Einführung in die Bayes Statistik- Slides Begleitung zum Online Kurs

Bayes-Faktor und MCMC

Martje Rave

Sommersemester 2025

ACHTUNG!

Diese Folien sind **NUR** Kurs begleitend und können noch viele Fehler beinhalten. Für die Klausurvorbereitung benutzt bitte den Online-Kurs und die Übungsblätter.
Bitte gebt mir Bescheid, wenn ihr Fehler findet.

Modellwahl

Wir haben verschiedene Modelle kennengelernt. Wenn wir uns bezüglich des Modells nicht sicher sind: wie können wir die Modelle dann vergleichen, welches Modell ist das Beste? Natürlich wollen wir die Schlüsse wieder aus der Posteriori ziehen. Idee: Das Modell ist eigentlich nur ein weiterer uns unbekannter Parameter.

Allgemein

Gegeben:

- Daten x
- K verschiedene Modelle
- Wahrscheinlichkeit, x unter Modell M_k zu beobachten: $p(x \mid M_k)$
- Prior auf Modelle: $p(M_k)$

Posterior-Odds (für zwei Modelle M_1 und M_2):

$$\frac{p(M_1 \mid x)}{p(M_2 \mid x)} = \frac{p(x \mid M_1)}{p(x \mid M_2)} \cdot \frac{p(M_1)}{p(M_2)}$$

Prior-Odds

Die **Prior-Odds** sind:

$$\frac{p(M_1)}{p(M_2)}$$

- Wie wahrscheinlich ist Modell M_1 im Vergleich zu Modell M_2 a priori?
- Unsere Vorannahmen über die Modelle beeinflussen die Entscheidung a posteriori direkt.
- Daten verändern die Prior-Odds zu Posterior-Odds.

Bayes-Faktor

Der Quotient

$$B(x) = \frac{p(x \mid M_1)}{p(x \mid M_2)}$$

heißt **Bayes-Faktor** zugunsten von M_1.

- **Posterior-Odds = Bayes-Faktor $\times$ Prior-Odds**
- Der Bayes-Faktor hängt nicht von der Prior auf die Modelle ab.
- Wohl aber von der Prior auf die Daten.

Marginale Likelihood

Zähler/Nenner des Bayes-Faktors ist die **marginale Likelihood** (oder **marginale Dichte**):

$$p(x \mid M_k) = \int p(x \mid \theta_k, M_k) p(\theta_k \mid M_k) \, d\theta_k$$

- Gibt an, wie wahrscheinlich die Daten insgesamt unter einem Modell sind.
- Erlaubt den Vergleich verschiedener Modelle, auch mit unterschiedlicher Komplexität.

Anmerkungen

- Für einfache Hypothesen ($H_0 : \theta = \theta_0, H_1 : \theta = \theta_1$) ist der Bayes-Faktor gleich dem Likelihood-Ratio.
- Die numerische Berechnung der marginalen Likelihood ist oft schwierig $\rightarrow$ moderne Methoden nötig.

Laplace-Approximation

Die Berechnung von Integralen ist z. B. über **Laplace-Integration** möglich.
Betrachte Integrale der Form:

$$\int_a^b g(x)\,dx = \int_a^b \exp(h(x))\,dx$$

mit $h(x)$: zweimal differenzierbare Funktion.
Ziel: Approximation des Integrals durch eine Normalverteilung.

Idee

- Ziel: $g(x)$ durch Normalverteilungsdichte approximieren.
- Setze $h(x) = \log(g(x))$: dann ist $h(x)$ näherungsweise quadratisch.
- Verwende Taylor-Entwicklung bis zum zweiten Term um Punkt x_0:

$$h(x) \approx h(x_0) + h'(x_0)(x - x_0) + \frac{h''(x_0)}{2}(x - x_0)^2$$

- $x_0 = \arg\max g(x) = \arg\max h(x)$
- Am Maximum: $h'(x_0) = 0$

Näherung des Integrals

Da $h'(x_0) = 0$, folgt:

$$h(x) \approx h(x_0) + \frac{h''(x_0)}{2}(x - x_0)^2$$

$$\int g(x)\, dx \approx \exp(h(x_0)) \cdot \int \exp\left(\frac{h''(x_0)}{2}(x - x_0)^2\right) dx$$

Die Funktion im Integral ist nun die Dichte einer **Normalverteilung** mit Erwartungswert x_0 und Varianz $-h(x_0)$. Das Integral über die Dichte ist 1.
Damit bleibt als Laplace-Approximation

$$\approx \exp(h(x_0)) \cdot \sqrt{2\pi \cdot (-h''(x_0))}$$

Ergebnis:

$$\int g(x)\, dx \approx \exp(h(x_0))\sqrt{2\pi(-h''(x_0))}$$

Algorithmus

Schritte:

1. Finde den Posteriori-Modus x_0
2. Berechne die zweite Ableitung der Log-Dichte an x_0
3. Setze in die Taylor-Approximation ein

Hinweis: Für numerische Optimierung werden Modus und zweite Ableitung benötigt.

Laplace-Approximation in R

```
laplace <- function(logpost, mode, y){ [basicstyle=] fit = optim(mode, logpost, method =
"Brent", y = y, hessian = TRUE, control = list(fnscale = -1), lower =
10^( - 6), upper = 10^6)mode = fitpar h =
-1/fithessianint = 0.5 * log(2 * pi) + 0.5 * log(h) + logpost(mode, y)return(int) }
```

- Numerisch oft besser: mit Log-Dichte statt Dichte rechnen.

Anwendung im Online Kurs

Laplace-Integration

Die Berechnung des Integrals ist z. B. über **Laplace-Integration** möglich (nach Pierre-Simon Laplace).

$$\int_a^b g(x)\,dx = \int_a^b \exp(h(x))\,dx$$

Dabei ist $h(x)$ eine zweimal differenzierbare Funktion.

Diese Methode wird verwendet zur Approximation marginaler Likelihoods.

Übersicht der Log Marginalen Likelihood

Modell	Log Marginale Likelihood
1	-137.7395
2	-151.0382
3	-138.8251

Bayesfaktor

Vergleiche zweier Modelle mit:

$$B_{12} = \exp\left(\log(p(x|M_1)) - \log(p(x|M_2))\right)$$

z. B.:

$$B_{12} \approx \exp(-137.7395 + 151.0382) \approx 596{,}450.8$$

Vergleich	**Bayesfaktor**
Modell 1 zu Modell 2	596450.8
Modell 1 zu Modell 3	3.0
Modell 3 zu Modell 2	201406.6

Skala des Bayes-Faktors

Nach **Jeffreys (1961)**:

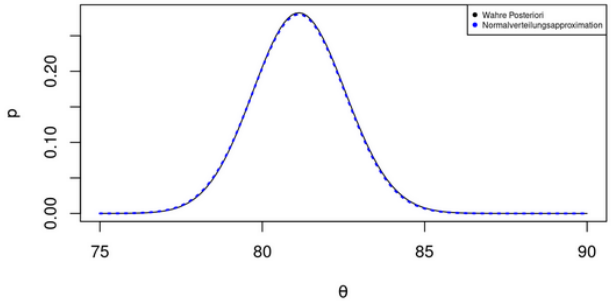
Bayes-Faktor B	**Interpretation**
$B < 1$	Modell I wird gestützt
$B \in [1, 10^{1/2}]$	Anzeichen für Modell k gegen Modell I, aber kaum erwähnenswert
$B \in [10^{1/2}, 10]$	**beachtliche** Anzeichen für Modell k gegen Modell I
$B \in [10, 10^{3/2}]$	**starke** Anzeichen
$B \in [10^{3/2}, 100]$	**sehr starke** Anzeichen
$B > 100$	**ausschlaggebende** Anzeichen

Dabei ist $10^{1/2} \approx 3.16$, $10^{3/2} \approx 31.6$

- Modell 1 und 3 beide mit ausschlaggebenden Anzeichen gegenüber Modell 2.
- $B_{13} = 3.0 \Rightarrow$ leichte Anzeichen für Modell 1 gegenüber Modell 3.

Güte der Laplace-Approximation

- Die Posteriori ist fast eine Normalverteilung – Approximation funktioniert gut.
- Für $n \to \infty$ konvergiert die Posteriori (unter Regularitätsbedingungen) gegen eine Normalverteilung.
- Bei asymmetrischer Posteriori: Laplace-Approximation schlechter, ggf. dritten Term der Taylor-Entwicklung einbeziehen.



Poisson-Regression

- Da wir Zähldaten haben, gehen wir von einem Poisson-Modell aus. Für jedes Jahr setzen wir an:

$$y_t \sim \text{Po}(\lambda_t e_t)$$

- y_t: Anzahl der Fälle im Jahr t
- e_t: Anzahl der erwarteten Fälle im Jahr t
- λ_t: unbekannte Rate im Jahr t

Poisson-Regression

Nun wollen wir den Einfluss der Kovariablen Tabakkonsum auf die Fälle modellieren:

- Im linearen Regressionsmodell mit Normalverteilungsannahme haben wir ein lineares Modell auf den Erwartungswert von y_t konstruiert.
- Der Erwartungswert der Poisson-Verteilung ist $\lambda_t e_t$. e_t ist gegeben, also modellieren wir λ_t.
- Ein lineares Modell ist problematisch, da die Rate λ positiv ist.
- **Idee:** Wir transformieren λ mit dem Logarithmus und machen ein lineares Modell auf den Logarithmus:

$$\log(\lambda_t) = \alpha + \beta x_t$$

- α: Intercept (Grundrate über alle Jahre)
- β: linearer Einfluss der Kovariablen auf die Log-Rate
- x_t: zeitverzögerter Tabakkonsum (z.B. 20 Jahre zuvor)

Priori:
Im Poisson-Modell ist die Gamma-Verteilung die konjugierte Priori-Verteilung für λ:

$$\lambda \sim \mathrm{Ga}(a, b)$$

Hier allerdings haben wir λ transformiert. Wir übernehmen stattdessen die Prioris aus der linearen Regression:

$$\alpha \sim N(m_\alpha, v_\alpha^2)$$
$$\beta \sim N(m_\beta, v_\beta^2)$$

Dichten

Datendichte:
Für eine einzelne Beobachtung gilt:

$$f(y_i \mid \lambda_i) = \frac{\lambda_i^{y_i}}{y_i!} \exp(-\lambda_i)$$

Für die gesamte Datendichte:

$$f(y \mid \lambda) = \prod_{t=1}^{T} \frac{\lambda_t^{y_t}}{y_t!} \exp(-\lambda_t)$$

mit $\lambda_t = \exp(\alpha + \beta x_t)$.

$$f(y \mid \alpha, \beta) \propto \left(\prod_{t=1}^{T} \exp(\alpha + \beta x_t)^{y_t} \right) \exp\left(- \sum_{t=1}^{T} \exp(\alpha + \beta x_t) \right)$$

$$= \exp\left(\alpha \sum_{t=1}^{T} y_t + \beta \sum_{t=1}^{T} x_t y_t - \sum_{t=1}^{T} \exp(\alpha + \beta x_t) \right)$$

Dichten

Posteriori: Zusammen mit den Prioris ergibt sich die Posteriori:

$$p(\alpha, \beta \mid y) \propto \exp\left(\alpha \sum y_t + \beta \sum x_t y_t - \sum \exp(\alpha + \beta x_t)\right) \tag{1}$$

$$\cdot \exp\left(-\frac{1}{2v_\alpha}(\alpha - m_\alpha)^2\right) \cdot \exp\left(-\frac{1}{2v_\beta}(\beta - m_\beta)^2\right) \tag{2}$$

$$= \exp\left(-\frac{\alpha^2}{2\nu_\alpha} + \alpha\left(\frac{m_\alpha}{\nu_\alpha} + \sum y_t\right)\right) \tag{3}$$

$$\cdot \exp\left(-\frac{\beta^2}{2\nu_\beta} + \beta\left(\frac{m_\beta}{\nu_\beta} + \sum x_t y_t\right)\right) \tag{4}$$

$$\cdot \exp\left(-\sum \exp(\alpha + \beta x_t)\right) \tag{5}$$

Man sieht, dass bei diesem Modell Datendichte und Priori nicht im (semi-)konjugierten Sinne zusammenpassen“. Der Gibbs-Sampler ist also nicht ohne Weiteres anwendbar.

Sampling-Methoden

Um Zufallszahlen aus einer Verteilung zu ziehen, gibt es diverse Methoden. Inversionsmethode
Gegeben sei die Verteilungsfunktion $F(x)$ einer Zufallsvariablen X. Sei $u \sim U[0,1]$. Dann ist

$$Y = F^{-1}(u) = inf\{y : F(y)u\} \sim X$$

Acceptance-Rejection-Methode:

- Ziel: Ziehen aus einer schwierigen Dichte $f(x)$ mithilfe einer einfach zu simulierenden Dichte $g(x)$.
- Es existiert ein $c > 0$ mit $cg(z) \geq f(z)$ für alle z mit $f(z) > 0$.
- Algorithmus:
  1. Ziehe Z gemäß $g(z)$.
  2. Ziehe $U \sim \mathrm{U}[0,1]$ unabhängig.
  3. Akzeptiere Z als Stichprobe aus $f(x)$, falls $U \leq \frac{f(Z)}{cg(Z)}$.
  4. Sonst: Wiederhole ab Schritt 1.
- Wähle c möglichst klein, um die Akzeptanzrate zu maximieren.
- Die Methode funktioniert auch, wenn $f(x)$ nicht normiert ist.

Squeezed Acceptance-Rejection-Sampling:

- Gegeben eine untere Schranke $s(z) \leq f(z)$.
- Akzeptiere Z sofort, wenn $U \leq \frac{s(z)}{cg(z)}$.
- Prüfe nur, ob $U \leq \frac{f(z)}{cg(z)}$, falls im ersten Schritt nicht akzeptiert wurde.

Importance Sampling:

- Ziel: Erwartungswert unter $f(\theta)$ bestimmen, Ziehungen aber aus $q(\theta)$.
- Zusammenhang:

$$\mathbb{E}_f[g(\theta)] = \int g(\theta) f(\theta) d\theta = \int g(\theta) \frac{f(\theta)}{q(\theta)} q(\theta) d\theta = \mathbb{E}_q \left[g(\theta) \frac{f(\theta)}{q(\theta)} \right]$$

- Schätzer mit Ziehungen $\theta_1, \ldots, \theta_m$ aus q:

$$\hat{g}^{IS} = \frac{1}{m} \sum_{i=1}^{m} g(\theta_i) \frac{f(\theta_i)}{q(\theta_i)}$$

- Die Varianz des Importanceschätzers ist

$$\mathrm{Var}(\hat{g}^{IS}) = \frac{1}{m} \mathrm{Var}_q \left(g(\theta) \frac{f(\theta)}{q(\theta)} \right)$$

- Die Normierungskonstante von f kann ebenfalls via Importance Sampling geschätzt werden.

Markov Chain Monte Carlo (MCMC)

Idee:

- Erzeuge eine Markovkette, deren stationäre Verteilung die gewünschte Posteriorverteilung ist.
- Die erzeugten Ziehungen sind voneinander abhängig.
- MCMC funktioniert auch für komplexe und hochdimensionale Probleme.

Markov-Eigenschaft:

- Eine Folge von Zufallsvariablen $Y = \{Y_t, t \in \mathbb{N}_0\}$ mit Zustandsraum S ist eine Markovkette, wenn gilt:

$$P(Y_t = k \mid Y_0 = j_0, \ldots, Y_{t-1} = j_{t-1}) = P(Y_t = k \mid Y_{t-1} = j_{t-1})$$

- Der nächste Zustand hängt nur vom aktuellen Zustand ab, nicht von der Vergangenheit.

Definitionen:

- $P(Y_t = k \mid Y_{t-1} = j)$ ist die Übergangswahrscheinlichkeit.
- Die Kette ist **homogen**, wenn die Übergangswahrscheinlichkeiten nicht von t abhängen.
- Die Übergangsmatrix $P = (p_{jk})$ sammelt alle Übergangswahrscheinlichkeiten.
- Eine Markovkette ist **irreduzibel**, wenn jeder Zustand von jedem anderen erreicht werden kann.
- Die **Periode** eines Zustands ist der größte gemeinsame Teiler der möglichen Rückkehrzeiten. Ist sie 1, ist der Zustand aperiodisch.

Markov-Kette: Wettermodell (Sonnenschein und Regen)

Übergangsmatrix:

$$A = \begin{pmatrix} 0.5 & 0.5 \\ 0.1 & 0.9 \end{pmatrix}$$

- Erster Eintrag: Sonnenschein
- Zweiter Eintrag: Regen

R-Berechnungen:

```
> a <- matrix(c(0.5, 0.5, 0.1, 0.9), nrow = 2, byrow = TRUE)
```

Markov-Kette: Wettermodell (Sonnenschein und Regen)

```
> a %*% a
     [,1] [,2]
[1,] 0.30 0.70
[2,] 0.14 0.86
```

```
> a %*% a %*% a %*% a %*% a %*% a %*% a %*% a %*% a %*% a %*% a %*% a %*% a
          [,1]      [,2]
[1,] 0.1666723 0.8333277
[2,] 0.1666655 0.8333345
```

```
> a %*% a %*% a %*% a %*% a %*% a %*% a %*% a %*% a %*% a %*% a %*% a %*% a %
          [,1]      [,2]
[1,] 0.1666676 0.8333324
[2,] 0.1666665 0.8333335
```

Interpretation:

- Nach vielen Schritten konvergiert die Verteilung gegen die stationäre Verteilung:

$$\pi = (0.1667,\ 0.8333)$$

- Das bedeutet: Langfristig ist die Wahrscheinlichkeit für Sonnenschein ca. 17%, für Regen ca. 83%.
- Dies illustriert die Motivation für MCMC: Die Markov-Kette konvergiert zur invarianten Verteilung, unabhängig vom Startzustand.

Invariante Verteilung:

- Eine Verteilung π ist invariant, wenn $\pi = \pi P$ gilt.
- Nach genügend vielen Schritten entspricht die Verteilung der Zustände der invarianten Verteilung π.
- Beispiel: Für eine Übergangsmatrix P kann man π berechnen, sodass $\pi = \pi P$.

Ergodizität:

- Eine Markovkette ist **ergodisch**, wenn sie irreduzibel und aperiodisch ist.
- Für jede Startverteilung π_0 gilt:

$$\lim_{t \to \infty} \pi_0 P^t = \pi$$

- Die Kette konvergiert gegen die stationäre Verteilung, der Einfluss der Startverteilung verschwindet.

Praxis:

- Im MCMC-Algorithmus muss die Markovkette homogen, irreduzibel und aperiodisch sein, damit die Zielverteilung erreicht wird.
- Die bekannten MCMC-Methoden gewährleisten diese Eigenschaften.

Invariante Verteilung einer Markovkette

Definition:

- Eine **invariante** (oder **stationäre**) Verteilung π einer homogenen Markovkette mit Übergangsmatrix P erfüllt:

$$\pi = \pi P$$

- Das bedeutet: Startet die Kette in der Verteilung π, bleibt sie für alle Zeitpunkte in dieser Verteilung.

Ergodizität und Konvergenz:

- Statt eines festen Startzustands kann man eine beliebige Startverteilung π_0 wählen.
- Eine Markovkette heißt **ergodisch**, wenn für jede Startverteilung π_0 gilt:

$$\lim_{t\to\infty} \pi_t = \lim_{t\to\infty} \pi_0 P^t = \pi$$

- Die Grenzverteilung einer ergodischen Markovkette ist die invariante Verteilung π.
- Eine homogene Markovkette ist ergodisch, wenn sie **irreduzibel** und **aperiodisch** ist.
- In diesem Fall konvergiert die Zustandsverteilung gegen π, unabhängig von der Startverteilung.
- Der Einfluss der Startverteilung verschwindet asymptotisch.

Metropolis-Algorithmus

Wir kommen daher zu einem weiteren MCMC-Verfahren: dem **Metropolis-Algorithmus**. Um aus einer mehrdimensionalen Verteilung $f(\theta)$ zu ziehen, geht man wie folgt vor: **Algorithmus:**

1. Setze Startwert $\theta^{(0)}$
2. Setze $k = 1$
3. Ziehe einen Vorschlag θ^* aus einer symmetrischen Vorschlagsverteilung mit Dichte $q(\theta^* \mid \theta^{(k-1)})$
4. Berechne die Akzeptanzwahrscheinlichkeit:

$$\alpha = \min\left(1, \frac{f(\theta^*)}{f(\theta^{(k-1)})}\right)$$

5. Ziehe $u \sim \mathcal{U}[0, 1]$
6. Falls $u \leq \alpha$, dann setze $\theta^{(k)} = \theta^*$
7. Falls $u > \alpha$, dann setze $\theta^{(k)} = \theta^{(k-1)}$
8. Setze $k = k + 1$

Wiederhole Schritte 3 bis 8 so oft wie nötig.

Bemerkungen zum Metropolis-Algorithmus

- Symmetrische Vorschlagsverteilung bedeutet:

$$q(\theta^* \mid \theta^{(k-1)}) = q(\theta^{(k-1)} \mid \theta^*)$$

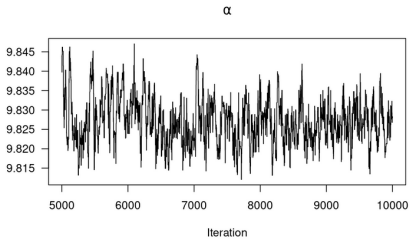
- Einfache Wahl: Normalverteilung mit Erwartungswert $\theta^{(k-1)}$ – sogenannter *Random Walk*.
- Ist $f(\theta^*) > f(\theta^{(k-1)})$, wird der Vorschlag immer akzeptiert.
- Ist $f(\theta^*) < f(\theta^{(k-1)})$, kann der Vorschlag verworfen werden.
- Wird der Vorschlag verworfen, wird der bisherige Wert erneut übernommen: auch bei stetigen Verteilungen können gleiche Werte mehrfach auftreten.
- Bei der Berechnung von α entfallen Konstanten, die nicht von θ abhängen:

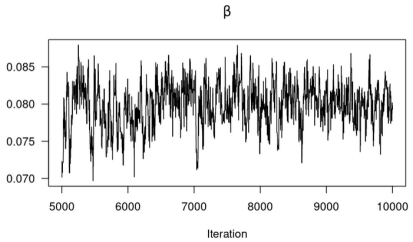
$$\alpha = \frac{f(\theta^*)}{f(\theta^{(k-1)})}$$

- Daher reicht es, die Posteriorverteilung nur bis auf eine Proportionalitätskonstante zu kennen.

Anwedung in R im Online Kurs

Auswertung des Metropolis-Algorithmus (Diagnostik)





Auswertung des Metropolis-Algorithmus (Diagnostik)

- **Konvergenz:**
  - Die Kette sollte nach einer gewissen Zeit (Burn-in) stabil um einen Mittelwert schwanken.
  - Keine sichtbaren Trends oder systematischen Drifts. Die können auf Modellfehler oder numerische Probleme hinweisen
- **Mischung (Mixing):**
  - Die Kette sollte den Parameterraum gut durchmischen.
  - Schnelle Wechsel zwischen verschiedenen Bereichen ohne lange Verweildauer in bestimmten Regionen.
  - Nur bei guter Mischung und Konvergenz ist die Posteriorverteilung korrekt repräsentiert.
- **Autokorrelation:**
  - Die Werte aufeinanderfolgender Ziehungen sollten möglichst wenig voneinander abhängen.
  - Hohe Autokorrelation deutet auf mangelnde Konvergenz führen zu verzerrten oder ineffizienten Schätzungen.

Auswertung des Metropolis-Algorithmus (Diagnostik)

- **Burn-in-Phase:**
  - Anfangsbereich (Burn-in) sollte entfernt werden, da die Kette dort noch nicht stationär ist.
- **Ausreißer:**
  - Keine plötzlichen Sprünge oder Ausreißer, die auf Probleme beim Sampling hinweisen könnten.
- **Vergleich mehrerer Ketten:**
  - Bei mehreren parallelen Ketten sollten alle Ketten ähnliche Verläufe und Mittelwerte zeigen.

MCMC-Algorithmen: Überblick und Prinzipien

MCMC steht für Markov Chain Monte Carlo:

- **Monte Carlo:** Numerische Algorithmen, die auf dem Ziehen von Zufallszahlen beruhen.
- **Markov Chain:** Jede Ziehung hängt nur vom vorherigen Zustand ab (Markov-Eigenschaft), nicht von weiter zurückliegenden Zuständen.
- Die Folge der Ziehungen wird als *Kette* bezeichnet.

MCMC-Algorithmen: Überblick und Prinzipien

Zentrale MCMC-Algorithmen:

- **Gibbs-Sampler:**
  - Ziehe jeden Parameter abwechselnd aus seiner bedingten Verteilung (*full conditional*).
  - Funktioniert nur, wenn diese Verteilungen bekannt und direkt simulierbar sind.
- **Metropolis-Algorithmus:**
  - Ziehe einen Vorschlag für alle Parameter (oder einen Teil) und akzeptiere diesen mit einer bestimmten Wahrscheinlichkeit.
  - Nur das Verhältnis der Posterior-Dichten (bis auf Konstanten) muss bekannt sein.
  - Die Wahl der Vorschlagsdichte beeinflusst die Effizienz des Algorithmus.
- **Metropolis-within-Gibbs:**
  - Ziehe wie beim Gibbs-Sampler aus $\theta_i | \theta_{-i}$.
  - **Falls die bedingte Verteilung nur bis auf eine Proportionalitätskonstante bekannt ist, nutze für diesen Schritt den Metropolis-Algorithmus.**

Poisson-Regression mit Überdispersion

Standardmodell:

- Bisher: $y_t \sim \text{Po}(\lambda_t e_t)$
- Bei der Poisson-Verteilung gilt: Erwartungswert = Varianz.
- In der Praxis ist die Varianz oft größer als der Erwartungswert (**Überdispersion**).

Erweiterung um Überdispersion:

- Einführung eines Überdispersionsparameters:

$$\log(\lambda_t) = \alpha + \beta x_t + \epsilon_t$$

- ϵ_t ist ein Fehlerterm, analog zur linearen Regression.
- Annahme: $\epsilon_t \sim N(0, \sigma^2)$ mit z.B. $\sigma^2 = 0{,}001$ (Regularisierung).
- Kleine Varianz drückt ϵ_t Richtung 0: Die Überdispersion erklärt nur zusätzliche Varianz, die das Poisson-Modell nicht abbildet.

Warum Regularisierung?

- Bei vielen Parametern und wenigen Daten verhindert die Regularisierung Überanpassung.
- Ohne Regularisierung könnte das Modell die Daten deterministisch festlegen.
- Die Regularisierungspriori zwingt ϵ_t nahe 0 und sorgt für stabile Schätzungen.

Posteriori-Verteilung:

- Die Posteriori für α, β, η (mit $\eta_t = \log(\lambda_t)$) kombiniert Likelihood und Prioris:

$$p(\alpha, \beta, \eta \mid y) \propto \exp\left(\sum_{t=1}^{T} \eta_t y_t - \sum_{t=1}^{T} \exp(\eta_t)\right) \exp\left(-\frac{1}{2v_\alpha}(\alpha - m_\alpha)^2\right) \tag{6}$$

$$\exp\left(-\frac{1}{2v_\beta}(\beta - m_\beta)^2\right) \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(\eta_t - \alpha - \beta x_t)^2\right) \tag{7}$$

- Die Full conditionals für α und β sind Normalverteilungen; für η_t ist ein Metropolis-Schritt nötig.

Full conditional für α:

- In der Posteriori hängen von α nur noch zwei Terme ab:

$$p(\alpha \mid \beta, \lambda; y) \propto \exp\left(-\frac{1}{2v_\alpha}(\alpha - m_\alpha)^2\right) \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(\eta_t - \alpha - \beta x_t)^2\right)$$

- Die Summe zweier quadratischer Terme ergibt wieder eine Normalverteilung:

$$\alpha \sim N(\tilde{m}_\alpha, \tilde{v}_\alpha)$$

$$\tilde{v}_\alpha = \left(\frac{1}{v_\alpha} + \frac{T}{\sigma^2}\right)^{-1}$$

$$\tilde{m}_\alpha = \tilde{v}_\alpha\left(\frac{m_\alpha}{v_\alpha} + \frac{1}{\sigma^2}\sum_{t=1}^{T}(\eta_t - \beta x_t)\right)$$

Full conditional für β:

- Analog ergibt sich für β:

$$\beta \sim N(\tilde{m}_\beta, \tilde{v}_\beta)$$

$$\tilde{v}_\beta = \left(\frac{1}{v_\beta} + \frac{1}{\sigma^2} \sum_{t=1}^{T} x_t^2 \right)^{-1}$$

$$\tilde{m}_\beta = \tilde{v}_\beta \left(\frac{m_\beta}{v_\beta} + \frac{1}{\sigma^2} \sum_{t=1}^{T} x_t (\eta_t - \alpha) \right)$$

Full conditional für η_t:

- Für η_t ergibt sich:

$$p(\eta_t \mid \alpha, \beta, \eta_{-t}; y) \propto \exp(\eta_t y_t) \exp(-\exp(\eta_t)) \cdot \exp\left(-\frac{1}{2\sigma^2}(\eta_t - \alpha - \beta x_t)^2\right)$$

- Dies ist keine Standardverteilung; ein Metropolis-Schritt ist erforderlich.

Metropolis-Hastings

- Wir verwenden hier eine Erweiterung des Metropolis-Algorithmus, der von Hastings vorgeschlagen wurde und daher Metropolis-Hastings-Algorithmus genannt wird.
- Dafür verwenden wir eine nicht-symmetrische Vorschlagsverteilung mit Dichte $q(\theta^*|\theta^{(k-1)})$.
- Die Vorschlagsverteilung kann von der letzten Ziehung $\theta^{(k-1)}$ abhängen, muss sie aber nicht.
- Die Dichte der Vorschlagsverteilung muss dann in der Akzeptanzwahrscheinlichkeit berücksichtigt werden:

$$\alpha = min\left(1, \frac{f(\theta)q(\theta^{(k-1)}|\theta)}{f(\theta^{(k-1)})q(\theta|\theta^{(k-1)})}\right)$$

- Für eine symmetrische Vorschlagverteilung $q(\theta^{(k-1)}|\theta) = q(\theta|\theta^{(k-1)})$ wird der Metropolis-Hastings-Algorithmus zum Metropolis-Algorithmus Nimmt man die full conditional als Vorschlagsverteilung, dann wird die Akzeptanzwahrscheinlichkeit gleich 1, wir sind dann beim Gibbs-Sampler.

Metropolis-Hastings

- Eine Idee, eine sinnvolle Vorschlagsverteilung zu konstruieren ist, die full conditional möglichst gut zu approximieren. Ist $q(\theta|\theta^{(k-1)}) \approx f(\theta)$, dann wird die Akzeptanzwahrscheinlichkeit fast 1, wir sind also "fast" beim Gibbs-Sampler.

Metropolis-Hastings-Algorithmus: Vorschlagsverteilung und Anwendung

Idee der Vorschlagsverteilung:

- Die Akzeptanzwahrscheinlichkeit im Metropolis-Hastings-Algorithmus ist am größten, wenn die Vorschlagsverteilung $q(\theta^*|\theta^{(k-1)})$ die full conditional möglichst gut approximiert.
- Ist $q(\theta^*|\theta^{(k-1)}) \approx f(\theta^*)$, dann ist die Akzeptanzrate fast 1 – wir sind also fast“ beim Gibbs-Sampler.
- Gleichzeitig muss man effizient aus q ziehen können.

Anwendung im Poisson-Modell:

- Die full conditional für η_t lautet:

$$p(\eta_t|\alpha, \beta, \eta_{-t}, y) \propto \exp(\eta_t y_t) \exp(-\exp(\eta_t)) \exp\left(-\frac{1}{2\sigma^2}(\eta_t - \alpha - \beta x_t)^2\right)$$

- Der erste und dritte Term passen zu einer Normalverteilung, nur $\exp(-\exp(\eta_t))$ nicht.
- Lösung: Taylor-Approximation von $\exp(\eta_t)$ um einen Punkt a bis zur 2. Ordnung:

$$\exp(\eta_t) \approx \exp(a) + \exp(a)(\eta_t - a) + \frac{\exp(a)}{2}(\eta_t - a)^2$$

Anwendung im Poisson-Modell:

- Einsetzen ergibt eine (approximative) Normalverteilung für η_t:

$$\eta_t \sim N\left(m, \frac{1}{v}\right)$$

mit

$$m = y_t - \exp(a) + a\exp(a) + \sigma^{-2}(\alpha + \beta x_t)$$

$$v = \sigma^{-2} + \exp(a)$$

- Diese Normalverteilung wird als Vorschlagsverteilung im Metropolis-Hastings-Schritt verwendet.
- Die η_t sind dabei voneinander unabhängig.

Monte-Carlo-Integration

Gegeben sei eine beliebige stetige Funktion $f(x) > 0$ mit bekanntem Wertebereich $[0, Y]$. Das Integral

$$\int_a^b f(x)\,dx$$

kann mit Monte-Carlo-Integration wie folgt approximiert werden:

1. Ziehe n gleichverteilte Zufallszahlen x_i aus $[a, b]$.
2. Ziehe unabhängig davon n gleichverteilte Zufallszahlen y_i aus $[0, Y]$.
3. Berechne den Anteil h der Punkte (x_i, y_i), die unterhalb der Funktion f liegen, d.h. $y_i < f(x_i)$.
4. Das Integral wird approximiert durch:

$$\int_a^b f(x)\,dx \approx h \cdot (b - a) \cdot Y$$

Monte-Carlo-Schätzer

Gegeben sei eine Dichte $p(x)$. Integrale der Form

$$\mathbb{E}(g(x)) = \int g(x)p(x)\,dx$$

können mit einer Stichprobe $x_1, \ldots, x_m$ aus $p(x)$ durch den Stichprobenmittelwert

$$\overline{g}_m = \frac{1}{m}\sum_{i=1}^{m} g(x_i)$$

approximiert werden.

Gesetz der großen Zahlen:

- Nach dem starken Gesetz der großen Zahlen gilt:

$$\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} g(x_i) = \int g(x)p(x)\,dx$$

- Das Gesetz gilt auch für nicht unabhängige Zufallszahlen (Ergodensatz).

Varianz des Monte-Carlo-Schätzers:

- Die Varianz von $\overline{g}_m$ ist:

$$\mathrm{Var}(\overline{g}_m) = \frac{1}{m} \int (g(x) - \mathbb{E}(g(x)))^2 p(x) dx = \frac{1}{m} \mathrm{Var}(g)$$

- Der Approximationsfehler verringert sich mit steigendem m.
- Nach dem zentralen Grenzwertsatz gilt:

$$\sqrt{m}\,(\overline{g}_m - \mathbb{E}(g(x))) \sim N(0, \mathrm{Var}(g))$$

Schätzer für die Varianz:

$$\widehat{\mathrm{Var}}(\overline{g}_m) = \frac{1}{m-1} \sum_{i=1}^{m} (g(x_i) - \overline{g}_m)^2$$

Anwendung:

- Die Definitionen gelten sowohl für diskrete als auch für stetige Zustände (dort spricht man vom Übergangskern statt von der Übergangsmatrix).
- Beim MCMC-Algorithmus muss die Markovkette homogen, irreduzibel und aperiodisch sein, damit die Zielverteilung als stationäre Verteilung erreicht wird.

Metropolis-Hastings-Algorithmus

Algorithmus:

- Ausgangspunkt: Aktueller Wert $\theta^{(k-1)}$.
- Ziehe einen Vorschlag θ^* aus einer Vorschlagsdichte $q(\theta^*|\theta^{(k-1)})$.
- Akzeptiere θ^* mit Wahrscheinlichkeit

$$\alpha = \min\left(1, \frac{p(\theta^*|x)\, q(\theta^{(k-1)}|\theta^*)}{p(\theta^{(k-1)}|x)\, q(\theta^*|\theta^{(k-1)})}\right)$$

- Falls akzeptiert: $\theta^{(k)} = \theta^*$, sonst $\theta^{(k)} = \theta^{(k-1)}$.
- Der Algorithmus erzeugt eine homogene Markovkette mit der gewünschten Zielverteilung als stationäre Verteilung.

Typen von Vorschlagsdichten:

- **Independence Proposal:** $q(\theta^*|\theta^{(k-1)})$ ist unabhängig von $\theta^{(k-1)}$.
- **Symmetrisches Proposal:** $q(\theta^*|\theta^{(k-1)}) = q(\theta^{(k-1)}|\theta^*)$ (z.B. Normalverteilung). Dann kürzt sich q in α heraus:

$$\alpha = \frac{p(\theta^*|x)}{p(\theta^{(k-1)}|x)}$$

- **Random Walk Proposal:** $\theta^* = \theta^{(k-1)} + \epsilon$, $\epsilon \sim f$. Typisch: $\theta^* \sim N(\theta^{(k-1)}, C)$ mit Kovarianzmatrix C.

Tuning:

- Zu kleine Varianz C: hohe Akzeptanzrate, aber starke Autokorrelation und ineffiziente Erkundung.
- Zu große Varianz C: niedrige Akzeptanzrate, viele Vorschläge werden abgelehnt.
- Tuning der Kovarianzmatrix ist entscheidend für die Effizienz.

Multivariate und komponentenweise Updates:

- Der Algorithmus kann für Vektoren θ angewandt werden.
- In hoher Dimension oft geringere Akzeptanzraten.
- Alternative: Komponentenweise Metropolis-Hastings, d.h. jede Komponente (skalar oder blockweise) wird einzeln aktualisiert:

$$\alpha = \min\left(1, \frac{p(\theta_1^*|x, \theta_2^{(k-1)})q(\theta_1^{(k-1)}|\theta_1^*)}{p(\theta_1^{(k-1)}|x, \theta_2^{(k-1)})q(\theta_1^*|\theta_1^{(k-1)})}\right)$$

- Updates können in fester oder zufälliger Reihenfolge erfolgen.

Gibbs-Sampling

Prinzip:

- Gibbs-Sampling ist ein Spezialfall des Metropolis-Hastings-Algorithmus, bei dem die Vorschlagsverteilung die vollständige bedingte Posteriorverteilung (full conditional) ist.
- Für multivariate Parametervektoren $\theta = (\theta_1, \ldots, \theta_p)$ wird iterativ jede Komponente (oder jeder Block) aus ihrer bedingten Verteilung $p(\theta_j | x, \theta_{-j})$ gezogen.
- Die Akzeptanzwahrscheinlichkeit ist immer 1, da der Vorschlag direkt aus der Zielverteilung gezogen wird.

Algorithmus:

- Initialisiere $\theta^{(0)}$.
- Für jede Iteration k und jede Komponente j:
  1. Ziehe $\theta_j^{(k)} \sim p(\theta_j | x, \theta_{-j}^{(k)})$
- Ein vollständiger Durchlauf über alle Komponenten ist ein "Zyklus" des Gibbs-Samplers.

Block-Gibbs-Sampling:

- Komponenten können auch zu Blöcken zusammengefasst werden: $\theta = (\theta_1, \ldots, \theta_p)$, wobei jeder Block gemeinsam aktualisiert wird.
- Jeder Block wird dann aus der bedingten Verteilung $p(\theta_j | x, \theta_{-j})$ gezogen.

Vorteile/Nachteile

Vorteile:

- Keine Notwendigkeit, eine Vorschlagsverteilung zu tunen.
- Proposals werden immer akzeptiert.
- Besonders effizient, wenn die full conditionals Standardverteilungen sind.

Nachteile:

- Die vollständigen bedingten Verteilungen müssen bekannt und einfach simulierbar sein.
- Kann bei stark korrelierten Parametern langsam konvergieren.