Albert-Ludwigs-Universität Freiburg Sommersemester 2009

Statistische Methoden der Datenanalyse

Markus Schumacher

Übung X

Matthew Beckingham und Markus Warsinsky

8.7.2009

Computerübung

Aufgabe 44 *Profile Likelihood für die Entdeckung eines neuen Teilchens*

Betrachtet wird folgendes Szenario: Eine Theorie sagt die Existenz eines neuen Teilchens mit einer Masse von 8 GeV vorher, welches im Experiment als eine resonante Überhöhung über einem exponentiell verteilten Untergrund ($\tau = 10\,\text{GeV}$) beobachtet werden könnte. Die Wahrscheinlichkeitsdichtefunktion für den Untergrund sei also eine Exponentialverteilung, und die für das Signal eine Gaussfunktion mit Mittelwert 8 GeV und Standardabweichung 0,5 GeV, da wir weiterhin annehmen, dass die durch die Detektorauflösung beobachtete Breite der Resonanz – sofern sie existiert – 0,5 GeV betrage. Des weiteren sagt unsere bisherige Standardtheorie eine Gesamtanzahl von Untergrundereignissen von $N_{\text{UG}} = 10000$ voraus, sowie unsere neue Theorie $N_{\text{Sig}} = 175$ Signalereignisse. Im folgenden soll mittels der Profile-Likelihood-Methode, die in der Vorlesung und der letzten Anwesenheitsaufgabe besprochen wurde, die Sensitivität des Experiments auf eine eventuelle Entdeckung untersucht werden. Die Profile-Likelihood ist definiert über das Verhältnis

$$\lambda = \frac{L(\vec{x}|H_0)}{L(\vec{x}|H_1)}, \tag{1}$$

wobei $\vec{x}$ die beobachteten Daten, L die unter der betreffenden Hypothese maximierte Likelihoodfunktion, H_0 die „nur-Untergrund“ Hypothese und H_1 die „Signal und Untergund“ Hypothese sind. Zumeist wird dann die Größe

$$q = -2\ln\lambda \tag{2}$$

betrachtet. Für ein Experiment mit „nur-Untergrund“ sollte $q(\vec{x}_{\text{UG}})$ verteilt sein wie eine χ^2-Verteilung mit einem Freiheitsgrad.

Im folgenden soll die Monte-Carlo-Methode benutzt werden, um Pseudoexperiment einerseits nur mit Untergrund, als auch mit Signal- und Untergrund durchzuführen. Mittels diese Pseudoexperimente können dann die Verteilungen von $q(\vec{x}_{\text{UG}}) \equiv q_0$ und von $q(\vec{x}_{\text{Sig.+UG}}) \equiv q_1$ erzeugt werden, um festzustellen, wie sensitiv das Experiment auf das vorhergesagte neue Teilchen ist.

Zur Durchführung der Peudoexperimente sollen mit

```
double data = Funk->GetRandom();
```

Zufallszahlen erzeugt werden, die nach der Untergrund- bzw. Signal-WDF verteilt sind. Die im jeweiligen Pseudoexperiment zu generierende Anzahl von Untergrund- bzw. Signalereignissen bestimmen Sie nach der Poissonverteilung mittels `gRandom->Poisson(nbkg);`. Achten Sie jeweils darauf, das jeweils sowohl ein Pseudoexperiment mit nur Untergrund und eines mit Signal und Untergrund gemacht wird. Im weiteren sollen drei verschiedene Suchstrategien nach diesem neuen Teilchen besprochen werden. Im Makro `/home/warsinsk/stat/uebung10/aufgabe44_anfang.C` befindet sich ein Beispielmakro, in dem einige der (auch später) benötigten Funktionen und Histogramme schon vordefiniert sind. Die Binbreiten und Vorgabewerte sind hier bereits aufeinander angepasst, so dass später Zeit gespart werden kann.

(i) Als ein erster Ansatz soll ein reines Zählexperiment in einem sogenannten Massenfenster gemacht werden, d.h. man zählt nur die Anzahl der Ereignisse in einem bestimmten Massengebiet. Im folgenden soll dieses Massengebiet die 2σ-Umgebung um die Position des Signals sein, also das Intervall zwischen 7 und 9 GeV. Weiterhin wollen wir unserer bisherigen Theorie in Bezug auf

die Untergrundvorhersage absolut vertrauen, die erwartete Anzahl an Untergrundereignissen im Massenfenster B ist also gegeben durch das Integral über die Untergrund-WDF multipliziert mit der mittleren Gesamtanzahl von Untergrundereignissen.

Wenn man in der so definierten Signalregion x Ereignisse beobachtet, ergibt sich nach einer einfachen Rechnung der q-Wert zu:

$$q = -2x \ln B + 2B + 2x \ln x - 2x. \tag{3}$$

Gehen Sie nun wie folgt vor:

a) Bestimmen Sie die Anzahl der Untergrundereignisse im Massenfenster mittels `TF1::Integral(Double_t low, Double_t high)`.

b) Führen Sie 10000 Pseudoexperimente nur mit Untergrund durch. Zählen Sie jedesmal die Anzahl von Ereignissen im Massenfenster und berechnen Sie für jedes Experiment den q-Wert, im folgenden q_0 genannt. Füllen Sie diese in ein Histogramm. Im Beispielmakro ist eines vorgegeben.

c) Führen Sie dasselbe für Pseudoexperimente mit Signal- und Untergrund durch. Sie können hier die gleichen Untergrundereignisse wie vorher verwenden und nur Signalereignisse hinzufügen. Füllen Sie die erhaltenen q-Werte (q_1) in ein weiteres Histogramm (ebenfalls vorgegeben).

d) Stellen Sie die Verteilungen von q_0 bzw. q_1 graphisch dar.

e) Verifizieren Sie, dass es sich bei der Verteilung von q_0 um eine χ^2-Verteilung mit einem Freiheitsgrad handelt. In `chi2snippet.C` befindet sich eine definierte Funktion, nebst geigneten Startwerten für eine Anpassung. Sie können auch die Normierung und die Anzahl der Freiheitsgrade fixieren (`FixParameter` statt `SetParameter`) und diese Kurve zum Vergleich in einer anderen Farbe (z.B. `SetLineColor(kRed)`) mit einzeichnen.

f) Berechnen Sie den Median der Verteilung von q_1. Dies können Sie z.B. wie folgt machen:

```
const Int_t nq = 1;
Double_t xq[nq];  // position where to compute the quantiles in [0,1]
Double_t yq[nq];  // array to contain the quantiles
xq[0]=0.5;
qvalue_sigplusbkg->GetQuantiles(nq,yq,xq);
Double_t median=yq[0];
```

g) Wie groß wäre also für Experimente mit Signal- und Untergrund im Median der q-Wert? Warum könnte man in diesem Fall den p-Wert für die Hypothese H_0 (nur Untergrund) nicht so einfach mit solchen Pseudoexperimenten ermitteln?

(ii) Als nächstes wollen wir von der Annahme, dass wir den Untergrund im Massenfenster exakt kennen, was nicht besonders realistisch ist, abrücken, und stattdessen annehmen, dass wir nur die Form des Untergrundes perfekt kennen. Man defniert sich dann beispielsweise ein Seitenband über die Forderung, mehr als 4σ von der Signalposition entfernt zu sein. Das Verhältnis τ zwischen Seitenband- und Signalregion ist dann gegeben durch das Verhältnis der Integrale der Untergrund-WDF in diesen beiden Gebieten. Dieses Seitenband wird dann zur Messung des Untergrundes in den Daten verwendet. Wenn dann in der Signalregion x und im Seitenband y Ereignisse gesehen werden, ergibt sich der q-Wert zu:

$$2(x \ln(x) + y \ln(y) - (x+y) \ln\left(\frac{x+y}{1+\tau}\right) - y \ln(\tau)). \tag{4}$$

a) Bestimmen Sie τ für den Fall der beschriebenen Signal- und Seitenbandregion.

b) Führen Sie wieder 10000 Pseudoexperimentze mit nur Untergrund sowie Signal- und Untergrund durch und füllen Sie q_0 bzw. q_1 in ein Histogramm.

c) Verifizieren Sie wieder das Verhalten von q_0 sowie den Median der q_1-Verteilung. Was fällt Ihnen auf?

(iii) Als letztes soll nun die Möglichkeit ins Auge gefasst werden, kein reines Zählexperiment durchzuführen, sondern die gesamte Form von Signal- und Untergrund zu berücksichtigen. Dies kann zum Beispiel durch eine gebinnte Maximum-Likelihood-Anpassung an die generierten Massenspektren geschehen. Im Falle der Hypothese H_0 würde man also nur eine Exponentialfunktion, im Falle der Hypothese H_1 eine Exponential- plus eine Gaussfunktion benutzen.

Die Profile-Likelihood ergibt sich dann über das Verhältnis der maximierten Likelihoodfunktionen. In ROOT kann auf die Likelihood wie folgt zugegriffen werden:

```
histogram->Fit("funk","LQ0");
TVirtualFitter *fitter = TVirtualFitter::GetFitter();
Double_t logl;
Double_t edm;
Double_t errdef;
Int_t nvpar;
Int_t nparx;
fitter->GetStats(logl,edm,errdef,nvpar,nparx);
```

Bei dem erhaltenen `logl` handelt es sich dann um $-2\ln L$.

Gehen Sie nun wie folgt vor:

a) Führen Sie wieder je 10000 Pseudoexperimente durch. Füllen Sie jeweils die Massenhistogramme. Führen Sie in jedem Pseudoexperiment eine Likelihood-Anpassung sowohl der reinen Untergrund- als auch der Signal- plus Untergrund-Funktion durch. Achten Sie darauf, nur die jeweiligen Normierungen frei zu lassen, die Form von Signal- und Untergrund setzen wir weiterhin als bekannt voraus. Achten Sie darauf, vor jeder Anpassung die Startwerte neu zu setzen. Ermitteln Sie aus den Ergebnissen der Anpassungen die Werte q_0 und q_1 (Hinweis: Für jeden q-Wert sind zwei Anpassungen notwendig.) Benutzen Sie für die Anpassung die Option `"LQ0"`, die Option `"LIQ0"` liefert eigentlich die exakteren Ergebnisse, ist aber auch erheblich langsamer.

b) Verfizieren Sie wieder die Eigenschaften von q_0 und bestimmen Sie den Median von q_1. Was fällt Ihnen auf?

(iv) Wir haben also gesehen, dass die Profile-Likelihood es ermöglicht, sehr einfach p-Werte auszurechnen, da der Satz von Wilkes für die Nullhypothese die Vorhersage macht, dass die q-Werte nach χ^2 verteilt sein sollen. Da für Entdeckungen im allgemeinen p-Werte in der Größenordnung von 10^{-7} (entsprechend q-Werten um 25) betrachtet werden, wäre eine MC-Simulation dieser Verteilung sehr zeitaufwändig. Wie viele Pseudoexperimente müßte man beispielsweise durchführen, wenn man bei einem erwartetem q von 25 den p-Wert auf 10% genau bestimmen wollte?

Aufgabe 45 *Fisher-Diskriminante*

Im folgenden soll die sogenannte Fisher-Diskriminante zur Trennung zwischen zwei Ereignisklassen benutzt werden. Dabei werden i.a. n trennende Variablen $x_1, \ldots x_n$ benutzt. Die Fisher-Diskriminante ist dann gegeben durch

$$t = \sum_{i=1}^{n} a_i x_i - \frac{1}{2}\sum_{i=1}^{n} a_i(\bar{x}_i^{(1)} + \bar{x}_i^{(2)}), \tag{5}$$

wobei $\bar{x}_i^{(1,2)}$ die Mittelwerte der Observablen der Klasse 1 bzw. 2 sind. Die a_i ergeben sich aus den Kovarianzmatrizen der Klasse j

$$V_{km}^{(j)} = \frac{1}{N}\sum_{N}(x_m^{(j)} - \bar{x}_m^{(1)})(x_k^{(j)} - \bar{x}_k^{(1)}), \tag{6}$$

wobei über einen Trainingsdatensatz der Größe N summiert wird, zu:

$$a_i = \sum_k (V^{-1})_{ik}(\bar{x}_k^{(1)} - \bar{x}_k^{(2)}) \qquad \text{mit:} \quad V_{mk} = \frac{1}{2}\left(V_{mk}^{(1)} + V_{mk}^{(2)}\right).$$

Im folgenden wollen wir die folgende Situation betrachten: Eine Signalklasse C_S und eine Untergrundklasse C_B sollen durch eine Fisher-Diskrimante in den Variablen (x_1,x_2) optimal getrennt werden. Die Signalereignisse sollen gemäß einer zweidimensionalen Gaussverteilung mit Mittelwerten (0. 7,0. 35) und $\sigma = 0.15$, und die Untergrundereignisse ebenfalls nach einer zweidimensionalen Gaußverteilung mit gleichem σ und Mittelwerten (0. 4,0. 75) verteilt sein.

Das Makro `aufgabe45_anfang.C` enthält bereits die Konstruktion der Fisher-Diskrimanten gemäß obiger Vorschrift und nach den beschriebenen Wahrscheinlichkeitsdichtefunktionen. Sie sind gerne eingeladen, sich von der Richtigkeit dieses Makros zu überzeugen. Benutzen Sie dieses Makro, um folgende Aufgaben durchzuführen.

(i) Erzeugen Sie nochmals 10000 Signal- und Untergrundereignisse, ermitteln Sie jeweils den Wert der Fisher-Diskriminante und tragen ihn in Histogramme ein. Stellen Sie die Histogramme für Signal- und Untergrundereignisse in verschiedenen Farben übereinandergelegt dar.

(ii) Zur Separation zwischen Signal- und Untergrund kann man also Schnitte der Form $t > t_{\text{cut}}$ anwenden. Fahren Sie einen solchen Schnitt durch und ermitteln Sie für jeden Schnittwert t_{cut} die Effizienz ϵ und Reinheit π, die definiert sind durch

$$\epsilon = \frac{N_S(t > t_{\text{cut}})}{N_S} \qquad \pi = \frac{N_S(t > t_{\text{cut}})}{N_S(t > t_{\text{cut}}) + N_B(t > t_{\text{cut}})}. \tag{7}$$

(iii) Stellen Sie die Reinheit in Abhängigkeit von der Effizienz mittels eines `TGraph` graphisch dar.

(iv) Erstellen Sie einen „Scatterplot“ der Meßgrößen x_1 und x_2 für Signal- und Untergrundereignisse.

(v) Zeichnen Sie in diesen Graphen Linien mit konstantem t ein.

(vi) Sollte noch Zeit bleiben: Führen Sie die obigen Aufgaben auch durch, wenn Sie nur in einer Variablen, z.B. x_1 schneiden. Vergleichen Sie die Effizienz-Reinheitskurve mit derjenigen der Fisher-Diskriminanten.