Hauptkomponentenanalyse, PCA

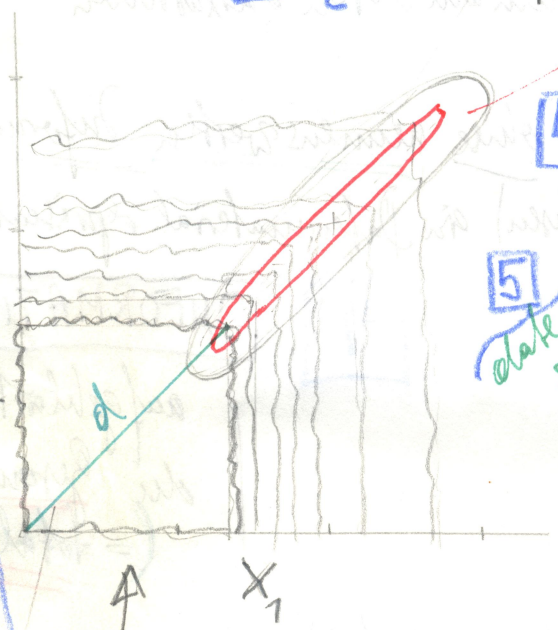
(principal components analysis)

(A) Zweck: Datenreduktion > einfachere Darstellung d. ursprünglich ermittelten Daten ohne nennenswerten Verlust ~~Daten~~ d. ursprünglichen Informationen

Variablen x_1, x_2

3 Centroid $(\bar{x}_1 | \bar{x}_2)$

2 Darstellung im bivariaten Koordinatensystem



4 falls es sich um eine sehr flache Punktwolke handelt,

> anstelle einer Vermessung von x_1 und x_2 könnte man ~~um eine größere~~ die Diagonale d vermessen:

$$d = \sqrt{x_1^2 + x_2^2}$$

5 datenreduzierender Effekt

1 gegeben ist eine Stichprobe zum Beispiel von (quadratische Briefmarken als Beispiel ähnlich Fliesen, Dachziegel, Schildkrötenpanzer, Zeiten, Blätter)

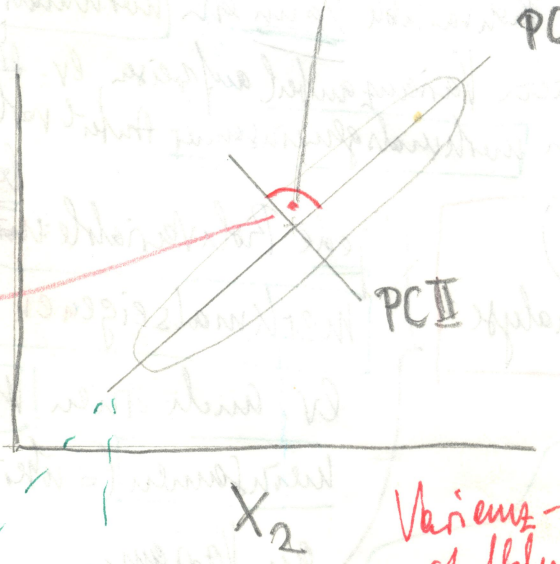
6 stellt eine LINEARKOMBINATION dar (linear compound, linear combination)

(B) Vorgehensweise der PCA:

PC... principal component = Hauptachse

ORTHOGONALITÄT

Die PCI und PCII stehen aufeinander normal



Varianz d. Komponente: eigenvalue | Richtung wird von eigenvector angegeben

Varianz-Staffelung

(A) Es wird ein neues Koordinatensystem nach folgenden Kriterien gesucht:

(1) die 1. Achse soll ein Maximum an Varianz d. Punktwolke „aufklären" [Varianz d. PCI ist Eigenwert 1]

(2) die folgende(n) Achse erklärt von der verbleibenden Varianz wieder ein Maximum ...

(3) alle neuen Achsen stehen

aufeinander normal (> ORTHOGONAL), d.h. die neuen Achse = neue Variablen, sind nicht miteinander korreliert. ≙ UNKORRELIERTHEIT

8

(4) Wenn die Rohdaten miteinander korreliert waren, so ist mit der PCA ein datenreduzierender Effekt zu erzielen, d.h. anstatt, daß die ursprüngl. z.B. 10 Rohvariablen zur Datenrepräsentation vorliegen können ev. 2 Hauptachsen (PCs) ausreichen, um die Daten darzustellen

der datenred. Effekt liegt darin, daß ohne nennenswerten Informationsverlust mit geringerem Aufwand (Achsen) ein Datenmaterial repräsentiert werden kann

9

INFORMATION VON:

aufgeklärt wird die Gesamtvarianz (= total variance)

(C) Ausgang der PCA von:

10

S (Var-Cov.-Matrix) oder R (Korrelationsmatrix)

11

der unterschiedliche Ausgang bedingt, daß die Ergebnisse d. PCA in der Regel unterschiedlich sind

(D) Die PCA ist auch Grundlage für Faktorenanalysen (factor analysis)

Konzeptionelle Differenzen zwischen

PCA: eine Rohvariable kann ein merkmalseigenen und einen Varianzanteil aufweisen, ev. ist aber nur ein merkmalsgemeinsamer Anteil vorhanden

(ortho)gonale Faktorenanalyse: jede Rohvariable besitzt einen merkmalseigenen Anteil und ev. auch einen merkmalsgemeinsamen (= überlappenden) Anteil an Varianz

merkmalseig. Anteil: ϵ_i ... der Variable i (Error)

≙ exklusiv, auspartialisiert ≙ unkorreliert mit allen anderen „errors"

(E) Beispiel aus der Morphometrie:

z.B. >40 Skelettmerkmale, transspezifisch erhoben, logarithmisch transformiert, lassen sich auf 2 Hauptachsen reduzieren > Aufklärungsgrad ~98% der Totalvarianz

1. Achse: sog. „Größenachse" ('Size axis')

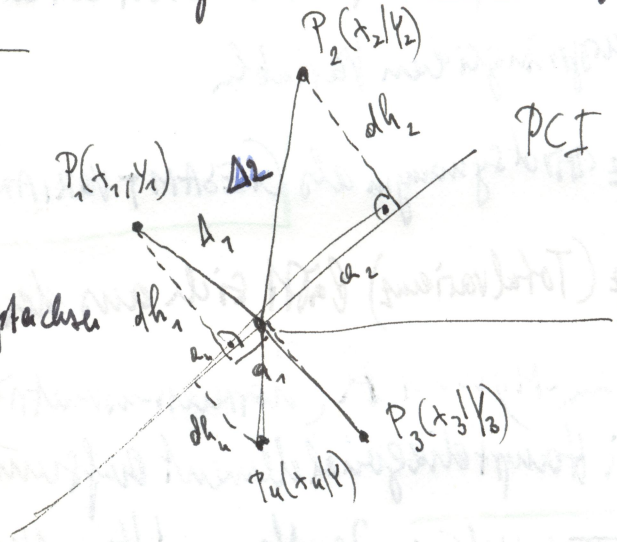
2. Achse: sog. „Proportionsachse" ('Shape axis')

Nachtrag zu (B): Nachtrag > $\boxed{\text{orthogonale Regression}}$ als alternativer Ansatz für die Hauptkomponentenanalyse

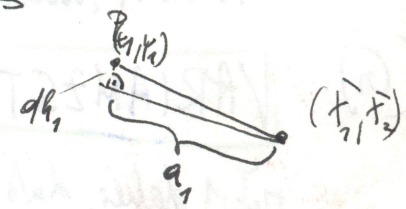
Koordinaten:

① ursprüngl. Koordinaten $P(x_i|y_i)$

② Koordinaten auf d. Hauptachsen $P(a_i|dh_i)$



→ minimaleres dh_1:



a) es gilt d. pythagor. Lehrsatz in d. Ebene:

$$\Delta_1^2 = dh_1^2 + a_1^2$$

b) wenn gilt: $\sum_{i=1}^{n} dh_i^2 \rightarrow \min$

dann gilt: $dh_i^2 = \Delta_i^2 - a_i^2$

der Δ_i^2 als Centroidentfernung eines Punktes konstant ist > wird a_i^2 maximiert

→ folglich die Varianz auf PCI maximal!

(F) a) Die Varianzen d. Hauptachsen werden EIGENWERTE (eigenvalues) genannt Symbol λ ... lambda d.h. $\lambda_1, \lambda_2, \ldots \lambda_m$

b) die Richtungen der Hauptachsen werden durch die EIGENVEKTOREN (eigenvectors) angegeben

c) Die Summe aller Eigenwerte eines Systems entspricht der TOTALVARIANZ (total variance) d. Systems

d) Die TOTALVARIANZ entspricht zum anderen der Summe der Varianzwerte d. ursprünglichen Variablen

e) Die Totalvarianz wird synonym als GESAMTVARIANZ bezeichnet.

f) Die Gesamtvarianz (Totalvarianz) läßt sich aus den Matrizen S (Varianz-Kovarianz-M.) und R (Korrelationsmatrix) leicht ermitteln, indem man die Hauptdiagonalelement aufsummiert

(G) VARIANZSTAFFELUNG: Der Name 'Hauptkomponenten-Analyse' rührt daher, daß im Zuge d. Verfahrensverlaufes beginnend mit den wichtigsten Achsen (d.h. Achsen, welche die größte ursprüngliche Varianz darstellen) gefolgt von den weniger wichtigen Achsen (mit geringerer Varianzaufklärung) die Achsenextraktion durchgeführt wird bis bei einer vollständigen Extraktion die am wenigsten wichtige Achse am Ende kommt.
Der Extraktionsfortschritt ist gestaffelt.
Eine VOLLSTÄNDIGE EXTRAKTION ergibt VARIANZSTAFFELUNG, die Anzahl der neuen Achsen (Hauptachsen) stimmt aber zahlenmäßig mit der Anzahl d. ursprünglichen Achsen (= ROHACHSEN) überein.

Extraktionsfolge: PC I → PC II → PC III ... PC_m bei m ROHACHSEN

VARIANZSTAFFELUNG: $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \lambda_m$