Conjugate priors in exponential families

Anirban Bhattacharya

http://www.stat.tamu.edu/~anirbanb

February 13, 2018

1 General setup

Let $x = (x_1, \ldots, x_n)$ denote the data. Assume a parametric family $\{f(x \mid \theta) : \theta \in \Theta\}$. We shall typically deal with situations where $\Theta \subset \mathbb{R}^d$. A Bayesian specification is completed by assigning a *prior distribution* $\pi(\cdot)$ on Θ. To begin with, we shall assume π is a probability distribution on Θ with $\int_\Theta \pi(\theta)d\theta = 1$. We shall relax this later when we talk about improper priors.

Key quantities:

(a) $\phi(x, \theta) = f(x \mid \theta)\pi(\theta)$.

(b) $m(x) = \int f(x \mid \theta)\pi(\theta)d\theta$ (normalizing constant/evidence).

(c) $\pi(\theta \mid x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}$ (posterior distribution of θ given data x). Let $B \subset \Theta$ be a subset of the parameter space. Then, upon observing the data x, our belief that the unknown parameter is in the set B is

$$\pi(B \mid x) = \frac{\int_B f(x \mid \theta)\pi(\theta)d\theta}{\int f(x \mid \theta)\pi(\theta)d\theta}.$$

If B_1 and B_2 are two subsets of the parameter space, the relative odds that the unknown parameter is in B_1 versus it is in B_2 is

$$\frac{\int_{B_1} f(x \mid \theta)\pi(\theta)d\theta}{\int_{B_2} f(x \mid \theta)\pi(\theta)d\theta}.$$

(d) If $y \sim g(\cdot \mid \theta, x)$, then $g(y \mid x) = \int g(y \mid \theta, x)\pi(\theta \mid x)d\theta$.

If the data are i.i.d., then $f(x \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta)$ and $m(x) = \int \prod_{i=1}^n f(x_i \mid \theta)\pi(\theta)d\theta$. If x_{n+1} is a future observation, then the predictive distribution

$$g(x_{n+1} \mid x_{1:n}) = \int f(x_{n+1} \mid \theta)\pi(\theta \mid x_{1:n})d\theta.$$

Thus, we average the likelihood of the future observation with the posterior distribution given the observed data as weight. This automatically takes into account the uncertainty in θ while making a prediction.

A nice feature of the posterior distribution is that it returns the same answer irrespective of whether the data is available in batch or sequentially. If the data arrives sequentially, we can continue updating our belief using the observed data up to any time point, and use the updated belief as the prior for the next time point. Once we have observed all the data, we get back the same posterior.

Initialize: Set $t = 0$ and $\pi_0 = \pi$.

Iterate: For $t = 1 : n$, $\pi_t(\theta \mid x_{1:t}) \propto \big[\prod_{i=1}^t f(x_i \mid \theta)\big]\pi_{t-1}(\theta)$.

Show that π_n is the usual posterior given $x = x_{1:n}$.

Example: bent coin. A bent coin is tossed n times. a Hs and $b = n - a$ Ts are observed. We wish to estimate the bias of the coin and predict the probability that the next toss will result in an H. This is a prototype for many statistical problems where only two types of outcomes are possible. For example: (1) a wellness survey randomly selects 200 TAMU students and asks whether they are generally happy with the quality of education received. 182 respond YES. What proportion of TAMU students are generally happy with the quality of education? (2) a new look of an email app is about to be unveiled. 1000 users are randomly selected and notified to participate in a survey to comment on the new look. 35 of them take the survey. what is the effectiveness of the survey mechanism?

We can formulate the problem as: $x_1, \ldots, x_n \mid p$ are independent Bernoulli(θ). Let us assign a $U(0, 1)$ prior to θ, which is a special case of a Beta(α, β) prior with $\alpha = \beta = 1$. The prior choice is a subjective assumption; we shall shortly see examples of other priors.

Likelihood: $f(x \mid \theta) = \theta^a(1 - \theta)^b$, where $a = \sum_{i=1}^n x_i$ and $b = n - a$.

Marginal likelihood:

$$\begin{aligned} m(x) &= \frac{1}{B(\alpha, \beta)} \int_0^1 \theta^{a+\alpha-1}(1 - \theta)^{b+\beta-1} d\theta \\ &= \frac{B(a + \alpha, b + \beta)}{B(\alpha, \beta)}. \end{aligned}$$

Posterior distribution: $\pi(\theta \mid x_{1:n}) \propto \theta^{a+\alpha-1}(1 - \theta)^{b+\beta-1}\mathbb{1}_{(0,1)}(\theta) \sim \text{Beta}(a + \alpha, b + \beta)$. The prior and posterior are in the same distribution family (Beta): we say that the Beta family is a conjugate prior for the Binomial likelihood.

Predictive distribution:

$$\begin{aligned} &P(x_{n+1} = 1 \mid x_1, \ldots, x_n) \\ &= \int P(x_{n+1} = 1 \mid \theta)\pi(\theta \mid x_{1:n}) d\theta \\ &= \int \theta\pi(\theta \mid x_{1:n}) d\theta \\ &= \frac{a + \alpha}{n + \alpha + \beta}. \end{aligned}$$

With $\alpha = \beta = 1$, the predictive probability is $(a + 1)/(n + 2)$, which is known as Laplace's rule.

Point and interval estimation: The posterior distribution summarizes the entire information regarding the unknown parameter upon observing the data. This distribution can now be summarized in various ways. For example, if only a point estimate is required, we can use some measure of central tendency of the posterior, e.g., the mean or the mode[1]. In the bent coin example, the posterior mean

$$\hat{\theta} = E(\theta \mid x) = \frac{a + \alpha}{n + \alpha + \beta} = \frac{n}{n + \alpha + \beta} \underbrace{\frac{a}{n}}_{m.l.e.} + \frac{\alpha + \beta}{n + \alpha + \beta} \underbrace{\frac{\alpha}{\alpha + \beta}}_{\text{prior mean}}$$

[1]When the posterior mode is used as a point estimate, the resulting estimate is often called a MAP (maximum a posteriori) estimate

is a convex combination of the m.l.e. (sample mean) and the prior mean (we shall see this repeatedly in many examples). The MAP estimate is

$$\hat{\theta}_{MAP} = \frac{a + \alpha - 1}{n + \alpha + \beta - 1}.$$

Observe that as n gets larger, both the posterior mean and the MAP estimate get closer and closer to the maximum likelihood estimate irrespective of the values of α, β (as long as they are fixed). This is again a more general phenomenon: in regular models, the m.l.e. and the posterior center merge asymptotically.

The Bayesian approach also provides an automatic characterization of uncertainty. No asymptotic arguments are required unlike the classical counterpart.

Definition. A $100 \times (1-\alpha)\%$ credible set for θ is any set $\mathcal{C} \subset \Theta$ such that $\pi(\mathcal{C} \mid x) \geq (1-\alpha)$. Typically, when θ is a one-dimensional parameter, $\mathcal{C}$ is taken to be an interval of the form $[l(x), u(x)]$. The shortest such interval is called the HPD (highest posterior density) credible region. If the posterior is unimodal, the HPD typically corresponds to the $100 \times \alpha/2\%$ and $100 \times (1-\alpha/2)\%$ quantiles.

Recall that the interpretation of a frequentist confidence set (say, 95%) is that under repeated replications of the experiment, the interval in question contains the true parameter value 95% of the times. A Bayesian credible set on the other hand makes a statement conditioned on the *observed data*; imaginary datasets that we haven't observed do not enter the picture. The statistician believes that given the particular dataset we have observed, the unknown parameter is 95% likely to be in the credible interval. For example, consider yourself to be an astronaut traveling to Mars on a five year flight. Compare the following two statements: (A) "in an imaginary population of spaceships like yours, the average life is greater than 10 years 95% of times", and (B) 11There is 95% probability that *this* spaceship will last at least 10 years". Which statement you will be more comforted with?

Albeit coming from philosophically different standpoints, Bayesian credible intervals often have the correct frequentist coverage (we shall make this concrete later on). This is very useful from a frequentist sense because we can obtain a uncertainty characterization without resorting to asymptotic arguments, whose assumptions may not hold in finite samples.

A comment on the marginal likelihood. Later, when we talk about model selection.

As we have seen above, the beta family is conjugate to the Bernoulli (equivalently Binomial) likelihood. Many other members of the exponential family have such conjugate priors. Consider two examples below:

Poisson example. Suppose $x_1, \ldots, x_n \mid \lambda \sim \text{Poisson}(\lambda)$. Consider a gamma prior on λ: $\lambda \sim \text{Gamma}(\alpha, \beta)$ with density

$$\frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta\lambda} \lambda^{\alpha-1} \mathbb{1}_{(0,\infty)}(\lambda).$$

Verify that

1. The posterior distribution $\lambda \mid x_{1:n} \sim \Gamma(T + \alpha, n + \beta)$, where $T = \sum_{i=1}^n x_i$.
2. The posterior mean $E(\lambda \mid x) = (T + \alpha)/(n + \beta) = (T/n) \times n/(n + \beta) + (\alpha/\beta) \times \beta/(n + \beta)$.

3. The marginal likelihood

$$m(x) = \Big[\prod_{i=1}^{n} x_i!\Big]^{-1} \frac{\beta^\alpha}{\Gamma(\alpha)} \times \frac{\Gamma(T+\alpha)}{(n+\beta)^{(T+\alpha)}}.$$

4. Find the predictive distribution.

Normal example (known variance). Suppose $x_1, \ldots, x_n \mid \mu \sim N(\mu, 1)$. Assume a normal prior $\mu \sim N(\xi, \tau^{-1})$. Verify that

1. The posterior distribution

$$\mu \mid x_{1:n} \sim N\left(\frac{n\bar{x} + \xi\lambda}{n+\lambda}, \frac{1}{n+\lambda}\right).$$

There is a simple way to remember this posterior in normal models with normal priors. The posterior precision (inverse variance) equals the data precision plus the prior precision. The posterior mean is a convex combination of the prior mean and the data mean, with the weights proportional to the respective precisions:

$$E(\mu \mid x_{1:n}) = \frac{n}{n+\lambda}\bar{x} + \frac{\lambda}{n+\lambda}\xi.$$

2. The marginal likelihood

$$\log m(x) = -\frac{n}{2}\log(2\pi) - \frac{S}{2} - \frac{1}{2}\,\frac{n\lambda}{n+\lambda}(\bar{x}-\xi)^2 - \frac{1}{2}\log(n+\lambda) + \frac{1}{2}\log(\lambda).$$

3. The predictive distribution:

$$f(x_{n+1} \mid x_{1:n}) \sim N(\hat{\mu}, 1 + \hat{\sigma^2}),$$

where $\hat{\mu}$ and $\hat{\sigma^2}$ are the posterior mean and variance respectively from the previous display.

Transformations. Consider $x_1, \ldots, x_n \mid \theta \sim \text{Bernoulli}(\theta)$ and suppose the parameter of interest is $\eta = \text{logit}(\theta)$. This is trivial in a Bayesian setting as we can obtain the posterior distribution of $\eta \mid x_{1:n}$ from the posterior of $\theta \mid x_{1:n}$ by a simple change of variable. We can now obtain point and interval estimates from the posterior distribution of $\eta \mid x_{1:n}$.

In general, if the parameter of interest is $\eta = g(\theta)$, we can use Monte Carlo (MC) to make approximate inference about η. Suppose we are given independent samples $\{\theta_t\}_{t=1}^T$ from $\pi(\theta \mid x_{1:n})$ (this may not be a trivial task unless the prior is conjugate - we shall see how to address this later on). Then, $\{\eta_t = g(\theta_t)\}_{t=1}^T$ are independent samples from $\pi(\eta \mid x_{1:n})$. The sample mean $\bar{\theta} = T^{-1}\sum_{t=1}^T \theta_t$ provides an approximation to the posterior mean $\int \eta\pi(\eta \mid x_{1:n})d\eta$ by the strong law of large numbers. Similarly, the sample quantiles of the $\{\eta_t\}$s can be used to construct credible intervals for η.

Dirichlet-multinomial example.

The two-parameter normal family. A random variable $Z \sim t_\nu$ (t distribution with $\nu > 0$ degrees of freedom) has density

$$f(z) = \frac{\Gamma(\nu+1/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\,\frac{1}{(1+t^2/\nu)^{(\nu+1)/2}}, \quad z \in \mathbb{R}.$$

For $d_1 \in \mathbb{R}$ and $d_2 > 0$, we say that $Z \sim t_\nu(d_1, d_2^2)$ if $(Z - d_1)/d_2 \sim t_\nu$. We shall call d_1 the center and d_2 the scale.

Suppose $x_1, \ldots, x_n \mid \mu, \sigma^2 \sim N(\mu, \tau^{-1})$, and the goal is to perform inference on the unknown mean μ and variance $\sigma^2 = \tau^{-1}$. We have separately seen so far: (i) if σ^2 is known, then a normal prior on μ is conjugate, (ii) if μ is known, then a gamma prior on the precision τ (VFY) is conjugate (this is equivalent to an inverse-gamma prior on σ^2). When both μ and τ are unknown, a natural idea is to use independent normal and gamma priors on μ and τ respectively. However, this prior is not conjugate.

A conjugate prior for (μ, τ) is defined hierarchically as: $\mu \mid \tau \sim (\xi, \lambda^{-1}\tau^{-1})$ and $\tau \sim$ gamma(a, b). This is called a normal-gamma (NG) prior, denoted $NG(\xi, \lambda, a, b)$, with prior hyperparameters $\xi, \lambda > 0, > 0, b > 0$. The joint prior distribution

$$\pi(\mu, \tau) \propto \tau^{1/2} e^{-\lambda\tau(\mu-\xi)^2/2} \tau^{a-1} e^{-b\tau}, \quad \mu \in \mathbb{R}, \tau > 0.$$

Note. For random variables (U, V), the joint distribution is uniquely specified by the conditional distribution of $U \mid V$ and the marginal distribution of V. (or vice versa). This extends to any number of random variables, where the joint distribution is determined by a cascading sequence of conditional distributions. Expressing a prior distribution for two or more parameters in this fashion is often called a hierarchical prior specification. Note that at any stage, we only have to specify a univariate distribution.

Apart from conjugacy, the conditional prior specification of $\mu \mid \tau$ leads to an easy interpretation of λ^{-1} as a signal-to-noise ratio: it quantifies how variable *a priori* the mean is relative to the variance.

The marginal prior of μ can be calculated by integrating over τ. We have

$$\begin{aligned}
&\pi(\mu) \\
&= \int \pi(\mu, \tau) d\tau \\
&= \frac{b^a}{\Gamma(a)} \frac{\lambda^{1/2}}{\sqrt{2\pi}} \int_0^\infty \tau^{a+1/2-1} e^{-[\lambda(\mu-\xi)^2/2+b]\tau} d\tau \\
&= \frac{\lambda^{1/2}}{\sqrt{2\pi}} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+1/2)}{[b + \lambda(\mu-\xi)^2/2]^{(a+1/2)}} \\
&= \frac{\Gamma(a+1/2)}{\Gamma(a)} \frac{1}{\sqrt{(2a)\pi}} \frac{1}{\sqrt{\{b/(a\lambda)\}}} \left[1 + \frac{(\mu-\xi)^2}{(2a) \times \{b/(a\lambda)\}}\right]^{-(2a+1)/2}.
\end{aligned}$$

Hence, the marginal prior for μ is a t distribution with center ξ, scale $\sqrt{b/(a\lambda)}$ and degrees of freedom $2a$. The uncertainty in the precision renders the marginal prior for μ to have heavier tails than the conditional prior.

The joint likelihood

$$L(\mu, \tau) \propto \tau^{n/2} \exp\left(-\frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2 \right).$$

Letting $S = \sum_{i=1}^n (x_i - \bar{x})^2$, we can write

$$L(\mu, \tau) \propto \tau^{n/2} e^{-\tau S/2} \; e^{-n\tau(\mu-\bar{x})^2/2}.$$

Based on what we have done for the normal mean model in the previous week, it is straightforward to see that

$$\pi(\mu \mid \tau, x_{1:n}) \sim N(\mu_n, \tau_n^{-1}),$$

where

$$\tau_n = (\lambda + n)\tau,\ \ \mu_n = \frac{\lambda}{n+\lambda}\xi + \frac{n}{n+\lambda}\bar{x} = \frac{n\bar{x} + \lambda\xi}{n+\lambda}.$$

Moreover, the joint posterior

$$\begin{aligned}\pi(\mu, \tau \mid x_{1:n}) &\propto L(\mu, \tau)\pi(\mu, \tau) \\ &\propto \tau^{n/2+a-1} e^{-\tau(S/2+b)} \tau^{1/2} \exp\left\{ -\frac{\tau}{2}\left[n(\bar{x} - \mu)^2 + \lambda(\mu - \xi)^2 \right] \right\}.\end{aligned}$$

Now,

$$\begin{aligned}&n(\bar{x} - \mu)^2 + \lambda(\mu - \xi)^2 \\ &= n\bar{x}^2 + \lambda\xi^2 + \left[\mu^2(n + \lambda) - 2\mu(\lambda\xi + n\bar{x})\right] \\ &= n\bar{x}^2 + \lambda\xi^2 + (n + \lambda)\left[\mu - \frac{n\bar{x} + \lambda\xi}{n+\lambda}\right]^2 - \frac{(n\bar{x} + \lambda\xi)^2}{(n + \lambda} \\ &= (n + \lambda)(\mu - \mu_n)^2 + \frac{n\lambda}{n+\lambda}(\bar{x} - \xi)^2.\end{aligned}$$

Let $Z_n = [n\lambda/(n + \lambda)](\bar{x} - \xi)^2$. Substituting in the previous display,

$$\begin{aligned}&\pi(\mu, \tau \mid x_{1:n}) \\ &\propto \tau^{n/2+a-1} e^{-\tau(S/2+Z_n/2+b)} \tau^{1/2} \exp\left\{ -\frac{(n + \lambda)\tau(\mu - \mu_n)^2}{2} \right].\end{aligned}$$

Comparing with the NG density, it is evident that $\mu, \tau \mid x_{1:n} \sim NG(\mu_n, n + \lambda, n/2 + a, S/2 + Z_n/2 + b)$. This proves that the NG prior is conjugate for the two parameter normal family.

To summarize:

1. The joint posterior $\mu, \tau \mid x_{1:n} \sim NG(\mu_n, n + \lambda, n/2 + a, S/2 + Z_n/2 + b)$.
2. The marginal posterior $\tau \mid x_{1:n} \sim \text{gamma}(n/2 + a, S/2 + Z_n/2 + b)$.
3. The conditional posterior $\mu \mid \tau, x_{1:n} \sim N(\mu_n, \tau_n^{-1})$.
4. The marginal posterior $\mu \mid x_{1:n}$ is a t distribution with center μ_n, scale

$$\sqrt{\frac{(S/2 + Z_n/2 + b)}{(n/2 + a)(n + \lambda)}}$$

and degrees of freedom $2(n/2 + a)$.

We have seen previously that if we are able to draw independent samples from the posterior distribution, then we can use Monte Carlo to obtain point and interval estimates for various functionals of the parameters. For example, we may be interested in doing inference on the coefficient of variation $c_v = \sigma/\mu = \tau^{-1/2}/\mu$. One option is to explicitly calculate the posterior distribution of c_v from the joint posterior of (μ, τ) via change of variable (check if this is tractable). Alternatively, we can:

a. For $t = 1, \ldots, T$, independently draw $(\mu^{(t)}, \tau^{(t)})$ from $\pi(\mu, \tau \mid x_{1:n})$.

b. Calculate $c_v^{(t)} = (\tau^{(t)})^{-1/2}/\mu^{(t)}$ for $t = 1, \ldots, T$.

c. Report sample quantiles of $\{c_v^{(t)}\}_{t=1}^T$ as an approximate credible interval for c_v.

How would you achieve step (a)? This is trivially achieved by first drawing a sample $\tau^{(t)}$ from a gamma$(n/2 + a, S/2 + Z_n/2 + b)$ distribution (the marginal posterior of $\tau \mid x_{1:n}$), and then

given $\tau^{(t)}$, drawing $\mu^{(t)} \mid \tau^{(t)} \sim N(\mu_n, \tau_n^{-1})$ (the conditional posterior of $\mu \mid \tau, x_{1:n}$).

[In general, if (U, V) have joint density $f_{U,V}$ with conditional $f_{U|V}$ and marginal f_V, then we can obtain a joint sample by first sampling $v \sim f_V$, and then $u \mid v \sim f_{U|V}(\cdot \mid V = v)$.]

A joint credible region for (μ, τ) can also be obtained from the MC samples. Please look up Figure 5.4. of Hoff for a plot of the joint posterior & read the accompanying example.

Normal-linear regression. Consider the homoskedastic normal linear regression model $Y \mid \beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 \mathrm{I}_n)$ for a $n \times d$ matrix of predictors X and $\beta \in \mathbb{R}^p$ a vector of regression coefficients. Elementwise, $y_i \sim N(x_i^{\mathrm{T}}\beta, \sigma^2)$, where $x_i^{\mathrm{T}} = (x_{i1}, \ldots, x_{id})^{\mathrm{T}}$ is the ith row of X, which contains the values of the d predictors for the ith observation. The goal is to estimate β from the data. We assume $n > d$ and that X has full column rank.

Let us initially assume σ^2 is known. We define two different conjugate priors for β:
(i) g-prior: $\beta \mid \sigma^2 \sim \mathcal{N}(0, g\,\sigma^2 (X^{\mathrm{T}}X)^{-1})$.
(ii) Ridge-prior: $\beta \mid \sigma^2 \sim N(0, \lambda^{-1}\,\sigma^2 \mathrm{I}_d)$.

Both priors lead to conjugate posteriors. g and λ are the respective hyperparameters for both prior. Note that for the g-prior, the prior covariance can be non-diagonal.

For the g-prior,

$$\beta \mid \sigma^2, y \sim \mathcal{N}\Bigg(\frac{g}{1+g} \underbrace{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y}_{\hat{\beta}_{OLS}}, \frac{g}{1+g}\,\sigma^2\,(X^{\mathrm{T}}X)^{-1} \Bigg),$$

while for the ridge prior,

$$\beta \mid \sigma^2, y \sim \mathcal{N}\Bigg(\underbrace{(X^{\mathrm{T}}X + \lambda \mathrm{I}_n)^{-1}X^{\mathrm{T}}y}_{\hat{\beta}_R}, \sigma^2\,(X^{\mathrm{T}}X + \lambda \mathrm{I}_n)^{-1} \Bigg).$$

Observe that the posterior mean with the ridge prior is the classical ridge estimator $\hat{\beta}_R$. Classically, the ridge estimator is defined as the solution to the convex optimization problem[2]

$$\hat{\beta}_R = \underset{\beta}{\operatorname{argmin}} \ \Big[\|y - X\beta\|^2 + \lambda \|\beta\|^2 \Big],$$

where λ is a tuning parameter. The Bayesian perspective lets us interpret λ as a precision parameter and learn it from the data (more later), and also provides uncertainty characterization for the ridge estimator. The posterior mean for the g-prior shrinks the OLS estimator towards the origin, and the amount of shrinkage is dictated by g. More on choice of g later.

The unknown σ^2 case. Taking cue from the analysis of the two-parameter normal model, it is not difficult to see that a multivariate version of the NG (since we are in the variance parameterization now) prior will continue to be conjugate. Specifically, if you want to use a g-prior for β, then the hierarchical prior is specified as

$$\beta \mid \sigma^2 \sim \mathcal{N}(0, g\,\sigma^2\,(X^{\mathrm{T}}X)^{-1}), \quad \sigma^2 \sim \mathrm{IG}(a/2, b/2).$$

[2] Although, in the Bayesian literature, the usage of such an estimator dates back earlier (early 1960s, at least) than the ridge estimator was coined from an optimization perspective.

The marginal prior for β can be exactly calculated as

$$\pi(\beta) = \int \pi(\beta \mid \sigma^2)\pi(\sigma^2)d\sigma^2 \propto \left[1 + \frac{\beta^{\mathrm{T}} X^{\mathrm{T}} X \beta}{gb}\right]^{-(a+d)/2}.$$

This can be recognized to be a multivariate t distribution.

Aside. A multivariate t distribution $t_{\nu,d}(\mu, \Sigma)$ has pdf proportional to

$$\left[1 + \frac{1}{\nu}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right]^{-(\nu+d)/2}.$$

For $\nu > 2$, the mean and covariance are given by μ and $\{\nu/(\nu-2)\}\,\Sigma$.

This implies the marginal prior for β is $t_{a,d}(0, (gb/a)\,(X^{\mathrm{T}}X)^{-1})$.

The joint posterior

$$\begin{aligned}\pi(\beta, \sigma^2 \mid y) &\propto f(y \mid \beta, \sigma^2)\,\pi(\beta \mid \sigma^2)\,\pi(\sigma^2) \\ &\propto (\sigma^2)^{-n/2}\, e^{-\frac{\|y - X\beta\|^2}{2\sigma^2}}\ (\sigma^2)^{-d/2} e^{-\frac{\beta^{\mathrm{T}} X^{\mathrm{T}} X \beta}{2g\sigma^2}}\ (\sigma^2)^{-(a/2+1)} e^{-\frac{b}{2\sigma^2}}.\end{aligned}$$

One way to proceed now is to do a bunch of algebra, simplify the expressions, and match with the prior to see if we can identify the parameters as before. I outline a slightly different approach here which is very handy in various problems.

Decompose $\pi(\beta, \sigma^2 \mid y) = \pi(\beta \mid \sigma^2, y)\,\pi(\sigma^2 \mid y)$. We already know, from our previous analysis, that,

$$\beta \mid \sigma^2, y \sim N\left(\frac{g}{1+g}\,\hat{\beta}_{OLS},\ \frac{g}{1+g}\,\sigma^2\,(X^{\mathrm{T}}X)^{-1}\right)$$

Hence, it remains to figure out the marginal posterior $\pi(\sigma^2 \mid y)$. Clearly,

$$\pi(\sigma^2 \mid y) \propto f(y \mid \sigma^2)\,\pi(\sigma^2), \quad f(y \mid \sigma^2) = \int f(y \mid \beta, \sigma^2)\pi(\beta \mid \sigma^2)d\beta.$$

Usually, you have to compute the integral over β to obtain the marginal density of $y \mid \sigma^2$. However, in this specific case, using normal distribution theory, we get (VFY!)

$$y \mid \sigma^2 \sim \mathcal{N}\Big(0, \sigma^2\{\mathrm{I}_n + gP_X\}\Big),$$

where $P_X = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is the projection matrix of X. Now, things have gotten much simpler. Using the fact (VFY!)

$$(\mathrm{I}_n + gP_X)^{-1} = \mathrm{I}_n - \frac{g}{1+g}P_X,$$

we get

$$\sigma^2 \mid y \sim \mathrm{IG}\left(\frac{a+n}{2}, \frac{b + y^{\mathrm{T}}\big(\mathrm{I}_n - \frac{g}{1+g}P_X\big)y}{2}\right).$$

The marginal posterior of $\beta \mid y$ is a multivariate t distribution; identify its parameters.