Kapitel I

Score und Information

Reinhard Höpfner

Vorlesung Mathematische Statistik

Wintersemester 2004/2005 und Wintersemester 2007/2008

Institut für Mathematik, Johannes Gutenberg Universität Mainz

08.11.04 und 26.05.08

Übersicht zu Kapitel I :

A. Score, Information, zwei Informationsschranken

In diesem Kapitel behandeln wir den Begriff der 'Information' in einem statistischen Modell in seiner klassischen Fassung. 'Information' erlaubt, Schranken für die Güte von Schätzern zu formulieren. Insbesondere – dies wird in diesem Kapitel aber nur heuristisch andiskutiert – besteht ein enger Zusammenhang zwischen Limesvarianzen für Maximum-Likelihood Schätzer und der 'Information' im statistischen Modell.

1.1 Definition: a) Ein *statistisches Modell* (oder *Experiment*) ist ein Tripel

$$(\Omega, \mathcal{A}, \mathcal{P})$$

wobei $\mathcal{P}$ eine festgelegte Familie von Wahrscheinlichkeitsmaßen auf einem meßbaren Raum $(\Omega, \mathcal{A})$ bezeichnet. Ein statistisches Modell heißt *parametrisch*, falls die Familie $\mathcal{P}$ bijektiv durch einen endlichdimensionalen Parameter beschrieben werden kann:

$$\mathcal{P} \;=\; \{P_\vartheta : \vartheta \in \Theta\}\;, \quad \Theta \subset I\!R^d\;,$$

und *dominiert*, wenn es ein σ-endliches Maß μ auf $(\Omega, \mathcal{A})$ gibt so daß

$$P \ll \mu \quad \text{für jedes Wahrscheinlichkeitsmaß } P \in \mathcal{P}.$$

b) Eine meßbare Abbildung von $(\Omega, \mathcal{A})$ in einen meßbaren Raum $(G, \mathcal{G})$ nennt man eine *Statistik auf* $(\Omega, \mathcal{A}, \mathcal{P})$. Eine Statistik T mit Werten in $(I\!R^k, \mathcal{B}(I\!R^k))$ heißt *q-fach integrierbar* (für $q \geq 1$ fest) falls

$$T \in L^q(P) \quad \text{für jedes } P \in \mathcal{P}.$$

c) Sei $(\Omega, \mathcal{A}, \{P_\vartheta : \vartheta \in \Theta\})$ ein parametrisches Experiment, $\Theta \subset I\!R^d$. Ein *Schätzer für den unbekannten Parameter* ist eine Statistik T mit Werten in $(I\!R^d, \mathcal{B}(I\!R^d))$. T heißt *erwartungstreu* falls

$$T \in L^1(P) \quad \text{und} \quad E_\vartheta(T) \;=\; \vartheta \quad \text{für jedes } \vartheta \in \Theta.$$

d) Sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein beliebiges Experiment, sei $\gamma : \mathcal{P} \to I\!R^k$ eine Abbildung. Ein *Schätzer für γ* ist eine Statistik T mit Werten in $(I\!R^k, \mathcal{B}(I\!R^k))$; T heißt *erwartungtreu für γ* falls

$$T \in L^1(P) \quad \text{und} \quad E_P(T) \;=\; \gamma(P) \quad \text{für jedes } P \in \mathcal{P}\;.$$

1.2 Definition: (Score und Information, klassische Definition) Betrachte ein Experiment

$$\mathcal{E} \;:=\; (\Omega, \mathcal{A}, \{P_\vartheta : \vartheta \in \Theta\}) \;, \quad \Theta \subset I\!R^d \text{ offen}$$

mit dominierendem Maß μ und mit Dichten

$$\frac{d\,P_\vartheta}{d\,\mu}\,(\omega) \;:=\; f(\vartheta,\omega) \;=\; f_\vartheta(\omega)\;, \quad \vartheta \in \Theta\;,\; \omega \in \Omega\;.$$

Für jedes feste $\omega \in \Omega$ sei $f(\cdot,\omega)$ stetig und partiell differenzierbar auf Θ. Die partiellen Ableitungen, punktweiser Limes meßbarer Funktionen $\frac{f(\vartheta+he_i,\cdot)-f(\vartheta,\cdot)}{h}$ für $h \to 0$, sind dann meßbar.

a) Schreibe ∇ für den Vektor der partiellen Ableitungen nach ϑ und definiere

$$M_\vartheta \;:=\; (\nabla \log f)(\vartheta,\omega) \;:=\; 1_{\{f_\vartheta>0\}}(\omega) \begin{pmatrix} \frac{\partial}{\partial\vartheta_1}\log f \\ \cdots \\ \frac{\partial}{\partial\vartheta_d}\log f \end{pmatrix}(\vartheta,\omega)\;, \quad \vartheta \in \Theta\;,\; \omega \in \Omega\;.$$

Dies ist eine wohldefinierte Zufallsvariable auf $(\Omega,\mathcal{A})$ mit Werten in $(I\!R^d, \mathcal{B}(I\!R^d))$.
Setze weiter voraus

$(*)$ für jedes $\vartheta \in \Theta$ gilt $M_\vartheta \in L^2(P_\vartheta)$ und $E_\vartheta(M_\vartheta) = 0$.

Sind alle diese Bedingungen erfüllt, so heißt M_ϑ *Score in* ϑ; die Kovarianzmatrix unter P_ϑ

$$I_\vartheta \;:=\; Cov_\vartheta(M_\vartheta) \;=\; E_\vartheta\left(M_\vartheta M_\vartheta^\top\right)$$

heißt *Fisher-Information in* ϑ; $\mathcal{E}$ nennt man ein *Experiment mit Score und Fisher-Information*.

b) Allgemeiner erlaubt man Abänderungen der in a) definierten M_ϑ auf P_ϑ-Nullmengen, und nennt jede Familie meßbarer Abbildungen $\{\widetilde{M}_\vartheta : \vartheta \in \Theta\}$ von $(\Omega,\mathcal{A})$ nach $(I\!R^d, \mathcal{B}(I\!R^d))$ mit der Eigenschaft

$$\text{für jedes } \vartheta \text{ in } \Theta \text{ gilt} \quad \widetilde{M}_\vartheta = M_\vartheta \quad P_\vartheta\text{–fast sicher}$$

eine *Festlegung des Score* im Modell $\{P_\vartheta : \vartheta \in \Theta\}$.

1.2' Bemerkungen: a) Score und Information hängen wesentlich ab von der für ein statistisches Modell $\mathcal{P}$ gewählten Parametrisierung $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$.
b) Score und Information sind unabhängig von der Wahl des dominierenden Maßes:

sind μ_1 und μ_2 verschiedene Wahlen eines die Familie $\{P_\vartheta : \vartheta \in \Theta\}$ dominierenden Maßes, so dominiert auch $\mu_1 + \mu_2$ diese Familie, und

$$\frac{d\,P_\vartheta}{d\,\mu_1} \cdot \frac{d\,\mu_1}{d(\mu_1+\mu_2)} \;=\; \frac{d\,P_\vartheta}{d(\mu_1+\mu_2)} \;=\; \frac{d\,P_\vartheta}{d\,\mu_2} \cdot \frac{d\,\mu_2}{d(\mu_1+\mu_2)}\;.$$

Dabei hängt der zweite Faktor auf der rechten und auf der linken Seite nicht von ϑ ab, tritt also im Score $(\nabla \log f)(\vartheta,\cdot)$ nicht mehr in Erscheinung.

1.3 Beispiel: ($d = 1$) Fixiere ein beliebiges Wahrscheinlichkeitsmaß F auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, wähle eine Funktion $h : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit den Eigenschaften

$$(\diamond) \qquad \int h dF \;=\; 0\,, \quad 0 < \int h^2 dF < \infty\,.$$

a) Sei zuerst h beschränkt und $\sup_{x\in\mathbb{R}} |h(x)| \le M < \infty$. Dann erfüllt die durch

$$(*) \qquad \big(\,\mathbb{R},\, \mathcal{B}(\mathbb{R}),\, \{F_\vartheta : |\vartheta| < M^{-1}\}\,\big) \quad \text{mit} \quad F_\vartheta(d\omega) \;:=\; (1+\vartheta h(\omega))\,F(d\omega)$$

definierte einparametrige Familie von Wahrscheinlichkeitsmaßen alle in 1.2 gemachten Voraussetzungen. Die Familie ist dominiert durch $\mu := F$, mit strikt positiven Dichten

$$f(\vartheta,\omega) \;=\; \frac{d\,F_\vartheta}{d\,F}(\omega) \;=\; 1+\vartheta h(\omega)\,, \quad \vartheta \in \Theta\,,\ \omega \in \mathbb{R}$$

(insbesondere sind alle Wahrscheinlichkeitsmaße F_ϑ, $|\vartheta| < M^{-1}$, äquivalent). An jeder Stelle $\vartheta \in \Theta$ sind Score M_ϑ und Information I_ϑ gegeben durch

$$M_\vartheta(\omega) \;=\; \left(\frac{h}{1+\vartheta h}\right)(\omega)\,, \quad I_\vartheta \;=\; \int \left(\frac{h}{1+\vartheta h}\right)^2 dF_\vartheta \;=\; \int \frac{h^2}{1+\vartheta h}\,dF\,.$$

Man nennt die Familie $(*)$ einen *einparametrigen Pfad durch den Punkt P in Richtung h*, aufgefaßt als Submodell im nichtparametrischen Modell $\mathcal{F}$ aller auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ existierenden Wahrscheinlichkeitsmaße. Insbesondere gilt an der Stelle $\vartheta = 0$

$$M_0 \;=\; h\,, \quad I_0 \;=\; \int h^2\,dF\,.$$

b) Allgemein kann man zu *jeder* Richtung $h : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit den Eigenschaften $(\diamond)$ einen einparametrigen Pfad durch den Punkt P in Richtung h festlegen:
Ist h unbeschränkt, so verallgemeinert man die Konstruktion aus a) folgendermaßen. Man wählt eine Trunkationsfunktion $\psi \in \mathcal{C}_0^1(\mathbb{R})$ (Klasse der auf $\mathbb{R}$ stetig differenzierbaren Funktionen mit kompaktem Träger) mit den Eigenschaften

$$\psi(x) = x \quad \text{falls } |x| < \frac{1}{3}\,, \quad \max|\psi| < \frac{1}{2}$$

und definiert eine Familie von Wahrscheinlichkeitsmaßen durch

$$(**) \quad (\,\mathbb{R},\, \mathcal{B}(\mathbb{R}),\, \{F_\vartheta : |\vartheta| < 1\}\,) \quad \text{mit} \quad F_\vartheta(d\omega) := \left(1 + [\psi(\vartheta h(\omega)) - \int \psi(\vartheta h) dF]\right) F(d\omega)\,.$$

(Falls insbesondere wie in a) h eine beschränkte Funktion ist, stimmen auf einer hinreichend kleinen Umgebung von $\vartheta = 0$ die Pfade $(**)$ und $(*)$ überein.) Wir zeigen, daß das Modell $(**)$ alle Voraussetzungen aus 1.2 erfüllt. Wegen $\max|\psi'| < \infty$ zeigt dominierte Konvergenz

$$\frac{d}{d\vartheta} \int \psi(\vartheta h)\, dF \;=\; \int h\, \psi'(\vartheta h)\, dF\,,$$

folglich sind mit strikt positiven Dichten $f(\vartheta, \omega) = 1 + [\psi(\vartheta h(\omega)) - \int \psi(\vartheta h) dF]$ die Scores

$$M_\vartheta(\omega) \;=\; \left(\frac{d}{d\vartheta} \log f\right)(\vartheta, \omega) \;=\; \frac{h(\omega)\psi'(\vartheta h(\omega)) - \int h\, \psi'(\vartheta h)\, dF}{1 + [\psi(\vartheta h(\omega)) - \int \psi(\vartheta h) dF]}$$

Zufallsvariable in $L^2(F_\vartheta)$ mit $E(M_\vartheta) = 0$ für alle $|\vartheta| < 1$, wegen

$$\int \frac{h(\omega)\psi'(\vartheta h(\omega)) - \int h\, \psi'(\vartheta h)\, dF}{1 + [\psi(\vartheta h(\omega)) - \int \psi(\vartheta h) dF]}\, F_\vartheta(d\omega) = \int \left[h(\omega)\psi'(\vartheta h(\omega)) - \int h\, \psi'(\vartheta h)\, dF\right] F(d\omega) = 0\,.$$

Dabei erhält man an der Stelle $\vartheta = 0$ wieder wie im Modell aus a)

$$M_0 \;=\; h\,, \quad I_0 \;=\; \int h^2 dF\,,$$

also ist auch $(**)$ als Teilfamilie von $\mathcal{F}$ ein einparametriger Pfad durch P in Richtung h. □

1.4 Beispiel: (Lokationsmodell) Sei F ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mit Dichte f bezüglich des Lebesguemaßes $\lambda\!\!\lambda$. Es gelte

$$f \text{ ist differenzierbar auf } \mathbb{R} \text{ mit Ableitung } f'$$

$$f \text{ ist strikt positiv auf } (a, b), \text{ und} \equiv 0 \text{ sonst}$$

für ein offenes Intervall (a, b) in $\mathbb{R}$ (mit $-\infty \le a,\, b \le +\infty$). Weiter sei

$$\int \left(\frac{f'}{f}\right)^2 dF \;<\; \infty\,.$$

Für das von F erzeugte Lokationsmodell

$$(\,\mathbb{R},\, \mathcal{B}(\mathbb{R}),\, \{F_\vartheta : \vartheta \in \mathbb{R}\}\,) \quad \text{mit} \quad dF_\vartheta \;:=\; f(\cdot - \vartheta) d\lambda\!\!\lambda$$

gelten dann alle in 1.2 gemachten Voraussetzungen. Das Modell ist dominiert durch $\lambda\!\!\lambda$. Für $\vartheta \in \Theta$ ist die Verteilung F_ϑ konzentriert auf das Intervall $(a + \vartheta, b + \vartheta)$, und Score und Information sind gegeben durch

$$M_\vartheta(\omega) \;=\; 1_{(a+\vartheta, b+\vartheta)}(\omega) \left(-\frac{f'}{f}\right)(\omega - \vartheta)\,, \quad \vartheta \in \Theta\,,\, \omega \in \mathbb{R}$$

sowie

$$I_\vartheta \;=\; E_\vartheta\left(M_\vartheta^2\right) \;=\; \int_a^b \left(\frac{f'}{f}\right)^2 dF\;, \quad \vartheta\in\Theta\;:$$

insbesondere hängt im Lokationsmodell die Information nicht vom Parameter ab. □

1.5 Hilfssatz: (Produktmodelle) Betrachte ein Experiment

$$\mathcal{E} \;:=\; (\,\Omega,\,\mathcal{A},\,\{P_\vartheta:\vartheta\in\Theta\}\,)\;, \quad \Theta\subset\mathbb{R}^d \text{ offen}$$

mit Score $\{M_\vartheta:\vartheta\in\Theta\}$ und Fisher-Information $\{I_\vartheta:\vartheta\in\Theta\}$ wie in 1.2. Dann gilt:

a) Alle Produktmodelle

$$\mathcal{E}_n \;:=\; \left(\bigtimes_{i=1}^n \Omega,\;\bigotimes_{i=1}^n \mathcal{A},\;\{P_{n,\vartheta}:=\bigotimes_{i=1}^n P_\vartheta:\vartheta\in\Theta\}\right)\;, \quad n\geq 1$$

erfüllen ebenfalls die in 1.2 gemachten Voraussetzungen:

in $\mathcal{E}_n$ sind Score $\{M_{n,\vartheta}:\vartheta\in\Theta\}$ und Information $\{I_{n,\vartheta}:\vartheta\in\Theta\}$ gegeben durch

$$M_{n,\vartheta}(\omega_1,\ldots,\omega_n) \;=\; \sum_{i=1}^n M_\vartheta(\omega_i) \quad \text{für } P_{n,\vartheta}\text{-fast alle} \quad (\omega_1,\ldots,\omega_n)\in\bigtimes_{i=1}^n\Omega\;, \quad \vartheta\in\Theta$$

$$I_{n,\vartheta} \;=\; n\cdot I_\vartheta\;, \quad \vartheta\in\Theta\;.$$

b) Man kann auch ausgehen vom unendlichen Produktexperiment

$$\mathcal{E}_\infty \;:=\; \left(\bigtimes_{i=1}^\infty \Omega,\;\bigotimes_{i=1}^\infty \mathcal{A},\;\{Q_\vartheta:=\bigotimes_{i=1}^\infty P_\vartheta:\vartheta\in\Theta\}\right)$$

mit Koordinatenprojektionen $X_i:(\omega_1,\omega_2,\ldots)\to\omega_i\,,\ i\in\mathbb{N}$; n-fache Versuchswiederholung ist dann das Experiment

$$\widetilde{\mathcal{E}}_n \;:=\; \left(\bigtimes_{i=1}^\infty \Omega,\;\mathcal{F}_n:=\sigma(X_1,\ldots,X_n),\;\{P_{n,\vartheta}:=Q_\vartheta|\mathcal{F}_n:\vartheta\in\Theta\}\right)$$

wobei $Q_\vartheta|\mathcal{F}_n$ die Restriktion von Q_ϑ auf die Sub-σ-Algebra $\mathcal{F}_n=\sigma(X_1,\ldots,X_n)$ von $\bigotimes_{i=1}^\infty\mathcal{A}$ bezeichnet. Für jedes feste $n<\infty$ schreiben sich Score und Information auf $\widetilde{\mathcal{E}}_n$ als

$$M_{n,\vartheta}(\omega_1,\omega_2,\ldots) \;=\; \sum_{i=1}^n M_\vartheta(\omega_i) \quad \text{für } P_{n,\vartheta}\text{-fast alle} \quad (\omega_1,\omega_2,\ldots)\in\bigtimes_{i=1}^\infty\Omega\;, \quad \vartheta\in\Theta$$

$$I_{n,\vartheta} \;=\; n\cdot I_\vartheta\;, \quad \vartheta\in\Theta\;.$$

Beweis: 1) Sei μ ein dominierendes Maß für das Experiment $\mathcal{E}$, seien $f_\vartheta=\frac{dP_\vartheta}{d\mu}$ Festlegungen der Dichten, die den Voraussetzungen aus 1.2 genügen. Das Produktexperiment $\mathcal{E}_n$ wird dann

dominiert durch $\mu_n := \bigotimes_{i=1}^{n} \mu$, mit Dichten

$$f_{n,\vartheta}(\omega_1,\ldots,\omega_n) \;=\; \frac{dP_{n,\vartheta}}{d\mu_n}(\omega_1,\ldots,\omega_n) \;=\; \prod_{i=1}^{n} \frac{dP_\vartheta}{d\mu}(\omega_i) \;=\; \prod_{i=1}^{n} f_\vartheta(\omega_i)\;.$$

Betrachte das Rechteck

$$A_{n,\vartheta} \;:=\; \{(\omega_1,\ldots,\omega_n) : f_{n,\vartheta}(\omega_1,\ldots,\omega_n) > 0\} \;=\; \bigtimes_{i=1}^{n} \{\omega_i : f_\vartheta(\omega_i) > 0\}$$

in $\bigotimes_{i=1}^{n} \mathcal{A}$. Mit den Konventionen aus 1.2 a) erhält man auf dem Produktraum $(\bigtimes_{i=1}^{n} \Omega, \bigotimes_{i=1}^{n} \mathcal{A})$

$$M_{n,\vartheta}((\omega_1,\ldots,\omega_n)) \;=\; 1_{A_{n,\vartheta}}((\omega_1,\ldots,\omega_n)) \sum_{i=1}^{n} M_\vartheta(\omega_i)\;.$$

Wegen $P_{n,\vartheta}\,(A_{n,\vartheta}) = 1$ stimmen auf $(\bigtimes_{i=1}^{n} \Omega, \bigotimes_{i=1}^{n} \mathcal{A})$ die meßbaren Abbildungen

$$(\omega_1,\ldots,\omega_n) \;\to\; M_{n,\vartheta}((\omega_1,\ldots,\omega_n))\;, \quad (\omega_1,\ldots,\omega_n) \;\to\; \sum_{i=1}^{n} M_\vartheta(\omega_i)$$

$P_{n,\vartheta}$-fast sicher überein, und werden wie in 1.2 b) unter $P_{n,\vartheta}$ identifiziert.

2) Unter $P_{n,\vartheta}$ sind die Ergebnisse $\omega_1,\ldots,\omega_n$ der Einzelversuche unabhängig, daher sind

$$(\omega_1,\ldots,\omega_n) \;\to\; M_\vartheta(\omega_l)\;, \quad 1 \le l \le d$$

i.i.d. auf $(\bigtimes_{i=1}^{n} \Omega, \bigotimes_{i=1}^{n} \mathcal{A})$. Also gilt für die Komponenten $M_{n,\vartheta,i} = 1_{\{f_{n,\vartheta}>0\}} \left(\frac{\partial}{\partial\vartheta_i} \log f_{n,\vartheta}\right)$ von $M_{n,\vartheta}$ und für die entsprechend bezeichneten Komponenten von M_ϑ

$$\begin{aligned} E_{P_{n,\vartheta}}\,(M_{n,\vartheta,i}\, M_{n,\vartheta,j}) \;&=\; \int \left(\sum_{l=1}^{n} M_{\vartheta,i}(\omega_l)\right)\left(\sum_{k=1}^{n} M_{\vartheta,j}(\omega_k)\right) P_{n,\vartheta}(d\omega_1,\ldots,d\omega_n) \\ &=\; n \cdot \int M_{\vartheta,i}(\omega_1)\, M_{\vartheta,j}(\omega_1)\, P_{n,\vartheta}(d\omega_1,\ldots,d\omega_n) \quad = \quad n \cdot (I_\vartheta)_{i,j}\;, \quad i,j = 1,\ldots,d \end{aligned}$$

wegen Unabhängigkeit der Einzelbeobachtungen, und wegen $(*)$ in 1.2. Damit ist a) bewiesen.

3) Aussage b) erhält man analog, da auf dem unendlichen Produktraum $\left(\bigtimes_{i=1}^{\infty} \Omega, \bigotimes_{i=1}^{\infty} \mathcal{A}\right)$ die Dichte von Q_ϑ bezüglich des unendlichen Produktmaßes $\bigotimes_{i=1}^{\infty} \mu$ in Einschränkung auf eine Sub-σ-Algebra $\mathcal{F}_n$ *für endliches* n durch $(\omega_1,\omega_2,\ldots) \longrightarrow \prod_{i=1}^{n} f_\vartheta(\omega_i)$ gegeben ist. □

Die Fisher-Information spielt eine wichtige Rolle in verschiedenen Schranken für die Güte von Schätzern. Die bekannteste Schranke ist die folgende.

1.6 Satz: (Cramér-Rao-Schranke, um 1940) Sei $\mathcal{E} = (\Omega, \mathcal{A}, \{P_\vartheta : \vartheta \in \Theta\})$, $\Theta \in \mathbb{R}^d$ offen, ein Experiment mit Score $\{M_\vartheta : \vartheta \in \Theta\}$ und Fisher-Information $\{I_\vartheta : \vartheta \in \Theta\}$ wie in 1.2. Die zu schätzende Kenngröße $\gamma : \Theta \to \mathbb{R}^k$ sei partiell differenzierbar, und

$$Y : (\Omega, \mathcal{A}) \to (\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$$

sei ein quadratintegrierbarer erwartungstreuer Schätzer für γ.

a) An jeder Stelle $\vartheta \in \Theta$, an der die Bedingungen

(+) die Informationsmatrix I_ϑ ist invertierbar

(++)
$$\begin{pmatrix} \frac{\partial}{\partial \vartheta_1}\gamma_1 & \cdots & \frac{\partial}{\partial \vartheta_d}\gamma_1 \\ \cdots & \cdots & \cdots \\ \frac{\partial}{\partial \vartheta_1}\gamma_k & \cdots & \frac{\partial}{\partial \vartheta_d}\gamma_k \end{pmatrix}(\vartheta) \;=\; E_\vartheta\left(Y\, M_\vartheta^\top\right)$$

erfüllt sind, gilt für den Schätzfehler von Y die Schranke

$$Cov_\vartheta(Y) \;\geq\; V_\vartheta\, I_\vartheta^{-1}\, V_\vartheta^\top$$

wobei V_ϑ die in der Vertauschungsbedingung (++) betrachtete Jacobi-Matrix von γ bezeichnet.

b) Schöpft ein Schätzer Y die Schranke aus a) an einer Stelle $\vartheta \in \Theta$ aus, d.h. gilt

$$Cov_\vartheta(Y) \;=\; V_\vartheta\, I_\vartheta^{-1}\, V_\vartheta^\top$$

für ein $\vartheta \in \Theta$, so besitzt der Schätzfehler von Y unter ϑ die Darstellung

$$Y - \gamma(\vartheta) \;=\; V_\vartheta\, I_\vartheta^{-1}\, M_\vartheta \quad P_\vartheta\text{–fast sicher}\,.$$

c) Spezialfall $\gamma = id$ unter (+) und (++): ein quadratintegrierbarer erwartungstreuer Schätzer Y für den unbekannten Parameter, welcher (++) erfüllt, erlaubt unter ϑ eine Darstellung

$$Y - \vartheta \;=\; I_\vartheta^{-1}\, M_\vartheta \;+\; \text{ orthogonale Terme in } L^2(P_\vartheta)$$

der Schätzfehler, und damit gilt die Schranke

$$Cov_\vartheta(Y) \;\geq\; I_\vartheta^{-1}\,.$$

Beweis: Fixiere einen Punkt $\vartheta \in \Theta$ mit (+) und (++). Nach Definition des Score in 1.2 gilt $M_\vartheta \in L^2(P_\vartheta)$ und $E_\vartheta(M_\vartheta) = 0$, und I_ϑ ist als Kovarianzmatrix symmetrisch und nichtnegativ definit. Betrachte $Y \in L^2(P_\vartheta)$ mit $E_\vartheta(Y) = \gamma(\vartheta)$.

1) Definiere V_ϑ durch die rechte Seite von (++), i.e. $V_\vartheta := E_\vartheta\left(Y\, M_\vartheta^\top\right)$, und betrachte die $\mathbb{R}^k$-wertige Zufallsvariable

$$W \;:=\; (Y - E_\vartheta(Y)) - V_\vartheta\, I_\vartheta^{-1}\, M_\vartheta \;.$$

Dann gilt $W \in L^2(P_\vartheta)$, $E_\vartheta(W) = 0$, und man berechnet die Kovarianzmatrix von W unter P_ϑ:

$$\begin{aligned} Cov_\vartheta(W) &= E_\vartheta(W\, W^\top) \\ &= Cov_\vartheta(Y) \;-\; E_\vartheta\left(\,(Y - \gamma(\vartheta))\, M_\vartheta^\top\, I_\vartheta^{-1}\, V_\vartheta^\top\,\right) \\ &\quad - E_\vartheta\left(V_\vartheta\, I_\vartheta^{-1}\, M_\vartheta\, (Y - \gamma(\vartheta))^\top\right) \;+\; E_\vartheta\left(V_\vartheta\, I_\vartheta^{-1}\, M_\vartheta\, M_\vartheta^\top\, I_\vartheta^{-1}\, V_\vartheta^\top\right) \;. \end{aligned}$$

Wegen $E_\vartheta(M_\vartheta) = 0$ reduziert sich der zweite Summand auf der rechten Seite dieser Gleichung zu $-V_\vartheta\, I_\vartheta^{-1}\, V_\vartheta^\top$, der dritte liefert dasselbe, der vierte ergibt $+V_\vartheta\, I_\vartheta^{-1}\, V_\vartheta^\top$ nach Definition der Information. Daher gilt

$$0 \;\le\; Cov_\vartheta(W) \;=\; Cov_\vartheta(Y) \;-\; V_\vartheta\, I_\vartheta^{-1}\, V_\vartheta^\top$$

im Sinne der Halbordnung auf symmetrischen und nichtnegativ definiten Matrizen, und damit die Behauptung a). Ein Gleichheitszeichen in der letzten Ungleichung erzwingt $W = E_\vartheta(W) = 0$ P_ϑ-fast sicher, woraus für Y unter P_ϑ die in b) genannte Darstellung folgt.
2) Das in 1) gegebene Argument zeigt ebenfalls

$$W \;\perp\; V_\vartheta\, I_\vartheta^{-1}\, M_\vartheta \quad \text{in } L^2(P_\vartheta)$$

und impliziert damit nach Definition von W eine Darstellung des Schätzfehlers unter ϑ als

$$Y - \gamma(\vartheta) \;=\; V_\vartheta\, I_\vartheta^{-1}\, M_\vartheta \;+\; \text{ orthogonale Terme in } L^2(P_\vartheta) \;.$$

Beachte, daß bis jetzt nur (+), nicht aber (++) benutzt wurde.
2) Im Spezialfall $k = d$ und $\gamma = id$ kommt nun erstmalig die Voraussetzung (++) wirklich ins Spiel: unter (++) ist V_ϑ die Identitätsmatrix I_d, und damit folgt c) aus 1) und 2).
3) Im allgemeinen Fall wie in Beweisschritt 1) erlaubt die Voraussetzung (++) eine anschauliche Interpretation der erzielten Schranke $V_\vartheta\, I_\vartheta^{-1}\, V_\vartheta^\top$, in Termen der Fisher-Information und der Jacobi-Matrix der zu schätzenden Kenngröße. □

1.7 Bemerkung: In natürlich parametrisierten d-parametrigen Exponentialfamilien mit offenem Parameterraum ist die Vertauschungsbedingung (++) erfüllt, siehe Barra (1981, Kap.

X und Lemma 1 in Kap. XI.1) oder Witting (1985, S. 153), und der Score in ϑ ist durch die bezüglich ϑ zentrierte kanonische Statistik der Exponentialfamilie gegeben. Siehe auch Barndorff-Nielsen (1988, Kap. 2.4). Ein heuristisches Argument für (++) lautet so: im Experiment $\mathcal{E}$ sollte die Funktion

$$\vartheta \;\longrightarrow\; \gamma(\vartheta) \;=\; E_\vartheta(Y) \;=\; \int \mu(d\omega)\, f(\vartheta,\omega)\, Y(\omega)$$

unter dem Integralzeichen differenziert werden dürfen, mit partiellen Ableitungen der Gestalt

$$\begin{aligned} \frac{\partial}{\partial\vartheta_j} E_\vartheta(Y_i) &= \int \mu(d\omega)\, \frac{\partial}{\partial\vartheta_j} f(\vartheta,\omega)\; Y_i(\omega) \\ &= \int P_\vartheta(d\omega)\, \frac{\partial}{\partial\vartheta_j} \log f(\vartheta,\omega)\; Y_i(\omega) \quad = \quad E_\vartheta\left((Y\, M_\vartheta^\top)_{i,j} \right) . \end{aligned}$$

Auf die Vertauschungsbedingung (++) werden wir noch in 1.10 und 1.12 ausführlich eingehen.

1.7' Bemerkung: a) Die Kovarianzmatrix $Cov_\vartheta(Y) \;=\; E_\vartheta\left((Y-\gamma(\vartheta))(Y-\gamma(\vartheta))^\top\right)$ beschreibt die Streuung von Y um $\gamma(\vartheta)$ in $\mathbb{R}^k$, unter wahrem ϑ, und liefert damit ein Maß für den Schätzfehler eines erwartungstreuen und quadratintegrierbaren Schätzers für γ an der Stelle ϑ. Die *Inverse der Fisher-Information* I_ϑ^{-1} beschreibt nach 1.6 c) eine minimale Streuung bzw. eine optimale Konzentration der Verteilung eines erwartungstreuen und quadratintegrierbaren Schätzers für den unbekannten Parameter an der Stelle ϑ. Dabei handelt es sich allerdings um eine Schranke, die nur in wenigen klassischen Modellen eine *erreichbare Schranke* ist, siehe Georgii (2002, S. 198), oder Witting (1985, S. 312–317).

b) Nur in wenigen klassischen Schätzproblemen ist die Eigenschaft eines Schätzers relevant, erwartungstreu für den unbekannten Parameter zu sein. In vielen statistischen Problemen hat man keinen Schätzer mit dieser Eigenschaft zur Hand. Darüberhinaus zeigt ein berühmtes Beispiel von Stein (Ibragimov und Has'minskii 1981, S. 26), daß sogar in Normalverteilungsmodellen

$$\left(\, \bigtimes_{i=1}^n \mathbb{R}^k,\; \bigotimes_{i=1}^n \mathcal{B}(\mathbb{R}^k),\; \{ \bigotimes_{i=1}^n \mathcal{N}(\vartheta, I_k) : \vartheta\in\Theta\} \,\right) , \quad \Theta := \mathbb{R}^k \;,\; k\geq 3$$

nicht-erwartungtreue Schätzer existieren, welche weniger breit um den wahren Parameterwert ϑ streuen als der in diesem Modell beste erwartungstreue quadratintegrierbare Schätzer (der empirische Mittelwert).

In der folgenden Ungleichung dagegen dürfen *beliebige* Schätzer betrachtet werden. Wir geben sie in Dimension $d = 1$ (multivariate Verallgemeinerungen existieren, siehe Gill und Levit (1995)) und für strikt positive Dichten. Man liest die unten erzielte Schranke als ein integriertes Risiko,

oder man geht aus von einem 'Bayes-Ansatz', indem man den wahren der Beobachtung zugrundeliegenden Parameter als eine Größe auffaßt, die von einem Wahrscheinlichkeitsmaß auf dem Parameterraum ausgewürfelt wird.

1.8 Satz: (van Trees Ungleichung, um 1968, Beweis nach Gill und Levit (1995)) Betrachte ein Experiment $\mathcal{E} := (\Omega, \mathcal{A}, \{P_\vartheta : \vartheta \in \Theta\})$ mit Score $\{M_\vartheta : \vartheta \in \Theta\}$ und Fisher-Information $\{I_\vartheta : \vartheta \in \Theta\}$ wie in 1.2, mit strikt positiven Dichten $f_\vartheta = \frac{dP_\vartheta}{d\mu}$ bezüglich des dominierenden Maßes μ. Sei Θ ein offenes Intervall in $\mathbb{R}$; die zu schätzende Kenngröße $\gamma : \Theta \to \mathbb{R}$ sei differenzierbar. Lege eine Lebesgue-dominierte *a priori Verteilung* π auf ein Teilintervall (a, b) von Θ so daß

i) $\frac{d\pi}{d\lambda} =: g$ ist differenzierbar auf $\mathbb{R}$, strikt positiv auf (a, b), und $\equiv 0$ außerhalb

ii) $$\int_{(a,b)} \left(\frac{g'}{g}\right)^2 d\pi \;=:\; J \;<\; \infty$$

gelten (J ist die Fisher–Information in dem von π auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ erzeugten Lokationsmodell, vgl. 1.4). Ist einer der Punkte a oder b zugleich Randpunkt von Θ, so sei $\gamma : \Theta \to \mathbb{R}$ an dieser Stelle stetig fortsetzbar durch einen Grenzwert in $\mathbb{R}$.
Dann gilt für jeden im Experiment $\mathcal{E}$ möglichen Schätzer T für γ die Abschätzung

$$\int_{(a,b)} E_\vartheta\left((T - \gamma(\vartheta))^2\right)\, \pi(d\vartheta) \quad \geq \quad \frac{\left(\int_{(a,b)} \gamma'(\vartheta)\, \pi(d\vartheta)\right)^2}{\int_{(a,b)} I_\vartheta\, \pi(d\vartheta) \;+\; J}\,.$$

Beweis: 1) Wir zeigen zuerst, daß die in 1.2 gemachten Annahmen die Produktmeßbarkeit

(∗) $$\Theta \times \Omega \;\ni\; (\vartheta, \omega) \;\to\; f(\vartheta, \omega) \;\in\; (0, \infty) \quad \text{ist } \mathcal{B}(\Theta) \otimes \mathcal{A}\text{- meßbar}$$

der Dichten implizieren: nach 1.2 gilt

$$\forall\, \vartheta \in \Theta: \quad f(\vartheta, \cdot) : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})) \quad \text{ist meßbar}$$

$$\forall\, \omega \in \Omega: \quad f(\cdot, \omega) : \Theta \to \mathbb{R} \quad \text{ist stetig}$$

wobei Θ offen in $\mathbb{R}$. Folglich hat man Messbarkeit in (ϑ, ω) aller Abbildungen

$$\varphi_k(\vartheta, \omega) \;:=\; \sum_{j \in \mathbb{Z},\, j2^{-k} \in \Theta} 1_{]\frac{j}{2^k}, \frac{j+1}{2^k}] \cap \Theta}(\vartheta)\; f(\frac{j}{2^k}, \omega)\,, \quad k \geq 1$$

und damit Meßbarkeit in (ϑ, ω) des punktweisen Limes $f = \lim_{k\to\infty} \varphi_k$.

2) Unter Produktmeßbarkeit der Dichten gemäß Schritt 1) wird

$$(\vartheta, A) \;\longrightarrow\; P_\vartheta(A) \;=\; \int 1_A(\omega)\, f(\vartheta,\omega)\, \mu(d\omega)\;, \quad \vartheta \in \Theta\;,\; A \in \mathcal{A}$$

zu einer Übergangswahrscheinlichkeit von $(\Theta, \mathcal{B}(\Theta))$ nach $(\Omega, \mathcal{A})$.

3) Sei $T : (\Omega, \mathcal{A}) \to (I\!R, \mathcal{B}(I\!R))$ als Schätzer für γ quadratintegrierbar mit

$$\int_{(a,b)} E_\vartheta\left((T-\gamma(\vartheta))^2\right)\, \pi(d\vartheta) \;<\; \infty$$

(sonst wäre nichts zu zeigen). Auch reicht es für den Beweis, nur die Einschränkung des Parameterraums Θ auf seine Teilmenge (a,b) zu betrachten. Wir identifizieren also Θ mit (a,b) – damit ist auch g strikt positiv auf Θ – und arbeiten auf dem Produktraum

$$\left(\overline{\Omega}, \overline{\mathcal{A}}\right) \;:=\; (\Theta\times\Omega, \mathcal{B}(\Theta)\otimes\mathcal{A})$$

versehen mit dem Wahrscheinlichkeitsmaß

$$\overline{P}(d\vartheta, d\omega) \;:=\; \pi(d\vartheta)\, P_\vartheta(d\omega) \;:=\; (\lambda\!\!\!\lambda\otimes\mu)(d\vartheta, d\omega)\; g(\vartheta) f(\vartheta,\omega)\;, \quad \vartheta\in\Theta\;,\; \omega\in\Omega\;.$$

4) Bezeichne $'$ die Ableitung nach dem Parameter. Für festes $\omega \in \Omega$ gilt

$$(+) \qquad \int_\Theta d\vartheta\; \left(f(\vartheta,\omega) g(\vartheta)\right)' \;=\; f(b,\omega)g(b) - f(a,\omega)g(a) \;=\; 0$$

mit $g(a) = 0 = g(b)$ nach Voraussetzung. Auch $\gamma(a), \gamma(b) \in I\!R$ sind nach Voraussetzung wohldefiniert, und partielle Integration zeigt wegen $(+)$

$$(++) \qquad \int_\Theta d\vartheta\, \gamma(\vartheta)\; \left(f(\vartheta,\omega)g(\vartheta)\right)' \;=\; -\int_\Theta d\vartheta\, \gamma'(\vartheta)\; \left(f(\vartheta,\omega)g(\vartheta)\right)$$

bei festem $\omega \in \Omega$. Aus $(+)$ und $(++)$ erhält man die Gleichung

$$\begin{aligned}
&\int_{\Theta\times\Omega} (\lambda\!\!\!\lambda\otimes\mu)(d\vartheta, d\omega)\; \left(f(\vartheta,\omega)g(\vartheta)\right)'\, (T(\omega)-\gamma(\vartheta)) \\
&= \;0 \;+\; \int_\Omega \mu(d\omega) \int_\Theta d\vartheta\; \gamma'(\vartheta)\, f(\vartheta,\omega)\, g(\vartheta) \\
&= \int_\Theta \pi(d\vartheta)\, \gamma'(\vartheta)\;.
\end{aligned}$$

Die linke Seite dieser Gleichungskette kann man aber umschreiben in

$$\begin{aligned}
&\int_{\Theta\times\Omega} (\lambda\!\!\!\lambda\otimes\mu)(d\vartheta, d\omega)\; \left(f(\vartheta,\omega)g(\vartheta)\right)'\, (T(\omega)-\gamma(\vartheta)) \\
&= \int_\Theta \int_\Omega \pi(d\vartheta) P_\vartheta(d\omega)\; \frac{\left(f(\vartheta,\omega)g(\vartheta)\right)'}{f(\vartheta,\omega)g(\vartheta)}\; (T(\omega)-\gamma(\vartheta)) \\
&= \int_{\Theta\times\Omega} \overline{P}(d\vartheta, d\omega)\; \left(\frac{g'}{g}(\vartheta) + \frac{f'}{f}(\vartheta,\omega)\right)\; (T(\omega)-\gamma(\vartheta))
\end{aligned}$$

hier wurde die Positivität der Dichten benutzt. Im letzten Integranden sind die durch runde Klammern bezeichneten Faktoren in $L^2(\overline{P})$: der zweite Faktor ist Schätzfehler eines nach Voraussetzung quadratintegrierbaren Schätzers; der erste Faktor ist vom Typ Score. Im ersten Faktor sind die logarithmischen Ableitungen orthogonal in $L^2(\overline{P})$:

$$(+++) \qquad \int_{\Theta\times\Omega} \overline{P}(d\vartheta, d\omega)\, \frac{g'}{g}(\vartheta)\, \frac{f'}{f}(\vartheta,\omega) \;=\; \int_\Theta \pi(d\vartheta)\, \frac{g'}{g}(\vartheta) \int_\Omega P_\vartheta(d\omega)\, \frac{f'}{f}(\vartheta,\omega) \;=\; 0\,.$$

Zusammen erhält man wegen (+++) nun mit Cauchy-Schwartz:

$$\begin{aligned}
\left(\int_\Theta \pi(d\vartheta)\, \gamma'(\vartheta)\right)^2 \;&=\; \left(\int_{\Theta\times\Omega} \overline{P}(d\vartheta, d\omega)\; \left(\frac{g'}{g}(\vartheta) + \frac{f'}{f}(\vartheta,\omega)\right)\; (T(\omega) - \gamma(\vartheta))\right)^2 \\
&\le\; \int_{\Theta\times\Omega} \overline{P}(d\vartheta, d\omega)\; \left(\frac{g'}{g}(\vartheta) + \frac{f'}{f}(\vartheta,\omega)\right)^2 \;\cdot\; \int_{\Theta\times\Omega} \overline{P}(d\vartheta, d\omega)\; (T(\omega) - \gamma(\vartheta))^2 \\
&=\; \left(J + \int_\Theta \pi(d\vartheta)\, I_\vartheta\right) \;\cdot\; \int_\Theta \pi(d\vartheta)\, E_\vartheta\left((T - \gamma(\vartheta))^2\right)
\end{aligned}$$

und damit die Behauptung. □

B. Schätzfolgen und Asymptotik der Informationsschranken

1.9 Definition: Betrachte eine Folge von Experimenten

$$\mathcal{E}_n \;=\; (\Omega_n,\, \mathcal{A}_n,\, \{P_{n,\vartheta} : \vartheta \in \Theta\})\,, \quad n \ge 1$$

wobei alle $\mathcal{E}_n$ durch *dieselbe* Menge $\Theta \subset \mathbb{R}^d$ parametrisiert sind. Sei $\gamma : \Theta \to \mathbb{R}^k$ eine Abbildung. Eine *Schätzfolge für* γ ist eine Folge $(Y_n)_n$ meßbarer Abbildungen

$$Y_n : (\Omega_n,\, \mathcal{A}_n) \to (\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))\,, \quad n \ge 1\,.$$

a) Eine Schätzfolge $(Y_n)_n$ heißt *konsistent für* γ falls gilt

$$\text{für jedes } \vartheta \in \Theta, \text{ jedes } \varepsilon > 0\text{:} \quad \lim_{n\to\infty} P_{n,\vartheta}\left(|Y_n - \gamma(\vartheta)| > \varepsilon\right) \;=\; 0$$

($(P_{n,\vartheta})_n$-stochastische Konvergenz der Folge $(Y_n)_n$ gegen $\gamma(\vartheta)$, für jedes $\vartheta \in \Theta$).

b) Eine Schätzfolge $(Y_n)_n$ heißt $(\varphi_n)_n$*-konsistent für* γ falls gilt

$$\text{für jedes } \vartheta \in \Theta \text{ ist die Familie} \quad \mathcal{L}\left(\varphi_n(\vartheta)(Y_n - \gamma(\vartheta)) \mid P_{n,\vartheta}\right)\,,\; n \ge 1\,, \quad \text{straff in } \mathbb{R}^k\,.$$

Hierbei bezeichnet $(\varphi_n)_n$ eine (im allgemeinen parameterabhängige) Schar von normierenden Folgen, entweder in $(0,\infty)$ mit der Eigenschaft $\varphi_n(\vartheta) \uparrow \infty$ für $n \to \infty$, für jedes $\vartheta \in \Theta$, oder allgemeiner im Raum der invertierbaren $k\times k$-Matrizen mit $|\det \varphi_n(\vartheta)| \uparrow \infty$ für $n \to \infty$.

c) Eine $(\varphi_n)_n$-konsistente Schätzfolge $(Y_n)_n$ für γ heißt *asymptotisch normal* falls gilt

$$\text{für jedes } \vartheta \in \Theta: \quad \mathcal{L}\left(\varphi_n(\vartheta)(Y_n - \gamma(\vartheta)) \mid P_{n,\vartheta}\right) \longrightarrow \mathcal{N}(0, \Sigma(\vartheta)) \ , \quad n \to \infty$$

(schwache Konvergenz in $\mathbb{R}^k$), mit geeigneten Normalverteilungen $\mathcal{N}(0, \Sigma(\vartheta))$, $\vartheta \in \Theta$.

Für den Rest des Teilkapitels konzentrieren wir uns auf unabhängige Versuchswiederholung. Sei

$$\mathcal{E} \ := \ (\Omega, \mathcal{A}, \{P_\vartheta : \vartheta \in \Theta\}) \ , \quad \Theta \subset \mathbb{R}^d \text{ offen}$$

ein Experiment mit Score $\{M_\vartheta : \vartheta \in \Theta\}$ und Fisher-Information $\{I_\vartheta : \vartheta \in \Theta\}$ wie in 1.2; für $n \geq 1$ betrachten wir Produktmodelle

$$(\diamond) \qquad \mathcal{E}_n \ := \ (\Omega_n, \mathcal{A}_n, \{P_{n,\vartheta} : \vartheta \in \Theta\}) \ = \ \left(\bigtimes_{i=1}^n \Omega, \ \bigotimes_{i=1}^n \mathcal{A}, \ \{P_{n,\vartheta} := \bigotimes_{i=1}^n P_\vartheta : \vartheta \in \Theta\} \right)$$

mit Score $M_{n,\vartheta}$ in $\vartheta \in \Theta$ und Information $I_{n,\vartheta} = n\, I_\vartheta$ wie in 1.5 a). Zunächst sei die beliebte Aussage 'die Cramér-Rao-Schranke ist asymptotisch scharf' kommentiert.

1.10 Bemerkung: (Asymptotische Cramér-Rao-Schranke) Betrachte $(\mathcal{E}_n)_n$ wie in $(\diamond)$. Sei I_ϑ invertierbar für alle $\vartheta \in \Theta$. Sei $(T_n)_n$ eine Folge von erwartungstreuen und quadratintegrierbaren Schätzern für den unbekannten Parameter, $\sqrt{n}$-konsistent und asymptotisch normal

$$(\circ) \qquad \text{für jedes } \vartheta \in \Theta: \quad \mathcal{L}\left(\sqrt{n}(T_n - \vartheta) \mid P_{n,\vartheta}\right) \longrightarrow \mathcal{N}(0, \Sigma(\vartheta))$$

(schwache Konvergenz in $\mathbb{R}^d$, für $n \to \infty$). Die Cramér-Rao-Schranke in $\mathcal{E}_n$ besagt

$$E_\vartheta\left([\sqrt{n}(T_n - \vartheta)][\sqrt{n}(T_n - \vartheta)]^\top\right) \ = \ n\, Cov_\vartheta(T_n) \ \geq \ n\, I_{n,\vartheta}^{-1} \ = \ I_\vartheta^{-1} \ .$$

Dies wird man dahingehend interpretieren, daß die 'bestmögliche' Limesvarianz in $(\circ)$ eben genau durch $\Sigma(\vartheta) = I_\vartheta^{-1}$ gegeben ist, $\vartheta \in \Theta$. Da man die Grenzverteilung $\Sigma(\vartheta) = I_\vartheta^{-1}$, $\vartheta \in \Theta$, tatsächlich in vielen Modellen durch geeignete Wahl einer Schätzfolge $(T_n)_n$ erreichen kann, wird man die Cramér-Rao-Schranke als 'asymptotisch scharf' bezeichnen.
Im Grunde ist dieses Argument ein einziger wohlversteckter Zirkelschluß: wir erläutern dies im einfachsten Fall $d = 1$. Fixiere $\vartheta \in \Theta$. Cramér-Rao braucht als Voraussetzung die Vertauschungsbedingung $(++)$ aus 1.6, hier für $\gamma = id$, und natürlich $(*)$ aus 1.2, also

$$E_\vartheta\,(M_{n,\vartheta}) = 0 \ , \quad E_\vartheta\,(T_n\, M_{n,\vartheta}) = 1 \ , \quad n \geq 1 \ ,$$

und damit insbesondere

$$(*) \qquad E_\vartheta\left([T_n - \vartheta]\, M_{n,\vartheta}\right) \;=\; 1\;, \quad n \geq 1\;.$$

Damit stellt die Vertauschungsbedingung (++) aus 1.6 einen *äußerst engen Zusammenhang unter jedem ϑ zwischen der Folge der Scores $(M_{n,\vartheta})_n$ und denjenigen Schätzfolgen $(T_n)_n$ her, auf die Cramér-Rao überhaupt angewandt werden darf*:
zerlegt man den Schätzfehler von T_n, indem man den Raum $L^2(\Omega_n, \mathcal{A}_n, P_{n,\vartheta})$ in den von $M_{n,\vartheta}$ erzeugten eindimensionalen Unterraum $U_{n,\vartheta}$ und dessen orthogonales Komplement $U_{n,\vartheta}^{\perp}$ aufspaltet, erhält man eine Darstellung

$$[T_n - \vartheta] \;=\; \alpha \cdot M_{n,\vartheta} \;+\; R(n,\vartheta)\;, \quad \alpha \in I\!R\;, \quad R(n,\vartheta) \in U_{n,\vartheta}^{\perp}\;.$$

Die Vertauschungsbedingung (++) von Cramér-Rao – hier in Form $(*)$, wobei gleichzeitig $E_\vartheta(M_{n,\vartheta}^2) = I_{n,\vartheta}$ gilt – erzwingt in dieser Darstellung $\alpha = I_{n,\vartheta}^{-1}$. Damit kann man mit Cramér-Rao nur Schätzfolgen $(T_n)_n$ betrachten, die unter $\vartheta \in \Theta$ gemäß

$$(\circ\circ) \qquad T_n - \vartheta \;=\; I_{n,\vartheta}^{-1}\, M_{n,\vartheta} \;+\; \text{ orthogonale Terme in } L^2(\Omega_n, \mathcal{A}_n, P_{n,\vartheta})\;, \quad n \geq 1$$

darstellbar sind: für diese aber ist die Aussage von Cramér-Rao trivial. In die Vertauschungsbedingung (++) aus 1.6 hat man also sehr viel hereingesteckt, vielleicht sogar mehr, als man am Ende in der Cramér-Rao-Schranke wieder herausholt.

Wir werden in späteren Kapiteln sorgfältig herausarbeiten, wie der – tatsächlich entscheidend wichtige – enge Zusammenhang zwischen der Folge der Scores (genauer: 'Score/Information') und 'optimalen' Schätzern (in einem zu präzisierenden Sinn) zu formulieren ist.

Im Unterschied zu Cramér-Rao liefert van Trees eine asymptotisch nützliche Schranke:

1.11 Satz: (Asymptotische van Trees-Schranke) Sei $d = 1$, Θ ein offenes Intervall in $I\!R$, betrachte unabhängige Versuchswiederholung $(\diamond)$ unter der Zusatzvoraussetzung strikt positiver Dichten $(\vartheta, \omega) \to f(\vartheta, \omega)$ wie in 1.8; setze weiter voraus

$$(\diamond\diamond) \qquad \Theta \;\ni\; \vartheta \;\to\; I_\vartheta \;\in\; (0, \infty) \quad \text{ist stetig.}$$

Dann gilt unter Berücksichtigung *aller* möglichen Schätzfolgen $(T_n)_n$ für den unbekannten Parameter an jeder Stelle $\vartheta_0 \in \Theta$

$$\lim_{c\downarrow 0} \liminf_{n\to\infty} \inf_{T_n\ \mathcal{A}_n\text{-mb}} \sup_{|\vartheta-\vartheta_0|<c} E_\vartheta\left([\sqrt{n}(T_n-\vartheta)]^2\right) \quad \geq \quad I_{\vartheta_0}^{-1}\,.$$

Beweis: Bis auf notationelle Änderungen genügt es, den Fall $\Theta = I\!R$ und $\vartheta_0 = 0$ zu betrachten. Wähle ein Wahrscheinlichkeitsmaß π auf $(I\!R, \mathcal{B}(I\!R))$ mit differenzierbarer Lebesgue-Dichte, welche auf $(-1,+1)$ strikt positiv und außerhalb $\equiv 0$ ist. Mit $g := \frac{d\pi}{d\lambda}$ setze voraus $J := \int (\frac{g'}{g})^2 d\pi < \infty$. Dann hat π alle in 1.8 i)–ii) geforderten Eigenschaften.
Dasselbe gilt dann für alle durch Skalieren entstehenden Wahrscheinlichkeitsmaße: für $c > 0$ ist $\pi_c(d\vartheta) := \frac{1}{c}\, g(\frac{\vartheta}{c})\, d\vartheta$ konzentriert auf $(-c,+c)$, und die Fisher-Information $J_c := \frac{1}{c^2}\, J$ in dem von π_c erzeugten Lokationsmodell ist endlich.
Für einen quadratintegrablen Schätzer T_n für $\gamma = id$ im Modell $\mathcal{E}_n$ liefert van Trees 1.8

$$\begin{aligned}\sup_{|\vartheta-\vartheta_0|<c} E_\vartheta\left([\sqrt{n}(T_n-\vartheta)]^2\right) &\geq \int_{-c}^{+c} \pi_c(d\vartheta)\; E_\vartheta\left([\sqrt{n}(T_n-\vartheta)]^2\right)\\ &\geq n\,\frac{1}{\int_{-c}^{+c} I_{n,\vartheta}\,\pi_c(d\vartheta)\;+\;J_c}\\ &= \frac{1}{\int_{-c}^{+c} I_\vartheta\,\pi_c(d\vartheta)\;+\;\frac{1}{c^2 n}\,J}\,;\end{aligned}$$

läßt man darüberhinaus nicht-quadratintegrierbare $\mathcal{A}_n$-meßbare Schätzer zur Konkurrenz zu, wird die linke Seite der eben gegebenen Abschätzung nur in trivialer Weise vergrößert. Also hat man

$$\inf_{T_n\ \mathcal{A}_n\text{-mb}} \sup_{|\vartheta-\vartheta_0|<c} E_\vartheta\left([\sqrt{n}(T_n-\vartheta)]^2\right) \quad \geq \quad \frac{1}{\int_{-c}^{+c} I_\vartheta\,\pi_c(d\vartheta)\;+\;\frac{1}{c^2 n}\,J}$$

für festes $c > 0$, folglich für $n \to \infty$

$$\liminf_{n\to\infty} \inf_{T_n\ \mathcal{A}_n\text{-mb}} \sup_{|\vartheta-\vartheta_0|<c} E_\vartheta\left([\sqrt{n}(T_n-\vartheta)]^2\right) \quad \geq \quad \frac{1}{\int_{-c}^{+c} I_\vartheta\,\pi_c(d\vartheta)}\,.$$

Läßt man nun c gegen 0 streben, erhält man wegen $(\diamond\diamond)$ die Behauptung. □

Satz 1.11 liefert eine Abschätzung vom 'Minimax'-Typ: bei Verwendung bestmöglicher Schätzer (wobei auf der Stufe n der Asymptotik *alle* im Modell $\mathcal{E}_n$ zur Verfügung stehenden Schätzer T_n zur Konkurrenz zugelassen sind) *minimiert man ein maximales Risiko* auf kleinen Kugeln um einen fest herausgegriffenen Punkt ϑ_0 des Parameterraums Θ. Die asymptotische van Trees-Schranke zeigt bei unabhängiger Versuchwiederholung, daß man in diesem Sinn beim Schätzen

des unbekannten Parameters in der Nähe eines beliebigen Parameterwertes ϑ_0 die Risikoschranke $I_{\vartheta_0}^{-1}$ asymptotisch nicht unterbieten kann.

Wir illustrieren nun die Nützlichkeit der asymptotischen van Trees Schranke durch das folgende Beispiel. Im nichtparametrischen Modell $\mathcal{F}$ aller Verteilungsfunktionen auf $\mathbb{R}$ betrachten wir für einen festgelegten Punkt $x \in \mathbb{R}$ die zu schätzende Kenngröße

$$\gamma \; : \quad \mathcal{F} \; \ni \; F \; \longrightarrow \; \gamma(F) := F(x) \; \in \; [0,1]$$

Wir wollen nachweisen, daß die empirische Verteilungsfunktion via $T_n := \widehat{F}_n(x)$ einen asymptotischen Minimax-Schätzer liefert, wobei '... max' auf das maximale quadratische Risiko auf Umgebungen von F der Form

$$V_\delta(F) \; := \; \left\{ \widetilde{F} \in \mathcal{F} : \; \sup_{y \in \mathbb{R}} |\widetilde{F}(y) - F(y)| < \delta \right\}$$

mit 'kleinem' Radius $\delta > 0$ Bezug nimmt. Seien X_i, $i \geq 1$, i.i.d ZV auf $(\Omega, \mathcal{A})$ mit unbekannter Verteilung $F \in \mathcal{F}$, und $\mathcal{A}_n := \sigma(X_1, \ldots, X_n)$ die von den ersten n Beobachtungen erzeugte Sub-σ-Algebra auf $(\Omega, \mathcal{A})$, $n \geq 1$. Eine Aussage der Art 'bei unbekanntem $F \in \mathcal{F}$ liefert die empirische Verteilungsfunktion einen asymptotisch optimalen Schätzer für F' wurde zuerst von Beran (1977) bewiesen. In der hier gegebenen Form (Schätzung von $F(x)$ an einer festen Stelle x unter unbekanntem $F \in \mathcal{F}$, mit van Trees 1.11) stammen Satz und Beweis von Kutoyants (1997).

1.11' Satz: (Kutoyants) Für $F \in \mathcal{F}$ und $x \in \mathbb{R}$ beliebig gilt mit $V_\delta(F) \subset \mathcal{F}$ wie oben

i) $$\lim_{\delta \downarrow 0} \liminf_{n \to \infty} \inf_{T_n \, \mathcal{A}_n\text{-mb}} \sup_{\widetilde{F} \in V_\delta(F)} E_{\widetilde{F}} \left(n \, [T_n - \widetilde{F}(x)]^2 \right) \quad \geq \quad F(x)(1 - F(x)) \; ,$$

ii) $$\lim_{\delta \downarrow 0} \lim_{n \to \infty} \sup_{\widetilde{F} \in V_\delta(F)} E_{\widetilde{F}} \left(n \, [\widehat{F}_n(x) - \widetilde{F}(x)]^2 \right) \quad = \quad F(x)(1 - F(x)) \; .$$

Beweis: Fixiere $F \in \mathcal{F}$ und $x \in \mathbb{R}$. Zuerst ist ii) einfach einzusehen: aus

$$\widehat{F}_n(x) - \widetilde{F}(x) \; = \; \frac{1}{n} \sum_{i=1}^{n} \left(1_{(-\infty, x]}(Y_i) - \widetilde{F}(x) \right)$$

und Unabhängigkeit der $Y_i \sim \widetilde{F}$ folgt sofort

$$E_{\widetilde{F}} \left((\widehat{F}_n(x) - \widetilde{F}(x))^2 \right) \; = \; \frac{1}{n} \, \widetilde{F}(x)(1 - \widetilde{F}(x)) \; , \quad n \geq 1$$

wobei nach Definition von $V_\delta(F)$ für kleines δ

$$\sup_{\widetilde{F}\in V_\delta(F)} \widetilde{F}(x)(1-\widetilde{F}(x))$$

nahe bei $F(x)(1-F(x))$ liegt. Im Rest des Beweises zeigen wir ii).

1) Wir definieren geeignete einparametrige Pfade $\mathcal{S}^h$ durch F in Richtungen h. Betrachte

$$h \text{ beschränkt auf } I\!R \ , \quad \int_{-\infty}^{\infty} h\,dF \ = \ 0 \ , \quad \int_{-\infty}^{x} h\,dF \ \neq \ 0$$

und wähle einen Skalierungsfaktor für h so daß gilt

$$\int_{-\infty}^{x} h\,dF \ = \ 1 \ .$$

Sei $\mathcal{H}$ die durch diese vier Eigenschaften bestimmte Teilklasse von $L^2(F)$. Für $h \in \mathcal{H}$ setze $\delta^h := \frac{\delta}{\sup|h|}$, wobei $\delta > 0$ der Radius der Umgebung $V_\delta(F)$ von F ist, und definiere

$$\mathcal{S}^h \ := \ \{\, F^h_\vartheta : |\vartheta| < \delta^h \,\} \ , \quad dF^h_\vartheta \ := \ (1+\vartheta h)\,dF \ .$$

Dies ist ein einparametriger Pfad durch F (es gilt $F^h_0 = F$ für alle $h \in \mathcal{H}$) in $V_\delta(F)$ (wegen $|\vartheta| < \delta^h$). Nach 1.3 ist die Information in $\mathcal{S}^h$ gegeben durch

$$I^h_\vartheta \ := \ \int \frac{h^2}{1+\vartheta h}\,dF \ , \quad |\vartheta| < \delta^h \ .$$

2) Fixiere $h \in \mathcal{H}$. Wegen der für h gewählten Skalierung gilt im Pfad $\mathcal{S}^h$

$$F^h_\vartheta(x) \ = \ \int_{-\infty}^{x} (1+\vartheta h)\,dF \ = \ F(x) + \vartheta \ , \quad |\vartheta| < \delta^h \ .$$

Also arbeitet jeder $\mathcal{A}_n$-meßbare Schätzer T_n für $\gamma : \mathcal{F} \ni \widetilde{F} \longrightarrow \widetilde{F}(x) \in [0,1]$ in Einschränkung auf $\mathcal{S}^h$ als Schätzer für

$$(+) \qquad \gamma : \quad \mathcal{S}^h \ \ni \ F^h_\vartheta \ \longrightarrow \ F(x) + \vartheta \ .$$

In Einschränkung auf $\mathcal{S}^h$ kann man assoziiert zu γ eine neue Kenngröße

$$\overline{\gamma}^h(\vartheta) \ := \ \gamma(F^h_\vartheta) - F(x) \ = \ \vartheta \ , \quad F^h_\vartheta \in \mathcal{S}^h$$

betrachten, also

$$\overline{\gamma}^h := id \quad \text{auf} \quad \{\vartheta : |\vartheta| < \delta^h\} \ ,$$

und jeden $\mathcal{A}_n$-meßbaren Schätzer T_n für γ in Restriktion auf $\mathcal{S}^h$ via

$$\overline{T}_n \ := \ T_n - F(x)$$

zu einem Schätzer für den unbekannten Parameter ϑ im Modell $\mathcal{S}^h$ umbauen. Die asymptotische van Trees Schranke 1.11 für das Schätzen des Parameters im einparametrigen Pfad $\mathcal{S}^h$ lautet

$$\lim_{c\downarrow 0}\ \liminf_{n\to\infty}\ \inf_{\overline{T}_n\ \mathcal{A}_n\text{-mb}}\ \sup_{|\vartheta|<c}\ E_\vartheta\left([\sqrt{n}(\overline{T}_n-\vartheta)]^2\right)\quad\geq\quad\left(I_0^h\right)^{-1}\ .$$

Nun transformiert man das Paar $\overline{T}_n, \overline{\gamma}^h{=}id$ in die ursprüngliche Form T_n, γ zurück und erhält in $\mathcal{S}^h$ die asymptotische Minimax-Schranke

$$(++)\qquad \lim_{c\downarrow 0}\ \liminf_{n\to\infty}\ \inf_{T_n\ \mathcal{A}_n\text{-mb}}\ \sup_{F_\vartheta^h:|\vartheta|<c}\ E_{F_\vartheta^h}\left([\sqrt{n}(T_n-\gamma(F_\vartheta^h))]^2\right)\quad\geq\quad\left(I_0^h\right)^{-1}\ .$$

3) Betrachten wir nun die Schar aller Funktionen $h\in\mathcal{H}$ und die ihnen zugeordneten einparametrigen Pfade $\mathcal{S}^h$ durch F, gilt wegen $(++)$ und wegen $\mathcal{S}^h\subset V_\delta(F)$

$$\begin{aligned}&\lim_{\delta\downarrow 0}\ \liminf_{n\to\infty}\ \inf_{T_n\ \mathcal{A}_n\text{-mb}}\ \sup_{\widetilde{F}\in V_\delta(F)}\ E_{\widetilde{F}}\left([\sqrt{n}(T_n-\gamma(\widetilde{F}))]^2\right)\\ \geq\quad&\lim_{c\downarrow 0}\ \liminf_{n\to\infty}\ \inf_{\overline{T}_n\ \mathcal{A}_n\text{-mb}}\ \sup_{F_\vartheta^h:|\vartheta|<c}\ E_{F_\vartheta^h}\left([\sqrt{n}(T_n-\gamma(F_\vartheta^h))]^2\right)\quad\geq\quad\left(I_0^h\right)^{-1}\end{aligned}$$

für jedes $h\in\mathcal{H}$. Damit ist gezeigt

$$\lim_{\delta\downarrow 0}\ \liminf_{n\to\infty}\ \inf_{T_n\ \mathcal{A}_n\text{-mb}}\ \sup_{\widetilde{F}\in V_\delta(F)}\ E_{\widetilde{F}}\left([\sqrt{n}(T_n-\widetilde{F}(x))]^2\right)\quad\geq\quad\sup\left\{\left(I_0^h\right)^{-1}:h\in\mathcal{H}\right\}\ .$$

4) Wir zeigen nun

$$(+++)\qquad \sup\left\{\left(I_0^h\right)^{-1}:h\in\mathcal{H}\right\}\ =\ F(x)(1-F(x)$$

und schließen damit den Beweis der Schranke i) ab. Unter Ausnutzung der für h gewählten Normierungen gilt in $\mathcal{S}^h$ mit Cauchy-Schwartz

$$\begin{aligned}1\ &=\ \int_{-\infty}^{x}h\,dF(x)\ -\ 0\ =\ \int_{-\infty}^{+\infty}h(y)\,[1_{(-\infty,x]}(y)-F(x)]\,F(dy)\\ &\leq\ \left(\int h^2\,dF\right)^{\frac{1}{2}}\left(\int[1_{(-\infty,x]}(y)-F(x)]^2\,F(dy)\right)^{\frac{1}{2}}\\ &=\ \left(I_0^h\right)^{\frac{1}{2}}\,(\,F(x)(1-F(x))\,)^{\frac{1}{2}}\ .\end{aligned}$$

Gleichheit in Cauchy-Schwartz gilt genau falls

$$(\circ)\qquad h(y)\ =\ c\cdot[1_{(-\infty,x]}(y)-F(x)]\quad\text{für ein } c\in\mathbb{R}\,,$$

was aufgrund der in der Definition der Klasse $\mathcal{H}$ vorgenommenen Normierungen genau im Fall

$$(\circ\circ)\qquad c\ =\ \frac{1}{F(x)(1-F(x))}$$

erreicht wird. Folglich enthält $\mathcal{H}$ genau ein Element $\overline{h}$, gegeben durch ($\circ$) und ($\circ\circ$), mit

$$F(x)(1-F(x)) \;=\; \left(I_0^{\overline{h}}\right)^{-1} \;=\; \sup\left\{\left(I_0^h\right)^{-1} : h \in \mathcal{H}\right\} .$$

Das ist (+++), und 1.11' ist vollständig bewiesen.

Man nennt $\overline{h}$ eine *ungünstigste Richtung* und $\mathcal{S}^{\overline{h}}$ einen *ungünstigsten einparametrigen Pfad* durch F: $\mathcal{S}^{\overline{h}}$ minimiert unter allen Pfaden $\mathcal{S}^h$, $h \in \mathcal{H}$, die Fisher-Information in 0. □

C. Heuristik zu Maximum-Likelihood-Schätzfolgen

Der große englische Statistiker R.A. Fisher, zwischen 1915 und 1950 Vater vieler wichtiger statistischer Verfahren, war davon überzeugt, daß in fast allen relevanten statistischen Problemen ein *Maximum-Likelihood*-Verfahren (ML) zu Schätzern führen sollte, die bei unabhängiger Versuchswiederholung $\sqrt{n}$-konsistent und asymptotisch normal sind, wobei

- als Kovarianzmatrix der Grenzverteilung die Inverse der Fisher-Information auftritt,
- keine Schätzfolgen mit besser konzentrierter Grenzverteilung existieren.

Er begündete das mit heuristischen Argumenten, die in diesem Teilabschnitt geschildert werden sollen. Kern des Arguments ist eine Darstellung der reskalierten ML-Schätzfehler in Termen von Score und Information, siehe (h8) in 1.14 unten. Die in seiner Heuristik für ML-Schätzfolgen 'zu sehende' Darstellung reskalierter Schätzfehler wird sich später (siehe Kapitel VII, mit mathematisch jedoch ganz anders ausgerichteten Argumenten) als Schlüssel zur lokalasymptotischen Charakterisierung optimaler Schätzfolgen erweisen.

1.12 Heuristik I: (Vertauschungsbedingungen) Wir betrachten ein Experiment

$$\mathcal{E} \;:=\; (\,\Omega,\, \mathcal{A},\, \{P_\vartheta : \vartheta \in \Theta\}\,) , \quad \Theta \subset \mathbb{R}^d \text{ offen}$$

mit Score $\{M_\vartheta : \vartheta \in \Theta\}$ und Fisher-Information $\{I_\vartheta : \vartheta \in \Theta\}$ wie in 1.2. Dabei seien die Dichten $f_\vartheta = \frac{dP_\vartheta}{d\mu}$ strikt positiv, und die Parametrisierung $\vartheta \to f(\vartheta, \cdot)$ sei hinreichend glatt auf Θ.

Dann sollten im Experiment $\mathcal{E}$ die folgenden *Vertauschungsbedingungen* (h1) und (h2) gelten, wobei wir ein nur heuristikgestütztes erwünschtes Gleichheitszeichen durch die Schreibweise $\overset{(!!)}{=}$ hervorheben. Für geeignete $Y : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ sollte gelten

$$\frac{\partial}{\partial\vartheta_i} E_\vartheta Y \;=\; \frac{\partial}{\partial\vartheta_i} \int Y\, f_\vartheta\, d\mu \overset{(!!)}{=} \int Y\, \frac{\partial}{\partial\vartheta_i} f_\vartheta\, d\mu \;=\; \int Y \left(\frac{\partial}{\partial\vartheta_i} \log f_\vartheta\right) f_\vartheta\, d\mu$$

und damit

$$\nabla(E_\vartheta Y) \overset{(!!)}{=} E_\vartheta(Y \cdot M_\vartheta)\,. \tag{h1}$$

Akzeptiert man (h1), so sollte weiter gelten

$$\begin{aligned}
\frac{\partial}{\partial\vartheta_j}\frac{\partial}{\partial\vartheta_i}E_\vartheta Y &\overset{(!!)}{=} \int Y\,\frac{\partial}{\partial\vartheta_j}\left(\left(\frac{\partial}{\partial\vartheta_i}\log f_\vartheta\right)\cdot f_\vartheta\right)d\mu \\
&= \int Y\left(\left(\frac{\partial}{\partial\vartheta_j}\frac{\partial}{\partial\vartheta_i}\log f_\vartheta\right)\cdot f_\vartheta + \left(\frac{\partial}{\partial\vartheta_i}\log f_\vartheta\right)\left(\frac{\partial}{\partial\vartheta_j}\log f_\vartheta\right)\cdot f_\vartheta\right)d\mu \\
&= E_\vartheta\left(Y\left(\frac{\partial}{\partial\vartheta_j}\frac{\partial}{\partial\vartheta_i}\log f_\vartheta \;+\; (M_\vartheta)_i(M_\vartheta)_j\;\right)\right)\,.
\end{aligned}$$

Im Spezialfall $Y \equiv 1$ ist die linke Seite dieser Gleichungskette gleich 0, und dies führt nach Definition von Score und Fisher-Information auf

$$I_\vartheta \;=\; E_\vartheta\left(M_\vartheta\, M_\vartheta^\top\right) \quad\overset{(!!)}{=}\quad -\,E_\vartheta\left((\nabla\nabla^\top \log f)(\vartheta,\cdot)\right) \tag{h2}$$

wobei man $(\nabla\nabla^\top \log f)(\vartheta,\cdot)$ als P_ϑ-integrierbar versteht. Indem man die erste Zeile der Gleichungskette oben alternativ in der Form

$$\frac{\partial}{\partial\vartheta_j}\frac{\partial}{\partial\vartheta_i}E_\vartheta Y \overset{(!!)}{=} \int Y\,\frac{\partial}{\partial\vartheta_j}\left(\frac{\partial}{\partial\vartheta_i}f(\vartheta,\cdot)\right)d\mu$$

schreibt, sieht man, daß die mittels (h2) ausgedrückte Vertauschungseigenschaft äquivalent als

$$E_\mu\left((\nabla\nabla^\top f)(\vartheta,\cdot)\right) \quad\overset{(!!)}{=}\quad 0 \tag{h2$'$}$$

formuliert werden kann. □

1.13 Heuristik II: (ML-Schätzer) Ein Maximum-Likelihood Schätzer für den unbekannten Parameter in einem statistischen Modell $\mathcal{E}$, das die in 1.12 geforderten Eigenschaften besitzt, benutzt bei Vorliegen der Beobachtung $\omega \in \Omega$ eine Maximalstelle der 'Likelihood-Funktion'

$$\Theta \;\ni\; \vartheta \;\longrightarrow\; f(\vartheta,\omega) \;\in (0,\infty)$$

als Schätzwert für den unbekannten Parameter.

1.14 Heuristik III: (Asymptotik von ML-Schätzern) Betrachte zu $\mathcal{E}$ aus 1.12 und 1.13 Produktmodelle

$$\mathcal{E}_n \;=\; (\,\Omega_n,\,\mathcal{A}_n,\,\{P_{n,\vartheta} : \vartheta\in\Theta\}\,) \;=\; \left(\bigtimes_{i=1}^{n}\Omega,\;\bigotimes_{i=1}^{n}\mathcal{A},\;\{P_{n,\vartheta} := \bigotimes_{i=1}^{n}P_\vartheta : \vartheta\in\Theta\}\right)$$

mit Score $M_{n,\vartheta}$ in $\vartheta \in \Theta$ und Information $I_{n,\vartheta} = n\, I_\vartheta$ wie wie in 1.5 a). Bezeichne $\widehat{\vartheta}_n$ den (!!) Maximum-Likelihood-Schätzer im Produktmodell $\mathcal{E}_n$, von dessen Konsistenz für $n \to \infty$ wir ausgehen (!!). Wieder wird Heuristik mit (!!) hervorgehoben. Dann sollten die zweiten Ableitungen der Log-Likelihoodfunktion in der Nähe der Maximalstelle Entwicklungen

$$\xi \;\longrightarrow\; \left(\nabla^\top \log f_n\right)(\xi) \;\approx\; \left(\nabla^\top \log f_n\right)(\widehat{\vartheta}_n) + \left(\xi - \widehat{\vartheta}_n\right)^\top \left(\nabla\nabla^\top \log f_n\right)(\widehat{\vartheta}_n)$$

zulassen, genauer: man wünscht in der Folge der Produktexperimente $\mathcal{E}_n$ unter wahrem ϑ

$$\text{(h3)}\quad \left(\nabla^\top \log f_n\right)(\vartheta) \overset{(!!)}{=} \left[\left(\nabla^\top \log f_n\right)(\widehat{\vartheta}_n) + \left(\vartheta - \widehat{\vartheta}_n\right)^\top \left(\nabla\nabla^\top \log f_n\right)(\widehat{\vartheta}_n)\right]\left[1 + o_{P_{n,\vartheta}}(1)\right] .$$

Dabei steht $o_{P_{n,\vartheta}}(1)$ für Terme, die $(P_{n,\vartheta})_n$-stochastisch für $n \to \infty$ verschwinden; wir schreiben kurz $\left[1 + o_{P_{n,\vartheta}}(1)\right]$ für eine Diagonalmatrix mit Einträgen $1 + o_{P_{n,\vartheta}}(1)$ auf der Diagonalen. Weiter wünscht man, daß zweite Ableitungen der log-Likelihoodratios in den Produktmodellen $\mathcal{E}_n$ nur wenig auf kleine Veränderungen des Parameterwerts reagieren (im klassischen Normalverteilungsmodell mit unbekanntem Mittelwert und gegebener Kovarianzmatrix sind log-Likelihoodratios quadratisch im Parameter und folglich die zweiten Ableitungen parameterfrei):

$$\text{(h4)}\qquad \left(\nabla\nabla^\top \log f_n\right)(\widehat{\vartheta}_n) \overset{(!!)}{=} \left(\nabla\nabla^\top \log f_n\right)(\vartheta)\;\left[1 + o_{P_{n,\vartheta}}(1)\right] .$$

Da $\widehat{\vartheta}_n$ eine Maximalstelle der Likelihoodfunktion ist, muß in der Entwicklung (h3) gelten

$$\left(\nabla^\top \log f_n\right)(\widehat{\vartheta}_n) \;=\; 0\,,$$

so daß (h3)+(h4) zusammen in die Aussage

$$\text{(h5)}\qquad \left(\nabla^\top \log f_n\right)(\vartheta) \overset{(!!)}{=} \left(\vartheta - \widehat{\vartheta}_n\right)^\top \left(\nabla\nabla^\top \log f_n\right)(\vartheta)\;\left[1 + o_{P_{n,\vartheta}}(1)\right]$$

münden. Nun gilt aber im Produktmodell wegen

$$\frac{1}{n}\left(\nabla\nabla^\top \log f_n\right)(\vartheta, (\omega_1, ..., \omega_n)) \;=\; \frac{1}{n}\sum_{i=1}^n \left(\nabla\nabla^\top \log f\right)(\vartheta, \omega_i)$$

für $n \to \infty$ bei Gültigkeit von (h2) das starke Gesetz der großen Zahlen, also unter wahrem ϑ

$$(\diamond)\qquad \frac{1}{n}\left(\nabla\nabla^\top \log f_n\right)(\vartheta, \cdot) \;=\; E_\vartheta\left(\nabla\nabla^\top \log f(\vartheta, \cdot)\right) \;+\; o_{P_{n,\vartheta}}(1) \;=\; -I_\vartheta \;+\; o_{P_{n,\vartheta}}(1)\,.$$

Hier schreiben wir $o_{P_{n,\vartheta}}(1)$ für einen Vektor, dessen Einträge $P_{n,\vartheta}$-stochastisch für $n \to \infty$ verschwinden. Zugleich mit $(\diamond)$ aber hat man wegen der Struktur des Score in Produktmodellen

$$\left(\nabla^\top \log f_n\right)(\vartheta, (\omega_1, ..., \omega_n)) \;=\; M_{n,\vartheta}^\top(\omega_1, ..., \omega_n) \;=\; \sum_{i=1}^n M_\vartheta^\top(\omega_i)$$

nach Zentralem Grenzwertsatz schwache Konvergenz für $n \to \infty$

$$(\diamond\diamond) \qquad \mathcal{L}\left(\frac{1}{\sqrt{n}} M_{n,\vartheta} \mid P_{n,\vartheta}\right) \longrightarrow \mathcal{N}(0, I_\vartheta) \quad \text{(schwach in } \mathbb{R}^d) \,.$$

Wegen ($\diamond$) und ($\diamond\diamond$) wird aus (h5)

$$\frac{1}{\sqrt{n}} \left(\nabla^\top \log f_n\right)(\vartheta) \overset{(!!)}{=} \sqrt{n}\left(\vartheta - \widehat{\vartheta}_n\right)^\top \frac{1}{n}\left(\nabla\nabla^\top \log f_n\right)(\vartheta) \left[1 + o_{P_{n,\vartheta}}(1)\right]$$

die Aussage

$$\frac{1}{\sqrt{n}} M_{n,\vartheta}^\top \overset{(!!)}{=} \left(\sqrt{n}(\widehat{\vartheta}_n - \vartheta)\right)^\top I_\vartheta \left[1 + o_{P_{n,\vartheta}}(1)\right] \,.$$

Wegen ($\diamond\diamond$) sind die Terme auf der linken Seite dieser Gleichung straff in $\mathbb{R}^d$ für $n \to \infty$ unter ϑ. Bei Invertierbarkeit der Fisher-Information ist dann auch der reskalierte ML-Schätzfehler selbst straff in $\mathbb{R}^d$ für $n \to \infty$. Damit nimmt die letzte Zeile die Gestalt

$$\frac{1}{\sqrt{n}} M_{n,\vartheta}^\top \overset{(!!)}{=} \left(\sqrt{n}(\widehat{\vartheta}_n - \vartheta)\right)^\top I_\vartheta + o_{P_{n,\vartheta}}(1)$$

und nach Multiplikation mit der Inversen der Fisher-Information des Einzelversuchs die Gestalt

$$(\text{h6}) \qquad \left(\sqrt{n}(\widehat{\vartheta}_n - \vartheta)\right) \overset{(!!)}{=} I_\vartheta^{-1} \cdot \frac{1}{\sqrt{n}} M_{n,\vartheta} + o_{P_{n,\vartheta}}(1)$$

an. Hat man aber für eine Schätzfolge eine stochastische Entwicklung (h6) der reskalierten Schätzfehler an der Stelle ϑ, so liefern ($\diamond$) und ($\diamond\diamond$) asymptotische Normalität

$$(\text{h7}) \qquad \mathcal{L}\left(\sqrt{n}(\widehat{\vartheta}_n - \vartheta) \mid P_{n,\vartheta}\right) \longrightarrow \mathcal{N}\left(0, I_\vartheta^{-1}\right) \quad \text{(schwach in } \mathbb{R}^d)$$

für $n \to \infty$. In (h7) erscheint die Inverse der Fisher-Information im Einzelexperiment $\mathcal{E}$ als Kovarianzmatrix der Grenzverteilung. Beachte, daß die stochastische Entwicklung (h6) der ML–Schätzfehler äquivalent in Form

$$(\text{h8}) \qquad \widehat{\vartheta}_n - \vartheta \overset{(!!)}{=} I_{n,\vartheta}^{-1} M_{n,\vartheta} + o_{P_{n,\vartheta}}\left(\frac{1}{\sqrt{n}}\right)$$

geschrieben werden kann. □

Wir begleiten die heuristischen Überlegungen in 1.12–1.14 durch zwei Beispiele.

1.15 Beispiel: Betrachte n-fache Versuchswiederholung

$$\mathcal{E}_n := \left(\Omega_n := \bigtimes_{i=1}^n \mathbb{R}^k,\ \mathcal{A}_n := \bigotimes_{i=1}^n \mathcal{B}(\mathbb{R}^k),\ \{P_{n,\vartheta} := \bigotimes_{i=1}^n \mathcal{N}(\vartheta, \Lambda) : \vartheta \in \Theta\}\right), \quad \Theta := \mathbb{R}^k$$

im Normalverteilungsmodell mit bekannter $k\times k$-Kovarianzmatrix Λ, symmetrisch und strikt positiv definit. Die Dimension der Einzelbeobachtung ist $k \geq 1$. Zu schätzen ist der unbekannte Mittelwert $\vartheta \in I\!R^k$.

a) Die Likelihoodfunktion bei Vorliegen der Einzelbeobachtungen $\omega_1, ..., \omega_n$ aus $I\!R^k$

$$\Theta \ni \vartheta \quad \longrightarrow \quad f_n(\vartheta, (\omega_1, ..., \omega_n)) \;=\; \left(\frac{1}{(2\pi)^{\frac{k}{2}} \, (\det \Lambda)^{\frac{1}{2}}} \right)^n \exp \left(-\frac{1}{2} \sum_{i=1}^n (\omega_i - \vartheta)^\top \Lambda^{-1} \, (\omega_i - \vartheta) \right)$$

hat als eindeutige Maximalstelle den empirischen Mittelwert

$$\widehat{\vartheta}_n(\omega_1, ..., \omega_n) \;:=\; \frac{1}{n} \sum_{i=1}^n \omega_i$$

für jedes $(\omega_1, ..., \omega_n) \in \Omega_n$. Mit Schreibweise $(X_1, \ldots, X_n) := id \,|_{\Omega_n}$ hat der Maximum-Likelihood-Schätzer für den unbekannten Parameter im Modell $\mathcal{E}_n$ die Form $\widehat{\vartheta}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

b) Die Folge $(\widehat{\vartheta}_n)_n$ ist $\sqrt{n}$-konsistent, dabei gilt sogar

$$(\circ) \qquad \mathcal{L}\left(\sqrt{n}(\widehat{\vartheta}_n - \vartheta) \mid P_{n,\vartheta} \right) \;=\; \mathcal{L}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \vartheta) \mid P_{n,\vartheta} \right) \;=\; \mathcal{N}(\, 0_k \,, \, \Lambda \,)$$

für jedes feste n. Score und Information in $\mathcal{E}_n$ sind gegeben durch

$$M_{n,\vartheta} = \Lambda^{-1} \sum_{i=1}^n (X_i - \vartheta) \;, \quad I_{n,\vartheta} \;=\; n \, \Lambda^{-1} \;.$$

Insbesondere ist die Fisher-Information in diesem Experiment unabhängig vom Parameter, und die Kovarianzmatrix des reskalierten ML-Schätzfehlers in $(\circ)$ stimmt für jedes n überein mit der Inversen der Fisher-Information im Einzelversuch. Auch sieht man sofort die (h8) in 1.14 entsprechende Darstellung der reskalierten ML-Schätzfehler

$$(\circ\circ) \qquad \widehat{\vartheta}_n - \vartheta \;=\; I_{n,\vartheta}^{-1} \, M_{n,\vartheta}$$

sogar ohne die in (h8) vorgesehenen $o_{P_{n,\vartheta}}(\frac{1}{\sqrt{n}})$–Restterme. □

Im Normalverteilungsbeispiel 1.15 ist die log-Likelihood-Funktion

$$\vartheta \quad \longrightarrow \quad -\frac{1}{2} \sum_{i=1}^n (X_i(\omega) - \vartheta)^\top \Lambda^{-1} \, (X_i(\omega) - \vartheta) \;\; + \;\; \text{nicht von } \vartheta \text{ abhängende Terme}$$

bei gegebener Beobachtung $\omega \in \Omega_n$ ein nach unten sich öffnendes quadratisches Polynom in $\vartheta \in \Theta$, zentriert um den Maximum-Likelihood-Schätzer $\widehat{\vartheta}_n(\omega)$ für den unbekannten Parameter (damit wegen $(\circ\circ)$ zentriert in der Nähe des wahren ϑ). Lokal in der Nähe der Maximalstelle

der Likelihoodfunktion trifft man in vielen statistischen Modellen auf ähnlich klare Bilder.

Dies muß jedoch keineswegs immer so sein, und LeCam (1990) sammelte mit Lust an der Provokation ML-Fehlschläge. Harmlos auf den ersten Blick wirkt das folgende Beispiel.

1.16 Beispiel: (Neyman und Scott $\approx$ 1950) Wir betrachten ein Lokations- und Skalenmodell mit Normalverteilungen, in dem ein Maximum-Likelihood-Schätzer nicht einmal konsistent ist. Definiere auf $(\Omega, \mathcal{A}) := (I\!R^{2k}, \mathcal{B}(I\!R^{2k}))$ ein Experiment $\{P_\vartheta : \vartheta \in \Theta\}$ durch

$$\Theta := (0,\infty) \times I\!R^k \ , \quad \vartheta =: (\sigma^2, \xi_1, \xi_2, ..., \xi_k) \ , \quad P_\vartheta := \mathcal{N}(\, \phi(\vartheta), \sigma^2 I_{2k}\,)$$

wobei $\phi(\vartheta)$ den Vektor $(\xi_1, \xi_1, \xi_2, \xi_2, ..., \xi_k, \xi_k) \in I\!R^{2k}$ bezeichnet. Für die kanonische Variable auf $(\Omega, \mathcal{A})$ schreiben wir $(X_1, Y_1, X_2, Y_2, ..., X_k, Y_k) := id\,|_\Omega$.
1) Für $i = 1, .., k$ sind die Zufallsvariablen $\frac{X_i - Y_i}{\sqrt{2}}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$ verteilt. Also ist

$$T \ := \ \frac{1}{k} \sum_{i=1}^{k} \left(\frac{X_i - Y_i}{\sqrt{2}} \right)^2$$

ein vernünftiger Schätzer für $\gamma(\vartheta) = \sigma^2$. Wegen $\mathcal{L}(Z^2) = \Gamma(\frac{1}{2}, \frac{1}{2})$ für $Z \sim \mathcal{N}(0,1)$ hat man

$$\mathcal{L}\,(T \mid P_\vartheta) = \Gamma(\, \frac{k}{2}, \, \frac{k}{2\sigma^2}\,) \ , \quad E_\vartheta(T) = \sigma^2 \ , \quad Var_\vartheta(T) = \frac{2\sigma^4}{k}$$

nach den üblichen Skalierungs- und Faltungseigenschaften von Gammaverteilungen.
2) Betrachte nun die Likelihoodfunktion

$$\vartheta \ \longrightarrow \ \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^{2k} \prod_{i=1}^{k} \exp\left(-\frac{1}{2\sigma^2} \left\{ (X_i - \xi_i)^2 + (Y_i - \xi_i)^2 \right\} \right) \ .$$

Einen Maximum-Likelihood-Schätzer $\widehat{\vartheta}$ für ϑ mit Komponenten $\widehat{\sigma^2}$, $\widehat{\xi}_1, ..., \widehat{\xi}_k$ bestimmt man so: für festes $i = 1, ..., k$ besitzt die Abbildung

$$\xi_i \ \longrightarrow \ (X_i - \xi_i)^2 + (Y_i - \xi_i)^2$$

ein eindeutiges Minimum an der Stelle

$$\widehat{\xi}_i \ := \ \frac{X_i + Y_i}{2} \ .$$

In der Log-Likelihoodfunktion bleibt also zu maximieren

$$\sigma^2 \quad \longrightarrow \quad -k \log(\sigma^2) - \sum_{i=1}^{k} \frac{1}{2\sigma^2} \left\{ 2 \left(\frac{X_i - Y_i}{2} \right)^2 \right\}$$

was wieder zu einer eindeutigen Maximalstelle führt:

$$\widehat{\sigma^2} \;:=\; \frac{1}{k}\sum_{i=1}^{k}\left(\frac{X_i - Y_i}{2}\right)^2 \;=\; \frac{1}{2}\,T\,.$$

Benutzt man also die erste Komponente $\widehat{\sigma^2}$ des ML-Schätzers $\widehat{\vartheta}$ zum Schätzen von σ^2, so ist die Verteilung von $\widehat{\sigma^2}$ auf Umgebungen von $\frac{1}{2}\sigma^2$ konzentriert, wegen

$$\mathcal{L}\left(\frac{1}{2}\,T \mid P_\vartheta\right) = \Gamma(\,\frac{k}{2},\,\frac{k}{\sigma^2}\,)\;,\quad E_\vartheta(T) = \frac{\sigma^2}{2}\;,\quad Var_\vartheta(T) = \frac{\sigma^4}{2k}$$

und dies um so schärfer, je größer k ist. Diese Schätzung geht also systematisch fehl. □

Die Heuristik aus 1.12–1.14 kann also auf keinen Fall für sich den Status einer universal richtig liegenden Intuition beanspruchen. In einer großen Zahl statistischer Modelle kann man sich jedoch mit Erfolg an ihr orientieren, um – unter angemessenen Voraussetzungen, und mit rigorosen Beweisen – das asymptotische Verhalten von ML-Schätzern zu beschreiben.

D. Konsistenz von Maximum-Likelihood-Schätzfolgen

Eine Likelihood-Funktion $\vartheta \to f(\vartheta,\omega)$ muß nicht notwendig für alle $\omega \in \Omega$ Maximalstellen besitzen, und wenn es Maximalstellen gibt, müssen diese nicht eindeutig sein. Dies erklärt die im folgenden gegebene 'vorsichtige' Definition eines Maximum-Likelihood-Schätzers und einer Maximum-Likelihood-Schätzfolge.

1.17 Definition: a) Betrachte ein statistisches Modell

$$\mathcal{E} := (\Omega, \mathcal{A}, \{P_\vartheta : \vartheta \in \Theta\})\;,\quad \Theta \subset I\!R^d$$

mit Dichten $f_\vartheta = \frac{dP_\vartheta}{d\mu}$ bezüglich eines dominierenden Maßes μ. Sei $T : (\Omega, \mathcal{A}) \to (I\!R^d, \mathcal{B}(I\!R^d))$ ein Schätzer für den unbekannten Parameter.
i) Sei $A \in \mathcal{A}$ ein Ereignis. Nenne T kurz *ML auf A* falls gilt

$$\text{für jedes } \omega \in A\text{: } T(\omega) \in \Theta\,,\;\; f(T(\omega),\omega) \;=\; \sup\{f(\xi,\omega) : \xi \in \Theta\}\;.$$

ii) T heißt Maximum-Likelihood-Schätzer für den unbekannten Parameter, falls gilt:

$$\text{es gibt ein } A \in \mathcal{A} \text{ mit } \; P_\vartheta(A) = 1 \;\text{ für alle } \vartheta \in \Theta\text{, und } T \text{ ist ML auf } A\;.$$

b) Betrachte wie in 1.9 eine Folge von Experimenten

$$\mathcal{E}_n \;=\; (\Omega_n, \mathcal{A}_n, \{P_{n,\vartheta} : \vartheta \in \Theta\}) \;, \quad n \geq 1 \;,$$

die durch dasselbe $\Theta \in I\!R^d$ parametrisiert sind. Eine Schätzfolge $(T_n)_n$ für den unbekannten Parameter heißt Maximum-Likelihood-Schätzfolge falls eine Folge 'guter Mengen' $A_n \in \mathcal{A}_n$ existiert so daß

$$\text{für } n \geq 1 \text{ fest : } T_n \text{ ist ML auf } A_n \text{ ; für } \vartheta \in \Theta \text{ fest : es gilt } \lim_{n\to\infty} P_{n,\vartheta}(A_n) \;=\; 1 \;\; .$$

Eine häufige Bezeichnung für Maximum-Likelihood-Schätzfolgen ist $(\widehat{\vartheta}_n)_n$. In hinreichend einfachen parametrischen Modellen gibt es explizite Darstellungen für ML-Schätzer. Im allgemeinen ist dies nicht der Fall, und ein ML-Schätzer kann als Maximalstelle der Likelihoodfunktion bei festem $\omega \in \Omega$ nur numerisch bestimmt werden.

In 'guten' Produktmodellen kann man auf verschiedene Weise aus Fishers Programm einen rigorosen Beweisgang machen. Einen Satz sehr starker hinreichender Voraussetzungen (vom Typ gleichgradige Integrierbarkeit der in 1.12–1-14 wichtigen Terme) findet man z.B. in Witting und Müller-Funk (1995, Kap. 6.1.3), in Anlehnung an einen von Wald um 1945 gegebenen Beweis. Schwächere Voraussetzungen findet man in Pfanzagl (1994); wir verdeutlichen den Typ der hier benutzten Voraussetzungen am Beispiel der Konsistenz von ML-Schätzfolgen.

1.18 Beispiel: (siehe Pfanzagl 1994, Kapitel 6.5) Betrachte ein Experiment

$$\mathcal{E} \;:=\; (\,\Omega,\, \mathcal{A},\, \{P_\vartheta : \vartheta \in \Theta\}\,) \;, \quad \Theta \subset I\!R^d \text{ offen}$$

mit Score $\{M_\vartheta : \vartheta \in \Theta\}$ und Fisher-Information $\{I_\vartheta : \vartheta \in \Theta\}$ wie in 1.2, und Produktmodelle

$$\mathcal{E}_n \;=\; (\,\Omega_n,\, \mathcal{A}_n,\, \{P_{n,\vartheta} : \vartheta \in \Theta\}\,) \;=\; \left(\, \bigtimes_{i=1}^{n} \Omega,\; \bigotimes_{i=1}^{n} \mathcal{A},\; \{P_{n,\vartheta} := \bigotimes_{i=1}^{n} P_\vartheta : \vartheta \in \Theta\} \right)$$

wie in 1.5 a). Dabei seien die Dichten $f(\vartheta, \omega)$, $\vartheta \in \Theta$, $\omega \in \Omega$, strikt positiv.
Für jedes $\vartheta \in \Theta$ und jedes $\varepsilon > 0$ gebe es endlich viele offene Teilmengen $V_1, \ldots, V_l$ von Θ (mit l, $V_1, \ldots, V_l$ abhängig von ϑ und ε) so daß folgende zwei Aussagen gelten:

$$(*) \qquad \vartheta \;\notin\; \bigcup_{i=1}^{l} V_i \;\; , \quad \{\xi \in \Theta : |\xi - \vartheta| > \varepsilon\} \;\subset\; \bigcup_{i=1}^{l} V_i \;,$$

$$(**) \qquad \left(\sup_{\xi \in V_i} \log \frac{f(\xi,\cdot)}{f(\vartheta,\cdot)} \right) \in L^1(P_\vartheta) \quad \text{und} \quad E_\vartheta \left(\sup_{\xi \in V_i} \log \frac{f(\xi,\cdot)}{f(\vartheta,\cdot)} \right) < 0 \quad \text{für alle } j = 1, ..., l.$$

Dann ist jede ML-Schätzfolge $(\widehat{\vartheta}_n)_n$ für den unbekannten Parameter konsistent.

Fixiere zum Beweis eine ML–Schätzfolge $(T_n)_n$ für den unbekannten Parameter, sei $(A_n)_n$ eine zugehörige Folge 'guter Mengen'. Betrachte ein festes $\vartheta \in \Theta$. Sei $\varepsilon > 0$ beliebig klein, wähle zu ϑ und ε offene Mengen $V_1, ..., V_l$ gemäß $(*)$ und $(**)$. Da T_n ML auf A_n ist, gilt nach Definition für festes $n \geq 1$

$$\bigcap_{j=1}^{l} \left\{ \sup_{\xi \in V_j} f_n(\xi,\cdot) < \max_{\xi \in \Theta : |\xi - \vartheta| \leq \varepsilon} f_n(\xi,\cdot) \right\} \bigcap A_n \quad \subset \quad \{ \, |T_n - \vartheta| \leq \varepsilon \, \}$$

und also

$$\begin{aligned} \{ \, |T_n - \vartheta| > \varepsilon \, \} \quad &\subset \quad \bigcup_{j=1}^{l} \left\{ \sup_{\xi \in V_j} f_n(\xi,\cdot) \geq \max_{\xi : |\xi - \vartheta| \leq \varepsilon} f_n(\xi,\cdot) \right\} \bigcup A_n^c \\ &\subset \quad \bigcup_{j=1}^{l} \left\{ \sup_{\xi \in V_j} f_n(\xi,\cdot) \geq f_n(\vartheta,\cdot) \right\} \bigcup A_n^c \end{aligned}$$

Dabei gilt $\lim_{n\to\infty} P_{n,\vartheta}(A_n^c) = 0$ nach Definition der 'guten Mengen'. In $\mathcal{E}_n$ schätzt man

$$P_{n,\vartheta} \left(\sup_{\xi \in V_j} f_n(\xi,\cdot) \geq f_n(\vartheta,\cdot) \right)$$

für jedes feste $j = 1, ..., l$ nach oben ab durch

$$P_{n,\vartheta} \left(\left\{ (\omega_1, ..., \omega_n) : \frac{1}{n} \sum_{i=1}^{n} \left(\sup_{\xi \in V_j} \log \frac{f(\xi,\omega_i)}{f(\vartheta,\omega_i)} \right) \geq 0 \right\} \right) ,$$

was wegen Voraussetzung $(**)$ nach dem starken Gesetz der großen Zahlen für $n \to \infty$ verschwindet. Also ist die Behauptung $\lim\limits_{n\to\infty} P_{n,\vartheta} \, (\, |T_n - \vartheta| > \varepsilon \,) = 0$ bewiesen. □

Im Beweis zu 1.18 steht hinter der Suche nach Bedingungen für gleichgradige Integrierbarkeit der in Fishers Programm wichtigen log-Likelihood Ratios der Begriff der Kullback-Divergenz

$$K(P_\vartheta, P_\xi) := \int - \log \left(\frac{f_\xi}{f_\vartheta} \right) dP_\vartheta$$

zwischen P_ϑ und $P_\xi \sim P_\vartheta$ (siehe z.B. Tsybakov 2004, Ch. 2.4, Äquivalenz der Wahrscheinlichkeitsmaße ist dabei nicht notwendig); mit Jensen gilt für die konvexe Funktion $-\log(\cdot)$

$$0 = -\log(1) = -\log E_\vartheta \left(\frac{f(\xi,\cdot)}{f(\vartheta,\cdot)} \right) \leq E_\vartheta \left(-\log \frac{f(\xi,\cdot)}{f(\vartheta,\cdot)} \right) \leq +\infty \, .$$

Die Kullback-Divergenz ist keine Metrik auf $\{P_\vartheta : \vartheta \in \Theta\}$. Um die 'Geometrie' eines dominierten Experiments $\{P_\vartheta : \vartheta \in \Theta\}$ (im Unterschied zur Euklidischen Geometrie von Θ) mit einer geeigneten Metrik zu beschreiben, benutzt man den Hellinger-Abstand $H(\cdot,\cdot)$, definiert durch

$$H^2(P_\xi, P_\vartheta) \;:=\; \frac{1}{2}\int \left(f_\xi^{1/2} - f_\vartheta^{1/2}\right)^2 d\mu \quad \in\; [0,1]$$

(Tsybakov 2004, Ch. 2.4, Strasser 1985, Ch. I.2), zusammen mit der Affinität

$$A(P_\xi, P_\vartheta) \;:=\; \int f_\xi^{1/2} f_\vartheta^{1/2}\, d\mu \;=\; 1 - H^2(P_\xi, P_\vartheta)\;.$$

Mit der Metrik $H(\cdot,\cdot)$ auf $\{P_\vartheta : \vartheta \in \Theta\}$) formuliert man einen deutlich allgemeineren Weg zur Konsistenz und $\sqrt{n}$-Konsistenz von ML-Schätzfolgen. Die Sätze 1.20 und 1.25 unten folgen Ibragimov und Has'minskii (1981). Wir listen zunächst die benötigen Voraussetzungen auf.

1.19 Voraussetzungen und Bezeichnungen: $\mathcal{E} = (\Omega, \mathcal{A}, \{P_\vartheta : \vartheta \in \Theta\})$ sei ein dominiertes Experiment mit Dichten $f_\vartheta = \frac{dP_\vartheta}{d\mu}$ und Likelihood Ratios $L^{\xi/\vartheta} = \infty \cdot 1_{\{f_\vartheta = 0\}} + \frac{f_\xi}{f_\vartheta} 1_{\{f_\vartheta > 0\}}$, für $\xi, \vartheta \in \Theta$. $\Theta \subset \mathbb{R}^d$ sei offen, und für alle $\omega \in \Omega$ sei $\Theta \ni \xi \longrightarrow f(\xi,\omega) \in [0,\infty)$ stetig.
Wir schreiben $\mathcal{K}$ für die Klasse aller Kompakta K in $\mathbb{R}^d$ mit $K \subset \Theta$. Mit den Notationen

$$\begin{aligned}
\underline{h}(\xi,\gamma,K) &:= \inf_{\xi' \in K \setminus B_\gamma(\xi)} H^2(P_{\xi'}, P_\xi)\;, \quad K \in \mathcal{K}\;,\; \xi \in K\;,\; \gamma > 0\;,\\
\overline{a}(\xi, K^c) &:= \int \sup_{\xi' \in K^c} f_{\xi'}^{1/2} f_\xi^{1/2}\, d\mu\;, \quad K \in \mathcal{K}\;,\; \xi \in \mathrm{int}(K)\;,\\
\varrho^2(\xi,\delta) &:= \int \sup_{\xi' \in B_\delta(\xi)} \left|f_{\xi'}^{1/2} - f_\xi^{1/2}\right|^2 d\mu\;, \quad \xi \in \Theta\;,\; \delta > 0
\end{aligned}$$

setzen wir die Gültigkeit der folgenden Bedingungen i) und ii) voraus:
i) *Hellinger-Stetigkeitsbedingung:* für jedes $\xi \in \Theta$ gilt

$$\lim_{\delta \downarrow 0} \varrho(\xi,\delta) \;=\; 0\;;$$

ii) *Identifizierungsbedingung:* für jedes $\xi \in \Theta$ und eine (damit: jede) kompakte Ausschöpfung $(K_m)_m$ von Θ gilt

$$\lim_{m\to\infty} \overline{a}(\xi, K_m^c) \;<\; 1\;.$$

1.20 Satz: Unter den Voraussetzungen aus 1.19 ist jede ML-Schätzfolge konsistent. Dabei verschwinden für jedes $\vartheta \in \Theta$ und jedes $\gamma > 0$ die Wahrscheinlichkeiten $P_{n,\vartheta}\left(|\widehat{\vartheta}_n - \vartheta| > \gamma\right)$ für

$n \to \infty$ sogar exponentiell schnell.

Der Beweis von 1.20 wird – nach einer Serie von Hilfssätzen – in 2.24 gegeben werden.

1.21 Hilfssatz: Die Hellinger-Stetigkeitsbedingung aus 1.19 impliziert

$$\underline{h}(\xi, \gamma, K) \quad > \quad 0$$

für beliebige Wahl von $K \in \mathcal{K},\ \xi \in K$ und $\gamma > 0$ (*Identifizierbarkeit auf Kompakta*).

Beweis: Die Hellinger-Stetigkeitsbedingung

$$\lim_{\delta \downarrow 0} \varrho(\xi_0, \delta) \ = \ 0$$

an jeder Stelle $\xi_0 \in \Theta$ impliziert zunächst

$$H(P_{\xi'}, P_{\xi_0}) \ \longrightarrow \ 0 \quad \text{für} \ \ \xi' \ \to \ \xi_0 \, .$$

Die umgekehrte Dreiecksungleichung $|d(x,z) - d(y,z)| \leq d(x,y)$ angewandt mit $d(\cdot,\cdot) = H(\cdot,\cdot)$ und $x = P_{\xi'}$, $y = P_{\xi_0}$, $z = P_\xi$ erlaubt, daraus auf Stetigkeit der Abbildung

$$(\circ) \qquad \Theta \ \ni \ \xi' \ \longrightarrow \ H^2(P_{\xi'}, P_\xi) \ \in \ [0,1]$$

bei beliebigem festem $\xi \in \Theta$ zu schließen. Wegen $P_{\xi'} \neq P_\xi$ für $\xi' \neq \xi$ hat $(\circ)$ keine andere Nullstelle als $\xi' = \xi$. Damit ist für festes $K \in \mathcal{K},\ \xi \in K,\ \gamma > 0$ die Abbildung $(\circ)$ stetig und strikt positiv auf dem Kompaktum $K \setminus B_\gamma(\xi)$. Also gilt $\min\limits_{\xi' \in K \setminus B_\gamma(\xi)} H^2(P_{\xi'}, P_\xi) > 0$ für jedes $\gamma > 0$, und damit die Behauptung. □

Wir fixieren nun eine ML-Schätzfolge $(\widehat{\vartheta}_n)_n$ mit 'guten Mengen' $(A_n)_n$. Wir fixieren auch $\vartheta \in \Theta$ als wahren Parameter. Der Nachweis der Konsistenz von $(\widehat{\vartheta}_n)_n$ unter ϑ geht von einem ähnlichen Ansatz wie in 1.18 aus, und wieder soll für gewisse Teilmengen $V \subset \Theta$ mit $\operatorname{dist}(V, \vartheta) > 0$

$$P_{n,\vartheta}\left(\sup_{\xi \in V} f_{n,\xi} \geq f_{n,\vartheta}\right) \ = \ P_{n,\vartheta}\left(\sup_{\xi \in V} \sqrt{L_n^{\xi/\vartheta}} \geq 1\right)$$

nach oben abgeschätzt werden. Dazu möchte man aber nun die Markov-Ungleichung

$$\leq \ E_{n,\vartheta}\left(\sup_{\xi \in V} \sqrt{L_n^{\xi/\vartheta}}\right)$$

benutzen. Die folgenden Hilfsätze formulieren Bedingungen für exponentiell schnelles Abklingen rechter Seiten in derartigen Ungleichungen.

1.22 Hilfssatz: Fixiere $K \in \mathcal{K},\ \vartheta \in K,\ \gamma > 0$, und betrachte Punkte $\xi_0 \in K \setminus B_\gamma(\xi_0)$.

a) Wählt man zu ξ_0 ein $\delta = \delta(\xi_0) > 0$ klein genug für

$$\varrho(\xi_0, \delta) \ < \ \underline{h}(\vartheta, \gamma, K) \ ,$$

so gilt für alle $n \geq 1$ die exponentielle Abschätzung

$$E_{n,\vartheta}\left(\sup_{\xi \in B_\delta(\xi_0) \cap K} \sqrt{L_n^{\xi/\vartheta}} \right) \ \leq \ e^{-n\,[\,\underline{h}(\vartheta,\gamma,K) - \varrho(\xi_0,\delta)\,]} \ .$$

b) Für alle $n \geq 1$ gilt die Abschätzung

$$E_{n,\vartheta}\left(\sup_{\xi \in K \setminus B_\gamma(\vartheta)} \sqrt{L_n^{\xi/\vartheta}} \right) \ \leq \ C\, e^{-n\,\frac{1}{2}\,\underline{h}(\vartheta,\gamma,K)}$$

mit einer geeigneten (von ϑ, γ, K, nicht aber von $n \geq 1$ abhängenden) Konstante $C < \infty$.

Beweis: 1) Schreibe kurz $V := B_\delta(\xi_0) \cap K$. Zunächst kann man für $(\omega_1, \ldots, \omega_n) \in \{f_{n,\vartheta} > 0\}$

$$\left(\sup_{\xi \in V} \sqrt{L_n^{\xi/\vartheta}} \right)(\omega_1, \ldots, \omega_n) \ = \ \sup_{\xi \in V} \prod_{i=1}^n \frac{f^{1/2}(\xi, \omega_i)}{f^{1/2}(\vartheta, \omega_i)}$$

durch

$$\prod_{i=1}^n f^{-1/2}(\vartheta, \omega_i) \left[f^{1/2}(\xi_0, \omega_i) + \sup_{\xi \in V} \left| f^{1/2}(\xi, \omega_i) - f^{1/2}(\xi_0, \omega_i) \right| \right]$$

abschätzen, was nach Integration $\int_{\Omega_n} (\bigotimes_{i=1}^n \mu)(d\omega_1, \ldots, d\omega_n) \prod_{i=1}^n f(\vartheta, \omega_i) \ldots$ zu

$$E_{n,\vartheta}\left(\sup_{\xi \in V} \sqrt{L_n^{\xi/\vartheta}} \right) \ \leq \ \left\{ \int_\Omega f_\vartheta^{1/2} \left[f_{\xi_0}^{1/2} + \sup_{\xi \in V} \left| f_\xi^{1/2} - f_{\xi_0}^{1/2} \right| \right] d\mu \right\}^n$$

führt. Nach Definition der Affinität und nach 1.19 hat man aber

$$\int f_\vartheta^{1/2} f_{\xi_0}^{1/2}\, d\mu \ = \ 1 - H^2(P_{\xi_0}, P_\vartheta) \ \leq \ 1 - \underline{h}(\vartheta, \gamma, K)$$

sowie mit Cauchy-Schwartz wegen $V = B_\delta(\xi_0) \cap K$

$$\int f_\vartheta^{1/2} \sup_{\xi \in V} \left| f_\xi^{1/2} - f_{\xi_0}^{1/2} \right| d\mu \ \leq \ 1 \cdot \varrho(\xi_0, \delta) \ = \ \varrho(\xi_0, \delta) \ .$$

Zusammen ergibt sich also mit $1 + x \leq e^x$ für $x \in I\!R$

$$E_{n,\vartheta}\left(\sup_{\xi \in B_\delta(\xi_0) \cap K} \sqrt{L_n^{\xi/\vartheta}} \right) \ \leq \ \{\, 1 - \underline{h}(\vartheta, \gamma, K) + \varrho(\xi_0, \delta) \,\}^n \ \leq \ e^{-n\,[\,\underline{h}(\vartheta,\gamma,K) - \varrho(\xi_0,\delta)\,]}$$

wobei $\underline{h}(\vartheta,\gamma,K) - \varrho(\xi_0,\delta) > 0$ nach Voraussetzung. Das zeigt a).

2) Wähle zu jedem Punkt ξ_0 in $K \setminus B_\gamma(\vartheta)$ einen Radius $\eta(\xi_0) > 0$ mit der Eigenschaft

$$\varrho(\xi_0, \eta(\xi_0)) \;<\; \frac{1}{2}\,\underline{h}(\vartheta,\gamma,K)\;,$$

Die Gesamtheit aller offenen Kugeln $\big\{B_{\eta(\xi_0)}(\xi_0) : \xi_0 \in K \setminus B_\gamma(\vartheta)\big\}$ bildet eine offene Überdeckung von $K \setminus B_\gamma(\vartheta)$. Da $K \setminus B_\gamma(\vartheta)$ kompakt in $\mathbb{R}^d$, kann man eine endliche Teilüberdeckung

$$B_{\eta(\xi_{0,i})}(\xi_{0,i}) \;:\; i = 1,\ldots,\ell = \ell(\vartheta,\gamma,K)$$

auswählen. Wendet man Schritt 1) auf alle $V_i := B_{\eta(\xi_{0,i})}(\xi_{0,i}) \cap K$ an, so ergibt sich

$$E_{n,\vartheta}\left(\sup_{\xi\,\in\,K\setminus B_\gamma(\vartheta)} \sqrt{L_n^{\xi/\vartheta}}\right) \;\le\; \sum_{i=1}^{\ell} E_{n,\vartheta}\left(\sup_{\xi\,\in\,V_i} \sqrt{L_n^{\xi/\vartheta}}\right) \;\le\; \ell\; e^{-n\,\frac{1}{2}\,\underline{h}(\vartheta,\gamma,K)}\;.$$

Das ist b), wobei die Konstante C durch die Zahl $\ell = \ell(\vartheta,\gamma,K)$ der zur Überdeckung von $K \setminus B_\gamma(\vartheta)$ benötigten Kugeln gegeben ist. □

1.23 Hilfssatz: Für $K \in \mathcal{K}$ und $\vartheta \in \mathrm{int}(K)$ sei die Bedingung

$$\overline{a}(\vartheta, K^c) \;<\; 1$$

erfüllt. Dann gilt für alle $n \ge 1$ die exponentielle Ungleichung

$$E_{n,\vartheta}\left(\sup_{\xi\,\in\,\Theta\cap K^c} \sqrt{L_n^{\xi/\vartheta}}\right) \;\le\; [\overline{a}(\vartheta, K^c)]^n\;.$$

Beweis: Schreibe kurz $V := K^c \cap \Theta$, und wie im Beweis von 1.22 für $(\omega_1,\ldots,\omega_n) \in \{f_{n,\vartheta} > 0\}$

$$\left(\sup_{\xi\,\in\,V} \sqrt{L_n^{\xi/\vartheta}}\right)(\omega_1,\ldots,\omega_n) \;=\; \sup_{\xi\,\in\,V} \prod_{i=1}^{n} \frac{f^{1/2}(\xi,\omega_i)}{f^{1/2}(\vartheta,\omega_i)} \;\le\; \prod_{i=1}^{n} f^{-1/2}(\vartheta,\omega_i)\left[\sup_{\xi\,\in\,V} f^{1/2}(\xi,\omega_i)\right]\;.$$

Daraus wird wie dort

$$E_{n,\vartheta}\left(\sup_{\xi\,\in\,V} \sqrt{L_n^{\xi/\vartheta}}\right) \;\le\; \left[\int \sup_{\xi\,\in\,V} f_\vartheta^{1/2} f_\xi^{1/2}\, d\mu\right]^n \;\le\; [\overline{a}(\vartheta, K^c)]^n$$

für beliebiges $n \ge 1$. □

1.24 Beweis von Satz 1.20: Für Teilmengen $U \subset \Theta$ mit $\mathrm{dist}(U,\vartheta) > 0$ liefert die Definition einer ML-Schätzfolge

$$A_n \bigcap \left\{\sup_{\xi\in U} f_{n,\xi} < f_{n,\vartheta}\right\} \;\subset\; \left\{\widehat{\vartheta}_n \notin U\right\}$$

und daraus

$$\left\{\widehat{\vartheta}_n \in U\right\} \;\subset\; \left\{\sup_{\xi\in U} f_{n,\xi} \geq f_{n,\vartheta}\right\} \bigcup A_n^c \,.$$

Für die 'guten Mengen' A_n gilt nach Definition einer ML-Schätzfolge

$$\lim_{n\to\infty} P_{n,\vartheta}\,(A_n^c) \;=\; 0\,;$$

Folglich bleibt von $P_{n,\vartheta}\left(\widehat{\vartheta}_n \in U\right)$ nur noch mit Markov-Ungleichung

$$P_{n,\vartheta}\left(\sup_{\xi\in U}\sqrt{\frac{f_{n,\xi}}{f_{n,\vartheta}}} \geq 1\right) \;\leq\; P_{n,\vartheta}\left(\sup_{\xi\in U}\sqrt{L_n^{\xi/\vartheta}} \geq 1\right) \;\leq\; E_{n,\vartheta}\left(\sup_{\xi\in U}\sqrt{L_n^{\xi/\vartheta}}\right)$$

zu betrachten. Setze nun $U := \Theta \setminus B_\gamma(\vartheta)$ für ein beliebig kleines $\gamma > 0$. Wähle eine kompakte Ausschöpfung $(K_m)_m$ für Θ

$$K_m \text{ kompakt in } I\!R^d\,,\; K_m \subset \mathrm{int}(K_{m+1})\,,\; \Theta \;=\; \bigcup_m K_m\,.$$

Für hinreichend großes m gilt dann $\vartheta \in \mathrm{int}(K_m)$ und $\overline{a}(\vartheta, K_m^c) < 1$. Für solche m zerlegt man

$$U \;=\; (K_m \setminus B_\gamma(\vartheta)) \bigcup (\Theta \cap K_m^c)$$

und hat entsprechend nach 1.22 b) und 1.23

$$\begin{aligned} P_{n,\vartheta}\left(|\widehat{\vartheta}_n - \vartheta| > \gamma\right) \;&\leq\; E_{n,\vartheta}\left(\sup_{\xi\,\in\,\Theta\cap K_m^c}\sqrt{L_n^{\xi/\vartheta}}\right) \;+\; E_{n,\vartheta}\left(\sup_{\xi\,\in\,K_m\setminus B_\gamma(\vartheta)}\sqrt{L_n^{\xi/\vartheta}}\right) \\ &\leq\; [\overline{a}(\vartheta, K_m^c)]^n \;+\; C\; e^{-n\,\frac{1}{2}\,\underline{h}(\vartheta,\gamma,K_m)} \end{aligned}$$

für alle $n \geq 1$. Auf der rechten Seite dieser Ungleichung fallen aber beide Terme exponentiell schnell für $n \to \infty$. Damit ist Satz 1.20 bewiesen. □

Um $\sqrt{n}$–Konsistenz von ML–Schätzfolgen zu zeigen, wendet man wieder dasselbe Beweisschema wie in 1.22 an, aber jetzt 'lokal' in Einschränkung auf kleine Kugeln mit Radius $n^{-1/2}$ um den 'wahren' Parameter ϑ (vgl. Ibragimov und Has'minskii 1981, S. 42 ff, S. 51 ff). Auf solchen Kugeln nutzt man zusätzlich zu 1.19 eine weitere Voraussetzung, die nahe an eine Differenzierbarkeitsbedingung für $\xi \to f_\xi^{1/2}$ in einem $L^2(\mu)$-Sinn herankommt. Darüber wird in Kapitel IV unten mehr gesagt werden.

1.25 Satz: Es gelte 1.19. Relativ zu jedem $\vartheta \in \Theta$ setzen wir weiter voraus: zum Fußpunkt ϑ existiere ein Radius $\gamma = \gamma(\vartheta) > 0$ (mit $\overline{B_\gamma(\vartheta)} \subset \Theta$) und eine Konstante $c(\vartheta, \gamma) > 0$ so daß die folgenden Aussagen i) und ii) gelten:

i) $$\overline{\lambda}^2(\vartheta, \gamma) \;:=\; \int \sup_{\xi : |\xi - \vartheta| \le \gamma} \frac{|f_\xi^{1/2} - f_\vartheta^{1/2}|^2}{|\xi - \vartheta|^2}\, d\mu \quad < \quad \infty \,;$$

ii) $$H^2(P_{\vartheta+h}, P_\vartheta) \;\ge\; c(\vartheta, \gamma)\, |h|^2 \quad \text{für alle } |h| \le \gamma \,.$$

Dann ist jede ML–Schätzfolge für den unbekannten Parameter $\sqrt{n}$–konsistent.

Beweis: Fixiere $\vartheta \in \Theta$. Wähle $\gamma = \gamma(\vartheta) > 0$ so, daß die in 1.25 gemachte Zusatzvoraussetzung erfüllt ist. Nach 1.20 bleibt zum Nachweis der $\sqrt{n}$–Konsistenz von $(\widehat{\vartheta}_n)_n$

$$\text{zu jedem } \varepsilon > 0 \text{ existiert } N < \infty \text{ so daß} \quad \limsup_{n\to\infty}\; P_{n,\vartheta}\Big(\; |\sqrt{n}\,(\widehat{\vartheta}_n - \vartheta)| > N \;\Big) \quad < \quad \varepsilon$$

nur noch zu zeigen: zu jedem $\varepsilon > 0$ existiert ein $N < \infty$ so daß

(∗) $$\limsup_{n\to\infty}\; P_{n,\vartheta}\Big(\; \widehat{\vartheta}_n \;\in\; B_\gamma(\vartheta) \setminus B_{n^{-1/2}N}(\vartheta) \;\Big) \quad < \quad \varepsilon \,.$$

1) Wir betrachten in Θ Kugelschalen mit Maßstab $n^{-1/2}$ um das Zentrum ϑ

$$R_{n,k} \;:=\; \Big\{\xi \in \Theta : \xi = \vartheta + n^{-1/2}h \text{ mit } k-1 < |h| \le k\Big\}$$

für $k \ge 1$. Zum Überdecken von $\overline{B_\gamma(\vartheta)} \setminus \{\vartheta\}$ braucht man die Schalen

$$R_{n,k}\,, \quad 1 \le k \le L_n(\vartheta, \gamma)\,, \quad L_n(\vartheta, \gamma) = O(n^{1/2})\,.$$

2) Für jeden in $\overline{R_{n,k}}$ ausgewählten Punkt $\xi_{n,0}$ gilt analog zu Beweisschritt 1) von 1.22

$$\int f_{\xi_{n,0}}^{1/2}\, f_\vartheta^{1/2}\, d\mu \;=\; 1 - H^2(P_{\xi_{n,0}}, P_\vartheta) \;\le\; 1 - c(\vartheta, \gamma)\,\frac{(k-1)^2}{n}$$

gemäß Voraussetzung ii) aus 1.25.

3) Im $\mathbb{R}^d$ braucht man eine zu $\frac{k^d}{\delta^d}$ proportionale Zahl von Kugeln mit (kleinem) Radius $\delta > 0$ zum Überdecken von $\{h \in \mathbb{R}^d : |h| \le k\}$. Daraus folgt, daß $O\left(\frac{k^{d-1}}{\delta^d}\right)$ δ–Kugeln zur Überdeckung von $\{h \in \mathbb{R}^d : k-1 \le |h| \le k\}$ in $\mathbb{R}^d$ ausreichen.

4) Betrachtet man in der durch $n^{-1/2}$ skalierten Kugelschale $R_{n,k}$ mit Zentrum ϑ Punkte der Form $\xi = \vartheta + n^{-1/2}h$, und läßt man den 'lokalen Parameter' $h \in \mathbb{R}^d$ eine Kugel mit Radius

$\delta > 0$ und Zentrum h_0 durchlaufen, so zeigt Voraussetzung i) aus 1.25

$$\int \sup_{\substack{\xi \in R_{n,k}\ ,\ \xi_0 \in \overline{R_{n,k}} \\ \xi = \vartheta + n^{-1/2} h \\ \xi_0 = \vartheta + n^{-1/2} h_0 \\ h \in B_\delta(h_0)}} |f_\xi^{1/2} - f_{\xi_0}^{1/2}|^2 \; d\mu \quad \leq \quad \delta^2 \, \overline{\lambda}^2(\vartheta, \gamma) \; \frac{1}{n} \, .$$

5) Man wählt nun für jedes n und jede der Kugelschalen $R_{n,k},\ 1 \leq k \leq L_n(\vartheta, \gamma)$, eine Überdeckung

$$V_{n,k,i} \ , \ 1 \leq i \leq \ell(n, k, \vartheta, \delta)$$

von $\overline{R_{n,k}}$ durch Kugeln mit Zentrum $\xi_{n,k,i,0} \in \overline{R_{n,k}}$ und Radius $n^{-1/2}\delta$. Schreibt man die Punkte $\xi \in R_{n,k}$ in Gestalt $\xi = \vartheta + n^{-1/2}h$, entspricht $V_{n,k,i}$ wie in Schritt 4) einer δ-Kugel im lokalen Parameter h um ein geeignetes Zentrum $h_{n,k,i,0}$. Nach Schritt 3) gilt daher für die Anzahl der zur Überdeckung von $\overline{R_{n,k}}$ benötigten $n^{-1/2}\delta$–Kugeln

$$\ell(n, k, \vartheta, \delta) \; \leq \; M \; \frac{k^{d-1}}{\delta^d}$$

mit einer geeigneten Konstante M, welche von der Dimension d des Parameterraumes, aber nicht von n und nicht von $\delta > 0$ abhängt.

6) Nun können wir diese Abschätzungen ähnlich wie in 1.24 und wie im Beweis von 1.22 zusammensetzen. Analog zu 1.24 startet man, indem man

$$(+) \qquad P_{n,\vartheta} \left(\sup_{\substack{\xi \in B_\gamma(\vartheta) \\ |\xi - \vartheta| > n^{-1/2} N}} \sqrt{L_n^{\xi/\vartheta}} \; \geq \; 1 \right)$$

mit einer beliebigen natürlichen Zahl N erst abschätzt zu

$$E_{n,\vartheta} \left(\sup_{\substack{\xi \in B_\gamma(\vartheta) \\ |\xi - \vartheta| > n^{-1/2} N}} \sqrt{L_n^{\xi/\vartheta}} \right) \, ,$$

und dann weiter mit den Überdeckungen aus 1) und 5) zu

$$\sum_{k=N}^{L_n(\vartheta,\gamma)} E_{n,\vartheta} \left(\sup_{\xi \in R_{n,k}} \sqrt{L_n^{\xi/\vartheta}} \right) \quad \leq \quad \sum_{k=N}^{L_n(\vartheta,\gamma)} \sum_{i=1}^{\ell(n,k,\vartheta,\delta)} E_{n,\vartheta} \left(\sup_{\xi \in V_{n,k,i}} \sqrt{L_n^{\xi/\vartheta}} \right) \, .$$

Den einzelnen Term auf der rechten Seite

$$E_{n,\vartheta} \left(\sup_{\xi \in V_{n,k,i}} \sqrt{L_n^{\xi/\vartheta}} \right)$$

schätzt man völlig analog zum Beweis von 1.22 wieder in der Form

$$\left[\int f_\vartheta^{1/2} \, f_{\xi_{n,k,i,0}}^{1/2} \, d\mu \; + \; \int f_\vartheta^{1/2} \sup_{\xi \in V_{n,k,i}} |f_\xi^{1/2} - f_{\xi_{n,k,i,0}}^{1/2}| \, d\mu \right]^n$$

nach oben ab, was nach 2), nach Cauchy-Schwartz kombiniert mit 4), sowie mit $1 + x \leq e^x$ zu

$$\left[\, 1 - c(\vartheta,\gamma)\,\frac{(k-1)^2}{n} \;+\; \delta\,\overline{\lambda}(\vartheta,\gamma)\,\frac{1}{\sqrt{n}}\,\right]^n \quad\leq\quad e^{-\,c_1(k-1)^2\,+\,c_2\,\delta\sqrt{n}}$$

führt, mit geeigneten Konstanten c_1, c_2. Mit der unter 5) gegebenen Schranke für $\ell(n,k,\vartheta,\delta)$ setzen wir alles zusammen und erhalten als obere Schranke für $(+)$

$$(++) \qquad \sum_{k=N}^{L_n(\vartheta,\gamma)\,\ell(n,k,\vartheta,\delta)} \sum_{i=1} E_{n,\vartheta}\left(\sup_{\xi\in V_{n,k,i}} \sqrt{L_n^{\xi/\vartheta}}\right) \quad\leq\quad \widetilde{M}\,\sum_{k=N}^{\infty}\frac{k^{d-1}}{\delta^d}\;e^{-\,c_1(k-1)^2\,+\,c_2\,\delta\sqrt{n}}\;.$$

7) Bis jetzt war $\delta > 0$ noch nicht festgelegt worden. Wir wählen nun $\delta = \delta_n > 0$ in Abhängigkeit von n so daß gilt

$$(+++) \qquad\qquad \delta_n^d \;=\; e^{\,c_2\,\delta_n\,\sqrt{n}} \quad \text{für alle } n \geq 1\;:$$

dies ist möglich, denn die Funktion $\frac{x}{\log x}$, definiert auf $0 < x < 1$, besitzt die Eigenschaft $\frac{x}{\log x} \to 0$ für $x \downarrow 0$, und die Aussage $(+++)$ ist äquivalent zu $\frac{d}{c_2\sqrt{n}} = \frac{\delta_n}{\log \delta_n}$. Mit Wahl $(+++)$ von $\delta = \delta_n > 0$ bleibt auf der rechten Seite von $(++)$ nur

$$\widetilde{M}\,\sum_{k=N}^{\infty} k^{d-1}\;e^{-\,c_1(k-1)^2}$$

zu betrachten. Hier summiert man aber über die Glieder einer konvergenten Reihe, folglich kann der letzte Ausdruck durch große Wahl von $N = N(\varepsilon) < \infty$ unter jede zu Beginn des Beweises gewählte Schwelle $\varepsilon > 0$ gedrückt werden. Damit ist die Aussage $(*)$ und damit die Behauptung des Satzes 1.25 bewiesen. □