STA 602 - Intro to Bayesian Statistics

Lecture 15

Li Ma

Duke University

Two-sample comparison

- In many inference problems we are interested in comparing multiple samples of data to identify difference among them.
- The most typical problem is the two-sample problem that compares two-groups of observations
  - E.g., patients vs healthy, treatment vs control, etc.
- The most classical version of the two-sample problem focuses on comparing the mean of some measurement between the groups.
- The modern version of this problem is generalized to identifying a variety of differences in the underlying distributions (e.g., mean, variance, tail, local, . . .)
- We will look at the Bayesian approach to comparing the mean.

The two-sample problem

- Suppose there are two samples of data $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$ each can be considered i.i.d. from their respective sampling distribution

$$X_i \overset{\text{iid}}{\sim} F_1 \quad \text{and} \quad Y_j \overset{\text{iid}}{\sim} F_2.$$

- In the most simple version, we assume that F_1 and F_2 are both Gaussian with equal variance

$$F_1 = \mathrm{N}(\theta_1, \sigma^2) \quad \text{and} \quad F_2 = \mathrm{N}(\theta_2, \sigma^2).$$

- A generalization allows the variance to be different for the two groups σ_1^2 and σ_2^2.
- The interest is in the difference between the two means $\theta_1 - \theta_2$. For example, one might be interested in testing the null hypothesis

$$H_0 : \theta_1 - \theta_2 = 0 \quad \text{vs} \quad H_1 : \theta_1 - \theta_2 \neq 0.$$

The two-sample t-test

- The classical test for testing H_0 is the *t-test*.

$$t_{pool} = \frac{\bar{x} - \bar{y}}{s_{pool}\sqrt{1/n + 1/m}}$$

where s^2_{pool} is the sample variance estimate based on the pooled sample combining the two groups of observations:

$$s^2_{pool} = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2} = \frac{\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2}{n+m-2}.$$

- The pooled sample is meaningful only due to our assumption that the variance is equal for the two groups.
- t_{pool} has a t-distribution with $n+m-2$ degrees of freedom under H_0.
- Equivalent, a test for H_0 can be achieved by estimating $\theta_1 - \theta_2$ using $\bar{x} - \bar{y}$ and construct a (frequentist) confidence interval:

$$\left[\bar{x} - \bar{y} - t_{1-\alpha/2} \cdot s_{pool}\sqrt{1/n + 1/m}, \bar{x} - \bar{y} + t_{1-\alpha/2} \cdot s_{pool}\sqrt{1/n + 1/m}\right].$$

- A p-value can be computed under H_0.

Two-sample t-test with unequal variances

- If one suspects that the variance is not equal, this test cannot be used. A modified version when the variance is unequal is called Welch's t-test

$$t_{welch} = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n + s_y^2/m}},$$

which has approximately (not exactly) a t-distribution.

The Bayesian approach to two-sample comparison

- Let's now try to quantify our uncertainty about the underlying parameter of interest $\theta_1 - \theta_2$ using probability distribution.
- We could place priors on θ_1 and θ_2 and find the induced posterior on $\theta_1 - \theta_2$.
- What prior would be appropriate?

Prior specification

- How about a prior with independence

$$p(\theta_1, \theta_2) = p(\theta_1)p(\theta_2).$$

- Such a prior is usually unreasonable in two-sample problems because the fact that we are comparing the two samples in the first place is because we *a priori* conjectured that the two samples might be similar to each other.
- For example, if we are comparing the survival rate of cancer patients with or without a treatment at a hospital, or the SAT scores of two groups of students from two classes in the same high school.
- In other words, an appropriate prior usually should induce a positive correlation between θ_1 and θ_2.
- But incorporating prior belief about such dependence appears hard.

A helpful reparameterization

$\mu = \frac{\theta_1 + \theta_2}{2}$

$\delta = \frac{\theta_1 - \theta_2}{2}$

- Consider a reparameterization of the model

$$\theta_1 = \mu + \delta \quad \text{and} \quad \theta_2 = \mu - \delta.$$

- We replaced two parameters (θ_1, θ_2) with two new parameters (μ, δ).
- The sampling model is completely equivalent!
- But specifying a prior on (μ, δ) is easier now:
  - μ represents the overall average level.
  - δ represents difference between the two groups.
- It is now much more reasonable to assume prior independence between (μ, δ), if we are comfortable assuming that the level of difference doesn't depend on the overall level.
- Question: What if we want to assume that the difference is proportional to the overall level?
  - In that case, we may want to reparametrize as $\theta_1 = \mu(1+\delta)$ and $\theta_2 = \mu(1-\delta)$.

$X_{11}, X_{12}, \ldots, X_{1n} \sim N(\theta_1 = \mu + \delta, \sigma^2)$

$X_{21}, X_{22}, \ldots, X_{2n} \sim N(\theta_2 = \mu + \delta, \sigma^2)$

Prior specification

Decouple θ_1, θ_2 by specifying μ, δ

① Easy to specify priors

② Increase sampling efficiency

- So we adopt the following prior

$$p(\mu, \delta, \sigma^2) = p(\mu)p(\delta)p(\sigma^2)$$

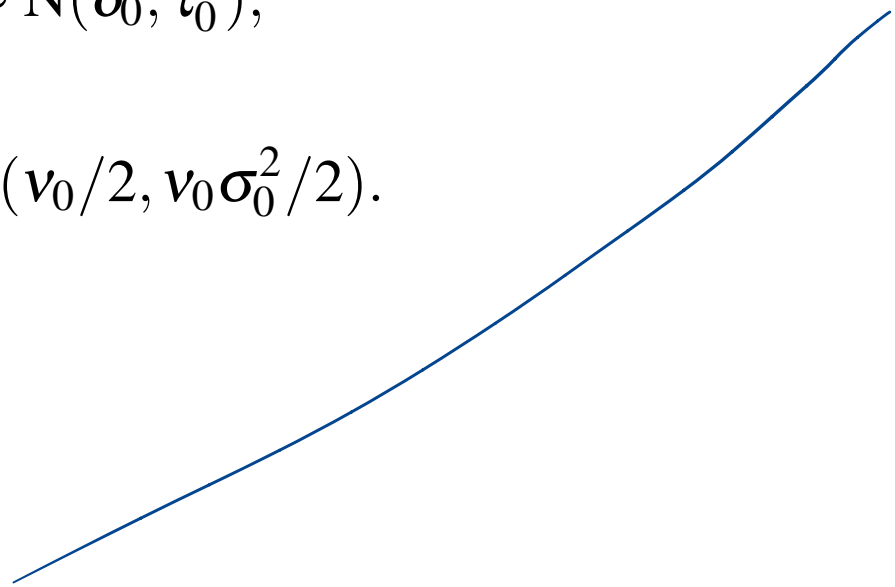
where

$$\mu \sim \mathrm{N}(\mu_0, \lambda_0^2),$$

$$\delta \sim \mathrm{N}(\delta_0, \tau_0^2),$$

and

$$\sigma^2 \sim \mathrm{IG}(\nu_0/2, \nu_0\sigma_0^2/2).$$



Bayesian inference on δ

Can be an exercise

- Now we can proceed as usual by first finding the posterior

$$p(\mu, \delta, \gamma | \mathbf{x}, \mathbf{y}) \propto p(\mathbf{x}, \mathbf{y} | \mu, \delta, \gamma) p(\mu, \delta, \gamma)$$

  where we let $\gamma = 1/\sigma^2$ for notational simplicity.
- The likelihood

$$\begin{aligned} p(\mathbf{x}, \mathbf{y} | \mu, \delta, \gamma) &\propto p(\mathbf{x} | \mu, \delta, \gamma) p(\mathbf{y} | \mu, \delta, \gamma) \\ &\propto \gamma^{n/2} e^{-\frac{\gamma}{2} \sum_i (x_i - \mu - \delta)^2} \cdot \gamma^{m/2} e^{-\frac{\gamma}{2} \sum_j (y_j - \mu + \delta)^2} \\ &\propto \gamma^{(n+m)/2} e^{-\frac{\gamma}{2} \left[\sum_i (x_i - \mu - \delta)^2 + \sum_j (y_j - \mu + \delta)^2 \right]}. \end{aligned}$$

The joint probability

- By Bayes theorem

$$
\begin{aligned}
& p(\mu, \delta, \gamma | \mathbf{x}, \mathbf{y}) \\
\propto & p(\mathbf{x}, \mathbf{y} | \mu, \delta, \gamma) p(\mu) p(\delta) p(\gamma) \\
\propto & \gamma^{\frac{n+m}{2}} e^{-\frac{\gamma}{2}\left[\sum_i (x_i - \mu - \delta)^2 + \sum_j (y_j - \mu + \delta)^2\right]} \cdot e^{-\frac{(\mu-\mu_0)^2}{2\lambda_0^2}} \cdot e^{-\frac{(\delta-\delta_0)^2}{2\tau_0^2}} \cdot \gamma^{\frac{\nu_0}{2}-1} e^{-\frac{\gamma}{2}\cdot\nu_0\sigma_0^2}
\end{aligned}
$$

The full conditional of σ^2

- The full conditional of $\gamma = 1/\sigma^2$ is

$$p(\gamma|\mu,\delta,\mathbf{x},\mathbf{y}) \propto \gamma^{\frac{\nu_0+n+m}{2}-1} e^{-\frac{\gamma}{2}\left[\nu_0\sigma_0^2+\sum_i(x_i-\mu-\delta)^2+\sum_j(y_j-\mu+\delta)^2\right]}.$$

That is,

$$\gamma|\,\mu,\delta,\mathbf{x},\mathbf{y} \sim \text{Gamma}\left(\frac{\nu_{n,m}}{2}, \frac{\nu_{n,m}\sigma_{n,m}^2}{2}\right)$$

or

$$\sigma^2|\,\mu,\delta,\mathbf{x},\mathbf{y} \sim \text{IG}\left(\frac{\nu_{n,m}}{2}, \frac{\nu_{n,m}\sigma_{n,m}^2}{2}\right).$$

where $\nu_{n,m} = \nu_0 + n + m$ and
$\nu_{n,m}\sigma_{n,m}^2 = \nu_0\sigma_0^2 + \sum_i(x_i-\mu-\delta)^2 + \sum_j(y_j-\mu+\delta)^2$.

If $\sigma_1^2 \neq \sigma_2^2$, consider $\sigma_1^2, \sigma_2^2 \sim \text{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$

The full conditional of μ

- The full conditional of μ is

$$p(\mu|\delta,\gamma,\mathbf{x},\mathbf{y}) \propto e^{-\frac{\gamma}{2}\left[\sum_i (x_i-\mu-\delta)^2+\sum_j (y_j-\mu+\delta)^2\right]} \cdot e^{-\frac{(\mu-\mu_0)^2}{2\lambda_0^2}}$$

$$\propto e^{-\frac{\gamma}{2}\left[\sum_i (\tilde{x}_i-\mu)^2+\sum_j (\tilde{y}_j-\mu)^2\right]} \cdot e^{-\frac{(\mu-\mu_0)^2}{2\lambda_0^2}}$$

  where $\tilde{x}_i = x_i - \delta$ and $\tilde{y}_j = y_j + \delta$.

- Now from our earlier results for a single Gaussian sample, we know the full conditional is given by

$$\mu \,|\, \delta,\gamma,\mathbf{x},\mathbf{y} \sim \mathrm{N}(\mu_{n,m}, \lambda_{n,m}^2)$$

  where

$$\mu_{n,m} = \lambda_{n,m}^2 \left(\frac{\sum_i \tilde{x}_i + \sum_j \tilde{y}_j}{n+m} \cdot \frac{n+m}{\sigma^2} + \frac{\mu_0}{\lambda_0^2} \right)$$

$$\frac{1}{\lambda_{n,m}^2} = (n+m)\gamma + \frac{1}{\lambda_0^2} = \frac{n+m}{\sigma^2} + \frac{1}{\lambda_0^2}.$$

The full conditional of δ

- Finally, the full conditional of δ is

$$p(\delta|\mu,\gamma,\mathbf{x},by) \propto e^{-\frac{\gamma}{2}\left[\sum_i (x_i-\mu-\delta)^2+\sum_j (y_j-\mu+\delta)^2\right]} \cdot e^{-\frac{(\delta-\delta_0)^2}{2\tau_0^2}}$$

$$\propto e^{-\frac{\gamma}{2}\left[\sum_i (\hat{x}_i-\delta)^2+\sum_j (\hat{y}_j-\delta)^2\right]} \cdot e^{-\frac{(\delta-\delta_0)^2}{2\tau_0^2}}$$

  where

$$\hat{x}_i = x_i - \mu \quad \text{and} \quad \hat{y}_j = \mu - y_j.$$

- Thus we can again draw from our earlier results for a single Gaussian sample,

$$\delta \,|\, \mu,\gamma,\mathbf{x},\mathbf{y} \sim \mathrm{N}(\delta_{n,m}, \tau^2_{n,m})$$

  where

$$\delta_{n,m} = \tau^2_{n,m}\left(\frac{\sum_i \hat{x}_i + \sum_j \hat{y}_j}{n+m} \cdot \frac{n+m}{\sigma^2} + \frac{\delta_0}{\tau_0^2}\right)$$

$$\frac{1}{\tau^2_{n,m}} = (n+m)\gamma + \frac{1}{\tau_0^2} = \frac{n+m}{\sigma^2} + \frac{1}{\tau_0^2}.$$

Gibbs sampling

- Initialize $(\mu^{(0)}, \delta^{(0)}, \sigma^{2\,(0)})$
- For $t = 1, 2, \ldots$
  - Update μ:
    - Compute $\lambda_{n,m}^{2\,(t)}$ and $\mu_{n,m}^{(t)}$.
    - Draw
    $$\mu^{(t)} \sim \mathrm{N}(\mu_{n,m}^{(t)}, \lambda_{n,m}^{2\,(t)}).$$
  - Update δ
    - Compute $\tau_{n,m}^{2\,(t)}$ and $\delta_{n,m}^{(t)}$.
    - Draw
    $$\delta^{(t)} \sim \mathrm{N}(\delta_{n,m}^{(t)}, \tau_{n,m}^{2\,(t)}).$$
  - Update σ^2
    - Compute $\nu_{n,m}\sigma_{n,m}^{2\,(t)}$.
    - Draw
    $$\sigma^2 \sim \mathrm{IG}(\nu_{n,m}/2, \nu_{n,m}\sigma_{n,m}^{2\,(t)}/2).$$

Bayesian hypothesis testing

$\delta^{(1)}, \delta^{(2)}, \ldots$

then $\hat{P}(\delta > 0 | x, y) = \frac{1}{S} \sum_{t=1}^{S} \mathbb{1}(\delta^{(s)} > 0)$

Both are wrong: ① Look at the tail prob; ② See whether 95% credible covers 0.

- It is tempting but *wrong* to "reject" H_0 at level α, say, if $\mathrm{P}(\delta > 0 | \mathbf{x}, \mathbf{y}) > 1 - \alpha$.
- A test like this will not provide nearly comparable inference under frequentest criteria (e.g., Type I error) compared to a frequentest test, e.g., a one-sided t-test, that rejects at level α.
- In one extreme, notice that $P(\delta = 0 | \mathbf{x}, \mathbf{y}) = 0$ always since δ is continuous!

Strategies:

* If $H_0: \delta \neq 0$, then consider $\hat{P}(|\delta - 0| \leq \Delta | x, y)$
* Spike and slab.

For example, here

* $H_0: d=0, \quad H_1: d=1.$

Then the 0-1 Loss function is

$$L(H_0; d) = \mathbb{1}(d=1)$$
$$L(H_1; d) = \mathbb{1}(d=0)$$

* For frequentist:

FP Type-I error $R(H_0, \delta) = \mathbb{E}[\mathbb{1}(\delta=1) \mid H_0] = \mathbb{P}(\delta=1 \mid H_0)$

FN 1-Power : $R(H_1, \delta) = 1 - \mathbb{E}[\mathbb{1}(\delta=1) \mid H_1] = \mathbb{P}(\delta=0 \mid H_1)$

Ideally you want risk to be low for both truths.

However it's impossible to minimize two types of error simultaneously.

Since in order to make less FP, you try not to reject so easily.

But in order to have high power, you want to reject more.

Neyman-Pearson Constrained optimization problem.

Minimize type-II error, while constraining type-I error $\leq \alpha$.

$$\arg\min_{\delta} R(H_1, \delta) \quad \text{s.t.} \quad R(H_0, \delta) \leq \alpha.$$

level alpha most powerful test.

* For Bayesian :

Set priors $p(H_0) = p_0$ and $p(H_1) = p_1$

Compute $p(H_0|\underline{x})$ and $p(H_1|\underline{x})$

and we have

$$\delta^*(x) = \arg\min_a \mathbb{E}[L(\theta, a)|\underline{x}] \qquad \theta \in \{0,1\}, \; a \in \{0,1\}$$

↑ Bayesian decision.

Predictive inference

One layer of uncertainty: $\mu, \delta, \sigma^2 \mid \underline{x}, \underline{y}$

The other layer of uncertainty: $(X_{n+1}, Y_{m+1} \mid \underline{x}, \underline{y})$

Where $X_{n+1} \mid \mu, \delta, \sigma^2 \sim N(\mu+\delta, \sigma^2)$

$Y_{m+1} \mid \mu, \delta, \sigma^2 \sim N(\mu-\delta, \sigma^2)$

These prob can be more meaningful

- Sometimes people are interested in predictive quantities such as

$$\begin{aligned} &\mathrm{P}(X_{n+1} - Y_{m+1} > 0 \,|\, \mathbf{x}, \mathbf{y}) \\ &= \int \mathrm{P}(X_{n+1} - Y_{m+1} > 0 | \mu, \delta, \sigma^2) p(\mu, \delta, \sigma^2 \,|\, \mathbf{x}, \mathbf{y}) d\mu d\delta d\sigma^2, \end{aligned}$$

  which can also be evaluated through MCMC with a Gibbs sampler.

$(X_{n+1} - Y_{m+1}) \mid \mu, \delta, \sigma^2 \sim N(2\delta, 2\sigma^2)$

Therefore, given Gibbs sampler $\delta^{(s)}, \sigma^{2(s)}$, draw $z_i \sim N(2\delta^{(s)}, 2\sigma^{2(s)})$

Then the integral is:

$$\frac{1}{S-B} \sum_{i=B+1}^{S} \mathbb{1}\{z^{(i)} > 0\}$$

Example: Air pollutant measurements

- Suppose we took 7 measurements in the morning and 9 measurements in the afternoon.

$$\mathbf{x} = (104, 105, 103, 102, 105, 107, 106)$$
$$\mathbf{y} = (104, 103, 106, 105, 102, 102, 108, 105, 104)$$

  and we are interested in the change in the pollution level from morning to afternoon.

- Suppose we are using the same device, and so σ^2 is assumed to be the same in the morning and afternoon.

Prior specification

- Based on historical data, *a priori* we think the average pollution level

$$\mu \sim \mathrm{N}(100, 25)$$

  That is, $\mu_0 = 100$ and $\lambda_0 = 5$.

- We think that the difference between morning and afternoon is typically close to 0, with standard deviation about 2,

$$\delta \sim \mathrm{N}(0, 4)$$

  That is, $\delta_0 = 0$ and $\tau_0 = 2$.

- We again adopt a weak prior on σ^2

$$\sigma^2 \sim \mathrm{IG}(\nu_0/2, \nu_0\sigma_0^2/2)$$

  where $\nu_0 = 1$ and $\sigma_0^2 = 4$.

Example: Air pollutant measurements

```
x <- c(104,105,103,102,105,107,106) # the data
y <- c(104,103,106,105,102,102,108,105,104)
n <- length(x) # sample size
m <- length(y)

# Prior specification
mu.0 <- 100; lambda2.0 <- 25;
delta.0 <- 0; tau2.0 <- 4;
nu.0 <- 1; sigma2.0 <- 4

# Initialization
niter <- 10000
nburnin <- 1000

xbar <- mean(x); ybar <- mean(y)
sx2 <- var(x); sy2 <- var(y)
s2.pool <- ((n-1)*sx2 + (m-1)*sy2)/(n+m-2)

mu.curr <- (xbar+ybar)/2
delta.curr <- (xbar-ybar)/2
sigma2.curr <- s2.pool

THETA <- matrix(NA,nrow=niter,ncol=3,dimnames=list(1:niter,c("mu","delta","sigma2")))
```

Start Gibbs sampling

```
for (t in 1:niter) {

  ## Update mu
  x.tilde <- x - delta.curr
  y.tilde <- y + delta.curr
  lambda2.n.m <- 1/((n+m)/sigma2.curr+1/lambda2.0)
  mu.n.m <- lambda2.n.m*(mean(c(x.tilde,y.tilde))*(n+m)/sigma2.curr + mu.0/lambda2.0)
  mu.curr <- rnorm(1,mean=mu.n.m,sd=sqrt(lambda2.n.m))

  ## Update delta
  x.hat <- x - mu.curr
  y.hat <- mu.curr - y
  tau2.n.m <- 1/((n+m)/sigma2.curr+1/tau2.0)
  delta.n.m <- tau2.n.m*(mean(c(x.hat,y.hat))*(n+m)/sigma2.curr + delta.0/tau2.0)
  delta.curr <- rnorm(1,mean=delta.n.m,sd=sqrt(tau2.n.m))

  ## Update sigma2
  sigma2.curr <-
    1/rgamma(1, shape=(nu.0+n+m)/2,
                rate=1/2*(nu.0*sigma2.0+sum((x-mu.curr-delta.curr)^2)+
                                        sum((y-mu.curr+delta.curr)^2)))

  ## Save the current iteration
  THETA[t,] <- c(mu.curr,delta.curr,sigma2.curr)
}
```

MCMC diagnostics

```
library(coda)
THETA.coda <- mcmc(THETA[-(1:nburnin),], start = 1+nburnin) # no burn-in steps
options(digits=3)
summary(THETA.coda)
```

```
##
## Iterations = 1001:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 9000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean    SD Naive SE Time-series SE
## mu     104.413 0.495  0.00522        0.00522
## delta    0.104 0.486  0.00512        0.00512
## sigma2   3.963 1.667  0.01758        0.01937
##
## 2. Quantiles for each variable:
##
##           2.5%     25%     50%     75%  97.5%
## mu     103.447 104.091 104.401 104.731 105.40
## delta   -0.854  -0.208   0.105   0.412   1.07
## sigma2   1.865   2.856   3.606   4.643   8.25
```

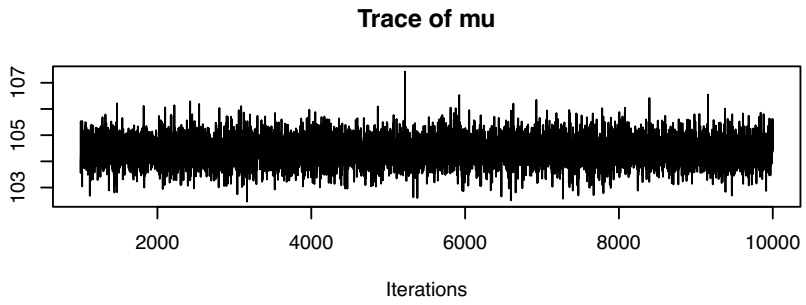
Autocorrelation and ESS

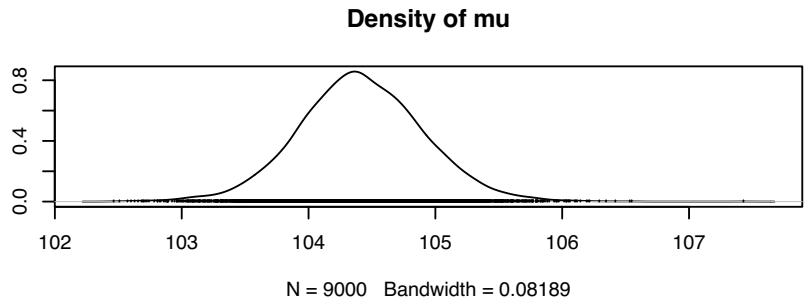
```
effectiveSize(THETA.coda)
```

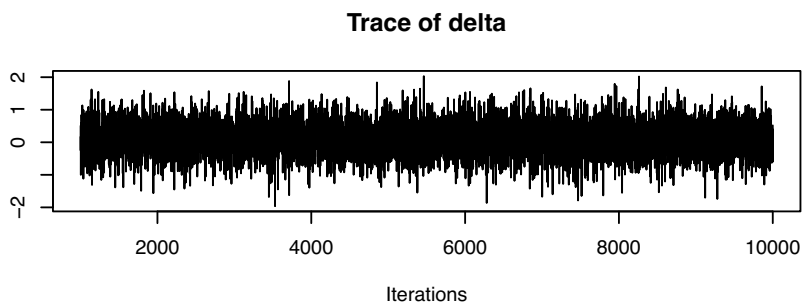
```
##     mu  delta sigma2
##   9000   9000   7411
```

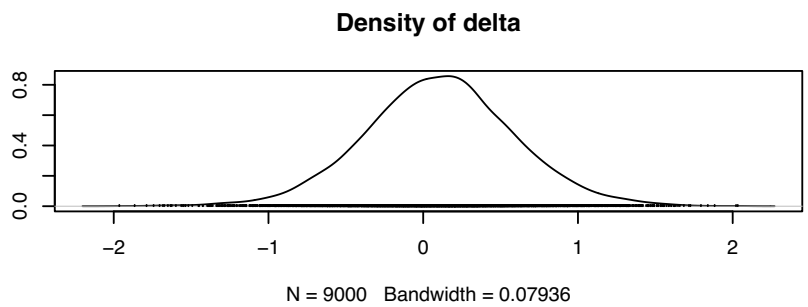
Trace plots

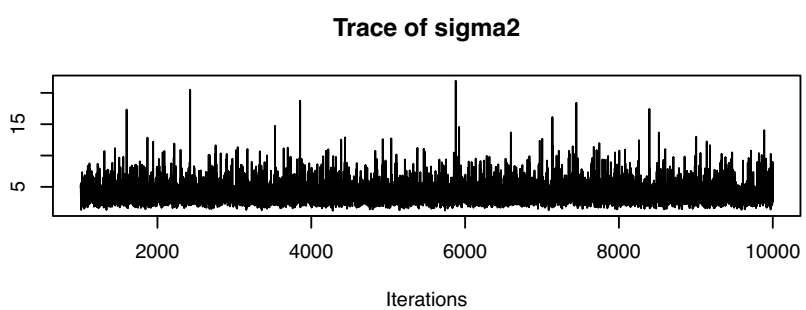
```
plot(THETA.coda)
```

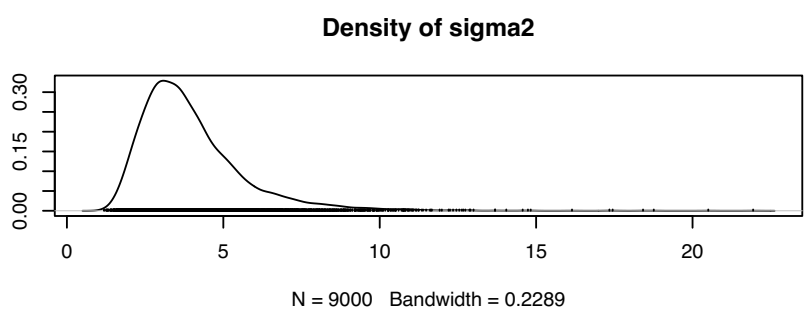












Autocorrelation plots

```
autocorr.plot(THETA.coda,lag.max=100)
```