

# Anpassungsrechnungen mit kleinsten Quadraten und Maximum Likelihood

Hauptseminar - Methoden der experimentellen Teilchenphysik  
WS 2011/2012

Fabian Hoffmann

2. Dezember 2011

## Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>2</b>
<b>2</b>	<b>Grundlagen</b>	<b>2</b>
2.1	Wahrscheinlichkeitsverteilung . . . . .	2
2.2	Erwartungswert und Varianz . . . . .	3
2.3	Zentraler Grenzwertsatz . . . . .	3
2.4	Eigenschaften von Schätzern . . . . .	4
<b>3</b>	<b>Maximum-Likelihood-Methode</b>	<b>4</b>
3.1	Fehlerberechnung . . . . .	5
3.2	Eigenschaften . . . . .	6
<b>4</b>	<b>Methode der kleinsten Quadrate</b>	<b>7</b>
4.1	Lineare kleinste Quadrate . . . . .	7
4.2	Fehlerfortpflanzung . . . . .	8
4.3	$\chi^2$ -Test . . . . .	9
4.4	Eigenschaften . . . . .	9
<b>5</b>	<b>Zusammenfassung</b>	<b>10</b>
<b>6</b>	<b>Literatur</b>	<b>10</b>

# 1 Einleitung

Eine wichtige Aufgabe von Experimentalphysikern ist die Auswertung von Daten. Gesucht sind zumeist die Parameter einer Theorie. Das Problem hierbei ist die Menge der Daten, im Vergleich zu den Parametern. Da es meist sehr viel mehr Daten als Parameter gibt, muss ein überbestimmtes Gleichungssystem gelöst werden. Dieses Problem wird mittels Anpassungsrechnung gelöst. Im Folgenden sollen zwei Verfahren vorgestellt werden:

- Maximum-Likelihood-Methode und
- Methode der kleinsten Quadrate.

Zunächst werden aber im Folgenden einige Grundlagen behandelt.

## 2 Grundlagen

### 2.1 Wahrscheinlichkeitsverteilung

Es gibt zwei grundlegende Arten von Wahrscheinlichkeitsverteilung. Zum einen die diskrete Wahrscheinlichkeitsverteilung, zum anderen die kontinuierliche.

Bei der diskreten Wahrscheinlichkeitsverteilung kann eine Zufallsvariable  $x$  nur diskrete Werte annehmen. Hierbei ist die Wahrscheinlichkeitsverteilung vollständig beschrieben, wenn zu allen der diskreten Werte  $x_i$  die Wahrscheinlichkeit  $c_i$  bekannt ist.

Bei der kontinuierlichen Wahrscheinlichkeitsverteilung kann eine Zufallsvariable  $x$  kontinuierlich Werte annehmen. Sie wird durch eine Wahrscheinlichkeitsdichtefunktion (engl. probability density function, pdf)  $f(x)$  beschrieben. Dabei ist die Wahrscheinlichkeit  $x$  zwischen  $a$  und  $b$  zu finden gegeben durch:

$$P(a < x < b) = \int_a^b f(x) dx$$

Wichtig ist, dass Wahrscheinlichkeitsverteilungen immer normiert sind. Das bedeutet

- für diskrete Wahrscheinlichkeit:  $\sum_i c_i = 1$
- für kontinuierliche Wahrscheinlichkeit:  $\int_{-\infty}^{\infty} f(x) dx = 1$

Im folgenden werden nur noch kontinuierliche Wahrscheinlichkeiten betrachtet. Die Betrachtungen lassen sich aber auch auf diskrete Wahrscheinlichkeitsverteilungen übertragen.

## 2.2 Erwartungswert und Varianz

Der **Erwartungswert**  $E[h(x)]$  einer Funktion  $h(x)$ , wobei  $x$  eine Zufallsvariable mit pdf  $f(x)$  ist, ist wie folgt definiert:

$$E[h(x)] = \int_{-\infty}^{\infty} h(x)f(x)dx$$

Der **Mittelwert** ist der Erwartungswert von  $x$ :

$$\langle x \rangle = E[x] = \int_{-\infty}^{\infty} xf(x)dx$$

**Varianz** von  $x$ :

$$V[x] = E[(x - \langle x \rangle)^2] = E[x^2] - E[x]^2$$

**Kovarianz** von  $x$  und  $y$ :

$$\text{cov}[x, y] = E[(x - \langle x \rangle)(y - \langle y \rangle)] = E[x \cdot y] - E[x]E[y]$$

Kovarianzmatrix:

$$\mathbf{V}[\mathbf{x}] = E[(\mathbf{x} - \langle \mathbf{x} \rangle)(\mathbf{x} - \langle \mathbf{x} \rangle)^T]$$
$$\mathbf{V} = \begin{pmatrix} \text{cov}(x_1, x_1) & \text{cov}(x_1, x_2) & \dots & \text{cov}(x_1, x_n) \\ \text{cov}(x_2, x_1) & \text{cov}(x_2, x_2) & \dots & \text{cov}(x_2, x_n) \\ \vdots & \vdots & \dots & \vdots \\ \text{cov}(x_n, x_1) & \text{cov}(x_n, x_2) & \dots & \text{cov}(x_n, x_n) \end{pmatrix}$$

## 2.3 Zentraler Grenzwertsatz

$S$  sei die Summe aus  $N$  unabhängigen Zufallsvariablen mit Mittelwert  $a_i$  und Varianz  $V_i$ . Dann gilt:

- $E[S] = \sum_{i=1}^N a_i$
- $V[S] = \sum_{i=1}^N V_i$
- Für  $N \rightarrow \infty$  ist die pdf von  $S$  eine Normalverteilung

## 2.4 Eigenschaften von Schätzern

- **Konsistenz**

Ein Schätzer  $\hat{\vec{a}}$  heißt konsistent, wenn er für größer werdende Stichproben, dem wahren Wert  $\vec{a}_0$  beliebig nahe kommt:

$$\lim_{n \rightarrow \infty} \hat{\vec{a}} = \vec{a}_0$$

Diese Eigenschaft wird im Allgemeinen von allen Schätzern gefordert.

- **Erwartungstreue**

Ein Schätzer ist eine Zufallsvariable, die einer gewissen Verteilung folgt. Bei einem erwartungstreuen Schätzer ist der Erwartungswert gleich dem wahren Wert.

$$E[\hat{a}] = a_0$$

Ein Schätzer, der nicht erwartungstreu ist, heißt verzerrt (engl. biased). Die Verzerrung  $b = E[\hat{a}] - a_0$  kann, wenn sie bekannt ist, korrigiert werden.

- **Effizienz**

Die Varianz eines Schätzers sollte möglichst klein sein, damit die Wahrscheinlichkeit, dass der Schätzer nahe am Erwartungswert liegt, möglichst groß ist. Je kleiner die Varianz, desto größer ist die Effizienz.

- **Robustheit**

Ein Schätzer ist robust, wenn er weder von falschen Daten, noch von falschen Annahmen beeinflusst wird. Meist wird Robustheit gegen Ausreißer in den Daten erreicht, indem ein gewisser Teil der Daten verworfen wird.

Die Eigenschaften Effizienz und Robustheit stehen oft in Widerspruch zueinander, sodass ein für das Problem angemessener Kompromiss gefunden werden muss.

### 3 Maximum-Likelihood-Methode

Für die Maximum-Likelihood-Methode wird die der Verteilung zugrunde liegende Wahrscheinlichkeitsdichte  $f(x, \mathbf{a})$  benötigt, wobei  $\mathbf{a}$  ein Vektor von zu bestimmenden Parametern ist. Die Likelihood-Funktion ist dann einfach das Produkt der Wahrscheinlichkeitsdichten der gemessenen Werte.

$$L(\mathbf{a}) = \prod_i f(x_i, \mathbf{a})$$

Die Likelihood-Funktion gibt die Wahrscheinlichkeit an, bei gegebenen Parametern  $\mathbf{a}$  die Messwerte  $\{x_i\}$  zu erhalten. Sie ist also eine Wahrscheinlichkeitsdichte in  $x$  und nicht in  $\mathbf{a}$ .

Bei der Maximum-Likelihood-Methode wird nun die Likelihood-Funktion bezüglich der Parameter  $\mathbf{a}$  maximiert.

$$L(\hat{\mathbf{a}}) = \text{Maximum}$$

$\hat{\mathbf{a}}$  ist also das  $\mathbf{a}$ , für das die Messwerte  $\{x_i\}$  am Wahrscheinlichsten sind. Es werden also nicht die Wahrscheinlichsten  $\mathbf{a}$  gefunden.

In der Praxis wird meist die negative Log-Likelihood-Funktion maximiert.

$$F(\mathbf{a}) = -\ln L(\mathbf{a}) = -\sum_i \ln f(x_i, \mathbf{a})$$

Der Logarithmus wird verwendet, da dieser für numerische Berechnungen wesentlich besser geeignet ist. Die Verwendung der negativen Funktion hat historische Gründe.

### 3.1 Fehlerberechnung

Bei großen Stichproben nähert sich die Likelihood-Funktion aufgrund des Zentralen Grenzwertsatzes einer Gaußverteilung an. Dies macht die Fehlerbestimmung einfach. Die negative Log-Likelihood-Funktion nähert sich einer Parabel, eine Taylorentwicklung kann somit beim quadratischen Term abgebrochen werden.

$$F(a) = F(\hat{a}) + \frac{1}{2} \cdot \frac{d^2 F}{da^2} \cdot (\hat{a} - a)^2 + \mathcal{O}(a^3)$$

Ein Vergleich mit dem Exponenten  $\frac{1}{2} \frac{(\hat{a}-a)^2}{\sigma^2}$  einer Gaußfunktion ergibt:

$$\Rightarrow F(\hat{a} \pm r \cdot \sigma) = F(\hat{a}) + \frac{r^2}{2}$$

Diese einfache Beziehung kann auch auf kleinere Stichproben angewandt werden, wenn Terme höhere Ordnung nicht vernachlässigt werden können. Hierbei entstehen dann jedoch im Allgemeinen asymmetrische Konfidenzgrenzen:

- $F(\hat{a} + r \cdot \sigma_+) = F(\hat{a}) + \frac{r^2}{2}$
- $F(\hat{a} - r \cdot \sigma_-) = F(\hat{a}) + \frac{r^2}{2}$

Ein Beispiel für beide Varianten ist in Abbildung 1 zu sehen.

### 3.2 Eigenschaften

Die Maximum-Likelihood-Methode ist, wie zu erwarten, konsistent. Allerdings ist sie im allgemeinen nicht erwartungstreu. Die Verzerrung nimmt aber mit steigender Anzahl von Messungen ab, sodass die Methode asymptotisch Erwartungstreu wird. Bei großen Messreihen kann also von Erwartungstreue ausgegangen werden, bei wenig Messwerten muss die Verzerrung aber gegebenenfalls berücksichtigt werden.

Eine wichtige Eigenschaft ist die maximale Effizienz der Methode. Kein Schätzer ist effizienter als der Maximum-Likelihood-Schätzer. Die geht jedoch wie oben erwähnt auf Kosten der Robustheit. Wenn eine falsche pdf zugrunde gelegt wird, dann ergeben sich für den Maximum-Likelihood-Schätzer extrem falsche Werte.

Ein weiterer Nachteil der Methode ist der oft große benötigte Rechenaufwand. Nur in seltenen Fällen gibt es eine analytische Lösung, meist muss das Minimum der negativen Log-Likelihood-Funktion numerisch bestimmt werden.

## 4 Methode der kleinsten Quadrate

Wir definieren uns für  $N$  Messwerte  $y_i$  mit Varianz  $\sigma_i^2$  eine Funktion  $S$ :

$$S = \sum_{i=1}^N \frac{(y_i - a_i)^2}{\sigma_i^2} = \sum_{i=1}^N \frac{\Delta y_i^2}{\sigma_i^2}.$$

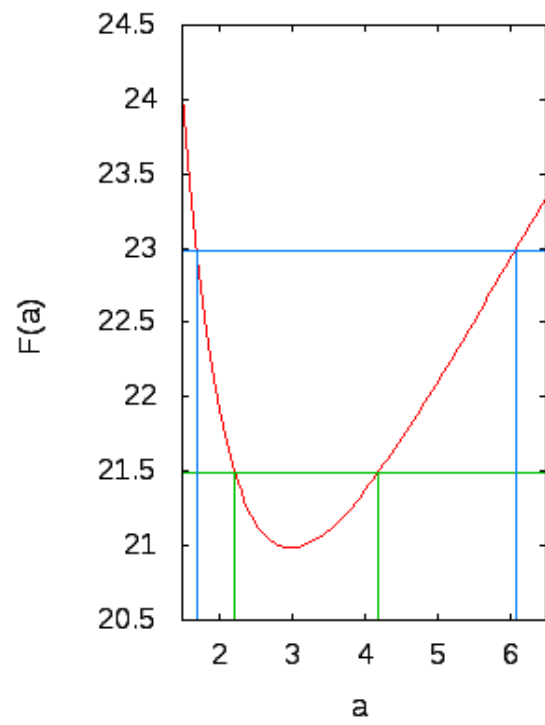
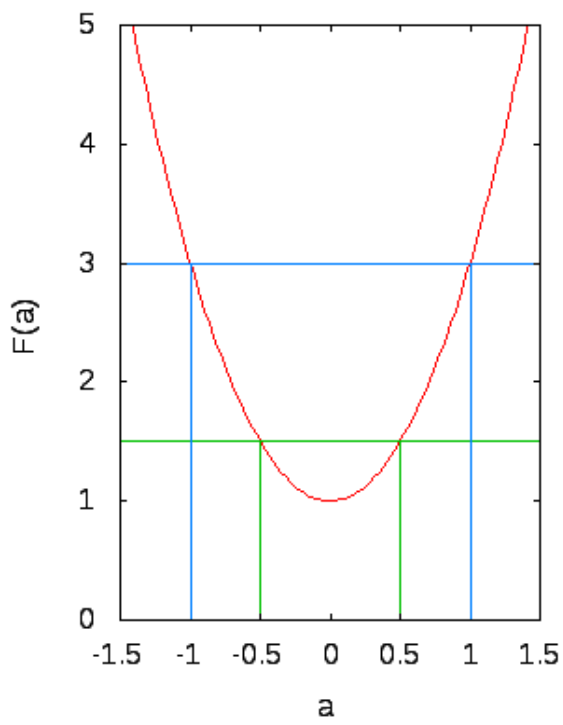


Abbildung 1: Symmetrischer und Asymmetrischer Fehler einer negativen Log-Likelihood-Funktion

Die Parameter  $a_i$  kommt hierbei aus der Theorie. Die Methode der kleinsten Quadrate besagt nun, dass die Funktion  $S$  nach den  $a_i$  minimiert werden muss. Die  $\Delta y_i$  werden Residuen genannt. Bei der Methode der kleinsten Quadrate wird also die Summe der Quadrate der Residuen minimiert.

Im allgemeinen Fall mit Kovarianzmatrix  $\mathbf{V}$  und  $\Delta \mathbf{y} = \begin{pmatrix} \Delta y_1 \\ \Delta y_2 \\ \vdots \\ \Delta y_N \end{pmatrix}$  ergibt sich:

$$S = \Delta \mathbf{y}^T \mathbf{V}^{-1} \Delta \mathbf{y}$$

Diese einfache Behandlung von Kovarianzen mit der Kovarianzmatrix ist ein Vorteil der Methode der kleinsten Quadrate.

Im Fall von gaußschen Fehlern, ist diese Methode identisch mit der Maximum-Likelihood-Methode. Dies ist schnell ersichtlich, wenn man eine gaußsche Wahrscheinlichkeitsdichte in die log-Likelihood-Funktion einsetzt:

$$\begin{aligned} F(a) &= - \sum_i \ln \left( \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \frac{(x_i - a)^2}{\sigma^2}} \right) \\ &= \text{const} + \frac{1}{2} \sum_i \frac{(x_i - a)^2}{\sigma^2} \\ &= \text{const} + \frac{1}{2} S(a) \end{aligned}$$

Bis auf eine unwesentliche Konstante und den Faktor  $\frac{1}{2}$  sind beide Methoden also identisch.

## 4.1 Lineare kleinste Quadrate

Wenn eine lineare Theorie  $t(x|\mathbf{a}) = \sum_i a_i t_i(x)$  vorliegt, lässt sich das Minimum analytisch finden. Der Einfachheit halber gehen wir erst mal davon aus, dass wir identische Varianzen haben.  $S$  ergibt sich nun zu

$$S = \frac{1}{\sigma^2} \sum_i (y_i - t(x_i))^2$$

Ableitung Null setzen ergibt die Normalengleichung:

$$\frac{\partial S}{\partial a_j} = -\frac{2}{\sigma^2} \sum_i t_j(x_i) (y_i - t(x_i)) \stackrel{!}{=} 0$$

$$\text{Normalengleichung: } \sum_i t_j(x_i) \sum_k \hat{a}_k t_k(x_i) = \sum_i y_i t_j(x_i)$$

Wir gehen in Matrixschreibweise über:

$$\mathbf{A} = \begin{pmatrix} t_1(x_1) & t_2(x_1) & \dots & t_p(x_1) \\ t_1(x_2) & t_2(x_2) & \dots & t_p(x_2) \\ \vdots & & & \vdots \\ t_1(x_n) & t_2(x_n) & \dots & t_p(x_n) \end{pmatrix}$$

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$\text{Normalengleichung: } (\mathbf{A}^T \mathbf{A}) \hat{\mathbf{a}} = \mathbf{A}^T \mathbf{y}$$

Diese Gleichung lässt sich lösen:

$$\hat{\mathbf{a}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{y}$$

Hier muss zwar eine meist sehr große Matrix invertiert werden, dies ist jedoch machbar und es müssen keine Näherungsmethoden angewendet werden.

## 4.2 Fehlerfortpflanzung

Im allgemeinen Fall mit Gewichtsmatrix  $\mathbf{W} = \mathbf{V}^{-1}$  lautet obige Lösungsformel:

$$\hat{\mathbf{a}} = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W} \mathbf{y} = \mathbf{B} \mathbf{y}$$

Damit lässt sich das Fehlerfortpflanzungsgesetz herleiten:

$$\mathbf{V}_a = \mathbf{B} \mathbf{y} \mathbf{B}^T$$

$$\mathbf{V}_a = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W} \mathbf{V} ((\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W})^T$$

$$\mathbf{V}_a = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W} \mathbf{A} (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1}$$

$$\mathbf{V}_a = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1}$$

Auch hier erhalten wir also eine einfache analytische Formel.

## 4.3 $\chi^2$ -Test

Bisher haben wir Fehlerbetrachtung nur im Sinne der Fortpflanzung der Fehler der Messwerte betrieben. Eine wichtige Frage ist aber auch oft, ob die gewählte Theorie überhaupt richtig ist. Im Fall von gaußschen Fehlern lässt sich dies bei der Methode der kleinsten Quadrate mithilfe des  $\chi^2$ -Tests machen.

Eine  $\chi_k^2$ -Verteilung gibt die Wahrscheinlichkeitsdichte für die Summe der Quadrate von  $k$  standard-normalverteilten Zufallsvariablen ( $k$  Freiheitsgrade) an. Für gaußsche

Fehler folgt  $S$  gerade solch einer  $\chi^2$  Verteilung mit  $n - p$  Freiheitsgraden. Die Wahrscheinlichkeit gerade den Wert  $S$  oder einen größeren zu bekommen ist dann gerade

$$\int_S^\infty = f(\chi^2)d\chi^2.$$

Beim  $\chi^2$ -Test wählt man nun eine Wahrscheinlichkeit die noch akzeptabel ist. Bei allem was unterhalb dieser Wahrscheinlichkeit ist wird die Theorie verworfen. Dies bedeutet aber auch, dass bei einer gewählten Wahrscheinlichkeit von 1% auch in 1% der Fälle eine wahre Theorie verworfen wird.

## 4.4 Eigenschaften

Auch die Methode der kleinsten Quadrate ist konsistent. Im linearen Fall ist sie zudem Erwartungstreu. Sie ist zudem die effizienteste erwartungstreue Methode. Dies geht auch hier auf Kosten der Robustheit. Vor allem Ausreißer in den Daten führen zu völlig falschen Werten, zu sehen in Abbildung 2. Im gaußschen Fall kann der  $\chi^2$ -Test angewendet werden, der die zugrunde liegende Theorie überprüfen kann.

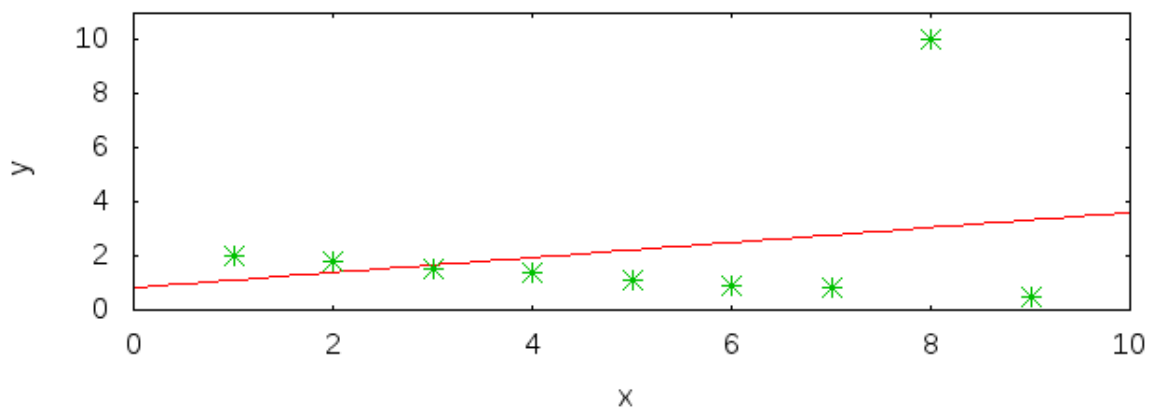


Abbildung 2: Ausreißer in den Daten führen zu falschem Kleinste-Quadrate-Fit

## 5 Zusammenfassung

In Tabelle 1 sind nochmal die Eigenschaften der beiden betrachteten Methoden aufgelistet. Viele Programme, wie zum Beispiel LibreOffice Calc, Microsoft Excel, oder auch Gnuplot, benutzen standardmäßig die Methode der kleinsten Quadrate, da für diese lediglich die Werte mit den jeweiligen Fehlern (bei gleichen Fehlern reichen die Werte) und die zu fittende Funktion benötigt werden und mit wenig Rechenaufwand gelöst werden kann.

	<b>Maximum-Likelihood</b>	<b>Kleinste Quadrate</b>
Voraussetzung	pdf exakt bekannt	Mittelwerte und Varianzen
Konsistent	Ja	Ja
Erwartungstreu	Nur asymptotisch	Im linearen Fall
Effizient	maximal	maximal
Robust	Nein (pdf muss exakt bekannt sein)	Nein (Ausreißer)
Rechenaufwand	kann sehr hoch werden	im linearen Fall gering
Fit-Qualität	nein	bei gaußschen Fehlern

Tabelle 1: Vergleich zwischen Maximum-Likelihood-Methode und Methode der kleinsten Quadrate

## 6 Literatur

- Volker Blobel, Erich Lohrmann: *Statistische und numerische Methoden der Datenanalyse*
- Gerhard Bohm, Günter Zech: *Einführung in Statistik und Messwertanalyse für Physiker* <http://www-library.desy.de/preparch/books/vstatmp.pdf>
- Roger Barlow: *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*