SDS 383D: Homework 2

Giorgio Paulon

February 5, 2017

Problem 1. A simple Gaussian location model

Take a simple Gaussian model with unknown mean and variance:

$$(Y_i|\theta, \sigma^2) \sim N(\theta, \sigma^2), \quad i = 1, \dots, n \tag{1}$$

Let $\boldsymbol{Y}$ be the vector of observations $\boldsymbol{Y} = (Y_1, \dots, Y_n)^T$. Suppose we place conjugate normal and inverse-gamma priors on θ and σ^2, respectively:

$$\begin{aligned}(\theta|\sigma^2) &\sim N(\mu, \tau^2\sigma^2)\\ \sigma^2 &\sim \textit{Inv-Gamma}(d/2, \eta/2)\end{aligned}$$

where $\mu, \tau > 0$, $d > 0$ and $\eta > 0$ are fixed scalar hyperparameters. Note a crucial choice here: the error variance 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.

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 $\omega = 1/\sigma^2$ and $\kappa = 1/\tau^2$ instead.

$$\begin{aligned}(Y_i|\theta, \omega) &\overset{\text{iid}}{\sim} N(\theta, \omega^{-1}), \quad i = 1, \dots, n\\ (\theta|\omega) &\sim N(\mu, (\omega\kappa)^{-1})\\ \omega &\sim \textit{Gamma}(d/2, \eta/2).\end{aligned} \tag{2}$$

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

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

This is often called the normal/gamma prior for (θ, ω) with parameters (μ, κ, d, η), and it's 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).

(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:*

$$p(\theta) \propto \left(1 + \frac{1}{\nu} \cdot \frac{(x-m)^2}{s^2}\right)^{-\frac{\nu+1}{2}}$$

with center m, scale s, and degrees of freedom ν, where you fill in the blank for m, s^2, and ν in terms of the four parameters of the normal-gamma family.

The joint prior distribution is

$$f_{(\theta,\omega)}(\theta, \omega) = f_{(\theta|\omega)}(\theta|\omega) f_\omega(\omega)$$

and the marginal prior for θ is

$$\begin{aligned}f_\theta(\theta) &= \int_\Omega f_{(\theta,\omega)}(\theta, \omega) d\omega\\ &= \int_\Omega \left(\frac{\omega\kappa}{2\pi}\right)^{1/2} e^{-\frac{\omega\kappa}{2}(\theta-\mu)^2} \frac{(\eta/2)^{d/2}}{\Gamma(d/2)} \omega^{d/2-1} e^{-\frac{\eta}{2}\omega} d\omega\\ &\propto \int_\Omega \exp\left\{-\omega\left(\frac{\kappa}{2}(\theta-\mu)^2\right) + \frac{\eta}{2}\right\} \omega^{\frac{d+1}{2}-1} d\omega\end{aligned}$$

By recognizing the kernel of a Gamma distribution, we can write

$$
\begin{aligned}
f_\theta(\theta) &\propto \Gamma\left(\frac{d+1}{2}\right)\left[\frac{\eta}{2}+\frac{\kappa}{2}(\theta-\mu)^2\right]^{-\frac{d+1}{2}} \\
&\propto \left[1+\frac{\kappa}{\eta}(\theta-\mu)^2\right]^{-\frac{d+1}{2}} \\
&\propto \left[1+\frac{1}{d}\cdot\frac{\kappa d}{\eta}(\theta-\mu)^2\right]^{-\frac{d+1}{2}} \\
&\sim t_d\left(\mu;\frac{\eta}{\kappa d}\right)
\end{aligned}
$$

(B) *Assume the normal sampling model in* (1) *and the normal-gamma prior in* (2)*. Calculate the joint posterior density $p(\theta,\omega|\boldsymbol{y})$, 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:*

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

From this form of the posterior, you should able to read off the new updated parameters, by pattern-matching against the functional form in 2. To make the calculations go more easily, you might first show (or recall, from a previous exercise) that the likelihood can be written in the form

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

where $S_y=\sum_{i=1}^n(y_i-\bar{\boldsymbol{y}})^2$ is the sum of squares for the $\boldsymbol{y}$ vector. This expresses the likelihood in terms of the two statistics $\bar{\boldsymbol{y}}$ and S_y, which you may recall from your math-stat course are sufficient statistics for (θ,σ^2). Take care in ignoring constants here: some term that is constant in θ may not be constant in ω, and vice versa.

First of all, let us rewrite the likelihood: the exponent can be expressed as

$$
\begin{aligned}
-\frac{\omega}{2}\sum_{i=1}^n(y_i-\theta)^2 &= -\frac{\omega}{2}\sum_{i=1}^n(y_i-\bar{\boldsymbol{y}}+\bar{\boldsymbol{y}}-\theta)^2 \\
&= -\frac{\omega}{2}\left\{\sum_{i=1}^n(y_i-\bar{\boldsymbol{y}})^2+\sum_{i=1}^n(\bar{\boldsymbol{y}}-\theta)^2+2(\bar{\boldsymbol{y}}-\theta)\sum_{i=1}^n(y_i-\bar{\boldsymbol{y}})\right\} \\
&= -\frac{\omega}{2}\left\{S_y+n(\bar{\boldsymbol{y}}-\theta)^2\right\}.
\end{aligned}
$$

Therefore, the posterior distribution is

$$
\begin{aligned}
p(\theta,\omega|\boldsymbol{y}) &\propto p(\boldsymbol{Y}|\theta,\omega)p(\theta,\omega)\\
&= \prod_{i=1}^{n}\left\{p(Y_i|\theta,\omega)\right\}p(\theta|\omega)p(\omega)\\
&= \left(\frac{\omega}{2\pi}\right)^{\frac{n}{2}}\exp\left\{-\frac{\omega}{2}\sum_{i=1}^{n}(y_i-\theta)^2\right\}\left(\frac{\omega\kappa}{2\pi}\right)^{\frac{1}{2}}\exp\left\{-\frac{\omega\kappa}{2}(\theta-\mu)^2\right\}\frac{\left(\frac{\eta}{2}\right)^{d/2}}{\Gamma\left(\frac{d}{2}\right)}\omega^{\frac{d}{2}-1}\exp\left\{-\frac{\eta}{2}\omega\right\}\\
&\propto \omega^{\frac{n}{2}}\exp\left\{-\frac{\omega}{2}S_y\right\}\exp\left\{-\frac{\omega}{2}n(\theta-\bar{\boldsymbol{y}})^2\right\}\omega^{\frac{1}{2}}\exp\left\{-\frac{\omega\kappa}{2}(\theta-\mu)^2\right\}\omega^{\frac{d}{2}-1}\exp\left\{-\frac{\eta}{2}\omega\right\}\\
&\propto \omega^{\frac{1}{2}}\exp\underbrace{\left\{-\frac{\omega}{2}\left[n(\theta-\bar{\boldsymbol{y}})^2+\kappa(\theta-\mu)^2\right]\right\}}_{\text{(i)}}\omega^{\frac{n+d}{2}-1}\exp\left\{-\omega\left(\frac{\eta}{2}+\frac{S_y}{2}\right)\right\}.
\end{aligned}
$$

We now rewrite (i) using the trick of completing the square. In fact

$$
\begin{aligned}
n(\theta-\bar{\boldsymbol{y}})^2+\kappa(\theta-\mu)^2 &= n\theta^2+n\bar{\boldsymbol{y}}^2-2n\bar{\boldsymbol{y}}\theta+\kappa\theta^2+\kappa\mu^2-2\kappa\mu\theta\\
&= (n+\kappa)\theta^2-2(n\bar{\boldsymbol{y}}+\kappa\mu)\theta+(n\bar{\boldsymbol{y}}^2+\kappa\mu^2)\\
&= (n+\kappa)\left[\theta^2-2\frac{n\bar{\boldsymbol{y}}+\kappa\mu}{n+\kappa}\theta+\frac{n\bar{\boldsymbol{y}}^2+\kappa\mu^2}{n+\kappa}\right]\\
&= (n+\kappa)\left[(\theta-\mu^*)^2+\frac{n\bar{\boldsymbol{y}}^2+\kappa\mu^2}{n+\kappa}-\frac{(n\bar{\boldsymbol{y}}+\kappa\mu)^2}{(n+\kappa)^2}\right]\\
&= (n+\kappa)\left[(\theta-\mu^*)^2+\frac{n\kappa}{(n+\kappa)^2}(\mu-\bar{\boldsymbol{y}})^2\right],
\end{aligned}
$$

where $\mu^*=\dfrac{n\bar{\boldsymbol{y}}+\kappa\mu}{n+\kappa}$.

The joint posterior distribution becomes

$$
\begin{aligned}
p(\theta,\omega|\boldsymbol{y}) &\propto \omega^{\frac{1}{2}}\exp\left\{-\frac{\omega(n+\kappa)}{2}(\theta-\mu^*)^2\right\}\exp\left\{-\frac{\omega n\kappa}{2(n+\kappa)}(\mu-\bar{\boldsymbol{y}})^2\right\}\omega^{\frac{n+d}{2}-1}\exp\left\{-\frac{\omega}{2}(\eta+S_y)\right\}\\
&= \omega^{\frac{n+d+1}{2}-1}\exp\left\{-\omega\frac{(n+\kappa)(\theta-\mu^*)^2}{2}\right\}\exp\left\{-\frac{\omega}{2}\left[\eta+S_y+\frac{n\kappa}{n+\kappa}(\mu-\bar{\boldsymbol{y}})^2\right]\right\}.
\end{aligned}
$$

Therefore $d^*=n+d$, $\quad\kappa^*=n+\kappa$, $\quad\eta^*=\eta+S_y+\dfrac{n\kappa}{n+\kappa}(\mu-\bar{\boldsymbol{y}})^2$.

(C) *From the joint posterior you just derived, what is the conditional posterior distribution $p(\theta|\boldsymbol{y},\omega)$? Note: this should require no calculation - you should just be able to read it off directly from the joint distribution.*

Let us remark that the joint posterior factors in the two terms

$$
p(\theta,\omega|\boldsymbol{y}) \propto \underbrace{\omega^{\frac{1}{2}}\exp\left\{-\frac{\omega\kappa^*}{2}(\theta-\mu^*)^2\right\}}_{\text{(ii)}}\underbrace{\omega^{\frac{d^*}{2}-1}\exp\left\{-\frac{\omega}{2}\eta^*\right\}}_{\text{(iii)}}.
$$

Therefore, (ii) represents the distribution

$$
\theta|\omega,\boldsymbol{y}\sim N\left(\mu^*,(\omega\kappa^*)^{-1}\right).
$$

(D) *From the joint posterior you calculated in (B), what is the marginal posterior distribution $p(\omega|\boldsymbol{y})$? Unlike the previous question, this one doesn't come 100% for free - you have to integrate over θ. But it shouldn't be too hard, since you can ignore constants not depending on ω in calculating this integral.*

Since we factored the joint posterior density, we know that (iii) represents the distribution

$$\omega|\boldsymbol{y} \sim \text{Gamma}\left(\frac{d^*}{2}, \frac{\eta^*}{2}\right).$$

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

$$p(\theta|\boldsymbol{y}) \propto \left(1 + \frac{1}{\nu} \cdot \frac{(x-m)^2}{s^2}\right)^{-\frac{\nu+1}{2}}$$

with center m^*, scale s^*, and degrees of freedom ν^* (where you fill in the blank for m^*, s^{*2}, and ν^*).

Since

$$\theta|\omega, \boldsymbol{y} \sim N\left(\mu^*, (\omega\kappa^*)^{-1}\right)$$
$$\omega|\boldsymbol{y} \sim \text{Gamma}\left(\frac{d^*}{2}, \frac{\eta^*}{2}\right),$$

in analogy with part (A), we find that

$$\theta|\boldsymbol{y} \sim t_{d^*}\left(\mu^*, \frac{\eta^*}{\kappa^* d^*}\right)$$

(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. Remember that a valid probability distribution must integrate to 1 (or something finite, so that it can normalized to integrate to 1) over its domain.*

As the κ, d and η approach 0, the priors

$$\theta|\omega \sim N\left(\mu, (\omega\kappa)^{-1}\right) \xrightarrow{d} 0$$
$$\omega \sim \text{Gamma}\left(\frac{d}{2}, \frac{\eta}{2}\right) \xrightarrow{d} 0$$

are not valid distributions.

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

As the κ, d and η approach 0, the posteriors

$$\theta|\omega, \boldsymbol{y} \sim N\left(\mu^*, (\omega\kappa^*)^{-1}\right) \xrightarrow{d} N\left(\bar{\boldsymbol{y}}; (\omega n)^{-1}\right)$$
$$\omega|\boldsymbol{y} \sim \text{Gamma}\left(\frac{d^*}{2}, \frac{\eta^*}{2}\right) \xrightarrow{d} \text{Gamma}\left(\frac{n}{2}; S_y\right)$$

are valid distributions.

(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 (E), 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.

A Bayesian credible interval for $\theta|\boldsymbol{y}$ is

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

where $m = \mu^*$, $s = \sqrt{\dfrac{\eta^*}{\kappa^* d^*}}$. As the prior parameters κ, d and η approach zero, we obtain $m \to \bar{\boldsymbol{y}}$ and $s \to \dfrac{\sqrt{S_y}}{n}$.

Therefore the Bayesian credible interval converges to the frequentist confidence interval, that is

$$\theta \in \bar{\boldsymbol{y}} \pm t^* \frac{\sqrt{S_y}}{n}.$$

Problem 2. The conjugate Gaussian linear model

Now consider the Gaussian linear model

$$(\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2) \sim N(X\boldsymbol{\beta}, (\omega\Lambda)^{-1}),$$

where $\boldsymbol{Y}$ is an n vector of responses, X is an $n \times p$ matrix of features, and $\omega = 1/\sigma^2$ is the error precision, and Λ is some known matrix. A typical setup would be $\Lambda = \mathcal{I}$, the $n \times n$ 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 $\boldsymbol{Y}$ 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 that the first entry in $\boldsymbol{\beta}$ is actually the intercept. We'll again work in terms of the precision $\omega = 1/\sigma^2$, and consider a normal–gamma prior for $\boldsymbol{\beta}$:

$$(\boldsymbol{\beta}|\omega) \sim N(\boldsymbol{m}, (\omega K)^{-1})$$
$$\omega \sim \textit{Gamma}\left(\frac{d}{2}, \frac{\eta}{2}\right)$$

Here K is a $p \times p$ precision matrix in the multivariate normal prior for $\boldsymbol{\beta}$, which we assume 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.

Basics

(A) *Derive the conditional posterior* $p(\boldsymbol{\beta}|\boldsymbol{y}, \omega)$.

(B) *Derive the marginal posterior* $p(\omega|\boldsymbol{y})$.

(C) *Putting these together, derive the marginal posterior* $p(\boldsymbol{\beta}|\boldsymbol{y})$.

(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* $\Lambda = \mathcal{I}$ *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?*

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:

Appendix A

R code