Bayes-Schätzung und Maximum-Likelihood-Schätzung

Notiz

Wolfgang Förstner
Photogrammetry & Robotics
Bonn University

29. 01. 2021

Inhaltsverzeichnis

1 Ziel

Das Ziel dieser Notiz ist das Prinzip der Bayes-Schätzung und der Maximum-Likelihood-Schätzung zu erläutern. Im einzelnen formalisieren wir eine Reihe von unsicheren Aussagen mit Hilfe der Wahrscheinlichkeitstheorie:

- Man kann *Vorinformation* oder seine *Überzeugung* (belief) über unbekannte Parameter, die man vor der Durchführung von Beobachtungen hat, als *a priori* Wahrscheinlichkeitsdichte modellieren.
- Man kann den Messprozess und seine Unsicherheit als bedingte Wahrscheinlichkeitsdichte modellieren. Sie kann gleichzeitig dazu verwendet werden um die Sicherheit (*likelihood*) darzustellen, wie gegebene Beobachtungen durch die Parameter erklärt werden.

- Man kann die Vorinformation und die Kenntnis über den Messprozess zusammenfassen und in der bedingten Wahrscheinlichkeitsdichte, der *a posteriori* Wahrscheinlichkeitsdichte, darstellen und so eine *verbesserten Überzeugung* über die Parameter nach der Durchführung von Beobachtungen zu modellieren.

- Aus dem Modell des Beobachtungsprozesses und den vorliegenden Beobachtungen kann man eine optimalen Schätzung für die Parameter ableiten, die sog. *Maximum-Likelihood-Schätzung* oder ML-Schätzung.

- Bei Vorliegen von Vorinformation, kann man aus der Verknüpfung dieser Vorinformation, dem Beobachtungsmodell und den Beobachtungen, zu einer optimalen Schätzung für die Parameter kommen, die sog. *Maximum-a-posteriori-Schätzung*, oder MAP-Schätzung. Sie wird, da das sog. Bayes-Theorem verwendet wird, auch als *Bayes-Schätzung* bezeichnet.

- Der Bayes-Schätzer kann als Erklärung für ein von Levenberg und Marquardt vorgeschlagenes Schätzverfahren dienen, das auch im Fall unvollständiger Beobachtung des Parametervektors eine Lösung ermöglicht.

Wir verdeutlichen die Zusammenhänge an der Bestimmung einer Strecke und an dem klassischen Gauß-Markov-Modell.

2 Die Modellierung

Wir beziehen uns auf folgende Beispiele.

1. Es geht um die Seitenlänge x einer Schachtel, deren Länge wir durch Messung bestimmen wollen. Dazu nehmen wir an, dass wir Vorwissen über die Länge x haben und wir dieses Vorwissen zusammen mit dem Wissen über den Messprozess und die tatsächlich durchgeführte Messung verschmelzen wollen. Anders ausgedrückt: Unsere Überzeugung über x vor der Messung, wollen wir durch Beobachtungen korrigieren. Das Ziel des Modellierung ist diese Situation statistisch zu beschreiben und dann für die Schätzung zu nutzen.

2. Wir verallgemeinern die Situation für den Fall des linearen Gauß-Markoff Modells.

2.1 Das Vorwissen

Als Vorwissen nehmen wir folgendes an: Wir wissen, dass die Seitenlänge 1.2 dm oder 1.6 dm ist, allerdings mit einer Ungenauigkeit von jeweils 0.1 dm. Dieses Wissen können wir so formalisieren: Mit einer Wahrscheinlichkeit von $P_1 = 1/2$ ist die Länge normal verteilt mit Mittelwert $\mu_1 = 1.2$ dm und Streuung $\sigma_1 = 0.1$ cm und mit der gleichen Wahrscheinlichkeit normal verteilt mit Mittelwert $\mu_2 = 1.6$ cm und $\sigma_2 = 0.1$ dm. Dieses Vorwissen können wir in einer Dichtefunktion darstellen wie in der folgenden Figur. Die Fläche unter der Dichte ist natürlich 1, sodass jede der beiden Dichten um 1.2 dm bzw. um 1.6 dm zur Hälfte beitragen. Mit der Dichtefunktion

$$g(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2\right) \tag{1}$$

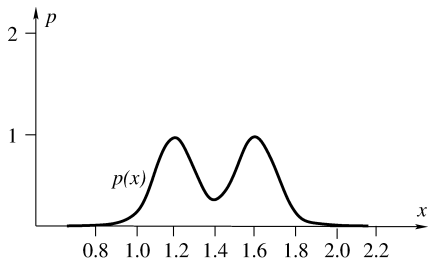

Abbildung 1: Vorwissen über x repräsentiert als Wahrscheinlichkeitsdichte

der Gauß-Verteilung würde das formal lauten

$$\underline{x} \sim p(x) = \frac{1}{2} g(x \mid 1.2, 0.1^2) + \frac{1}{2} g(x \mid 1.6, 0.1^2)\,. \tag{2}$$

Die Gleichung besagt folgendes: die Zufallsvariable $\underline{x}$ repräsentiert unser unsicheres Vorwissen über die Länge. Dieses Vorwissen der Seitenlänge folgt der Verteilung, die durch die Dichtefunktion $p(x)$ charakterisiert ist.

In der Statistik sind zwei Sprechweisen hierfür üblich:

1. Die *prior Verteilung* von $\underline{x}$ wird durch die Dichte $p(x)$ charakterisiert.
2. Die *Überzeugung*, die wir von $\underline{x}$ vor der Messung haben wird durch die Dichte $p(x)$ charakterisiert.

Die zweite Formulierung hat den Vorteil, dass wir uns vorstellen können, dass sich unsere Überzeugung durch weitere zukünftige Messungen schrittweise verbessern lässt.

2.2 Der Messprozess

Nun modellieren wir den Messprozess. Er beschreibt den Zusammenhang zwischen den Beobachtungen und den Parametern. Dieses Modell gibt uns zwei Auskünfte:

1. Wie groß sind die Messabweichungen? Das ist gleichbedeutend mit der Frage: Wie beschreiben wir die Unsicherheit der Messwerte, falls die unbekannten Parameter bekannt sind?
2. Wie gut erklären die vorliegenden Beobachtungen die unbekannten Parameter? Das ist gleichbedeutend mit der Frage: Wie groß ist die Sicherheit, die wir über die Parameter haben, falls die Beobachtungen vorliegen.

2.2.1 Die Messung

Im Beispiel habe das Messgerät eine Ungenauigkeit von 0.15 dm. Die Abweichungen der Messung l von der Länge x sind daher im Mittel 0.15 dm ist. Wenn wir 1.5 dm messen, ist damit die Unsicherheit der Messung so wie in der folgenden Figur dargestellt. Diese Normalverteilung hat ihren Mittelwert bei 1.5 dm. Das bedeutet, wir nehmen den Messwert 1.5 dm, der ja eine fixe Größe ist, als besten Wert für die Seitenlänge x und argumentieren, wenn wir die Messung wiederholen würden und die tatsächliche Seitenlänge 1.5 dm ist. Formal stellen wir dies als eine bedingte

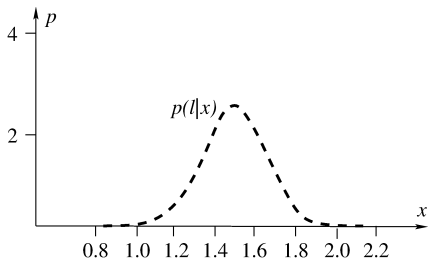

Abbildung 2: Unsicherheit der Messung $l = 1.5$, Streuung 0.015 dm

Dichtefunktion $p(l \mid x)$ dar, die sagt, wie die Zufallsvariable $\underline{l}$ verteilt ist,[1] falls man x kennt:

$$\underline{l} \mid x \sim p(l \mid x) = g(l \mid x, \sigma^2) = g(l \mid 1.5, 0.15^2)\,. \tag{3}$$

Der Senkrechtstrich $\mid$ zeigt an, dass alles was rechts von $\mid$ steht als bekannt, also fest angesehen wird, unabhängig davon, ob es eine Variable x ist, also die Aussage "Die Zufallsvariable $\underline{x}$ hat den Wert x", d.h. $\underline{x} = x$ ist oder ob es eine Realisierung $x = 1.5$ ist. Die Dichtefunktion $g(x, 1.5, 0.15^2)$, die in Fig. 2 darstellt ist, zeigt wie diese Messungen aussehen würden.

2.2.2 Erklärung der Parameter durch die Beobachtungen – likelihood

Man kann die bedingte Dichte $p(l \mid x)$ auch anders lesen. Falls wir den Messwert l wissen, gibt uns $p(l \mid x)$ in gewisser Weise an, wie sicher wir die Messung durch x erklären können.

Im Deutschen gibt es leider nur einen Namen für "wahrscheinlich" bzw 'Wahrscheinlichkeit". Wenn wir im Deutschen dagegen das Wort "Sicherheit" verwenden, nehmen wir implizit an, dass die Sicherheit (subjektiv die Wahrscheinlichkeit) groß ist, sonst würden wir das Wort "Unsicherheit" verwenden.

Im Englischen gibt es mehrere Worte für "wahrscheinlich" bzw 'Wahrscheinlichkeit", z.B auch "likely" oder "likelihood". Das Wort "likely" wird innerhalb der Statistik verwendet, wenn die Sicherheit für das Auftreten einer bestimmten Größe gemeint ist, man aber das Wort "Wahrscheinlichkeit" vermeiden will, da der Sicherheitswert i.a. keine Wahrscheinlichkeit (oder -sdichte) im strengen Sinne ist.

Die Dichte $p(l \mid x)$ wird daher, bei festem l auch als "Likelihood" oder 'Likelihood Funktion" von x verwendet:

$$L(x) = p(l \mid x)\,, \tag{4}$$

wobei man implizit annimmt, dass $L(x)$ sich auf den Messwert l bezieht. Man beachte, die Funktion $L(x)$ ist keine Dichte, da das Integral über x i.a. nicht 1 ist.

2.3 Die Verschmelzung von Vorwissen und Messung

Nun führen wir diese beiden Informationen zusammen und interpretieren das Ergebnis auf zwei Weisen.

[1] Zufallsvariable sind unterstrichen.

2.3.1 Die Bayesformel

Wir nehmen an, dass die Vorinformation keinen Einfluss auf die Messung hat. Damit haben wir zwei unabhängige Aussagen über x durch ihre Wahrscheinlichkeitsdichten, direkt durch $p(x)$ und indirekt durch $p(l \mid x)$ spezifiziert.

Die Aussage "Ich weiß dass x gemäß $p(x)$ verteilt ist *und* (gleichzeitig), dass die Unsicherheit meiner Messung durch $p(l \mid x)$ charakterisiert wird" hat demnach eine Wahrscheinlichkeitsdichte, die sich aus dem Produkt der beiden Wahrscheinlichkeitsdichten[2] ergibt:

$$p(x)\, p(l \mid x)\,. \tag{5}$$

Dies ist aber, nach den Axiomen der Wahrscheinlichkeitstheorie die Verbund-Wahrscheinlichkeitsdichte von l und x

$$p(x, l) = p(x)\, p(l \mid x)\,. \tag{6}$$

Diese Dichte ist nicht unmittelbar interpretierbar. Eigentlich möchten wir den wahrscheinlichsten Wert für x wissen, falls uns die Beobachtung l vorliegt, bzw. wir wollen die Überzeugung über x nach der Durchführung der Messung charakterisieren. Diese Überzeugung können wir mit der bedingten Dichte $p(x \mid l)$ charakterisieren, deren Maximum eine plausible Schätzung für x ergibt.

Um dies zu erreichen, wenden wir die Relation $p(x, l) = p(x)\, p(l \mid x)$ auch in der Form an, bei der l und x vertauscht sind und erhalten zunächst

$$p(x, l) = p(x)\, p(l \mid x) = p(l)\, p(x \mid l)\,. \tag{7}$$

Die Auflösung nach $p(x \mid l)$ liefert direkt

$$\boxed{p(x \mid l) = \frac{p(x)\, p(l \mid x)}{p(l)}\,.} \tag{8}$$

Dies ist die *posteriori Dichte* von $\underline{x}$ für eine gegebene Beobachtung l. Die Funktion ist für das Beispiel in der folgenden Figur zusammen mit $p(x)$ und $p(l \mid x)$ dargestellt

Man sieht dass die Unsicherheit über die Seitenlänge x kleiner geworden ist. Statt zweier gleich hoher Maxima, sind es jetzt ein höheres Maximum und ein kleineres lokales Maximum. Es erscheint daher nicht ganz sicher als besten Wert für x das absolute Optimum in der Nähe von $x = 1.57$ zu wählen. Dieses Maximum liegt zwischen dem Messwert und dem rechten Maximum der prior Verteilung.

Die Gl. (8) wird als *Bayesformel* oder *Bayes-Theorem* bezeichnet und ist zentrales Element der statistischen Mustererkennung. Hier verwenden wir sie zur Verknüpfung von Vorwissen und Messungen. Sie gilt natürlich auch für Vektoren $\boldsymbol{l}$ und $\boldsymbol{x}$:

$$\boxed{p(\boldsymbol{x} \mid \boldsymbol{l}) = \frac{p(\boldsymbol{x})\, p(\boldsymbol{l} \mid \boldsymbol{x})}{p(\boldsymbol{l})}\,.} \tag{9}$$

Sie enthält – in unserem Kontext der Schätzung von Parametern:

[2] Falls A und B zwei sich unabhängige Ereignisse sind, die mit Wahrscheinlichkeiten $P(A)$ und $P(B)$ auftreten, dann tritt das Ereignis $A\&B$, dass beide Ereignisse gleichzeitig auftreten, mit der Wahrscheinlichkeit $P(A\&B) = P(A)P(B)$ auf, s. Schuh (2020, Satz 3.6, S. 82)

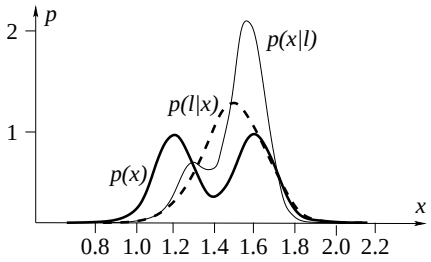

Abbildung 3: Die prior Dichte $p(x)$ (dick durchgezogen) zeigt die Unsicherheit unserer Überzeugung, nämlich dass x ungefähr 1.2 oder 1.6 ist. Die Likelihood-Funktion $L(x) = p(l \mid x)$ (gestrichelt) folgt aus der unsicheren Messung von 1.5. Die posteriori Dichte $p(x \mid l)$ (dünn durchgezogen), charakterisiert die Unsicherheit unserer Überzeugung von x nach der Messung. Die beste Schätzung, die Bayes-Schätzung oder Maximum a posteiori Schätzung, die unserer Vorinformation und die Messung berücksichtigt liegt beim Maximum von $p(x \mid l)$ also etwa bei $\widehat{x}_{\text{Bayes}} \approx 1.57$, näher bei dem rechten lokalen Maximum der Vorinformation, da die Messung ungenauer als die Vorinformation ist. Falls wir die Vorinformation nicht verwenden, sondern nur unsere Kenntnis über das Messverfahren, also die Likelihood-Funktion, dann ist die beste Schätzung das Maximum von $p(l \mid x)$ also $\widehat{x}_{\text{ML}} = 1.5$

1. Die *a priori Dichte* $p(\boldsymbol{x})$ für die Parameter, die wir bestimmen wollen. Diese Vorinformation kann aus Abschätzungen, dem Abgreifen aus Karten, oder durch früher durchgeführte Messungen entstehen. Sie spezifiziert die Überzeugung, die man über den Wert von $\boldsymbol{x}$ hat, bevor man die Messung von $\boldsymbol{l}$ durchführt.

2. Die bedingte Dichte $p(\boldsymbol{l} \mid \boldsymbol{x})$ charakterisiert die Unsicherheit des Beobachtungsprozesses. Sie ist die (bedingte) Dichte der Beobachtungen, falls die Parameter $\boldsymbol{x}$ bekannt sind, modellieren also im Wesentlichen die Abhängigkeit von den Parametern und die zufälligen Abweichungen beim Beobachten. Falls man sie als Likelihood-Funktion $L(\boldsymbol{x}) := p(\boldsymbol{l} \mid \boldsymbol{x})$ interpretiert, sagt sie wie gut die mit dem modellierten Messgerät durchgeführten Beobachtungen durch die Parameter erklärt werden können.

3. Die bedingte Dichte $p(\boldsymbol{x} \mid \boldsymbol{l})$ ist die a posteriori Dichte. Sie charakterisiert die Unsicherheit der Parameter $\boldsymbol{x}$, falls die Beobachtungen $\boldsymbol{l}$ vorliegen. Gleichbedeutend, spezifiziert sie die durch die Beobachtungen korrigierte Überzeugung, die man über den Wert $\boldsymbol{x}$ hat.

4. Die Dichte $p(\boldsymbol{l})$ ist bei vorliegenden Beobachtungen eine Konstante. Ansonsten dient der Wert dazu, dass der Ausdruck auf der rechten Seite eine Dichtefunktion ist, also das Integral über die Variable $\boldsymbol{x}$ Eins ist.

2.3.2 Belief propagation – Korrektur der Überzeugung

Die Überzeugung, die man über $\boldsymbol{x}$ hat, kommt in der Bayesformel zweimal vor. Wir wollen sie mit $\text{bel}_t(\boldsymbol{x})$ – für engl. *belief* – kennzeichnen, wobei der Index angibt, auf welche Situation, z.B. welchen Zeitpunkt, sich die Überzeugung bezieht:

1. Die Überzeugung über $\boldsymbol{x}$ vor der Messung ist dann z.B.

$$\text{bel}_{\text{a priori}}(\boldsymbol{x}) := p(\boldsymbol{x})\,. \tag{10}$$

2. Die Überzeugung über $\boldsymbol{x}$ nach der Messung ist entsprechend

$$\mathrm{bel}_{\text{a posteriori}}(\boldsymbol{x}) := p(\boldsymbol{x} \mid \boldsymbol{l})\,. \tag{11}$$

Die beiden sind durch die Sicherheit (likelihood) $p(\boldsymbol{l} \mid \boldsymbol{x})$ wie gut $\boldsymbol{x}$ die Beobachtungen erklärt verknüpft:

$$\mathrm{bel}_{\text{a posteriori}}(\boldsymbol{x}) = k\; p(\boldsymbol{l} \mid \boldsymbol{x})\; \mathrm{bel}_{\text{a priori}}(\boldsymbol{x})\,. \tag{12}$$

Falls wir die Situation verallgemeinern und sie uns auf der Zeitachse vorstellen, können wir auch schreiben:

$$\mathrm{bel}_t(\boldsymbol{x}) = k\; p(\boldsymbol{l}_t \mid \boldsymbol{x})\; \mathrm{bel}(\boldsymbol{x}_{t-1})\,. \tag{13}$$

Die Konstante dient zur Normierung. Der Index t der Beobachtungen $\boldsymbol{l}_t$ bezieht sich auf den Zeitraum vor t und nach $t-1$, wie etwa in dem folgenden Diagramm mit nach rechts laufender Zeitachse.

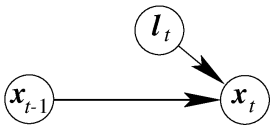

Abbildung 4: Belief propagation, Korrektur der Überzeugung über $\boldsymbol{x}$ durch Beobachtungen $\boldsymbol{l}$

Generell können wir ja nicht davon ausgehen, dass die Überzeugung, die wir vor der Beobachtung über $\boldsymbol{x}$ haben, durch die Beobachtungen gestützt wird. Falls die Beobachtungen und die Vorinformation sich widersprechen, wir also z.B. $l = 3.0$ dm gemessen hätten, hätte sich die Überzeugung, die wir von x haben deutlich geändert, ungefähr auf den gewogenen Mittelwert der Vorinformation und der Messung – es sei denn, dass wir beim Messprozess Ausreißer zugelassen hätten oder die Diskrepanz dazu verwendet hätten, die Genauigkeit unserer Vorinformation und unserer Messung anzupassen.[3]

3 Schätzprinzipien

Wir betrachten nun zwei Schätzverfahren, die sich auf die Likelihood-Funktion $L(\boldsymbol{x})$ und die a posteriori Verteilung $p(\boldsymbol{x} \mid \boldsymbol{l})$ beziehen.

3.1 Die Maximum-Likelihood-Schätzung

Sollten wir nur den Messprozess und die Beobachtung(-en) kennen, ist es plausibel, denjenigen Wert für x zu wählen, der l am besten erklärt, d. h. bei dem $L(x)$ maximal ist. Also erhalten wir die sog. *Maximum-Likelihood-Schätzung* (ML Schätzung) aus

$$\widehat{x}_{\mathrm{ML}} = \mathrm{argmax}_x\; L(x) = \mathrm{argmax}_x\; p(l \mid x)\,. \tag{14}$$

Im Grunde haben wir diese Argumentation im letzten Abschnitt intuitiv als plausibel verwendet, als wir aus der Messung $l = 1.5$ dm auf den Mittelwert von $\underline{x}$ geschlossen

[3] Es ist interessant zu überlegen, wie sich die Bedeutung der Aussagen verändert, falls wir den Begriff "Überzeugung" durch den Begriff "Vorurteil" ersetzen.

haben. D.h. wir haben x im vorigen Abschnitt als ML Schätzung aus der Beobachtung l abgeleitet.

Falls wir einen unbekannten U-Vektor $\boldsymbol{x}$ aus den N Beobachtungen $\boldsymbol{l}$ ableiten wollen und den Messprozess probabilistisch charakterisieren können, d.h. die Verteilung der Beobachtung bei gegebenen Parametern durch die bedingte Dichte beschreiben können

$$\underline{\boldsymbol{l}} \mid \boldsymbol{x} \sim p(\boldsymbol{l} \mid \boldsymbol{x}) \tag{15}$$

erhalten wir mit der Likelihood-Funktion

$$L(\boldsymbol{x}) = p(\boldsymbol{l} \mid \boldsymbol{x}) \tag{16}$$

die ML Schätzung aus

$$\boxed{\widehat{\boldsymbol{x}}_{\mathrm{ML}} = \mathrm{argmax}_x \; L(\boldsymbol{x}) = \mathrm{argmax}_x \; p(\boldsymbol{l} \mid \boldsymbol{x})\,.} \tag{17}$$

Das Prinzip ist sehr allgemein. Nehmen wir als zweites Beispiel das Gauss-Markov Model mit dem mathematischen Modell – in der uns üblichen Schreibweise

$$\underline{\boldsymbol{l}} = \boldsymbol{A}\boldsymbol{x} + \underline{\boldsymbol{e}} \quad \text{und} \quad \underline{\boldsymbol{e}} \sim p(\boldsymbol{e}) = g(\boldsymbol{0}, \Sigma_{ee})\,. \tag{18}$$

Wenn wir die Abhängigkeiten explizit machen und die Gauß-Verteilung durch die ersten beiden Moment, d.h. Mittelwert und Kovarianzmtrix, spezifizieren, erhalten wir

$$\underline{\boldsymbol{l}} \mid \boldsymbol{x} \sim g(\mathbb{E}(\underline{\boldsymbol{l}} \mid \boldsymbol{x}), \mathbb{D}(\underline{\boldsymbol{l}} \mid \boldsymbol{x})) \quad \text{mit} \quad \mathbb{E}(\underline{\boldsymbol{l}} \mid \boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} \quad \text{und} \quad \mathbb{D}(\underline{\boldsymbol{l}} \mid \boldsymbol{x}) = \Sigma_{ee} \tag{19}$$

Auf der linken Seite steht die Zufallsvariable "$\underline{\boldsymbol{l}}$ gegeben $\boldsymbol{x}$", deren Verteilung sich also auf die Situation bezieht, dass $\boldsymbol{x}$ bekannt ist. Dann können wir das auch schreiben als

$$\underline{\boldsymbol{l}} \mid \boldsymbol{x} \sim p(\boldsymbol{l} \mid \boldsymbol{x}) = g(\boldsymbol{l} \mid \boldsymbol{A}\boldsymbol{x}, \Sigma_{ee})\,. \tag{20}$$

oder ausführlich

$$\underline{\boldsymbol{l}} \mid \boldsymbol{x} \sim p(\boldsymbol{l} \mid \boldsymbol{x}) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}(\boldsymbol{l} - \boldsymbol{A}\boldsymbol{x})^{\mathsf{T}} \Sigma_{ee}^{-1} (\boldsymbol{l} - \boldsymbol{A}\boldsymbol{x})\right)\,. \tag{21}$$

Die ML Schätzung für $\boldsymbol{x}$ ergibt sich aus dem Maximum von $p(\boldsymbol{l} \mid \boldsymbol{x})$. Da die Exponentialfunktion monoton steigend ist, können wir statt dessen auch das Minimum des negativen Logarithmus $-\log p(\boldsymbol{l} \mid \boldsymbol{x})$ als ML Schätzung verwenden:

$$\widehat{\boldsymbol{x}}_{\mathrm{ML}} = \mathrm{argmin}_x \; (-\log L(\boldsymbol{x})) = \mathrm{argmin}_x \; (-\log p(\boldsymbol{l} \mid \boldsymbol{x}))\,. \tag{22}$$

oder für den Fall des Gauß-Markov Modells, nach Vernachlässigung konstanter Terme,

$$\boxed{\widehat{\boldsymbol{x}}_{\mathrm{ML}} = \mathrm{argmin}_x \; (\boldsymbol{l} - \boldsymbol{A}\boldsymbol{x})^{\mathsf{T}} \Sigma_{ee}^{-1} (\boldsymbol{l} - \boldsymbol{A}\boldsymbol{x})\,.} \tag{23}$$

Dies ist aber gleichzeitig eine gewogene Kleinste Quadrate-Schätzung (KQ-Schätzung), bei der die Gewichtsmatrix als inverse der Kovarianzmatrix Σ_{ee} verwendet wird:

$$\boxed{\widehat{\boldsymbol{x}}_{\mathrm{KQ}} = \mathrm{argmin}_x \; (\boldsymbol{l} - \boldsymbol{A}\boldsymbol{x})^{\mathsf{T}} \boldsymbol{W} (\boldsymbol{l} - \boldsymbol{A}\boldsymbol{x}) \quad \text{mit} \quad \boldsymbol{W} := \Sigma_{ee}^{-1}\,.} \tag{24}$$

Wenn wir also rückwärts argumentieren, können wir jede gewogene KQ-Schätzung als ML-Schätzung interpretieren indem wir sagen: Hier wurde – implizit, d.h. ohne dies zu sagen oder zu meinen – eine Gaußverteilung mit Kovarianzmatrix $\Sigma_{ee} = \boldsymbol{W}^{-1}$ als statistisches Modell angenommen. Damit kann man die Plausibilität des Modells für die KQ Schätzung, speziell die Plausibilität der Gewichtsmatrix, prüfen.

3.2 Die Bayes-Schätzung

Nun verwenden wir für die Schätzung nicht nur die Beobachtungen, sondern auch die Vorinformation. Dann erhalten wir folgenden Schätzer aus der Maximierung der a posteriori Dichtefunktion der Parameter bei gegebenen Beobachtungen

$$\boxed{\widehat{\boldsymbol{x}}_{\mathrm{MAP}} = \widehat{\boldsymbol{x}}_{\mathrm{Bayes}} = \mathrm{argmax}_x\, p(\boldsymbol{x} \mid \boldsymbol{l})\,.} \tag{25}$$

Es ist der *maximum a posteriori Schätzer* oder kurz MAP-Schätzer. Da die a posteriori Dichte über die Bayesformel mit den Beobachtungen verknüpft ist, wird dieser Schätzer oft auch *Bayes-Schätzer* genannt.[4]

Wie wir aus der Bayesformel (9) erkennen, ist der Nenner $p(\boldsymbol{l})$ irrelevant, da er bei gegebenen Beobachtungen eine Konstante ist. Daher können wir den Bayes-Schätzer auch folgendermaßen definieren:

$$\boxed{\widehat{\boldsymbol{x}}_{\mathrm{MAP}} = \widehat{\boldsymbol{x}}_{\mathrm{Bayes}} = \mathrm{argmax}_x\, p(\boldsymbol{x}) p(\boldsymbol{l} \mid \boldsymbol{x})\,.} \tag{26}$$

Für unser Ausgleichungsbeispiel nehmen wir der Einfachheit halber an, dass die Vorinformation sich durch eine Gauß-Verteilung modellieren lässt, etwa wenn wir die Vorinformation für die Position $\boldsymbol{x}_0$ von Punkten aus einer Karte mit der Ungenauigkeit $\Sigma_{x_0x_0}$ abgreifen. Damit liegen uns zwei Typen von Beobachtungen vor: (1) die – möglicherweise sehr ungenaue – Vorinformation $\{\boldsymbol{x}_0, \Sigma_{x_0x_0})\}$ und (2) die mit einem Messinstrument erzeugten Beobachtungen $(\{\boldsymbol{l}, \Sigma_{ll})$, jetzt $\Sigma_{ee} = \Sigma_{ll}$ nutzend:

$$\underline{\boldsymbol{x}}_0 \quad \sim p(\boldsymbol{x}) \quad = g(\boldsymbol{x} \mid \boldsymbol{x}_0, \Sigma_{x_0x_0}) \tag{27}$$

$$\underline{\boldsymbol{l}} \mid \boldsymbol{x} \quad \sim p(\boldsymbol{l} \mid \boldsymbol{x}) \quad = g(\boldsymbol{l} \mid \boldsymbol{A}\boldsymbol{x}, \Sigma_{ll}) \tag{28}$$

Man kann, mit einigen Umformungen des Produkts (26) der beiden Normalverteilungen in (27) und (28) zeigen, dass die optimale Bayes-Schätzung identisch mit der MAP-Schätzung ist, wenn wir beide Beobachtungsgruppen verwenden um die Parameter zu bestimmen, s. Schuh (2020, Kap. 5.3.4), Bishop (2006, Sect. 2.3.2). Das Modell lautet also

$$\begin{bmatrix} \underline{\boldsymbol{x}}_0 \\ \underline{\boldsymbol{l}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{A} \end{bmatrix} \boldsymbol{x} + \begin{bmatrix} \underline{\boldsymbol{e}}_{x_0} \\ \underline{\boldsymbol{e}} \end{bmatrix} \quad \text{mit} \quad \mathbb{D}\left(\begin{bmatrix} \underline{\boldsymbol{e}}_{x_0} \\ \underline{\boldsymbol{e}} \end{bmatrix} \right) = \begin{bmatrix} \Sigma_{x_0x_0} & \boldsymbol{0} \\ \boldsymbol{0} & \Sigma_{ll} \end{bmatrix}. \tag{29}$$

Das führt auf folgende Schätzung

$$\widehat{\boldsymbol{x}}_{\mathrm{Bayes}} = (\Sigma^{-1}_{x_0x_0} + \boldsymbol{A}^{\mathsf{T}}\Sigma_{ll}\boldsymbol{A})^{-1}(\Sigma^{-1}_{x_0x_0}\boldsymbol{x}_0 + \boldsymbol{A}^{\mathsf{T}}\Sigma_{ll}\boldsymbol{l})\,. \tag{30}$$

Man erhält also wegen

$$\widehat{\boldsymbol{x}}_{\mathrm{Bayes}} = (\Sigma^{-1}_{x_0x_0} + \Sigma^{-1}_{\widehat{x}\widehat{x},\mathrm{ML}})^{-1}(\Sigma^{-1}_{x_0x_0}\boldsymbol{x}_0 + \Sigma^{-1}_{\widehat{x}\widehat{x},\mathrm{ML}}\widehat{\boldsymbol{x}}_{\mathrm{ML}})\,. \tag{31}$$

das Ergebnis: *Die Bayes-Schätzung ist das mit den inversen Kovarianzmatrizen* gewogene Mittel *der Vorinformation und der ML-Schätzung aus den Beobachtungen.*

[4] In der Statistik wird manchmal zwischen dem MAP-Schätzer und dem Bayes-Schätzer unterschieden: Wenn nur die a posteriori Dichte maximiert wird, also nur Wahrscheinlichkeiten ein Rolle spielen, spricht man von MAP Schätzer. Wenn zusätzlich die erwarteten Kosten für den Schätzwert, z.B. über dessen Abweichung vom wahren Wert, berücksichtigt wird, spricht man dann von Bayes-Schätzer. Wir setzen die Begriffe im Folgenden gleich.

Wir können die Bayesformel auch folgendermaßen schreiben

$$p(\boldsymbol{x} \mid \boldsymbol{l}) = \frac{p(\boldsymbol{x})}{p(\boldsymbol{l})}\, p(\boldsymbol{l} \mid \boldsymbol{x}) \tag{32}$$

Falls die Dichtefunktion $p(\boldsymbol{x})$ der Vorinformation eine Konstante ist, ist das gleichbedeutend mit der Aussage: Wir haben keine Vorinformation über $\boldsymbol{x}$.[5] Dann sind die beiden bedingten Dichten proportional.

$$p(\boldsymbol{x} \mid \boldsymbol{l}) = k\, p(\boldsymbol{l} \mid \boldsymbol{x})\,. \tag{33}$$

Dann gilt

$$\widehat{\boldsymbol{x}}_{\text{Bayes}} = \text{argmax}_x\, p(\boldsymbol{x} \mid \boldsymbol{l}) \equiv \widehat{\boldsymbol{x}}_{\text{ML}} = \text{argmax}_x\, p(\boldsymbol{l} \mid \boldsymbol{x})\,. \tag{34}$$

Die Maximum-Likelihood-Schätzung kann als Bayes-Schätzung ohne Vorinformation interpretiert werden

Im Beispiel des Gauß-Markov Modells ist in (21) die Dichte auf der rechten Seite bis auf einen konstanten Faktor identisch mit $p(\boldsymbol{x} \mid \boldsymbol{l})$.

Die gewogene Kleinste-Quadrate-Schätzung ist bei Linearität des Modells und Gauß-Verteilung der Messabweichungen identisch mit einer Bayes-Schätzung ohne Vorinformation.

3.3 Unvollständige Beobachtungen und Levenberg-Marquadt Verfahren

Die Bayes-Schätzung ermöglicht die Schätzung der Parameter, wenn die Beobachtungen unvollständig sind, also die zugehörige Normalgleichungsmatrix $A^{\mathsf{T}}\Sigma_{ll}^{-1}A$ der ML-Schätzung singulär ist:

- Falls die Beobachtungen keine Schätzung der Parameter erlauben, etwa weil es zu wenige sind, kann man (36) auch in der Form

  $$\widehat{\boldsymbol{x}}_{\text{Bayes}} = (\Sigma_{x_0x_0}^{-1} + A^{\mathsf{T}}\Sigma_{ll}^{-1}A)^{-1}(\Sigma_{x_0x_0}^{-1}\boldsymbol{x}_0 + A^{\mathsf{T}}\Sigma_{ll}^{-1}\boldsymbol{l})\,. \tag{35}$$

  verwenden. Falls etwa einer der Punkte, für den Vorinformation vorliegt, nicht beobachtetet wird, ist trotz Singularität von $A^{\mathsf{T}}\Sigma_{ll}^{-1}A$ weiterhin eine Bayes-Schätzung möglich. Sie führt aber zu keiner Veränderung der vor der Messung bekannten/angenommenen Koordinaten dieses Punktes.

- Entsprechend kann man im Falle, dass man bei einer ML-Schätzung nicht weiß, ob alle Parameter durch die Beobachtungen bestimmbar sind, für die Parameter fiktive Vorinformation vorsehen, um die Lösung des Normalgleichungssystems innerhalb einer Bayes-Schätzung zu garantieren. Dies ist innerhalb eines Iterationsverfahrens besonders einfach, da man die fiktiven Zuschläge mit Null ansetzen kann und die Kovarianzmatrix $\Sigma_{x_0x_0}$ als Vielfache einer Einheitsmatrix wählen kann. So erhält man die Zuschläge $\widehat{\Delta\boldsymbol{x}}$ für die Parameter aus

  $$\widehat{\Delta\boldsymbol{x}}_{\text{LM}} = (\lambda I + A^{\mathsf{T}}\Sigma_{ll}^{-1}A)^{-1}\, A^{\mathsf{T}}\Sigma_{ll}^{-1}\Delta\boldsymbol{l}\,. \tag{36}$$

[5] In einem endlichen Universum ist der Wertebereich von $\boldsymbol{x}$ beschränkt, sodass $p(\boldsymbol{x}) = c > 0$ ist. Falls wir den Wertebereich einer z.B. normal-verteilten Größe entsprechend einschränken, hat dies einen vernachlässigbaren Einfluss auf die Verteilung.

Levenberg und Marquardt haben vorgeschlagen, den Faktor $\lambda > 0$, der ja ein Gewicht für die fiktive Vorinformation darstellt, im Lauf der Iterationen klein werden zu lassen, sodass die fiktive Vorinformation die Lösung einerseits ermöglicht wird sie andererseits auch nicht gestört wird, s. Levenberg (1944); Marquardt (1963).

Literatur

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer. 9

Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics 2*(2), 164–168. 11

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics 11*(2), 431–441. 11

Schuh, W. D. (2020). *Ausgleichungsrechnung und Angewandte Statistik*. 5, 9