Bewertete Hausaufgabe 2020 in Statistisches Data Mining

Philipp Rieser

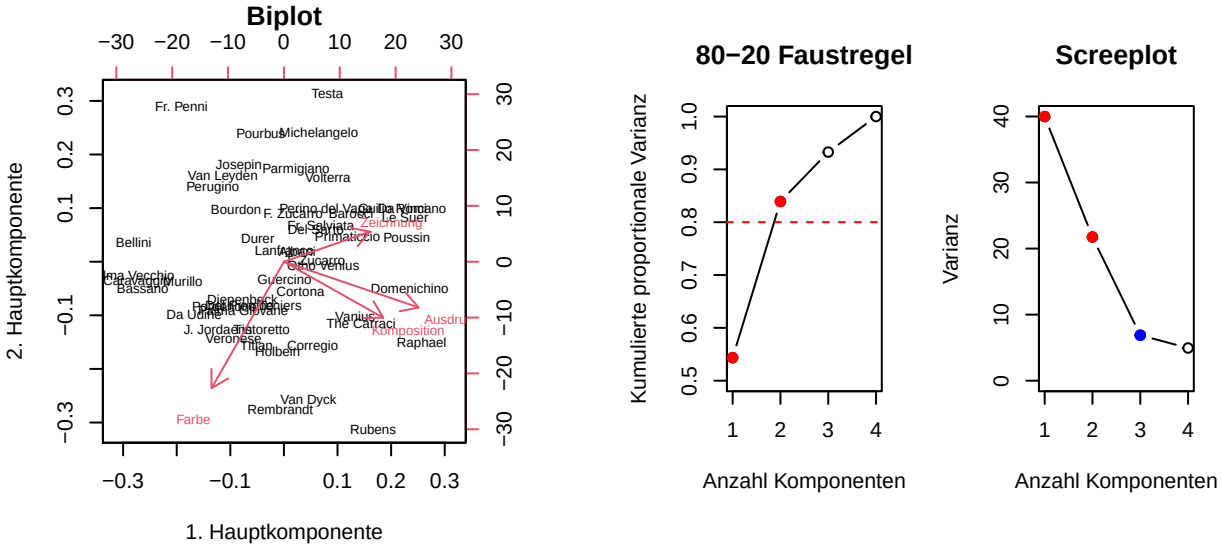
Aufgabe 1: Hauptkomponenten-Analyse

```
km.pc <- prcomp(KMha20, scale = FALSE)
```

i) Wieso Kovarianzmatrix?

Da die Bewertungen der Künstler alle die gleichen Einheiten aufweisen, ist es nicht nötig die Daten zu standardisieren. Dies bedeutet man verwendet die Kovarianzmatrix für die Hauptkomponenten-Analyse. Standardisierung könnte zum Verlust von wichtigen Informationen führen.

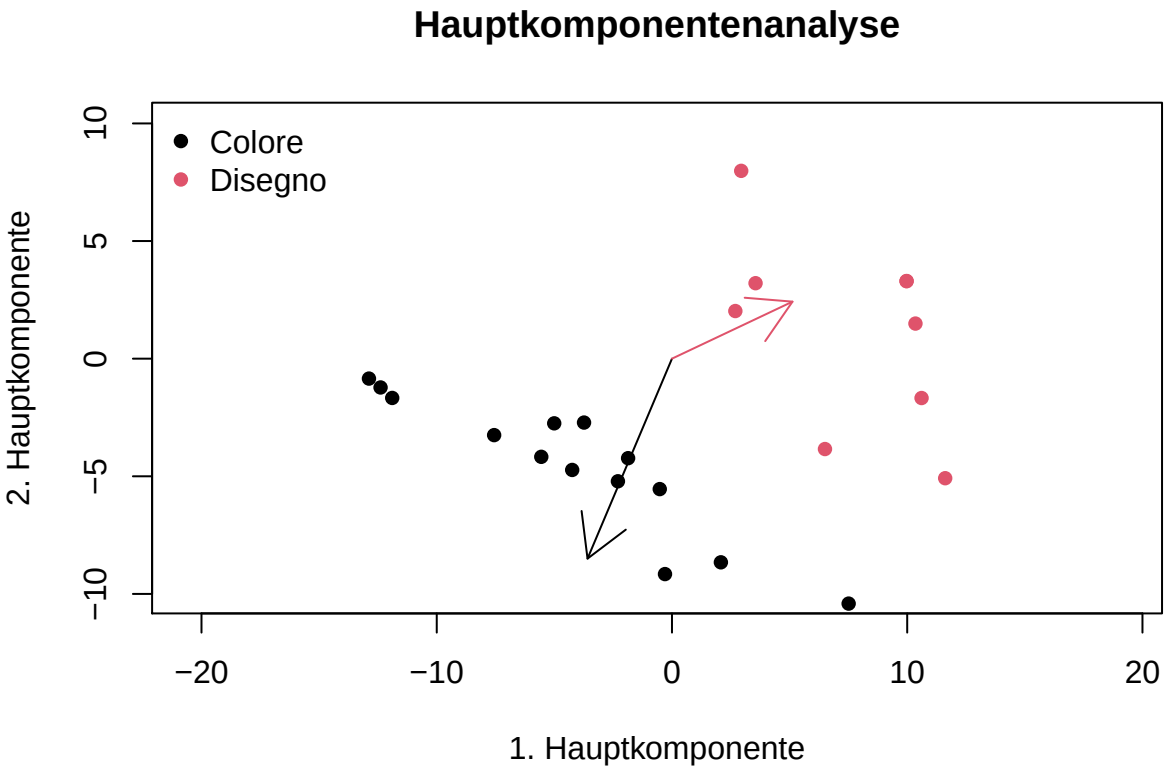
ii) Visuelle Darstellung der Hauptkomponenten-Analyse



Güte: In der obigen Darstellung ist der Biplot und rechts die Plots für die Güte der Analyse aufgeführt. Laut der 80-20 Faustregel (mittlerer Plot) liegt man mit 2 Komponenten oberhalb der rot gestrichelten 80%-Linie, somit würden 2 Hauptkomponenten für die ädaquate Darstellung der Variabilität genügen. Der Ellbogen im Screeplot (rechts) liegt bei der 3. Komponente (blauer Punkt). Folglich genügen die ersten beiden Hauptkomponenten.

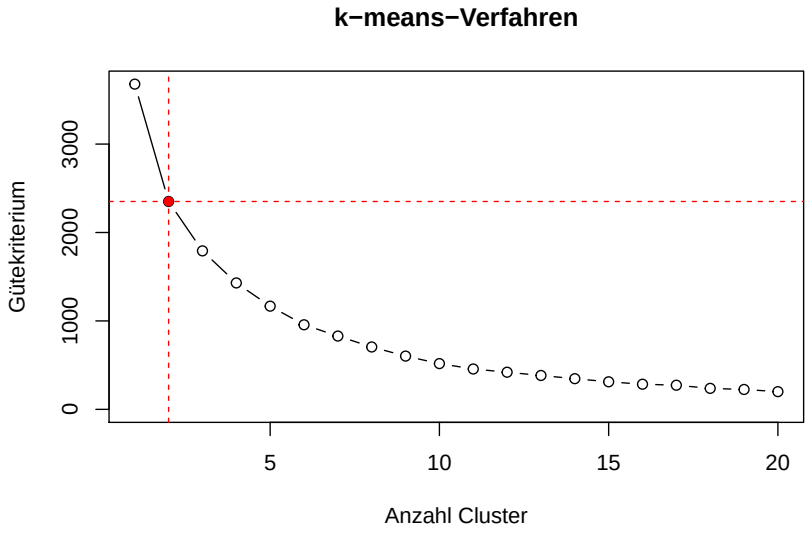
Interpretation: In der ersten Hauptkomponente scheint der Einfluss der Variable *Ausdruck* am grössten zu sein. *Farbe* hat einen leicht negativen Wert. Die zweite Hauptkomponente weist bei Variable *Farbe* einen starken negativen Wert auf, somit ist diese Variable bei beiden Hauptkomponenten negativ. Künstler die in der Kategorie *Farbe* besser bewertet wurden zieht es im Diagramm nach unten links. Die Variable *Zeichnung* ist in beiden Hauptkomponenten positiv und gut bewertete Künstler dieser Kategorie zieht es in diese Richtung. Recherchen in der Kunsttheorie ergaben, dass unter den kontroversen Begriffen *Disegno* und *Colore* die Stile der klassischen Künstler eben in diese beiden Kategorien *Zeichnung* und *Farbe* einteilen lassen. In der nachfolgenden Darstellung sind Künstler mit einer Farbe- bzw. Zeichnung-Bewertung von

grösser als 15 abgebildet. Man kann eine Unterteilung der beiden Kategorien feststellen. Somit lässt sich festhalten, dass negative Werte der ersten beiden Hauptkomponenten eher auf *Colore*-Künstler und positive Werte eher auf *Disegno*-Künstler hinweisen



Aufgabe 2: k-means Verfahren

Man suche die optimale Anzahl an Clustern mit dem k-means Verfahren. Im Plot unten kann man keinen eindeutigen Knick erkennen. Jedoch wird mit wachsender Anzahl Cluster das Gütekriterium nicht wesentlich kleiner. Das Gütekriterium nimmt von $k = 1$ nach $k = 2$ weitaus mehr ab als von $k = 2$ nach $k = 3$. Man kann hier also die optimale Anzahl an Clustern von 2 wählen.




```
KMha20.km2 <- kmeans(KMha20, centers=2, nstart=1111)
```

Aufgabe 3: Cluster-Analyse

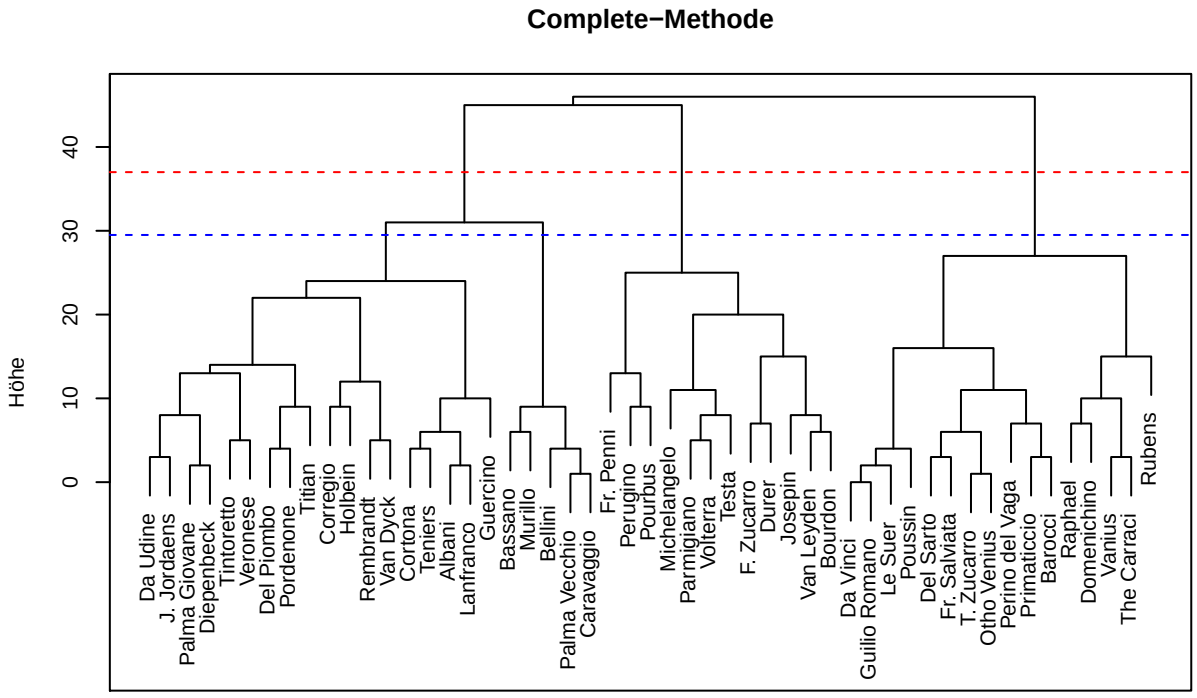
i) Wieso Manhatten-Distanzen

```
km.mh <- dist(KMha20, method="manhattan")
```

Bei den Daten handelt es sich um Bewertungen auf einer Skala von 0-20, sie weisen also auf eine Rangfolge hin. Wendet man die klassischen euklidischen Distanzen auf diese ordinalen Daten an, dann würden Informationen verloren gehen. Die Manhatten-Methode ist robuster und beachtet die Grundstruktur der Daten.

ii) Hierarchische Cluster-Analyse

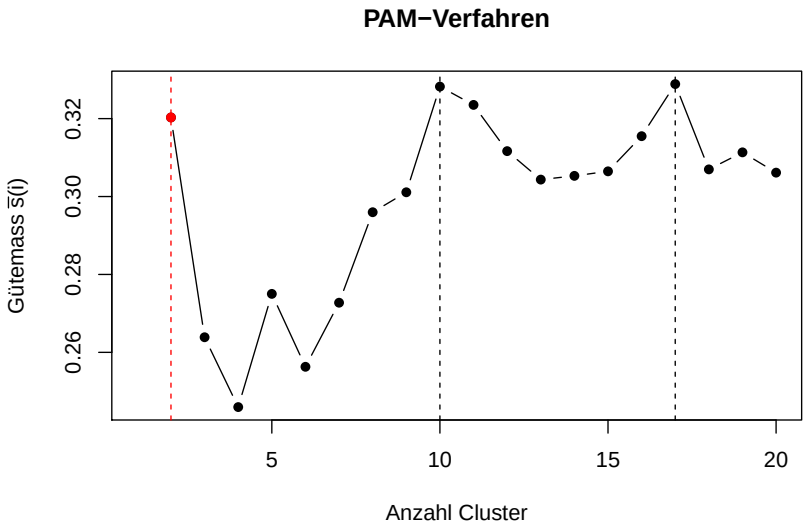
Die *Single*-Methode führt auf ein eher ungünstiges treppenstufiges Dendrogramm. Die *Average*- und die *Ward*-Methode liefern passable Ergebnisse, wobei es bei der *Average*-Methode einen Ausreisser hat (Fr. Penni). Im Vergleich der agglomerativen Koeffizienten ($0 < AC < 1$) weist die *Ward*-Methode mit einem Wert von $AC = 0.94$ den höchsten auf. Diese Methode liefert aber tendenziell nur scheinbar sinnvolle Lösungen. Die *Complete*-Methode weist einen AC von 0.89 auf und wird für die weitere Analyse verwendet. Die *Average*-Methode weist mit einem von $AC = 0.8$ zwar auch einen hohen Wert auf, könnte aber von dem Ausreisser beeinflusst sein. Die *Single*-Methode mit einem AC von 0.63 wird nicht in Erwägung gezogen.



Zwei sinnvolle Clusterbildungen sind im obigen Plot visualisiert. Beim durchtrennen bei der roten Linie würden 3 Cluster entstehen und beim druchtrennen bei blau würde man 4 Cluster erhalten. Mit der durchschnittlichen Silhouette width ($0 < \bar{s}_i < 1$) lässt sich die Güte der gebildeten Cluster vergleichen. Ein durchtrennen bei blau führt auf ein $\bar{s}_i$ von 0.28 welches leicht besser ist als ein durchtrennen bei rot, $\bar{s}_i = 0.26$. Erhöhung der Anzahl Cluster führt auf eine Verringerung des Gütemasses. Laut Kaufmann und Rousseeuw sollte jedoch Berücksichtigt werden, dass dieser tiefe Wert nur auf ein scheinbares Vorhandensein von Clustern hinweist.

```
km.com <- agnes(km.mh, method="complete")
km.com4 <- cutree(km.com, 4)
```

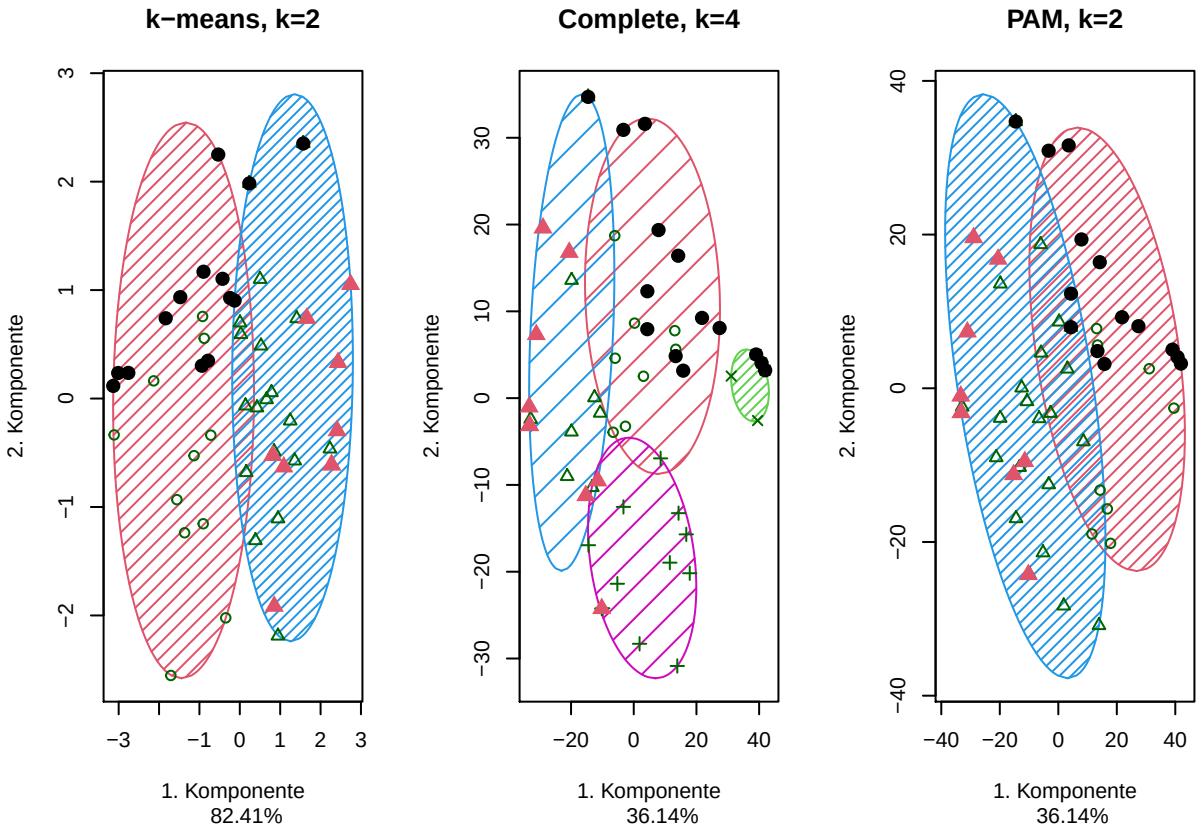
iii) PAM



Im obigen Plot sind auf der Abszisse die Anzahl Cluster aufgetragen und auf der Ordinate die dazu entsprechenden, mit der PAM-Verfahren berechneten $\bar{s}\langle i\rangle$-Werte. Für das Verfahren sind die mit der Manhatten-Methode berechneten Distanzen verwendet worden. Das Maximum liegt bei 17 Clustern, welches bei einem Datensatz mit 51 Beobachtungen nach einer etwas zu hohen Anzahl erscheint, auch das bilden von 10 Clustern (zweit höchster Wert) ist noch etwas zu grosszügig. Die weitere Analyse wird mit 2 Clustern fortgeführt, welches im Vergleich zu den anderen auch einen relativ hohen Wert aufweist. Man beachte das der $\bar{s}\langle i\rangle$-Wert von 0.32 bei 2 Clustern auf ein nur scheinbares Vorhandensein von Clustern hinweist.

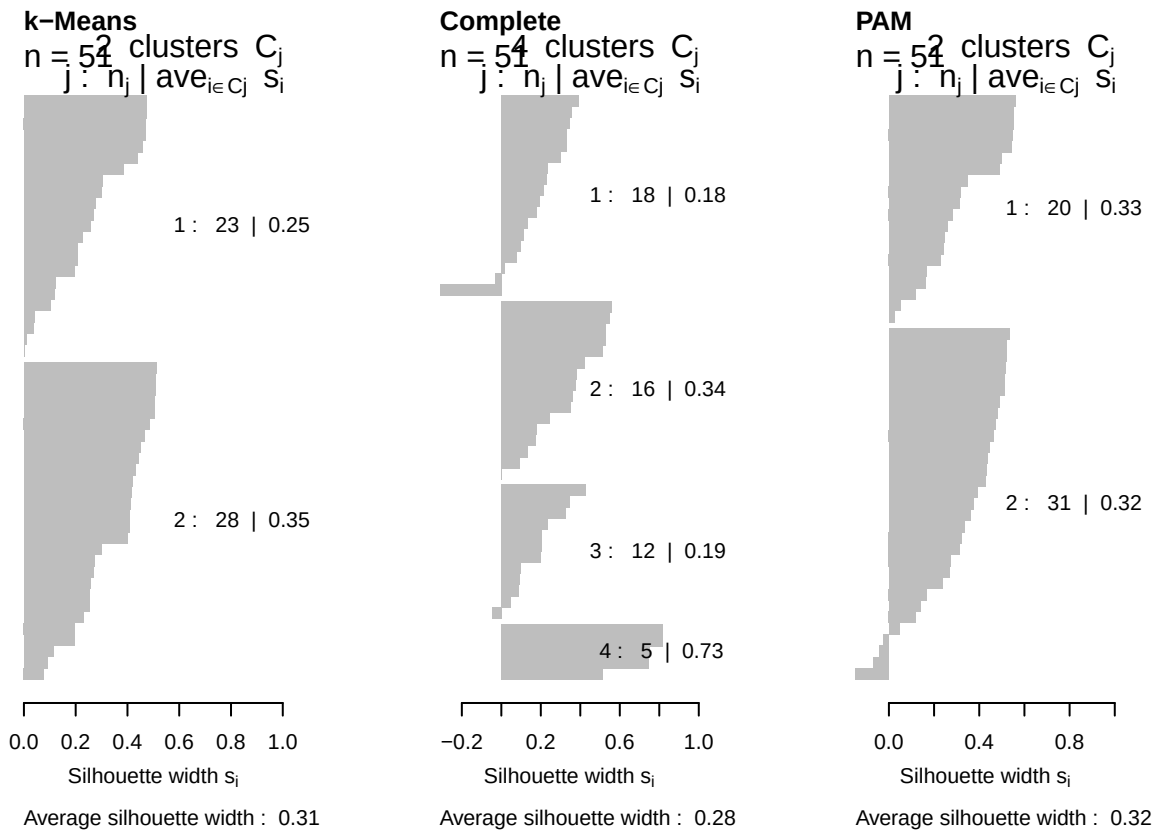
```
km.pam2 <- pam(km.mh, k=2, diss=TRUE)
```

iv) Clusplot, Silhouette-Diagramm



Die obige Darstellung zeigt die verschiedenen angwendeten Methoden mit der jeweiligen optimalen Anzahl von Clustern. Rein aus visueller Betrachtung scheint das k-means Verfahren (linker Plot) auf eine brauchbarere Lösung als die anderen beiden Verfahren zu führen. Die Überschneidung der 2 Cluster ist minimaler als beispielweise bei den 2 Clustern im PAM-Verfahren (rechter Plot). Das Complete-Verfahren (mittlere Plot) führt auf Überschneidungen von 3 der 4 Clustern. Dies weist darauf hin, dass nicht alle Künstler eindeutig in bestimmte Cluster eingeteilt werden können. Zusätzlich wurden die Plots mit der aus Aufgabe 1 erwähnten *Disegno* und *Colore* Unterscheidung ergänzt. Schwarze Punkte sind Künstler mit Farbe > 15 und rote Dreiecke mit Zeichnung > 15. Vorallem in den Verfahren k-means und PAM lässt sich diese Unterscheidung gut erkennen.

Bei allen 3 Verfahren ist die Silhouette width eher klein. Beim Complete- und PAM-Verfahren kann man Balken erkennen die nach links zeigen, diese weisen auf eine falsche Zuweisung der Künstler hin. Das k-means Verfahren führt zu den brauchbarsten Clustern. Im Vergleich zum PAM-Verfahren ist das Gütemass $\bar{s}\langle i\rangle$ nur um einen Hunderdstel kleiner, weist aber keine Künstler einem falschen Cluster zu.



v) Vergleich mit 2-dim. Darstellung aus Aufgabe 1

Der Biplot aus der Aufgabe 1 ist eine 2-dim. Darstellung der ersten beiden Komponenten aus der Hauptkomponenten-Analyse, bei welcher die unveränderten Daten verwendet wurden. In dieser Darstellung lassen sich über 80% der Variabilität der Daten ädaquat darstellen. Über 80% der ursprünglichen Strukturen in den Daten lassen sich auf 2 Dimensionen darstellen. In den Clusplots für die hierarchische Cluster-Analyse sowie für das PAM-Verfahren wurden zur Berechnung die Manhatten-Distanzen verwendet. In ihrer 2-dim. Darstellung lassen sich nur noch etwas mehr als 36% der ursprünglichen Variabilität der Daten darstellen, da die Daten "verändert" wurden. Dies bedeutet, dass ein grosser Teil der Informationen in einer höheren Dimension auftreten und hier nicht dargestellt werden können. Beim k-means Verfahren wurden die ursprünglichen Daten verwendet, weshalb im Clusplot wie auch in der Hauptkomponenten-Analyse ein hoher Erklärungsanteil der Daten steckt.

Aus der Erkenntnis der Kunsttheorie-Recherche macht eine Einteilung in 2 Clustern für *Disegno* und *Colore* am meisten Sinn.