R. Höpfner: Schätzer und Tests, Sommersemester 2012

Übungsblatt 4

Abgabe per mail an `hoepfner@uni-mainz.de` bis **SA 09.05.12**

Besprechung voraussichtlich DI 12.06.12

Aufgabe 4.1 (Eine ungewöhnliche Likelihoodfläche): Man betrachte ein 'gestörtes' Lokations- und Skalenmodell für iid Beobachtungen, in dem die Verteilung der Einzelbeobachtung durch ihre Lebesgue-Dichte

$$f_{\mu,\sigma}(x) \;=\; (1-\alpha)\,\frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \;+\; \alpha\,\frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}(x-\mu)^2} \quad , \quad x \in I\!R$$

gegeben sei. Dabei seien $\mu \in I\!R$ und $\sigma > 0$ die unbekannten Parameter der Verteilung; der Mischungsparameter $0 < \alpha < \frac{1}{2}$ sei fest und bekannt (LeCam 1990 schlägt boshafterweise vor, $\alpha = 10^{-10^{137}}$ zu wählen).

a) Man generiere einen kleinen Datensatz von Beobachtungen $X_1, \ldots, X_n$ (etwa: $n = 5$) unter $\mu_0 = 0.5$, $\sigma_0 = 0.5$, $\alpha = 0.3$. Man visualisiere die log-likelihood Fläche gegeben diese Beobachtungen über einem geeigneten Gitter von (μ, σ)-Werten. Man benutze dabei σ-Werte, die immer näher an 0 herankommen, um in der log-likelihood Fläche die für $(\mu, \sigma) \to (X_i, 0^+)$ entstehenden Singularitäten sichtbar zu machen.

b) Man überlege sich aus der Gestalt der Produktdichten, warum in diesem Modell die log-likelihood Funktion basierend auf Beobachtungen $X_1, \ldots, X_n$ Singularitäten zu den Randpunkten

$$(\mu, \sigma) \in \big\{(X_1, 0^+), \ldots, (X_n, 0^+)\big\}$$

hin besitzen muss. Was folgt daraus für Maximum-Likelihood-Schätzverfahren?

c) In welcher Weise ändert sich die log-likelihood Fläche bei wachsender Zahl n von Beobachtungen ?

Hinweis: Man runde in a) die Beobachtungen $X_1, \ldots, X_n$ auf eine Nachkommastelle, und verwende ein μ-Gitter mit ebenfalls auf eine Nachkommastelle gerundeten Werten ... (warum?)

(Abgabe: nur a): Programm und ein guter 'persp'-plot als pdf)

Aufgabe 4.2 (Konfidenzintervalle I): Man betrachte das von der Doppelexponentialverteilung erzeugte Lokations- und Skalenmodell

$$(\diamond) \qquad \left(\mathbb{R}^n \,,\, \mathcal{B}(\mathbb{R}^n)) \,,\, \left\{ P_{m,c}^n = \bigotimes_{i=1}^n P_{m,c} : m \in \mathbb{R} \,,\, c > 0 \right\} \right) \;, \quad P_{m,c}(dx) \;:=\; \frac{1}{2c}\, e^{-|\frac{x-m}{c}|}\, dx$$

aus Aufgabe 3.2 a). Man betrachte Schätzer

$$\widehat{m}_n \;:=\; median(X_1, \ldots, X_n)$$

für den Lokationsparameter $m \in \mathbb{R}$ und

$$\begin{aligned} \widehat{c}_n \;&:=\; \frac{1}{n} \sum_{i=1}^n |X_i - \widehat{m}_n| \;, \\ \widetilde{c}_n \;&:=\; median(Y_1, \ldots, Y_n) \quad \text{wobei} \quad Y_i := |X_i - \widehat{m}_n| \end{aligned}$$

für den Skalenparameter $c > 0$, wie in den Aufgaben 3.3 und 3.4.

a) Man gebe in Analogie zu 1.16 A) der Vorlesung drei Typen von Konfidenzintervallen für den unbekannten Lokationsparameter an, die im Modell $(\diamond)$ zu vorgegebenem $\beta \in (0,1)$ den Konfidenzkoeffizient β liefern. Diese drei basiere man auf

$$(\widehat{m}_n, \widehat{c}_n) \quad , \quad (\widehat{m}_n, \widetilde{c}_n) \quad , \quad \left(\overline{X}_n, \sqrt{\widetilde{S}_n^2} \right)$$

als Schätzer für das Paar $(m, c) \in \mathbb{R} \times (0, \infty)$. Warum sind alle drei Paare im betrachteten Modell sinnvolle Schätzer?

b) Man scheibe eine Funktion im Argument n, mit der man sich die Verteilungen

$$r_n \;:=\; \mathcal{L}\left(\frac{\sqrt{n}\,\widehat{m}_n}{\widehat{c}_n} \mid P_{0,1}^n \right) \;, \quad v_n \;:=\; \mathcal{L}\left(\frac{\sqrt{n}\,\widehat{m}_n}{\widetilde{c}_n} \mid P_{0,1}^n \right) \;, \quad w_n \;:=\; \mathcal{L}\left(\frac{\sqrt{n}\,\overline{X}_n}{\sqrt{\widetilde{S}_n^2}} \mid P_{0,1}^n \right)$$

und ihre Quantile approximativ verschafft, in der folgenden Weise: aufgrund von $N = 10000$ Simulationsläufen mit je n doppelexponentialverteilten iid ZV bilde man eine empirische Verteilung für jede der drei genannten Grössen, ermittle empirische Quantile, und zeichne diese in ein gutes Histogramm ein. Idealerweise setze man in gleichem Massstab drei Graphiken in Querformat auf eine Seite, oben ein gutes Histogramm der empirischen Verteilung von r_n mit empirischen Quantilen, mittig dasselbe für v_n, unten für w_n (beachte: von einer Verteilung t_{n-1} der Statistik w_n kann hier natürlich keine Rede mehr sein: man ist nicht mehr im Normalverteilungsmodell, es gibt keine Sätze über die Verteilung des Paares (empirischer Mittelwert, empirische Varianz) in Lokations- und Skalenmodellen, die von einem allgemeinen F mit endlichen zweiten Momenten

erzeugt werden: also ist es sinnvoll, die genannten Verteilungen empirisch zu ermitteln!). Man setze dabei den Masstab so an, dass obere und untere 1%, 2.5%, 5% Quantile eingezeichnet sind und optisch verglichen werden können. Auch erwähne man in einer Überschrift, dass es sich um Ergebnisse für das von der Doppelexponentialverteilung erzeugte Lokations- und Skalenmodell handelt.

c) Anhand des Datensatzes

6.0490067,	20.0324319,	8.0917892,	5.3001702,	8.4231110,	8.4178589,
7.9412088,	6.0004114,	7.6231672,	3.2640157,	7.9500893,	7.1471330,
−3.7385962,	13.1849629,	15.3080767,	6.6531491,	9.1205368,	0.9610968,
4.9916255,	11.5371074,	3.1072696,	5.7450847,	5.0906057,	9.5253568,
4.7945269,	4.0560317,	10.1374325			

vergleiche man die drei nach a) und b) gebildeten Konfidenzintervalle für den unbekannten Lokationsparameter $m \in \mathbb{R}$. Man simuliere sich selbst weitere Datensätze (variieren von n, m, c ...) und führe auf diesen denselben Vergleich aus. Welches Verfahren ist vorzuziehen?

(Abgabe: zu b) und c): Programm und Graphiken)

Aufgabe 4.3 (Konfidenzintervalle II): Man betrachte das Modell ($\diamond$) aus Aufgabe 4.2.

a) Man gebe in Analogie zu 1.16 B) der Vorlesung drei Typen von Konfidenzintervallen für den unbekannten Skalenparameter an, die zu vorgegebenem $\beta \in (0,1)$ den Konfidenzkoeffizient β liefern. Diese drei basiere man auf $\widehat{c}_n$, $\widetilde{c}_n$ und $\sqrt{\widetilde{S}_n^2}$.

e) Man schreibe eine Funktion im Argument n, mit der man sich die Verteilungen

$$\mathcal{L}\left(\widehat{c}_n \mid P_{0,1}^n\right) \ , \quad \mathcal{L}\left(\widetilde{c}_n \mid P_{0,1}^n\right) \ , \quad \mathcal{L}\left(\sqrt{\widetilde{S}_n^2} \mid P_{0,1}^n\right)$$

und ihre Quantile approximativ verschafft, und erstelle Graphiken ähnlich wie in Aufgabe 4.2 b).

f) Anhand des Datensatzes aus Aufgabe 4.2 c) vergleiche man die drei gebildeten Konfidenzintervalle für den unbekannten Skalenparameter $c > 0$. Man simuliere sich selbst weitere Datensätze (variieren von n, m, c ...) und führe auf diesen denselben Vergleich aus.

(Abgabe: zu b) und c): Programm und Graphiken)

Aufgabe 4.4 (Konfidenzintervalle III): Man betrachte man nun ein Lokations- und Skalenmodell, das von einer Verteilung F erzeugt wird, wobei

i) F endliche erste Momente besitzt, aber keine endlichen zweiten Momente;

ii) F nicht einmal endliche ersten Momente besitzt.

Welche der in den Aufgaben 4.2 und 4.3 benutzten Schätzer bleiben sinnvoll unter i), und welche bleiben sinnvoll unter ii)? Natürlich muss man dann in Abhängigkeit von F die Verteilungen dieser Schätzer jeweils neu simulieren ...