Statistics

Xinyu Zhong
Queens' College

March 7, 2024

Contents

Abstract

1 Bayesian Statistics

1.1 Likelihood Function

Likelihood is the probability of the data given the parameters.:

$$L(X|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{1}$$

It is a function of θ.

Example: Sky survey path area of sky, count all starts above certain brightness threshold
Data n, representing the number of stars in an area, is discrete and follows a Poisson distribution:

$$f(n|s) = \frac{(As)^n e^{-As}}{n!}$$

Here s is the number density of the stars, and A is the area of the sky.

1.2 Conditional Probability

Conditional probability is the probability of an event given that another event has occurred.:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2}$$

Now replace A with model, $\mathcal{M}$ and B with data, $\mathcal{D}$:

$$P(\mathcal{M}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M})P(\mathcal{M})}{P(\mathcal{D})}$$

Here:

1. $P(\mathcal{M}|\mathcal{D})$ is the posterior probability, which is the probability of the model given the data.
2. $P(\mathcal{D}|\mathcal{M})$ is the likelihood function, which is the probability of the data given the model.
3. $P(\mathcal{M})$ is the prior probability, which is the probability of the model.
4. $P(\mathcal{D})$ is the evidence, which is the probability of the data.
5. The evidence is the normalisation constant, which is the probability of the data averaged over all possible models.

Now rewrite the notation to avoid confusion:

1. Likelihood:
$$P(\mathcal{D}|\mathcal{M}) \equiv \mathcal{L}(\mathcal{D}|\mathcal{M}) \equiv \mathcal{L}(x|\theta)$$
2. Prior:
$$P(\mathcal{M}) \equiv \pi(\mathcal{M}) \equiv \pi(\theta)$$

3. Posterior:
$$P(\mathcal{M}|\mathcal{D}) \equiv \mathcal{P}(\mathcal{M}|\mathcal{D}) \equiv \mathcal{P}(\theta|x)$$

4. Evidence:
$$P(\mathcal{D}) \equiv \mathcal{Z}$$

1.3 Selection of Prior

To start a Bayesian analysis, we need to choose a prior. There are many ways to do so. We can choose prior based on the physical symmetry or easiness of analytical calculation.

1.3.1 Uniform Prior

Uniform prior is a prior that is constant.:

$$\pi(s) = \begin{cases} 1/S_{max} & \text{if } 0 \leq s \leq S_{max} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

This prior is translational invariant.

1.3.2 Jeffreys Prior

Jeffreys prior is a prior that is invariant under reparameterisation(scaling of parameters).:

$$\pi(s) = \frac{1}{\log\left(\frac{S_{max}}{S_{min}}\right)} \begin{cases} 1/S & \text{if } S_{min} \leq s \leq S_{max} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

T prior is scaling invariant, i.e.
$$\pi(s)ds = \pi(\alpha s)d(\alpha s)$$

1.3.3 Conjugate Prior

Conjugate prior is a prior that is chosen to be the same form as the posterior.:

$$\pi(s, k, \theta) = \frac{s^{k-1}e^{-s/\theta}}{\theta^k \Gamma(k)} \tag{5}$$

This is one particular example of conjugate prior, for likelihood that follows the Poisson distribution. In the example of the sky survey, the likelihood is a Poisson distribution:

$$\mathcal{L}(n|s) = \frac{(As)^n e^{-As}}{n!}$$

The conjugate prior of Poisson distribution is the Gamma distribution:

$$\pi(s, k, \theta) = \frac{s^{k-1}e^{-s/\theta}}{\theta^k \Gamma(k)}$$

This gives the posterior:

$$\mathcal{P}(s|n, k', \theta') = \frac{s^{k'-1}e^{-s/\theta'}}{\theta'^{k'} \Gamma(k')}$$

where $k' = k + n$ and $\theta' = \frac{\theta}{A\theta+1}$. This can be done iteratively, i.e. we can use the posterior as the prior for the next iteration. Then the new parameters in the posterior will be

$$k' = k + n_1 + n_2 + ...$$

and

$$\theta' = \frac{\theta}{(A_1 + A_2 + ...)\theta + 1}$$

When the number of data points, N, is large,

$$k_N \to \sum_1^N n$$

and

$$\theta_N \to \frac{1}{\sum_1^N A}$$

the choice of first parameters becomes unimportant. i.e. the posterior will converge to the same value regardless of the choice of the first prior. As $N \to \infty$ The new posterior is a Gamma distribution with a narrower peak that is almost a delta function.

2 Fisher Information

2.1 Score

Score is the derivative of the log likelihood function with respect to the parameter.:

$$S(\theta) = \frac{d}{d\theta} \log \mathcal{L}(x|\theta) \tag{6}$$

A good quality estimator?
The expected value of the score is zero:

$$E_x[S(\theta)] = \langle S(\theta) \rangle = 0$$

Proof:

$$\begin{aligned}
\langle S(\theta) \rangle &= \int S(\theta)\mathcal{L}(x|\theta)dx \\
&= \int \frac{d}{d\theta} \log \mathcal{L}(x|\theta)\mathcal{L}(x|\theta)dx \\
&= \int \frac{1}{\mathcal{L}(x|\theta)} \frac{d}{d\theta}\mathcal{L}(x|\theta)\mathcal{L}(x|\theta)dx \\
&= \int \frac{d}{d\theta}\mathcal{L}(x|\theta)dx \\
&= \frac{d}{d\theta} \int \mathcal{L}(x|\theta)dx \\
&= \frac{d}{d\theta} 1 \\
&= 0
\end{aligned}$$

2.2 Fisher Information

Fisher information is the variance of the score.:

$$I(\theta) = \langle S(\theta)^2 \rangle = \int S(\theta)^2 \mathcal{L}(x|\theta)dx \tag{7}$$

A good quality estimator should have a high Fisher information.

Note that information is not invariant under reparameterisation. For example, if we have a parameter s and we change it to $\theta = \log s$, then the Fisher information will change:

$$I(\phi) = I(\theta)\left(\frac{d\theta}{d\phi}\right)^2$$

Usually, the Fisher information is expressed in a matrix form:

$$I_{ij} = -\int dx[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathcal{L}(x|\theta)]\mathcal{L}(x|\theta)$$

3 Which survey is better?

Now we have two surveys, survey 1 and survey 2. Survey 2 has a larger area, but a lower sensitivity. We want to know which survey is better.

Survey 1

- Area: A_1
- Random variable: $s1$
- Number of stars: $n_1 \sim \text{Poisson}(A_1 s_1)$
- Likelihood: $\mathcal{L}_1(n_1|s_1) = \frac{(A_1 s_1)^{n_1} e^{-A_1 s_1}}{n_1!}$
- Fisher information: $I_1(s_1) = \frac{1}{s_1}$

Survey 2

- Area: A_2
- Random variable: s_2
- Number of stars: $n_1 \sim \text{Poisson}(A_2 s_2)$
- Number of stars that was detected: $n_2 \sim \text{Binomial}(m,\, p)$
- Likelihood:

$$\begin{aligned}\mathcal{L}_2(n_1, n_2|s_2) &= \sum_{m=n_2}^{\infty} P(n_2|m)P(m|s) \\ &= \sum_{m=n_2}^{\infty} \binom{m}{n_2} p^{n_2}(1-p)^{m-n_2}\frac{(A_2 s_2)^m e^{-A_2 s_2}}{m!} \\ &= \dots \\ &= \frac{(A_2 s_2)^{n_2} e^{-A_2 p s_2}}{n_2!}\end{aligned}$$

which is another Poisson distribution with $A_2 p s_2$ as the parameter.

$$\mathcal{L}_2(n_2|s) = \frac{(A_2 p s)^{n_2} e^{-A_2 p s}}{n_2!}$$

- Fisher information: $I_2(s_2) = \frac{A_2 p}{s_2}$

To compare whether survey 1 or survey 2 is better, we need to compare the Fisher information of the two surveys.

4 Post-Posterior

In

5 Sampling methods

The two sampling methods covered are the Metropolis-Hastings algorithm and the Gibbs sampling algorithm. Before the two algorithms are introduced, we discussed the Markov Chain and the detailed balance condition.

5.1 Markov Chain

Markov chain is a stochastic process that satisfies the Markov property.:

$$P(X_{n+1} = x | X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P(X_{n+1} = x | X_n = x_n) \tag{8}$$

The probability of the next state only depends on the current state.

5.1.1 Time homogeneous Markov Chain

Time homogeneous Markov chain is a Markov chain that the transition probability is independent of time.:

$$P(X_{n+1} = x | X_n = x_n) = P(X_{n+1} = x | X_n = x_n, X_{n-1} = x_{n-1}, ..., X_1 = x_1) \tag{9}$$

The probability of the next state only depends on the current state.

5.1.2 Auto-correlation

In the context of Markov chain, the auto-correlation is the correlation of the state at time t with the state at time $t + \tau$.
A very long auto-correlation time means that the Markov chain is not efficient, it brings problem of slow convergence and in-efficient sampling.
A very short auto-correlation is usually preferred. The problem of slow convergence can be explored to solve by reparametrisation of the model such that the shape of the PDF has is variance along a certain direction.

5.1.3 Stationary Distribution

Stationary distribution is a distribution that does not change over time.:

$$\pi(x) = \int \pi(x')\rho(x, x')dx' \tag{10}$$

The probability of being in state x is the same as the probability of being in state x'.

5.1.4 Detailed Balance condition

Detailed balance condition is a condition that the probability of transitioning from state i to state j is the same as the probability of transitioning from state j to state i.:

$$\pi(x)\rho(x', x) = \pi(x')\rho(x, x') \tag{11}$$

The satisfaction of D.B. condition is a sufficient condition for the Markov chain to have a stationary distribution.
To demonstrate that D.B. condition satisfies stationary MC, we integrate both sides of the equation with respect to x':

$$\int \pi(x)\rho(x', x)dx' = \int \pi(x')\rho(x, x')dx'$$
$$\pi(x) = \int \pi(x')\rho(x, x')dx'$$

5.2 Metroplis-Hastings Algorithm

Proposed distribution:

$$Q(x'|x) \tag{12}$$

It is a distribution that is used to propose a new state, usually a Gaussian distribution. It is then rejected or accepted based on the acceptance probability.

Acceptance probability:

$$a(x, y) = \min\left(1, \frac{P(y)Q(x|y)}{P(x)Q(y|x)}\right) \tag{13}$$

It is the probability of accepting the new state. It can be thought as the ratio of the flux of traveling from 1 state to another. The probabilistic rule for accepting/rejecting proposed points, is designed such that the resulting Markov chain satisfies detailed balance with $\pi = p$.

Algorithm 1 Metropolis-Hastings Algorithm

$x_0 \sim \alpha$ ▷ Initialization
$i \leftarrow 0$ ▷ Iteration index
while $i \geq 0$ **do** ▷ Iterate $i = 0, 1, 2, \ldots$
  $y \sim Q(y \mid x_i)$ ▷ Proposal
  $a \leftarrow \frac{P(y)Q(x_i|y)}{P(x_i)Q(y|x_i)}$ ▷ MH acceptance probability
  $u \sim \mathcal{U}(0, 1)$ ▷ Uniform random number
  if $u < a$ **then**
    $x_{i+1} \leftarrow y$ ▷ Markov transition (accept)
  else
    $x_{i+1} \leftarrow x_i$ ▷ Markov transition (reject)
  end if
  $i \leftarrow i + 1$
end while

We only accept the new state with a probability of a. The acceptance probability is the ratio of the target distribution and the proposal distribution.

$$a(x, y) = \min\left(1, \frac{P(y)Q(x|y)}{P(x)Q(y|x)}\right)$$

5.2.1 D.B. condition in Metropolis-Hastings Algorithm

Now, we prove that the Metropolis-Hastings algorithm satisfies the detailed balance condition, and hence has a stationary distribution.
Recall the detailed balance condition:

$$\pi(x)\rho(x', x) = \pi(x')\rho(x, x')$$

LHS of the equation is

$$\pi(x)\rho(x', x) = \pi(x)Q(x'|x)$$

RHS of the equation is

$$\begin{aligned}\pi(x')\rho(x, x') &= \pi(x')Q(x'|x) \\ &= \pi(x')() \\ &= p(x)Q(x'|x)\end{aligned}$$

5.2.2 Enhancement of Metropolis-Hastings Algorithm

Choice of proposal distribution A bad choice of proposal distribution can lead to a very long time of convergence. The MH algorithm generally works best if the proposal closely matches the target distribution. For example in 2D, we can choose a proposal distribution that is a Gaussian distribution with a covariance matrix that is the same as the covariance matrix of the target distribution.
A few choices for the proposal distribution are:

$$\mathbf{\Sigma}_{\text{Small}} = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}, \quad \mathbf{\Sigma}_{\text{Large}} = \begin{pmatrix} 1.5 & 0 \\ 0 & 1.5 \end{pmatrix}, \quad \text{and} \quad \mathbf{\Sigma}_{\text{Corr}} = \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix} \tag{14}$$

1. A smaller covariance matrix will lead to a smaller step size, which will lead to a longer auto-correlation time.
2. A large covariance matrix will lead to a larger step size, which will make the algorithm less efficient, as rejection rate will be higher.
3. A correlated covariance matrix that is close to the true covariance matrix will lead to a more efficient algorithm. However, this requires prior knowledge of the covariance matrix.

5.3 Gibbs Sampling

The algorithm that turns the N-dimensional problem into 1-dimensional problems.

5.3.1 1D Conditional Probability

Given that

$$p(\mathbf{x}) = p(x_1, x_2, ..., x_d)$$

The 1-dimensional conditional probability is

$$p(x_i|x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_d)$$

or

$$p(x^k|x^{-k})$$

To evaluate the normalised conditional probability, we can use the Bayes' theorem:

$$p(x^k|k^{-1}) = \frac{P(\mathbf{x})}{\int P(x^1, ..., x^{k-1}, y^k, x^{k+1}..., x^d)dy^k}$$

5.3.2 Gibbs Sampling Algorithm

Each time, 1 component of the vector is updated. The algorithm is as follows:

1. Start with an initial guess of the vector $\mathbf{x}$
2. Randomly choose a component x^k
3. Update x^k using the conditional probability $p(x^k|x^{-k})$
4. Repeat step 2 and 3 for a large number of times

5.3.3 Example

Given a 2D distribution:

$$p(x, y) \propto ye^{-xy-y}$$

The conditional probability of x given y is

$$p(x|y) \propto e^{-yx}$$

The conditional probability of y given x is

$$p(y|x) \propto ye^{-(x+1)y}$$

5.3.4 DB condition in Gibbs Sampling

The probability defined in Gibbs sampling is:

$$p(y|x) = \rho(y, x) = w_k\delta^{d-1}(x^{-k} - y^{-k})P(y^k|x^{-k})$$

and

$$p(x|y) = \rho(x, y) = w_k\delta^{d-1}(y^{-k} - x^{-k})P(x^k|y^{-k})$$

where w_k is the normalisation constant ?? .
Recall the detailed balance condition:

$$\pi(x)\rho(x', x) = \pi(x')\rho(x, x')$$

LHS of the equation is

$$\begin{aligned}\pi(x)\rho(x', x) &= \pi(x)w_k\delta^{d-1}(x^{-k} - x'^{-k})P(x'^k|x^{-k}) \\ &= p(x^k|x^{-k})p(x^{-k})w_k\delta^{d-1}(x^{-k} - x'^{-k})P(x'^k|x^{-k})\end{aligned}$$

RHS of the equation is

$$\begin{aligned}\pi(x')\rho(x, x') &= \pi(x')w_k\delta^{d-1}(x^{-k} - x'^{-k})P(x^k|x'^{-k}) \\ &= p(x'^k|x^{-k})p(x^{-k})w_k\delta^{d-1}(x^{-k} - x'^{-k})P(x^k|x'^{-k})\end{aligned}$$

Given that the property of the delta function that LHS and RHS are non-zero only when $x^k = x'^k$, we see that the detailed balance condition is satisfied.

5.3.5 Problem of Gibbs Sampling

The problem of Gibbs sampling is that it is slow to converge.

5.3.6 Alternatives to Gibbs Sampling: Gibbs Sweep

The Gibbs sweep is a method that updates all the components of the vector at once. The algorithm is as follows:

1. Start with an initial guess of the vector $\mathbf{x}$
2. Start for $k = 1$ to d
3. Update x^k using the conditional probability $p(x^k|x^{-k})$, where x^{-k} is the vector without the k-th component. Use the updated x^k for the next iteration.
4. Repeat step 2 and 3 for all components of the vector

Problem of this method is that it does not satisfy the detailed balance condition.

5.4 Hamiltonian Monte Carlo

5.4.1 Momentum P

For momentum P, we introduce a new distribution(which can be chosen arbitrarily) that is independent of the position x. The distribution is usually chosen to be a Gaussian distribution:

$$Q(p) \sim \mathcal{N}(0, M)$$

where M is the mass matrix, which is symmetric positive definite.

$$\log Q(p) = -k(p) + const.$$

where $k(p)$ is the kinetic energy.
For a Gaussian distribution, the kinetic energy is

$$k(p) = \frac{1}{2} p^T M^{-1} p$$

We define $R(x, p)$ such that:

$$\log R(x, p) = -H(x, p)$$

where

$$H(x, p) = U(x) + k(p)$$

is the Hamiltonian, which is the sum of the potential energy $U(x)$ and the kinetic energy $k(p)$.

The target distribution is uncovered by the marginalisation of the joint distribution:

$$P(x) = \int R(x, p) dp$$

Note a few conditional probabilities:

$$x, p \sim R(x, p)$$
$$x|p \sim P(x|p)$$
$$p|x \sim Q(p|x)$$

HMC introduces a fictitious parameter t. The position $x(t)$ and the momentum $p(t)$ are functions of time and evolves according to the Hamiltonian equations of motion:

$$\frac{dx(t)}{dt} = \frac{\partial H}{\partial p} = M^{-1}p$$
$$\frac{dp(t)}{dt} = -\frac{\partial H}{\partial x} = -\nabla U(x)$$

Hence, these equations defines a map from states at time t to states at time $t+s$:

$$T_s : (x(t), p(t)) \to (x(t+s), p(t+s))$$

5.4.2 LeapFrog Algorithm

Leapfrog algorithm is a numerical method to evolve the Hamiltonian equations of motion. It is a second order symplectic integrator.

Algorithm 2 Leapfrog Step

procedure LEAPFROG($x, p, \Delta t, M$)
  $p \leftarrow p - \frac{1}{2}\Delta t \nabla|_x E(x)$ ▷ Half step for momentum
  $x \leftarrow x + \Delta t M^{-1} \cdot p$ ▷ Full step for position
  $p \leftarrow p - \frac{1}{2}\Delta t \nabla|_x E(x)$ ▷ Half step for momentum
  return x, p
end procedure

5.4.3 Time-Reversibility,Volume-Preserving and Energy Conservation

The choice of the leapfrog algorithm is to ensure that it is time reversible and volume preserving.

Energy conservation is not guaranteed, but the error is bounded with order of Δt^2 .

$$\mathcal{H}(x(t+s), p(t+s)) - \mathcal{H}(x(t), p(t)) = \mathcal{O}(\Delta t^2)$$

5.4.4 Property of Hamiltonian Monte Carlo

1. Reversible
2. Volume preserving
3. Conserves energy

5.4.5 HMC Algorithm

5.4.6 Relation to blocked Gibbs sampling and MH algorithm

5.5 Slice Sampling

6 Bayesian Model Comparison

Recall that posterior:

$$P(\theta|\mathcal{D}, \mathcal{M}) = \frac{\mathcal{L}(\mathcal{D}|\theta, \mathcal{M})\pi(\theta|\mathcal{M})}{\mathcal{Z}}$$

where $\mathcal{Z}$ is the evidence, which is the probability of the data averaged over all possible models.

$$\mathcal{Z} = \int \mathcal{L}(\mathcal{D}|\theta, \mathcal{M})\pi(\theta|\mathcal{M})d\theta$$

6.1 Frequentist Approach

Given two models for the same data, $\mathcal{M}_1(\lambda_A)$ and $\mathcal{M}_2(\lambda_B)$, we want to know which model is better. The frequentist approach is to use the maximum likelihood ratio test:

$$MLR(Maximum\ Likelihood\ Ratio) = \frac{\mathcal{L}(\mathcal{D}|\lambda_A, \mathcal{M}_1)}{\mathcal{L}(\mathcal{D}|\lambda_B, \mathcal{M}_2)}$$

6.2 Baysian Approach

We use the posterior odds ratio for the same problem:
Posterior Odds Ratio:

$$O_{A,B} = \frac{P(\mathcal{M}_1|\mathcal{D})}{P(\mathcal{M}_2|\mathcal{D})} = \frac{\mathcal{Z}_1}{\mathcal{Z}_2}\frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} \tag{15}$$

We are updating the prior odds ratio with the Bayes factor into the posterior odds ratio.

6.3 Example

6.4 Occam Penalty

The ideal of Occam's razor is that the simplest model that fits the data is the best.
Occam Penalty:

Models that are more flexible (more complicated or more parameters that can take values spanning a wider range), have lower evidence.

Baysian Inference incorporates the Occam penalty, which is the penalty for having more parameters.

6.5 Computation of Z

The computation of the evidence is important in the Baysian model comparison.Nonetheless, it is often difficult to compute.

6.5.1 Analytical Solution

An analytical solution is possible for some simple models, such as the Gaussian distribution. Given that in Baysian:

$$P(\mathcal{M}|\mathcal{D}) = \frac{\mathcal{L}(\mathcal{D}|\mathcal{M})\pi(\mathcal{M})}{\mathcal{Z}}$$

If we know the analytical form of the likelihood and the prior, and hence the posterior, we can compute the evidence analytically.

6.5.2 Laplace Approximation

However, it is usually difficult to compute the evidence analytically. The Laplace approximation is a method to approximate the evidence.
Given that evidence, $\mathcal{Z}$:

$$\begin{aligned}\mathcal{Z} &= \int \mathcal{L}(\mathcal{D}|\theta, \mathcal{M})\pi(\theta|\mathcal{M})d\theta \\ &= \int P^*(\theta|\mathcal{D}, \mathcal{M})d\theta\end{aligned}$$

where $P^*(\mathcal{D}|\theta)$ is the unnormalised posterior.
Now we can taylor expand the log likelihood around the maximum posterior:

$$\log P^*(\theta|\mathcal{D},\mathcal{M}) \approx \log P^*(\theta_{max}|\mathcal{D},\mathcal{M}) - \frac{1}{2}c(\theta-\theta_{max})^2$$

where c is the second derivative of the log likelihood at the maximum posterior.
i.e.

$$c = -\frac{d^2}{d\theta^2}\log P^*(\theta|\mathcal{D},\mathcal{M})|_{\theta_{max}}$$

Taking exponential of both sides, we have:

$$P^*(\mathcal{D}|\theta,\mathcal{M}) \approx P^*(\mathcal{D}|\theta_{max},\mathcal{M})\exp\left(-\frac{1}{2}c(\theta-\theta_{max})^2\right)$$

Therefore, such this to the integral, we have:

$$\begin{aligned}\mathcal{Z} &\approx P^*(\mathcal{D}|\theta_{max},\mathcal{M})\int \exp\left(-\frac{1}{2}c(\theta-\theta_{max})^2\right)d\theta \\ &\approx P^*(\mathcal{D}|\theta_{max},\mathcal{M})\sqrt{\frac{2\pi}{c}}\end{aligned}$$

Hence, we change the nature of the problem from integrate to find the maximum likelihood and the evaluation of the second derivative of the log likelihood at maximum likelihood.

Note that by changing the parameterisation of the model, the evidence will change.i.e. the evidence is not invariant under reparameterisation.

$$x \to \theta = f(x)$$

In the case of multidimensional parameter space $\theta \in \mathbb{R}^D$

$$\log P^*(\mathcal{D}|\theta,\mathcal{M}) \approx \log P^*(\mathcal{D}|\theta_{max},\mathcal{M}) - \frac{1}{2}(\theta-\theta_{max})^T\nabla^2\log\mathcal{L}(\mathcal{D}|\theta,\mathcal{M})|_{\theta_{max}}(\theta-\theta_{max}) + ...$$

where $\nabla^2\log\mathcal{L}(\mathcal{D}|\theta,\mathcal{M})|_{\theta_{max}}$ is the Hessian matrix.
and

$$\mathcal{Z} \approx P^*(\mathcal{D}|\theta_{max},\mathcal{M})\sqrt{\frac{(2\pi)^D}{\det|\nabla^2\log\mathcal{L}(\mathcal{D}|\theta,\mathcal{M})|_{\theta_{max}}|}}$$

If the posterior has more than one peak?

6.5.3 Nested Sampling

Nested sampling is a method to compute the evidence. It is a method that is based on the idea of the prior volume.
The prior volume is the volume of the prior space that is enclosed by the likelihood contour.

???

6.5.4 Thermodynamic Integration

Thermodynamic integration is another method to compute the evidence.
Annealed likelihood:

$$\mathcal{L}(\mathcal{D}|\theta,\mathcal{M})^\beta = \mathcal{L}(\mathcal{D}|\theta,\mathcal{M})^\beta \tag{16}$$

where β is the inverse temperature, it gets the inspiration from the Boltzmann distribution, i.e. $\beta = 1/kT$
Annealed posterior:

$$\pi(\theta|\mathcal{D},\mathcal{M})^\beta = \frac{\mathcal{L}(\mathcal{D}|\theta,\mathcal{M})^\beta \pi(\theta|\mathcal{M})}{\mathcal{Z}_\beta} \tag{17}$$

where $\mathcal{Z}^\beta$ is the evidence at the inverse temperature β, i.e. $\mathcal{Z}_\beta = \int \mathcal{L}(\mathcal{D}|\theta,\mathcal{M})^\beta \pi(\theta|\mathcal{M}) d\theta$
Note that at

1. $\beta = 0$, the annealed likelihood is flat: $\mathcal{Z}_\beta = 1$
2. $\beta = 1$, the annealed likelihood is the likelihood:

Note this algebraic identity:

$$\begin{aligned}
\frac{d}{d\beta}\log \mathcal{Z}_\beta &= \frac{1}{\mathcal{Z}_\beta}\frac{d}{d\beta}\mathcal{Z}_\beta \\
&= \frac{1}{\mathcal{Z}_\beta}\frac{d}{d\beta}\int \mathcal{L}(\mathcal{D}|\theta,\mathcal{M})^\beta \pi(\theta|\mathcal{M}) d\theta \\
&= \frac{1}{\mathcal{Z}_\beta}\int \frac{d}{d\beta}\mathcal{L}(\mathcal{D}|\theta,\mathcal{M})^\beta \pi(\theta|\mathcal{M}) d\theta \\
&= \frac{1}{\mathcal{Z}_\beta}\int \log \mathcal{L}(\mathcal{D}|\theta,\mathcal{M})\mathcal{L}(\mathcal{D}|\theta,\mathcal{M})^\beta \pi(\theta|\mathcal{M}) d\theta \\
&= \int \log \mathcal{L}(\mathcal{D}|\theta,\mathcal{M}) P(\theta|\mathcal{D},\mathcal{M})^\beta d\theta \\
&= \langle \log \mathcal{L}(\mathcal{D}|\theta,\mathcal{M})\rangle_\beta
\end{aligned}$$

Numerically, the expected value can be approximated by the average over a large number of samples.

$$\log \mathcal{Z}_1 - \log \mathcal{Z}_0 = \int_0^1 \langle \log \mathcal{L}(\mathcal{D}|\theta,\mathcal{M})\rangle_\beta d\beta$$

where the integral can be approximated by the trapezoidal rule.

6.6 Savage-Dickey Density Ratio

Given that we are only interested in the evidence ratio, we can use the Savage-Dickey density ratio to compute the evidence ratio.

$$B_{1,2} = \frac{P(\theta|\mathcal{D},\mathcal{M}_1)}{P(\theta|\mathcal{D},\mathcal{M}_2)} = \frac{\mathcal{Z}_1}{\mathcal{Z}_2}$$

Given two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, where $\mathcal{M}_1$ is a sub-model of $\mathcal{M}_2$, i.e. $\mathcal{M}_1$ is a nested in $\mathcal{M}_2$.
Suppose that $\mathcal{M}_1$ has a parameter ϵ and $\mathcal{M}_2$ has a parameter ϵ and additional parameters ϕ, where $\epsilon \in \mathbb{R}$ and $\phi \in \mathbb{R}^D$.
We note:

$$\begin{aligned}
\mathcal{L}(\mathcal{D}|\phi,\mathcal{M}_1) &= \mathcal{L}(\mathcal{D}|\phi,\mathcal{M}_2,\epsilon=0) \\
\pi(\epsilon|\mathcal{M}_1) &= \pi(\epsilon|\mathcal{M}_2,\epsilon=0)
\end{aligned}$$

Then the evidence ratio is:

$$\begin{aligned}
\mathcal{Z}_1 &= P(\mathcal{D}|\mathcal{M}_1) \\
&= \int \mathcal{L}(\mathcal{D}|\phi, \mathcal{M}_1)\pi(\phi|\mathcal{M}_1)d\phi \\
&= \int \mathcal{L}(\mathcal{D}|\phi, \mathcal{M}_2, \epsilon = 0)\pi(\phi|\mathcal{M}_2, \epsilon = 0)d\phi \\
&= P(\mathcal{D}|\mathcal{M}_2, \epsilon = 0) \\
&= \frac{P(\epsilon = 0|\mathcal{D}, \mathcal{M}_2)}{P(\epsilon = 0|\mathcal{M}_2)} P(\mathcal{D}|\mathcal{M}_2) \\
&= \frac{P(\epsilon = 0|\mathcal{D}, \mathcal{M}_2)}{P(\epsilon = 0|\mathcal{M}_2)} \mathcal{Z}_{\in}
\end{aligned}$$

Hence, the evidence ratio is:

$$B_{1,2} = \frac{P(\epsilon = 0|\mathcal{D}, \mathcal{M}_2)}{P(\epsilon = 0|\mathcal{M}_2)}$$

which is just the ratio of the posterior density at $\epsilon = 0$ under $\mathcal{M}_2$ to the prior density at $\epsilon = 0$ under $\mathcal{M}_2$.
It is independent of $\mathcal{M}_2$