Bayesian inference

When facing an inference problem that we wish to tackle with a Bayesian method, we start with the following steps.

1. Collect the data $D = \{d_1, d_2, \dots d_N\}$, where each d_i can be a vector or even a matrix of observations, but each d_i is independent from every other d_j (given fixed model parameters).
2. Come up with a generative model $P(d \,|\, \theta)$, where $\theta = (\theta_1, \theta_2, \dots \theta_K)$ stands for the collection of all unknown model parameters.
3. Use our best judgment to come up with a prior distribution for θ, $P_0(\theta)$.
4. Write down the posterior
$$P(\theta \,|\, D) = \frac{P_0(\theta) \prod_i P(d_i \,|\, \theta)}{Z} = \frac{\tilde{P}(\theta \,|\, D)}{Z},$$
where $\tilde{P}(\theta \,|\, D)$ is the unnormalized posterior of θ and Z is the necessary normalization factor (independent of θ).
5. Calculate or approximate relevant characteristics of the posterior. These often include
   - median of each θ_k, mean and mode of θ,
   - percentiles and standard deviation of each θ_k, and covariance matrix of θ,
   - histogram of each θ_k.

This 5th step is hard. Exact analytical solutions exist for a handful of widely-used models, most of which assume (multivariate) Gaussian distributions, and exact numerical solutions are feasible if model parameters are few, $K \lesssim 5$. Most non-Gaussian, real-life inference problems, however, require approximations or sampling to carry out the 5th step.

Exact solutions

- **Analytical.** In rare cases (typically Gaussian models), Z can be expressed as an analytical formula of the data D, and the posterior $P(\theta \,|\, D)$ is a known distribution, the formulas for whose mean and variance can be looked up.
- **Numerical.** If no analytical solution exists, one can numerically evaluate $w_\theta = \tilde{P}(\theta \,|\, D)$ on a *regular* grid of θ values. Posterior distribution, expectation value and variance can be calculated as averages:
$$P(\theta) \approx \frac{w_\theta}{\Delta\theta \sum_\theta w_\theta}, \qquad \mathbb{E}(\theta) \approx \frac{\sum_\theta w_\theta \theta}{\sum_\theta w_\theta}, \qquad \mathrm{Var}(\theta) \approx \frac{\sum_\theta w_\theta \theta^2}{\sum_\theta w_\theta} - \left[\mathbb{E}(\theta)\right]^2,$$
where $\Delta\theta$ is the volume of a grid cell on the regular θ grid. Median, mode and covariance can also be directly evaluated from the posterior, $P(\theta)$. This is feasible only if the dimension of θ, i.e. K, is small. Making the grid finer makes this approximation more accurate.

MCMC sampling

As the dimension of θ increases, evaluating $\tilde{P}$ on a regular grid becomes unrealistic. This is called the curse of dimensionality. Monte-Carlo Markov Chain (MCMC) sampling provides an alternative strategy: Instead of scanning all values of θ and weighting them with their (unnormalized) posterior value w_θ, MCMC methods obtain samples from the posterior.

The most notable MCMC sampling methods are **Metropolis-Hastings** sampling, **Hamiltonian** sampling and **Gibbs** sampling. Common to all sampling methods, they produce a series of $\theta^{(t)}$ samples for $t = 0, 1, \dots T$, drawn approximately from the posterior $P(\theta \,|\, D)$. Methods differ in how they draw $\theta^{(t+1)}$ using $\theta^{(t)}$ and the data D. All MCMC methods have a warm-up time t_0, during which samples are similar to

the starting value $\theta^{(0)}$, and a thinning period τ, which sets the minimal necessary distance between samples so they are not correlated significantly. From the resulting samples, expectation value and variance can be approximated as

$$\mathbb{E}(\theta) \approx \frac{1}{L}\sum_{l=1}^{L}\theta^{(t_0+l\tau)}, \qquad \mathrm{Var}(\theta) \approx \frac{1}{L}\sum_{l=1}^{L}\left[\theta^{(t_0+l\tau)}\right]^2 - \left[\mathbb{E}(\theta)\right]^2,$$

where $L = (T - t_0)/\tau$ is the total number of (approximately) uncorrelated samples. Median, mode and covariance can also be directly approximated by their empirical values from the samples $\{\theta^{(t_0+l\tau)} : l = 1, 2 \dots L\}$.

To find acceptable values for warm-up time t_0, we can plot traces of $\theta^{(t)}$ as a function of t. We should choose t_0 such that all initial transient motion is gone for $t > t_0$. The thinning period can be choosen by plotting thinned traces $\theta^{(t_0+l\tau)}$ as a function of l for different τ values, and choosing the smallest τ for which the line-plot of the trace looks like a "hairy caterpillar".

Disclaimer: MCMC sampling methods are prone to the problem of isolated local maxima in the posterior: If two or more regions of the posterior are equally likely but separated by deep valleys, no sampler will be able to make jumps between them. To mitigate this, running multiple chains, started from different $\theta^{(0)}$ values can be run. We can monitor the log-posterior $\log \tilde{P}(\theta^{(t)} \,|\, D)$ for each chain, which, if it is significantly lower than the other chains, indicate that the chain in question is stuck in a sub-optimal local maximum. Then we can discard it, and combine the samples only from chains that have similarly high log-posterior.

MAP + Laplace approximation

MCMC sampling methods require drawing a new sample $\theta^{(t+1)}$ in every iteration, for which the algorithm needs to have efficient access to all data D. This becomes a logistic difficulty as the size of the data N increases, resulting in slowdown. Sampling methods are practical only for inference problems, where the data (not the raw data, but the features relevant for the inference problem) fit comfortably in memory. When both exact methods and sampling methods prove to be infeasible, maximum a posteriori (MAP) estimation combined with Laplace approximation may provide acceptable solutions.

Strictly speaking MAP finds only the mode of the posterior distribution,

$$\theta_{\mathrm{MAP}} = \mathrm{mode}(\theta) = \arg\max_{\theta}\left[\tilde{P}(\theta \,|\, D)\right] = \arg\max_{\theta}\left[\log \tilde{P}(\theta \,|\, D)\right].$$

This can be calculated by one of the following methods.

- **Analytically**, if the unique solution to the algebraic equations,

  $$0 = \partial_{\theta_k} \log \tilde{P}(\theta \,|\, D) \qquad k = 1, 2 \dots K,$$

  can be found in closed form.

- **Gradient descent** starts from $\theta^{(0)}$, and continually moves θ towards the direction in which $-\log \tilde{P}(\theta \,|\, D)$ is decreasing most steeply. This is implemented by python's `scipy.optimize.minimize()` function.

- **Expectation-Maximization methods** separate the parameters in θ into two groups (usually: latent variables and model parameters), in such a way that allows one-step optimization of one group while the other is fixed. Alternating which group to fix and which to optimize, eventually, results in converging to a local maximum of the log posterior.

Laplace approximation can always be carried out, after the MAP estimation is found. It approximates the posterior with a Gaussian distribution centered at the MAP estimate. The covariance of this Gaussian is

$$\mathrm{Cov}_{\mathrm{Laplace}}(\theta) = \Big[-\nabla_\theta \nabla_\theta \log \tilde{P}(\theta \,|\, D)\Big]^{-1},$$

i.e. the inverse of the negative second-order derivative (aka Hessian matrix) of the log posterior. For 1-dimensional θ, it gives a single number, the "Laplace variance" of θ.

Example

The hand of an assembly-line robot needed to be replaced 31 days, 82 days, 122 days and 186 days after its initial purchase. Question: What is the estimated lifetime of the robot's hand?

Data and Model

The data consists of $N = 4$ waiting times between consecutive installations:

$$D = \{t_1, t_2, t_3, t_4\} = \{31 - 0, 82 - 31, 122 - 82, 186 - 122\} = \{31, 51, 40, 64\}.$$

Assuming that the workload was constant, we model t_i with an exponential distribution with lifetime of τ days.

$$P(t \,|\, \tau) = \frac{1}{\tau} \exp\left(-\frac{t}{\tau}\right).$$

We use an uninformative prior for the lifetime:

$$P_0(\tau) = \frac{\text{const.}}{\tau},$$

where const. stands for an appropriate constant that normalizes the prior in a plausible region.

The posterior can be written as

$$P(\tau \,|\, D) = \frac{P_0(\tau) \prod_i P(t_i \,|\, \tau)}{Z} = \frac{1}{Z} \frac{1}{\tau} \prod_i \left[\frac{1}{\tau} \exp\left(-\frac{t_i}{\tau}\right)\right] = \frac{1}{Z} \frac{1}{\tau^{N+1}} \exp\left(-\frac{1}{\tau} \sum_i t_i\right).$$

MAP + Laplace approximation

The unnormalized log-posterior is

$$\log \tilde{P} = \log \left[P_0(\tau) \prod_i P(t_i \,|\, \tau)\right] = -(N+1) \log \tau - \frac{\sum_i t_i}{\tau} + \text{const.}$$

Requiring that the first derivative is zero yields an algebraic equation for τ, which we can solve for the MAP estimate, τ_{MAP}.

$$0 = \partial_\tau \log \tilde{P} = -\frac{N+1}{\tau} + \frac{\sum_i t_i}{\tau^2} \qquad \Rightarrow \qquad \tau_{\text{MAP}} = \text{mode}(\tau) = \frac{\sum_i t_i}{N+1} = 37.20$$

The second-order derivative allows us to calculate the "Laplace standard deviation" $\text{std}_{\text{Laplace}}(\tau)$.

$$\begin{aligned} (\partial_\tau)^2 \log \tilde{P} &= \frac{N+1}{\tau^2} - \frac{2\sum_i t_i}{\tau^3} = -\frac{\sum_i t_i}{(\tau_{\text{MAP}})^3} \\ \text{Var}_{\text{Laplace}} &= \left[-(\partial_\tau)^2 \log \tilde{P}\right]^{-1} = \frac{(\tau_{\text{MAP}})^3}{\sum_i t_i} \qquad \Rightarrow \qquad \text{std}_{\text{Laplace}}(\tau) = \sqrt{\frac{(\tau_{\text{MAP}})^3}{\sum_i t_i}} = \frac{\sum_i t_i}{\sqrt{(N+1)^3}} = 16.64 \end{aligned}$$

MCMC sampling

Let's use STAN (http://mc-stan.org/) to run Hamiltonian sampling. The python API (pystan) is great. Let's run 4 chains with the NUTS algorithm, each with a total of 2000 iterations, the first 1000 of which we drop, but keep every one of the remaining samples. Here is the full code.

```
1  #!/usr/bin/python
2  import numpy
3  import pystan
4
5  # Define model
6  model_code = """
7  data {
8      int<lower=1> N;        // number of waiting times
9      real<lower=0> t[N];    // waiting times [in days]
10 }
11 parameters {
12     real<lower=0> tau;     // lifetime [in days]
13 }
14 model {
15     target += log(1/tau);  // log-prior
16     for (n in 1:N){
17         target += log(1/tau) - t[n] / tau;  // log-likelihood
18     }
19 }
20 """
21 model = pystan.StanModel(model_code=model_code)
22
23 # Get data and run sampling
24 data = {'N': 4,
25         't': [31, 51, 40, 64]}
26 fit = model.sampling(data=data,
27                      algorithm='NUTS',
28                      chains=4,
29                      iter=2000,
30                      warmup=1000,
31                      thin=1,
32                      seed=123456)
33
34 # Post-process the results
35 tau_samples = fit.extract()['tau']
36 print('mean(tau): ' + str(numpy.mean(tau_samples)))
37 print('std(tau):  ' + str(numpy.std(tau_samples)))
38 fit.plot().savefig('./tau.pdf', bbox_inches='tight')
```

Line 21 compiles the STAN model, defined by Lines 6-20, then it runs the sampling (Lines 26-32) and displays the results: the estimated mean and standard deviation of τ (Lines 36-37),

```
1 mean(tau): 62.845930336113604
2 std(tau):  41.65566288308886
```

and an estimate of the posterior and traces of τ in the figure tau.pdf (Line 38).

Exact solution

The normalization constant Z can be calculated by integrating the unnormalized posterior over the entire domain of τ (here we apply the limits $\tau_{\min} \to 0$ and $\tau_{\max} \to \infty$ to obtain closed forms):

$$Z = \int_0^\infty d\tau \; \frac{1}{\tau^{N+1}} \exp\left(-\frac{1}{\tau}\sum_i t_i\right) = \frac{1}{[\sum_i t_i]^N} \int_0^\infty dz \; z^{N-1} \exp(-z) = \frac{\Gamma(N)}{[\sum_i t_i]^N},$$

where we used a new variable $z = \frac{1}{\tau}\sum_i t_i$, and expressed $\tau = \frac{1}{z}\sum_i t_i$ and $d\tau = -\frac{dz}{z^2}\sum_i t_i$. The analytical form of the normalized posterior is

$$P(\tau \,|\, D) = \frac{[\sum_i t_i]^N}{\Gamma(N)} \tau^{-N-1} \exp\left(-\frac{\sum_i t_i}{\tau}\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{-\alpha-1} \exp\left(-\frac{\beta}{\tau}\right) = \mathrm{IG}(\tau \,|\, \alpha, \beta),$$

which is the inverse-gamma distribution for $\alpha = N$ and $\beta = \sum_i t_i$, the mean, mode and standard deviation are

$$\begin{aligned}
\mathbb{E}(\tau) &= \frac{\beta}{\alpha - 1} = \frac{\sum_i t_i}{N - 1} = 62.00 \\
\mathrm{mode}(\tau) &= \frac{\beta}{\alpha + 1} = \frac{\sum_i t_i}{N + 1} = 37.20 \\
\mathrm{std}(\tau) &= \frac{\beta}{(\alpha - 1)\sqrt{\alpha - 2}} = \frac{\sum_i t_i}{(N - 1)\sqrt{N - 2}} = 43.84
\end{aligned}$$

Comparison

We can plot the posterior densities, produced by the three methods. While the histogram of the Monte Carlo Markov Chain samples (Line 35 in code) reproduce the exact posterior quite well. The MAP + Laplace approximation is accurate only in finding the mode and estimating the order of magnitude of std, but it fails to capture the skewness of the posterior, and underestimates τ.

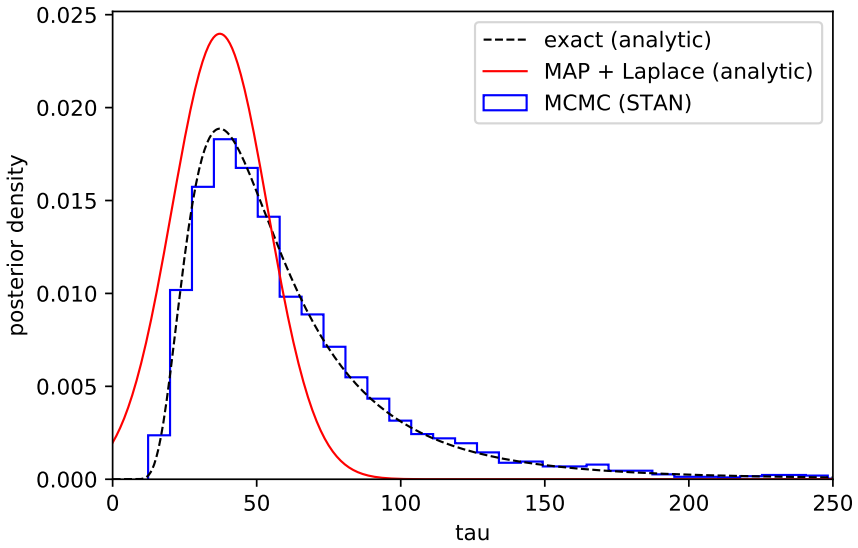


Figure 1: Posterior of τ as estimated by MAP + Laplace approximation and MCMC sampling, as well as the exact curve.

Problems

All of the following problems can be solved with MCMC sampling. Apart from one, they are also solvable with MAP+Laplace approximation, either analytically or numerically. Several of them also have exact analytical solutions. In each problem, we are interested in a "best guess" (e.g. mean or mode) and some "error bars" (e.g. std) for the parameters in question.

Problem 1 - Linear regression

Given (x, y) data ($x = [1, 2, 3, 4, 5, 6, 7], y = [3.23, 5.38, 6.54, 7.99, 9.86, 11.35, 11.98]$), assume a linear model $y = ax + c + \varepsilon$, where ε is a Gaussian variable with zero mean and variance σ^2. Estimate the parameters (a, c, σ^2).

Problem 2 - Negative binomial

An experiment with binary outcome is performed until it fails $r = 5$ times; the recorded number of successes is $k = 21$ (see negative binomial distribution). Estimate the success probability p.

Problem 3 - Normal

A person took seven different IQ tests and got the following scores, $x = [112, 111, 115, 131, 124, 114, 96]$. Assume normal distribution of the test results $x \sim \mathcal{N}(\mu, \sigma^2)$, and estimate the parameters (μ, σ^2). What is the probability that μ, is below 100?

Problem 4 - Fixed effect model

In a greenhouse, tomato plants are separated into three (unequal) groups, each of which is treated with a different fertilizer. Yield of each plant, y (in grams) is measured to be

$$\begin{aligned} &\text{group A:} && y_A = [2604, 2665, 2251, 2510] \\ &\text{group B:} && y_B = [1559, 1729, 1866, 1413, 1159, 1692] \\ &\text{group C:} && y_C = [1528, 1444, 1041] \end{aligned}$$

Let's model yield as $y_{g,i} = \mu_g + \varepsilon_{g,i}$, where $g \in \{A, B, C\}$ is the group index, $i = 1, 2, \ldots$ is the plant index, μ_g are fixed and unknown, and $\varepsilon_{g,i}$ are iid. Gaussian variables with zero mean and σ^2 variance. Estimate the parameters $(\mu_A, \mu_B, \mu_C, \sigma^2)$.

Problem 5 - Random effect model

Model the data in pProblem 4 as $y_{g,i} = \mu + \delta_g + \varepsilon_{g,i}$, where μ is fixed and unknown, δ_g are iid. Gaussian variables with zero mean and variance σ_δ^2, and $\varepsilon_{g,i} \sim \mathcal{N}(0, \sigma^2)$ iid. as before. Estimate the parameters $(\mu, \sigma_\delta^2, \sigma^2)$. What percentage of the observed variance is explained by the difference in fertilizers, i.e. what is $\sigma_\delta^2/(\sigma_\delta^2 + \sigma^2)$?

Problem 6 - $\exp(-x^4)$

In a park, the heights of 10 pine trees are measured $x = [435, 491, 620, 356, 443, 398, 383, 475, 554, 639]$ in centimeters. Assume the following distribution for the heights,

$$P(x \,|\, c, s) = \frac{1}{Z}\frac{1}{s}\exp\left(-\frac{(x-c)^4}{s^4}\right), \qquad \text{where } Z = \frac{\Gamma(\frac{1}{4})}{2} = 1.8128,$$

and estimate the parameters (c, s). [Extra: Fit a normal distribution too, and calculate which model has higher likelihood.]

Problem 7 - Uniform

A pseudo-random number generator produces the following sequence $x = [8.155, 6.272, 3.583, 7.096, 6.469]$. Assume that it draws samples uniformly from an interval $l \leq x \leq h$, and estimate the parameters (l, h).

Problem 8 - Mixed error profile

Twelve screws produced by a machine are sent to quality control, and measured how much their length deviates from the specification. The measured deviations are $x =$ [-1.32, -8.17, -0.21, -0.73, +5.78, +1.75, -0.29, -0.34, -0.52, +1.29, -0.60, +0.28] in percentages. We suspect that some unknown fraction (f) of the screws are affected by a much larger errors than the rest, and we model this with the following mixture distribution,

$$P(x \mid f, \sigma, s) = (1-f) \times \text{Normal}(x \mid 0, \sigma) + f \times \text{Cauchy}(x \mid 0, s) = (1-f) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) + f \frac{1}{s\pi} \frac{1}{1+(x/s)^2}.$$

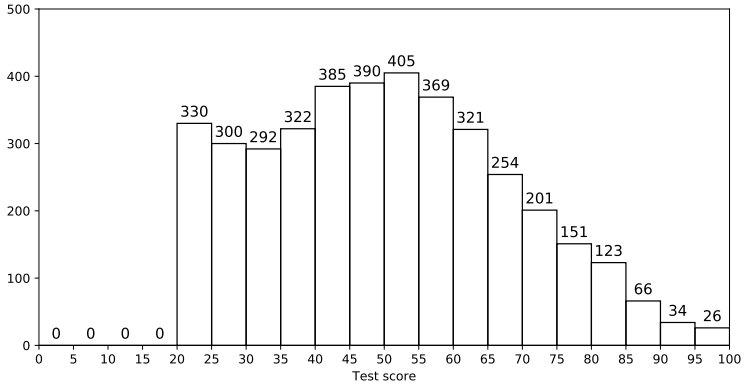
Estimate the parameters (f, σ, s). [Extra: For each measured deviation, estimate the probability that it is a result of the Cauchy-distributed error.]

Problem 9 - Multinomial

In an egg-processing unit, the quality control mechanism takes 20 eggs from the conveyor belt, and grades their quality, i.e. giving them labels A, B, C, D or F. The observed counts are $(n_A, n_B, n_C, n_D, n_F) = (7, 8, 3, 1, 1)$. Assuming multinomial distribution, estimate the fractions of each quality grades in this batch, i.e. the parameters $(p_A, p_B, p_C, p_D, p_F)$.

Problem 10 - Threshold effect

In a school district, 3955 seniors took a standardized math test. The histogram of the scores are published, as shown below. A striking feature of this data is that no students received scores below 20. This is likely an indication that the minimal passing score was 20, and grading was done in a way that ensured that no student failed. This hypothesis is also supported by the visible abundance just above the score 20.



Let's estimate the distribution of the "unadjusted" scores by assuming the following process.

- True scores are distributed according to a normal distribution with mean μ and standard deviation σ, truncated to the $[0, 100]$ interval.
- If the true score is below 20, it gets adjusted. Let's model this by a replacement of the true score a score drawn from an exponential distribution with scale parameter b, truncated to the [20, 100] interval.

Estimate the parameters (μ, σ, b), and the proportion of the students that would have failed if their scores had not been adjusted.