Université Claude Bernard Lyon 1
Master Physique

Novembre 2024

Rapport

Méthodes statistiques de traitement des données

TP Minimisation

Auteurs : *BUFFAT Dimitri, MAAMAATUAIAHUTAPU Tetautahi*

1. Introduction

Aujourd'hui, nous disposons d'une quantité importante de données issue de mesure que l'on doit pouvoir exploiter. Mais celles-ci comportent systématiquement du bruit, les rendant inexploitables sans un traitement adapté. L'objectif de ce TP est de simulé une mesure contenant un signal et du bruit dont on connaîtra toutes les caractéristiques à l'avance. Ensuite, à partir de ces données, on appliquera une méthode de traitement vu en cours et on comparera les résultats attendues avec deux utiliser pour la simulation. D'autre part, nous feront la distribution du pull. Et enfin, on utilisera notre code sur des données similaires mais dont on ne connaîtra pas les caractéristiques afin de savoir si oui ou non il y a un signal et qu'elle forme il possède.

2. Génération des données

Pour générer nos pseudo-données, nous allons utiliser la fonction de distribution suivante :

$$f(x) = (1-k)Nae^{-bx} + kN\frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

Où k est la fraction de signal, N le nombre total d'événements, a un facteur de normalisation à calculer pour que ae^{-bx} soit une densité de probabilité, b la pente du bruit de fond exponentiel, μ la position du pic du signal et σ sa largeur.

On a donc le premier terme qui correspond au bruit et le second à notre signal. On s'attend alors à avoir, graphiquement, une courbe dominé par une diminution inversement exponentielle comportant à un moment une gaussienne.

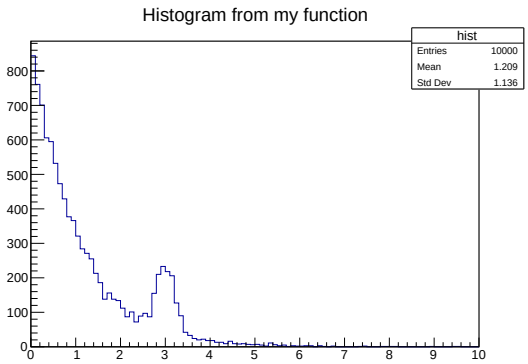

Figure 1 – Histogramme généré.

Nous avons obtenus ce que nous attendons, on va maintenant essayer de retrouver la fraction de signal qui est la paramètre caractérisant nos pseudos-données.

3. Scan du Likelihood

Pour retrouver le paramètre k, nous allons utiliser la méthode de maximum de vraisemblance.

Pour cela, nous allons définir un modèle dont la formule sera la même que précédement à l'exception que k sera maintenant une variable. On va ensuite définir la fonction de vraisemblance.

$$\mathcal{L} = \prod_{i=1}^{N} \text{Poisson}(d_i; f(x_i; k))$$

Numériquement, nous avons besoin de minimiser le *negative log likelihood*

$$l = -log\mathcal{L}$$

Autour du minimum on aura :

$$l(k) = l(\hat{k}) + \frac{1}{2}(k-\hat{k})^2 \left.\frac{d^2 l}{dk^2}\right|_{k=\hat{k}} + ...$$

Avec :

$$\sigma^2 = \left(\frac{d^2 l}{dk^2}\right)^{-1}_{k=\hat{k}}$$

En pratique on a donc une parabole autour de la valeur recherchée. On va donc réaliser un scan du likelihood sur l'ensemble de notre plage de données et voir si une parabole en ressort à un endroit.

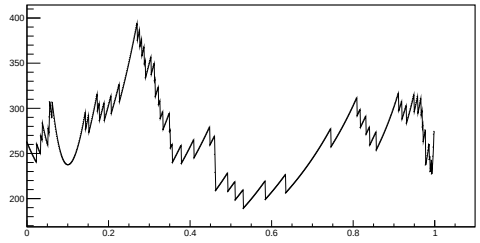

FIGURE 2 – Scan complet du likelihood.

On remarque une parabole autour de $k = 0.1$, on va donc zoomer autour de ce point et réaliser un fit avec une parabole.

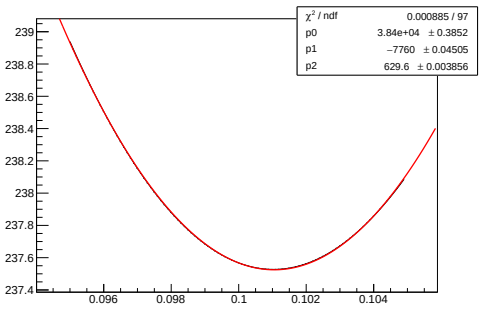

FIGURE 3 – Fit sur la partie parabolique du likelihood.

Ce fit va nous fournir les coefficients a, b et c que l'on fixera dans cette parabole afin d'en déduire la valeur de k pour laquelle on est à son minimum.

Pour l'incertitude de k, on a vu que σ est en réalité contenu dans le coefficient a :

$$a = \frac{1}{2\sigma^2}$$

Ainsi on obtient :

$$k = 0.101053 \pm 0.003609$$

La valeur utilisée pour générer les pseudo-données était $k = 0.1$ validant ainsi notre méthode pour le scan du likelihood.

4. Distribution du pull

L'histogramme des pulls nous permet d'évaluer l'écart entre les valeurs ajustées et la valeur "vraie" du paramètre k du modèle. Nous avons donc généré 10000 lots de pseudo-données, et pour chaque lot, effectué un scan du likelihood et calculé le pull à l'aide de la formule :

$$pull = \frac{k_{exp} - k_{true}}{\sigma_{exp}}$$

Nous attendons que la distribution des pulls soit une gaussienne centrée autour de 0 avec un écart-type de 1 si le modèle décrit correctement les données.

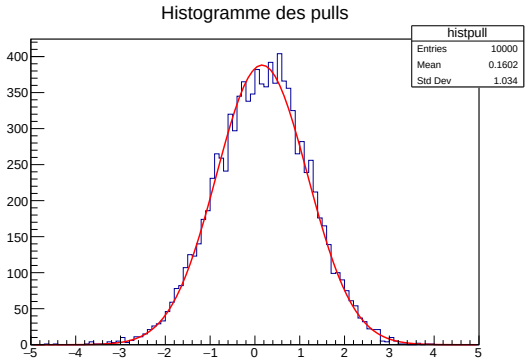

FIGURE 4 – Distribution des pulls pour 10000 lots de pseudo-données générés à partir du modèle f.

chi2	NDf	Mean	Sigma
103.395	74	0.158465 ± 0.0103374	1.01749 ± 0.0070142

TABLE 1 – Paramètres obtenus lors de l'ajustement de la distribution des pulls avec une fonction gaussienne.

Un test de Kolmogorov-Smirnov ($p - value = 0.31 > 0.05$) suggère que la distribution suit bien une loi normale. Bien que ce résultat soit attendu, il faudrait réaliser d'autres tests de normalités pour affirmer ce résultat sans aucun doutes.

Néanmoins, nous avons réalisé un fit gaussien de cette histogramme et obtenu les résultats du tableau 1. La valeur moyenne décalée de 0 suggère un léger biais dans le modèle, le fit ou le calcul des pulls. Un calcul additionnel de la somme

$$\frac{1}{N}\sum_i k^i_{exp} - k_{true}$$

nous a permis d'estimer le biais à $5.3 \cdot 10^{-4}$. Cela suggère que la procédure d'ajustement et le modèle adopté sont globalement précis, avec une déviation systématique très faible.

5. Bonus : Challenge.root

Pour déterminer le nombre d'événement de signal, on répète le même procédé. En faisant un scan du likelihood sur toute la plage de données, on trouve une parabole à l'extrémité gauche sur lequel on va zoomé.

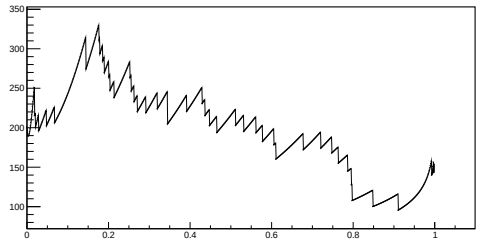

FIGURE 5 – Scan complet du likelihood.

En faisant un fit et en appliquant les mêmes formules que dans les précédantes parties, on trouve :

$$k = 0.00349627 \pm 0.000882082$$

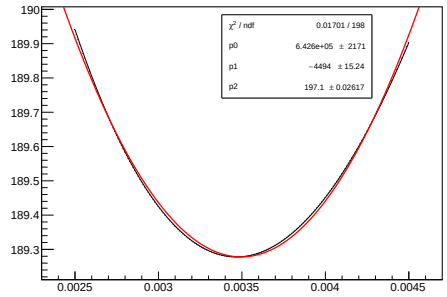

FIGURE 6 – Fit sur la partie parabolique du likelihood.

Comme on a :

$$k = \frac{N_k}{N}$$

Avec $N = 10000$, on a :

$$N_k = 349.627 \pm 88.208$$