Gemischte Modelle

Oktober 2016

Sonja Greven (V) und David Rügamer (Ü)
Institut für Statistik
Ludwig-Maximilians-Universität München
https://www.elab.moodle.elearning.lmu.de/course/view.php?id=1074

Mit Dank an Susanne Konrath und Fabian Scheipl für Material vergangener Jahre.

Literatur Gemischte Modelle

- Demidenko, E. (2004). *Mixed Models: Theory and Applications.* Wiley.
- Diggle, P. J.; Heagerty, P.; Liang, K. L.; Zeger, S. L. (2002). *Analysis of Longitudinal Data.* Oxford University Press, Oxford.
- **Fahrmeir, L., Kneib, T. und Lang, S. (2009)**. *Regression: Modelle, Methoden und Anwendungen (2. Auflage).* Springer. - Begleitend zur Vorlesung. Erhältlich als Ebook bei der Universitätsbibliothek: `https://opac.ub.uni-muenchen.de/`
- McCulloch, C. E.; Searle, S. R. (2001). *Generalized, Linear, and Mixed Models.* John Wiley.
- Pinheiro, J. C.; Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS.* Springer, New York. - Praxisorientierte Einführung in die Analyse gemischter Modelle und ausführliche Beschreibung des R-Pakets `nlme` für LMMs.
- Ruppert, D.; Wand, M. P.; Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge University Press. - Verbindung gemischte Modelle und Penalisierung.
- Verbeke, G.; Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer, New York.
- Wood, S. (2006). *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC. - Verbindung gemischte Modelle und Penalisierung, R-Paket `mgcv`.

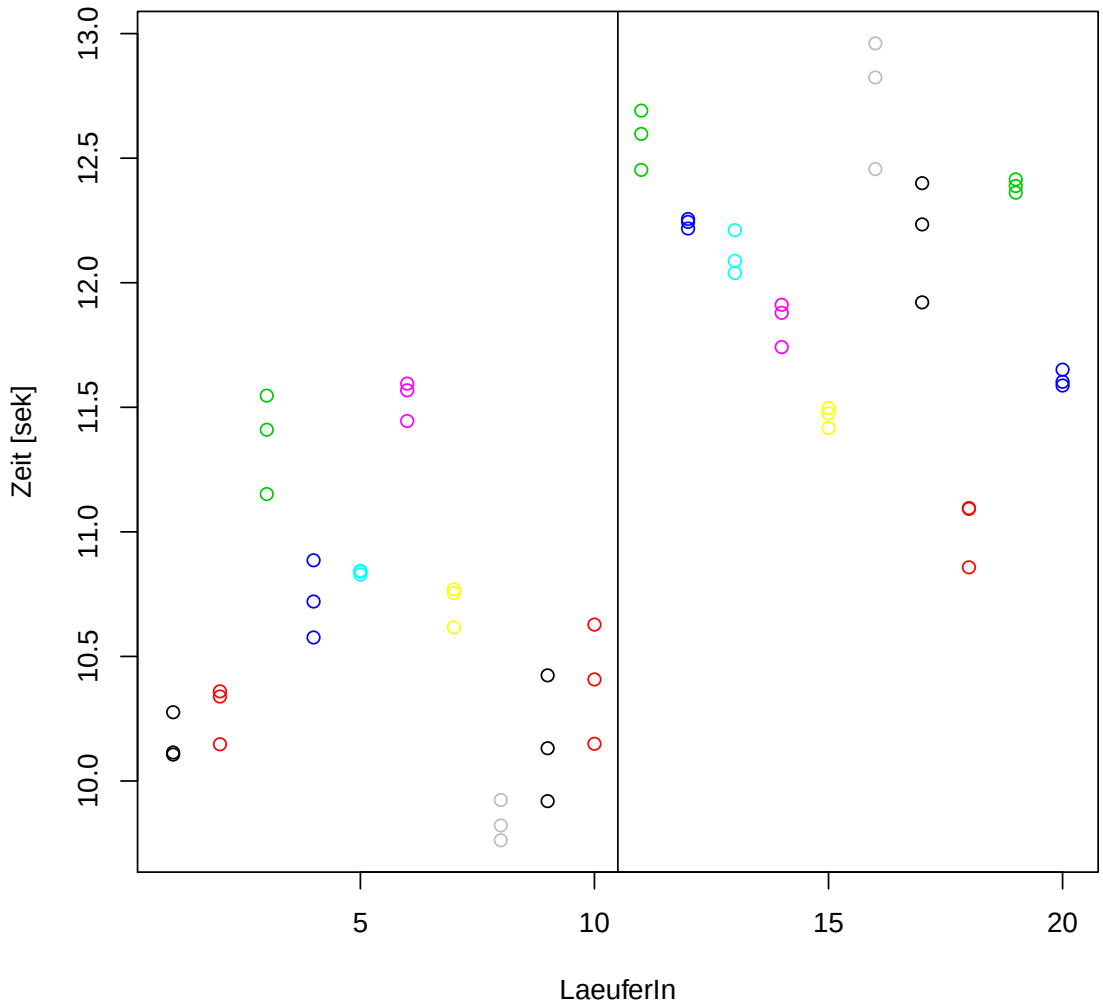
Inhalt der Vorlesung Gemischte Modelle

1. Das lineare gemischte Modell
2. Likelihood-Schätzung für lineare gemischte Modelle
3. Likelihood-Inferenz im linearen gemischten Modell
4. Bayes-Schätzung für lineare gemischte Modelle
5. Additive gemischte Modelle
6. Das generalisierte lineare gemischte Modell
7. Likelihood-Schätzung für generalisierte lineare gemischte Modelle

Inhalt

Motivation: 100-Meter-Lauf

Drei 100-Meter-Läufe mit 10 Männern und 10 Frauen. Die Zeiten sehen so aus:



Wie könnte man diese Daten mit bekannten Methoden (lineare Modelle) modellieren? → **Diskussion**

Motivation: 100-Meter-Lauf

Wir wollen ein Modell, das uns ermöglicht:

- die Schätzung des Geschlechtseffekts (Populationsparameter)
- die Schätzung der LäuferInneneffekte (individuelle Effekte)
- die Schätzung der Korrelationsstruktur
- valide Inferenz.

100-Meter-Lauf: Ein lineares gemischtes Modell

Überlegung: Hätten wir pro Person ihre Durchschnittszeit μ_i (d.h. ein Wert pro Person), wäre ein sinnvolles Modell (iid = unabhängig identisch verteilt)

$$\mu_i = \beta_{g_i} + b_i, \qquad b_i \overset{iid}{\sim} N(0, \tau^2), \tag{1}$$

da die Durchschnittszeiten einzelner Personen um das Geschlechtsmittel β_{g_i}, $g_i \in \{1, 2\}$, variieren.

Die beobachteten Zeiten pro Lauf streuen um die individuelle Durchschnittszeit:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2). \tag{2}$$

Zusammen genommen ergibt sich damit das Modell ($\perp$ = unabhängig)

$$y_{ij} = \beta_{g_i} + b_i + \varepsilon_{ij}, \qquad b_i \overset{iid}{\sim} N(0, \tau^2), \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \qquad b_i \perp \varepsilon_{ij}. \tag{3}$$

Ein Blick auf das Modell

$$y_{ij} = \beta_{g_i} + b_i + \varepsilon_{ij}, \qquad b_i \overset{iid}{\sim} N(0, \tau^2), \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2). \tag{3}$$

Die Effekte b_i sind

- die „Fehler"terme in (1)
- Teil des Erwartungswertes in (2)
- sogenannte **zufällige Effekte** in Modell (3). Sie spiegeln hier wieder, dass die LäuferInnen aus einer Population kommen (in der wir die individuellen Durchschnittszeiten als normalverteilt um das Geschlechtsmittel annehmen).

Ein Modell mit zufälligen und festen Effekten nennt man **gemischtes Modell**.

Interpretation der Parameter

$$y_{ij} = \beta_{g_i} + b_i + \varepsilon_{ij}, \qquad b_i \overset{iid}{\sim} N(0, \tau^2), \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2). \tag{3}$$

- β_1, β_2 die Durchschnittszeiten für Männer / Frauen (Populationsparameter)
- b_i die Abweichung der Durchschnittszeit von Person i vom Mittel der Geschlechtsgruppe (individuelle Effekte)
- τ^2 die Varianz der Durchschnittszeiten pro Geschlechtsgruppe (in den zwei Gruppen als gleich **angenommen**)
- ε_{ij} die Abweichung der j-ten Laufzeit von der Durchschnittszeit für Person i
- σ^2 die Varianz der persönlichen Laufzeiten (für alle i gleich **angenommen**)

Bedingte Sicht

$$y_{ij} = \beta_{g_i} + b_i + \varepsilon_{ij}, \qquad b_i \overset{iid}{\sim} N(0, \tau^2), \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2). \tag{3}$$

Betrachte die Verteilung von y_{ij} bedingt auf b_i ($\rightarrow$ **überlegen**):

$$y_{ij}|b_i \overset{iid}{\sim} N(\beta_{g_i} + b_i, \sigma^2)$$

b_i modelliert individuelle Effekte im Erwartungswert (EW) analog zum linearen Modell mit festen Effekten

$$y_{ij} = \beta_{g_i} + \beta_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

Aber:

- Die Geschlechtseffekte sind identifizierbar (nicht kollinear mit den β_i).
- Das Modell ist bei 1-2 Messungen für einige i schätzbar (mehr später).
- Da wir im Wesentlichen τ^2 schätzen (zur Vorhersage der b_i mehr später), wächst die Zahl der Parameter nicht mit der Anzahl der LäuferInnen.

Marginale Sicht

$$y_{ij} = \beta_{g_i} + b_i + \varepsilon_{ij}, \qquad b_i \overset{iid}{\sim} N(0, \tau^2), \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2). \tag{3}$$

Betrachte die marginale Verteilung der y_{ij} ($\rightarrow$ **überlegen**):

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} \beta_{g_i} \\ \beta_{g_i} \\ \beta_{g_i} \end{pmatrix}, \begin{pmatrix} \tau^2 + \sigma^2 & \tau^2 & \tau^2 \\ \tau^2 & \tau^2 + \sigma^2 & \tau^2 \\ \tau^2 & \tau^2 & \tau^2 + \sigma^2 \end{pmatrix} \right) \tag{4}$$

b_i induziert eine Kovarianzstruktur analog zum allgemeinen linearen Modell

$$y_{ij} = \beta_{g_i} + \varepsilon_{ij}, \qquad \boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}) \sim N(\mathbf{0}, \boldsymbol{V}_i). \tag{5}$$

Hier:

- Begründung für eine mögliche Kovarianzstruktur $\boldsymbol{V}_i$ (auch unbalanziert).
- Vorhersage individueller Effekte möglich (mehr später).

(3) impliziert das allgemeine lineare Modell (5), aber nicht umgekehrt!

Wozu gemischte Modelle?

Anhand des Beispiels zeigt sich bereits: Gemischte Modelle werden gerne verwendet für die Analyse korrelierter Daten. Z.B.

- Longitudinaldaten: Wiederholte Beobachtungen in zeitlicher Abfolge an denselben Subjekten/Beobachtungseinheiten. (z.B. Patienten über die Zeit)
- Clusterdaten / gruppierte Daten: Gruppen (Cluster) mit mehreren Beobachtungen pro Gruppe. (z.B. Läuferdaten, Daten für Familien)
- Hierarchische Daten: Gruppierte Daten mit mehreren geschachtelten Ebenen. (z.B. Schüler in Klassen in Schulen, Patienten in Ärzten in Krankenhäusern)
- Gekreuzte Designs: Gruppierte Daten mit mehreren gekreuzten Ebenen. (z.B. alle Personen bearbeiten die gleichen Aufgaben)

Daten vom gleichen Subjekt/Cluster/Beobachtungseinheit sind sich tendenziell ähnlicher als Daten verschiedener Subjekte/Cluster/Beobachungseinheiten.

Das allgemeine lineare gemischte Modell (LMM)

Definition: In allgemeiner Form ist das **lineare gemischte Modell** gegeben durch

$$\underset{n\times 1}{\boldsymbol{y}} = \underset{n\times p}{\boldsymbol{X}}\,\underset{p\times 1}{\beta} + \underset{n\times r}{\boldsymbol{Z}}\,\underset{r\times 1}{\boldsymbol{b}} + \underset{n\times 1}{\varepsilon}. \tag{6}$$

Mit $\varepsilon \sim N(\mathbf{0}, \boldsymbol{R})$, $\boldsymbol{b} \sim N(\mathbf{0}, \boldsymbol{G})$ und der Annahme, dass die zufälligen Effekte $\boldsymbol{b}$ und die Fehler ε **unabhängig** sind, ist die **Verteilungsannahme** gegeben durch

$$\begin{pmatrix} \boldsymbol{b} \\ \varepsilon \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R} \end{pmatrix}\right). \tag{7}$$

Die Kovarianzmatrizen $\boldsymbol{G}$ für $\boldsymbol{b}$ und $\boldsymbol{R}$ für ε werden als **positiv semi-definit** bzw. **positiv definit** angenommen.

Zu den Verteilungsannahmen

- Die **Normalverteilungsannahme der Fehler** ε ist nicht für alle Aussagen zur Inferenz in LMMs notwendig. Da Likelihood-basierte Schätzung üblich ist, nehmen wir sie jedoch in die Definition mit auf.

- Für die **zufälligen Effekte** $\boldsymbol{b}$ ist die **Normalverteilungsannahme** nicht zwingend. Alternative Verteilungen, z.B. Mischungsverteilungen, möglich. I.d.R. werden dann die Algorithmen zur Berechnung der Schätzer komplexer.

- ε und $\boldsymbol{b}$ sind als unabhängig angenommen.

Vorteile der Analyse mit gemischten Modellen

Zufällige Effekte können als Platzhalter für die Effekte von unbeobachteten oder unzureichend gemessenen Kovariablen dienen, die **Korrelation zwischen Beobachtungen an den gleichen Beobachtungseinheiten** verursachen.

Im Gegensatz zum linearen Regressionsmodell mit unabhängigen Fehlern führt die Berücksichtigung dieser Korrelation

- zu einer **verbesserten Schätzgenauigkeit** (kleineren wahren Standardfehlern) $\rightarrow$ **Übung**
- zu validen modellbasierten Standardfehlern und damit Konfidenzintervallen und Tests $\rightarrow$ **Übung**

Vorteile der Analyse mit gemischten Modellen

Im Gegensatz zum linearen Regressionsmodell mit festen Effekten für die Beobachtungseinheiten

- können Effekte für Kovariablen (z.B. Geschlecht), die nur zwischen Beobachtungseinheiten variieren, geschätzt werden.
- werden die festen Effekte wegen der kleineren Anzahl von Modellparametern effizienter geschätzt

Im Gegensatz zum linearen Regressionsmodell mit allgemeiner Kovarianz

- erlauben die Vorhersagen für die zufälligen Effekte **individuelle Prognosen**.

Weitere Annahmen in gemischten Modellen

Auch wenn die zufälligen Effekte die Effekte von unbeobachteten Kovariablen auffangen, so können Sie **nicht** Confounding durch solche Kovariablen verhindern.

Dies liegt daran, dass die zufälligen Effekte (marginal, siehe später), die Kovarianzstruktur beeinflussen, nicht jedoch den Erwartungswert.

Streng genommen, lautet in (7) die Annahme an ε und $\boldsymbol{b}$:

$$\mathrm{E}(\varepsilon|\boldsymbol{X}, \boldsymbol{b}) = \mathbf{0}, \qquad \mathrm{E}(\boldsymbol{b}|\boldsymbol{X}) = \mathbf{0}$$

(Regressionsannahme und Random effects-Annahme), so dass $\mathrm{E}(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$.
Dies ist insbesondere verletzt, wenn Confounding vorliegt.

Beispiel Confounding

Betrachte das Modell

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + u_{ij}\gamma + b_i + \varepsilon_{ij}, \qquad b_i \overset{iid}{\sim} N(0, \tau^2), \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

y_{ij} misst den Schulerfolg von Kind j in Schule i; u_{ij} ist a) Förderbedarf des Kindes oder b) Familieneinkommen. Beispiele für mögliches Confounding wären

a) Bei Förderbedarf wählen Eltern Schulen mit unterstützender Schulkultur, die sich auch im Schulerfolg wiederspiegelt (Random effects-Annahme verletzt).

b) Familieneinkommen korreliert mit Sprachkenntnissen, Ziele der Eltern für ihre Kinder etc., die mit Schulerfolg korrelieren (Regressionsannahme verletzt).

Ein Modell mit festen Effekten b_i würde den Bias durch a), jedoch nicht den Bias durch b) verhindern. Ziel muss auf jeden Fall sein, in $\boldsymbol{x}_{ij}$ mögliche Confounder möglichst gut abzubilden.

Frei nach Clarke, Crawford, Steele & Vignoles (2010): *The Choice Between Fixed and Random Effects Models: Some Considerations for Educational Research.* IZA Discussion Paper No. 5287

Beispiel 100-Meter-Lauf

Das Modell für die 100-Meter-Lauf-Daten

$$y_{ij} = \beta_{g_i} + b_i + \varepsilon_{ij}, \qquad b_i \overset{iid}{\sim} N(0, \tau^2), \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \qquad b_i \perp \varepsilon_{ij},$$

lässt sich in die Form

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \varepsilon, \qquad \boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G}),\ \varepsilon \sim N(\boldsymbol{0}, \boldsymbol{R}),\ \boldsymbol{b} \perp \varepsilon$$

bringen mit

$$\boldsymbol{y} = (y_{11}, \ldots, y_{20,3})' \qquad \boldsymbol{b} = (b_1, \ldots, b_{20})$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 1 & \ldots & 1 \end{pmatrix}' \qquad \boldsymbol{G} = \tau^2 \boldsymbol{I}_{20}$$

$$\beta = (\beta_1, \beta_2)' \qquad \varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{20,3})$$

$$\boldsymbol{Z} = \begin{pmatrix} 1 & 1 & 1 & \ldots & 0 & 0 & 0 \\ & & & \ddots & & & \\ 0 & 0 & 0 & \ldots & 1 & 1 & 1 \end{pmatrix}' \qquad \boldsymbol{R} = \sigma^2 \boldsymbol{I}_{60}.$$

Spezialfall Longitudinal- und Clusterdaten

Wiederholte Beobachtungen y_{ij} der Zielvariablen $\rightarrow$ **Beispiel in Übung**

- von Subjekt i zum Zeitpunkt t_{ij} bei Longitudinaldaten
- für das j-te Objekt aus dem Cluster i bei Clusterdaten

mit jeweils Kovariablenvektor $(\boldsymbol{x}'_{ij}, \boldsymbol{z}'_{ij})'$, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$.

Verschiedene Variabilitätsquellen in den Daten:

- Zwischen den Subjekten/Clustern, Abweichungen vom Populationsmittel.
- Innerhalb des Subjekts/Clusters, Abweichungen einer Messung vom Mittelwert des entsprechenden Subjekts/Clusters.

Spezialfall Longitudinal- und Clusterdaten

Das lineare gemischte Modell auf Beobachtungs- bzw. Cluster/Subjekt-Ebene ist

$$\begin{aligned} y_{ij} &= \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i + \varepsilon_{ij}, \quad j = 1, \ldots, n_i, i = 1, \ldots, N \qquad \text{bzw.} \\ \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, N. \end{aligned}$$

wobei $\boldsymbol{X}_i$ und $\boldsymbol{Z}_i$ die n_i Zeilen $\boldsymbol{x}_{ij}'$ bzw. $\boldsymbol{z}_{ij}'$ enthalten und $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})$.

Annahmen:

- Unabhängig und identisch normalverteilte zufällige Effekte $\boldsymbol{b}_i \overset{iid}{\sim} N(\boldsymbol{0}, \boldsymbol{D})$,
- unabhängig und normalverteilte Fehler $\varepsilon_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$,
- $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \varepsilon_1, \ldots, \varepsilon_N$ unabhängig,
- $\boldsymbol{D}$ positiv semi-definit und $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_N$ positiv definit.

Spezialfall Longitudinal- und Clusterdaten

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, N.$$

- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ entspricht den (festen) **Populationseffekten.**
- $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iq})'$ entspricht den (zufällligen) **subjekt- / clusterspezifischen Effekten.**
- $\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}'_{i1} \\ \vdots \\ \boldsymbol{x}'_{in_i} \end{pmatrix}$, $\boldsymbol{Z}_i = \begin{pmatrix} \boldsymbol{z}'_{i1} \\ \vdots \\ \boldsymbol{z}'_{in_i} \end{pmatrix}$ sind die **Designmatrizen** für die p populationsspezifischen bzw. q subjektspezifischen Kovariablen.

Dabei können die Kovariablen mit j (bzw. t_{ij}, *zeitvariierend*) variieren oder nicht.

Falls die Variablen $\boldsymbol{z}_{ij}$ in $\boldsymbol{x}_{ij}$ enthalten sind, lassen sich die $\boldsymbol{b}_i$ mit $\mathrm{E}(\boldsymbol{b}_i) = \boldsymbol{0}$ als **individuelle Abweichungen** vom Populationsmittel interpretieren.

Spezialfall Longitudinal- und Clusterdaten

Das lineare gemischte Modell für Longitudinal- und Clusterdaten ist ein Spezialfall des allgemeinen LMMs

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \varepsilon, \qquad \boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G}),\ \varepsilon \sim N(\boldsymbol{0}, \boldsymbol{R}),\ \boldsymbol{b} \perp \varepsilon$$

mit

- $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N')'$ und $\varepsilon = (\varepsilon_1', \ldots, \varepsilon_N')'$ der Länge $n = \sum_{i=1}^{N} n_i$,
- β der Länge p,
- $\boldsymbol{b} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_N')'$ der Länge $r = Nq$,
- $\boldsymbol{X} = (\boldsymbol{X}_1' | \ldots | \boldsymbol{X}_N')'$ der Dimension $n \times p$,
- $\boldsymbol{Z} = \text{blockdiag}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)$ der Dimension $n \times Nq$,
- $\boldsymbol{G} = \text{blockdiag}(\boldsymbol{D}, \ldots, \boldsymbol{D}, \ldots, \boldsymbol{D})$ der Dimension $Nq \times Nq$,
- $\boldsymbol{R} = \text{blockdiag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_i, \ldots, \boldsymbol{\Sigma}_N)$ der Dimension $n \times n$.

Die blockdiagonalen Kovarianzmatrizen resultieren aus der Unabhängigkeitsannahme für Beobachtungen an verschiedenen Individuen / Clustern.

Spezialfall hierarchische Struktur: Beispiel

Die Lesefähigkeit von 875 achtjährigen Schülern in 29 Klassen in 11 Schulen wird anhand eines standardisierten Scores y gemessen. Mögliches Modell:

$$\begin{aligned} y_{ijk} &= \beta_0 + b_i + b_{ij} + \varepsilon_{ijk}, \\ & \quad b_i \overset{iid}{\sim} N(0, \tau_1^2),\ b_{ij} \overset{iid}{\sim} N(0, \tau_2^2),\ \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2),\ b_i, b_{ij}, \varepsilon_{ijk} \text{ unabh.} \end{aligned}$$

wobei

- y_{ijk}: der Lesescore des k-ten Kindes in der j-ten Klasse der i-ten Schule
- β_0: die allgemeine mittlere Lesefähigkeit von Achjährigen
- b_i: die Abweichung der mittleren Lesefähigkeit in Schule i vom allgemeinen Mittel
- b_{ij}: die Abweichung der mittleren Lesefähigkeit der Klasse j vom Mittel ihrer Schule i
- ε_{ijk}: die Abweichung der Lesefähigkeit von Kind k von der mittleren Lesefähigkeit seiner Klasse.

Spezialfall hierarchische Struktur: Beispiel

Das Modell lässt sich wieder in die allgemeine LMM-Form

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \varepsilon, \qquad \boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G}),\ \varepsilon \sim N(\boldsymbol{0}, \boldsymbol{R}),\ \boldsymbol{b} \perp \varepsilon$$

bringen mit

- $\boldsymbol{y} = (y_{1,1,1}, \ldots, y_{11,3,30})'$ und $\boldsymbol{\varepsilon} = (\varepsilon_{1,1,1}, \ldots, \varepsilon_{11,3,30})'$ der Länge $n = 875$,
- $\boldsymbol{X} = (1, \ldots, 1)'$ der Dimension 875×1,
- $\boldsymbol{\beta} = \beta_0$,
- $\boldsymbol{b} = (b_1, \ldots, b_{11} | b_{1,1}, \ldots, b_{11,3})'$ der Länge $r = 11 + 29 = 40$,
- $\boldsymbol{Z} = (\text{blockdiag}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{11}) | \text{blockdiag}(\boldsymbol{Z}_{1,1}, \ldots, \boldsymbol{Z}_{11,3}))$ der Dimension 875×40, wobei $\boldsymbol{Z}_i$ bzw. $\boldsymbol{Z}_{ij}$ Einservektoren sind mit Länge gleich der Anzahl der Schüler (bzw. der Schüler in Klasse j) in Schule i,
- $\boldsymbol{G} = \text{blockdiag}(\tau_1^2 \boldsymbol{I}_{11}, \tau_2^2 \boldsymbol{I}_{29})$ der Dimension 40×40,
- $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_{875}$ der Dimension 875×875.

Spezialfall gekreuzte Struktur: Beispiel

In einem Phonetik-Experiment sprechen 9 Subjekte 140 Worte, in denen *s*- und *sch*-Laute vorkommen, je 5 mal. Ein akustischer Index y misst, ob der gesprochene Laut einem *s* oder *sch* näher ist.

Mögliches Modell für y_{ijk} ($i$te Person, $j$tes Wort, $k$te Wiederholung):

$$y_{ijk} = \boldsymbol{x}_j'\boldsymbol{\beta} + b_i + c_j + d_{ij} + \varepsilon_{ijk},$$

$$b_i \overset{iid}{\sim} N(0, \tau_b^2),\ c_j \overset{iid}{\sim} N(0, \tau_c^2),\ d_{ij} \overset{iid}{\sim} N(0, \tau_d^2),\ \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2),$$

$$b_i, c_j, d_{ij}, \varepsilon_{ijk} \text{ unabhängig},\ i = 1, \ldots, 9;\ j = 1, \ldots, 140;\ k = 1, \ldots, 5,$$

mit

- $\boldsymbol{\beta}$ Effekte von Wortmerkmalen wie Betonung etc.
- zufällige Effekte b_i für Person i, c_j für Wort j und d_{ij} für deren Interaktion.

Ein einfacheres Modell wäre das Modell ohne Interaktion d_{ij}.

Spezialfall gekreuzte Struktur: Beispiel

Das Modell lässt sich wieder in die allgemeine LMM-Form

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \varepsilon, \qquad \boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G}),\ \varepsilon \sim N(\boldsymbol{0}, \boldsymbol{R}),\ \boldsymbol{b} \perp \varepsilon$$

bringen mit

- $\boldsymbol{y} = (y_{1,1,1}, \ldots, y_{9,140,5})'$ und $\varepsilon = (\varepsilon_{1,1,1}, \ldots, \varepsilon_{9,140,5})'$ der Länge $n = 9 \cdot 140 \cdot 5 = 6300$,
- $\boldsymbol{X}$ der Dimension $6300 \times p$ enthält $\boldsymbol{x}_j$ zeilenweise,
- $\boldsymbol{Z} = (\boldsymbol{I}_9 \otimes \boldsymbol{1}_{140 \cdot 5} | \boldsymbol{1}_9 \otimes \boldsymbol{I}_{140} \otimes \boldsymbol{1}_5 | \boldsymbol{I}_{9 \cdot 140} \otimes \boldsymbol{1}_5)$ der Dimension 6300×1409, wobei $\boldsymbol{1}_l$ der Einservektor der Länge l ist und $\otimes$ das Kroneckerprodukt.
- $\boldsymbol{b} = (b_1, \ldots, b_9 | c_1, \ldots, c_{140} | d_{1,1}, \ldots, d_{9,140})'$ der Länge $r = 9 + 140 + 9 \cdot 140 = 1409$,
- $\boldsymbol{G} = \text{blockdiag}(\tau_b^2 \boldsymbol{I}_9, \tau_c^2 \boldsymbol{I}_{140}, \tau_d^2 \boldsymbol{I}_{9 \cdot 140})$ der Dimension 1409×1409,
- $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_{6300}$ der Dimension 6300×6300.

Die Kovarianzstruktur

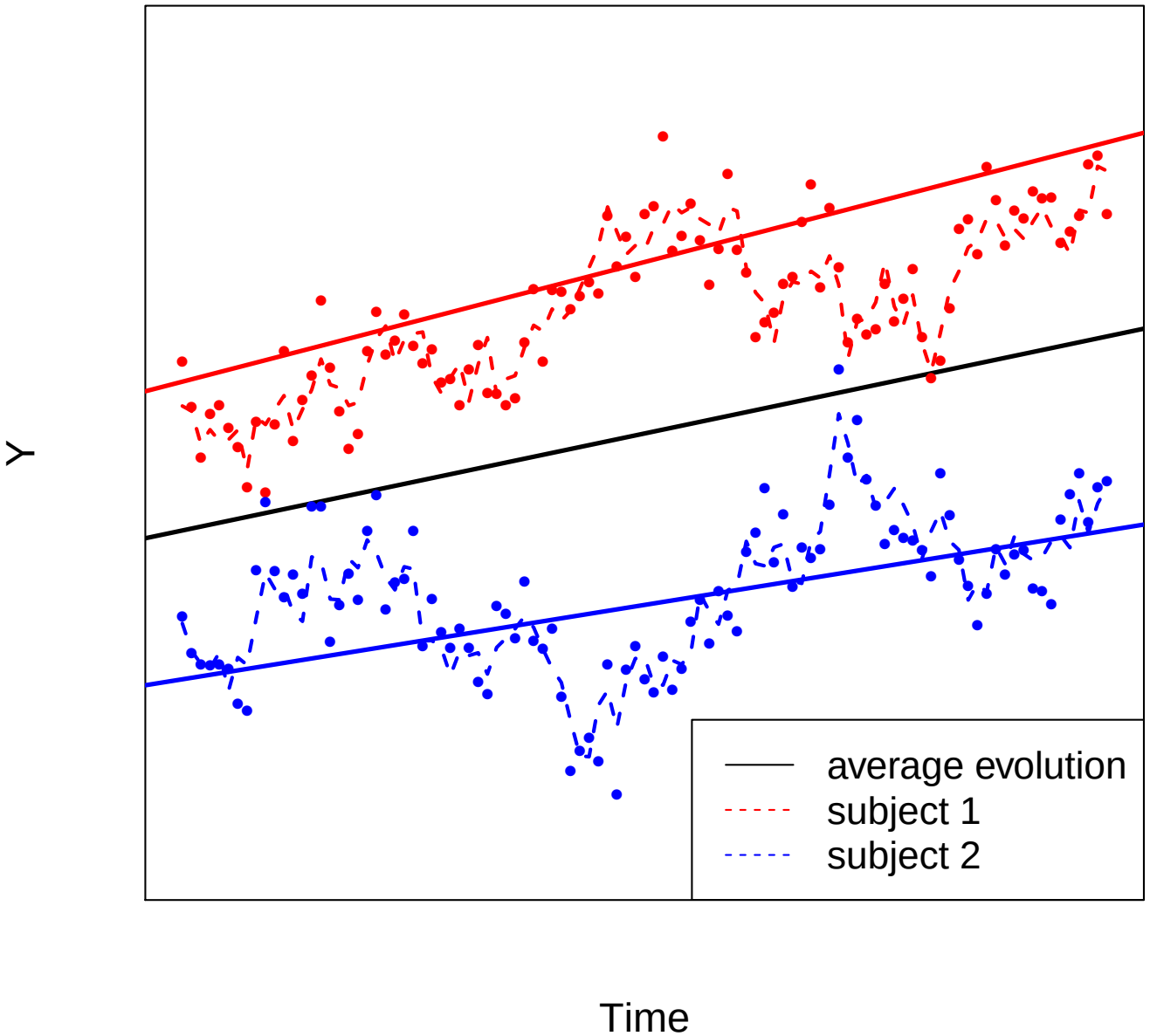
- $\boldsymbol{G}$ und $\boldsymbol{R}$ modellieren die Abhängigkeitsstruktur von $\boldsymbol{b}$ bzw. ε. Zusätzliche Annahmen (z.B. Diagonalmatrix) - unabhängig für $\boldsymbol{G}$ und $\boldsymbol{R}$ - ergeben Modelle verschiedener Komplexität und Flexibilität.

- $\boldsymbol{G}$, $\boldsymbol{R}$ und $\boldsymbol{Z}$ implizieren zusammen die Kovarianzstruktur für $\boldsymbol{y}$, $\boldsymbol{V} = \text{Cov}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}$.

- $\boldsymbol{Z}\boldsymbol{b}$ zusammen mit $\boldsymbol{G}$ modelliert Unterschiede zwischen Beobachtungseinheiten - z.B. zwischen Schülern und Klassen.

- ε ist der Fehlerterm. $\boldsymbol{R}$ fängt möglicherweise verbleibende Autokorrelation auf, die nicht durch $\boldsymbol{Z}\boldsymbol{b}$ erklärt wird.

Die Kovarianzstruktur - longitudinales Beispiel



Conditional Independence Model

Die stärkste Annahme für die Fehler ist, dass sie unabhängig und identisch normalverteilt sind, $\boldsymbol{R} = \sigma^2 \boldsymbol{I_n}$ oder

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n \quad \Leftrightarrow \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n).$$

Aus der Unabhängigkeit der Fehler folgt die bedingte Unabhängigkeit der y_i gegeben $\boldsymbol{b}$, also von $y_i|\boldsymbol{b}, \ldots, y_n|\boldsymbol{b}$.

Die **Korrelation** zwischen den Beobachtungen y_i wird im **Conditional Independence Model** nur durch den Vektor $\boldsymbol{b}$ der **zufälligen Effekten** erzeugt.

Werden die zufälligen Effekte zusätzlich unabhängig angenommen, (bei Longitudinal-/Clusterdaten $\boldsymbol{D} = \text{diag}(\tau_1^2, ..., \tau_q^2)$ diagonal) so spricht man von einem **Varianzkomponentenmodell**.

Spezialfall Random Intercept Modell

Bei Designvektor $\boldsymbol{z}'_{ij} = 1$ ergibt sich das ***Random Intercept Modell***

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + b_i + \varepsilon_{ij}, \quad b_i \overset{iid}{\sim} N(0, \tau^2)$$

mit individuellen Interzepten (Beispiel 100-Meter-Lauf mit $\boldsymbol{x}'_{ij}\boldsymbol{\beta} = \beta_{g_i}$).

In Kombination mit $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2 \boldsymbol{I}_{n_i})$ führt dies zur marginalen Kovarianzstruktur

$$\text{Cov}(y_{ij}, y_{ik}) = \tau^2 + \sigma^2 \delta_{jk}$$

$$\Rightarrow \text{Corr}(y_{ij}, y_{ik}) = \frac{\tau^2}{\sigma^2 + \tau^2} =: \rho \geq 0, \quad j \neq k$$

(mit Kronecker-Delta $\delta_{jk} = 1$ für $j = k$, $\delta_{jk} = 0$ sonst.)
Block-konstante Korrelationsstruktur der Zielvariablen (***Compound Symmetry***).
ρ groß wenn interindividuelle Varianz groß relativ zur intraindividuellen Varianz.

Random Intercept-Random Slope Modell

Beim ***Random Intercept-Random Slope Modell***

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij}, \quad (b_{0i}, b_{1i})' \overset{iid}{\sim} N\left(\mathbf{0}, \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix}\right)$$

unterscheiden sich individuelle Interzepte und Steigungen, z.B. über die Zeit. Bei $\varepsilon_i \sim N(0, \sigma^2 \boldsymbol{I}_{n_i})$ ergibt sich eine quadratische Varianzfunktion:

$$\begin{aligned} \text{Var}(y_{ij}) &= \tau_1^2 + 2\tau_{12}t_{ij} + \tau_2^2 t_{ij}^2 + \sigma^2 \quad \text{und} \\ \text{Cov}(y_{ij}, y_{ik}) &= \tau_1^2 + \tau_{12}t_{ij} + \tau_{12}t_{ik} + \tau_2^2 t_{ij}t_{ik}, \quad j \neq k. \end{aligned}$$

Bei zusätzlichem quadratischen Term $b_{2i}t_{ij}^2$ ergibt sich ein Polynom 4. Ordnung.

Nicht-longitudinales Beispiel: t_{ij} Dosis eines Medikaments - b_{1i} berücksichtigt individuelle Unterschiede in der Reaktion auf das Medikament.

Allgemeines R

Manchmal ist die Annahme $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_n$ zu vereinfachend.

Autokorrelation
Bei Longitudinaldaten ($\rightarrow$ ALD) z.B. wird $\boldsymbol{R}$ block-diagonal gewählt, mit (bei balanzierten Daten) unstrukturierten Kovarianzen $\boldsymbol{\Sigma}_i$ oder

$$\text{Cov}(\varepsilon_{ij}, \varepsilon_{ik}) = \underbrace{\sigma_1^2 \delta_{jk}}_{\text{„Messfehler“}} + \underbrace{\sigma_2^2 g(|t_{ij} - t_{ik}|)}_{\text{Autokorrelation}}$$

für eine monoton fallende Funktion $g(\cdot)$ mit $g(0) = 1$ und $\lim_{u \to \infty} g(u) = 0$.

Häufig nur entweder unabhängiger oder autokorrelierter Teil gut schätzbar.

Heteroskedastizität
Bei den 100-Meter-Lauf-Daten könnte man z.B. zulassen, dass die Varianz $\sigma^2_{g_i}$ der individuellen Zeiten vom Geschlecht abhängt.

Konditionale und marginale Perspektive

Konditionale oder bedingte Perspektive auf das gemischte Modell:

$$\boldsymbol{y}|\boldsymbol{b} \sim N(\boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b}, \boldsymbol{R}), \quad \boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G}). \tag{8}$$

Interpretation: Die zufälligen Effekte sind individuelle Effekte (von Beobachtungseinheiten), die in der Population variieren und unter Normalverteilungsannahme geschätzt werden.

In dieser hierarchischen Formulierung des LMM wird der Erwartungswert von y_i als Funktion von Populationseffekten und individuellen Effekten modelliert.

Marginale Perspektive auf das gemischte Modell:

$$\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \boldsymbol{V}) \qquad \text{mit} \quad \boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R} \tag{9}$$

Interpretation: Die zufälligen Effekte induzieren eine Korrelationsstruktur und ermöglichen so eine valide statistische Analyse korrelierter Daten.

In der marginalen Formulierung des LMM wird der marginale, über die Population gemittelte Erwartungswert von y_i als Funktion von Populationseffekten modelliert.

Konditionale und marginale Perspektive

- Aus der hierarchischen Darstellung (8) folgt die marginale Darstellung (9).
- Bezeichne mit p die Dichten der entsprechenden Verteilungen. Dann ist einfach zu zeigen, dass die Dichte der marginalen Verteilung

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{b})p(\boldsymbol{b})d\boldsymbol{b}$$

  die Dichte einer Normalverteilung (NV) mit EW $\boldsymbol{X}\beta$ und Kovarianz $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}$ ist.
- Im LMM, wenn $\boldsymbol{y}|\boldsymbol{b}$ normalverteilt ist, lässt sich diese Integration analytisch durchführen. Dass dies für andere Verteilungen der Exponentialfamilie nicht geht, ist ein wesentlicher Grund dafür, dass die Inferenz für **generalisierte lineare gemischte Modelle** (GLMMs) schwieriger ist als für LMMs.

Konditionale und marginale Perspektive

- Aus der marginalen Verteilung von $\boldsymbol{y}$ alleine folgt nicht die bedingte Verteilung für $\boldsymbol{y}$ gegeben $\boldsymbol{b}$ und die Verteilung von $\boldsymbol{b}$.
- Das marginale Modell für sich betrachtet nimmt keine zufälligen Effekte an, um Heterogenitäten darzustellen.
- Nicht jede Kovarianz $\boldsymbol{V}$ erlaubt hierarchische Interpretation, $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}$.

Hierarchisches und **marginales Modell** sind also **nicht äquivalent**. Beispiel:

- Random Intercept Modell hierarchisch: τ^2 als Varianz muss nicht-negativ sein.
- Marginal Compound Symmetry: τ^2 als Kovarianz könnte negativ werden.

Dennoch gleiche Interpretation der festen Effekte $\boldsymbol{\beta}$ in hierarchischer und marginaler Formulierung des LMM. Dies ist aber i.A. für GLMMs nicht der Fall! (Mehr dazu in Kapiteln 6+7.)

Konsequenzen für die Schätzung

- Hauptinteresse nur an festen Effekten β
  $\Rightarrow$ Verwendung des marginalen Modells.

- Interesse an festen und zufälligen Effekten β, $\boldsymbol{b}$ sowie den Varianz-/Kovarianzkomponenten in der Kovarianzmatrix $\boldsymbol{D}$
  $\Rightarrow$ Verwendung der hierarchischen Darstellung.

Inhalt

Schätzung der festen Effekte

Für die Schätzung der festen Effekte β verwenden wir die marginale Verteilung (9), $\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \boldsymbol{V})$.

Wir nehmen zunächst an, dass die Kovarianzen $\text{Cov}(\boldsymbol{b}) = \boldsymbol{G}$ und $\text{Cov}(\varepsilon) = \boldsymbol{R}$, und damit $\text{Cov}(\boldsymbol{y}) = \boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R}$, bekannt sind. Dann ist (9) ein allgemeines lineares Modell. Das verallgemeinerte Kleinste-Quadrate (KQ)-Kriterium

$$\min_{\beta} (\boldsymbol{y} - \boldsymbol{X}\beta)' \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\beta)$$

für β ergibt den verallgemeinerten KQ-Schätzer (Aitken-Schätzer)

$$\widehat{\beta} = \left(\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{y}. \tag{10}$$

(Wir nehmen hier an, dass die Inversen $\boldsymbol{V}^{-1}$ und $\left(\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1}$ existieren. Verallgemeinerungen mit generalisierten Inversen existieren.)

Schätzung der festen Effekte

Dies ist gleichzeitig der Maximum-Likelihood (ML)-Schätzer, wenn wir die Normalverteilungsannahme treffen: Die log-Likelihood für β aus dem marginalen Modell (9) ist

$$l(\beta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\beta)'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta). \quad (11)$$

$$\frac{d}{d\beta}l(\beta) = \boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta) \stackrel{!}{=} \boldsymbol{0} \quad \Rightarrow \quad \widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}.$$

Optimalität des Schätzers für β

$\widehat{\beta}$ ist der BLUE, der beste lineare erwartungstreue Schätzer (best linear unbiased estimator) für β. **Beweis:**

Erwartungstreue $\mathrm{E}(\widehat{\beta}) = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\beta = \beta$ ist klar.

Sei $\widetilde{\beta} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{y}$ ein anderer linearer erwartungstreuer Schätzer für β.
$\Rightarrow \mathrm{E}(\widetilde{\beta}) = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}\beta \stackrel{!}{=} \beta \ \forall\beta \quad \Rightarrow \quad \boldsymbol{a} = \boldsymbol{0}$ und $\boldsymbol{B}\boldsymbol{X} = \boldsymbol{I}_p$.
$\Rightarrow$ Für alle $\boldsymbol{c} \in \mathbb{R}^p$ gilt

$$
\begin{aligned}
& \mathrm{Var}(\boldsymbol{c}'\widetilde{\beta}) - \mathrm{Var}(\boldsymbol{c}'\widehat{\beta}) \\
= \ & \boldsymbol{c}'\boldsymbol{B}\boldsymbol{V}\boldsymbol{B}'\boldsymbol{c} - \boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c} \\
= \ & \boldsymbol{c}'\boldsymbol{B}\boldsymbol{V}\boldsymbol{B}'\boldsymbol{c} - \boldsymbol{c}'\boldsymbol{B}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{B}'\boldsymbol{c} \\
= \ & \boldsymbol{c}'\boldsymbol{B}\boldsymbol{V}^{1/2}[\boldsymbol{I}_n - \boldsymbol{V}^{-1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1/2}]\boldsymbol{V}^{1/2}\boldsymbol{B}'\boldsymbol{c} \geq 0,
\end{aligned}
$$

da die mittlere Matrix eine Projektionsmatrix (idempotent) und damit positiv semi-definit ist. Damit ist $\widehat{\beta}$ der lineare erwartungstreue Schätzer mit der kleinsten Varianz. □

Optimalität des Schätzers für β

- Unter der Normalverteilungsannahme kann man sogar zeigen, dass $\widehat{\beta}$ der beste erwartungstreue Schätzer ist.

- Mit der Transformation $\boldsymbol{X}^* = \boldsymbol{V}^{-1/2}\boldsymbol{X}$, $\boldsymbol{y}^* = \boldsymbol{V}^{-1/2}\boldsymbol{y}$, $\varepsilon^* = \boldsymbol{V}^{-1/2}(\boldsymbol{Z}\boldsymbol{b} + \varepsilon)$ gilt $\boldsymbol{y}^* = \boldsymbol{X}^*\beta + \varepsilon^*$ mit $\varepsilon^* \sim N(\boldsymbol{0}, \boldsymbol{I_n})$.

  Somit kann das marginale Modell $\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \boldsymbol{V})$ auf das gewöhnliche lineare Modell $\boldsymbol{y}^* \sim N(\boldsymbol{X}^*\beta, \boldsymbol{I_n})$ zurückgeführt werden.

  Die Optimalitätseigenschaften von $\widehat{\beta}$ ergeben sich dann unmittelbar aus dem Gauß-Markov-Theorem für das gewöhnliche lineare Modell.

Prädiktion der zufälligen Effekte

Oft ist es von Interesse, auch Vorhersagen für die zufälligen Effekte $\boldsymbol{b}$ zu erhalten. Dazu benötigen wir die hierarchische Modellformulierung (8).

Unter der Normalverteilungsannahme ist der bedingte Erwartungswert $\widehat{\boldsymbol{b}} := \mathrm{E}(\boldsymbol{b}|\boldsymbol{y})$ von $\boldsymbol{b}$, gegeben die Daten $\boldsymbol{y}$, die beste lineare erwartungstreue Vorhersage (BLUP, best linear unbiased prediction) für $\boldsymbol{b}$.

- $\widehat{\boldsymbol{b}}$ ist wegen des Satzes vom iterierten Erwartungswert unverzerrt:

$$\mathrm{E}(\widehat{\boldsymbol{b}}) = \mathrm{E}(\mathrm{E}(\boldsymbol{b}|\boldsymbol{y})) = \mathrm{E}(\boldsymbol{b}) = \mathbf{0}.$$

  Für die zufälligen Effekte wird hierbei Unverzerrtheit als $\mathrm{E}(\widehat{\boldsymbol{b}}) = \mathrm{E}(\boldsymbol{b})$ definiert und nicht etwa als $\mathrm{E}(\widehat{\boldsymbol{b}}|\boldsymbol{b}) = \boldsymbol{b}$ für alle $\boldsymbol{b}$.

Prädiktion der zufälligen Effekte

- $\widehat{\boldsymbol{b}} = \mathrm{E}(\boldsymbol{b}|\boldsymbol{y})$ minimiert $\mathrm{E}[(\tilde{\boldsymbol{b}} - \boldsymbol{b})'(\tilde{\boldsymbol{b}} - \boldsymbol{b})]$ in der Klasse der unverzerrten, linearen Schätzer $\tilde{\boldsymbol{b}}$:

$$\begin{aligned}
\mathrm{E}[(\tilde{\boldsymbol{b}} - \boldsymbol{b})'(\tilde{\boldsymbol{b}} - \boldsymbol{b})] &= \mathrm{E}[(\tilde{\boldsymbol{b}} - \widehat{\boldsymbol{b}} + \widehat{\boldsymbol{b}} - \boldsymbol{b})'(\tilde{\boldsymbol{b}} - \widehat{\boldsymbol{b}} + \widehat{\boldsymbol{b}} - \boldsymbol{b})] \\
&= \mathrm{E}[(\tilde{\boldsymbol{b}} - \widehat{\boldsymbol{b}})'(\tilde{\boldsymbol{b}} - \widehat{\boldsymbol{b}})] + 2\,\mathrm{E}[(\tilde{\boldsymbol{b}} - \widehat{\boldsymbol{b}})'(\widehat{\boldsymbol{b}} - \boldsymbol{b})] \\
&\quad + \mathrm{E}[(\widehat{\boldsymbol{b}} - \boldsymbol{b})'(\widehat{\boldsymbol{b}} - \boldsymbol{b})] \\
&= \mathrm{E}[(\tilde{\boldsymbol{b}} - \widehat{\boldsymbol{b}})'(\tilde{\boldsymbol{b}} - \widehat{\boldsymbol{b}})] + \mathrm{E}[(\widehat{\boldsymbol{b}} - \boldsymbol{b})'(\widehat{\boldsymbol{b}} - \boldsymbol{b})] \\
&\geq \mathrm{E}[(\widehat{\boldsymbol{b}} - \boldsymbol{b})'(\widehat{\boldsymbol{b}} - \boldsymbol{b})],
\end{aligned}$$

da $\mathrm{E}[(\tilde{\boldsymbol{b}} - \widehat{\boldsymbol{b}})'(\widehat{\boldsymbol{b}} - \boldsymbol{b})] = \mathrm{E}_{\boldsymbol{y}}\,\mathrm{E}_{\boldsymbol{b}|\boldsymbol{y}}[(\tilde{\boldsymbol{b}} - \widehat{\boldsymbol{b}})'(\mathrm{E}(\boldsymbol{b}|\boldsymbol{y}) - \boldsymbol{b})] = \boldsymbol{0}$.

Gleichheit gilt für $\tilde{\boldsymbol{b}} = \widehat{\boldsymbol{b}}$.

Prädiktion der zufälligen Effekte

Um $\widehat{\boldsymbol{b}}$ explizit zu bestimmen, betrachte die gemeinsame Verteilung von $\boldsymbol{y}$ und $\boldsymbol{b}$. Diese ergibt sich aus dem zweistufigen hierarchischen Modell für $\boldsymbol{y}|\boldsymbol{b}$ und $\boldsymbol{b}$ zu

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{b} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{X}\beta \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{V} & \boldsymbol{ZG} \\ \boldsymbol{GZ}' & \boldsymbol{G} \end{pmatrix} \right). \tag{12}$$

Herleitung für Interessierte: $\rightarrow$ **Extrablatt**.

Klar: Aus (12) folgt wiederum (Satz B.4, Fahrmeir et al., 2009)

- $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G})$
- $\boldsymbol{y}|\boldsymbol{b} \sim N(\boldsymbol{X}\beta + \boldsymbol{Zb}, \boldsymbol{V} - \boldsymbol{ZGG}^{-1}\boldsymbol{GZ}' = \boldsymbol{R})$.

Aus (12) folgt (Satz B.4, Fahrmeir et al., 2009) der bedingte Erwartungswert

$$\mathrm{E}(\boldsymbol{b}|\boldsymbol{y}) = \boldsymbol{GZ}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta). \tag{13}$$

In der Praxis wird der unbekannten Vektor β in (13) durch den verallgemeinerten KQ-Schätzer $\widehat{\beta}$ aus (10) ersetzt,

$$\widehat{\boldsymbol{b}} = \boldsymbol{GZ}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\beta}) = (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\beta}).$$

Henderson's mixed model equations

BLUE und BLUP (gemeinsam auch mit BLUP bezeichnet) $\widehat{\boldsymbol{\beta}}$ und $\widehat{\boldsymbol{b}}$ sind die Lösung der sogenannten Henderson's mixed model equations

$$\begin{aligned} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\widehat{\boldsymbol{b}} &= \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\widehat{\boldsymbol{b}} &= \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{aligned} \tag{14}$$

Für diese Schätzgleichungen gibt es mehrere Herleitungen, u.a. folgende drei:

1. $\widehat{\boldsymbol{\beta}}$ und $\widehat{\boldsymbol{b}}$ haben minimale erwartete quadratische Abweichung unter den linearen erwartungstreuen Schätzern/Vorhersagen.
   (Analog: Linearkombinationen, z.B. $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{b}}$.)
2. Eine bayesianische Herleitung (mehr in Kapitel 4).
3. Die Schätzgleichungen ergeben sich bei simultaner Maximierung (bzgl. $\boldsymbol{\beta}$ und $\boldsymbol{b}$) der sogenannten **gemeinsamen Likelihood** $L(\boldsymbol{\beta}, \boldsymbol{b})$ basierend auf der gemeinsamen Dichte von $\boldsymbol{y}$ und $\boldsymbol{b}$.

Gemeinsame / penalisierte Likelihood

Wegen $p(\boldsymbol{y}, \boldsymbol{b}) = p(\boldsymbol{y}|\boldsymbol{b})p(\boldsymbol{b})$ ergibt sich (bis auf additive Konstanten)

$$l(\beta, \boldsymbol{b}) = \log L(\beta, \boldsymbol{b}) = -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b})' \boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}' \boldsymbol{G}^{-1}\boldsymbol{b}. \quad (15)$$

Diese kann als penalisierte Likelihood aufgefasst werden mit Strafterm $\boldsymbol{b}' \boldsymbol{G}^{-1}\boldsymbol{b}$.
Die Maximierung ist äquivalent zur Minimierung des penalisierten KQ-Kriteriums

$$KQ_{pen}(\beta, \boldsymbol{b}) = (\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b})' \boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b}) + \boldsymbol{b}' \boldsymbol{G}^{-1}\boldsymbol{b}. \quad (16)$$

- Der erste Term entspricht einem verallgemeinerten KQ-Kriterium.
- Der zweite Term berücksichtigt die Verteilungsannahme für $\boldsymbol{b}$ und bestraft ($\boldsymbol{b}' \boldsymbol{G}^{-1}\boldsymbol{b} \geq 0$ wegen $\boldsymbol{G} \geq 0$ als Kovarianz) Abweichungen von $\mathrm{E}(\boldsymbol{b}) = \boldsymbol{0}$.
  - Für $\boldsymbol{G} = \boldsymbol{0}$ gilt $\widehat{\boldsymbol{b}} = \boldsymbol{0}$.
  - Für $\boldsymbol{G}^{-1} \to \boldsymbol{0}$ geht $\boldsymbol{b}' \boldsymbol{G}^{-1}\boldsymbol{b} \to 0$ und $\boldsymbol{b}$ wird wie ein fester Effekt geschätzt.

Differenzieren von $KQ_{pen}(\beta, \boldsymbol{b})$ und Nullsetzen der Ableitungen ergibt (14).

Definiert man $\boldsymbol{C} = (\boldsymbol{X}|\boldsymbol{Z})$ und die partitionierte Matrix

$$\boldsymbol{A} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G}^{-1} \end{pmatrix}, \tag{17}$$

so lässt sich die Lösung der Schätzgleichung (14) auch in der folgenden kompakten Form schreiben

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{b}} \end{pmatrix} = (\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{A})^{-1}\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{y}. \tag{18}$$

In dieser Form erkennt man die enge Beziehung zur Ridge-Schätzung.
Ebenso besteht ein enger Zusammenhang zur (empirischen) Bayes-Schätzung (mehr dazu später).

Shrinkage

$\widehat{\boldsymbol{b}}$ ist ein Shrinkage-Schätzer für $\boldsymbol{b}$, d.h. seine Komponenten haben eine geringere Streuung als sie hätten, wenn $\boldsymbol{b}$ als feste Effekte behandelt würden.

Dies scheint auf den ersten Blick der Eigenschaft der Unverzerrtheit des BLUP zu widersprechen. Erwartungstreue bedeutet hier jedoch

$$\mathrm{E}(\widehat{\boldsymbol{b}}) = \mathrm{E}(\boldsymbol{b}) = \mathbf{0},$$

und nicht

$$\mathrm{E}(\widehat{\boldsymbol{b}}|\boldsymbol{b}) = \boldsymbol{b} \text{ für alle } \boldsymbol{b}.$$

Shrinkage

$$
\begin{aligned}
\widehat{\boldsymbol{y}} &= \boldsymbol{X}\widehat{\beta} + \boldsymbol{Z}\widehat{\boldsymbol{b}} \\
&= \boldsymbol{X}\widehat{\beta} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\beta}) \\
&= (\boldsymbol{V} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}')\boldsymbol{V}^{-1}\boldsymbol{X}\widehat{\beta} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{y} \\
&= \boldsymbol{R}\boldsymbol{V}^{-1}\boldsymbol{X}\widehat{\beta} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{y}.
\end{aligned}
$$

Der BLUP für $\boldsymbol{y}$ ist also ein gewichtetes Mittel des Populationsmittels $\boldsymbol{X}\widehat{\beta}$ und der individuellen Daten $\boldsymbol{y}$. Die beobachteten Daten werden gegen das Populationsmittel „geschrumpft“ (“borrowing of strength”).

Beachte, dass $\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$.

- Wenn die Varianz der Residuen $\boldsymbol{R}$ im Verhältnis zu $\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$ groß ist, wird viel Gewicht auf das Populationsmittel $\boldsymbol{X}\widehat{\beta}$ gelegt.
- Wenn $\boldsymbol{R}$ klein ist, gilt das Gegenteil.

Shrinkage: Varianz-Bias-Tradeoff

Beispiel Random-Intercept-Modell (ohne Kovariablen)

$$y_{ij} = b_i + \varepsilon_{ij}, \quad b_i \overset{iid}{\sim} N(0, \tau^2),\ \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2),$$
$$j = 1, \ldots, n_i, i = 1, \ldots, N.$$

Für den BLUP $\hat{\boldsymbol{b}}^{BLUP}$ und den KQ-Schätzer $\hat{\boldsymbol{b}}^{KQ}$, der $\boldsymbol{b}$ als feste Effekte schätzt, gilt

	$\hat{\boldsymbol{b}}^{BLUP}$	$\hat{\boldsymbol{b}}^{KQ}$	
$\hat{b}_i$			
$\mathrm{E}(\hat{b}_i \mid b_i)$			
$\mathrm{Var}(\hat{b}_i \mid b_i)$			
$\mathrm{MSE}(\hat{b}_i)$			

$$\mathrm{MSE}(\hat{b}_i) = \mathrm{E}_b\, \mathrm{E}_{y|b}[(\hat{b}_i - b_i)^2] = \mathrm{E}_b[\mathrm{Var}(\hat{b}_i|b_i) + (E(\hat{b}_i|b_i) - b_i)^2].$$

Schätzung der Kovarianzstruktur

Sei ϑ der Parametervektor, der alle unbekannten Parameter in den Kovarianzmatrizen $\boldsymbol{R} = \boldsymbol{R}(\vartheta)$, $\boldsymbol{G} = \boldsymbol{G}(\vartheta)$ und $\boldsymbol{V} = \boldsymbol{V}(\vartheta)$ enthält.

Indem man einen (konsistenten) Schätzer $\widehat{\vartheta}$ einsetzt, erhält man die geschätzten Kovarianzmatrizen

$$\widehat{\boldsymbol{R}} = \boldsymbol{R}(\widehat{\vartheta}), \quad \widehat{\boldsymbol{G}} = \boldsymbol{G}(\widehat{\vartheta}), \quad \widehat{\boldsymbol{V}} = \boldsymbol{V}(\widehat{\vartheta}),$$

die für die Berechnung der BLUPs von β und $\boldsymbol{b}$ verwendet werden können.

Maximum-Likelihood (ML)-Schätzung von ϑ

Die **ML-Schätzung** von ϑ basiert auf der Likelihood des marginalen Modells

$$\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \boldsymbol{V}(\vartheta)).$$

Die log-Likelihood für β und ϑ ist bis auf additive Konstanten gegeben durch

$$l(\beta, \vartheta) = -\frac{1}{2}\left\{\log|\boldsymbol{V}(\vartheta)| + (\boldsymbol{y} - \boldsymbol{X}\beta)'\boldsymbol{V}(\vartheta)^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta)\right\}.$$

Maximiert man $l(\beta, \vartheta)$ für festes ϑ bezüglich β, so erhält man (wieder)

$$\widehat{\beta}(\vartheta) = \arg\max_{\beta}\left(l(\beta, \vartheta)\right) = \left(\boldsymbol{X}'\boldsymbol{V}(\vartheta)^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}(\vartheta)^{-1}\boldsymbol{y}.$$

Einsetzen von $\widehat{\beta}(\vartheta)$ in $l(\beta, \vartheta)$ liefert die nur von ϑ abhängige Profil-log-Likelihood

$$l_P(\vartheta) = l(\widehat{\beta}(\vartheta), \vartheta) = -\frac{1}{2}\left\{\log|\boldsymbol{V}(\vartheta)| + (\boldsymbol{y} - \boldsymbol{X}\widehat{\beta}(\vartheta))'\boldsymbol{V}(\vartheta)^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\beta}(\vartheta))\right\}.$$

Maximierung von $l_P(\vartheta)$ bezüglich ϑ liefert den ML-Schätzer $\widehat{\vartheta}_{ML}$.

Motivation REML-Schätzung

ML-Schätzer für Varianzen sind (nach unten) verzerrt, da der Verlust von Freiheitsgraden durch die Schätzung von $\widehat{\boldsymbol{\beta}}$ unberücksichtigt bleibt.

Beispiel lineares Modell $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I_n})$

- ML-Schätzer ist $\widehat{\sigma}^2 = \frac{1}{n}\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}$, aber $\mathrm{E}(\widehat{\sigma}^2) = \frac{n-p}{n}\sigma^2$.
- Benutzt wird üblicherweise der erwartungstreue Schätzer $\widehat{\sigma}^2 = \frac{1}{n-p}\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}$.

  Dies ist der restringierte Maximum Likelihood-Schätzer. Er berücksichtigt den Verlust an Freiheitsgraden durch Schätzung von $\boldsymbol{\beta}$.

REstringierte Maximum-Likelihood (REML) Schätzung von ϑ

Grundidee: Basiere Likelihood nicht auf $\boldsymbol{y}$, sondern auf $(n-p)$ linear unabhängigen Fehlerkontrasten $\boldsymbol{Ay}$, deren Verteilung von β unabhängig ist.

Teile die log-Likelihood in zwei Teile auf für $\boldsymbol{Ay}$ (für ϑ) und $\boldsymbol{By}$ (für β) mit

1. $\boldsymbol{A}$ von Rang $n-p$ und $\boldsymbol{B}$ von Rang p.
2. Die zwei Teile sind statistisch unabhängig, also hier $\text{Cov}(\boldsymbol{Ay}, \boldsymbol{By}) = \boldsymbol{0}$.
3. $\boldsymbol{Ay}$ sind Fehlerkontraste, d.h. $\text{E}(\boldsymbol{Ay}) = \boldsymbol{0}$.
4. $\boldsymbol{BX}$ hat Rang p ($\Rightarrow \boldsymbol{By}$ schätzt eine eineindeutige Funktion von β).

Geeignete Matrizen: $\boldsymbol{A} = \boldsymbol{I_n} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}$, $\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}$.

$\rightarrow$ Englischer Name residual oder restricted maximum likelihood estimation.

Aus den ersten zwei Punkten folgt, dass die log-Likelihood $l = l' + l^*$ ist (bis auf additive Konstanten), l' für $\boldsymbol{Ay}$ und l^* für $\boldsymbol{By}$.

Maximierung von l^* (aus $\boldsymbol{By}$) bzgl. β ergibt den MLE/BLUE $\widehat{\beta}$ (abhängig von ϑ).

Maximierung von l' (aus $\boldsymbol{Ay}$) bzgl. ϑ: Die restringierte log-Likelihood ist (bis auf additive Konstanten) unabhängig von dem gewählten Fehlerkontrast

$$\begin{aligned} l_R(\vartheta) &= -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\widehat{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\beta}) \\ &= l(\widehat{\beta}(\vartheta),\vartheta) - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}| \end{aligned} \tag{19}$$

Herleitung für Interessierte: → **Extrablatt**
Maximierung ergibt den restringierten ML- (REML)-Schätzer $\widehat{\vartheta}_{REML}$ für ϑ.

Alternative Herleitung des REML-Schätzers

Alternativ maximiert der REML-Schätzer $\widehat{\boldsymbol{\vartheta}}_{REML}$ die marginale log-Likelihood für $\boldsymbol{\vartheta}$, bei der $\boldsymbol{\beta}$ aus der Likelihood $L(\boldsymbol{\beta}, \boldsymbol{\vartheta})$ des marginalen Modells herausintegriert wird

$$l_R(\boldsymbol{\vartheta}) = \log \left(\int L(\boldsymbol{\beta}, \boldsymbol{\vartheta}) d\boldsymbol{\beta} \right). \tag{20}$$

Diese ergibt sich bis auf additive Konstanten wieder als (19). → **Übung**

Auch beim LMM ist die Reduktion der Verzerrung von $\widehat{\boldsymbol{\vartheta}}_{ML}$ der Hauptgrund für die Verwendung von $\hat{\boldsymbol{\vartheta}}_{REML}$ als Schätzer für $\boldsymbol{\vartheta}$. Es ist jedoch allgemein nicht gesichert, ob auch der mean squared error (MSE) geringer wird.

Empirische Versionen von BLUP und BLUE

Mit den durch **(RE)ML** geschätzten Kovarianzmatrizen $\widehat{\boldsymbol{R}}$, $\widehat{\boldsymbol{G}}$ und $\widehat{\boldsymbol{V}}$ ergeben sich empirische Versionen von BLUP und BLUE, EBLUP und EBLUE, für $\boldsymbol{b}$ und $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{y}, \qquad \widehat{\boldsymbol{b}} = \widehat{\boldsymbol{G}}\boldsymbol{Z}'\widehat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}).$$

- Im Gegensatz zum linearen Modell sind die Schätzer der festen Effekte $\boldsymbol{\beta}$ im LMM abhängig von der Kovarianzmatrix $\boldsymbol{V}$. Daher sind die Schätzer $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\vartheta}}_{REML})$ und $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\vartheta}}_{ML})$ **nicht identisch**.
- Nach Einsetzen von $\widehat{\boldsymbol{\vartheta}}$ gelten die Optimalitätseigenschaften nicht mehr exakt.
- Ebenso sind die Varianzen bzw. Kovarianzmatrizen der Schätzer analytisch nicht exakt zugänglich (mehr später).

Numerische Berechnung der Schätzer

- Die Maximierung von $l_R(\vartheta)$ $[l_P(\vartheta)]$ zur Berechnung von $\hat{\vartheta}_{REML}$ $[\hat{\vartheta}_{ML}]$ erfolgt wegen Nichtlinearität in ϑ meist per Newton-Raphson oder Fisher-Scoring.
- Dabei können diese Algorithmen keine Restriktion von Parametern auf $[0, \infty)$ berücksichtigen, sodass negative Schätzer für Varianzen entstehen könnten.
- Daher maximiert `lme()` in R-Paket `nlme` die (restringierte) log-Likelihood bzgl. der skalierten log-Varianzen. Allerdings kann diese Methode ein Maximum in der Null (kein zufälliger Effekt) nicht finden.
- `lmer()` in R-Paket `lme4` verwendet eine Parametrisierung, die über eine Nebenbedingung positiv (semi-)definites $\boldsymbol{G}$ sicherstellt und singuläres $\boldsymbol{G}$ erlaubt. Die Funktion verwendet Optimierung mit Nebenbedingungen statt Newton-Raphson.

Numerische Berechnung der Schätzer

- Die Bedingungen an ϑ unterscheiden sich zwischen marginalem ($\boldsymbol{V}$ positiv (semi-)definit) und konditionalem Modell ($\boldsymbol{G}$ und $\boldsymbol{R}$ positiv (semi-)definit). Softwarepakete maximieren bzgl. ϑ zum Teil über einen größeren Parameterraum als das konditionale Modell implizieren würde.
- Die historisch erste Alternative ist der EM-Algorithmus, der die zufälligen Effekte als fehlende Daten behandelt. Die Konvergenz kann langsam sein.

Likelihood-Schätzung in R

Das klassische Paket zur LMM-Schätzung in R ist `nlme`. Eine ausführliche Beschreibung findet man im Buch der Autoren, Pinheiro & Bates (2000).

Der iterative Optimierungsalgorithmus ist eine Hybridversion aus zunächst **EM-** und im Anschluss **Newton-Raphson(NR)-Algorithmus**.

- **EM-Algorithmus**: Iterationen schnell und einfach zu berechnen. Schnell nahe ans Optimum, Konvergenz nahe des Optimums potentiell sehr langsam.
- **NR-Algorithmus**: Iterationen sehr rechenintensiv. (Berechnung 1. und 2. Ableitung, Scorefunktion, Hessematrix, in jedem Schritt.) Möglicherweise instabil entfernt vom Optimum. Sehr schnelle Konvergenz nahe des Optimums.

Alternativen

- Für additive und generalisierte additive gemischte Modelle (mehr dazu in Kapitel 5) eignet sich das R-Paket `mgcv`, siehe auch Wood (2006).
- `lme4` ist eine neuere Weiterentwicklung von Douglas Bates und Koautoren, das lineare und generalisierte lineare gemischte Modelle fitten kann (unter Benutzung von S4-Klassen). Siehe auch Bates, Maechler, Bolker & Walker (2015). *Fitting Linear Mixed-Effects Models using lme4.* Journal of Statistical Software, angenommen. `http://arxiv.org/abs/1406.5823`.
- Viele Beispiele für lineare gemischte Modelle mit `SAS proc mixed` finden sich in Verbeke & Molenberghs, 2000.

Inhalt

Schätzung der Kovarianzmatrix der Schätzer

Bei bekannter Kovarianzstruktur ist wegen (18) $\rightarrow$ **Übung**

$$\text{Cov}\begin{pmatrix} \widehat{\beta} \\ \widehat{\boldsymbol{b}} - \boldsymbol{b} \end{pmatrix} = (\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{A})^{-1}. \tag{21}$$

Beachte, dass $\text{Cov}(\widehat{\boldsymbol{b}} - \boldsymbol{b})$ statt $\text{Cov}(\widehat{\boldsymbol{b}})$ verwendet wird, was den Bias durch Shrinkage und die Variabilität in $\boldsymbol{b}$ berücksichtigt.

Geschätzte Versionen mit eingesetztem $\widehat{\boldsymbol{R}}$ und $\widehat{\boldsymbol{A}}$ können zur Konstruktion approximativer Konfidenz- oder Vorhersageintervalle für Ausdrücke in β und $\boldsymbol{b}$ verwendet werden (z.B. für neue Beobachtungen $\boldsymbol{y}_i$); die Diagonalelemente ergeben geschätzte Varianzen für die Komponenten von $\widehat{\beta}$ bzw. $\widehat{\boldsymbol{b}}$.

Beachte, dass die Variabilität durch die Schätzung der Kovarianzparameter nicht berücksichtigt wird und die tatsächlichen Varianzen daher unterschätzt werden.

Schätzung der Kovarianzmatrix der Schätzer

Manchmal wird auch eine Alternative verwendet, bei der $\boldsymbol{b}$ wie ein fester Effekt behandelt wird, indem man auf $\boldsymbol{b}$ bedingt. Dann erhält man $\rightarrow$ **Übung**

$$\text{Cov}\left(\left.\begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{b}}\end{pmatrix}\right|\boldsymbol{b}\right) = \left(\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C}+\boldsymbol{A}\right)^{-1}\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C}\left(\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C}+\boldsymbol{A}\right)^{-1}. \quad (22)$$

Diese Kovarianzmatrix ist von der Sandwich-Matrix-Form, da sie wie ein Sandwich aus $(\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C}+\boldsymbol{A})^{-1}$ und $\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C}$ zusammengesetzt ist.

Nach Einsetzen der Schätzer $\widehat{\boldsymbol{R}}$, $\widehat{\boldsymbol{A}}$ erhält man die geschätzte Kovarianzmatrix.

Wir präferieren (21). (22) führt zu kleineren Standardfehlern,

$$\begin{aligned}&\text{Cov}\begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{b}}-\boldsymbol{b}\end{pmatrix} - \text{Cov}\left(\begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{b}}\end{pmatrix}|\boldsymbol{b}\right)\\ &= \left(\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C}+\boldsymbol{A}\right)^{-1}\boldsymbol{A}\left(\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C}+\boldsymbol{A}\right)^{-1} \geq \boldsymbol{0}.\end{aligned}$$

Tests für die festen Effekte

Mit den geschätzten Kovarianzmatrizen lassen sich Wald-Tests für

$$H_0 : \boldsymbol{L}\beta = \mathbf{0} \text{ gegen } H_A : \boldsymbol{L}\beta \neq \mathbf{0} \tag{23}$$

oder Konfidenzintervalle konstruieren. Dazu benutzen wir, dass

$$T_W = (\widehat{\beta} - \beta)' \boldsymbol{L}' [\boldsymbol{L} \operatorname{Cov}(\widehat{\beta}) \boldsymbol{L}']^{-1} \boldsymbol{L} (\widehat{\beta} - \beta)$$

asymptotisch $\chi^2_{\text{Rang}(\boldsymbol{L})}$-verteilt ist. Es können auch robuste Kovarianzmatrix-Schätzer verwendet werden, siehe z.B. Liang & Zeger (1986). *Longitudinal Data Analysis Using Generalized Linear Models*. Biometrika, 73 (1), 13-22.

Wichtig: Alle asymptotischen Ergebnisse in diesem Kapitel gelten in LMMs für Longitudinal- oder Cluster-Daten, in denen $\boldsymbol{y}$ aus N unabhängigen Teilvektoren $\boldsymbol{y}_i$ besteht, für $N \to \infty$. Resultate für allgemeine LMMs sind schwierig zu zeigen (vgl. die Diskussion in Ruppert, Wand & Carroll (2003), Kapitel 4.8).

Die Wald-Teststatistik basiert auf Standardfehlern (21), die die wahre Variabilität in $\widehat{\boldsymbol{\beta}}$ durch Vernachlässigung der Schätzung von $\boldsymbol{\vartheta}$ unterschätzen.

Dieses Problem wird häufig durch approximative *t*- oder *F*-Statistiken verringert:

- Für einen einzelnen Parameter β_j in $\boldsymbol{\beta}$ wird die Verteilung von $(\widehat{\beta}_j - \beta_j)/\widehat{SE}(\widehat{\beta}_j)$ durch eine *t*-Verteilung approximiert.
- Für generelle Hypothesen (23) wird die Verteilung von

$$F = \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \boldsymbol{L}' [\boldsymbol{L} \, \mathrm{Cov}(\widehat{\boldsymbol{\beta}}) \boldsymbol{L}']^{-1} \boldsymbol{L} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\mathrm{Rang}(\mathbf{L})}$$

  durch eine *F*-Verteilung mit Rang($\mathbf{L}$) Zähler-Freiheitsgraden approximiert.

Die Freiheitsgrade für die t-Verteilung bzw. die Nenner-Freiheitsgrade für die F-Verteilung werden aus den Daten geschätzt. Häufig wird die Satterthwaite-Approximation verwendet, die die Momente der Verteilungen matcht.

Für kleine Stichproben gibt die Methode nach Kenward und Rogers bessere Ergebnisse. Diese berücksichtigt die zusätzliche Unsicherheit durch Schätzung von ϑ.

Die Berechnung von Satterthwaite-artigen Freiheitsgraden ist recht rechenintensiv.

In R: R-Paket `pbkrtest`, siehe Halekoh & Højsgaard (2014), A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest, Journal of Statistical Software, 59 (9).

Alternativ können wir den Likelihood-Quotienten-Test (LQT) verwenden. Bei nicht so großem n gibt der LQT meist genauere Ergebnisse als der Wald-Test.

Die Teststatistik des LQT

$$T_{LQT} = 2 \sup_{H_A} l(\boldsymbol{\beta}, \boldsymbol{\vartheta}) - 2 \sup_{H_0} l(\boldsymbol{\beta}, \boldsymbol{\vartheta})$$

ist asymptotisch unter der Nullhypothese $\chi^2_{\text{Rang}(\boldsymbol{L})}$-verteilt. Voraussetzungen:

- genestete Modelle
- gleiche Kovarianzstruktur in jedem Modell
- keine REML-Schätzung! (Bei verschiedenen festen Effekten unter H_0 und H_A unterscheiden sich die Fehlerkontraste $\boldsymbol{A}_0\boldsymbol{y}$ und $\boldsymbol{A}_A\boldsymbol{y}$ $\Rightarrow$ die Likelihoods für $\boldsymbol{A}_0\boldsymbol{y}$ und $\boldsymbol{A}_A\boldsymbol{y}$ sind nicht vergleichbar.)

Beachte außerdem, dass die Approximation bei kleinem $n - p$ schlecht sein kann.

Tests für die (Ko-)Varianzkomponenten in ϑ

Bei Longitudinal- oder Clusterdaten mit

$$\boldsymbol{D} = \begin{pmatrix} d_{11} & \dots & d_{1q} \\ \vdots & \ddots & \vdots \\ d_{1q} & \dots & d_{qq} \end{pmatrix} = \left(\begin{array}{ccc} & \boldsymbol{D}_1 & \\ & & \\ \hline d_{1q} & \dots & d_{qq} \end{array}\right)$$

könnte z.B. ein Test für

$$H_0: \quad \boldsymbol{D} = \left(\begin{array}{ccc} & \boldsymbol{D}_1 & \\ & & \\ \hline 0 & \dots & 0 \end{array}\right) \begin{array}{c} 0 \\ \vdots \\ 0 \end{array} \right) \text{ mit } \boldsymbol{D}_1 \text{ positiv-definit } (q-1 \times q-1) \quad \text{gegen}$$

$$H_A: \quad \boldsymbol{D} \text{ positiv-semidefinit } (q \times q) \tag{24}$$

von Interesse sein. Dies entspricht einem Test für den $q$ten zufälligen Effekt, z.B.

- H_0: Random Intercept gegen H_A: Random Intercept und Slope ($q = 2$)
- H_0: keine zufälligen Effekte gegen H_A: Random Intercept ($q = 1$).

Asymptotische Verteilung des LQT

(24) verletzt die Standardannahme der getesteten Parameter im Innern des Parameterraums. Z.B. $d_{qq} \geq 0$ als Varianz und $d_{qq} = 0$ (unter H_0) daher am Rand des Parameterraums. (Generell: komplizierte Restriktionen an $\boldsymbol{D} \geq 0$ und $\boldsymbol{D}_1 \geq 0$.)

Für Longitudinal- oder Cluster-Daten mit $N \to \infty$ hat T_{LQT} für (24) asymptotisch unter H_0 eine 0.5:0.5-Mischungsverteilung aus zwei χ^2_{q-1}- und χ^2_q-Verteilungen (χ^2_0-Verteilung = Punktmasse an der Null).

Für komplexere Hypothesen für $\boldsymbol{D}$ können komplexere χ^2-Mischungen entstehen.

Diese Mischungsverteilung gilt nicht, wenn ein Nuisance-Parameter (Störparameter) am Rand des Parameterraums liegt, also z.B. die Varianz der Random Intercepts (nahe) Null ist, während die Varianz der Random Slopes getestet wird.

Exakte Verteilung des LQT

Kann $\boldsymbol{y}$ nicht in viele unabhängige Teilvektoren $\boldsymbol{y}_i, i = 1, \ldots, N$ unterteilt werden, so approximiert die Mischungsverteilung die Verteilung von T_{LQT} meist schlecht.

Die exakte Verteilung von T_{LRT} für $H_0 : \tau_1^2 = 0$ im allgemeinen LMM mit $\boldsymbol{G} = \tau_1^2 \boldsymbol{I}_{q_1}$ und $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_n$ ist im R-Paket `RLRsim` (F. Scheipl) implementiert. Approximationen für $\boldsymbol{G} = \text{diag}(\tau_1^2 \boldsymbol{I}_{q_1}, \ldots, \tau_s^2 \boldsymbol{I}_{q_s})$, $s > 1$, beruhen auf

- Greven, S., Crainiceanu, C., Küchenhoff, H. & Peters, A. (2008). *Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models.* JCGS, 17 (4): 870-891.
- Scheipl, F., Greven, S. & Küchenhoff, H. (2008). *Size and Power of Tests for a Zero Random Effect Variance or Polynomial Regression in Additive and Linear Mixed Models.* CSDA, 52 (7): 3283-3299.

Wenn nur Parameter in ϑ getestet werden, kann auch ein restringierter LQT basierend auf der REML-Schätzung verwendet werden. Die Asymptotik ist entsprechend, die exakte Verteilung ist ebenfalls in `RLRsim` implementiert.

Modellselektion

AIC- und BIC-Werte in Modelloutputs, z.B. von `lme()` in R, basieren meist auf der marginalen Likelihood (9), z.B. das Akaike Informationskriterium (AIC)

$$AIC = -2 \log l(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\vartheta}}) + 2(p + d),$$

mit d die Anzahl der Parameter in $\boldsymbol{\vartheta}$. Ein AIC basierend auf der marginalen restringierten log-Likelihood kann analog definiert werden.

Die Herleitung des AIC nimmt unabhängig und identisch verteilte Beobachungen sowie einen Parameterraum $\mathbb{R}^{p+d}$ an. Bei der Selektion von Parametern in $\boldsymbol{\vartheta}$ (mit Restriktionen, z.B. Varianzen aus $[0, \infty)$) führt dies zu einem Bias und tendenziell zu kleineren Modellen ohne zufällige Effekte.

Modellselektion

Neuere Ansätze verwenden für die Selektion von zufälligen Effekten ein AIC basierend auf der konditionalen log-Likelihood für $\boldsymbol{y}|\boldsymbol{b} \sim N(\boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b}, \boldsymbol{R})$. Die korrekten Freiheitsgrade werden hergeleitet in

- Greven, S. und Kneib, T. (2010). *On the Behaviour of Marginal and Conditional AIC in Linear Mixed Models.* Biometrika, 97(4): 773-789

und das konditionale AIC ist implementiert im R-Paket `cAIC4`.

Inhalt

Wiederholung: Bayes-Inferenz

- Parameter $\theta \in \mathbf{\Theta}$ nicht deterministisch, sondern als zufällig angenommen.
- Volles Wahrscheinlichkeitsmodell für alle beobachteten und unbeobachteten Größen bestehend aus:
  1. ***Beobachtungsmodell:*** bedingte Verteilung der Daten gegeben unbekannte Parameter θ, $p(\boldsymbol{y}|\theta)$.
  2. ***Priori-Verteilung*** $p(\theta)$: drückt Vorwissen/Annahmen über θ aus.
- Mathematisch günstig: zum Beobachtungsmodell ***konjugierte Prioris***. (Prioris und Posterioris in der gleichen Verteilungsfamilie.)
- Informationsgehalt der Priori: ***nicht*** oder ***schwach informative Prioris***.

Wiederholung: Bayes-Inferenz

- **Statistische Schlüsse** basieren auf der **Posteriori-Verteilung** $p(\theta|\boldsymbol{y})$, $\theta \in \boldsymbol{\Theta}$, der bedingten Verteilung der unbeobachteten Größen gegeben die beobachteten Daten.

- **Berechnung** mit dem Satz von Bayes:

$$p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)\, p(\theta)}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{y}|\theta)\, p(\theta)\, d\theta} \propto p(\boldsymbol{y}|\theta)\, p(\theta)$$

  mit der **Normierungskonstanten** der Posteriori-Dichte im Nenner.

Wiederholung: Bayes-Inferenz

- Übliche **Punktschätzer** der Bayesianischen Inferenz:
  - Posteriori-Erwartungswert $\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\boldsymbol{y})$,
  - Posteriori-Median $\hat{\boldsymbol{\theta}} = \inf\{\boldsymbol{\theta} : F(\boldsymbol{\theta}|\boldsymbol{y}) \geq 0.5\}$,
  - Posteriori-Modus $\hat{\boldsymbol{\theta}} = \arg\max\{p(\boldsymbol{\theta}|\boldsymbol{y})\}$.

- **Problem:** Posteriori-Verteilung meist analytisch unzugänglich

- **Lösung**: Verwendung von ***MCMC-Verfahren***, mit denen (abhängige) Zufallszahlen aus der Posteriori Verteilung gezogen werden können.

Grundidee MCMC

- Konstruiere Markov-Kette (MK), deren stationäre Verteilung mit der Posteriori Verteilung übereinstimmt.
- Zustände der MK entsprechen gezogenen Zufallszahlen, die (nach entsprechender Konvergenzzeit (Burn-In-Phase) der MK) abhängige Stichprobe aus Posteriori darstellen.
- Abhängigkeit kann durch geeignetes Ausdünnen der Stichprobe reduziert werden.
- Interessierende Größen (z.B. P.-Erwartungswert), werden dann aus dieser (ausgedünnten) Stichprobe durch die **empirischen Analoga** geschätzt.
- Bekanntester Algorithmus: ***Metropolis-Hastings-Algorithmus***, Spezialfall: ***Gibbs-Sampler***

Beobachtungsmodell

Beobachtungsmodell:

$$\boldsymbol{y}|\beta, \boldsymbol{b}, \vartheta \sim N(\boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b}, \boldsymbol{R}(\vartheta))$$

entspricht $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \varepsilon$ mit $\varepsilon \sim N(\boldsymbol{0}, \boldsymbol{R}(\vartheta))$.

Priori für β

- Jetzt auch „**feste**“ **Effekte** β als **Zufallsgrößen**
- kein Vorwissen über $\beta \Rightarrow$ ***nichtinformative Priori***, d.h.

$$p(\beta) \propto const,$$

sonst

$$\beta \sim N(\boldsymbol{m}, \boldsymbol{M}) \text{ mit bekanntem EWert } \boldsymbol{m}, \text{ Kovarianz } \boldsymbol{M}.$$

- nichtinformative Priori $p(\beta) \propto const$ ergibt sich als Grenzfall der Priori-Normalverteilung (NV) für ***Präzisionsmatrix*** $\boldsymbol{M}^{-1} \to \mathbf{0}$.
- zum Beobachtungsmodell **konjugierte Normalverteilung** für β $\Rightarrow$ Posteriori-Inferenz vergleichsweise einfach.

Prioris für b, ε

Üblicherweise:

$$\boldsymbol{b} \sim N(\mathbf{0}, \boldsymbol{G}(\vartheta)); \quad \varepsilon \sim N(\mathbf{0}, \boldsymbol{R}(\vartheta))$$

- Kovarianzmatrizen $\boldsymbol{G} = \text{Cov}(\boldsymbol{b})$ und $\boldsymbol{R} = \text{Cov}(\varepsilon)$ hängen i.A. von unbekannten ***Hyperparametern*** im Vektor ϑ ab.
- ***Voller Bayes-Ansatz:***
  ϑ ebenfalls **Zufallsvariable**, mit (Hyper-)Priori $p(\vartheta)$, die in Ermittlung der Posteriori mit einfließt.
- ***empirischer Bayes-Ansatz:***
  ϑ als unbekannter, aber **fester Parameter**.

Weitere Annahme: Zufallsgrößen β, $\boldsymbol{b}$ und ε a priori **unabhängig**.

Gemeinsame Posteriori bei NV-Priori für β und b

$$p(\beta, \boldsymbol{b}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\beta, \boldsymbol{b})p(\beta)p(\boldsymbol{b})$$
$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b})\right.$$
$$\left. - \frac{1}{2}(\beta - \boldsymbol{m})'\boldsymbol{M}^{-1}(\beta - \boldsymbol{m}) - \frac{1}{2}\boldsymbol{b}'\boldsymbol{G}^{-1}\boldsymbol{b}\right).$$

$$\Rightarrow \left.\begin{pmatrix} \beta \\ \boldsymbol{b} \end{pmatrix}\right| \boldsymbol{y} \sim N(\boldsymbol{\mu}_{\beta,b}, \boldsymbol{\Sigma}_{\beta,b}) \text{ (Übung) mit}$$

$$\boldsymbol{\Sigma}_{\beta,b} = \left(\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{A}\right)^{-1}; \quad \boldsymbol{C} = [\boldsymbol{X}|\boldsymbol{Z}]; \quad \boldsymbol{A} = \begin{bmatrix} \boldsymbol{M}^{-1} & 0 \\ 0 & \boldsymbol{G}^{-1} \end{bmatrix}$$

$$\boldsymbol{\mu}_{\beta,b} = \boldsymbol{\Sigma}_{\beta,b}\left(\tilde{\boldsymbol{m}} + \boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{y}\right); \quad \tilde{\boldsymbol{m}} = \begin{pmatrix} \boldsymbol{M}^{-1}\boldsymbol{m} \\ \boldsymbol{0} \end{pmatrix}$$

Posteriori bei nichtinformativer Priori für β

Nichtinformative Priori $p(\beta) \propto const$ entspricht Präzisionsmatrix $\boldsymbol{M}^{-1} = \boldsymbol{0}$:

$$
\begin{aligned}
p(\beta, \boldsymbol{b}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\beta, \boldsymbol{b})p(\boldsymbol{b}) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}'\boldsymbol{G}^{-1}\boldsymbol{b}\right) \\
&\Rightarrow \text{Posteriori-EW für } \begin{pmatrix} \beta \\ \boldsymbol{b} \end{pmatrix}\text{: } \left(\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{A}\right)^{-1}\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{y} \\
&\Rightarrow \text{Posteriori-Kov. für } \begin{pmatrix} \beta \\ \boldsymbol{b} \end{pmatrix}\text{: } \left(\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{A}\right)^{-1}
\end{aligned}
$$

- **Posteriori** äquivalent zu **penalisiertem KQ-Kriterium** $KQ_{pen}(\beta, \boldsymbol{b})$ in (16)
- **Posteriori-Modus** als Maximierer identisch mit **BLUP-Schätzern**, die $KQ_{pen}(\beta, \boldsymbol{b})$ minimieren. **Posteriori-EW** = Posteriori-Modus wegen NV.
- Posteriori-Kov. identisch mit Kovarianz in (21).

Empirische Bayes-Schätzung

- ***Empirische Bayes-Schätzer*** für $\boldsymbol{\beta}, \boldsymbol{b}$ durch Einsetzen der geschätzten Kovarianzmatrizen $\hat{\boldsymbol{G}} = \boldsymbol{G}(\hat{\boldsymbol{\vartheta}})$ und $\hat{\boldsymbol{R}} = \boldsymbol{R}(\hat{\boldsymbol{\vartheta}})$ in vorige Ausdrücke.

- Schätzung von $\hat{\boldsymbol{\vartheta}}$ z.B. durch Maximieren der **marginalen Likelihood** für $\boldsymbol{\vartheta}$:

$$\hat{\boldsymbol{\vartheta}} = \arg\max p(\boldsymbol{y}|\boldsymbol{\vartheta}) = \arg\max \int p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\vartheta}) p(\boldsymbol{b}|\boldsymbol{\vartheta}) p(\boldsymbol{\beta}) d\boldsymbol{\beta} d\boldsymbol{b}.$$

- Für nicht-informative $p(\boldsymbol{\beta}) \propto const$ ist die **marginale Likelihood**

$$p(\boldsymbol{y}|\boldsymbol{\vartheta}) = \int p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\vartheta}) p(\boldsymbol{b}|\boldsymbol{\vartheta}) d\boldsymbol{\beta} d\boldsymbol{b} = \int p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\vartheta}) d\boldsymbol{\beta}$$

proportional zur **restringierten Likelihood** $\exp\{l_R(\boldsymbol{\vartheta})\}$, siehe (20).
$\Rightarrow$ emp. Bayes-Schätzer in diesem Fall äquivalent zu **REML-Schätzer** $\hat{\boldsymbol{\vartheta}}_{REML}$ und den dazugehörigen EBLUPs für $\boldsymbol{\beta}$ und $\boldsymbol{b}$.

Volle Bayes-Inferenz

- **Voller Bayes-Ansatz**: Priori-Verteilung $p(\boldsymbol{\vartheta})$ auch für unbekannte Parameter $\boldsymbol{\vartheta}$; $\boldsymbol{\beta}$, $\boldsymbol{b}|\boldsymbol{\vartheta}$ und $\boldsymbol{\vartheta}$ als unabhängig angenommen.

- Inferenz basiert auf **Posteriori-Verteilung**

$$p(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\vartheta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\vartheta})p(\boldsymbol{\beta})p(\boldsymbol{b}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta})$$

- $p(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\vartheta}|y)$ echte Posteriori-Dichte, wenn zur **Normierung** gilt:

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\vartheta})p(\boldsymbol{\beta})p(\boldsymbol{b}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta})d\boldsymbol{\beta}d\boldsymbol{b}d\boldsymbol{\vartheta} < \infty.$$

- Bei echter, **informativer** Priori mit $\int p(\boldsymbol{\vartheta})d\boldsymbol{\vartheta} = 1$ existiert auch $p(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\vartheta}|\boldsymbol{y})$.

- Für **nichtinformative** Priori mit $\int p(\boldsymbol{\vartheta})d\boldsymbol{\vartheta} = \infty$ Existenz der Posteriori nicht allgemein gesichert.

Beziehungen zur Likelihood-Inferenz

Bei nichtinformativer Priori $p(\boldsymbol{\vartheta}) \propto const$ und Existenz der Posteriori:

- REML-Schätzer $\hat{\boldsymbol{\vartheta}}_{REML}$ als Maximierer der marginalen Likelihood = Posteriori-Modus der marginalen Posteriori von $\boldsymbol{\vartheta}$ wegen

$$p(\boldsymbol{\vartheta}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\vartheta})\frac{p(\boldsymbol{\vartheta})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y}|\boldsymbol{\vartheta}).$$

- Bei zusätzlich nichtinformativer Priori $p(\boldsymbol{\beta}) \propto const$: ML-Schätzer $\hat{\boldsymbol{\vartheta}}_{ML}$ als Maximierer der Likelihood = $\boldsymbol{\vartheta}$-Komponente des Posteriori-Modus der gemeinsamen Posteriori von $\boldsymbol{\beta}$ und $\boldsymbol{\vartheta}$ wegen

$$p(\boldsymbol{\beta}, \boldsymbol{\vartheta}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\vartheta}, \boldsymbol{\beta})\frac{p(\boldsymbol{\vartheta})p(\boldsymbol{\beta})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y}|\boldsymbol{\vartheta}, \boldsymbol{\beta}).$$

Inferenz

- Normierungskonstante der Posteriori i.A. nicht analytisch zugänglich
  $\Rightarrow$ Posteriori-Dichte $p(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\vartheta}|\boldsymbol{y})$ **nicht in geschlossener Form** darstellbar

- volle Bayes-Inferenz daher üblicherweise mittels ***MCMC-Simulation***
  (Details siehe z.B. Fahrmeir et al. (2007, Abschnitt B.5.3))

- moderne (approximative) Alternativen:
  INLA (Integrated Nested Laplace Approximation),
  Variational Bayes-Verfahren

MCMC mit blockweisem Gibbs-Sampling

Vorgehen: Teile Parametervektor $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\vartheta})$ in **Teilvektoren zusammengehöriger Parameter**, d.h. üblicherweise $\boldsymbol{\beta}$, $\boldsymbol{b}$ und $\boldsymbol{\vartheta}$ auf.

- Wähle Startwerte $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{b}^{(0)}$, $\boldsymbol{\vartheta}^{(0)}$ und Anzahl der Iterationen T
- Bilde ***vollständig bedingte Dichten*** (full conditionals) gegeben der restlichen Parameter und $\boldsymbol{y}$
- Ziehe **sequentiell Zufallszahlen** $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{b}^{(t)}$, $\boldsymbol{\vartheta}^{(t)}$ aus diesen (geg. jeweils die momentan aktuellen Zustände) bis T erreicht.

Nach einer gewissen Konvergenzphase können die Zufallszahlen als Ziehungen aus den Marginalverteilungen von $\boldsymbol{\beta}|\boldsymbol{y}$, $\boldsymbol{b}|\boldsymbol{y}$ und $\boldsymbol{\vartheta}|\boldsymbol{y}$ angesehen werden.

Vollständig bedingte Dichten

$$p(\beta|\boldsymbol{b},\vartheta,\boldsymbol{y}) \propto p(\boldsymbol{y}|\beta,\boldsymbol{b},\vartheta)p(\beta)$$
$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\beta-\boldsymbol{Z}\boldsymbol{b})'\boldsymbol{R}(\vartheta)^{-1}(\boldsymbol{y}-\boldsymbol{X}\beta-\boldsymbol{Z}\boldsymbol{b})\right.$$
$$\left.-\frac{1}{2}(\beta-\boldsymbol{m})'\boldsymbol{M}^{-1}(\beta-\boldsymbol{m})\right)$$

$$p(\boldsymbol{b}|\beta,\vartheta,\boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\beta-\boldsymbol{Z}\boldsymbol{b})'\boldsymbol{R}(\vartheta)^{-1}(\boldsymbol{y}-\boldsymbol{X}\beta-\boldsymbol{Z}\boldsymbol{b})\right.$$
$$\left.-\frac{1}{2}\boldsymbol{b}'\boldsymbol{G}(\vartheta)^{-1}\boldsymbol{b}\right)$$

$$p(\vartheta|\beta,\boldsymbol{b},\boldsymbol{y}) \propto p(\boldsymbol{y}|\beta,\boldsymbol{b},\vartheta)p(\boldsymbol{b}|\vartheta)p(\vartheta)$$
$$\propto |\boldsymbol{R}(\vartheta)|^{-1/2}\,|\boldsymbol{G}(\vartheta)|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{b}'\boldsymbol{G}(\vartheta)^{-1}\boldsymbol{b}\right.$$
$$\left.-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\beta-\boldsymbol{Z}\boldsymbol{b})'\boldsymbol{R}(\vartheta)^{-1}(\boldsymbol{y}-\boldsymbol{X}\beta-\boldsymbol{Z}\boldsymbol{b})\right)p(\vartheta)$$

Vollständig bedingte Dichten von β, b

$$\beta|\cdot \sim N(\mu_\beta, \Sigma_\beta) \text{ mit } (\textbf{Übung})$$
$$\Sigma_\beta = \left(X'R(\vartheta)^{-1}X + M^{-1}\right)^{-1}$$
$$\mu_\beta = \Sigma_\beta \left(M^{-1}m + X'R(\vartheta)^{-1}(y - Zb)\right)$$

$$b|\cdot \sim N(\mu_b, \Sigma_b) \text{ mit (analog)}$$
$$\Sigma_b = \left(Z'R(\vartheta)^{-1}Z + G(\vartheta)^{-1}\right)^{-1}$$
$$\mu_b = \Sigma_b \left(Z'R(\vartheta)^{-1}(y - X\beta)\right)$$

- **nichtinformative Priori** mit $M^{-1} = \mathbf{0}$: Erwartungswert $\mu_\beta = (X'R(\vartheta)^{-1}X)^{-1}X'R(\vartheta)^{-1}(y - Zb)$ ist gewichteter KQ-Schätzer angewandt auf die um Zb bereinigten Daten
- **informative Priori** $\Rightarrow$ Erwartungswert μ_β ist gewichtetes Mittel aus KQ-Schätzer und Priori-Erwartungswert.
- Analog für μ_b.

Vollständig bedingte Dichten von ϑ im Spezialfall

LMM für Clusterdaten mit $\text{Cov}(\varepsilon) = \sigma^2 I_n \Rightarrow \boldsymbol{\vartheta}$ enthält σ^2 und Parameter in $\boldsymbol{D}$.

- nichtinformative Jeffreys Prioris $p(\sigma^2) \propto \sigma^{-2}$, $p(\boldsymbol{D}) \propto |\boldsymbol{D}|^{-\frac{q+1}{2}}$ führen i.A. zu uneigentlichen (d.h. nicht normierbaren) Posteriori-Verteilungen

$\Rightarrow$ Schwach informative inverse Gammaverteilung $\sigma^2 \sim IG(a_\sigma, b_\sigma)$; a_σ, b_σ klein

- Für $\boldsymbol{D}$ oft inverse Wishart-Verteilung. Bei $\boldsymbol{D} = \text{diag}(\tau_1^2, \ldots, \tau_q^2)$ mit unabh. τ_j^2 ergibt diese ein Produkt von IGs mit $\tau_j^2 \sim IG(a_{\tau_j}, b_{\tau_j}), j = 1, \ldots, q$.

Dann full conditionals bei zusätzlich $p(\boldsymbol{\beta}) \propto const$ (nichtinformative Priori):

$$\sigma^2|\cdot \sim IG(\tilde{a}_\sigma, \tilde{b}_\sigma) \text{ mit } \tilde{a}_\sigma = a_\sigma + \frac{1}{2},\ \tilde{b}_\sigma = b_\sigma + \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b}||_2^2$$

$$\tau_j^2|\cdot \sim IG(\tilde{a}_{\tau_j}, \tilde{b}_{\tau_j}) \text{ mit } \tilde{a}_{\tau_j} = a_{\tau_j} + \frac{N}{2},\ \tilde{b}_{\tau_j} = b_{\tau_j} + \frac{1}{2}\sum_{i=1}^{N} b_{ij}^2$$

Probleme der NV-Annahme für b

NV-Annahme für zufällige Effekte $\boldsymbol{b} \sim N(\mathbf{0}, \boldsymbol{G})$ mathematisch günstig. Aber wie sensitiv ist die Schätzung bei Fehlspezifikation?

- Die Schätzer der festen Effekte β und der Kovarianzparameter ϑ sind meist sehr robust gegenüber Fehlspezifikation der Verteilung der $\boldsymbol{b}$.
  Die Standardfehler können jedoch über/unterschätzt werden.

- Durch den Shrinkage-Effekt können die EBLUPs der zufälligen Effekte $\boldsymbol{b}$ normalverteilt aussehen, selbst wenn die Verteilung der $\boldsymbol{b}$ z.B. bimodal / schief ist / hohe Wahrscheinlichkeitsmasse an den Rändern hat (heavy tails)
  $\Rightarrow$ Schlechte Vorhersagen $\hat{\boldsymbol{b}}$. q-q-Plots der $\hat{\boldsymbol{b}}$ eignen sich nicht zur Diagnose.
  Diagnose durch Fitten eines flexibleren Modells.

Alternative Prioris für b: Skalenmischungen

- Verwendung von Prioris mit mehr Masse auf den Rändern (heavy-tailed): z.B. t-Verteilung mit niedrigen Freiheitsgraden, Laplace-Verteilung. Diese sind oft darstellbar als Skalenmischung von Normalverteilungen

$$p(b_i) = \int \phi(b_i|\mu, \sigma^2) p(\sigma^2|\theta) d\sigma^2$$

- Darstellbar als Skalenmischung $\Rightarrow$ sehr leicht in Modellhierarchie für LMM einzubauen

- Beispiel: t-Verteilung mit $df = \nu$ ist Skalenmischung aus $N(0, \sigma^2)$ mit $\sigma^{-2} \sim \Gamma(\nu/2, \nu/2)$

Alternative Prioris für b: Finite Mischungen

- Zur Aufdeckung von Clustern (z.B. durch unbeobachtete Kovariablen oder latente Subpopulationen) können multimodale Verteilungen für die zufälligen Effekte verwendet werden, z.B. ***finite Mischverteilungsmodelle***:

$$p(\boldsymbol{b}_i|\boldsymbol{\pi}, \boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{b}_i|\boldsymbol{\phi}_k),$$

  mit Gewichten $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\sum_k^K \pi_k = 1$, und parametrischer Verteilungsfamilie $p(\boldsymbol{b}_i|\boldsymbol{\phi}_k)$ mit Parametern $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K)$ (z.B. multivariate NV $p(\boldsymbol{b}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ mit Erwartungswerten $\boldsymbol{\mu}_k$ und homogenen Kovarianzen $\boldsymbol{\Sigma}$)

- nichtparametrische Erweiterung: K nicht fest/konstant, wird mitgeschätzt.

- Noch flexibler: nichtparametrische Bayes-Ansätze, z.B. **Dirichlet-Prozess-**, Dirichlet-Prozess-Mischungs-Prioris.

- Inferenz für diese Modelle basiert i.d.R. auf MCMC-Techniken.

Inhalt

Motivation

Die gemeinsame log-Likelihood (basierend auf der gemeinsamen Dichte von $\boldsymbol{y}$ und $\boldsymbol{b}$) im LMM, aus der sich die BLUPs für β und $\boldsymbol{b}$ ergeben, hat die Form (s. (15))

$$l(\beta, \boldsymbol{b}) = \log L(\beta, \boldsymbol{b}) = -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b})' \boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}' \boldsymbol{G}^{-1}\boldsymbol{b}.$$

Diese kann als penalisierte Likelihood aufgefasst werden mit Strafterm $\boldsymbol{b}' \boldsymbol{G}^{-1}\boldsymbol{b}$.

Dies öffnet eine Verbindung zu penalisierten Regressionsmethoden (mit quadratischen Straftermen) und damit zu einer sehr allgemeinen Modellklasse.

Wir beginnen mit dem ausführlichen Beispiel der penalisierten Splines; weitere knappe Beispiele zeigen dann die Größe der flexiblen Modellklasse auf.

Das nichtparametrische Regressionsproblem

Modell: Die metrische Zielvariable y läßt sich durch eine deterministische Funktion $f(z)$ der metrischen Kovariablen z und einen additiven Fehler ε erklären:

$$y_i = f(z_i) + \varepsilon_i, i = 1, \ldots, n.$$

- **Annahmen über f:**
  Glattheitsanforderungen, z.B. via Stetigkeit oder Differenzierbarkeit von f.
  Deutlich flexibler als Linearitätsannahme $f(z) = \beta_0 + \beta_1 z$ (lineares Modell).

- **Annahmen über ε:**
  Wie im linearen Modell $\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right) \Rightarrow E(y_i) = f(z_i)$.
  D.h. die Funktion f modelliert den Erwartungswert der Zielvariablen.

Splines

Idee: Modelliere f stückweise durch Polynome, ohne dass Sprungstellen / Kanten entstehen $\rightarrow$ Splines auf dem Definitionsbereich $[a, b]$ von z.

Definition von Splines:

Wähle Knoten $a = \kappa_1 < \ldots < \kappa_m = b$. f ist ein Spline vom Grad $\ell \geq 0$, falls gilt

- $f(z)$ ist auf den Intervallen $[\kappa_j, \kappa_{j+1})$ ein Polynom vom Grad ℓ.
- $f(z)$ ist stetig (falls $\ell > 0$) und $(\ell - 1)$-mal stetig differenzierbar. (Glattheit)

Konstruktive Darstellung durch Basisfunktionen:

Jeder Spline f vom Grad ℓ mit Knoten $\kappa_1 < \ldots < \kappa_m$ kann eindeutig als Linearkombination $f(z) = \sum_{j=1}^{d} \gamma_j B_j(z)$ von $d = \ell + m - 1$ Basisfunktionen B_j dargestellt werden, die den d dimensionalen Splineraum $S_\ell(\kappa_1, ..., \kappa_m)$ aufspannen.

Bekannte Basen: Truncated-Power-Series (TP) und Basic-Splines (B-Splines).

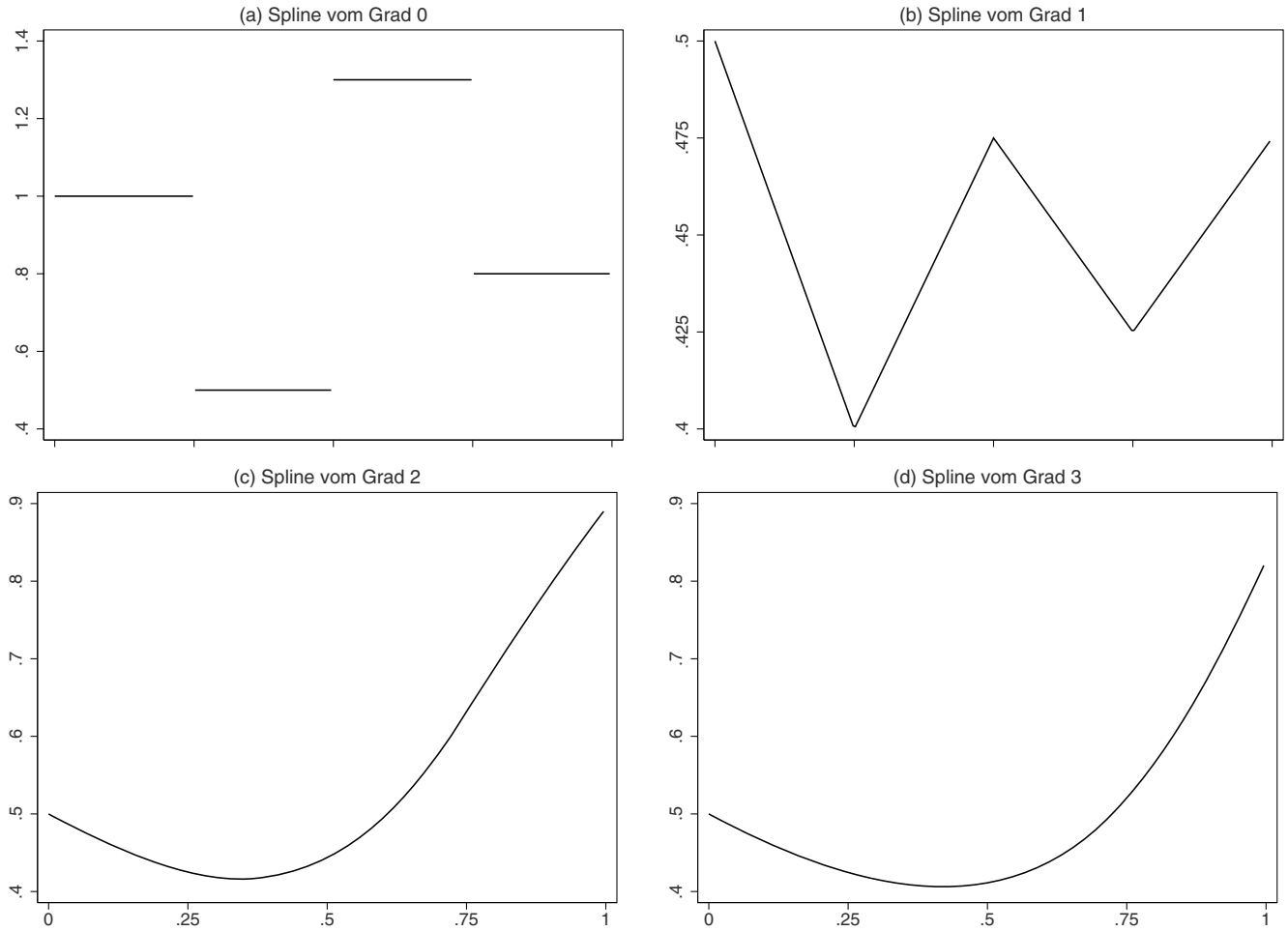


Abbildung 1: Beispiele für Splines vom Grad 0, 1, 2 und 3 zu den Knoten $\kappa_1 = 0, \kappa_2 = 0.25, \kappa_3 = 0.5, \kappa_4 = 0.75$ und $\kappa_5 = 1$.
Quelle: Fahrmeir et al (2009, Abb. 7.5).

Definition der d Basisfunktionen

Für eine Knotenmenge $\{\kappa_1, ..., \kappa_m\}$ und einen Grad ℓ:

TP-Basisfunktionen

$$B_j(z) = z^{j-1}, \qquad j = 1, \ldots, \ell + 1 \qquad \text{(Polynome)}$$
$$B_{\ell+j}(z) = (z - \kappa_j)_+^{\ell}, \qquad j = 2, \ldots, m - 1 \qquad \text{(abgeschnittene Potenzen).}$$

mit $(u)_+^{\ell} = [\max(u, 0)]^{\ell}$.

B-Spline Basisfunktionen (rekursiv)

$$B_j^0(z) = \mathbb{1}_{[\kappa_j, \kappa_{j+\mathbf{1}})}(z), \qquad j = 1, \ldots, d,$$
$$B_j^{\ell}(z) = \frac{z - \kappa_{j-\ell}}{\kappa_j - \kappa_{j-\ell}} B_{j-1}^{\ell-1}(z) + \frac{\kappa_{j+1} - z}{\kappa_{j+1} - \kappa_{j+1-\ell}} B_j^{\ell-1}(z), \qquad j = 1, \ldots, d$$

mit zusätzlichen Knoten $\kappa_{1-\ell}, ..., \kappa_0 < a$ und $b < \kappa_{m+1}, ..., \kappa_{m+\ell}$.

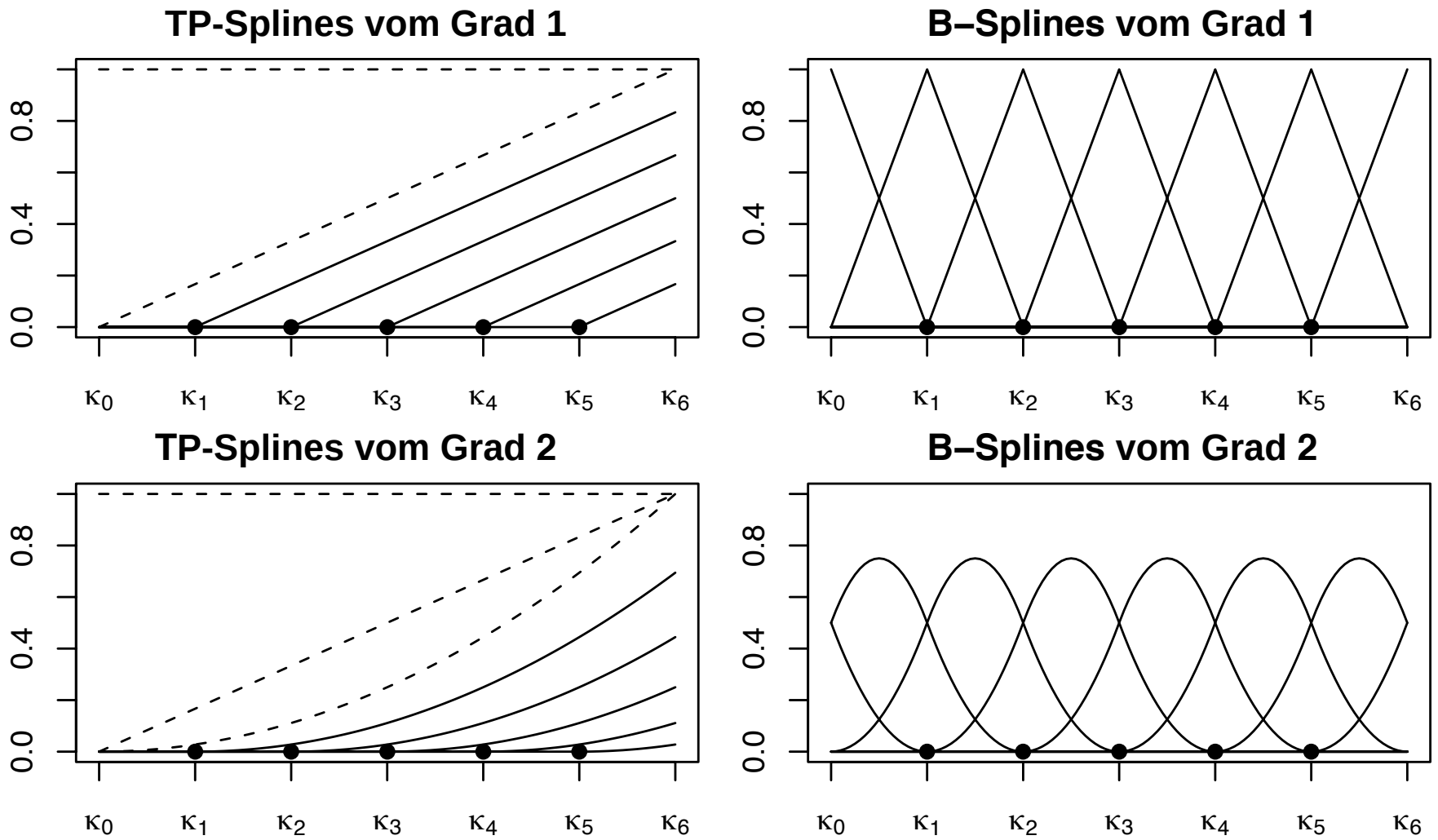


Abbildung 2: Lineare (oben) und quadratische (unten) TP- (links) und B-Spline-Basifunktionen (rechts) für $K = 5$ äquidistante Knoten mit $\kappa_0 = 0$. Für die TP-Basis sind die Polynomterme gestrichelt dargestellt.

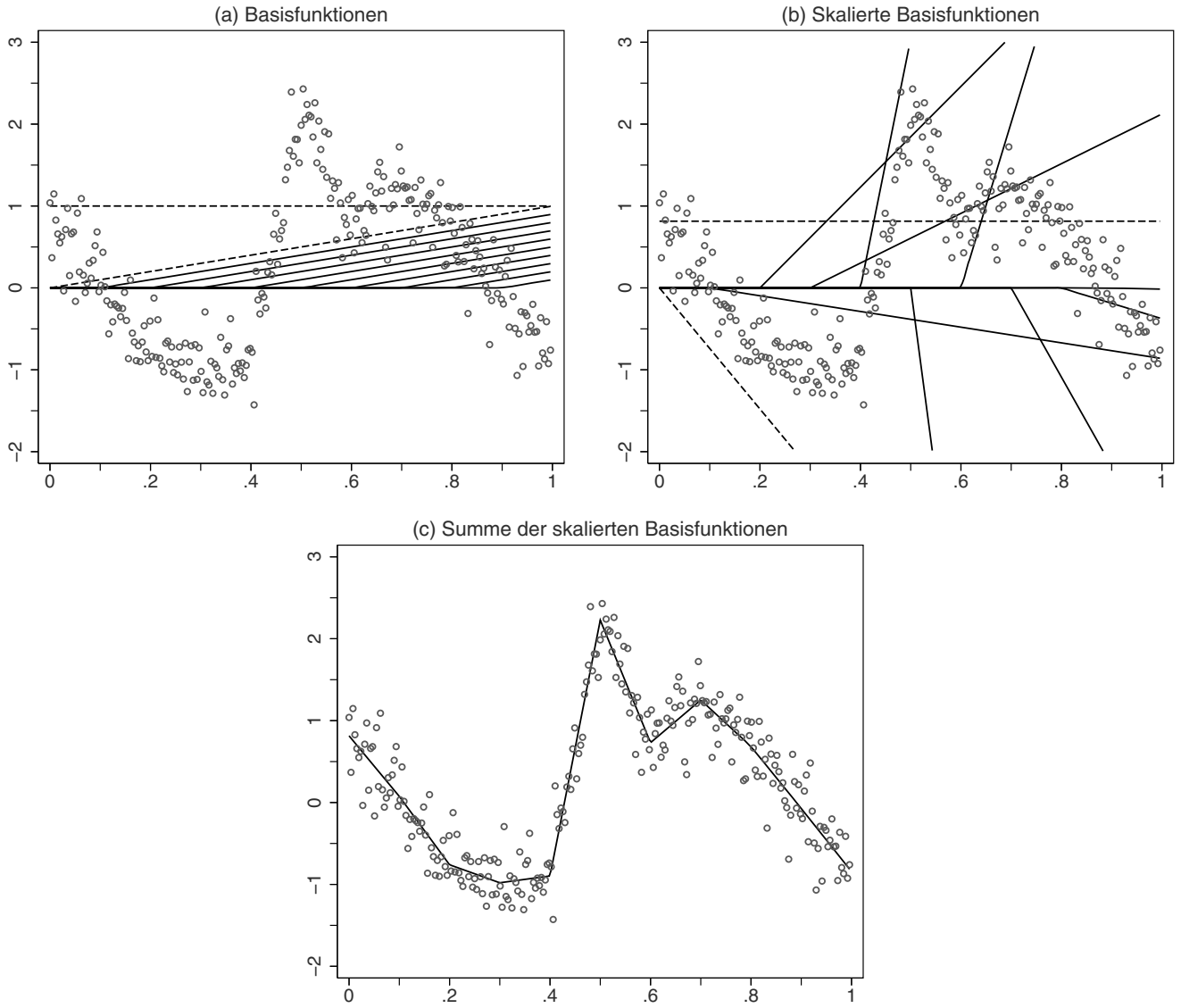


Abbildung 3: Konstruktion einer Funktion aus skalierten TP-Basisfunktionen für lineare Splines. Quelle: Fahrmeir et al (2009, Abb. 7.6).

Basisdarstellung

Damit ergibt sich die Darstellung des nichtparametrischen Regressionsproblems als

$$y_i = f(z_i) + \varepsilon_i \approx \sum_{j=1}^{d} \gamma_j B_j(z_i) + \varepsilon_i.$$

Man erhält ein **lineares Modell** in dem Parametervektor $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)'$

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

mit Designmatrix aus TP- bzw. B-Spline-Basis

$$\boldsymbol{B} = \begin{pmatrix} B_1(z_1) & \ldots & B_d(z_1) \\ \vdots & & \vdots \\ B_1(z_n) & \ldots & B_d(z_n) \end{pmatrix}.$$

Mit KQ-Schätzer ergibt sich $\widehat{f}(z) = (B_1(z), \ldots, B_d(z))(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{y}$.

Einfluss der Knotenzahl

Es gilt der Varianz-Bias-Trade Off:

- große Knotenzahl: flexible Funktionsschätzung, aber große Variabilität (Overfitting)
- kleine Knotenzahl: glattere Schätzung, aber möglicher Bias

$\Rightarrow$ Wesentliche Frage nach der Anzahl der zu verwendenden Knoten!

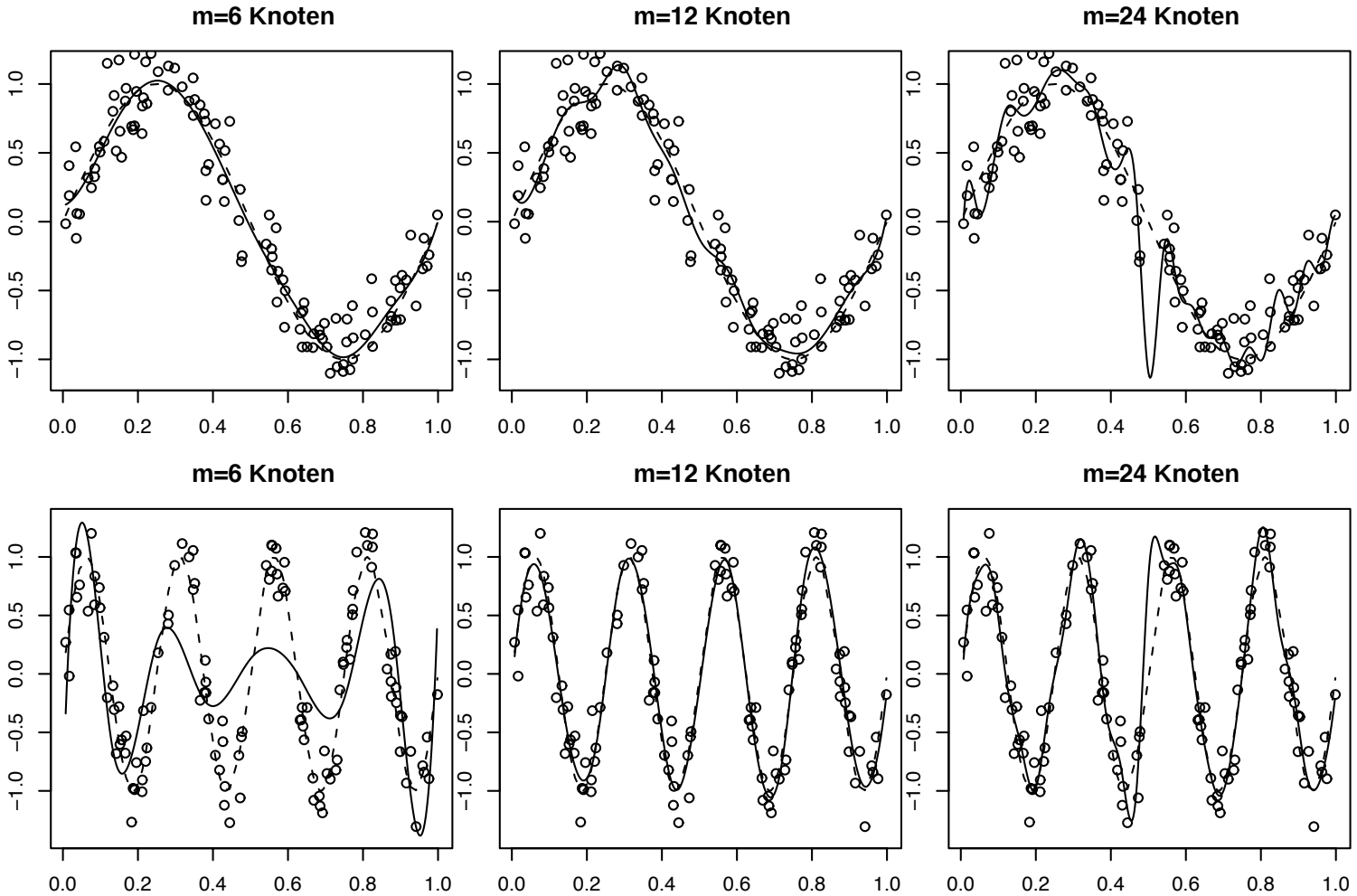


Abbildung 4: Abhängigkeit der Glättung mit Splines von der Knotenwahl. Die gestrichelte Linie gibt jeweils die wahre Funktion wieder, die durchgezogene Linie die Schätzung mit B-Splines mit 6, 12 bzw. 24 Knoten.

Grundidee penalisierte Splines

- Approximiere $f(z)$ durch einen Spline mit einer **großen Zahl von Knoten** (üblicherweise ca. 20 bis 40) $\Rightarrow$ prinzipiell flexibel genug, auch stark variierende Funktionen $f(z)$ darzustellen.
- Führe zusätzlich einen Strafterm ein, der eine zu große **Variabilität der Schätzung bestraft**. (Regularisierung)
- Minimiere anstelle des üblichen KQ-Kriteriums ein um diesen Strafterm erweitertes penalisiertes KQ-Kriterium (äquivalent: Maximierung einer penalisierten log-Likelihood).

Penalisierte Splines auf TP-Basis

$$f(z) = \underbrace{\gamma_1 + \gamma_2 z + \ldots + \gamma_{\ell+1} z^\ell}_{\text{1.Teil: globale polynomiale Form}} + \underbrace{\gamma_{\ell+2}(z-\kappa_2)_+^\ell + \ldots + \gamma_d (z-\kappa_{m-1})_+^\ell}_{\text{2.Teil: Abweichung von globaler polynomialer Form}} .$$

Rauhe Funktionsschätzungen bei großer Variabilität des 2. Modellteils.

Da die ℓte Ableitung von $(z-\kappa_j)_+^\ell$ gleich $\ell!$ für $z > \kappa_j$ und 0 sonst ist, $j \geq \ell + 2$, sind die γ_j proportional zu Sprüngen in der ℓten Ableitung.

$\Rightarrow$ Strafterm für betragsmäßig große Koeffizienten der abgeschnittenen Potenzen:

$$pen(\boldsymbol{\gamma}, \boldsymbol{K}) = \lambda \sum_{j=l+2}^{d} \gamma_j^2 = \lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}$$

mit $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)'$ und $\boldsymbol{K} = diag(\underbrace{0, \ldots, 0}_{(\ell+1)}, \underbrace{1, \ldots, 1}_{(m-2)})$.

Penalisierte Splines auf B-Spline-Basis: P-Splines

- Ein geeigneter Strafterm lässt sich über das Integral der (quadrierten) k-ten Ableitung definieren. Häufig wird dabei $k = 2$ verwendet:

$$\int (f''(z))^2 dz = \sum_{i=1}^{d} \sum_{j=1}^{d} \gamma_i \gamma_j \int B_i''(z) B_j''(z) dz = \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}$$

  mit Strafmatrix $\boldsymbol{K} = (\int B_i''(z) B_j''(z) dz)_{i,j}$.

- Für die k-ten Ableitungen lässt sich der Strafterm über Differenzen k-ter Ordnung Δ^kder Parameter γ approximieren. Diese sind rekursiv definiert

$$\begin{aligned}
\Delta^1 \gamma_j &= \gamma_j - \gamma_{j-1}, \\
\Delta^2 \gamma_j &= \Delta^1 \Delta^1 \gamma_j = \Delta^1 \gamma_j - \Delta^1 \gamma_{j-1} = \gamma_j - 2\gamma_{j-1} + \gamma_{j-2}, \\
&\vdots \\
\Delta^k \gamma_j &= \Delta^{k-1} \gamma_j - \Delta^{k-1} \gamma_{j-1}.
\end{aligned}$$

P-Splines

Der Strafterm für B-Splines mit $k$ten Differenzen $\Delta^k \gamma_j$ hat die Gestalt

$$pen(\gamma, \boldsymbol{K}_k) = \lambda \gamma' \boldsymbol{K}_k \gamma = \lambda \gamma' \boldsymbol{D}_k' \boldsymbol{D}_k \gamma$$

mit der $(d-k) \times d$ ***Differenzenmatrix*** $\boldsymbol{D}_k$ und der $d \times d$ Strafmatrix $\boldsymbol{K}_k$

$$\boldsymbol{D}_1 = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}, \quad \boldsymbol{K}_1 = \begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 1 \end{pmatrix},$$

$$\boldsymbol{D}_k = \boldsymbol{D}_1 \boldsymbol{D}_{k-1}, \qquad \boldsymbol{K}_k = \boldsymbol{D}_k' \boldsymbol{D}_k$$

(bei entsprechend angepassten Dimensionen).

Penalisiertes KQ-Kriterium

Minimiere das penalisierte KQ-Kriterium über γ in Abhängigkeit vom $\lambda \geq 0$,

$$PKQ(\lambda, \gamma) = (\boldsymbol{y} - \boldsymbol{B}\gamma)'(\boldsymbol{y} - \boldsymbol{B}\gamma) + \lambda\gamma'\boldsymbol{K}\gamma.$$

- Der Glättungsparameter $\lambda \geq 0$ steuert dabei den **Kompromiss zwischen Datentreue** und **Glattheit** der Funktionsschätzung.
- $\lambda \to 0$: Geringes Gewicht für den Strafterm. $PKQ(\lambda, \gamma) \approx KQ(\gamma)$ und $\hat{\gamma}$ nahe beim KQ-Schätzer.
- $\lambda \to \infty$: Schätzung vollständig durch den Strafterm dominiert. Resultat:
  - Bei der TP-Basis ein Polynom vom Grad ℓ als Schätzung für $f(z)$.
  - Bei P-Splines des Grades $\ell \geq k - 1$, $k$te Differenzen, ein Polynom vom Grad $k - 1$.
- Durch Wahl von $\lambda \geq 0$ Kontinuum zwischen diesen beiden Extremen.

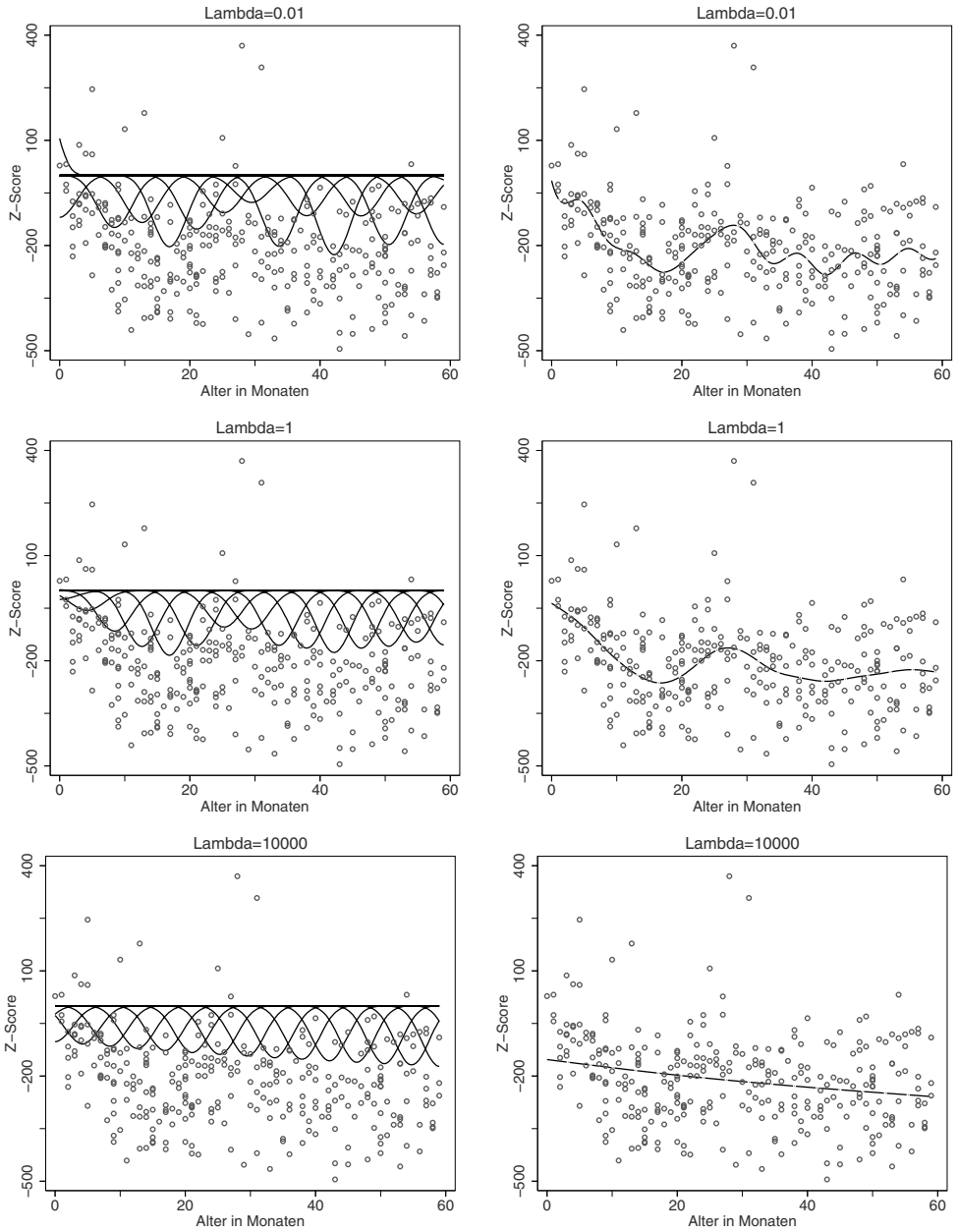


Abbildung 5: Einfluss des Glättungsparameters auf die Schätzung eines P-Splines basierend auf zweiten Differenzen. Quelle: Fahrmeir (2009, Abb. 7.15).

Die Rolle des Glättungsparameters

- Vorteil: Die Glattheit der Schätzung wird nicht mehr über Anzahl und Position der Knoten, sondern nur durch den Glättungsparameter λ gesteuert.
  $\Rightarrow$ Frage nach der Wahl **geeigneter Werte für den Glättungsparameter** λ!

Optimalitätskriterien zur Bestimmung von Glättungsparametern, z.B.:

- Überlegungen zum **Bias-Varianz-Trade Off**. Einen solchen Kompromiss erhält man durch die Betrachtung des mittleren quadratischen Fehlers.
- Approximation des quadratischen Prognosefehlers durch **Kreuzvalidierung**.
- Verwendung von anderen aus der **Modellwahl** bekannten Kriterien, z.B. AIC.
- Hier: **Darstellung der Penalisierungsansätze als gemischtes Modell**.

TP-Penalisierungsansatz als gemischtes Modell

Für penalisierte Splines mit TP-Basis ist das penalisierte KQ-Kriterium

$$PKQ(\lambda, \gamma) = (\boldsymbol{y} - \boldsymbol{B}\gamma)'(\boldsymbol{y} - \boldsymbol{B}\gamma) + \lambda \sum_{j=l+2}^{d} \gamma_j^2.$$

Zerlege die Designmatrix $\boldsymbol{B} = (\boldsymbol{X}|\boldsymbol{Z})$ und die Koeffizienten γ in

- $\beta = (\gamma_1, \ldots, \gamma_{\ell+1})'$, die nicht penalisierten Polynomkoeffizienten,
- $\boldsymbol{b} = (\gamma_{\ell+2}, \ldots, \gamma_d)'$, die penalisierten Koeff. der abgeschnittenen Potenzen.

$$\Rightarrow PKQ(\lambda, \gamma)/\sigma^2 = (\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b})'(\frac{1}{\sigma^2}\boldsymbol{I_n})(\boldsymbol{y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b}) + \boldsymbol{b}'(\frac{\lambda}{\sigma^2}\boldsymbol{I_{m-2}})\boldsymbol{b}.$$

Dies ist das penalisierte KQ-Kriterium (16) für ein LMM

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \varepsilon, \quad \varepsilon \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I_n}), \quad \boldsymbol{b} \sim N(\boldsymbol{0}, \frac{\sigma^2}{\lambda}\boldsymbol{I_{m-2}}). \tag{25}$$

- β und $\boldsymbol{b}$ können als Vektoren fester / zufälliger Effekte aufgefasst werden.
- Die Varianz τ^2 der zufälligen Effekte ist dabei $\tau^2 = \sigma^2/\lambda$.

Schätzung über gemischte Modelle

- Im gemischten Modell (25) können die Varianzen σ^2 und τ^2 über ML- oder REML-Schätzung bestimmt werden.
- Den optimalen Glättungsparameter erhält man dann als $\hat{\lambda} = \hat{\sigma}^2/\hat{\tau}^2$.
- Die Hypothese

$$H_0 : f \text{ Polynom vom Grad } \ell \text{ bzw. } k-1 \quad \text{gegen} \quad H_A : f \text{ flexiblere Funktion}$$

lässt sich durch einen Test überprüfen für

$$H_0 : \tau^2 = 0 \quad \text{gegen} \quad H_A : \tau^2 > 0.$$

(Beachte die Hinweise zum Testen von Varianzkomponenten, Folien 66 ff.! Hier lässt sich $\boldsymbol{y}$ nicht in N unabhängige Teilvektoren zerlegen, sodass die asymptotischen Mischungsverteilungen für T_{LQT} nicht gültig sind. Die exakte Verteilung oder Bootstrap-Verfahren bieten Alternativen.)

Punktweise Konfidenzbänder

Der BLUP für $f(z) = \boldsymbol{C}_z(\beta', \boldsymbol{b}')'$ mit $\boldsymbol{C}_z = (B_1(z), \ldots, B_d(z))$ ergibt sich aus (18) mit $\boldsymbol{C} = \boldsymbol{B} = [\boldsymbol{X}|\boldsymbol{Z}]$ als

$$\hat{f}(z) = \boldsymbol{C}_z(\hat{\beta}', \hat{\boldsymbol{b}}')' = \boldsymbol{C}_z(\boldsymbol{C}'\boldsymbol{C} + \lambda \operatorname{diag}(\boldsymbol{0}, \boldsymbol{I_{m-2}}))^{-1}\boldsymbol{C}'\boldsymbol{y},$$

mit der (bayesianischen) Kovarianz aus (21)

$$\operatorname{Cov}(\hat{f}(z) - f(z)) = \sigma^2 \boldsymbol{C}_z(\boldsymbol{C}'\boldsymbol{C} + \lambda \operatorname{diag}(\boldsymbol{0}, \boldsymbol{I_{m-2}}))^{-1}\boldsymbol{C}_z'.$$

Diese Kovarianz berücksichtigt möglichen Bias in $\hat{f}(z)$. → **Übung**

Unter der Annahme asymptotischer Normalverteilung können hiermit approximative punktweise Konfidenzbänder konstruiert werden. → **Übung**

Für simultane Konfidenzbänder siehe Ruppert, Wand und Carroll (2003, 6.5).

Penalisierte Regression und LMMs

Für allgemeine Penalisierungsansätze mit penalisierter Residuenquadratsumme

$$PKQ(\lambda) = (\boldsymbol{y} - \boldsymbol{B}\gamma)'(\boldsymbol{y} - \boldsymbol{B}\gamma) + \lambda\gamma'\boldsymbol{K}\gamma \tag{26}$$

könnte man nun versuchen, mit $\tau^2 = \sigma^2/\lambda$ ein LMM zu definieren der Form

$$\boldsymbol{y} = \boldsymbol{B}\gamma + \varepsilon, \quad \varepsilon \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I_n}), \quad \gamma \sim N(\boldsymbol{0}, \tau^2\boldsymbol{K}^{-1}).$$

Problem: Die Inverse $\boldsymbol{K}^{-1}$ existiert nicht immer, $0 < \text{Rang}(\boldsymbol{K}) < \dim(\gamma)$.

- **Beispiel:** Für penalisierte B-Splines hat $\boldsymbol{K}_k = \boldsymbol{D}_k'\boldsymbol{D}_k$ keinen vollen Rang.

Konsequenz: Die aus (26) resultierende Dichte der zufälligen Effekte

$$p(\gamma) \propto \exp\left(-\frac{1}{2\tau^2}\gamma'\boldsymbol{K}\gamma\right)$$

ist damit **teilweise uneigentlich**, d.h. sie lässt sich nicht normieren.

Konstruktion einer geeigneten Zerlegung

Sei $h = \text{Rang}(\boldsymbol{K})$ und $d = \dim(\gamma)$. Zerlege γ in Teilvektoren β und $\boldsymbol{b}$, so dass

$$\gamma = \underset{d\times(d-h)}{\tilde{\boldsymbol{X}}} \underset{(d-h)\times 1}{\beta} + \underset{d\times h}{\tilde{\boldsymbol{Z}}} \underset{h\times 1}{\boldsymbol{b}}.$$

Wähle $\tilde{\boldsymbol{X}}$ und $\tilde{\boldsymbol{Z}}$, so dass

- $\boldsymbol{K}\tilde{\boldsymbol{X}} = \boldsymbol{0}$, so dass β nicht durch $\boldsymbol{K}$ penalisiert wird,
- $\tilde{\boldsymbol{Z}}'\boldsymbol{K}\tilde{\boldsymbol{Z}} = \boldsymbol{I_h}$, so dass sich der Strafterm als $\lambda \boldsymbol{b}'\boldsymbol{b}$ schreiben lässt:

$$\gamma'\boldsymbol{K}\gamma = (\tilde{\boldsymbol{X}}\beta + \tilde{\boldsymbol{Z}}\boldsymbol{b})' \underbrace{\boldsymbol{K}(\tilde{\boldsymbol{X}}}_{\boldsymbol{0}}\beta + \tilde{\boldsymbol{Z}}\boldsymbol{b}) = \boldsymbol{b}' \underbrace{\tilde{\boldsymbol{Z}}'\boldsymbol{K}\tilde{\boldsymbol{Z}}}_{\boldsymbol{I_h}}\boldsymbol{b} = \boldsymbol{b}'\boldsymbol{b}.$$

β, $\boldsymbol{b}$ können dann als feste und (i.i.d.) zufällige Effekte aufgefasst werden im LMM

$$\boldsymbol{y} = \boldsymbol{B}\gamma + \varepsilon = \boldsymbol{B}\tilde{\boldsymbol{X}}\beta + \boldsymbol{B}\tilde{\boldsymbol{Z}}\boldsymbol{b} + \varepsilon =: \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \varepsilon, \quad \boldsymbol{b} \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{I_h}).$$

$\lambda = \sigma^2/\tau^2$ lässt sich wieder über REML oder ML schätzen.

Details zur Konstruktion

- Verwende eine Basis des Nullraums von $\boldsymbol{K}$ als Spalten von $\tilde{\boldsymbol{X}} \Rightarrow \boldsymbol{K}\tilde{\boldsymbol{X}} = \boldsymbol{0}$. Z.B. Polynome vom Grad 0 bis $k-1$ für B-Splines mit $k$ten Differenzen.
- Für $\tilde{\boldsymbol{Z}}$, verwende Spektralzerlegung $\boldsymbol{K} = \underset{d\times h}{\boldsymbol{\Gamma}}\ \underset{h\times h}{\boldsymbol{\Omega}_+}\ \underset{h\times d}{\boldsymbol{\Gamma}'}$ mit $\boldsymbol{\Omega}_+$ Diagonalmatrix der positiven Eigenwerte, $\boldsymbol{\Gamma}$ orthonormale Matrix der Eigenvektoren.

  Definiere $\tilde{\boldsymbol{Z}} = \boldsymbol{L}(\boldsymbol{L}'\boldsymbol{L})^{-1}$ mit $\boldsymbol{L} = \boldsymbol{\Gamma}{\boldsymbol{\Omega}_+}^{1/2}$ (also $\boldsymbol{K} = \boldsymbol{L}\boldsymbol{L}'$).

  $\Rightarrow \tilde{\boldsymbol{Z}}'\boldsymbol{K}\tilde{\boldsymbol{Z}} = \left(\boldsymbol{L}'\boldsymbol{L}\right)^{-1}\boldsymbol{L}'\boldsymbol{L}\boldsymbol{L}'\boldsymbol{L}\left(\boldsymbol{L}'\boldsymbol{L}\right)^{-1} = \boldsymbol{I_h}$.
- Die Spektralzerlegung ist nicht immer notwendig. So kann beispielsweise für P-Splines auch $\boldsymbol{L} = \boldsymbol{D}'$ mit der Differenzenmatrix $\boldsymbol{D}$ gewählt werden.

$\Rightarrow$ Die Zerlegung von γ ist nicht eindeutig.

Frequentistische und Bayesianische Sicht

- Frequentistisch betrachtet sind die Parameter γ feste, unbekannte Parameter.
- Im umformulierten gemischten Modell enthält der $\boldsymbol{b}$-Teil von γ zufällige Effekte - keine (festen) Parameter, sondern Zufallsgrößen!
- Streng genommen wäre die Darstellung als LMM nur als algorithmischer Trick zu betrachten, nicht als tatsächliche Umformulierung des Modells.
- Bayesianisch betrachtet ist dies kein Problem, da eh alle Parameter Zufallsgrößen sind. Die zwei Darstellungen sind äquivalente Formulierungen der gleichen Priori-Annahmen.

Bayesianische Sicht auf Penalisierungsansätze

Beobachtungsmodell

$$\boldsymbol{y} = \boldsymbol{B}\gamma + \varepsilon, \quad \varepsilon \sim N\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$$

mit Designmatrix $\boldsymbol{B}$, die entsprechende Basisfunktionen enthält.

Penalisierung der Form $pen(\lambda, \boldsymbol{K}) = \lambda\gamma'\boldsymbol{K}\gamma$ entspricht (improperer) multivariater **Priori-Normalverteilung** $p(\gamma|\boldsymbol{K}, \tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\gamma'\boldsymbol{K}\gamma\right)$ für $\tau^2 = \sigma^2/\lambda$ oder

$$\gamma \sim N(\boldsymbol{0}, \tau^2\boldsymbol{K}^{-1}).$$

Beispiel: TP-Basis

- **Priori-Normalverteilung** für die Koeffizienten der abgeschnittenen Polynome: $\gamma_{l+2}, \ldots, \gamma_d \overset{iid}{\sim} N(0, \tau^2)$.
- **Nichtinformative Priori** für die Koeffizienten der globalen Polynome: $p(\gamma_j) \propto const$, $j = 1, \ldots, l+1$.

Die beiden verschiedenen Typen von Priori-Verteilungen spiegeln die Unterscheidung in unpenalisiert und penalisiert zu schätzende Parameter wieder.

Beispiel: B-Spline-Basis

- Priori-Annahme für γ: Random Walks (Irrfahrten) der Ordnung k als stochastisches Analogon zur Differenzenbestrafung. Z.B. $k = 1$ (RW1):

$$\gamma_j = \gamma_{j-1} + u_j, \quad u_j \sim N(0, \tau^2), \quad j = 2, \ldots, d.$$

- nichtinformative Priori-Verteilung $p(\gamma_1) \propto const$
- $\Rightarrow$ bedingte Verteilungen $\gamma_j|\gamma_{j-1}, \ldots, \gamma_1 \sim N(\gamma_{j-1}, \tau^2)$
- gemeinsame Verteilung für γ: NV mit EW $\mathbf{0}$ und Präzisionsmatrix $\boldsymbol{K}/\tau^2$, $\boldsymbol{K} = \boldsymbol{D}_1'\boldsymbol{D}_1$
- Präzisionsmatrix hat keinen vollen Rang $\Rightarrow$ Kovarianzmatrix $\tau^2\boldsymbol{K}^{-1}$ existiert nicht bzw. gemeinsame Priori ist eine teilweise uneigentlichen Verteilung
- trotzdem eigentliche Posteriori-Verteilung (NV) mit

$$E(\gamma|\boldsymbol{y}) = \left(\boldsymbol{B}'\boldsymbol{B} + \frac{\sigma^2}{\tau^2}\boldsymbol{K}\right)^{-1}\boldsymbol{B}'\boldsymbol{y}, \quad \text{Cov}(\gamma|\boldsymbol{y}) = \left(\boldsymbol{B}'\boldsymbol{B} + \frac{\sigma^2}{\tau^2}\boldsymbol{K}\right)^{-1}.$$

Variierende Koeffizienten

- Glatte Interaktion $g(z)x$ zwischen metrischer Variable z, binärer Variable x.
- Das Modell sollte dann auch die Haupteffekte $x\beta_1$ und $f(z)$ enthalten,

$$y_i = \beta_0 + x_i\beta_1 + f(z_i) + g(z_i)x_i + \varepsilon_i.$$

  - $f(z)$ ist der nichtlineare Effekt von z, falls $x = 0$,
  - $f(z) + g(z) + \beta_1$ ist der nichtlineare Effekt von z, falls $x = 1$.

  Zentriere $g(z)$ um Null, damit das Modell identifizierbar ist:
  $g(z)x + \beta_1 x = (g(z) + c)x + (\beta_1 - c)x$.

- $g(z)$ lässt sich wieder über Basisfunktionen approximieren. Die Designmatrix enthält das Produkt der Werte der Basisfunktionen und der x-Werte.
- Eine geeignete Strafmatrix ist wie zuvor z.B. $\boldsymbol{K}_k = \boldsymbol{D}_k'\boldsymbol{D}_k$ für B-Splines.
- Schätzung über das lineare gemischte Modell analog zu penalisierten Splines. Analog für x stetig oder kategorial. Details in Fahrmeir et al (2009, 8.3-8.4).

Bivariate Glättung und räumliche Effekte

Für die Darstellung von räumlichen Effekten oder nichtparametrischen Interaktionsoberflächen lässt sich die Methodik der P-Splines verallgemeinern:

$$y_i = f(z_{i1}, z_{i2}) + \varepsilon_i \approx \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \gamma_{jk} B_{jk}(z_{i1}, z_{i2}) + \varepsilon_i$$

mit z.B.

$$B_{jk}(z_1, z_2) = B_j^{(1)}(z_1) B_k^{(2)}(z_2), \quad j = 1, \ldots, d_1, \quad k = 1, \ldots, d_2$$

eine Tensorprodukt-Basis (Produkte aller univariater B-Spline-Basisfunktionen). Ein geeigneter Strafterm ist hier z.B. (mit Kronecker-Produkt $\otimes$, analog für $\boldsymbol{D}_k$)

$$\begin{aligned} \lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma} &= \lambda \Big[\sum_{j=1}^{d_1} \sum_{k=2}^{d_2} (\gamma_{jk} - \gamma_{j,k-1})^2 + \sum_{k=1}^{d_2} \sum_{j=2}^{d_1} (\gamma_{jk} - \gamma_{j-1,k})^2 \Big] \\ &= \lambda \boldsymbol{\gamma}' [(\boldsymbol{I}_{\boldsymbol{d_2}} \otimes \boldsymbol{D}_1)' (\boldsymbol{I}_{\boldsymbol{d_2}} \otimes \boldsymbol{D}_1) + (\boldsymbol{D}_1 \otimes \boldsymbol{I}_{\boldsymbol{d_1}})' (\boldsymbol{D}_1 \otimes \boldsymbol{I}_{\boldsymbol{d_1}})] \boldsymbol{\gamma}. \end{aligned}$$

Die Schätzung über gemischte Modelle läuft wie für allgemeine Penalisierungsansätze beschrieben. Für Details, weitere (z.B. radiale) Basisfunktionen und Beziehungen zum Kriging in der räumlichen Statistik, siehe Fahrmeir et al (2009, Abschnitt 7.2).

Für höherdimensionale Glättung, siehe Fahrmeir et al (2009, Abschnitt 7.3).

Markov-Zufallsfelder

Bei räumlichen Daten in Regionen oder auf einem Gitter wird räumliche Glättung meist über Nachbarschaften umgesetzt:

$$r \in N(s) \Leftrightarrow r \text{ ist zu } s \text{ benachbart.}$$

Verwenden wir ein Intercept γ_s pro Region s, so ist ein möglicher Strafterm

$$\lambda \sum_{s=1}^{d} \sum_{r \in N(s), r<s} (\gamma_r - \gamma_s)^2 = \lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}$$

mit $\boldsymbol{K} = (K_{sr})_{sr}$, $K_{sr} = -1$ für $r \in N(s)$, $K_{sr} = |N(s)|$ für $r = s$, $K_{sr} = 0$, sonst. Eine Schätzung über gemischte Modelle ist wieder möglich, für Details und die Beziehung zu Markov-Zufallsfeldern siehe Fahrmeir et al (2009, Abschnitt 7.2.4).

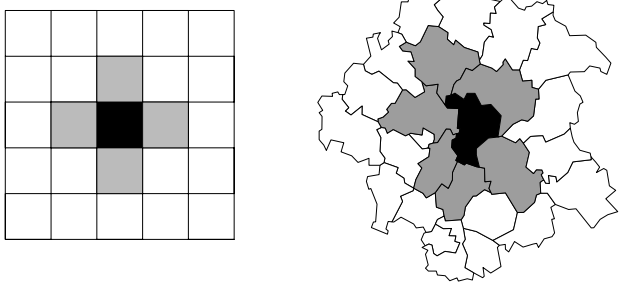


Abbildung 6: Nachbarschaften 1. Ordnung für reguläres Gitter und irreguläre Regionen. Die Nachbarn der schwarz gekennzeichneten Region sind in grau wiedergegeben. Quelle: Fahrmeir et al (2009, Abb. 7.45).

Ordinale Kovariablen

Bei einer Kovariablen mit geordneten Kategorien $k = 0, \ldots, K$, bei der ein glatter Anstieg/Abfall der Koeffizienten $\gamma_0 = 0, \gamma_1, \ldots, \gamma_K$ angenommen wird, schlagen

- Gertheiss, J. and G. Tutz (2009). *Penalized regression with ordinal predictors.* International Statistical Review 77, 345–365.

den Strafterm

$$\lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma} \quad = \quad \lambda \sum_{j=1}^{K} (\gamma_j - \gamma_{j-1})^2 = \lambda \boldsymbol{\gamma}' \boldsymbol{D}_1' \boldsymbol{D}_1 \boldsymbol{\gamma}$$

vor. Die Schätzung über ein LMM läuft analog zu penalisierten Splines. Der Test

$$H_0 : \gamma_1 = \cdots = \gamma_K = \gamma_0 = 0 \quad \text{gegen} \quad H_A : \gamma_k \neq 0 \text{ für mindestens ein } k \qquad (27)$$

auf Einfluss der Kovariablen über einen (R)LRT

$$H_0 : \tau^2 = 0 \quad \text{gegen} \quad H_A : \tau^2 > 0$$

im zugehörigen LMM hat größere Power als ein F-Test für (27) im linearen Modell:

- Gertheiss, J. and F. Oehrlein (2011). *Testing linearity and relevance of ordinal predictors.* Electronic Journal of Statistics 5, 1935-1959.

Lineare Gemischte Modelle als Modulare Modelle

Die Beispiele zeigen, dass LMMs ein mächtiges Werkzeug sind, mit dem flexibel Modellbausteine in strukturiert-additiven Regressionsmodellen modular kombiniert werden können:

- zufällige Effekte für z.B. Longitudinal- oder Clusterdaten
- räumliche Terme
- glatte Terme oder Interaktionen, variierende Koeffizienten
- ...

Damit sind z.B. Modelle der Form

$$y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + b_{i0} + u_{ij1}b_{i1} + f_1(z_{ij1}) + f_2(z_{ij2}, z_{ij3}) + f_3(\boldsymbol{s}_{ij}) + f_4(z_{ij4})x_{ij,p+1} + \varepsilon_{ij}$$

mit Kovariablen $\boldsymbol{x}$ und $\boldsymbol{u}$, metrischen Kovariablen $\boldsymbol{z}$, räumlichen Koordinaten $\boldsymbol{s}$, festen Effekten $\boldsymbol{\beta}$ und zufälligen Effekten $\boldsymbol{b}$ möglich.

Software

Penalisierungsansätze können formal als LMM aufgeschrieben werden, weisen jedoch nicht die sonst typische Gruppierungsstruktur auf.
$\Rightarrow$ Verwende zur Schätzung spezialisierte Software, die an diese Struktur angepasst ist (und die nötigen Designmatrizen $\boldsymbol{X}$ und $\boldsymbol{Z}$ automatisch konstruiert).

Hier ist insbesonder das Paket `mgcv` von Simon Wood zu nennen.
Die `gamm()`-Funktion oder `gam()` mit Option `method = "(RE)ML"` fitten (generalisierte) additive gemischte Modelle mit (RE)ML-Schätzung.

- Feste und zufällige Effekte werden wie in `lme()` spezifiziert.
- Glatte Effekte $f(z)$ über `s(z)`.
- Glatte räumliche oder Interaktions-Oberflächen über `s(`z_1`,`z_2`)` oder `te(`z_1`,`z_2`)`.
- Für variierende Koeffizienten ist das `by`-Argument vorgesehen, `s(z, by=x)`.
- Weitere Möglichkeiten und Optionen, siehe die Manual sowie Wood (2006).

Inhalt

Generalisiertes Lineares Modell (GLM): Strukturannahme

- Strukturannahme: Verknüpfung von Erwartungswert und linearem Prädiktor $\eta_i = \boldsymbol{x}_i'\beta$ durch die Linkfunktion $g(\cdot)$:

$$E(y_i|\boldsymbol{x}_i) = g^{-1}(\eta_i) \Leftrightarrow g(E(y_i|\boldsymbol{x}_i)) = \eta_i$$

  - Binäre Daten: z.B. Logitmodell: $\mathrm{E}(y_i|\boldsymbol{x}_i) = P(y_i = 1|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i'\beta)}{1+\exp(\boldsymbol{x}_i'\beta)}$
  - Zähldaten: z.B. log-lineares Modell: $\mathrm{E}(y_i|\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i'\beta)$

Generalisiertes Lineares Modell (GLM): Verteilungsannahme

- Verteilungsannahme: Gegeben η_i sind die y_i unabhängig verteilt und ihre Dichte $f(y_i)$ gehört zu einer Exponentialfamilie:

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

$$\text{mit } \mathrm{E}(y_i|\boldsymbol{x}_i) = b'(\theta_i) = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta});\ \mathrm{Var}(y_i|\boldsymbol{x}_i) = \phi b''(\theta_i);$$

  - θ_i: natürlicher/kanonischer Parameter
  - ϕ_i: Skalenparameter ($\phi_i = 1$ für Binär-/Poisson-Daten)

Die Exponentialfamilie - Beispiele

Dichten / Wahrscheinlichkeitsfunktionen:

$$\text{Normal}: \quad f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y-\mu)^2}{2\sigma^2})$$

$$\text{Binomial}: \quad f(y|\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$$

$$\text{Poisson}: \quad f(y|\lambda) = \frac{\lambda^y}{y!} \exp(-\lambda)$$

	μ	θ	ϕ	$b(\theta)$
Normal	μ	μ	σ^2	$\frac{1}{2}\theta^2$
Binomial	π	$\log(\frac{\pi}{1-\pi})$	1	$n\log(1+\exp(\theta))$
Poisson	λ	$\log(\lambda)$	1	$\exp(\theta)$

GLM

- Strukturannahme:

$$E(y_i|\boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta})$$

- Verteilungsannahme:

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

$\Rightarrow$ Verknüpfung von Erwartungswertstruktur und Varianzstruktur

- $\mathrm{E}(y_i|\boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}) = b'(\theta_i)$
- $\mathrm{Var}(y_i|\boldsymbol{x}_i) = \phi_i b''(\theta_i)$

GLM

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

- $\mathrm{E}(y_i|\boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}) = b'(\theta_i)$
- $\mathrm{Var}(y_i|\boldsymbol{x}_i) = \phi_i b''(\theta_i)$

Also:

$$\begin{array}{ccccc} \boldsymbol{\eta} & & & & \boldsymbol{\theta} \\ & \overset{g^{-1}}{\searrow} & & \overset{b'}{\swarrow} \quad \overset{\phi b''}{\searrow} & \\ & \mathrm{E}(\boldsymbol{y}|\boldsymbol{\theta} \text{ bzw. } \boldsymbol{\eta}) & & & \mathrm{Var}(\boldsymbol{y}|\boldsymbol{\theta}) \end{array}$$

GLM mit kanonischem Link

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

- $\mathrm{E}(y_i|\boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}) = b'(\theta_i)$
- $\mathrm{Var}(y_i|\boldsymbol{x}_i) = \phi b''(\theta_i)$
- kanonische Linkfunktion: $g^{-1}(\cdot) = b'(\cdot)$

Also:

$$\boldsymbol{\eta} \quad \left(\overset{\text{kanonische Linkfkt.}}{=}\right) \quad \boldsymbol{\theta}$$

$$\overset{g^{-1}}{\searrow} \qquad \overset{b'}{\swarrow} \qquad \overset{\phi b''}{\searrow}$$

$$\mathrm{E}(\boldsymbol{y}|\boldsymbol{\theta} \text{ bzw. } \boldsymbol{\eta}) \qquad \mathrm{Var}(\boldsymbol{y}|\boldsymbol{\theta})$$

- Beispiele: $g(\cdot) = \mathrm{id}(\cdot)$ für Normalverteilung, $g(\cdot) = \mathrm{logit}(\cdot)$ für Bernoulliverteilung, $g(\cdot) = \log(\cdot)$ für Poissonverteilung

LM $\Rightarrow$ LMM

- Was ist anders/neu im gemischten Modell?
- Wozu gemischte Modelle?
- Warum sind gemischte Modelle so beliebt?

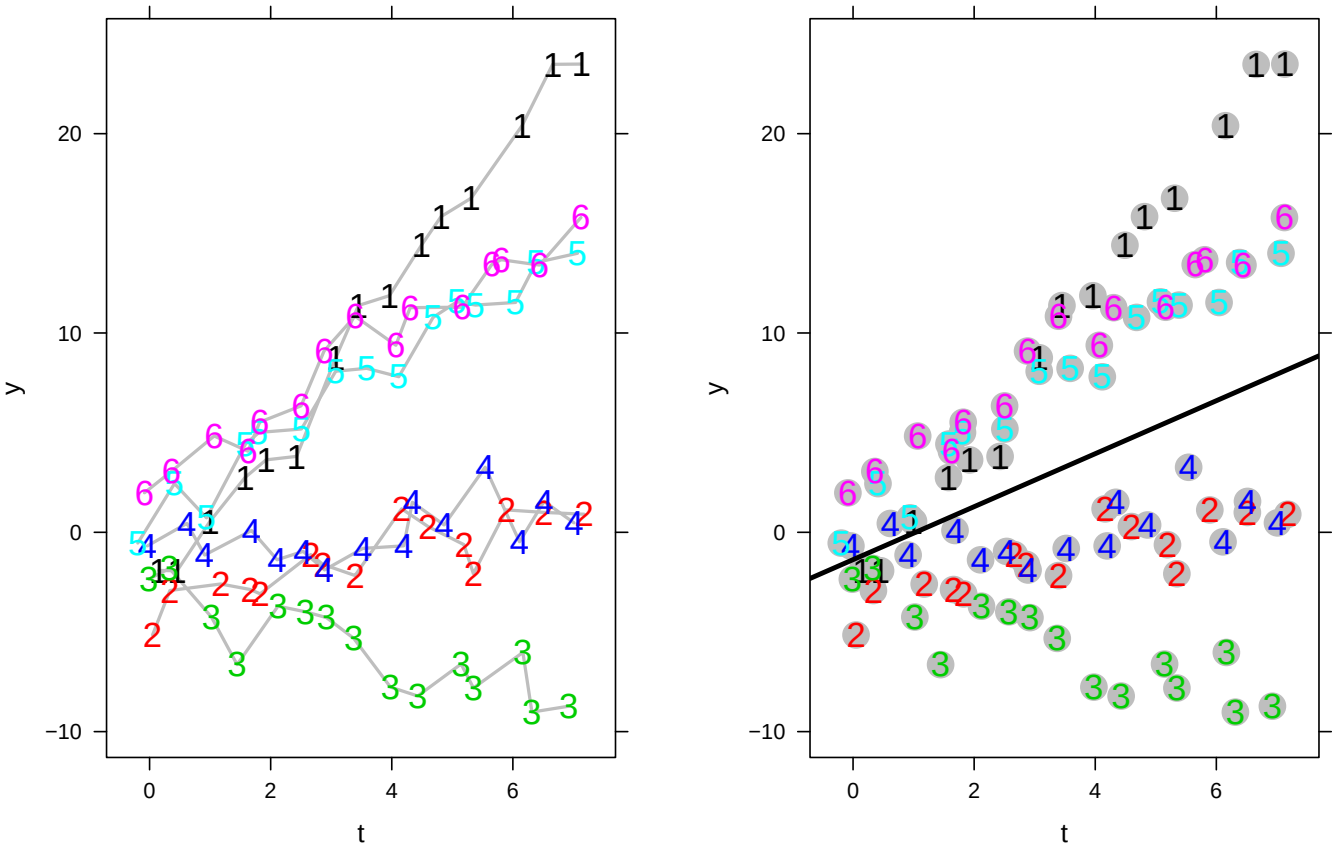
Formal: Was ist anders/neu im LMM?

- Erweiterung des linearen Prädiktors um zufällige/penalisierte Effekte $\boldsymbol{b}$:
  $\eta = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b}$
- Verteilungsannahme über $\boldsymbol{b}$: $\boldsymbol{b} \sim \mathcal{P}(\vartheta)$
  üblich: $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G}(\vartheta))$
- Inferenz über gemeinsame / penalisierte (Log-)Likelihood:
  $l_{pen}(\boldsymbol{y}|\beta, \boldsymbol{b}, \vartheta) = l(\boldsymbol{y}|\beta, \boldsymbol{b}, \vartheta) + l(\boldsymbol{b}|\vartheta)$
- konditionales Modell / marginales Modell:
  $\boldsymbol{Z}$ und $\boldsymbol{G}(\vartheta)$ bestimmen marginale Kovarianzstruktur von $\boldsymbol{y}$.

$$\boldsymbol{y}|\boldsymbol{b} \sim N(\boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b}, \boldsymbol{R}(\vartheta)); \; \boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G}(\vartheta))$$
$$\Rightarrow \boldsymbol{y} \sim N(\boldsymbol{X}\beta, \boldsymbol{Z}\boldsymbol{G}(\vartheta)\boldsymbol{Z}' + \boldsymbol{R}(\vartheta))$$

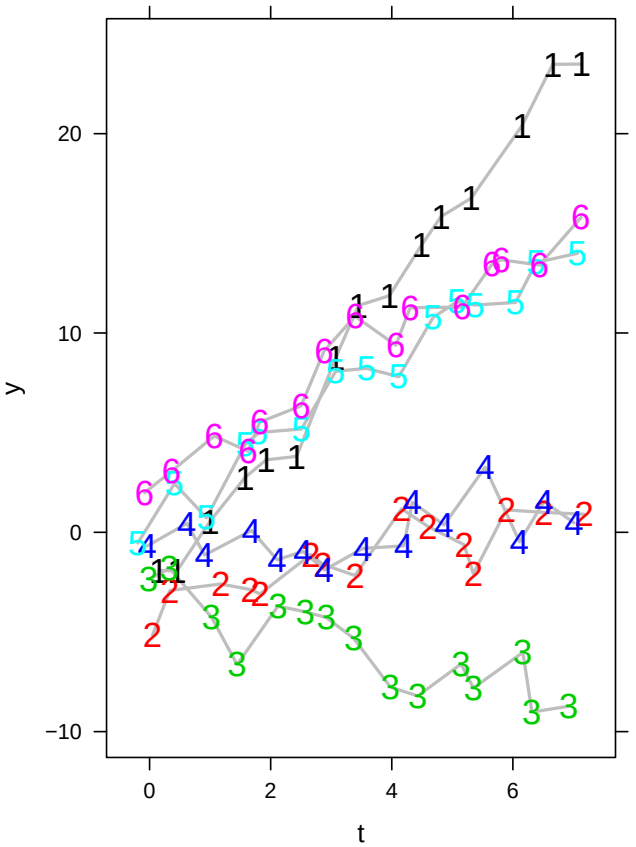
Inhaltlich: Wozu LMMs?

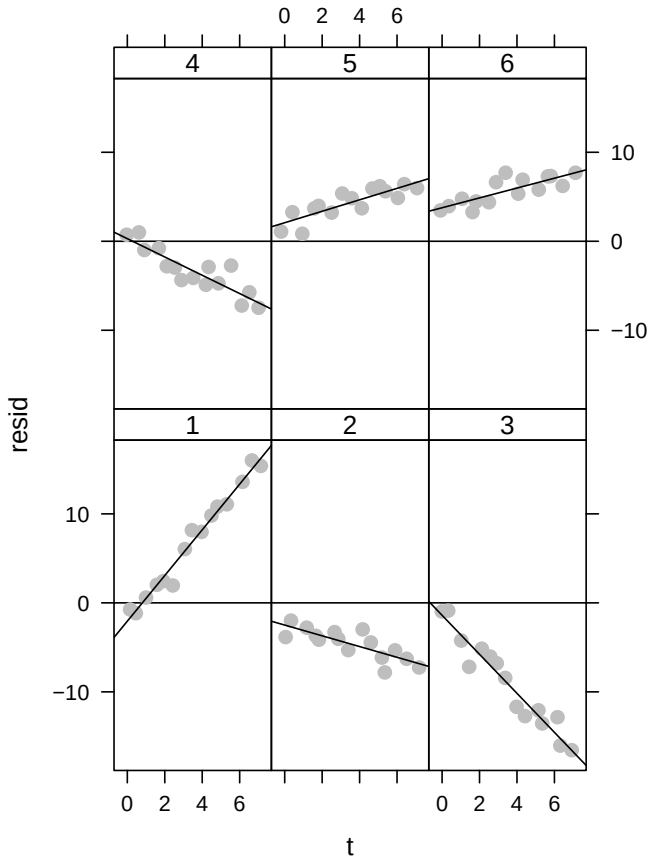
Konditional betrachtet:
(zeitliche/räumliche/hierarchische) Struktur der Daten legt bestimmte Struktur der Abweichungen der Beobachtungen $\boldsymbol{y}$ von $\boldsymbol{X}\beta$ nahe:



Inhaltlich: Wozu LMMs?

Konditional betrachtet:
(zeitliche/räumliche/hierarchische) Struktur der Daten legt bestimmte Struktur der Abweichungen der Beobachtungen $\boldsymbol{y}$ von $\boldsymbol{X}\beta$ nahe:

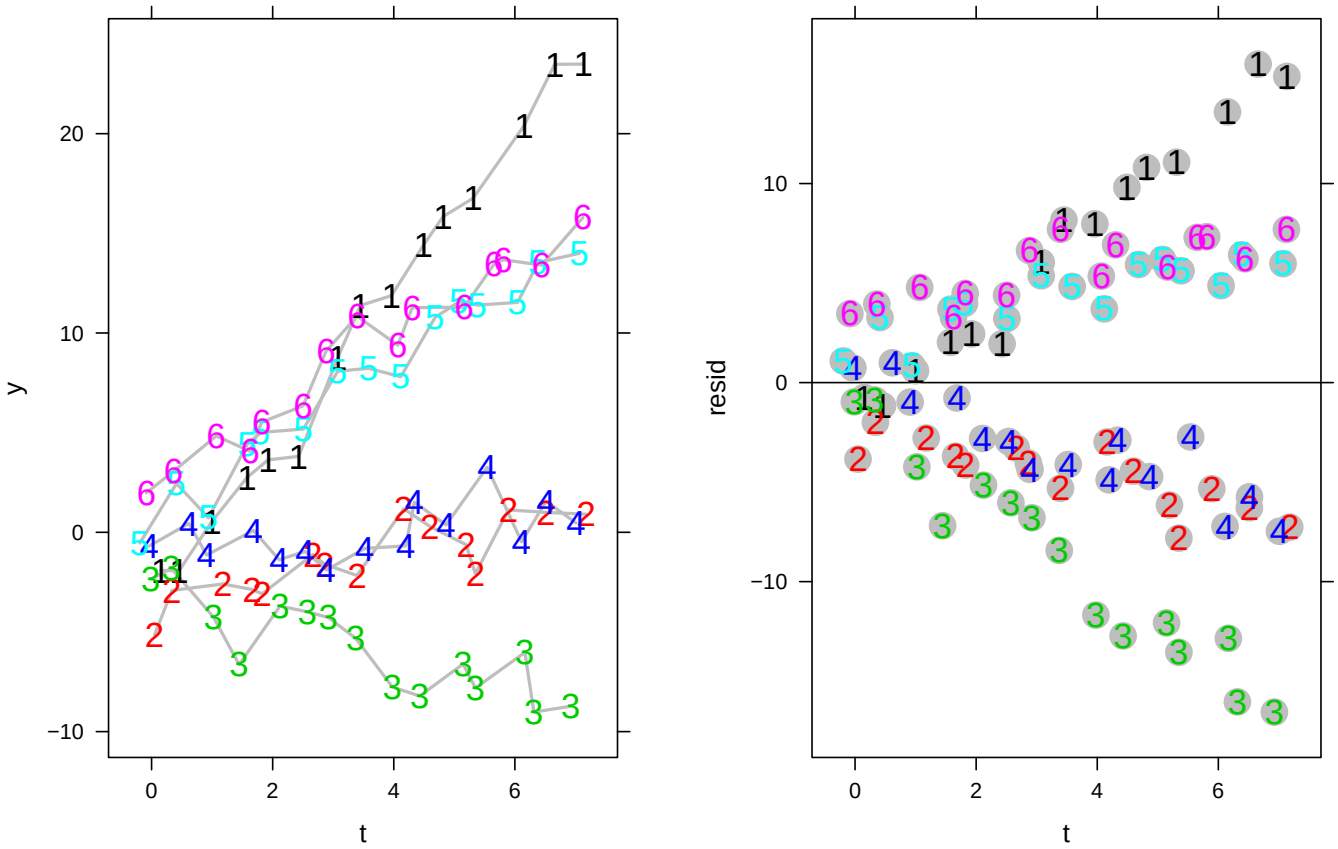




Inhaltlich: Wozu LMMs?

Marginal betrachtet:

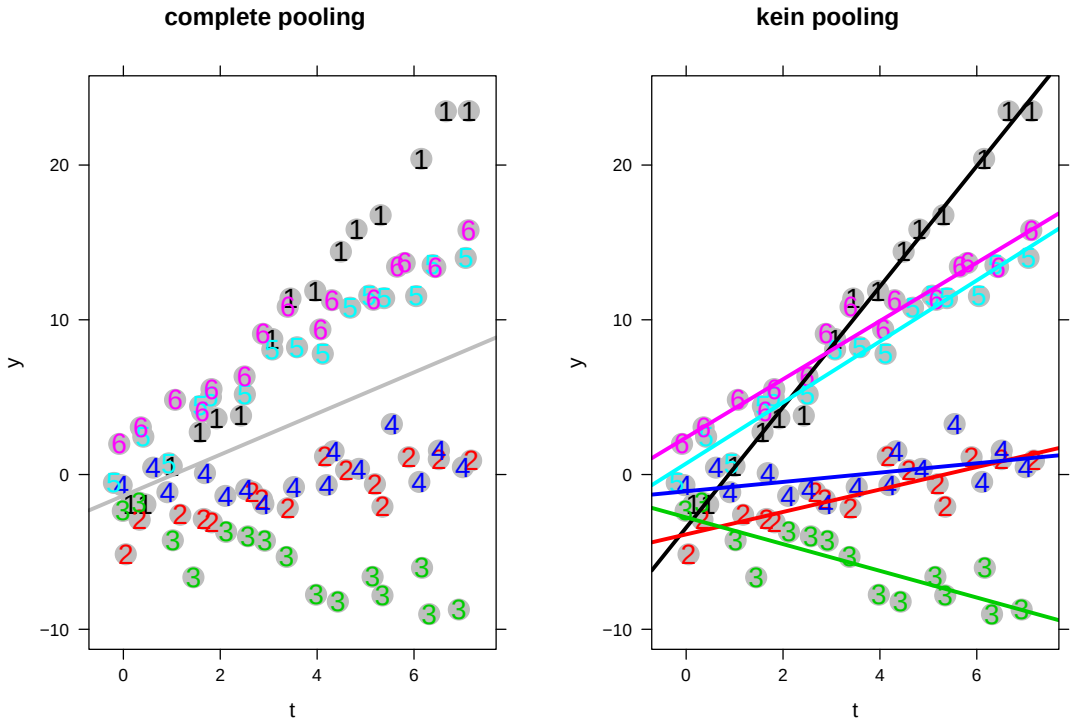
(zeitliche/räumliche/hierarchische) Struktur der Daten legt bestimmte Kovarianzstruktur der Beobachtungen nahe:



Warum sind LMMs so beliebt?

LMM bilden Kompromiss zwischen *komplettem pooling* und *keinem pooling*

- *komplettes pooling*: gewöhnliches LM ohne Einbezug der Gruppierungsstruktur
- *kein pooling*: separates LM (oder ein fester Effekt) für jede Ausprägung der Gruppierung



Warum sind LMMs so beliebt?

- LMM für hierarchische Daten als Kompromiss zwischen *komplettem pooling* und *keinem pooling*: Schätzung der Varianzparameter $\boldsymbol{\vartheta}$ erfasst Heterogenität zwischen Gruppen auf Basis des gesamten Datensatzes $\rightarrow$ Shrinkage der Koeffizienten zum Populationsmittel entsprechend stärker oder schwächer.
- penalisiertes Likelihoodkriterium stabilisiert Parameterschätzung (aber: Bias-Varianz-Tradeoff!)
- bayesianisch: priori-Annahmen über $\boldsymbol{b}$ bilden Vorwissen über Struktur der Daten ab. Einführung dieser zusätzlichen Information erlaubt stabile Schätzung von Modellen mit sehr vielen Parametern.
- sehr vielseitig: Modelle für zeitliche, räumliche, hierarchische und gruppierte Datenstrukturen sowie nichtlineare Effekte lassen sich als (G)LMM auffassen.

LMM $\Rightarrow$ GLMM: Flexibilisierung der Verteilungsannahme

- LMM: nur normalverteilter Response: $\boldsymbol{y}|\eta \sim N(\eta, \boldsymbol{R})$
- GLMM: Response aus beliebiger Exponentialfamilie:

$$\mathrm{E}(\boldsymbol{y}|\eta) = g^{-1}(\boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b})$$
$$\boldsymbol{y}|\eta \sim \text{Expo.fam.}$$

GLMM = GLM + LMM

- 1. Stufe:
  y_i bedingt auf den erweiterten linearen Prädiktor $\boldsymbol{X}\beta + \boldsymbol{Zb}$
  unabhängig verteilt nach Exponentialfamilie (analog GLM)

- 2. Stufe:
  Verteilungsannahme über $\boldsymbol{b} \Rightarrow$ regularisierte Schätzung (analog LMM)
  Interpretation der Verteilungsannahme als Strafterm oder Eigenschaft der Grundgesamtheit (freq.) oder Priori-Annahme (bayes.)

GLMM für Longitudinal-/Clusterdaten

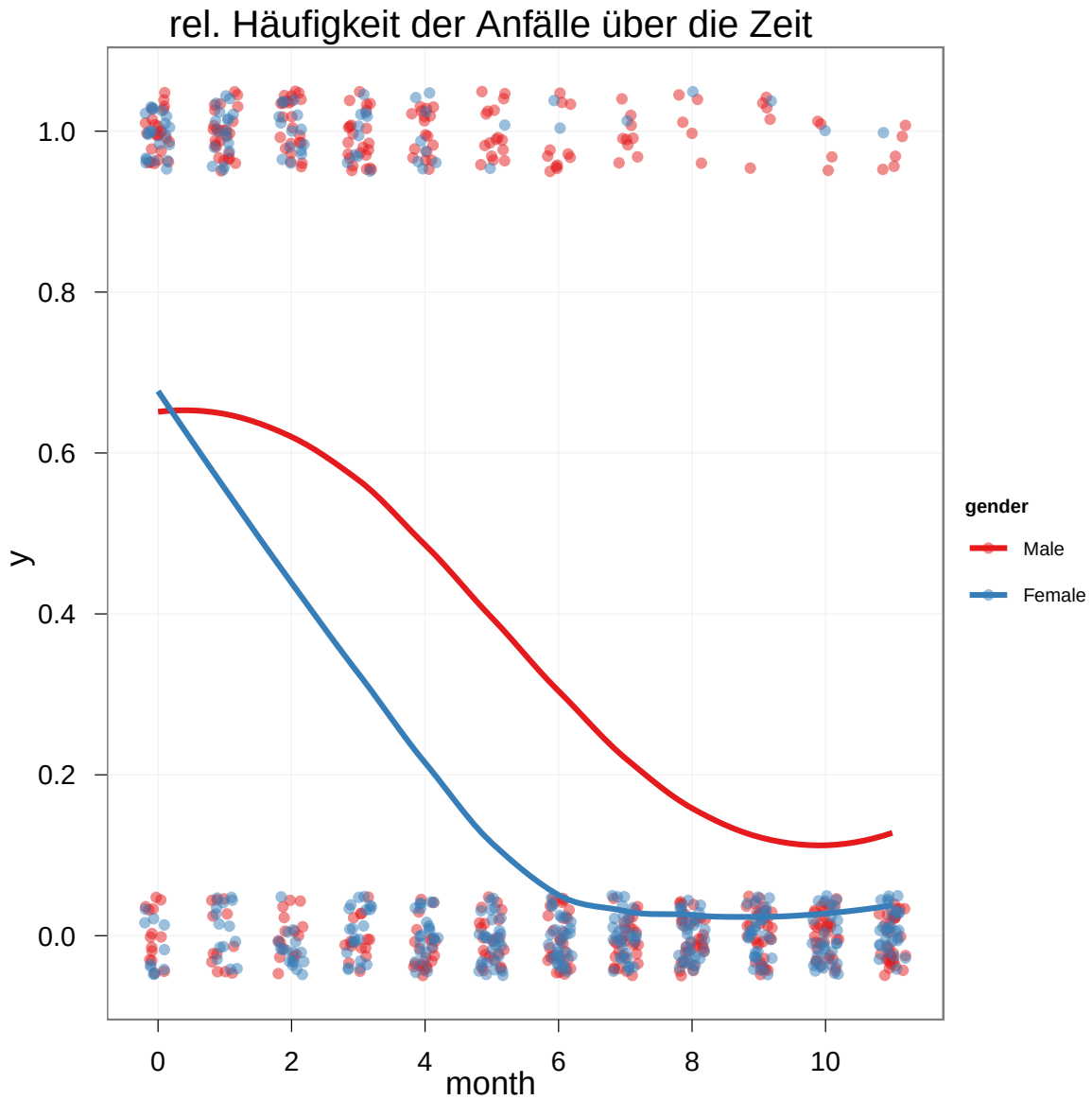
- y_{ij}; $i = 1, \ldots, N$; $j = 1, \ldots, n_i$ mit je n_i Messwiederholungen an N Beobachtungseinheiten (z.B. Binär-/Zähldaten)
- Bedingte Verteilung von $y_{ij}|\boldsymbol{b}_i$ ist aus Exponentialfamilie (z.B. Bernoulli/Poisson)
- Bedingter Erwartungswert $\mu_{ij} = \mathrm{E}(y_{ij}|\boldsymbol{b}) = g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i)$
- Die zufälligen Effekte $\boldsymbol{b}_i$ sind normalverteilt,

$$\boldsymbol{b_i} \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{D}).$$

▸ Allg. GLMM

Beispiel: Madras-Daten

- 69 stationäre Schizophrenie-Patienten
- 1 Jahr lang nach Ersteinlieferung beobachtet
- 12 monatliche Beobachtungen pro Patient ob Schub (ja/nein)



Beispiel: Madras-Daten

- Response: $\mathtt{y}_{it} \in 0, 1$: Patient/in i hat Schub im Monat t
- Kovariablen:
  - `month`: Zeit in Monaten nach Einlieferung (0-11)
  - `gender`: Geschlecht
  - `id`: Patient
- Logit-Modell mit zufälligem Intercept: $\rightarrow$ **Übung**

$$\text{logit}(P(y_{it} = 1|b_i)) = \beta_0 + b_i + \beta_M \text{month}_t + \beta_G \text{gender}_i$$

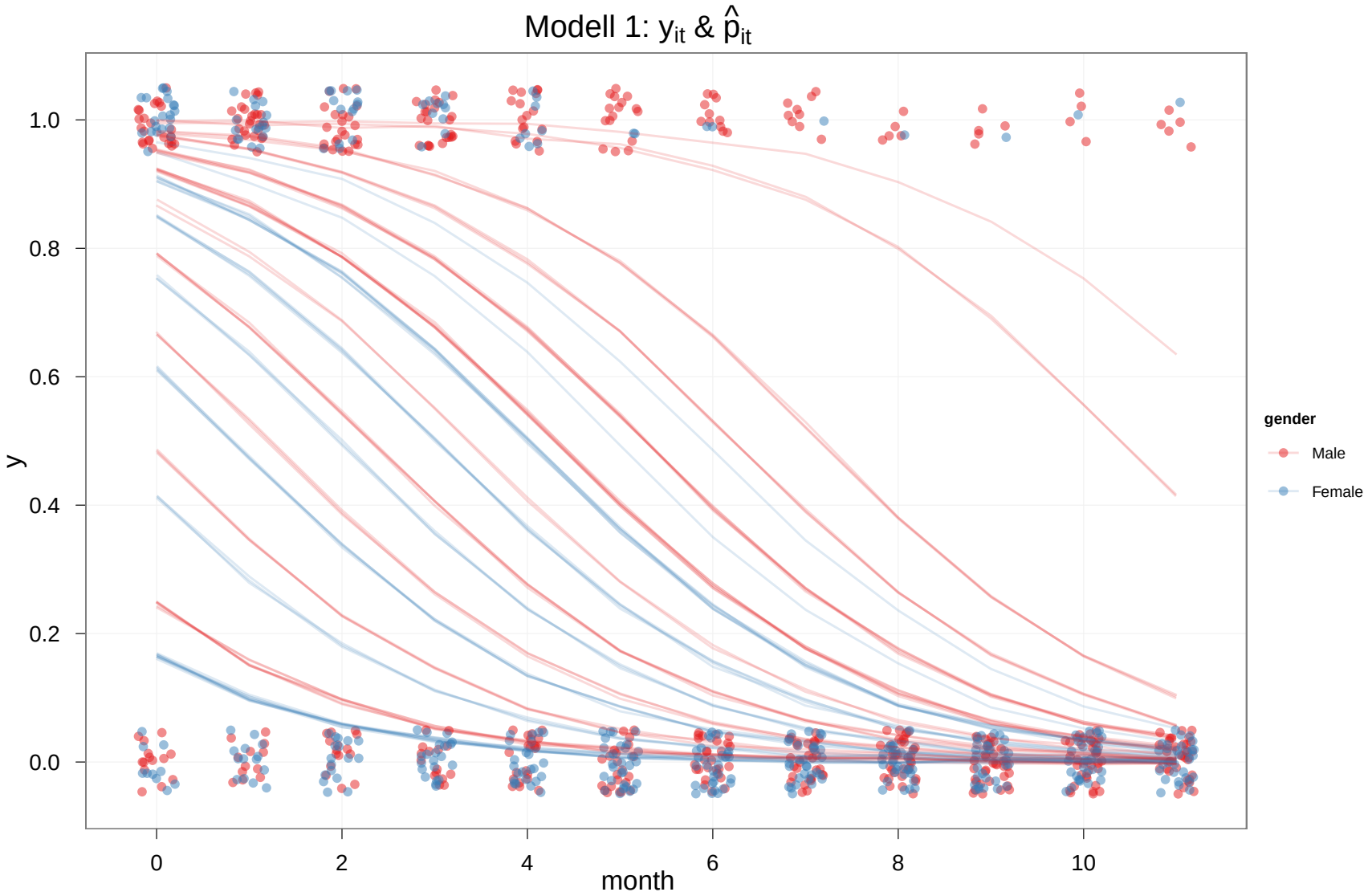
Beispiel: Madras-Daten - Ergebnisse

```
Random effects:
Groups Name        Variance Std.Dev.
id     (Intercept) 4.91     2.22
Number of obs: 828, groups: id, 69

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.933      0.437    4.42  9.8e-06 ***
month          -0.573      0.045  -12.72  < 2e-16 ***
genderFemale   -1.574      0.595   -2.65   0.0081 **
```

```
R> quantile(exp(ranef(m1)$id$'(Intercept)'), p=c(.2,.4,.6,.8))

   20%   40%   60%   80%
0.138 0.634 2.167 5.828
```

$\Rightarrow$ relativ starke individuelle Heterogenität

Beispiel: Madras-Daten - Ergebnisse



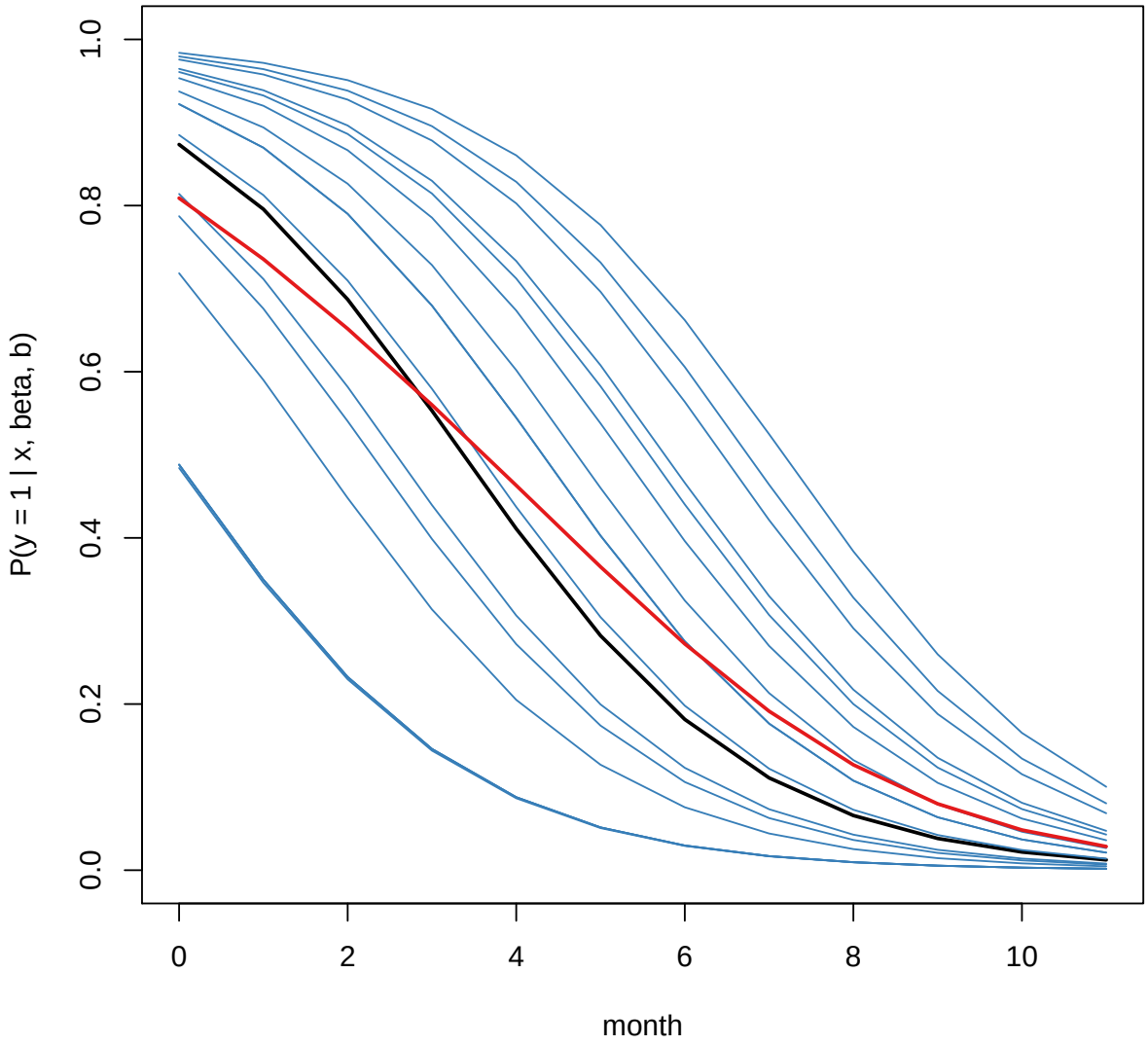
Beispiel: Madras-Daten - Interpretation

Wie sind die Parameter in diesem Logit-Modell mit Random Intercept zu interpretieren?

	`beta`	`exp(beta)`
`month`	-0.573	0.564

Graphisch:

- $\frac{\exp(\beta_0+\beta_1 \text{month}_t)}{1+\exp(\beta_0+\beta_1 \text{month}_t)}$
- $P(y=1|x,\beta,b)$ für Quantile von b
- $\mathrm{E}_b[P(y=1|x,\beta,b) = P(y=1|x,\beta)$



Interpretation Random-Intercept-Logitmodell

Konditional: Odds-Ratio

$$\frac{\frac{P(y_{it}=1|x_{it}+1,b_i)}{P(y_{it}=0|x_{it}+1,b_i)}}{\frac{P(y_{it}=1|x_{it},b_i)}{P(y_{it}=0|x_{it},b_i)}} = \frac{\exp((x_{it}+1)\beta+b_i)}{\exp(x_{it}\beta+b_i)} = \exp(\beta)$$

Marginal: Odds-Ratio

$$\frac{\frac{P(y_{it}=1|x_{it}+1)}{P(y_{it}=0|x_{it}+1)}}{\frac{P(y_{it}=1|x_{it})}{P(y_{it}=0|x_{it})}} = \frac{\frac{\mathrm{E}_{b_i}[P(y_{it}=1|x_{it}+1,b_i)]}{\mathrm{E}_{b_i}[P(y_{it}=0|x_{it}+1,b_i)]}}{\frac{\mathrm{E}_{b_i}[P(y_{it}=1|x_{it},b_i)]}{\mathrm{E}_{b_i}[P(y_{it}=0|x_{it},b_i)]}} \neq \exp(\beta)$$

$$\text{mit } \mathrm{E}_{b_i}[P(y_{it}=1|x_{it},b_i)] = \int \frac{\exp(x_{it}\beta+b_i)}{1+\exp(x_{it}\beta+b_i)} p(b_i)db_i \quad \text{etc.}$$

$\Rightarrow$ Interpretation von β nur bedingt auf zufällige Effekte b_i zulässig

$\Rightarrow$ β ist individuenspezifischer Parameter, nicht Populationsparameter

Beispiel: Madras-Daten - Interpretation

Wie sind die Parameter in diesem Logit-Modell mit Random Intercept zu interpretieren?

	`beta`	`exp(beta)`
`month`	-0.573	0.564

GLMM in allgemeiner Form

- Verteilungsannahme I: Gegeben die zufälligen Effekte $\boldsymbol{b}$ sind die y_i bedingt unabhängig und die bedingte Dichte $f(y_i|\boldsymbol{b})$ gehört zur Exponentialfamilie
- Strukturannahme: Der **bedingte** Erwartungswert $\mu_i = \mathrm{E}(y_i|\boldsymbol{b}_i)$ ist mit dem linearen Prädiktor

$$\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{b}$$

  durch $\mu_i = g^{-1}(\eta_i)$ verknüpft mit bekannter Linkfunktion $g(\cdot)$.
- Verteilungsannahme II: Die zufälligen Effekte $\boldsymbol{b}$ sind normalverteilt,

$$\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G})$$

  (auch andere Verteilungen möglich).

Hierarchisches GLMM Spezialfall mit $\boldsymbol{X} = (\boldsymbol{X}_1'|\ldots|\boldsymbol{X}_N')'$, $\boldsymbol{Z} = \mathrm{diag}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)$ und $\boldsymbol{G} = \mathrm{diag}(\boldsymbol{D}, \ldots, \boldsymbol{D})$ analog zum hierarchischen LMM. ▸ Hierarch. GLMM

GLMM: Marginale und konditionale Verteilung

Erinnerung:

$$\mathsf{E}(y_i|\eta_i) = g^{-1}(\boldsymbol{x_i'}\boldsymbol{\beta} + \boldsymbol{z_i'}\boldsymbol{b})$$

Für $g^{-1}(x) \neq x$ gilt:

$$\begin{aligned}\mathsf{E}_b(\mathsf{E}(y_i|\eta_i)) &= \int \mathsf{E}(y_i|\eta_i)p(\boldsymbol{b}|\boldsymbol{\vartheta})d\boldsymbol{b} \\ &= \int g^{-1}(\boldsymbol{x_i'}\boldsymbol{\beta} + \boldsymbol{z_i'}\boldsymbol{b})p(\boldsymbol{b}|\boldsymbol{\vartheta})d\boldsymbol{b} \\ &\neq g^{-1}(\boldsymbol{x_i'}\boldsymbol{\beta})\end{aligned}$$

GLMM mit $g^{-1}(x) \neq x$
$\Rightarrow$ marginaler Erwartungswert $\mathsf{E}_b(\mathsf{E}(y_i|\eta_i)) \neq g^{-1}(\boldsymbol{x_i'}\boldsymbol{\beta})$
Parameter $\boldsymbol{\beta}$ haben nur konditionale Interpretation.

GLMM: Marginale und konditionale Verteilung

- Konditionale Kovarianzstruktur: Gegeben die zufälligen Effekte $\boldsymbol{b}$ sind die y_i bedingt unabhängig

$$\text{Cov}(y_i, y_{i'}|\boldsymbol{b}) = 0, \quad i \neq i'.$$

(Es gibt keine Residuenkovarianzmatrix $\boldsymbol{R}$ wie im LMM.)

- Marginale Kovarianzstruktur: Es folgt für die marginale Kovarianz

$$\begin{aligned}
\text{Cov}(y_i, y_{i'}) &= \text{Cov}(\text{E}(y_i|\boldsymbol{b}), \text{E}(y_{i'}|\boldsymbol{b})) + \text{E}(\text{Cov}(y_i, y_{i'}|\boldsymbol{b})) \\
&= \text{Cov}(\mu_i, \mu_{i'}) + 0 \\
&= \text{Cov}(g^{-1}(\boldsymbol{x_i'}\beta + \boldsymbol{z_i'}\boldsymbol{b}), g^{-1}(\boldsymbol{x_{i'}'}\beta + \boldsymbol{z_{i'}'}\boldsymbol{b})), i \neq i'.
\end{aligned}$$

Im Allgemeinen hängt die marginale Korrelation von den Kovariablen $\boldsymbol{x_i}$ ab.

GLMM + AMM = GAMM

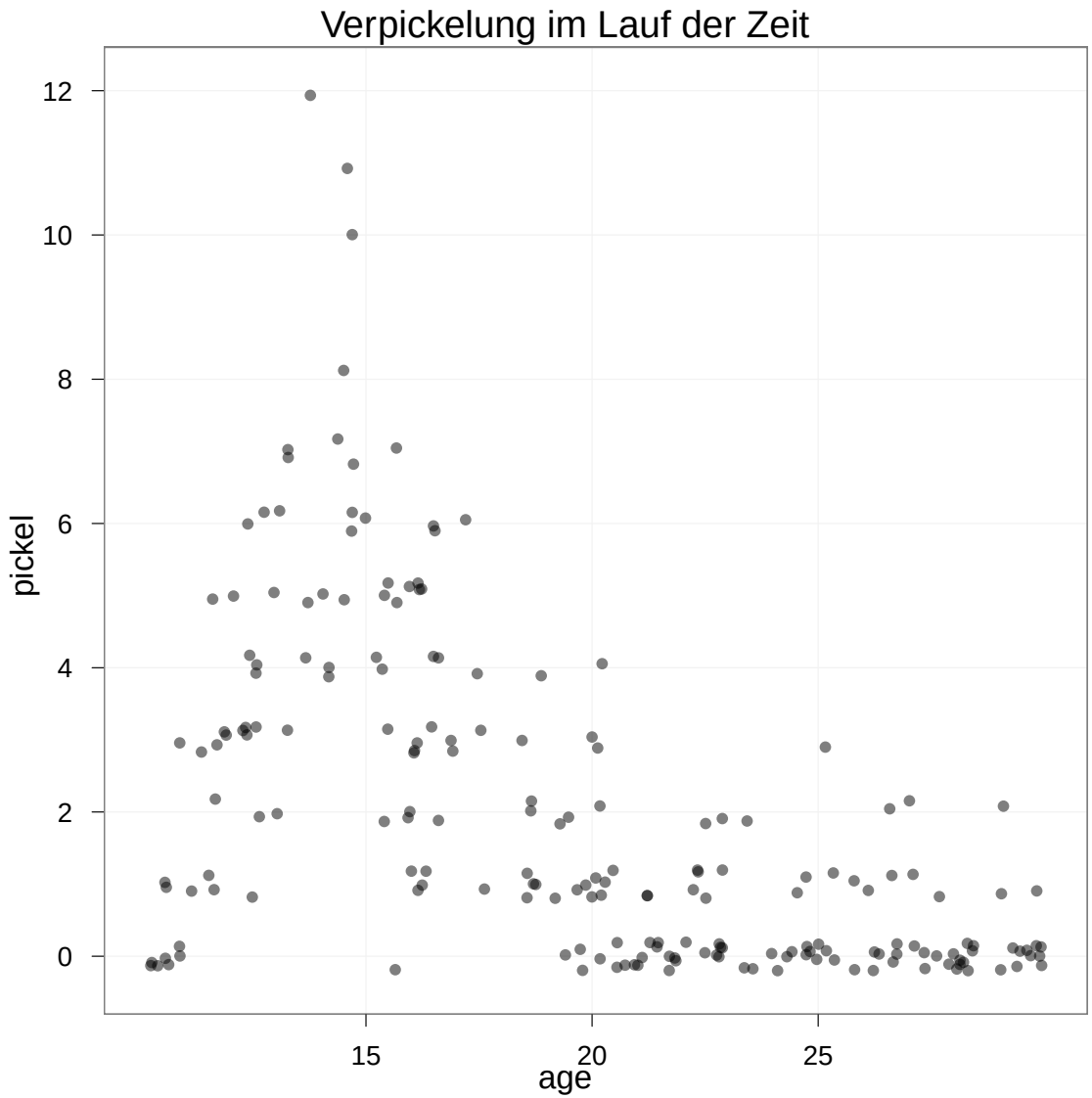
- Additive gemischte Modelle (AMMs) und
- Generalisierte lineare gemischte Modelle (GLMMs)

lassen sich zu generalisierten additiven gemischten Modellen (GAMMs) verbinden.

Hierzu werden die Designmatrizen $\boldsymbol{X}$ und $\boldsymbol{Z}$ und die Kovarianz $\boldsymbol{G}$ der zufälligen Effekte $\boldsymbol{b}$ im GLMM analog zu den additiven gemischten Modellen gewählt, siehe Kapitel 5.

Beispiel Verpickelung

- keine echten Daten
- Pickel auf den Gesichtern von 200 Teens und Twens wurden gezählt (1 Beob./Person)
- Verlauf der Verpickelung über die Zeit



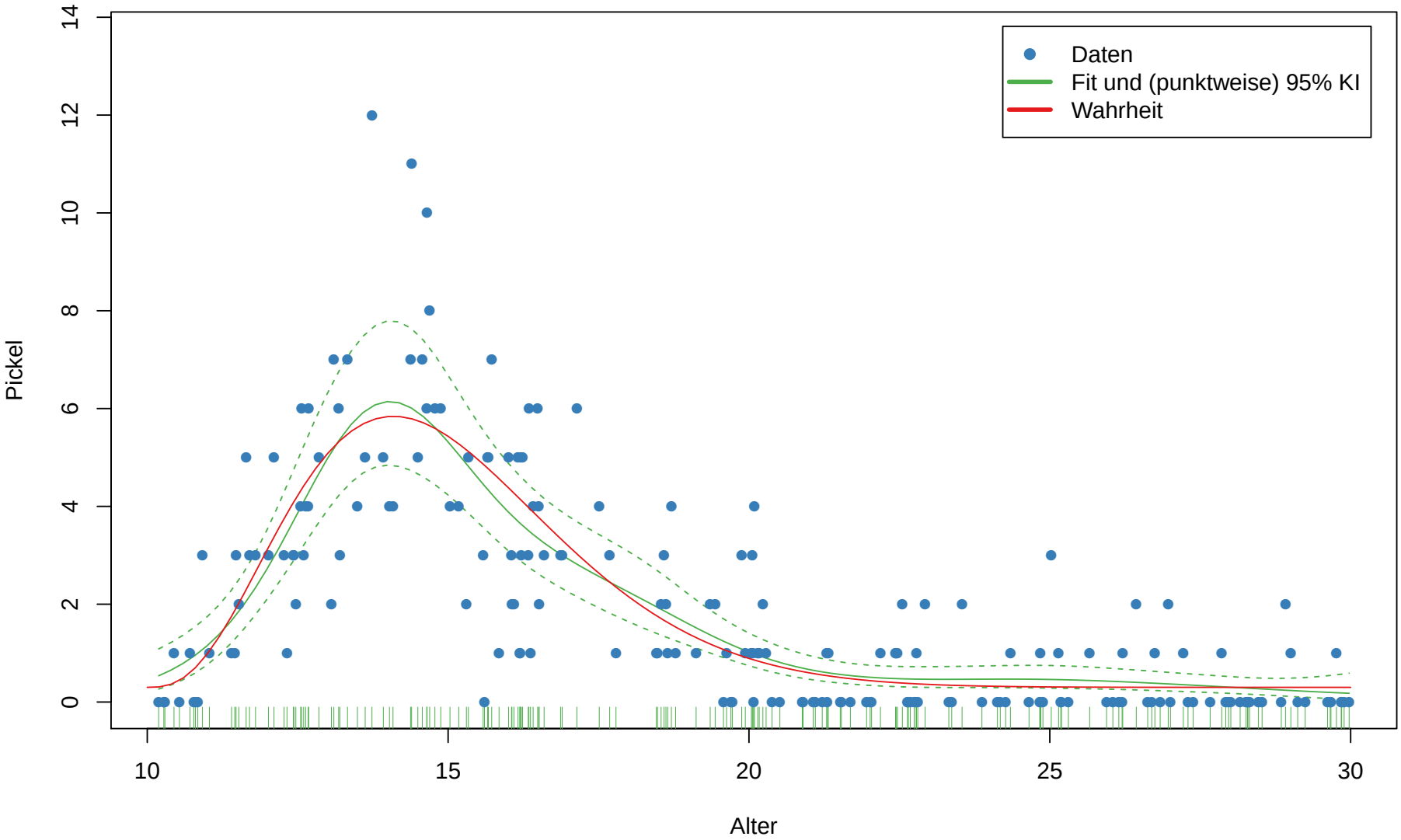
Beispiel Verpickelung: Modell

- Response $\texttt{pickel}_i$: Person i hat $\texttt{pickel}_i$ Pickel im Gesicht
- Kovariable $\texttt{alter}_i$: Alter in Jahren $(10 - 30)$
- Modell für $\textbf{pickel} = (\texttt{pickel}_1, \ldots, \texttt{pickel}_n)$ mit Splines aus der TP(2)-Basis für $\textbf{alter} = (\texttt{alter}_1, \ldots, \texttt{alter}_n)$:

$$
\begin{aligned}
\textbf{pickel}|\boldsymbol{b} &\sim \textit{Poisson}\left(\exp\left(f(\textbf{alter})\right)\right) \\
f(\textbf{alter}) &= \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} \\
\boldsymbol{X} &= (\mathbf{1}, \textbf{alter}, \textbf{alter}^2) \\
\boldsymbol{Z} &= \left((\texttt{alter}_i - \kappa_k)_+^2\right)_{i=1,\ldots,n;k=1,\ldots,m} \\
\boldsymbol{b} &\sim N(\mathbf{0}, \tau^2 \boldsymbol{I}_m)
\end{aligned}
$$

- Wir verwenden für den Fit auf der nächsten Folie quadratische B-Splines mit Differenzen 2. Ordnung. $\rightarrow$ **Übung**

Beispiel Verpickelung: Ergebnis



Inhalt

Likelihood eines GLMM

Verteilungsannahme für $\boldsymbol{y}|\boldsymbol{b}$: $y_i|\boldsymbol{b} \sim$ Expo.fam.(θ, ϕ) unabhängig
Verteilungsannahme für $\boldsymbol{b}|\vartheta$: $\boldsymbol{b}|\vartheta \sim N(\boldsymbol{0}, \boldsymbol{G}(\vartheta))$

$\Rightarrow$ gemeinsame / penalisierte Likelihood:

$$L(\beta, \phi, \vartheta, \boldsymbol{b}) = f(\boldsymbol{y}, \boldsymbol{b}|\beta, \phi, \vartheta) = \left(\prod_{i=1}^{n} f(y_i|\beta, \phi, \boldsymbol{b}, \vartheta) \right) f(\boldsymbol{b}|\vartheta)$$

bzw. marginale Likelihood:

$$L(\beta, \phi, \vartheta) = f(\boldsymbol{y}|\beta, \phi, \vartheta) = \int \left(\prod_{i=1}^{n} f(y_i|\beta, \phi, \boldsymbol{b}, \vartheta) \right) f(\boldsymbol{b}|\vartheta) d\boldsymbol{b}.$$

Erinnerung: im LMM ist das Integral analytisch lösbar.
$\int f(\boldsymbol{y}, \boldsymbol{b}|\beta, \phi, \vartheta) d\boldsymbol{b}$ ist die Dichte einer $N(\boldsymbol{X}\beta, \boldsymbol{Z}\boldsymbol{G}(\vartheta)\boldsymbol{Z}' + \boldsymbol{R}(\phi, \vartheta))$-Verteilung.

Im GLMM: ??

ML-Inferenz

Interessante Parameter: primär $\boldsymbol{\beta}, \boldsymbol{\vartheta}$, evtl. ϕ
$\Rightarrow$ Problem: finde argmax $L(\boldsymbol{\beta}, \boldsymbol{\vartheta}, \phi)$, wobei

$$L(\boldsymbol{\beta}, \boldsymbol{\vartheta}, \phi) = \int \left(\prod_{i=1}^{n} f(y_i | \boldsymbol{\beta}, \phi, \boldsymbol{b}, \boldsymbol{\vartheta}) \right) f(\boldsymbol{b} | \boldsymbol{\vartheta}) d\boldsymbol{b}$$

$$= \int \left(\prod_{i=1}^{n} \exp\left(\frac{y_i \theta_i - b(\theta_i)}{\phi} - c(y_i, \phi) \right) \right) |\boldsymbol{G}(\boldsymbol{\vartheta})|^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{b}' \boldsymbol{G}(\boldsymbol{\vartheta})^{-1} \boldsymbol{b} \right) d\boldsymbol{b}$$

mit $\theta_i = (b')^{-1}(g^{-1}(\eta_i))$, $\eta_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \boldsymbol{z}_i' \boldsymbol{b}$.

Hoch-dimensionales Integral, i.A. nicht analytisch lösbar
$\Rightarrow$ iterative Optimierung einer Approximation der Likelihood

Laplace-Approximation

Problem: Löse r-dimensionales Integral $H = \int \exp(h(\boldsymbol{\theta}))d\boldsymbol{\theta}$

Ansatz:

- bestimme $\boldsymbol{\theta}_0 = \arg\max h(\boldsymbol{\theta})$
- quadratische Taylor-Entwicklung von $h(\boldsymbol{\theta})$ um $\boldsymbol{\theta}_0$:

$$h(\boldsymbol{\theta}) \approx h(\boldsymbol{\theta}_0) + \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \underbrace{\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} h(\boldsymbol{\theta}_0)\right)}_{=-\boldsymbol{P}} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

- $\Rightarrow \int \exp(h(\boldsymbol{\theta}))d\boldsymbol{\theta} \approx \int \exp(h(\boldsymbol{\theta}_0) \underbrace{-\tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\boldsymbol{P}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)}_{\text{wie bei } N(\boldsymbol{\theta_0}, \boldsymbol{P^{-1}})})d\boldsymbol{\theta}$
- $\Rightarrow H \approx \exp\left(h(\boldsymbol{\theta}_0)\right) \underbrace{\left((2\pi)^{r/2}|\boldsymbol{P}|^{-1/2}\right)}_{\text{1/Normierung der } N(\boldsymbol{\theta_0}, \boldsymbol{P^{-1}})}$

GLMM-Likelihood: Laplace-Approximation

Verwende eine Laplace-Approximation, mit Entwickelung um den Maximierer $\hat{\boldsymbol{b}}$ von $L(\boldsymbol{\beta}, \phi, \boldsymbol{\vartheta}, \boldsymbol{b}) = f(\boldsymbol{y}|\boldsymbol{\beta}, \phi, \boldsymbol{b}) f(\boldsymbol{b}|\boldsymbol{\vartheta})$,

$$
\begin{aligned}
\log(L(\boldsymbol{\beta}, \phi, \boldsymbol{\vartheta})) &= \log\left(\int f(\boldsymbol{y}|\boldsymbol{\beta}, \phi, \boldsymbol{b}) f(\boldsymbol{b}|\boldsymbol{\vartheta}) d\boldsymbol{b}\right) \\
&= \log\left(\int L(\boldsymbol{\beta}, \phi, \boldsymbol{b}) (2\pi)^{-r/2} |\boldsymbol{G}(\boldsymbol{\vartheta})|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} \boldsymbol{b}' \boldsymbol{G}(\boldsymbol{\vartheta})^{-1} \boldsymbol{b}\right) d\boldsymbol{b}\right) \\
&\approx \log(L(\boldsymbol{\beta}, \hat{\boldsymbol{b}}, \phi)) - \tfrac{r}{2} \log(2\pi) - \tfrac{1}{2} \log|\boldsymbol{G}(\boldsymbol{\vartheta})| - \tfrac{1}{2} \hat{\boldsymbol{b}}' \boldsymbol{G}(\boldsymbol{\vartheta})^{-1} \hat{\boldsymbol{b}} \\
&\quad + \log\left(\int \exp\left(-\tfrac{1}{2} (\boldsymbol{b} - \hat{\boldsymbol{b}})' \mathcal{I}(\hat{\boldsymbol{b}}) (\boldsymbol{b} - \hat{\boldsymbol{b}})\right) d\boldsymbol{b}\right) \\
&= l(\boldsymbol{\beta}, \hat{\boldsymbol{b}}, \phi) - \tfrac{1}{2} \log|\mathcal{I}(\hat{\boldsymbol{b}})| - \tfrac{1}{2} \log|\boldsymbol{G}(\boldsymbol{\vartheta})| - \tfrac{1}{2} \hat{\boldsymbol{b}}' \boldsymbol{G}(\boldsymbol{\vartheta})^{-1} \hat{\boldsymbol{b}}.
\end{aligned}
$$

Dabei ist $\mathcal{I}(\boldsymbol{b}) = -\,\mathrm{E}\left(\frac{\partial^2}{\partial \boldsymbol{b} \partial \boldsymbol{b}'} l(\boldsymbol{\beta}, \phi, \boldsymbol{\vartheta}, \boldsymbol{b})\right)$ (mit zusätzlichem Erwartungswert), Herleitung siehe S. 166.

Schaukel-Algorithmus

In der Laplace-Approximation wird der Maximierer $\hat{\boldsymbol{b}}$ von $L(\boldsymbol{\beta}, \phi, \boldsymbol{\vartheta}, \boldsymbol{b})$ benötigt.
$\Rightarrow$ iterativer, zweistufiger Schaukel-Algorithmus:

1. Für gegebene $\boldsymbol{\beta}, \phi, \boldsymbol{\vartheta}$ bestimme $\hat{\boldsymbol{b}} = \arg\max_b L(\boldsymbol{\beta}, \phi, \boldsymbol{\vartheta}, \boldsymbol{b})$ über penalisierten IRLS-Algorithmus (P-IRLS).
2. Maximiere Laplace-Approximation $\tilde{L}(\boldsymbol{\beta}, \phi, \boldsymbol{\vartheta})$ von $L(\boldsymbol{\beta}, \phi, \boldsymbol{\vartheta})$ in $\hat{\boldsymbol{b}}$ mit numerischer Optimierung (Pseudo-Newton-Algorithmen wie BFGS).

Iteriere bis zur Konvergenz der Devianz $-2 \log L(\boldsymbol{\beta}, \phi, \boldsymbol{\vartheta}, \boldsymbol{b})$.

Dieser Schaukel-Algorithmus ist im R-Paket `lme4` implementiert.

Grundidee: IRLS-Algorithmus

- IRLS = Fisher-Scoring für GLM
  - IRLS: **I**teratively **R**e-Weighted **L**east **S**quares
  - Führt GLM-Schätzproblem auf iterierte gewichtete KQ-Schätzung zurück.
- Fisher-Scoring zur Lösung des Score-Gleichungssystems $s(\boldsymbol{\theta}) \stackrel{!}{=} \mathbf{0}$
  - lineare Taylor-Entwicklung $s(\boldsymbol{\theta}) \approx s(\boldsymbol{\theta}_0) - \mathcal{J}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \stackrel{!}{=} \mathbf{0}$ mit der beobachteten Informationsmatrix $\mathcal{J}(\boldsymbol{\theta})$.
  - Ersetzen von $\mathcal{J}(\boldsymbol{\theta})$ (Newton-Raphson) durch die erwartete Informationsmatrix $\mathcal{I}(\boldsymbol{\theta})$ (Fisher-Scoring) liefert

$$\mathcal{I}(\boldsymbol{\theta}_0)\boldsymbol{\theta} = \mathcal{I}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0 + s(\boldsymbol{\theta}_0).$$

$s(b)$ und $\mathcal{I}(b)$ für den kanonischen Link

Für den kanonischen Link ist $\boldsymbol{\theta} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b}$.

$$\begin{aligned}
\Rightarrow s(\boldsymbol{b}) &= \frac{\partial}{\partial \boldsymbol{b}} l(\beta, \phi, \vartheta, \boldsymbol{b}) = \frac{\partial}{\partial \boldsymbol{b}} \left(const + \frac{\boldsymbol{\theta}'\boldsymbol{y} - b(\boldsymbol{\theta})'\mathbf{1}}{\phi} - \frac{1}{2}\boldsymbol{b}'\boldsymbol{G}(\vartheta)^{-1}\boldsymbol{b} \right) \\
&= \frac{1}{\phi}(\boldsymbol{Z}'\boldsymbol{y} - \boldsymbol{Z}' \operatorname{diag}(b'(\boldsymbol{\theta}))\mathbf{1}) - \boldsymbol{G}(\vartheta)^{-1}\boldsymbol{b} \\
&= \frac{1}{\phi}\boldsymbol{Z}'(\boldsymbol{y} - \mu) - \boldsymbol{G}(\vartheta)^{-1}\boldsymbol{b} \quad \text{und}
\end{aligned}$$

$$\begin{aligned}
\mathcal{I}(\boldsymbol{b}) &= -\operatorname{E}\left(\frac{\partial^2}{\partial \boldsymbol{b} \partial \boldsymbol{b}'} l(\beta, \phi, \vartheta, \boldsymbol{b}) \right) = -\left(\frac{\partial}{\partial \boldsymbol{b}'} s(\boldsymbol{b}) \right) \\
&= \frac{1}{\phi}\boldsymbol{Z} \operatorname{diag}(b''(\boldsymbol{\theta}))\boldsymbol{Z}' + \boldsymbol{G}(\vartheta)^{-1} =: \boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}' + \boldsymbol{G}(\vartheta)^{-1}
\end{aligned}$$

P-IRLS

Fisher-Scoring ausgehend von Startwert $\boldsymbol{b}_0$:

$$\mathcal{I}(\boldsymbol{b}_0)\boldsymbol{b} = \mathcal{I}(\boldsymbol{b}_0)\boldsymbol{b}_0 + s(\boldsymbol{b}_0)$$

mit im GLMM: $$s(\boldsymbol{b}) = \frac{1}{\phi}\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{G}(\boldsymbol{\vartheta})^{-1}\boldsymbol{b};$$

$$\mathcal{I}(\boldsymbol{b}) = \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}(\boldsymbol{\vartheta})^{-1}$$

liefert mit $\boldsymbol{W}_0 = \boldsymbol{W}(\boldsymbol{b}_0)$, $\boldsymbol{\mu}_0 = \boldsymbol{\mu}(\boldsymbol{b}_0)$:

$$(\boldsymbol{Z}'\boldsymbol{W}_0\boldsymbol{Z} + \boldsymbol{G}(\boldsymbol{\vartheta})^{-1})\boldsymbol{b} = (\boldsymbol{Z}'\boldsymbol{W}_0\boldsymbol{Z} + \boldsymbol{G}(\boldsymbol{\vartheta})^{-1})\boldsymbol{b}_0 + \frac{1}{\phi}\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{\mu}_0) - \boldsymbol{G}(\boldsymbol{\vartheta})^{-1}\boldsymbol{b}_0$$

$$\Leftrightarrow (\boldsymbol{Z}'\boldsymbol{W}_0\boldsymbol{Z} + \boldsymbol{G}(\boldsymbol{\vartheta})^{-1})\boldsymbol{b} = \boldsymbol{Z}'\boldsymbol{W}_0 \underbrace{(\boldsymbol{Z}\boldsymbol{b}_0 + \frac{1}{\phi}\boldsymbol{W}_0^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_0))}_{\text{working response } \tilde{\boldsymbol{y}}}$$

$\Rightarrow$ Schätzgleichung eines LMM mit bekanntem $\boldsymbol{W}_0, \boldsymbol{G}(\boldsymbol{\vartheta})$, vgl. (14):

$$\tilde{\boldsymbol{y}}|\boldsymbol{b} \sim N(\boldsymbol{Z}\boldsymbol{b}, \boldsymbol{W}_0^{-1});\ \boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G}(\boldsymbol{\vartheta}))$$

P-IRLS

P-IRLS-Algorithmus iteriert folgende Schritte bis zur Konvergenz von $\widehat{\boldsymbol{b}}$:

i. setze Iterationswert $\boldsymbol{b}_0 = \hat{\boldsymbol{b}}^{(k)}$, berechne mit $\boldsymbol{\beta}$, $\boldsymbol{b}_0$ die working responses $\tilde{\boldsymbol{y}}$ und Gewichte $\boldsymbol{W}_0$

ii. berechne $\hat{\boldsymbol{b}}^{(k+1)}$ als Lösung des daraus abgeleiteten gewichteten, penalisierten KQ-Problems

(Adaptive) Gauss'sche Quadratur

- Laplace-Approximation recht schnell, aber ungenau besonders für kleine Clustergrößen und starke „Diskretheit" (logistische Regression schlechter als Poisson-Regression).
- Gauss-Quadratur genauere Methode, um $\int f(\boldsymbol{y}|\boldsymbol{\beta}, \phi, \boldsymbol{b}) f(\boldsymbol{b}|\boldsymbol{\vartheta}) d\boldsymbol{b}$ zu approximieren, aber wesentlich rechenaufwändiger.
- Benutzt orthonormalisierte zufällige Effekte $\boldsymbol{b}^\star = \boldsymbol{G}(\boldsymbol{\vartheta})^{-1/2}\boldsymbol{b}$.
- $\int f(\boldsymbol{y}|\boldsymbol{\beta}, \phi, \boldsymbol{\vartheta}, \boldsymbol{b}^\star) f(\boldsymbol{b}^\star) d\boldsymbol{b}^\star \approx \sum_{k=1}^{Q} w_k f(\boldsymbol{y}|\boldsymbol{\beta}, \phi, \boldsymbol{\vartheta}, \boldsymbol{b}_k^\star)$
- Die Stützstellen $\boldsymbol{b}_k^\star$ ergeben sich aus den Nullstellen des Q-ten Hermite-Polynoms und bestimmen auch die Gewichte w_k.

(Adaptive) Gauss'sche Quadratur

- Genauigkeit der Aproximation steigt mit wachsendem Q
  $\Rightarrow$ erhöhe Q solange bis keine Änderung in Schätzung mehr zu beobachten
- in `lme4` nur implementiert für Modelle mit *einer* einzigen Gruppierungsvariable (option: `nAGQ`)
- Laplace-Approximation ergibt sich als Spezialfall $Q = 1$.

Penalized Quasi-Likelihood (PQL):

Ähnliche Idee wie P-IRLS: Approximiere $\boldsymbol{y}$ durch $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{y}|\boldsymbol{b})$ plus Fehler mit Varianz $\mathrm{Var}(\boldsymbol{y}|\boldsymbol{b})$. Eine Taylor-Approximation von $\boldsymbol{\mu}$ um die aktuellen Werte β_0 und $\boldsymbol{b}_0$ und Umsortieren ergibt

$$\tilde{\boldsymbol{y}} := \boldsymbol{X}\beta_0 + \boldsymbol{Z}\boldsymbol{b}_0 + \frac{1}{\phi}\boldsymbol{W}_0^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_0) = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \varepsilon.$$

Algorithmus:

1. Für gegebene Werte $\beta_0 = \hat{\beta}^{(k)}$, $\boldsymbol{b}_0 = \hat{\boldsymbol{b}}^{(k)}$ bestimme *working responses* $\tilde{\boldsymbol{y}}$.
2. Bestimme $\hat{\beta}^{(k+1)}, \hat{\boldsymbol{b}}^{(k+1)}$ und $\hat{\vartheta}^{(k+1)}, \hat{\phi}^{(k+1)}$ aus dem LMM $\tilde{\boldsymbol{y}} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \varepsilon$ mit $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G}(\vartheta))$ und $\varepsilon \sim N(\boldsymbol{0}, \boldsymbol{W}_0^{-1})$

Iteriere bis zur Konvergenz.

PQL

- Alternative Herleitung für PQL: ergibt sich ebenfalls bei Fisher-Scoring für die gemeinsame/penalisierte log-Likelihood (die im LMM die BLUPs ergibt). Daher auch mögliche Schätzart bei GAMMs.
- in R implementiert in Funktion `glmmPQL` im Paket `MASS` (`glmmPQL` wird auch von Funktion `gamm` im Paket `mgcv` benutzt)
- benutzt iterierte ML- oder REML-Schätzung
- Schätzung der Varianzparameter nach unten verzerrt, v.a. falls n_i klein. Bessere Approximation, je näher Daten an Normalverteilung (z.B. Poisson-Verteilung mit größeren λ's)
- Konvergenz nicht garantiert
- Kein AIC, BIC oder Devianz berechenbar

Vergleich glmmPQL/lme vs. (g)lmer

	`MASS::glmmPQL / nlme::lme`	`lme4::(g)lmer`
Daten	nur genestete Daten; grosse Datensätze/Modelle oft nicht zu fitten	genestete & gekreuzte Datenstrukturen; auch riesige Datensätze werden gefittet
(G)AMs	via `mgcv::gamm`	via `gamm4`
Cov($\boldsymbol{b}_i$)	sehr breite, erweiterbare Klasse von Kovarianzstrukturen (s. `nlme::pdMat`)	nur unstrukturierte oder diagonale Kovarianzen (in `gamm4` beliebige Präzisionsmatrizen $\boldsymbol{P}$ mit $\boldsymbol{G}(\vartheta) = \vartheta\boldsymbol{P}^{-1}$)
Cov(ε)	sehr breite, erweiterbare Klasse von Kovarianzstrukturen (s. `nlme::varFunc`)	Cov(ε) $= \sigma^2\boldsymbol{I}_n$ oder $\sigma^2\,\mathrm{diag}(\boldsymbol{w}^{-1})$ mit bekannten Gewichten $\boldsymbol{w}$
Stabilität	sehr instabil für komplexe Strukturen von $\boldsymbol{b}$	sehr stabil für LMMs; für GLMMs ab > 3 zuf. Effekten oft kritisch
Speed	relativ langsam	LA sehr schnell (benutzt Sparse-Matrix-Algorithmen); AGQ deutlich langsamer

Vorhersage der zufälligen Effekte

Die beste Vorhersage für $\boldsymbol{b}$ (minimaler mean squared error of prediction $\mathrm{E}(\boldsymbol{b} - \widehat{\boldsymbol{b}})^2$) wäre wieder

$$\widehat{\boldsymbol{b}} = \mathrm{E}(\boldsymbol{b}|\boldsymbol{y}).$$

In der Praxis würde man $\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\vartheta}}$ einsetzen. Allerdings erfordert

$$\widehat{\boldsymbol{b}} = \int \boldsymbol{b} f(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\vartheta}) d\boldsymbol{b} = \frac{\int \boldsymbol{b} f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \phi) f(\boldsymbol{b}|\boldsymbol{\vartheta}) d\boldsymbol{b}}{\int f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \phi) f(\boldsymbol{b}|\boldsymbol{\vartheta}) d\boldsymbol{b}}$$

wieder numerische Integration.

Alternativ zum Posteriori-EW wird häufig (z.B. in `lme4` als Teil von P-IRLS) der Posteriori-Modus berechnet, der $f(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\vartheta}) \propto f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \phi) f(\boldsymbol{b}|\boldsymbol{\vartheta})$ maximiert. Die beiden unterscheiden sich i.A. außer im LMM unter NV.

PQL schätzt $\boldsymbol{b}$ direkt im Schätzalgorithmus.

Hypothesen-Tests für β

- Wegen der Maximum-Likelihood-Schätzung von β können prinzipiell Wald-, Likelihood-Quotienten- (LQT) oder Score-Tests verwendet werden mit einer entsprechenden χ^2-Verteilung als Referenzverteilung.

- Die Güte der Approximation hängt ab von
  - der Güte der Approximation an die Likelihood in der Schätzung
  - der Güte der asymptotischen Approximation, die nur für Longitudinal/Clusterdaten bei $N \to \infty$ greift.

- PQL-Schätzung beruht auf der Likelihood von Pseudodaten und daher ist kein Likelihood-Quotiententest möglich. Inferenz für PQL wird meist basiert auf dem LMM in der PQL-Schätzung. Allerdings ist $\widehat{\beta}$ i.A. nicht konsistent.

Inferenz für D

Hypothesen-Tests

- Für das longitudinale/Cluster-GLMM gelten die gleichen asymptotischen Ergebnisse für Tests von Parametern in $\boldsymbol{D}$ wie im LMM (Rand des Parameterraums).
- Es ist keine exakte Verteilung vorhanden.

Modellselektion

- Ein konditionales AIC für Poisson-, binomial- oder normalverteilte Zielgrößen ist im R-Paket `cAIC4` implementiert, siehe auch
  Saefken, Kneib, van Waveren & Greven (2014). A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models. Electronic Journal Statistics, 8, 201-225.

Konfidenzintervalle

- Generell gilt für einen Parameter θ:
  Ein Wert θ_0 ist im $(1-\alpha)$%-Konfidenzintervall für θ $\Leftrightarrow$ Die Hypothese $H_0 : \theta = \theta_0$ gegen $H_A : \theta \neq \theta_0$ wird zum Level α nicht verworfen.
  (Sofern Test und Konfidenzintervall auf der gleichen Statistik beruhen.)

- Im GLMM: Für Parameter θ in $\boldsymbol{\beta}$ oder $\boldsymbol{D}$, betrachte ein Gitter von θ_0-Werten um den Schätzer $\hat{\theta}$.

- Konstruiere das Konfidenzintervall für θ so, dass alle Werte auf dem Gitter enthalten sind, für die ein LQT mit Referenzverteilung χ^2_1 H_0 nicht ablehnt.

- Dieser Ansatz funktioniert auch bei nicht-symmetrischen Verteilungen von Schätzern, jedoch nicht bei Parametern θ_0 in $\boldsymbol{D}$ nahe des Randes des Parameterraums.

- Der Ansatz ist in `lme4` implementiert in der Funktion `confint`. Alternativen in `confint` sind Wald- (für $\boldsymbol{\beta}$) und Bootstrap-basierte Konfidenzintervalle.