Stochastik für Info SoSe 2023
Asymptotische Maximum Likelihood Theorie

Hanno Gottschalk

June 21, 2023

Inhaltsverzeichnis der Vorlesung

Fragestellung 3 / 27

Quantifizierung des Schätzfehlers

Bisher wissen wir nun, dass der ML-Schätzer sich 'irgendwann mal' an den wahren Wert annähert.

Da wir in der Praxis keine unendlich langen Messreihen durchführen können, brauchen wir eine Beschreibung des Schätzfehlers $\hat{\theta}_{ML} - \theta_0$.

Der Schäzfehler ist eine Zufallsvariable, d.h. wir können den Fehler über die Verteilung der Z.V. charakterisieren.

Kann man diese Verteilung bestimmen?

Mögliche Verfahren zur Bestimmung des Schätzfehlers

- Lösung der ML-Gleichungen und analytische Berechnung der Dichtefunktion für die Parameter mit Transformationssätzen für W.-keitsdichten
- MC Simulation (auch parametrisches bootstrapping genannt)
- Asymptotische Theorie (wie Zentraler Grenzwertsatz)

Das erste Verfahren ist nur in Spezialfällen (etwa: Lineares Modell) tatsächlich lösbar...
Das MC-Verfahren haben wir schon kennen gelernt (vg. Vorl. 13) - oft mit Implementierungsaufwand und langen Simulationszeiten verbunden – pro Simulation eine nicht lineare Optimierung $\rightarrow$ Stabilitätsprobleme!
Wollen nun die asymptotische Theorie erforschen!

Vorbild Zentraler Grenzwertsatz

'Vorbild aller Schätzer' – das arithmetische Mittel im Produktmodell:

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n).$$

Der Zentrale Grenzwertsatz kann als Aussage über den Schätzfehler verstanden werden:

$$\sqrt{n}(\bar{X}_n - \mu) \longrightarrow N(0, \sigma^2), \quad \sigma^2 = \mathrm{Var}[X]$$

(Konvergenz nach Verteilung)

Wünschen ein ähnliches Resultat für ML-Schätzer

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \longrightarrow N(0, \Sigma).$$

Probleme bei der Formulierung

$\theta_0, \hat{\theta}_{ML}$ sind q-Vektoren, daher brauchen wir einen multivariaten Zentralen Grenzwertsatz.

Dazu müssen wir erstmal verstehen, was eine multivariate Normalverteilung ist $N(\mu, \Sigma)$, $\mu \in \mathbb{R}^q$, $\Sigma \in \mathrm{Mat}_{q\times q}(\mathbb{R})$.

Dazu müssen wir erstmal verstehen, was eine Kovarianzmatrix Σ ist.

Wenn das erledigt ist, wollen wir die Konvergenz der Schätzfehlerverteilung gegen $N(0, \Sigma)$ beweisen und Σ als Funktion von θ_0 berechnen!

Kovarianzmatrizen und Multivariate Normalverteilung 8 / 27

Kovarianzmatrizen

Es sei $X : \Omega \to \mathbb{R}^q$ eine $\mathbb{R}^q$-wertige Zufallsvariable (Zufallsvektor).

Def.: Die Kovarianzmatrix von $X = (X_1, \ldots, X_q)$ ist definert als

$$\Sigma = \mathrm{Cov}[X] = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \cdots & \mathrm{Cov}(X_1, X_q) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_q, X_1) & \cdots & \mathrm{Cov}(X_q, X_q) \end{pmatrix} \in \mathrm{Mat}_{q\times q}(\mathbb{R})$$

Wh.: $\mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$.

Eigenschaften von Kovarianzmatrizen

Es sei Σ die Kovarianzmatrix eines Zufallsvektors X. Dann gilt:

- Σ ist eine symmetrische Matrix $\Sigma' = \Sigma \Leftrightarrow \Sigma_{i,j} = \Sigma_{j,i}$.
- Σ ist positiv semidefinit, d.h. $v'\Sigma v \geq 0 \; \forall v \in \mathbb{R}^q$.

Denn: (i) $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$ ✓

(ii)

$$\begin{aligned} v'\Sigma v &= \sum_{i,j=1}^{q} v_i \mathrm{Cov}(X_i, X_j) v_j \\ &= \mathrm{Cov}\left(\sum_{i=1}^{q} v_i X_i, \sum_{i=1}^{q} v_j X_j\right) = \mathrm{Var}\left[\sum_{i=1}^{q} v_i X_i\right] \geq 0 \quad ✓ \end{aligned}$$

Transformationsverhalten der Kovarianzmatrix

Satz: Sei X ein q-dimensionaler Zufallsvektor mit Erwartungswert $\mu \in \mathbb{R}^q$ und Kovarianzmatrix $\Sigma_X \in \mathrm{Mat}_{q\times q}$.
Sei A eine $n \times q$-Matrix und $Y = AX$ ein n-dimensionaler Z.V. .
Dann ist der Erwartungswert von Y gegeben als $A\mu \in \mathbb{R}^n$ und für die Kovarianzmatrix Σ_Y gilt

$$\Sigma_Y = A\Sigma_X A'$$

Denn:

$$\mathbb{E}[Y_j] = \mathbb{E}[\sum_{k=1}^{q} A_{j,k} X_k] = \sum_{k=1}^{q} A_{j,k} \mathbb{E}[X_k] = \sum_{k=1}^{q} A_{j,k}\mu_k = (A\mu)_j$$

und

$$\begin{aligned} \mathrm{Cov}(Y_i, Y_j) &= \mathrm{Cov}\left(\sum_{k=1}^{q} A_{i,k} X_k, \sum_{k=1}^{q} A_{j,k} X_k\right) \\ &= \sum_{k,l=1}^{q} A_{i,k} A_{j,l} \mathrm{Cov}(X_k, X_l) = (A\Sigma_X A')_{i,j} \end{aligned}$$

Multivariate Normalverteilung

Def.: Es sei Σ eine positiv definite $q \times q$-Matrix (damit auch symmetrisch) und $\mu \in \mathbb{R}^q$.

Dann ist die multivariate Normalverteilung $N(\mu, \Sigma)$ mit Kovarianzmatrix Σ und Erwartungswert μ gegeben durch die Dichte

$$f(x|\mu, \Sigma) = \frac{1}{(2\pi)^{q/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}, x \in \mathbb{R}^q.$$

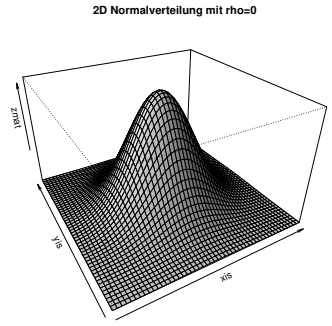
$|\Sigma|$ ist die Determinante von Σ.

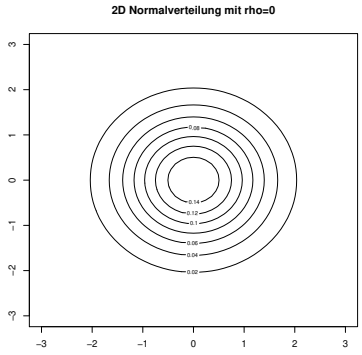
Dass dies eine multivariate W.-keitsdichte ist, wurde in EinfStoch bewiesen (Übung). Trick: Betrachte $X \sim N(0, \mathbf{1})$ multivariat Standardnormalverteilt und $Y = \sqrt{\Sigma}X + \mu$. Dann $Y \sim N(\mu, \Sigma)$.

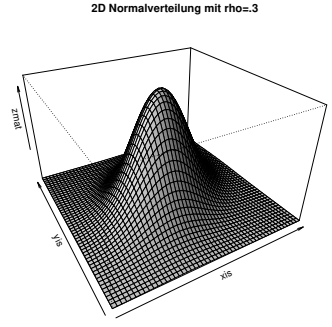
2D Normalverteilung

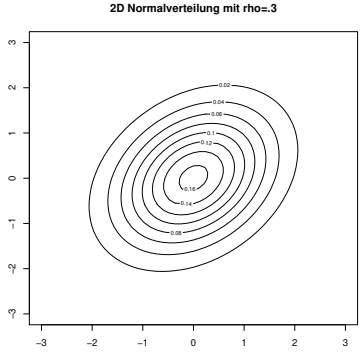
In 2 Dimensionen sid die Kovarianzmatritzen parametrisiert durch $\sigma_1, \sigma_2 > 0$ und $\rho \in (-1, 1)$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad |\Sigma| = \sigma_1^2\sigma_2^2(1-\rho^2) \tag{1}$$









Asymptotische Normalität

Def.: Eine (multivariate) Schätzstatistik $\hat{\theta}$ für θ_0 in $\mathbb{R}^q$ heißt asymptotisch normal, falls $\exists a_n \to \infty$ und Σ positiv definit so dass

$$a_n(\hat{\theta} - \theta_0) \longrightarrow N(0, \Sigma) \text{ nach Verteilung}$$

In diesem Fall schreibe auch $a_n(\hat{\theta} - \theta_0) \sim_a N(0, \Sigma)$.

Beispiel: Im Produktmodell mit quadratintegrierbarem X gilt $\sqrt{n}(\bar{X} - \mu) \sim_a N(0, \sigma^2)$

(Zentraler Grenzwertsatz)

Fisher-Information 15 / 27

Motivation

Wir sind schon guten Mutes (wegen der Simulationen) dass ML-Schätzer (und auch MM-Schätzer) asymptotisch normal sind.

Wir brauchen noch eine gute Idee für die Kovarianzmatrix.

- Die Kovarianzmatrix soll eine klar definierte Formel haben.
- Die Kovarianzmatrix soll möglichst klein sein (geringe Fluktuation)
- Die Kovarianzmatrix soll bequem aus den Daten schätzber sein, auch wenn wir sie nicht explizit berechnen können.

Fisher-Information

Def.: Gegeben sei ein statistisches Modell P_θ mit multivariater Dichte $f(\underline{x}|\theta)$ (diskreter Fall Analog).

$$l'(x|\theta) = \nabla_\theta l(x|\theta) = \nabla_\theta \log f(x|\theta) \text{ score-Funktion}, x \in \mathbb{R}^{nd}.$$

und $X \sim f(x, \theta)$ Dann ist die Fisher-Information $I(\theta_0)$, $\theta_0 \in \Theta$, die $q \times q$-Matrix definiert durch

$$I_{i,j}(\theta_0) = \mathbb{E}_{\theta_0}[l_i'(X, \theta_0) l_j'(X, \theta_0)] = \int l_i'(x, \theta_0) l_j'(x, \theta_0) f(x|\theta_0)\, dx \tag{2}$$

Eigenschaften der Fisher-Info

Lemma: Unter Voraussetzung (*)- s.u.-:

$$I(\theta_0) = \mathrm{Cov}_{\theta_0}[l'(X|\theta_0)]$$

Beweis: Zu zeigen $\mathbb{E}_{\theta_0}[l'(X|\theta_0)] = 0$

$$\Rightarrow \mathbb{E}_{\theta_0}[l_i'(X, \theta_0) l_j'(X, \theta_0)] = \mathrm{Cov}_{\theta_0}(l_i'(X, \theta_0), l_j'(X, \theta_0)).$$

$$\begin{aligned} 0 &= \nabla_{\theta_0} 1 = \nabla_{\theta_0} \int f(x|\theta_0) dx \\ &\overset{(*)}{=} \int \nabla_{\theta_0} f(x|\theta_0) dx = \int \frac{\nabla_\theta f(x|\theta_0)}{f(x|\theta_0)} f(x|\theta_0) dx \\ &= \int l'(x|\theta_0) f(x|\theta_0) dx = \mathbb{E}_{\theta_0}[l'(X|\theta_0)]. \end{aligned}$$

Eigenschaften der Fisher-Info II

Satz: Unter Voraussetzung (*) - s.o. - und (**)- s.u.-

$$I(\theta_0) = -\mathbb{E}_{\theta_0}[l''(X|\theta_0)] \;\; l''_{i,j}(x|\theta) = \frac{\partial^2}{\partial\theta_i \partial\theta_j} l(x|\theta). \tag{3}$$

Beweis:

$$\begin{aligned} 0 &= \nabla^2_{\theta_0} 1 = \nabla_{\theta_0} \int l'(x|\theta_0) f(x|\theta_0) dx \\ &\overset{(**)}{=} \int \nabla_{\theta_0}[l'(x|\theta_0) f(x|\theta_0)] dx \\ &= \int l''(x|\theta_0) f(x|\theta_0) dx + \int l'(x|\theta_0) l'(x|\theta_0) f(x|\theta_0) dx \\ &= \mathbb{E}_{\theta_0}[l''(X|\theta_0)] + I(\theta_0). \end{aligned}$$

qed.

Fisher-Information im Produktmodell

Lemma: Es sei $P_\theta^{(n)}$, $f(x|\theta) = \prod_{j=1}^n f(x_j|\theta)$ das Produktmodell, $I^{(n)}(\theta_0)$ die Fisher-Info des Produktmodells und $I^{(1)}(\theta_0)$ die Fischer-Info für $X_j \sim f(x_j|\theta)$.
Dann gilt

$$I^{(n)}(\theta_0) = nI^{(1)}(\theta_0)$$

Beweis:

$$l'(x|\theta) = \nabla_\theta \log f(x|\theta) = \sum_{j=1}^n \nabla_\theta \log f(x_j|\theta) = \sum_{j=1}^n l'(x_j|\theta) \;\; \Rightarrow$$

$$I^{(n)}(\theta_0) = \mathrm{Cov}_{\theta_0}[l'(X|\theta_0)] = \sum_{j=1}^n \mathrm{Cov}_{\theta_0}[l'(X_j|\theta_0)] = nI^{(1)}(\theta_0)$$

qed.

Asymtotische Normalität für ML - der Satz:

Satz Gegeben ein Produktmodell mit konsistenetm ML Schätzer $\hat{\theta}_{ML}$, bei dem auch die Annahmen (*) und (**) gelten.
Dann ist der Schätzfehler $\hat{\theta} - \theta_0$ asymptotisch normal, und es gilt

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \sim_a N(0, I^{(1)}(\theta_0)^{-1}). \tag{4}$$

Bemerkung: Auch in sehr vielen nicht-Produktmodellen gilt annäherungsweise

$$(\hat{\theta}_{ML} - \theta_0) \sim N(0, I(\theta_0)^{-1}).$$

D.h. wenn die Fisher-Info gegen unendlich geht, verschwindet das Gewackel der Parameterschätzer!

Asymtotische Normalität für ML - der Beweis

Hier nur die Beweisidee: Taylor-Entwicklung der ML-Gleichungen um θ_0. Schreibe $I(\theta_0) = I^{(1)}(\theta_0)$:

$$\sum_{j=0} l'(X_j|\hat{\theta}_{ML}) = 0$$

$$0 = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l'(X_i|\theta_0) + \left(\frac{1}{n} \sum_{j=1}^{n} l''(X_i|\theta^*) \right) \sqrt{n}(\hat{\theta}_{ML} - \theta_0)$$

$$0 = \underbrace{\frac{1}{\sqrt{n}} \sum_{j=1}^{n} l'(X_i|\theta_0)}_{\longrightarrow N(0, I(\theta_0)) \text{ (ZGWS)}} + \underbrace{\left(\frac{1}{n} \sum_{j=1}^{n} l''(X_i|\theta^*) \right)}_{\longrightarrow \mathbb{E}_{\theta_0}[l''(X|\theta_0)] = -I(\theta_0)} \sqrt{n}(\hat{\theta}_{ML} - \theta_0)$$

$$\begin{aligned} \sqrt{n}(\hat{\theta}_{ML} - \theta_0) &= -\left(\frac{1}{n} \sum_{j=1}^{n} l''(X_i|\theta^*) \right)^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} l'(X_i|\theta_0) \\ &\longrightarrow N(0, I(\theta_0)^{-1} I(\theta_0) I(\theta_0)^{-1}) = N(0, I(\theta_0)^{-1}) \end{aligned}$$

Schätzung der Fisher-Inormationsmatrix

Problem:

- $I(\theta_0)$ hängt von dem unbekannten Parameter θ_0 ab...
- Die Formeln für die Fisher-Information sind evtl. kompliziert ... und ich bin zu faul, das auszurechnen...

Lösung:

- Wähle $I(\hat{\theta}_{ML})$ als Schätzer für $I(\theta_0)$ (ist unter geeigneten Vorr. konsistent) .
- Nütze aus, dass nach dem Gesetz der gr. Zahlen

$$I^{(n)}(\theta_0) \approx -n\mathbb{E}_{\theta 0}[l''(x|\hat{\theta}_{ML})] \approx -\sum_{j=1}^{n} l''(X_j|\hat{\theta}_{ML})$$

Konsequenzen für die Praxis von ML

Da man ohnehin schon $-\log \mathscr{L}$ minimiert, kann man sich vom Optimierer auch gleich die Hessematrix am Optimum mit ausgeben lassen!

$$-\sum_{j=1}^{n} l''(X_j|\hat{\theta}_{ML}) = \mathrm{Hess}(-\log \mathscr{L})$$

Oft arbeitet der Optimierer ja sowieso mit der Hessematrix!

Hier bekommt man endlich mal was umsonst!

Algorithmisches Vorgehen

- Minimiere die neg. log Likelihood numerisch
- Lasse die Hessematrix ausgeben
- Invertiere die Hessematrix H, um die multivariate Parameter-Kovarianz zu schätzen.
- Die Standardabweichung für Parameter θ_j ergibt sich als $\sqrt{(H^{-1})_{j,j}}$, $j = 1, \ldots, q$.

So geht man auch vor, wenn man kein Produktmodell vorliegen hat!