Bayesian Learning, 7.5 hp

Home assignment - Part A

Mattias Villani

2025-09-17

Table of contents

Problem 1 - Bernoulli data with a Beta prior 1
Problem 2 - Modeling stock returns with a student-t distribution 2
Problem 3 - Making decisions 3
Problem 4 - Polynomial regression 5

Problem 1 - Bernoulli data with a Beta prior

Let $y_1, ..., y_n|\theta \sim \text{Bern}(\theta)$, and assume that you have obtained a sample with $s = 14$ successes in $n = 20$ trials. Assume a $\text{Beta}(\alpha_0, \beta_0)$ prior for θ and let $\alpha_0 = \beta_0 = 2$.

Problem 1a)

Draw random numbers from the posterior $\theta|y \sim \text{Beta}(\alpha_0+s, \beta_0+f)$, where $y = (y_1, ..., y_n)$, and verify graphically that the Monte Carlo (MC) estimates of the posterior mean and standard deviation converges to the true values as the number of random draws grows large.

Problem 1b)

Use simulation (`nDraws = 10000`) to compute the posterior probability $\Pr(\theta < 0.5|y)$ and compare with the exact value [hint: `pbeta()`].

Problem 1c)

Compute the posterior distribution of the log-odds $\phi = \log\left(\frac{\theta}{1-\theta}\right)$ by simulation.

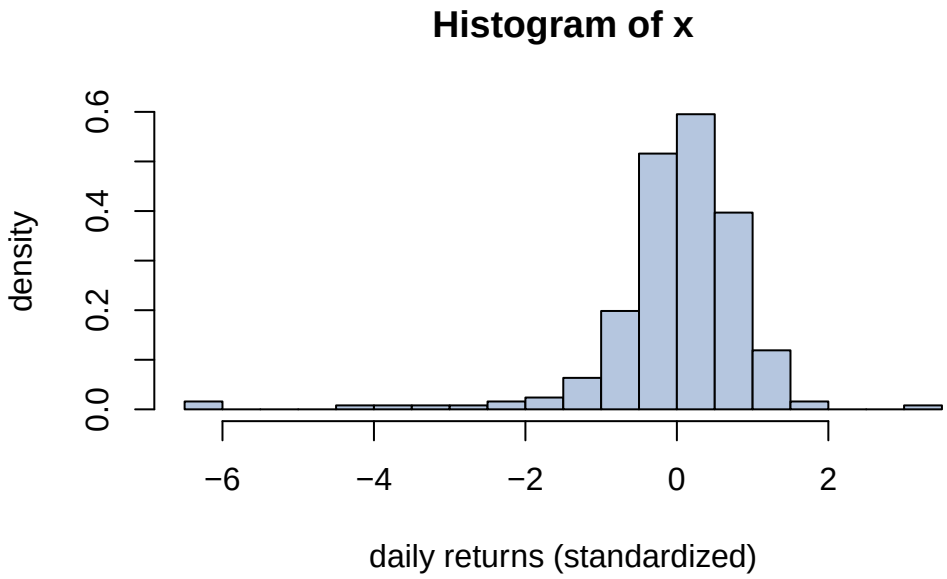
Problem 2 - Modeling stock returns with a student-t distribution

The vector `returns` in the dataset `ericsson` (load the data with `load(ericsson.RData)`) contains 252 observations on daily percentage returns on the Ericsson stock. In this exercise we analyze the standardized returns:

```
load("ericsson.RData") # Place the data file in the same directory as your Quarto file.
x = (returns - mean(returns)) / sd(returns) # standardized returns
```

Always a good idea to plot the data before the analysis:

```
hist(x, 30, freq = FALSE, xlab = "daily returns (standardized)", ylab = "density",
     col = prettycols[5])
```



The heavy tails with occasional extreme returns are evident from the histogram. Let X_i be the standardized returns on the ith day. We will here use the heavy-tailed student-t distribution, $t(\mu = 0, \sigma^2 = 1, \nu)$, to model the returns, and we will for simplicity assume that the returns are independent. Note that the location is zero and the variance one, since we have standardized the data. The only unknown parameter is the degrees of freedom $\nu > 0$ which models the tails (smaller ν gives heavier tails); note that ν has to be positive, but does not need to be an integer. In summary, we assume the following model for the standardized Ericsson daily stock returns

$$X_1, \dots, X_n | \nu \overset{\text{iid}}{\sim} t(0, 1, \nu)$$

Problem 2a)

Plot the log-likelihood function for ν based on the 252 data observations. Note that the `dt` function in R gives the density for the standard $t(0, 1, \nu)$ distribution, and that R uses `df` for the degrees of freedom parameter ν. From the graph, what would you say is the maximum likelihood estimate of ν ?

Problem 2b)

Plot the likelihood function $p(x_1, \ldots, x_n|\nu)$ for ν. Compare $p(x_1, \ldots, x_n|\nu = 1)$ and $p(x_1, \ldots, x_n|\nu = 10)$, what do you conclude from this comparison?

Problem 2c)

Plot the logarithm of the posterior distribution for ν

$$\log p(\nu|x_1, \ldots, x_n) \propto \log p(x_1, \ldots, x_n|\nu) + \log p(\nu)$$

where $p(\nu)$ is density of the prior $\nu \sim \text{Expon}(0.25)$ in the rate parametrization (i.e. $\mathbb{E}(\nu) = 4$ *a priori*). [hint: the posterior distribution is not a known distribution, so you have to plot the log-posterior over a grid of values for ν).

Problem 2d)

Plot the posterior distribution of ν. [hint: don't forget to normalize numerically so that the posterior is true density].
Overlay the prior density [hint: `lines()` adds lines to existing plot)

Problem 2e)

Compute the posterior mean of ν using numerical integration.

Problem 3 - Making decisions

You are the manager of a small fruit shop that sells a particular exclusive mango fruit. You buy each mango for \$10 and sell them for \$20. One reason for the large mark-up is that some of the mango may go unsold, and must then be used for mango smoothies which only brings in \$3 per mango, i.e. a loss of \$7 on each mango that goes unsold. Each week you must decide on how many mangoes to bring into the shop, and the demand is uncertain.

Problem 3a)

Let X_i be the demand for the mango on the ith week, and assume the model

$$X_1, \dots, X_n | \theta \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$$

where θ is the mean demand, and $\theta \sim \text{Gamma}(\alpha = 7, \beta = 2)$ *a priori*. The manager has collected data on the number of sold mangoes in the previous ten weeks: `x = c(3, 5, 4, 3, 6, 8, 6, 1, 14, 3)`. Simulate 10000 draws from the posterior $p(\theta|x_1, \dots, x_{10})$ and plot a histogram to represent the posterior density. Use the posterior draws to compute the **posterior** probability $\Pr(\theta > 8|x_1, \dots, x_{10})$. Compare with the exact result using the `pgamma` function.

Problem 3b)

The predictive distribution for the next week's demand, X_{n+1}, can be shown to follow a negative binomial distribution (see Chapter 6 in the Bayesian Learning book)

$$X_{n+1}|x_1, \dots, x_n \sim \text{NegBin}\Big(\alpha + \sum_{i=1}^{n} x_i, \frac{\beta + n}{\beta + n + 1}\Big)$$

where $\text{NegBin}(r, \theta)$ is the negative binomial distribution with support $x \in \{0, 1, 2, ...\}$. The `rnbinom`, `dnbinom` and `pnbinom` functions in R implements the negative binomial distribution. Note that the first argument r is called `size` in R, and the second argument θ is called `prob`. Simulate 10000 draws from the predictive distribution and plot the distribution. Use the predictive draws to compute the predictive probability that at least 8 mangoes are sold next week. Compare with the exact result using the `pnbinom` function.

Problem 3c)

You need to decide on how many mangoes to order for the coming week (week 11). Call this action a_{11}. Make this choice based on maximizing posterior expected utility (or predictive expected utility, since the uncertain demand is in the future). Use profit as the utility:

- The profit from the sold mangoes is $10 \cdot \min(X_{11}, a_{11})$ (we cannot sell more than we have in stock)
- the loss from the unsold mangoes is $-7 \cdot \max(0, a_{11} - X_{11})$ (we loose \$7 on each mango left at the end of the week).

So, the utility is

$$U(X_{11}, a_{11}) = 10 \cdot \min(X_{11}, a_{11}) - 7 \cdot \max(0, a_{11} - X_{11})$$

Use simulation to find the optimal number of mangoes to buy for week 11. [hint: re-use the same predictive draws for all values of a_{11}. Maybe the `sapply` function can be useful, but it is not strictly necessary].

Problem 4 - Polynomial regression

The dataset `tempLinkoping` in the package `SUdatasets` contains daily temperatures (in Celcius degrees) at Malmslätt, Linköping over the course of the year 2016 (366 days since 2016 was a leap year). The response variable is `temp` and the covariate is

$$\texttt{time} = \frac{\text{the number of days since beginning of year}}{366}$$

Your task is to perform a Bayesian analysis of a quadratic regression

$$\texttt{temp} = \beta_0 + \beta_1 \cdot \texttt{time} + \beta_2 \cdot \texttt{time}^2 + \varepsilon, \quad \varepsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

You can access the data from Github as follows

```
#install.packages("remotes")                # Uncomment this the first time
library(remotes)
#install_github("StatisticsSU/SUdatasets")  # Uncomment this the first time
library(SUdatasets)
head(tempLinkoping)
```

```
         time  temp
1 0.002732240   0.1
2 0.005464481  -4.5
3 0.008196721  -6.3
4 0.010928962  -9.6
5 0.013661202  -9.9
6 0.016393443 -17.1
```

Problem 4a) Determine a suitable prior distribution

Use the conjugate prior

$$\beta|\sigma^2 \sim N(\mu_0, \sigma^2\Omega_0^{-1})$$

$$\sigma^2 \sim \text{inv}{-}\chi^2(\nu_0, \sigma_0^2)$$

You need to determine suitable prior hyperparameters μ_0, Ω_0, σ_0^2 and ν_0. Start with $\mu_0 = (-10, 100, -100)^\top$, $\Omega_0 = 0.01 \cdot I_3$, where I_3 is the 3×3 identity matrix, $\nu_0 = 3$ and $\sigma_0^2 = 1$. Check if this prior agrees with your prior opinions by simulating draws from the joint prior of all parameters and for every draw compute the regression curve. This gives a collection of regression curves, one for each draw from the prior. Do the collection of curves look reasonable? If not, change the prior hyperparameters until the results agree with your prior beliefs about the regression curve. The `mvtnorm` package contains the multivariate normal distribution, and here is an implementation of a random number generator for the inv$-\chi^2(\nu_0, \sigma_0^2)$ distribution:

```
# Simulator for the scaled inverse Chi-square distribution
rScaledInvChi2 <- function(n, v_0, sigma2_0){
  return((v_0*sigma2_0)/rchisq(n, df = v_0))
}
```

Problem 4b) Simulating from the posterior

Write a program that *simulates from the joint posterior distribution* of β_0, β_1,β_2 and σ^2. Plot the marginal posteriors for each parameter as a histogram. Also produce another figure with a scatter plot of the temperature data and overlay a curve for the posterior median of the regression function $f(time) = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot time^2$, computed for every value of *time*. Also overlay curves for the lower 2.5% and upper 97.5% posterior credible interval for $f(time)$. That is, compute the 95% equal tail posterior probability intervals for every value of *time* and then connect the lower and upper limits of the interval by curves. Does the interval bands contain most of the data points? Should they?

Problem 4c) Locating the day with the highest expected temperature

It is of interest to locate the *time* with the highest expected temperature (that is, the *time* where $f(time)$ is maximal). Let's call this value $\tilde{x}$. Use the simulations in b) to simulate from the posterior distribution of $\tilde{x}$. [Hint: since the regression curve is a quadratic, the maximum on the curve is given by $\tilde{x} = -\frac{\beta_1}{2\beta_2}$]