.

U. Mortensen

Multiple Regression, Multikollinearität, und PCA

21. 05. 2013

Inhaltsverzeichnis

1 Multiple Regression

1.1 Der Ansatz

Gegeben seien p Prädiktorvariablen (entsprechend "Symptomen") $X_1, \ldots, X_p$ und eine Kriteriumsvariable Y, die anhand der Gleichung

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

vorhergesagt werden soll. Dies ist ein *linearer* Ansatz: die unabhängigen Variablen werden nach einer möglichen Gewichtung (Multiplikation mit den b_j) nur addiert. Der Ansatz ist allerdings verallgemeinerbar, indem man nichtlineare Terme wie $X_j X_k j$, $x_j^2 X_k^3$ etc hinzufügt. Solche Terme ergeben sich eventuell aus theoretischen Betrachtungen oder aus anderen empirischen Beobachtungen. Man hat dann zB

$$Y = b_0 + b_1 X_1 + \cdots + b_p X_p + b_{p+1} X_j X_k + b_{p+2} X_l^2 X_m^4 + e \tag{2}$$

Durch Umbenennung ("Reparametrisierung") $Z_j = X_j$, $Z_{p+1} = X_j X_k$, $Z_{p+2} = X_l^2 X_m^4$ etc wird daraus wieder ein in den Regressionskoeffizienten $b_0, b_1, \ldots$ linearer Ansatz:

$$Y = b_0 + b_1 Z_1 + \cdots + b_p Z_p + b_{p+1} Z_{p+1} + b_{p+2} Z_{p+2} + e \tag{3}$$

Die Z_j bedeuten hier also nicht, dass die Variablen standardisiert wurden. Diese Gleichung ist wieder linear in den Z_j.

Für eine gegebene Stichprobe von m Fällen kann man zur Vektor- bzw. Matrixnotation übergehen:

$$\mathbf{Y} = b_0 \vec{1} + b_1 \mathbf{Z}_1 + \cdots + b_p \mathbf{Z}_p + \mathbf{e}, \tag{4}$$

wobei $\vec{1} = (1, 1, \ldots, 1)'$ (m-dimensional) ist. Nochmalige Umbenennung der $\mathbf{Z}_j$ in $\mathbf{X}_j$ liefert also

$$\mathbf{Y} = b_0 \vec{1} + b_1 \mathbf{X}_1 + \cdots + b_p \mathbf{X}_p + \mathbf{e}, \tag{5}$$

wobei die Möglichkeit, dass einige der $\mathbf{X}_j$ sich aus nichtlinearen Termen anderer Prädiktoren zusammemsetzen, zugelassen wird, – die $\mathbf{Z}$-Notation kann dann für standardisierte Variablen reserviert werden.

Es ist of sinnvoll, zu zentrierten Vektoren überzugehen, also

$$\mathbf{X}_j \to \mathbf{x}_j = \mathbf{X}_j - \bar{x}_j \vec{1}. \tag{6}$$

Die additive Konstante b_0 entfällt dann (wie bei der Standardisierung). In noch kompakterer Schreibweise hat man

$$\mathbf{Y} = X\mathbf{b} + \mathbf{e}, \tag{7}$$

$\mathbf{b} = (b_0, b_1, \ldots, b_p)'$, $\mathbf{e} = (e_1, \ldots, e_m)'$. Wie schon angemerkt entfällt b_0, wenn statt $\mathbf{Y}$ der zentrierte Vektor $\mathbf{y}$ betrachtet wird (wobei dann auch die $\mathbf{x}_j$ und nicht die $\mathbf{X}_j$ die Matrix X definieren. Im nichtzentrierten Fall ist also

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ & & & \vdots & \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mp} \end{pmatrix}, \tag{8}$$

im zentrierten Fall entfällt der 1-Vektor.

1.2 Schätzung der Parameter mit der Methode der Kleinsten Quadrate

Die KQ-Methode[1] besteht darin, die Summe der Fehlerquadrate, also $\mathbf{e}'\mathbf{e}$ zu minimieren. Der Wert dieser Summe ist eine Funktion des Vektors $\mathbf{b}$. Man hat demnach

$$Q(\mathbf{b}) = (\mathbf{Y} - X\mathbf{b})'(\mathbf{Y} - X\mathbf{b}) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'X\mathbf{b} - \mathbf{b}'X'\mathbf{Y} + \mathbf{b}'X'X\mathbf{b}. \tag{9}$$

Da $X\mathbf{b}$ ein Vektor ist folgt, dass $\mathbf{Y}'X\mathbf{b} = \mathbf{b}'X'\mathbf{Y}$ Skalare sind. Man bildet nun die partiellen Ableitungen[2] nach den einzelnen Komponenten b_i des Vektors $\mathbf{b}$:

$$\frac{\partial \mathbf{b}}{\partial b_i} = (0, \ldots, 0, 1, 0, \ldots, 0)',$$

wobei die 1 asn der i-ten Stelle steht. Dann hat man[3]

$$\frac{\partial Q(\mathbf{b})}{\partial b_i} = -2\mathbf{Y}'X\mathbf{e}_i + \mathbf{e}_i'X'X\mathbf{b} + \mathbf{b}'X'X\mathbf{e}_i = -2\mathbf{Y}'X\mathbf{e}_i + 2\mathbf{e}_i'X'X\mathbf{b}$$

denn $\mathbf{e}_i'X'X\mathbf{b} = \mathbf{b}'X'X\mathbf{e}_i$, weil beide Ausdrücke Skalare sind. Die Gleichungen werden gleich Null gesetzt, um den Vektor $\hat{\mathbf{b}}$ zu bestimmen, für den $\mathbf{e}'\mathbf{e}$ minimal wird. Man kann die Ausdrücke für alle $i = 1, \ldots, p$ zusammenfassen; dies läuft darauf hinaus, dass die $\mathbf{e}_i$ durch die Einheitsmatrix I ersetzt werden, so dass man die Gleichung

$$\mathbf{Y}'X = (X'X)\hat{\mathbf{b}} \tag{10}$$

erhält, woraus durch Multiplikation von links mit $(X'X)^{-1}$

$$\hat{\mathbf{b}} = (X'X)^{-1}X'\mathbf{Y} \tag{11}$$

[1] KQ = Kleinste Quadrate

[2] Die Methode wird explizit im Skriptum *Vektoren und Matrizen für Multivariate Verfahren* (VMMVA) dargestellt.

[3] Anwendung der Produkt- und der Kettenregel

folgt. Diese Lösung setzt voraus, dass die Inverse $(X'X)^{-1}$ existiert, – die Annahme ist aber nicht unvernünftig, da die Messungen ja im Allgemeinen fehlerbehaftet und die Vektoren in der Datenmatrix X deshalb linear unabhängig sind (eine notwendige Bedingung für die Existenz der Inversen). Trotzdem können sich Probleme ergeben, die im Zusammenhang mit dem Begriff der Multikollinearität diskutiert werden (s. Abschnitt 1.4). Hier werden zunächst noch einige Implikationen von (11) betrachtet.

Projektion Setzt man in (7) die Schätzung $\hat{\mathbf{b}}$ für $\mathbf{b}$ ein, so erhält man

$$\mathbf{Y} = X\hat{\mathbf{b}} + \hat{\mathbf{e}} = \hat{\mathbf{Y}} + \hat{\mathbf{e}}, \tag{12}$$

wobei $\hat{\mathbf{Y}} = X\hat{\mathbf{b}}$ und $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$. Der geschätzte Fehlervektor $\hat{\mathbf{e}}$ unterscheidet sich von dem in (7) auftretenden Fehlervektor $\mathbf{e}$: $\hat{\mathbf{e}}$ ist der Fehler, der bei der "Vorhersage" von $\mathbf{Y}$ durch die KQ-Schätzung $\hat{\mathbf{Y}}$ auftritt. Bei Verwendung einer anderen Schätzmethode[4] kann ein anderer Fehlervektor $\hat{\mathbf{e}}$ entstehen. Setzt man in (7) den Ausdruck (11) für $\hat{\mathbf{b}}$ ein, erhält man

$$\mathbf{Y} = X(X'X)^{-1}X'\mathbf{Y} + \hat{\mathbf{b}} \tag{13}$$

In diesem Ausdruck tritt die Matrix

$$P_r = X(X'X)^{-1}X' \tag{14}$$

auf; diese Matrix heißt *Projektionsmatrix*. Durch Nachrechnen[5] findet man, dass P_R symmetrisch ist:

$$P_r = P_r'. \tag{15}$$

P_r hat die merkwürdige Eigenschaft der *Idempotenz*, dh

$$P_r^2 = P_r P_r = (X(X'X)^{-1}X')(X(X'X)^{-1}X') = X(X'X)^{-1}X' = P_r \tag{16}$$

gilt[6]. Daraus folgt $P_r^n = P_r$ für alle $n > 0$. Man findet, dass $\hat{\mathbf{Y}}$ und $\hat{\mathbf{e}}$ orthogonal sind:

$$\hat{\mathbf{Y}}'\hat{\mathbf{e}} = X'P_r(\mathbf{Y} - \hat{\mathbf{Y}}) = X'P_r\mathbf{Y} - X'P_rP_r\mathbf{Y} = 0 \tag{17}$$

wegen der Idempotenz von P_r. Da Orthogonalität lineare Unabhängigkeit impliziert folgt, dass $\hat{\mathbf{Y}}$ und $\hat{\mathbf{e}}$ unkorreliert und deshalb linear unabhängig sind. Weiter gilt

$$\hat{\mathbf{Y}} = X\hat{\mathbf{b}} = P_r\mathbf{Y}. \tag{18}$$

[4]Zum Beispiel die Maximum Likelihood-Methode, – die aber bei normalverteilten Daten die gleichen Resultate liefert wie die KQ-Methode, bei anderen Verteilungen aber zu anderen Schätzungen und damit zu einem anderen Fehlervektor $\hat{\mathbf{e}}$ führt.

[5]$(AB)' = B'A'$

[6]$(AB)C = A(BC)$

Der erste Teil, $\hat{\mathbf{Y}} = X\hat{\mathbf{b}}$, bedeutet, dass $\hat{\mathbf{Y}}$ eine Linearkombination der Spalten von X ist, dh $\hat{\mathbf{Y}}$ ist ein Element des Vektorraums $\mathcal{C}(X)$, also des Vektorraums, der durch die Spalten von X *aufgespannt* wird (die Spalten von X können linear abhängig sein, dann ist $\mathcal{C}(X) = \mathcal{C}(B)$, B eine Menge von linear unabhängigen Vektoren, die es ermöglichen, die Spalten von X als Linearkombinationen darzustellen). Der zweite Teil, $\hat{\mathbf{Y}} = P_r$ bedeutet, dass $\hat{\mathbf{Y}}$ eine Projektion des Vektors $\mathbf{Y}$ auf $\mathcal{C}(X)$ ist, was den Ausdruck 'Projektionsmatrix' erklärt.

1.3 Der Multiple Korrelationskoeffizient

Es wird zuerst eine Zerlegung von $\|\mathbf{Y}\|^2$ betrachtet. Denn $\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\mathbf{e}}$, so dass

$$\begin{aligned} \mathbf{Y}'\mathbf{Y} &= \|\mathbf{Y}\|^2 = (\hat{\mathbf{Y}} + \mathbf{e})'(\hat{\mathbf{Y}} + \mathbf{e}) \\ &= \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\mathbf{e}}'\hat{\mathbf{e}} + 2\hat{\mathbf{Y}}'\hat{\mathbf{e}}, \end{aligned}$$

d.h. es ist

$$\|\mathbf{Y}\|^2 = \|\hat{\mathbf{Y}}\|^2 + \|\hat{\mathbf{e}}\|^2 \tag{19}$$

da nach (17) $\hat{\mathbf{Y}}'\hat{\mathbf{e}} = 0$. Für zentrierte Werte ist $\|\mathbf{Y}\|^2$ proportional zur Varianz der Y-Werte, und $\|\hat{\mathbf{Y}}\|^2$ ist proportional zur Varianz der vorhergesagten Werte $\hat{\mathbf{Y}}$; $\|\hat{\mathbf{e}}\|^2$ ist die Varianz der Fehler. Dividiert man beide Seiten von (51) durch $\|\mathbf{Y}\|^2$, so erhält man

$$1 = \frac{\|\hat{\mathbf{Y}}\|^2}{\|\mathbf{Y}\|^2} + \frac{\|\hat{\mathbf{e}}\|^2}{\|\mathbf{Y}\|^2} \tag{20}$$

oder

$$\frac{\|\hat{\mathbf{Y}}\|^2}{\|\mathbf{Y}\|^2} = 1 - \frac{\|\hat{\mathbf{e}}\|^2}{\|\mathbf{Y}\|^2}, \tag{21}$$

dh der Anteil der Varianz der vorgesagten Werte an der Gesamtvarianz ist gleich 1 minus dem Anteil der Fehlervarianz an der Gesamtvarianz der Y-Werte.

Nun werde das Skalarprodukt $\hat{\mathbf{Y}}'\mathbf{Y}$ betrachtet. Bei zentrierten Werten ist es gleich der Kovarianz zwischen den gemessenen Werten $\mathbf{Y}$ und den vorhergesagten Werten $\hat{\mathbf{Y}}$. Es ist

$$\begin{aligned} \mathbf{Y}'\hat{\mathbf{Y}} &= \mathbf{Y}'(\mathbf{Y} - \hat{\mathbf{e}}) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\hat{\mathbf{e}} \\ &= \|\mathbf{Y}\|^2 - (\hat{\mathbf{Y}} + \hat{\mathbf{e}})'\hat{\mathbf{e}} = \|\mathbf{Y}\|^2 - \hat{\mathbf{Y}}'\hat{\mathbf{e}} - \hat{\mathbf{e}}'\hat{\mathbf{e}} \\ &= \|\mathbf{Y}\|^2 - \|\hat{\mathbf{e}}\|^2, \end{aligned} \tag{22}$$

denn $\hat{\mathbf{Y}}'\hat{\mathbf{e}} = 0$ nach (17). Die Gleichung (22) bedeutet $\|\mathbf{Y}\|^2 = \mathbf{Y}'\hat{\mathbf{Y}} + \|\hat{\mathbf{e}}\|^2$, dh

$$\mathbf{Y}'\hat{\mathbf{Y}} = \|\mathbf{Y}\|^2 - \|\hat{\mathbf{e}}\|^2 \tag{23}$$

Nach einer Division durch $\|\mathbf{Y}\|^2$ erhält man

$$\frac{\mathbf{Y}'\hat{\mathbf{Y}}}{\|\mathbf{Y}\|^2} = 1 - \frac{\|\hat{\mathbf{e}}\|^2}{\|\mathbf{Y}\|^2}, \tag{24}$$

und in Kombination mit (21) hat man

$$\frac{\mathbf{Y}'\hat{\mathbf{Y}}}{\|\mathbf{Y}\|^2} = \frac{\|\hat{\mathbf{Y}}\|^2}{\|\mathbf{Y}\|^2}. \tag{25}$$

Für nicht nur zentrierte, sondern standardisierte Werte ist $\|\mathbf{Y}\|^2 = 1$; dann ist $\hat{\mathbf{Y}}'\mathbf{Y} = R^2$. R ist der *multiple Regressionskoeffizient*. R^2 ist also der Anteil der Varianz der vorhergesagten Werte an der Varianz der Y-Werte. Wieder in Kombination mit (21) kann man

$$R^2 = 1 - \frac{\|\hat{\mathbf{e}}\|^2}{\|\mathbf{Y}\|^2} \tag{26}$$

schreiben; dieser Ausdruck ist aus der einfachen Regressionsrechnung bekannt, wo r_{xy}^2 als *Determinationskoeffizient* bezeichnet wird.

Von Interesse ist die Interpretation der Komponenten $\hat{b}_j$ von $\hat{\mathbf{b}}$. Sie könnten ja Aufschluß über den Anteil geben, mit dem der Prädiktor $\mathbf{X}_j$ in die abhängige Variable, also die Kriteriumsvariable $\mathbf{Y}$ eingeht. Aber:

> "The problem of interpreting the regression function is a thorny one. It ist interesting to speculate about the relative contributions of the predictor elements to the prediction of the criterion element, but we habe to temper our interpretations with the realization that the obtained prediction results from a *system of predictors* in which the elements interact in a complex fashion." Aus: Cooley, W.W., Lohnes, P. R.: Multivariate Data Analysis. New York 1971.

Seit 1971 hat man sich zu diesem Problem einige Gedanken gemacht, und in den folgenden Abschnitten wird auf Lösungsmöhlichkeiten eingegangen. Um das Problem anzugehen soll zunäschst ein Blick auf die Eigenschaften der Schätzung $\hat{\mathbf{b}}$ geworfen werden.

1.4 Multikollinearität und Eigenschaften der Schätzungen

Es werden der Einfachheit halber zentrierte Vektoren $\mathbf{x}_j = \mathbf{X}_j - \bar{x}_j\vec{1}$ betrachtet.

Die Prädiktoren $\mathbf{x}_j$ heißt *multikollinear*, wenn für mindestens zwei von ihnen die Beziehung

$$\mathbf{x}_k = a_{jk}\mathbf{x}_j + e_{jk}, \quad b_{jk} \neq 0 \tag{27}$$

besteht. Wäre der Fehler $e_{jk} = 0$, so wären $\mathbf{x}_j$ und $\mathbf{x}_k$ also linear abhängig. Dann ist der Rang von X und damit der von $X'X$ kleiner als $\min(m, n)$ und $(X'X)^{-1}$ existiert nicht, d.h. kann nicht bertechnet werden. Allerings sind wegen des Messfehlers alle Vektoren von X *numerisch* linear unabhängig, so dass X und $X'X$ vollen Rang haben, – aber dennoch ist $X'X$ *ill conditioned*, so dass die Berechnung von $(X'X)^{-1}$ führt zu schlechten Schätzungen der Parameter. Die Implikationen der Multikollinearität können an einem einfachen Beispiel, aufgezeigt werden. Dazu soll der Fall von nur zwei Prädiktoren betrachtet werden. Von den zentrierten Vektoren $\mathbf{x}_1, \mathbf{x}_2$ soll dabei zu standardisierten Vektoren übergegangen werden; $a = a_{12}$ wird dann zu einem Korrelationskoeffizienten $r = r_{12}$:

$$\mathbf{z}_y = b_1\mathbf{z}_1 + b_2\mathbf{z}_2 + \mathbf{e} \tag{28}$$

Der Matrix X entspricht dann die $(m \times 2)$-Matrix Z, und man hat

$$R = \frac{1}{m}Z'Z = \frac{1}{m}\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \tag{29}$$

mit $r = r_{12} = a$, und $z_2 = rz_1 + e$. Gesucht ist die inverse Matrix $(Z'Z)^{-1}$. Man findet

$$(Z'Z)^{-1} = \frac{1}{1-r^2}\begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}. \tag{30}$$

Für unabhängige Prädiktoren hat man im Idealfall $r = 0$ und $Z'Z$ ist einfach die Einheitsmatrix. Für $r \to 1$ hingegen strebt $1-r^2 \to 0$ und $1/(1-r^2)$ und damit die Elemente von $Z'Z$ gegen unendlich. Dieser Sachverhalt erzeugt interpretatorische Schwierigkeiten, wie im folgenden Abschnitt deutlich wird.

Eigenschaften der Schätzungen: Von Interesse sind stets die Eigenschaften von Schätzungen: (i) sind die Schätzungen verzerrt (haben sie einen "Bias") oder nicht, (ii) wie groß sind die Varianzen und Kovarianzen, – hier zwischen verschiedenen Komponenten von $\mathbf{b}$. Man findet (hier ohne Beweis):

$$\mathbb{E}(\hat{\mathbf{b}}) = \mathbf{b} \tag{31}$$

$$Kov(\hat{\mathbf{b}}) = \sigma^2(X'X)^{-1} \tag{32}$$

$\mathbb{E}$ steht für 'Erwartungswert', Kov für die Varianz-Kovarianzmatrix der Komponentenschätzungen. (31) besagt, dass die Schätzung *unverzerrt* (biasfrei) ist, dh die Schätzungen weichen nur zufällig vom wahren Vektor $\mathbf{b}$ ab, es gibt keine systematische Unter- oder Überschätzung. (47) besagt, dass die Varianzen und Kovarianzen (i) von σ^2, der Varianz des Fehlers e, (ii) von der Inversen $(X'X)^{-1}$ abhängen. Sind die Elemente von X standardisiert, so ist $X'X = R$ die Korrelationsmatrix (bis auf einen Faktor $1/m$), $(X'X)^{-1}$ ist dann die Inverse R^{-1} der Korrelationsmatrix. Um den Effekt von $(X'X)^{-1}$ zu sehen, wird diese Matrix für

den Fall einer (2×2)-Matrix illustriert, allerdings soll ein in diesem Zusammenhang auftretender Begriff eingeführt werden.

Es genügt, den Fall standardisierter Messwerte zu betrachten. In diesem Fall gilt also

$$Kov(\hat{\mathbf{b}}) = \frac{\sigma^2}{1-r^2} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix} \tag{33}$$

Die Diagonalzellen von $Z'Z$ geben die Varianzen der Schätzungen der Komponenten an: sie sind gleich $\sigma^2/(1-r^2)$ und gehen für $r \to 1$ gegen unendlich, dh man kann Regressionsgewichte erhalten, die beliebig weit von den wahren Werten abweichen, – unabhängig davon, wie groß der Stichprobenumfang auch ist. Die Kovarianz der Komponenten von $\hat{\mathbf{b}}$ ist *negativ*: ist $\hat{b}_1 > 0$, so ist $\hat{b}_2 < 0$ und umgekehrt. Diese Gegenläufigkeit der Schätzungen resultiert aus den Eigenschaften von $(X'X)^{-1}$ bzw $(Z'Z)^{-1}$.

Diese Eigenschaften der Schätzung $\hat{\mathbf{b}}$ machen das Problem der Interpretation so dornig, um an die Bemerkung Cooley & Lohnes' zu erinnern.

1.5 Ridge-Regression und andere Verfahren

Es gibt verschiedene Methoden, das Problem korrelierender Prädiktoren anzugehen. Ein Ansatz ist die *Stepwise Regression*, bei der aus der Menge der gegebenen Prädiktoren diejenigen ausgewählt werden, die mög-lichst wenig miteinander korrelieren und die die Kriteriumsvariable hinreichend gut voraussagen. Man kann die (i) *forward selection* wählen, bei der man mit einer Prädiktorvariablen anfängt und sukzessive weitere hinzufügt, bis ein akzeptabler Fit erreicht ist, oder (ii) man geht umgekehrt vor, indem man mit allen Variablen beginnt und sukzessive Prädiktorvariablen entfernt, bis die erwünschte Unkorreliertheit bei gutem Fit erreicht worden ist, oder (iii) man kombiniert beide Ansätze. Das Verfahren wird mittlerweile computerisiert nahezu automatisch durchgeführt. Die mit dem Ansatz verbundenen Probleme sind vielfach[7]; ein Hauptargument der Kritik ist das *data dredging* – man fischt so lange, bis man ein Modell gefunden hat, das einem plausibel oder erwünscht erscheint.

Ein anderer Ansatz besteht darin, eine PCA auf die Matrix X der Prädiktoren anzuwenden. Man erhält latente Variable, die unkorreliert sind und behält nur so viele von ihnen als Prädiktoren, wie für eine gute Vorhersage notwendig sind. Das Verfahren wird später ausführlich vorgestellt.

Eine andere Klasse von Verfahren sind die *Shrinkage*-Verfahren, bei denen die geschätzten Regressionskoeffizienten "geschrumpft" werden. Der Ansatz geht auf

[7] Chatfield, C. (1995) Model uncertainty, data mining and statistical inference, *J. R. Statist. Soc. A 158*, Part 3, pp. 419–466.

den russischen Mathematiker Andrey Nikolayevich Tikhonov (1906 – 1993) zurück, der es im Zusammenhang mit Untersuchungen zur Lösung von Gleichungen entwickelte. So hat ein lineares Gleichungssystem die Form $A\mathbf{x} = \mathbf{y}$, und $\mathbf{x}$ ist der Vektor der Unbekannten. Die Lösung liegt eindeutig fest, wenn die Inverse A^{-1} existiert. Die Lösung ist dann

$$A^{-1}A\mathbf{x} = \mathbf{x} = A^{-1}\mathbf{y}.$$

Probleme treten auf, wenn entweder A^{-1} gar nicht existiert, wenn also etwa die Spaltenvektoren von A linear abhängig sind, oder wenn ausgeprägte Multikollinearitäten existieren, in welchem Fall A^{-1} "schlecht konditioniert"[8] ist und die Schätzungen instabil sind. Tychonoff ging von dem Ansatz

$$(A\mathbf{x} - \mathbf{y})'(A\mathbf{x} - \mathbf{y}) + \mathbf{x}'\Gamma'\Gamma\mathbf{x} \stackrel{!}{=} \min \tag{34}$$

aus, wobei Γ eine geeignet zu wählende Matrix (Tychonoff-Matrix) ist. Man spricht von der *Tychonoff-Regularisierung*, nud $\mathbf{x}'\Gamma'\Gamma\mathbf{x} = \|\Gamma\mathbf{x}|^2$ heißt *Regularisierungsterm*. Tychonoffs Ansatz war lange nicht bekannt, da er nur auf russisch veröffentlicht worden war. Hoerl & Kennard (1970)[9] entwickelten einen analogen Ansatz für den Spezialfall $\Gamma = \lambda I$, I die Einheitsmatrix. Γ ist dann eine Diagonalmatrix, deren Diagonalelemente gleich λ sind, und λ muß für ein gegebenes Problem geschätzt werden. Wendet man diesen Ansatz auf die Schätzung des Parametervektors $\mathbf{b}$ an, mit $A = X'X$, so findet man nach entsprechender Umformulierung

$$\hat{\mathbf{b}}_\lambda = (X'X + \lambda^2 I)^{-1}X'\mathbf{Y}. \tag{35}$$

Dieser Ansatz wurde von Hoerl & Kennard *Ridge-Regression* genannt, weil der "ridge"[10] der Matrix, also die Diagonalzellen, um den Wert λ erhöht werden. Der Effekt dieser Erhöhung ist eine "Schrumpfung" (shrinkage) der Schätzungen und damit eine Reduktion ihrer Varianz (s. unten). Die folgenden Aussagen lassen sich beweisen:

1. $X'X + \lambda^2 I$ hat dieselben Eigenvektoren wir $X'X$.
2. $X'X + \lambda^2 I$ hat die Eigenvektoren $\lambda_j + \lambda^2$.

Die symmetrische Matrix $X'X$ möge die Matrix der Eigenvektoren P haben, assoziiert mit der Diagonalmatrix Λ der Eigenwerte, so dass

$$X'X = P\Lambda P'. \tag{36}$$

[8] Mathematiker- und Statistikerjargon

[9] Hoerl, A. E., Kennard, R. W. (1970) Ridge Regression: Biased Estimation for Nonorthognal Problems. *Technometrics*, 12(1), 55–67

[10] ridge = Dachfirst, Felsgrat, Hügelkette, ...

Dann ist

$$(X'X)^{-1} = (P\Lambda P')^{-1} = (P')^{-1}\Lambda^{-1}P^{-1}.$$

Aber wegen der Orthonormalität von P ist $P^{-1} = P'$, deshalb folgt $(P')^{-1} = P$, so dass

$$(X'X)^{-1} = P\Lambda^{-1}P'. \tag{37}$$

Man findet dann

$$(X'X + \lambda^2 I)^{-1} = PD^{-1}P', \tag{38}$$

wobei

$$D^{-1} = \text{diag}\left(\frac{1}{\lambda_1 + \lambda_1}, \ldots, \frac{1}{\lambda_n + \lambda^2}\right). \tag{39}$$

Unter Anwendung des Äußeren Vektorprodukts hat man dann für die Schätzung von $\mathbf{b}$

$$\hat{\mathbf{b}}_\lambda = \left(\sum_{j=1}^{n} \frac{\mathbf{P}_j \mathbf{P}_j'}{\lambda_j + \lambda^2}\right) X'\mathbf{Y}. \tag{40}$$

Hohe Korrelationen zwischen den Spalten von X implizieren kleine Eigenwerte λ_j. Ist $\lambda = 0$, so sieht man, dass kleine Eigenwerte große Komponenten von $\hat{\mathbf{b}}$ implizieren, – dies geschieht, wenn man die multiple Regression direkt anwendet. ein positiver Wert von λ impliziert aber, dass die Komponenten von $\hat{\mathbf{b}}$ klein bleiben. $\lambda \neq 0$ "regularisiert" also die Schätzung der Regressionsgewichte.

Nach der Singularwertzerlegung von X gilt bekanntlich

$$X = Q\Lambda^{1/2}P', \tag{41}$$

wobei Q die orthonormale Matrix der Eigenvektoren von XX' ist, P ist die orthonormale Matrix der Eigenvektoren von $X'X$, und $\Lambda^{1/2}$ ist die Diagonalmatrix der Wurzeln aus den von Null verschiedenen Eigenwerten von $X'X$ bzw XX'. Setzt man die rechte Seite von (72) in (40) ein, so erhält man

$$\hat{\mathbf{b}}_\lambda = \left(\sum_{j=1}^{n} \frac{\mathbf{P}_j \mathbf{P}_j'}{\lambda_j + \lambda^2}\right) P\Lambda^{1/2}Q'\mathbf{Y} = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \lambda^2}\mathbf{P}_j Q_j' Y \tag{42}$$

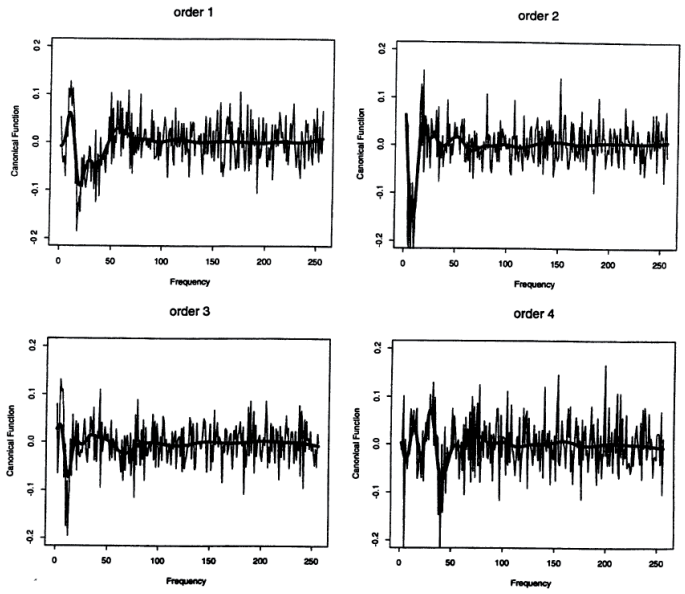
Weiter ist $\hat{\mathbf{Y}} = X\hat{\mathbf{b}}_\lambda$. Macht man wieder von von der SVD $X = Q\Lambda^{1/2}P'$ Gebrauch, so erhält man

$$\hat{\mathbf{Y}} = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \lambda^2} Q_j Q_j' \mathbf{Y} \tag{43}$$

Man kann jetzt den Effekt der Regularisierung verdeutlichen. Es sei ein Wert für λ gegeben. $g(\lambda_j) = \sqrt{\lambda_j}/(\lambda_j + \lambda^2)$; dann ist

$$\lim_{\lambda_j \to 0} g(\lambda_j) = \lim_{\lambda_j \to 0} \frac{\sqrt{\lambda_j}}{\lambda_j + \lambda^2} = \lim_{\lambda_j \to 0} \frac{1}{\lambda_j + \lambda^2/\sqrt{\lambda_j}} = 0. \tag{44}$$

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



Analog dazu findet man

$$\lim_{\lambda_j \to 0} f(\lambda_j) = \lim_{\lambda_j \to 0} \frac{\lambda_j}{\lambda_j + \lambda} = \lim_{\lambda_j \to 0} \frac{1}{1 + \lambda^2/\lambda_j} = 0. \tag{45}$$

Die hier eingeführte Funktion $f(\lambda_j) = \lambda_j/(\lambda_j + \lambda^2)$ heißt *Filter*. Der Grenzwert (44) zeigt, wie sich die Regularisierung auf die Schätzung der Komponenten von $\mathbf{b}$ auswirken: die Faktoren $g(\lambda_j)$ werden um so kleiner, je kleiner λ_j, und damit werden auch die geschätzten Komponenten von $\mathbf{b}$ kleiner. (45) zeigt, wie sich diese Schrumpfung auf die Schätzung $\hat{\mathbf{Y}}$ auswirkt – ebenfalls dämpfend. Mit kleiner werdenden λ_j-Werten werden die Komponenten von $\hat{\mathbf{b}}$ nicht mehr größer, sondern kleiner werden – daher der Ausdruck Shrinkage (Schrumpfung). Das Ausmaß der Schrumpfung hängt vom Wert von λ ab. Die Regularisierung reduziert damit die Varianz der Schätzungen und führt zu stabileren Schätzungen als die unregularisierte KQ-Schätzung. Insbesondere werden diejenigen Komponenten, die zu Prädiktoren mit insignifikantem Beitrag zur Vorhersage von Y korrespondieren, gegen Null geschrumpft, was die Interpetierbarkeit der Schätzung $\hat{\mathbf{b}}_\lambda$ erhöht. Abb.[11] 1 illustriert sowohl den Effekt hoher Korrelationen zwischen den Prädikto-

[11] aus: Hastie, T., Buja, A., Tibshirani, R.; (1995) Penalized Discriminant Analysis. *The Annals of Statistics* 23 (1), 73 – 102

ren: die Schätzungen oszillieren wegen der negativen Vorzeichen in $(X'X)^{-1}$ und haben stark überhöhte Werte. Die Schrumpfung bewirkt, dass irrelevante Komponenten gegen Null gehen und nur die relevanten "überleben": die durchgezogene Kurve repräsentiert die korrigierten Schätzungen.

Nun seien die Eigenwerte λ_j gegeben und es werde der Effekt von λ betrachtet. Offenbar gilt

1. $\lambda \to 0$ impliziert $f(\lambda_j) \to 1$ für alle j, d.h. $\hat{\mathbf{b}}_\lambda \to \hat{b}b$, d.h. man erhält den Fall der gewöhnlichen KQ-Schätzung (OLS = Ordinary Least Square),

2. $\lambda \to \infty$ impliziert $f(\lambda_j) \to 0$, d.h. $\hat{\mathbf{b}}_\lambda \to 0$; d.h. je größer der Regularisierungsterm λ, desto kleiner die Werte der geschätzten Regressionskoeffizienten.

1.6 Modellkomplexität und Bias-Varianz-Tradeoff

Es werde noch einmal die einfache (nicht regularisierte oder penalisierte) multiple Regression betrachtet:

$$\mathbf{Y} = X\mathbf{b} + \mathbf{e},$$

X eine $N \times p$-Matrix (n Fälle, p Prädiktoren) Die Komponenten des Vektors $\mathbf{b}$, also die Regressionsgewichte, sind nicht bekannt, aber es liegen Messungen $\mathbf{Y}$ für die Kriteriumsvariable und X für die Prädiktorvariablen vor. Ein wichtiger Satz ist nun das Gauß-Markov-Theorem:

Satz 1.1 *Es gelte (i) der Erwartungswert der Fehler sei gleich Null, d.h.* $\mathbb{E}(e) = \vec{0}$*, und (ii) die Kovarianzen zwischen den Fehlerkomponenten seien gleich Null und die Fehler seien homoszedastisch (haben die gleiche Varianz), d.h.* $Kov(e) = \sigma^2 I$*, I die Einheitsmatrix. Dann gilt*

$$\hat{\boldsymbol{b}} = (X'X)^{-1}X'\boldsymbol{Y} \tag{46}$$

$$Kov(\hat{\boldsymbol{b}}) = \sigma^2(X'X)^{-1} \tag{47}$$

$$s^2 = \frac{1}{n-p}QS_{res} \tag{48}$$

mit $\mathbb{E}(\boldsymbol{b}) = \boldsymbol{b}$*, d.h. die Schätzung ist verzerrungsfrei (biasfrei, keine* systematische *unter- oder Überschätzung), und* $\hat{\boldsymbol{b}}$ *hat minimale Varianz, wobei die Varianz und die Kovarianz der Schätzungen durch (47) gegeben sind (die Schätzungen sind BLUE = Best Linear Unbiased Estimates). Die Schätzung von* σ^2 *ist durch (48) gegeben.*

Beweis: Lehrbücher der mathematischen Statistik.

Die Schätzung $\hat{\mathbf{b}}$ ist die KQ-Schätzung und das Bemerkenswerte an dem Satz ist, dass keine Verteilungsannahmen gemacht werden. Wichtig ist, dass die Schätzung $\hat{\mathbf{b}}$ unverzerrt ist, und natürlich dass sie eine minimale Varianz hat, – vorausgesetzt wird aber eben die Unkorreliertheit der Fehler und ihre Homoszedastizität. Der Satz sagt *nicht* aus, dass die Varianzen der Komponenten von $\hat{\mathbf{b}}$ notwendig klein sind, es wird nur ausgesagt, dass sie unter den genannten Nebenbedingungen die kleinstmöglichen sind. Welche unangenehmen Implikationen die Beziehung $Kov(\hat{\mathbf{b}}) = \sigma^2(X'X)^{-1}$ hat, ist bereits im Abschnitt ber die Multikollinearität angedeutet worden. Der Effekt der Kollinearität kann mit Regularisierungs- oder Penalisierungsmethoden reduziert werden; wie bei der Ridge-Regression betrachtet man dazu nicht die Schätzung (46), sondern

$$\hat{\mathbf{b}}_\lambda = (X'X + \lambda I)^{-1}X'\mathbf{Y}.$$

Damit wird in Abhängigkeit vom Wert von λ die Varianz der Schätzungen reduziert, – aber die Verzerrungsfreiheit für $\hat{\mathbf{b}}_\lambda$ geht verloren. Das Phänomen, mit dem man es nun zu tun bekommt, ist der *Bias-Variance-Tradeoff*[12]

Bias-Variance-Tradeoff Mit $\mathbb{E}$ werde der Erwartungswert, mit Var die Varíanz einer zufälligen Veränderlichen bezeichnet. Es sei zunächst an die Definition des Bias einer Parameterschätzung erinnert. Ist θ ein zu schätzender Parameter und $\hat{\theta}$ eine Schätzung, so ist der Bias $\mathbb{B}$ durch

$$\mathbb{B}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta) = \mathbb{E}(\hat{\theta}) - \theta \tag{49}$$

definiert. Die (bedingte) Varianz ist

$$MSE(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2|\theta) \tag{50}$$

(im Englischen Mean square error (MSE)). Es gilt der

Satz 1.2 *Es sei* $\mu = \mathbb{E}(\hat{\theta})$*,* $Var(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \mu)^2$*. Dann ist*

$$MSE(\hat{\theta}) = \mathbb{B}^2(\hat{\theta}) + Var(\hat{\theta}). \tag{51}$$

Beweis: Zur Vereinfachung wird $MSE(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2)$ für $MSE(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2|\theta)$ geschrieben. Es ist

$$\begin{aligned}
\mathbb{E}(\hat{\theta} - \theta)^2 &= \mathbb{E}(\hat{\theta} - \mu + \mu - \theta)^2 \\
&= \mathbb{E}((\hat{\theta} - \mu)^2 + (\mu - \theta)^2 + 2(\theta - \mu)(\mu - \theta) \\
&= \mathbb{E}(\hat{\theta} - \mu)^2 + \mathbb{E}(\mu - \theta)^2 + 2(\mu - \theta)\mathbb{E}((\hat{\theta} - \mu) \\
&= \mathbb{E}(\hat{\theta} - \mu)^2 + (\mu - \theta)^2 = \mathrm{Var}(\hat{\theta}) + \mathbb{B}^2
\end{aligned}$$

[12]"There is no free lunch", wie es der Ökonom Milton Friedman, einer der geistigen Väter Liberalisierungs-, Deregulierungs- und Steuersenkungsbewegung der vergangenen Jahrzehnte, einst formulierte, – soll heißen, dass man den Preis erhöhter Varianz zahlen muß, wenn man einen kleinen Bias wünscht, und einen erhöhten Bias in Kauf nehmen muß, wenn man eine kleine Varianz wünscht.

denn $(\mu - \theta)$ ist eine Konstante und keine zufällige Veränderliche, und $\mathbb{E}(\hat{\theta} - \mu) = 0$. □

Unter den Bedingungen des Gauß-Markov-Theorems ist also $\mathbb{B}^2 = 0$.

Bias-Variance-Tradeoff und Ridge-Regression Es wird der Bias bei der Ridge-Regression bestimmt. Es gilt der

Satz 1.3 *Es sei $R = Z'Z$ die Matrix der Korrelationen zwischen den Prädiktoren*[13]*. Dann ist*

$$\begin{aligned}
\hat{\boldsymbol{b}}_\lambda &= (I + \lambda R^{-1})^{-1}\hat{\boldsymbol{b}}, && (52)\\
\mathbb{E}(\hat{\boldsymbol{b}}_\lambda) &= (I + \lambda R^{-1})^{-1}\boldsymbol{b} && (53)\\
\mathbb{B}(\hat{\boldsymbol{b}}_\lambda) &= \mathbb{E}(\mathbb{E}(\hat{\boldsymbol{b}}_\lambda) - \boldsymbol{b}) = -\lambda R^{-1}\boldsymbol{b} && (54)\\
Var(\hat{\boldsymbol{b}}_\lambda) &= \sigma^2(R + \lambda I)^{-1}R^{-1}(R + \lambda I)^{-1} && (55)
\end{aligned}$$

Beweis: Es ist

$$\begin{aligned}
\hat{\mathbf{b}}_\lambda &= (Z'Z + \lambda I)^{-1}Z'\mathbf{y} = (R - \lambda I)^{-1}Z'\mathbf{y}\\
&= (R + \lambda I)^{-1}RR^{-1}Z'\mathbf{y}\\
&= (R(I + \lambda R^{-1}))^{-1}R\underbrace{(Z'Z)^{-1}Z'\mathbf{y}}_{\hat{\mathbf{b}}}\\
&= (I + \lambda R^{-1})^{-1}R^{-1}R\hat{\mathbf{b}}\\
&= (I + \lambda R^{-1})^{-1}\hat{\mathbf{b}}.
\end{aligned}$$

wobei $\hat{\mathbf{b}}$ die gewöhnliche (OLS-)KQ-Schätzung von $\mathbf{b}$ ist[14]. Nun ist

$$\mathbb{E}(\hat{\mathbf{b}}_\lambda) = \mathbb{E}((I + \lambda R^{-1})^{-1}\hat{\mathbf{b}}) = (I - \lambda R^{-1})^{-1}\mathbb{E}(\hat{\mathbf{b}}) = (I + \lambda R^{-1})^{-1}\mathbf{b},$$

und dies wurde behauptet.

(54) folgt sofort aus (53). (55) sieht man wie folgt: Es ist

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_\lambda) &= \mathbb{E}(\hat{\theta}_\lambda - \mathbb{E}(\hat{\theta}_\lambda)^2\\
&= \mathbb{E}(W\hat{\theta}_\lambda - W\mathbf{b})^2,\ \ W = (I - \lambda R^{-1})^{-1}\\
&= \mathbb{E}(W(\hat{b}b_\lambda - \mathbf{b})(W(\hat{b}b_\lambda - \mathbf{b}))')\\
&= \mathbb{E}(W(R^{-1}Z'\mathbf{y} - \mathbf{b})(R^{-1}Z'\mathbf{y} - \mathbf{b})')\\
&= \mathbb{E}(WR^{-1}Z'\mathbf{e}\mathbf{e}'ZR^{-1}W')\\
&= \sigma^2WR^{-1}Z'ZR^{-1}W' = \sigma^2WR^{-1}W',
\end{aligned}$$

denn $\mathbb{E}(\mathbf{e}\mathbf{e}') = \sigma^2 I$. □

[13] Die Division durch m wird hier zur Vereinfachung vernachlässigt

[14] zur Erinnerung: $(R(I - \lambda R^{-1}))^{-1} = (I + \lambda R^{-1})^{-1}R^{-1}$ nach der Regel $(AB)^{-1} = B^{-1}A^{-1}$!

Man sieht sofort, dass für $\lambda = 0$ der Bias verschwindet, denn dann ist $(I - \lambda R^{-1})^{-1} = I^{-1} = I$. Für größer werdenden Wert von λ dagegen wird λR^{-1} größer und damit der Bias.

2 Faktorenanalyse und Hauptkomponenten (PCA)

2.1 Faktorenanalyse

Der Vektor $\mathbf{X} = (X_1, \ldots, X_n)'$ heißt *zufälliger Vektor* oder auch einfach *Zufallsvektor*, wenn die Komponenten zufällige Variablen sind. Ist $\mathbb{E}(X_j) = \mu_j$ der Erwartungswert von X_j, so ist

$$\mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_n))' = (\mu_1, \ldots, \mu_n)' \tag{56}$$

Mit $x_j = X_j - \mu_j$ ist

$$\mathbf{x} = (x_1, \ldots, x_n)' \tag{57}$$

der Vektor der Abweichungen vom jeweiligen Erwartungswert. Die Matrix der Varianzen und Kovarianzen der Komponenten von $\mathbf{x}$ ist durch

$$\mathrm{Kov}(\mathbf{X}) = \mathbb{E}(\mathbf{x}\mathbf{x}') = \begin{pmatrix} \mathbb{E}(x_1^2) & \mathbb{E}(x_1x_2) & \cdots & \mathbb{E}(x_1x_n) \\ \mathbb{E}(x_2x_1) & \mathbb{E}(x_2^2) & \cdots & \mathbb{E}(x_2x_n) \\ & & \ddots & \\ \mathbb{E}(x_nx_1) & \mathbb{E}(x_nx_2) & \cdots & \mathbb{E}(x_n^2) \end{pmatrix}. \tag{58}$$

gegeben. die Matrix der Varianzen $\mathbb{E}(x_j^2)$ und Kovarianzen $\mathbb{E}(x_ix_j)$ der Komponenten von $\mathbf{X}$.

Das Modell der Faktorenanalyse ist durch

$$x_{ij} = a_{j1}L_{i1} + a_{j2}L_{i2} + \cdots + a_{js}L_{is} + e_{ij} \tag{59}$$

definiert. Die $L_{i1}, \ldots, L_{is}$ sind zufällige Variablen, ebenso wie der Fehler e_{ij}. Die $a_{j1}, \ldots, a_{js}$ sind "Ladungen". Die L_{i1} die die m Komponenten des Vektors $\mathbf{L}_1$; die übringen $\mathbf{L}_2, \ldots, \mathbf{L}_s$ sind analog definiert. Diese Vektoren sind ebenfalls zufällige Vektoren, weil die Stichpribe der Fälle – Personen oder Objekte – eben zufällig ist.

Es werden die folgenden Annahmen gemacht:

$$\mathbb{E}(\mathbf{L}) = \vec{0} \tag{60}$$

$$\mathbb{E}(\mathbf{e}) = \vec{0} \tag{61}$$

$$Kov(\mathbf{L}_j) = \mathbb{E}(\mathbf{L}_j\mathbf{L}_j') \tag{62}$$

$$Kov(\mathbf{e}) = \mathbb{E}(\mathbf{e}\mathbf{e}') = \Psi = \mathrm{diag}(\psi_1^2, \ldots, \psi_n^2) \tag{63}$$

$$Kov(\mathbf{e}, \mathbf{L}) = \mathbb{E}(\mathbf{e}\mathbf{L}') = 0 \tag{64}$$

Nach (63) sollen die "Fehler" (auch: "spezifische Faktoren") unkorreliert sein, und nach (64) sind Fehler und Faktoren unkorreliert. Die ψ_k^2, $k = 1, \ldots, n$ heißen auch *Restvarianzen*. Die Ladungen a_{jk} werden hzu einer Matrix A zusammengefasst, und man hat dann

$$\mathbf{x} = A\mathbf{L} + \mathbf{e}. \tag{65}$$

Für die Varianz-Kovarianzstruktur ergibt sich

$$\mathbf{x}\mathbf{x}' = (A\mathbf{L} + \mathbf{e})(A\mathbf{L} + \mathbf{e})' = A\mathbf{L}\mathbf{L}'A' + \mathbf{e}A'\mathbf{L}' + A\mathbf{L}\mathbf{e}' + \mathbf{e}\mathbf{e}'. \tag{66}$$

Für die Varianz-Kovarianz-Matrix Σ erhält man unter Berücksichtigung der Annahmen

$$\Sigma := Kov(\mathbf{x}) = A\mathbb{E}(\mathbf{L}\mathbf{L}')A' + \mathbb{E}(\mathbf{e}\mathbf{L}')A' + A\mathbb{E}(\mathbf{L}\mathbf{e}) + \mathbb{E}(\mathbf{e}\mathbf{e}')$$

$$\Sigma = AA' + \Psi; \tag{67}$$

diese Gleilchung gilt nach Thurstone (1935) als das *Fundamentaltheorem* der Faktorenanalyse. Für die Kovarianz zwischen der j-ten und der k-ten Variablen findet man demnach

$$c_{jk} = \sum_{r=1}^{s} a_{jr}a_{kr} + \begin{cases} \psi_j^2, & j = k \\ 0, & j \neq k \end{cases} \tag{68}$$

Die Ladungen a_{jk}, also die Elemente von A, können als Kovarianzen zwischen den Variablen V_j und den latenten Faktoren interpretiert werden:

$$\begin{aligned} Kov(\mathbf{x}, \mathbf{L}) &= \mathbb{E}(\mathbf{x}\mathbf{L}') = \mathbb{E}[(A\mathbf{L} - \mathbf{e})\mathbf{L}'] \\ &= \mathbb{E}(A\mathbf{L}\mathbf{L}' + \mathbf{e}\mathbf{L}') = A\mathbb{E}(\mathbf{L}\mathbf{L}') + \mathbf{e}(\mathbf{e}\mathbf{L}') \\ &= A \end{aligned} \tag{69}$$

Die Diagnonalzellen von $Kov(\mathbf{x})$ enthalten die Varianzen der X_j. Man erhält

$$\sigma_j^2 = c_j j = a_{j1}^2 + a_{j2}^2 + \cdots + a_{js}^2 + \psi_j^2. \tag{70}$$

Von Interesse ist die Teilsumme $a_{j1}^2 + a_{j2}^2 + \cdots + a_{js}^2$:

$$h^2 := a_{j1}^2 + a_{j2}^2 + \cdots + a_{js}^2 \tag{71}$$

heißt *Kommunalität*. h^2 ist der Teil der Varianz σ_j^2, der durch die latenten Variablen zustande kommt, die in *alle* gemessenen Variablen eingehen. h^2 ist natürlich nicht bekannt, da zunächst die Anzahl der latenten Faktoren nicht bekannt ist, d.h. h^2 muß aus den Daten geschätzt werden. Das Ziel der Faktorenanalyse ist die Analyse bzw. die Erklärung der Kovarianzen zwischen den Variablen; dieses Ziel unterscheidet sie von der PCA, die die Varianzen in den Daten erklären will. Es zeigt sich aber, dass die Faktorenanalyse durch eine PCA approximiert werden kann.

2.2 Die Hauptachsentransformation (PCA)

Das Prinzip der PCA wurde bereits im Skriptum *Vektoren und Matrizen in der Multivariaten Analyse* (VMMVA) vorgestellt, so dass hier nur der Kern des Ansatzes noch einmal präsentiert wird.

Gegeben sei eine (m, n)-Datenmatrix X (m Zeilen, n Spalten). Die Spalten repräsentieren Variablen $V_1, \ldots, V_n$, die Zeilen Fälle (Personen, Objekte, Zeitpunkte). Wie in VMMVA gezeigt wurde, läßt sich für jede Matrix die Singularwertzerlegung (SVD)

$$X = Q\Lambda^{1/2}P' \tag{72}$$

angeben, wobei Q die orthonormale (m, n)-Matrix der Eigenvektoren von XX' ist, P ist die (n, n)-Matrix der Eigenvektoren von $X'X$, und Λ ist die (n, n)-Diagonalmatrix der Eigenwerte (die von Null verschiedenen Eigenwerte sind für XX' und $X'X$ identisch).

Es ist sinnvoll, die Spaltenvektoren von X zu zentrieren, d.h. die Vektoren

$$\mathbf{x}_j = \mathbf{X}_j - \bar{x}_j\vec{1}, \quad j = 1, \ldots, n \tag{73}$$

zu betrachten; falls die V_j in verschiedenen Maßeinheiten gemessen werden, ist es darüber hinaus sinnvoll, die Messwerte zu standardisieren:

$$z_{ij} = \frac{X_{ij} - \bar{x}_j}{s_j},\ i = 1, \ldots, m; j = 1, \ldots, n \tag{74}$$

Statt der Matrix X wird dann die Matrix $Z = (z_{ij})$ analysiert. Bei der Zerlegung

$$Z = Q\Lambda^{1/2}P' \tag{75}$$

sind die Matrizen Q, Λ, P nicht mit den entsprechenden Matrizen in (72) identisch!

Für die x_{ij}-Werte liefert die SVD $X = Q\Lambda^{1/2}P'$ eine Zerlegung, die dem faktorenanalytischen Ansatz (59) bis auf die Fehler formal entspricht:

$$x_{ij} = a_{j1}q_{i1} + a_{j2}q_{i2} + \cdots + a_{jn}q_{in}, \tag{76}$$

wobei angenommen wird, dass die Matrix X den Rang $r = n = \min(m, n)$ hat. Man kann eine Rang $r < n$ annehmen und die Restsumme

$$a_{j,r+1}q_{i,r+1} + \cdots + a_{jn}q_{in}$$

mit dem Fehler e_{ij} gleichsetzen. Das Wesentliche ist aber die Rotation der Punktekonfiguration in ein Achsensystem, das neue latente Variable postuliert, die nicht miteinander korrelieren. Die Varianz der Koordinaten auf der ersten Achse ist

dann die maximal mögliche Varianz, die auf der dazu orthogonalen zweiten Achse ist die zweitgrößte Varianz, etc. Davon abweichende Rotationen repräsentieren dann notwendigerweise miteinander korrelierende latente Variable.

Zur Entscheidung über die zu berücksichtigende Anzahl latenter Dimensionen oder Variablen wird im Allgemeinen der Scree-Test verwendet. SCREE-TEST

Wie in VMMVA in Abschnitt 3.9 gezeigt wurde, bedeutet

$$ZP = L = Q\Lambda^{1/2} \tag{77}$$

den Übergang zu einem achsenparallelen Ellipsoid: die m-dimensionalen Spaltenvektoren von L sind orthogonal, die Korrelationen zwischen den "latenten" Variablen $\mathbf{L}_j$ und $\mathbf{L}_k$ sind also gleich Null für alle $j, k = 1, \ldots, n,\ j \neq k$. Ob zentriert oder spaltenstandardisiert: die Varianz λ_1 der Komponenten von $\mathbf{L}_1$ ist stets maximal, die von $\mathbf{L}_2$ ist die zweitgrößte, etc.:

$$\lambda_1 = \mathbf{L}_1'\mathbf{L}_1 = \|\mathbf{L}_1\|^2 \tag{78}$$

ist dann die maximal mögliche Varianz der Koordinaten der Fälle, allgemein ist $\lambda_j = \|\mathbf{L}_j\|^2$ die Varianz der Koordinaten auf der j-ten latenten Variablen (vergl. auch Satz 3.5, Abschnitt 3.11 in VMMVA).

Im Allgemeinen ist man an der Struktur der Variablen interessiert. Statt $L = Q\Lambda^{1/2}$ betrachtet man dann die *Ladungen*

$$A = P\Lambda^{1/2}, \tag{79}$$

so dass $Z = QA'$. a_{jk} ist die Ladung der j-ten Variablen auf der k-ten latenten Variablen. Es folgt

$$Z'Z = P\lambda P' = \sum_{j=1}^{n} \lambda_j \mathbf{P}_j \mathbf{P}_j', \tag{80}$$

insbesondere

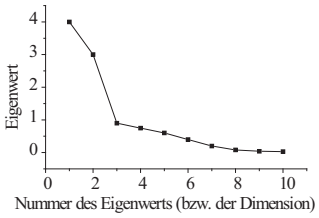
$$nr_{jk} = \sum_{s=1}^{n} a_{js}a_{ks}, \tag{81}$$

und wegen $r_{jj} = 1$ für alle j gilt insbesondere

$$nr_{jj} = \sum_{s=1}^{n} a_{js}^2 = 1. \tag{82}$$

Genügt eine 2-dimensionale Lösung, so bedeutet diese Gleichung, dass die Endpunkte der 2-dimensionalen Vektoren, die die Variablen repräsentieren, alle auf einem Kreis liegen. Im 3-dimensionalen Fall liegen die Endpunkte alle auf einer Kugel, etc.

Abbildung 2: Scree-Test



Abschätzung der Anzahl latenter Dimensionen Es gibt eine Reihe von Möglichkeiten, die Anzahl der latenten Dimensionen abzuschätzen. Eine der bekanntesten und einfachsten ist der Scree-Test (vergl. Abb. 2). Er basiert auf der Tatsache, dass die Eigenwerte $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_1$ jeweils zu den Varianzanteilen, die durch eine latente Dimension erklärt werden, korrespondieren. "Grosse" Eigenwerte, legen dann die Existenz einer latenten Variablen nahe, während kleine nur zufällige Effekte reflektieren. Wenn man Glück hat, gibt es zwischen einem "großen" und den "kleinen" Eigenwerten einen deutlichen Unterschied, der nahelegt, dass die Anzahl der großen Eigenwerte gleich der Anzahl der latenten Variablen ist. In staark verrauschten Daten hat man allerdings oft einen allmählichen Übergang von den großen zu den kleinen Eigenwerten, so dass auch die Entscheidung für eine bestimmte Anzahl von latenten Variablen eine zufällige Komponente enthält.

Eine andere bzw. zusätzliche Möglichkeit der Entscheidung für einen bestimmten r-Wert kann man daraus ableiten, dass für den Fall $r < n$ die Endpunkte der Merkmalsvektoren auf einer r-dimensionalen Hyperkugel liegen. Hat man also

$$\sum_{s=1}^{r} a_{js} \approx 1 \text{ für alle } j \tag{83}$$

so ist dies ein Hinweis, dass r die Anzahl der latenten Variablen ist.

Beliebt ist Kaiser's Kriterium, demzufolge Eigenwerte $\lambda_k < 1$ nur zufällige Effekte anzeigen. Man sollte dieses Kriterium mit Vorsicht anwenden (Cliff's Argument).

Die PCA als Modell Die PCA wird oft – im Gegensatz zur Faktorenanalyse – als ein rein deskriptives Verfahren angesehen. Ein Grund für diese Interpretation mag darin liegen, dass die SVD $X = Q\Lambda^{1/2}P'$ für *jede* Matrix X berechnet werden kann, d.h. es kann nicht vorkommen, dass die Berechnung nicht möglich ist, wenn irgendeine Annahme nicht gilt. Ob die Beschreibung in Termen der Matrizen Q, Λ und P eine sinnvolle Interpretation zuläßt, ist eine andere Frage.

Es ist allerdings nicht klar, warum die PCA *nicht auch* als ein Modell angesehen werden kann: die Unterscheidung zwischen Theorie (hier: Modell) und Deskription ist ohnehin fließend, wie auch aus allgemeinen wissenschaftstheoretischen Betrachtungen hervorgeht. Die SVD läßt sich mit der Hypothese kombinieren, dass die Variablen durch die Matrix P bzw. durch $A = P\Lambda^{1/2}$ repräsentiert werden können unabhängig von den Fällen, d.h. unabhängig von der Matrix Q, die eben die Fälle, also etwa Personen, abbilde. Diese Annahme läßt sich testen: die Ladungen der Variablen sollten für verschiedene Stichproben von Fällen in guter Näherung gleich sein. Das folgende Beispiel illustriert einen solchen Ansatz:

Beispiel 2.1 Raterobjektivität und Spezifische Objektivität Die Beurteilerübereinstimmung kann als Korrelation zwischen den Einschätzungen durch verschiedene Urteiler bestimmt werden. Gegeben seien Daten wie in der Tabelle 1. Das Ausmaß der Übereinstimmung sollte sich in den Korrelationen zwischen

Tabelle 1: Einschätzungen von n Personen bzw. Objekten ω_i durch K Experten

Person	Experten			
ω_i	R_1	R_2	$\cdots$	R_k
1	x_{11}	x_{12}	$\cdots$	x_{1K}
2	x_{21}	x_{22}	$\cdots$	x_{2K}
$\vdots$			$\vdots$	
n	x_{n1}	x_{n2}	$\cdots$	x_{nK}

ihren Bewertungen ausdrücken. Die Matrix der Korrelationen sei R. Da R symmetrisch ist, gilt $R = P\Lambda P'$, P die orthonormale Matrix der Eigenvektoren von R. Im Idealfall korrelieren die Ratings perfekt, so dass $r_{jk} = 1$ für alle j, k. Für die Spaltenvektoren der (spaltenstandardisierten) Matrix Z gilt dann $\mathbf{z}_j = \mathbf{z}_k$ für alle j, k. Generell gilt $ZP = L$, L die Matrix der latenten Vektoren. Insbesondere für $\mathbf{L}_1$ gilt dann

$$\mathbf{z}_1 p_{11} + \mathbf{z}_2 p_{21} + \cdots + \mathbf{z}_n p_{n1} = \mathbf{z}_1 \alpha_1 = \mathbf{L}_1,$$

und für $\mathbf{L}_2$ findet man analog

$$\mathbf{z}_1 p_{12} + \mathbf{z}_2 p_{22} + \cdots + \mathbf{z}_n p_{n2} = \mathbf{z}_1 \alpha_2 = \mathbf{L}_2,$$

da ja $\mathbf{z}_j$ alle identisch sind. Das heißt aber, dass $\mathbf{L}_1$ und $\mathbf{L}_2$ dieselbe Orientierung wie $\mathbf{z}_1$ haben, mithin können $\mathbf{L}_1$ und $\mathbf{L}_2$ nicht orthogonal und damit nicht linear unabhängig sein, woraus folgt, dass $\alpha_2 = 0$ sein muß. Es gibt also nur eine latente Dimension. Umgekehrt folgt aus der Eindimensionalität, dass alle Korrelationen

gleich 1 sind. Denn dann ist $\mathbf{z}_j = \mathbf{LP} = p\mathbf{L}$, denn $\mathbf{P}$ kann nur aus einer Komponente – p – bestehen (s. oben). Dann folgt unmittelbar, dass $\mathbf{z}_j'\mathbf{z}_k = p^2\|\mathbf{L}\|^2 = m$, so dass $p = m/\|\mathbf{L}\|$. Natürlich gibt es keine fehlerfreien Ratings, so dass der Fall $\mathbf{z}_1 = \cdots = \mathbf{z}_n$ nur angenähert gelten kann, aber hohe Korrelationen legen den Fall der Eindimensionalität nahe. Dies bedeutet, dass die verschiedenen Rater die beurteilten Eigenschaften in (angenähert) gleicher Weise gewichten. Eine ungleiche Gewichtung der Merkmale bewirkt eine Reduktion der Korrelationen und damit eine reduzierte Reliabilität der Rater.

Es sei Z die Matrix der standardisierten Ratings aus der Tabelle 1. Die Singularwertzerlegung dieser Matrix sei $Z = Q\Lambda^{1/2}P'$. Bei der Berechnung der Korrelationen $Z'Z/m$ kürzt sich die Matrix Q der Personenscores wegen $Q'Q = I$ heraus:

$$Z'Z = P\Lambda^{1/2}Q'Q\Lambda^{1/2}P' = P\Lambda P'.$$

Bei "objektiven" Ratern bleibt die Matrix P invariant gegenüber verschiedenen Stichproben von zu beurteilenden Personen, d.h. P sollte stichprobeninvariant sein. Aber es gilt auch

$$Z'Q'\Lambda^{-1} = P,$$

so dass die Invarianz sofort getestet werden kann, wenn man Q und Λ als Eigenvektoren und Eigenwerte von ZZ' berechnet.

2.3 PCA-Regression

Es soll noch einmal die multiple Regression (7), also

$$\mathbf{Y} = X\mathbf{b} + \mathbf{e}$$

betrachtet werden. Wie bereits diskutiert wurde, bedeutet die Kollinearität der Prädiktorvariablen, dass die Schätzung $\hat{\mathbf{b}}$ für $\mathbf{b}$ schwer zu interpretieren ist. Eine Möglichkeit, diesen Schwierigkeiten zu entgehen, besteht in der PCA-Regression. DAbei substituiert man zunächst die SVD $X = Q\Lambda^{1/2}P'$ für X:

$$\mathbf{Y} = X\mathbf{b} + \mathbf{e} = Q\Lambda^{1/2}P'\mathbf{b} + \mathbf{e}. \tag{84}$$

Die Spaltenvektoren $\mathbf{x}_j$ von X werden hierbei durch die orthonormalen Vektoren in Q bzw in $L = Q\Lambda^{1/2}$ ersetzt, und der Vektor der Regressionsgewichte ist nun

$$\mathbf{b}_0 = \Lambda^{1/2}P'\mathbf{b} \text{ bzw. } \mathbf{b}_0 = P'\mathbf{b}. \tag{85}$$

Es werde insbesondere der erste Fall angenommen. Dann ist

$$\mathbf{Y} = Q\mathbf{b}_0 + \mathbf{e}. \tag{86}$$

Die KQ-Schätzung für $\mathbf{b}_0$ ist dann

$$\hat{\mathbf{b}}_0 = (Q'Q)^{-1}Q'\mathbf{Y} = Q'\mathbf{Y}, \tag{87}$$

da ja $Q'Q) = I$ die Einheitsmatrix ist. Es folgt

$$\hat{\mathbf{b}}_0 = Q'(Q\mathbf{b}_0 + \mathbf{e}) = \mathbf{b}_0 + Q'\mathbf{e}, \tag{88}$$

d.h. $\hat{\mathbf{b}}_0$ setzt sich aus dem "wahren" Wert $\mathbf{b}_0$ und einem Fehlervektor $Q'\mathbf{e}$ zusammen. Man erhält

$$\hat{\mathbf{b}}_0 - \mathbf{b}_0 = Q'\mathbf{e}. \tag{89}$$

Für den Erwartungswert dieser Differenz findet man

$$\mathbb{E}(\hat{\mathbf{b}}_0 - \mathbf{b}_0) = 0, \tag{90}$$

d.h. die Schätzung ist erwartungstreu ("biasfrei"). Für die Varianz der Schätzungen erhält man wegen (89)

$$\begin{aligned} Var(\hat{\mathbf{b}}_0) &= \mathbb{E}(\hat{\mathbf{b}}_0 - \mathbf{b}_0)^2 = \mathbb{E}(Q'\mathbf{e})^2 \\ &= \mathbb{E}\left(\sum_i \sum_{i'} q_{ij} q_{i'j} e_i e_{i'}\right) = \left(\sum_{ij} q_{ij}^2\right) \sigma^2 = \sigma^2 \end{aligned} \tag{91}$$

also

$$Var(\hat{\mathbf{b}}_0) = \sigma^2 \tag{92}$$

Wegen (85) ist

$$\hat{\mathbf{b}}0\Lambda^{1/2}P'\hat{\mathbf{b}} \tag{93}$$

Man kann daraus die Schätzung $\hat{\mathbf{b}}$ der ursprünglichen Regressionskoeffizienten erhalten:

$$\hat{\mathbf{b}} = P\Lambda^{-1/2}\hat{\mathbf{b}}_0. \tag{94}$$

$P\Lambda^{-1/2}$ enthält die Linearkombinationen der Spaltenvektoren $\mathbf{P}_j$ von P mit den jeweiligen Gewichten $1/\sqrt{\lambda_j}$, so dass man (94) in der Form

$$\hat{\mathbf{b}} = \left(\sum_{j=1}^{n} \frac{\mathbf{P}_j}{\sqrt{\lambda_j}}\right) \hat{\mathbf{b}}_0 \tag{95}$$

schreiben kann. Man sieht wieder, dass kleine λ_j-Werte den Wert der Komponenten von $\hat{\mathbf{b}}$ aufblähen.