Bayesianische Inferenz

Sylvia Frühwirth-Schnatter und Gregor Kastner

Inhaltsverzeichnis

1 Einleitung . . . 2
2 Der bayesianische Anteilstest . . . 3
3 Lineare Regression . . . 6
4 Bayesianische Punkt- und Intervallschätzung . . . 11
5 Vorhersagen mithilfe der bayesianischen Statistik . . . 16
6 Bayesianisches Testen von Hypothesen und Modellwahl . . . 22
7 Jenseits konjugierter Bayes-Analysen . . . 29
8 Zusammenfassung und Ausblick . . . 32
Literatur . . . 33

Zusammenfassung

Das Kapitel versucht sich in einer konzisen Einführung in die Welt der Bayes-Statistik. Schwerpunkte bilden dabei das statistische Modellieren mit Likelihood *und* Prior, bayesianisches Punkt- und Intervallschätzen sowie das Schätzen nichtlinearer Funktionale, Vorhersagen, bayesianisches Testen und Modellwahl via Bayes-Faktoren und dem Dichtequotienten nach Savage-Dickey sowie ein kurzer Abriss über nichtkonjugierte Bayes-Analysen via Monte-Carlo-Methoden. Die theoretischen Ausführungen werden begleitet von zwei konkreten Beispielen: dem Schätzen, Testen und Vorhersagen von Anteilen sowie der Regressionsanalyse mit und ohne Regularisierung.

S. Frühwirth-Schnatter
Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien, Wien, Österreich
E-Mail: sylvia.fruehwirth-schnatter@wu.ac.at

G. Kastner (✉)
Institut für Statistik, Universität Klagenfurt, Klagenfurt, Österreich
E-Mail: gregor.kastner@aau.at

© Springer-Verlag GmbH Deutschland, ein Teil von Springer Nature 2024

J. Gertheiss et al. (Hrsg.), *Moderne Verfahren der Angewandten Statistik*,
https://doi.org/10.1007/978-3-662-63496-7_23-2

1 Einleitung

Mithilfe des bayesianischen Ansatzes lassen sich zahlreiche komplexe Probleme der statistischen Inferenz, die innerhalb der Standard-Statistik nicht oder nur wenig überzeugend zu bewältigen sind, auf kohärente und elegante Weise lösen. Neuere Beispiele hierfür sind die Inferenz aufgrund nichtnormalverteilter Daten (etwa bei sehr kleinen Trefferzahlen, darunter viele Nullen), Robustheitsbetrachtungen beim statistischen Schließen, falls die Bruchpunkte (bekannter oder unbekannter Zahl) nicht beobachtet werden, oder die regularisierte Schätzung hochdimensionaler Modelle. Die Vorteile der Bayes-Statistik, wie sie grundsätzlich seit Laplace bekannt sind, wurden besonders durch die technologischen Entwicklungen in der zweiten Hälfte des 20. Jahrhunderts gefördert; erhebliche Innovationen in zahlreichen Bereichen der Schätz- und Testtheorie waren die Folge. Der bayesianische Ansatz findet heute breite Anwendung in vielen Gebieten wie Biologie, Finanzwirtschaft, Genetik, Marketing, Medizin, Ökonomie et cetera.

Bayes-Statistik wird also aus mehreren Gründen stetig populärer, findet jedoch in vielen Statistik-Einführungskursen an Universitäten und anderen Bildungseinrichtungen nach wie vor kaum Platz. Ziel dieses Kapitels ist es also, die Grundlagen bayesianischen Modellierens vorzustellen, wichtige Konzepte des bayesianischen Schätzens, Vorhersagens und Testens zu erläutern sowie Hilfestellungen für die praktische Anwendung anhand von konkreten Beispielen zu geben.

Bayesianisches Modellieren bedeutet im Wesentlichen eine (im Kontext der Aufgabenstellung sinnvolle) Spezifikation der *gemeinsamen* Verteilung von beobachtbaren Daten $\boldsymbol{y} = (y_1, \ldots, y_N)$ und dem Modellparameter(vektor) $\boldsymbol{\vartheta}$. Aufgrund des fundamentalen Zusammenhanges

$$p(\boldsymbol{y}, \boldsymbol{\vartheta}) = p(\boldsymbol{y}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}) \tag{1}$$

ist dies eindeutig möglich, wenn die aus der frequentistischen Statistik bekannte *Likelihood* $p(\boldsymbol{y}|\boldsymbol{\vartheta})$ und die sogenannte *Prior* $p(\boldsymbol{\vartheta})$ festgelegt werden.

Wir beginnen die Erkundung des fundamentalen Ansatzes (1) in Abschn. 2 mit einem konkreten Beispiel, dem Modellieren von Anteilen, gefolgt von Abschn. 3, der sich mit der Bayes-Regression auseinandersetzt. In Abschn. 4 behandeln wir Punkt- und Intervallschätzer sowie das Schätzen nichtlinearer Funktionale. Abschn. 5 hat das Vorhersagen zukünftiger Beobachtungen zum Thema; Abschn. 6 die bayesianische Modellwahl sowie das Testen von Hypothesen. In Abschn. 7 gehen wir kurz auf Probleme ein, die sich einer rein analytischen Lösung weitgehend entziehen. Abschn. 8 schließt zusammenfassend und gibt einen Ausblick.[1]

[1] R-Code für sämtliche Beispiele und Abbildungen in diesem Kapitel ist unter https://github.com/gregorkastner/bayes-inferenz zu finden.

2 Der bayesianische Anteilstest

Eine der klassischen Fragestellungen in der Statistik ist das Bestimmen eines relativen *Anteils* ϑ in einer (großen) Grundgesamtheit. Dies kann etwa der Anteil an defekten Bauteilen, der Anteil an Wählern einer bestimmten Partei oder der Anteil an Personen, die ein gewisses Produkt kaufen, sein.

Typischerweise wird angenommen, dass N Personen (mit „Zurücklegen“) in unabhängiger Weise aus dieser Grundgesamtheit gezogen werden. Für jede dieser Personen wird beobachtet, ob sie die interessierende Eigenschaft erfüllt oder nicht. Formalisiert beobachten wir also $\boldsymbol{y} = (y_1, \ldots, y_N)$, wobei $y_i = 1$ gilt, wenn Person i die Eigenschaft erfüllt. Umgekehrt ist $y_i = 0$, wenn Person i die Eigenschaft nicht erfüllt.

2.1 Probabilistische Modellformulierung

Das wohl wichtigste Wesensmerkmal der (bayesianischen wie nichtbayesianischen) Statistik ist ein probabilistisches Modell zur Formalisierung der dem Auswahlprozess innewohnenden Unsicherheit. Im vorliegenden Fall könnte dies wie folgt aussehen.

Die Beobachtungen $\boldsymbol{y} = (y_1, \ldots, y_N)$ seien N unabhängige Realisationen einer dichotomen Zufallsvariable Y mit Erfolgswahrscheinlichkeit ϑ. Es gilt also

$$\Pr(Y = 1|\vartheta) = \vartheta,$$

und die Wahrscheinlichkeitsfunktion einer einzelnen Beobachtung y_i von Y ist für alle $i = 1, \ldots, N$ gegeben durch

$$p(y_i|\vartheta) = \vartheta^{y_i}(1-\vartheta)^{1-y_i}.$$

Aufgrund der Unabhängigkeit der Beobachtungen gilt für die gemeinsame Wahrscheinlichkeitsfunktion aller Beobachtungen

$$p(\boldsymbol{y}|\vartheta) = \prod_{i=1}^{N} p(y_i|\vartheta) = \vartheta^{S_N}(1-\vartheta)^{N-S_N}, \tag{2}$$

wobei $S_N = \sum_{i=1}^{N} y_i$ die Anzahl der die Eigenschaft erfüllenden Personen bezeichnet.

2.2 Herleitung der Posterior

Um eine auf die Beobachtungen bedingte Aussage über den Anteil ϑ treffen zu können, gilt es, die sogenannte *Posterior* (auch: *A-posteriori-Verteilung*) $p(\vartheta|\boldsymbol{y})$ zu bestimmen. Dies gelingt durch das Bayes-Theorem

$$p(\vartheta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\vartheta)p(\vartheta)}{p(\boldsymbol{y})} \propto p(\boldsymbol{y}|\vartheta)p(\vartheta). \tag{3}$$

Hierbei ist $p(\boldsymbol{y}|\vartheta)$ die Likelihood (2) und $p(\boldsymbol{y})$ die Normalisierungskonstante (*Marginal Likelihood*), welche nicht vom Parameter ϑ abhängt und daher gerne unter Verwendung des Proportionalitätssymbols $\propto$ weggelassen wird. Die dritte Größe in (3), $p(\vartheta)$, ist die sogenannte *Prior* (auch: *A-priori-Verteilung*), welche eine wesentliche Einflussgröße darstellen kann und im Zuge der Modellierung festgelegt werden muss.

Beginnen wir mit einer konstanten Prior $p(\vartheta) = 1$. Diese Wahl spiegelt die Annahme wider, dass sämtliche gleich große Bereiche des Parameterraumes $\Theta = [0, 1]$ a priori (also *vor* dem Betrachten der Daten) gleich wahrscheinlich sein sollen.

Aufgrund dieser speziellen Wahl ergibt sich umgehend durch Einsetzen in (3)

$$p(\vartheta|\boldsymbol{y}) \propto \vartheta^{S_N}(1-\vartheta)^{N-S_N}.$$

Betrachtet man diese Gleichung als Funktion von ϑ, so wird klar, dass sowohl Likelihood als auch Posterior proportional zur Dichtefunktion einer Beta-Verteilung sind,[2] und damit

$$\begin{aligned} &\vartheta|\boldsymbol{y} \sim \mathcal{B}(a_N, b_N), \\ &a_N = S_N + 1, \quad b_N = N - S_N + 1 \end{aligned} \tag{4}$$

gilt. Es zeigt sich, dass nur zwei suffiziente Statistiken für die Berechnung der Posterior nötig sind: die Gesamtzahl der Beobachtungen N und die Anzahl der Erfolge S_N.

Will man andere Priors wählen, zum Beispiel, weil aus vergangenen Studien schon Wissen über den Parameter vorliegt, so bietet sich als Verallgemeinerung der Gleichverteilung die Beta-Verteilung an. Kombiniert man nämlich die Likelihood mit einer $\mathcal{B}(a_0, b_0)$-Prior, so erhält man

$$p(\vartheta|\boldsymbol{y}) \propto \vartheta^{S_N+a_0-1}(1-\vartheta)^{N-S_N+b_0-1},$$

also wieder eine Beta-Posterior, deren Parameter a_N und b_N die suffizienten Statistiken N und S_N mit den sogenannten Hyperparametern a_0 und b_0 verknüpfen:

$$\vartheta|\boldsymbol{y} \sim \mathcal{B}(a_N, b_N), \quad a_N = a_0 + S_N, \quad b_N = b_0 + N - S_N. \tag{5}$$

[2] Die Dichtefunktion der Beta-Verteilung für $y \in [0, 1]$ mit den Parametern $a \in \Re^+$ und $b \in \Re^+$ ist gegeben durch

$$f_B(y; a, b) = \frac{1}{B(a, b)}\vartheta^{a-1}(1-\vartheta)^{b-1}.$$

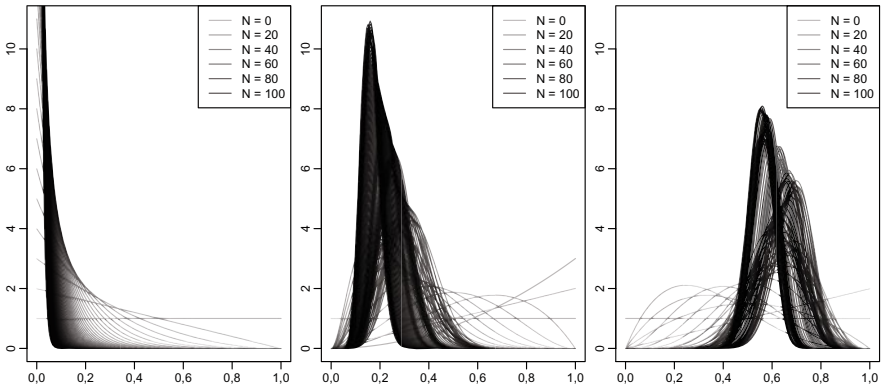

Abb. 1 Dichten der Posterior $p(\vartheta|\boldsymbol{y})$ bei konstanter Prior für zufällig generierte Daten mit verschiedenen Werten von ϑ. Links: $\vartheta = 0$. Mitte: $\vartheta = 0{,}1$. Rechts: $\vartheta = 0{,}5$. Zur Visualisierung wird die Stichprobe sequentiell vergrößert: $N = 0, 1, \ldots, 100$

Aus (5) erkennt man leicht, wie die Hyperparameter a_0 und b_0 zu interpretieren sind: a_0 erhöht die Anzahl der Erfolge und b_0 erhöht die Anzahl der Misserfolge, wenn in einem imaginären Experiment weitere $a_0 + b_0$ Versuche unternommen werden.

Durch Visualisierung der Posterior gewinnt man rasch einen Eindruck, welche Wertebereiche von ϑ nach dem Beobachten der Daten wie wahrscheinlich sind. Abb. 1 lässt erkennen, dass die Posterior unterschiedlichste Formen annehmen kann. Gilt beispielsweise $a_N > 1$ und $b_N > 1$, so liegt der Modus der Posterior im Inneren des Parameterraums. Ist $b_N > 1$ und $a_N = 1$, so liegt der Modus der Posterior bei 0. Gilt $0 < a_N < 1$, so ist die Dichte der Posterior sogar unbeschränkt. Gilt umgekehrt $a_N > 1$ und $b_N = 1$, so liegt der Modus der Posterior bei 1, und wenn $0 < b_N < 1$ gilt, dann ist die Dichte wiederum unbeschränkt.

Punktschätzer für den Parameter ϑ lassen sich nun einfach durch verschiedene Lagemaße der Posterior bestimmen. Typisch sind etwa der Modus der Posterior (kann hier zu $(a_N - 1)/(a_N + b_N - 2)$ errechnet werden), der Erwartungswert $\mathrm{E}(\vartheta|\boldsymbol{y})$ (hier $a_N/(a_N + b_N)$) oder der Median (in diesem Beispiel nicht in geschlossener Form verfügbar). Wählt man eine gleichverteilte Prior ($a_0 = b_0 = 1$), so erhält man beispielsweise als Modus der Posterior die relative Häufigkeit S_N/N, während sich der Erwartungswert der Posterior mit $(S_N + 1)/(N + 2)$ im Allgemeinen (also für $S_N \neq N/2$) davon unterscheidet. Dies macht deutlich, dass für die Bestimmung von Punktschätzern nicht nur die Wahl der Prior, sondern auch die Wahl des Lagemaßes eine Rolle spielt.[3]

[3] Ob der gebotenen Kürze an dieser Stelle nur erwähnt: Sowohl die Wahl der Prior sowie die Wahl des Lagemaßes kann in direkten Zusammenhang mit Verlustfunktionen gesetzt werden (siehe etwa Robert 2001).

Tab. 1 Zusammenfassung dreier Priors und der dadurch implizierten Posteriors für das Anteilsbeispiel

	a_0	b_0	a_N	b_N	$\mathrm{E}(\vartheta)$	$\mathrm{SD}(\vartheta)$	$\mathrm{E}(\vartheta\|\boldsymbol{y})$	$\mathrm{SD}(\vartheta\|\boldsymbol{y})$
Gleichverteilung	1	1	7	1539	0,5	0,2887	0,0045	0,0017
Jeffreys-Prior	0,5	0,5	6,5	1538,5	0,5	0,3536	0,0042	0,0016
Informative Prior	0,98	97,02	6,98	1635,02	0,01	0,01	0,0043	0,0016

2.3 Eine SARS-CoV-2-Prävalenzstudie

Um die Prävalenz von SARS-CoV-2 inklusive der Dunkelziffer, also der Anzahl der undetektiert Infizierten, besser abschätzen zu können, wurde im April 2020 eine für Österreich repräsentative Stichprobe ausgewählt.[4] Es wurden $N = 1544$ sogenannte Polymerase-Kettenreaktionstests durchgeführt, von denen $S_N = 6$ ein positives Ergebnis lieferten. Zur Illustration sollen drei verschiedene Priors näher analysiert werden. Erstens die eingangs diskutierte *flache* Prior $\mathscr{B}(1, 1)$, zweitens die sogenannte *Jeffreys-Prior* $\mathscr{B}(0{,}5, 0{,}5)$ und drittens eine informative Prior, bei der die Hyperparameter so gewählt werden, dass sowohl A-priori-Erwartungswert als auch A-priori-Standardabweichung gleich 0,01 sind. Man geht also im letzteren Fall mit recht hoher Sicherheit vor dem Betrachten der Daten davon aus, dass der Anteil bei etwa 1 % liegt. Diese Annahmen implizieren bei Verwendung einer Beta-Prior $a_0 = 0{,}98$ und $b_0 = 97{,}02$.

Tab. 1 fasst die Parameter sowie Erwartungswert und Standardabweichung der Priors und der Posteriors zusammen. Es wird ersichtlich, dass der Einfluss der Priorwahl im konkreten Beispiel aufgrund der Stichprobengröße nicht besonders stark ausfällt. Dennoch gibt es kleinere Unterschiede: Die gleichverteilte Prior impliziert einen etwas höheren Erwartungswert der Posterior als die anderen beiden Priors. Weiterhin fällt auf, dass die von der Jeffreys-Prior implizierte Posterior eine niedrigere Standardabweichung aufweist als die von der gleichverteilten Prior implizierte Posterior, obwohl die Jeffreys-Prior selbst eine höhere Standardabweichung aufweist als die gleichverteilte.

Abb. 2 stellt die Dichten der Priors und Posteriors graphisch dar. Erneut wird ersichtlich, dass sich die Priors deutlich, die Posteriors aber nur wenig unterscheiden. Man beachte auch, dass die Posteriors rechtsschief sind.

3 Lineare Regression

Als zweites Beispiel für den bayesianischen Ansatz dient das lineare Regressionsmodell. Gegeben sei ein Vektor der Länge N von Responsevariablen $y_1, \ldots, y_N$,

[4] Die Ergebnisse der Studie sind im Internet unter https://www.sora.at/fileadmin/downloads/projekte/Austria_Spread_of_SARS-CoV-2_Study_Report.pdf abrufbar. Für alternative Methoden zur Prävalenzschätzung in Österreich siehe etwa Hirk et al. (2020).

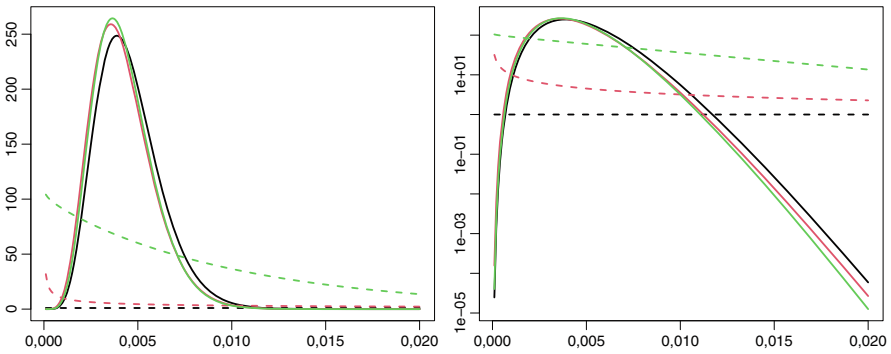

Abb. 2 Visualisierung dreier Priors (strichlierte Linien) und dreier Posteriors (durchgezogene Linien) für die Anteilsschätzung in der ersten SORA-Prävalenzstudie. Schwarz: gleichverteilte Prior. Rot: Jeffreys-Prior. Grün: informative Prior. Links: lineare Skala. Rechts: logarithmische Skala

die in (möglicher) Abhängigkeit von d erklärenden Variablen $x_{i,1}, \ldots, x_{i,d}$ stehen. Hierbei sei $x_{i,1} := 1$ für $i = 1, \ldots, N$ der konstante Prädiktor. Es wird weiterhin angenommen, dass die Fehler unabhängig voneinander (homoskedastisch) normalverteilt seien, also

$$y_i = \boldsymbol{x}_i \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right), \tag{6}$$

für $i = 1, \ldots, N$ gelte. Hierbei beinhalte der $(1 \times d)$-dimensionale Zeilenvektor $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$ die jeweils zu y_i gehörigen erklärenden Variablen, und $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top$ bestehe aus den unbekannten Regressionskoeffizienten zur Mittelwertvorhersage von y_i durch die Prädiktoren $\boldsymbol{x}_i$. Anmerkung: Weil $x_{i,1} = 1$ für alle i gilt, ist β_1 der y-Achsenabschnitt.

3.1 Zur Likelihood

Ziel ist es, die unbekannten Parameter $\boldsymbol{\beta}$ und σ^2 in (6) aus den Daten zu lernen. Das in (3) für einen univariaten Parameter ϑ formulierte Bayes-Theorem gilt gleichermaßen für den multivariaten Parameter $\boldsymbol{\vartheta} = (\boldsymbol{\beta}, \sigma^2)$ und man erhält

$$p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}) \propto p(\boldsymbol{y} | \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2).$$

Ein erster wichtiger Schritt besteht darin, die Likelihood zu konkretisieren. Unter Verwendung der multivariaten Normalverteilung[5] gilt wegen der Annahme unab-

[5] Die Dichtefunktion der multivariaten Normalverteilung für $\boldsymbol{y} \in \Re^d$ mit Erwartungswert $\boldsymbol{\mu} \in \Re^d$ und positiv definiter Varianz-Kovarianz-Matrix $\boldsymbol{\Sigma} \in \Re^{d \times d}$ ist gegeben durch

$$f_N(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})\right).$$

hängiger, homoskedastischer Fehler

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}_N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_N\right), \tag{7}$$

wobei $\boldsymbol{I}_N$ die N-dimensionale Einheitsmatrix bezeichne und

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_N \end{pmatrix}$$

gelte. Die Likelihood $p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)$ von $\boldsymbol{y}$ ist also

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right). \tag{8}$$

Spätestens an dieser Stelle wird klar, dass die Prior $p(\boldsymbol{\beta}, \sigma^2)$ spezifiziert werden muss, um ein vollständiges Modell zu erhalten. Zur Illustration diene im Folgenden eine sogenannte *uninformative* Prior.

3.2 Regression mit einer uninformativen Prior

Während die in Abschn. 2.2 verwendete $\mathcal{B}(a_0, b_0)$-Prior eine gültige Wahrscheinlichkeitsdichte definiert, solange $a_0 > 0$ und $b_0 > 0$ gilt, wird in der Regressionsanalyse häufig mit sogenannten *uneigentlichen (improper)* Priors gearbeitet, die keine gültige Wahrscheinlichkeitsdichte definieren. Wir verwenden zur Illustration den Klassiker (Jeffreys 1939)

$$p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}. \tag{9}$$

Diese Prior definiert keine Wahrscheinlichkeitsdichte, da das Integral auf $\Re^d \times \Re^+$ unendlich ist, eignet sich jedoch gut, um A-priori-Unwissen insbesondere in Bezug auf den Parametervektor $\boldsymbol{\beta}$ auszudrücken.

Zur Vereinfachung der Likelihood (8) in Kombination mit dieser Prior ist es hilfreich, sich dreier aus der klassischen Regressionsanalyse bekannter Statistiken zu bedienen. Erstens ist das der Kleinste-Quadrate-Schätzer $\hat{\boldsymbol{\beta}}$ der unbekannten Regressionskoeffizienten $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y},$$

zweitens die Quadratsumme der zugehörigen Residuen

$$\text{SSR} = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}},$$

und drittens die Freiheitsgrade $\text{df} = N - d$. Der im frequentistischen Sinne unverzerrte Schätzer der Fehlervarianz σ^2 ist $\hat{\sigma}^2 = \text{SSR}/\text{df}$.

Unter Verwendung der Identität $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ lässt sich die Quadratsumme in (8) in der folgenden Art zerlegen:

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \text{SSR},$$

was in Kombination mit der Prior (9) zur folgenden Posterior führt:

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) &\propto \left(\frac{1}{\sigma^2}\right)^{N/2+1} \exp\left(-\frac{\text{SSR}}{2\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right) \\ &= \left(\frac{1}{\sigma^2}\right)^{(N-d)/2+1} \exp\left(-\frac{\text{SSR}}{2\sigma^2}\right) \left(\frac{1}{\sigma^2}\right)^{d/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right). \end{aligned}$$

Mit etwas Erfahrung erkennt man, dass es sich hierbei um ein Produkt einer inversen Gammaverteilung[6] und einer multivariaten Normalverteilung handelt, womit die Posterior durch

$$\begin{aligned} &\boldsymbol{\beta}|\sigma^2, \boldsymbol{y} \sim \mathcal{N}_d\left(\boldsymbol{b}_N, \sigma^2 \boldsymbol{B}_N\right), \quad \sigma^2|\boldsymbol{y} \sim \mathcal{G}^{-1}\left(c_N, C_N\right), \\ &\boldsymbol{b}_N = \hat{\boldsymbol{\beta}}, \quad \boldsymbol{B}_N = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \quad c_N = \frac{\text{df}}{2}, \quad C_N = \frac{\text{SSR}}{2} \end{aligned} \tag{10}$$

charakterisiert wird. Die Posterior existiert nur, wenn $\boldsymbol{X}$ vollen Rang hat.

Für die Berechnung der marginalen Posterior $\boldsymbol{\beta}|\boldsymbol{y}$ muss über $\sigma^2|\boldsymbol{y}$ integriert werden, also

$$p(\boldsymbol{\beta}|\boldsymbol{y}) = \int_0^\infty p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}) p(\sigma^2|\boldsymbol{y}) d\sigma^2.$$

Dieses Integral ist analytisch lösbar, man erhält

$$\boldsymbol{\beta}|\boldsymbol{y} \sim t_{2c_N}\left(\boldsymbol{b}_N, \boldsymbol{B}_N \hat{\sigma}^2\right), \tag{11}$$

[6] Die Dichtefunktion der inversen Gammaverteilung für $y \in \Re^+$ mit den Parametern $\alpha \in \Re^+$ und $\beta \in \Re^+$ ist

$$f_{IG}(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{y}\right)^{\alpha+1} e^{-\beta/y}.$$

wobei $\hat{\sigma}^2 = \text{SSR}/\text{df}$ den aus der frequentistischen Statistik bekannten unverzerrten Varianzschätzer bezeichnet.[7]

Die obigen Resultate legen zwei äquivalente Wege nahe, mithilfe eines Computers M Ziehungen aus der Posterior der Regressionsparameter $\boldsymbol{\beta}$ zu generieren.

1. Für $m = 1, \ldots, M$, ziehe $\sigma^{2,(m)}$ aus der inversen Gammaverteilung in (10) und dann, unter Verwendung dieser Ziehung, $\boldsymbol{\beta}^{(m)}$ aus der multivariaten Normalverteilung $\mathcal{N}_d\left(\boldsymbol{b}_N, \sigma^{2,(m)}\boldsymbol{B}_N\right)$.
2. Ziehe M-mal direkt aus der multivariaten Student-t-Verteilung (11).

Diese Ziehungen können dann zur Monte-Carlo-Schätzung interessierender Größen verwendet werden.

3.3 Ein AR(1)-Modell für Wechselkursdaten

Als Anwendungsbeispiel für das bayesianische Regressionsmodell soll im Folgenden ein sogenanntes autoregressives Modell erster Ordnung, das AR(1)-Modell, dienen. Bei diesem wird angenommen, dass der Wert einer bestimmten Größe zum Zeitpunkt t in linearer Abhängigkeit vom Wert zum Zeitpunkt $t-1$ steht:

$$y_t = \beta_1 + \beta_2 y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \sigma^2\right). \tag{12}$$

In anderen Worten wird hier y_t bedingt auf y_{t-1} als normalverteilt modelliert:

$$y_t | y_{t-1}, \beta_1, \beta_2, \sigma^2 \sim \mathcal{N}\left(\beta_1 + \beta_2 y_{t-1}, \sigma^2\right),$$

wobei y_{t-1} bei der Modellierung von y_t jeweils als beobachtet angenommen wird. Das Modell ist ein Spezialfall eines linearen Regressionsmodells mit

$$\boldsymbol{y} = \begin{pmatrix} y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & y_1 \\ \vdots & \vdots \\ 1 & y_{N-1} \end{pmatrix}.$$

[7] Die Dichtefunktion der multivariaten Student-t-Verteilung mit $\boldsymbol{\mu} \in \Re^d$ und positiv definiter Matrix $\boldsymbol{\Sigma} \in \Re^{d \times d}$ ist gegeben durch

$$f_{t_\nu}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)} (\nu\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \left(1 + \frac{1}{\nu}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)^{-(\nu+d)/2}.$$

Ist $\nu > 1$, dann ist $\boldsymbol{\mu}$ der Erwartungswert; ist $\nu > 2$, dann ist $\frac{\nu}{\nu-2}\boldsymbol{\Sigma}$ die Varianz-Kovarianz-Matrix.



Abb. 3 Tägliche Wechselkursdaten: der Preis eines Euros in US-Dollar im Zeitverlauf

Wir wollen dieses Modell nun mit logarithmischen Wechselkursdaten zwischen Euro und US-Dollar schätzen; y_t ist also der Logarithmus des Preises eines Euros in US-Dollar zum Zeitpunkt t. Die verwendeten Daten finden sich im R-Paket `stochvol` (R Core Team 2023; Hosszejni und Kastner 2021) und sind in Abb. 3 visualisiert.

Eine Regression mit der uninformativen Prior (9) liefert eine trivariate Posterior für β_1, β_2 und σ^2, welche in Abb. 4 dargestellt wird. Die Posterior $p(\beta_1, \beta_2, \sigma^2|\boldsymbol{y})$ beziehungsweise Monte-Carlo-Ziehungen $(\beta_1, \beta_2, \sigma^2)^{(m)}$, die mittels des oben beschriebenen Algorithmus 1 aus der Posterior gezogen wurden, bilden die Grundlage für alle weiteren Schritte statistischer Inferenz wie Schätzen, Vorhersagen und Testen, die in den folgenden Abschnitten im Detail diskutiert werden.

4 Bayesianische Punkt- und Intervallschätzung

Einer der Vorzüge der Bayes-Statistik ist die Tatsache, dass mit (Ziehungen aus) der Posterior bereits die wichtigste Ingredienz für Schätzen, Testen und Vorhersagen bereitsteht. Im Folgenden wird dies etwas näher erläutert.

4.1 Punktschätzer

Ein Punktschätzer fasst die Information in der Posterior mithilfe eines geeigneten Lagemaßes zusammen. Der geläufigste Punktschätzer ist der Erwartungswert der Posterior $\mathrm{E}(\boldsymbol{\vartheta}|\boldsymbol{y})$. Dieser ist ein Spaltenvektor der Länge $\dim \boldsymbol{\vartheta}$, definiert durch

$$\mathrm{E}(\boldsymbol{\vartheta}|\boldsymbol{y}) = \int \boldsymbol{\vartheta}\, p(\boldsymbol{\vartheta}|\boldsymbol{y}) d\boldsymbol{\vartheta}. \tag{13}$$

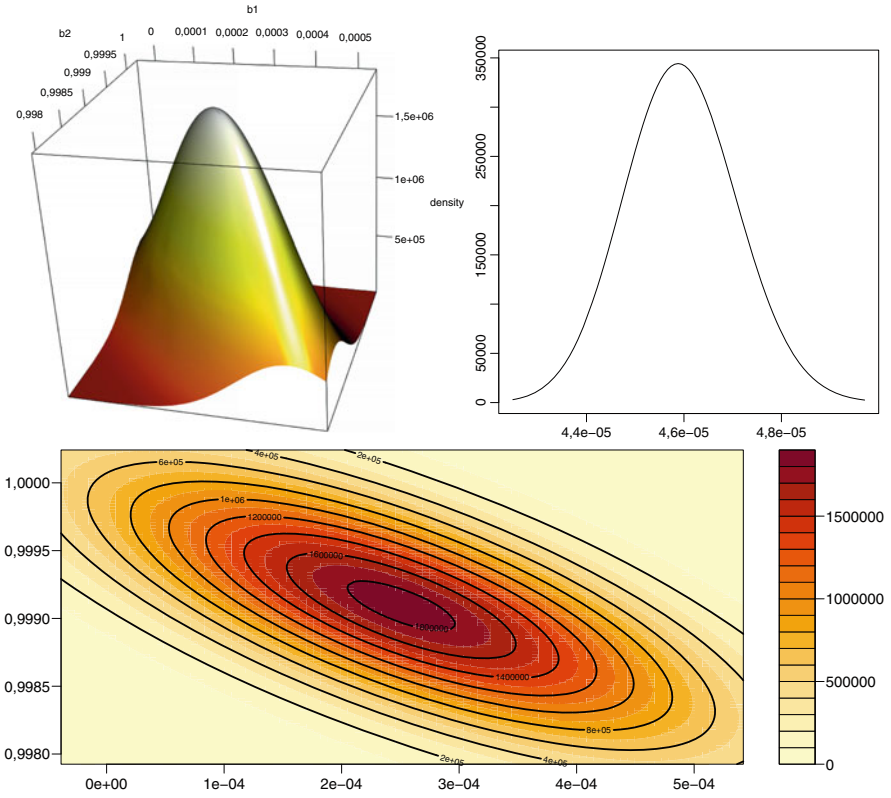

Abb. 4 Darstellung der Dichten der bivariaten marginalen Posterior der Regressionsparameter $\beta_1, \beta_2|\boldsymbol{y}$ (links oben und unten) sowie der univariaten marginalen Posterior der Fehlervarianz $\sigma^2|\boldsymbol{y}$ (rechts oben)

Für den j-ten Eintrag ϑ_j von $\boldsymbol{\vartheta}$ stellt $\mathrm{E}(\boldsymbol{\vartheta}|\boldsymbol{y})_j = \mathrm{E}(\vartheta_j|\boldsymbol{y})$ damit ein Lagemaß der univariaten marginalen Dichte $p(\vartheta_j|\boldsymbol{y})$ dar. Alternative und häufig verwendete Lagemaße sind der Modus sowie der Median der Posterior.

Eine der großen Stärken der Bayes-Statistik liegt in der ihr inhärenten Quantifizierung von (Schätz-)Unsicherheit, welche sich in der Breite der Posterior niederschlägt. Zu ihrer Berechnung wird oft die Varianz-Kovarianz-Matrix der Posterior $\mathrm{Var}(\boldsymbol{\vartheta}|\boldsymbol{y})$ herangezogen. Diese ist eine quadratische $(\dim \boldsymbol{\vartheta} \times \dim \boldsymbol{\vartheta})$-Matrix, definiert durch

$$\mathrm{Var}(\boldsymbol{\vartheta}|\boldsymbol{y}) = \int (\boldsymbol{\vartheta} - \mathrm{E}(\boldsymbol{\vartheta}|\boldsymbol{y}))(\boldsymbol{\vartheta} - \mathrm{E}(\boldsymbol{\vartheta}|\boldsymbol{y}))^\top p(\boldsymbol{\vartheta}|\boldsymbol{y}) d\boldsymbol{\vartheta}. \tag{14}$$

Jeder Hauptdiagonaleintrag $\mathrm{Var}(\boldsymbol{\vartheta}|\boldsymbol{y})_{jj} = \mathrm{Var}(\vartheta_j|\boldsymbol{y})$ stellt ein Streuungsmaß der marginalen Dichte $p(\vartheta_j|\boldsymbol{y})$ dar. Skalierungsbedingt wird jedoch meist die zugehö-

rige Standardabweichung $\mathrm{SD}\left(\vartheta_j|\boldsymbol{y}\right) = \sqrt{\mathrm{Var}(\vartheta_j|\boldsymbol{y})}$ verwendet; sie ist in der Regel leichter zu interpretieren, da sie auf derselben Skala gemessen wird wie ϑ_j selbst.

Von Interesse sind weiterhin die Korrelationen der unterschiedlichen Parameter, die sich aus den außerdiagonalen Einträgen von (14) ergeben. Für viele (insbesondere annähernd normalverteilte) Posteriors stellt der Korrelationskoeffizient

$$\mathrm{Corr}(\vartheta_j, \vartheta_k|\boldsymbol{y}) = \frac{\mathrm{Var}(\boldsymbol{\vartheta}|\boldsymbol{y})_{jk}}{\mathrm{SD}\left(\vartheta_j|\boldsymbol{y}\right)\mathrm{SD}\left(\vartheta_k|\boldsymbol{y}\right)}$$

ein gutes Abhängigkeitsmaß zwischen den Posterior-Komponenten ϑ_j und ϑ_k von $\boldsymbol{\vartheta}$ dar.

Hat man Ziehungen aus der Posterior zur Verfügung, so lassen sich obige Größen via Monte-Carlo-Integration berechnen. Ausgehend von einer hinreichend großen Stichprobe $\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(M)}$ von $p(\boldsymbol{\vartheta}|\boldsymbol{y})$ können die Integrale (13) und (14) via

$$\mathrm{E}(h(\boldsymbol{\vartheta})|\boldsymbol{y}) = \int h(\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|\boldsymbol{y})d\boldsymbol{\vartheta} \approx \frac{1}{M}\sum_{m=1}^{M} h(\boldsymbol{\vartheta}^{(m)}) \tag{15}$$

approximiert werden, wobei $h(\boldsymbol{\vartheta}) = \boldsymbol{\vartheta}$ für (13) beziehungsweise $h(\boldsymbol{\vartheta}) = (\boldsymbol{\vartheta} - \mathrm{E}(\boldsymbol{\vartheta}|\boldsymbol{y}))(\boldsymbol{\vartheta} - \mathrm{E}(\boldsymbol{\vartheta}|\boldsymbol{y}))^\top$ für (14) zu wählen ist. Das Gesetz der großen Zahlen stellt dabei sicher, dass der Stichprobenmittelwert von $h(\boldsymbol{\vartheta}^{(1)}), \ldots, h(\boldsymbol{\vartheta}^{(M)})$ für $M \to \infty$ gegen den Erwartungswert von $h(\boldsymbol{\vartheta})$ unter der Posterior konvergiert.

Solche Monte-Carlo-Schätzer sind insbesondere dann eine große Hilfe, wenn h nichtlinear ist. In diesem Fall ist das Integral in (15) häufig nicht analytisch lösbar, während sich der zugehörige Monte-Carlo-Schätzer auf nahezu triviale Weise als Mittelwert der transformierten Ziehungen $h(\boldsymbol{\vartheta}^{(1)}), \ldots, h(\boldsymbol{\vartheta}^{(M)})$ ergibt.

4.2 Intervallschätzer

Eine häufig verwendete Alternative zur Berechnung von Lage- und Streuungsmaßen ist die Angabe sogenannter *Kredibilitätsbereiche*. Diese werden oft als „(un)gleiche Zwillinge“ der aus der frequentistischen Statistik bekannten Konfidenzbereiche charakterisiert. Kredibilitätsbereiche sind Bereiche, die den unbekannten Parameter mit einer festgelegten A-posteriori-Wahrscheinlichkeit α beinhalten.

Für Analysen eindimensionaler Parameter bieten sich hierzu in natürlicher Weise Quantile der Posterior an, etwa das $\alpha/2$- und das $(1-\alpha/2)$-Quantil für ein in Bezug auf die Über- beziehungsweise Unterschreitungswahrscheinlichkeit symmetrisches α-Kredibilitätsintervall. Hat man M Ziehungen aus der j-ten Komponente der Posterior, $\vartheta_j^{(m)}$ mit $m = 1, \ldots, M$, zur Verfügung, so können diese Quantile einfach durch die empirischen Quantile der Ziehungen geschätzt werden.

Ein spezielles und oft verwendetes Kredibilitätsintervall ist das sogenannte *Highest-Posterior-Density-(HPD-)Intervall*. Es stellt unter allen α-Kredibilitäts-

Tab. 2 95 %-Kredibilitätsintervalle für das Anteilsbeispiel

	symmetrisch	Breite	HPD	Breite
Gleichverteilung	[0,18 %, 0,84 %]	0,66 %	[0,15 %, 0,79 %]	0,64 %
Jeffreys	[0,16 %, 0,80 %]	0,64 %	[0,13 %, 0,75 %]	0,61 %
Informativ	[0,17 %, 0,79 %]	0,62 %	[0,14 %, 0,74 %]	0,60 %

Tab. 3 95 %-Kredibilitätsintervalle für das Regressionsbeispiel

	symmetrisch	Breite	HPD	Breite
β_1	[−0,000097131, 0,000598763]	0,000695894	[−0,000097131, 0,000598763]	0,000695894
β_2	[0,997706526, 1,000479137]	0,002772611	[0,997706526, 1,000479137]	0,002772611
σ^2	[0,000043721, 0,000048271]	0,000004550	[0,000043683, 0,000048230]	0,000004547

intervallen das kürzestmögliche dar. Im Allgemeinen ist das HPD-Intervall nicht immer einfach zu finden; bei unimodalen Posteriors bietet es sich jedoch an, dieses numerisch zu approximieren, indem das kürzeste Intervall $\left[\vartheta_{j,(l+1)}, \vartheta_{j,(\lceil M\alpha\rceil+l)}\right]$ für $l=0,\ldots,M-\lceil M\alpha\rceil$ ausgewählt wird. Hierbei steht $\vartheta_{j,(k)}$ für die k-te Ordnungsstatistik (also die k-tgrößte) der Ziehungen $\vartheta_j^{(m)}$, $m=1,\ldots,M$. Multivariate Kredibilitätsbereiche (insbesondere multivariate HPD-Bereiche) sind im Allgemeinen schwieriger zu bestimmen. Hierzu sei aus Platzgründen auf die Literatur verwiesen, siehe etwa Tanner (1993) oder Frühwirth-Schnatter (1994).

Für das Anteilsbeispiel errechnete 95 %-Kredibilitätsintervalle und ihre jeweiligen Breiten sind in Tab. 2 zusammengefasst. Wie oben erläutert, sind zur Berechnung der symmetrischen Intervalle die entsprechenden Quantile der Beta-Posterior zu verwenden. Die HPD-Intervalle werden, wie oben beschrieben, numerisch aus den Ordnungsstatistiken der M Ziehungen ermittelt.

In der Tabelle ist deutlich erkennbar, dass für jeden der drei Priors das HDP-Intervall kürzer ist als das symmetrische Intervall. Zugleich zeigt sich, dass die Verwendung der informativen Prior die Kredibilitätsintervalle verkürzt. Andererseits wird deutlich, dass unabhängig von der gewählten Prior beziehungsweise der gewählten Berechnungsmethode die Unsicherheit in der Anteilsschätzung aufgrund der geringen Anzahl an Tests sehr groß ist.

Betrachtet man das AR(1)-Modell für die Wechselkursdaten, so ergeben sich die in Tab. 3 angeführten Intervalle für die drei Modellparameter. Das Intervall für β_1 beinhaltet 0 knapp und $\Pr(\beta_1 < 0) \approx 0{,}08$. Dies deutet darauf hin, dass möglicherweise ein Modell mit bei 0 fixiertem Achsenabschnitt besser geeignet gewesen wäre. Das Intervall für β_2 wiederum beinhaltet 1 knapp und $\Pr(\beta_2 > 1) \approx 0{,}10$. Dies deutet darauf hin, dass wir es möglicherweise mit einer nichtstationären Zeitreihe zu tun haben. Wir werden im Abschn. 6 noch näher auf diese Fragen eingehen. Da die Posteriors für die Regressionskoeffizienten symmetrisch sind, ist das HPD-Intervall und das symmetrische Intervall für diese beiden Parameter identisch. Die Fehlervarianzverteilung ist leicht rechtsschief, weswegen HPD- und symmetrisches Intervall nicht identisch (aber sehr ähnlich) ausfallen.

Abschließend sei noch angemerkt, dass sich die Interpretation von Kredibilitätsbereichen in aller Regel einfacher als die Interpretation von frequentistischen Konfidenzbereichen gestaltet. Für an Anwendungen Interessierte oder Statistik-Neulinge oft künstlich wirkende Aussagen wie: „Bei oftmaliger Wiederholung des Experiments unter gleichen Bedingungen enthalten 95 % der Intervalle den wahren Parameter“, oder häufig gebrauchte Umschreibungen dessen, etwa „Das Konfidenzintervall überdeckt mit 95 %iger Sicherheit den wahren Parameter“, sind bei Kredibilitätsbereichen nicht nötig. Sie erlauben eine direkte Wahrscheinlichkeitsaussage über die interessierende Größe selbst. Nach der Berechnung der konkreten Intervalle in Tab. 3 kann zum Beispiel die Aussage getätigt werden, dass der gesuchte Parameter β_2 mit 95 %iger Wahrscheinlichkeit zwischen 0,9977 und 1,0005 liegt.

4.3 Schätzen von nichtlinearen Funktionalen

Hat man, wie bereits erwähnt, Ziehungen aus der Posterior, so ist es ein Leichtes, Ziehungen von beliebigen Funktionalen interessierender Parameter abzuleiten. Dies ist insbesondere bei nichtlinearen Funktionalen von großem Vorteil, wie etwa beim Schätzen einer Standardabweichung (der Wurzel einer Varianz), beim Berechnen von Portfoliogewichten im Bereich der optimalen Vermögensallokation oder bei der Analyse von Zeitreihen. Im Folgenden diene als Illustration die Bestimmung der stationären Verteilung eines AR(1)-Prozesses.

Ist ein AR(1)-Prozess stationär (also $|\beta_2| < 1$), so ergibt sich die langfristige (marginale) Verteilung als

$$y \sim \mathcal{N}\left(\mu, \tilde{\sigma}^2\right), \quad \text{wobei} \quad \mu = \beta_1/(1-\beta_2) \quad \text{und} \quad \tilde{\sigma}^2 = \sigma^2/(1-\beta_2^2).$$

Wie im vorausgehenden Abschnitt diskutiert wurde und zum Beispiel aus Abb. 4 ersichtlich ist, kann die Posterior von β_2 jedoch mit positiver Wahrscheinlichkeit auch Werte größer gleich 1 annehmen; in diesen Fällen existiert keine langfristige Verteilung. Eine Möglichkeit, dieses Problem zu umgehen, ergibt sich durch Einschränkung der Prior auf Werte innerhalb von $(-1, 1)$, etwa durch die Modifikation der Prior auf eine Gleichverteilung im Intervall $[-0{,}9999, 0{,}9999]$. Aus praktischer Sicht kann dies mit den bereits in Abschn. 3.2 erzeugten Ziehungen erzielt werden, indem einfach alle Ziehungen außerhalb des erlaubten Intervalls verworfen werden – ein erneutes Schätzen ist somit nicht nötig.

Um langfristigen Erwartungswert und Standardabweichung der Wechselkursdaten basierend auf den Ziehungen des oben beschriebenen Algorithmus zu schätzen, kann man wie folgt vorgehen.

1. Für $m = 1, \ldots, M$, verwerfe die m-ten Ziehungen aller Parameter, falls die m-te Ziehung von β_2 betragsmäßig größer als 0,9999 ist.

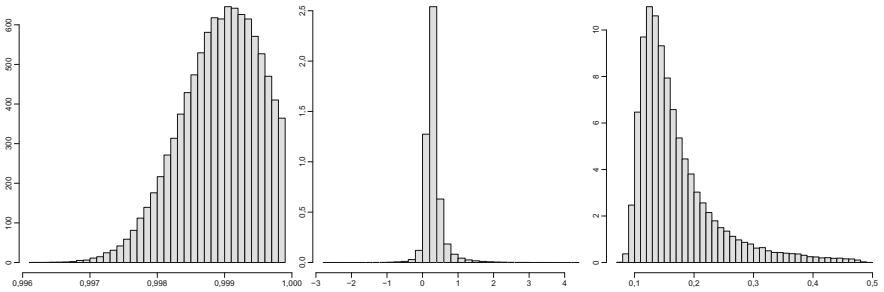

Abb. 5 Links: Histogramm von Ziehungen aus $p(\beta_2|\boldsymbol{y})$ unter der stationaritätsforcierenden gleichverteilten Prior $p(\beta_2) \sim \mathscr{U}[-0{,}9999, 0{,}9999]$. Mitte: Histogramm von Ziehungen aus der Posterior des langfristigen Erwartungswerts $p(\mu|\boldsymbol{y}) = p(\beta_1/(1-\beta_2)|\boldsymbol{y})$. Rechts: Histogramm von Ziehungen aus der Posterior der langfristigen Standardabweichung $p(\tilde{\sigma}|\boldsymbol{y}) = p\left(\sqrt{\sigma^2/(1-\beta_2^2)}\middle|\boldsymbol{y}\right)$

2. Für die verbleibenden Ziehungen berechne $\mu^{(m)} = \beta_1^{(m)}/(1-\beta_2^{(m)})$ und $\tilde{\sigma}^{(m)} = \sqrt{\sigma^{2,(m)}/(1-(\beta_2^{(m)})^2)}$.

Abb. 5 illustriert die „neue“ Posterior von β_2 und die dadurch implizierten marginalen Posteriors von μ und $\tilde{\sigma}$. Zu beachten ist insbesondere die Schiefe und die augenscheinlich sehr hohe Kurtosis des langfristigen Erwartungswerts sowie die klar erkennbare Schiefe der langfristigen Standardabweichung.

5 Vorhersagen mithilfe der bayesianischen Statistik

Eine der wichtigsten Anwendungen statistischer Inferenz ist die Vorhersage zukünftiger Ereignisse. Im Folgenden beleuchten wir die Konstruktion von Vorhersageverteilungen und sich daraus ergebenden Punkt- und Intervallvorhersagen aus bayesianischer Sicht.

5.1 Vorhersageverteilungen

Ziel sei die Vorhersage eines zukünftigen Datenpunkts y_f bei gegebenen Beobachtungen $\boldsymbol{y} = (y_1, \ldots, y_N)$ einer Zufallsvariable Y. Der genaue Wert von y_f ist unbekannt, und die Vorhersage wird durch die Bestimmung der *Vorhersageverteilung* (auch: *Posterior-Predictive*) von y_f möglich. Im Bayes-Kontext sind nicht nur Punktvorhersagen von y_f von Interesse. Vielmehr ist es möglich, die gesamte Dichtefunktion $p(y_f|\boldsymbol{y})$ zu bestimmen.

Dazu ist natürlich ein Modell vonnöten, etwa das eingangs diskutierte Modell für Anteile $Y \sim \text{BiNom}(N, \vartheta)$. Wäre ϑ (über die ohnehin als gegeben angenommene

Stichprobengröße N hinaus) bekannt, so ließe sich die Vorhersage direkt aus der Dichte $p(y_f|N, \vartheta)$ der BiNom (N, ϑ)-Verteilung bestimmen. Da jedoch ϑ nicht bekannt ist, wird in nichtbayesianischen Ansätzen ϑ in der Regel durch einen Schätzwert $\hat{\vartheta}$ ersetzt und y_f aus BiNom $(N, \hat{\vartheta})$ bestimmt. Bei diesem Zugang wird also die Schätzunsicherheit vernachlässigt; dies wiederum impliziert jedoch, dass Vorhersageintervalle in der Regel zu eng sind und nur asymptotisch korrekte Überdeckungswahrscheinlichkeiten liefern. Der Bayes-Ansatz hingegen bezieht die Schätzunsicherheit durch die Einbeziehung der Posterior in natürlicher und kohärenter Weise ein.

Unter der Annahme, dass das Modell vom unbekannten Parameter(-vektor) $\boldsymbol{\vartheta}$ abhängt, lässt sich $p(y_f|\boldsymbol{y})$ auf natürliche Weise als marginale Verteilung der gemeinsamen Verteilung $p(y_f, \boldsymbol{\vartheta}|\boldsymbol{y})$ schreiben:

$$p(y_f|\boldsymbol{y}) = \int p(y_f, \boldsymbol{\vartheta}|\boldsymbol{y}) d\boldsymbol{\vartheta}.$$

Außerdem lässt sich $p(y_f, \boldsymbol{\vartheta}|\boldsymbol{y})$ als das Produkt einer bedingten und einer marginalen Verteilung schreiben:

$$p(y_f, \boldsymbol{\vartheta}|\boldsymbol{y}) = p(y_f|\boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}|\boldsymbol{y}), \tag{16}$$

vorausgesetzt, die Stichprobenverteilung $p(y_f|\boldsymbol{\vartheta})$ ist unabhängig von den beobachteten Realisierungen $\boldsymbol{y}$. In diesem Fall folgt aus (16), dass sich die Vorhersageverteilung einer zukünftigen Beobachtung y_f bedingt auf die Beobachtungen $\boldsymbol{y}$ in folgender Weise schreiben lässt:

$$p(y_f|\boldsymbol{y}) = \int p(y_f|\boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}|\boldsymbol{y}) d\boldsymbol{\vartheta}. \tag{17}$$

Die Stichprobenverteilung $p(y_f|\boldsymbol{\vartheta})$ geht also gemeinsam mit der Posterior $p(\boldsymbol{\vartheta}|\boldsymbol{y})$ in die Bestimmung der Vorhersageverteilung ein. Gleichung (17) ist insofern von großer praktischer Bedeutung, weil klar wird, dass die Posterior $p(\boldsymbol{\vartheta}|\boldsymbol{y})$ auch für die Vorhersage zukünftiger Werte suffizient ist. Der Wert des Integrals in (17) kann jedoch nur in wenigen Fällen analytisch bestimmt werden; im Allgemeinen werden hierzu Monte-Carlo-Methoden verwendet, die sich bereits bei der Schätzung nichtlinearer Funktionale als nützlich erwiesen haben.

Für eine Sequenz von Vorhersagen $\boldsymbol{y}_f = (y_{f,1}, \ldots, y_{f,H})$ der Länge H, gegeben Daten $\boldsymbol{y}$ und ein Modell, welches von unbekannten Parameter $\boldsymbol{\vartheta}$ abhängt, errechnet sich die Vorhersageverteilung ganz analog zum univariaten Fall:

$$p(\boldsymbol{y}_f|\boldsymbol{y}) = \int p(\boldsymbol{y}_f|\boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}|\boldsymbol{y}) d\boldsymbol{\vartheta}. \tag{18}$$

5.2 Vorhersageverteilung der Binomialverteilung

Betrachten wir zur Exemplifizierung wieder unser Anteilsbeispiel. Will man etwa für H zukünftige Subjekte $\boldsymbol{y}_f = (y_{f,1}, \ldots, y_{f,H})$ vorhersagen, wie viele Subjekte $S_f = \sum_{h=1}^{H} y_{f,h}$ die interessierende Eigenschaft besitzen, so kann man für $k = 0, \ldots, H$ die Vorhersageverteilung bestimmen, evaluieren und natürlich auch graphisch darstellen, wie Abb. 6 für das einführende Beispiel der Prävalenzstudie mit $H = 100$ ungetesteten Personen illustriert. Die dargestellten Wahrscheinlichkeiten lassen sich für dieses Beispiel explizit berechnen:

$$\begin{aligned}\Pr(S_f = k|\boldsymbol{y}) &= \int \Pr(S_f = k|\vartheta)\, p(\vartheta|\boldsymbol{y})\, d\vartheta \\ &= \int \binom{H}{k} \vartheta^k (1-\vartheta)^{H-k} \frac{1}{B(a_N, b_N)} \vartheta^{a_N - 1} (1-\vartheta)^{b_N - 1} d\vartheta \\ &= \binom{H}{k} \frac{B(a_N + k, b_N + H - k)}{B(a_N, b_N)}.\end{aligned}$$

In diesem Beispiel ergibt sich als exakte Vorhersageverteilung die sogenannte Beta-Binomialverteilung, welche sich für wachsende Stichprobengröße N recht schnell der approximierenden Binomialverteilung nähert, sodass im vorliegenden Fall die aus der Anteilsschätzung resultierende Unsicherheit in Bezug auf die Vorhersageunsicherheit gering und in Abb. 6 nur auf der logarithmischen Skala deutlich sichtbar ist. Wäre es etwa von Interesse, die Wahrscheinlichkeit abzuschätzen, dass sich unter 100 nicht getesteten Personen mehr als zwei infizierte Personen befinden (sie beträgt etwa 0,0119), so würde die approximierende Binomialverteilung diese Wahrscheinlichkeit mit etwa 0,0091 um immerhin etwa 23 % unterschätzen.

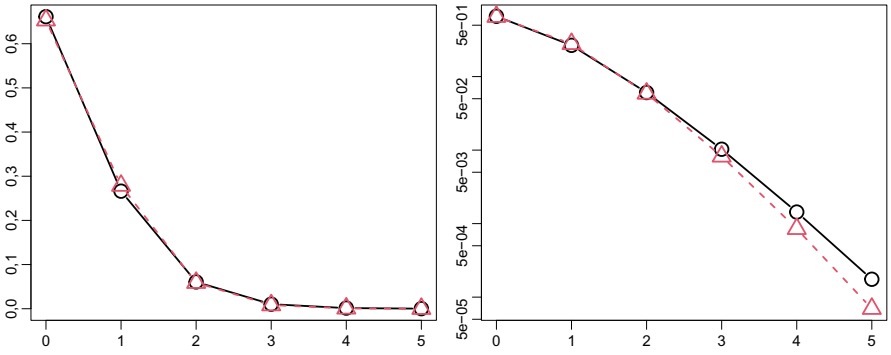

Abb. 6 Vorhersage der Fallzahlen bei $H = 100$ für $k = 0, 1, \ldots, 5$. Schwarz durchgezogen (Kreise): Vorhersageverteilung. Rot strichliert: approximierende Binomialverteilung BiNom $(H, \hat{\vartheta})$ mit $\hat{\vartheta} = \mathrm{E}(\vartheta|\boldsymbol{y}) = a_N/(a_N + b_N)$. Links: lineare Skala. Rechts: logarithmische Skala

Andererseits ergibt sich beispielsweise für die Wahrscheinlichkeit, dass gar keine dieser Personen infiziert ist, mit etwa 0,6612 beziehungsweise 0,6531 annähernd derselbe Wert, was einem relativen Fehler von etwas mehr als 1 % entspricht.

5.3 Vorhersageverteilung im Regressionsmodell

Eine häufige Anwendung des in Abschn. 3 vorgestellten Regressionsmodells ist die Vorhersage einer Responsevariablen y_{N+1} für eine neue Realisation von erklärenden Variablen, im Folgenden zusammengefasst im Zeilenvektor $\boldsymbol{x}_{N+1}$. Unter der Annahme von homoskedastisch normalverteilten Fehlern, der Prior (9) und der Notation von (10) ergibt sich als Vorhersagedichte

$$p(y_{N+1}|\boldsymbol{y}) = \int f_N\left(y_{N+1}; \boldsymbol{x}_{N+1}\boldsymbol{b}_N, (\boldsymbol{x}_{N+1}\boldsymbol{B}_N\boldsymbol{x}_{N+1}^\top + 1)\sigma^2\right) f_{IG}(\sigma^2; c_N, C_N)d\sigma^2,$$

wobei f_N die auf Seite 7 definierte Dichtefunktion der Normalverteilung und f_{IG} die auf Seite 9 definierte Dichtefunktion der inversen Gammaverteilung bezeichnet. Dieses Integral ist analytisch lösbar und führt auf eine Student-t-Verteilung:

$$y_{N+1}|\boldsymbol{y} \sim t_{2c_N}\left(\boldsymbol{x}_{N+1}\boldsymbol{b}_N, (\boldsymbol{x}_{N+1}\boldsymbol{B}_N\boldsymbol{x}_{N+1}^\top + 1)C_N/c_N\right). \tag{19}$$

Die Vorhersageverteilung ist in Abb. 7 für einen kleinen Ausschnitt dargestellt. Viele der beobachteten Datenpunkte befinden sich in einem Bereich mit hoher Dichte, einige wenige reißen jedoch aus.

Im AR(1)-Modell lässt sich also unmittelbar für eine gegebene Zeitreihe $\boldsymbol{y} = (y_1, \ldots, y_N)$ eine Einschrittvorhersage für y_{N+1} durchführen, da der künftige

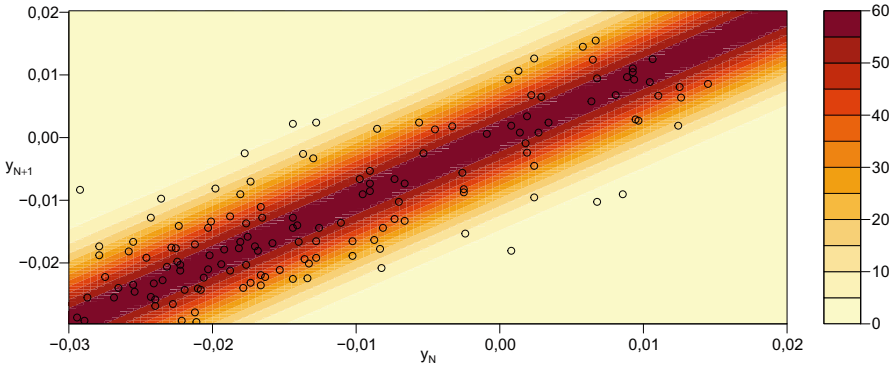

Abb. 7 Visualisierung der Vorhersagedichte $p(y_{N+1}|\boldsymbol{y})$ für $y_N \in [-0{,}03, 0{,}02]$. Beobachtete Datenpunkte in diesem Intervall sind mit schwarzen Kreisen eingezeichnet

Prädiktor $\boldsymbol{x}_{N+1} = (1, y_N)$ beobachtbar ist. Für Mehrschrittvorhersagen ergibt sich allerdings das Problem der dynamischen Vorhersage, da die künftigen Prädiktoren $\boldsymbol{x}_{N+2} = (1, y_{N+1})$, $\boldsymbol{x}_{N+3} = (1, y_{N+2})$, et cetera, von den künftigen Werten der Zeitreihe abhängen. Die entsprechenden Vorhersagedichten können mathematisch nur durch ein Integral charakterisiert werden, das nicht analytisch lösbar ist. Auch in diesem Fall ermöglichen Monte-Carlo-Methoden eine einfache Approximation der Vorhersagedichten.

5.4 Vorhersage via Monte-Carlo

Möchte man mittels des in Abschn. 3.3 vorgestellten AR(1)-Modells den Verlauf der nächsten H Tage $\boldsymbol{y}_{N+1}^{N+H} := (y_{N+1}, \ldots, y_{N+H})$ vorhersagen, so gilt es, das folgende Integral zu lösen:

$$p(\boldsymbol{y}_{N+1}^{N+H}|\boldsymbol{y}) = \int p(\boldsymbol{y}_{N+1}^{N+H}|\boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) d(\boldsymbol{\beta}, \sigma^2). \tag{20}$$

Diese Dichte ist nicht in geschlossener Form verfügbar und analytische Integration wie in (19) ist unmöglich. Allerdings ist es ein Leichtes, Ziehungen aus (20) mithilfe der in Abschn. 3.2 beziehungsweise Abschn. 4.3 generierten Ziehungen aus der Posterior zu erzeugen.

Definiere $\boldsymbol{x}_{N+1}^{(m)} := (1, y_N)$ für $m = 1, \ldots, M$ und wiederhole für $h = 1, \ldots, H$:

1. Ziehe $y_{N+h}^{(m)}$ aus $\mathcal{N}\left(\boldsymbol{x}_{N+h}^{(m)}\boldsymbol{\beta}^{(m)}, \sigma^{2,(m)}\right)$, $m = 1, \ldots, M$.
2. Definiere $\boldsymbol{x}_{N+h+1}^{(m)} := (1, y_{N+h}^{(m)})$, $m = 1, \ldots, M$.

Die so erlangten Ziehungen sind Ziehungen aus der multivariaten Vorhersageverteilung für die nächsten H Zeitpunkte. In Abb. 8 sind dadurch geschätzte

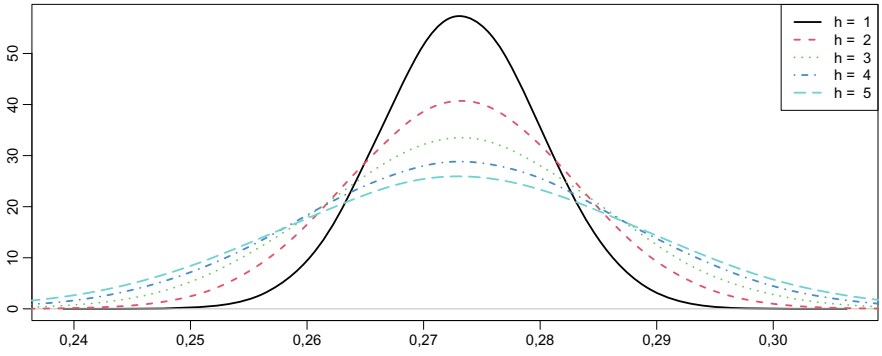

Abb. 8 AR(1)-Modell für Wechselkurse: geschätzte Vorhersagedichten für die nächsten fünf Zeitpunkte $N + h$, $h = 1, \ldots, 5$

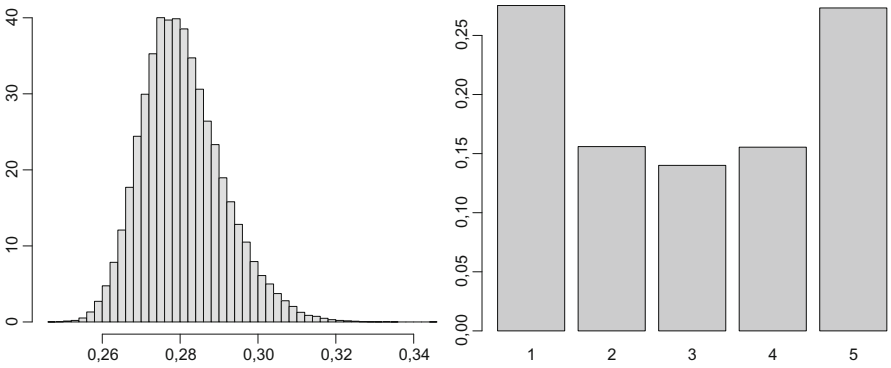

Abb. 9 Links: Ziehungen aus der Verteilung des für die nächsten fünf Zeitpunkte $N+1, \ldots, N+5$ vorhergesagten Maximums. Rechts: Wahrscheinlichkeitsverteilung, wann dieses Maximum angenommen wird

marginale Dichten für $h = 1, \ldots, 5$ visualisiert. Hier wird deutlich, wie die Vorhersageunsicherheit mit steigendem Vorhersagehorizont h größer wird.

Die Ziehungen können auch zur Vorhersage (nichtlinearer) Transformationen $g(\boldsymbol{y}_{N+1}^{N+H})$ von $\boldsymbol{y}_{N+1}^{N+H}$, wie etwa des Maximums U von fünf in der Zukunft liegenden Werten $y_{N+1}, \ldots, y_{N+5}$, verwendet werden. Eine Stichprobe dieses Maximums ist durch

$$U^{(m)} = \max\left(y_{N+1}^{(m)}, \ldots, y_{N+5}^{(m)}\right)$$

gegeben und in Abb. 9 visualisiert.

5.5 Vorhersageintervalle und Punktvorhersagen

In Analogie zu Punkt- und Intervallschätzern, die sich aus der Posterior ableiten lassen, werden Punkt- und Intervallvorhersagen aus der Vorhersageverteilung konstruiert. Eine Vorhersageregion ist demnach eine Teilmenge S_α des Stichprobenraumes $\mathscr{Y}$, der den unbekannten Wert y_f mit Wahrscheinlichkeit α enthält:

$$\Pr(y_f \in S_\alpha | \boldsymbol{y}) = \int_{S_\alpha} p(y_f | \boldsymbol{y}) dy_f = \alpha. \tag{21}$$

Wie bei der Parameterschätzung kann man auch hier nach Vorhersageintervallen mit kleinstmöglicher Länge suchen.

Im Rahmen der bayesianischen Entscheidungstheorie lässt sich ein Punktschätzer aus der Vorhersageverteilung konstruieren, der eine bestimmte Verlustfunktion minimiert. Sei $R(\hat{y}_f, y_f)$ eine Funktion, die den Verlust der Vorhersage von y_f

durch den Punktschätzer $\hat{y}_f$ quantifiziert. Daraus lässt sich der erwartete Verlust in Bezug auf die Vorhersageverteilung berechnen:

$$\mathrm{E}(R(\hat{y}_f, y_f)|\boldsymbol{y}) = \int R(\hat{y}_f, y_f) p(y_f|\boldsymbol{y}) dy_f.$$

Als allgemeines Prinzip kann man diejenige Punktvorhersage $\hat{y}_f$ wählen, welche den erwarteten Verlust minimiert. Aus der quadratischen Verlustfunktion $R(\hat{y}_f, y_f) = (\hat{y}_f - y_f)^2$ ergibt sich beispielsweise der Erwartungswert $\mathrm{E}(y_f|\boldsymbol{y})$ der Vorhersageverteilung als optimale Wahl, während für die absolute Verlustfunktion $R(\hat{y}_f, y_f) = |\hat{y}_f - y_f|$ der Median und für die 0-1-Verlustfunktion $R(\hat{y}_f, y_f) = 1 - I\{\hat{y}_f = y_f\}$ der Modus der Vorhersageverteilung zu wählen ist.

6 Bayesianisches Testen von Hypothesen und Modellwahl

Werden für einen Datensatz mehrere Modelle geschätzt, so stellt sich die natürliche Frage: Welches dieser Modelle passt am besten zu den Daten? Im Folgenden soll kurz die bayesianische Antwort auf Fragen dieser Art skizziert und exemplifiziert werden.

6.1 Modellwahrscheinlichkeiten

Der klassische Vergleich zweier Hypothesen im Bayes-Kontext geht auf Jeffreys (1939) zurück, welcher sich mit Vorhersagen zweier konkurrierender wissenschaftlicher Theorien beschäftigte. Kann jede dieser zwei Theorien durch je ein stochastisches Datenmodell $p(\boldsymbol{y}|\mathscr{M}_1)$ beziehungsweise $p(\boldsymbol{y}|\mathscr{M}_2)$ dargestellt werden, so kann das Bayes-Theorem zur Berechnung der Wahrscheinlichkeit verwendet werden, mit der jede dieser beiden Theorien zutrifft.[8]

Zur Illustration diene der Vergleich zweier *einfacher* Hypothesen. Damit sind Hypothesen gemeint, bei denen keine unbekannten Parameter mehr verbleiben, sobald eine der beiden Hypothesen als wahr angenommen wird. Nehmen wir an, dass laut Modell $\mathscr{M}_1$ der wahre Wert von $\boldsymbol{\vartheta}$ gleich $\boldsymbol{\vartheta}_1$ ist, während laut Modell $\mathscr{M}_2$ $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_2$ gilt. Sei weiterhin $\mathscr{M}$ ein zufälliger dichotomer Indikator, der den Wert l mit $l \in \{1, 2\}$ annimmt, wenn Modell $\mathscr{M}_l$ das wahre Modell darstellt.

Bayesianische Inferenz basiert auf der Spezifikation der gemeinsamen Verteilung von $(\boldsymbol{y}, \mathscr{M})$, deren Dichte die folgende Zerlegung zulässt:

$$p(\boldsymbol{y}, \mathscr{M}) = p(\boldsymbol{y}|\mathscr{M}) p(\mathscr{M}).$$

[8] Eine wichtige Einschränkung sei hier nicht unerwähnt: Für den in diesem Abschnitt vorgestellten Zugang zur Modellwahl ist es notwendig, sogenannte *eigentliche (proper)* Priors, also solche, die eine gültige Wahrscheinlichkeitsdichte definieren, zu verwenden. Die für die in Abschn. 3.2 vorgestellte Regression verwendete uninformative Prior (9) ist hier beispielsweise nicht geeignet.

Hier bezeichnet $p(\mathscr{M})$ die A-priori-Wahrscheinlichkeit für die beiden Hypothesen und $p(\boldsymbol{y}|\mathscr{M})$ die Likelihood von $\boldsymbol{y}$ gegeben $\mathscr{M}$. Es folgt in natürlicher Weise:

$$p(\mathscr{M}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathscr{M})p(\mathscr{M})}{p(\boldsymbol{y})}, \tag{22}$$

mit $p(\boldsymbol{y}) = p(\boldsymbol{y}|\mathscr{M}=1)p(\mathscr{M}=1) + p(\boldsymbol{y}|\mathscr{M}=2)p(\mathscr{M}=2)$ (Satz der totalen Wahrscheinlichkeit), wobei $p(\mathscr{M}=1|\boldsymbol{y})$ beziehungsweise $p(\mathscr{M}=2|\boldsymbol{y})$ die A-posteriori-Wahrscheinlichkeit der jeweiligen Hypothesen bezeichnet. Je näher die A-posteriori-Wahrscheinlichkeit $p(\mathscr{M}=1|\boldsymbol{y})$ bei 1 liegt, desto wahrscheinlicher ist $\mathscr{M}_1$ im Vergleich zu $\mathscr{M}_2$ (und vice versa).

Diese Idee lässt sich leicht verallgemeinern. Seien $\boldsymbol{\vartheta}_1$ und $\boldsymbol{\vartheta}_2$ die unbekannten Parameter unter den Hypothesen $\mathscr{M}_1$ beziehungsweise $\mathscr{M}_2$. Bayes-Inferenz basiert auf der Spezifikation der gemeinsamen Verteilung von $(\boldsymbol{y}, \mathscr{M}, \boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2)$. Es gilt:

$$p(\boldsymbol{y}, \mathscr{M}, \boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2) = p(\boldsymbol{y}|\mathscr{M}, \boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2)p(\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2|\mathscr{M})p(\mathscr{M}),$$

wobei $p(\boldsymbol{y}|\mathscr{M}, \boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2)$ die Likelihood der Daten $\boldsymbol{y}$ unter der durch $\mathscr{M}$ beschriebenen Hypothese beschreibt. Unter Modell $\mathscr{M}_1$ ($\mathscr{M}=1$) gilt $p(\boldsymbol{y}|\mathscr{M}=1, \boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2) = p(\boldsymbol{y}|\mathscr{M}_1, \boldsymbol{\vartheta}_1)$; unter Modell $\mathscr{M}_2$ ($\mathscr{M}=2$) gilt $p(\boldsymbol{y}|\mathscr{M}=2, \boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2) = p(\boldsymbol{y}|\mathscr{M}_2, \boldsymbol{\vartheta}_2)$. Die auf die Hypothesen bedingte Prior für $(\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2)$ ist $p(\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2|\mathscr{M})$, und $p(\mathscr{M}=l) = \Pr(\mathscr{M}_l)$ für $l \in \{1, 2\}$ sind die Priors für die Modelle (Hypothesen) selbst.

Der Satz von Bayes impliziert:

$$p(\mathscr{M}, \boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathscr{M}, \boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2)p(\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2|\mathscr{M})p(\mathscr{M})}{p(\boldsymbol{y})}, \tag{23}$$

wobei die Normalisierungskonstante $p(\boldsymbol{y})$ weiter unten in (24) definiert wird. Die A-posteriori-Wahrscheinlichkeiten $\Pr(\mathscr{M}_1|\boldsymbol{y})$ und $\Pr(\mathscr{M}_2|\boldsymbol{y})$ der Hypothesen $\mathscr{M}_1$ und $\mathscr{M}_2$ lassen sich aus (23) durch Marginalisierung bestimmen:

$$p(\mathscr{M}=1|\boldsymbol{y}) = \frac{\Pr(\mathscr{M}_1)}{p(\boldsymbol{y})} \int p(\boldsymbol{y}|\boldsymbol{\vartheta}_1, \mathscr{M}_1)p(\boldsymbol{\vartheta}_1|\mathscr{M}_1) \int p(\boldsymbol{\vartheta}_2|\boldsymbol{\vartheta}_1, \mathscr{M}_1)d\boldsymbol{\vartheta}_2 d\boldsymbol{\vartheta}_1,$$

$$p(\mathscr{M}=2|\boldsymbol{y}) = \frac{\Pr(\mathscr{M}_2)}{p(\boldsymbol{y})} \int p(\boldsymbol{y}|\boldsymbol{\vartheta}_2, \mathscr{M}_2)p(\boldsymbol{\vartheta}_2|\mathscr{M}_2) \int p(\boldsymbol{\vartheta}_1|\boldsymbol{\vartheta}_2, \mathscr{M}_2)d\boldsymbol{\vartheta}_1 d\boldsymbol{\vartheta}_2.$$

Also gilt

$$\Pr(\mathscr{M}_1|\boldsymbol{y}) \propto \Pr(\mathscr{M}_1)p(\boldsymbol{y}|\mathscr{M}_1), \quad p(\boldsymbol{y}|\mathscr{M}_1) = \int p(\boldsymbol{y}|\boldsymbol{\vartheta}_1, \mathscr{M}_1)p(\boldsymbol{\vartheta}_1|\mathscr{M}_1)d\boldsymbol{\vartheta}_1,$$

$$\Pr(\mathscr{M}_2|\boldsymbol{y}) \propto \Pr(\mathscr{M}_2)p(\boldsymbol{y}|\mathscr{M}_2), \quad p(\boldsymbol{y}|\mathscr{M}_2) = \int p(\boldsymbol{y}|\boldsymbol{\vartheta}_2, \mathscr{M}_2)p(\boldsymbol{\vartheta}_2|\mathscr{M}_2)d\boldsymbol{\vartheta}_2,$$

unter der Bedingung, dass die Normalisierungskonstante

$$p(\boldsymbol{y}) = \Pr(\mathscr{M}_1)p(\boldsymbol{y}|\mathscr{M}_1) + \Pr(\mathscr{M}_2)p(\boldsymbol{y}|\mathscr{M}_2) \tag{24}$$

endlich ist. Dieses Resultat zeigt deutlich, wie die A-priori-Wahrscheinlichkeiten für jede der Hypothesen $\mathscr{M}_l$, $l \in \{1, 2\}$, durch das Beobachten der Daten $\boldsymbol{y}$ aktualisiert werden. Die Daten kommen dabei über die sogenannte *Prior-Predictive-Probability* (auch *Marginal Likelihood*) $p(\boldsymbol{y}|\mathscr{M}_l)$ zur Geltung. Im Kontext von Modellvergleichen wird die Marginal Likelihood $p(\boldsymbol{y}|\mathscr{M}_l)$ typischerweise als *Modell-Likelihood* bezeichnet.

6.2 Bayes-Faktoren

Die wohl häufigste Form des Vergleiches zweier Hypothesen erfolgt durch Berechnen des Verhältnisses ihrer auf die Daten bedingten Wahrscheinlichkeiten. Die Evidenz für $\mathscr{M}_1$ (gegenüber $\mathscr{M}_2$) ist etwa definiert durch das Verhältnis der A-posteriori-Wahrscheinlichkeiten:

$$\frac{\Pr(\mathscr{M}_1|\boldsymbol{y})}{\Pr(\mathscr{M}_2|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\mathscr{M}_1)}{p(\boldsymbol{y}|\mathscr{M}_2)} \frac{\Pr(\mathscr{M}_1)}{\Pr(\mathscr{M}_2)}. \tag{25}$$

In dieser Gleichung findet sich einerseits der Term $\Pr(\mathscr{M}_1)/\Pr(\mathscr{M}_2)$, also das Verhältnis der A-priori-Wahrscheinlichkeiten, sowie andererseits der sogenannte Bayes-Faktor $B_{12} = p(\boldsymbol{y}|\mathscr{M}_1)/p(\boldsymbol{y}|\mathscr{M}_2)$, definiert als das Verhältnis der Marginal Likelihoods. Der Bayes-Faktor gibt also an, wie stark die Daten $\boldsymbol{y}$ die A-priori-Wahrscheinlichkeit für $\mathscr{M}_1$ (im Verhältnis zu $\mathscr{M}_2$) verändern. So bedeutet etwa $B_{12} > 1$, dass $\mathscr{M}_1$ im Licht der beobachteten Daten verhältnismäßig plausibler ist als $\mathscr{M}_2$. Nimmt man für beide Modelle dieselbe A-priori-Wahrscheinlichkeit an, so reduziert sich das Verhältnis der A-posteriori-Wahrscheinlichkeiten auf den Bayes-Faktor und somit auf den Quotienten der Marginal Likelihoods.[9]

Zur Interpretation der Bayes-Faktoren (Jeffreys 1939; Kass und Raftery 1995) kann etwa Tab. 4 herangezogen werden. Liegt, bei gleichen A-priori-Wahrscheinlichkeiten, $\log B_{12}$ zwischen -1 und 1 (und B_{12} dementsprechend zwischen etwa 0,37 und 2,71), so sind beide Hypothesen recht wahrscheinlich. Liegt jedoch beispielsweise $\log B_{12}$ bei 3, so entspricht dies einer 95 %igen Wahrscheinlichkeit für Modell $\mathscr{M}_1$, und $\log B_{12} \approx -5$ entspricht einer 99 %igen Wahrscheinlichkeit für Modell $\mathscr{M}_2$.

Einer der wesentlichen Unterschiede zum frequentistischen Testen besteht darin, dass nicht nur Evidenz *gegen* eine zu verwerfende Hypothese (die Nullhypothese) gesammelt wird, sondern alle Hypothesen durch Festlegen der jeweiligen A-priori-

[9] Die formale Ähnlichkeit zum „klassischen" Likelihood-Quotienten-Test ist augenscheinlich. Zu beachten ist jedoch, dass bei diesem die Maxima der Likelihoods, nicht die (gewichteten) „Durchschnittslikelihoods" in ein Verhältnis gesetzt werden.

Tab. 4 Log-Bayes-Faktoren und Hypothesenwahrscheinlichkeiten bei gleichen A-priori-Wahrscheinlichkeiten

$\log B_{12}$	$\Pr(\mathscr{M}_1 \mid \boldsymbol{y})$	$\Pr(\mathscr{M}_2 \mid \boldsymbol{y})$
−7	0,000911	0,999
−6	0,00247	0,998
−5	0,00669	0,993
−4	0,018	0,982
−3	0,0474	0,953
−2	0,119	0,881
−1	0,269	0,731
0	0,5	0,5
1	0,731	0,269
2	0,881	0,119
3	0,953	0,0474
4	0,982	0,018
5	0,993	0,00669
6	0,998	0,00247
7	0,999	0,000911

Wahrscheinlichkeiten gleichermaßen in Betracht gezogen werden. Während also beim klassischen Testen die Nullhypothese konstruktionsbedingt nie (auch nicht probabilistisch) „bewiesen“ werden kann, so liefert im Gegensatz dazu der Bayes-Ansatz A-posteriori-Wahrscheinlichkeiten für beide (im Allgemeinen *sämtliche*) in Betracht gezogenen Hypothesen. Geht etwa $\log B_{12}$ im Lichte neuer Daten gegen *minus* unendlich, dann wird $\mathscr{M}_1$ zugunsten von $\mathscr{M}_2$ verworfen, während $\mathscr{M}_2$ zugunsten von $\mathscr{M}_1$ zu verwerfen ist, wenn $\log B_{12}$ im Lichte neuer Daten gegen *plus* unendlich geht. Weiterführendes zum Themenkomplex „frequentistisches und bayesianisches Testen“ findet sich etwa in Berger und Sellke (1987) und Berger und Delampady (1987).

Zum Abschluss sei noch erwähnt, dass Modellwahl via größtmöglicher Likelihood und Modellwahl via größtmöglicher Marginal Likelihood im Allgemeinen andere Ergebnisse liefern. So ist etwa bei geschachtelten Modellen (man denke beispielsweise an Variablenselektion im Kontext von Regressionsmodellen) mit steigender Anzahl an Parametern ein größeres Maximum der Likelihood zu erwarten. Ist dieses Maximum jedoch nur unwesentlich größer, so wird die zugehörige Marginal Likelihood kleiner sein (Atkinson 1978; Smith und Spiegelhalter 1980). In anderen Worten: Sind zwei geschachtelte Modelle, ein einfaches und ein komplexeres, vergleichbar gut geeignet, um einen gewissen Datensatz zu beschreiben, so wird bayesianische Modellwahl eher das einfache Modell bevorzugen, obwohl das komplexere Modell in aller Regel ein größeres Maximum der Likelihood aufweist. Das ist auch der Grund, warum Strafterme für die Anzahl der Parameter, wie etwa von Akaike (1974) oder Schwarz (1978) vorgeschlagen, im Bayes-Kontext normalerweise nicht explizit benötigt – sondern sogar implizit berücksichtigt – werden.

6.3 Dichtequotienten nach Savage-Dickey

Seien im Folgenden zwei geschachtelte Modelle $\mathscr{M}_1$ und $\mathscr{M}_2$ gegeben und sei ohne Beschränkung der Allgemeinheit $\mathscr{M}_1$ das einfachere (sparsamere) Modell. Die unbekannten Parameter werden mit $\boldsymbol{\vartheta}_1 \in \Theta_1$ sowie $\boldsymbol{\vartheta}_2 \in \Theta_2$ bezeichnet, die jeweiligen Likelihoods der Daten $\boldsymbol{y}$ unter Modell $\mathscr{M}_1$ beziehungsweise $\mathscr{M}_2$ mit $p(\boldsymbol{y}|\mathscr{M}_1, \boldsymbol{\vartheta}_1)$ sowie $p(\boldsymbol{y}|\mathscr{M}_2, \boldsymbol{\vartheta}_2)$. Da die Modelle geschachtelt sind, lässt sich $\boldsymbol{\vartheta}_2$ als $\boldsymbol{\vartheta}_2 = (\boldsymbol{\vartheta}_1, \boldsymbol{\psi})$ schreiben, wobei $\boldsymbol{\vartheta}_1$ der den beiden Modellen gemeine Parameter ist. Unter Modell $\mathscr{M}_2$ ist der Parameter $\boldsymbol{\psi}$ unbekannt, während unter Modell $\mathscr{M}_1$ die Einschränkung $\boldsymbol{\psi} = \boldsymbol{\psi}_0$ gelte.

In diesem Setting lässt sich der Bayes-Faktor zugunsten von Modell $\mathscr{M}_1$ als

$$B_{12} = \frac{p(\boldsymbol{y}|\mathscr{M}_1)}{p(\boldsymbol{y}|\mathscr{M}_2)} = \frac{p(\boldsymbol{\psi}_0|\boldsymbol{y}, \mathscr{M}_2)}{p(\boldsymbol{\psi}_0|\mathscr{M}_2)} \tag{26}$$

schreiben, gegeben, dass sich auch die Priors „schachteln“ lassen, also

$$p(\boldsymbol{\vartheta}_1|\mathscr{M}_1) = p(\boldsymbol{\vartheta}_1|\boldsymbol{\psi} = \boldsymbol{\psi}_0, \mathscr{M}_2) = \frac{p(\boldsymbol{\vartheta}_1, \boldsymbol{\psi}_0|\mathscr{M}_2)}{\int p(\boldsymbol{\vartheta}_1, \boldsymbol{\psi}_0|\mathscr{M}_2) d\boldsymbol{\vartheta}_1}$$

gilt.[10] Dickey (1971) führte dieses Resultat auf Leonard J. Savage zurück und nannte den letzten Term in (26) „Savage's density ratio“. Heutzutage wird der Dichtequotient typischerweise *Savage-Dickey-Density-Ratio* genannt.

Der Dichtequotient nach Savage-Dickey hat eine sehr anschauliche Interpretation, wie Abb. 10 weiter unten illustriert. Dargestellt ist die marginale Prior $p(\psi|\mathscr{M}_2)$ und die marginale Posterior $p(\psi|\boldsymbol{y}, \mathscr{M}_2)$ für ein Problem eines univariaten Parameters ψ, der unter $\mathscr{M}_1$ auf ψ_0 fixiert ist. Der Bayes-Faktor berechnet sich dann als Quotient der A-posteriori-Ordinate $p(\psi_0|\boldsymbol{y}, \mathscr{M}_2)$ und der A-priori-Ordinate $p(\psi_0|\mathscr{M}_2)$. Ist die A-posteriori-Ordinate viel kleiner als die A-priori-Ordinate, $p(\psi_0|\boldsymbol{y}, \mathscr{M}_2) \ll p(\psi_0|\mathscr{M}_2)$, so ist nach (26) der Bayes-Faktor für Modell $\mathscr{M}_1$ deutlich kleiner als 1, was wiederum für das unrestringierte Modell $\mathscr{M}_2$ spricht. Dies ist dann und nur dann der Fall, wenn die Posterior $p(\psi|\boldsymbol{y}, \mathscr{M}_2)$ den Großteil ihrer Masse auf Werten konzentriert, die deutlich von $\boldsymbol{\psi}_0$ abweichen. Ist umgekehrt die A-posteriori-Ordinate deutlich größer als die A-priori-Ordinate, $p(\psi_0|\boldsymbol{y}, \mathscr{M}_2) \gg p(\psi_0|\mathscr{M}_2)$, so ist nach (26) der Bayes-Faktor für Modell $\mathscr{M}_1$ deutlich größer als 1; dies spricht nun für das restringierte Modell $\mathscr{M}_1$ und ist dann und nur dann der Fall, wenn die Posterior $p(\psi|\boldsymbol{y}, \mathscr{M}_2)$ den Großteil ihrer Masse bei Werten um ψ_0 hat.

Für den Vergleich geschachtelter Hypothesen erweist sich der Dichtequotient nach Savage-Dickey also aus zweierlei Gründen als praktisch. Erstens: Da der

[10] Diese Eigenschaft impliziert in der Regel insbesondere, dass $\boldsymbol{\vartheta}_1$ und $\boldsymbol{\psi}$ a priori nicht voneinander abhängen dürfen, was in der Praxis eine gewisse Einschränkung darstellt, siehe etwa Heck (2019). Sind $\boldsymbol{\vartheta}_1$ und $\boldsymbol{\psi}$ voneinander abhängig, so besteht die Möglichkeit, den verallgemeinerten Dichtequotienten nach Savage-Dickey (Verdinelli und Wasserman 1995) zu verwenden.

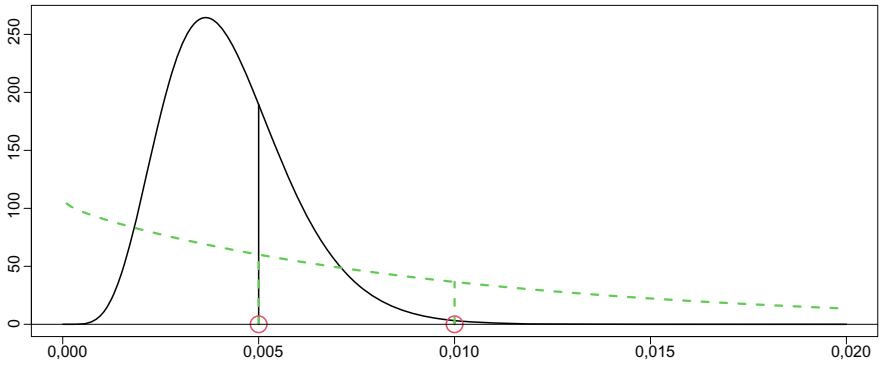

Abb. 10 Illustration des Dichtequotienten nach Savage-Dickey. Gezeigt ist die Prior in Grün strichliert und die Posterior in Schwarz. Für $\psi_0 = 0{,}005$ (linker roter Kreis) gilt $p(\psi_0|\boldsymbol{y}, \mathscr{M}_2) \approx 189{,}48 > p(\psi_0|\mathscr{M}_2) \approx 60{,}11$ und $\log B_{12} \approx 1{,}15$; für $\psi_0 = 0{,}01$ (rechter roter Kreis) gilt $p(\psi_0|\boldsymbol{y}, \mathscr{M}_2) \approx 3{,}18 < p(\psi_0|\mathscr{M}_2) \approx 35{,}54$ und $\log B_{12} \approx -2{,}44$

letzte Term in (26) nur das unrestringierte Modell $\mathscr{M}_2$ beinhaltet, muss das restringierte Modell $\mathscr{M}_1$ zur Berechnung gar nicht erst geschätzt werden. Zweitens: Man braucht zur Schätzung von (26) lediglich die Ordinaten der marginalen A-priori-Dichte $p(\boldsymbol{\psi}_0|\mathscr{M}_2)$ sowie der marginalen A-posteriori-Dichte $p(\boldsymbol{\psi}_0|\boldsymbol{y}, \mathscr{M}_2)$, welche in verschiedenen Anwendungen in geschlossener Form zur Verfügung steht. Ist hingegen bei der praktischen Berechnung des Dichtequotienten in (26) die Ordinate der marginalen A-posteriori-Dichte $p(\boldsymbol{\psi}_0|\boldsymbol{y}, \mathscr{M}_2)$ nicht in geschlossener Form verfügbar, die bedingte Posterior $p(\boldsymbol{\psi}_0|\boldsymbol{\vartheta}_1, \boldsymbol{y}, \mathscr{M}_2)$ jedoch schon, so lässt sich Monte-Carlo-Integration einsetzen:

$$p(\boldsymbol{\psi}_0|\boldsymbol{y}, \mathscr{M}_2) \approx \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{\psi}_0|\boldsymbol{\vartheta}_1^{(m)}, \boldsymbol{y}, \mathscr{M}_2),$$

wobei $\boldsymbol{\vartheta}_1^{(m)}$ für $m = 1, \ldots, M$ A-posteriori-Ziehungen von $\boldsymbol{\vartheta}_1$ unter Modell $\mathscr{M}_2$ sind. Steht selbst $p(\boldsymbol{\psi}_0|\boldsymbol{\vartheta}_1, \boldsymbol{y}, \mathscr{M}_2)$ nicht analytisch zur Verfügung, so sei an dieser Stelle auf Chen (1994) verwiesen.

6.4 Anteilsbeispiel

Zurückkommend auf das einführende Beispiel der Prävalenzstudie, wird im Folgenden die bayesianische Modellwahl anhand der informativen Prior mit A-priori-Erwartungswert und A-priori-Standardabweichung gleich 0,01 illustriert. Möchte man nun etwa testen, ob nach dem Beobachten der Daten die Hypothese, dass der Anteil 1 % beträgt, beibehalten werden kann, so kann man die Modellwahrscheinlichkeit dieses Modells ($\mathscr{M}_2$) mit der Modellwahrscheinlichkeit eines Modells mit einem bei 0,01 fixierten Anteil ($\mathscr{M}_1$) vergleichen.

Für die Modellwahrscheinlichkeiten ergibt sich folgendes Bild:

$$p(\boldsymbol{y}|\mathscr{M}_1) = \binom{N}{S_N} 0{,}01^{S_N} 0{,}99^{N-S_N} \approx 0{,}0036,$$

$$\begin{aligned} p(\boldsymbol{y}|\mathscr{M}_2) &= \int_0^1 \binom{N}{S_N} \vartheta^{S_N} (1-\vartheta)^{N-S_N} \frac{1}{B(a_0, b_0)} \vartheta^{a_0-1} (1-\vartheta)^{b_0-1} d\vartheta \\ &= \binom{N}{S_N} \frac{B(S_N + a_0, N - S_N + b_0)}{B(a_0, b_0)} \approx 0{,}0414, \end{aligned}$$

und damit ein Bayes-Faktor von etwa 0,087 beziehungsweise ein Log-Bayes-Faktor von $\log B_{12} \approx -2{,}441$: deutliche Evidenz gegen die Hypothese, dass der tatsächliche Anteil bei 1 % liegen könnte!

Etwas einfacher lässt sich für dieses Beispiel der Dichtequotient nach Savage-Dickey bestimmen, da sowohl Prior als auch Posterior einfach evaluiert werden können. Einsetzen in (26), wobei ψ_0 dem Anteil ϑ entspricht, der unter Modell $\mathscr{M}_2$ unbekannt ist, ergibt:

$$B_{12} = \frac{p(0{,}01|\boldsymbol{y}, \mathscr{M}_2)}{p(0{,}01|\mathscr{M}_2)} = \frac{\frac{1}{B(a_N, b_N)} 0{,}01^{a_N-1} 0{,}99^{b_N-1}}{\frac{1}{B(a_0, b_0)} 0{,}01^{a_0-1} 0{,}99^{b_0-1}} \approx 0{,}087.$$

Zur Illustration diene Abb. 10. Es ist klar ersichtlich, dass die Dichte der Prior bei dem hypothetischen Anteilswert von 0,01 deutlich größer ist als die Dichtefunktion der Posterior (während bei einem hypothetischen Anteilswert von 0,005 das Gegenteil der Fall ist).

6.5 Wechselkursbeispiel

Wie bereits weiter oben erwähnt, sind für den Modellvergleich mittels Bayes-Faktoren eigentliche Priors vonnöten. Mit der Wahl $p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}|\sigma^2) p(\sigma^2)$ und

$$\boldsymbol{\beta}|\sigma^2 \sim \mathscr{N}_d \left(\boldsymbol{b}_0, \sigma^2 \boldsymbol{B}_0\right), \quad \sigma^2 \sim \mathscr{G}^{-1}(c_0, C_0), \tag{27}$$

lässt sich analog zu Abschn. 3 die Posterior als Produkt einer Normal- und einer inversen Gammaverteilung darstellen:

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{y} \sim \mathscr{N}_d \left(\boldsymbol{b}_N, \sigma^2 \boldsymbol{B}_N\right), \quad \sigma^2|\boldsymbol{y} \sim \mathscr{G}^{-1}(c_N, C_N), \tag{28}$$

wobei

$$\boldsymbol{B}_N^{-1} = \boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{X}, \quad \boldsymbol{b}_N = \boldsymbol{B}_N (\boldsymbol{B}_0^{-1} \boldsymbol{b}_0 + \boldsymbol{X}^\top \boldsymbol{y}),$$
$$c_N = c_0 + N/2, \quad C_N = C_0 + (\boldsymbol{y}^\top \boldsymbol{y} + \boldsymbol{b}_0^\top \boldsymbol{B}_0^{-1} \boldsymbol{b}_0 - \boldsymbol{b}_N^\top \boldsymbol{B}_N^{-1} \boldsymbol{b}_N)/2.$$

Tab. 5 Logarithmen der Marginal Likelihoods für vier Modelle und Differenzen zum Nullmodell

	$\mathscr{M}_0$	$\mathscr{M}_1$	$\mathscr{M}_2$	$\mathscr{M}_{12}$
Log-ML	7048,28	7040,24	7042,99	7035,76
Differenzen	0	−8,0356	−5,2875	−12,5117

Für dieses Modell steht eine geschlossene Form für die Marginal Likelihood zur Verfügung:

$$p(\boldsymbol{y}) = \frac{1}{(2\pi)^{N/2}} \sqrt{\frac{|\boldsymbol{B}_N|}{|\boldsymbol{B}_0|}} \frac{C_0^{c_0}}{C_N^{c_N}} \frac{\Gamma(c_N)}{\Gamma(c_0)},$$

was bequemes Modellvergleichen beziehungsweise Hypothesentesten ermöglicht. Zur Illustration dienen die folgenden Fragen:

1. Kann der Drift aus dem Modell entfernt werden oder: $\beta_1 = 0$?
2. Liegt eine sogenannte Einheitswurzel vor oder: $\beta_2 = 1$?
3. Handelt es sich um einen Random Walk ohne Drift oder: $\beta_1 = 0$ und $\beta_2 = 1$?

Dazu schätzen wir vier Modelle mit folgenden Hyperparametern: $c_0 = 11$, $C_0 = 1$, $\boldsymbol{b}_0 = (0,1)^\top$, $\boldsymbol{B}_0 = \boldsymbol{I}_2$, nämlich:

$\mathscr{M}_0$, ein Random Walk ohne Drift, also $\beta_1 = 0$ und $\beta_2 = 1$;
$\mathscr{M}_1$, ein Random Walk mit Drift, also β_1 unbekannt und $\beta_2 = 1$;
$\mathscr{M}_2$, ein AR(1)-Modell ohne Drift, also $\beta_1 = 0$ und β_2 unbekannt;
$\mathscr{M}_{12}$, ein AR(1)-Modell mit Drift, also β_1 unbekannt und β_2 unbekannt.

Die logarithmischen Marginal Likelihoods (Log-MLs) sind in Tab. 5 dargestellt. Unter der Annahme, dass alle vier Modelle a priori gleich wahrscheinlich sind, also $\Pr(\mathscr{M}_0) = \Pr(\mathscr{M}_1) = \Pr(\mathscr{M}_2) = \Pr(\mathscr{M}_{12})$ gilt, besitzt das Nullmodell $\mathscr{M}_0$ die höchste Modellevidenz, gefolgt von $\mathscr{M}_2$ und $\mathscr{M}_1$. Erst an letzter Stelle steht das volle Modell $\mathscr{M}_{12}$ mit einem Log-Bayes-Faktor von $\log B_{12,0} = -12{,}51$ gegenüber dem Nullmodell. Das einfache AR(1)-Modell scheint also nicht besser zu passen als ein Random Walk.[11]

7 Jenseits konjugierter Bayes-Analysen

Viele der wirklich interessanten Probleme statistischen Modellierens entziehen sich einer geschlossenen Bayes-Analyse, bei der die Posterior, die sich aus dem

[11] Wie schon im Abschn. 6.3 angesprochen, ist eine Verwendung des Dichtequotienten nach Savage-Dickey für dieses Modell nicht zulässig, da in der Prior (27) Abhängigkeit zwischen σ^2 und $\boldsymbol{\beta}$ besteht. Verwendete man diesen dennoch, so erhielte man einen (hier nur leicht verzerrten) Schätzwert für $\log B_{0,12}$ von $-11{,}12$.

Bayes-Theorem (3) ergibt, einer Verteilungsfamilie mit geschlossen darstellbarer Dichte angehört. Zahlreiche in der Praxis relevante Kombinationen aus Prior und Likelihood führen zu einer sogenannten nichtkonjugierten Bayes-Analyse, bei der die Posterior nicht „von Hand“ berechnet werden kann.

7.1 Nichtkonjugierte Bayes-Analysen

Mehrere Faktoren können die Ursache sein, dass wir mit einer nichtkonjugierten Bayes-Analyse konfrontiert sind. Zum einen kann die Wahl einer Prior außerhalb der Familie der sogenannten konjugierten Priors opportun erscheinen. Konjugierte Priors führen nach Kombination mit der Likelihood zu einer Posterior aus derselben Verteilungsfamilie. So kommen bei der beschriebenen Anteilsschätzung sowohl die Prior als auch die Posterior aus der Familie der Beta-Verteilungen. Wenngleich konjugierte Priors vorteilhaft erscheinen, gibt es viele Gründe, allgemeinere Priors zu verwenden. Ein erwähnenswertes Beispiel sind nichtkonjugierte Priors zur bayesianischen Variablenselektion in Regressionsmodellen, wie sie im Detail im Kapitel zur bayesianischen Modellwahl von Wagner Hirk et al. (2020) beschrieben sind und in Abschn. 7.2 zur Illustration einer nichtkonjugierten Bayes-Analyse kurz diskutiert werden.

Häufiger aber liegt die Ursache in einer Likelihood, die sich der Wahl einer konjugierten Prior prinzipiell entzieht. Das ist im Speziellen für Stichprobenverteilungen außerhalb der Familie der Exponentialverteilungen der Fall. Als prominentes Beispiel sei hier auf die Student-t-Verteilung verwiesen, bei der angenommen wird, dass die Daten unabhängige Realisationen einer Zufallsvariablen $Y \sim t_\nu\left(\mu, \sigma^2\right)$ mit unbekanntem Parametervektor $\boldsymbol{\vartheta} = (\nu, \mu, \sigma^2)$ sind. Zudem führen hierarchische Modelle, die von latenten Variablen abhängen, im Allgemeinen zu nichtkonjugierten Problemen. Als gängige Beispiel seien Modelle mit zufälligen Effekten, Zustandsraummodelle oder stochastische Volatilitätsmodelle erwähnt.

Nichtkonjugierte Probleme, bei denen die Posterior $p(\boldsymbol{\vartheta}|\boldsymbol{y})$ keiner geschlossenen Familie von Wahrscheinlichkeitsverteilungen angehört, werden in der modernen Bayes-Statistik mithilfe von sogenannten Markov-Chain-Monte-Carlo-Verfahren (MCMC-Verfahren) behandelt, die – ähnlich wie klassische Monte-Carlo-Verfahren (MC-Verfahren) – eine Folge von zufälligen Ziehungen $\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(M)}$ aus der Posterior $p(\boldsymbol{\vartheta}|\boldsymbol{y})$ verwenden. Im Unterschied zu klassischen MC-Verfahren sind MCMC-Ziehungen in der Regel autokorreliert, weil sie mittels eines Markovprozesses generiert werden. Der Markovprozess ist durch die Übergangsdichte $k(\boldsymbol{\vartheta}^{(m)}|\boldsymbol{\vartheta}^{(m-1)})$ charakterisiert, die beschreibt, wie $\boldsymbol{\vartheta}^{(m)}$ generiert wird, wenn bei der vorangegangenen Ziehung $\boldsymbol{\vartheta}^{(m-1)}$ realisiert wurde. Bei geeigneter Wahl von $k(\boldsymbol{\vartheta}^{(m)}|\boldsymbol{\vartheta}^{(m-1)})$ ergibt sich die Posterior $p(\boldsymbol{\vartheta}|\boldsymbol{y})$ als stationäre Verteilung dieses Markovprozesses, sodass, ausgehend von einem beliebigen Startwert $\boldsymbol{\vartheta}^{(0)}$, die generierten Ziehungen ab einem gewissen M_0 als Realisationen der Posterior $p(\boldsymbol{\vartheta}|\boldsymbol{y})$ betrachtet werden können. Die Ziehungen $\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(M_0)}$ werden als Aufwärmphase (*Burn-In*) betrachtet und verworfen. Die verbleibenden MCMC-Ziehungen können in gleicher Weise wie klassische MC-Ziehungen verwendet werden, etwa

zur Schätzung von nichtlinearen Funktionalen. Für eine rigorose mathematische Darstellung von MCMC-Verfahren sei auf Tierney (1994) verwiesen.

Unter den gängigen MCMC-Verfahren erfreuen sich Gibbs-Sampling (Casella und George 1992) und der Metropolis-Hastings-Algorithmus (Chib und Greenberg 1995) großer Beliebtheit. Vor allem Gibbs-Sampling in Kombination mit *Data-Augmentation* (Tanner und Wong 1987) gehört zu den wichtigsten algorithmischen Innovationen des 20. Jahrhunderts und erlaubt die Schätzung zahlreicher interessanter statistischer Modelle. Als Beispiele seien angeführt: finite Mischungsmodelle (Frühwirth-Schnatter 2001; Frühwirth-Schnatter et al. 2021), Zustandsraummodelle (Frühwirth-Schnatter 1994; Bitto und Frühwirth-Schnatter 2019; Kastner und Huber 2020) sowie stochastische Volatilitätsmodelle (Kastner et al. 2017; Kastner 2019).

7.2 Regressionsanalyse mit Lasso-Prior

Bei einer Regressionsanalyse führt die Wahl der bayesianischen Lasso-Prior mit globalem Skalenparameter τ,

$$p(\boldsymbol{\beta}|\sigma^2) \propto \prod_{j=1}^{d} \exp(-|\beta_j|/(\sigma^2\tau)),$$

zu Variablenselektion im Regressionsmodell. Während die Posterior, die sich aus der Kombination der Likelihood mit der bayesianischen Lasso-Prior ergibt, keine unmittelbare konjugierte Analyse erlaubt, führt Data-Augmentation zu einer Darstellung des Modells, die die Anwendung von Gibbs-Sampling möglich macht. Dabei verwendet man folgende Darstellung der bayesianischen Lasso-Prior als Skalenmischung einer Normalverteilung:

$$\beta_j|\sigma^2, \psi_j \sim \mathcal{N}\left(0, \sigma^2\tau\psi_j\right), \quad \psi_j \sim \mathcal{E}(1), \quad j = 1, \ldots, d, \tag{29}$$

mit lokalem Skalenparameter ψ_j, der als Prior eine Exponentialverteilung besitzt.

Kennt man den Vektor der Skalenparameter $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_d)$, so ist die bedingte Prior $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\psi})$ mit der konjugierten Prior (27) identisch:

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{\psi} \sim \mathcal{N}_d\left(\mathbf{0}, \sigma^2\boldsymbol{B}_0\right), \quad \boldsymbol{B}_0 = \tau\,\mathrm{Diag}(\psi_1, \ldots, \psi_d), \quad \sigma^2 \sim \mathcal{G}^{-1}(c_0, C_0),$$

wobei als A-priori-Erwartungswert $\boldsymbol{b}_0 = \mathbf{0}$ angenommen wird und die Matrix $\boldsymbol{B}_0$ eine Diagonalmatrix ist, die vom latenten Skalenparametervektor $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_d)$ abhängt. Die entsprechende bedingte Posterior $\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{\psi}$ ist durch (28) gegeben, wobei zu beachten ist, dass die Parameter C_N, $\boldsymbol{B}_N$ und $\boldsymbol{b}_N$ von $\boldsymbol{\psi}$ abhängen. Somit können $\boldsymbol{\beta}$ und σ^2 bedingt auf $\boldsymbol{\psi}$ gemeinsam aus der bedingten Posterior $\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{\psi}$ gezogen werden. Data-Augmentation ermöglicht zudem, den Vektor der unbeobachtbaren Skalenparameter $\boldsymbol{\psi}$ aus der bedingten Dichte $p(\boldsymbol{\psi}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y})$ zu imputieren.

Diese beiden bedingten Dichten sind die Grundlage für einen zweistufigen Gibbs-Sampler, der, ausgehend von einem Startwert für $\psi_j^{(0)}$, $j = 1, \ldots, d$, zum Beispiel 1, für $m = 1, \ldots, M$ folgende Schritte wiederholt:

1. Definiere $\boldsymbol{B}_0 = \tau \text{Diag}\left(\psi_1^{(m-1)}, \ldots, \psi_d^{(m-1)}\right)$ und berechne die zugehörigen Momente $c_N, C_N, \boldsymbol{b}_N$ und $\boldsymbol{B}_N$ der bedingten Posterior gemäß (28).
2. Ziehe $\sigma^{2,(m)}$ aus der inversen Gammaverteilung $\mathscr{G}^{-1}(c_N, C_N)$ und dann, unter Verwendung dieser Ziehung, den Parameter $\boldsymbol{\beta}^{(m)}$ aus der multivariaten Normalverteilung $\mathscr{N}_d\left(\boldsymbol{b}_N, \sigma^{2,(m)} \boldsymbol{B}_N\right)$.
3. Ziehe die lokalen Skalenparameter $\psi_j^{(m)}$, $j = 1, \ldots, d$, aus der bedingten Dichte $p(\psi_j | \boldsymbol{\beta}^{(m)}, \sigma^{2,(m)}, \boldsymbol{y})$, die die Form einer inversen Gauß-Dichte annimmt.

Die ersten M_0 Ziehungen werden verworfen und die verbleibenden Ziehungen sind die Grundlage für die weitere Inferenz. Hier sind naturgemäß die Ziehungen für die Regressionsparameter $\boldsymbol{\beta}$ von vorrangigem Interesse. Darüber hinaus enthalten die A-posteriori-Ziehungen für die lokalen Skalenparameter ψ_j Evidenz auf Basis der beobachteten Daten, in welchem Maße der zugehörige Prädiktor relevant ist. Wird die marginale Posterior von ψ_j im Vergleich zur Prior stark gegen 0 gezogen, so wird β_j entsprechend nach 0 geschrumpft und der zugehörige Prädiktor trägt wenig zur Erklärung der Varianz in den Daten bei. Wenn andererseits die Posterior von ψ_j wesentlich größere Werte als die Prior zulässt, liegt ein relevanter Prädiktor vor.

8 Zusammenfassung und Ausblick

In diesem Kapitel wurde versucht, anhand von Beispielen in die grundlegenden Paradigmen der bayesianischen Inferenz einzuführen. Die dargestellten Prinzipien erfreuen sich insbesondere seit der Verwendung leistungsstarker Computer und dem damit verbundenen Einzug von MCMC-Methoden großer Beliebtheit. Auf viele zeitgenössische Problemstellungen im Kontext von hochdimensionalen Parameterräumen, (tiefen) neuronalen Netzen, Regressions- und Klassifikationsforests sowie anderen, heute oft im Schlüsselwort *Machine-Learning* subsumierten Methoden liefert der Bayes-Ansatz fundierte und praktikable Antworten.

Für die Schätzung bayesianischer Modelle steht heute eine Reihe an Software zur Verfügung. Als Beispiele seien nur zwei erwähnt: JASP (JASP Team 2023), als Repräsentant für eine graphische Statistik-Software, und STAN (Carpenter et al. 2017), als Repräsentant für eine generische Modellierungssprache; eine Zusammenfassung der in R implementierten bayesianischen Modelle bietet Park et al. (2023).

Die gigantische Anzahl an Daten, welche, angetrieben durch die fortschreitende Digitalisierung, mehr und mehr anfallen, stellt die Bayes-Statistik nun vor neue Herausforderungen. Es drängt sich unter anderem die Frage auf, wie mit den massiven Datenmengen aus computationaler Sicht im Rahmen des bayesianischen

Ansatzes umgegangen werden kann und wie sich MCMC-Verfahren auf diese riesigen Informationsmengen skalieren lassen.

Literatur

Akaike H (1974) A new look at statistical model identification. IEEE Trans Autom Control 19: 716–723

Atkinson AC (1978) Posterior probabilities for choosing a regression model. Biometrika 65:39–48

Berger JO, Delampady M (1987) Testing precise hypotheses. Stat Sci 2:317–352

Berger JO, Sellke T (1987) Testing a point null hypothesis: the irreconcilability of *P* values and evidence. J Am Stat Assoc 82:112–139

Bitto A, Frühwirth-Schnatter S (2019) Achieving shrinkage in a time-varying parameter model framework. J Econ 210:75–97

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017) Stan: A probabilistic programming language. J Stat Softw 76(1):1–32

Casella G, George EI (1992) Explaining the Gibbs sampler. Am Stat 46:167–174

Chen M-H (1994) Importance-weighted marginal Bayesian posterior density estimation. J Am Stat Assoc 89:818–824

Chib S, Greenberg E (1995) Understanding the Metropolis-Hastings algorithm. J Am Stat Assoc 49:327–335

Dickey JM (1971) The weighted likelihood ratio, linear hypotheses on normal location parameters. Ann Math Stat 42:204–23

Frühwirth-Schnatter S (1994) Data augmentation and dynamic linear models. J Time Ser Anal 15:183–202

Frühwirth-Schnatter S (2001) Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models. J Am Stat Assoc 96:194–209

Frühwirth-Schnatter S, Walli GM, Grün B (2021) Generalized mixtures of finite mixtures and telescoping sampling. Bayesian Anal 16:1279–1307

Heck DW (2019) A caveat on the Savage-Dickey Density ratio: the case of computing Bayes factors for regression parameters. Br J Math Stat Psychol 72(2):316–333

Hirk R, Kastner G, Vana L (2020) Investigating the dark figure of COVID-19 cases in Austria: borrowing from the Decode Genetics Study in Iceland. Austrian J Stat 49(5):1–17

Hosszejni D, Kastner G (2021) Modeling univariate and multivariate stochastic volatility in R with stochvol and factorstochvol. J Stat Softw 100(12):1–34

JASP Team (2023) JASP (Version 0.17.3)[Computer software]. https://jasp-stats.org/

Jeffreys H (1939) Theory of Probability. Clarendon, Oxford

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795

Kastner G (2019) Sparse Bayesian time-varying covariance estimation in many dimensions. J Econ 210(1):98–115

Kastner G, Frühwirth-Schnatter S, Lopes HF (2017) Efficient Bayesian inference for multivariate factor stochastic volatility models. J Comput Graph Stat 26(4):905–917

Kastner G, Huber F (2020) Sparse Bayesian vector autoregressions in huge dimensions. J Forecast 39(7):1142–1165

Park JH, Cameletti M, Pang X, Quinn KM (2023) CRAN task view: Bayesian inference. Version 2023-07-17. https://CRAN.R-project.org/view=Bayesian

R Core Team (2023) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Robert CP (2001) The Bayesian Choice, 2. Aufl. Springer, New York/Berlin/Heidelberg

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Smith AFM, Spiegelhalter DJ (1980) Bayes factors and choice criteria for linear models. J R Stat Soc Ser B 42:213–220

Tanner MA (1993) Tools for statistical inference. Methods of the exploration of posterior distributions and likelihood functions, 2. Aufl. Springer, New York
Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. J Am Stat Assoc 82:528–540
Tierney L (1994) Markov Chain for exploring posterior distributions. Ann Stat 22:1701–1762
Verdinelli I, Wasserman L (1995) Computing Bayes factors using a generalization of the Savage-Dickey density ratio. J Am Stat Assoc 90:614–618