4.6 Bayesianisches lineares Modell

Modell:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ ,$$

wobei $\boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{X} \in \mathbb{R}^{n \times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\varepsilon} \in \mathbb{R}^n$

Annahmen und Notation:

$$\begin{aligned} p &= \text{rang}(\boldsymbol{X}) \\ \boldsymbol{\varepsilon} &= (\varepsilon_1, \ldots, \varepsilon_n)^\top, \quad \varepsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2) \end{aligned}$$

4.6 Bayesianisches lineares Modell

Bayesianisch:

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X} \sim \text{MVN}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$$

Likelihood:

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) &\propto |\sigma^2 \boldsymbol{I}_n|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\sigma^2 \boldsymbol{I}_n)^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \\ &= (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \end{aligned}$$

4.6 Bayesianisches lineares Modell

4.6.1 Nichtinformative Prioriverteilung

Die **nichtinformative Priori**

$$p(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-1}$$

ist insbesondere im Fall $p \ll n$ nützlich.
Für die **gemeinsame Posteriori** folgt:

$$p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) \propto (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right).$$

Sei

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \\
\widehat{\boldsymbol{y}} &= \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top}_{\boldsymbol{H}} \boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}, \\
\widehat{\boldsymbol{\varepsilon}} &= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{y} - \widehat{\boldsymbol{y}}.
\end{aligned}$$

4.6 Bayesianisches lineares Modell

4.6.1 Nichtinformative Prioriverteilung

Aus der Theorie linearer Modelle ist bekannt, dass

$$\boldsymbol{X}^\top \widehat{\varepsilon} = 0, \quad \widehat{\boldsymbol{y}}^\top \widehat{\varepsilon} = 0.$$

Daraus ergeben sich folgende Umformungen:

$$\begin{aligned}
(\boldsymbol{y} - \boldsymbol{X}\beta)^\top (\boldsymbol{y} - \boldsymbol{X}\beta) &= [(\boldsymbol{y} - \widehat{\boldsymbol{y}}) + (\widehat{\boldsymbol{y}} - \boldsymbol{X}\beta)]^\top [(\boldsymbol{y} - \widehat{\boldsymbol{y}}) + (\widehat{\boldsymbol{y}} - \boldsymbol{X}\beta)] \\
&= \widehat{\varepsilon}^\top \widehat{\varepsilon} + (\widehat{\boldsymbol{y}} - \boldsymbol{X}\beta)^\top (\widehat{\boldsymbol{y}} - \boldsymbol{X}\beta) + 2(\widehat{\boldsymbol{y}} - \boldsymbol{X}\beta)^\top \widehat{\varepsilon} \\
&= \widehat{\varepsilon}^\top \widehat{\varepsilon} + (\widehat{\boldsymbol{y}} - \boldsymbol{X}\beta)^\top (\widehat{\boldsymbol{y}} - \boldsymbol{X}\beta) \\
&\overset{\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\beta}}{=} \widehat{\varepsilon}^\top \widehat{\varepsilon} + (\widehat{\beta} - \beta)^\top \boldsymbol{X}^\top \boldsymbol{X} (\widehat{\beta} - \beta),
\end{aligned}$$

so dass sich die Posteriori schreiben lässt als

$$p(\beta, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) \propto (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2} \left(\widehat{\varepsilon}^\top \widehat{\varepsilon} + (\widehat{\beta} - \beta)^\top \boldsymbol{X}^\top \boldsymbol{X} (\widehat{\beta} - \beta) \right) \right.$$

4.6 Bayesianisches lineares Modell

4.6.1 Nichtinformative Prioriverteilung

Die **bedingte Posteriori** von $\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}$ ist

$$p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}) \propto \exp\left(-\frac{1}{2\sigma^2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right),$$

da $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ nicht von $\boldsymbol{\beta}$ abhängt.
Man identifiziert obigen Ausdruck als Kern einer multivariaten Normalverteilung, genauer ist

$$p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}) \sim \text{MVN}(\widehat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}),$$

ein bekanntes Resultat aus der Theorie linearer Modelle.

4.6 Bayesianisches lineares Modell

4.6.1 Nichtinformative Prioriverteilung

Die **marginale Posteriori** von σ^2 erhält man über Herausintegrieren von $\boldsymbol{\beta}$ bzw. einfacher über den Satz von Bayes

$$f(\sigma^2|\boldsymbol{y},\boldsymbol{X}) = \frac{f(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X})}{f(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X})} \ .$$

Die Normalisierungskonstante für die bedingte Posteriori von $\boldsymbol{\beta}$ ist $\sigma^{-p/2}$, also

$$\begin{aligned} f(\sigma^2|\boldsymbol{y},\boldsymbol{X}) &\propto (\sigma^2)^{-(\frac{n}{2}+1)}\sigma^{\frac{p}{2}} \exp\left(-\frac{1}{2\sigma^2}\,\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}\right) \\ &= (\sigma^2)^{-(\frac{n-p}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}\,\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}\right). \end{aligned}$$

Dies ist der Kern einer inv-$\chi^2\left(n-p, \frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}}{n-p}\right)$ bzw. IG$\left(\frac{n-p}{2}, \frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}}{2}\right)$-Verteilung. Es gilt:

$$\mathbb{E}[\sigma^2|\boldsymbol{y},\boldsymbol{X}] = \frac{n-p}{n-p-2}\cdot\frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}}{n-p} = \frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}}{n-p-2} \ .$$

4.6 Bayesianisches lineares Modell

4.6.1 Nichtinformative Prioriverteilung

Algorithmus 9 : Direkte Simulation von $\boldsymbol{\beta}$ und $\boldsymbol{\sigma^2}$ im bayesianischen linearen Modell

---

Für $t = 1, \ldots, T$:

1. Ziehe $\left(\sigma^2\right)^{(t)}$ aus $f\left(\sigma^2 | \boldsymbol{y}, \boldsymbol{X}\right)$, d.h. aus inv-$\chi^2\left(n-p, \frac{\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}}}{n-p}\right)$.
2. Ziehe $\boldsymbol{\beta}^{(t)}$ aus $f\left(\boldsymbol{\beta} | \left(\sigma^2\right)^{(t)}, \boldsymbol{y}, \boldsymbol{X}\right)$, d.h. aus MVN $\left(\widehat{\boldsymbol{\beta}}, \left(\sigma^2\right)^{(t)} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right)$, wobei $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

4.6 Bayesianisches lineares Modell

4.6.1 Nichtinformative Prioriverteilung

Eine Alternative zur direkten Simulation besteht in der Verwendung von Gibbs-Sampling, indem zusätzlich zur vollständig bedingten Dichte von $\boldsymbol{\beta}$ die vollständig bedingte Dichte von σ^2,

$$
\begin{aligned}
f(\sigma^2|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X}) &\propto p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) \\
&\propto (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}\left(\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}} + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)\right),
\end{aligned}
$$

zur Simulation verwendet wird.
Dies ist für festes $\boldsymbol{\beta}$ der Kern einer skalierten inv-$\chi^2\left(n, \frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{n}\right)$-Verteilung.

4.6 Bayesianisches lineares Modell

4.6.1 Nichtinformative Prioriverteilung

Damit lässt sich auch die **marginale Posteriori** von $\boldsymbol{\beta}$ herleiten:

$$f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \frac{f(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X})}{f(\sigma^2|\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta})} = \frac{f(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X})\cdot f(\sigma^2|\boldsymbol{y},\boldsymbol{X})}{f(\sigma^2|\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X})}$$

$$\propto \frac{\exp\left(-\frac{1}{2\sigma^2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)}{\left[\frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{n}\right]^{n/2} \exp\left(-\frac{1}{2\sigma^2}[\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})]\right)}$$

Damit:

$$f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) \propto \left[\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right]^{-n/2}.$$

4.6 Bayesianisches lineares Modell

4.6.1 Nichtinformative Prioriverteilung

Setzt man

$$\widehat{\sigma}_\varepsilon^2 = \frac{\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}}}{n-p} \Leftrightarrow \widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}} = (n-p)\widehat{\sigma}_\varepsilon^2,$$

so ist

$$\begin{aligned} f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) &\propto \left[(n-p)\widehat{\sigma}_\varepsilon^2 + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right]^{-n/2} \\ &= \left((n-p)\widehat{\sigma}_\varepsilon^2 \cdot \left[1 + \frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(n-p)\widehat{\sigma}_\varepsilon^2}\right]\right)^{-\frac{n}{2}} \\ &\propto \left[1 + \frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(n-p)\widehat{\sigma}_\varepsilon^2}\right]^{-\frac{(n-p)+p}{2}}. \end{aligned}$$

4.6 Bayesianisches lineares Modell

4.6.1 Nichtinformative Prioriverteilung

Dies entspricht dem Kern einer multivariaten t-Verteilung mit $n - p$ Freiheitsgraden, Lokationsparameter $\widehat{\boldsymbol{\beta}}$ und Skalenparameter $\sigma_\varepsilon^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, also

$$\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X} \sim \text{mv-t}_{n-p}\left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}_\varepsilon^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right).$$

Abschließend betrachten wir noch die **prädiktive Verteilung** für $\widetilde{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}$. Seien

- m die Anzahl neuer Beobachtungen,
- $\widetilde{\boldsymbol{X}}$ neue Beobachtungen von Regressoren der Dimension $m \times p$,
- $\widetilde{\boldsymbol{y}}$ der Vektor der Prognosen der Dimension $m \times 1$.

4.6 Bayesianisches lineares Modell

4.6.1 Nichtinformative Prioriverteilung

Zur Simulation können wir Algorithmus 9 wie folgt erweitern:

Algorithmus 10 : Direkte Simulation der prädiktiven Verteilung im bayesianischen linearen Modell

---

Für $t = 1, \ldots, T$:

1. Ziehe $\left(\sigma^2\right)^{(t)}$ aus $f\left(\sigma^2 | \boldsymbol{y}, \boldsymbol{X}\right)$, d.h. aus inv-$\chi^2\left(n-p, \frac{\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}}}{n-p}\right)$.
2. Ziehe $\boldsymbol{\beta}^{(t)}$ aus $f\left(\boldsymbol{\beta} | \left(\sigma^2\right)^{(t)}, \boldsymbol{y}, \boldsymbol{X}\right)$, d.h. aus MVN $\left(\widehat{\boldsymbol{\beta}}, \left(\sigma^2\right)^{(t)} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right)$, wobei $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.
3. Für $i = 1, \ldots, m$:

   Ziehe $\widetilde{\boldsymbol{y}}_i \sim MVN\left(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}^{(t)}, \left(\sigma^2\right)^{(t)}\right)$, wobei $\widetilde{\boldsymbol{x}}_i^\top$ die i-te Zeile von $\widetilde{\boldsymbol{X}}$ bezeichnet.

4.6 Bayesianisches lineares Modell

4.6.1 Nichtinformative Prioriverteilung

Es ist sogar eine analytische Berechnung möglich:

$$f(\widetilde{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}) \sim \text{mv-t}_{n-p}\left[\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}, \hat{\sigma}_{\varepsilon}^2\left(\widetilde{\boldsymbol{X}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\widetilde{\boldsymbol{X}}^\top + \boldsymbol{I}_n\right)\right]$$

in Analogie zur Berechnung von Prognose und Prognoseintervallen für lineare Modelle aus frequentistischer Sicht.

4.6 Bayesianisches lineares Modell

4.6.2 Konjugierte Prioriverteilung

Im Falle der **konjugierten Priori**

$$\sigma^2 \sim \text{inv-}\chi^2(\kappa_0, \sigma_0^2) \,, \quad \boldsymbol{\beta}|\sigma^2 \sim \text{MVN}(\boldsymbol{\beta}_0, \sigma^2 \boldsymbol{\Sigma}_0)$$

bzw.

$$\boldsymbol{\beta}, \sigma^2 \sim \text{MVN-inv-}\chi^2(\boldsymbol{\beta}_0, \sigma_0^2 \boldsymbol{\Sigma}_0; \kappa_0, \sigma_0^2)$$

ergibt sich die **gemeinsame Posteriori**

$$\sigma^2|\boldsymbol{y}, \boldsymbol{X} \sim \text{inv-}\chi^2(\kappa_n, \sigma_n^2) \,, \quad \boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X} \sim \text{MVN}(\boldsymbol{\beta}_n, \sigma^2 \boldsymbol{\Sigma}_n)$$

bzw.

$$\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X} \sim \text{MVN-inv-}\chi^2(\boldsymbol{\beta}_n, \sigma_n^2 \boldsymbol{\Sigma}_n; \kappa_n, \sigma_n^2) \,,$$

wobei

$$\begin{aligned}
\boldsymbol{\beta}_n &= (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{X})^{-1}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^\top \boldsymbol{y}) \ , \\
\boldsymbol{\Sigma}_n &= (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{X})^{-1} \ , \\
\kappa_n &= \kappa_0 + n \ , \\
\sigma_n^2 &= (\boldsymbol{\beta}_0^\top \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 - \boldsymbol{\beta}_n^\top \boldsymbol{\Sigma}_n^{-1}\boldsymbol{\beta}_n + \boldsymbol{y}^\top \boldsymbol{y} + \kappa_0\sigma_0^2)/(\kappa_0 + n) \ .
\end{aligned}$$

Als **bedingte Posteriori** von $\boldsymbol{\beta}$ ergibt sich

$$\boldsymbol{\beta} \,|\sigma^2, \boldsymbol{y}, \boldsymbol{X} \quad \sim \quad \text{MVN}(\boldsymbol{\beta}_n, \sigma^2\boldsymbol{\Sigma}_n),$$

als **marginale Posteriori** von σ^2

$$\sigma^2|\boldsymbol{y}, \boldsymbol{X} \quad \sim \quad \text{inv-}\chi^2(\kappa_n, \sigma_n^2).$$

1. **Ridge-Regression**
   Hinweis: Es ist im Allgemeinen sinnvoll, die „echten" Kovariablen (ohne Intercept) zu standardisieren, um Unterschiede in der Skala zu beseitigen. Ferner geht man zum zentrierten Response über, so dass der Intercept entfällt. Man erhält $\boldsymbol{X}^*, \boldsymbol{y}^*$. Betrachte nun

$$\boldsymbol{y}^* = \boldsymbol{X}^* \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \ \boldsymbol{\varepsilon} \ \sim \ N(0, \sigma^2 \boldsymbol{I}_n).$$

   Der Ridge-Schätzer ist durch

$$[(\boldsymbol{X}^*)^\top \boldsymbol{X}^* + \lambda \boldsymbol{I}_n]^{-1} (\boldsymbol{X}^*)^\top \boldsymbol{y}^*$$

   mit $\lambda > 0$ gegeben.

   ...

1. ...

   Dieser lässt sich wie folgt bayesianisch interpretieren: Sei

$$p(\boldsymbol{\beta}^*) \sim N(0, \tau^2 \boldsymbol{I}_n),$$

   d.h. die Komponenten von $\boldsymbol{\beta}^*$ sind a priori unkorreliert (also wegen der Normalverteilung auch unabhängig).
   Dann ist die bedingte Posteriori $f\left(\boldsymbol{\beta}^* | \boldsymbol{y}^*, \boldsymbol{X}^*, \sigma^2, \tau^2\right)$ gleich

$$\text{MVN}\left(\left[(\boldsymbol{X}^*)^\top \boldsymbol{X}^* + \frac{\sigma^2}{\tau^2}\boldsymbol{I}_n\right]^{-1} (\boldsymbol{X}^*)^\top \boldsymbol{y}^*, \sigma^2 \left((\boldsymbol{X}^*)^\top \boldsymbol{X}^* + \frac{\sigma^2}{\tau^2}\boldsymbol{I}_n\right)^{-1}\right).$$

   Der Parameter λ enstpricht dabei dem Quotienten σ^2/τ^2.

2. **Ungleiche Varianzen der Störvariablen / abhängige Störvariablen**

   Allgemein:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \ \sim \ N(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$$

$$\boldsymbol{y} \ \sim \ \text{MVN}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{y}}),\ \boldsymbol{\Sigma}_{\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$$

   Problem: Spezifikation der Prioriverteilung für $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$.

Mögliche Auswege sind:

(a) Parametrisiere

$$\boldsymbol{\Sigma}_{\boldsymbol{y}} = \sigma^2 \boldsymbol{Q}_{\boldsymbol{y}}$$

mit $\boldsymbol{Q}_{\boldsymbol{y}}$ bekannt und $p(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-1}$.
Dieser Fall ist auf das Modell aus Abschnitt 4.6.1 reduzierbar, indem man das Modell

$$\underbrace{\boldsymbol{Q}^{-1/2}\boldsymbol{y}}_{\boldsymbol{y}^*} = \underbrace{\boldsymbol{Q}^{-1/2}\boldsymbol{X}}_{\boldsymbol{X}^*}\boldsymbol{\beta} + \underbrace{\boldsymbol{Q}^{-1/2}\boldsymbol{\varepsilon}}_{\boldsymbol{\varepsilon}^*}$$

betrachtet. Man erhält dann wieder ein homoskedastisches Regressionsmodell in den Größen $\boldsymbol{y}^*, \boldsymbol{X}^*, \boldsymbol{\varepsilon}^*$.

(b) Gewichtete Regression:

$$\boldsymbol{\Sigma}_{\boldsymbol{y}} = \text{diag}(\sigma^2 w_i^{-1})_{1 \leq i \leq n}$$

lässt sich als Spezialfall von (a) auffassen.

(c) Korrelationen: Schreibe

$$\boldsymbol{\Sigma_y} = \boldsymbol{SRS} \text{ mit } \boldsymbol{S} = \text{diag}(\sigma_1, \ldots, \sigma_p)$$

mit beispielweise

$$p(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2) = \prod_{j=1}^{p} p(\sigma_j^2) \quad \text{und} \quad p(\sigma_j^2) \sim \text{inv-}\chi^2(\nu_j, \sigma_{0j}^2).$$

Es bleibt die Spezifikation der Korrelationsmatrix.
Priori-Spezifikationen müssen insbesondere positive Definitheit gewährleisten. Eine einfache Variante besteht in der Verwendung von (positiver) „Äqui-Korrelation", was zum Beispiel bei Clusterdaten eine vernünftige Annahme darstellt:

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix},$$

4.6 Bayesianisches lineares Modell

4.6.3 Spezialfälle und Erweiterungen

(c) ...

wobei $\rho \sim \mathit{Unif}[0;1]$ eine positive Korrelation erzwingt. Bei Messwiederholungen greift man oft auf eine autoregressive Kovarianzstruktur zurück. Für in 1. Ordnung autokorrelierte Residuen

$$\varepsilon_t = \rho\varepsilon_{t-1} + Z_t,\ Z_t \sim N(0, \sigma^2),$$

erhält man

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \cdots & 1 \end{bmatrix}.$$

Andere Zerlegungen basieren auf der Cholesky- oder Spektralzerlegung und sind relativ komplex.

(d) Übergang zu Modellen mit Zufallseffekten: Modelle mit Zufallseffekten (linear mixed models, generalized linear mixed models) führen zu strukturierten, meist parametersparsamen Kovarianzmatrizen. *Aber*: Die Modellgleichung ändert sich und $\boldsymbol{\Sigma}_{\boldsymbol{y}} \neq \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, d.h. man kommt in eine andere Modellklasse.

4.7 Bayesianisches generalisiertes lineares Modell

Struktur von GLMs: Der Response folgt einer Verteilung aus einer einfachen Exponentialfamilie ($i = 1, \ldots, n$)

$$f(y_i|\theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i}\right) \cdot c(y_i, \phi_i) \tag{4.1}$$

oder

$$f(y_i|\theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)}\right) \cdot c(y_i, \phi_i),$$

wobei in vielen Fällen $\phi_i \equiv \phi$ (Bernoulli-, Poissonverteilung). Es ist

$$\mu_i = \mathbb{E}[y_i|\theta_i] = b'(\theta_i) \ , \quad \mathrm{Var}(y_i|\theta_i) = b''(\theta_i)\phi_i$$

und θ_i der kanonische Parameter.
Mit einer Linkfunktion g bzw. Responsefunktion $h = g^{-1}$ gelte

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}. \tag{4.2}$$

4.7 Bayesianisches generalisiertes lineares Modell

Beispiel 4.3 (Logit-Modell)
Mit $\mu_i = \mathbb{P}(y_i = 1)$ ist

$$\begin{aligned} f(y_i|\mu_i) &= \mu_i^{y_i}(1-\mu_i)^{1-y_i} \\ &= \exp\big(y_i \log(\mu_i) + (1-y_i)\log(1-\mu_i)\big) \\ &= \exp\Bigg(y_i \underbrace{\log\left(\frac{\mu_i}{1-\mu_i}\right)}_{\theta_i} + \log(1-\mu_i)\Bigg) \end{aligned}$$

mit

$$\theta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right) \;\Leftrightarrow\; \mu_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)} \,.$$

...

...

Dies entspricht (4.1) mit $\phi_i = 1,\ c(y_i, \phi_i) = 1,$

$$\begin{aligned} b(\theta_i) &= -\log\left(1 - \frac{\exp(\theta_i)}{1 + \exp(\theta_i)}\right) \\ &= -\log\left(\frac{1}{1 + \exp(\theta_i)}\right) \\ &= \log\left(1 + \exp(\theta_i)\right) \end{aligned}$$

und

$$b'(\theta_i) = \frac{1}{1 + \exp(\theta_i)} \cdot \exp(\theta_i) = \mu_i.$$

4.7 Bayesianisches generalisiertes lineares Modell

Als Prioriverteilung für $\boldsymbol{\beta}$ in (4.2) eignet sich

$$\boldsymbol{\beta} \sim \text{MVN}(\boldsymbol{\beta}_0, \boldsymbol{B}_0)$$

mit $\boldsymbol{B}_0 > 0$ (vgl. Abschnitt 4.5.3 zur multivariaten Normalverteilung bei bekannter Kovarianzmatrix).
$\boldsymbol{\beta}$ beeinflusst $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ über den Prädiktor $\boldsymbol{\mu}(\boldsymbol{\beta}) = \boldsymbol{h}(\boldsymbol{X}\boldsymbol{\beta})$, wobei die Auswertung komponentenweise zu verstehen ist. Für $\boldsymbol{B}_0^{-1} \to \boldsymbol{0}$ erhält man eine nichtinformative Priori.

4.7 Bayesianisches generalisiertes lineares Modell

Über die Darstellung (4.1) als Exponentialfamile mit kanonischen Parametern erhält man als Posterioriverteilung

$$
\begin{aligned}
f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right)\cdot\prod_{i=1}^{n}\exp\left(\frac{y_i\theta_i-b(\theta_i)}{\phi_i}\right)\\
&= \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right)\exp\left(\sum_{i=1}^{n}\frac{y_i\theta_i-b(\theta_i)}{\phi_i}\right)\\
&= \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+\sum_{i=1}^{n}\frac{y_i\theta_i-b(\theta_i)}{\phi_i}\right).
\end{aligned}
$$

4.7 Bayesianisches generalisiertes lineares Modell

Beispiel 4.4 (Logit-Modell)

Im Falle des Logit-Modells erhält man als Posteriori

$$
\begin{aligned}
f(\beta|\boldsymbol{y},\boldsymbol{X}) &\propto \exp\left(-\frac{1}{2}(\beta-\beta_0)^\top \boldsymbol{B}_0^{-1}(\beta-\beta_0)\right) \\
&\quad \times \prod_{i=1}^{n} \mu_i(\beta)^{y_i}(1-\mu_i(\beta))^{1-y_i} \\
&= \exp\left(-\frac{1}{2}(\beta-\beta_0)^\top \boldsymbol{B}_0^{-1}(\beta-\beta_0)\right) \\
&\quad \times \prod_{i=1}^{n} h(\boldsymbol{x}_i^\top\beta)^{y_i}(1-h(\boldsymbol{x}_i^\top\beta))^{1-y_i} \\
&= \exp\left(-\frac{1}{2}(\beta-\beta_0)^\top \boldsymbol{B}_0^{-1}(\beta-\beta_0)\right) \\
&\quad \times \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{x}_i^\top\beta)}{1+\exp(\boldsymbol{x}_i^\top\beta)}\right)^{y_i} \left(\frac{1}{1+\exp(\boldsymbol{x}_i^\top\beta)}\right)^{1-y_i}.
\end{aligned}
$$

4.7 Bayesianisches generalisiertes lineares Modell

Problem: Der Posteriori-Kern entspricht keinem Kern einer bekannten Verteilung. Die Posteriori-Verteilung ist demnach nicht analytisch zugänglich.
Mögliche Auswege sind

1. Approximation oder
2. Exploration der Posteriori durch Generierung von Samples aus der Posteriori.

Wir betrachten im Folgenden Lösung 2. Hier gibt es mehrere Möglichkeiten; sehr etabliert ist ein Vorschlag von Gamerman (1997) , eine Variante des *Metropolis-Hastings-Algorithmus*.

4.7 Bayesianisches generalisiertes lineares Modell

4.7.1 Ein MCMC-Algorithmus: Metropolis-Hastings

Zunächst folgt eine Darstellung des Grundproblems, ohne näher auf die zugrundeliegende mathematische Theorie einzugehen. Bekannt ist, dass für $X_i \overset{i.i.d}{\sim} \pi$, $i = 1, \ldots, n$, wobei π eine Verteilung bezeichnet, interessierende Kennzeichen dieser Verteilung (Momente, Dichte etc.) — Existenz vorausgesetzt — durch Simulation von Zufallszahlen gemäß π als Monte-Carlo-Schätzung gewonnen werden können, z.B.

$$\widehat{\mathbb{E}[X]} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

4.7 Bayesianisches generalisiertes lineares Modell

4.7.1 Ein MCMC-Algorithmus: Metropolis-Hastings

Dies ist wenig „spannend", da, wenn π bekannt ist, in der Regel auch der Erwartungswert zugänglich ist. Angenommen jedoch, man betrachtet eine (nichtlineare) Funktion von X, zum Beispiel $g(X) = X^2$. Dann ist möglicherweise die Dichte der Transformation $g(X)$ noch analytisch bestimmbar, aber der Erwartungswert komplex zu berechnen.
Im Fall, dass X mehrdimensional ist, kann die analytische Bestimmung derartiger Kenngrößen analytisch unmöglich und bei höherer Dimension mittels numerischer Integration zu instabil sein.

4.7 Bayesianisches generalisiertes lineares Modell

4.7.1 Ein MCMC-Algorithmus: Metropolis-Hastings

Unter geeigneten Voraussetzungen lässt sich obige Monte-Carlo Schätzung erweitern zu

$$\widehat{\mathbb{E}[g(X)]} = \frac{1}{n} \sum_{i=1}^{n} g(x_i).$$

(Dies ist ein allgemeines Prinzip, also nicht notwendigerweise bayesianisch, solange es sich bei π nicht zum Beispiel um eine Posterioriverteilung handelt.)

Es sei allerdings bemerkt, dass dieses Vorgehen im Vergleich zur exakten Lösung mit einem Monte-Carlo Fehler behaftet ist.
Wesentliche Voraussetzung ist zudem, dass Zufallszahlen aus π gezogen werden können.
Verfahren zur Generierung von i.i.d. Zufallszahlen sind beispielsweise das Inversionsverfahren, Rejection Sampling oder Importance Sampling (vgl. Vorlesung Computerintensive Methoden). Gerade bei höherer Dimension sind diese jedoch zum Teil nicht oder nur sehr kompliziert anwendbar.

4.7 Bayesianisches generalisiertes lineares Modell

4.7.1 Ein MCMC-Algorithmus: Metropolis-Hastings

Eine Alternative stellen *Markov Chain Monte Carlo (MCMC)*-Verfahren dar. Ziel ist die Generierung einer Markov-Kette $(X_0, \ldots, X_n)$ von (abhängigen!) Zufallszahlen, deren Verteilung gegen die interessierende Verteilung konvergiert, d.h. π ist die stationäre oder invariante Verteilung der Markov-Kette. Der *Ergodensatz* erlaubt dann Schätzungen der Form

$$\widehat{\mathbb{E}[X]} = \frac{1}{n - \text{burnin}} \sum_{i=\text{burnin}+1}^{n} x_i \quad \text{bzw.} \quad \widehat{\mathbb{E}[g(X)]} = \frac{1}{n - \text{burnin}} \sum_{i=\text{burnin}+1}^{n} g(x_i),$$

wobei $x_0, \ldots, x_{\text{burnin}}$ Werte am Anfang der Sequenz bezeichnen, bevor sich die Kette in der stationären Verteilung befindet, und die deshalb „weggeworfen" werden.

4.7 Bayesianisches generalisiertes lineares Modell

4.7.1 Ein MCMC-Algorithmus: Metropolis-Hastings

Praktische Umsetzung: Starte mit einem Startwert x_0 und ziehe dann für $i = 1, \ldots, n$ Werte $X_i \sim P(\cdot|X_{i-1})$, wobei P den Markov-Übergangskern bezeichnet, der nur vom aktuellen Zustand der Kette abhängt. An ihn bzw. die Markov-Kette werden die folgenden Anforderungen gestellt:

1. Die Markov-Kette ist homogen.
2. Die Markov-Kette ist irreduzibel.
3. Die Markov-Kette ist aperiodisch.
4. Die Markov-Kette ist positiv rekurrent.

4.7 Bayesianisches generalisiertes lineares Modell

4.7.1 Ein MCMC-Algorithmus: Metropolis-Hastings

Wir betrachten hier Markov-Ketten in diskreter Zeit bei diskretem oder stetigem Zustandsraum, gewöhnlich eine Teilmenge des $\mathbb{R}^p$. Für allgemeine Zustandsräume ist „mehr Technik" erforderlich, aber keine „neuen Ideen". Für den hier betrachteten Fall ist die Zielverteilung π immer gegeben.

Univariater Metropolis-Hastings

Wir beschreiben nun den *Metropolis-Hastings-Algorithmus* (kurz: *MH*) zur Generierung einer wie oben beschriebenen Markov-Kette für den univariaten Fall; dieser Algorithmus enthält den Gibbs-Sampler als Spezialfall. Sei π die Dichte der Zielverteilung, aus der wir simulieren möchten, und q eine geeignete Vorschlagsdichte, aus der neue Zustände der Kette generiert werden, d.h.

$$X_i \sim q(\cdot|x_{i-1}),$$

zum Beispiel $q_{x_i|x_{i-1}} = N(x_{i-1}, 1)$ oder
$q_{x_i|x_{i-1}} = Unif[x_{i-1} - c, x_{i-1} + c]$.

4.7 Bayesianisches generalisiertes lineares Modell

4.7.1 Ein MCMC-Algorithmus: Metropolis-Hastings

Die Vorschläge werden nicht immer, sondern nur mit einer gewissen *Akzeptanzwahrscheinlichkeit* $\alpha(x_{i-1}, x_i)$ akzeptiert. Für den MH-Algorithmus hat diese die Gestalt

$$\alpha(x_{i-1}, x_i) = \min\left(1, \frac{\pi(x_i) \cdot q(x_{i-1}|x_i)}{\pi(x_{i-1}) \cdot q(x_i|x_{i-1})}\right).$$

Wird x_i nicht akzeptiert, so setzt man $x_i \leftarrow x_{i-1}$, d.h. der alte Zustand wird beibehalten. Ein wesentlicher Vorteil dieses Verfahrens besteht darin, dass sich die (meist unbekannte) Normalisierungskonstante von π herauskürzt, d.h. der MH-Algorithmus kann auch (bzw. gerade) für diese Fälle angewendet werden. Die Konstruktion von α gewährleistet, dass die Bedingungen 1. bis 4. eingehalten werden.

4.7 Bayesianisches generalisiertes lineares Modell

4.7.1 Ein MCMC-Algorithmus: Metropolis-Hastings

Für $q(x_{i-1}|x_i) = q(x_i|x_{i-1})$ reduziert sich der MH-Algorithmus auf den *Metropolis-Algorithmus* mit

$$\alpha(x_{i-1}, x_i) = \min\left(1, \frac{\pi(x_i)}{\pi(x_{i-1})}\right),$$

d.h. wenn die Zieldichte an der Stelle x_i größer als an x_{i-1} ist, wird der neue Vorschlag stets akzeptiert, andernfalls nur im Verhältnis $\pi(x_i)/\pi(x_{i-1})$. Setzt man die Akzeptanzwahrscheinlichkeit konstant gleich eins, erhält man den Gibbs-Sampler.

Der MH-Algorithmus akzeptiert tendenziell neue Werte in Bereichen mit hoher Dichte (relevante Bereiche). Die Akzeptanzwahrscheinlichkeit sollte nicht zu gering sein, um regelmäßige Zustandsänderungen in der Kette zu erhalten. Sie sollte allerdings auch nicht zu hoch sein, d.h. die Varianz der Vorschlagsverteilung sollte nicht zu niedrig sein, damit der Träger von π ausreichend gut exploriert wird.

4.7 Bayesianisches generalisiertes lineares Modell

4.7.1 Ein MCMC-Algorithmus: Metropolis-Hastings

Algorithmus 11 : Univariater Metropolis-Hastings-Algorithmus

Setze Startwert x_0.

Für $i = 1, \ldots, n$:

1. Ziehe X_i aus $q(\cdot|x_{i-1})$.
2. Ziehe $U \sim Unif[0; 1]$; akzeptiere, wenn

$$U \leq \alpha(x_{i-1}, x_i),$$

ansonsten setze $x_i \leftarrow x_{i-1}$.

Multivariater Metropolis-Hastings

Die Verallgemeinerung auf den multivariaten Fall ist im Prinzip einfach, zum Beispiel mit

$$q(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}) = \text{MVN}(\boldsymbol{x}_{i-1}, \boldsymbol{\Sigma}).$$

Ein Problem stellt hier die Wahl der „Tuning-Matrix" $\boldsymbol{\Sigma}$ dar, die die Akzeptanzwahrscheinlichkeit steuert. Meist ist $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$; man startet mehrere Läufe und berechnet die Akzeptanzrate. Die Varianzen der Vorschlagsdichte werden dann solange variiert, bis „angemessene" Akzeptanzraten erreicht werden. Im Fall bayesianischer GLMs existiert eine Variante, die automatisch brauchbare Vorschlagsdichten berechnet, wie im folgenden Abschnitt beschrieben wird.

4.7 Bayesianisches generalisiertes lineares Modell

4.7.2 Metropolis-Hastings mit IWLS-Vorschlagsdichte

Aus der Vorlesung Generalisierte Regression ist das *Fisher-Scoring* bekannt:

4.7 Bayesianisches generalisiertes lineares Modell

4.7.2 Metropolis-Hastings mit IWLS-Vorschlagsdichte

Beispiel 4.5 (Fisher-Scoring beim Logit-Modell)
Die Scorefunktion im Logit-Modell (bei kanonischer Linkfunktion) hat die Form

$$\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})).$$

Die Fisher-Information ist

$$\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \text{diag}\Big(\mu_i(\boldsymbol{\beta})(1-\mu_i(\boldsymbol{\beta}))\Big)\boldsymbol{X}.$$

Bezeichnet $\widehat{\boldsymbol{\beta}}$ den ML-Schätzer, so ist

$$\text{Cov}(\widehat{\boldsymbol{\beta}}) = \Big[\boldsymbol{X}^\top \text{diag}\Big(\mu_i(\widehat{\boldsymbol{\beta}})(1-\mu_i(\widehat{\boldsymbol{\beta}}))\Big)\boldsymbol{X}\Big]^{-1}.$$

Der Fisher-Scoring Algorithmus hat dann die Form

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{(k+1)} &= \widehat{\boldsymbol{\beta}}^{(k)} + \boldsymbol{F}^{-1}(\widehat{\boldsymbol{\beta}}^{(k)})\boldsymbol{s}(\widehat{\boldsymbol{\beta}}^{(k)}) \\ &= \widehat{\boldsymbol{\beta}}^{(k)} + \Big[\boldsymbol{X}^\top \text{diag}\Big(\mu_i(\widehat{\boldsymbol{\beta}}^{(k)})(1-\mu_i(\widehat{\boldsymbol{\beta}}^{(k)}))\Big)\boldsymbol{X}\Big]^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}^{(k)})).\end{aligned}$$

4.7 Bayesianisches generalisiertes lineares Modell

4.7.2 Metropolis-Hastings mit IWLS-Vorschlagsdichte

Allgemein lässt sich das Fisher-Scoring wie folgt umschreiben:
Definiere Pseudo-Beobachtungen $\widetilde{\boldsymbol{y}} = (\widetilde{y}_1(\boldsymbol{\beta}), \ldots, \widetilde{y}_n(\boldsymbol{\beta}))^\top$, wobei

$$\widetilde{y}_i(\boldsymbol{\beta}) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + D_i^{-1}(y_i - \mu_i)$$

mit

$$D_i(\boldsymbol{\beta}) = \frac{\partial h(\eta_i)}{\partial \eta_i} = \frac{\partial h(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{\partial \boldsymbol{x}_i^\top \boldsymbol{\beta}} \quad \text{und } \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

Im Spezialfall des Logit-Modells ist

$$D_i(\boldsymbol{\beta}) = \mu_i(1 - \mu_i) = \mu_i(\boldsymbol{\beta})(1 - \mu_i(\boldsymbol{\beta})).$$

Fasse diese Einträge zu $\boldsymbol{D} = \text{diag}(D_1, \ldots, D_n)$ zusammen.
Definiere weiter

$$w_i(\boldsymbol{\beta}) = D_i^2(\boldsymbol{\beta})[\text{Var}(y_i)]^{-1} \text{ und } \boldsymbol{W} = \text{diag}(w_1(\boldsymbol{\beta}), \ldots, w_n(\boldsymbol{\beta})).$$

Im Logit-Modell:

$$w_i(\boldsymbol{\beta}) = \frac{[\mu_i(1-\mu_i)]^2}{\mu_i(1-\mu_i)} = \mu_i(1-\mu_i).$$

Dann lässt sich das Fisher-Scoring als *iterierte kleinste Quadrate-Schätzung (IWLS, iteratively (re)-weighted least squares)* schreiben:

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\boldsymbol{X}^\top \boldsymbol{W}^{(k)} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(k)} \tilde{\boldsymbol{y}}^{(k)}.$$

4.7 Bayesianisches generalisiertes lineares Modell

4.7.2 Metropolis-Hastings mit IWLS-Vorschlagsdichte

Aus der Analogie von kleinster Quadrate- und Maximum-Likelihood-Schätzung im Normalverteilungsfall lässt sich dies interpretieren als

$$\widetilde{\boldsymbol{y}}^{(k)} \sim \text{MVN}\left(\boldsymbol{X}\boldsymbol{\beta}, (\boldsymbol{W}^{-1})^{(k)}\right).$$

Bayesianische Version: Kombiniere das Ganze mit der Prioriverteilung $\boldsymbol{\beta} \sim \text{MVN}(\boldsymbol{\beta}_0, \boldsymbol{B}_0)$.

Iteriere dazu:

1. Aktueller Zustand sei $\boldsymbol{\beta}^{(t)}$; berechne

$$\widetilde{\boldsymbol{y}}^{(t)} = \boldsymbol{X}^\top \boldsymbol{\beta}^{(t)} + \boldsymbol{D}^{-1}(\boldsymbol{\beta}^{(t)})(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(t)})).$$

2. Ziehe $\boldsymbol{\beta}^* \sim \text{MVN}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{C}^{(t+1)})$ mit

$$\begin{aligned}\boldsymbol{\beta}^{(t+1)} &= (\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^{(t)})\boldsymbol{X})^{-1} \\ &\qquad \cdot [\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^{(t)})\widetilde{\boldsymbol{y}}(\boldsymbol{\beta}^{(t)})], \\ \boldsymbol{C}^{(t+1)} &= (\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^{(t)})\boldsymbol{X})^{-1}.\end{aligned}$$

3. Akzeptiere $\boldsymbol{\beta}^*$ mit Wahrscheinlichkeit

$$\alpha(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^*) = \min\left(1, \frac{f(\boldsymbol{\beta}^*|\boldsymbol{X})}{f(\boldsymbol{\beta}^{(t)}|\boldsymbol{X})} \times \frac{q(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^*)}{q(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(t)})}\right),$$

wobei $q(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^*)$ dem Wert der Dichte von

$$\text{MVN}\Big(\Big(\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^*)\boldsymbol{X}\Big)^{-1}\Big(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^*)\tilde{\boldsymbol{y}}(\boldsymbol{\beta}^*)\Big), \Big(\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^*)\boldsymbol{X}\Big)^{-1}\Big)$$

an der Stelle $\boldsymbol{\beta}^{(t)}$ entspricht ...

3. . . . und analog $q(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(t)})$ dem Wert der Dichte von

$$\text{MVN}\Bigg(\Big(\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^{(t)})\boldsymbol{X}\Big)^{-1}\Big(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^{(t)})\tilde{\boldsymbol{y}}(\boldsymbol{\beta}^{(t)})\Big), \Big(\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^{(t)})\boldsymbol{X}\Big)^{-1}\Bigg)$$

an der Stelle $\boldsymbol{\beta}^*$.

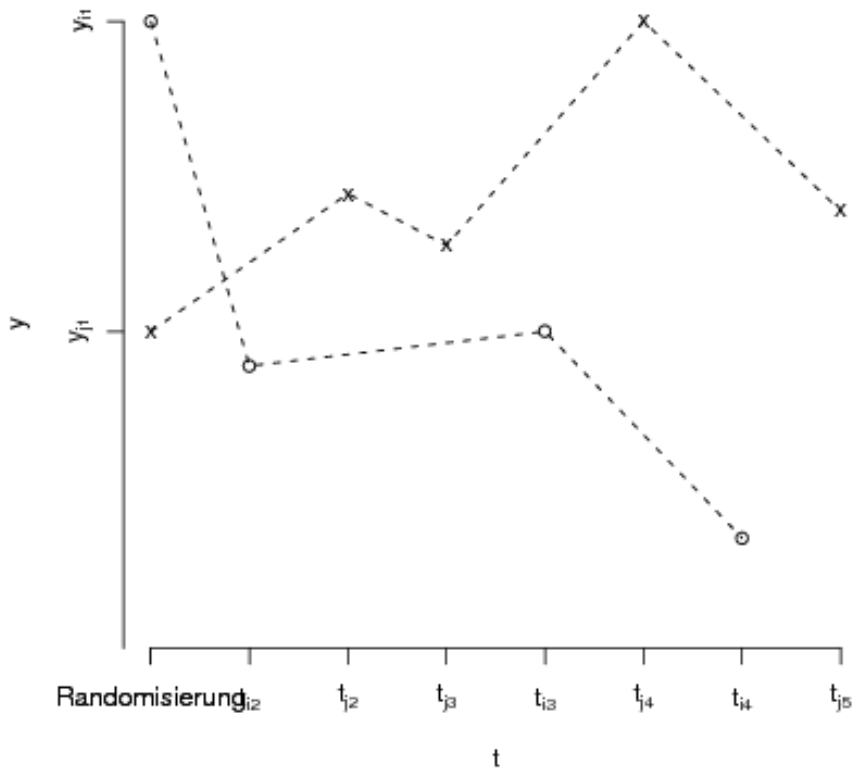
4.8 Bayesianische generalisierte lineare gemischte Modelle

Der Prädiktor des GLM aus dem vorherigen Abschnitt wird hier erweitert. Im Folgenden konzentrieren wir uns auf Cluster- und Longitudinaldaten. Bei letzteren lassen sich die Daten für ein Individuum i wie folgt strukturieren:

Response	Kovariablen			Zeitpunkt der Beobachtung
y_{i1}	x_{i11}	$\dots$	x_{ip1}	t_{i1}
$\vdots$				$\vdots$
y_{iT_i}	x_{i1T_i}	$\dots$	x_{ipT_i}	t_{iT_i}

4.8 Bayesianische generalisierte lineare gemischte Modelle

Dabei sind $y_{i1}, \ldots, y_{iT_i}$ korreliert, T_i kann variieren und die Beobachtungszeitpunkte können von Individuum zu Individuum variieren. Die Beobachtungszeitpunkte sollten jedoch nicht informativ für den Response sein. Die folgende Abbildung zeigt schematisch eine solche Datensituation, wie sie zum Beispiel bei einer kontrollierten Studie auftauchen könnte.



4.8 Bayesianische generalisierte lineare gemischte Modelle

Diese Art der Daten stellt nicht nur wegen der Abhängigkeit der Beobachtungen eine Herausforderung dar. Häufig sind hier zum Beispiel Drop-Outs und damit fehlende Werte.
Oft treten Longitudinaldaten zudem in Kombination mit Survival-Daten auf, zum Beispiel kann $y_{it_i}, \ldots, y_{iT_i}$ der (mit Messfehlern behaftete) Verlauf eines Biomarkers sein. Die Frage ist dann, ob der Biomarkerverlauf prognostisch für die Überlebenszeit ist. Dies führt zu sogenannten *joint models* (siehe Modelle für Longitudinal- und Überlebenszeitdaten).
GLMMs können mit diesem Datentyp gut umgehen, wenn man die sogenannte „bedingte Unabhängigkeitsannahme" akzeptiert:
...

4.8 Bayesianische generalisierte lineare gemischte Modelle

1. Erweitere den Prädiktor zu

$$\eta_{it} = \boldsymbol{x}_{it}^\top \underbrace{\boldsymbol{\beta}}_{\text{feste Effekte}} + \boldsymbol{z}_{it}^\top \underbrace{\boldsymbol{\alpha}_i}_{\text{zufällige Effekte}}$$

in der Annahme, dass

$$\boldsymbol{\alpha}_i \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}).$$

Dabei ist $\boldsymbol{x}_{it}^\top = (x_{i1}, \ldots, x_{ip_t})$ der Kovariablenvektor, und $\boldsymbol{z}_{it}^\top \in \mathbb{R}^{1 \times q}$ kann Kovariablen aus $\boldsymbol{x}_{it}$ enthalten und zum Beispiel die Zeit t selbst.

Beispiel 4.6 (Random Intercept Modell)
Sei

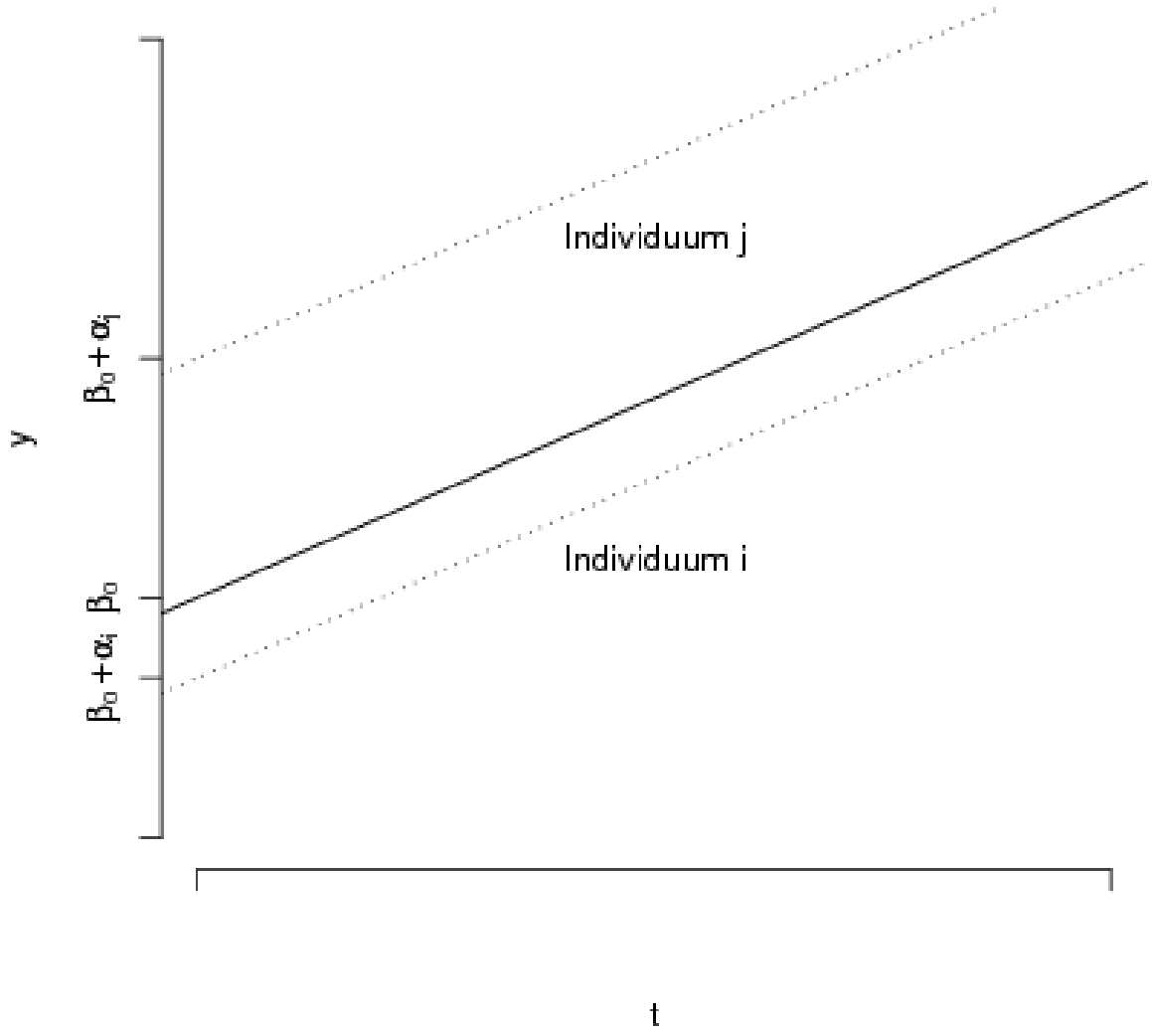
$$\eta_{it_i} = \beta_0 + \beta_1 t_i + \alpha_i,\ \alpha_i \ \sim \ N(0, \sigma_\alpha^2),$$

dann haben wir für ein Individuum i:

$$\begin{pmatrix} \eta_{it_{i_1}} \\ \vdots \\ \eta_{it_{T_i}} \end{pmatrix} = \begin{pmatrix} 1 & t_{i_1} \\ \vdots & \vdots \\ 1 & t_{i_{T_i}} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \alpha_i.$$

...

4.8 Bayesianische generalisierte lineare gemischte Modelle



2. Wir treffen die bedingte Unabhängigkeitsannahme

$$y_{it} \perp\!\!\!\perp y_{it'} \mid \boldsymbol{\alpha}_i, \boldsymbol{\beta}$$

für alle $t \neq t'$.
Diese erlaubt die Darstellung der gemeinsamen Verteilung von $(y_{i1}, \ldots, y_{iT_i})$ als Produkt der bedingten Verteilungen

$$f(y_{i1}, \ldots, y_{iT_i}) = \prod_{t=1}^{T_i} f(y_{it}|\boldsymbol{\alpha}_i).$$

4.8 Bayesianische generalisierte lineare gemischte Modelle

Ohne diese bedingte Unabhängigkeitsannahme verlieren GLMMs deutlich an Attraktivität. Das volle Setup bei n Individuen sieht wie folgt aus:

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT_i} \end{pmatrix}, \quad \boldsymbol{X}_i = \begin{pmatrix} x_{i11} & \dots & x_{ip1} \\ \vdots & & \\ x_{i1T_i} & \dots & x_{ipT_i} \end{pmatrix},$$

$$\boldsymbol{Z}_i = \begin{pmatrix} z_{i11} & \dots & z_{iq1} \\ \vdots & & \\ z_{i1T_i} & \dots & z_{iqT_i} \end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_n \end{pmatrix}$$

und

$$\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\alpha}_i.$$

4.8 Bayesianische generalisierte lineare gemischte Modelle

Zusammenfassend in Matrixnotation ergibt sich:

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{Z}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{Z}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_n \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{Z}_1 & \cdots & \boldsymbol{0} \\ \boldsymbol{X}_2 & \boldsymbol{0} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \boldsymbol{0} \\ \boldsymbol{X}_n & \boldsymbol{0} & \cdots & \boldsymbol{Z}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_n \end{pmatrix}.$$

4.8 Bayesianische generalisierte lineare gemischte Modelle

Die bayesianische Herangehensweise ist hier im Prinzip wie im GLM mit

$$\boldsymbol{\beta} \sim \text{MVN}(\boldsymbol{\beta}_0, \boldsymbol{B}_0) \quad \text{und} \quad \boldsymbol{\alpha} \sim \text{MVN}(\mathbf{0}, \underbrace{\text{diag}(\boldsymbol{\Sigma}, \ldots, \boldsymbol{\Sigma})}_{(n \cdot q) \times (n \cdot q)})$$

bzw.

$$\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \boldsymbol{\beta}_0 \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \underbrace{\text{diag}(\boldsymbol{B}_0, \boldsymbol{\Sigma}, \ldots, \boldsymbol{\Sigma})}_{(p+nq) \times (p+nq)} \right).$$

Dabei bezeichnen $\text{diag}(\boldsymbol{\Sigma}, \ldots, \boldsymbol{\Sigma})$ bzw. $\text{diag}(\boldsymbol{B}_0, \boldsymbol{\Sigma}, \ldots, \boldsymbol{\Sigma})$ Blockdiagonalmatrizen.

Bemerkung.

(i) Bei komplexeren Situationen, zum Beispiel bei nicht unabhängigen Individuen, kann $\operatorname{diag}(\boldsymbol{B}_0, \boldsymbol{\Sigma}, \ldots, \boldsymbol{\Sigma})$ durch eine nicht-blockweise Matrix ersetzt werden.

(ii) GLMMs sind hochdimensional, wenn n groß ist. Spezielle Algorithmen zur Optimierung sind notwendig.

4.8 Bayesianische generalisierte lineare gemischte Modelle

Zusätzlich wird eine (Hyper-) Prioriverteilung für $\boldsymbol{\Sigma}$ konstruiert, weil die $\boldsymbol{\alpha}_i$ unbeobachtete, latente Variablen sind, d.h. die $\boldsymbol{\alpha}_i$ sind zu behandeln wie die ε_i im linearen Modell; auch dort haben wir für die Varianz eine Priori angenommen.

Die Priori könnte zum Beispiel $\boldsymbol{\Sigma} \sim \text{inv-Wishart}_{\nu_0}(\boldsymbol{\Lambda}_0^{-1})$, also

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu_0+q+1)/2} \exp\left(-\frac{1}{2}\,\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_0)\right),$$

...

4.8 Bayesianische generalisierte lineare gemischte Modelle

...

mit resultierender Posteriori

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma} | \boldsymbol{y}, \boldsymbol{X}) \propto & \left[\prod_{i=1}^{n} \prod_{j=1}^{T_i} f(y_{it_j} | \cdot) \right] \times \\
& \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{B}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right) \times \\
& |\boldsymbol{\Sigma}|^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \boldsymbol{\alpha}_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}_i \right) \times \\
& |\boldsymbol{\Sigma}|^{-(\nu_0 + q + 1)/2} \exp\left(-\frac{1}{2} \operatorname{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda}_0) \right) .
\end{aligned}
$$

4.8 Bayesianische generalisierte lineare gemischte Modelle

Ein möglicher Algorithmus zur Simulation der Posteriori ist dann ein blockweiser Gibbs-Sampler.

(i) *Full-Conditional für den β-Block:*

$$f(\beta|\alpha, \Sigma, y, X) \propto \left[\prod_{i=1}^{n}\prod_{j=1}^{T_i} f(y_{it_j}|\cdot)\right] \cdot \exp\left(-\frac{1}{2}(\beta-\beta_0)^\top B_0^{-1}(\beta-\beta_0)\right)$$

lässt sich wie im bayesianischen GLM behandeln, wenn man zusätzlich einen Offset $z_i^T \alpha_i$ verwendet:

$$\widetilde{y}_i(\beta^{(t-1)}|\alpha_i) = x_i^\top \beta^{(t-1)} + z_i^\top \alpha_i + D_i^{-1}[y_i - \mu_i(\beta^{(t-1)}, \alpha_i)]$$

bzw.

$$\widetilde{y}_i(\beta^{(t-1)}|\alpha_i) - z_i^\top \alpha_i = x_i^\top \beta^{(t-1)} + D_i^{-1}[y_i - \mu_i(\beta^{(t-1)}, \alpha_i)]$$

Definiert man $\widetilde{\widetilde{y}}_i(\beta^{(t-1)}|\alpha_i) = \widetilde{y}_i(\beta^{(t-1)}|\alpha_i) - z_i^\top \alpha_i$, so lässt sich der IWLS-Metropolis-Hastings-Algorithmus aus 4.7.2 zum Ziehen aus dieser Full-Conditional anwenden.

4.8 Bayesianische generalisierte lineare gemischte Modelle

2. *Full-Conditional für den α_i-Block:* Für $i = 1, \ldots, n$ erhält man

$$f(\boldsymbol{\alpha}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{y}, \boldsymbol{X}) \propto \left[\prod_{i=1}^{n}\prod_{j=1}^{T_i} f(y_{it_j}|\cdot)\right] \exp\left(-\boldsymbol{\alpha}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}_i\right).$$

Dies lässt sich wieder wie ein GLM mit Offset $\boldsymbol{x}_i^\top\boldsymbol{\beta}$ interpretieren und mit dem IWLS-MH-Algorithmus mit Proposal-Verteilung

$$\text{MVN}\Big([\boldsymbol{\Sigma}^{-1} + \boldsymbol{z}_i\boldsymbol{W}_i(\boldsymbol{\alpha}_i^{(t-1)})\boldsymbol{z}_i^\top]^{-1}\boldsymbol{z}_i\boldsymbol{W}_i(\boldsymbol{\alpha}_i^{(t-1)})[\widetilde{y}_i(\boldsymbol{\alpha}_i^{(t-1)}) - \boldsymbol{x}_i^\top\boldsymbol{\beta}],$$
$$[\boldsymbol{\Sigma}^{-1} + \boldsymbol{z}_i\boldsymbol{W}_i(\boldsymbol{\alpha}_i^{(t-1)})\boldsymbol{z}_i^\top]^{-1}\Big)$$

behandeln.

3. *Full-Conditional für* $\boldsymbol{\Sigma}$*:*

$$f(\boldsymbol{\Sigma}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{y}, \boldsymbol{X}) \propto |\boldsymbol{\Sigma}|^{-(n+\nu_0+q+1)/2} \cdot \exp\left(-\frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_0 + \sum_{i=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^\top\right)\right)$$

entspricht dem Kern einer inv-Wishart$_{\nu_0+n}(\boldsymbol{\Lambda}_0 + \sum_{i=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^\top)$-Verteilung (implizite Dimension $q \times q$); diese lässt sich direkt mit einem geeigneten Zufallszahlengenerator simulieren.

4.8 Bayesianische generalisierte lineare gemischte Modelle

Insgesamt hat man also einen blockweisen Gibbs-Sampler mit

$$\begin{array}{ccccccc} 1 & + & n & + & 1 & = & n+2 \\ \boldsymbol{\beta} & & \{\boldsymbol{\alpha}_i\}_{i=1}^n & & \boldsymbol{\Sigma} & & \end{array}$$

Blöcken.
Verwendet man beim Update von $\boldsymbol{\beta}$ und $\boldsymbol{\alpha}_i$ einen Akzeptanzmechanismus, dann handelt es sich bei dem Gibbs-Sampler genauer gesagt um einen *Metropolis-Hastings-within-Gibbs*-Algorithmus, da die einzelnen Blöcke jeweils mit Metropolis-Hastings erzeugt, am Ende des Durchgangs aber die $n+2$ Blöcke noch einmal (mit Wahrscheinlichkeit 1) formal akzeptiert werden.