Statistische Methoden der Datenanalyse

Markus Schumacher, Stan Lai, Florian Kiss

Übung XII

28.1.2014, 29.1.2014

Anwesenheitsaufgaben

Aufgabe 57 *Anpassungsgüte bei einer Maximum Likelihood Anpassung*

In dieser Übung sollen Sie die Ergebnisse der "Binned"Maximum Likelihood Anpassung an die $e^+e^- \to \mu^+\mu^-$-Ereignisse aus Übung VII benutzen. Betrachtet werden sollen die Verhältnisse der Likelihoodfunktionen

$$\lambda = \frac{\mathcal{L}(\vec{n}|\vec{\nu})}{\mathcal{L}(\vec{n}|\vec{n})}.$$

Damit soll gezeigt werden, dass

$$\chi^2_M = -2\ln\lambda_M$$

gemäß einer χ^2-Verteilung mit $N - m - 1$ Freiheitsgraden verteilt ist. Dabei ist N die Anzahl der Datenbins und m die Anzahl der angepassten Parameter.

Das Makro `/home/slai/StatisticsCourse/PS12/aufgabe57_anfang.C` gibt ein Beispiel, wie man die generierten $e^+e^- \to \mu^+\mu^-$-Ereignisse (Anzahl N_{tot}) einliest, ihre $\cos\theta$-Verteilung berechnet und in ein Histogramm namens `hist` einfüllt.

(i) Definieren Sie eine `TF1`-Funktion gemäß

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

und führen Sie eine Anpassung dieser Funktion an das $\cos\theta$-Histogramm mit dem Befehl

```
hist.Fit("FunkName","IL");
```

durch. Die Option `"IL"` steht für eine Maximum Likelihood Anpassung unter Benutzung der Integrale über die Bins des Histogramms `hist`.

(ii) Definieren Sie als nächstes eine weitere `TF1`-Funktion gemäß $f(x;\alpha,\beta)$ unter Benutzung der angepassten Werte $\hat{\alpha}$ und $\hat{\beta}$. Die Werte von $\hat{\alpha}$ und $\hat{\beta}$ können Sie mit

```
TF1::GetParameter(int i)
```

ermitteln, wobei der Wert des `i`-ten Parameters zurückgegeben wird. Zum Fixieren der Parameter in der neuen Funktion benutzen Sie

```
TF1::FixParameter(int i, float wert)
```

um den Wert des `i`-ten Parameters auf `wert` zu setzen. Benutzen Sie dann

```
TF1::GetRandom(),
```

um eine gemäß der so definierten Funktion verteilte Zufallszahl zu erhalten.

(iii) Erzeugen Sie N_{tot} neue Werte für $\cos\theta$ und füllen Sie diese in ein neues Histogramm. Führen Sie eine Anpassung der ursprünglichen Funktion $f(x;\alpha,\beta)$ an das neue $cos(\theta)$-Histogramm durch.

(iv) Ermitteln Sie als nächstes den χ^2-Wert für das erzeugte Histogramm, indem Sie die Formel

$$\chi^2_M = 2\sum_{i=1}^{N}\left(n_i \ln\frac{n_i}{\hat{\nu}_i}\right)$$

benutzen. Dabei sind N die Anzahl der Histogrammbins und $\hat{\nu}_i$ die Erwartungswerte der angepassten Funktion $f(x;\hat{\alpha},\hat{\beta})$, die gegeben sind durch

$$\hat{\nu}_i = N_{\text{tot}} \int_{x_i^{\min}}^{x_i^{\max}} f(x;\hat{\alpha},\hat{\beta})\mathrm{d}x.$$

Zum Integrieren der Funktion $f(x;\hat{\alpha},\hat{\beta})$ sollten Sie die Methode

```
TF1::Integral(int binMin,int binMax)
```

verwenden. Die untere Bingrenze des `i`-ten Bins erhalten Sie mittels

```
TH1::GetBinLowEdge(int i).
```

Schreiben Sie eine Schleife über alle Histogrammbins, in der Sie die Einzelbeiträge zu χ^2_M aufsummieren.

(v) Wiederholen Sie dieses Vorgehen 1000 mal und füllen Sie jedesmal den Wert von χ^2_M in ein Histogramm ein. Denken Sie daran, das neue Histogramm für die zufällig ermittelten Werte von $\cos\theta$ am Anfang jedes Zufallsexperiments wieder zurückzusetzen. Dies geschieht am einfachsten mit der Methode

```
TH1::Reset();
```

(vi) Stellen Sie das Histogramm am Bildschirm dar und überzeugen Sie sich davon, dass es einer χ^2-Verteilung folgt. Führen Sie, falls Sie noch Zeit haben, eine Anpassung einer χ^2-Funktion an das Histogramm durch. Stimmt die Anzahl der Freiheitsgrade mit der Erwartung überein?

Hausaufgaben

Aufgabe 58 *Zählexperiment für eine Signal- und Untergrundmessung - Teil 2* **10 Punkte**

Monte-Carlo-Studien von Proton-Proton-Kollisionen im ATLAS-Detektor haben gezeigt, dass der Wirkungsquerschnitt für $pp \to H+X \to \gamma\gamma+X$-Ereignisse, die die Ereignisselektion passieren, gegeben ist durch $\sigma_S = 25.4\,\text{fb}$. Der Wirkungsquerschnitt für Untergrundereignisse, die dieselbe Ereignisselektion passieren, beträgt $\sigma_B = 947\,\text{fb}$. In einer weiteren Analyse kann ein reiner Untergrunddatensatz mit einem Wirkungsquerschnitt von $\sigma_T = 10300\,\text{fb}$ ausgewählt werden.

(i) Benutzen Sie die Relation

$$N = \mathcal{L}\sigma,$$

um die erwarteten Anzahlen von Signal- (x_s), Untergrundereignissen (x_b) im Signaldatensatz sowie die Anzahl von Ereignissen in der Seitenbandregion ($y = \tau b$) für eine integrierte Luminosität von $\mathcal{L} = 10\,\text{fb}^{-1}$ auszurechnen.

(ii) Berechnen Sie die Schätzer für die Anzahl der Signalereignisse $\hat{s}$, der Untergrundereignisse $\hat{b}$ unter der Hypothese von Signal plus Untergrund , sowie den Schätzer $\hat{\hat{b}}$ auf die Anzahl der Untergrundereignisse in der Nur-Untergrund Hypothese. (Annahme: Die Anzahl der beobachteten Ereignissen entsprechen genau der Anzahl der erwarteten Ergebnissen)

(iii) Berechnen Sie die Größe

$$q = -2\ln\lambda.$$

(iv) Berechnen Sie daraus die Signifikanz des vorhergesagten Signals.

(v) Nehmen Sie nun an, dass die Untergrundrate in der Signalregion mit einer relativen Genauigkeit von $\Delta b/b = 5\%$ abgeschätzt werden kann. Wenn man annimmt, dass es sich dabei um einen Poissonfehler handelt, kann man eine effektive Seitenbandregion konstruieren mit Ereignisanzahl $y' = \tau' b$. Dazu setzt man die relative statistische Unsicherheit in der hypothetischen Seitenbandregion (Poissonfehler) gleich der relativen Genauigkeit der Untergrundvorhersage:

$$\frac{\sqrt{\tau' b}}{\tau' b} = \frac{\Delta b}{b} \Leftrightarrow \tau' = \frac{b}{(\Delta b)^2}$$

Berechnen Sie die Werte für τ' und y' für die effektive Seitenbandregion.

(vi) Benutzen Sie die Werte für τ' und y' sowie die ursprüngliche Anzahl von Ereignissen in der Signalregion x, um den neuen Wert für q und daher der Signifikanz für eine Messung mit einem Fehler von 5% auf die Untergrundvorhersage zu bekommen.

Aufgabe 59 *Maximale Separation der Fisherdiskriminante* **10 Punkte**

Betrachten Sie eine Teststatistik t basierend auf einer Linearkombination der Eingangsvariablen $\vec{x} = (x_1, \ldots, x_n)$ mit Koeffizienten $\vec{a} = (a_1, \ldots, a_n)$,

$$t(\vec{x}) = \vec{a}^T \vec{x}.$$

Unter den zwei Hypothesen H_0 und H_1 sind dann die Mittelwerte und Kovarianzen der Daten $\vec{x}$ gegeben durch

$$(\mu_k)_i = \int x_i f(\vec{x}|H_k) dx_1 \ldots dx_n, \qquad k \in \{0,1\},$$

$$(V_k)_{ij} = \int (x-\mu_k)_i (x-\mu_k)_j f(\vec{x}|H_k) dx_1 \ldots dx_n, \qquad k \in \{0,1\}.$$

(i) Zeigen Sie, dass sich die Erwartungswerte und Varianzen von t unter den beiden Hypothesen dann ergeben zu:

$$\tau_k = \vec{a}^T \vec{\mu}_k, \qquad k \in \{0,1\},$$

$$\Sigma_k^2 = \vec{a}^T V_k \vec{a}, \qquad k \in \{0,1\}.$$

(ii) Ein Maß für die Separation der zwei Hypothesen unter Verwendung der Teststatistik t ist dann gegeben durch

$$J(\vec{a}) = \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2}.$$

Zeigen Sie unter Benutzung von (i), dass sich dieses Separationsmaß auch schreiben lässt als

$$J(\vec{a}) = \frac{\vec{a}^T B \vec{a}}{\vec{a}^T W \vec{a}}.$$

mit

$$B_{ij} = (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j,$$

und

$$W_{ij} = (V_0 + V_1)_{ij}.$$

(iii) Bilden Sie die Ableitung $\partial J(\vec{a})/\partial \vec{a}$ von $J(\vec{a})$ nach $\vec{a}$ und zeigen Sie, dass das Maximum von $J(\vec{a})$ durch die Eigenwertgleichungen

$$W^{-1} B \vec{a} = \lambda \vec{a}$$

gegeben ist.

(iv) Zeigen Sie, dass für einen beliebigen Vektor $\vec{a}$ der Vektor $B\vec{a}$ parallel zu $(\vec{\mu}_0 - \vec{\mu}_1)$ ist.

(v) Zeigen Sie damit, dass

$$\vec{a} \propto W^{-1}(\vec{\mu}_0 - \vec{\mu}_1)$$

eine Lösung der Eigenwertgleichungen aus (iii) ist und daher $J(\vec{a})$ maximiert.