

Statistische Methoden der Datenanalyse

Markus Schumacher, Stan Lai, Florian Kiss

Übung XII

29.1.2013, 01.2.2013

Anwesenheitsaufgaben

Aufgabe 64 *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 ($\tau = 10$ 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 Detektoraufösung 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 Untergrund“ 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 $q(\vec{x}_{\text{Sig.}+\text{UG}}) \equiv q_1$ 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 Massegebiet. Im folgenden soll dieses Massegebiet 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 erwarteten 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 Pseudoexperimente 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 65 *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 $pp \rightarrow H + X \rightarrow \gamma\gamma + X$ -Ereignisse, die die Ereignis Selektion passieren, gegeben ist durch $\sigma_S = 25.4 \text{ fb}$. Der Wirkungsquerschnitt für Untergrundeignisse, die dieselbe Ereignis Selektion 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), Untergrundeignissen (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 Untergrundeignisse \hat{b} unter der Hypothese von Signal plus Untergrund, sowie den Schätzer $\hat{\hat{b}}$ auf die Anzahl der Untergrundeignisse 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 66 *Maximale Separation der Fisherdiskriminante*

10 Punkte

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

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

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

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

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

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

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

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

- (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.