4.9 Hierarchische Modelle

Dieser Abschnitt behandelt ein Beispiel aus Gelman, Carlin, Stern und Rubin (2003) zum Tumorrisiko bei Ratten. Durchgeführt wurden 71 Experimente i mit jeweils n_i Ratten j:

$$y_{ij} = \begin{cases} 1, & \text{Ratte entwickelt Tumor,} \\ 0, & \text{Ratte entwickelt keinen Tumor.} \end{cases}$$

Also ist $y_i = \sum_{j=1}^{n_i} y_{ij}$ die Anzahl an Ratten in Experiment i, die einen Tumor entwickeln.

4.9 Hierarchische Modelle

Ideen:

1. Experiment-spezifische Wahrscheinlichkeiten $\theta_i,\ i = 1, \ldots, n$, für Tumorentwicklung betrachten, potentiell zurückzuführen auf Heterogenität der Ratten, unterschiedliche experimentelle Bedingungen usw.
2. Alle θ_i stammen aus einer Population, zum Beispiel einer Beta(α, β)-Verteilung.
3. Anstatt 71 Parameter $\theta_1, \ldots, \theta_{71}$ direkt aus den Daten zu schätzen, nehmen wir eine Verteilung für die θ_i an.

4.9 Hierarchische Modelle

Ideen:

4. Ohne weitere Information sind $\theta_1, \ldots, \theta_{71}$ als exchangeable zu betrachten, d.h. für die gemeinsame Priori von $\theta = (\theta_1, \ldots, \theta_{71})$ gilt

$$p(\theta|\phi) = \prod_{i=1}^{71} p(\theta_i|\phi),$$

wobei ϕ die *Hyperparameter* bezeichnet. „Averaging" über ϕ liefert die marginale Priori

$$p(\theta) = \int \left(\prod_{i=1}^{n} p(\theta_i|\phi) \right) p(\phi)\, d\phi.$$

Da ϕ nicht bekannt ist, erhält es eine eigene (Hyper-) Prioriverteilung $p(\phi)$, im Beispiel eine Priori für α und β.

4.9 Hierarchische Modelle

Struktur des hierarchischen Modells:

$$\phi = (\alpha, \beta) \sim p(\alpha, \beta)$$

θ_1	θ_2	$\ldots$	θ_{70}	θ_{71}
$\downarrow$	$\downarrow$		$\downarrow$	$\downarrow$
$y_1 \vert \theta_1, n_1$	$y_2 \vert \theta_2, n_2$	$\ldots$	$y_{70} \vert \theta_{70}, n_{70}$	$y_{71} \vert \theta_{71}, n_{71}$
$0/20$	$0/20$	$\ldots$	$9/24$	$4/14$
		$\ldots$		

Hierarchisches Modell des Beispiels in der top-down Schreibweise:

$$\text{tauchen in der Likelihood auf} \quad \begin{cases} y_{ij} \vert n_i, \theta_i & \sim & \text{Binomial}(n_i, \theta_i) \\ \theta_i \vert \alpha, \beta & \sim & \text{Beta}(\alpha, \beta) \end{cases}$$

$$\text{taucht nicht in der Likelihood auf} \quad \{ \; (\alpha, \beta) \sim p(\alpha, \beta)$$

Die **Posteriori** für alle Parameter lautet

$$f(\theta, \alpha, \beta | y_1, \ldots, y_{71}) \propto$$
$$\left(\prod_{i=1}^{71} \theta_i^{y_i}(1-\theta_i)^{n_i - y_i}\right)\left(\prod_{i=1}^{71} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}\right) \cdot p(\alpha, \beta).$$

Für feste α, β ist die Posteriori von $(\theta_1, \ldots, \theta_{71})$ das Produkt unabhängiger Posterioris, die jeweils einer Beta$(\widetilde{\alpha}_i, \widetilde{\beta}_i)$-Verteilung folgen mit $\widetilde{\alpha}_i = \alpha + y_i$, $\widetilde{\beta}_i = \beta + (n_i - y_i)$. Genauer:

$$f(\theta | \alpha, \beta, \{y_{ij}\}) = \prod_{i=1}^{71} \frac{\Gamma(\alpha+\beta+n_i)}{\Gamma(\alpha+y_i)\Gamma(\beta+n_i-y_i)} \cdot \theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1}.$$

4.9 Hierarchische Modelle

Die **marginale Posteriori** $f(\theta|\alpha,\beta,\{y_{ij}\})$ von (α,β) ergibt sich über die bereits vielfach verwandte Formel

$$f(\alpha,\beta|\{y_{ij}\}) = \frac{f(\theta,\alpha,\beta|\{y_{ij}\})}{f(\theta|\alpha,\beta,\{y_{ij}\})}.$$

Im Beispiel erhält man

$$f(\alpha,\beta|\{y_{ij}\}) \propto p(\alpha,\beta) \cdot \prod_{i=1}^{71} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_i)\Gamma(\beta+n_i-y_i)}{\Gamma(\alpha+\beta+n_i)}.$$

„Knackpunkt" ist die **nicht-triviale Wahl der Hyperpriori** $p(\alpha, \beta)$: Oft sind keine Anhaltspunkte für die Verteilung der Hyperparameter vorhanden. Die Hyperpriori sollte in diesem Fall möglichst wenig informativ sein, allerdings gleichzeitig so, dass die Posteriori proper ist.

4.9 Hierarchische Modelle

Für das Beispiel schlagen Gelman, Carlin, Stern und Rubin (2003) vor, die Hyperparameter in den (kompletten) $\mathbb{R}^2$ zu transformieren, zum Beispiel durch

$$(\alpha, \beta) \mapsto \left[\text{logit} \left(\frac{\alpha}{\alpha + \beta} \right), \log(\alpha + \beta) \right],$$

wobei hier $\text{logit}(\alpha/(\alpha + \beta)) = \log(\alpha/\beta)$.

Zur Interpretation:

- $\alpha/(\alpha + \beta)$ ist (Priori-) Erwartungswert der Betaverteilung
- $\alpha + \beta$ lässt sich als Priori-Stichprobengröße auffassen
- Aber: Gleichverteilung auf der transformierten Skala führt zu uneigentlicher Posteriori

4.9 Hierarchische Modelle

Alternativ schlagen obige Autoren folgende Priori vor:

$$p\left(\log\left(\frac{\alpha}{\beta}\right), \log(\alpha+\beta)\right) \propto \alpha\beta(\alpha+\beta)^{-5/2}. \tag{1}$$

Diese ergibt sich aus einer Gleichverteilung für eine Approximation der Priori-Standardabweichung $(\alpha+\beta)^{-1/2}$, die unabhängig mit einer Gleichverteilung auf dem Priori-Erwartungswert kombiniert wird, d.h.

$$p\left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-\frac{1}{2}}\right) \propto 1.$$

4.9 Hierarchische Modelle

Die Simulation der Posterioriverteilung $f(\theta|\alpha, \beta, \{y_{ij}\})$ ist dann wie folgt:

1. Ziehe Zufallszahlen aus $f(\log(\alpha/\beta), \log(\alpha+\beta)|\{y_{ij}\})$: Dazu wird $f(\alpha, \beta|\{y_{ij}\})$ gemäß Dichtetransformationssatz transformiert (unter anderem wird also $p(\alpha, \beta)$ durch die Hyperpriori (1) ersetzt) und dann auf einem feinen Gitter berechnet. $(\log(\alpha/\beta), \log(\alpha+\beta))|\{y_{ij}\}$ kann nun unter Verwendung des CDF-Samplers (Algorithmus 5 auf Seite 52) simuliert werden.
2. Transformiere die in Schritt 1 gezogenen Zufallszahlen auf die ursprüngliche (α, β)-Skala zurück.
3. Ziehe dann für $i = 1, \ldots, 71 \quad \theta_i|\alpha, \beta, \{y_{ij}\}$ gemäß einer Beta$(\alpha + y_i, \beta + n - y_i)$-Verteilung.

4.10 Konvergenzdiagnostik

Schwierigkeiten bei statistischer Inferenz entstehen durch iterative Simulation:

1. Simulierte Zufallszahlen aus der Posteriori **repräsentieren die Zielverteilung eventuell unzureichend** (Problem der Startwertewahl, Einfluss der Startwerte auf spätere Ziehungen).
2. MCMC: **Zufallszahlen sind korreliert** — die Inferenz ist ungenauer, als wenn die gleiche Anzahl unabhängiger Zufallszahlen verwandt würde. Stichwort: „effektive Sitchprobengröße".

4.10 Konvergenzdiagnostik

Strategien:

1. Multiple Sequenzen, die stark über den Parameterraum streuen.
2. Konvergenz-Monitoring.
3. Falls das „Mixing" schlecht ist (der Parameterraum wird unzureichend exploriert, vorwiegend Bewegung entlang weniger lokaler Maxima der Zielverteilung usw.), sollte der Algorithmus geändert werden.

4.10 Konvergenzdiagnostik

Bezüglich Punkt 2 schlagen Gelman, Carlin, Stern und Rubin (2003) für skalaren Parameter vor:

- Generiere m parallele Sequenzen der Länge $n \geq 2$ nach Entfernen der Burnin-Werte, also

$$\{\psi_{ij}\} \text{ für } i = 1, \ldots, n,\ j = 1, \ldots, m.$$

4.10 Konvergenzdiagnostik

- Berechne die Varianz zwischen den Sequenzen und innerhalb jeder Sequenz: Setze

$$\overline{\psi}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij} \quad , \quad \overline{\psi}_{\cdot\cdot} = \frac{1}{m}\sum_{j=1}^{m}\overline{\psi}_{\cdot j}$$

und definiere

$$B = n \cdot \frac{1}{m-1}\sum_{j=1}^{m}(\overline{\psi}_{\cdot j} - \overline{\psi}_{\cdot\cdot})^2$$

sowie

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2 \quad \text{mit} \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\psi_{ij} - \overline{\psi}_{\cdot j})^2.$$

4.10 Konvergenzdiagnostik

Betrachte dann folgende Schätzung für die marginale Posteriori-Varianz:

$$\widehat{\mathrm{Var}}^{+}(\psi|y) = \frac{n-1}{n} W + \frac{1}{n} B.$$

Die Varianz für ψ wird in der Regel überschätzt. Die Schätzung ist jedoch für
$n \to \infty$ unverzerrt. W allein unterschätzt in der Regel die Varianz, aber

$$\lim_{n \to \infty} \mathbb{E}[W] = \mathrm{Var}(\psi|y).$$

4.10 Konvergenzdiagnostik

- Führe nun ein Monitoring für die Größe

$$\widehat{R} = \sqrt{\frac{\widehat{\mathrm{Var}}^{+}(\psi|y)}{W}}$$

mit $R \to 1$ für $n \to \infty$ durch, d.h. wenn $\widehat{R}$ groß ist, erhält man potentiell eine Verbesserung der Inferenz, wenn weitere Simulationen durchgeführt werden.

4.10 Konvergenzdiagnostik

- Also:
  (a) Wenn $\widehat{R}$ groß: Lasse Simulation weiter laufen.
  (b) Wenn $\widehat{R}$ „nahe" 1: Verwende alle $m \cdot n$ Werte für Posteriori-Inferenz. Die optimale Größe von $\widehat{R}$ ist wiederum ein Tuning-Parameter, zum Beispiel $\widehat{R} < 1.1$.

  Erfolg ist allerdings nicht garantiert. Die Methode funktioniert gut bei approximativ normaler marginaler Posteriori, ist jedoch weniger geeignet, wenn Interesse an den extremen Quantilen besteht.
- Effektive Anzahl unabhängiger Ziehungen:

$$n_{\text{eff}} = m \cdot n \cdot \frac{\widehat{\text{Var}}^{+}(\psi)}{W}.$$

  Dabei sollte m nicht zu klein sein, da sonst B wiederum schlecht geschätzt wird.

Obige Methode ist im `R`-Paket `coda` (convergence diagnostics and output analysis) implementiert.

4.11 Modellwahl und Modellkritik

Modellwahl ist ein weites Feld, auch in Likelihood-basierter Inferenz.

Generelle Strategien sind schwierig zu finden, da sie in der Regel abhängen von

- Kontext (Substanzwissenschaft)
- Randbedingungen (zum Beispiel randomisierte Studie oder Beobachtungsstudie)
- *signal-to-noise-ratio* (kontrolliertes experimentelles Umfeld oder starke Heterogenität)

4.11 Modellwahl und Modellkritik

Sensitivitätsanalyse:

- Einfluss der Priori auf Ergebnisse.
- Einfluss des Modells (Likelihood) auf Güte der Vorhersagen. Welche Fälle werden schlecht durch das Modell beschrieben?

Modellwahl:

Typischerweise, zum Beispiel bei einem Regressionsmodell, erfolgt eine Auswahl der Kovariablen; „nested versus non-nested" Situation.

4.11 Modellwahl und Modellkritik

Ein populärer Vorschlag zur Modellwahl ist das **DIC (Deviance Information Criterion)**.

Wie das AIC und BIC ist das DIC eine asymptotische Approximation und nur anwendbar, falls die Posteriori approximativ multivariat normal ist.

Allgemein ist die Devianz für Daten $\boldsymbol{y}$ und Parameter(-vektor) $\boldsymbol{\theta}$ definiert als

$$D(\boldsymbol{y}, \boldsymbol{\theta}) = -2 \log f(\boldsymbol{y}|\boldsymbol{\theta}) + C(\boldsymbol{y}).$$

Beurteilt man Modelle nach der Devianz, so ist ein Modell umso besser, je kleiner diese ist.

4.11 Modellwahl und Modellkritik

Das DIC kann aus den generierten Samples der MCMC-Simulation berechnet werden. Sei $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ eine generierte Sequenz. Die **erwartete Devianz** bezüglich der Posterioriverteilung von $\boldsymbol{\theta}$ ist

$$\mathbb{E}[D(\boldsymbol{y}, \boldsymbol{\theta})|\boldsymbol{y}]$$

und wird durch

$$\overline{D} = \frac{1}{L} \sum_{l=1}^{L} D(\boldsymbol{y}, \boldsymbol{\theta}_l)$$

geschätzt.

Dies ist ein Maß dafür, wie gut das Modell an die Daten angepasst ist (je kleiner, umso besser).

4.11 Modellwahl und Modellkritik

Die **effektive Anzahl der Parameter** in einem bayesianischen Modell ist

$$p_D = \overline{D} - D(\overline{\boldsymbol{\theta}}),$$

wobei

$$\overline{\boldsymbol{\theta}} = \frac{1}{L} \sum_{l=1}^{L} \boldsymbol{\theta}_l$$

die Schätzung des Posteriori-Erwartungswertes von $\boldsymbol{\theta}$ und $D(\overline{\boldsymbol{\theta}})$ die Devianz ausgewertet an $\overline{\boldsymbol{\theta}}$ ist.

4.11 Modellwahl und Modellkritik

Für ein **lineares Modell** mit Normalverteilungsannahme entspricht p_D der **Anzahl unrestringierter** Parameter im Modell. Das DIC ist dann definiert als

$$\text{DIC} = 2\overline{D} - D(\overline{\boldsymbol{\theta}}) = \overline{D} + p_D.$$

Hinweis:

$$\begin{aligned} \overline{D}^{\text{prediction}} &= \mathbb{E}[D(\boldsymbol{y}^{\text{rep}}, \overline{\boldsymbol{\theta}})] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i^{\text{rep}} - \mathbb{E}[y_i^{\text{rep}}|\boldsymbol{y}])^2\right], \end{aligned}$$

wobei der Erwartungswert bezüglich der a posteriori prädiktiven Verteilung zu verstehen ist. Es ist $\overline{D}^{\text{prediction}} \approx \text{DIC}$.