1 Ingredients of Bayesian Inference

• The *model* for a typical observation X conditional on unknown parameter θ is $f(x|\theta)$. As a function of θ, $f(x|\theta) = \ell(\theta)$ is called *likelihood.* The functional form of f is fully specified up to a parameter θ.

• The parameter θ is supported by the parameter space Θ and considered a random variable. The random variable θ has a distribution $\pi(\theta)$ that is called the *prior.*

• If the prior for θ is specified up to a parameter τ, $\pi(\theta|\tau)$, τ is called a *hyperparameter.*

• The distribution $h(x,\theta) = f(x|\theta)\pi(\theta)$ is called the *joint* distribution for X and θ.

• The joint distribution can also be factorized as, $h(x,\theta) = \pi(\theta|x)m(x)$.

• The distribution $\pi(\theta|x)$ is called the *posterior* distribution for θ, given $X = x$.

• The *marginal* distribution $m(x)$ can be obtained by integrating out θ from the joint distribution $h(x,\theta)$,

$$m(x) = \int_\Theta h(x,\theta)d\theta = \int_\Theta f(x|\theta)\pi(\theta)d\theta.$$

• Therefore, the posterior $\pi(\theta|x)$ can be expressed as

$$\pi(\theta|x) = \frac{h(x,\theta)}{m(x)} = \frac{f(x|\theta)\pi(\theta)}{m(x)} = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}.$$

• Suppose $Y \sim f(y|\theta)$ is to be observed. The (posterior) predictive distribution of Y, given observed $X = x$ is

$$f(y|x) = \int_\Theta f(y|\theta)\pi(\theta|x)d\theta.$$

The marginal distribution $m(y) = \int_\Theta f(y|\theta)\pi(\theta)d\theta$ is sometimes called the prior predictive distribution.

name	notation	equal to
model, likelihood	$f(x\|\theta)$	
prior	$\pi(\theta)$	
joint	$h(x,\theta)$	$f(x\|\theta)\pi(\theta)$
marginal	$m(x)$	$\int_\Theta f(x\|\theta)\pi(\theta)d\theta$
posterior	$\pi(\theta\|x)$	$f(x\|\theta)\pi(\theta)/m(x)$
predictive	$f(y\|x)$	$\int_\Theta f(y\|\theta)\pi(\theta\|x)d\theta$

Example 1: Suppose that the likelihood (model) for X given θ is binomial $\mathcal{B}(n,\theta)$, i.e.,

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x},\ x = 0,1,\ldots,n,$$

and the prior is beta, $\mathcal{B}e(\alpha,\beta)$, where the hyperparameters α and β are known,

$$\pi(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},\ 0 \le \theta \le 1.$$

Find the joint, marginal, posterior, and predictive distributions.

- $h(x,\theta) = \frac{\binom{n}{x}}{B(\alpha,\beta)}\theta^{\alpha+x-1}(1-\theta)^{n-x+\beta-1},\ 0 \le \theta \le 1, x = 0,1,\ldots,n.$
- $m(x) = \frac{\binom{n}{x}B(x+\alpha,n-x+\beta)}{B(\alpha,\beta)},\ x = 0,1,\ldots,n.$
- $\pi(\theta|x) = \frac{1}{B(x+\alpha,n-x+\beta)}\theta^{\alpha+x-1}(1-\theta)^{n-x+\beta-1},\ 0 \le \theta \le 1,$ which is $\mathcal{B}e(\alpha+x, n-x+\beta).$
- $f(y|x) = \frac{\binom{n}{y}B(x+y+\alpha,2n-x-y+\beta)}{B(x+\alpha,n-x+\beta)},\ y = 0,1,\ldots,n.$

Example 2: This example is important because it addresses the normal likelihood and normal prior combination often used in practice. Assume that an observation, X is normally distributed with mean θ and known variance σ^2. The parameter of interest, θ also has normal distribution with parameters μ and τ^2. The Bayesian model is $X|\theta \sim \mathcal{N}(\theta,\sigma^2)$ and $\theta \sim \mathcal{N}(\mu,\tau^2)$. Find the marginal, posterior, and predictive distributions.

The exponent E in the joint distribution $h(x,\theta)$ is

$$E = -\frac{1}{2\sigma^2}(x-\theta)^2 - \frac{1}{2\tau}(\theta-\mu)^2.$$

After straightforward but tedious algebra,

$$E = -\frac{1}{2\rho}\left(\theta - \rho\left(\frac{x}{\sigma^2}+\frac{\mu}{\tau^2}\right)\right)^2 - \frac{1}{2(\sigma^2+\tau^2)}(x-\mu)^2,$$

where $\rho = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}$. Thus, since $h(x,\theta) = f(x|\theta)\pi(\theta) = \pi(\theta|x)m(x)$, the posterior distribution is normal with mean $\rho(\frac{x}{\sigma^2}+\frac{\mu}{\tau^2}) = \frac{\tau^2}{\sigma^2+\tau^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\mu$, and variance $\rho = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}$. The marginal is also normal with mean μ and variance $\sigma^2+\tau^2$. Therefore, $\theta|X \sim \mathcal{N}(\frac{\tau^2}{\sigma^2+\tau^2}X + \frac{\sigma^2}{\sigma^2+\tau^2}\mu, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$ and $X \sim \mathcal{N}(\mu,\sigma^2+\tau^2)$.

If $X_1, X_2, \ldots, X_n$ are observed instead of a single observation X, then the sufficiency of $\bar{X}$ implies that the Bayesian model for θ is the same as for X with σ^2 replaced by $\frac{\sigma^2}{n}$. In other words, the Bayesian model is

$$\bar{X}|\theta \sim \mathcal{N}(\theta, \frac{\sigma^2}{n}) \text{ and } \theta \sim \mathcal{N}(\mu,\tau^2).$$

producing

$$\theta|\bar{X} \sim \mathcal{N}(\frac{\tau^2}{\frac{\sigma^2}{n}+\tau^2}\bar{X} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+\tau^2}\mu, \frac{\frac{\sigma^2}{n}\tau^2}{\frac{\sigma^2}{n}+\tau^2}).$$

Show that predictive distribution of Y, given $X_1, \ldots, X_n$ is

$$Y|\bar{X} \sim \mathcal{N}(\frac{\tau^2}{\frac{\sigma^2}{n}+\tau^2}\bar{X} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+\tau^2}\mu, \sigma^2 + \frac{\frac{\sigma^2}{n}\tau^2}{\frac{\sigma^2}{n}+\tau^2}).$$

Notice that the posterior mean

$$\frac{\tau^2}{\frac{\sigma^2}{n}+\tau^2}\bar{X} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+\tau^2}\mu$$

is linear combination of the MLE estimator $\bar{X}$ and the prior mean μ with weights $\lambda = \frac{\tau^2}{\frac{\sigma^2}{n}+\tau^2}$ and $1-\lambda = \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+\tau^2}$. When the sample size n increases, the influence of the prior mean diminishes as $\lambda \to 1$. On the other hand if n is small and our prior opinion about μ is strong (τ^2 small) the posterior mean is close to μ. We will see later several more cases in which the posterior mean is a linear combination of its frequentist estimate and the prior mean.

Instead of completing the squares in E, one may first conclude that the any marginal distribution from multivariate normal will be normal as well, and to find $m(x)$ only the marginal mean and variance are needed.

By the elementary properties of conditional expectation, $EX = E(E(X|\theta)) = E\theta = \mu$, and $VarX = E(Var(X|\theta)) + Var(E(X|\theta)) = E\sigma^2 + Var(\theta) = \sigma^2 + \tau^2$. Now, the posterior can be obtained from the defining equation, $\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}$.

Example 3: Suppose that the normal observations

$$2.9441, -13.3618, 7.1432, 16.2356, -6.9178, 8.5800, 12.5400, -15.9373, -14.4096, 5.7115$$

are coming from $\mathcal{N}(\theta, 100)$ distribution. Assume that our prior on θ is $\mathcal{N}(20, 20)$ The posterior is normal $\mathcal{N}(6.8352, 6.6667)$. The three densities are shown in Figure 1. The Figure 1 is produced by `BA_nornor2.m`

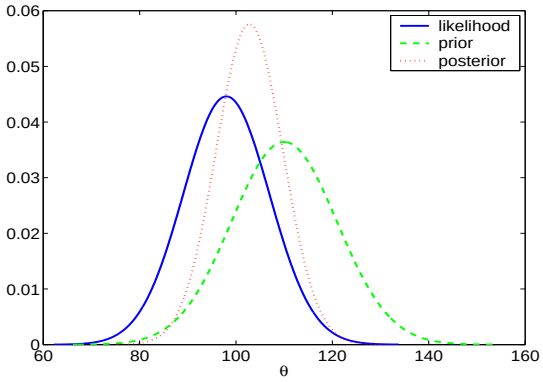

Figure 1: Likelihood, prior and posterior for the model and data in Example 3.

and commends:

```
> dat=[2.9441,-13.3618,7.1432,16.2356,-6.9178,8.5800,...
>      12.5400,-15.9373,-14.4096,5.7115];
> [m,v] = BA_nornor2(dat,100,20,20)
```

The fact that the normal/normal model, the posterior $\pi(\theta|X_1, \ldots, X_n) = \pi(\theta|\bar{X})$ is a special case of a general property given by Lemma bellow.

Lemma. Suppose the sufficient statistics $T = T(X_1, \ldots, X_n)$ exist. Then $\pi(\theta|X_1, \ldots, X_n) = \pi(\theta|T)$.

Proof: Factorization theorem for sufficient statistics is

$$f(x|\theta) = g(t, \theta)h(x),$$

where $t = T(x)$, and $h(x)$ does not depend on θ. Thus

$$\begin{aligned}\pi(\theta|x) &= \frac{\pi(\theta)f(x|\theta)}{\int_\theta \pi(\theta)f(x|\theta)d\theta} = \frac{\pi(\theta)g(t,\theta)h(x)}{\int_\theta \pi(\theta)g(t,\theta)h(x)d\theta} \\ &= \frac{\pi(\theta)g(t,\theta)}{\int_\theta \pi(\theta)g(t,\theta)d\theta} = \frac{\pi(\theta)g(t,\theta)\phi(t)}{\int_\theta \pi(\theta)g(t,\theta)\phi(t)d\theta} = \frac{\pi(\theta)f(t|\theta)}{\int_\theta \pi(\theta)f(t|\theta)d\theta} = \pi(\theta|t),\end{aligned}$$

where $f(t|\theta) = \int_{x:T(x)=t} f(x|\theta)dx = \int_{x:T(x)=t} g(t,\theta)h(x)dx = g(t,\theta)[\int_{x:T(x)=t} h(x)dx] = g(t,\theta)\phi(t)$.

Definition: The statistics $T = T(X)$ is sufficient (in the Bayesian sense) if for any prior the resulting posterior satisfies

$$\pi(\theta|X) = \pi(\theta|T).$$

Theorem: T is sufficient in the Bayesian sense if and only if it is sufficient in the usual sense.

The posterior is the ultimate experimental summary for a Bayesian. The location measures (especially the mean) of the posterior are of importance. The posterior mean represents one possible Bayes estimator of the parameter. We will see later, while discussing the decision theoretic setup, that posterior mode and median are also Bayes estimators under different loss functions.

Generalized Maximum Likelihood Estimator. The *generalized MLE* is the largest mode of the $\pi(\theta|x)$. The standard MLE maximizes $\ell(\theta)$, while the generalized MLE maximizes $\pi(\theta)\ell(\theta)$. Bayesians prefer the name MAP (maximum aposteriori) estimator or simply posterior mode.

The MAP estimator is popular in Bayesian analysis since it is often calculationaly less demanding that the posterior mean or median. The reason is simple, the posterior need not to be fully specified since $\text{argmax}_\theta \pi(\theta|x) = \text{argmax}_\theta f(x|\theta)\pi(\theta)$.

IQ Example. Jeremy, an enthusiastic GaTech student, poses a statistical model for his scores on standard IQ tests. He thinks that, in general, his scores are normally distributed with unknown mean θ and the variance 80. Prior (and expert) opinion is that the IQ of GaTech students, θ, is a normal random variable, with mean 110 and the variance 120. Jeremy took the test and scored 98. Estimate his true IQ θ in a Bayesian manner. [Solution: Posterior is $\mathcal{N}(102.8, 48)$. Use `BA_nornor1.m` as `[u,v] = BA_nornor1(98,80,110,120)` to obtain Figure 2.]

Example 4: Let $X_1, \ldots, X_n$, given θ are Poisson $\mathcal{P}(\theta)$ with probability mass function $f(x|\theta) = \frac{\theta^{x_i}}{x_i!}e^{-\theta}$, and $\theta \sim \mathcal{G}(\alpha, \beta)$ given by $\pi(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta}$ (See alternative parametrization of gamma distribution in Handout 0). Then, $\pi(\theta|X_1, \ldots, X_n) = \pi(\theta|\sum X_i) \propto \theta^{\sum X_i+\alpha-1}e^{-(n+\beta)\theta}$, which is $\mathcal{G}(\sum_i X_i + \alpha - 1, n + \beta)$. The mean is (see Handout 0), $E\theta|X = \frac{\sum X_i+\alpha}{n+\beta}$, and it can be represented in the form which compromises the two estimators, MLE and the prior mean,

$$E\theta|X = \frac{n}{n+\beta}\frac{\sum X_i}{n} + \frac{\beta}{\beta+n}\frac{\alpha}{\beta}.$$

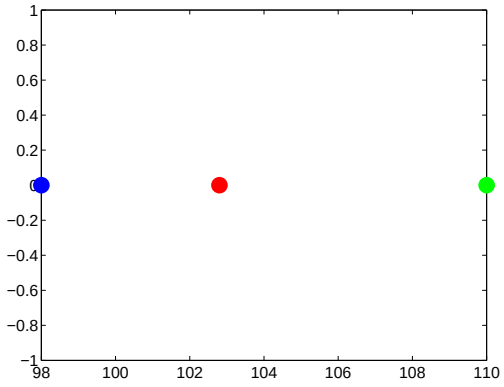

Figure 2: IQ Example. Extreme dots are the means (medians, modes) of the likelihood and prior. The dot in the middle is the posterior mean.

Show that the marginal is Negative Binomial.

Example 1 (continued): The mean of random variable with Beta $\mathcal{B}e(\alpha,\beta)$ distribution is $\frac{\alpha}{\alpha+\beta}$. Its mode is $\frac{\alpha-1}{\alpha+\beta-2}, \alpha>1, \beta>1$.

The posterior mean is

$$\frac{X+\alpha}{n+\alpha+\beta} = \frac{n}{n+\alpha+\beta}\frac{X}{n} + \frac{\alpha+\beta}{n+\alpha+\beta}\frac{\alpha}{\alpha+\beta},$$

yet another example in which the posterior mean is a weighted average of MLE, X/n, and the prior mean $\frac{\alpha}{\alpha+\beta}$. Notice, as in the normal/normal and Poisson/gamma cases, that when n is large the posterior mean is close to MLE, i.e. the weight $\frac{n}{n+\alpha+\beta}$ is close to 1, while when α is large, the posterior mean is close to the prior mean. Large α indicates small prior variance [for fixed β, the variance of $\mathcal{B}e(\alpha,\beta)$ behaves as $O(1/\alpha^2)$] and the prior is concentrated about its mean.

Example Probability of Heads (Handout 1, continued): A coin is flipped 4 times and tails showed up every single time. What is the probability of heads, θ? The MLE $\hat{\theta}=0$ is quite unreasonable estimator. If the prior is uniform on [0,1], the posterior is proportional to $\theta^0(1-\theta)^4$ which is Beta $\mathcal{B}e(1,5)$. The mean is $\hat{\theta}_B = \frac{1}{5+1} = \frac{1}{6}$, which is a more reasonable estimator of θ.

1.1 Exercises

1. Show that $\hat{\theta}=X$ is the MAP estimator for $X|\theta$ distributed as $f(x|\theta)=e^{-(x-\theta)}\mathbf{1}(x\in[\theta,\infty))$, and the prior on θ is standard Cauchy $\mathcal{C}a(0,1)$, $\pi(\theta)=\frac{1}{\pi(1+\theta^2)}$.

2. Suppose $X=(X_1,\ldots,X_n)$ is a sample from uniform $\mathcal{U}(0,\theta)$ distribution. Let θ have Pareto $\mathcal{P}a(\theta_0,\alpha)$ distribution (see Handout 0). Show that the posterior is Pareto $\mathcal{P}a(\max\{\theta_0,x_1,\ldots,x_n\}\alpha+n)$.

3. Let $X\sim\mathcal{G}(n/2,2\theta)$ (so that X/θ is χ^2_n). Let $\theta\sim\mathcal{IG}(\alpha,\beta)$. Show that the posterior is $\mathcal{IG}(n/2+$

$\alpha, (x/2 + \beta^{-1})^{-1})$.

4. If $X = (X_1, \ldots, X_n)$ is a sample from negative binomial $\mathcal{NB}(m, \theta)$ distribution and θ has beta $\mathcal{B}e(\alpha, \beta)$ distribution. Show that the posterior is beta $\mathcal{B}e(\alpha + mn, \beta + \sum_{i=1}^{n} x_i)$.