Asymptotisch gedrag van de Maximum Likelihood-schatter

1 Wald's consistentie bewijs

Laat $X_1, X_2, \ldots, X_n$ onafhankelijke stochastische variabelen zijn met dichtheid p_θ, voor $\theta \in \Theta \subset \mathbb{R}^k$. We nemen aan dat de verzameling $\mathcal{X} \subset \mathbb{R}$ waarin X_i onder p_θ waarden aanneemt, niet afhangt van θ.

In de onderstaande stelling nemen we onder andere aan dat de functie $(x, \theta) \mapsto p_\theta(x)$ continu is op $\mathcal{X} \times \Theta$. De likelihood

$$L(\theta; X_1, \ldots, X_n) = \prod_{i=1}^{n} p_\theta(X_i)$$

is dan een continue functie van θ (die afhangt van de observaties X_i). Onder de aanname dat Θ compact is, neemt de likelihood dus een maximale waarde aan in een zeker punt in Θ. Dit punt is, per definitie, de Maximum Likelihood-schatter $\hat{\theta}_n$. Merk op dat de likelihood in meerdere punten een maximum kan hebben. In dat geval definiëren we $\hat{\theta}_n$ als één van die punten.

Stelling 1.1. *Stel dat Θ compact is, dat $(x, \theta) \mapsto p_\theta(x)$ positief, continu en begrensd is op $\mathcal{X} \times \Theta$, en dat $p_\theta \neq p_{\theta_0}$ als $\theta \neq \theta_0$. Dan geldt voor alle $\theta_0 \in \Theta$ dat*

$$\hat{\theta}_n \xrightarrow{\mathbb{P}_{\theta_0}} \theta_0$$

als $n \to \infty$. Dat wil zeggen dat voor alle $\varepsilon > 0$, $\mathbb{P}_{\theta_0}(|\hat{\theta}_n - \theta_0| > \varepsilon) \to 0$ als $n \to \infty$.

Bewijs. Neem $\theta_0 \in \Theta$ vast. Definieer $m_\theta(x) = \log p_\theta(x)$ en

$$\mathbb{M}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i), \qquad M(\theta) = \mathbb{E}_{\theta_0} m_\theta(X_i).$$

Merk op dat $\mathbb{M}_n$ gelijk is aan n^{-1} maal de log-likelihood, dus $\hat{\theta}_n$ is het (beter: een) punt waar $\mathbb{M}_n$ maximaal is. Vanwege de wet van grote aantallen convergeert $\mathbb{M}_n(\theta)$ onder $\mathbb{P}_{\theta_0}$ met kans één naar $M(\theta)$ voor iedere θ. In dit bewijs laten we zien dat het punt waar $\mathbb{M}_n$ maximaal is "mee convergeert" naar het punt waar M maximaal is.

We bewijzen eerst dat M een uniek maximum heeft in het punt θ_0. Calculus laat zien dat voor $x > -1$ geldt dat $x \leq (1+x)\log(1+x)$. Voor willekeurige positieve kansdichtheden g en h hebben we daarom

$$\int (\sqrt{g} - \sqrt{h})^2 = 2 - 2\int \sqrt{gh} = 2\int (g - \sqrt{gh}) =$$
$$2\int \frac{\sqrt{g}}{\sqrt{h}}\Big(\frac{\sqrt{g}}{\sqrt{h}} - 1\Big)h \leq 2\int \frac{\sqrt{g}}{\sqrt{h}}\Big(\frac{\sqrt{g}}{\sqrt{h}} \log \frac{\sqrt{g}}{\sqrt{h}}\Big)h = \int g \log \frac{g}{h}.$$

Er volgt dat voor alle $\theta \in \Theta$,

$$\begin{aligned} M(\theta_0) - M(\theta) &= \int p_{\theta_0}(x) \log \frac{p_{\theta_0}(x)}{p_\theta(x)}\, dx \\ &\geq \int (\sqrt{p_\theta} - \sqrt{p_{\theta_0}})^2 \geq 0. \end{aligned}$$

Hieruit volgt ten eerste dat M maximaal is in θ_0. Verder zien we dat als $M(\theta) = M(\theta_0)$ voor zekere $\theta \in \Theta$, dan is

$$\int (\sqrt{p_\theta(x)} - \sqrt{p_{\theta_0}(x)})^2\, dx = 0.$$

Vanwege de identificeerbaarheidsaanname kan dat alleen als $\theta = \theta_0$ (check!). We concluderen dat M een uniek maximum heeft in θ_0.

Voor $U \subset \Theta$ definiëren we nu $m_U(x) = \sup_{\theta \in U} m_\theta(x)$. We laten nu eerst zien dat er voor iedere $\theta \neq \theta_0$ een open omgeving $U_\theta \subset \Theta$ bestaat zodanig dat

$$\mathbb{E}_{\theta_0} m_{U_\theta}(X_1) < M(\theta_0).$$

Neem voor het bewijs van deze bewering $\theta \neq \theta_0$ en een rij open omgevingen V_l van θ zodanig dat $\mathrm{diam}(V_l) \to 0$. Vanwege de aangenomen continuïteit geldt dat voor iedere $x \in \mathcal{X}$,

$$m_{V_l}(x) = \sup_{\theta' \in V_l} m_{\theta'}(x) \to m_\theta(x)$$

voor $l \to \infty$. De aannames impliceren dat puntsgewijs convergerende functies aan de linkerkant uniform begrensd zijn door een constante. Vanwege de gedomineerde convergentie stelling en de vorige alinea volgt daarom dat

$$\mathbb{E}_{\theta_0} m_{V_l}(X_1) \to \mathbb{E}_{\theta_0} m_\theta(X_1) = M(\theta) < M(\theta_0)$$

als $l \to \infty$. Voor l groot genoeg is de linkerkant dus strikt kleiner dan $M(\theta_0)$, waaruit de bewering volgt.

Voor het slot van het bewijs nemen we $\varepsilon > 0$ vast en definiëren we $B = \{\theta \in \Theta : |\theta - \theta_0| \geq \varepsilon\}$. Deze verzameling is compact en wordt overdekt door de collectie van open verzamelingen $\{U_\theta,\, \theta \in \Theta\}$. Zij $U_{\theta_1}, \ldots, U_{\theta_p}$ een eindige deeloverdekking. Dan geldt voor $\theta \in B$ dat er een $j \in \{1, \ldots, p\}$ is zodanig dat $\theta \in U_{\theta_j}$, en dus $m_\theta(x) \leq m_{U_{\theta_j}}(x)$. Het volgt dat

$$\mathbb{M}_n(\theta) = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i) \leq \frac{1}{n}\sum_{i=1}^n m_{U_{\theta_j}}(X_i) \leq \max_{j=1,\ldots,p} \frac{1}{n}\sum_{i=1}^n m_{U_{\theta_j}}(X_i).$$

De rechterkant hangt niet af van θ, dus we hebben ook

$$\sup_{\theta \in B} \mathbb{M}_n(\theta) \leq \max_{j=1,\ldots,p} \frac{1}{n}\sum_{i=1}^n m_{U_{\theta_j}}(X_i) \overset{\mathbb{P}_{\theta_0}-\text{b.z.}}{\to} \max_{j=1,\ldots,p} \mathbb{E}_{\theta_0} m_{U_{\theta_j}}(X_1) < M(\theta_0)$$

als $n \to \infty$. Omdat vanwege de wet van grote aantallen $\mathbb{M}_n(\theta_0) \to M(\theta_0)$ in kans, concluderen we dat

$$\mathbb{P}_{\theta_0}(|\hat\theta_n - \theta_0| \geq \varepsilon) \leq \mathbb{P}_{\theta_0}\Big(\sup_{\theta \in B} \mathbb{M}_n(\theta) \geq \mathbb{M}_n(\theta_0)\Big) \to 0$$

als $n \to \infty$. □

Het bewijs van Stelling 1.1 gaat terug tot Wald (1949). De voorwaarden van de stelling kunnen voor sommige concrete voorbeelden te restrictief zijn. Het is mogelijk om consistentie van de meest aannemelijke schatter af te leiden onder zwakkere voorwaarden. Dit vergt echter het gebruik van wiskundige technieken die buiten het bestek van dit college vallen. Zie bijvoorbeeld Van der Vaart (1998) voor meer informatie.

2 Asymptotische normaliteit

Laat weer $X_1, X_2, \dots, X_n$ onafhankelijke stochastische variabelen zijn met dichtheid p_θ, voor $\theta \in \Theta \subset \mathbb{R}$. We nemen nu meer "regulariteitsvoorwaarden" aan dan in de vorige paragraaf. In het bijzonder nemen we aan dat $\log p_\theta$ differentieerbaar is en dat de Fisher informatie (in één waarneming)

$$i_\theta = \mathbb{V}\mathrm{ar}_\theta \frac{\partial}{\partial\theta} \log p_\theta(X_1)$$

eindig is.

Stelling 2.1. *Onder regulariteitsvoorwaarden geldt onder $\mathbb{P}_{\theta_0}$ dat*

$$\sqrt{n}(\hat\theta_n - \theta_0) \xrightarrow{\mathrm{d}} N(0, i_{\theta_0}^{-1})$$

als $n \to \infty$. Dat wil zeggen dat voor alle $x \in \mathbb{R}$,

$$\mathbb{P}_{\theta_0}\Big(\sqrt{n i_{\theta_0}}(\hat\theta_n - \theta_0) \le x\Big) \to \Phi(x),$$

als $n \to \infty$, met $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy$ de standaard normale verdelingsfunctie.

Schets van bewijs. Definieer $\ell_\theta(x) = \log p_\theta(x)$, $\dot\ell_\theta(x) = \frac{\partial}{\partial\theta}\ell_\theta(x)$ en $\ddot\ell_\theta(x) = \frac{\partial}{\partial\theta}\dot\ell_\theta(x)$. Onder voorwaarden geldt voor de MLS dat $S_n(\hat\theta_n) = 0$, met

$$S_n(\theta) = \sum_{i=1}^{n} \dot\ell_\theta(X_i) \tag{1}$$

de score functie. Een Taylor expansie van S_n rond θ_0 geeft, met $\dot S_n(\theta) = \frac{\partial}{\partial\theta} S_n(\theta)$,

$$0 = S_n(\hat\theta_n) = S_n(\theta_0) + \dot S_n(\bar\theta_n)(\hat\theta_n - \theta_0),$$

met $\bar\theta_n$ een punt tussen θ_0 en $\hat\theta_n$. Er volgt dat

$$\sqrt{n}(\hat\theta_n - \theta_0) = \Big(-\frac{1}{n}\dot S_n(\bar\theta_n)\Big)^{-1} \frac{1}{\sqrt{n}} S_n(\theta_0).$$

Voor de termen in de som (1) geldt (zie het bewijs van Cramér-Rao) $\mathbb{E}_{\theta_0}\dot\ell_{\theta_0}(X_1) = 0$ en $\mathbb{V}\mathrm{ar}_{\theta_0}\dot\ell_{\theta_0}(X_1) = i_{\theta_0}$. De centrale limietstelling impliceert dat

$$\frac{1}{\sqrt{n}} S_n(\theta_0) \xrightarrow{\mathrm{d}} N(0, i_{\theta_0})$$

onder $\mathbb{P}_{\theta_0}$. De stelling in de vorige paragraaf laat zien dat the MLE consistent is onder voorwaarden. In dat geval convergeert $\bar{\theta}_n$ in kans naar θ_0 onder $\mathbb{P}_{\theta_0}$ en dus

$$-\frac{1}{n}\dot{S}_n(\bar{\theta}_n) \approx -\frac{1}{n}\dot{S}_n(\theta_0)$$

voor grote n. Vanwege de wet van grote aantallen convergeert dit naar $-\mathbb{E}_{\theta_0}\ddot{\ell}_{\theta_0}(X_1)$. Het differentiëren van de relatie

$$\int \dot{\ell}_\theta(x)p_\theta(x)\,dx = 0$$

naar θ laat zien dat $-\mathbb{E}_{\theta_0}\ddot{\ell}_{\theta_0}(X_1) = i_{\theta_0}$ (ga na!).

Alles samenvoegend zien we dat

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathrm{d}} i_{\theta_0}^{-1}N(0, i_{\theta_0}) = N(0, i_{\theta_0}^{-1})$$

onder $\mathbb{P}_{\theta_0}$. □

Een precies bewijs langs de geschetste lijnen kan gevonden worden in Lehmann and Casella (1998). Ook hier geldt dat met behulp van meer geavanceerde wiskundige middelen resultaten kunnen worden afgeleid onder zwakkere voorwaarden (cf. Van der Vaart (1998)).

Bibliografie

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer-Verlag, New York, second edition.

Van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press, Cambridge.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statistics* **20**, 595–601.