UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4020/9020 – Bayesian Statistics**
	Part I of two parts: The project
WITH:	**Nils Lid Hjort**
TIME FOR EXAM:	**3.xii.–14.xii.2012**

This is the exam project set for STK 4020/9020, autumn semester 2012. It is made available on the course website as of *Monday 3 December 12:00*, and candidates must submit their written reports by *Friday 14 December 11:59* (or earlier), to the reception office at the Department of Mathematics, in duplicate. The supplementary oral examinations take place *Monday December 17* (practical details concerning this are provided elsewhere). Reports may be written in nynorsk, bokmål, riksmål, English or Latin, and should preferably be text-processed (TeX, LaTeX, Word), but may also be hand-processed. Give your name on the first page. Write concisely (in der Beschränkung zeigt sich erst der Meister; brevity is the soul of wit; краткость – сестра таланта). Relevant figures need to be included in the report. Copies of machine programmes used (in `R`, or `matlab`, or similar) are also to be included, perhaps as an Appendix to the report. Candidates are required to work on their own (i.e. without cooperation with any others), but are graciously allowed not to despair if they do not manage to answer all questions well.

Importantly, each student needs to submit *two special extra pages* with her or his report. *The first* (page A) is the 'erklæring' (self-declaration form), properly signed, and with the appropriate course form STK 4020 (master level) or STK 9020 (PhD level) clearly marked; it is available at the webpage as 'Exam Project, page A, declaration form'. *The second* (page B) is the student's one-page summary of the exam project report, which should also contain a brief self-assessment of its quality.

This exam set contains four exercises and comprises eight pages (including a one-page Appendix). Note that *the STK 9020 students need to answer also Exercise 4*, whereas the STK 4020 students can confine their attention to Exercises 1–3.

Exercise 1

ARE BAD-TEMPERED MEN BETTER at finding good-tempered women than the good-tempered men are? Or, to rephrase such a delicate and intricate question, do good-tempered women in their good-temperedness have a certain tendency to penetrate the shields of even bad-tempered men? Sir Francis Galton did not merely invent fingerprinting and correlation and regression and the two-dimensional normal distribution while working on anthropology and genetics and meteorology or exploring the tropics, but had a formidable appetite for even arcane psychometrics and for actually attempting to answer half-imprecise but good questions like the above in meaningful ways – by going out in the world to observe, to note, to think, to analyse (just as his perhaps even more famous cousin did).

On an inspired day in 1887 he therefore sat down and examined interview results pertaining to 111 married couples (see the Appendix), and classified the wives and husbands into 'bad-tempered' and 'good-tempered', reaching the following table:

		wife : good-tempered	wife : bad-tempered
husband :	good-tempered	24	27
	bad-tempered	34	26

He did not merely compute the proportions of relevance to the question raised above, but speculated about methods for answering whether the observed deviations from 'there is no difference' were statistically significant (some ten years before such concepts slowly began to take precise form in the statistics community). – Below it is your task to use the above simple dataset to help illustrate certain Bayesian techniques which might have interested Galton.

(a) Let us write

$$\begin{pmatrix} N_{0,0} & N_{0,1} \\ N_{1,0} & N_{1,1} \end{pmatrix} = \begin{pmatrix} 24 & 27 \\ 34 & 26 \end{pmatrix}$$

for the counts $N_{i,j} = \#\{X = i, Y = j\}$ for $i, j = 0, 1$, with X the good- (0) or bad-tempered (1) category for the husband and Y similarly the category for the wife. We shall take take the observed counts to be a random sample from the multinomial model with parameters $(n, p_{0,0}, p_{0,1}, p_{1,0}, p_{1,1})$, with $p_{i,j}$ interpreted as $\Pr\{X = i, Y = j\}$ for a randomly selected married couple (X, Y). Briefly discuss the validity of this assumption. Also give clear interpretations to the quantities

$$\alpha_i = p_{i,0} + p_{i,1} \quad \text{for } i = 0, 1,$$
$$\beta_j = p_{0,j} + p_{1,j} \quad \text{for } j = 0, 1.$$

(b) Your Bayesian duty is now to come up with a prior for $(p_{0,0}, p_{0,1}, p_{1,0}, p_{1,1})$ which matches *your* prior beliefs concerning the world of married couples. For simplicity you are asked to choose your prior from the class of Dirichlet distributions, say $\mathrm{Dir}(a_{0,0}, a_{0,1}, a_{1,0}, a_{1,1})$ with density

$$\frac{\Gamma(k)}{\Gamma(a_{0,0})\Gamma(a_{0,1})\Gamma(a_{1,0})\Gamma(a_{1,1})} p_{0,0}^{a_{0,0}-1} p_{0,1}^{a_{0,1}-1} p_{1,0}^{a_{1,0}-1} (1 - p_{0,0} - p_{0,1} - p_{1,0})^{a_{1,1}-1}$$

over the simplex where the $p_{i,j}$ are positive with $p_{0,0} + p_{0,1} + p_{1,0} < 1$; also, $k = a_{0,0} + a_{0,1} + a_{1,0} + a_{1,1}$. Discuss, but briefly, how you arrived at your prior.

(c) We shall take an interest in the four parameters

$$\phi = \sum_{i,j} (p_{i,j} - \alpha_i \beta_j)^2 / p_{i,j}, \quad \kappa = \max_{i,j} |p_{i,j} - \alpha_i \beta_j|,$$
$$\gamma = p_{0,0} + p_{1,1}, \quad \delta = p_{0,1} p_{1,0} / (p_{0,0} p_{1,1}).$$

Explain how these parameters may be interpreted in the present context. For your chosen prior, use simulation to display the 0.05, 0.50, 0.95 quantiles of these parameters.

(d) Using Galton's data and your prior, derive the posterior distribution of the $(p_{0,0}, p_{0,1}, p_{1,0}, p_{1,1})$. Again via simulation, display the 0.05, 0.50, 0.95 quantiles for the posterior distribution of the four parameters $\phi, \kappa, \gamma, \delta$. Sum up your findings.

(e) With the multinomial model used here, derive the so-called Jeffreys prior. Please redo part of or all of the analysis using this prior to complement your own. Briefly discuss whether there are any noticeable discrepancies between the Jeffreys based analysis and that based on your own prior.

Exercise 2

A THING OF BEAUTY IS A JOY FOR EVER, but even a good car loses market and monetary value as time marches on. The data below are a partial reconstruction and transformation from my side of a bigger data set where I focus attention on how much a car is being sold for, as a function of its age x (measured in years) since the initial sale as brand new, in terms of this initial sale value, in percent y. Thus car no. 2 of these thirty cars was sold after 1.90 years at a price 62.1% of its original price, etc.

x	y	x	y	x	y
1.56	47.6	4.54	29.3	6.03	19.8
1.90	62.1	4.65	22.4	6.15	14.4
1.92	47.3	4.75	23.0	6.23	15.7
1.99	53.7	5.05	26.0	6.43	15.9
2.02	40.7	5.31	18.7	6.65	17.9
2.26	47.9	5.38	15.6	6.89	9.7
2.93	38.5	5.56	19.4	7.86	9.4
3.16	48.4	5.61	16.5	8.16	10.7
4.26	25.0	5.91	14.0	9.10	5.4
4.34	25.6	5.95	17.9	9.36	6.0

The model we shall use for these $n = 30$ pairs (x_i, y_i) is to treat the y_i as conditionally independent given the x_i, and with

$$y_i = a \exp(-b x_i) \varepsilon_i \quad \text{for } i = 1, \ldots, 30, \tag{1}$$

where the ε_i are taken as i.i.d. from the gamma distribution with parameters (c, c), i.e. with density

$$g_{c,c}(u) = \frac{c^c}{\Gamma(c)} u^{c-1} \exp(-cu) \qquad \text{for } u > 0.$$

(a) First plot the data, along with an ordinarily estimated linear regression line. Briefly discuss the differences between ordinary linear regression and the current model (1), and explain how one may interpret its three parameters. If you see reasons why the model (1) might be expected to work better than ordinary linear regression, give them.

(b) Spend a little time generating and then looking at pseudo-data (x_i, y_i') in your computer, for different choices of the parameter triple (a, b, c), where you keep the x_i as given above but generate y_i' from the model. Use this to put up *some* not unreasonable prior for (a, b, c). It does not need to be particularly good, but you ought to demonstrate that when you generate data from a model sampled from your prior, then these are at least not very unreasonable.

(c) Show that the likelihood function may be expressed as

$$L_n(a, b, c) = \prod_{i=1}^{n} \Big\{ g_{c,c}\Big(\frac{y_i}{a\exp(-bx_i)}\Big) \frac{1}{a\exp(-bx_i)} \Big\}.$$

Find the maximum likelihood estimates of the three model parameters. (You are expected to do this by programming the log-likelihood function and then using suitable software, and where you may need to spend a little time finding a well-working starting point for the algorithm in question. I find parameter estimates 90.8942, 0.2849, 44.0710.) Display the estimated regression curve with the data, and comment.

(d) Regardless of your efforts under point (b) you shall now work with the perhaps crude prior that takes a, b, c as independent and uniform over the intervals $[30, 100]$, $[0.1, 5.0]$, $[0.5, 90.0]$, respectively. Set up a Markov Chain Monte Carlo scheme to generate samples (a, b, c) from the resulting posterior distribution. Use output from such a chain to give [1] 0.05, 0.50, 0.95 quantiles for the posterior distribution for the three parameters and [2] estimated correlations between them. Briefly compare these results with those for the 'lazy Bayesian' who uses the normal approximation associated with maximum likelihood analysis.

(e) Compute 0.05, 0.50, 0.95 quantiles of the posterior distribution of t_0, the 'half-time' for a car, i.e. the time point where its value is expected to be half the original price.

(f) Suppose you have a car that is $x_0 = 5.0$ years old. You contemplate selling it and wonder what price you will get. You are asked now to work with and display *two* distributions, along with their 0.05, 0.50, 0.95 quantiles. The first is the posterior distribution of μ, its expected price, expressed as percentage of the original price of your car. The second is the predictive distribution of this not-yet-observed price Y itself. Comment briefly on your findings.

Exercise 3

The tendency in modern civilisation is to make the world uniform, observes the first non-European winner of the Nobel Prize in literature (in a conversation in 1930). This exercise investigates ways of estimating uniformity parameters across different experiments.

(a) Suppose Y is a single observation from the uniform distribution on $[0, \theta]$, with θ an unknown parameter. Find the mean and standard deviation of Y.

(b) Then find a formula for the risk function $\mathrm{E}_\theta(cY-\theta)^2$ of the estimator cY as an estimator of θ under squared error loss. Show that the best estimator of this type is $\widehat\theta = 1.5\,Y$ and give a formula for its risk function.

(c) Suppose now that θ has a prior density of the form

$$p_b(\theta) = b^2\theta\exp(-b\theta) \quad \text{for } \theta > 0,$$

where b is a positive prior parameter (taken so far to be a known number). Find the explicit posterior density when Y is observed to be some value y. Show in fact that $\theta\,|\,y$ can be represented as $y + \mathrm{Expo}(b)$, where $\mathrm{Expo}(b)$ is an exponential variable with parameter b (i.e. with density $b\exp(-bz)$ for $z > 0$).

(d) Find the Bayes estimator $\widehat\theta_B$ for the prior above, under squared error loss. Find the risk function also of this estimator. For what range of θ values is the Bayes estimator performing better than the natural frequentist estimator $1.5\,Y$ found above?

(e) In the setup above, with θ having the $p_b(\theta)$ prior and $Y\,|\,\theta$ is uniform $[0,\theta]$, find the mean and variance of Y. Find also the full marginal distribution of Y. (You may or may not find it convenient to answer the second question first.)

(f) Assume now that there are independent observations $Y_1,\ldots,Y_n$ from n different uniform experiments, over intervals $[0,\theta_1],\ldots,[0,\theta_n]$. The risk of a procedure estimating this ensemble of uniform parameters, with $\widetilde\theta_i$ estimating θ_i, is to be measured via the average squared error loss function

$$L(\theta,\widetilde\theta) = n^{-1}\sum_{i=1}^n(\widetilde\theta_i-\theta_i)^2.$$

Find first the risk function for the frequentist method using $\widehat\theta_i = 1.5\,Y_i$ for $i = 1,\ldots,n$.

(g) For the prior that takes $\theta_1,\ldots,\theta_n$ independent from the same prior $b^2\theta\exp(-b\theta)$, what is the Bayes estimator $(\widehat\theta_{B,1},\ldots,\widehat\theta_{B,n})$? When the fine-tuning parameter b of the prior cannot be easily agreed upon, argue from an empirical Bayes perspective that a natural estimator is in fact

$$\theta_i^* = Y_i + \bar Y, \quad \text{where } \bar Y = n^{-1}\sum_{i=1}^n Y_i.$$

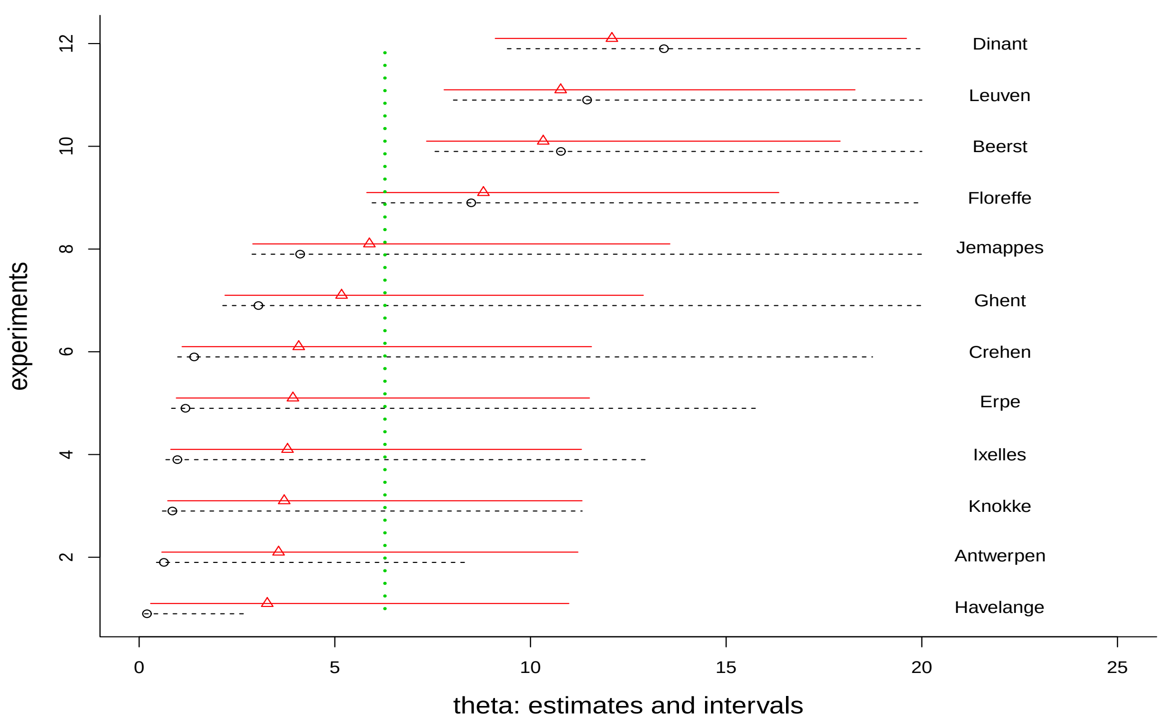
(h) Find the risk function for the empirical Bayes method. Try to characterise the set of parameter values $(\theta_1,\ldots,\theta_n)$ for which the empirical Bayes method performs better than the frequentist 'each experiment separately' method. For this you may find it fruitful to analyse risk functions in terms of average value $\bar\theta$ and spread $\kappa^2 = n^{-1}\sum_{i=1}^n(\theta_i-\bar\theta)^2$. Briefly discuss your findings.

(i) For the purposes of illustration, suppose now that there were twelve different experiments as described above, carried out in twelve different Belgian laboratories and associated with unknown parameters θ_i, yielding data y_i as follows:

0.422, 7.185, 0.937, 8.940, 0.789, 5.657, 2.032, 0.133, 0.651, 2.743, 0.566, 7.631.

In addition to computing the frequentist and empirical Bayes estimates one wishes a full Bayesian analysis, including 90% credibility intervals for each θ_i. Try to carry out such an analysis, where the setup is as above, but with a noninformative start prior $\pi(b) = b^{-1}$ for the hyperparameter b of the $p_b(\theta_i) = b^2\theta_i \exp(-b\theta_i)$ prior used for the twelve θ_i, etc. You are not necessarily required to construct an identical version of the plot below, but you should aim at computing and displaying such 90% credibility intervals for the θ_i.

(j) Finally include in your analysis 0.05, 0.50, 0.95 quantiles of the posterior distributions of [1] overall mean $\bar\theta$, [2] overall spread κ, and [3] the ratio $\kappa/\bar\theta$. Comment on your findings.



For the twelve uniform experiments, ordered here from the smallest to the largest observation, the figure displays point estimates $\widehat\theta_i$ (small circles, on the dotted lines) and empirical Bayes estimates θ_i^* (small triangles, on the full lines). Also indicated are ordinary experiment-by-experiment 90% confidence intervals (dotted lines; the six top confidence intervals actually extend rather longer than to what I chose as upper plotting point limit for the figure), and 90% Bayes credibility intervals (full lines). The vertical line indicates the overall estimate $n^{-1}\sum_{i=1}^n \theta_i^*$.

Exercise 4 – for the PhD students taking STK 9020 only

The number of PhD candidates in the kingdom of Norway has more than doubled over the past ten years (from 4124 in 2002 to 9041 in 2011, actually). This is mindboggingly spellbindingly fantastic.

By the general rules of the Faculty of Mathematics and Natural Sciences those taking the PhD STK 9020 version of this course are required to be examined and evaluated in a somewhat different manner from those taking the STK 4020 version. We solve this here by demanding that the STK 9020 candidates work also with the present Exercise 4 (those among the STK 4020 students eager to work with this exercise too are however welcomed to do so). This exercise is as follows.

I have uploaded Andrew Gelman's 2008 article *Objections to Bayesian Statistics* to the course website, taken from the *Bayesian Analysis* online journal, along with discussion contributions by José Bernardo, Jay Kadane, Stephen Senn, Larry Wasserman, and Gelman's rejoinder (Gelman is himself a prominent Bayesian, but chose nevertheless to air some of his objections to parts of Bayesian practice in this manner).

Read through the Gelman 2008 paper and ensuing discussion, and write up a short essay (perhaps three pages?) where you (a) briefly sum up just a few points from this discussion and (b) choose one or twho of these themes for further elaboration from your side. You are very much invited to present your own views as relevant for your own work (ongoing or prospective). I emphasise that you are not necessarily required to care about all of the Gelman 2008 discussion; you are instead supposed to find something there worth discussing further from your own perspectives or tastes.

Appendix: Measuring good-temperedness and bad-temperedness

To help assess or ascertain whether you or persons near you are good-tempered or bad-tempered, perhaps via interviews with members of your own family (across several generations), below are the criteria Galton instructed his data compilers to use. Matching a high enough number of epithets on the 'good' list makes you a good-tempered person, and correspondingly with the 'bad' list. Galton had such data for nearly two thousand individuals, and appears to have been specifically interested in the inheritance aspect, how and to what degree character traits are passed on to the next generation. The data set used for this project's Exercise 1 are extracted from these data, using the 111 married couples.

Good temper: amiable, buoyant, calm, cool, equable, forbearing, gentle, good, mild, placid, self-controlled, submissive, sunny, timid, yielding. (15 epithets in all.)

Bad temper: acrimonious, aggressive, arbitrary, bickering, capricious, captious, choleric, contentious, crotchety, decisive, despotic, domineering, easily offended, fiery, fits of anger, gloomy, grumpy, harsh, hasty, headstrong, huffy, impatient, imperative, impetuous, insane temper, irritable, morose, nagging, obstinate, odd-tempered, passionate, peevish, peppery, proud, pugnacious, quarrelsome, quick-tempered, scolding, short, sharp, sulky, sullen, surly, uncertain, vicious, vindictive. (46 epithets in all.)

Discussing these criteria at some length, Galton includes the following comment: 'We can hardly, too, help speculating uneasily upon the terms that our own relatives would select as most appropriate to our particular selves.'