Kapitel 3

Likelihood-Inferenz

3.1 Parametrische Likelihood-Inferenz

Situation: $\mathcal{P}_\theta = \{f(\boldsymbol{x}|\theta) : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^p$, $p \ll n$, p konstant für $n \to \infty$. $f(\boldsymbol{x}|\theta)$ ist eine diskrete oder stetige oder allgemeiner eine Radon-Nikodym-Dichte.

Definition 3.1 (Likelihoodfunktion)**.** *Die* Likelihoodfunktion *von* $\theta \in \Theta$*,*

$$L(\theta) = f(\boldsymbol{x}|\theta),$$

ist definiert als die Dichte der beobachteten Daten $\boldsymbol{X} = (X_1, \ldots, X_n) = \boldsymbol{x} = (x_1, \ldots, x_n)$*, betrachtet als Funktion von* θ*. Mit* $L(\theta)$ *ist auch* $\widetilde{L}(\theta) = const \times L(\theta)$ *eine Likelihoodfunktion.*

Zu unterscheiden sind folgende Situationen:

1. $X_1, \ldots, X_n$ sind i.i.d. wie $X_i \sim f_1(x|\theta)$ (Statistik IV). Es gilt die Faktorisierung

$$L(\theta) = \prod_{i=1}^{n} f_1(x_i|\theta).$$

2. $X_1, \ldots, X_n$ — bzw. $Y_1|\boldsymbol{z}_1, \ldots, Y_n|\boldsymbol{z}_n$ im Regressionsfall bei einer Zielvariable $\boldsymbol{Y}$ und Kovariablenvektor $\boldsymbol{z}$ — sind unabhängig, aber nicht mehr identisch verteilt. Es gilt die Faktorisierung

$$L(\theta) = \prod_{i=1}^{n} f_i(x_i|\theta).$$

3. Die Paare $(X_1^d, X_1^s), \ldots, (X_i^d, X_i^s), \ldots, (X_n^d, X_n^s)$ sind unabhängig, die einzelnen Komponenten innerhalb eines Paares unter Umständen abhängig. Die Indizes s, d beziehen sich auf stetige bzw. diskrete Variablen. Eine derartige Datenlage ergibt sich beispielsweise bei Survivaldaten mit stetigen Überlebenszeiten und einem diskreten Zensierungsindikator $X_i^d = I(C_i \leq T_i)$, wobei C_i bzw. T_i den Zensierungs- bzw. Verweildauerprozess bezeichnen. Unter obige Situation fallen auch Mischverteilungsmodelle. X_i^d enspricht dann einer Klassenzugehörigkeit und X_i^s einem stetigen Merkmal(svektor).

4. Zeitlich korrelierte Daten / Stichprobenvariablen $X_1, \ldots, X_t, \ldots, X_n$ mit Dichtefunktion

$$f(x_1, \ldots, x_t, \ldots, x_n|\theta) = \\ f(x_n|x_{n-1}, \ldots, x_t, \ldots, x_1; \theta) \cdot \ldots \cdot f(x_{n-1}|x_{n-2}, \ldots, x_1; \theta) \cdot \ldots \cdot f(x_2|x_1; \theta) f(x_1|\theta).$$

Bei Markov-Ketten erster Ordnung mit der Eigenschaft

$$f(x_n|x_{n-1}, \ldots, x_1; \theta) = f(x_n|x_{n-1}; \theta)$$

vereinfacht sich die Likelihood zu

$$L(\theta) = \left(\prod_{i=2}^{n} f(x_i|x_{i-1}; \theta) \right) f(x_1|\theta).$$

Beispiel 3.1 (zu diesen vier Situationen)**.**

1. *Siehe Statistik IV bzw. Grundstudium.*
2. *Regressionssituationen (Querschnittsdaten) mit unabhängigen Zielvariablen $Y_1|\boldsymbol{z}_1, \ldots, Y_n|\boldsymbol{z}_n$ und festen Kovariablen $\boldsymbol{z}_i$:*
   - *klassisches lineares Modell: $Y_i|\boldsymbol{z}_i \sim N(\boldsymbol{z}_i^\top \boldsymbol{\beta}, \sigma^2)$,*
   - *Logit- oder Probitmodell: $Y_i|\boldsymbol{z}_i \sim Bin(1, \pi_i = h(\boldsymbol{z}_i^\top \boldsymbol{\beta}))$,*
   - *Poisson-Regression: $Y_i|\boldsymbol{z}_i \sim Po(\lambda_i = h(\boldsymbol{z}_i^\top \boldsymbol{\beta}))$.*
3. *Markov-Ketten, autoregressive Modelle für Zeitreihen/Longitudinaldaten.*
4. *Autoregressiver Prozess 1. Ordnung: Sei*

$$X_t = \alpha + \gamma X_{t-1} + \varepsilon_t$$

mit $\varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$ oder — mit zusätzlichem (zeitabhängigen) Kovariablenvektor $\boldsymbol{z}_t$ —

$$X_t = \alpha + \gamma X_{t-1} + \boldsymbol{z}_t^\top \boldsymbol{\beta} + \varepsilon_t.$$

In letzterem Fall hat die Likelihood die Form

$$L(\theta) = \left(\prod_{i=2}^{n} f_i(x_i|x_{i-1}; \theta) \right) f_1(x_1)$$

mit

$$f_i(x_i|x_{i-1}; \theta) = \phi(x_i|\alpha + \gamma x_{i-1} + \boldsymbol{z}_i^\top \boldsymbol{\beta}, \sigma^2),$$

wobei $\phi(x|\mu, \tau^2)$ den Wert der Normalverteilungsdichte mit Erwartungswert μ und Varianz τ^2 an der Stelle x bezeichnet.

Beispiel 3.2. *Wir betrachten unabhängige, aber (teils) unvollständige Ziehungen aus $N(\theta, 1)$.*

1. *Ziehung: Es sei $x_1 = 2.45$. Dann ist*

$$L_1(\theta) = \phi(x_1 - \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(2.45 - \theta)^2 \right).$$

2. Ziehung: *Es sei nur* $0.9 < x_2 < 4$ *bekannt (unvollständige oder intervallzensierte Beobachtung). Die Likelihood lautet dann:*

$$L_2(\theta) = \mathbb{P}_\theta(0.9 < X_2 < 4) = \Phi(4-\theta) - \Phi(0.9-\theta).$$

Formal könnte man auch eine binäre Variable

$$X_2^d = \begin{cases} 1, \ 0.9 < X_2 < 4, \\ 0, \ \textit{sonst} \end{cases}$$

mit Dichtefunktion

$$f_2^d(1) = \mathbb{P}(X_2^d = 1) = \Phi(4-\theta) - \Phi(0.9-\theta)$$

definieren.

3. Ziehung: $z_1, \ldots, z_n$ *seien i.i.d. Realisierungen aus* $N(\theta, 1)$. *Bekannt sei aber nur*

$$x_3 = \max_{1 \le i \le n} z_i = z_{(n)}.$$

Der Rest sind fehlende Werte („missing values"). Die Verteilungsfunktion von $X_3 = Z_{(n)}$ *ist*

$$\begin{aligned} F_\theta(z_{(n)}) &= \mathbb{P}_\theta(Z_{(n)} \le z_{(n)}) = \mathbb{P}_\theta(Z_i \le z_{(n)} \ \forall \ i) \\ &= [\Phi(z_{(n)} - \theta)]^n. \end{aligned}$$

Die Dichte ergibt sich über Differentiation bezüglich θ*:*

$$f_\theta(z_{(n)}) = n[\Phi(z_{(n)} - \theta)]^{n-1} \phi(z_{(n)} - \theta),$$

d.h. für zum Beispiel $n = 5$ *und* $z_{(n)} = x_3 = 3.5$ *gilt*

$$L_3(\theta) = 5[\Phi(3.5 - \theta)]^4 \phi(3.5 - \theta).$$

Die gesamte Likelihood der drei Beobachtungen ist

$$L(\theta) = L_1(\theta) \cdot L_2(\theta) \cdot L_3(\theta),$$

also das Produkt der Likelihoodfunktionen L_1*,* L_2 *und* L_3*.*

Fazit: Die Likelihood ist sehr allgemein definiert.

Beziehung zur Bayes-Inferenz

- $p(\theta)$ sei die Prioriverteilung,
- $f(x|\theta) = L(\theta)$ die Likelihood.
- Dann ist

$$\begin{aligned} p(\theta|x) &\propto p(\theta) \cdot L(\theta) \\ \text{„Posteriori"} &\propto \text{„Priori"} \times \text{Likelihood}. \end{aligned}$$

Likelihood-Quotient

Frage: Wie vergleicht man die Likelihoods $L(\theta_1)$ und $L(\theta_2)$ für $\theta_1 \neq \theta_2$?

Antwort: Man betrachtet den Quotienten (nicht die Differenz), da dieser invariant gegenüber eineindeutigen Transformationen

$$x \mapsto y = y(x) \;\Leftrightarrow\; y \mapsto x(y)$$

ist. Für stetige x, y gilt mit dem Transformationssatz für Dichten:

$$f_Y(y|\theta) = f_X(x(y)|\theta) \left|\det\left(\frac{\partial x}{\partial y}\right)\right|$$

und somit

$$L(\theta; y) = L(\theta; x) \left|\det\left(\frac{\partial x}{\partial y}\right)\right| \;\Rightarrow\; \frac{L(\theta_2; y)}{L(\theta_1; y)} = \frac{L(\theta_2; x)}{L(\theta_1; x)}.$$

Satz 3.2.

1. *Sei $T = T(X)$ suffizient für θ. Dann gilt $L(\theta; x) = const \times L(\theta; t)$ mit $t = T(x)$, d.h. $L(\theta; x)$ und $L(\theta; t)$ sind äquivalent.*
2. *$L(\theta; x)$ ist minimalsuffzient.*

Beweis. Folgt unmittelbar aus den Resultaten aus Abschnitt 2. □

3.2 Maximum-Likelihood-Schätzung

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

3.2.1 Schätzkonzept

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

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

Dazu äquivalent ist

$$\ell(\widehat{\theta}; x) \geq \ell(\theta; x), \;\; \ell(\theta) = \log L(\theta)$$

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

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

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

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

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:

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

heißt *beobachtete Informationsmatrix*. Bildet man den Erwartungswert bezüglich allen 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 2):

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

Beispiel 3.3 (Lineares Modell)**.** *Betrachte*

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

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

- *Log-Likelihood:*
$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\underbrace{\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2}_{\textit{KQ-Kriterium}}$$

- *Score-Funktion:*
$$\begin{aligned} s_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma^2) &= \frac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\boldsymbol{Z}^\top(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) \\ s_{\sigma^2}(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2 \end{aligned}$$

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

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{ML} &= (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{y}, \\ \sigma^2_{ML} &= \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2. \end{aligned}$$

  Der ML-Schätzer für $\boldsymbol{\beta}$ *entspricht also dem KQ-Schätzer. Der ML-Schätzer für* σ^2 *ist verzerrt, aber asymptotisch erwartungstreu. Der Restricted Maximum Likelihood (REML) Schätzer*

$$\sigma^2_{REML} = \frac{1}{n-p}\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2$$

  ist erwartungstreu für σ^2*. Dabei ist* p *die Dimension von* $\boldsymbol{\beta}$*.*

- *Informationsmatrizen:*

$$
\begin{aligned}
-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= -\frac{\partial s_{\boldsymbol{\beta}}}{\partial \boldsymbol{\beta}^\top} = \frac{1}{\sigma^2} \boldsymbol{Z}^\top \boldsymbol{Z} = \left(\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) \right)^{-1} && \textit{(von } \boldsymbol{y} \textit{ unabhängig)} \\
-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \sigma^2} &= \frac{1}{\sigma^4} \boldsymbol{Z}^\top (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) && \Rightarrow \ \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \sigma^2} \right] = 0 \\
-\frac{\partial^2 \ell}{\partial \sigma^2 \partial \sigma^2} &= -\frac{n}{2(\sigma^2)^2} + \frac{\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2}{(\sigma^2)^3} && \Rightarrow \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \sigma^2 \partial \sigma^2} \right] = \frac{n}{2\sigma^4}
\end{aligned}
$$

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

Beispiel 3.4 (GLM)**.** *Seien* $y_i \overset{unabh.}{\sim} f(y_i|\mu_i)$ *für* $i = 1, \ldots, n$ *mit* $\mu_i = h(\boldsymbol{x}_i^\top \boldsymbol{\beta})$*, etwa* $y_i \ \sim Po(\lambda_i)$ *und* $\lambda_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ *(loglineares Poisson-Modell, vgl. Übung/generalisierte Regression).*

Beispiel 3.5 (GLMM für Longitudinaldaten)**.** *Sei* $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{it}, \ldots, y_{iT})$ *mit bedingt unabhängigen Komponenten* $y_{it} \sim f(y_{it}|\mu_{it})$ *und* $\mu_{it} = h(\boldsymbol{z}_i^\top \boldsymbol{\beta} + \boldsymbol{w}_i^\top \boldsymbol{\gamma}_i)$*. Die* $\boldsymbol{\gamma}_i$ *sind individuenspezifische Intercepts (*$\boldsymbol{w}_i \equiv \boldsymbol{1}$*) mit Prioriverteilung* $\gamma_{it} \overset{i.i.d.}{\sim} N(0, \tau^2)$*. Die Likelihood des Parameters* $\theta = (\boldsymbol{\beta}, \tau^2)$ *lautet*

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

Lösungsansätze für die Maximierung der Likelihood: EM-Algorithmus mit REML bzw. Bayes-Inferenz.

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).

Notation:

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

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

Algorithmus 1 : **EM-Algorithmus**

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

- **E**-Schritt: Berechne

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

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

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

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

Satz 3.3. *Unter relativ allgemeinen Annahmen gilt* $\theta^{(k)} \to \widehat{\theta}_{ML}$ *für* $k \to \infty$.

Eigenschaften des EM-Algorithmus:

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

Eine Alternative bietet die Bayes-Inferenz.

Beispiel 3.6 (Mischverteilungen)**.** *Seien* $X_1, \ldots, X_n$ *i.i.d. wie* $X \sim f(x|\theta)$. *Betrachte die Mischverteilung*

$$f(x|\theta) = \sum_{j=1}^{J} \pi_j f_j(x|\theta_j) \quad \textit{mit} \quad \theta = (\{\theta\}_{j=1}^{J}, \{\pi_j\}_{j=1}^{J}). \tag{3.1}$$

Dabei sind

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

Speziell: Bei einer Mischung von Normalverteilungen erhalten wir

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

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 (3.1): x_i entstammt einer von J Subpopulationen, wobei in Subpopulation j gilt:

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

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

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

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

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

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)

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

***E**-Schritt:*

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

wobei wir nur

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

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

***M**-Schritt: Berechne*

$$\begin{aligned} \pi_j^{(k+1)} &= \underset{\pi_j}{\operatorname{argmax}} \, Q(\theta) \overset{1.}{=} \frac{1}{n} \sum_{i=1}^{n} p_{ij}^{(k)} \\ \mu_j^{(k+1)} &= \underset{\mu_j}{\operatorname{argmax}} \, Q(\theta) \overset{2.}{=} \sum_{i=1}^{n} w_{ij}^{(k)} x_i \\ \Sigma_j^{(k+1)} &= \underset{\Sigma_j}{\operatorname{argmax}} \, Q(\theta) \overset{3.}{=} \sum_{i=1}^{n} w_{ij}^{(k)} (x_i - \mu_j^{(k+1)})(x_i - \mu_j^{(k+1)})^T \end{aligned}$$

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 Normalverteilungslikelihood.

Beispiel 3.7 (Gemischte Modelle)**.** *Eine Herleitung für E- und M-Schritt in linearen gemischten Modellen findet sich in Pawitan, Kapitel 12.8.*

3.2.3 Asymptotische Eigenschaften

Satz 3.4. *Seien $X_1, \ldots, X_n$ i.i.d. aus einer Dichte $f(x|\theta)$, die folgenden Annahmen genügt:*

- *$f(x|\theta)$ ist Fisher-regulär.*
- *Die Informationsmatrix $\mathbf{I}(\theta)$ ist positiv definit im Inneren von Θ.*
- *Es existieren Funktionen M_{jkl} derart, dass*

$$\left| \frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} \log f(x|\theta) \right| \leq M_{jkl}(x)$$

und

$$\mathbb{E}_{\theta_0}[M_{jkl}(X)] < \infty$$

für alle j, k und l, wobei θ_0 den wahren Wert des Parameters bezeichnet.

Dann gilt unter weiteren, relativ schwachen Regularitätsannahmen:

- *Die Likelihood-(ML-)Gleichungen haben für $n \to \infty$ mit Wahrscheinlichkeit 1 eine Lösung $\widehat{\theta}_n$ (d.h. $\mathbb{P}(\widehat{\theta}_n$ existiert$) \to 1$) mit $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$; die konsistente Lösung $\widehat{\theta}_n$ ist eindeutig und $\mathbb{P}(\widehat{\theta}_n$ ist (lokales) Maximum$) \to 1$.*
- *$\widehat{\theta}_n \overset{a}{\sim} N(\theta_0, \mathbf{I}_n^{-1}(\theta_0))$ bzw. $\mathbf{I}_n^{1/2}(\theta_0)(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I)$,*
- *$\widehat{\theta}_n \overset{a}{\sim} N(\theta_0, \mathbf{J}_n^{-1}(\theta_0))$ bzw. $\mathbf{J}_n^{1/2}(\theta_0)(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I)$,*

d.h. ML-Schätzer sind asymptotisch erwartungstreue BAN-Schätzer.

Bemerkung.

1. *Es sind auch andere Varianten von Regularitätsannahmen möglich.*
2. *Der Satz gilt unter stärkeren Regularitätsannahmen auch für i.n.i.d. und abhängige $X_1, \ldots, X_n$.*
3. *$\mathbf{I}(\theta_0)$ und $\mathbf{J}(\theta_0)$ können auch durch $\mathbf{I}(\widehat{\theta}_n)$ bzw. $\mathbf{J}(\widehat{\theta}_n)$ ersetzt werden.*

Beweis. Erfolgt lediglich skizzenhaft.

- *Existenz (für skalares θ):* Es gilt, dass

$$\mathbb{P}_{\theta_0}\left(\prod_{i=1}^n f(x_i|\theta_0) > \prod_{i=1}^n f(x_i|\theta) \right) \to 1 \text{ für } n \to \infty \text{ für alle } \theta \neq \theta_0.$$

Beweis: Logarithmieren liefert

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(f(x_i|\theta)/f(x_i|\theta_0)\right) < 0.$$

Nach dem Gesetz der großen Zahlen konvergiert die linke Seite in Wahrscheinlichkeit gegen die Kullback-Leibler-Distanz

$$\mathbb{E}_{\theta_0}[\log\left(f(x|\theta)/f(x|\theta_0)\right)].$$

Anwendung der Ungleichung von Jensen liefert, dass

$$\mathbb{E}_{\theta_0}[\log\left(f(x|\theta)/f(x|\theta_0)\right)] < \log E_{\theta_0}[f(x|\theta)/f(x|\theta_0)] = 0,$$

woraus die Behauptung folgt. Wähle nun $a > 0$ klein genug, so dass $(\theta_0 - a; \theta_0 + a)$ vollständig in Θ enthalten ist. Setze

$$S_n = \{x : L(\theta_0; x) > L(\theta_0 - a; x) \text{ und } L(\theta_0; x) > L(\theta_0 + a; x)\}.$$

Für beliebige Stichproben $x \in S_n$ existiert somit ein Punkt $\widehat{\theta}_n \in (\theta_0 - a; \theta_0 + a)$, der die Likelihood (lokal) maximiert, d.h. $s(\widehat{\theta}_n) = 0$. Aus eben bewiesener Hilfsaussage folgt, dass $\mathbb{P}_{\theta_0}(S_n) \to 1$ für jedes beliebige a.

- *Konsistenz und Eindeutigkeit (für skalares θ):* siehe Huzurbazar (1948).
- *Asymptotische Normalität der Score-Funktion:* Aus der Fisher-Regularität folgt, dass der Erwartungswert und die Kovarianzmatrix existieren und durch $\mathbb{E}[s_i(\theta)] = 0$ und $\mathrm{Cov}(s_i(\theta)) = \mathbf{i}(\theta)$ gegeben sind. Der zentrale Grenzwertsatz liefert $s(\theta) \overset{a}{\sim} N(0, \mathbf{I}(\theta))$ bzw.
$$\mathbf{I}(\theta)^{-1/2}s(\theta) = (n\mathbf{i}(\theta))^{-1/2}(\sum_{i=1}^{n} s_i(\theta) - 0) \xrightarrow{d} N(0, I).$$
- *Asymptotische Normalität von $\widehat{\theta}_n$:* Eine Taylorentwicklung von $s(\widehat{\theta}_n) = 0$ um θ führt zu
$$0 = s(\widehat{\theta}_n) = s(\theta) - \mathbf{J}(\theta)(\widehat{\theta}_n - \theta) + o(\widehat{\theta}_n - \theta).$$
Wegen dem Satz von Slutsky können wir im Folgenden auch $\mathbf{J}(\theta)$ durch $\mathbf{I}(\theta) = \mathbb{E}[\mathbf{J}(\theta)]$ ersetzen, da $\frac{1}{n}\mathbf{J}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{j}_i(\theta) \xrightarrow{\mathbb{P}} \mathbf{i}(\theta)$. Dies liefert
$$s(\theta) \overset{a}{\sim} \mathbf{I}(\theta)(\widehat{\theta}_n - \theta) \quad \text{bzw.} \quad \widehat{\theta}_n - \theta \overset{a}{\sim} \mathbf{I}^{-1}(\theta)s(\theta)$$
und somit
$$\widehat{\theta}_n - \theta \overset{a}{\sim} N(0, \mathbf{I}^{-1}(\theta)\mathbf{I}(\theta)\mathbf{I}^{-1}(\theta)) = N(0, \mathbf{I}^{-1}(\theta)).$$

□

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

Betrachte lineare Hypothesen

$$H_0 : \boldsymbol{C}\theta = \boldsymbol{d} \quad \text{vs.} \quad H_1 : \boldsymbol{C}\theta \neq \boldsymbol{d},$$

wobei $\boldsymbol{C}$ vollen Zeilenrang $s \leq p = \dim(\theta)$ besitze.

Wichtiger Spezialfall:

$$H_0 : \theta_s = 0 \quad \text{vs.} \quad H_1 : \theta_s \neq 0,$$

wobei θ_s einen beliebigen s-dimensionalen Subvektor von θ bezeichnet, zum Beispiel in einem GLM, wo $\boldsymbol{\beta}_s = \mathbf{0}$ bedeutet, dass die zugehörigen Kovariablen nicht signifikant sind.

Likelihood-Quotienten-Statistik

Die Likelihood-Quotienten-Statistik

$$\lambda = 2\left(\ell(\widehat{\theta}) - \ell(\widetilde{\theta})\right) = 2\log\left[\frac{L(\widehat{\theta})}{L(\widetilde{\theta})}\right]$$

vergleicht das unrestringierte Maximum der Log-Likelihood $\ell(\widehat{\theta})$ (über Θ) mit dem Maximum der Log-Likelihood unter der H_0-Restriktion, d.h. $\widetilde{\theta}$ maximiert $\ell(\theta)$ unter der Nebenbedingung $\boldsymbol{C}\theta = \boldsymbol{d}$. Die Struktur eines zugehörigen Tests lautet:

$$\lambda \text{ zu groß} \Rightarrow H_0 \text{ ablehnen.}$$

Nachteil: Es ist eine numerische Maximierung von $\ell(\theta)$ unter linearer Nebenbedingung notwendig, um $\widetilde{\theta}$ zu erhalten.

Wald-Statistik

Die Wald-Statistik

$$w = (\boldsymbol{C}\widehat{\theta} - \boldsymbol{d})^\top (\boldsymbol{C}\mathbf{I}^{-1}(\widehat{\theta})\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\widehat{\theta} - \boldsymbol{d})$$

misst die (gewichtete) Distanz zwischen der unrestringierten Schätzung $\boldsymbol{C}\widehat{\theta}$ von $\boldsymbol{C}\theta$ und dem hypothetischen Wert $\boldsymbol{d}$ unter H_0. Ein Test wird so konstruiert, dass

$$w \text{ zu groß} \Rightarrow H_0 \text{ ablehnen.}$$

Vorteil gegenüber λ: Keine Berechnung von $\widetilde{\theta}$ nötig.

Score- (oder Rao-) Statistik

Die Score-Statistik lautet

$$u = s(\widetilde{\theta})^\top \mathbf{I}^{-1}(\widetilde{\theta}) s(\widetilde{\theta}).$$

Idee: Für $\widehat{\theta}$ gilt $s(\widehat{\theta}) = 0$. Falls H_1 richtig ist, wird $s(\widetilde{\theta})$ deutlich von $0 = s(\widehat{\theta})$ verschieden sein, d.h.

$$u \text{ wird groß} \Rightarrow H_0 \text{ ablehnen.}$$

Die Statistik berechnet also den Abstand $s(\widetilde{\theta})$ vom Ursprung, gewichtet mit $\mathbf{I}^{-1}(\widetilde{\theta})$.

Beispiel 3.8 (Test für einen Subvektor)**.** *Betrachte*

- $H_1 : \ \eta = \boldsymbol{x}^\top \boldsymbol{\beta}$ *Prädiktor in vollem GLM,*
- $H_0 : \ \eta_s = \boldsymbol{x}_s^\top \boldsymbol{\beta}_s$ *Prädiktor in reduziertem GLM (nach Weglassen von Kovariablen).*

Die Log-Likelihood $\ell(\boldsymbol{\beta}_s)$ im reduzierten Submodell werde durch $\widehat{\boldsymbol{\beta}}_s$ maximiert. Mit $\widehat{\boldsymbol{\beta}}_s$ und $\widehat{\boldsymbol{\beta}}$ lässt sich die Likelihood-Quotienten-Statistik bestimmen. Für die Wald-Statistik ergibt sich

$$\boldsymbol{w} = (\widehat{\boldsymbol{\beta}})_s^\top \widehat{\boldsymbol{A}}_s^{-1} (\widehat{\boldsymbol{\beta}})_s,$$

dabei bezeichne $(\widehat{\boldsymbol{\beta}})_s$ die Elemente des Subvektors $\boldsymbol{\beta}_s$ in $\widehat{\boldsymbol{\beta}}$ und $\widehat{\boldsymbol{A}}_s$ sei die Teilmatrix von $\widehat{\boldsymbol{A}} = \mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}})$, die diesen Elementen entspricht.

Satz 3.5. *Unter H_0 und den gleichen Regularitätsannahmen wie in Satz 3.4 gilt:*

$$\lambda, \ w, \ u \ \overset{a}{\sim} \ \chi^2(s).$$

D.h. man lehnt H_0 ab, falls $\lambda, w, u > \chi^2_{1-\alpha}(s)$ ist. Für finite Stichproben besitzen λ, w, u aber unterschiedliche Werte; im Zweifelsfall sollte man λ bevorzugen.

Beweis.

- *Beweis für w:* Es gilt

  $$\widehat{\theta} \ \overset{a}{\sim} N(\theta, \mathbf{I}^{-1}(\widehat{\theta}))$$

  und damit

  $$\boldsymbol{C}\widehat{\theta} \ \overset{a}{\sim} N(\boldsymbol{C}\theta, \boldsymbol{C}\mathbf{I}^{-1}(\widehat{\theta})\boldsymbol{C}^\top).$$

  Unter H_0 folgt

  $$\boldsymbol{C}\widehat{\theta} - \underbrace{\boldsymbol{C}\theta}_{\boldsymbol{d}} \ \overset{a}{\sim} \ N(\boldsymbol{0}, \underbrace{\boldsymbol{C}\mathbf{I}^{-1}(\widehat{\theta})\boldsymbol{C}^\top}_{\boldsymbol{A}}),$$

  also

  $$\boldsymbol{A}^{-1/2}(\boldsymbol{C}\widehat{\theta} - \boldsymbol{d}) \ \overset{a}{\sim} \ N(0, I)$$

  und somit

  $$\boldsymbol{w} = (\boldsymbol{C}\widehat{\theta} - \boldsymbol{d})^T \boldsymbol{A}^{-1} (\boldsymbol{C}\widehat{\theta} - \boldsymbol{d}) \ \overset{a}{\sim} \ \chi^2(s).$$

- *Beweis für λ:* Durch Taylorentwicklung kann gezeigt werden, dass $w \overset{a}{\sim} \lambda$ und somit $\lambda \overset{a}{\sim} \chi^2(s)$. Die Beweisskizze wird hier lediglich für den Spezialfall

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$$

geführt (das entspricht $\boldsymbol{C} = I, \boldsymbol{d} = \theta_0$, $\text{rang}(\boldsymbol{C}) = p = \dim(\theta)$). Eine Taylorentwicklung 2. Ordnung von $\ell(\theta_0)$ um den unrestringierten Maximum-Likelihood-Schätzer $\widehat{\theta}$ liefert

$$\ell(\theta_0) \approx \ell(\widehat{\theta}) + s(\widehat{\theta})^\top(\theta_0 - \widehat{\theta}) - \frac{1}{2}(\theta_0 - \widehat{\theta})^\top \mathbf{J}(\widehat{\theta})(\theta_0 - \widehat{\theta}),$$

also wegen $s(\widehat{\theta}) = 0$

$$\lambda = 2\left(\ell(\widehat{\theta}) - \ell(\theta_0)\right) \approx (\widehat{\theta} - \theta_0)^\top \mathbf{J}(\widehat{\theta})(\widehat{\theta} - \theta_0) \approx (\widehat{\theta} - \theta_0)^\top \mathbf{I}(\widehat{\theta})(\widehat{\theta} - \theta_0) \overset{a}{\sim} \chi^2(p).$$

- *Beweis für u:* Wir nehmen denselben Speziallfall wie im Beweis für λ an, also $\widetilde{\theta} = \theta_0$. Es ist

$$s(\theta_0) \overset{a}{\sim} N(\mathbf{0}, \boldsymbol{I}(\theta_0))$$

bzw.

$$\boldsymbol{I}^{-1/2}(\theta_0) s(\theta_0) \overset{a}{\sim} N(\mathbf{0}, I),$$

also

$$s(\theta_0)^\top \underbrace{\mathbf{I}^{-\top/2}(\theta_0)\mathbf{I}^{-1/2}(\theta_0)}_{\mathbf{I}(\theta_0)^{-1}} s(\theta_0) \overset{a}{\sim} \chi^2(p).$$

□

3.3.2 Konfidenzintervalle

- Gemeinsamer Konfidenzbereich:

$$(\widehat{\theta} - \theta)^\top \mathbf{I}(\widehat{\theta})(\widehat{\theta} - \theta) \overset{a}{\sim} \chi^2(p)$$

$$\Rightarrow \mathbb{P}_\theta\left((\widehat{\theta} - \theta)^\top \mathbf{I}(\widehat{\theta})(\widehat{\theta} - \theta) \leq \chi^2_{1-\alpha}(p)\right) \overset{a}{\approx} 1 - \alpha.$$

Daraus lässt sich ein $(1-\alpha)$-Konfidenz-Ellipsoid konstruieren.

- Komponentenweise Konfidenzintervalle für θ_j, $j = 1, \ldots, p$:

$$\frac{\widehat{\theta}_j - \theta_j}{\widehat{\sigma}_j} \overset{a}{\sim} N(0,1),$$

wobei $\widehat{\sigma}_j^2$ das j-te Diagonalelement von $\widehat{\text{Cov}}(\widehat{\theta}) = \mathbf{I}^{-1}(\widehat{\theta})$ ist. Das zugehörige approximative $(1-\alpha)$-Konfidenzintervall lautet:

$$\widehat{\theta}_j \pm z_{1-\alpha/2}\, \widehat{\sigma}_j.$$

3.3.3 Modellwahl

Zum Vergleich verschiedener Modelle existieren Modellwahlkriterien, die die Güte der Anpassung, gemessen durch $\ell(\hat{\theta})$, und die Modellkomplexität $p = \dim(\theta)$ bewerten, indem sie die beiden Größen durch eine Straffunktion $\text{pen}(p)$ in einem Kompromiss zu

$$-\ell(\hat{\theta}) + \text{pen}(p)$$

zusammenführen. Dabei wird $-\ell(\hat{\theta})$ klein bei guter Anpassung, $\text{pen}(p)$ groß bei stark bzw. überparametrisierten Modellen. Am bekanntesten ist *Akaikes Informationskriterium*

$$\text{AIC} = -2\ell(\hat{\theta}) + 2p$$

mit $\text{pen}(p) = 2p$.

Motivation: $\{f_\theta(x) = f(x|\theta), \theta \in \Theta\}$ parametrisiere die betrachteten Modelle und $g(x)$ sei die wahre Dichte für X. Ziel: Minimiere die Kullback-Leibler-Distanz

$$D(g, f_\theta) = \mathbb{E}_X \left(\log \frac{g(X)}{f(X|\theta)} \right) \geq 0,$$

bzw. $\mathbb{E}_Z[K(f_{\hat{\theta}(Z)}, g)] = \mathbb{E}_Z \mathbb{E}_X[\log g(X) - \log f_{\hat{\theta}(Z)}(X)]$, wenn θ aus gegebenen Daten $Z = z$ geschätzt wird, $X, Z \overset{i.i.d}{\sim} g$. Die Akaike Information (ohne Konstanten) $\mathbb{E}_Z \mathbb{E}_X[-\log f_{\hat{\theta}(Z)}(X)]$ ist ein prädiktives Maß für zwei unabhängige Realisationen x und z aus g. Zur Schätzung liegt die maximierte Loglikelihood $-\log f_{\hat{\theta}(Z)}(Z)$ vor, die jedoch nicht erwartungstreu ist, sondern durch die doppelte Verwendung von Z „überoptimistisch“ bzgl. der Anpassung des Modells. Unter den Regularitätsbedingungen von Satz 3.4 lässt sich zeigen, dass der Bias genau durch $2p$ ausgeglichen wird.

Eine Alternative ist zum Beispiel das *Schwartz- (Bayes-) Informationskriterium*

$$\text{BIC} = -2\ell(\hat{\theta}) + p \log n$$

wobei n die Größe des Datensatzes ist. Für $n \geq 8$ „bestraft“ das BIC die Modellkomplexität stärker als das AIC.

Es lässt sich zeigen, dass die Modellwahl basierend auf dem BIC asymptotisch äquivalent ist zur Modellwahl basierend auf sogenannten Bayes-Faktoren, siehe Held, Kapitel 7.2, für eine Herleitung. Die Bayes-Faktoren vergleichen die Posteriori-Modellwahrscheinlichkeiten mit den Priori-Modellwahrscheinlichkeiten.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

Bisher haben wir volle (*genuine*) Likelihood-Inferenz betrieben: Gegeben war ein parametrisches statistisches Modell, das heißt eine Familie von Verteilungen oder Dichten mit Parameter $\theta \in \Theta$.

Bisherige Grundannahme: Es existiert ein „wahres” $\theta_0 \in \Theta$ derart, dass $\mathbb{P}_{\theta_0}$ die Verteilung des datengenerierenden Prozesses $\mathbb{P}_0$ ist, das heißt $\mathbb{P}_{\theta_0} = \mathbb{P}_0$ gilt.