6, auf dem Niveau d = 2 zusammengeschlossen. Nun müssen noch Gruppe 5 und 6 zusammengeschlossen werden. Um das Niveau des Zusammenschlusses zu finden, ist die gesamte Distanzmatrix zu durchsuchen. Für $d_{5,6}$ kommen folgende Werte in Frage:

$$d_{1,3} = 5$$
$$d_{1,4} = 6$$
$$d_{2,3} = 3$$
$$d_{2,4} = 4$$

Als nächststehende Nachbarn der beiden Gruppen qualifizieren sich A2 und A3 mit $d_{2,3} = 3$. Auf diesem Niveau wird in Abb. 5.5, B die neue Gruppe 7 gebildet. Damit ist die Analyse abgeschlossen. Ihr Resultat ist ein Dendrogramm, welches über die Gruppenstruktur der Stichprobe (A1,A2,A3,A4) Aufschluss gibt. Meist besteht das Ziel der Analyse darin, eine bestimmte Anzahl von Gruppen, sagen wir 2, zu generieren. Zu diesem Zwecke ist das Dendrogramm zu zerschneiden, und zwar

zwischen d = 2 und d = 3 (gestrichelte Linie in Abb. 5.5, B). Es resultieren (A3,A4) und (A1,A2) als Gruppen.

Die Anwendungsmöglichkeiten der Single Linkage Analysis sollen beim Vergleich verschiedener Methoden erörtert werden (Kap. 5.4.5). Hier ist noch zu vermerken, dass es zahlreiche Varianten und Erweiterungen gibt. JANCEY (1974) schlägt eine Methode vor, bei welcher die Anzahl resultierender Gruppen vorzugeben ist. Kann aufgrund der Stichprobenstruktur eine Lösung mit natürlichen Gruppen gefunden werden, so erfolgt die Unterteilung des Dendrogrammes automatisch. Andernfalls wird die Zahl der Gruppen durch den Algorithmus selbst verändert.

Eine herausragende Rolle spielt die Single Linkage Analysis in der Geographie, und zwar vor allem im zweidimensionalen Fall zur Lösung des "Nächster Nachbar"- Problems. Abb. 5.6 zeigt eine Karte mit 5 Ortschaften. Diese sind so miteinander zu verbinden, dass

1. jeder Punkt mindestens einmal verbunden wird;
2. keine Schleifen auftreten;
3. die Summe aller Verbindungsstrecken minimal ist.

Nach GOWER und ROSS (1969) liefert die Single Linkage Analysis unmittelbar die Lösung. Das so entstehende Gebilde (Abb. 5.6) heisst Minimalbaum. Es kann derart unterteilt werden, dass bei pflanzensoziologischen Datensätzen ausgesprochene Gradientenstrukturen aufzufinden sind (KUHN 1983).

5.4.2 Complete Linkage Analysis

Die Complete Linkage Analysis ergibt sich durch ganz geringe Aenderung des Single Linkage Algorithmus. Die Vorschrift lautet: Es sind stets jene 2 Gruppen zu einer einzigen, neuen Gruppe zusammenzufassen, welche sich bezüglich ihrer **unähnlichsten** Individuen am nächsten stehen. Am Beispiel der Abb. 5.5, C, soll der Ablauf verfolgt werden. Der erste Schritt (Gruppe 5) verläuft dabei genau gleich wie bei der

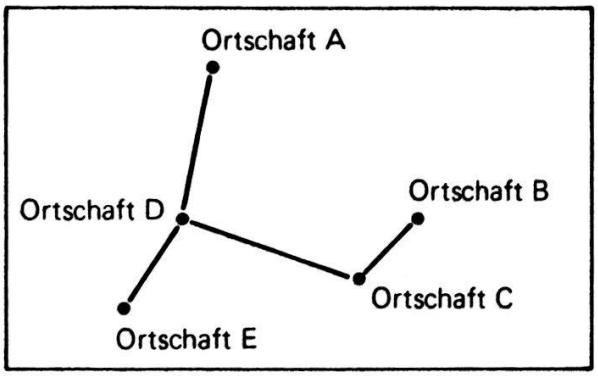

Abb. 5.6 Lösung des Problems "nächster Nachbar" im zweidimensionalen (geografischen) Falle.

Single Linkage Analysis, was sich auch im Dendrogramm niederschlägt. Auch im zweiten Schritt sind wieder die 3 Distanzen $d_{1,2}$, $d_{1,5}$ und $d_{2,5}$ zu prüfen. $d_{1,2}$ kann direkt an der Distanzmatrix abgelesen werden. Für $d_{1,5}$ ist im Gegensatz zur Single Linkage Analysis der Maximalabstand massgebend, nämlich $d_{1,4} = 6$. Entsprechend gilt neu $d_{2,5} = d_{2,4} = 4$. Zusammenfassend gilt:

$$\begin{aligned} d_{1,2} &= 2 \\ d_{1,5} &= 6 \\ d_{2,5} &= 4 \end{aligned}$$

A1 und A2 werden zur Gruppe 6 zusammengeschlossen, und zwar auf dem Niveau d = 2. Für die Bildung der Gruppe 7 müssen wiederum die Maximalabstände gesucht werden. Wir finden

$$d_{5,6} = \max\ (d_{1,3}, d_{1,4}, d_{2,3}, d_{2,4})$$

$$= \max\ (5,6,3,4) = d_{1,4} = 6$$

Dieser Wert wird zur Vollendung des Dendrogrammes in Abb. 5.5, C, verwendet.

5.4.3 Average Linkage Analysis

Im Vergleich zu den eben beschriebenen Methoden handelt es sich dabei um eine mittlere, gemässigte Lösung. Statt der Maximal- oder Minimaldistanz zwischen Individuen verschiedener Gruppen, wird als Kriterium für den Zusammenschluss die mittlere Distanz gewählt. So beträgt dann der massgebende Abstand zwischen den Gruppen 2 (A2) und 5 (A3,A4) $d_{2,5} = 3.5$, wie aus Abb. 5.5, A sofort ersichtlich wird. Alle übrigen Operationen sind mit denjenigen der schon beschriebenen Verfahren identisch. Für den Leser, der die Analyse nachvollziehen will, wird in Abb. 5.5, D das resultierende Dendrogramm gegeben.

5.4.4 Minimalvarianz-Analyse

Im Gegensatz zu den bisher besprochenen Methoden beruht die Minimalvarianz - Analyse (ORLOCI 1967) auf den Streuungsverhältnissen der Gruppen. Der Zusammenschluss bestehender Gruppen zu grösseren, neuen, erfolgt stets so, dass die gruppeninterne Varianz möglichst wenig zunimmt. Das Verfahren beruht mithin auf den Konzepten der Varianzanalyse.

Wir beginnen die Betrachtung mit der Definition der gruppeninternen Varianz Q_g (vergleiche dazu auch die Ausführungen in Kapitel 6.3). Diese ist gleich der Summe der quadrierten Abstände jedes Gruppenindividuums zum Gruppenzentrum:

$$Q_g = \sum_{i=1}^{p} \left(\sum_{j=1}^{n_g} (x_{ij} - \bar{x}_i)^2 \right)$$

Darin ist x_{ij} die Koordinate (Artmächtigkeit) der Art i in der Aufnahme j, n_g die Anzahl Aufnahmen der Gruppe g, $\bar{x}_i$ der Mittelwert aller Arten in g und p die Anzahl Arten. Q_g lässt sich rascher berechnen aus der Matrix der quadrierten Distanzen:

$$Q_g = \frac{1}{n_g} \sum_{i<j} d_{ij}^2$$

Den formalen Nachweis für diesen Zusammenhang zeigt z.B. PIELOU (1977), S. 319 f. Σ für $i<j$ bedeutet, dass alle $n_g*(n_g-1)/2$ Elemente der D^2 Matrix, welche sich auf die Individuen der Gruppe g beziehen, summiert werden. Verwenden wir wieder das Beispiel aus Abb. 5.5, so müssen die Distanzen zuerst quadriert werden:

$$D^2 = \begin{matrix} 0 & 4 & 25 & 36 \\ 4 & 0 & 9 & 16 \\ 25 & 9 & 0 & 1 \\ 36 & 16 & 1 & 0 \end{matrix}$$

Die linke untere Hälfte der Matrix ist zu vernachlässigen, sodass jedes Element nur einfach gezählt wird. Als Beispiel erhalten wir für

$$Q_{2,3,4} = 1/3\ (9+16+1) = 26/3 = 8\ 2/3.$$

$Q_{2,3,4}$ ist die interne Varianz der Gruppe (2,3,4). ORLOCIs (1967) Kriterium zur Fusion zweier Gruppen A und B lautet nun, dass die Zunahme der Varianz Q(A,B) minimal sein soll, wobei gilt

$$Q(A,B) = Q(A+B) - Q(A) - Q(B).$$

Q(A+B) ist die Varianz der neu zu bildenden Gruppe, Q(A) und Q(B) sind diejenigen der alten Gruppen. Damit ist Q(A,B) jener Betrag, um welchen die Varianz beim Zusammenschluss von A und B vermehrt wird.

Für den ersten Zusammenschluss brauchen bloss die Elemente der D^2-Matrix nach dem kleinsten Wert abgesucht zu werden. Als Minimum qualifiziert sich

$$Q(3,4) = 1/2\ (1) = 0,5 = Q(5)\ .$$

Für den nächsten Zusammenschluss müssen zuerst alle Q(A,B)-Werte berechnet werden. Man erhält:

$$Q(1,2) = 1/2\ (4) = 2$$

$$Q(1,5) = \frac{1}{n_1+n_5}\left(d_{1,3}^2+d_{1,4}^2+d_{3,4}^2\right) - Q(1) - Q(5)$$

$$= 1/3\ (25+36+1) - 0 - 1/2 = 20\ 2/3 - 1/2 = 20\ 1/6$$

$$Q(2,5) = 1/3\ (9+16+1) - 0 - 1/2 = 8\ 2/3 - 1/2 = 8\ 1/6$$

In der Matrix-Schreibweise gilt

$$Q = \begin{matrix} 0 & 2 & 20\ 1/6 \\ 2 & 0 & 8\ 1/6 \\ 20\ 1/6 & 8\ 1/6 & 0 \end{matrix}$$

Als neue Gruppe 6 qualifizieren sich die Individuen (1,2) mit Q = 2. Schliesslich ist für Gruppe 7 Q(5,6) zu berechnen:

$$Q(5,6) = 1/4\ (4+25+36+9+16+1)\ -1/2\ -2 = 20\ 1/4$$

Werden die eben gefundenen Q-Werte auf der y-Achse aufgetragen, so erhält man das Dendrogramm in Abb. 5.5, E.

5.4.5 Besonderheiten agglomerativer Verfahren

Beim Vergleich der hier gezeigten Methoden anhand des kleinen 4-Punkte Beispiels (Abb. 5.5) ist bemerkenswert, dass das Resultat stets gleich ausfällt. In der Tat kann generell gesagt werden, dass die zu erwartenden Unterschiede bei den meisten Datenstrukturen klein sind. In der Anfangsphase funktionieren alle vier besprochenen Methoden gleich: Zuerst werden die nächsten Nachbarn zu Zweiergruppen zusammengefasst. Dies ist auch der konzeptionell schwächste Teil agglomerativer Verfahren. Die Lage einzelner Punkte, welche stets gewissen Zufälligkeiten unterworfen ist, entscheidet wesentlich über das Resultat. Erst bei steigender Individuenzahl pro Gruppe treten die Verschiedenheiten deutlicher hervor. Die wesentlichsten Unterschiede betreffen:

a) <u>Die Tendenz zur Kettenbildung</u>
Es handelt sich um eine typische Eigenschaft der Single Linkage Analysis. Gradienten bildende, beliebig lange Reihen von Aufnahmen können als eigenständige Gruppen erkannt werden. Gerade gegenteilig verhält sich die Complete Linkage Analysis. Bei ihr wachsen die Gruppen fast gleichförmig um die zu Beginn gefundenen Zentren. Sie unterteilt langgestreckte Reihen in mehrere, gedrungene Gruppen und eignet sich damit besser für Daten, die aus natürlichen Gruppen

zusammengesetzt sind. Abb. 5.7 illustriert solche Verhältnisse. Sie zeigt den (künstlich konstruierten Fall) einer Stichprobe von drei Gruppen, wobei eine davon eine lange Kette bildet. Die einzige Methode, welche diese Struktur erfolgreich aufdeckt, ist die Single Linkage Analysis. Complete Linkage, Average Linkage und Minimalvarianz-Analyse sind dazu nicht in der Lage.

Die Vermutung liegt nahe, die Single-Linkage Analysis könnte sich zur Strukturierung von Daten einer Gradientenanalyse eignen. Es ist aber zu beachten, dass in der Praxis selten einfache Punkteketten auftreten. Sobald kompliziertere Konfigurationen vorliegen, lassen sich sphärische oder elliptische Gruppen leichter interpretieren.

b) Reaktion auf Einzelindividuen

Wie oben dargelegt, bestimmt in der Anfangsphase stets die Lage einzelner Individuen den Gang der Analyse. Erst bei fortschreitendem Gruppierungsprozess unterscheiden sich die Methoden. Single und Complete Linkage Analysis verwenden als Kriterium für Zusammenschlüsse bloss die Distanz zu einem einzigen Individuum einer Gruppe. Beide Methoden sind deshalb wenig robust. Average Linkage Analysis und das Minimalvarianzverfahren berücksichtigen dagegen die Lage aller Individuen einer Gruppe. Sie tragen mithin der gesamten Gruppenstruktur Rechnung.

c) Berücksichtigung der Gruppengrösse

Die einzige der hier erwähnten Methoden, welche auch die Anzahl der Punkte einer Gruppe als Zusammenschlusskriterium verwendet, ist die Minimalvarianzanalyse. Abb. 5.8 zeigt zwei Fälle, die von der Average Linkage Analyse gleich interpretiert werden. Wir nehmen an, dass die beiden Gruppen A und B zusammengeschlossen werden sollen. Die alten Gruppenzentren liegen im rechten und linken Beispiel gleich weit auseinander. In der Minimalvarianzanalyse ist jedoch nicht dieser Abstand massgebend, sondern die Varianz der neuen Gruppe (A,B). Da die Lage des Zentrums von der Anzahl Individuen in A und B abhängig ist, liegt es rechts näher bei der grösseren Gruppe, A. In der Minimalvarianzanalyse wird

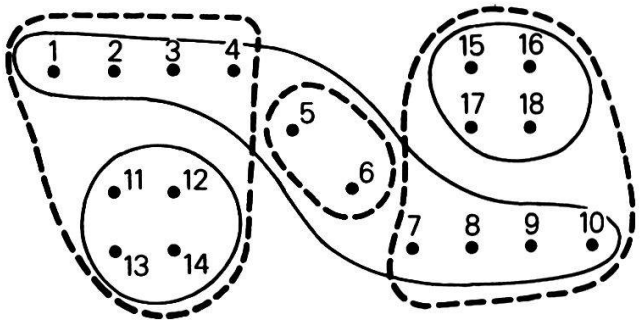

Abb. 5.7 Unterschiede zwischen verschiedenen agglomerativen Clusterverfahren am Beispiel der Aehnlichkeitsstruktur von Abb. 5.2, C: Gruppenbildung von Single Linkage Analysis (geschlossene Linie), Gruppenbildung von Complete Linkage Analysis (gestrichelt).

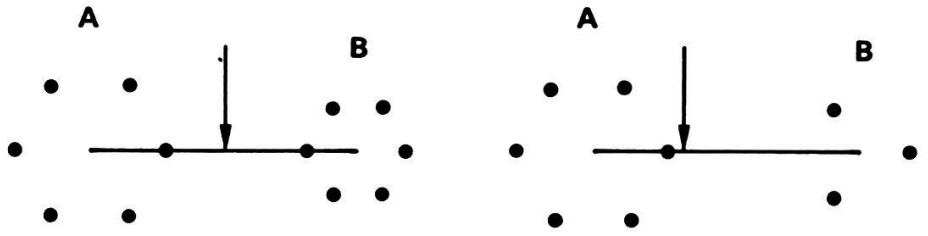

Abb. 5.8 Durch Average Linkage- und Minimalvarianzanalyse verschieden behandelte Fälle eines Gruppenzusammenschlusses. Für die Average Linkage Analyse sind die Zusammenschlüsse links und rechts gleich günstig. Aus der Sicht der Minimalvarianzanalyse erhält derjenige rechts den Vorzug.

der Zusammenschluss rechts wegen der kleineren Varianz die günstigere Lösung sein. Das ist für die meisten Anwendungen sinnvoll. Je mehr Punkte an einer Gruppe beteiligt sind, desto eher kann davon ausgegangen werden, dass es sich nicht um einen Artefakt handelt und dass sie deshalb eigenständig ist. Erwartet man dagegen in einem Datensatz Ausreisser und möchte diese sicher von echten Gruppen abgetrennt haben, so ist die Complete Linkage Analyse angemessen.