4

Bayes-Schätzung

4.1 Überblick

Die Bayes-Schätzung gehört zu den wichtigsten Konzepten der Signalverarbeitung. Sie stellt die Verallgemeinerung und damit ein Rahmenwerk für einen Großteil klassischer und moderner Schätzalgorithmen dar, so unter anderem für die *Maximum-Likelihood*-Schätzung, die Schätzung nach dem Prinzip des kleinsten mittleren Fehlerquadrats und die *Maximum-a-posteriori*-Schätzung.

Die Bayes-Schätzung basiert auf dem Satz von Bayes:

> Mit den Auftretenswahrscheinlichkeiten $P(A)$ und $P(B)$ für die Ereignisse A und B, den bedingten Wahrscheinlichkeiten[1] $P(A|B)$ und $P(B|A)$ sowie der Wahrscheinlichkeit $P(A, B)$ des gemeinsamen Auftretens der Ereignisse A und B gilt
>
> $$P(A|B)P(B) = P(B|A)P(A) = P(A, B). \tag{4.1}$$

Direkt aus dem Satz von Bayes folgt

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{4.2}$$

Die Wahrscheinlichkeit $P(A)$ in Gleichung 4.2 wird als *A-priori*-Wahrscheinlichkeit bezeichnet, da sie die Wahrscheinlichkeit des Auftretens des Ereignisses A ohne Wissen um das Ereignis B wiedergibt. Die *A-priori*-Wahrscheinlichkeit repräsentiert im Bayes'schen Ansatz das Vorwissen. $P(A|B)$ ist die

[1] Die *bedingte Wahrscheinlichkeit* $P(A|B)$ ist die Auftretenswahrscheinlichkeit von Ereignis A, wenn das Ereignis B bei einem anderen Zufallsexperiment oder bei einer vorherigen Durchführung des gleichen Zufallsexperimentes bereits eingetreten ist.

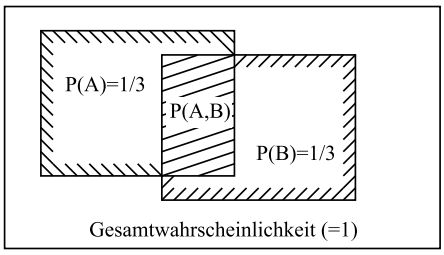

Abb. 4.1. Veranschaulichung des Satzes von Bayes ($P(B|A) = 1/4$, $P(A|B) = 1/4$, $P(A, B) = 1/12$)

A-posteriori-Wahrscheinlichkeit, die die Wahrscheinlichkeit des Ereignisses A nach Eintreffen des Ereignisses B angibt.

Analog zu den Auftretenswahrscheinlichkeiten in Gleichung 4.2 gilt für Verteilungsdichten der kontinuierlichen Zufallsvariablen X und Y

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}. \tag{4.3}$$

4.2 Schätztheorie

4.2.1 Zielstellung

Die Schätztheorie beschäftigt sich zum einen mit der Entwicklung von Schätzalgorithmen, zum anderen aber auch mit der Bewertung und dem Vergleich verschiedener Schätzverfahren. Die Aufgabe von Schätzverfahren besteht in der Schätzung von Parametern von z.B. Verteilungsdichtefunktionen[2] bzw. statistischen Modellen[3], der Schätzung der Koeffizienten prädiktiver Modelle[4] oder auch der Schätzung von Signalen aus getätigten Beobachtungen.

[2] z.B. Mittelwert und Varianz

[3] Ein statistisches bzw. probabilistisches Modell beschreibt die zufälligen Signalfluktuationen mit Verteilungsdichtefunktionen und deren Parametern wie z.B. Mittelwert, Varianz, Kovarianz. Die Motivation für die Anwendung von Modellen in der Signalverarbeitung ergibt sich aus der Möglichkeit der Verknüpfung von im Modell enthaltenem Vorwissen mit den beobachteten Signalwerten. Auf diesem Wege ist zum Beispiel der Ausschluss abwegiger Messdaten möglich bzw. allgemein die Gewichtung von Vorwissen und beobachteten Signalwerten entsprechend ihrer statistischen Sicherheit. Darüber hinaus geben Modelle die Möglichkeit, Struktur, Zusammensetzung und Entstehung von Signalen zu verstehen und gezielt zu beeinflussen.

[4] Ein prädiktives Modell beschreibt die Korrelationsstruktur eines Signals, siehe z.B. lineare Prädiktion. Ebenso können auch konditionale probabilistische Modelle die Korrelationsstruktur erfassen.

4.2.2 Bewertungskriterien für Schätzer

Schätzalgorithmen[5] verarbeiten beobachtete Signalwerte. Diese Signalwerte sind in gewissem Grade zufällig oder unterliegen zufälligen Schwankungen. Deshalb sind die Ergebnisse eines Schätzers gleichfalls zufällig und können z.B. mit einer Verteilungsfunktion statistisch beschrieben werden. Mit Hilfe dieser statistischen Beschreibung der Schätzergebnisse ist deren Bewertung möglich.

Wichtig für die Bewertung eines Schätzverfahrens sind vor allem folgende Kennwerte:

- der Erwartungswert $E[\hat{\boldsymbol{\theta}}]$: Mit ihm wird das mittlere Ergebnis eines Schätzalgorithmus angegeben[6].
- der systematische Fehler[7] $E[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}] = E[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}$: Mit ihm wird die durchschnittliche Abweichung des Schätzergebnisses vom wahren Parameterwert bemessen.
- die Kovarianz der Schätzung $\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = E[(\hat{\boldsymbol{\theta}} - E[\hat{\boldsymbol{\theta}}])(\hat{\boldsymbol{\theta}} - E[\hat{\boldsymbol{\theta}}])^T]$: Sie ist ein Index für die Schwankung des Schätzergebnisses um den mittleren Schätzwert sowie Abhängigkeiten der verschiedenen Schätzwertschwankungen untereinander.

Ein guter Schätzer weist keinen systematischen Fehler sowie eine möglichst kleine Kovarianz auf. Mit Bezug auf die Bewertungskriterien können folgende Eigenschaften von Schätzalgorithmen definiert werden:

- erwartungstreuer Schätzer[8]:

$$E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta} \tag{4.4}$$

  Der Schätzalgorithmus ermittelt den wahren Parameterwert.

- asymptotisch erwartungstreuer Schätzer[9]:

$$\lim_{N \to \infty} E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta} \tag{4.5}$$

  Das Ergebnis des Schätzalgorithmus nähert sich mit größerer Datenanzahl N an den wahren Parameterwert an.

[5] Schätzalgorithmen werden auch als Schätzer bezeichnet.
[6] $\hat{\boldsymbol{\theta}}$ bezeichnet den Schätzwert für den zu schätzenden Parameter $\boldsymbol{\theta}$.
[7] engl.: bias
[8] engl.: unbiased estimator
[9] engl.: asymptotically unbiased estimator

- konsistenter Schätzer[10]:

$$\lim_{N\to\infty} P[|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| > \varepsilon] = 0, \tag{4.6}$$

mit $\varepsilon > 0$ und beliebig klein. Gleichung 4.6 beschreibt eine Konvergenz in Wahrscheinlichkeit und definiert die einfache bzw. schwache Konsistenz. Von starker Konsistenz spricht man bei Konvergenz mit Wahrscheinlichkeit 1 [13]

$$P[\lim_{N\to\infty} \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}] = 1 \tag{4.7}$$

und von Konsistenz im quadratischen Mittel bei

$$\lim_{N\to\infty} E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})] = 0. \tag{4.8}$$

Starke Konsistenz und Konsistenz im quadratischen Mittel schließen einfache Konsistenz mit ein.

- wirksamer Schätzer[11]:

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}_{wirksam}) \leq \mathrm{Cov}(\hat{\boldsymbol{\theta}}) \tag{4.9}$$

Die Varianz des Schätzergebnisses besitzt im eindimensionalen Fall, d.h. $\boldsymbol{\theta} = \theta$, den kleinstmöglichen Wert. Im Falle mehrdimensionaler Parametervektoren berührt $\mathrm{Cov}(\hat{\boldsymbol{\theta}}_{wirksam})$ die Cramér-Rao-Schranke (siehe Abschnitt 4.2.3).

Beim Entwurf von Schätzalgorithmen ist man im Allgemeinen bemüht, wirksame Verfahren zu entwickeln. Dies ist mit den im Abschnitt 4.4 beschriebenen Maximum-Likelihood-Schätzern (ML-Schätzern) prinzipiell möglich. Allerdings sind ML-Schätzer oft schwierig zu berechnen, so dass in diesen Fällen auf einfachere, meist nicht-wirksame Methoden, z.B. Momenten-Verfahren, zurückgegriffen werden muss.

4.2.3 Die Cramér-Rao-Schranke

Das Kriterium der Wirksamkeit soll im Folgenden noch einmal näher betrachtet werden. Prinzipiell können Schätzalgorithmen einen Parameterwert nicht beliebig genau bestimmen. Es gibt vielmehr eine untere Schranke für die Varianz einer Schätzung. Die Varianz $\mathrm{var}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$ eines erwartungstreuen Schätzers ist von unten beschränkt durch die Cramér-Rao-Schranke (CR-Schranke), die sich aus der Verteilungsdichte des beobachteten Signals[12] $f(y|\theta)$ mit

[10] engl.: consistent estimator

[11] engl.: efficient estimator

[12] Aus Übersichtlichkeitsgründen wird die bedingte Verteilungsdichte $f_{Y|\Theta}(y|\theta)$ im Folgenden mit $f(y|\theta)$ notiert.

$$\mathrm{var}(\hat{\theta}) \geq \left. \frac{1}{E\left[(\partial \log f(y|\theta)/\partial\theta)^2\right]} \right|_{\theta=\substack{\text{wahrer}\\ \text{Parameterwert}}} = I^{-1}(\theta) \tag{4.10}$$

ergibt. $I(\theta)$ ist die Fisher-Information.

Zur Herleitung von Gleichung 4.10 wendet man die Cauchy-Schwarz'sche Ungleichung[13] auf $E[(\partial \log f(y|\theta)/\partial\theta)(\hat{\theta}-\theta)]$ an und erhält

$$(E[(\partial \log f(y|\theta)/\partial\theta)(\hat{\theta}-\theta)])^2 \leq E[(\partial \log f(y|\theta)/\partial\theta)^2]E[(\hat{\theta}-\theta)^2]. \tag{4.11}$$

Ferner gilt bei vorausgesetzter Vertauschbarkeit von Differentiation und Integration sowie $E[\hat{\theta}] = \theta$

$$E\left[\frac{\partial \log f(y|\theta)}{\partial\theta}(\hat{\theta}-\theta)\right] = \int \frac{\partial \log f(y|\theta)}{\partial\theta}(\hat{\theta}-\theta)f(y|\theta)dy \tag{4.12}$$

$$= \int \frac{\partial f(y|\theta)}{\partial\theta}(\hat{\theta}-\theta)dy \tag{4.13}$$

$$= \int \frac{\partial f(y|\theta)}{\partial\theta}\hat{\theta}dy - \theta\int \frac{\partial f(y|\theta)}{\partial\theta}dy \tag{4.14}$$

$$= \frac{\partial}{\partial\theta}\int f(y|\theta)\hat{\theta}dy - \theta\frac{\partial}{\partial\theta}\int f(y|\theta)dy \tag{4.15}$$

$$= \frac{\partial}{\partial\theta}E[\hat{\theta}] - \theta\frac{\partial}{\partial\theta}1 = \frac{\partial}{\partial\theta}\theta - \theta\cdot 0 = 1. \tag{4.16}$$

Wird dieses Ergebnis in Gleichung 4.11 eingesetzt, ergibt sich sofort die Cramér-Rao-Schranke. Ein Schätzer, der die CR-Schranke erreicht, ist wirksam.

Soll eine differenzierbare Funktion $g(\theta)$ geschätzt werden, ergibt sich die CR-Schranke mit

$$\mathrm{var}(\hat{g}) \geq \frac{(\partial g/\partial\theta)^2}{E\left[(\partial \log f(y|\theta)/\partial\theta)^2\right]}. \tag{4.17}$$

Für Parameter-Vektoren ist die CR-Schranke für die Kovarianz-Matrix des Parametervektors $\mathbf{C}_{\hat{\boldsymbol{\theta}}}$ gegeben mit

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} \geq \mathbf{I}_{\boldsymbol{\theta}}^{-1}, \tag{4.18}$$

wobei $\mathbf{I}_\theta$ die Informations-Matrix bezeichnet und Gleichung 4.18 die positive Semidefinitheit von $\mathbf{C}_{\hat{\boldsymbol{\theta}}} - \mathbf{I}_{\boldsymbol{\theta}}^{-1}$ bedeutet. Die Informations-Matrix ist definiert mit

$$\mathbf{I}_{\boldsymbol{\theta}} = E\left[\left(\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\right]\Bigg|_{\boldsymbol{\theta}=\substack{\text{wahrer}\\ \text{Parameterwert}}}. \tag{4.19}$$

[13] $(E[x_1x_2])^2 \leq E[x_1^2]E[x_2^2]$

Soll eine differenzierbare vektorwertige Funktion $\mathbf{g}(\boldsymbol{\theta})$ geschätzt werden, ergibt sich die CR-Schranke aus

$$\mathbf{C}(\hat{\mathbf{g}}) \geq \mathbf{J}\mathbf{I}_{\boldsymbol{\theta}}^{-1}\mathbf{J}^T, \tag{4.20}$$

mit $\mathbf{J}$ als Jacobi-Matrix der Funktion $\mathbf{g}(\boldsymbol{\theta})$.

4.3 Verfahren der Bayes-Schätzung

4.3.1 Die Berechnung der A-posteriori-Verteilungsdichte

Das Bayes'sche Konzept beruht auf der folgenden Idee: Vorhandenes Wissen über mögliche Signal- oder Parameterwerte kann zur Bewertung beobachteter Signal- oder geschätzter Parameterwerte genutzt werden. Das Vorwissen[14] legt einen *A-priori*-Raum aller möglichen Werte des Signals bzw. des zu schätzenden Parameters fest. Nach erfolgter Beobachtung, d.h. mit Kenntnis des gemessenen Signals entsteht ein *A-posteriori*-Raum, in dem alle sowohl mit dem Vorwissen als auch mit der Beobachtung konsistenten Werte enthalten sind.

Die Umsetzung dieser Idee erfolgt mit der Verknüpfung von *A-priori*- und *A-posteriori*-Verteilungsdichte gemäß dem Bayes'schen Satz. Zum Beispiel erhält man für einen beobachteten Signalvektor $\mathbf{y} = [y_1, y_2, \ldots, y_N]$ und den gesuchten Parametervektor $\boldsymbol{\theta}$

$$f_{\Theta|Y}(\boldsymbol{\theta}|\mathbf{y}) = \frac{f_{Y|\Theta}(\mathbf{y}|\boldsymbol{\theta}) f_{\Theta}(\boldsymbol{\theta})}{f_Y(\mathbf{y})} = \frac{f_{Y|\Theta}(\mathbf{y}|\boldsymbol{\theta}) f_{\Theta}(\boldsymbol{\theta})}{\int\limits_{\theta} f_{Y|\Theta}(\mathbf{y}|\boldsymbol{\theta}) f_{\Theta}(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{4.21}$$

Die *A-posteriori*-Verteilungsdichte $f_{\Theta|Y}(\boldsymbol{\theta}|\mathbf{y})$ ist proportional zum Produkt von *A-priori*-Verteilungsdichte $f_{\Theta}(\boldsymbol{\theta})$ und der aus der Beobachtung resultierenden Likelihood-Funktion $f_{Y|\Theta}(\mathbf{y}|\boldsymbol{\theta})$. Der Prior wichtet also die Beobachtung und sorgt so für deren Konsistenz mit dem Vorwissen.

Die Verteilungsdichte $f_Y(\mathbf{y})$ ist für eine konkrete Beobachtung konstant und besitzt lediglich die Aufgabe der Normierung. Allerdings ist $f_Y(\mathbf{y})$ wegen des Integrals im Nenner des rechten Terms von Gleichung 4.21 oft schwierig zu berechnen. Im Falle gaußscher Verteilungsdichten ist jedoch die analytische Berechnung von $f_Y(\mathbf{y})$ und damit auch von $f_{\Theta|Y}(\boldsymbol{\theta}|\mathbf{y})$ relativ einfach, wie das folgende Beispiel zeigt.

Beispiel 4.1. Gesucht ist die *A-posteriori*-Verteilungsdichte für den Mittelwert μ bei gaußschem Prior und N gaußverteilten, voneinander statistisch unabhängigen Messdaten. Die Varianz der Messdaten wird als bekannt vorausgesetzt.

[14] engl.: prior

Der Prior für den Mittelwert ist gegeben mit

$$f(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right), \tag{4.22}$$

wobei die Vorinformation sowohl in der gaußschen Verteilungsdichte als auch in den konkreten Werten für Varianz und Mittelwert, d.h. σ_0^2 bzw. μ_0 besteht. Da die N beobachteten Werte y_i laut Aufgabenstellung einer gaußschen Verteilungsdichte mit bekannter Varianz σ^2 entstammen und darüber hinaus untereinander statistisch unabhängig sind, ergibt sich die Likelihood-Funktion mit

$$f_{Y|\mu}(\mathbf{y}|\mu) = \prod_{i=1}^{N} f(y_i|\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right) \tag{4.23}$$

$$= (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{\sum_{i=1}^{N}(y_i-\mu)^2}{2\sigma^2}\right). \tag{4.24}$$

Die Verteilungsdichte $f(\mathbf{y})$ ergibt sich unter Berücksichtigung der Gleichungen 4.22 und 4.24 zu

$$f(\mathbf{y}) = \int_{-\infty}^{\infty} f_{Y|\mu}(\mathbf{y}|\mu) f(\mu) d\mu = \int_{-\infty}^{\infty} \prod_{i=1}^{N} f(y_i|\mu) f(\mu) d\mu$$

$$= \frac{(2\pi\sigma^2)^{-N/2}}{\sqrt{2\pi\sigma_0^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{\sigma^2(\mu-\mu_0)^2 + \sigma_0^2\sum_{i=1}^{N}(y_i-\mu)^2}{2\sigma_0^2\sigma^2}\right) d\mu$$

$$= \frac{(2\pi\sigma^2)^{-N/2}}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{\sigma^2\mu_0^2 + \sigma_0^2\sum_{i=1}^{N} y_i^2}{2\sigma_0^2\sigma^2}\right)$$

$$\times \int_{-\infty}^{\infty} \exp\left(-\frac{\mu^2(\sigma^2+N\sigma_0^2)}{2\sigma_0^2\sigma^2} + \frac{\mu(\mu_0\sigma^2 + \sigma_0^2\sum_{i=1}^{N} y_i)}{\sigma_0^2\sigma^2}\right) d\mu. \tag{4.25}$$

Mit [14]

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}A\cdot\mu^2 + h\cdot\mu\right) d\mu = \sqrt{\frac{2\pi}{A}} \exp\left(\frac{h^2}{2A}\right) \tag{4.26}$$

und

$$A = \frac{\sigma^2 + N\sigma_0^2}{\sigma_0^2\sigma^2} \qquad h = \frac{\sigma^2\mu_0 + \sigma_0^2\sum_{i=1}^{N} y_i}{\sigma_0^2\sigma^2} \tag{4.27}$$

erhält man

$$f(\mathbf{y}) \propto \exp\left(-\frac{1}{2}\frac{\sigma^2\mu_0^2 + \sigma_0^2\sum_{i=1}^N y_i^2}{\sigma_0^2\sigma^2} + \frac{1}{2}\frac{(\sigma^2\mu_0 + \sigma_0^2\sum_{i=1}^N y_i)^2}{\sigma_0^2\sigma^2(\sigma^2 + N\sigma_0^2)}\right). \quad (4.28)$$

Nach Einsetzen der Gleichungen 4.22, 4.24 und 4.28 in Gleichung 4.21 ergibt sich die gaußsche *A-posteriori*-Verteilungsdichte $f_{\mu|Y}(\mu|\mathbf{y})$ mit dem Mittelwert und der Varianz

$$\mu_N = \mu_0\frac{\sigma^2}{\sigma^2 + N\sigma_0^2} + \frac{\sigma_0^2}{\sigma^2 + N\sigma_0^2}\sum_{i=1}^N y_i \quad (4.29)$$

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \qquad \text{bzw.} \quad \sigma_N^2 = \sigma_0^2\frac{\sigma^2}{N\sigma_0^2 + \sigma^2}. \quad (4.30)$$

Aus den Gleichungen 4.29 und 4.30 wird die Abhängigkeit der Parameterschätzwerte von der Datenanzahl deutlich. Wenn keine Messdaten vorhanden sind, gleichen die Parameterschätzwerte den Parameterwerten der *A-priori*-Verteilungsdichte. Für große N schrumpft jedoch der Einfluss des Vorwissens. Abbildung 4.2 zeigt dieses Prinzip qualitativ.

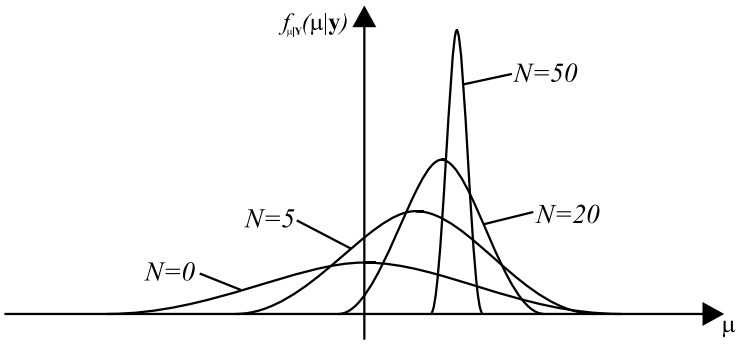

Abb. 4.2. Bayes-Lernen: *A-posteriori*-Verteilungsdichte in Abhängigkeit von der Datenanzahl N (qualitative Darstellung)

□

4.3.2 Klassische Schätzverfahren im Kontext der Bayes-Schätzung

Die Bayes'sche Risikofunktion

Während im vorigen Abschnitt die *A-posteriori*-Verteilungsdichte mit ihren Parametern direkt berechnet wurde, erfolgt die Parameterschätzung nun durch die Minimierung der Bayes'schen Risikofunktion

$$R(\hat{\theta}) = E[C(\hat{\theta}, \theta)] \tag{4.31}$$

$$= \int_{\theta} \int_{y} C(\hat{\theta}, \theta) f_{Y,\Theta}(y, \theta) dy d\theta \tag{4.32}$$

$$= \int_{\theta} \int_{y} C(\hat{\theta}, \theta) f_{\Theta|Y}(\theta|y) f_Y(y) dy d\theta, \tag{4.33}$$

mit $C(\hat{\theta}, \theta)$ als gewichtete Fehlerfunktion, die die um einen Gewichtsfaktor ergänzten Schätzfehler enthält.[15] Da bei gegebenem Beobachtungsvektor die Verteilungsdichte $f_Y(y)$ konstant, d.h. ohne Einfluss auf die Optimierung ist, kann die Risikofunktion reduziert werden auf (bedingte Risikofunktion)

$$R(\hat{\theta}|y) = \int_{\theta} C(\hat{\theta}, \theta) f_{\Theta|Y}(\theta|y) d\theta. \tag{4.34}$$

Die Minimierung der bedingten Risikofunktion führt zum Bayes-Schätzwert des Parameters θ

$$\hat{\theta}_{Bayes} = \underset{\hat{\theta}}{\operatorname{argmin}}\ R(\hat{\theta}|y) \tag{4.35}$$

$$= \underset{\hat{\theta}}{\operatorname{argmin}} \left(\int_{\theta} C(\hat{\theta}, \theta) f_{\Theta|Y}(\theta|y) d\theta \right) \tag{4.36}$$

$$= \underset{\hat{\theta}}{\operatorname{argmin}} \left(\int_{\theta} C(\hat{\theta}, \theta) f_{Y|\Theta}(y|\theta) f_{\Theta}(\theta) d\theta \right). \tag{4.37}$$

In Abhängigkeit von gewichteter Fehlerfunktion $C(\hat{\theta}, \theta)$ und Prior $f_{\Theta}(\theta)$ ergeben sich verschiedene Schätzungen, unter anderem die im Folgenden besprochene *Maximum-a-posteriori*-Schätzung (MAP), die *Maximum-Likelihood*-Schätzung (ML), die Schätzung nach dem kleinsten mittleren Fehlerquadrat (minimum mean square error, MMSE) und die Schätzung nach dem kleinsten mittleren Absolutfehler (minimum mean absolute value of error, MAVE).

[15] Die Berechnung der Risikofunktion kann als dreistufiges Verfahren interpretiert werden: Zunächst werden für alle denkbaren Kombinationen von y und θ die Schätzwerte $\hat{\theta}$ berechnet. Anschließend bewertet die gewichtete Fehlerfunktion $C(\hat{\theta}, \theta)$ die entstandenen Schätzfehler. Zum Beispiel können größere Fehler stärker gewichtet werden als kleine. Die Mittelung der bewerteten Schätzfehler im dritten Schritt liefert schließlich *einen einzigen* Kennwert für die Güte der Schätzung. Schätzverfahren können nun so entworfen werden, dass sie diesen Kennwert minimieren.

Maximum-a-posteriori-Schätzung

Die gewichtete Fehlerfunktion der MAP-Schätzung ist gegeben mit

$$C_{MAP}(\hat{\theta}, \theta) = 1 - \delta(\hat{\theta}, \theta). \tag{4.38}$$

Aus Gleichung 4.34 folgt somit für die bedingte Risikofunktion[16]

$$R_{MAP}(\hat{\theta}|y) = \int_{\theta} [1 - \delta(\hat{\theta}, \theta)] f_{\Theta|Y}(\theta|y) d\theta \tag{4.39}$$

$$= 1 - f_{\Theta|Y}(\hat{\theta}|y). \tag{4.40}$$

Das Minimum der bedingten Risikofunktion $R_{MAP}(\hat{\theta}|y)$ ist genau dort, wo die bedingte Verteilungsdichte $f_{\Theta|Y}(\hat{\theta}|y)$ maximal wird. Für den optimalen MAP-Schätzwert $\hat{\theta}_{MAP}$ folgt deshalb

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\ f_{\Theta|Y}(\theta|y) = \underset{\theta}{\operatorname{argmax}}\ f_{Y|\Theta}(y|\theta) f_{\Theta}(\theta). \tag{4.41}$$

Maximum-Likelihood-Schätzung

Die ML-Schätzung besitzt die gleiche gewichtete Fehlerfunktion wie die MAP-Schätzung

$$C_{ML}(\hat{\theta}, \theta) = 1 - \delta(\hat{\theta}, \theta). \tag{4.42}$$

Folglich erhält man die gleiche bedingte Risikofunktion

$$R_{ML}(\hat{\theta}|y) = \int_{\theta} [1 - \delta(\hat{\theta}, \theta)] f_{\Theta|Y}(\theta|y) d\theta \tag{4.43}$$

$$= 1 - f_{\Theta|Y}(\hat{\theta}|y). \tag{4.44}$$

Analog zur MAP-Schätzung liegt das Minimum der bedingten Risikofunktion $R_{ML}(\hat{\theta}|y)$ beim Maximum der bedingten Verteilungsdichte $f_{\Theta|Y}(\hat{\theta}|y)$, d.h.

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\ f_{\Theta|Y}(\theta|y) = \underset{\theta}{\operatorname{argmax}}\ f_{Y|\Theta}(y|\theta) f_{\Theta}(\theta). \tag{4.45}$$

[16] unter Berücksichtigung der Ausblendeigenschaft der Deltafunktion

$$f(\hat{\theta}) = \int_{-\infty}^{\infty} f(\theta) \delta(\hat{\theta}, \theta) d\theta$$

und der Flächeneigenschaft der Verteilungsdichte

$$\int_{-\infty}^{\infty} f(\theta) d\theta = 1$$

Der Unterschied zur MAP-Schätzung resultiert aus einem anderen Prior. Bei der ML-Schätzung wird eine konstante *A-priori*-Verteilungsdichte $f_\Theta(\theta) = const.$ angenommen. Alle Werte von θ sind aufgrund dieser *A-priori*-Verteilungsdichte zunächst gleichwahrscheinlich. Es fließt demnach kein Vorwissen über den Parameter θ in die Schätzung ein. Der ML-Schätzer ist demnach gegeben mit

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\ f_{Y|\Theta}(y|\theta). \tag{4.46}$$

Die Maximum-Likelihood-Schätzung wird im Abschnitt 4.4 vertieft.

Schätzung nach dem kleinsten mittleren Fehlerquadrat

Die gewichtete Fehlerfunktion ergibt sich bei diesem Schätzverfahren aus der quadrierten Differenz zwischen Schätzwert $\hat{\theta}$ und wahrem Parameter θ

$$C_{MSE}(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2. \tag{4.47}$$

Daraus folgt die bedingte Risikofunktion mit

$$R_{MSE}(\hat{\theta}|y) = E[(\hat{\theta} - \theta)^2|y] \tag{4.48}$$

$$= \int_\theta (\hat{\theta} - \theta)^2 f_{\Theta|Y}(\theta|y) d\theta. \tag{4.49}$$

Das Minimum der bedingten Risikofunktion erhält man mit

$$\frac{\partial R_{MSE}(\hat{\theta}|y)}{\partial \hat{\theta}} = 2\hat{\theta} \underbrace{\int_\theta f_{\Theta|Y}(\theta|y) d\theta}_{=1} - 2 \int_\theta \theta f_{\Theta|Y}(\theta|y) d\theta \tag{4.50}$$

$$= 2\hat{\theta} - 2 \int_\theta \theta f_{\Theta|Y}(\theta|y) d\theta \stackrel{!}{=} 0. \tag{4.51}$$

Somit entspricht die MMSE-Schätzung dem bedingten Erwartungswert

$$\hat{\theta}_{MMSE} = \int_\theta \theta f_{\Theta|Y}(\theta|y) d\theta = E[\theta|y]. \tag{4.52}$$

Schätzung nach dem kleinsten mittleren Absolutfehler

Als gewichtete Fehlerfunktion wird der Absolutwert der Abweichung zwischen Schätzwert $\hat{\theta}$ und wahrem Parameter θ genutzt

$$C_{MAVE}(\hat{\theta}, \theta) = |\hat{\theta} - \theta|. \tag{4.53}$$

Die daraus resultierende bedingte Risikofunktion ist

$$R_{MAVE}(\hat{\theta}|y) = E[|\hat{\theta} - \theta|] \tag{4.54}$$

$$= \int_{\theta} |\hat{\theta} - \theta| f_{\Theta|Y}(\theta|y) d\theta \tag{4.55}$$

$$= \int_{-\infty}^{\hat{\theta}(y)} [\hat{\theta} - \theta] f_{\Theta|Y}(\theta|y) d\theta + \int_{\hat{\theta}(y)}^{\infty} [\theta - \hat{\theta}] f_{\Theta|Y}(\theta|y) d\theta. \tag{4.56}$$

Die Minimierung erfolgt mit

$$\frac{\partial R_{MAVE}(\hat{\theta}|y)}{\partial \hat{\theta}} = \int_{-\infty}^{\hat{\theta}(y)} f_{\theta|Y}(\theta|y) d\theta - \int_{\hat{\theta}(y)}^{\infty} f_{\theta|Y}(\theta|y) d\theta \stackrel{!}{=} 0. \tag{4.57}$$

Durch Umstellen ergibt sich unmittelbar

$$\int_{-\infty}^{\hat{\theta}(y)} f_{\theta|Y}(\theta|y) d\theta = \int_{\hat{\theta}(y)}^{\infty} f_{\theta|Y}(\theta|y) d\theta. \tag{4.58}$$

Die MAVE-Schätzung $\hat{\theta}_{MAVE}$ ergibt somit den Median[17] der *A-posteriori*-Verteilungsdichte für den Parameter θ.

Die Abbildungen 4.3 und 4.4 fassen die verschiedenen gewichteten Fehlerfunktionen und die daraus resultierenden Schätzwerte am Beispiel einer bimodalen Verteilungsdichte noch einmal zusammen.

4.4 Die Maximum-Likelihood-Schätzung

Aufgrund ihrer hervorragenden Eigenschaften ist die ML-Schätzung ein sehr beliebtes Schätzverfahren. Sie ist *asymptotisch* erwartungstreu und gleichfalls wirksam. Die Schätzergebnisse sind gaußverteilt. Die ML-Schätzung liefert in der Regel auch bei kurzen Datensätzen gute Ergebnisse, obwohl die asymptotischen Eigenschaften dann keine Gültigkeit besitzen.

Die ML-Schätzung besitzt die Eigenschaft der Invarianz [57]. Dies bedeutet, dass die ML-Schätzung einer Funktion $\mathbf{g}(\boldsymbol{\theta})$ auch durch die Transformation des geschätzten Parametervektors $\hat{\boldsymbol{\theta}}$ berechnet werden kann, d.h.

[17] Der Median ist der Punkt einer Verteilung, für den größere Werte und kleinere Werte gleichwahrscheinlich sind, also $P(x < x_{Median}) = P(x > x_{Median})$.

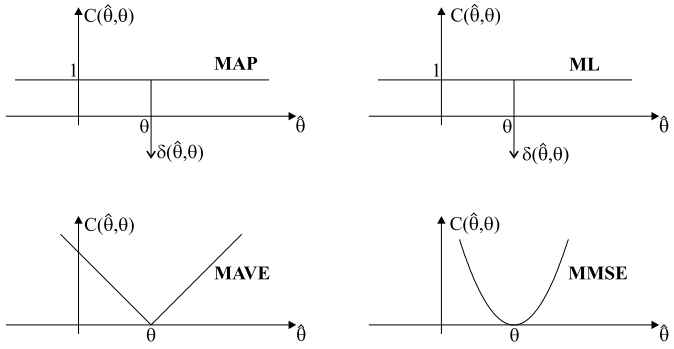

Abb. 4.3. Klassische gewichtete Fehlerfunktionen im Kontext der Bayes-Schätzung

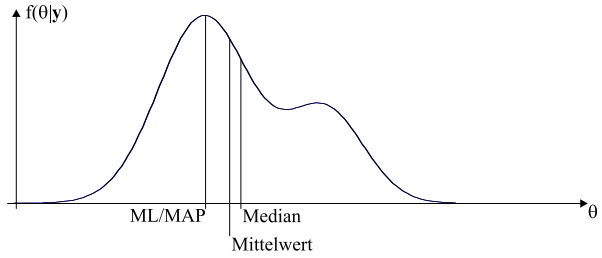

Abb. 4.4. Beispielhafte Resultate der klassischen Schätzverfahren im Kontext der Bayes-Schätzung (MAP für den Fall fehlender Vorinformation)

$$\hat{\mathbf{g}} = \mathbf{g}(\hat{\boldsymbol{\theta}}). \tag{4.59}$$

Gemäß Gleichung 4.46 muss im Rahmen der ML-Schätzung *der* Parametervektor $\hat{\boldsymbol{\theta}}$ gefunden werden, der die Likelihood-Funktion maximiert. Die vorliegenden Messwerte repräsentieren dann den wahrscheinlichsten Messwertvektor, der mit diesem gefundenen Parametervektor $\hat{\boldsymbol{\theta}}$ und der angenommenen Verteilungsdichtefunktion gemessen werden kann.[18]

Insbesondere bei Verteilungsdichten der exponentiellen Familie wird anstelle der Likelihood-Funktion L die Log-Likelihood-Funktion $\log L$ verwendet. Wegen der strengen Monotonie der Logarithmusfunktion ergeben sich für Likelihood- und Log-Likelihood-Funktion gleiche Optimalstellen. Der Grund für die Nutzung der Log-Likelihood-Funktion besteht in der sich vereinfachenden Rechnung, da zumindest bei exponentiellen Verteilungsdichten die Exponentialfunktion entfällt. Durch Produktbildung entstandene Verbund-

[18] Die der ML-Schätzung zugrunde liegende Idee ist: Der Messwertvektor $\mathbf{y}$ wurde gemessen; deshalb muss sein Auftreten sehr wahrscheinlich sein. Die ML-Schätzung gibt den Parametervektor, der den Messwertvektor zum Wahrscheinlichsten aller möglichen Messwertvektoren werden lässt.

verteilungsdichten gehen bei der Logarithmierung in einfache Summen über. Ferner wird durch die Logarithmierung der Dynamikbereich der Verteilungsdichten eingeschränkt, womit eine höhere Rechengenauigkeit erzielt werden kann. Nachteilig an der Log-Likelihood-Funktion ist die mitunter, im Vergleich zur Likelihood-Funktion, weniger deutliche Ausprägung der Optima.

Die Durchführung einer ML-Schätzung ist nicht immer praktikabel. Im Allgemeinen ergeben sich im Verlaufe der Schätzung nichtlineare Gleichungen, die gegebenenfalls aufwändig mit nummerischen Verfahren gelöst werden müssen. In einigen Fällen ist eine analytische Rechnung jedoch möglich. Dazu nutzt man häufig die Annahme, dass N voneinander statistisch unabhängige Messwerte $y_1, y_2, \ldots, y_N$ der kontinuierlichen Zufallsvariable Y beobachtet werden können[19]. Aufgrund der angenommenen statistischen Unabhängigkeit der Messwerte y_i resultiert die Likelihood-Funktion $L = f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta})$ aus der Multiplikation der Werte der Verteilungsdichtefunktion $f_{Y|\boldsymbol{\Theta}}(y_i|\boldsymbol{\theta})$ für die beobachteten y_i

$$L(\mathbf{y}|\boldsymbol{\theta}) = f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} f_{Y|\boldsymbol{\Theta}}(y_i|\boldsymbol{\theta}). \tag{4.60}$$

Beispiel 4.2. Aus N Messwerten y_i ist der wahrscheinlichste Mittelwert zu berechnen. Als Voraussetzung wird angenommen, dass alle Messwerte die gleiche gaußsche Verteilungsdichtefunktion mit gleicher unbekannter Varianz und gleichem unbekannten Mittelwert besitzen.

Die Likelihood-Funktion erhält man mit

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-(y_i - \hat{\mu})^2/(2\sigma^2)\right] \tag{4.61}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-1/(2\sigma^2)\sum_{i=1}^{N}(y_i - \hat{\mu})^2\right]. \tag{4.62}$$

Daraus ergibt sich die Log-Likelihood-Funktion

$$\log L = -\frac{N}{2}\ln(2\pi) - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{\mu})^2. \tag{4.63}$$

Aus der Ableitung der Log-Likelihood-Funktion

$$\frac{\partial \log L}{\partial \hat{\mu}} = 0 = \sum_{i=1}^{N}(y_i - \hat{\mu}) \tag{4.64}$$

entsteht die bekannte Gleichung für den Mittelwert

[19] Die Messwerte können zu einem Messwertvektor zusammengestellt werden, d.h. $\mathbf{y} = [y_1, y_2, \ldots, y_N]^T$.

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} y_i. \tag{4.65}$$

□

4.5 Der Expectation-Maximization-Algorithmus

4.5.1 Überblick

Der Expectation-Maximization-Algorithmus (EM-Algorithmus) gehört zu den wichtigsten Schätzalgorithmen der Signalverarbeitung. Er ist ein iteratives Verfahren und wird zur Maximum-Likelihood-Schätzung aus unvollständigen Daten verwendet [69]. Die Konvergenz des EM-Algorithmus wird unter anderem in [28] gezeigt.

4.5.2 Maximum-Likelihood-Schätzung mit unvollständigen Daten

Der EM-Algorithmus ermöglicht eine Parameter- oder Signalschätzung auf der Basis unvollständiger Daten. Unvollständig bedeutet in diesem Zusammenhang, dass nicht alle zur Schätzung notwendigen Informationen vorliegen. Der EM-Algorithmus versucht, diese fehlenden Informationen aus den vorhandenen Daten und vorhandenen Schätzwerten für die gesuchten Parameter zu gewinnen. Anschließend wird auf der Grundlage der beobachteten Daten und den Schätzwerten für die fehlenden Informationen eine Maximierung der Likelihood-Funktion bezüglich der gesuchten Parameter durchgeführt. Dies führt zu verbesserten Schätzwerten für die gesuchten Parameter. Da im Allgemeinen die Schätzung für die gesuchten Parameter zu diesem Zeitpunkt noch nicht ausreichend gut ist, wird diese Prozedur mehrfach wiederholt, wobei die verbesserten Parameterschätzwerte jeweils im nächsten Durchlauf zur Schätzung der fehlenden Information genutzt werden. Diese Vorgehensweise führt zu der in Abbildung 4.5 dargestellten Iteration.

Der E-Schritt

Die Schätzung der fehlenden Informationen erfolgt indirekt, indem der Erwartungswert der Log-Likelihood-Funktion[20] bezüglich der Verteilungsdichte der fehlenden Daten berechnet wird. Dieser Schritt wird deshalb als E-Schritt[21] bezeichnet. Dies wird im Folgenden vertieft.

[20] Die Log-Likelihood-Funktion hängt sowohl von den beobachteten Daten als auch von den fehlenden Informationen ab.

[21] Expectation (Erwartungswertbildung)

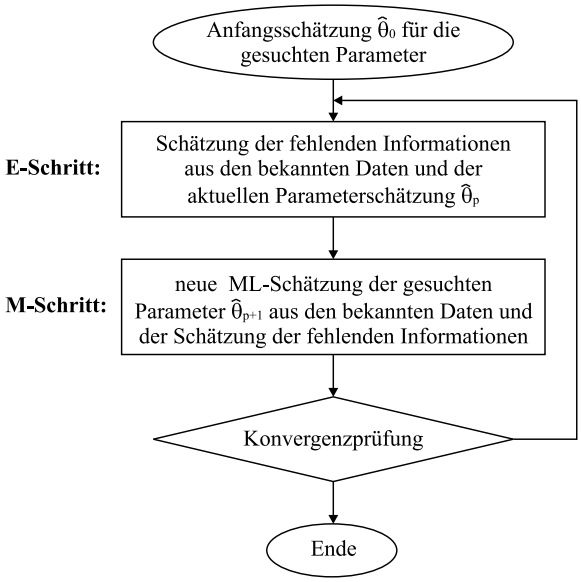

Abb. 4.5. Ablaufschema des EM-Algorithmus

Die unvollständigen[22] Daten werden mit $\mathbf{y}$ bezeichnet, die fehlenden Informationen bzw. fehlenden Daten mit $\mathbf{h}$. Unvollständige und fehlende Daten vereinigen sich zu den vollständigen Daten $\mathbf{z}$.

Die Log-Likelihood-Funktion der vollständigen Daten ist gegeben mit

$$\log L = \log f_Z(\mathbf{z}|\boldsymbol{\theta}) = \log f_Z(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta}). \tag{4.66}$$

Durch das Fehlen der Daten $\mathbf{h}$ ist die Maximum-Likelihood-Schätzung meist erheblich erschwert. Deshalb wird zunächst eine Hilfsfunktion Q eingeführt, mit der der Erwartungswert der Log-Likelihood-Funktion der vollständigen Daten bezüglich der Verteilungsdichte der fehlenden Daten berechnet wird

$$Q = E[\log f_Z(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta})]. \tag{4.67}$$

Die Verteilungsdichte der fehlenden Daten $f_H(\mathbf{h}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p)$ wird hierbei unter Nutzung der bereits aus der Iteration p vorhandenen Schätzwerte $\hat{\boldsymbol{\theta}}_p$ der gesuchten Parameter berechnet. Damit ergibt sich

$$Q = \int_{\mathbf{h}} [\log f_Z(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta})] f_H(\mathbf{h}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p) d\mathbf{h}. \tag{4.68}$$

[22] d.h. die beobachteten Daten

Der M-Schritt

Die Hilfsfunktion Q ist der Erwartungswert der Log-Likelihood-Funktion. Das heißt, sie ist die Log-Likelihood-Funktion, die man im Mittel für die verschiedenen Werte der fehlenden Daten $\mathbf{h}$ erwarten kann. Deshalb ist es sinnvoll, mit ihr die Maximum-Likelihood-Schätzung für die gesuchten Parameter $\boldsymbol{\theta}$ durchzuführen

$$\hat{\boldsymbol{\theta}}_{neu} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q. \tag{4.69}$$

Die neu berechnete ML-Schätzung der gesuchten Parameter kann in der nächsten Iteration $p + 1$ zur Berechnung der Verteilungsdichte $f_H(\mathbf{h}|\mathbf{y}, \hat{\boldsymbol{\theta}}_{p+1})$ genutzt werden.

4.5.3 Die mathematischen Grundlagen des EM-Algorithmus

Der EM-Algorithmus kann als Maximierung einer unteren Schranke der Likelihood-Funktion angesehen werden. Die untere Schranke wird so gewählt, dass sie im Punkt der aktuellen Parameterschätzung die Likelihood-Funktion berührt. Die Maximierung dieser unteren Schranke führt automatisch zu einem Punkt im Parameterraum, bei dem die Likelihood-Funktion einen

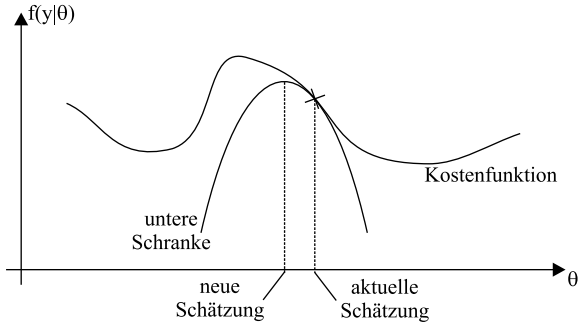

Abb. 4.6. Maximierung der Likelihood-Funktion mit einer unteren Schranke (nach [67])

größeren Wert einnimmt[23]. Dies veranschaulicht Abbildung 4.6. Der EM-Algorithmus entspricht der iterativen Durchführung dieses Zweischritt-Verfahrens aus Berechnung und anschließender Maximierung der unteren Schranke. Dies wird im Folgenden vertieft.

Die Likelihood-Funktion $f(\mathbf{y}|\boldsymbol{\theta}) = k(\boldsymbol{\theta})$ ist die zu maximierende Kostenfunktion. Sie ist gleichzeitig die Randdichte von $f_Z(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta})$

[23] vorausgesetzt, dass der Gradient an der aktuellen Stelle ungleich Null ist

$$k(\boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta}) = \int\limits_{\mathbf{h}} f_Z(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta}) d\mathbf{h}. \tag{4.70}$$

Nach der Erweiterung um die weiter unten definierte Verteilungsdichte $q(\mathbf{h})$ kann die Jensen-Ungleichung[24] auf die Kostenfunktion angewendet werden [67]

$$k(\boldsymbol{\theta}) = \int\limits_{\mathbf{h}} \frac{f_Z(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta})}{q(\mathbf{h})} q(\mathbf{h}) d\mathbf{h} \geq g(\boldsymbol{\theta}, q(\mathbf{h})) = \prod_{\mathbf{h}} \left(\frac{f_Z(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta})}{q(\mathbf{h})} \right)^{q(\mathbf{h}) d\mathbf{h}}, \tag{4.71}$$

mit

$$\int\limits_{\mathbf{h}} q(\mathbf{h}) d\mathbf{h} = 1. \tag{4.72}$$

Mit $g(\boldsymbol{\theta}, q(\mathbf{h}))$ ist man im Besitz einer unteren Schranke für die Kostenfunktion. Die Verteilungsdichte $q(\mathbf{h})$ muss nun so bestimmt werden, dass die untere Schranke die Kostenfunktion im Punkt der aktuellen Parameterschätzung $\hat{\boldsymbol{\theta}}_p$ berührt. Dazu wird die Funktion $g(\boldsymbol{\theta}, q(\mathbf{h}))$ zunächst logarithmiert

$$G(\boldsymbol{\theta}, q) = \log g(\boldsymbol{\theta}, q) = \int\limits_{\mathbf{h}} [q(\mathbf{h}) \log f_Z(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta}) - q(\mathbf{h}) \log q(\mathbf{h})] d\mathbf{h}. \tag{4.73}$$

Die Maximierung von G bezüglich $q(\mathbf{h})$ erfolgt unter zusätzlicher Berücksichtigung der Normierungsnebenbedingung für $q(\mathbf{h})$[25]

$$G(\boldsymbol{\theta}, q) = \lambda (1 - \int\limits_{\mathbf{h}} q(\mathbf{h}) d\mathbf{h}) + \int\limits_{\mathbf{h}} [q(\mathbf{h}) \log f_Z(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta}) - q(\mathbf{h}) \log q(\mathbf{h})] d\mathbf{h}. \tag{4.74}$$

Die Ableitung nach $q(\mathbf{h})$ für ein beliebiges, aber festes $\mathbf{h}$ [67][26] sowie an der Stelle $\hat{\boldsymbol{\theta}}_p$ ergibt schließlich die Optimalitätsbedingung

$$\frac{\partial G}{\partial q(\mathbf{h})} = -\lambda - 1 + \log f_Z(\mathbf{y}, \mathbf{h}|\hat{\boldsymbol{\theta}}_p) - \log q(\mathbf{h}) = 0. \tag{4.75}$$

Durch Umstellen der letzten Gleichung erhält man

$$\lambda + 1 = \log(\frac{f_Z(\mathbf{y}, \mathbf{h}|\hat{\boldsymbol{\theta}}_p)}{q(\mathbf{h})}) \tag{4.76}$$

bzw.

[24] Jensen-Ungleichung: $\sum_j g(j) a_j \geq \prod_j g^{a_j}(j)$ mit $\sum_j a_j = 1$, $a_j \geq 0$ und $g(j) \geq 0$ (Das arithmetische Mittel ist nie kleiner als das geometrische Mittel [67].)

[25] Zur Optimierung mit Gleichungsnebenbedingungen (Lagrangesche Multiplikatormethode): siehe Abschnitt 3.5.

[26] Aufgrund $\mathbf{h}$ beliebig, aber fest entfällt das Integral beim Differenzieren.

$$\exp(\lambda+1)q(\mathbf{h}) = f_Z(\mathbf{y},\mathbf{h}|\hat{\boldsymbol{\theta}}_p). \tag{4.77}$$

Die Integration beider Seiten bezüglich $\mathbf{h}$ führt zu

$$\exp(\lambda+1) = \int_{\mathbf{h}} f_Z(\mathbf{y},\mathbf{h}|\hat{\boldsymbol{\theta}}_p)d\mathbf{h}. \tag{4.78}$$

Substituiert man in dieser letzten Gleichung $\exp(\lambda+1) = f_Z(\mathbf{y},\mathbf{h}|\hat{\boldsymbol{\theta}}_p)/q(\mathbf{h})$, erhält man für $q(\mathbf{h})$

$$q(\mathbf{h}) = \frac{f_Z(\mathbf{y},\mathbf{h}|\hat{\boldsymbol{\theta}}_p)}{\int_{\mathbf{h}} f_Z(\mathbf{y},\mathbf{h}|\hat{\boldsymbol{\theta}}_p)d\mathbf{h}} = f_H(\mathbf{h}|\mathbf{y},\hat{\boldsymbol{\theta}}_p). \tag{4.79}$$

Für dieses $q(\mathbf{h})$ ergibt sich für den Wert der unteren Schranke $g(\boldsymbol{\theta}, q(\mathbf{h}))$ an der Stelle $\hat{\boldsymbol{\theta}}_p$

$$g(\hat{\boldsymbol{\theta}}_p, q) = \prod_{\mathbf{h}} \left(\frac{f_Z(\mathbf{y},\mathbf{h}|\hat{\boldsymbol{\theta}}_p)}{q(\mathbf{h})}\right)^{q(\mathbf{h})d\mathbf{h}} = \prod_{\mathbf{h}} \left(\frac{f_Z(\mathbf{y},\mathbf{h}|\hat{\boldsymbol{\theta}}_p)}{f_H(\mathbf{h}|\mathbf{y},\hat{\boldsymbol{\theta}}_p)}\right)^{q(\mathbf{h})d\mathbf{h}} \tag{4.80}$$

$$= \prod_{\mathbf{h}} \Big(f(\mathbf{y}|\hat{\boldsymbol{\theta}}_p)\Big)^{q(\mathbf{h})d\mathbf{h}} \tag{4.81}$$

$$= [f(\mathbf{y}|\hat{\boldsymbol{\theta}}_p)]^{\int_{\mathbf{h}} q(\mathbf{h})d\mathbf{h}} = f(\mathbf{y}|\hat{\boldsymbol{\theta}}_p). \tag{4.82}$$

Das Ergebnis[27] in Gleichung 4.82 bedeutet, dass die untere Schranke $g(\boldsymbol{\theta}, q(\mathbf{h}))$ die Kostenfunktion $f(\mathbf{y}|\boldsymbol{\theta})$ im Punkt $\hat{\boldsymbol{\theta}}_p$ berührt.

Um das Maximum der untere Schranke zu finden, muss in $G(\boldsymbol{\theta}, q)$ lediglich der Term

$$\int_{\mathbf{h}} q(\mathbf{h}) \log f_Z(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})d\mathbf{h} = \int_{\mathbf{h}} f_H(\mathbf{h}|\mathbf{y},\hat{\boldsymbol{\theta}}_p) \log f_Z(\mathbf{y},\mathbf{h}|\boldsymbol{\theta})d\mathbf{h} \tag{4.83}$$

bezüglich $\boldsymbol{\theta}$ maximiert werden. Dies aber entspricht exakt dem EM-Algorithmus in den Gleichungen 4.68 und 4.69.

4.5.4 Der EM-Algorithmus am Beispiel von MOG-Modellen

MOG-Modelle

Kompliziert strukturierte Verteilungsdichten werden häufig mit einer Summe gewichteter Gaußverteilungsdichten, sogenannten MOG[28]-Modelle approximiert. Diese Modelle sind einfach handhabbar und ermöglichen die statistische Erfassung sehr komplexer Prozesse.

[27] Beim Schritt von Gleichung 4.81 zu Gleichung 4.82 wurde die Beziehung $\prod_{i=1}^{N} a^{b_i} = a^{\sum_{i=1}^{N} b_i}$ genutzt.

[28] MOG: Mixture of Gaussians (Mischung gaußscher Verteilungsdichten)

MOG-Modelle besitzen die Form

$$f(\mathbf{y}) = \sum_{m=1}^{M} f(\mathbf{y}|m)P(m), \tag{4.84}$$

mit $P(m)$ als Gewichtungsfaktoren für die Gaußverteilungsdichten $f(\mathbf{y}|m)$. Die Gewichtungsfaktoren $P(m)$ erfüllen die Randbedingung

$$\sum_{m=1}^{M} P(m) = 1, \qquad 0 \leq P(m) \leq 1. \tag{4.85}$$

Sie können als Wahrscheinlichkeiten interpretiert werden, mit der ein Signalsample $\mathbf{y}_i$ von der Komponente m des MOG-Modells generiert wurde.

Die Verteilungsdichten $f(\mathbf{y}|m)$ sind im Allgemeinen multivariat und besitzen als Parameter jeweils den Mittelwertvektor $\boldsymbol{\mu}_m$ und die Kovarianzmatrix $\mathbf{R}_m$. Die Gaußverteilungsdichten bzw. Komponenten des MOG-Modells sind somit

$$f(\mathbf{y}|m) = \frac{1}{(2\pi)^{d/2}|\mathbf{R}_m|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_m)^T\mathbf{R}_m^{-1}(\mathbf{y}-\boldsymbol{\mu}_m)\right), \tag{4.86}$$

wobei d die Dimension von $\mathbf{y}$ angibt.

In der Nomenklatur des EM-Algorithmus repräsentieren die jeweils zum Zeitpunkt i beobachteten Signalsamples $\mathbf{y}_i$ die unvollständigen Daten, während die vollständigen Daten aus den Wertepaaren $\{\mathbf{y}_i, m_i\}$ bestehen. Die Komponenten-Label m_i können nicht direkt beobachtet werden.

Der EM-Algorithmus für MOG-Modelle

Die Log-Likelihood-Funktion der vollständigen Daten

Die Verbundverteilungsdichte aller beobachteten Daten bei gegebenem Parametervektor $\boldsymbol{\theta}$ und angenommener statistischer Unabhängigkeit zwischen den Signalsamples ist

$$\tilde{f}(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} f(\mathbf{y}_i|\boldsymbol{\theta}). \tag{4.87}$$

Die Kombination der Gleichungen 4.84 und 4.87 sowie die anschließende Logarithmierung ergeben die Log-Likelihood-Funktion der unvollständigen Daten

$$L_u = \log \tilde{f}(\mathbf{y}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log f(\mathbf{y}_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left\{ \sum_{m=1}^{M} f(\mathbf{y}_i|m, \boldsymbol{\theta}_m)P(m) \right\}. \tag{4.88}$$

Mit dem EM-Algorithmus sollen nun die Parameter der einzelnen Normalverteilungen $\boldsymbol{\theta}_m = \{\boldsymbol{\mu}_m, \mathbf{R}_m\}$ sowie die Wahrscheinlichkeiten $P(m)$ geschätzt werden.[29]

Aufgrund der Summe innerhalb des Logarithmus ist dieses Schätzproblem schwierig zu lösen. Wenn die fehlenden Daten, d.h. das Komponenten-Label m für jeden Zeitpunkt i, bekannt wären, würde sich Gleichung 4.88 erheblich vereinfachen, da so die Summe innerhalb des Logarithmus entfällt. Die Log-Likelihood-Funktion der dann vollständigen Daten wäre [11]

$$L_v = \sum_{i=1}^{N} \log\left[f(\mathbf{y}_i, m_i|\boldsymbol{\theta})\right] = \sum_{i=1}^{N} \log\left[f(\mathbf{y}_i|m_i, \boldsymbol{\theta})P(m_i)\right]. \tag{4.89}$$

Der E-Schritt

Im E-Schritt des EM-Algorithmus wird zunächst die diskrete Wahrscheinlichkeitsverteilung der fehlenden Daten $f(\mathbf{m}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p)$ für die aktuelle Iteration p des EM-Algorithmus geschätzt[30]

$$f(\mathbf{m}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p) = \prod_{i=1}^{N} f(m_i|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p), \tag{4.90}$$

wobei die diskreten Wahrscheinlichkeitsverteilungen $f(m_i|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p)$ unter Verwendung des Bayes'schen Satzes sowie der Schätzwerte $\hat{P}_p(m_i)$ und $\hat{\boldsymbol{\theta}}_p$ mit

$$f(m_i|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) = \frac{f(\mathbf{y}_i|m_i, \hat{\boldsymbol{\theta}}_p)\hat{P}_p(m_i)}{f(\mathbf{y}_i|\hat{\boldsymbol{\theta}}_p)} = \frac{f(\mathbf{y}_i|m_i, \hat{\boldsymbol{\theta}}_p)\hat{P}_p(m_i)}{\sum_{m=1}^{M} f_m(\mathbf{y}_i|m_i, \hat{\boldsymbol{\theta}}_p)\hat{P}_p(m_i)}. \tag{4.91}$$

ermittelt werden. Der Erwartungswert folgt mit

$$Q = \sum_{\mathbf{m}} L_v f(\mathbf{m}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p) \tag{4.92}$$

$$= \sum_{\mathbf{m}} \left[\sum_{i=1}^{N} \log f(\mathbf{y}_i|m_i, \boldsymbol{\theta})P(m_i)\right] \left[\prod_{k=1}^{N} f(m_k|\mathbf{y}_k, \hat{\boldsymbol{\theta}}_p)\right] \tag{4.93}$$

$$= \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} \cdots$$

$$\cdots \sum_{m_N=1}^{M} \sum_{i=1}^{N} \log\left[f(\mathbf{y}_i|m_i, \boldsymbol{\theta})P(m_i)\right] \left[\prod_{k=1}^{N} f(m_k|\mathbf{y}_k, \hat{\boldsymbol{\theta}}_p)\right]. \tag{4.94}$$

[29] Aus Übersichtlichkeitsgründen wird der Index von $\boldsymbol{\theta}_m$ im Folgenden nicht weiter mitgeschrieben. $\boldsymbol{\theta}$ symbolisiert den im jeweiligen Kontext zu bestimmenden Parametervektor.

[30] Zu beachten ist, dass $f(\mathbf{m}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p)$ eine multivariate diskrete Wahrscheinlichkeitsverteilung der vektoriellen Zufallsvariable $\mathbf{m} = [m_1, m_2, \ldots, m_N]$ ist. Dies muss unter anderem bei der Bildung des Erwartungswertes berücksichtigt werden, in dem über jedes einzelne Element von $\mathbf{m}$ summiert wird.

Die Struktur dieses Ausdrucks soll anhand eines Beispiels veranschaulicht werden.

Beispiel 4.3. Gleichung 4.94 wird für den Fall $N = 3$ untersucht.

Die diskrete Wahrscheinlichkeitsverteilung der fehlenden Daten ist

$$f(\mathbf{m}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p) = f(m_1|\mathbf{y}_1, \hat{\boldsymbol{\theta}}_p) \cdot f(m_2|\mathbf{y}_2, \hat{\boldsymbol{\theta}}_p) \cdot f(m_3|\mathbf{y}_3, \hat{\boldsymbol{\theta}}_p). \tag{4.95}$$

Als Log-Likelihood-Funktion der vollständigen Daten bekommt man

$$\begin{aligned} L_v = &\log\left[f(\mathbf{y}_1|m_1, \boldsymbol{\theta})P(m_1)\right] + \log\left[f(\mathbf{y}_2|m_2, \boldsymbol{\theta})P(m_2)\right] \\ &+ \log\left[f(\mathbf{y}_3|m_3, \boldsymbol{\theta})P(m_3)\right]. \end{aligned} \tag{4.96}$$

Zur Berechnung des Erwartungswertes Q werden nun $f(\mathbf{m}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p)$ und L_v in Gleichung 4.92 eingesetzt

$$Q = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} \sum_{m_3=1}^{M} L_v f(\mathbf{m}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p) \tag{4.97}$$

$$\begin{aligned} = &\sum_{m_1=1}^{M} \sum_{m_2=1}^{M} \sum_{m_3=1}^{M} \log\left[f(\mathbf{y}_1|m_1, \boldsymbol{\theta})P(m_1)\right] f(\mathbf{m}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p) \\ &+ \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} \sum_{m_3=1}^{M} \log\left[f(\mathbf{y}_2|m_2, \boldsymbol{\theta})P(m_2)\right] f(\mathbf{m}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p) \\ &+ \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} \sum_{m_3=1}^{M} \log\left[f(\mathbf{y}_3|m_3, \boldsymbol{\theta})P(m_3)\right] f(\mathbf{m}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p). \end{aligned} \tag{4.98}$$

Bei genauer Betrachtung der drei Terme in Gleichung 4.98 können Vereinfachungen vorgenommen werden. Dies soll zunächst am Beispiel des ersten Terms gezeigt werden. Es gilt

$$\begin{aligned} &\sum_{m_1=1}^{M} \sum_{m_2=1}^{M} \sum_{m_3=1}^{M} \log\left[f(\mathbf{y}_1|m_1, \boldsymbol{\theta})P(m_1)\right] f(\mathbf{m}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p) \\ = &\sum_{m_1=1}^{M} \sum_{m_2=1}^{M} \sum_{m_3=1}^{M} \log\left[f(\mathbf{y}_1|m_1, \boldsymbol{\theta})P(m_1)\right] f(m_1|\mathbf{y}_1, \hat{\boldsymbol{\theta}}_p) f(m_2|\mathbf{y}_2, \hat{\boldsymbol{\theta}}_p) \\ &\times f(m_3|\mathbf{y}_3, \hat{\boldsymbol{\theta}}_p) \end{aligned} \tag{4.99}$$

$$\begin{aligned} = &\sum_{m_1=1}^{M} \log\left[f(\mathbf{y}_1|m_1, \boldsymbol{\theta})P(m_1)\right] f(m_1|\mathbf{y}_1, \hat{\boldsymbol{\theta}}_p) \\ &\times \sum_{m_2=1}^{M} f(m_2|\mathbf{y}_2, \hat{\boldsymbol{\theta}}_p) \sum_{m_3=1}^{M} f(m_3|\mathbf{y}_3, \hat{\boldsymbol{\theta}}_p). \end{aligned} \tag{4.100}$$

Da jedoch auch

$$\sum_{m_2=1}^{M} f(m_2|\mathbf{y}_2, \hat{\boldsymbol{\theta}}_p) = \sum_{m_3=1}^{M} f(m_3|\mathbf{y}_3, \hat{\boldsymbol{\theta}}_p) = 1 \tag{4.101}$$

gilt, erhält man für den ersten Term von Gleichung 4.98

$$\begin{aligned} &\sum_{m_1=1}^{M} \sum_{m_2=1}^{M} \sum_{m_3=1}^{M} \log\left[f(\mathbf{y}_1|m_1, \boldsymbol{\theta})P(m_1)\right] f(\mathbf{m}|\mathbf{y}, \hat{\boldsymbol{\theta}}_p) \\ &= \sum_{m_1=1}^{M} \log\left[f(\mathbf{y}_1|m_1, \boldsymbol{\theta})P(m_1)\right] f(m_1|\mathbf{y}_1, \hat{\boldsymbol{\theta}}_p). \end{aligned} \tag{4.102}$$

Überträgt man dieses Ergebnis auf die anderen Terme in Gleichung 4.98, ergibt sich für den Erwartungswert

$$\begin{aligned} Q = &\sum_{m_1=1}^{M} \log\left[f(\mathbf{y}_1|m_1, \boldsymbol{\theta})P(m_1)\right] f(m_1|\mathbf{y}_1, \hat{\boldsymbol{\theta}}_p) \\ &+ \sum_{m_2=1}^{M} \log\left[f(\mathbf{y}_2|m_2, \boldsymbol{\theta})P(m_2)\right] f(m_2|\mathbf{y}_2, \hat{\boldsymbol{\theta}}_p) \\ &+ \sum_{m_3=1}^{M} \log\left[f(\mathbf{y}_3|m_3, \boldsymbol{\theta})P(m_3)\right] f(m_3|\mathbf{y}_3, \hat{\boldsymbol{\theta}}_p). \end{aligned} \tag{4.103}$$

Durch die Umbenennung $m = m_i$ bekommt man schließlich

$$Q = \sum_{i=1}^{3} \sum_{m=1}^{M} \log\left[f(\mathbf{y}_i|m, \boldsymbol{\theta})P(m)\right] f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p). \tag{4.104}$$

□

Das Ergebnis aus Gleichung 4.104 kann auf einen beliebigen Wert N verallgemeinert werden und man erhält für den E-Schritt

$$Q = \sum_{i=1}^{N} \sum_{m=1}^{M} \log(f(\mathbf{y}_i|m, \boldsymbol{\theta})P(m)) f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) \tag{4.105}$$

$$\begin{aligned} &= \sum_{i=1}^{N} \sum_{m=1}^{M} \log[f(\mathbf{y}_i|m, \boldsymbol{\theta})] f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) \\ &\quad + \sum_{i=1}^{N} \sum_{m=1}^{M} \log[P(m)] f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p). \end{aligned} \tag{4.106}$$

Die Maximierung von Gleichung 4.106 kann nach $P(m)$ und $\boldsymbol{\theta}$ getrennt erfolgen, da beide unabhängig voneinander in verschiedenen Termen von Gleichung 4.106 existieren.

Der M-Schritt

Zunächst wird Gleichung 4.106 bezüglich $P(m)$ maximiert. Dabei ist die Gleichungsnebenbedingung $\sum P(m) = 1$ zu berücksichtigen. Zusammen mit der Nebenbedingung erhält man für die Ableitung bezüglich $P(m)$

$$\frac{\partial}{\partial P(m)} \left[\sum_{m=1}^{M} \sum_{i=1}^{N} \log[P(m)] f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) + \lambda \left((\sum_{m=1}^{M} P(m)) - 1 \right) \right] \stackrel{!}{=} 0. \tag{4.107}$$

Daraus folgen

$$\sum_{i=1}^{N} \frac{1}{P(m)} f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) + \lambda = 0 \tag{4.108}$$

und weiter

$$-\lambda P(m) = \sum_{i=1}^{N} f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p). \tag{4.109}$$

Wird die letzte Gleichung über m summiert, ergibt sich wegen $\sum_m P(m) = 1$ für λ der Wert $\lambda = -N$ und somit[31]

$$P(m) = \frac{1}{N} \sum_{i=1}^{N} f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) \stackrel{!}{=} \hat{P}_{p+1}(m). \tag{4.110}$$

Bei der Maximierung des verbleibenden Terms von Gleichung 4.106 gilt zunächst für die gaußschen Verteilungsdichten

$$\begin{aligned} &\sum_{m=1}^{M} \sum_{i=1}^{N} \log[f(\mathbf{y}_i|m, \boldsymbol{\theta})] f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) \\ &= \sum_{m=1}^{M} \sum_{i=1}^{N} \left(-\frac{d}{2} \log(2\pi) - \frac{1}{2} \log|\mathbf{R}_m| - \frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_m)^T \mathbf{R}_m^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_m) \right) \\ &\quad \times f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p). \end{aligned} \tag{4.111}$$

Das Nullsetzen der Ableitung nach $\boldsymbol{\mu}_m$ führt auf

$$\sum_{i=1}^{N} \mathbf{R}_m^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_m) f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) = \mathbf{0} \tag{4.112}$$

und

$$\boldsymbol{\mu}_m = \frac{\sum_{i=1}^{N} \mathbf{y}_i f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p)}{\sum_{i=1}^{N} f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p)} \stackrel{!}{=} \hat{\boldsymbol{\mu}}_{m,p+1}. \tag{4.113}$$

[31] Die diskreten Wahrscheinlichkeitsverteilungen $f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p)$ ergeben sich aus Gleichung 4.91, wobei auch dort die Substitution $m = m_i$ vorgenommen werden muss.

Die Bestimmung der Kovarianzmatrix $\mathbf{R}_m$ erfolgt aus Gleichung 4.111 unter Nutzung von $\sum_i \mathbf{y}_i^T \mathbf{R}^{-1} \mathbf{y}_i = tr(\mathbf{R}^{-1} \sum_i \mathbf{y}_i \mathbf{y}_i^T)$

$$\begin{aligned}
&\sum_{m=1}^{M} \sum_{i=1}^{N} \log[f(\mathbf{y}_i|m, \boldsymbol{\theta})] f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) \\
&= \sum_{m=1}^{M} \sum_{i=1}^{N} \left[-\frac{d}{2} \log(2\pi) + \frac{1}{2} \log(|\mathbf{R}_m^{-1}|) \right] f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) \\
&\quad -\frac{1}{2} \sum_{m=1}^{M} \sum_{i=1}^{N} (\mathbf{y}_i - \boldsymbol{\mu}_m)^T \mathbf{R}_m^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_m) f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) \qquad (4.114) \\
&= \sum_{m=1}^{M} \sum_{i=1}^{N} \left[-\frac{d}{2} \log(2\pi) + \frac{1}{2} \log(|\mathbf{R}_m^{-1}|) \right] f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) \\
&\quad -\frac{1}{2} \sum_{m=1}^{M} tr \left[\mathbf{R}_m^{-1} \sum_{i=1}^{N} (\mathbf{y}_i - \boldsymbol{\mu}_m)(\mathbf{y}_i - \boldsymbol{\mu}_m)^T f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) \right]. \qquad (4.115)
\end{aligned}$$

Die Ableitung von Gleichung 4.115 nach $\mathbf{R}_m^{-1}$ erfolgt unter Berücksichtigung der Beziehungen

$$\frac{\partial \log |det(\mathbf{A})|}{\partial \mathbf{A}} = \mathbf{A}^{-T} \qquad \text{und} \qquad \frac{\partial tr(\mathbf{AB})}{\partial \mathbf{A}} = \mathbf{B}^T \qquad (4.116)$$

sowie mit Hilfe der Symmetrie-Eigenschaft der Kovarianzmatrix $\mathbf{R}_m = \mathbf{R}_m^T$. Damit bekommt man

$$\mathbf{R}_m \sum_{i=1}^{N} f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) - \sum_{i=1}^{N} (\mathbf{y}_i - \boldsymbol{\mu}_m)(\mathbf{y}_i - \boldsymbol{\mu}_m)^T f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) = \mathbf{0} \qquad (4.117)$$

und schließlich

$$\mathbf{R}_m = \frac{\sum_{i=1}^{N} f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p)(\mathbf{y}_i - \boldsymbol{\mu}_m)(\mathbf{y}_i - \boldsymbol{\mu}_m)^T}{\sum_{i=1}^{N} f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p)} \stackrel{!}{=} \hat{\mathbf{R}}_{m,p+1}. \qquad (4.118)$$

Die Adaptionsgleichungen in jeder Iteration sind somit

$$\hat{P}_{p+1}(m) = \frac{1}{N} \sum_{i=1}^{N} f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p) \qquad (4.119)$$

$$\hat{\boldsymbol{\mu}}_{m,p+1} = \frac{\sum_{i=1}^{N} \mathbf{y}_i f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p)}{\sum_{i=1}^{N} f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p)} \qquad (4.120)$$

$$\hat{\mathbf{R}}_{m,p+1} = \frac{\sum_{i=1}^{N} f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p)(\mathbf{y}_i - \boldsymbol{\mu}_{m,p+1})(\mathbf{y}_i - \boldsymbol{\mu}_{m,p+1})^T}{\sum_{i=1}^{N} f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p)}. \qquad (4.121)$$

Beispiel 4.4. Ein Beispiel-Datensatz wurde mit folgenden Parametern generiert:

$$P(1) = 0.7, \mu_1 = -1, \sigma_1^2 = 0.1; P(2) = 0.3, \mu_2 = 1, \sigma_2^2 = 0.5.$$

Die Anzahl der generierten Daten beträgt $N = 10000$. Die Iteration erfolgt mit den Gleichungen 4.119 bis 4.121. Die diskreten Wahrscheinlichkeitsverteilungen $f(m|\mathbf{y}_i, \hat{\boldsymbol{\theta}}_p)$ werden mit Gleichung 4.91 berechnet. Tabelle 4.1 und die folgenden Abbildungen zeigen den Verlauf der Iteration.

Iteration	$P(1)$	μ_1	σ_1^2	$P(2)$	μ_2	σ_2^2
Startwert	0.1	0.0	0.1	0.9	10.0	1.0
1	0.990996565	-0.430898603	0.986740580	0.009003435	2.668559073	0.076061023
10	0.883977031	-0.670766489	0.556306624	0.116022969	1.637170064	0.225741481
20	0.709794409	-0.993638696	0.103459381	0.290205591	1.041628295	0.474012113
30	0.706746846	-0.996091761	0.102033667	0.293253154	1.026389209	0.492333730
40	0.706686218	-0.996138296	0.102007413	0.293313782	1.026083285	0.492708181
50	0.706684955	-0.996139265	0.102006866	0.293315045	1.026076907	0.492715989
60	0.706684928	-0.996139286	0.102006855	0.293315072	1.026076774	0.492716152

Tabelle 4.1. Konvergenz des Verfahrens

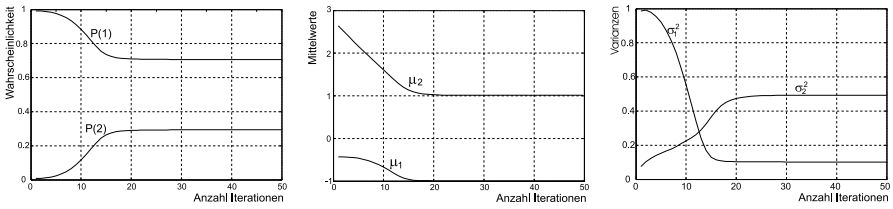

Abb. 4.7. Konvergenz der Komponentenwahrscheinlichkeiten (links), der Mittelwerte (mitte) und der Varianzen (rechts)

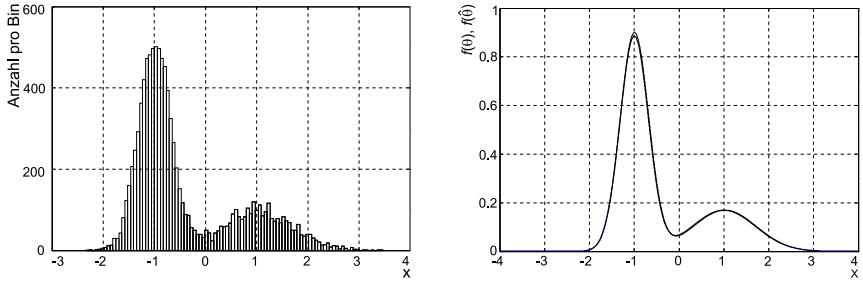

Abb. 4.8. Histogramm der beobachteten Daten (links) sowie die theoretische und die berechnete Verteilungsdichte (beide sind nahezu überlappend) der beobachteten Daten (rechts).

Die Konvergenz des Verfahrens ist in diesem Falle unproblematisch. Dies liegt vor allem an der Gültigkeit des Modells, der geringen Parameterzahl, der einfachen Struktur der zu approximierenden Verteilungsdichte sowie an der hohen Anzahl von Datensamples. □