Hauptkomponentenanalyse

Voraussetzung und verwandte Themen

Für diese Beschreibungen sind Grundlagen der Statistik vorteilhaft. Weiterführende und verwandte Themen sind:

www.versuchsmethoden.de/PLS.pdf

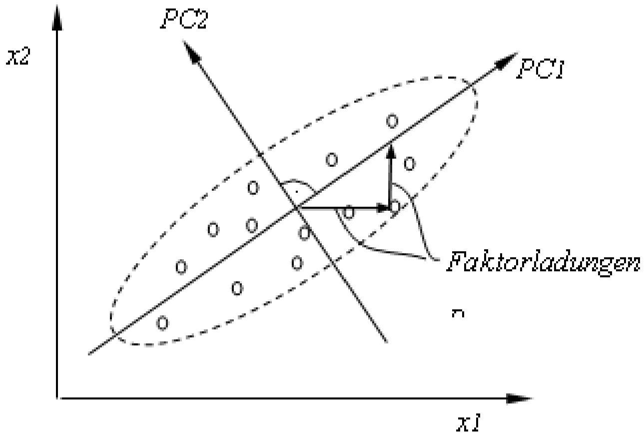
Einführung

Die Hauptkomponentenanalyse berechnet aus ursprünglichen Daten (Variablen x) neue sogenannte latente Variablen, die man kurz als Faktoren bezeichnet und die die **Hauptkomponenten** PC darstellen.

Ziel und Nutzen

Das Ziel ist eine Datenreduktion, wobei mit wenigen Faktoren alle Ausgangsvariablen zu beschrieben werden sollen. Dabei müssen die Ausgangsvariablen genügend korrelieren.

Grundlagen



Anhand von 2 Ausgangsvariablen x_1 und x_2 und deren Messpunkte, soll das Prinzip wie rechts im Bild dargestellt werden. Die Messpunkte liegen in einer Ellipse, die in Ihrer Form und Neigung von der Korrelation zwischen den Variablen abhängen. Durch Verschiebung des Nullpunktes und Drehung des Koordinatensystems entsteht ein neues Achsensystem. Die erste so genannte Hauptachse weist in Richtung der größten Streuung der standardisierten Werte von x_1 und x_2. Die zweite Hauptachse steht senkrecht auf der ersten, wobei hier der nächst geringere Varianzanteil erklärt wird. Die Hauptkomponenten bezeichnet man deshalb auch als Eigenvektoren.

Zur Bestimmung von Hauptkomponenten werden so genannte Faktorladungen P (Loadings) und Score-Werte T gebildet. Die Faktorladungen ergeben auf dem ursprünglichen Koordinatensystem von x_1 und x_2 die Lage der PC. Die Dimension der Faktorladungen ist Anzahl Komponenten x Anzahl Variable x. Die Score-Werte T beschreiben die Projektionen auf die Hauptachsen für jeden Punkt. Die Dimension von T ist Anzahl Komponente x Anzahl Messwerte. In der Matrixschreibweise ist der Zusammenhang:

$$X = T\,P^T$$

Für die Faktorladungen gilt die Bedingung:

$$p_1^2 + p_2^2 + \ldots p_k^2 = 1$$

Die Hauptkomponenten werden über die Score-Werte t_i und die so genannten Eigenwerte λ_i berechnet:

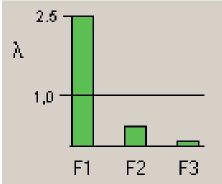
$$PC_i = t_i / \sqrt{\lambda_i}$$

Der Eigenwert λ_i gibt an, wie viel von der Gesamtvarianz aller Variablen durch diesen Faktor erfasst wird. Die Eigenwerte dienen auch zur Entscheidung, ob Faktoren im Modell beibehalten oder weggelassen werden können. Ist der Eigenwert kleiner oder gleich 1, erklärt er weniger oder nur gleich viel wie die Varianz einer einzigen Variablen. Damit kann der Faktor weggelassen werden. Eigenwerte und Eigenvektoren ergeben zusammen eine voneinander unabhängige (orthogonale) Struktur. Die Eigenwerte lassen sich nicht analytisch berechnen, sondern müssen iterativ bestimmt werden (Eigenwertproblem). Für weitere Details sei auf die einschlägige Literatur verwiesen (insbesondere *Multivariate Datenanalyse, Kessler*).

Beispiel: Gegeben sind die Variablen x_1, x_2 und x_3. Berechnet wurde der Faktor F:

x_1	x_2	x_3	F
1	3	4	-1,00
2	4	3	-0,70
3	1	1	1,00
4	2	2	0,70

Für diese Daten reicht ein Faktor aus (λ für den zweiten und dritten Faktor ist unter 1). Neben den Faktorladungen gibt es noch die so genannten Korrelations-Ladungen. Das sind die Korrelationen zwischen den Faktoren und den ursprünglichen Variablen.



Betrachtet man die Korrelationen zueinander, so zeigt sich, dass der neue Faktor mit allen Ausgangsvariablen deutlich höher korreliert, als die Variablen untereinander. Ziel des Faktors ist es ja gerade eine möglichst gute „Beschreibung“ aller Variablen gemeinsam zu erreichen.

x_3: F	-0,958
x_2: F	-0,881
x_1: F	0,881
x_1: x_3	-0,800
x_2: x_3	0,800
x_1: x_2	0,600

Korrelationen zwischen Faktor und Ausgangsvariablen

Korrelations-Ladungen

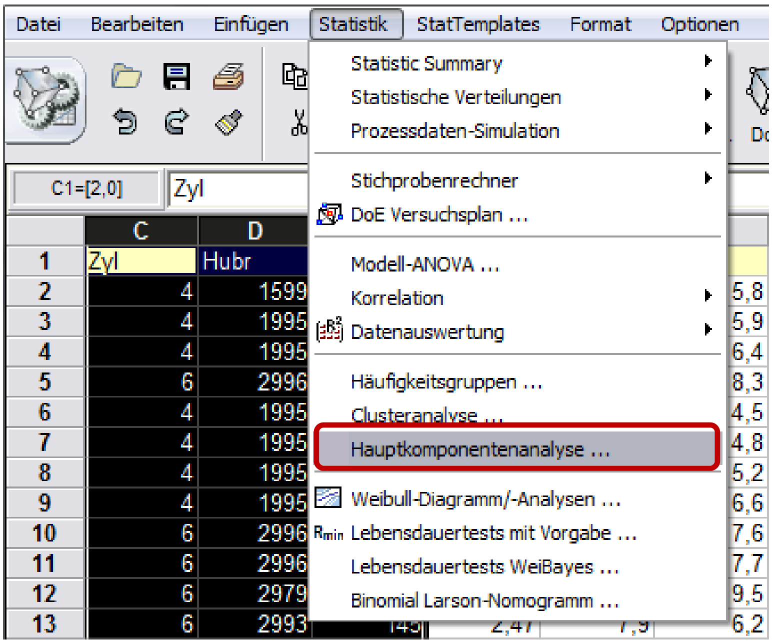
Zu beachten ist hier für die Interpretation, dass der Faktor mit Variable x_2 und x_3 negativ und mit Variable x_1 positiv korreliert. Es bedeutet eine negative Korrelation, dass für die Veränderung der Faktorwerte in einer Untersuchung von kleinen Werten auf große Werte die Richtung für Variable x_3 und x_2 umgekehrt ist.

Eine Erweiterung der Hauptkomponentenanalyse ist das PLS-Verfahren.

Hauptkomponentenanalyse

Anwendung in Visual-XSel 14.0

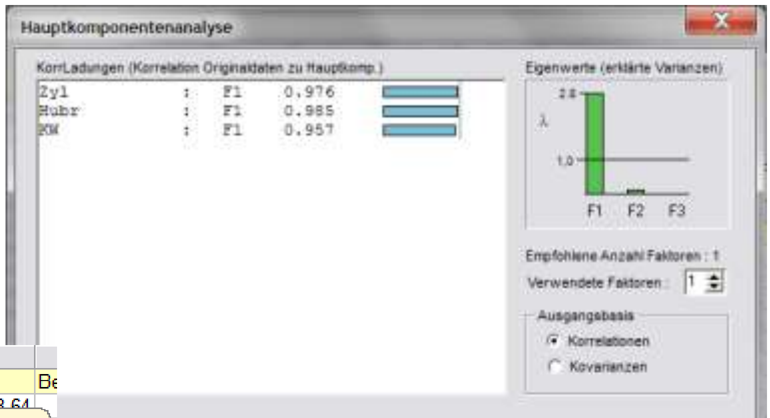
www.crgraph.de



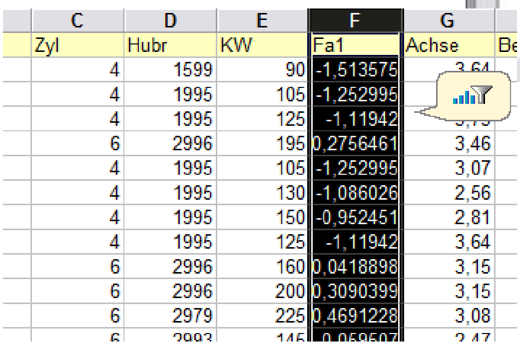
Verwenden Sie den Menüpunkt ***Statistik / Hauptkomponentenanalyse***, wobei die gewünschten Variablen (Datenspalten) vorher zu markieren sind. Diese müssen direkt nebeneinander stehen.

Für die Daten im *Beispiel_Verbrauch.vxt* unter *Datei / Beispieldaten* sollen die Datenreihen Zyl, Hubr und KW zu einem Faktor zusammengefasst werden, denn diese Eigenschaften hängen stark miteinander zusammen.

In der rechten Balkengrafik ist zu sehen, dass sich die drei Spalten mit einer Komponente darstellen lassen, denn nur F1 liegt hinsichtlich der Eigenwerte über 1.0.



Rechts neben den markierten Spalten wird die ermittelte Faktorspalte eingefügt.



C	D	E	F	G
Zyl	Hubr	KW	Fa1	Achse
4	1599	90	-1,513575	3,64
4	1995	105	-1,252995	
4	1995	125	-1,11942	
6	2996	195	0,2756461	3,46
4	1995	105	-1,252995	3,07
4	1995	130	-1,086026	2,56
4	1995	150	-0,952451	2,81
4	1995	125	-1,11942	3,64
6	2996	160	0,0418898	3,15
6	2996	200	0,3090399	3,15
6	2979	225	0,4691228	3,08
6	2993	145	0,059507	2,47

Diese neue Datenspalte kann nun für weitere Analysen anstelle der 3 Ausgangsdaten verwendet werden. Die Daten sind damit ideal reduziert worden.

Der Nachteil ist jedoch, dass man es dann mit einer abstrakten Bezeichnung zu tun hat. Man könnte diesen Faktor aber als leistungsstarke Fahrzeuge Interpretieren, um damit einen Zusammenhang zum Verbrauch herzustellen. Eine multiple Regression zwischen Zyl, Hubr, KW und dem Verbrauch ist hier nicht möglich, da die Korrelation der Variablen zu groß ist. Dafür war diese Korrelation aber die Voraussetzung für die Bildung eines neuen Faktors.

Das PLS – Verfahren vereint beide Verfahren.

Literatur
Taschenbuch der statistischen Qualitäts- und Zuverlässigkeitsmethoden

Die wichtigsten Methoden und Verfahren für die Praxis.

Beinhaltet statistische Methoden für Versuchsplanung & Datenanalyse, sowie Zuverlässigkeit & Weibull.

- Statistische Verteilungen und Tests & Mischverteilungen
- Six Sigma Einführung und Zyklen
- Systemanalysen Wirkdiagramm, FMEA, FTA, Matrizen-Methoden
- Shainin- und Taguchi-Methoden
- Versuchsplanung DoE, D-Optimal
- Korrelations- und Regressionsverfahren
- Multivariate Datenauswertungen
- Prozessfähigkeit – Messmittelfähigkeit MSA 4 und VDA 5
- Regelkarten
- Toleranzrechnung und Monte-Carlo-Simulation
- Statistische Hypothesentests
- Weibull und Lebensdaueranalysen
- Stichprobengröße

190 Seiten, Ringbuch

ISBN: 978-3-00-043678-9