Seminararbeit:

Konzepte einer nichtinformativen priori

Seminar Information und statistische Inferenz
im Wintersemester 2014/2015

Autor: Alexander Neumaier

Betreuer: Paul Fink / PD Dr. Dr. Christina Schneider

19. Januar 2015

Inhaltsverzeichnis

1 Definition und Einführung

Die Bayes-Inferenz ist ein sinnvolles Instrument zur Schätzung von Parameterwerten. Dabei hat die a priori Verteilung die Aufgabe, das Vorwissen, das man über den interessierenden Parameter θ hat oder nicht hat, in geeigneter Weise anzugeben. Hat man kein Vorwissen über θ, so versucht man dieses „Nichtwissen“ in Form von sogenannten nichtinformativen a priori Verteilungen darzustellen. Dies ist jedoch nicht so einfach, wie man zunächst meinen könnte. Rüger (1999) verwendet unter anderem folgende Beispiele um die grundsätzlichen Komplexitäten kurz darstellen:
Geht es im Bernoulli-Experiment darum den Parameterwert p zu schätzen, so ist es zunächst naheliegend ohne Vorwissen als a priori Verteilung eine stetige Gleichverteilung auf dem Parameterraum $\Theta = [0; 1]$ anzunehmen, da so kein Parameterwert gegenüber einem anderen als wahrscheinlicher betrachtet wird. Betrachtet man nun den transformierten Parameter $\eta = \varphi(p) = \frac{1}{p}$, so geht diese Gleichverteilung nach dem Dichtetransformationssatz in eine Verteilung mit der Dichte der Form $\pi(\eta) = \frac{1}{\eta^2}$ über. Da dies keiner Gleichverteilung enstpricht, ist der „Informationsgehalt“ über η ein anderer als der über p. Da die Transformation $\eta = \varphi(\theta) = \frac{1}{p}$ eineindeutig ist, sollte der Informationsgehalt jedoch in beiden Fällen gleich sein, da man ja „nur“ den Parameter eineindeutig tranformiert und allein dadurch sich die Informationen darüber nicht verändern sollten. (Vgl. Information einer Verteilung). Es ergibt sich also ein Widerspruch.(Vgl. Rüger; 1999, S.213)
In einem weiteren Beispiel wird nun wieder der Parameterwert p des Bernoulli-Experiments betrachtet, diesmal jedoch keine Gleichverteilung, sondern eine Verteilung mit der Dichte:

$$\pi(p) = \frac{1}{p(1-p)} \quad (0 < p < 1)$$

Grund für diese Wahl sind einerseits Varianzmaximierung, was eine Informationsminimierung zufolge hat. Andererseits erhält man dadurch auf dem transformierten Parameterraum $\gamma(\Theta) = \mathbb{R}$ mit $\gamma = \gamma(p) = log[p/(1-p)]$ (Exponentialfamilie der Verteilungsannahme des Bernoulli-Experiments) eine Gleichverteilung. Bei z Erfolgen unter n Versuchen erhält man als a posteriori-Verteilung:

$$\pi(p|z) = c(z)p^{z-1}(1-p)^{n-z-1}$$

Diese hat nun das Problem, dass für $z = 0$ oder $z = n$ das Integral über die a posteriori Dichte auf dem Paramterraum Θ nicht mehr endlich ist. Dies ist jedoch nicht zulässig, weshalb die Wahl dieser a priori Verteilung ein Probelm darstellt.(Vgl. Rüger; 1999, S.218/219)

Wie man anhand der Beispiele erkennt, treten bei der Konstruktion nichtinformativer a prioris durchaus Probleme und Paradoxa auf. Diese gilt es selbstverständlich zu minimieren. Im Folgenden werden nun Regeln betrachtet, die helfen sollen, nichtinformative a prioris festzulgen und enstehende Paradoxa möglichst abzuschwächen.

2 Bayes-Laplace-Regel

Die Bayes-Laplace-Regel basiert auf dem Prinzip vom unzureichenden Grund. Dieses besagt, dass kein Grund dazu bestehe, einen Parameterwert gegenüber einem anderen als wahrscheinlicher zu betrachten, falls von einer vollkommenen Unkenntnis über den interessierenden Parameter θ ausgegangen wird. In so einem Fall sei eine Gleichverteilung auf dem Parameterraum Θ anzunehmen.
Dieses Prinzip führt jedoch zu bereits genannten Widersprüchen, welche ersichtlich werden, wenn man eineindeutige Transformation betrachtet. Betrachtet man nun die Annahme einer Gleichverteilung über Θ, so führt die Umparametrisierung $\eta = \varphi(\theta)$ nicht im Allgemeinen zu einer Gleichverteilung auf dem transformierten Parameterraum, was uns bereits das erste Beispiel in der Einführung gezeigt hat. Somit ist die Gleichverteilungsannahme im Allgemeinen nicht invariant gegenüber eineindeutigen Transformationen. Dies sollte jedoch gegeben sein, da sich der Informationsgehalt nicht allein durch eine eineindeutige Transformation verändern sollte.
Dieses Problem tritt nicht auf, wenn Θ endlich ist, man also einen Parameterraum $\Theta = \{\theta_1, ..., \theta_N\}$ betrachtet. In diesem Fall gilt die Beziehung

$$\pi(\eta_j) = \pi(\theta_j) \qquad j = 1, ..., N$$

Die Formel besagt nichts anderes, als dass die Verteilungsannahme durch die Transformation unverändert bleibt, inbesondere also, dass die Gleichverteilungsannahme vor der Transformation in eine Gleichverteilungsannahme nach der Transformation übergeht. Somit ist bei einem endlichem Parameterraum die Gleichverteilung invariant gegenüber eineindeutigen Transformationen.
Nach Rüger (1999) liefert die Bayes-Laplace-Regel also ein zufrieden stellendes Ergebnis, wenn Θ nur aus endlich vielen Elementen besteht. In anderen Fällen sei sie aufgrund der beschriebenen Paradoxa nicht widerspruchsfrei anwendbar.
(Vgl. Rüger; 1999, S.220)
Die Regel ist auf den ersten Blick sehr einleuchtend und nachvollziehbar, da aufgrund des vorausgesetzten Nichtwissens kein Grund dazu besteht, einem Parameter eine höhere Wahrscheinlichkeit als einem beliebigen anderen zuzuordnen. Ist die Voraussetzung des endlichen Parameterraums gegeben, so ist die Regel somit auch sehr einfach und unkom-

pliziert anwendbar. Jedoch wird man in der Realität wohl öfter mit Paraneterräumen zu tun haben, die mehr als endliche viele Möglichkeiten zulassen. Von daher wird uns die Bayes-Laplace-Regel alleine wohl nicht genügen, da man dem entstehenden Paradoxon durchaus Beachtung schenken sollte.

3 Jaynes-Regel

Um das vorhandene Vorwissen, das man über θ hat zu messen, gibt es ein eine Größe, die ein Maß dafür darstelllt - die sogenannte Entropie $H(\pi)$ der Verteilung π. Je größer die Entropie in π ist, desto geringer sind die darin enthaltenen Informationen über θ. Hier setzt die Jaynes-Regel an: Da es um nichtinformative a prioris geht, ist das Vorwissen über θ minimal und die Entropie der zugehörigen a priori Verteilung π sollte somit maximal sein.(Vgl. Rüger; 1999, S.220/221)

3.1 diskreter Fall

Ist der Parameterraum abzählbar, also $\Theta = \{\theta_1, \theta_2, ..., \theta_j, ...\}$, so kann die Entropie nach der Shannon'schen Formel gemessen werden:

$$H(\pi) = -\sum \pi(\theta_j) log \pi(\theta_j)$$

Im Fall eines endlichen Parameterraums wird $H(\pi)$ durch eine Gleichverteilung auf Θ maximiert. Somit erhält man in diesem Fall dasselbe Ergebnis wie bei der Bayes-Laplace-Regel.
Ist Θ nicht mehr endlich, jedoch immer noch diskret, so ist es zur Entropiemaximierung nötig, die Menge der möglichen a priori Verteilungen auf eine Teilklasse $\mathfrak{V}$ zu beschränken. Um dies zu erreichen, wird eine Vorinformation vorausgesetzt, die aus m Restriktionen der Form:

$$E_\pi[\psi_k(\theta)] = \sum_{j=1}^{\infty} \psi_k(\theta_j)\pi(\theta_j) = \mu_k \qquad (k = 1, ..., m)$$

besteht. Hierbei sind $\psi_k(\theta)$ Funktionen und μ_k ihre bekannten a priori Erwartungswerte. Rüger (1999) führt als Beispiel den Paramterraum $\Theta = \mathbb{N}$ an. Die bekannte Vorinformation sei hierbei der Erwartungswert μ des Paramters θ

$$E_\pi[\theta] = \sum_{\theta} \theta\pi(\theta) = \mu$$

Will man unter den beschriebenen Bedingungen die Entropie $H(\pi)$ maximieren, so erhält man (vgl. Variationsrechnung) als a priori Verteilung:

$$\pi(\theta_j) = \frac{exp\{\sum_{k=1}^{m} c_k \psi_k(\theta_j)\}}{\sum_{j=1}^{\infty} exp\{\sum_{k=1}^{m} c_k \psi_k(\theta_j)\}}$$

Die Werte c_k werden dabei aus den Restriktionen bestimmt. So lange der Nenner endlich ist, hat man also eine Lösung gefunden.
Im beschriebenen Beispiel erhält man dadurch als Lösung:

$$\pi(\theta) = \frac{exp\{c\theta\}}{\sum_{\theta=1}^{\infty} exp\{c\theta\}}$$

Der Nenner ist in diesem Fall endlich, solange $exp(c) < 1$ gilt. In diesem Fall lässt sich die a priori Dichte vereinfachen:

$$\pi(\theta) = (1 - exp(c))exp(c)^{\theta - 1}$$

Im Endeffekt bekommt man also für $\Theta = \mathbb{N}$ und Erwartungswert μ mit dem Ziel der Entropiemaximierung als a priori Dichte eine Geometrische Verteilung mit Parameter $p = exp(c) = 1/\mu$ (Vgl. Erwartungswert der Geometrischen Verteilung: $\mu = 1/p$). Somit ergibt sich c in diesem Fall zu $-log\ \mu$.
(Vgl. Rüger; 1999, S.221-223)

3.2 stetiger Fall

Im stetigen Fall werden für Θ Intervalle betrachtet. Die Entropie ergibt sich nach der Shannon'schen Formel zu:

$$H(\pi) = -\int \pi(\theta) log \pi(\theta) d\theta$$

Für beschränkte Intervalle, also $\Theta = [a; b]$ wird auch hier die Entropie durch die Gleichverteilung auf Θ maximiert. Da man hier jedoch im stetigen Fall dasselbe Ergebnis wie bei der Bayes-Laplace-Regel erhält, treten dadurch logischerweise auch wieder die mit der Invarianz verbundenen Paradoxa auf.
Für unbeschränkte Intervalle sind analog zum diskreten Fall Restriktionen der Form:

$$E_\pi[\psi_k(\theta)] = \int_\Theta \psi_k(\theta)\pi(\theta)d\Theta = \mu_k \qquad (k = 1, ..., m)$$

nötig um die Verteilungen auf eine Teilklasse $\mathfrak{V}$ zu beschränken. Für $\Theta = \mathbb{R}_+$ und bekannten Erwartungswert μ als angeführtes Beispiel von Rüger (1999) lautet die Restriktion:

$$E_\pi[\theta] = \int_\Theta \theta\pi(\theta)d\theta = \mu$$

Die Maximierung der Entropie ergibt sich unter den gegebenen Bedingungen zu:

$$\pi(\theta_j) = \frac{exp\{\sum_{k=1}^m c_k\psi_k(\theta)\}}{\int_\Theta exp\{\sum_{k=1}^m c_k\psi_k(\theta)\}d\theta}$$

was wieder im Falle des endlichen Nenners die Lösung darstellt.
Im Beispiel erhält man

$$\pi(\theta) = \frac{exp\{c\theta\}}{\int_0^\infty exp\{c\theta\}d\theta}$$

als a priori Verteilung. Der Nenner ist endlich, solange $c < 0$ gilt. Dadurch geht die a priori Verteilung über in

$$\pi(\theta) = -cexp(c\theta)$$

Man erhält so für $\Theta = \mathbb{R}_+$ und bekannten Erwartungswert μ durch Entropiemaximierung eine Exponentialverteilung mit Parameter $\lambda = -c = 1/\mu$ (Vgl. Erwartungswert der Exponentialverteilung: $\mu = 1/\lambda$)
Erweitert man Θ auf $\mathbb{R}$ und setzt zusätzlich zum bekannten Erwartungswert die Varianz σ^2 als bekannt voraus, was als weiteres Beispiel von Rüger (1999) dargestellt wird, so erhält man durch selbes Vorgehen eine Normalverteilung mit Parametern μ und σ^2, also die $N(\mu, \sigma^2)$-Verteilung, als a priori Verteilung, welche die Entropie maximiert. Da in diesem Fall in der Vorinformation sowohl der bekannte Mittelwert als auch die bekannte Varianz zu beachten ist, benötigt man hier zwei Restriktionen. Diese haben die folgende Form:

$$E_\pi[\theta] = \int \theta\pi(\theta)d\theta = \mu$$

$$E_\pi[\theta^2] = \int \theta^2\pi(\theta)d\theta = \sigma^2 + \mu^2$$

Die zweite Restriktion ergibt sich hier aus dem Varianzverschiebungssatz.
(Vgl. Rüger; 1999, S.223-225)
Leider tritt auch bei offenen Intervallen genau wie bei den Geschlossenen und der Bayes-Laplace-Regel wieder das Problem der Invarianz auf: Maximiert man zuerst die Entropie der Verteilung von θ und transformiert anschließend in $\varphi(\theta)$, so erhält man nicht die selbe Verteilung, wie wenn zuerst transformiert und dann die Entropie maximiert wird. Damit entsteht wieder das Paradoxon, dass sich der Informationsgehalt der a priori Verteilung allein durch eine eineindeutige Transformation verändert. Nach Rüger (1999) sei somit

die Jaynes-Regel ein gutes Konzept, solange Θ diskret ist, da man in diesem Fall auf das eben genannte Problem nicht stößt. Geht man jedoch auf den stetigen Fall über, so sei die Regel nicht mehr befriedigend.(Vgl. Rüger; 1999, S.226)
In meinen Augen stellt die Jaynes-Regel durch die Anwendung des Entropiebegriffs einen ebenso nachvollziehbaren Ansatz dar wie die Bayes-Laplace-Regel, aber da sie im Prinzip auch diesselben Problem aufweist, kann Rüger zugestimmt werden. Was im Falle von abzählbar unendlichen Parameterräumen und bei offenen Intervallen zusätzlich noch zu bemängeln ist, ist die Tatsache, dass man Vorinformationen (Restriktionen) voraussetzt, obwohl man ja eigentlich von nichtinformativen a prioris spricht, also eigentlich gar kein Vorwissen hat.

4 Jeffreys-Regel

Nun wird nach einer Regel gesucht, die das bisher auftretende Problem der Invarianz berücksichtigen kann und somit folgendem Prinzip, dem sogenannten Prinzip der Invarianz, folgen soll: „Wird für den Parameter θ nach der betreffenden Regel die nichtinformative a priori Verteilung $\pi(\theta)$ bestimmt und anschließend θ in $\varphi(\theta)$ transformiert, so soll die so entstehende Verteilung von $\varphi(\theta)$ dieselbe sein wie diejenige, die sich ergibt, wenn zuerst θ in $\varphi(\theta)$ transformiert und dann die Regel auf $\varphi(\theta)$ angewandt wird.“
Die Jeffreys-Regel versucht nun genau diesen Ansatz sowohl für reele θ als auch für vektorielle θ zu berücksichtigen. Dazu wird diesmal nicht nur der Parameterraum Θ beachtet, sondern auch die zugrundeliegende Verteilungsannahme $\mathfrak{P}$, insbesondere im Bezug auf die Fisher-Information. Es werden also Fisher-reguläre Verteilungsannahmen vorausgesetzt.(Vgl. Rüger; 1999, S.226/227)

4.1 eindimensionaler Fall

Hat man einen reelen Parameter θ und eine Fisher-reguläre Verteilungsannahme $\mathfrak{P} = \{f(x;\theta) : \theta \in \Theta\}$ mit der Fisher-Information $I(\theta)$ der i.i.d. Stichprobe X gegeben, so ist nach der Jeffreys-Regel die nichtinformative priori Verteilung zu wählen, die proportional zu $\sqrt{I(\theta)}$ ist.
Statt $I_X(\theta)$ darf auch immer $I_{X1}(\theta)$ verwendet werden, da die a priori Verteilung unabhängig von n ist. Die Invarianz der Jeffreys-Regel im eindimensionalen Fall folgt unter diesen Bedingungen aus folgender Formel:

$$\tilde{I}_X[\varphi(\theta)] \cdot [\varphi'(\theta)]^2 = I_X(\theta)$$

Dabei entspricht $\varphi(\theta)$ dem transformierten Parameter und $\tilde{I}_X$ dessen Fisher-Information. Durch die Invarianzeigenschaft alleine ergibt sich jedoch noch keine Begründung dieser Regel. Diese soll nun erfolgen:
Die Fisher-Information kann als Maß dafür interpretiert werden, wie gut man in der Umgebung von θ die Dichten $f(x;\theta)$ trennen kann (Vgl. Information einer Stichprobe). Betrachtet man kleine Unterschiede $\triangle\theta$ zwischen zwei Parameterwerten, so gilt die

Beziehung

$$logf(x;\theta + \triangle\theta) - logf(x;\theta) \approx K(\theta;x)\triangle\theta$$

K entspricht dabei der Score-Funktion und stellt quasi das Änderungsverhalten der Log-Likelihood in Abhängigkeit vom unbekannten Parameter θ und dem Ergebnis der Stichprobe x relativ zu $\triangle\theta$ dar. Die Fisher-Information hingegen ist als quadrierter Erwartungswert der Score-Funktion auffassbar, also:

$$I(\theta) = E_\theta K^2(\theta;X)$$

Aus diesen beiden Formeln folgt, dass man die Fisher-Information als mittleres lokales Maß dafür interpretieren kann, wie sich ein kleiner Unterschied $\triangle\theta$ zwischen zwei Parameterwerten auf die Log-Likelihood-Funktion auswirkt. Die Änderung des Funktionswertes ist natürlich relativ zu $\triangle\theta$ zu sehen. Somit beschreibt die Fisher-Information im Endeffekt nichts anderes an, als die informationstheoretische Bedeutung eines möglichen Parameters θ. Es scheint also sinnvoll, Parametern mit höherer informationstheoretischer Bedeutung eine höhere Dichte zuzuweisen. Somit wählt man die a priori Dichte $\pi(\theta)$ proportional zu $\sqrt{I(\theta)}$. (Wurzelziehen erfolgt um die Skala des Parameters einzuhalten)
(Vgl. Rüger; 1999, S.228/229)
Betrachtet wird nun als Beispiel wieder das Bernoulli-Experiment. Somit lautet die Fisher-Information zum unbekannten Parameter p:

$$I(p) = \frac{n}{p(1-p)}$$

Als a priori Verteilung, die proportional zur Wurzel dieser Fisher-Information ist (n kann wie erläutert gleich 1 gesetzt werden), ergibt sich:

$$\pi(p) = \frac{1}{\pi\sqrt{p(1-p)}}$$

(Das π im Nenner entspricht der Kreiszahl und stellt dabei den Normierungsfaktor dar).Diese nichtinformative a priori Verteilung stellt eine Beta(1/2;1/2)-Verteilung dar. Ein weiteres Beispiel wird angegeben durch den unbekannten Parameter λ einer Poisson-Verteilung. Hier ergibt sich die Fisher-Information zu $I(\lambda) = n/\lambda$. Als a priori Verteilung ergibt sich:

$$\pi(\lambda) = \frac{1}{\sqrt{\lambda}}$$

(Vgl. Rüger; 1999, S.102/231)

4.2 mehrdimensionaler Fall

Betrachtet man nun einen vektoriellen Parameter $\theta = (\theta_1, \theta_2, ..., \theta_r)$ wieder mit der Fisher-regulären Verteilungsannahme $\mathfrak{P} = \{f(x;\theta) : \theta \in \Theta\}$ und Fisher-Informationsmatrix $I(\theta)$ der i.i.d. Stichprobe X, so ist nach der Jeffreys-Regel die a priori Verteilung zu wählen, deren Dichte proportional zu $\sqrt{detI(\theta)}$ ist.
Auch hier darf wieder $I_X(\theta)$ durch $I_{X1}(\theta)$ ersetzt werden. Die Invarianz der Jeffreys-Regel im mehrdimensionalen Fall folgt aus

$$\triangle \tilde{I}_X(\varphi) \triangle' = I_X(\theta)$$

Dabei entspricht φ dem transformierten Parameter,$\tilde{I}$ der Fisher-Information des transformierten Parameters und $\triangle$ der Hesse-Matrix.
Auch hier soll wieder eine Begründung der Regel erfolgen:
Im Gegensatz zum eindimensionalen Fall betrachtet man nun einen Vektor für kleine Unterschiede zwischen den Parameterwerten $\triangle\theta = (\triangle\theta_1, ..., \triangle\theta_r)$, ebenso einen r-dimensionalen Vektor für die Score-Funktion $K(\theta; x)$ und eine (r x r)-dimensionierte Fisher-Informationsmatrix. Die Beziehung lautet analog zum eindimensionalen Fall:

$$log f(x; \theta + \triangle\theta) - log f(x; \theta) \approx K(\theta; x)(\triangle\theta)'$$

Um nun den Unterchied zwischen $f(x;\theta)$ und $f(x;\theta + \triangle\theta)$ informationstheoretisch zu erfassen und proportional zu $\triangle\theta$ zu messen (wie im eindimensionalen Fall) ist es hier noch zusätzlich nötig aus der Fisher-Informationsmatrix eine Gewichtsfunktion zu bilden, die die Skala des Parameters einhält. Hierbei bietet sich $\sqrt{detI(\theta)}$ an, was zur Jeffreys-Regel führt:

$$\pi(\theta) \sim \sqrt{detI(\theta)} = \prod_{i=1}^{r} \sqrt{\lambda_i(\theta)}$$

$\lambda_1(\theta), ..., \lambda_r(\theta)$ bezeichnen dabei die Eigenwerte der Matrix $I(\theta)$
Eine weitere Möglichkeit um eine Gewichtsfunktion zu bilden, wäre beispielsweise die Bildung der Spur statt der Determinante.
(Vgl. Rüger; 1999, S.230/231)
Als Beispiel für einen mehrdimensionalen Paramterraum betrachtet Rüger (1999) eine Multinomialverteilung: Ein Zufallsexperiment mit r Ereignissen $A_1, ..., A_r$ und Wahrscheinlichkeiten $p_1 = P(A_1), ..., p_r = P(A_r)$ mit $p_1 + ... + p_r = 1$ wird n mal unabhängig voneinander durchgeführt. Betrachtet man die Statistik $Z = (Z_1, ..., Z_r)$ mit Z_i = Hufigkeit von A_i unter den n Versuchen, so erhält man eine Multinomialverteilung mit den Paramtern n und $p_1, ..., p_r$.(Kurz gesagt handelt es sich bei der Multinomialverteilung um eine allgemeinere Binomialverteilung, die mehr als zwei Ausgänge des Zufallsex-

periments zulässt) Die Fisher-Informationsmatrix dazu ist eine (r-1)x(r-1)-Matrix mit Determinante $n^{r-1}/(p_1...p_r)$. Die Jeffreys-Regel liefert somit folgende a priori Verteilung:

$$\pi(p_1, ..., p_{r-1}) = c/\sqrt{p_1...p_r}$$

Dabei ergibt sich $p_r = 1 - (p_1 + ... + p_{r-1})$ und c stellt den Normierungsfaktor dar. Als a priori Randverteilung eines p_i erhält man die Beta(1/2;1/2)-Verteilung wie schon beim Bernoulli-Experiment im eindimensiolen Fall. Somit ist in diesem Fall die Jeffreys-Regel auch invariant gegenüber einer Veränderung der Dimension des Parameters.(Vgl. Rüger; 1999, S.106/232)

4.3 Paradoxa der Jeffreys-Regel

Um die sich ergebenden Paradoxa zu veranschaulichen verwendet Rüger (1999) das Gauß-Experiment. D.h. $X = (X_1, ..., X_n)$ stellt i.i.d. Stichproben eines $N(\mu, \sigma^2)$-verteilten Merkmals dar. Betrachtet werden dabei 3 Fälle:

1.μ unbekannt, σ^2 bekannt:

In diesem Fall lautet die Fisher-Information für den unbekannten Parameter $I(\mu) = n/\sigma^2$. Da sowohl n als auch σ^2 bekannt sind, handelt es sich dabei um eine Konstante. Somit ergibt sich als a priori Verteilung ebenfalls eine Konstante (also eine Gleichverteilung), die gleich 1 gesetzt werden kann, also $\pi(\mu) = 1$

2.μ bekannt, σ^2 unbekannt:

Hier ergibt sich die Fisher-Information (diesmal für σ^2) zu $I(\sigma^2) = n/2\sigma^4$. Nach der Jeffreys-Regel erhält man als a priori Dichte als eine Verteilung, die proportional zu $1/\sigma^2$ ist. Der Proportionalitätsfaktor kann gleich 1 gesetzt werden. Somit erhält man:

$$\pi_{\sigma^2}(\sigma^2) = 1/\sigma^2$$

Daraus ergibt sich, wenn man statt σ^2 den Parameter σ betrachtet, die a priori Verteilung $\pi_\sigma = 1/\sigma$, welche man aufgrund der Invarianz auch erhalten hätte, wenn man von vornherein durch σ parametrisiert hätte.

Die a posteriori Verteilung nach Beobachtung x ergibt sich zu:

$$\pi(\sigma^2|x) \sim (\frac{1}{\sigma^2})^{\frac{n+2}{2}} exp\{-ns_\mu^2/2\sigma^2\}$$

mit $s_\mu^2 = \frac{1}{n}\sum(x_i - \mu)^2$.

3.μ unbekannt, σ^2 unbekannt:

Diesmal hat man einen vektoriellen unbekannten Parameter $\theta = (\mu, \sigma^2)$ und es ist somit

diesmal nötig die Fisher-Informationsmatrix zu bestimmen. Diese ergibt sich zu:

$$\begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/2\sigma^4 \end{pmatrix}$$

mit der Determinante $n^2/2\sigma^6$. Nach der Jeffreys-Regel ergibt sich dadurch eine a priori Verteilung der Form

$$\pi_\theta(\mu, \sigma^2) = 1/\sigma^3$$

Daraus erhält man für den Parameter $\eta = (\mu, \sigma)$ eine a priori Verteilung der Form $\pi_\eta = 1/\sigma^2$, welche man auch wieder aufgrund der Invarianz erhalten hätte, wenn man von vornherein durch η parametrisiert hätte.
Die a posteriori Verteilung nach Beobachtung x ergibt sich zu:

$$\pi(\sigma^2|x) \sim (\frac{1}{\sigma^2}^{\frac{n+2}{2}} exp\{-(n-1)s^2/2\sigma^2\}\pi(\mu, \sigma^2|x) \sim (\frac{1}{\sigma^2})^{\frac{n+3}{2}} exp\{-(n-1)s^2/2\sigma^2\}exp\{-n(\bar{x}-\mu)^2/2\sigma^2\}$$

mit $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$.
Die Randdichte von σ^2 hat dabei die Form

$$\pi(\sigma^2|x) \sim (\frac{1}{\sigma^2})^{\frac{n+2}{2}} exp\{-(n-1)s^2/2\sigma^2\}$$

(Vgl. Rüger; 1999, S.233/234)
Bei genauerer Betrachtung dieser Fälle entstehen Probleme:
Die bedingte Dichte für $\eta = (\mu, \sigma)$ lautet $\pi(\mu, \sigma) = 1/\sigma^2$. Da diese Dichte unabhängig von μ ist, müsste σ als unabhängig von μ gelten, was zur Folge hätte, dass die bedingte Verteilung von σ unter μ ebenfalls die Dichte $\pi(\sigma|\mu) = 1/\sigma^2$ hätte. Das bedeutet nun, dass sich die Dichte auch dann nicht ändern würde, wenn μ bekannt wäre, da sie unabhängig von μ ist. Im 2.Fall jedoch, wo μ als bekannt vorausgesetzt war, ergibt sich als a priori Dichte für σ $\pi_\sigma = 1/\sigma$. Diese stimmt jedoch nicht mit der obigen bedingten Dichte überein. Es ergibt sich also ein Widerspruch. Somit wären die Regeln der Wahrscheinlichkeitsrechnung bei bedingungsloser Anwendung der Jeffreys-Regel nicht allgemeingültig.
Ein weiteres Problem wird ersichtlich, wenn man die a posteriori Dichten im 2. und 3.Fall betrachtet. Mit Hilfe des Dichtetransformationssatzes ergibt sich, dass ns_μ^2/σ^2 im Fall 2 a posteriori nach $\mathcal{X}_n^2$ verteilt ist. Im Fall 3 bei Betrachtung der Randdichte ergibt sich, dass $(n-1)s^2/\sigma^2$ a posteriori ebenfalls $\mathcal{X}_n^2$ verteilt ist. D.h. bei Übergang vom bekannten auf den unbekannten Fall verliert man keinen Freiheitsgrad, was aber eigentlich der Fall sein sollte. Das wiederum hätte zur Folge, dass man zum Beispiel bei

der Schätzung von Konfidenzintervallen dasselbe Ergebnis erhält, wenn μ bekannt ist wie wenn μ unbekannt ist. Somit ist auch hier wieder ein Verstoß gegen die Regeln der Wahrscheinlichkeitsrechung ersichtlich.
Im Allgemeinen lassen sich die Probleme durch das sogennante Marginalisierungsproblem zusammenfassen. Dieses lautet:"Ist die Regel zur Festlegung einer a priori Verteilung invariant gegenüber einer Bildung von Marginalverteilungen (d.h. gegenüber Projektionen des Parameters)?"
Betrachtet man beispielsweise den Parameter $\theta = (\theta_1, \theta_2)$, die Projektionsabbildung $(\theta_1, \theta_2) \rightarrow \theta_1$ und eine Regel J zur Festlegung der a priori Verteilung. Geht man gleich von θ auf θ_1 über und wendet dann die Regel J an, so sollte man im Falle der Invarianz beim Marginalisierungsproblem diesselbe Verteilung erhalten, wie wenn man zuerst die Regel J anwendet, und anschließend aus der gemeinsamen a priori Dichte $\pi(\theta_1, \theta_2)$ die Marginalverteilung $\pi_1(\theta_1)$ festlegt. Im vorhin erläuterten Gauß-Experiment ist diese Invarianz, wie bereits gezeigt, nicht gegeben. Es wird dabei auch sehr deutlich, welche Probleme daraus enstehen. Somit sei es nach Rüger sehr wichtig, diese Invarianzeigenschaft einzufordern.
(Vgl. Rüger; 1999, S.235-237)
Für den eindimensionalen Fall stellt die Jeffreys-Regel nach Rüger ein zufriedenstellendes Konzept dar, im mehrdimensionalen Fall weise es jedoch aus eben beschriebenen Gründen noch Mängel auf. Dort soll die Anwendung auf Fälle beschränkt werden, in denen die Invarianz gegeben ist (wie z.B. im obigen Fall bei der Multinomialverteilung). Zudem merkt Rüger an, dass diese Fälle sich offenbar dadurch auszeichnen, dass die einzelnen Komponenten $\theta_1, ..., \theta_r$ des Parameters θ eine gleiche inhaltliche Bedeutung hätten und somit in gewisser Weise vertauschbar seien.(Vgl. Rüger; 1999, S.237)
Die Jeffreys-Regel stellt wohl einen guten Ansatz dar, wenn es um eindimensionale Intervalle geht. Dadurch können auf jeden Fall schon mal Fälle mit diesem Konzept bearbeitet werden, die nur mit Hilfe der Jaynes- oder Bayes-Laplace-Regel nicht zu lösen sind, ohne dass man auf mathematische Paradoxa stößt. Steht man nun vor dem Problem, dass Θ mehrdimensional ist, so sollte man die Regel nicht bedenkenlos anwenden und sich gegenfalls bei jedem Fall genau überlegen, ob und inwieweit man von der Regel Gebrauch machen kann. Dazu kann man auf jeden Fall Rügers Idee beachten, indem man sich die inhaltliche Bedeutung der einzelnen Parameter klar macht.

5 Zusammenfassung und Ausblick

Im Folgenden seien die einzelnen Regeln mit ihren Möglichkeiten und Paradoxa nach Rüger (1999) noch einmal kurz zusammengefasst:
Die Bayes-Laplace-Regel eigne sich besonders gut, wenn man einen diskreten, endlichen Parameterraum $\Theta = \{\Theta_1, ..., \Theta_n\}$ zugrunde liegen hat. In diesem Fall erhält man als nichtinformative a priori Verteilung eine diskrete Gleichverteilung auf Θ. Im Falle von anderen Parameterräumen sei die Anwendung der Regel aufgrund ihrer Invarianzprobleme bei eineindeutigen Umparametrisierungen nicht optimal.
Auch die Jaynes-Regel, die das Ziel verfolgt, die Entropie, ein Maß für das Nichtwissen, zu maximieren, sei bei diskretem, endlichem Θ gut anwendbar und führt in diesem Fall auch zum selben Ergebnis wie die Bayes-Laplace-Regel. Ist Θ nicht mehr endlich, aber noch diskret, also $\Theta = \{\Theta_1, ..., \Theta_k, ...\}$, so sei die Regel zwar grundsätzlich anwendbar, jedoch benötige man in diesem Fall gewisse Vorinformationen, um die Menge aller möglichen a priori Verteilungen auf eine Teiklasse $\mathfrak{V}$ zu beschränken. Im stetigen Fall, also bei Intervallen, erscheine sie aufgrund der selben Probleme wie bei der Bayes-Laplace-Regel ebenfalls unbefriedigend.
Für den stetigen Fall eigne sich die Jeffreys-Regel gut, da die Invarianzprobleme bei eineindeutigen Umparametrisierungen hier nicht auftauchen, wobei man beachten müsse, dass man bei mehrdimensionalen Parameterräumen hier mit dem Marginalisierungsproblem konfrontiert wird. Bei dieser Regel wird zusätzlich zu Θ auch noch die zugrundeliegende Verteilungsannahme $\mathfrak{P}$ betrachtet. Man erhält eine nichtinformative a priori Verteilung, die proportional zu $\sqrt{I(\theta)}$ ist (bzw. proportional zu $\sqrt{detI(\theta)}$ im mehrdimensionalen Fall). Im diskreten Fall ist die Regel aufgrund der Fisher-Regularitätsvoraussetzungen nicht anwendbar.

Konzepte zur Konstruktion von nichtinformativen a priori Verteilungen gibt es, wie gezeigt, mehrere. Jedoch weist jede von ihnen, was uns auch die Zusammenfassung zeigt, Schwachpunkte auf. Was leider somit fehlt, ist eine allgemeine Regel, welche man bei einem vorhandenen „Nichtwissen“ immer anwenden kann, um dieses wiederzugeben. Des Weiteren wäre es wohl durchaus noch sinnvoll, jeweils für den unendlichen, diskreten Fall und den mehrdimensionalen stetigen Fall noch ein zufriedenstellendes Konzept zu finden, ohne auf zu große mathematische Problematiken zu stoßen

Auch sollte man bedenken, dass man bei der Konstruktion dieser prioris immer von einem kompletten a priori Nichtwissen ausgeht. Jedoch gibt es wohl unterscheidliche Möglichkeiten, ein Nichtwissen zu definieren oder aufzufassen. Von daher spielt die subjektive Sichtweise hier wohl eine große Rolle. Da wir an mehreren Stellen dieser Arbeit auf Widersprüche gestoßen sind, kann man sich durchaus fragen, ob es überhaupt sinnvoll ist, nichtinformative a prioris in dieser Form zu konstruieren und den Informationsbegriff für solche statistischen Fragestellungen zu verwenden.

Literaturverzeichnis

Rüger (1999). *Test- und Schätztheorie Band I Grundlagen*, Oldenbourg.