Statistische Methoden der Datenanalyse

Markus Schumacher, Stan Lai, Florian Kiss

Übung IX

07.01.2014, 08.01.2014

Anwesenheitsaufgaben

Aufgabe 45 *Kleinste Quadrat Anpassung an $e^+e^- \to \mu^+\mu^-$ Monte Carlo Daten*

Diese Übung baut auf der in Aufgabe 36 durchgeführten Maximum-Likelihood-Anpassung an "binned"Histogramme auf.

Im Streuungsprozess $e^+e^- \to \mu^+\mu^-$ folgt der Winkel θ des μ^+ Teilchens der theoretischen Verteilung:

$$N(\cos\theta) \propto 1 + \alpha\cos\theta + \beta\cos^2\theta,$$

Der Winkel θ wird relativ zur Strahlachse gemessen. In der Übung sollen Sie einen 'Binned' Kleinste-Quadrat-Fit durchführen, um die Parameter zu finden, welche zur Erstellung der Monte Carlo Ereignisse benutzt wurden. Ein solcher simulierter Datensatz befindet sich in `/home/slai/StatisticsCourse/PS9/MuMuEvGen.root`.

(i) Am besten starten Sie von Ihrer Lösung zu Aufgabe 36. Berechnen Sie zusätzlich zum Wert der log-Likelihood für jedes angenommene Wertepaar (α_i,β_j) den Wert des χ^2, indem Sie folgende Gleichung benutzen:

$$\chi^2(\alpha_i,\beta_j) = \sum_{k=1}^{N} \frac{(n_k - \nu_k(\alpha_i,\beta_j))^2}{\sigma_k^2}$$

wobei σ_k der erwartete Fehler auf die Anzahl der Einträge im Bin k ist und sich ergibt zu $\sqrt{\nu_k}$. Die Theorievorhersage ν_k ist gegeben durch:

$$\nu_k(\alpha_i,\beta_j) = n_{tot} \int_{x_k^{min}}^{x_k^{max}} f(x;\alpha_i,\beta_j)dx$$

Beachten Sie, dass im Gegensatz zur log-Likelihoodfunktion die Gesamtanzahl der Einträge des Histogramms bekannt sein muss, um eine Anpassung durchzuführen. Diese können Sie mittels der Methode `TH1::Integral()` ermitteln.

(ii) Füllen Sie für jedes α_i und β_i den χ^2-Wert in einen zweidimensionalen Graph. Dies funktioniert analog zu einem eindimensionalen Graphen:

```
TGraph2D chi2Graph = TGraph2D(likepoints);
```

wobei `likepoints` die Gesamtanzahl an Kombinationen von α und β ist, für die die Likelihood ausgewertet werden soll.

Jeder Punkt des Graphen kann gesetzt werden durch die `TGraph2D` Funktion:

```
SetPoint(int pointNumber, float alpha, float beta, float chi2)
```

Denken Sie daran, `pointNumber` auszurechnen (dies ist nicht die Zählvariable in den Schleifen über die Werte von α und β!).

(iii) Zeichnen Sie den Graphen (χ^2 gegen α und β) unter Benutzung von

```
chi2Graph.Draw("colz");
```

wobei die Option `"colz"` ROOT die Anweisung gibt, einen Surface-Plot zu erstellen. Weitere Optionen des Draw-Befehls können im 'ROOT User's Guide' in der `TGraph2D` Sektion gefunden werden.

(iv) Ermitteln Sie aus den erzeugten Graph eine Abschätzung für $\hat{\alpha}$ und $\hat{\beta}$ und deren Standardabweichungen $\hat{\sigma}_\alpha$ und $\hat{\sigma}_\beta$. Dazu könnten folgende Befehle nützlich sein:

```
float minimum=chi2Graph.GetHistogram().GetMinimum();
```

gibt das Minimum des Graphen zurück.

```
chi2Graph.GetHistogram().SetMaximum(minimum+1);
```

setzt für die Darstellung die maximale Zeichenebene auf das Minimum des Graphen plus 1.0.

(v) Was ändert sich, wenn anstatt der erwarteten Fehler für σ_k die Fehler auf die Histogrammeinträge benutzt werden? Diese können mittels der Methode `TH1::GetBinError(int k)` ermittelt werden.

(vi) Vergleichen Sie die Werte Ihres Fits mit denen des ROOT Fitting Paketes, indem Sie die Histogramm-Anpassungsfunktion

```
hist.Fit("functionName", "I");
```

benutzen, wobei `"functionName"` der Name der WDF Funktion ist und die Option `"I"` eine Anpassung mittels der Methode der kleinsten Quadrate ausführt, indem die gegebene Funktion über jedes Bin des Histogramms integriert wird. Beachten Sie, dass die für die Maximum Likelihood Anpassung verwendete Funktion auf Eins normiert ist (was in diesem Fall egal war). Definieren Sie sich deshalb eine zweite Funktion, die auf die Anzahl der Histogrammeinträge geteilt durch die Binbreite normiert ist. Sie können dazu einen weiteren Funktionsparameter einführen und diesen mit der Methode `FixParameter` fest setzen, so dass er in der Anpassung nicht variiert wird.

Hausaufgaben

Aufgabe 46 *Geradenanpassung in Matrixnotation* **8 Punkte**

Betrachten Sie eine Stichprobe vom Umfang N von Messwerten $(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$, die alle denselben Fehler σ auf die unabhängigen Messung von y_i haben sollen. Benutzen Sie die Matrixnotation der Methode der kleinsten Quadrate, um Schätzer $\hat{\theta}_0$ und $\hat{\theta}_1$ auf die Parameter einer anzupassenden Gerade der Form

$$\lambda_i = \theta_0 + \theta_1 x_i = \sum_{j=0}^{1} a_j(x_i)\theta_j = \sum_{j=0}^{1} A_{ij}\theta_j$$

mit $A_{ij} = a_j(x_i)$ anzugeben.

(i) Schreiben Sie zunächst die Matrix A und die Kovarianzmatrix V auf.

(ii) Berechnen Sie die Matrix für die Schätzer $\hat{\theta}$, indem Sie die Matrizen $A^T V^{-1} A$ und $A^T V^{-1}\vec{y}$ berechnen.

(iii) Schreiben Sie die Schätzer $\hat{\theta}_0$ und $\hat{\theta}_1$ in Abhängigkeit der Erwartungswerte und Varianzen von x, y, xy sowie der Kovarianz von x und y auf.

(iv) Wie unterscheidet sich die Rechnung, wenn jeder Messwert einen nicht korrelierten, aber unterschiedlichen Fehler σ_i auf y_i hat? Schreiben Sie die Matrizen für V, $A^T V^{-1} A$, $A^T V^{-1}\vec{y}$ und somit $\hat{\theta}$ auf.

Aufgabe 47 *Methode der KQ mit Zwangsbedingungen: Winkelmessung im Dreieck* **8 Punkte**

In der Vorlesung wurde gezeigt, dass, an einen Satz von Messungen $\vec{y} = (y_1,y_2,\ldots,y_N)$ eine lineare Funktion $A\vec{\theta}$ anzupassen, ein Fit mit der Methode der Linearen Kleinsten Quadrate durchgeführt werden kann. Dieser minimiert die Größe

$$\chi^2 = (\vec{y} - A\vec{\theta})^T V^{-1} (\vec{y} - A\vec{\theta}),$$

wobei V die Kovarianzmatrix der Messungen $\vec{y}$ ist.

Weiterhin wurde gezeigt, dass unter Berücksichtigung eines Satzes von K Randbedingungen $\vec{b} = (b_1,b_2,\ldots,b_K)$, welche die Gleichungen $B\vec{\theta} - \vec{b} = 0$ erfüllen, die Methode der kleinsten Quadrate durch Minimierung der Größe

$$\chi^2 = (\vec{y} - A\vec{\theta})^T V^{-1} (\vec{y} - A\vec{\theta}) + 2\vec{\lambda}^T (B\vec{\theta} - \vec{b})$$

verbessert werden kann, wobei $\vec{\lambda} = (\lambda_1,\lambda_2,\ldots,\lambda_K)$ ein Vektor von Lagrange-Multiplikatoren ist. Folglich muss χ^2 minimiert werden in Bezug auf $\vec{\theta}$ und $\vec{\lambda}$. Die Lösung dieses Minimierungsproblems ergibt die Schätzer

$$\hat{\vec{\theta}} = C^{-1}\vec{c} - C^{-1}B^T V_B^{-1}(BC^{-1}\vec{c} - \vec{b}) = C^{-1}\vec{c} - C^{-1}B^T\hat{\vec{\lambda}}$$

wobei $C \equiv A^T V^{-1} A$, $\vec{c} \equiv A^T V^{-1}\vec{y}$ und $V_B \equiv BC^{-1}B^T$. Die Varianz errechnet sich daher zu

$$V[\hat{\vec{\theta}}] = C^{-1} - (BC^{-1})^T V_B^{-1}(BC^{-1}).$$

Die verbesserten Schätzer für die Messungen sind gegeben durch:

$$\hat{\vec{\eta}} = A\hat{\vec{\theta}} = A\left[C^{-1}\vec{c} - C^{-1}B^T V_B^{-1}(BC^{-1}\vec{c} - \vec{b})\right],$$

mit der Varianz

$$V[\hat{\vec{\eta}}] = AV[\hat{\vec{\theta}}]A^T = A\left[C^{-1} - (BC^{-1})^T V_B^{-1} BC^{-1}\right]A^T.$$

Betrachten Sie nun die Messung von drei Winkeln eines Dreiecks analog zu dem Beispiel aus der Vorlesung. Die drei unkorrelierten Messungen sind gegeben durch $\vec{y} = (y_1,y_2,y_3)$ mit $\sigma_i = \sigma$ und an diese soll die Funktion $A\vec{\theta}$ mit $\vec{\theta} = (\theta_1,\theta_2,\theta_3)$ angepasst werden, wobei A die Einheitsmatrix in drei Dimensionen ist.

(i) Was sind die Werte für B und $\vec{b}$?

(ii) Berechnen Sie $\hat{\vec{\theta}}$ und $V[\hat{\vec{\theta}}]$ und somit $\hat{\vec{\eta}}$ und $V[\hat{\vec{\eta}}]$.

(iii) Wie verbessert die Zwangsbedingung auf die Messungen die Schätzungen der gemessenen Werte der Dreieckswinkel?

Aufgabe 48 *Fabrikü̈berprüfung* **4 Punkte**

Eine Behörde kontrolliert aufgrund eines Umweltschutzgesetzes, ob Fabriken verbotene Abfälle in Gewässer einleiten. Mittels eines Hypothesentests, basierend auf der gemessenen Schadstoffkonzentration in der Nähe einer Fabrik, wird entschieden, ob jene überprüft werden soll.

Es wird angenommen, dass die Schadstoffkonzentration in der Nähe einer korrekt betriebenen Fabrik gaußverteilt sei mit Mittwelwert $\mu_0 = 0.1$ mg/L und Standardabweichung $\sigma_0 = 0.02$ mg/L. Die Werte im Fall einer Fabrik, welche unrechtmäßig Abfälle entsorgt, seien $\mu_1 = 0.19$ mg/L und $\sigma_1 = 0.05$ mg/L (ebenfalls gaußverteilt).

Man fordert die Überprüfung einer Fabrik, wenn die Konzentration größer als 0.14 mg/L ist.

(i) Was ist die Signifikanz des Tests? (D.h. die Wahrscheinlichkeit, dass eine Fabrik überprüft wird, obwohl sie keine Schadstoffe deponiert.)

(ii) Wie groß ist die Mächtigkeit des Tests gegen die Hypothese, dass die Fabrik nicht kontrolliert werden muss? Wie groß ist die Wahrscheinlichkeit, dass die Fabrik nicht kontrolliert wird, obwohl sie Schadstoffe deponiert?

(iii) Nehmen Sie jetzt an, dass 40% der Fabriken nach Vorschrift betrieben werden. Was ist der Anteil der Fabriken, die nach Vorschrift betrieben werden, aber dennoch von der Behörde überprüft werden müssen?