gleich sein. Die Lösung dieses Problems ist klar: Man nimmt den Mittelwert der Quotienten. Schwieriger wird es, wenn mehrere Parameter ins Spiel kommen, wenn die Meßreihe als mehr als nur einen Parameter bestimmen soll.

Solche Fälle treten nicht nur auf in Naturwissenschaft und Technik, sondern auch in den Wirtschafts- und Sozialwissenschaften, wo es zwar selten *exakte* Gesetze gibt, man den Zusammenhang zwischen verschiedenen Größen aber trotzdem zumindest näherungsweise durch eine mathematische Formel beschreiben will – auch wenn diese in konkreten Einzelfällen gelegentlich ziemlich falsch sein kann.

Als Beispiel dieser Art können wir den Zusammenhang zwischen Korruption und Wohlstand in verschiedenen Staaten betrachten: edes Jahr veröffentlicht die Organisation *Transparancy International* ihren *corruption perceptions index (CPI)*, in dem jedem Land eine Zahl zwischen null und zehn zugeordnet wird, je nachdem, wie stark Geschäftsleute, Risikospezialisten und die Bevölkerung die Korruption im betreffenden Land einschätzen: Ein Index von zehn bedeutet, daß es praktische keine Korruption gibt, während bei null nichts läuft ohne Bimbes. Die neuesten Daten stammen vom 18. Oktober 2005 und sind unter

`http://www.transparency.org/cpi/`

zu finden. Die Zahlen werden als Mittelwerte über die letzten drei Jahren berechnet, so daß singuläre Ereignisse eines Jahres nicht zu sehr ins Gewicht fallen. Wir vergleichen diese Zahlen mit dem Bruttonationaleinkommen pro Einwohner, das auf dem Server des Statistischen Bundesamtes unter

`http://www.destatis.de/ausl_prog/suche_ausland.htm`

zu finden ist, indem man unter „Indikatoren“ das Feld „BNE je Einwohner“ auswählt. Es ist in sogenannten „Internationalen Dollar“ angegeben, das sind von der Weltbank mit einem Kaufkraftfaktor korrigierte US-$. Die meisten Werten beziehen sich auf das Jahr 2004, in einigen Fällen sind allerdings auch nur Daten für 2003 oder gar 2002 verfügbar. In der folgenden Tabelle sind alle Staaten aufgelistet, für die sowohl das Bruttonationaleinkommen pro Einwohner als auch der CPI für 2005 vorliegt;

das Bruttonationaleinkommen ist kursiv gedruckt, der Korruptionsindex fett:

Ägypten	*4120*	**3,4**
Albanien	*5070*	**2,4**
Algerien	*6260*	**2,8**
Angola	*2030*	**2,0**
Argentinien	*12460*	**2,8**
Armenien	*4270*	**2,0**
Aserbaidschan	*3830*	**2,2**
Äthiopien	*810*	**2,2**
Australien	*29200*	**8,8**
Bahrain	*18070*	**5,8**
Bangladesch	*1980*	**1,7**
Barbados	*15060*	**6,9**
Belgien	*31360*	**7,4**
Belize	*6510*	**3,7**
Benin	*1120*	**2,9**
Bolivien	*2590*	**2,5**
Bosnien und Herzegowina	*7430*	**2,9**
Botsuana	*8920*	**5,9**
Brasilien	*8020*	**3,7**
Bulgarien	*7870*	**4,0**
Burkina Faso	*1220*	**3,4**
Burundi	*660*	**2,3**
Chile	*10500*	**7,3**
China	*5530*	**3,2**
Costa Rica	*9530*	**4,2**
Côte d’Ivoire	*1390*	**1,9**
Dänemark	*31550*	**9,5**
Deutschland	*27950*	**8,2**
Dominikanische Republik	*6750*	**3,0**
Ecuador	*3690*	**2,5**
El Salvador	*4980*	**4,2**
Eritrea	*1050*	**2,6**
Estland	*13190*	**6,4**
Finnland	*29560*	**9,6**

Frankreich	*29320*	**7,5**
Gabun	*5600*	**2,9**
Gambia	*1900*	**2,7**
Georgien	*2930*	**2,3**
Ghana	*2280*	**3,5**
Griechenland	*22000*	**4,3**
Guatemala	*4140*	**2,5**
Guyana	*4110*	**2,5**
Haiti	*1680*	**1,8**
Honduras	*2710*	**2,6**
Indien	*3100*	**2,9**
Indonesien	*3460*	**2,2**
Iran	*7550*	**2,9**
Irland	*33170*	**7,4**
Island	*32360*	**9,7**
Israel	*23510*	**6,3**
Italien	*27860*	**5,0**
Jamaika	*3630*	**3,6**
Japan	*30040*	**7,3**
Jemen	*820*	**2,7**
Jordanien	*4640*	**5,7**
Kambodscha	*2180*	**2,3**
Kamerun	*2090*	**2,2**
Kanada	*30660*	**8,4**
Kasachstan	*6980*	**2,6**
Kenia	*1050*	**2,1**
Kirgisistan	*1840*	**2,3**
Kolumbien	*6820*	**4,0**
Kongo	*750*	**2,3**
Kongo, Dem. Republik	*680*	**2,1**
Korea, Republik	*20400*	**5,0**
Kroatien	*11670*	**3,4**
Kuwait	*19510*	**4,7**
Laos, Dem. Volksrepublik	*1850*	**3,3**
Lesotho	*3210*	**3,4**
Lettland	*11850*	**4,2**

Libanon	*5380*	**3,1**
Litauen	*12610*	**4,8**
Luxemburg	*61220*	**8,5**
Madagaskar	*830*	**2,8**
Malawi	*620*	**2,8**
Malaysia	*9630*	**5,1**
Mali	*980*	**2,9**
Malta	*18720*	**6,6**
Marokko	*4100*	**3,2**
Mauritius	*11870*	**4,2**
Mazedonien	*6480*	**2,7**
Mexiko	*9590*	**3,5**
Moldau, Republik	*1930*	**2,9**
Mongolei	*2020*	**3,0**
Mosambik	*1160*	**2,8**
Namibia	*6960*	**4,3**
Nepal	*1470*	**2,5**
Neuseeland	*22130*	**9,6**
Nicaragua	*3300*	**2,6**
Niederlande	*31220*	**8,6**
Niger	*830*	**2,4**
Nigeria	*930*	**1,9**
Norwegen	*38550*	**8,9**
Oman	*13250*	**6,3**
Österreich	*31790*	**8,7**
Pakistan	*2160*	**2,1**
Panama	*6870*	**3,5**
Papua-Neuguinea	*2300*	**2,3**
Paraguay	*4870*	**2,1**
Peru	*5370*	**3,5**
Philippinen	*4890*	**2,5**
Polen	*12640*	**3,4**
Portugal	*19250*	**6,5**
Ruanda	*1300*	**3,1**
Rumänien	*8190*	**3,0**
Russische Föderation	*9620*	**2,4**

Sambia	*890*	**2,6**
Saudi-Arabien	*14010*	**3,4**
Schweden	*29770*	**9,2**
Schweiz	*35370*	**9,1**
Senegal	*1720*	**3,2**
SierraLeone	*790*	**2,4**
Simbabwe	*2180*	**2,6**
Singapur	*26590*	**9,4**
Slowakei	*14370*	**4,3**
Slowenien	*20730*	**6,1**
Spanien	*25070*	**7,0**
Sri Lanka	*4000*	**3,2**
Südafrika	*10960*	**4,5**
Sudan	*1870*	**2,1**
Swasiland	*4970*	**2,7**
Syrien, Arabische Republik	*3550*	**3,4**
Tadschikistan	*1150*	**2,1**
Tansania, Vereinigte Republik	*660*	**2,9**
Thailand	*8020*	**3,8**
Trinidad und Tobago	*11180*	**3,8**
Tschad	*1420*	**1,7**
Tschechische Republik	*18400*	**4,3**
Tunesien	*7310*	**4,9**
Türkei	*7680*	**3,5**
Turkmenistan	*6910*	**1,8**
Uganda	*1520*	**2,5**
Ukraine	*6250*	**2,6**
Ungarn	*15620*	**5,0**
Uruguay	*9070*	**5,9**
Usbekistan	*1860*	**2,2**
Venezuela	*5760*	**2,3**
Vereinigte Staaten	*39710*	**7,6**
Vereinigtes Köenigreich	*31460*	**8,6**
Vietnam	*2700*	**2,6**
Weißrußland	*6900*	**2,6**
Zypern	*22330*	**5,7**

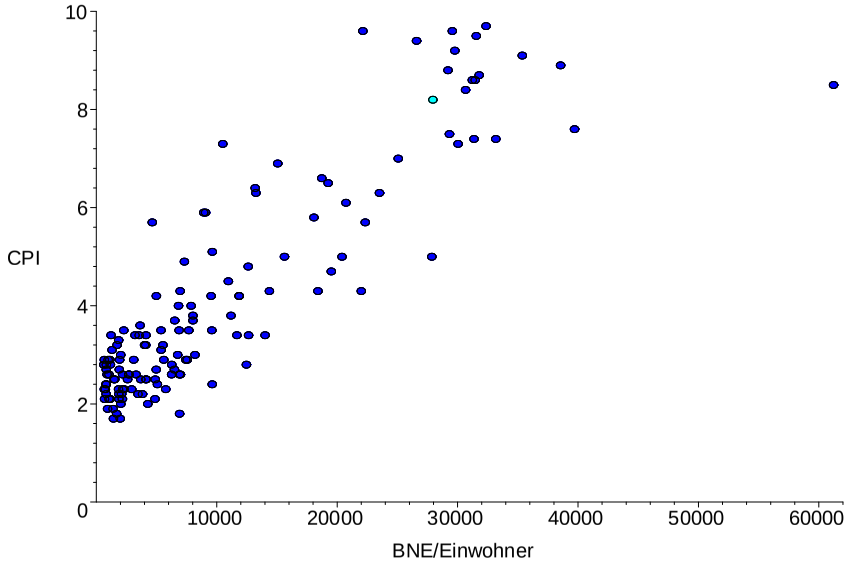

Abb. 18: Zusammenhang zwischen Korruption und Bruttonationaleinkommen je Einwohner

Abbildung 18 zeigt die 142 Datenpunkte zu dieser Liste graphisch, wobei der Punkt für Deutschland etwas heller eingezeichnet ist.

Der erste Augenschein zeigt, daß korruptionsärmere Länder oftmals reicher sind: Das weitgehend korruptionsfreie Island hat ein Bruttonationaleinkommen von 32 360 $ pro Einwohner, das deutlich korruptere Deutschland nur 27 950 $ und ein stark korruptes Land wie Tansania nur 660 $. Allerdings gibt es auch Ausnahmen: Beispielsweise hat Italien mit 27 860 $ pro Einwohner zwar fast das gleiche Bruttonationaleinkommen wie Deutschland, ist aber deutlich korrupter. Es gibt also sicherlich keinen deterministischen Zusammenhang zwischen Korruption und Wohlstand, aber doch eine Tendenz.

Falls wir nun versuchen, beispielsweise einen linearen Zusammenhang der Form

$$CPI = a + b \cdot BNE$$

zu finden, so haben wir 142 Gleichungen für die beiden unbekannten

Koeffizienten a und b, und ein kurzer Blick auf Abbildung 18 zeigt, daß dieses lineare Gleichungssystem keine Lösung haben kann.

Wir suchen also keine Lösung, sondern zwei Zahlen a und b derart, daß die 142 Gleichungen „möglichst gut“ gelten. Was das bedeuten soll läßt sich mathematisch auf verschiedene, nicht äquivalente Weisen definieren; da wir uns im Augenblick mit Skalarprodukten beschäftigen, bietet sich an, die 142 Bruttoinlandsprodukte pro Einwohner und die 142 Korruptionsindizes zu zwei Vektoren $\vec{x}, \vec{y} \in \mathbb{R}^{142}$ zusammenzufassen, und nach Zahlen a, b zu suchen, so daß die Länge des Differenzvektors $\vec{y} - a\vec{x} - b$ möglichst klein wird. Ausgeschrieben bedeutet dies, wenn wir die Komponenten von $\vec{x}$ mit x_i und die von $\vec{y}$ mit y_i bezeichnen, daß die Summe

$$\sum_{i=1}^{142} (y_i - ax_i - b)^2$$

der Abweichungsquadrate möglichst klein sein soll – von daher der Name „Methode der kleinsten Quadrate“ für diesen Ansatz, mit dessen Hilfe sein Schöpfer GAUSS sowohl die Position des Planetoiden Ceres vorhersagte als auch die Vermessung und Kartierung des Königreichs Hannover durchführte.

Derselbe Ansatz läßt sich natürlich auf jedes lineare Gleichungssystem über den reellen oder komplexen Zahlen anwenden: Wir haben ein möglicherweise unlösbares lineares Gleichungssystem $A\vec{x} = \vec{b}$ und wollen einen Vektor $\vec{x}$ so bestimmen, daß der Vektor $A\vec{x} - \vec{b}$ minimale Länge hat.

Falls das lineare Gleichungssystem lösbar ist, gibt es damit kein Problem: Wir bestimmen irgendeine Lösung $\vec{x}$ und haben damit einen Vektor gefunden, für den $A\vec{x} - \vec{b}$ die Länge null hat – kürzer geht es nicht.

Im allgemeinen ist aber für den gesuchten Vektor $\vec{x}$ das Produkt $A\vec{x}$ von $\vec{b}$ verschieden; es sei etwa gleich $\vec{c}$. Dann ist $\vec{c}$ ein Vektor, der sich in der Form $A\vec{x}$ darstellen läßt, und unter allen solchen Vektoren ist es derjenige, für den die Länge des Differenzvektors zu $\vec{b}$ minimal ist. Dies erinnert an die orthogonalen Projektionen aus dem vorigen Abschnitt, und in der Tat läßt sich das Problem damit lösen:

Nehmen wir an, wir haben n Gleichungen in m Unbekannten mit Koeffizienten aus $k = \mathbb{R}$ oder $k = \mathbb{C}$. Dann definiert die Matrix $A \in k^{n\times m}$ des Gleichungssystems eine lineare Abbildung

$$\varphi\colon k^m \to k^n; \quad \vec{v} \mapsto A\vec{v};$$

deren Bildraum sei U. Falls die rechte Seite $\vec{b}$ in U liegt, ist das Gleichungssystem lösbar; andernfalls suchen wir einen Vektor $\vec{x} \in k^m$, für den die Länge des Vektors $A\vec{x} - \vec{b}$ minimal wird. Da die Vektoren, die sich in der Form $A\vec{x}$ darstellen lassen, genau die Vektoren aus U sind, ist somit $A\vec{x} = \pi_U(\vec{b})$ die orthogonale Projektion von $\vec{b}$ nach U. Diese könnten wir *im Prinzip* bestimmen, indem wir die QR-Zerlegung von A berechnen, denn dann sind die ersten Spalten von Q eine Basis von U, die durch die weiteren Spalten zu einer Basis von ganz k^n ergänzt wird; danach haben wir ein lösbares lineares Gleichungssystem.

Wir wollen uns überlegen, wie wir $\vec{x}$ auch ohne die rechnerisch aufwendige QR-Zerlegung bestimmen können.

Für den gesuchten Vektor $\vec{x}$ (oder für die gesuchten Vektoren $\vec{x}$) ist $A\vec{x} = \varphi_U(\vec{b})$. Da $A\vec{x}$ bereits in U liegt, ist $\pi_U(A\vec{x}) = A\vec{x}$, also ist die Gleichung $A\vec{x} = \pi_U(\vec{b})$ äquivalent zu

$$\pi_U(A\vec{x}) = \pi_U(\vec{b}) \quad \text{oder} \quad A\vec{x} - \vec{b} \in \operatorname{Kern} \pi_U = U^\perp .$$

Das orthogonale Komplement $U^\perp$ von U besteht aus allen Vektoren $\vec{y} \in k^n$, die senkrecht stehen auf U, für die also gilt

$$(A\vec{x}) \cdot \vec{y} = 0 \quad \text{für alle } \vec{x} \in k^m .$$

Wie wir im vorletzten Abschnitt gesehen haben, ist

$$(A\vec{x}) \cdot \vec{y} = \vec{x} \cdot A^*\vec{y} \quad \text{für alle } \vec{x} \in k^m, \vec{y} \in k^n ,$$

$\vec{y}$ liegt also genau dann in $U^\perp$, wenn $A^*\vec{y}$ senkrecht steht auf allen Vektoren $\vec{x} \in k^m$. Ein solcher Vektor aus k^m ist insbesondere $A^*\vec{y}$ selbst; wegen der positiven Definitheit des (HERMITEschen) Skalarprodukts ist also $A^*\vec{y} = \vec{0}$. Da aus $A^*\vec{y} = \vec{0}$ für alle $\vec{x} \in k^m$ folgt, daß $\vec{x} \cdot A^*\vec{y}$ verschwindet, ist damit

$$U^\perp = \left\{ \vec{y} \in k^n \;\middle|\; A^*\vec{y} = \vec{0} \right\} .$$

$A\vec{x} - \vec{b}$ liegt also genau dann im Kern von π_U, wenn $A^*(A\vec{x} - \vec{b}) = \vec{0}$ ist oder, anders ausgedrückt, wenn $\vec{x}$ eine Lösung des linearen Gleichungssystems

$$(A^*A)\,\vec{x} = A^*\vec{b}$$

ist. Da die adjungierte Matrix A^* einfach die transponierte Matrix zur komplex konjugierten Matrix zu A ist, wobei die komplexe Konjugation über $\mathbb{R}$ natürlich entfällt, läßt sich dieses Gleichungssystem schnell aufstellen und dann nach GAUSS lösen.

Betrachten wir dies konkret im eingangs diskutierten Fall eines linearen Zusammenhangs $y = ax + b$ zu N Wertepaaren $(x_i, y_i) \in \mathbb{R}^2$, wobei N sinnvollerweise größer als zwei sein sollte. Wir haben dann N Gleichungen

$$y_i = ax_i + b \quad \text{oder} \quad x_i a + b = y_i\,,$$

wobei hier im Gegensatz zu unserer sonstigen Gewohnheit die Parameter a und b unbekannt sind, während die x_i und die y_i bekannt sind. Wir haben also ein lineares Geichungssystem von N Gleichungen in den beiden Variablen a und b.

Fassen wir die Werte x_i zusammen zu einem Vektor $\vec{x} \in \mathbb{R}^N$ und die y_i zu einem Vektor $\vec{y} \in \mathbb{R}^n$, so läßt sich dieses Gleichungssystem kurz schreiben als

$$\vec{x} \cdot a + \vec{1} \cdot b = \vec{y}\,,$$

wobei $\vec{1} \in \mathbb{R}^N$ jenen Vektor bezeichnen soll, dessen sämtliche Komponenten eins sind.

Die Matrix des Gleichungssystems ist somit die $N \times 2$-Matrix A mit Spalten $\vec{x}$ und $\vec{1}$. Da wir mit reellen Zahlen rechnen, ist A^* einfach die transponierte Matrix dazu, also die $2 \times N$-Matrix, in deren erster Zeile die x_i stehen, während in der zweiten lauter Einser stehen. Somit ist

$${}^tAA = \begin{pmatrix} \vec{x}\cdot\vec{x} & \vec{x}\cdot\vec{1} \\ \vec{x}\cdot\vec{1} & \vec{1}\cdot\vec{1} \end{pmatrix} \quad \text{und} \quad {}^tA\vec{b} = \begin{pmatrix} \vec{x}\cdot\vec{y} \\ \vec{1}\cdot\vec{y} \end{pmatrix}\,,$$

das Gleichungssystem wird also zu

$$(\vec{x}\cdot\vec{x})a + (\vec{x}\cdot\vec{1})b = \vec{x}\cdot\vec{y} \quad \text{und} \quad (\vec{x}\cdot\vec{1})a + Nb = \vec{1}\cdot\vec{y}\,.$$

Seine Matrix ist genau dann singulär, wenn die Determinante verschwindet, wenn also $N(\vec{x}\cdot\vec{x}) = (\vec{x}\cdot\vec{1})^2$ ist. Nach der CAUCHY-SCHWARZschen Ungleichung ist

$$\left|\vec{1}\cdot\vec{x}\right| \le \left|\vec{1}\right|\cdot|\vec{x}| = \sqrt{N}\,|\vec{x}|\,, \quad \text{also} \quad \left|\vec{1}\cdot\vec{x}\right|^2 \le N(\vec{x}\cdot\vec{x})$$

mit Gleichheit nur dann, wenn die Vektoren $\vec{x}$ und $\vec{1}$ linear abhängig sind, wenn also alle x_i denselben Wert x haben. In diesem Fall ist die erste Gleichung das x-fache der zweiten, es gibt also unendlich viele Lösungen.

Andernfalls ist die Matrix invertierbar, die Lösung also eindeutig.

Führen wir die (in der Ausgleichsrechnung ziemlich verbreiteten) Abkürzungen

$$[x^r] = \sum_{i=1}^{N} x_i^r\,, \quad [y^r] = \sum_{i=1}^{N} x_i^r \quad \text{und} \quad [x^r y^s] = \sum_{i=1}^{N} x_i^r y_i^s$$

ein, so erhält das Gleichungssystem die übersichtlichere Gestalt

$$[x^2]a + [x]b = [xy] \quad \text{und} \quad [x]a + Nb = [y]\,.$$

Subtraktion von $[x]/[x^2]$ mal der ersten Gleichung von der zweiten führt auf

$$\left(N - \frac{[x]^2}{[x^2]}\right) b = [y] - \frac{[x]}{[x^2]}[xy]$$

oder $(N[x^2] - [x]^2)b = [y][x^2] - [x][xy]$, d.h.

$$b = \frac{[y][x^2] - [x][xy]}{N[x^2] - [x]^2}\,.$$

(Man beachte, daß im Falle der eindeutigen Lösbarkeit sowohl $[x^2] > 0$ als auch $N[x^2] - [x]^2 > 0$ ist.)

Einsetzen von b in die erste Gleichung ergibt dann auch

$$a = \frac{[xy] - [x]b}{[x^2]}\,.$$

Im Falle des Zusammenhangs zwischen Korruptionsindex CPI und Bruttonationaleinkommen pro Einwohner BNE erhalten wir nach diesen Formeln die Ausgleichsgerade

$$\text{CPI} = 2{,}29265 + 0{,}00017682\, i\text{BNE}\,,$$

die Steigung ist also erwartungsgemäß positiv. Der relativ große konstante Term zeigt, daß *im Mittel* Korruption selbst bei sehr armen Ländern deutlich über dem unteren Ende der Skala liegt. Abbildung 19 zeigt die Ausgleichsgerade zusammen mit den Daten.

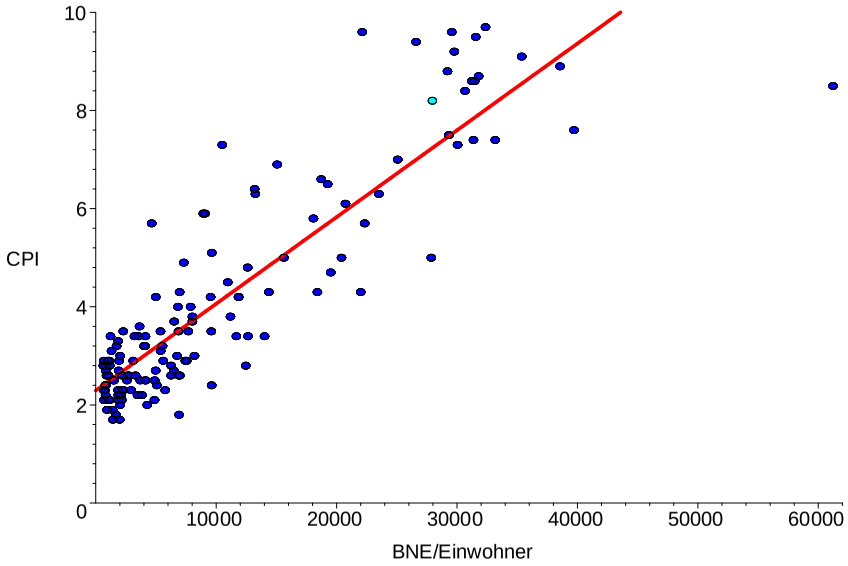

Abb. 19: Ausgleichsgerade zu Abbildung 18

Natürlich sind die Datenpunkte relativ breit gestreut um die Ausgleichsgerade; der Zusammenhang zwischen Korruption und Wohlstand ist schließlich zum Glück kein unausweichliches deterministisches Gesetz, sondern nur eine empirische Beobachtung.

Auch bei Messungen physikalischer Größen, wo die verschiedenen Meßgrößen meist durch wohlbekannte Naturgesetze miteinander verbunden sind, gibt es praktisch immer eine Streuung der Daten um die theoretisch richtige Meßkurve; absolut fehlerfreie Messungen sind, trotz aller Mühe der Experimentatoren, fast nie möglich, da es praktisch immer ein Grundrauschen der Meßgeräte und/oder nicht in ihrer Gesamtheit erfaßbare Umgebungseinflüsse *usw.* gibt. Vor allem bei Messungen, mit denen Konstanten für Naturgesetze ermittelt werden sollen oder gar ein Experiment zwischen zwei oder mehr Hypothesen entscheiden soll, ist es daher wichtig zu wissen, wie gut die Übereinstimmung zwischen den Daten und der berechneten Kurve (oder Fläche *usw.*) wirklich ist.

Solche Maße stellt die Statistik zur Verfügung; für ihr Verständnis sind daher meist zumindest Grundlagenkenntnisse der Statistik notwendig, wie wir sie (wenn auch nur kurz) im nächsten Semester behandeln werden. Im einfachsten und zugleich wichtigsten Fall eines linearen Zusammenhangs zwischen zwei Größen allerdings reicht die lineare Algebra, um das sowohl in der Theorie wie auch den Anwendungen wichtigste Qualitätsmaß zu definieren, den Korrelationskoeffizienten.

Angenommen, wir haben N Datenpaare (x_i, y_i), zwischen denen ein perfekter linearer Zusammenhang besteht, d.h.

$$y_i = ax_i + b \quad \text{für alle } i = 1, \ldots, N\,.$$

Wir wollen den Datenvektoren $\vec{x} \in \mathbb{R}^n$ mit Komponenten x_i und $\vec{y} \in \mathbb{R}^N$ mit Komponenten y_i Vektoren zuordnen, die nicht nur in einem linearen Zusammenhang stehen, sondern sogar gleich sind; mit anderen Worten, wir wollen die Parameter a und b aus obiger Gleichung eliminieren.

Dazu betrachten wir als erstes die Mittelwerte

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{und} \quad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i\,.$$

Da $y_i = ax_i + b$ ist für alle i, folgt sofort, daß auch $\overline{y} = a\overline{x} + b$ ist, und damit $(y_i - \overline{y}) = a(x_i - \overline{x})$ für alle $i = 1, \ldots, N$.

Damit ist der Parameter b eliminiert. Bezeichnen wir wieder mit $\vec{1} \in \mathbb{R}^N$ den Vektor, dessen sämtliche Komponenten Einser sind, ist nun also $(\vec{y} - \overline{y}\vec{1}) = a(\vec{x} - \overline{x}\vec{1})$. Aus dieser Gleichung können wir nun leicht a bis

auf sein Vorzeichen eliminieren, indem wir die beiden Vektoren durch ihre Länge dividieren. Dies ist natürlich nur möglich, wenn keiner der beiden Vektoren gleich dem Nullvektor ist, wenn also nicht alle $x_i = \overline{x}$ oder alle $y_i = \overline{y}$ sind. Bei nicht getürkten Messungen ist dies allerdings *praktisch* nie der Fall, so daß die Nützlichkeit der folgenden Diskussion und Definition nicht darunter leidet, daß wir diesen Fall ausschließen müssen.

Falls also weder $\vec{y} - \overline{y}\vec{1}$ noch $\vec{x} - \overline{x}\vec{1}$ der Nullvektor ist, betrachten wir die beiden auf Länge eins normierten Vektoren

$$\frac{\vec{y} - \overline{y}\vec{1}}{\left|\vec{y} - \overline{y}\vec{1}\right|} \quad \text{und} \quad \frac{\vec{x} - \overline{x}\vec{1}}{\left|\vec{x} - \overline{x}\vec{1}\right|} .$$

Diese sind nun offensichtlich entweder gleich (für $a > 0$) oder entgegengesetzt gleich (für $a < 0$).

Wenn (wie in der Realität meist der Fall) *kein* perfekter linearer Zusammenhang zwischen den x_i und den y_i besteht, können wir trotzdem – falls weder $\vec{y} - \overline{y}\vec{1}$ noch $\vec{x} - \overline{x}\vec{1}$ der Nullvektor ist – die beiden Vektoren

$$\frac{\vec{y} - \overline{y}\vec{1}}{\left|\vec{y} - \overline{y}\vec{1}\right|} \quad \text{und} \quad \frac{\vec{x} - \overline{x}\vec{1}}{\left|\vec{x} - \overline{x}\vec{1}\right|}$$

betrachten. Da beides Einheitsvektoren sind, unterscheiden sie sich nur in der Richtung; als Maß für ihren Unterschied bietet sich daher den Winkel zwischen $\vec{x}$ und $\vec{y}$ an. Rechnerisch einfacher ist der Cosinus dieses Winkels, denn der ist bei Einheitsvektoren einfach gleich dem Skalarprodukt.

Definition: Der Korrelationskoeffizient zwischen zwei Datenvektoren $\vec{x}$ und $\vec{y} \in \mathbb{R}^n$, die keine Vielfachen des Vektors $\vec{1} \in \mathbb{R}^N$ sind, ist

$$\rho = \frac{(\vec{x} - \overline{x} \cdot \vec{1}) \cdot (\vec{y} - \overline{y} \cdot \vec{1})}{\left|\vec{x} - \overline{x} \cdot \vec{1}\right| \cdot \left|\vec{y} - \overline{y} \cdot \vec{1}\right|} .$$

Damit ist also $\rho = \pm 1$ genau dann, wenn es einen perfekten linearen Zusammenhang $y_i = ax_i + b$ zwischen den beiden Größen gibt, mit $\rho = 1$ für $a > 0$ und $\rho = -1$ für $a < 0$. Ansonsten ist der Zusammenhang

umso besser, je größer der Betrag von ρ ist. Für $\rho = 0$ stehen die beiden Vektoren $\vec{x} - \overline{x}\vec{1}$ und $\vec{y} - \overline{y}\vec{1}$ senkrecht aufeinander, d.h. wenn x_i größer ist als der Mittelwert $\overline{x}$, kann y_i im Mittel genauso gut größer wie auch kleiner als der Mittelwert $\overline{y}$ sein. (in der Statistik ist dies die *Definition* für die Unabhängigkeit von Daten.)

Definition: Zwei Größen x und y heißen $\left\{\begin{matrix}\text{positiv}\\\text{negativ}\end{matrix}\right\}$ korreliert, wenn $\rho \left\{\begin{matrix}>\\<\end{matrix}\right\} 0$ ist. Sie heißen unkorreliert oder voneinander unabhängig, wenn $\rho = 0$ ist.

Im Beispiel der Korruption erhalten wir einen Korrelationskoeffizienten von $\rho \approx 0{,}885395$; dies entspricht einem Winkel von etwa $27{,}7^\circ$ zwischen den oben definierten Vektoren.

Um ein Gefühl für Korrelationskoeffizienten zu bekommen, wollen wir zwei Beispiele betrachten, die sich zumindest visuell sehr unterscheiden: Der CPI für Deutschland hatte in den letzten Jahren folgende Werte:

Jahr:	1980–1985	1988–1992	1995	1996	1997	1998	1999
CPI:	8,14	8,13	8,14	8,27	8,23	7,9	8,0
Jahr:	2000	2001	2002	2003	2004	2005	
CPI:	7,6	7,4	7,3	7,7	8,2	8,2	

Wie Abbildung 20 zeigt, sieht der Zusammenhang zwischen Jahr und CPI nicht sonderlich linear aus: Der Bimbesknick ist unverkennbar, jedoch scheint die Talsohle inzwischen durchschritten, so daß die abwärts gehende Ausgleichsgerade wohl er nicht den derzeitigen Trend beschreibt. Der Korrelationskoeffizient $\kappa \approx -0{,}357$ ist demnach auch ziemlich schlecht: Er ist der Kosinus eines Winkels von knapp 111°.

Vergleichen wir dagegen die Mannheimer Ergebnisse von Europawahl und Gemeinderatswahl vom 13. Juni 2004 miteinander, so gibt es bei keiner der vier Parteien, die zu beiden Wahlen angetreten ist, dramatische Unterschiede zwischen ihrem Stimmanteil bei den beiden Wahlen, obwohl gewisse Abweichungen unverkennbar sind.

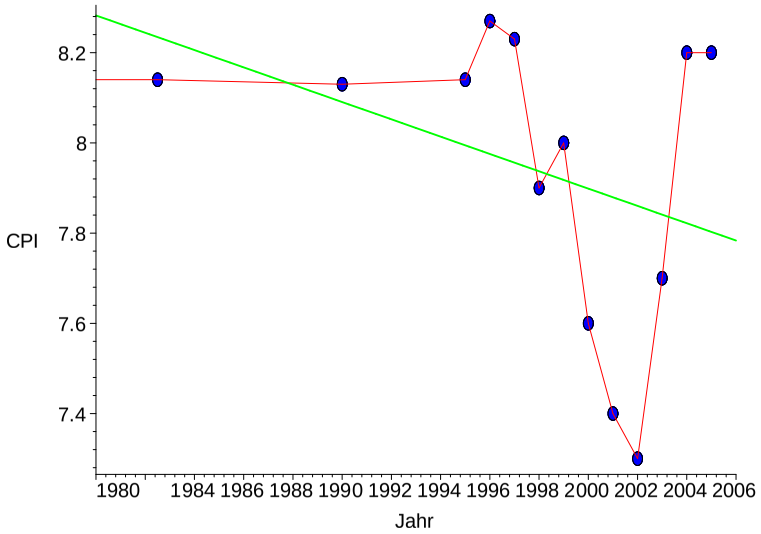

Abb. 20: Zeitabhängigkeit des CPI für Deutschland

	Europawahl	*Gemeinderatswahl*
CDU	38,14	40,41
SPD	28,91	33,38
Grüne	14,72	10,19
FDP	5,86	3,43

Wie Abbildung 21 zeigt, kann man den Zusammenhang in recht guter Näherung durch eine Gerade beschreiben, und in der Tat erhalten wir hier $\kappa \approx 0{,}9900614$, was einem Winkel von etwa acht Grad entspricht.

Korrelationskoeffizienten mit kleinem Betrag müssen nicht unbedingt bedeuten, daß kein deterministischer Zusammenhang zwischen den Daten besteht: Sie besagen nur, daß es keinen *linearen* Zusammenhang gibt. Betrachtet man etwa Wertepaare $(x_i, \sin x_i)$, so erhält man für Werte x_i, die einigermaßen gleichmäßig über eine oder mehrere Perioden der Sinusfunktktion verteilt sind, einen Korrelationskoeffizienten nahe Null, obwohl der Zusammenhang zwischen den beiden werten eines jeden Paars strikt deterministisch ist. In so einem Fall ist einfach der lineare Ansatz die falsche Strategie und man muß alternative Ansätze finden.

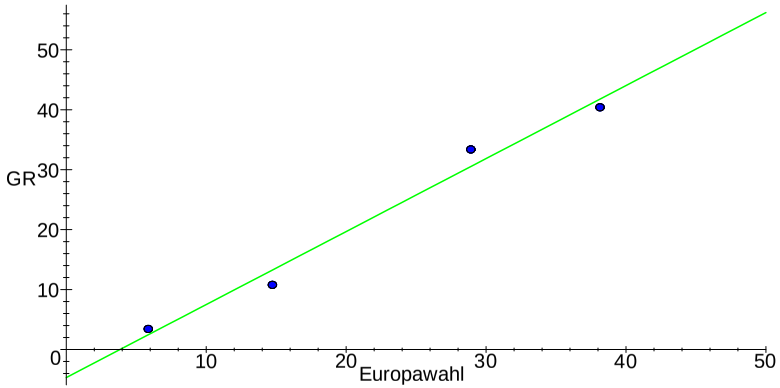

Abb. 21: Zusammenhang Europawahl/Gemeinderatswahl

j) Euklidische Vektorräume in der Informationssuche

In §4*j)* haben wir gesehen, wie Google die Unzahl von Internetseiten völlig unabhängig von jeder Suchanfrage nach ihrer Wichtigkeit ordnet. Hier soll es kurz skizziert werden, wie die Lineare Algebra auch hilft beim Problem, zu einer Suchanfrage geeignete Dokumente zu finden. Dabei muß es nicht unbedingt um Suche im Internet gehen; mindestens genauso wichtig sind wissenschaftliche Literaturdatenbanken, in denen zumindest Tausende (meist deutlich mehr) wissenschaftlicher Arbeiten gespeichert sind, aus denen ein Anwender die für seine Forschung relevanten finden möchte. Am einfachsten geht das, wenn entweder der Autor oder ein Berichterstatter die Arbeiten nach Themengebieten ordnet: In der Mathematik etwa gibt es dazu ein umfangreiches Klassifikationsschema der beiden westlichen Referatorgane *Zentralblatt für Mathematik und ihre Grenzgebiete* und *Mathematical Reviews*, das auch die meisten Fachzeitschriften verwenden; in anderen Wissenschaften ist es ähnlich.

Solche Zuordnungen sind meistens recht genau, da sie von Experten der jeweiligen *Teil*gebiete vorgenommen werden; andererseits birgt natürlich auch gerade das die Gefahr in sich, daß eine Arbeit, die für mehrere Gebiete relevant ist, möglicherweise nur denen zugeordnet wird, für

die sich der Autor oder Berichterstatter interessiert. Außerdem ist selbst eine sehr detailierte Einteilung, die im Falle der Mathematik immerhin 35 Seiten Kleingedrucktes benötigt, immer noch zu grob, um genau *die* drei Arbeiten zu finden, in denen ein sehr spezielles Problem behandelt wird. Im Internet mit seiner Vielzahl von teilweise sehr schnell variierenden Informationsangeboten ist ein solcher Ansatz von vornherein chancenlos.

Zusätzlich zur Klassifikation durch menschliche Experten braucht man daher bei der Informationssuche auch Algorithmen, die gelesene Informationen automatisch klassifizieren und bezüglich ihrer Relevanz zu einer konkreten Suchanfrage beurteilen können.

Große Internetsuchmaschinen verwenden dazu eine Vielzahl von Algorithmen; mit Ausnahme von `google.com`, die ihr Rangbildungsverfahren unter

`http://www.google.com/technology/pigeonrank.html`

mehr oder weniger ausführlich beschreiben, schweigen sie sich allerdings aus über die genauen Einzelheiten und Parameter: Schließlich sollen die vielen unseriösen Anbieter, die mit allen Tricks Besucher auf ihre Webseiten locken wollen, nicht auch noch unterstützt werden.

Wir müssen uns daher auf die grundlegenden mathematischen Algorithmen beschränken, die wohl in der einen oder andereren Form in praktisch jeer Suchmaschine zu finden sind und die, als Gegenstand wissenschaftlicher Forschung, natürlich öffentlich bekannt sind.

Die ersten Systeme arbeiteten mit den üblichen Suchalgorithmen aus der Textverarbeitung, durchsuchten also alle gespeicherten Dokumente nach dem Vorkommen einer oder mehrerer vorgegebener Zeichenketten. Auch wenn es dafür sehr effiziente Algorithmen gibt, ist dieses Verfahren bei wirklich großen Datenmengen nicht mehr mit realistischem Aufwand durchführbar, so daß nun meist Verfahren aus der linearen Algebra verwendet werden.

Dazu wird eine Liste von Suchbegriffen s_i, $i = 1, \cdots, n$ festgelegt – beispielsweise die Wörter aus einem Wörterbuch der Dokumentsprache. Oftmals werden darauf noch geeignete Operationen angewandt wie

stemming, d.h. Wörter mit gleichem Stamm werden miteinander identifiziert, oder *latent semantic indexing*, wo durch Clusterbildung bei den vorhandenen Dokumenten Begriffspaare identifiziert werden, die im allgemeinen im gleichen Kontext auftreten und die dann auch bei Suchanfragen als äquivalent betrachtet werden; außerdem werden sogenannte „Nullwörter“, die für Suchanfragen typischerweise ohne Bedeutung sind, eliminiert. Dabei handelt es sich beispielsweise um Artikel und Praepositionen, gelegentlich aber auch um spezifische Wörter aus dem Kontext des jeweiligen Systems: Bei Boeing, die ein solches System zur Verwaltung ihrer Wartungshandbücher aufbauten, ist etwa das Wort „aeroplane“ ein Nullwort – die Gesellschaft verkauft schließlich keine Rasenmäher.

Sind nun m Dokumente zu betrachten, so bildet man eine $n \times m$-Matrix A, deren Eintrag a_{ij} etwas über das Vorkommen des i-ten Suchbegriffs im j-ten Dokument aussagt. Im einfachsten Fall setzt man einfach $a_{ij} = 1$, falls der Begriff vorkommt und null sonst, alternativ kann a_{ij} auch die Häufigkeit des Begriff im Dokument sein, wobei diese Häufigkeit oft noch gewichtet wird, indem beispielsweise Vorkommen im vorderen Teil des Dokuments höher gewichtet wird oder aber die Suchmaschine ohnehin nur den Anfangsteil des Dokuments bis zu einer gewissen Maximallänge berücksichtigt. Auch das Vorkommen in Überschriften oder zwischen `<META>`-tags kann eventuell gesondert behandelt werden, indem man beispielsweise Inhalte, die im Browserfenster nicht sichtbar werden, wegen der damit verbundenen Mißbrauchsmöglichkeit ignoriert. Gelegentlich wird auf das Ergebnis noch eine Skalierungsfunktion wie etwa $\log(1 + x)$ angewendet.

Die entstehende Matrix ist natürlich riesig; schon 1998 wurde geschätzt, daß allein für englischsprachige Dokumente bis zu 300 000 Suchbegriffe notwendig sind, die in etwa 300 Millionen Dokumenten gesucht werden müssen; die Matrix hat also knapp hundert Billionen Einträge. Bei nur einem Byte pro Einträg hätte man also bei der Speicherung als Feld einen Platzbedarf von etwa 90 Terabyte.

Nun kommt allerdings in fast jedem Dokument nur ein verschwindend geringer Bruchteil der Suchbegriffe vor, so daß die meisten Einträge von A Nullen sind. Die Matrix läßt sich daher erheblich kompakter

speichern, wenn man beispielsweise nur die Tripel (i, j, a_{ij}) notiert, für die $a_{ij} \neq 0$ ist. Die numerische Mathematik kennt eine ganze Reihe von Algorithmen, mit denen man auch solche sogenannte „spärlich besetzte“ Matrizen effizient behandeln kann.

Der Inhalt des j-ten Dokuments wird nun also kodiert durch den j-ten Spaltenvektor der Matrix A, einen Vektor aus $\mathbb{R}^n$. Auch eine Suchanfrage läßt sich durch einen solchen Vektor kodieren, indem man die j-te Komponente auf eins setzt, falls der j-te Suchbegriff in der Anfrage vorkommt, und auf null sonst. (Man kann natürlich auch andere Werte wählen und beispielsweise seltene Wörter höher gewichten als häufige *usw.*)

Ein Dokument sollte umso besser zu einer Suchanfrage passen, je weniger sich die dazu gehörigen Vektoren voneinander unterscheiden. Als Maß für den Unterschied zweier Vektoren haben wir im vorigen Abschnitt den Cosinus des eingeschlossenen Winkels kennengelernt; falls man die Spaltenvektoren der Matrix auf Länge Eins normiert, läst sich dieser durch eine einziges Skalarprodukt berechnen. Ein Dokument wird dann als relevant für die Suchanfrage betrachtet, wenn dieser Wert über einer festzulegenden Schranke liegt, und die so gefundenen Dateien können dann eventuell noch mit anderen Methoden (Volltextsuche, Links von anderern Seiten, ...) weiter untersucht werden zur Festlegung der endgültigen Reihenfolge, in der sie dem Benutzer gezeigt werden.

Für sehr große Datenmengen ist allerdings die Matrix A trotz ihrer spärlichen Besetzung immer noch zu groß; wie bei der Komprimierung von Bilddaten sucht man daher nach einer Art und Weise, sie bei möglichst geringem Informationsverlust deutlich zu komprimieren. Ein angenehmer Nebeneffekt dabei ist, wie experimentelle Untersuchungen zeigen, auch eine gewisse „Rauschunterdrückung“: Es ist zwar schwierig, exakt zu definieren, was „Rauschen“ in einer Term-Dokument-Matrix sein soll, aber jeder wird wohl damit übereinstimmen, daß etwa dieses Skriptum nicht die ideale Referenz zum Thema „Rasenmäher“ ist, obwohl dieses Wort hier nun schon zum zweiten Mal vorkommt.

Einen Ansatz zur Datenreduktion liefert die QR-Zerlegung: Ist $A = QR$

und $\vec{a} \in \mathbb{R}^n$ eine Suchanfrage, so ist für die j-ten Spalten $\vec{a}_j$ von A und $\vec{r}_j$ von R

$$\vec{a} \cdot \vec{a}_j = \vec{a} \cdot (Q\vec{r}_j) = ({}^tQ\vec{a}) \cdot \vec{r}_j \,,$$

und die Matrix R wird im allgemeinen deutlich mehr Nullen enthalten als A, da der Rang von A wohl deutlich unter n liegen dürfte. Eine weitere Komprimierung wird dadurch erreicht, daß Einträge von R, die unterhalb einer gewissen Schranke liegen, auf Null gesetzt werden; dadurch ändert sich bei hinreichend kleiner Schranke an den meisten Skalarprodukten nicht viel, dafur verringert sich aber der Speicherbedarf noch einmal beträchtlich.

Oft verwendet man anstelle der QR-Zerlegung auch die hier nicht behandelte Singulärwertzerlegung von A: Danach läßt sich A schreiben als Produkt UDV mit orthogonalen Matrizen $U \in \mathbb{R}^{n\times n}$ und $V \in \mathbb{R}^{m\times m}$ sowie einer Diagonalmatrix $D \in \mathbb{R}^{n\times m}$. U und V können so gewählt werden, daß die Diagonaleinträge der Größe nach angeordnet sind, und man erhält die gewünschte Rangreduktion, indem man alle Einträge unterhalb einer gewissen Größe auf Null setzt.

Eine ausführlichere Darstellung der Verfahren zur Textsuche, die keine über den Inhalt dieses Skriptums hinausgehende Mathematikkenntnisse voraussetzt, findet man beispielsweise in

MICHAEL W. BERRY, MURRAY BROWNE: Understanding Search Engines: Mathematical Modeling and Text Retrieval, SIAM, 1999, ^{2}2005,

Fallstudien im Tagungsband

MICHAEL W. BERRY [Hrsg.]: Computational Information Retrieval, SIAM, 2001.