8.4 Multiparameter Models

8.4.1 Multinomial distribution

Example 8.4.1 (Pre-election Polling)**.** *In late October 1988, a survey was conducted by CBS News of 1447 adults in the US to find out their preferences in the upcoming Presidential election. Out of 1447 persons,* $y_1 = 727$ *supported George Bush,* $y_2 = 583$ *supported Michael Dukakis, and* $y_3 = 137$ *supported other candidates or expressed no opinion. Assuming no other information on the respondents, the 1447 observations are exchangeable. An estimated of interest is* $\theta_1 - \theta_2$*, the population difference in support for the two major candidates.*

Let $\boldsymbol{y} = (y_1, y_2, y_3)$. Then $\boldsymbol{y}$ can be modeled as multinomial distribution

$$\boldsymbol{y} \sim MN(1447, \boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$. The conjugate prior of the multinomial distribution is the Dirichlet distribution $Dir(\boldsymbol{\alpha})$ defined as

$$f(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_i \theta_i^{\alpha_i - 1}.$$

Consider the non-informative prior where $\alpha_1 = \alpha_2 = \alpha_3 = 1$, then

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \sim Dir(728, 584, 138).$$

8.4.2 Normal model with a nuisance parameter

Suppose the parameter $\theta = (\theta_1, \theta_2)$ has two parts and we are only interested in inference for θ_1, so θ_2 is considered as a "nuisance" parameter. Given

the observation y, we seek the conditional distribution of the parameter of interest given the observed data $p(\theta_1|y)$. It is known that

$$f(\theta_1, \theta_2|y) \propto f(y|\theta_1, \theta_2) f(\theta_1, \theta_2).$$

By averaging over θ_2,

$$f(\theta_1|y) = \int p(\theta, \theta_2|y) d\theta_2.$$

Alternatively, the join posterior density can be factored to yield

$$f(\theta_1|y) = \int p(\theta_1|\theta_2, y) p(\theta_2|y) d\theta_2,$$

indicating that the posterior distribution $p(\theta_1|y)$ is a mixture of the conditional posterior distributions given the nuisance parameter.

Example 8.4.2 (Estimating the speed of light)**.** *Simon Newcomb set up an experiment in 1882 to measure the speed of light. Newcomb measured the amount of time required for light to travel a distance of 7442 meters.*

Assume that we have a random sample $Y_1, \cdots, Y_n$. Assume that $Y_i \sim N(\mu, \sigma^2)$.

Conjugate model

$$\begin{cases} Y_i \sim N(\mu, \sigma^2), \\ \mu \propto 1, \\ \sigma^2 \propto (\sigma^2)^{-1}. \end{cases}$$

Model checking: Is the model consistent with data? In Bayes statistics, one can use posterior predictive checking. If the model fits, then replicated data generated under the model should look similar to observed data. An observed discrepancy can be due to model misfit or chance.

Let y be the observed data and θ be the vector of parameters. Let y^{rep} be the replicated data that could have been "observed" assuming the same model mechanism. Then

$$f(y^{rep}|y) = \int f(y^{rep}|\theta)p(\theta|y)d\theta.$$

We measure the discrepancy between model and data by defining test quantities $T(y, \theta)$, the aspects of the data we wish to check. Unlike in the frequentist test statistic, $T(y, \theta)$ can depends on the parameter. For each replicated observation, one can compute the *posterior predictive p-values*, defined as

$$p_B = P(T(y^{rep}, \theta) \geq T(y, \theta)) = \int\int I_{T(y^{rep},\theta)\geq T(y,\theta)} p(y^{rep}|\theta)p(\theta|y)dy^{rep}d\theta.$$

Example 8.4.3 (Check the assumption of the speed light data)**.** *(a). Use the histogram;*

(b). Use the $min(y)$;

(c). Consider $T(y, \theta) = |y_{(61)} - \theta| - |y_{(6)} - \theta|$.

8.5 Hierarchical Models and MCMC

Many statistical applications involve multiple parameters that can be regarded as related or connected in some way by the structure of the problem, implying that a joint probability model for these parameters should reflect the dependence among them. For example, in a study of the effectiveness of cardiac treatments, with the patients in hospital j having survival probability θ_j, it might be reasonable to expect that estimates of the θ_j's, which represent a sample of hospitals, should be related to each other. This is achieved in a natural way if we use a prior distribution in which the θ_j's are viewed as a sample from a common population distribution.

It is natural to model such a problem hierarchically, with observable outcomes modeled conditionally on certain parameters, which themselves are given a probabilistic specification in terms of further parameters, known as *hyperparameters*.

In practice, the nonhierarchical models are usually inappropriate for hierarchical data: with few parameters, they generally cannot fit large datasets accurately, whereas with many parameters, they tend to "overfit" such data.

Example 8.5.1. Baseball example (Efron and Morris) *In 1977, Bradley Efron and Carl Morris published a paper about the shrinkage estimator that has better mean squared error than the simple average. Their prime example was the batting averages of 18 player in the 1970 season: they considered trying to estimate the players' average over the remainder of the season, based on their first 45 at-bats. Let Y_i be the batting average of the i-th player on their first 45 at-bats, and θ_i be the true batting average.*

Normal Hierarchical Model:

$$\begin{cases} Y_i|\theta_i, \sigma^2 \sim N(\theta_i, \sigma^2) \\ \theta_i|\mu, \tau^2 \sim N(\mu, \tau^2) \\ \mu \propto 1 (\sim N(0, 1000)) \\ (\sigma^2, \tau^2) \propto (\sigma^2)^{-1}(\tau^2)^{-1} (\sim Gamma(0.001, 0.001)) \end{cases}$$

θ_i's are the parameters of interest, σ^2, μ, τ^2 are not, they are called a nuisance parameters: parameters which are not of immediate interest but which must be accounted for in the analysis of those parameters which are of interest.

It is difficult (impossible) to calculate the posterior distribution $p(\theta_i|\boldsymbol{Y})$ explicitly. In hierarchical modeling, it is generally impossible to derive the posterior distribution explicitly. One must rely on the numerical methods

to calculate the posterior distribution. Markov chain Monte Carlo (MCMC) is a general method based on drawing values of θ from approximate distributions and then correcting those draws to better approximate the target posterior distribution, $p(\theta|y)$. The samples are drawn sequentially, with the distribution of the sampled draws depending on the last value drawn; hence, the draws form a Markov chain. Gibbs sampler is one of the most important method using MCMC.

Example 8.5.2. *Toy Example Consider a single observation* (y_1, y_2) *from a bivariate normally distributed population with unknown mean* $\theta = (\theta_1, \theta_2)$ *and known covariance matrix* $\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. *Assume that* $\pi(\theta_1)\pi(\theta_2) \propto 1$. *Consider the observation* $(y_1, y_2) = (-1, 5)$.

Gibbs Sampler

1. $\theta_1|\theta_2, y \sim N(y_1 + 0.5(\theta_2 - y_2), 1 - 0.5^2)$;
2. $\theta_2|\theta_1, y \sim N(y_2 + 0.5(\theta_1 - y_1), 1 - 0.5^2)$;

The basic method of inference from iterative simulation is the same as for Bayesian simulation in general: use the collection of all the simulated draws from $p(\theta|y)$ to summarize the posterior density and to compute quantiles, moments, and other summarizes of interest as needed.

Iterative simulation adds two difficulties to the problem of simulation inference. First, if the iterations have not proceeded long enough, the simulations may be grossly unrepresentative of the target distribution. The second problem is their within-sequence correlation. Simulation inference from correlated draws is generally les precise than from the same number of independent draws.

Methods:

1. To diminish the effect of the starting distribution, we generally discard the first k simulation, referred as *burn-in*.
2. We *thin* the sequences by keeping every k-th simulation draw from each sequence and discarding the rest.
3. Monitoring convergence based on multiple sequences with overdispersed starting points.

Gibbs sampler for the multinomial model

Gibbs sampler for the Baseball dataset

We now use BUGS[1] to draw samples from the posterior distribution. The BUGS project is concerned with flexible software for the Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) methods. The project began in 1989 in the MRC Biostatistics Unit, Cambridge, and led initially to the 'Classic' BUGS program, and then onto the WinBUGS software developed jointly with the Imperial College School of Medicine at St Mary's, London.

Baseball example:
BUGS model specification:

```
model{
for( j in 1:J){
y[j] ~ dnorm( theta[j], prec.y )
     theta[j] ~ dnorm( mu.theta, prec.theta )
}
       mu.theta ~ dnorm(0, 0.0001)
       prec.y ~ dgamma( 0.001, 0.001 )

       prec.theta ~ dgamma( 0.001, 0.001)
}

Data:
list(   J=18, y=c(0.400, 0.378, 0.356, 0.333, 0.311, 0.311, 0.289, 0.267,
0.244, 0.244, 0.222, 0.222, 0.222, 0.222, 0.222, 0.200, 0.178, 0.156 ) )

Inits:
list( theta=c(0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2,
0.2, 0.2, 0.2, 0.2, 0.2, 0.2), prec.y=1, prec.theta=1, mu.theta=0)
```

IN BUGS, the function *dnorm()* takes two arguments, the mean and the precision which is the inverse of the variance.

One can use WinBUGS to run the Gibbs sampler to obtain the posterior draw. Other commonly used software, such as R, Stata, SAS, provide API (application programming interface) function to use WinBUGS. The packages required is "R2OpenBUGS".

```
library(R2WinBUGS)
baseball <- read.table("EfronMorrisBB.txt", header=TRUE)
y <- baseball[,5]
J <- length(y)

data <- list("J", "y")
inits <- function(){
```

[1]Bayesian inference Using Gibbs Sampling

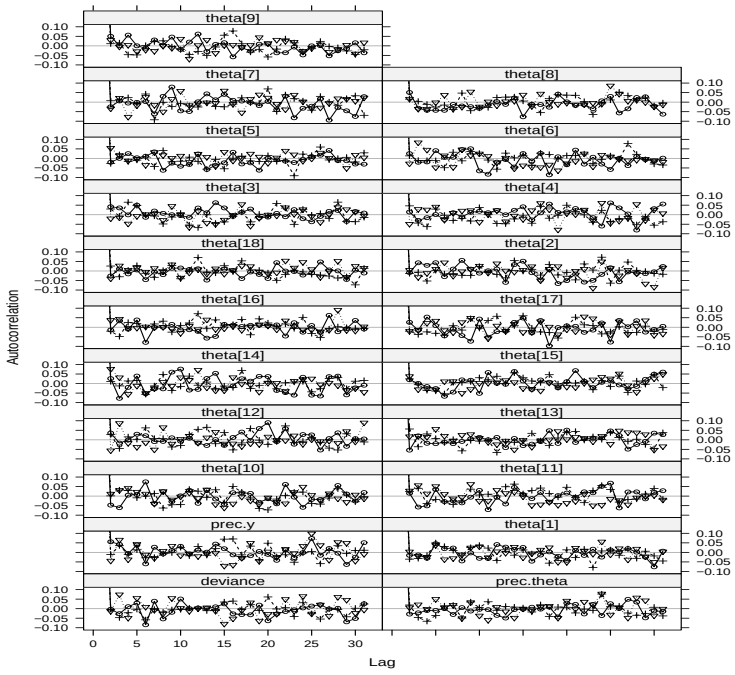

Figure 8.4: Auto correlation.

```
  list( theta= rnorm(18, 0.2, 0.1), sigma.y=rgamma(1,2,2),
        sigma.theta=rgamma(1,2,2), mu.theta=rnorm(1,0,1))
                    }
parameters <- c("theta", "prec.y", "sigma.y", "prec.theta", "sigma.theta")

baseball.gibbs <- bugs( data, inits, parameters, "baseball.txt", n.iter=50000,
                   n.burnin=10000, n.thin=40, n.chains=3 )
```

We plot the autocorrelation function as in Figure 8.4

In Figure 8.5, we plot the density function for each parameters.

We should perform the Gelman-Rubin convergence diagnostic with gelman.diag(). The shrink factors should be below 1.05.

```
> gelman.diag( mcmc.baseball )
Multivariate psrf
1.02
```

With MCMC convergence assured, we can retrieve the point estimates and 95% credible intervals of θ_i.

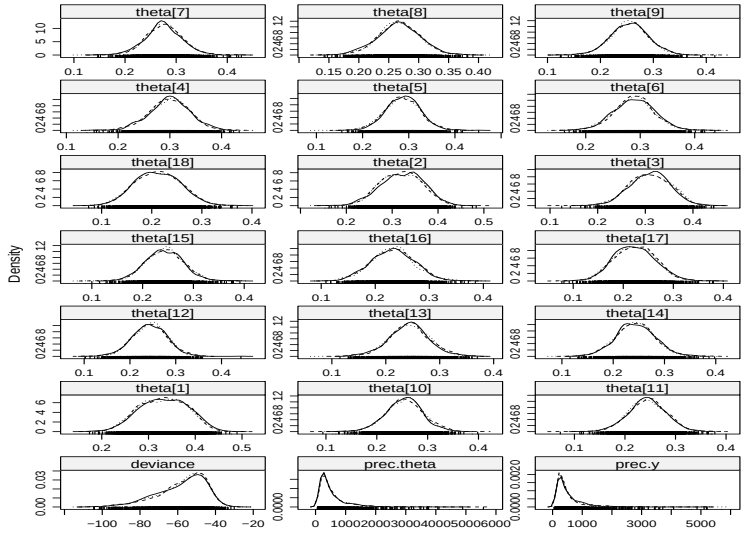

Figure 8.5: Auto correlation.

It is clearly seen that the Bayes estimator is better in terms of predicting the season batting average. We have plotted the Batting average (in the first 45 at.bats), the estimated batting average, and season batting average. There is a clearly shrinkage effect.