UCSB, Physics 129L, Computational Physics
Lecture notes, Week 7

Zihang Wang (UCSB), zihangwang@ucsb.edu

February 22, 2025

Contents

1 Statistical inference

1.1 Gaussian fluctuation, likelihood function, and maximum likelihood estimator

Let's say you have N measurements from an experiment.

- We are given N number of data measurements $\mathbf{x}, \mathbf{y} = (x_i, y_i)$.
- We have a proposed model for the data $y = y(x_i|\boldsymbol{\theta})$, governed by M hyper parameters $\boldsymbol{\theta}$.
- Let's say the fluctuation associated with each measurement is Gaussian distributed with standard deviation, σ_i and the measurements are uncorrelated.

Then, the probability of the probability associated with the Gaussian fluctuation at i is given by,

$$p(y_i|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-[y(x_i|\boldsymbol{\theta}) - y_i]^2}{2\sigma_i^2}\right), \tag{1}$$

and the joint probability over all measurements is the **likelihood function** $\mathcal{L}(\boldsymbol{\theta})$, which has the following product form,

$$\mathcal{L}(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta}) = \prod_i p(y_i|\boldsymbol{\theta}). \tag{2}$$

In calculations, since the above products are hard to evaluate, we define the **log-likelihood function**,

$$\mathcal{L}(\boldsymbol{\theta}) \to \log[\mathcal{L}(\boldsymbol{\theta})], \tag{3}$$

such that at the saddle points, derivatives become zero with unique correspondence,

$$\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = 0 \to \mathbf{S}_{\boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \log[\mathcal{L}(\boldsymbol{\theta})]) = \frac{1}{\mathcal{L}(\boldsymbol{\theta})}\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = 0, \tag{4}$$

where $\mathbf{S}_{\boldsymbol{\theta}}$ is the **score function**, defined as the gradient of the log-likelihood function. This is called the **maximum likelihood estimation** (MLE). You may also notice that in general,

$$\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta})\mathbf{S}_{\boldsymbol{\theta}}. \tag{5}$$

For Gaussian fluctuation, we have the **χ-square distribution**,

$$-2\log[\mathcal{L}(\boldsymbol{\theta})] \sim \sum_i \left(\frac{[y(x_i|\boldsymbol{\theta}) - y_i]^2}{\sigma_i^2} \right) = \chi^2, \tag{6}$$

and we set the derivative to zero,

$$0 = \frac{\partial}{\partial \boldsymbol{\theta}} \log[\mathcal{L}(\boldsymbol{\theta})]) = \frac{\partial}{\partial \boldsymbol{\theta}}\chi^2 = \sum_i \left(\frac{[y(x_i|\boldsymbol{\theta}) - y_i]}{\sigma_i^2} \frac{\partial y(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) = 0. \tag{7}$$

The above expression suggests that, for Gaussian distributed fluctuations, the maximum likelihood estimator $\boldsymbol{\theta}_M$ is the solution that minimizes the χ-square distributions (system of equations).

Let's consider the following example: You are tossing a coin and want to determine whether it is fair, i.e., the probability of getting heads is 0.5. One can sample the "fairness" of the coin by sampling an underlying distribution by repeatedly tossing the coin. In this case, the "Data" $\mathbf{y}$ will be a sequence of upside and downsides, and the hyper parameters will be the probability $\boldsymbol{\theta}$. The likelihood function will be a product of N Bernoulli trials $y_i = 0, 1$, leading to the binomial distribution with k upside coins,

$$\mathcal{L}(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta}) = \prod_i p(y_i|\boldsymbol{\theta}) = \binom{N}{k}\theta^k(1-\theta)^{N-k}, \tag{8}$$

and the log-likelihood function is,

$$\log L(\theta \mid k) = \log\left(\binom{n}{k}\right) + k\log\theta + (n-k)\log(1-\theta), \tag{9}$$

and,

$$\frac{d}{d\theta}\log L(\theta \mid k) = \frac{k}{\theta} - \frac{N-k}{1-\theta} = 0. \tag{10}$$

This gives the usual result we see when estimating the probability for coin tossing,

$$\theta = \frac{k}{N}. \tag{11}$$

1.2 Maximum Likelihood Estimation: Linear Regression

Linear regression is one of the most important statical model in physics, and it can be derived via the maximum likelihood. The model has the form,

$$y(x) = a + bx, \tag{12}$$

and we want to compare this model against a tuple (x_i, y_i).

The log-likelihood function is given by,

$$-2\log(\mathcal{L}) \sim -2\log\left(\prod_i p(x_i, y_i \mid a, b)\right) = \sum_i \frac{(y_i - y(x_i \mid a, b))^2}{\sigma_i^2} = \chi^2(a, b). \tag{13}$$

We minimize the χ^2 with respect to the controlling parameters a, b,

$$\begin{aligned} 0 &= \frac{\partial \chi^2}{\partial a} = -2\sum_{i=1}^{N} \frac{y_i - a - bx_i}{\sigma_i^2}, \\ 0 &= \frac{\partial \chi^2}{\partial b} = -2\sum_{i=1}^{N} \frac{x_i(y_i - a - bx_i)}{\sigma_i^2}. \end{aligned} \tag{14}$$

For simplicity, let's define the following quantities,

$$U_1 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}, \quad U_2 = \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}, \quad U_3 = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}, \tag{15}$$

$$U_4 = \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2}, \quad U_5 = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}. \tag{16}$$

The system of equations becomes,

$$aU_1 + bU_2 = U_3, \quad aU_2 + bU_4 = U_5, \tag{17}$$

and in matrix form, it becomes,

$$\begin{pmatrix} U_1 & U_2 \\ U_2 & U_4 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} U_3 \\ U_5 \end{pmatrix}. \tag{18}$$

We solve the above equation,

$$a = \frac{U_4U_3 - U_2U_5}{\Delta}, \quad b = \frac{U_1U_5 - U_2U_3}{\Delta}, \quad \Delta = U_1U_4 - (U_2)^2. \tag{19}$$

This is the **least squares method** or maximum likelihood estimation. It gives the best-fit values of the y-intersection and slope a, b : $\hat{a}$ and $\hat{b}$, and it is the **MLE fixed point**.

1.3 Fisher Information Matrix and Score function, and Cramer-Rao Bound

Since $L(\theta) = p(x_i, y_i \mid a, b)$ is a probability density, we have

$$\int p(x_i, y_i \mid a, b)\, dx = 1. \tag{20}$$

Differentiate both sides with respect to $\theta = a, b$ (assuming we can interchange differentiation and integration under regularity conditions),

$$\frac{d}{d\theta} \int p(x_i, y_i \mid a, b)\, dx = \int \frac{\partial}{\partial \phi} p(x_i, y_i \mid a, b)\, dx = 0. \tag{21}$$

We notice that,

$$\frac{\partial}{\partial \phi} p(x_i, y_i \mid a, b) = p(x_i, y_i \mid a, b) \frac{\partial}{\partial \phi} \log p(x_i, y_i \mid a, b) = p(x_i, y_i \mid a, b) S(\theta). \tag{22}$$

In other words, we can write,

$$\int p(x_i, y_i \mid a, b) S(\theta)\, dx = \mathbb{E}[S(\theta)] = 0. \tag{23}$$

This means that the expectation value of the score function is zero. We take a derivative with respect to a different parameter $\phi = a, b$,

$$\frac{d}{d\phi} \mathbb{E}[S(\phi)] = \frac{d}{d\phi} \int p(x_i, y_i \mid a, b) S(\theta)\, dx = 0. \tag{24}$$

We can differentiate under the integral sign,

$$\int \frac{\partial}{\partial \phi} \left(p(x_i, y_i \mid a, b) S(\theta) \right) dx = 0, \tag{25}$$

and the product rule in the integrand,

$$\frac{\partial}{\partial \phi} \left(p(x_i, y_i \mid a, b) S(\theta) \right) = \frac{\partial p(x_i, y_i \mid a, b)}{\partial \phi} S(\theta) + p(x_i, y_i \mid a, b) \frac{\partial S(\theta)}{\partial \phi}. \tag{26}$$

Recall that,

$$\frac{\partial p(x_i, y_i \mid a, b)}{\partial \phi} = p(x_i, y_i \mid a, b) S(\phi). \tag{27}$$

Plug this back into the integral:

$$\int p(x_i, y_i \mid a, b) \left[S(\theta)S(\phi) + \frac{\partial S(\theta)}{\partial \phi} \right] dx = 0. \tag{28}$$

Recall that the integral implies the expectation value over all samples, and those defines the components of the **Fisher information matrix**,

$$\mathcal{I}_{\theta\phi} = \mathbb{E}\left[S(\theta)S(\phi)\right] = -\mathbb{E}\left[\frac{\partial S(\theta)}{\partial \phi}\right]. \tag{29}$$

Fisher information matrix captures the uncertainty associated with a given maximum likelihood estimation (MLE), and it is directly related to the covariance matrix $C_{\theta\phi}$ when expanding at the MLE fixed point $\hat{a}, \hat{b}$.

The fixed point solution is given by $\hat{a}, \hat{b}$. What about the second order variation of the likelihood function? This is called asymptotic covariance matrix. With Taylor expansion near the fixed point, we have the **Hessian**,

$$\begin{aligned}
&-2\log(\mathcal{L}(a,b)) = -2\log(\mathcal{L}(\hat{a},\hat{b})) \\
&+\frac{1}{2}\left((a-\hat{a})^2\frac{\partial^2}{\partial a^2} + (a-\hat{a})(b-\hat{b})\frac{\partial}{\partial a}\frac{\partial}{\partial b} + (b-\hat{b})(a-\hat{a})\frac{\partial}{\partial b}\frac{\partial}{\partial a} + (b-\hat{b})^2\frac{\partial^2}{\partial b^2}\right)\Bigg|_{\hat{a},\hat{b}}(-2\log(\mathcal{L})) \\
&= -2\log(\mathcal{L}(\hat{a},\hat{b})) - \frac{1}{2}\begin{bmatrix} a-\hat{a} \\ b-\hat{b} \end{bmatrix}^\top \begin{bmatrix} \mathcal{I}_{aa} & \mathcal{I}_{ab} \\ \mathcal{I}_{ba} & \mathcal{I}_{bb} \end{bmatrix} \begin{bmatrix} a-\hat{a} \\ b-\hat{b} \end{bmatrix},
\end{aligned}$$

where $\mathcal{I}$ is the fisher matrix, and the **asymptotic covariance matrix** of its MLE is given by the inverse (remember the standard deviation is at the denominator in a Gaussian),

$$C(\hat{a},\hat{b}) = \mathcal{I}(\hat{a},\hat{b})^{-1}. \tag{30}$$

This shows that a higher Fisher Information implies more precise parameter estimates. If the Fisher Information Matrix were infinitely large, meaning perfect information, the uncertainty would be zero, and we would have an exact estimate with no variation.

From the above fixed point expression, we are able to estimate the uncertainty associated with a given MLE fixed point. This matrix is called **asymptotic covariance matrix** because for large sample sizes, MLEs are asymptotically normally distributed (central limit theorem),

$$\boldsymbol{\theta} \sim \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \frac{\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}}{N}\right), \quad \text{or} \quad \hat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\boldsymbol{\theta}, \frac{\mathcal{I}(\boldsymbol{\theta})^{-1}}{N}\right), \tag{31}$$

where $\mathcal{N}$ is normal distribution, and the former expression is usually used in the **posterior distribution** under **Bayesian inference**, while the latter usually used by **Frequentist** where $\boldsymbol{\theta}$ is taken to be the **true parameter**.

When we **assume $\boldsymbol{\theta}$ is the true parameter**, we can define the **Cramer-Rao Bound**, which establishes the **minimum achievable variance** for any unbiased estimator $\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ of a parameter. It quantifies the precision limit imposed by statistical properties of the underlying data. This implies the covariance matrix is larger or equal than the true fisher information matrix,

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \geq \mathcal{I}^{-1}(\boldsymbol{\theta}), \tag{32}$$

where $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher Information Matrix (FIM), and $\geq$ denotes that $\mathrm{Cov}(\hat{\boldsymbol{\theta}}) - \mathcal{I}^{-1}(\boldsymbol{\theta})$ is positive semi-definite: a matrix M is positive semi-definite if,

$$\mathbf{x}^\top M \mathbf{x} \geq 0 \quad \text{for all nonzero vectors } \mathbf{x}. \tag{33}$$

This means that the exponent is negatively bounded in the Gaussian exponent.

On the other hand, **if we assume θ is a random variable** , we have the **Bayesian Cramer-Rao Bound**,

$$\mathrm{Cov}(\boldsymbol{\theta}) \geq \mathcal{I}^{-1}_{\text{posterior}}(\hat{\boldsymbol{\theta}}) = \left[\mathcal{I}_{\text{likelihood}}(\hat{\boldsymbol{\theta}}) + \mathcal{I}_{\text{prior}}(\hat{\boldsymbol{\theta}})\right]^{-1}, \tag{34}$$

where $\mathcal{I}^{-1}_{\text{likelihood}}(\hat{\boldsymbol{\theta}})$ has the same definition as $\mathcal{I}^{-1}(\boldsymbol{\theta})$ previously.

In the frequentist approach, this bound provides a lower bound on the variance of any unbiased estimator based **solely on the likelihood function**. On the other hand, in the Bayesian approach, if we have strong prior knowledge about θ, this information reduces posterior uncertainty (inversely summed), potentially leading to posterior variances smaller than the frequentist.

Let's go back to the example. Recall that the log-likelihood function in the linear regression model is the χ-square distribution,

$$-2\log(\mathcal{L}) = \chi^2(a, b). \tag{35}$$

This means that the fisher information matrix defines an **ellipse**,

$$\hat{\boldsymbol{\theta}}^T \mathcal{I}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\theta}} = \begin{bmatrix} a - \hat{a} & b - \hat{b} \end{bmatrix} \mathcal{I}(a, b) \begin{bmatrix} a - \hat{a} \\ b - \hat{b} \end{bmatrix} = \chi^2(a, b) = \chi^2_\alpha, \tag{36}$$

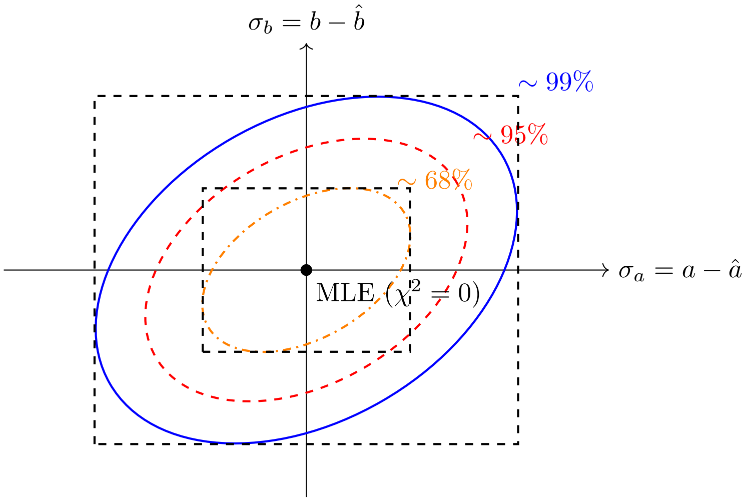
where we define the **confidence levels** for a, b,

$$P(\chi^2 \leq \chi^2_\alpha) = 1 - \alpha. \tag{37}$$

$1 - \alpha$ is the desired confidence level (e.g., 68%, 95%, or 99%). The common confidence levels are obtained from a chi-squared distribution table with a given degree's of freedom. For example, in a chi-squared distribution with 10 degrees of freedom:

- For 68% confidence: $\chi^2 \leq 14.14$
- For 95% confidence: $\chi^2 \leq 18.31$
- For 99% confidence: $\chi^2 \leq 23.21$

In the frequentist approach (or Bayesian will uniform prior), we get the MLE $\hat{a}, \hat{b}$, and we want to find how far the estimation is within a given **confidence level** χ^2_α, as shown in the figure below. To estimate the variance σ_a, σ_b in a, b with respect to the MLE parameter $\hat{a}, \hat{b}$, we can preform the marginalization that projects the ellipse onto the $a - \hat{a}$ and $b - \hat{b}$ axes (black rectangles in the figure below).



As a side note, the **correlation matrix** is derived from the covariance matrix by normalizing the covariances. For a parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_k)^T$, the correlation matrix is defined as,

$$\mathrm{Corr}(\hat{\theta})_{i,j} = \frac{\mathrm{Cov}(\hat{\theta}_i, \hat{\theta}_j)}{\sqrt{\mathrm{Var}(\hat{\theta}_i) \cdot \mathrm{Var}(\hat{\theta}_j)}} = \frac{\left[\mathcal{I}(\hat{\theta}_i)^{-1}\right]_{i,j}}{\sqrt{\left[\mathcal{I}(\hat{\theta}_i)^{-1}\right]_{i,i} \cdot \left[\mathcal{I}(\hat{\theta}_i)^{-1}\right]_{j,j}}}. \tag{38}$$

2 Statistical inference

Statistical inference provides methods for interpreting probability and updating beliefs about populations based on sampled data (or the likelihood function). There are two major approaches to inference **Frequentist** and **Bayesian**.

2.1 Frequentist Interpretation

The frequentist uses the concept of frequency or repeated sampling. It focuses on estimating statistical parameters, e.g. means and variances and making statistical inferences based on the given data. In frequentist statistics, parameters governing the underlying distribution are treated as fixed, unknown values, and the goal is to estimate these parameters using point estimates (e.g., maximum likelihood) or confidence intervals. For example, in the Stern-Gerlach experiment, the underlying distribution is **assumed** to be the binomial distribution and the hyper parameter is $p = p_0$. The frequentist approach **does not incorporate prior beliefs** on the parameter p (it is given and fixed), and it relies solely on the data at hand.

Hypothesis testing is a key component of frequentist statistics: we start with two competing hypotheses, null and alternative hypothesis. The former is a declaration of no difference, whereas the later stands in direct opposition to the null, suggesting a significant difference as an alternative. As an example, in the Stern-Gerlach experiment, the frequentist would first propose a binomial parameter $p_0 = 0.5$ with a null hypothesis: "the probability of electron with spin up configuration is equal to 0.5". Using statistical tests and analysis, such as confidence intervals, the frequentist approach would then assess whether the observed outcomes significantly deviate from the null hypothesis's expected probability of 0.5. The alternative hypothesis usually rejects the null, e.g. "the probability of electron with spin up configuration is *not* equal to 0.5".

For example, for a linear regression model, the null hypothesis (H_0) is that the slope is 0 (no relationship between x and y):

$$H_0 : b = 0 \tag{39}$$

The alternative hypothesis (H_a) is that the slope is not 0 (there is a relationship between X and Y):

$$H_a : b \neq 0 \tag{40}$$

To test the hypothesis, we need to calculate the t-statistic for the slope:

$$t = \frac{b}{\sigma_b} \tag{41}$$

If the absolute value of the t-statistic exceeds the critical value (from a known table), we reject the null hypothesis H_0.

2.2 Bayesian Inference

The Bayesian approach is based on Bayes' rule,

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}. \tag{42}$$

To be explicit, we have,

$$P(\text{Parameter}|\text{Data}) = \frac{P(\text{Data}|\text{Parameter})P(\text{Parameter})}{P(\text{Data})}, \tag{43}$$

where "Data" represents the observed data, and "Parameter" represents the variables that govern the underlying probability distribution function from which the observed data is drawn.

$P(\text{Parameter}|\text{Data})$ is referred to as the **posterior** probability, representing the conditional probability of governing parameters given a set of observed data points. $P(\text{Data}|\text{Parameter})$ is the **likelihood** function that captures the probability of obtaining the observed data with the specified set of controlling parameters. $P(\text{Parameter})$ is referred to as a **prior** distribution, which contains the assumptions regarding the probability distribution of the controlling parameters. $P(\text{Data})$ is known as the **evidence**, representing the overall probability of observing a particular dataset. Since we are mainly interested in the relative probability of various controlling parameters. Therefore, we absorb it into the normalization coefficient. It is worth noting that the Bayes' rule is independent of any inference we will soon discuss.

Bayesian inference treats **both observed data and parameters as random variables** with probability distributions, in contrast to the frequentist approach. Bayesian inference typically begins with prior beliefs or knowledge about the controlling parameters. For instance, in the Stern-Gerlach experiment, one might assume a prior probability distribution for the parameter p, such as a uniform distribution. The likelihood function quantifies the probability of observing the data with a specific set of controlling parameters and often relies on certain assumptions. For example, it might be described by a binomial distribution with the controlling parameter p in the Stern-Gerlach experiment. The posterior probability represents the updated beliefs about controlling parameters after integrating the observed data. It can then be employed as the new prior for subsequent observations, enabling the continuous refinement of beliefs in the presence of additional data.

Bayesian inference incorporates subjective prior beliefs and provides posterior probability distributions of the controlling parameters, while frequentist inference focuses on objective measures based solely on observed data. The choice between the two approaches often depends on the specific problem and available data.