week11: Die Maximum Likelihood Methode, Teil1

Die Maximum Likelihood Methode ist eine Standard-Methode in der Statistik, um die Modell-Parameter eines statistischen Modells zu bestimmen. Maximum-Likelihood-Schätzer haben in der Regel gute statistische Eigenschaften. Die Funktionsweise der Methode versteht man am besten, wenn man ein paar Beispiele durchrechnet. Wir betrachten die folgenden 3 Beispiele:

- In Beispiel 1 sind n Zufallszahlen gegeben, die von einer Normalverteilung generiert worden sind, es sind aber der Mittelwert und die Standardabweichung der Normalverteilung nicht bekannt. Diese Parameter sollen aus den gegebenen Daten zurückgewonnen werden.
- In Beispiel 2 sind n Poisson-verteilte Zufallszahlen gegeben, der Parameter λ der Poisson-Verteilung ist aber nicht bekannt und soll geschätzt werden. Konzeptionell ist das nicht wesentlich anders als das Beispiel 1.
- In Beispiel 3 sind die Returns einer Finanzzeitreihe gegeben und wir möchten einen Zeithorizont d so wählen, dass die mit einer d-Tages-Volatilität normierten Returns (die hatten wir schon im week2.pdf im letzten Beispiel 3 definiert und in der ersten Aufgabe auf dem 6. Übungsblatt genauer analysiert) möglichst gut durch eine Standard-Normalverteilung approximiert werden können. Wie können wir das optimale d bestimmen?

Die Beispiele 1 und 2 haben die schöne Eigenschaft, dass die Maximum-Likelihood-Schätzer noch analytisch in geschlossener Form berechnet werden können. Im Allgemeinen ist das typischerweise nicht mehr der Fall, das Beispiel 3 führt in natürlicher Weise auf ARCH-Zeitreihenmodelle und für ARCH- oder auch GARCH-Zeitreihenmodelle können die Modellparameter nur numerisch durch Maximieren der log-Likelihood-Funktion bestimmt werden. Ok, schauen wir uns das jetzt im Detail an:

Beispiel 1: Gegeben seien n Zufallszahlen $x_1, x_2, \cdots, x_n$. Wir wissen, dass diese Zahlen mit Hilfe einer Normalverteilung generiert worden sind, kennen aber nicht den Mittelwert μ und die Standardabweichung σ dieser Normalverteilung. Welche Werte von μ und σ passen 'am besten' zu den gegebenen Zahlen $x_1, \cdots, x_n$? Und welches Kriterium nehmen wir für 'am besten'?

Lösung 1: Die Aussage: "Die x_i sind normalverteilt mit Mittelwert μ und Standardabweichung σ" ist äquivalent zu

$$\mathsf{Prob}\Big[\, x_i \in [\tilde{x}_i, \tilde{x}_i + d\tilde{x}_i) \,\Big] \;=\; \tfrac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(\tilde{x}_i-\mu)^2}{2\sigma^2}}\, d\tilde{x}_i \qquad (1)$$

Wir nehmen an, dass wir unabhängige Zufallszahlen haben, und bekommen dann aus (1)

$$\begin{aligned}
&\mathsf{Prob}\Big[\,x_1 \in [\tilde{x}_1, \tilde{x}_1 + d\tilde{x}_1),\; x_2 \in [\tilde{x}_2, \tilde{x}_2 + d\tilde{x}_2),\; \cdots\;,\; x_n \in [\tilde{x}_n, \tilde{x}_n + d\tilde{x}_n)\,\Big] \;= \\
&\overset{\text{unabh.}}{=} \quad \mathsf{Prob}\Big[\,x_1 \in [\tilde{x}_1, \tilde{x}_1 + d\tilde{x}_1)\,\Big]\; \mathsf{Prob}\Big[\,x_2 \in [\tilde{x}_2, \tilde{x}_2 + d\tilde{x}_2)\,\Big]\; \cdots\; \mathsf{Prob}\Big[\,x_n \in [\tilde{x}_n, \tilde{x}_n + d\tilde{x}_n)\,\Big] \\
&\overset{(1)}{=} \quad \prod_{i=1}^{n} \tfrac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(\tilde{x}_i-\mu)^2}{2\sigma^2}}\, d\tilde{x}_i \\
&= \quad (2\pi\sigma^2)^{-\frac{n}{2}}\; \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\tilde{x}_i - \mu)^2 \Big\}\, d\tilde{x}_1 \cdots d\tilde{x}_n
\end{aligned} \tag{2}$$

Jetzt haben wir ja konkrete Werte für die Zufallszahlen, nämlich $x_1, \cdots, x_n$. Diese Werte können wir auf der rechten Seite von (2) für die $\tilde{x}_i$ einsetzen. Die $d\tilde{x}_i$ sind Intervallbreiten, etwa $d\tilde{x}_i = 0.1$ oder $d\tilde{x}_i = 0.01$. Wir werden gleich sehen, dass diese Intervallbreiten für die eigentliche Rechnung keine Rolle spielen, aber um eine konkrete Vorstellung zu haben, wählen wir etwa $d\tilde{x}_i = 0.01$ und wir lassen die Tilde weg, schreiben einfach dx_i. Dann bekommen wir eine Funktion, die nur noch von μ und σ abhängt,

$$L(\mu, \sigma) \;:=\; (2\pi\sigma^2)^{-\frac{n}{2}}\; \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \Big\}\, dx_1 \cdots dx_n \tag{3}$$

Diese Funktion heisst **Likelihood-Funktion**, sie gibt also die W'keit an, dass, unter der Annahme, die Zufallszahlen sind von einer Normalverteilung mit Mittelwert μ und Standardabweichung σ generiert worden, dass sich dann konkret Werte in den Intervallen $[x_i, x_i + dx_i)$ realisieren. Also ist es plausibel, das μ und das σ dann so zu wählen, dass diese W'keit maximal wird,

$$L(\mu, \sigma) \overset{!}{\to} \max \qquad \Rightarrow \qquad \mu = \mu_{\mathrm{ML}},\; \sigma = \sigma_{\mathrm{ML}} \tag{4}$$

Die so gewonnenen Werte von μ und σ heissen dann die **Maximum-Likelihood-Schätzer** von μ und σ.

In unserem konkreten Beispiel können wir das Maximum der Likelihood-Funktion (3) analytisch in geschlossener Form berechnen, das machen wir gleich in dem Theorem 11.1 weiter unten. In vielen Anwendungssituationen ist jedoch eine analytische Berechnung in geschlossener Form nicht möglich und man ist auf numerische Prozeduren angewiesen. Anstatt die Likelihood-Funktion L selber zu maximieren, betrachtet man dann typischerweise den Logarithmus der Likelihood-Funktion $\log L$. Da der Logarithmus eine streng monoton steigende Funktion ist, gilt

$$(\mu, \sigma) \text{ maximiert } L(\mu, \sigma) \qquad \Leftrightarrow \qquad (\mu, \sigma) \text{ maximiert } \log L(\mu, \sigma)$$

Das macht man jetzt aus folgendem Grund: Das L ist typischerweise durch ein Produkt gegeben und die Anzahl der Faktoren in diesem Produkt ist gleich der Anzahl der verfügbaren Daten. Wenn man etwa eine Finanz-Zeitreihe hat mit täglichen Schlusskursen der letzten 10 Jahre, dann sind das etwa $N = 2500$ Daten und das L ist dann im wesentlichen

$$L \;\sim\; \mathrm{prob}^N \;=\; \mathrm{prob}^{2500}$$

Auf dem Computer ist die Zahl prob^{2500} entweder 0 oder unendlich, je nachdem, ob das prob kleiner oder grösser ist als 1 (probieren Sie das ruhig mal aus), also das ist numerisch nicht handhabbar. In Gegensatz dazu ist

$$\log L \;\sim\; \log\big[\,\text{prob}^N\,\big] \;=\; N\log[\text{prob}] \;=\; 2500\,\log[\text{prob}]$$

und das ist numerisch unproblematisch.

Kehren wir jetzt zu unserem konkreten Beispiel 1 zurück. Auch für die analytische Rechnung ist es günstig, den Logarithmus zu betrachten,

$$\begin{aligned}
\log L(\mu,\sigma) &= \log\Big[\,(2\pi\sigma^2)^{-\frac{n}{2}}\;\exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big\}\,dx_1\cdots dx_n\Big] \\
&= -\frac{n}{2}\log[2\pi\sigma^2] \;-\; \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \;+\; \log[dx_1\cdots dx_n] \\
&= -n\,\log\sigma \;-\; \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \;+\; \text{const} \qquad (5)
\end{aligned}$$

wobei die Grösse

$$\text{const} \;:=\; -\frac{n}{2}\log[2\pi] \;+\; \log[dx_1\cdots dx_n]$$

eine von μ und σ unabhängige Zahl ist. Es ist also hinreichend, die Funktion

$$\log\tilde{L}(\mu,\sigma) \;:=\; -n\,\log\sigma \;-\; \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \qquad (6)$$

zu betrachten. Wie oben bereits angekündigt, stellen wir also fest, dass die Intervallbreiten $dx_1,\cdots,dx_n$ für die eigentliche Rechnung irrelevant sind. Es gilt das folgende

Theorem 11.1: Das $\log\tilde{L}(\mu,\sigma)$ aus Gleichung (6) wird maximiert durch

$$\mu \;=\; \frac{1}{n}\sum_{i=1}^{n}x_i \;=:\; \mu_{\text{ML}}(x_1,\cdots,x_n) \qquad (7)$$

$$\sigma^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \;=:\; \sigma^2_{\text{ML}}(x_1,\cdots,x_n) \qquad (8)$$

mit $\bar{x} := \frac{1}{n}\sum_{i=1}^{n}x_i = \mu_{\text{ML}}(x_1,\cdots,x_n)$.

Beweis: Notwendige Bedingung für ein Maximum ist

$$\tfrac{\partial}{\partial\mu}\log\tilde{L}(\mu,\sigma) \;=\; 0 \qquad (9)$$

$$\tfrac{\partial}{\partial\sigma}\log\tilde{L}(\mu,\sigma) \;=\; 0 \qquad (10)$$

Nun ist

$$\tfrac{\partial}{\partial\mu}\log\tilde{L}(\mu,\sigma) \;=\; +\frac{2}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

also folgt aus Gleichung (9)

$$\sum_{i=1}^{n}(x_i-\mu) \;=\; \sum_{i=1}^{n}x_i \;-\; n\,\mu \;=\; 0$$

oder

$$\mu \;=\; \frac{1}{n}\sum_{i=1}^{n}x_i$$

Für die Ableitung nach σ erhalten wir

$$\tfrac{\partial}{\partial\sigma}\log\tilde{L}(\mu,\sigma) \;=\; -\,\frac{n}{\sigma} \;+\; \frac{2}{2\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

und Gleichung (10) liefert

$$\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 \;=\; \frac{n}{\sigma}$$

oder

$$\sigma^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Damit ist das Theorem bewiesen. ■

Machen wir gleich das Beispiel 2, wie gesagt, das geht fast genauso:

Beispiel 2: Gegeben seien n Zufallszahlen $x_1, x_2, \cdots, x_n$. Wir wissen, dass diese Zahlen mit Hilfe einer Poisson-Verteilung generiert worden sind, kennen aber nicht den Parameter λ der Poisson-Verteilung. Welcher Wert von λ passt am besten, im Sinne der Maximum Likelihood Methode, zu den gegebenen Zahlen $x_1, \cdots, x_n$?

Lösung 2: Die Aussage: "Die x_i sind Poisson-verteilt mit Parameter $\lambda \in \mathbb{R}$" ist äquivalent zu

$$\mathsf{Prob}\big[\,x_i \;=\; k_i\,\big] \;=\; \frac{\lambda^{k_i}}{k_i!}\,e^{-\lambda} \tag{11}$$

mit natürlichen Zahlen $k_i \in \mathbb{N}$. Wir nehmen an, dass wir unabhängige Zufallszahlen haben, und bekommen dann aus (11)

$$\begin{aligned}
&\mathsf{Prob}\Big[\,x_1 = k_1\,,\; x_2 = k_2\,,\; \cdots\;,\; x_n = k_n\Big] \\
&\overset{\text{unabh.}}{=} \;\mathsf{Prob}\big[\,x_1 = k_1\,\big]\;\mathsf{Prob}\big[\,x_2 = k_2\,\big]\;\cdots\;\mathsf{Prob}\big[\,x_n = k_n\,\big] \\
&\overset{(11)}{=} \;\prod_{i=1}^{n}\Big\{\,\frac{\lambda^{k_i}}{k_i!}\,e^{-\lambda}\,\Big\} \;=\; \frac{\lambda^{k_1+\cdots+k_n}}{k_1!\cdots k_n!}\,e^{-n\lambda}
\end{aligned} \tag{12}$$

Jetzt haben wir ja konkrete Werte für die Zufallszahlen, nämlich $x_1, \cdots, x_n$. Diese Werte können wir auf der rechten Seite von (12) für die k_i einsetzen. Dann bekommen wir wieder eine Funktion, die nur noch von dem gesuchten λ abhängt,

$$L(\lambda) \;:=\; \frac{\lambda^{x_1+\cdots+x_n}}{x_1!\cdots x_n!}\, e^{-n\lambda} \tag{13}$$

und das ist dann die Likelihood-Funktion für dieses konkrete Beispiel. Wir wollen das Maximum finden und betrachten dazu wieder den Logarithmus der Likelihood-Funktion,

$$\begin{aligned}
\log L(\lambda) \;&=\; \log\left[\,\frac{\lambda^{x_1+\cdots+x_n}}{x_1!\cdots x_n!}\, e^{-n\lambda}\,\right] \\
&=\; \log\big[\,\lambda^{x_1+\cdots+x_n}\,\big] \;-\; \log\big[\,x_1!\cdots x_n!\,\big] \;+\; \log\big[\,e^{-n\lambda}\,\big] \\
&=\; (x_1+\cdots+x_n)\,\log\lambda \;-\; n\,\lambda \;+\; \text{const}
\end{aligned} \tag{14}$$

wobei die Zahl

$$\text{const} \;:=\; -\,\log\big[\,x_1!\cdots x_n!\,\big]$$

wieder nicht von dem gesuchten Parameter, in diesem Fall das λ, abhängt, so dass sie für das Maximieren vernachlässigt werden kann, da sie etwa beim Ableiten nach λ wegfällt. Also:

$$\frac{d}{d\lambda}\log L(\lambda) \;=\; \frac{x_1+\cdots+x_n}{\lambda} \;-\; n \;\stackrel{!}{=}\; 0$$

und wir bekommen

$$\lambda \;=\; \frac{x_1+\cdots+x_n}{n} \;=:\; \hat{\lambda}_{\mathrm{ML}}(x_1,\cdots,x_n) \tag{15}$$

Die Funktion $\hat{\lambda}_{\mathrm{ML}} : \mathbb{R}^n \to \mathbb{R}$ gegeben durch (15) heisst dann der Maxium Likelihood Schätzer für das λ. Dass das in diesem Fall gerade der Mittelwert der x_i ist, ist jetzt nicht weiter erstaunlich, da für eine mit Parameter λ Poisson-verteilte Zufallszahl x die Gleichung $\mathsf{E}[x] = \lambda$ gilt.

Das Beispiel 3 diskutieren wir dann in der nächsten Veranstaltung.