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* $\sigma^2$ *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* $\sigma^2$*. 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*

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

*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.*