Statistische Methoden der Datenanalyse

Markus Schumacher, Stan Lai, Florian Kiss

Übung XIII

05.2.2013, 08.2.2013

Anwesenheitsaufgaben

Aufgabe 67 *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

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

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

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

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 dieser 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 Pseudoexperimente 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, dass 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/slai/StatisticsCourse/PS12/aufgabe64_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:

$$q = -2x \ln B + 2B + 2x \ln x - 2x.$$

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 das selbe 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/slai/StatisticsCourse/PS12/chi2snippet.C` befindet sich eine definierte Funktion, nebst geeigneten 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 definiert 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:

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

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?

Aufgabe 68 *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

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

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

$$V_{km}^{(j)} = \frac{1}{N}\sum_{N}(x_m^{(j)} - \bar{x}_m^{(1)})(x_k^{(j)} - \bar{x}_k^{(1)}), \tag{2}$$

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-Diskriminante 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 `/home/slai/StatisticsCourse/PS13/aufgabe68_anfang.C` enthält bereits die Konstruktion der Fisher-Diskriminanten gemäß obiger Vorschrift und nach den beschriebenen Wahrscheinlichkeitsdichtefunktionen. Sie sind gerne eingeladen, sich von der Richtigkeit dieses Makros zu überzeugen. Erweitern Sie dieses Makro, um folgende Aufgaben durchzuführen.

(i) Erzeugen Sie nochmals 10000 Signal- und Untergrundereignisse, z.B. für Signalereignisse mittels:

```
x[0]=rnd.Gaus(signalmeanx,signalsigmax);
x[1]=rnd.Gaus(signalmeany,signalsigmay);
```

Ermitteln Sie für jedes erzeugte Wertepaar den Wert der Fisher-Diskriminante und tragen ihn in Histogramme für Signal- und Untergrund ein. Benutzen Sie für Signal- und Untergrund dasselbe "Binning". Stellen Sie die Histogramme für Signal- und Untergrundereignisse in verschiedenen Farben übereinandergelegt dar. Die Linienfarbe des Histogramms können Sie beispielsweise mit `TH1::SetLineColor(kRed)` ändern.

(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 die Reinheit π, die definiert sind durch

$$\epsilon = \frac{N_S(t > t_{\text{cut}})}{N_S} \qquad \pi = \frac{N_S(t > t_{\text{cut}})}{N_S(t > t_{\text{cut}}) + N_B(t > t_{\text{cut}})}. \tag{3}$$

Das Durchfahren des Schnittes kann über die bereits erstellten Histogramme der Fisherdiskriminanten für Signal- und Untergrund erfolgen. Schreiben Sie dazu einfach eine Schleife über die Histogrammbins, ermitteln mittels `hist.Integral(i,nbins+1)` die Anzahl an Signal- bzw. Untergrundereignissen die oberhalb des gerade aktuell betrachteten Bins liegen. Dies entspricht genau der Anwendung eines Schnittes mit dem Schnittwert beim unteren Rand des Histogrammbins. Stellen Sie einen passenden `TGraph` bereit, in den Sie die ermittelten Werte für Effizienz und Reinheit als Wertepaare eingeben. Stellen Sie damit die Reinheit in Abhängigkeit von der Effizienz graphisch dar. Beachten Sie, dass Sie den Fall von 0 Signal- und 0 Untergrundereignissen nach dem Schnitt auf t abfangen müssen, da es sonst beim Ausrechnen der Reinheit zu einer Division durch Null kommt.

(iii) Erstellen Sie einen „Scatterplot“ der Meßgrößen x_1 und x_2 für Signal- und Untergrundereignisse. Benutzen Sie dazu zwei weitere `TGraph`-Objekte. Sie können die Farbe der Punkte mittels `graph.SetMarkerColor(kRed);` verändern um Signal- und Untergrund unterscheiden zu können.

(iv) Zeichnen Sie in diesen Graphen Linien mit konstantem t ein. Benutzen Sie dazu eine Funktion vom Typ `TF1`. Überlegen Sie sich, wie diese Funktion aussehen muss. Betrachten Sie dazu, wie die Fisherdiskriminante berechnet wird. Schreiben Sie sich diese auf, setzen Sie sie auf einen beliebigen Wert t_{cut} und lösen nach y auf. Die übrigbleibenden Parameter sind als Parameter der `TF1`-Funktion zu benutzen und entsprechend mit `SetParameter` zu setzen. Die Linienfarbe der Funktion können Sie mittels `funk.SetLineColor(kGreen)` verändern.

(v) 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.

Hausaufgaben

Aufgabe 69 *Fisherdiskrimante und Likelihoodverhältnis* **0 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 neue Definition des Schwellenwertes von der Form $a_0' = a_0 + \log\frac{\pi_0}{\pi_1}$.