SDS 383D Exercises 3: Bayes and the Gaussian linear model

Tianqi Chen (tc34927)

1 A simple Gaussian location model

Take a simple Gaussian model with unknown mean and variance:

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

Let $\mathbf{y}$ be the vector of observations $\mathbf{y} = (y_1, \ldots, y_n)^T$.

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

$$\begin{aligned}(\theta \mid \sigma^2) &\sim N(\mu, \tau^2\sigma^2) \\ \sigma^2 &\sim \text{Inv-Gamma}\left(\frac{d}{2}, \frac{\eta}{2}\right)\,,\end{aligned}$$

where μ, $\tau > 0$, $d > 0$ 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. 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

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

  if d is large.[1]

[1] This expression is only valid if $d > 2$.

What is meant by the phrases "prior sample size" and "prior sum of squares"? Well, 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 the sample size and the sums of squares. The prior here is designed to mimic the likelihood function for σ^2 that you'd get if you hallucinated 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 $\omega = 1/\sigma^2$ and $\kappa = 1/\tau^2$ instead.

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

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

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

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.) Note: you can obviously write this joint density in Equation 2 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.[2]

(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. Note: you did a problem just like this on a previous exercise! This shouldn't be a lengthy re-derivation.

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

$$p(\theta, \omega \mid \mathbf{y}) \propto \omega^{(d^\star+1)/2-1} \exp\left\{-\omega \cdot \frac{\kappa^\star(\theta-\mu^\star)^2}{2}\right\} \cdot \exp\left\{-\omega \cdot \frac{\eta^\star}{2}\right\} \tag{3}$$

From this form of the posterior, you should able to read off the new updated parameters, by pattern-matching against the functional form in Equation 2:

- $\mu \longrightarrow \mu^\star = ?$

[2]The term "kernel" is heavily overloaded in statistics. See Wikipedia for the sense in which I mean the term.

- $\kappa \longrightarrow \kappa^\star = ?$
- $d \longrightarrow d^\star = ?$
- $\eta \longrightarrow \eta^\star = ?$

You may notice that my parameterization of the normal-gamma in Equation 2 differs from, say, the one you might find in textbooks or on websites. I've chosen this parameterization in order to make these four updates for the parameters, above, as simple-looking and intuitive as possible.

Tip: this one is a bit of an algebra slog, with a lot of completing the square, collecting common terms, and cancelling positives with negatives. For example, 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

$$p(\mathbf{y} \mid \theta, \omega) \propto \omega^{n/2} \exp\left\{-\omega \cdot \left(\frac{S_y + n(\bar{y} - \theta)^2}{2}\right)\right\},$$

where $S_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the sum of squares for the $\mathbf{y}$ vector. This expresses the likelihood in terms of the two statistics $\bar{y}$ 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. You're focusing on the joint posterior, so you can't ignore anything that has a θ or ω in it.

(C) From the joint posterior you just derived, what is the conditional posterior distribution $p(\theta \mid \mathbf{y}, \omega)$? Note: this should require no calculation—you should just be able to 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.

(D) From the joint posterior you calculated in (A), what is the marginal posterior distribution $p(\omega \mid \mathbf{y})$? 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.

(E) Show that the marginal posterior $p(\theta \mid \mathbf{y})$ takes the form of a centered, scaled t distribution and 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 posterior rather than the prior.

(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.)

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

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

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

where m and s are the posterior center and scale parameters from (E), and $t^\star$ 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.

Answer

The joint prior distribution is:

$$\begin{aligned} p(\theta,\omega) &= p(\theta|\omega)p(\omega) \\ &\propto (\kappa\omega)^{1/2} exp(-\kappa\omega(\theta-\mu)^2/2)\omega^{d/2-1} exp(-\eta\omega/2) \\ &\propto \omega^{(d+1)/2-1} exp(-(\kappa(\theta-\mu)^2+\eta)\omega/2) \end{aligned}$$

(A) Integrate out ω, we have

$$\begin{aligned} p(\theta) &\propto \int_{\mathbb{R}} \omega^{(d+1)/2-1} exp(-(\kappa(\theta-\mu)^2+\eta)\omega/2) d\omega \\ &= \frac{\Gamma((d+1)/2)}{(\kappa(\theta-\mu)^2+\eta)^{(d+1)/2}} \\ &\propto \left(1+\frac{1}{d}\frac{(\theta-\mu)^2}{\eta/d\kappa}\right)^{-(d+1)/2} \end{aligned} \quad .$$

Therefore, the center $m=\mu$, scale $s=\sqrt{\eta/d\kappa}$ and degrees of freedom $\nu=d$.

(B) The joint distribution of θ, ω, y is:

$$\begin{aligned} p(\theta,\omega|y) \propto p(y,\theta,\omega) &\propto \omega^{(d+1)/2-1} exp(-(\kappa(\theta-\mu)^2+\eta)\omega/2)\Pi_{i=1}^n \omega^{1/2} exp(-\omega(y_i-\theta)^2/2) \\ &= \omega^{(d+n+1)/2-1} exp\left(-\left(\kappa(\theta-\mu)^2+\sum_{i=1}^n (y_i-\theta)^2+\eta\right)\omega/2\right) \\ &\propto \omega^{(d+n+1)/2-1} exp\left(-\left((\kappa+n)\left(\theta-\frac{\kappa\mu+n\bar{y}}{\kappa+n}\right)^2+\frac{\kappa n(\mu-\bar{y})^2}{\kappa+n}+s_y+\eta\right)\omega/2\right) \end{aligned}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, $s_y = \sum_{i=1}^n (y_i-\bar{y})^2$, $d^* = d+n$ $\kappa^* = \kappa+n$ and $\mu^* = \frac{\kappa\mu+n\bar{y}}{\kappa+n}$ and $\eta^* = \eta + \frac{\kappa n(\mu-\bar{y})^2}{\kappa+n} + s_y$.

(C) The conditional posterior distribution $\theta|y,\omega$ is $\mathcal{N}(\mu^*, (\kappa^*\omega)^{-1})$.

(D) Integrate out θ, we have

$$\begin{aligned} p(\omega|y) &\propto \omega^{(d^*+1)/2-1} exp(-\eta^*\omega/2)) \int_{\mathbb{R}} exp(-\kappa^*\omega(\theta-\mu^*)^2) d\theta \\ &\propto \omega^{d^*/2-1} \int_{\mathbb{R}} \frac{(\kappa^*\omega)^{1/2}}{\sqrt{2\pi}} exp(-\kappa^*\omega(\theta-\mu^*)^2) d\theta \\ &= \omega^{d^*/2-1} exp(-\eta^*\omega/2)) \end{aligned}$$

(E) Since the posterior is also a normal/gamma distribution, similar to (A), we know the center $m^* = \mu^*$, scale $s = \sqrt{\eta^*/d^*\kappa^*}$ and degrees of freedom $\nu^* = d^*$.

(F) False. When $d, \eta \to 0$, the Gamma kernel $\omega^{d/2-1} exp(-\eta\omega/2) \to \omega^{-1}$, which is not integrable on $(0, 1)$.

(G) True. When κ, d and η approach zero, none of κ^*, d and η^* approaches zero. Thus, the posterior distribution is non-degenerate.

(H) False. The degrees of freedom of the frequentist t-test is $n-1$ instead of n.

2 The conjugate Gaussian linear model

Now consider the homoskedastic Gaussian linear model

$$(\mathbf{y} \mid \beta, \sigma^2) \sim \mathrm{N}(X\beta, (\omega\Lambda)^{-1}),$$

where $\mathbf{y}$ 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 $\Lambda = I$, the $n \times n$ identity matrix, so that the residuals of the model are i.i.d. normal with variance σ^2 (hence homoskedastic). But we'll want to 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 that the first entry in β is actually the intercept.

We'll again work in terms of the precision $\omega = \sigma^2$, and consider a normal–gamma prior for β:

$$(\beta \mid \omega) \sim N(m, (\omega K)^{-1}) \tag{4}$$

$$\omega \sim \mathrm{Gamma}(d/2, \eta/2). \tag{5}$$

Here K is a $p \times p$ precision matrix in the multivariate normal prior for β, 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.[3]

2.1 Basics

(A) Derive the conditional posterior $p(\beta \mid \mathbf{y}, \omega)$.

(B) Derive the marginal posterior $p(\omega \mid \mathbf{y})$.

(C) Putting these together, derive the marginal posterior $p(\beta \mid \mathbf{y})$.

Answer

The joint p.d.f. is

$$p(\mathbf{y}, \beta, \omega) = \omega^{(d+n+p)/2-1} exp(\eta\omega/2) exp(-\omega((\beta - m)^T K(\beta - m) + (y - X\beta)^T \Lambda(y - X\beta))/2)$$

(A) $p(\beta|\mathbf{y}, \omega) \propto exp(-\beta^T\omega(K + X^T\Lambda X)\beta/2 + \omega(Km + X^T\Lambda y)^T\beta) \Rightarrow \beta|\mathbf{y}, \omega \sim \mathcal{N}(\hat{m}, (\omega\hat{K})^{-1})$, where $\hat{K} = K + X^T\Lambda X$ and $\hat{m} = \hat{K}^{-1}(Km + X^T\Lambda y)$

(B) $p(\omega|\mathbf{y}) = p(\mathbf{y}, \beta, \omega)/p(\beta|\mathbf{y}, \omega) \propto \omega^{(d+n)/2-1} exp(\eta\omega/2) exp(-\omega(m^T Km + y^T\Lambda y - \hat{m}^T\hat{\Lambda}\hat{m})/2)$

It is clear that $\omega|\mathbf{y}$ follows a Gamma distribution, i.e. $\Gamma(d^*/2, \eta^*/2)$, where $d^* = d + n$ and $\eta^* = \eta + m^T Km + y^T\Lambda y - \hat{m}^T\hat{K}\hat{m}$

[3] 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.

(C) $p(\beta|\mathbf{y}) = p(\mathbf{y},\beta,\omega)/p(\omega|\mathbf{y},\beta) \propto \frac{\Gamma((d+n+1)/2)}{(\eta^*+(\beta-\hat{m})^T\hat{\Lambda}(\beta-\hat{m}))^{(d^*+p)/2}}$

The p.d.f. has all the characteristics of a location-scale student t distribution. Specifically, the degree of freedom $\nu^* = d^*$, the center $m^* = \hat{m}$ and the scale matrix $\Sigma = (K^*)^{-1}$, $K^* = \nu^*\hat{K}/\eta^*$.

2.2 Example

Take a look at the data in "greenbuildings.csv" from the class website, and consider the following short case study in regression. In commercial real estate, issues of eco-friendliness are intimately tied up with ordinary decisions about how to allocate capital. Every new project involves negotiating a trade-off between costs incurred and benefits realized over the lifetime of the building. In this context, the decision to invest in eco-friendly buildings could pay off in many possible ways. Of course, at the end of the day, tenants may or may not be willing to pay a premium for rental space in green buildings. We can only find out by carefully examining data on the commercial real-estate market.

The file "greenbuildings.csv" contains data on over 7,000 commercial rental properties from across the United States. Of these, 685 properties have been awarded either LEED or EnergyStar certification as a green building. The basic idea is that a commercial property can receive a green certification if its energy efficiency, carbon footprint, site selection, and building materials meet certain environmental benchmarks. A group of real estate economists constructed the data set in the following way. They first sampled 679 green-certified buildings listed on the LEED or EnergyStar websites, and cross-referenced those buildings with data from the CoStar Group (www.costar.com), a standard data vendor in commercial real estate. In order to provide a control population, each of these 685 buildings was matched to a cluster of nearby commercial buildings in the CoStar database. Each small cluster contains one green-certified building, and all non-rated buildings within a quarter-mile radius of the certified building. On average, each of the 679 clusters contains roughly 12 buildings.

For our purposes, the variables we'll use are:

- Rent: rent in dollars per square foot per year (the standard measure in commercial real estate)
- leasing rate: what percentage (0-100 scale) of the building is actually leased to a tenant
- green rating: 0/1, whether the building has a green certification
- CityMarketRent: an index of commercial real estate prices for the market in which the building is located
- age: age of the building in years
- class a, class b: indicators for whether the building is class A or class B (if both are 0, the building is class C)

To build a model here, first define a new response variable, revenue per square foot, defined as the Rent divided by the leasing rate, times 100. This represents how much actual revenue the property brings in per square foot. This will be your response variable. As predictors, using green rating, City Market Rent, age, class a, and class b (and an intercept, of course). Fit the (homoskedastic) Bayesian linear model to this data set, choosing $\Lambda = I$ and something diagonal and pretty vague for the prior precision matrix $K = \text{diag}(\kappa_1, \kappa_2)$.

What is a 95% Bayesian credible interval for the coefficient on the green rating variable? How does your result compare to the classical 95% confidence interval, e.g. from lm in R? What does a histogram of the model residuals reveal? (Here you can define the residual vector as $y - X\hat{\beta}$, where $\hat{\beta}$ is the posterior mean.) Are you happy with your model? Why or why not?

Answer

2.3 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:

$$(\mathbf{y} \mid \beta, \omega, \Lambda) \sim N(X\beta, (\omega\Lambda)^{-1}) \tag{6}$$
$$\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n) \tag{7}$$
$$\lambda_i \overset{iid}{\sim} \text{Gamma}(h/2, h/2) \tag{8}$$
$$(\beta \mid \omega) \sim N(m, (\omega K)^{-1}) \tag{9}$$
$$\omega \sim \text{Gamma}(d/2, \eta/2)\,. \tag{10}$$

where h is a fixed hyperparameter.

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

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

(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.

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

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 $p(\beta, \omega, \Lambda \mid \mathbf{y})$. 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 green buildings 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)?[4] How does the 95% credible interval on each model term compare with the credible intervals you got under the homoskedastic linear model in the previous section? Are there certain regions of predictor space that seem to be associated with higher variance residuals?

Answer

The joint p.d.f. of $\mathbf{y}, \beta, \omega, \Lambda$ is:

[4] An interesting plot will be a plot of the posterior mean of $1/\lambda_i$ for each data point.

$$p(\mathbf{y},\beta,\omega,\Lambda) = \omega^{(d+n+p)/2-1} exp(\eta\omega/2) exp(-\omega((\beta-m)^T K(\beta-m) + (\mathbf{y}-X\beta)^T \Lambda(y-X\beta))/2)$$
$$\prod_{i=1}^{n} \lambda^{(h+1)/2-1} exp(-h\lambda_i/2)$$

(A) The p.d.f. of conditional distribution $\mathbf{y}|X,\beta,\omega$ can be derived as follows:

$$\begin{aligned} p(y_i|\beta,\omega;x_i) &\propto p(y_i,\beta,\omega;x_i) \\ &= \int_{\lambda_i} P(y_i,\lambda_i,\beta,\omega;x_i) d\lambda_i \\ &\propto \int_{\lambda_i} \lambda_i^{(h+1)/2-1} exp(-\lambda_i(h+\omega(y_i - x_i^T\beta)^2/2) d\lambda_i \\ &\propto (h+\omega(y_i - x_i^T\beta)^2)^{-(h+1)/2} \end{aligned}.$$

Therefore, we know that $y_i|\beta,\omega \sim \omega^{-1/2}(t_h + x_i^T\beta)$.

(B) The p.d.f. of conditional posterior distribution $\lambda_i|\mathbf{y},\beta,\omega$ is:

$$p(\lambda_i|\mathbf{y},\beta.\omega) \propto p(\mathbf{y},\beta,\omega,\lambda_i) \propto \lambda_i^{(h+1)/2-1} exp(-(h+(y_i - x_i^T\beta)^2)\lambda_i/2)$$

Therefore, we know that $\lambda_i|\mathbf{y},\beta.\omega \sim \Gamma((h+1)/2, (h+\omega(y_i - x_i^T\beta)^2)/2)$.