Jörg Meyer, Hameln

Stochastik und Vektorgeometrie[1]

***Zusammenfassung*:** *In der Beschreibenden Statistik und in der Wahrscheinlichkeitsrechnung kommen häufig Quadratsummen vor. Deutet man diese als Skalarprodukte, so lassen sich manche Aussagen über Mittel- und Erwartungswerte, Varianzen oder über Regressionskoeffizienten in durchsichtiger Weise vektorgeometrisch deuten und beweisen.*
Behandelt werden auch die Lagebeziehung von arithmetischem Mittel und Median sowie Regressionsparabeln.

0 Zur Notation

Im Folgenden werden Punkte mit ihren zugehörigen Ortsvektoren identifiziert.

1 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 Skalarprodukt $\begin{pmatrix} x_1 \\ ... \\ x_n \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix} := \frac{\sum_{i=1}^{n} x_i \cdot y_i}{n}$ gilt $E^2 = 1$ sowie $\alpha = D \cdot E$, und die Aufgabe schreibt sich 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).

---

[1] Eine etwas andere und weniger umfangreiche Version erschien unter dem Titel „Vernetzungen zwischen Vektorgeometrie und Beschreibender Statistik“ in Stochastik in der Schule **24** (2); S. 24 - 29 (2004).

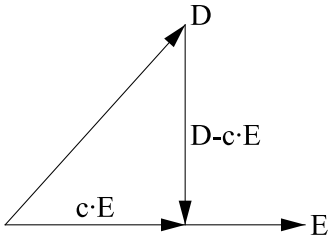

Abb. 1

Die Länge von $c \cdot E - D$ ist minimal, wenn $c \cdot E - D$ auf E senkrecht steht, wenn also $c \cdot E^2 = D \cdot E$ und deswegen $c = \alpha$ gilt.

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

Die bei der Berechnung der empirischen Varianz häufig verwendete Formel

$$\text{Var} = \sigma^2 = \frac{\sum_{i=1}^{n} (d_i - \alpha)^2}{n} = \frac{\sum_{i=1}^{n} d_i^2}{n} - \alpha^2 .$$

ist nur der Satz des *Pythagoras* in der Form $(\alpha \cdot E - D)^2 = D^2 - (\alpha \cdot E)^2$.

1a Vernetzungen zwischen Vektorgeometrie und Wahrscheinlichkeitsrechnung

Eine Analogisierung der obigen Betrachtungen in Richtung Verknüpfung zwischen Vektorgeometrie und Wahrscheinlichkeitsrechnung ist auch leicht möglich:

Die Ergebnismenge sei endlich, und eine darauf definierte Zufallsvariable Z nehme die Werte d_i mit den Wahrscheinlichkeiten p_i an ($i = 1, ..., n$). Mit dem Skalarprodukt $\begin{pmatrix} x_1 \\ ... \\ x_n \end{pmatrix} * \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix} := \sum_{i=1}^{n} p_i \cdot x_i \cdot y_i$,

dem Wertevektor $D = \begin{pmatrix} d_1 \\ ... \\ d_n \end{pmatrix}$ sowie dem Einsenvektor $E = \begin{pmatrix} 1 \\ ... \\ 1 \end{pmatrix}$ gilt $E * E = 1$, $D * E = \mu$ (Erwartungswert $E(Z)$), $(D - \mu \cdot E) * E = D * E - \mu \cdot E * E = \mu - \mu = 0$ sowie $|D - \mu \cdot E| = \sqrt{\sum_{i=1}^{n} p_i \cdot (d_i - \mu)^2} = \sigma$ (Standardabweichung $\sqrt{\text{Var}(Z)}$). Wegen

$$\begin{aligned} (D - \mu \cdot E)^2 &= \sum_{i=1}^{n} p_i \cdot (d_i - \mu)^2 \\ &= \sum_{i=1}^{n} p_i \cdot d_i^2 - 2 \cdot \mu \cdot \sum_{i=1}^{n} p_i \cdot d_i + \mu^2 \cdot \sum_{i=1}^{n} p_i \\ &= \sum_{i=1}^{n} p_i \cdot d_i^2 - \mu^2 \end{aligned}$$

gilt auch hier der Satz des Pythagoras in der Form $(D - \mu \cdot E)^2 = D^2 - \mu^2$ und führt zur Fomel

$$\mathrm{Var}(Z) = \mathrm{E}\left(Z^2\right) - \left(\mathrm{E}(Z)\right)^2 .$$

2 Der Regressionskoeffizient und Projektionen

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

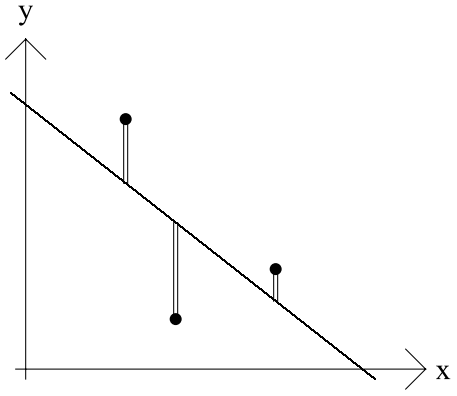

Abb. 2

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

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

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

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

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*

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

schreibt sich dann als

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

die gesuchte Gerade geht somit durch den Schwerpunkt. 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 $U = \begin{pmatrix} u_1 \\ ... \\ u_n \end{pmatrix}$, $V = \begin{pmatrix} v_1 \\ ... \\ v_n \end{pmatrix}$ und $E = \begin{pmatrix} 1 \\ ... \\ 1 \end{pmatrix}$ als *Orthogonalitätsrelationen*

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

deuten. Die *zweite Forderung*

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

bedeutet: Wähle a so, dass $a \cdot U - V$ möglichst kurz ist. Man bekommt dieses a, wenn man V auf U senkrecht projiziert (Abb. 3).

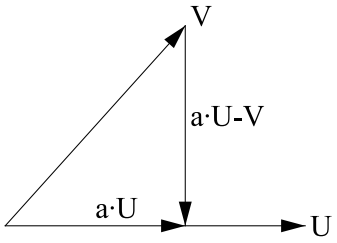

Abb. 3

Es ist dann a so zu bestimmen, dass

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

ist, was auf $a = \dfrac{U \cdot V}{U^2}$ 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* $\cos(U, V) = \dfrac{U \cdot V}{|U| \cdot |V|}$ 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

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

Wir haben dann das Problem: Bestimme a und b so, dass

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

mit $U = \begin{pmatrix} u_1 \\ ... \\ u_n \end{pmatrix}$, $E = \begin{pmatrix} 1 \\ ... \\ n \end{pmatrix}$ und $V = \begin{pmatrix} v_1 \\ ... \\ v_n \end{pmatrix}$ 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. 4).

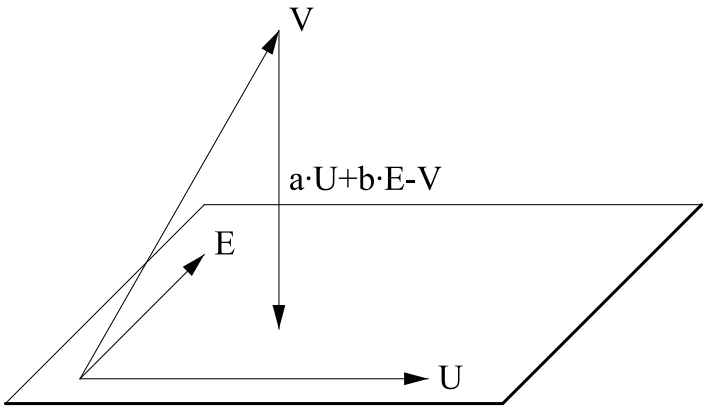

Abb. 4

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 $b = E \cdot V = 0$.

3 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 $S = \frac{U}{|U|}$ und $T = \frac{V}{|V|}$.

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

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

Für normierte Daten stimmen also Regressions- und Korrelationskoeffizient überein.

4 Der Zusammenhang zwischen arithmetischem Mittel, Median und Standardabweichung

Es seien $d_1, d_2, ..., d_n$ wie in Abschnitt 1 irgendwelche numerischen Daten und $\alpha := \frac{d_1 + d_2 + ... + d_n}{n}$ 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

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

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 die Abschätzung

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

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.

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

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

$E = \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}$ sowie dem Skalarprodukt $X\cdot Y = \frac{\sum_{i=1}^{n} x_i\cdot y_i}{n}$ ist $E^2 = 1$, und (U) schreibt sich als $\left(Z\cdot E\right)^2 \le Z^2$

; das ist aber richtig wegen $\left(Z\cdot E\right)^2 = Z^2\cdot E^2\cdot\cos^2\left(E, Z\right)$.

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 *weitere Aussagen* machen lassen: Die Dreiecksungleichung für Beträge liefert (DrU), und deren 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

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

Wegen (Absch) hat man insgesamt die Ungleichung

$$|\alpha - \beta| \leq \tau \leq \sigma.$$

5 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 $\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} x_i - \bar{x} \\ y_i - \bar{y} \end{pmatrix}$ 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 $Q \cdot E = U \cdot U$ und $E \cdot E = 1$ schreibt sich das Gleichungssystem einfacher als

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

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. Für ungerade Werte von n ist dann U von der Form $U = \begin{pmatrix} u-k\cdot\delta \\ \dots \\ u-\delta \\ u \\ u+\delta \\ \dots \\ u+k\cdot\delta \end{pmatrix} = \begin{pmatrix} -k\cdot\delta \\ \dots \\ -\delta \\ 0 \\ \delta \\ \dots \\ k\cdot\delta \end{pmatrix}$ (wegen $U\cdot E = 0$), also $U\cdot Q = 0$. Für gerade Werte ist

$U = \begin{pmatrix} u-k\cdot\delta \\ \dots \\ u-3\cdot\delta \\ u-\delta \\ u+\delta \\ u+3\cdot\delta \\ \dots \\ u+k\cdot\delta \end{pmatrix} = \begin{pmatrix} -k\cdot\delta \\ \dots \\ -3\cdot\delta \\ -\delta \\ \delta \\ 3\cdot\delta \\ \dots \\ k\cdot\delta \end{pmatrix}$, und hier folgt ebenso $U\cdot Q = 0$. Man bekommt das recht übersichtliche System

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

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

$$U = \frac{1}{4}\cdot\begin{pmatrix} -3 \\ -1 \\ 1 \\ 3 \end{pmatrix},\ Q = \frac{1}{8}\cdot\begin{pmatrix} 9 \\ 1 \\ 1 \\ 9 \end{pmatrix} \text{ und } V = \frac{1}{8}\cdot\begin{pmatrix} -1 \\ -5 \\ 3 \\ 3 \end{pmatrix}.$$

Das Gleichungssystem ist

$$\begin{array}{rcrcl} a\cdot\frac{164}{64} & & + & c\cdot\frac{20}{16} & = & \frac{16}{64} \\ & b\cdot\frac{20}{16} & & & = & \frac{20}{32} \\ a\cdot\frac{20}{16} & & + & c & = & 0 \end{array}$$

und 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. 5 zeigt die Situation.

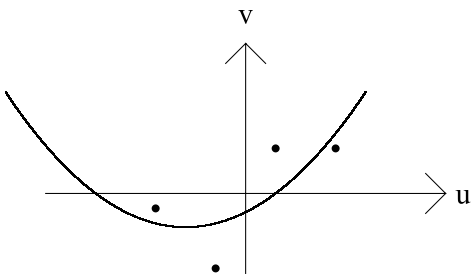

Abb. 5

Die Vorgehensweise überträgt sich auf Polynome höheren Grades.