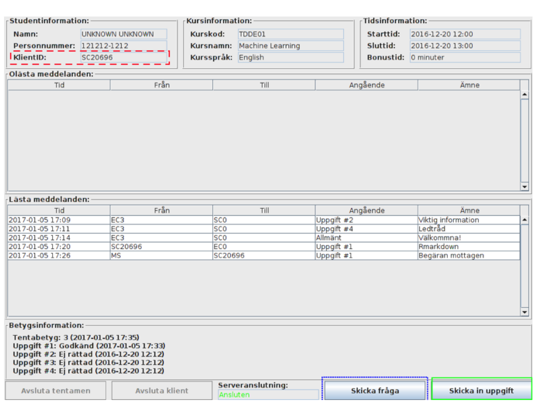
LINKÖPING UNIVERSITY
Dept. of Computer and Information Science
Division of Statistics and Machine Learning
Mattias Villani

2017-08-16
Bayesian Learning, 6 hp
732A91/TDDE07

Computer Exam - Bayesian Learning (732A91/TDDE07), 6 hp

Time:	2-6 PM
Allowable material:	- The allowed material in the folder given_files in the exam system. - Calculator with erased memory.
Teacher:	Mattias Villani. Phone: 070 − 0895205 and through the Communication client.
Exam scores:	Maximum number of credits on the exam: 40. Maximum number of credits on each exam question: 10.
Grades (732A91):	A: 36 points B: 32 points C: 24 points D: 20 points E: 16 points F: <16 points
Grades (TDDE07):	5: 34 points 4: 26 points 3: 18 points U: <18 points

INSTRUCTIONS:
When asked to give a solution on **Paper**, give that answer on physical papers supplied with the exam.
Each submitted sheet of paper should be marked with your *Client ID* from the *Communication Client*.
The client ID is the code in the `red` dashed rectangle in figure below.
All other answers should be submitted in a single PDF file using the *Communication Client.*
Include important code needed to grade the exam (inline or at the end of the PDF).
Submission starts by clicking the button in the `green` solid rectangle in figure below.
The submitted PDF file should be named *BayesExam.pdf*
Questions can be asked through the Communication client (`blue` dotted rectangle in figure below).
Full score requires clear and well motivated answers.



1. Bayesian inference for Cauchy data

The Cauchy distribution has density

$$p(y) = \frac{1}{\pi\gamma}\left(\frac{1}{1+\left(\frac{y-\theta}{\gamma}\right)^2}\right) \quad -\infty < y < \infty,$$

where $-\infty < \theta < \infty$ is the location parameter and $\gamma > 0$ is the scale parameter. The file `ExamData.R` contains code for the Cauchy density.

(a) *Credits: 3p.* Assume for now that we know that $\gamma = 1$. Plot the posterior distribution of θ based on the sample in the supplied data file `CauchyData.RData`. For simplicity, let $\theta \sim N(0, 10^2)$.
Solution: See `Exam732A91_170816_Sol.R`.

(b) *Credits: 5p.* Now assume that also γ is unknown and that $\gamma \sim \text{lognormal}(0, 1)$ a priori independently from θ (the lognormal density is given the file `ExamData.R`). Use numerical optimization to obtain a normal approximation of the *joint* posterior distribution of θ and γ. You don't need to plot the distribution, just provide its mean and covariance matrix. [Hints: use the argument `lower` in `optim`, and `method=c("L-BFGS-B")`].
Solution: See `Exam732A91_170816_Sol.R`.

(c) *Credits: 2p.* Use the normal approximation in 1(b) to obtain the marginal posterior for the 99% percentile of the Cauchy distribution $\theta + \gamma \cdot \tan(\pi(0.99 - 0.5))$. [Hint: `rmvnorm` in the `mvtnorm` package].
Solution: See `Exam732A91_170816_Sol.R`.

2. Regression

The Boston housing data contains characteristics of 506 houses in the Boston suburbs and their selling price. The dataset is loaded by the `ExamData.R` file. The original data is in `Boston` and `?Boston` will present the help file with information on all variables. We are here interested in modelling the response variable `medv` (median value of the house in 1000$) as a function of all the other variables in the dataset. The `ExamData.R` also prepares the data so that the vector **y** contains the response variable and the matrix **X** contains the covariates (with the first column being ones to model the intercept term). The vector `covNames` contains the names of all the covariates. Use the conjugate prior

$$\begin{aligned}\beta|\sigma^2 &\sim N(0, 10^2\sigma^2 I)\\ \sigma^2 &\sim Inv-\chi^2(1, 6^2).\end{aligned}$$

(a) *Credits: 4p.* Use the function `BayesLinReg` supplied in `ExamData.R` to simulate 5000 draws from the posterior distribution of all regression coefficients and the error variance. Summarize the posterior by the point estimate under the quadratic loss function, and by 95% equal-tail credible intervals. Interpret the credible interval for the regression coefficient on the number of rooms (`rm`).

Solution: See `Exam732A91_170816_Sol.R`.

(b) *Credits: 4p.* The owners of house no. 381 is considering selling their house. They bought the house for \$10400 (`medv`=10.4). The real estate agent says that because the crime rate has gone down dramatically in the area (`crim` has decreased from 88.9762 to 10), the house is expected to sell for around \$20000 now, and there is even a good chance of getting as much as \$30000. Do a Bayesian analysis (using simulation methods) to determine how reasonable the claims of the agent are.
Solution: See `Exam732A91_170816_Sol.R`.

(c) *Credits: 2p.* The linear Gaussian regression model analyzed by `BayesLinReg.R` makes a number of assumptions, and also assumes that we know the correct covariates to use. Discuss on **Paper** how a Bayesian can proceed if has been established that one or several of these assumptions are not fulfilled in the data.

Solution: The model assumes (see Slide 5 on Lecture 5): i) linear relationship, ii) same variance for all observations, iii) independent errors, iv) normality (Gaussianity) of the errors, v) the covariates are non-random (known). If one or several of the assumptions are wrong, we can generalize the model. I have mentioned most of these extensions, in varying degree of detail, and you don't need to list all of them here to get full score. The important thing is the understanding that we need a model general model and that this will affect how easy it is to analyze the data (we often need MCMC). We can generalize i) by using polynomial or *splines*, ii) by modeling the *error variance as a function of covariates*, e.g. $\sigma^2 = \exp(\mathbf{x}_i^T \gamma)$, iii) by having a *time series model for the errors* (AR(1)-process for example), iv) by using *mixture of normals for the errors* instead of a single normal and v) by modeling the covariates by a probability model. Many of these model extensions will require Gibbs sampling or Metropolis-Hastings for the simulation since it will not be possible to find conjugate priors and do direct simulation like in `BayesLinReg`. The fact that we don't know which covariates to use can be handled by *Bayesian variable selection*, or clever shrinkage priors.

3. Exponential data

Let $X_1, ..., X_n|\theta \overset{iid}{\sim} \text{Expon}(\theta)$ be exponentially distributed data. This problem should only be solved on **Paper**.

(a) *Credits: 4p.* Show that the Gamma distribution is the conjugate prior for independent exponential data.
Solution: Let $\theta \sim \text{Gamma}(\alpha, \beta)$ with density

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\theta\beta}.$$

By Bayes' theorem and independence

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

which is proportional to the $\text{Gamma}(\alpha + n, \beta + n\bar{x})$ density. Since the posterior belongs to the same distributional family (Gamma) as the prior, the prior is indeed conjugate to the exponential data model (likelihood).

(b) *Credits: 4p.* Derive the predictive distribution of a new observation x_{n+1} from the same model as in 3(a).
Solution: The predictive distribution of a new observation x_{n+1} is

$$\begin{aligned}
p(x_{n+1}|x_1, ..., x_n) &= \int p(x_{n+1}|\theta)p(\theta|x_1, ..., x_n)d\theta \text{ [given } \theta, x_{n+1} \text{ is indep of } x_1, ..., x_n] \\
&= \int \theta e^{-x_{n+1}\theta} \frac{(\beta + n\bar{x})^{\alpha+n}}{\Gamma(\alpha + n)} \theta^{\alpha+n-1} e^{-\theta(\beta+n\bar{x})} d\theta \\
&= \frac{(\beta + n\bar{x})^{\alpha+n}}{\Gamma(\alpha + n)} \int \theta^{\alpha+n+1-1} e^{-\theta(x_{n+1}+\beta+n\bar{x})} d\theta \\
&= \frac{(\beta + n\bar{x})^{\alpha+n}}{\Gamma(\alpha + n)} \frac{\Gamma(\alpha + n + 1)}{(x_{n+1} + \beta + n\bar{x})^{\alpha+n+1}} \\
&= (\alpha + n) (\beta + n\bar{x})^{\alpha+n} \frac{1}{(x_{n+1} + \beta + n\bar{x})^{\alpha+n+1}}.
\end{aligned}$$

This distribution is sometimes called the Gamma-Gamma distribution. It is enough for full score here to arrive at a distribution that is proportional to

$$\frac{1}{(x_{n+1} + \beta + n\bar{x})^{\alpha+n+1}}$$

(c) *Credits: 2p.* Suppose that you may have doubts on whether the exponential distribution really is appropriate for your data. Propose two alternative models that may be good candidates here, and discuss how a Bayesian can handle a situation where three candidate models are plausible, but one does not know which model is the best. This is a discussion question to be answered on **Paper**. Be brief and concise in your answer.
Solution: i) Bayesian model comparison (posterior distribution or marginal likelihoods over the three models). ii) Model averaging if one is interested in prediction. iii) Posterior predictive checks can be used to rule out models that can't have generated the data. For full points two out of three of these should be mentioned.

4. Prediction and decision

A firm produces a product. Let X_t denote the quantity demanded of the product in quarter t, which is assumed to follow a Poisson distribution: $X_t|\theta \overset{iid}{\sim} \text{Poisson}(\theta)$. The demand in the four quarters in the previous year was: $x_1 = 220, x_2 = 323, x_3 = 174, x_4 = 229$.

(a) *Credits: 2p.* Simulate 1000 draws from the posterior distribution of θ using a conjugate prior for θ with mean 250 and a standard deviation of 50.
Solution: We know from the lecture slides that the conjugate prior is $\theta \sim \text{Gamma}(\alpha, \beta)$ giving the posterior $\theta|x_1, ..., x_4 \sim \text{Gamma}(\alpha + \sum_{i=1}^{4} x_i, \beta + n)$, and the data is: $n = 4$ and $\sum_{i=1}^{4} x_i = 946$. We need to determine α and β in the prior. From the properties of the Gamma distribution

$$E(\theta) = \frac{\alpha}{\beta} = 250$$
$$Var(\theta) = \frac{\alpha}{\beta^2} = 50^2.$$

Now, solving this system of equation gives $\alpha = 25$ and $\beta = 0.1$. The posterior is therefore $\theta|x_1, ..., x_4 \sim \text{Gamma}(971, 4.1)$. See also `Exam732A91_170816_Sol.R`.

(b) *Credits: 3p.* Simulate 1000 draws from the predictive distribution of next quarter's demand, X_5, and plot the draws as a histogram. What is $\Pr(X_5 \leq 200|x_1, ..., x_4)$?
Solution: The predictive distribution is

$$p(X_5|x_1, ..., x_4) = \int p(X_5|\theta)p(\theta|x_1, ..., x_4)d\theta$$

We can therefore simulate draws from the predictive distribution of X_5 by repeatedly simulating i) from the posterior $\theta|x_1, ..., x_4 \sim \text{Gamma}(\alpha + \sum_{i=1}^{4} x_i, \beta + n)$ followed by ii) simulation from the model $X_5|\theta \sim \text{Poisson}(\theta)$. See also `Exam732A91_170816_Sol.R`.

(c) *Credits: 5p.* The firm needs to decide how much of the product to keep in stock for next quarters sale. Its utility function is of the form

$$u(a, X_5) = \begin{cases} p \cdot X_5 - (a - X_5) & \text{if } X_5 \leq a \\ p \cdot a - 0.05\,(X_5 - a)^2 & \text{if } X_5 > a \end{cases}$$

where $p = 10$ is the sale price for the product and a (positive integer) is the stock held for next quarter. This utility function is given in the file `ExamData.R` (note that `X5` can be a vector of values, but a needs to a scalar in the code). Use simulation to find the optimal a from a Bayesian point of view. Explain (argue) why the optimal value is larger than the expected value of X_5. [Hint: use a grid of a values around the expected value for X_5.]
Solution: See `Exam732A91_170816_Sol.R` and Figure 1 which plots the posterior expected utility as a function of a and marks out the maximum.

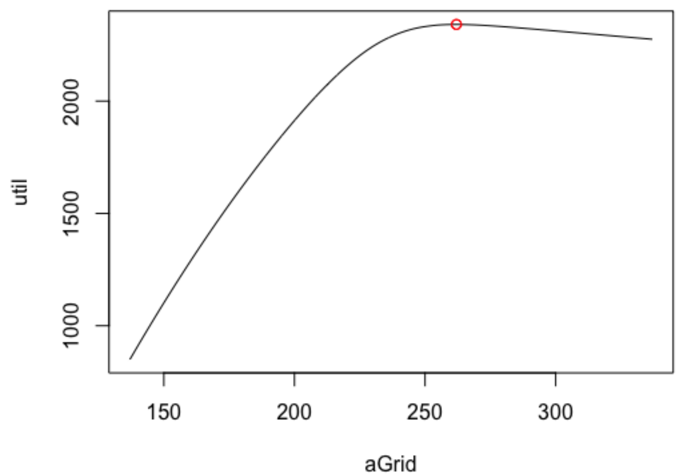


Figure 1: Posterior expected utility as a function of stock size, a.

Good luck!

Mattias