week11b: Kapitel 7: Effizienz von Schätzern, Teil 2
Beweis der Cramer-Rao Abschätzung

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

$$\int_{\mathbb{R}^n} p_\theta(x)\, d^nx \;=\; \int_{\mathbb{R}^n} p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n)\, dx_1dx_2\cdots dx_n \;=\; 1$$

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

$$\int_{\mathbb{R}^n} p_\theta(x)\, d^nx \;=\; \int_{\mathbb{R}^n} p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n)\, dx_1dx_2\cdots dx_n \;=\; 1$$

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

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

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-Informationsmatrix $I(\theta)$ 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

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

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,

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

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

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

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

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

Wir setzen jetzt

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

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

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

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

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

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

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

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

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

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

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

Wir wählen jetzt

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

und erhalten aus (10)

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

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. ■