6 Approximation von Daten, Ausgleichsrechnung

6.1 Lineare Datenmodelle

Ein Beispiel dazu hat schon Kapitel 5.4 gebracht. Die Vorlesungsfolien und die Übungen bringen weitere Beispiele dazu. Das best-angepasste Modell ergibt sich dabei immer aus der kleinste-Quadrate-Näherung an ein überbestimmtes Gleichungssystem.

Aber nicht immer liefert die Methode der kleinsten Fehlerquadrate eine plausible Anpassung. Einige wenige grob falsche Werte in den Daten („Ausreißer“) können das Ergebnis gewaltig verzerren. Im Kapitel 6.6 wird eine robuste Methode vorgestellt. MATLAB bietet in seinen Toolboxen verschiedene Methoden zur robusten Anpassung (*robust fit*) an.

6.2 Polynomiale Regression

Hier handelt es sich um einen wichtigen Spezialfall der linearen Datenmodelle, für den sich die allgemeinen Formeln etwas vereinfachen.

Eine typische Aufgabe zu diesem Abschnitt könnte lauten: Gegeben sind Messergebnisse für einen Satz von Temperaturwerten T und die entsprechenden Widerstandswerte R eines Temperaturfühlers. Der Zusammenhang zwischen R und T lässt sich näherungsweise in der Form $R = a + bT + cT^2$ beschreiben. Wenn für genau drei (verschiedene) T-Werte Daten vorliegen, lassen sich die drei Parameter a, b und c eindeutig bestimmen. Es ist aber sinnvoll, mehr Messungen durchführen, damit Messfehler der Einzelmessungen nicht so stark ins Gewicht fallen. Die Parameter a, b und c werden dann durch *Ausgleichsrechnung* (man sagt auch *Regression*) bestimmt.

Polynomiale Regression (Ausgleich durch ein Polynom)
Gegeben: $m + 1$ Wertepaare $(x_i, y_i), i = 0, \ldots, m$
Gesucht: $p(x)$, ein Polynom n-ten Grades, $n < m$, so dass die Summe der Fehlerquadrate

$$\sum_{i=0}^{m} (p(x_i) - y_i)^2$$

minimal wird. Locker formuliert: $y = p(x)$ approximiert möglichst gut die Datenpunkte.

Nicht immer ist ein Polynomansatz ein geeignetes Modell, aber oft lässt sich ein scheinbar komplizierteres Modell auf ein polynomiales Modell zurückführen (Beispiele in den Übungen).

Direkter Lösungsweg: Ansatz des Polynoms mit unbestimmten Koeffizienten,

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1} x^{n-1} + a_n x_n \ .$$

Einsetzen der gegebenen Wertepaare führt auf ein System von $m + 1$ linearen Gleichungen in den $n + 1$ unbekannten Koeffizienten $a_0, a_1, \ldots, a_n$.

Sofern $n < m$, liegt ein *überbestimmtes System* vor.

- klassische Lösung: Näherung nach der Methode der Normalengleichungen: lässt sich bei kleinen Beispielen mit Papier und Stift rechnen; bei großen Datenmengen Gefahr von Rundungsfehlern.

  Die Normalengleichungen sind aber nur dann eindeutig lösbar, wenn in den insgesamt $m+1$ Wertepaaren mindestens $n+1$ der x-Werte verschieden sind.

- moderner Lösungsweg: QR-Zerlegung. Praktisch nur am Rechner durchführbar. Bessere Konditionszahl, weniger anfällig für Rundungsfehler.

Dieser Lösungsweg (Ansatz mit unbestimmten Koeffizienten, Aufstellen des überbestimmten Systems, Bilden der Normalengleichungen) lässt sich für die polynomiale Regression etwas abkürzen. Setzt man

$$s_0 = m+1, \quad t_0 = \sum_{i=0}^{m} y_i$$

und

$$s_k = \sum_{i=0}^{m} x_i^k, \quad t_k = \sum_{i=0}^{m} x_i^k y_i \quad \text{für } k > 0,$$

so lassen sich, wie man leicht herleiten kann, die Normalengleichungen in folgender Gestalt schreiben:

$$\begin{array}{lllllll}
s_0 a_0 & + & s_1 a_1 & + & \ldots\ldots & + & s_n a_n = t_0 \\
s_1 a_0 & + & s_2 a_1 & + & \ldots\ldots & + & s_{n+1} a_n = t_1 \\
 & & & & \ldots\ldots & & \\
s_n a_0 & + & s_{n+1} a_1 & + & \ldots\ldots & + & s_{2n} a_n = t_n
\end{array}$$

Die Normalengleichungen können für größere n ziemlich schlecht konditioniert sein. Bessere Resultate lassen sich in solchen Fällen durch Entwicklung nach orthogonalen Polynomen erzielen.
Verwendet man beispielsweise als Datensatz die Punkte

$$(x, \exp(x)) \quad \text{für } x = 0;\ 0{,}01;\ 0{,}02; \ldots;\ 3{,}99; 4$$

und approximiert die Daten durch Regressionspolynome verschieden hohen Grades (Rechnung mit vierzehnstelliger Genauigkeit), so wird die Güte der Approximation vorerst mit steigendem Grad besser. Ab dem dreizehnten Grad aber wachsen die Fehler wieder an. Das aus den Normalengleichungen berechnete Polynom 25-ten Grades hat kaum mehr Ähnlichkeit mit der approximierten Funktion. Verwendet man orthogonale Polynome (Tschebyscheff-Polynome), treten diese numerischen Probleme nicht auf.

6.3 Ausgleichsgerade

Ein wichtiger Spezialfall der obigen Problemstellung: An $m+1$ (mehr als zwei) gegebene Datenpunkte soll eine Gerade mit Gleichung $y = a + bx$ so angepasst werden, dass die Summe der Fehlerquadrate minimal wird.

$$a = \frac{s_2 t_0 - s_1 t_1}{s_0 s_2 - s_1^2}$$

$$b = \frac{s_0 t_1 - s_1 t_0}{s_0 s_2 - s_1^2}$$

Das Finden einer Ausgleichsgeraden wird oft auch als „lineare Regression“ bezeichnet. Das ist aber ein mißverständlicher Terminus, weil er leicht zur Verwechslung mit „linearen Datenmodellen“ führt.

6.4 Nichtlineare Datenmodelle

Dazu gibt es Material auf den Vorlesungsfolien und ausführlicher in den Übungsunterlagen. Kurzfassung:Jacobi-Matrix bilden und iterieren. Im Unterschied zum Newton-Verfahren, das Sie schon kennen, ist das lineare Gleichungssystem mit der Jacobimatrix nun überbestimmt und wird näherungsweise im Sinn der kleinsten Fehlerquadrate gelöst.

Dieses Verfahren wird als *Gauß-Newton-Verfahren* bezeichnet.

Bei nichtlinearen Ausgleichsproblemen gibt es neben dem Gauß-Newton-Verfahren noch weitere Verfahren (z. B. Levenberg-Marquardt-Algorithmus). Damit beschäftigt sich die Optimierung als Teilgebiet der Angewandten Mathematik. Dieses Skript kann nicht näher darauf eingehen.

6.5 Warum „kleinste Quadrate“

Die Methode der kleinsten Quadrate minimiert die euklidische Länge des Residuenvektors. Man kann aber die Größe des Residuenvektors durchaus auch anders messen und entsprechend andere Minimalbedingungen fordern. Wichtige Beispiele: Die Summe der *Absolutbeträge* der Fehler soll minimal werden, oder der *maximale Fehler* soll minimal werden („Minimax-Approximation“). Zwei Gründe sprechen für die kleinsten Quadrate:

- Einfache Herleitung und Durchführung: die Minimalwert-Aufgabe lässt sich mit elementarer Differentialrechnung lösen und führt auf ein einfaches algebraisches Problem

- Statistische Überlegungen: Wenn die Daten mit unabhängigen, zufälligen, normalverteilten Fehlern mit gleicher Standardabweichung behaftet sind, sind kleinste Quadrate in gewissem Sinn optimal (genauer: Die Methode liefert eine *maximum likelihood*-Schätzung der Parameter). Umgekehrt gilt aber: wenn die Fehler in den Daten *nicht* normalverteilt etc. sind, dann sind kleinste Quadrate unter Umständen ziemlich schlecht; siehe unten.

- Weitere statistische Eigenschaften: Wenn man eine Abschätzung für die Genauigkeit der Daten hat, kann man auf die Genauigkeit der berechneten Modellparameter schließen.

  Angenommen, die Daten sind so skaliert, dass die Varianz der Messfehler gleich 1 ist. Sei $C = (A^T A)^{-1}$ die inverse Matrix des Systems der Normalengleichungen. Die Diagonalelemente von C sind die Varianzen der entsprechenden Modellparameter; die Elemente außerhalb der Hauptdiagonale sind die entsprechenden Kovarianzen.

Und was spricht dagegen?

- Die Methode reagiert empfindlich auf „Ausreißer“ in den Daten. Das Quadrieren der Fehler bestraft grosse Abweichungen streng. Deswegen ist die Methode der kleinsten Quadrate bereit, eine Kurve wild zu verzerren, um ein paar weit außen liegende Datenpunkte auch noch annähernd zu erreichen. Ein paar Ausreißer in ansonsten vernünftigen Daten können so eine völlig unsinnige Approximation bewirken.

6.6 Ausgleichsgerade mit Minimieren der absoluten Fehler

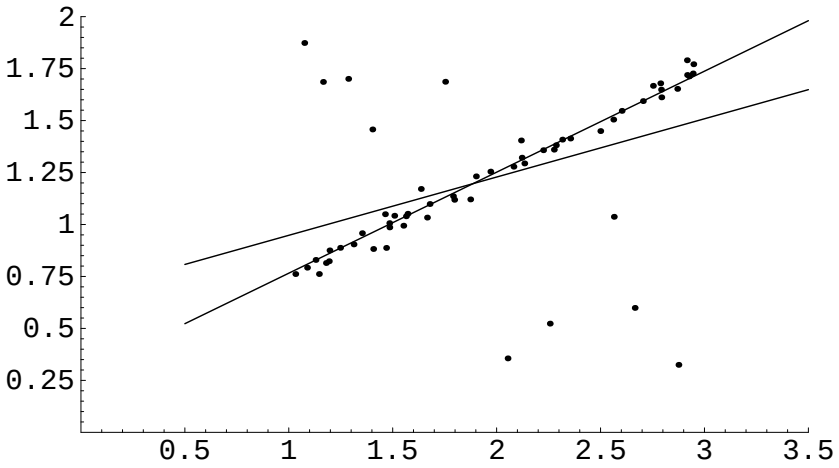

Abbildung 8: Anpassen einer Geraden an Datenpunkte. Die Ausgleichsgerade nach der Methode der kleinsten Quadrate lässt sich von den wenigen Ausreißern stark ablenken. Minimieren des absoluten Fehlers legt eine wesentlich plausiblere Gerade durch die Daten.

Gegeben: m Wertepaare $(x_i, y_i), i = 1, \ldots, m$
Gesucht: Eine Gerade in der Form $y = a + bx$, so dass die Summe der absoluten Fehler

$$\sum_{i=1}^{m} |p(x_i) - y_i|$$

minimal wird.
Die Lösung lautet (für die dahinterliegende Theorie sei nochmals auf das Buch *Numerical Recipes* verwiesen): Für gegebenes b ergibt sich a als Median eines Datenfeldes,

$$a = \text{median}\{y_i - bx_i\}$$

Den Parameter b findet man als Lösung der Gleichung

$$0 = \sum_{i=1}^{m} x_i \,\text{sgn}(y_i - a - bx_i)$$

(wobei sgn(0) als Null interpretiert werden soll). Wenn man für a in dieser Gleichung die durch die vorigen Gleichung bestimmte Funktion $a(b)$ einsetzt, bleibt eine Gleichung in einer Unbekannten übrig. Intervallhalbierung (siehe Kapitel 1.7) ist die geeignete Lösungsmethode dafür.

6.7 Ausgleichsgerade mit Minimieren der Normalabstände

Dazu gibt es unter dem Titel *Total Least Squares* Material auf den Vorlesungsfolien und den Übungsunterlagen.