

Mathematische Statistik
Skript zur Vorlesung
im Sommersemester 2026

Markus Reiß
Humboldt-Universität zu Berlin
mreiss@math.hu-berlin.de

VORLÄUFIGE FASSUNG: 5. Juni 2026

Inhaltsverzeichnis

1	Entscheidungstheorie	1
1.1	Formalisierung eines statistischen Problems	1
1.2	Minimax- und Bayes-Ansatz	3
1.3	Das Stein-Phänomen	10
1.4	Ergänzungen*	12
2	Dominierte Modelle und Suffizienz	14
2.1	Dominierte Modelle	14
2.2	Exponentialfamilien	15
2.3	Suffizienz	17
2.4	Vollständigkeit	21
2.5	Cramér-Rao-Schranke	23
2.6	Äquivarianz	30
3	Asymptotische Schätztheorie	33
3.1	Momentenschätzer	33
3.2	Maximum-Likelihood- und M-Schätzer	37
3.3	Asymptotik	41

1 Entscheidungstheorie

1.1 Formalisierung eines statistischen Problems

1.1 Definition. Ein Messraum $(\mathcal{X}, \mathcal{F})$ versehen mit einer Familie $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ von Wahrscheinlichkeitsmaßen, $\Theta \neq \emptyset$ beliebige Parametermenge, heißt statistisches Experiment oder statistisches Modell. \mathcal{X} heißt Stichprobenraum. Jede $(\mathcal{F}, \mathcal{S})$ -messbare Funktion $Y : \mathcal{X} \rightarrow S$ heißt Beobachtung oder Statistik mit Werten in (S, \mathcal{S}) und induziert das statistische Modell $(S, \mathcal{S}, (\mathbb{P}_\vartheta^Y)_{\vartheta \in \Theta})$. Sind die Beobachtungen Y_1, \dots, Y_n für jedes \mathbb{P}_ϑ unabhängig und identisch verteilt, so nennt man Y_1, \dots, Y_n eine mathematische Stichprobe.

1.2 Definition. Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell. Eine Entscheidungsregel ist eine messbare Abbildung $\rho : \mathcal{X} \rightarrow A$, wobei der Messraum (A, \mathcal{A}) der sogenannte Aktionsraum ist. Jede Funktion $l : \Theta \times A \rightarrow [0, \infty) =: \mathbb{R}^+$, die messbar im zweiten Argument ist, heißt Verlustfunktion. Das Risiko einer Entscheidungsregel ρ bei Vorliegen des Parameters $\vartheta \in \Theta$ ist

$$R(\vartheta, \rho) := \mathbb{E}_\vartheta[l(\vartheta, \rho)] = \int_{\mathcal{X}} l(\vartheta, \rho(x)) \mathbb{P}_\vartheta(dx).$$

1.3 Beispiele.

(a) Wir formalisieren das Beobachtungsmodell

$$Y_i = \mu + \varepsilon_i, \quad i = 1, \dots, n,$$

mit unabhängigen Fehlervariablen $\varepsilon_1, \dots, \varepsilon_n \sim N(0, \sigma^2)$. Dann ist der Beobachtungsvektor $Y = (Y_1, \dots, Y_n)^\top \sim N(\mu \mathbf{1}_n, \sigma^2 E_n)$ -verteilt mit $\mathbf{1}_n = (1, \dots, 1)^\top \in \mathbb{R}^n$ und n -dimensionaler Einheitsmatrix E_n . Als statistisches Modell wählen wir daher $(\mathbb{R}^n, \mathfrak{B}_{\mathbb{R}^n}, (N(\mu \mathbf{1}_n, \sigma^2 E_n))_{\mu \in \mathbb{R}, \sigma > 0})$. Die Parametermenge ist $\Theta = \mathbb{R} \times (0, \infty)$ mit Parametern $\vartheta = (\mu, \sigma)$. Alternativ können wir sagen, dass Y_1, \dots, Y_n eine $N(\mu, \sigma^2)$ -verteilte mathematische Stichprobe ist.

Um das Stichprobenmittel $\bar{Y} := \rho(Y_1, \dots, Y_n) := \frac{1}{n} \sum_{i=1}^n Y_i$ als Entscheidungsregel zu interpretieren und seine Güte bei der Schätzung von μ zu messen, betrachtet man den Aktionsraum $A = \mathbb{R}$ und beispielsweise die quadratische Verlustfunktion $l(\vartheta, a) = l((\mu, \sigma), a) = (\mu - a)^2$. Beim Verlust ist σ irrelevant; da aber die Verteilung \mathbb{P}_ϑ von σ abhängt, spricht man von einem Störparameter. Das quadratische Risiko (auch MSE: mean squared error) ist $R((\mu, \sigma), \rho) = \mathbb{E}_{\mu, \sigma}[(\mu - \bar{Y})^2] = \sigma^2 n^{-1}$, da ja $\bar{Y} - \mu \sim N(0, \sigma^2 n^{-1})$.

(b*) Allgemeiner können wir das Beobachtungsmodell

$$Y_i = \mu + \varepsilon_i, \quad i = 1, \dots, n,$$

mit zentrierten und unkorrelierten Fehlervariablen $\varepsilon_1, \dots, \varepsilon_n$ betrachten. Ist die Art der Verteilung der (ε_i) unbekannt, sollte man auf dem Stichprobenraum $(\mathbb{R}^n, \mathfrak{B}_{\mathbb{R}^n})$ die Familie $\mathcal{P} = \{\mathbb{P} \text{ W-Maß auf } \mathfrak{B}_{\mathbb{R}^n} \mid \int_{\mathbb{R}^n} x \mathbb{P}(dx) =$

$\mu \mathbf{1}_n, \int_{\mathbb{R}^n} (x - \mu \mathbf{1}_n)(x - \mu \mathbf{1}_n)^\top \mathbb{P}(dx) = \sigma^2 E_n, \mu \in \mathbb{R}, \sigma > 0\}$ betrachten. In dieser Betrachtungsweise bleibt von einem unendlich-dimensionalen Parameterraum \mathcal{P} maximal ein zweidimensionaler interessierender Parameter $\vartheta = (\mu, \sigma)$ übrig. Interessanterweise ändert sich das quadratische Risiko des Stichprobenmittels in diesem allgemeineren Modell nicht.

(c*) Im Gaußschen multivariaten linearen Modell beobachten wir

$$Y_i = \langle x_i, \beta \rangle + \varepsilon_i, \quad i = 1, \dots, n,$$

mit gegebenen Kovariablen $x_1, \dots, x_n \in \mathbb{R}^p$, interessierendem Parameter $\beta \in \mathbb{R}^p$ und $\varepsilon_1, \dots, \varepsilon_n \sim N(0, \sigma^2)$ unabhängig. Als statistisches Modell ergibt sich $(\mathbb{R}^n, \mathfrak{B}_{\mathbb{R}^n}, (\otimes_{i=1}^n N(\langle x_i, \beta \rangle, \sigma^2))_{\beta \in \mathbb{R}^p, \sigma > 0})$. Mit der Designmatrix $X = (x_1^\top, \dots, x_n^\top)^\top \in \mathbb{R}^{n \times p}$ gilt äquivalent $\otimes_{i=1}^n N(\langle x_i, \beta \rangle, \sigma^2) = N(X\beta, \sigma^2 E_n)$. Der Kleinste-Quadrate-Schätzer ist $\hat{\beta} = (X^\top X)^{-1} X^\top Y$, sofern X Rang p besitzt (x_1, \dots, x_n spannen den \mathbb{R}^p auf). Mit Aktionsraum $A = \mathbb{R}^p$ und quadratischem Verlust $l((\beta, \sigma), a) = |\beta - a|^2$ (mit Euklidischer Norm $|\bullet|$) erhalten wir das quadratische Risiko des Kleinste-Quadrate-Schätzers

$$R((\beta, \sigma), \hat{\beta}) = \mathbb{E}_{\beta, \sigma}[|\beta - \hat{\beta}|^2] = \mathbb{E}[|(X^\top X)^{-1} X^\top \varepsilon|^2] = \sigma^2 \text{trace}((X^\top X)^{-1})$$

mit der Spur $\text{trace}(M) := \sum_i M_{i,i}^2$.

- (d) Für einen Test auf Wirksamkeit eines neuen Medikaments werden 100 Versuchspersonen mit diesem behandelt. Unter der (stark vereinfachenden) Annahme, dass alle Personen identisch und unabhängig auf das Medikament reagieren, wird in Abhängigkeit von der Anzahl N der erfolgreichen Behandlungen entschieden, ob die Erfolgsquote höher ist als diejenige einer klassischen Behandlung. Als Stichprobenraum wähle $\mathcal{X} = \{0, 1, \dots, 100\}$ mit der Potenzmenge $\mathcal{P}(\mathcal{X})$ als σ -Algebra und $\mathbb{P}_p = \text{Bin}(100, p), p \in \Theta = [0, 1]$, als mögliche Verteilungen. Die Nullhypothese ist $H_0 : p \leq p_0$ für den unbekanntem Parameter p . Als Aktionsraum dient $A = \{0, 1\}$ (H_0 annehmen bzw. verwerfen), und wir wählen den Verlust $l(p, a) = \ell_0 \mathbf{1}(p \leq p_0, a = 1) + \ell_1 \mathbf{1}(p > p_0, a = 0)$ mit Konstanten $\ell_0, \ell_1 \geq 0$. Dies führt auf das Risiko einer Entscheidungsregel (eines Tests)

$$R(p, \rho) = \begin{cases} \ell_0 \mathbb{P}_p(\rho = 1), & p \leq p_0 \\ \ell_1 \mathbb{P}_p(\rho = 0), & p > p_0 \end{cases}$$

und die Fehlerwahrscheinlichkeit erster Art wird mit ℓ_0 , die zweiter Art mit ℓ_1 gewichtet.

1.4 Definition. Die Entscheidungsregel ρ heißt besser als eine Entscheidungsregel ρ' , falls $R(\vartheta, \rho) \leq R(\vartheta, \rho')$ für alle $\vartheta \in \Theta$ gilt und falls ein $\vartheta_0 \in \Theta$ mit $R(\vartheta_0, \rho) < R(\vartheta_0, \rho')$ existiert. Eine Entscheidungsregel heißt zulässig, wenn es keine bessere Entscheidungsregel gibt.

1.5 Bemerkung. Häufig wird die Menge der betrachteten Entscheidungsregeln eingeschränkt. Bei Schätzern wird beispielsweise Erwartungstreue, Linearität

oder allgemeiner Invarianz bezüglich gewisser Gruppenoperationen (vergleiche *Äquivarianz* weiter unten) gefordert. So ist der Kleinste-Quadrate-Schätzer im linearen Modell nach dem Satz von Gauß-Markov zulässig unter quadratischem Verlust in der Klasse der erwartungstreuen und linearen Schätzern.

1.6 Beispiel. Es sei Y_1, \dots, Y_n eine $N(\vartheta, 1)$ -verteilte mathematische Stichprobe mit $\vartheta \in \mathbb{R}$. Betrachte $\hat{\vartheta}_1 = \bar{Y}$, $\hat{\vartheta}_2 = \bar{Y} + 0.5$, $\hat{\vartheta}_3 = 6$ unter quadratischem Verlust $l(\vartheta, a) = (\vartheta - a)^2$. Wegen $R(\vartheta, \hat{\vartheta}_1) = 1/n$, $R(\vartheta, \hat{\vartheta}_2) = 0.25 + 1/n$ ist $\hat{\vartheta}_1$ besser als $\hat{\vartheta}_2$, allerdings ist weder $\hat{\vartheta}_1$ besser als $\hat{\vartheta}_3$ noch umgekehrt. In der Tat ist $\hat{\vartheta}_3$ zulässig, weil $R(\vartheta, \hat{\vartheta}_3) = 0$ für $\vartheta = 6$ gilt und jeder Schätzer mit dieser Eigenschaft Lebesgue-fast überall mit $\hat{\vartheta}_3$ übereinstimmt. Später werden wir sehen, dass auch $\hat{\vartheta}_1$ zulässig ist.

1.2 Minimax- und Bayes-Ansatz

1.7 Bemerkung. Da das Risiko $R(\vartheta, \rho)$ einer Entscheidungsregel ρ im Allgemeinen vom unbekanntem wahren Parameter ϑ abhängt, werden Entscheidungsregeln üblicherweise gemäß ihrem maximalen Risiko in ϑ oder einem geeignet über ϑ gemittelten Risiko beurteilt.

1.8 Definition. Eine Entscheidungsregel ρ heißt minimax, falls

$$\sup_{\vartheta \in \Theta} R(\vartheta, \rho) = \inf_{\rho'} \sup_{\vartheta \in \Theta} R(\vartheta, \rho'),$$

wobei sich das Infimum über alle Entscheidungsregeln ρ' erstreckt.

1.9 Definition. Der Parameterraum Θ trage die σ -Algebra \mathcal{F}_Θ , die Verlustfunktion l sei produktmessbar und $\vartheta \mapsto \mathbb{P}_\vartheta(B)$ sei messbar für alle $B \in \mathcal{F}$. Die a-priori-Verteilung π des Parameters ϑ ist gegeben durch ein Wahrscheinlichkeitsmaß auf $(\Theta, \mathcal{F}_\Theta)$. Das zu π assoziierte Bayesrisiko einer Entscheidungsregel ρ ist

$$R_\pi(\rho) := \mathbb{E}_\pi[R(T, \rho)] = \int_{\Theta} R(\vartheta, \rho) \pi(d\vartheta) = \int_{\Theta} \int_{\mathcal{X}} l(\vartheta, \rho(x)) \mathbb{P}_\vartheta(dx) \pi(d\vartheta).$$

ρ heißt Bayesregel oder Bayes-optimal (bezüglich π), falls

$$R_\pi(\rho) = \inf_{\rho'} R_\pi(\rho')$$

gilt, wobei sich das Infimum über alle Entscheidungsregeln ρ' erstreckt.

1.10 Definition. Es sei X eine (S, \mathcal{S}) -wertige Zufallsvariable auf $(\Omega, \mathcal{F}, \mathbb{P})$. Eine Abbildung $K : S \times \mathcal{F} \rightarrow [0, 1]$ heißt reguläre bedingte Wahrscheinlichkeit oder Markovkern bezüglich X , falls

- (a) $A \mapsto K(x, A)$ ist Wahrscheinlichkeitsmaß für alle $x \in S$;
- (b) $x \mapsto K(x, A)$ ist messbar für alle $A \in \mathcal{F}$;
- (c) $K(X, A) = \mathbb{P}(A | X) := \mathbb{E}[\mathbf{1}_A | X]$ \mathbb{P} -f.s. für alle $A \in \mathcal{F}$.

1.11 Bemerkung. Statt $K(x, A)$ schreiben wir im folgenden meist $\mathbb{P}(A | X = x)$ und nehmen dann Eigenschaften (a), (b) implizit an. Eigenschaft (c) besagt gerade, dass $\mathbb{E}[\mathbb{P}(A | X)\mathbf{1}(X \in B)] = \mathbb{P}(A \cap \{X \in B\})$ gilt für alle messbaren Mengen A und B .

1.12 Satz. *Es sei (Ω, d) ein vollständiger, separabler Raum mit Metrik d und Borel- σ -Algebra \mathcal{F} (polnischer Raum). Für jede Zufallsvariable X auf $(\Omega, \mathcal{F}, \mathbb{P})$ existiert eine reguläre bedingte Wahrscheinlichkeit K bezüglich X . K ist \mathbb{P} -f.s. eindeutig bestimmt, d.h. für eine zweite solche reguläre bedingte Wahrscheinlichkeit K' gilt $\mathbb{P}(\forall A \in \mathcal{F} : K(X, A) = K'(X, A)) = 1$.*

Beweis. Siehe z.B. Gänsler, Stute (1977): Wahrscheinlichkeitstheorie, Springer. \square

1.13 Bemerkung. Während eine Minimaxregel den maximal zu erwartenden Verlust minimiert, kann das Bayesrisiko als ein (mittels π) gewichtetes Mittel der zu erwartenden Verluste angesehen werden. Alternativ wird π als die subjektive Einschätzung der Verteilung des zugrundeliegenden Parameters interpretiert. Daher wird das Bayesrisiko auch als insgesamt zu erwartender Verlust in folgendem Sinne verstanden.

1.14 Definition. Definiere $\Omega := \mathcal{X} \times \Theta$ und $\tilde{\mathbb{P}}$ auf $(\Omega, \mathcal{F} \otimes \mathcal{F}_\Theta)$ gemäß

$$\tilde{\mathbb{P}}(A \times B) := \iint \mathbf{1}_{A \times B}(x, \vartheta) \mathbb{P}_\vartheta(dx) \pi(d\vartheta) = \int_B \mathbb{P}_\vartheta(A) \pi(d\vartheta), \quad A \in \mathcal{F}, B \in \mathcal{F}_\Theta,$$

und Fortsetzung auf $\mathcal{F} \otimes \mathcal{F}_\Theta$ (gemeinsame Verteilung von Beobachtung und Parameter), wobei π eine a-priori-Verteilung auf \mathcal{F}_Θ und $(\vartheta, A) \mapsto \mathbb{P}_\vartheta(A)$ ein Markovkern sei. Bezeichne mit X und T die Koordinatenprojektionen von Ω auf \mathcal{X} bzw. Θ . Dann gilt $R_\pi(\rho) = \mathbb{E}_{\tilde{\mathbb{P}}}[l(T, \rho(X))]$.

Die Verteilung von T unter der regulären bedingten Wahrscheinlichkeit $\tilde{\mathbb{P}}(\bullet | X = x)$ von $\tilde{\mathbb{P}}$ heißt a-posteriori-Verteilung des Parameters gegeben die Beobachtung $X = x$.

1.15 Bemerkung. Im Gegensatz zum wahren Parameter ϑ und der zugehörigen Verteilung \mathbb{P}_ϑ ist $\tilde{\mathbb{P}}$ durch das Modell bekannt und damit auch die a-posteriori-Verteilung gegeben die Beobachtung von X .

1.16 Satz. (Bayesformel) *Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell sowie π eine a-priori-Verteilung auf $(\Theta, \mathcal{F}_\Theta)$, so dass \mathbb{P}_ϑ für alle $\vartheta \in \Theta$ μ -Dichten $f^{X|T=\vartheta}$ sowie π eine ν -Dichte f^T besitzt mit entsprechenden Maßen μ und ν . Ist $f^{X|T=\bullet} : \mathcal{X} \times \Theta \rightarrow \mathbb{R}^+$ ($\mathcal{F} \otimes \mathcal{F}_\Theta$)-messbar, so besitzt die a-posteriori-Verteilung $\mathbb{P}^{T|X=x}$ des Parameters für $\tilde{\mathbb{P}}^X$ -fast alle $x \in \mathcal{X}$ eine ν -Dichte, nämlich*

$$f^{T|X=x}(\vartheta) = \frac{f^{X|T=\vartheta}(x) f^T(\vartheta)}{f^X(x)} \quad \text{mit } f^X(x) := \int_{\Theta} f^{X|T=\vartheta'}(x) f^T(\vartheta') \nu(d\vartheta').$$

Beweis. Übung! \square

1.17 Beispiele.

- (a) Für einen Bayestest (oder auch ein Bayes-Klassifikationsproblem) setze $\Theta = \{0, 1\}$ und betrachte eine a-priori-Verteilung π mit $\pi(\{0\}) =: \pi_0$, $\pi(\{1\}) =: \pi_1$. Die Wahrscheinlichkeitsmaße $\mathbb{P}_0, \mathbb{P}_1$ auf $(\mathcal{X}, \mathcal{F})$ mögen die Dichten p_0, p_1 bezüglich einem Maß μ besitzen (z.B. $\mu = \mathbb{P}_0 + \mathbb{P}_1$). Nach der Bayesformel (mit Zählmaß ν) erhalten wir die a-posteriori-Verteilung

$$\tilde{\mathbb{P}}(T = i | X = x) = \frac{\pi_i p_i(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)}, \quad i = 0, 1 \quad \tilde{\mathbb{P}}^X\text{-f.ü.}$$

- (b) Es sei X_1, \dots, X_n eine $N(\mu, E_d)$ -verteilte mathematische Stichprobe im \mathbb{R}^d und $\pi = N(a, \sigma^2 E_d)$ eine a-priori-Verteilung für $\mu \in \mathbb{R}^d$ mit $a \in \mathbb{R}^d$, $\sigma > 0$. Dann liefert die Bayesformel bezüglich Lebesguemaß und mit $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$:

$$\begin{aligned} f^{T|X=x}(\mu) &\propto f^{X|T=\mu}(x) f^T(\mu) \\ &\propto \exp\left(-\frac{|\mu - a|^2}{2\sigma^2} - \frac{1}{2} \sum_{i=1}^n |x_i - \mu|^2\right) \\ &\propto \exp\left(\langle \mu, \sigma^{-2}a + n\bar{x} \rangle - \frac{1}{2} |\mu|^2 (\sigma^{-2} + n)\right) \\ &\propto \exp\left(-\frac{1}{2} (\sigma^{-2} + n) \left|\mu - \frac{a + n\sigma^2 \bar{x}}{1 + n\sigma^2}\right|^2\right). \end{aligned}$$

Die a-posteriori-Verteilung ist also wiederum eine Normalverteilung:

$$\tilde{\mathbb{P}}^{T|X=x} = N\left(\frac{a + n\sigma^2 \bar{x}}{1 + n\sigma^2}, \frac{\sigma^2}{1 + n\sigma^2} E_d\right).$$

Beachte, dass für großen Stichprobenumfang n oder große a-priori-Varianz σ^2 sich die a-posteriori-Verteilung um das Stichprobenmittel konzentriert, während sie für sehr kleine a-priori-Varianz und geringen Stichprobenumfang, so dass $n\sigma^2 \ll 1$, nahe bei der a-priori-Verteilung bleibt.

1.18 Satz. *Eine Regel ρ ist Bayes-optimal, falls gilt*

$$\rho(X) \in \operatorname{argmin}_{a \in A} \mathbb{E}_{\tilde{\mathbb{P}}} [l(T, a) | X] \quad \tilde{\mathbb{P}}\text{-f.s.},$$

d.h. $\mathbb{E}_{\tilde{\mathbb{P}}} [l(T, \rho(x)) | X = x] \leq \mathbb{E}_{\tilde{\mathbb{P}}} [l(T, a) | X = x]$ für alle $a \in A$ und $\tilde{\mathbb{P}}^X$ -fast alle $x \in \mathcal{X}$.

Beweis. Für eine beliebige Entscheidungsregel ρ' gilt

$$R_\pi(\rho') = \mathbb{E}_{\tilde{\mathbb{P}}} [\mathbb{E}_{\tilde{\mathbb{P}}} [l(T, \rho'(X)) | X]] \geq \mathbb{E}_{\tilde{\mathbb{P}}} [\mathbb{E}_{\tilde{\mathbb{P}}} [l(T, \rho(X)) | X]] = R_\pi(\rho).$$

□

1.19 Satz. *Für $\Theta \subseteq \mathbb{R}^d$, $A = \mathbb{R}^d$ und quadratisches Risiko (d.h. $l(\vartheta, a) = |a - \vartheta|^2$) ist die (vektorwertige) bedingte Erwartung $\hat{\vartheta}_\pi := \mathbb{E}_{\tilde{\mathbb{P}}} [T | X]$ Bayes-optimaler Schätzer von ϑ bezüglich der a-priori-Verteilung π , sofern $T \in L^2(\tilde{\mathbb{P}})$ gilt.*

Beweis. Dies folgt aus der L^2 -Projektionseigenschaft der bedingten Erwartung

$$\mathbb{E}_{\tilde{\mathbb{P}}}[|\mathbb{E}_{\tilde{\mathbb{P}}}[T | X] - T|^2] = \inf_{g: \mathcal{X} \rightarrow \mathbb{R}^d \text{ messbar}} \mathbb{E}_{\tilde{\mathbb{P}}} [|g(X) - T|^2],$$

vgl. Stochastik II (Fall $d > 1$ folgt analog), via

$$R_\pi(\mathbb{E}_{\tilde{\mathbb{P}}}[T | X]) = \mathbb{E}_{\tilde{\mathbb{P}}} [|\mathbb{E}_{\tilde{\mathbb{P}}}[T | X] - T|^2] \leq \mathbb{E}_{\tilde{\mathbb{P}}} [|\rho(X) - T|^2] = R_\pi(\rho)$$

für alle Entscheidungsregeln ρ . □

1.20 Bemerkung. Für $d = 1$ und den Absolutbetrag $l(\vartheta, a) = |\vartheta - a|$ ist jeder a-posteriori-Median $\hat{\vartheta}_\pi$, d.h. $\tilde{\mathbb{P}}(T \leq \hat{\vartheta}_\pi | X) \geq 1/2$ und $\tilde{\mathbb{P}}(T \geq \hat{\vartheta}_\pi | X) \geq 1/2$, Bayes-optimaler Schätzer. Dies folgt aus der L^1 -Minimierung des Medians.

1.21 Beispiele. (Fortsetzung)

- (a) Nach Beispiel 1.17(a) und Satz 1.18 finden wir einen Bayestest $\varphi(x)$ für den 0-1-Verlust $l(\vartheta, a) = \mathbf{1}(a \neq \vartheta)$ als Minimalstelle von

$$a \mapsto \mathbb{E}_{\tilde{\mathbb{P}}}[l(T, a) | X = x] = \frac{\pi_0 p_0(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)} a + \frac{\pi_1 p_1(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)} (1 - a).$$

Daher ist ein Bayestest (Bayesklassifizierer) gegeben durch

$$\varphi(x) = \begin{cases} 0, & \pi_0 p_0(x) > \pi_1 p_1(x) \\ 1, & \pi_1 p_1(x) > \pi_0 p_0(x) \\ \text{beliebig,} & \pi_0 p_0(x) = \pi_1 p_1(x) \end{cases}$$

und wir entscheiden uns für dasjenige $\vartheta \in \{0, 1\}$, dessen a-posteriori-Wahrscheinlichkeit am größten ist (“MAP-estimator: maximum a posteriori estimator“). Für später sei bereits auf die Neyman-Pearson-Struktur von φ in Abhängigkeit von $p_1(x)/p_0(x)$ hingewiesen.

- (b) Nach Beispiel 1.17 und Satz 1.19 ist ein der Bayesschätzer unter quadratischem Risiko für $X_1, \dots, X_n \sim N(\mu, E_d)$ und $\pi = N(a, \sigma^2 E_d)$ gegeben durch die bedingte Erwartung

$$\hat{\mu}_{a, \sigma^2} = \mathbb{E}_{\tilde{\mathbb{P}}}[T | X] = \frac{a + n\sigma^2 \bar{X}}{1 + n\sigma^2},$$

wie sofort aus der Normalverteilung der a-posteriori-Verteilung folgt. Man beachte, dass $\hat{\mu}_{a, \sigma^2}$ eine Konvexkombination vom a-priori-Mittelwert a und dem Stichprobenmittel \bar{X} ist.

1.22 Lemma. *Es liege die Situation aus Definition 1.9 vor. Für jede Entscheidungsregel ρ gilt*

$$\sup_{\vartheta \in \Theta} R(\vartheta, \rho) = \sup_{\pi} R_\pi(\rho),$$

wobei sich das zweite Supremum über alle a-priori-Verteilungen π erstreckt. Insbesondere ist das Risiko einer Bayesregel stets kleiner oder gleich dem Minimalrisiko.

Beweis. Natürlich gilt $R_\pi(\rho) = \int_\Theta R(\vartheta, \rho) \pi(d\vartheta) \leq \sup_{\vartheta \in \Theta} R(\vartheta, \rho)$. Durch Betrachtung der a-priori-Verteilungen δ_ϑ (Diracmaß im Punkt $\vartheta \in \Theta$) folgt daher die Behauptung. \square

1.23 Bemerkung. Man kann dieses Lemma insbesondere dazu verwenden, untere Schranken für das Minimax-Risiko durch das Bayesrisiko abzuschätzen.

1.24 Satz. Für jede Entscheidungsregel ρ gilt:

- (a) Ist ρ minimax und eindeutig in dem Sinn, dass jede andere Minimax-Regel die gleiche Risikofunktion besitzt, so ist ρ zulässig.
- (b) Ist ρ zulässig mit konstanter Risikofunktion, so ist ρ minimax.
- (c) Ist ρ eine Bayesregel (bzgl. π) und eindeutig in dem Sinn, dass jede andere Bayesregel (bzgl. π) die gleiche Risikofunktion besitzt, so ist ρ zulässig.
- (d) Die Parametermenge Θ bilde einen metrischen Raum mit Borel- σ -Algebra \mathcal{F}_Θ . Ist ρ eine Bayesregel (bzgl. π), so ist ρ zulässig, falls (i) $R_\pi(\rho) < \infty$; (ii) für jede nichtleere offene Menge U in Θ gilt $\pi(U) > 0$; (iii) für jede Regel ρ' mit $R_\pi(\rho') \leq R_\pi(\rho)$ ist $\vartheta \mapsto R(\vartheta, \rho')$ stetig.

Beweis. Übung! \square

1.25 Satz. Es sei X_1, \dots, X_n eine $N(\mu, E_d)$ -verteilte d -dimensionale mathematische Stichprobe mit $\mu \in \mathbb{R}^d$ unbekannt. Bezüglich quadratischem Risiko ist das arithmetische Mittel $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ minimax als Schätzer von μ .

1.26 Bemerkung. Die Beweisidee ist, dass \bar{X} ein sogenannter “improper Bayes“-Schätzer ist mit dem Lebesguemaß als a-priori-Verteilung. Dies wird mit einem Grenzwertargument formal umgesetzt.

Beweis. Zunächst beachte, dass $\bar{X} - \mu \sim N(0, \frac{1}{n} E_d)$ gilt, so dass

$$R(\mu, \bar{X}) = \sum_{i=1}^d \mathbb{E}_\mu[(\bar{X}_i - \mu_i)^2] = \frac{d}{n}$$

folgt. Betrachte nun die a-priori-Verteilung $\pi = N(0, \sigma^2 E_d)$ für μ . Gemäß Beispiel 1.21 ist der Bayes-optimale Schätzer $\hat{\mu}_{\sigma, n} = \frac{n\sigma^2}{1+n\sigma^2} \bar{X}$. Seine Risikofunktion ist (gemäß Bias-Varianz-Zerlegung)

$$\begin{aligned} R(\mu, \hat{\mu}_{\sigma, n}) &= (\mathbb{E}_\mu[\hat{\mu}_{\sigma, n}] - \mu)^2 + \text{Var}_\mu(\hat{\mu}_{\sigma, n}) \\ &= \left(\frac{1}{1+n\sigma^2}\right)^2 |\mu|^2 + \left(\frac{n\sigma^2}{1+n\sigma^2}\right)^2 \mathbb{E}[|\bar{X} - \mu|^2] \\ &= \frac{|\mu|^2 + nd\sigma^4}{(1+n\sigma^2)^2}. \end{aligned}$$

Somit können wir das Minimax-Risiko von unten abschätzen:

$$\begin{aligned}
\inf_{\rho} \sup_{\mu} R(\mu, \rho) &= \inf_{\rho} \sup_{\pi} R_{\pi}(\rho) \\
&\geq \inf_{\rho} \sup_{\sigma > 0} R_{N(0, \sigma^2 E_d)}(\rho) \\
&\geq \sup_{\sigma > 0} \inf_{\rho} R_{N(0, \sigma^2 E_d)}(\rho) \\
&= \sup_{\sigma > 0} \mathbb{E}_{\pi} \left[\frac{|\mu|^2 + nd\sigma^4}{(1 + n\sigma^2)^2} \right] \\
&= \sup_{\sigma > 0} \frac{d\sigma^2 + nd\sigma^4}{(1 + n\sigma^2)^2} = \sup_{\sigma > 0} \frac{d\sigma^2}{1 + n\sigma^2} = \frac{d}{n},
\end{aligned}$$

wie behauptet.

Anmerkung: da die bedingte Kovarianzmatrix $\text{Var}_{\mathbb{P}}(T | X) = \frac{\sigma^2}{1+n\sigma^2} E_d$ (s.o.) nicht von X abhängt, ergibt sich das Bayesrisiko alternativ auch direkt aus

$$R_{N(0, \sigma^2 E_d)}(\hat{\mu}_{\sigma, n}) = \mathbb{E}_{\mathbb{P}}[|\mathbb{E}_{\mathbb{P}}[T | X] - T|^2] = \sum_{i=1}^d \mathbb{E}_{\mathbb{P}}[\text{Var}_{\mathbb{P}}(T_i | X)] = \frac{d\sigma^2}{1 + n\sigma^2}.$$

□

1.27 Satz. *Es sei X_1, \dots, X_n eine $N(\mu, 1)$ -verteilte skalare mathematische Stichprobe mit $\mu \in \mathbb{R}$ unbekannt. Bezüglich quadratischem Risiko ist das arithmetische Mittel $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ zulässig als Schätzer von μ .*

Beweis. Gäbe es einen Schätzer $\hat{\mu}$ mit $R(\mu, \hat{\mu}) \leq \frac{1}{n}$ und $R(\mu_0, \hat{\mu}) < \frac{1}{n}$ für ein $\mu_0 \in \mathbb{R}$, so wäre wegen Stetigkeit der Risikofunktion $\mu \mapsto R(\mu, \hat{\mu})$ (Übung!) sogar $R(\mu, \hat{\mu}) \leq \frac{1}{n} - \varepsilon$ für alle $|\mu - \mu_0| < \delta$ mit $\varepsilon, \delta > 0$ geeignet. Damit hätte $\hat{\mu}$ ein Bayesrisiko $R_{N(0, \sigma^2)}(\hat{\mu}) \leq \frac{1}{n} - \varepsilon \int_{\mu_0 - \delta}^{\mu_0 + \delta} \varphi_{0, \sigma^2}$. Also wäre für $\sigma \rightarrow \infty$

$$\frac{1}{n} - R_{N(0, \sigma^2)} \geq \frac{2\varepsilon\delta}{\sigma\sqrt{2\pi}} \exp\left(-\frac{((\mu_0 - \delta) \vee (\mu_0 + \delta))^2}{(2\sigma^2)}\right) \asymp \frac{2\varepsilon\delta}{\sigma\sqrt{2\pi}}$$

größer als ein Vielfaches von σ^{-1} , während für den Bayesschätzer (siehe oben)

$$\frac{1}{n} - R_{N(0, \sigma^2)}(\hat{\mu}_{\sigma, n}) = \frac{1}{n} - \frac{\sigma^2}{1 + n\sigma^2} = \frac{\sigma^{-2}}{n(n + \sigma^{-2})}$$

von der Ordnung σ^{-2} ist. Dies widerspricht der Optimalität des Bayesschätzers bei einer hinreichend großen Wahl von σ . Also ist \bar{X} zulässig. □

1.28 Bemerkung. Beachte, dass aus der Zulässigkeit und der konstanten Risikofunktion von \bar{X} die Minimaleigenschaft auch direkt aus Satz 1.24(b) folgt.

Liegt eine andere Verteilung mit Erwartungswert μ und Varianz eins vor als die Normalverteilung, so ist \bar{X} weder zulässig noch minimax (sofern $n \geq 3$), vergleiche Lehmann/Casella, Seite 153. Für $d = 2$ ist \bar{X} weiterhin zulässig unter Normalverteilungsannahme, allerdings gilt das für $d \geq 3$ nicht mehr: Stein-Phänomen s.u.

1.29 Definition (*). Eine Verteilung π auf $(\Theta, \mathcal{F}_\Theta)$ heißt ungünstigste a-priori-Verteilung zu einer gegebenen Verlustfunktion, falls

$$\inf_{\rho} R_{\pi}(\rho) = \sup_{\pi'} \inf_{\rho} R_{\pi'}(\rho).$$

1.30 Satz (*). Es sei eine a-priori-Verteilung π mit zugehöriger Bayesregel ρ_{π} gegeben. Dann ist die Eigenschaft $R_{\pi}(\rho_{\pi}) = \sup_{\vartheta \in \Theta} R(\vartheta, \rho_{\pi})$ äquivalent zu folgender Sattelpunkteigenschaft

$$\forall \pi' \forall \rho' : R_{\pi'}(\rho_{\pi}) \leq R_{\pi}(\rho_{\pi}) \leq R_{\pi}(\rho').$$

Aus jeder dieser Eigenschaften folgt, dass ρ_{π} minimax und π ungünstigste a-priori-Verteilung ist.

Beweis. Wegen $\sup_{\vartheta} R(\vartheta, \rho_{\pi}) = \sup_{\pi'} R_{\pi'}(\rho_{\pi})$ folgt aus der Sattelpunkteigenschaft $R_{\pi}(\rho_{\pi}) \geq \sup_{\vartheta} R(\vartheta, \rho_{\pi})$. Da in jedem Fall ' \leq ' gilt, folgt $R_{\pi}(\rho_{\pi}) = \sup_{\vartheta} R(\vartheta, \rho_{\pi})$.

Andererseits bedeutet die Eigenschaft von ρ_{π} , Bayesschätzer zu sein, gerade dass $R_{\pi}(\rho_{\pi}) \leq R_{\pi}(\rho')$ für alle ρ' gilt. Mit $R_{\pi}(\rho_{\pi}) = \sup_{\vartheta \in \Theta} R(\vartheta, \rho_{\pi})$ schließen wir dann auch

$$R_{\pi'}(\rho_{\pi}) = \int_{\Theta} R(\vartheta, \rho_{\pi}) \pi'(d\vartheta) \leq \int_{\Theta} R_{\pi}(\rho_{\pi}) \pi'(d\vartheta) = R_{\pi}(\rho_{\pi}).$$

Aus der Sattelpunkteigenschaft folgt direkt die Minimaxeigenschaft:

$$\sup_{\vartheta} R(\vartheta, \rho_{\pi}) = \sup_{\pi'} R_{\pi'}(\rho_{\pi}) = \inf_{\rho'} R_{\pi}(\rho') \leq \inf_{\rho'} \sup_{\vartheta} R(\vartheta, \rho').$$

Analog erhalten wir $\inf_{\rho'} R_{\pi}(\rho') = \sup_{\pi'} R_{\pi'}(\rho_{\pi}) \geq \sup_{\pi'} \inf_{\rho} R_{\pi'}(\rho)$, so dass π ungünstigste a-priori-Verteilung ist. □

1.31 Beispiel. Es werde $X \sim \text{Bin}(n, p)$ mit $n \geq 1$ bekannt und $p \in [0, 1]$ unbekannt beobachtet. Gesucht wird ein Bayesschätzer $\hat{p}_{a,b}$ von p unter quadratischem Risiko für die a-priori-Verteilung $p \sim B(a, b)$, wobei $B(a, b)$ die Beta-Verteilung mit Parametern $a, b > 0$ auf $[0, 1]$ bezeichnet. Die a-posteriori-Verteilung berechnet sich zu $p \sim B(a + X, b + n - X)$ und der Bayesschätzer als $\hat{p}_{a,b} = \frac{a+X}{a+b+n}$ (Übung!). Als Risiko ergibt sich $\mathbb{E}_p[(\hat{p}_{a,b} - p)^2] = \frac{(a-ap-bp)^2 + np(1-p)}{(a+b+n)^2}$. Im Fall $a^* = b^* = \sqrt{n}/2$ erhält man das Risiko $(2\sqrt{n} + 2)^{-2}$ für $\hat{p}_{a^*,b^*} = \frac{X + \sqrt{n}/2}{n + \sqrt{n}} = \frac{X}{n} - \frac{X - \frac{n}{2}}{n(\sqrt{n} + 1)}$ (unabhängig von p !), woraus die Sattelpunkteigenschaft folgt:

$$\forall \pi \forall \hat{p} : R_{\pi}(\hat{p}_{a^*,b^*}) \leq R_{B(a^*,b^*)}(\hat{p}_{a^*,b^*}) \leq R_{B(a^*,b^*)}(\hat{p}).$$

Damit ist $B(a^*, b^*)$ ungünstigste a-priori-Verteilung und \hat{p}_{a^*,b^*} Minimax-Schätzer von p . Insbesondere ist der natürliche Schätzer $\hat{p} = X/n$ mit $\mathbb{E}_p[(\hat{p} - p)^2] = p(1-p)/n$ nicht minimax (er ist jedoch zulässig).

1.32 Bemerkung. Erhalten wir bei Wahl einer Klasse von a-priori-Verteilungen für ein statistisches Modell dieselbe Klasse (i.A. mit anderen Parametern) als a-posteriori-Verteilungen zurück, so nennt man die entsprechenden Verteilungsklassen konjugiert. An den Beispielen sehen wir, dass die Beta-Verteilungen zur Binomialverteilung konjugiert sind und die Normalverteilungen zu den Normalverteilungen (genauer müsste man spezifizieren, dass für unbekanntem Mittelwert in der Normalverteilung a-priori-Normalverteilungen konjugiert sind). Konjugierte Verteilungen sind die Ausnahme, nicht die Regel, und für komplexere Modelle werden häufig computer-intensive Methoden wie MCMC (Markov Chain Monte Carlo) verwendet, um die a-posteriori-Verteilung zu berechnen (Problem: i.A. hochdimensionale Integration).

1.3 Das Stein-Phänomen

Wir betrachten folgendes grundlegendes Problem: Anhand einer mathematischen Stichprobe $X_1, \dots, X_n \sim N(\mu, E_d)$ im \mathbb{R}^d soll $\mu \in \mathbb{R}^d$ möglichst gut bezüglich quadratischem Verlust $l(\mu, \hat{\mu}) = |\hat{\mu} - \mu|^2$ geschätzt werden. Intuitiv wegen Unabhängigkeit der Koordinaten ist das (koordinatenweise) arithmetische Mittel \bar{X} . Ein anderer, sogenannter empirischer Bayesansatz, beruht auf der Familie der a-priori-Verteilungen $\mu \sim N(0, \sigma^2 E_d)$. In den zugehörigen Bayesschätzern setzen wir dann allerdings statt σ^2 die Schätzung

$$\hat{\sigma}^2 = \frac{|\bar{X}|^2}{d} - n^{-1} \text{ (erwartungstreu wegen } X_i \sim N(0, (\sigma^2 + n^{-1})E_d) \text{ unter } \tilde{\mathbb{P}})$$

ein und erhalten

$$\hat{\mu} = \left(1 - \frac{1}{1 + n\hat{\sigma}^2}\right) \bar{X} = \left(1 - \frac{d}{n|\bar{X}|^2}\right) \bar{X}.$$

Der Ansatz lässt vermuten, dass $\hat{\mu}$ kleineres Risiko hat als \bar{X} , wann immer $|\mu|$ klein ist. Überraschenderweise gilt für Dimension $d \geq 3$ sogar, dass $\hat{\mu}$ besser ist als \bar{X} . Das folgende Steinsche Lemma ist der Schlüssel für den Beweis.

1.33 Lemma (Stein). *Es sei $f : \mathbb{R}^d \rightarrow \mathbb{R}$ eine Funktion, die Lebesgue-f.ü. absolut stetig in jeder Koordinate ist. Dann gilt für $X \sim N(\mu, \sigma^2 E_d)$ mit $\mu \in \mathbb{R}^d$, $\sigma > 0$,*

$$\mathbb{E}[(X - \mu)f(X)] = \sigma^2 \mathbb{E}[\nabla f(X)],$$

sofern $\mathbb{E}[|\frac{\partial f}{\partial x_i}(X)|] < \infty$ für alle $i = 1, \dots, d$ gilt.

Beweis. Ohne Einschränkung der Allgemeinheit betrachte die Koordinate $i = 1$ sowie $\mu = 0$, $\sigma = 1$; sonst setze $\tilde{f}(x) = f(\sigma x + \mu)$. Es genügt dann,

$$\mathbb{E}[X_1 f(X) | X_2 = x_2, \dots, X_d = x_d] = \mathbb{E}[\frac{\partial f}{\partial x_1}(X) | X_2 = x_2, \dots, X_d = x_d]$$

zu zeigen für Lebesgue-fast alle $x_2, \dots, x_d \in \mathbb{R}$, was wegen Unabhängigkeit gerade für $f_x(u) := f(u, x_2, \dots, x_d)$ die Identität $\int u f_x(u) e^{-u^2/2} du = \int f'_x(u) e^{-u^2/2} du$ ist. Dies folgt durch partielle Integration, sofern die Randterme

verschwinden; ein geschickter Einsatz des Satzes von Fubini zeigt dies jedoch ohne weitere Voraussetzungen:

$$\begin{aligned}
\int_{-\infty}^{\infty} f'_x(u) e^{-u^2/2} du &= \int_0^{\infty} f'_x(u) \int_u^{\infty} z e^{-z^2/2} dz du - \int_{-\infty}^0 f'_x(u) \int_{-\infty}^u z e^{-z^2/2} dz du \\
&= \int_0^{\infty} \left(\int_0^z f'_x \right) z e^{-z^2/2} dz - \int_{-\infty}^0 \left(\int_z^0 f'_x \right) z e^{-z^2/2} dz \\
&= \int_{-\infty}^{\infty} z e^{-z^2/2} (f_x(z) - f_x(0)) dz \\
&= \int_{-\infty}^{\infty} f_x(z) z e^{-z^2/2} dz.
\end{aligned}$$

Die Anwendung von Fubini in der zweiten Zeile wird gerechtfertigt durch dieselbe Rechnung mit $|f'_x|$ statt f'_x , da nach Voraussetzung $\int_0^{\infty} \int_u^{\infty} |f'_x(u)| z e^{-z^2/2} dz du$ und das analoge Integral über $u < 0$ endlich sind. \square

Betrachten wir nun allgemeine Schätzer der Form $\hat{\mu} = \bar{X} - g(\bar{X})$, so gilt

$$\mathbb{E}_{\mu}[|\hat{\mu} - \mu|^2] = \mathbb{E}_{\mu} \left[|\bar{X} - \mu|^2 + |g(\bar{X})|^2 - 2\langle \bar{X} - \mu, g(\bar{X}) \rangle \right].$$

Kann man nun auf $g = (g_1, \dots, g_d) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ das Steinsche Lemma koordinatenweise anwenden, so erhalten wir einen Ausdruck $W(\bar{X})$ unabhängig von μ :

$$\begin{aligned}
\mathbb{E}_{\mu}[|\hat{\mu} - \mu|^2] &= \frac{d}{n} + \mathbb{E}_{\mu}[W(\bar{X})] \text{ mit} \\
W(x) &:= |g(x)|^2 - \frac{2}{n} \sum_{i=1}^d \frac{\partial g_i(x)}{\partial x_i} = |g(x)|^2 - \frac{2}{n} \operatorname{div}(g(x)).
\end{aligned}$$

Für $g(x) = \frac{cx}{|x|^2}$, $c > 0$ eine Konstante, ist das Steinsche Lemma anwendbar. Wir erhalten

$$\operatorname{div}(g(x)) = c \sum_{i=1}^d \frac{|x|^2 - 2x_i^2}{|x|^4} = c(d-2)|x|^{-2}$$

und

$$W(x) = \frac{c^2}{|x|^2} - \frac{2c(d-2)}{n|x|^2} < 0 \text{ falls } c \in (0, 2(d-2)n^{-1}), d \geq 3.$$

Beachte, dass $g(x) = \frac{2(d-2)x}{n|x|^2}$ gerade $W(x) = 0$ löst, was a posteriori den Ansatz für g plausibel macht. Der minimale Wert $W(x) = -(d-2)^2/(n^2|x|^2)$ wird für $c = (d-2)/n$ erreicht, und wir haben folgendes bemerkenswertes Resultat bewiesen.

1.34 Satz. *Es sei $d \geq 3$ und X_1, \dots, X_n eine $N(\mu, E_d)$ -verteilte mathematische Stichprobe mit $\mu \in \mathbb{R}^d$ unbekannt. Dann gilt für den James-Stein-Schätzer*

$$\hat{\mu}_{JS} := \left(1 - \frac{d-2}{n|\bar{X}|^2}\right) \bar{X}$$

mit $\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i$, dass

$$\mathbb{E}_\mu[|\hat{\mu}_{JS} - \mu|^2] = \frac{d}{n} - \mathbb{E}_\mu \left[\frac{(d-2)^2}{n^2 |\bar{X}|^2} \right] < \frac{d}{n} = \mathbb{E}_\mu[|\bar{X} - \mu|^2].$$

Insbesondere ist \bar{X} bei quadratischem Risiko kein zulässiger Schätzer von μ im Fall $d \geq 3$!

1.35 Bemerkungen.

- (a) Die Abbildung $\mu \mapsto \mathbb{E}_\mu[|\bar{X}|^{-2}]$ ist monoton fallend in $|\mu|$ und erfüllt $\mathbb{E}_0[|\bar{X}|^{-2}] = n/(d-2)$, $\mathbb{E}_0[|\hat{\mu}_{JS} - \mu|^2] = 2/n$. Daher ist $\hat{\mu}_{JS}$ nur für μ nahe 0, große Dimensionen d und kleine Stichprobenumfänge n eine bedeutende Verbesserung von \bar{X} . Der James-Stein-Schätzer heißt auch Shrinkage-Schätzer, weil er die Beobachtungen zur Null hinzieht (wobei auch jeder andere Wert möglich wäre). In aktuellen hochdimensionalen Problemen findet diese Idee breite Anwendung.
- (b) Die k -te Koordinate $\hat{\mu}_{JS,k}$ des James-Stein-Schätzers verwendet zur Schätzung von μ_k auch die anderen Koordinaten $X_{i,l}$, $l \neq k$, obwohl diese unabhängig von $X_{i,k}$ sind. Eine Erklärung für diese zunächst paradoxe Situation ist, dass zwar $\sum_{k=1}^d \mathbb{E}_\mu[(\hat{\mu}_{JS,k} - \mu_k)^2] < \sum_{k=1}^d \mathbb{E}_\mu[(\bar{X}_k - \mu_k)^2]$ gilt, jedoch im Allgemeinen eine Koordinate k_0 existieren wird mit $\mathbb{E}_\mu[(\hat{\mu}_{JS,k_0} - \mu_{k_0})^2] > \mathbb{E}_\mu[(\bar{X}_{k_0} - \mu_{k_0})^2]$. Man beachte auch, dass der stochastische Fehler (die Varianz) von \bar{X} linear mit der Dimension d wächst, so dass es sich auszahlt, diesen Fehler auf Kosten einer Verzerrung (Bias) zu verringern, vgl. Übung.
- (c) Selbst der James-Stein-Schätzer (sogar mit positivem Gewicht, s.u.) ist unzulässig. Die Konstruktion eines zulässigen Minimax-Schätzers ist sehr schwierig (gelöst für $d \geq 6$, vgl. Lehmann/Casella, S. 358).

1.36 Satz. *Es sei $d \geq 3$ und X_1, \dots, X_n eine $N(\mu, E_d)$ -verteilte mathematische Stichprobe mit $\mu \in \mathbb{R}^d$ unbekannt. Dann ist der James-Stein-Schätzer mit positivem Gewicht*

$$\hat{\mu}_{JS+} := \left(1 - \frac{d-2}{n|\bar{X}|^2}\right)_+ \bar{X}, \quad a_+ := \max(a, 0),$$

bei quadratischem Risiko besser als der James-Stein-Schätzer $\hat{\mu}_{JS}$.

Beweis. Übung! □

1.4 Ergänzungen*

1.37 Definition. Ein Entscheidungskern oder eine randomisierte Entscheidungsregel $\rho : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$ ist ein Markovkern auf dem Aktionsraum (A, \mathcal{A}) mit der Interpretation, dass bei Vorliegen der Beobachtung x gemäß $\rho(x, \bullet)$ eine Entscheidung zufällig ausgewählt wird. Das zugehörige Risiko ist

$$R(\vartheta, \rho) := \mathbb{E}_\vartheta \left[\int_A l(\vartheta, a) \rho(da) \right] = \int_{\mathcal{X}} \int_A l(\vartheta, a) \rho(x, da) \mathbb{P}_\vartheta(dx).$$

1.38 Beispiel. Es sei $\Theta = \Theta_0 \dot{\cup} \Theta_1$, $A = [0, 1]$ und der Verlust $l(\vartheta, a) = l_0 a \mathbf{1}_{\Theta_0}(\vartheta) + l_1(1 - a) \mathbf{1}_{\Theta_1}(\vartheta)$ vorgegeben. In diesem Rahmen kann eine Entscheidungsregel ρ als randomisierter Test (oder Entscheidungskern) ρ' von $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1$ aufgefasst werden. Dazu setze $A' := \{0, 1\}$, $\mathcal{F}_{A'} := \mathcal{P}(A')$, benutze den gleichen Verlust l (eingeschränkt auf A') und definiere die bedingten Wahrscheinlichkeiten $\rho'(x, \{1\}) := \rho(x)$, $\rho'(x, \{0\}) := 1 - \rho(x, \{1\})$. Dies bedeutet also, dass $\rho(x)$ die Wahrscheinlichkeit angibt, mit der bei der Beobachtung x die Hypothese abgelehnt wird.

1.39 Lemma. *Es sei $A \subseteq \mathbb{R}^d$ konvex sowie $l(\vartheta, a)$ eine im zweiten Argument konvexe Verlustfunktion. Dann gibt es zu jeder randomisierten Entscheidungsregel ρ eine deterministische Entscheidungsregel ρ' , deren Risiko nicht größer ist.*

Beweis. Aus der Jensenschen Ungleichung folgt wegen Konvexität von $l(\vartheta, \bullet)$

$$R(\vartheta, \rho) = \int_{\mathcal{X}} \int_A l(\vartheta, a) \rho(x, da) \mathbb{P}_{\vartheta}(dx) \geq \int_{\mathcal{X}} l\left(\vartheta, \int_A a \rho(x, da)\right) \mathbb{P}_{\vartheta}(dx).$$

Da A konvex ist, gilt $\rho'(x) := \int_A a \rho(x, da) \in A$ und somit $R(\vartheta, \rho) \geq R(\vartheta, \rho')$. \square

1.40 Definition. Zu vorgegebener Verlustfunktion l heißt eine Entscheidungsregel ρ unverzerrt, falls

$$\forall \vartheta, \vartheta' \in \Theta : \mathbb{E}_{\vartheta}[l(\vartheta', \rho)] \geq \mathbb{E}_{\vartheta}[l(\vartheta, \rho)] =: R(\vartheta, \rho).$$

1.41 Lemma. *Es seien $g : \Theta \rightarrow A \subseteq \mathbb{R}$ und $l(\vartheta, \rho) = (\rho - g(\vartheta))^2$ der quadratische Verlust. Dann ist eine Entscheidungsregel (ein Schätzer von $g(\vartheta)$) $\hat{g} : \mathcal{X} \rightarrow A$ mit $\mathbb{E}_{\vartheta}[\hat{g}^2] < \infty$ und $\mathbb{E}_{\vartheta}[\hat{g}] \in g(\Theta)$ für alle $\vartheta \in \Theta$ genau dann unverzerrt, wenn sie erwartungstreu ist, d.h. $\mathbb{E}_{\vartheta}[\hat{g}] = g(\vartheta)$ für alle $\vartheta \in \Theta$ gilt.*

Beweis. Es gelte $\mathbb{E}_{\vartheta}[\hat{g}] = g(\vartheta')$ mit Parametern $\vartheta', \vartheta \in \Theta$. Dann ist

$$\mathbb{E}_{\vartheta}[(\hat{g} - g(\vartheta'))^2] = \text{Var}_{\vartheta}(\hat{g}) \leq (\mathbb{E}_{\vartheta}[\hat{g}] - g(\vartheta))^2 + \text{Var}_{\vartheta}(\hat{g}) = \mathbb{E}_{\vartheta}[(\hat{g} - g(\vartheta))^2]$$

und Gleichheit gilt genau dann, wenn $g(\vartheta) = \mathbb{E}_{\vartheta}[\hat{g}]$. Ist \hat{g} unverzerrt, so gilt $\mathbb{E}_{\vartheta}[(\hat{g} - g(\vartheta'))^2] \geq \mathbb{E}_{\vartheta}[(\hat{g} - g(\vartheta))^2]$, also $\mathbb{E}_{\vartheta}[\hat{g}] = g(\vartheta)$.

Ist \hat{g} andererseits erwartungstreu, so folgt für alle ϑ, ϑ' analog

$$\mathbb{E}_{\vartheta}[(\hat{g} - g(\vartheta))^2] = \text{Var}_{\vartheta}(\hat{g}) \leq (\mathbb{E}_{\vartheta}[\hat{g}] - g(\vartheta'))^2 + \text{Var}_{\vartheta}(\hat{g}) = \mathbb{E}_{\vartheta}[(\hat{g} - g(\vartheta'))^2],$$

also Unverzerrtheit. \square

1.42 Lemma. *Es sei $\Theta = \Theta_0 \dot{\cup} \Theta_1$, $A = [0, 1]$. Für den Verlust $l(\vartheta, a) = l_0 a \mathbf{1}_{\Theta_0}(\vartheta) + l_1(1 - a) \mathbf{1}_{\Theta_1}(\vartheta)$ mit $l_0, l_1 > 0$ ist eine Entscheidungsregel ρ (ein randomisierter Test von $H_0 : \vartheta \in \Theta_0$ gegen $H_1 : \vartheta \in \Theta_1$) genau dann unverzerrt, wenn sie zum Niveau $\alpha := \frac{l_1}{l_0 + l_1}$ unverfälscht ist, d.h.*

$$\forall \vartheta \in \Theta_0 : \mathbb{E}_{\vartheta}[\rho] \leq \alpha, \quad \forall \vartheta \in \Theta_1 : \mathbb{E}_{\vartheta}[\rho] \geq \alpha.$$

Beweis. Übung! \square

2 Dominierte Modelle und Suffizienz

2.1 Dominierte Modelle

2.1 Bemerkung. Wir sagen, dass ein Maß ν absolutstetig bezüglich einem Maß μ auf (Ω, \mathcal{F}) ist (Notation $\nu \ll \mu$), wenn $\mu(A) = 0 \Rightarrow \nu(A) = 0$ für alle $A \in \mathcal{F}$ gilt. Der Satz von Radon-Nikodym (Stochastik II, Funktionalanalysis) zeigt, dass dann für σ -endliches μ stets eine (μ -f.ü. eindeutige) μ -Dichte f_ν von ν existiert, das heißt eine messbare Funktion $f_\nu : \Omega \rightarrow \mathbb{R}^+$ mit $\nu(A) = \int_A f_\nu(x) \mu(dx)$, $A \in \mathcal{F}$. f_ν heißt auch Radon-Nikodym-Dichte von ν bezüglich μ und man schreibt $f_\nu = \frac{d\nu}{d\mu}$.

2.2 Definition. Ein statistisches Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ heißt dominiert (von μ), falls es ein σ -endliches Maß μ auf \mathcal{F} gibt, so dass \mathbb{P}_ϑ absolutstetig bezüglich μ ist ($\mathbb{P}_\vartheta \ll \mu$) für alle $\vartheta \in \Theta$. Die durch ϑ parametrisierte Radon-Nikodym-Dichte

$$L(\vartheta, x) := \frac{d\mathbb{P}_\vartheta}{d\mu}(x), \quad \vartheta \in \Theta, x \in \mathcal{X},$$

heißt auch Likelihoodfunktion, wobei diese meist als durch x parametrisierte Funktion in ϑ aufgefasst wird.

2.3 Beispiele.

- (a) $\mathcal{X} = \mathbb{R}$, $\mathcal{F} = \mathcal{B}_\mathbb{R}$, \mathbb{P}_ϑ ist gegeben durch eine Lebesguedichte f_ϑ , beispielsweise $\mathbb{P}_{(\mu, \sigma)} = N(\mu, \sigma^2)$ oder $\mathbb{P}_\vartheta = U([0, \vartheta])$. Dann ist das Modell vom Lebesguemaß dominiert mit Likelihoodfunktion $L(\vartheta, x) = f_\vartheta(x)$.
- (b) Jedes statistische Modell auf dem Stichprobenraum $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$ oder allgemeiner auf einem abzählbaren Raum $(\mathcal{X}, \mathcal{P}(\mathcal{X}))$ ist vom Zählmaß dominiert mit Likelihoodfunktion $L(\vartheta, x) = \mathbb{P}_\vartheta(\{x\})$.
- (c) Ist $\Theta = \{\vartheta_1, \vartheta_2, \dots\}$ abzählbar, so ist $\mu = \sum_i c_i \mathbb{P}_{\vartheta_i}$ mit $c_i > 0$, $\sum_i c_i = 1$ ein dominierendes Maß.
- (d) $\mathcal{X} = \mathbb{R}$, $\mathcal{F} = \mathcal{B}_\mathbb{R}$, $\mathbb{P}_\vartheta = \delta_\vartheta$ für $\vartheta \in \Theta = \mathbb{R}$ (δ_ϑ ist Punktmaß in ϑ) ist nicht dominiert. Ein dominierendes Maß μ müsste nämlich $\mu(\{\vartheta\}) > 0$ für alle $\vartheta \in \Theta$ und damit $\mu(A) = \infty$ für jede überabzählbare Borelmenge $A \subseteq \mathbb{R}$ erfüllen (sonst folgte aus $|\{x \in A \mid \mu(\{x\}) \geq 1/n\}| \leq n\mu(A) < \infty$, dass $A = \bigcup_{n \geq 1} \{x \in A \mid \mu(\{x\}) \geq 1/n\}$ abzählbar ist). Damit kann μ nicht σ -endlich sein.

2.4 Satz. *Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein dominiertes Modell. Dann gibt es ein Wahrscheinlichkeitsmaß \mathbb{Q} der Form $\mathbb{Q} = \sum_{i=1}^{\infty} c_i \mathbb{P}_{\vartheta_i}$ mit $c_i \geq 0$, $\sum_i c_i = 1$, $\vartheta_i \in \Theta$, so dass $\mathbb{P}_\vartheta \ll \mathbb{Q}$ für alle $\vartheta \in \Theta$ gilt.*

2.5 Bemerkung. Ein solches Wahrscheinlichkeitsmaß \mathbb{Q} heißt auch privilegiertes dominierendes Maß.

Beweis. Sei zunächst das dominierende Maß μ endlich sowie

$$\mathcal{P}_0 := \left\{ \sum_i c_i \mathbb{P}_{\vartheta_i} \mid \vartheta_i \in \Theta, c_i \geq 0, \sum_i c_i = 1 \right\} \text{ (konvexe Hülle von } (\mathbb{P}_{\vartheta}) \text{),}$$

$$\mathcal{A} := \left\{ A \in \mathcal{F} \mid \exists \mathbb{P} \in \mathcal{P}_0 : \mathbb{P}(A) > 0 \text{ und } \frac{d\mathbb{P}}{d\mu} > 0 \text{ } \mu\text{-f.ü. auf } A \right\}.$$

Wähle nun eine Folge (A_n) in \mathcal{A} mit $\mu(A_n) \rightarrow \sup_{A \in \mathcal{A}} \mu(A) < \infty$. Setze $A_\infty := \bigcup_n A_n$ und bezeichne \mathbb{P}_n ein Element in \mathcal{P}_0 mit $\mathbb{P}_n(A_n) > 0$, $\frac{d\mathbb{P}_n}{d\mu} > 0$ μ -f.ü. auf A_n . Für beliebige $c_n > 0$ mit $\sum_n c_n = 1$ setze $\mathbb{Q} := \sum_n c_n \mathbb{P}_n \in \mathcal{P}_0$.

Aus der Wahl von \mathbb{P}_n folgt $\frac{d\mathbb{Q}}{d\mu} \geq c_n \frac{d\mathbb{P}_n}{d\mu} > 0$ μ -f.ü. auf A_n und somit $\frac{d\mathbb{Q}}{d\mu} > 0$ μ -f.ü. auf A_∞ und $\mathbb{Q}(A_\infty) > 0$, so dass A_∞ ebenfalls in \mathcal{A} liegt und wegen $\mu(A_\infty) \geq \mu(A_n)$ also $\mu(A_\infty) = \sup_{A \in \mathcal{A}} \mu(A)$ erfüllt.

Zeige $\mathbb{P} \ll \mathbb{Q}$ für alle $\mathbb{P} \in \mathcal{P}_0$, woraus direkt $\mathbb{P}_{\vartheta} \ll \mathbb{Q}$ für alle ϑ folgt. Sonst gilt $\mathbb{P}(A) > 0$ und $\mathbb{Q}(A) = 0$ für ein $\mathbb{P} \in \mathcal{P}_0$ und ein $A \in \mathcal{F}$. Dies impliziert $\mathbb{Q}(A \cap A_\infty) = 0 \Rightarrow \mu(A \cap A_\infty) = 0$ (da $\frac{d\mathbb{Q}}{d\mu} > 0$ auf A_∞) und weiter $\mathbb{P}(A \cap A_\infty) = 0$ (da $\mathbb{P} \ll \mu$). Für $B := \{ \frac{d\mathbb{P}}{d\mu} > 0 \}$ gilt $\mathbb{P}(B) = 1$, und wir erhalten $\mathbb{P}(A \cap A_\infty^C \cap B) = \mathbb{P}(A) > 0$. Aus $\mathbb{P} \ll \mu$ folgt $\mu(A \cap A_\infty^C \cap B) > 0$ und somit $\mu(A_\infty \dot{\cup} (A \cap A_\infty^C \cap B)) > \mu(A_\infty)$. Nun ist aber $(\mathbb{P} + \mathbb{Q})/2 \in \mathcal{P}_0$ sowie $\frac{d(\mathbb{P} + \mathbb{Q})}{2d\mu} > 0$ μ -f.ü. auf $A_\infty \dot{\cup} (A \cap A_\infty^C \cap B)$, was $A_\infty \dot{\cup} (A \cap A_\infty^C \cap B) \in \mathcal{A}$ zeigt. Dies widerspricht aber der Eigenschaft $\mu(A_\infty) = \sup_{A \in \mathcal{A}} \mu(A)$.

Ist μ σ -endlich, so zerlege $\mathcal{X} := \bigcup_{m \geq 1} \mathcal{X}_m$ mit $\mu(\mathcal{X}_m) < \infty$, definiere das Maß \mathbb{Q}_m wie oben \mathbb{Q} , wobei im Fall $\mathbb{P}_{\vartheta}(\mathcal{X}_m) = 0$ für alle $\vartheta \in \Theta$ einfach $\mathbb{Q}_m = \mathbb{P}_{\vartheta}$ für ein beliebiges $\vartheta \in \Theta$ gesetzt wird. Dann leistet $\sum_{m \geq 1} 2^{-m} \mathbb{Q}_m$ das Gewünschte. \square

2.2 Exponentialfamilien

2.6 Definition. Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_{\vartheta})_{\vartheta \in \Theta})$ ein von μ dominiertes Modell. Dann heißt $(\mathbb{P}_{\vartheta})_{\vartheta \in \Theta}$ Exponentialfamilie (in $\eta(\vartheta)$ und T), wenn $k \in \mathbb{N}$, $\eta : \Theta \rightarrow \mathbb{R}^k$, $C : \Theta \rightarrow \mathbb{R}^+$, $T : \mathcal{X} \rightarrow \mathbb{R}^k$ messbar und $h : \mathcal{X} \rightarrow \mathbb{R}^+$ messbar existieren, so dass

$$\frac{d\mathbb{P}_{\vartheta}}{d\mu}(x) = C(\vartheta) h(x) \exp(\langle \eta(\vartheta), T(x) \rangle_{\mathbb{R}^k}), \quad x \in \mathcal{X}, \vartheta \in \Theta.$$

T wird natürliche suffiziente Statistik von $(\mathbb{P}_{\vartheta})_{\vartheta \in \Theta}$ genannt. Sind η_1, \dots, η_k linear unabhängige Funktionen und gilt für alle $\vartheta \in \Theta$ die Implikation

$$\lambda_0 + \lambda_1 T_1 + \dots + \lambda_k T_k = 0 \text{ } \mathbb{P}_{\vartheta}\text{-f.s.} \Rightarrow \lambda_0 = \lambda_1 = \dots = \lambda_k = 0$$

($1, T_1, \dots, T_k$ sind \mathbb{P}_{ϑ} -f.s. linear unabhängig), so heißt die Exponentialfamilie (strikt) k -parametrisch.

2.7 Bemerkungen.

- (a) $C(\vartheta)$ ist nur Normierungskonstante: $C(\vartheta) = (\int h(x) e^{\langle \eta(\vartheta), T(x) \rangle} \mu(dx))^{-1}$.
- (b) Die Darstellung ist nicht eindeutig, mit einer invertierbaren Matrix $A \in \mathbb{R}^{k \times k}$ erhält man beispielsweise eine Exponentialfamilie in $\tilde{\eta}(\vartheta) = A\eta(\vartheta)$ und $\tilde{T}(x) = (A^\top)^{-1}T(x)$.

- (c) Die Funktion h kann in das dominierende Maß absorbiert werden, indem man $\tilde{\mu}(dx) = h(x)\mu(dx)$ statt μ betrachtet. Da $C(\vartheta) > 0$ gilt, ist dann $\frac{d\mathbb{P}_\vartheta}{d\tilde{\mu}} > 0$ $\tilde{\mu}$ -f.s. und alle Verteilungen $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ sind untereinander und mit $\tilde{\mu}$ äquivalent (gegenseitig absolut-stetig). Insbesondere bildet für ein ϑ_0 die Familie $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ auch eine Exponentialfamilie bezüglich \mathbb{P}_{ϑ_0} in $\tilde{\eta}(\vartheta) = \eta(\vartheta) - \eta(\vartheta_0)$ und $T(x)$.
- (d) Aus der Identifizierbarkeitsforderung $\mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta'}$ für alle $\vartheta \neq \vartheta'$ folgt die Injektivität von η . Andererseits impliziert die Injektivität von η bei einer k -parametrischen Exponentialfamilie die Identifizierbarkeitsforderung.

2.8 Definition. Bildet $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ eine Exponentialfamilie (mit obiger Notation), so heißt

$$\mathcal{Z} := \left\{ u \in \mathbb{R}^k \mid \int_{\mathcal{X}} e^{\langle u, T(x) \rangle} h(x) \mu(dx) \in (0, \infty) \right\}$$

ihr natürlicher Parameterraum. Die entsprechend mit $u \in \mathcal{Z}$ parametrisierte Familie wird natürliche Exponentialfamilie in T genannt.

2.9 Beispiele.

- (a) $(N(\mu, \sigma^2))_{\mu \in \mathbb{R}, \sigma > 0}$ ist zweiparametrische Exponentialfamilie in $\eta(\mu, \sigma) = (\mu/\sigma^2, 1/(2\sigma^2))^\top$ und $T(x) = (x, -x^2)^\top$ unter dem Lebesguemaß als dominierendem Maß. Jedes u der Form $u = (\mu/\sigma^2, 1/(2\sigma^2))^\top$ ist natürlicher Parameter, und der natürliche Parameterraum ist gegeben durch $\mathcal{Z} = \mathbb{R} \times (0, \infty)$. Ist $\sigma > 0$ bekannt, so liegt eine einparametrische Exponentialfamilie in $\eta(\mu) = \mu/\sigma^2$ und $T(x) = x$ vor.
- (b) $(\text{Bin}(n, p))_{p \in (0,1)}$ bildet eine Exponentialfamilie in $\eta(p) = \log(p/(1-p))$ (auch logit-Funktion genannt) und $T(x) = x$ bezüglich dem Zählmaß μ auf $\{0, 1, \dots, n\}$. Der natürliche Parameterraum ist \mathbb{R} . Beachte, dass für den Parameterbereich $p = [0, 1]$ keine Exponentialfamilie vorliegt, da $(\text{Bin}(n, p))_{p \in [0,1]}$ keine äquivalenten Wahrscheinlichkeitsmaße sind.

2.10 Lemma. Bildet $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ eine (k -parametrische) Exponentialfamilie in $\eta(\vartheta)$ und $T(x)$, so bilden auch die Produktmaße $(\mathbb{P}_\vartheta^{\otimes n})_{\vartheta \in \Theta}$ eine (k -parametrische) Exponentialfamilie in $\eta(\vartheta)$ und $\sum_{i=1}^n T(x_i)$ mit

$$\frac{d\mathbb{P}_\vartheta^{\otimes n}}{d\mu^{\otimes n}}(x) = C(\vartheta)^n \left(\prod_{i=1}^n h(x_i) \right) \exp \left(\langle \eta(\vartheta), \sum_{i=1}^n T(x_i) \rangle \right), \quad x \in \mathcal{X}^n, \vartheta \in \Theta.$$

Beweis. Dies folgt sofort aus der Produktformel $\frac{d\mathbb{P}_\vartheta^{\otimes n}}{d\mu^{\otimes n}}(x) = \prod_{i=1}^n \frac{d\mathbb{P}_\vartheta}{d\mu}(x_i)$. \square

2.11 Satz. Es sei $(\mathbb{P}_\vartheta)_{\vartheta \in \mathcal{Z}}$ eine Exponentialfamilie mit natürlichem Parameterraum $\mathcal{Z} \subseteq \mathbb{R}^k$ und Darstellung

$$\frac{d\mathbb{P}_\vartheta}{d\mu}(x) = C(\vartheta) h(x) \exp(\langle \vartheta, T(x) \rangle) = h(x) \exp(\langle \vartheta, T(x) \rangle - A(\vartheta)), \quad \vartheta \in \mathcal{Z},$$

wobei $A(\vartheta) = \log \left(\int h(x) \exp(\langle \vartheta, T(x) \rangle) \mu(dx) \right)$. Ist ϑ^0 ein innerer Punkt von \mathcal{Z} , so ist die erzeugende Funktion $\psi_{\vartheta^0}(s) = \mathbb{E}_{\vartheta^0}[e^{\langle T, s \rangle}]$ in einer Umgebung der Null wohldefiniert und beliebig oft differenzierbar.

Es gilt $\psi_{\vartheta^0}(s) = \exp(A(\vartheta^0 + s) - A(\vartheta^0))$ für alle s mit $\vartheta^0 + s \in \mathcal{Z}$. Für $i, j = 1, \dots, k$ folgt $\mathbb{E}_{\vartheta^0}[T_i] = \frac{dA}{d\vartheta_i}(\vartheta^0)$ und $\text{Cov}_{\vartheta^0}(T_i, T_j) = \frac{d^2A}{d\vartheta_i d\vartheta_j}(\vartheta^0)$.

Beweis. Für alle $s \in \mathbb{R}^k$ mit $\vartheta^0 + s \in \mathcal{Z}$ gilt

$$\psi_{\vartheta^0}(s) = \int e^{\langle T, s \rangle} e^{\langle \vartheta^0, T \rangle - A(\vartheta^0)} h \, d\mu = \int e^{\langle \vartheta^0 + s, T \rangle - A(\vartheta^0)} h \, d\mu = e^{A(\vartheta^0 + s) - A(\vartheta^0)}.$$

Insbesondere ist ψ_{ϑ^0} in einer Umgebung von $s = 0$ endlich und somit wohldefiniert.

Für $v \in \mathbb{R}^k$ und $\varepsilon > 0$ hinreichend klein, betrachte den Differenzenquotienten

$$\frac{\psi_{\vartheta^0}(\varepsilon v) - \psi_{\vartheta^0}(0)}{\varepsilon} = \int \frac{e^{\varepsilon \langle T, v \rangle} - 1}{\varepsilon} e^{\langle \vartheta^0, T \rangle - A(\vartheta^0)} h \, d\mu.$$

Der Bruch im Integranden konvergiert für $\varepsilon \rightarrow 0$ punktweise gegen $\langle T, v \rangle$. Aus der Ungleichung $|\frac{e^{az} - 1}{z}| \leq \frac{e^{\delta|a|}}{\delta}$ für $|z| \leq \delta$, $a \in \mathbb{R}$, ergibt sich $\frac{e^{\varepsilon_0 \langle T, v \rangle} + e^{-\varepsilon_0 \langle T, v \rangle}}{\varepsilon_0}$ als Majorante des Bruchs für alle $\varepsilon \leq \varepsilon_0$. Nach dem ersten Schritt gilt für $\varepsilon_0 > 0$ mit $\vartheta^0 \pm \varepsilon_0 v \in \mathcal{Z}$, dass die Majorante integrierbar ist, und wir schließen mittels dominierter Konvergenz auf die Richtungsableitung

$$\lim_{\varepsilon \rightarrow 0} \frac{\psi_{\vartheta^0}(\varepsilon v) - \psi_{\vartheta^0}(0)}{\varepsilon} = \mathbb{E}_{\vartheta^0}[\langle T, v \rangle], \quad v \in \mathbb{R}^k.$$

Also ist ψ_{ϑ^0} differenzierbar bei Null mit Gradienten $\mathbb{E}_{\vartheta^0}[T]$. Wegen $A(\vartheta^0 + s) = A(\vartheta^0) + \log(\psi_{\vartheta^0}(s))$ für s in einer Nullumgebung ist also auch A differenzierbar bei ϑ^0 mit Gradienten $\mathbb{E}_{\vartheta^0}[T]$ (beachte $\psi_{\vartheta^0}(0) = 1$).

Analog ergibt sich, dass ψ_{ϑ^0} beliebig oft differenzierbar ist mit höheren partielle Ableitungen

$$\left. \frac{d^{i_1}}{ds_1^{i_1}} \cdots \frac{d^{i_k}}{ds_k^{i_k}} \psi_{\vartheta^0}(s) \right|_{s=0} = \int T_1^{i_1} \cdots T_k^{i_k} e^{\langle \vartheta^0, T \rangle - A(\vartheta^0)} d\mu = \mathbb{E}_{\vartheta^0}[T_1^{i_1} \cdots T_k^{i_k}].$$

Wir erhalten insbesondere

$$\frac{d^2A}{d\vartheta_i d\vartheta_j}(\vartheta^0) = \frac{d^2 \log(\psi_{\vartheta^0})}{ds_i ds_j}(0) = \left(\frac{d^2 \psi_{\vartheta^0}}{ds_i ds_j} - \frac{d\psi_{\vartheta^0}}{ds_i} \frac{d\psi_{\vartheta^0}}{ds_j} \right)(0) = \text{Cov}_{\vartheta^0}(T_i, T_j).$$

□

2.12 Beispiel. Für $\mathbb{P}_{\vartheta} = N(\vartheta, 1)^{\otimes n}$ bildet $(\mathbb{P}_{\vartheta})_{\vartheta \in \mathbb{R}}$ eine natürliche Exponentialfamilie in $T(x) = \sum_{i=1}^n x_i$, $x \in \mathbb{R}^n$, mit $A(\vartheta) = n\vartheta^2/2$. Wir erhalten $\mathbb{E}_{\vartheta}[T] = A'(\vartheta)$, d.h. $\mathbb{E}_{\vartheta}[\sum_{i=1}^n X_i] = n\vartheta$, sowie $\text{Var}_{\vartheta}(T) = A''(\vartheta)$, d.h. $\text{Var}_{\vartheta}(\sum_{i=1}^n X_i) = n$.

2.3 Suffizienz

2.13 Beispiel. Es sei X_1, \dots, X_n eine gemäß der Lebesgue-dichte $f_{\vartheta} : \mathbb{R} \rightarrow \mathbb{R}^+$ verteilte mathematische Stichprobe. Dann liefern die Statistiken \bar{X} oder

$\max(X_1, \dots, X_n)$ im Allgemeinen Information über f_ϑ und damit ϑ . Hingegen sind $\mathbf{1}(X_1 < X_2)$ oder $\mathbf{1}(X_1 = \max(X_1, \dots, X_n))$ Statistiken, deren Verteilung nicht von f_ϑ abhängt (sofern die i.i.d.-Annahme gültig ist) und somit keinerlei Informationen über ϑ beinhalten. Allgemein heißt eine Statistik V *ancillary*, wenn ihre Verteilung nicht von ϑ abhängt. Also ist beispielsweise $V = \mathbf{1}(X_1 < X_2)$ ancillary, weil stets $\text{Bin}(1, 1/2)$ -verteilt. Intuitiv ist alle Information bereits in der Ordnungsstatistik $X_{(1)}, \dots, X_{(n)}$ enthalten mit $X_{(1)} = \min\{X_1, \dots, X_n\}$, $X_{(k+1)} := \min\{X_1, \dots, X_n\} \setminus \{X_{(1)}, \dots, X_{(k)}\}$ oder äquivalent in der empirischen Verteilungsfunktion $\hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_i \leq x)$, $x \in \mathbb{R}$. Die Ordnungsstatistik und die empirische Verteilungsfunktion reduzieren die Datenmenge (da nicht injektiv) und sind in folgendem Sinne suffizient, vgl. Beispiel 2.18(c) unten.

2.14 Definition. Eine (S, \mathcal{F}) -wertige Statistik T auf $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ heißt suffizient (für $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$), falls für jedes $\vartheta \in \Theta$ die bedingte Wahrscheinlichkeit von \mathbb{P}_ϑ gegeben T nicht von ϑ abhängt, d.h. es existiert ein Markovkern $k : S \times \mathcal{F} \rightarrow [0, 1]$, so dass

$$\forall \vartheta \in \Theta, B \in \mathcal{F} : k(T, B) = \mathbb{P}_\vartheta(B | T) := \mathbb{E}_\vartheta[\mathbf{1}_B | T] \quad \mathbb{P}_\vartheta\text{-f.s.}$$

Statt $k(t, B)$ schreiben wir $\mathbb{P}_\bullet(B | T = t)$ bzw. $\mathbb{E}_\bullet[\mathbf{1}_B | T = t]$.

2.15 Satz (Faktorisierungskriterium von Neyman). *Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein von μ dominiertes Modell mit Likelihoodfunktion L sowie T eine (S, \mathcal{F}) -wertige Statistik. Dann ist T genau dann suffizient, wenn eine messbare Funktion $h : \mathcal{X} \rightarrow \mathbb{R}^+$ existiert, so dass für alle $\vartheta \in \Theta$ eine messbare Funktion $g_\vartheta : S \rightarrow \mathbb{R}^+$ existiert mit*

$$L(\vartheta, x) = g_\vartheta(T(x))h(x) \quad \text{für } \mu\text{-f.a. } x \in \mathcal{X}.$$

2.16 Lemma. *Es seien \mathbb{P} und μ Wahrscheinlichkeitsmaße mit $\mathbb{P} \ll \mu$ und T eine messbare Abbildung auf $(\mathcal{X}, \mathcal{F})$. Dann gilt für alle $B \in \mathcal{F}$*

$$\mathbb{P}(B | T) = \mathbb{E}_\mathbb{P}[\mathbf{1}_B | T] = \frac{\mathbb{E}_\mu[\mathbf{1}_B \frac{d\mathbb{P}}{d\mu} | T]}{\mathbb{E}_\mu[\frac{d\mathbb{P}}{d\mu} | T]} \quad \mathbb{P}\text{-f.s.}$$

Beweis. Für jede beschränkte messbare Funktion φ erfüllt die rechte Seite

$$\begin{aligned} \mathbb{E}_\mathbb{P} \left[\frac{\mathbb{E}_\mu[\mathbf{1}_B \frac{d\mathbb{P}}{d\mu} | T]}{\mathbb{E}_\mu[\frac{d\mathbb{P}}{d\mu} | T]} \varphi(T) \right] &= \mathbb{E}_\mu \left[\frac{\mathbb{E}_\mu[\mathbf{1}_B \frac{d\mathbb{P}}{d\mu} | T]}{\mathbb{E}_\mu[\frac{d\mathbb{P}}{d\mu} | T]} \varphi(T) \frac{d\mathbb{P}}{d\mu} \right] \\ &= \mathbb{E}_\mu \left[\frac{\mathbb{E}_\mu[\mathbf{1}_B \frac{d\mathbb{P}}{d\mu} | T]}{\mathbb{E}_\mu[\frac{d\mathbb{P}}{d\mu} | T]} \varphi(T) \mathbb{E}_\mu[\frac{d\mathbb{P}}{d\mu} | T] \right] \\ &= \mathbb{E}_\mu[\mathbf{1}_B \frac{d\mathbb{P}}{d\mu} \varphi(T)] \\ &= \mathbb{E}_\mathbb{P}[\mathbf{1}_B \varphi(T)]. \end{aligned}$$

Zusammen mit der $\sigma(T)$ -Messbarkeit ist dies genau die Charakterisierung dafür, dass die rechte Seite eine Version der bedingten Erwartung $\mathbb{E}_\mathbb{P}[\mathbf{1}_B | T]$ ist. \square

2.17 Bemerkung. Mit den üblichen Approximationsargumenten lässt sich dies zu $\mathbb{E}_{\mathbb{P}}[f | T] = \mathbb{E}_{\mu}[f \frac{d\mathbb{P}}{d\mu} | T] / \mathbb{E}_{\mu}[\frac{d\mathbb{P}}{d\mu} | T]$ für $f \in L^1(\mathbb{P})$ verallgemeinern.

Beweis des Faktorisierungssatzes. Ohne Einschränkung sei μ ein Wahrscheinlichkeitsmaß, sonst betrachte das äquivalente Wahrscheinlichkeitsmaß $\tilde{\mu}(dx) = z(x)\mu(dx)$ mit $z = \sum_{m \geq 1} 2^{-m} \mu(\mathcal{X}_m)^{-1} \mathbf{1}_{\mathcal{X}_m}$, wobei die Zerlegung $\mathcal{X} := \bigcup_{m \geq 1} \mathcal{X}_m$ mit $\mu(\mathcal{X}_m) < \infty$ wegen der σ -Endlichkeit von μ existiert.

Aus dem Lemma und der Form von $L(\vartheta, x)$ folgt daher

$$\mathbb{P}_{\vartheta}(B | T) = \frac{g_{\vartheta}(T) \mathbb{E}_{\mu}[\mathbf{1}_B h | T]}{g_{\vartheta}(T) \mathbb{E}_{\mu}[h | T]} = \frac{\mathbb{E}_{\mu}[\mathbf{1}_B h | T]}{\mathbb{E}_{\mu}[h | T]} \quad \mathbb{P}_{\vartheta}\text{-f.s.}$$

Da die rechte Seite unabhängig von ϑ ist, ist T suffizient.

Ist nun andererseits T suffizient, so setze $k(T, B) := \mathbb{P}_{\bullet}(B | T)$. Für das privilegierte dominierende Maß \mathbb{Q} gemäß Satz 2.4 gilt dann ebenfalls $\mathbb{Q}(B | T) = \sum_i c_i \mathbb{P}_{\vartheta_i}(B | T) = k(T, B)$ \mathbb{Q} -f.s. Nach dem Satz von Radon-Nikodym gilt auf dem Teilraum $(\mathcal{X}, \sigma(T))$

$$\forall \vartheta \exists f_{\vartheta} : \mathcal{X} \rightarrow \mathbb{R}^+ \quad \sigma(T)\text{-messbar} : \frac{d\mathbb{P}_{\vartheta} |_{\sigma(T)}}{d\mathbb{Q} |_{\sigma(T)}} = f_{\vartheta}.$$

Nach Stochastik II gibt es eine messbare Funktion g_{ϑ} , so dass $f_{\vartheta} = g_{\vartheta} \circ T$, und für beliebiges $B \in \mathcal{F}$ erhalten wir

$$\mathbb{P}_{\vartheta}(B) = \mathbb{E}_{\vartheta}[\mathbb{E}_{\bullet}[\mathbf{1}_B | T]] = \mathbb{E}_{\mathbb{Q}}[\mathbb{E}_{\mathbb{Q}}[\mathbf{1}_B | T] g_{\vartheta}(T)] = \mathbb{E}_{\mathbb{Q}}[\mathbf{1}_B g_{\vartheta}(T)],$$

so dass $g_{\vartheta} \circ T$ auch die Radon-Nikodym-Dichte $\frac{d\mathbb{P}_{\vartheta}}{d\mathbb{Q}}$ auf ganz \mathcal{F} ist. Mit $\frac{d\mathbb{P}_{\vartheta}}{d\mu} = \frac{d\mathbb{P}_{\vartheta}}{d\mathbb{Q}} \frac{d\mathbb{Q}}{d\mu}$ erhalten wir den Ausdruck von $L(\vartheta, x)$, wobei $h(x) = \frac{d\mathbb{Q}}{d\mu}(x)$. \square

2.18 Beispiele.

- (a) Die Identität $T(x) = x$ und allgemein jede bijektive, bi-messbare Transformation T ist suffizient.
- (b) Die natürliche suffiziente Statistik T einer Exponentialfamilie ist in der Tat suffizient. Im Normalverteilungsmodell $(N(\mu, \sigma^2)^{\otimes n})_{\mu \in \mathbb{R}, \sigma > 0}$ ist damit $T_1(x) = (\sum_{i=1}^n x_i, -\sum_{i=1}^n x_i^2)^{\top}$ suffizient, aber durch Transformation auch $T_2(x) = (\bar{x}, \bar{x}^2)$ oder $T_3(x) = (\bar{x}, \bar{s}^2)$ mit der empirischen Varianz $\bar{s}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$. Bei einer Bernoullikette $(\text{Bin}(1, p)^{\otimes n})_{p \in (0,1)}$ ist $T(x) = \sum_{i=1}^n x_i$ (die Anzahl der Erfolge) suffizient.

Betrachten wir den Fall $(N(\vartheta, 1)^{\otimes n})_{\vartheta \in \mathbb{R}}$ eines unbekanntes Mittelwerts, so ist bereits $T(x) = \bar{x}$ suffizient. In diesem Fall können wir die bedingte Verteilung $\mathbb{P}_{\bullet}(B | T)$ einfach generieren: für (auch von T) unabhängige $N(0, 1)$ -Zufallsvariablen Z_1, \dots, Z_n setze $\tilde{X}_i := \bar{X} + Z_i - \bar{Z}$. Dann ist

$(\tilde{X}_1, \dots, \tilde{X}_n)$ normalverteilt mit $\mathbb{E}[\tilde{X}_i] = \vartheta$ sowie für $i \neq j$

$$\begin{aligned}\text{Var}(\tilde{X}_i) &= \text{Var}(\bar{X}) + \text{Var}((1 - 1/n)Z_i) + \sum_{j \neq i} \text{Var}(Z_j/n) \\ &= \frac{1}{n} + \frac{(n-1)^2}{n^2} + \frac{n-1}{n^2} = 1, \\ \text{Cov}(\tilde{X}_i, \tilde{X}_j) &= \text{Var}(\bar{X}) - \text{Cov}(Z_i, \bar{Z}) - \text{Cov}(\bar{Z}, Z_j) + \text{Var}(\bar{Z}) \\ &= \frac{1}{n} - \frac{1}{n} - \frac{1}{n} + \frac{1}{n} = 0.\end{aligned}$$

Es gilt also $\tilde{X} \sim N(\vartheta, 1)^{\otimes n}$, und wir haben eine Stichprobe des Modells basierend auf der suffizienten Statistik generiert, ohne den Parameter ϑ zu kennen.

- (c) Ist X_1, \dots, X_n eine mathematische Stichprobe, wobei X_i gemäß der Lebesgue-dichte $f_\vartheta : \mathbb{R} \rightarrow \mathbb{R}^+$ verteilt ist, so ist die Ordnungsstatistik $(X_{(1)}, \dots, X_{(n)})$ suffizient. Die Likelihoodfunktion lässt sich nämlich in der Form $L(\vartheta, x) = \prod_{i=1}^n f_\vartheta(x_{(i)})$ schreiben.
- (d) Es wird die Realisierung $(N_t, t \in [0, T])$ eines Poissonprozesses zum unbekanntem Parameter $\lambda > 0$ kontinuierlich auf $[0, T]$ beobachtet (man denke an Geigerzähleraufzeichnungen). Mit $S_k = \inf\{t \geq 0 \mid N_t = k\}$ werden die Sprungzeiten bezeichnet. In der Wahrscheinlichkeitstheorie wird gezeigt, dass bedingt auf das Ereignis $\{N_T = n\}$ die Sprungzeiten (S_1, \dots, S_n) dieselbe Verteilung haben wie die Ordnungsstatistik $(X_{(1)}, \dots, X_{(n)})$ mit unabhängigen $X_i \sim U([0, T])$. Da sich die Beobachtung $(N_t, t \in [0, T])$ eindeutig aus den S_k rekonstruieren lässt, ist die Verteilung dieser Beobachtung gegeben $\{N_T = n\}$ unabhängig von λ , und N_T ist somit eine suffiziente Statistik (die Kenntnis der Gesamtzahl der gemessenen radioaktiven Zerfälle liefert bereits die maximal mögliche Information über die Intensität λ).

2.19 Satz (Rao-Blackwell). *Es seien $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell, der Aktionsraum $A \subseteq \mathbb{R}^k$ konvex und die Verlustfunktion $l(\vartheta, a)$ im zweiten Argument konvex. Ist T eine für $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ suffiziente Statistik, so gilt für jede Entscheidungsregel ρ und für $\tilde{\rho} := \mathbb{E}_\bullet[\rho \mid T]$ die Risikoabschätzung*

$$\forall \vartheta \in \Theta : R(\vartheta, \tilde{\rho}) \leq R(\vartheta, \rho).$$

Beweis. Dies folgt aus der Jensenschen Ungleichung für bedingte Erwartungen:

$$R(\vartheta, \tilde{\rho}) = \mathbb{E}_\vartheta[l(\vartheta, \mathbb{E}_\vartheta[\rho \mid T])] \leq \mathbb{E}_\vartheta[\mathbb{E}_\vartheta[l(\vartheta, \rho) \mid T]] = R(\vartheta, \rho).$$

□

2.20 Bemerkung. Ist l sogar strikt konvex sowie $\mathbb{P}_\vartheta(\tilde{\rho} = \rho) < 1$, so gilt in der Jensenschen Ungleichung sogar die strikte Ungleichung und $\tilde{\rho}$ ist besser als ρ .

2.21 Satz (*). *Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell und T eine suffiziente Statistik. Dann gibt es zu jedem randomisierten Test φ einen randomisierten Test $\tilde{\varphi}$, der nur von T abhängt und dieselben Fehlerwahrscheinlichkeiten erster und zweiter Art besitzt, nämlich $\tilde{\varphi} = \mathbb{E}_\bullet[\varphi \mid T]$.*

Beweis. Dies folgt jeweils aus $\mathbb{E}_\vartheta[\tilde{\varphi}] = \mathbb{E}_\vartheta[\varphi]$. \square

2.22 Beispiel. Es sei X_1, \dots, X_n eine $U([0, \vartheta])$ -verteilte mathematische Stichprobe mit $\vartheta > 0$ unbekannt. Dann ist \bar{X} ein erwartungstreuer Schätzer des Erwartungswerts $\frac{\vartheta}{2}$, so dass $\hat{\vartheta} = 2\bar{X}$ ein plausibler Schätzer von ϑ ist mit quadratischem Risiko $R(\vartheta, \hat{\vartheta}) = 4 \text{Var}_\vartheta(\bar{X}) = \frac{4\vartheta^2}{12n}$. Nun ist jedoch (bezüglich Lebesguemaß auf $(\mathbb{R}^+)^n$) die Likelihoodfunktion

$$L(\vartheta, x) = \prod_{i=1}^n (\vartheta^{-1} \mathbf{1}_{[0, \vartheta]}(x_i)) = \vartheta^{-n} \mathbf{1}_{[0, \vartheta]}(\max_{i=1, \dots, n} x_i).$$

Demnach ist $X_{(n)} = \max_{i=1, \dots, n} X_i$ eine suffiziente Statistik, und wir bilden

$$\tilde{\vartheta} := \mathbb{E}_\bullet[\hat{\vartheta} | X_{(n)}] = \frac{2}{n} \sum_{i=1}^n \mathbb{E}_\bullet[X_i | X_{(n)}].$$

Aus Symmetriegründen reicht es, $\mathbb{E}_\bullet[X_1 | X_{(n)}]$ zu bestimmen. Als bedingte Verteilung von X_1 gegeben $\{X_{(n)} = m\}$ vermuten wir $\frac{1}{n}\delta_m + \frac{n-1}{n}U([0, m])$ wegen

$$\begin{aligned} \mathbb{P}_\vartheta(X_1 \leq x | X_{(n)} \in [m, m+h]) &= \frac{(x \wedge (m+h))(m+h)^{n-1} - (x \wedge m)m^{n-1}}{(m+h)^n - m^n} \\ &\xrightarrow{h \rightarrow 0} \frac{1}{n} \mathbf{1}(\{m < x\}) + \frac{n-1}{n} \frac{x \wedge m}{m}. \end{aligned}$$

In der Tat gilt für $x \in [0, \vartheta]$:

$$\begin{aligned} &\int_0^\vartheta \left(\frac{1}{n} \delta_m + \frac{n-1}{n} U([0, m]) \right) ([0, x]) \mathbb{P}_\vartheta^{X_{(n)}}(dm) \\ &= \int_0^\vartheta \left(\frac{1}{n} \mathbf{1}_{[0, x]}(m) + \frac{n-1}{n} \frac{x \wedge m}{m} \right) n m^{n-1} \vartheta^{-n} dm \\ &= \frac{1}{n} (x/\vartheta)^n + \frac{n-1}{n} \left((x/\vartheta)^n + \frac{n x (\vartheta^{n-1} - x^{n-1})}{(n-1) \vartheta^n} \right) \\ &= \frac{x}{\vartheta} = \mathbb{P}_\vartheta(X_1 \leq x). \end{aligned}$$

Es folgt $\mathbb{E}[X_1 | X_{(n)}] = \frac{1}{n} X_{(n)} + \frac{n-1}{n} \frac{X_{(n)}}{2} = \frac{n+1}{2n} X_{(n)}$. Wir erhalten $\tilde{\vartheta} = \frac{n+1}{n} X_{(n)}$. Natürlich ist $\tilde{\vartheta}$ auch erwartungstreu und als quadratisches Risiko ergibt eine kurze Rechnung $R(\vartheta, \tilde{\vartheta}) = \frac{\vartheta^2}{n^2+2n}$. Wir sehen, dass $\tilde{\vartheta}$ bedeutend besser als $\hat{\vartheta}$ ist, für $n \rightarrow \infty$ erhalten wir die Ordnung $O(n^{-2})$ anstelle $O(n^{-1})$. Es bleibt, die Frage zu klären, ob auch $\tilde{\vartheta}$ noch weiter verbessert werden kann (s.u.).

2.4 Vollständigkeit

2.23 Definition. Eine (S, \mathcal{S}) -wertige Statistik T auf $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ heißt vollständig, falls für alle messbaren Funktionen $f: S \rightarrow \mathbb{R}$ gilt

$$\forall \vartheta \in \Theta : \mathbb{E}_\vartheta[f(T)] = 0 \implies \forall \vartheta \in \Theta : f(T) = 0 \quad \mathbb{P}_\vartheta\text{-f.s.}$$

2.24 Bemerkung. Ist T vollständig und g messbar, so ist auch $g(T)$ vollständig, wie sofort aus der Definition folgt. Dieses Verhalten ist genau entgegengesetzt zur Suffizienz, wo aus $g(T)$ suffizient folgt, dass T suffizient ist.

Eine Statistik V heißt *ancillary*, falls ihre Verteilung \mathbb{P}_ϑ^V nicht von ϑ abhängt. Ist T vollständig und $V = f(T)$ integrierbar und ancillary, so hängt die Verteilung von V nicht von ϑ ab und damit ist $\mathbb{E}_\vartheta[V] = c$, c eine Konstante, und wegen Vollständigkeit $V = c$ \mathbb{P}_ϑ -f.s. (betrachte $\tilde{f}(x) = f(x) - c$). Vollständigkeit von T impliziert also, dass jede ancillary Statistik der Form $V = f(T)$ trivial (d.h. fast sicher konstant) ist. Es ist keine redundante Information mehr in T enthalten. Da $V = \mathbf{1}(X_1 < X_2)$ in Beispiel 2.13 ancillary und nicht-trivial ist, sind weder $T = (X_1, \dots, X_n)$ (Identität) noch $T = (X_1, X_2)$ vollständig.

2.25 Satz (Lehmann-Scheffé). *Es seien $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell und $\gamma(\vartheta) \in \mathbb{R}$, $\vartheta \in \Theta$, der jeweils interessierende Parameter. Es existiere ein erwartungstreuer Schätzer $\hat{\gamma}$ von $\gamma(\vartheta)$ mit endlicher Varianz. Ist T eine suffiziente und vollständige Statistik, so ist $\tilde{\gamma} = \mathbb{E}_\bullet[\hat{\gamma} | T]$ ein Schätzer von gleichmäßig kleinster Varianz in der Klasse aller erwartungstreuen Schätzer (UMVU: uniformly minimum variance unbiased).*

Beweis. Zunächst ist klar, dass $\tilde{\gamma}$ wiederum erwartungstreu ist. Außerdem ist $\tilde{\gamma}$ der f.s. einzige erwartungstreue Schätzer, der $\sigma(T)$ -messbar ist, weil jeder andere solche Schätzer $\bar{\gamma}$ wegen Vollständigkeit $\mathbb{E}[\tilde{\gamma} - \bar{\gamma}] = 0 \Rightarrow \tilde{\gamma} = \bar{\gamma}$ \mathbb{P}_ϑ -f.s. erfüllt. Nach dem Satz von Rao-Blackwell besitzt $\tilde{\gamma}$ damit kleineres quadratisches Risiko als jeder andere erwartungstreue Schätzer. Nach der Bias-Varianz-Zerlegung ist das quadratische Risiko bei erwartungstreuen Schätzern gleich der Varianz. \square

2.26 Bemerkung. Beachte, dass die Aussage des Satzes von Lehmann-Scheffé sogar analog für das Risiko bei beliebigen im zweiten Argument konvexen Verlustfunktionen gilt, wie sofort aus dem Satz von Rao-Blackwell folgt.

2.27 Satz. *Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ eine k -parametrische Exponentialfamilie in T mit natürlichem Parameter $\vartheta \in \Theta \subseteq \mathbb{R}^k$. Besitzt Θ ein nichtleeres Inneres, so ist T suffizient und vollständig.*

Beweis. Es bleibt, die Vollständigkeit zu beweisen. Ohne Einschränkung sei $[-a, a]^k \subseteq \Theta$ für ein $a > 0$ (sonst verschiebe entsprechend) sowie $h(x) = 1$ (sonst betrachte $\tilde{\mu}(dx) = h(x)\mu(dx)$). Für alle $\vartheta \in \Theta$ gelte $\mathbb{E}_\vartheta[f(T)] = 0$ für ein $f \in \bigcap_{\vartheta \in \Theta} L^1(\mathbb{R}^k, \mathbb{P}_\vartheta^T)$. Mit $f^+ = \max(f, 0)$, $f^- = \max(-f, 0)$ sowie mit dem Bildmaß μ^T des dominierenden Maßes μ unter T folgt

$$\forall \vartheta \in [-a, a]^k : \int_{\mathbb{R}^k} \exp(\langle \vartheta, t \rangle) f^+(t) \mu^T(dt) = \int_{\mathbb{R}^k} \exp(\langle \vartheta, t \rangle) f^-(t) \mu^T(dt).$$

Insbesondere gilt $\int f^+(t) \mu^T(dt) = \int f^-(t) \mu^T(dt) =: M \in [0, \infty)$. Ist $M = 0$, so ist $f^+ = f^- = 0$ μ^T -f.ü. und somit $f(T) = 0$ \mathbb{P}_ϑ -f.s. für alle $\vartheta \in \Theta$. Dann ist die Vollständigkeit von T nachgewiesen.

Betrachte nun den Fall $M > 0$. Dann definieren $\mathbb{P}^+(dt) := M^{-1} f^+(t) \mu^T(dt)$, $\mathbb{P}^-(dt) := M^{-1} f^-(t) \mu^T(dt)$ Wahrscheinlichkeitsmaße auf $(\mathbb{R}^k, \mathfrak{B}_{\mathbb{R}^k})$. Die obige Identität bedeutet gerade, dass die Laplace-Transformierten $\chi^\pm(\vartheta) :=$

$\int_{\mathbb{R}^k} \exp(\langle \vartheta, t \rangle) \mathbb{P}^\pm(dt)$ für $\vartheta \in [-a, a]^k$ übereinstimmen. χ^+ und χ^- sind darüberhinaus auf dem k -dimensionalen komplexen Streifen $\{\vartheta \in \mathbb{C}^k \mid |\operatorname{Re}(\vartheta_j)| < a\}$ wohldefiniert und analytisch (Potenzreihen). Der Eindeutigkeitssatz für analytische Funktionen impliziert daher $\chi^+(iu) = \chi^-(iu)$ für alle $u \in \mathbb{R}^k$. Also besitzen \mathbb{P}^+ und \mathbb{P}^- dieselben charakteristischen Funktionen, so dass $\mathbb{P}^+ = \mathbb{P}^-$ folgt (Eindeutigkeitssatz für charakteristische Funktionen). Dies liefert $f^+ = f^-$ μ^T -f.ü. und somit $f(T) = 0$ \mathbb{P}_ϑ -f.s. für alle $\vartheta \in \Theta$. T ist vollständig. \square

2.28 Beispiele.

- (a) Das lineare Modell $Y = X\beta + \sigma\varepsilon$ mit Designmatrix $X \in \mathbb{R}^{n \times p}$ vom Rang p , Gaußschen Fehlern $\varepsilon \sim N(0, E_n)$ bildet eine $(p+1)$ -parametrische Exponentialfamilie in $\eta(\beta, \sigma) = \sigma^{-2}(\beta, -1/2)^\top \in \mathbb{R}^p \times \mathbb{R}^-$ und $T(Y) = (X^\top Y, |Y|^2)^\top \in \mathbb{R}^p \times \mathbb{R}^+$. Der natürliche Parameterbereich $\mathcal{X} = \mathbb{R}^p \times \mathbb{R}^-$ besitzt nichtleeres Inneres in \mathbb{R}^{p+1} , so dass T suffizient und vollständig ist. Durch bijektive Transformation ergibt sich, dass dies auch für $((X^\top X)^{-1}X^\top Y, |Y|^2) = (\hat{\beta}, |\Pi_X Y|^2 + (n-p)\hat{\sigma}^2)$ mit dem Kleinst-Quadrat-Schätzer $\hat{\beta}$ und $\hat{\sigma}^2 = \frac{|Y - X\hat{\beta}|^2}{n-p}$ gilt. Wegen $\Pi_X Y = X\hat{\beta}$ ist also a fortiori auch $(\hat{\beta}, \hat{\sigma}^2)$ suffizient und vollständig. Damit besitzen beide Schätzer jeweils minimale Varianz in der Klasse aller (!) erwartungstreuen Schätzer (von β bzw. σ^2), in Erweiterung des Satzes von Gauß-Markov, der sich auf die Klasse der erwartungstreuen, linearen Schätzer bezieht. Hierfür ist die Normalverteilungsannahme essentiell.

Im Spezialfall $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$ i.i.d. ist also $\hat{\mu} = \bar{Y}$ UMVU. Auch wenn wir bereits wussten, dass $\hat{\mu}$ minimax und zulässig ist, zeigt die UMVU-Eigenschaft, dass es keinen erwartungstreuen Schätzer $\tilde{\mu}$ von μ geben kann, der $\operatorname{Var}_{\mu_0}(\tilde{\mu}) < \operatorname{Var}_{\mu_0}(\hat{\mu}) = \frac{\sigma^2}{n}$ für irgendein $\mu_0 \in \mathbb{R}$ erfüllt.

- (b) Es sei $X_1, \dots, X_n \sim U([0, \vartheta])$ eine mathematische Stichprobe mit $\vartheta > 0$ unbekannt. Aus der Form $L(x, \vartheta) = \vartheta^{-n} \mathbf{1}(x_{(n)} \leq \vartheta)$ für $x \in (\mathbb{R}^+)^n$ der Likelihoodfunktion folgt, dass das Maximum $X_{(n)}$ der Beobachtungen suffizient ist (Beispiel 2.22). Gilt für $f: \mathbb{R}^+ \rightarrow \mathbb{R}$, integrierbar auf jedem Intervall $[0, \vartheta]$, und für alle $\vartheta > 0$

$$\mathbb{E}_\vartheta[f(X_{(n)})] = \int_0^\vartheta f(t) n \vartheta^{-n} t^{n-1} dt = 0,$$

so muss $f = 0$ Lebesgue-fast überall gelten, woraus die Vollständigkeit von $X_{(n)}$ folgt. Andererseits gilt $\mathbb{E}_\vartheta[X_{(n)}] = \frac{n}{n+1}\vartheta$. Also ist $\hat{\vartheta} = \frac{n+1}{n}X_{(n)}$ erwartungstreuer Schätzer von ϑ mit gleichmäßig kleinster Varianz.

2.5 Cramér-Rao-Schranke

2.29 Lemma (Chapman-Robbins-Ungleichung). *Es seien $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell, \hat{g} ein erwartungstreuer Schätzer von $g(\vartheta) \in \mathbb{R}$ und $\vartheta_0 \in \Theta$.*

Dann gilt für jedes $\vartheta \in \Theta$ mit $\mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta_0}$, $\mathbb{P}_\vartheta \ll \mathbb{P}_{\vartheta_0}$, $\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}} \in L^2(\mathbb{P}_{\vartheta_0})$

$$\text{Var}_{\vartheta_0}(\hat{g}) = \mathbb{E}_{\vartheta_0}[(\hat{g} - g(\vartheta_0))^2] \geq \frac{(g(\vartheta) - g(\vartheta_0))^2}{\text{Var}_{\vartheta_0}\left(\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}}\right)}.$$

Beweis. Dies folgt wegen $\mathbb{E}_{\vartheta_0}\left[\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}}\right] = 1$ aus

$$\begin{aligned} |g(\vartheta) - g(\vartheta_0)| &= |\mathbb{E}_\vartheta[\hat{g} - g(\vartheta_0)] - \mathbb{E}_{\vartheta_0}[\hat{g} - g(\vartheta_0)]| \\ &= \left| \mathbb{E}_{\vartheta_0} \left[(\hat{g} - g(\vartheta_0)) \left(\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}} - 1 \right) \right] \right| \\ &\leq \mathbb{E}_{\vartheta_0}[(\hat{g} - g(\vartheta_0))^2]^{1/2} \mathbb{E}_{\vartheta_0} \left[\left(\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}} - 1 \right)^2 \right]^{1/2}, \end{aligned}$$

wobei zuletzt die Cauchy-Schwarz-Ungleichung angewendet wurde. \square

2.30 Bemerkung. Beachte, dass die untere Schranke der Chapman-Robbins-Ungleichung unabhängig vom Schätzer \hat{g} ist. Ziel ist es, eine möglichst große untere Schranke herzuleiten, also das Supremum der Schranke über alle zulässigen Werte von ϑ zu bilden.

2.31 Beispiele.

- (a) Wir beobachten $X \sim \text{Exp}(\vartheta)$ mit $\vartheta > 0$ unbekannt. Dann ist die Likelihoodfunktion gegeben durch $\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}}(x) = (\vartheta/\vartheta_0)e^{-(\vartheta-\vartheta_0)x}$, $x \geq 0$. Diese ist in $L^2(\mathbb{P}_{\vartheta_0})$ nur im Fall $\vartheta > \vartheta_0/2$ und besitzt dann die Varianz $\text{Var}_{\vartheta_0}\left(\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}}\right) = \frac{(\vartheta-\vartheta_0)^2}{\vartheta_0(2\vartheta-\vartheta_0)}$.

Im Fall erwartungstreuer Schätzer \hat{g} für $g(\vartheta) = \vartheta$ ergibt die Chapman-Robbins-Gleichung $\mathbb{E}_{\vartheta_0}[(\hat{g} - g(\vartheta_0))^2] \geq \sup_{\vartheta > \vartheta_0/2} \vartheta_0(2\vartheta - \vartheta_0) = \infty$. Sofern wir also beliebig große Werte ϑ zulassen, existiert kein erwartungstreuer Schätzer von ϑ mit endlicher Varianz.

Im Fall $g(\vartheta) = \vartheta^{-1}$ hingegen liefert die Chapman-Robbins-Ungleichung $\mathbb{E}_{\vartheta_0}[(\hat{g} - g(\vartheta_0))^2] \geq \sup_{\vartheta > \vartheta_0/2} \frac{2\vartheta - \vartheta_0}{\vartheta^2 \vartheta_0} = \vartheta_0^{-2}$, und die Identität $\hat{g} = X$ erreicht auch diese Schranke.

- (b) Wir beobachten $X \sim N(\vartheta, \frac{1}{n})$ mit $\vartheta \in \mathbb{R}$ unbekannt. Dann ist

$$\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}} = \exp\left(-\frac{n}{2}(X-\vartheta)^2 + \frac{n}{2}(X-\vartheta_0)^2\right) = \exp\left(n(\vartheta-\vartheta_0)X - \frac{n}{2}(\vartheta^2 - \vartheta_0^2)\right).$$

Wir berechnen mit $\tilde{\vartheta} := \vartheta_0 + 2(\vartheta - \vartheta_0)^2$ und $\mathbb{E}_{\vartheta_0}\left[\frac{d\mathbb{P}_{\tilde{\vartheta}}}{d\mathbb{P}_{\vartheta_0}}\right] = 1$

$$\begin{aligned} \mathbb{E}_{\vartheta_0} \left[\left(\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}} \right)^2 \right] &= \mathbb{E}_{\vartheta_0}[\exp(2n(\vartheta - \vartheta_0)X - n(\vartheta^2 - \vartheta_0^2))] \\ &= \mathbb{E}_{\vartheta_0} \left[\frac{d\mathbb{P}_{\tilde{\vartheta}}}{d\mathbb{P}_{\vartheta_0}} \exp\left(\frac{n}{2}(\tilde{\vartheta}^2 - \vartheta_0^2) - n(\vartheta^2 - \vartheta_0^2)\right) \right] \\ &= \exp\left(\frac{n}{2}((\vartheta_0 + 2(\vartheta - \vartheta_0)^2)^2 - \vartheta_0^2) - n(\vartheta^2 - \vartheta_0^2)\right) \\ &= \exp(n(\vartheta - \vartheta_0)^2). \end{aligned}$$

Die Chapman-Robbins-Schranke für erwartungstreue Schätzer $\hat{\vartheta}$ von ϑ (also $g(\vartheta) = \vartheta$) ist

$$\text{Var}_{\vartheta_0}(\hat{\vartheta}) \geq \sup_{\vartheta \neq \vartheta_0} \frac{(\vartheta - \vartheta_0)^2}{\exp(n(\vartheta - \vartheta_0)^2) - 1} = \frac{1}{n},$$

wobei das Supremum für $\vartheta \rightarrow \vartheta_0$ erhalten wird. Diese untere Schranke wird natürlich von $\hat{\vartheta} = X$ auch erreicht. Beachte, dass $\text{Var}_{\vartheta_0}(\frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}})$ exponentiell wächst in $(\vartheta - \vartheta_0)^2$, was typisch ist und erklärt, warum es meist reicht, die Chapman-Robbins-Schranke für $\vartheta \rightarrow \vartheta_0$ zu betrachten.

2.32 Definition. Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ mit $\Theta \subseteq \mathbb{R}^k$ ein von μ dominiertes Modell mit Likelihoodfunktion L . Das Modell heißt Hellinger-differenzierbar bei $\vartheta_0 \in \text{int}(\Theta)$, wenn es einen Zufallsvektor $\dot{\ell}(\vartheta_0) \in L^2(\mathbb{P}_{\vartheta_0}; \mathbb{R}^k)$ gibt mit

$$\lim_{\vartheta \rightarrow \vartheta_0} \int \left(\frac{\sqrt{L(\vartheta, x)} - \sqrt{L(\vartheta_0, x)} - \frac{1}{2} \langle \dot{\ell}(\vartheta_0, x), \vartheta - \vartheta_0 \rangle \sqrt{L(\vartheta_0, x)}}{|\vartheta - \vartheta_0|} \right)^2 d\mu(x) = 0.$$

Die Fisher-Informationsmatrix bei $\vartheta_0 \in \text{int}(\Theta)$ ist gegeben durch

$$I(\vartheta_0) = \mathbb{E}_{\vartheta_0}[\dot{\ell}(\vartheta_0) \dot{\ell}(\vartheta_0)^\top].$$

Mit $\ell(\vartheta, x) := \log(L(\vartheta, x))$ ($\log 0 := -\infty$) wird die Loglikelihood-Funktion bezeichnet. Man nennt $\vartheta \mapsto \dot{\ell}(\vartheta)$ auch Score-Funktion.

2.33 Bemerkungen.

(a) Sofern alle folgenden Ausdrücke klassisch differenzierbar sind, gilt

$$\nabla_\vartheta \sqrt{L(\vartheta)} = \frac{\nabla_\vartheta L(\vartheta)}{2\sqrt{L(\vartheta)}} = \frac{1}{2} \sqrt{L(\vartheta)} \nabla_\vartheta \log(L(\vartheta)) = \frac{1}{2} \sqrt{L(\vartheta)} \dot{\ell}(\vartheta).$$

Insbesondere ist die Score-Funktion $\dot{\ell}$ die Ableitung der Loglikelihood-Funktion ℓ . In der Statistik wird die Ableitung bezüglich dem Parameter traditionell mit einem Punkt bezeichnet.

(b) Die Differenzierbarkeit im quadratischen $L^2(\mu)$ -Mittel ist sehr viel allgemeiner und recht natürlich. Wegen $\sqrt{L(\vartheta)} \in L^2(\mu)$, was sofort aus $\int L(\vartheta) d\mu = 1 < \infty$ folgt, kann man $\vartheta \mapsto \sqrt{L(\vartheta)}$ als $L^2(\mu)$ -wertige Abbildung auffassen, so dass die Verteilungen (\mathbb{P}_ϑ) im geometrischen Sinne eine Untermannigfaltigkeit des Hilbertraums $L^2(\mu)$ bilden. Insbesondere gilt

$$\mathbb{E}_{\vartheta_0}[|\dot{\ell}(\vartheta_0)|^2] = \int |\dot{\ell}(\vartheta_0)(x)|^2 L(\vartheta_0, x) \mu(dx) = \int |\dot{\ell}(\vartheta_0)(x) \sqrt{L(\vartheta_0, x)}|^2 \mu(dx),$$

so dass $\dot{\ell}(\vartheta) \in L^2(\mathbb{P}_{\vartheta_0}; \mathbb{R}^k) \iff \dot{\ell}(\vartheta) \sqrt{L(\vartheta_0)} \in L^2(\mu; \mathbb{R}^k)$ und das Integral bei der Definition von Hellinger-Differenzierbarkeit wohldefiniert ist.

- (c) Nach Definition ist die Fisher-Informationsmatrix symmetrisch. Wegen $\langle I(\vartheta_0)v, v \rangle = \mathbb{E}_{\vartheta_0}[\langle \dot{\ell}(\vartheta_0), v \rangle^2] \geq 0$ für beliebige $v \in \mathbb{R}^k$ ist die Fisher-Informationsmatrix auch stets positiv-semidefinit.
- (d) Die Score-Funktion und die Fisher-Information sind unabhängig vom dominierenden Maß; denn mit einem privilegierten dominierenden Maß \mathbb{Q} gilt $L(\vartheta) = \frac{d\mathbb{P}_\vartheta}{d\mathbb{Q}} \frac{d\mathbb{Q}}{d\mu}$, so dass in der Definition von $\dot{\ell}$ der Faktor $\frac{d\mathbb{Q}}{d\mu}$ aus dem Integranden ausgeklammert werden kann und somit $\dot{\ell}$ ebenso die Definition bezüglich dem dominierenden Maß \mathbb{Q} erfüllt.

2.34 Lemma. Für alle $\vartheta \in \Theta \subseteq \mathbb{R}^k$ in einer Umgebung von $\vartheta_0 \in \Theta$ gelte $\mathbb{P}_\vartheta \ll \mathbb{P}_{\vartheta_0}$ sowie die $L^2(\mathbb{P}_{\vartheta_0})$ -Differenzierbarkeit der Likelihoodfunktion $L_{\vartheta_0}(\vartheta, x) := \frac{d\mathbb{P}_\vartheta}{d\mathbb{P}_{\vartheta_0}}(x)$ bei ϑ_0 , d.h. mit dem Gradienten $\dot{L}_{\vartheta_0}(\vartheta_0) \in L^2(\mathbb{P}_{\vartheta_0}; \mathbb{R}^k)$ gilt

$$\lim_{\vartheta \rightarrow \vartheta_0} \mathbb{E}_{\vartheta_0} \left[\left(\frac{L_{\vartheta_0}(\vartheta) - L_{\vartheta_0}(\vartheta_0) - \langle \dot{L}_{\vartheta_0}(\vartheta_0), \vartheta - \vartheta_0 \rangle}{|\vartheta - \vartheta_0|} \right)^2 \right] = 0.$$

Dann ist (\mathbb{P}_ϑ) Hellinger-differenzierbar bei ϑ_0 mit $\dot{\ell}(\vartheta_0) = \dot{L}_{\vartheta_0}(\vartheta_0)$.

Beweis. Aus obiger Bemerkung folgt, dass es genügt, \mathbb{P}_{ϑ_0} als dominierendes Maß und die Likelihoodfunktion $L_{\vartheta_0}(\vartheta)$ zu betrachten. Wir erhalten mit $L_{\vartheta_0}(\vartheta_0) = 1$ und der Minkowski-Ungleichung in $L^2(\mathbb{P}_{\vartheta_0})$:

$$\begin{aligned} & \left\| \sqrt{L_{\vartheta_0}(\vartheta)} - 1 - \frac{1}{2} \langle \dot{L}_{\vartheta_0}(\vartheta_0), \vartheta - \vartheta_0 \rangle \right\|_{L^2(\mathbb{P}_{\vartheta_0})} \\ & \leq \left\| \frac{L_{\vartheta_0}(\vartheta) - 1 - \langle \dot{L}_{\vartheta_0}(\vartheta_0), \vartheta - \vartheta_0 \rangle}{\sqrt{L_{\vartheta_0}(\vartheta)} + 1} \right\|_{L^2(\mathbb{P}_{\vartheta_0})} + \left\| \langle \dot{L}_{\vartheta_0}(\vartheta_0), \vartheta - \vartheta_0 \rangle \left(\frac{1}{\sqrt{L_{\vartheta_0}(\vartheta)} + 1} - \frac{1}{2} \right) \right\|_{L^2(\mathbb{P}_{\vartheta_0})} \\ & \leq \left\| L_{\vartheta_0}(\vartheta) - 1 - \langle \dot{L}_{\vartheta_0}(\vartheta_0), \vartheta - \vartheta_0 \rangle \right\|_{L^2(\mathbb{P}_{\vartheta_0})} + \frac{|\vartheta - \vartheta_0|}{2} \left\| |\dot{L}_{\vartheta_0}(\vartheta_0)| \frac{1 - \sqrt{L_{\vartheta_0}(\vartheta)}}{\sqrt{L_{\vartheta_0}(\vartheta)} + 1} \right\|_{L^2(\mathbb{P}_{\vartheta_0})}. \end{aligned}$$

Nach Voraussetzung besitzt der erste Summand die Ordnung $o(|\vartheta - \vartheta_0|)$. Außerdem gilt insbesondere $L_{\vartheta_0}(\vartheta) \rightarrow L_{\vartheta_0}(\vartheta_0) = 1$ in $L^2(\mathbb{P}_{\vartheta_0})$ und damit auch in \mathbb{P}_{ϑ_0} -Wahrscheinlichkeit. Weil nun $G(x) := \frac{1 - \sqrt{x}}{\sqrt{x} + 1}$ für $x \geq 0$ im Betrag durch 1 beschränkt ist und $\lim_{x \rightarrow 1} G(x) = 0$ gilt, folgt mit dominierter Konvergenz (unter stochastischer Konvergenz), dass die zweite $L^2(\mathbb{P}_{\vartheta_0})$ -Norm gegen Null konvergiert. Damit ist der gesamte Ausdruck von der Ordnung $o(|\vartheta - \vartheta_0|)$. \square

2.35 Beispiele.

- (a) Es sei X_1, \dots, X_n eine mathematische Stichprobe gemäß der Lebesgue-dichte $f_\vartheta(x) = \frac{1}{2\sigma} e^{-|x - \vartheta|/\sigma}$, $x \in \mathbb{R}$, $\sigma > 0$ bekannt und $\vartheta \in \mathbb{R}$ unbekannt. Für beliebige $\vartheta_0, \vartheta \in \mathbb{R}$ gilt

$$L_{\vartheta_0}(\vartheta) = \exp \left(- \sum_{i=1}^n (|X_i - \vartheta| - |X_i - \vartheta_0|) / \sigma \right)$$

und L_{ϑ_0} ist $L^2(\mathbb{P}_{\vartheta_0})$ -differenzierbar (Nachweis!) mit

$$\dot{L}_{\vartheta_0}(\vartheta_0) = \dot{\ell}(\vartheta_0) = \sum_{i=1}^n (\mathbf{1}(X_i - \vartheta_0 > 0) - \mathbf{1}(X_i - \vartheta_0 < 0)) / \sigma.$$

Die Fisher-Information ist

$$I(\vartheta_0) = \sum_{i=1}^n \text{Var}_{\vartheta_0} \left(\mathbf{1}(X_i - \vartheta_0 > 0) - \mathbf{1}(X_i - \vartheta_0 < 0) \right) \sigma^{-2} = n\sigma^{-2}.$$

Beachte, dass der eher seltene Fall vorliegt, dass die Fisher-Information nicht vom unbekanntem Parameter abhängt.

- (b) Es sei $f(x) = \frac{1}{2}\Gamma(a)^{-1}|x|^{a-1}e^{-|x|}$, $x \in \mathbb{R}$, eine zweiseitige $\Gamma(a, 1)$ -Dichte für festes $a > 0$ und X_1, \dots, X_n eine gemäß $f(\bullet - \vartheta)$ -verteilte mathematische Stichprobe mit $\vartheta \in \mathbb{R}$ unbekannt. Bemerke, dass sich Beispiel (a) mit $\sigma = 1$ im Spezialfall $a = 1$ ergibt. Dann ist

$$L_{\vartheta_0}(\vartheta) = \left(\prod_{i=1}^n \frac{|X_i - \vartheta|}{|X_i - \vartheta_0|} \right)^{a-1} \exp \left(- \sum_{i=1}^n (|X_i - \vartheta| - |X_i - \vartheta_0|) \right),$$

$$\dot{L}_{\vartheta_0}(\vartheta_0) = \sum_{i=1}^n \left(((a-1)|X_i - \vartheta_0|^{-1} - 1) \text{sgn}(\vartheta_0 - X_i) \right).$$

Wegen der exponentiell abfallenden Dichte gilt $L_{\vartheta_0}(\vartheta) \in L^2(\mathbb{P}_{\vartheta_0})$ genau dann, wenn $\int_{[-K, K]^n} \left(\prod_{i=1}^n \frac{|x_i - \vartheta|^2}{|x_i - \vartheta_0|} \right)^{a-1} dx < \infty$ für alle $K > 0$, also genau dann, wenn $2(a-1) > -1$ und $a-1 < 1$, also $a \in (1/2, 2)$. $\dot{L}_{\vartheta_0}(\vartheta_0)$ liegt genau dann in $L^2(\mathbb{P}_{\vartheta_0})$, wenn $a-3 > -1$ oder $a = 1$, also $a \in \{1\} \cup (2, \infty)$ gilt. Dieses Modell ist demnach im obigen Sinn $L^2(\mathbb{P}_{\vartheta_0})$ -differenzierbar genau im Fall $a = 1$, jedoch ist es für alle $a \in \{1\} \cup (2, \infty)$ Hellinger-differenzierbar. Allgemein erfordert Hellinger-Differenzierbarkeit in einem solchen Lokationsmodell, dass jede Nullstelle x_0 von f eine Ordnung größer eins besitzt im Sinn von $\limsup_{x \rightarrow x_0} f(x)|x - x_0|^{-\gamma} < \infty$ für ein $\gamma > 1$.

2.36 Satz (Cramér-Rao-Schranke). *Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_{\vartheta})_{\vartheta \in \Theta})$ mit $\Theta \subseteq \mathbb{R}^k$ ein statistisches Modell, $g : \Theta \rightarrow \mathbb{R}$ besitze bei $\vartheta_0 \in \text{int}(\Theta)$ die Ableitung $\dot{g}(\vartheta_0)$ und \hat{g} sei ein erwartungstreuer Schätzer von $g(\vartheta)$. Für alle ϑ in einer Umgebung von ϑ_0 gelte $\mathbb{P}_{\vartheta} \ll \mathbb{P}_{\vartheta_0}$ sowie die $L^2(\mathbb{P}_{\vartheta_0})$ -Differenzierbarkeit der Likelihoodfunktion $L_{\vartheta_0}(\vartheta) := \frac{d\mathbb{P}_{\vartheta}}{d\mathbb{P}_{\vartheta_0}}$ bei ϑ_0 . Falls die Fisher-Informationsmatrix $I(\vartheta_0)$ strikt positiv-definit ist, gilt die Cramér-Rao-Ungleichung als untere Schranke für das quadratische Risiko*

$$\mathbb{E}_{\vartheta_0} [(\hat{g} - g(\vartheta_0))^2] = \text{Var}_{\vartheta_0}(\hat{g}) \geq \langle I(\vartheta_0)^{-1} \dot{g}(\vartheta_0), \dot{g}(\vartheta_0) \rangle.$$

Beweis. Zunächst beachte die Hellinger-Differenzierbarkeit des Modells by ϑ_0 und betrachte $\vartheta_h = \vartheta_0 + hv$ mit $h \downarrow 0$ und $v \in \mathbb{R}^k$, $v \neq 0$. Dann folgt aus der Chapman-Robbins-Ungleichung, $L_{\vartheta_0}(\vartheta_0) = 1 = \mathbb{E}_{\vartheta_0}[L_{\vartheta_0}(\vartheta)]$ und $\dot{\ell}(\vartheta_0) = \dot{L}_{\vartheta_0}(\vartheta_0)$

$$\mathbb{E}_{\vartheta_0} \left[(\hat{g} - g(\vartheta_0))^2 \right] \geq \limsup_{h \downarrow 0} \frac{((g(\vartheta_h) - g(\vartheta_0))/h)^2}{\mathbb{E}_{\vartheta_0} [((L_{\vartheta_0}(\vartheta_h) - L_{\vartheta_0}(\vartheta_0))/h)^2]} = \frac{(\langle \dot{g}(\vartheta_0), v \rangle)^2}{\langle I(\vartheta_0)v, v \rangle}.$$

Das Supremum der rechten Seite über Richtungen v wird bei $v = I(\vartheta_0)^{-1} \dot{g}(\vartheta_0)$ angenommen, was die Behauptung zeigt. \square

2.37 Bemerkungen.

- (a) Im eindimensionalen Fall $k = 1$ ist die Cramér-Rao-Schranke gerade $\frac{(\dot{g}(\vartheta_0))^2}{I(\vartheta_0)}$ und insbesondere $\frac{1}{I(\vartheta_0)}$ für die Identität $g(\vartheta) = \vartheta$.
- (b) Die Cramér-Rao-Schranke zeigt, dass es umso schwieriger ist $g(\vartheta)$ zu schätzen, je stärker g variiert (d.h. bei großem $|\dot{g}(\vartheta)|$) und je kleiner die Fisher-Information ist. Aus der Definition sieht man, dass die Fisher-Information klein ist, wenn die Likelihood-Funktion wenig variiert, also die Verteilung der Beobachtungen \mathbb{P}_ϑ sehr nahe bei \mathbb{P}_{ϑ_0} liegt für ϑ nahe bei ϑ_0 .
- (c) Ist \hat{g} kein erwartungstreuer Schätzer von $g(\vartheta)$, so doch von $\gamma(\vartheta) := \mathbb{E}_\vartheta[\hat{g}]$ (so existent). Mit der Bias-Varianz-Zerlegung liefert die Cramér-Rao-Ungleichung bei Existenz von $\dot{\gamma}(\vartheta_0)$ für diesen Fall

$$\mathbb{E}_{\vartheta_0}[(\hat{g} - g(\vartheta_0))^2] \geq (g(\vartheta_0) - \gamma(\vartheta_0))^2 + \langle I(\vartheta_0)^{-1} \dot{\gamma}(\vartheta_0), \dot{\gamma}(\vartheta_0) \rangle.$$

Beachte dazu auch, dass erwartungstreue Schätzer von $g(\vartheta)$ nicht existieren müssen bzw. oftmals keine weiteren erstrebenswerten Eigenschaften besitzen.

2.38 Lemma. *Bildet (\mathbb{P}_ϑ) eine Exponentialfamilie in T mit natürlichem Parameterbereich Θ , so ist (\mathbb{P}_ϑ) im Innern von Θ L^2 - und Hellinger-differenzierbar mit Fisher-Information $I(\vartheta) = \dot{A}(\vartheta)$ (Notation aus Satz 2.11).*

Sofern $I(\vartheta_0)$ strikt positiv-definit ist, erreicht T_i , $i = 1, \dots, k$, als erwartungstreuer Schätzer von $g_i(\vartheta) = \mathbb{E}_\vartheta[T_i]$ die Cramér-Rao-Schranke (ist Cramér-Rao-effizient) bei $\vartheta_0 \in \text{int}(\Theta)$.

Beweis. Nach Satz 2.11 gilt $g(\vartheta) = E_\vartheta[T] = \dot{A}(\vartheta)$ und $\text{Cov}_\vartheta(T) = \ddot{A}(\vartheta)$ (Kovarianzmatrix). Andererseits ist die Loglikelihoodfunktion $\ell_{\vartheta_0}(\vartheta) = \langle \vartheta - \vartheta_0, T \rangle - (A(\vartheta) - A(\vartheta_0))$, so dass die Scorefunktion $\dot{\ell}_{\vartheta_0}(\vartheta) = T - \dot{A}(\vartheta)$ im klassischen Sinn existiert und

$$I(\vartheta) = \mathbb{E}_\vartheta[(\dot{\ell}(\vartheta))(\dot{\ell}(\vartheta))^\top] = \text{Var}_\vartheta(T) = \ddot{A}(\vartheta)$$

gelten sollte. Da $L_{\vartheta_0}(\vartheta) = \exp(\langle \vartheta - \vartheta_0, T \rangle - A(\vartheta) + A(\vartheta_0))$ klassisch differenzierbar ist, folgt die $L^2(\mathbb{P}_{\vartheta_0})$ -Differenzierbarkeit bei ϑ_0 sofern Integration und Grenzwert vertauscht werden dürfen, was wie in Satz 2.11 nachgewiesen wird und die Korrektheit der obigen Rechnungen bestätigt.

Wegen $\dot{g}_i(\vartheta_0) = (\ddot{A}_{ij}(\vartheta_0))_j =: \ddot{A}_{i\bullet}(\vartheta_0)$ ist die Cramér-Rao-Schranke gerade

$$\langle \ddot{A}(\vartheta_0)^{-1} \ddot{A}_{i\bullet}(\vartheta_0), \ddot{A}_{i\bullet}(\vartheta_0) \rangle = \langle e_i, \ddot{A}_{i\bullet}(\vartheta_0) \rangle = \ddot{A}_{ii}(\vartheta_0),$$

was gleich der Varianz von T_i unter \mathbb{P}_{ϑ_0} ist. □

2.39 Beispiel. Es sei X_1, \dots, X_n eine $N(\mu, \sigma^2)$ -verteilte mathematische Stichprobe mit $\mu \in \mathbb{R}$ unbekannt und $\sigma > 0$ bekannt. Zur erwartungstreuen Schätzung von μ betrachte $\hat{\mu} = \bar{X}$. Dann gilt $\text{Var}_\mu(\hat{\mu}) = \sigma^2/n$ sowie für die Fisher-Information $I(\mu) = n/\sigma^2$ (beachte $A(\mu) = \frac{n\mu^2}{2\sigma^2}$, $\ddot{A}(\mu) = n/\sigma^2$).

Also ist $\hat{\mu}$ effizient im Sinne der Cramér-Rao-Ungleichung. Um nun μ^2 zu schätzen, betrachte den erwartungstreuen (!) Schätzer $\widehat{\mu^2} = (\bar{X})^2 - \sigma^2/n$. Es gilt $\text{Var}_\mu(\widehat{\mu^2}) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}$, während die Cramér-Rao-Ungleichung die untere Schranke $\frac{4\mu^2\sigma^2}{n}$ liefert. Damit ist $\widehat{\mu^2}$ nicht Cramér-Rao-effizient. Allerdings ist \bar{X} eine suffiziente und vollständige Statistik, so dass der Satz von Lehmann-Scheffé zeigt, dass $\widehat{\mu^2}$ minimale Varianz unter allen erwartungstreuen Schätzern besitzt. Demnach ist die Cramér-Rao-Schranke hier nicht scharf.

2.40 Bemerkung. In der Tat wird die Cramér-Rao-Schranke nur erreicht, wenn (\mathbb{P}_ϑ) eine Exponentialfamilie in T bildet und $g(\vartheta) = \mathbb{E}_\vartheta[T]$ oder eine lineare Funktion davon zu schätzen ist. Wegen der Vollständigkeit der Statistik T könnte man in diesen Fällen alternativ auch mit dem Satz von Lehmann-Scheffé argumentieren. Später werden wir sehen, dass in allgemeineren Modellen immerhin asymptotisch die Cramér-Rao-Schranke erreichbar ist.

2.41 Lemma. *Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ mit $\Theta \subseteq \mathbb{R}^k$ ein bei $\vartheta_0 \in \Theta$ Hellinger-differenzierbares statistisches Modell. Dann ist die Likelihood-Funktion L im $L^1(\mu)$ -Sinn differenzierbar mit Ableitung $\dot{\ell}(\vartheta)L(\vartheta)$, und es gilt $\mathbb{E}_{\vartheta_0}[\dot{\ell}(\vartheta_0)] = 0$.*

Beweis. Betrachte den Zähler im Kriterium für $L^1(\mu)$ -Differenzierbarkeit mit $\dot{L}(\vartheta_0) = \dot{\ell}(\vartheta_0)L(\vartheta_0)$:

$$\begin{aligned} & \left\| L(\vartheta) - L(\vartheta_0) - \langle \dot{\ell}(\vartheta_0), \vartheta - \vartheta_0 \rangle L(\vartheta_0) \right\|_{L^1(\mu)} \\ & \leq \left\| \left(\sqrt{L(\vartheta)} - \sqrt{L(\vartheta_0)} - \frac{1}{2} \langle \dot{\ell}(\vartheta_0), \vartheta - \vartheta_0 \rangle \sqrt{L(\vartheta_0)} \right) \left(\sqrt{L(\vartheta)} + \sqrt{L(\vartheta_0)} \right) \right\|_{L^1(\mu)} \\ & \quad + \frac{1}{2} \left\| \left(\langle \dot{\ell}(\vartheta_0), \vartheta - \vartheta_0 \rangle \sqrt{L(\vartheta_0)} \right) \left(\sqrt{L(\vartheta)} - \sqrt{L(\vartheta_0)} \right) \right\|_{L^1(\mu)}. \end{aligned}$$

Im ersten Ausdruck konvergiert der erste Faktor nach Voraussetzung in $L^2(\mu)$ mit der Ordnung $o(|\vartheta - \vartheta_0|)$ gegen Null und der zweite Faktor in $L^2(\mu)$ gegen $2\sqrt{L(\vartheta_0)}$. Mit der Cauchy-Schwarz-Ungleichung folgt also, dass dieser Ausdruck von der Ordnung $o(|\vartheta - \vartheta_0|)$ ist. Im zweiten Ausdruck besitzt der erste Faktor eine $L^2(\mu)$ -Norm der Ordnung $O(|\vartheta - \vartheta_0|)$, während der zweite Faktor in $L^2(\mu)$ gegen Null konvergiert. Damit ist der gesamte Term von der Ordnung $o(|\vartheta - \vartheta_0|)$ und L somit $L^1(\mu)$ -differenzierbar bei ϑ_0 .

Aus L^1 -Konvergenz folgt Konvergenz der entsprechenden Integrale. Wegen $\int (L(\vartheta, x) - L(\vartheta_0, x)) d\mu(x) = 1 - 1 = 0$ schließen wir durch Einsetzen von $\vartheta = \vartheta_0 + he_i$ ($h \rightarrow 0$, e_i i -ter Einheitsvektor) $0 = \int \langle \dot{\ell}(\vartheta_0), e_i \rangle L(\vartheta_0) d\mu(x) = \mathbb{E}_{\vartheta_0}[\dot{\ell}_i(\vartheta_0)]$ für alle $i = 1, \dots, k$. \square

2.42 Lemma. *Es seien X_1, \dots, X_n Beobachtungen aus unabhängigen Hellinger-differenzierbaren Modellen $\mathcal{E}_1, \dots, \mathcal{E}_n$ mit derselben Parametermenge $\Theta \subseteq \mathbb{R}^k$. Bezeichnet I_j die entsprechende Fisher-Information, erzeugt von der Beobachtung X_j , so ist das Produktmodell, erzeugt von X_1, \dots, X_n , Hellinger-differenzierbar mit Fisher-Information*

$$\forall \vartheta \in \Theta : I(\vartheta) = \sum_{j=1}^n I_j(\vartheta).$$

Beweis. Nach Annahme sind die entsprechenden Likelihoodfunktionen L_1, \dots, L_n bezüglich der dominierenden Maße μ_1, \dots, μ_n Hellinger-differenzierbar mit Score-Funktionen $\dot{\ell}_1, \dots, \dot{\ell}_n$. Also ist auch die gemeinsame Likelihoodfunktion $L(\vartheta, x) = \prod_{j=1}^n L_j(\vartheta, x_j)$ bezüglich $\mu = \mu_1 \otimes \dots \otimes \mu_n$ Hellinger-differenzierbar mit Score-Funktion $\dot{\ell}(\vartheta, x) = \sum_{j=1}^n \dot{\ell}_j(\vartheta, x_j)$, wie für $n = 2$ mit dem Satz von Fubini folgt:

$$\begin{aligned} & \left\| \sqrt{L_1(\vartheta)L_2(\vartheta)} - \sqrt{L_1(\vartheta_0)L_2(\vartheta_0)} - \frac{1}{2} \langle \dot{\ell}_1(\vartheta) + \dot{\ell}_2(\vartheta), \vartheta - \vartheta_0 \rangle \sqrt{L_1(\vartheta_0)L_2(\vartheta_0)} \right\|_{L^2(\mu)} \\ & \leq \left\| \sqrt{L_1(\vartheta_0)} \right\|_{L^2(\mu_1)} \left\| \sqrt{L_2(\vartheta)} - \sqrt{L_2(\vartheta_0)} - \frac{1}{2} \langle \dot{\ell}_2(\vartheta), \vartheta - \vartheta_0 \rangle \sqrt{L_2(\vartheta_0)} \right\|_{L^2(\mu_2)} \\ & + \left\| \sqrt{L_2(\vartheta_0)} \right\|_{L^2(\mu_2)} \left\| \sqrt{L_1(\vartheta)} - \sqrt{L_1(\vartheta_0)} - \frac{1}{2} \langle \dot{\ell}_1(\vartheta), \vartheta - \vartheta_0 \rangle \sqrt{L_1(\vartheta_0)} \right\|_{L^2(\mu_1)} \\ & + \left\| \sqrt{L_2(\vartheta)} - \sqrt{L_2(\vartheta_0)} \right\|_{L^2(\mu_2)} \left\| \sqrt{L_1(\vartheta)} - \sqrt{L_1(\vartheta_0)} \right\|_{L^2(\mu_1)} \\ & = o(|\vartheta - \vartheta_0|) + o(|\vartheta - \vartheta_0|) + O(|\vartheta - \vartheta_0|^2). \end{aligned}$$

Für allgemeine $n \geq 2$ verwende vollständige Induktion.

Wegen Unabhängigkeit der X_1, \dots, X_n sowie $\mathbb{E}_\vartheta[\dot{\ell}_j(\vartheta, X_j)] = 0$ gilt daher

$$\begin{aligned} & \mathbb{E}_\vartheta \left[\dot{\ell}(\vartheta, (X_1, \dots, X_n)) \dot{\ell}(\vartheta, (X_1, \dots, X_n))^\top \right] \\ & = \mathbb{E}_\vartheta \left[\left(\sum_{j=1}^n \dot{\ell}_j(\vartheta, X_j) \right) \left(\sum_{j=1}^n \dot{\ell}_j(\vartheta, X_j)^\top \right) \right] \\ & = \sum_{j,m=1}^n \mathbb{E}_\vartheta \left[\dot{\ell}_j(\vartheta, X_j) \dot{\ell}_m(\vartheta, X_m)^\top \right] = \sum_{j=1}^n \mathbb{E}_\vartheta \left[\dot{\ell}_j(\vartheta, X_j) \dot{\ell}_j(\vartheta, X_j)^\top \right]. \end{aligned}$$

□

2.43 Bemerkung. Das Lemma zeigt also, dass die Fisher-Information unter unabhängigen Beobachtungen additiv ist, so dass sie bei einer mathematischen Stichprobe gerade gleich dem Stichprobenumfang n mal der Fisher-Information bei einer Beobachtung ist.

2.6 Äquivarianz

Mit dem Satz von Lehmann-Scheffé und der Cramér-Rao-Schranke konnten wir beste erwartungstreue Schätzer unter quadratischem Risiko verstehen. Statt Erwartungstreue ist es oft angemessen, gewisse Invarianzeigenschaften von Schätzern zu fordern. Dies führt auf den Begriff der Äquivarianz, der allgemein für Gruppenoperationen existiert, aber hier nur im Fall einer einparametrischen Translationsinvarianz dargestellt wird.

2.44 Definition. Im Lokationsmodell beobachten wir einen \mathbb{R}^n -wertigen Zufallsvektor $X = (X_1, \dots, X_n)$, welcher eine gemeinsame Verteilung \mathbb{P}_ϑ mit Lebesgue-Dichte $f(x_1 - \vartheta, \dots, x_n - \vartheta)$, $(x_1, \dots, x_n) \in \mathbb{R}^n$, besitzt, wobei f bekannt und $\vartheta \in \mathbb{R}$ ein unbekannter Lokationsparameter ist.

2.45 Bemerkung. Gilt $\mathbb{E}_0[X_i] = 0$, so folgt $\mathbb{E}_\vartheta[X_i] = \vartheta$, $i = 1, \dots, n$, und der unbekannte Parameter ϑ ist gerade der Erwartungswert der Beobachtungen.

2.46 Definition. In Lokationsmodell heißt ein Schätzer $\hat{\vartheta} : \mathbb{R}^n \rightarrow \mathbb{R}$ äquivariant, wenn gilt

$$\forall x_1, \dots, x_n \in \mathbb{R} \quad \forall \vartheta \in \mathbb{R} : \hat{\vartheta}(x_1 + \vartheta, \dots, x_n + \vartheta) = \hat{\vartheta}(x_1, \dots, x_n) + \vartheta$$

und die Verlustfunktion l invariant ist, also $l(\vartheta, a) = \ell(\vartheta - a)$ gilt mit $\ell : \mathbb{R} \rightarrow [0, \infty)$ messbar.

2.47 Lemma. Ist $\hat{\vartheta}$ äquivarianter Schätzer und l eine invariante Verlustfunktion, so besitzt $\hat{\vartheta}$ konstantes Risiko: $R(\vartheta, \hat{\vartheta}) = R(0, \hat{\vartheta})$ für alle $\vartheta \in \mathbb{R}$.

Beweis. Es gilt

$$\begin{aligned} R(\vartheta, \hat{\vartheta}) &= \mathbb{E}_\vartheta[\ell(\hat{\vartheta}(X_1, \dots, X_n) - \vartheta)] \\ &= \mathbb{E}_0[\ell(\hat{\vartheta}(X_1 + \vartheta, \dots, X_n + \vartheta) - \vartheta)] \\ &= \mathbb{E}_0[\ell(\hat{\vartheta}(X_1, \dots, X_n) + \vartheta - \vartheta)] = R(0, \hat{\vartheta}), \end{aligned}$$

wobei wir in der zweiten Gleichheit benutzen, dass X unter \mathbb{P}_ϑ wie $X + \vartheta$ unter \mathbb{P}_0 verteilt ist, und in der dritten Gleichheit die Äquivarianz von $\hat{\vartheta}$ verwendet wird. \square

2.48 Definition. Ein äquivarianter Schätzer $\hat{\vartheta}$ heißt bester äquivarianter Schätzer (MRIE: minimum risk invariant estimator) im Lokationsmodell, falls

$$R(0, \hat{\vartheta}) = \inf_{\tilde{\vartheta} \text{ äquivariant}} R(0, \tilde{\vartheta})$$

gilt, wobei sich das Infimum über alle äquivarianten Schätzer $\tilde{\vartheta}$ erstreckt.

2.49 Satz. Es sei $\hat{\vartheta}_0$ ein äquivarianter Schätzer und setze $T(x_1, \dots, x_n) := (x_1 - x_n, \dots, x_{n-1} - x_n) \in \mathbb{R}^{n-1}$, $(x_1, \dots, x_n) \in \mathbb{R}^n$.

- (a) Ein Schätzer $\hat{\vartheta}$ ist genau dann äquivariant, wenn es eine messbare Funktion $u : \mathbb{R}^{n-1} \rightarrow \mathbb{R}$ gibt mit $\hat{\vartheta}(x) = \hat{\vartheta}_0(x) - u(T(x))$, $x \in \mathbb{R}^n$.
- (b) Es gelte zusätzlich $\mathbb{E}_0[\hat{\vartheta}_0^2] < \infty$. Dann ist durch $\hat{\vartheta} := \hat{\vartheta}_0 - \mathbb{E}_0[\hat{\vartheta}_0 | T]$ ein bester äquivarianter Schätzer bei quadratischem Verlust gegeben.

Beweis. Für (a) folgt durch Einsetzen, dass $\hat{\vartheta}_0 + u(T)$ äquivariant ist. Andererseits folgt aus der Äquivarianz von $\hat{\vartheta}$ und $\hat{\vartheta}_0$

$$\forall x_1, \dots, x_n, \vartheta \in \mathbb{R} : (\hat{\vartheta} - \hat{\vartheta}_0)(x_1 + \vartheta, \dots, x_n + \vartheta) = (\hat{\vartheta} - \hat{\vartheta}_0)(x_1, \dots, x_n).$$

Mit $\vartheta = -x_n$ und $u(t_1, \dots, t_{n-1}) = (\hat{\vartheta} - \hat{\vartheta}_0)(t_1, \dots, t_{n-1}, 0)$ folgt die behauptete Darstellung in (a).

Jeder äquivariante Schätzer $\tilde{\vartheta}$ mit der Darstellung in (a) erfüllt $R(0, \tilde{\vartheta}) = \mathbb{E}_0[(\hat{\vartheta}_0 - u(T))^2]$ für eine messbare Funktion u von T . Nach Charakterisierung der bedingten Erwartung wird dies durch $u(T) = \mathbb{E}_0[\hat{\vartheta}_0 | T]$ minimiert und nach (a) ist $\hat{\vartheta} = \hat{\vartheta}_0 - \mathbb{E}_0[\hat{\vartheta}_0 | T]$ wiederum äquivariant. \square

2.50 Korollar. *Im Lokationsmodell gelte $\mathbb{E}_0[X_n^2] < \infty$. Dann ist der beste äquivariante Schätzer bezüglich quadratischem Risiko gegeben durch den Pitman-Schätzer*

$$\hat{\vartheta} = \frac{\int_{-\infty}^{\infty} z f(X_1 - z, \dots, X_n - z) dz}{\int_{-\infty}^{\infty} f(X_1 - z, \dots, X_n - z) dz}.$$

Beweis. Der Schätzer $\hat{\vartheta}_0 = X_n$ ist äquivariant mit $\mathbb{E}_0[\hat{\vartheta}_0^2] < \infty$ nach Voraussetzung. Aus dem Satz folgt daher, dass $\hat{\vartheta} = X_n - \mathbb{E}_0[X_n | T]$ bester äquivarianter Schätzer ist. Nach dem Dichtetransformationssatz besitzt (T, X_n) unter \mathbb{P}_0 die gemeinsame Lebesgue-dichte

$$f^{(T, X_n)}(t, x_n) = f(t_1 + x_n, \dots, t_{n-1} + x_n, x_n), \quad t \in \mathbb{R}^{n-1}, x_n \in \mathbb{R}.$$

Nach der Bayesformel erhalten wir die bedingte Dichte

$$f^{X_n | T=t}(x_n) = \frac{f(t_1 + x_n, \dots, t_{n-1} + x_n, x_n)}{\int_{-\infty}^{\infty} f(t_1 + \xi, \dots, t_{n-1} + \xi, \xi) d\xi}, \quad x_n \in \mathbb{R} \quad \text{für } \mathbb{P}_0^T\text{-f.a. } t.$$

Wir erhalten die bedingte Erwartung durch Integration und substituieren $z = X_n - x_n$ (für jede Realisierung von X_n):

$$\begin{aligned} X_n - \mathbb{E}_0[X_n | T] &= \frac{\int_{-\infty}^{\infty} (X_n - x_n) f(T_1 + x_n, \dots, T_{n-1} + x_n, x_n) dx_n}{\int_{-\infty}^{\infty} f(T_1 + x_n, \dots, T_{n-1} + x_n, x_n) dx_n} \\ &= \frac{\int_{-\infty}^{\infty} z f(T_1 + X_n - z, \dots, T_{n-1} + X_n - z, X_n - z) dz}{\int_{-\infty}^{\infty} f(T_1 + X_n - z, \dots, T_{n-1} + X_n - z, X_n - z) dz} \\ &= \frac{\int_{-\infty}^{\infty} z f(X_1 - z, \dots, X_{n-1} - z, X_n - z) dz}{\int_{-\infty}^{\infty} f(X_1 - z, \dots, X_{n-1} - z, X_n - z) dz}. \end{aligned}$$

Dies zeigt die Behauptung. □

2.51 Beispiel. Im Lokationsmodell mit Produktdichte $f(x) = \prod_{i=1}^n f_1(x_i)$ kann man den Pitman-Schätzer $\hat{\vartheta}$ in folgenden Fällen leicht berechnen (Übung!):

- (a) $f_1(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$, $\hat{\vartheta} = \bar{X}$;
- (b) $f_1(x) = a^{-1} \mathbf{1}_{[-\frac{a}{2}, \frac{a}{2}]}(x)$ für $a > 0$, $\hat{\vartheta} = \frac{X_{(1)} + X_{(n)}}{2}$;
- (c) $f_1(x) = \lambda e^{-\lambda(x+1)} \mathbf{1}_{[-1, \infty)}(x)$ für $\lambda > 0$, $\hat{\vartheta} = X_{(1)} + 1 - \frac{1}{n\lambda}$.

Für $\sigma = 1$, $a = \sqrt{3}$ und $\lambda = 1$ gilt in allen drei Fällen, dass f_1 Erwartungswert 0 und Varianz 1 besitzt. Weil \bar{X} äquivariant ist jeweils mit $\mathbb{E}_{\vartheta}[(\bar{X} - \vartheta)^2] = \frac{1}{n}$, muss das quadratische Risiko von $\hat{\vartheta}$ in (b) und (c) kleiner als $\frac{1}{n}$ sein. Allgemein ist unter allen Dichten f_1 mit Erwartungswert μ und endlicher Varianz σ^2 das quadratische Risiko eines besten äquivarianten Schätzers maximal bei der Normalverteilung $N(\mu, \sigma^2)$. Die Normalverteilung ist also ungünstigste Verteilung in dieser Klasse.

2.52 Bemerkungen.

- (a) Man kann die Bedingung $\mathbb{E}_0[X_n^2] < \infty$ durch die Existenz eines äquivarianten Schätzers mit endlichem quadratischen Risiko ersetzen.
- (b) Der Pitman-Schätzer kann auch als *uneigentlicher Bayes-Schätzer* verstanden werden mit dem translationsinvarianten Lebesguemaß als a-priori-Verteilung für ϑ . Die a-posteriori-Dichte ist dann

$$f^{T|X=x}(\vartheta) = \frac{f(x_1 - \vartheta, \dots, x_n - \vartheta)}{\int_{-\infty}^{\infty} f(x_1 - \vartheta', \dots, x_n - \vartheta') d\vartheta'}, \quad \vartheta \in \mathbb{R}.$$

Der Bayesschätzer bezüglich quadratischem Risiko ergibt sich als bedingte Erwartung

$$\hat{\vartheta} = \frac{\int_{-\infty}^{\infty} \vartheta f(X_1 - \vartheta, \dots, X_n - \vartheta) d\vartheta}{\int_{-\infty}^{\infty} f(X_1 - \vartheta', \dots, X_n - \vartheta') d\vartheta'},$$

was gerade der Pitman-Schätzer ist.

- (c) Der Pitman-Schätzer ist minimax bezüglich quadratischem Risiko. Wie (b) suggeriert, kann dies über das Bayesrisiko bei a-priori-Verteilung $\pi = U([-R, R])$ und den Grenzübergang $R \rightarrow \infty$ gezeigt werden.

3 Asymptotische Schätztheorie

3.1 Momentenschätzer

3.1 Definition. Es seien $(\mathcal{X}^n, \mathcal{F}^{\otimes n}, (\mathbb{P}_{\vartheta}^{\otimes n})_{\vartheta \in \Theta})$ ein statistisches (Produkt-)Modell und $g(\vartheta)$ mit $g : \Theta \rightarrow \mathbb{R}^p$ ein abgeleiteter Parameter. Ferner sei $\psi = (\psi_1, \dots, \psi_q) : \mathcal{X} \rightarrow \mathbb{R}^q$ derart, dass $\varphi(\vartheta) := \mathbb{E}_{\vartheta}[\psi]$ für alle $\vartheta \in \Theta$ existiert. Gibt es nun eine messbare Funktion $G : \mathbb{R}^q \rightarrow \mathbb{R}^p$ mit $G \circ \varphi = g$, so heißt $\hat{g}_n := G(\frac{1}{n} \sum_{i=1}^n \psi(x_i))$ (verallgemeinerter) Momentenschätzer für $g(\vartheta)$ mit Momentenfunktionen ψ_1, \dots, ψ_q .

3.2 Beispiele.

- (a) Es sei $X_1, \dots, X_n \sim \text{Exp}(\lambda)$ eine mathematische Stichprobe mit $\lambda > 0$ unbekannt. Betrachte die klassische Momentenfunktion $\psi(x) = x^k$ für ein $k \in \mathbb{N}$.
Mit $g(\lambda) = \lambda$ und $\varphi(\lambda) = \mathbb{E}_{\lambda}[X_i^k] = \lambda^{-k} k!$ ergibt sich $G(x) = (k!/x)^{1/k}$ und als Momentenschätzer für λ

$$\hat{\lambda}_{k,n} := \left(\frac{k!}{\frac{1}{n} \sum_{i=1}^n X_i^k} \right)^{1/k}.$$

- (b*) Betrachte einen autoregressiven Prozess der Ordnung 1 (AR(1)-Prozess):

$$X_n = aX_{n-1} + \varepsilon_n, \quad n \geq 1,$$

mit (ε_n) i.i.d., $\mathbb{E}[\varepsilon_n] = 0$, $\text{Var}(\varepsilon_n) = \sigma^2 < \infty$ und $X_0 = x_0 \in \mathbb{R}$. Um a zu schätzen, betrachte folgende Identität für das bedingte gemeinsame Moment:

$$\mathbb{E}[X_{n-1}X_n \mid \varepsilon_1, \dots, \varepsilon_{n-1}] = aX_{n-1}^2.$$

Dies führt auf eine modifizierte Momentenmethode als Schätzidee (Yule-Walker-Schätzer):

$$\hat{a}_n := \frac{\frac{1}{n} \sum_{k=1}^n X_{k-1}X_k}{\frac{1}{n} \sum_{k=1}^n X_{k-1}^2} = a + \frac{\sum_{k=1}^n X_{k-1}\varepsilon_k}{\sum_{k=1}^n X_{k-1}^2}.$$

Im Fall $|a| < 1$ kann man mit Hilfe des Ergodensatzes auf die Konsistenz von \hat{a}_n für $n \rightarrow \infty$ schließen. Allgemeiner zeigt man leicht, dass $M_n := \sum_{k=1}^n X_{k-1}\varepsilon_k$ ein Martingal bezüglich $\mathcal{F}_n := \sigma(\varepsilon_1, \dots, \varepsilon_n)$ ist mit quadratischer Variation $\langle M \rangle_n := \sum_{k=1}^n X_{k-1}^2$. Das starke Gesetz der großen Zahlen für L^2 -Martingale liefert daher die Konsistenz

$$\hat{a}_n = a + \frac{M_n}{\langle M \rangle_n} \xrightarrow{\text{f.s.}} a.$$

3.3 Lemma. *Ist G stetig beim Momentenschätzer $\hat{g}_n = G(\frac{1}{n} \sum_{i=1}^n \psi(x_i))$, so ist \hat{g}_n ein (stark) konsistenter Schätzer von $g(\vartheta)$, d.h. $\lim_{n \rightarrow \infty} \hat{g}_n = g(\vartheta)$ $\mathbb{P}_\vartheta^{\otimes \mathbb{N}}$ -f.s.*

Beweis. Nach dem starken Gesetz der großen Zahlen gilt wegen der Stetigkeit von G $\mathbb{P}_\vartheta^{\otimes \mathbb{N}}$ -fast sicher:

$$\lim_{n \rightarrow \infty} G\left(\frac{1}{n} \sum_{i=1}^n \psi(X_i)\right) = G\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \psi(X_i)\right) = G(\varphi(\vartheta)) = g(\vartheta).$$

□

3.4 Satz (Δ -Methode). *Es seien (X_n) eine Folge von Zufallsvektoren im \mathbb{R}^k , $\sigma_n > 0$, $\sigma_n \rightarrow 0$, $\vartheta_0 \in \mathbb{R}^k$ sowie $\Sigma \in \mathbb{R}^{k \times k}$ positiv semi-definit und es gelte*

$$\sigma_n^{-1}(X_n - \vartheta_0) \xrightarrow{d} N(0, \Sigma).$$

Ist $f : \mathbb{R}^k \rightarrow \mathbb{R}$ in einer Umgebung von ϑ_0 stetig differenzierbar mit Gradienten \dot{f} , so folgt

$$\sigma_n^{-1}(f(X_n) - f(\vartheta_0)) \xrightarrow{d} N(0, \langle \Sigma \dot{f}(\vartheta_0), \dot{f}(\vartheta_0) \rangle),$$

wobei $N(0, 0)$ gegebenenfalls als Punktmaß δ_0 in der Null zu verstehen ist.

Beweis. Nach dem Lemma von Slutsky (vgl. Stochastik II) gilt $X_n - \vartheta_0 = \sigma_n \frac{X_n - \vartheta_0}{\sigma_n} \xrightarrow{d} 0$ und somit (Stochastik I) $X_n \xrightarrow{\mathbb{P}} \vartheta_0$ für $n \rightarrow \infty$. Eine Taylorentwicklung ergibt

$$f(X_n) = f(\vartheta_0) + \langle \dot{f}(\vartheta_0), X_n - \vartheta_0 \rangle + R_n$$

mit $R_n/|X_n - \vartheta_0| \rightarrow 0$ für $X_n \rightarrow \vartheta_0$ jeweils bezüglich fast sicherer und damit auch jeweils bezüglich stochastischer Konvergenz (Teilteilfolgenargument; Stochastik II). Wiederum mittels Slutsky-Lemma folgt

$$\frac{R_n}{\sigma_n} = \left| \frac{X_n - \vartheta_0}{\sigma_n} \right| \frac{R_n}{|X_n - \vartheta_0|} \xrightarrow{d} 0$$

und also auch bezüglich stochastischer Konvergenz. Eine dritte Anwendung des Slutsky-Lemmas gibt daher

$$\sigma_n^{-1}(f(X_n) - f(\vartheta_0)) = \langle \dot{f}(\vartheta_0), \sigma_n^{-1}(X_n - \vartheta_0) \rangle + \sigma_n^{-1}R_n \xrightarrow{d} N(0, \dot{f}(\vartheta_0)^\top \Sigma \dot{f}(\vartheta_0));$$

denn es gilt $\langle \dot{f}(\vartheta_0), \sigma_n^{-1}(X_n - \vartheta_0) \rangle \xrightarrow{d} \langle \dot{f}(\vartheta_0), \Sigma^{1/2}Z \rangle \sim N(0, \langle \Sigma \dot{f}(\vartheta_0), \dot{f}(\vartheta_0) \rangle)$ mit $Z \sim N(0, E_k)$. \square

3.5 Beispiel. Aus einer mathematischen Stichprobe $X_1, \dots, X_n \sim \text{Pois}(\lambda)$ bestimmt man den UMVU-Schätzer $\hat{\lambda}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Nach dem zentralen Grenzwertsatz gilt $\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{d} N(0, \lambda)$ unter $\mathbb{P}_\lambda^{\otimes \mathbb{N}}$. Um asymptotisch ein Konfidenzintervall herzuleiten, stört es, dass die asymptotische Varianz vom Parameter selbst abhängt. Betrachtet man nun $f(x) = 2x^{1/2}$ mit $\dot{f}(x) = x^{-1/2}$ in der Δ -Methode, so folgt $\sqrt{n}(2\hat{\lambda}_n^{1/2} - 2\lambda^{1/2}) \xrightarrow{d} N(0, 1)$, so dass $[2\hat{\lambda}_n^{1/2} - n^{-1/2}q_{1-\alpha/2}, 2\hat{\lambda}_n^{1/2} + n^{-1/2}q_{1-\alpha/2}]$ mit dem $(1 - \alpha/2)$ -Quantil $q_{1-\alpha/2}$, $\alpha \in (0, 1)$, von $N(0, 1)$ ein asymptotisches $(1 - \alpha)$ -Konfidenzintervall für $2\lambda^{1/2}$ bildet. Rücktransformation ergibt dann für λ selbst das asymptotische $(1 - \alpha)$ -Konfidenzintervall $[(\hat{\lambda}_n^{1/2} - (4n)^{-1/2}q_{1-\alpha/2})_+^2, (\hat{\lambda}_n^{1/2} + (4n)^{-1/2}q_{1-\alpha/2})^2]$. Die Idee, mittels Δ -Transformation eine asymptotische Varianz unabhängig vom unbekanntem zu erhalten, ist in vielen Situationen sehr fruchtbar und nennt sich Varianz-stabilisierende Transformation.

Alternativ kann man die asymptotische Varianz durch $\hat{\lambda}_n$ konsistent schätzen und mittels Slutsky-Lemma auf $(n/\hat{\lambda}_n)^{1/2}(\hat{\lambda}_n - \lambda) \xrightarrow{d} N(0, 1)$ schließen. Daraus ergibt sich $[\hat{\lambda}_n - (\hat{\lambda}_n/n)^{1/2}q_{1-\alpha/2}, \hat{\lambda}_n + (\hat{\lambda}_n/n)^{1/2}q_{1-\alpha/2}]$ als asymptotisches $(1 - \alpha)$ -Konfidenzintervall. Ein solches über Varianzschätzung erhaltenes Konfidenzintervall hat zwar den Vorteil, symmetrisch um $\hat{\lambda}_n$ zu sein, weist aber im Allgemeinen eine schlechtere Normalapproximation auf als die Konstruktion via Varianz-stabilisierender Transformation, d.h. die Überdeckungswahrscheinlichkeit für festes n weicht stärker von $1 - \alpha$ ab.

3.6 Satz. *Es seien $\vartheta_0 \in \Theta$, $g : \Theta \rightarrow \mathbb{R}$ und für hinreichend großes n existiere der Momentenschätzer $\hat{g}_n = G(\frac{1}{n} \sum_{i=1}^n \psi(x_i))$ mit Momentenfunktionen $\psi_j \in L^2(\mathbb{P}_{\vartheta_0})$, $j = 1, \dots, q$. Betrachte $\text{Var}_{\vartheta_0}(\psi) := (\text{Cov}_{\vartheta_0}(\psi_i, \psi_j))_{i,j=1,\dots,q} \in \mathbb{R}^{q \times q}$. Sofern G in einer Umgebung von $\varphi(\vartheta_0)$ stetig differenzierbar ist, ist \hat{g}_n unter $\mathbb{P}_{\vartheta_0}^{\otimes \mathbb{N}}$ asymptotisch normalverteilt mit Rate $n^{-1/2}$, asymptotischem Mittelwert Null und Varianz $\langle \text{Var}_{\vartheta_0}(\psi) \dot{G}(\varphi(\vartheta_0)), \dot{G}(\varphi(\vartheta_0)) \rangle$:*

$$\sqrt{n}(\hat{g}_n - g(\vartheta_0)) \xrightarrow{d} N(0, \langle \text{Var}_{\vartheta_0}(\psi) \dot{G}(\varphi(\vartheta_0)), \dot{G}(\varphi(\vartheta_0)) \rangle) \text{ (unter } \mathbb{P}_{\vartheta_0}^{\otimes \mathbb{N}}).$$

3.7 Bemerkung. Die Begriffe *asymptotischer Mittelwert* und *asymptotische Varianz* sind leicht irreführend: es gilt nicht notwendigerweise, dass

die Momente von $\sqrt{n}(\hat{g}_n - g(\vartheta_0))$ gegen die entsprechenden Momente von $N(0, \langle \text{Var}_{\vartheta_0}(\psi) \dot{G}(\varphi(\vartheta_0)), \dot{G}(\varphi(\vartheta_0)) \rangle)$ konvergieren (dafür wird gleichgradige Integrierbarkeit benötigt).

Beweis. Nach dem multivariaten zentralen Grenzwertsatz gilt unter $\mathbb{P}_{\vartheta_0}^{\otimes \mathbb{N}}$

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \psi(X_i) - \varphi(\vartheta_0) \right) \xrightarrow{d} N(0, \text{Var}_{\vartheta_0}(\psi)).$$

Die Behauptung folgt daher unmittelbar mit der Δ -Methode (setze $\sigma_n = n^{-1/2}$, $f = G$). \square

3.8 Beispiel. Im Exponentialverteilungsmodell aus Beispiel 3.2 gilt $G'(x) = -(k!/x)^{1/k} (kx)^{-1}$ und $\Sigma(\lambda_0) = \text{Var}_{\lambda_0}(X_i^k) = ((2k)! - (k!)^2) / \lambda_0^{2k}$. Alle Momentenschätzer $\hat{\lambda}_{k,n}$ sind asymptotisch normalverteilt mit Rate $n^{-1/2}$ und Varianz $\sigma_k^2 = \lambda_0^2 k^{-2} ((2k)! / (k!)^2 - 1)$. Da $\hat{\lambda}_{1,n}$ die gleichmäßig kleinste asymptotische Varianz besitzt und auf der suffizienten Statistik \bar{X} basiert, wird dieser Schätzer im Allgemeinen vorgezogen.

3.9 Bemerkung (*). Die Momentenmethode kann unter folgendem allgemeinen Gesichtspunkt verstanden werden: Ist X_1, \dots, X_n eine mathematische Stichprobe mit Werten in \mathbb{R} , so ist die empirische Verteilungsfunktion $F_n(x) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_i \leq x)$ eine suffiziente Statistik und nach dem Satz von Glivenko-Cantelli gilt \mathbb{P}_{ϑ} -f.s. $F_n(x) \rightarrow F_{\vartheta}(x) = \mathbb{P}_{\vartheta}(X_i \leq x)$ gleichmäßig in $x \in \mathbb{R}$. Ist nun $g(\vartheta)$ als Funktional $G(F_{\vartheta}(x), x \in \mathbb{R})$ darstellbar, so verwende die empirische Version $G(F_n(x), x \in \mathbb{R})$ als Schätzer von $g(\vartheta)$. Falls das Funktional G stetig bezüglich der Supremumsnorm ist, so folgt die Konsistenz.

Der Satz von Donsker für empirische Prozesse zeigt $\sqrt{n}(F_n - F_{\vartheta}) \xrightarrow{d} \Gamma_{\vartheta}$ gleichmäßig auf \mathbb{R} mit einem zentrierten Gaußprozess Γ_{ϑ} von der Kovarianzstruktur $\text{Cov}(\Gamma_{\vartheta}(x), \Gamma_{\vartheta}(y)) = F_{\vartheta}(x \wedge y) - F_{\vartheta}(x)F_{\vartheta}(y)$. Ist G ein *Hadamard-differenzierbares* Funktional, so folgt $\sqrt{n}(G(F_n(x), x \in \mathbb{R}) - g(\vartheta)) \xrightarrow{d} \dot{G}(F_{\vartheta})\Gamma_{\vartheta}$ unter \mathbb{P}_{ϑ} , also insbesondere asymptotische Normalverteilung mit Rate $n^{-1/2}$ und explizit bestimmbarer asymptotischer Varianz, siehe z.B. das Buch von van der Vaart für mehr Details.

Als einfaches (lineares) Beispiel sei $g(\vartheta) = \mathbb{E}_{\vartheta}[\psi(X_i)]$ zu schätzen und $X_i \geq 0$ \mathbb{P}_{ϑ} -f.s. Dann folgt informell $G(F_{\vartheta}) = \int_0^{\infty} \psi(x) dF_{\vartheta}(x) = \int_0^{\infty} \psi'(x)(1 - F_{\vartheta}(x)) dx$. Aus der Linearität erhalten wir $\dot{G}(F_{\vartheta})\Gamma_{\vartheta} = \int_0^{\infty} \psi'(x)(-\Gamma_{\vartheta}(x)) dx$. Dies ist normalverteilt mit Erwartungswert Null und Varianz

$$\begin{aligned} & \int_0^{\infty} \int_0^{\infty} \psi'(x)\psi'(y)(F_{\vartheta}(x \wedge y) - F_{\vartheta}(x)F_{\vartheta}(y)) dx dy \\ &= \int_0^{\infty} \int_0^{\infty} \psi(x)\psi(y)\partial_{xy}(F_{\vartheta}(x \wedge y) - F_{\vartheta}(x)F_{\vartheta}(y)) dx dy \\ &= \int_0^{\infty} \psi^2(x) dF_{\vartheta}(x) - \left(\int_0^{\infty} \psi(x) dF_{\vartheta}(x) \right)^2, \end{aligned}$$

was natürlich gerade der Varianz von $G(F_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(X_i)$ entspricht.

3.2 Maximum-Likelihood- und M-Schätzer

3.10 Beispiele.

- (a) Auf dem diskreten Stichprobenraum \mathcal{X} seien Verteilungen $(P_\vartheta)_{\vartheta \in \Theta}$ gegeben. Bezeichnet p_ϑ die zugehörige Zähldichte und ist die Verlustfunktion $l(\vartheta, \rho)$ homogen in $\vartheta \in \Theta$, so ist es für die Schätzung von ϑ plausibel, bei Vorliegen des Versuchsausgangs x für einen Schätzer $\hat{\vartheta}(x)$ denjenigen Parameter $\vartheta \in \Theta$ zu wählen, für den die Wahrscheinlichkeit $p_\vartheta(x)$ des Eintretens von x maximal ist: $\hat{\vartheta}(x) := \operatorname{argmax}_{\vartheta \in \Theta} p_\vartheta(x)$. Dieser Schätzer heißt Maximum-Likelihood-Schätzer (MLE). Bereits im vorliegenden Fall ist weder Existenz noch Eindeutigkeit ohne Weiteres garantiert. Bei Nicht-Eindeutigkeit wählt man einen maximierenden Parameter ϑ nach Belieben aus. Im Fall einer mathematischen Stichprobe $X_1, \dots, X_n \sim \text{Poiss}(\lambda)$ mit $\lambda > 0$ unbekannt, ergibt sich beispielsweise

$$\hat{\lambda} = \operatorname{argmax}_{\lambda > 0} \prod_{i=1}^n \left(e^{-\lambda} \frac{\lambda^{X_i}}{X_i!} \right) = \bar{X}$$

im Fall $\bar{X} > 0$. Ist $\bar{X} = 0$, d.h. $X_1 = \dots = X_n = 0$, so wird das Supremum nur asymptotisch für $\lambda \rightarrow 0$ erreicht. Hier könnte man sich behelfen, indem man $\text{Poiss}(0)$ als Punktmaß in der Null stetig ergänzt.

- (b) Besitzen die Verteilungen \mathbb{P}_ϑ Lebesguedichten f_ϑ , so führt der Maximum-Likelihood-Ansatz analog auf $\hat{\vartheta}(x) = \operatorname{argmax}_{\vartheta \in \Theta} f_\vartheta(x)$. Betrachte die Stichprobe Y der Form $Y = e^X$ mit $X \sim N(\mu, 1)$ mit $\mu \in \mathbb{R}$ unbekannt. Dann ist Y log-normalverteilt, und es gilt

$$\hat{\mu}(Y) = \operatorname{argmax}_{\mu \in \mathbb{R}} \frac{e^{-(\log(Y) - \mu)^2/2}}{\sqrt{2\pi}Y} = \log(Y).$$

Man sieht, dass der MLE invariant unter Parametertransformation ist: bei Beobachtung von $X \sim N(\mu, 1)$ erhält man den MLE $\tilde{\mu}(X) = X$ und Einsetzen von $X = \log(Y)$ führt auf dasselbe Ergebnis. Interessanterweise führt die Momentenmethode unter Benutzung von $\mathbb{E}_\mu[Y] = e^{\mu+1/2}$ auf den Schätzer $\bar{\mu}(Y) = \log(Y) - 1/2$, während $\mathbb{E}_\mu[X] = \mu$ auf $\tilde{\mu}(X) = X$ führt; Momentenschätzer, beruhend auf demselben Moment, sind also im Allgemeinen nicht transformationsinvariant.

3.11 Definition. Es sei $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein von μ dominiertes Modell mit Likelihoodfunktion $L(\vartheta, x)$. Eine Statistik $\hat{\vartheta} : \mathcal{X} \rightarrow \Theta$ (Θ trage eine σ -Algebra \mathcal{F}_Θ) heißt Maximum-Likelihood-Schätzer (MLE) von ϑ , falls $L(\hat{\vartheta}(x), x) = \sup_{\vartheta \in \Theta} L(\vartheta, x)$ für μ -fast alle $x \in \mathcal{X}$ gilt.

3.12 Bemerkung. Der MLE braucht weder zu existieren noch eindeutig zu sein, falls er existiert. Er hängt von der gewählten Version der Radon-Nikodym-Dichte ab; es gibt jedoch häufig eine kanonische Wahl, wie beispielsweise bei stetigen Lebesguedichten. Außerdem ist eine Abänderung auf einer Nullmenge bezüglich aller \mathbb{P}_ϑ irrelevant, weil der Schätzer vor Realisierung des Experiments

festgelegt wird und diese Realisierung damit fast sicher zum selben Schätzwert führen wird.

Bei einer eindeutigen Parametrisierung $\vartheta \mapsto h(\vartheta)$ ergibt sich $\hat{h} := h(\hat{\vartheta})$ als MLE für $h(\vartheta)$.

3.13 Lemma. Für eine natürliche Exponentialfamilie $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ in $T(x)$ ist der MLE $\hat{\vartheta}$ implizit gegeben durch die Momentengleichung $\mathbb{E}_{\hat{\vartheta}(x)}[T] = T(x)$, vorausgesetzt der MLE existiert und $\hat{\vartheta}(x) \in \text{int}(\Theta)$.

Beweis. Schreiben wir die Loglikelihoodfunktion in der Form $\ell(\vartheta, x) = \log(h(x)) + \langle \vartheta, T(x) \rangle - A(\vartheta)$, so folgt (vgl. Satz 2.11) wegen der Differenzierbarkeit im Innern $\dot{\ell}(\hat{\vartheta}(x), x) = T(x) - \dot{A}(\hat{\vartheta}(x)) = 0$ und somit $\mathbb{E}_{\hat{\vartheta}(x)}[T] = T(x)$. \square

3.14 Beispiele.

- (a) Es sei $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ eine mathematische Stichprobe. Dann ist der MLE für $\vartheta = (\mu/\sigma^2, 1/(2\sigma^2))^\top$ gegeben durch $\mathbb{E}_{\hat{\vartheta}}[(\bar{X}, \overline{X^2})^\top] = (\bar{X}, \overline{X^2})^\top$, also $\hat{\mu} = \bar{X}$, $\widehat{\mu^2 + \sigma^2} = \overline{X^2}$. Durch Reparametrisierung $(\mu, \mu^2 + \sigma^2) \mapsto (\mu, \sigma^2)$ erhalten wir $\hat{\sigma}^2 = \overline{X^2} - (\bar{X})^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$. Beachte, dass der MLE $\hat{\sigma}^2$ nicht erwartungstreu ist.
- (b) Bei Beobachtung einer Markovkette (X_0, X_1, \dots, X_n) auf dem Zustandsraum $S = \{1, \dots, M\}$ mit parameterunabhängigem Anfangswert $X_0 = x_0$ und unbekanntem Übergangswahrscheinlichkeiten $\mathbb{P}(X_{k+1} = j | X_k = i) = p_{ij}$ ergibt sich die Likelihoodfunktion (bzgl. Zählmaß) durch

$$L((p_{kl}), X) = \prod_{i=1}^n p_{X_{i-1}, X_i} = \prod_{k,l=1}^M p_{kl}^{N_{kl}(X)},$$

wobei $N_{kl}(X) = |\{i = 1, \dots, n \mid X_{i-1} = k, X_i = l\}|$ die Anzahl der beobachteten Übergänge von Zustand k nach Zustand l angibt. Als MLE ergibt sich nach kurzer Rechnung die relative Häufigkeit $\hat{p}_{ij} = N_{ij} / (\sum_{m \in S} N_{im})$ der Übergänge (beliebig, falls der Nenner null ist).

- (c) Beim allgemeinen parametrischen Regressionsmodell mit Beobachtungen

$$Y_i = g_\vartheta(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

ergibt sich unter der Normalverteilungsannahme $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d. als MLE der Kleinste-Quadrate-Schätzer $\hat{\vartheta} = \text{argmin}_{\vartheta \in \Theta} \sum_{i=1}^n (Y_i - g_\vartheta(x_i))^2$.

3.15 Definition. Für zwei Wahrscheinlichkeitsmaße \mathbb{P} und \mathbb{Q} auf demselben Messraum $(\mathcal{X}, \mathcal{F})$ heißt die Funktion

$$\text{KL}(\mathbb{P} \mid \mathbb{Q}) = \begin{cases} \int_{\mathcal{X}} \log \left(\frac{d\mathbb{P}}{d\mathbb{Q}}(x) \right) \mathbb{P}(dx), & \text{falls } \mathbb{P} \ll \mathbb{Q}, \\ +\infty, & \text{sonst} \end{cases}$$

Kullback-Leibler-Divergenz (oder auch Kullback-Leibler-Abstand, relative Entropie) von \mathbb{P} bezüglich \mathbb{Q} .

3.16 Lemma. Für die Kullback-Leibler-Divergenz gilt:

(a) $\text{KL}(\mathbb{P} \mid \mathbb{Q}) \geq 0$ und $\text{KL}(\mathbb{P} \mid \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$;

(b) für Produktmaße ist KL additiv:

$$\text{KL}(\mathbb{P}_1 \otimes \mathbb{P}_2 \mid \mathbb{Q}_1 \otimes \mathbb{Q}_2) = \text{KL}(\mathbb{P}_1 \mid \mathbb{Q}_1) + \text{KL}(\mathbb{P}_2 \mid \mathbb{Q}_2);$$

(c) bildet $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ eine natürliche Exponentialfamilie und ist ϑ_0 innerer Punkt von Θ , so gilt

$$\text{KL}(\mathbb{P}_{\vartheta_0} \mid \mathbb{P}_\vartheta) = A(\vartheta) - A(\vartheta_0) + \langle \dot{A}(\vartheta_0), \vartheta_0 - \vartheta \rangle.$$

3.17 Bemerkung. Im Allgemeinen ist KL nicht symmetrisch und damit keine Metrik. Trotzdem spielt die Kullback-Leibler-Divergenz eine Hauptrolle in der Asymptotik Likelihood-basierter Verfahren.

Beweis. Für (a) können wir o.B.d.A. $\mathbb{P} \ll \mathbb{Q}$ annehmen. Dann folgt aus der strikten Konvexität von $h(x) = x \log(x)$ auf $[0, \infty)$ (mit stetiger Ergänzung $h(0) = 0$) mittels Jensen-Ungleichung

$$\text{KL}(\mathbb{P} \mid \mathbb{Q}) = \int h\left(\frac{d\mathbb{P}}{d\mathbb{Q}}(x)\right) \mathbb{Q}(dx) \geq h\left(\int \frac{d\mathbb{P}}{d\mathbb{Q}}(x) \mathbb{Q}(dx)\right) = h(1) = 0$$

mit Gleichheit genau dann, wenn $\frac{d\mathbb{P}}{d\mathbb{Q}}$ \mathbb{Q} -f.s. konstant ist. Da eine konstante Dichte zwischen Wahrscheinlichkeitsmaßen notwendigerweise gleich Eins sein muss, folgt Aussage (a) aus $\frac{d\mathbb{P}}{d\mathbb{Q}} = 1$ \mathbb{Q} -f.s. $\iff \mathbb{P} = \mathbb{Q}$.

Für (b) benutze die Produktdichte und Fubini im Fall $\text{KL}(\mathbb{P}_1 \mid \mathbb{Q}_1) < \infty$ und $\text{KL}(\mathbb{P}_2, \mathbb{Q}_2) < \infty$:

$$\begin{aligned} \text{KL}(\mathbb{P}_1 \otimes \mathbb{P}_2 \mid \mathbb{Q}_1 \otimes \mathbb{Q}_2) &= \int \int \log\left(\frac{d\mathbb{P}_1}{d\mathbb{Q}_1}(x_1) \frac{d\mathbb{P}_2}{d\mathbb{Q}_2}(x_2)\right) \mathbb{P}_1(dx_1) \mathbb{P}_2(dx_2) \\ &= \int \log\left(\frac{d\mathbb{P}_1}{d\mathbb{Q}_1}(x_1)\right) \mathbb{P}_1(dx_1) + \int \log\left(\frac{d\mathbb{P}_2}{d\mathbb{Q}_2}(x_2)\right) \mathbb{P}_2(dx_2) \end{aligned}$$

und wir erhalten $\text{KL}(\mathbb{P}_1 \mid \mathbb{Q}_1) + \text{KL}(\mathbb{P}_2 \mid \mathbb{Q}_2)$. Um die Anwendung von Fubini zu begründen, müssen wir noch

$$\int \int \left| \log\left(\frac{d\mathbb{P}_1}{d\mathbb{Q}_1}(x_1) \frac{d\mathbb{P}_2}{d\mathbb{Q}_2}(x_2)\right) \right| \mathbb{P}_1(dx_1) \mathbb{P}_2(dx_2) < \infty$$

zeigen. Das Doppelintegral kann mittels Dreiecksungleichung und Verwendung der Funktion h abgeschätzt werden durch

$$\int \left| h\left(\frac{d\mathbb{P}_1}{d\mathbb{Q}_1}(x_1)\right) \right| \mathbb{Q}_1(dx_1) + \int \left| h\left(\frac{d\mathbb{P}_2}{d\mathbb{Q}_2}(x_2)\right) \right| \mathbb{Q}_2(dx_2).$$

Da h nach unten beschränkt ist ($h(x) \geq -e^{-1}$) und $\mathbb{Q}_1, \mathbb{Q}_2$ Wahrscheinlichkeitsmaße sind, sind die Integrale endlich genau dann, wenn die Integrale über die Integranden ohne Absolutwerte endlich sind. Letzteres folgt aus $\text{KL}(\mathbb{P}_1 \mid \mathbb{Q}_1) < \infty$ und $\text{KL}(\mathbb{P}_2, \mathbb{Q}_2) < \infty$. Im Fall $\text{KL}(\mathbb{P}_1 \mid \mathbb{Q}_1) = \infty$ oder $\text{KL}(\mathbb{P}_2, \mathbb{Q}_2) = \infty$ folgt $\text{KL}(\mathbb{P}_1 \otimes \mathbb{P}_2 \mid \mathbb{Q}_1 \otimes \mathbb{Q}_2) = \infty$ auf ähnliche Weise.

Behauptung (c) folgt durch Einsetzen von $\log\left(\frac{d\mathbb{P}_{\vartheta_0}}{d\mathbb{P}_\vartheta}(x)\right) = \langle T(x), \vartheta_0 - \vartheta \rangle + A(\vartheta) - A(\vartheta_0)$ sowie $\mathbb{E}_{\vartheta_0}[T] = \dot{A}(\vartheta_0)$, vergleiche Satz 2.11. \square

3.18 Bemerkung. Wegen $\ddot{A}(\vartheta_0) = \text{Var}_{\vartheta_0}(T)$ in (c) erhalten wir für $\vartheta, \vartheta_0 \in \text{int}(\Theta)$ mit einer Taylorentwicklung $\text{KL}(\mathbb{P}_{\vartheta_0} | \mathbb{P}_{\vartheta}) = \frac{1}{2} \langle \text{Var}_{\bar{\vartheta}}(T)(\vartheta - \vartheta_0), \vartheta - \vartheta_0 \rangle$ mit einer Zwischenstelle $\bar{\vartheta}$ zwischen ϑ und ϑ_0 . Beachte, dass $\text{Var}_{\bar{\vartheta}}(T)$ gerade die Fisher-Information bei $\bar{\vartheta}$ angibt. Im Fall der mehrdimensionalen Normalverteilung $N(\mu, \Sigma)$ mit strikt positiv-definiter Kovarianzmatrix folgt aus $A(\mu) = \langle \Sigma^{-1}\mu, \mu \rangle / 2$, dass $\ddot{A}(\mu) = \Sigma^{-1}$ unabhängig von μ ist und somit $\text{KL}(N(\vartheta_0, \Sigma) | N(\vartheta, \Sigma)) = \frac{1}{2} \langle \Sigma^{-1}(\vartheta - \vartheta_0), \vartheta - \vartheta_0 \rangle$ gilt.

3.19 Definition. Es sei $(\mathcal{X}_n, \mathcal{F}_n, (\mathbb{P}_{\vartheta}^n)_{\vartheta \in \Theta})_{n \geq 1}$ eine Folge statistischer Modelle sowie $g(\vartheta)$ mit $g : \Theta \rightarrow \Gamma$ der interessierende Parameter. Eine Funktion $K : \Theta \times \Gamma \rightarrow \mathbb{R} \cup \{+\infty\}$ heißt Kontrastfunktion, falls $\gamma \mapsto K(\vartheta_0, \gamma)$ ein eindeutiges Minimum bei $g(\vartheta_0)$ besitzt für alle $\vartheta_0 \in \Theta$. Eine Folge $K_n : \Gamma \times \mathcal{X}_n \rightarrow \mathbb{R} \cup \{+\infty\}$ heißt zugehöriger Kontrastprozess (oder bloß Kontrast), falls folgende Bedingungen gelten:

- (a) $K_n(\gamma, \bullet)$ ist \mathcal{F}_n -messbar für alle $\gamma \in \Gamma$;
- (b) $\forall \gamma \in \Gamma, \vartheta_0 \in \Theta : K_n(\gamma) \rightarrow K(\vartheta_0, \gamma)$ $\mathbb{P}_{\vartheta_0}^n$ -stochastisch für $n \rightarrow \infty$.

Ein zugehöriger Minimum-Kontrast-Schätzer oder M-Schätzer von $g(\vartheta)$ ist gegeben durch $\hat{\gamma}_n(x_n) := \text{argmin}_{\gamma \in \Gamma} K_n(\gamma, x_n)$ (sofern existent; nicht notwendigerweise eindeutig).

3.20 Beispiele.

- (a) Es sei $g(\vartheta) = \vartheta$, $\Gamma = \Theta$. Beim Produktexperiment $(\mathcal{X}^n, \mathcal{F}^{\otimes n}, (\mathbb{P}_{\vartheta}^{\otimes n})_{\vartheta \in \Theta})$ mit $\mathbb{P}_{\vartheta} \sim \mathbb{P}_{\vartheta'}$ für alle $\vartheta, \vartheta' \in \Theta$ ist

$$K_n(\vartheta, x) = -\frac{1}{n} \sum_{i=1}^n \ell(\vartheta, x_i)$$

mit der Loglikelihood-Funktion $\ell(\vartheta) = \log\left(\frac{d\mathbb{P}_{\vartheta}}{d\mu}\right)$ bezüglich einem dominierenden Maß μ ein Kontrastprozess zur Kontrastfunktion

$$\begin{aligned} K(\vartheta_0, \vartheta) &= \mathbb{E}_{\vartheta_0}[-\ell(\vartheta)] = \mathbb{E}_{\vartheta_0} \left[\log \left(\frac{L(\vartheta_0)}{L(\vartheta)} \right) \right] - \mathbb{E}_{\vartheta_0}[\log(L(\vartheta_0))] \\ &= \text{KL}(\mathbb{P}_{\vartheta_0} | \mathbb{P}_{\vartheta}) - \mathbb{E}_{\vartheta_0}[\ell(\vartheta_0)], \end{aligned}$$

sofern $\ell(\vartheta_0) \in L^1(\mathbb{P}_{\vartheta_0})$ gilt. Der zugehörige M-Schätzer ist der MLE.

- (b) Es sei $g(\vartheta) = \vartheta$, $\Gamma = \Theta$. Betrachte das Regressionsmodell $Y_i = f_{\vartheta}(i/n) + \varepsilon_i$, $i = 1, \dots, n$, mit $f_{\vartheta} : [0, 1] \rightarrow \mathbb{R}$ stetig, (ε_i) i.i.d. mit $\mathbb{E}[\varepsilon_i] = 0$ und $\mathbb{E}[\varepsilon_i^2] < \infty$. Dann folgt leicht mit Riemannscher Summen-Approximation, dass $K_n(\vartheta) = \frac{1}{n} \sum_{i=1}^n (Y_i - f_{\vartheta}(x_i))^2$ einen Kontrastprozess zur Kontrastfunktion $K(\vartheta_0, \vartheta) = \int_0^1 (f_{\vartheta_0}(x) - f_{\vartheta}(x))^2 dx + \mathbb{E}[\varepsilon_i^2]$ bildet. Dabei muss natürlich die Identifizierbarkeitsbedingung $f_{\vartheta} \neq f_{\vartheta'}$ für alle $\vartheta \neq \vartheta'$ gelten. Also ist der Kleinste-Quadrate-Schätzer hier ebenfalls M-Schätzer.
- (c) Im Regressionsmodell aus (b) liege nun eine Modellmisspezifikation vor in dem Sinne, dass die Beobachtungen gemäß $Y_i = f^0(i/n) + \varepsilon_i$ generiert werden, wobei $f^0 : [0, 1] \rightarrow \mathbb{R}$ nicht notwendigerweise gleich einem

f_{ϑ} ist. Nimmt man an, dass die Funktion selbst der Parameter ϑ im Kleinste-Quadrate-Ansatz ist, d.h. $\hat{\vartheta}_n = \operatorname{argmin}_{\vartheta \in \Theta} \frac{1}{n} \sum_{i=1}^n (Y_i - \vartheta(i/n))^2$ mit $\Theta \subseteq L^2([0, 1])$, so erhalten wir nach obiger Herleitung im Grenzwert die 'Kontrast-Typ-Funktion' $K(f^0, \vartheta) = \int_0^1 (f^0(x) - \vartheta(x))^2 dx + \mathbb{E}[\varepsilon_i^2]$. Für $f^0 \notin \Theta$ wird das Minimum nun natürlich nicht in f^0 angenommen, so dass in der Kontrasttheorie die Funktion g wesentlich wird.

Dazu nehmen wir an, dass die parametrische Funktionenmenge Γ (vormals Θ) Riemann-integrierbare Funktionen enthält sowie abgeschlossen in $L^2([0, 1])$ und konvex ist, so dass für jede Funktion $\vartheta \in L^2([0, 1])$ eine eindeutige L^2 -Orthogonalprojektion $g(\vartheta)$ auf Γ existiert. Beispielsweise kann Γ die Menge aller Polynome vom Grad $\leq d$ sein. Bezeichnet Θ die Menge der quadratisch Riemann-integrierbaren Funktionen in $L^2([0, 1])$, so ist $K_n(\gamma) = \frac{1}{n} \sum_{i=1}^n (Y_i - \gamma(i/n))^2$, $\gamma \in \Gamma$, Kontrastprozess zur Kontrastfunktion $K(\vartheta_0, \gamma) = \|\vartheta_0 - \gamma\|_{L^2}^2 + \mathbb{E}[\varepsilon_i^2]$, welche genau bei $\gamma = g(\vartheta_0)$ ihr Minimum in Γ annimmt. Es ist zu erwarten (vgl. Übungen), dass unter geeigneten Bedingungen der Kleinste-Quadrate-Schätzer $\hat{\gamma}_n = \operatorname{argmin}_{\gamma \in \Gamma} \frac{1}{n} \sum_{i=1}^n (Y_i - \gamma(i/n))^2$ unter $\mathbb{P}_{\vartheta_0}^n$ gegen $g(\vartheta_0)$ konvergiert. Im derart misspezifizierten Modell wird also die beste L^2 -Approximation an die wahre Funktion ϑ_0 geschätzt, z.B. das best approximierende Polynom vom Grad $\leq d$.

3.3 Asymptotik

3.21 Satz. *Es sei $(K_n)_{n \geq 1}$ ein Kontrastprozess zur Kontrastfunktion K . Dann ist der zugehörige M -Schätzer $\hat{\gamma}_n$ konsistent für $g(\vartheta_0)$, $\vartheta_0 \in \Theta$, unter folgenden Bedingungen:*

- (A1) Γ ist ein kompakter Raum;
- (A2) $\gamma \mapsto K(\vartheta_0, \gamma)$ ist stetig und $\gamma \mapsto K_n(\gamma)$ ist $\mathbb{P}_{\vartheta_0}^n$ -f.s. stetig für alle $n \geq 1$;
- (A3) $\sup_{\gamma \in \Gamma} |K_n(\gamma) - K(\vartheta_0, \gamma)| \rightarrow 0$ $\mathbb{P}_{\vartheta_0}^n$ -stochastisch.

3.22 Bemerkung. Beachte, dass $\hat{\gamma}_n$ als Minimum einer fast sicher stetigen Funktion auf einem Kompaktum stets fast sicher existiert. Es kann außerdem messbar gewählt werden (vgl. Witting, 2. Band, Satz 6.7).

Bedingungen (A1) und (A2) können ersetzt werden durch die schwächere (wieso?) Bedingung

$$(A1') : \forall \varepsilon > 0 \quad \inf_{d(\gamma, g(\vartheta_0)) \geq \varepsilon} K(\vartheta_0, \gamma) > K(\vartheta_0, g(\vartheta_0))$$

mit der Metrik d von Γ , vergleiche Übungen.

Beweis. Zeige, dass die Funktion $\operatorname{argmin} : C(\Gamma) \rightarrow \Gamma$, wobei eine Minimalstelle ausgewählt sei bei Nichteindeutigkeit, stetig bezüglich Maximumsnorm auf $C(\Gamma)$ ist an den Stellen f , wo $m_f := \operatorname{argmin}_{\gamma} f(\gamma)$ eindeutig ist. Betrachte $f_n \in C(\Gamma)$ mit $\|f_n - f\|_{\infty} \rightarrow 0$. Dann konvergieren auch die Minima

$f_n(m_{f_n}) \rightarrow f(m_f)$ wegen

$$\begin{aligned} f_n(m_{f_n}) - f(m_f) &\geq f(m_{f_n}) - f(m_f) - \|f_n - f\|_\infty \geq -\|f_n - f\|_\infty \rightarrow 0, \\ f_n(m_{f_n}) - f(m_f) &\leq f_n(m_{f_n}) - f_n(m_f) + \|f_n - f\|_\infty \leq \|f_n - f\|_\infty \rightarrow 0. \end{aligned}$$

Ist nun $m \in \Gamma$ (Γ kompakt) ein Häufungspunkt von (m_{f_n}) , so folgt mit gleichmäßiger Konvergenz $f(m) = \lim_{n \rightarrow \infty} f_n(m_{f_n}) = f(m_f)$. Eindeutigkeit des Minimums liefert $m = m_f$, und daher besitzt (m_{f_n}) als einzigen Häufungspunkt notwendigerweise den Grenzwert m_f .

Das *Continuous-Mapping-Theorem* für stochastische Konvergenz liefert mit (A3) die Behauptung, weil argmin stetig ist auf dem deterministischen Grenzwert $K(\vartheta_0, \bullet)$. \square

3.23 Satz. Ist $\Gamma \subseteq \mathbb{R}^k$ kompakt, $(X_n(\gamma), \gamma \in \Gamma)_{n \geq 1}$ eine Folge stetiger Prozesse mit $X_n(\gamma) \xrightarrow{\mathbb{P}} X(\gamma)$ für alle $\gamma \in \Gamma$ und stetigem Grenzprozess $(X(\gamma), \gamma \in \Gamma)$, so gilt $\max_{\gamma \in \Gamma} |X_n(\gamma) - X(\gamma)| \xrightarrow{\mathbb{P}} 0$ genau dann, wenn

$$\forall \varepsilon > 0 : \lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P} \left(\sup_{|\gamma_1 - \gamma_2| < \delta} |X_n(\gamma_1) - X_n(\gamma_2)| \geq \varepsilon \right) = 0.$$

Die Bedingung in der vorigen Zeile (Straffheit) folgt aus

$$\exists \alpha, \beta > 0 \ K > 0 \ \forall n \geq 1, \gamma_1, \gamma_2 \in \Gamma : \mathbb{E}[|X_n(\gamma_1) - X_n(\gamma_2)|^\alpha] \leq K |\gamma_1 - \gamma_2|^{k+\beta}.$$

Beweis. Siehe Stochastik II bzw. Übung. \square