SDS 383D: Exercise 2

Diego Garcia-Olano

February 15, 2017

Problem 1. Bayes and the Gaussian linear model

A simple Gaussian location model

Take a simple Gaussian model with unknown mean and variance:

$$(yi|\theta, \sigma^2) \sim N(\theta, \sigma^2), i = 1, \ldots, n. \tag{1}$$

Let y be the vector of observations $y = (y_1, ..., y_n)^T$.

Suppose we place conjugate normal and inverse-gamma priors on θ and σ^2, respectively:

$$p(\theta|\sigma^2) \sim N(\mu, \tau^2\sigma^2)$$

$$\sigma^2 \sim InvGamma(\frac{d}{2}, \frac{\eta}{2})$$

where $\mu, \tau > 0$, $d > 0$ and $\eta > 0$ are fixed scalar hyperparameters.

*Note a crucial choice here: the error variance σ^2 appears in the prior for θ.
This affects the interpretation of the hyperparameter τ,
which is not the prior variance of θ, but rather the prior signal-to-noise ratio.
This is pretty common thing to do in setting up priors for location parameters:
to *scale the prior by the error variance*. There are a few good reasons to do this,
but historically the primary one has been analytical convenience (as you'll now see).

Here's a sensible way to interpret each of these four parameters:

- μ is a prior guess for θ.
- τ is a prior signal-to-noise ratio
  - that is, how disperse your prior is for θ, relative to the error standard deviation σ.
- d is like a *"prior sample size"* for the error variance σ^2.
- η is like a *"prior sum of squares"* for the error variance σ^2.
  More transparently, η/d is like a **"prior guess"** for the error variance σ^2. It's not exactly the prior mean for σ^2, but it's close to the prior mean as d gets larger, since the inverse-gamma(a,b) prior has expected value

$$E(\sigma^2) = \frac{b}{a-1} = \frac{\eta/2}{d/2-1} = \frac{\eta}{d-2}$$

  if d is large. This expression is only valid if d > 2.

What is meant by *"prior sample size"* (d) and *"prior sum of squares"* (η)?
Remember that **conjugate priors always resemble the likelihood functions** that they're intended to play nicely with. The two relevant quantities in the likelihood function for σ^2 are (i) the sample

size and (ii) the sums of squares. The prior here is designed to mimic the likelihood function for σ^2 that you'd get if you had a previous data set with sample size d and sums of squares η.

Precisions are easier than variances. It's perfectly fine to work with this form of the prior, and it's easier to interpret this way. But it turns out that we can make the algebra a bit cleaner by working with the precisions: $\omega = \frac{1}{\sigma^2}$ and $\kappa = \frac{1}{\tau^2}$ instead.

$$p(\theta|\omega) \sim N\big(\mu, (\omega\kappa)^{-1}\big)$$
$$\omega \sim Gamma(\frac{d}{2}, \frac{\eta}{2})$$

This means that the joint prior for (θ, ω) has the form:

$$p(\theta, \omega) \propto \omega^{\frac{d+1}{2}-1} \cdot exp\big(-\omega\frac{\kappa(\theta-\mu)^2}{2}\big) \cdot exp\big(-\omega \cdot \frac{\eta}{2}\big) \tag{2}$$

This is often called the $normal/gamma$ prior for (θ, ω) with parameters (μ, κ, d, η),
and its equivalent to a normal/inverse-gamma prior for (θ, σ^2).
The interpretation of κ is like a *prior sample size* for the mean θ

Note: you can obviously write this joint density for $p(\theta|\omega)$ in a way that combines the exponential terms, but this way keeps the bit involving θ separate, so that you can recognize the normal kernel. The term "kernel" is heavily overloaded in statistics so see `https://en.wikipedia.org/wiki/Kernel_(statistics)#In_Bayesian_statistics`.

(A) By construction, we know that the marginal prior distribution $p(\theta)$ is a gamma mixture of normals. Show that this takes the form of a centered, scaled t distribution:

$$p(\theta) \propto \big(1 + \frac{1}{v} \cdot \frac{(x-m)^2}{s^2}\big)^{-\frac{v+1}{2}}$$

with center m, scale s, and degrees of freedom v,
where you fill in the blank for m, s^2, and v in terms of the four parameters
of the normal-gamma family. * you did a problem like this in exercises 1!

By definition, the marginal of θ is the integral of the joint distribution $p(\theta, \omega)$ with respect to ω (ie, we integrate out ω) so

$$\begin{aligned} p(\theta) &= \int p(\theta, \omega) d\omega \\ &\propto \int \omega^{\frac{d+1}{2}-1} \cdot exp\big(-\omega\frac{\kappa(\theta-\mu)^2}{2}\big) \cdot exp\big(-\omega \cdot \frac{\eta}{2}\big) \\ &= \int \omega^{\frac{d+1}{2}-1} \cdot exp\big(\frac{\omega\kappa(\theta-\mu)^2 + -\omega\eta}{2}\big) \\ &= \int \omega^{(\frac{d+1}{2}-1)} \cdot exp\big(-\omega(\frac{\kappa(\theta-\mu)^2 + \eta}{2})\big) \end{aligned} \tag{3}$$

This is the kernel for gamma(a,b) for $a = \frac{d+1}{2}$ and $b = \frac{\kappa(\theta-\mu)^2+\eta}{2}$
As a probability distribution, it must integrate to 1/c where c is a constant.
In the case for a gamma distribution the leading constant for the pdf is
$\frac{b^a}{\Gamma(a)} = c$ thus $1/c = \frac{\Gamma(a)}{b^a}$ so we can rewrite (1) as:

$$\begin{aligned} &= \Gamma(\frac{d+1}{2})(\frac{\kappa(\theta-\mu)^2+\eta}{2})^{-\frac{d+1}{2}} \\ &\propto (\frac{\kappa(\theta-\mu)^2+\eta}{2})^{-\frac{d+1}{2}} \\ &= (\frac{\eta}{2} + \frac{\kappa(\theta-\mu)^2}{2})^{-\frac{d+1}{2}} \\ &= \big(\frac{\eta}{2} \cdot (1 + \frac{\kappa(\theta-\mu)^2}{\eta})\big)^{-\frac{d+1}{2}} \\ &= \frac{\eta}{2}^{-\frac{d+1}{2}} \cdot (1 + \frac{\kappa(\theta-\mu)^2}{\eta})^{-\frac{d+1}{2}} \end{aligned} \tag{4}$$

Since we only care about the value of the equation with respect to θ,
we can treat the first part of the equation as constant

$$\begin{aligned} &= (1 + \frac{\kappa(\theta-\mu)^2}{\eta})^{-\frac{d+1}{2}} \\ &= (1 + \frac{d}{d} \cdot \frac{\kappa(\theta-\mu)^2}{\eta})^{-\frac{d+1}{2}} \\ &= (1 + \frac{1}{d} \cdot \frac{(\theta-\mu)^2}{\eta d/\kappa})^{-\frac{d+1}{2}} \end{aligned} \tag{5}$$

This is close to the form of our solution,

$$p(\theta) \propto \big(1 + \frac{1}{v} \cdot \frac{(x-m)^2}{s^2}\big)^{-\frac{v+1}{2}}$$

Thus set $v = d$, $m = \mu$ and set $s^2 = \eta d/\kappa$ to get the centered, scaled t-student form.

(B) Assume the normal sampling model in Eq 1 and the normal-gamma prior in Eq 2.
Calculate joint posterior density $p(\theta, \omega|\mathbf{y})$, up to constant factors not depending on ω or θ.
Show that this is also a normal/gamma prior in the same form as above:

$$p(\theta, \omega|\mathbf{y}) \propto \omega^{(d^*+1)/2-1} exp\{-\omega \cdot \frac{\kappa^*(\theta-\mu^*)^2}{2}\} exp\{-\omega \cdot \frac{\eta^*}{2}\}$$

We have normal sampling model (eq 1): $(\mathbf{y}|\theta, \sigma^2) \sim N(\theta, \sigma^2)$
and also normal-gamma prior (eq 2) : $p(\theta, \omega) \propto \omega^{\frac{d+1}{2}-1} \cdot exp\big(-\omega\frac{\kappa(\theta-\mu)^2}{2}\big) \cdot exp\big(-\omega \cdot \frac{\eta}{2}\big)$

Calculate the joint posterior density,

$$
\begin{aligned}
p(\theta,\omega|\mathbf{y}) &\propto p(\mathbf{y}|\theta,\omega)p(\theta,\omega) \\
&= \omega^{n/2} exp\{-\omega\cdot(\frac{S_y+n(\bar{y}-\theta)^2}{2})\}\cdot\omega^{\frac{d+1}{2}-1}\cdot exp\big(-\omega\frac{\kappa(\theta-\mu)^2}{2}\big)\cdot exp\big(-\omega\cdot\frac{\eta}{2}\big) \\
&= \omega^{(\frac{n+d+1}{2}-1)}\cdot exp\{-\omega(\frac{S_y+n(\bar{y}-\theta)^2+\kappa(\theta-\mu)^2+\eta}{2})\} \\
&= \omega^{(\frac{n+d+1}{2}-1)}\cdot exp\{-\frac{\omega}{2}(S_y+n\bar{y}^2+n\theta^2-2n\bar{y}\theta+\kappa\theta^2+\kappa\mu^2-2\kappa\theta\mu+\eta)\} \\
&= \omega^{(\frac{n+d+1}{2}-1)}\cdot exp\{-\frac{\omega}{2}\big((n+\kappa)\theta^2-2(n\bar{y}+\kappa\mu)\theta+(S_y+n\bar{y}^2+\kappa\mu^2+\eta)\big)\}
\end{aligned}
\tag{6}
$$

The term in 2nd part that is multipled by $-\frac{\omega}{2}$ is of the form $ax^2-2bx+c$
so we can complete the square as such:

$$
\begin{aligned}
ax^2-2bx+c &= a[x^2-2(\frac{b}{a})x+\frac{c}{a}] \\
&= a[x^2-2(\frac{b}{a})x+(\frac{b}{a})^2-(\frac{b}{a})^2+\frac{c}{a}] \\
&= a[(x-\frac{b}{a})^2-(\frac{b}{a})^2+\frac{c}{a}] \\
&= a(x-\frac{b}{a})^2-\frac{b^2}{a}+c
\end{aligned}
\tag{7}
$$

Now plugging in for $a=n+\kappa$, $x=\theta$, $b=n\bar{y}+\kappa\mu$, and c = the final term in (6), we get:

$$
\begin{aligned}
&= \omega^{(\frac{n+d+1}{2}-1)}\cdot exp\{-\frac{\omega}{2}\big((n+\kappa)(\theta-\frac{n\bar{y}+\kappa\mu}{n+\kappa})^2-\frac{(n\bar{y}+\kappa\mu)^2}{n+\kappa}+(S_y+n\bar{y}^2+\kappa\mu^2+\eta)\big)\} \\
&= \omega^{(\frac{n+d+1}{2}-1)}\cdot exp\{-\frac{\omega}{2}\big((n+\kappa)(\theta-\frac{n\bar{y}+\kappa\mu}{n+\kappa})^2-\frac{(n^2\bar{y}^2+\kappa^2\mu^2+2n\bar{y}\kappa\mu)}{n+\kappa}+(n\bar{y}^2+\kappa\mu^2)+S_y+\eta\big)\} \\
&= \omega^{(\frac{n+d+1}{2}-1)}\cdot exp\{-\frac{\omega}{2}\big((n+\kappa)(\theta-\frac{n\bar{y}+\kappa\mu}{n+\kappa})^2-\frac{(n^2\bar{y}^2+\kappa^2\mu^2+2n\bar{y}\kappa\mu)}{n+\kappa}+\frac{n+\kappa}{n+\kappa}(n\bar{y}^2+\kappa\mu^2)+S_y+\eta\big)\} \\
&= \omega^{(\frac{n+d+1}{2}-1)}\cdot exp\{-\frac{\omega}{2}\big((n+\kappa)(\theta-\frac{n\bar{y}+\kappa\mu}{n+\kappa})^2-\frac{(n^2\bar{y}^2+\kappa^2\mu^2+2n\bar{y}\kappa\mu)}{n+\kappa}+\frac{n^2\bar{y}^2+\kappa n\bar{y}^2+n\kappa\mu^2+\kappa^2\mu^2}{n+\kappa}+ \\
&= \omega^{(\frac{n+d+1}{2}-1)}\cdot exp\{-\frac{\omega}{2}\big((n+\kappa)(\theta-\frac{n\bar{y}+\kappa\mu}{n+\kappa})^2+\frac{(-n^2\bar{y}^2-\kappa^2\mu^2-2n\bar{y}\kappa\mu)}{n+\kappa}+\frac{n^2\bar{y}^2+\kappa n\bar{y}^2+n\kappa\mu^2+\kappa^2\mu^2}{n+\kappa}- \\
&= \omega^{(\frac{n+d+1}{2}-1)}\cdot exp\{-\frac{\omega}{2}\big((n+\kappa)(\theta-\frac{n\bar{y}+\kappa\mu}{n+\kappa})^2+\frac{(-n^2\bar{y}^2-\kappa^2\mu^2-2n\bar{y}\kappa\mu)}{n+\kappa}+\frac{n^2\bar{y}^2+\kappa n\bar{y}^2+n\kappa\mu^2+\kappa^2\mu^2}{n+\kappa}- \\
&= \omega^{(\frac{n+d+1}{2}-1)}\cdot exp\{-\frac{\omega}{2}\big((n+\kappa)(\theta-\frac{n\bar{y}+\kappa\mu}{n+\kappa})^2+\frac{(-\kappa^2\mu^2-2n\bar{y}\kappa\mu+\kappa n\bar{y}^2+n\kappa\mu^2)}{n+\kappa}+S_y+\eta\big)\} \\
&= \omega^{(\frac{n+d+1}{2}-1)}\cdot exp\{-\frac{\omega}{2}\big((n+\kappa)(\theta-\frac{n\bar{y}+\kappa\mu}{n+\kappa})^2+\frac{n\kappa(-\kappa\mu^2-2\bar{y}\mu+\bar{y}^2+\mu^2)}{n+\kappa}+S_y+\eta\big)\} \\
&= \omega^{(\frac{n+d+1}{2}-1)}\cdot exp\{-\frac{\omega}{2}\big((n+\kappa)(\theta-\frac{n\bar{y}+\kappa\mu}{n+\kappa})^2+\frac{-n\kappa^2\mu^2\cdot n\kappa(\bar{y}-\mu)^2}{n+\kappa}+S_y+\eta\big)\}
\end{aligned}
\tag{8}
$$

So we have

$$
p(\theta,\omega|\mathbf{y}) \propto \omega^{(\frac{n+d+1}{2}-1)}\cdot exp\{-\frac{\omega}{2}\big((n+\kappa)(\theta-\frac{n\bar{y}+\kappa\mu}{n+\kappa})^2\}\cdot exp\{-\frac{\omega}{2}(\frac{-n\kappa^2\mu^2\cdot n\kappa(\bar{y}-\mu)^2}{n+\kappa}+S_y+\eta)\}
$$

which has the form of the normal-gamma

$$p(\theta, \omega|\mathbf{y}) \propto \omega^{(d^*+1)/2-1} exp\{-\omega \cdot \frac{\kappa^*(\theta-\mu^*)^2}{2}\} exp\{-\omega \cdot \frac{\eta^*}{2}\}$$

From this form of the posterior, the new updated parameters are

- $\mu \rightarrow \mu^* = \frac{n\bar{y}+\kappa\mu}{n+\kappa}$
- $\kappa \rightarrow \kappa^* = n + \kappa$ AND $d \rightarrow d^* = n + d$
- $\eta \rightarrow \eta^* = \frac{-n\kappa^2\mu^2 \cdot n\kappa(\bar{y}-\mu)^2}{n+\kappa} + S_y + \eta$

(C) From the joint posterior just derived, what is the conditional posterior distribution $p(\theta|\mathbf{y}, \omega)$? - you can read it off directly from the joint distribution, since you took care to set up things so that the joint posterior was in the same form as Equation 2.

* normal - gamma (normal sampling model with gamma prior on precision) from Eq 2:

$$p(\theta, \omega) \propto \omega^{\frac{d+1}{2}-1} \cdot exp\big(-\omega \frac{\kappa(\theta-\mu)^2}{2}\big) \cdot exp\big(-\omega \cdot \frac{\eta}{2}\big)$$

In general $p(\theta, \omega) \propto p(\omega)p(\theta|\omega)$
and $p(\omega) = \omega^{\frac{d+1}{2}-1} \cdot \big(-\omega \cdot \frac{\eta}{2}\big)$ and $p(\theta|\omega) = exp\big(-\omega \frac{\kappa(\theta-\mu)^2}{2}\big)$
Thus reading off from B,

$$p(\theta|\mathbf{y}, \omega) \propto exp\{-\frac{\omega}{2}\big((n+\kappa)(\theta - \frac{n\bar{y}+\kappa\mu}{n+\kappa})^2\}$$

This is the Normal distribution form, $p(\theta|\mathbf{y}, \omega) \sim N(\frac{n\bar{y}+\kappa\mu}{n+\kappa}, -\omega(n+\kappa))$

(D) From the joint posterior calculated in (B), what is the marginal posterior distribution $p(\omega|y)$? - Unlike the previous question, where you could just read it off, here you have to integrate over θ. Ignore constants not depending on ω in calculating this integral.

$$\begin{aligned} p(\omega|y) &= \int p(\theta, \omega|y) d\theta \\ &\propto \int \omega^{\frac{d^*+1}{2}-1} exp\{-\omega \cdot \frac{\kappa^*(\theta-\mu^*)^2}{2}\} exp\{-\omega \cdot \frac{\eta^*}{2}\} d\theta \\ &= \omega^{\frac{d^*+1}{2}-1} \cdot exp\{-\omega \cdot \frac{\eta^*}{2}\} \int exp\{-\omega \cdot \frac{\kappa^*(\theta-\mu^*)^2}{2}\} d\theta \end{aligned} \tag{9}$$

The integral here is a Normal kernel: $N(\mu^*, \omega\kappa^*)$ and as a probability distribution must integrate to $\frac{1}{c}$, where c is the constant of proportionality for the Normal density, which in this case is $\omega^{1/2}$. Thus $\frac{1}{c} = \frac{1}{\omega^{1/2}} = \omega^{-1/2}$

$$\begin{aligned} &= \omega^{\frac{d^*+1}{2}-1} \cdot exp\{-\omega \cdot \frac{\eta^*}{2}\} \cdot \omega^{-1/2} \\ &= \omega^{\frac{d^*}{2}-1} \cdot exp\{-\omega \cdot \frac{\eta^*}{2}\} \end{aligned} \tag{10}$$

This is the kernel for the gamma distribution, $Gamma(\frac{d^*}{2}, \frac{\eta^*}{2})$ as seen in Eq 2.

(E) From (C) and (D), we know that the marginal posterior distribution $p(\theta|\mathbf{y})$ is a gamma mixture of normals. Show that this takes the form of a centered, scaled t distribution:

$$p(\theta) \propto \big(1 + \frac{1}{v} \cdot \frac{(x-m)^2}{s^2}\big)^{-\frac{v+1}{2}}$$

with center m, scale s, and degrees of freedom v (fill in the blank for m, s^2, and v). express the parameters of this t distribution in terms of the four parameters of the normal-gamma posterior for (θ, ω). Note: since you've set up the normal-gamma family in this careful conjugate form, this should require no extra work. It's just part (A), except for the prior rather than the posterior.

$$\begin{aligned} p(\theta) &= \int p(\theta, \omega|\mathbf{y}) d\omega \\ &\propto \int \omega^{\frac{d^*+1}{2}-1} \cdot exp\big(-\omega \frac{\kappa^*(\theta-\mu^*)^2}{2}\big) \cdot exp\big(-\omega \cdot \frac{\eta^*}{2}\big) d\omega \end{aligned} \tag{11}$$

This is the kernel for gamma(a,b) for $a = \frac{d^*+1}{2}$ and $b = \frac{\kappa^*(\theta-\mu^*)^2+\eta^*}{2}$
As a probability distribution, it must integrate to 1/c where c is a constant, in this case that for the gamma distribution pdf, $c = \frac{b^a}{\Gamma(a)}$ thus $1/c = \frac{\Gamma(a)}{b^a}$ so we can rewrite (1) as:

$$\begin{aligned} &= \Gamma(\frac{d+1}{2})(\frac{\kappa^*(\theta-\mu^*)^2+\eta^*}{2})^{-\frac{d^*+1}{2}} \\ &\propto (\frac{\kappa^*(\theta-\mu^*)^2+\eta}{2})^{-\frac{d^*+1}{2}} \\ &= (\frac{\eta^*}{2} + \frac{\kappa^*(\theta-\mu^*)^2}{2})^{-\frac{d^*+1}{2}} \\ &= \big(\frac{\eta^*}{2} \cdot (1 + \frac{\kappa^*(\theta-\mu^*)^2}{\eta^*})\big)^{-\frac{d^*+1}{2}} \\ &= \frac{\eta^*}{2}^{-\frac{d^*+1}{2}} \cdot (1 + \frac{\kappa^*(\theta-\mu^*)^2}{\eta^*})^{-\frac{d^*+1}{2}} \end{aligned} \tag{12}$$

Since we only care about the value of the equation with respect to θ,
we can treat the first part of the equation as constant

$$\begin{aligned} &= (1 + \frac{\kappa^*(\theta-\mu^*)^2}{\eta^*})^{-\frac{d^*+1}{2}} \\ &= (1 + \frac{d^*}{d^*} \cdot \frac{\kappa^*(\theta-\mu^*)^2}{\eta^*})^{-\frac{d^*+1}{2}} \\ &= (1 + \frac{1}{d^*} \cdot \frac{(\theta-\mu^*)^2}{\frac{\eta^*}{d^*\kappa^*}})^{-\frac{d^*+1}{2}} \end{aligned} \tag{13}$$

This is close to the form of our solution, $p(\theta) \propto \big(1 + \frac{1}{v} \cdot \frac{(x-m)^2}{s^2}\big)^{-\frac{v+1}{2}}$
Thus set $v = d^*$, $m = \mu^*$ and set $s^2 = \frac{\eta^*}{d^*\kappa^*}$ to get the centered, scaled t-student form.

(F) *True or false*: in the limit as the prior parameters [κ, d, and η] approach zero, the priors $p(\theta)$ and $p(\omega)$ are valid probability distributions.
- a valid probability distribution must integrate to 1 (or something finite, so that it can normalized to integrate to 1) over its domain.

$$p(\omega) = \frac{(\frac{\eta}{2})^{\frac{1}{2}}}{\Gamma(\frac{d}{2})}\omega^{\frac{d}{2}-1} \cdot exp^{-\frac{\eta}{2}\omega}$$

For the ω factor, in the left part of the RHS,
as $d, \eta \to 0$, $\omega^{\frac{d}{2}-1}$ approaches $\frac{1}{\omega}$ and its scalar $\frac{(\frac{\eta}{2})^{\frac{1}{2}}}{\Gamma(\frac{d}{2})}$ approaches $\frac{0}{\Gamma(0)} = \frac{0}{\infty}$ which is ill defined, hence $p(\omega)$ is not a valid probability distribution.

(G) *True or false*: in the limit as the prior parameters [κ, d, and η] approach zero, the posteriors $p(\theta|\mathbf{y})$ and $p(\omega|\mathbf{y})$ are valid probability distributions.

From (E) we have,

$$p(\theta|\mathbf{y}) = (1 + \frac{1}{d^*} \cdot \frac{(\theta - \mu^*)^2}{\frac{\eta^*}{d^*\kappa^*}})^{-\frac{d^*+1}{2}}$$

where the posterior parameters are

- $\mu \to \mu^* = \frac{n\bar{y}+\kappa\mu}{n+\kappa}$
- $\kappa \to \kappa^* = n + \kappa$
- $d \to d^* = n + d$
- $\eta \to \eta^* = \frac{n\kappa(\bar{y}-\mu)^2}{n+\kappa} + S_y + \eta$

As these posteriors $\to 0$,

- $\mu^* = \frac{n\bar{y}+\kappa\mu}{n+\kappa} \to \frac{n\bar{y}+(0)\mu}{n+(0)} = \frac{n\bar{y}}{n} = \bar{y}$
- $\kappa^* = n + \kappa \to n + (0) = n$
- $d^* = n + d \to n + (0) = n$
- $\eta^* = \frac{n\kappa(\bar{y}-\mu)^2}{n+\kappa} + S_y + \eta \to 0 + S_y + 0 = S_y$

Thus plugging into $p(\theta|\mathbf{y})$ above we get

$$(1 + \frac{1}{n} \cdot \frac{(\theta - \bar{y})^2}{\frac{S_y}{n^2}})^{-\frac{n+1}{2}}$$

which is a valid probability distribution with the form of the centered, scaled t distribution, with parameters

- $m = \bar{y}$
- $v = n$
- $s^2 = \frac{S_y}{n^2}$

s^2 is what we would expect if we had no prior knowledge of θ Thus from a philosophical view point, our posterior distributions do not have to rely on valid prior probability distributions. This is a contentious issue in the debate between frequentists and bayesians.

From (D) we have,

$$p(\omega|\mathbf{y}) = \omega^{\frac{d^*}{2}-1} \cdot exp\{-\omega \cdot \frac{\eta^*}{2}\}$$

plugging in the posterior variables here we get

$$p(\omega|\mathbf{y}) = \omega^{\frac{n}{2}-1} \cdot exp\{-\omega \cdot \frac{S_y}{2}\}$$

which is a valid gamma distribution, $Gamma(\frac{n}{2}, \frac{S_y}{2})$

(H) Your result in (E) implies that a Bayesian credible interval for θ takes the form

$$\theta \in m \pm t^* \cdot s,$$

where m and s are the posterior center and scale parameters from (F),
and t^* is the appropriate critical value of the t distribution for your coverage level and degrees of freedom (e.g. it would be 1.96 for a 95% interval under the normal distribution).

True or false: In the limit as the prior parameters κ, d, and η approach zero,
the Bayesian credible interval for θ becomes identical to the classical (frequentist) confidence interval for θ at the same confidence level.

Problem 2. The conjugate Gaussian linear model

Now consider the Gaussian linear model,

$$(\mathbf{y}|B, \sigma^2) \sim N(XB, (\omega\Lambda)^{-1}),$$

where:

- **y** is an n vector of responses,
- X is an nxp matrix of features,
- $\omega = 1/\sigma^2$ is the error precision, and
- Λ is some known matrix.

A typical setup would be $\Lambda = I$, the nxn identity matrix, so that the residuals of the model are i.i.d. normal with variance σ^2. But we'll consider other setups as well, so we'll leave a generic Λ matrix in the sampling model for now. Note that when we write the model this way, we typically assume one of two things: either

(1) that both the y variable and all the X variables have been centered to have mean zero, so that an intercept is unnecessary; or

(2) that X has a vector of 1's as its first column, so the first entry in B is actually the intercept. We'll again work in terms of the precision $\omega = \sigma^2$, and consider a normal-gamma prior for B:

$$(B|\omega) \sim N(m, (\omega K)^{-1}) \tag{14}$$

$$\omega \sim Gamma(d/2, \eta/2) \tag{15}$$

Here K is a pxp precision matrix in the multivariate normal prior for B, assumed to be known. The items below follow a parallel path to the derivations you did for the Gaussian location model - except for the multivariate case. Don't reinvent the wheel if you don?t have to: you should be relying heavily on your previous results about the multivariate normal distribution. That is, if you find yourself completing the square over and over again with matrices and vectors, you should stop and revisit your Exercises 1 solutions.

Basics

(A) Derive the conditional posterior $p(B|\mathbf{y}, \omega)$.
Given $(y|\beta, \omega) \sim N(X\beta, (\omega\Lambda)^{-1})$ and priors for $(B|\omega)$ and ω, we may write:

$$\begin{aligned} p(\beta, \omega|y) &\propto p(y|\beta, \omega) \cdot p(\beta, \omega) \\ &= p(y|\beta, \omega) \cdot p(\beta|\omega) \cdot p(\omega) \\ &= \overbrace{(\omega^{n/2} \exp[-\frac{1}{2}(y - X\beta)^T \omega\Lambda(y - X\beta)])}^{\text{(i)}} \overbrace{(\omega^{p/2} \exp[-\frac{1}{2}(\beta - m)^T \omega K(\beta - m)])}^{\text{(ii)}} \overbrace{(\omega^{d/2-1} \exp[-\frac{\eta}{2}])}^{\text{(iii)}} \\ &= \omega^{(d+p+n)/2-1} \exp\left(-\frac{1}{2}\omega\left[(y - X\beta)^T\Lambda(y - X\beta) + (\beta - m)^T K(\beta - m) + \eta\right]\right) \end{aligned} \tag{16}$$

Focusing on second exp term we get:
$\exp\left(-\frac{1}{2}\omega\left[(y^T - \beta^T X^T)(\Lambda y - \Lambda X\beta) + (\beta^T - m^T)(K\beta - Km) + \eta\right]\right)$
$\exp\left(-\frac{1}{2}\omega\left[(y^T\Lambda y - \beta^T X^T\Lambda y - y^T\Lambda X\beta + \beta^T X^T\Lambda X\beta) + (\beta^T K\beta - m^T K\beta - \beta^T Km + m^T Km) + \eta\right]\right)$
$\exp\left(-\frac{1}{2}\omega\left[y^T\Lambda y - 2y^T\Lambda X\beta + \beta^T X^T\Lambda X\beta + \beta^T K\beta - 2m^T K\beta + m^T Km + \eta\right]\right)$
$\exp(-\frac{1}{2}\omega \underbrace{\left[\beta^T(X^T\Lambda X + K)\beta - 2(y^T\Lambda X + m^T K)\beta + y^T\Lambda y + m^T Km + \eta\right]}_{\text{(iv)}})$

Now (16) is

$$\omega^{(d+p+n)/2-1} \exp\left(-\frac{1}{2}\omega\left[\beta^T(X^T\Lambda X + K)\beta - 2(y^T\Lambda X + m^T K)\beta + y^T\Lambda y + m^T Km + \eta\right]\right)$$

And we let,

- $A = X^T\Lambda X + K$

- $b^T = y^T \Lambda X + m^T K \Rightarrow b = X^T \Lambda y + Km$
- $c = y^T \Lambda y + m^T K m + \eta$

and the expression in (iv) once the square is completed becomes,

$$\begin{aligned} \beta^T A \beta - 2b^T \beta + c &= \beta^T A \beta - 2b^T \beta + b^T A^{-1} b - b^T A^{-1} b + c \\ &= (\beta - A^{-1}b)^T A (\beta - A^{-1}b) - b^T A^{-1} b + c, \end{aligned} \tag{17}$$

Now let

$$\begin{aligned} m^\star &= A^{-1}b = (X^T \Lambda X + K)^{-1}(X^T \Lambda y + Km) \\ K^\star &= A = X^T \Lambda X + K, \end{aligned} \tag{18}$$

and we can also simplify the term

$$\begin{aligned} b^T A^{-1} b &= b^T I A^{-1} \\ &= b^T A^{-1} A A^{-1} b \\ &= m^{\star T} K^\star m^\star, \end{aligned} \tag{19}$$

and finally let

$$\begin{aligned} \eta^\star &= c - m^{\star T} K^\star m^\star \\ &= \eta + y^T \Lambda y + m^T K m - m^{\star T} K^\star m^\star \end{aligned} \tag{20}$$

We can at last express (iv) as

$$(\beta - m^\star)^T K^\star (\beta - m^\star) + \eta^\star.$$

Now the joint posterior distribution may be written as

$$\begin{aligned} p(\beta, \omega | y) &\propto \omega^{(d+p+n)/2-1} \exp\left(-\frac{1}{2}\omega\left[(\beta - m^\star)^T K^\star (\beta - m^\star) + \eta^\star\right]\right) \\ &= \underbrace{\omega^{p/2} \exp\left[-\frac{1}{2}(\beta - m^\star)^T \omega K^\star (\beta - m^\star)\right]}_{p(\beta|y,\omega) \sim N(m^*, (\omega K^\star)^{-1})} \underbrace{\omega^{d^\star/2-1} \exp\left[-\frac{1}{2}\eta^\star \omega\right]}_{p(\omega|y) \sim \text{Gamma}\left(\frac{d^\star}{2}, \frac{\eta^\star}{2}\right)}, \end{aligned} \tag{21}$$

with

$$\begin{aligned} m^\star &= (X^T \Lambda X + K)^{-1}(X^T \Lambda y + Km) \\ K^\star &= X^T \Lambda X + K \\ \eta^\star &= \eta + y^T \Lambda y + m^T K m - m^{\star T} K^\star m^\star \\ d^\star &= d + n \end{aligned} \tag{22}$$

(B) Derive the marginal posterior $p(\omega|\mathbf{y})$
$p(\omega|y) \sim \text{Gamma}\left(\frac{d^\star}{2}, \frac{\eta^\star}{2}\right)$

(C) Putting these together, derive the marginal posterior $p(B|\mathbf{y})$. The marginal posterior for β may be found with

$$\begin{aligned} p(\beta|y) &= \int_0^\infty p(\beta,\omega|y)d\omega \\ &\propto \int_0^\infty \omega^{(d+p+n)/2-1}\exp\left(-\frac{1}{2}\omega\left[(\beta-m^\star)^T K^\star(\beta-m^\star)+\eta^\star\right]\right)d\omega \\ &\propto \left[\frac{1}{2}\left[(\beta-m^\star)^T K^\star(\beta-m^\star)+\eta^\star\right]\right)^{-\frac{d+p+n}{2}} \\ &\propto \left[1+(\beta-m^\star)^T\frac{K^\star}{\eta^\star}(\beta-m^\star)\right]^{-\frac{d+n+p}{2}} \\ &\propto \left[1+\frac{1}{d+n}\cdot(\beta-m^\star)^T\cdot\frac{d+n}{\eta^\star}K^\star\cdot(\beta-m^\star)\right]^{-\frac{d+n+p}{2}}, \end{aligned} \tag{23}$$

which a Student's t-distribution with $d+n$ degrees of freedom, mean vector $m^\star$, and covariance matrix $\frac{d+n}{\eta^\star}K^\star$.

(D) Take a look at the data in "gdpgrowth.csv" from the class website, which has macroeconomic variables for several dozen countries. In particular, consider a linear model (with intercept) for a country?s GDP growth rate (GR6096) versus its level of defense spending as a fraction of its GDP (DEF60).
Fit the Bayesian linear model to this data set, choosing $\Lambda = I$ and something diagonal and pretty vague for the prior precision matrix $K = diag(\kappa_1, \kappa_2)$. Inspect the fitted line (graphically). Are you happy with the fit? Why or why not?
see exercises2/Ex2Dbayes-gdpgrowth-R

A heavy-tailed error model

Now it's time for your first "real" use of the hierarchical modeling formalism to do something cool. Here's the full model you'll be working with:

$$\begin{aligned} (\mathbf{y}|B,\omega,\Lambda) &\sim N(XB,(\omega\Lambda)^{-1}) \\ \Lambda &= diag(\lambda_1,...,\lambda_n) \\ \lambda_i &\overset{iid}{\sim} Gamma(h/2,h/2) \\ (B|\omega) &\sim N(m,(\omega K)^{-1} \\ \omega &\sim Gamma(d/2,\eta/2) \end{aligned} \tag{24}$$

where h is a fixed hyper parameter.

(A) Under this model, what is the implied conditional distribution $p(y_i|X,\beta,\omega)$?

Notice that λ_i has been marginalized out. This should look familiar.

$$\begin{aligned} p(y_i|X,\beta,\omega) &= \int p(y_i|X,\beta,\omega,\lambda_i)\cdot p(\lambda_i)d\lambda_i \\ &\propto \int \underbrace{(\omega\lambda_i)^{1/2} exp\left[\frac{-\omega\lambda_i}{2}\cdot(y_i - x_i^T\beta)^2\right]}_{p(y_i|X,\beta,\omega,\lambda_i)} \cdot \underbrace{\lambda_i^{h/2-1} exp[-\frac{h}{2}\lambda_i]}_{p(y_i)} d\lambda_i \\ &= \int \lambda_i^{(h+1)/2-1} \exp\left[-\frac{1}{2}\left(\omega(y_i - x_i^T\beta)^2 + h\right)\lambda_i\right] d\lambda_i \\ &\propto \left[\frac{1}{2}\left(\omega(y_i - x_i^T\beta)^2 + h\right)\right]^{-(h+1)/2} \\ &\propto \left[1 + \frac{1}{h}\cdot\frac{(y_i - x_i^T\beta)^2}{\omega^{-1}}\right]^{-(h+1)/2}, \end{aligned} \tag{25}$$

This has the form of the students t-distribution with mean $x_i^T\beta$ and scale $1/\omega$ thus

$$p(y_i|X,\beta,\omega) \sim t_n(x_i^T\beta, 1/\omega)$$

(B) What is the conditional posterior distribution $p(\lambda_i|\mathbf{y},\beta,\omega)$?

$$\begin{aligned} p(\lambda_i|\mathbf{y},B,\omega) &\propto p(\lambda_i, y_i, \beta, \omega) \\ &= p(y_i|\lambda_i,\beta,\omega)\cdot p(\lambda_i,\beta,\omega) \\ &= p(y_i|\lambda_i,\beta,\omega)\cdot p(\lambda_i|\beta,\omega)\cdot p(\beta,\omega) \\ &\propto p(y_i|\lambda_i,\beta,\omega)\cdot p(\lambda_i) \end{aligned} \tag{26}$$

We calculated this in the first part of A, so we may write:

$$\begin{aligned} &\propto \lambda_i^{(h+1)/2-1}\exp\left[-\frac{1}{2}\left(\omega(y_i - x_i^T\beta)^2 + h\right)\lambda_i\right] \\ &\sim \text{Gamma}\left(\frac{h+1}{2}, \frac{h+\omega(y_i - x_i^T\beta)^2}{2}\right) \end{aligned} \tag{27}$$

(C) Combining these results with those from the "Basics" subsection above, code up a Gibbs sampler that repeatedly cycles through sampling the following three sets of conditional distributions.

- $p(\beta|\mathbf{y},\omega,\Lambda)$
- $p(\omega|\mathbf{y},\Lambda)$
- $p(\lambda_i|\mathbf{y},\beta,\omega)$

The first two should follow identically from previous results, except that we are explicitly conditioning on Λ, which is now a random variable rather than a fixed hyperparameter.

If you cycle through these conditional posterior draws a few thousand times, you will build

up a **Markov-chain Monte Carlo (MCMC)** sample
from the joint posterior distribution $p(B, \omega, \Lambda|\mathbf{y})$.

Now use your Gibbs sampler (with at least a few thousand draws) to fit this model to the GDP growth rate data for an appropriate choice of h. Are you happier with the fit? What's going on here (i.e. what makes the model more or less appropriate for the data)? An interesting plot will be the posterior mean of $1/\lambda_i$ for each country.

From our prior results we have,

$$\begin{aligned} p(\beta|\mathbf{y}, \omega, \Lambda) &\sim N(m^*, (\omega K^\star)^{-1}) \\ p(\omega|\mathbf{y}, \Lambda) &\sim \text{Gamma}\left(\frac{d^\star}{2}, \frac{\eta^\star}{2}\right) \\ p(\lambda_i|\mathbf{y}, \beta, \omega) &\sim \text{Gamma}\left(\frac{h+1}{2}, \frac{h + \omega(y_i - x_i^T\beta)^2}{2}\right) \end{aligned} \tag{28}$$

Appendix A

R code