SDS 383D: Exercise 2

Diego Garcia-Olano

February 5, 2017

Problem 1. Bayes and the Gaussian linear model

A simple Gaussian location model

Take a simple Gaussian model with unknown mean and variance:

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

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:

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

$$\sigma^2 \sim Inv - Gamma(\frac{d}{2}, \frac{\eta}{2})$$

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:

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

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!

Appendix A

R code