Vorlesung über Mathematische Statistik
Sommersemester 2004

Wolfgang Wefelmeyer

Inhaltsverzeichnis

1 Exponentielle Familien

Unter den parametrischen Verteilungsfamilien sind exponentielle Familien im wesentlichen die einzigen, für die nichtasymptotisch optimale Schätzer existieren. Für nicht-exponentielle Familien und für nichtparametrische und semiparametrische Modelle lassen sich nur Schätzer finden, die asymptotisch, also mit gegen unendlich wachsendem Stichprobenumfang optimal sind. (Das geht auch nur dann, wenn solche Modelle in einem geeigneten Sinne gegen exponentielle Familien konvergieren.)

Gegeben sei ein meßbarer Raum $(\Omega, \mathcal{F})$ und eine Familie $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, von Wahrscheinlichkeitsmaßen.

Definition. Eine Familie P_ϑ, $\vartheta \in \Theta$, heißt *exponentielle Familie* in $\eta(\vartheta)$ und T, wenn sie bezüglich eines dominierenden Maßes $\mu|\mathcal{F}$ Dichten folgender Form hat:

$$f_\vartheta(x) = c(\vartheta) \exp(\eta(\vartheta)^\top T(x)),$$

wobei sowohl $1, \eta_1, \ldots, \eta_d$ als auch $1, T_1, \ldots, T_n$ linear unabhängig sind. Sie heißt *kanonisch*, wenn $\eta(\vartheta) = \vartheta$ gilt. Dann besteht der *natürliche Parameter-Raum* (das heißt: der maximale Parameter-Raum) aus den ϑ mit

$$\int \exp(\vartheta^\top T(x))\mu(dx) < \infty.$$

Damit sich die Dichten zu 1 integrieren, muß gelten:

$$c(\vartheta) = \Big(\int \exp(\eta(\vartheta)^\top T(x))\mu(dx) \Big)^{-1}.$$

Nützlich ist die Struktur einer exponentiellen Familie insbesondere, wenn die Verteilung von T einfacher als P_ϑ ist; zum Beispiel, wenn T eine kleinere Dimension als x hat.

Gelegentlich wird die affine Unabhängigkeit der η_j und der T_j nicht in die Definition genommen. Dann läßt sich im allgemeinen die Dimension reduzieren. Trotzdem ist die Darstellung nicht eindeutig.

Satz 1 *Sei P_ϑ, $\vartheta \in \Theta$, eine exponentielle Familie in ϑ und T, und sei Θ offen. Sei h eine Funktion, für die $H(\xi) = \int h \exp(\xi^\top T) d\mu$ für $\xi = (\xi_1, \ldots, \xi_d)$ mit $\xi_j = \vartheta_j + i\tau_j$ und $\vartheta \in \Theta$ existiert und endlich ist. Dann ist H analytisch in jedem ξ_j, und die Ableitungen können unter das Integral gezogen werden.*

Beweis. Schreibe $h\exp(\xi^\top T) = \exp(\xi_1 T_1)h\exp(\sum_{j=2}^d \xi_j T_j)$. Zerlege das Produkt aus den letzten beiden Faktoren in Real- und Imaginärteil und dann jeweils in Positiv- und Negativteil und schlage diese Funktionen zu μ. Dann läßt sich H schreiben als $H(\xi) = \int \exp(\xi_1 T_1) d(\mu_1 - \mu_2 + i(\mu_3 - \mu_4))$. Es reicht also, ein Integral der Form $H(\xi) = \int \exp(\xi T) d\mu$ zu betrachten. Schreibe den Differenzenquotienten als

$$\frac{H(\zeta) - H(\xi)}{\zeta - \xi} = \int \frac{e^{\zeta T} - e^{\xi T}}{\zeta - \xi} d\mu.$$

Für $|z| \le \delta$ gilt

$$\Big|\frac{e^{az} - 1}{z}\Big| = \Big|\sum_{m=1}^{\infty} \frac{z^{m-1} a^m}{m!}\Big| \le \frac{e^{\delta|a|}}{\delta}.$$

Für $|\zeta - \xi| \le \delta$ läßt sich also der Integrand abschätzen durch

$$e^{\xi T}\Big|\frac{e^{(\zeta-\xi)T} - 1}{\zeta - \xi}\Big| \le \frac{1}{\delta} e^{\xi T + \delta|T|} \le \frac{1}{\delta}\big|e^{(\xi+\delta)T} + e^{(\xi-\delta)T}\big|.$$

Die rechte Seite ist integrierbar. Mit dem Satz von der dominierten Konvergenz folgt für $\zeta \to \xi$

$$\frac{H(\zeta) - H(\xi)}{\zeta - \xi} \to \int T e^{\xi T} d\mu.$$

Die höheren Ableitungen erhält man durch Induktion.

2 Suffiziente Statistiken

Gegeben sei ein meßbarer Raum $(\Omega, \mathcal{F})$ und eine Familie $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, von Wahrscheinlichkeitsmaßen. Sei $T : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ eine Statistik. Bezeichne P_ϑ^T die induzierte Verteilung von T unter P_ϑ. Es existiere eine reguläre bedingte Verteilung $P_\vartheta(\cdot|T)$ gegeben T. Dann gilt

$$P_\vartheta A = \int P_\vartheta(A|T = t) P_\vartheta^T(dt), \quad A \in \mathcal{F}.$$

Definition. Die Statistik T heißt suffizient für ϑ, wenn es für alle $A \in \mathcal{F}$ eine von ϑ unabhängige Version der bedingten Wahrscheinlichkeit $P_\vartheta(A|T)$ gibt.

Wir werden in Kapitel 4 sehen, daß dann die besten Schätzer Funktionen von T sein müssen. Das vereinfacht die Suche nach guten Schätzern.

Beispiel. Seien $X_1, \ldots, X_n$ unabhängig und $Bi_{1,p}$-verteilt. Dann ist $T = \sum_{i=1}^n X_i$ suffizient für p.

Setze $X = (X_1, \ldots, X_n)$. Es gilt

$$P(X = x|T = t) = \frac{P(X = x, T = t)}{P(T = t)} = \frac{p^t(1-p)^{n-t}}{\binom{n}{t}p^t(1-p)^{n-t}} = \frac{1}{\binom{n}{t}}.$$

Das ist unabhängig von p.

Satz 2 *Sei $(\Omega, \mathcal{F}) = (\mathbb{R}^k, \mathcal{B}^k)$ und T suffizient für ϑ. Dann existiert eine von ϑ unabhängige reguläre bedingte Verteilung $P(\cdot|T)$.*

Beweis. Für $x \in \mathbb{Q}^k$ existiert eine von ϑ unabhängige Version der bedingten Wahrscheinlichkeit $P(X \leq x|T)$. Dadurch sind die Verteilungsfunktionen $P(X \leq x|T = t)$ für $x \in \mathbb{R}$ festgelegt und von ϑ unabhängig.

Hauptergebnis dieses Kapitels ist das Faktorisierungskriterium von Neyman. Es besagt, daß für dominierte Familien T genau dann suffizient ist, wenn sich die Dichte bis auf einen von ϑ unabhängigen Faktor als Funktion von T schreiben läßt. Wir benötigen zwei Hilfsresultate. Das erste beweisen wir nicht.

Definition. Zwei Familien $\mathcal{M}$ und $\mathcal{N}$ von Maßen auf $(\Omega, \mathcal{F})$ heißen *äquivalent*, wenn $\mu A = 0$ für alle $\mu \in \mathcal{M}$ genau dann gilt, wenn $\mu A = 0$ für alle $\nu \in \mathcal{N}$.

Satz 3 *(Halmos und Savage) Eine Familie von Wahrscheinlichkeitsmaßen ist durch ein σ-endliches Maß dominiert genau dann, wenn sie eine abzählbare äquivalente Teilfamilie besitzt.*

Satz 4 *Sei P_ϑ, $\vartheta \in \Theta$, dominiert durch ein σ-endliches Maß. Sei $Q = \sum_{n=1}^\infty c_n P_{\vartheta_n}$ äquivalent zu P_ϑ, $\vartheta \in \Theta$. Dann ist T suffizient für ϑ genau dann, wenn eine nichtnegative, $\mathcal{E}$-meßbare Funktion g_ϑ existiert mit*

$$dP_\vartheta = g_\vartheta \circ T dQ, \quad \vartheta \in \Theta.$$

Beweis. Sei $\mathcal{F}_0$ die von T induzierte σ-Algebra.

Notwendig. Sei T suffizient für ϑ. Nach Definition des bedingten Erwartungswerts gilt für $A \in \mathcal{F}$ und $A_0 \in \mathcal{F}_0$):

$$\int_{A_0} P(A|T) dP_\vartheta = P_\vartheta A \cup A_0, \quad \vartheta \in \Theta.$$

Weil der Integrand nicht von ϑ abhängt, gilt auch

$$\int_{A_0} P(A|T)dQ = QA \cup A_0,$$

also $P(A|T) = Q(A|T)$. Nach dem Satz von Radon–Nikodỳm und dem Faktorisierungslemma existiert eine $Q|\mathcal{F}_0$-Dichte von $P_\vartheta|\mathcal{F}_0$ der Form $g_\vartheta \circ T$. Es bleibt zu zeigen, daß $g_\vartheta \circ T$ auch $Q|\mathcal{F}$-Dichte von $P_\vartheta|\mathcal{F}$ ist. Für $A \in \mathcal{F}$ gilt

$$\begin{aligned} P_\vartheta A &= \int P(A|T)dP_\vartheta = \int Q(A|T)dP_\vartheta \\ &= \int E_Q(Ag_\vartheta \circ T|T)dQ = \int_A g_\vartheta \circ T dQ. \end{aligned}$$

Hinreichend. Gelte $dP_\vartheta = g_\vartheta \circ T dQ$. Sei $A \in \mathcal{F}$. Wir zeigen: $Q(A|T) = P_\vartheta(A|T)$ Q-f.s. Definiere $d\nu = 1_A dP_\vartheta$. Wegen

$$\frac{d\nu}{dQ} = \frac{d\nu}{dP_\vartheta}\frac{dP_\vartheta}{dQ}$$

gilt einerseits

$$\frac{d\nu|\mathcal{F}_0}{dQ|\mathcal{F}_0} = P_\vartheta(A|T)g_\vartheta \circ T,$$

andererseits

$$\frac{d\nu|\mathcal{F}_0}{dQ|\mathcal{F}_0} = E_Q(Ag_\vartheta \circ T|T) = Q(A|T)g_\vartheta \circ T,$$

also $Q(A|T) = P_\vartheta(A|T)$ $P_\vartheta|\mathcal{F}_0$-f.s. wegen $g_\vartheta \circ T \neq 0$ $P_\vartheta|\mathcal{F}_0$-f.s. Das heißt: T ist suffizient.

Satz 5 *(Faktorisierungskriterium von Neyman) Sei P_ϑ, $\vartheta \in \Theta$, dominiert durch ein σ-endliches Maß μ. Dann ist T suffizient für ϑ genau dann, wenn eine nichtnegative, $\mathcal{E}$-meßbare Funktion g_ϑ und eine nichtnegative, $\mathcal{F}$-meßbare Funktion h existieren mit*

$$dP_\vartheta = hg_\vartheta \circ T dQ, \quad \vartheta \in \Theta.$$

Beweis. Notwendig. Nach Satz 3 existiert ein $Q = \sum c_n P_{\vartheta_n}$ äquivalent zu P_ϑ, $\vartheta \in \Theta$. Wegen

$$\frac{dP_\vartheta}{d\mu} = \frac{dP_\vartheta}{dQ}\frac{dQ}{d\mu}$$

folgt die Behauptung mit $h = dQ/d\mu$ aus Satz 4.

Hinreichend. Es gilt nach Voraussetzung:

$$\frac{dQ}{d\mu} = \sum c_n \frac{dP_{\vartheta_n}}{d\mu} = h \sum c_n g_{\vartheta_n} \circ T,$$

also

$$\frac{dP_\vartheta}{dQ} = \frac{dP_\vartheta}{d\mu}\frac{d\mu}{dQ} = \frac{g_\vartheta \circ T}{\sum c_n g_{\vartheta_n} \circ T}.$$

Aus Satz 4 folgt, daß T suffizient ist.

Korollar 1 *Bei einer exponentiellen Familie in $\eta(\vartheta)$ und T ist T suffizient für ϑ.*

Beweis. Wir haben Dichten der Form $f_\vartheta(x) = c(\vartheta)\exp(\eta(\vartheta)^\top T(x))$, also wie im Satz 5.

Beispiel. Sind $X_1, \ldots, X_n$ unabhängig und $Bi_{1,p}$-verteilt, so ist $\sum_{i=1}^n X_i$ suffizient für p.

Die Zähldichte von $X = (X_1, \ldots, X_n)$ ist

$$P(X = x) = p^{\sum_{i=1}^n x_i}(1-p)n - \sum_{i=1}^n x_i.$$

Das ist von der Form des Faktorisierungskriteriums.

Eine suffiziente Statistik ist nicht eindeutig. Suffizienz bleibt unter allen (meßbaren) eineindeutigen Transformationen erhalten.

3 Vollständige und anzilläre Statistiken

Gegeben sei ein meßbarer Raum $(\Omega, \mathcal{F})$ und eine Familie $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, von Wahrscheinlichkeitsmaßen. Sei $T : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ eine Statistik. Wir sind an möglichst einfachen suffizienten Statistiken interessiert.

Definition. Eine suffiziente Statistik T heißt *minimal suffizient*, wenn für jede suffiziente Statistik S eine meßbare Funktion h existiert, so daß $T = h \circ S P_\vartheta$-f.s. für $\vartheta \in \Theta$.

Ist $(\Omega, \mathcal{F}) = (\mathbb{R}^k, \mathcal{B}^k)$ und ist P_ϑ, $\vartheta \in \Theta$, dominiert, so existiert eine minimale suffiziente Statistik (Bahadur, 1957). Das zeigen wir hier nicht.

Um zu beschreiben, was an einer suffizienten Statistik noch überflüssig ist, verwenden wir folgenden Begriff.

Definition. Eine Statistik T heißt *anzillär*, wenn ihre Verteilung P_ϑ^T nicht von ϑ abhängt. Sie heißt *anzillär erster Ordnung*, wenn ihr Erwartungswert $E_\vartheta T$ nicht von ϑ abhängt.

An der Faktorisierung $P_\vartheta A = \int P_\vartheta(A|T = t)P_\vartheta^T(dt)$ sieht man, daß anzillär komplementär zu suffizient ist. Eine anzilläre Statistik enthält keine Information über ϑ. Es ist plausibel, daß eine suffiziente Statistik minimal ist, wenn keine nichtkonstante Funktion von ihr anzillär ist, oder auch nur anzillär erster Ordnung.

Definition. Eine Statistik T heißt *(beschränkt) vollständig* für ϑ, wenn für jede (beschränkte) meßbare Funktion $g : E \to \mathbb{R}$ gilt:

$$E_\vartheta g(T) = 0 \text{ für } \vartheta \in \Theta \text{ impliziert: } g \text{ ist konstant } P_\vartheta^T\text{-f.s. für } \vartheta \in \Theta.$$

In anderen Worten: Die Familie P_ϑ^T, $\vartheta \in \Theta$, ist so reichhaltig, daß eine nicht identisch verschwindende Funktion zumindest für ein ϑ einen nicht verschwindenden Erwartungswert unter P_ϑ^T hat.

Es gibt sehr große Familien, die nicht vollständig sind. Zum Beispiel verschwinden für alle um Null symmetrischen Verteilungen die Erwartungswerte aller um Null antisymmetrischen Funktionen.

Eine vollständige suffiziente Statistik ist minimal (Lehmann und Scheffé 1950, Bahadur 1957). Das zeigen wir hier nicht.

Beispiel. Seien $X_1, \dots, X_n$ unabhängig und verteilt nach der Gleichverteilung auf $(0, \vartheta)$. Dann ist $T = \max X_i$ vollständig und suffizient für $\vartheta > 0$.

Suffizienz. Die Dichte von $(X_1, \dots, X_n)$ ist

$$\frac{1}{\vartheta^n} \prod_{i=1}^n 1_{(0,\vartheta)}(x_i) = \frac{1}{\vartheta^n} 1_{(0,\vartheta)}(T).$$

Also ist T suffizient nach Satz 5.

Vollständigkeit. Die Verteilungsfunktion von T ist $P(T \le x) = x^n/\vartheta^n$ für $x \in (0, \vartheta)$, die Dichte also nx^{n-1}/ϑ^n. Sei g eine Funktion mit

$$0 = E_\vartheta g(T) = \frac{n}{\vartheta^n} \int_0^\vartheta g(x) x^{n-1} dx, \quad \vartheta > 0.$$

Das Integral läßt sich auffassen als Verteilungsfunktion eines signierten Maßes. Dessen Dichte muß also f.s. verschwinden.

Für das folgende schöne Ergebnis haben wir zunächst keine Verwendung. Es ist aber in der asymptotischen Statistik nützlich.

Satz 6 *(Lemma von Basu) Ist T beschränkt vollständig und suffizient, dann ist jede anzilläre Statistik unabhängig von T.*

Beweis. Sei V anzillär. Setze $a(t) = P_\vartheta(V \in A|T = t)$. Weil T suffizient ist, hängt a nicht von ϑ ab. Weil V anzillär ist, hängt $E_\vartheta a(T) = P_\vartheta(V \in A)$ nicht von ϑ ab. Also gilt $E_\vartheta g(T) = 0$ für $g(t) = a(t) - P(V \in A)$. Weil T beschränkt vollständig ist, gilt also $a = P(V \in A)$ P_ϑ^T-f.s. Das heißt: V und T sind unabhängig unter P_ϑ.

Satz 7 *Sei $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, eine exponentielle Familie in $\eta(\vartheta)$ und T. Ist das Innere von $\eta(\Theta)$ nichtleer, so ist T vollständig für $\vartheta \in \Theta$.*

Beweis. Sei ohne Einschränkung $\eta(\vartheta) = \vartheta$ und $I = (-a, a)^d \subset \Theta$. Sei g eine Funktion mit $E_\vartheta g(T) = 0$ für $\vartheta \in \Theta$. Schreibe

$$E_\vartheta g(T) = \int g(t) \exp(\vartheta^\top t)\mu^T(dt).$$

Es gilt

$$\int g^+(t) \exp(\vartheta^\top t)\mu^T(dt) = \int g^-(t) \exp(\vartheta^\top t)\mu^T(dt). \tag{3.1}$$

Für $\vartheta = 0$ erhält man insbesondere $\int g^+ d\mu^T = \int g^- d\mu^T$. Ohne Einschränkung nehmen wir $\int g^+ d\mu^T = 1$ an. Definiere Wahrscheinlichkeitsmaße $dP^+ = g^+ d\mu^T$ und $dP^- = g^- d\mu^T$. Dann läßt sich (3.1) schreiben als $\int \exp(\vartheta^\top t)P^+(t) = \int \exp(\vartheta^\top t)P^-(t)$. Für $\xi_j = \vartheta_j + i\tau_j$ mit $\vartheta_j \in (-a, a)$ sind die Funktionen $H^+(\xi) = \int \exp(\xi^\top t)P^+(t)$ und $H^-(\xi) = \int \exp(\xi^\top t)P^-(t)$ nach Satz 1 analytisch in jedem ξ_j, stimmen also überein. Für $\xi_j = i\vartheta_j$ gilt also insbesondere $\int \exp(i\vartheta^\top t)P^+(t) = \int \exp(i\vartheta^\top t)P^-(t)$. Nach dem Eindeutigkeitssatz für charakteristische Funktionen gilt deshalb $P^+ = P^-$, also $g^+ = g^-$ μ^T-f.s., also $g = 0$ μ^T-f.s., also P_ϑ^T-f.s. für $\vartheta \in \Theta$.

4 Konvexe Verlustfunktionen

Gegeben sei wieder ein meßbarer Raum $(\Omega, \mathcal{F})$ und eine Familie $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, von Wahrscheinlichkeitsmaßen. Bezeichne X die Identität auf Ω. Sei $s : \Theta \to \mathbb{R}$ eine Funktion. Wir beobachten $X = x$ und wollen den Wert $s(\vartheta)$ schätzen.

Definition. Ein *Schätzer* (für s) ist eine meßbare Abbildung $S : \Omega \to \mathbb{R}$. Eine *Verlustfunktion* ist eine Abbildung $L : \Theta \times \mathbb{R} \to [0, \infty)$ mit $L(\vartheta, \cdot)$ meßbar und $L(\vartheta, s(\vartheta)) = 0$ für $\vartheta \in \Theta$. Die Verlustfunktion heißt *(strikt)*

konvex, wenn $L(\vartheta, \cdot)$ (strikt) konvex für jedes $\vartheta \in \Theta$ ist. Die Verlustfunktion $L(\vartheta, s) = (s - s(\vartheta))^2$ heißt *quadratisch*.

Es ist kein realistisches Ziel, S so zu wählen, daß $L(\vartheta, S(x))$ für alle ϑ und x minimiert wird. Wie versuchen stattdessen, den mittleren Verlust zu minimieren.

Definition. Die *Risikofunktion* von S ist $R(\vartheta, S) = E_\vartheta L(\vartheta, S(X))$.

Satz 8 *(Rao und Blackwell) Sei T suffizient für ϑ, S ein P_ϑ-integrierbarer Schätzer und L eine konvexe Verlustfunktion für s. Dann gilt $R(\vartheta, S^T) \leq R(\vartheta, S)$ für $S^T = E(S|T)$. Ist die Verlustfunktion strikt konvex und $L(\vartheta, S)$ P_ϑ-integrierbar, so ist die Ungleichung strikt, falls nicht $S = S^T$ f.s.*

Beweis. Nach der Jensenschen Ungleichung gilt

$$L(\vartheta, E(S|t)) \leq E(L(\vartheta, S)|t) \quad P_\vartheta^T\text{-f.s.}$$

Die Ungleichung ist strikt, wenn $L(\vartheta, \cdot)$ strikt konvex ist. Die Behauptung folgt durch Integration nach P_ϑ^T.

5 Erwartungstreue Schätzer

Die Minimierung von $R(\vartheta, S)$ ist immer noch kein realistisches Ziel, wenn man keine Einschränkungen an die konkurrierenden Schätzer macht. Zum Beispiel haben die konstanten Schätzer $S(x) = c$ Risiko Null, falls $s(\vartheta) = c$ für das wahre ϑ gilt. Für einen sinnvollen Optimalitätsbegriff muß man solche "voreingenommenen" Schätzer ausschließen.

Definition. Ein Schätzer S ist *erwartungstreu* (für s), wenn $E_\vartheta S = s(\vartheta)$ für jedes $\vartheta \in \Theta$ gilt.

Jede (meßbare) Funktion S ist ein erwartungstreuer Schätzer ihres Erwartungswerts (falls er existiert).

Beispiel. Sei X verteilt nach $Bi_{n,p}$, $p \in (0,1)$, und sei $s(p) = 1/p$. Ein Schätzer S ist erwartungstreu für s, wenn

$$\sum_{k=0}^{n} S(k) \binom{n}{k} p^k (1-p)^{n-k} = \frac{1}{p}, \quad p \in (0,1).$$

Für $p \to 0$ geht die linke Seite gegen $S(0)$, die rechte gegen unendlich. Also existiert kein erwartungstreuer Schätzer.

Es gibt Funktionen s, die sich auf großen Verteilungsfamilien erwartungstreu schätzen lassen.

Beispiel. Seien $X_1, \ldots, X_n$ unabhängig und verteilt nach P aus einer Familie $\mathcal{P}$ auf $\mathcal{F}$. Sei f integrierbar für $P \in \mathcal{P}$. Setze $\vartheta = P$ und $s(P) = E_P f$. Ein erwartungstreuer Schätzer für s ist der *empirische Schätzer*

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Definition. Sei L eine Verlustfunktion für s. Ein erwartungstreuer Schätzer S für s heißt *erwartungstreu mit gleichmäßig kleinstem Risiko* für L, wenn für jeden erwartungstreuen Schätzer S' gilt:

$$R(\vartheta, S) \leq R(\vartheta, S'), \quad \vartheta \in \Theta.$$

Ist die Verlustfunktion die quadratische, so heißt S *UMVU* (*uniformly minimum variance unbiased*).

Definition. Eine Funktion s heißt *erwartungstreu schätzbar*, wenn ein erwartungstreuer Schätzer für s existiert.

Lemma 1 *Sei T vollständig und suffizient für $\vartheta \in \Theta$. Ist s erwartungstreu schätzbar, so existiert (f.s.) genau ein erwartungstreuer Schätzer, der eine Funktion von T ist.*

Beweis. Existenz. Nach Voraussetzung existiert ein erwartungstreuer Schätzer S. Setze $S^T = E_\vartheta(S|T)$.

Eindeutigkeit. Seien $S \circ T$ und $S' \circ T$ zwei erwartungstreue Schätzer, die Funktionen von T sind. Dann gilt $E_\vartheta(S \circ T - S' \circ T = 0$ für $\vartheta \in \Theta$. Wegen der Vollständigkeit folgt $S \circ T = S' \circ T$ P_ϑ-f.s. für $\vartheta \in \Theta$.

Satz 9 *Sei T vollständig und suffizient für $\vartheta \in \Theta$. Sei s erwartungstreu schätzbar.*

a) Es existiert ein erwartungstreuer Schätzer S, der gleichmäßig kleinstes Risiko für jede konvexe Verlustfunktion hat.

b) Er ist der eindeutige erwartungstreue Schätzer, der eine Funktion von T ist.

c) Ist $L(\vartheta, \cdot)$ strikt konvex und $L(\vartheta, S)$ integrierbar, so ist S der eindeutige erwartungstreue Schätzer mit minimalem Risiko.

Beweis. a) Sei S erwartungstreu für s. Der Schätzer $S^T = E_\vartheta(S|T)$ ist ebenfalls erwartungstreu und hat gleichmäßig kleineres Risiko nach Satz 8.

b) Obiges Lemma.

c) Nach Zusatz zu Satz 8 wird das minimale Risiko nur von S^T angenommen.

Wenn wir eine vollständige und suffiziente Statistik T haben, finden wir einen gleichmäßig besten erwartungstreuen Schätzer durch Lösen von $E_\vartheta(S|T) = s(\vartheta)$, $\vartheta \in \Theta$, nach S. Falls wir schon einen erwartungstreuen Schätzer S gefunden haben, brauchen wir nur noch $S^T = E_\vartheta(S|T)$ zu berechnen. Das folgende Beispiel zeigt, daß UMVU-Schätzer ziemlich schlecht sein können.

Beispiel. Sei X verteilt nach P_λ, $\lambda > 0$. Der (eindeutige) UMVU-Schätzer für $\exp(-2\lambda)$ ist

$$S^T = \begin{cases} 1 & X \text{ gerade} \\ -1 & X \text{ ungerade.} \end{cases}$$

P_λ, $\lambda > 0$, ist eine exponentielle Familie in $T = X$. Also ist X vollständig nach Satz 7 und dem Korollar zu Satz 5. Nach Satz 9 (Lehmann– Scheffé) ist der UMVU-Schätzer der (eindeutige) erwartungstreue Schätzer der Form $S \circ X$. Nach der obigen Methode finden wir S, indem wir $E_\lambda S = \exp(-2\lambda)$ nach S lösen. Es gilt

$$E_\lambda S = e^{-\lambda} \sum_{k=0}^{\infty} S(k) \frac{\lambda^k}{k!}, \quad e^{-\lambda} = \sum_{k=0}^{\infty} (-1)^k \frac{\lambda^k}{k!}.$$

Die Behauptung folgt durch Koeffizientenvergleich.

6 Cramér–Rao-Ungleichung

Es gilt $E_\vartheta(S - s(\vartheta))^2 = E_\vartheta(S - E_\vartheta S)^2 + (E_\vartheta S - s(\vartheta))^2$. Für die quadratische Verlustfunktion ist das Risiko also die Summe aus der Varianz des Schätzers und dem Quadrat des *Bias* $E_\vartheta S - s(\vartheta)$. Für erwartungstreue Schätzer verschwindet der Bias, und das Risiko ist gleich der Varianz des Schätzers. In diesem Kapitel geben wir eine untere Schranke dafür an.

Sei $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, eine Familie äquivalenter Wahrscheinlichkeitsmaße und $s : \Theta \in \mathbb{R}$ eine Funktion. Der *Dichtequotient* von P_τ nach P_ϑ ist $L_{\vartheta\tau} = dP_\tau / dP_\vartheta$. Es gilt $E_\vartheta L_{\vartheta\tau} = 1$. Sei $\vartheta \in \Theta$ fest.

Satz 10 *(Hammersley–Chapman–Robbins-Ungleichung) Sei S erwartungstreu für s. Dann gilt (mit $0/0 = 0$)*

$$Var_\vartheta S \geq \sup_{\tau \in \Theta} \frac{(s(\tau) - s(\vartheta))^2}{E_\vartheta (L_{\vartheta\tau} - 1)^2}.$$

Beweis. Mit $E_\vartheta(L_{\vartheta\tau} - 1) = 0$ gilt

$$s(\tau) - s(\vartheta) = E_\tau S - E_\vartheta S = E_\vartheta((L_{\vartheta\tau} - 1)S) = E_\vartheta((L_{\vartheta\tau} - 1)(S - s(\vartheta)).$$

Jetzt die Schwarzsche Ungleichung anwenden.

Im folgenden sei $\Theta \subset \mathbb{R}$, und ϑ liege im Inneren von Θ.

Definition. $L_{\vartheta\tau}$ heißt im *quadratischen Mittel differenzierbar* in $\tau = \vartheta$ mit Ableitung $\dot{\ell}_\vartheta \in L_2(P_\vartheta)$, wenn

$$\big(E_\vartheta(L_{\vartheta\tau} - 1 - (\tau - \vartheta)\dot{\ell}_\vartheta)^2\big)^{1/2} = o(\tau - \vartheta).$$

Wegen $E_\vartheta(L_{\vartheta\tau} - 1) = 0$ gilt $E_\vartheta \dot{\ell}_\vartheta = 0$. Die *Fisher-Information* in ϑ ist $I_\vartheta = E_\vartheta \ell_\vartheta'^2 = Var_\vartheta \dot{\ell}_\vartheta$.

Satz 11 *(Cramér–Rao-Ungleichung) Sei $L_{\vartheta\tau}$ im quadratischen Mittel differenzierbar in $\tau = \vartheta$. Die Fisher-Information in ϑ sei positiv. Sei s differenzierbar in ϑ mit Ableitung $s'(\vartheta)$, und sei S erwartungstreu für s. Dann gilt*

$$Var_\vartheta S \geq \frac{s'(\vartheta)^2}{I_\vartheta}.$$

Beweis. Es gilt nach Voraussetzung

$$E_\vartheta\Big(\frac{L_{\vartheta\tau} - 1}{\tau - \vartheta}\Big)^2 \to E_\vartheta \ell_\vartheta'^2 = 0.$$

Wegen $I_\vartheta > 0$ folgt die Behauptung.

Die Cramér–Rao-Schranke ist offensichtlich i.a. schlechter als die Hammersley–Chapman–Robbins-Schranke. Insbesondere ist sie i.a. nicht scharf.

7 Neyman–Pearson-Lemma

Gegeben sei eine Hypothese $H \subset \Theta$. Die Alternative sei $K = \Theta \setminus H$. Wir beobachten $X = x$ und wollen testen, ob H zutrifft oder nicht.

Definition. Ein *(randomisierter) Test* (für H gegen K) ist eine meßbare Abbildung $\varphi : \Omega \to [0, 1]$.

Wird x beobachtet, so entscheiden wir uns mit Wahrscheinlichkeit $\varphi(x)$ für K. Ist $\varphi = 1_C$, so heißt φ *nichtrandomisiert* und C *kritischer Bereich*.

Definition. Die *Gütefunktion* von φ ist die Abbildung $\vartheta \to E_\vartheta \varphi$.

Die Alternative beschreibt gewöhnlich die riskanteren Entscheidungen. Eine fälschliche Entscheidung dafür heißt *Fehler erster Art*; der umgekehrte Fehler heißt *Fehler zweiter Art*. Für $\vartheta \in H$ ist $E_\vartheta \varphi$ die Wahrscheinlichkeit für den Fehler erster Art; für $\vartheta \in K$ ist $1 - E_\vartheta \varphi$ die für den Fehler zweiter Art. Wir können nicht beide Wahrscheinlichkeiten gleichzeitig minimieren. Deshalb beschränken wir die Wahrscheinlichkeit für den Fehler erster Art und versuchen einen Test zu finden, der unter dieser Nebenbedingung die Wahrscheinlichkeit für den Fehler zweiter Art minimiert, also die Güte $E_\vartheta \varphi$ für $\vartheta \in K$ maximiert.

Definition. Ein Test φ hat das Niveau α (für H), wenn $E_\vartheta \varphi \leq \alpha$ für $\vartheta \in H$.

Definition. Ein Test ψ zum Niveau α (für H) heißt *bester Test* zum Niveau α gegen $\vartheta \in K$, wenn für jeden Test φ zum Niveau α gilt: $E_\vartheta \psi \geq E_\vartheta \varphi$. Der Test ψ heißt *gleichmäßig bester Test* zum Niveau α (gegen K), wenn er optimal gegen alle $\vartheta \in K$ ist.

Das ist im allgemeinen nicht gleichzeitig für alle $\vartheta \in K$ möglich. Zunächst betrachten wir den Fall, daß H und K *einfach*, nämlich einpunktig sind, und schreiben P und Q für Hypothese und Alternative und p und q für ihre Dichten bezüglich eines dominierenden Maßes μ. Dann läßt sich $E_Q \varphi$ maximieren. Das α-*Quantil* einer Verteilung $P|\mathcal{B}$ mit Verteilungsfunktion $F(x) = P(-\infty, x]$ ist

$$F^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\}.$$

Satz 12 *(Neyman–Pearson-Lemma) Sei P ein Wahrscheinlichkeitsmaß und Q ein endliches signiertes Maß.*

a) Sei ψ ein Test mit

$$\psi = \begin{cases} 1 & q > cp \\ 0 & q < cp. \end{cases}$$

Dann gilt für jeden Test φ:

$$E_Q\varphi \leq E_Q\psi + E_P(\varphi - \psi).$$

b) Gilt $E_P\varphi \leq E_P\psi$ und $E_Q\varphi = E_Q\psi$, so gilt

$$\varphi = \begin{cases} 1 & q > cp \\ 0 & q < cp. \end{cases}$$

c) Sei $\alpha \in (0,1)$. Sei c das $(1-\alpha)$-Quantil der Verteilung von q/p unter P, also

$$c = \inf\{y : P(q > yp) \leq \alpha\}.$$

Setze

$$a = \begin{cases} \frac{\alpha - P(q>cp)}{P(q=cp)} & P(q = cp) > 0 \\ 0 & P(q = cp) = 0, \end{cases}$$

und

$$\psi = \begin{cases} 1 & q > cp \\ a & q = cp \\ 0 & q < cp. \end{cases}$$

Dann gilt $E_P\psi = \alpha$.

Beweis. a) Es gilt $(q - cp)(\psi - \varphi) \geq 0$.

b) Nach a) gilt $c = 0$ oder $E_P\varphi = E_P\psi$, also $\int (q - cp)(\psi - \varphi)d\mu = 0$. Da der Integrand nichtnegativ ist, folgt die Behauptung.

c) Es gilt $E_P\psi = P(q > cp) + aP(q = cp) = \alpha$.

Die Randomisierung ist nötig, wenn die Verteilungsfunktion von q/p unter P in c den Wert $1-\alpha$ nicht annimmt. Die Randomisierung ist dieselbe für alle x, für die $q(x) = cp(x)$ gilt.

Ist $\alpha \in (0,1)$ und ψ der *Neyman–Pearson-Test* aus Satz 12c, so gilt $E_Q\varphi \leq E_Q\psi$ für jeden Test φ zum Niveau α. Ist umgekehrt φ ein bester Test zum Niveau α, so ist φ nach Satz 12b $(P+Q)$-f.s. von der Form

$$\varphi = \begin{cases} 1 & q > cp \\ 0 & q < cp. \end{cases}$$

Also unterscheidet sich φ höchstens auf $\{q = cp\}$ von ψ.

8 Monotone Dichtequotienten und gleichmäßig beste Tests

Sei $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, eine Familie von Wahrscheinlichkeitsmaßen. Wir hatten in Kapitel 6 Dichtequotienten für äquivalente Maße eingeführt. Im allgemeinen dominiert P_ϑ nicht P_τ. Für $\mu = P_\vartheta + P_\tau$ setzen wir dann

$$p_\vartheta = \frac{dP_\vartheta}{d\mu}, \quad p_\tau = \frac{dP_\tau}{d\mu}, \quad L_{\vartheta\tau} = \frac{p_\tau}{p_\vartheta}\mathbf{1}(p_\vartheta > 0) + \infty\mathbf{1}(p_\vartheta = 0, p_\tau > 0).$$

Im folgenden nehmen wir $\Theta \subset \mathbb{R}$ an.

Definition. Sei $T : \Omega \in \mathbb{R}$ eine Zufallsvariable. Die Familie P_ϑ, $\vartheta \in \Theta$, hat *monotone Dichtequotienten* in T, wenn für $\vartheta, \tau \in \Theta$ mit $\vartheta < \tau$ eine nichtfallende Funktion $H_{\vartheta\tau}$ existiert mit $L_{\vartheta\tau} = H_{\vartheta\tau} \circ T$ $(P_\vartheta + P_\tau)$-f.s.

Satz 13 *Hat P_ϑ, $\vartheta \in \Theta$, monotone Dichtequotienten in T, und ist $\alpha \in (0,1)$, so existiert ein gleichmäßig bester Test zum Niveau α für $\tau \leq \vartheta$ gegen $\tau > \vartheta$, nämlich*

$$\psi = \begin{cases} 1 & T > b \\ a & T = b \\ 0 & T < b, \end{cases}$$

wobei a und b bestimmt sind durch $E_\vartheta\psi = \alpha$.

Für $\tau < \vartheta$ minimiert $\varphi = \psi$ das Niveau $E_\tau\varphi$ unter allen Tests φ mit $E_\vartheta\varphi = \alpha$.

Beweis. Wie im Beweis von Satz 12b wähle a und b mit $E_\vartheta\psi = \alpha$. Sei $\tau > \vartheta$. Nach Vorausetzung hat ψ die Form

$$\psi = \begin{cases} 1 & L_{\vartheta\tau} > c = H_{\vartheta\tau}(b) \\ 0 & L_{\vartheta\tau} < c = H_{\vartheta\tau}(b). \end{cases}$$

Nach Satz 12a ist ψ bester Test zum Niveau α für ϑ gegen τ. Analog ist $1-\psi$ bester Test zum Niveau $1-\alpha$ für für ϑ gegen $\tau < \vartheta$. Insbesondere gilt für $\tau < \vartheta$:

$$E_\tau(1-\psi) \geq E_\tau(1-\alpha) = 1-\alpha,$$

also $E_\tau\psi \leq \alpha$.

Satz 14 *Sei P_ϑ, $\vartheta \in \Theta$, eine eindimensionale exponentielle Familie in $\eta(\vartheta)$ und T. Ist η nichtfallend (nichtwachsend), so hat die Familie monotone Dichtequotienten in T $(-T)$.*

Beweis. P_ϑ hat μ-Dichte der Form $c(\vartheta)\exp(\eta(\vartheta)T)$. Also gilt

$$L_{\vartheta\tau} = \frac{c(\tau)}{c(\vartheta)} e^{(\eta(\tau)-\eta(\vartheta))T}.$$

Ist η nichtfallend und $\tau > \vartheta$, so gilt $\eta(\tau) - \eta(\vartheta) \geq 0$. Also ist

$$H_{\vartheta\tau}(t) = \frac{c(\tau)}{c(\vartheta)} e^{(\eta(\tau)-\eta(\vartheta))t}$$

nichtfallend in t.

9 Lokal beste Tests

Sei $P_\vartheta|\mathcal{F}$, $\vartheta \in \theta \subset \mathbb{R}$, eine Familie äquivalenter Wahrscheinlichkeitsmaße. Sei ϑ im Inneren von Θ.

Definition. $L_{\vartheta\tau}$ ist *im Mittel differenzierbar* in $\tau = \vartheta$ mit Ableitung $\dot{\ell}_\vartheta \in L_1(P_\vartheta)$, wenn

$$E_\vartheta \left| L_{\vartheta\tau} - 1 - (\tau - \vartheta)\dot{\ell}_\vartheta \right| = o(\tau - \vartheta).$$

Lemma 2 *Sei $L_{\vartheta\tau}$ im Mittel differenzierbar in $\tau = \vartheta$. Dann gilt für jeden Test φ, daß $E_\tau\varphi$ differenzierbar in $\tau = \vartheta$ mit Ableitung $E_\vartheta(\dot{\ell}_\vartheta\varphi)$ ist.*

Beweis. Es gilt

$$\frac{E_\tau\varphi - E_\vartheta\varphi}{\tau - \vartheta} = \frac{E_\vartheta((L_{\vartheta\tau} - 1)\varphi)}{\tau - \vartheta} \to E_\vartheta(\dot{\ell}_\vartheta\varphi).$$

Definition. Ein Test φ heißt *α-ähnlich* für ϑ, wenn $E_\vartheta\varphi = \alpha$. Ein α-ähnlicher Test ψ heißt *lokal bester ähnlicher Test* zum Niveau α gegen $\tau > \vartheta$, wenn für jeden α-ähnlichen Test φ gilt:

$$\partial_{\tau=\vartheta} E_\tau\psi \geq \partial_{\tau=\vartheta} E_\tau\varphi.$$

Satz 15 *Sei $L_{\vartheta\tau}$ im Mittel differenzierbar in $\tau = \vartheta$ mit Ableitung $\dot{\ell}_\vartheta$. Dann existiert ein lokal bester ähnlicher Test zum Niveau α gegen $\tau > \vartheta$, nämlich*

$$\psi = \begin{cases} 1 & \dot{\ell}_\vartheta > c \\ a & \dot{\ell}_\vartheta = c \\ 0 & \dot{\ell}_\vartheta < c, \end{cases}$$

wobei a und c bestimmt sind durch $E_\vartheta\psi = \alpha$.

Beweis. Es ist zu zeigen: $\varphi = \psi$ maximiert $E_\vartheta(\dot{\ell}_\vartheta\varphi)$ unter den α-ähnlichen Tests. Dazu wenden wir Satz 12a (Neyman–Pearson-Lemma) an für $P = P_\vartheta$, $dQ = \dot{\ell}_\vartheta dP_\vartheta$ und $\mu = P_\vartheta$.

10 Konfidenzbereiche

Sei $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta \subset \mathbb{R}$, eine Familie von Wahrscheinlichkeitsmaßen. Bezeichne X die Identität auf Ω und $\mathfrak{P}$ die Menge der Teilmengen von $\mathbb{R}$.

Definition. Ein *Konfidenzbereich* (für ϑ) ist eine Abbildung $B : \Omega \to \mathfrak{P}$ mit $\{x \in \Omega : \vartheta \in B(x)\} \in \mathcal{B}$ für $\vartheta \in \mathbb{R}$.

Wird $X = x$ beobachtet, so nehmen wir an, daß der wahre Parameter in $B(x)$ liegt. Manchmal akzeptieren wir auch zu große (oder zu kleine) Parameter. Das entspricht den Hypothesen $\tau \leq \vartheta$ (oder $\tau \geq \vartheta$) statt $\{\vartheta\}$. Allgemein bezeichne H_ϑ die Menge der akzeptablen Parameter, wenn ϑ wahr ist.

Definition. Ein Konfidenzbereich B hat das *Niveau* $1-\alpha$ (für ϑ und H), wenn

$$P_\tau\{x \in \Omega : \vartheta \in B(x)\} \geq 1-\alpha, \quad \tau \in H_\vartheta, \vartheta \in \Theta.$$

Definition. Ein Konfidenzbereich B^* zum Niveau $1-\alpha$ heißt *gleichmäßig bester Konfidenzbereich zum Niveau* $1-\alpha$, wenn für jeden Konfidenzbereich B zum Niveau $1-\alpha$ gilt:

$$P_\tau(\vartheta \in B^*) \leq P_\tau(\vartheta \in B), \quad \tau \notin H_\vartheta, \vartheta \in \Theta.$$

Satz 16 *a) Hat der Konfidenzbereich B das Niveau $1-\alpha$ für H, so hat für jedes ϑ der kritische Bereich $C_\vartheta = \{x \in \Omega : \vartheta \notin B(x)\}$ das Niveau α für H_ϑ. Ist B gleichmäßig bester Konfidenzbereich, so ist C_ϑ gleichmäßig bester kritischer Bereich für jedes ϑ.*

b) Hat C_ϑ das Niveau α für H_ϑ, so hat der Konfidenzbereich $B(x) = \{\vartheta : x \notin C_\vartheta\}$ das Niveau $1-\alpha$ für H. Ist C_ϑ gleichmäßig bester kritischer Bereich für jedes ϑ, so ist B gleichmäßig bester Konfidenzbereich.

Beweis. Es gilt $P_\tau C_\vartheta = P_\tau(\vartheta \notin B(x))$.

11 M-Schätzer und Maximum-Likelihood-Schätzer

Beispiel. Sei $\mathcal{P}$ eine Familie von Wahrscheinlichkeitsmaßen auf $(\mathbb{R}, \mathcal{B})$ mit endlicher Varianz. Seien $X_1, \ldots, X_n$ unabhängig mit Verteilung $P \in \mathcal{P}$. Ein Schätzer für den Erwartungswert $s(P) = E_P(X) = PX$ ist das Stichprobenmittel $\frac{1}{n}\sum_{i=1}^n X_i$. Sei $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ die *empirische Verteilung.*

Hier bezeichnet δ_x das *Dirac-Maß* in x. Das Stichprobenmittel läßt sich schreiben als $s(\mathbb{P}_n)$. Ist s geeignet stetig, so ist $s(\mathbb{P}_n)$ konsistent; ist s geeignet differenzierbar, so ist $n^{1/2}(s(\mathbb{P}_n) - s(P))$ asymptotisch normal. In diesem Beispiel wissen wir das schon:

$$n^{1/2}(s(\mathbb{P}_n) - s(P)) = n^{-1/2} \sum_{i=1}^{n} (X_i - PX).$$

Der Erwartungswert ist das Minimum von $P(X-s)^2$ in s, denn $P(X-s)^2 = P(X-PX)^2 + (PX-s)^2$. Entsprechend ist das Stichprobenmittel $s(\mathbb{P}_n)$ das Minimum der stochastischen Version $\mathbb{P}(X-s)^2$. Ein solcher Schätzer heißt M-Schätzer. Für $s(P)$ muß gelten:

$$\partial_{s=s(P)} P(X-s)^2 = 0.$$

Ebenso muß für $s(\mathbb{P}_n)$ gelten:

$$\partial_{s=s(\mathbb{P}_n)} \mathbb{P}_n(X-s)^2 = 0.$$

In unserem Beispiel wissen wir das schon:

$$P(X - s(P)) = 0, \quad \mathbb{P}_n(X - s(\mathbb{P}_n)) = 0.$$

Das läßt sich verallgemeinern.

Satz 17 *Sei $P|\mathcal{F}$ ein Wahrscheinlichkeitsmaß, $\Theta \subset \mathbb{R}^d$ und ϑ im Inneren von Θ. Für τ in einer Umgebung von ϑ sei ψ_τ eine d-dimensionale meßbare Funktion und $\psi_\tau(x)$ stetig differenzierbar. Für die Matrix der partiellen Ableitungen gelte $|\dot{\psi}_\vartheta| \leq H$ für ein P-integrierbares H, und $E_P\dot{\psi}_\vartheta$ sei invertierbar. Sei $\hat{\vartheta} = \vartheta + o_P(1)$ eine (konsistente) Lösung der Schätzgleichung $\sum \psi_\tau(X_i) = o_P(n^{1/2})$. Dann gilt*

$$n^{1/2}(\hat{\vartheta} - \vartheta) = -(P\dot{\psi}_\vartheta)^{-1} n^{-1/2} \sum_{i=1}^{n} \psi_\vartheta(X_i) + o_P(1).$$

Beweis. Mit einer Taylorentwicklung erhalten wir

$$\psi_\tau(x) = \psi_\vartheta(x)(\tau - \vartheta) + \int_0^1 (\dot{\psi}_{\vartheta + s(\tau-\vartheta)}(x) - \dot{\psi}_\vartheta(x)) ds (\tau - \vartheta).$$

Sei

$$h_a(x) = \sup_{|\tau - \vartheta| \leq a} |\dot{\psi}_\tau(x) - \dot{\psi}_\vartheta(x)|.$$

Weil $\dot\psi_\tau(x)$ stetig in $\tau=\vartheta$ ist, gilt $h_a \downarrow 0$ punktweise für $a \downarrow 0$. Also gilt mit dem Satz von der dominierten Konvergenz $Ph_a \downarrow 0$ für $a \downarrow 0$, also mit dem Schwachen Gesetz der großen Zahl

$$\limsup_n \frac{1}{n}\sum_{i=1}^n h_{a_n}(X_i) \le \limsup_n \frac{1}{n}\sum_{i=1}^n h_a(X_i) = Ph_a + o_P(1).$$

Daraus folgt $\frac{1}{n}\sum_{i=1}^n h_{a_n}(X_i) = o_P(1)$ für $a \downarrow 0$. Mit der Taylorentwicklung gilt also

$$o_P(1) = \frac{1}{n}\sum_{i=1}^n \psi_{\hat\vartheta}(X_i) = \frac{1}{n}\sum_{i=1}^n \psi_\vartheta(X_i) + \frac{1}{n}\sum_{i=1}^n \dot\psi_\vartheta(X_i) + o_P(1)(\hat\vartheta-\vartheta).$$

Die Behauptung folgt durch Auflösen nach $\hat\vartheta-\vartheta$ und nochmalige Anwendung des Schwachen Gesetzes der großen Zahl.

Da $\frac{1}{n}\sum_{i=1}^n \psi_\vartheta(X_i) = P\psi_\vartheta + o_P(1)$, können wir

$$\frac{1}{n}\sum_{i=1}^n \psi_{\hat\vartheta}(X_i) = o_P(n^{-1/2})$$

nur erwarten, wenn $P\psi_\vartheta = 0$. Unter dieser Bedingung (und recht starken weiteren technischen Bedingungen) läßt sich die Existenz konsistenter Lösungen der Schätzgleichung zeigen; das behandeln wir hier aber nicht. Im folgenden wenden wir Satz 17 auf eine parametrische Familie an.

Die Dichte von $X_1,\dots,X_n$ ist $\prod_{i=1}^n f_\vartheta(X_i)$. Der *Maximum-Likelihood-Schätzer* $\hat\vartheta$ maximiert diese Dichte. Insbesondere gilt $\sum \partial_{\vartheta=\hat\vartheta} \log f_\vartheta(X_i) = 0$. Wir zeigen, daß die asymptotische Kovarianzmatrix dieses Schätzers gleich der Inversen der Fisher-Information ist.

Definition. Sei P_ϑ, $\vartheta\in\Theta\subset\mathbb{R}^d$, eine Familie von Wahrscheinlichkeitsmaßen und ϑ im Inneren von Θ. Die Familie ist *Cramér-regulär* bei ϑ, wenn P_τ für τ in einer Umgebung von ϑ eine positive μ-Dichte f_τ hat und f_τ und $\ell_\tau = \log f_\tau$ zweimal stetig differenzierbar sind mit $|\ddot f_\tau| \le A$ und $|\ddot\ell_\tau| \le H$ für μ-integrierbares A und P_ϑ-integrierbares H, und wenn die *Informationsmatrix* $I_\vartheta = E_\vartheta \dot\ell_\vartheta \dot\ell_\vartheta^\top$ positiv definit ist.

Lemma 3 *Für bei ϑ Cramér-reguläre Familien gilt $E_\tau\dot\ell_\tau = 0$ und $E_\tau\ddot\ell_\tau = -I_\tau$ für τ in einer Umgebung von ϑ.*

Beweis. Wir skizzieren den Beweis nur. Die Argumente sind ähnlich wie in Satz 17. Es gilt $\mu f_\tau = 1$, also

$$0 = \mu(f_{\tau'} - f_\tau) = (\tau' - \tau) \int_0^1 \dot{f}_{\tau + s(\tau' - \tau)} ds.$$

Mit dem Satz von der dominierten Konvergenz konvergiert das Integral für $\tau' \to \tau$ gegen $\dot{f}_\tau$. Also gilt $E_\tau \dot{\ell}_\tau = \mu \dot{f}_\tau = 0$. Mit demselben Argument gilt $\mu \ddot{f}_\tau = 0$. Wegen $\dot{f} = \dot{\ell} f$ gilt $\ddot{f} = \ddot{\ell} f + \dot{\ell} \dot{f}^\top = \ddot{\ell} f + \dot{\ell} \dot{\ell}^\top f$, insbesondere also $E_\tau \ddot{\ell}_\tau + E_\tau \dot{\ell}_\tau \dot{\ell}_\tau^\top = 0$.

Satz 18 *Sei P_ϑ, $\vartheta \in \Theta \subset \mathbb{R}^d$, Cramér-regulär. Sei $\hat{\vartheta} = \vartheta + o_P(1)$ eine Lösung der Schätzgleichung $\sum_{i=1}^n \dot{\ell}_\tau(X_i) = o_P(n^{-1/2})$. Dann gilt*

$$n^{1/2}(\hat{\vartheta} - \vartheta) = I_\vartheta^{-1} n^{-1/2} \sum_{i=1}^n \dot{\ell}_\vartheta(X_i) + o_P(1).$$

Insbesondere ist $\hat{\vartheta}$ asymptotisch normal mit Kovarianz-Matrix I_ϑ^{-1}.

Beweis. Wende Satz 17 für $\psi = \dot{\ell}$ und $P = P_\vartheta$ an. Nach Lemma 3 ist $E_\vartheta \ddot{\ell}_\vartheta = -I_\vartheta$.

12 Empirische Schätzer und lineare Regression

Sei $\mathcal{P}$ eine Familie von Wahrscheinlichkeitsmaßen auf $(\Omega, \mathcal{F})$, und $X_1, \ldots, X_n$ seien unabhängig mit Verteilung P. Für $A \in \mathcal{F}$ gilt nach dem Gesetz der großen Zahl

$$\mathbb{P}_n A = \frac{1}{n} \sum_{i=1}^n 1_A(X_i) \to PA \quad \text{f.s.}$$

Nach dem zentralen Grenzwertsatz gilt

$$n^{1/2}(\mathbb{P}_n A - PA) = n^{-1/2} \sum_{i=1}^n (1_A(X_i) - PA) \Rightarrow N(0, PA(1 - PA)).$$

Wir sagen (etwas mißverständlich): $\mathbb{P}_n A$ ist *asymptotisch normal* mit Varianz $PA(1 - PA)$. Sei $f : \mathbb{R} \to \mathbb{R}^k$ mit $Pf_r^2 = E_P f_r^2 = \int f_r^2 dP < \infty$ für $r = 1, \ldots, k$. Der *empirische Schätzer* für Pf ist

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^n f(X_i).$$

Wieder gilt

$$\mathbb{P}_n f \to Pf \quad \text{f.s.}$$

und

$$n^{1/2}(\mathbb{P}_n f - Pf) = n^{-1/2} \sum_{i=1}^{n} (f(X_i) - Pf) \Rightarrow N(0, \Sigma)$$

mit Kovarianzmatrix $\Sigma = (\sigma_{rs})$ und

$$\sigma_{rs} = P(f_r - Pf_r)(f_s - Pf_s) = Pf_r f_s - Pf_r Pf_s.$$

Bemerkung. Ist $t : \mathbb{R}^k \to \mathbb{R}^m$ stetig differenzierbar mit $m \times k$-Matrix $T = (t_a^b)_{a,b}$ von partiellen Ableitungen, so erhält man durch Taylor-Entwicklung

$$t(\mathbb{P}_n f) = t(Pf) + T(\mathbb{P}_n f - Pf) + o_P(n^{-1/2}).$$

Also ist $t(\mathbb{P}_n f)$ asymptotisch normal mit Kovarianzmatrix $T\Sigma T^\top$.

Beispiel. (*Lineare Einschränkung.*) Sei $h : \Omega \to \mathbb{R}^m$ meßbar mit $Ph_r^2 < \infty$ für $r = 1, \ldots, m$. Für $P \in \mathcal{P}$ gelte die lineare Einschränkung $Ph = 0$. Sei $f : \Omega \to \mathbb{R}^k$ meßbar mit $Pf_s^2 < \infty$ für $s = 1, \ldots, k$. Außer dem empirischen Schätzer $\mathbb{P}_n f$ finden wir dann noch andere erwartungstreue Schätzer für Pf: Für jede $k \times m$-Matrix A ergibt sich der erwartungstreue Schätzer

$$\mathbb{P}_n f - A\mathbb{P}_n h = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Ah(X_i)).$$

Seine asymptotische Kovarianzmatrix is $P(f - Pf - Ah)(f - Pf - Ah)^\top$. Nach der Schwarzschen Ungleichung wird die Kovarianzmatrix minimiert durch

$$A^* = Pfh^\top (Phh^\top)^{-1}.$$

Das hängt vom unbekannten P ab. Wir schätzen A^*, indem wir die Erwartungswerte durch empirische Schätzer ersetzen:

$$\hat{A} = \mathbb{P}_n f h^\top (\mathbb{P}_n h h^\top)^{-1} = \sum_{i=1}^{n} f(X_i) h^\top(X_i) \Big(\sum_{i=1}^{n} h(X_i) h^\top(X_i) \Big)^{-1}.$$

Mit dem Gesetz der großen Zahl gilt $\hat{A} = A + o_P(1)$. Also ist der Schätzer $\mathbb{P}_n f - \hat{A}\mathbb{P}_n h$ asymptotisch äquivalent zu $\mathbb{P}_n f - A^*\mathbb{P}_n h$. Aus dem zentralen Grenzwertsatz und $Ph = 0$ ergibt sich $n^{1/2}\mathbb{P}_n h = O_P(1)$, also

$$\begin{aligned} n^{1/2}(\mathbb{P}_n f - \hat{A}\mathbb{P}_n h - Pf) &= n^{1/2}(\mathbb{P}_n f - A^*\mathbb{P}_n h - Pf) - (\hat{A} - A^*) n^{1/2} \mathbb{P}_n h \\ &= n^{1/2}(\mathbb{P}_n f - A^*\mathbb{P}_n h - Pf) + o_P(1). \end{aligned}$$

Insbesondere hat $\mathbb{P}_n f - \hat{A}\mathbb{P}_n h - Pf$ dieselbe asymptotische Kovarianzmatrix wie $\mathbb{P}_n f - A^*\mathbb{P}_n h - Pf$, nämlich

$$Pff^\top - PfPf^\top - Pfh^\top(Phh^\top)^{-1}Phf^\top.$$

Man kann zeigen, daß $\mathbb{P}_n f - \hat{A}\mathbb{P}_n h - Pf$ asymptotisch nicht zu übertreffen ist, es sei denn, man weiß mehr über $\mathcal{P}$.

Problem. Es gelte die Einschränkung $Ph_\vartheta = 0$ mit unbekanntem d-dimensionalen Parameter ϑ. Gesucht werden asymptotisch optimale Schätzer für ϑ (das ist i.w. bekannt) und für Pf. Zusatz: Was schätzen diese Schätzer, wenn $Ph_\vartheta = 0$ nicht gilt, und schätzen sie es optimal?

Beispiel. (*Lineare Regression.*) Sei X eine k-dimensionale Zufallsvariable und Y eine reelle Zufallsvariable. Wir vermuten einen linearen Zusammenhang und schreiben $Y = \vartheta^\top X + \varepsilon$. Sei P die Verteilung von (X, Y). Es gelte $P|X|^4 < \infty$ und $PY^4 < \infty$, und $PXX^\top$ sei positiv definit. Das *Kleinste-Quadrate-Funktional* $\vartheta(P)$ minimiert $P(Y - \vartheta^\top X)^2$ in ϑ. Durch Differenzieren erhalten wir $PX(Y - \vartheta^\top(P)X) = 0$, also

$$\vartheta(P) = (PXX^\top)^{-1}PXY.$$

Der *Kleinste-Quadrate-Schätzer* für $\vartheta(P)$ ist

$$\vartheta(\mathbb{P}_n) = \Big(\sum_{i=1}^n X_i X_i^\top\Big)^{-1} \sum_{i=1}^n X_i Y_i.$$

Wir erhalten ihn natürlich auch als Lösung der empirischen Version von $PX(Y - \vartheta^\top(P)X) = 0$, also als Lösung der Schätzgleichung $\mathbb{P}_n X(Y - \vartheta^\top X) = 0$. Da sich das nach ϑ auflösen läßt, brauchen wir die Theorie der M-Schätzer nicht zu bemühen. Der Kleinste-Quadrate-Schätzer ist eine Funktion zweier empirischer Schätzer. Wir könnten deshalb die asymptotische Verteilung von $\vartheta(\mathbb{P}_n)$ aus der obigen Bemerkung herleiten. Hier geht es aber einfacher. Schreibe

$$n^{1/2}(\vartheta(\mathbb{P}_n) - \vartheta(P)) = \Big(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\Big)^{-1} n^{-1/2} \sum_{i=1}^n X_i(Y_i - \vartheta(P)^\top X_i).$$

Die zweite Summe ist zentriert. Mit dem Gesetz der großen Zahl erhalten wir

$$n^{1/2}(\vartheta(\mathbb{P}_n) - \vartheta(P)) = (PXX^\top)^{-1} n^{-1/2} \sum_{i=1}^n X_i(Y_i - \vartheta(P)^\top X_i) + o_p(1)$$

Also ist $\vartheta(\mathbb{P})$ asymptotisch normal mit Kovarianzmatrix

$$(PXX^\top)^{-1}P[XX^\top(Y-\vartheta^\top X)^2)](PXX^\top)^{-1}.$$

Bis jetzt haben wir keine Modellannahmen gemacht. Nehmen wir nun an, daß ein linearer Zusammenhang $Y = \vartheta^\top X + \varepsilon$ existiert in dem (sehr schwachen) Sinn, daß $E(\varepsilon|X) = 0$ gilt; anders gesagt, daß $E(Y|X) = \vartheta^\top X$ gilt. Dann gilt $E(XY|X) = (XX^\top)\vartheta$, also $\vartheta = \vartheta(P)$. Läßt sich dann der Kleinste-Quadrate-Schätzer verbessern? Die Modellannahme bedeutet, daß für alle (quadratintegrierbaren) k-dimensionalen Zufallsvektoren $W(X)$ gilt:

$$PW(X)(Y-\vartheta^\top X) = 0.$$

Wir erhalten also M-Schätzer als Lösungen $\hat\vartheta_W$ von $\mathbb{P}_nW(X)(Y-\vartheta^\top X) = 0$, also

$$\hat\vartheta_W = (\mathbb{P}_nW(X)X^\top)^{-1}\mathbb{P}_nW(X)Y = \Big(\sum_{i=1}^n W(X_i)X_i^\top\Big)^{-1}\sum_{i=1}^n W(X_i)Y_i.$$

Solche Schätzer heißen *gewichtete* Kleinste-Quadrate-Schätzer. Wie oben erhält man

$$\hat\vartheta_W - \vartheta = \Big(\sum_{i=1}^n W(X_i)X_i^\top\Big)^{-1}\sum_{i=1}^n W(X_i)(Y_i-\vartheta^\top X_i).$$

Also ist $\hat\vartheta_W$ asymptotisch normal mit Kovarianzmatrix

$$P(W(X)X^\top)^{-1}P(W(X)W(X)^\top\rho^2(X))P(XW(X)^\top)^{-1},$$

wobei $\rho^2(X)$ die bedingte Varianz von ε gegeben X ist. Nach der Schwarzschen Ungleichung wird die Kovarianzmatrix minimiert durch

$$W^*(X) = \rho^{-2}(X)X.$$

Das hängt vom unbekannten P ab. Wir müssen ρ durch einen geeigneten Schätzer $\hat\rho$ ersetzen. Solche Schätzer werden wir erst später kennenlernen. Wir erhalten

$$\hat\vartheta_* = \Big(\sum_{i=1}^n \hat\rho^{-2}(X_i)X_iX_i^\top\Big)^{-1}\sum_{i=1}^n \hat\rho^{-2}(X_i)X_iY_i$$

mit asymptotischer Kovarianzmatrix $(P\rho^{-2}(X)XX^\top)^{-1}$.

Problem. Es gelte die Einschränkung $E(h_\vartheta(X,Y)|X) = 0$ mit unbekanntem d-dimensionalen Parameter ϑ. (Oben hatten wir $h_\vartheta(X,Y) = Y-\vartheta^\top X$.) Gesucht werden asymptotisch optimale Schätzer für ϑ (das ist i.w. bekannt) und für Pf. Zusatz: Was schätzen diese Schätzer, wenn die Einschränkung nicht gilt, und schätzen sie es optimal?

13 Stichprobenquantile

Seien $X_1, \ldots, X_n$ unabhängig mit stetiger Verteilungsfunktion F. Dann heißt $R_j = \sum_{i=1}^n \mathbf{1}(X_i \le X_j)$ der *Rang* von X_j. Die der Größe (dem Rang) nach geordneten Beobachtungen $X_{1:n} \le \cdots \le X_{n:n}$ heißen *Ordnungsstatistiken*. Das *p-Quantil* ist $\xi_p = F^{-1}(p) = \inf\{x : F(x) \ge p\}$. Die *empirische Verteilungsfunktion* ist

$$\mathbb{F}_n(t) = \mathbb{P}_n(-\infty, t] = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i \le t) = \frac{1}{n}|\{i : X_i \le t\}|.$$

Das *Stichproben-p-Quantil* ist $\hat{\xi}_p = \mathbb{F}_n^{-1}(p)$.

Satz 19 *Sei ξ_p eindeutig und $k = np + o(n)$. Dann gilt $X_{k:n} \to \xi_p$ in Wahrscheinlichkeit.*

Beweis. Sei $p_\varepsilon = F(\xi_p + \varepsilon)$. Dann gilt $p_\varepsilon > p$. Schreibe

$$\begin{aligned} P(X_{k:n} \le \xi_p + \varepsilon) &= P\Big(\sum_{i=1}^n \mathbf{1}(X_i \le \xi_p + \varepsilon) \ge k\Big) \\ &= P\Big(\frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i \le \xi_p + \varepsilon) - p_\varepsilon \ge \frac{k}{n} - p_\varepsilon)\Big). \end{aligned}$$

Das konvergiert gegen 1 nach dem Gesetz der großen Zahl. Analog für $\xi_p - \varepsilon$ statt $\xi_p + \varepsilon$.

Satz 20 *Es habe F eine Dichte f, die in ξ_p stetig und positiv ist. Sei $k = np + o(n^{-1/2})$. Dann gilt*

$$n^{1/2}(X_{k:n} - \xi_p) \Rightarrow N(0, p(1-p)/f^2(\xi_p)).$$

Beweis.

$$\begin{aligned} P(n^{1/2}(X_{k:n} - \xi_p) \le t) &= P(X_{k:n} \le \xi_p + n^{-1/2}t) \\ &= P\Big(\sum_{i=1}^n \mathbf{1}(X_i \le \xi_p + n^{-1/2}t) \ge k\Big) \\ &= P\Big(\frac{1}{n}\sum_{i=1}^n Y_{ni} \ge u_n)\Big) \end{aligned}$$

mit

$$Y_{ni} = \mathbf{1}(X_i \le \xi_p + n^{-1/2}t) - F(\xi_p + n^{-1/2}t),$$
$$u_n = n^{-1/2}(k - nF(\xi_p + n^{-1/2}t)).$$

Die Zufallsvariable Y_{ni} nimmt den Wert $1 - F(\xi_p + n^{-1/2}t)$ mit Wahrscheinlichkeit $F(\xi_p + n^{-1/2}t)$ an, und den Wert $-F(\xi_p + n^{-1/2}t)$ mit Wahrscheinlichkeit $1 - F(\xi_p + n^{-1/2}t)$. Es gilt $EY_{ni} = 0$, $EY_{ni}^2 \to p(1-p)$ und $u_n \to -tf(\xi_p)$. Nach einer Version des Zentralen Grenzwertsatzes für Dreiecksschemata gilt also

$$P\Big(n^{-1/2}\sum_{i=1}^{n} Y_{ni} \ge u_n\Big) \to \Phi(-xf(\xi_p)/(p(1-p))^{1/2}).$$

Hier ist Φ die Verteilungsfunktion der Standard-Normalverteilung $N(0,1)$.

14 Punktweise Konvergenz von Kernschätzern

Seien $X_1, \dots, X_n$ unabhängig mit Dichte f. Ein *Kernschätzer* für $f(x)$ ist

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_b(x - X_i)$$

mit $K_b(t) = K(t/b)/b$, *Kern* K und Bandweite b. Der *mittlere quadratische Fehler* (MSE) von $\hat{f}(x)$ ist

$$E(\hat{f}(x) - f(x))^2 = \operatorname{Var}\hat{f}(x) + (E\hat{f}(x) - f(x))^2,$$

die Summe aus Varianz und Quadrat des *Bias*. Für $r = 0, 1, 2, \dots$ sei $\mathrm{Lip}_{r,\alpha}(L)$ die Klasse der Funktionen, die beschränkt und r-mal differenzierbar sind und deren r-te Ableitungen in x *Lipschitz der Ordnung* α mit Konstante L sind:

$$|f^{(r)}(x) - f^{(r)}(y)| \le L|x - y|^{\alpha}.$$

Sei $\mathcal{K}_{r,\alpha}$ die Klasse der Funktionen, die beschränkt sind mit

$$\int K(t)\,dt = 1,$$
$$\int t^j K(t)\,dt = 0, \quad j = 1, \dots, 1 \vee r,$$
$$\int |t^{r+\alpha}K(t)|\,dt < \infty.$$

Satz 21 *Ist $f \in \mathrm{Lip}_{r,\alpha}(L)$ und $K \in \mathcal{K}_{r,\alpha}$, so gilt für $b \downarrow 0$:*

$$E(\hat{f}(x) - f(x))^2 \le \frac{1}{nb} f(x) \int K^2(t)\,dt + b^{2(r+\alpha)} \Big(\frac{L}{r!} \int |t^{r+\alpha} K(t)|\,dt \Big)^2.$$

Also wird die optimale Rate $n^{-2(r+\alpha)/(2(r+\alpha)+1)}$ erzielt, wenn b proportional zu $n^{-1/(2(r+\alpha)+1)}$ ist.

Beweis. Schreibe Varianz und Bias als

$$\begin{aligned}
\operatorname{Var} \hat{f}(x) &= \frac{1}{n} \operatorname{Var} K_b(x - X) \\
&= \frac{1}{n} \big(E K_b^2(x - X) - (E K_b(x - X))^2 \big) \\
&= \frac{1}{n} \Big(\int K_b^2(x - u) f(u)\,du - \Big(\int K_b(x - u) f(u)\,du \Big)^2 \Big); \\
E\hat{f}(x) - f(x) &= \int K_b(x - u)(f(u) - f(x))\,du.
\end{aligned}$$

Für $m = 1, 2$ gilt

$$\begin{aligned}
\int K_b^m(x - u) f(u)\,du &= b^{-m} \int K^m \Big(\frac{x - u}{b} \Big) f(u)\,du \\
&= b^{-m+1} \int K^m(t) f(x - bt)\,dt.
\end{aligned}$$

Also

$$\operatorname{Var} \hat{f}(x) = \frac{1}{nb} \int K^2(t) f(x - bt)\,dt - \frac{1}{n} \Big(\int K(t) f(x - bt)\,dt \Big)^2.$$

Für $m = 1, 2$ gilt mit Beschränktheit von K:

$$\int |K^m(t)||f(x - bt) - f(x)|\,dt \le L b^\alpha \int |K^m(t)||t|^\alpha\,dt,$$

also

$$\operatorname{Var} \hat{f}(x) = \frac{1}{nb} f(x) \int K^2(t)\,dt + \frac{1}{nb} o_b(1).$$

Für $r = 0$ gilt

$$|E\hat{f}(x) - f(x)| = \Big| \int K(t)(f(x - bt) - f(x))\,dt \Big| \le b^\alpha L \int |K(t) t^\alpha|\,dt.$$

Für $r = 1$ gilt

$$\begin{aligned}|E\hat{f}(x) - f(x)| &= \Big| \int K(t)\Big(\sum_{j=1}^{r} \frac{-(bt)^j}{j!} f^{(j)}(x) \\ &\qquad + \frac{-(bt)^r}{r!}(f^{(r)}(z) - f^{(r)}(x))\Big)\, dt\Big| \\ &\le b^{r+\alpha}\frac{L}{r!}\int |K(t)t^{r+\alpha}|\, dt.\end{aligned}$$

Der *integrierte mittlere quadratische Fehler* (MISE) von $\hat{f}$ ist $E\int(\hat{f}(x)-f(x))^2\,dx$. Eine Schranke dafür läßt sich nicht einfach aus aus Satz 21 gewinnen, es sei denn, wir wüßten, daß die Dichte f auf einer beschränkten Menge lebt.

15 Konvergenz von Kernschätzern in L_1

Sei λ das Lebesgue-Maß auf $\mathbb{R}$. Wir setzen $L_1 = L_1(\lambda)$ und $\int f = \int f(x)\,dx$. Die L_1-Norm einer Funktion f ist $\|f\|_1 = \int |f|$. Wir benötigen einige Begriffe und Ergebnisse aus der reellen Analysis. Die *Translation* von f um y ist definiert durch $f_y(x) = f(x-y)$. Für $f \in L_1$ gilt $\|f_y\|_1 = \|f\|_1$.

Lemma 4 *(L_1-Stetigkeit der Translation) Für $f \in L_1$ ist $y \to f_y$ gleichmäßig L_1-stetig.*

Beweis. Sei $\varepsilon > 0$. Wähle g stetig mit Träger in $[-A, A]$ und $\|f-g\|_1 < \varepsilon$. Dann ist g gleichmäßig stetig. Also existiert $\delta > 0$ mit $|g(y)-g(z)| < e/(3A)$ für $|y-z| < \delta$. Für $|y-z| < \delta$ gilt also $\|g_y - g_z\|_1 < (2A+\delta)\varepsilon/(3A) < \varepsilon$, also

$$\begin{aligned}\|f_y - f_z\|_1 &\le \|f_y - g_z\|_1 + \|g_y - g_z\|_1 + \|g_z - f_z\|_1 \\ &= \|f-g\|_1 + \|g_y - g_z\|_1 + \|g-f\|_1 < 3\varepsilon.\end{aligned}$$

Lemma 5 *Für $h \in L_1$ gilt*

$$\iint |g(x-bu) - g(x)||h(u)|\,du\,dx \to 0, \quad b \to 0.$$

Beweis. Es gilt $\|g_y - g\|_1 \le 2\|g\|_1$. Nach Lemma 4 ist $y \to g_y$ L_1-stetig. Nach dem Satz von der monotonen Konvergenz also

$$\iint |g(x-bu) - g(x)|\,dx|h(u)|\,du = \int \|g_{bu} - g\|_1 |h(u)|\,du \to 0, \quad b \to 0.$$

Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ heißt *absolutstetig* auf dem endlichen Intervall $[a, b]$, wenn für alle $\varepsilon > 0$ ein $\delta > 0$ existiert, so daß für alle endlichen Familien von disjunkten Intervallen $(a_j, b_j] \subset (a, b]$ mit $\sum_j (b_j - a_j) < \delta$ gilt: $\sum_j |f(b_j) - f(a_j)| < \varepsilon$. Die Funktion heißt *absolutstetig*, wenn sie auf allen endlichen Intervallen absolutstetig ist.

Für eine Funktion $g : \mathbb{R} \to \mathbb{R}$ ist die *Totalvariation* auf einem Intervall $[a, b]$ definiert als

$$\sup_{(t_j)} \sum_j |f(t_j) - f(t_{j-1})|,$$

wobei sich das Supremum über alle endlichen Partitionen $a = t_0 < \cdots < t_k = b$ erstreckt. Eine auf $[a, b]$ absolutstetige Funktion hat endliche Totalvariation auf $[a, b]$.

Wir benötigen den Fundamentalsatz für Lebesgue-Integrale: Ist f absolutstetig auf $[a, b]$, so existiert f.s. die Ableitung f' und ist in L_1, und

$$f(x) - f(a) = \int_a^x f'(t)\, dt, \quad x \in [a, b].$$

Für $r = 0, 1, \ldots$ sei $\mathcal{L}_{r,0}$ die Klasse der Funktionen f, die $(r-1)$-mal differenzierbar sind mit $f^{(r-1)}$ absolutstetig und $f^{(r)} \in L_1$. Für $0 < \alpha \leq 1$ sei $\mathcal{L}_{r,\alpha}$ die Klasse der Funktionen $f \in \mathcal{L}_{r,0}$, für die $y \to f_y^{(r)}$ L_1-*Lipschitz der Ordnung* α ist: Für ein $L > 0$ gilt

$$\|f_y^{(r)} - f^{(r)}\|_1 \leq L|y|^\alpha, \quad y \in \mathbb{R}.$$

Die *Faltung* $f * g$ zweier Funktionen $f, g \in L_1$ ist definiert durch

$$f * g(x) = \int f(x - u) g(u)\, du.$$

Es gilt

$$\|f * g\|_1 \leq \iint |f(x - u) g(u)|\, du\, dx = \|f\|_1 \|g\|_1.$$

Den Erwartungswert eines Kernschätzers können wir wie folgt schreiben:

$$\begin{aligned} E\hat{f}(x) &= EK_b(x - X) \\ &= \int K_b(x - y) f(y)\, dy = f * K_b(x) = \int f(x - bu) K(u)\, du. \end{aligned}$$

Es gilt $\|K_b\|_1 = \|K\|_1$, also

$$\|E\hat{f}\|_1 = \|f * K_b\|_1 \leq \|f\|_1 \|K_b\|_1 = \|K\|_1.$$

Satz 22 *Ist $f \in \mathcal{L}_{r,0}$ und $K \in \mathcal{K}_{r,0}$, so gilt*

$$\|f * K_b - f\|_1 = o(b^r).$$

Ist $0 < \alpha \leq 1$ und $f \in \mathcal{L}_{r,\alpha}$, $K \in \mathcal{K}_{r,\alpha}$, so gilt

$$\|f * K_b - f\|_1 = O(b^{r+\alpha}).$$

Beweis. Durch Taylor-Entwicklung:

$$f(x-bu) - f(x) = \sum_{j+1}^{r} \frac{(-bu)^j}{j!} f^{(j)}(x) + \frac{(-bu)^r}{r!} \int_0^1 (f^{(r)}(x - tbu) - f(x))\, dt.$$

Mit $\int u^j K(u)\, du = 0$, $j = 1, \ldots, r$, gilt

$$f * K_b(x) - f(x) = \frac{(-b)^r}{r!} \int_0^1 \int (f^{(r)}(x - tbu) - f(x)) u^r K(u)\, du\, dt.$$

Für $\alpha = 0$ wenden wir Lemma 5 und den Satz von der dominierten Konvergenz füe das dt-Integral an. Für $\alpha > 0$ verwenden wir

$$\int |f^{(r)}(x - tbu) - f(x)\, dx \leq Lb^\alpha |u|^\alpha.$$

Satz 23 *Hat f endliches zweites Moment und ist $K \in \mathcal{L}_{1,1}$, so gilt*

$$E\|\hat{f} - f * K_b\|_1 = O((nb)^{-1/2}).$$

Beweis. Sei $V(x) = (1 + |x|)^2$. Mit der Schwarzschen Ungleichung gilt für jedes meßbare g:

$$\|g\|_1^2 = \Big(\int \frac{V^{1/2}|g|}{V^{1/2}} \Big)^2 \leq \|V^{-1}\|_1 \|Vg^2\|_1,$$

und $\|V^{-1}\|_1$ ist endlich. Es gilt also

$$\begin{aligned} E\|\hat{f} - f * K_b\|_1^2 &\leq \|V^{-1}\|_1 E\|V(\hat{f} - f * K_b)^2\|_1 \\ &= \|V^{-1}\|_1 \|EV(\hat{f} - f * K_b)^2\|_1 \leq \frac{1}{n} \|V^{-1}\|_1 \|V \cdot f * K_b^2\|_1. \end{aligned}$$

Es gilt $V(x + y) \leq V(x)W(y)$ für $W(y) = 1 + 2|y|(1 + |y|)$ und

$$f * K_b^2(x) = \frac{1}{b^2} \int K^2\Big(\frac{x-y}{b}\Big) f(y)\, dy = \frac{1}{b} \int f(x - bu) K^2(u)\, du\, dx,$$

also

$$\begin{aligned}\|V \cdot f * K_b^2\|_1 &= \frac{1}{b} \iint V(x+bu) f(x) K^2(u)\, du\, dx \\ &\leq \frac{1}{b} \int V(x) f(x)\, dx \int W(bu) K^2)\, du.\end{aligned}$$

Weil K beschränkt ist, ist das letzte Integral endlich.

Der *erwartete L_1-Fehler* von $\hat{f}$ ist $E \int |\hat{f} - f|$. Es gilt

$$E \int |\hat{f} - f| = E\|\hat{f} - f\|_1 \leq E\|\hat{f} - f * K_b\|_1 + \|f * K_b - f\|_1.$$

Ist $f \in \mathcal{L}_{r,\alpha}$ mit endlichem zweiten Moment und $K \in \mathcal{K}_{r,\alpha}$ mit $\alpha = 1$, falls $r = 1$, so gilt nach den Sätzen 22 und 23:

$$E \int |\hat{f} - f| = O(b^{r+\alpha}) + O((nb)^{-1/2}).$$

Die optimale Bandweite und Konvergenzrate sind also dieselben wie für die Konvergenz der Wurzel des punktweisen mittleren quadratischen Fehlers, Satz 21.

Wir erwarten, daß $(nb)^{1/2}(\hat{f}(x) - f(x))$ asymptotisch normal mit Mittelwert 0 ist, solange b schneller als die optimale Bandweite gegen 0 geht, denn dann ist der Bias vernachlässigbar. Für die optimale Bandweite b wird $(nb)^{1/2}(\hat{f}(x) - f(x))$ unter geeigneten Annahmen immer noch normal sein, aber der Mittelwert der Grenzverteilung wird nicht 0 sein.

16 Plug-in-Schätzer für Faltungsdichten

Setzt man den Dichteschätzer in ein "glattes Funktional" der Dichte ein, so erhält man i.a. eine schnellere Rate; sie kann sogar die "parametrische" Rate $n^{-1/2}$ sein. Das einfachste Funktional ist ein lineares Funktional $Eg(X) = \int g(x) f(x)\, dx$. Der "Plug-in-Schätzer" dafür ist

$$\begin{aligned}\int g(x) \hat{f}(x)\, dx &= \int g(x) \frac{1}{n} \sum_{i=1}^{n} K_b(x - X_i)\, dx \\ &= \frac{1}{n} \sum_{i=1}^{n} \int g(x + X_i) K_b(x)\, dx \\ &= \iint g(x+u) d\hat{F}(u) K_b(x)\, dx.\end{aligned}$$

Der standardisierte Fehler dieses Schätzers ist ein geglättetes Funktional des empirischen Prozesses $\Delta = n^{1/2}(\hat{F} - F)$:

$$n^{1/2}\Big(\int g(x)\hat{f}(x)\,dx - Eg(X)\Big) = \iint g(x+u)d\Delta(u)K_b(x)\,dx.$$

Wir erwarten die Rate $n^{-1/2}$. Im folgenden behandeln wir ein weniger offensichtliches Beispiel.

Seien $X_1, \ldots, X_n$ unabhängig und identisch verteilt. Für $m \geq 2$ seien $u_1, \ldots, u_m$ reellwertige Funktionen, so daß $u_j(X_j)$ Dichte f_j hat. Wir wollen die Dichte g der Summe $u_1(X_1) + \cdots + u_m(X_m)$ schätzen. Es gilt

$$\begin{aligned} g(y) &= f_1 * \cdots * f_m(y) \\ &= \int \ldots \int f_1(y - y_2 - \cdots - y_m) f_2(y_2) \ldots f_m(y_m)\,dy_2 \ldots dy_m. \end{aligned}$$

Also ist g ein (glattes) Funktional von $f_1, \ldots, f_m$. Einen *Plug-in-Schätzer* für g erhält man, indem man Schätzer für $f_1, \ldots, f_m$ einsetzt. Wir verwenden Kernschätzer

$$\hat{f}_j = \frac{1}{n}\sum_{i=1}^{n} K_{jb_j}(y - u_j(X_j))$$

mit Kernen K_j, Bandweiten b_j, und $K_{jb_j}(y) = K_j(y/b_j)/b_j$. Es gilt

$$E\hat{f}_j = f_j * K_{jb_j}.$$

Sei $\mathcal{A}$ die Klasse der Funktionen $f \in L_1$, für die $f' \in L_1$ existiert mit $f(x) = \int_{-\infty}^{x} f'(t)\,dt$ für $x \in \mathbb{R}$.

Das folgende Lemma beweisen wir nicht.

Lemma 6 *(Satz von Fréchet–Kolmogorov) Eine abgeschlossene Teilmenge von L_1 ist kompakt genau dann, wenn*

$$\begin{aligned} &\sup_{h\in H} \|h\|_1 < \infty, \\ &\lim_{\delta\downarrow 0} \sup_{|t|<\delta} \sup_{h\in H} \int |h(x-t) - h(x)|\,dx = 0, \\ &\lim_{c\uparrow\infty} \sup_{h\in H} \int_{|x|>c} |h(x)|\,dx = 0. \end{aligned}$$

Sei g_j die Dichte von

$$u_1(X_1) + \cdots + u_{j-1}(X_{j-1}) + u_{j+1}(X_{j+1}) + \cdots + u_m(X_m).$$

Setze

$$\begin{aligned}\mathbb{H}_{nj}(y) &= n^{-1/2}\sum_{i=1}^{n}(g_j(y-u_j(X_i))-g(y)) \\ &= n^{-1/2}\sum_{i=1}^{n}(g_j(y-u_j(X_i))-Eg_j(y-u_j(X))).\end{aligned}$$

Definition. Eine Familie von Zufallselementen $H_1, H_2, \ldots$ in L_1 ist *straff*, wenn für jedes $\varepsilon > 0$ eine kompakte Menge $K \subset L_1$ existiert mit

$$P(H_n \in K) \geq 1-\varepsilon, \quad n = 1, 2, \ldots.$$

Wir zeigen, daß $\mathbb{H}_{1j}, \mathbb{H}_{2j}, \ldots$ straff in L_1 ist. Dazu schreiben wir $\mathbb{H}_{nj}$ als Glättung $n^{1/2}(\mathbb{F}_j - F_j) * g_j'$ des empirischen Prozesses der $u_j(X_1), \ldots, u_J(X_n)$ und verwenden die Straffheit des empirischen Prozesses.

Lemma 7 *Hat f_j endliches $(2+\varepsilon)$-Moment für ein $\varepsilon > 0$ und ist $g_j \in \mathcal{A}$, so ist $\mathbb{H}_{1j}, \mathbb{H}_{2j}, \ldots$ straff in L_1.*

Beweis. Sei F_j die Verteilungsfunktion von f_j und $\mathbb{F}_j$ die empirische Verteilungsfunktion von $u_j(X_1), \ldots, u_j(X_n)$. Setze $\Delta_j = n^{1/2}(\mathbb{F}_j - F_j)$. Dann gilt

$$\begin{aligned}E\|\Delta\|_1 &= \int E(n^{1/2}|\mathbb{F}_j - F_j|)\,dx \\ &\leq (E(\mathbf{1}(u_j(X) \leq x) - F_j(x))^2)^{1/2} \\ &= \int F_j^{1/2}(x)(1-F_j^{1/2}(x))\,dx < \infty.\end{aligned}$$

Also gilt $\Delta_j \in L_1$ (f.s.). Ebenso gilt für $c \to \infty$:

$$E\int_{|x|>c}|\Delta_j(x)|\,dx \leq \int_{|x|>c} F_j^{1/2}(x)(1-F_j^{1/2}(x)) \to 0.$$

Wir können $\mathbb{H}_{nj}$ wie folgt schreiben:

$$\mathbb{H}_{nj}(y) = \int g_j(y-u)d\Delta_j(u).$$

Mit partieller Integration und Variablentransformation:

$$\mathbb{H}_{nj}(y) = \int \Delta_j(u)g_j'(y-u)\,du = \Delta_j * g_j'(y).$$

Also gilt

$$\|\mathbb{H}_{nj}\|_1 \le \|\Delta_j\|_1 \|g_j'\|_1,$$
$$\sup_{|t|<\delta} \int |\mathbb{H}_{nj}(x-t) - \mathbb{H}_{nj}(x)|\, dx \le \sup_{|t|<\delta} \int |g_j(x-t) - g_j(x)|\, dx \|\Delta_j\|_1,$$
$$\int_{|x|>2c} |\mathbb{H}_{nj}(x)|\, dx \le \int_{|x|>c} |\Delta_j(x)|\, dx \|g_j'\| + \int_{|x|>c} |g_j'(x)|\, dx \|\Delta_j\|_1.$$

Weil g' integrierbar ist, folgt aus Lemma 4 für $\delta \downarrow 0$:

$$\sup_{|t|<\delta} \int |g_j(x-t) - g_j(x)|\, dx \to 0.$$

Die Behauptung folgt jetzt aus Lemma 6.

Satz 24 *Es habe $f_1, \dots, f_m$ endliches $(2+\varepsilon)$-Moment, und es gelte*

$$g \in \mathcal{L}_{r+1,0}, \quad g_1, \dots, g_m \in \mathcal{A}, \quad K_1, \dots, K_m \in \mathcal{K}_{r,1}.$$

Gilt $n(\min b_j)^2 \to \infty$ and $n(\max b_j)^{2(r+1)} \to 0$, dann gilt

$$\|n^{1/2}(\hat{g} - g) - (\mathbb{H}_{n1} + \dots + \mathbb{H}_{nm})\|_1 = o_p(1).$$

Beweis. Nach Satz 23 gilt

$$E\|\hat{f}_j - E\hat{f}_j\|_1 = O((nb_j)^{-1/2}),$$

also

$$\|\hat{f}_j - E\hat{f}_j\|_1 = O_p((nb_j)^{-1/2}).$$

Setze $\bar{g} = E\hat{f}_1 * \dots * E\hat{f}_m = f_1 * \dots * f_m * K_{1b_1} * \dots * K_{mb_m} = g * K_*$ mit $K_* = K_{1b_1} * \dots * K_{mb_m}$. Wie in Satz 22 gilt

$$\|\bar{g} - g\|_1 = O((\max b_j)^{r+1}).$$

Für $A \subset \{1, \dots, m\}$ setze $\gamma_A = *_{j\in A}(\hat{f}_j - E\hat{f}_j) * *_{j\notin A} E\hat{f}_j$ mit $\gamma_\emptyset = E\hat{g}$ und $\gamma_{\{1,\dots,m\}} = *_{j=1}^m(\hat{f}_j - E\hat{f}_j)$. Es gilt

$$\hat{g} = *_{j=1}^m \hat{f}_j = *_{j=1}^m (E\hat{f}_j + \hat{f}_j - E\hat{f}_j) = \sum_A \gamma_A = \sum_{r=0}^m \Gamma_r$$

mit $\Gamma_r = \sum_{|A|=r} \gamma_A$. Insbesondere haben wir $\Gamma_0 = \gamma_\emptyset = E\hat{g}$ und $\Gamma_1 = \sum_{j=1}^m (\hat{f}_j - E\hat{f}_j) * \bar{g}_j$ mit $\bar{g}_j = *_{i\neq j} E\hat{f}_i = g_j * *_{i\neq j} K_{ib_i}$. Wegen $\|a * b\|_1 \leq \|a\|_1 \|b\|_1$ gilt für $A \subset \{1, \ldots, m\}$ mit mindestens zwei Elementen j, k:

$$\|\gamma_A\|_1 \leq \|\hat{f}_j - E\hat{f}_j\|_1 \|\hat{f}_k - E\hat{f}_k\|_1 \prod_{i\in A\setminus\{j,k\}} \|\hat{f}_i - E\hat{f}_i\|_1 \prod_{i\in A} \|E\hat{f}_i\|_1.$$

Also gilt mit obiger Darstellung für $\hat{g}$:

$$\left\|\hat{g} - \bar{g} - \sum_{j=1}^m (\hat{f}_j - E\hat{f}_j) * \bar{g}_j\right\|_1 \leq \sum_{r=2}^m \|\Gamma_r\|_1 = O_p((n \min b_j)^{-1}).$$

Es gilt

$$n^{1/2}(\hat{f}_j - E\hat{f}_j) * \bar{g}_j = \mathbb{H}_{nj} * K_*.$$

Also bleibt zu zeigen:

$$\|\mathbb{H}_{nj} * K_* - \mathbb{H}_{nj}\|_1 = o_p(1).$$

Dazu schätzen wir ab:

$$\begin{aligned}
\|\mathbb{H}_{nj} * K_* - \mathbb{H}_{nj}\|_1 &= \int \Big| \int (\mathbb{H}_{nj}(x-u) - \mathbb{H}_{nj}(x)) K_*(u)\, du \Big|\, dx \\
&\leq \sup_{|u|<\delta} \int |\mathbb{H}_{nj}(x-u) - \mathbb{H}_{nj}(x)|\, dx \int_{|u|<\delta} |K_*(u)|\, du \\
&\quad + 2\|\mathbb{H}_{nj}\|_1 \int_{|u|\geq\delta} |K_*(u)|\, du \\
&\leq \sup_{|u|<\delta} \int |\mathbb{H}_{nj}(x-u) - \mathbb{H}_{nj}(x)|\, dx \|K_*\|_1 \\
&\quad + 2\|\mathbb{H}_{nj}\|_1 \delta^{r+1} \int |u|^{r+1} |K_*(u)|\, du.
\end{aligned}$$

Mit der Abschätzung aus dem Beweis von Lemma 7 folgt die Behauptung.

Für einen Zentralen Grenzwertsatz über $\hat{g}$ benötigen wir einen Zentralen Grenzwertsatz für Banachraum-wertige Zufallselemente. Wir geben hier ohne Beweis eine Version für L_1.

Lemma 8 *Sei $H_1, H_2, \ldots$ eine straffe Folge L_1-wertiger Zufallselemente. Sei $\int \varphi(x) H_n(x)\, dx$ asymptotisch normal für alle stetigen φ mit kompaktem Träger. Dann konvergiert H_n schwach in L_1 gegen einen Gaußschen Prozeß.*

Korollar 2 *Unter den Voraussetzungen von Satz 24 konvergiert* $n^{1/2}(\hat{g}-g)$ *schwach in* L_1 *gegen einen zentrierten Gaußschen Prozeß mit Kovarianzfunktion*

$$(x,y) \to \sum_{i,j=1}^{m} E\big(g_i(x-u_i(X))-g(x)\big)\big(g_j(x-u_j(X))-g(x)\big).$$

Für $m=2$ gilt $g_1=f_2$ und $g_2=f_1$. Die Voraussetzungen von Satz 24 implizieren also, daß f_1, f_2 differenzierbar sind. Setzen wir $r=1$, so ist nach Kapitel 14 die optimale Bandweite für $\hat{f}_j$ von der Ordnung $n^{-1/3}$. In Satz 24 lassen wir Bandweiten $n^{-1/2} \ll b \ll n^{-1/4}$ zu. Wir brauchen also weder zu überglätten noch zu unterglätten.

17 Nichtparametrische Regression und Nadaraya–Watson-Schätzer

Sei (X,Y) eien zweidimensionale Zufallsvariable. Die *Regressionsfunktion* von Y auf X ist der bedingte Erwartungswert

$$g(X) = E(Y|X).$$

Wir beobachten unabhängige Realisationen (X_i,Y_i), $i=1,\dots,n$. Der *Nadaraya–Watson-Schätzer* für $g(x)$ ist

$$\frac{\sum_{i=1}^n K_b(x-X_i)Y_i}{\sum_{i=1}^n K_b(x-X_i)}.$$

Hier ist K ein Kern und b eine Bandweite.

Sei $f(x,y)$ die Dichte von (X,Y), und $f(x)=\int f(x,y)\,dy$ die Dichte von X. Dann ist $f(x,y)/f(x)$ die bedingte Dichte von Y gegeben $X=x$, also $g(x)=\int yf(x,y)\,dy/f(x)$. Ist f stetig, so konvergiert $\frac{1}{n}\sum_{i=1}^n K_b(x-X_i)$ gegen $f(x)$; siehe Kapitel 14. Für den Zähler von $\hat{g}(x)$ haben wir unter geeigneten Annahmen

$$\begin{aligned} EK_b(x-X)Y &= \iint K_b(x-z)yf(z,y)\,dz\,dy \\ &= \iint K(u)f(x-bu,y)y\,du\,dy \\ &\to \int K(u)\,du \int yf(x,y)dy = \int yf(x,y)\,dy. \end{aligned}$$

Konvergenzraten erhält man wie bei Dichteschätzern.

18 Lokale polynomiale Glätter

Wir bleiben beim Schätzen der Regressionsfunktion g. Ist g ein (unbekanntes) Polynom mit (bekanntem) Grad r, so können wir g schätzen, indem wir die Beobachtungen (X_i, Y_i) durch ein Polynom vom Grad r approximieren, das den mittleren quadratischen Fehler

$$\sum_{i=1}^{n} \Big(Y_i - \sum_{k=0}^{r} \vartheta_k X_i^k \Big)^2$$

minimiert. Durch Differentiation erhalten wir

$$\sum_{i=1}^{n} X_i^k \Big(Y_i - \sum_{k=0}^{r} \hat{\vartheta}_k X_i^k \Big) = 0,$$

also

$$\hat{\vartheta} = \Big(\sum_{i=1}^{n} Z_i Z_i^\top \Big)^{-1} \sum_{i=1}^{n} Z_i Y_i$$

mit $Z_i = (X_i^0, \ldots, X_i^r)^\top$. Der Schätzer $\hat{g}(x) = \sum_{k=0}^{r} \hat{\vartheta}_k x^k$ ist ein Kleinste-Quadrate-Schätzer; wir nennen ihn *polynomialen Glätter*.

Für das *Lageparameter-Modell* $Y = \vartheta + \varepsilon$ mit $E\varepsilon = 0$ ist $g(x) = \vartheta$ und $\hat{\vartheta} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

Für das *lineare Regressionsmodell* mit *Interzept* ϑ_0, $Y = \vartheta_0 + \vartheta_1 X + \varepsilon$ mit $E(\varepsilon|X) = 0$, ist $g(X) = \vartheta_0 + \vartheta_1 X$ und

$$ZZ^\top = \begin{pmatrix} 1 & X \\ X & X^2 \end{pmatrix}.$$

Für das *lineare Regressionsmodell ohne Interzept*, $Y = \vartheta X + \varepsilon$ mit $E(\varepsilon|X) = 0$, ist $g(X) = \vartheta X$. Wir können also den Index 0 weglassen und erhalten

$$\hat{\vartheta} = \Big(\sum_{i=1}^{n} X_i^2 \Big)^{-1} \sum_{i=1}^{n} X_i Y_i.$$

Das hatten wir in Kapitel 12 schon allgemeiner hergeleitet, nämlich für $Y = \vartheta^\top X + \varepsilon$ mit Kovariablen*vektor* X.

Ist die Funktion g kein Polynom der Ordnung r, aber r-mal differenzierbar, so können wir g lokal durch ein Polynom approximieren und obige

Methode lokal anwenden. Sei x fest. Eine Taylor-Entwicklung liefert für z nahe x:

$$g(z) \approx \sum_{k=0}^{r} \frac{g^{(k)}(x)}{k!}(z-x)^k.$$

Den Koeffizienten $(\vartheta_0, \ldots, \vartheta_r)$ entsprechen $(g(x), g'(x), \ldots, g^{(r)}(x)/r!)$. Wir minimieren den mittleren quadratischen Fehler

$$\sum_{i=1}^{n} \Big(\frac{g^{(k)}(x)}{k!}(X_i - x)^k \Big)^2 K_b(X_i - x).$$

Sei $\hat{Q} = (\hat{Q}_{jk})_{j,k=0,\ldots,r}$ die Matrix mit Elementen

$$\hat{Q}_{jk} = \sum_{i=1}^{n} (X_i - x)^{j+k} K_b(X_i - x),$$

und sei $\hat{W} = (\hat{W}_0, \ldots, \hat{W}_r)^\top$ der Vektor mit Elementen

$$\hat{W}_k = \sum_{i=1}^{n} (X_i - x)^k K_b(X_i - x) Y_i.$$

Wir erhalten also einen Schätzer für $g(x), g'(x), \ldots, g^{(r)}(x)$, den *lokalen polynomialen Glätter*, durch

$$\Big(\hat{g}(x), \hat{g}'(x), \ldots, \frac{\hat{g}^{(r)}(x)}{r!} \Big) = \hat{Q}^{-1} \hat{W}.$$

Für $r = 0$ ergibt sich der Nadaraya–Watson-Schätzer

$$\hat{g}(x) = \frac{\hat{W}_0}{\hat{Q}_0} = \frac{\sum_{i=1}^{n} K_b(X_i - x) Y_i}{\sum_{i=1}^{n} K_b(X_i - x)}.$$

Dieser lokal konstante Glätter für eine Regressionsfunktion entspricht dem Kernschätzer $\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_b(X_i - x)$ für eine Dichte.

19 Kontiguität

Wir wissen schon, daß wir glatte Funktionale einer unbekannten Verteilung mit der Rate $n^{-1/2}$ schätzen können, wobei n der Stichprobenumfang ist. Das gilt für parametrische Modelle (Maximum-Likelihood-Schätzer) und

nichtparametrische Modelle (M-Schätzer, empirische Schätzer). Für parametrische Modelle bedeutet das insbesondere, daß wir zwei Parameter mit Sicherheit asymptotisch unterscheiden können, wenn ihr Abstand langsamer als $n^{-1/2}$ schrumpft. Interessant sind also die Abst"ande der Ordnung $n^{-1/2}$. Wir zeigen im folgenden, daß $n^{-1/2}$ tatsächlich die beste Konvergenzrate für Schätzer glatter Funktionale ist. Dazu brauchen wir nur eine wie $n^{-1/2}$ schrumpfende Umgebung eines Parameters zu betrachten. Das sieht man auch heuristisch, wenn man den Likelihoodquotienten ansieht. Ist f_ϑ die Dichte eines Wahrscheinlichkeitsmaßes P_ϑ, ist Θ eine offene Teilmenge von $\mathbb{R}$, und sind n unabhängige Beobachtungen $X_1, \dots, X_n$ aus einem der P_ϑ gegeben, so ist der Log-Likelihoodquotient

$$\begin{aligned}\log \frac{dP^n_{\vartheta+n^{-1/2}t}}{dP^n_\vartheta}(X_1,\dots,X_n) &= \sum_{i=1}^n (\ell_{\vartheta+n^{-1/2}t}(X_i) - \ell_\vartheta(X_i)) \\ &\approx tn^{-1/2}\sum_{i=1}^n \dot\ell_\vartheta(X_i) + \frac{1}{2}t^2\frac{1}{n}\sum_{i=1}^n \ddot\ell(X_i) \\ &= tn^{-1/2}\sum_{i=1}^n \dot\ell_\vartheta(X_i) - \frac{1}{2}t^2 I_\vartheta\end{aligned}$$

mit $\ell_\vartheta = \log f_\vartheta$; $\dot\ell_\vartheta$, $\ddot\ell_\vartheta$ den Ableitungen nach ϑ; und $I_\vartheta = E\dot\ell^2_\vartheta$ der Fisher-Information wie in Kapitel 11. Also hat $\log dP^n_\tau/dP^n_\vartheta$ genau dann eine nicht ausgeartete Verteilung, wenn $\tau - \vartheta$ von der Ordnung $n^{-1/2}$ ist. Dann sind P^n_τ und P^n_ϑ "benachbart". Für $\tau = \vartheta + n^{-1/2}t$ ist die Grenzverteilung eine Normalverteilung mit Mittelwert $-\frac{1}{2}t^2 I_\vartheta$ und Varianz $t^2 I_\vartheta$. Wir sagen dann, wir haben "lokale asymptotische Normalität" (LAN). Daraus werden wir im folgenden schließen, daß $n^{-1/2}$ die otimale Konvergenzrate für Schätzer glatter Funktionale von ϑ ist. Genauer werden wir zeigen, daß es keinen "regulären" Schätzer gibt, der eine bessere asymptotische Varianz als I_ϑ^{-1} hat. Das ist die asymptotische Varianz des Maximum-Likelihood-Schätzers.

Für eine Normalverteilung $N_{tI,I}$ mit Lageparameter tI gilt

$$\log \frac{dN_{tI,I}}{dN_{0,I}}(X) = \frac{(X-tI)^2}{2I} + \frac{X^2}{2I} = tX - \frac{1}{2}t^2 I.$$

Asymptotisch ist der obige Log-Likelihoodquotient von dieser Form. Also reduziert sich das ursprüngliche Schätzproblem asymptotisch auf diesen Fall.

Definition. Eine Folge reeller Zufallsvariablen V_n heißt *gleichmäßig integrierbar* unter P_n, wenn

$$\sup_n P_n(|V_n|\mathbf{1}(|V_n| > c)) \to 0, \quad c \to \infty.$$

Die folgenden Lemmas zitieren wir ohne Beweis.

Lemma 9 *Die Folge V_n ist gleichmäßig integrierbar genau dann, wenn $P_n|V_n|$ beschränkt ist und*

$$P_n|V_n|A_n \to 0, \quad \textit{wenn} \quad P_nA_n \to 0.$$

Lemma 10 *Gelte $V_n \Rightarrow V$. Dann ist V_n ist gleichmäßig integrierbar genau dann, wenn $P_n|V_n|$ beschränkt ist und*

$$P|V| < \infty \quad \textit{und} \quad P_n|V_n| \to P|V|.$$

Lemma 11 *Gilt $V_n \Rightarrow V$, so auch* $\liminf_n P_n|V_n| \geq P|V|$.

Gegeben seien Wahrscheinlichkeitsmaße P_n, Q_n auf $(\Omega_n, \mathcal{F}_n)$. Wie in Kapitel 8 definieren wir mit $p_n = dP_n/d(P_n+Q_n)$ und $q_n = dQ_n/d(P_n+Q_n)$ den Dichtequotienten

$$L_n = \frac{dQ_n}{dP_n} = \frac{q_n}{p_n}\mathbf{1}(p_n > 0) + \infty\mathbf{1}(p_n = 0, q_n = 0).$$

Definition. Q_n ist zu P_n *benachbart* ($Q_n \lhd P_n$), wenn für $A_n \in \mathcal{F}_n$ gilt:

$$P_nA_n \to 0 \quad \text{impliziert} \quad Q_nA_n \to 0.$$

Lemma 12 *Die folgenden Bedingungen sind äquivalent:*
(1) $Q_n \lhd P_n$.
(2) L_n ist straff unter P_n, und wenn $L_n \Rightarrow L$ für eine Teilfolge, dann gilt $PL = 1$.
(3) L_n ist gleichmäßig integrierbar unter P_n, und $Q_n(L_n = \infty) \to 0$.
(4) L_n ist straff unter Q_n.

Beweis. *(3) und (1) sind äquivalent:* Schreibe

$$\begin{aligned} Q_nA_n &= Q_n\mathbf{1}(L_n < \infty)A_n + Q_n\mathbf{1}(L_n = \infty)A_n \\ &= P_nL_nA_n + Q_n\mathbf{1}(L_n = \infty)A_n. \end{aligned} \tag{19.1}$$

Hier haben wir $\mathbf{1}(L_n < \infty)$ weglassen können, weil $P_n(L_n = \infty) = 0$, also $P_n(L_n < \infty) = 1$.

(3) impliziert (1): Seien $A_n \in \mathcal{F}_n$ mit $P_nA_n \to 0$. Es gilt $P_nL_nA_n \to 0$ nach Lemma 9 "$\Rightarrow$" für $V_n = L_n$. Nach Voraussetzung gilt $Q_n(L_n = \infty) \to 0$, also $Q_n\mathbf{1}(L_n = \infty)A_n \to 0$, also $Q_nA_n \to 0$ nach (19.1).

(1) impliziert (3): Seien $A_n \in \mathcal{F}_n$ mit $P_n A_n \to 0$. Nach Voraussetzung gilt $Q_n A_n \to 0$, also $Q_n \mathbf{1}(L_n = \infty) A_n \to 0$. Nach (19.1) gilt $P_n L_n A_n \to 0$. Außerdem gilt $P_n L_n = Q_n(L_n < \infty) \leq 1$. Nach Lemma 9 "$\Leftarrow$" für $V_n = L_n$ ist L_n gleichmäßig integrierbar unter P_n. Außerdem gilt $P_n(L_n = \infty) = 0$, also $Q_n(L_n = \infty) \to 0$.

(2) impliziert (3): Es genügt zu zeigen: Jede Teilfolge hat eine Teilfolge mit (3). Nach dem Satz von Prohorov besitzt jede Teilfolge eine konvergente Teilfolge mit $L_n \Rightarrow L$. Mit Lemma 11 gilt $\liminf_n P_n L_n \geq PL = 1$. Anderseits ist $P_n L_n = Q_n(L_n < \infty) \leq 1$. Also gilt $P_n L_n \to PL$ und $Q_n(L_n < \infty) \to 1$. Also ist L_n gleichmäßig integrierbar unter P_n nach Lemma 10 "$\Leftarrow$", und $Q_n(L_n = \infty) = 1 - Q_n(L_n < \infty) \to 1$.

(3) impliziert (2): Ist P_n gleichmäßig integrierbar unter P_n, so auch straff, denn für $c > 0$ gilt

$$P_n(L_n > c) \leq \frac{1}{c} P_n L_n \mathbf{1}(L_n > c).$$

Gelte nun $L_n \Rightarrow L$ für eine Teilfolge. Dann gilt $P_n L_n \to PL$ nach Lemma 10 "$\Rightarrow$". Außerdem gilt

$$P_n L_n = Q_n(L_n < \infty) = 1 - Q_n(L_n = \infty) \to 1,$$

also $PL = 1$.

(3) impliziert (4): Wegen

$$Q_n(L_n > c) = P_n L_n \mathbf{1}(L_n > c) + Q_n(L_n = \infty).$$

(4) impliziert (1): Wegen

$$\begin{aligned} Q_n A_n &= Q_n(L_n \leq c) A_n + Q_n(L_n > c) \\ &\leq P_n L_n \mathbf{1}(L_n \leq c) A_n + Q_n(L_n > c) \\ &\leq c P_n A_n + Q_n(L_n > c). \end{aligned}$$

Lemma 13 *(Drittes Lemma von Le Cam) Gilt $Q_n \lhd P_n$ und $(L_n, V_n) \Rightarrow (L, V)$ unter P_n, dann auch $(L_n, V_n) \Rightarrow (L, V)$ unter Q_n, und $dQ = L dP$.*

Beweis. Sei $f : [0, \infty] \times \mathbb{R} \to \mathbb{R}$ stetig und beschränkt. Schreibe

$$Q_n f(L_n, V_n) = P_n L_n f(L_n, V_n) + Q_n \mathbf{1}(L_n = \infty) f(L_n, V_n).$$

Der zweite Term geht gegen 0, weil $Q_n(L_n = \infty) \to 0$ nach Lemma 9 "(1) $\Rightarrow$ (3)". Schreibe den ersten Term als

$$P_n(L_n \wedge c) f(L_n, V_n) + P_n(L_n - c)^+ f(L_n, V_n).$$

Der linke Summand konvergiert gegen $P(L \wedge c)f(L,V)$; für große c geht das gegen $PLf(L,V)$. Der rechte Summand ist klein für großes c nach Lemma 9 "(1) $\Rightarrow$ (3)" mit $Q_n(L_n - c)^+ \leq Q_n L_n(L_n \geq c)$.

Für den Beweis des nächsten Lemmas erinnern wir daran ,daß die charakteristische Funktion einer mehrdimensionlaen Normalverteilung $N_{\mu,\Sigma}$ die Form $\varphi(t) = \exp(i\mu^\top t - \frac{1}{2}t^\top \Sigma t)$ hat.

Lemma 14 *Gelte $Q_n \lhd P_n$. Setze $\Lambda_n = \log L_n$. Seien V_n Zufallsvariablen mit $(\Lambda_n, V_n) \Rightarrow (\Lambda, V)$ unter P_n, wo (Λ, V) einer zweidimensionalen Normalverteilung folgt mit Mittelwertvektor und Kovarianzmatrix*

$$\begin{pmatrix} -\frac{1}{2}a \\ b \end{pmatrix} \quad \textit{und} \qquad \begin{pmatrix} a & c \\ c & d \end{pmatrix}.$$

Dann gilt $(\Lambda_n, V_n) \Rightarrow N$ unter Q_n, wo N eine zweidimensionale Normalverteilung mit derselben Kovarianzmatrix ist wie (Λ, V), und mit Mittelwertvektor

$$\begin{pmatrix} \frac{1}{2}a \\ b + c \end{pmatrix}.$$

Beweis. Unter P_n hat die charakteristische Funktion der Grenzverteilung die Form $\varphi(s,t) = P\exp(is\Lambda + itV)$. Nach Lemma 13 hat die charakteristische Funktion der Grenzverteilung unter Q_n die Form

$$PL\exp(is\Lambda + itV) = P\exp((is+1)\Lambda + itV) = \varphi(s-i,t).$$

Es gilt

$$\varphi(s,t) = \exp\Big(-\frac{1}{2}isa + itb - \frac{1}{2}(s^2a + t^2d + 2stc)\Big).$$

Also ist die charakteristische Funktion unter Q_n von der Form

$$\begin{aligned}\varphi(s-i,t) &= \exp\Big(-\frac{1}{2}isa - \frac{1}{2}a + itb - \frac{1}{2}(s^2a - 2isa - a + t^2d + 2stc - 2itc)\Big) \\ &= \exp\Big(\frac{1}{2}isa + it(b+c) - \frac{1}{2}(s^2a + t^2d + 2stc)\Big).\end{aligned}$$

20 Faltungssatz

Sei f stetig und $f_n \to f$ punktweise. Im "regulären" Fall, für "gutartige" f_n, wird die Konvergenz sogar *stetig* sein:

$$f_n(x_n) \to f(x), \quad \text{wenn} \quad x_n \to x.$$

Für die Verteilungen der Schätzerfolgen werden wir nur eine abgeschwächte Version der stetigen Konvegenz annehmen:

$$n^{1/2}(\hat{\vartheta} - (\vartheta + n^{-1/2}t)) \Rightarrow V \quad \text{unter } P^n_{\vartheta+n^{-1/2}t}, \quad t \in \mathbb{R}.$$

Solche Schätzer werden wir *regulär* nennen. Für sie werden wir eine (scharfe) untere Schranke für die asymptotische Varianz angeben. Für Anwendungen auf nichtparametrische Modelle ist es von Vorteil, alle folgenden Aussagen zunächst für *lokale* Modelle zu formulieren, also für P_{nt} statt $P^n_{\vartheta+n^{-1/2}t}$, und für V_n statt $n^{1/2}(\hat{\vartheta} - \vartheta)$.

Setze $L_{nt} = dP_{nt}/dP_n$ und $\Lambda_{nt} = \log L_{nt}$.

Definition. Die Familie P_{nt}, $t \in \mathbb{R}$, heißt *asymptotisch normal*, wenn

$$\Lambda_{nt} = tS_n - \frac{1}{2}t^2 J + o_{P_n}(1), \quad t \in \mathbb{R},$$
$$S_n \Rightarrow J^{1/2}N,$$

wobei N eine standardnormalverteilte Zufallsvariable und J eine positive Zahl ist.

Lemma 15 *Ist P_{nt}, $t \in \mathbb{R}$, asymptotisch normal, so sind P_{ns} und P_{nt} für alle s und t benachbart.*

Beweis. Weil $\lhd$ transitiv ist, genügt es, $t = 0$ zu betrachten. Es gilt $P_{ns} \lhd P_{n0}$ nach Lemma 12 "(4) $\Rightarrow$ (1)". Es gilt $P_{n0} \lhd P_{ns}$ nach Lemma 12 "(2) $\Rightarrow$ (1)". wenn dort die Rollen von P_n und Q_n vertauscht werden.

Definition. V_n heißt *regulär* mit *Limes* V, wenn

$$V_n - t \Rightarrow V \quad \text{unter } P_{nt}, \quad t \in \mathbb{R}.$$

Satz 25 *(Faltungssatz von Hájek und Le Cam) Ist P_{nt} asymptotisch normal und V_n regulär, so gilt*

$$(J^{-1}S_n, V_n - J^{-1}S_n) \Rightarrow (J^{-1/2}N, U) \quad \textit{unter } P_n$$

für eine Zufallsvariable U, die unabhängig von N ist.

Beweis. Da (S_n, V_n) straff unter P_n ist, gilt $(S_n, V_n) \Rightarrow (J^{-1/2}N, V)$ unter P_n für eine Teilfolge. Nach Lemma 15 sind P_{ns} und P_{nt} benachbart. Nach Lemma 13 gilt für jedes stetige und beschränkte f:

$$P_t f(V) = P_s \exp(\Lambda_t - \Lambda_s) f(V).$$

Insbesondere gilt

$$P_t \exp(iu(V-t)) = P_s \exp\Big(iuV - iut + tJ^{1/2}N - \frac{1}{2}t^2 J - sJ^{1/2}N + \frac{1}{2}s^2 J\Big).$$

Da V_n regulär ist, hängt die linke Seite nicht von t ab. Die rechte Seite ist analytisch in t. Ersetze t durch $s - iJ^{-1}u$:

$$P \exp(iuV) = \exp\Big(-\frac{1}{2}u^2 J^{-1}\Big) P_s \exp(iu(V - J^{-1/2}N)).$$

Also hängt die Verteilung von $V - J^{-1/2}N$ unter P_s nicht von s ab. Nach Satz 6 ist also $V - J^{-1/2}N$ unabhängig von N. Da S_n und V_n in Verteilung konvergieren, hängt die Grenzverteilung von $V - J^{-1/2}N$ nicht von der Teilfolge ab.

Insbesondere ist $V_n = J^{-1}S_n + V_n - J^{-1}S_n$ asymptotisch wie die Faltung $J^{-1/2}N + U$ verteilt. Daher der Name "Faltungssatz". Ist M unabhängig von N, so gilt

$$P(-a < N + M < a) \le P(-a < N < a), \quad a > 0.$$

Also ist V_n bestenfalls asymptotisch wie $J^{-1/2}N$ in symmetrischen Intervallen um 0 konzentriert. V_n ist asymptotisch optimal genau dann, wenn U nach δ_0 verteilt ist. Genau dann gilt $V_n = J^{-1}S_n + o_{P_n}(1)$.

21 Lokale asymptotische Normalität für unabhängige Beobachtungen

Gegeben sei eine Familie von äquivalenten Wahrscheinlichkeitsmaßen P_ϑ, $\vartheta \in \Theta$, mit $\Theta \subset \mathbb{R}$ offen. Sei ϑ fest.

Definition. P_τ heißt *Hellinger-differenzierbar* in $\tau = \vartheta$ mit *Ableitung* $\dot{\ell}_\vartheta \in L_{2,0}(P_\vartheta)$, wenn

$$\Big\| L_{\tau\vartheta}^{1/2} - 1 - \frac{1}{2}(\tau - \vartheta)\dot{\ell}_\vartheta \Big\|_2 = o(\tau - \vartheta).$$

Wegen $P_\vartheta L_{\tau\vartheta} = P_\tau \Omega = 1$ und $x - 1 = 2(x^{1/2} - 1) + (x^{1/2} - 1)^2$ gilt

$$P_\vartheta(L_{\tau\vartheta} - 1) = -\frac{1}{2} P_\vartheta (L_{\tau\vartheta} - 1)^2. \tag{21.1}$$

Bezeichne $I_\vartheta = P_\vartheta \dot\ell^2$ die *Fisher-Information*. Setze

$$L_{nt} = \frac{dP^n_{\vartheta+n^{-1/2}t}}{dP^n_\vartheta}, \quad \Lambda_{nt} = \log L_{nt}.$$

Satz 26 *Ist P_τ Hellinger-differenzierbar in $\tau = \vartheta$ mit* Ableitung $\dot\ell_\vartheta$*, so gilt*

$$\Lambda_{nt} = tn^{-1/2}\sum_{i=1}^n \dot\ell_\vartheta(X_i) - \frac{1}{2}t^2 I_\vartheta + o_{P_\vartheta}(1),$$
$$n^{-1/2}\sum_{i=1}^n \dot\ell_\vartheta(X_i) \Rightarrow I_\vartheta^{1/2} N.$$

Wir sagen dann: P_τ ist *lokal asymptotisch normal* (LAN) in ϑ.

Beweis. Schreibe $\log x = 2\log(1 + x^{1/2} - 1)$. Wir verwenden die Taylor-Entwicklung $\log(1+x) = x - \frac{1}{2}x^2 + r(x)$. Für $Z = dP_{\vartheta+n^{-1/2}t}/dP_\vartheta$ ergibt sich

$$\begin{aligned}\log Z &= 2\log(1 + Z^{1/2} - 1) = 2(Z^{1/2}-1) - (Z^{1/2}-1)^2 + 2r(Z^{1/2}-1)\\ &= 2\big(Z^{1/2} - 1 - P_\vartheta(Z^{1/2}-1)\big) - 2P_\vartheta(Z^{1/2}-1)^2\\ &\quad + 2r(Z^{1/2}-1) - \big((Z^{1/2}-1)^2 - P_\vartheta(Z^{1/2}-1)^2\big).\end{aligned}$$

Mit $Z_i = Z(X_i)$ und (21.1) also

$$\begin{aligned}\Lambda_{nt} &= \sum_{i=1}^n \log Z_i = 2\sum_{i=1}^n \big((Z_i^{1/2}-1)^2 - P_\vartheta(Z_i^{1/2}-1)^2\big) - 2nP_\vartheta(Z_i^{1/2}-1)^2\\ &\quad + 2\sum_{i=1}^n r(Z_i^{1/2}-1) - \sum_{i=1}^n \big((Z_i^{1/2}-1)^2 - P_\vartheta(Z_i^{1/2}-1)^2\big).\end{aligned}$$

Der erste Term ist $tn^{-1/2}\sum_{i=1}^n \dot\ell_\vartheta(X_i)$ bis auf $o_{P_\vartheta}(1)$. Der zweite Term konvergiert in Wahrscheinlichkeit gegen $-\frac{1}{2}P_\vartheta\dot\ell_\vartheta^2 = -\frac{1}{2}t^2 I_\vartheta$. Für den dritten Term haben wir

$$\begin{aligned}nP_\vartheta(|Z^{1/2}-1| > \varepsilon) &\le \frac{n}{\varepsilon^2}P_\vartheta(Z^{1/2}-1)^2\mathbf{1}(|Z^{1/2}-1|>\varepsilon)\\ &= \frac{t^2}{4\varepsilon^2}P_\vartheta\dot\ell_\vartheta^2\mathbf{1}(|\dot\ell_\vartheta| > 2t^{-1}n^{1/2}\varepsilon) + o(1) \to 0.\end{aligned}$$

Das gilt auch für eine Folge $\varepsilon_n \to 0$. Mit $|r(x)| \le 2|x|^3$ für $|x| \le \frac{1}{2}$ und dem Gesetz der großen Zahl gilt

$$\sum_{i=1}^n (Z_i^{1/2} - 1)\mathbf{1}(|Z_i^{1/2} - 1| \le \varepsilon_n) \le 2\varepsilon_n \sum_{i=1}^n (Z_i^{1/2} - 1)^2$$
$$= \frac{t^2}{2n}\varepsilon_n \sum_{i=1}^n \dot{\ell}_\vartheta^2(X_i) + o_{P_\vartheta}(1) = o_{P_\vartheta}(1).$$

Der vierte Term ist $\frac{t^2}{4n}\sum_{i=1}^n (\dot{\ell}_\vartheta^2(X_i) - P_\vartheta \dot{\ell}_\vartheta^2)$ bis auf $o_{P_\vartheta}(1)$, also von der Ordnung $o_{P_\vartheta}(1)$ nach dem Gesetz der großen Zahl.

22 Effiziente Schätzer für parametrische Familien

Gegeben sei eine Familie von äquivalenten Wahrscheinlichkeitsmaßen P_ϑ, $\vartheta \in \Theta$, mit $\Theta \subset \mathbb{R}$ offen. Sei ϑ fest. Seien $X_1, \dots, X_n$ nach P_ϑ verteilt.

Definition. Ein Schätzer $\hat{\vartheta}$ heißt *regulär* in ϑ mit *Limes* V, wenn

$$n^{1/2}(\hat{\vartheta} - (\vartheta + n^{-1/2}t)) \Rightarrow V \quad \text{unter } P_{\vartheta + n^{-1/2}t}, \quad t \in \mathbb{R}.$$

Definition. Ein Schätzer $\hat{\vartheta}$ heißt *asymptotisch linear* in ϑ mit *Einflußfunktion* h, wenn $h \in L_{2,0}(P_\vartheta)$ und

$$n^{1/2}(\hat{\vartheta} - \vartheta) = n^{-1/2} \sum_{i=1}^n h(X_i) + o_{P_\vartheta}(1).$$

Ein solcher Schätzer ist asymptotisch normal mit Varianz $Eh^2(X)$.

Satz 27 *Sei P_τ Hellinger-differenzierbar in ϑ und $\hat{\vartheta}$ asymptotisch linear in ϑ mit Einflußfunktion h. Dann ist $\hat{\vartheta}$ regulär genau dann, wenn $h - I_\vartheta^{-1}\dot{\ell}_\vartheta \perp \dot{\ell}_\vartheta$ unter P_ϑ.*

Beweis. Es ist $(tn^{-1/2}\sum_{i=1}^n \dot{\ell}_\vartheta(X_i), n^{-1/2}\sum_{i=1}^n h(X_i))$ asymptotisch normal mit Kovarianz $tP_\vartheta \dot{\ell}_\vartheta h$. Nach Satz 26 und Lemma 14 gilt

$$n^{1/2}(\hat{\vartheta} - (\vartheta + n^{-1/2}t)) \Rightarrow t(P_\vartheta h^2)^{1/2} N + tP_\vartheta \dot{\ell}_\vartheta h - t \quad \text{unter } P_{\vartheta + n^{-1/2}t}.$$

Es ist also $\hat{\vartheta}$ regulär genau dann, wenn $P_\vartheta \dot{\ell}_\vartheta h = 1$, das heißt

$$P_\vartheta(h - I_\vartheta^{-1}\dot{\ell}_\vartheta)\dot{\ell}_\vartheta = P_\vartheta \dot{\ell}_\vartheta h - 1 = 0.$$

Satz 28 *Sei P_τ Hellinger-differenzierbar in ϑ und $\hat{\vartheta}$ regulär mit Limes V. Dann gilt $V = I_\vartheta^{-1/2} N + U$ in Verteilung für eine Zufallsvariable U, die unabhängig von N ist.*

Ist U verteilt nach δ_0, so ist $\hat{\vartheta}$ asymptotisch linear in ϑ mit Einflußfunktion $I_\vartheta^{-1}\dot{\ell}_\vartheta$.

Beweis. Der erste Teil folgt aus Satz 25 mit $V_n = J^{-1}S_n + V_n - J^{-1}S_n$. Ist U verteilt nach δ_0, so folgt aus Satz 25, daß $V_n - J^{-1}S_n \Rightarrow \delta_0$. Das ist der zweite Teil der Behauptung.

Wenn man sich auf asymptotisch lineare Schätzer beschränkt, braucht man den Faltungssatz nicht. Ist $\hat{\vartheta}$ asymptotisch linear in ϑ mit Einflußfunktion $h \in L_{2,0(P_\vartheta)}$ und regulär, dann gilt $h - I_\vartheta^{-1}\dot{\ell}_\vartheta \perp \dot{\ell}_\vartheta$ unter P_ϑ nach Satz 27. Weil $\hat{\vartheta}$ asymptotisch linear ist, ist $\hat{\vartheta}$ asymptotisch normal mit Varianz $\|h\|_2^2$. Es gilt mit Pythagoras:

$$\|h\|_2^2 = \|I_\vartheta^{-1}\dot{\ell}_\vartheta + h - I_\vartheta^{-1}\dot{\ell}_\vartheta\|_2^2 = I_\vartheta^{-1} + \|h - I_\vartheta^{-1}\dot{\ell}_\vartheta\|_2^2 \geq I_\vartheta^{-1}.$$

Insbesondere ist $\hat{\vartheta}$ nicht nur in allen symmetrischen, sondern sogar in *allen* 0 enthaltenden Intervallen asymptotisch höchstens wie $I_\vartheta^{-1/2}N$ konzentriert.

23 Effiziente Schätzer für nichtparametrische Familien

Sei $\mathcal{P}$ eine Familie von Wahrscheinlichkeitsmaßen auf $(\Omega, \mathcal{F})$. Sei $P \in \mathcal{P}$ fest. Seien $X_1, \ldots, X_n$ nach P verteilt.

Sei $U_0 \subset L_{2,0}(P)$. Für $u \in U_0$ existiere P_{nu} äquivalent zu P mit

$$\left\| \frac{dP_{nu}}{dP}^{1/2} - 1 - \frac{1}{2} n^{-1/2} u \right\|_2 = o(n^{-1/2}).$$

U_0 heißt *Tangentialraum* von $\mathcal{P}$ in P. Nach Satz 26 gilt LAN in P:

$$\log \frac{dP_{nu}^n}{dP^n} = n^{-1/2} \sum_{i=1}^n u(X_i) - \frac{1}{2}\|u\|_2^2 + o_p(1), \qquad (23.1)$$

$$n^{-1/2} \sum_{i=1}^n u(X_i) \Rightarrow \|u\|_2 N.$$

Definition. Ein Funktional $\tau : \mathcal{P} \to \mathbb{R}$ heißt *differenzierbar* in P mit *Gradient* g, wenn $g \in L_{2,0}(P)$ und

$$n^{1/2}(\tau(P_{nu}) - \tau(P)) \to Pgu, \quad u \in U_0.$$

Die Projektion g_0 von g in U_0 heißt *kanonischer* Gradient.

Das lokale Modell P_{n,tg_0}, $t \in \mathbb{R}$, heißt *ungünstigstes* eindimensionales lokales Modell durch P. Es ist g_0 die Richtung des steilsten Anstiegs von τ. Eine gegebene Änderung von τ in dieser Richtung bedeutet die geringste Veränderung von P, also die schlechteste statistische Unterscheidbarkeit.

Definition. Ein Schätzer T heißt *regulär* (für τ) in P mit *Limes* V, wenn

$$n^{1/2}(T - \tau(P_{nu})) \Rightarrow V \quad \text{unter } P_{nu}, \quad u \in U_0.$$

Definition. Ein Schätzer T heißt *asymptotisch linear* (für τ) in P mit *Einflußfunktion* h, wenn $h \in L_{2,0}$ und

$$n^{1/2}(T - \tau(P)) = n^{-1/2} \sum_{i=1}^{n} h(X_i) + o_p(1).$$

Ein solcher Schätzer ist asymptotisch normal mit Varianz $Eh^2(X) = \|h\|_2^2$.

Satz 29 *Sei $\mathcal{P}$ LAN in P, τ differenzierbar in P mit Gradient g und T asymptotisch linear mit Einflußfunktion h. Dann ist T regulär genau dann, wenn $h - g_0 \perp U_0$ unter P.*

Beweis. Sei $u \in U_0$. Es ist $(n^{-1/2} \sum_{i=1}^n u(X_i), n^{-1/2} \sum_{i=1}^n h(X_i))$ asymptotisch normal mit Kovarianz Puh. Nach (23.1) und Lemma 14 gilt

$$n^{1/2}(T - \tau(P_{nu})) \Rightarrow \|h\|_2 N + Puh - Pg_0u \quad \text{unter } P_{nu}, \quad u \in U_0.$$

Also ist T regulär genau dann, wenn $P(h - g_0)u = 0$ für alle $u \in U_0$.

Satz 30 *Sei $\mathcal{P}$ LAN in P, τ differnzierbar in P mit Gradient g und T regulär mit Limes V. Dann gilt $V = \|g_0\|_2 N + U$ in Verteilung für eine Zufallsvariable U, die unabhängig von V ist.*

Ist U verteilt nach δ_0, so ist T asymptotisch linear in P mit Einflußfunktion g_0.

Beweis. Wende Satz 28 für das ungünstigste eindimensionale lokale Modell P_{n,tg_0}, $t \in \mathbb{R}$, an.

Wir haben Regularität von T nur in der ungünstigsten Richtung gebraucht. Satz 28 sagt uns aber insbesondere, daß wir in Satz 30 keine bessere

(also kleinere) Varianzschranke als $\|g_0\|_2^2$ bekommen, wenn wir Regularität abschwächen zu

$$n^{1/2}(T - \tau(P_{n,tg_0})) \Rightarrow V \quad \text{unter } P_{n,tg_0}, \quad t \in \mathbb{R}.$$

Sei $\mathcal{P}$ die Familie *aller* Wahrscheinlichkeitsmaße auf $(\Omega, \mathcal{F})$, das *nichtparametrische Modell*. Für $u \in L_{2,0}(P)$ setze

$$\begin{aligned} \bar{u}_n &= u\mathbf{1}\Big(|u| < \frac{1}{2}n^{1/4}\Big), \\ u_n &= \bar{u}_n - P\bar{u}_n, \\ P_{nu}(dx) &= P(dx)(1 + n^{-1/2}u_n(x)). \end{aligned}$$

Das folgende Lemma beweisen wir nicht.

Lemma 16 *P_{nu} ist Hellinger-differenzierbar in P mit Ableitung u.*

Der Tangentialraum des nichtparametrischen Modells ist also $L_{2,0}(P)$.

Satz 31 *Sei $h \in L_2(P)$. Im nichtparametrischen Modell ist der empirische Schätzer $T = \frac{1}{n}\sum_{i=1}^n h(X_i)$ effizient für $\tau(P) = Ph$.*

Beweis. Mit Lemma 16 und Satz 26 gilt (23.1) für $u \in L_{2,0}(P)$. Nach Konstruktion von P_{nu} gilt

$$n^{1/2}(P_{nu}h - Ph) \to Puh = Pu(h - Ph), \quad u \in L_{2,0}(P).$$

Also ist $h - Ph$ der kanonische Gradient. Für den empirischen Schätzer gilt

$$n^{1/2}(T - Ph) = n^{-1/2}\sum_{i=1}^n (h(X_i) - Ph).$$

Also ist T (exakt) linear mit Einflußfunktion $h - Ph$, also effizient.

Für nichtparametrische Modelle wird der Effizienzbegriff trivial, wenn man sich auf *asymptotisch lineare* reguläre Schätzer beschränkt. Der Tangentialraum ist $L_{2,0}(P)$. Ist τ ein differenzierbares Funktional mit Gradient $g \in L_{2,0}(P)$, so ist g eindeutig bestimmt und kanonisch. Ist T ein Schätzer für τ, der regulär und asymptotisch linear mit Einflußfunktion h ist, so gilt $h = g$ nach Satz 29. Alle regulären asymptotisch linearen Schätzer haben also dieselbe Einflußfunktion, nämlich den kanonischen Gradienten, und sind deshalb effizient.