Kapitel 6

Die Maximum-Likelihood-Methode

In diesem und dem nächsten Kapitel werden wir Methoden untersuchen, mit denen für Daten von Stichproben eine möglichst optimale theoretische Beschreibung beziehungsweise ein passendes Modell gefunden werden kann. Es kann sich dabei um diskrete Modell-Hypothesen oder um Funktionen der Messwerte handeln. Funktionen werden im allgemeinen durch geeignete Wahl von Parametern an die Messungen angepasst. Die Prozedur der Anpassung optimaler Parameter oder der Wahl einer Hypothese sollte gleichzeitig ein quantitatives Kriterium für die Güte der Beschreibung der Daten im Vergleich zu anderen möglichen Hypothesen bieten.

Die 'Maximum-Likelihood-Methode' (ML-Methode) ist in verschiedener Hinsicht die allgemeinste Methode zur Parameterschätzung mit vielen optimalen Eigenschaften. Eine speziellere Methode ist die sogenannte 'Methode der kleinsten Quadrate', die auf dem χ^2-Test für normal-verteilte Messwerte beruht (siehe nächstes Kapitel). Die 'Methode der kleinsten Quadrate' entspricht der 'Maximum-Likelihood-Methode' für den Spezialfall, dass die Stichproben aus Normalverteilungen stammen. Deshalb diskutieren wir im folgenden zunächst das ML-Prinzip.

6.1 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:

$$L(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{6.1}$$

Die so definierte Stichprobenfunktion heisst **Likelihood-Funktion** und ist als Wahrscheinlichkeitsdichte für Stichproben $x_1, \ldots, x_n$ auf deren Definitionsbereich Ω nor-

miert:

$$\int_\Omega L(x_1, \ldots, x_n|\theta)\, dx_1 \ldots dx_n = 1 \tag{6.2}$$

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:
>
> $$L(x_1, \ldots, x_n|\hat{\theta}) \geq L(x_1, \ldots, x_n|\theta) \qquad \forall\, \theta \tag{6.3}$$

Das Prinzip läuft also auf die Aufgabe hinaus, das Maximum von L in bezug auf die Parameter zu finden. Die Parameter können diskret oder kontinuierlich sein. Im diskreten Fall muss das die maximale Likelihood-Funktion bezüglich diskreter Hypothesen gefunden werden. Wenn die Parameter kontinuierlich sind kann man gängige numerische Methoden zum Auffinden des Maximums als Funktion der Parameter benutzten. 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:

$$\mathcal{L}(x_1, \ldots, x_n|\theta) = \ln L(x_1, \ldots, x_n|\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta) \tag{6.4}$$

Die Maximierungsbedingungen (bei kontinuierlichen Parametern) lauten dann für die Log-Likelihood-Funktion, zunächst für nur einen Parameter θ:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln f(x_i|\theta) = 0 \qquad \Longrightarrow\ \hat{\theta} \tag{6.5}$$

$$\left.\frac{\partial^2 \mathcal{L}}{\partial \theta^2}\right|_{\theta=\hat{\theta}} < 0 \tag{6.6}$$

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

$$\frac{\partial \mathcal{L}}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \ln f(x_i|\theta) = 0 \qquad \Longrightarrow\ \hat{\theta} \tag{6.7}$$

$$\left.\frac{\partial^2 \mathcal{L}}{\partial \theta_i \theta_j}\right|_{\theta=\hat{\theta}} = U_{ij}(\hat{\theta}) \text{ negativ definit} \tag{6.8}$$

Die Matrix U ist negativ definit, wenn alle Eigenwerte kleiner 0 sind. Falls Gleichung (6.7) 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

$$f(t|\tau) = \frac{1}{\tau}\, e^{-t/\tau}, \tag{6.9}$$

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:

$$L(t_1, \ldots, t_n|\tau) = \prod_{i=1}^{n} \frac{1}{\tau}\, e^{-t_i/\tau} \qquad \Longrightarrow \qquad \mathcal{L}(t_1, \ldots, t_n|\tau) = \sum_{i=1}^{n} \left(-\ln \tau - \frac{t_i}{\tau} \right) \tag{6.10}$$

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

$$\frac{\partial \mathcal{L}}{\partial \tau} = \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0 \qquad \Longrightarrow \qquad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i = \bar{t} \tag{6.11}$$

mit

$$\left. \frac{\partial^2 \mathcal{L}}{\partial \tau^2} \right|_{\tau = \hat{\tau}} = -\frac{n}{\hat{\tau}^2} < o \tag{6.12}$$

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:

$$L(x_1, \ldots, x_n|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x_i-\mu)^2}{2\,\sigma^2}} \tag{6.13}$$

$$\Longrightarrow \mathcal{L}(x_1, \ldots, x_n|\mu, \sigma^2) = \frac{1}{2} \sum_{i=1}^{n} \left(-\ln \sigma^2 - \ln 2\pi - \frac{(x_i - \mu)^2}{2\,\sigma^2} \right) \tag{6.14}$$

Die Maximierung in Bezug auf beide Parameter fordert:

$$\frac{\partial \mathcal{L}}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0 \tag{6.15}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = \sum_{i=1}^{n} \left(-\frac{1}{2\,\sigma^2} + \frac{1}{2\,\sigma^4} (x_i - \mu)^2 \right) = 0 \tag{6.16}$$

Die Lösung des Gleichungssystems ergibt:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \tag{6.17}$$

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{6.18}$$

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):

$$E(\hat{\sigma^2}) = \left(1 - \frac{1}{n}\right) \sigma^2 \tag{6.19}$$

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

6.2 ML-Methode für Histogramme

In den Beispielen im vorigen Abschnitt wurde die Likelihood-Funktion als Produkt der Wahrscheinlichkeiten der einzelnen Ereignisse konstruiert ('unbinned likelihood'). Häufig werden Messdaten auch als Histogramme dargestellt, das heißt, die Häufigkeit von Ereignissen als Funktion einer Variablen wird für endliche Intervalle ('bins') dieser Variablen aufgetragen.

Beispiel: In Abb. 6.1 sind die Raten von beobachteten Myonpaaren, die in Proton-Kern-Reaktionen von einem separierten Vertex kommen, gegen deren invariante Masse pro Masseninterval aufgetragen. Die einzelnen Zählraten sind hier als Punkte mit Fehlerbalken eingezeichnet (könnten aber auch als Histogrammbalken dargestellt werden), ein getrennt gemessener Untergrund wird zusätzlich als Histogramm eingezeichnet. Man beobachtet bei etwa 3.1 GeV das Signal für den Zerfall $J/\psi \to \mu^+\mu^-$ mit einer etwa gauss-förmigen Massenverteilung auf einem näherungsweise konstanten Untergrund. Eine Funktion bestehend aus der Summe einer Normalverteilung und einem konstanten Untergrund wurde mit der ML-Methode an die Verteilung angepasst. Die Funktion hat bis zu 4 Parameter: Höhe, Breite und Lage der Normalverteilung und eine Konstante für den Untergrund. Statt der Höhe der Normalverteilung definiert mal vorteilhafter das Integral unter der Signalkurve, weil das direkt die gesuchte Anzahl der J/ψ-Mesonen ergibt und sich damit eine Umrechnung mit eventuell korrelierten Parameterfehlern vermeiden läßt.

Für die Bestimmung der Likelihood-Funktion, die wir für die Anpassung brauchen, nehmen wir an, dass die Raten N_i in jedem Interval i poisson-verteilt sind. Wir vergleichen diese Raten mit der Hypothese $\lambda_i(\theta)$, die wir als Mittelwert der Anpassungsfunktion $f(x|\theta)$ um die Intervalmitte x_i bestimmen:

$$\lambda_i(\theta) = \langle f(x|\theta) \rangle_{[x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2}]} \tag{6.20}$$

Die Likelihood-Funktion wird dann aus den Poisson-Wahrscheinlichkeiten für die Beobachtung von N_i Ereignissen bei gegebenem Erwartungswert λ_i in jedem Interval i konstruiert:

$$L(\theta) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-\lambda_i} \lambda_i^{N_i}}{N_i!} \quad \Rightarrow \ln L(\theta) = \sum_{i=1}^{n} \left(-\lambda_i + N_i \ln \lambda_i - \ln(N_i!)\right) \tag{6.21}$$

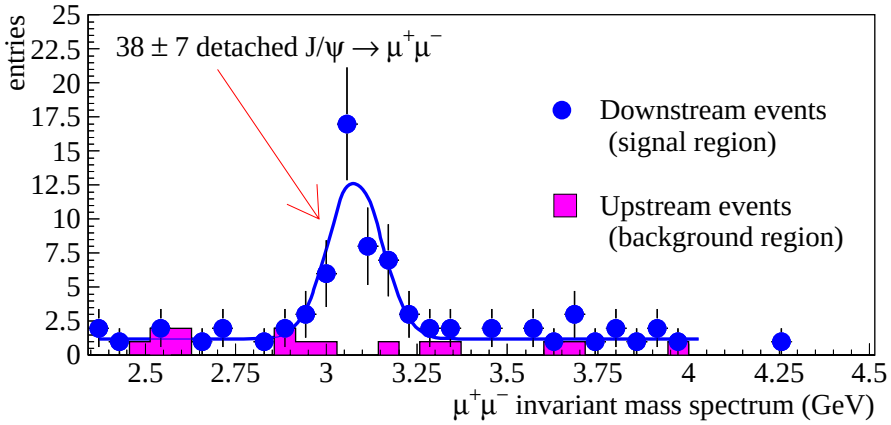

Abbildung 6.1: Massenverteilung von Myonpaaren in Proton-Kern-Reaktionen (HERA-B-Experiment), die einen gemeinsamen Vertex mit Abstand ('detached') zum Primärvertex haben. Die Myonpaare wurden als Kandidaten für Zerfälle von J/ψ-Mesonen, die wiederum aus Zerfällen von langlebigen B-Mesonen stammen, selektiert. Die Verteilung wird durch eine Normalverteilung für das J/ψ-Signal über einem konstanten Untergrund beschrieben.

Der letzte Term ist durch die Messung gegeben und hängt nicht von den zu optimierenden Parametern θ ab. Die zu maximierende Log-Likelihood-Funktion reduziert sich deshalb auf:

$$\ln L(\theta) = \sum_{i=1}^{n} (-\lambda_i + N_i \ln \lambda_i) \tag{6.22}$$

Wenn jedes einzelne Ereignis tatsächlich gemessen wurde und nicht durch den Messprozess bereits der Eintrag in Histogramme erfolgt, kann man alternativ zu dieser 'binned likelihood' Methode natürlich auch die Likelihood-Funktion mit den Wahrscheinlichkeiten der einzelnen Ereignisse konstruieren ('unbinned likelihood'). Die 'unbinned likelihood' kann im Allgemeinen mehr Information ausnutzen.

Bemerkung: Häufig wird die Poisson-Verteilung für die Raten durch eine Normalverteilung approximiert, um dann als Log-Likelihood-Funktion die χ^2-Funktion anpassen zu können (siehe nächstes Kapitel). Bei kleinen Zählraten, insbesondere mit Null-Einträgen in Intervallen, führt das in der Regel zu verfälschten Ergebnissen. Aber auch bei Zählraten, für die die Gauss-Approximation gut ist, gibt es ein Problem: Das Integral unter der Anpassungskurve wird regelmässig unterschätzt, wenn die Fehler durch $1/\sqrt{N_i}$ abgeschätzt werden. Damit werden Fluktuationen nach unten durch einen kleineren Fehler stärker bewichtet als Fluktuationen nach oben. Im Mittel zieht das dann die Anpassungskurve nach unten. Wenn man unbedingt eine χ^2-Anpassung machen will, kann man als Abhilfe den Fehler iterativ mit dem aktuellen Anpassungswert λ_i als $1/\sqrt{\lambda_i}$ festlegen.

6.3 Berücksichtigung von Zwangsbedingungen

Oft sind bei einer Anpassung einer Funktion an Messdaten Zwangsbedingungen zu berücksichtigen. Zwangsbedingungen kommen häufig bei kinematischen Anpassun-

gen vor: zum Beispiel ist in einer e^+e^--Annihilation im Schwerpunktsystem die Summe der Impulse gleich null und die Summe der Energien gleich zweimal die Strahlenergie. Daraus resultieren 4 Zwangsbedingungen, die durch weitere Bedingungen, wie Massen- oder Vertexbedingungen an Untersysteme von Teilchen, ergänzt werden können. Jede Zwangsbedingung kann zur Eliminierung eines Parameters benutzt werden, zum Beispiel kann man mit der gerade erwähnten Impulserhaltung 3 Impulskomponenten eliminieren. Häufig ist das aber nicht erwünscht, zum Beispiel um bei der Anpassung die äquvalente Behandlung der Parameter zu gewährleisten oder um schwierigen Eliminierungs-Algorithmen aus dem Weg zu gehen.

6.3.1 Methode der Lagrange-Multiplikatoren

Die k_c Zwangbedingungen ('constraints') eines Anpassungsproblems werden als Funktionen $c_j(\theta)$ $(j = 1, \ldots, k_c)$ definiert, die verschwinden, wenn die jeweilige Bedingung erfüllt ist. Wie in der klassischen Mechanik lassen sich die Bedingungen mit der Methode der Lagrange-Multiplikatoren in die Likelihood-Funktion einbeziehen:

$$\mathcal{L} = \ln L = \sum_{i=1}^{m} \ln f(x_i|\theta) \; - \; \sum_{j=1}^{k_c} \lambda_j \, c_j(\theta). \tag{6.23}$$

Die k_c Lagrange-Multiplikatoren λ_j werden wie zusätzliche Parameter behandelt, bezüglich der die Likelihood-Funktion ebenfalls zu minimieren ist. Zu den m Maximierungsbedingungen in (6.7)

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = 0 \tag{6.24}$$

kommen noch die k_c Bedingungen

$$\frac{\partial \mathcal{L}}{\partial \lambda_j} = c_j(\theta) = 0. \tag{6.25}$$

Das Verschwinden der Funktionen $c_j(\theta)$ ergibt sich also aus der Maximierungsbedingung bezüglich der Lagrange-Multiplikatoren.

6.3.2 Zwangsbedingungen als Zufallsverteilungen

Insbesondere wenn Zwangsbedingungen nicht scharf definiert sind oder nur mit begrenzter Genauigkeit bekannt sind, kann man die Abweichungen als Zufallsverteilung behandeln. Mit einer angenommenen Normalverteilung mit der Breite δ_j für die Verteilung von c_j um Null ergibt sich in der Log-Likelihood-Funktion ein χ^2-artiger Zusatz:

$$\mathcal{L} = \ln L = \sum_{i=1}^{m} \ln f(x_i|\theta) \; - \; \frac{1}{2} \sum_{j=1}^{k_c} \frac{c_j^2(\theta)}{\delta_j^2}. \tag{6.26}$$

Diese Art der Implementierung der Zwangbedingungen kann auch im Falle scharf definierter Zwangsbedingungen vorteilhaft sein, weil die Anzahl der Parameter kleiner wird. In diesem Fall würde man die δ_j genügend klein machen (eventuell auch adaptiv während des Maximierungsprozesses).

6.3.3 Erweiterte ML-Methode

Es gibt Probleme, bei denen sich aus einer ML-Anpassung gleichzeitig die Anzahl der zu erwartenden Ereignisse ergibt und diese Anzahl mit der Anzahl der tatsächlich beobachteten Ereignisse in Übereinstimmung gebracht werden soll. Will man zum Beispiel von n Ereignissen bestimmen, welcher Bruchteil jeweils aus einer von drei angenommenen Reaktionen stammt, sollte gleichzeitig die Summe der jeweiligen Anzahlen gleich n sein: $n = n_1 + n_2 + n_3$. Man kann nun diese Bedingung als einen zusätzlichen Faktor in die Likelihood-Funktion einsetzen, und zwar entsprechend der Poisson-Verteilung als Wahrscheinlichkeit, dass bei einem Erwartungswert λ tatsächlich n Ereignisse beobachtet werden. Die Likelihood-Funktion (6.1) mit normierten Wahrscheinlichkeiten $f(x|\theta)$ wird dann erweitert zu:

$$L(x_1, \ldots, x_n|\theta) = \frac{\lambda^n \, \mathrm{e}^{-\lambda}}{n!} \prod_{i=1}^{n} f(x_i|\theta) \tag{6.27}$$

Daraus folgt für die Log-Likelihood-Funktion:

$$\mathcal{L}(x_1, \ldots, x_n|\theta) = n \ln \lambda - \lambda + \sum_{i=1}^{n} \ln f(x_i|\theta), \tag{6.28}$$

wobei der für die Maximierung irrelevante Term $(-\ln n!)$ weggelassen wurde.

Mit der Umrechnung

$$n \ln \lambda + \sum_{i=1}^{n} \ln f(x_i|\theta) = \sum_{i=1}^{n} (\ln f(x_i|\theta) + \ln \lambda) = \sum_{i=1}^{n} \ln (\lambda f(x_i|\theta)) \tag{6.29}$$

kann eine Funktion $g(x|\theta) = \lambda f(x|\theta)$ definiert werden, deren Normierung λ ist:

$$\int_\Omega g(x|\theta) dx = \lambda \int_\Omega f(x|\theta) dx = \lambda \tag{6.30}$$

Damit wird aus (6.28) die gängige Form der erweiterten Likelihood-Funktion (EML):

$$\mathcal{L}(x_1, \ldots, x_n|\theta) = \sum_{i=1}^{n} \ln g(x_i|\theta) - \int_\Omega g(x|\theta) dx \tag{6.31}$$

Dass $\mathcal{L}$ tatsächlich maximal wird, wenn der zusätzliche Term in (6.31) n ergibt, kann man sich folgendermaßen klar machen: Wir skalieren die Funktion $g(x|\theta)$ mit einem Faktor β und fragen uns, für welchen Wert von β die Likelihood-Funktion maximal wird:

$$\mathcal{L} = \sum_{i=1}^{n} \ln (\beta \, g(x_i|\theta)) - \int_\Omega \beta \, g(x|\theta) dx \tag{6.32}$$

Die Maximierungsbedingung bezüglich β lautet:

$$\frac{\partial \mathcal{L}}{\partial \beta} = \frac{n}{\beta} - \int_\Omega g(x|\theta) dx = 0 \qquad \Longrightarrow \qquad \beta = \frac{n}{\int_\Omega g(x|\theta) dx} \tag{6.33}$$

Man sieht also, dass für das tatsächlich gewählte $\beta = 1$ die Likelihood-Funktion für

$$n = \int_\Omega g(x|\theta)dx \tag{6.34}$$

maximal wird. Man kann sich vergewissern, dass diese Normierungsbedingung sogar exakt erfüllt wird, obwohl wir bei der Herleitung der EML von einer Poisson-Verteilung ausgegangen waren. Zusätzlich lernt man von diesem Beweis, dass man β auch anders wählen und damit andere Normierungsbedingungen erhalten kann. Naheliegend wäre zum Beispiel $\beta = 1/n$, womit sich nach (6.33) $\int_\Omega g(x|\theta)dx = 1$ ergibt[1].

Beispiel: Wir greifen das oben angeführte Beispiel auf: n gemessene Ereignisse sollen m verschiedenen Reaktionen zugeordnet werden, für jede Reaktion j gibt es die normierte Wahrscheinlichkeit $f_j(x)$, dass das Ereignis aus dieser Reaktion stammt. Die Funktion g wird dann definiert:

$$g(x|n_1,\ldots,n_m) = \sum_{j=1}^{m} n_j f_j(x) \quad \Longrightarrow \quad \int_\Omega g(x|n_1,\ldots,n_m)dx = \sum_{j=1}^{m} n_j \tag{6.35}$$

Mit der erweiterten Likelihood-Funktion

$$\mathcal{L}(x_1,\ldots,x_n|n_1,\ldots,n_m) = \sum_{i=1}^{n} \ln g(x_i|n_1,\ldots,n_m) - \sum_{j=1}^{m} n_j \tag{6.36}$$

wird die Bedingung $n = \sum_{j=1}^{m} n_j$ erfüllt.

Diesen Ansatz kann man für das Beispiel der Abb. 6.1 anwenden, wenn man aus den einzelnen Ereignissen eine Likelihood-Funktion ('unbinned likelihood') konstruieren will (statt aus den Histogrammeinträgen, wie vorher behandelt): die Anpassung soll dann n_S Signalereignisse und n_B Untergrundereignisse mit der Bedingung für die Gesamtzahl $n = n_S + n_B$ ergeben.

6.3.4 Freiheitsgrade und Zwangsbedingungen

Eine Anpassung einer Hypothese an eine Stichprobe kann nur gemacht werden, wenn die Anzahl der Parameter m höchstens gleich der Anzahl der Messwerte n ist. Die Anzahl der Freiheitsgrade ergeben sich dann zu:

$$n_F = n - m. \tag{6.37}$$

Jede unabhängige Zwangsbedingung trägt wie ein zusätzlicher Messwert bei, so dass sich für k_c Bedingungen ergibt:

$$n_F = n - m + k_c. \tag{6.38}$$

Ein positiver Wert von n_F erlaubt eine Verbesserung der Messung durch Ausgleich zwischen den Messwerten. Bei kinematischen Anpassungen spricht man von n_FC-Fit

[1] Diese Normierung ist zum Beispiel bei der in [Z. Phys. C16 (1982) 13] dargestellten Analyse benutzt worden.

('constrained fit'). Ein 4C-Fit ('four-C fit') ergibt sich zum Beispiel, wenn man $3n$ Impulskomponenten eines n-Teilchensystems gemessen hat, das System 4 Zwangsbedingungen durch die Viererimpuls-Erhaltung unterliegt und die $3n$ Impulskomponenten als Parameter des Systems angepasst werden. Es wäre nur ein 1C-Fit, wenn ein Teilchen nicht beobachtet würde (das man aber dann wegen der Zwangsbedingungen rekonstruieren kann).

6.4 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.

6.4.1 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:

$$V_{ij}(\theta) = \int (\hat{\theta}_i - \theta_i)\,(\hat{\theta}_j - \theta_j)\,L(x_1, \ldots, x_n|\theta)\,dx_1 \ldots dx_n \tag{6.39}$$

Hier ist also $\theta = (\theta_1, \ldots, \theta_m)$ der 'wahre' Parametersatz und $\hat{\theta}(x_1, \ldots, x_n)$ sind die Schätzungen, die man jeweils für eine Stichprobe erhält. Die Stichproben, 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:

$$V_{ij}(\theta) = \frac{\int (\theta_i - \hat{\theta}_i)\,(\theta_j - \hat{\theta}_j)\,L(x_1, \ldots, x_n|\theta)\,d\theta_1 \ldots d\theta_m}{\int L(x_1, \ldots, x_n|\theta)\,d\theta_1 \ldots d\theta_m} \tag{6.40}$$

Hier ist $\hat{\theta}$ die ML-Schätzung, die aus der einen gemessenen Stichprobe $(x_1, \ldots, x_n)$ bestimmt wurde. In der Formel (6.40) 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.

6.4.2 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:

$$\mathcal{L}(x_1,\ldots,x_n|\theta) = \mathcal{L}(x_1,\ldots,x_n|\hat{\theta}) + (\theta-\hat{\theta})\left.\frac{\partial\mathcal{L}}{\partial\theta}\right|_{\theta=\hat{\theta}} + \frac{1}{2}(\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j)\left.\frac{\partial^2\mathcal{L}}{\partial\theta_i\partial\theta_j}\right|_{\theta=\hat{\theta}} + \ldots \tag{6.41}$$

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

$$V_{ij}^{-1} = -\left.\frac{\partial^2\mathcal{L}}{\partial\theta_i\partial\theta_j}\right|_{\theta=\hat{\theta}} \tag{6.42}$$

Damit ergibt sich in der Umgebung des Maximums:

$$\mathcal{L}((x_1,\ldots,x_n|\theta) \approx \mathcal{L}_{max} - \frac{1}{2}(\theta-\hat{\theta})^T V^{-1}(\theta-\hat{\theta}) \tag{6.43}$$

und für die Likelihood-Funktion L folgt:

$$L((x_1,\ldots,x_n|\theta) \approx L_{max}\, e^{-\frac{1}{2}(\theta-\hat{\theta})^T V^{-1}(\theta-\hat{\theta})} \tag{6.44}$$

Das heisst, wenn die Likelihood-Funktion als Funktion der Parameter ein annähernd gaussisches Verhalten zeigt, kann die Varianz durch die zweiten Ableitungen entsprechend (6.42) 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:

$$\sigma_i^2 = \frac{1}{V_{ii}^{-1}} = \left(-\left.\frac{\partial^2\mathcal{L}}{\partial{\theta_i}^2}\right|_{\theta=\hat{\theta}}\right)^{-1} \tag{6.45}$$

6.4.3 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

$$L((x_1,\ldots,x_n|\theta) = const \tag{6.46}$$

festgelegt sind und einen bestimmten Wahrscheinlichkeitsinhalt η, entprechend einem Vertrauensniveau, haben. Bei zwei Parametern (θ_i,θ_j) ergibt sich zum Beispiel

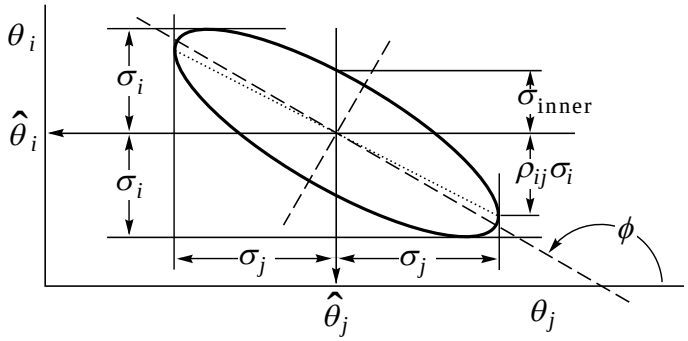

Abbildung 6.2: Standard-Fehlerellipse für die Schätzwerte $\hat{\theta}_i$ und $\hat{\theta}_j$.

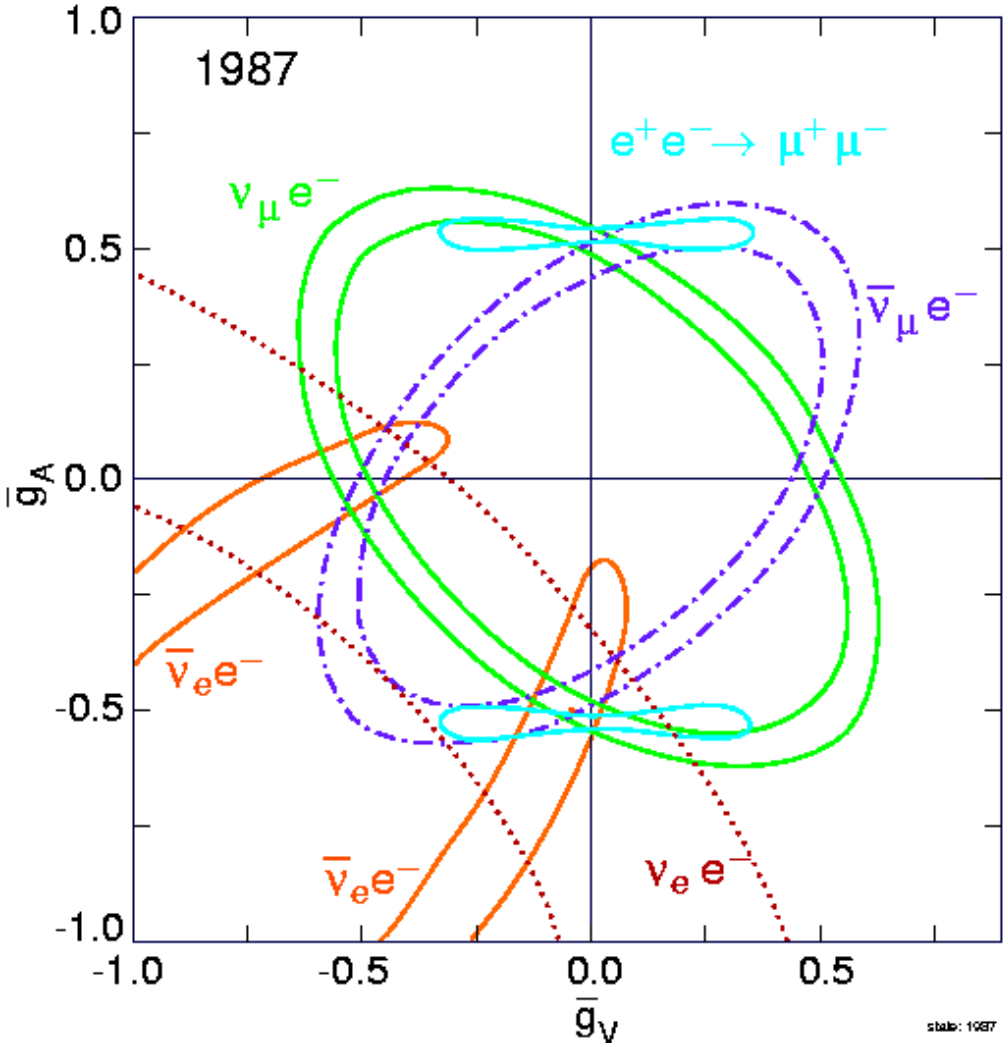

Abbildung 6.3: Beispiel für Likelihood-Kontouren: Die zwei Konstanten g_V und g_A (Kopplung von Leptonen an das Z^0-Boson) werden in verschiedenen Teilchenreaktionen gemessen, die sehr unterschiedliche Likelihood-Kontouren liefern. Die besten Schätzwerte liegen innerhalb der Kontouren, die ringförmige 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.

in der Regel eine geschlossene, zwei-dimensionale Raumkurve um die Schätzwerte $(\hat{\theta}_i, \hat{\theta}_j)$ der Parameter (Abb. 6.2). Im allgemeinen können die Hyperflächen beliebige Volumina im Parameterraum einschliessen, zum Beispiel brauchen diese Volumina auch nicht zusammenzuhängen (ein Beispiel ist in Abb. 6.3 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 (6.44) beschrieben werden kann, folgt

$$2\,\Delta\mathcal{L} \;=\; 2\,[\mathcal{L}_{max} \;-\; \mathcal{L}((x_1,\ldots,x_n|\theta)] \;=\; (\theta-\hat{\theta})^T\,V^{-1}\,(\theta-\hat{\theta}) \tag{6.47}$$

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. 6.2) nur 39.4% Wahrscheinlichkeit, während das für $m = 1$ bekanntlich 68.3% sind.

6.5 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
$$\theta \rightarrow \phi \tag{6.48}$$
ergibt sich:
$$\hat{\phi} = \phi(\hat{\theta}) \tag{6.49}$$
Zum Beispiel kann man für die Schätzung einer mittleren Lebensdauer τ auch die Zerfallswahrscheinlichkeit $\lambda = 1/\tau$ benutzen, denn aus
$$\frac{\partial L}{\partial \lambda}(\hat{\lambda}) = 0 \tag{6.50}$$
folgt
$$\frac{\partial L}{\partial \tau}\frac{\partial \tau}{\partial \lambda}(\hat{\lambda}) = 0 \;\Rightarrow\; \frac{\partial L}{\partial \tau}(\tau(\hat{\lambda})) = 0 \qquad (\frac{\partial \tau}{\partial \lambda} \neq 0) \tag{6.51}$$

2. Konsistenz: Für große Stichproben geht der Schätzwert in den tatsächlichen Wert über:
$$\lim_{n\rightarrow\infty} \hat{\theta} = \theta \tag{6.52}$$

3. Verzerrung: Wir hatten am Beispiel der Schätzung der Varianz einer Gauss-Verteilung gesehen (siehe (6.19)), 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:

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta \tag{6.53}$$

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 oder 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:

$$P(\theta_i|\vec{x}) = \frac{P(\vec{x}|\theta_i) \cdot P(\theta_i)}{\sum_j P(\vec{x}|\theta_j) \cdot P(\theta_j)} \tag{6.54}$$

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 verschiedener 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:

$$P(i|m) = \frac{P(m|i) \cdot P(i)}{\sum_j P(m|j) \cdot P(j)} \tag{6.55}$$

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 einer Teilchensorte 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

$$\theta \to \phi(\theta) \tag{6.56}$$

ergibt sich für die Berechnung eines Vertrauensintervalls:

$$\int_{\theta_1}^{\theta_2} L(\vec{x}|\theta)\, d\theta = \int_{\phi(\theta_1)}^{\phi(\theta_2)} L(\vec{x}|\phi(\theta)) \left|\frac{\partial\theta}{\partial\phi}\right| d\phi \neq \int_{\phi_1}^{\phi_2} L(\vec{x}|\phi)\, d\phi. \tag{6.57}$$

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