TMA4300 Computer Intensive Statistical Methods

Exercise 2, Spring 2024

Solutions must be handed in no later than Sunday the **17$^{\text{th}}$ of March 2024, at 23:59**. All answers including derivations, computer code and graphics (all in one pdf document!) should be submitted in ovsys as specified on the course home page.

Problem 1

In this problem, we will look at a portion of the Tokyo rainfall dataset, a famous dataset with daily rainfall data from 1951–1989. We will consider the response to be whether the amount of rainfall exceeded 1mm over the given time period:

$$y_t|x_t \sim \text{Bin}(n_t, \pi(x_t)), \quad \pi(x_t) = \frac{\exp\{x_t\}}{1+\exp\{x_t\}} = \frac{1}{1+\exp\{-x_t\}},$$

for n_t being 10 for $t = 60$ (February 29th) and 39 for all other days in the year, and $\pi(x_t)$ being the probability of rainfall exceeding 1mm for days $t = 1, \ldots, T$ and $T = 366$. Note that x_t is the logit probability of exceedence and can be obtained from $\pi(x_t)$ ($\pi(\cdot)$ is known as the 'expit' or 'inverse logit' function) via the logit function: $x_t = log(\pi(x_t)/(1-\pi(x_t)))$. We assume conditional independence among the $y_t|x_t$ for all $t = 1, \ldots, 366$.

a) Begin by downloading the Tokyo rainfall dataset from the course wiki page. Explore the dataset, plot the response as a function of t, and describe any patterns that you see.

b) Obtain the likelihood of y_t depending on parameters $\pi(x_t)$ for $t = 1, \ldots, 366$.

We will apply a Bayesian hierarchical model to the dataset, using a random walk of order 1 (RW(1)) to model the trend on a logit scale,

$$x_t = x_{t-1} + u_t,$$

for $u_t \overset{iid}{\sim} N(0, \sigma_u^2)$ so that,

$$p(\mathbf{x} \mid \sigma_u^2) \propto \prod_{t=2}^{T} \frac{1}{\sigma_u} \exp\left\{-\frac{1}{2\sigma_u^2}(x_t - x_{t-1})^2\right\}. \tag{1}$$

We will place the following inverse gamma prior on σ_u^2,

$$p(\sigma_u^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}(1/\sigma_u^2)^{\alpha+1}\exp\{-\beta/\sigma_u^2\},$$

for shape α and scale β. Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)^T$, $\mathbf{x} = (x_1 \; \ldots \; x_T)^T$, and $\boldsymbol{\pi} = (\pi(x_1) \; \ldots \; \pi(x_T))^T$.

c) Find the conditional $p(\sigma_u^2|\boldsymbol{y}, \mathbf{x})$. If the conditional is a named distribution, name it along with its associated parameters.

d) Consider the conditional prior proposal distribution, $Q(\mathbf{x}'_{\mathcal{I}}|\mathbf{x}_{-\mathcal{I}}, \sigma_u^2, \boldsymbol{y}) = p(\mathbf{x}'_{\mathcal{I}}|\mathbf{x}_{-\mathcal{I}}, \sigma_u^2)$, where $\mathbf{x}'_{\mathcal{I}}$ is the proposed values for $\mathbf{x}_{\mathcal{I}}$, $\mathcal{I} \subseteq \{1, \ldots, 366\}$ is a set of time indices, and $\mathbf{x}_{-\mathcal{I}} = \mathbf{x}_{\{1,\ldots,366\}\setminus\mathcal{I}}$ is $\mathbf{x}$ subset to include all indices other than those in $\mathcal{I}$. Show that the resulting acceptance probability is given by the ratio of likelihoods:

$$\alpha(\mathbf{x}'_{\mathcal{I}}|\mathbf{x}_{-\mathcal{I}}, \sigma_u^2, \boldsymbol{y}) = \min\left\{1, \; \frac{p(\boldsymbol{y}_{\mathcal{I}}|\mathbf{x}'_{\mathcal{I}})}{p(\boldsymbol{y}_{\mathcal{I}}|\mathbf{x}_{\mathcal{I}})}\right\}.$$

Note that the density specified by (1) is improper (as discussed by e.g. Lavine & Hodges 2012, https://doi.org/10.1080/00031305.2012.654746), for example for $T = 2$, the density takes the shape of a infinite 'Gaussian ridge' centered around the line given by $x_1 = x_2$. It follows that (1) can be rewritten as

$$p(\mathbf{x}|\sigma_u^2) \propto \exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}\right\} \tag{2}$$

resembling a multivariate normal density but where the impropriety of the density translates into the precision matrix

$$\mathbf{Q} = \frac{1}{\sigma_u^2}\begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & & \ddots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}.$$

having one zero eigenvalue.

e) Partitioning the components of $\mathbf{x}$ into two subvectors writing

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{pmatrix},$$

and partitioning the precision matrix in the same way as

$$\mathbf{Q} = \begin{pmatrix} \boldsymbol{Q}_{AA} & \boldsymbol{Q}_{AB} \\ \boldsymbol{Q}_{BA} & \boldsymbol{Q}_{BB} \end{pmatrix},$$

show that the $\mathbf{x}_A$ has a proper multivariate normal distribution conditional on $\mathbf{x}_B$

with conditional mean and precision for given by

$$\boldsymbol{\mu}_{A|B} = -\boldsymbol{Q}_{AA}^{-1}\boldsymbol{Q}_{AB}\mathbf{x}_B$$
$$\boldsymbol{Q}_{A|B} = \boldsymbol{Q}_{AA}.$$

Hint: Do blockwise matrix multiplication and use the fact that the conditional density is (as always) proportional to the joint (improper) density.

f) Implement an MCMC sampler for the posterior $p(\boldsymbol{\pi}, \sigma_u^2|\boldsymbol{y})$ using MH steps for individual x_t parameters using the conditional prior, $p(x_t|\mathbf{x}_{-t}, \sigma_u)$, and Gibbs steps for σ_u^2. Assume $\alpha = 2$ and $\beta = 0.05$, which is an informative prior placing approximately 95% of the prior mass of σ_u^2 between 0.01 and 0.25. Run the MCMC algorithm for 50,000 iterations, where every element of $\mathbf{x}$ is updated per iteration, and use `proc.time()[3]` to calculate the computation time and acceptance rates. Give traceplots, histograms, and estimated autocorrelation functions for σ_u^2, $\pi(x_1)$, $\pi(x_{201})$, and $\pi(x_{366})$. Provide central estimates and 95% credible intervals for σ_u^2, and a graph showing 95& credible limits and the posterior mean $\pi(x_t)$ as function of $t = 1, 2, \ldots, 366$ and compare predictions for $\boldsymbol{\pi}$ and associated uncertainties to y_t/n_t as a function of t. Do the traceplots show the Markov chain converged? Describe your findings. (**Hint:** multiplying many probabilities together in R generally leads to poor results. Instead, it is better to work on a log scale and simplify analytically when possible, exponentiating at the end.)

g) Repeat the previous exercise, except now rather than updating individual x_t parameters, use a conditional prior proposal involving $p(\mathbf{x}_{(a,b)}|\mathbf{x}_{-(a,b)}, \sigma_u^2)$, where $\mathbf{x}_{(a,b)} = (x_a \ \ldots \ x_b)^T$ choosing intervals of length M. Explore different values of tuning parameter M, and explain your choice M. Why might incorporating a block step over the $\mathbf{x}_{(a,b)}$ parameters in this way improve efficiency of your MCMC sampler? Why might it do the opposite, depending on M? (**Hint:** precomputing $\boldsymbol{Q}_{AA}^{-1}\boldsymbol{Q}_{AB}$ as well as the Cholesky decomposition of $\boldsymbol{Q}_{AA}^{-1}$ will speed up results. You may need to do this 3 times: once for $a = 1$ and $b = M$, once for $a > 1$ and $b < 366$, and once for $b = 366$. As in *d*), every element of $\mathbf{x}$ should be updated per iteration, but when $b = 366$ you may need either a smaller block or for it to overlap with the previous block.)

Problem 2

We will continue looking at the Tokyo rainfall dataset, only now using INLA rather than MCMC. INLA can be installed and loaded into R with the following code:

```
install.packages("INLA",repos=c(getOption("repos"),
```

```
INLA="https://inla.r-inla-download.org/R/stable"), dep=TRUE)
library("INLA")
```

After loading in the Tokyo rainfall dataset, we can fit the same model as in the previous problem with the following code in R:

```
control.inla = list(strategy="simplified.laplace", int.strategy="ccd")
mod <- inla(n.rain ~ -1 + f(day, model="rw1", constr=FALSE),
            data=rain, Ntrials=n.years, control.compute=list(config = TRUE),
            family="binomial", verbose=TRUE, control.inla=control.inla)
```

Note that here we use a simplified Laplace approximation and 'ccd' integration rather than Laplace approximation and grid integration respectively. Run `?control.inla` for more information on the two options. We remove the intercept with the `-1` option, and use a RW(1) by passing the `model="rw1"` option to the `f()` function. Run `inla.doc("rw1")` for documentation provided by INLA on its built-in RW(1) model.

a) Compare the predictions and uncertainties of INLA with those of your previous Markov chains, and again use `proc.time()[3]` to calculate the computation time. Describe your findings. Note that `mod$summary.fitted.values` contains predictions and 95% CIs (what do the other 'summary' objects in `mod` contain?). Also, make sure to use the same priors as in problem 1. (**hint**: What should the prior on the intercept be? See `?control.fixed` for information on the prior for the intercept, and `?f` along with `inla.doc("rw1")` and `inla.doc(X)` where `X` is a character vector giving INLA's name for a prior for information on how to include a prior for σ_u^2. INLA places a prior on the log precision rather than the variance, so make sure to transform the prior accordingly.)

b) How robust are the results to the two `control.inla` inputs we have used? See `?control.inla`.

c) Consider the following model in INLA:

```
mod <- inla(n.rain ~ f(day, model="rw1", constr=TRUE),
            data=rain, Ntrials=n.years, control.compute=list(config = TRUE),
            family="binomial", verbose=TRUE, control.inla=control.inla)
```

   How is this different from the previous model mathematically, assuming the prior for σ_u^2 is set to be the same? Are its predictions significantly different, and why are they/are they not different?

Problem 3

Implement the same model as in problem 1 and 2 using the R-package RTMB. You will first need to write R-function computing (negative log of) the joint likelihood $\pi(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})$. Passing this function to `MakeADFun` produces a new R-function that computes the (negative log) of the Laplace approximation of the marginal likelihood $\pi(\mathbf{y}|\boldsymbol{\theta}) = \int \pi(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}$ using automatic differentiation. Then use e.g. `nlminb` on the approximate marginal likelihood to obtain maximum likelihood estimates of the model parameters. See `?sdreport` and the RTMB-documentation for details. Briefly compare the results to those obtained in problem 1 and 2.

Oral presentations

Date	Problem	Team
19.03	Ex2: Problem 1 a)-c)	TBD
19.03	Ex2: Problem 1 d)	TBD
19.03	Ex2: Problem 1 e)	TBD
19.03	Ex2: Problem 1 f)	TBD
19.03	Ex2: Problem 1 g)	TBD
19.03	Ex2: Problem 2 a)-b)	TBD
19.03	Ex2: Problem 2 c)	TBD
19.03	EX2: Problem 3	TBD