Stochastik für Info
Statistische Modelle und ihre Kalibrierung

Hanno Gottschalk

June 21, 2023

Inhaltsverzeichnis der Vorlesung

Statistische Modelle

Was muß ein ’statistisches Modell’ leisten?

- Begrifflicher Rahmen für stochastischen Ausgang von Zufallsexperimenten
- (Unabhängige) Wiederholung von (identischen) Zufallsexperimenten
- Modell für Stichprobe und Daten (uni- und multivariat)
- Vorauswahl für vermutete Verteilungsfamilie
- Anpassung der ’am besten passenden’ Verteilung aus der Familie
- Konzept für Validierung der Vorauswahl

Statistische Modelle

Def. Ein *statistisches Modell* ist gegeben durch

- Eine Ereignismenge Ω
- Zufallsvariablen $X, \dots, X_n : \Omega \to \mathbb{R}^d$ oder $\to \mathbb{Z}^d$ oder andere Zustandsräume
- Eine Familie von W-Maßen P_θ auf Ω mit Parameterraum $\theta \in \Theta$

Bemerkung: Anstelle der W-Maße werden zumeist die Verteilungsfunktionen/Dichten $F_{X,\theta}$ / $f_{X,\theta}$ angegeben für

$$F_{X,\theta}(x) = P_\theta(X_{i,j} \le x_{i,j} \forall i, j) \quad X = \begin{pmatrix} X_{1,1} & \cdots & X_{1,d} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,d} \end{pmatrix}$$

Beispiele für Statistische Modelle

Das Produktmodell

Def.: Im *Produktmodell* modellieren wir die unabhängige Wiederholung von Zufallsexperimenten unter denselben Rahmenbedingungen.

Nehmen an, dass $X_j \sim F_{X,\theta}$ und (Unabhängigkeit der Zufallsexperimente – nicht der Merkmale!)

$$F_{X,\theta}(x) = F_{(X_1,\dots,X_n)',\theta}(x_{1,1}, \dots, x_{n,d}) = \prod_{j=1}^{n} F_{X,\theta}(x_j)$$

bzw. dasselbe für Dichten (hier kontinuierlich)

$$f_{X,\theta}(x) = f_{(X_1,\dots,X_n)',\theta}(x_{1,1}, \dots, x_{n,d}) = \prod_{j=1}^{n} f_{X,\theta}(\underline{x}_j)$$

Das Produktmodell - univariates Beispiel

Huhn legt an einem Tag mit W-keit p ein Ei oder nicht (W-keit $1-p$)

$n = 200$ Hühner im Stall

Das Produktmodell - univariates Beispiel

- $\Omega = \{0,1\}^{\times 200}$, $\omega = (0,1,1,1,0,\ldots) = (\omega_1, \omega_2, \ldots)$
- $\Theta = [0,1], \theta = p$, $P_p(\omega) = p^{\sum \omega_i}(1-p)^{n-\sum \omega_i}$
- $X_j(\omega) = \omega_j$ Legeerfolg Huhn j ($\omega_j \in \{0,1\}$)
- $f_{X,p}(0) = (1-p)$, $f_{X,p}(1) = p$ diskrete Dichte
- Dichtefunktion

$$\begin{aligned} f_{\underline{X},p}(x_1,\ldots,x_{200}) &= \begin{cases} p^{\sum x_i}(1-p)^{n-\sum x_i} & x \in \{0,1\}^{\times 200} \\ 0 & x \in \mathbb{Z}^{200} \text{ sonst} \end{cases} \\ &= \begin{cases} \prod_{j=1}^n p^{x_j}(1-p)^{1-x_j} & x \in \{0,1\}^{\times 200} \\ 0 & \text{sonst} \end{cases} \end{aligned}$$

Zufallsvariable und Statistik

Def. Eine *Statistik* ist mathematisch einfach eine Zufallsvariable $S : \Omega \to \mathbb{R}$ (Messbarkeit weggelassen).

Warum unterscheidet man 'Zufallsvariale' und 'Statistik'?

Begriffe zwar math. identisch, aber mit unterschiedlicher Interpretation!

Zufallsvariable: Ergebnis von Zufallsexperiment $\longrightarrow$ 'Datensatz'

Statistik: Möglichst aussagekräftige Kenngröße, die auf *mehreren* Datensätzen beruht

Beispiel Statistik: *arith. Mittel von Z.V.:* $\bar{X} = \frac{X_1+\cdots+X_n}{n}$.

Statistik und Punktschätzer

Def. Ein *Punktschätzer* ist eine Staistik, die speziell der Bestimmung eines Parameters θ_i eines statistischen Modells mit Parametern $\theta = (\theta_1, \ldots, \theta_q) \in \Theta$ dient.

Beispiel: Das arithmetische Mittel $\bar{Y}$ von Y is im lin. Modell zwar stets eine wichtige Statistik, im allgemeinen aber *kein* Punktschätzer für $\underline{\beta}$.

Eigenschaften von Punktschätzern

Def.: In einem Stat. Modell heißt ein Punktschätzer $S : \Omega \to \mathbb{R}$ heißt *erwartungstreu* für einen Parameter θ_i, wenn

$$\mathbb{E}_\theta[S] = \theta_i, \;\; \forall \theta = (\theta_1, \ldots, \theta_i, \ldots, \theta_q) \in \Theta \tag{1}$$

$\mathbb{E}_\theta$ ist Erwartungswert bezüglich P_θ

Beispiel: Gegeben sei das univariate Produktmodell $X \sim N(\mu, \sigma^2)$ (oder andere Vert. mit μ als Parameter), dann ist $\bar{X}$ ein erwartungstreuer Schätzer für μ

Denn:

$$\mathbb{E}_{\mu,\sigma^2}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{\mathbb{E}_{\mu,\sigma^2}[X_1] + \cdots + \mathbb{E}_{\mu,\sigma^2}[X_n]}{n} = \frac{n\mu}{n} = \mu$$

Erwartungstreue der Empirischen Varianz

Gegeben sei das univariate Produktmodell $X \sim N(\mu, \sigma^2)$ (oder andere Vert. mit der Varianz σ^2 als Parameter), dann ist $\hat{\sigma}^2$ erwartungstreuer Schätzer für σ^2

$$\hat{\sigma}^2 = \hat{\sigma}^2(X_1, \ldots, X_n) = \frac{1}{n-1}\sum_{j=1}^n (X_j - \bar{X})^2 \tag{2}$$

$$\begin{aligned}(n-1)\mathbb{E}_{\mu,\sigma^2}[\hat{\sigma}^2] &= \sum_{j=1}^n \mathbb{E}_{\mu,\sigma^2}\left[\big((X_j-\mu) - (\bar{X}-\mu)\big)^2\right] \\ &= \sum_{j=1}^n \big\{\mathbb{E}_{\sigma^2,\mu}[(X_j-\mu)^2] + \mathbb{E}_{\sigma^2,\mu}[(\bar{X}-\mu)^2] \\ &\quad - 2\mathbb{E}[(X_j-\mu)(\bar{X}-\mu)]\big\}\end{aligned}$$

Erwartungstreue der Empirischen Varianz II

$$\begin{aligned} &= \sum_{j=1}^{n}\left[\sigma^2 + \frac{\sigma^2}{n} - 2\mathrm{Cov}(X_j, \bar{X})\right] \\ &= (n+1-2)\sigma^2 = (n-1)\sigma^2 \end{aligned}$$

Denn:

$$\begin{aligned} \mathrm{Cov}(X_j, \bar{X}) &= \frac{1}{n}\sum_{k=1}^{n}\mathrm{Cov}(X_j, X_k) \\ &= \frac{1}{n}\sigma^2 \end{aligned}$$

qed.

Normierung $1/(n-1)$ nicht $1/n$!!!!

Bias und asymptotische Erwartungstreue

Def.: Der *Bias* (Verzerrung) eines Schätzers S für θ_i in einem statistischen Modell ist

$$\mathrm{Bias}_\theta(S) = \mathbb{E}_\theta[S] - \theta_i, \quad \theta = (\theta_1, \ldots, \theta_i, \ldots, \theta_q) \in \Theta \tag{3}$$

Der Bias ist der *systematische Schätzfehler*

Def. Ein Schätzer S (oder genauer eine Schätzerfamilie S_n) heißt *asymptotisch Erwartungstreu*, wenn

$$\mathrm{Bias}_\theta(S) = \mathrm{Bias}_\theta(S_n) \longrightarrow 0 \text{ für } n \to \infty \tag{4}$$

Beispiel: $\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2$ is asymptotisch Erwartungstreu für σ^2, denn der Bias ist $-\frac{\sigma^2}{n}$.

Konsistenz von Schätzern

Def.: Ein Schätzer S (oder genauer eine Schätzerfamilie S_n) für den Parameter θ_i heißt *konsistent*, wenn $S_n \to \theta_i$ nach Wahrscheinlichkeit

$$P(|S_n - \theta_i| > \epsilon) \to 0 \text{ für } n \to \infty \quad \forall \epsilon > 0$$

Beispiel: $\bar{X}$ ist konsistent für μ (schwaches Gesetz der gr. Zahlen)

Es gilt: S_n asymptotisch erwartungstreu und $\mathrm{Var}(S_n) \to 0 \Rightarrow S_n$ ist konsistent.

Beweis: $\epsilon > 0$ und $n > n_0$ so dass $|\mathrm{Bias}(S_n)| < \epsilon/2 \Rightarrow$

$$\begin{aligned} P(|S_n - \theta_i| > \epsilon) &\leq P(|S_n - \mathbb{E}[S_n]| + |\mathrm{Bias}_\theta(S_n)| > \epsilon/2 + \epsilon/2) \\ &\leq P(|S_n - \mathbb{E}[S_n]| > \epsilon/2) \leq \frac{4\mathrm{Var}(S_n)}{\epsilon^2} \to 0 \end{aligned}$$

MSE - Mean Square Error

Def.: Der MSE eines Schätzers S (oder genauer einer Schätzefamilie S_n) ist definiert als

$$\mathrm{MSE}_\theta(S) = \mathbb{E}_\theta[(S_n - \theta_i)^2] \tag{5}$$

Es gilt die Bias-Varianz-Zerlegung:

$$\mathrm{MSE}_\theta(S) = \mathrm{Var}_\theta(S) + \mathrm{Bias}_\theta(S)^2 \tag{6}$$

Denn: vgl. Skript Einf. Stoch.

Es gilt: Falls für einen Schätzer S gilt $\mathrm{MSE}(S) \to 0 \Rightarrow S$ ist konsistent.

Denn: Der Schätzer ist dann asymptotisch Erwartungstreu und die Varianz verschwindet für $n \to \infty$

Das Maximum Likelihood Prinzip

- Beobachte Daten $x_1, \ldots, x_n$ aus Zufallsexperiment
- Habe statistisches Modell $P_\theta^{(n)}$, $\theta \in \Theta$, für die Verteilung der n-fachen Wiederholung des Z.E. .
- Unter allen möglichen Modellen $P_\theta^{(n)}$ wähle dasjenige, welches den **beobachteten Daten** $x_1, \ldots, x_n$ die höchste W.-keit zuordnet.

Dies kann als eine Schätzprozedur verstanden werden:

$$\hat{\theta}_{ML} = \hat{\theta}_{ML}(x_1, \ldots, x_n) = \arg\max_{\theta \in \Theta} P_\theta(\{(x_1, \ldots, x_n)\}) \tag{7}$$

Argmax und Argmin

Def.: Das argmax einer Funktion $f : \Theta \to \mathbb{R}$ ist definiert als

$$\arg\max_{\theta \in \Theta} f(\theta) = \{\theta^* \in \Theta : f(\theta^*) \geq f(\theta)\ \forall \theta \in \Theta\} \tag{8}$$

Analog gilt für das argmin

$$\arg\min_{\theta \in \Theta} f(\theta) = \{\theta^* \in \Theta : f(\theta^*) \leq f(\theta)\ \forall \theta \in \Theta\} \tag{9}$$

- Das argmax und argmin sind i.A. Mengen, d.h. mehrere Lösungen (oder keine) sind möglich
- Ist f stetig und Θ kompakt, so gibt es mindestens eine Lösung
- Sind argmin/argmax eindeutig, so identifiziere die Menge mit ihrem einzigen Element.

Maximum Likelihood Schätzer - Definition

Def.: $P_\theta^{(n)}$, $\theta \in \Theta$, sei ein statistisches Modell für das n-fach durchgeführte Zufallsexperiment $X_1, \dots, X_n$.

(i) Sind die X_j diskret verteilt, so heißt eine Z.V. $\hat{\theta}_{ML}$ Maximum-Likelihood-Schätzer, falls

$$\hat{\theta}_{ML} = \hat{\theta}_{ML}(X_1, \dots, X_n) \in \arg\max_{\theta\in\Theta} P_\theta(\{(X_1, \dots, X_n)\}). \tag{10}$$

(ii) Sind die X_j kontinuierlich verteilt mit gemeinsamer Dichte $f(x_1, \dots, x_n|\theta)$, so heißt eine Z.V. $\hat{\theta}_{ML}$ Maximum-Likelihood-Schätzer, falls

$$\hat{\theta}_{ML} = \hat{\theta}_{ML}(X_1, \dots, X_n) \in \arg\max_{\theta\in\Theta} f(X_1, \dots, X_n|\theta). \tag{11}$$

Hier setzen wir nicht unbedingt die unabhängige oder identische Wiederholung der Zufallsexperimente voraus.

Likelihood und Log-Likelihood

Def.: Die **Likelihood** des statistischen Modells $P_\theta^{(n)}$, gegeben die Daten $x_1, \dots, x_n$ ist gegeben durch

$$\mathscr{L}(x_1, \dots, x_n|\theta) = \begin{cases} P_\theta(\{(x_1, \dots, x_n)\}) & \text{falls} X_j \text{ diskret} \\ f(x_1, \dots, x_n|\theta) & \text{falls} X_j \text{ kontinuierlich} \end{cases} \tag{12}$$

Der Logarithmus ist eine streng monoton steigende Funktion $\Rightarrow$ (kontinuierlicher Fall)

$$\arg\max_{\theta\in\Theta} f(X_1, \dots, X_n|\theta) = \arg\max_{\theta\in\Theta} \log\left(f(X_1, \dots, X_n|\theta)\right) \tag{13}$$

Def. Die log-Likelihood ist definiert als der Logarithmus der Likelihood-Funktion.

Unabhängige Wiederholungen

Wir betrachten den Spezialfall, dass $X_1, \ldots, X_n$ unabhängig voneinander sind.

Insbesondere gilt im Produktmodell

$$\log \mathscr{L}(x_1, \ldots, x_n|\theta) = \sum_{j=1}^{n} \log f(x_j|\theta) \tag{14}$$

Maximum Likelihood Gleichungen

Wie löst man das Optimierungsproblem in θ?

- Ableiten in θ, Null setzen und Auflösen (ML-Gleichungen)
- Numerische Optimierungsmethoden

Def.: Die Maximum-Likelihood Gleichungen sind gegeben durch
($k = 1, \ldots, q,\ \theta = (\theta_1, \ldots, \theta_q)$):

$$0 = \frac{\partial}{\partial \theta_k} \log \mathscr{L}(x_j|\theta) = \sum_{j=1}^{n} \frac{\partial}{\partial \theta_k} \log f(x_j|\theta), \tag{15}$$

Die score-Funktion ist gegeben als $l'(y|x, \theta) = \nabla_\theta l(x|y, \theta)$ mit $l(x|\theta) = \log f(x|\theta)$.

Maximum-Likelihood-Gleichungen II

Also können die ML-Gleichungen auch äquivalent über die Score-Funktionen formuliert werden:
Im Produktmodell gilt:

$$0 = \sum_{j=1}^{n} l'(x_j|\theta), \quad l(x,\theta) = \log f(x|\theta),$$

$$l'(x,\theta) = \nabla_\theta l(x,\theta) = \nabla_\theta \log f(x|\theta) = \frac{\nabla_\theta f(x|\theta)}{f(x|\theta)}.$$

Beispiel

$X_j \sim \mathrm{Exp}(\lambda)$ sei exponentialverteilt, $x_1, \ldots, x_n$ Messwerte.

$$f(x|\lambda) = \lambda e^{-\lambda x}, \quad x > 0.$$

$$l(x|\lambda) = \log(\lambda) - \lambda x$$

$$l'(x|\lambda) = \frac{1}{\lambda} - x$$

$$0 = \sum_{j=1}^{n} [\frac{1}{\lambda} - x_j] = \frac{n}{\lambda} - n\bar{x}$$

$$\hat{\lambda}_{ML} = \frac{1}{\bar{x}}$$

KL-Information

Def.: Gegeben sei ein statistisches Modell, das der Einfahheit halber mit kontinuierlichen Dichten $X_j \sim f(x,\theta)$ i.i.d. angenommen wird (Rechnungen für diskrete Diche analog), $\theta \neq \theta' \Rightarrow f(.|\theta) \neq f(.|\theta')$ (als L^1-fkt.)

Die Kullback-Leibler Infomation ist gegeben durch

$$\begin{aligned} K(\theta_0|\theta) &= -\mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X|\theta)}{f(X|\theta_0)}\right)\right] \\ &= -\int \log\left(\frac{f(x|\theta)}{f(x|\theta_0)}\right) f(x|\theta_0)dx \qquad (16) \end{aligned}$$

Satz: $K(\theta_0|\theta) \geq 0$ und Gleichheit gilt genau dann wenn $\theta = \theta_0$.

Jensensche Ungleichung

$\log(t)$ ist strikt konkav $\Rightarrow -\log(t)$ ist strikt konvex.

Lemma: (Jensensche Ungleichung)
Y reelwertige Z.V. F konvex $\Rightarrow$

$$\mathbb{E}[F(Y)] \geq F(\mathbb{E}[Y])$$

Ist F strikt konvex, $F'' > 0$, dann gilt Gleichheit genau dann wenn $Y = c$ fast sicher (Y deterministisch).

Bew.: Siehe Skript Einf. Stochastik.

Beweis eind. Minimum für die KL-Info

Anwendung: $Y = \frac{f(X|\theta)}{f(X|\theta_0)} \Rightarrow$

$$\begin{aligned} K(\theta_0|\theta) &= \mathbb{E}_{\theta_0}[-\log(Y)] \geq -\log(\mathbb{E}_{\theta_0}[Y]) \\ &= -\log\left(\int \frac{f(x|\theta)}{f(x|\theta_0)} f(x|\theta_0)dx\right) \\ &= -\log\left(\int f(x,\theta)\right) = -\log(1) = 0 \end{aligned}$$

Gleichheit gilt genau dann (Jensen) wenn $\frac{f(x|\theta)}{f(x|\theta_0)} = c \Leftrightarrow f(x|\theta) = cf(x|\theta_0)$.

Es muss gelten $c = 1$ (Normiertheit des Integrals über beide W.-keitsdichten), also

$$f(x|\theta) = f(x|\theta_0) \quad \Rightarrow \theta = \theta_0 \quad \textbf{qed.}$$

Konsistenz von Maximum Likelihood

Welche Rolle spielt die KL-Info?

Beobachtung: Sei wieder $P_\theta^{(n)}$ das Produktmodell zu $X_j \sim f(x|\theta_0)$.
D.h. θ_0 ist der wahre Parameter.

$$\begin{aligned} & \arg\max_{\theta\in\Theta} \mathscr{L}(X_1,\ldots,X_n|\theta) \\ = & \arg\max_{theta\in\Theta} \log \mathscr{L}(X_1,\ldots,X_n|\theta) \\ = & \arg\max_{\theta\in\Theta} \frac{1}{n} \log \mathscr{L}(X_1,\ldots,X_n|\theta) \\ = & \arg\max_{\theta\in\Theta} \frac{1}{n} (\log \mathscr{L}(X_1,\ldots,X_n|\theta) - \log \mathscr{L}(X_1,\ldots,X_n|\theta_0)) \\ = & \arg\max_{\theta\in\Theta} \frac{1}{n} \sum_{j=1}^{n} [l(X_j|\theta) - l(X_j|\theta_0)] \\ \stackrel{?}{\to} & \arg\max_{\theta\in\Theta} \mathbb{E}_{\theta_0} [\log f(X|\theta) - \log f(X|\theta_0)] = -K(\theta_0|\theta) = \theta_0 \end{aligned}$$

Mathematische Feinheiten

Warum das Fragezeichen?

Zwar konvergiert die (redefinierte) log-Likelihood-Funktion gegen (minus) die KL-Information für jedes $\theta, \theta_0 \dots$

Die minus KL-Info hat 1-deutiges Maximum in $\theta = \theta_0 \dots$

Aber dürfen wir $\arg\max$ und Grenzwert ($\to$) einfach vertauschen?

Ja, unter geeigneten Voraussetzungen!

Numerische Beispiele

ML für die Weibullverteilung

Betrachten das Weibull-Produktmodell

$$f(x|\theta) = f(x|\eta, m) = \left(\frac{m}{\eta}\right)\left(\frac{x}{\eta}\right)^{m-1} \exp\left\{-\left(\frac{x}{\eta}\right)^m\right\}.$$

$$l(x|\theta) = l(x|\eta, m) = \log\left(\frac{m}{\eta}\right) + (m-1)\log\left(\frac{x}{\eta}\right) - \left(\frac{x}{\eta}\right)^m$$

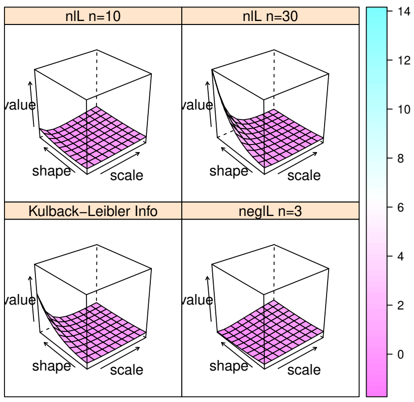
$$\log \mathscr{L}(x_1, \dots, x_n|\eta, m) = \sum_{j=1}^{n} l(x_j|\eta, m)$$

Numerische Lösung

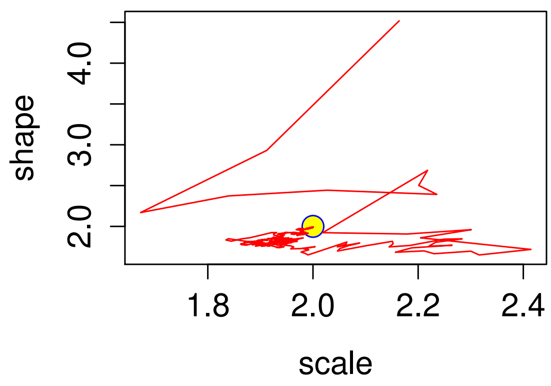
Da die ML-Gleichungen des Weibull-Produktmodells keine geschlossene Lösung haben, verwenden wir ein Verfahren mit numerischer Optimierung:

```
n=100                                            # set number of
                                                 # Wei(2,2) Pseudos
theta0=c(2,2)                                    # set θ0
nlL=function(theta,X) -sum(dweibull(X,           # def.
scale=theta[1],shape=theta[2],log=TRUE))         # negLogLikelihood
X=rweibull(n,scale=theta0[1],shape=theta0[2])    # generate sample
thetaS=c(1,1)                                    # set start as Exp(1)
optim(thetaS,nlL,X=X)                            # calculate estimate
                                                 # by optimization of θ
```


Konvergenz zu KL



Konsistenz in der Simulation I



Hanno Gottschalk

Stochastik für Informatiker – 39 / 39