Chi-square fitting - 2

Chih-Keng Hung

May 5, 2025

1 Practice 1

Purpose:

1. To connect the concept of χ^2 with curve fitting.
2. To understand the degree of freedom of the model.

Task:

1. Generate 2000 sets of data points defined by a function $y = a * x + b + noise$ with `x = np.linspace(1, 10, 10)`.
2. The noise is sampled from a normal distribution with mean $= 0$ and standard deviation $= \sigma_0 = 3$.
3. Fit the data points set by set with the function $y = a * x + b$ using `curve_fit`.
4. Now calculate the χ^2 for each set of data points. The standard deviation σ_j may vary slightly but remains close to σ_0.
5. $\chi_i^2 = \sum_{j=1}^{10} \frac{(data\_point_{i,j} - fitted\_result_{i,j})^2}{{\sigma_j}^2}$ $(i = 1 \sim 2000)$, where i is the index of data set and j is the index of x value in one data set.($\texttt{len}(\chi^2) = 2000$)
6. You might find out that this χ^2 is just the sum of the square of the residuals divided by the standard deviation, and it is anologous to previous practice: summing up the χ^2 of each normal distribution sample.
7. Plot all data points you have generated by $y = a * x + b + noise$.
8. Plot the histogram of these χ^2 obtained from each set of data points.(A set of data points is a 10 points line with noise)
9. Compare the histogram with the chi-square distribution.
10. The term **model** is often used in the statement of goodness of fit. What does **model** mean in this case?
11. State what **overfitting** is.
12. Design an experiment to illustrate the **overfitting** phenomenon.
13. By making slight adjustments to your code, you can effectively achieve this goal.
14. Describe your observations based on the modifications made.
15. Will this phenomenon occur when analyzing the data from the General Physics Laboratory experiments, such as the pendulum or gravity experiment?
16. Can we determine if the model is overfitting the data points obtained by unknown functions? Design your own experiment to prove your point of view.

Hint:

1. The histogram of χ^2 should be similar to the chi-square distribution.
2. The DOF of the model used for fitting plays a crucial role in shaping the histogram's distribution.
3. The original data points are generated using the function $y = a * x + b + noise$.

2 Practice 2

Purpose:

1. To clarify the concept of goodness of fit.

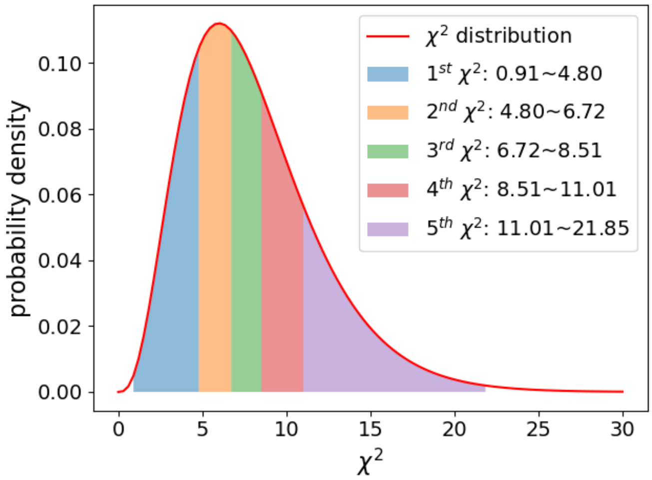

Figure 1: The χ^2 distribution in Practice 1 separated in five sections. Each section contains 400 sets of data points. This figure is just an example, you probably get different χ^2 values in each section.

Task:

1. Show the histogram of a and b in **Practice 1** fitted in separated five sections shown in Figure 1.
2. Each section contains 400 sets of data points, grouped as follows: 1–400, 401–800, 801–1200, 1201–1600, and 1601–2000, ordered by χ^2 value.
3. Remember the dataset or the fitted result is not sorted by the χ^2 value at the beginning.
4. You'll need to get those a and b values sorted by the χ^2 value and choose certain indexes that are in the section to plot the histogram.
5. Will a and b be closer to what you originally used to generate the data points in the section with smaller χ^2 value?
6. Is it appropriate to assess the goodness of fit based solely on the accuracy of the fitted a and b values?

3 Practice 3

Purpose:

1. To understand the covariance matrix and the correlation coefficient more deeply.

Task:

1. By using the data points generated in **Practice 1**, get the mean and standard deviation of these data points.
2. Fit these mean value of y with the function $y = a * x + b$ using `curve_fit`, and remember to set *absolute_sigma=True*.
3. By using the result in **Practice 1**, you can show the histogram of those a and b values via `plt.hist` and `plt.hist2d` methods.(2000 sets of data points)
4. Compare the standard deviation and correlation coefficient obtained from the fitted covariance matrix with those derived from statistical graphs of a and b.
5. Perform a linear fit on the 2D histogram of parameters a and b to obtain the slope, which characterizes certain properties useful for calculating the correlation coefficient.
6. Ensure the definition of those quantities you are calculating is correct.
7. Finally, you should be able to find everything is consistent with the covariance matrix.(the values will be close to each other)
8. Explain how you calculate the correlation coefficient from a and b values.