

# Statistische Methoden der Datenanalyse

Markus Schumacher, Stan Lai, Florian Kiss

## Übung VIII

17.12.2013, 18.12.2013

### Anwesenheitsaufgaben

#### Aufgabe 41 *Erweiterter Maximum Likelihood Fit für eine Signal- und Untergrundverteilung*

In dieser Übung werden wir einen Beispieldatensatz betrachten, welcher aus zwei verschiedenen Arten von Ereignissen besteht: Signalereignisse, welche in einer Gaussverteilung  $f_s(x)$  vorliegen und Untergrundereignisse, welche nach einer Exponentialfunktion  $f_b(x)$  verteilt sind. Aufgabe ist es, die erweiterte Maximum Likelihood Methode zu benutzen, um die Anzahl von Signal- und Untergrundereignissen aus dem Beispieldatensatz zu errechnen.

Betrachten Sie die WDF  $f(x; \mu_s, \mu_b)$  für die Signal- und Untergrundverteilung:

$$f(x; \mu_s, \mu_b) = \frac{\mu_s}{\mu_s + \mu_b} f_s(x) + \frac{\mu_b}{\mu_s + \mu_b} f_b(x)$$

wobei entsprechend  $\mu_s$  und  $\mu_b$  für die Anzahl von Signal- und Untergrundereignissen stehen. Um das Beispiel zu vereinfachen wird angenommen, dass die Parameter der Signal- und Untergrund-WDF bekannt sind. Daher sind  $f_s$  und  $f_b$  Funktionen nur einer Variablen  $x$ .

Falls die Gesamtzahl an Ereignissen poissonverteilt ist, ergibt sich für die WDF:

$$P(n; \mu_s, \mu_b) = \frac{(\mu_s + \mu_b)^n}{n!} \exp(-(\mu_s + \mu_b))$$

Daher ist die log-Likelihoodfunktion gegeben durch

$$\ln \mathcal{L} = -(\mu_s + \mu_b) + \sum_{i=1}^n \ln[(\mu_s + \mu_b) f(x_i; \mu_s, \mu_b)]$$

Diese Funktion sollen Sie maximieren, um die Parameter der Signal- und Untergrundnormalisierung zu finden.

Zuerst muss ein Datensatz mit der Signal- und der Untergrund-WDF erzeugt werden. Dies wurde bereits in `/home/slai/StatisticsCourse/PS8/aufgabe41_i.C` vorbereitet. Kopieren Sie sich dieses Skript in ein beliebiges Verzeichnis und lassen es laufen. Es erledigt die folgenden Aufgaben:

- (i) Eine TF1 Funktion, welche die Summe einer Gauss- und einer Exponentialfunktion darstellt wird bereitgestellt:

$$f(x; \mu, \sigma, \tau, \mu_s, \mu_b) = \frac{\mu_s}{\mu_s + \mu_b} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) + \frac{\mu_b}{\mu_s + \mu_b} \frac{1}{\tau} \exp\left(-\frac{x}{\tau}\right)$$

in dem Intervall  $0.0 < x < 100.0$ . Die Parameter werden auf  $\mu = 5.0$ ,  $\sigma = 0.5$  und  $\tau = 10.0$  gesetzt.

- (ii) Die TF1-Funktion `GetRandom` wird benutzt, um einen Satz von 1000 Werten der obigen Verteilungsfunktion zu erhalten und diese in ein Array gefüllt. Ein Array ist eine feste Aneinanderreihung von (in diesem Fall) Fließkommazahlen. Um auf das  $i$ -te Element zuzugreifen, wird einfach `data[i]` benutzt.
- (iii) Die Daten werden weiterhin in ein Histogramm gefüllt und graphisch dargestellt.

Als nächstes führen Sie nun selbst einen erweiterten Maximum Likelihood Fit auf dem erstellten Datensatz durch. Ergänzen Sie hierzu das Skript um folgende Komponenten (fügen Sie die Komponenten am Ende des Skriptes hinzu):

- (iv) Definieren sie eine weitere TF1 Funktion, wie oben, der Form  $f(x; \mu, \sigma, \tau, \mu_s, \mu_b)$ . Benutzen Sie die TF1 Funktion `TF1::FixParameter(Int_t parNum, Double_t value)`, um die Funktionsparameter auf die Werte  $\mu = 5.0$ ,  $\sigma = 0.5$  und  $\tau = 10.0$  zu setzen (man könnte diese Parameter variabel lassen, jedoch müsste man dann einen fünfdimensionalen Fit durchführen!).
- (v) Erstellen Sie, analog zu Aufgabe 36, eine Schleife über mögliche Werte des Parameters  $\mu_s$ . Definieren Sie innerhalb dieser Schleife eine weitere Schleife über mögliche Werte von  $\mu_b$ . Sinnvolle Bereiche für  $\mu_s$  und  $\mu_b$  sind zwischen 450 und 550 mit 50 äquidistanten Schritten.
- (vi) Setzen Sie innerhalb der Schleife die Parameter der neuen TF1-Funktion auf die gerade aktuellen Werte von  $\mu_s$  und  $\mu_b$  mittels der `SetParameter(int parnumber, float parwert)`-Funktion.
- (vii) Erstellen Sie für jeden Wert von  $\mu_s$  und  $\mu_b$  eine Schleife über das Array mit den generierten Daten und berechnen Sie den aufsummierten log-Likelihood Wert. Hier kann es nützlich sein, mit der TF1 Funktion `Eval(Double_t x)` den Wert der Funktion am Punkt  $x$  zu berechnen.
- (viii) Befüllen Sie ein `TGraph2D` Objekt mit den aufsummierten log-Likelihood Werten für jeden Wert von  $\mu_s$  und  $\mu_b$ .
- (ix) Zeichnen Sie den Graphen und schätzen Sie damit die Parameter  $\hat{\mu}_s$  und  $\hat{\mu}_b$  und deren Standardabweichungen  $\sigma_{\hat{\mu}_s}$  und  $\sigma_{\hat{\mu}_b}$  ab.
- (x) Vergleichen Sie die Werte Ihres Fits mit denen des ROOT fitting Paketes (benutzen Sie `dataHist.Fit("fitFunk2", "LI");`). Definieren Sie sich zuvor `fitFunk2` als separate Funktion und fixieren Sie wieder  $\mu = 5.0$ ,  $\sigma = 0.5$  und  $\tau = 10.0$ . Stellen Sie diese Funktion wie folgt bereit:

```
TF1 fitFunk2 = TF1("fitFunk2",
  "([0]*1.0/TMath::Sqrt(2*TMath::Pi()*[2]*[2])
  *exp(-0.5*((x-[1])/[2])**2)
  + [3]*(1.0/[4])*exp(-x/[4]))", 0.0, 100.0);
fitFunk2.FixParameter(1, 5.0);
fitFunk2.FixParameter(2, 0.5);
fitFunk2.FixParameter(4, 10.0);
fitFunk2.SetParameter(0, 1000.0);
fitFunk2.SetParameter(3, 1000.0);
```

Beachten Sie, dass die Fitparameter 0 und 3 mit einem Faktor 10 zu multiplizieren sind, da die Binbreite nicht 1 ist.

# Hausaufgaben

## Aufgabe 42 Kombination von Messungen mit der Methode der kleinsten Quadrate

11 Punkte

Es ist möglich, einen Spezialfall der Methode der kleinsten Quadrate zu benutzen, um Messungen von derselben Messgröße zu kombinieren. Betrachten Sie  $N$  Messungen,  $y_i$ , welche den wahren, aber unbekanntem Wert  $\lambda$  bestimmen sollen. Jede Messung  $y_i$  hat einen geschätzten Fehler von  $\sigma_i$ . Da  $\lambda$  für alle Ereignisse konstant ist, ergibt sich die  $\chi^2$ -Variable zu:

$$\chi^2(\lambda) = \sum_{i=1}^N \frac{(y_i - \lambda)^2}{\sigma_i^2}$$

- (i) Wenn jedoch die Messungen von  $y_i$  nicht unabhängig sind, sondern eine Korrelation, gegeben durch die Kovarianzmatrix  $V$ , besitzen, ergibt sich:

$$\chi^2(\lambda) = \sum_{i,j=1}^N (y_i - \lambda)(V^{-1})_{ij}(y_j - \lambda)$$

Zeigen Sie, dass der Schätzer der Methode der kleinsten Quadrate (dieser minimiert  $\chi^2(\lambda)$ ) für  $\lambda$  gegeben ist durch

$$\hat{\lambda} = \sum_{i=1}^N w_i y_i$$

wobei die Gewichtungen  $w_i$  gegeben sind durch

$$w_i = \frac{\sum_{j=1}^N (V^{-1})_{ij}}{\sum_{k,l=1}^N (V^{-1})_{kl}},$$

(offensichtlich ist die Summe über alle  $w_i$  gleich 1) und dass die Varianz gegeben ist durch

$$V[\hat{\lambda}] = \sum_{i,j=1}^N w_i V_{ij} w_j.$$

Tipp: Benutzen Sie Fehlerfortpflanzung!

- (ii) Betrachten Sie jetzt zwei Messungen,  $y_1$  und  $y_2$ , mit einer zugehörigen Kovarianzmatrix

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

mit dem Korrelationskoeffizienten  $\rho = V_{12}/(\sigma_1\sigma_2)$ . Zeigen Sie durch Berechnung des Inversen der Kovarianzmatrix  $V^{-1}$ , dass der Schätzer für  $\lambda$  gegeben ist durch

$$\hat{\lambda} = w y_1 + (1 - w) y_2$$

mit

$$w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}.$$

- (iii) Zeigen Sie, dass die Varianz von  $\hat{\lambda}$  gegeben ist durch

$$\frac{1}{V[\hat{\lambda}]} = \frac{1}{1 - \rho^2} \left[ \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - \frac{2\rho}{\sigma_1\sigma_2} \right] = \frac{1}{\sigma^2}$$

und zeigen Sie dass folglich

$$\frac{1}{\sigma^2} - \frac{1}{\sigma_1^2} \geq 0$$

was bedeutet, dass die Kombination von zwei Messungen immer zu einer Verbesserung der Varianz des Schätzers führt.

Betrachten Sie eine Anpassung mittels der Methode der kleinsten Quadrate an ein Histogramm mit Einträgen  $y_i$  in Bins  $i = 1, \dots, N$  und vorhergesagten Werten

$$\lambda_i(\vec{\theta}) = n \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\theta}) dx,$$

wobei die WDF  $f(x; \vec{\theta})$  von unbekanntem Parametern  $\vec{\theta}$  abhängt. Nehmen Sie an, dass man die Gesamtanzahl  $n$  ersetzt durch einen Parameter  $\nu$ , sodass

$$\lambda_i(\vec{\theta}) = \nu \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\theta}) dx,$$

und diesen simultan mit den anderen Parametern anpasst, indem man die Größe

$$\chi^2(\vec{\theta}, \nu) = \sum_{i=1}^N \frac{(y_i - \lambda_i(\vec{\theta}, \nu))^2}{\sigma_i^2}$$

minimiert.

- (i) Zeigen Sie, dass für die Wahl  $\sigma_i^2 = \lambda_i$  sich der Schätzer

$$\hat{\nu}_{\text{KQ}} = n + \frac{\chi_{\min}^2}{2}$$

ergibt.

- (ii) Zeigen Sie, dass sich für  $\sigma_i^2 = y_i$  (modifizierte kleinste Quadrate) der Schätzer

$$\hat{\nu}_{\text{MKQ}} = n - \chi_{\min}^2$$

ergibt.

**Aufgabe 44** *Unsicherheiten in der Abschätzung von Parametern*

Die Lösungen für die Schätzer einer Anpassung auf der Basis der Methode der kleinsten Quadrate  $\hat{\vec{\theta}}$  können in Matrixform geschrieben werden als

$$\hat{\vec{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1} \vec{y}.$$

Zeigen Sie durch Anwenden der allgemeinen Formel für Fehlerfortpflanzung, dass die Kovarianzmatrix geschrieben werden kann als

$$V(\hat{\vec{\theta}}) = (A^T V^{-1} A)^{-1}$$

mit  $V = V(\vec{y})$ .