SDS 383D: Homework 2

Giorgio Paulon

February 10, 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

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

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

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

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

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

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

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

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

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 $p(\boldsymbol{\beta}|\boldsymbol{y}, \omega)$.*

Let us start by finding the joint posterior density

$$p(\boldsymbol{\beta}, \omega|\boldsymbol{y}) \propto \omega^{\frac{n}{2}} \exp\left\{-\frac{\omega}{2}(\boldsymbol{Y} - X\boldsymbol{\beta})^T\Lambda(\boldsymbol{Y} - X\boldsymbol{\beta})\right\} \omega^{\frac{p}{2}} \exp\left\{-\frac{\omega}{2}(\boldsymbol{\beta} - \boldsymbol{m})^T K(\boldsymbol{\beta} - \boldsymbol{m})\right\} \omega^{\frac{d}{2}-1} \exp\left\{-\omega\frac{\eta}{2}\right\}.$$

We can rewrite the first exponential term, analogously with the univariate case, in terms of the MLE $\hat{\boldsymbol{\beta}} = (X^T\Lambda X)^{-1}X^T\Lambda\boldsymbol{Y}$. In fact,

$$\begin{aligned}
&-\frac{\omega}{2}\left[(\boldsymbol{Y} - X\boldsymbol{\beta})^T\Lambda(\boldsymbol{Y} - X\boldsymbol{\beta})\right] \\
=&-\frac{\omega}{2}\left\{(\boldsymbol{Y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})^T\Lambda(\boldsymbol{Y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})\right\} \\
=&-\frac{\omega}{2}\left\{(\boldsymbol{Y} - X\hat{\boldsymbol{\beta}})^T\Lambda(\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T\Lambda X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + 2(\boldsymbol{Y} - X\hat{\boldsymbol{\beta}})^T\Lambda X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\} \\
=&-\frac{\omega}{2}\left\{S_y + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T\Lambda X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\},
\end{aligned}$$

where the cross product term cancels because $\boldsymbol{Y}^T\Lambda X = \hat{\boldsymbol{\beta}}^T X^T\Lambda X$ from the MLE equation.

Therefore,

$$p(\boldsymbol{\beta}, \omega|\boldsymbol{y}) \propto \omega^{\frac{n}{2}} \exp\left\{-\frac{\omega}{2}\left[S_y + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T\Lambda X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right]\right\} \exp\left\{-\frac{\omega}{2}(\boldsymbol{\beta} - \boldsymbol{m})^T K(\boldsymbol{\beta} - \boldsymbol{m})\right\} \omega^{\frac{d+p}{2}-1} \exp\left\{-\omega\frac{\eta}{2}\right\}.$$

We can complete the square by grouping the terms containing $\boldsymbol{\beta}$, that is

$$\begin{aligned}
&(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T\Lambda X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{m})^T K(\boldsymbol{\beta} - \boldsymbol{m}) \\
&= \boldsymbol{\beta}^T X^T\Lambda X\boldsymbol{\beta} + \boldsymbol{\beta}^T K\boldsymbol{\beta} - 2\hat{\boldsymbol{\beta}}^T X^T\Lambda X\boldsymbol{\beta} - 2\boldsymbol{m}^T K\boldsymbol{\beta} + \hat{\boldsymbol{\beta}}^T X^T\Lambda X\hat{\boldsymbol{\beta}} + \boldsymbol{m}^T K\boldsymbol{m} \\
&= \boldsymbol{\beta}^T(X^T\Lambda X + K)\boldsymbol{\beta} - 2(\hat{\boldsymbol{\beta}}^T X^T\Lambda X + \boldsymbol{m}^T K)\boldsymbol{\beta} + \hat{\boldsymbol{\beta}}^T X^T\Lambda X\hat{\boldsymbol{\beta}} + \boldsymbol{m}^T K\boldsymbol{m}.
\end{aligned}$$

By completing the square, we get

$$(\boldsymbol{\beta} - \boldsymbol{m}^*)^T K^*(\boldsymbol{\beta} - \boldsymbol{m}^*) + r$$

where

$$\begin{aligned}
\boldsymbol{m}^* &= K^{*^{-1}}(K\boldsymbol{m} + X^T\Lambda X\hat{\boldsymbol{\beta}}) \\
K^* &= X^T\Lambda X + K \\
r &= \hat{\boldsymbol{\beta}}^T X^T\Lambda X\hat{\boldsymbol{\beta}} + \boldsymbol{m}^T K\boldsymbol{m} - (K\boldsymbol{m} + X^T\Lambda X\hat{\boldsymbol{\beta}})^T K^{*^{-1}}(K\boldsymbol{m} + X^T\Lambda X\hat{\boldsymbol{\beta}}) \\
&= \hat{\boldsymbol{\beta}}^T X^T\Lambda X\hat{\boldsymbol{\beta}} + \boldsymbol{m}^T K\boldsymbol{m} - \boldsymbol{m}^{*^T} K^*\boldsymbol{m}^*.
\end{aligned}$$

Thus, the joint posterior density is

$$p(\boldsymbol{\beta}, \omega|\boldsymbol{y}) \propto \underbrace{\omega^{\frac{p}{2}} \exp\left\{-\frac{\omega}{2}(\boldsymbol{\beta} - \boldsymbol{m}^*)^T K^*(\boldsymbol{\beta} - \boldsymbol{m}^*)\right\}}_{\text{(i): } p(\boldsymbol{\beta}|\omega, \boldsymbol{y}) = N(\boldsymbol{m}^*, (\omega K^*)^{-1})} \underbrace{\omega^{\frac{d+n}{2}-1} \exp\left\{-\omega\left(\frac{\eta + S_y + r}{2}\right)\right\}}_{\text{(ii): } p(\omega|\boldsymbol{y}) = \text{Gamma}(\frac{d^*}{2}, \frac{\eta^*}{2})},$$

where $d^* = d + n$ and $\eta^* = \eta + S_y + r$.

Since the joint posterior factors in the two terms, (i) represents the distribution

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

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

Since the joint posterior factors in the two terms, (ii) represents the distribution

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

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

As usual, the marginal posterior $p(\boldsymbol{\beta}|\boldsymbol{y})$ can be obtained by integrating out ω from the joint posterior density. Mathematically speaking,

$$\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{y}) &= \int p(\boldsymbol{\beta}|\omega, \boldsymbol{y}) p(\omega|\boldsymbol{y}) d\omega \\
&= \int \left(\frac{\omega}{2\pi}\right)^{\frac{p}{2}} |K^*|^{\frac{1}{2}} \exp\left\{-\frac{\omega}{2}(\boldsymbol{\beta}-\boldsymbol{m}^*)^T K^*(\boldsymbol{\beta}-\boldsymbol{m}^*)\right\} \frac{\left(\frac{\eta^*}{2}\right)^{\frac{d^*}{2}}}{\Gamma\left(\frac{d^*}{2}\right)} \omega^{\frac{d^*}{2}-1} \exp\left\{-\frac{\eta^*}{2}\omega\right\} d\omega \\
&= \left(\frac{1}{2\pi}\right)^{\frac{p}{2}} |K^*|^{\frac{1}{2}} \frac{\left(\frac{\eta^*}{2}\right)^{\frac{d^*}{2}}}{\Gamma\left(\frac{d^*}{2}\right)} \int \underbrace{\exp\left\{-\frac{\omega}{2}(\boldsymbol{\beta}-\boldsymbol{m}^*)^T K^*(\boldsymbol{\beta}-\boldsymbol{m}^*)\right\} \omega^{\frac{d^*+p}{2}-1} \exp\left\{-\frac{\eta^*}{2}\omega\right\}}_{\propto \text{Gamma}\left(\frac{d^*+p}{2}, \frac{\eta^*}{2} + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{m}^*)^T K^*(\boldsymbol{\beta}-\boldsymbol{m}^*)\right)} d\omega \\
&= \left(\frac{1}{2\pi}\right)^{\frac{p}{2}} |K^*|^{\frac{1}{2}} \frac{\left(\frac{\eta^*}{2}\right)^{\frac{d^*}{2}}}{\Gamma\left(\frac{d^*}{2}\right)} \frac{\Gamma\left(\frac{d^*+p}{2}\right)}{\left[\frac{\eta^*}{2} + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{m}^*)^T K^*(\boldsymbol{\beta}-\boldsymbol{m}^*)\right]^{\frac{d^*+p}{2}}} \\
&= \frac{\Gamma\left(\frac{d^*+p}{2}\right)}{\Gamma\left(\frac{d^*}{2}\right)} |K^*|^{\frac{1}{2}} \left(\frac{1}{\pi\eta^*}\right)^{\frac{p}{2}} \left[1 + \frac{1}{\eta^*}(\boldsymbol{\beta}-\boldsymbol{m}^*)^T K^*(\boldsymbol{\beta}-\boldsymbol{m}^*)\right]^{-\frac{d^*+p}{2}} \\
&= \frac{\Gamma\left(\frac{d^*+p}{2}\right)}{\Gamma\left(\frac{d^*}{2}\right)} |K^*|^{\frac{1}{2}} \left(\frac{1}{\pi d^* \cdot \frac{\eta^*}{d^*}}\right)^{\frac{p}{2}} \left[1 + \frac{1}{d^*} \cdot \frac{1}{\frac{\eta^*}{d^*}}(\boldsymbol{\beta}-\boldsymbol{m}^*)^T K^*(\boldsymbol{\beta}-\boldsymbol{m}^*)\right]^{-\frac{d^*+p}{2}}.
\end{aligned}$$

Therefore

$$(\boldsymbol{\beta}|\boldsymbol{y}) \sim t_{d^*}\left(\boldsymbol{m}^*; \frac{\eta^*}{d^*} {K^*}^{-1}\right),$$

where we denote with t the p-multivariate Student's t distribution.

(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?*

We apply a bayesian regression analysis to the data in "gdpgrowth.csv". The regression tries to explain the GDP growth rate as a level of defense spending.

We specified the hyperparameters in such a way that the prior specification is vague, i.e.

$$\begin{aligned}
\Lambda &= \text{diag}\,\{1, \dots, 1\} \\
\boldsymbol{m} &= (0, 0)^T \\
K &= \text{diag}\{0.01, 0.01\} \\
d &= 0.02 \\
\eta &= 0.02.
\end{aligned}$$

In this way, $E[\omega] = 1$ and $\text{Var}(\omega) = 100$ (vague precision parameter). Moreover, the precision matrices Λ and K have small values.

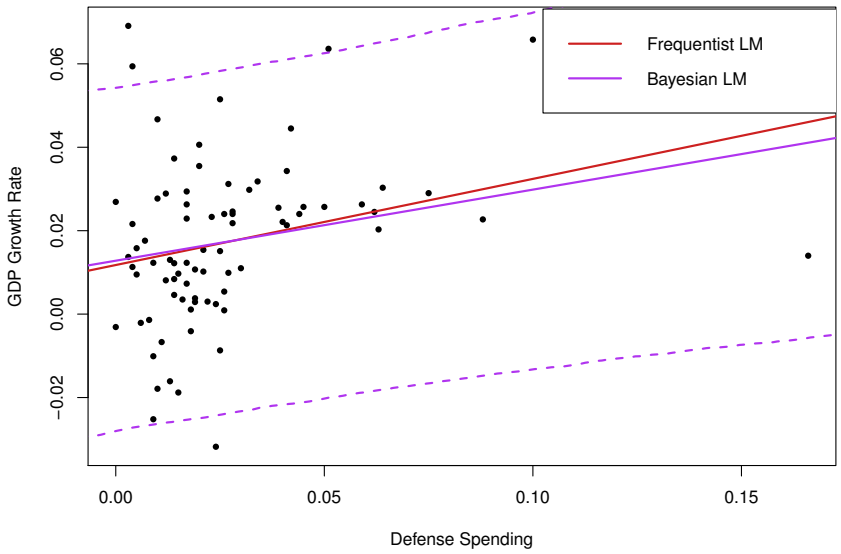

Figure 1: Frequentist and Bayesian regression lines, along with the 95% predictive intervals.

In Figure 1, both the frequentist and the bayesian regression lines are displayed. As one can see, in the bayesian framework there is a shrinkage effect towards the prior (in this case of $\boldsymbol{\beta}$, which is $(0, 0)^T$). In Table 1 the frequentist estimate $\hat{\boldsymbol{\beta}}$ and the bayesian posterior mean are reported.

	Frequentist	Bayesian
$\hat{\beta}_0$	0.011768	0.012674
$\hat{\beta}_1$	0.206506	0.172013

Table 1: Frequentist estimate for the regression coefficients, along with the Bayesian posterior mean.

Although both of the coefficients are significant, the model seems to be misspecified because of the presence of outliers in the left part of the X domain. However, it is still possible to perform a posterior inference for the parameters. In particular, in Figure 2, it is possible to

appreciate how much the data "shift" the prior distribution to a more precise and localized posterior distribution.

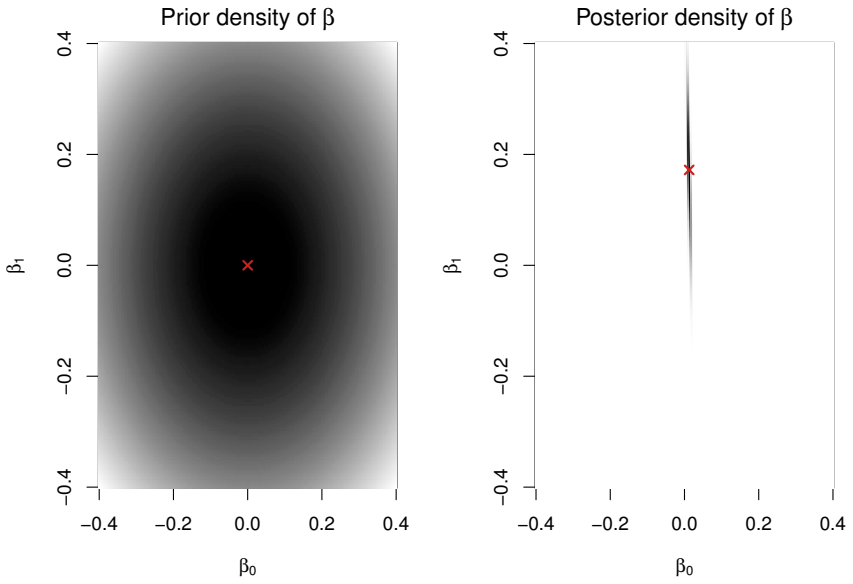

Figure 2: Joint prior and joint posterior densities of the parameters $\boldsymbol{\beta}$.

Moreover, we can also compute the 95% predictive bounds for the regression line. In fact, since we know the posterior distribution of the parameters in closed form, we can obtain a i.i.d. sample of regression lines evaluated on a grid of x values from it. Subsequently, we can consider their quantiles as the credible bounds for the linear regression line. Mathematically, speaking,

$$\begin{aligned} p(y^*|\boldsymbol{y}) &= \int p(y^*|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}) p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) d\boldsymbol{\beta} d\sigma^2 \\ &= \int p(y^*|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) d\boldsymbol{\beta} d\sigma^2 \\ &\approx \sum_{g=1}^{G} p(y^*|\boldsymbol{\beta}^{(g)}, \sigma^{2(g)}), \end{aligned}$$

where $\left(\boldsymbol{\beta}^{(g)}, \sigma^{2(g)}\right)_{g=1}^{G}$ is a sample from the posterior distribution $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})$.

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:

$$
\begin{aligned}
&(\boldsymbol{Y}|\boldsymbol{\beta},\omega,\Lambda) \sim N(X\boldsymbol{\beta},(\omega\Lambda)^{-1}) \\
&\Lambda = \textit{diag}(\lambda_1,\dots,\lambda_n) \\
&\lambda_i \overset{\text{iid}}{\sim} \textit{Gamma}\left(\frac{h}{2},\frac{h}{2}\right) \\
&(\boldsymbol{\beta}|\omega) \sim N(\boldsymbol{m},(\omega K)^{-1}) \\
&\omega \sim \textit{Gamma}\left(\frac{d}{2},\frac{\eta}{2}\right).
\end{aligned}
$$

where h is a fixed hyperparameter.

(A) *Under this model, what is the implied conditional distribution $p(y_i|X,\boldsymbol{\beta},\omega)$? Notice that λ_i has been marginalized out. This should look familiar.*

The conditional distribution $p(y_i|X,\boldsymbol{\beta},\omega)$ can be found by marginalizing λ_i out, that is

$$
\begin{aligned}
p(y_i|X,\boldsymbol{\beta},\omega) &= \int p(y_i|X,\boldsymbol{\beta},\omega,\lambda_i)p(\lambda_i|X,\boldsymbol{\beta},\omega)p(\boldsymbol{\beta}|\omega)p(\omega)d\lambda_i \\
&\propto \int p(y_i|X,\boldsymbol{\beta},\omega,\lambda_i)p(\lambda_i)d\lambda_i \\
&\propto \int \underbrace{(\omega\lambda_i)^{\frac{1}{2}}\exp\left\{-\frac{\omega\lambda_i}{2}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2\right\}\lambda_i^{\frac{h}{2}-1}\exp\left\{-\frac{h}{2}\lambda_i\right\}}_{\text{Gamma}\left(\frac{h+1}{2},\frac{h}{2}+\frac{\omega}{2}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2\right)} d\lambda_i \\
&\propto \frac{\Gamma\left(\frac{h+1}{2}\right)}{\left[\frac{h}{2}+\frac{1}{2}\omega(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2\right]^{\frac{h+1}{2}}} \\
&\propto \left[1+\frac{\omega}{h}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2\right]^{-\frac{h+1}{2}}.
\end{aligned}
$$

Therefore we recognize that

$$
p(y_i|X,\boldsymbol{\beta},\omega) \sim t_h\left(\boldsymbol{x}_i^T\boldsymbol{\beta},\frac{1}{\omega}\right).
$$

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

We can find the conditional posterior distribution

$$
\begin{aligned}
p(\lambda_i|\boldsymbol{y},\boldsymbol{\beta},\omega) &\propto p(\boldsymbol{y}|\boldsymbol{\beta},\omega,\lambda_i)p(\lambda_i|\boldsymbol{\beta},\omega)p(\boldsymbol{\beta},\omega) \\
&\propto p(y_i|\boldsymbol{\beta},\omega,\lambda_i)p(\lambda_i) \\
&= \left(\frac{\omega\lambda_i}{2\pi}\right)^{\frac{1}{2}}\exp\left\{-\frac{\omega\lambda_i}{2}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2\right\}\frac{\left(\frac{h}{2}\right)^{\frac{h}{2}}}{\Gamma\left(\frac{h}{2}\right)}\lambda_i^{\frac{h}{2}-1}\exp\left\{-\frac{h}{2}\lambda_i\right\} \\
&\propto \lambda_i^{\frac{h+1}{2}-1}\exp\left\{-\lambda_i\left[\frac{\omega}{2}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2+\frac{h}{2}\right]\right\}.
\end{aligned}
$$

Therefore we notice that

$$
p(\lambda_i|\boldsymbol{y},\boldsymbol{\beta},\omega) \sim \text{Gamma}\left(\frac{h+1}{2},\frac{1}{2}\left[\omega(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2+h\right]\right).
$$

(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:*

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

The first two should follow identically from your 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 $p(\boldsymbol{\beta}, \omega, \Lambda|\boldsymbol{y})$. Why this technique works for getting posterior draws is the subject of a different course, but hopefully it is reasonably intuitive. 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)?

The full conditionals are the following:

- $p(\boldsymbol{\beta}|\boldsymbol{y}, \omega, \Lambda) = N_p(\boldsymbol{m}^*, K^*)$, where

$$\begin{aligned} \boldsymbol{m}^* &= {K^*}^{-1}(K\boldsymbol{m} + X^T\Lambda X\hat{\boldsymbol{\beta}}) \\ K^* &= K + X^T\Lambda X \\ \hat{\boldsymbol{\beta}} &= (X^T\Lambda X)^{-1}X^T\Lambda\boldsymbol{y}. \end{aligned}$$

- $p(\omega|\boldsymbol{y}, \Lambda) = \text{Gamma}\left(\frac{d^*}{2}, \frac{\eta^*}{2}\right)$, where

$$\begin{aligned} d^* &= n + d \\ \eta^* &= \eta + S_y + r \\ S_y &= (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) \\ r &= (\boldsymbol{m} - \hat{\boldsymbol{\beta}})^T(K^{-1} + (X^T\Lambda X)^{-1})^{-1}(\boldsymbol{m} - \hat{\boldsymbol{\beta}}) \end{aligned}$$

- $p(\lambda_i|\boldsymbol{y}, \boldsymbol{\beta}, \omega) = \text{Gamma}\left(\frac{h+1}{2}, \frac{1}{2}\left[\omega(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + h\right]\right)$.

We can iteratively sample from these full conditionals in order to obtain chains from the joint posterior distribution. For this simulation, we run the algorithm for $12,000$ iterations, while the first 2000 iterations were discarded and we use a thinning of 2 to reduce the autocorrelation of the Markov chain. The final sample size is then 5000.

A simple check of the chains relative to the $\boldsymbol{\beta}$ parameters show that the Gibbs sampler has converged, as one can see in Figure 3.

The posterior estimates of this model should be more robust with respect to the previous ones. In fact, since there are several outliers, those observations will be associated to smaller

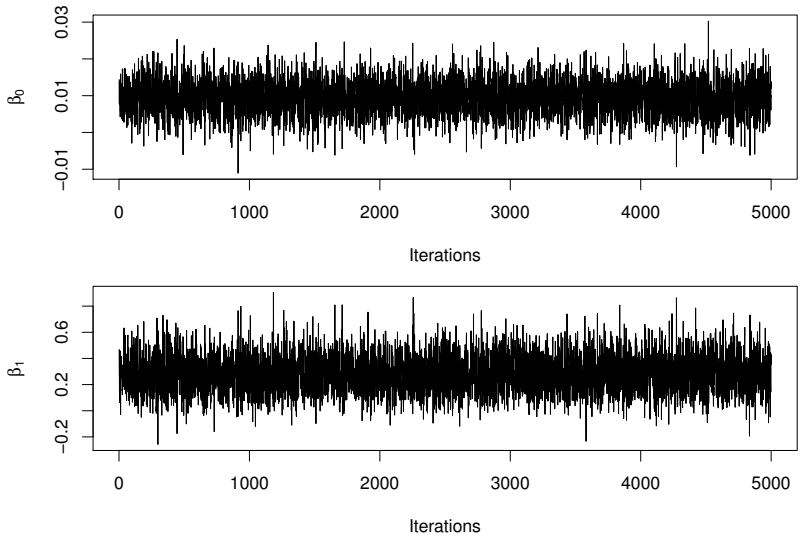

Figure 3: Traceplots of the regression parameters $\boldsymbol{\beta}$.

individual precisions λ_i, thus affecting less severely the regression line. Therefore, this problem can be reinterpreted as an outlier detection. By examining the posterior estimates of the different λ_i's, we remark that some of them are much smaller than the others. By selecting the 5 observations with lower precision and displaying them on the scatterplot of the data, we can recognize these as outliers.

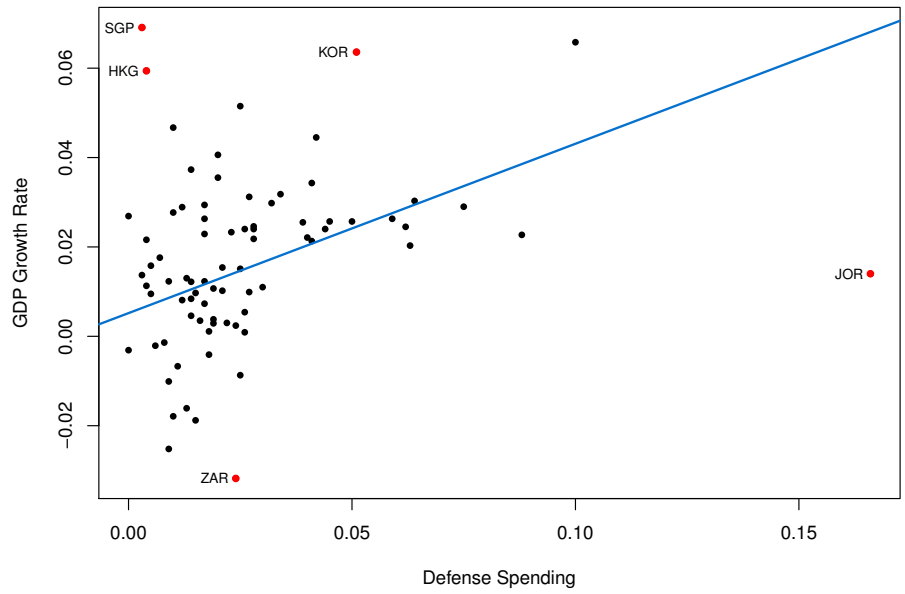

Figure 4: Outlier detection for the heavy tailed Bayesian regression line.

In Figure 4, the outliers are shown in red. Let us remark that these observations are exactly the ones that lie at a further distance from the regression line.

We can also show, in Figure 5, the marginal posterior densities for the $\boldsymbol{\beta}$ parameters. The point estimate are the posterior medians, and we can also compute the 95% posterior credible intervals.

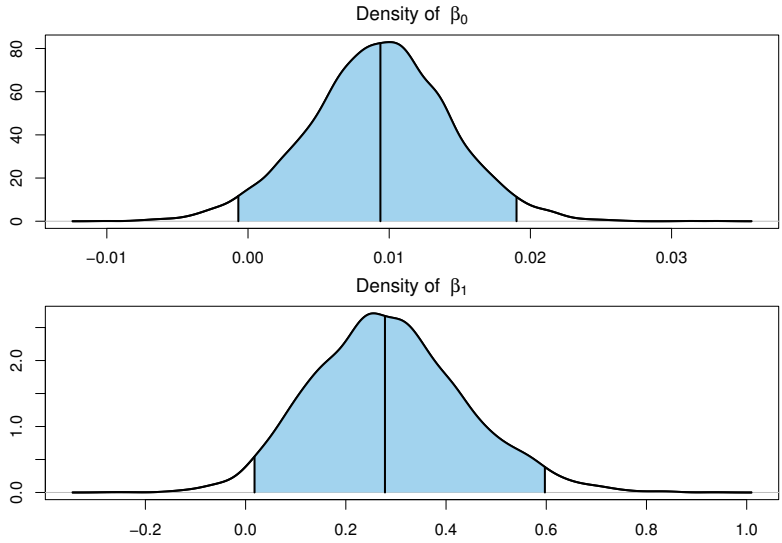

Figure 5: Marginal posterior distributions of the regression parameters $\boldsymbol{\beta}$. The kernel density estimate is shown in black solid line, along with the posterior median and the posterior 95% credible intervals.

Appendix A

R code

```r
library(mvtnorm)

##########################################
### THE CONJUGATE GAUSSIAN LINEAR MODEL ###
##########################################
gdp <- read.csv(file = 'SDS383D-master/data/gdpgrowth.csv')

n <- dim(gdp)[1]
X <- cbind(rep(1, n), gdp$DEF60)
y <- gdp$GR6096
p <- dim(X)[2] - 1

# FREQUENTIST LINEAR REGRESSION
lm1 = lm(y ~ X-1)
summary(lm1)

par(mar=c(4,4,2,2))
plot(X[,2], y, pch = 16, cex = 0.8, asp = 1, xlab = 'Defense Spending', ylab = 'GDP Growth Rate')
abline(a = lm1$coefficients[1], b = lm1$coefficients[2], lwd = 2, col = 'firebrick3')

# BAYESIAN LINEAR REGRESSION
# Y | beta, omega \sim N(X %*% beta; (omega * Lambda^{-1}))
# beta | omega \sim N(m; (omega * K^{-1}))
# omega \sim Gamma(d/2; eta/2)

# Hyperparameters
Lambda <- diag(rep(1, n))
m <- rep(0, p + 1)
K <- diag(c(0.01, 0.01))
d <- 0.02
eta <- 0.02


# Compute the MLE
XtLambdaX <- t(X) %*% Lambda %*% X
beta.hat <- solve(XtLambdaX) %*% crossprod(X, Lambda) %*% y

# Update the hyperparameters for the posterior distribution
K.new <- K + XtLambdaX
m.new <- solve(K.new) %*% (K %*% m + XtLambdaX %*% beta.hat)
```

```r
d.new <- d + n
S <- t(y - X %*% beta.hat) %*% Lambda %*% (y - X %*% beta.hat)
r <- t(m - beta.hat) %*% solve(solve(K) + solve(XtLambdaX)) %*% (m - beta.hat)
# r <- t(beta.hat) %*% XtLambdaX %*% beta.hat + t(m) %*% K %*% m - t(m.new) %*% K.new %*% m.new
eta.new <- as.numeric(eta + S + r)

# Plot the regression line: the model is conjugate so we already have the MAP estimates
par(mar=c(4,4,2,2))
plot(X[,2], y, pch = 16, cex = 0.8, asp = 1, xlab = 'Defense Spending', ylab = 'GDP Growth Rate')
abline(a = lm1$coefficients[1], b = lm1$coefficients[2], lwd = 2, col = 'firebrick3')
abline(a = m.new[1], b = m.new[2], lwd = 2, col = 'dodgerblue3')
legend('topright', legend = c('Frequentist LM','Bayesian LM'), lwd = 2, lty = 1, col = c('
    firebrick3', 'dodgerblue3'))


# If we did not have a closed form solution for the posterior, we should have integrated
# the MCMC sample.
# Grid on which we evaluate the bayesian regression line, that is the line
# y = E[beta[1] | data] + E[beta[2] | data] * x
samp.size <- 5000
x.gr <- seq(min(X[,p+1]) - 0.1*diff(range(X[,p+1])), max(X[,p+1]) + 0.1*diff(range(X[,p+1])),
    length.out = 200)

pred <- matrix(ncol = 200, nrow = samp.size)
beta.post <- rmvt(samp.size, m.new, sigma = (eta.new/d.new) * solve(K.new), df = d.new)
sig.post <- 1/rgamma(samp.size, d.new/2, eta.new/2)

for (i in 1:samp.size){
  pred[i,] <- beta.post[i,1] + beta.post[i,2] * x.gr + rnorm(1, mean = 0, sd = sqrt(sig.post))
}

# Draw the average regression line
predit <- apply(pred, 2, mean)
lines(x.gr, predit, type="l", col = 'darkorchid2', lwd=2)
predit.quant <- apply(pred, 2, quantile, prob=c(0.05,0.95))
lines(x.gr,predit.quant[1,], col = 'darkorchid2', lwd = 2,lty = 2)
lines(x.gr,predit.quant[2,], col = 'darkorchid2', lwd = 2,lty = 2)

# Contour plot for the marginals of the beta parameters
xgrid <- seq(-0.4, 0.4, length.out = 200)
ygrid <- seq(-0.4, 0.4, length.out = 200)
z.prior <- matrix(nrow = length(xgrid), ncol = length(ygrid))
z.post <- matrix(nrow = length(xgrid), ncol = length(ygrid))
for (i in 1:length(xgrid)){
  for (j in 1:length(ygrid)){
    z.prior[i,j] <- dmvt(c(xgrid[i], ygrid[j]), m, sigma = (eta/d) * solve(K), df = d)
    z.post[i,j] <- dmvt(c(xgrid[i], ygrid[j]), m.new, sigma = (eta.new/d.new) * solve(K.new), df
        = d.new)
  }
}

par(mfrow=c(1,2), mar=c(4,4,2,2), cex = 1.1)
grays <- gray((200:0)/200)
image(xgrid, ygrid, exp(z.prior), col = grays, xlab = bquote(beta[0]), ylab = bquote(beta[1]),
    main = bquote("Prior density of"~beta))
points(m[1], m[2], pch = 4, col = 'firebrick3', lwd = 2)
```

```
93 image(xgrid, ygrid, exp(z.post), col = grays, xlab = bquote(beta[0]), ylab = bquote(beta[1]),
       main = bquote("Posterior density of"~beta))
94 points(m.new[1], m.new[2], pch = 4, col = 'firebrick3', lwd = 2)
```

Listing A.1: Code to perform a conjugte linear regression in the Bayesian setting.