

## 3.2 Maximum-Likelihood-Schätzung

Die Maximum-Likelihood-Schätzung ist die populärste Methode zur Konstruktion von Punktschätzern bei rein parametrischen Problemstellungen.

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.1 Schätzkonzept

Maximum-Likelihood-Prinzip: Finde Maximum-Likelihood-Schätzwert  $\hat{\theta}$ , so dass

$$L(\hat{\theta}; \mathbf{x}) \geq L(\theta; \mathbf{x}) \text{ für alle } \theta \in \Theta.$$

Dazu äquivalent ist

$$\ell(\hat{\theta}; \mathbf{x}) \geq \ell(\theta; \mathbf{x}), \quad \ell(\theta) = \log L(\theta)$$

mit der Log-Likelihood  $\ell$ . Meist sucht man nach (lokalen) Maxima von  $\ell(\theta)$  durch Nullsetzen der Score-Funktion

$$s(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \left( \frac{\partial \ell(\theta)}{\partial \theta_1}, \dots, \frac{\partial \ell(\theta)}{\partial \theta_p} \right)^\top$$

(soweit die 1. Ableitung der Log-Likelihood existiert!) als Lösung der sogenannten *ML-Gleichung*

$$s(\hat{\theta}) = 0.$$

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.1 Schätzkonzept

Dies funktioniert (meist) unter Annahme von Fisher-Regularität. Nur in einfachen Fällen ist die Lösung analytisch zugänglich. Die numerische Lösung geschieht über Verfahren wie Newton-Raphson, Fisher-Scoring, Quasi-Newton oder über den EM-Algorithmus. Erstere drei Verfahren arbeiten mit der Hesse-Matrix der Log-Likelihood bzw. Approximationen an diese:

$$J(\theta; \mathbf{x}) = -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^\top} = \left( -\frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j} \right)$$

heißt *beobachtete Informationsmatrix*. Bildet man den Erwartungswert bezüglich aller möglichen Stichproben  $X$  aus  $\mathcal{X}$ , so erhält man die *erwartete Informationsmatrix*

$$I(\theta) = \mathbb{E}_\theta[J(\theta; X)].$$

Unter Fisher-Regularität gilt (vgl. Abschnitt 86):

$$\mathbb{E}_\theta[s(\theta)] = 0 \quad \text{und} \quad \text{Cov}_\theta(s(\theta)) = \mathbb{E}_\theta[s(\theta)s(\theta)^\top] = I(\theta).$$

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.1 Schätzkonzept

#### Beispiel 3.3 (Lineares Modell)

Betrachte

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{mit} \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}).$$

- ▶ Likelihood:

$$L(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|^2\right)$$

- ▶ Log-Likelihood:

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \underbrace{\|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|^2}_{\text{KQ-Kriterium}}$$

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.1 Schätzkonzept

- ▶ Score-Funktion:

$$s_{\beta}(\beta, \sigma^2) = \frac{\partial \ell(\beta, \sigma^2)}{\partial \beta} = \frac{1}{\sigma^2} \mathbf{Z}^{\top} (\mathbf{y} - \mathbf{Z}\beta)$$

$$s_{\sigma^2}(\beta, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \|\mathbf{y} - \mathbf{Z}\beta\|^2$$

Man verifiziert leicht, dass  $\mathbb{E}[s_{\beta}] = \mathbb{E}[s_{\sigma^2}] = 0$  ist. Aus den ML-Gleichungen, die sich durch Nullsetzen der Score-Funktionen ergeben, folgt:

$$\hat{\beta}_{\text{ML}} = (\mathbf{Z}^{\top} \mathbf{Z})^{-1} \mathbf{Z}^{\top} \mathbf{y},$$

$$\sigma_{\text{ML}}^2 = \frac{1}{n} \|\mathbf{y} - \mathbf{Z}\hat{\beta}_{\text{ML}}\|^2.$$

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.1 Schätzkonzept

- ▶ Score-Funktion (fortgeführt):

Der ML-Schätzer für  $\beta$  entspricht also dem KQ-Schätzer. Der ML-Schätzer für  $\sigma^2$  ist verzerrt, aber asymptotisch erwartungstreu.

Der Restricted Maximum Likelihood (REML) Schätzer

$$\sigma_{\text{REML}}^2 = \frac{1}{n - p} \|\mathbf{y} - \mathbf{Z}\beta\|^2$$

ist erwartungstreu für  $\sigma^2$ . Dabei ist  $p$  die Dimension von  $\beta$ .

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.1 Schätzkonzept

- ▶ Informationsmatrizen:

$$-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\frac{\partial s_{\boldsymbol{\beta}}}{\partial \boldsymbol{\beta}^\top} = \frac{1}{\sigma^2} \mathbf{Z}^\top \mathbf{Z} = \left( \text{Cov}(\hat{\boldsymbol{\beta}}) \right)^{-1} \quad (\text{von } \mathbf{y} \text{ unabhängig})$$

$$-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \sigma^2} = \frac{1}{\sigma^4} \mathbf{Z}^\top (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) \quad \Rightarrow \quad \mathbb{E} \left[ -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \sigma^2} \right] = 0$$

$$-\frac{\partial^2 \ell}{\partial \sigma^2 \partial \sigma^2} = -\frac{n}{2(\sigma^2)^2} + \frac{\|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|^2}{(\sigma^2)^3} \quad \Rightarrow \quad \mathbb{E} \left[ -\frac{\partial^2 \ell}{\partial \sigma^2 \partial \sigma^2} \right] = \frac{n}{2\sigma^4}$$

Der letzte Erwartungswert folgt aus

$$\|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|^2 = \sum_{i=1}^n \varepsilon_i^2 \sim \sigma^2 \chi^2(n).$$

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.1 Schätzkonzept

#### Beispiel 3.4 (GLM)

Seien  $y_i \stackrel{\text{unabh.}}{\sim} f(y_i|\mu_i)$  für  $i = 1, \dots, n$  mit  $\mu_i = h(\mathbf{x}_i^\top \boldsymbol{\beta})$ , etwa  $y_i \sim \text{Po}(\lambda_i)$  und  $\lambda_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$  (loglineares Poisson-Modell, vgl. generalisierte Regression).

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.1 Schätzkonzept

#### Beispiel 3.5 (GLMM für Longitudinaldaten)

Sei  $\mathbf{y}_i = (y_{i1}, \dots, y_{it}, \dots, y_{iT})$  mit bedingt unabhängigen Komponenten  $y_{it} \sim f(y_{it}|\mu_{it})$  und  $\mu_{it} = h(\mathbf{z}_{it}^\top \boldsymbol{\beta} + \mathbf{w}_{it}^\top \boldsymbol{\gamma}_i)$ . Die  $\boldsymbol{\gamma}_i$  sind z.B. individuenspezifische Intercepts ( $\mathbf{w}_{it} \equiv 1$ ) mit Priorverteilung  $\boldsymbol{\gamma}_i \stackrel{\text{i.i.d.}}{\sim} N(0, \tau^2)$ . Die Likelihood des Parameters  $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau^2)$  lautet

$$L(\boldsymbol{\beta}, \tau^2) = \int \prod_{i=1}^n \prod_{t=1}^T f(y_{it}|\boldsymbol{\beta}, \tau^2, \boldsymbol{\gamma}_i) p(\boldsymbol{\gamma}_i) d\boldsymbol{\gamma}_i.$$

Lösungsansätze für die Maximierung der Likelihood:

EM-Algorithmus oder numerische Integration mit REML bzw. Bayes-Inferenz. Siehe die Vorlesungen Gemischte Modelle und Analyse Longitudinaler Daten.

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

#### **EM (Expectation-Maximization)-Algorithmus**

Der EM-Algorithmus ist eine Alternative zu Newton-Raphson, Fisher-Scoring usw., vor allem in Modellen mit unvollständigen Daten oder latenten (nicht direkt beobachtbaren) Variablen oder Faktoren (vgl. Computerintensive Methoden).

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

#### Beispiele:

- ▶ Bei einer Screening-Untersuchung werden positive Testergebnisse eines Schnelltests durch den Goldstandard geprüft, negative Testergebnisse jedoch nicht. Daher ist die Verteilung der Anzahl positiver Testergebnisse aus  $k$  Tests bei den wirklich Kranken bekannt, nicht jedoch die Häufigkeit von 0 positiven Testergebnissen aus  $k$ . Für die Ermittlung der Falsch-Negativrate wäre dieser nicht beobachtete Wert jedoch relevant.
- ▶ Die Daten stammen aus einer Mischverteilung, aus welcher Mischungskomponente jede Beobachtung stammt, ist jedoch unbekannt.
- ▶ Gemischte Modelle

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

*Notation:*

- ▶  $\mathbf{x}$  beobachtbare („unvollständige“) Daten
- ▶  $\mathbf{z}$  unbeobachtbare Daten/latente Variablen
- ▶  $(\mathbf{x}, \mathbf{z})$  vollständige Daten
- ▶  $L(\theta; \mathbf{x}) = f(\mathbf{x}|\theta)$  Likelihood der beobachtbaren Daten
- ▶  $L(\theta; \mathbf{x}, \mathbf{z}) = f(\mathbf{x}, \mathbf{z}|\theta)$  Likelihood der vollständigen Daten

Der EM-Algorithmus ist insbesondere nützlich, wenn  $L(\theta; \mathbf{x})$  schwierig zu berechnen und  $L(\theta; \mathbf{x}, \mathbf{z})$  leichter zu handhaben ist.

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

#### Intuition:

$$\begin{aligned}\ell(\theta; \mathbf{x}, \mathbf{z}) &= \ell(\theta; \mathbf{z}|\mathbf{x}) + \ell(\theta; \mathbf{x}) \\ \Rightarrow \underbrace{\mathbb{E}[\ell(\theta; \mathbf{x}, \mathbf{Z})]}_{Q(\theta; \theta^{(k)})} &= \underbrace{\mathbb{E}[\ell(\theta; \mathbf{Z}|\mathbf{x})]}_{C(\theta; \theta^{(k)})} + \ell(\theta; \mathbf{x}),\end{aligned}$$

wobei der Erwartungswert  $\mathbb{E}$  bzgl.  $f(\mathbf{Z}|\mathbf{x}; \theta^{(k)})$  genommen wird.

Es gilt  $C(\theta^{(k)}; \theta^{(k)}) \geq C(\theta^{(k+1)}; \theta^{(k)})$  wegen der Informationsungleichung.

Daher folgt aus  $Q(\theta^{(k+1)}; \theta^{(k)}) \geq Q(\theta^{(k)}; \theta^{(k)})$

$$\ell(\theta^{(k+1)}; \mathbf{x}) \geq \ell(\theta^{(k)}; \mathbf{x}).$$

Vorgehen: Für gegebenes  $\theta^{(k)}$  maximiere  $Q(\theta; \theta^{(k)})$  über  $\theta$ . Neues  $\theta^{(k+1)}$  erhöht die Likelihood. Iteriere.

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

---

#### Algorithmus 1 : EM-Algorithmus

---

Startwert:  $\theta^{(0)}$

- ▶ **E**-Schritt: Berechne

$$Q(\theta) = Q(\theta; \theta^{(0)}) = \mathbb{E}_{\mathbf{z}|\mathbf{x}}[\ell(\theta; \mathbf{x}, \mathbf{Z})|\mathbf{x}, \theta^{(0)}].$$

- ▶ **M**-Schritt: Berechne  $\theta^{(1)}$ , so dass  $Q(\theta)$  maximiert wird:

$$\theta^{(1)} = \underset{\theta}{\operatorname{argmax}} Q(\theta).$$

Iteriere **E**/**M**-Schritte:  $\theta^{(0)}, \theta^{(1)}, \dots, \theta^{(k)}$  bis zur Konvergenz.

---

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

#### Satz 3.3

Unter relativ allgemeinen Annahmen gilt  $\theta^{(k)} \rightarrow \hat{\theta}_{ML}$  für  $k \rightarrow \infty$ .

*Eigenschaften des EM-Algorithmus:*

- ▶ Monotonie:  $\ell(\theta^{(k+1)}; \mathbf{x}) \geq \ell(\theta^{(k)}; \mathbf{x})$ .
- ▶ Langsame Konvergenz.
- ▶ Der Standardfehler des resultierenden Schätzers ist schwierig zu bestimmen, die Informationsmatrix ist nicht direkt zugänglich wie beim Fisher-Scoring.

Eine Alternative bietet u.a. die Bayes-Inferenz.

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

#### **Beispiel 3.6** (Mischverteilungen)

Seien  $X_1, \dots, X_n$  i.i.d. wie  $X \sim f(x|\theta)$ . Betrachte die Mischverteilung

$$f(x|\theta) = \sum_{j=1}^J \pi_j f_j(x|\theta_j) \quad \text{mit} \quad \theta = (\{\theta_j\}_{j=1}^J, \{\pi_j\}_{j=1}^J). \quad (3.2)$$

Dabei sind

- ▶  $\pi_j$  unbekannte Mischungsanteile,  $\sum_{j=1}^J \pi_j = 1$ ,
- ▶  $f_j(x|\theta_j)$  die  $j$ -te Mischungskomponente,
- ▶  $\theta_j$  der unbekannte Parameter(-vektor) .

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

Speziell: Bei einer Mischung von Normalverteilungen erhalten wir

$$f_j(x|\theta_j) \propto |\Sigma_j|^{-1/2} \exp\left(-\frac{1}{2}(x - \mu_j)^\top \Sigma_j^{-1}(x - \mu_j)\right)$$
$$X \sim \pi_1 N(\mu_1, \Sigma_1) + \pi_2 N(\mu_2, \Sigma_2) + \dots + \pi_J N(\mu_J, \Sigma_J).$$

Im univariaten Fall mit zwei Mischungskomponenten also:

$$X \sim \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2).$$

Interpretation des Mischungsmodells (3.2):  $x_i$  entstammt einer von  $J$  Subpopulationen, wobei in Subpopulation  $j$  gilt:

$$X_i|j \sim f_j(x_i|\theta_j).$$

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

Definiere die unbeobachtete (latente) Indikatorvariable  $Z_i$  für  $j = 1, \dots, J$  durch

$$Z_i = j \Leftrightarrow x_i \text{ ist aus Population } j.$$

Die Randverteilung sei  $\mathbb{P}(Z_i = j) = \pi_j$ ,  $j = 1, \dots, J$ . Dann lautet die bedingte Verteilung von  $x_i|Z_i$ :

$$x_i|Z_i = j \sim f_j(x_i|\theta_j).$$

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

Die Log-Likelihood der beobachteten Daten  $x$  ist

$$\ell(\theta; x) = \sum_{i=1}^n \log \left( \sum_{j=1}^J \pi_j f_j(x_i | \theta_j) \right),$$

die der vollständigen Daten  $(x, z)$

$$\begin{aligned} \ell(\theta; x, z) &= \sum_{i=1}^n \log f(x_i, z_i | \theta) \\ &= \sum_{i=1}^n \log (f(x_i | z_i; \theta) \cdot f(z_i)) \\ &= \sum_{i=1}^n (\log f_{z_i}(x_i | \theta_{z_i}) + \log \pi_{z_i}). \end{aligned}$$

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

E-Schritt:

$$\begin{aligned} Q(\theta) &= \mathbb{E}_{\mathbf{z}|\mathbf{x}}[\ell(\theta; \mathbf{x}, \mathbf{Z})|\mathbf{x}, \theta^{(k)}] \\ &= \sum_{i=1}^n \sum_j^J p_{ij}^{(k)} \left\{ \log \pi_j - \frac{1}{2} \log |\Sigma_j| - \frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \right\} \end{aligned}$$

wobei wir nur

$$p_{ij}^{(k)} = \mathbb{P}(Z_i = j | x_i, \theta^{(k)}) \stackrel{\text{Bayes}}{=} \frac{\pi_j^{(k)} f_j(x_i | \theta_j^{(k)})}{\sum_{s=1}^J \pi_s^{(k)} f_j(x_i | \theta_s^{(k)})}.$$

für  $i = 1, \dots, n$ ,  $j = 1, \dots, J$  tatsächlich in der Praxis berechnen müssen.

## 3.2 Maximum-Likelihood-Schätzung

### 3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

$$Q(\theta) = \sum_{i=1}^n \sum_j^J p_{ij}^{(k)} \left\{ \log \pi_j - \frac{1}{2} \log |\Sigma_j| - \frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \right\}$$

**M**-Schritt: Berechne

$$\pi_j^{(k+1)} = \operatorname{argmax}_{\pi_j} Q(\theta) \stackrel{1.}{=} \frac{1}{n} \sum_{i=1}^n p_{ij}^{(k)}$$

$$\mu_j^{(k+1)} = \operatorname{argmax}_{\mu_j} Q(\theta) \stackrel{2.}{=} \sum_{i=1}^n w_{ij}^{(k)} x_i$$

$$\Sigma_j^{(k+1)} = \operatorname{argmax}_{\Sigma_j} Q(\theta) \stackrel{3.}{=} \sum_{i=1}^n w_{ij}^{(k)} (x_i - \mu_j^{(k+1)})(x_i - \mu_j^{(k+1)})^T$$

mit  $w_{ij}^{(k)} = \frac{p_{ij}^{(k)}}{\sum_{s=1}^J p_{is}^{(k)}}$ . 1. folgt für  $J = 2$  als Maximierer der binomialen Likelihood (für  $J > 2$  braucht man Lagrange). 2.+3. folgt als Maximierer der gewichteten Normalverteilungsl likelihood.