Mattias Villani
Department of Statistics
Stockholm University

Computer Lab 2

You are recommended to use R for solving the labs.
You work and submit your labs in pairs, but both of you should contribute equally and understand all parts of your solutions.
`It is not allowed to share exact solutions` with other student pairs.
Submit your solutions via Athena.

---

1. **Linear and polynomial regression**
   The dataset `tempLinkoping` contains daily temperatures (in Celcius degrees) at Malmslätt, Linköping over the course of the year 2016 (366 days since 2016 was a leap year). The response variable is *temp* and the covariate is

$$time = \frac{\text{the number of days since beginning of year}}{366}.$$

   The task is to perform a Bayesian analysis of a quadratic regression

$$temp = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot time^2 + \varepsilon, \; \varepsilon \overset{iid}{\sim} N(0, \sigma^2).$$

   (a) *Determining the prior distribution of the model parameters.* Your task is to set the hyperparameters μ_0, Ω_0, ν_0 and σ_0^2 in the conjugate prior to sensible values. Start with $\mu_0 = (-10, 100, -100)^T$, $\Omega_0 = 0.01 \cdot I_3$, $\nu_0 = 4$ and $\sigma_0^2 = 1$. Check if this prior agrees with your prior opinions by simulating draws from the joint prior of all parameters and for every draw compute the regression curve. This gives a collection of regression curves, one for each draw from the prior. Do the collection of curves look reasonable? If not, change the prior hyperparameters until the results agree with your prior beliefs about the regression curve. [Hint: the R package `mvtnorm` will be handy. And use your *Inv*-χ^2 simulator from Lab 1.]

   (b) Write a program that *simulates from the joint posterior distribution* of β_0, β_1,β_2 and σ^2. Plot the marginal posteriors for each parameter as a histogram. Also produce another figure with a scatter plot of the temperature data and overlay a curve for the posterior median of the regression function $f(time) = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot time^2$, computed for every value of *time*. Also overlay curves for the lower 2.5% and upper 97.5% posterior credible interval for $f(time)$. That is, compute the 95% equal tail posterior probability intervals for every value of *time* and then connect the lower and upper limits of the interval by curves. Does the interval bands contain most of the data points? Should they?

   (c) It is of interest to locate the *time* with the highest expected temperature (that is, the *time* where $f(time)$ is maximal). Let's call this value $\tilde{x}$. Use the simulations in b) to simulate from the *posterior distribution of* $\tilde{x}$. [Hint: the regression curve is a quadratic. You can find a simple formula for $\tilde{x}$ given β_0, β_1 and β_2.]

   (d) Say now that you want to *estimate a polynomial model of order* 7, but you suspect that higher order terms may not be needed, and you worry about over-fitting. Suggest a suitable prior that mitigates this potential problem. You do not need to compute the posterior, just write down your prior. [Hint: the task is to specify μ_0 and Ω_0 in a smart way.]

2. **Posterior approximation for classification with logistic regression**
The dataset `womenWork` contains $n = 200$ observations on the following nine variables:

Variable	Data type	Meaning	Role
`work`	Binary	Whether or not the woman works	Response
`constant`	1	Constant to the intercept	Feature
`husbandInc`	Numeric	Husband's income	Feature
`educYears`	Counts	Years of education	Feature
`expYears`	Counts	Years of experience	Feature
`expYears2`	Numeric	(Years of experience/10)^2	Feature
`age`	Counts	Age	Feature
`nSmallChild`	Counts	Number of child ≤ 6 years	Feature
`nBigChild`	Counts	Number of child > 6 years	Feature

(a) Consider the logistic regression

$$\Pr(y = 1|\mathbf{x}) = \frac{\exp\left(\mathbf{x}^T\beta\right)}{1 + \exp\left(\mathbf{x}^T\beta\right)},$$

where y is the binary variable with $y = 1$ if the woman works and $y = 0$ if she does not. $\mathbf{x}$ is a 8-dimensional vector containing the eight features (including a one for the constant term that models the intercept). Fit the logistic regression using maximum likelihood estimation by the command: `glmModel <- glm(work ~ 0 + ., data = womenWork, family = binomial)`. Note how I added a zero in the model formula so that R doesn't add an extra intercept (we already have an intercept term from the `constant` feature). Note also that a dot (.) in the model formula means to add all other variables in the dataset as features. `family = binomial` tells R that we want to fit a logistic regression.

(b) Now the fun begins. Our goal is to approximate the posterior distribution of the 8-dim parameter vector β with a multivariate normal distribution

$$\beta|\mathbf{y}, \mathbf{X} \sim N\left(\tilde{\beta}, J_{\mathbf{y}}^{-1}(\tilde{\beta})\right),$$

where $\tilde{\beta}$ is the posterior mode and $J(\tilde{\beta}) = -\frac{\partial^2 \ln p(\beta|\mathbf{y})}{\partial\beta\partial\beta^T}|_{\beta=\tilde{\beta}}$ is the observed Hessian evaluated at the posterior mode. Note that $\frac{\partial^2 \ln p(\beta|\mathbf{y})}{\partial\beta\partial\beta^T}$ is an 8×8 matrix with second derivatives on the diagonal and cross-derivatives $\frac{\partial^2 \ln p(\beta|\mathbf{y})}{\partial\beta_i\partial\beta_j}$ on the off-diagonal. Now, both $\tilde{\beta}$ and $J(\tilde{\beta})$ are computed by the `optim` function in R. You can use my R notebook code `https://github.com/mattiasvillani/BayesLearnCourse/raw/master/Notebooks/R/SpamOptim.Rmd` as a guide, but don't copy my code, write your own. Use the prior $\beta \sim \mathcal{N}(0, \tau^2 I)$, with $\tau = 10$. Your report should include your code as well as numerical values for $\tilde{\beta}$ and $J_{\mathbf{y}}^{-1}(\tilde{\beta})$ for the `womenWork` data. Compute an approximate 95% credible interval for the coefficient on `nSmallChild`. Would you say that this feature is an important determinant of the probability that a women works?

(c) Write a function that simulates from the predictive distribution of the response variable in a logistic regression. Use your normal approximation from 2(b). Use that function to simulate and plot the predictive distribution for the `work` variable for a 40 year old woman, with two children (3 and 9 years old), 8 years of education, 10 years of experience, and a husband with an income of 10.
[Hint: the R package `mvtnorm` will again be handy. And remember my discussion on how Bayesian prediction can be done by simulation.]