PCA1

Zur Erläuterung und theoretischen Vertiefung der Methode der Hauptkomponentenanalyse (= PCA) sollte die Durchrechnung von konkreten Beispielen hilfreich sein.

In der Folge werden 2 Beispiele präsentiert, wovon das 2. Beispiel ein umfangreiches nach Anzahl der Merkmalsvariablen wie nach Zahl der zugrundeliegenden Arten und Individuen ist.

A) Beispiel-1 'BOHNEN' für PCA:

a) Das Datenmaterial stammt von einer Stichprobenziehung einer Bohnensorte und hatte den Umfang $n = 50$

[siehe Cbo.tls]

Aus Gründen allometrischer Zusammenhänge war zu vermuten, daß die zu vermessenden Merkmalsvariablen ohne Transformation nicht primär linear zueinander bezogen sind.

Aus diesem Grund wurden die abgenommenen Meßdaten d. Variablen logarithmiert (zur Basis e, log. naturalis).

Als Meßvariablen wurden ursprünglich LÄNGE, BREITE und GEWICHT abgenommen, mit den physikalischen Dimensionen LÄNGE [mm], BREITE [mm] und GEWICHT [g]. Nach der logarithmischen Transformation lagen die Variablen als LLG, LBR und LGEW vor.

Die logarithmische Transformation führt im Fall allometrischer Zusammenhänge zu einer LINEARISIERUNG der Daten zusammenhänge, welche GRUNDVORAUSSETZUNG für das ALM (= Allgemeine Lineare MODELL) sind.

b) Als Grundlage für die nachfolgende Durchführung einer PCA wurde zunächst eine Produkt-Moment-Korrelationsmatrix R zwischen d. 3 Merkmalsvariablen berechnet [siehe KORRMA.PDF].
Da die Rohvariablen in unterschiedlichen physikalischen Meßdimensionen, u. zw. [mm] und [g] erhoben wurden, war für die Auswahl der zugrundeliegenden Matrix d. PCA nur die Korrelationsmatrix aufgrund der damit erzielten Dimensionslosigkeit d. Parameter möglich [WICHTIG! Der Produkt-Moment-Korrelationskoeffizient führt zur NORMIERUNG d. Variablen und zu deren Dimensionslosigkeit bezogen auf die physikalische Dimension]

c) Das Protokoll für die VOLLSTÄNDIGE LÖSUNG der PCA findet sich unter RLBOPCA.pdf.
Nachstehende Punkte sind zu beachten:

1. Den Ausgang nimmt das Verfahren von der Korrelationsmatrix (3-dimensional, KORRMA.PDF), $n = 50$ (Individuen als OTUs bezeichnet).
2. Die Gesamtvarianz (= 'total variance') kann im Falle einer zugrundeliegenden Korrelationsmatrix leicht angegeben werden: tv ... total variance; t_v = Anzahl d. Merkmalsvariablen = Summe d. Hauptdiagonale in der Korrelationsmatrix
3. Auf die Gesamtvarianz bezieht sich d. gesamte Varianzbilanzierung d. PCA.
4. Das Auffinden d. 1. Hauptachse (Hauptkomponente, PC-1) bedeutet, den Hauptanteil d. Gesamtvarianz

PCA2

insofern ausfindig zu machen, als eine ACHSE zu finden ist, welche diesen Hauptanteil repräsentiert.
Man kann auch formulieren: Es ist eine Achse im Merkmalsraum zu finden, welche ein MAXIMUM an der GESAMTVARIANZ darstellt.

5 Die Varianz d. neuen Achse im Merkmalsraum wird als EIGENWERT (='eigenvalue') bezeichnet.

6 Die POSITION d. neuen Achse im Merkmalsraum wird durch den EIGENVEKTOR ('eigenvector') bestimmt.

7 Die sukzessive extrahierten Hauptachsen erklären in der Folge immer geringere Anteile an d. Gesamtvarianz
[VARIANZSTAFFELUNG $\lambda_1 \geq \lambda_2 \geq \lambda_3$ ...]

8 Wenn man anstatt vom normierten Eigenvektor auszugehen, den mit der Standardabweichung der Komponente [$\sqrt{\lambda_i}$] skalierten Eigenvektor zugrunde legt, so erhält man einen sog. FAKTOREN(EIGEN)-VEKTOR.
Die KOEFFIZIENTEN dieses Vektors werden als LADUNGEN bezeichnet.

9 Die Ladungen sind einer sehr einfachen Interpretation zugänglich. Es handelt sich um EINFACHE KORRELATIONSKOEFFIZIENTEN der jeweiligen ROHVARIABLEN (= Variablen, die d. Analyse zugrunde liegen) mit d. Haupt-komponente.

(10) Das Quadrat dieser FAKTORENLADUNGEN (= LDGQUA) entspricht d. einfachen BESTIMMTHEITSMASS und gibt damit Auskunft darüber, welche Anteile der Varianzen d. einzelnen ROHVARIABLEN in einer Komponente repräsentiert sind. (≙ AMVAR)

(11) Es läßt sich ebenso bilanzieren, welchen Anteil eine bestimmte Rohvariable an der Aufklärung der Gesamtvarianz an einer bestimmten Hauptkomponente hat [≙ MAGVAR]

d) Die Interpretation d. vorliegenden Beispiels soll folgende Punkte unbedingt umfassen:

(1) Die 1. Hauptachse wird durch einen Eigenvektor charakterisiert, dessen Koeffizienten allesamt Werte mit positivem Vorzeichen und sehr ähnlicher Mächtigkeit zeigen. Variieren auf dieser PC-1 bedeutet, daß die sie zusammensetzenden Rohvariablen auf dieser Achse GLEICHSINNIG variieren. Das bedeutet, wenn z.B. die Länge einer Bohne in Sample zunimmt, so handelt es sich um Bohnen, deren Breite ebenso wie deren Gewicht zunehmen. Umgekehrt, wenn z.B. die Länge einer Bohne in Sample abnimmt, so ist zu erwarten, daß es sich um eine Bohne handelt, deren Breite und deren Gewicht ebenso abnehmen. Die PC-I ist als sog. 'Größen-Variation-Achse' (= 'Size axis') zu interpretieren.

PCA 3

(2) Die 1. Hauptachse und ihre Interpretation haben aufgrund der Tatsache, dass diese 1. Hauptachse den „Löwenanteil" an der Gesamtvarianz, nämlich ~ 78%, repräsentiert, das größte GEWICHT (im Sinne von Bedeutung für die Interpretation) Hingegen haben alle folgenden Achsen viel weniger Gewicht.

(3) Die PC-2 wird durch einen Eigenvektor charakterisiert, welcher durch eine Vorzeichenkontrastierung der Koeffizienten für Länge und Breite auffällt. Der niedrige absolute Koeffizientenwert für das Gewicht erlaubt es diesen Koeffizienten außer Acht zu lassen, weil ihm an dieser Achse offenbar keine nennenswerte Bedeutung zukommt.
Auf dieser Achse variieren die Bohnen der Stichprobe gegengleich, auf LÄNGE und BREITE, bezogen:
D.h., findet man Bohnen im Sample mit größerer Länge, so zeichnen sich diese Bohnen auch durch geringere Breite aus (> 'schlanke Bohnen'!).
Umgekehrt gilt: Bohnen mit geringerer Länge sind im Sample eher breiter (> 'dickere Bohnen').

(4) Interessant ist die Koeffizientencharakteristik des 3. Eigenvektors. Hier werden LÄNGE und BREITE zum GEWICHT kontrastiert. D.h. Bei (wenigen) Bohnen d. Stichprobe findet man Individuen, die sehr groß erscheinen (LÄNGE u. BREITE), aber geringeres Gewicht aufweisen. Umgekehrt gibt es auch kleinere Bohnen, die relativ schwer sind.

e) Als letzter Teil d. Analysen in RLBOPCA.pdf findet sich unter 'Ordination der Daten aus der taxonometrischen Matrix' die Repräsentation der Koordinatenwerte für die Daten der 50 Individuen auf den neuen Achsen (= 3 Hauptachsen).