UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK4021/9021 — Applied Bayesian statistics

Day of examination: November 30 - 2022

Examination hours: 15.00 – 19.00.

This problem set consists of 6 pages.

Appendices: None

Permitted aids: Approved calculator

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Notation/formulas:

- $I(A)$ is the indicator function, equal to 1 if the event A is true.
- The Gamma density function is given by

$$\text{Gamma}(y;a,b) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}, \quad y > 0$$

  with $a, b > 0$ and with expectation a/b and variance a/b^2.
- The Poisson distribution is given by

$$\text{Poisson}(y;\lambda) = \frac{\lambda^y}{y!} e^{-\lambda}, \quad y = 0, 1, 2, ...$$

  with $\lambda > 0$ and expectation and variance equal to λ.
- The Negative binomial distribution is given by

$$\text{Neg-Bin}(y;\alpha,\beta) = \binom{y+\alpha-1}{\alpha-1} \left(\frac{\beta}{\beta+1}\right)^{\alpha} \left(\frac{1}{\beta+1}\right)^{y}, \quad y = 0, 1, 2, ...$$

  with $\alpha, \beta > 0$ and expectation equal to $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}(\beta + 1)$.
- Bold $\boldsymbol{b}$ is used for a vector.

(Continued on page 2.)

Problem 1

Consider a classification setting where $C \in \{1, 2, 3\}$ and

$$Y|C = c \sim \text{Gamma}(y; a_c, b_c).$$

(a) Assume for all c:

- $\Pr(C = c) = 1/3$;
- $a_c = a$.

Derive $\Pr(C = c|y)$ for all c. Here y is a single observation.

(b) Assume $a = 3, b_1 = 1, b_2 = 0.5$ and $b_3 = 0.25$. Assume further a loss function $L(c, \hat{c}) = I(\hat{c} \neq c)$. Specify for which values of y you obtain $\hat{c} = 1, 2$ or 3.

(c) Consider now the alternative loss function

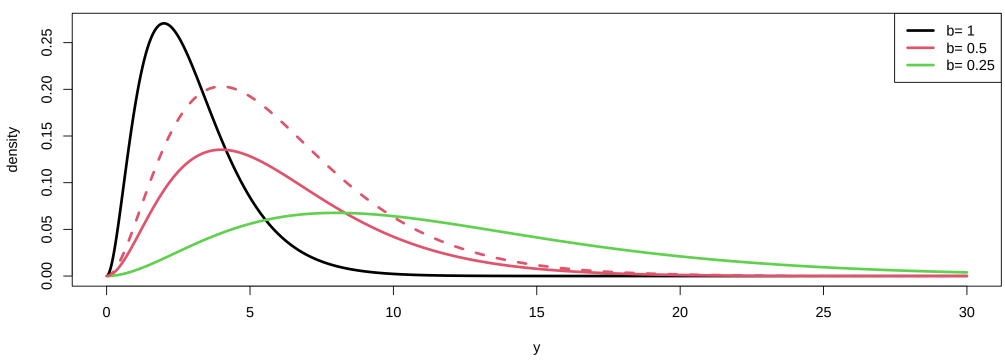
$$L(c, \hat{c}) = \begin{cases} I(\hat{c} \neq 1) & \text{if } c = 1; \\ 1.5 \cdot I(\hat{c} \neq 2) & \text{if } c = 2; \\ I(\hat{c} \neq 3) & \text{if } c = 3. \end{cases}$$

Show that the classification rule in this case becomes

$$\hat{c} = \begin{cases} 1 & \text{if } \Pr(C = 1|y) > \max\{\frac{3}{2}\Pr(C = 2|y), \Pr(C = 3|y)\}; \\ 2 & \text{if } \frac{3}{2}\Pr(C = 2|y) > \max\{\Pr(C = 1|y), \Pr(C = 3|y)\}; \\ 3 & \text{if } \Pr(C = 3|y) > \max\{\frac{3}{2}\Pr(C = 2|y), \Pr(C = 1|y)\}. \end{cases}$$

(d) The plot below shows the three Gamma densities. In addition, the dashed line shows the density for $C = 2$ multiplied by 1.5. Discuss how this plot is related to the decision rules obtained in (b) and in (c).

Given the form of the new loss function, do you find the changes in decision boundaries reasonable?



(Continued on page 3.)

Problem 2

We will continue with the same classification setting as for Problem 1, that is where $C \in \{1,2,3\}$, $\Pr(C=c)=1/3$ and

$$Y|C=c \sim \text{Gamma}(y;a,b_c)$$

for $c=1,2,3$. In this problem (which can be solved independently from Problem 1) we will however see how we can perform inference on the now unkown parameters $\boldsymbol{\theta}=(a,b_1,b_2,b_3)$. Assume we have data $\{y_{c,i}, c=1,2,3, i=1,...,n_c\}$ where you know that $y_{c,i}$ is from class c. One can show that the likelihood function for $\boldsymbol{\theta}$ is, up to some proportionality constant, equal to

$$L(\boldsymbol{\theta}) = \frac{\prod_{c=1}^3 b_c^{an_c}}{\Gamma(a)^n}(\prod_{c=1}^3\prod_{i=1}^{n_c} y_{c,i})^{a-1}\prod_{c=1}^3 e^{-b_c\sum_{i=1}^{n_c} y_{c,i}}$$

where $n=\sum_{i=1}^3 n_c$. This you do not need to show.

(a) What are the sufficient statistics in this case? Why are sufficient statistics useful?

(b) Define a prior

$$p(\boldsymbol{\theta}) = p(a)\prod_{c=1}^3 \text{Gamma}(b_c;\alpha,\beta)$$

where we assume α,β and possible parameters in $p(a)$ are fixed.

Show that conditional on a and $\boldsymbol{y}$ (the whole set of observations), b_1,b_2,b_3 are independent.

Derive also these conditional distributions, that is $p(b_c|a,\boldsymbol{y})$.

(c) Show how you can compute the marginal posterior distribution $p(a|\boldsymbol{y})$.

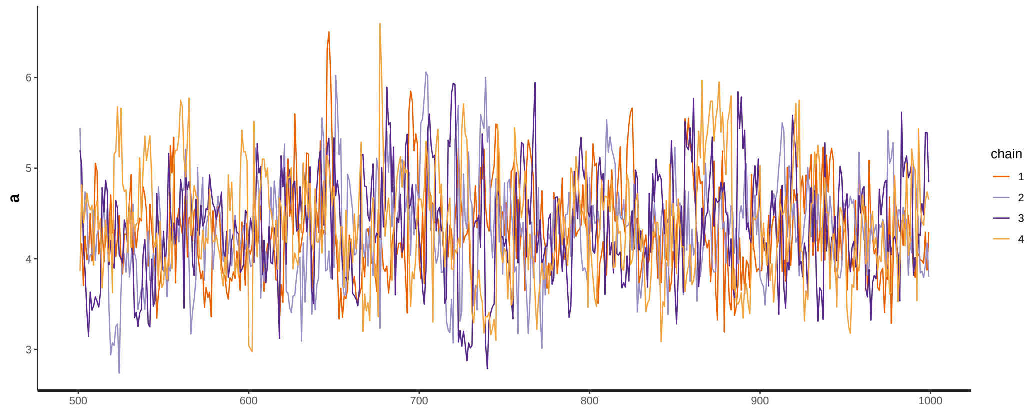
Use this to explain how you can simulate from the posterior distribution $p(a,b_1,b_2,b_3|\boldsymbol{y})$.

(d) The table below shows the output from a call to `stan` based on a set of 100 simulated observations with 1000 MCMC iterations of which 500 are used as burnin/warmup.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
a	4.34	0.03	0.57	3.31	3.94	4.31	4.69	5.54	462	1.00
lp__	−246.36	0.06	1.45	−250.06	−247.05	−246.06	−245.29	−244.50	606	1.01

Further, the plot below shows traceplots from 4 parallel chains. Both the table and the plot are for the parameter a.

(Continued on page 4.)



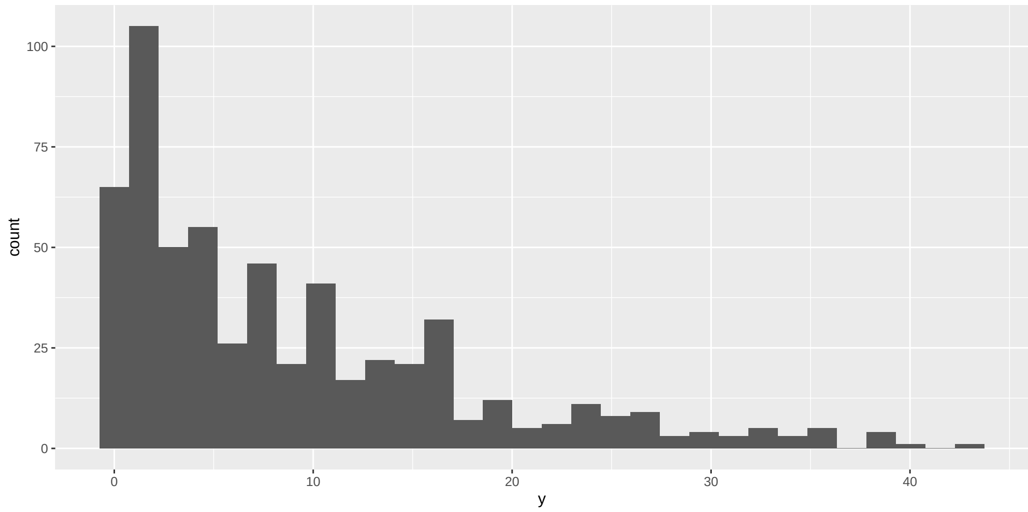
Explain what `n_eff` (the effective number of samples) and `Rhat` are in this table (you do not need explicit formulas but explain it in general terms). Also explain what we mean by "burnin".

Based on the table and the plot, do you feel confident on that there are enough iterations here?

(e) Assume now that you have some additional observations $y_i, i = 1, ..., n_2$ where the corresponding classes are unknown. Explain how you can use MCMC algorithms to include these observations in the inference procedure.

Problem 3

The histogram below shows the new admissions at hospital due to Covid in the period Feb 17 2020 - Sep 17 2021, a total of $n = 588$ observations.



Since these are count data, a natural model would be the Poisson model. However, the empirical mean of the data are 9.00 while the empirical variance is 79.38, indicating that there is some extra variability involved (overdispersion).

(Continued on page 5.)

An alternative model approach is to assume

$$Y_t|\lambda_t \overset{ind}{\sim} \text{Poisson}(\lambda_t);$$
$$\lambda_t \overset{ind}{\sim} p(\lambda|\boldsymbol{\theta}),$$

where $p(\lambda|\boldsymbol{\theta})$ is some distribution defined through some hyperparameters $\boldsymbol{\theta}$. Here, $\overset{ind}{\sim}$ means that the variables are (conditionally) independent. We will consider two possible models for λ_t:

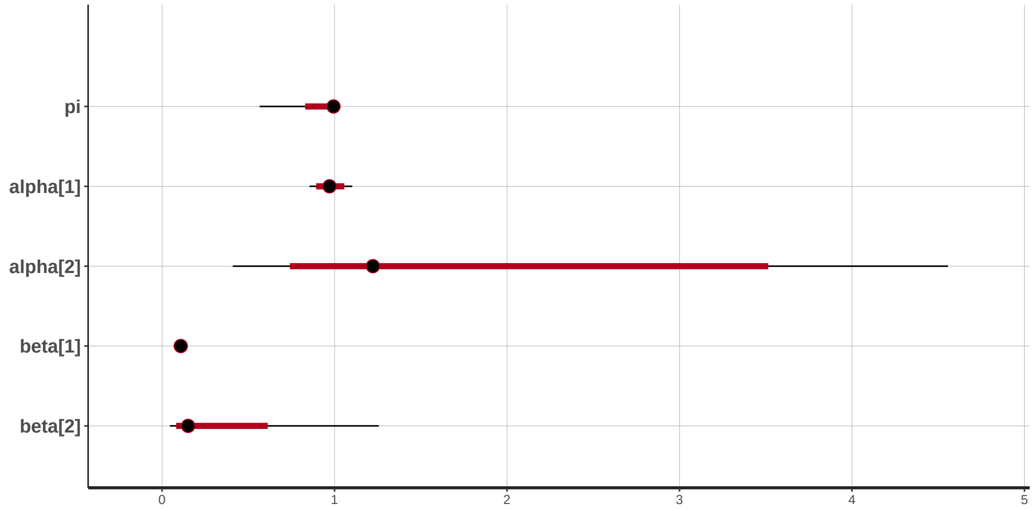
$$\lambda_t \sim \text{Gamma}(\alpha, \beta) \qquad \text{Model 1;}$$
$$\lambda_t \sim \pi \cdot \text{Gamma}(\alpha_1, \beta_1) + (1-\pi) \cdot \text{Gamma}(\alpha_2, \beta_2) \qquad \text{Model 2,}$$

where $\boldsymbol{\theta}_1 = (\alpha, \beta)$ and $\boldsymbol{\theta}_2 = (\pi, \alpha_1, \beta_1, \alpha_2, \beta_2)$ are hyperparameters for models 1 and 2, respectively. Model 2 is a special case of a *mixture distribution.*

(a) During the course we have shown that for model 1, the marginal distribution for Y_t (given $\boldsymbol{\theta}_1$) becomes Neg-Bin$(y; \alpha, \beta)$. This you do not need to show.

Based on this, derive the marginal distribution for Y_t under model 2 (given $\boldsymbol{\theta}_2$).

(b) Leave-one-out cross-validation (LOO-CV) gave values of -1619.8 and -1673.7 for models 1 and 2 respectively. The plot below shows estimates and 80% and 95% credibilitility intervals for the hyperparameters in Model 2. These results are based on uniform distributions for the α's, Gamma priors for the β's and a Beta$(3, 1)$ prior for π.



Discuss how LOO-CV can be computed. Based on the LOO values and the results displayed in the plot, which model would you prefer?

(c) In order to check the model(s), we will look at posterior predictive checking, using the test quantity

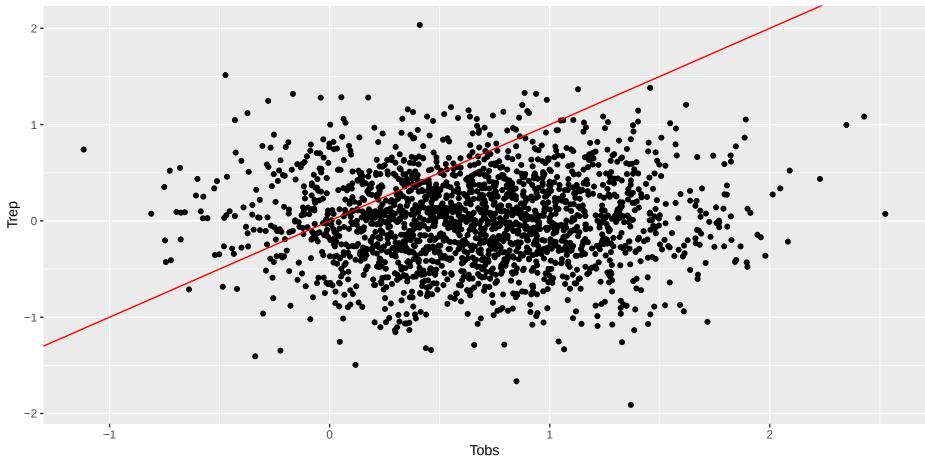
$$T(\boldsymbol{y}, \boldsymbol{\lambda}) = \frac{1}{n-1} \sum_{t=2}^{n} (y_t - \lambda_t)(y_{t-1} - \lambda_{t-1}).$$

(Continued on page 6.)

Here we assue that the observations are ordered according to dates.

The plot below shows plot of $T(\boldsymbol{y}, \boldsymbol{\lambda})$ against $T(\boldsymbol{y}^{rep}, \boldsymbol{\lambda})$. The red line is the line correponding to $y = x$. Further, the Bayesian p-value based on this test quantity is 0.159. Both the plot and the p value is based on the best model from (b).

Discuss these results, including some comments on why this specific test statistic is reasonable for the given problem.



THE END