Vernetzungen zwischen Vektorgeometrie und Beschreibender Statistik

JÖRG MEYER, HAMELN

Zusammenfassung: *In der Beschreibenden Statistik kommen häufig Quadratsummen vor. Deutet man diese als Skalarprodukte, so lassen sich manche Aussagen über Mittelwerte, Varianzen oder über Regressionskoeffizienten in durchsichtiger Weise vektorgeometrisch deuten und beweisen. Auch zur Matrizenrechnung wird ein Zusammenhang hergestellt. Behandelt werden die Themenkomplexe Minimalität des arithmetischen Mittels, Regressionsgeraden und Regressionsparabeln. Auch zur Lagebeziehung von arithmetischem Mittel und Median wird ein Zusammenhang hergestellt.*

1. Einleitung

In den neuen Bundes-Einheitlichen Prüfungsanforderungen im Fach Mathematik steht:

„Die Prüfungsaufgaben im Abitur erfordern einen Unterricht, der in den drei Sachgebieten (Analysis, Lineare Algebra / Analytische Geometrie und Stochastik) den Aufbau adäquater Grundvorstellungen der zentralen Begriffe und Methoden als Schwerpunkt hat [...].“

Das ist neu! Und das wird nur gelingen, wenn man die drei Gebiete vielfältig miteinander vernetzt. Hier geht es um Vernetzungen zwischen der Beschreibenden Statistik und der Vektorgeometrie.

Zur Notation: Im Folgenden werden Punkte mit ihren zugehörigen Ortsvektoren identifiziert.

2. Zur Minimalitätseigenschaft des arithmetischen Mittels

Es seien $d_1, d_2, ..., d_n$ irgendwelche numerischen Daten und $\alpha := \frac{d_1 + d_2 + ... + d_n}{n}$ deren arithmetisches Mittel. Die Minimalitätseigenschaft des arithmetischen Mittels lautet so:

$$\text{Für alle c ist } \sum_{i=1}^{n}(\alpha - d_i)^2 \le \sum_{i=1}^{n}(c - d_i)^2 .$$

Anders formuliert: Die Aufgabe

$$\text{„Bestimme c so, dass } \sum_{i=1}^{n}(c - d_i)^2 \text{ minimal ist“}$$

wird durch $c = \alpha$ gelöst.

Dies lässt sich auch *vektorgeometrisch* beweisen! Dazu führen wir zwei n-dimensionale Vektoren ein, und zwar den Datenvektor $D = \begin{pmatrix} d_1 \\ ... \\ d_n \end{pmatrix}$ sowie den Einsenvektor $E = \begin{pmatrix} 1 \\ ... \\ 1 \end{pmatrix}$. Mit dem Standard-Skalarprodukt $X \cdot Y = \sum_{i=1}^{n} x_i \cdot y_i$ schreibt sich die Aufgabe folgendermaßen: „Bestimme c so, dass der Vektor $\begin{pmatrix} c - d_1 \\ ... \\ c - d_n \end{pmatrix} = c \cdot E - D$ minimale Länge hat“.

Die (geometrische) Lösung ist offensichtlich: Man muss nur D auf E senkrecht projizieren (Abb. 1).

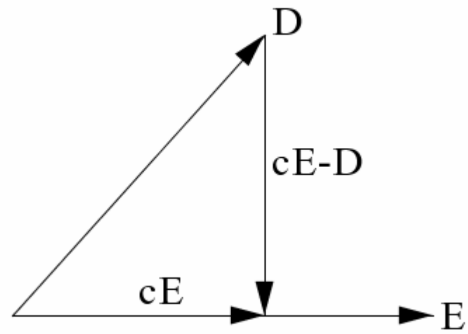

Abb. 1

Die Länge von $c \cdot E - D$ ist minimal, wenn $c \cdot E - D$ auf E senkrecht steht, wenn also $c \cdot \underbrace{E \cdot E}_{n} = \underbrace{D \cdot E}_{d_1 + ... + d_n}$ und deswegen $c = \frac{D \cdot E}{E^2} = \alpha$ gilt.

Die Länge $\left|\alpha \cdot E - D\right| = \sqrt{\sum_{i=1}^{n} (\alpha - d_i)^2}$ des Abweichungsvektors ist Wesentlichen die Standardabweichung.

Die bei der Berechnung der empirischen Varianz häufig verwendete Formel

$$\sum_{i=1}^{n} (\alpha - d_i)^2 = \sum_{i=1}^{n} d_i^2 - n \cdot \alpha^2 .$$

ist nur der Satz des *Pythagoras* in der Form

$$(\alpha \cdot E - D)^2 = D^2 - (\alpha \cdot E)^2 .$$

Eine Analogisierung dieser Betrachtungen in Richtung Verknüpfung zwischen Vektorgeometrie und Wahrscheinlichkeitsrechnung findet sich bei Scheid, H. (1986): Stochastik in der Kollegstufe. BI: Mannheim.

3. Der Regressionskoeffizient und Projektionen

Gegeben sind n Datenpaare $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$ $(i = 1, ..., n)$.

Gesucht ist diejenige Gerade („*Ausgleichs*-“ oder „*Regressionsgerade*“) mit der Gleichung $y = a \cdot x + b$, die die Daten möglichst „gut“ annähert. Die y-Werte sind möglicherweise messfehlerbehaftet, die x-Werte nicht.

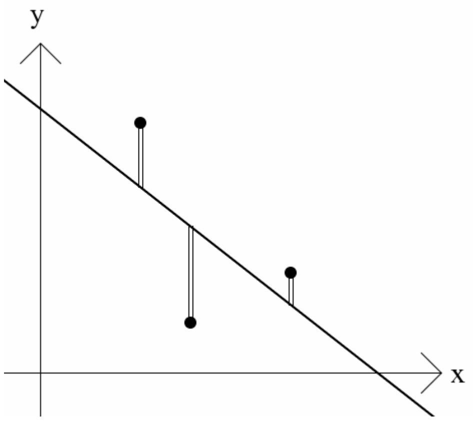

Abb. 2

Was heißt „gut“? Sicherlich ist es sinnvoll zu fordern, dass die Summe der vertikalen Abstände (in Abb. 2 durch Doppellinien gekennzeichnet) verschwindet, d. h. dass

$$\sum_{i=1}^{n} (a \cdot x_i + b - y_i) = 0$$

ist. Damit ist aber die Regressionsgerade noch nicht eindeutig bestimmt. *Eine* weitere (fruchtbare) Forderung an die zu findende Gerade besteht darin, dass

$$\sum_{i=1}^{n} (a \cdot x_i + b - y_i)^2 \quad \text{minimal}$$

wird. Abweichend von Abschnitt 1 kürzen wir die arithmetischen Mittel hier als $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

und $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ ab.

Die *erste Forderung*

$$\sum_{i=1}^{n} (a \cdot x_i + b - y_i) = 0$$

schreibt sich dann als

$$a \cdot \bar{x} + b = \bar{y} ;$$

die gesuchte Gerade geht somit durch den Schwerpunkt. (Das ist auch der Fall, wenn nicht die Summe der vertikalen Abstände, sondern die Summe der zur Ausgleichsgerade senkrechten Abstände - in Abb. 3 durch Einfachlinien gekennzeichnet - verschwinden soll.)

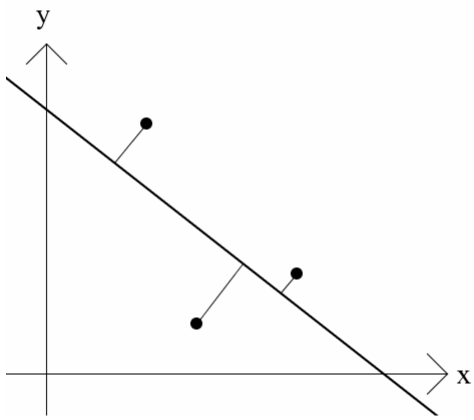

Abb. 3

Daher liegt eine Koordinatenverschiebung $u := x - \bar{x};\ v := y - \bar{y}$ nahe; sie führt zu

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} x_i - \bar{x} \\ y_i - \bar{y} \end{pmatrix},$$

und die Regressionsgerade bekommt die einfache Gleichung $v = a \cdot u$.

Natürlich ist $\bar{u} = \bar{v} = 0$, diese Gleichungen lassen sich mit $U = \begin{pmatrix} u_1 \\ \dots \\ u_n \end{pmatrix}$, $V = \begin{pmatrix} v_1 \\ \dots \\ v_n \end{pmatrix}$ und $E = \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}$ als *Orthogonalitätsrelationen*

$$U \cdot E = V \cdot E = 0 \qquad \text{(OR)}$$

deuten.

Die zweite Forderung

$$\sum_{i=1}^{n} (v_i - a \cdot u_i)^2 \text{ minimal}$$

bedeutet: Wähle a so, dass $a \cdot U - V$ möglichst kurz ist. Man bekommt dieses a, wenn man V auf U senkrecht projiziert (Abb. 4).

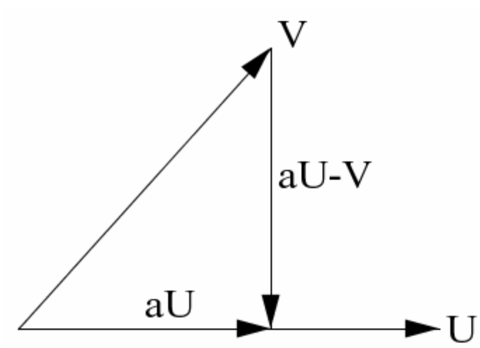

Abb. 4

Es ist dann a so zu bestimmen, dass

$$(V - a \cdot U) \cdot U = 0$$

ist, was auf $a = \frac{U \cdot V}{U^2}$ führt. Bekanntlich heißt a *Regressionskoeffizient*.

An dieser Stelle sollte man der Frage nachgehen, *wie gut* die Datenpunkte durch eine Gerade beschrieben werden.

Wenn alle Daten genau auf einer Geraden liegen, ist $V = a \cdot U$. Genau dann ist $\cos(U, V) = \pm 1$.

Auf der anderen Seite hat man die maximale Abweichung von einer Geradenform, falls U und V zueinander senkrecht stehen. Genau dann ist $\cos(U, V) = 0$.

Daher ist der *Korrelationskoeffizient* $\cos(U, V) = \frac{U \cdot V}{|U| \cdot |V|}$ ein gutes Maß dafür, wie gut die Datenpunkte durch eine Gerade beschrieben werden können.

Nebenbei: Die vorgängige Bestimmung von $b = 0$ ist sachlich überflüssig (allerdings didaktisch sinnvoll), es reicht die zweite Forderung

$$\sum_{i=1}^{n} \left(v_i - (a \cdot u_i + b)\right)^2 \text{ minimal}.$$

Wir haben dann das Problem:

Bestimme a und b so, dass

$$a \cdot \begin{pmatrix} u_1 \\ \dots \\ u_n \end{pmatrix} + b \cdot \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix} - \begin{pmatrix} v_1 \\ \dots \\ v_n \end{pmatrix} = a \cdot U + b \cdot E - V$$

möglichst kurz ist!

U und E spannen eine Ebene auf. Gesucht sind dann a und b so, dass der Abstand zwischen $a \cdot U + b \cdot E$ und V minimal ist.

Das erreicht man, wenn man den Vektor V auf die von U und E aufgespannte Ebene senkrecht projiziert (Abb. 5).

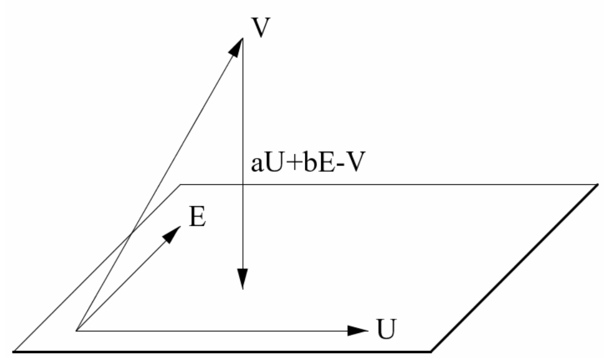

Abb. 5

Wie bestimmt man die Projektion? Es muss sein:

$$(a \cdot U + b \cdot E - V) \cdot U = 0 \text{ und}$$
$$(a \cdot U + b \cdot E - V) \cdot E = 0.$$

Dies Gleichungssystem lässt sich besonders einfach lösen, falls (E, U) eine Orthogonalbasis ist. Dies ist aber hier wegen (OR) der Fall. Dann ist $a = \frac{U \cdot V}{U \cdot U}$ und $b = \frac{E \cdot V}{E \cdot E} = 0$.

4. Der Regressionskoeffizient der standardisierten Daten

Wir hatten die Daten $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$ zentralisiert zu $\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} x_i - \bar{x} \\ y_i - \bar{y} \end{pmatrix}$. Nun ist es eine sinnvolle Idee, die Daten auch zu *normieren* zu $S = \frac{U}{|U|} \cdot \lambda$ und $T = \frac{V}{|V|} \cdot \lambda$. Dabei ist natürlich $\lambda > 0$.

Vom Standpunkt der Vektorgeometrie aus ist $\lambda = 1$ naheliegend, vom Standpunkt der Stochastik ist es $\lambda = \sqrt{n}$ oder $\lambda = \sqrt{n-1}$. Alsdann ist $|S| = |T| = \lambda$.

Berechnet man für diese normierten Daten den Regressionskoeffizienten, so bekommt man

$$a = \frac{S \cdot T}{S^2} = \frac{\frac{U}{|U|} \cdot \lambda \cdot \frac{V}{|V|} \cdot \lambda}{\frac{U}{|U|} \cdot \lambda \cdot \frac{U}{|U|} \cdot \lambda} = \frac{U \cdot V}{|U| \cdot |V|} = \cos(U, V).$$

Für (irgendwie) normierte Daten stimmen also Regressions- und Korrelationskoeffizient überein.

5. Der Regressionskoeffizient und Matrizen

Das Problem bei der Regressionsgerade war:

Bestimme a und b so, dass

$$a \cdot \begin{pmatrix} u_1 \\ \dots \\ u_n \end{pmatrix} + b \cdot \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix} = a \cdot U + b \cdot E = \underbrace{(U, E)}_{=:M} \cdot \begin{pmatrix} a \\ b \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} u_1 & 1 \\ \dots & \dots \\ u_n & 1 \end{pmatrix}}_{=M} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = M \cdot \begin{pmatrix} a \\ b \end{pmatrix}$$

sich möglichst wenig von $\begin{pmatrix} v_1 \\ \dots \\ v_n \end{pmatrix} = V$ unterscheidet.

Zu lösen wäre also das *überbestimmte* System $M \cdot \begin{pmatrix} a \\ b \end{pmatrix} = V$. Hier liegt ein anderer Repräsentationswechsel als bei den Projektionen vor.

Man multipliziert auf beiden Seiten mit der transponierten Matrix M^t und erhält $M^t \cdot M \cdot \begin{pmatrix} a \\ b \end{pmatrix} = M^t \cdot V$. Hier ist $M^t \cdot M$ eine quadratische und symmetrische Matrix, und dies Gleichungssystem ist lösbar! Man erhält $M^t \cdot M = \begin{pmatrix} U^2 & 0 \\ 0 & n \end{pmatrix}$ und $M^t \cdot V = \begin{pmatrix} U \cdot V \\ 0 \end{pmatrix}$.

Das Gleichungssystem lautet also

$$\begin{pmatrix} U^2 & 0 \\ 0 & n \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} U \cdot V \\ 0 \end{pmatrix}$$

und hat die Lösung $a = \frac{U \cdot V}{U^2}$ und $b = 0$.

6. Der Zusammenhang zwischen arithmetischem Mittel, Median und Standardabweichung

Es seien $d_1, d_2, \dots, d_n$ wie in Abschnitt 1 irgendwelche numerischen Daten und $\alpha := \frac{d_1 + d_2 + \dots + d_n}{n}$ deren *arithmetisches Mittel*, $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \alpha)^2}{n}}$ deren *Standardabweichung* sowie β deren *Median*.

Dann gilt: Der Abstand der beiden Mittel α und β ist durch σ beschränkt, d. h. es gilt:

$$|\alpha - \beta| \le \sigma.$$

Wie kann man das beweisen? Wir fangen mit dem linken Term an. Nach der Dreiecksungleichung für Beträge gilt

$$|\alpha - \beta| = \frac{1}{n} \cdot \left| \sum_{i=1}^{n} (d_i - \beta) \right| \le \frac{1}{n} \cdot \sum_{i=1}^{n} |d_i - \beta|$$

und aufgrund der Minimalitätseigenschaft des Medians ist

$$\frac{1}{n} \cdot \sum_{i=1}^{n} |d_i - \beta| \le \frac{1}{n} \cdot \sum_{i=1}^{n} |d_i - \alpha|.$$

Wenn nun noch

$$\frac{1}{n} \cdot \sum_{i=1}^{n} |d_i - \alpha| \le \sqrt{\frac{\sum_{i=1}^{n} (d_i - \alpha)^2}{n}} = \sigma$$

gelten würde, hätte man die Behauptung bewiesen. Schreibt man, um die Struktur des zu Beweisenden klarer zu sehen, z_i für $|d_i - \alpha|$, so muss

$$\frac{\sum_{i=1}^{n} z_i}{n} \le \sqrt{\frac{\sum_{i=1}^{n} z_i^2}{n}}$$

bzw.

$$\left(\sum_{i=1}^{n} z_i \right)^2 \le n \cdot \sum_{i=1}^{n} z_i^2 \qquad \text{(U)}$$

gelten. Nun kann man in der rechten Seite von (U) ein *Skalarprodukt* zu erkennen. Mit $Z = \begin{pmatrix} z_1 \\ \dots \\ z_n \end{pmatrix}$ und

$E = \begin{pmatrix} 1 \\ ... \\ 1 \end{pmatrix}$ sowie dem Standard-Skalarprodukt $X \cdot Y = \sum_{i=1}^{n} x_i \cdot y_i$ ist $n = E \cdot E$, und (U) schreibt sich als $(Z \cdot E)^2 \le E^2 \cdot Z^2$; das ist aber richtig wegen $(Z \cdot E)^2 = E^2 \cdot Z^2 \cdot \cos^2(E, Z)$.

Vernetzungen zwischen Stochastik und Vektorgeometrie lohnen sich also! Der Repräsentationswechsel

Quadratsumme $\Rightarrow$ Skalarprodukt

ist häufig fruchtbar, wie an den Beispielen wohl deutlich geworden ist.

Übrigens: Analysiert man den Beweis zu $|\alpha - \beta| \le \sigma$, so stellt man fest, dass sich *genauere Aussagen* machen lassen:

Die Dreiecksungleichung für Beträge liefert

$$|\alpha - \beta| = \frac{1}{n} \cdot \left| \sum_{i=1}^{n} (d_i - \beta) \right| \le \frac{1}{n} \cdot \sum_{i=1}^{n} |d_i - \beta| .$$

Die rechte Seite ist das zum Median gehörige Streuungsmaß $\tau := \frac{\sum_{i=1}^{n} |d_i - \beta|}{n}$, die *mittlere absolute Abweichung*. Damit ist

$$|\alpha - \beta| \le \tau .$$

Aufgrund der Minimalitätseigenschaft des Medians gilt

$$\tau = \frac{\sum_{i=1}^{n} |d_i - \beta|}{n} \le \frac{\sum_{i=1}^{n} |d_i - \alpha|}{n} .$$

Nun ist wie oben

$$\frac{\sum_{i=1}^{n} |d_i - \alpha|}{n} \le \sqrt{\frac{\sum_{i=1}^{n} (d_i - \alpha)^2}{n}} = \sigma ,$$

so dass man insgesamt die Ungleichung

$$|\alpha - \beta| \le \tau \le \sigma$$

hat.

7. Zur quadratischen Regression

Die Methoden von Abschnitt 2 lassen sich fruchtbar machen zur Erläuterung der quadratischen Regression. Wieder haben wir n Datenpaare $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$ ($i = 1, ..., n$), und gesucht ist diejenige Parabel, die die Daten möglichst „gut" annähert. Wie im Fall der linearen Regression wird es sich als vorteilhaft erweisen, wenn man eine Schwerpunktstranslation vornimmt und zu $\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} x_i - \overline{x} \\ y_i - \overline{y} \end{pmatrix}$ übergeht. Es sind dann a, b und c so zu bestimmen, dass die v_i möglichst dicht bei den jeweiligen Werten für $a \cdot u_i^2 + b \cdot u_i + c$ liegen. Mit

$$U = \begin{pmatrix} u_1 \\ u_2 \\ ... \\ u_n \end{pmatrix}, \ Q := \begin{pmatrix} u_1^2 \\ u_2^2 \\ ... \\ u_n^2 \end{pmatrix}, \ V = \begin{pmatrix} v_1 \\ v_2 \\ ... \\ v_n \end{pmatrix} \text{ und } E = \begin{pmatrix} 1 \\ 1 \\ ... \\ 1 \end{pmatrix}$$

heißt das: V soll möglichst dicht bei $a \cdot Q + b \cdot U + c \cdot E$ liegen.

Hier ist der Anlass, geometrische Grundvorstellungen auf den nicht mehr vorstellbaren vierdimensionalen Raum zu erweitern:

- Wenn $a \cdot U - V$ möglichst kurz sein soll, muss man V auf die durch den Richtungsvektor U aufgespannte Ursprungsgerade projizieren (Abschnitt 1).
- Wenn $a \cdot U + b \cdot E - V$ möglichst kurz sein soll, muss man V auf die durch die Richtungsvektoren U und E aufgespannte Ursprungsebene projizieren (Abschnitt 2).
- Wenn $a \cdot Q + b \cdot U + c \cdot E - V$ möglichst kurz sein soll, so sollte man analog V auf denjenigen dreidimensionalen Raum projizieren, der durch den Ursprung geht und durch die drei Richtungsvektoren Q, U und E aufgespannt wird.

Analog zu den Abschnitten 1 und 2 führt das auf die drei Bedingungen

$$(a \cdot Q + b \cdot U + c \cdot E - V) \cdot Q = 0$$
$$(a \cdot Q + b \cdot U + c \cdot E - V) \cdot U = 0$$
$$(a \cdot Q + b \cdot U + c \cdot E - V) \cdot E = 0 .$$

(Man gelangt übrigens zu den gleichen Termen, wenn man $\Delta := \sum_{i=1}^{n} \left(a \cdot u_i^2 + b \cdot u_i + c - v_i \right)^2$ nach a, b und nach c ableitet, dieser Weg hätte natürlich auch schon früher offen gestanden.)

Aufrund der Orthogonalitätsrelationen (OR) und wegen $Q \cdot E = U \cdot U$ schreibt sich das Gleichungssystem einfacher als

$$\begin{array}{lllllll} a \cdot Q \cdot Q & + & b \cdot U \cdot Q & + & c \cdot U \cdot U & = & V \cdot Q \\ a \cdot Q \cdot U & + & b \cdot U \cdot U & & & = & U \cdot V \\ a \cdot U \cdot U & & & + & c \cdot E \cdot E & = & 0 \end{array}$$

Das Gleichungssystem wird noch etwas einfacher, wenn man die x-Werte als *äquidistant* annimmt, wenn also $x_{i+1} - x_i = u_{i+1} - u_i$ von i unabhängig ist. Unter dieser Voraussetzung ist nämlich $U \cdot Q = 0$, und man bekommt das recht übersichtliche System

$$\begin{array}{lllllll} a \cdot Q \cdot Q & & & + & c \cdot U \cdot U & = & V \cdot Q \\ & & b \cdot U \cdot U & & & = & U \cdot V \\ a \cdot U \cdot U & & & + & c \cdot E \cdot E & = & 0 \end{array}$$

Beispiel: Gegeben seien die 4 Punkte $\begin{pmatrix} -1 \\ 2 \end{pmatrix}$, $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$ und $\begin{pmatrix} 2 \\ 3 \end{pmatrix}$, für die die Ausgleichsparabel gesucht ist. Wegen $\bar{x} = \frac{1}{2}$ und $\bar{y} = \frac{9}{4}$ ist

$$U = \begin{pmatrix} -3/2 \\ -1/2 \\ 1/2 \\ 3/2 \end{pmatrix}, \ Q = \begin{pmatrix} 9/4 \\ 1/4 \\ 1/4 \\ 9/4 \end{pmatrix} \text{ und } V = \begin{pmatrix} -1/4 \\ -5/4 \\ 3/4 \\ 3/4 \end{pmatrix}.$$

Das Gleichungssystem hat die Lösung

$$a = \frac{1}{4}, \ b = \frac{1}{2}, \ c = -\frac{5}{16};$$

die Ausgleichsparabel hat somit die Gleichung $v = \frac{u^2}{4} + \frac{u}{2} - \frac{5}{16}$; Abb. 6 zeigt die Situation.

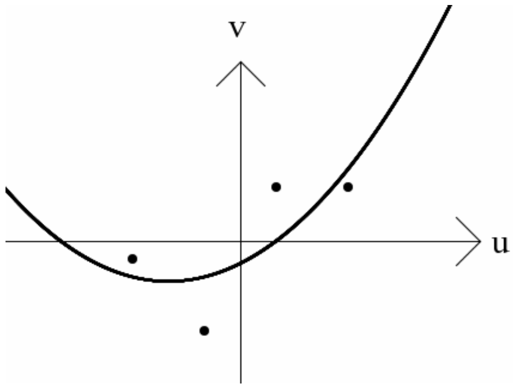

Abb. 6

Die Vorgehensweise überträgt sich auf Polynome höheren Grades.

Bei der Parabelregression kann es natürlich passieren, dass der führende Koeffizient a verschwindet. Man sieht am Gleichungssystem, dass das genau dann der Fall ist, wenn Q auf V senkrecht steht (wie es auch zu erwarten ist). Alsdann ist natürlich auch $c = 0$.

Abb. 7 zeigt das Beispiel $X = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}$, $Y = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 3 \end{pmatrix}$. Man muss Y nur an einer Stelle geringfügig verändern, um aus der Geraden eine nach oben oder eine nach unten geöffnete Parabel zu erzeugen.

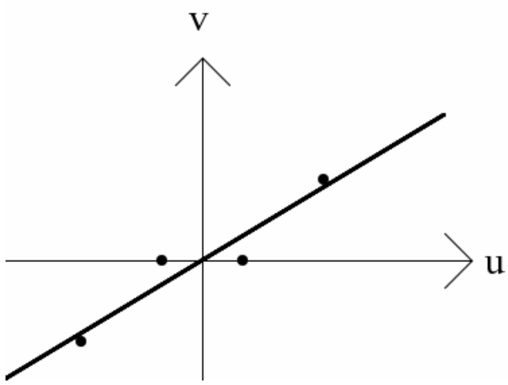

Abb. 7

Anschrift des Verfassers

Jörg Meyer
Schäfertrift 16
31789 Hameln
J.M.Meyer@t-online.de