3 Weitere Aspekte der statistischen Entscheidungstheorie

3.1 Bayes-Verfahren in der statistischen Entscheidungstheorie

3.1.1 Begrifflicher Hintergrund und „Erinnerungen“

Proposition 3.1 (Allgemeines Theorem von Bayes)

Seien X und U zwei Zufallsvariablen mit gemeinsamer Wahrscheinlichkeitsfunktion $f_{X,U}(\cdot)$ bzw. Dichte $f_{X,U}(\cdot)$ (bezüglich eines dominierenden σ-finiten Maßes $\nu \otimes \lambda$) und den bedingten Wahrscheinlichkeitsfunktionen bzw. bedingten Dichten $f_{X|U}(\cdot|u)$ und $f_{U|X}(\cdot|x)$ (bezüglich ν bzw. λ).
Dann gilt:

$$f_{U|X}(u|x) = \frac{f_{X|U}(x|u) \cdot f_U(u)}{f_X(x)} \tag{3.2}$$

mit

$$f_X(x) = \int f_{X|U}(x|u) \cdot f_U(u)\, d\nu(u). \tag{3.3}$$

Bem. 3.2

Bei stetigem U mit Dichte $f_U(u)$ erhält man also Proposition 3.1 mit

$$f_X(x) = \int f_{X|U}(x|u) \cdot f_U(u) du. \tag{3.4}$$

Im Fall von diskreten Zufallsvariablen X und U – mit $\mathcal{U}$ als Träger von U – ergibt sich

$$p(\{U = u\}|\{X = x\}) = \frac{p(\{X = x\}|\{U = u\}) \cdot p(\{U = u\})}{p(\{X = x\})}$$

mit

$$p(\{X = x\}) = \sum_{u \in \mathcal{U}} p(\{X = x\}|\{U = u\}) \cdot p(\{U = u\}).$$

Bem. 3.3 (Normierungskonstante)

$f_X(x)$ aus (3.3) spielt die Rolle einer reinen Normierungskonstante, die nicht von u abhängt. Häufig reicht es daher, $f_{X|U}(x|u) \cdot f_U(u)$ zu berechnen. Da man weiß, dass sich insgesamt eine Wahrscheinlichkeitsdichte ergeben muss, kennt man implizit auch die Normierungskonstante.

Bem. 3.4 (Def. Posteriori-Verteilung und ergänzende Bem.)

Gegeben sei ein datengestütztes Entscheidungsproblem $((\mathbb{A}, \Theta, l(\cdot)); (\mathcal{X}, \sigma(\mathcal{X}), p_\vartheta(\cdot))$, wobei $p_\vartheta(\cdot)$ die Wahrscheinlichkeitsfunktion bzw. Dichte $f_\vartheta(\cdot)$ besitze, und eine Priori-Verteilung auf $(\Theta, \sigma(\Theta))$ mit Dichte bzw. Wahrscheinlichkeitsfunktion $\pi(\cdot)$.
Dann heißt für jedes $x \in \mathcal{X}$ (vgl. (3.1)) mit $c(x)$ als geeigneter Normierungskonstante gemäß (3.3)

$$\pi(\vartheta|x) = c(x) \cdot f_\vartheta(x) \cdot \pi(\vartheta) \propto f_\vartheta(x) \cdot \pi(\vartheta) \tag{3.5}$$

die *Posteriori(-Verteilung) (von ϑ) gegeben x bezüglich der Priori $\pi(\cdot)$.*

Stellt man sich eine Zufallsgröße U ("Umwelt", "Nature") vor, die den konkreten Wert ϑ bestimmt und interpretiert die Stichprobenverteilung $p_\vartheta(\cdot)$ als bedingte Verteilung von X gegeben U, z.B. mit Dichte $f_{X|U}(x|\vartheta)$ bzw. Wahrscheinlichkeitsfunktion $p(\{X = x\}|\{U = \vartheta\}), x \in \mathcal{X}$, so entsteht (3.5) durch Anwenden der Proposition 3.1. Die Posteriori $\pi(\vartheta|x)$ ist dann die Dichte (bzw. Wahrscheinlichkeitsfunktion) der bedingten Verteilung von U gegeben X.

Für die Normierungskonstante c(x) gilt dann

$$\frac{1}{c(x)} = f_X(x) = \int f_{X|U}(x|\vartheta)\pi(\vartheta)d\vartheta$$

im Falle von stetigem X und U, und bei diskretem X und U

$$\frac{1}{c(x)} = P(\{X = x\}) = \sum_{j=1}^{m} P(\{X = x\}|\{U = \vartheta_j\}) \cdot \pi(\{U = \vartheta_j\})$$

$$= \sum_{j=1}^{m} p(\{X = x\}|\{U = \vartheta_j\}) \cdot \pi(\vartheta_j)$$

$f_X(x)$ und $p(\{X = x\})$, nicht zu verwechseln mit den als bedingte Verteilungen interpretierten $f_\vartheta(x)$ und $p_\vartheta(\{X = x\})$, heißen (als Funktion von x) *priori-prädiktive Verteilung*.

Analog gibt es auch eine *posteriori-prädiktive Verteilung*, wenn man in analoger Weise über die Posteriori-Verteilung herausintegriert bzw. -summiert.

Dies ist dann die Wahrscheinlichkeitsverteilung der nächsten Beobachtung, basierend auf dem aktuellen Wissensstand.

Bem. 3.5 (Robuste Bayes-Analyse, Generalized Bayes Rule)

In der Situation von Bem. 3.4 kann man auch mit *Priori-Credalmengen* $\mathcal{M}$ arbeiten. Dann heißt

$$\mathcal{M}_{.|x} = \Big\{ \pi(\cdot|x) \;\Big|\; \exists \pi(\cdot) \in \mathcal{M} : \pi(\cdot|x) \text{ ist Posteriori-Verteilung (3.6) von } \vartheta \text{ gegeben } x \text{ bezüglich } \pi(\cdot) \Big\}$$

Posteriori-Credalmenge gegeben x bezüglich $\mathcal{M}$. Man spricht dann von *einer robusten Bayes-Analyse;* (3.7) wird dann oft als *Generalized Bayes Rule* (GBR) bezeichnet.

Bem. 3.6 (Bayes Postulat (nicht entscheidungstheoretisch))
Nach der Beobachtung der Stichprobe enthält die (klassische) Posteriori-Verteilung bzw. die Posteriori-Credalmenge die volle Information, d.h. sie beschreibt das Wissen über den unbekannten Parameter vollständig.
Alle statistischen Analysen haben sich ausschließlich auf die Posteriori zu stützen; darauf aufbauend insbesondere Konstruktion von

- Bayesschen-Punktschätzungen: *MPD-Schätzer (Maximum Posteriori Density-Schätzer)*

- Bayessche-Intervallschätzung: *HPD-Intervalle (Highest posterior density-Intervalle)*

- Bayes-Tests

Proposition 3.7 (Suffizienz und Posteriori-Verteilung)

Ist in der Situation von Bemerkung 3.4 T eine für ϑ suffiziente Statistik mit Wahrscheinlichkeitsfunktion bzw. Dichte $g_\vartheta(\cdot)$, so hängt die Posteriori $\pi(\vartheta|x)$ nur mehr über $t = T(x)$ von x ab. Es gilt

$$\pi(\vartheta|x) \propto g_\vartheta(t) \cdot \pi(\vartheta)$$

Beweis:
Gemäß (3.5) ist

$$\pi(\vartheta|x) \propto f_\vartheta(x) \cdot \pi(\vartheta)$$

wobei wegen der Suffizienz von T sich $f_\vartheta(x)$ schreiben lässt als $f_\vartheta(x) = h_{X|T}(x) \cdot g_\vartheta(t)$. Einsetzen liefert die Behauptung.

Def. 3.8 (Erinnerung: Exponentialfamilien)

Sei $(\mathcal{X}, \sigma(\mathcal{X}), (p_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\Theta \subseteq \mathbb{R}^q$.

- $(p_\vartheta)_{\vartheta\in\Theta}$ bildet eine (oder ist ein Mitglied der) *q-parametrige(n) Exponentialfamilie* in $(T_1, ..., T_q)$ mit *natürlichem Parameter* $(c_1(\vartheta), ..., c_q(\vartheta))$, wenn sich die Dichte $f_\vartheta(\cdot) = f_{X||\vartheta}(\cdot)$ bezüglich eines dominierenden σ-finiten Maßes (also insbesondere Dichte/Wahrscheinlichkeitsfunktion) in die folgende Form bringen läßt: Mit $t_1 := T_1(\vec{x}), ..., t_q := T_q(\vec{x})$ ist

$$f_\vartheta(\vec{x}) = h(\vec{x}) \cdot g(\vartheta) \cdot \exp\Big(\sum_{\ell=1}^{q} c_\ell(\vartheta) t_\ell\Big).$$

- Enthält Θ echt innere Punkte und sind $1, c_1(\vartheta), c_2(\vartheta), ..., c_q(\vartheta)$ und $1, T_1(x), T_2(x), ..., T_q(x)$ (f.-s.) jeweils linear unabhängig, so spricht man von einer *strikt* q-parametrigen Exponentialfamilie. (Der „natürliche Parameterraum“ hat wirklich die Dimension q.)

3.1.2 Konjugierte Verteilungen, Bayes-Lernen

a) Ein Motivationsbeispiel

Bsp. 3.9 (Beta-Binomialmodell)

Absolutes Standardbeispiel

- Stichprobenmodell: Bernoulliverteilung (allgemein: Binomialverteilung) zu Parameter ϑ

$$P_\vartheta\left(\{X_i = x_i\}\right) = \vartheta^{x_i}(1-\vartheta)^{1-x_i}$$

jetzt im Bayes Kontext als bedingte Verteilung schreiben (wieder mit „Hilfsvariable“ U:

$$P\left(\{X_i = x_i\} \mid \{U = \vartheta\}\right) = \vartheta^{x_i}(1-\vartheta)^{1-x_i}$$

- gebräuchliche Priori-Verteilung:
  Betaverteilung, gilt als sehr flexibel, zwei Parameter $a > 0$, $b > 0$ hier als Priori verwendet, Bezeichnung $\pi(\cdot)$

$$\pi(\vartheta) = \frac{\vartheta^{a-1}(1-\vartheta)^{b-1}}{B(a,b)} \cdot I_{[0;1]}(\vartheta)$$

  $B(a,b)$ ist eine reine Normierungskonstante.

Es gilt:

Erwartungswert: $\frac{a}{a+b}$ Modus: $\frac{a-1}{a+b-2}$, $a > 1, b > 1$

Abbildung 1: Rüger, (1999) Test- und Schätztheorie I, Seite 193

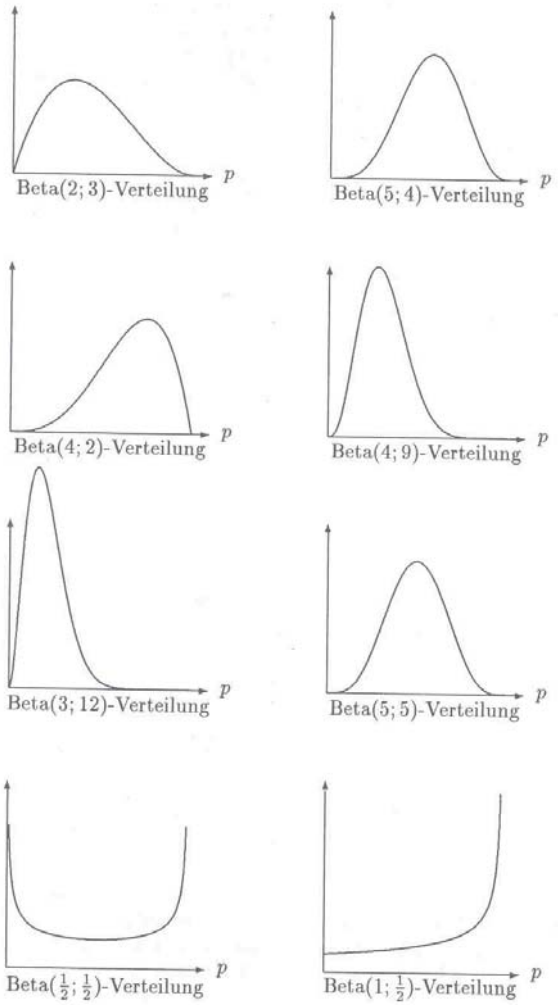


Abbildung 2.17: Einige Beta$(a; b)$-Verteilungen.
Die Beta$(1; 1)$-Verteilung ist die Gleichverteilung. Die an der Senkrechten im Punkt 0.5 gespiegelte Dichte einer Beta$(a; b)$-Verteilung ist die Beta$(b; a)$-Verteilung.

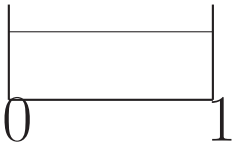
Jetzt Satz von Bayes anwenden: Posteriori nach einer Beobachtung berechnen.

$$\begin{aligned}
\pi(\vartheta|x_i) &= \frac{\vartheta^{x_i}(1-\vartheta)^{1-x_i} \cdot \vartheta^{a-1}(1-\vartheta)^{b-1}}{\underbrace{\text{Norm.} \cdot B(a,b)}_{\text{Normierung}}} \cdot I_{[0;1]}(\vartheta) \\
&\propto \vartheta^{x_i+a-1} \cdot (1-\vartheta)^{b-x_i} \cdot I_{[0;1]}(\vartheta) \\
&= \vartheta^{a'-1} \cdot (1-\vartheta)^{b'-1} \cdot I_{[0;1]}(\vartheta)
\end{aligned}$$

Posteriori ist also wieder eine Betaverteilung, nun mit den Parametern

$$a' = a + x_i \quad \text{und} \quad b' = b - x_i + 1 = b + (1 - x_i).$$

Start z.B. mit $a^{(0)} = 1$, $b^{(0)} = 1$:



Gleichverteilung (als Nichtwissen verkaufbar?)

$\boxed{x_1 = 1}$ beobachtet $\Rightarrow$ $a^{(1)} = a^{(0)} + 1 = 2,\ b^{(1)} = b^{(0)} + 0 = 1$

$Beta(2, 1)$-Verteilung

$\pi(\vartheta \mid x_1) \propto \vartheta I_{[0;1]}(\vartheta)$

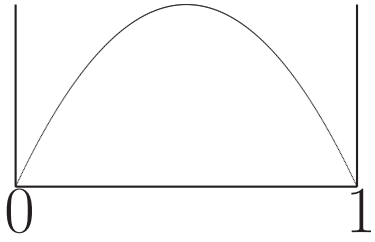
Jetzt weiteres Experiment:

(neue) Priori (alte Posteriori) Beta(2,1) Verteilung
neue Stichprobe x_2

neue Posteriori: $Beta(a^{(1)}alt + x_i, b^{(1)} + (1 - x_i)) - Beta(a^{(2)}, b^{(2)})$

z.B. $\boxed{x_2 = 0} \rightarrow a^{(2)} = 2 + 0 = 2,\ b^{(2)} = 1 + 1 - 0 = 2$

$$\pi_2(\vartheta|(x_1, x_2)') = \frac{\vartheta^{2-1} \cdot (1-\vartheta)^1}{\text{Norm.}} = \frac{\vartheta^1 \cdot (1-\vartheta)^1}{\text{Norm.}} = \frac{\vartheta - \vartheta^2}{\text{Norm.}} \quad \text{für} \quad \vartheta \in [0;1]$$



$$\pi_2(0|x) = \pi_2(1|x) = 0$$

Weitere Beobachtung $\boxed{x_3 = 1}$

neue Posteriori: $'Beta(a^{(2)} + x_i, b^{(2)} + (1 - x_i))'$
$a^{(3)} = 2 + 1,\ b^{(3)} = b_{alt} = 2$

$$\pi_3(\vartheta|(x_1, x_2, x_3)') \propto \vartheta^2(1-\vartheta)^1 I_{[0;1]}(\vartheta) = \vartheta^2 - \vartheta^3 I_{[0;1]}(\vartheta)$$

$\boxed{x_4 = 1}$:

$$\pi_4(\vartheta|(x_1, x_2, x_3, x_4)') \propto \vartheta^3(1-\vartheta)^1$$

Allgemein gilt bei n unabhängigen Wiederholungen:

Die Posteriori $\pi_n(\vartheta|(x_1,\ldots,x_n)')$ ist eine

$$B\left(a^{(0)}+\sum_{i=1}^{n}x_i;b^{(0)}+n-\sum_{i=1}^{n}x_i\right) \text{ Verteilung.} \tag{3.7}$$

Man kann zeigen: Dasselbe Ergebnis erhält man, wenn man $x_1, ..., x_n$ auf einmal verarbeitet.

In diesem Beispiel gilt für die posteriori-prädiktive Verteilung

$$\begin{aligned} & P(X_{n+1} = 1 | X_1 = x_1, \ldots, X_n = x_n) \\ & = \mathbb{E}(\pi_n(\vartheta | x_1, \ldots, x_n)) \\ & = \frac{a + \sum_{i=1}^{n} x_i}{a + b + n}. \end{aligned}$$

Für die Gleichverteilung (vgl. oben) als Ausgangspriori ergibt sich wegen $a^{(0)} = b^{(0)} = 1$

$$P(X_{n+1} = 1 | X_n = x_n, \ldots, X_1 = x_1)$$

$$= \frac{\left(\sum_{i=1}^{n} x_i\right) + 1}{n+2}$$

und damit im Spezialfall $\sum_{i=1}^{n} x_i = \frac{n}{2}$

$$P(X_{n+1} = 1 | X_n = x_n, \ldots, X_1 = x_1)$$

$$= \frac{\frac{n}{2} + 1}{\left(\frac{n}{2} + 1\right) + \left(\frac{n}{2} + 1\right)} = \frac{1}{2},$$

unabhänging vom Stichprobenumfang n; die Anzahl an Beobachtungen beeinflusst hier nicht das Wettverhalten auf das nächste Ereignis. $\rightarrow$ Man setzt also, unabhänging von der Menge an Stichprobenbeobachtungen (ob $0, 10, 100$, oder 10000), stets $\frac{1}{2}$ als Wahrscheinlichkeit an.
Dies mag sinnvoll sein, wenn man sich auf eine einzige Zahl festlegen muss – auf welche sonst, zumal wenn man die Symmetrie zwischen den Beobachtungen 0 und 1 berücksichtigt.
Läßt man Unentschlossenheit/Indifferenz im Weltverhalten zu, so wird man bei kleinem n einen großen Indifferenzbereich haben; bei sehr großen n hingegen verschwindet dieser Bereich (fast völlig).

Eine geeignete Formalisierung ist durch Priori-Credalmengen möglich.

Bem. 3.10 (Zur Kritik des Ansatzes, Priori-Credal-Mengen)

Natürlich hängt (3.7) neben den Stichprobenergebnis auch noch von $a^{(0)}$ und $b^{(0)}$ entscheidend ab.

\+ „Vorwissen kommt mit herein“

- Aber: Wann hat man schon so präzises Vorwissen und braucht dann noch eine Stichprobe?

- Was tut man bei Nichtwissen?
  Die Gleichverteilung ist als Modell für Nichtwissen nicht unproblematisch. (vgl. auch Kritik Laplace-Regel) Man hat dann also „keine Information“ über ϑ, aber eine informative Priori z.B. über eine bijektive Transformation von ϑ.

- Modellierung von partiellem Vorwissen und Nichtwissen durch Credalmengen:

  * Lasse a und/oder b in Bereich variieren und betrachte die (konvexe Hülle aller) entstehenden Verteilungen als Priori-Credalmenge. („robuste Bayes-Analyse“)

  * Idee: Fast „völliges Nichtwissen“, alle möglichen Werte von a und b $\rightarrow$ „*near ignorance prior*“;
    in anderer Parametrisierung besser darstellbar, siehe Bem. 3.17

b) Konjugiertheit: Definition und klassische Ergebnisse

Def. 3.11 (Konjugiertheit)

Eine Verteilungsfamilie Π von Priori-Verteilungen heißt zu einer Menge $\mathcal{P}$ von Stichprobenverteilungen *konjugiert*, wenn für jede Priori $\pi(\cdot) \in \Pi$ und jedes $P(\cdot) \in \mathcal{P}$ die zugehörige Posteriori-Verteilung wieder ein Element von Π ist. Man sagt dann auch, dass jedes Element $\pi(\cdot) \in \Pi$ zu $\mathcal{P}$ konjugiert ist.

Bem. 3.12 (Konjugiertheit und Suffizienz)

Seien Π und $\mathcal{P}$ konjugiert. Beschreibt man die Elemente von $\mathcal{P}$ mit einer suffizienten Statistik T (vgl. Prop. 3.7), so bleibt die Posteriori-Verteilung von ϑ gegeben t in Π.

Proposition 3.13 (Beispiele für Konjugiertheit: Beta-Binomial/Dirichlet-Multinomial-Modell/Gamma-Poisson-Modell, Selbstkonjugiertheit der Normalverteilung)

a) Die Menge der Betaverteilungen als Priori ist zur Menge der Bernoulliverteilungen konjugiert (vgl. Bsp. 3.9).
Allgemeiner gilt:
Ist $\vec{X} = (X_1, \ldots, X_k)$ eine Stichprobe eines zum Parameter $\vec{\vartheta} = (\vartheta_1, \ldots, \vartheta_k)$ multinomial-verteilten Untersuchungsmerkmals, besitzt $\vec{X}$ also die Wahrscheinlichkeitsfunktion

$$f(x|\vec{\vartheta}) \propto \prod_{j=1}^{k} \vartheta_j^{x_j}$$

und wählt man die sog. *Dirichlet-Verteilung* zum Parameter $\vec{\alpha} = (\alpha_1, \ldots, \alpha_k)^T$

$$\pi(\vec{\vartheta}) = \prod_{j=1}^{k} \vartheta_j^{(\alpha_j - 1)},$$

so ist die Posteriori-Verteilung eine Dirichlet-Verteilung mit dem Parameter $\alpha' = (\alpha'_1, \ldots, \alpha'_k)^T$, wobei

$$\alpha'_j = \alpha_j + x_j - 1, \quad j = 1, \ldots, k.$$

b) Ist $\vec{X} = (X_1, \ldots, X_n)$ eine i.i.d. Stichprobe eines zum Parameter λ Poisson verteilten Untersuchungsmerkmals, besitzt $\vec{X}$ also die Wahrscheinlichkeitsfunktion

$$f(x|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! x_2! \ldots x_n!} e^{-n\lambda},$$

und wählt man als Priori-Verteilung eine Gamma-Verteilung mit Parametern a und b, d.h. eine Verteilung mit der Dichte

$$\pi(\lambda) = \frac{b^a}{\underbrace{\Gamma(a)}_{\text{Norm.konst.}}} \lambda^{a-1} e^{-b\lambda}, \tag{3.8}$$

so ist die Posteriori-Verteilung eine Gamma-Verteilung mit den Parametern

$$a + \sum_{i=1}^n x_i \quad \text{und} \quad b + n.$$

Bsp. 3.14 (Normalverteilung)

Ist $\vec{X} = (X_1, \ldots, X_n)$ eine i.i.d. Stichprobe eines mit den Parametern μ und σ^2 normalverteilten Untersuchungsmerkmals, so gilt:

(i) Ist σ^2 bekannt und wählt man als priori Verteilung für μ eine Normalverteilung mit den Parametern ν und ρ^2, so ist die a posteriori Verteilung $\pi(\mu|\vec{x})$ eine Normalverteilung mit den Parametern ν' und $\rho^{2'}$ mit

$$\nu' = \frac{\bar{x}\rho^2 + \nu\dfrac{\sigma^2}{n}}{\rho^2 + \dfrac{\sigma^2}{n}} \tag{3.9}$$

und

$$\rho^{2'} = \frac{\rho^2 \cdot \dfrac{\sigma^2}{n}}{\rho^2 + \dfrac{\sigma^2}{n}} \, . \tag{3.10}$$

(ii) Ist μ bekannt, aber σ^2 unbekannt, so erhält man die konjugierte Verteilung, indem man $\frac{1}{\sigma^2}$ als gammaverteilt annimmt. Man sagt dann, σ^2 sei *invers gammaverteilt*.

Wie findet man solche konjugierten Paare?

Satz 3.15 (Zur Konjugiertheit in Exponentialfamilien)

Hat in der Situation von Def. 3.4 jedes Element der Menge $\mathcal{P}$ der Stichprobenverteilungen eine Dichte bzw. Wahrscheinlichkeitsfunktion $f(x|\vartheta)$ der Form

$$f(x|\vartheta) \propto h(\vartheta) \exp\big(T(x) \cdot b(\vartheta)\big) \tag{3.11}$$

und jedes Element der Menge Π, aus der die Priori-Verteilung stammt, eine Dichte bzw. Wahrscheinlichkeitsfunktion der Form

$$\pi(\vartheta) \propto [h(\vartheta)]^{\alpha} \exp\big(b(\vartheta) \cdot \beta\big), \tag{3.12}$$

so sind Π und $\mathcal{P}$ konjugiert. Es gilt dann

$$\pi(\vartheta|x) \propto [h(\vartheta)]^{\alpha+1} \cdot \exp\big((T(x) + \beta) \cdot b(\vartheta)\big). \tag{3.13}$$

Beweis:

(3.13) ergibt sich unmittelbar durch Anwenden der Formel für die Posteriori-Verteilung auf (3.11) und (3.12). Dann ist (3.13) mit $\alpha' := \alpha + 1$ und $\beta' := \beta + T(x)$ von der Form (3.12), also sind tatsächlich Π und $\mathcal{P}$ konjugiert.

Bem. 3.16 (zu Satz 3.15)

- Der Satz kann also direkt zur Konstruktion geeigneter, konjugierter Priori-Verteilungen verwendet werden, indem man die Stichprobenverteilung in die Form (3.11) bringt und dann eine Priori gemäß (3.12) wählt.

- $b(\vartheta)$ spielte in (3.11) und in (3.12) eine ganz unterschiedliche Rolle:
  In (3.11) ist $b(\vartheta)$ der natürliche Parameter der Exponentialfamilie, aus der die Likelihood / Stichprobenverteilung stammt.
  In (3.12) hingegen ist $b(\vartheta)$ die suffiziente Statistik für den natürlichen Parameter β der Exponentialfamilie, aus der die Priori stammt. (Bei der Priori ist ja der Wert von ϑ „zufällig“.)

- Ähnliches gilt für $h(\vartheta)$.

c) Konjugiertheit und verallgemeinerte Bayes-Inferenz

Bem. 3.17 (Eine alternative Darstellung von Satz 3.15)

Eine zum Nachweis meist umständlichere, aber für die Interpretation oft anschaulichere und für die Verallgemeinerung besser geeignete, alternative Darstellung von Satz 3.15 lautetet:

Hat in der Situation von Def. 3.4 jedes Element der Menge $\mathcal{P}$ der Stichprobenverteilungen eine Dichte bzw. Wahrscheinlichkeitsfunktion $f(x|\vartheta)$ der Form

$$f(x|\vartheta) \propto \exp\Big(\psi(\vartheta)\tau(\vec{x}) - n \cdot d(\vartheta)\Big) \tag{3.14}$$

und jedes Element der Menge Π, aus der die Priori-Verteilung stammt, eine Dichte bzw. Wahrscheinlichkeitsfunktion der Form

$$\pi(\vartheta) \propto \exp\left(n^{(0)}\left(\psi(\vartheta)y^{(0)} - d(\vartheta)\right)\right) \tag{3.15}$$

so sind Π und $\mathcal{P}$ konjugiert. Es gilt dann

$$\pi(x|\vartheta) \propto \exp\left(n^{(1)}\left(\psi(\vartheta)y^{(1)} - d(\vartheta)\right)\right) \tag{3.16}$$

mit

$$n^{(1)} = n^{(0)} + n \tag{3.17}$$

und

$$y^{(1)} = \frac{n^{(0)}y^{(0)} + \tau(x)}{n^{(0)} + n}. \tag{3.18}$$

$y^{(0)}$ ist typischerweise ein Lageparameter, $n^{(0)}$ kann man als virtuelle Beobachtungen interpretieren, auf denen das Priori-Wissen beruht.

Mit $\bar{\tau}(\vec{x}) = \frac{1}{n}\tau(\vec{x})$ läßt sich die Aufdatierung des Parameters schreiben als

$$y^{(1)} = \frac{n^{(0)}}{n^{(0)} + n} y^{(0)} + \frac{n}{n^{(0)} + n} \bar{\tau}(\vec{x}) \tag{3.19}$$

also als gewichtetes Mittel der Priori-Vermutung und des Stichprobenmittels.

Beispielsweise ist dann die Beta-priori aus Abschnitt 3.4.2

$$\begin{aligned}\pi(\vartheta) &\propto \vartheta^{a-1}(1-\vartheta)^{b-1} I_{[0;1]}(\vartheta) \\ &= \vartheta^{n^{(0)}y^{(0)}-1}(1-\vartheta)^{(1-y^{(0)})-1} I_{[0;1]}(\vartheta)\end{aligned}$$

mit (festem) $n^{(0)} > 0$ und (festem) $y^{(0)} \in (0;1)$. $y^{(0)}$ ist dann genau der Erwartungswert der Priori; ferner gilt für die priori-prädiktive Verteilung der nächsten austauschbaren Beobachtung X_{neu}:

$$P(X_{neu} = 1) = y^{(0)}$$

.

Bem. 3.18 (Robuste Bayes-Analyse in konjugierten Modellen)

- Die Darstellung in Bem. 3.17 ermöglicht eine elegante robuste Bayes-Analyse. Priori-Credalmengen erzeugt man durch intervallwertige Priori-Parameter:[15]

  $\left[\underline{y}^{(0)}, \bar{y}^{(0)}\right]$ typischerweise intervallwertiger Priori-Mittelwert bzw. bzw. Lageparameter

  und/oder

  $\left[\underline{n}^{(0)}, \bar{n}^{(0)}\right]$ intervallwertige virtuelle Beobachtungen, auf denen das Priori-Wissen beruht

[15] Walley (1991/1996): Binomial-/Multinomialmodell. Quaeghebeur & deCooman(2005): Exponentialfamilien mit festem $n^{(0)}$; $n^{(0)}$ zusätzlich variabel. Walter & Augustin (2009)

Lässt man $\underline{y}^{(0)}$ und $\bar{y}^{(0)}$ gegen die Grenzen des zulässigen Priori-Parameterbereichs gehen, so erhält man sogenannte *„near ignorance“*-Modelle.

Beispielsweise gilt dann für die priori-prädiktive Verteilung

$$P(X = 1) = [\lim_{y^{(0)} \downarrow 0} y^{(0)}, \lim_{y^{(0)} \uparrow 1} y^{(0)}] = [0; 1]$$

und analog für $P(X = 0)$, was Nichtwissen über ϑ deutlich ausdrückt.

- Für die posteriori-prädiktive Verteilung ergibt sich mit festem „Priori-Gewichts-Parameter“ $n^{(0)}$ nach n Beobachtungen $x_1, \ldots x_n$:

$$
\begin{aligned}
& P(X_{neu} = 1 | X_1 = x_1, \ldots, X_n = x_n) \\
&= \left[\lim_{y^{(0)} \downarrow 0} \frac{n^{(0)} y^{(0)} + \sum_{i=1}^{n} x_i}{n^{(0)} + n}; \lim_{y^{(0)} \uparrow 1} \frac{n^{(0)} y^{(0)} + \sum_{i=1}^{n} x_i}{n^{(0)} + 1} \right] \\
&= \left[\frac{\sum_{i=1}^{n} x_i}{n^{(0)} + n}; \frac{n^{(0)} + \sum_{i=1}^{n} x_i}{n^{(0)} + n} \right] .
\end{aligned}
$$

Die Breite

$$\frac{n^{(0)}}{n^{(0)}+n}$$

des Intervalls, nimmt in n monoton ab: Für kleines n sind aus Nichtwissen nur schwache Folgerungen ziehbar; für größeres n wird man präziser.
Beachte, dies ist kein Konfidenzintervall, sondern eine „intervallwertige Punktschätzung“! Die entsprechende Verallgemeinerung auf mehrkategoriale Beobachtungen ist das sog. *Imprecise-Dirichlet-Model* (IDM, Walley (1996, J. Royal Statistical Society B)). Es gilt als Grundlage vieler weiterführender Anwendungen. Man kann zeigen: Die Priori- und Posteriori-Verteilungen sind unabhänging von der Kategorisierung; Zusammenfassen/Präzisieren der Kategorien ändert die Inferenz nicht (vgl. im Gegensatz dazu die Auseinandersetzung mit der Laplace-Regel in Kap. 2.3)

3.1.3 (Reine) Bayes-Punktschätzung

Def. 3.19 (MPD-Schätzung)

Gegeben eine Beobachtung $\vec{x}$ und die Posteriori-Verteilung mit Dichte bzw. Wahrscheinlichkeitsfunktion $\pi(\vartheta|\vec{x})$ heißt $\widehat{\vartheta}$ mit

$$\pi(\hat{\vartheta}|\vec{x}) = \max_{\vartheta \in \Theta} \pi(\vartheta|\vec{x})$$

(reiner) Bayes-Schätzwert oder *Maximum (bzw Highest) Posteriori Density Schätzwert* (MPD- (bzw. HPD-) Schätzwert) oder *Posteriori-Modus-Schätzwert*. Die zugehörige Schätzfunktion $\hat{\vartheta}(\vec{X})$ heißt reine *Bayes-Schätzung* oder *MPD- (bzw. HPD-) Schätzung* bzw. *Posteriori-Modus-Schätzung*.

Bem. 3.20 (Zur MPD-Schätzung)

a) Ist die Posteriori-Verteilung unimodal, so ist $\widehat{\vartheta}$ der Modus der Posteriori.

b) Ist der Zustandsraum Θbeschränkt und liegt dem Schätzproblem als Priori-Verteilung eine Gleichverteilung zugrunde, so gilt

$$\pi(\vartheta|\vec{x}) \propto f(\vec{x}|\vartheta) \cdot \pi(\vartheta) = f(\vec{x}|\vartheta) \cdot \text{ Konstante}$$

D.h. der MPD-Schätzer ist dasjenige ϑ, das $f(x|\vartheta)$ maximiert, also der Maximum-Likelihood-Schätzwert.

c) Im Falle $\Theta = \mathbb{R}^+$ oder $\Theta = \mathbb{R}$ gibt es keine Gleichverteilung auf Θ, denn mit

$$f(x) = c \quad \text{ist} \qquad \int_0^\infty f(x)dx = \int_0^\infty c\, dx = \big[x\big]_0^\infty = \infty$$

unabhängig von $c > 0$.

Man kann aber zeigen, dass viele der zentralen Ergebnisse der Bayes-Theorie erhalten bleiben, wenn man auch nicht normierbare σ-finite Maße als Prioris zu lässt (z.B. Lebesque Maß $\lambda(\cdot)$; $\lambda([a,b]) := b - a$: „*improper priors*“)

Bsp. 3.21 (Beta-Binomialmodell)

$\pi(\vartheta|x_1,...,x_n)$ ist $B(a+\sum_{i=1}^n x_i; b+n-\sum_{i=1}^n x_i) =: B(a',b')$ -verteilt.
Hat man bei der Priori a=1=b gewählt, so ergibt sich mit $\frac{a'-1}{a'+b'-2}$ als Modus der Beta(a',b')-Verteilung der MPD-Schätzwert

$$\hat{\vartheta} = \frac{1+\sum_{i=1}^n x_i - 1}{1+\sum_{i=1}^n x_i + 1 + n - \sum_{i=1}^n x_i - 2} = \frac{1}{n}\sum_{i=1}^n x_i,$$

also in der Tat der ML-Schätzwert.

Bem. 3.22 (Fortsetzung von Beispiel 3.14)

Man betrachte die Bayes-Inferenz für den Mittelwert einer Normalverteilung aus einer i.i.d. Stichprobe. Geht man im Sinne von Bem. 3.18 zu Priori-Credalmengen über, wobei ν in einem Intervall $[\underline{\nu}^{(0)}, \bar{\nu}^{(0)}]$ und $n^{(0)}$ in $[\underline{n}^{(0)}, \bar{n}^{(0)}]$ variiert, so gilt[16]

[16] Walter&Augustin(2009, Remark 4.1)

$$\underline{\nu}^{(1)} = \begin{cases} \dfrac{\bar{n}^{(0)}\underline{\nu}^{(0)} + \sum_{i=1}^{n} x_i}{\bar{n}^{(0)} + n}, & \text{falls } \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i \geq \underline{\nu}^{(0)} \\ \dfrac{\underline{n}^{(0)}\bar{\nu}^{(0)} + \sum_{i=1}^{n} x_i}{\underline{n}^{(0)} + n}, & \text{falls } \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i < \underline{\nu}^{(0)} \end{cases}$$

$$\underline{\nu}^{(1)} = \begin{cases} \dfrac{\bar{n}^{(0)}\bar{\nu}^{(0)} + \sum_{i=1}^{n} x_i}{\bar{n}^{(0)} + n}, & \text{falls } \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i \leq \bar{\nu}^{(0)} \\ \dfrac{\underline{n}^{(0)}\bar{\nu}^{(0)} + \sum_{i=1}^{n} x_i}{\underline{n}^{(0)} + n}, & \text{falls } \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i > \bar{\nu}^{(0)} \end{cases}$$

Daraus ergibt sich insbesondere, dass

$$\bar{\nu}^{(1)} - \underline{\nu}^{(1)} = \frac{\bar{n}^{(0)}\big(\bar{\nu}^{(0)} - \underline{\nu}^{(0)}\big)}{\bar{n}^{(0)} + n}$$

$$+\mathrm{pdc}\left(\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i}; \underline{\nu}^{(0)}, \bar{\nu}^{(0)}\right) \frac{\mathrm{n}\big(\bar{\mathrm{n}}^{(0)} - \underline{\mathrm{n}}^{(0)}\big)}{\big(\bar{\mathrm{n}}^{(0)} + \mathrm{n}\big)\big(\underline{\mathrm{n}}^{(0)} + \mathrm{n}\big)}$$

Dabei ist

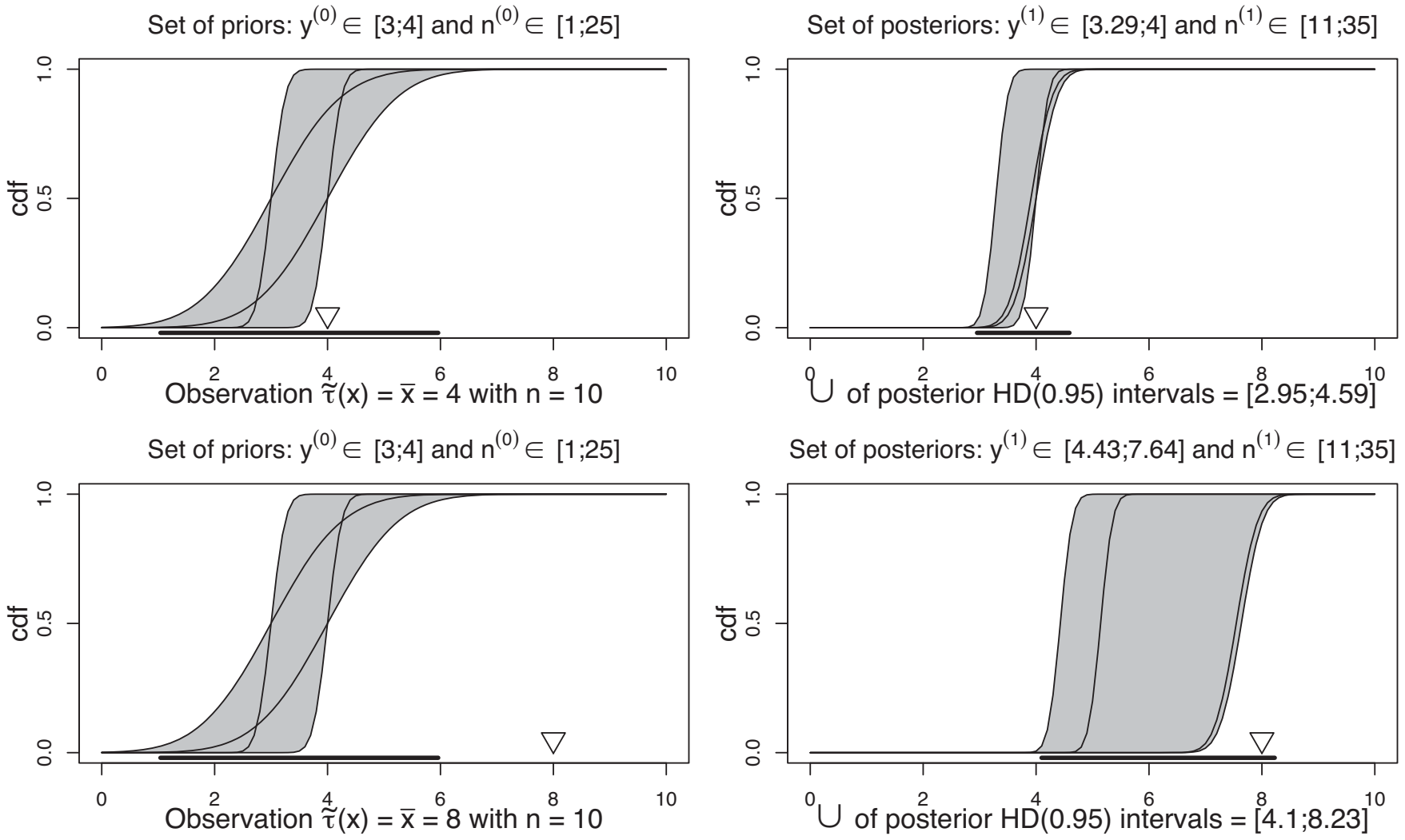
$$\mathrm{pdc}\left(\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i}; \underline{\nu}^{(0)}, \bar{\nu}^{(0)}\right) := \inf\left\{\left|\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i} - \nu^{(0)}\right| \;\middle|\; \underline{\nu}^{(0)} \leq \nu^{(0)} \leq \bar{\nu}^{(0)}\right\}.$$

pdc misst das Ausmaß des Priori-Daten-Konflikts, also wie weit die Stichprobenbeobachtung $\frac{1}{n}\sum_{i=1}^{n} x_i$ von dem Priori-Mittelwert $[\underline{\nu}^{(0)}, \bar{\nu}^{(0)}]$ entfernt ist. (0, wenn innerhalb des Intervalls, sonst Differenz zur nächsten Grenze.)
Damit gilt also:

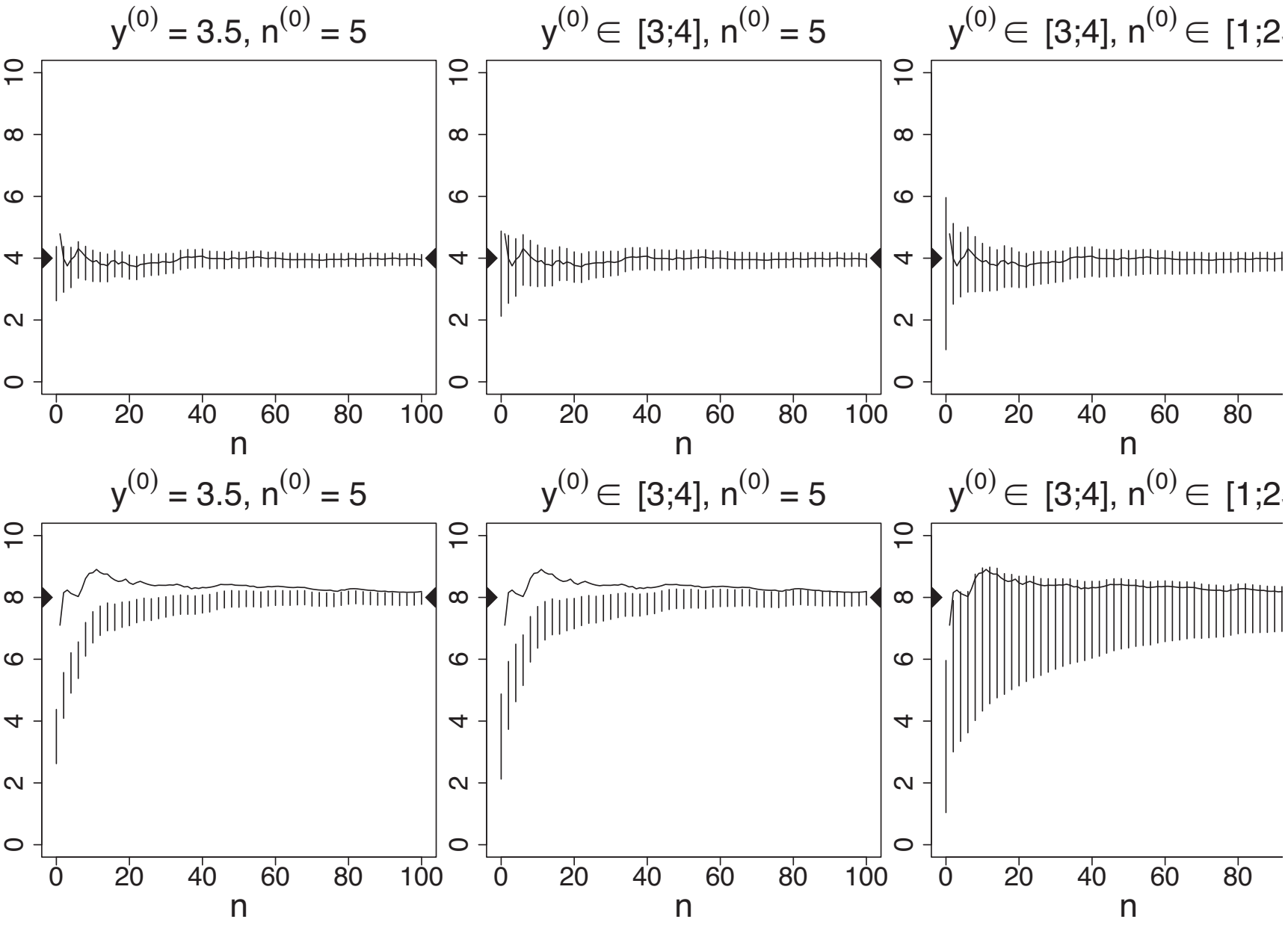
- Bei „nicht überraschenden Beobachtungen“ist die Posteriori-Unschärfe klein.

- Bei „überraschenden Beobachtungen“ hingegen gilt: Die Posteriori-Unschärfe ist groß; bei abgeleiteten Folgerungen ist man sehr vorsichtig.

Quelle: Walter & Augustin (2009, S. 268)



Quelle: Walter & Augustin (2009, S. 268)



3.1.4 Der Hauptsatz der Bayes-Entscheidungstheorie

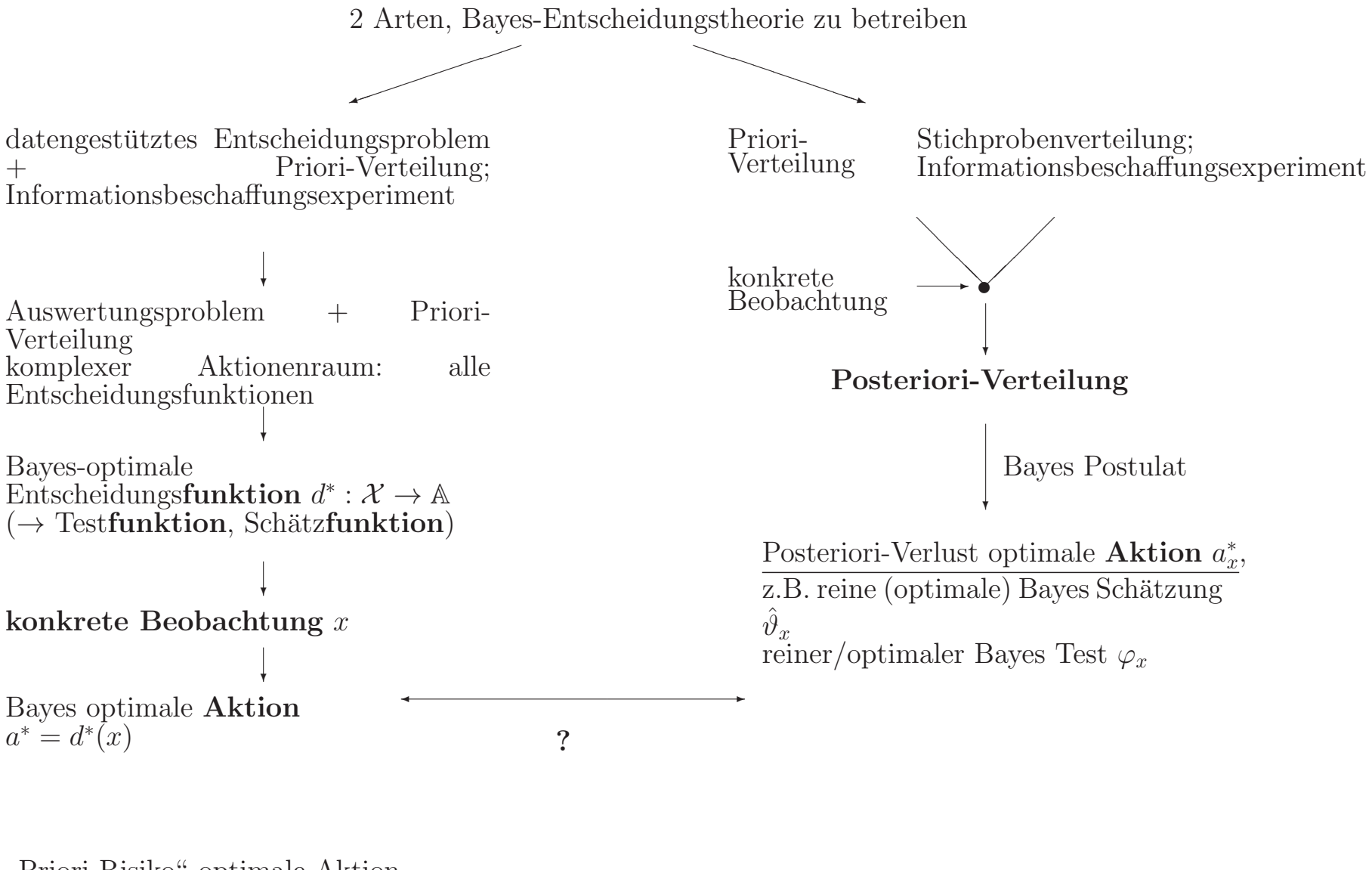
Def. 3.23 (Bayes-Analyse mit Verlustfunktion)

Gegeben sei ein datenbasiertes Entscheidungsproblem $((\mathbb{A}, \Theta, l(\cdot)); (\mathcal{X}, A, (P_\vartheta)_{\vartheta\in\Theta}))$ und eine Priori-Verteilung $\pi(\cdot)$ über $(\Theta, \sigma(\Theta))$.
Eine Aktion $a_x^* \in \mathbb{A}$ heißt *Posteriori-Verlust optimal* zur *Beobachtung* $x \in \mathcal{X}$, wenn gilt

$$\mathbb{E}_{\pi(\cdot|x)} l(a_x^*, \vartheta) \leq \mathbb{E}_{\pi(\cdot|x)} l(a, \vartheta) \quad \forall a \in \mathbb{A}.$$

a_x^* ist also sozusagen Bayes-Aktion zur „aufdatierten Priori-Verteilung“ $\pi(\cdot|x)$.

Analog definiert man eine *Posteriori-Nutzen-Optimalität.*


2 Arten, Bayes-Entscheidungstheorie zu betreiben
datengestütztes Entscheidungsproblem + Priori-Verteilung; Informationsbeschaffungsexperiment
Priori-Verteilung
Stichprobenverteilung; Informationsbeschaffungsexperiment
konkrete Beobachtung
Auswertungsproblem + Priori-Verteilung
komplexer Aktionenraum: alle Entscheidungsfunktionen
Posteriori-Verteilung
Bayes Postulat
Bayes-optimale Entscheidungsfunktion $d^* : \mathcal{X} \to \mathbb{A}$
($\to$ Testfunktion, Schätzfunktion)
Posteriori-Verlust optimale Aktion a_x^*,
z.B. reine (optimale) Bayes Schätzung $\hat{\vartheta}_x$
reiner/optimaler Bayes Test φ_x
konkrete Beobachtung x
Bayes optimale Aktion $a^* = d^*(x)$
?
„Priori-Risiko“ optimale Aktion

Satz 3.24 (Hauptsatz der Bayes-Entscheidungstheorie)

Gegeben sei ein datengestütztes Entscheidungsproblem $((\mathbb{A}, \Theta, \ell(\cdot)); (\mathcal{X}, \mathcal{A}, (P_\vartheta)_{\vartheta\in\Theta}))$, bestehend aus einem *datenfreien Entscheidungsproblem* $(\mathbb{A}, \Theta, \ell(\cdot))$ und einer Informationsstruktur $(\mathcal{X}, \mathcal{A}, (P_\vartheta)_{\vartheta\in\Theta})$ sowie eine a Priori-Verteilung $\pi(\cdot)$ über $(\Theta, \sigma(\Theta))$.

Eine Entscheidungs**funktion**

$$\begin{aligned} d^* : \mathcal{X} &\longrightarrow \mathbb{A} \\ x &\longmapsto d^*(x) \end{aligned}$$

ist genau dann Bayes-optimal im zugehörigen Auswertungsproblem, wenn für jedes $x \in \mathcal{X}$ die zugehörige Aktion $d^*(x)$ Posteriori-Verlust optimal zur Beobachtung x ist.

Beweis: Für den diskreten Fall [17]

- Vorneweg eine Hilfsüberlegung: Suche die Lage des Minimums $\vec{z}_{min}$ einer Funktion $f(\vec{z})$ mit $\vec{z} = (z_1, ..., z_n)$, wobei $f(\vec{z}) = \sum_{i=1}^{n} c_i f_i(z_i)$, also die i-te Komponente von z nur im i-ten Summanden auftritt.

$$f(\vec{z}) = \sum_{i=1}^{n} c_i f_i(z_i) \to \min_{\vec{z}}$$

$$\iff f_i(z_i) \longrightarrow \min_{z_i} \quad \text{für jedes i unabhängig von den anderen Summanden.}$$

- Der Deutlichkeit halber wird wieder eine Hilfsvariable U eingeführt und $P_\vartheta(\{X = x\})$ wird als $P(\{X = x\}|\{U = \vartheta\})$ geschrieben.

[17] für den allgemeinen Fall: siehe z.B. Rüger (1999, S. 283f.)

Angewendet auf Entscheidungsprobleme mit der Posteriori $\pi(\vartheta|x)$ ergibt sich mit dieser Notation

$$\pi(\vartheta|x) = \frac{P(\{X = x\}|\{U = \vartheta\}) \cdot \pi(\vartheta)}{P(\{X = x\})}. \tag{3.20}$$

Betrachte Entscheidungsfunktion $d(\cdot)$ im Auswertungsproblem:

$$\begin{aligned} R(d, \vartheta) &= \mathbb{E}_{P_\vartheta}(\ell(d(x), \vartheta) \\ \text{also hier } R(d, \vartheta) &= \sum_{x \in \mathcal{X}} \ell(d(x), \vartheta) \cdot P_\vartheta(\{X = x\}). \end{aligned}$$

Die optimale Entscheidungsfunktion zur Priori $\pi(\cdot)$ minimiert unter allen d

$$\mathbb{E}_\pi(R(d,\vartheta)),$$

löst also

$$\sum_{\vartheta\in\Theta}\left(\sum_{x\in\mathcal{X}}\ell(d(x),\vartheta)\cdot P_\vartheta(\{X=x\})\right)\cdot\pi(\vartheta)\to\min_d$$

$$\Longleftrightarrow \quad \sum_{\vartheta\in\Theta}\sum_{x\in\mathcal{X}}\Big(\ell(d(x),\vartheta)\cdot \underbrace{P(\{X=x\}|U=\vartheta)\cdot\pi(\vartheta)}_{=\pi(\vartheta|x)\cdot P(\{X=x\})}\Big)\to\min_d$$

$$\overset{\text{wegen (3.20)}}{\Longleftrightarrow} \underbrace{\sum_{x\in\mathcal{X}}}_{\hat{=}\sum_{i=1}^n}\Big(\underbrace{\sum_{\vartheta\in\Theta}\ell(d(x),\vartheta)\cdot\pi(\vartheta|x)}_{\hat{=}f_i(z_i)}\Big)\cdot \underbrace{P(\{X=x\})}_{\hat{=}c_i;\text{ priori-prädiktiv, marginal}} \to\min_d$$

- Wegen der Hilfsüberlegung ist dies äquivalent dazu, für jedes feste x

$$\sum_{\vartheta\in\Theta} l(d(x),\vartheta)\cdot\pi(\vartheta|x)$$

separat zu minimieren nach $a := d(x)$ für festes x.
Dies liefert jeweils genau die Posteriori-Verlust optimale Aktion, also die Bayes-Aktion zur Posteriori als neuer Priori.

Satz 3.25 (Bestimmung von Bayes-optimalen Entscheidungsfunktionen, z.B. Rüger (1999, Satz 2.20))

Gegeben sei das Schätzproblem als datengestütztes Entscheidungsproblem gemäß Kapitel 1.5 sowie eine Priori-Verteilung $\pi(\cdot)$.
Dann gilt:

i) Wählt man die absolute bzw. quadratische Verlustfunktion, so gilt für die Bayes-optimale Entscheidungsfunktion $d^*_{quad}(\cdot)$ bzw. $d^*_{abs}(\cdot)$:

ii) Für jedes x ist $d^*_{quad}(\cdot)$ genau der Erwartungswert und $d^*_{abs}(\cdot)$ der Median der Posteriori-Verteilung $\pi(\vartheta|x)$.

iii) Die HPD-Schätzung (Modus) ergibt sich näherungsweise für kleine ϵ, wenn man die sogenannte Toleranzverlustfunktion zum Grade ϵ verwendet:

$$\ell_\epsilon(\hat{\vartheta}, \vartheta) = \begin{cases} 1 & |\hat{\vartheta} - \vartheta| > \epsilon \\ 0 & |\hat{\vartheta} - \vartheta| \leq \epsilon \end{cases}$$

3.1.5 Zum Einfluss der Priori-Verteilung; „Asymptotische Objektivität“, uninformative Prioris und ihr Informationsgehalt

Satz 3.26 („Asymptotische Objektivität von Bayes-Verfahren“, „Konsistenzsatz“)

Sei $\Theta = \{\vartheta_1, \ldots, \vartheta_m\}$ ein endlicher Parameterraum und $\vec{X} = (X_1, \ldots, X_n)$ eine i.i.d. Stichprobe eines beliebig verteilten (reellwertigen) Untersuchungsmerkmals mit Dichten $f(x_i|\vartheta_{wahr}),\ \vartheta_{wahr} \in \Theta$.
Sei $\pi(\vartheta)$ die Wahrscheinlichkeitsfunktion der Priori-Verteilung auf Θ mit $\pi(\vartheta) > 0$ für alle ϑ. Dann gilt für die Wahrscheinlichkeitsfunktion der nach n Beobachtungen gebildeten Posteriori-Verteilung $\pi_n(\vartheta|x)$

$$\lim_{n\to\infty} \pi_n(\vartheta|x) = \begin{cases} 1 & \text{falls } \vartheta = \vartheta_{wahr} \\ 0 & \text{falls } \vartheta \neq \vartheta_{wahr} \end{cases}$$

Korollar 3.27 (Konsistenzsatz für Credal-Bayes-Verfahren)

In der Situation von Satz 3.26 gilt:
Ist $\mathcal{M}$ eine Priori-Credalmenge mit $\pi(\cdot) > 0, \forall \pi \in \mathcal{M}$, so zieht sich die nach n Beobachtungen gebildete Posteriori-Credalmenge $\mathcal{M}_{|x}^{(n)}$ im Punkt ϑ_{wahr} zusammen:

$$\lim_{n\to\infty} \left(\inf_{\pi(\cdot|x)\in\mathcal{M}_{|x}^{(n)}} \pi(\vartheta|x) \right)$$

$$= \lim_{n\to\infty} \left(\sup_{\pi(\cdot|x)\in\mathcal{M}_{|x}^{(n)}} \pi(\vartheta|x) \right) = \begin{cases} 1 & \text{falls } \vartheta = \vartheta_{wahr} \\ 0 & \text{falls } \vartheta. \end{cases}$$

Bem. 3.28 („Nichtinformative“ Priori-Verteilung)

- Es gibt verschiedene Versuche, ähnlich der Laplace-Regel, „nichtinformative“ Priori-Verteilungen zu definieren und diese dann als Standardbewertungen heranzuziehen.
  - z.B. die Gleichverteilung, diese ist aber nicht invariant gegenüber Transformationen des Parameters. Man hat dann also „keine Information“ über ϑ, aber eine informative Priori z.B. über eine bijektive Transformation von ϑ.
  - z.B. Verteilungen, die invariant bezüglich bijektiver Transformationen des Parameters sind (Jeffrey-Regel).
  - z.B. Verteilungen, die die Entropie maximieren (Jaynes-Regel)

Bem. 3.29 (Erneute kritische Diskussion des Bayes-Ansatzes)