

Funktionale Datenanalyse

(Abgabe: Mo., 01.07.2013, vor 13:15)

1. Lade in R das Paket `fda`, und betrachte wieder den Datensatz `growth`. Wir möchten nun die Funktionen aus diesem Paket benutzen, um eine funktionale Hauptkomponentenanalyse durchzuführen.
 - (a) Berechne (wieder) die Mittelwertfunktion für die 93 Wachstumskurven von Jungs und Mädchen zusammen, und ziehe diese bei jeder Kurve ab. Speichere den neuen Datensatz als `wach`.
 - (b) Benutze den Befehl `base<-create.bspline.basis(...)`, um eine (Spline) Basis auf $[0, 18]$ mit 12 Funktionen zu erzeugen, und plote diese.
 - (c) Die meisten Funktionen aus `fda` arbeiten mit Objekten einer bestimmte Form (*functional data object*). Verwende die Funktion `Data2fd(...)`, um `wach` in die gewünschte Form zu bringen, und plote das Ergebnis (`wachfd`). Transformiere auch die Mittelwertfunktion. Plote nun die erste Wachstumskurve (man kann auf die erste Kurve in `wachfd` durch `wachfd[1]` zugreifen).
 - (d) Jetzt kann man `pca.fd(...)` anwenden, um die funktionale Hauptkomponentenanalyse durchzuführen. Lasse 10 Hauptfunktionen (*harmonics*) berechnen. Speichere das Objekt als `gfda` und plote es: Was wird dargestellt?
 - (e) Untersuche die Struktur von `gfda`, um folgendes zu bekommen:
 - Die *scores* $\xi_k^{(i)}$, $i = 1, \dots, 93$, $k = 1, \dots, 10$.
 - Der Anteil an 'Variation', die jede Hauptfunktion erklärt: Mit wie vielen Hauptfunktionen kann man 90% der Variation innerhalb der Daten erklären?
 - Die Hauptfunktionen: Speichere sie als `harmfd` (sie sind immer noch *functional data objects*).
 - (f) Wir versuchen jetzt, den Effekt jeder Hauptfunktion zu bestimmen. Hierfür kann man sie (oder ein Vielfaches davon) zu der Mittelwertfunktion addieren und subtrahieren, und die zwei Kurven, die man bekommt, in einem gemeinsamen Bild darstellen. Führe dies für die ersten 4 Hauptfunktionen durch.
 - (g) Plote nun die Schätzungen, die man für die Wachstumskurven bekommt, wenn man
 - Nur die erste Hauptfunktion betrachtet (und natürlich die Mittelwertfunktion)
 - Die ersten zwei Hauptfunktionen (und Mittelwertfunktion)
 - Die ersten 3 Hauptfunktionen (und Mittelwertfunktion)und vergleiche sie mit den ursprünglichen Daten.
 - (h) Plote $\xi^{(1)}$ (die Koeffizienten der 93 Kurven bezüglich der ersten Hauptfunktion) gegen $\xi^{(2)}$. *K-means* ist ein bekanntes Verfahren zur Clusteranalyse: Ein Datensatz wird auf einer 'optimalen' Art (bezüglich der euklidischen Norm) in k (k gegeben) Untergruppen geteilt.
Wir wissen, dass die Wachstumskurven aus 2 Gruppen (Mädchen und Jungs) kommen: Verwende die R-Funktion `kmeans(...)`, um die 2 Untergruppen zu bestimmen. Plote die Elemente aus der Gruppe 1 in einer farbe und die Elemente aus Gruppe 2 in einer anderen Farbe. Warum bekommt man nicht, was man erwartet hatte? Wie sollte man vorgehen?

(0,5+1+2+1+1+2+2,5+2 Punkte)