UNIVERSITY OF BRISTOL

School of Mathematics

Theory of Inference

MATH35600/MATHM0019

(Paper code MATH–35600)

Solutions Included

May/June 2019   2 hours 30 minutes

This paper contains FOUR questions. All answers will be used for assessment.

Calculators are not permitted in this examination.

The marking scheme is indicative and is intended only as a guide to the relative weighting of the questions. Answers should be concise and to the point. Lengthy imprecise answers and irrelevant information will lose marks. No marks will be lost for minor R errors if the statistical meaning is clear.

Do not turn over until instructed.

1. Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$. If $n \times p$ model matrix $\mathbf{X}$ has QR decomposition $\mathbf{X} = \mathbf{Q}\left[\begin{array}{c}\mathbf{R}\\ \mathbf{0}\end{array}\right]$ and $\mathbf{q} = \mathbf{Q}^T\mathbf{y}$ then the least squares estimator of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{q}_{1:p}$, where $\mathbf{q}_{i:j}$ denotes elements i to j of $\mathbf{q}$.

   Solution comments. *This is mostly standard stuff, but mixing up the notation from the notes a little so that it can't be done by rote-learning. (e)iii is supposed to be a bit trickier.*

   (a) Show that $\mathbf{q} \sim N\left(\left[\begin{array}{c}\mathbf{R}\boldsymbol{\beta}\\ \mathbf{0}\end{array}\right], \mathbf{I}\sigma^2\right)$. [5 marks]

   Solution

   $$E(\mathbf{q}) = E(\mathbf{Q}^T\mathbf{y}) = \mathbf{Q}^T E(\mathbf{y}) = \mathbf{Q}^T\mathbf{X}\boldsymbol{\beta} = \left[\begin{array}{c}\mathbf{R}\boldsymbol{\beta}\\ \mathbf{0}\end{array}\right]$$

   and the covariance matrix of $\mathbf{q}$ is just $\mathbf{V}_q = \mathbf{Q}^T\mathbf{I}\mathbf{Q}\sigma^2 = \mathbf{I}\sigma^2$. Hence since $\mathbf{q}$ is just a linear transformation of a normal random vector, the result is proved.

   (b) Find the distribution of $\|\mathbf{q}_{p+1:n}\|^2/\sigma^2$ and hence or otherwise find an unbiased estimator of σ^2. [3 marks]

   Solution From part a we know that the elements of $\mathbf{q}_{p+1:n}$ are i.i.d. $N(0, \sigma^2)$ hence $\|\mathbf{q}_{p+1:n}\|^2/\sigma^2 = \sum_{j=p+1}^{n} q_j^2/\sigma^2 \sim \sum_{i=1}^{n-p} N(0,1)^2 \sim \chi^2_{n-p}$. $E(\chi^2_{n-p}) = n - p$, so $\hat{\sigma}^2 = \|\mathbf{q}_{p+1:n}\|^2/(n-p)$ is an unbiased estimator of σ^2.

   (c) Explain whether or not $\hat{\boldsymbol{\beta}}$ and $\|\mathbf{q}_{p+1:n}\|^2$ are independent. [3 marks]

   Solution For multivariate Gaussian random variables only, zero covariance implies independence. So, from the results of a, $\mathbf{q}_{p+1:n}$ and $\mathbf{q}_{1:p}$ are independent. Since $\hat{\boldsymbol{\beta}}$ depends only on the former and $\|\mathbf{q}_{p+1:n}\|^2$ only on the latter, then $\hat{\boldsymbol{\beta}}$ and $\|\mathbf{q}_{p+1:n}\|^2$ are also independent.

   (d) Write out the model matrix for the linear model used in the following R code

```
> a
[1] 2 1 1 3 1 1 2 3
Levels: 1 2 3
> x
[1] 4.5 9.5 0.8 2.1 3.2 9.0 6.8 7.8
> mod <- lm(y ~ a + x)
```

   [4 marks]

   Solution Lose marks for unidentifiable or failing to include an intercept

```
> model.matrix(mod)
  (Intercept) a2 a3   x
1           1  1  0 4.5
2           1  0  0 9.5
3           1  0  0 0.8
4           1  0  1 2.1
5           1  0  0 3.2
6           1  0  0 9.0
7           1  1  0 6.8
8           1  0  1 7.8
```

(e) A researcher recruits a group of volunteers to a study and on the basis of a 6 month diary gives them a score for digestive system health. Several covariates are recorded alongside whether or not the volunteers drink grapefruit juice at least twice a week. Modelling of the resulting data finds a strong positive association between drinking grapefruit juice and digestive health.

i. Briefly explain why it is not legitimate to conclude from this study that drinking grapefruit juice improves digestive health. [2 marks]

Solution It is an observational study, and we can not tell the difference between grapefruit juice promoting digestive health and grapefruit juice consumption being correlated with things that cause good digestive health that we have not measured.

ii. Explain, concisely, two ways in which we might be able to modify the study and/or analysis to find out whether drinking grapefruit juice improves digestive health, giving brief explanations of how these approaches allow this. [5 marks]

Solution

A. We could perform a study in which participants are randomly allocated to grapefruit juice drinking or not. The randomization breaks any possible correlation between grapefruit juice consumption and other drivers of digestive health, allowing us to establish causation.

B. Alternatively we could attempt to find an *instrumental variable* which is independent of all plausible direct drivers of digestive health, but is correlated with tendency to drink grapefruit juice. Replacing the grapefruit juice drinking covariate with grapefruit juice drinking regressed on the instrument also breaks the link between the grapefruit juice drinking and the confounders. The problem is finding a valid instrument.

iii. When the study is re-run to find out whether drinking grapefruit juice improves digestive health, a small but highly statistically significant negative effect of grapefruit juice drinking on digestive health is found. How is this possible?

[3 marks]

Solution It could be that grapefruit juice consumption is positively correlated with unmeasured variables that strongly positively affect digestive health, but that if you controlled for those variables, grapefruit juice consumption is harmful. This would yield the observed result.

2. **Solution comments**. *This is mostly about testing understanding of how the Bayesian and frequentist approaches compare. (f) and (g) are supposed to be slightly trickier than the others.* As a rough guide the answers to this question should involve about one sentence per mark.

   (a) Briefly explain the key difference between the treatment of model parameters in Bayesian and frequentist statistical inference. [2 marks]

   Solution In the Bayesian case parameters are treated as random variables, in the frequentist case they are fixed states of nature.

   (b) If $\mathbf{y}$ and $\boldsymbol{\theta}$ denote data and parameter vectors, briefly state how the likelihood $f(\mathbf{y}|\boldsymbol{\theta})$ is used to estimate parameters in frequentist inference, and contrast this with its use in Bayesian inference. [3 marks]

   Solution The observed $\mathbf{y}$ are plugged in and the resulting likelihood function maximized w.r.t. θ. The maximizing θ gives the estimate. In the Bayesian case the likelihood is used to update the prior distribution on θ to a posterior, using Bayes theorem.

   (c) Define a p-value, and give a concise explanation of how to interpret high and low p-values. [3 marks]

   Solution The probability under the null hypothesis of obtaining a test statistic at least as favourable to the alternative hypothesis as that actually observed. Low values are evidence against the null and for the alternative. High values indicate insufficient evidence to reject the null.

   (d) Give a brief explanation of how Newton's method of maximizing a function works. [3 marks]

   Solution Evaluate the function and its first two derivatives at parameter guess θ. Approximate the function by the quadratic matching the derivatives at θ. Maximize the approximation w.r.t. θ to obtain the next guess at the optimum. (bonus for mentioning positive definiteness and step halving).

   (e) State, as concisely as possible how a generalized likelihood ratio test is conducted, including the large sample distributional result used. State the main conditions for the large sample result to be valid. [4 marks]

   Solution The models to be compared must be nested (the null model is a restricted version of the alternative), and the restriction must not amount to placing parameters of the alternative model at the edge of the parameter space. Likelihood must also be sufficiently smooth. We estimate both models by maximum likelihood estimation. Then under the null model in the large sample limit $2(l(\hat{\theta}_1) - l(\hat{\theta}_0)) \sim \chi^2_{p_1-p_0}$, where $p_j = \dim(\theta_j)$. This result can be used to compute a p-value.

   (f) Give the Bayesian equivalent of the likelihood ratio statistic, and briefly explain why it can be interpreted directly, without requiring something like a p-value. [3 marks]

   Solution The Bayes factor is the ratio of the *marginal likelihoods* of the models being compared, that is the expected likelihoods according to the priors. The fact that we look at average likelihoods rather than maximized likelihoods removes the feature that the larger model always has the higher likelihood, so we don't need a p-value to judge whether the alternative likelihood is 'larger enough' to be worth taking notice of. (Bonus mark for mentioning the can of worms this opens.)

(g) Suppose that you have a sample of count data, $y_1, y_2, \ldots y_n$ from a wildlife survey and want to establish whether a Poisson model $y_i \sim \text{Poi}(\lambda)$ or negative binomial model $y_i \sim \text{NB}(\lambda, \theta)$ is more appropriate. The negative binomial distribution is often used for count data that have a higher variance than a Poisson distribution would suggest. When the extra negative binomial parameter, θ, tends to infinity the negative binomial distribution tends to the Poisson distribution.

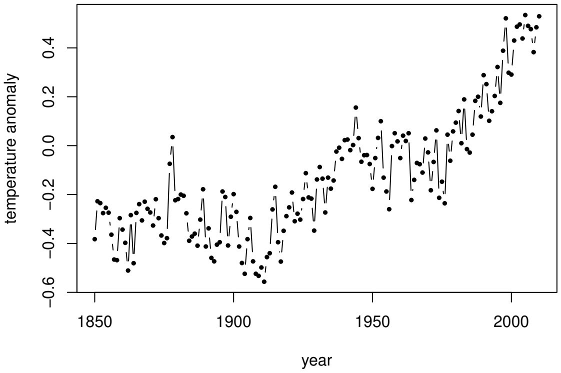
i. Suppose that you fit the two alternative models to the data by maximum likelihood estimation. State with reasons whether AIC, a generalized likelihood ratio test, both or neither can be used to compare the models. [3 marks]

Solution The null model is restricting θ to the edge of the parameter space, so we can not use the GLRT. No such problem with AIC, so we could use that.

ii. Suppose that you instead decide to take a Bayesian approach. From previous surveys you have a well defined prior distribution for λ, but no real information on θ, so decide on a vague exponential prior on $1/\theta$. Briefly explain the main advantages of using Bayes Factors, BIC or DIC for deciding between the models in this case. [4 marks]

Solution The Bayes factor is problematic - we have used an essentially arbitrary vague prior on θ, which makes the marginal likelihood essentially meaningless as the basis for model comparison. BIC or DIC can be used. We'd probably favour the latter if using stochastic simulation, since then the posterior mode is not directly accessible, and the latter otherwise.

3. This question is about modelling the global mean temperature series over the last 150 years, shown here.



Note that what is plotted is difference between average temperature in degrees centigrade and the average temperature in a reference period around 1950. The R session at the end of this question fits models to these data. Answers to the questions should be as concise as possible: as a rough guide answers should involve about one sentence per mark.

Solution comments. *This question is designed to be generally more challenging and to build somewhat in difficulty. The models and data set are unseen.*

(a) State the mathematical form of the model being fitted in part 1, and comment on its adequacy in the light of the plots shown. [6 marks]

Solution $T_i = \beta_0 + \beta_1 Y_i + \beta_2(Y_i - 1950)_+ + \epsilon_i$, where $(x)_+ = x$ if $x > 0$ and 0 otherwise. The ϵ_i are independent zero mean random variables, with constant variance σ^2 (also Gaussian for AIC). The model is not really adequate. There is strong residual pattern evident in both the middle plot and the ACF: the independence assumption on the ϵ_i is not valid with this model structure.

(b) Describe the purpose of the R function defined in part 2, including a mathematical statement of the model it implements for the temperature data, and a statement of what the function returns. [7 marks]

Solution A model $T_i = \beta_0 + \sum_{j=1}^4 \beta_j Y_i^j + \epsilon_i$ is used where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{V})$ and $\mathbf{V}_{ij} = \rho^{|i-j|}\sigma^2$ (Y_i is used in centred, linearly rescaled form, presumably to avoid co-linearity problems). The function evaluates the log likelihood for this model (using the multivariate normal p.d.f.), and returns twice the negative log likelihood (to facilitate use of a function minimizer and AIC computation).

(c) Give a concise description of the purpose of part 3 and comment on the adequacy of the model involved. [4 marks]

Solution `optim` is used to fit the model given above by maximum likelihood estimation. The estimate of ρ is consistent with the ACF of the residuals, and the estimate of σ consistent with the data and residual plots. The fitted values in the left plot look highly plausible as well, so $\hat{\boldsymbol{\beta}}$ seems reasonable.

(d) Use an appropriate statistical procedure to compare the two models fitted, indicating which is preferable. What would be wrong with using a generalized likelihood ratio test for this purpose? [4 marks]

Solution The AIC for the linear model was -188 while the AIC for the second model is easily calculated to be -290 + 12 = -278, so the second model is a substantial improvement according to AIC. The GLRT is invalid as the models are not nested.

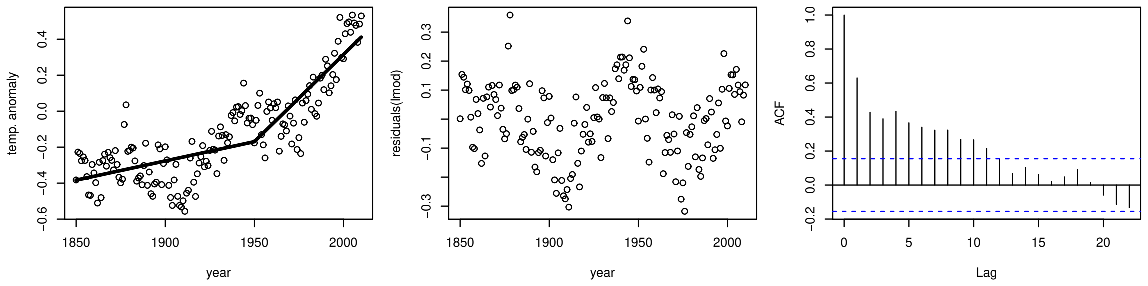
(e) The climate change scientists who gathered the data, prefer the model in part 1, because they have been using it for a long time. They would like confidence intervals for the model fit. Explain briefly why the intervals calculated by usual linear model methods are not appropriate here, and suggest an alternative for obtaining intervals for this model. [4 marks]

Solution The model assumptions are obviously not met, so intervals that rely on them being met must be suspect. We could bootstrap to obtain uncertainty estimates, i.e. repeatedly resample a data set of T_i, Y_i pairs with replacement from the original data, refitting the model to each re-sample, to build up a distribution of fitted values for the model, from which intervals can be computed. (An objection is that we repeatedly reproduce the same realization of the correlation in this way. Bonus mark for mentioning this)

Part 1

```
> dat$t50 <- dat$time - 1950
> dat$t50[dat$t50<0] <- 0
> lmod <- lm(temp~time+t50,data=dat)
> par(mfrow=c(1,3))
> plot(dat$time,dat$temp,xlab="year",ylab="temp. anomaly")
> lines(dat$time,fitted(lmod))
> plot(dat$time,residuals(lmod),xlab="year")
> acf(residuals(lmod))
```

Continued...



```
> AIC(lmod)
[1] -187.9727
```

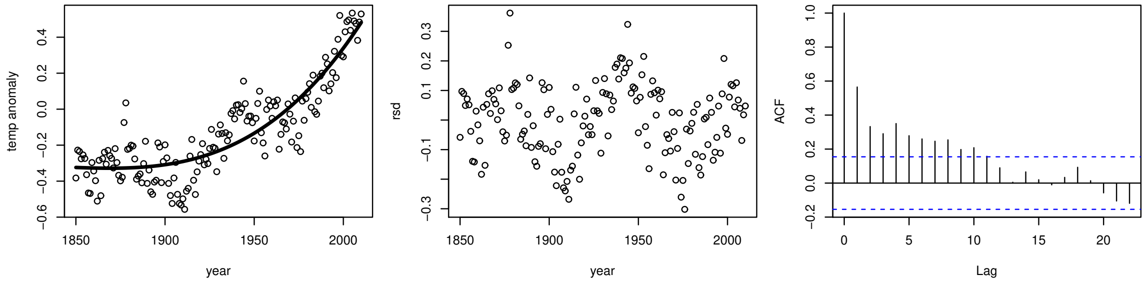
Part 2

```
ll <- function(theta,temp,time,return.mu = FALSE) {
  n <- length(temp)
  rho <- exp(theta[1])/(1+exp(theta[1]))
  sigma <- exp(theta[2])
  theta <- theta[-c(1,2)]
  ## following is efficient version of
  ## V <- matrix(0,n,n);
  ## for (i in 1:n) for (j in 1:n) V[i,j] <- rho^abs(i-j)*sigma^2
  V <- outer(time,time,function(x,y,rho) rho^abs(x-y),rho=rho)*sigma^2
  time <- (time - mean(time))/sd(time)
  mu.temp <- theta[1]
  for (i in 2:length(theta)) mu.temp <- mu.temp + theta[i]*time^(i-1)
  if (return.mu) return(mu.temp)
  t(temp-mu.temp)%*%solve(V,temp-mu.temp) +
     as.numeric(determinant(V,log=TRUE)$modulus) + n * log(2*pi)
}
```

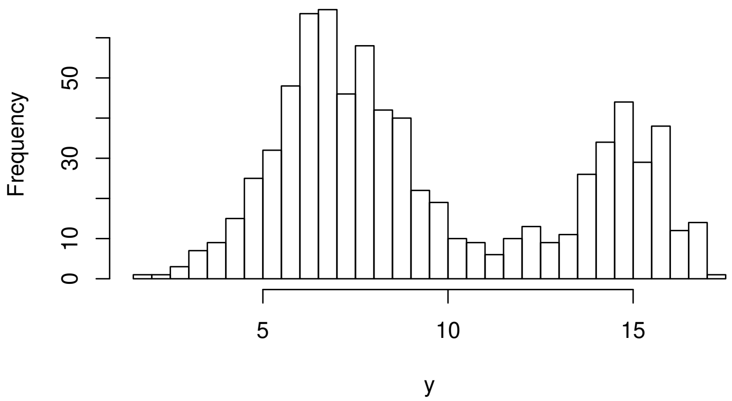
Continued...

Part 3

```
> theta0 <- c(1,-2,-.25,.25,.1,0)
> ll(theta0,dat$temp,dat$time)
 -283.6932
> fit <- optim(theta0,ll,method="BFGS",temp=dat$temp,time=dat$time)
> fit
$par
[1]  0.25390325 -2.11291358 -0.23162481  0.17957294  0.10542016  0.01872984

$value
[1] -290.1887

$counts
function gradient
      60       13

$convergence
[1] 0

> rho <- exp(fit$par[1])/(1+exp(fit$par[1]))
> sigma <- exp(fit$par[2])
> rho;sigma
[1] 0.563137
[1] 0.1208852
> mu.temp <- ll(fit$par,dat$termp,dat$time,return.mu = TRUE)
> rsd <- dat$temp - mu.temp
> par(mfrow=c(1,3),mar=c(4,4,1,1))
> plot(dat$time,dat$temp,xlab="year",ylab="temp anomaly")
> lines(dat$time,mu.temp,lwd=3)
> plot(dat$time,rsd,xlab="year")
> acf(rsd,main="")
```



4. A group of high energy physicists observe a set of particle energies in a series of experiments. Here is a histogram of the energies:



The two large peaks are explained by well established theory, but the experiments are testing a theory which predicts a small peak between the large peaks. There is a suggestion of an extra peak in the histogram, but it is not clear if it is real or just a chance occurrence. Since established theory provides quite accurate information on the location of the outer peaks, and the range of possibilities for the middle peak (if it exists) is also well defined by the proposed theory, it is decided to take a Bayesian approach to analysis. The JAGS and R code at the end of this question aims to perform the analysis. Answers should be as concise as possible: as a rough guide answers should involve about one sentence per mark.

Solution comments. *Again a more challenging question with unseen model and data. The final part on model comparison requires students to have really go to grips with this material.*

(a) Give a concise mathematical statement of the model for the observed energies, y_i, implemented in the JAGS code. You need not include the value of every parameter of the priors. [4 marks]

Solution $y_i \sim N(\mu_{k(i)}, \sigma^2_{k(i)})$, $k(i) \sim \text{dcat}(\mathbf{p})$, $\mu_k \sim N(m_k, s_k^2)$, $\sigma_k^2 \sim \text{gamma}(r_k, \lambda_k)$, $\mathbf{p} \sim \text{Dirichlet}(\boldsymbol{\alpha})$.

(b) Briefly explain the purpose of the R session labelled `## PART A`, and the meaning of the plot. [3 marks]

Solution The shape of the prior distributions for the precision of the peak widths is being plotted. The priors for the precisions of the two outer peaks are quite narrow, while the prior for the peak that is being searched for is much wider.

(c) Explain what is being done in `## PART B` of the session, and where appropriate why, including interpretation of the plots. [4 marks]

Solution The session is using JAGS to Gibbs sample from the model defined in the JAGS code. Trace plots are then used to check on how well the chains are mixing and how quickly they converge (only every 10th simulation has been stored). The precisions have been transformed to standard deviations for this purpose, to aid interpretation. Convergence appears rapid, well within the first 100 steps, and mixing looks reasonable. ACF plots are produced, discarding the first 100 observations as burn-in, these confirm the impression of rapid mixing. The values of the parameters all look sensible, given the initial histogram.

(d) Give the R code for computing a 95% credible interval for the standard deviation of the middle peak. [3 marks]

Solution

```
> quantile(sim$comp.tau[2,100:2000,1]^-.5,c(.025,0.975))
     2.5%     97.5%
0.6244703 0.8260642  ## not required, of course!!
```

(e) Explain the statistical purpose of `## PART C` and briefly interpret the plots. [3 marks]

Solution The code is evaluating the proportion of observations in each peak for each simulation in the chain, and examining trace plots of this quantity (post burn-in). The values seem consistent with the priors, with about 4.5% of observations attributed to the middle peak.

(f) In `## PART D` the original 3 peak model is compared to a model with the two outer component peaks, but no middle peak. Write out the JAGS model code for the 2 component model, as a modification of the original 3 component model. You need only show the modified code lines, and can put '· · ·' for any lines that are identical to the given code. [4 marks]

Solution

```
model {
 for (i in 1:N) {
  comp[i] ~ dcat(pc[1:2]) ## assign obs. to components
  ...
  ...
  ...
 }
 ## set up priors...
 p.mean <- c(7,15)
 sd.mean <- c(.1,.1)
 shape.tau <- c(34,246)
 rate.tau <- c(100,200)
 pc[1:3] ~ ddirich(c(.70,.3)) ## Dirichlet prior
 for (i in 1:2) {
   ...
   ...
 }
}
```

(g) Briefly explain how the alternative models are compared in `## PART D`, and which one you would select. How else might the models be compared in this case? Give reasons for favouring one or the other approaches (you can assume either can be computed). [4 marks]

Solution The two models are compared by DIC and the three component version is heavily favoured. In this case all the priors in the models are real well justified descriptions of prior uncertainty so comparing the models using Bayes factors would be fully justified.

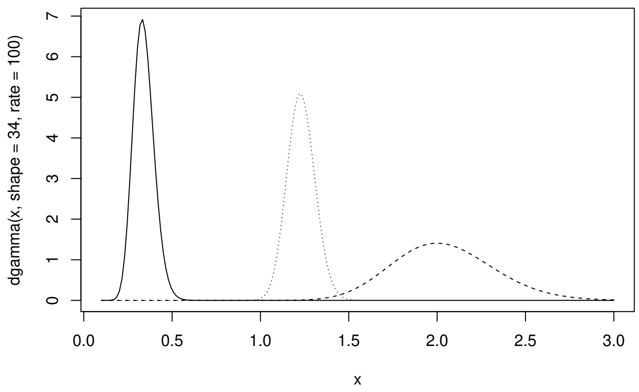
Continued...

The JAGS file (mix3.JAGS) implementing the model contains the following code. Note that the dcat distribution has a k dimensional parameter vector $\mathbf{p}$, where $\sum_{i=1}^{k} p_i = 1$, and describes a random variable that takes integer value $i \in [1, k]$ with probability p_i. The ddirich distribution is a suitable prior for $\mathbf{p}$: its parameter vector gives the prior expectation of $\mathbf{p}$.

```
model {
 for (i in 1:N) {
  comp[i] ~ dcat(pc[1:3]) ## assign obs. to components
  mu[i] <- comp.mu[comp[i]] ## component mean for ith obs
  tau[i] <- comp.tau[comp[i]] ## comp. precision for ith obs
  y[i] ~ dnorm(mu[i],tau[i]) ## obs density, given component
 }
 ## set up priors...
 p.mean <- c(7,12,15)
 sd.mean <- c(.1,.5,.1)
 shape.tau <- c(34,51,246)
 rate.tau <- c(100,25,200)
 pc[1:3] ~ ddirich(c(.68,.04,.28)) ## Dirichlet prior
 for (i in 1:3) {
  comp.tau[i] ~ dgamma(shape.tau[i],rate.tau[i])
  comp.mu[i] ~ dnorm(p.mean[i],1/sd.mean[i])
 }
}
```

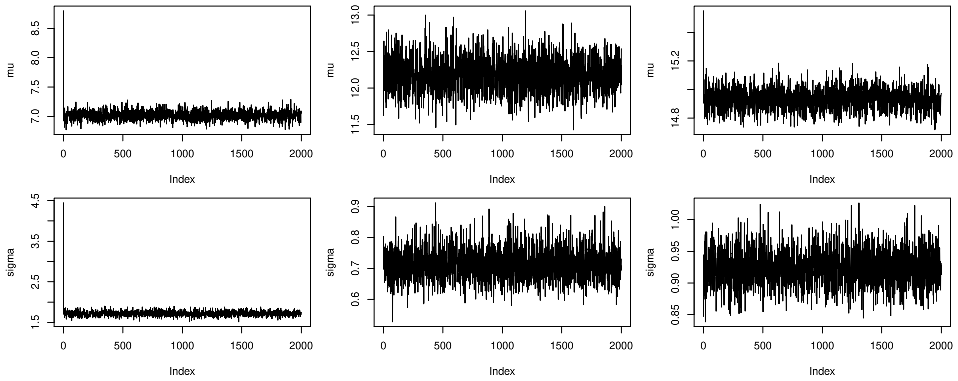
The R session using this model follows.

```
## PART A
> x <- seq(.1,3,length=200)
> plot(x,dgamma(x,shape=34,rate=100),type="l")
> lines(x,dgamma(x,shape=51,rate=25),lty=2)
> lines(x,dgamma(x,shape=246,rate=200),lty=3)
```

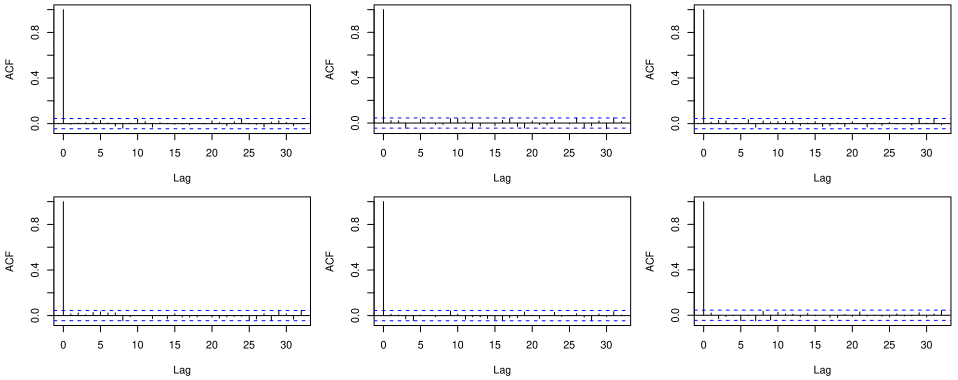


Continued...

```
> ## PART B
> library(rjags)
> n.sim <- 20000
> jam <- jags.model("mix3.JAGS",data=list(y=y,N=length(y)))
Compiling model graph
   Resolving undeclared variables
   Allocating nodes
Graph information:
   Observed stochastic nodes: 767
   Unobserved stochastic nodes: 774
   Total graph size: 3116

Initializing model

> sim <- jags.samples(jam,c("comp","comp.mu","comp.tau"),
+                      n.iter=n.sim,thin=10)
  |**************************************************| 100%
> par(mfrow=c(2,3))
> for (i in 1:3) plot(sim$comp.mu[i,,],type="l",ylab="mu")
> for (i in 1:3) plot(1/sim$comp.tau[i,,]^.5,type="l",ylab="sigma")
```

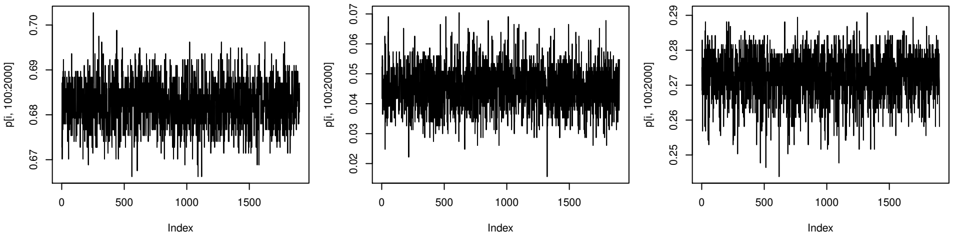


```
> par(mfrow=c(2,3))
> for (i in 1:3) acf(sim$comp.mu[i,100:2000,],main="mu")
> for (i in 1:3) acf(1/sim$comp.tau[i,100:2000,]^.5,main="sigma")
```



Continued...

```
> ## PART C
> p <- apply(sim$comp[,,],2,tabulate)/length(y)
> par(mfrow=c(1,3))
> for (i in 1:3) plot(p[i,100:2000],type="l")
```



```
> ## PART D
> jam <- jags.model("mix3.JAGS",data=list(y=y,N=length(y)),n.chains=2)
> dic.samples(jam,n.iter=10000)
  |**************************************************| 100%
Mean deviance:  2711
penalty 268.9
DIC: 2980
>
> jam0 <- jags.model("mix2.JAGS",data=list(y=y,N=length(y)),n.chains=2)
> dic.samples(jam0,n.iter=10000)
  |**************************************************| 100%
Mean deviance:  2926
penalty 219.6
DIC: 3146
```

End of examination.