Mattias Villani
Dept of Statistics, Stockholm University
Dept of Computer and Information Science, Linköping University

Bayesian Statistics I
Bayesian Learning

Mathematical Exercises 1

Try to solve the problems before class. Don't worry if you fail, the important thing is trying.
You should not hand in any solutions.
This part of the course is not obligatory and is not graded.

---

1. What matters is the likelihood
   (a) Assume that you want to investigate the proportion (θ) of defective items manufactured at a production line. Your colleague takes a random sample of 30 items and tells you for each item whether or not it was defective. So she records the data as $x_1 = 0, x_2 = 1, ..., x_n = 0$, where $x_i = 1$ if the item is defective and $x_i = 0$ otherwise. There were three defective items in the sample. Assume a uniform prior for θ. Compute the posterior of θ.
   (b) Assume your colleague only told you that there were three defective items in the sample of 30 items, but she did not tell you which specific items that were defective. Assume a uniform prior for θ. Compute the posterior of θ.
   (c) Your colleague now tells you that she did not decide on the sample size before the sampling was performed. Her sampling plan was to keep on sampling items until she had found three defective ones. It just happened that the 30th item was the third one to be defective. Redo the posterior calculation, under the new sampling scheme. Compare the results with Problem 1(a). [Hint: negative binomial distribution]

2. Normal with a normal is normal
   (a) Let $x_1, ..., x_n \overset{iid}{\sim} \mathrm{N}(\theta, \sigma^2)$ and assume that σ^2 is known. Assume a uniform prior:
   $$p(\theta) \propto c.$$
   Derive the posterior distribution of θ.
   (b) Assume now a normal prior
   $$\theta \sim N(\mu_0, \tau_0^2).$$
   Derive the posterior distribution of θ.

3. IT ENDS WITH AN ABNORMAL TWIST

   (a) Let $x_1, ..., x_{10} \overset{iid}{\sim} N(\theta, 1)$. Let the sample mean be $\bar{x} = 1.873$. Assume that $\theta \sim N(0, 5)$ apriori. Compute the posterior distribution of θ.

   (b) Assume now that you have a second sample $y_1, ..., y_{10} \overset{iid}{\sim} N(\theta, 2)$, where θ is the same quantity as in 3a. The sample mean in this second sample is $\bar{y} = 0.582$. Compute the posterior distribution of θ using both samples (the x's and the y's) under the assumption that the two samples are indepedent.

   (c) You have now managed to obtain a third sample $z_1, ..., z_{10} \overset{iid}{\sim} N(\theta, 3)$, with mean $\bar{z} = 1.221$. Unfortunately, the measuring device for this latter sample was defective: any measurement above 3 was recorded as 3. There were two such measurements. Compute the posterior distribution based on all three samples (x, y and z). [Hint: in this case the posterior distribution is not a known distribution (it is not normal for example). It is enough to give an expression for the (unnormalized) posterior. You can also plot this over a grid on your computer, if you like.]

4. SOLVING THIS PROBLEM SHOULD NOT TAKE EXPONENTIALLY LONG TIME

   (a) Let $x_1, ..., x_n \overset{iid}{\sim} \text{Expon}(\theta)$. We use the parametrization of the exponential distribution where if $X \sim \text{Expon}(\theta)$ then $E(X) = 1/\theta$. Show that the conjugate prior for the exponential model is $\theta \sim \text{Gamma}(\alpha, \beta)$. Derive the posterior distribution for θ .

Have fun!

- Mattias

Problem 2a)

Model: $X_1, \dots, X_n \mid \theta \overset{iid}{\sim} N(\theta, \sigma^2)$ $\quad \sigma^2$ known

Prior: $P(\theta) \propto$ constant

Posterior:
$$p(\theta \mid x_1, \dots, x_n) \propto p(x_1, \dots, x_n \mid \theta) \cdot p(\theta)$$
$$= p(x_1, \dots, x_n \mid \theta) \qquad \text{[prior is constant]}$$
$$= \prod_{i=1}^{n} p(x_i \mid \theta) \qquad \text{[Independence]}$$
$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \theta)^2\right)$$
$$\propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(x_i - \theta)^2\right)$$
$$= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right)$$

Now,
$$\sum_{i=1}^{n}(x_i - \theta)^2 = \sum_{i=1}^{n}\left((x_i - \bar{x}) - (\theta - \bar{x})\right)^2$$
$$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\theta - \bar{x})^2 - 2(\theta - \bar{x})\sum_{i=1}^{n}(x_i - \bar{x})$$
$$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\theta - \bar{x})^2 - 2(\theta - \bar{x})\underbrace{\left(\sum_{i=1}^{n} x_i - n\bar{x}\right)}_{0}$$
$$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\theta - \bar{x})^2$$

$$\text{So,} \quad p(\theta \mid x_1, \dots, x_n) \propto \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i - x)^2 - n(\theta - \bar{x})^2\right)\right)$$

$$= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) \cdot \exp\left(-\frac{1}{2\sigma^2} n(\theta - \bar{x})^2\right)$$

$$\propto \exp\left(-\frac{1}{2(\sigma^2/n)}(\theta - \bar{x})^2\right)$$

$$\propto N\left(\bar{x}, \frac{\sigma^2}{n}\right)$$

Problem 2b)

<u>prior:</u> $\theta \sim N(\mu_0, \tau_0^2)$

<u>posterior:</u> $p(\theta \mid x_1, \dots, x_n) \propto p(x_1, \dots, x_n \mid \theta) \cdot p(\theta$

$$\propto \exp\left(-\frac{1}{2(\sigma^2/n)}(\theta - \bar{x})^2\right) \cdot \exp\left(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right)$$

$$= \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2/n}(\theta - \bar{x})^2 + \frac{1}{\tau_0^2}(\theta - \mu_0)^2\right)\right)$$

$$\frac{n}{\sigma^2}(\theta^2 + \bar{x}^2 - 2\theta\bar{x}) + \frac{1}{\tau_0^2}(\theta^2 + \mu_0^2 - 2\theta\mu_0)$$

$$= \text{const} + \theta^2\left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}\right) - 2\left(\frac{n}{\sigma^2}\bar{x} + \mu_0\right)\theta$$

Not important. It will end up in the normalization constant.

we want this to be of the form

$\frac{1}{\tau_n^2}(\theta - \mu_n)^2$ since then the posterior will be $N(\mu_n, \tau_n^2)$

this is acheived if

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}$$ (since this is the coefficient on θ^2 above)

and

$$\frac{\mu_n}{\tau_n^2} = \frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau_0^2}$$ (since this is the coefficient on -2θ above)

So,
$$\mu_n = \tau_n^2 \left(\frac{n}{\sigma^2}\bar{x} + \frac{1}{\tau_0^2}\mu_0 \right)$$

$$= \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}} \left(\frac{n}{\sigma^2}\bar{x} + \frac{1}{\tau_0^2}\mu_0 \right)$$

$$= \omega\bar{x} + (1-\omega)\mu_0$$

with $\omega = \dfrac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}$

Pheew!!!

Problem 4b)

Model: $x_1, \dots, x_n \sim Exp(\theta)$

Prior: $\theta \sim Gamma(\alpha, \beta)$

Posterior: $p(\theta \mid x_1, \dots, x_n) \propto p(x_1, \dots x_n \mid \theta) \cdot p(\theta)$

$$\propto \prod_{i=1}^{n} \underbrace{\theta e^{-\theta x_i}}_{\text{Exponential distribution}} \cdot \underbrace{\theta^{\alpha-1} e^{-\beta\theta}}_{\text{proportional to Gamma}(\alpha,\beta)\text{ density.}}$$

$$\propto \theta^{n} e^{-\theta \sum_{i=1}^{n} x_i} \theta^{\alpha-1} e^{-\beta\theta}$$

$$= \theta^{n+\alpha-1} e^{-(\sum_{i=1}^{n} x_i + \beta)\theta}$$

$$\propto Gamma(\alpha+n, \beta + \sum_{i=1}^{n} x_i)$$

Mattias Villani
Dept of Statistics, Stockholm University
Dept of Computer and Information Science, Linköping University

Mathematical Exercises 2

Try to solve the problems before class. Don't worry if you fail, the important thing is trying.
You should not hand in any solutions.
This part of the course is not obligatory and is not graded.

1. TAU-CHI.
   (a) Let $x_1, ..., x_n \overset{iid}{\sim} N(\theta, \sigma^2)$. Assume that θ is known, but σ^2 unknown. Derive the posterior distribution for σ^2. Use the conjugate prior.
   (b) Assume that $\theta = 1$ and that you have observed the data $x_1 = 0.6, x_2 = 3.2, x_3 = 1.2$. Compute the posterior of σ^2 based on these three data points. Use a prior with very little information (it is up to you how to define little information).

2. FEEL THE BERN.
   (a) Let $x_1, ..., x_n \overset{iid}{\sim} Bern(\theta)$, with a $Beta(\alpha, \beta)$ prior for θ. Derive the predictive distribution for x_{n+1}.
   (b) You need to decide if you bring your umbrella during your daily walk. It has rained on two days during the last ten days, and you assess those ten days to be representative also for the weather today, the 11th day. Your utility for the action-state combinations are given in the table below. Assume a $Beta(1, 1)$ prior for θ. Compute the Bayesian decision.
   (c) How sensitive is your decision in (b) to the changes in the prior hyperparameters, α and β?

	Rainy	Sunny
Bring umbrella	10	20
Leave umbrella	-50	50

3. Campaign or no campaign - that is the question.

(a) Let x_i be the number of sales of a product on month i. Let $x_1, ..., x_n \overset{iid}{\sim} N(\theta, \sigma^2)$ be the (approximate) distribution for the sales, and let $\theta \sim N(200, 50^2)$ a priori. Assume that $\sigma^2 = 25^2$ and that we have observed $n = 5$ and $\bar{x} = 320.4$. Compute the predictive distribution for x_6.

(b) The company has the choice of performing a marketing campaign for their product. The marketing campaign costs \$300 and is believed to increase sales by 20% compared to when no campaign is performed. The company sells the product for $p = 10$ dollar and the cost of producing the product is $q = 5$ dollar. There are no fixed production costs. Assume that the company's utility is described by $U(y) = 1 - \exp(-y/1000)$, where y is the total profit from sales in the next month. Should the company perform the marketing campaign? [Hint: the expected value of the exponential function of a normal random variable $S \sim N(\mu, \sigma^2)$ is $E(\exp(S)) = \exp(\mu + \sigma^2/2)$.]

4. Predictive distribution for a Poisson model

(a) Do Exercise 13(a) in Chapter 2 of the course book. That is, assume that the number of fatal accidents on scheduled airline flights each year are independent with a Poisson(θ) distribution. Set a prior distribution for θ and determine the posterior distribution based on the data from 1976 through 1985, given below. Under this model, give a 95% predictive interval for the number of fatal accidents in 1986. You can use the normal approximation to the gamma and Poisson or compute using simulation.

24, 25, 31, 31, 22, 21, 26, 20, 16, 22

Have fun!

- Mattias

EXERCISE SET NO. 2
BAYESIAN LEARNING

1a) $X_1, \dots, X_n \mid \theta, \sigma^2 \sim N(\theta, \sigma^2)$

θ known

$\sigma^2 \sim \mathrm{Inv}\chi^2(\nu_0, \sigma_0^2)$

Posterior: (implicit conditioning on θ)

$$p(\sigma^2 \mid x_1, \dots, x_n) \propto p(x_1 \dots, x_n \mid \sigma^2)\, p(\sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \theta)^2\right) \cdot p(\sigma^2)$$

$$\left[\text{define } s^2 = \frac{\sum_{i=1}^{n}(x_i - \theta)^2}{n} \right]$$

$$\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left(-\frac{ns^2}{2\sigma^2}\right) \boxed{\frac{\exp\left(\frac{-\nu_0\sigma_0^2}{2\sigma^2}\right)}{(\sigma^2)^{\nu_0/2+1}}}$$

Density (pdf) of $\mathrm{Inv}\chi^2(\nu_0, \sigma_0^2)$ prior.

$$= \frac{\exp\left(-\frac{ns^2 + \nu_0\sigma_0^2}{2\sigma^2}\right)}{(\sigma^2)^{(n+\nu_0)/2+1}}$$

So $\sigma^2 \mid x_1 \dots, x_n, \theta \sim \mathrm{Inv}\chi^2(\nu_n, \sigma_n^2)$

$$\nu_n = \nu_0 + n$$

$$\sigma_n^2 = \frac{ns^2 + \nu_0\sigma_0^2}{\nu_0 + n}$$

1b) $s^2 = \frac{\sum_{i=1}^{3}(x_i-\theta)^2}{3} = \frac{(0.6-1)^2+(3.2-1)^2+(1.2-1)^2}{3} = 1.68$

Non-informative : $v_0 \to 0$

why is this non-informative?

Reason 1: v_n becomes n

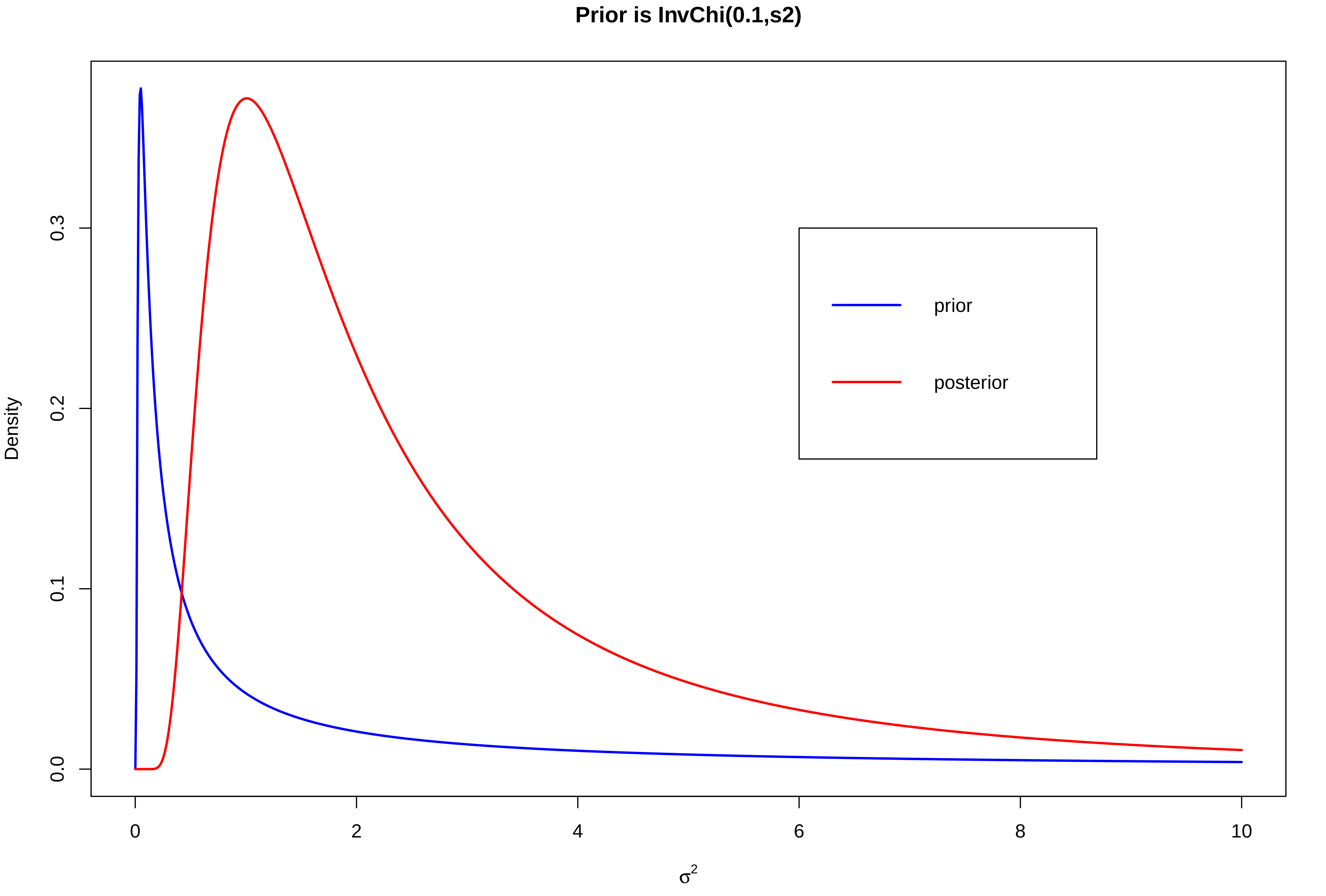
Reason 2: $\text{Inv}\chi^2(v_0, \sigma_0^2)$ becomes $\frac{1}{\sigma^2}$

when $v_0 \to 0$.

Note that as $v_0 \to 0$ the posterior approaches the $\text{Inv}\chi^2(n, s^2)$ density.

So,

$\sigma^2 | x_1, x_2, x_3 \sim \text{Inv}\chi^2(3, 1.68)$


Prior is InvChi(0.1,s2)
prior
posterior
Density
0.0
0.1
0.2
0.3
0
2
4
6
8
10
σ2

2a Prediction of Bernoulli data

The predictive distribution of x_{n+1} given the first n trials ($x_{1:n}$) is

$$\begin{aligned}
p(x_{n+1}|x_{1:n}) &= \int p(x_{n+1}|\theta)p(\theta|x_{1:n})d\theta && x_{n+1} \text{ is indep. of } x_{1:n} \text{ given } \theta \\
&= \int \theta^{x_{n+1}}(1-\theta)^{1-x_{n+1}}p(\theta|x_{1:n})d\theta && \theta|x_{1:n} \sim \text{Beta}(\alpha+s,\beta+f) \\
&= \int \theta^{x_{n+1}}(1-\theta)^{1-x_{n+1}}\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+s)\Gamma(\beta+f)}\theta^{\alpha+s-1}(1-\theta)^{\beta+f-1}d\theta \\
&= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+s)\Gamma(\beta+f)}\int \theta^{x_{n+1}+\alpha+s-1}(1-\theta)^{1-x_{n+1}+\beta+f-1}d\theta \\
&= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+s)\Gamma(\beta+f)}\frac{\Gamma(x_{n+1}+\alpha+s)\Gamma(1-x_{n+1}+\beta+f)}{\Gamma(1+\alpha+\beta+n)} \\
&= \frac{\Gamma(x_{n+1}+\alpha+s)\Gamma(1-x_{n+1}+\beta+f)}{\Gamma(\alpha+s)\Gamma(\beta+f)(\alpha+\beta+n)} && \text{using } \Gamma(y+1)=y\Gamma(y)
\end{aligned}$$

So,

$$p(x_{n+1}=1|x_{1:n}) = \frac{\Gamma(1+\alpha+s)}{\Gamma(\alpha+s)(\alpha+\beta+n)} = \frac{(\alpha+s)\Gamma(\alpha+s)}{\Gamma(\alpha+s)(\alpha+\beta+n)} = \frac{\alpha+s}{\alpha+\beta+n}$$

and therefore [since $p(x_{n+1}=0|x_{1:n}) = 1 - p(x_{n+1}=1|x_{1:n})$]

$$p(x_{n+1}=0|x_{1:n}) = \frac{\beta+f}{\alpha+\beta+n}.$$

The predictive distribution is therefore

$$x_{n+1}|x_{1:n} \sim \text{Bern}\left(\frac{\alpha+s}{\alpha+\beta+n}\right).$$

2b Umbrella decision

(a) Let x_{11} be the binary variable indicating rain on the 11th day. From Problem 2a, the predictive distribution for the $(n+1)$th Bernoulli trial is

$$x_{n+1}|x_{1:n} \sim \text{Bern}\left(\frac{\alpha+s}{\alpha+\beta+n}\right).$$

and the predictive probability for rain is therefore here

$$\Pr(x_{11}=1|x_{1:10}) = \frac{1+2}{1+1+10} = 0.25.$$

The expected utility from the decision to bring the umbrella is then

$$\mathrm{EU}_{\mathrm{bring}} = \mathrm{Pr}(\mathrm{sunny}) \cdot \mathrm{U}(\mathrm{bring}, \mathrm{sunny}) + \mathrm{Pr}(\mathrm{rain}) \cdot \mathrm{U}(\mathrm{bring}, \mathrm{rain}) = 0.75 \cdot 20 + 0.25 \cdot 10 = 17.5$$

and the expected utility of leaving the umbrella at home is

$$\mathrm{EU}_{\mathrm{leave}} = \mathrm{Pr}(\mathrm{sunny}) \cdot \mathrm{U}(\mathrm{leave}, \mathrm{sunny}) + \mathrm{Pr}(\mathrm{rain}) \cdot \mathrm{U}(\mathrm{leave}, \mathrm{rain}) = 0.75 \cdot 50 + 0.25 \cdot (-50) = 25.0.$$

The expected utility is therefore maximized by leaving the umbrella at home.
This is the Bayesian decision.
2c Figure 15.1 shows how the optimal Bayesian decision varies for different combinations of the prior hyperparameters.
Figure 15.2 shows how the optimal Bayesian decision varies for different combinations of the prior hyperparameters when $s = 16$ and $f = 64$.

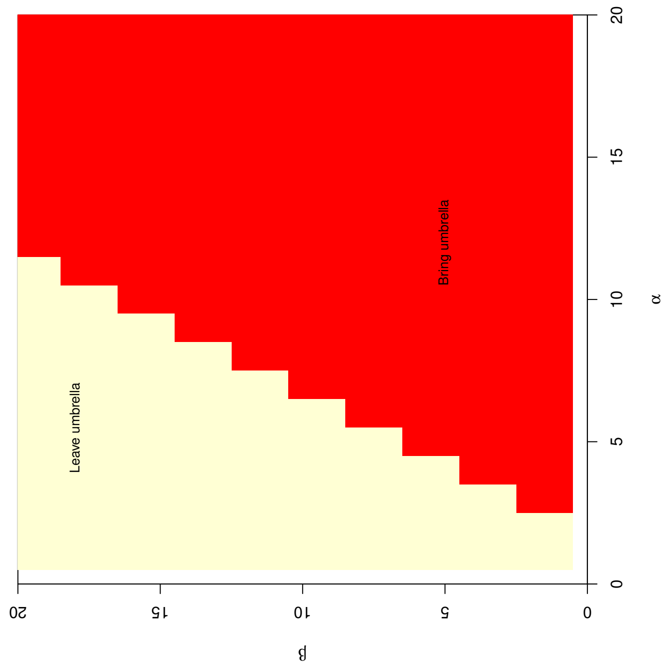


Fig. 15.1. How the Bayesian decision depends on the prior hyperparameters when $s = 2$ and $f = 8$

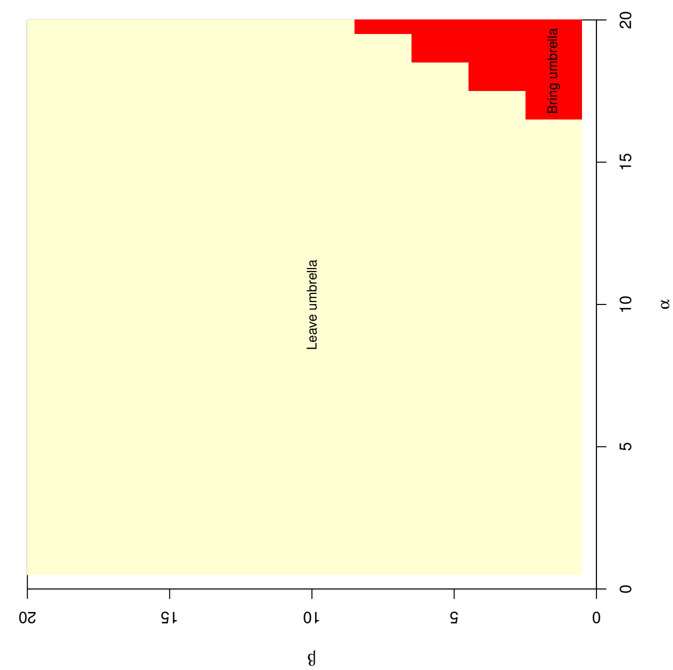


Fig. 15.2. How the Bayesian decision depends on the prior hyperparameters when $s = 16$ and $f = 64$

Solution 3a): The predictive distribution of x_6:

$$x_6|x_{1:5} \sim \mathcal{N}\left(\mu_n, \sigma^2 + \tau_n^2\right)$$

as shown in Lecture 4, slide 6. Here μ_n and ${\tau_n}^2$ are the posterior mean and variance of θ, which were derived in Lecture 2. So,

$$\tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{1}{\frac{1}{50^2} + \frac{5}{25^2}} = 119,$$

$$\mu_n = w\bar{x} + (1-w)\,\mu_0$$

$$\text{with } w = \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{\frac{5}{25^2}}{\frac{1}{50^2} + \frac{5}{25^2}} = 0.95$$

$$\text{so } \mu_n = 0.95 \cdot 320.4 + 0.05 \cdot 200 = 315.$$

So,

$$x_6|x_{1:5} \sim \mathcal{N}\left(315, 25^2 + 119\right) = \mathcal{N}\left(315, 27.3^2\right)$$

Solution 3b): The expected utility when there is no campaign is

$$E\left[U\left((p-q)\,x_6\right)|x_{1:5}\right] = E\left[1 - \exp\left(-5x_6/1000\right)|x_{1:5}\right] = 1 - E\left[\exp\left(S_1\right)\right],$$

where S_1 is a normal random variable with mean $-5 \cdot 315/1000 = -1.575$ and standard deviation $5 \cdot 27.3/1000 = 0.1365$. So the expected utility is

$$1 - E\left[\exp\left(S_1\right)\right] = 1 - \exp\left(-1.575 + 0.1365^2/2\right) = 0.7911.$$

The expected utility when there is a campaign is

$$E\left[U\left(1.2\,(p-q)\,x_6 - 300\right)|x_{1:5}\right] = E\left[1 - \exp\left(-\left(1.2 \cdot 5x_6 - 300\right)/1000\right)|x_{1:5}\right] = 1 - E\left[\exp\left(S_2\right)\right],$$

where S_2 is a normal random variable with mean $-\left(1.2 \cdot 5 \cdot 315 - 300\right)/1000 = -1.59$ and standard deviation $1.2 \cdot 5 \cdot 27.3/1000 = 0.1638$. So the expected utility is

$$1 - E\left[\exp\left(S_2\right)\right] = 1 - \exp\left(-1.59 + 0.1638^2/2\right) = 0.7933.$$

Since the expected utility of running the campaign is higher, this is what the company should do.

Solution 4a): Let $y_i \overset{iid}{\sim} Poi\left(\theta\right)$ be the number of fatal accidents for year $i = 1, \ldots, 10$. Then,

$$p\left(y_i|\theta\right) = \frac{1}{y!}\theta^{y_i}\exp\left(-\theta\right).$$

Use a conjugate gamma prior for $\theta \sim Gamma\left(\alpha, \beta\right)$ so that $p\left(\theta\right) \propto \theta^{\alpha-1}\exp\left(-\beta\theta\right).$ Now

$$p\left(\theta|y\right) \propto p\left(\theta\right)\prod_{i=1}^{10} p\left(y_i|\theta\right) \propto \theta^{\alpha-1}\exp\left(-\beta\theta\right)\theta^{\sum_{i=1}^{10} y_i}\exp\left(-10\theta\right)$$

$$\propto \theta^{\alpha+10\bar{y}-1}\exp\left(-\left(\beta+10\right)\theta\right).$$

So, we get the posterior $\theta|y \sim Gamma\left(\alpha + 10\bar{y}, \beta + 10\right).$ We compute $\bar{y} = 23.8$ and a non-informative (non-proper) gamma prior is obtained from $\alpha = 0,\ \beta = 0$, so that $\theta|y \sim Gamma\left(238, 10\right).$ To get the 95% predictive bands we can use either normal approximation or simulation. For the normal approximation, we need to compute the predictive mean and variance for a new observation

$$\tilde{y} : E\left(\tilde{y}|y\right) = E\left[E\left(\tilde{y}|\theta, y\right)|y\right] = E\left(\theta|y\right) = \frac{238}{10} = 23.8,$$

$$\begin{aligned} Var\left(\tilde{y}|y\right) &= E\left[Var\left(\tilde{y}|\theta, y\right)|y\right] + Var\left[E\left(\tilde{y}|\theta, y\right)|y\right] \\ &= E\left[\theta|y\right] + Var\left[\theta|y\right] = \frac{238}{10} + \frac{238}{10^2} = 26.18, \end{aligned}$$

where we have used the formulas for means and variances of conditional distributions on page 21 in the coursebook, and that $E(X) = Var(X) = \theta$ for $X \sim Poi(\theta)$ and $E(X) = \alpha/\beta$ and $Var(X) = \alpha/\beta^2$ for $X \sim Gamma\left(\alpha, \beta\right)$. A normal approximation of the posterior is

$$\tilde{y}|y \sim N\left(E\left(\tilde{y}|y\right), Var\left(\tilde{y}|y\right)\right) = N\left(23.8, 26.18\right).$$

A 95% predictive interval is

$$23.8 \pm 1.96 \cdot \sqrt{26.18} \Rightarrow 13.77 < \tilde{y} < 33.83.$$

Alternatively, using simulation, a predictive interval can be computed by repeatedly simulating $\theta^{(j)}$ from $\theta|y$ and then $\tilde{y}^{(j)} \sim \tilde{y}|\theta^{(j)}$ for $j = 1, \ldots, 1000$ and extracting the 2.5th and 97.5th percentile from the samples. The following code does this in R:

```
theta = rgamma(1000,238,10)
y = rpois(1000,theta)
quantile(y,probs=c(0.025,0.975))
```

LINKÖPING UNIVERSITY
Division of Statistics and Machine Learning
Department of Computer and Information Science
Mattias Villani and Per Sidén

Mathematical Exercises 3

Try to solve the problems before class. Don't worry if you fail, the important thing is trying.
You should not hand in any solutions.
This part of the course is not obligatory and is not graded.

---

1. Fill in the blanks
   (a) Show that the full conditional posterior of I_i on Lecture 7, Slide 21, is correct.

2. Frequentist meltdown or Bayesian breakdown?
   (a) Let $x_1, ..., x_n \overset{iid}{\sim} Uniform(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. Let $\hat{\theta} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ be an estimator of θ. Derive an expression for the (repeated) sampling variance of $\hat{\theta}$.
   (b) Derive the posterior distribution for θ assuming a uniform prior distribution. [Hint: Here it absolutely crucial to think about the support for the data distribution. Once you have observed some data, some θ values are no longer possible. I strongly suggest that you plot some imaginary data on the real line and plot the data distribution in the same graph for some made up values of θ. Just to make you think in the right direction.]
   (c) Assume that you have observed three data observations: $x_1 = 1.1, x_2 = 2.09, x_3 = 1.4$. What would a frequestist conclude about θ? What would a Bayesian conclude? Discuss.

3. Who doesn't want to be Normal?
   (a) Let $x_1, ..., x_n \overset{iid}{\sim} \text{Bern}(\theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$ a priori. Find the posterior mode of θ.
   (b) Approximate the posterior distribution of θ by a normal distribution.
   (c) Assume now that you have the data $n = 6$ and $s = 1$. Plot the true posterior distribution and the normal approximation in the same graph. Assume a uniform prior for θ.
   (d) Redo the previous exercise, but this time with twice the data size: $n = 12$ and $s = 2$.

4. Naive doctors

(a) Three diseases (A,B and C) have very common symptoms and are therefore hard to distinguish between for a doctor. A medical company has developed two different tests (T1 and T2) to discriminate between the three diseases. A training data from $n = 20$ patients was collected to learn a predictive model that can be used to classify a patient into disease A-C on the basis of the results from both T1 and T2. $n_A = 5$ of the patients had disease A, $n_B = 5$ of the patients had disease B and $n_C = 10$ of the patients had disease C. The table below gives the mean measurement in each patient group for both tests. The test measurements can be assumed to follow a normal distribution with variance $\sigma^2 = 1$ for all patient groups, and for both tests. Develop a Naive Bayes classifier based on this training data. You can assume uniform priors in any place you needs a prior. Make a prediction for a new patient with measurement 1.3 on T1 and 4.2 on T2.

	$\bar{X}_1$	$\bar{X}_2$
Disease A	1.2	2.1
Disease B	1.4	3.5
Disease C	0.7	4.7

Have fun!

LINKÖPING UNIVERSITY
Division of Statistics and Machine Learning
Department of Computer and Information Science
Mattias Villani and Per Sidén

Bayesian Learning

Mathematical Exercises 3

Try to solve the problems before class. Don't worry if you fail, the important thing is trying.
You should not hand in any solutions.
This part of the course is not obligatory and is not graded.

---

1. Fill in the blanks

   (a) Show that the full conditional posterior of I_i on Lecture 7, Slide 21, is correct.
   Solution: (a) By Bayes' theorem, the full conditional posterior of I_i is

$$p(I_i = 1|\mathbf{x}, \cdot) \propto p(\mathbf{x}|I_i = 1, I_{-i}, \cdot)p(I_i = 1),$$

   where $\cdot$ is a shorthand for all other model parameters: $\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2$. The symbol I_{-i} denotes all allocation variables except for I_i. Now, since the data are iid we have $p(\mathbf{x}|I_i, I_{-i}, \cdot) = \prod_{j=1}^{n} p(x_j|I_j, \cdot)$. Note that only the factor $p(x_i|I_i, \cdot)$ in the product depends on I_i, and the other $n-1$ factors can therefore be moved into the proportionality constant. We therefore have

$$p(I_i = 1|\mathbf{x}, I_{-i}, \cdot) \propto \phi(x_i|\mu_1, \sigma_1^2)\pi,$$

   where $\phi(x_i|\mu_1, \sigma_1^2)$ denotes the pdf of a $N(\mu_1, \sigma_1^2)$ variable evaluated in the point x_i. By exactly the same reasoning

$$p(I_i = 2|\mathbf{x}, I_{-i}, \cdot) \propto \phi(x_i|\mu_2, \sigma_2^2)(1 - \pi).$$

   All that remains to do is to normalize these two probabilities so that they sum to one, and we get the required result

$$p(I_i = 1|\mathbf{x}, I_{-i}, \cdot) = \frac{\phi(x_i|\mu_1, \sigma_1^2)\pi}{\phi(x_i|\mu_1, \sigma_1^2)\pi + \phi(x_i|\mu_2, \sigma_2^2)(1 - \pi)}.$$

2. Frequentist meltdown or Bayesian breakdown?

   (a) Let $x_1, ..., x_n \overset{iid}{\sim} Uniform(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. Let $\hat{\theta} = \bar{x} = \sum_{i=1}^{n}$ be an estimator of θ. Derive an expression for the (repeated) sampling variance of $\hat{\theta}$.

   (b) Derive the posterior distribution for θ assuming a uniform prior distribution. [Hint: Here it absolutely crucial to think about the support for the data distribution. Once you have observed some data, some θ values are no longer possible. I strongly suggest that you plot some imaginary data on the real line and plot the data distribution in the same graph for some made up values of θ. Just to make you think in the right direction.]

(c) Assume that you have observed three data observations: $x_1 = 1.1, x_2 = 2.09, x_3 = 1.4$. What would a frequestist conclude about θ? What would a Bayesian conclude? Discuss.

Solution: See sketched solution in the end of this document.

3. Who doesn't want to be Normal?

(a) Let $x_1, ..., x_n \stackrel{iid}{\sim} \text{Bern}(\theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$ a priori. Find the posterior mode of θ.

(b) Approximate the posterior distribution of θ by a normal distribution.

(c) Assume now that you have the data $n = 6$ and $s = 1$. Plot the true posterior distribution and the normal approximation in the same graph. Assume a uniform prior for θ.

(d) Redo the previous exercise, but this time with twice the data size: $n = 12$ and $s = 2$.

Solution: See sketched solution in the end of this document.

4. Naive doctors

(a) Three diseases (A,B and C) have very common symptoms and are therefore hard to distinguish between for a doctor. A medical company has developed two different tests (T1 and T2) to discriminate between the three diseases. A training data from $n = 20$ patients was collected to learn a predictive model that can be used to classify a patient into disease A-C on the basis of the results from both T1 and T2. $n_A = 5$ of the patients had disease A, $n_B = 5$ of the patients had disease B and $n_C = 10$ of the patients had disease C. The table below gives the mean measurement in each patient group for both tests. The test measurements can be assumed to follow a normal distribution with variance $\sigma^2 = 1$ for all patient groups, and for both tests. Develop a Naive Bayes classifier based on this training data. You can assume uniform priors in any place you needs a prior. Make a prediction for a new patient with measurement 1.3 on T1 and 4.2 on T2.

	$\bar{X}_1$	$\bar{X}_2$
Disease A	1.2	2.1
Disease B	1.4	3.5
Disease C	0.7	4.7

Solution: Let us first focus on the predictive probability for disease A given the outcomes from the two tests T_1 and T_2

$$\begin{aligned}\Pr(A|T_1, T_2) &\propto \Pr(T_1, T_2|A)\Pr(A) \\ &= \Pr(T_1|A)\Pr(T_2|A)\Pr(A),\end{aligned}$$

where the latter equality is a result of the simplifying naive Bayes assumption that features (the tests) are independent conditional on the class (the disease). Similar expression hold for $\Pr(B|T_1, T_2)$ and $\Pr(C|T_1, T_2)$. Let us first estimate the class probabilities $P(A)$, $P(B)$ and $P(C)$. This estimation problem is a Multinomial-Dirichlet problem with $n = 20$ trials ending up in the categories: $n_A = 5, n_B = 5$ and $n_C = 10$. We can therefore obtain the posterior distribution for the vector $(\Pr(A), \Pr(B), \Pr(C))$ as a Dirichlet distribution. However, Naive Bayes typically uses just a point estimate of

$(\Pr(A), \Pr(B), \Pr(C))$, and one immediate candidate for a point estimator is the posterior mean. The uniform prior distribution here is the Dirichlet$(1, 1, 1)$ distribution, and for the observed data, the posterior mean vector of $(\Pr(A), \Pr(B), \Pr(C))$ is

$$\left(\frac{n_A + 1}{n + 3}, \frac{n_B + 1}{n + 3}, \frac{n_C + 1}{n + 3}\right) = \left(\frac{6}{23}, \frac{6}{23}, \frac{11}{23}\right).$$

Now, the class conditional feature distributions are $T_1|A \sim N(\bar{x}_{1A}, 1{+}1/n_A)$ and similarly for disease B and C. Note that this is the predictive distribution in the normal model with known variance equal to one, and a uniform prior for mean. [That is, the old result from Lecture 4: $\tilde{y}|\mathbf{y} \sim N(\bar{y}, \sigma^2(1 + 1/n_A))$ with $\sigma = 1$]. So, we have for disease A:

$$\begin{aligned}\Pr(A|T_1, T_2) &\propto \Pr(T_1|A)\Pr(T_2|A)\Pr(A) \\ &= \phi(1.3, \mu = 1.2, \sigma^2 = 1 + 1/5) \cdot \phi(4.2, \mu = 2.1, \sigma^2 = 1 + 1/5) \cdot \frac{6}{23},\end{aligned}$$

where $\phi(x, \mu, \sigma^2)$ is the pdf of a normal distribution with mean μ and variance σ^2 evaluated in the point x. The predictive distribution for the new patient is obtained by repeating this type of calculation also for $\Pr(B|T_1, T_2)$ and $\Pr(C|T_1, T_2)$ and normalizing so that the three probabilities sum to one. The predictive distribution for the new patient is plotted in Figure 1. Figure 2 displays the estimated feature distribution for each disease and the observed T_1 (left) and T_2 (right) for the new patient. Note how T_1 gives somewhat more support for disease A and B, but that the observed value for T_2 is extremely improbable for a patient with disease A. This explains why disease B gets the highest predictive probability, and A gets a predictive probability close to zero.

Have fun!

- Mattias

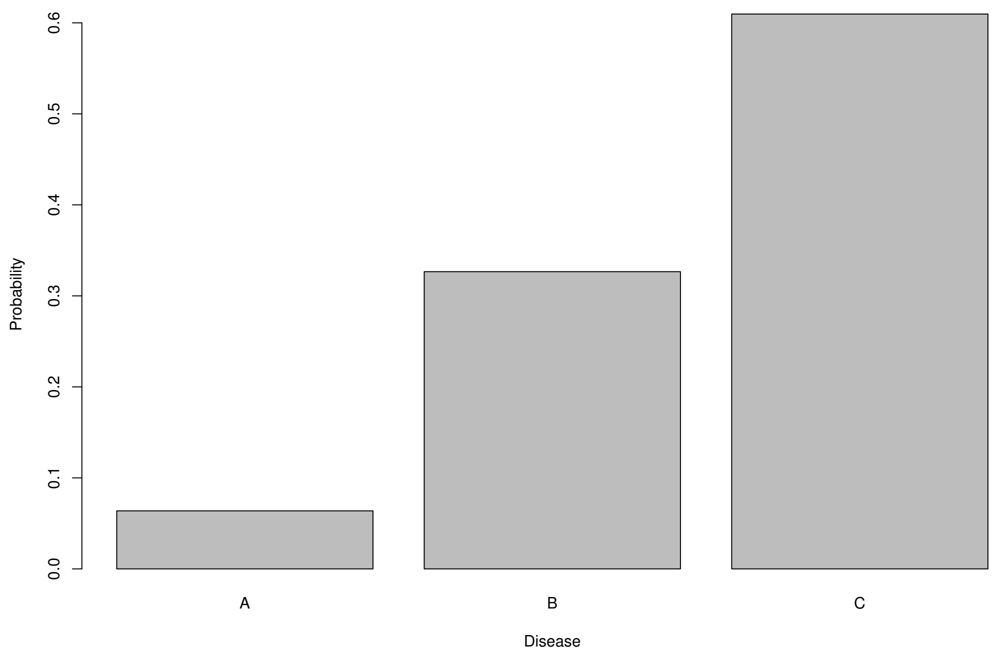
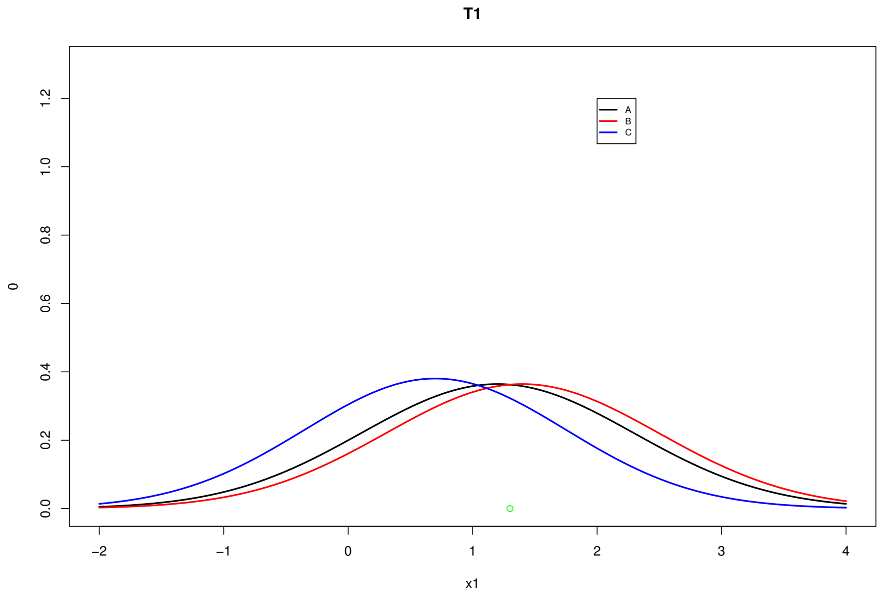


Figure 1: Estimated class conditional distributions for test T_1 and the value for T_1 for the predicted patient (as green circle).



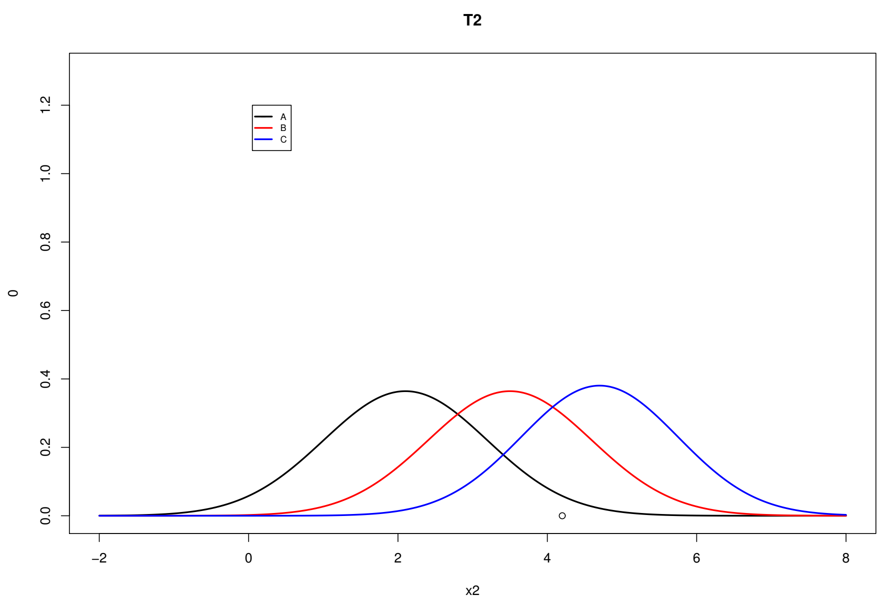
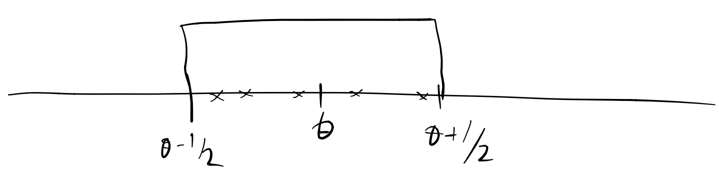


Figure 2: Estimated class conditional distributions for test T_1 (left) and T_2 (right). The value for T_1 and T_2 for the predicted patient are plotted as a green circles in respective graphs.

2a)



$\mathrm{Var}(\bar{X}) = \dfrac{\sigma^2}{n}$ $\quad \sigma^2 = \mathrm{Var}(X_i) \quad X_i \sim U(\theta - 1/2, \theta + 1/2)$

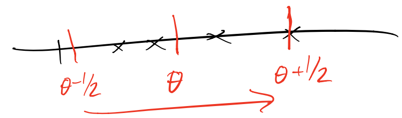
$X \sim U(0,1) \quad \mathrm{Var}(X) = \dfrac{1}{12}$

So $\mathrm{Var}(\bar{x}) = \dfrac{1}{12n}$

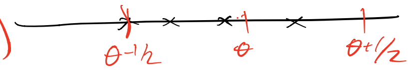
2b) $P(\theta | x_1, \dots, x_n) \propto P(x_1 \dots x_n | \theta) P(\theta)$

$$= \prod_{i=1}^{n} P(x_i | \theta) P(\theta)$$

$$= \prod_{i=1}^{n} I\left(\theta - \tfrac{1}{2} \le x_i \le \theta + \tfrac{1}{2}\right) \cdot 1$$



$\theta + 1/2 \ge x_{max}$
$\theta - 1/2 \le x_{min}$ $\Rightarrow \theta \in [x_{max} - 1/2, x_{min} + 1/2]$



$$P(\theta | x_1 \dots x_n) \propto 1 \quad \text{for } \theta \in [x_{max} - 1/2, x_{min} + 1/2]$$
$$= 0 \quad \text{otherwise.}$$

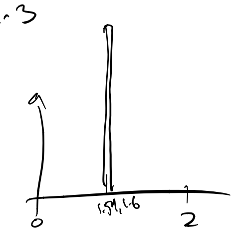
$$\theta | x_1 \dots x_n \sim U(x_{max} - 1/2, x_{min} + 1/2)$$

2c) Frequentist : $\hat{\theta} = \bar{X} = 1.53$

$$Var(\hat{\theta}) = \frac{1}{12n} = \frac{1}{12 \cdot 3} = 0.027777$$

$$SD(\hat{\theta}) = 0.1666$$

Bayesian : $\theta | x_1 x_2 x_3 \sim U(1.59, 1.6)$



3 a) $\theta | x_1 \ldots x_n \sim Beta(\alpha+s, \beta+f)$

$$p(\theta|x) \propto \theta^{\alpha+s-1}(1-\theta)^{\beta+f-1} \Rightarrow \ln p(\theta|x) \propto (\alpha+s-1)\ln\theta + (\beta+f-1)\ln(1-\theta)$$

$$\frac{\partial \ln p(\theta|x)}{\partial\theta} = \frac{\alpha+s-1}{\theta} + \frac{\beta+f-1}{1-\theta}(-1)$$

$$\frac{\partial \ln p(\theta|x)}{\partial\theta} = 0 \Rightarrow \frac{\alpha+s-1}{\theta} = \frac{\beta+f-1}{1-\theta}$$

$$\Rightarrow \hat{\theta} = \frac{\alpha+s-1}{\alpha+\beta+n-2}$$

3b) $\theta|x_1 \ldots x_n \overset{approx}{\sim} N(\hat{\theta}, -J_{\hat{\theta}|x}^{-1})$ $\quad 1-\hat{\theta} = \frac{\beta+f-1}{\alpha+\beta+n-2}$

$$\frac{\partial^2 \ln p(x|\theta)}{\partial\theta^2} = -\frac{\alpha+s-1}{\theta^2} + \frac{\beta+f-1}{(1-\theta)^2}(-1)$$

$$\left.\frac{\partial^2 \ln p(x|\theta)}{\partial\theta^2}\right|_{\theta=\hat{\theta}} = -\left(\frac{\alpha+s-1}{\left(\frac{\alpha+s-1}{\alpha+\beta+n-2}\right)^2} + \frac{\beta+f-1}{\left(\frac{\beta+f-1}{\alpha+\beta+n-2}\right)^2}\right)$$

$$= -(\alpha+\beta+n-2)^2\left(\frac{1}{\alpha+s-1} + \frac{1}{\beta+f-1}\right)$$

$$= -(\alpha+\beta+n-2)^2\left(\frac{\alpha+\beta+n-2}{(\alpha+s-1)(\beta+f-1)}\right)$$

$$= -\frac{(\alpha+\beta+n-2)^3}{(\alpha+s-1)(\beta+f-1)}$$

$$\theta \mid x_1 \dots x_n \overset{\text{approx}}{\sim} N\left(\hat{\theta} = \frac{\alpha + s - 1}{\alpha + \beta + n - 2},\ -J^{-1}_{\hat{\theta}, x} = \frac{(\alpha+s-1)(\beta+f-1)}{(\alpha+\beta+n-2)^3} \right)$$

$\theta \sim \text{Beta}(\alpha+s, \beta+f)$

Check: $$\text{Var}(\theta) = \frac{(\alpha+s)(\beta+f)}{(\alpha+s+\beta+f)^2(\alpha+s+\beta+f+1)}$$

$$= \frac{(\alpha+s)(\beta+f)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$$

3 c) Se R-code

3 d) —— " ——

LINKÖPING UNIVERSITY
Division of Statistics and Machine Learning
Department of Computer and Information Science
Mattias Villani and Per Sidén

Bayesian Learning

Mathematical Exercises 4

Try to solve the problems before class. Don't worry if you fail, the important thing is trying.
You should not hand in any solutions.
This part of the course is not obligatory and is not graded.

---

1. Bernoulli meets Laplace
   (a) Let $x_1, ..., x_n \overset{iid}{\sim} Bernoulli(\theta)$ and assume the prior $\theta \sim Beta(\alpha, \beta)$. Derive the marginal likelihood for this model.
   (b) Compute the marginal likelihood of the model in a) using the Laplace approximation.
   (c) Is this approximation accurate if $\alpha = \beta = 1$ and you have observed $s = 6$ success in $n = 10$ trials?

2. Fill in the blanks - again
   (a) Derive the marginal likelihood for the Poisson model with Gamma prior at the end of Slide 6 at Lecture 10.

3. Pareto
   (a) Let $x_1, \ldots, x_n \overset{iid}{\sim} \text{Uniform}(0, \theta)$. Let $\theta \sim \text{Pareto}(\alpha, \beta)$, that is
   $$p(\theta) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}},\ \theta \geq \beta.$$
   Show that this is a conjugate prior to this Uniform model and derive the posterior for θ. [Hint: Don't forget to include an indicator function when you write up the likelihood function. The $\text{Uniform}(0, \theta)$ distribution is zero for outcomes larger than θ.]
   (b) Derive the predictive distribution of x_{n+1} given $x_1, \ldots, x_n$. [Hint: It is wise to break up the integrals in two parts.]

Have fun!

- Mattias

LINKÖPING UNIVERSITY
Division of Statistics and Machine Learning
Department of Computer and Information Science
Mattias Villani and Per Sidén

Bayesian Learning

Mathematical Exercises 4

Try to solve the problems before class. Don't worry if you fail, the important thing is trying.
You should not hand in any solutions.
This part of the course is not obligatory and is not graded.

1. Bernoulli meets Laplace

(a) Let $x_1, ..., x_n \overset{iid}{\sim} Bernoulli(\theta)$ and assume the prior $\theta \sim Beta(\alpha, \beta)$. Derive the marginal likelihood for this model.

Solution: Let $\boldsymbol{y} = (y_1, \ldots, y_n)$. The marginal likelihood is

$$\begin{aligned}
p(\boldsymbol{y}) &= \int p(\boldsymbol{y}|\theta)p(\theta)d\theta \\
&= \int \theta^s(1-\theta)^f \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\
&= \frac{1}{B(\alpha,\beta)}\int \theta^{\alpha+s-1}(1-\theta)^{\beta+f-1}d\theta \\
&\text{[The integral is with respect to the kernel of a Beta}(a+s, b+f)\text{ density]} \\
&= \frac{B(\alpha+s,\beta+f)}{B(\alpha,\beta)}
\end{aligned}$$

(b) Compute the marginal likelihood of the model in a) using the Laplace approximation.

Solution: The Laplace approximation of a log marginal likelihood is

$$\ln \hat{p}(\boldsymbol{y}) = \ln p(\boldsymbol{y}|\hat{\theta}) + \ln p(\hat{\theta}) + (1/2)\ln\left|J^{-1}_{\boldsymbol{y},\hat{\theta}}\right| + (1/2)\ln(2\pi),$$

where $\hat{\theta}$ is the posterior mode and $J_{\boldsymbol{y},\hat{\theta}}$ is minus the second derivative at the mode. Now,

$$\begin{aligned}
\ln p(\theta|\boldsymbol{y}) &= -\text{Beta}(a+s, b+f) + (\alpha+s-1)\ln\theta + (\beta+f-1)\ln(1-\theta) \\
\frac{d\ln p(\theta|\boldsymbol{y})}{d\theta} &= \frac{\alpha+s-1}{\theta} - \frac{\beta+f-1}{1-\theta} \\
\frac{d^2\ln p(\theta|\boldsymbol{y})}{d\theta^2} &= -\frac{\alpha+s-1}{\theta^2} - \frac{\beta+f-1}{(1-\theta)^2}
\end{aligned}$$

Solving $d\ln p(\theta|\boldsymbol{y})/d\theta = 0$ for θ gives the posterior mode

$$\hat{\theta} = \frac{\alpha+s-1}{\alpha+\beta+n-2}.$$

and therefore

$$J_{\boldsymbol{y},\hat{\theta}}^{-1} = -\left[\frac{d^2 \ln p(\theta|\boldsymbol{y})}{d\theta^2}|_{\theta=\hat{\theta}}\right]^{-1} = \frac{(\alpha+s-1)(\beta+f-1)}{(\alpha+\beta+n-2)^3}.$$

(c) Is this approximation accurate if $\alpha = \beta = 1$ and you have observed $s = 6$ success in $n = 10$ trials?

Solution: For this data we have $\hat{\theta} = s/n = 0.6$ and $J_{\boldsymbol{y},\hat{\theta}}^{-1} = sf/n^3 = 0.024$. So, $\ln \hat{p}(\boldsymbol{y}) = -7.676$ which is quite close to the true log marginal likelihood $\ln p(\boldsymbol{y}) = -7.745$. One way to see that the approximation is accurate is for example to see the differences in comparing a model with a null model where $\theta = 1/2$.

2. Fill in the blanks - again

(a) Derive the marginal likelihood for the Poisson model with Gamma prior at the end of Slide 6 at Lecture 10.

Solution: The marginal likelihood is

$$\begin{aligned}
p(x_1,\ldots,x_n) &= \int p(x_1,\ldots,x_n|\theta)p(\theta)d\theta \\
&= \int \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} d\theta \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)\prod_{i=1}^n x_i!} \int \theta^{\alpha+\sum_{i=1}^n x_i - 1} e^{-(\beta+n)\theta} d\theta \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)\prod_{i=1}^n x_i!} \frac{\Gamma(\alpha+\sum_{i=1}^n x_i)}{(\beta+n)^{\alpha+\sum_{i=1}^n x_i}} \\
&= \frac{\beta^\alpha}{(\beta+n)^{\alpha+n\bar{x}}} \frac{\Gamma(\alpha+n\bar{x})}{\Gamma(\alpha)\prod_{i=1}^n x_i!}
\end{aligned}$$

3. Pareto

(a) Let $x_1,\ldots,x_n \overset{iid}{\sim} \text{Uniform}(0,\theta)$. Let $\theta \sim \text{Pareto}(\alpha,\beta)$, that is

$$p(\theta) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}},\ \theta \geq \beta.$$

Show that this is a conjugate prior to this Uniform model and derive the posterior for θ. [Hint: Don't forget to include an indicator function when you write up the likelihood function. The Uniform$(0,\theta)$ distribution is zero for outcomes larger than θ.]

Solution: The likelihood function is of the form

$$\prod_{i=1}^{n} \frac{1}{\theta} I(x_i \leq \theta) = \left(\frac{1}{\theta}\right)^n I(x_{\max} \leq \theta)$$

where $x_{\max} = \max(x_1, \ldots, x_n)$, and the Pareto prior is

$$p(\theta) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}} \cdot I(\beta \le \theta).$$

By Bayes' theorem we therefore have

$$\begin{aligned} p(\theta|x_1, \ldots, x_n) &\propto p(x_1, \ldots, x_n|\theta)p(\theta) \\ &= \left(\frac{1}{\theta}\right)^n I(x_{\max} \le \theta)\frac{\alpha\beta^\alpha}{\theta^{\alpha+1}} \cdot I(\beta \le \theta) \\ &= \frac{\alpha\beta^\alpha}{\theta^{n+\alpha+1}} \cdot I\big(\tilde{\beta} \le \theta\big) \end{aligned}$$

where $\tilde{\beta} = \max(x_{\max}, \beta)$. This is proportional to a Pareto$\big(\alpha + n, \tilde{\beta}\big)$ density.

(b) Derive the predictive distribution of x_{n+1} given $x_1, \ldots, x_n$. [Hint: It is wise to break up the integrals in two parts.]

Solution: From a) the posterior distribution is

$$\theta|x_1, \ldots, x_n \sim \text{Pareto}(\alpha + n, \tilde{\beta}),$$

where $\tilde{\beta} = \max(x_{\max}, \beta)$. The predictive distribution is

$$\begin{aligned} p(x_{n+1}|x_{1:n}) &= \int_0^\infty p(x_{n+1}|\theta)p(\theta|x_{1:n})d\theta \\ &= \int_0^\infty \frac{1}{\theta}I(x_{n+1} \le \theta)\frac{(\alpha+n)\tilde{\beta}^{(\alpha+n)}}{\theta^{n+\alpha+1}} \cdot I\big(\tilde{\beta} \le \theta\big)d\theta \\ &= (\alpha+n)\int_0^\infty \frac{\tilde{\beta}^{(\alpha+n)}}{\theta^{n+\alpha+2}}I\big(\max(x_{n+1}, \tilde{\beta}) \le \theta\big)d\theta \end{aligned}$$

In order to compute this integral we will separate the integral in two cases: i) $x_{n+1} \le \tilde{\beta}$ where $\max(x_{n+1}, \tilde{\beta}) = \tilde{\beta}$ and ii) $x_{n+1} > \tilde{\beta}$ where $\max(x_{n+1}, \tilde{\beta}) = x_{n+1}$. Now, when $x_{n+1} \le \tilde{\beta}$, we have

$$\begin{aligned} p(x_{n+1}|x_{1:n}) &= (\alpha+n)\int_0^\infty \frac{\tilde{\beta}^{(\alpha+n)}}{\theta^{n+\alpha+2}}I\big(\max(x_{n+1}, \tilde{\beta}) \le \theta\big)d\theta \\ &= (\alpha+n)\int_{\tilde{\beta}}^\infty \frac{\tilde{\beta}^{(\alpha+n)}}{\theta^{n+\alpha+2}}d\theta \\ &= \frac{\alpha+n}{(\alpha+n+1)\tilde{\beta}}\int_{\tilde{\beta}}^\infty \frac{(\alpha+n+1)\tilde{\beta}^{(\alpha+n+1)}}{\theta^{(n+\alpha+1)+1}}d\theta \\ &= \frac{\alpha+n}{\alpha+n+1}\frac{1}{\tilde{\beta}} \end{aligned}$$

This shows that the predictive distribution for x_{n+1} is $\frac{\alpha+n}{\alpha+n+1} \cdot \text{Uniform}(x_{n+1}|0, \tilde{\beta})$ when

$x_{n+1} \leq \tilde{\beta}$. Turning now to the other case when $x_{n+1} > \tilde{\beta}$ we have

$$
\begin{aligned}
p(x_{n+1}|x_{1:n}) &= (\alpha+n)\int_0^\infty \frac{\tilde{\beta}^{(\alpha+n)}}{\theta^{n+\alpha+2}} I\big(\max(x_{n+1},\tilde{\beta}) \leq \theta\big) d\theta \\
&= (\alpha+n)\int_{x_{n+1}}^\infty \frac{\tilde{\beta}^{(\alpha+n)}}{\theta^{n+\alpha+2}} d\theta \\
&= \frac{(\alpha+n)\tilde{\beta}^{(\alpha+n)}}{(\alpha+n+1)x_{n+1}^{\alpha+n+1}} \int_{x_{n+1}}^\infty \frac{(\alpha+n+1)x_{n+1}^{\alpha+n+1}}{\theta^{n+\alpha+2}} d\theta \\
&= \frac{1}{(\alpha+n+1)} \frac{(\alpha+n)\tilde{\beta}^{(\alpha+n)}}{x_{n+1}^{\alpha+n+1}},
\end{aligned}
$$

which can be recognized as $\frac{1}{\alpha+n+1} \cdot \text{Pareto}(x_{n+1}|\alpha+n,\tilde{\beta})$ In summary,

$$
x_{n+1}|x_{1:n} \sim \begin{cases} \frac{\alpha+n}{\alpha+n+1} \cdot \text{Uniform}(x_{n+1}|0,\tilde{\beta}), & \text{if } x_{n+1} \leq \tilde{\beta} \\ \frac{1}{\alpha+n+1} \cdot \text{Pareto}(x_{n+1}|\alpha+n,\tilde{\beta}), & \text{if } x_{n+1} > \tilde{\beta}, \end{cases}
$$

where $\tilde{\beta} = \max(x_{\max},\beta)$.

Have fun!

- Mattias