Albert-Ludwigs-Universität Freiburg Wintersemester 2011/2012

Statistische Methoden der Datenanalyse

Markus Schumacher

Übung XII

Markus Warsinsky

30.1.2012

Anwesenheitsaufgaben

Aufgabe 66 *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 Hausaufgabe 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 Pseudoexperimente 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

```
float 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 `myrandom.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/sd_wise1112/ueb12/aufgabe66_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 B von Untergrundereignissen im Massenfenster 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 B der erwarteteten Untergrundereignisse im Massenfenster mittels `FunkUG.Integral(Double_t low, Double_t high)` und der bekannten erwarteten Gesamtanzahl `nbkg`.

b) Führen Sie 10000 Pseudoexperimente nur mit Untergrund durch. Ermitteln Sie die Anzahl der zu nehmenden Messwerte in jedem Zufallsexperiment mittels `int nbkg_diesesexperiment=myrandom.Poisson(nbkg);`. Würfeln Sie dann entsprechend `nbkg_diesesexperiment`-mal zufällig einen Wert gemäß `FunkUG` und zählen die Anzahl von Ereignissen im Massenfenster. Wenn das Zufallsexperiment vollständig erfolgt ist (also die Schleife über die `nbkg_diesesexperiment` Zufallszahlen beendet ist), berechnen Sie für jedes Experiment den q-Wert, im folgenden q_0 genannt. Füllen Sie diesen in ein Histogramm. Im Beispielmakro ist eines vorgegeben (`qvalue_bkgonly`).

c) Führen Sie dasselbe für Pseudoexperimente mit Signal- und Untergrund durch. Sie können hier die gleichen Untergrundereignisse wie vorher verwenden und nur Signalereignisse hinzufügen. Ermitteln Sie die Anzahl an Signalereignissen mittels `int nsig_diesesexperiment=myrandom.Poisson(nsig);`. Würfeln Sie dann entsprechend `nsig_diesesexperiment`-mal zufällig einen Wert gemäß `FunkSig` und zählen die Anzahl von Ereignissen im Massenfenster. Beachten Sie, dass zur Ermittlung der q-Werte nun die Summe von Signal- und Untergrundereignissen benötigt wird, da die Hypothese H_1 simuliert wird. Ermitteln Sie die erhaltenen q-Werte (q_1) und füllen Sie sie in ein weiteres Histogramm (ebenfalls vorgegeben: `qvalue_sigplusbkg`).

d) Stellen Sie die Verteilungen von q_0 bzw. q_1 graphisch dar, nachdem Sie 10000 Zufallsexperimente durchgeführt haben.

e) Verifizieren Sie, dass es sich bei der Verteilung von q_0 um eine χ^2-Verteilung mit einem Freiheitsgrad handelt. In `/home/warsinsk/sd_wise1112/ueb12/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. Um die Verteilung besser sehen zu können, können Sie mittels `gPad.SetLogy();` eine halblogarithmische Darstellung wählen.

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

```
double xq[1];  // position where to compute the quantiles in [0,1]
double yq[1];  // array to contain the quantiles
xq[0]=0.5;
qvalue_sigplusbkg.GetQuantiles(1,yq,xq);
float 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) Wir haben also gesehen, dass es die Profile-Likelihood 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 erwarteten q von 25 den p-Wert auf 10% genau bestimmen wollte?

Hausaufgaben

Aufgabe 67 *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 $\vec{a} = (a_1, \ldots, a_n)$,

$$t(\vec{x}) = \vec{a}^T \vec{x}. \tag{5}$$

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

$$(\mu_k)_i = \int x_i f(\vec{x}|H_k) dx_1 \ldots dx_n, \qquad k \in \{0,1\}, \tag{6}$$

$$(V_k)_{ij} = \int (x - \mu_k)_i (x - \mu_k)_j f(\vec{x}|H_k) dx_1 \ldots dx_n, \qquad k \in \{0,1\}. \tag{7}$$

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

$$\tau_k = \vec{a}^T \vec{\mu}_k, \qquad k \in \{0,1\}, \tag{8}$$

$$\Sigma_k^2 = \vec{a}^T V_k \vec{a}, \qquad k \in \{0,1\}. \tag{9}$$

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

$$J(\vec{a}) = \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2}. \tag{10}$$

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

$$J(\vec{a}) = \frac{\vec{a}^T B \vec{a}}{\vec{a}^T W \vec{a}}. \tag{11}$$

mit

$$B_{ij} = (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j, \tag{12}$$

und

$$W_{ij} = (V_0 + V_1)_{ij}. \tag{13}$$

(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

$$W^{-1} B \vec{a} = \lambda \vec{a} \tag{14}$$

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

$$\vec{y} \propto W^{-1}(\vec{\mu}_0 - \vec{\mu}_1) \tag{15}$$

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

Aufgabe 68 *Fisherdiskrimante und Likelihoodverhältnis* **10 Punkte**

Betrachten sie einen Satz von Observablen $\vec{x}$, die unter den Hypothesen H_0 und H_1 durch zwei multidimensionale Gaussverteilungen mit identischen Kovarianzmatrizen $V_0 = V_1 = V$ als Wahrscheinluchkeitsdichtefunktionen beschrieben werden sollen. Die WDFs unter den verschiedenen Hypothesen lauten also:

$$f(\vec{x}|H_k) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu}_k)^T V^{-1} (\vec{x} - \vec{\mu}_k)\right] \qquad k \in \{0,1\}.$$

(i) Zeigen Sie, dass das Likelihoodverhältnis gegeben ist durch

$$r = \frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} = \exp(t),$$

wobei t die Fisherdiskriminante

$$t(\vec{x}) = a_0 + (\vec{\mu}_0 - \vec{\mu}_1)^T V^{-1} \vec{x}$$

mit einem beliebigen Schwellenwert a_0 ist. Dementsprechend ist eine Optimierung des Likelihoodverhältnisses äquivalent zu einer Optimierung der Fisherdiskriminante.

(ii) Benutzen Sie das Bayes-Theorem mit A-Prioriwahrscheinlichkeiten π_0 und π_1 für H_0 und H_1, um zu zeigen, dass die bedingte Wahrscheinlichkeit für H_0 bei gegebenen Daten $\vec{x}$ gegeben ist durch

$$P(H_0|\vec{x}) = \frac{1}{1 + \exp(-t)} = s(t)$$

wobei die $s(t)$ die logistische Funktion ist. Betrachten Sie dazu eine Redefinition des Schwellenwertes von der Form $a_0' = a_0 + \log \frac{\pi_0}{\pi_1}$.