# Assignment #7 – Machine Learning – Professor Haugh

## **Jaime Gacitua**

## **jg3499**

## **Due Friday April 22, 2016**

## Question 1

1. We have the following distribution

It can be re-written as,

Characterizing this function as an exponential family of the form

We can identify,

1. Substituting in the expression for the conjugate prior,

In order to transform this expression to the form,

We have to create the following equivalences,

From the first equation,

From the second relationship,

From here, we can state the following equations and solutions,

## Question 2

1. Let x follow a Bernoulli distribution, with parameter.
2. If the conjugate prior for is a beta distribution, it must have the form.
3. Computing the posterior distribution, considering i.i.d realizations of the r.v.

We recognize from here that the posterior is

## Question 3

1. We start with the expression

Using the definition of, and the variance function of correlated values,

Using the definition of the autocorrelation function, over the,

We know that, and, therefore,

From the definition of,

1. If we assume that follows an AR(1) process, , and .

From the definition of, and using this particular case for

From the definition of known series,

As n grows, becomes smaller and smaller.

We will assume n is large, yielding the convenient approximation,

1. We will start on the left hand side of the relationship we want to prove, and using the identity provided in the homework,

Using (2),

Since the are i.i.d. ,

With this result, we can re-arrange the terms

If we consider as data points, we can remove the expectation, yielding an estimator for,

This is an unbiased estimator only if is an unbiased estimator for. This condition must be true because if we take expectation of the right hand side, it will yield zero only if.

1. From part (a) we demonstrated that

I will construct an unbiased estimator for, using realizations of the process (each of length).

Where,

The average of therealization, the overall average value across all the realizations.

To justify (1), we can see that we found an unbiased estimator,

Relating this estimator to (1), we can see that the expressions are in a sense equivalent,

The termis represented by

The term is represented by, but now we will gather the average variance, expanding this term to.

## Question 4

1. The implemented Gibbs sampler can be found in the appendix, in the code for question 4.
2. The Gelman-Rubin diagnostic was implemented and can be found in the appendix, in the code for question 4.

The output of the diagnostic can be found in part (c)

1. A total of chains were created, and. The output table is presented below. The statistic achieved is better than the textbook. It is interesting to see the high dispersion of. It seems to be a parameter more difficult to estimate with certainty.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Parameter | 2.5% | 25% | 50% | 75% | 97.5% | R |
| Theta1 | 56.3 | 60.0 | 61.5 | 62.8 | 66.1 | 1.00002 |
| Theta2 | 61.4 | 64.5 | 65.7 | 67.0 | 70.4 | 1.00000 |
| Theta3 | 63.2 | 66.3 | 67.5 | 68.8 | 72.4 | 1.00001 |
| Theta4 | 56.4 | 59.7 | 60.8 | 61.9 | 64.8 | 1.00001 |
| Mu | 53.4 | 62.0 | 63.9 | 65.8 | 74.5 | 1.00000 |
| Sigma | 2.3 | 2.9 | 3.3 | 3.8 | 5.9 | 1.00004 |
| Tau | 1.6 | 3.5 | 5.4 | 8.7 | 30.6 | 1.00001 |

## Question 5

I programmed the MCMC, using the bigram matrix.

The MCMC allowed me to uncode the following message:

**it bos o pright cald dof in ouril ond the clacks bere striking thirteen**

This is pretty close to English language, allowing to assume the following translation:

**It was a bright cold day in April and the clocks were striking thirteen.**

I ran a total of chains, and each one of them to a total of iterations. I used the log plausibility to assess the likelihood of the resulting function f. I recorded the best log plausibility attained throughout all the chains and iterations, and used the respective f function to translate the coded message.

The highest log plausibility achieved was -158.10, which translates to a plausibility of . It is interesting to note that we deal with very small levels of plausibility. To avoid numerical problems, it is useful to work with the log plausibility.

It is interesting to note that as iterations progress, the log plausibility can go up or down. The probability of going down is lower than going up, because we flip a “biased coin” every time we achieve a worse plausibility.

A summary of the iterations is displayed in the following graph.



# Appendix

## Question 4 – Code

rm(list=ls())

require(stats)

require(geoR)

#### Dataset ----

dataA = matrix(data = c(62, 60, 63, 59))

dataB = matrix(data = c(63, 67, 71, 64, 65, 66))

dataC = matrix(data = c(68, 66, 71, 67, 68, 68))

dataD = matrix(data = c(52, 62, 60, 61, 63, 64, 63, 59))

# Calculate total number of data points

n <- length(dataA) + length(dataB) + length(dataC) + length(dataD)

#### Functions ----

calc.sigma2 <- function(dataA, dataB, dataC, dataD, thetas){

# initialize

sigma2 <- 0

for(i in dataA){

sigma2 <- sigma2 + (i-thetas[1])^2

}

for(i in dataB){

sigma2 <- sigma2 + (i-thetas[2])^2

}

for(i in dataC){

sigma2 <- sigma2 + (i-thetas[3])^2

}

for(i in dataD){

sigma2 <- sigma2 + (i-thetas[4])^2

}

n <- length(dataA) + length(dataB) + length(dataC) + length(dataD)

sigma2 <- sigma2 / n

## Generate inv-chi squared to get sigma2

sigma2 <- rinvchisq(n = 1, df = n, scale = sigma2)

return(sigma2)

}

calc.tau2 <- function(thetas.temp, mu.temp){

# initialize

tau2.temp <- 0

for(j in thetas.temp){

tau2.temp <- tau2.temp + (j-mu.temp)^2

}

J <- length(thetas.temp)

tau2.temp <- tau2.temp / (J - 1)

## Generate chi squared

tau2.temp <- rinvchisq(n = 1, df = (J-1), scale = tau2.temp+0.001)

return(tau2.temp)

}

calc.theta.param <- function(data, mu, sigma2, tau2){

theta.param <- ( mu / tau2 + length(data) \* mean(data) / sigma2 ) /

(1 / tau2 + length(data) / sigma2)

return(theta.param)

}

calc.vega.param <- function(data, tau2, sigma2){

vega.param <- 1 / (1 / tau2 + length(data) / sigma2)

return(vega.param)

}

calc.thetas <- function(dataA, dataB, dataC, dataD, mu, sigma2, tau2){

theta <- mat.or.vec(4,1)

# For each theta draw from its distribution

theta.param <- calc.theta.param(dataA, mu, sigma2, tau2)

vega.param <- calc.vega.param(dataA, sigma2, tau2)

theta[1] <- rnorm(1, mean = theta.param, sd = sqrt(vega.param))

theta.param <- calc.theta.param(dataB, mu, sigma2, tau2)

vega.param <- calc.vega.param(dataB, sigma2, tau2)

theta[2] <- rnorm(1, mean = theta.param, sd = sqrt(vega.param))

theta.param <- calc.theta.param(dataC, mu, sigma2, tau2)

vega.param <- calc.vega.param(dataC, sigma2, tau2)

theta[3] <- rnorm(1, mean = theta.param, sd = sqrt(vega.param))

theta.param <- calc.theta.param(dataD, mu, sigma2, tau2)

vega.param <- calc.vega.param(dataD, sigma2, tau2)

theta[4] <- rnorm(1, mean = theta.param, sd = sqrt(vega.param))

return(theta)

}

calc.mu <- function(thetas, tau2){

mu.param <- sum(thetas) / length(thetas)

mu <- rnorm(1, mean = mu.param, sd = sqrt(tau2 / length(thetas)))

return(mu)

}

iterate <- function(thetas, mu, sigma2, tau2, dataA, dataB, dataC, dataD){

thetas <- calc.thetas(dataA = dataA, dataB = dataB, dataC = dataC,

dataD = dataD, mu = mu, sigma2 = sigma2, tau2 = tau2)

mu <- calc.mu(thetas = thetas, tau2 = tau2)

sigma2 <- calc.sigma2(dataA = dataA, dataB = dataB, dataC = dataC,

dataD = dataD, thetas = thetas)

tau2 <- calc.tau2(thetas = thetas, mu = mu)

return.list <- list("sigma2" = sigma2, "tau2" = tau2, "thetas" = thetas, "mu" = mu)

return(return.list)

}

calc.var.psi.chains <- function(psi.chains){

psi.col.means <- colMeans(psi.chains)

psi.mean <- mean(psi.chains)

n <- nrow(psi.chains)

m <- ncol(psi.chains)

B <- n / (m - 1) \* sum((psi.col.means - psi.mean)^2)

W <- 1 / m \* sum( n / (n-1) \* (colMeans(psi.chains\*psi.chains) - colMeans(psi.chains)^2))

var.psi.chains <- (n-1)/n\*W + B/n

R.convergence <- sqrt(var.psi.chains / W)

output = list("var" = var.psi.chains, "B" = B, "W" = W, "R.conv" = R.convergence)

return(output)

}

### Initialization Values ----

m <- 4 # number of chains

n <- 100000 # length of each chain

n0 <- 2\*n # = 2n, 50% of each chain is discarded

chains <- list()

iterations1 <- list()

iterations2 <- list()

psi.sigma2 <- matrix(0, nrow = n, ncol = m)

psi.tau2 <- matrix(0, nrow = n, ncol = m)

psi.mu <- matrix(0, nrow = n, ncol = m)

psi.theta1 <- matrix(0, nrow = n, ncol = m)

psi.theta2 <- matrix(0, nrow = n, ncol = m)

psi.theta3 <- matrix(0, nrow = n, ncol = m)

psi.theta4 <- matrix(0, nrow = n, ncol = m)

chain.number <- 0

### Run the chains ----

for(iter.m in 1:(m/2)){

# Initialize thetas randomly

thetas = mat.or.vec(1,4)

thetas[1] = sample(dataA,1)

thetas[2] = sample(dataB,1)

thetas[3] = sample(dataC,1)

thetas[4] = sample(dataD,1)

# Initialize mu

mu = mean(thetas)

# initial sigma2

sigma2 <- calc.sigma2(dataA, dataB, dataC, dataD, thetas)

# initial tau2

tau2 <- calc.tau2(thetas = thetas, mu = mu)

# collect first iteration parameters

parameters <- list("sigma2" = sigma2, "tau2" = tau2, "thetas" = thetas, "mu" = mu)

# iterate for a total of n0+2n times

counter <- 1

for(iter.n in 1:(n0+2\*n)){

# cat('Iter: ', iter.n, '\n')

# update all parameters

parameters <- iterate(thetas = parameters$thetas,

mu = parameters$mu,

sigma2 = parameters$sigma2,

tau2 = parameters$tau2,

dataA = dataA,

dataB = dataB,

dataC = dataC,

dataD = dataD)

## Save iterations after we have 'burned in' n0

# Save the first m iterations in one chain

if(iter.n == 1){

chain.number <- chain.number + 1

cat('Chain Number: ', chain.number, '\n')

}

if((counter>n0) && (counter<=(n0+n))){

psi.sigma2[counter - n0, chain.number] <- parameters$sigma2

psi.tau2[counter - n0, chain.number] <- parameters$tau2

psi.mu[counter - n0, chain.number] <- parameters$mu

psi.theta1[counter - n0, chain.number] <- parameters$thetas[1]

psi.theta2[counter - n0, chain.number] <- parameters$thetas[2]

psi.theta3[counter - n0, chain.number] <- parameters$thetas[3]

psi.theta4[counter - n0, chain.number] <- parameters$thetas[4]

}

# Save the next m iterations in a second chain

if(iter.n == (n0+n+1)){

chain.number <- chain.number + 1

cat('Chain Number: ', chain.number, '\n')

}

if(counter>(n0+n)){

psi.sigma2[counter - n0 - n, chain.number] <- parameters$sigma2

psi.tau2[counter - n0 - n, chain.number] <- parameters$tau2

psi.mu[counter - n0 - n, chain.number] <- parameters$mu

psi.theta1[counter - n0 - n, chain.number] <- parameters$thetas[1]

psi.theta2[counter - n0 - n, chain.number] <- parameters$thetas[2]

psi.theta3[counter - n0 - n, chain.number] <- parameters$thetas[3]

psi.theta4[counter - n0 - n, chain.number] <- parameters$thetas[4]

}

counter <- counter + 1

}

}

var.psi.theta1 <- calc.var.psi.chains(psi.theta1)

var.psi.theta2 <- calc.var.psi.chains(psi.theta2)

var.psi.theta3 <- calc.var.psi.chains(psi.theta3)

var.psi.theta4 <- calc.var.psi.chains(psi.theta4)

var.psi.sigma2 <- calc.var.psi.chains(psi.sigma2)

var.psi.tau2 <- calc.var.psi.chains(psi.tau2)

var.psi.mu <- calc.var.psi.chains(psi.mu)

var.psi.theta1$R.conv

var.psi.theta2$R.conv

var.psi.theta3$R.conv

var.psi.theta4$R.conv

var.psi.mu$R.conv

var.psi.sigma2$R.conv

var.psi.tau2$R.conv

quants <- c(0.025, 0.25, 0.5, 0.75, 0.975)

quantile(as.vector(psi.theta1, mode = 'numeric'), quants)

quantile(as.vector(psi.theta2, mode = 'numeric'), quants)

quantile(as.vector(psi.theta3, mode = 'numeric'), quants)

quantile(as.vector(psi.theta4, mode = 'numeric'), quants)

quantile(as.vector(psi.mu, mode = 'numeric'), quants)

quantile(as.vector(sqrt(psi.sigma2), mode = 'numeric'), quants)

quantile(as.vector(sqrt(psi.tau2), mode = 'numeric'), quants)

## Question 5 – Code

require(R.matlab)

# text to decode

cvtext = c(19,15,16,24,18,17,16,18,16,8,1,19,20,26,15,16,6,9,14,2,16,2,18,21,16,19,3,16,18,13,1,19,14,16,18,3,

2,16,15,26,27,16,6,14,9,6,10,17,16,24,27,1,27,16,17,15,1,19,10,19,3,20,16,15,26,19,1,15,27,27,3)

## Import Data ----

transition.matrix <- readMat('English\_trans.mat')

# A = bigram transition A(i,j) = Prob(x\_t = i| x\_{t-1} = j)

A <- transition.matrix$A

# S = trigram transition A(i,j,k) = Prob(x\_t = i| x\_{t-1} = j, x\_{t-2} = k)

S <- transition.matrix$S

## Functions -----

calc.log.plaus.bi <- function(A, f, cvtext){

log.plaus <- 0

for (k in 2:length(cvtext)){

log.plaus = log.plaus + log(A[f[cvtext[k]],f[cvtext[k-1]]])

}

return(log.plaus)

}

calc.log.plaus.tri <- function(A, f, cvtext){

log.plaus <- 0

for (k in 3:length(cvtext)){

log.plaus = log.plaus + log(S[f[cvtext[k]],f[cvtext[k-1]],f[cvtext[k-2]]])

}

return(log.plaus)

}

## Clean Data: remove zeros from the transition matrix and renormalize ----

tol <- 1e-6

# Clean A

colSums(A)

min(A)

A <- A + tol

for (j in 1:ncol(A)){

A[,j] = A[,j]/colSums(A)[j]

}

colSums(A)

min(A)

# Clean S

S = S + tol

for (k in 1:27) {

for (l in 1:27) {

sumprob <- 0

for (m in 1:27) {

sumprob <- sumprob + S[m,l,k];

}

if (sumprob>0) {

for (m in 1:27) {

S[m,l,k] <- S[m,l,k] / sumprob;

}

}

}

}

## Settings ----

log.plaus.max <- -99999999

n <- 500000 # Number of iterations per chain

m <- 10 # Number of Chains to simulate

freq.report <- 20000 # Reporting frequency

plaus.hist <- matrix(data = 0, nrow = n / freq.report, ncol = m)

for (chains in 1:m) {

cat('\n Chain # ', chains, '\n')

# Initial function, picked at random

f <- sample(x = c(1:27), size = 27, replace = FALSE)

# Numbrer of iterations

## Loop ----

for (counter in 1:n) {

# Calculate Log likelihood of plausibility of function f

log.plaus <- calc.log.plaus.bi(A, f, cvtext) #+ calc.log.plaus.bi(A, f, cvtext)

if (log.plaus > log.plaus.max) {

log.plaus.max <- log.plaus

f.max.plaus <- f

}

## Make random transposition of 2 letters of f

# Draw 2 random numbers

change <- sample(1:27, 2)

# Transpose those positions

temp <- f[change[2]]

f[change[2]] <- f[change[1]]

f[change[1]] <- temp

# Calculate new log likelihood of plausibility of new function f\*

log.plaus.2 <- calc.log.plaus.bi(A, f, cvtext) #+ calc.log.plaus.bi(A, f, cvtext)

# If the plaus is worse, we flip a biased coin to see if we accept the new function.

if (log.plaus.2 < log.plaus) {

coin <- runif(n = 1, min = 0, max = 1)

if (coin > exp(x = log.plaus.2 ) / exp(x = log.plaus )) {

# We switch back because we got tails

temp <- f[change[2]]

f[change[2]] <- f[change[1]]

f[change[1]] <- temp

}

}

if (counter %% freq.report == 0) {

cat('Progress: ', (counter / n), ', Log Plausibility: ', log.plaus, 'Max: ',log.plaus.max, '\n')

pos <- floor(counter / freq.report)

plaus.hist[pos,chains] <- log.plaus

}

}

}

## Convert to text ----

ctext <- matrix(0, nrow = 1, ncol = length(cvtext))

# Apply the function f

for (k in 1:length(cvtext)) {

ctext[k] <- f.max.plaus[cvtext[k]]

}

# Do the transformation

ctextconv <- matrix('0', nrow = 1, ncol = length(ctext))

for(i in 1:length(ctext)){

if (ctext[i] == 27) {

ctext[i] <- -64

}

ctext[i] <- ctext[i] + 96

}

## Final output ----

cat('Best Log Plausibiilty: ', log.plaus.max, '\n')

ctextorig <- rawToChar(as.raw(cvtext+96))

ctextconv <- rawToChar(as.raw(ctext))

for (k in 1:length(ctextorig)) {

cat(ctextorig[k])

}

cat('\n')

for (k in 1:length(ctextconv)) {

cat(ctextconv[k])

}

## Plot ----

# get the range for the x and y axis

yrange <- range(plaus.hist[1:(n/freq.report),])

# set up the plot

plot(c(1, n/freq.report), yrange, type="n", xlab="Iteration (x20000)",

ylab="Log Plausibility" )

linetype <- c(1:m)

plotchar <- seq(18,18+m,1)

# add lines

for (i in 1:m) {

lines(c(1:(n/freq.report)), plaus.hist[1:(n/freq.report),i], type="b", lwd=1.5,

lty=linetype[i], pch=plotchar[i])

}

# add a title and subtitle

title("Results of Iterations, 10 different starting points, Log Plausibility Function")