Sieber, Christina:

# Quantifizierung des Priori-Einflusses in Bayesianischen Modellen mittels der Kullback-Leibler-Distanz

**Bachelorarbeit, Sommersemester 2017**

Fakultät für Mathematik, Informatik und Statistik

Ludwig-Maximilians-Universität München

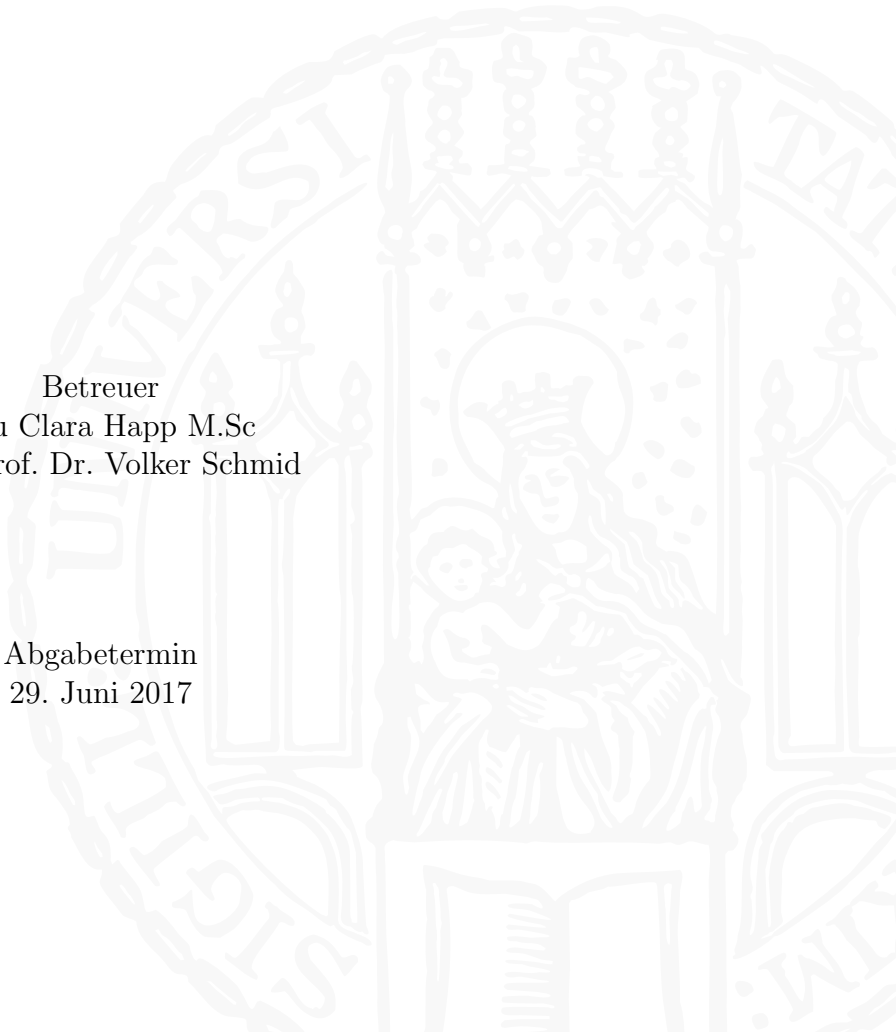
https://doi.org/10.5282/ubm/epub.41011

Institut für Statistik
LMU München

# Bachelorarbeit

## Quantifizierung des Priori-Einflusses in Bayesianischen Modellen mittels der Kullback-Leibler-Distanz

*Christina Sieber*



Betreuer
Frau Clara Happ M.Sc
Herr Prof. Dr. Volker Schmid

Abgabetermin
29. Juni 2017

# Vorwort

Eine sogenannte Priori-Verteilung drückt in der Bayes-Inferenz den Grad der Unsicherheit über einen unbekannten Parameter vor Berücksichtigung einer Beobachtung aus. Nachdem eine Datenerhebung durchgeführt wurde, gelangt die Information zu dem unbekannten Parameter über die Likelihoodfunktion mit dem Satz von Bayes in die Posteriori-Verteilung.
Dabei haben die Likelihood, wie auch die Priori einen Einfluss auf die Posteriori. Anhand der Kullback-Leibler-Distanz, die als Messgröße zwischen zwei Dichten verwendet wird, kann der Abstand zwischen Priori und Posteriori betrachtet werden.

Diese Arbeit setzt sich der Untersuchung der Quantifizierung von Priori-Verteilungen mittels der Kullback-Leibler-Distanz zum Ziel.
Dafür werden in einer kurzen Einführung Grundbegriffe erläutert, die im Laufe der Arbeit für das weitere Verständnis unabdingbar sind.

Werden zunächst eine Normalverteilung und eine Binomialverteilung mit skalarem Parameter betrachtet, wird dieser im Anschluss auf einen bivariaten Parameter für eine Normalverteilung erweitert.
Für die Wahl der Priori-Parameter wird eine Grundstruktur betrachtet, auf die in jedem Kapitel zugegriffen wird. Die Wahl einer schwachen und einer informativen Priori, führt zu einer weiteren Untergliederung der Likelihood mit großer bzw. kleiner Streuung. Dabei wird der Einfluss der Priori auf die Posteriori zunächst theoretisch betrachtet und im Anschluss simulativ überprüft.
Die anschließende Betrachtung der Kullback-Leibler-Distanz soll einen numerische Wert für den Abstand zwischen der Priori und der Posteriori geben.

Eine Anwendung auf die historische Daten der BMW Aktie runden die Arbeit mit einem praktischen Beispiel ab.

Ein besonderer Dank gilt meiner Betreuerin Frau Clara Happ, die sich trotz der Endphase Ihrer Promotion, Zeit und Mühe für Erklärungen, Rechnungen und Hilfestellungen genommen hat.

# Inhaltsverzeichnis

# Kapitel 1

# Einleitung

Im Gegensatz zur frequentischen Statistik, bei der Wahrscheinlichkeit als relative Häufigkeit interpretiert wird, ist in der Bayesianischen Inferenz der subjektive Wahrscheinlichkeitsbegriff von Bedeutung. Die subjektive Wahrscheinlichkeit wird basierend auf einer Vermutung bzw. dem Glaubensgrad über den Eintritt eines Ereignisses gebildet.
Die Priori-Verteilung drückt dabei den Grad der Unsicherheit über einen unbekannten Parameter aus.[Voß, 2004, Vgl. S. 383-389]
Aufgrund unterschiedlicher persönlicher Überzeugungen bzw. Vorwissen kann es dabei zu abweichenden Priori-Verteilungen kommen. Wird die Vermutung durch eine Datenbeobachtung bestätigt, fällt ein sogenannter Überraschungseffekt für einen Betrachter wesentlich geringer aus, als bei einem nicht erwarteten Ausgang und einer damit verbundenen starken Veränderung der Posteriori.
Die Kullback-Leibler-Distanz wird dabei verwendet, um einen eventuellen existierenden Überraschungseffekt als Distanzmaß zwischen der Priori und Posteriori auszudrücken.
[Itti and Baldi, 2009, Vgl. S. 1296]

# Kapitel 2

# Einführung in die Bayes-Statistik

Im folgenden Kapitel werden wichtige Grundbegriffe der Wahrscheinlichkeitstheorie und der Bayes-Statistik im Überblick zusammengetragen, die für den Verlauf der Arbeit ein grundlegendes Verständnis geben sollen.

## 2.1 Der Satz von Bayes

Mit dem Satz von Bayes wird jeder Student mindestens einmal im Studium innerhalb der Wahrscheinlichkeitsrechnung konfrontiert werden. Eine Grundmenge $\Omega$ wird in die Ereignisse $A_1, A_2, \ldots, A_n$ partitioniert mit der Wahrscheinlichkeit $P(A_i) > 0$ für $i = 1, \ldots, n$.

Betrachtet wird nun die Wahrscheinlichkeit von Ereignis $A_i$ unter der Bedingung, dass eine bereits bekannte Information für Ereignis B hinzugezogen wird - dargestellt durch die bedingte Wahrscheinlichkeit mit

$$P(A_i|B) = \frac{P(B|A_i)P(B)}{P(B)}, \qquad i = 1 \ldots, n$$

Dabei sollte $P(B|A_i) > 0$ erfüllt sein. Mit dem Satz der totalen Wahrscheinlichkeit für $P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$ ergibt sich der Satz von Bayes

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}, \qquad i = 1 \ldots, n$$

mit dem die Wahrscheinlichkeit für ein bestimmtes Ereignis berechnet wird.
Im Kontext dazu, gibt $P(A_i)$ die Wahrscheinlichkeit für ein bestimmtes Ereignis $A_i$ an. Durch Eintritt von Ereignis B verändert sich die Wahrscheinlichkeit und wird dargestellt in $P(A_i|B)$.[Fahrmeir et al., 2011, Vgl. S. 209-213]

## 2.2 Grundmodell der Bayesianischen Statistik

Das Grundmodell, bestehend aus der Likelihood, Priori-/ und Posteriori-Verteilung, sind Bestandteile der Bayesianischen Inferenz. Bevor Daten aus einer Beobachtung vorliegen,

wird die Vermutung über einen unbekannten Parameter in Form der Priori-Verteilung ausgedrückt. Verläuft diese sehr flach, so besteht eine unsichere Vermutung und die Priori ist dementsprechend wenig informativ. Mit Berücksichtigung einer Datenbeobachtung fließt die Information, die in den Daten zu dem unbekannten Parameter vorliegen, über den Satz von Bayes in die Posteriori-Verteilung.

### 2.2.1 Likelihood-Funktion

Seien $x = (x_1, x_2, \ldots, x_n)$ beobachtete Realisationen eines Zufallsvektor $X$ vom Umfang $n$, die zusammengefasst in der Dichtefunktion $f(x|\theta)$ sind. Von Interesse ist der unbekannte Parameter $\theta$ mit dem Ziel, die Information aus der Datenbeobachtung x zu nutzen, um Aussagen über den Parameter machen zu können. Die Wahrscheinlichkeit der beobachteten Daten $x$ wird nun als Funktion von $\theta$ in der sogenannten Likelihoodfunktion $L(\theta|x) = f(x|\theta)$ dargestellt.[Held and Bové, 2014, Vgl. S. 14]

### 2.2.2 A-Priori-Verteilung

Ein typisches Merkmal der Bayes-Statistik ist das subjektive Wissen, welches das Maß der persönlichen Überzeugung für das Eintreten eines Ereignisses darstellt.[Voß, 2004, Vgl. S. 384] Diese Vermutung über einen unbekannten Parameter vor einer Datenbeobachtung, in Form einer Stichprobe, wird durch die A-Priori-Verteilung $p(\theta)$ ausgedrückt. Für die Wahl der richtigen Priori-Verteilung werden drei Möglichkeiten erwähnt:

**Nichtinformative Priori**
Ist aufgrund fehlendem bzw. zu schwachem Wissen keine Vorinfomation zur Festlegung einer Priori-Verteilung möglich, kann zum Beispiel eine stetige Gleichverteilung mit konstanter Dichte als nichtinformative Priori gewählt werden. [Held, 2008, Vgl. S. 150]

**Uneigentliche Priori-Verteilung**
Die Wahl einer groß gewählten Priori-Varianz ist eine mögliche Alternative, um den Einfluss der Priori-Verteilung auf die Posteriori zu minimieren. Eine sogenannte impropere Priori-Verteilung führt dazu, dass die Priori für einen stetigen Parameter $\theta$ divergiert

$$\int_{\Theta} f(\theta) d\theta = \infty$$

und nicht mehr integrierbar ist. [Held and Bové, 2014, Vgl. S. 184]

**Konjugierte Priori**
Die rechentechnisch einfachste Methode stellt die Wahl einer konjugierten Priori-Verteilung dar. Multipliziert mit der Likelihood, ist die Posteriori-Verteilung der gleichen Verteilungsfamilie wie die Priori zugehörig.[Held, 2008, Vgl. S. 150]

Im Folgenden wird sich auf die konjugierte Priori beschränkt und diese durch $p_k(\theta)$ gekennzeichnet.

### 2.2.3 A Posteriori-Verteilung

Die Posteriori-Verteilung beinhaltet die gesamte Information über den unbekannten Parameter $\theta$, nachdem die Beobachtung x eines Zufallsvektors X, die zusammengefasst in der Dichtefunktion $f(\theta|x)$ ist, erhoben und beobachtet wurde. Mit der Priori-Verteilung und deren Dichtefunktion $p(\theta)$ ergibt sich mit dem Satz von Bayes für einen stetigen Parameter $\theta$ die Posteriori-Verteilung

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{\int_\Theta f(x|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}$$

Nach Integration des Nenners

$$\int_\Theta f(x|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta} = \int f(x|\tilde{\theta})d\tilde{\theta} = f(x)$$

ergibt sich die marginale Likelihood $f(x)$, welche nicht mehr abhängig von $\tilde{\theta}$ ist. Die Posteriori ist somit proportional zum Produkt der Likelihood und der Dichte der Priori

$$p(\theta|x) \propto f(x|\theta)p(\theta)$$

mit der Normalisierungskonstante $\frac{1}{f(x)}$, so dass die Eigenschaft einer Dichtefunktion $\int f(x|\tilde{\theta})d\tilde{\theta} = 1$ erfüllt ist.[Held and Bové, 2014, Vgl. S. 170f.]

## 2.3 Maximum-Likelihood-Schätzung

Eine Methode über die Schätzung eines unbekannten Parameters liefert die Maximum-Likelihood-Schätzung. Für das nicht-bayesianische Verfahren wird sich dabei nur auf die Likelihood beschränkt.

Durch Maximierung der Likelihoodfunktion erhält man für die Realisationen $x_1, \ldots, x_n$ den Parameter, für den die Likelihood innerhalb des Parameterraumes maximal wird. Der sogenannten Maximum-Likelihood-Schätzer, im Folgenden abgekürzt als ML-Schätzer, wird durch die erste Ableitung der Likelihood und Nullsetzen der Score-Funktion berechnet.

### 2.3.1 Log-Likelihood-Funktion

Der Logarithmus der Likelihoodfunktion, die sogenannte Log-Likelihood-Funktion $l(\theta|x)$, kann als einfachere Variante zur Lösung des Optimierungsproblemes hinzugezogen wer-

den. Aufgrund der streng monotonen Transformation der Logarithmierung, trägt die Log-Likelihood die gleiche Information wie die Likelihood und es gilt [Voß, 2004, Vgl. S. 392]

$$\hat{\theta}_{ML} = \underset{\theta}{argmax}\ L(\theta|x) = \underset{\theta}{argmax}\ \log L(\theta|x) = \underset{\theta}{argmax}\ l(\theta|x).$$

### 2.3.2 Score-Funktion

Die erste Ableitung der Log-Likelihood-Funktion wird als Score-Funktion bezeichnet und es gilt

$$s(\theta|x) = \frac{\partial l(\theta|x)}{d\theta}.$$

Durch Nullsetzen und Auflösen von $s(\theta|x)$ erhält man den ML-Schätzer $\hat{\theta}_{ML}$.

### 2.3.3 Fisher-Information und asymptotische Verteilung des ML-Schätzer

Durch die zweite negative Ableitung der Log-Likelihood-Funktion bzw. der negativen Ableitung der Scorefunktion berechnet sich die Fisher-Information durch

$$I(\theta|x) = -\frac{\partial^2 l(\theta|x)}{\partial^2 \theta} = -\frac{\partial s(\theta|x)}{\partial \theta}.$$

Wird diese an der Stelle $\hat{\theta}_{ML}$ betrachtet, so wird $I(\hat{\theta}_{ML}|x)$ als beobachtete Fisher-Information bezeichnet.[Held, 2008, Vgl. S. 29]
Oftmals ist $I(\theta|x)$ nicht nur vom Parameter, sondern auch von den Daten abhängig, weshalb der Erwartungswert der Fisher-Information $I(\theta|x)$ betrachtet wird

$$J(\theta) = \mathbb{E}[I(\theta|x)].$$

$J(\theta)$ wird dabei als erwartete Fisher-Information bezeichnet. [Held, 2008, Vgl. S. 64]
Für die asymptotische Verteilung des ML-Schätzers gilt

$$\hat{\theta}_{ML} \overset{a}{\sim} N\left(\theta, (J(\theta))^{-1}\right)$$

und die Varianz des ML-Schätzers wird durch die Inverse der erwarteten Fisher-Information $J(\theta)$ geschätzt.[Held, 2008, Vgl. S. 80]

## 2.4 Bayesianische Punktschätzer

Für die Schätzung des unbekannten Parameters in der Bayesianischen Inferenz, werden die aus der frequentischen Statistik bereits bekannten Lageparameter Erwartungswert, Modus und Median verwendet. In der weiteren Betrachtung wird sich auf den Erwar-

tungswert und den Modus beschränkt, der auch als Maximum-a-Posteriori-Schätzer bekannt ist. Als Bayesianische Punktschätzer $\hat{\theta}$ bezeichnet, definieren sich diese wie folgt:

### 2.4.1 Posteriori-Erwartungswert

Der Posteriori-Erwartungswert

$$\hat{\theta}_{PE} = \mathbb{E}(\theta|x) = \int \theta p(\theta|x) d\theta$$

ist der erwartete Wert der Posteriori-Verteilung $p(\theta|x)$.

### 2.4.2 Maximum-a-Posteriori-Schätzer

Ähnlich zur Maximum-Likelihood-Methode wird der unbekannte Parameter bei der Maximum-a-Posteriori-Schätzung (MAP) durch das globale Maximum der Posteriori-Verteilung berechnet.

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

Aufgrund der zur Berechnung nicht benötigten Normierungskonstante, wird $\hat{\theta}_{MAP}$ gegenüber $\hat{\theta}_{PE}$ bevorzugt.[Fahrmeir et al., 2009, Vgl. S. 480f.]

## 2.5 Markov-Chain-Monte-Carlo-Methode

Die Normalisierungskonstante für die Posteriori-Verteilung wird durch Integration der Parameter berechnet. Für ein hochdimensionales Integral ist das analytisch und numerisch kaum möglich, weshalb sogenannte Markov-Chain-Monte-Carlo-Methoden, abgekürzt MCMC, angewendet werden. Dabei wird eine Markov-Kette erzeugt, die gegen eine interessierende Posteriori-Verteilung konvergiert. Nach einer gewissen Konvergenzzeit, der sogennanten „Burn In"-Phase, bilden die gezogenen Zufallszahlen eine abhängige Stichprobe, die als Posteriori-Verteilung dargestellt werden kann.[Fahrmeir et al., 2009, Vgl. S. 482]

## 2.6 Full-Conditional

Falls kein skalarer sondern ein multivariater Parameter vorliegt, wird eine einzelne Komponente $\theta_j$ betrachtet. Aus diesen sogenannten Full-Conditionals $p(\theta_j|x, \theta_{-j})$, gegeben den Daten $x$ und $\theta_{-j}$, werden nun Zufallszahlen gezogen. Dabei bezeichnet $\theta_{-j}$ den Vektor $\theta$ ohne die Komponente $\theta_j$.

Wird wiederum für die einzelnen Parameter $\theta_j$ eine konjugierte Priori gewählt, so stammen die Full-Conditional und die als semikonjugiert bezeichneten Prioris $p(\theta_j)$ aus der gleichen Verteilungsfamilie.[Held and Bové, 2014, Vgl. S. 270]

## 2.7 Gibbs-Sampling

Ein Algorithmus für eine Markov-Kette ist das Gibbs-Sampling, bei dem der Parametervektor in n Komponenten unterteilt wird mit $\theta = (\theta_1, \ldots, \theta_n)$. Iterativ werden nun Zufallszahlen aus den Full Conditionals der Parameterkomponenten gezogen. Dabei wird der erhaltene Wert einer Komponente als bedingte Größe in die Full Conditional der folgenden Komponete eingesetzt, so dass jeder Untervektor $\theta_j$ bedingt auf den letzen Wert der vorherigen Komponente aktualisiert wird.[Gelman et al., 2014, Vgl.S. 277]

## 2.8 Theorie der Bayesianischen „Surprise“

An der Stärke eines „Überraschungs-Effektes“ gemessen, entsteht eine nicht erwartete Erkenntnis oftmals vor allem in der Gegenwart von Unsicherheit, beispielsweise aufgrund fehlender oder falscher Informationen. Ebenso können identische Daten für diverse Personen eine unterschiedlich starke Auswirkung haben. Auch der Zeitraum kann von Bedeutung sein.
Eine Möglichkeit zur Quantifizierung dieses Effektes ist die Kullback-Leibler-Distanz.
[Itti and Baldi, 2009, Vgl. S.1295-1299]

## 2.9 Kullback-Leibler-Distanz

Die Kullback-Leibler-Distanz

$$D_{KL}(p\big\|q) = \mathbb{E}\left(ln\frac{p}{q}\right) = \begin{cases} \int_{\mathcal{X}} p(x) ln\left(\frac{p(x)}{q(x)}\right) dx, & \text{wenn } \mathcal{X} \text{ ist stetig} \\ \sum\limits_{x_i \in \mathcal{X}} p(x_i) ln\left(\frac{p(x_i)}{q(x_i)}\right), & \text{wenn } \mathcal{X} \text{ ist diskret} \end{cases}$$

ist eine Messgröße zwischen zwei Dichtefunktionen $p$ und $q$. Im Kontext zur Bayesianischen Statistik kann die Kullback-Leibler-Distanz als Kriterium genutzt werden, wie weit die Priori(=$q$) von der Posteriori(=$p$) entfernt liegt.
Diese nimmt keine negativen Werte an $D_{KL}(p\big\|q) \geq 0$ und Gleichheit $D_{KL}(p\big\|q) = 0$ ist nur erfüllt, wenn p und q identisch sind.[Calin and Udrişste, 2014, Vgl. S. 113]
Aufgrund der fehlenden Symmetrie

$$D_{KL}(p\big\|q) \neq D_{KL}(q\big\|p)$$

kann diese nicht als gewöhnliches Distanzmaß bezeichnet werden.[Held and Bové, 2014, Vgl. S 329]

# Kapitel 3

# Anwendung auf Wahrscheinlichkeitsverteilungen

Die im vorangegangenem Kapitel aufgeführte Theorie soll nun auf zwei Wahrscheinlichkeitsverteilungen angewendet werden. Wird vorerst nur ein eindimensionaler Parameter betrachtet, wird dieser im Anschluss auf einen zweidimensionalen Parameter erweitert.

## 3.1 Normalverteilung mit skalarem Parameter

Die Normalverteilung ist eine der am Wichtigsten in der Statistik und gehört zu den stetigen Wahrscheinlichkeitsverteilungen. Abhängig vom Mittelwert $\mu \in \mathbb{R}$ und der Varianz $\sigma^2 > 0$, wird sie dargestellt durch die Dichtefunktion

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \qquad \text{mit } x \in \mathbb{R}.$$

Das Maximum besitzt die auch als Gauß´sche Glockenkurve bezeichnete Dichte an der Stelle $x = \mu$, und fällt symmetrisch zum Maximum gegen Null ab. Dabei bestimmt die Varianz $\sigma^2$ die Konzentration um den Mittelwert. Ist eine Zufallsvariable X normalverteilt, wird diese in der Form $X \sim N(\mu,\sigma^2)$ dargestellt.[Voß, 2004, Vgl. S. 338f.]
Im weiteren Verlauf wird nun die Annahme getroffen, dass Daten x=$(x_1, x_2, \ldots, x_n)$ aus einer Stichprobe unabhängig und identisch verteilt sind und ein skalarer unbekannter Parameter $\theta$ vorliegt.
Dabei lassen sich bei der Normalverteilung zwei Fälle unterscheiden.

### 3.1.1 Erwartungswert $\mu \in \mathbb{R}$ unbekannt

Für den ersten Fall wird die Normalverteilung mit bekannter Varianz betrachtet. Der unbekannte Parameter stellt den Erwartungswert $\mu \in \mathbb{R}$ dar und das Grundmodell, bestehend aus der Likelihood, Priori und Posteriori, ergibt sich wie folgt.

#### 3.1.1.1 Grundmodell

Seien $x = (x_1, \ldots, x_n)$ unabhängig und identisch verteilte Beobachtungen von $X \sim N(\mu, \sigma^2)$ mit zu schätzendem Parameter $\mu$. Die Likelihood ergibt sich durch

$$\begin{aligned} f(x_1, \ldots, x_n|\mu) &\overset{i.i.d}{=} \prod_{i=1}^{n} f(x_i|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\Big(-\frac{1}{2\sigma^2}(x_i-\mu)^2\Big) \\ &= \underbrace{\Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n}_{\text{konstant}} exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big) \\ &\propto exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big). \end{aligned}$$

Die dazugehörige konjugierte Priori-Verteilung

$$\begin{aligned} p_k(\mu) &= \frac{1}{\sqrt{2\pi}\sigma_0} exp\Big(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\Big) \\ &\propto exp\Big(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\Big) \end{aligned}$$

ist ebenfalls normalverteilt [Held and Bové, 2014, Vgl. S. 181] und der unbekannte Parameter $\mu$ wird als Wahrscheinlichkeitsverteilung mit den sogenannten Hyperparametern $\mu_0$ und $\sigma_0^2$ dargestellt. Diese werden als bekannt angenommen und drücken die Vermutung über die Lage des unbekannten Parameters $\mu$ aus.[Gelman et al., 2014, Vgl. S. 40] Die Posteriori ergibt sich aus der Multiplikation der Likelihood mit der konjugierten Priori und ist einer Normalverteilung zugehörig

$$\begin{aligned} p(\mu|x) &\propto f(x|\mu)p_k(\mu) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \end{aligned}$$

$$\propto \exp\left(-\frac{\sigma_0^2 \sum_{i=1}^{n}(x_i-\mu)^2 + \sigma^2(\mu-\mu_0)^2}{2\sigma^2\sigma_0^2}\right)$$

$$\propto \exp\left(-\frac{1}{2\sigma^2\sigma_0^2}\left(\sigma_0^2\left(\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + n\mu^2\right) + \sigma^2\left(\mu^2 - 2\mu\mu_0 + \mu_0^2\right)\right)\right)$$

$$\propto \exp\left(-\frac{1}{2\sigma^2\sigma_0^2}\left(\mu^2\left(n\sigma_0^2+\sigma^2\right) - 2\mu\left(\sigma_0^2\sum_{i=1}^{n}x_i + \sigma^2\mu_0\right) + \sigma_0^2\sum_{i=1}^{n}x_i^2 + \sigma^2\mu_0^2\right)\right)$$

$$\propto \exp\left(-\frac{(n\sigma_0^2+\sigma^2)}{2\sigma^2\sigma_0^2}\left(\mu^2 - 2\mu\underbrace{\frac{(\sigma_0^2\sum_{i=1}^{n}x_i + \sigma^2\mu_0)}{(n\sigma_0^2+\sigma^2)}}_{:=y} + \underbrace{\frac{(\sigma_0^2\sum_{i=1}^{n}x_i + \sigma^2\mu_0)}{(n\sigma_0^2+\sigma^2)}}_{\text{konstant in }\mu}\right)\right)$$

$$\propto \exp\left(-\frac{(n\sigma_0^2+\sigma^2)}{2\sigma^2\sigma_0^2}\left(\mu^2 - 2\mu y + y^2 - y^2\right)\right) \qquad \text{Quadratische Ergänzung}$$

$$\propto \exp\left(-\frac{(n\sigma_0^2+\sigma^2)}{2\sigma^2\sigma_0^2}(\mu-y)^2\right)\underbrace{\exp\left(-y^2\right)}_{\text{konstant in }\mu} \qquad \text{Binomische Formel}$$

$$\propto \exp\left(-\frac{(n\sigma_0^2+\sigma^2)}{2\sigma^2\sigma_0^2}(\mu-y)^2\right)$$

$$\underbrace{\propto \exp\left(-\frac{1}{2}\left(\frac{n\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\right)^{-1}\left(\mu - \frac{\sigma_0^2\sum_{i=1}^{n}x_i + \sigma^2\mu_0}{n\sigma_0^2+\sigma^2}\right)^2\right)}_{\text{Kern einer } N(\mu_{post},\sigma_{post}^2)\text{-Verteilung}}$$

mit den Parametern

$$\mu_{post} = \frac{\sigma_0^2\sum_{i=1}^{n}x_i + \sigma^2\mu_0}{n\sigma_0^2+\sigma^2} \overset{(*)}{=} \frac{\sigma_0^2 n\bar{x} + \sigma^2\mu_0}{n\sigma_0^2+\sigma^2} \qquad (*) \text{ mit } \sum_{i=1}^{n}x_i = n\frac{1}{n}\sum_{i=1}^{n}x_i = n\bar{x}$$

und

$$\sigma_{post}^2 = \left(\frac{n\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\right)^{-1} = \left(\frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}\right).$$

Bei einem Vergleich zwischen der Priori-/ und der Posteriori-Varianz ist erkennbar, dass mit zunehmendem Stichprobenumfang ($n \to \infty$) der Nenner von $\sigma^2_{post}$ größere Werte annimmt und die Posteriori-Varianz gegen Null konvergiert. Die aus der Datenbeobachtung erhaltene Sicherheit über den unbekannten Parameter dominiert und die Varianz für die Posteriori fällt kleiner aus, als die der Priori. Für klein gewähltes n wird ein Vergleich zwischen $\sigma^2_{post}$ und $\sigma^2_0$ von der Beobachtung und der Stärke des Vorwissens eines Betrachters beeinflusst.[Rüger, 1999, Vgl. S. 207]
Eine Alternative zur Darstellung der Posteriori-Varianz ist die Präzision

$$\frac{1}{\sigma^2_{post}} = \left(\frac{n\sigma^2_0 + \sigma^2}{\sigma^2\sigma^2_0}\right) = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma^2_0}\right) = (n\tau + \kappa)^{-1},$$

die vor allem eine übersichtlichere Interpretation ermöglicht. Als Inverse der Varianz wird die Daten-Präzision durch $\tau = \frac{1}{\sigma^2}$ dargestellt, welche im Vergleich zur Priori-Präzision $\kappa = \frac{1}{\sigma^2_0}$ mit dem Stichprobenumfang n multipliziert wird.[Held, 2008, Vgl. S. 148]

#### 3.1.1.2 Maximum-Likelihood-Schätzung

Der ML-Schätzer $\hat{\mu}_{ML}$ ergibt sich durch die Ableitungen der Likelihood und Lösung der Score-Gleichung s($\mu|x$)=0.

$$L(\mu, \sigma^2) = f(x|\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

$$l(\mu, \sigma^2) = \log L(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}$$

$$s(\mu|x) = \frac{\partial l(\mu, \sigma^2)}{\partial \mu} = \frac{\sum_{i=1}^{n} x_i - n\mu}{\sigma^2} \stackrel{!}{=} 0 \qquad \Rightarrow \hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$I(\mu|x) = -\frac{\partial s(\mu|x)}{\partial \mu} = \frac{n}{\sigma^2}$$

Da die Fisher-Information $I(\mu|x)$ nicht von $\mu$ abhängt, kann diese mit der erwarteten Fisher-Information $J(\mu)$ gleichgesetzt werden.[Held, 2008, Vgl. S. 68]
Die asymptotische Varianz des ML-Schätzers wird durch die Inverse der Fisher-Information $(J(\mu))^{-1}$ geschäzt und es gilt

$$\hat{\mu}_{ML} \stackrel{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right).$$

#### 3.1.1.3 Bayes-Schätzer

Durch die Symmetrie-Eigenschaft der Normalverteilung[Voß, 2004, Vgl. S. 339] und der Erwartungswert und der Modus somit dem Parameter $\mu$ entsprechen, kann der Posteriori-Mittelwert den Bayesianischen Punktschätzern gleichgesetzt werden:

$$\mu_{post} = \hat{\mu}_{PE} = \hat{\mu}_{MAP} = \frac{\sigma_0^2 n\bar{x} + \sigma^2\mu_0}{n\sigma_0^2 + \sigma^2} = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x} + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0.$$

Nach Umstellung des Bayes-Schätzers lässt sich eine Gewichtung zwischen dem Stichprobenmittelwert $\bar{x}$ und dem Priori-Erwartungswert $\mu_0$ erkennen. Eine schwache Priori-Varianz mit $\sigma_0^2 \to \infty$ führt zur einer starken Gewichtung von $\bar{x}$ und $\mu_{post}$ strebt gegen den ML-Schätzer. Analoges Verfahren gilt bei einem größer werdenden Datenumfang $n \to \infty$. Tendiert die Vermutung über die Lage des unbekannten Parameters stark mit $\sigma_0^2 \to 0$, dominiert der Priori-Erwartungswert hauptsächlich in kleinen Stichproben.[Rüger, 1999, Vgl. S. 206]

#### 3.1.1.4 Visualisierung der Normalverteilung

Abbildung 3.1 zeigt die Auswirkung unterschiedlich gewählter Varianz-Parameter auf den Mittelwert $\mu = 0$, der in diesem Beispiel frei gewählt wird. Bei klein gewählten Werten für die Varianz $\sigma^2$, verläuft die Dichte der Normalverteilung steiler um $\mu$ herum, dargestellt durch die braun eingefärbte $N(0, 0.04)$-Verteilung. Je größer die Varianz, desto weniger Vorwissen über die Lage des unbekannten Parameters wird ausgedrückt und die Dichte verläuft flach. Eine größer werdende Varianz verdeutlicht somit eine hohe Unsicherheit und ist damit wenig informativ.

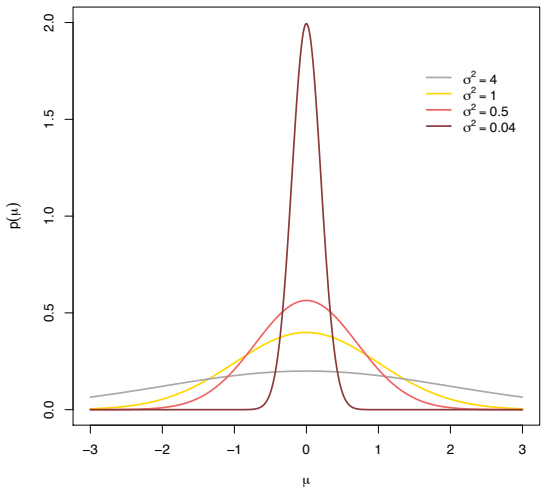



Abbildung 3.1: Visualisierung einer Normalverteilung

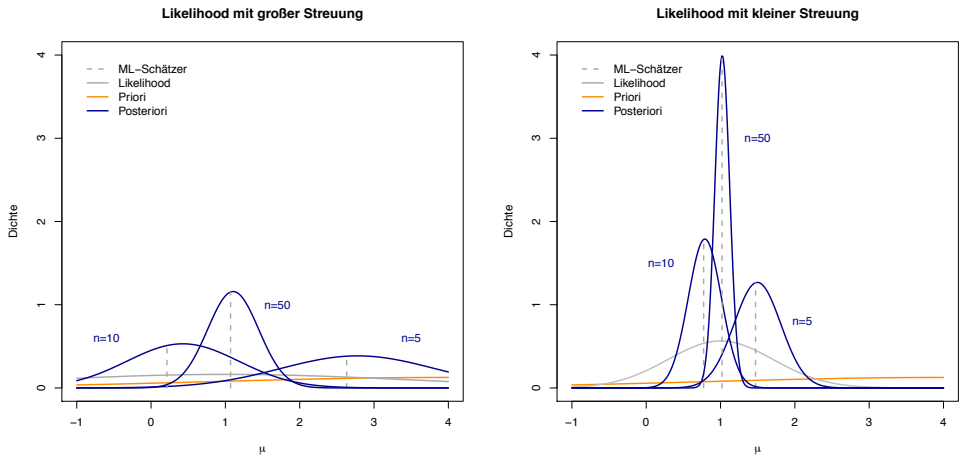
Für die weitere Betrachtung wird eine Fallunterscheidung mit folgender Grundstruktur betrachtet, die sich als roter Faden durch die Arbeit ziehen wird. Dabei wird zwischen einer schwachen bzw. informativen Priori mit weiterer Untergliederung in eine Likelihood mit großer bzw. kleiner Streuung unterschieden.

#### 3.1.1.5 Graphische Darstellung und Interpretation

Aufgrund der dominierenden Rolle der Likelihood für große n, werden im Folgenden vorerst kleine Stichproben vom Umfang $n$=5, $n$=10 und $n$=50 betrachtet. Die Erwartungswerte der Priori mit $\mu_0$=4 und der Likelihood mit $\mu$=1 sind dabei beliebig gewählt. Um den Einfluss der Priori besser darzustellen, werden diese bewusst mit einem größeren Abstand zueinander gewählt.

Fall 1: Wahl einer schwachen Priori

Basierend auf der Visualisierung der Normalverteilung in Abbildung 3.1 wird für eine schwache Priori die Varianz $\sigma_0^2$=10 gewählt. Die Gewichtung von $\bar{x}$ fällt stärker aus und der ML-Schätzer wird in Abbildung 3.2 durch die senkrecht verlaufenden grauen Linien dargestellt. Die jeweiligen Posterioris erreichen für kleines n bereits ihr Maximum am ML-Schätzer. Die Dichtefunktionen für eine Likelihood mit kleiner Streuung in der rechten Grafik, fallen im Vergleich deutlich höher aus.



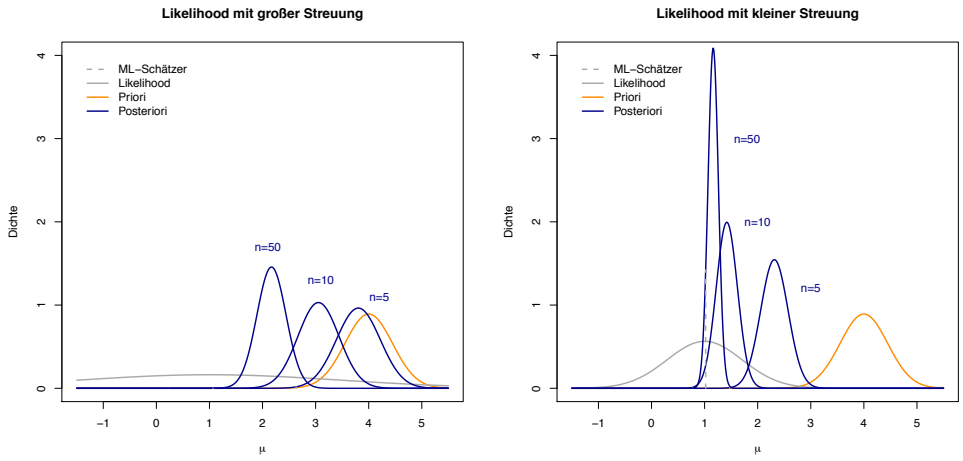

(a) X ∼ N(1,6) und $\mu_0 \sim N(4,10)$ (b) X ∼ N(1,0.5) und $\mu_0 \sim N(4,10)$

Abbildung 3.2: Wahl einer schwachen Priori

Fall 2: Wahl einer informativen Priori

Bei der Wahl einer informativen Priori mit $\sigma_0^2 = 0.2$ fällt dieser eine höhere Gewichtung zu, wenn die vorliegenden Daten wenig Information tragen. Im linken Bild von Abbildung 3.3 deutlich erkennbar, dass die Posteriori-Verteilungen für n=5 und n=10 nahe an der Priori liegen und einen ähnlichen Verlauf aufweisen. Die Posteriori wird aufgrund der informativ gewählten Priori bestimmt und der ML-Schätzer der jeweiligen Stichpro-

be ist nicht ersichtlich, da dieser weiter links liegende Werte für $\mu$ aufweist. Für zunehmendes n wird der Einfluss der Priori geringer und die Posteriori strebt gegen den Mittelwertschätzer. Bei Betrachtung der rechten Grafik und einer klein gewählten Varianz für Likelihood und Priori erreichen Priori und Posteriori in unterschiedlichen Wertebereichen ihr Maximum. Die gewonnene Information aus den Daten ist nicht informativ und es entsteht ein Widerspruch zur gemachten Priori-Annahme.




(a) X ∼ N(1,6) und $\mu_0 \sim N(4, 0.2)$ (b) X ∼ N(1,0.5) und $\mu_0 \sim N(4, 0.2)$

Abbildung 3.3: Wahl einer informativen Priori

In allen vier Betrachtungen ist deutlich erkennbar, dass mit zunehmendem Stichprobenumfang der Einfluss der Priori geringer wird, die Posteriori eine geringere Varianz aufweist und die Posteriori-Mittelwerte gegen das Maximum der Likelihood streben. Der Bayes-Schätzer entspricht somit approximativ dem ML-Schätzer.

Eine Alternative für die Wahl der Priori-Parameter wäre, wenn der Erwartungswerte dem ML-Schätzer entspricht und die Varianz der asymptotischen Varianz des ML-Schätzers. Dabei ist die Varianz einer Normalverteilung durch $\mathbb{V}(X) = \sigma^2$ gegeben [Voß, 2004, Vgl. S. 341] und für die folgende Ungleichung gilt:

$$\mathbb{E}(\mu_0|x, \sigma^2) \stackrel{!}{=} \hat{\mu}_{ML} \Leftrightarrow \mu_0 = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\mathbb{V}(\mu_0|x, \sigma^2) \stackrel{!}{=} (J(\mu))^{-1} \Leftrightarrow \sigma_0^2 = \frac{\sigma^2}{n}$$

Für die Betrachtung einer großen Streuung der Likelihood mit $X \sim N(1, 6)$ in Abbildung 3.4 ist dabei aufgrund der Parameterwahl ein ähnlicher Verlauf von Priori und Posteriori um den ML-Schätzer sichtbar. Die Streuung wird mit zunehmendem n kleiner, wobei die Posteriori eine geringere Varianz im Vergleich zur Priori aufweist.

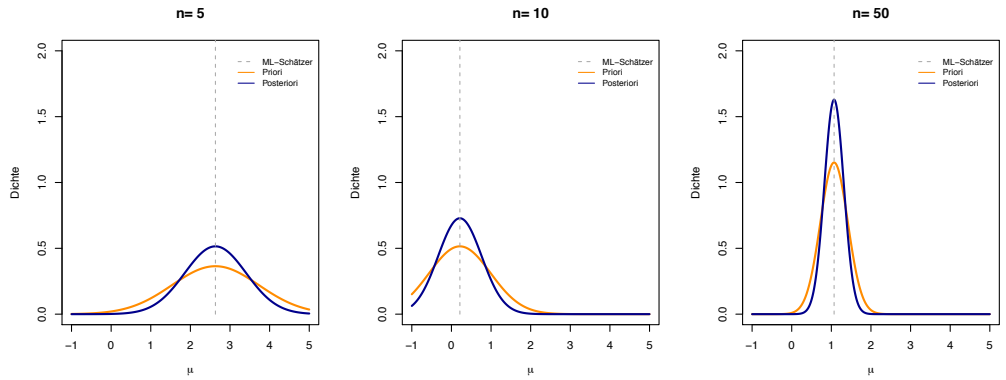
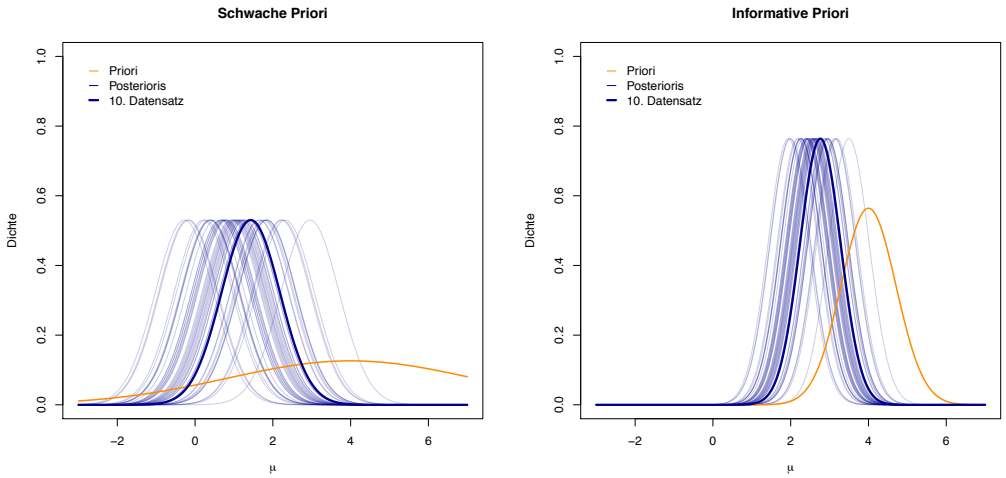



Abbildung 3.4: Große Streuung der Likelihood mit $X \sim N(1,6)$

Vor allem für kleine Stichproben sind die Parameter abhängig von den zufällig gezogenen Daten aus der Likelihood. Abbildung 3.5 visualisiert 50 Datensätze mit einem Stichprobenumfang von jeweils 10 Beobachtungen. Ein frei gewählter Datensatz, hier der 10. Datensatz, wird mit einer dickeren Linie kenntlich gemacht. Ausgehend von einer Likelihood mit $X \sim N(1,6)$ wird nun zwischen einer schwachen und einer informativen Priori der Vergleich dargestellt. Die Erwartungswerte der Posteriori nehmen bei einer flachen Priori mit $\mu_0 \sim N(4,10)$ Werte zwischen -0.5 und 3.5 an. Bei einer informativen Priori $\mu_0 \sim N(4,0.5)$ reduziert sich dieser auf einen Bereich von etwa [1.8,4]. Im Vergleich ist in der rechten Abbildung eine höhere Dichte mit kleiner Varianz bei Priori und Posteriori erkennbar.




(a) X ∼ N(1,6) und $\mu_0 \sim N(4,10)$ (b) X ∼ N(1,6) und $\mu_0 \sim N(4,0.5)$

Abbildung 3.5: Simulation von 50 Datensätzen

### 3.1.2 Varianz $\sigma^2 > 0$ unbekannt

Der zweite Fall einer Normalverteilung mit skalarem Parameter ist die unbekannte Varianz $\sigma^2 > 0$. Der Erwartungswert wird nun als bekannt vorausgesetzt.

#### 3.1.2.1 Grundmodell

Für einen Vektor $x = (x_1, x_2, \ldots, x_n)$ von n unabhängig und identisch verteilten Beobachtungen aus einer Normalverteilung ergibt sich für die Dichte einer Normalverteilung mit unbekannter Varianz

$$f(x_1, \ldots, x_n|\sigma^2) \overset{i.i.d}{=} \prod_{i=1}^{n} f(x_i|\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(-\frac{1}{2\sigma^2}(x_i-\mu)^2\Big)$$

$$= \underbrace{(2\pi)^{-\frac{n}{2}}}_{konstant} (\sigma^2)^{-\frac{n}{2}} exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big)$$

$$\propto (\sigma^2)^{-\frac{n}{2}} exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big).$$

Ist die Likelihood normalverteilt mit $\sigma^2$ unbekannt, stellt die konjugierte Priori-Verteilung eine Invers-Gamma-Verteilung $\sigma^2 \sim IG(a,b)$ dar, [Held and Bové, 2014, S. 181]

$$p_k(\sigma^2|a,b) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)} exp\Big(-\frac{b}{\sigma^2}\Big)$$

$$\propto (\sigma^2)^{-(a+1)} exp\Big(-\frac{b}{\sigma^2}\Big)$$

mit Erwartungswert und Modus [Held, 2008, Vgl. S. 272]

$$\mathbb{E}(\sigma^2|a,b) = \frac{b}{a-1} \text{ für } a > 1 \qquad \text{und} \qquad mod(\sigma^2|a,b) = \frac{b}{a+1}.$$

Die Dichte der Posteriori

$$p(\sigma^2|x) \propto f(x|\sigma^2) p_k(\sigma^2|a,b)$$

$$\propto (\sigma^2)^{-\frac{n}{2}} exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big)(\sigma^2)^{-(a+1)} exp\Big(-\frac{b}{\sigma^2}\Big)$$

$$\underbrace{\propto (\sigma^2)^{-(a+\frac{n}{2}+1)} exp\Big[-\frac{1}{\sigma^2}\Big(b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\Big)\Big]}_{\text{Kern einer Invers-Gamma-Verteilung}}$$

ist eine Invers-Gamma-Verteilung $\sigma^2|x \sim IG(\tilde{a}, \tilde{b})$ mit den Parametern

$$\tilde{a} = a + \frac{n}{2} \qquad und \qquad \tilde{b} = b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Diese sind abhängig von Umfang $n$ und den Daten aus der Stichprobe, sowie den Priori-Parametern $a$ und $b$.

#### 3.1.2.2 Maximum-Likelihood-Schätzung

Durch Ableitung der Likelihoodfunktion

$$L(\mu, \sigma^2) = f(x|\mu, \sigma^2) = \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\Big)$$

$$l(\mu, \sigma^2) = \log L(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}$$

$$s(\sigma^2|x) = \frac{\partial l(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{2\sigma^2}\Big(-n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\Big)$$

und Auflösen der Score-Gleichung, erhält man als ML-Schätzer:

$$s(\sigma^2|x) \stackrel{!}{=} 0 \quad \Rightarrow \quad \widehat{\sigma^2}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Dieser wird als Stichprobenvarianz bezeichnet und ist ein Maß für die Unsicherheit der erhobenen Daten. [Rüger, 1999, Vgl. S. 206]

#### 3.1.2.3 Bayes-Schätzer

Werden die Posteriori-Parameter $\tilde{a}$ und $\tilde{b}$ in den Modus einer Invers-Gamma-Verteilung eingesetzt, so ergibt sich:

$$\widehat{\sigma^2}_{MAP} = mod(X|\tilde{a}, \tilde{b}) = \frac{\tilde{b}}{\tilde{a}+1} = \frac{b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2}{a + \frac{n}{2} + 1} = \frac{b}{a + \frac{n}{2} + 1} + \frac{\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2}{a + \frac{n}{2} + 1}$$

$$= \frac{2b}{n + 2(a+1)} + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n + 2(a+1)}$$

$$= \frac{2(a+1)}{n + 2(a+1)} \underbrace{\frac{b}{a+1}}_{Priori-Modus} + \frac{n}{n + 2(a+1)} \underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}_{Stichprobenvarianz}$$

Der MAP-Schätzer ist ein gewichtetes Mittel des Modalwertes der Priori und die für großes n dominierende Stichprobenvarianz mit $\sigma^2_{MAP} \overset{n\to\infty}{\longrightarrow} \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$.

Analoges Verfahren für den Posteriori-Erwartungswert führt zu:

$$\widehat{\sigma^2}_{PE} = \mathbb{E}(X|\tilde{a},\tilde{b}) = \frac{\tilde{b}}{\tilde{a}-1} = \frac{b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}{a+\frac{n}{2}-1}$$

$$= \frac{2(a-1)}{n+2(a-1)} \underbrace{\frac{b}{a-1}}_{Priori-EW} + \frac{n}{n+2(a-1)} \underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}_{Stichprobenvarianz}.$$

#### 3.1.2.4 Visualisierung einer Invers-Gamma-Verteilung

Ausgehend vom bereits definierten Erwartungswert und der Varianz einer Invers-Gamma-Verteilungen

$$\mathbb{V}(X|a,b) = \frac{b^2}{(a-1)^2(a-2)} \qquad \text{mit } a>1, a>2$$

werden die Parameter a und b so gewählt, dass für die weitere Betrachtung eine Unterscheidung zwischen informativer und schwacher Priori möglich ist.[Gelman et al., 2014, S. 577] Die dunkelbraun eingefärbte Dichtefunktion in Abbildung 3.6 mit $\mathbb{E}(\sigma^2|3,1)$=0.5 und $\mathbb{V}(\sigma^2|3,1)$=0.25 weist eine geringere Varianz auf, als die rötlich markierte Invers-Gamma-Verteilung mit $\sigma^2 \sim IG(4,3)$. Wohingegen die graue Dichtefunktion mit $\mathbb{E}(\sigma^2|4,10) = \frac{10}{3}$ und $\mathbb{V}(\sigma^2|4,10) \approx$5.55 einen flachen Verlauf und eine deutlich höhere Varianz im Vergleich aufweist.

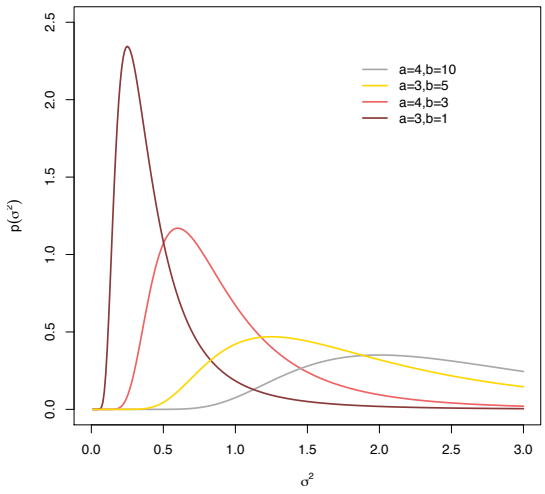


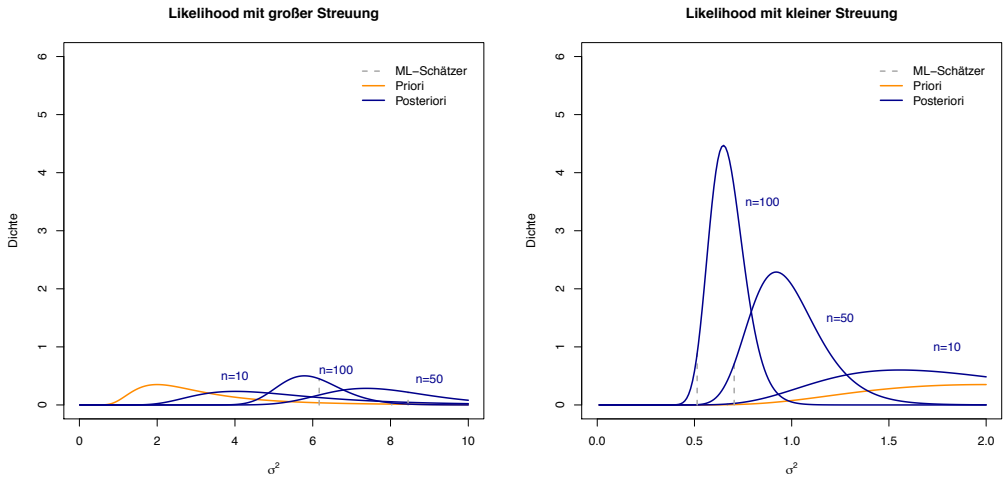

Abbildung 3.6: Visualisierung einer Invers-Gamma-Verteilung

#### 3.1.2.5 Graphische Darstellung und Interpretation

Betrachtet wird auch hier eine Fallunterscheidung zwischen schwacher und informativer Priori mit weiterer Untergliederung der Likelihood mit großer und kleiner Varianz. Aufgrund starker Gewichtung der Stichprobenvarianz bei großen Datenbeobachtungen werden in diesem Unterkapitel vorerst nur Umfänge von $n = 10$, $n = 50$ und $n = 100$ betrachtet.

Fall 1: Wahl einer schwachen Priori

Der direkte Vergleich zwischen den beiden Grafiken in Abbildung 3.7 ist nicht möglich, da die x-Achse für eine bessere Darstellung angepasst wird. Die Skalierung der y-Achse bleibt jedoch konstant. Bei der Wahl einer großen Varianz für die Likelihood ($\sigma^2 = 6$) und Priori ($\mathbb{V}(\sigma^2|4,10) \approx 5.55$), ist ein flacher Verlauf der Verteilungen erkennbar. Trägt die Likelihood viel Information, so erhöhen sich die Posteriori-Dichten und die Streuung wird mit größer werdendem Stichprobenumfang kleiner. Die jeweiligen MAP-Schätzer streben dabei gegen die Stichprobenvarianz. Für $n = 10$ nimmt die Posteriori jedoch einen ähnlichen Verlauf wie die Priori an.

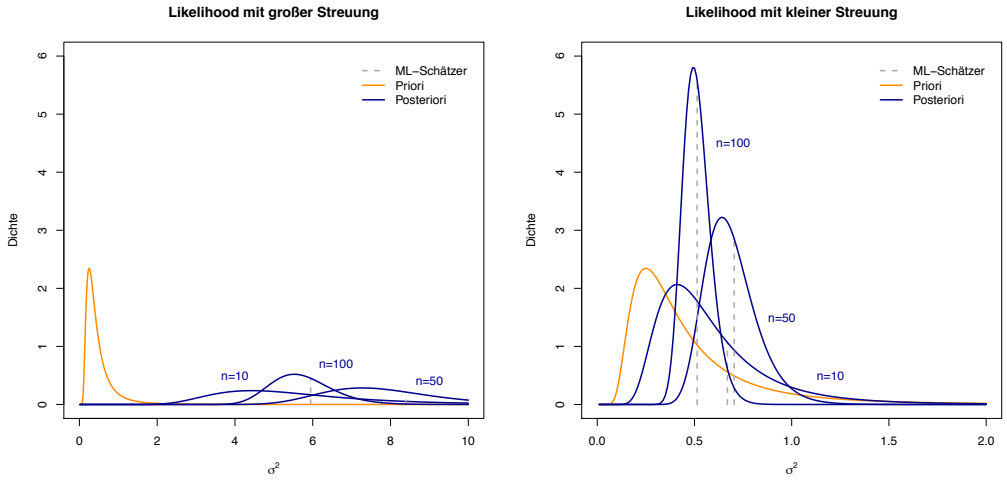



(a) $X \sim N(1,6)$ (b) $X \sim N(1,0.5)$

Abbildung 3.7: Wahl einer schwachen Priori $\sigma^2 \sim IG(4,10)$

Fall 2: Wahl einer informativen Priori

Ein Vergleich der Likelihood mit großer Streuung zwischen den Abbildungen 3.7 und 3.8 zeigt kaum Veränderung am Verlauf der Posterioris und die Priori hat für die gewählten Parameter keinen Einfluss. Bei Betrachtung der Abbildung 3.8 lässt sich jedoch eine leichte Tendenz der Posteriori für einen Stichprobenumfang $n = 10$ zur Priori erkennen. Für eine Likelihood mit kleiner Streuung weisen die Posterioris eine hohe Dichte auf und die Streuung verringert sich. Je größer n, desto mehr dominiert die Information aus den Daten und der MAP-Schätzer strebt gegen die Stichprobenvarianz.




(a) $X \sim N(1,6)$ (b) $X \sim N(1,0.5)$

Abbildung 3.8: Wahl einer informativen Priori $\sigma^2 \sim IG(3,1)$

Da die Parameterwahl beliebig gewählt wurde, ist eine Alternative für die Wahl der Hyperparameter a und b die Gleichsetzung der Erwartungswerte und Varianzen der Verteilungen.[Llera and Beckmann, 2016, Vgl. S. 3]
Der Erwartungswert einer Invers-Gamma-Verteilung wird mit dem Erwartungswert $\mu$ einer Normalverteilung gleichgesetzt. Analoges Verfahren für die Varianz und anschließende Auflösung der Ungleichung ergeben die Priori-Parameter $a'$ und $b'$:

$$\frac{b}{a-1} \stackrel{!}{=} \mu \Leftrightarrow b = \mu(a-1)$$

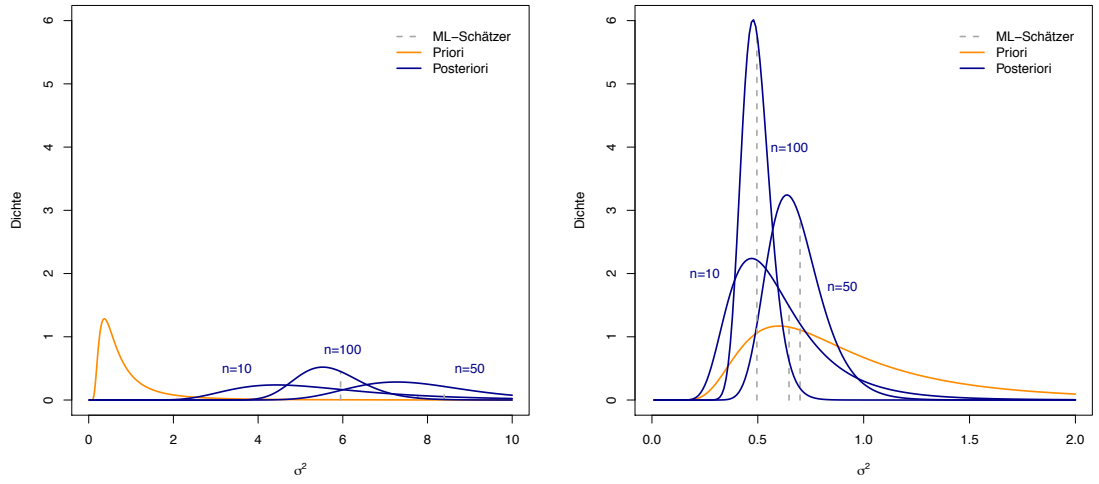
$$\frac{b^2}{(a-1)^2(a-2)} \stackrel{!}{=} \sigma^2 \stackrel{b\ eingesetzt}{\Leftrightarrow} \frac{\mu^2(a-1)^2}{(a-1)^2(a-2)} = \sigma^2 \Leftrightarrow \frac{\mu^2}{(a-2)} = \sigma^2 \Leftrightarrow a = \frac{\mu^2}{\sigma^2} + 2$$

mit

$$a' = \frac{\mu^2}{\sigma^2} + 2 \qquad\qquad b' = \mu\Big(\frac{\mu^2}{\sigma^2} + 1\Big)$$

Nun entfällt die Wahl der Parameter einer Invers-Gamma-Verteilung und es wird zwischen einer Likelihood mit großer und kleiner Varianz unterschieden.

In Abbildung 3.9 ist erkennbar, dass bei der Wahl von $\sigma^2 = 6$ die Dichten der Posteriori einen flachen Verlauf gegenüber der Priori aufweisen. Eine ähnliche Darstellung zur Abbildung 3.8 ist erkennbar. Wird die Varianz der Likelihood klein gewählt, werden die Streuungen der Posteriori geringer und die Dichten nehmen höhere Werte an. Eine Überlappung der Verteilung von Priori und Posteriori ist bei einer Likelihood mit einer hohen gewählten Varianz nicht gegeben und die Priori besitzt einen geringen Einfluss. Zu beachten gilt, dass die Skalierung der x-Achse für eine bessere Betrachtung angepasst

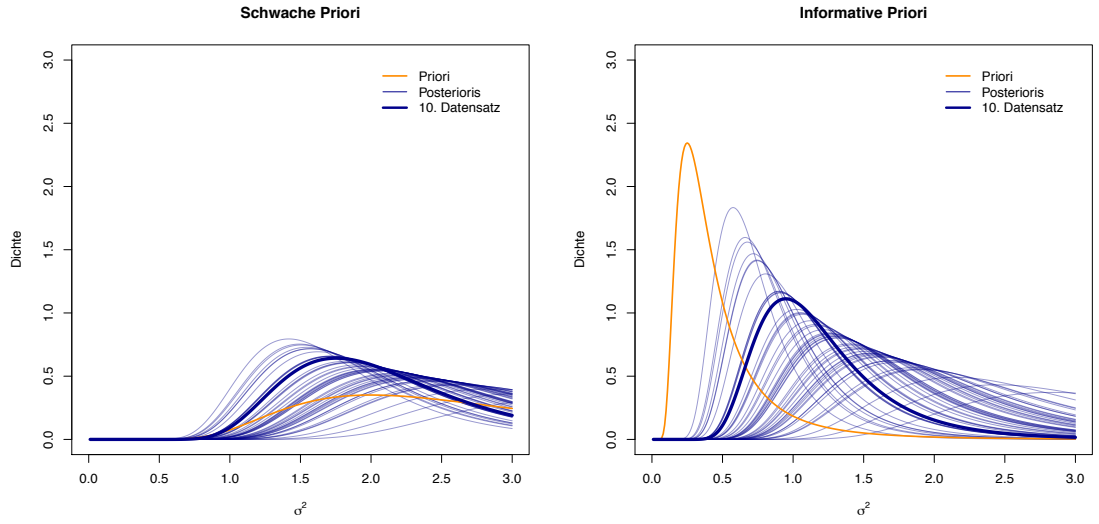



(a) X ~ N(1,6) (b) X ~ N(1,0.5)

Abbildung 3.9: Gleichsetzung der Erwartungswerte und Varianzen

wurde und somit für den ersten direkten Vergleich nicht täuschen sollte.

Da die Datenbeobachtungen erneut aus einer Normalverteilung gezogen werden, zeigt Abbildung 3.10 eine Simulation von 50 gezogenen Datensätzen von jeweils 10 Beobachtungen einer Normalverteilung mit $X \sim N(1,2)$ und einer Unterscheidung zwischen schwacher und informativer Priori. Auffallend dabei ist der nur gering höhere, aber simultane Verlauf der Posteriori-Dichte ab $\sigma^2$=1 in der rechten Grafik.




(a) X ~ N(1,2) und $\sigma_0^2 \sim IG(4,10)$ (b) X ~ N(1,2) und $\sigma_0^2 \sim IG(3,1)$

Abbildung 3.10: Simulation von 50 Datensätzen

## 3.2 Binomialverteilung

Die Binomialverteilung ist eine diskrete Wahrscheinlichkeitsverteilung. Sie basiert auf einem Zufallsexperiment mit nur zwei sich gegenseitig ausschließenden Ereignissen und der Wahrscheinlichkeit $\pi$, dass ein Erfolg eintritt. Wird das sogenannte Bernoulli-Experiment mit Wahrscheinlichkeitsfunktion

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

für n unabhängige Wiederholungen mit konstanter Wahrscheinlichkeit $\pi$ durchgeführt, ergibt sich die Binomialverteilung

$$f(x|\pi) = \binom{n}{x}\pi^x(1-\pi)^{n-x}, \qquad x = 0, 1, \ldots, n.$$

Der unbekannte Parameter $\pi$ gibt die Wahrscheinlichkeit an, dass nach n Wiederholungen ein Erfolg mit der Anzahl x eingetreten ist oder die restlichen (n-x)-male nicht.
[Voß, 2004, Vgl. S. 323ff.]

#### 3.2.0.1 Grundmodell

Sei $x_1, \ldots, x_n$ eine Zufallsstichprobe aus einer Bernoulli-Verteilung B(1,$\pi$) mit dem unbekannten Parameter $\pi \in [0, 1]$. Dann ist

$$f(x_1, \ldots, x_n|\pi) \overset{i.i.d}{=} \prod_{i=1}^{n} f(x_i|\pi) = \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i}$$

$$\propto \pi^{\sum_{i=1}^{n} x_i}(1-\pi)^{n-\sum_{i=1}^{n} x_i}$$

die auf den Daten basierende Likelihood.
Für die konjugierte Priori wird eine Betaverteilung

$$p_k(\pi) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} = \underbrace{\frac{1}{B(\alpha,\beta)}}_{konstant}\pi^{\alpha-1}(1-\pi)^{\beta-1}$$

$$\propto \pi^{\alpha-1}(1-\pi)^{\beta-1}$$

mit Parametern $\alpha, \beta > 0$ gewählt.[Held and Bové, 2014, Vgl. S. 181]
Dabei wird $\Gamma(\cdot)$ als Gammafunktion bezeichnet und der Vorfaktor $\frac{1}{B(\alpha,\beta)}$ entspricht dem

Kehrwert der unvollständigen Beta-Funktion[Rüger, 1999, Vgl. S. 201] mit

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt.$$

Die Dichtefunktion der Posteriori-Verteilung ist somit

$$\begin{aligned} p(\pi|x) &\propto f(x|\pi)p_k(\pi) \\ &\propto \pi^{\sum_{i=1}^n x_i}(1-\pi)^{n-\sum_{i=1}^n x_i}\pi^{\alpha-1}(1-\pi)^{\beta-1} \\ &\propto \underbrace{\pi^{(\alpha+\sum_{i=1}^n x_i)-1}(1-\pi)^{(\beta+n-\sum_{i=1}^n x)-1}}_{Kern\ einer\ Beta-Verteilung} \end{aligned}$$

und entspricht dem Kern einer Betaverteilung mit den Parametern

$$\tilde{\alpha} = \alpha + \sum_{i=1}^n x_i \qquad und \qquad \tilde{\beta} = \beta + n - \sum_{i=1}^n x_i.$$

Bei Betrachtung der Posteriori-Parameter setzen sich diese aus den Priori-Parametern $\alpha$ und $\beta$, der Versuche n und den erhaltenen Erfolgen $\sum_{i=1}^n x_i$ zusammen. Wird zu $\tilde{\alpha}$ die Anzahl der Erfolge hinzuaddiert, sind es bei $\tilde{\beta}$ die Misserfolge.

#### 3.2.0.2 Maximum-Likelihood-Schätzung

Durch die Ableitung der Likelihood

$$L(\pi) = f(x_1,\ldots,x_n|\pi) = \prod_{i=1}^n f(x_i|\pi) = \pi^{\sum_{i=1}^n x_i}(1-\pi)^{n-\sum_{i=1}^n x}$$

$$l(\pi) = \log L(\pi) = \sum_{i=1}^n x_i ln(\pi) + \left(n - \sum_{i=1}^n x_i\right) ln(1-\pi)$$

$$s(\pi) = \frac{\partial l(\pi)}{\partial \pi} = \frac{\sum_{i=1}^n x_i}{\pi} + \frac{n-\sum_{i=1}^n x_i}{1-\pi}$$

ergibt sich der ML-Schätzer durch Nullsetzen der Score-Funktion

$$s(\pi) \stackrel{!}{=} 0 \qquad \Rightarrow \qquad \hat{\pi}_{ML} = \frac{\sum_{i=1}^n x_i}{n}$$

und die asymptotische Verteilung des ML-Schätzers durch die Inverse von $J(\pi)$:

$$I(\pi) = -\frac{\partial s(\pi)}{\partial \pi} = \frac{x}{\pi^2} + \frac{n-x}{(1-\pi)^2}$$

$$J(\pi) = \mathbb{E}(I(\pi)) = \frac{n}{\pi(1-\pi)}$$

$$\hat{\pi}_{ML} \overset{a}{\sim} Be\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

#### 3.2.0.3 Bayes-Schätzer

Die Bayes-Schätzer einer Betaverteilung sind durch den Modus

$$\hat{\pi}_{MAP} = \frac{\tilde{\alpha} - 1}{\tilde{\alpha} + \tilde{\beta} - 2} = \frac{\alpha + \sum_{i=1}^{n} x_i - 1}{\alpha + \beta + n - 2}$$

und dem Posteriori-Erwartungswert gegeben.[Gelman et al., 2014, S. 579]

$$\hat{\pi}_{PE} = \frac{\tilde{\alpha}}{\tilde{\alpha} + \tilde{\beta}} = \frac{\alpha + \sum_{i=1}^{n} x_i}{\alpha + \beta + n} = \frac{\alpha + \beta}{(\alpha + \beta + n)} \underbrace{\frac{\alpha}{(\alpha + \beta)}}_{Priori-EW} + \frac{n}{(\alpha + \beta + n)} \underbrace{\frac{\sum_{i=1}^{n} x_i}{n}}_{\hat{\pi}_{ML}}$$

Eine Umformung stellt eine Gewichtung proportional zu $\alpha + \beta$ bzw. zu $n$ zwischen dem Erwartungswert der Priori und dem ML-Schätzer einer Binomialverteilung dar. Je größer der Datenumfang n wird, desto stärker wird $\hat{\pi}_{ML}$ gewichtet. Der Priori-Erwartungswert hingegen dominiert, wenn die Summe von $\alpha$ und $\beta$ größer als der Stichprobenumfang $n$ ist, mit $(\alpha + \beta) > n$.

#### 3.2.0.4 Visualisierung einer Beta-Verteilung

Um beide Parameter $\alpha$ und $\beta$ besser spezifizieren zu können, wird die Betaverteilung für verschiedene gewählte Parameter in Abbildung 3.11 grafisch veranschaulicht. Dabei werden der Erwartungswert und die Varianz betrachtet

$$\mathbb{E}(\pi|\alpha,\beta) = \frac{\alpha}{\alpha+\beta} \qquad \text{und} \qquad \mathbb{V}(\pi|\alpha,\beta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \qquad \text{mit } \alpha,\beta > 0$$

und basierend darauf ein Unterschied zwischen einer schwachen und informativen Priori getroffen.[Gelman et al., 2014, S.579] Den identischen Erwartungswert $\pi = 0.25$ besitzen die Beta-Verteilungen $\pi \sim Be(2,6)$ mit einer Varianz von $\mathbb{V}(\pi|2,6) \approx 0.021$ und $\pi \sim Be(10,30)$ mit $\mathbb{V}(\pi|10,30) \approx 0.0046$. Durch die Multiplikation der Parameter mit der Konstanten 5, reduziert sich die Varianz und die Priori weist eine weitaus höhere Dichte auf, in 3.11 dargestellt durch die dunkelbraune Dichtefunktion. Je größer die Parameter

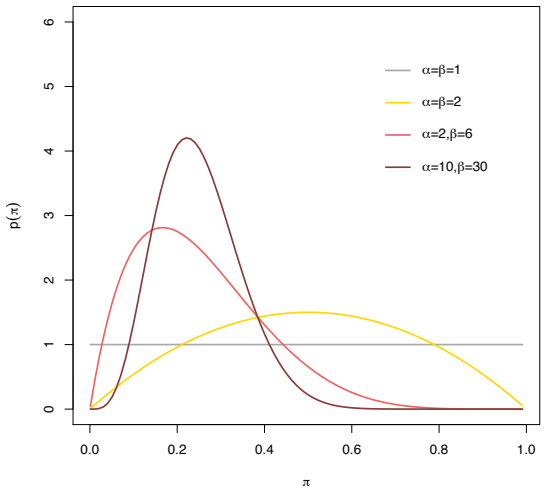
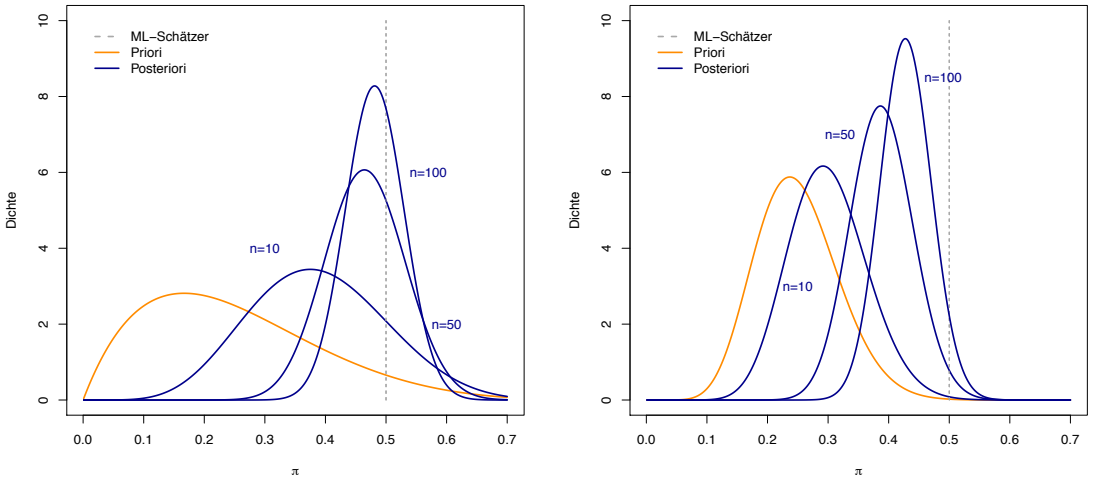



Abbildung 3.11: Visualisierung einer Beta-Verteilung

gewählt werden, desto informativer wird somit die Dichte. Sollte die Priori eine Beta-Verteilung mit $\pi \sim Be(1,1)$ darstellen, würde eine Gleichverteilung auf dem Intevall [0,1] entstehen und die Posteriori wäre komplett abhängig von den Daten.

#### 3.2.0.5 Graphische Darstellung und Interpretation

Für die graphische Darstellung in Abbildung 3.12 werden Stichprobenumfänge von $n = 10$, $n = 50$ und $n = 100$ betrachtet. Sei $\pi = 0.25$ der vermutete Parameter vor Erfassung einer Stichprobe und der ML-Schätzer mit $\hat{\pi}_{ML} = 0.5$ beliebig gewählt. In der linken Grafik wird eine schwache Priori mit $Be(2,6)$ betrachtet. Bei einem Stichprobenumfang von $n = 10$ ergeben sich die Posteriori-Parameter einer Beta-Verteilung mit $\tilde{\alpha} = \alpha + \sum_{i=1}^{n} x_i \Leftrightarrow 2 + (10 * 0.5) = 7$ und $\tilde{\beta} = \beta + n - \sum_{i=1}^{n} x_i \Leftrightarrow 6 + 10 - (10 * 0.5) = 11$. Der Posteriori-Erwartungswert $\mathbb{E}(\hat{\pi}_{ML}|7,11) \approx 0.38$ liegt zwischen den zwei gewählten Parametern $\pi$ und $\hat{\pi}_{ML}$ und die Priori besitzt einen geringen Einfluss. Für große n nehmen die Posteriori in den Bereichen ihr Maximum an, während die Priori bereits flach abfällt. Ein starker Einfluss der Priori liegt in diesem Fall nicht vor. Bei der Wahl einer informativen Priori mit $\alpha = 10$ und $\beta = 30$ dominiert vor allem in den kleinen Stichproben, hinsichtlich der Gewichtung aus 3.1.3.3., mit $(\alpha + \beta) > n \Leftrightarrow 40 > 10$ die Priori. Erkennbar daran, dass im Vergleich der folgenden zwei Abbildung Bild b) eine leichte Linksverschiebung zur Priori vorweist und für $n = 100$ die Posteriori ihr Maximum noch nicht am ML-Schätzer erreicht.




(a) Schwache Priori mit $\alpha = 2, \beta = 6$

(b) Informative Priori mit $\alpha = 10, \beta = 30$

Abbildung 3.12: Beta-Verteilung

Auch in diesem Fall fällt die Wahl der Priori-Parameter beliebig aus, weshalb im weiteren Verlauf der Priori-Erwartungswert einer Invers-Gamma-Verteilung mit dem ML-Schätzer einer Binomialverteilung gleichgesetzt wird. Analog wird für die Varianz verfahren und anschließend die Ungleichung nach a und b aufgelöst.
Es soll gelten

$$\mathbb{E}(\pi|a,b) \stackrel{!}{=} \hat{\pi}_{ML} \Leftrightarrow \frac{a}{a+b} = \pi$$

$$\text{und } \mathbb{V}(\pi|a,b) \stackrel{!}{=} \mathbb{V}(\hat{\pi}_{ML}) \Leftrightarrow \frac{ab}{(a+b)^2(a+b+1)} = \frac{\pi(1-\pi)}{n}$$

und nach Auflösung der Ungleichung ergeben sich die neuen Parameter

$$a = \pi(n-1)$$

$$b = (1-\pi)(n-1)$$

Die vollständige Auflösung der Gleichung mit Rechenweg kann im Anhang nachvollzogen werden.
Für einen beliebig gewählten Parameter $\pi$=0.5 verlaufen Priori- und Posteriori-Verteilung ähnlich und erreichen jeweils ihr Maximum am ML-Schätzer. Die Streuung der Posteriori ist dabei geringfügig kleiner. Mit zunehmender Datenbeobachtung nehmen beide Verteilungen eine höhere Dichte an.

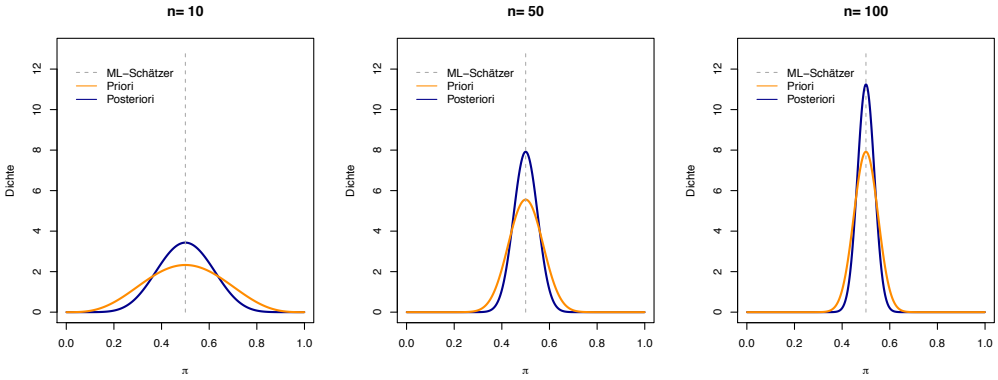



Abbildung 3.13: Gleichsetzung der Parameter

## 3.3 Normalverteilung mit bivariatem Parameter

Im folgenden Abschnitt wird der skalare Parameter auf einen bivariaten Parameter erweitert.

### 3.3.1 Erwartungswert und Varianz unbekannt

Erneut wird Erwartungwert und Varianz einer Normalverteilung betrachtet, von der beide Parameter nun als unbekannt angenommen werden - mit dem Parameter $\theta = (\mu, \sigma^2)$. Für die Priori wird a-priori-Unabhängigkeit angenommen, so dass sich für die gemeinsame Priori-Verteilung das Produkt der einzelnen Prioris ergibt:

$$p(\mu, \sigma^2) = p(\mu)p(\sigma^2).$$

Die Vermutung über den Erwartungswert hängt somit nicht von der Varianz ab und vice versa.[Rüger, 1999, Vgl. S. 208f.]
Für die Posteriori-Verteilung gilt somit

$$p(\theta|x) = p(\mu, \sigma^2|x) \propto f(x|\mu, \sigma^2)p(\mu)p(\sigma^2).$$

#### 3.3.1.1 Grundmodell

Für die Dichte der Likelihood mit $x|\mu, \sigma^2 \overset{i.i.d}{\sim} N(\mu, \sigma^2)$ gilt

$$f(x|\mu, \sigma^2) \overset{i.i.d}{=} \prod_{i=1}^{n} f(x_i|\sigma^2) = (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\Big)$$

$$\propto (\sigma^2)^{-\frac{n}{2}} exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\Big).$$

Um die Konjugiertheit zu erhalten, kann für die Verteilung von $\mu$ eine Normalverteilung und für $\sigma^2$ eine Invers-Gamma-Verteilung gewählt werden, was bereits in Kapitel 3.1 behandelt wurde. Durch die Unabhängigkeit der Prioris ergeben sich die jeweiligen semikonjugierten Prioris mit

$$p(\mu) \propto exp\Big(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\Big)$$

$$p(\sigma^2) \propto (\sigma^2)^{-(a+1)} exp\Big(-\frac{b}{\sigma^2}\Big)$$

Die Posteriori ist das Produkt der Likelihood mit den semi-konjugierten Prioris:

$$p(\mu,\sigma^2|x) \propto f(x|\mu,\sigma^2)p(\mu)p(\sigma^2)$$

$$\propto (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)(\sigma^2)^{-(a+1)}\exp\left(-\frac{b}{\sigma^2}\right)$$

$$\propto (\sigma^2)^{-(a+\frac{n}{2}+1)} \exp\left(-\frac{1}{\sigma^2}\left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}+b\right)\right)\exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

Dabei entspricht diese dem Kern einer Normal-Invers-Gamma-Verteilung und somit ergibt sich für $\theta = (\mu, \sigma^2)$:

$$\theta \sim Normal - IG\left(\frac{n\sigma_0^2\bar{x}+\sigma^2\mu_0}{n\sigma_0^2+\sigma^2}, \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}, a+\frac{n}{2}, b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}\frac{(\mu_0-\bar{x})^2}{2}\right)$$

Aufgrund einer langen Rechnung ist diese im Anhang zu finden, jedoch für die weitere Betrachtung nicht relevant.

Aus der Posteriori $p(\mu, \sigma^2|x)$ erhält man die Full-Conditional von $\mu$ mit

$$p(\mu|x,\sigma^2) \propto \exp\left(-\frac{1}{2}\left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}+\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right)\right),$$

wobei es sich um den Kern der $N\left(\frac{n\sigma_0^2\bar{x}+\sigma^2\mu_0}{n\sigma_0^2+\sigma^2}, \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)$-Verteilung handelt. Die vollständige Rechnung wurde bereits im Kapitel 3.1.1 sorgfältig aufgeführt.
Analog ergibt sich für die vollständig bedingte Dichte von $\sigma^2$

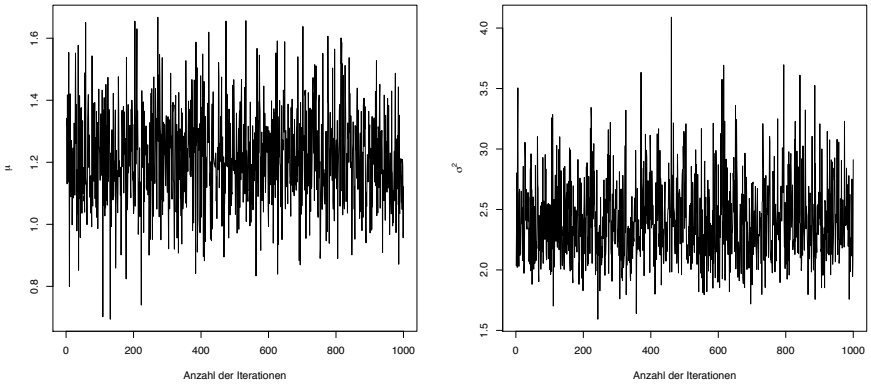
$$p(\sigma^2|x,\mu) \propto (\sigma^2)^{-(a+\frac{n}{2}+1)} \exp\left(-\frac{1}{\sigma^2}\left(b+\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}\right)\right)$$

Diese entspricht dem Kern einer $IG\left(a+\frac{n}{2}, b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$-Verteilung.

Obwohl beide Parameter als unbekannt angenommen werden, ist $\sigma^2$ in der Full-Conditional von $\mu$ bzw. $\mu$ in $p(\sigma^2|x,\mu)$ enthalten. Als eine mögliche Lösung wird dafür das Verfahren von Gibbs-Sampling im Rahmen der MCMC-Verfahren implementiert.

#### 3.3.1.2 Gibbs-Sampling

Als Startwert wird nun der Mittelwertschätzer aus einer Stichprobe vom Umfang 100 mit $X \sim N(1,2)$ gezogen. Die Parameter werden in diesem Fall beliebig gewählt. Für 1000 Iterationen wird abwechselnd aus den Full Conditionals $p(\mu|x,\sigma^2)$ und $p(\sigma^2|x,\mu)$ gezogen. Eine bereits gezogene Zufallszahl wird dabei sofort als bedingende Größe in die Full-

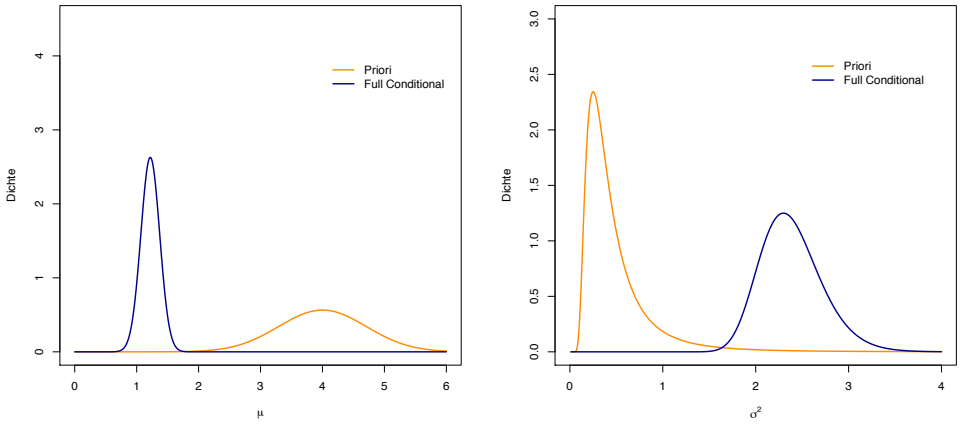



(a) Sampling von $\mu$ (b) Sampling von $\sigma^2$

Abbildung 3.14: Gibbs Sampling

Conditional der anderen Komponente eingesetzt.[Held, 2008, Vgl. S. 200] Zieht man im ersten Schritt beispielsweise $\boldsymbol{\mu}$ aus der Full-Conditional von $p(\mu|x,\sigma^2)$, so wird im darauffolgenden Schritt aus $p(\sigma^2|x,\boldsymbol{\mu})$ gezogen. Die markierte bedingte Größe $\boldsymbol{\mu}$ entspricht dabei dem gerade entstandenem Wert aus vorherigem Schritt. Nachdem für diesen Prozess 1000 Wiederholungen simuliert werden, wird die „Burn In"-Phase entfernt. In diesem Beispiel werden aufgrund der Abhängigkeit zur Wahl des Startwertes die ersten 10 Iterationen entfernt.

Aus den Samplings, graphisch dargestellt in Abbildung 3.14, werden die Schätzer für $\mu$ und $\sigma^2$ durch Berechnung der Mittelwerte ermittelt und anschließend in die jeweilige Full-Conditional als unbekannter Parameter eingesetzt. Für die Prioris $\mu_0 \sim N(4, 0.5)$ und $\sigma^2 \sim IG(3,1)$ wird die Full-Conditional in folgender Grafik verschaulicht.




(a) Full Conditional für $\mu$ (b) Full Conditional für $\sigma^2$

Abbildung 3.15: Full Conditionals mit Priori

# Kapitel 4

# Quantifizierung des Priori-Einflusses

Bisher wurde bei jeder Betrachtung die beliebige Wahl der Priori-Parameter erwähnt. Durch die Kullback-Leibler-Distanz soll im folgenden Kapitel näher darauf eingegangen werden, welchen Einfluss die Wahl der Priori-Parameter auf die Posteriori hat.
Ist es eventuell ausreichend, die Betrachtung auf die Full-Conditionals zu beschränken?

## 4.1 Normalverteilung mit unbekanntem Erwartungswert

Für normalverteilte Priori und Posteriori mit den Parametern

$$\mu_0 \sim N\left(\mu_0, \sigma_0^2\right)$$

$$\mu_{post}|\sigma^2, x \sim N\left(\mu_{post} = \frac{\sigma_0^2 n\bar{x} + \sigma^2\mu_0}{n\sigma_0^2 + \sigma^2}, \sigma_{post}^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)$$

und der Dichtefunktion

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

ergibt sich für die Kullback-Leibler-Distanz

$$D_{KL}\left(p(\mu|x)\middle\|p_k(\mu)\right) = \int_{-\infty}^{\infty} p(\mu|x) ln\left(\frac{p(\mu|x)}{p_k(\mu)}\right) d\mu = \mathbb{E}\left(ln\left(\frac{p(\mu|x)}{p_k(\mu)}\right)\right)$$

mit

$$ln\left(\frac{p(\mu|x)}{p_k(\mu)}\right) = ln\left(p(\mu|x)\right) - ln\left(p_k(\mu)\right)$$

$$= ln\left(\frac{1}{\sqrt{2\pi\sigma_{post}^2}}\exp\left(-\frac{1}{2}\frac{(\mu-\mu_{post})^2}{\sigma_n^2}\right)\right) - ln\left(\frac{1}{\sqrt{2\pi\sigma_{post}^2}}\exp\left(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right)\right)$$

$$= ln\left((\sigma_{post})^{-1}(2\pi)^{-\frac{1}{2}}\right) - \frac{1}{2}\frac{(\mu-\mu_{post})^2}{\sigma_{post}^2} - ln\left((\sigma_0)^{-1}(2\pi)^{-\frac{1}{2}}\right) + \frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}$$

$$= ln\left(\frac{\sigma_0}{\sigma_{post}}\right) + \frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2} - \frac{1}{2}\frac{(\mu-\mu_{post})^2}{\sigma_{post}^2}$$

Somit folgt

$$D_{KL}\left(p(\mu|x)\middle\|p_k(\mu)\right) = \mathbb{E}\left(ln\left(\frac{p(\mu|x)}{p_k(\mu)}\right)\right)$$

$$= ln\left(\frac{\sigma_0}{\sigma_{post}}\right) + \frac{1}{2\sigma_0^2}\mathbb{E}(\mu-\mu_0)^2 - \frac{1}{2\sigma_{post}^2}\mathbb{E}(\mu-\mu_{post})^2$$

mit den Erwartungswerten

$$\mathbb{E}(\mu-\mu_{post})^2 = \sigma_{post}^2$$

und

$$\mathbb{E}(\mu-\mu_0)^2 = \sigma_n^2 + (\mu_{post}-\mu_0)^2$$

Die Werte für die Posteriori-Parameter $\mu_{post}$ und $\sigma_{post}^2$ eingesetzt, berechnet sich die Kullback-Leibler-Distanz wie folgt:

$$D_{KL}\Big(p(\mu|x)\Big|\Big|p_k(\mu)\Big) = ln\left(\frac{\sigma_0}{\sigma_{post}}\right) + \frac{1}{2\sigma_0^2}(\sigma_{post}^2 + (\mu_{post} - \mu_0)^2) - \frac{1}{2\sigma_{post}^2}\sigma_{post}^2$$

$$= ln\left(\frac{\sigma_0}{\frac{\sigma\sigma_0}{\sqrt{n\sigma_0^2+\sigma^2}}}\right) + \frac{1}{2\sigma_0^2}\left(\frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2} + \left(\frac{\sigma_0^2 n\bar{x} + \sigma^2\mu_0}{n\sigma_0^2+\sigma^2} - \mu_0\right)^2\right) - \frac{1}{2}$$

$$= \quad \dots$$

$$= ln\left(\frac{\sqrt{n\sigma_0^2+\sigma^2}}{\sigma}\right) + \frac{\sigma^2}{2(n\sigma_0^2+\sigma^2)} + \frac{1}{2(n\sigma_0^2+\sigma^2)^2}(\sigma_0 n\bar{x} - n\mu_0\sigma_0)^2 - \frac{1}{2}$$

Ein vollständig und ausführlicher Rechenweg mit Berechnung der Erwartungswerte ist im Anhang zu finden.

### 4.1.1 Minimierung der KL-Distanz

Durch Ableitung nach $\mu_0$ und Nullsetzung erhält man das Minimum.

$$\frac{\partial D_{KL}(\,\cdot\,\|\,\cdot\,)}{\partial\mu_0} = \frac{1}{2(n\sigma_0^2+\sigma^2)^2}\left(\sigma_0^2 n^2\bar{x}^2 - 2n^2\mu_0\sigma_0^2 n\bar{x} + n^2\mu_0^2\sigma_0^2\right)$$

$$= \frac{2n^2\mu_0\sigma_0^2}{2(n\sigma_0^2+\sigma^2)^2} - \frac{2n^2\sigma_0^2\bar{x}}{2(n\sigma_0^2+\sigma^2)^2} = \frac{\mu_0 n^2\sigma_0^2 - n^2\sigma^2\bar{x}}{(n\sigma_0^2+\sigma^2)^2} \stackrel{!}{=} 0$$

$$\Leftrightarrow \mu_0 n^2\sigma_0^2 - n^2\sigma_0^2\bar{x} = 0 \Leftrightarrow n^2\sigma_0^2(\mu_0 - \bar{x}) = 0$$

$$\Rightarrow \mu_0 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial^2 D_{KL}(\,\cdot\,\|\,\cdot\,)}{\partial^2\mu_0} = \frac{n^2\sigma_0^2}{(n\sigma_0^2+\sigma^2)^2} > 0$$

Dabei handelt es sich bei $\bar{x}$ um das Minimum, da die Quadrierung nicht negativ werden kann. Abbildung 4.1 verdeutlicht, dass bei einer Stichprobe vom Umfang n=100 mit frei gewählten Parametern $\mu \sim N(1,2)$ und $\sigma_0^2 = 2$ das Minimum der Kullback-Leibler-Distanz bei etwa 1 liegt. Der Wert des Maximum-Likelihood-Schätzers ist durch die graugestrichelte Linie angedeutet. Vorgegeben durch die ausgewählten Parameter, hätte man den Priori-Erwartungswert um diesen Wert legen sollen, damit die Distanz zur Posteriori minimal wird und die Datendichte den geringsten Einfluss bewirkt. Zu beachten gilt, dass die Priori in diesem Fall abhängig von den Daten ist. Hierbei handelt es sich um ei-

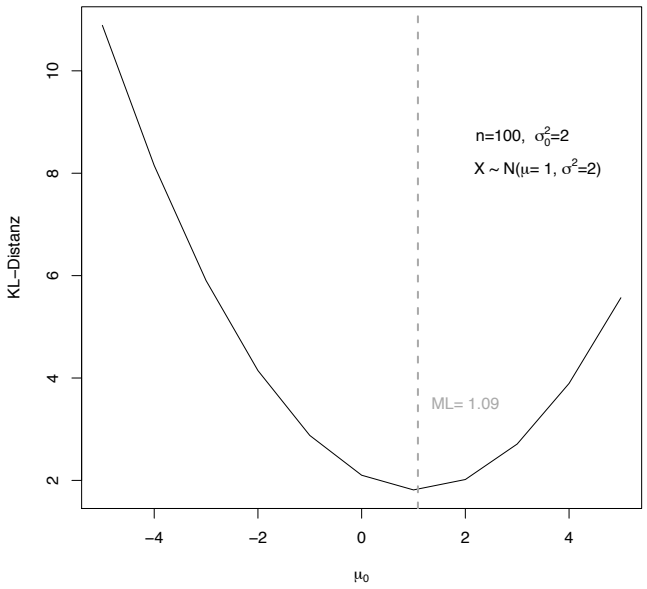



Abbildung 4.1: Minimierung der KL-Distanz

ne nachträgliche Betrachtung, welche Auswirkung die Wahl von $\mu$ bei gegebenen Daten gehabt haben könnte.

### 4.1.2 Betrachtung der Kullback-Leibler-Distanz

Auch bei der graphischen Untersuchung der Kullback-Leibler-Distanz wird bereits bekannte Grundstruktur angewendet und zwischen der Beobachtung einer schwachen bzw. informativen Priori nochmals untergliedert in eine Likelihood mit großer und kleiner Streuung. In Illustration 4.2 wird auf einen der vier Fälle näher eingegangen und eine Likelihood $X \sim N(1,6)$ mit informativ gewählter Priori $\mu_0 \sim N(4,0.2)$ betrachtet. Wurde sich im dritten Kapitel auf kleine Stichproben beschränkt, wird im Folgenden auch ein Datenumfang bis einschließlich $n = 1000$ betrachtet. Vor allem in kleinen Stichproben fällt die Distanz sehr gering aus und die Dichtefunktionen der Priori und Posteriori weisen einen ähnlichen Verlauf auf. Obwohl eine Likelihood mit großer Streuung gewählt wird und damit wenig Informationen tragen soll, dominiert diese mit zunehmendem Stichprobenumfang. Der Einfluss der Priori sinkt mit ansteigender Datenbeobachtung und die Kullback-Leibler-Distanz nimmt höhere Werte an. Die prozentuale Steigerung bei Verzehnfachung des Datenumfanges wird mit größer werdendem n geringer. Dabei verändert sich die KL-Distanz zwischen $n = 10$ und $n = 100$ um knapp das Siebenfache, während zwischen $n = 100$ und $n = 1000$ der Anstieg bei knapp 70% liegt. Der Einfluss der Priori bewirkt somit für großes n kaum eine Veränderung.

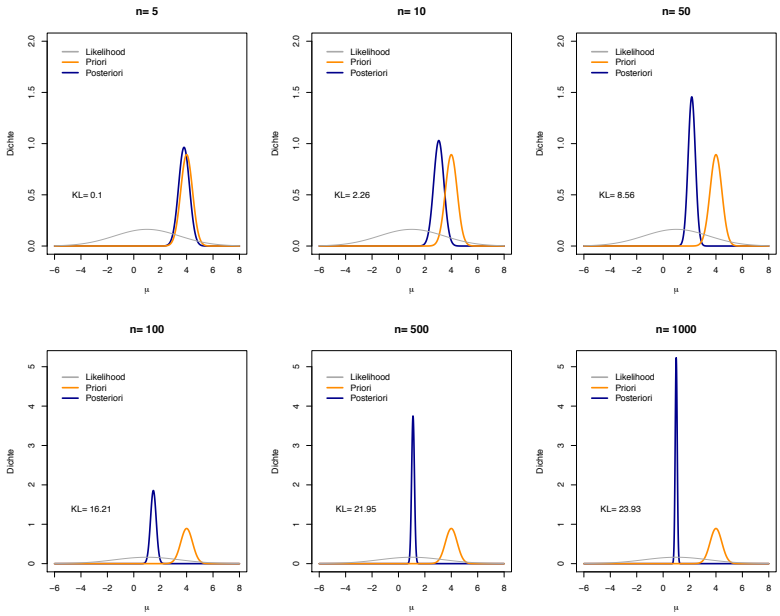



Abbildung 4.2: KL-Distanz mit $X \sim N(1,6)$ und $\mu_0 \sim N(4,0.2)$

Eine Simulation von 100 Datensätzen soll die Verteilung der Kullback-Leibler-Distanz in den folgenden Boxplots veranschaulichen. Dabei werden die gleichen Parameter wie in Abbildung 4.2 gewählt und die Simulation für verschiedene Umfänge der Datenbeobachtungen betrachtet. Der Stichprobenumfang $n = 10$ ist am kompaktesten, während $n = 50$ und $n = 100$ die breitesten Streuungen aufweisen. Mit zunehmendem n wird der Wert der KL-Distanz größer, ersichtlich dadurch, dass $n = 500$ und $n = 1000$ relativ hohe Werte haben. Die Differenz der beiden Mediane fällt im Vergleich zu den anderen Beobachtungen geringer aus, da wie bereits erwähnt, der Anstieg in den großen Stichproben geringer ausfällt.

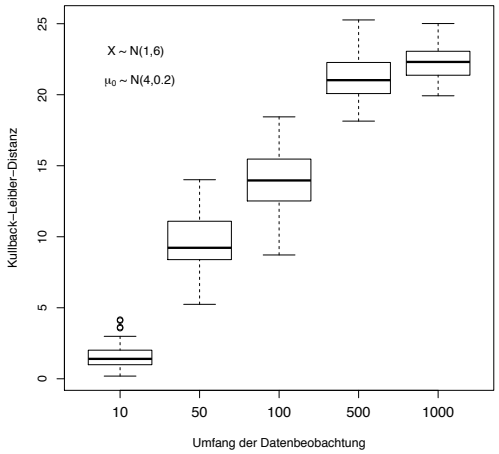



Abbildung 4.3: Verteilung der KL-Distanz

Um deshalb das Verhalten in großen Stichproben zu betrachten, wird die Kullback-Leibler-Distanz nochmals aufgeführt mit

$$D_{KL}(\,\cdot\,\|\,\cdot\,) = ln\left(\frac{\sqrt{n\sigma_0^2+\sigma^2}}{\sigma}\right) + \frac{\sigma^2}{2(n\sigma_0^2+\sigma^2)} + \frac{1}{2(n\sigma_0^2+\sigma^2)^2}\left(\sigma_0 n\bar{x} - n\mu_0\sigma_0\right)^2 - \frac{1}{2}$$

$$= \underbrace{ln\left(\sqrt{n\sigma_0^2+\sigma^2}\right)}_{\overset{n\to\infty}{\to}\infty} - ln(\sigma) + \underbrace{\frac{\sigma^2}{2(n\sigma_0^2+\sigma^2)}}_{\overset{n\to\infty}{\to}0} + \underbrace{\frac{\overbrace{\left(\sigma_0 n\bar{x} - n\mu_0\sigma_0\right)^2}^{\overset{n\to\infty}{\to}\infty}}{2\left(n\sigma_0^2+\sigma^2\right)^2}}_{\overset{n\to\infty}{\to}\infty} - \frac{1}{2}$$

Mit der Regel von L'Hospital ist die Bestimmung eines unbestimmten Ausdruckes, hier $\frac{\infty}{\infty}$, möglich. Wenn beide Funktionen im Nenner und im Zähler gegen unendlich konvergieren, kann jeweils die Ableitung berechnet werden und der Grenzwert erneut betrachtet werden.[Forster, 2004, Vgl. S. 171]
Nach zweimaliger Anwendung der Regel von de l'Hospital:

$$\lim_{n\to\infty}\frac{\left(\sigma_0 n\bar{x} - n\mu_0\sigma_0\right)^2}{2\left(n\sigma_0^2+\sigma^2\right)^2} = \frac{\sigma_0^2 n^2\bar{x}^2 - 2\sigma_0^2 n^2\bar{x}\mu_0 + n^2\mu_0^2\sigma_0^2}{2(n^2\sigma_0^4 + 2n\sigma_0^2\sigma^2 + \sigma^4} \overset{l'Hospital}{=} \frac{2n\sigma_0^2\bar{x}^2 - 4n\sigma_0^2\bar{x}\mu_0 + 2n\mu_0^2\sigma_0^2}{4n\sigma_0^4 + 2\sigma_0^2\sigma^2}$$

$$\overset{l'Hospital}{=} \frac{2\sigma_0^2\bar{x}^2 - 4\sigma_0^2\bar{x}\mu_0 + 2\mu_0^2\sigma_0^2}{4\sigma_0^4} = \frac{(\bar{x}-\mu_0)^2}{\sigma_0^2}$$

folgt für die asymptotische Verteilung der Kullback-Leibler-Distanz:

$$D_{KL}(\,\cdot\,|\,\cdot\,) \overset{n\to\infty}{\to} ln\left(\sqrt{n\sigma_0^2+\sigma^2}\right) - ln(\sigma) + \frac{(\bar{x}-\mu_0)^2}{\sigma_0^2} - \frac{1}{2}$$

Die Kullback-Leibler-Distanz konvergiert für große Datenbeobachtungen mit $n \to \infty$ gegen unendlich, nimmt jedoch aufgrund der Wurzel im ersten Term in großen Stichprobenumfängen verhältnismäßig geringere Steigerungen der Werte an.

## 4.2 Normalverteilung mit unbekannter Varianz

Für eine Normalverteilung mit unbekannter Varianz wird der Abstand zwischen zwei Invers-Gamma-Verteilungen mit Dichtefunktion

$$f(x|a,b) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\left(-\frac{b}{x}\right) \qquad \text{für } a,b > 0.$$

gemessen.[Gelman et al., 2014, S. 577]
Die Kullback-Leibler-Distanz zwischen der bereits bekannten konjugierten Priori und Posteriori

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

$$\sigma^2|x \sim IG\left(\tilde{a} = a + \frac{n}{2}, \tilde{b} = b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

berechnet sich wie folgt:

$$D_{KL}\left(p(\sigma^2|x)\middle\|p_k(\sigma^2)\right) = \mathbb{E}\left(ln\left(\frac{p(\sigma^2|x)}{p_k(\sigma^2)}\right)\right)$$

$$= \tilde{a}ln(\tilde{b}) - aln(b) + ln\left(\frac{\Gamma(a)}{\Gamma(\tilde{a})}\right) + (a-\tilde{a})\mathbb{E}\left(ln(\sigma^2)\right) - \mathbb{E}\left(\frac{1}{\sigma^2}\right)(\tilde{b}-b)$$

$$= \tilde{a}ln(\tilde{b}) - aln(b) + ln\left(\frac{\Gamma(a)}{\Gamma(\tilde{a})}\right) + (a-\tilde{a})\left(ln(\tilde{b}) - \Psi(\tilde{a})\right) - \frac{\tilde{a}}{\tilde{b}}(\tilde{b}-b)$$

$$= \tilde{a}ln(\tilde{b}) - aln(b) + ln\left(\frac{\Gamma(a)}{\Gamma(\tilde{a})}\right) + aln(\tilde{b}) - a\Psi(\tilde{a}) - \tilde{a}ln(\tilde{b}) + \tilde{a}\Psi(\tilde{a}) - \frac{\tilde{a}}{\tilde{b}}(\tilde{b}-b)$$

$$= aln\left(\frac{\tilde{b}}{b}\right) + ln\left(\frac{\Gamma(a)}{\Gamma(\tilde{a})}\right) + \Psi(\tilde{a})(\tilde{a}-a) - \frac{\tilde{a}}{\tilde{b}}(\tilde{b}-b)$$

Wobei

$$ln\left(\frac{p(\sigma^2|x)}{p_k(\sigma^2)}\right) = ln\left(\frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}(\sigma^2)^{-(\tilde{a}+1)}\exp(-\frac{\tilde{b}}{\sigma^2})\right) - ln\left(\frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)}\exp(-\frac{b}{\sigma^2})\right)$$

$$= \tilde{a}ln(\tilde{b}) - ln(\Gamma(\tilde{a})) - (\tilde{a}+1)ln(\sigma^2) - \frac{\tilde{b}}{\sigma^2} - aln(b) + ln(\Gamma(a)) + (a+1)ln(\sigma^2) + \frac{b}{\sigma^2}$$

$$= \tilde{a}ln(\tilde{b}) - aln(b) + ln\left(\frac{\Gamma(a)}{\Gamma(\tilde{a})}\right) + (a-\tilde{a})ln(\sigma^2) - \frac{1}{\sigma^2}(\tilde{b}-b)$$

und für die Erwartungswerte

$$\mathbb{E}\Big(ln(\sigma^2)\Big) = ln(\tilde{b}) - \Psi(\tilde{a}) \qquad \text{und} \qquad \mathbb{E}\Big(\frac{1}{\sigma^2}\Big) = \frac{\tilde{a}}{\tilde{b}}$$

gilt. Dabei wird $\Psi(\cdot) = \frac{\Gamma(\cdot)'}{\Gamma(\cdot)}$ als Digamma-Funktion bezeichnet. [Calin and Udrişste, 2014, Vgl. S. 60] Für die vollständige Ausführung der Kullback-Leibler-Distanz sei auf den Anhang verwiesen.

Die Werte für $\tilde{a}$ und $\tilde{b}$ eingesetzt, ergibt sich die Kullback-Leibler-Distanz zwischen zwei Invers-Gamma-Verteilungen mit

$$\begin{aligned} D_{KL}\Big(p(\sigma^2|x)\Big\|p_k(\sigma^2)\Big) &= aln\Big(\frac{b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}{b}\Big) + ln\Big(\frac{\Gamma(a)}{\Gamma(a+\frac{n}{2})}\Big) + \Psi\Big(a+\frac{n}{2}\Big)\frac{n}{2} \\ &\quad - \frac{a+\frac{n}{2}}{b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}\Big(b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 - b\Big) \\ &= aln\Big(\frac{2b+\sum_{i=1}^{n}(x_i-\mu)^2}{2b}\Big) + ln\Big(\frac{\Gamma(a)}{\Gamma(a+\frac{n}{2})}\Big) + \Psi\Big(a+\frac{n}{2}\Big)\frac{n}{2} \\ &\quad - \frac{(2a+n)\sum_{i=1}^{n}(x_i-\mu)^2}{4b+2\sum_{i=1}^{n}(x_i-\mu)^2} \end{aligned}$$

### 4.2.1 Betrachtung der Kullback-Leibler-Distanz

Wird der gleiche Fall wie im vorherigen Unterkapitel betrachtet und eine informative Priori $\sigma^2 \sim IG(3,1)$ mit einer hohen Varianz der Likelihood $X \sim N(1,6)$ gewählt, so liegt der KL-Abstand bei einer kleinen Stichprobe mit $n = 5$ bereits bei einem Wert von 4.67 und steigt mit größer werdendem n. Bei einem kleinen Datenumfang verläuft die Posteriori relativ flach, strebt jedoch ab einem Umfang von $n = 50$ gegen den in graugekennzeichneten ML-Schätzer und die Streuung verringert sich. Der Einfluss der Priori fällt dabei gering aus und die Likelihood dominiert erneut für großes n. Dabei verändert sich der Anstieg der Kullback-Leibler-Distanz in den großen Stichproben nur noch geringfügig.

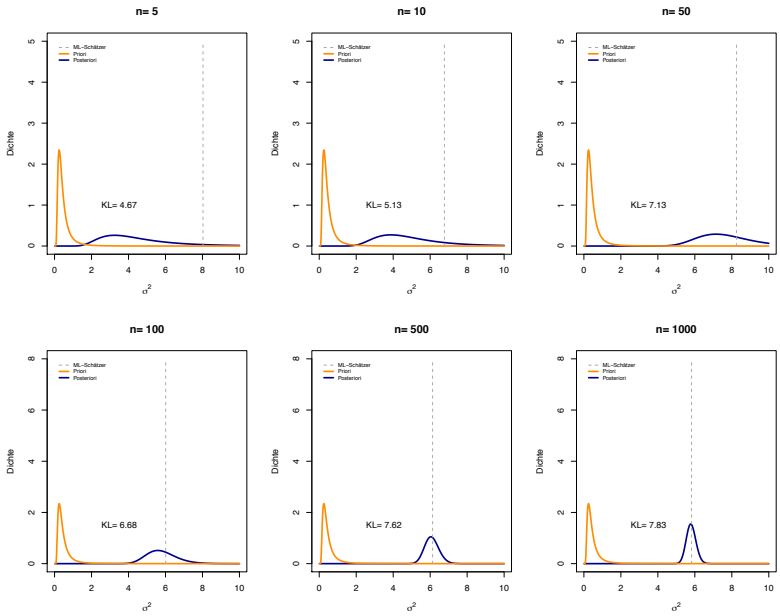



Abbildung 4.4: KL-Distanz mit $X \sim N(1,6)$ und $\sigma^2 \sim IG(3,1)$

Aufgrund der dominierenden Rolle der Likelihood für großes n, wird die Betrachtung einer klein gewählten Varianz für Priori und Likelihood auf einen kleinen Stichprobenumfang bis einschließlich $n = 50$ beschränkt. Im Vergleich zu Abbildung 4.4 fällt die Kullback-Leibler-Distanz deutlich geringer aus und auch für $n = 50$ bewirkt die Information aus den Daten nur eine minimale Veränderung. Aufgrund der Wahl einer informativen Priori mit $\sigma^2 \sim IG(3,1)$ verlaufen die Dichtefunktionen der Priori und der Posteriori ähnlich und sind um den ML-Schätzer verteilt. Für steigendes n nimmt die Posteriori-Dichte höhere Werte und eine kleiner werdende Streuung an.

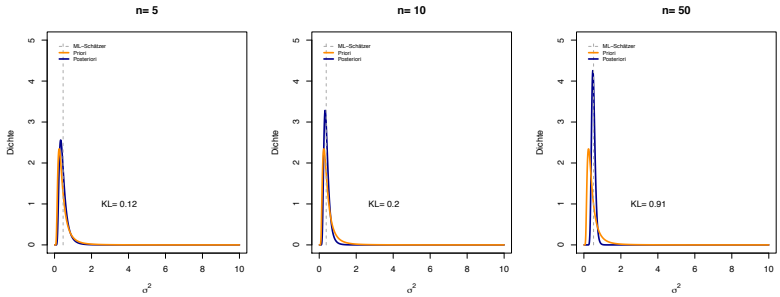
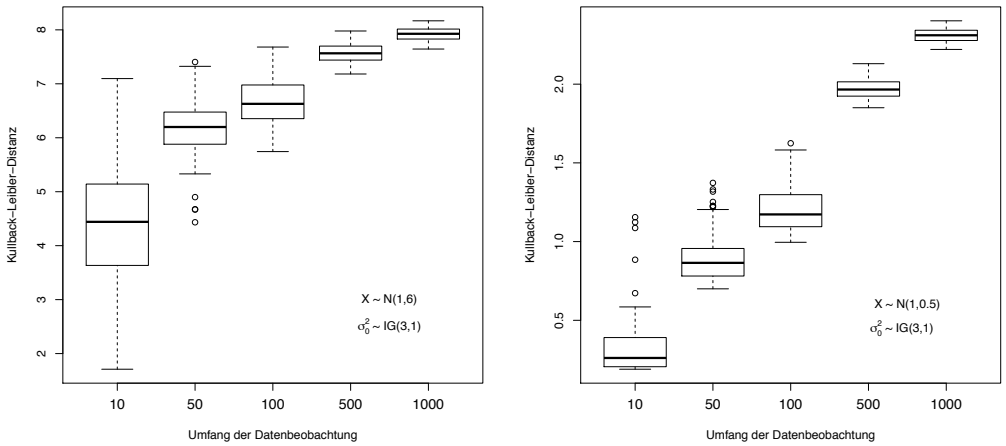



Abbildung 4.5: KL-Distanz mit $X \sim N(1,0,5)$ und $\sigma^2 \sim IG(3,1)$

Eine Übersicht über die Verteilung der KL-Distanzen veranschaulicht Abbildung 4.6. Dafür werden N=100 Datensätze aus der Normalverteilung mit der jeweiligen Größe n gezogen. Eine Verteilung der KL-Distanz über die Datensätze wird in Form von Boxplots dargestellt - dabei wurde eine Priori mit $\sigma^2 \sim IG(3,1)$ gewählt. Ein direkter Vergleich der Boxplots ist aufgrund der angepassten Skalierung der y-Achse nicht gegeben. Jedoch ist bei einem Vergleich für eine Stichprobe von $n = 10$ eine deutlich größere Streuung in a) erkennbar. In beiden Abbildungen steigt die KL-Distanz mit größer werdendem n und der Boxplot ist für $n = 1000$ am kompaktesten. Jedoch nimmt der Kullback-Leibler-Abstand für die Wahl einer Likelihood mit großer Streuung höhere Werte an.




(a) Likelihood mit großer Streuung (b) Likelihood mit kleiner Streuung

Abbildung 4.6: Informative Priori $\sigma^2 \sim IG(3,1)$

## 4.3 Binomialverteilung

Aus Kapitel 3.2. ist bereits bekannt, dass die Beta-Verteilung mit Dichtefunktion

$$f(\pi) = \frac{1}{B(\alpha,\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} \qquad \text{mit } 0 < \pi < 1$$

als konjugierte Priori-Verteilung für $\pi$ gewählt wird.

Für die konjugierte Priori und Posteriori ergibt sich

$$\pi \sim Be(\alpha,\beta)$$

$$\pi|x \sim Be\left(\tilde{\alpha} = \alpha + \sum_{i=1}^{n} x_i, \tilde{\beta} = \beta + n - \sum_{i=1}^{n} x_i\right)$$

Die Kullback-Leibler-Distanz zwischen zwei Beta-Verteilungen berechnet sich wie folgt:

$$D_{KL}\left(p(\pi|x)\middle\|p_k(\pi)\right) = \int_0^1 p(\pi|x) ln\left(\frac{p(\pi|x)}{p_k(\pi)}\right) d\pi = \mathbb{E}\left(ln\left(\frac{p(\pi|x)}{p_k(\pi)}\right)\right)$$

$$= ln\left(\frac{\tilde{C}}{C}\right) + (\tilde{\alpha}-\alpha)\mathbb{E}\left(ln(\pi)\right) + (\tilde{\beta}-\beta)\mathbb{E}\left(ln(1-\pi)\right)$$

$$= ln\left(\frac{\tilde{C}}{C}\right) + (\tilde{\alpha}-\alpha)\Big(\Psi(\tilde{\alpha}) - \Psi(\tilde{\alpha}+\tilde{\beta})\Big) + (\tilde{\beta}-\beta)\Big(\Psi(\tilde{\beta}) - \Psi(\tilde{\alpha}+\tilde{\beta})\Big)$$

$$= ln\left(\frac{\tilde{C}}{C}\right) + (\tilde{\alpha}-\alpha)\Psi(\tilde{\alpha}) + (\tilde{\beta}-\beta)\Psi(\tilde{\beta}) + (\tilde{\alpha}-\alpha+\tilde{\beta}-\beta)\Psi(\tilde{\alpha}+\tilde{\beta})$$

mit

$$
\begin{aligned}
ln\left(\frac{p(\pi|x)}{p_k(\pi)}\right) &= ln\Big(p(\pi|x)\Big) - ln\Big(p_k(\pi)\Big) \\
&= \left(ln\left(\frac{\Gamma(\tilde{\alpha}+\tilde{\beta})}{\Gamma(\tilde{\alpha})\Gamma(\tilde{\beta})}\pi^{\tilde{\alpha}-1}(1-\pi)^{\tilde{\beta}-1}\right)\right) - \left(ln\left(\frac{\Gamma(a+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}\right)\right) \\
&= \underbrace{ln\left(\frac{\Gamma(\tilde{\alpha}+\tilde{\beta})}{\Gamma(\tilde{\alpha})\Gamma(\tilde{\beta})}\right)}_{:=\tilde{C}} - \underbrace{ln\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)}_{:=C} \\
&\qquad + (\tilde{\alpha}-1-\alpha+1)ln(\pi) + (\tilde{\beta}-1-\beta+1)ln(1-\pi) \\
&= ln\left(\frac{\tilde{C}}{C}\right) + (\tilde{\alpha}-\alpha)ln(\pi) + (\tilde{\beta}-\beta)ln(1-\pi)
\end{aligned}
$$

und den Erwartungswerten

$$
\begin{aligned}
\mathbb{E}\,(ln(\pi)) &= \Psi(\alpha+x) - \Psi(\alpha+\beta+n) \\
\mathbb{E}\,(ln(1-\pi)) &= \Psi(\beta+n-x) - \Psi(\alpha+\beta+n)
\end{aligned}
$$

Auch hier ist der vollständige Rechenweg im Anhang nachvollziehbar.

Werden die Werte für $\tilde{\alpha}$ und $\tilde{\beta}$ in die Kullback-Leibler-Distanz eingesetzt, so erhält man

$$
\begin{aligned}
D_{KL}\Big(p(\pi|x)\Big\|p_k(\pi)\Big) &= ln\left(\frac{\tilde{C}}{C}\right) + x\,\Psi(\alpha+x) + (n-x)\,\Psi(\beta+n-x) + n\,\Psi(\alpha+\beta+n) \\
&= ln\left(\frac{\Gamma(\alpha+\beta+n)\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)\Gamma(\alpha+x)\Gamma(\beta+n-x)}\right) \\
&\qquad + x\,\Psi(\alpha+x) + (n-x)\,\Psi(\beta+n-x) + n\,\Psi(\alpha+\beta+n)
\end{aligned}
$$

### 4.3.1 Betrachtung der Kullback-Leibler-Distanz

In Abbildung 4.7 wird der ML-Schätzer $\hat{\pi}_{ML} = 0.5$ beliebig gewählt und durch die grau gestrichelte Linie in den Grafiken kenntlich gemacht. Für die Priori-Parameter mit $\alpha = 10$ und $\beta = 30$ beträgt der Erwartungswert $\mathbb{E}(\pi|\alpha,\beta) = 0.25$. Einer informativen Priori fällt in kleinen Stichproben, in der die Daten wenig Information tragen, mehr Gewichtung zu. Erkenntlich dadurch, dass Priori und Posteriori einen ähnlichen Verlauf aufweisen und

die KL-Distanz sehr gering ausfällt. Je mehr Information den Daten zugetragen wird, desto weniger dominiert die Priori und ein Anstieg der Kullback-Leibler-Distanz ist die Folge. In den großen Stichproben konzentriert sich die Posteriori um den ML-Schätzer und die Vermutung über den unbekannten Parameter konnte nicht bestätigt werden.

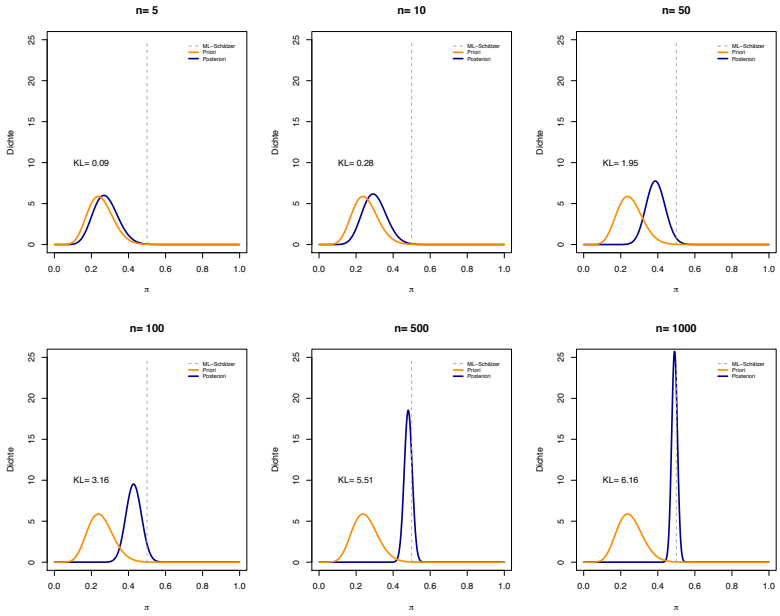



Abbildung 4.7: Informative Priori mit $\alpha$=10 und $\beta$=30

Bei der Wahl einer schwachen Priori wird das Verhalten in folgender Grafik nur in kleinen Stichproben betrachtet, da bereits für einen Stichprobenumfang von n=50 die Posteriori ihr Maximum beinahe am ML-Schätzer erreicht und der Einfluss der Priori geringer wird. Ein Vergleich beider Abbildungen weist für eine Priori mit $\alpha = 2$ und $\beta = 6$ eine größere Streuung auf und der Priori fällt weniger Gewichtung zu, als bei der Wahl einer informativen Priori.
Die Skalierung der y-Achse wurde für eine bessere Veranschaulichung in Abbildung 4.8 angepasst.

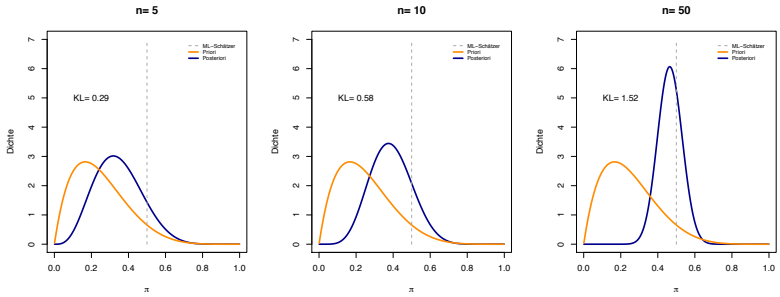
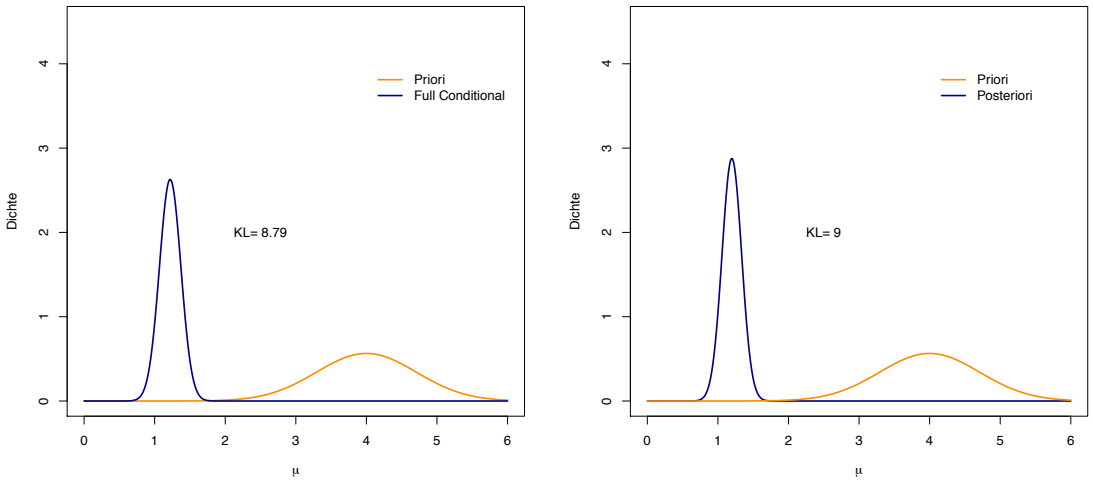



Abbildung 4.8: Schwache Priori mit $\alpha$=2 und $\beta$=6

## 4.4 Full Conditional

Zur anfangs aufgeführten Frage, ob eine Ziehung aus der Full-Conditional ausreichend wäre, wird im Folgenden ein Vergleich zwischen einer Full Conditional und einer Normalverteilung mit unbekanntem Parameter $\mu$ betrachtet. Dabei wird eine Priori $\mu_0 \sim N(4, 0.5)$ und eine Likelihood $X \sim N(1, 2)$ mit $n = 100$ beliebig gewählt. Die normalverteilte konjugierte Priori und Posteriori mit unbekanntem Parameter $\mu$ und einer KL-Distanz von 9 werden in der rechten Grafik veranschaulicht. Für die bereits bekannte Full-Conditional von $\mu$ aus Abbildung 3.15 und den identisch gewählten Parameter wie beim Gibbs-Sampling, wird die Kullback-Leibler-Distanz zwischen zwei Normalverteilungen berechnet. Dabei ist ein minimaler Rückgang der KL-Distanz, verbunden mit einer niedrigeren Dichte der Full Conditional erkennbar.




(a) Full Conditional für $\mu$

(b) Priori und Posteriori sind normalverteilt

Abbildung 4.9: Vergleich

# Kapitel 5

# Anwendung auf Realdaten

Abschließend soll ein praktisches Beispiel in diese Arbeit miteinfließen. Dabei werden die historischen Daten des Münchner Unternehmens Bayerische Motoren Werke Aktiengesellschaft näher betrachtet. Die Daten wurden über die Internetseite www.finance.yahoo.com am 02. Juni 2017 bezogen.[Finance, oD]

## 5.1 Kursverlauf der BMW-Aktie

Der zeitliche Rahmen zur Beobachtung der BMW-Aktie umfasst den 03. Januar 2000 bis einschließlich 28. April 2017. Werden vorerst die täglichen Aktienwerte betrachtet, so kommt es in der alljährlichen Weihnachtszeit und im turbulenten Börsenjahr 2008 vermehrt zu fehlenden Daten.[1]

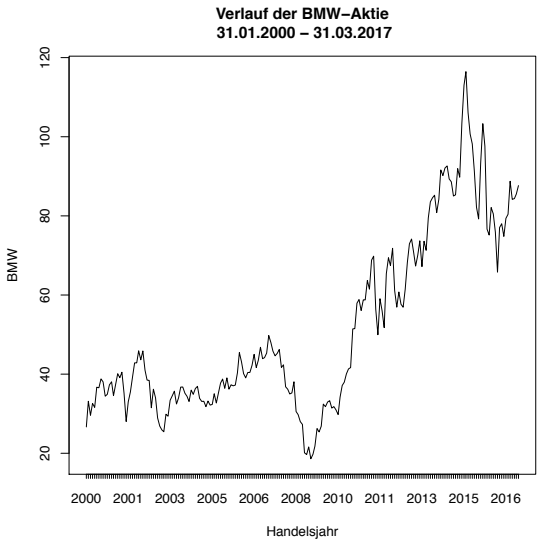



Abbildung 5.1: Zeitlicher Verlauf

[1] Aufgrund der Finanzmarktkrise musste die New Yorker Investmentbank Lehman Brother Insolvenz anmelden und es kam zu starken Kurseinbrüchen.[Bundesbank, oD]

Infolgedessen werden die Aktienkurse jeweils zum Monatsende betrachtet und man erhält 207 Beobachtungen mit jeweils sieben Variablen. Wird sich die anschließende Betrachtung vor allem auf den Schlusskurs beschränken, werden in den historischen Daten unter anderem der Eröffnungskurs, wie auch das Tagestief, bzw. -hoch aufgeführt.

Von Interesse ist der Return einer Aktie. Dieser bildet sich aus der Differenz zwischen den Schlusskursen an zwei aufeinanderfolgenden Monaten, mit Voraussetzung, dass zwischenzeitlich keine Dividende ausgezahlt wurde. [Heldt, oD]

Die logarithmierte Aktienrendite wird in Abbildung 5.2 in vier verschiedenen Grafiken visualisiert. Dabei lassen sich Zeiten mit niedriger wie auch hoher Volatilität [2] erkennen, weshalb die Rendite auch negative Werte annehmen kann. Der Boxplot der monatlichen Rendite und der Q-Q-Plot weisen auf eine Normalverteilung hin. Bei abschließender Überprüfung auf Autokorrelation für die folgende Analyse, ist eine Unkorreliertheit im Zeitverlauf der Renditen erkennbar.

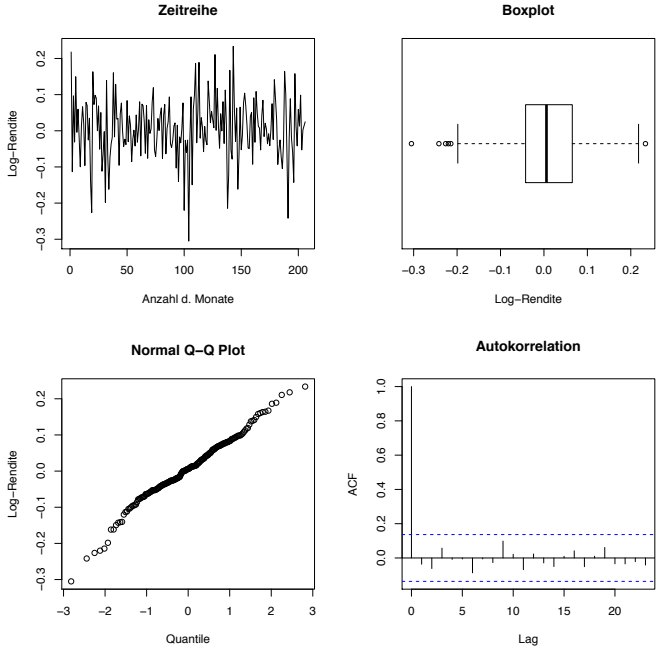
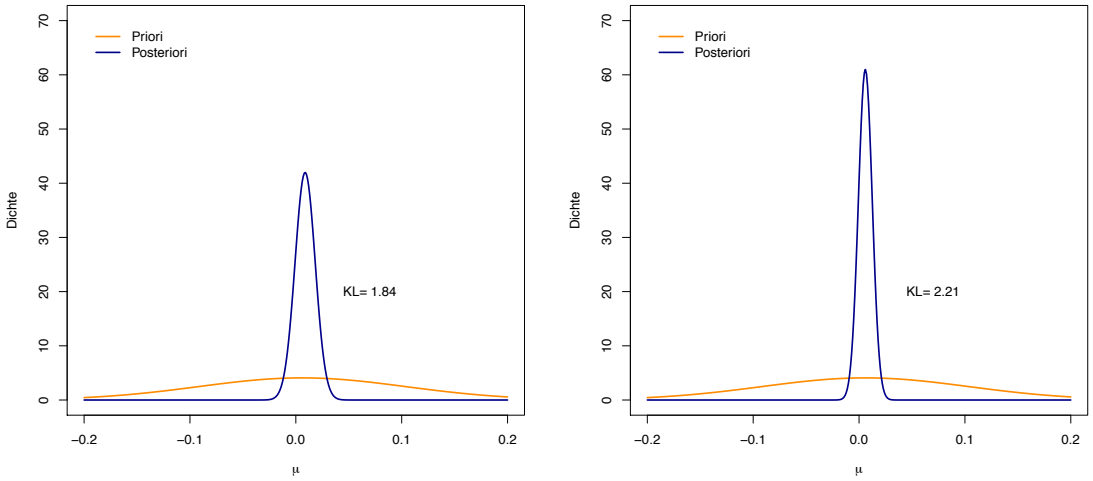



Abbildung 5.2: Überprüfung auf Normalverteilung

## 5.2 Anwendung auf die Bayesianische Statistik

Um die Thematik der Priori-/ und Posteriori-Verteilung in das Beispiel einfließen zu lassen, werden die Parameter der konjugierten Priori aus dem Zeitraum 2000 bis 2003 gewählt. Es entstehen 35 Beobachtungen mit Priori-Parametern $\mu_0 \approx 0.0002$ und $\sigma_0^2 \approx 0.011$.

[2] Volatilität bezeichnet die Schwankung einer Zeitreihe




(a) Likelihood 2004-2006

(b) Likelihood 2004-2017

Abbildung 5.3: Priori 2000-2003

Aufgrund der wiederholt genannten dominierenden Rolle der Likelihood in großen Stichproben, wird in Abbildung 5.3 zwischen zwei unterschiedlichen Stichprobenumfängen unterschieden. Die kleine Stichprobe vom Umfang n=35 in der linken Grafik umschließt das Jahr 2004 bis einschließlich Dezember 2006 mit den Parametern $\mu \approx 0.0088$ und $\sigma^2 \approx 0.0032$. Bei der grafischen Darstellung von Priori und Posteriori beträgt die Kullback-Leibler-Distanz den auf zwei Kommastellen gerundeten Wert 1.84. Wird der Zeitraum 2004 bis 2017 als Likelihood gewählt, werden 158 Beobachtungen mit Parametern $\mu \approx 0.0059$ und $\sigma^2 \approx 0.0068$ und KL=2.21 betrachtet.
Im Vergleich mit der Posteriori ist in beiden Grafiken ein flacher Verlauf der Priori sichtbar. Ein zunehmender Stichprobenumfang in b) ist mit einer geringen Streuung und einem Anstieg der Kullback-Leibler-Distanz verbunden.

# Kapitel 6

# Schlussbetrachtung

Im Rahmen der Fragestellung wurde eine Grundstruktur gewählt, die zwischen einer schwachen und informativen Priori unterscheiden sollte. Diese wurde basierend auf dem Erwartungswert und der Varianz der jeweiligen Wahrscheinlichkeitsverteilung gewählt. Eine weitere Untergliederung der Likelihood mit einer starken und geringen Streuung rundete die Fallunterscheidung ab.

Dabei konnte eine höhere Gewichtung der Likelihood in großen Stichproben durchgehend betrachtet werden. Eine falsche Vermutung bzw. Einschätzung der Priori würde somit bei großen Datenumfängen keine starke Veränderung der Posteriori bewirken.
Deshalb wurde sich oftmals nur auf eine kleine Stichprobe bis einschließlich einem Umfang von n=50 konzentriert. Sowohl bei einer informativ, wie auch bei einer schwach gewählten Priori konnte ein Einfluss auf die Posteriori festgestellt werden. Dabei fiel diesem eine unterschiedliche Gewichtung zu, je nachdem welcher Stichprobenumfang und Wahl der Priori-Parameter getroffen wurde.
Die Kullback-Leibler-Distanz wurde anschließend als Kriterium für die Entfernung zwischen Priori und Posteriori verwendet und bestätigte die theoretisch und simulativ überprüften Vermutungen.
Bei Betrachtung einer Priori und dem Vergleich zwischen einer Full Conditional und einer Posteriori konnte die Full-Conditional nur einen minimalen Anstieg der Kullback-Leibler-Distanz aufweisen.

Im Hinblick auf die persönliche Parameterwahl konnte die Kullback-Leibler-Distanz, als Kriterium für die Entfernung zwischen Priori und Posteriori, in ihrer Aussage gerechtfertigt werden.

# Literaturverzeichnis


[Bundesbank, oD] Bundesbank, D. (o.D.). https://www.bundesbank.de/Redaktion/DE/Dossier/Service/schule_und_bildung_kapitel_4.html?notFirst=true&docId=147560. Eingesehen am 24.07.2017.

[Calin and Udrişste, 2014] Calin, O. and Udrişste, C. (2014). *Geometric Modeling in Probability and Statistics*. Springer International, Switzerland.

[Fahrmeir et al., 2009] Fahrmeir, L., Kneib, T., and Lang, S. (2009). *Regression - Modelle, Methoden und Anwendungen*. Springer-Verlag Berlin Heidelberg. 2. Auflage.

[Fahrmeir et al., 2011] Fahrmeir, L., Künstler, R., and Pigeot, I. (2011). *Statistik-Der Weg zur Datenanalyse*. Springer-Verlag. Siebte Auflage.

[Finance, oD] Finance, Y. (o.D.). https://finance.yahoo.com/quote/BMW.DE/history?period1=946854000&period2=1493330400&interval=1d&filter=history&frequency=1d. Eingesehen am 02.06.2017.

[Forster, 2004] Forster, O. (2004). *Analysis 1*. Friedr. Vieweg & Sohn Verlag. 7. verbesserte Auflage.

[Gelman et al., 2014] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Taylor & Francis Group, LLC. Third Edition.

[Held, 2008] Held, L. (2008). *Methoden der statistischen Inferenz, Likelihood und Bayes*. Spektrum Akademischer Verlag.

[Held and Bové, 2014] Held, L. and Bové, D. S. (2014). *Applied Statistical Inference*. Springer-Verlag.

[Heldt, oD] Heldt, D. C. (o.D.). Aktienrendite. http://wirtschaftslexikon.gabler.de/Archiv/4294/aktienrendite-v11.html. Eingesehen am 22.07.2017.

[Itti and Baldi, 2009] Itti, L. and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, pages 1295–1306.

[Koch, 2000] Koch, K.-R. (2000). *Einführung in die Bayes-Statistik*. Springer-Verlag, Berlin, Heidelberg.

[Llera and Beckmann, 2016] Llera, A. and Beckmann, C. (2016). Estimating an inverse gamma distribution. https://www.researchgate.net/publication/301819694_Estimating_an_Inverse_Gamma_distribution.

[Rüger, 1999] Rüger, B. D. (1999). *Test- und Schätztheorie, Band I: Grundlagen*. Lehr- und Handbücher der Statistik. R.Oldenbourg Verlag München Wien, München, Wien.

[Voß, 2004] Voß, P. D. W. (2004). *Taschenbuch der Statistik*. Fachbuchverlag Leipzig im Carl Hanser Verlag. 2. verbesserte Auflage.

# Anhang

## Binomialverteilung - Alternative zur Parameterwahl der Priori

Der Erwartungswert einer Beta-Verteilung wird mit dem ML-Schätzer gleichgesetzt. Analog für die Varianz mit der asymptotischen asymptotischen Varianz des ML-Schätzers.
Es soll gelten

$$\mathbb{E}(\pi|a,b) \stackrel{!}{=} \hat{\pi}_{ML} \Leftrightarrow \frac{a}{a+b} = \pi$$

$$\text{und } \mathbb{V}(\pi|a,b) \stackrel{!}{=} \mathbb{V}(\hat{\pi}_{ML}) \Leftrightarrow \frac{ab}{(a+b)^2(a+b+1)} = \frac{\pi(1-\pi)}{n}$$

Auflösung der Ungleichung für die Erwartungswerte

$$\frac{a}{a+b} \stackrel{!}{=} \pi \Leftrightarrow a = \pi(a+b) \Rightarrow b = \frac{a-\pi a}{\pi} = a\frac{1-\pi}{\pi} = az \qquad \text{mit } z := \frac{1-\pi}{\pi}$$

Dabei gilt für $z = \frac{1-\pi}{\pi} \Leftrightarrow z\pi = 1-\pi \Leftrightarrow z\pi+\pi = 1 \Leftrightarrow (1+z)\pi = 1 \Leftrightarrow 1+z = \frac{1}{\pi}$ $(*)$
und für die Varianz

$$\frac{ab}{(a+b)^2(a+b+1)} \stackrel{!}{=} \frac{\pi(1-\pi)}{n}$$

$$\Leftrightarrow \frac{a^2z}{(a+az)^2(a+az+1)} = \frac{\pi(1-\pi)}{n}$$

$$\Leftrightarrow \frac{a^2z}{(a^2(1+z)^2)(a(1+z)+1)} = \frac{\pi(1-\pi)}{n}$$

$$\stackrel{(*)}{\Leftrightarrow} \frac{a^2z}{\frac{a^2}{\pi^2}(\frac{a}{\pi}+1)} = \frac{\pi(1-\pi)}{n} \Leftrightarrow \frac{\pi^2a^2z}{a^2(\frac{a}{\pi}+1)} = \frac{\pi(1-\pi)}{n}$$

$$\Leftrightarrow \frac{\pi^2z}{(\frac{a}{\pi}+1)} = \frac{\pi(1-\pi)}{n} \Leftrightarrow \frac{n\pi^2z}{\pi(1-\pi)} = \frac{a}{\pi}+1$$

$$\Leftrightarrow \frac{n\pi^2z}{\pi(1-\pi)} = \frac{a+\pi}{\pi}$$

$$\Leftrightarrow a = \frac{n\pi^3z}{\pi(1-\pi)} - \pi \qquad \text{für } z = \frac{1-\pi}{\pi}$$

$$= \frac{n\pi^3(1-\pi)}{\pi^2(1-\pi)} - \pi = n\pi - \pi = \pi(n-1)$$

Eingesetzt in b:

$$b = a\frac{(1-\pi)}{\pi} = \pi(n-1)\frac{(1-\pi)}{\pi} = (n-1)(1-\pi)$$

# Berechnung der Kullback-Leibler Distanz

## Normalverteilung

Für normalverteilte Priori und Posteriori mit den Parametern

$$\mu_0 \sim N\left(\mu_0, \sigma_0^2\right)$$

$$\mu_{post}|\sigma^2, x \sim N\left(\mu_{post} = \frac{\sigma_0^2 n\bar{x} + \sigma^2\mu_0}{n\sigma_0^2 + \sigma^2}, \sigma_{post}^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)$$

und der Dichtefunktion

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

berechnet sich die Kullback-Leibler-Distanz wie folgt:

$$D_{KL}\Big(p(\mu|x)\Big\|p_k(\mu)\Big) = \int_{-\infty}^{\infty} p(\mu|x) ln\left(\frac{p(\mu|x)}{p_k(\mu)}\right) d\mu = \mathbb{E}\left(ln\left(\frac{p(\mu|x)}{p_k(\mu)}\right)\right)$$

mit

$$\begin{aligned}
ln\left(\frac{p(\mu|x)}{p_k(\mu)}\right) &= ln\left(p(\mu|x)\right) - ln\left(p_k(\mu)\right) \\
&= ln\left(\frac{1}{\sqrt{2\pi\sigma_{post}^2}} \exp\left(-\frac{1}{2}\frac{(\mu-\mu_{post})^2}{\sigma_n^2}\right)\right) - ln\left(\frac{1}{\sqrt{2\pi\sigma_{post}^2}} \exp\left(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right)\right) \\
&= ln\Big((\sigma_{post})^{-1}(2\pi)^{-\frac{1}{2}}\Big) - \frac{1}{2}\frac{(\mu-\mu_{post})^2}{\sigma_{post}^2} - ln\Big((\sigma_0)^{-1}(2\pi)^{-\frac{1}{2}}\Big) + \frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2} \\
&= -ln\Big(\sigma_{post}\Big) - \frac{1}{2}\frac{(\mu-\mu_{post})^2}{\sigma_{post}^2} + ln\left(\sigma_0\right) + \frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2} \\
&= ln\left(\frac{\sigma_0}{\sigma_{post}}\right) + \frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2} - \frac{1}{2}\frac{(\mu-\mu_{post})^2}{\sigma_{post}^2}
\end{aligned}$$

Somit folgt

$$\begin{aligned}
D_{KL}\Big(p(\mu|x)\Big\|p_k(\mu)\Big) &= \mathbb{E}\left(ln\left(\frac{p(\mu|x)}{p_k(\mu)}\right)\right) \\
&= ln\left(\frac{\sigma_0}{\sigma_{post}}\right) + \frac{1}{2\sigma_0^2}\mathbb{E}(\mu-\mu_0)^2 - \frac{1}{2\sigma_{post}^2}\mathbb{E}(\mu-\mu_{post})^2
\end{aligned}$$

mit den Erwartungswerten

$$\mathbb{E}(\mu-\mu_{post})^2 = \mathbb{E}(\mu^2 - 2\mu\mu_{post} + \mu_{post}^2) = \mathbb{E}(\mu^2) - 2\mu_{post}\underbrace{\mathbb{E}(\mu)}_{\mu_{post}} + \mu_{post}^2$$

$$= \mathbb{E}(\mu^2) - 2\mu_{post}^2 + \mu_{post}^2 \qquad \text{Varianzverschiebungssatz}$$

$$= \mathbb{V}(\mu) = \sigma_{post}^2$$

und

$$\mathbb{E}(\mu-\mu_0)^2 = \mathbb{E}(\mu-\mu_{post}+\mu_{post}-\mu_0)^2 = \underbrace{\mathbb{E}(\mu-\mu_{post})^2}_{=\sigma_{post}^2} + \underbrace{2(\mu_{post}+\mu_0)\mathbb{E}(\mu-\mu_{post})}_{=0} + (\mu_{post}-\mu_0)^2$$

$$= \sigma_n^2 + (\mu_{post}-\mu_0)^2$$

Die Werte für die Posteriori-Parameter $\mu_{post}$ und $\sigma_{post}^2$ eingesetzt

$$D_{KL}\left(p(\mu|x)\middle\|p_k(\mu)\right) = ln\left(\frac{\sigma_0}{\sigma_{post}}\right) + \frac{1}{2\sigma_0^2}\left(\sigma_n^2 + (\mu_{post}-\mu_0)^2\right) - \frac{1}{2\sigma_{post}^2}\sigma_{post}^2$$

$$= ln\left(\frac{\sigma_0}{\frac{\sigma\sigma_0}{\sqrt{n\sigma_0^2+\sigma^2}}}\right) + \frac{1}{2\sigma_0^2}\left(\frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2} + \left(\frac{\sigma_0^2 n\bar{x}+\sigma^2\mu_0}{n\sigma_0^2+\sigma^2} - \mu_0\right)^2\right) - \frac{1}{2}$$

$$= ln\left(\frac{\sigma_0\sqrt{n\sigma_0^2+\sigma^2}}{\sigma\sigma_0}\right) + \frac{1}{2\sigma_0^2}\left(\frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2} + \left(\frac{\sigma_0^2 n\bar{x}+\sigma^2\mu_0-\mu_0(n\sigma_0^2+\sigma^2)}{n\sigma_0^2+\sigma^2}\right)^2\right) - \frac{1}{2}$$

$$= ln\left(\frac{\sqrt{n\sigma_0^2+\sigma^2}}{\sigma}\right) + \frac{1}{2\sigma_0^2(n\sigma_0^2+\sigma^2)^2}\left(\sigma^2\sigma_0^2(n\sigma_0^2+\sigma^2) + \left(\sigma_0^2 n\bar{x} - n\mu_0\sigma_0^2\right)^2\right) - \frac{1}{2}$$

$$= ln\left(\frac{\sqrt{n\sigma_0^2+\sigma^2}}{\sigma}\right) + \frac{\sigma^2\sigma_0^2(n\sigma_0^2+\sigma^2)}{2\sigma_0^2(n\sigma_0^2+\sigma^2)^2}$$

$$+ \frac{1}{2\sigma_0^2(n\sigma_0^2+\sigma^2)^2}\left(\sigma_0^4 n^2\bar{x}^2 - 2\sigma_0^4\bar{x}n^2\mu_0 + n^2\mu_0^2\sigma_0^4\right) - \frac{1}{2}$$

$$= ln\left(\frac{\sqrt{n\sigma_0^2+\sigma^2}}{\sigma}\right) + \frac{\sigma^2}{2(n\sigma_0^2+\sigma^2)} + \frac{1}{2(n\sigma_0^2+\sigma^2)^2}\left(\sigma_0 n\bar{x} - n\mu_0\sigma_0\right)^2 - \frac{1}{2}$$

# Berechnung der Kullback-Leibler-Distanz

## Invers-Gamma-Verteilung

Für eine Normalverteilung mit unbekannter Varianz wird der Abstand zwischen zwei Invers-Gamma-Verteilungen mit Dichtefunktion

$$f(x|a,b) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\left(-\frac{b}{x}\right) \qquad \text{für } a,b > 0.$$

gemessen.[Gelman et al., 2014, S. 577]
Bereits bekannte konjugierte Priori und Posteriori

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

$$\sigma^2|x \sim IG\left(\tilde{a} = a + \frac{n}{2}, \tilde{b} = b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

Es gilt

$$D_{KL}\left(p(\sigma^2|x)\middle\|p_k(\sigma^2)\right) = \int_{-\infty}^{\infty} p(\sigma^2|x) ln\left(\frac{p(\sigma^2|x)}{p_k(\sigma^2)}\right) d\sigma^2 = \mathbb{E}\left(ln\left(\frac{p(\sigma^2|x)}{p_k(\sigma^2)}\right)\right)$$

mit

$$\begin{aligned} ln\left(\frac{p(\sigma^2|x)}{p_k(\sigma^2)}\right) &= ln\left(\frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}(\sigma^2)^{-(\tilde{a}+1)} exp(-\frac{\tilde{b}}{\sigma^2})\right) - ln\left(\frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)} exp(-\frac{b}{\sigma^2})\right) \\ &= \tilde{a} ln(\tilde{b}) - ln(\Gamma(\tilde{a})) - (\tilde{a}+1) ln(\sigma^2) - \frac{\tilde{b}}{\sigma^2} - a ln(b) + ln(\Gamma(a)) + (a+1) ln(\sigma^2) + \frac{b}{\sigma^2} \\ &= \tilde{a} ln(\tilde{b}) - a ln(b) + ln\left(\frac{\Gamma(a)}{\Gamma(\tilde{a})}\right) + (a - \tilde{a}) ln(\sigma^2) - \frac{1}{\sigma^2}(\tilde{b} - b) \end{aligned}$$

Und somit

$$\begin{aligned} D_{KL}\left(p(\sigma^2|x)\middle\|p_k(\sigma^2)\right) &= \mathbb{E}\left(ln\left(\frac{p(\sigma^2|x)}{p_k(\sigma^2)}\right)\right) \\ &= \tilde{a} ln(\tilde{b}) - a ln(b) + ln\left(\frac{\Gamma(a)}{\Gamma(\tilde{a})}\right) + (a - \tilde{a})\mathbb{E}\left(ln(\sigma^2)\right) - \mathbb{E}\left(\frac{1}{\sigma^2}\right)(\tilde{b} - b) \end{aligned}$$

Die Erwartungswerte berechnen sich wie folgt:

$$\mathbb{E}\left(ln(\sigma^2)\right) = \int_{\Theta} p(\sigma^2|\tilde{a},\tilde{b}) ln(\sigma^2) d\sigma^2 = \int_0^{\infty} \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}(\sigma^2)^{-(\tilde{a}+1)} \exp(-\frac{\tilde{b}}{\sigma^2}) ln(\sigma^2) d\sigma^2$$

Substitution mit t:=$\frac{\tilde{b}}{\sigma^2} \Leftrightarrow \sigma^2 = \frac{\tilde{b}}{t}$.
Abgeleitet nach t ergibt sich $d\sigma^2 = -\frac{\tilde{b}}{t^2}dt \Leftrightarrow d\sigma^2 = \frac{t^2}{\tilde{b}}dt$

$$= \int_0^\infty \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\tilde{b}^{-(\tilde{a}+1)}t^{\tilde{a}+1}\exp(-t)ln(\frac{\tilde{b}}{t})\frac{\tilde{b}}{t^2}dt$$

$$= \frac{1}{\Gamma(\tilde{a})}\int_0^\infty t^{\tilde{a}+1-2}\exp(-t)\Big(ln(\tilde{b}) - ln(t)\Big)dt$$

$$= \frac{1}{\Gamma(\tilde{a})}\left(ln(\tilde{b})\underbrace{\int_0^\infty t^{\tilde{a}-1}\exp(-t)dt}_{:=\text{Gammafunktion } \Gamma(\tilde{a})} - \underbrace{\int_0^\infty t^{a-1}\exp(-t)ln(t)dt}_{=\text{Ableitung der Gammafunktion } \Gamma'(\tilde{a})}\right)$$

Die Ableitung der Gammafunktion ergibt sich dadurch, da
$t^{\tilde{a}-1} = \exp\left((\tilde{a}-1)\log(t)\right) := f(t)$
Abgeleitet nach t ergibt sich $\frac{d}{dt}f(t) = t^{\tilde{a}-1}ln(t)$

$$= \frac{1}{\Gamma(\tilde{a})}\Big(ln(\tilde{b})\Gamma(\tilde{a}) - \Gamma'(\tilde{a})\Big) = ln(\tilde{b}) - \underbrace{\frac{\Gamma(\tilde{a})}{\Gamma'(\tilde{a})}}_{=Digamma-Fkt} = ln(\tilde{b}) - \Psi(\tilde{a})$$

Und somit gilt für

$$\mathbb{E}\Big(ln(\sigma^2)\Big) = ln(\tilde{b}) - \Psi(\tilde{a})$$

Des weiteren:

$$\mathbb{E}\Big((\sigma^2)^{-1}\Big) = \int_\Theta p(\sigma^2|\tilde{a},\tilde{b})(\sigma^2)^{-1}d\sigma^2 = \int_0^\infty \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}(\sigma^2)^{-(\tilde{a}+1)}exp(-\frac{\tilde{b}}{\sigma^2})(\sigma^2)^{-1}d\sigma^2$$

$$= \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a}}\int_0^\infty (\sigma^2)^{-(\tilde{a}+1)-1}\exp(-\frac{\tilde{b}}{\sigma^2})d\sigma^2$$

$$= \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\underbrace{\int_0^\infty \frac{\tilde{b}^{\tilde{a}+1}}{\Gamma(\tilde{a}+1)}(\sigma^2)^{-(\tilde{a}+1)-1}\exp(-\frac{\tilde{b}}{\sigma^2})d\sigma^2}_{=1,\text{Dichte einer Invers-Gamma-Verteilung}}\frac{\Gamma(\tilde{a}+1)}{\tilde{b}^{\tilde{a}+1}}$$

$$= \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\frac{\tilde{a}\Gamma(\tilde{a})}{\tilde{b}^{\tilde{a}}\tilde{b}} = \frac{\tilde{a}}{\tilde{b}}$$

und damit gilt

$$\mathbb{E}\Big((\sigma^2)^{-1}\Big) = \frac{\tilde{a}}{\tilde{b}}$$

# Berechnung der Kullback-Leibler-Distanz

## Beta-Verteilung

Aus Kapitel 3.2. ist bereits bekannt, dass die Beta-Verteilung mit Dichtefunktion

$$f(\pi) = \frac{1}{B(\alpha,\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} \qquad \text{mit } 0 < \pi < 1$$

als konjugierte Priori-Verteilung für $\pi$ gewählt wird.
Für konjugierte Priori und Posteriori ergibt sich

$$\pi \sim Be(\alpha,\beta)$$

$$\pi|x \sim Be\left(\tilde{\alpha} = \alpha + \sum_{i=1}^{n} x_i, \tilde{\beta} = \beta + n - \sum_{i=1}^{n} x_i\right)$$

Die Kullback-Leibler-Distanz zwischen zwei Beta-Verteilungen berechnet sich wie folgt:

$$D_{KL}\left(p(\pi|x)\middle\|p_k(\pi)\right) = \int_{-\infty}^{\infty} p(\pi|x) ln\left(\frac{p(\pi|x)}{p_k(\pi)}\right) d\pi = \mathbb{E}\left(ln\left(\frac{p(\pi|x)}{p_k(\pi)}\right)\right)$$

mit

$$\begin{aligned}
ln\left(\frac{p(\pi|x)}{p_k(\pi)}\right) &= ln\left(p(\pi|x)\right) - ln\left(p_k(\pi)\right) \\
&= \left(ln\left(\frac{\Gamma(\tilde{\alpha}+\tilde{\beta})}{\Gamma(\tilde{\alpha})\Gamma(\tilde{\beta})}\pi^{\tilde{\alpha}-1}(1-\pi)^{\tilde{\beta}-1}\right)\right) - \left(ln\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}\right)\right) \\
&= \underbrace{ln\left(\frac{\Gamma(\tilde{\alpha}+\tilde{\beta})}{\Gamma(\tilde{\alpha})\Gamma(\tilde{\beta})}\right)}_{\tilde{C}} - \underbrace{ln\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)}_{C} + (\tilde{\alpha}-1-\alpha+1)ln(\pi) + (\tilde{\beta}-1-\beta+1)ln(1-\pi) \\
&= ln\left(\frac{\tilde{C}}{C}\right) + (\tilde{\alpha}-\alpha)ln(\pi) + (\tilde{\beta}-\beta)ln(1-\pi)
\end{aligned}$$

Und somit

$$D_{KL}\left(p(\pi|x)\middle\|p_k(\pi)\right) = ln\left(\frac{\tilde{C}}{C}\right) + (\tilde{\alpha}-\alpha)\mathbb{E}\left(ln(\pi)\right) + (\tilde{\beta}-\beta)\mathbb{E}\left(ln(1-\pi)\right)$$

Die Erwartungswerte berechnen sich wie folgt:

$$\mathbb{E}\left(ln(\pi)\right) = \int_0^1 ln(\pi)p(\pi|x)d\pi = \int_0^1 ln(\pi)\frac{\pi^{\tilde{\alpha}-1}(1-\pi)^{\tilde{\beta}-1}}{B(\tilde{\alpha},\tilde{\beta})}d\pi$$

$$= \frac{1}{B(\tilde{\alpha},\tilde{\beta})}\int_0^1 \frac{\partial}{\partial(\tilde{\alpha})}\pi^{\tilde{\alpha}-1}(1-\pi)^{\tilde{\beta}-1}d\pi = \frac{1}{B(\tilde{\alpha},\tilde{\beta})}\frac{\partial}{\partial(\tilde{\alpha})}\underbrace{\int_0^1 \pi^{\tilde{\alpha}-1}(1-\pi)^{\tilde{\beta}-1}d\pi}_{=B(\tilde{\alpha},\tilde{\beta})}$$

$$= \frac{1}{B(\tilde{\alpha},\tilde{\beta})}\left(\frac{\partial}{\partial(\tilde{\alpha})}B(\tilde{\alpha},\tilde{\beta})\right) \overset{\text{Ableitung von } ln(\cdot)}{=} \frac{\partial}{\partial(\tilde{\alpha})}\left(lnB(\tilde{\alpha},\tilde{\beta})\right)$$

$$= \frac{\partial}{\partial(\tilde{\alpha})}ln\left(\frac{\Gamma(\tilde{\alpha})\Gamma(\tilde{\beta})}{\Gamma(\alpha+\beta+n)}\right) \overset{\Gamma(\tilde{\beta}) \text{ fällt weg}}{=} \frac{\partial}{\partial(\tilde{\alpha})}\left(ln\Gamma(\tilde{\alpha}) - ln(\Gamma(\tilde{\alpha}+\tilde{\beta})\right)$$

$$= \frac{\partial}{\partial(\tilde{\alpha})}\left(ln\Gamma(\tilde{\alpha})\right) - \frac{\partial}{\partial(\tilde{\alpha})}\left(ln\Gamma(\tilde{\alpha}+\tilde{\beta})\right)$$

$$= \frac{\Gamma'(\tilde{\alpha})}{\Gamma(\tilde{\alpha}} - \frac{\Gamma'(\tilde{\alpha}+\tilde{\beta})}{\Gamma(\tilde{\alpha})}$$

$$= \Psi(\tilde{\alpha}) - \Psi(\tilde{\alpha}+\tilde{\beta}) = \Psi\left(\alpha + \sum_{i=1}^{n} x_i\right) - \Psi(\alpha+\beta+n)$$

und für $\mathbb{E}(ln(1-\pi))$ gilt mit $\pi \sim Be\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i\right)$

$$\mathbb{E}\left(ln(1-\pi)\right) = \int_0^1 ln(1-\pi)f\left(\pi|\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i\right)d\pi$$

Es gilt:

$$\mathbb{E}(ln(\pi)) = \Psi(\tilde{\alpha}) - \Psi(\tilde{\alpha}+\tilde{\beta})$$

$$\pi \sim Be(\tilde{\alpha},\tilde{\beta}) \Leftrightarrow 1-\pi \sim Be(\tilde{\beta},\tilde{\alpha})$$

Damit gilt mit $\pi \sim Be\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i\right)$

$$\mathbb{E}(ln(1-\pi)) = \Psi\left(\beta + n - \sum_{i=1}^{n} x_i\right) - \Psi(\alpha + \beta + n)$$

Eine Alternative wäre analoge Berechnung wie oben, nur abgeleitet nach $\tilde{\beta}$:

$$\mathbb{E}\,(ln(1-\pi)) = \int_0^1 ln(1-\pi)p(\pi|x)d\pi = \int_0^1 ln(1-\pi)\frac{\pi^{\tilde{\alpha}-1}(1-\pi)^{\tilde{\beta}-1}}{B(\tilde{\alpha},\tilde{\beta})}d\pi$$

$$= \frac{1}{B(\tilde{\alpha},\tilde{\beta})}\int_0^1 \frac{\partial}{\partial(\tilde{\beta})}\pi^{\tilde{\alpha}-1}(1-\pi)^{\tilde{\beta}-1}d\pi$$

$$= \ldots \text{analog, nun Ableitung nach } \tilde{\beta} \ldots$$

$$= \frac{\Gamma'(\tilde{\beta})}{\Gamma(\tilde{\beta})} - \frac{\Gamma'(\tilde{\alpha}+\tilde{\beta})}{\Gamma(\tilde{\beta})}$$

$$= \Psi(\tilde{\beta}) - \Psi(\tilde{\alpha}+\tilde{\beta}) = \Psi\left(\beta + n - \sum_{i=1}^{n} x_i\right) - \Psi(\alpha + \beta + n)$$

## Berechnung der gemeinsamen Posteriori

**Gemeinsame konjugierte Priori für $(\mu, \sigma^2)$**
Aus den vorherigen Unterkapiteln ist bereits bekannt, dass der Erwartungswert bedingt der Varianz einer Normalverteilung folgt und die konjugierte Priori der Varianz einer Invers-Gamma-Verteilung unterliegt.

$$\mu|\sigma^2 \sim N(\mu_0, \sigma_0^2)$$

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

Die Bedingung auf $\sigma^2$ in der bedingten Verteilung $\mu|\sigma^2$ weist auf die Abhängigkeit zwischen den beiden Parameter in der gemeinsamen konjugierten a-Priori-Verteilung hin. [Gelman et al., 2014, Vgl. S. 68]

$$p_k(\mu, \sigma^2) \propto p_k(\mu|\sigma^2, x) p_k(\sigma^2|a, b)$$

$$\propto (\sigma)^{-1} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right)(\sigma^2)^{-(a+1)} \exp\left(-\frac{b}{\sigma^2}\right)$$

**Gemeinsame Posteriori-Verteilung**
Multipliziert man die konjugierte gemeinsame Priori-Verteilung auf die Likelihood, so ergibt sich die gemeinsame Posteriori-Verteilung.

$$p(\mu, \sigma^2|x) \propto f(x|\mu, \sigma^2) p(\mu|\sigma^2) p(\sigma^2)$$

$$\propto (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\mu - \bar{x})^2\right)(\sigma)^{-1} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right)(\sigma^2)^{-(a+1)} \exp\left(-\frac{b}{\sigma^2}\right)$$

$$\propto (\sigma^2)^{-(a+\frac{n}{2}+1)} \exp\left(-\frac{1}{\sigma^2}\left(b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\right)(\sigma)^{-1} \exp\left(-\frac{1}{2}\left(\frac{n(\mu - \bar{x})^2}{\sigma^2} + \frac{(\mu - \mu_0)^2}{\sigma_0^2}\right)\right)$$

$$\propto (\sigma^2)^{-(a+\frac{n}{2}+1)} \exp\left(-\frac{1}{\sigma^2}\left(b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\right)$$

$$\cdot (\sigma)^{-1} \exp\left(-\frac{1}{2\sigma^2\sigma_0^2}\left((n\sigma_0^2 + \sigma^2)\left(\mu - \frac{n\sigma_0^2\bar{x} + \sigma^2\mu_0}{n\sigma_0^2 + \sigma^2}\right)^2 + \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}(\mu_0 - \bar{x})^2\right)\right)$$

$$\propto (\sigma^2)^{-(a+\frac{n}{2}+1)} \exp\left(-\frac{1}{\sigma^2}\left(b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}\frac{(\mu_0-\bar{x})^2}{2}\right)\right)$$

$$\cdot (\sigma)^{-1} \exp\left(-\frac{1}{2\sigma^2\sigma_0^2}\left((n\sigma_0^2+\sigma^2)\left(\mu-\frac{n\sigma_0^2\bar{x}+\sigma^2\mu_0}{n\sigma_0^2+\sigma^2}\right)^2\right)\right)$$

Gemeinsame Posteriori besitzt also die Parameter

$$\mu|\sigma^2, x \sim N\left(\frac{n\sigma_0^2\bar{x}+\sigma^2\mu_0}{n\sigma_0^2+\sigma^2}, \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)$$

$$\sigma^2|x \sim IG\left(a+\frac{n}{2}, b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}\frac{(\mu_0-\bar{x})^2}{2}\right)$$

Hiermit versichere ich, die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt zu haben und alle Ausführungen, die anderen Schriften wörtlich oder sinngemäß entnommen wurden kenntlich gemacht zu haben.

Gröbenzell, den 29. Juni 2017

Christina Sieber