Bayesian inference fitting

2023

Introduction: what is a probability?

Introduction: what is a probability?

Frequenquist view: frequency of the outcomes for repeated trials

Introduction: what is a probability?

Frequenquist view: frequency of the outcomes for repeated trials

Bayesian view: degree of belief (or how one would bet)

Introduction: what is a probability?

Frequenquist view: frequency of the outcomes for repeated trials

Bayesian view: degree of belief (or how one would bet)

Advantage of the Bayesian view: probability distributions can be assigned to the parameters we wish to fit

Some definitions

Measurement X:
vector consisting of measured values

Model $P(X|\theta, I)$:
In general a probability distribution for the measured values X. It depends on a number of parameters represented by the vector $\theta = (\theta_1, \theta_2, \ldots$
The symbol I represents all other possible a priori knowledge or assumptions about the system. For instance, different models M_1, M_2 can be compared, such that $P(X|\theta, M_1)$ and $P(X|\theta, M_2)$ differ.

Bayes's theorem

From a measurement X and a model $P(X|\theta, I)$ (called the *global likelihood*), we want $P(\theta|X, I)$, the *posterior* probability distribution for the model parameters θ.

Bayes's theorem

From a measurement X and a model $P(X|\theta, I)$ (called the *global likelihood*), we want $P(\theta|X, I)$, the *posterior* probability distribution for the model parameters θ.

Bayes's theorem is an application of conditional probabilities:

$$P(X, \theta|I) = P(X|\theta, I)P(\theta|I) = P(\theta|X, I)P(X|I)$$

Bayes's theorem

From a measurement X and a model $P(X|\theta, I)$ (called the *global likelihood*), we want $P(\theta|X, I)$, the *posterior* probability distribution for the model parameters θ.

Bayes's theorem is an application of conditional probabilities:

$$P(X, \theta|I) = P(X|\theta, I)P(\theta|I) = P(\theta|X, I)P(X|I)$$

$$\Rightarrow P(\theta|X, I) = \frac{P(X|\theta, I)P(\theta|I)}{P(X|I)}$$

Bayes's theorem

From a measurement X and a model $P(X|\theta, I)$ (called the *global likelihood*), we want $P(\theta|X, I)$, the *posterior* probability distribution for the model parameters θ.

Bayes's theorem is an application of conditional probabilities:

$$P(X, \theta|I) = P(X|\theta, I)P(\theta|I) = P(\theta|X, I)P(X|I)$$

$$\Rightarrow P(\theta|X, I) = \frac{P(X|\theta, I)P(\theta|I)}{P(X|I)}$$

normalization constant $P(X|I) = \int P(X|\theta, I)P(\theta|I)\, d\theta$
prior distribution: $P(\theta|I)$

Model comparison with Bayes's theorem

Suppose we have two models M_1 and M_2 that both explain the data X and want to chose which is better. We can compute the ratio of probabilities for the models

$$\frac{P(M_2|X,I)}{P(M_1|X,I)} = \frac{P(X|M_2,I)P(M_2|I)}{P(X|M_1,I)P(M_1|I)} = \frac{P(X|M_2,I)}{P(X|M_1,I)}$$

if we give equal priors to the models, such that $P(M_1|I) = P(M_2|I)$

Model comparison: Occam's razor

Suppose that model M_2 has a free parameter θ while M_1 has none. We have

$$P(X|M_2, I) = \int \underbrace{P(D|\theta, M_2, I)}_{\text{peaked at } \tilde{\theta} \text{ with width } \delta\theta} \quad \underbrace{P(\theta|M_2, I)}_{\text{uniform in interval } 1/\Delta\theta} \quad d\theta$$

$$= P(D|\tilde{\theta}, M_2, I)\frac{\delta\theta}{\Delta\theta}$$

We then have

$$\frac{P(M_2|X, I)}{P(M_1|X, I)} = \frac{P(X|M_2, I)}{P(X|M_1, I)} = \frac{P(X|\tilde{\theta}, M_2, I)}{P(X|M_1, I)}\frac{\delta\theta}{\Delta\theta}$$

The small factor $\frac{\delta\theta}{\Delta\theta} \ll 1$ penalizes the model with the free parameter. This is a natural emergence of Occam's razor that privileges simple models.

Toy example: repeated measurement of X

Measure N times $X \sim \mathcal{N}(\mu, \sigma)$, with known σ but unknown μ

Toy example: repeated measurement of X

Measure N times $X \sim \mathcal{N}(\mu, \sigma)$, with known σ but unknown μ

Obtain the average $\bar{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i \sim \mathcal{N}(\mu, \sigma_N)$
with $\sigma_N = \sigma / \sqrt{N}$

Toy example: repeated measurement of X

Measure N times $X \sim \mathcal{N}(\mu, \sigma)$, with known σ but unknown μ

Obtain the average $\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i \sim \mathcal{N}(\mu, \sigma_N)$
with $\sigma_N = \sigma/\sqrt{N}$

Model: $P(\bar{X}_N|\mu, I) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left(-\frac{(\bar{X}_N - \mu)^2}{2\sigma_N^2}\right)$

Toy example: repeated measurement of X

Measure N times $X \sim \mathcal{N}(\mu, \sigma)$, with known σ but unknown μ

Obtain the average $\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i \sim \mathcal{N}(\mu, \sigma_N)$
with $\sigma_N = \sigma/\sqrt{N}$

Model: $P(\bar{X}_N|\mu, I) = \frac{1}{\sqrt{2\pi}\sigma_N}\exp\left(-\frac{(\bar{X}_N-\mu)^2}{2\sigma_N^2}\right)$

$$P(\mu|\bar{X}_N, I) = \frac{P(\bar{X}_N|\mu, I)P(\mu|I)}{P(\bar{X}_N|I)}$$

Toy example: repeated measurement of X

Measure N times $X \sim \mathcal{N}(\mu, \sigma)$, with known σ but unknown μ

Obtain the average $\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i \sim \mathcal{N}(\mu, \sigma_N)$
with $\sigma_N = \sigma/\sqrt{N}$

Model: $P(\bar{X}_N|\mu, I) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left(-\frac{(\bar{X}_N-\mu)^2}{2\sigma_N^2}\right)$

$$
\begin{aligned}
P(\mu|\bar{X}_N, I) &= \frac{P(\bar{X}_N|\mu, I)P(\mu|I)}{P(\bar{X}_N|I)} \\
&= \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left(-\frac{(\mu - \bar{X}_N)^2}{2\sigma_N^2}\right)
\end{aligned}
$$

Toy example: repeated measurement of X

Measure N times $X \sim \mathcal{N}(\mu, \sigma)$, now both σ and μ are unknown

Toy example: repeated measurement of X

Measure N times $X \sim \mathcal{N}(\mu, \sigma)$, now both σ and μ are unknown

Model: $P(\vec{X}|\mu, \sigma, I) = (\sqrt{2\pi}\sigma^2)^{-N/2} \exp\left(-\frac{\sum_i (X_i - \mu)^2}{2\sigma^2}\right)$

Toy example: repeated measurement of X

Measure N times $X \sim \mathcal{N}(\mu, \sigma)$, now both σ and μ are unknown

Model: $P(\vec{X}|\mu, \sigma, I) = (\sqrt{2\pi}\sigma^2)^{-N/2} \exp\left(-\frac{\sum_i (X_i - \mu)^2}{2\sigma^2}\right)$

$$P(\mu, \sigma|\vec{X}, I) = \frac{P(\vec{X}|\mu, \sigma, I)P(\mu, \sigma|I)}{P(\vec{X}|I)}$$

Toy example: repeated measurement of X

Measure N times $X \sim \mathcal{N}(\mu, \sigma)$, now both σ and μ are unknown

Model: $P(\vec{X}|\mu, \sigma, I) = (\sqrt{2\pi}\sigma^2)^{-N/2} \exp\left(-\frac{\sum_i (X_i - \mu)^2}{2\sigma^2}\right)$

$$P(\mu, \sigma|\vec{X}, I) = \frac{P(\vec{X}|\mu, \sigma, I)P(\mu, \sigma|I)}{P(\vec{X}|I)}$$

To obtain the normalization $P(\vec{X}|I)$, one needs to integrate over both μ and σ ...

Bayesian fitting

Consider the simplest probabilistic model M for the measurement process:

$$y = f(x, \vec{\theta}) + e \quad \text{with } e \sim \mathcal{N}(0, \sigma)$$

Bayesian fitting

Consider the simplest probabilistic model M for the measurement process:

$$y = f(x, \vec{\theta}) + e \quad \text{with } e \sim \mathcal{N}(0, \sigma)$$

Measure N data points $\vec{x}$, $\vec{y}$:

$$P(\vec{x}, \vec{y}|M, \vec{\theta}, \sigma, I) = (\sqrt{2\pi}\sigma^2)^{-N/2} \exp\left(-\frac{\sum_i (y_i - f(x_i, \vec{\theta}))^2}{2\sigma^2}\right)$$

Bayesian fitting

Consider the simplest probabilistic model M for the measurement process:

$$y = f(x, \vec{\theta}) + e \quad \text{with } e \sim \mathcal{N}(0, \sigma)$$

Measure N data points $\vec{x}$, $\vec{y}$:

$$P(\vec{x}, \vec{y} | M, \vec{\theta}, \sigma, I) = (\sqrt{2\pi}\sigma^2)^{-N/2} \exp\left(-\frac{\sum_i (y_i - f(x_i, \vec{\theta}))^2}{2\sigma^2}\right)$$

$$P(\vec{\theta}, \sigma | \vec{x}, \vec{y}, M, I) = \frac{P(\vec{x}, \vec{y} | M, \vec{\theta}, \sigma, I) P(\vec{\theta}, \sigma | I, M)}{P(\vec{x}, \vec{y} | M, I)}$$

Monte Carlo Markov Chains: Metropolis algorithm

Goal: sample a non-normalized probability distribution $P(\vec{\lambda})$ in a high-dimensional space $\vec{\lambda} = (\vec{\theta}, \sigma)$ without any integrals

Monte Carlo Markov Chains: Metropolis algorithm

Goal: sample a non-normalized probability distribution $P(\vec{\lambda})$ in a high-dimensional space $\vec{\lambda} = (\vec{\theta}, \sigma)$ without any integrals

Construct chains $\vec{\lambda}_1, \vec{\lambda}_2, \ldots, \vec{\lambda}_N$
with the following update rule for $\vec{\lambda}_i \to \vec{\lambda}_{i+1}$:

- Randomly pick one component of $\vec{\lambda}_i$
- sample an easy symmetric distribution around the previous value $q(\vec{\lambda}_{\text{new}}|\vec{\lambda}_i)$
- accept the new value $\vec{\lambda}_{\text{new}}$ with probability $\alpha(\vec{\lambda}_{\text{new}}|\vec{\lambda}_i) = \min(1, (q(\vec{\lambda}_i|\vec{\lambda}_{\text{new}})P(\vec{\lambda}_{\text{new}})/(q(\vec{\lambda}_{\text{new}}|\vec{\lambda}_i)P(\lambda_i)))$

Monte Carlo Markov Chains: Metropolis algorithm

Goal: sample a non-normalized probability distribution $P(\vec{\lambda})$ in a high-dimensional space $\vec{\lambda} = (\vec{\theta}, \sigma)$ without any integrals

Construct chains $\vec{\lambda}_1, \vec{\lambda}_2, \ldots, \vec{\lambda}_N$
with the following update rule for $\vec{\lambda}_i \to \vec{\lambda}_{i+1}$:

- Randomly pick one component of $\vec{\lambda}_i$
- sample an easy symmetric distribution around the previous value $q(\vec{\lambda}_{\mathrm{new}}|\vec{\lambda}_i)$
- accept the new value $\vec{\lambda}_{\mathrm{new}}$ with probability $\alpha(\vec{\lambda}_{\mathrm{new}}|\vec{\lambda}_i) = \min(1, (q(\vec{\lambda}_i|\vec{\lambda}_{\mathrm{new}})P(\vec{\lambda}_{\mathrm{new}})/(q(\vec{\lambda}_{\mathrm{new}}|\vec{\lambda}_i)P(\lambda_i)))$

The stationary state of the chain can be proven to sample the distribution $P(\vec{\lambda})$

Demonstration of the Metropolis algorithm

Let's show that $P(\vec{\lambda})$ is the stationary distribution of the chain. First we show detailed balance by computing

Suppose we draw $\vec{\lambda}_i$ from the final distribution $P(\vec{\lambda}_i)$. we can then compute the joint distribution to have $\vec{\lambda}$, then pick $\vec{\lambda}_{i+1}$

$$
\begin{aligned}
P(\vec{\lambda}_i, \vec{\lambda}_{i+1}) &= P(\vec{\lambda}_i) q(\vec{\lambda}_{i+1}|\vec{\lambda}_i) \alpha(\vec{\lambda}_{i+1}|\vec{\lambda}_i) \\
&= P(\vec{\lambda}_i) q(\vec{\lambda}_{i+1}|\vec{\lambda}_i) \min(1, \frac{q(\vec{\lambda}_i|\vec{\lambda}_{i+1} P(\vec{\lambda}_{i+1})}{q(\vec{\lambda}_{i+1}|\vec{\lambda}_i))P(\lambda_i)}) \\
&= \min(P(\vec{\lambda}_i) q(\vec{\lambda}_{i+1}|\vec{\lambda}_i), P(\vec{\lambda}_{i+1}) q(\vec{\lambda}_i|\vec{\lambda}_{i+1})) \\
&= \ldots = P(\vec{\lambda}_{i+1}) q(\vec{\lambda}_i|\vec{\lambda}_{i+1}) \alpha(\vec{\lambda}_i|\vec{\lambda}_{i+1})
\end{aligned}
$$

$\Rightarrow$ detailed balance.

Demonstration of the Metropolis algorithm

Now it's easy to show that

$$\begin{aligned}\int P(\vec{\lambda}_i)q(\vec{\lambda}_{i+1}|\vec{\lambda}_i)\alpha(\vec{\lambda}_{i+1}|\vec{\lambda}_i)d\vec{\lambda}_i &= \int P(\vec{\lambda}_{i+1})q(\vec{\lambda}_i|\vec{\lambda}_{i+1})\alpha(\vec{\lambda}_i|\vec{\lambda}_{i+1})d\vec{\lambda}_i \\ &= P(\vec{\lambda}_{i+1})\underbrace{\int q(\vec{\lambda}_i|\vec{\lambda}_{i+1})\alpha(\vec{\lambda}_i|\vec{\lambda}_{i+1})d\vec{\lambda}_i}_{=1} \\ &= P(\vec{\lambda}_{i+1})\end{aligned}$$

In conclusion, if we sample the desired distribution $P(\vec{\lambda})$, then we always will sample it, i.e. it is the stationary distribution.

Examples of priors: uniform

Uniform prior:
If we know that a parameter lies inside a interval $T_1 \leq T \leq T_2$, then we can set the prior to

$$P(T|I) = \frac{1}{T_2 - T_1}$$

if $T_1 \leq T \leq T_2$ and zero otherwise.

Note that if we “forget” the prior in Bayes’s theorem, we are effectively choosing a uniform prior.

Examples of priors: Jeffreys

In many cases, we might not have a range of values for the parameter T and not even a scale. Then an uninformed prior should be one that gives equal probability for T to lie at different scales, such as the Jeffreys prior:

$$P(T|I) = \frac{1}{\ln(T_{\max}/T_{\min})T}$$

where $0 < T_{\min} \leq T \leq T_{\max}$

This has the property that each decade has the same probability:

$$\int_{0.1}^{1} P(T|I)dT = \int_{1}^{10} P(T|I)dT$$