week10b: Kapitel 5.3: Effizienz und Konsistenz, Teil 2
Beweis der Cramer-Rao Abschätzung und Folgerungen

In der letzten Veranstaltung haben wir das folgende Setting betrachtet: Wir haben Zufallszahlen oder zufällige Grössen $x_1, x_2, \cdots, x_n$, die von einer Wahrscheinlichkeitsverteilung

$$p_\theta(x) \;=\; p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n)$$

generiert worden sind. Es gelte also

$$\int_{\mathbb{R}^n} p_\theta(x)\; d^n x \;=\; \int_{\mathbb{R}^n} p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n)\; dx_1 dx_2 \cdots dx_n \;=\; 1$$

Die theta's $\theta_1, \cdots, \theta_m$ sind die Modellparameter. Wenn wir das $p_\theta(x)$ nur als Funktion von θ auffassen, weil wir für die $x_1, \cdots, x_n$ die uns gegebenen realisierten Daten einsetzen, dann ist das genau die Likelihood-Funktion,

$$L(\theta) \;=\; L(\theta_1,\cdots,\theta_m) \;=\; p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n)$$

Wir hatten das $p_\theta(x)$ für die 3 Beispiele aus dem Kapitel 5.1 noch einmal hingeschrieben und in allen 3 Beispielen konnten wir Ableitungen nach den Modellparametern mit den Integralen oder den Summen über die x_i vertauschen. Das ist nicht immer ganz selbstverständlich, da man etwa bei gleichverteilten Zufallszahlen ja eine Rechteck-Funktion als W'keitsdichte hat, die im klassischen Sinne nicht differenzierbar ist, und solche Fälle hatten wir ausgeschlossen. Die Cramer-Rao Abschätzung war dann:

Theorem 5.3.2 (Cramer-Rao Lower Bound): Gegeben sei eine Wahrscheinlichkeitsverteilung $p_\theta(x) = p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n)$ mit

$$\int_{\mathbb{R}^n} p_\theta(x)\; d^n x \;=\; \int_{\mathbb{R}^n} p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n)\; dx_1 dx_2 \cdots dx_n \;=\; 1$$

und Ableitungen nach den Modellparametern $\theta_1, ..., \theta_m$ seien vertauschbar mit den x-Integralen. Für $k = 1, \cdots, m$ seien

$$\hat\theta \;=\; \begin{pmatrix} \hat\theta_1 \\ \vdots \\ \hat\theta_m \end{pmatrix} \;:\; \mathbb{R}^n \;\to\; \mathbb{R}^m$$

erwartungstreue Schätzer, d.h. es gilt

$$\mathsf{E}[\,\hat\theta_k\,] \;:=\; \int_{\mathbb{R}^n} \hat\theta_k(x)\; p_\theta(x)\; d^n x \;=\; \theta_k$$

Weiter sei

$$\mathsf{Cov}(\hat{\theta}) \;:=\; \Big(\mathsf{Cov}[\hat{\theta}_k, \hat{\theta}_\ell] \Big)_{k,\ell=1,\ldots,m} \;\in\; \mathbb{R}^{m\times m}$$

die Covarianz-Matrix von $\hat{\theta}$, insbesondere also

$$\mathsf{V}[\,\hat{\theta}_k\,] \;=\; \mathsf{Cov}(\hat{\theta})_{k,k}$$

Wir definieren die sogenannte Fisher-Informations-Matrix $I(\theta)$ durch

$$I(\theta) \;:=\; \Big(-\mathsf{E}\Big[\, \tfrac{\partial^2 \log p_\theta}{\partial\theta_k \partial\theta_\ell} \,\Big] \Big)_{k,\ell=1,\ldots,m} \;\in\; \mathbb{R}^{m\times m}$$

mit

$$\mathsf{E}\Big[\, \tfrac{\partial^2 \log p_\theta}{\partial\theta_k \partial\theta_\ell} \,\Big] \;:=\; \int_{\mathbb{R}^n} \tfrac{\partial^2 \log p_\theta}{\partial\theta_k \partial\theta_\ell}(x_1, \cdots, x_n)\; p_\theta(x_1, \cdots, x_n)\; d^n x$$

und $I^{-1}(\theta)$ sei das Inverse von $I(\theta)$. Dann gilt:

$$\big\langle\, v\,, \mathsf{Cov}(\hat{\theta})\, v \,\big\rangle \;\geq\; \big\langle\, v\,, I^{-1}(\theta)\, v \,\big\rangle \qquad \forall v \in \mathbb{R}^m$$

Insbesondere,

$$\mathsf{V}[\,\hat{\theta}_k\,] \;\geq\; \big[\, I^{-1}(\theta) \,\big]_{k,k}$$

für jeden erwartungstreuen Schätzer $\hat{\theta}_k : \mathbb{R}^n \to \mathbb{R}$.

Beweis: Für $1 \leq k, \ell \leq m$ betrachten wir das Integral

$$\begin{aligned}
\int_{\mathbb{R}^n} \big\{\, \hat{\theta}_k(x) \,-\, \theta_k \,\big\}\; \frac{\partial \log p_\theta}{\partial\theta_\ell}\; p_\theta(x)\; d^n x \;&=\; \int_{\mathbb{R}^n} \big\{\, \hat{\theta}_k(x) \,-\, \theta_k \,\big\}\; \frac{\partial p_\theta}{\partial\theta_\ell}\; d^n x \\
&=\; \int_{\mathbb{R}^n} \hat{\theta}_k(x)\; \frac{\partial p_\theta}{\partial\theta_\ell}\; d^n x \;-\; \int_{\mathbb{R}^n} \theta_k\; \frac{\partial p_\theta}{\partial\theta_\ell}\; d^n x \\
&=\; \frac{\partial}{\partial\theta_\ell} \int_{\mathbb{R}^n} \hat{\theta}_k(x)\; p_\theta(x)\; d^n x \;-\; \theta_k\; \frac{\partial}{\partial\theta_\ell} \int_{\mathbb{R}^n} p_\theta(x)\; d^n x \\
&=\; \frac{\partial}{\partial\theta_\ell}\, \theta_k \;-\; \theta_k\; \frac{\partial}{\partial\theta_\ell}\, 1 \;=\; \delta_{k,\ell} \qquad (1)
\end{aligned}$$

Es seien jetzt

$$v, w \;\in\; \mathbb{R}^m$$

beliebige Vektoren und etwa

$$\langle v, \hat{\theta} \rangle \;=\; \textstyle\sum_{k=1}^m v_k\, \hat{\theta}_k$$

also $\langle \cdot, \cdot \rangle$ bezeichnet das Standardskalarprodukt im $\mathbb{R}^m$. Dann folgt aus (1), wenn wir von links mit v und von rechts mit w skalar multiplizieren,

$$\int_{\mathbb{R}^n} \left\langle v, \hat{\theta}(x) - \theta \right\rangle \left\langle \nabla_\theta \log p_\theta(x), w \right\rangle \, p_\theta(x) \, d^n x \;\; = \;\; \langle v, w \rangle \tag{2}$$

Für beliebige Funktionen $f, g : \mathbb{R}^n \to \mathbb{R}$, $f = f(x)$ und $g = g(x)$, wird durch

$$(f, g) \;\; := \;\; \int_{\mathbb{R}^n} f(x) \, g(x) \, p_\theta(x) \, d^n x$$

ein Skalarprodukt auf der Menge der bezüglich $p_\theta(x) d^n x$ quadratintegrablen Funktionen definiert und es gilt die Cauchy-Schwarz'sche Ungleichung,

$$\begin{aligned} \left| \int_{\mathbb{R}^n} f(x) \, g(x) \, p_\theta(x) \, d^n x \right|^2 &= \left| \, (f, g) \, \right|^2 \\ &\leq \|f\|^2 \, \|g\|^2 \\ &= \int_{\mathbb{R}^n} f(x)^2 \, p_\theta(x) \, d^n x \;\; \int_{\mathbb{R}^n} g(x)^2 \, p_\theta(x) \, d^n x \end{aligned} \tag{3}$$

Wir setzen jetzt

$$\begin{aligned} f(x) &:= \left\langle v, \hat{\theta}(x) - \theta \right\rangle \\ g(x) &:= \left\langle \nabla_\theta \log p_\theta(x), w \right\rangle \end{aligned}$$

und bekommen aus Gleichung (2) und der Cauchy-Schwarz'schen Ungleichung (3)

$$\begin{aligned} \left| \, \langle v, w \rangle \, \right|^2 &= \left| \int_{\mathbb{R}^n} \left\langle v, \hat{\theta}(x) - \theta \right\rangle \left\langle \nabla_\theta \log p_\theta(x), w \right\rangle \, p_\theta(x) \, d^n x \right|^2 \\ &\leq \int_{\mathbb{R}^n} \left| \left\langle v, \hat{\theta}(x) - \theta \right\rangle \right|^2 p_\theta(x) \, d^n x \\ &\quad \times \int_{\mathbb{R}^n} \left| \left\langle \nabla_\theta \log p_\theta(x), w \right\rangle \right|^2 p_\theta(x) \, d^n x \end{aligned} \tag{4}$$

Das erste Integral auf der rechten Seite von (4) können wir auch folgendermassen schreiben:

$$\begin{aligned} \int_{\mathbb{R}^n} \left| \left\langle v, \hat{\theta}(x) - \theta \right\rangle \right|^2 p_\theta(x) \, d^n x &= \sum_{k,\ell=1}^m v_k v_\ell \int_{\mathbb{R}^n} \left[\hat{\theta}_k(x) - \theta_k \right] \left[\hat{\theta}_\ell(x) - \theta_\ell \right] \, p_\theta(x) \, d^n x \\ &= \sum_{k,\ell=1}^m v_k v_\ell \, \mathsf{Cov}[\hat{\theta}_k, \hat{\theta}_\ell] \\ &= \langle v, \mathsf{Cov}(\hat{\theta}) v \rangle \end{aligned} \tag{5}$$

Und für das zweite Integral auf der rechten Seite von (4) erhalten wir

$$\int_{\mathbb{R}^n} \left| \left\langle \nabla_\theta \log p_\theta(x), w \right\rangle \right|^2 p_\theta(x) \, d^n x \;\; = \;\; \sum_{k,\ell=1}^m w_k w_\ell \int_{\mathbb{R}^n} \frac{\partial \log p_\theta}{\partial \theta_k} \, \frac{\partial \log p_\theta}{\partial \theta_\ell} \, \; p_\theta(x) \, d^n x \tag{6}$$

Das Integral auf der rechten Seite von (6) können wir etwas umformen: Wegen

$$\int_{\mathbb{R}^n} p_\theta(x)\, d^n x \;=\; 1$$

ist

$$\begin{aligned} 0 \;=\; \frac{\partial}{\partial\theta_k} \int_{\mathbb{R}^n} p_\theta(x)\, d^n x \;&=\; \int_{\mathbb{R}^n} \frac{\partial p_\theta}{\partial\theta_k}(x)\, d^n x \\ &=\; \int_{\mathbb{R}^n} \frac{\frac{\partial p_\theta}{\partial\theta_k}(x)}{p_\theta(x)}\, p_\theta(x)\, d^n x \\ &=\; \int_{\mathbb{R}^n} \frac{\partial \log p_\theta(x)}{\partial\theta_k}\, p_\theta(x)\, d^n x \end{aligned} \tag{7}$$

Diese Identität tun wir noch einmal nach θ_ℓ ableiten und bekommen

$$\begin{aligned} 0 \;&=\; \frac{\partial}{\partial\theta_\ell} \int_{\mathbb{R}^n} \frac{\partial \log p_\theta(x)}{\partial\theta_k}\, p_\theta(x)\, d^n x \\ &=\; \int_{\mathbb{R}^n} \frac{\partial^2 \log p_\theta(x)}{\partial\theta_\ell\partial\theta_k}\, p_\theta(x)\, d^n x \;+\; \int_{\mathbb{R}^n} \frac{\partial \log p_\theta(x)}{\partial\theta_k}\, \frac{\partial p_\theta}{\partial\theta_\ell}(x)\, d^n x \\ &=\; \int_{\mathbb{R}^n} \frac{\partial^2 \log p_\theta(x)}{\partial\theta_\ell\partial\theta_k}\, p_\theta(x)\, d^n x \;+\; \int_{\mathbb{R}^n} \frac{\partial \log p_\theta(x)}{\partial\theta_k}\, \frac{\partial \log p_\theta(x)}{\partial\theta_\ell}\, p_\theta(x)\, d^n x \end{aligned}$$

Also haben wir

$$\begin{aligned} \int_{\mathbb{R}^n} \frac{\partial \log p_\theta(x)}{\partial\theta_k}\, \frac{\partial \log p_\theta(x)}{\partial\theta_\ell}\, p_\theta(x)\, d^n x \;&=\; -\int_{\mathbb{R}^n} \frac{\partial^2 \log p_\theta(x)}{\partial\theta_\ell\partial\theta_k}\, p_\theta(x)\, d^n x \\ &=\; -\,\mathsf{E}\left[\, \frac{\partial^2 \log p_\theta}{\partial\theta_\ell\partial\theta_k} \,\right] \\ &=\; +\, I(\theta)_{\ell,k} \;=\; +\, I(\theta)_{k,\ell} \end{aligned} \tag{8}$$

Das können wir auf der rechten Seite von (6) einsetzen und bekommen

$$\begin{aligned} \int_{\mathbb{R}^n} \big|\, \langle \nabla_\theta \log p_\theta(x), w \rangle \,\big|^2\, p_\theta(x)\, d^n x \;&=\; \sum_{k,\ell=1}^{m} w_k w_\ell \int_{\mathbb{R}^n} \frac{\partial \log p_\theta}{\partial\theta_k}\, \frac{\partial \log p_\theta}{\partial\theta_\ell}\; p_\theta(x)\, d^n x \\ &=\; \sum_{k,\ell=1}^{m} w_k w_\ell\, I(\theta)_{k,\ell} \\ &=\; \langle w, I(\theta) w \rangle \end{aligned} \tag{9}$$

Insbesondere ist das $I(\theta)$ und damit auch $I^{-1}(\theta)$ eine positiv definite Matrix. Wir setzen (5)

und (9) auf der rechten Seite von (4) ein und bekommen

$$\begin{aligned} \big|\,\langle v, w\rangle\,\big|^2 &\leq \int_{\mathbb{R}^n} \big|\,\big\langle v, \hat\theta(x) - \theta\big\rangle\,\big|^2\, p_\theta(x)\; d^n x \;\times\; \int_{\mathbb{R}^n} \big|\,\big\langle \nabla_\theta \log p_\theta(x), w\big\rangle\,\big|^2\, p_\theta(x)\; d^n x \\ &= \langle v, \mathsf{Cov}(\hat\theta) v\rangle \;\times\; \langle w, I(\theta) w\rangle \end{aligned} \tag{10}$$

Wir wählen jetzt

$$w \;:=\; I^{-1}(\theta)\, v$$

und erhalten aus (10)

$$\big|\,\langle v, I^{-1}(\theta) v\rangle\,\big|^2 \;\leq\; \langle v, \mathsf{Cov}(\hat\theta) v\rangle \;\times\; \langle I^{-1}(\theta) v, v\rangle \tag{11}$$

oder, mit $\langle I^{-1}(\theta)v, v\rangle = \langle v, I^{-1}(\theta)v\rangle \geq 0$,

$$\langle v, I^{-1}(\theta) v\rangle \;\leq\; \langle v, \mathsf{Cov}(\hat\theta) v\rangle \tag{12}$$

Das aber genau war zu zeigen. ■

Aus dem obigen Beweis können wir noch eine interessante Folgerung ziehen: Man ist natürlich interessiert an erwartungstreuen Schätzern mit minimaler Varianz. Schätzer, für die in Gleichung (12) das Gleichheitszeichen gilt, sind offensichtlich Schätzer mit minimaler Varianz. Wir haben in Gleichung (12) ein Gleichheitszeichen genau dann, wenn wir in der Cauchy-Schwarzschen Ungleichung (3), das war

$$\big|\,\textstyle\int_{\mathbb{R}^n} f(x)\, g(x)\, p_\theta(x)\, d^n x\,\big|^2 \;\leq\; \int_{\mathbb{R}^n} f(x)^2\, p_\theta(x)\, d^n x \;\int_{\mathbb{R}^n} g(x)^2\, p_\theta(x)\, d^n x$$

ein Gleichheitszeichen haben, mit

$$\begin{aligned} f(x) &= \big\langle v, \hat\theta(x) - \theta\big\rangle \\ g(x) &= \big\langle \nabla_\theta \log p_\theta(x), w\big\rangle \;=\; \big\langle \nabla_\theta \log p_\theta(x), I^{-1}(\theta) v\big\rangle \end{aligned}$$

In der Cauchy-Schwarzschen Ungleichung hat man genau dann ein Gleichheitszeichen, wenn die Funktionen oder Vektoren linear abhängig sind,

$$g(x) \;=\; c\, f(x)$$

mit einer von x unabhängigen Konstanten c. Ein erwartungstreuer Schätzer, dessen Varianz durch die Cramer-Rao Lower Bound gegeben ist, kann also genau dann gefunden werden, wenn die Gleichung

$$\big\langle \nabla_\theta \log p_\theta(x), I^{-1}(\theta) v\big\rangle \;=\; \big\langle I^{-1}(\theta) \nabla_\theta \log p_\theta(x), v\big\rangle \;=\; c\,\big\langle \hat\theta(x) - \theta, v\big\rangle \quad \forall\, x \in \mathbb{R}^n$$

erfüllt ist für beliebige $v \in \mathbb{R}^m$. Also müssen wir haben

$$c\,\big(\,\hat\theta(x) - \theta\,\big) \;=\; I^{-1}(\theta) \nabla_\theta \log p_\theta(x) \tag{13}$$

wobei das $c = c_\theta$ von θ abhängen kann, aber nicht von den x. Tatsächlich muss das $c = 1$ sein, denn: Aus (13) folgt

$$\nabla_\theta \log p_\theta(x) \;=\; c_\theta \; I(\theta) \, \big(\hat{\theta}(x) - \theta \big)$$

oder

$$\frac{\partial \log p_\theta}{\partial \theta_j} \;=\; \sum_{k=1}^{m} c_\theta \, I(\theta)_{j,k} \, \big(\hat{\theta}_k(x) - \theta_k \big)$$

Wir differenzieren das nach θ_ℓ und bekommen mit der Produktregel

$$\begin{aligned}
\frac{\partial^2 \log p_\theta}{\partial \theta_\ell \partial \theta_j} &= \sum_{k=1}^{m} \frac{\partial [c_\theta I(\theta)_{j,k}]}{\partial \theta_\ell} \, \big(\hat{\theta}_k(x) - \theta_k \big) \;+\; \sum_{k=1}^{m} c_\theta \, I(\theta)_{j,k} \, \frac{\partial [\hat{\theta}_k(x) - \theta_k]}{\partial \theta_\ell} \\
&= \sum_{k=1}^{m} \frac{\partial [c_\theta I(\theta)_{j,k}]}{\partial \theta_\ell} \, \big(\hat{\theta}_k(x) - \theta_k \big) \;+\; \sum_{k=1}^{m} c_\theta \, I(\theta)_{j,k} \, (-\delta_{k,\ell}) \\
&= \sum_{k=1}^{m} \frac{\partial [c_\theta I(\theta)_{j,k}]}{\partial \theta_\ell} \, \big(\hat{\theta}_k(x) - \theta_k \big) \;-\; c_\theta \, I(\theta)_{j,\ell}
\end{aligned}$$

Von dieser Gleichung nehmen wir den Erwartungswert und bekommen

$$\begin{aligned}
-I(\theta)_{j,\ell} \;=\; \mathsf{E}\left[\frac{\partial^2 \log p_\theta}{\partial \theta_\ell \partial \theta_j} \right] \;&=\; \sum_{k=1}^{m} \frac{\partial [c_\theta I(\theta)_{j,k}]}{\partial \theta_\ell} \, \underbrace{\mathsf{E}[\, \hat{\theta}_k - \theta_k \,]}_{=\,0} \;-\; c_\theta \, I(\theta)_{j,\ell} \\
&=\; -\; c_\theta \, I(\theta)_{j,\ell}
\end{aligned}$$

also muss das $c_\theta = 1$ sein. Aus der Gleichung (13) ergibt sich dann also die

Folgerung 5.3.3: Erwartungstreue Minimum-Varianz-Schätzer sind von der Form

$$\hat{\theta}(x) \;=\; \theta \;+\; I^{-1}(\theta) \, \nabla_\theta \log p_\theta(x) \tag{14}$$

Da die Modellparameter $\theta = (\theta_1, \cdots, \theta_m)$ ja gerade geschätzt werden sollen, also nicht bekannt sind, stellt Gleichung (14) keine explizite Formel für einen Minimum-Varianz-Schätzer dar. In einem benutzbaren Schätzer dürfen die Modellparameter nicht explizit auftauchen.