Grundbegriffe der multivariaten Datenanalyse

Johannes Andres

Bei Fragestellungen, die die Beziehungen zwischen mehreren Variablen zum Thema haben, wird bei der statistischen Analyse oft von Konzepten der linearen Algebra Gebrauch gemacht. Der vorliegende Artikel soll zunächst in knapper Form die wichtigsten Begriffe und Tatsachen aus der Vektor- und Matrizenrechnung einführen. Danach soll die Fruchtbarkeit der mathematischen Begriffsbildungen für die Datenaufbereitung an der Hauptkomponentenanalyse gezeigt werden. Schließlich wird kurz auf mehrdimensionale wahrscheinlichkeitstheoretische Konzepte eingegangen.

1 Wichtige Begriffe der linearen Algebra

Zu Beginn sollen zur Motivation mit *Variablen-* und *Personenraum* zwei geometrische Darstellungsweisen multivariater Daten erläutert werden. Als Beispiel seien an drei Personen zwei Variablen X_1 und X_2 erhoben worden, z.B. Ergebnisse aus zwei Untertests eines Intelligenztests. Üblicherweise schreibt man solche Datenmengen geordnet in Form einer sogenannten *Datenmatrix*, in der für jede Person eine Zeile und für jede Variable eine Spalte vorgesehen ist. Hier könnte die Datenmatrix folgendermaßen aussehen:

$$\begin{pmatrix} 1 & -2 \\ 2 & 1 \\ -3 & 1 \end{pmatrix}$$

Die erste Person hat also die Werte 1 und -2 in den beiden Tests erzielt, in der zweiten Variablen hatten die drei Personen die Werte -2, 1 und 1, etc. Eine Möglichkeit, solche Daten geometrisch darzustellen, besteht darin, für jede Person einen Punkt in einem Koordinatensystem einzutragen, dessen Achsen den Variablen entsprechen. Man nennt dies die Darstellung im *Variablenraum*. Als „duale" Möglichkeit bietet sich an, jeder Person eine Achse zuzuordnen und die Variablen als Punkte oder noch besser als „Vektoren" im so entstehenden *Personenraum* einzutragen (der Begriff ist dann etwas irreführend, wenn die den Zeilen der Datenmatrix entsprechenden experimentellen Einheiten nicht Personen sind). Für die Daten des Beispiels sind die beiden Möglichkeiten einander in Abbildung 1 gegenübergestellt.

Die Darstellung der Variablen (genauer der Werte der Variablen in der gegebenen Stichprobe) als Vektoren bietet viele Vorteile bei der Veranschaulichung komplexerer Zusammenhänge, von denen einige im folgenden beschrieben werden sollen. Dazu sollen zunächst wichtige Begriffe der *Vektorrechnung* eingeführt werden.

Etwas vereinfachend sei ein Vektor ein n-Tupel von reellen Zahlen, die entweder untereinander oder nebeneinander notiert werden. Im ersten Fall spricht man

https://doi.org/10.25521/HQM14

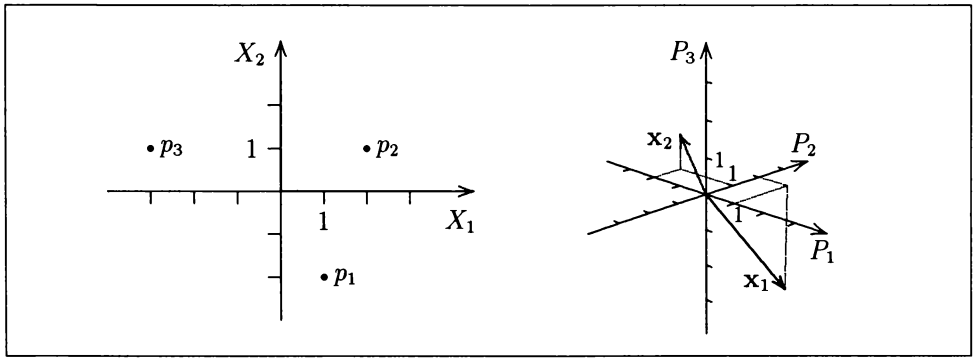

ABBILDUNG 1. **Darstellung von Daten im Variablenraum und im Personenraum.**

von einem Spaltenvektor, im zweiten von einem Zeilenvektor. Abkürzend sei die Sprechweise n-Vektor vereinbart, um die Anzahl der „Komponenten" des Vektors zu bezeichnen. Die Datenmatrix des Beispiels kann man sich also zusammengesetzt denken aus drei den Personen entsprechenden 2-Vektoren (Zeilenvektoren) oder zwei den Variablen entsprechenden 3-Vektoren (Spaltenvektoren). Ein Zeilen-n-Vektor kann in der Form $(x_1, \ldots, x_n)$ notiert werden, während für einen Spalten-n-Vektor oft die platzsparende Schreibweise $(x_1, \ldots, x_n)'$ gewählt wird, die andeutet, daß die Zahlen eigentlich untereinander zu schreiben sind. Wenn nichts weiter vermerkt ist, ist mit dem Wort „Vektor" im folgenden stets ein Spalten-n-Vektor gemeint, wobei n eine feste Zahl sei; die Aussagen bis zur Einführung der Matrizen gelten jedoch analog immer auch für Zeilenvektoren. Geometrisch kann man sich einen Vektor repräsentiert denken als einen Pfeil in einem n-dimensionalen Koordinatensystem, der seinen Ursprung im Nullpunkt $\mathbf{0}$ $(:=(0, \ldots, 0)')$ hat und mit der Spitze in dem Punkt endet, dessen Koordinaten die Komponenten des Vektors sind.

Die Menge aller n-Vektoren sei mit $\mathbb{R}^n$ bezeichnet ($\mathbb{R}$ steht für die Menge der reellen Zahlen). Wichtige Operationen mit Vektoren sind die skalare Multiplikation und die Addition. Bei der skalaren Multiplikation wird eine reelle Zahl a mit einem Vektor $\mathbf{x}$ „multipliziert", das Ergebnis $a\mathbf{x}$ ist ein neuer Vektor, dessen Komponenten die Produkte der Komponenten von $\mathbf{x}$ mit a sind (also $a(x_1, \ldots, x_n)' = (ax_1, \ldots, ax_n)'$). Die Summe zweier Vektoren $\mathbf{x}$ und $\mathbf{y}$ ist der Vektor $\mathbf{x} + \mathbf{y}$, dessen Komponenten die Summen der entsprechenden Komponenten von $\mathbf{x}$ und $\mathbf{y}$ sind (also $(x_1, \ldots, x_n)' + (y_1, \ldots, y_n)' = (x_1 + y_1, \ldots, x_n + y_n)'$). Die Differenz $\mathbf{x} - \mathbf{y}$ zweier Vektoren ist gleich $\mathbf{x} + (-1)\mathbf{y}$. Durch diese beiden Operationen wird der $\mathbb{R}^n$ zu einem (reellen) Vektorraum. Geometrisch entspricht der skalaren Multiplikation die „Streckung" des Vektors um den Faktor a (bei negativem a ist noch am Nullpunkt zu spiegeln) und der Addition das „Aneinanderlegen" von Vektoren (der eine Vektor wird so verschoben, daß sein Ursprung in der Spitze des andern zu liegen kommt; das Ergebnis ist dann der Vektor von $\mathbf{0}$ zur Spitze des verschobenen Vektors).

Eine Summe der Form $\sum_{i=1}^{m} a_i \mathbf{x}_i$ heißt auch *Linearkombination* der Vektoren $\mathbf{x}_1, \ldots, \mathbf{x}_m$ mit Koeffizienten a_i. Linearkombinationen erhält man also durch mehrfache Anwendung der genannten Operationen. Entsprechen Vektoren im Personenraum den in einer Stichprobe erhobenen Variablen, so entsprechen Linearkombinationen dieser Vektoren den in gleicher Weise gebildeten Linearkombinationen der

Variablen, die eine eigene Bedeutung haben können, wie die der Summe der Testergebnisse der Untertests eines Intelligenztests oder der Differenz der Ergebnisse einer Variablen vor und nach einem *treatment* (zur etwas anderen Verwendung des Ausdrucks „Linearkombination" bei Variablen vgl. die Ausführungen weiter unten).

Zentrale Begriffe der Vektorrechnung sind die der *linearen Unabhängigkeit* und der *linearen Abhängigkeit.* Ein System $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ von Vektoren heißt linear abhängig, wenn es möglich ist, den Nullvektor $\mathbf{0}$ als eine Linearkombination der $\mathbf{x}_i$ darzustellen, bei der nicht alle Koeffizienten gleich 0 sind. Ist dies nicht möglich, so heißen die Vektoren linear unabhängig. Bei mindestens zwei linear abhängigen Vektoren gibt es dann übrigens mindestens einen, der als Linearkombination der restlichen geschrieben werden kann. Der *Rang* eines Systems von Vektoren ist definiert als die maximale Anzahl von linear unabhängigen Vektoren in dem System (der Vollständigkeit halber: Der Rang eines Systems von Nullvektoren ist Null). Anschaulich gesprochen läßt man in dem System solange Vektoren weg, die Linearkombinationen der übrigbleibenden sind, bis die zuletzt noch übrigen Vektoren linear unabhängig sind. Die Anzahl dieser Vektoren ist dann der Rang.

Eine Teilmenge $V \neq \emptyset$ des $\mathbb{R}^n$, die mit einem Vektor auch jedes skalare Vielfache dieses Vektors und mit zwei Vektoren auch deren Summe enthält, heißt *(linearer) Unterraum.* Enthält ein Unterraum ein System von Vektoren, so enthält er auch jede Linearkombination dieser Vektoren. Darüber hinaus bildet die Menge aller Linearkombinationen eines Systems von Vektoren stets einen Unterraum, der als der von diesem System erzeugte Unterraum bezeichnet wird. Für jeden Unterraum kann man ein linear unabhängiges System von Vektoren finden, das ihn erzeugt, und es stellt sich heraus, daß alle derartigen Systeme gleich viele Vektoren enthalten. Diese Anzahl nennt man dann die *Dimension* des Unterraums. Beipiele von Unterräumen sind der nulldimensionale Unterraum $\{\mathbf{0}\}$, (eindimensionale) Geraden und (zweidimensionale) Ebenen durch den Nullpunkt oder auch $\mathbb{R}^n$ selber, dessen Dimension n ist. Der Rang eines Systems von Vektoren erhält eine anschauliche Deutung dadurch, daß er gleich der Dimension des durch das System erzeugten Unterraums ist.

Hat man zu einem Unterraum ein erzeugendes linear unabhängiges System von Vektoren gefunden, so läßt sich jeder Vektor des Unterraums in genau einer Weise als Linearkombination der Vektoren des Systems schreiben. Ein solches System nennt man auch *Basis* des Unterraums. Die Koeffizienten der Darstellung eines Vektors als Linearkombination der Basisvektoren lassen sich als Koordinaten des Vektors in dem neuen Koordinatensystem deuten, dessen Achsen durch $\mathbf{0}$ in Richtung der Basisvektoren laufen und dessen Einheiten an den Spitzen dieser Vektoren abgetragen sind. In diesem Sinne entsprechen die Basen des $\mathbb{R}^n$ selbst gerade den möglichen alternativen Koordinatensystemen mit gleichem Nullpunkt. Eine spezielle Basis ist die der sogenannten *Einheitsvektoren*; dabei ist der i-te Einheitsvektor der Vektor, der an der i-ten Stelle eine Eins hat und sonst nur aus Nullen besteht.

Neben linearen Unterräumen spielen oft auch sogenannte *affine Unterräume* eine wichtige Rolle. Einen affinen Unterraum erhält man dadurch, daß man einen linearen Unterraum in einer festen Richtung um einen festen Betrag verschiebt oder, anders ausgedrückt, indem man zu allen seinen Punkten den gleichen Vektor hinzuaddiert. Beispiele sind Punkte sowie Geraden und Ebenen, die nun nicht mehr unbedingt durch den Nullpunkt gehen müssen. Die Dimension eines affinen Unterraums ist die

Dimension des linearen Unterraums, aus dem er durch Verschiebung hervorgeht.

Je zwei Vektoren des $\mathbb{R}^n$ kann man noch auf eine weitere Art verknüpfen, nämlich durch das *Skalarprodukt*, das zwei Vektoren $\mathbf{x} = (x_1, \ldots, x_n)'$ und $\mathbf{y} = (y_1, \ldots, y_n)'$ die Zahl $\langle \mathbf{x}, \mathbf{y} \rangle := \sum x_i y_i$ zuordnet. Durch iterierte Anwendung des Satzes von Pythagoras erkennt man, daß $|\mathbf{x}|^2 := \langle \mathbf{x}, \mathbf{x} \rangle = \sum x_i^2$ die quadrierte Länge des Vektors $\mathbf{x}$ ist (wenn man, was vorausgesetzt sei, das Koordinatensystem so gewählt hat, daß die Achsen senkrecht aufeinanderstehen und die Einheiten aller Achsen gleich sind). Auch das Skalarprodukt hat eine geometrische Bedeutung: Die Zahl $\langle \mathbf{x}, \mathbf{y} \rangle / (|\mathbf{x}||\mathbf{y}|)$ ist der Kosinus des zwischen den Vektoren eingeschlossenen Winkels. Insbesondere stehen $\mathbf{x}$ und $\mathbf{y}$ genau dann senkrecht aufeinander, wenn $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ ist. Ein System von Vektoren der Länge 1, die paarweise senkrecht sind, nennt man auch *orthonormal*. Eine Basis aus paarweise senkrechten Vektoren der Länge 1 heißt *Orthonormalbasis*. Bei der Darstellung von Daten im Personenraum haben diese Begriffe auch eine statistische Bedeutung: Sind die den Variablen entsprechenden Spalten zentriert, d.h. haben sie Mittelwert 0 (wie im Beispiel), so ist, jeweils bis auf einen Faktor $1/n$, die quadrierte Länge eines Vektors gleich der Varianz der entsprechenden Variablen in der Stichprobe und das Skalarprodukt zweier Vektoren gleich der Kovarianz der zugehörigen Variablen. Die Korrelation ist dann der Kosinus des Winkels zwischen den Vektoren. Diese Tatsachen sind oft hilfreich für das Verständnis komplexerer Zusammenhänge wie z.B. von Eigenschaften der Partialkorrelation.

Es folgen nun wichtige Grundbegriffe der *Matrizenrechnung*. Eine *Matrix* besteht aus in einem rechteckigen Schema angeordneten Zahlen. Eine Matrix mit n Zeilen und m Spalten heißt $(n \times m)$-Matrix; die Datenmatrix des Beispiels ist also eine (3×2)-Matrix. Matrizen sollen mit fetten Großbuchstaben bezeichnet werden, für ihre Elemente wird dann meist der entsprechende Kleinbuchstabe mit einem Doppelindex aus Zeilen- und Spaltennummer verwendet. Wenn z.B. $\mathbf{X}$ die Datenmatrix des Beispiels ist, so heißt das Element -2 in der ersten Zeile und zweiten Spalte x_{12}; zur Einführung dieser Notation dient die Kurzschreibweise $\mathbf{X} = (x_{ij})$. Vektoren können als spezielle Matrizen aufgefaßt werden, nämlich Spalten-n-Vektoren als $(n \times 1)$-Matrizen und Zeilen-n-Vektoren als $(1 \times n)$-Matrizen.

Wie Vektoren können Matrizen mit gleicher Zeilen- und Spaltenzahl (elementweise) addiert werden, und das Produkt einer Zahl mit einer Matrix ist genau wie bei Vektoren definiert. Darüber hinaus gibt es bei Matrizen eine Multiplikation: Ist $\mathbf{A}$ eine $(n \times l)$-Matrix und $\mathbf{B}$ eine $(l \times m)$-Matrix, so ist das Produkt $\mathbf{AB}$ diejenige $(n \times m)$-Matrix $\mathbf{C} = (c_{ij})$ mit $c_{ij} = \sum_{k=1}^{l} a_{ik} b_{kj}$. Die Multiplikation ist also nur definiert, wenn die Spaltenzahl der ersten Matrix mit der Zeilenzahl der zweiten übereinstimmt, wofür die abkürzende Sprechweise vereinbart sei, daß die Matrizen passende Größe haben. Nützlich ist die Merkregel, nach der das Element c_{ij} dadurch entsteht, daß die i-te Zeile von $\mathbf{A}$ mit der j-ten Spalte von $\mathbf{B}$ „multipliziert" wird.

Die Matrizenmultiplikation ist assoziativ, es gilt also für Matrizen $\mathbf{A}, \mathbf{B}, \mathbf{C}$ passender Größe die Gleichung $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$. Ebenso gelten bei jeweils passender Größe die Distributivgesetze $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$ und $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{AC}$. Hingegen ist die Matrizenmultiplikation im allgemeinen nicht kommutativ. Selbst wenn die Matrizen passende Größe haben, so daß beide Produkte definiert sind, ist also im allgemeinen $\mathbf{AB} \neq \mathbf{BA}$.

Eine weitere Operation für Matrizen ist das *Transponieren*: Für eine $(n \times m)$-

Matrix $\mathbf{A} = (a_{ij})$ ist die transponierte Matrix $\mathbf{A}'$ diejenige $(m \times n)$-Matrix, die in der i-ten Zeile und j-ten Spalte das Element a_{ji} enthält. In gewisser Weise werden also beim Transponieren aus Zeilen Spalten gemacht und umgekehrt. Offenbar gilt $\mathbf{A}'' = \mathbf{A}$. Zwischen Multiplikation und Transposition gilt folgender Zusammenhang: Für Matrizen passender Größe ist $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$. Das Skalarprodukt zweier Spaltenvektoren $\mathbf{x}$ und $\mathbf{y}$ läßt sich jetzt auch als Matrizenprodukt $\mathbf{x}'\mathbf{y}$ schreiben.

Auch für Matrizen kann ein *Rang* definiert werden. Der Zeilenrang einer Matrix ist der Rang des Systems der als Zeilenvektoren aufgefaßten Zeilen der Matrix, analog ist der Spaltenrang definiert. Es stellt sich heraus, daß Zeilen- und Spaltenrang stets gleich sind, weshalb man hierfür auch kurz die Bezeichnung Rang verwendet. Insbesondere stimmen die Ränge einer Matrix und ihrer Transponierten überein.

Matrizen, deren Zeilen- und Spaltenzahl übereinstimmen, heißen quadratisch. Eine quadratische Matrix $\mathbf{A} = (a_{ij})$, bei der alle Elemente a_{ij} mit $i \neq j$, also alle außerhalb der „Diagonale", gleich 0 sind, heißt *Diagonalmatrix*. Eine Diagonalmatrix, die in der Diagonale nur Einsen enthält, heißt *Einheitsmatrix*. Die Abkürzung für die $(n \times n)$-Einheitsmatrix ist $\mathbf{I}_n$ oder kurz $\mathbf{I}$. Die Einheitsmatrix spielt beim Multiplizieren die Rolle der Eins: Für jede $(n \times m)$-Matrix $\mathbf{A}$ gilt nämlich $\mathbf{I}_n\mathbf{A} = \mathbf{A} = \mathbf{A}\mathbf{I}_m$. Gibt es für eine quadratische Matrix $\mathbf{A}$ eine weitere quadratische Matrix $\mathbf{B}$ mit $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$, so heißt $\mathbf{A}$ *invertierbar* oder *regulär*. Die Matrix $\mathbf{B}$ ist durch jede der Bedingungen $\mathbf{AB} = \mathbf{I}$ und $\mathbf{BA} = \mathbf{I}$ eindeutig bestimmt und heißt die *Inverse* von $\mathbf{A}$. Sie wird auch als $\mathbf{A}^{-1}$ notiert und ist ihrerseits invertierbar, ihre Inverse ist wieder $\mathbf{A}$. Nicht invertierbare Matrizen heißen auch *singulär*.

Eine $(n \times n)$-Matrix $\mathbf{A}$ ist genau dann invertierbar, wenn ihr Rang gleich n ist. Die wichtigsten Regeln lauten: Sind $\mathbf{A}$ und $\mathbf{B}$ invertierbar, so auch $\mathbf{AB}$; es gilt dann $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. Mit $\mathbf{A}$ ist auch $\mathbf{A}'$ invertierbar, es gilt $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$. Ferner gilt: Ist $\mathbf{A}$ eine $n \times n$-Matrix, $\mathbf{x}$ ein Vektor mit Unbekannten $x_1, \ldots, x_n$ und $\mathbf{y}$ ein n-Vektor, so ist $\mathbf{Ax} = \mathbf{y}$ ein lineares Gleichungssystem mit n Gleichungen und n Unbekannten. Ist $\mathbf{A}$ invertierbar, so ist das Gleichungssystem eindeutig lösbar, und die Lösung ist $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$.

Matrizen treten oft in der Modellierung psychologischer Zusammenhänge auf. So stellt man sich oft vor, daß eine Gruppe von m Variablen $X_1, \ldots, X_m$ (z.B. Intelligenzfaktoren) eine andere Gruppe von n Variablen $Y_1, \ldots, Y_n$ (z.B. Leistungen in praktischen Aufgaben) beeinflußt. Die Vorstellung der Beeinflussung wird dann meist so präzisiert, daß der Wert einer Y-Variablen sich ergeben soll, indem die Werte der X-Variablen mit jeweils spezifischen Koeffizienten multipliziert und die Ergebnisse aufaddiert werden; dazu kommt meist noch ein „Fehler" und unter Umständen eine Konstante. Man hat dann für jede Y-Variable eine Gleichung der Art, wie sie aus der multiplen Regression bekannt ist. Ob solche Modelle angemessen sind, muß im Einzelfall diskutiert werden, ihre statistische Behandlung ist bei aller Primitivität der Modellvorstellung jedenfalls kompliziert genug (was vielleicht manchmal den Blick auf die Frage nach der Angemessenheit verstellt). Stellt man die Werte einer Versuchsperson in den X-Variablen zu einem m-Vektor $\mathbf{x}$ und die Werte in den Y-Variablen zu einem n-Vektor $\mathbf{y}$ zusammen, so ergibt sich in einem Modell ohne Fehler und Konstante $\mathbf{y}$ als $\mathbf{Ax}$, wobei $\mathbf{A}$ die (für alle Personen gleiche) Matrix ist, deren Elemente a_{ij} jeweils das „Gewicht" angeben, mit der die j-te X-Variable in die Bildung der i-ten Y-Variablen eingeht. Kommt noch eine Konstante hinzu, so lautet

die entsprechende Gleichung $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$, wobei $\mathbf{b}$ ein n-Vektor von Konstanten ist.

Abbildungen der Form $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ heißen *lineare Abbildungen*. Sie respektieren die Vektoroperationen wegen $\mathbf{A}(\mathbf{x}_1 + \mathbf{x}_2) = \mathbf{A}\mathbf{x}_1 + \mathbf{A}\mathbf{x}_2$ und $\mathbf{A}(a\mathbf{x}) = a(\mathbf{A}\mathbf{x})$ (durch die entsprechenden Forderungen sind lineare Abbildungen allgemein definiert). Abbildungen der Form $\mathbf{x} \mapsto \mathbf{A}\mathbf{x} + \mathbf{b}$ heißen *affin*; sie sind die natürlichen Verallgemeinerungen der univariaten „linearen Transformationen" (wie man sieht, ist die Terminologie leider nicht ganz konsistent). Solche Abbildungen sind z.B. bei Transformationen von Variablen wichtig, wie in der später besprochenen Hauptkomponentenanalyse; auch der Übergang zu Koordinaten bezüglich eines neuen Koordinatensystems wird durch sie beschrieben. Lineare und affine Abbildungen sind dann umkehrbar, wenn $\mathbf{A}$ eine invertierbare quadratische Matrix ist.

Für quadratische Matrizen sind zwei Kennwerte von besonderer Bedeutung in der multivariaten Statistik, nämlich die Determinante und die Spur. Die *Spur* ist einfach die Summe der Diagonalelemente. Eine wichtige Regel: Ist $\mathbf{A}$ eine $(n \times m)$ Matrix und $\mathbf{B}$ eine $(m \times n)$-Matrix, so sind die Spuren von $\mathbf{AB}$ und $\mathbf{BA}$ gleich. Die Definition der *Determinante* $\det(\mathbf{A})$ einer Matrix $\mathbf{A}$ ist etwas komplizierter. Sie kann induktiv geschehen, durch Angabe der Determinante von (1×1)-Matrizen und einer Vorschrift, wie man die Determinante von $(n \times n)$-Matrizen berechnet, wenn man die von $((n-1) \times (n-1))$-Matrizen schon berechnen kann. Die Determinante einer (1×1)-Matrix ist einfach die Zahl, aus der die Matrix besteht. Ist $\mathbf{A}$ eine $(n \times n)$-Matrix und i eine Zahl zwischen 1 und n, so bezeichnet man mit $\mathbf{A}_{1i}$ die $((n-1) \times (n-1))$-Matrix, die entsteht, wenn man in $\mathbf{A}$ die erste Zeile und die i-te Spalte wegläßt, und definiert $\det(\mathbf{A}) := \sum_{i=1}^{n} (-1)^{(1+i)} a_{1i} \det(\mathbf{A}_{1i})$. Die Determinante einer (2×2)-Matrix $\mathbf{A} = (a_{ij})$ ist z.B. dann gleich $a_{11}a_{22} - a_{12}a_{21}$. Geometrisch ist der Betrag von $\det(\mathbf{A})$ das Volumen des „Parallelepipeds", das man erhält, wenn man die Menge aller Linearkombinationen der Spalten von $\mathbf{A}$ mit Koeffizienten zwischen 0 und 1 bildet; das Vorzeichen gibt die „Orientierung" wieder. Die Determinante läßt erkennen, ob eine Matrix regulär ist oder nicht, denn sie ist genau für reguläre Matrizen von Null verschieden. Wichtige Regeln sind: $\det(\mathbf{I}) = 1$, $\det(\mathbf{A}') = \det(\mathbf{A})$, $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$.

Die quadratischen $(n \times n)$-Matrizen repräsentieren lineare Abbildungen des $\mathbb{R}^n$ in sich selbst. Es kann vorkommen, daß dabei ein Vektor auf ein Vielfaches von sich selber abgebildet wird. Dies führt zu einer wichtigen Begriffsbildung: Ist $\mathbf{A}$ eine quadratische Matrix und gilt für eine Zahl λ und einen Vektor $\mathbf{x} \neq \mathbf{0}$ die Beziehung $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, so heißt λ *Eigenwert* von $\mathbf{A}$ und $\mathbf{x}$ zugehöriger *Eigenvektor*. Die Eigenvektoren zu einem festen Eigenwert bilden zusammen mit dem Nullvektor einen Unterraum, den sogenannten Eigenraum des Eigenwertes. Die Eigenwerte einer $(n \times n)$-Matrix $\mathbf{A}$ erhält man als die reellen Nullstellen von $\det(\lambda\mathbf{I} - \mathbf{A})$, einem Polynom n-ten Grades in λ, das auch charakteristisches Polynom von $\mathbf{A}$ heißt. Anstelle von reellen Vektorräumen kann man auch komplexe betrachten (man rechnet dann mit komplexen Zahlen und Vektoren); dann können zu den reellen Eigenwerten einer Matrix noch komplexe hinzukommen. In diesem Fall ist jede Nullstelle des charakteristischen Polynoms ein Eigenwert, und man bezeichnet die Vielfachheit der Nullstelle auch als die *Vielfachheit* oder *Multiplizität* des Eigenwerts. Nach einem wichtigen Satz (dem Fundamentalsatz der Algebra) addieren sich die Multiplizitäten aller Eigenwerte zu n, und daher kann es nie mehr als n Eigenwerte

geben. Die Dimensionen der Eigenräume können nicht größer sein als die Multiplizitäten der entsprechenden Eigenwerte. Die Determinante ist das Produkt und die Spur ist die Summe der Eigenwerte, wenn man jeden (möglicherweise komplexen) Eigenwert dabei so oft als Faktor bzw. Summand aufführt, wie seine Multiplizität angibt. Häufig wird folgende Tatsache verwendet: Ist $\mathbf{A}$ eine $(n \times m)$-Matrix und $\mathbf{B}$ eine $(m \times n)$-Matrix, so stimmen die Eigenwerte $\neq 0$ von $\mathbf{AB}$ mit den Eigenwerten $\neq 0$ von $\mathbf{BA}$ samt ihren Multiplizitäten überein, und die entsprechenden Eigenräume haben gleiche Dimension.

Diese im allgemeinen Fall etwas verwickelten Verhältnisse vereinfachen sich, wenn man symmetrische Matrizen betrachtet. Eine $(n \times n)$-Matrix $\mathbf{A}$ heißt *symmetrisch*, wenn $\mathbf{A} = \mathbf{A}'$ gilt. Nach dem sogenannten *Spektralsatz* haben symmetrische Matrizen nur reelle Eigenwerte, wobei die Multiplizität eines Eigenwerts gleich der Dimension des zugehörigen Eigenraums ist und Eigenvektoren zu verschiedenen Eigenwerten orthogonal sind. Man kann diese für die multivariate Statistik zentralen Sachverhalte auch noch anders ausdrücken: Für eine symmetrische Matrix $\mathbf{A}$ kann man eine Orthonormalbasis finden, die nur aus Eigenvektoren besteht. Bildet man mit diesen Eigenvektoren als Spalten eine Matrix $\mathbf{G}$, so gilt $\mathbf{G'G} = \mathbf{I}$ und $\mathbf{G'AG}$ ist eine Diagonalmatrix, deren Diagonalelemente die Eigenwerte entsprechend ihrer Vielfachheit sind (durch geeignete Anordnung der Spalten von $\mathbf{G}$ erreicht man auch, daß die Diagonalelemente der Größe nach geordnet sind).

Symmetrisch sind zum Beispiel Kovarianzmatrizen und Korrelationsmatrizen, die noch eine weitere wichtige Eigenschaft haben: Sie sind positiv semidefinit. Dabei heißt eine symmetrische $(n \times n)$-Matrix $\mathbf{A}$ positiv semidefinit, wenn für alle n-Vektoren $\mathbf{x}$ die Beziehung $\mathbf{x'Ax} \geq 0$ gilt. Gilt für alle $\mathbf{x} \neq \mathbf{0}$ sogar $\mathbf{x'Ax} > 0$, so heißt $\mathbf{A}$ *positiv definit*. Eine symmetrische Matrix ist genau dann positiv semidefinit, wenn alle Eigenwerte ≥ 0, und positiv definit, wenn alle Eigenwerte > 0 sind. Eine positiv semidefinite Matrix ist genau dann positiv definit, wenn sie regulär ist.

Positiv definite Matrizen spielen z.B. bei der Konstruktion von mehrdimensionalen Konfidenz„ellipsoiden" eine Rolle. Ist $\mathbf{A}$ eine positiv definite $(n \times n)$-Matrix, $\mathbf{b}$ ein fester n-Vektor und $k > 0$ eine reelle Zahl, so sei mit $\mathcal{E}(\mathbf{A}, \mathbf{b}, k)$ abkürzend die Menge $\{\mathbf{x} \in \mathbb{R}^n | (\mathbf{x} - \mathbf{b})'\mathbf{A}^{-1}(\mathbf{x} - \mathbf{b}) \leq k^2\}$ bezeichnet. Diese Menge ist ein *Ellipsoid*, eine n-dimensionale Kugel, die in mehreren aufeinander senkrecht stehenden Richtungen „gestaucht" oder „gestreckt" worden ist (für $n = 2$ eine Ellipse). Der Mittelpunkt des Ellipsoids ist der Punkt $\mathbf{b}$, und die Hauptachsen (das sind die „Durchmesser" in den Streckungsrichtungen) haben als Richtung die Eigenvektoren von $\mathbf{A}$ und als halbe Längen das k-fache der Wurzeln der zugehörigen Eigenwerte.

2 Begriffe der deskriptiven Statistik

Ausgangspunkt bei der Anwendung multivariater Verfahren ist meistens eine Situation, in der bei mehreren Personen nicht nur eine, sondern mehrere Variablen erhoben worden sind. So stehen am Anfang der Entwicklung des Gebiets unter anderem Analysen von Datensammlungen, in denen bei einer großen Anzahl von Verbrechern zum Zwecke der Identifizierung verschiedene Körpermaße wie Gesamtlänge, Kopfbreite, Länge des linken Fußes etc. registriert worden waren. Ziel war es dabei, aus den vielen ursprünglichen Variablen wenige neue zu konstruieren, die bis auf geringe

Verluste die „wesentliche Information" über die Maße der Personen enthielten.

Zunächst werden die Begriffe *Zentroid* und *Kovarianzmatrix* diskutiert. Üblicherweise faßt man Daten in einer Situation wie oben in einer Datenmatrix zusammen, in der die Zeilen für die Personen und die Spalten für die Variablen $X_1, \ldots, X_p$ stehen. Die Anzahl der Variablen sei p, die der Personen n; die Datenmatrix $\mathbf{X}$ ist dann eine $(n \times p)$-Matrix. In der i-ten Zeile stehen die Werte, die die i-te Person in den Variablen $X_1, \ldots, X_p$ hat. Zur Abkürzung für die transponierte i-te Zeile verwendet man meist das Symbol $\mathbf{x}_i$ (das also hier eine andere Bedeutung besitzt als im ersten Abschnitt). In einer Untersuchung, in der bei fünf Personen zwei Variablen X_1 und X_2 erhoben wurden, könnte sich die folgende 5×2-Matrix ergeben haben:

$$\begin{pmatrix} 2 & 1 \\ 4 & 3 \\ 1 & 1 \\ 5 & 2 \\ 3 & 3 \end{pmatrix}.$$

In der univariaten Statistik werden zur Charakterisierung einer Verteilung meistens der Mittelwert und die Varianz benutzt. Im multivariaten Fall benutzt man statt dessen den (p-dimensionalen) *Mittelwertsvektor* $\bar{\mathbf{x}}$ (manchmal auch *Zentroid* genannt), der als Komponenten die Mittelwerte der einzelnen Variablen besitzt, und die $(p \times p)$-*Kovarianzmatrix* $\mathbf{S}$, bei der in der i-ten Zeile und j-ten Spalte die Kovarianz zwischen der i-ten und der j-ten Variablen steht, wobei die in der Diagonale stehenden Varianzen und die Kovarianzen hier die unkorrigierten, also mit Division durch n ermittelten, seien. In unserem Beispiel ergeben sich

$$\begin{pmatrix} 3 \\ 2 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} 2 & 0.8 \\ 0.8 & 0.8 \end{pmatrix}$$

als Mittelwertsvektor und Kovarianzmatrix der Daten. Neben der Kovarianzmatrix spielt oft auch die *Korrelationsmatrix* eine Rolle, die entsprechend aufgebaut ist: Die Kovarianzen sind durch Korrelationen ersetzt, und in der Diagonale stehen Einsen.

Stellt man sich die Daten als Punktwolke im p-dimensionalen Variablenraum vor, so ist $\bar{\mathbf{x}}$ der „Schwerpunkt" der Punktwolke, wenn alle Datenpunkte die gleiche Masse besitzen, und daher zur Kennzeichnung ihrer Lage gut geeignet.

Die Kovarianzmatrix $\mathbf{S}$ enthält Informationen über die Variabilität der Daten. An ihrem Rang r kann man die „Dimension" der Punktwolke erkennen: Ist $r < p$, so liegen alle Datenpunkte in einem r-dimensionalen affinen Unterraum des Variablenraums, und zwar in dem, dessen zugehöriger linearer Unterraum durch die Spalten von $\mathbf{S}$ aufgespannt werden und der $\bar{\mathbf{x}}$ enthält. Beispielsweise bedeutet ein Rang von 1, daß alle Datenpunkte auf einer Geraden liegen, ein Rang von 0, daß sie alle gleich sind. Falls der Rang gleich p, $\mathbf{S}$ also regulär ist, so erhält man einen Eindruck von der Lage der Punkte, wenn man das (p-dimensionale) Ellipsoid $\mathcal{E}(\mathbf{S}, \bar{\mathbf{x}}, 1)$ betrachtet. Im Falle einer Variablen ist dies das Intervall mit den Punkten, die vom Mittelwert höchstens eine Standardabweichung entfernt liegen. Das Ellipsoid gibt häufig einen guten Eindruck von der Form und Ausdehnung der Punktwolke, was anschließend und weiter unten noch präzisiert wird. So liefert z.B. eine Verallgemeinerung der Tschebyscheffschen Ungleichung (vgl. z.B. Hays, 1988, S.182), daß der relative Anteil

der Punkte, die nicht im Innern von $\mathcal{E}(\mathbf{S}, \bar{\mathbf{x}}, k)$ (dem um den Faktor k vergrößerten $\mathcal{E}(\mathbf{S}, \bar{\mathbf{x}}, 1)$) liegen, höchstens p/k^2 ist. Man überzeugt sich leicht, daß im Beispiel .4 und 2.4 die Eigenwerte und $(-.5, 1)'$ und $(2, 1)'$ zugehörige Eigenvektoren sind. In der folgenden Abbildung 2 sind die Datenpunkte, das Zentroid und die Ellipse eingezeichnet. Die angesprochene Ungleichung besagt z.B. für $k = 2$, daß im Innern der um den Faktor 2 vergrößerten Ellipse mindestens die Hälfte der Daten liegt. In der Tat liegen im Beispiel sogar schon alle Punkte darin; die Ungleichung rechnet, wie im Univariaten, mit dem schlimmsten Fall.

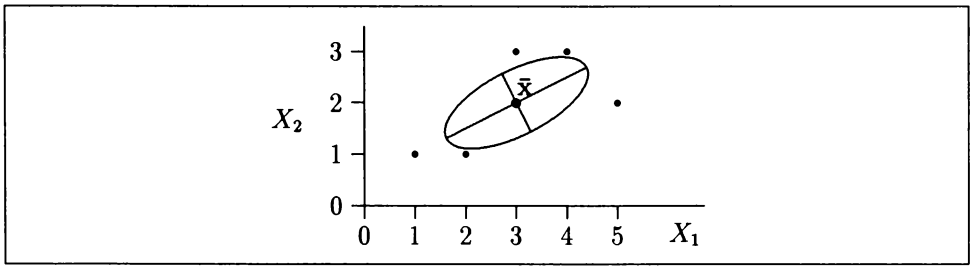

ABBILDUNG 2. **Die fünf Datenpunkte des Beispiels mit Zentroid $\bar{\mathbf{x}}$ und der zur Kovarianzmatrix gehörenden Ellipse $\mathcal{E}(\mathbf{S}, \bar{\mathbf{x}}, 1)$; die beiden Hauptachsen sind eingezeichnet.**

Häufig möchte man zusätzlich die Datenvariabilität durch eine einzige Zahl kennzeichnen. Dazu bieten sich als Möglichkeiten die Determinante und die Spur der Kovarianzmatrix an, die auch geometrische Interpretationen haben. Zieht man die Wurzel aus der Determinante, so erhält man bis auf einen von der Zahl p der Variablen abhängigen Faktor (für $p = 2$ z.B. π) das Volumen des Ellipsoids, was nicht unplausibel ist, wenn man bedenkt, daß die Determinante gleich dem Produkt der Eigenwerte ist; die Wurzeln der Eigenwerte waren ja gerade die halben Längen der Hauptachsen. Die Spur, die auch gleich der Summe der Eigenwerte ist, ist gleichzeitig der durchschnittliche quadrierte Abstand der Datenpunkte vom Zentroid, und damit eine unmittelbar naheliegende Verallgemeinerung der Varianz. Genaueren Aufschluß als diese beiden „summarischen" Werte geben allerdings die Eigenwerte selber, denen man zusätzlich Information über die Form des Ellipsoids entnehmen kann: Sind alle etwa gleich groß, so ist das Ellipsoid etwa kugelförmig, ist einer groß und alle anderen vergleichsweise klein und etwa gleich groß, so könnte man das Ellipsoid als „zigarrenförmig" bezeichnen etc. Vor diesem Hintergrund ist es verständlich, daß in die Bildung vieler Statistiken, mit denen multivariate Hypothesen getestet werden, die Eigenwerte von Kovarianzmatrizen, vielleicht auf dem Umweg über die Spur oder die Determinante, entscheidend eingehen.

Nun soll der zentrale Begriff der *Linearkombination* besprochen werden. Charakteristisch für multivariate Verfahren ist es, daß häufig fast gleichberechtigt mit den Variablen selbst beliebige Linearkombinationen der Variablen betrachtet werden. Sind die Ausgangsvariablen $X_1, \ldots, X_p$, so ist eine Linearkombination dieser Variablen eine neue Variable von der Form $\sum_{i=1}^{p} a_i X_i + b$, wobei die a_i und b fest vorgegebene reelle Zahlen sind; die a_i heißen auch Koeffizienten der Linearkombination. Im Gegensatz zu dem gleichlautenden Begriff aus der linearen Algebra ist hier also auch noch die Addition einer Konstanten b vorgesehen. Der Wert einer solchen

neuen Variable für eine Person wird natürlich durch Einsetzen der Werte der X_i ermittelt. Gelegentlich haben solche Linearkombinationen auch eine inhaltliche Bedeutung. Geben z.B. X_1 und X_2 die Werte einer Variablen vor und nach einem *treatment* an, so liefert die Linearkombination $X_2 - X_1$ die Veränderung. Meistens kann man allerdings Linearkombinationen nur mit hinreichender Phantasie einen inhaltlichen Sinn geben, was jedoch ihre Nützlichkeit für klar umrissene Zwecke nicht schmälert. So wird zum Beispiel bei der multiplen Regression diejenige Linearkombination einer Menge von „Prädiktoren" gesucht, die eine „Kriteriumsvariable" (im Sinne der „kleinsten Quadrate") optimal vorhersagt; Prädiktoren könnten hier ausgewählte Schulnoten von Studierenden sein, Kriterium die Endnote im Studium. Obwohl hier die gefundene Lösung kaum inhaltlich sinnvoll interpretierbar sein dürfte, leistet sie doch ihren Zweck, nämlich die Vorhersage, optimal.

Werden für eine beliebige Linearkombination die Koeffizienten a_i zu einem Vektor $\mathbf{a}$ zusammengefaßt, so ergibt sich der zu einem Datenvektor $\mathbf{x}$ gehörende Wert y der neuen Variablen Y als $y = \mathbf{a}'\mathbf{x} + b$. Die bekannten Formeln für Mittelwert $\bar{y}$ und Varianz S_Y^2 von Y lassen sich in Matrizenschreibweise kurz so formulieren: $\bar{y} = \mathbf{a}'\bar{\mathbf{x}} + b$ und $S_Y^2 = \mathbf{a}'\mathbf{S}\mathbf{a}$.

Im Variablenraum lassen sich spezielle Linearkombinationen, nämlich solche, bei denen die Summe der quadrierten Koeffizienten gleich 1 ist, geometrisch besonders schön veranschaulichen; solche Linearkombinationen nennt man auch *standardisiert*. Es sei hierzu Y eine solche standardisierte Linearkombination, die zusätzlich 0 als additive Konstante besitzt. Zeichnet man eine Gerade durch den Nullpunkt in Richtung des durch die Koeffizienten der Linearkombination gegebenen Vektors und macht sie zu einer Zahlengeraden, indem man die Spitze des Vektors zur Eins macht, so erhält man den Wert einer Versuchsperson in der neuen Variablen Y, indem man von dem zu dieser Person gehörenden Datenpunkt das Lot auf die Gerade fällt und dort die Zahl im Fußpunkt abliest. Den Mittelwert von Y erhält man, wenn man in dieser Weise das Zentroid auf sie projiziert, und als Bild des oben beschriebenen Ellipsoids unter der Projektion erhält man ein Intervall, dessen Endpunkte jeweils eine Standardabweichung der neuen Variablen vom Mittelwert entfernt sind (vgl. Abbildung 3a). Spezialfälle sind die Variablen selbst, die ja als Linearkombinationen mit Koeffizienten, die alle bis auf eine Eins gleich Null sind, aufgefaßt werden können; projiziert man also das Ellipsoid auf die Achsen, so kann man dort unmittelbar die Streuungen der Variablen ablesen. Da übrigens jede Linearkombination, bei der nicht alle Koeffizienten Null sind, durch eine geeignete lineare Transformation zu einer standardisierten gemacht werden kann, gelten auch für sie entsprechende Aussagen; man hat nur die Skala der Gerade, also den Nullpunkt und die Einheit entsprechend anders zu wählen, bei standardisierten Linearkombinationen, deren additive Konstante nicht Null ist, ist nur der Nullpunkt der neuen Zahlengerade zu verschieben.

Das nächste Thema ist die *Hauptkomponentenanalyse*. Eine naheliegende Frage bei der Betrachtung der geometrischen Situation ist die, in welcher Richtung die Streuung der Daten maximal bzw. minimal ist. Übersetzt ist dies die Frage nach standardisierten Linearkombinationen mit maximaler bzw. minimaler Streuung. Die Anschauung legt nahe, daß diese Richtungen die Richtungen der Achsen des Ellipsoids mit größter und kleinster Länge sind, also die der Eigenvektoren zum größten und kleinsten Eigenwert. Eine umfassende Antwort auf die Frage gibt die Hauptkompo-

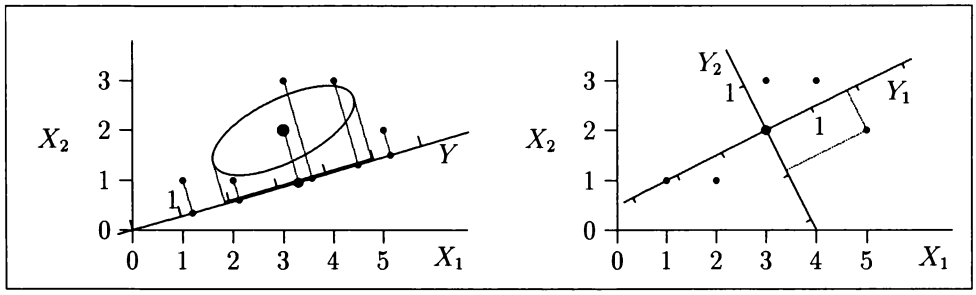

ABBILDUNG 3. a: Veranschaulichung der standardisierten Linearkombination $Y = .96X_1 + .28X_2$ durch eine orthogonale Projektion. Auf der ebenfalls mit Y bezeichneten Geraden können in den Lotfußpunkten die Werte von Y für die Datenpunkte aus Abbildung 2 abgelesen werden. Das Bild der Ellipse ist die Menge der Punkte, die höchstens eine Standardabweichung vom Mittelwert von Y entfernt sind. b: Geometrische Darstellung der Transformation der Daten aus Abbildung 2 in Hauptkomponenten Y_1 und Y_2. Die Hilfslinien zum Ablesen der neuen Koordinaten sind für den Punkt $(5,2)'$ eingezeichnet.

nentenanalyse. Man wählt hierbei zunächst ein System orthonormaler Eigenvektoren $\mathbf{g}_1, \ldots, \mathbf{g}_p$ zu den der Größe nach absteigend geordneten Eigenwerten von $\mathbf{S}$. Dies ist nach dem Spektralsatz stets möglich, und die Eigenvektoren sind sogar bis auf das Vorzeichen eindeutig bestimmt, falls alle Eigenwerte verschieden sind. Sodann bildet man mit den Komponenten dieser Vektoren als Koeffizienten standardisierte Linearkombinationen und wählt die additive Konstante so, daß dem Zentroid jeweils die Null zugeordnet wird. Die entstehenden neuen p Variablen heißen dann auch *Hauptkomponenten* und seien mit $Y_1, \ldots, Y_p$ bezeichnet.

Mit diesen Hauptkomponenten läßt sich die oben gestellte Frage beantworten: Unter allen standardisierten Linearkombinationen hat die erste Hauptkomponente (also die zum größten Eigenwert gehörende) die größte und die letzte die kleinste Varianz. Es gilt sogar noch mehr: Die Hauptkomponenten sind unkorreliert, unter allen standardisierten Linearkombinationen, die mit Y_1 unkorreliert sind, hat Y_2 größte Varianz, und entsprechend für die folgenden: Unter allen standardisierten Linearkombinationen, die zu $Y_1, \ldots, Y_k$ unkorreliert sind, hat Y_{k+1} größte Varianz. Die Varianzen sind dabei die zugehörigen Eigenwerte, die auf diese Weise auch statistische Bedeutung erhalten. Die Hauptkomponenten können alternativ auch durch die letztgenannten Eigenschaften definiert werden, immer mit dem Zusatz, daß die additive Konstante so zu wählen ist, daß das Zentroid auf Null abgebildet wird, daß also der Mittelwert der Hauptkomponenten Null sein soll. (Die Hauptkomponenten sind übrigens nicht eindeutig bestimmt, vor allem nicht, wenn mehrere Eigenwerte zusammenfallen, insofern ist das Wort „Definition" nicht ganz korrekt, aber Unterschiede zwischen verschiedenen Auswahlen sind unerheblich.)

Faßt man die Hauptkomponenten zu einem Vektor zusammen, so erhält man zu dem Datenpunkt $\mathbf{x}$ einer Person den Vektor $\mathbf{y}$ der Hauptkomponenten $Y_1, \ldots, Y_p$ durch die Vorschrift $\mathbf{y} = \mathbf{G}'(\mathbf{x} - \bar{\mathbf{x}}) = \mathbf{G}'\mathbf{x} - \mathbf{G}'\bar{\mathbf{x}}$, wo $\mathbf{G}$ die Matrix ist, deren Spalten die Eigenvektoren $\mathbf{g}_1, \ldots, \mathbf{g}_p$ sind. Geometrisch kann man den Vektor $\mathbf{y}$ als Koordinatenvektor des ursprünglichen Datenpunktes bezüglich desjenigen neuen Koordinatensystems auffassen, das seinen Nullpunkt in $\bar{\mathbf{x}}$ hat und dessen Achsen man erhält,

wenn man die Vektoren $\mathbf{g}_1, \ldots, \mathbf{g}_p$ in diesen Punkt verschiebt und nach beiden Seiten verlängert. Die Einheiten sind dabei durch die Spitzen der verschobenen Vektoren definiert (vgl. Abbildung 3b).

Mit Hilfe der Hauptkomponenten können viele weitere interessante Fragen beantwortet werden. So kann man zum Beispiel für ein vorgegebenes $k < p$ nach dem k-dimensionalen affinen Unterraum fragen, von dem die Datenpunkte minimalen durchschnittlichen quadrierten Abstand haben. Ist z.B. $p = 2$ und $k = 1$, so ist dies die Frage nach derjenigen Geraden, zu der der durchschnittliche quadrierte Abstand der Datenpunkte minimal ist. (Dies ist wohl zu unterscheiden von der ähnlichen Aufgabe bei der linearen Regression: Dort sind die „Abstandsmessungen" parallel zur Ordinatenachse vorzunehmen, hier senkrecht zur gesuchten Geraden.) Der gesuchte affine Raum ist derjenige, der das Zentroid enthält und von dort aus durch die ersten k Eigenvektoren aufgespannt wird – in dem Beispiel ist er also die Gerade durch das Zentroid in Richtung des ersten Eigenvektors, also der größten Hauptachse der Ellipse. Anders ausgedrückt ist der gesuchte Raum in dem neuen Koordinatensystem die Menge der Punkte, deren letzte $(p - k)$ Koordinaten alle Null sind. Fragt man nach dem Punkt des Unterraums, der zu einem gegebenen Datenpunkt den kleinsten Abstand hat, so erhält man dessen Koordinaten in dem neuen System einfach dadurch, daß man bei den Koordinaten des gegeben Punktes im neuen System die letzten $(p - k)$ Koordinaten durch Null ersetzt.

Dies hat auch praktische Konsequenzen: Will man aus Gründen der Ökonomie oder der Anschaulichkeit die Information, die über eine Person in den p Daten steckt, mit möglichst geringem „Informationsverlust" auf k Angaben reduzieren, so wählt man zweckmäßigerweise die Werte der ersten k Hauptkomponenten. Aus diesen kann man die ursprünglichen Daten zwar nicht genau rekonstruieren, man erhält jedoch eine ungefähre Vorstellung von der Lage des eigentlichen Datenpunktes, wenn man in dem k-dimensionalen Unterraum den Punkt einträgt, der die Hauptkomponenten als Koordinaten besitzt. Natürlich wird dieser Punkt meistens vom ursprünglichen Datenpunkt verschieden sein, und man kann auch den durchschnittlichen Fehler angeben: Wählt man als Maß für die durchschnittliche Abweichung den Durchschnitt der quadrierten Abstände der ursprünglichen und rekonstruierten Punkte, so ergibt sich die Summe der letzten $(p - k)$ Eigenwerte. Für alle anderen Möglichkeiten, die Daten durch k Linearkombinationen zu ersetzen, ist das entsprechende Maß mindestens ebensogroß, in diesem Sinne sind also die ersten k Hauptkomponenten optimal. Wählt man die Spur der Kovarianzmatrix, also den durchschnittlichen quadrierten Abstand zum Zentroid, als verallgemeinertes Maß für die Varianz, so ergibt sich, daß die Gesamtvarianz der Datenpunkte gleich der Summe aus der Varianz der rekonstruierten Punkte und der durchschnittlichen quadrierten Abweichung ist. In diesem Sinn kann man, wenn man will, das Verhältnis der Summe der ersten k Eigenwerte zur Spur von $\mathbf{S}$ als relativen Anteil der durch die ersten k Hauptkomponenten aufgeklärten Varianz ansehen und als Kriterium für die Wahl eines geeigneten k heranziehen, wenn man in diesem Sinne die Anzahl der Daten reduzieren will.

Die Vorteile eines solchen Vorgehens bestehen zunächst darin, daß jede Person, wenn auch mit Informationsverlust, durch weniger Werte charakterisiert ist (man denke an das Beispiel der Identifizierung von Verbrechern durch Körpermaße oder an Möglichkeiten der graphischen Datendarstellung). Ein weiterer Vorzug besteht in

der Unkorreliertheit der Hauptkomponenten. Diese sorgt nämlich für eine „Entkopplung" der Information in dem Sinn, daß bei einer linearen Regression keine Hauptkomponente etwas zur Vorhersage einer anderen beiträgt. Da alle Mittelwerte Null sind, sieht man darüber hinaus einem Wert sofort an, ob er über- oder unterdurchschnittlich ist. Die Unkorreliertheit und die geringere Anzahl der Variablen kann auch ein Vorteil sein, wenn man weitere statistische Verfahren damit durchführt, sie zum Beispiel als Prädiktoren in einer multiplen Regression einsetzt. Ein Nachteil liegt darin, daß die neuen Variablen keine inhaltliche Bedeutung haben; oft kann man jedoch bei Betrachtung der Gewichte, mit denen die Ausgangsvariablen in ihre Bildung eingehen, mit etwas Einfallsreichtum griffige Namen für sie finden.

Ein Problem bei Hauptkomponentenanalysen kann in der Skalierung der p Ausgangsvariablen liegen, von der das Ergebnis entscheidend abhängt. Unterwirft man die Variablen unterschiedlichen linearen Transformationen, so sind die Hauptkomponenten der umskalierten Variablen praktisch nicht mit denen der Originalvariablen vergleichbar, in dem Sinne, daß es keine bequeme Möglichkeit gibt, die beiden Sätze von Hauptkomponenten ineinander umzurechnen. Gibt es keinen Grund, eine bestimmte Kombination von Skalen zu bevorzugen, so ist eine inhaltliche Interpretation der Hauptkomponenten somit fragwürdig. Oft versucht man sich aus diesem Problem dadurch zu retten, daß man die Variablen vor der Analyse standardisiert, die Analyse also auf die Korrelationsmatrix anwendet.

Die Hauptkomponentenanalyse hat oberflächlich, z.B. in den benutzten Rechenverfahren, eine gewisse Ähnlichkeit mit der von Schönemann und Borg (in diesem Band) behandelten Faktorenanalyse. Wesensmäßig sind diese Verfahren jedoch grundverschieden: Während die Hauptkomponentenanalyse lediglich eine praktische Transformation der Variablen liefert, versucht die Faktorenanalyse, hinter den Variablen verborgene, vielleicht sogar „erklärende", Strukturen aufzufinden. Um so befremdlicher ist die leider häufig anzutreffende fehlende Differenzierung zwischen diesen beiden Verfahren.

Zum Schluß werden einige weitere wichtige *Datentransformationen* behandelt. Die Transformation der ursprünglichen Daten in die Hauptkomponenten, ebenso die Abbildungen, die Datenpunkten die Koordinaten der nächstgelegenen Punkte in einem Unterraum zuordnen, wie auch Linearkombinationen sind Beispiele für die Verwendung affiner Abbildungen und belegen deren Wichtigkeit für die multivariate Statistik. Ist $\mathbf{x}$ der Vektor der Werte einer Person in p Variablen $X_1, \ldots, X_p$, so kann man die Komponenten des mit Hilfe einer affinen Abbildung gebildeten Vektors $\mathbf{y} := \mathbf{A}\mathbf{x} + \mathbf{b}$ als Werte der Person in auf diese Weise neu gebildeten Variablen $Y_1, \ldots, Y_q$ ansehen. Die neue Variable Y_i ist dabei eine Linearkombination der ursprünglichen Variablen, und zwar diejenige, die als Koeffizienten die Elemente der i-ten Zeile von $\mathbf{A}$ und als Konstante die i-te Komponente von $\mathbf{b}$ besitzt. Als Zentroid der neuen Variablen in der Stichprobe ergibt sich dann $\bar{\mathbf{y}} = \mathbf{A}\bar{\mathbf{x}} + \mathbf{b}$ und als Kovarianzmatrix $\mathbf{ASA}'$. Diese grundlegenden Formeln sind Verallgemeinerungen der entsprechenden univariaten Formeln für Mittelwerte und Varianzen/Kovarianzen von Linearkombinationen, aus denen sie auch unmittelbar mit Hilfe elementarer Rechenregeln folgen.

Die Transformation der ursprünglichen Daten in die Hauptkomponenten hat die zusätzliche Eigenschaft, daß sie durch eine leicht angebbare weitere affine Transfor-

mation wieder rückgängig gemacht werden kann; aus den Hauptkomponenten kann man also die ursprünglichen Daten zurückrechnen. Solche Transformationen, die im Grunde nur die Daten anders und für einige Zwecke geeigneter darstellen, sind in den multivariaten Verfahren von großer Bedeutung. Man kann sich, als weiteres Beispiel, fragen, ob man analog zur z-Transformation im Univariaten die Daten so transformieren („standardisieren") kann, daß die transformierten Daten Mittelwerte von 0 und Varianzen von 1 haben und zusätzlich unkorreliert sind. In einem solchen Fall wäre das Zentroid nach der Transformation der Nullpunkt, die Kovarianzmatrix die Einheitsmatrix und das oben beschriebene Ellipsoid eine p-dimensionale Kugel. In der Tat gibt es, falls $\mathbf{S}$ invertierbar ist, mehrere mögliche Transformationen, die dies leisten. Eine davon erhält man, wenn man die Hauptkomponenten noch durch die Wurzeln aus den zugehörigen Eigenwerten teilt, eine andere, die sogenannte *Mahalanobis-Transformation*, erhält man mit der Vorschrift $\mathbf{y} := \mathbf{S}^{-1/2}(\mathbf{x} - \bar{\mathbf{x}})$ (für eine positiv definite Matrix $\mathbf{S}$ ist dabei $\mathbf{S}^{-1/2}$ diejenige – eindeutig bestimmte – positiv definite Matrix, die mit sich selbst multipliziert $\mathbf{S}^{-1}$ ergibt).

Auch für diese standardisierenden Transformationen sei gleich eine praktische Anwendung geschildert. Hat man für eine Stichprobe von Personen die Werte in einer Testbatterie, bestehend aus psychometrischen Tests $X_1, \ldots, X_p$, ermittelt, so möchte man oft gerne ein Maß dafür haben, wie stark sich eine Person von einer anderen oder vom Durchschnitt unterscheidet („Profilvergleiche"). Eine naheliegende Möglichkeit wäre es nun, den (euklidischen) Abstand der zugehörigen Punkte im Variablenraum zu benutzen, den man als Wurzel aus der Summe der quadrierten Differenzen in den einzelnen Untertests erhält. Bei diesem Vorgehen kann aber (z.B. bei hoch korrelierenden Tests) die unbefriedigende Situation auftreten, daß zwei Punkte zwar den gleichen Abstand vom Zentroid haben und damit das gleiche Maß der Verschiedenheit vom Durchschnitt erhalten, jedoch in bezug auf das die Kovarianzmatrix repräsentierende Ellipsoid ganz unterschiedliche Lagen haben, indem der eine deutlich innerhalb, der andere deutlich außerhalb liegt, weshalb unter Berücksichtigung der Form der Punktwolke der eine als normal, der andere als eher ungewöhnlich zu klassifizieren wäre. Ein Ausweg aus dieser Situation besteht darin, daß man euklidische Abstände erst bildet, nachdem man die Variablen nach einem der beschriebenen Verfahren in standardisierte transformiert hat. Es stellt sich heraus, daß das so definierte Abstandsmaß sogar unabhängig von der gewählten Standardisierung ist und sich für zwei Punkte $\mathbf{x_1}$ und $\mathbf{x_2}$ zu $\{(\mathbf{x_1} - \mathbf{x_2})'\mathbf{S}^{-1}(\mathbf{x_1} - \mathbf{x_2})\}^{1/2}$ berechnet. Die auf diese Weise definierte Distanz heißt auch *Mahalanobis-Distanz*; die Menge der Punkte, die vom Zentroid Mahalanobis-Distanz r haben, ist gerade der Rand von $\mathcal{E}(\mathbf{S}, \bar{\mathbf{x}}, r)$, was diesem Ellipsoid eine zusätzliche anschauliche Bedeutung verleiht. Viele Formeln der multivariaten Statistik lassen sich als Anwendung naheliegender Operationen auf Daten deuten, die man zunächst mit einer geeigneten Transformation bezüglich einer Fehlerkovarianzmatrix „standardisiert" hat, und werden so leichter durchschaubar.

3 Wahrscheinlichkeitstheoretische Begriffe

In diesem Abschnitt geht es um den Begriff des *Zufallsvektors* und um die *multivariate Normalverteilung*. Parallel zu den beschriebenen deskriptiven Konzepten

für Stichprobendaten lassen sich entsprechende wahrscheinlichkeitstheoretische definieren. Hat man p Zufallsvariablen $x_1, \ldots, x_p$ gegeben, die auf einem gemeinsamen Wahrscheinlichkeitsraum definiert sind, so faßt man sie oft zweckmäßigerweise zu einem p-dimensionalen Zufallsvektor $\mathbf{x}$ zusammen. Solche Zufallsvariablen könnten beispielsweise die Werte sein, die eine zufällig aus einer Population zu ziehende Person in p Variablen aufweist, z.B. in Persönlichkeitsvariablen, für deren Zusammenhang man sich interessiert. Den *Erwartungswert(vektor)* eines solchen Zufallsvektors definiert man als Vektor der Erwartungswerte der Einzelvariablen, und seine *Kovarianzmatrix* analog wie oben als Matrix der Varianzen und Kovarianzen der Zufallsvariablen. Entsprechend definiert man Zufallsmatrizen und deren Erwartungswerte. Traditionellerweise benutzt man übrigens in der multivariaten Statistik zur Bezeichnung von Zufallsvektoren und ihren Bestandteilen meist die gleichen Symbole wie zur Bezeichnung möglicher Werte (Kleinbuchstaben, z.B. $\mathbf{x}$ und $x_1, \ldots, x_p$); die Bedeutung eines Symbols ist jeweils dem Kontext zu entnehmen. Das Symbol für den Erwartungswert ist E, ist also $\mathbf{x}$ ein Zufallsvektor der Länge p, so gilt $E(\mathbf{x}) = (E(x_1), \ldots, E(x_p))'$. Die Kovarianzmatrix eines Zufallsvektors $\mathbf{x}$ sei mit $V(\mathbf{x})$ bezeichnet.

Ist nun $\mathbf{x}$ ein p-dimensionaler Zufallsvektor mit Erwartungswert $\boldsymbol{\mu}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$, so charakterisiert analog wie oben das Ellipsoid $\mathcal{E}(\boldsymbol{\Sigma}, \boldsymbol{\mu}, 1)$ die Verteilung von $\mathbf{x}$; die Wahrscheinlichkeit dafür, daß $\mathbf{x}$ Werte nicht im Innern von $\mathcal{E}(\boldsymbol{\Sigma}, \boldsymbol{\mu}, k)$ annimmt, ist beispielsweise höchstens p/k^2. Ist $\mathbf{A}$ eine $(q \times p)$-Matrix und $\mathbf{b}$ ein q-Vektor, so wird durch die Vorschrift $\mathbf{y} := \mathbf{A}\mathbf{x} + \mathbf{b}$ ein neuer q-dimensionaler Zufallsvektor definiert, dessen Erwartungswert gleich $\mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ und dessen Kovarianzmatrix gleich $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ ist. Linearkombinationen sind Spezialfälle ($q = 1$). Analog zur Stichprobensituation kann man auch Hauptkomponenten definieren.

Meistens werden in einer Untersuchung die Variablen mehrfach erhoben. So wird man in dem Beispiel mit den Persönlichkeitsvariablen eine größere Stichprobe von n Personen planen. Im wahrscheinlichkeitstheoretischen Modell der Stichprobenziehung setzt man dann im allgemeinen voraus, daß die n Zufallsvektoren $\mathbf{x}_1, \ldots, \mathbf{x}_n$, die aus den Ergebnissen der n Personen der Zufallsstichprobe bestehen, gemeinsam unabhängig sind und die gleiche Verteilung haben. Aus den erhobenen Werten wird man dann den Mittelwertsvektor $\bar{\mathbf{x}}$ und die Kovarianzmatrix $\mathbf{S}$ der Stichprobe berechnen. Unter den angegebenen Voraussetzungen ($\mathbf{x}_1, \ldots, \mathbf{x}_n$ unabhängig, $E(\mathbf{x}_i) = \boldsymbol{\mu}$, $V(\mathbf{x}_i) = \boldsymbol{\Sigma}$) gilt dann: $E(\bar{\mathbf{x}}) = \boldsymbol{\mu}$, $V(\bar{\mathbf{x}}) = (1/n)\boldsymbol{\Sigma}$ und $E(\mathbf{S}) = ((n-1)/n)\boldsymbol{\Sigma}$. Es gelten also ähnliche Beziehungen wie im Univariaten: Der Mittelwertsvektor ist ein erwartungstreuer Schätzer von $\boldsymbol{\mu}$, die „Varianz" der Schätzung wird mit dem Faktor $1/n$ kleiner, die Stichprobenkovarianzmatrix $\mathbf{S}$ ist nicht erwartungstreu für $\boldsymbol{\Sigma}$, wird es aber, wenn man sie mit dem Korrekturfaktor $n/(n-1)$ multipliziert. Bei komplizierteren Begriffsbildungen sind die Verhältnisse etwas verwickelter: Die Eigenwerte der Stichprobenkovarianzmatrix sind immerhin konsistente Schätzer der Eigenwerte der theoretischen Kovarianzmatrix (gleichgültig, ob man die Korrektur mit $n/(n-1)$ vornimmt oder nicht), während man für die Konsistenz der Koeffizientenvektoren der Hauptkomponenten noch Zusatzvoraussetzungen machen muß, um die Mehrdeutigkeiten in der Definition zu beseitigen.

Für Zufallsvektoren wird meistens die Annahme der Multinormalverteiltheit gemacht, eine Annahme, die vielleicht nicht realistisch ist, jedoch häufig eine bequeme

statistische Behandlung überhaupt erst ermöglicht. In vielen Fällen hat man auch Gründe zu der Hoffnung, daß Verletzungen sich nicht allzu gravierend auswirken. Die *multivariate Normalverteilung* (kurz: *Multinormalverteilung*) wird meistens in folgender Weise definiert: Ein Zufallsvektor $\mathbf{x} = (x_1, \ldots, x_p)'$ heißt multinormalverteilt, falls jede Linearkombination seiner Komponenten $x_1, \ldots, x_p$ normalverteilt oder (fast sicher) konstant ist. Diese Definition hat den Nachteil, daß man dann zum Beispiel konstante „Zufallsvariablen" ebenfalls zu den normalverteilten rechnet, was jedoch durch das Wegfallen lästiger Fallunterscheidungen kompensiert wird. Eine Multinormalverteilung ist durch ihren Erwartungswert und ihre Kovarianzmatrix bereits eindeutig charakterisiert. Ist ein p-dimensionaler Zufallsvektor $\mathbf{x}$ multinormalverteilt mit $E(\mathbf{x}) = \boldsymbol{\mu}$ und $V(\mathbf{x}) = \boldsymbol{\Sigma}$, so wird dies mit $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ abgekürzt. Ist $\boldsymbol{\Sigma}$ regulär, so besitzt die $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-Verteilung eine Wahrscheinlichkeitsdichte, und zwar die, die in einem Punkt $\mathbf{x}$ den Wert

$$\frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

annimmt. Man erkennt als wesentlichen Bestandteil des Exponenten die quadrierte Mahalanobis-Distanz von $\mathbf{x}$ und $\boldsymbol{\mu}$ und sieht ferner, daß die Mengen, auf denen die Dichtefunktion konstant ist (im Zweidimensionalen die „Höhenlinien") die Ränder der bekannten Ellipsoide $\mathcal{E}(\boldsymbol{\Sigma}, \boldsymbol{\mu}, r)$ für feste Werte von r sind.

Unmittelbar aus der Definition der Multinormalverteiltheit folgt, daß diese Eigenschaft bei affinen Abbildungen nicht verlorengeht: Ist $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ und $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ ($\mathbf{A}$ sei dabei eine $q \times p$-Matrix und $\mathbf{b}$ ein q-Vektor), so ist $\mathbf{y} \sim N_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$. Außerdem sind Summen unabhängiger multinormalverteilter Zufallsvariablen ebenfalls wieder multinormalverteilt; als Spezialfall vererbt sich die Multinormalverteiltheit auf den Mittelwertsvektor: Sind $\mathbf{x}_1, \ldots, \mathbf{x}_n$ unabhängig, $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, so ist $\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, (1/n)\boldsymbol{\Sigma})$. Die Kovarianzmatrix $\mathbf{S}$ hat in dieser Situation übrigens eine wichtige Bedeutung: Sie ist Maximum-Likelihood-Schätzer für $\boldsymbol{\Sigma}$.

4 Weiterführende Literatur

Einige mathematischen Grundlagen für die multivariate Datenanalyse sind für Psychologen dargestellt in dem Buch von Rhenius (1983). Ausführlicher sind Green und Carroll (1976). Kurze Einführungen in Vektor- und Matrizenrechnung gibt es in mehreren Büchern über multivariate Statistik, vgl. hierfür die Angaben in dem Kapitel über multivariate Verfahren. Dort finden sich auch Verweise auf Texte zu den wahrscheinlichkeitstheoretischen Begriffen und zur Hauptkomponentenanalyse. Für den mathematisch interessierten Leser kann zur linearen Algebra die noch recht elementare Darstellung von Lang (1970) empfohlen werden.

Literaturverzeichnis

Green, P. E. & Carroll, J. D. (1976). *Mathematical tools for applied multivariate analysis.* New York: Academic Press.

Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth: Holt, Rinehart and Winston.

Lang, S. (1970). *Linear algebra* (2nd ed.). Reading: Addison-Wesley.

Rhenius, D. (1983). *Mathematik für die Psychologie: Eine Einführung.* Bern: Huber.