



Statistik III: Multivariate Stat. Verfahren

Übungsblatt 7

Bearbeitung: Do. 1.7.2004, 16.00 Uhr.

Alle Aufgaben können im Mac-Cip Pool bearbeitet werden.

1. Korrigieren Sie im Datensatz **Eurodist** die Entfernung zwischen Rom und Athen auf 1420km.
 - (a) Berechnen Sie für diese Daten eine hierarchische Clusterung mittels average Linkage und Wards Methode, und plotten Sie diese.
 - (b) Vergleichen Sie die Ergebnisse untereinander, sowie zum Ergebnis aus der Vorlesung S. 197.
 - (c) Plotten Sie die Daten via MDS in 2 Dimensionen, und färben Sie die Städte nach der 4 Cluster Lösung des average Linkage ein.
2. Betrachtet werden die fünf Messgrößen des **Crabs** Datensatz.
 - (a) Berechnen Sie Clusterungen mittels hierarchischem Clustering mit den 4 in R implementierten Linkage Kriterien (single, complete, average und Ward).
 - (b) Gehen Sie wie in (a) vor, nur verwenden Sie nun nicht die Rohdaten, sondern deren Hauptkomponenten.
 - (c) Vergleichen Sie die erhaltenen 4 Cluster Lösungen. Welches Linkage Methode erscheint Ihnen am besten?
3. Wir betrachten den Fall des Model Based Clustering von multivariat normalverteilten Zufallsvariablen. Gegeben sei folgende Spektralzerlegung der Varianz Kovarianz Matrix

$$\Sigma_k = \delta_k \mathbf{P}_k \Lambda_k \mathbf{P}_k'$$

mit Proportionalitätskonstanten δ_k und zu den Eigenwerten proportionalen Einträgen in der Diagonalmatrix Λ_k .

Leiten Sie möglichst viele Einschränkungen der Σ_k aus dieser Darstellung ab, und beschreiben Sie diese verbal über die möglichen Formen der k anzupassenden Cluster.

4. Berechnen Sie mittels der Funktion `mclust` aus der `mclust` Library (muss ggf. noch auf Ihrem Computer installiert werden) eine Clusterung der **Crabs** Daten.
 - (a) Für welche Anzahl von Clustern, und welche Varianz Kovarianz Einschränkung ergibt sich der größte BIC Wert?
 - (b) Plotten Sie eine beliebige 2-d Projektion der Klassifikation. Welche Form nehmen die Cluster Ellipsen an?
 - (c) Welche Gruppen in den Daten werden durch die gefundenen Cluster approximiert?
5. Geben Sie für die Daten aus Aufgabe 4 der Funktion `mclust` zwei Cluster vor, und berechnen Sie die zugehörigen Cluster. Arbeiten Sie nun mit den zwei getrennten Clustern jeweils separat weiter, und passen Sie erneut je 2 Cluster über die Funktion `mclust` an.

Vergleichen Sie nun die 4 Gruppen im **Crabs** Datensatz mit den 4 gefundenen Clustern, und vergleichen Sie diese Cluster mit der 4 Cluster Lösung des gesamten Datensatz.