Einführung in die Bayes-Statistik

Volker Schmid, Clara Happ

27.–30. März 2017

Institut für Statistik
Ludwig–Maximilians–Universität München

Inhaltsverzeichnis

1 Der Satz von Bayes **3**
- 1.1 Wahrscheinlichkeitsrechnung . . . 3
- 1.2 Satz von Bayes . . . 4
- 1.3 Bayesianisches Lernen . . . 6

2 Bayes-Inferenz **7**
- 2.1 Zuvallsvariablen und Dichten . . . 7
- 2.2 Wichtige Verteilungen . . . 8
- 2.3 Bayes und die Billardkugeln . . . 10
- 2.4 Bayesianische Inferenz . . . 12

3 Prioris **13**
- 3.1 Subjektive Priori . . . 13
- 3.2 Konjugierte Priori . . . 14
- 3.3 Nichtinformative Priori . . . 15
- 3.4 Regularisierungspriori . . . 18

4 Bayesianisches Schätzen und Testen **19**
- 4.1 Punktschätzer . . . 19
- 4.2 Intervallschätzer . . . 21
- 4.3 Optimalität der Schätzer* . . . 22
- 4.4 Bayes-Tests* . . . 24

5 Bayesianische Modellierung **26**
- 5.1 Normalverteilungsmodel . . . 26
- 5.2 Regression . . . 29
- 5.3 Hierarchische Modellierung . . . 31
- 5.4 Empirischer Bayes* . . . 34

6 Numerische Verfahren zur Bestimmung der Posteriori **34**
- 6.1 Numerische Integration* . . . 35
- 6.2 Laplace-Approximation . . . 36
- 6.3 Monte-Carlo-Integration . . . 38

7 Markov Chain Monte Carlo **41**
- 7.1 Markovketten* . . . 42
- 7.2 Metropolis-Hastings-Algorithmus . . . 45
- 7.3 Gibbs-Sampling . . . 47

Einleitung

Thomas Bayes

Geschichte

- Mitte 18. Jahrhundert: Bayes entwickelt die Bayes-Formel
- Anfang 19. Jahrhundert: Pierre-Simon Laplace entwickelt die Bayes-Formel, prägt den Begriff "Inverse Wahrscheinlichkeit"
- Anfang 20. Jahrhundert: Ronald Fisher entwickelt den Frequentismus, Maximum-Likelihood-Schätzer, prägt den Begriff "Bayes-Statistik"
- Ende des 20. Jahrhunderts: Bayes-Verfahren werden wieder aktuell, komplexe Modelle dank Computer möglich

Literatur: Stephen E. Fienberg: When Did Bayesian Inference Become "Bayesian"? Bayesian Analysis (2006) 1(1), pp. 1–40.

1 Der Satz von Bayes

1.1 Wahrscheinlichkeitsrechnung

Wir interessieren uns (vorerst) für zufällige Ereignisse in einem Ereignisraum Ω. Jedem Ereignis A kann eine Wahrscheinlichkeit $P(A)$ zugeordnet werden, für die die Axiome von Kolmogorov gelten:

- $0 \leq P(A) \leq 1$
- Für das das sichere Ereignis Ω gilt $P(\Omega) = 1$
- Für beliebige disjunkte Ereignisse A und B gilt $P(A \cup B) = P(A) + P(B)$

Daraus folgt direkt:

- $P(\emptyset) = 0$
- $P(\overline{A}) = 1 - P(A)$

Wahrscheinlichkeit kann unterschiedlich interpretiert werden:

- klassisch: Alle Elementarereignisse haben die selbe Wahrscheinlichkeit
- als relative Häufigkeit bei unendlicher Wiederholung
- subjektiv als Maß für den Glauben an das Eintreten
- als Eigenschaft eines physikalischen Systems (Propensität)

Beispiel 1.1 (Würfel)**.** *Wir werfen einen fairen sechsseitigen Würfel. Die Wahrscheinlichkeit des Ereignisses A: "Es fällt eine 6" ist* $P(A) = 1/6$.

Alternativ lassen sich Wahrscheinlichkeiten auch als Odds darstellen:

Definition 1.1. $Odds(A) = \frac{P(A)}{P(\overline{A})}$

Beispiel 1.2 (Fortsetzung Bsp.1.1)**.** *der Odds für A ist*

$$Odds(A) = \frac{1/6}{5/6} = 1 : 5$$

Die Wettquote für A ist hier dementsprechend $5 : 1$. *Wettet man erfolgreich auf A, erhält man neben dem Einsatz das Fünffache des Einsatzes zurück, zusammen also das Sechsfache des Einsatzes.*

Zwei Ereignisse heißen unabhängig, wenn

$$P(A \cap B) = P(A) \cdot P(B)$$

Beispiel 1.3 (Fairer Würfel)**.** *Sei A: "Es fällt eine 5 oder 6", B: "Es fällt eine gerade Zahl". Dann ist $A \cap B$: "Es fällt eine 6". Es gilt*

$$P(A) \cdot P(B) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6} = P(A \cap B)$$

Als bedingte Wahrscheinlichkeit $P(A|B)$ bezeichnen wir das Eintreten eines Ereignisses A unter der Bedingung, das ein Ereignis B eingetreten ist.

Beispiel 1.4 (Fairer Würfel)**.** *Sei A: "Es fällt eine 4, 5 oder 6", B: "Es fällt eine gerade Zahl". Dann ist $P(A|B) = \frac{2}{3}$*

Definition 1.2. *Wir nennen*

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

bedingte Wahrscheinlichkeit von A gegeben B. $P(B)$ heißt marginale Wahrscheinlichkeit von B.

1.2 Satz von Bayes

Es gilt also:

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ P(A \cap B) &= P(B|A)P(A) \\ P(A|B)P(B) &= P(B|A)P(A) \end{aligned}$$

Daraus folgt direkt:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Mit dem Satz von der totalen Wahrscheinlichkeit:

$$P(B) = P(B|A)P(A) + P(B|\overline{A})P(\overline{A})$$

erhalten wir

Definition 1.3. *Satz von Bayes*

$$\boxed{P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B|A)P(A)+P(B|\overline{A})P(\overline{A})}}$$

Beispiel 1.5 (Würfel)**.** *Sei B: "Eine 6 fällt" und*

$$\begin{array}{ll} A: & \textit{Würfel ist fair;} \quad P(B|A) = 1/6 \\ \bar{A}: & \textit{Würfel ist unfair;} \quad P(B|\bar{A}) = 1 \end{array}$$

Nach Laplace gehen wir von $P(A) = P(\bar{A}) = 1/2$ *aus. Dann gilt*

$$\begin{aligned} P(A|B) &= \frac{\frac{1}{6}\cdot\frac{1}{2}}{\frac{1}{6}\cdot\frac{1}{2}+1\cdot\frac{1}{2}} = \frac{2}{5} \\ P(A|\bar{B}) &= \frac{\frac{5}{6}\cdot\frac{1}{2}}{\frac{5}{6}\cdot\frac{1}{2}+0\cdot\frac{1}{2}} = 1 \end{aligned}$$

Beispiel 1.6 (Medizinische Tests)**.** *Ein medizinischer Test soll das vorliegen einer Krankheit feststellen. Solche Tests sind nicht ganz fehlerfrei, es kommt zu falsch positiven und falsch negativen Ergebnissen. Wir definieren uns folgende Ereignisse*

A Eine Person ist krank

B Der Test zeigt ein positives Ergebnis

Es gelte

$P(A) = 0.02, P(\overline{A}) = 0.98$

$P(B|A) = 0.95, P(\overline{B}|A) = 0.05$ *(falsch negativ)*

$P(B|\overline{A}) = 0.1$ *(falsch positiv),* $P(\overline{B}|\overline{A}) = 0.9$

Wie groß ist die Wahrscheinlichkeit, krank zu sein, wenn der Test positiv ausfällt?

$$\begin{aligned} P(A|B) &= \frac{P(B|A)\cdot P(A)}{P(B|A)P(A)+P(B|\overline{A})P(\overline{A})} \\ &= \frac{0.95\cdot 0.02}{0.95\cdot 0.02+0.1\cdot 0.98} \\ &= \frac{0.019}{0.019+0.098} = 0.162\ldots \end{aligned}$$

1.3 Bayesianisches Lernen

Der Satz von Bayes verhilft uns, aus B für A zu lernen. Allerdings nur, wenn A und B nicht (stochastisch) unabhängig sind. Sonst folgt:

$$P(A|B) = P(A|\overline{B}) = P(A)$$

Sind A und B dagegen abhängig, dann folgt aus

$$P(A \cap B) \neq P(A)P(B)$$

und der Definition der bedingten Wahrscheinlichkeit

$$P(A|B) \neq P(A),$$

dass heißt, das Eintreten von B ändert die Wahrscheinlichkeit für das Eintreten von A, "wir lernen aus B".

Beispiel 1.7. *Wir betrachten im Beispiel 1.6 die Odds:*

$$\begin{aligned} \frac{P(A|B)}{P(\overline{A}|B)} &= \frac{P(B|A)}{P(B|\overline{A})} \cdot \frac{P(A)}{P(\overline{A})} \\ &= \frac{0.95}{0.1} \cdot \frac{0.02}{0.98} = \frac{0.019}{0.098} = 0.194 \end{aligned}$$

Beispiel 1.8. *Ein Los von Stücken wurde mit Wahrscheinlichkeit 0.7 in Firma A und mit Wahrscheinlichkeit 0.3 in Firma B produziert.*

Die Ausschußquoten betragen 1% in Firma A und 5% in Firma B.

Die Information über den Produzenten ist verloren gegangen.

Bei einer Kontrolle von n Stücken werden y Ausschußstücke entdeckt. Ist aus diesem Ergebnis ein Rückschluss auf den Produzenten möglich?

Für $n = 100$ ergeben sich folgende Posteriori-Wahrscheinlichkeiten:

y	0	1	2	3	4	5	6
$P(A\|y)$	0.993	0.965	0.842	0.505	0.164	0.036	0.007
$P(\overline{A}\|y)$	0.007	0.035	0.158	0.495	0.836	0.964	0.993

Für $P(A) = P(\overline{A}) = 0.5$ ergeben sich folgende Posteriori-Wahrscheinlichkeiten:

y	0	1	2	3	4	5	6
$P(A\|y)$	0.984	0.922	0.695	0.304	0.077	0.016	0.003
$P(\overline{A}\|y)$	0.016	0.077	0.305	0.696	0.923	0.984	0.997

2 Bayes-Inferenz

2.1 Zuvallsvariablen und Dichten

Statt Ereignissen betrachten wir nun Zufallsvariablen, also zufällige Größen. In Bsp. 1.1 ist der Würfelwurf eine Zufallsvariable. Damit können wir Wahrscheinlichkeiten von Zufallsvariablen definieren:

$$P(X = x) = P(\text{"Die Zufallsvariable } X \text{ nimmt den Wert } x\text{an.)}$$

Jede Zufallsvariable hat eine Verteilung. Wir unterscheiden

- Diskrete Verteilungen: Zufallsvariable nimmt ganzzahlige Werte an (z.B. Würfel)
- Stetige Verteilungen: Zufallsvariable ist reellwertig (z.B. Normalverteilung)

Verteilungen können wir über ihre Dichten f definieren. Es gilt:

- im diskreten Fall $P(X = a) = f(a)$
- im stetigen Fall $P(a < X < b) = \int_a^b f(x)dx$
- $f(x) \geq 0$
- *Verteilungsfunktion*: $F(x) = P(x \leq X)$
- $f(x) = \frac{d}{dx}F(x)$ (falls Verteilungsfunktion differenzierbar)
- Normierung: es gilt immer $\int_{-\infty}^{\infty} f(x) = 1$

Für diskrete Zufallsvariablen wird das Integral durch die Summe ersetzt: $\int f(x)dx := \sum_x f(x)$.

Satz 2.1 (Transformationssatz*)**.** *Sei f_X die Dichte der Zuvallsvariablen X und $T : x \to T(x)$ eine bijektive Abbildung. Dann gilt für die Dichte f_T von $T(X)$*

$$f_T(y) = |(T^{-1})'(y)| \cdot f\left(T^{-1}(y)\right)$$

Beispiel 2.1. *Quadrierte Standardnormalverteilung**

Seien X_1 und X_2 zwei Zufallsvariablen. Dann ist

- $f(\mathbf{x}) = f(x_1, x_2)$ die Dichte der gemeinsamen Verteilung von X_1 und X_2, wenn gilt
$$P((X_1, X_2) \in A) = \int_A f(x_1, x_2)dx_1x_2$$

- die Dichte der marginalen Verteilung von X_1 ist

$$f(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2$$

- die Dichte der bedingten Verteilung von X_1 gegeben X_2 hat den Wert x_2 ist

$$f(x_1|X_2 = x_2) = f(x_1|x_2) = \frac{f(x_1, x_2)}{f(x_2)}$$

Beispiel 2.2. *Zweidimensionale Normalverteilung*

Definition 2.1 (Satz von Bayes für Dichten)**.**

$$\boxed{f(\theta|x) = \frac{f(x|\theta) \cdot f(\theta)}{f(x)} = \frac{f(x|\theta) \cdot f(\theta)}{\int f(x|\tilde{\theta}) \cdot f(\tilde{\theta}) d\tilde{\theta}}}$$

2.2 Wichtige Verteilungen

Gleichverteilung Laplace prägte das "Prinzip vom unzureichenden Grund", jedes Elementarereignis ist gleich wahrscheinlich.

Diskrete Gleichverteilung $X \sim U(1, \ldots, n)$:

$$f(i) = P(X = i) = 1/n \text{für alle} i = 1, \ldots, n.$$

Stetige Gleichverteilung $X \sim U[a, b]$:

$$f(x) = \frac{1}{b-a} \text{für} a \leq x \leq b.$$

Damit gilt für $a < c < d < b$

$$P(c < x < d) = \frac{d-c}{b-a}.$$

Binomialverteilung Führt man ein Zufallsexperiment mit zwei möglichen Ausgängen (z.B. Erfolg/Mißerfolg, allgemein 1/0) durch, so spricht man von einem Bernoulliexperiment mit (Erfolgs-)Wahrscheinlichkeit π. Wiederholt man n Bernoulliexperimente und zählt die Erfolge/1en, so spricht man von einem Binomalexperiment.

$X \sim B(1, \pi)$:

$$f(x) = P(X = x) = \pi^x (1-\pi)^{1-x}$$

$X \sim B(n, \pi)$:

$$f(x) = P(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

Dabei ist $\binom{n}{x}$ die Anzahl der möglichen Reihenfolgen von genau x Einsen bei n Versuchen.

Normalverteilung Die Normalverteilung $X \sim N(\mu, \sigma^2)$ ist die wichtigste Verteilung der Statistik (Zentraler Grenzwertsatz!), z.B. als Verteilung des Beobachtungsfehlers im linearen Modell. Es gilt:

$$f(x) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Es gilt:

$$(x-\mu)^2 = x^2 - 2x\mu + \mu^2$$

und damit

$$\begin{aligned} f(x) &= \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left(-\frac{1}{2\sigma^2}x^2 + x\frac{\mu}{\sigma 2} - \frac{\mu}{\sigma^2}\right) \\ &= \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left(-\frac{1}{2}\tau x^2 + x \cdot m - \frac{\mu}{\sigma^2}\right) \end{aligned}$$

Diese Form der Dichte nennt man **kanonische Form der Normalverteilung** mit Parametern m und τ (Präzision). Es gilt: $E(X) = m/\tau$ und $Var(X) = 1/\tau$.

Betaverteilung Die Betaverteilung $X \sim Beta(a, b,)$ ist eine stetige Verteilung auf dem Intervall $[0, 1]$. Die Dichte ist

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

mit $B(a, b)$ die Betafunktion – diese garantiert uns hier die Normiertheit.

Die stetige Gleichverteilung ist ein Spezialfall der Betaverteilung: $B(1, 1) \sim U[0, 1]$

Gammaverteilung Die Gamma-Verteilung ist eine stetige Verteilung auf dem Intervall $[0, \infty)$. Die Dichte ist

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-xb)$$

mit $\Gamma(a, b)$ die Gammafunktion. Vorsicht, es existiert eine alternative Parametrisierung der Dichte mit $k = a, \theta = 1/b$.

2.3 Bayes und die Billardkugeln

Beispiel 2.3 (An Essay towards solving a Problem in the Doctrine of Chance (1763))**.** *Eine weiße Billiardkugel wird auf eine Gerade der Länge 1 gerollt. Die Wahrscheinlichkeit dafür, dass sie an einem Punkt π zu liegen kommt, ist konstant für alle $\pi \in [0,1]$. Eine rote Kugel wird unter den selben Bedinungen n-mal gerollt. Sei x die Zahl der Versuche, in denen die rote Kugel links von der ersten Kugel, also links von π zu liegen kommt.*

Welche Information über π erhalten wir aus der Beobachtung x?

Sei die weiße Kugel bereits gerollt und liege auf dem Punkt π. Die rote Kugel gerollt. Dann ist die Wahrscheinlichkeit, dass die rote Kugel links von der weißen zu liegen kommt gleich π. Rollen wir n-mal, so handelt es sich um ein Binomialexperiment mit Erfolgswahrscheinlichkeit π. Gegeben $\Pi = \pi$ ist also:

$$P(X = x|\Pi = \pi) = f(x|\pi) = \binom{n}{x}\pi^x(1-\pi)^{n-x}.$$

Um nun mit dem Satz von Bayes eine Aussage über *pi* gegeben x zu machen, brauchen wir $f(\pi)$. Was wissen wir über π vor der Beobachtung?

Annahme: Vor der Beobachtung, lateinisch *a priori*, wissen wir nichts über π. Wir folgen dem Prinzip von unzureichenden Grund und nehmen $\pi \sim U[0,1]$.

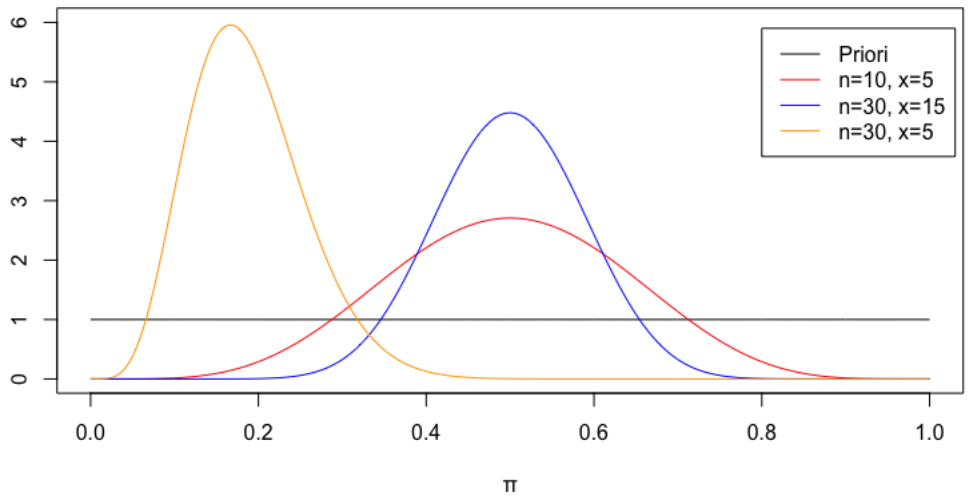
Dann erhalten wir nach der Beobachtung, lateinisch *a posteriori*, mit dem Satz von Bayes

$$\begin{aligned} f(\pi|x) &= \frac{f(x|\pi)f(\pi)}{\int f(x|\tilde{\pi})f(\tilde{\pi})d\tilde{\pi}} = \frac{\binom{n}{x}\pi^x(1-\pi)^{n-x}\cdot 1}{f(x)} \\ &= C(x)\cdot\pi^x(1-\pi)^{n-x} = C(x)\cdot\pi^{(x+1)-1}(1-\pi)^{(n-x+1)-1} \qquad (1) \end{aligned}$$

Dabei ist $C(x)$ eine Konstante bezüglich π (hängt nicht von π, nur von x ab). (1) sieht bis auf die Konstante aus wie die Dichte der Beta$(x+1, n-x+1)$-Verteilung. Wir sagen: $\pi^{(x+1)-1}\pi^{(n-x+1)-1}$ ist der "Kern" der Beta-Verteilung.

- Wir haben Vorwissen über die Wahrscheinlichkeit π in Form einer **Priori-Verteilung** bzw. Priori -Dichte formuliert.
- Nach der Beobachtung x wissen wir mehr über π; wir haben die **Posteriori-Verteilung** bzw. Posteriori-Dichte erhalten.
- Bayes-Prinzip: Alle Schlüsse werden aus der Posteriori-Verteilung gezogen.
- Zur Berechnung der Posteriori brauchen wir zudem das Beobachtungsmodell bzw. **Datendichte** $f(x|\pi)$ (auch als Likelihood bezeichnet)

- und die **Normalisierungskonstante** (auch marginale Likelihood), die wir hier nicht explizit berechnen mussten.



Sei $n = 30$ und $x = 5$. Wie lautet dann unser Schätzer für π?

- Posteriori-Erwartungswert: Nach Beobachtung von x, welchen Wert erwarten wir für π?

$$E(\pi) = \frac{x+1}{n+1} = \frac{6}{31} \approx 0.194$$

- Posteriori-Modus: Welcher Wert von π maximiert die Posteriori?

$$\hat{\pi}_{\text{MAP}} = \frac{x}{n} = \frac{5}{30} \approx 0.167$$

- Posteriori-Median: Welcher Wert hat die mittlere Wahrscheinlichkeit?

$$\hat{\pi}_{\text{med}} \approx 0.181$$

- Kredibilitätsintervall: In welchem Intervall liegt π mit Wahrscheinlichkeit (z.B.) 0.95?

$$\text{HPD} = [0.0644, 0.3220]$$

2.4 Bayesianische Inferenz

- Wir beobachten n Daten x_i, die aus einem Zufallsprozess entstanden sind
- **Annahme**: x_i ist die Realisierung einer Zufallsvariable X_i
- **Annahme**: X_i hat Verteilung F mit Dichte $f(x)$
- **Parametrische Annahme**: Die Dichte ist bekannt bis auf einen Parameter θ: $f(x|\theta)$
- θ ist unbekannt und die Information über θ lässt sich in Form einer Wahrscheinlichkeitsverteilung mit Dichte darstellen
- Vor der Beobachtung (*a priori*) ist unsere Information $p(\theta)$
- Durch Beobachtung erhalten wir mehr Information, ausgedrückt durch die *a posteriori*-Verteilung $\theta|x$
- Sowohl x als auch θ können mehrdimensional sein!
- Die Dichte der Posteriori-Verteilung erhalten wir über die Bayes-Formel

$$\boxed{f(\theta|x) = \frac{f(x|\theta) \cdot f(\theta)}{\int f(x|\tilde{\theta}) f(\tilde{\theta}) d\tilde{\theta}}}$$

- Bayes-Prinzip: Alle Schlüsse werden **nur** aus der Posteriori-Verteilung gezogen
- Festlegung des statistischen Modells für x, Datendichte (Likelihood) $f(x|\theta)$
- Festlegung des *a priori*-Wissens über θ, Priori-Dichte $p(\theta)$
- Berechnung der Posteriori $p(\theta|x)$ (insbesondere Normalisierungskonstante)

"Bayes-Prinzip: Alle Schlüsse werden **nur** aus der Posteriori-Verteilung gezogen" – Was machen wir nun mit der Posteriori?

- Grundsätzlich: Komplette Posteriori wichtig (Darstellung bei hochdimensionalem Parameter θ aber schwierig)
- Punktschätzer (Posterior-Erwartungswert, Maximum-a-Posteriori-Schätzer, Posteriori-Modus)
- Intervallschätzer
- Testen
- Modellvergleich (d.h. verschiedene Annahmen für $f(x)$!)

3 Prioris

Für eindimensionale Parameter:

- subjektive Prioris
- nicht-informative Prioris
- konjugierte Prioris

Bei mehrdimensionalen Parametern benutzt man die Grundkonzepte und erweitert sie.

3.1 Subjektive Priori

Beispiel 3.1 (Subjektive Prioris, Dupuis 1995)**.** *In einem biologischen Experiment werden Echsen markiert und später nochmals eingefangen. Unbekannter Parameter ist die Einfangwahrscheinlichkeit p_t. Die Biologen nennen folgende Zahlen:*

Zeitpunkt	*2*	*3*	*4*	*5*	*6*
Mittelwert	*0.3*	*0.4*	*0.5*	*0.2*	*0.2*
95% Kred.int.	*[0.1,0.5]*	*[0.2,0.6]*	*[0.3,0.7]*	*[0.05,0.4]*	*[0.05,0.4]*

- Subjektive Prioris basieren auf Vorwissen des Experten (oder Statistikers) über die Verteilung des Parameters
- Problem: Das Vorwissen liegt selten in Form einer Dichte vor
- Für diskrete Parameter noch relativ einfach: Abschätzung der einzelnen Wahrscheinlichkeiten aus bisherigen Experimenten (bei Reliabilität!)
- Übertragung auf stetige Parameter: Histogramm
- Die Art (Form) der Priori ist oft bekannt (oder die genaue Form nicht relevant), aber die Parameter der Verteilung nicht
- Dann kann man die Parameter z.B. aus dem Mittelwert und der Varianz schätzen
- Möglichst robuste Statistiken verwenden (Median, Quantile)
- Soweit möglich Überprüfung der Verteilung (z.B. Schiefe bei der Normalverteilung)

In Beispiel 3.1: Mögliche Priori: Betaverteilung; für $\theta \sim \text{Beta}(a, b)$ gilt $E(\theta) = \frac{a}{a+b}$, über (konsistentes) Kredibilitätsintervall sind a und b zu ermitteln .

Definition 3.1. *Die **Entropie** ist ein Maß für die (Shannon-)Information einer Dichte*

$$H(p) = -\sum p(\theta_i)\log(p(\theta_i))$$

- Will man eine Priori mit möglichst wenig Information, so maximiert man die Entropie bezüglich der Dichte.
- Sind Momente $E(g_k(X))$ bekannt, dann lässt sich eine MEP in der Form

$$p_{\text{MEP}} = c\exp\left(\sum_{k=1}^{K}\lambda_k g_k(x)\right)$$

finden, so dass die Momente erhalten bleiben.

Beispiel 3.2 (Maximum Entropy Priori)**.**

3.2 Konjugierte Priori

Definition 3.2 (Konjugierte Verteilungen)**.** *Eine Familie $\mathcal{F}$ von Verteilungen auf Θ heißt **konjugiert**, zu einer Dichte $f(x|\theta)$, wenn für jede Priori $p(\theta)$ auf $\mathcal{F}$ die Posteriori $p(\theta|x)$ ebenfalls zu* F *gehört*

Vergleiche Bsp. 2.3.

Beispiel 3.3. *Normalverteilung bei bekannter Varianz.*

Beispiel 3.4. *Seien $X_1,\ldots,X_n \sim X$ unabhängige Poisson-verteilte Zufallszahlen mit unbekanntem Parameter λ. Die Datendichte ist dann*

$$f(\mathbf{x}|\lambda) = \prod_{i=1}^{n}\left[\frac{\lambda_i^x}{x_i!}\exp(-\lambda)\right] = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n(x_i!)}\exp\left(-n\lambda\right)$$

Wir nehmen für λ a priori eine Gammaverteilung an:

$$p(\lambda) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda).$$

Dann ist die Posteriori-Dichte:

$$\begin{aligned} p(\lambda|\mathbf{x}) &= \frac{f(\mathbf{x}|\lambda)p(\lambda)}{\int f(\mathbf{x}|\tilde{\lambda})p(\tilde{\lambda})d\tilde{\lambda}} \\ &= C(\mathbf{x})\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n(x_i!)}\exp\left(-n\lambda\right)\frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda) \\ &= \tilde{C}(\mathbf{x})\lambda^{\left(a+\sum_{i=1}^n x_i\right)-1}\exp(-(b+n)\lambda) \end{aligned}$$

Das heißt, $\lambda|\mathbf{x}$ ist Gammaverteilt mit Parametern $a+\sum_{i=1}^n x_i$ und $b+n$.

Definition 3.3. *Die Exponentialfamilie ist eine Familie von Verteilungen mit ähnlicher ("schöner") Form*

$$f(x|\theta) = h(x)\exp(b(\theta) + \nu(\theta) \cdot T(x))$$

- $\nu(\theta)$ *ist der natürliche (kanonische) Parameter*
- $T(x)$ *ist eine suffiziente Statistik*

Analoge Definition für mehrdimensionale θ.

Satz 3.1. *Gegeben sei eine Datendichte aus der Exponentialfamilie mit natürlichem Parameter* θ

$$f(x|\theta) = h(x)\exp\left(b(\theta) + \theta \cdot T(x)\right).$$

Dann gibt es (unter Bedingungen) eine konjugierte Priori mit Dichte

$$p(\theta|\mu,\lambda) = C(\mu,\lambda)\exp\left(\lambda b(\theta) + \theta \cdot \mu\right)$$

und die Posteriori hat die aufdatierten Parameter $\mu + T(x)$ *und* $\lambda + 1$.

Datendichte	Priori	Posteriori-Parameter
Binomial	Beta	$a + x, b + n - x$
Poisson	Gamma	$a + x, b + n$
Negative Binomial	Beta	$a + \sum x_i, b + rn$
Geometrisch	Beta	$a + n, b + \sum x_i$
Multinomial	Dirichlet	$\alpha + (c_1, \ldots, c_k)$

Datendichte	Priori	Posteriori-Parameter
Normal (σ^2 bekannt)	Normal	$\tilde{m}/s, 1/s$
Normal (μ bekannt)	Invers Gamma	$a + 0.5 \cdot n,$ $b + 0.5 \cdot \sum(x_i - \mu)^2$
Log-Normal (σ^2 bekannt)	Normal	$\tilde{m}/s, 1/s$
Log-Normal (μ bekannt)	Invers Gamma	$a + 0.5 \cdot n,$ $b + 0.5 \cdot \sum(\log(x_i) - \mu)^2$
Gleich auf (0,θ)	Pareto	$\max x_i, k + n$
Exponential	Gamma	$a + n, b + \sum x_i$

3.3 Nichtinformative Priori

Bei subjektiven und konjugierten Prioris enthält die Priori Information, die in die Posteriori eingeht. Laplace benutzte nach dem *Prinzip vom unzureichenden Grund*

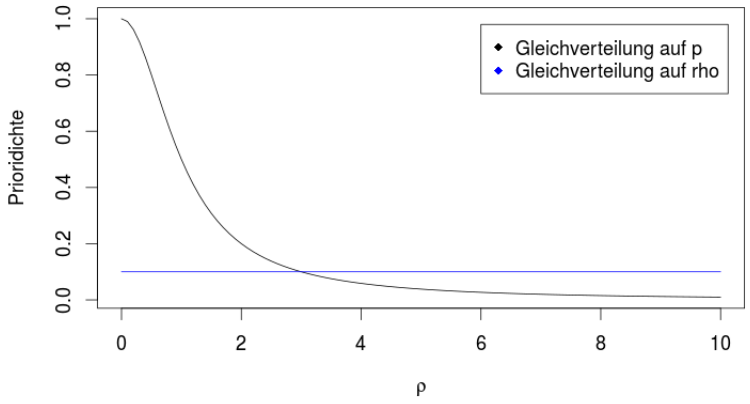
als Erster die Gleichverteilung für Fälle, in denen keine Information vorliegt. Es gilt dann

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{\int f(x|\theta)p(\theta)d\theta} \propto f(x|\theta).$$

Die Posteriori entspricht dann (bis auf die Normalisierungskonstante) der Likelihood. Aber:

- Gleichverteilung auf $(-\infty, +\infty)$ ist keine (propere) Verteilung
- Die Gleichverteilung ist nicht invariant gegenüber Reparametrisierungen

Beispiel 3.5. *Sei p ein unbekannter Wahrscheinlichkeitsparameter mit Gleichverteilung auf $[0, 1]$ als Priori. Dann hat die Priori des odds $\rho = \frac{p}{1-p}$ die Dichte $p(\rho) = \frac{1}{1+\rho^2}$. Dies ergibt sich direkt aus dem Transformationssatz für Dichten.*



Sei $f(X|\theta)$ die (zweimal differenzierbare) Dichte der Zufallsvariable X gegeben dem (skalaren) Parameter θ.

Definition 3.4 (Beobachtete Information)**.**

$$J(\theta) = -\left(\frac{d\log(f(X|\theta))}{d\theta}\right)^2$$

Die beobachtete Information beschreibt, wieviel Information über den Parameter in der Stichprobe vorliegt.

Definition 3.5 (Fisher-Information)**.**

$$I(\theta) = E[J(\theta)] = -E\left[\left(\frac{d\log(f(X|\theta))}{d\theta}\right)^2\right]$$

Die Fisher-Information beschreibt, wieviel Information über den Parameter man in der Stichprobe erwarten kann. Für erwartungstreue Schätzer gilt $\text{Var}(\hat{\theta}) \geq I(\theta)^{-1}$ (Cramer-Rao-Schranke).

Satz 3.2. *Jeffreys Priori mit der Dichte*

$$p^*(\theta) \propto I^{1/2}(\theta)$$

ist invariant gegenüber Reparametrisierungen.

Beweis: Sei $\phi(\theta)$ eine Reparametrisierung des Parameters θ. Dann ist die Dichte von ϕ nach dem Transformationssatz für Dichten:

$$\begin{aligned} p(\phi) &= p(\theta)\left|\frac{d\theta}{d\phi}\right| \\ &\propto \sqrt{-E\left[\left(\frac{d\log(f(X|\theta))}{d\theta}\right)^2\right]\left(\frac{d\theta}{d\phi}\right)^2} \\ &= \sqrt{-E\left[\left(\frac{d\log(f(X|\theta))}{d\theta}\frac{d\theta}{d\phi}\right)^2\right]} \\ &= \sqrt{-E\left[\left(\frac{d\log(f(X|\theta(\phi)))}{d\phi}\right)^2\right]} = \sqrt{I(\phi)} \end{aligned}$$

Bernardo (1979) schlägt vor: Falls $\theta = (\theta_1, \theta_2)$, θ_1 ist der interessierende Parameter und θ_2 ist *nuisance*, berechne Jeffreys Priori für feste θ_1: $p(\theta_2|\theta_1)$ und benutze

$$\tilde{f}(x|\theta_1) = \int f(x|\theta_1, \theta_2)p(\theta_2|\theta_1)d\theta_2$$

um Jeffreys Priori für θ_1 zu erhalten. Die Dichte der gemeinsame Priori ist dann

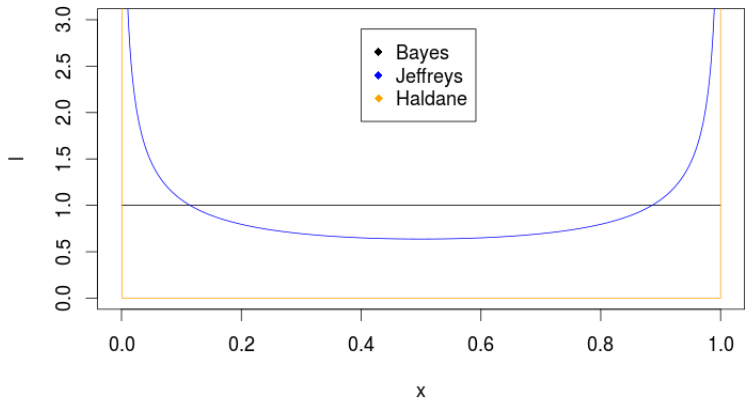
$$p(\theta_1, \theta_2) = p(\theta_2|\theta_1)p(\theta_1).$$

Die **Referenz-Priori** ist die am wenigsten informative Priori, d.h. sie maximimiert den Informationsgewinn durch die Beobachtung.

- Die nichtinformative Priori kann auch die konjugierte sein
- Viele flache (nichtinformative) Prioris sind nicht proper ($\int p(\theta) = \infty$). Dies ist unproblematisch, solange die Posteriori proper ist (Bsp. 3.6)
- Gerne benutzt man auch "relativ"flache Prioris, z.B. $\theta \sim N(0, 10^{12})$

- Für konjugierte und subjektive Prioris empfiehlt sich eine Sensitivitätsanalyse

Beispiel 3.6 (Binomialverteilung mit Betapriori)**.**



3.4 Regularisierungspriori

Oft setzt man Prioris bewusst ein, um den Parameterraum (stochastisch) zu verkleinern. Beispiele:

- A Priori sind Parameter identisch
- Regression mit mehr Kovariablen als Beobachtungen → Variablenselektion
- "Glatte" zeitliche oder räumliche Trends

Beispiel 3.7. *Gegeben seien Beobachtungen einer normalverteilten Zufallsvariablen X an zwei Zeitpunkten mit Erwartungswert μ_1 und μ_2. Varianz σ^2 sei bekannt. A priori gehen wir davon aus, dass μ_1 und μ_2 gleich sind. Mögliche Prioris:*

- $\mu_1 \sim N(\mu_0, \tau^2), \mu_2 \sim N(\mu_0, \tau^2)$
- $\mu_1 \sim N(\mu_2, \tau^2)$

  identisch zu: $\mu_2 \sim N(\mu_1, \tau^2)$

  identisch zu: $(\mu_1 - \mu_2) \sim N(0, \tau^2)$
- $(\mu_1 - \mu_2) \sim \nu\delta_0 + (1 - \nu)N(0, \tau^2)$ *(Spike and Slab)*

alternativ: $(\mu_1 - \mu_2) \sim \nu N(0, \tau_0^2) + (1 - \nu) N(0, \tau^2)$ *mit* τ_0^2 *sehr groß*

Gegeben sei ein lineares Regressionsmodell mit Zielvariable y und P Kovariablen $x_1, \ldots, x_P$:

$$y_i = \beta_0 + \sum_{p=1}^{P} \beta_p x_{ip}$$

- Ridge-Priori: $\beta_p \sim N(0, \tau^2)$
- Lasso-Priori: $\beta_p| \sim$ *Laplace* identisch mit: $\beta_p|\tau_p^2 \sim N(0, \tau_p^2), \tau_p^2 \sim Exp(\lambda)$

Zusammenhang mit penalisierter Likelihood

- Bayesianische Regularisierung hat Ähnlichkeiten zum frequentistischen Ansatz der Penalisierung.
- Penalisierte Log-Likelihood: $\ell_p en(\theta) = \ell(\theta|x) + pen(\theta)$
- Log-Posteriori: $\log(p(\theta|x) = \log(f(\theta|x)) + \log(p(\theta)) + C$
- Regularisierte Bayes-Ansätze und penalisierte Log-Likelihood ergeben identische Punktschätzer (mit MAP)
- Penalisierte Log-Likelihood-Ansätze fehlen asymptotische Aussagen, damit keine theoretisch gut fundierten Intervallschätzer und Tests
- Bayesianischer Ridge und Bayesianischer Lasso selektieren keine Variablen wegen *Model Averaging*

4 Bayesianisches Schätzen und Testen

4.1 Punktschätzer

Definition 4.1 (MAP-Schätzer)**.** *Der* ***Maximum-A-Posteriori-Schätzer*** *(oder Posteriori-Modus) ist derjenige Wert, der die Posteriori maximiert.*

$$\hat{\theta}_{MAP} = argmax\ p(\theta|x)$$

- Für flache Prioris ($p(\theta) \propto$ const.) entspricht der MAP-Schätzer dem Maximum-Likelihood-Schätzer
- Für andere Prioris entspricht der MAP-Schätzer einen penalisierten ML-Schätzer oder eines restringierten ML-Schätzers

- Eigenschaften des (P-)ML-Schätzers gelten auch für den MAP-Schätzer

Beispiel 4.1 (Schätzer im Beta-Binomial-Modell)**.**

Definition 4.2 (Posteriori-Erwartungswert)**.** *Der erwartete Wert des Parameters gegeben der Beobachtung*

$$\hat{\theta}_{PE} = E(\theta|x)$$

ist der ***Posteriori-Erwartungswert-Schätzer****.*

- Der Posteriori-Erwartungswert ergibt sich aus der Bayesianischen Entscheidungstheorie, falls man die (übliche) quadratische Verlustfunktion benutzt (siehe dort)
- Optimalitätseigenschaften dort
- Für konjugierte Prioris kann der Posteriori-Erwartungswert i.d.R. einfach durch Aufdatierung berechnet werden

Fortsetzung Bsp. 4.1

Definition 4.3 (Posteriori-Median)**.** *Der Median der Posteriori-Verteilung des Parameters gegeben der Beobachtung*

$$\hat{\theta}_{med} = Median(\theta|x)$$

ist der ***Posteriori-Median-Schätzer****.*

- Ergibt sich aus der Bayesianischen Entscheidungstheorie, falls man die absolute Verlustfunktion benutzt
- Robuster gegenüber Ausreißern

Fortsetzung Bsp. 4.1

Vergleich der Punktschätzer

- Der Posteriori-Erwartungswert ist nicht unbedingt unverzerrt
- MAP und PE sind nicht invariant bezüglich streng monotoner Transformationen
- Für (fast) symmetrische Posterioris sind $\hat{\theta}_{MAP}, \hat{\theta}_{PE}$ und $\hat{\theta}_{med}$ (fast) identisch

4.2 Intervallschätzer

Aus der Posteriori-Verteilung lässt sich ein Intervall $[\theta_u, \theta_o]$ bestimmen, das den Parameter θ mit Wahrscheinlichkeit $(1-\alpha)$ enthält.

Definition 4.4. *Ein Intervall $I = [\theta_u, \theta_o]$, für das gilt*

$$\int_{\theta_u}^{\theta_o} p(\theta|x)d\theta = 1 - \alpha$$

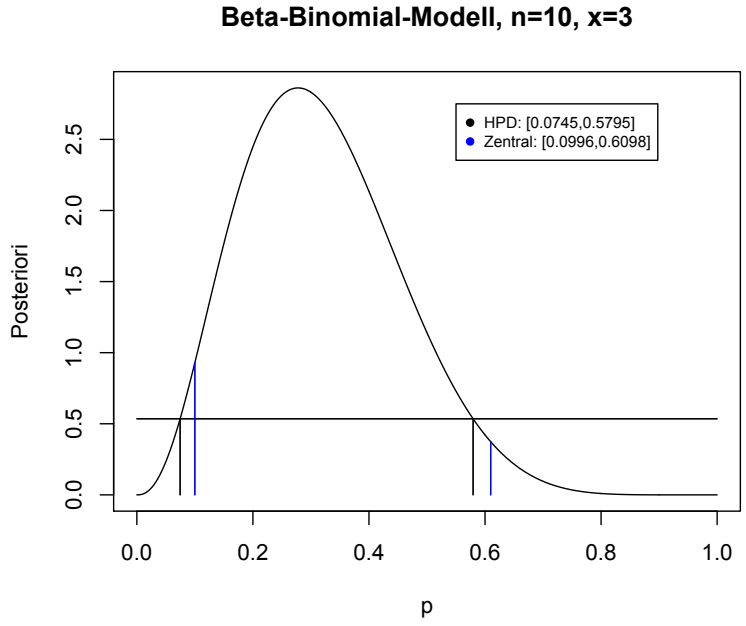
*nennt man $(1-\alpha)$-**Kredibilitätsintervall**.*

Kredibilitätsintervalle sind nicht eindeutig.

Definition 4.5. *Ein $(1-\alpha)$-Kredibilitätsintervall H heisst **Highest Posteriori Density Interval (HPD-Intervall)**, wenn für alle $\theta \in H$ und alle $\theta^* \notin H$ gilt:*

$$p(\theta|x) \geq p(\theta^*|x)$$

Beispiel 4.2 (Kredibilitätsintervalle Betabinomialmodell)**.**



Sei $[a, b]$ ein Bayesianischer a%-Intervallschätzer für θ. Dann gilt:

Die Wahrscheinlichkeit, dass θ zwischen a und b liegt, ist α%.

Vergleiche dazu Interpretation des frequentistischen Konfidenzintervalls:

> Das Konfidenzintervall schließt bei unendlicher Wiederholung des Zufallsexperiments in $\alpha\%$ der Fälle den wahren Parameters ein.

Beide Aussagen gelten natürlich gegeben den beobachteten Daten.

- $(1-\alpha)$ HPD-Bereiche haben minimale Länge unter allen $(1-\alpha)$-Kredibilitätsintervallen
- HPD-Bereiche müssen keine Intervalle sein
- HPD-Bereiche sind nicht invariant gegenüber streng monotonen Transformationen
- Gleichendige Intervalle (Posteriori-Quantile) sind invariant gegenüber streng monotonen Transformationen
- Bei konjugierten Prioris lassen sich Intervalle relativ einfach über Posteriori-Varianz bzw. -Standardabweichung berechnen
- Über die Dualität von Intervallschätzern und Tests lässt sich mit Bayesianischen Intervallen auch testen

4.3 Optimalität der Schätzer*

Im Folgenden formulieren wir "Schätzen eines Parameters" als Entscheidungsproblem. Ziel ist es, den optimalen (Punkt- bzw. Intervall-)Schätzer zu finden.

Ziel der Entscheidungstheorie ist es,die optimale Entscheidung zu finden. Dafür definieren wir

Definition 4.6 (Entscheidungsfunktion)**.** *Eine Entscheidungsfunktion ist eine Abbildung vom Stichprobenraum $\mathcal{X}$ in den Entscheidungsraum D*

$$d : \mathcal{X} \to D; x \mapsto d(x)$$

Oft ist $D = \Theta$.

Definition 4.7 (Verlustfunktion – loss function)**.** *Die Verlustfunktion ordnet jeder Entscheidung einen Verlust zu*

$$L : D \times \Theta \to \mathbb{R}; (d, \theta) \mapsto L(d, \theta)$$

Beispiel 4.3 (Sankt Petersburg Paradoxon*)**.**

Beispiel 4.4 (Schätzen als Entscheidungsraum)**.** *Sei* $X_i \overset{\text{iid}}{\sim} X \sim N(\mu, \sigma^2)$ *mit* σ^2 *bekannt und* $\theta = \mu$ *zu schätzen. Der Entscheidungsraum* D *ist der Raum der möglichen Schätzungen* $d = \hat{\mu} \in \mathbb{R}$*. Üblich ist die quadratische Verlustfunktion* $L(d(x), \mu) = (d(x) - \mu)^2$*. Robuster ist der absolute Verlust* $L(d(x) - \mu) = |d(x) - \mu|$*.*

Problem der statistischen Entscheidungstheorie: Der Verlust hängt von der zufälligen Beobachtung x und vom unbekannten Parameter θ. **Frequentistischer Ansatz:** Wir betrachten den erwarteten (mittleren) Verlust

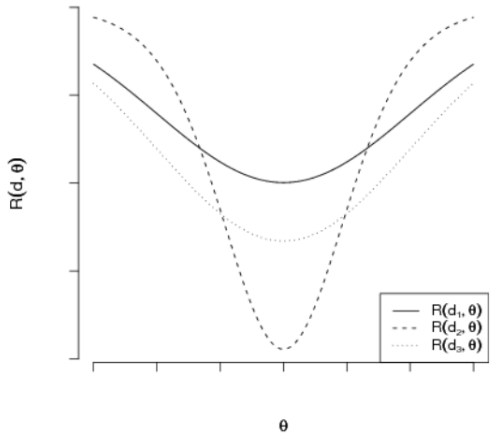
Definition 4.8 (Risiko)**.** *Die Risikofunktion ist der mittlere Verlust bei Entscheidungsfunktion* d *für einen wahren Parameter* θ

$$R(d, \theta) = E\left[L(d(X), \theta)\right] = \int_{\mathcal{X}} L(d(x), \theta) f(x|\theta) dx$$

Das Risiko ist abhängig vom unbekannten Parameter θ. Lösungsansätze:

- Beschränkung auf **zulässige Entscheidungen**. d ist zulässig, wenn es kein d^* gibt, für das gilt $R(d^*, \theta) < R(d, \theta)$ für alle θ
- **Minimax-Entscheidung**: Suche die Entscheidung, die das Maximum der Risikofunktion minimiert

Beispiel 4.5 (*)**.**



Bayesianischer Ansatz: Betrachte den a posteriori erwarteten Verlust

Definition 4.9. *Der a posteriori erwartete Verlust ist*

$$r(d, p|x) = E\left[L(d(X), \theta)|x\right] = \int_{\Theta} L(d(x), \theta) p(\theta|x) d\theta$$

Es gilt:

$$r(d, p|x) = E\left[R(d, \theta)\right] = \int_{\Theta} \int_{\mathcal{X}} L(d(x), \theta) f(x|\theta) dx \, p(\theta) d\theta$$

Definition 4.10. *Für eine Verlustfunktion L und eine Priori-Verteilung p ist jede Entscheidung d^*, welche das den a posteriori erwarteten Verlust $r(d, p|x)$ minimiert* ***Bayes-optimal****. Der Wert $r^*(p) = r(d^*, p|x)$ heißt dann* ***Bayes-Risiko****.*

Satz 4.1. *Eine Entscheidung d^* ist genau dann Bayes-optimal, wenn d^* für jede Beobachtung $x \in \mathcal{X}$ den a posteriori erwarteten Verlust minimiert.*

- Bayes-optimale Entscheidungen sind immer im frequentistischen Sinn zulässig
- Das Bayes-Risiko ist immer kleiner gleich dem Minimax-Risiko
- Wenn d_0 eine Bayesoptimale Entscheidung ist und $R(d, \theta) \leq r^*(p_0)$ für alle θ im Träger von p_0, dann ist d_0 die Minimax-Entscheidung und p_0 die ungünstigste Prioriverteilung

Für das statistische Entscheidungsproblem $d = \hat{\theta}$ gilt

- Bei quadratischer Verlustfunktion ist der Posteriori-Erwartungswert Bayes-optimal
- Bei absoluter Verlustfunktion ist der Posteriori-Median Bayes-optimal
- Bei 0/1-Verlustfunktion ist (für $\epsilon \to 0$) der Posteriori-Modus Bayes-optimal
- Bei 0/1-Verlustfunktion und flacher Priori ($p(\theta) \propto 1$) ist der ML-Schätzer Bayes-optimal

$$L_\epsilon(d, \theta) = \begin{cases} 1, & \text{falls} |d - \theta| \geq \epsilon, \\ 0, & \text{falls} |d - \theta| < \epsilon. \end{cases}$$

4.4 Bayes-Tests*

Gegeben sei eine Nullhypothese

$$H_0 : \theta \in \Theta_0$$

und für einen Test ϕ die a_0–a_1-gewichtete 0/1-Verlustfunktion

$$L(\theta|\phi) = \begin{cases} 0 & \text{für } \phi = \mathrm{I}_{\Theta_0}(\theta) \\ a_0 & \text{für } \theta \in \Theta_0 \text{ und } \phi = 0 \\ a_1 & \text{für } \theta \notin \Theta_0 \text{ und } \phi = 1 \end{cases}$$

Dann ist der Bayestest (mit der Priori p)

$$\phi_p(x) = \begin{cases} 1 & \text{für } P(\theta \in \Theta_0|x) > \frac{a_1}{a_0+a_1} \\ 0 & \text{sonst} \end{cases}$$

Der Bayestest hängt von der Priori, den Daten und vom Akzeptanzlevel $a_1/(a_0 + a_1)$ ab.

- Problematisch sind zweiseitige Tests bei reelwertigen Parametern, also $H_0 : \theta = \theta_0$.
- In der Regel ist $P(\theta_0) = 0$ und damit auch $P(\theta_0|x) = 0$.
- Hier ist also eine modifizierte Priori notwendig, die zusätzliche Dirac-Masse auf die Nullhypothese legt.
- Dies kann jedoch zu anderen Problemen führen (Lindley's Paradoxon)

Praxis-Ansatz: Benutze die Dualität von Tests und Intervallen wie im Frequentistischen. Betrachte also die Frage: Liegt θ_0 im Kredibilitätsintervall?

Beispiel 4.6 (Testen als Entscheidungsproblem)**.** *Sei $x \sim N(\mu, \sigma^2)$, $\mu \sim N(\mu_0, \nu_0^2)$, σ^2 bekannt und $H_0 : \mu < 0$. Die Posteriori ist $\mu \sim N(\hat{\mu}_{PE}, \nu^2)$ mit*

$$\hat{\mu}_{PE} = \frac{\sigma^2 + \mu_0 + \nu_0^2 x}{\sigma^2 + \nu_0^2}, \nu^2 = \frac{\sigma^2 \nu_0^2}{\sigma^2 + \nu_0^2}$$

Damit ist

$$P(H_0|x) = P(\mu < 0|x) = p\left(\frac{\mu - \hat{\mu}_{PE}}{\nu^2} < \frac{-\hat{\mu}_{PE}}{\nu^2}\right) = \Phi\left(-\frac{\hat{\mu}_{PE}}{\nu^2}\right)$$

Also, akzeptiere H_0, falls $\hat{\mu}_{PE} < -z_{a_1/(a_0+a_1)}\nu^2$.

In einer Stadt wurden $n = 98.451$ Babys geboren, davon $k = 49.581$ Jungen und $n-k = 48.870$ Mädchen. Testen Sie die Nullhypothese H_0: "Die Wahrscheinlichkeit dass ein Neugeborenes ein Junge ist, ist $p = 0.5$."

Frequentistischer Ansatz: Die Binomialverteilung kann über die Normalverteilung approximiert werden. Dann ist die Wahrscheinlichkeit für die Beobachtung unter der Nullhypothese:

$$\begin{aligned} P(X \geq x \mid \mu = np) &= \sum_{x=k}^{n} \frac{1}{\sqrt{2\pi(np(1-p))}} exp\left(-\frac{(u-np)^2}{2np(1-p)}\right) du \\ &\approx 0.0117. \end{aligned}$$

Wir lehnen die Nullhypothese auf dem 5%-Niveau ab.

Bayesianischer Ansatz: Ohne weitere Information ist die $P(H_0) = P(H_1) = 0.5$. Uns interessiert die Posteriori-Wahrscheinlichkeit von H_0:

$$P(H_0 \mid k) = \frac{P(k \mid H_0)\pi(H_0)}{P(k \mid H_0)\pi(H_0) + P(k \mid H_1)\pi(H_1)}.$$

Es gilt:

$$\begin{aligned} P(k \mid H_0) &= \binom{n}{k}(p)^k(1-p)^{n-k} \approx 1.95 \times 10^{-4} \\ P(k \mid H_1) &= \int_0^1 \binom{n}{k} p^k(1-p)^{n-k} dp \approx 1.02 \times 10^{-5} \end{aligned}$$

Einsetzen ergibt: $P(H_0 \mid k) \approx 0.95$, wir bevorzugen also die Nullhypothese.

Woher kommen die Unterschiede?

- Der frequentistische Ansatz basiert nur auf der Nullhypothese H_0.
- Der Bayesianische Ansatz berücksichtigt die Gegenhypothese H_1, mit ziemlich hohen Priori-Wahrscheinlichkeiten für relativ unwahrscheinliches Fälle.
- Die Priori hat hier die Form einer Mischung aus einer Dirac-Verteilung auf 0.5 und einer Gleich-Verteilung auf [0,1]. D.h. unter H_1 wird hier eine weitere Priori-Verteilung für p verwendet.
- Mit einer uninformativen Priori kommt der Bayesianische Ansatz zum selben Ergebnis wie der frequentistische.

5 Bayesianische Modellierung

5.1 Normalverteilungsmodel

Gegeben seien n unabhängig normalverteilte Beobachtungen

$$X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n.$$

Die gemeinsame Datendichte lautet

$$\begin{aligned} f(x|\theta) &= \prod (f(x_i|\theta)) \\ &= \left(\frac{1}{\sigma\sqrt{(2\pi)}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right) \end{aligned}$$

σ^2 bekannt Konjugierte Priori $\mu \sim N(\mu_0, \sigma_0^2)$. Damit ist die Posteriori

$$\begin{aligned} \mu|(x_1, \ldots, x_n) &\sim N(m/s, 1/s) \\ m &= \frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n x_i}{\sigma^2} \\ s &= \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \end{aligned}$$

Jeffreys Priori: $p(\mu) \propto$ const. entspricht $\sigma_0^2 \to \infty$.

μ bekannt Konjugierte Priori: $\sigma^2 \sim IG(a, b)$ führt zur Posteriori

$$\sigma^2|(x_1, \ldots, x_n) \sim IG\left(a + n/2, b + 0.5 \sum_{i=1}^n (x_i - \mu)^2\right)$$

Jeffreys Priori $p(\sigma^2) \propto \sigma^{-1}$ entspricht "IG(0,0)".

Alternativ lässt sich die Inferenz für die Präzision gleich Inverse der Varianz betreiben:

$$\tau = \sigma^{-2}.$$

Die konjugierte Priori ist dann die Gamma-Verteilung

$$p(\tau) = \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp -b\tau.$$

Die Posteriori lautet

$$p(\tau|x) \propto \tau^{n/2} \exp\left(-\tau/2 \sum_{i=1}^n (x - \mu)^2\right) \cdot \tau^{a-1} \exp(-b\tau),$$

ist also die $\mathrm{Ga}(a + n/2, b + 0.5 \sum_{i=1}^n (x_i - \mu)^2)$-Verteilung.

Zwei unbekannte Parameter Bei jeweils einem unbekanntem Parameter und unter Benutzung der konjugierten Verteilung kennen wir die Posteriori vollständig. Im Folgenden seien beide Parameter (μ und τ) unbekannt.

Wir wollen die selben Prioris wie oben benutzen und gehen von *a priori*-Unabhängigkeit der Parameter aus:

$$p(\mu, \tau) = p(\mu) \cdot p(\tau)$$

Die Posteriori lautet bis auf Konstanten:

$$\begin{aligned} p(\mu, \tau|x) &\propto \exp\left(-\tau_0/2(\mu - \mu_0)^2\right) \\ &\quad \cdot \tau^{n/2} \exp\left(-\tau/2 \sum_{i=1}^n (x - \mu)^2\right) \cdot \tau^{a-1} \exp(-b\tau) \end{aligned}$$

Wir betrachten die bedingte Posteriori eines Parameters, z.B. $p(\mu|\tau, x)$. Nach der Definition der bedingten Dichte gilt

$$p(\mu|\tau, x) = \frac{p(\mu, \tau|x)}{p(\tau|x)} \propto p(\mu, \tau|x)$$

Hier also: Die bedingte Posteriori-Dichte von μ gegeben τ ist die Normalverteilung. Das ergibt sich automatisch aus dem Normalverteilungsmodell mit bekannter Varianz!

Die bedingte Posteriori hilft uns aber nicht weiter, weil wir den Parameter τ nicht kennen.

Definition 5.1. *(Vollständig bedingte Posterior) Sei $\theta = (\theta_\mathbf{1}, \ldots \theta_\mathbf{p})$. Als vollständig bedingte Posteriori ("full conditional posterior") bezeichnen wir die Verteilung eines Parameters θ_i gegeben allen anderen Parametern θ_{-i} und den Daten x. Es gilt:*

$$p(\theta_i|\theta_{-i}, x) \propto p(\theta|\mathbf{x}).$$

Definition 5.2. *(Semikonjugierte Priori) Eine Familie $\mathcal{F}$ von Verteilungen auf Θ heißt **semikonjugiert** wenn für jede Priori $p(\theta)$ auf $\mathcal{F}$ die vollständig bedingte Posteriori $p(\theta_i|\theta_{-i}, x)$ ebenfalls zu $\mathcal{F}$ gehört.*

Wir betrachten die marginale Posteriori eine Parameters, also z.B. $p(\tau|x)$. Diese erhalten wir durch marginalisieren der gemeinsamen Posteriori

$$p(\tau|x) = \int p(\mu, \tau|x) d\mu.$$

Alternativ kann man folgende Formel ausnutzen

$$p(\tau|x) = \frac{p(\mu, \tau|x)}{p(\mu|\tau, x)}$$

Interessiert uns in mehrparametrischen Modellen nur ein Parameter, so ziehen wir die Schlüsse aus der marginalen Posteriori (*"Model averaging"*).

Im Normalverteilungsmodell gilt:

- $\tau|\mathbf{x} \sim Ga$
- $\mu|\mathbf{x} \sim t$

5.2 Regression

Lineare Regression Übliches Lineares Regressionsmodell:

$$\begin{aligned} y_i &= \alpha + \beta x_i + \epsilon_i \\ \mathrm{E}(\epsilon) &= 0 \\ \mathrm{Var}(\epsilon) &= \sigma^2 \\ \mathrm{Cov}(\epsilon_i, \epsilon_j) &= 0 \end{aligned}$$

Bayesianisches lineares Regressionsmodell

$$\begin{aligned} y_i &\sim \mathrm{N}(\alpha + \beta x_i, \sigma^2) \\ \alpha &\sim \mathrm{N}(m_\alpha, v_\alpha^2) \\ \beta &\sim \mathrm{N}(m_\beta, v_\beta^2) \end{aligned}$$

Bei festem σ^2 sind dies die konjugierten Prioris. Wir kennen allerdings σ^2 in der Regel nicht. Wir nehmen zusätzlich an:

$$\sigma^2 \sim IG(a, b)$$

Generalisierte Lineare Regression Das Modell lässt sich relativ einfach auf beliebige Verteilungen verallgemeinern, z.B. ein Poisson-Modell

$$\begin{aligned} y_i &\sim Po(\lambda_i) \\ \log(\lambda_i) &= \alpha + \beta x_i \\ \alpha &\sim \mathrm{N}(m_\alpha, v_\alpha^2) \\ \beta &\sim \mathrm{N}(m_\beta, v_\beta^2) \end{aligned}$$

Die (vollständig bedingten) Posterioris sind jedoch keine Standardverteilungen mehr.

Multivariate Regression

$$y_i \sim \mathrm{N}(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2) \tag{2}$$

$$\sigma^2 \sim IG(a, b) \tag{3}$$

$$\boldsymbol{\beta} \sim \mathrm{N}(\mu_0, \Lambda_0^{-1}) \tag{4}$$

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}) p(\sigma^2) \\
&\propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \\
&\cdot \quad |\Lambda_0|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^{\mathrm{T}} \boldsymbol{\Lambda}_0 (\boldsymbol{\beta} - \boldsymbol{\mu}_0)\right) \\
&\cdot \quad (\sigma^2)^{-(a_0+1)} \exp\left(-\frac{b_0}{\sigma^2}\right)
\end{aligned}
$$

$$
p(\beta|\sigma^2, \mathbf{y}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}(\sigma^2 \mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\Lambda}_0)\boldsymbol{\beta} + (\sigma^2 \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\Lambda}_0)\boldsymbol{\beta})\right)
$$

Damit

$$
\begin{aligned}
\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{x} &\sim N(\boldsymbol{\mu}_n, \sigma^2 \boldsymbol{\Lambda}^{-1}) && (5) \\
\sigma^2|\boldsymbol{\beta}, \mathbf{y}, \mathbf{x} &\sim IG(a_n, b_n) && (6) \\
\boldsymbol{\mu}_n &= (\sigma^2 \mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\Lambda}_0)^{-1}(\boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 + \sigma^2 \mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}) && (7) \\
\boldsymbol{\Lambda}_n &= (\sigma^2 \mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\Lambda}_0) && (8) \\
a_n &= a_0 + \frac{n}{2} && (9) \\
b_n &= b_0 + \frac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 - \boldsymbol{\mu}_n^{\mathrm{T}}\boldsymbol{\Lambda}_n\boldsymbol{\mu}_n) && (10)
\end{aligned}
$$

Über die Kovarianz- oder die Präzisionsmatrix lassen sich Korrelationen zwischen den Kovariableneffekten modellieren. Z.B. ein zeitlich geglätterer Trend.

Beispiel 5.1 (Random Walk Priori)**.** *Gegeben sei eine Zeitreihe y_t mit $t = 1, \ldots, T$. Wir wollen die Zeitreihe glätten. Sei $\mathbf{X} = \mathbf{I}_T$, dann ist obiges Modell gleich*

$$
y_t = \beta_t + \epsilon_t \text{ für } t = 1, \ldots, T
$$

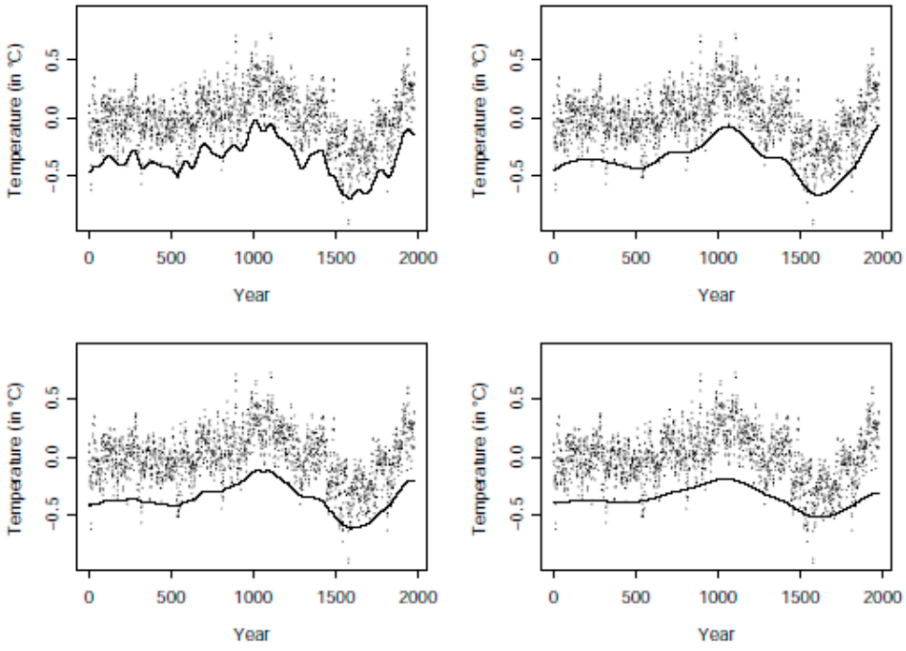
Als Priori für β nehmen wir einen "Random Walk":

$$
\begin{aligned}
\beta_1 &\sim N(0, \tau_0^2) \\
\beta_t &\sim N(\beta_{t-1}, \tau^2)
\end{aligned}
$$

Der Parameter τ steuert die Glattheit der Zeitreihe β_t.

Es lässt sich zeigen (mit $\tau_0 \to \infty$):

$$\Lambda = \tau^{-2} \begin{pmatrix} 1 & -1 & 0 & \cdots & & & 0 \\ -1 & 2 & -1 & 0 & \cdots & & 0 \\ 0 & -1 & 2 & -1 & 0 & \cdots & 0 \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ & & \cdots & 0 & -1 & 2 & -1 \\ & & & \cdots & 0 & -1 & 1 \end{pmatrix}$$



5.3 Hierarchische Modellierung

Level 1: Datenmodell, Definition der Likelihood

Level 2: Priori-Modell der unbekannten Parameter

Level 3: (Hyper-)Prioris der Prioriparameter in Level 2

Kann an sich beliebig erweitert werden, i.d.R. reichen aber drei Level. Inferenz üblicherweise mit MCMC.

Beispiel 5.2 (Räumliches APC-Modell)**.** *Anzahl männliche Magenkrebstote in Westdeutschland*

- *Jahre 1976 - 1990*
- *13 Altersgruppen a 5 Jahre (15-19 bis 85-89)*
- *Geburtskohorten von 1896-1975*
- *30 Regierungsbezirke*

$$
\begin{aligned}
y_{ijt} &\sim \mathrm{B}(n_{ijt}, \pi_{ijt}) && (11)\\
\mathrm{logit}(\pi_{jtl}) &= \xi_{jtl} && (12)\\
&= \mu + \theta_j + \phi_t + \psi_k + \alpha_l + \begin{bmatrix} \delta_{jl} \\ \delta_{kl} \end{bmatrix} + z_{jtl} && (13)
\end{aligned}
$$

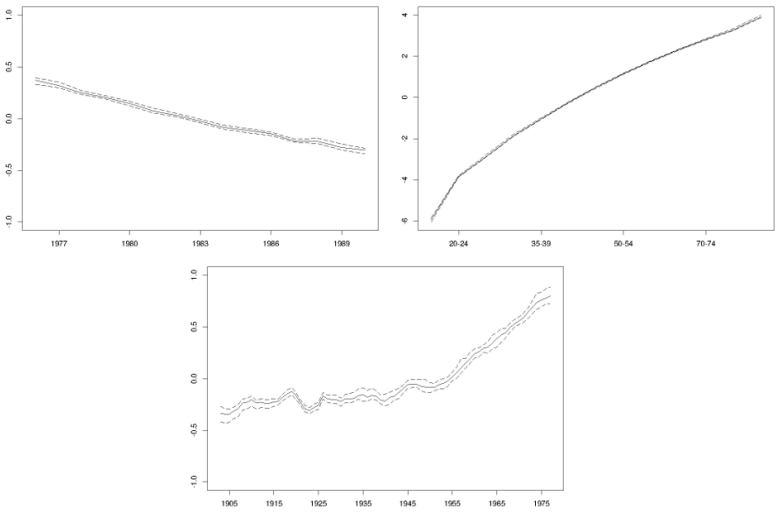
- μ: Intercept (1 Parameter)
- θ_j: Effekt der Altersgruppe j (15 Parameter)
- ϕ_t: Effekt der Periode t (13 Parameter)
- ψ_k: Effekt der Kohorte $k = k(j,t)$ (75 Parameter)
- α_l: Räumlicher Effekt l (30 Parameter)
- δ_{tl}: Interaktion zwischen Perioden- und räumlichen Effekt (390 Parameter)
- z_{jtl}: zufälliger Effekt (Überdispersion, 5850 Parameter)
- Random Walk Priori für APC-Effekte mit Glättungsparameter (Präzision)
- 2D-Random-Walk (Gauss-Markovzufallsfeld) für räumlichen Effekt mit Glättungsparameter
- Interaktion: Unabhängiger Random Walk pro Region oder 3D-GMZF mit Glättungsparameter
- iid normalverteilter zufälliger Effekt mit unbekannter Varianz/Präzision

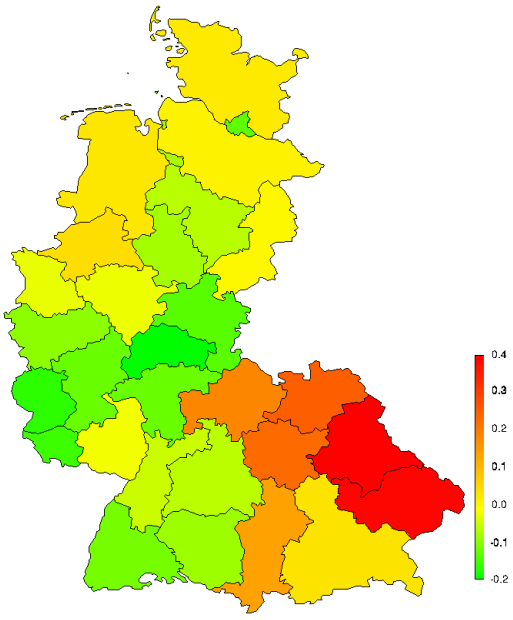
Alle Prioris haben die Form

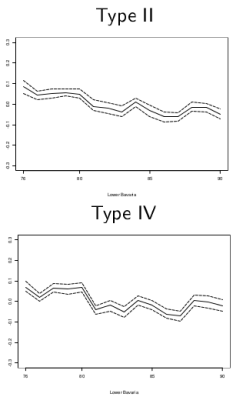
$$p(\theta|\kappa) \propto \exp\left(-\frac{\kappa}{2}\,\theta^T \mathbf{K}_\theta \theta\right) \qquad (14)$$

Gamma-Prioris auf alle Präzisionsparameter. Hyperprioriparameter können Ergebnis beeinflußen → Sensitivitätsanalyse.

- Sehr komplexe Posteriori
- Marginale Posterioris nicht geschlossen herleitbar
- Bedingte Posterioris leicht anzugeben und (mit Trick) Standardverteilungen





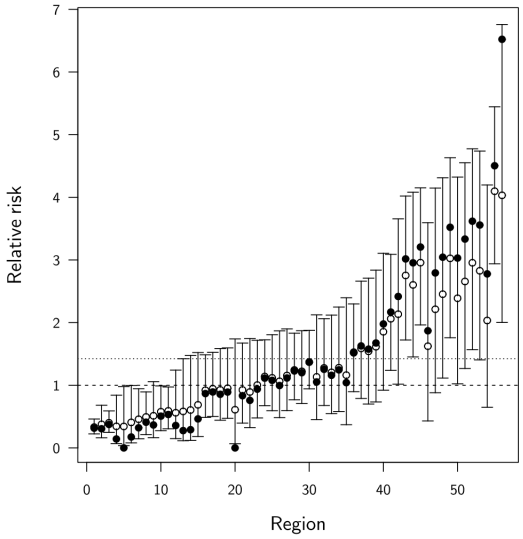


5.4 Empirischer Bayes*

- Bei den hierarchischen Modellen kann die Posteriori in der Regel nicht komplett geschlossen angegeben werden.
- Die vollständig bedingten Dichten sind aber meist bekannt.
- Idee des Empirischen Bayes-Schätzers:
  - Schätze die unbekannten Priori-Parameter aus der marginalen Posteriori (marginale Posteriori, Level 3)
  - Setze die Schätzer in die vollständig bedingte Posteriori der restlichen Parameter (Level 2, *Plug-In-Schätzer*)

Beispiel 5.3 (Lippenkrebs in Schottland)**.**

95 % equi-tailed credible intervals for Scottish lip cancer incidence rates λ_i ($i = 1, \ldots, 56$), calculated with an empirical Bayes approach. The *dotted line* marks the MLE $\hat{\alpha}_{\mathrm{ML}}/\hat{\beta}_{\mathrm{ML}} = 1.4240$ of the prior mean. *Open circles* denote the posterior mean estimates of λ_i, while *filled circles* denote the MLEs x_i/e_i.



6 Numerische Verfahren zur Bestimmung der Posteriori

Problem: Um die marginale Posteriori zu bestimmen, müssen wir alle anderen Parameter heraus integrieren. In hochdimensionalen Modellen wird das analytisch kaum möglich sein. Lösungsansätze:

- Numerische Integration

- Approximation der (marginalen) Posteriori
- Simulationsverfahren

6.1 Numerische Integration*

Die Trapezregel basiert auf stückweiser Integration

$$\int_a^b g(x)dx = \sum_{i=0}^{G-1} \int_{x_i}^{x_{i+1}} g(x)dx$$

mit Knoten $x_0 = a < x_1 < \cdots < x_G = b$. Approximation der Teilintegrale durch Fläche des Trapezes in einem Intervall:

$$\frac{1}{2}(x_{i+1} - x_i)(g(x_i) + g(x_i + 1))$$

Mit Intervallbreite $h = (x_{i+1} - x_i)$ ist

$$\int_a^b g(x)dx \approx h \left(\frac{1}{2}g(a) + \sum_{i=1}^{n-1} g(x_I) + \frac{1}{2}g(b) \right)$$

- Wahl von $m + 1$ äqudistanten Stützstellen $x_{i0} = x_i < x_{i1} < \cdots < x_{i,m+1} = x_{i+1}$ in jedem Intervall $[x_i, x_{i+1}]$
- Berechnung der Funktionswerte $g(x_{i,j}$
- Interpolation der $m+1$ Punkte $(x_{i,j}, g(x_{i,j}))$ durch ein Polynom $p_i(x_{ij})$ vom Grad m
- Integration des Näherungspolynoms

$$T_i = \int_{x_i}^{x_{i+1}} g(x)dx \approx \int_{x_i}^{x_{i+1}} p_i(x)dx = \sum_{j=0}^{m} w_{ij} g(x_{ij})$$

Simpson-Regel: $m = 2, x_{i1} = (x_i + x_{i+1})/2$

$$\int_{x_i}^{x_{i+1}} f(x)dx = \frac{x_{i+1} - x_1}{6} \left(f(x_i) + 4f((x_i + x_{i+1}/2) + f(x_{i+1}) \right)$$

Durch Newton-Cotes werden Polynome vom Grad $\leq m$ für ungerade m und vom Grad $\leq m + 1$ für gerade m exakt integriert. Sonst hängt der Integrationsfehler von der $(m + 1)$-ten Ableitung $g'(x)$ ab.

Adaptive Wahl der Knoten:

- Starte mit wenigen Knoten zur ersten Bestimmung des Integrale
- Einfügen von Zwischenknoten; Bereiche in denen sich die Integralschätzung ändert weiter zerlegen.

R-Funktionen im Package `cubature`:`integrate`, `adaptIntegrate`.

- Integrationsbereich muss beschränkt sein (Ausweg: abschneiden, Transformation
- Singularitäten
- Hochdimensionale Funktionen
- übliche numerische Probleme

6.2 Laplace-Approximation

Gesucht ist das Integral

$$\int_{-\infty}^{\infty} \exp(-nh(x))dx,$$

wobei $h(x)$ eine konvexe, zweimal differenzierbare Funktion ist, die ihr Minimum an der Stelle $x = \tilde{x}$ hat (bzw. $\exp(-nh(x))$ ihr Maximum). Es gilt: $h'(\tilde{x}) = 0, h''(\tilde{x}) > 0$. **Taylorentwicklung** von $h(x)$ um $\tilde{x}$ ergibt

$$\begin{aligned} h(x) &\approx h(\tilde{x}) + h'(\tilde{x})(x - \tilde{x}) + \frac{1}{2}h''(\tilde{x})(x - \tilde{x})^2 \\ &= h(\tilde{x}) + \frac{1}{2}h''(\tilde{x})(x - \tilde{x})^2 \end{aligned}$$

Definition 6.1. *Die **Laplace-Approximation** ist definiert als*

$$\begin{aligned} \int_{-\infty}^{\infty} \exp(-nh(x))dx &\approx \exp(-nh(\tilde{x})) \\ &\quad \cdot \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}nh''(\tilde{x})(x - \tilde{x})^2\right) dx \\ &= \exp(-nh(\tilde{x}))\sqrt{\frac{2\pi}{nh''(\tilde{x})}} \end{aligned}$$

Der relative Fehler der Laplace-Approximation ist $O(\frac{1}{n})$.
Berechnung des Posteriori-Erwartungswerts von $g(\theta)$

$$\begin{aligned} E(g(\theta|x)) &= \int_\Theta g(\theta)p(\theta|x)d\theta && (15)\\ &= \int_\Theta g(\theta)\frac{f(x|\theta)p(\theta)}{\int_\Theta f(x|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}d\theta && (16)\\ &= \frac{\int_\Theta g(\theta)f(x|\theta)p(\theta)d\theta}{\int_\Theta f(x|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}} && (17)\end{aligned}$$

erfordert die Berechnung des Quotienten zweier ähnlicher Integrale.
Mit

$$\begin{aligned} -nh(\theta) &= \log f(x|\theta) + \log p(\theta)\\ -nq(\theta) &= \log g(\theta) + \log f(x|\theta) + \log p(\theta)\end{aligned}$$

ist

$$E(g(\theta)|x) = \frac{\int \exp(-nq(\theta))d\theta)}{\int \exp(-nh(\theta))d\theta)}.$$

Seien $\hat{\theta}$ und $\tilde{\theta}$ die Minimumstellen von $h(\theta)$ und $q(\theta)$. Dann ist

$$E(g(\theta)|x) \approx \sqrt{\frac{h''(\hat{\theta})}{q''(\tilde{\theta})}}\exp\left(-n(q(\tilde{\theta}) - h(\hat{\theta}))\right).$$

Die Laplace-Approximation für $E(g(\theta)|x)$ hat einen relativen Fehler von $O(1/n^2)$.

Beispiel 6.1 (Laplace-Approximation für das Poisson-Gamma-Modell)**.** *Simulierte Daten mit $\lambda = 25$ und Jeffreys-Priori:*

n	*Posterior-Erwartungswert*	*Approximation*	*absoluter Fehler*
1	*27.500*	*25.703*	*1.79745*
2	*28.750*	*27.801*	*0.94933*
3	*30.167*	*29.522*	*0.64487*
5	*26.300*	*25.909*	*0.39092*
10	*25.450*	*25.252*	*0.19763*
25	*25.700*	*25.620*	*0.07962*
50	*25.950*	*25.910*	*0.03991*
100	*26.175*	*26.155*	*0.01998*

Im mehrdimensionalen wird die zweite Ableitung durch die Hesse-Matrix $\boldsymbol{H}_h = \frac{\partial^2 h(x)}{\partial x_i \partial x_j}\Big|_{\tilde{x}}$ ersetzt. Taylorentwicklung:

$$h(x) \approx h(\tilde{x}) + \frac{1}{2}(x-\tilde{x})^T \boldsymbol{H}_h (x-\tilde{x})$$

Sei

$$I = \int \exp -nh(\boldsymbol{x})d\boldsymbol{x}$$

und $\tilde{\boldsymbol{x}}$ die Minimalstelle von $h(\boldsymbol{x})$.

Definition 6.2. *Die Laplace-Approximation von I ist gegeben als*

$$I \approx \left(\frac{2\pi}{n}\right)^{p/2} |\boldsymbol{H}_h|^{1/2} \exp\left(-nh(\tilde{\boldsymbol{x}})\right)$$

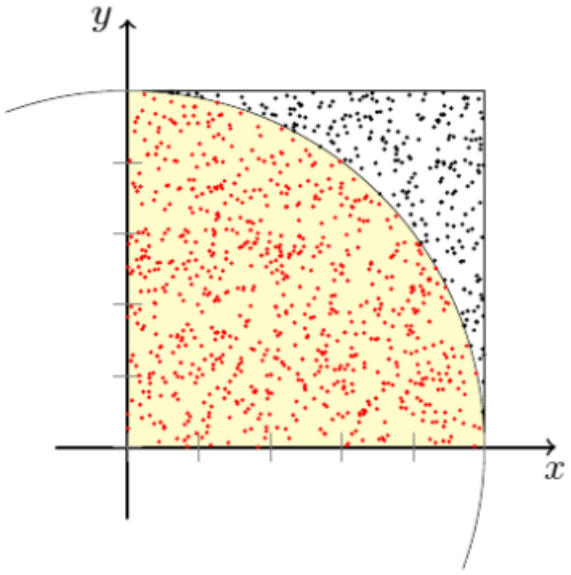
Setze

$$\begin{aligned} -nh(\boldsymbol{\theta}) &= \log p(\boldsymbol{x}|\boldsymbol{\theta}) + logp(\boldsymbol{\theta}) \\ -nq(\boldsymbol{\theta}) &= \log g(\boldsymbol{\theta}) + \log p(\boldsymbol{x}|\boldsymbol{\theta}) + logp(\boldsymbol{\theta}) \end{aligned}$$

und $\hat{\boldsymbol{\theta}}$ die Minimalstelle von $h(\boldsymbol{\theta})$ und $\tilde{\boldsymbol{\theta}}$ die Minimalstelle von $q(\boldsymbol{\theta})$, dann ist die Laplace-Approximation des Erwartungswerts von $g(\boldsymbol{\theta})$ gegeben als

$$E(g(\boldsymbol{\theta})) \approx \sqrt{\frac{\boldsymbol{H}_h}{\boldsymbol{H}_q}} \exp\left(-n\left(q(\tilde{\boldsymbol{\theta}}) - h(\hat{\boldsymbol{\theta}})\right)\right)$$

6.3 Monte-Carlo-Integration



Definition 6.3 (Monte-Carlo-Integration)**.** *Sei $f(x) > 0$ eine beliebige stetige Funktion mit bekanntem Wertebereich $[0, Y]$. $\int_a^b f(x)dx$ kann wie dann folgt approximiert werden.*

- *Ziehe n gleichverteilte Zufallszahlen x aus $[a, b]$*
- *Ziehe unabhängig davon n gleichverteilte Zufallszahlen y aus $[0, Y]$*
- *Berechne den Anteil h der Punkte (x_i, y_i), die unterhalb der Funktion f liegen*
- $\int_a^b f(x)dx \approx h(b-a)Y$

Ist $p(x)$ eine Dichte, so können Integrale der Form

$$E(g(x)) = \int g(x)p(x)dx$$

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

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

approximiert werden. Aus dem starken Gesetz der grossen Zahlen folgt

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

Die Varianz des Monte-Carlo-Schätzers $\bar{g}_m$ ist gegeben durch

$$Var(\bar{g}_m) = \frac{1}{m}\int \left(g(X) - E(g(x))\right)^2 p(x)dx = \frac{1}{m}Var(g).$$

Der Approximationsfehler verringert sich mit steigendem m. Es folgt sogar aus dem zentralen Grenzwertsatz

$$\sqrt{m}\left(g_m - E(g(x))\right) \sim N(0, Var(g)).$$

Schätzer für $Var(\bar{g}_m)$ ist

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

Zur Monte-Carlo-Schätzung muss aus der Posteriori gezogen werden. Methoden:

- Inversion
- Acception-Rejection
- Markov-Ketten (siehe Abschnitt 7)

Gegeben sei die Verteilungsfunktion $F(x)$ einer Zufallsvariablen X. Sei $u \sim U[0,1]$. Dann ist

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

Ziel: Wir wollen aus einer Dichtefunktion $f(x)$ ziehen. Gegeben sei eine Dichtefunktion $g(x)$, nach der wir problemlos Zufallszahlen ziehen können. Es existiere ein c, so dass

$$cg(z) \geq f(z)$$

für alle z mit $f(z) > 0$. Dann können Zufallszahlen gemäß $f(x)$ wie folgt gezogen werden:

- ziehe Z gemäß $g(z)$
- akzeptiere Z mit Wahrscheinlichkeit $\frac{f(z)}{g(z)c}$

Gegeben sei eine untere Schranke $s(z) \leq f(z)$. Für u Ziehung aus $U[0,1]$ akzeptiere Z

- wenn $u \leq \frac{s(z)}{cg(z)}$
- wenn $u \leq \frac{f(z)}{cg(z)}$

Der zweite Schritt kann ausgelassen werden, wenn bereits im ersten Schritt akzeptiert wurde.

Der Posteriori-Erwartungswert von $g(\theta)$ ist gegeben durch

$$\begin{aligned} E_{post}\left(g(\theta)|x\right) &= \int g(\theta)p(\theta|x) \\ &= \int \frac{g(\theta)f(x|\theta)p(\theta)}{f(x|\theta)p(\theta)}d\theta = \frac{E_{priori}\left(g(\theta)f(x|\theta)\right)}{E_{priori}\left(f(x|\theta)\right)} \end{aligned}$$

Mit i.i.d.-Ziehungen $\theta_1, \ldots, \theta_m$ aus der Priori-Verteilung gilt

$$E(g(\theta)|x) \approx \frac{\frac{1}{m}\sum_{i=1}^{m} g(\theta_i)f(x|\theta_i)}{\frac{1}{m}\sum_{i=1}^{m} f(x|\theta_i)}$$

Effizienteres Verfahren: Ziehe aus einer Importance Dichte q, die etwa der Posteriori entspricht.

$$E(g(\theta)|x) = \int \frac{g(\theta)p(\theta|x)}{q(\theta)} q(\theta) d\theta = E_q \left(\frac{g(\theta)p(\theta|x)}{q(\theta)} \right)$$

Mit Ziehungen $\theta_1, \ldots, \theta_m$ gemäß q gilt

$$\hat{g}_{IS} = \frac{1}{m} \sum_{i=1}^{m} g(\theta_i) \frac{p(\theta_i|x)}{q(\theta_i)}$$

wobei $\frac{p(\theta_i|x)}{q(\theta_i)}$ die Importancegewichte sind.

Die Berechnung des Importanceschätzers benötigt die normierte Posteriori-dichte. Die Normierungskonstante kann zuvor ebenfalls mit Importance Sampling geschätzt werden.

Die Varianz des Schätzers ist

$$Var(\hat{g}_{IS}) = \frac{1}{m} Var_q \left(\frac{g(\theta)p(\theta|x)}{q(\theta)} \right)$$

7 Markov Chain Monte Carlo

Idee: Erzeuge Ziehungen aus der Posteriori-Verteilung und approximiere daraus Statistiken der Posteriori-Verteilung

- Posteriori-Erwartungswert durch den Mittelwert
- Posteriori-Median über Median der Stichprobe
- Quantile der Posteriori-Verteilung über Quantile der Stichprobe
- HPD-Intervalle als kürzeste Intervalle, die $100(1-\alpha)\%$ der Stichprobe enthalten
- Idee: Erzeuge eine Markovkette, deren stationäre Verteilung die Posteriori-verteilung ist
- Ziehungen sind voneinander abhängig
- Funktioniert für komplexe und hochdimensionale Probleme

7.1 Markovketten*

Definition 7.1 (Markoveigenschaft)**.** *Ein zeitdiskreter stochastischer Prozess* $Y = \{Y_t, t \in \mathbb{N}_0\}$ *mit abzählbarem Zustandsraum* S *heisst Markov-Kette, wenn*

$$P(Y_t = k|Y_0 = j_0, Y_1 = j_1, \ldots, Y_{t-1} = j_{t-1}) = P(Y_t = k|Y_{t-1} = j_{t-1})$$

für alle $t \geq 0$ *und für alle* $k, j_0, \ldots, j_{t-1} \in S$*.*

- $P(Y_t = k|Y_{t-1} = j)$ heisst Übergangswahrscheinlichkeit
- die Markovkette ist homogen, wenn $P(Y_t = k|Y_{t-1} = j)$ nicht von t abhängt

$$p_{jk} = P(Y_t = k|Y_{t-1} = j) = P(Y_1 = k|Y_0 = j)$$

- Die Matrix $\boldsymbol{P} = (p_{jk})$ heisst Übergangsmatrix
- Eine Markovkette heißt irreduzibel, falls für alle $j, k \in S$ eine positive Zahl $1 \leq t \leq \infty$ existiert, so dass

$$P(Y_t = k|Y_0 = j) > 0,$$

  also jeder Zustand k von jedem Zustand j in endlicher Zeit erreicht werden kann

- Die Periode eines Zustands k ist der größte gemeinsame Teiler der Zeitpunkte n, zu denen eine Rückkehr möglich ist
- Falls alle Zustände einer Markov-Kette die Periode 1 haben, ist die Kette aperiodisch
- Die Wahrscheinlchkeit dafür, dass eine in k startende homogene Markov-Kette irgendwann wieder nach k zurückkehrt heißt Rückkehrwahrscheinlichkeit

$$f_{kk} = \sum_{t=1}^{\infty} P(Y_t = k; T_{t-1} \neq k, \ldots, Y_1 \neq k|Y_0 = k)$$

- der Zustand k heisst rekurrent, wenn $f_{kk} = 1$
- die Rückkehrzeit (Rekurrenzzeit) des Zustandes k ist

$$T_{kk} = \min(t \geq 1 : Y_t = k|Y_0 = k)$$

- ein rekurrenter Zustand ist positive rekurrent, wenn $E(T_{kk}) < \infty$, und null-rekurrent, wenn $E(T_{kk})$ nicht existiert.

Eine diskrete Wahrscheinlichkeitsverteilung π heisst invariante Verteilung der homogenen Markovkette Y_t bzw. ihrer Übergangsmatrix $\boldsymbol{P}$, falls gilt

$$\pi = \pi \boldsymbol{P}$$

Die invariante Verteilung einer irreduziblen Markovkette ist eindeutig, wenn alle Zustände positiv rekurrent sind.

- Eine Markovkette heisst ergodisch, wenn die Zustandsverteilung π_t von Y_t für jede beliebige Startverteilung π_0 gegen die selbe Wahrscheinlichkeitsverteilung π konvergiert:

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

- Die Grenzverteilung einer ergodischen Markovkette ist die invariante Verteilung π

- Eine homogene Markovkette mit Übergangsmatrix $\boldsymbol{P}$ ist ergodisch, wenn sie irreduzibel und aperiodisch ist

- Die Zustandsverteilung einer irreduziblen und aperiodischen, homogenen Markovkette kovergiert daher gegen die stationäre Verteilung π

- Eine ergodische Markovkette wird asymptotisch stationär, d.h., der Einfluss der Startverteilung geht verloren

Mit der Übergangsmatrix $\boldsymbol{P}$ einer irreduziblen, aperiodischen Markovkette, deren stationäre Verteilung $\boldsymbol{\pi}$ ist, können Zufallszahlen $Y \sim \pi$ wie folgt erzeugt werden:

- Wahl eines beliebigen Startwerts $y^{(0)}$

- Simulation der Realisierungen einer Markovkette der Länge m mit Übergangsmatrix $\boldsymbol{P}$: $(y^{(1)}, \ldots, y^{(m)})$

Ab einem gewissen Index b, dem *burn-in* geht der Einfluss der Startverteilung verloren und daher gilt approximativ

$$y^{(t)} \sim \pi, \text{ für } i = b, \ldots, m.$$

Die Ziehungen sind identisch verteilt, aber nicht unabhängig.

Die bisherigen Überlegungen können aus stetige Zustandsräume erweitert werden:

Definition 7.2. *Ein zeitdiskreter stochastischer Prozess* $Y = \{Y_t, t \in \mathbb{N}_0\}$ *mit Zustandsraum* S *heisst (allgemeine) Markovkette, wenn die Markoveigenschaft*

$$P(Y_t \in A | Y_0 \in A_0, Y_1 \in A_1, \ldots, Y_{t-2} \in A_{t-2}, Y_{t-1} = x) =$$
$$P(Y_t \in A | Y_{t-1} = x)$$

für beliebige $x \in S$ *und* $A_0, A_1, \ldots, A_{t-2}, A \subset S$ *gilt.*

Die Markovkette heisst homogen, falls die Wahrscheinlichkeit $P(Y_t \in A | Y_{t-1} = x)$ nicht vom Zeitpunkt t abhängt.

- Ist Y eine homogene Markovkette, so ist

  $$P(x, A) = P(Y_t \in A | Y_{t-1} = x) = P(Y_1 \in A | Y_0 = x)$$

  ihr Übergangskern.

- Für eine Markovkette mit endlichem Zustandsraum ist $P(x, A)$ durch die Übergangswahrscheinlichkeit $p_{j,k}$ bzw. die Übergangsmatrix bestimmt

- Für $S = \mathbb{R}$ ist der Übergangskern durch die Übergangsdichte $p(x, y)$ mit $P(x, A) = \int_A p(x, y)dy$ bestimmt.

- Eine Verteilung Π auf S mit Dichte π heisst invariant für den Übergangskern $P(x, A)$ genau dann, wenn für alle $A \in S$ gilt

  $$\Pi(A) = \int P(x, A)\pi(x)dx$$

- Eine allgemeine Markovkette ist irreduzibel, wenn von beliebigem Startwert x_0 aus jede Menge A mit $\Pi(A) > 0$ mit positiver Wahrscheinlichkeit in einer endlichen Zahl von Schritten erreicht werden kann

- $\{E_1, \ldots, E_{m-1}\}$ bilden einen m-Zyklus, wenn $P(x, E_{i+1 \, mod \, m}) = 1$ für alle $x \in E_i \subset S$ und alle i.

- Die Periode d der Markovkette ist der größte Wert m, für den ein m-Zyklus existiert und die Kette ist aperiodisch, wenn $d = 1$.

Ist Y eine irreduzible, aperiodische Markovkette mit Übergangskern P und invarianter Verteilung π, so ist π eindeutig bestimmt und es gilt

$$||P^t(x, .) - \pi(.)|| =\!\!\rightarrow 0,$$

mit $t \to \infty$.

- Ziel: finde für eine vorgegebene Dichte $\pi(\theta) = p(\theta|x)$ einen Übergangskern P mit invarianter Verteilung π.
- Für eine irreduzible Markovkette ist die invariante Verteilung eindeutig, aber nicht umgekehrt

Beispiel 7.1.

$$\boldsymbol{P} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

Mit $\boldsymbol{\pi} = (\pi, 1-\pi)$ als stationärer Verteilung ist das Gleichungssystem für p und q unterbestimmt, hat also unendlich viele Lösungen.

7.2 Metropolis-Hastings-Algorithmus

Im Metropolis-Hastings-Algorithmus wird der Übergangskern der Markovkette erzeugt, in dem ausgehen von $\theta^{(k-1)}$ eine Ziehung aus einer Vorschlagsdichte $q(\theta^*|\theta^{(k-1)})$ erfolgt.

Der Wert θ^* wird mit Wahrscheinlichkeit

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

akzeptiert, also

$$\theta^{(k)} = \theta^*$$

gesetzt. Andernfalls wird $\theta^{(k-1)}$ beibehalten, also

$$\theta^{(k)} = \theta^{(k-1)}.$$

Der Metropolis-Hastings-Algorithmus erzeugt eine homogene Markovkette.

Der Übergangskern der Markovkette ist

$$P(x, A) = \int_A p(x, z)dz + r(x)\delta_x(A)$$

mit

$$p(x, z) = \begin{cases} q(x, z)\alpha(x, z) & x \neq z \\ 0 & \text{sonst} \end{cases}$$

und

$$r(x) = 1 - \int p(x, z)dz.$$

$$\begin{aligned}\alpha &= \frac{p(\theta^*|x)q(\theta^{(k-1)}|\theta^*)}{p(\theta^{(k-1)}|x)q(\theta^*|\theta^{(k-1)})} \\ &= \frac{f(x|\theta^*)p(\theta^*)f(x)q(\theta^{(k-1)}|\theta^*)}{f(x|\theta^{(k-1)})p(\theta^{(k-1)})f(x)(q(\theta^*|\theta^{(k-1)})}\end{aligned}$$

- Die Normalisierungskonstante kürzt sich, muss also für den Algorithmus nicht bekannt sein.
- Für die Konvergenz müssen Irreduzibilität und Aperiodizität der Markovkette vorliegen. Diese sind schwer nachzuweisen.
- Z.B. der Posteriori-Erwartungswert kann über den Mittelwert der Stichprobe approxmiert werden.
- Die Ziehungen sind nicht unabhängig

Die Varianz des Schätzers von $E(g(\theta))$ ist

$$E((\hat{g}(\theta) - E(g(\theta)))^2) = \frac{1}{m}\Omega_0(g)$$

wobei $\Omega_0(g)$ die Spektraldichte des Prozesses $g(\theta^{(k)})$ ist und mit ρ_s der Autokorrelation des Prozzesses mit Lag s gilt

$$\tau = \frac{\Omega_o(g)}{Var(g)} = 1 + 2\sum_{s=1}^{\infty}\rho_s$$

τ heisst Ineffizienzfaktor ($\tau = 1$ für i.i.d-Ziehungen). m/τ heisst effektive Stichprobengröße.

Durch Ausdünnen der Ziehungen erhält man (fast) unabhängige Stichproben.

- Independence Proposal: Vorschlagsverteilung ist unabhängig vom aktuellen Wert
- Symetrisches Proposal: $q(\theta^*|\theta^{(k-1)}) = q(\theta^{(k-1)}|\theta^*)$, die Vorschlagsdichte kürzt sich aus der Akzeptanzwahrscheinlichkeit (Metropolis-Algorithmus):

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

Jeder Vorschlag mit $p(\theta^*|x) > p(\theta^{(k-1)}|x)$ wird angenommen!

- Random Walk Proposal: Vorschlagsverteilung ist ein Random Walk

$$\theta^* = \theta^{(k-1)} + \epsilon, \epsilon \sim f$$

also $q(\theta^*|\theta^{(k-1)}) = f(\theta^* - \theta^{(k-1)})$.

Random Walk wird in der Regel mit Normalverteilung konstruiert:

$$\theta^* \sim N(\theta^{(k-1)}, C)$$

mit vorgegebener Kovarianzmatrix C.

- Eine zu kleine Varianz führt zu hohen Akzeptanzraten, aber ineffizienten, da stark autokorrelierten Ziehungen. Im Extremfall $C \to 0$ führt zu $\alpha = 1$, $\tau \to \infty$.
- Eine zu große Varianz führt zu zu großen Schritten, Vorschläge liegen in den Enden der Posterioriverteilung, sehr kleine Akzeptanzraten.
- Tuning der Kovarianzmatrix notwendig
- Metropolis-Hastings-Algorithmus kann für θ-Vektoren durchgeführt werden
- Akzeptanzraten jedoch i.d.R. geringer mit höherer Dimension
- Alternative ist der komponentenweise Metropolis-Hastings: Jede Komponente des Parameters wird einzeln (skalar oder blockweise) aufdatiert. Sei $\theta = (\theta_1, \theta_2)$:

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

- Updates können in fester oder zufälliger Weise erfolgen

Beispiel 7.2 (Poisson-Gamma-Modell)**.**

7.3 Gibbs-Sampling

Für multivariate Parameter bietet es sich beim Gibbs-Sampler an, die **vollständig bedingte Posteriori** (full conditional) als Proposalverteilung zu benutzen - wenn es sich um eine Standardverteilung handelt

$$q(\theta_1^*) \propto p(\theta_1^*|x, \theta_2^{(k-1)}).$$

Damit wird die Akzeptanzwahrscheinlichkeit gleich 1.

Beispiel 7.3 (Normalverteilung)**.** *Modell:*

$$x_i \sim N(\mu, \sigma^2)$$

Likelihood:

$$p(\boldsymbol{x}|\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{\sigma^2}\sum(x_i - \mu)^2\right)$$

Semi-konjugierte Prioris:

$$\begin{aligned} \mu &\sim N(m_0, s_0) \\ \tau = \sigma^{-2} &\sim Ga(a, b) \\ \mu &\perp \tau \end{aligned}$$

Posteriori-Verteilung:

$$\begin{aligned} p(\mu, \tau|\boldsymbol{x}) &\propto \tau^{n/2} \exp\left(-\frac{\tau}{2}\sum(x_i - \mu)^2\right) \\ &\cdot \exp\left(-\frac{1}{2s_0}(\mu - m_0)^2\right) \tau^a \exp(-b\tau) \end{aligned}$$

Full conditional von μ:

$$p(\mu, |\boldsymbol{x}, \tau) \propto \exp\left(-\frac{\tau}{2}\sum(x_i - \mu)^2 - \frac{1}{2s_0}(\mu - m_0)^2\right)$$

Es handelt sich um den Kern der $\mathrm{N}(s^{-1}m, s^{-1})$-Verteilung mit $m = \tau\sum x_i + m_0/s_0$ und $s = n\tau + s_0^{-1}$.

Posteriori-Verteilung:

$$\begin{aligned} p(\mu, \tau|\boldsymbol{x}) &\propto \tau^{n/2} \exp\left(-\frac{\tau}{2}\sum(x_i - \mu)^2\right) \\ &\cdot \exp\left(-\frac{1}{2s_0}(\mu - m_0)^2\right) \tau^a \exp(-b\tau) \end{aligned}$$

Full conditional von τ:

$$p(\tau|\boldsymbol{x}, \mu) \propto \tau^{a+n/2} exp\left(-(b + 0.5\sum(x_i - \mu)^2)\right)$$

Es handelt sich um den Kern der $\mathrm{Ga}(a + n/2, b + 0.5\sum(x_i - \mu)^2)$-Verteilung.

Gibbs-Sampler:

1. Wähle Startwert τ_0
2. Ziehe $\mu \sim N(s^{-1}m, s^{-1})$

3. Ziehe $\tau \sim Ga(a + n/2, b + 0.5 \sum (x_i - \mu)^2)$

4. Iteriere 2 und 3 für $m = 1, \ldots, M$

Allgemeiner können die Komponenten des Parametervektors in Blöcke zerlegt: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$. Jeder Block wird dann aus der vollständig bedingten Dichte $p(\boldsymbol{\theta}_j | \boldsymbol{x}, \boldsymbol{\theta}_{-j})$ gezogen.