STK 4020: Bayesian Statistics
Autumn 2010

Course Notes and Exercises
by Nils Lid Hjort

– This version: as of 15 October 2010 –

1. Prior to posterior updating with Poisson data

This exercise illustrates the basic prior to posterior updating mechanism for Poisson data.

(a) First make sure that you are reasonably acquainted with the Gamma distribution. We say that $Z \sim \mathrm{Gamma}(a, b)$ if its density is

$$g(z) = \frac{b^a}{\Gamma(a)} z^{a-1} \exp(-bz) \quad \text{on } (0, \infty).$$

Here a and b are positive parameters. Show that

$$\mathrm{E}\, Z = \frac{a}{b} \quad \text{and} \quad \mathrm{Var}\, Z = \frac{a}{b^2} = \frac{\mathrm{E} Z}{b}.$$

In particular, low and high values of b signify high and low variability, respectively.

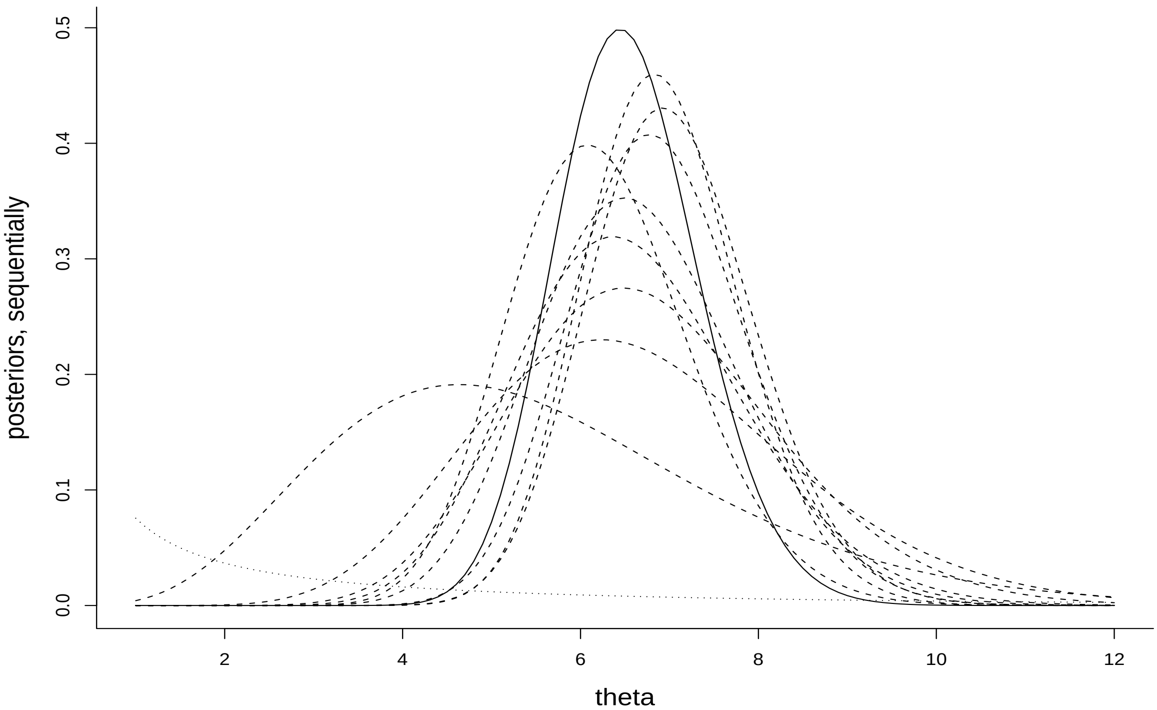

Figure 1: Eleven curves are displayed, corresponding to the $\mathrm{Gamma}(0.1, 0.1)$ *intial prior density for the Poisson parameter* θ *along with the ten updates following each of the observations 6, 8, 7, 6, 7, 4, 11, 8, 6, 3.*

(b) Now suppose $y \mid \theta$ is a Poisson with parameter θ, and that θ has the prior distribution $\mathrm{Gamma}(a, b)$. Show that $\theta \mid y \sim \mathrm{Gamma}(a + y, b + 1)$.

(c) Then suppose there are repeated Poisson observations $y_1, \ldots, y_n$, being i.i.d. $\sim \mathrm{Pois}(\theta)$ for given θ. Use the above result repeatedly, e.g. interpreting $p(\theta \mid y_1)$ as the new prior before observing y_2, etc., to show that

$$\theta \mid y_1, \ldots, y_n \sim \mathrm{Gamma}(a + y_1 + \cdots + y_n, b + n).$$

Also derive this result directly, i.e. without necessarily thinking about the data having emerged sequentially.

(d) Suppose the prior used is a rather flat $\mathrm{Gamma}(0.1, 0.1)$ and that the Poisson data are 6, 8, 7, 6, 7, 4, 11, 8, 6, 3. Reconstruct a version of Figure 1 in your computer, plotting the ten curves $p(\theta \mid \mathrm{data}_j)$, where data_j is $y_1, \ldots, y_j$, along with the prior density. Also compute the ten Bayes estimates $\widehat{\theta}_j = \mathrm{E}(\theta \mid \mathrm{data}_j)$ and the posterior standard deviations, for $j = 0, \ldots, 10$.

(e) The mathematics turned out to be rather uncomplicated in this situation, since the Gamma continuous density matches the Poisson discrete density so nicely. Suppose instead that the initial prior for θ is a uniform over $[0.5, 50]$. Try to compute posterior distributions, Bayes estimates and posterior standard deviations also in this case, and compare with you found above.

2. The Master Recipe for finding the Bayes solution

Consider a general framework with data y, in a suitable sample space $\mathcal{Y}$; having likelihood $p(y \mid \theta)$ for given parameter θ (stemming from an appropriate parametric model), with θ being inside a parameter space Ω; and with loss function $L(\theta, a)$ associate with decision or action a if the true parameter value is θ, with a belonging to a suitable action space $\mathcal{A}$. This could be the real line, if a parameter space is called for; or a two-valued set $\{\mathrm{reject}, \mathrm{accept}\}$ if a hypothesis test is being carried out; or the set of all intervals, if the statistician needs a confidence interval.

A statistical *decision function*, or procedure, is a function $\widehat{a}\colon \mathcal{Y} \to \mathcal{A}$, getting from data y the decision $\widehat{a}(y)$. Its *risk function* is the expected loss, as a function of the parameter:

$$R(\widehat{a}, \theta) = \mathrm{E}_\theta L(\theta, \widehat{a}) = \int L(\theta, \widehat{a}(y)) p(y \mid \theta) \,\mathrm{d}y.$$

(In particular, in this expectation operation the random element is y, having its $p(y \mid \theta)$ distribution for given parameter, and the integration range is that of the sample space $\mathcal{Y}$.)

So far the framework does not include Bayesian components per se, and is indeed a useful one for frequentist statistics, where risk functions for different decision functions (be they estimators, or tests, or confidence intervals, depending on the action space and the loss function) may be compared.

We are now adding one more component to the framework, however, which is that of a *prior distribution* $p(\theta)$ for the parameter. The overall risk, or *Bayes risk*, associated with a decision function $\widehat{a}$, is then the overall expected loss, i.e.

$$\mathrm{BR}(\widehat{a}, p) = \mathrm{E}\, R(\widehat{a}, \theta) = \int R(\widehat{a}, \theta) p(\theta)\, \mathrm{d}\theta.$$

(Here θ is the random quantity, having its prior distribution, making also the risk function $R(\widehat{a}, \theta)$ random.) The *minimum Bayes risk* is the smallest possible Bayes risk, i.e.

$$\mathrm{MBR}(p) = \min\{\mathrm{BR}(\widehat{a}, p) \colon \text{all decision functions } \widehat{a}\}.$$

Th *Bayes solution* for the problem is the strategy or decision function $\widehat{a}_B$ that succeeds in minimising the Bayes risk, with the given prior, i.e.

$$\mathrm{MBR}(p) = \mathrm{BR}(\widehat{a}_B, p).$$

The *Master Theorem* about Bayes procedures is that there is actually a recipe for finding the optimal Bayes solution $\widehat{a}_B(y)$, for the given data y (even without taking into account other values y' that could have been observed).

(a) Show that the *posterior density* of θ, i.e. the distribution of the parameter given the data, takes the form

$$p(\theta \,|\, y) = k(y)^{-1} p(\theta) p(y \,|\, \theta),$$

where $k(y)$ is the required integration constant $\int p(\theta) p(y \,|\, \theta)\, \mathrm{d}\theta$. This is the *Bayes theorem*.

(b) Show also that the *marginal distribution* of y becomes

$$p(y) = \int p(y \,|\, \theta) p(\theta)\, \mathrm{d}\theta.$$

(I follow the GCSR book's convention regarding using the 'p' multipurposedly.)

(c) Show that the overall risk may be expressed as

$$\begin{aligned}\mathrm{BR}(\widehat{a}, p) &= \mathrm{E}\, L(\theta, \widehat{a}(Y)) \\ &= \mathrm{E}\, \mathrm{E}\, \{L(\theta, \widehat{a}(Y)) \,|\, Y\} \\ &= \int \Big\{ \int L(\theta, \widehat{a}(y)) p(\theta \,|\, y)\, \mathrm{d}\theta \Big\}\, p(y)\, \mathrm{d}y.\end{aligned}$$

The inner integral, or 'inner expectation', is $\mathrm{E}\{L(\theta, \widehat{a}(y)) \,|\, y\}$, the expected loss given data.

(d) Show then that the optimal Bayes strategy, i.e. minimising the Bayes risk, is achieved by using

$$\widehat{a}_B(y) = \text{argmin} g = \text{the value } a_0 \text{ minimising the function } g,$$

where $g = g(a)$ is the expected posterior loss,

$$g(a) = \text{E}\{L(\theta, a) \,|\, y\}.$$

The g function is evaluated and mininised over all a, for the given data y. This is the Bayes recipe. – For examples and illustrations, with different loss functions, see the Nils 2008 Exercises.

3. Minimax estimators

For a decision function $\widehat{a}$, bringing data y into a decision $\widehat{a}(y)$, its max-risk is

$$R_{\max}(\widehat{a}) = \max_{\theta} R(\widehat{a}, \theta).$$

We say that a procedure a^* is *minimax* if it minimises the max-risk, i.e.

$$R_{\max}(a^*) \leq R_{\max}(\widehat{a}) \quad \text{for all competitors } \widehat{a}.$$

Here I give recipes (that often but not always work) for finding minimax strategies.

(a) For any prior p and strategy $\widehat{a}$, show that

$$\text{MBR}(p) \leq R_{\max}(\widehat{a}).$$

(b) Assume a^* is such that there is actually equality in (a), for a suitable prior p. Show that a^* is then minimax.

(c) Assume more generally that a^* is such that $\text{MBR}(p_m) \to R_{\max}(a^*)$, for a suitable sequence of priors p_m. Show that a^* is indeed minimax.

We note that minimax strategies often but not always have constant risk functions, and that they need not be unique – different minimax strategies for the same problem need to have identical max-risks, but the risk functions themselves need not be identical.

4. Minimax estimation of a normal mean [cf. Nils 2008 #3, 6, 9]

A prototype normal mean model is the simple one with a single observation $y \sim \text{N}(\theta, 1)$. We let the loss function be squared error, $L(\theta, a) = (a - \theta)^2$.

(a) Show that the maximum likelihood (ML) solution is simply $\theta^* = y$. Show that its risk function is $R(\theta^*, \theta) = 1$, i.e. constant.

(b) Let θ have the prior $\text{N}(0, \tau^2)$. Show that (θ, y) is binormal, and that $\theta \,|\, y \sim \text{N}(\rho y, \rho)$, with $\rho = \tau^2/(\tau^2 + 1)$. In particular, $\widehat{\theta}_B(y) = \rho y$ is the Bayes estimator.

(c) Find the risk function for the Bayes estimator, and identify where it is smaller than that of the ML solution, and where it is larger. Comment on the situation where τ is small (and hence ρ), as well as on the case of τ being big (and hence ρ close to 1).

(d) Show that $\mathrm{MBR}(\mathrm{N}(0,\tau^2)) = \rho = \tau^2/(\tau^2+1)$. Use the technique surveyed above to show that y is indeed minimax.

(e) This final point is to exhibit a technique for demonstrating, in this particular situation, that y is not only minimax, but the only minimax solution – this was given as Exercise #9(e) in the Nils 2008 collection, but without any hints. Assume that there is a competitor $\widehat\theta$ that is different from y and also a minimax estimator. Then, since risk functions are continuous (show this), there must be a positive ε and a non-empty interval $[c,d]$ with

$$R(\widehat\theta,\theta) \le \begin{cases} 1-\varepsilon & \text{on } [c,d], \\ 1 & \text{everywhere.} \end{cases}$$

Deduce from this that

$$\mathrm{MBR}(\mathrm{N}(0,p_\tau)) \le \mathrm{BR}(\widehat\theta,p_\tau) \le \int_{[c,d]} (1-\varepsilon)p_\tau(\theta)\,\mathrm{d}\theta + \int_{\text{elsewhere}} 1\cdot p_\tau(\theta)\,\mathrm{d}\theta,$$

writing p_τ for the $\mathrm{N}(0,\tau^2)$ prior. This leads to

$$\varepsilon(2\pi)^{-1/2}\frac{1}{\tau}\int_{[c,d]} \exp(-\tfrac12\theta^2/\tau^2)\,\mathrm{d}\theta \le 1-\mathrm{MBR}(p_\tau) = \frac{1}{\tau^2+1}.$$

Show that this leads to a contradiction: hence y is the single minimax estimator in this problem.

(f) Generalise the above to the situation with $y_1,\ldots,y_n \sim \mathrm{N}(\theta,\sigma^2)$.

5. Minimax estimation of a Poisson mean [cf. Nils 2008 #12]

Let $y\,|\,\theta$ be a Poisson with mean parameter θ, which is is to be estimated with weighted squared error loss $L(\theta,t)=(t-\theta)^2/\theta$. This case was treated in Nils 2008 #12, but here I add more, to take care of the more difficult admissibility point #12(g), where the task is to show that y is the only minimax estimator.

(a) Show that the maximum likelihood (ML) estimator is y itself, and that its risk function is the constant 1.

(b) Consider the prior distribution $\mathrm{Gamma}(a,b)$ for θ. Show that $\mathrm{E}\,\theta = a/b$ and that $\mathrm{E}\,\theta^{-1} = b/(a-1)$ if $a>1$, and infinite if $a\le 1$.

(c) Show that $\theta\,|\,y$ is a $\mathrm{Gamma}(a+y,b+1)$, from which follows

$$\mathrm{E}(\theta\,|\,y) = \frac{a+y}{b+1} \quad\text{and}\quad \mathrm{E}(\theta^{-1}\,|\,y) = \frac{b+1}{a-1+y}.$$

The latter formula holds if $a-1+y>0$, which means for all y if $a\ge 1$, but care is needed if $a<1$ and $y=0$. Show that the Bayes solution is

$$\widehat\theta = \frac{a-1+y}{b+1} \quad \text{for all } y\ge 0,$$

provided $a \geq 1$, but that we need the more careful formula

$$\widehat{\theta} = \begin{cases} (a-1+y)/(b+1) & \text{if } y \geq 1, \\ 0 & \text{if } y = 0, \end{cases}$$

in the case of $a < 1$.

(d) Taking care of the simplest case $a > 1$ first, show that

$$\mathrm{MBR}(p_{a,b}) = \frac{1}{b+1},$$

writing $p_{a,b}$ for the Gamma prior (a, b). This is enough to demonstrate that y is indeed minimax, cf. the Nils 2008 #12 Exercise.

(e) Attempt to show that y is the only minimax estimator via the technique of the previous exercise, starting with a competitor $\widetilde{\theta}$ with risk function always bounded by 1 and bounded by say $1 - \varepsilon$ on some non-empty parameter interval $[c, d]$. Show that this leads to

$$\varepsilon \int_{[c,d]} p_{a,b}(\theta)\,\mathrm{d}\theta \leq 1 - \mathrm{MBR}(p_{[a,b]}).$$

For the easier case of $a > 1$, this gives a simple right hand side, but, perhaps irritatingly, not a contradiction – one does not yet know, despite certain valid and bold mathematical efforts, whether y is the unique minimax method or not.

(f) Since the previous attempt ended with 'epic fail', we need to try out the more difficult case $a < 1$ too. Show that

$$\mathrm{E}\{L(\theta, \widehat{\theta}) \,|\, y\} = \begin{cases} 1/(b+1) & \text{if } y \geq 1, \\ a/(b+1) & \text{if } y = 0. \end{cases}$$

Deduce from this a minimum Bayes risk formula also for the case of $a < 1$:

$$\mathrm{MBR}(p_{a,b}) = \frac{1}{b+1}\Big\{1 - \Big(\frac{b}{b+1}\Big)^a\Big\} + \frac{a}{b+1}\Big(\frac{b}{b+1}\Big)^a.$$

(g) Find a sufficiently clever sequence of Gamma priors (a_m, b_m), with $a_m \to 1$ from the left and $b_m \to 0$ from the right, that succeeds in squeezing a contradiction out of equality in point(e). Conclude that y is not only minimax, but the only minimax strategy.

(h) Generalise these results to the situation where $y_1, \ldots, y_n$ are independent and Poisson with rates $c_1\theta, \ldots, c_n\theta$, and known multipliers $c_1, \ldots, c_n$. Indentify a minimax solution and show that it is the only one on board.