Kapitel 5

Die Maximum-Likelihood-Methode

Mit dem χ^2-Test kann man quantitativ bestimmen, ob die Elemente einer Stichprobe Normalverteilungen mit angenommenen oder geschätzten Parametern $\mu_i,\ \sigma_i$ folgen. Durch Minimierung von χ^2 als Funktion der Parameter kann man eine optimale Schätzung der Parameter erhalten. Das ist die sogenannte 'Methode der kleinsten Quadrate', die im nächsten Kapitel behandelt wird.

Die 'Methode der kleinsten Quadrate' entspricht der 'Maximum-Likelihood-Methode' für den Spezialfall, dass die Stichproben aus Normalverteilungen stammen. Die 'Maximum-Likelihood-Methode' (ML-Methode) ist eine allgemeine Methode zur Bestimmung von Parametern aus Stichproben für beliebige Wahrscheinlichkeitsverteilungen. Deshalb diskutieren wir im folgenden zunächst das ML-Prinzip.

5.1 Das Maximum-Likelihood-Prinzip

Es sei wieder eine Stichprobe $x_1, \ldots, x_n$ vom Umfang n gegeben, wobei jedes x_i im allgemeinen für einen ganzen Satz von Variablen stehen kann.

Wir wollen jetzt die Wahrscheinlichkeit für das Auftreten dieser Stichprobe berechnen unter der Annahme, dass die x_i einer Wahrscheinlichkeitsdichte $f(x|\theta)$ folgen, die durch einen Satz von Parametern $\theta = \theta_1, \ldots, \theta_m$ bestimmt ist. Wenn die Messungen zufällig sind (siehe die Gleichungen (4.4, 4.5) in Abschnitt 4.1), ist diese Wahrscheinlichkeit das Produkt der Wahrscheinlichkeiten für das Auftreten jedes einzelnen Elementes der Stichprobe:

$$L(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{5.1}$$

Die so definierte Stichprobenfunktion heisst **Likelihood-Funktion** und ist als Wahrscheinlichkeitsdichte für Stichproben $x_1, \ldots, x_n$ auf deren Definitionsbereich Ω normiert:

$$\int_\Omega L(x_1, \ldots, x_n|\theta)\, dx_1 \ldots dx_n = 1 \tag{5.2}$$

Das gilt für alle θ, solange $f(x_i|\theta)$ richtig normiert ist. Es ist wichtig zu realisieren, dass L nicht auf den θ-Bereich normiert ist. Andererseits betrachtet man L bei der Suche nach optimalen Parametern als eine Funktion der Parameter, die im Optimierungsprozess variiert werden.

Das ML-Prinzip lässt sich nun wie folgt formulieren:

Wähle aus allen möglichen Parametersätzen θ denjenigen Satz $\hat{\theta}$ als Schätzung, für den gilt:

$$L(x_1, \ldots, x_n|\hat{\theta}) \geq L(x_1, \ldots, x_n|\theta) \qquad \forall\, \theta \tag{5.3}$$

Das Prinzip läuft also auf die Aufgabe hinaus, das Maximum von L in bezug auf die Parameter zu finden. Da L als Produkt von Wahrscheinlichkeiten sehr kleine Zahlenwerte haben kann, benutzt man aus numerischen Gründen meistens den Logarithmus der Likelihood-Funktion, die sogenannte Log-Likelihood-Funktion:

$$\mathcal{L}(x_1, \ldots, x_n|\theta) = \log L(x_1, \ldots, x_n|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{5.4}$$

Die Maximierungsbedingungen lauten dann für die Log-Likelihood-Funktion, zunächst für nur einen Parameter θ:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log f(x_i|\theta) = 0 \qquad \Longrightarrow\ \hat{\theta} \tag{5.5}$$

$$\left.\frac{\partial^2 \mathcal{L}}{\partial \theta^2}\right|_{\theta=\hat{\theta}} < 0 \tag{5.6}$$

Die Verallgemeinerung auf mehrere Parameter $\theta = \theta_1, \ldots, \theta_m$ lautet:

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \sum_{i=1}^{n} \log f(x_i|\theta) = 0 \qquad \Longrightarrow\ \hat{\theta} \tag{5.7}$$

$$\left.\frac{\partial^2 \mathcal{L}}{\partial \theta_i \theta_j}\right|_{\theta=\hat{\theta}} = U_{ij}(\hat{\theta}) \text{ negativ definit} \tag{5.8}$$

Die Matrix U ist negativ definit, wenn alle Eigenwerte kleiner 0 sind. Falls Gleichung (5.7) auf ein lineares Gleichungssystem führt, kann man die Lösung durch Matrixinversion erhalten. Im allgemeinen sind die Gleichungen nicht-linear und man muss eine numerische, meistens iterative Methode zur Lösung finden. Wir werden Lösungsverfahren im Zusammenhang mit der 'Methode der kleinsten Quadrate' im nächsten Kapitel besprechen.

Beispiele:

1. Schätzung der mittleren Lebensdauer: Die Abfolge der Zerfälle eines radioaktiven Präparates habe die Wahrscheinlichkeitsdichte

$$f(t|\tau) = \frac{1}{\tau}\, e^{-t/\tau}, \tag{5.9}$$

die als einzigen Parameter die mittlere Lebensdauer τ enthält. In einer Messung werden n Zerfälle mit den Zeiten $t_i,\ i = 1, \ldots, n$ gemessen. Die Likelihood-Funktion dieser Stichprobe ist:

$$L(t_1, \ldots, t_n|\tau) = \prod_{i=1}^{n} \frac{1}{\tau}\, e^{-t_i/\tau} \qquad \Longrightarrow \qquad \mathcal{L}(t_1, \ldots, t_n|\tau) = \sum_{i=1}^{n} \left(-ln\,\tau - \frac{t_i}{\tau}\right) \tag{5.10}$$

Die Maximierung von $\mathcal{L}$ ergibt den ML-Schätzwert für τ:

$$\frac{\partial \mathcal{L}}{\partial \tau} = \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0 \qquad \Longrightarrow \qquad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i = \bar{t} \tag{5.11}$$

mit

$$\left. \frac{\partial^2 \mathcal{L}}{\partial \tau^2} \right|_{\tau=\hat{\tau}} = -\frac{n}{\hat{\tau}^2} < o \tag{5.12}$$

Die ML-Schätzung der mittleren Lebensdauer ist also das arithmetische Mittel der gemessenen Zeiten.

2. Schätzung der Parameter einer Gauss-Verteilung: Eine Stichprobe $x_i,\ i = 1, \ldots, n$ aus einer Normalverteilung $N(\mu,\, \sigma)$ hat die Likelihood-Funktion:

$$L(x_1, \ldots, x_n | \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x_i-\mu)^2}{2\,\sigma^2}} \tag{5.13}$$

Die Maximierung in Bezug auf beide Parameter fordert:

$$\frac{\partial \mathcal{L}}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \tag{5.14}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = \sum_{i=1}^{n} \left(-\frac{1}{2\,\sigma^2} + \frac{1}{2\,\sigma^4} \left(x_i - \mu\right)^2 \right) = 0 \tag{5.15}$$

Die Lösung des Gleichungssystems ergibt:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \tag{5.16}$$

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{5.17}$$

Die ML-Schätzung des Mittelwertes ist also wieder das arithmetische Mittel. Die Schätzung der Varianz ist allerdings verzerrt, denn der Erwartungswert ist nicht unabhängig von n (siehe dazu Abschnitt 4.2):

$$E(\hat{\sigma^2}) = \left(1 - \frac{1}{n} \right) \sigma^2 \tag{5.18}$$

Die Schätzung ist aber 'konsistent', weil der Erwartungswert der Schätzung für große n gegen den zu schätzenden Parameter konvergiert.

5.2 Fehlerbestimmung für ML-Schätzungen

Die Fehler oder Unsicherheiten in der Parameterbestimmung mit der ML-Methode lassen sich nur in speziellen Fällen explizit angeben, zum Beispiel wenn die Likelihood-Funktion normalverteilt in den Parametern ist (siehe unten). Andererseits ist eine Parameterbestimmung ohne Aussagekraft, wenn man nicht einen Fehler oder ein Vertrauensniveau angeben kann. Im allgemeinen wird die vollständige Kovarianzmatrix benötigt, wenn man ML-Ergebnisse für die weitere Auswertung braucht.

5.2.1 Allgemeine Methoden der Varianzabschätzung

Direkte Methode: Die direkte Methode gibt die Streuung der Schätzwerte $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ an, wenn man viele Messungen mit Stichproben $(x_1, \ldots, x_n)$ macht:

$$V_{ij}(\theta) = \int (\hat{\theta}_i - \theta_i)\,(\hat{\theta}_j - \theta_j)\, L(x_1, \ldots, x_n|\theta)\, dx_1 \ldots dx_n \tag{5.19}$$

Hier ist also θ der 'wahre' Parametersatz und $\hat{\theta}(x_1, \ldots, x_n)$ sind die Schätzungen, die man jeweils für eine Stichprobe erhält. Die Stichprobe, über die integriert wird, folgen der Wahrscheinlichkeitsdichte $L(x_1, \ldots, x_n)$.

Bei dieser Varianzbestimmung wird die Kenntnis des wahren Parametersatzes θ und der Verlauf von L als Funktion der x_i vorausgesetzt. Bei einer Messung weiss man in der Regel weder das eine noch das andere. Man kann diese Methode aber zum Beispiel zur Planung von Experimenten benutzen, um die zu erwartenden Fehler beim Testen eines Modells mit bestimmten Parametern auszuloten. Die Auswertung wird dann in der Regel mit Simulationen der Stichproben gemacht. Auch für experimentelle Messungen kann man diese Bestimmung der Varianzen benutzen. Für den geschätzten Parametersatz $\hat{\theta}$ simuliert man den Verlauf der Likelihood-Funktion durch die Simulation vieler Messungen, die man in der Praxis nicht durchführen könnte.

Praktische Methode: In der Praxis wird meistens $L(x_1, \ldots, x_n|\theta)$ bei fester Stichprobe $(x_1, \ldots, x_n)$ als Wahrscheinlichkeitsdichte für θ angenommen. Dann erhält man für die Varianzmatrix:

$$V_{ij}(\theta) = \frac{\int (\theta_i - \hat{\theta}_i)\,(\theta_j - \hat{\theta}_j)\, L(x_1, \ldots, x_n|\theta)\, d\theta_1 \ldots d\theta_m}{\int L(x_1, \ldots, x_n|\theta)\, d\theta_1 \ldots d\theta_m} \tag{5.20}$$

Hier ist $\hat{\theta}$ die ML-Schätzung, die aus der einen gemessenen Stichprobe $(x_1, \ldots, x_n)$ bestimmt wurde. In der Formel (5.20) ist berücksichtigt, dass L nicht auf den θ-Bereich normiert ist, wie bereits oben erwähnt wurde.

In der Regel werden die Integrationen numerisch durch Abtasten der Likelihood-Funktion für verschiedene Parameter θ durchgeführt.

5.2.2 Varianzabschätzung durch Entwicklung um das Maximum

Wenn die Likelihood-Funktion gewisse günstige Eigenschaften hat, insbesondere wenn der Verlauf um den optimalen Parametersatz als Funktion der Parameter ein ausgeprägtes Maximum hat und nach beiden Seiten monoton abfällt, kann man eine Entwicklung um das Maximum versuchen. Aus Gründen, die wir gleich verstehen werden, entwickeln wir die Log-Likelihood-Funktion:

$$\mathcal{L}(x_1, \ldots, x_n|\theta) = \mathcal{L}(x_1, \ldots, x_n|\hat{\theta}) + (\theta - \hat{\theta})\, \frac{\partial \mathcal{L}}{\partial \theta}\Big|_{\theta=\hat{\theta}} + \frac{1}{2}\,(\theta_i - \hat{\theta}_i)\,(\theta_j - \hat{\theta}_j)\, \frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j}\Big|_{\theta=\hat{\theta}} + \ldots \tag{5.21}$$

Wegen der Maximumbedingung verschwindet die erste Ableitung. Die zweiten Ableitungen werden zusammengefasst:

$$V_{ij}^{-1} \;=\; -\frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j}\Big|_{\theta=\hat{\theta}} \tag{5.22}$$

Damit ergibt sich in der Umgebung des Maximums:

$$\mathcal{L}((x_1, \ldots, x_n|\theta) \approx \mathcal{L}_{max} - \frac{1}{2}(\theta - \hat{\theta})^T V^{-1} (\theta - \hat{\theta}) \tag{5.23}$$

und für die Likelihood-Funktion L folgt:

$$L((x_1, \ldots, x_n|\theta) \approx L_{max}\, e^{-\frac{1}{2}(\theta-\hat{\theta})^T V^{-1} (\theta-\hat{\theta})} \tag{5.24}$$

Das heisst, wenn die Likelihood-Funktion als Funktion der Parameter ein annähernd gaussisches Verhalten zeigt, kann die Varianz durch die zweiten Ableitungen entsprechend (5.22) abgeschätzt werden. In der Praxis wird häufig angenommen, dass die Likelihood-Funktion einer (Multi)-Normalverteilung folgt.

Wenn die Parameter unkorreliert sind, ist V^{-1} diagonal und die Varianz der Parameter ist:

$$\sigma_i^2 = \frac{1}{V_{ii}^{-1}} = \left(-\frac{\partial^2 \mathcal{L}}{\partial {\theta_i}^2}\Big|_{\theta=\hat{\theta}} \right)^{-1} \tag{5.25}$$

5.2.3 Vertrauensintervalle und Likelihood-Kontouren

Die Fehler der Parameter werden häufig als die Wurzeln aus den Varianzen, wie sie im vorigen Abschnitt bestimmt wurden, angegeben. Wenn man genauer sein will, kann man Likelihood-Kontouren angeben. Das sind im allgemeinen Fall Hyperflächen im Parameterraum, die durch

$$L((x_1, \ldots, x_n|\theta) = const \tag{5.26}$$

festgelegt sind und einen bestimmten Wahrscheinlichkeitsinhalt η, entprechend einem Vertrauensniveau, haben. Bei zwei Parametern (θ_i, θ_j) ergibt sich zum Beispiel in der Regel eine geschlossene, zwei-dimensionale Raumkurve um die Schätzwerte $(\hat{\theta}_i, \hat{\theta}_j)$ der Parameter (Abb. 5.1). Im allgemeinen können die Hyperflächen beliebige Volumina im Parameterraum einschliessen, zum Beispiel brauchen diese Volumina auch nicht zusammenzuhängen (ein Beispiel ist in Abb. 5.2 gezeigt).

Als Vertrauensniveau können Werte wie 68%, 90%, 95% usw. angegeben werden. Im allgemeinen müssen die Likelihood-Kontouren dafür numerisch integriert werden. In dem speziellen Fall, dass die Likelihood-Funktion durch eine Normalverteilung entsprechend (5.24) beschrieben werden kann, folgt

$$2\,\Delta\mathcal{L} = 2\,[\mathcal{L}_{max} - \mathcal{L}((x_1, \ldots, x_n|\theta)] = (\theta - \hat{\theta})^T V^{-1} (\theta - \hat{\theta}) \tag{5.27}$$

einer χ^2-Verteilung mit m Freiheitsgraden (m = Anzahl der Parameter). In diesem Fall ergibt $2\,\Delta\mathcal{L} = 1$ die Kovarianzen der Parameter. Die Kontouren zu einem Vertrauensniveau η ergeben sich aus den Kurven in Abb. 4.3 durch $2\,\Delta\mathcal{L} = \chi^2 = const$ für $n_F = m$ und mit $\eta = 1 - \alpha$. Die Kontouren sind im Zweidimensionalen Ellipsen und im allgemeinen m-dimensionale Ellipsoide.

Zum Beispiel enthält die Kontour mit $m = 2$, $2\,\Delta\mathcal{L} = 1$ (das ist die Ellipse, die die $\pm 1\sigma$-Linien schneidet, siehe Abb. 5.1) nur 39.4% Wahrscheinlichkeit, während das für $m = 1$ bekanntlich 68.3% sind.

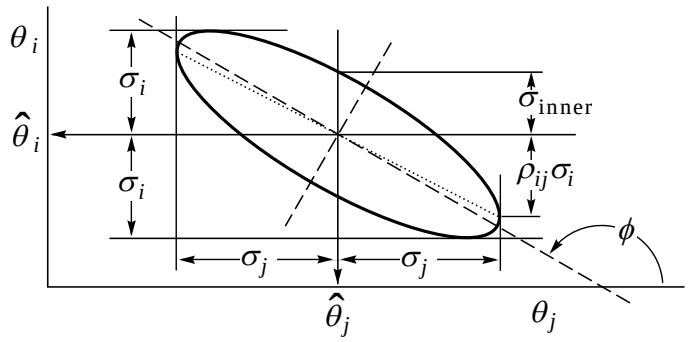

Abbildung 5.1: Standard-Fehlerellipse für die Schätzwerte $\hat{\theta}_i$ und $\hat{\theta}_j$.

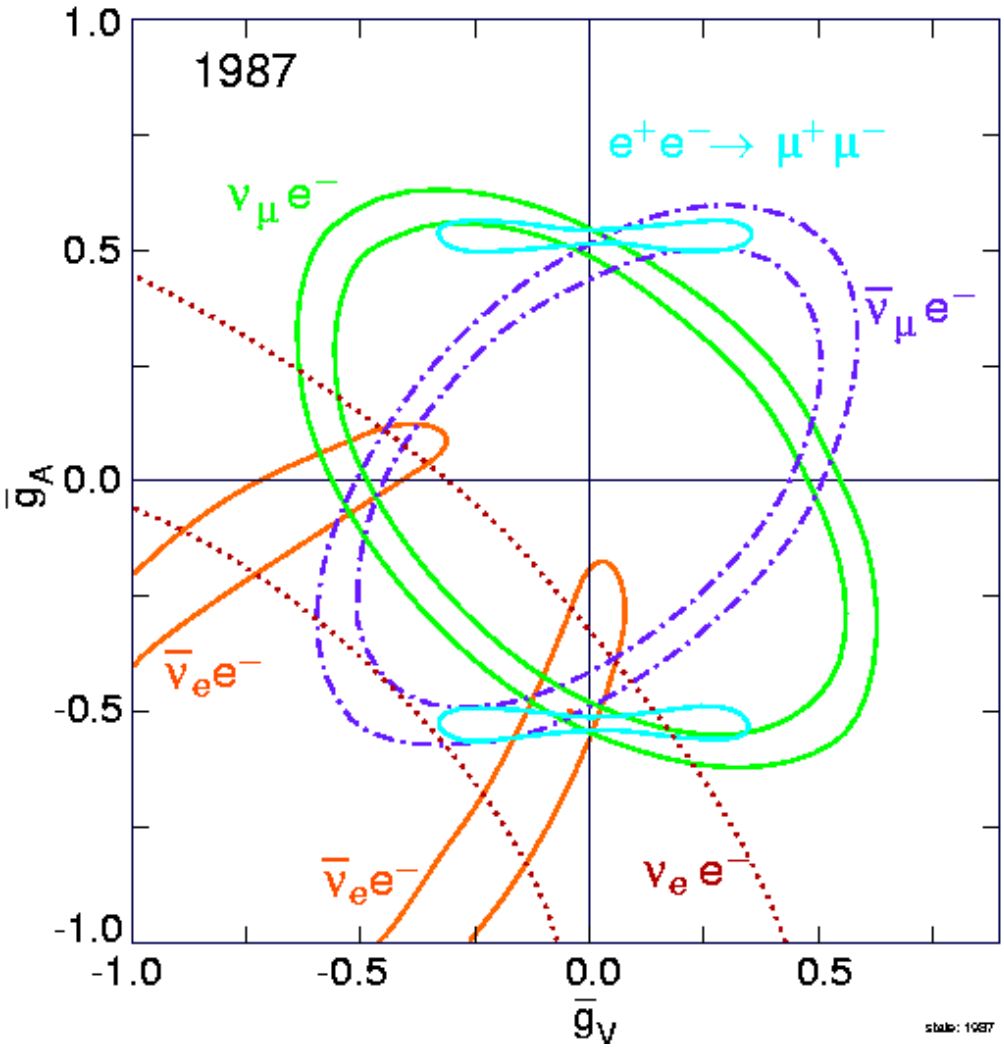

Abbildung 5.2: Beispiel für Likelihood-Kontouren: Die zwei Konstanten g_V und g_A (Kopplung von Leptonen an das Z^0-Boson) werden in verschiedenen Teilchenreaktionen gemessen, die sehr unterschiedliche Likelihood-Kontouren liefern. Die besten Schätzwerte liegen innerhalb der Kontouren, die ringförmige und zum Teil auch nicht zusammenhängende Gebiete beschreiben. Der einzige Bereich, den alle Likelihood-Kontouren umschreiben, ist nahe $g_V = 0,\ g_A = -0.5$. Für die genaue Analyse müssen alle Likelihood-Funktionen kombiniert werden.

5.3 Eigenschaften von ML-Schätzungen

Die Likelihood-Schätzung der Parameter hat in vieler Hinsicht optimale Eigenschaften. Im Rahmen dieser Vorlesung ist allerdings nicht ausreichend Zeit, in die Details und die mathematischen Beweise zu schauen. Einige dieser Eigenschaften sollen hier nur kurz erwähnt werden:

1. Invarianz gegenüber Parametertransformationen: Im allgemeinen ist die Schätzung unabhängig davon, wie die Parameter dargestellt werden. Für eine Transformation

$$\theta \to \phi \tag{5.28}$$

ergibt sich:

$$\hat{\phi} = \phi(\hat{\theta}) \tag{5.29}$$

Zum Beispiel kann man für die Schätzung einer mittleren Lebensdauer τ auch die Zerfallswahrscheinlichkeit $\lambda = 1/\tau$ benutzen, denn aus

$$\frac{\partial L}{\partial \lambda}(\hat{\lambda}) = 0 \tag{5.30}$$

folgt

$$\frac{\partial L}{\partial \tau}\frac{\partial \tau}{\partial \lambda}(\hat{\lambda}) = 0 \;\Rightarrow\; \frac{\partial L}{\partial \tau}(\tau(\hat{\lambda})) = 0 \qquad (\frac{\partial \tau}{\partial \lambda} \neq 0) \tag{5.31}$$

2. Konsistenz: Für große Stichproben geht der Schätzwert in den tatsächlichen Wert über:

$$\lim_{n\to\infty} \hat{\theta} = \theta \tag{5.32}$$

3. Verzerrung: Wir hatten am Beispiel der Schätzung der Varianz einer Gauss-Verteilung gesehen (siehe (5.18)), dass die ML-Schätzung nicht unbedingt verzerrungsfrei ist, d. h. es gilt nicht $E(\hat{\theta}) = \theta$ für alle n. Allgemein gilt allerdings, dass die ML-Schätzung asymptotisch verzerrungsfrei ist:

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta \tag{5.33}$$

4. Effizienz: In den meisten Fällen ist eine ML-Schätzung effizient, das heisst, die geschätzten Parameter haben minimale Varianz. Jedenfalls gilt das im Fall großer Stichproben: die ML-Schätzung ist asymptotisch effizient.

Schwieriger ist die Beurteilung der Fehler und Vertrauensintervalle einer Schätzung. Das Problem tritt dann auf, wenn man die Likelihood-Funktion als Wahrscheinlichkeitsdichte der Parameter interpretiert und benutzt. Zur Fehlerabschätzung braucht man eigentlich den Verlauf der gesamten Likelihood-Funktion. Wir hatten bereits darauf hingewiesen, dass die Likelihood-Funktion in Abhängigkeit von den Parametern nicht normiert ist. Um richtig normieren zu können, müsste man eigentlich den möglichen Bereich der Parameter genau kennen und auch, ob alle Parameter gleich wahrscheinlich sind, das heisst, was die 'a priori' Wahrscheinlichkeiten der Parameter sind.

Nach dem Bayes-Theorem (1.13) würde man bei einer gegebenen Stichprobe $\vec{x}$ und für diskrete Hypothesen θ_i folgende 'a posteriori' Wahrscheinlichkeit, dass die Hypothese θ_i wahr ist, erhalten:

$$P(\theta_i|\vec{x}) = \frac{P(\vec{x}|\theta_i) \cdot P(\theta_i)}{\sum_j P(\vec{x}|\theta_j) \cdot P(\theta_j)} \tag{5.34}$$

Hier entspricht $P(\vec{x}|\theta_i)$ der Likelihood-Funktion $L(\vec{x}|\theta_i)$ und $P(\theta_i)$ ist die 'a priori' Wahrscheinlichkeit der Hypothese θ_i. Der Nenner normiert auf alle möglichen Hypothesen (für kontinuierliche Hypothesen-Parameter ergibt sich ein Normierungsintegral).

Beispiel: In Teilchenexperimenten möchte man häufig die gemessenen langlebigen Teilchen identifizieren, typischerweise die 5 Teilchensorten $i,\ i = p,\ K,\ \pi,\ e,\ \mu$. Aus den Informationen verscheidener Detektoren, die uns hier nicht im Detail interessieren, kann man eine Masse m des Teilchens bestimmen (zum Beispiel aus der Messung von Impuls und Geschwindigkeit) und damit eine Wahrscheinlichkeit für eine Teilchenhypothese i:

$$P(i|m) = \frac{P(m|i) \cdot P(i)}{\sum_j P(m|j) \cdot P(j)} \tag{5.35}$$

Die Wahrscheinlichkeit $P(m|i)$, bei Vorliegen des Teilchens i eine Masse m zu messen, bestimmt man in der Regel experimentell mit bekannten Teilchenstrahlen. Die 'a priori' Wahrscheinlichkeit $P(i)$ für das Auftreten der Teilchensorte i entnimmt man dem gleichen Experiment, weil die Teilchenhäufigkeiten abhängig von der Energie der Reaktion (und eventuell noch anderen Parametern) sind. Die Teilchenhäufigkeiten sind im allgemeinen sehr unterschiedlich, mit starker Dominanz der Pionen. Wenn es zum Beispiel einen Faktor 10 mehr Pionen als Kaonen gibt, muss $P(m|K) > 10 \cdot P(m|\pi)$ sein, damit es als Kaon identifiziert wird. Die Kenntnis der 'a priori' Wahrscheinlichkeit ist also in diesem Fall besonders wichtig.

In vielen Fällen kennt man die 'a priori' Wahrscheinlichkeiten für die Hypothesen nicht und nimmt dann an, dass sie konstant sind. Dass das problematisch ist, sieht man auch daran, dass die Vertrauensintervalle nicht invariant gegen Transformationen der Parameter sind. Für die Transformation

$$\theta \to \phi(\theta) \tag{5.36}$$

ergibt sich für die Berechnung eines Vertrauensintervalls:

$$\int_{\theta_1}^{\theta_2} L(\vec{x}|\theta)\, d\theta = \int_{\phi(\theta_1)}^{\phi(\theta_2)} L(\vec{x}|\phi(\theta)) \left|\frac{\partial\theta}{\partial\phi}\right| d\phi \neq \int_{\phi_1}^{\phi_2} L(\vec{x}|\phi)\, d\phi \tag{5.37}$$

Das rechte Integral hätte man ja erhalten, wenn man von vornherein ϕ als Parameter gewählt hätte.