Inhalt

2 Klassische Schätz- und Testtheorie

Grundmodell:

Die Stichprobe $X = (X_1, \ldots, X_n)$ besitzt die Verteilung $\mathbb{P} \in \mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}, \Theta \subseteq \mathbb{R}^k$, wobei

- θ: k-dimensionaler Parameter
- Θ: Parameterraum
- $k < n$, oft $k \ll n$, mit $\dim(\theta) = k$ fest für asymptotische ($n \to \infty$)-Betrachtungen.
- In der Regel vorausgesetzt: Es existiert Dichte

$$f(x|\theta) = f(x_1, , \ldots, x_n|\theta) \text{ zu } \mathbb{P}_\theta,$$

so dass man analog schreiben kann:

$$\mathcal{P} = \{f(x|\theta) : \ \theta \in \Theta\}.$$

2 Klassische Schätz- und Testtheorie

- Klassische Schätz- und Testtheorie für finite (d.h. für festen Stichprobenumfang n) i.i.d.-Stichprobe von besonderer Relevanz; es gilt:

$$f(x|\theta) = f(x_1|\theta) \cdot \ldots \cdot f(x_n|\theta).$$

- Viele Begriffe, insbesondere der Schätztheorie, jedoch von genereller Bedeutung.
- Literatur: Lehmann & Casella (1998), Lehmann & Romano (2005), Rüger (1999, 2002) Band I+II

Definition 2.1 (Statistik)

Eine Statistik ist eine messbare Funktion

$$T : \begin{cases} \mathcal{X} & \longrightarrow & \mathbb{R}^l \\ x & \longmapsto & T(x). \end{cases}$$

Normalerweise ist $l < n$, da mit der Statistik T eine Dimensionsreduktion erzielt werden soll.

Beispiel 2.1

$\longrightarrow$ $T(x)$ Schätzfunktion

$\longrightarrow$ $T(x)$ Teststatistik

Gesucht: Punkt- oder Bereichsschätzung für θ oder einen transformierten Parametervektor $\tau(\theta)$.

Beispiel 2.2
$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ mit $\theta = (\mu, \sigma^2)^\top$. Hier könnte $\tau(\theta) = \mu$ sein (d.h. σ^2 ist Nuisance-Parameter) oder $\tau(\theta) = 1/\sigma^2$ (d.h. die Präzision ist von Interesse).

Definition 2.2 (Punktschätzung, Schätzer, Schätzfunktion)

Sei

$$T : \begin{cases} \mathcal{X} & \longrightarrow & \Theta \subseteq \mathbb{R}^k \\ x & \longmapsto & T(x) \end{cases}$$

eine messbare Abbildung.
Man bezeichnet mit $T(x)$ den Schätzwert oder die Punktschätzung (zu konkreter Realisation x) und mit $T(X)$ den Punktschätzer von θ, der eine Zufallsvariable ist (auch gebräuchlich: $\widehat{\theta}(x)$ oder kurz $\widehat{\theta}$, d.h. notationell wird nicht zwischen Schätzwert und Schätzfunktion unterschieden).

Der Begriff der Suffizienz ist von grundlegender Bedeutung in der klassischen parametrischen Inferenz; darüber hinaus ist die Bedeutung (stark) abgeschwächt, vgl. auch Statistik IV.

Definition 2.3

Eine Statistik T heißt suffizient für θ (oder auch für $\mathcal{P}$) $\overset{def}{\Leftrightarrow}$ die bedingte Verteilung bzw. Dichte von X gegeben $T(x) = t$ ist für alle Werte von $T(x) = t$ von θ unabhängig, d.h.

$$f_{X|T}(x|T(x) = t, \theta) = f_{X|T}(x|T(x) = t)$$

hängt nicht von θ ab.

2.1 Klassische Schätztheorie

2.1.1 Suffizienz

Idee: Zusätzliche Information in X, die nicht in T enthalten ist, ist durch $f_{X|T}$ gegeben. Falls $f_{X|T}$ von θ unabhängig ist, dann enthält die Stichprobe x nicht mehr Information über θ als $T(x)$.

Im Folgenden nehmen wir die Existenz einer Dichte für X an. Das Kriterium im folgenden Satz ist äquivalent und konstruktiv:

Satz 2.4 (Faktorisierungssatz, Neyman-Kriterium)

Eine Statistik T ist suffizient für θ genau dann wenn

$$f(x|\theta) = h(x)g(T(x)|\theta)$$

für fast alle x.

D.h. die Dichte lässt sich in zwei Teile faktorisieren, von denen ein Teil von x, aber nicht von θ, und der andere nur von θ und $T(x)$ abhängt.

Beweis.

„⇒": Falls T suffizient ist, gilt:

$$f_{X|T}(x|T(x) = t, \theta) = f_{X|T}(x|T(x) = t) = \frac{f_{X,T}(x, t|\theta)}{f_{T|\theta}(t|\theta)}.$$

Weiterhin ist

$$f_{X,T}(x, t|\theta) = \begin{cases} f_{X|\theta}(x|\theta) & \text{für } T(x) = t \\ 0 & \text{sonst,} \end{cases}$$

d.h. insgesamt

$$\underbrace{f_{X|T}(x|t)}_{h(x)} \cdot \underbrace{f_{T|\theta}(t|\theta)}_{g(T(x)|\theta)} = f_{X|\theta}(x|\theta).$$

Beweis fortgeführt

„$\Leftarrow$": Man erhält die Dichte von T, ausgewertet an t, indem man im obigen Faktorisierungskriterium über die x, für die $T(x) = t$ gilt, summiert (bzw. integriert). Im diskreten Fall also:

$$f_{T|\theta}(t|\theta) = \sum_{x:\,T(x)=t} h(x)g(T(x)|\theta) = g(t|\theta) \sum_{x:\,T(x)=t} h(x).$$

Damit ist die bedingte Dichte von X gegeben $T = t$,

$$\frac{f_{X|\theta}(x|\theta)}{f_{T|\theta}(t|\theta)} = \frac{h(x)g(T(x)|\theta)}{\sum_{x:\,T(x)=t} h(x)g(t|\theta)} = \frac{h(x)}{\sum_{x:\,T(x)=t} h(x)},$$

unabhängig von θ. Im stetigen Fall werden Summen durch Integrale ersetzt; im Detail werden Messbarkeitsbedingungen verwendet.

Beispiel 2.3 (Bernoulli-Experiment)

Seien $X_1, \ldots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bin}(1, \pi)$ und $Z = \sum_{i=1}^n X_i$ die Anzahl der Erfolge.
Dann ist Z suffizient für π, denn

$$\begin{aligned} f_{X|Z}(x|z, \pi) &= \mathbb{P}_\pi(X = x | Z = z) \\ &= \frac{\prod_{i=1}^n \pi^{x_i}(1-\pi)^{1-x_i}}{\binom{n}{z}\pi^z(1-\pi)^{n-z}}, \quad \text{wobei } \sum_{i=1}^n x_i = z \\ &= \binom{n}{z}^{-1} \end{aligned}$$

ist unabhängig von π.

Beispiel 2.3 (Bernoulli-Experiment) fortgeführt

Gemäß Faktorisierungssatz ist

$$f(x|\pi) = \underbrace{\frac{1}{\binom{n}{z}}}_{=h(x)} \underbrace{\binom{n}{z}\pi^z(1-\pi)^{n-z}}_{=g(z|\pi)} = \underbrace{1}_{=h^*(x)} \underbrace{\pi^z(1-\pi)^{n-z}}_{=g^*(z|\pi)}.$$

Beispiel 2.4 (Normalverteilung)

Sei $X = (X_1, \dots, X_n)$ mit $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ und $\theta = (\mu, \sigma^2)^\top$.

$$f_{X|\theta}(x|\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right)$$

$$= \underbrace{(2\pi)^{-n/2}}_{h(x)} \underbrace{(\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right)\right)}_{g((\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)|\theta)},$$

Beispiel 2.4 (Normalverteilung) fortgeführt

d.h. $T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ ist suffizient für $\theta = (\mu, \sigma^2)^\top$.
Aber: Die bijektive Transformation $\widetilde{T}(X) = (\bar{X}, S^2)$ ist auch suffizient für θ, wobei S^2 die Stichprobenvarianz bezeichnet.

Beispiel 2.5 (Exponentialverteilung)

Sei $X = (X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} \operatorname{Exp}(\lambda)$, dann

$$f(x|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \underbrace{1}_{h(x)} \cdot \underbrace{\lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)}_{g(T(x)|\lambda)}$$

mit $T(x) = \sum_{i=1}^{n} x_i$. Nach der ursprünglichen Definition ist

$$\frac{f_{X,T|\lambda}(x,t|\lambda)}{f_{T|\lambda}(t|\lambda)} = \frac{\lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)}{\frac{\lambda^n}{\Gamma(n)} \left(\sum_{i=1}^{n} x_i\right)^{n-1} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)} = \frac{\Gamma(n)}{\left(\sum_{i=1}^{n} x_i\right)^{n-1}},$$

unabhängig von λ. Dabei wird benutzt, dass die Summe von n unabhängigen und identisch exponentialverteilten Zufallsvariablen mit Parameter λ gammaverteilt ist mit Parametern n und λ.

Beispiel 2.6 (Ordnungsstatistik)
Sei $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x|\theta)$ (wobei f stetige Dichte ist) und $T(X) = X_{(\cdot)} = (X_{(1)}, \ldots, X_{(n)})$ die Ordnungsstatistik.

Dann gilt

$$f_{X|T,\theta}(x|T = x_{(\cdot)}, \theta) = \frac{1}{n!}\,.$$

Die Gleichheit folgt aus der Stetigkeit, denn $x_i \neq x_j \; \forall i \neq j$ (mit Wahrscheinlichkeit 1). $X_{(\cdot)}$ ist suffizient für θ. Wir haben also bei i.i.d.-Beobachtungen keinen Informationsverlust durch Ordnen der Daten.

Bemerkung.

- Offensichtlich ist $T(X) = X$, d.h. die Stichprobe selbst, suffizient.
- Ebenso ist jede eineindeutige Transformation von X oder von einer suffizienten Statistik $T(X)$ suffizient.
- Ist T suffizient, dann auch (T, T^*), wobei T^* eine beliebige weitere Statistik darstellt.

Dies zeigt: Die Dimension einer suffizienten Statistik sollte soweit wie möglich reduziert werden.

Definition 2.5 (Minimalsuffizienz)

Eine Statistik T heißt minimalsuffizient für $\theta \overset{def}{\Leftrightarrow}$ T ist suffizient, und zu jeder anderen suffizienten Statistik V existiert eine Funktion H mit

$$T(X) = H(V(X)) \;\; \mathcal{P} - \textit{fast überall.}$$

Frage: Existieren minimalsuffiziente Statistiken? Wenn ja, sind sie eindeutig?

Beispiel 2.7 (Normalverteilung)

Seien $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

1. $T(X) = \bar{X}$ ist minimalsuffizient für μ bei bekanntem σ^2.
2. $T(X) = \sum_{i=1}^n (X_i - \mu)^2$ ist minimalsuffizient für σ^2 bei bekanntem μ.
3. $T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ ist minimalsuffizient für μ und σ^2.
4. $T(X) = |X|$ ist minimalsuffizienz für σ^2 wenn $X \sim N(0, \sigma^2)$ $(n = 1, \mu = 0)$. X ist ebenfalls suffizient, aber nicht minimalsuffizient (trotz gleicher Dimension).

Lemma 2.6

Sind T und S minimalsuffiziente Statistiken, dann existieren injektive Funktionen g_1, g_2, so dass $T = g_1(S)$ und $S = g_2(T)$.

Satz 2.7 (Charakterisierung von Minimalsuffizienz durch Likelihood-Quotienten)

Definiere den Likelihood-Quotienten

$$\Lambda_x(\theta_1, \theta_2) = \frac{f(x|\theta_1)}{f(x|\theta_2)}.$$

Eine notwendige und hinreichende Bedingung für die Minimalsuffizienz einer Statistik T für θ ist, dass für alle θ_1 und θ_2 mit $\theta_1 \neq \theta_2$ gilt:

$$T(x) = T(x') \;\Leftrightarrow\; \Lambda_x(\theta_1, \theta_2) = \Lambda_{x'}(\theta_1, \theta_2).$$

Beispiel 2.8 (Suffizienz in Exponentialfamilien)

Die Dichte einer k-parametrischen Exponentialfamilie hat die Form

$$\begin{aligned} f(x|\theta) &= h(x) \cdot c(\boldsymbol{\theta}) \cdot \exp(\gamma_1(\boldsymbol{\theta}) T_1(x) + \ldots + \gamma_k(\boldsymbol{\theta}) T_k(x)) \\ &= h(x) \cdot \exp(b(\boldsymbol{\theta}) + \boldsymbol{\gamma}(\boldsymbol{\theta})^\top \mathbf{T}(x)), \end{aligned}$$

d.h. $\mathbf{T}(X) = (T_1(X), \ldots, T_k(X))^\top$ ist suffizient für $\boldsymbol{\theta}$ nach Faktorisierungssatz. Falls Θ ein offenes Rechteck in $\mathbb{R}^k$ enthält, ist $\mathbf{T}$ auch minimalsuffizient.

Es folgt nun die Charakterisierung der Minimalsuffizienz nach Lehmann-Scheffé. Dazu wird der Begriff der Vollständigkeit benötigt.

Definition 2.8

Eine Statistik T ist vollständig $\overset{def}{\Leftrightarrow}$ für jede reelle (messbare) Funktion g gilt:

$$\mathbb{E}_\theta[g(T)] = 0 \;\; \forall\theta \;\Rightarrow\; \mathbb{P}_\theta(g(T) = 0) = 1 \; \forall\theta.$$

Beispiel Vollständigkeit
Sei $X = (X_1, \ldots, X_n) \sim Po(\lambda)$ und $T = \sum_{i=1}^n X_i$. Nach Beispiel 2.8 ist T suffizient.
Sei h messbar. Da $T = \sum_i X_i \sim Po(n\lambda)$, gilt für

$$E_\lambda[h(T)] = \sum_{t=0}^{\infty} h(t) \frac{(n\lambda)^t}{t!} \exp(-n\lambda) = \exp(-n\lambda) \sum_{t=0}^{\infty} \underbrace{h(t) \frac{n^t}{t!}}_{c_t} \lambda^t = 0$$

$$\Rightarrow \sum_{t=0}^{\infty} c_t \lambda^t = \sum_{t=0}^{\infty} 0 \lambda^t$$

(zwei konvergierende Potenzreihen). Daraus folgt $c_t = 0$ für alle t und wegen $\frac{n^t}{t!} \neq 0$ also $h(t) = 0$ für alle t. Damit ist T vollständig (und mit dem folgenden Satz auch minimalsuffizient).

Aus der Definiton wird nicht unmittelbar klar, warum „Vollständigkeit" eine wünschenswerte Eigenschaft eines Schätzers sein sollte. Einen möglichen Grund liefert der folgende Satz.

Satz 2.9 (Lehmann-Scheffé)

Angenommen, X besitzt eine Dichte $f(x|\theta)$ und $T(X)$ ist suffizient und vollständig für θ. Dann ist $T(X)$ minimalsuffizient für θ.

Beweis des Satzes von Lehmann-Scheffé:

Vorausgesetzt wird, dass eine minimalsuffiziente Statistik existiert - bewiesen wurde dies von Lehmann und Scheffé (1950). Ist dies der Fall, so ist diese bis auf bijektive Transformationen eindeutig. Bezeichne $S = g_1(T)$ eine solche minimalsuffiziente Statistik für eine Funktion g_1. Definiere $g_2(S) = \mathbb{E}[T|S]$. Da S suffizient für θ ist, hängt $g_2(S)$ nicht von θ ab. Betrachte nun

$$g(T) = T - g_2(S) = T - g_2(g_1(T)).$$

Anwendung des Satzes von der iterierten Erwartung liefert:

$$\mathbb{E}_\theta[g(T)] = \mathbb{E}_\theta[T] - \mathbb{E}_\theta[\mathbb{E}[T|S]] = 0.$$

Da T vollständig ist, ist $g(T) = 0$ bzw. $g_2(S) = T$ mit Wahrscheinlichkeit 1, d.h. T ist eine Funktion von S. Wähle x und x', so dass $g_2(S(x)) = g_2(S(x'))$, dann ist $T(x) = T(x')$ und $S(x) = g_1(T(x)) = g_1(T(x')) = S(x')$, d.h. g_1 und g_2 sind injektiv, so dass T und S äquivalent. Somit ist auch T minimalsuffizient.

Bemerkung (Ancillarity einer Statistik)
Eine Statistik $V(X)$ heißt *ancillary* („Hilfsstatistik", „ankillar") für $\mathcal{P}$, wenn ihre Verteilung nicht von θ abhängt (also bekannt ist).
Häufiger Sachverhalt: $T = (U, V)$ ist suffizient für θ, V ancillary, U nicht suffizient.

Beispiel 2.9 (Ancillary Statistik)
$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} U\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$.
Man kann dann zeigen (Davison, 2004, Ex. 12.3), dass mit

$$\begin{aligned} U &= \frac{1}{2}(X_{(1)} + X_{(n)}) \\ V &= X_{(n)} - X_{(1)} \end{aligned}$$

$T = (U, V)$ suffizient, aber nicht vollständig für θ ist. Ferner ist U alleine nicht suffizient und V ancillary.

- Fehler eines Schätzers $\widehat{\theta} = \widehat{\theta}(X)$ ist $\widehat{\theta} - \theta$.
- Messung des Fehlers durch Verlustfunktion, zum Beispiel

$$
\begin{aligned}
L(\widehat{\theta}, \theta) &= |\widehat{\theta} - \theta| && \text{Abstand } (\theta \text{ skalar}),\\
L(\widehat{\theta}, \theta) &= \|\widehat{\theta} - \theta\|^2 && \text{quadratischer Fehler},\\
L(\widehat{\theta}, \theta) &= \frac{\|\widehat{\theta} - \theta\|^2}{\|\theta\|^2} && \text{relativer quadratischer Fehler},\\
L(\widehat{\theta}, \theta) &= (\widehat{\theta} - \theta)^\top \mathbf{D}(\widehat{\theta} - \theta) && \text{gewichteter quadratischer Fehler}\\
& && (\mathbf{D} \text{ positiv definit}).
\end{aligned}
$$

- Risikofunktion $R(\widehat{\theta}, \theta) = \mathbb{E}_\theta[L(\widehat{\theta}, \theta)]$.
- Hier wird (hauptsächlich) quadratischer Verlust betrachtet.

Definition 2.10 (Erwartungstreue, Bias, Varianz eines Schätzers)

- $\widehat{\theta}$ *heißt erwartungstreu* $\overset{def}{\Leftrightarrow}$ $\mathbb{E}_\theta[\widehat{\theta}] = \theta$.
- $\mathrm{Bias}_\theta(\widehat{\theta}) = \mathbb{E}_\theta[\widehat{\theta}] - \theta$.
- $\mathrm{Var}_\theta(\widehat{\theta}) = \mathbb{E}_\theta[(\widehat{\theta} - \mathbb{E}_\theta[\widehat{\theta}])^2]$, θ *skalar.*

Definition 2.11 (MSE)

Der mittlere quadratische Fehler (mean squared error) ist definiert als

$$\mathrm{MSE}_\theta(\widehat{\theta}) = \mathbb{E}_\theta[(\widehat{\theta} - \theta)^2] = \mathrm{Var}_\theta(\widehat{\theta}) + (\mathrm{Bias}_\theta(\widehat{\theta}))^2.$$

Der Gesamtfehler lässt sich also aufteilen in einen zufälligen Fehler (Varianz) und einen systematischen (quadrierter Bias).

Vergleicht man zwei Schätzer bezüglich ihres MSE, kann für einen Teilbereich von Θ der MSE des einen, für andere Teilbereiche der MSE des zweiten Schätzers kleiner sein:

Beispiel 2.10 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} B(1, \pi)$.

1. MSE von $\widehat{\pi} = \bar{X}$:

$$\mathbb{E}_\pi[(\bar{X} - \pi)^2] = \mathrm{Var}_\pi(\bar{X}) = \frac{\pi(1-\pi)}{n}.$$

Beispiel 2.10 fortgeführt

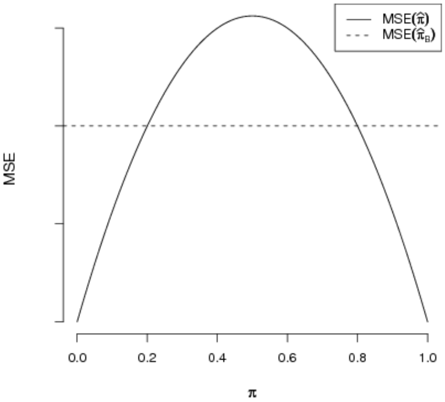
2. MSE des Bayes-Schätzers (Posteriori-Erwartungswert) bei einer Priori $p(\pi) \sim \text{Be}(\alpha, \beta)$:

$$\widehat{\pi}_B = \frac{Y + \alpha}{\alpha + \beta + n}, \quad Y = \sum_{i=1}^{n} X_i,$$

$$\begin{aligned} \text{MSE}(\widehat{\pi}_B) &= \text{Var}_\pi \left(\frac{Y + \alpha}{\alpha + \beta + n} \right) + \left(\mathbb{E}_\pi \left[\frac{Y + \alpha}{\alpha + \beta + n} - \pi \right] \right)^2 \\ &= \frac{n\pi(1 - \pi)}{(\alpha + \beta + n)^2} + \left(\frac{n\pi + \alpha}{\alpha + \beta + n} - \pi \right)^2 . \end{aligned}$$

Für $\alpha = \beta = \sqrt{n}/2$ ergibt sich

$$\mathrm{MSE}_\pi(\widehat{\pi}_B) = \mathbb{E}_\pi[(\widehat{\pi}_B - \pi)^2] = \frac{1}{4}\frac{n}{(n+\sqrt{n})^2} = \text{const bezüglich } \pi.$$



Fazit: In der Regel wird man keinen „MSE-optimalen" Schätzer $\widehat{\theta}^{\mathsf{opt}}$ finden in dem Sinne, dass $\mathrm{MSE}_\theta(\widehat{\theta}^{\mathsf{opt}}) \leq \mathrm{MSE}_\theta(\widehat{\theta})$ für alle θ und alle konkurrierenden $\widehat{\theta}$. Bei Einschränkung auf erwartungstreue Schätzer ist dies öfter möglich.
Deshalb die Forderung:

Definition 2.12 (zulässiger („admissible") Schätzer)

Ein Schätzer $\widehat{\theta}$ heißt zulässig $\overset{def}{\Leftrightarrow}$ es gibt keinen Schätzer $\widetilde{\theta}$ mit $\mathrm{MSE}_\theta(\widetilde{\theta}) \leq \mathrm{MSE}_\theta(\widehat{\theta})$ für alle θ und $\mathrm{MSE}_\theta(\widetilde{\theta}) < \mathrm{MSE}_\theta(\widehat{\theta})$ für mindestens ein θ, d.h. es gibt keinen Schätzer $\tilde{\theta}$, der $\hat{\theta}$ gleichmäßig/strikt „dominiert".

Definition 2.13 (Verallgemeinerungen des MSE auf $\theta \in \mathbb{R}^p$, $p > 1$)

Üblich sind die folgenden zwei Alternativen:

1. *MSE (skalar):*

$$
\begin{aligned}
\mathrm{MSE}_\theta^{(1)}(\widehat{\theta}) &= \mathbb{E}_\theta[\|\widehat{\theta} - \theta\|^2] \\
&= \sum_{j=1}^{p} \mathbb{E}_\theta[(\widehat{\theta}_j - \theta_j)^2] \\
&= \sum_{j=1}^{p} \mathrm{MSE}_\theta(\widehat{\theta}_j)
\end{aligned}
$$

2. *MSE-Matrix:*

$$
\begin{aligned}
\mathrm{MSE}_\theta^{(2)}(\widehat{\theta}) &= \mathbb{E}_\theta[(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^\top] \\
&= \mathrm{Cov}_\theta(\widehat{\theta}) + (\mathbb{E}_\theta[\widehat{\theta}] - \theta)(\mathbb{E}_\theta[\widehat{\theta}] - \theta)^\top
\end{aligned}
$$

Diese Variante wird häufig bei linearen Modellen betrachtet.

Bemerkung. Das j-te Diagonalelement der MSE-Matrix ist $\mathrm{MSE}_\theta(\widehat{\theta}_j)$. Vergleich von MSE-Matrizen gemäß *„Löwner"-Ordnung*:

$$\mathrm{MSE}_\theta(\widetilde{\theta}) \overset{(\leq)}{<} \mathrm{MSE}_\theta(\widehat{\theta})$$

bedeutet, dass die Differenz $\mathrm{MSE}_\theta(\widehat{\theta}) - \mathrm{MSE}_\theta(\widetilde{\theta})$ positiv (semi-)definit ist.
Man definiert allgemein für symmetrische $(p \times p)$-Matrizen $\mathbf{A}, \mathbf{B}$:

$$\begin{aligned} \mathbf{A} \leq \mathbf{B} \quad &\overset{\text{def}}{\Leftrightarrow} \quad \mathbf{B} - \mathbf{A} \text{ ist positiv semidefinit,} \\ \mathbf{A} < \mathbf{B} \quad &\overset{\text{def}}{\Leftrightarrow} \quad \mathbf{B} - \mathbf{A} \text{ ist positiv definit.} \end{aligned}$$

Beispiel 2.11 (Gauß-Experiment) Seien $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$.

- σ^2 bekannt, μ unbekannt: MSE-Vergleich von $\bar{X}$ und $T = a\bar{X} + b$.
- σ^2 unbekannt, μ bekannt:
  - Eine Möglichkeit:

$$S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2,\ \mathbb{E}_{\sigma^2}(S_\mu^2) = \sigma^2$$

  - Weitere Möglichkeit:

$$V_\mu^2 = \frac{1}{n+2}\sum_{i=1}^{n}(X_i - \mu)^2, \mathbb{E}_{\sigma^2}(V_\mu^2) = \frac{n}{n+2}\sigma^2$$

Es stellt sich heraus, dass $\mathrm{MSE}_{\sigma^2}(V_\mu^2) < \mathrm{MSE}_{\sigma^2}(S_\mu^2)$ ist.

Beispiel 2.11 (Gauß-Experiment) fortgeführt

- μ und σ^2 unbekannt:
  - Eine Möglichkeit:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

$$\mathbb{E}_{\sigma^2}(S^2) = \sigma^2, \;\; \mathrm{MSE}_{\sigma^2}(S^2) = \mathrm{Var}_{\sigma^2}(S^2) = \frac{2}{n-1}\, \sigma^4.$$

  - Weitere Möglichkeit:

$$V^2 = \frac{1}{n+1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

$$\mathbb{E}_{\sigma^2}(V^2) = \frac{n-1}{n+1}\, \sigma^2, \; \mathrm{MSE}_{\sigma^2}(V^2) = \frac{2}{n+1}\, \sigma^4,$$

d.h. V^2 dominiert S^2.

Beispiel 2.11 (Gauß-Experiment) fortgeführt

- μ und σ^2 (weiterhin) unbekannt:
  - Der sogenannte *Stein-Schätzer*

$$T = \min\left\{V^2, \frac{1}{n+2}\sum_{i=1}^{n} X_i^2\right\}$$

dominiert V^2 (und damit S^2).

Plausibilitätsbetrachtung: Ist $\mu = 0$, so ist $\sum_{i=1}^{n} X_i^2/(n+2)$ besserer Schätzer als V^2. Ist $\mu \neq 0$, so ist V^2 ein besserer Schätzer als $\sum_{i=1}^{n} X_i^2/(n+2)$. Beim Stein-Schätzer wird fallweise mit hoher Wahrscheinlichkeit der jeweils bessere Schätzer benutzt.

Beispiel 2.12 (Steins Paradoxon)
Seien $(X_1, \ldots, X_m)^\top \sim N_m(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_m)$ multivariat normalverteilt mit $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top$ und σ^2 bekannt. Es sollen simultan die Erwartungswerte $\mu_1, \ldots, \mu_m$ geschätzt werden. Man beachte dabei, dass die einzelnen Komponenten als unabhängig angenommen werden. Die Stichprobe hat die Form

$$X_{11}, \ldots, X_{1n_1}, \ldots, X_{m1}, \ldots, X_{mn_m}$$

(i.i.d. Stichproben aus „Gruppen" $1, \ldots, m$). Übliche Schätzer:

$$T_j = \bar{X}_j, \quad j = 1, \ldots, m, \qquad \mathbf{T} = (T_1, \ldots, T_m)^\top = (\bar{X}_1, \ldots, \bar{X}_m)^\top.$$

Der (skalare) MSE ist:

$$\mathbb{E}_\mu[\|\mathbf{T} - \boldsymbol{\mu}\|^2] = \sum_{j=1}^m \mathbb{E}_\mu[(\bar{X}_j - \mu_j)^2] = \sum_{j=1}^m \frac{\sigma_j^2}{n_j}.$$

Beispiel 2.12 (Steins Paradoxon) fortgeführt
Paradoxerweise gilt:

1. Für $m \leq 2$ ist $\mathbf{T}$ zulässig.
2. Für $m \geq 3$ ist $\mathbf{T}$ *nicht* zulässig und wird dominiert durch den Stein-Schätzer

$$\mathbf{T}^* = \left(1 - \frac{(m-2)\sigma^2}{\sum_{j=1}^{m} n_j \bar{X}_j^2}\right)\mathbf{T}.$$

Es lässt sich zeigen: T^* ist selbst unzulässig. Der Stein-Schätzer ist ein sogenannter *Shrinkage-Schätzer*.

Beispiel 2.13 (Lineares Modell)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \;\; \boldsymbol{\varepsilon} \;\sim\; (N)(\mathbf{0}, \sigma^2 \mathbf{I})$$

KQ-Schätzer: $\widehat{\boldsymbol{\beta}}_{\text{KQ}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$

Ridge-Schätzer: $\widehat{\boldsymbol{\beta}}_{\text{Ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{D})^{-1} \mathbf{X}^\top \mathbf{y}$,

wobei $\mathbf{D}$ eine Diagonalmatrix mit positiven Diagonalelementen ist.
Für einen MSE-Vergleich siehe Übung.

Fazit: Bereits im einfachen Beispiel der Schätzung von π in $B(1, \pi)$ (siehe Beispiel 2.10) zeigt sich, dass es im Allgemeinen keine MSE-optimalen Schätzer gibt.

Auswege:

1. Einschränkung auf Teilklasse von Schätzern, zum Beispiel erwartungstreue (und lineare) Schätzer, äquivariante Schätzer, ...
2. MSE-Kriterium verändern:
   - Ersetze $\mathrm{MSE}_\theta(\widehat{\theta})$ durch Minimierung von $\max_{\theta \in \Theta} \mathrm{MSE}_\theta(\widehat{\theta})$ (Minimax-Kriterium)
   - oder ersetzte $\mathrm{MSE}_\theta(\widehat{\theta})$ durch $\mathbb{E}_{p(\theta)}[\mathrm{MSE}_\theta(\widehat{\theta})]$ bei einer Priori-Verteilung $p(\theta)$ (Bayes-Schätzer).

Hier: Strategie 1 mit erwartungstreuen Schätzern, vgl. 2.1.4.

Definition 2.14 (Fisher-reguläre Verteilungsfamilien)

Eine Familie von Verteilungen $\mathcal{P}_\theta$ *mit Dichte* $f(x|\theta) = f(x_1, \ldots, x_n|\theta)$, $\theta \in \Theta$, *heißt Fisher-regulär, wenn Folgendes gilt:*

1. *Der Träger* $\{x \in \mathcal{X} : f(x|\theta) > 0\}$ *ist unabhängig von* θ *(dies ist zum Beispiel bei* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} U[0;\theta]$ *oder bei der Pareto-Verteilung verletzt).*
2. Θ *ist offen in* $\mathbb{R}^p$ *(verletzt zum Beispiel bei* $\sigma^2 \geq 0$*).*
3. *Die ersten und zweiten Ableitungen von* $f(x|\theta)$ *bzgl.* θ *existieren und sind für jedes* θ *endliche Funktionen von* x.
4. *Vertauschbarkeit: Sowohl für* $f(x|\theta)$ *als auch für* $\log(f(x|\theta))$ *kann erstes und zweites Differenzieren nach* θ *und Integration über* x *vertauscht werden.*

Definition 2.15 (Log-Likelihood, Scorefunktion und Information)

$\ell(\theta; x) = \log f(x|\theta)$ *(Log-Likelihood von θ bzgl. der Stichprobe x)*

$$s(\theta; x) = \frac{\partial}{\partial \theta}\ell(\theta; x) = \left(\frac{\partial}{\partial \theta_1}\ell(\theta; x), \dots, \frac{\partial}{\partial \theta_p}\ell(\theta; x)\right)^\top$$

(Score-Funktion)

$J(\theta; x) = -\dfrac{\partial^2 \ell(\theta; x)}{\partial \theta \partial \theta^\top}$ *(beobachtete Informationsmatrix der Stichprobe mit Elementen* $(J(\theta; x))_{ij} = -\dfrac{\partial^2 \log f(x|\theta)}{\partial \theta_i \partial \theta_j}$*)*

$I(\theta) = \mathrm{E}_\theta[J(\theta; X)]$ *(erwartete oder Fisher-Informationsmatrix)*

Satz 2.16

Ist $\mathcal{P}_\theta$ Fisher-regulär, so gilt:

1. $\mathbb{E}_\theta\left[s(\theta;X)\right] = 0$
2. $\mathbb{E}_\theta\left[-\frac{\partial^2 \ell(\theta;X)}{\partial\theta\partial\theta^\top}\right] = I(\theta) = \mathrm{Cov}_\theta(s(\theta;X))$

Beweis

Zu 1.:

$$\begin{aligned}
\mathbb{E}_\theta[s(\theta; X)] &= \int s(\theta; x) f(x|\theta)\, dx \\
&= \int \frac{\partial}{\partial\theta} \log(f(x|\theta)) f(x|\theta)\, dx \\
&= \int \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)} f(x|\theta) dx \\
&= \frac{\partial}{\partial\theta} \int f(x|\theta) dx = 0
\end{aligned}$$

Zu 2.:

$$\mathbb{E}_\theta\left[-\frac{\partial^2 \ell(\theta;X)}{\partial\theta\partial\theta^\top}\right] = -\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta^\top}f(X|\theta)}{f(X|\theta)}\right)\right]$$

$$= -\mathbb{E}_\theta\left[\frac{f(X|\theta)\frac{\partial^2}{\partial\theta\partial\theta^\top}f(X|\theta) - (\frac{\partial}{\partial\theta}f(X|\theta))(\frac{\partial}{\partial\theta^\top}f(X|\theta))}{f(X|\theta)^2}\right]$$

unter Verwendung der Quotientenregel der Differentiation. Dies ist

$$\begin{aligned} &= - \quad \mathbb{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta\partial\theta^\top}f(X|\theta)}{f(X|\theta)}\right] + \mathbb{E}_\theta\left[\frac{\frac{\partial}{\partial\theta}f(X|\theta)}{f(X|\theta)}\cdot\frac{\frac{\partial f(X|\theta)}{\partial\theta^\top}}{f(X|\theta)}\right] \\ &= - \quad \int \frac{\partial^2}{\partial\theta\partial\theta^\top}f(x|\theta)dx + \mathbb{E}_\theta[s(\theta;X)s(\theta;X)^\top] \end{aligned}$$

Zu 2. (fortgeführt):
Der erste Summand ist unter Vertauschung von Differentiation und Integration gleich null. Für den zweiten Teil ergibt sich mit Teil 1.

$$\mathbb{E}[s(\theta; X)s(\theta; X)^\top] = \mathrm{Cov}_\theta(s(\theta; X)).$$

$\square$

Weitere Eigenschaften:

- Sind $X_1, \dots, X_n$ unabhängig und gemäß $X_i \sim f_i(x|\theta)$, $i = 1, \dots, n$, verteilt, so gilt:

$$\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta) \quad , \quad \ell_i(\theta) = \log f_i(x_i|\theta)$$

$$s(\theta) = \sum_{i=1}^{n} s_i(\theta) \quad , \quad s_i(\theta) = \frac{\partial}{\partial\theta} \log f_i(x_i|\theta)$$

$$J(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^\top} = \sum_{i=1}^{n} -\frac{\partial^2 \log f_i(x_i|\theta)}{\partial\theta\partial\theta^\top}$$

Weitere Eigenschaften (fortgeführt):

- Für $X_1, \ldots, X_n$ i.i.d. wie $X_1 \sim f_1(x|\theta)$ folgt

  $$I(\theta) = \mathbb{E}_\theta[J(\theta)] = n \cdot i(\theta),$$

  wobei

  $$i(\theta) = \mathbb{E}_\theta\left[-\frac{\partial^2 \ell_1(\theta; X)}{\partial\theta\partial\theta^\top}\right] = \mathrm{Cov}_\theta\left(\frac{\partial \log f_1(X|\theta)}{\partial\theta}\right)$$

  die erwartete Information einer Einzelbeobachtung ist, d.h. die erwartete Informationsmatrix der Stichprobe $X_1, \ldots, X_n$ ist die n-fache erwartete Information einer (typischen) Stichprobenvariable X_1.

Weitere Eigenschaften (fortgeführt):

- Für eine Statistik $T = T(X), X = (X_1, \ldots, X_n)^\top$ mit $T \sim f_T(t|\theta)$ kann man die Begriffe Score-Funktion und Fisher-Information völlig analog definieren. Insbesondere ist

$$I_T(\theta) = \mathbb{E}_\theta \left[-\frac{\partial^2 \log f_T(t|\theta)}{\partial\theta\partial\theta^\top} \right].$$

Satz 2.17 (Suffizienz und Fisher-Information)

Sei $I(\theta)$ die Fisher-Information für X. Dann gilt unter Fisher-Regularität für jede Statistik T:

1. $I_T(\theta) \leq I(\theta)$.
2. $I_T(\theta) = I(\theta) \;\Leftrightarrow\;$ *T ist suffizient für θ.*

Also: Bei einer suffizienten Statistik T wird keine (erwartete) Information „verschenkt".