Albert-Ludwigs-Universität Freiburg Wintersemester 2011/2012

Statistische Methoden der Datenanalyse

Markus Schumacher

Übung VII

Markus Warsinsky

12.12.2011

Anwesenheitsaufgaben

Aufgabe 43 *'Binned' Maximum Likelihood Fit zu e + e− mit Monte Carlo Statistik*

Der differentielle Wirkungsquerschnitt für $e^+e^- \to \mu^+\mu^-$ folgt der theoretischen Verteilung

$$\frac{\mathrm{d}\sigma}{\mathrm{d}(\cos\theta)\mathrm{d}\phi} \propto 1 + \alpha\cos\theta + \beta\cos^2\theta, \tag{1}$$

wobei θ und ϕ für die Winkel des μ^+ Teilchens in Kugelkoordinaten stehen. Der Winkel θ wird relativ zur z-Achse und der Winkel ϕ relativ zur x-Achse gemessen. In der Übung sollen Sie einen 'Binned' Maximum Likelihood Fit durchführen, um die Parameter zu finden, welche zur Erstellung der Monte Carlo Ereignisse benutzt wurden. Ein solcher simulierter Datensatz befindet sich in `/home/warsinsk/sd_wise1112/ueb7/MuMuEvGen.root`.

Um eine 'Binned' Maximum Likelihood Statistik zu erstellen betrachte man ein Histogramm mit N Bins, wobei jedes n_i Einträge $\vec{n} = (n_1,\ldots,n_N)$ enthält. Für eine Wahrscheinlichkeitsdichtefunktion (WDF) $f(x;\vec{\theta})$ ist der Erwartungswert gegeben durch $\vec{\nu} = (\nu_1,\ldots,\nu_N)$, wobei $\nu_i(\vec{\theta})$ gegeben ist durch das Integral der WDF über die Breite des Bins

$$\nu_i(\vec{\theta}) = n_{tot}\int_{x_i^{min}}^{x_i^{max}} f(x;\vec{\theta})dx \tag{2}$$

wobei ${x_i}^{min}$ und ${x_i}^{max}$ für die Grenzen jedes Bins i stehen.

Daher ist die gemeinsame WDF, $f_g(n;\vec{\nu})$, gegeben durch die Multinomialverteilung

$$f_g(n;\vec{\nu}) = \frac{n_{tot}!}{n_1!\ldots n_N!}\left(\frac{\nu_1}{n_{tot}}\right)^{n_1}\ldots\left(\frac{\nu_N}{n_{tot}}\right)^{n_N} \tag{3}$$

und demzufolge ist die log-Likelihoodfunktion für die gemeinsame WDF gegeben durch

$$\ln\mathcal{L}(\vec{\theta}) = \sum_{i=1}^{N} n_i \ln\nu_i(\vec{\theta}) \tag{4}$$

Dies ist eine Funktion, welche maximiert werden muss, um die zur Erstellung des Monte Carlo Datensatzes verwendeten Parameter $\vec{\theta}$ zu finden. Beachten Sie diesbezüglich, dass n_{tot} nicht in der Berechnung der log-Likelihoodfunktion benötigt wird, da es sich hierbei um eine additive Konstante handelt.

Um die Parameter Ihrer Monte Carlo Verteilung zu berechnen gehen Sie folgende Schritte durch:

(i) Zum Berechnen der Variablen $\cos\theta$ findet sich in `/home/warsinsk/sd_wise1112/ueb7/aufgabe43_i.C` ein ROOT-Skript. Kopieren Sie sich dieses Skript in ein Verzeichnis Ihrer Wahl. Öffnen Sie es in einem Texteditor, lassen es laufen, und versuchen Sie die Funktionsweise zu verstehen.

(ii) Modifizieren Sie dieses Skript nun so, dass Sie die Werte von $\cos\theta$ in ein Histogramm mit 200 Bins zwischen -1 und +1 einfüllen. Stellen Sie das Histogramm am Ende des Skripts graphisch auf dem Bildschirm dar.

(iii) Definieren Sie eine `TF1` Funktion der Form

$$f(x;\alpha,\beta) = \frac{1+\alpha x+\beta x^2}{2+\frac{2\beta}{3}}, \tag{5}$$

welche in dem Maximum Likelihood Fit benutzt werden soll. Dies können Sie mit folgendem Befehl bewerkstelligen:

```
TF1 fitFunk = TF1("fitFunk",
            "( 1.0 + [0]*TMath::Power(x,2) + [1]*x)/(2.0+2.0*[0]/3.0)" , -1., 1.);
```

Um die Parameter auf bestimmte Werte zu setzen, können Sie die Funktion `SetParameter` wie folgt benutzen:

```
fitFunk.SetParameter(0, 1.0);
fitFunk.SetParameter(1, 0.1);
```

(iv) Definieren Sie - analog zu dem Maximum Likelihood Fit aus Aufgabe 38 - eine Schleife, um über mögliche Werte von α zu iterieren. Definieren Sie innerhalb dieser Schleife eine weitere Schleife über möglichen Werte von β. Sinnvolle Werte für die Variation von α sind zwischen 0 und 0,25 und für β zwischen 0,8 und 1,25. Nehmen Sie sowohl für α als auch für β 50 äquidistante Werte an (somit wird die Likelihood für 250 Kombinationen ausgerechnet).

(v) Erstellen Sie für jeden Wert von α_i und β_i eine Schleife über die Bins in Ihrem $\cos\theta$ Histogramm und berechnen Sie den log-Likelihood Wert. Gehen Sie wie folgt vor:

- Setzen Sie zunächst die Parameter in der bereitgestellten `TF1`-Funktion auf die gerade aktuellen Werte von α_i und β_i.
- Erstellen Sie eine Schleife von i=1 (das niedrigste Bin des Histogramms - Bin 0 - ist das Underflow Bin) bis i=`hist.GetNbinsX()`. Dies ist bereits die dritte und gleichzeitig innerste `for`-Schleife.
- Die Anzahl an Einträgen im $i$ten Bin erhalten Sie mit der `TH1` Funktion `GetBinContent(int i)`
- Um Ihre Funktion über die Binbreite zu integrieren benutzen Sie die `TF1` Funktion `Integral(float a, float b)`. `a` steht für die untere, `b` für die obere Integrationsgrenze. Um diese Werte aus dem $\cos\theta$ Histogramm zu erhalten, benutzen Sie die `TH1` Funktion `GetBinLowEdge(int i)`, welche die untere Grenze des `i`-ten Bins ausgibt.
- Summieren Sie die log-Likelihoodfunktion für alle Bins des Histogramms auf.

(vi) Füllen Sie für jedes α_i und β_i den log-Likelihood Wert in einen zweidimensionalen Graph. Dies funktioniert analog zu einem eindimensionalen Graphen:

```
TGraph2D likeGraph = TGraph2D(likepoints);
```

wobei `likepoints` die Gesamtanzahl an Kombinationen von α und β ist, für die die Likelihood ausgewertet werden soll.

Jeder Punkt des Graphen kann gesetzt werden durch die `TGraph2D` Funktion:

```
SetPoint(int pointNumber, float alpha, float beta, float loglikelihood)
```

Denken Sie daran, sich `pointNumber` auszurechnen (dies ist nicht die Zählvariable in den Schleifen über die Werte von α und β!).

(vii) Zeichnen Sie den Graphen (log-Likelihood gegen α und β) unter Benutzung von

```
likeGraph.Draw("colz");
```

wobei die Option `"colz"` ROOT die Anweisung gibt, einen Surface-Plot zu erstellen. Weitere Optionen des Draw-Befehls können im 'ROOT User's Guide' in der `TGraph2D` Sektion gefunden werden.

(viii) Ermitteln Sie aus den erzeugten Graph eine Abschätzung für $\hat{\alpha}$ und $\hat{\beta}$ und deren Standardabweichungen $\sigma_{\hat{\alpha}}$ und $\sigma_{\hat{\alpha}}$. Dazu könnten folgende Befehle nützlich sein:

`float maximum=likeGraph.GetHistogram().GetMaximum();`

gibt das Maximum des Graphen zurück.

`likeGraph.GetHistogram().SetMaximum(maximum-0.5);`

setzt für die Darstellung die minimale Zeichenebene auf das Maximum des Graphen minus 0.5.

(ix) Vergleichen Sie die Werte Ihres Fits mit denen des RooT Fitting Paketes, indem Sie die Histogramm-Anpassungsfunktion

`hist.Fit("functionName", "LI");`

benutzen, wobei `"functionName"` der Name der WDF Funktion ist und die Option `"LI"` einen Likelihood Fit ausführt, indem die gegebene Funktion über jedes Bin des Histogramms integriert wird.

Aufgabe 44 *Erweiterter Maximum Likelihood Fit für eine Signal- and 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:

$$f(x;\mu_s,\mu_b) = \frac{\mu_s}{\mu_s+\mu_b} f_s(x) + \frac{\mu_b}{\mu_s+\mu_b} f_b(x) \tag{6}$$

wobei entsprechend μ_s und μ_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:

$$P(n;\mu_s,\mu_b) = \frac{(\mu_s+\mu_b)^n}{n!} \exp(-(\mu_s+\mu_b)) \tag{7}$$

Daher ist die log-Likelihoodfunktion gegeben durch

$$\ln \mathcal{L} = -(\mu_s+\mu_b) + \sum_{i=1}^{n} \ln[(\mu_s+\mu_b)\ f(x_i;\mu_s,\mu_b)] \tag{8}$$

Diese Funktion sollen Sie minimieren, 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/warsinsk/sd_wise1112/ueb7/aufgabe44_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:

$$f(x;\ \mu,\ \sigma,\ \tau,\ \mu_s,\ \mu_b) = \frac{\mu_s}{\mu_s+\mu_b} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) + \frac{\mu_b}{\mu_s+\mu_b} \frac{1}{\tau} \exp\left(-\frac{x}{\tau}\right) \tag{9}$$

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 $\mu = 2.0$, $\sigma = 0.5$ 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 43, eine Schleife über mögliche Werte des Parameters μ_s. Definieren Sie innerhalb dieser Schleife eine weitere Schleife über mögliche Werte von μ_b. Sinnvolle Bereiche für μ_s und μ_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 μ_s und μ_b mittels der `SetParameter(int parnumber, float parwert)`-Funktion.

(vii) Erstellen Sie für jeden Wert von μ_s und μ_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 μ_s und μ_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 $\mu = 2.0$, $\sigma = 0.5$ 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 45 *Maximum Likelihood für die Poissonverteilung* **5 Punkte**

Die Anzahl k der innerhalb einer Minute das Höllental passierenden LKW sei poissonverteilt, d.h. es gilt:

$$f(k;\nu) = \frac{\nu^k}{k!}e^{-\nu}, \quad k = 0,1,2,\ldots, \quad \nu > 0.$$

Eine Stichprobe mit Umfang $n = 100 (x_1,\ldots,x_{100})$ liefert folgendes Ergebnis:

k	absolute Häufigkeit
0	10
1	20
2	30
3	20
4	10
5	5
6	5
7+	0

(i) Schätzen Sie den Wert des Parameters ν, indem Sie $L(\nu) = \prod_{i=1}^{100} f(k = x_i;\nu)$ maximieren für die gegebene Stichprobe $(x_1,\ldots,x_{100})$.

(ii) Berechnen Sie den ML-Schätzer $\hat{\nu}$ für eine allgemeine Stichprobe vom Umfang n, indem Sie $\ln L$ maximieren. Die Messwerte seien dabei Poisson-verteilt.

(iii) Ist der Schätzer erwartungstreu?

(iv) Berechnen Sie die Varianz $V(\hat{\nu})$ für den Fall $n \to \infty$ unter der Annahme, dass der ML-Schätzer effizient ist. Verwenden Sie die MVB für erwartungstreue Schätzer

$$V[\hat{\nu}] = E\left[-\frac{\partial^2 \ln L}{\partial \nu^2}\right]^{-1}$$

und dass für einen effizienten Schätzer gilt

$$E\left[\frac{\partial^2 \ln L}{\partial \nu^2}\right] = \left(\frac{\partial^2 \ln L}{\partial \nu^2}\right)_{\nu=\hat{\nu}}$$

.

Aufgabe 46 *Geradenanpassung* **7 Punkte**

Nehmen Sie an, dass Sie N Messwerte $(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$ mit jeweils verschiedenen, aber bekannten Fehlern σ_i auf den Wert von y_i vorliegen haben. Es soll mittels der Methode der kleinsten Quadrate diejenige Gerade ermittelt werden, die am besten durch diese Punkte geht.

Betrachten Sie die folgende Parametrisierung einer Geraden für die Datenpunkte (x_i,y_i):

$$f_i(\theta_1;\theta_2;x_i) = \theta_1 + \theta_2 x_i,$$

wobei θ_1 der Achsenabschnitt und θ_2 die Steigung der Geraden sind.

(i) Ermitteln Sie die kleinsten-Quadrat-Schätzer $\hat{\theta}_1$ und $\hat{\theta}_2$, indem Sie die Größe

$$X^2 = \sum_{i=1}^{N} w_i (y_i - f_i)^2 = \sum_{i=1}^{N} \left(\frac{y_i - f_i}{\sigma_i}\right)^2$$

minimieren.

(ii) Benutzen Sie die Fehlerfortpflanzungsformel um die Varianzen der beiden Schätzer $\sigma^2_{\hat{\theta}_1}$ und $\sigma^2_{\hat{\theta}_2}$ zu ermitteln.

Aufgabe 47 *Kombination von Messungen mit der Methode der kleinsten Quadrate* **8 Punkte**

Es ist möglich, einen Spezialfall der Methode der kleinsten Quadrate zu benutzen, um Messungen mit derselben Qualität zu kombinieren. Betrachten Sie N Messungen, y_i, welche den wahren, aber

unbekannten Wert λ bestimmen sollen. Jede Messung y_i hat einen geschätzten Fehler von σ_i. Da λ für alle Ereignisse konstant ist, ergibt sich die χ^2-Variable zu:

$$\chi^2(\lambda) = \sum_{i=1}^{N} \frac{(y_i - \lambda)^2}{\sigma_i^2} \tag{10}$$

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

$$\chi^2(\lambda) = \sum_{i,j=1}^{N} (y_i - \lambda)(V^{-1})_{ij}(y_j - \lambda) \tag{11}$$

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

$$\hat{\lambda} = \sum_{i=1}^{N} w_i y_i \tag{12}$$

wobei die Gewichtungen w_i gegeben sind durch

$$w_i = \frac{\sum_{j=1}^{N} (V^{-1})_{ij}}{\sum_{k,l=1}^{N} (V^{-1})_{kl}}, \tag{13}$$

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

$$V[\hat{\lambda}] = \sum_{i,j=1}^{N} w_i V_{ij} w_j. \tag{14}$$

Tipp: Benutzen Sie Fehlerfortpflanzung!

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

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{15}$$

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

$$\hat{\lambda} = wy_1 + (1 - w)y_2 \tag{16}$$

mit

$$w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}. \tag{17}$$

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

$$\frac{1}{V[\hat{\lambda}]} = \frac{1}{1 - \rho^2}\left[\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - \frac{2\rho}{\sigma_1\sigma_2}\right] = \frac{1}{\sigma^2} \tag{18}$$

und zeigen Sie dass folglich

$$\frac{1}{\sigma^2} - \frac{1}{\sigma_1^2} \geq 0 \tag{19}$$

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