Application of Maximum likelihood estimate in signal separation *

Ziwei, Huang[1†]
1 Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510275, China

Abstract: Separating signals of interest from background noise is essential for identifying and classifying different events in particle physics experiments. To effectively separate the two types of events mixed in one observation, we employed a fully statistically based method called the maximum likelihood estimate, which appears to have a high classification performance. A simulated observation containing two classes of normally distributed events was generated for demonstration. Then, the maximum likelihood function was defined based on the observation, following which we minimized the maximum likelihood function to get the estimated parameters. The result was further verified in the likelihood function space, and the confidence interval was calculated to evaluate the accuracy. We also changed the proportion or the total number of events contained in the simulated observation, and repeated the process to probe into how the confidence interval responds to the properties of the observation. We found that the maximum likelihood estimate is an effective framework for identifying and classifying different normally distributed events in a complex observation. The confidence interval expends almost linearly as the total number of events increases, and the difference in confidence intervals of the two classes of events remains nearly consistent if the proportion remains consistent. This study reveals the effectiveness of the maximum likelihood estimate in separating normally distributed events, providing a powerful tool for analyzing observation in particle physics experiments.

Keywords: Maximum likelihood estimate, Particle physics experiment, Signal separation

1 Introduction

DUE to some limitations of current commonly used particle signal detectors, the raw data obtained from a single observation may contain signals of many types of particle events. For example, many detectors collect signals restricted in a specific energy interval, but it is quite common that many different types of particle event may share a similar characteristic energy level, e.g., solar neutrinos signals and radioactive decay signals. This may lead to confusion when handling with these data.

To exclude the influence of background noise and get a specific type of signal of interest, it is essential to property classify these mixed events into different types. One robust approach is to take advantage of the characteristics of the statistic distribution of different types of events. Although signals of different classes of events may happen to have similar characteristic energy level, they may appear to show distinct statistic distributions. The determination of the classical model of statistic distribution of a certain type of event is knowledge-based, while the parameters are unknown. Previously, researchers might have to do this work manually according to their experience, but today with the aim of powerful computers, efficient algorithms can be applied to separate different signals accurately and automatically.

In this work, we employed a fully statistically based method called a maximum likelihood estimate to separate different types of signal from a single observation, and the model achieved high classification performance.

2 Experimental principles and method

2.1 Experimental principle[1]

2.1.1 Maximum likelihood estimate (MLE)

Maximum likelihood estimate, a fully statistical-based algorithm for parameters estimation based on frequency, is an efficient method for classifying two types of events with different distribution patterns. The core principle of this algorithm is to properly select a set of parameters θ^* to maximize the maximum likelihood function L, which is defined as the probability of joint distribution of multiple samplings:

$$\theta^* = argmax L(x_1, x_2, ..., x_n; \theta) \tag{1}$$

$$L(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{2}$$

while $f(x_i; \theta)$ is defined as the probability density function.

For example, assume that we have two types of

*Supported and taught by Luyoutang, School of Physics, Sun Yat-sen University

†Corresponding author. ID: 20980066 Email: huangzw29@mail2.sysu.edu.cn

normally distributed events, A and B:

$$A : f(x, \mu_1, \sigma_1) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$$
$$B : f(x, \mu_2, \sigma_2) = \frac{2}{\sigma_2\sqrt{2\pi}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$$

For a given energy interval, we can evenly divide the interval into small intervals M. Then, for the i th interval, we can calculate the number of events falling it. We can always choose a large enough M to promise that not too much events are included in every interval. If this condition is satisfied, we can assume that the distribution of events in every interval conforms to Poisson distribution.

$$\pi_i = \frac{\lambda_i^{k_i} e^{-\lambda_i}}{k_i!} \tag{3}$$

when

$$\lambda_i = \sum_j \mu_j H_{ij} \tag{4}$$

λ_i is the expected number of events included in the i th interval; k_i is the actual number of events included in the i th interval; H_{ij} is the normalization factor; μ_j is the expected number of events of type j included in the interval i, $j = A, B$.

Now, we can define the maximum likelihood function as:

$$\begin{aligned} L(k, \lambda) &= \prod_{i=1}^{n} \pi_i \\ &= \prod_{i=1}^{n} \frac{\lambda_i^{k_i} e^{-\lambda_i}}{k_i!} \end{aligned} \tag{5}$$

To reduce computational complexity, we can take the logarithm of L:

$$-lnL = \sum_i \lambda_i - k_i ln(\lambda_i) \tag{6}$$

Our goal is to estimate a suitable set of parameters θ^* to minimize $-lnL$. This can be done using a lot of existing well-founded methods. In our work, we chose to use the *scipy.optimize.minimize*() method implanted in Python Scipy modules (See scipy.optimize.minimize — SciPy v1.8.0 Manual.

2.2 The confidence interval

To evaluate the performance and accuracy of the maximum likelihood estimate on signal separation, we calculated the confidence interval of the result. When the total number of events is large enough, the distribution of the maximum likelihood function is similar to the χ^2 distribution:

$$\chi^2 = -2ln\frac{\widetilde{L}(k; \mu_1, \mu^*_{i\neq 1})}{L(k; \mu^*)} \tag{7}$$

L is the maximum distribution, while $\widetilde{L}$ is the distribution of other parameters when a specific parameter reaches its maximum distribution.

We have:

$$\chi^2 = 2(\widetilde{l} - l) = 2\Delta l \tag{8}$$
$$l = -lnL \tag{9}$$

To calculate the confidence interval, we can keep a specific parameter at its maximum distribution, while changing other parameters one by one to obtain the μ_l and μ_h that satisfies the condition $\chi^2 = 1$, i.e. $\Delta l = 0.5$. The confidence interval is defined as

$$[\mu_l, \mu_h] \tag{10}$$

2.3 Materials and instruments

2.3.1 Data

For demonstration, we generated two sets of normally distributed signals and concatenated them into one sequence.

1. Set A: 100 events; $\mu_1 = 10MeV$; $\sigma_1 = 2MeV$
2. Set B: 200 events; $\mu_2 = 15MeV$; $\sigma_2 = 3MeV$

To investigate how the confidence interval responds to the portion and total number of events of A and B, we also generated a sequence of data sets and repeated the process (Table 1 and Table 2, details in the Supplementary information).

	Class A	Class B
Data1	100	200
Data21	150	200
Data22	200	200
Data23	250	200
Data24	300	200

Table 1: **Change the proportion of the two events**

	Class A	Class B
Data31	50	100
Data32	75	150
Data1	100	200
Data33	125	250
Data34	150	300

Table 2: **Change the total number of the two events**

2.3.2 Platform

The analysis was conducted in python (3.9.7). The version of packages used in this study is described in detail in the supplemental information.

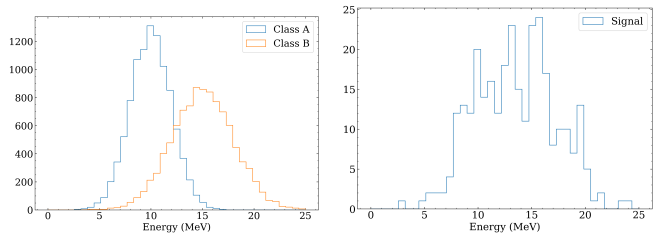
2.4 Method

1. We first generated two classes of events according to the parameters described above and concatenated them into one sequence.
2. Next, the maximum likelihood function $-lnL$ was defined based on the simulated data.
3. Then, we used *minimize* function in *Scipy* to minimized the maximum likelihood function and obtained the maximum likelihood estimated parameters.
4. Finally, the confidence interval was calculated to evaluate the accuracy.

3 Result

3.1 Generation of the simulated data

np.random.normal was used to generate two sets of simulated normally distributed events. Fig. 1a shows the distribution of 100,000 class A events and 100,000 class B events, showing that they actually conform to the normal distribution. 100 Class A events and 200 Class B events generated using the same model were then mixed to generate a simulated observation, which conformed the mixture normal distribution (Fig. 1b).



(a) (b)

Figure 1: **The distribution of the signals**

4 Estimate the maximum likelihood parameters and fit the model

According to Equation 5, the maximum likelihood function of the given data was initially defined. To reduce the computational complexity, we further took the logarithm of L (Equation 6). Then, the estimate parameters θ^* were calculated by minimizing the $-lnL$. The results were verified in the likelihood function space (Fig. 2).

The resulting estimated parameters θ^* show a high consistency with our initial setting (Table 3), showing that the maximum likelihood estimate is an efficient method for separating two normally distributed signals.

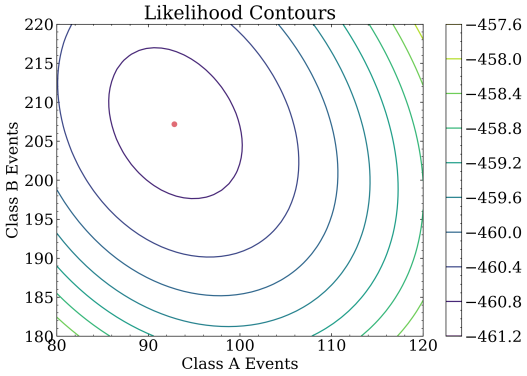


Figure 2: **Verify the results in the likelihood function space**

	Class A	Class B
Setting	100	200
Estimate	92.837	207.163

Table 3: **Compare the estimate to the setting**

Basing on the estimated parameters, the fitting signals of class A and class B together with the total fit were generated and compared with the original observation (Fig. 3). The performance of the model was satisfactory.

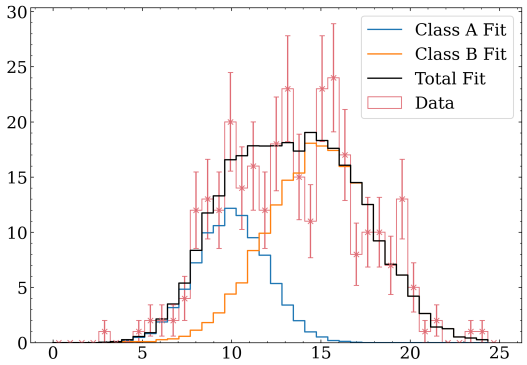


Figure 3: **Fitting the model**

5 Calculate the confidence interval

To further evaluate the accuracy of the model, the confidence interval was calculated according to Equation 9. The profiles of the event A and B were plotted (Fig. 4), and the confidence interval of the results are:

Class A $92.84^{+12.63}_{-11.89}$

Class B $207.16^{+16.65}_{-15.88}$

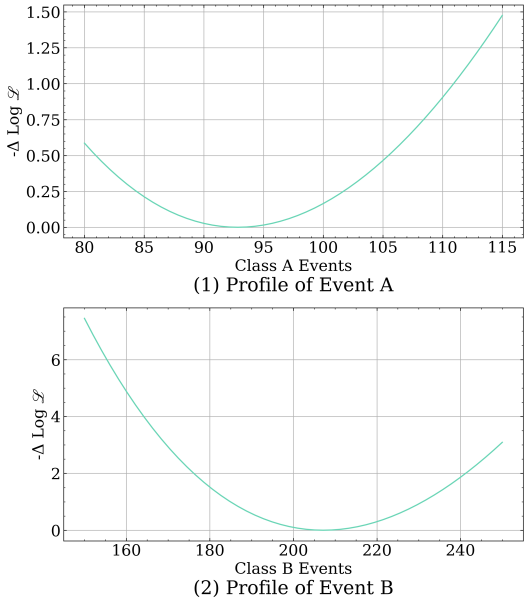
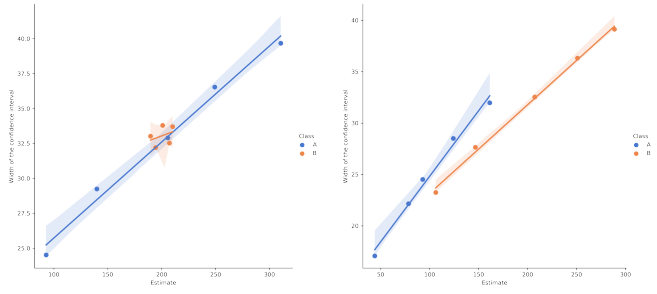


Figure 4: **Profiles of event A and B**

6 Confidence interval responds to the properties of the observation

To probe into how the proportion and the total number of events in the observation affect the confidence interval, we generated a series of observation (Table 1 and Table 2) and test the confidence intervals. The results show that the width of the confidence interval is perfectly positively correlated with the estimate number of the events (Fig. 5a, $r_A = 0.994$; and Fig. 5b $r_A = 0.994$, $r_B = 0.998$; See detailed regression parameters in the supplementary information)



(a) (b)

Figure 5: **Confidence interval responds to proportion and the total number of the observation**

7 Conclusion and Discussion

In this research we employed the maximum likelihood estimate on mixed observations and prove that it is an efficient algorithm to identify and classify different normally distributed events in a complex observation. We also found that the confidence interval expends almost linearly as the total number of events increases.

7.0.1 Known limitation

However, there are still a few known limitations of this method when dealing with the real-world data:

1. In our experiment, we found that when the total number of events increases to about 1000, the algorithm becomes unstable and sometimes the maximum likelihood function cannot converge, which shows that the robustness of the algorithm should be further interrogated.

2. As shown in the results, the width of the confidence interval expands almost linearly as the number of events increase, thus there is a consideration that whether the confidence interval is acceptable when dealing with real-world big data.

3. In this experiment we only demonstrated separating two normally distributed signals from an observation using MLE, but whether this algorithm is compatible with signals conforming other distribution patterns is still unknown, which requires further research.

Reference

[1] SHEN H. General Physics Laboratory[M]. [S.l.]: Science press, 2015 (: 1).