Bayesian Statistics

Wenjie Tu

Spring Semester 2022

- Base rate paradox
  - $P(B \mid A_1) \gg P(B \mid A_2) \not\Longrightarrow P(A_1 \mid B) > P(A_2 \mid B)$
  - Often $P(B \mid A_1)$ large and $P(A_1)$ very small
- Challenges of Bayesian inference
  - Finding a good model (both prior and likelihood)
  - Calculating the posterior
  - Assessing the fit of the model
- Comparison between Bayesian and frequentist
  - Assumption (parameters)
  - Estimation (MLE, MCMC)
  - Testing (NHST, Bayes factors)
  - Bayesian learning (few observations, prior belief)
- Bayesian point estimates
  - A Bayesian point estimate summarizes the posterior distribution in a number. The following estimates for the location are often used:
    - Posterior mean
    - Posterior median
    - Posterior mode
- Bayesian decision theory
  - Bayesian decision theory provides a unified approach for Bayesian point estimates
  - The posterior risk is the expected loss under the posterior:

$$\rho(T(x), \pi) = \mathbb{E}(L(T(X), \theta) \mid x) = \int_{\Theta} L(T(X), \theta)\pi(\theta \mid x)d\theta$$

    - It is obtained by integrating the loss function over the posterior of the parameter θ
    - It depends on the data x but not on the parameter θ
- Frequentist decision theory
  - The frequentist risk:

$$R(T, \theta) = \mathbb{E}_{\theta}(L(T(X), \theta)) = \int_{X} L(T(x), \theta)f(x \mid \theta)dx$$

    - It is obtained by integrating the loss function over the data x
    - It depends on the parameter θ but on the data
  - How to minimize the frequentist risk?

    - Minimax
    - Minimize weighted risk
    - Admissibility
- Testing: Frequentist vs. Bayesian statistics
- Decisions based on Bayes factors
- p-values vs. posterior probability (confidence interval interpretation)
  - In frequentist statistics, the p-value is taken as a measure of evidence against the null hypothesis.
  - p-value is not the same as the posterior probability of the null hypothesis.
  - Posterior probabilities can be substantially larger than p-values.
  - p-values can be misleading measures of evidence against the null hypothesis.
  - Do not confuse $P(H_0 \text{ true} \mid \text{data})$ with $P(\text{data} \mid H_0 \text{ true})$
- Highest posterior density credible set and central credible interval
  - The equi-tailed credible interval has $\frac{\alpha}{2}$ and $1 - \frac{1}{\alpha}$ quantiles of $\pi(\theta \mid x)$ at its endpoints.
    - Easy to compute from MC and MCMC samples
    - Nice invariance properties
  - The highest posterior density interval provides the shortest possible $(1 - \alpha)$ credible interval.
    - For symmetric distributions it coincides with equi-tailed credible interval
    - Hard to compute
    - Invariance property does not apply
- Frequentist asymptotics vs. Bayesian asymptotics
  - Frequentist asymptotics:

$$\widehat{\theta}_n \overset{\text{approx}}{\sim} \mathcal{N}\left(\theta_0, \frac{1}{n} I(\theta_0)^{-1}\right)$$

$$2\left(\log L_n(\widehat{\theta}_n) - \log L_n(\theta_0)\right) \xrightarrow{d} \chi^2_p$$

  - Bayesian asymptotics:

$$\theta \mid (x_1, \cdots, x_n) \overset{\text{approx}}{\sim} \mathcal{N}\left(\widehat{\theta}_n, \frac{1}{n} I(\widehat{\theta}_n)^{-1}\right)$$

    - Interpretation: the influence of the prior disappears asymptotically and the posterior is concentrated in a $\sqrt{\frac{1}{n}}$ neighborhood of the MLE.
- Likelihood principle
  - Conditionality principle
    - If an experiment for inference about a parameter θ is chosen independently from a collection of different possible experiments, then any experiment not chosen is irrelevant to the inference.
  - Sufficiency principle
    - If there are two observations x and y such that $T(x) = T(y)$ for a sufficient statistic T, then any conclusion about θ should be the same for x and y.
  - Likelihood principle
    - If there are two different experiments for inference about the same parameter θ and if the outcomes x and y from the two experiments are such that the likelihood functions differ only by a multiplicative constant, then the inference should be the same.
  - Conclusions
    - Frequentist tests can violate the likelihood principle
    - Bayesian tests do not suffer from this drawback
    - Point estimation by maximum likelihood does obey the likelihood principle

- Conjugate priors & sufficient statistics & exponential families?
  - If the posterior distribution $\pi(\theta \mid x)$ is in the same probability distribution family as the prior probability distribution $\pi(\theta)$, the prior and posterior are called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $f(x \mid \theta)$.
  - Exponential family is the only class of distributions which allow for sufficient statistics whose dimension is independent of n.
- Non-informative priors (uniform prior)
  - Finite volume
  - Not invariant under reparametrizations
- Improper priors
  - A prior $\pi(\theta)$ is called an improper prior if

$$\int_{\Theta} \pi(\theta) d\theta = \infty$$

    * Depending on the likelihood, $\pi(\theta) f(x \mid \theta)$ can have both finite or infinite total mass if $\pi(\theta)$ has infinite mass
  - Improper priors with proper posteriors can be justified as follows
    * Approximate an improper prior by a sequence of proper priors π_k
    * Show that the associated sequence of posteriors $\pi_k(\theta \mid x)$ converges to $\pi(\theta \mid x)$
- Equivariance of Jeffreys prior
  - Jeffreys prior:

$$\pi(\theta) \propto \det(I(\theta))^{1/2}$$

    where $I(\theta)$ is the Fisher information matrix

$$I(\theta) = -\mathbb{E}_{\theta}\left(\frac{\partial^2}{\partial\theta\partial\theta^T} \log f(X \mid \theta)\right)$$

  - Conclusions:
    * Jeffreys prior is usually a good choice for scalar parameters, but for vector parameters, it can have undesirable features.
    * It often leads to improper priors.
    * It violates the likelihood principle because the Fisher information contains an integral over X.
- Reference priors
  - A reference prior is a prior π for which the distance between the prior π and the posterior $\pi(\theta \mid x)$ is maximal. If the prior has a small influence on the posterior, the data x has the largest possible impact.
- Kullback-Leibler divergence

$$KL(f, g) = \int f(x) \log \frac{f(x)}{g(x)} dx$$

  - In general $KL(f, g) \neq KL(g, f)$
  - It satisfies $KL(f, g) \geq 0$ and $KL(f, g) = 0$ if and only if $f(x) = g(x)$ for almost all x
- Problem with maximization of mutual information $I(X, \theta)$
- Bernardo's approach for nuisance parameters
- Connection between regularization and prior
- Pros and cons of empirical Bayes method
  - Pros:

    - Do not need to compute the integral
    - Do not need to choose a hyperprior
  - Cons:
    - The data x is used twice
    - In general, $\pi(\theta \mid x, \widehat{\xi}(x))$ underestimates uncertainty in $\pi(\theta \mid x)$

- g-prior
  - The g-prior is a middle ground between being informative and completely non-informative. The idea is to introduce (possibly weak) prior information about β_γ but to bypass the prior correlation structure of β_γ
  - Since for the MLE $\hat{\beta}_\gamma$, $\text{Var}(\hat{\beta}_\gamma) = (\mathbf{X}_\gamma^\intercal \mathbf{X}_\gamma)^{-1}\hat{\sigma}^2$, the prior puts more mass in areas of the paramtere space where we expect the data to be more informative about β on average
  - $g > 0$ is a hyperparameter which can be interpreted as being inversely proportional to the amount of information available in the prior relative to the data
    - $g = 1$ gives the prior the same weight as the data
    - When g is large, the prior is weakly informative. For $g \to \infty, \pi(\beta_\gamma \mid \sigma^2) \propto 1$

- Model selection and improper priors
- Bayesian model averaging
  - Making predictions under every model
  - Averaging all predictions according to the posterior probability of each model
- How to choose g
  - Bartlett's paradox
  - Information paradox
- Laplace approximation
  - Laplace approximation refers to approximating the posterior normalizing constant with Laplace's method.
- Bayes factor and Bayesian information criterion (BIC)
- Independent Monte Carlo
  - Quantile transformation (inverse transform sampling)
- Rejection sampling
  - Rejection sampling is a Monte Carlo algorithm to sample data from a sophisticated distribution with the help of a proxy distribution.
- Importance sampling
  - Importance sampling is based on a similar idea as rejection sampling. Instead of rejecting some variables, we weight them with an appropriate weighting function.
  - Rejection area vs. acceptance area
- Sampling importance resampling
- Markov chain Monte Carlo
  - MCMC provides a class of algorithms for systematic random sampling from high-dimensional probability distributions. It works by constructing a Markov chain that eventually converges to the target distribution (stationary or equilibrium)
    - Irreducible (no matter where you start, the chain is always able to reach to other point in a finite number of iterations with positive probability)
    - Aperiodic (no periodic pattern)
    - Positive-recurrent (the expected return time to any state is finite)

  - Due to the Markov property, samples are not independent anymore
  - Independent Monte Carlo sampling is inefficient or even intractable for high-dimensional probabilistic models

- Gibbs sampler
  - The Gibbs sampler is a special case of the single-component Metropolis-Hastings algorithm, which uses the full conditional posterior distribution as the proposal distribution. For each time, the Gibbs sampler works with a univariate proposal distribution (all components except one are held fixed at their given values).
  - In each step, random values are generated from univariate distributions (easy to compute).
  - Require closed form of the full conditional posterior.
  - Acceptance rate is equal to 1.
  - Gibbs sampler does not require any tuning of proposal distribution.
  - Gibbs sampler can be ineffective when the parameter space is complicated or the parameters are highly correlated (slow).

- Metropolis-Hastings algorithm
  - Reversibility (the probability of transitioning from state A to state B is equal to the probability of transitioning from state B to state A)
  - A distribution π is called reversible for the transition kernel P if

$$\int_A \pi(x)P(x,B)dx = \int_B \pi(x)P(x,A)dx \qquad \forall A, B$$

$$\pi(x)p(x,y) = \pi(y)p(y,x)$$

  - Compact form of MH algorithm:

$$p(x,y) = q(x,y)a(x,y)$$

$$a(x,y) = \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right)$$

  - Random walk Metropolis algorithm

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

- Adaptive MCMC
  - How to choose Σ such that the average acceptance rate after the burn-in phase is 23.4%

- Hamiltonian Monte Carlo
  - Gibbs sampler makes only small moves
  - RWM algorithm makes either small moves (with higher acceptance rate and autocorrelation) or big moves with lower acceptance rate (inefficient)
  - Hamiltonian Monte Carlo allows for making big moves that are still accepted with high probability
  - Hamiltonian dynamics: using a time-reversible and volume-preserving numerical integrator to porpose a move to a new point in the state space.

- Sequential Monte Carlo

- Approximate Bayesian computation