4.3 Grundbegriffe verallgemeinerter Wahrscheinlichkeiten

4.3.1 Credalmengen, Intervallwahrscheinlichkeiten und ihre Strukturen

Ziel: Verallgemeinerte Wahrscheinlichkeitsrechnung:

- konsequente Verallgemeinerung der klassischen traditionellen Wahrscheinlichkeitsrechnung, d.h. des auf die Axiomatik von Kolmogorov aufbauenden Kalküls, das als Spezialfall mitenthalten sein soll und als Richtschnur dient. Jede Mengenfunktion $p(\cdot)$, die Axiome von Kolmogorov erfüllt, wird im folgenden zur Unterscheidung als *klassische Wahrscheinlichkeit* bezeichnet.

- angemessene Berücksichtigung das Ausmaßes an Ambiguität

- Zuverlässigkeit statt (Über)präzision.

Manski's *Law of Decreasing Credibility*: „The credibility of inferences decreases with the strength of the assumptions maintained." (Manski, 2003, p. 1)

Bem. 4.6 Grundlegende Ansätze, Credalmenge, Intervallwahrscheinlichkeit und Struktur

Ähnlich wie bei der Modellierung des Ellsberg-Experiments gibt es auch allgemein zwei (verwandte (s.u.)) Ansätze zu Verallgemeinerten Wahrscheinlichkeiten über einem Messraum $(\Omega, \mathcal{A})$:

a) *Credalmengen*

- *Mengen* $\mathcal{M}$ klassischer Wahrscheinlichkeiten als Grundentität: Die Menge in ihrer Gesamtheit beschreibt die Wahrscheinlichkeit; kein Element ist wahrscheinlicher als ein anderes. (Legen einer Verteilung über diese Menge würde nach Mischungsprozess auf eine klassische Wahrscheinlichkeit führen, also das Grundproblem nicht lösen.)

- $\mathcal{M}$ heißt dann *Credal-Menge* (Isaac Levi: Credal Set)

- „Größe“ der Menge reflektiert Ausmaß der Ambiguität:

- perfekte probabilistische Information, ideale Stochastizität, Risikosituation
  $\rightarrow$ Menge, die aus einem Punkt besteht; klassische Wahrscheinlichkeit als Spezialfall

- Unsicherheit i.e.S., volle Ambiguität
  $\rightarrow$ Menge *aller* Wahrscheinlichkeitsverteilungen auf dem gegebenem Messraum

- Sehr allgemeine Modellbildung möglich: Man trifft eine Aussage nur über die Wahrscheinlichkeitskomponenten derjenigen Ereignisse, über die man etwas weiß; die restlichen Werte ergeben sich implizit.

b) *Intervallwahrscheinlichkeit*

- Bei einem Meßraum $(\Omega, \mathcal{A})$ betrachtet man statt der Abbildung

$$\begin{aligned} p : \mathcal{A} &\to \mathbb{R} \\ A &\mapsto p(A) \end{aligned}$$

die Abbildung

$$\begin{aligned} P : \mathcal{A} &\to \mathcal{Z}_0([0;1]) \\ A &\mapsto P(A) = [\underbrace{L}_{lower}(A), \underbrace{U}_{upper}(A)], \end{aligned}$$

wobei $\mathcal{Z}_0([0;1])$ die Menge aller abgeschlossenen Intervalle in [0;1] ist.[10]

[10] Es sei nochmals an die Konvention erinnert: klassische Wahrscheinlichkeit wird mit Kleinbuchstaben symbolisiert, Intervallwahrscheinlichkeit mit Großbuchstaben

c) Formal sind beide Ansätze eng verwandt, denn es gilt:

- Aus jeder nichtleeren Menge $\mathcal{M}$ von klassischen Wahrscheinlichkeiten kann man eine Intervallbewertung erzeugen, indem man festsetzt:

$$L(A) = \inf_{p \in \mathcal{M}} p(A) \qquad U(A) = \sup_{p \in \mathcal{M}} p(A), \quad \forall A \in \mathcal{A} \tag{4.2}$$

- Umgekehrt kann man zu jeder intervallwertigen Bewertung $P(\cdot) = [L(\cdot), U(\cdot)]$ die Menge aller damit kompatiblen klassischen Wahrscheinlichkeiten betrachten:

$$\begin{aligned} \mathcal{M} := \{p(\cdot) \quad & | \; p(\cdot) \text{ ist klassische Wahrscheinlichkeit} \wedge \\ & | L(A) \leq p(A) \leq U(A), \forall A \in \mathcal{A}\}. \end{aligned} \tag{4.3}$$

* und dann heißt $\mathcal{M}$ *Struktur* (oder Core (Kern)) der Intervallbewertung $P(\cdot)$.

* Ist $\mathcal{M} \neq \emptyset$, so heißt $P(\cdot)$ *R-Wahrscheinlichkeit*

* Gilt zudem für alle $A \in \mathcal{A}$:

$$L(A) = \inf_{p \in \mathcal{M}} p(A) \qquad \text{und} \qquad U(A) = \sup_{p \in \mathcal{M}} p(A), \tag{4.4}$$

so spricht man von *F-Wahrscheinlichkeit* oder Intervallwahrscheinlichkeit (im eigentlichen Sinn).

Bem. 4.7 (Einige weitere Bemerkungen zur Intervallwahrscheinlichkeit)

- $L(\cdot)$ und $U(\cdot)$ werden häufig als *Kapazität* (oder *fuzzy-Maß*) bezeichnet.

- Innerhalb des Intervalls herrscht völlige Unsicherheit; kein Wert ist „wahrscheinlicher“.

- perfekte probabilistische Information, ideale Stochastizität: $L(\cdot) = U(\cdot)$, das Intervall besteht aus einem Punkt $\rightarrow$ klassische Wahrscheinlichkeit als Spezialfall.

- Unsicherheit i.e.S., volle Ambiguität führt auf das Intervall [0;1]. Dies ist das korrekte Modell für absolutes Nichtwissen bezüglich Wahrscheinlichkeiten; man weiß nur, dass die Wahrscheinlichkeitskomponente jedes Ereignissen zwischen 0 und 1 liegt.

- Obige Definition der Intervallwahrscheinlichkeit ist konsequent interpretationsunabhängig (Weichselberger (2000, 2001)): direkte Verallgemeinerung der Kolmogorovschen Axiomatik.

- Es gibt auch zwei operationale Definitionen von Intervallwahrscheinlichkeiten, die – ähnlich wie in der klassischen Theorie – umgekehrt auch als Interpretationen einer axiomatischen Vorgehensweise gesehen werden können:

  * subjektivisch: untere und obere Wettquotienten (behavioristisch, Walley, wohl am stärksten verbreitet), Verallgemeinerung des de Finetti-Ansatzes. Wetten mit unterschiedlichem Kauf- und Verkaufspreis.

- frequentistisch: Einhüllen der Häufigkeitspunkte nicht notwendig konvergierender Folgen relativer Häufigkeiten; verallgemeinerte von Mises-Kollektive.

- Unabhängig davon, ob man eine frequentistische oder subjektivistische Interpretation bevorzugt, gibt es zwei verschiedene Auffassungen wie man die Struktur und die Intervalle (bzw. auch eine Credalmenge) versteht:

  * Die sog. *epistemische* Sicht beruht auf der Vorstellung, es gebe eine wahre klassische Wahrscheinlichkeit, die man aber nur (noch?) nicht kennt. Die Credalmenge bzw. die Intervalle zeichnen dann einfach Bereiche aus, in denen die klassische Wahrscheinlichkeit liegt. (Situation im Ellsberg-Experiment; es gibt einen wahren Anteil der gelben Kugeln. Das ist auch die Sicht des Fundamentaltheorems von de Finetti).

  * Demgegenüber versteht der *ontologische* Ansatz Wahrscheinlichkeit grundsätzlich als intervallwertige/mengenwertige Entität.

  Für die Modellierung spielt die Unterscheidung erst bei komplexeren Modellen (z.B. unabhängige Koppelung) eine wichtige Rolle, weswegen hier nicht genauer darauf eingegangen wird.

- Die Idee mit intervallwertigen Wahrscheinlichkeiten zu arbeiten, ist natürlich naheliegend und beileibe nicht neu (spätestens seit Boole vor ca. 150 Jahren)
  Relativ neu ist aber die Existenz eines leistungsfähigen Kalküls basierend auf einer „sauberen“Axiomatik.

- Bemerkung: Die Begriffe R- und F-Wahrscheinlichkeiten stammen von Weichselberger (2000, 2001), der zeigt, dass man damit bereits eine interpretationsunabhängige tragfähige Axiomatik bilden kann. (Zusatzforderungen an $L(\cdot)$ und $U(\cdot)$, die die Flexibilität einschränken würden, sind nicht notwendig.)

- Bei R-Wahrscheinlichkeit ist per definitionem die Menge $\mathcal{M}$ aus (4.4) nicht leer. Die Intervallbewertung $P(\cdot) = [L(\cdot), U(\cdot)]$ ist im „wahrscheinlichkeitsbezogenen Sinn“ nicht widersprüchlich; es gibt klassische Wahrscheinlichkeiten, die mit den Intervallgrenzen $L(\cdot)$ und $U(\cdot)$ verträglich sind. Dies ist eine Minimalforderung an Bewertungen im Wahrscheinlichkeitskontext, allerdings können $L(\cdot)$ und/oder $U(\cdot)$ „zu weit“ sein, d.h. es kann Ereignisse $A \in \mathcal{A}$ geben mit.

$$L(A) < p(A) \qquad \forall p \in \mathcal{M}$$

und/oder

$$U(A) > p(A) \qquad \forall p \in \mathcal{M}.$$

* Allgemeine Überprüfung, ob R-Wsk und F-Wahrscheinlichkeit vorliegt: lineare Optimierung (Weichselberger 2001, Kap. 4.1)

* Subjektivistisch interpretiert bedeutet R-Wahrscheinlichkeit:
  Es können keine Wettsysteme aufgestellt werden, die zu sicherem Verlust führen (Walley (1991): *avoiding sure loss*), eventuell ist aber, wegen „zu weiter Grenzen“, die Indifferenz zu stark ausgeprägt, und es werden dann vorteilhafte Wetten nicht akzeptiert.
  Es gibt also mehrere R-Wahrscheinlichkeiten mit derselben Struktur.

- Bei F-Wahrscheinlichkeit passen hingegen Intervallgrenzen und Struktur in eineindeutiger Weise zusammen. Keine der Intervallgrenzen ist zu weit. Die in der Struktur enthaltenen probabilistischen Information wird in den Intervallgrenzen voll widergespiegelt. Das Weltverhalten ist „kohärent“ (Walley (1991), *coherent*).
  - ∗ Ermittelt man aus den Grenzen die Struktur, verwendet diese als Credalmenge und produziert über (4.2) eine Intervallbewertung $\widetilde{P}(\cdot)$, so ist $P(\cdot) = \widetilde{P}(\cdot)$; man erhält also in diesem Fall die Ausgangswahrscheinlichkeiten.
  - ∗ Eine sehr schwache notwendige Bedingung für F-Wahrscheinlichkeiten ist
$$L(A) = 1 - U(A^C), \quad \forall A \in \mathcal{A}.$$
Bei $|\Omega| > 2$ ist dies aber keineswegs hinreichend.

* Die Überprüfung, ob eine F-Wahrscheinlichkeit vorliegt, kann wieder mittels linearer Optimierung erfolgen (Weichselberger (2001, Kap. 4.2))

Bem. 4.8 Zentrale Grundlage zur numerischen Behandlung, Strukturen als konvexe Polyeder

Ist $\Omega = \{\omega_1, \ldots, \omega_q\}$ endlich mit $|\Omega| = q$, so ist die Struktur jeder F-Wahrscheinlichkeit $\Pi(\cdot)$ über $(\Omega, \mathcal{P}(\Omega))$ ein konvexes Polyeder im $\mathbb{R}^q$ (, wobei wieder klassische Wahrscheinlichkeit $p(\cdot)$ mit der Wahrscheinlichkeitsfunktion $f(\cdot)$ vermöge $f(\omega_\ell) = p(\{\omega_\ell\})$, $\ell = 1, ..., q$, mit einem Punkt des $\mathbb{R}^q$ identifiziert wird.).
Im Allgemeinen ist jede Credal-Menge, die durch *endlich* viele *lineare* Bedingungen an ihre Elemente, also an die klassische Wahrscheinlichkeiten $\pi(\cdot)$ beschrieben wird, ein konvexes Polyeder. Jede dieser Bedingungen hat somit die Form

$$\underline{\alpha} \leq \sum_{\ell=1}^{q} \alpha_\ell p(\{\omega_\ell\}) \leq \overline{\alpha}$$

für jeweils geeignete $\underline{\alpha}, \overline{\alpha}, \alpha_1, \ldots, \alpha_m$.
Insbesondere ist dann jeweils die Extremalpunktmenge $\mathcal{E}(\mathcal{M})$ nicht leer und endlich.

4.3.2 Typische Modellklassen und Anwendungen

- Ordinale Wahrscheinlichkeiten, z.B. $\pi(\{\vartheta_{(1)}\}) \leq \pi(\{\vartheta_{(2)}\}) \leq \pi(\{\vartheta_{(3)}\})$, beschreibt eine Credal-Menge.

- Eine andere Interpretation/Anwendung von Credal-Mengen ist die Modellierung von Gruppenentscheidungen: verschiedene klassische Priori-Verteilungen.

- Kontaminationsmodelle: Robuste Statistik (s.u.)

- *Robuste Bayes Analyse:* Mengen von Prioris

- Vorsichtige Analyse bei unvollständigen Daten (Manski (2003): *Partial Identification*). Betrachte die Menge aller mit den Daten kompatibler Modelle!

- Alternative Inferenzmethoden, insbesondere *logischer Wahrscheinlichkeitsbegriff* (Levi, Kyburg, Weichselberger/Wallner)

- Modellierung Unsicheren Wissens in Expertensystemen (Medizin, auch Wirtschaftswissenschaften); künstliche Intelligenz: Dempster-Shafer-Theorie

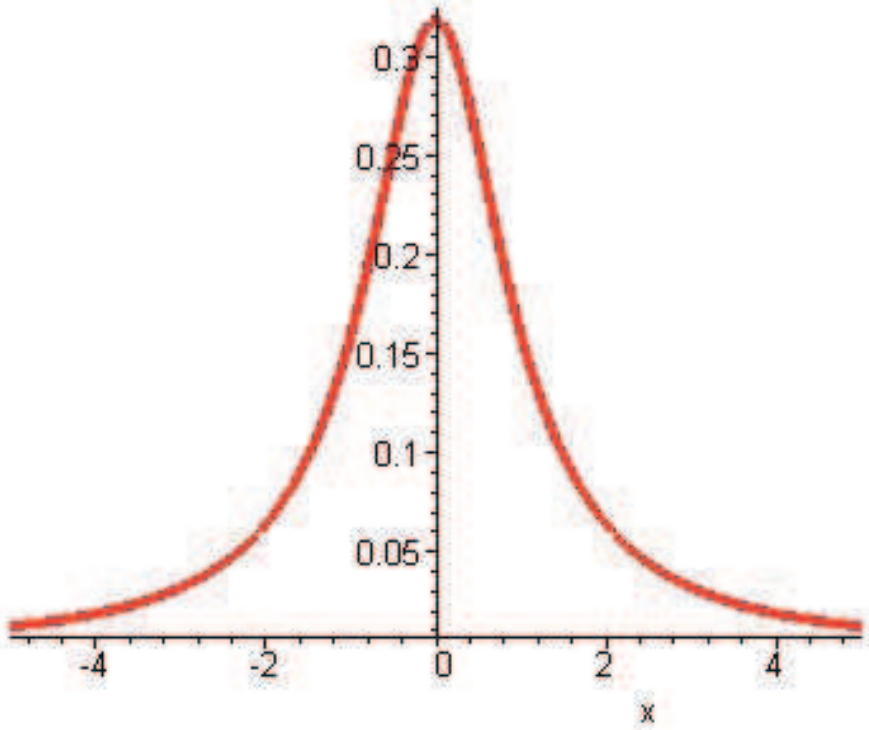
- Finanz- und Finanzierungsmathematik (auch enger Zusammenhang zur sog. Kohärenz von Risikomaßen)

- In manchen Anwendungen wird auch vorgeschlagen, bei der Modellbildung statt mit Plug-in-Punktschätzern mit Plug-in-Intervallschätzern zu arbeiten, um die Tatsache zu nutzen, dass eine Intervallschätzung für eine Wahrscheinlichkeit wesentlich mehr Information enthält als ein entsprechender Punktschätzer.

- Entscheidungstheorie (hier sogleich näher betrachtet)

Bem. 4.9 (Zum Hintergrund: Robustheit)

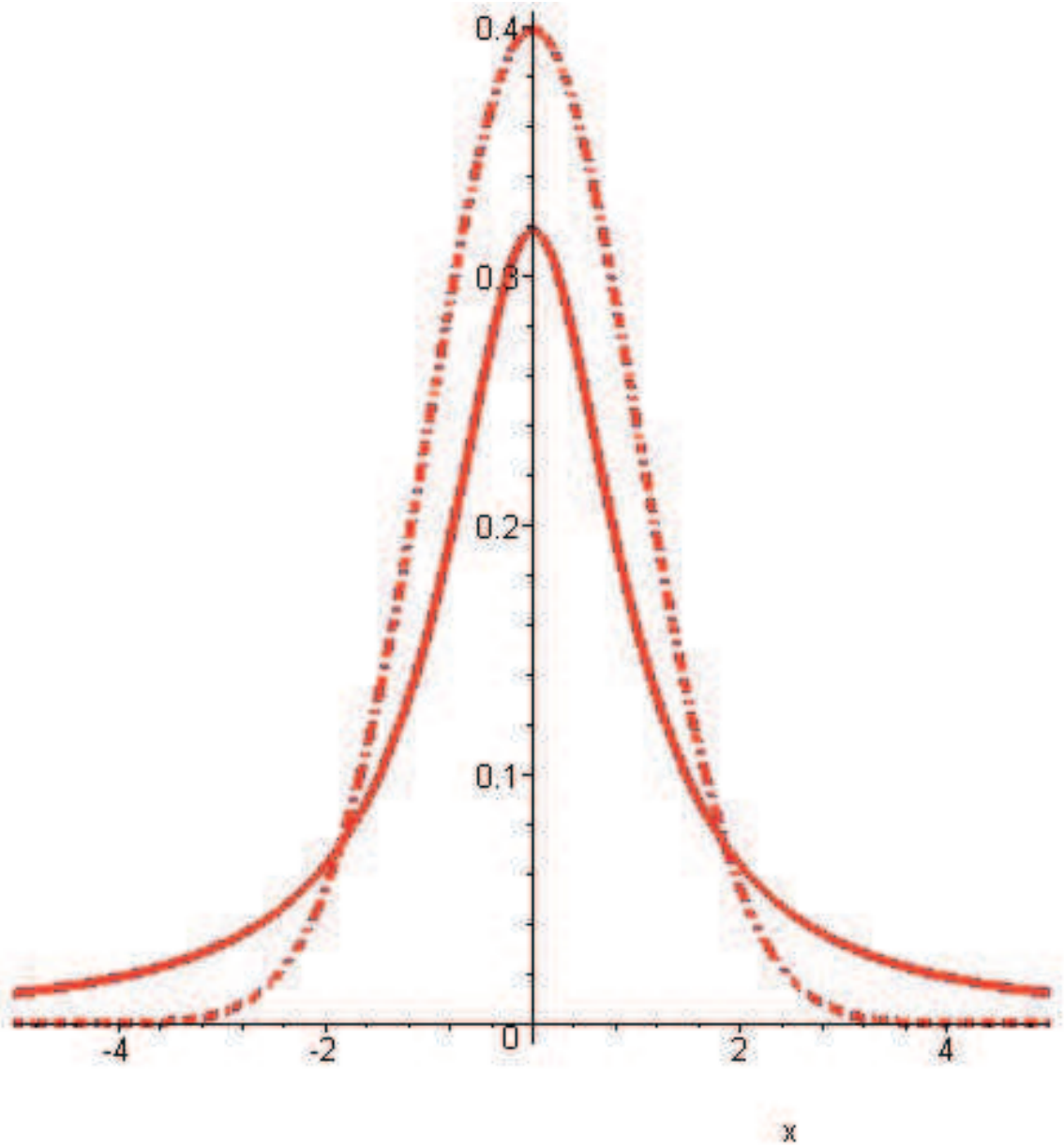
Grundlegend für die Entwicklung der robusten Statistik war die Erkenntnis, dass bei parametrischen Modellen optimale statistische Verfahren sich potentiell desaströs bei minimalen Abweichungen von der Verteilungsannahme verhalten.


0.3
0.25
0.2
0.15
0.1
0.05
-4
-2
0
2
4
x

Betrachtet man beispielsweise die Schätzung des Lageparameters aus einer i.i.d. Stichprobe $X_1, \ldots, X_n$, so gilt für das arithmetische Mittel $\bar{X}$:

i) $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots n$, $\longrightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

ii) $X_i \sim Cauchy(a, b)$, $i = 1, \ldots, n$ $\longrightarrow \bar{X} \sim Cauchy(a, b)$.


0.4
0.3
0.2
0.1
0
-4
-2
2
4
x

Interpretation:
Man versucht deshalb, sich gegen solche Effekte zu „versichern“, indem man Verfahren entwickelt, die im „idealen Modell“ nicht ganz so leistungsfähig sind („Versicherungsprämie“), dafür aber bei kleiner Abweichung vom idealen Modell nicht zusammenbrechen.

Bem. 4.10 (Typische Modellklassen der robusten Statistik)

Man ersetzt ein „ideales statistisches Modell“ $(\Omega, \mathcal{A}, (p_\vartheta)_{\vartheta\in\Theta})$ mit präzisen Wahrscheinlichkeiten p_ϑ durch entsprechende Credalmengen $\mathcal{M}_\vartheta$, die aus allen Verteilungen bestehen, die in einem gewissen Sinn nahe an $p_\vartheta(\cdot)$ sind. Die Credalmenge wird zum Beispiel über Wahrscheinlichkeitsmetriken (siehe z.B. Rüger (2002, S. 41ff)) beschrieben oder durch Schranken an die Dichten oder die Verteilungsfunktionen. Ein besonders gängiges Modell ist auch das ε-*Kontaminations-Modell*, bei dem man für „kleines“ $\varepsilon > 0$

betrachtet:

$$\begin{aligned} \mathcal{M}_{\vartheta,\varepsilon} = \{q(\cdot) \in \mathcal{P}(\Omega, \mathcal{A}) | \\ q(\cdot) = (1-\varepsilon)p_\vartheta(\cdot) + \varepsilon \cdot r(\cdot), \\ r \in \mathcal{P}(\Omega, \mathcal{A})\} \end{aligned} \tag{4.5}$$

mit $\mathcal{P}(\Omega, \mathcal{A})$ wieder als der Menge aller Wahrscheinlichkeitsmaße über $(\Omega, \mathcal{A})$.

4.4 Erste Anwendungen in der Entscheidungstheorie

4.4.1 Verallgemeinerte Erwartungswerte

intervallwertige Wahrscheinlichkeiten $\Rightarrow$ intervallwertige Erwartungswerte
Zwei Definitionen in der Literatur üblich.

Def. 4.11 (Verallgemeinerte Erwartungswerte)

Betrachtet werden eine Zufallsvariable X auf einem Meßraum $(\Omega, \mathcal{A})$ und eine F-Wahrscheinlichkeit $P(\cdot) = [L(\cdot); U(\cdot)]$ mit Struktur $\mathcal{M}$.

a) X heißt $\mathcal{M}$-*integrierbar*, wenn X für jedes $p \in \mathcal{M}$ p-integrierbar ist.

b) Dann heißt für $\mathcal{M}$-integrierbares X

$$\mathbb{E}_{\mathcal{M}} X := [\underline{\mathbb{E}}_{\mathcal{M}} X; \overline{\mathbb{E}}_{\mathcal{M}} X] := [\inf_{p \in \mathcal{M}} \mathbb{E}_p X; \sup_{p \in \mathcal{M}} \mathbb{E}_p X] \tag{4.3}$$

intervallwertiger Erwartungswert i.e.S.

b) Ist X zudem nicht negativ, so heißen die Ausdrücke

$$\int X \, \mathrm{dL} := \int L(\{X > x\}) \, \mathrm{dx} \tag{4.4}$$

$$\int X \, \mathrm{dU} := \int U(\{X > x\}) \, \mathrm{dx} \tag{4.5}$$

Choquet-Integrale zu $L(\cdot)$ und $U(\cdot)$.

Bem. 4.12 (Zu Def. 4.11)

i) Die Definitionen in Teil a) und b) lassen sich unmittelbar auf Credal-Mengen ausdehnen.

ii) Das Choquet-Integral ist v.a. in den Wirtschaftswissenschaften populär („*Choquet Expected Utility*“).

iii) Ist Ω endlich und $\mathcal{M}$ die Struktur einer F-Wahrscheinlichkeit oder allgemein ein konvexes Polyeder (vgl. Bemerkung 4.8), so können $\underline{\mathbb{E}}_{\mathcal{M}}X$ und $\bar{\mathbb{E}}_{\mathcal{M}}X$ einfach durch lineare Optimierung bestimmt werden:

$$\sum_{\omega\in\Omega} X(\omega)p(\{w\}) \longrightarrow \max_{p(\cdot)}$$

unter den durch $\mathcal{M}$ beschriebenen Nebenbedingungen.

iv) Allgemein gilt

$$\int X \, \mathrm{dL} \leq \underline{\mathbb{E}}_{\mathcal{M}} X \leq \overline{\mathbb{E}}_{\mathcal{M}} X \leq \int X \, \mathrm{dU} \tag{4.6}$$

v) Unter gewissen Regularitätsbedingenen, die zum Beispiel bei endlichem Ω erfüllt sind, gilt:

$$\int X \mathrm{dL} = \underline{\mathbb{E}}_{\mathcal{M}} X \qquad \text{für } \textit{alle } \mathcal{M}\text{-integrierbaren } X$$

genau dann, wenn $L(\cdot)$ *zwei-monoton*[11] ist, d.h. wenn für alle $A, B \in \mathcal{A}$

$$L(A \cup B) + L(A \cap B) \geq L(A) + L(B) \tag{4.7}$$

gilt.

[11] Der Begriff der Zweimonotonie ist rein technischer Natur; lässt sich inhaltlich nicht direkt interpretieren.

Bem. 4.13 (Exkurs: Woher kommen die Formeln (4.4) und (4.5)?)

Es gilt allgemein für nichtnegatives X

$$\mathbb{E}_P X = \int_0^\infty p(\{X > x\})\,\mathrm{dx}, \tag{4.8}$$

sofern die rechte Seite endlich ist.

In dieser Beziehung $p(\cdot)$ durch $L(\cdot)$ und $U(\cdot)$ ersetzen.

Beweis von (4.8) durch partielle Integration:

$$\int_0^\infty p(\{X > x\})\,\mathrm{dt} = \int_0^\infty \underbrace{(1 - F(x))}_{u} \cdot \underbrace{1}_{v'}\,\mathrm{dx} =$$

$$= \left[\underbrace{(1 - F(x))}_{u} \cdot \underbrace{x}_{v}\right]_0^\infty - \int_0^\infty - \underbrace{f(x)}_{u'}\underbrace{x}_{v}\,\mathrm{dx}.$$

Damit gilt, falls

$$\lim_{x\to\infty}(1 - F(x)) \cdot x = \lim_{x\to\infty} x - x \cdot F(x) = 0 \tag{4.9}$$

in der Tat

$$\int_0^\infty p(\{X > x\})\,\mathrm{dt} = 0 + \int_0^\infty x \cdot f(x)\,\mathrm{dx} = \mathbb{E}X.$$

(4.9) stellt nur sicher, dass die rechte Seite von (4.8) endlich ist. Angenommen, es gibt ein x_0 so, dass $x - x \cdot F(x) \geq c > 0$ für alle $x \geq x_0$, dann wäre

$$\begin{aligned}
&\int_0^\infty (1 - F(x)) \, \mathrm{dx} = \\
= &\int_0^{x_0} (1 - F(x)) \, \mathrm{dx} + \int_{x_0}^\infty (1 - F(x)) \, \mathrm{dx} = \\
\geq &\int_0^{x_0} (1 - F(x)) \, \mathrm{dx} + \int_{x_0}^\infty c \, \mathrm{dx} = \infty
\end{aligned}$$

Diskreter Spezialfall: kann X nur natürliche Zahlen annehmen, so ist

$$\mathbb{E}X = \sum_{n \in \mathbb{N}} P(\{X \geq n\})$$

Beweis:

$$\begin{aligned}
\mathbb{E}X &= \sum_{n\in\mathbb{N}} n \cdot P(\{X = n\}) = \\
&= 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + 3 \cdot P(X = 3) + \dots \\
&= P(X = 1) + \\
&\quad P(X = 2) + P(X = 2) + \\
&\quad \underbrace{P(X = 3)} + \underbrace{P(X = 3)} + \underbrace{P(X = 3)} + \\
&\quad \dots \\
&= P(X \geq 1) + P(X \geq 2) + P(X \geq 3) + \dots
\end{aligned}$$

4.4.2 Verallgemeinerter Erwartungsnutzen und Optimalitätskriterien

Einen guten Überblick bietet: Troffaes (2007, International Journal Approximate Reasoning).

Def. 4.14 (Verallgemeinerter Erwartungsnutzen)

Gegeben sei ein datenfreies Entscheidungsproblem $(\mathbb{A}, \Theta, u(\cdot))$, und eine *verallgemeinerte Priori-Bewertung*, also eine Intervallwahrscheinlichkeit $\Pi(\cdot)$ mit Struktur $\mathcal{M} \neq \emptyset$ oder eine konvexe Credal-Menge $\mathcal{M} \neq \emptyset$ auf $(\Theta, \sigma(\Theta))$. Dann heißt

$$\mathbb{E}_{\mathcal{M}}(u(a)) = [\inf_{\pi \in \mathcal{M}} \mathbb{E}_{\pi}(u(a)); \sup_{\pi \in \mathcal{M}} \mathbb{E}_{\pi}(u(a))]$$

der verallgemeinerte Erwartungsnutzen der Aktion a zur verallgemeinerten Priori-Bewertung $\Pi(\cdot)$ bzw. $\mathcal{M}$.

Bem. 4.15 (Intervallordnungen)

Da Intervalle in der Regel unvollständige Ordnungen produzieren, bedarf die Forderung „maximiere den verallgemeinerten Erwartungsnutzen $\mathbb{E}_{\mathcal{M}}(u(a))$ unter allen $a \in \mathbb{A}$“ noch der genaueren Präzisierung, die sich in verschiedenen Optimalitätsbegriffen niederschlägt:

- Entweder: Aussagen analog zur Zulässigkeit: *E-Zulässigkeit, Intervallordnungen im eigentlichen Sinn*, die i.A. keine vollständige Ordnungen produzieren.

- oder alternativ: Intervalle durch *eine* reelle Zahl beschreiben (Mittelpunkt, untere Intervallgrenze, etc. ...): *Repräsentationen* z.B. durch minimalen Erwartungsnutzen oder gewichtete Summe des maximalen und minimalen Erwartungsnutzens.

Bem. 4.11 (Optimale Aktionen)

a1) Betrachtet werde die Situation von Def. 4.14. Dann heißt

$$\begin{aligned} \Phi(\cdot, \mathcal{M}) : \mathbb{A} &\longrightarrow \mathbb{R} \\ a &\longmapsto \underline{\mathbb{E}}_{\mathcal{M}}(u(a)) \end{aligned}$$

Max E Min-Kriterium und jede Aktion $a^* \in \mathbb{A}$ mit

$$\Phi(a^*, \mathcal{M}) \geq \Phi(a, \mathcal{M}), \quad \forall a \in \mathbb{A},$$

also mit

$$\underline{\mathbb{E}}_{\mathcal{M}} u(a^*) \geq \underline{\mathbb{E}}_{\mathcal{M}}(u(a)), \quad \forall a \in \mathbb{A}$$

Max E Min Aktion zur verallgemeinerten Priori-Bewertung $\Pi(\cdot)$ *bzw.* $\mathcal{M}$.

Man würde dann genau erwarten, dass sich im Extremfall perfekte probabilistische Information des Bayeskriterium und bei völliger Ambigurität das Maximin-Kriterium ergibt. Dies ist in der Tat so, vgl. 4.4.4.

a2) Für $0 \leq \eta \leq 1$ heißt

$$\begin{aligned} \Phi_\eta(\mathcal{M}) : \mathbb{A} &\longrightarrow \mathbb{R} \\ \mathbb{A} &\longmapsto \eta \cdot \underline{\mathbb{E}}_{\mathcal{M}}(u(a)) + (1-\eta)\overline{\mathbb{E}}_{\mathcal{M}}(u(a)) \end{aligned}$$

lineare Repräsentation mit der Vorsicht η und jede Aktion $a^{*\eta}$ mit

$$\Phi_\eta(a^{*\eta}, \mathcal{M}) \geq \Phi_\eta(a, \mathcal{M}), \quad \forall a \in \mathbb{A},$$

also mit

$$\eta \cdot \underline{\mathbb{E}}_{\mathcal{M}} u(a^{*\eta}) + (1-\eta)\overline{\mathbb{E}}_{\mathcal{M}} u(a^{*\eta}) \geq$$
$$\eta \cdot \underline{\mathbb{E}}_{\mathcal{M}}(u(a)) + (1-\eta)\overline{\mathbb{E}}_{\mathcal{M}}(u(a)), \forall a \in \mathbb{A},$$

optimale Aktion zur verallgemeinerten Priori-Bewertung $\Pi(\cdot)$ *bzw.* $\mathcal{M}$ *bei der Vorsicht* η.

b) Eine Aktion $a_\pi \in \mathbb{A}$ heißt *E-zulässig (E-admissible)* bezüglich der verallgemeinerten Priori-Bewertung $\Pi(\cdot)$ bzw. $\mathcal{M}$, falls es ein $\pi(\cdot) \in \mathcal{M}$ gibt, so dass a_π Bayes-Aktion zu $\pi(\cdot)$ ist.

Bem. 4.12

- Der Begriff Max E Min (Maximiere den minimalen Erwartungswerts) geht auf die sogenannte Theorie der *linearen partiellen Information* von Kofler und Menges zurück, die in den Wirtschaftswissenschaften in den 80iger Jahren schon einmal populär war.
  Andere gängige Bezeichnungsweisen für dieses Kriterium sind insbesondere Γ-*Maximinkriterium* und in Situationen, in denen in (4.6) Gleichheit gilt, *Choquet-Erwartungsnutzen-Optimalität*.

- Dem Max E Min-Kriterium liegt eine skeptische Perspektive zugrunde, man führt sozusagen einen „lokalen Maximin-Ansatz“ durch, indem man sich auf das ungünstigste Element von $\mathcal{M}$ konzentriert.

- Analog entspricht die lineare Repräsentation mit der Vorsicht η einem „lokalen Hurwicz-Kriterium“. Die Vorsicht η modelliert die Einstellung zur Ambiguität; $\eta > \frac{1}{2}$ bedeutet Ambiguitätsaversion, $\eta < \frac{1}{2}$ Ambiguitätsfreudigkeit. Der Fall $\eta = 1$ führt auf das Max E Min-Prinzip. (Man vergleiche hierzu auch Satz 4.14.)

4.4.3 Zurück zum Ellsberg-Experiment

Bsp. 4.13 (Modellierung des Ellsberg-Experiments)

Situation 1
a_1 auf $\{r\}$ setzen
a_2 auf $\{s\}$ setzen

	ϑ_1	ϑ_2	ϑ_3
	r	g	s
a_1	1	0	0
a_2	0	0	1

vgl. vorne: nochmals wieder in Situation 1 für jedes π

$$\mathbb{E}_\pi(u(a_1)) = \sum_{j=1}^{3} u(a_1, \vartheta_j)\pi(\{\vartheta_j\}) = \pi(\{r\})$$

$$\mathbb{E}_\pi(u(a_2)) = \sum_{j=1}^{3} u(a_2, \vartheta_j)\pi(\{\vartheta_j\}) = \pi(\{s\})$$

also mit

$$\mathcal{M} = \left\{\pi(\cdot) \text{ Wsk auf } \mathcal{P}\left(\{r,s,g\}\right) \middle| \pi(\{r\}) = \frac{1}{3}; \pi(\{s,g\}) = \frac{2}{3}\right\}.$$

$$\underline{\mathbb{E}}_{\mathcal{M}}(u(a_1)) = \inf_{\pi\in\mathcal{M}} \mathbb{E}_\pi(u(a_1)) = \inf_{\pi\in\mathcal{M}} \pi(\{r\}) = \frac{1}{3}$$

und

$$\underline{\mathbb{E}}_{\mathcal{M}}(u(a_2)) = \inf_{\pi\in\mathcal{M}} \mathbb{E}_\pi(u(a_2)) = \inf_{\pi\in\mathcal{M}} \pi(\{s\}) = 0$$

sowie

$$\overline{\mathbb{E}}_{\mathcal{M}}(u(a_1)) = \sup_{\pi\in\mathcal{M}} \mathbb{E}_\pi(u(a_1)) = \sup_{\pi\in\mathcal{M}} \pi(\{r\}) = \frac{1}{3}$$

und

$$\overline{\mathbb{E}}_{\mathcal{M}}(u(a_2)) = \sup_{\pi\in\mathcal{M}} \mathbb{E}_\pi(u(a_2)) = \sup_{\pi\in\mathcal{M}} \pi(\{s\}) = \frac{2}{3}.$$

Folglich ist mit

$$\begin{aligned}
\Phi_\eta(a_i, \mathcal{M}) &:= \eta \cdot \underline{\mathbb{E}}_{\mathcal{M}}(u(a_i)) + (1-\eta)\overline{\mathbb{E}}_{\mathcal{M}}(u(a_i)) \\
\Phi_\eta(a_1, \mathcal{M}) &= \eta \cdot \frac{1}{3} + (1-\eta) \cdot \frac{1}{3} = \frac{1}{3} \\
\Phi_\eta(a_2, \mathcal{M}) &= \eta \cdot 0 + (1-\eta) \cdot \frac{2}{3} = (1-\eta) \cdot \frac{2}{3}
\end{aligned}$$

und damit

$$\Phi_\eta(a_1, \mathcal{M}) > \Phi_\eta(a_2, \mathcal{M}) \Longleftrightarrow \eta > \frac{1}{2}.$$

in Situation 2
a_3 auf $\{r\}$ oder $\{g\}$ setzen
a_4 auf $\{s\}$ oder $\{g\}$ setzen

	r	g	s
a_3	1	1	0
a_4	0	1	1

$$\mathbb{E}_\pi(u(a_3)) = \sum_{j=1}^{3} u(a_3, \vartheta_j)\pi(\{\vartheta_j\}) = \pi(\{r\}) + \pi(\{g\})$$

$$\mathbb{E}_\pi(u(a_4)) = \sum_{j=1}^{3} u(a_4, \vartheta_j)\pi(\{\vartheta_j\}) = \pi(\{g\}) + \pi(\{s\})$$

also

$$\underline{\mathbb{E}}_{\mathcal{M}}(u(a_3)) = \inf_{\pi\in\mathcal{M}} \mathbb{E}_\pi(u(a_3)) = \inf_{\pi\in\mathcal{M}} (\pi(\{r\}) + \pi(\{g\})) = \frac{1}{3}$$

und[12]

$$\underline{\mathbb{E}}_{\mathcal{M}}(u(a_4)) = \inf_{\pi\in\mathcal{M}} \mathbb{E}_\pi(u(a_4)) = \inf_{\pi\in\mathcal{M}} (\pi(\{g\}) + \pi(\{s\}) = \frac{2}{3}$$

sowie

$$\overline{\mathbb{E}}_{\mathcal{M}}(u(a_3)) = \sup_{\pi\in\mathcal{M}} \mathbb{E}_\pi(u(a_3)) = \sup_{\pi\in\mathcal{M}} (\pi(\{r\}) + \pi(\{g\})) = 1$$

[12]Hier ist Vorsicht geboten: Man muss die internen Restriktionen berücksichtigen; es muss „dasselbe $\pi(\cdot)$ sein, bezüglich dem das Infimum gebildet wird“:

$$\inf_{\pi\in\mathcal{M}} (\pi(\{g\}) + \pi(\{s\})) \neq \inf_{\pi\in\mathcal{M}} \pi(\{g\}) + \inf_{\pi\in\mathcal{M}} \pi(\{s\}).$$

Dies ist der wesentliche Unterschied zwischen Intervallwahrscheinlichkeiten und einer naiven Intervallarithmetik.

und

$$\overline{\mathbb{E}}_{\mathcal{M}}(u(a_4)) = \sup_{\pi\in\mathcal{M}} \mathbb{E}_\pi(u(a_4)) = \sup_{\pi\in\mathcal{M}} (\pi(\{g\}) + \pi(\{s\})) = \frac{2}{3}.$$

Damit ist hier mit dem Kriterium $\Phi_\eta(\cdot, \mathcal{M})$ von oben

$$\Phi(a_3, \mathcal{M}) = \eta \cdot \frac{1}{3} + (1-\eta) \cdot 1 = 1 - \frac{2}{3}\eta$$

und

$$\Phi(a_4, \mathcal{M}) = \eta \cdot \frac{2}{3} + (1-\eta) \cdot \frac{2}{3} = \frac{2}{3}$$

$$\Phi(a_4, \mathcal{M}) > \Phi(a_3, \mathcal{M}) \Longleftrightarrow \frac{2}{3} > 1 - \frac{2}{3}\eta \Longleftrightarrow \eta > \frac{1}{2}.$$

Die ambiguitätsaversen Personen ($\eta > \frac{1}{2}$) bevorzugen a_1 vor a_2 *und* a_4 vor a_3, die ambiguitätsfreudigen a_2 vor a_1 *und* a_3 vor a_4. Damit sind genau die am häufigsten geäußerten Präferenzen, die ja der klassischen Wahrscheinlichkeitsrechnung widersprochen haben, geeignet modelliert.

4.4.4 Verallgemeinerter Erwartungsnutzen als Überbau über die klassischen Kriterien der Entscheidungstheorie

Die Intervallwahrscheinlichkeit und der darauf aufbauende verallgemeinerter Erwartungsnutzen wurde als Mittelweg zwischen Bayes- und Maximin-Sicht motiviert. In der Tat gilt der folgende Satz:

Satz 4.14 (Klassische Entscheidungskriterien als Extrempole I)

In der Situation von Def. 4.16 gilt:

a) Herrscht ideale Stochastizität, ist also $\mathcal{M} = \{\pi\}$ (einelementig), so gilt:
Eine Aktion $a^* \in \mathbb{A}$ ist genau dann optimal im Sinne der in Definition 4.11 genannten Kriterien, wenn a^* Bayes-Aktion (im klassischen Sinn von Kapitel **??**) zu $\pi(\cdot)$ ist.

b) Herrscht Unsicherheit i.e.S., besteht $\mathcal{M}$ also aus allen klassischen Wahrscheinlichkeiten auf $(\Theta, \sigma(\Theta))$, so gilt:
Eine Aktion $a^* \in \mathbb{A}$ ist genau dann Max E Min Aktion zu $\mathcal{M}$, wenn a^* Maximin-Aktion im Sinne von Kapitel **??** ist, und optimale Aktion zu $\mathcal{M}$ bei der Vorsicht η, wenn a^* Hurwicz-Aktion zum Optimismusparameter $1 - \eta$ im Sinne von Kapitel **??** ist.

Es lassen sich ähnliche Aussagen für die E-Zulässigkeit treffen:

Proposition 4.15 Klassische Entscheidungskriterien als Extrempole II

i) Herrscht in der Situation von 4.16 ideale Stochastizität, ist also $\mathcal{M} = \{\pi\}$ (einelementig), so gilt: Die Menge der E-zulässigen Aktionen zu $\mathcal{M}$ ist identisch zu der Menge der Bayes-Aktionen zu $\pi(\cdot)$.

ii) Herrscht Unsicherheit i.e.S., besteht $\mathcal{M}$ also aus allen klassischen Wahrscheinlichkeiten auf $(\Theta, \sigma(\Theta))$ und ist $\mathbb{A}$ konvex, so gilt mit $\mathcal{M}^+ := \{\pi(\cdot) \in \mathcal{M} | \pi(\cdot) > 0\}$: $a^* \in \mathbb{A}$ ist genau dann zulässig (im Sinne von Kapitel 2.0), wenn a^* E-zulässig bezüglich $\mathcal{M}^+$ ist.

4.4.5 Berechnung optimaler Aktionen über die Extremalpunktmenge

Die Tatsache, dass gemaß Bem. 4.8 auf endlichen Räumen Intervall-Wahrscheinlichkeiten und weitere typische Credalmengen konvexe Polyeder sind, ist für die konkrete Berechnung optimaler Aktionen sehr hilfreich, denn sie erlaubt den Rückzug auf die endlichen, nichtleeren Extremalpunktmenge.

Satz 4.16 (Berechnung verallgemeinerter Erwartungsnutzen)

Gegeben sei ein datenfreies Entscheidungsproblem $(\mathbb{A}, \Theta, u(\cdot))$, und eine *verallgemeinerte Priori-Bewertung*, also eine Intervallwahrscheinlichkeit $\Pi(\cdot)$ mit Struktur $\mathcal{M} \neq \emptyset$ oder eine Credal-Menge $\mathcal{M} \neq \emptyset$ auf $(\Theta, \sigma(\Theta))$, die zusätzlich die Bedingungen von Bem. 4.8 erfüllt. Dann gilt mit $\mathcal{E}(\mathcal{M})$ als Extremalpunktmengen:

$$\underline{\mathbb{E}}_{\mathcal{M}}(u(a)) = \min_{\pi \in \mathcal{E}(\mathcal{M})} \sum_{j=1}^{m} u(a, \vartheta_j)\pi(\{\vartheta_j\})$$

und

$$\bar{\mathbb{E}}(u(a)) = \max_{\pi \in \mathcal{E}(\mathcal{M})} \sum_{j=1}^{m} u(a, \vartheta_j)u(\{\vartheta_j\}).$$

Bem. 4.17

- Dieser Satz ist eigentlich ein Korollar zu Satz 2.17 aus Kapitel 1, wenn man sich an Bemerkung 4.8 erinnert, demzufolge Strukturen und entsprechende Credalmengen konvexe Polyeder bilden.
- Zur Berechnung des verallgemeinerten Erwartungsnutzen brauchen also nur endlich viele klassische Erwartungswerte bestimmt werden.

Bsp. 4.18 (Investitionsproblem)

Gegeben sei das Investitionsproblem (vgl. das Beispiel aus Abschnitt 1.3.4)

	ϑ_1	ϑ_2	ϑ_3
a_1	10000	2000	-15000
a_2	1000	1000	0

und eine Experteneinschätzung in Form einer ordinalen Wahrscheinlichkeit, d.h. der Experte ordnet nur die Umweltzustände nach der Wahrscheinlichkeit ihres Eintretens. Gegeben sei die Credalmenge

$$\mathcal{M} = \{\pi(\cdot)|\pi(\{\vartheta_1\}) \geq \pi(\{\vartheta_2\}) \geq \pi(\{\vartheta_3\})\}.$$

Berechne die Max E Min-Aktion!

$\mathcal{M}$ besteht aus linearen Restriktionen, die sich aus der ordinalen Wahrscheinlichkeitsbewertung ergeben, z.B.

$$\pi(\{\vartheta_{j_1}\}) \geq \pi(\{\vartheta_{j_2}\}) \Longleftrightarrow 0 \leq \pi(\{\vartheta_{j_1}\}) - \pi(\{\vartheta_{j_2}\}),$$

und den trivialen Bedingungen:

$$0 \leq \pi(\{\vartheta_j\}) \leq 1,$$

$$\pi(\{\vartheta_1\}) + \pi(\{\vartheta_2\}) + \pi(\{\vartheta_3\}) = 1,$$

also ist $\mathcal{M}$ in der Tat ein konvexes Polyeder.

Berechnet man die Extremalpunktmenge $\mathcal{E}(\mathcal{M})$, so ergibt sich [13]

[13](siehe z. B. E. Kofler: Prognosen und Stabilität bei unvollständiger Information. Campus (Frankfurt; New York), 1989.)

$$\mathcal{E}(\mathcal{M}) = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix} \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \right\}.$$

Damit ist bei a_1

$$\underline{\mathbb{E}}_{\mathcal{M}}(u(a_1)) = \inf_{\pi \in \mathcal{M}} \mathbb{E}_\pi(u(a_1)) = \min_{\pi \in \mathcal{E}(\mathcal{M})} \left(\sum_{j=1}^{m} u(a_1, \vartheta_j)\pi(\{\vartheta_j\}) \right)$$

$$= \min \left(10000 \cdot 1 + 0; 10000 \cdot \frac{1}{2} + 2000 \cdot \frac{1}{2} + 0; \frac{1}{3}(10000 + 20000 - 15000) \right)$$

$$= \min (10000; 6000; -1000) = -1000.$$

und für a_2

$$\underline{\mathbb{E}}_{\mathcal{M}}(u(a_2)) = \inf_{\pi\in\mathcal{M}} \mathbb{E}_\pi(u(a_2)) = \min_{\pi\in\mathcal{E}(\mathcal{M})} \left(\sum_{j=1}^{m} u(a_2,\vartheta_j)\pi(\{\vartheta_j\}) \right)$$

$$= \min\left(1000\cdot 1; 1000\cdot\frac{1}{2} + 1000\cdot\frac{1}{2}; \frac{1}{3}(1000+1000+0)\right)$$

$$= \frac{2000}{3}$$

$$\underline{\mathbb{E}}_{\mathcal{M}}(u(a_1)) = \min_{\pi\in\mathcal{E}(\mathcal{M})} \mathbb{E}_\pi(u(a_1)) < \min_{\pi\in\mathcal{E}(\mathcal{M})} \mathbb{E}_\pi(u(a_2)) = \underline{\mathbb{E}}_{\mathcal{M}}(u(a_2)).$$

Also ist a_2 Max E Min-Aktion.

Bem. 4.19 (Praktische Berechnung)

- Bei größeren Problemen ist die Berechnung der Extremalpunktmenge durchaus aufwendig, aber es gibt spezielle Algorithmen (aus der „computational geometry“, die mittlerweile auch in R verwendbar sind).[14]

- Die Berechnung von $\underline{\mathbb{E}}_{\mathcal{M}}(u(a))$ und $\overline{\mathbb{E}}_{\mathcal{M}}(u(a))$ für festes $a \in \mathbb{A}$ kann effizient mit linearer Optimierung erfolgen (vgl. Bem. 4.8 und Bem. 4.12). In vielen Modellen kann auch auf die Darstellung über das Choquet-Integral zurückgegriffen werden, das eine explizite Berechnungsformel liefert. Man kann ferner sogar zeigen, dass sich die Berechnung der Max E Min optimalen randomisierten Aktionen auf ein *einziges* lineares Optimierungsproblem zurückführen lässt.[15]

---

[14] vgl. das RCDD-Package (Geyer and Meeden, 2013, CRAN) als Interface für Fukudas „cdd-Bibliothek“.

[15] Vgl. Utkin & Augustin (2005, Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications), wo auch Algorithmen für die anderen Kriterien entwickelt werden.

5 Konditionale Inferenz und Entscheidungstheorie, Ausblick

5.1 Konditionale Bayes-Inferenz: Begrifflicher Hintergrund und „Erinnerungen“

In Kapitel 1.5.4 wurde davon gesprochen, dass die Lösung/ Darstellung des datengestützten Entscheidungsproblems über das Auswertungsproblem und die damit verbunden Suche nach optimalen Entscheidungsfunktionen durchaus auch auf Kritik stösst. Diese stützt sich v.a. einerseits

- auf die immense computationale Komplexität

und andererseits ganz prinzipiell

- auf die Problematik kontrafaktischer Ereignisse bei der Bewertung von Entscheidungsfunktionen mittels der Risikofunktion.

Dies legt eine konditionale Sicht als mögliche Alternative nahe: Es werden optimale Lösungen für die konkret beobachtete Datenkonstellation $\{x\}$ gesucht („auf $\{x\}$ konditionierte Betrachtung“). Dies führt auf die „übliche Bayes-Inferenz“, die mit Hilfe der sogenannten Posteriori-Verteilung gegeben die Daten arbeitet, und in diesem Sinne hier zur Abgrenzung als „*konditionale Bayes-Inferenz*“ bezeichnet werde.
Zur Vorbereitung werde an das Theorem von Bayes in seiner allgemeinen Form erinnert:

Proposition 5.1 (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:

$$f_{U|X}(u|x) = \frac{f_{X|U}(x|u) \cdot f_U(u)}{f_X(x)} \tag{5.2}$$

mit

$$f_X(x) = \int f_{X|U}(x|u) \cdot f_U(u)\, d\nu(u). \tag{5.3}$$

Bem. 5.2

Bei stetigem U mit Dichte $f_U(u)$ erhält man also Proposition 5.1 mit

$$f_X(x) = \int f_{X|U}(x|u) \cdot f_U(u) du. \tag{5.4}$$

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

$$p(\{U = u\}|\{X = x\}) = \frac{p(\{X = x\}|\{U = u\}) \cdot p(\{U = u\})}{p(\{X = x\})} \tag{5.5}$$

mit

$$p(\{X = x\}) = \sum_{u \in \mathcal{U}} p(\{X = x\}|\{U = u\}) \cdot p(\{U = u\}) \tag{5.6}$$

Betrachtet werde im Folgenden immer einer dieser beiden Spezialfälle . Die allgemeinere Formulierung über beliebige Dichten bezüglich geeigneter dominierender Maße ist unproblematisch.

Bem. 5.3 (Normierungskonstante)

$f_X(x)$ aus (5.3) 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.
Man schreibt dann mit $\propto$ als Symbol für „proportional zu“

$$f_{X|U}(x|u) \propto f_{X|U}(u|x) \cdot f_U(u).$$

Bem. 5.4 (Konditionale Bayes-Inferenz: Konzeptionelle Hintergründe)

Gegeben sei ein datengestütztes Entscheidungsproblem $((\mathbb{A}, \Theta, l(\cdot)); (\mathcal{X}, \sigma(\mathcal{X}), (p_\vartheta(\cdot))_{\vartheta\in\Theta})$, wobei $p_\vartheta(\cdot)$ die Wahrscheinlichkeitsfunktion bzw. Dichte $f_\vartheta(\cdot)$ besitze, und eine Priori-Verteilung auf einem geeigneten Maßraum $(\Theta, \sigma(\Theta))$ mit Dichte bzw. Wahrscheinlichkeitsfunktion $\pi(\cdot)$.
Sei, als gedankliche Hilfskonstruktion, U („Umwelt “, „Natur “) eine „Zufallsgrösse“ (Zufallsvariable/-element), die das Eintreten des Umweltzustands ϑ beschreibt und X eine Zufallsgrösse, die den Ausgang des Informationbeschaffungsexperiments beschreibt. Dann gelte für die Dichte bzw. Wahrscheinlichkeitsfunktion $f_{X,U}(x, \vartheta)$ der gemeinsamen Verteilung von X und U für alle $x \in \mathcal{X}$ und $\vartheta \in \Theta$:

$$f_{X,U}(x, \vartheta) = \pi(\vartheta) \cdot f_\vartheta(x)$$

Das heißt für alle $x \in \mathcal{X}$ und $\vartheta \in \Theta$ gilt

$$f_\vartheta(x) = f_{X|U}(x|\vartheta) \ ;$$

die Verteilung der Zufallsgrösse aus der Informationsstruktur wird als bedingte Verteilung von X gegeben U interpretiert (!!).
Dann ergibt sich (!!!) für jedes x aus dem Satz von Bayes gemäß Proposition 5.1 für (eine Version der bzw.) die Dichte bzw. Wahrscheinlichkeitsfunktion $\pi(\vartheta|x)$ der bedingten Verteilung von U gegeben $X = x$:

$$\pi(\vartheta|x) = c(x) \cdot f_\vartheta(x) \cdot \pi(\vartheta) \tag{5.7}$$

mit

$$\frac{1}{c(x)} = f_X(x) = \int f_{X|U}(x|\vartheta)\pi(\vartheta)d\vartheta \tag{5.8}$$

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

$$\frac{1}{c(x)} = p(\{X = x\}) = \sum_{j=1}^{m} p(\{X = x\}|\{U = \vartheta_j\}) \cdot \pi(\{U = \vartheta_j\}) \tag{5.9}$$

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

Für jede Beobachtung $x \in \mathcal{X}$ wird die bedingte Verteilung von U gegeben $X = x$ als *Posteriori-Verteilung des Parameters ϑ nach der Beobachtung x* bezeichnet. Die zugehörige Dichte bzw. Wahrscheinlichkeitsfunktion $\pi(\vartheta|x)$ heißt *Posteriori-Dichte nach der Beobachtung*, und $f_\vartheta(x)$ heißt *Likelihood*.

Die marginale Verteilung von X mit Dichte $f_X(x)$ aus (5.8) bzw. Wahrscheinlichkeitsfunktion $(p(\{X = x\}))_{x\in\mathcal{X}}$ aus (5.9) heißt *Priori-Prädiktive-Verteilung*.

Die Größen$f_X(x)$ und $p(\{X = x\})$ sind nicht zu verwechseln mit den als bedingte Verteilungen interpretierten $f_\vartheta(x)$ und $p_\vartheta(\{X = x\})$.

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. 5.5 (Bayes-Postulat (nicht entscheidungstheoretisch))
Nach der Beobachtung der Stichprobe enthalte die (klassische) Posteriori-Verteilung die volle Information, d.h. sie beschreibe das Wissen über den unbekannten Parameter vollständig.
Alle statistischen Analysen mögen sich ausschließlich auf die Posteriori zu stützen; darauf baut insbesondere auch die Konstruktion von

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

- Bayesschen-Intervallschätzungen: *HPD-Intervalle (Highest posterior density-Intervalle)*

- Bayes-Tests

Proposition 5.6 (Suffizienz und Posteriori-Verteilung)/

Ist in der Situation von Bemerkung 5.4 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 [16]

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

Beweis:
Gemäß (5.7) 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.

[16] $\propto$ „proportional zu“, vgl. Bemerkung 5.3

Def. 5.7 (Vorbereitende 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 p_ϑ ist für jedes $\vartheta \in \Theta$ 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 $f_\vartheta(x)$ 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

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

- 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.)

5.1.1 Konjugierte Verteilungen, Bayes-Lernen

Wir betrachten hier, unter Rückbezug auf das Bayes-Postulat (vgl. Bem. 5.5), das sukzessive Lernen aus Stichproben.

a) Ein Motivationsbeispiel

Bsp. 5.8 (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 $\pi(\cdot)$

$$\pi(\vartheta) = \frac{\vartheta^{a-1}(1-\vartheta)^{b-1}}{B(a,b)} \cdot I_{[0;1]}(\vartheta) \tag{5.10}$$

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

Es gilt:

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

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

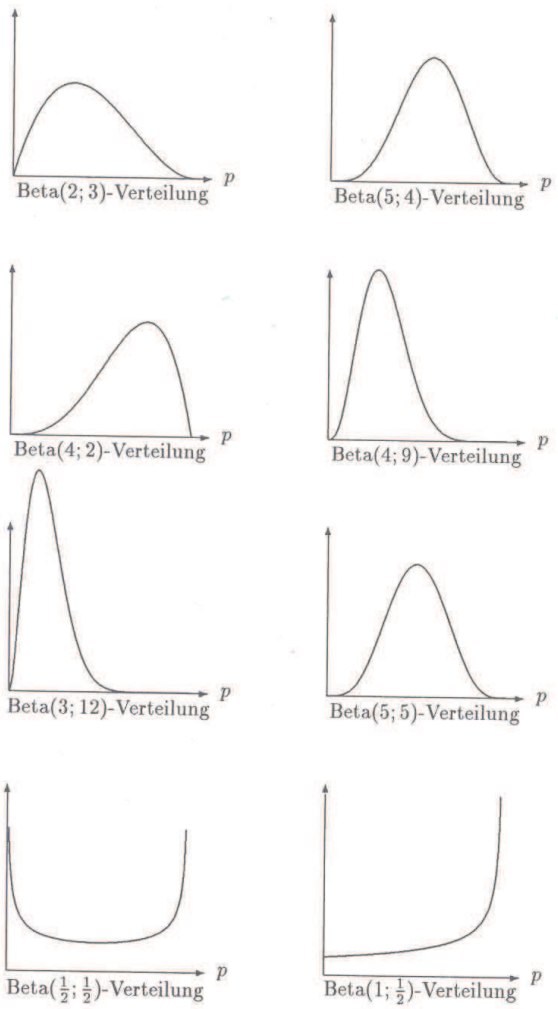


Abbildung 2.17: Einige Beta$(a; b)$-Verteilungen.
Die Beta$(1; 1)$-Verteilung ist die Gleichverteilung. Die an der Senkrechten im Punkt 0.5 gespiegelte Dichte einer Beta$(a; b)$-Verteilung ist die Beta$(b; a)$-Verteilung.

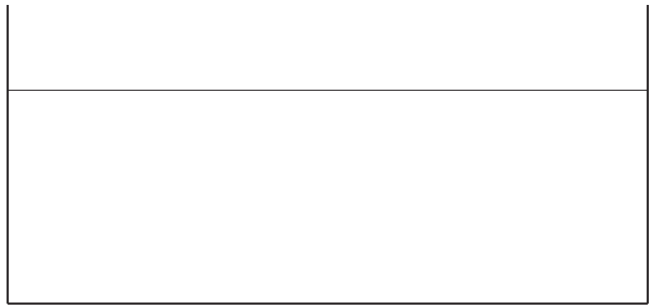
Jetzt Satz von Bayes anwenden: Posteriori nach einer Beobachtung berechnen.

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

- 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

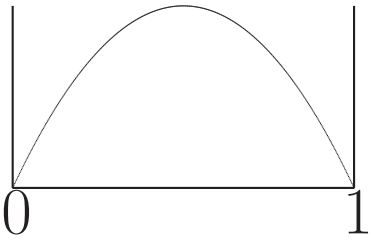
$\pi(\vartheta \mid x_1) \propto \vartheta I_{[0;1]}(\vartheta)$

- Jetzt weiteres Experiment:

  Updating-Prinzip: (neue) Priori = (alte Posteriori): Beta(2,1) Verteilung
  neue Stichprobe x_2
  neue Posteriori: $\text{Beta}(a^{(2)}, b^{(2)}) = \text{Beta}(a^{(1)} + x_i, b^{(1)} + (1 - x_i))$

  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: $\text{Beta}(a^{(3)}, b^{(3)}) = \text{Beta}(a^{(2)} + x_i, b^{(2)} + (1 - x_i))$
  $a^{(3)} = 2 + 1,\ b^{(3)} = 2$

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

- Weiter zusätzliche Beobachtung $\boxed{x_4 = 1}$:
  neue Posteriori: $\text{Beta}(a^{(4)}, b^{(4)}) = \text{Beta}(a^{(3)} + x_i, b^{(3)} + (1 - x_i))$
  $a^{(4)} = 3 + 1,\ b^{(4)} = 2$

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

- Allgemein gilt bei n unabhängigen Wiederholungen:

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

$$B\left(a^{(0)}+\sum_{i=1}^{n}x_i;b^{(0)}+n-\sum_{i=1}^{n}x_i\right) \text{ Verteilung.} \qquad (5.11)$$

  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 $a^{(0)} = b^{(0)} = 1$

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

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

b) Konjugiertheit: Definition und klassische Ergebnisse (vgl. „Schätzen und Testen I“)

Def. 5.9 (Konjugiertheit)

Eine Verteilungsfamilie Π von Priori-Verteilungen heißt zu einer Menge $\mathcal{P}$ von Stichprobenverteilungen *konjugiert*, wenn für jede Priori-Verteilung $\pi(\cdot) \in \Pi$ 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.

Proposition 5.10 (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. 5.8).
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

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

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

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

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

$$\pi(\lambda) = \frac{b^a}{\underbrace{\Gamma(a)}_{\text{Norm.konst.}}} \lambda^{a-1} e^{-b\lambda}, \tag{5.12}$$

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. 5.11 (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

$$\nu' = \frac{\bar{x}\rho^2 + \nu\dfrac{\sigma^2}{n}}{\rho^2 + \dfrac{\sigma^2}{n}} \tag{5.13}$$

und

$$\rho^{2'} = \frac{\rho^2 \cdot \dfrac{\sigma^2}{n}}{\rho^2 + \dfrac{\sigma^2}{n}} \, . \tag{5.14}$$

(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 5.12 (Zur Konjugiertheit in Exponentialfamilien)

Hat in der Situation von Def. 5.4 jedes Element der Menge $\mathcal{P}$ der Stichprobenverteilungen eine Dichte bzw. Wahrscheinlichkeitsfunktion $f(x|\vartheta)$ der Form

$$f(x|\vartheta) \propto h(\vartheta) \exp\big(T(x) \cdot b(\vartheta)\big) \tag{5.15}$$

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

$$\pi(\vartheta) \propto [h(\vartheta)]^{\alpha} \exp\big(b(\vartheta) \cdot \beta\big), \tag{5.16}$$

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

$$\pi(\vartheta|x) \propto [h(\vartheta)]^{\alpha+1} \cdot \exp\big((T(x) + \beta) \cdot b(\vartheta)\big). \tag{5.17}$$

Beweis:

(5.17) ergibt sich unmittelbar durch Anwenden der Formel für die Posteriori-Verteilung auf (5.15) und (5.16). Dann ist (5.17) mit $\alpha' := \alpha + 1$ und $\beta' := \beta + T(x)$ von der Form (5.16), also sind tatsächlich Π und $\mathcal{P}$ konjugiert.

Bem. 5.13 (zu Satz 5.12)

- Der Satz kann also direkt zur Konstruktion geeigneter, konjugierter Priori-Verteilungen verwendet werden, indem man die Stichprobenverteilung in die Form (5.15) bringt und dann eine Priori gemäß (5.16) wählt.
- $b(\vartheta)$ spielte in (5.15) und in (5.16) eine ganz unterschiedliche Rolle:
  In (5.15) ist $b(\vartheta)$ der natürliche Parameter der Exponentialfamilie, aus der die Likelihood / Stichprobenverteilung stammt.
  In (5.16) 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 $h(\vartheta)$.
- De facto „datiert man einfach mittels der suffizienten Statistik auf“ (vgl. 5.6).

Bsp. 5.14 (Beispiele zu Satz 5.12)

Man bestimme in folgenden Situationen unter Verwendung von Satz 5.12 jeweils eine konjugierte Priori-Verteilung:

a) $X_1, ...X_n$ ist i.i.d. normalverteilt mit unbekanntem μ und bekannter Varianz σ^2

b) X ist binomialverteilt zum unbekannten Parameter p

c) $X_1, ..., X_n$ ist i.i.d. Poisson-verteilt mit unbekanntem Parameter λ

5.1.2 (Reine) Bayes-Punktschätzung

Def. 5.15 (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. 5.16 (Zur MPD-Schätzung)

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

b) Ist der Zustandsraum Θ beschränkt 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. 5.17 (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.

5.1.3 Der Hauptsatz der Bayes-Entscheidungstheorie

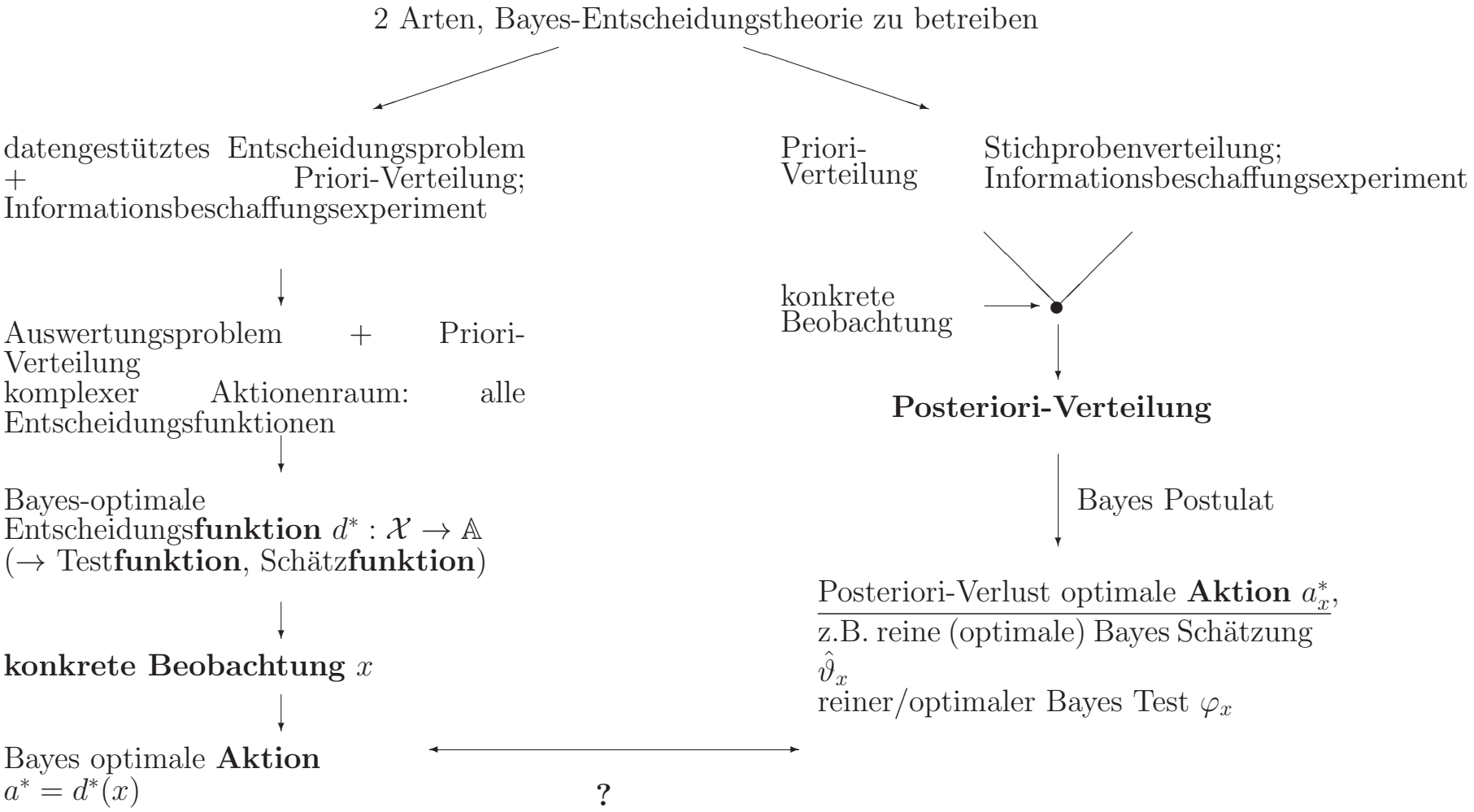
Def. 5.18 (Posteriori-Verlust-optimale Aktionen, konditionale Bayes-Aktionen)

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* $x \in \mathcal{X}$ oder *konditionale Bayes-Aktion zu* x *und der Priori-Verteilung* $\pi(\cdot)$, 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 Posteriori-Verteilung $\pi(\cdot|x)$ als „aufdatierter 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-optimale Entscheidungsfunktion $d^* : \mathcal{X} \to \mathbb{A}$
($\to$ Testfunktion, Schätzfunktion)
Bayes Postulat
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 5.19 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 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 $f(\vec{z})$ in $\vec{z} = (z_1, ..., z_n)$, wobei mit $c_i \geq 0\ ,\ i = 1, \ldots, n$ gilt: $f(\vec{z}) = \sum_{i=1}^{n} c_i f_i(z_i)$, 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 (vgl. Bem. 5.4) 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

$$\pi(\vartheta|x) = \frac{p(\{X = x\}|\{U = \vartheta\}) \cdot \pi(\vartheta)}{p(\{X = x\})}. \tag{5.18}$$

Nun betrachte man Entscheidungsfunktionen $d(\cdot)$ im Auswertungsproblem:
Für die Risikofunktion

$$R(d, \vartheta) = \mathbb{E}_{p_\vartheta}(\ell(d(x), \vartheta)$$

gilt hier

$$R(d, \vartheta) = \sum_{x \in \mathcal{X}} \ell(d(x), \vartheta) \cdot p_\vartheta(\{X = x\}).$$

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

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

löst also

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

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

$$\overset{(5.18)}{\iff} \underbrace{\sum_{x\in\mathcal{X}}}_{\hat{=}\sum_{i=1}^n}\underbrace{\Big(\sum_{\vartheta\in\Theta}\ell(d(x),\vartheta)\cdot\pi(\vartheta|x)\Big)}_{\hat{=}f_i(z_i)}\cdot \underbrace{p(\{X=x\})}_{\hat{=}c_i;\text{ priori-prädiktiv, marginal}} \to\min_d$$

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

Satz 5.20 (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 quadratische bzw. absolute Verlustfunktion, so gilt für die Bayes-optimale Entscheidungsfunktion $d^*_{quad}(\cdot)$ bzw. $d^*_{abs}(\cdot)$:
Für jedes x ist $d^*_{quad}(\cdot)$ genau der Erwartungswert und $d^*_{abs}(\cdot)$ der Median der Posteriori-Verteilung $\pi(\vartheta|x)$.

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

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

5.1.4 „Asymptotische Objektivität“ der konditionalen Bayes-Inferenz

Satz 5.21 („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 $\pi_n(\vartheta|x)$

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

Bem. 5.22 (Erneute kritische Diskussion des Bayes-Ansatzes)

5.2 (Robuste) Bayes-Inferenz aus entscheidungstheoretischer Sicht

5.2.1 Generalized Bayes Rule

Bem. 5.23 Zur Kritik des Ansatzes, Priori-Credal-Mengen

Natürlich hängt (5.11) 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 ist äusserst problematisch (vgl. Kap. 2.5.1).

- 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. 5.26

Bem. 5.24 Robuste Bayes-Analyse, Generalized Bayes Rule

In der Situation von Bem. 5.4 kann man auch mit *Priori-Credalmengen* $\mathcal{M}$ arbeiten. Dann heißt

$$\mathcal{M}_{\cdot|x} = \Big\{ \pi(\cdot|x) \;\Big|\; \exists \pi(\cdot) \in \mathcal{M} : \pi(\cdot|x) \text{ ist Posteriori-Verteilung} \tag{5.19}$$
$$\text{von } \vartheta \text{ gegeben } x \text{ bezüglich } \pi(\cdot) \Big\}$$

Posteriori-Credalmenge gegeben x bezüglich $\mathcal{M}$. Man spricht dann von *einer robusten Bayes-Analyse;* (5.20) wird dann oft als *Generalized Bayes Rule* (GBR) bezeichnet.

Bem. 5.25 (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

5.2.2 Konjugiertheit und verallgemeinerte Bayes-Inferenz

Bem. 5.26 (Eine alternative Darstellung von Satz 5.12)

Eine zum Nachweis meist umständlichere, aber für die Interpretation oft anschaulichere und für die Verallgemeinerung besser geeignete, alternative Darstellung von Satz 5.12 lautetet:

Hat in der Situation von Def. 5.4 jedes Element der Menge $\mathcal{P}$ der Stichprobenverteilungen eine Dichte bzw. Wahrscheinlichkeitsfunktion $f(x|\vartheta)$ der Form

$$f(x|\vartheta) \propto \exp\Big(\psi(\vartheta)\tau(\vec{x}) - n \cdot d(\vartheta)\Big) \tag{5.20}$$

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

$$\pi(\vartheta) \propto \exp\Big(n^{(0)}\Big(\psi(\vartheta)y^{(0)} - d(\vartheta)\Big)\Big) \tag{5.21}$$

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

$$\pi(x|\vartheta) \propto \exp\left(n^{(1)}\left(\psi(\vartheta)y^{(1)} - d(\vartheta)\right)\right) \tag{5.22}$$

mit

$$n^{(1)} = n^{(0)} + n \tag{5.23}$$

und

$$y^{(1)} = \frac{n^{(0)}y^{(0)} + \tau(x)}{n^{(0)} + n}. \tag{5.24}$$

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

$n^{(0)}$ tritt immer gemeinsam mit n auf, nämlich im Ausdruck $n^{(0)} + n$

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

$$y^{(1)} = \frac{n^{(0)}}{n^{(0)} + n}y^{(0)} + \frac{n}{n^{(0)} + n}\bar{\tau}(\vec{x}) \tag{5.25}$$

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

Beispielsweise ist dann die Beta-priori aus (5.10)

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

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

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

.

Bem. 5.27 (Robuste Bayes-Analyse in konjugierten Modellen)

- Die Darstellung in Bem. 5.26 ermöglicht eine elegante robuste Bayes-Analyse. Priori-Credalmengen erzeugt man durch intervallwertige Priori-Parameter:[18]

  $\left[\underline{y}^{(0)}, \bar{y}^{(0)}\right]$ 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

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

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)

Bem. 5.28 (Fortsetzung von Beispiel 5.11)

Man betrachte die Bayes-Inferenz für den Mittelwert einer Normalverteilung aus einer i.i.d. Stichprobe. Geht man im Sinne von Bem. 5.27 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[19]

[19] 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

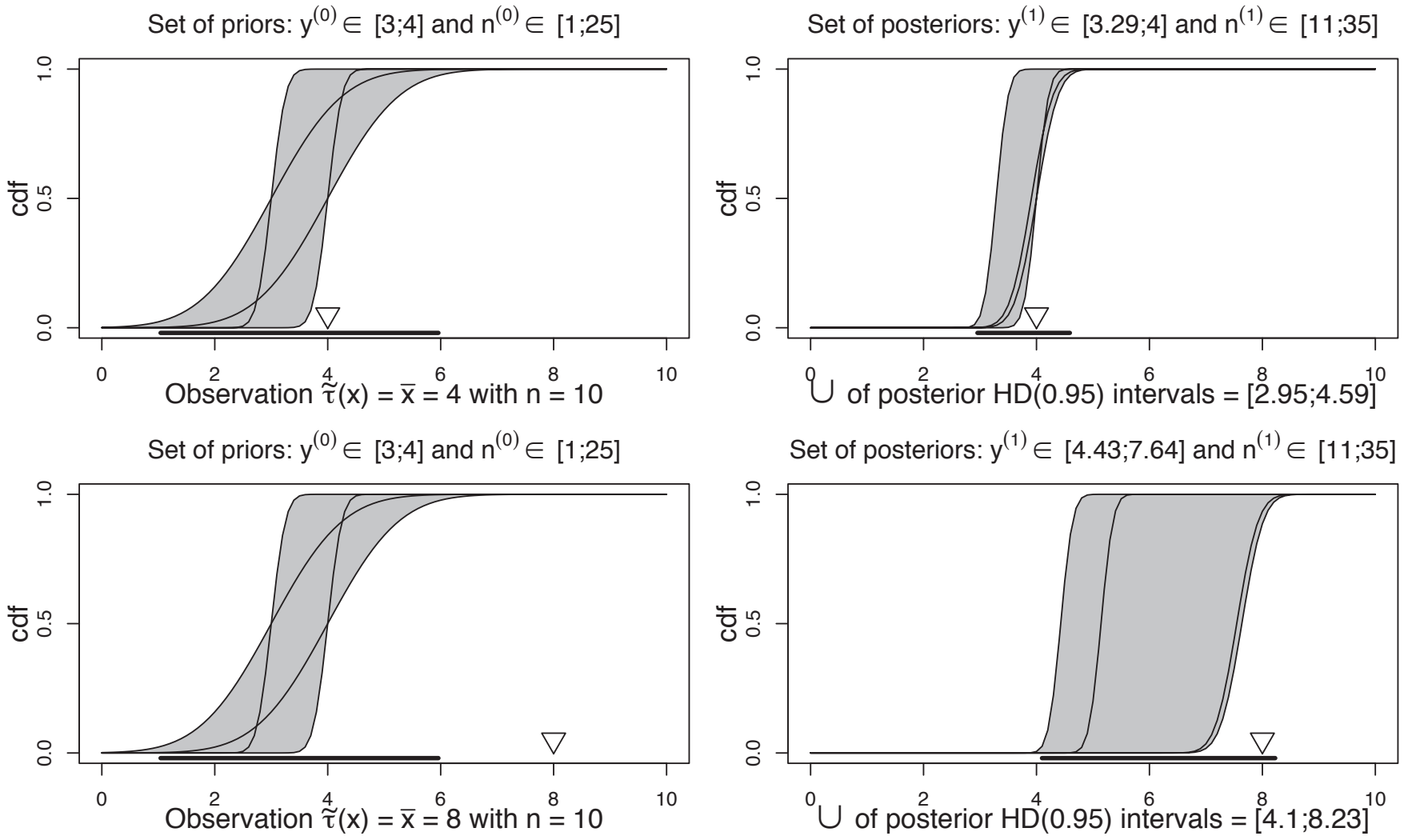
$$\mathrm{pdc}\left(\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i}; \underline{\nu}^{(0)}, \bar{\nu}^{(0)}\right) := \inf\left\{\left|\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i} - \nu^{(0)}\right| \;\middle|\; \underline{\nu}^{(0)} \leq \nu^{(0)} \leq \bar{\nu}^{(0)}\right\}.$$

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.

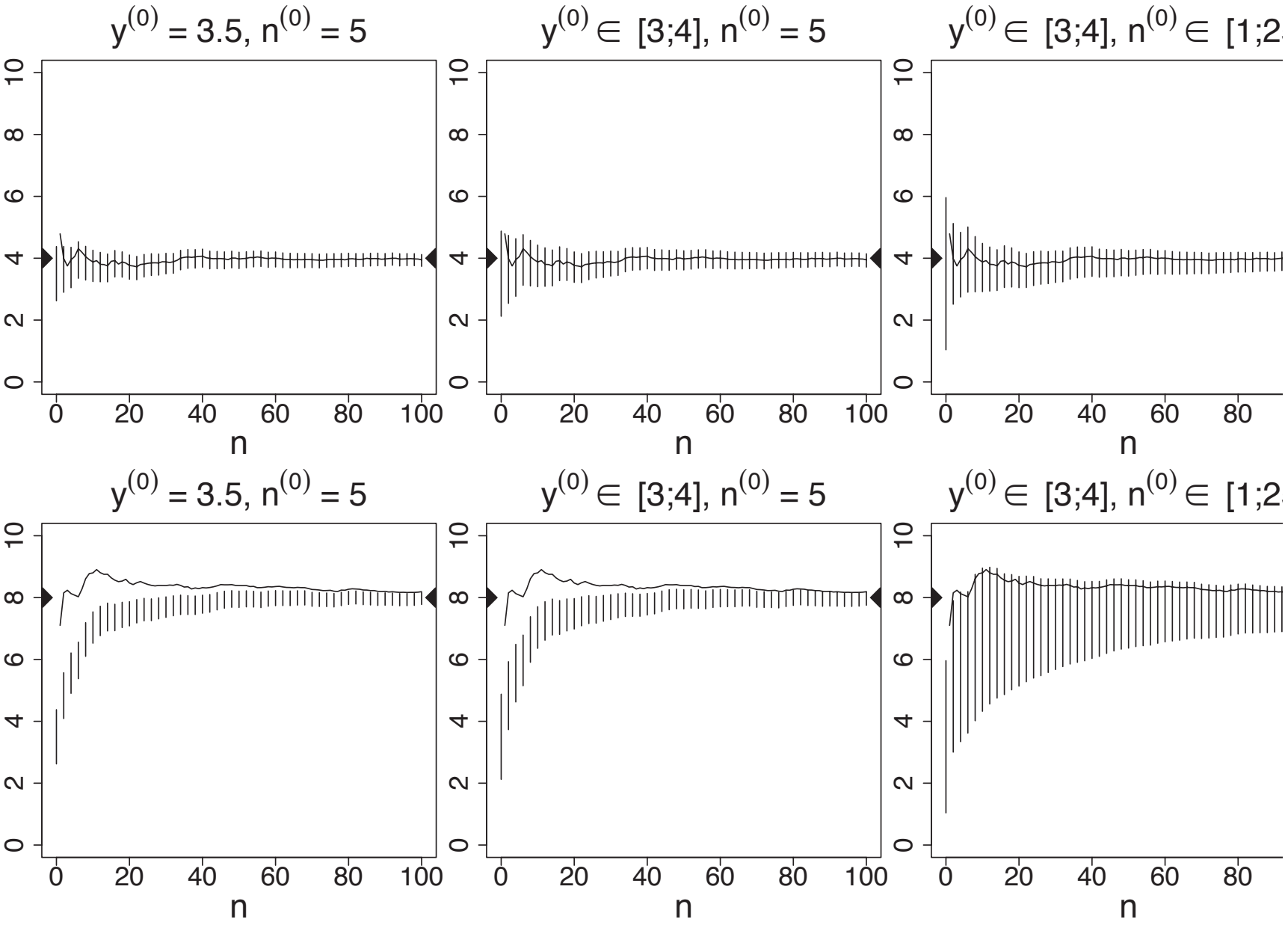
Quelle: Walter & Augustin (2009, S. 268)



Die Bilder zeigen die Inferenz mit einer Menge von Normalverteilungen (angedeutet durch die Verteilungsfunktionen) gebildeten Priori-Credalmenge (jeweils links) und die zugehörigen Posteriori Credalmengen, einmal im Fall ohne Priori-Daten-Konflikt (oben) und mit Priori-Daten-Konflikt (unten). In der Tat ist im zweiten Fall die Credalmenge sehr „gross“; man kann in dieser widersprüchlichen Situation kaum weitergehende Schlüsse tätigen, während die obere Aussage relativ präzise Aussagen ermöglicht.

Die Abbildungen zeigen das arithmetische Mittel (schwarze Linie) und HPD-Intervalle zum Sicherheitsgrad 95 Prozent bei der konjugierten Inferenz über den Mittelwert einer Normalverteilung für verschiedene Stichprobenumfänge in einer typischen Situation ohne Priori-Daten-Konflikt (oben) und mit Priori-Daten-Konflikt (unten). Links ist die klassische Modelierung basierend auf einer Priori-Normalverteilung, rechts eine geeignet konstruierte Credalmenge. Man sieht, dass die HPD Schätzungen hier deutlich breiter sind als im Fall ohne Priori-Daten Konflikt, während im klassischen Fall die Intervalle gleich lang bleiben und nur verschoben werden.

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



Korollar 5.29 (Korollar zu Satz 5.21: Konsistenzsatz für Credal-Bayes-Verfahren)

In der Situation von Satz 5.21 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:

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

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