.

02. 06. 2013

U. Mortensen

Was Sie wissen sollten

Von den folgenden Verfahren sollten Sie wissen, (i) wozu man sie braucht, (ii) was sie leisten, (iii) wo ihre Grenzen liegen.

1. Multiple Regression
2. PCA als Approximation für die Faktorenanalyse
3. Diskriminanzanalyse
4. Kanonische Korrelation

Multiple Regression: Sie spielt z.B. in der Diagnostik eine große Rolle. Gegeben sind messbare Variablen $X_1, \ldots, X_p$, die als Symptome oder Prädiktoren gelten oder von denen man vermutet, dass sie als solche gelten könnten. Y ist eine Weitere Variable, die man "voraussagen" will (zB Eignung, Depression). Der Ansatz ist

$$Y = b_0 + b_1 X_1 + \cdots + b_p X_p + e \tag{1}$$

Geht man von zentrierten Werte aus (der jeweilige Mittelwert wird von den Variablen abgezogen, schreibt man i. A. kleine Buchstaben:

$$y = b_1 x_1 + \cdots + b_p x_p + e \tag{2}$$

wobei e der Messfehler ist; er nimmt in (1) und (2) verschiedene Werte an. Es wird angenommen, dass die Fehler im Mittel gleich Null sind und sie nicht korreliert sind (vernünftige Annahme, wenn die Personen, bei denen die Daten erhoben werden, unabhängig voneinander ihre Antworten geben).

Die Regressionsgewichte werden mit der Methode der Kleinsten Quadrate (KQ) geschätzt. In Matrixschreibweise erhält man für den Vektor $\hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_p)'$ (es wird hier nur der zentrierte Fall betrachtet, – wenn der nicht vorliegt, erhält man noch die Schätzung $\hat{b}_0$) die Gleichung

$$\hat{\mathbf{b}} = (X'X)^{-1} X'\mathbf{y} \tag{3}$$

Sind die Messwerte standardisiert, so ist $X = Z$ und $X'X$ ist – bis auf den Faktor $1/m$, m die Anzahl der Fälle - gleichg der Matrix der Korrelationen zwischen den

Prädiktoren. Die Schätzung des Vektors $\mathbf{b}$ ist stichprobenabhängig, also fehlerbehaftet, d.h. die Schätzungen für die Komponenten b_j haben eine Varianz und sind u.U. korreliert. Es gilt

$$\mathbb{E}(\hat{\mathbf{b}} = \mathbf{b} \quad (4)$$

$$Kov(\hat{\mathbf{b}}) = \sigma^2(X'X)^{-1} \quad (5)$$

(4) besagt, dass die Schätzung für $\mathbf{b}$ keinen *Bias* hat, d.h. sie ist nicht mit einem *systematischen* Fehler behaftet, $\mathbf{b}$ wird also nicht systematisch unter- oder überschätzt. (5) besagt, dass der Stichprobenfehler, dh hier die Varianz der Schätzungen, proportional zur Varianz σ^2 der Fehler e ist. Weiter geht noch die inverse Matrix $X'X$ (eventuelle $= R^{-1}$) noch ein. Gelten die obigen Annahmen über die Fehler, läßt sich zeigen, dass die KQ-Schätzung $\hat{\mathbf{b}}$ die beste Schätzung ist, die man überhaupt bekommen kann (Gauß-Markov-Theorem), sie hat die kleinste Varianz relativ zu den Schätzungen, die anhand anderer Schätzmethoden erreicht werden können.

Das heißt *nicht*, dass die Schätzungen wirklich das leisten, was sie leisten sollen (zB die Vorhersage von Y für einen neuen Fall (= Patienten)). Denn wenn die Prädiktoren korreliert sind, ist $(X'X)^{-1}$ nicht gleich der Einheitsmatrix wie im unkorrelierten Fall. Sobald einige Prädiktoren relativ hoch korreliert sind, geht (i) die Varianz der Schätzungen hoch, und (ii) sind sie negativ korreliert – nimmt b_j einen großen positiven Wert an, so nimmt b_{j+1} einen entsprechend hohen negativen Wert an. Dies bedeutet, dass man für die *gegebenen* Daten oft einen guten Fit erhält (Y- und $\hat{Y}$-Werte weichen nicht sehr voneinander ab), aber für *neue* Prädiktorwerte erhält man große Abweichungen des vorhergesagten Wertes $\hat{Y}$ vom tatsächlichen Wert.

Ursache: Für $X'X$ existiert stets die Zerlegung in Eigenvektoren $X'X = P\Lambda P'$, P enthält die Eigenvektoren, Λ die Eigenwerte. P ist orthonormal (Spaltenvektoren sind auf die Länge 1 normiert und orthogonal zueinander). Dann hat man

$$(X'X)^{-1} = P\Lambda^{-1}P' = \sum_{j=1}^{n} \frac{\mathbf{P}_j\mathbf{P}'_j}{\lambda_j} \quad (6)$$

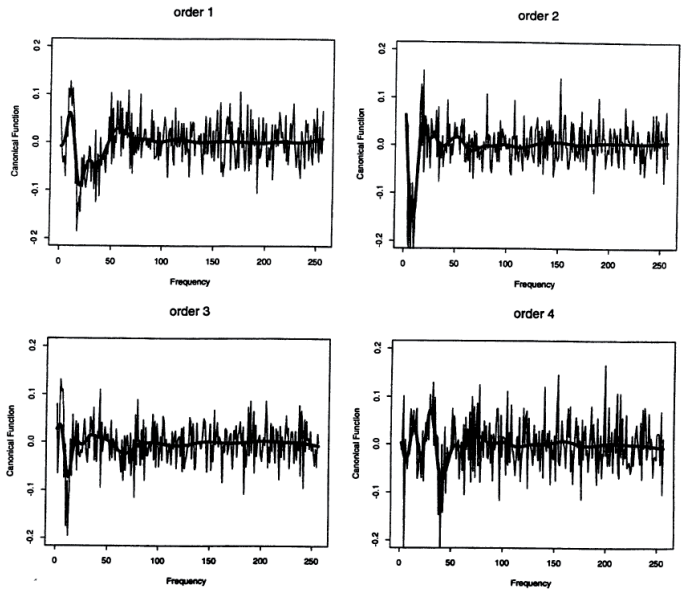
Große Korrelationen bedeuten kleine Eigenwerte, und da diese in den Nenner in (6) eingehen, blähen sie die Werte in $(X'X)^{-1}$ auf, mit der genannten Folge für die Schätzungen $\hat{\mathbf{b}}$.

Es gibt in der Literatur viele Vorschläge, wie man das Problem lösen kann:

1. **Stepweise Regression**, wobei man Prädiktoren eliminiert, ohne dabei den multiplen Regressionskoeffizienten allzu sehr zu reduzieren. Das Verfahren ist in einigen Statistikpaketen implementiert, verfährt gleichwohl ein wenig willkürlich und ist nicht notwendig optimal.

2. **PCA-Regression**, – hier wird die Matrix X der Prädiktorwerte einer PCA unterzogen. Man erhält p orthogonale latente Vektoren, die statt der ge-

Abbildung 1: Vorhersage von Phonemen anhand der Frequenzkomponenten als Prädiktoren: Effekt der Schrumpfung. Die durchgezogene Linie repräsentiert die korrigierten Schätzungen



messenen Vektoren als Prädiktoren verwendet werden. Dabei versucht man oft, die latenten Vektoren mit "kleinen" Eigenwerten zu vernachlässigen, um die Vorhersage so sparsam als möglich zu machen. Problem: die PCA-Regression funktioniert nicht immer, insbesondere, wenn nur ein reduzierte Anzahl latenter Vektoren berücksichtigt wird (die Vernachlässigung von latenten Variablen mit kleinen Eigenwerten kann suboptimal sein, es ist aber nicht klar, welche Auswahl man als Alternative wählen kann).

3. **Ridge-Regression**, – hier wir auf die Diagnoalelemente von $X'X$ ein Konstante h oder λ (die Bezeichnungen variieren) addiert. Der Effekt ist, dass die Komponenten von $\hat{\mathbf{b}}$ nun "geschrumpft" werden, – jo größer der Wert von h, desto mehr wird geschrumpft. Der optimale Wert von h wird durch Kreuzvalidierung geschätzt.

Zur Ridge-Regression Man geht also von $X'X + hI$ aus, I die Einheitsmatrix. Die Schätzung für $\mathbf{b}$ ist nun

$$\hat{\mathbf{b}}_h = (X'X + hI)^{-1}X'\mathbf{y} = \left(\sum_{j=1}^{n} \frac{\mathbf{P}_j\mathbf{P}'_j}{\lambda_j + h}\right) X'\mathbf{y}. \tag{7}$$

Man spricht auch von *regularisierten* Schätzungen.

Abb. 1 beschreibt das Resultat (Vorhersage von Lauten anhand von Frequenzkomponenten): die ursprünglichen Schätzungen haben eine große Varianz und oszillieren. Die durchgezogene Linie zeigt die regularisierten Schätzungen; sie zeigen, welche Komponenten die wirkliche Informatui zur Vorhersage enthalten.

PCA als Approximation der FA Das faktorenanalytische Modell wird überlicherweise in der Form

$$z_{ij} = a_{j1}F_{i1} + \cdots + a_{jr}F_{ir} + e_{ij} \tag{8}$$

vorgestellt: die F_{ik}, $k = 1, \ldots, r$ sind die Faktorenscores, dh die Ausprägungen der latenten Variablen (Faktoren) bei der i-ten Person, und die a_{jk} sind die Faktorladungen der Variablen ("Tests").

Die PCA geht davon aus, dass die Spaltenvektoren der Datenmatrix X als Linearkombinationen von orthogonalen Vektoren, die latente Variable repräsentieren, dargestellt werden können; diese Annahme führt zum Ansatz

$$X = LP', \tag{9}$$

wobei die Matrix L die latenten Vektoren enthält und die Spalten von P' die für die Repräsentation von $\mathbf{x}_j$ benötigten "Gewichte". Es folgt

$$X'X = PL'LP' = P\Lambda P', \tag{10}$$

dh die Matrix P ist die Matrix der orthonormalen Eigenvektoren von $X'X$ (u. U. $X'X = R$ die Matrix der Korrelationen). Normiert man L, dh bildet man die Matrix $Q = L\Lambda^{-1/2}$, so ist $L = Q\Lambda^{1/2}$ und in (9) eingesetzt ergibt sich die *Singularwertzerlegung*

$$X = Q\Lambda^{1/2}P'. \tag{11}$$

L bzw. Q enthält die Ausprägungen der Personen auf den latenten Dimensionen, entspricht also den Faktorscores. P repräsentiert die Ausprägungen der Variablen auf den latenten Dimensionen. Man findet sofort $XX' = Q\Lambda Q'$, dh die Matrix Q enthält die Eigenvektoren von XX'.

Setzt man

$$A = P\Lambda^{1/2} \tag{12}$$

so enthält A gerade die *Ladungen* der Variablen, dh das Element a_{jk} von A ist die Ladung der j-ten Variablen (des j-ten Tests) auf der k-ten latenten Dimension. Ist $X = Z$, dh enthält X standardisierte Werte, so findet man

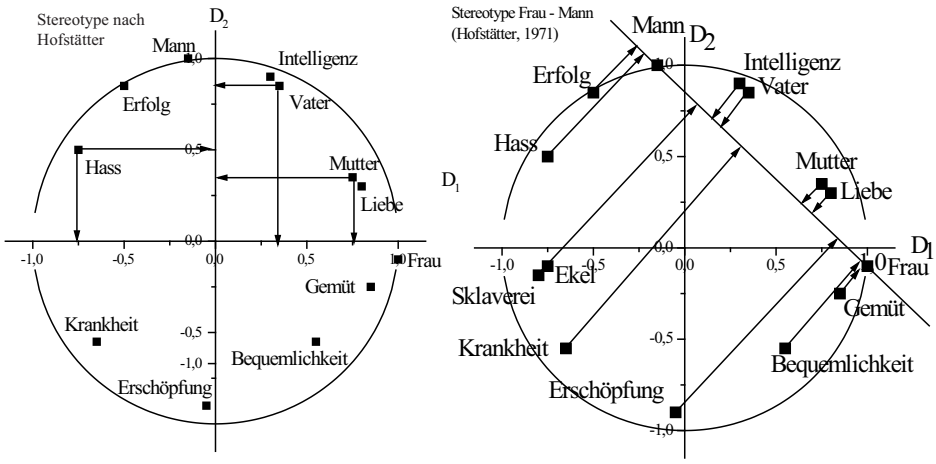
$$R = \frac{1}{m}Z'Z = \frac{1}{m}AA'. \tag{13}$$

Speziell erhält man für die Variablen j, k

$$r_{jk} = \begin{cases} \sum_{s=1}^{n} a_{js}a_{ks} = \cos\theta_{jk}, & j \neq k \\ \sum_{s=1}^{n} a_{js}^2 = 1, & j = k \end{cases} \tag{14}$$

θ_{jk} der Winkel zwischen den Variablenvektoren. Der Fall $j = k$ zeigt, dass die Vektoren, die die Variablen repräsentieren, alle die Länge 1 haben, mithin liegen die Endpunkte auf einer n-dimensionalen Hyperkugel. Werden statt der n latenten Varibalen nur 2 benötigt, so liegen die Endpunkte auf, bzw nahe dem Einheitskreis mit dem Radius 1, bei 3 latenten Variablen auf einer Kugel. Hofstätters (1959/1971) Befund aus der Stereotypenforschung zeigt, dass eine 2-dimensionale

Abbildung 2: 2-dimensionale Stereotypstruktur (nach Hofstätter 1959/71). Die Gerade im rechten Bild zeigt die männlich-weiblich Skala nach der Polaritätstheorie (Antike, Scholastiker, Goethe, Wellek)

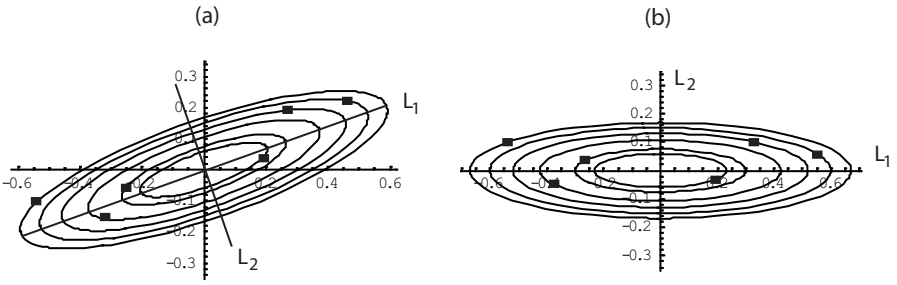


Struktur (nach der Methode des Polaritätsprofils) ausreicht. Die von Hermeneutikern gern zitierte Theorie der Polarität des Männlichen und Weiblichen ist (rechts) eingezeichnet worden, zusammen mit den Projektionen der Konzepte; die Polaritätstheorie erscheint ebenfalls plausibel und läßt sich wohl auch in ein System der Dialektik – in dem das eine stets das andere negiert – einfügen, ist aber sicherlich falsch.

Für die Konfigurationen der Fälle (üblicherweise Personen) zeigt sich, dass die einzelnen Fälle jeweils auf einem Ellipsoid (= Ellipse im 2-dimensionalen Fall) liegen, deren Orientierung durch $X'X$ gegeben ist: Der Ansatz (9) und damit die SVD (11) impliziert also die Darstellung 3 (für den 2-dimensionalen Fall), und der Übergang zu den orthogonalen latenten Dimensionen L bedeutet den Übergang zu Koordinaten, die unkorreliert (orthogonal) sind. Die PCA ist also nichts weiter als eine Koordinatentransformation (Rotation).

Der Unterschied zur Faktorenanalyse (FA) besteht darin, dass bei der PCA $r_{jj} = 1$ gesetzt wird, dh r_{jj} ist einfach die Korrelation des j-ten Datenvektors mit sich selbst. IN de Faktorenanalyse wird dafür die *Kommunalität* h_j eingesetzt, die der Reliabilität in der Testtheorie entspricht. Denn in der Tat ist es so, dass zweimalige Messungen derselben Variablen bei den gleichen Personen *nicht*

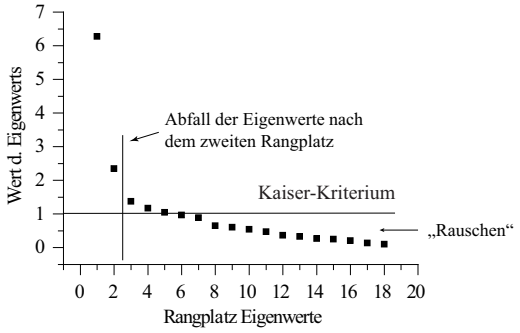
Abbildung 3: Punktekonfiguration und Ellipsen



bedeuten, dass die beiden Messwertreihen mit $r = 1$ korrelieren, da ja bei jeder Messung Messfehler in die Daten eingehen. Je geringer die Varinaz dieser Meßfehler, desto höher die Reliabilität. Im Allgemeinen werden aber bei der PCA nicht alle latenten Vektoren in L berücksichtigt, sondern nur $r < n$, so dass die Schätzung $\hat{r}_{jj}$ auf der Basis dieser ersten r latenten Vektoren (vergl. (14) mit n durch $r < n$ ersetzt) als eine erste Abschätzung der Kommunälität dienen kann.

Die Schätzuung von r ist nicht ganz einfach, man kann vom Scree-Test ausgehen, vergl. Abbildung 4. Derjenige Wert von r, bei dem ein mehr oder weni-

Abbildung 4: Scree-Test



ger scharfer Abfall der Eigenwerte zu beobachten ist, kann als Schätzung für r verwendet werden. Beliebt ist das Kaiser-Kriterium, nach dem alle Eigenwerte kleiner als 1 nur noch Rauschen anzeigen, weshalb die dazu korrespondierenden Eigenvektoren vernachlässigbar seien (s. Abb. 4), allerdings ist dieses Kriterium umstritten. Eine etwas ausführlichere Darstellung der Schätzproblematik findet man im Sktript *Multiple Regression, Multikollinearität, und PCA*.

Diskriminanzanalyse Man spricht auch von der Linearen Diskriminanzanalyse

(LDA) oder Fisherschen Linearen Disktriminanzanalyse (FLDA).

Eine Klassifikation von Fällen ist oft schon mit der multiplen Regression möglich, zB wenn man sagen kann, dass alle Personen mit einem Y-Wert größer als Y_{krit} "geeignet", oder "erkrankt" etc sind. Mehr als zwei Klassen können betrachtet werden, wenn man die Y-Werte eben Intervalle $I_1 : Y \leq Y_1$, $I_2 : Y_1 < Y \leq Y_2$ etc bis $I_k : Y > Y_{k-1}$ aufteilen kann. Aber oft liegen für eine derartige Klassifikation keine gemessenen Y-Werte, sondern nur Klassennmen vor: $Y = 1$ geeignet, $Y = 0$ nicht geeignet, oder $Y = 1$ der Patient leidet an Alzheimer, $Y = 0$ er leidet nicht an Alzheimer. Natürlich kann man auch mehr als nur zwei Kategorien betrachten. Y repräsentiert dann nur Kategoriennamen.

Fisher (1936) hatte die Idee, den Y tatsächlich "Messwerte" nach Maßgabe der Kategorienzugehörigkeit auf der Basis gemessener Prädiktorwerte $X_1, \ldots, X_p$ zuzuordnen:

$$Y = u_1 X_1 + u_2 X_2 + \cdots + u_p X_p \tag{15}$$

Dazu müssen die Gewichte $u_1, \ldots, u_p$ in geeigneter Weise geschätzt werden, und zwar so, dass die Y-Werte den oben genannten Intervallen entsprechen und ein Intervall zu einer Klasse korrespondiert. Es seien nun $\bar{y}_k$, $k = 1, \ldots, K$ die – noch nicht bekannten – Mittelwerte für die K betrachteten Klassen. Die Klassen könnten zB Studienfächer sein, und die X_j sind messbare Indikatoren für die Zugehörigkeit zu einem Studienfach: Kravatte und Anzug kann auf einen Juristen oder BWLer schließen lassen, – aber es gibt Ausnahmen, unübliche Introvertiertheit auf einen Mathematiker oder theoretischen Physiker, – aber auch hier gibt es Ausnahmen, etc. Um die $\mathbf{u} = (u_1, \ldots, u_p)'$ so zu bestimmen, dass die Zuordnung zu einem Studienfach möglichst fehlerfrei geschieht, setzte Fisher das Kriterium:

1. Die Varianz der $\mathbf{y}_k$ soll maximal sein relativ zur
2. Varianz innerhalb der Klassen.

Dazu kann man wie bei der Varianzanalyse vorgehen: man zerlegt die Varianz der y-Werte in eine Quadratsumme "innerhalb" und in eine Quadratsumme "zwischen":

$$\begin{aligned} QS_{total}(y) &= \sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ik} - \bar{y})^2 = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ik} - \bar{y}_k + \bar{y}_k - \bar{y})^2 \\ &= \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ik} - \bar{y}_k)^2}_{QS-innerhalb} + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\bar{y}_k - \bar{y})^2}_{QSzwischen} \end{aligned} \tag{16}$$

Fishers Kriterium ist nun: maximiere

$$\lambda = \frac{QS_{zwischen}}{QS_{innerhalb}} \tag{17}$$

denn wenn man (15) berücksichtigt, sieht man, dass die y-Werte ja von den unbekannten Gewichten $\mathbf{u}$ abhängen. Mit ein bißchen Algebra bestimmt man dan die Matrix B, deren Elemente die (pooled) Varianzen und Kovarianzen "zwischen" (between) sind, und die Matrix W, deren Elemente die Varianzen und Kovarianzen "innerhalb" sind. (17) nimmt dann die Form

$$\lambda(\mathbf{u}) = \frac{\mathbf{u}'B\mathbf{u}}{\mathbf{u}'W\mathbf{u}} \tag{18}$$

an. $\lambda(\mathbf{u})$ ist also eine Funktion der unbekannten Gewichte $u_1, \ldots, u_p$, und mit den Mitteln der Differentialrechung findet man, dass die beste Schätzung $\hat{\mathbf{u}}$ durch die Gleichung

$$W^{-1}B\hat{\mathbf{u}}_k = \lambda_k \mathbf{u}_k, \quad k = 1, \ldots, k_0 \tag{19}$$

d.h. der gesuchte Vektor der Gewichte ist ein Eigenvektor der Matrix $W^{-1}B$, W^{-1} die Inverse zur Matrix W. Der zugehörige Eigenvektor λ_k entspricht dem λ in (18): je größer der Wert von λ_k, desto besser werden die Kategorien auf der k-ten *Diskriminanzfunktion* (auch *kanonische Variate* genannt) getrennt. Denn die Maximierung von (18) zeigt, dass es möglicherweise nicht nur eine Variable Y gibt, auf der zwischen den Kategorien disktiminiert werden kann, sondern mehrere. Gibt es zwei, also $k_0 = 2$, so hat man ein Koordinatensystem $Y_1 \times Y_2$ von zwei Diskriminanzfunktionen Y_1 und und Y_2, und die Fälle liegen in einer Ebene, die durch Geraden in Teilmengen aufgeteilt wird, die zu den einzelnen Klassen korrespondieren. Meistens genügen eine oder zwei Diskriminanzfunktionen.

Im eindimemsionalen Fall (nur eine Y-Variable) entscheidet man nach der Regel

$$C_k \Leftrightarrow |y - \bar{y}_k| = \min, \tag{20}$$

dh man vergleicht den errechneten y-Wert mit den Mittelwerten für die einzelnen Klassen und entscheidet sich für diejenige Klasse, für die die Differenz $|y - \bar{y}_k|$ minimal ist. Die REgel überträgt sich auf den 2-dimensionalen Fall: man entscheidet sich für die Klasse, in deren Bereich y fällt, etc.

Anmerkung: Man beachte, dass Y eine *latente Variable* ist. Gibt es nur eine Diskriminanzfunktion, so definiert sie einen 1-dimensionalen Teilraum im p-dimensionalen Prädiktorraum. Die Orientierung dieser Geraden ist so bestimmt, dass die Projektion der Fälle auf diese Gerade so ausfällt, dass Fälle, die zu einer Klasse gehören, auch auf Y benachbart liegen. Im 2-dimensionalen Fall ergibt sich eine Ebene als Teilraum es p-dimensionalen Prädiktorraums, und die Ebene ist so orientiert, dass Projektion der Fälle auf diese Ebene in die entsprechenden Teilbereiche liefert. Natürlich kann es zu Fehlklassifikationen kommen: wenn die Varianz der Fälle so groß ist, dass zu einer klasse gehörende Fälle in das Intervall oder den Teilbereich einer Nachbarklasse fallen. Aber diese Teilbereiche werden so gwählt, dass die Wahrscheinlichkeit einer Fehlerklassifikation minimalisiert wird. Erlauben die Prädiktoren (Symptome) keine gute Diskriminierung, weil ihre Varianz relativ zur Klassenzugehörigkeit zu groß ist, so kann man nichts machen, – außer, sich andere Prädiktoren zu suchen.

Kanonische Korrelation Die Kanonische Korrelation ist eine Verallgemeinerung der multiplen Regression bzw. Korrelation: man hat zwei "Sätze" von Variablen, $X_1, \ldots, X_p$ und $Y_1, \ldots, Y_q$ und versucht, den einen Datensatz auf der Basis des anderen "vorherzusagen", wobei dann auf latente Variable zurückgegriffen wird. Das Verfahren leistet gute Dienste bei der Analyse von Fragebögen, aber auch bei der Analyse von Daten von Vorher-Nachher-Untersuchungen: man hat Daten von Patienten vor einer Intervention oder Therapie, und Daten für dieselben Variablen ($Y_j = X_j$) nach der Intervention oder Therapie und möchte wissen, ob die Intervention eine Wirkung hatte, die sich in den Daten ausdrückt.

Der Grundgedanke besteht darin, ähnlich wie bei der PCA jeden Datensatz durch bestimmte latente Variable zu "erklären", also eine Menge $\mathbf{P}_1, \ldots, \mathbf{P}_r$ von latenten Vektoren für die X-Variablen und eine Menge $\mathbf{Q}_1, \ldots, \mathbf{Q}_r$ für die Y-Variablen derart zu bestimmen (dies sind nicht die Vektoren aus den Q- und P-Matrizen der Singularwertzerlegung!), dass $\mathbf{P}_1$ und $\mathbf{Q}_1$ maximal korrelieren, $\mathbf{P}_2$ und $\mathbf{Q}_2$ zweitmaximal korrelieren, etc. Sind diese Korrelationen hoch, so kann man sagen, dass die X- und Y-Variablen eng zusammenhängen, sind sie niedrig, so sind X- und Y-Variablen weitgehend unabhängig. Hat also eine Therapie einen guten Erfolg, so sollten die X-Variablen (die Messungen bestimmter Symptome vor der Therapie) und die Y-Variablen (sind sind die X-Variablen nach der Therapie) einen möglichst geringen Zusammenhang aufweisen, hat sie keinen Erfolg, so haben sich die Symptome nicht oder nur wenig verändert und der Zusammenhang zwischen X (vorher) und Y (X-Werte nachher) ist hoch.

Aus der PCA ist bekannt, dass man vom Ansatz $X = LP'$ ausgehend zu $XP = L$ übergehen kann, dh man kann die Spaltenvektoren von L aus den Spaltenvektoren der Datenmatrix X ausrechnen, sobald man P (die Matrix der Eigenvektoren von $X'X$) berechnet hat. In analoger Weise geht man hier vor: Man sucht zwei Matrizen A und B derart, dass

$$XA = P = [\mathbf{P}_1, \ldots, \mathbf{P}_r] \tag{21}$$

$$YB = Q = [\mathbf{Q}_1, \ldots, \mathbf{Q}_R] \tag{22}$$

und

$$\rho(\mathbf{P}1, \mathbf{Q}_1) \geq \rho(\mathbf{P}_2, \mathbf{Q}_2) \geq \cdots \geq \rho(\mathbf{P}_r, \mathbf{Q}_r), \tag{23}$$

wobei ρ die jeweilige Korrelation bezeichnet. (21) bzw. (22) können für den s-ten Vektor in der Form

$$\mathbf{P}_s = X\mathbf{a}_s, \quad \mathbf{Q}_s = Y\mathbf{b}_s, \quad s = 1, \ldots, r \tag{24}$$

geschrieben werden, wobei $\mathbf{a}_s$ und $\mathbf{b}_s$ sind die s-ten Spaltenvektoren von A bzw. B. Dann ist

$$\rho(\mathbf{P}_s, \mathbf{Q}_s) = \frac{Kov(\mathbf{P}_s\mathbf{Q}_s)}{\sqrt{Var(\mathbf{P}_s)Var(\mathbf{Q}_s)}} \tag{25}$$

(dies ist einfach die Produkt-Moment-Korrelationsformel). Natürlich ist diese Ausdruck in Matrix- bzw. Vektorschreibweise gleich

$$\rho(\mathbf{P}_s, \mathbf{Q}_s) = \frac{\mathbf{P}_s' \mathbf{Q}_s}{\|\mathbf{P}_s\| \|\mathbf{Q}_s\|} = \frac{\mathbf{a}_s' X' Y \mathbf{b}_2}{} \tag{26}$$

Ein ausführliches Beispiel findet sich im Skriptum 'Kanonische Korrelation'.