Prof. Dr. E. Spodarev
Judith Olszewski

SoSe 2015
18.06.2015

Statistik-Praktikum/WiMa-Praktikum II - Übungsblatt 8

Vorstellung der Ergebnisse in der Übung am 25.06.2015

Aufgabe 1

Erzeuge 100 Zufallsvektoren (x, y) mit Korrelationskoeffizienten $\varrho = 0.9$ (Tipp: $Y = \varrho \cdot X + \sqrt{1 - \varrho^2} \cdot Z$ mit $X \sim \mathrm{N}(0, 1), Z \sim \mathrm{N}(0, 1)$) und plotte diese Werte. Führe anschließend eine Hauptkomponentenanalyse mit der Prozedur PRINCOMP (Principal-Component) durch und erzeuge einen Plot der Daten nach Transformation der Hauptachsen. Analysiere und interpretiere das Ergebnis. Sorge in allen Plots für einen Titel und eine angemessene Achseneinteilung.

Aufgabe 2 Bei der Frage, wie viele Hauptkomponenten verwendet werden sollen, kann ein sog. Scree-Test (scree=Geröllhalde) helfen. Hierbei wird ein Scatterplot erzeugt, bei dem die Eigenwerte der empirischen Kovarianzmatrix der Daten gegen deren Indizes abgetragen werden. Verbindet man die Punkte durch Streckenzüge, entsteht häufig eine Form, die an den Fuß eines Berges erinnert. Diejenigen Eigenwerte, die in etwa horizontal liegen (in der Geröllhalde), führen dazu, die zugehörigen Hauptkomponenten aus den weiteren Überlegungen auszuschließen. Als Wert für die Anzahl der Hauptkomponenten wählt man diejenige Nummer, deren zugehöriger Eigenwert der letzte am Berg vor der Geröllhalde ist. (Mathematisches Kriterium: Betrachte nur Hauptkomponenten, deren zugehöriger EW größer als das arithmetische Mittel sämtlicher EW sind, d.h. $\lambda_k > \sum_{j=1}^{p} \lambda_j / p$).
Führe eine Hauptkomponentenanalyse mittels der Prozedur PRINCOMP für die Variablen der Datei 'econom.sas7bdat' durch, d.h. für die Variablen Arbeitslosenquote (ALQ), Zunahme Bruttoinlandsprodukt (BIP), Inflationsrate (INFLA), Investitionsquote (INVEST), Steuerquote (STEUER), Bevölkerung in Mio (POPUL), Arbeitskosten je Std. (ARBKOST), Anzahl Streiktage je 1000 Beschäftigte (STREIKTG) sowie Anzahl der in Betrieb befindlichen Atomkraftwerke (AKW), welche für 20 Industriestaaten gegeben sind. Entscheide mittels eines Scree-Tests, wie viele Hauptkomponenten betrachtet werden sollten. Analysiere und interpretiere das Ergebnis. Sorge in deinem Plot für einen Titel und eine angemessene Achseneinteilung.

Bitte zweite Seite beachten

Nutze folgenden Beispielcode als Hilfestellung:

```
TITLE1 'Hauptkomponentenanalyse für 2 Var und Scatterplot';
PROC PRINCOMP DATA=.... OUT=pca2 COV;
   VAR ...;
RUN;

SYMBOL1 ...;
AXIS1 ...;
PROC GPLOT DATA=pca2;
   PLOT PRIN2*PRIN1 / VREF=0 HREF=0 HAXIS=AXIS1 VAXIS=AXIS1;
RUN;
```

Die Beobachtungen werden hierbei in einem neuen Koordinatensystem, welches durch die Hauptkomponenten gebildet werden, dargestellt. Die zugehörigen Koordinaten stehen in den Variablen PRIN1 und PRIN2 in der Datei, die im PRINCOMP-Statement mit der OUT-Option erstellt wird (hier: pca2). Mit der Option COV in PRINCOMP werden die EW und EV der Kovarianzmatrix (anstelle der Korrelationsmatrix) berechnet.

```
TITLE1 'Scree-Test';
PROC PRINCOMP DATA=.... OUTSTAT=stat1 ;

DATA stat2(KEEP=ev pc);
   SET stat1(WHERE=(_TYPE_='EIGENVAL'));
   LABEL  pc='Hauptkomponenten' ev='Eigenwerte';
   ARRAY pr {AnzahlVar} varerst--varletzt;
   DO i=1 TO AnzahlVar;
      pc='HK'||LEFT(i);
      ev=pr{i};
   OUTPUT;  END;
RUN;

SYMBOL1 ...;
PROC GPLOT DATA=stat2;
   PLOT ev*pc / VREF=1;
RUN;
```

Hier enthält die Datei stat1 eine Beobachtung (Zeile), in der die Eigenwerte der Hauptkomponenten stehen. Diese Beobachtung enthält in der automatischen SAS-Variablen TYPE den Wert 'Eigenval' und wird mit der entsprechenden DATA-SET-Option ausgewählt. Der Rest des Data-Steps besteht darin eine geeignete Beschriftung der x-Achse mit HK1, HK2 usw. zu erstellen (je nachdem wie viele Variablen man hat).