Mattias Villani
Stockholm University and Sveriges Riksbank
e-mail: mattias.villani@gmail.com

May 10, 2011

SOLUTIONS TO HOME EXAM NO. 1 IN BAYESIAN STATISTICS I

These are scetchy solutions to the problems in Exam 1.
I solved them by myself without any co-operation!

(1) Comment on the following statements:

(a) *'To a Bayesian, even the speed of light can be a random variable. This is clearly wrong since the speed of light is a natural constant'.*
All that matters for Bayesian inference is whether or not You know the value for the speed of light. Probability is subjective, anything that it at least partly unknown to You should be described by the language of uncertainty, that is by probabilities. So, any partially unknown quantity is treated by Bayesians **as if** it was a random variable. I may therefore even assign a probability distribution to an event or quantity that I had full knowledge about yesterday, but have forgotten today. Also a Bayesian may be interested in the philosophical questions whether or not a phenomena is random in a more deeper, intrinsic, way, but this is irrelevant to the inference question.

(b) *'Bayesianism has no place in science. Everything about it is subjective, everything is allowed.'*
Here one can make a philosophical argument for why we can never know if such a thing as objectivity even exists. One can also argue that supposedly objective statistical methods are riddled with subjective choices. Frequentists find it awkward to assign subjective priors on parameters, but have no problem with subjectively chosing a (class) of models. Also model choice is subjective. But most importantly, Bayesian inference is all about the transition from priors to posteriors. What can we learn from the data? This transition is as objective as it gets: it is dictated by Bayes theorem, hard-core math! So even when the prior is subjective, so You and I can have different priors, all Bayesians use the same rule for updating. Moreover, the subjective prior will have less and less influence as we get more data, so we can in addition find out when diverging prior opionions converge to the same posterior. "Objectivity by subjective consensus" (TM, patent pending).

(c) *'Noninformative priors do not exist'*
Well, they don't. All priors carry some information. It is very hard to come up with

a baseline to which all other prior can be compared to. If a prior is considered non-informative in one parametrization, it might not be for another parametrization. Jeffreys prior tries to solve this latter problem, but is suspicious since it violates the likelihood principle (see my Lecture notes). But some priors are clearly more informative than other, at least if we carefully specify the quantity/parameter of interest. But typically it doesn't really matter that much which "non-informative" prior you choose as they will all be quickly over-ruled by a few data points. It is therefore typically enough to know that a prior is not strongly informative (unless you want it to be, i.e. when you have strong prior information).

(d) *'A Bayesian analysis requires a fully specified likelihood. The Bayesian approach is therefore not applicable when some observations are missing'.*
Not true. Actually, the Bayesian treatment of missing information is very straightforward. A missing observation is simply another unknown quantity, just like a parameter. If something is unknown, just assign a prior to it and estimate it along with the other parameters. This is a sophisticated version of imputation. The computation can easily be solved by Gibbs sampling:
   (i) Update the missing data given the model parameters
   (ii) Update the model parameters given the complete data, including the missing, which you sampled in Step i.

(2) Consider a random sample from the exponential distribution: $x_1, ..., x_n|\theta \overset{iid}{\sim} Exponential(\theta)$, where $E(x) = 1/\theta$.

(a) *What is the natural conjugate prior for θ?*
The natural conjugate prior is a conjugate prior which is proportional to the likelihood function. Here the likelihood is $\prod_{i=1}^{n} \theta \exp(-\theta x_i) = \theta^n \exp(-\theta n\bar{x})$. The natural conjugate prior is therefore a density of the form $\theta^\alpha \exp(-\theta\beta)$, which is the $Gamma(\alpha+1, \beta)$ density. By convention one typically uses the $Gamma(\alpha, \beta)$ instead (because we get rid of the $+1$ in the first argument). In 2b, we will see that this prior actually is conjugate, i.e. that also the posterior belongs to the Gamma family.

(b) *Derive the posterior distribution of θ.*
By Bayes' rule we have

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

where $x = (x_1, ..., x_n)'$ is the full sample. So,

$$p(\theta|x) \propto \theta^n \exp(-\theta n\bar{x})\theta^{\alpha-1} \exp(-\theta\beta)$$
$$\theta^{n+\alpha-1} \exp[\theta(n\bar{x} + \beta)],$$

which is the $Gamma(\alpha+n, \beta+n\bar{x})$. This also shows that the $Gamma(\alpha, \beta)$ prior is conjugate.

(c) *Derive Jeffreys' prior for θ.*

Jeffreys prior is

$$p(\theta) \propto [I(\theta)]^{1/2}$$

where $I(\theta) = -E_{x|\theta}\left[\frac{\partial^2}{\partial\theta^2}\ln p(x|\theta)\right]$. Now,

$$\frac{\partial^2}{\partial\theta^2}\ln p(x|\theta) = \frac{\partial^2}{\partial\theta^2}(\ln\theta + x\theta) = -\frac{1}{\theta^2}.$$

The Jeffreys prior is therefore

$$p(\theta) \propto [I(\theta)]^{1/2} = \frac{1}{\theta}.$$

(3) Consider a sample from the uniform model: $x_1, ..., x_n|\theta \overset{iid}{\sim} U(0,\theta)$, where $\theta > 0$.

(a) *Show that the Pareto(α,β) density is the natural conjugate prior for θ.*

The likelihood function is of the form

$$p(x_1, ..., x_n|\theta) = \left(\frac{1}{\theta}\right)^n I(x_{max} \leq \theta),$$

where $I(A)$ is an indicator function for the event A and x_{max} is the largest of the observations. The Pareto density is of the form

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

that is the Pareto density has support $\theta \in [\beta, \infty)$. The posterior given the Pareto prior is

$$\begin{aligned} p(\theta|x_1, ..., x_n) &\propto \left(\frac{1}{\theta}\right)^n I(x_{max} \leq \theta)\frac{\alpha\beta^\alpha}{\theta^{\alpha+1}} \cdot I(\theta \geq \beta) \\ &= \frac{\alpha\beta^\alpha}{\theta^{n+\alpha+1}} \cdot I(\theta \geq x_{max} and \theta \geq \beta) \\ &= \frac{\alpha\beta^\alpha}{\theta^{n+\alpha+1}} \cdot I\left[\theta \geq \max(x_{max}, \beta)\right] \\ &\propto Pareto\left[\theta|\alpha + n, \tilde{\beta}\right], \end{aligned}$$

where $\tilde{\beta} = \max(x_{max}, \beta)$. Hence, the prior and posterior both belong to the Pareto family, and the Pareto prior is a natural conjugate prior for the iid $U(0,\theta)$ model.

(b) *Derive the posterior distribution of θ.*

The posterior for θ is the Pareto$\left[\theta|\alpha + n, \tilde{\beta}\right]$ density.

(c) *Derive Jeffreys' prior for θ.*

Jeffreys prior is

$$p(\theta) \propto [I(\theta)]^{1/2}$$

where $I(\theta) = -E_{x|\theta}\left[\frac{\partial^2}{\partial\theta^2}\ln p(x|\theta)\right]$. Now,

$$I(\theta) = -\int_0^\infty \left[\frac{\partial^2}{\partial\theta^2}\ln p(x|\theta)\right] p(x|\theta)dx = -\int_0^\theta \left[\frac{\partial^2}{\partial\theta^2}\ln p(x|\theta)\right] p(x|\theta)dx,$$

since $p(x|\theta) = 0$ for $x > \theta$. Now, for $\theta > x$ (the derivative does not exist at $\theta = x$, but this point is of measure zero, i.e. it does not matter here) we have

$$\frac{\partial^2}{\partial\theta^2} \ln p(x|\theta) = \frac{\partial^2}{\partial\theta^2}(-\ln\theta) = \frac{1}{\theta^2}.$$

The Jeffreys prior is therefore

$$p(\theta) \propto [I(\theta)]^{1/2} = \left(-\frac{1}{\theta}\right)^{1/2}.$$

This density is complex-valued, so it cannot be used. [Open question: Does this model violate regularity conditions that necessary for deriving Jeffrey's prior. I need to find out ... Or did I do an algebraic mistake?]

(d) *Derive the predictive distribution of x_{n+1} given $x_1, ..., x_n$, using the natural conjugate prior.*

The predictive density of x_{n+1} is

$$p(x_{n+1}|x_1, ..., x_n) = \int p(x_{n+1}|\theta)p(\theta|x_1, ..., x_n)d\theta$$

$$p(x_{n+1}|\theta)p(\theta|x_1, ..., x_n) = \left(\frac{1}{\theta}\right) I(\theta \geq x_{n+1}) \frac{(\alpha+n)\tilde{\beta}^{\alpha+n}}{\theta^{\alpha+n+1}} \cdot I(\theta \geq \tilde{\beta})$$

Integrating with respect to θ. Assume first that $x_{n+1} \leq \tilde{\beta}$, so that $I(\theta \geq x_{n+1})$ is unity for all $\theta \in [\tilde{\beta}, \infty)$. Thus, for $x_{n+1} \leq \tilde{\beta}$ we have

$$\begin{aligned} p(x_{n+1}|x_1, ..., x_n) &= \int_0^\infty \left(\frac{1}{\theta}\right) I(\theta \geq x_{n+1}) \frac{(\alpha+n)\tilde{\beta}^{\alpha+n}}{\theta^{\alpha+n+1}} \cdot I(\theta \geq \tilde{\beta}) d\theta \\ &= (\alpha+n) \int_0^\infty \frac{\tilde{\beta}^{\alpha+n}}{\theta^{\alpha+n+2}} \cdot I[\theta \geq \max(x_{n+1}, \tilde{\beta})] d\theta. \end{aligned}$$

To be able to perform this integration we need to separate the two cases: i) $x_{n+1} \leq \tilde{\beta}$ wheremax$(x_{n+1}, \tilde{\beta}) = \tilde{\beta}$ and ii) $x_{n+1} > \tilde{\beta}$ wheremax$(x_{n+1}, \tilde{\beta}) = x_{n+1}$. This means that the predictive distribution will have different shapes over the interval $x_{n+1} \leq \tilde{\beta}$ and the interval $x_{n+1} > \tilde{\beta}$.

First, we deal with the case $x_{n+1} \leq \tilde{\beta}$. Here

$$\begin{aligned} p(x_{n+1}|x_1, ..., x_n) &= (\alpha+n) \int_0^\infty \frac{\tilde{\beta}^{\alpha+n}}{\theta^{\alpha+n+2}} \cdot I[\theta \geq \max(x_{n+1}, \tilde{\beta})] d\theta. \\ &= (\alpha+n) \int_{\tilde{\beta}}^\infty \frac{\tilde{\beta}^{\alpha+n}}{\theta^{\alpha+n+2}} d\theta \\ &= \frac{\alpha+n}{(\alpha+n+1)\tilde{\beta}} \int_{\tilde{\beta}}^\infty \frac{(\alpha+n+1)\tilde{\beta}^{\alpha+n+1}}{\theta^{(\alpha+n+1)+1}} d\theta \\ &= \frac{\alpha+n}{(\alpha+n+1)} \frac{1}{\tilde{\beta}} \end{aligned}$$

since the integrand is recognized as Pareto$(\alpha+n+1, \tilde{\beta})$ pdf and therefore integrates to one. This shows that the predictive density for $x_{n+1} \leq \tilde{\beta}$ is $\frac{\alpha+n}{\alpha+n+1} U(x_{n+1}|0, \tilde{\beta})$.

For $x_{n+1} > \tilde{\beta}$ we have $I[\theta \geq \max(x_{n+1}, \tilde{\beta})] = I[\theta \geq \max(x_{n+1})]$, so

$$\begin{aligned} p(x_{n+1}|x_1, ..., x_n) &= (\alpha + n) \int_0^\infty \frac{\tilde{\beta}^{\alpha+n}}{\theta^{\alpha+n+2}} \cdot I[\theta \geq \max(x_{n+1}, \tilde{\beta})] d\theta \\ &= (\alpha + n) \int_{x_{n+1}}^\infty \frac{\tilde{\beta}^{\alpha+n}}{\theta^{\alpha+n+2}} d\theta \\ &= \frac{(\alpha + n)\tilde{\beta}^{\alpha+n}}{(\alpha + n + 1)x_{n+1}^{(\alpha+n+1)}} \int_{x_{n+1}}^\infty \frac{(\alpha + n + 1)x_{n+1}^{(\alpha+n+1)}}{\theta^{\alpha+n+2}} d\theta \\ &= \frac{1}{\alpha + n + 1} \frac{(\alpha + n)\tilde{\beta}^{\alpha+n}}{x_{n+1}^{(\alpha+n+1)}}, \end{aligned}$$

since the integrand can be recognized as a Pareto$(\alpha+n+1, x_{n+1})$ pdf an therefore integrates to one. From the expression above we recognize the predictive density when $x_{n+1} > \tilde{\beta}$ as the $\frac{1}{\alpha+n+1} Pareto(x_{n+1}|\alpha + n, \tilde{\beta})$ distribution.
In summary, the predictive density is

$$x_{n+1}|x_1, ..., x_n \sim \begin{cases} \frac{\alpha+n}{\alpha+n+1} Uniform(x_{n+1}|0, \tilde{\beta}) & \text{if } x_{n+1} \leq \tilde{\beta} \\ \frac{1}{\alpha+n+1} Pareto(x_{n+1}|\alpha + n, \tilde{\beta}) & \text{if } x_{n+1} > \tilde{\beta} \end{cases}$$

(4) Late train arrivals has been a nuisance in Sweden in recent years. Let θ be the probability that the morning train at 8.15 am from Nyköping arrives more than an hour late to Stockholm Central.

(a) *Assume a Beta(α, β) prior for θ. Let one of your friends play the role of an expert. Elicit you friend's values of α and β by letting him/her state his/her prior mean and standard deviation of θ.*
Let m and s denote your friend's mean and standard deviation. The prior hyperparameters can then be solved from the system of equations:

$$m = \frac{\alpha}{\alpha + \beta}$$

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

(b) *During the last 100 days, the train has never been more than an hour late. Update the prior from 4a) to a posterior distribution for θ based on these data.*
This gives the posterior $Beta(\alpha, \beta + 100)$.

(c) *When the train is more than an hour late, the railway company (SJ) needs to refund the ticket cost to every passenger in the train. The total cost for being late on a given day is approximately 296,000 Krona. SJ can avoid delays by servicing the train more frequently. The monthly service cost is 108,000 Krona, and would guarantee that the morning train would never be more than an hour late in that month. According to a Bayesian analysis, should SJ pay the extra service cost?*
This is decision problem with two actions: a_1 : Pay the service fee, a_2 : Do not pay the service fee. We choose the action that minimizes expected loss. Assume that the loss equals the cost (SJ doesn't per se care about people waiting ...). Expected

loss for a_1 is then $EL_1 = 108000 = -108000$. The expected loss for action a_2 is

$$EL_2 = E_{y|x}[y \cdot 296000] = 296000 \cdot E(y|x),$$

where $y =$#train within the month with more than one hour delay, and x is the data from 4b. $y|\theta \sim Bin(30, \theta)$ if we assume that a month has 30 days. The posterior for θ is $Beta(\alpha, \beta + 100)$, so

$$E(y|x) = E_{\theta|x}E(y|x, \theta) = E_{\theta|x}(30 \cdot \theta) = 30\frac{\alpha}{\alpha + \beta + 100}.$$

So, choose a_1 if

$$30\frac{\alpha}{\alpha + \beta + 100} \cdot 296000 > 108000.$$

For example, if $\alpha = \beta = 1$ we get $EL_2 = 87058$ krona, so we choose to not pay for the service.

(5) Let $y_1, ..., y_n$ be a random sample from the $N(\theta, \sigma^2)$ model, where both θ and σ^2 are unknown. Let $n = 25$, $\bar{y} = 33.75$ and $s^2 = 5^2$, where s is the sample standard deviation.

(a) *Compute the posterior distribution of θ based on some conjugate prior.*
$\theta|\sigma^2 \sim N(\mu_0, \sigma^2/\kappa_0)$ and $\sigma^2 \sim$Inv-$\chi^2(\nu_0, \sigma_0^2)$ is the natural conjugate prior. The marginal posterior of θ is a $t_{\nu_0}(\mu|\mu_n, \sigma_n^2/\kappa_n)$.

(b) *Investigate the sensitivity of the posterior to variations in the prior.*
There are many ways one can do this, preferably with plots. One good example is to plot e.g. μ_n as a function of κ_0 for different values of μ_0. Or the posterior variance, σ_n^2/κ_n, as function of κ_0 for given values of σ_0 and ν_0. We can also do 3D graphics, plotting μ_n as function of μ_0 and κ_0. You can also plot the whole posterior density for different values on the prior hyperparameters, e.g. κ_0.

(c) *Suppose that the data has two outlying (extreme) observations. Discuss how you would deal with this?*
Outliers can really affect inferences, especially in Gaussian/normal models. The usual way to deal with this is to relax the Gaussian assumption and to use a more heavy-tailed model, such as the student-t or a mixture distribution with Gaussian components.

- Mattias