STA 602 - Intro to Bayesian Statistics

Lecture 14

Li Ma

Duke University

Missing data

- Real data sets are almost never as "clean" as you would love them to be.
- One common issue is missingness, i.e., the value of some variables for some observations are not recorded for one reason or another.
- There are a number of different causes for missing values. Some examples include
  - Data lost by "accident" or unintended random errors: e.g., accidentally skipped a question on a survey, recording error during data compiling.
  - Data that are selectively recorded: e.g., an experimental measurement is very expensive and so one only carries it out on a subset of subjects, which are deemed to be more likely producing interesting results based on *other available measurements*.
  - Data that are not recorded due to its underlying value: e.g., a device that measures high temperatures might have a higher failure rate (producing missing values) at extremely high temperatures.
  - Many others possibilities …

Three types of missingness

- For a multivariate observation $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$, missing data scenarios can generally be characterized into the following types
  - Missing completely at random (MCAR): “the missingness has nothing to do we values of $\mathbf{X}$.”
  - Missing at random (MAR): “the missingness can depend on the observation $\mathbf{X}$ *only* through the observed part of $\mathbf{X}$.”
  - Missing not at random (MNAR): otherwise.

Examples from survey sampling

- 100 individuals are given a survey asking for their demographic information and a poll on their political interest such as whether they are supporters of a particular candidate.
- Suppose every one had completed their demographic questions, but some didn't respond to a few questions in the political section.
- MCAR: all respondants skipping questions on political information randomly, having nothing to do with either their political information or their demographic information.
- MAR: younger respondants skipped the political questions at a higer rate, and their age is recorded in the demographic section.
- MNAR: supporters for a politicular presendential candidate skipped more political questions than others.

(Trump supporters).

Mathematical formulation of missing data

- Suppose we observe p-dimensinoal random vectors on n samples

$$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$$

where $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ is the p variables for the ith sample.

- For example,

Subject ID	Age	Gender	Education	Political affiliation
1	36	F	Grad degree	Democrat
2	35	M	College	??
3	42	??	??	Republican
⋮	⋮	⋮	⋮	

- Let $\mathbf{X}$ be the entire data matrix. It can be represented as a matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \vdots & X_{np} \end{pmatrix}.$$

with possible NA's in some elements.

An alternative representation

- For Sample i, had there not been any missing values, we would observe the whole p-vector $\mathbf{X}_i$.
- In reality, we observe a subvector $\mathbf{X}_{i,obs}$ while the other elements $\mathbf{X}_{i,mis}$ are missing:

$$\mathbf{X}_i = (\mathbf{X}_{i,obs}, \mathbf{X}_{i,mis}).$$

- *Note that for differen i, the missing dimensions will be different, and so please don't confuse the notation as the observed values always proceed the missing ones.*
- Note that $\mathbf{X}_{i,mis}$ represents the values that *could have been observed* but are missing. So their values are actually not in our data. For example, the actually education level and gender of Sample 3 in the above.
- We can also write all the observed data collectively as $\mathbf{X}_{obs}$ and the missing data $\mathbf{X}_{mis}$.
- What is our data?

An alternative representation

- For each sample, not only do we observe the non-missing data $\mathbf{X}_{i,obs}$, we also observe whether each of its variable is missing or not.
- In other words, we observe an indicator vector for each sample

$$\boldsymbol{I}_i = (I_{i1}, I_{i2}, \ldots, I_{ip})'$$

where $I_{ij} \in \{0, 1\}$ such that I_{ij} indicates the jth variable is observed for Sample i.

- Collectively, let $\boldsymbol{I} = (\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_n)$ be the indicators for all observations.
- So an actual dataset can be rewritten in the form of

$$(\mathbf{X}_{obs}, \boldsymbol{I}).$$

← think this as a list

- Note that we don't get to observe the values of $\mathbf{X}_{mis}$ in the *complete* data

$$(\mathbf{X}_{obs}, \mathbf{X}_{mis}, \boldsymbol{I}).$$

Inference strategy

- Bayesian inference carries forward as usual.
- Some specific details: The sampling model should include both the model for $\mathbf{X}$ and that for $\boldsymbol{I}$, the missingness.
- Let $(\boldsymbol{\theta}, \boldsymbol{\psi})$ be the parameters for the sampling model, where $\boldsymbol{\theta}$ is for $\mathbf{X}$, and $\boldsymbol{\psi}$ are parameters for $\boldsymbol{I}$ given $\mathbf{X}$, that is we specify

sampling model.

$$p(\underline{\mathbf{x}}|\boldsymbol{\theta}) \quad \text{and} \quad p(\boldsymbol{I}|\mathbf{x}, \boldsymbol{\psi}).$$

complete data

MCAR : $p(I|\psi)$
MAR : $p(I|X_{obs}, \psi)$
MNAR: $p(I|X, \psi)$

- We can specify a prior $p(\boldsymbol{\theta}, \boldsymbol{\psi})$, note that overlap between $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is a special case of prior specification.

Inference strategy

- The likelihood for our data $(\mathbf{x}_{obs}, \boldsymbol{I})$ is

$$p(\mathbf{x}_{obs}, \boldsymbol{I}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \int p(\mathbf{x}_{obs}, \mathbf{x}_{mis}, \boldsymbol{I}|\boldsymbol{\theta}, \boldsymbol{\psi}) d\mathbf{x}_{mis}.$$

  This often does not have a simple closed form, not even known up to a normalizing constant.
- So applying Bayes theorem

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{x}_{obs}, \boldsymbol{I}) \propto p(\mathbf{x}_{obs}, \boldsymbol{I}|\boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{\theta}, \boldsymbol{\psi})$$

$$\propto \left[\int p(x_{obs}, x_{mis}, I|\theta, \psi)\, d x_{mis} \right] \cdot p(\theta, \psi)$$

  directly is not easy!

An important observation

- In many problems, since we usually directly specify a sampling model for the "complete data", the likelihood, *had there been no missing data at all,*

$$p(\mathbf{x}|\boldsymbol{\theta})$$

  often does have simple analytic forms.

- The difficulty arises from the inability to observe $\mathbf{x}_{mis}$.

Key idea

- There is nothing preventing us from treating $\mathbf{x}_{mis}$ just like the other unobserved quantities, e.g., $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, and sample from the posterior joint distribution of $(\mathbf{x}_{mis}, \boldsymbol{\theta}, \boldsymbol{\psi})$:

$$p(\mathbf{x}_{mis}, \boldsymbol{\theta}, \boldsymbol{\psi} | \mathbf{x}_{obs}, \boldsymbol{I}).$$

- Question: Do we need to specify a prior distribution for $\mathbf{X}_{mis}$?
  - No! The sampling model for the complete data already does the job! We have already treated X_{mis} as a random variable by specifying $p(X_{obs}, X_{mis}, I | \theta, \psi)$
  - We are already able to write down the joint probability.
- The joint probability is

$$\begin{aligned} p(\mathbf{x}_{obs}, \mathbf{x}_{mis}, \boldsymbol{I}, \boldsymbol{\theta}, \boldsymbol{\psi}) &= \overbrace{p(\mathbf{x}_{obs}, \mathbf{x}_{mis}, \boldsymbol{I} | \boldsymbol{\theta}, \boldsymbol{\psi})}^{\text{sampling model}} p(\boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= p(\mathbf{x}_{obs}, \mathbf{x}_{mis} | \boldsymbol{\theta}) p(\boldsymbol{I} | \mathbf{x}_{obs}, \mathbf{x}_{mis}, \boldsymbol{\psi}) p(\boldsymbol{\theta}, \boldsymbol{\psi}). \end{aligned}$$

A conceptual justification

Data augmentation, Latent variable:
aug → sample → discard

- The essence of this strategy is to do the integration over $\mathbf{x}_{mis}$ *after* applying Bayes theorem, rather than *before*.

$$p(\mathbf{x}_{mis}, \boldsymbol{\theta}, \boldsymbol{\psi} | \mathbf{x}_{obs}, \boldsymbol{I}) \propto p(\mathbf{x}_{obs}, \mathbf{x}_{mis} \,|\, \boldsymbol{\theta}) p(I | \mathbf{x}_{obs}, \mathbf{x}_{mis}, \boldsymbol{\psi}) p(\boldsymbol{\theta}, \boldsymbol{\psi}).$$

- After sampling from this posterior

$$(\mathbf{x}_{mis}^{(1)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\psi}^{(1)}), (\mathbf{x}_{mis}^{(2)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\psi}^{(2)}), \ldots$$

  Then discarding the sampled values $\mathbf{x}_{mis}^{(t)}$ corresponds to integrating out $\mathbf{x}_{mis}$ in the posterior

$$p(\boldsymbol{\theta}, \boldsymbol{\psi} | \mathbf{x}_{obs}, \boldsymbol{I}) = \int p(\mathbf{x}_{mis}, \boldsymbol{\theta}, \boldsymbol{\psi} | \mathbf{x}_{obs}, \boldsymbol{I}) d\mathbf{x}_{mis}.$$

- This avoids the difficulty we had previously in sampling from the marginal posterior of $(\boldsymbol{\theta}, \boldsymbol{\psi})$ directly.

Bayesian imputation

- Inspired by Gibbs sampling, if we can iteratively sample $\mathbf{x}_{mis}$, $\boldsymbol{\theta}$, and $\boldsymbol{\psi}$ from their full conditionals. Then the Markov Chain
$$(\mathbf{x}_{mis}^{(1)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\psi}^{(1)}), (\mathbf{x}_{mis}^{(2)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\psi}^{(2)}), \ldots$$
will eventually converge to the target distribution
$$p(\mathbf{x}_{mis}, \boldsymbol{\theta}, \boldsymbol{\psi} | \mathbf{x}_{obs}, \boldsymbol{I}).$$
- So the key is to derive the full conditional of $\mathbf{x}_{mis}$ given $\mathbf{x}_{obs}$, $\boldsymbol{I}$, $\boldsymbol{\theta}$, and $\boldsymbol{\psi}$.
- The step in the sampler that draws from the full conditional of $\mathbf{x}_{mis}$ given others is called *imputation*.
- Note the difference from the naive strategy of imputing the missing values once and for all, which ignores the uncertainty in the missing values.
- A frequentist counterpart is to repeatedly impute $\mathbf{x}_{mis}$, then maximize over the parameter values $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. (E.g., EM algorithm and others.)

Finding the joint probability

- The full joint probability of all random quantities is given by
  $p(X_{mis}, \theta, \psi \mid X_{obs}, I) \propto$
  $$p(\mathbf{x}_{obs}, \mathbf{x}_{mis}, \boldsymbol{I}, \boldsymbol{\theta}, \boldsymbol{\psi}) = p(\mathbf{x}_{obs}, \mathbf{x}_{mis}|\boldsymbol{\theta})p(\boldsymbol{I}|\mathbf{x}_{obs}, \mathbf{x}_{mis}, \boldsymbol{\psi})p(\boldsymbol{\theta}, \boldsymbol{\psi})$$
  based on which we can try to find the full conditionals of $\mathbf{x}_{mis}$ along with those for $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$.
- In case the full conditional is complicated, there are more advanced MCMC sampling strategies to sample from them.
- This is the general strategy for dealing with data that are missing not at random (MNAR), which is the case when the missingness model has the full form
  $$p(\boldsymbol{I}|\mathbf{x}_{obs}, \mathbf{x}_{mis}, \boldsymbol{\psi}).$$
- The missingness matrix $\boldsymbol{I}$ influences the full conditional for $\mathbf{x}_{mis}$ and thus cannot be ignored in dealing with missing data.
- Things are a bit nicer with MCAR and MAR.

Missing completely at random (MCAR)

- If, however, one is willing to assume MCAR, then

$$p(\boldsymbol{I}|\mathbf{x}_{obs},\mathbf{x}_{mis},\boldsymbol{\psi}) = p(\boldsymbol{I}|\boldsymbol{\psi}).$$

  In this case, the right hand side of the full joint probability becomes

$$p(\mathbf{x}_{obs},\mathbf{x}_{mis}|\boldsymbol{\theta})p(\boldsymbol{I}|\boldsymbol{\psi})p(\boldsymbol{\theta},\boldsymbol{\psi})$$

  In particular, if we are just interested in $\boldsymbol{\theta}$ alone, and willing to place *independent* priors

$$p(\boldsymbol{\theta},\boldsymbol{\psi}) = p(\boldsymbol{\theta})p(\boldsymbol{\psi})$$

  then the part of the joint probability that involves $\mathbf{x}_{mis}$ and $\boldsymbol{\theta}$ is

$$\boxed{p(\mathbf{x}_{obs},\mathbf{x}_{mis}|\boldsymbol{\theta})p(\boldsymbol{\theta}).}$$

- Thus we can safely *ignore* the likelihood contribution from the missingness $\boldsymbol{I}$ whatsoever in finding the full conditionals of $\mathbf{x}_{mis}$ and $\boldsymbol{\theta}$.

Missing at random (MAR)

- If one thinks MCAR is too strong an assumption to make, but willing to assume MAR, then

  $$p(\boldsymbol{I}|\mathbf{x}_{obs}, \mathbf{x}_{mis}, \boldsymbol{\psi}) = p(\boldsymbol{I}|\mathbf{x}_{obs}, \boldsymbol{\psi})$$

  then the right-hand side becomes

  $$p(\mathbf{x}_{obs}, \mathbf{x}_{mis}|\boldsymbol{\theta})p(\boldsymbol{I}|\mathbf{x}_{obs}, \boldsymbol{\psi})p(\boldsymbol{\theta})p(\boldsymbol{\psi}).$$

  In this case, the part that involves $\boldsymbol{\theta}$ and $\mathbf{x}_{mis}$ is

  $$\boxed{p(\mathbf{x}_{obs}, \mathbf{x}_{mis}|\boldsymbol{\theta})p(\boldsymbol{\theta})}$$

  which again does not depend on $\boldsymbol{I}$. So the full conditionals for $\mathbf{x}_{mis}$ and $\boldsymbol{\theta}$ also will *not* depend on $\boldsymbol{I}$!
- Therefore MCAR and MAR are called *ignorable* missingness.

Example: Multivariate normal model

- Suppose each $\mathbf{y}_i \in \mathbb{R}^p$ and we model them as multivariate normal

$$\mathbf{y}_i \overset{\text{iid}}{\sim} \mathrm{N}(\boldsymbol{\theta}, \Sigma).$$

sampling model for complete data

- Now for each observation we may have some missing data, and so we can write

$$\mathbf{y}_i = (\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis}).$$

- Assume that the missingness is ignorable, i.e., MCAR or MAR. As such, we can ignore the missingness indicator $\boldsymbol{I}_i$.
- Suppose we adopt the same independent prior for $\boldsymbol{\theta}$ and Σ as before

$$\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\mu}_0, \Lambda_0) \quad \text{and} \quad \Sigma \sim \mathrm{IW}(\nu_0, \boldsymbol{S}_0).$$

- Let's try to find the full conditional for $(\mathbf{y}_{mis}, \boldsymbol{\theta}, \Sigma)$.

The full conditionals

- The full conditionals for $\boldsymbol{\theta}$ and Σ given $\mathbf{y} = (\mathbf{y}_{obs}, \mathbf{y}_{mis})$ are exactly those when there are no missing data.
- For $\mathbf{y}_{mis}$, we have

$$
\begin{aligned}
p(\mathbf{y}_{mis} | \mathbf{y}_{obs}, \boldsymbol{\theta}, \Sigma) &= \frac{p(\mathbf{y}_{obs}, \mathbf{y}_{mis} | \boldsymbol{\theta}, \Sigma)}{p(\mathbf{y}_{obs} | \boldsymbol{\theta}, \Sigma)} \\
&= \frac{\prod_i p(\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis} | \boldsymbol{\theta}, \Sigma)}{\prod_i p(\mathbf{y}_{i,obs} | \boldsymbol{\theta}, \Sigma)} \\
&= \prod_i p(\underbrace{\mathbf{y}_{i,mis} | \mathbf{y}_{i,obs}, \boldsymbol{\theta}}_{\text{the form for } (\theta_1, \theta_2)^T \text{ is different for each } i.}, \Sigma).
\end{aligned}
$$

- For multivariate normal random vector

$$(\mathbf{y}_{obs}, \mathbf{y}_{mis}) \mid \boldsymbol{\theta}, \Sigma \sim \mathrm{N}(\boldsymbol{\theta}, \Sigma)$$

where

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} \begin{matrix} obs \\ mis \end{matrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

we have (*)

$$\mathbf{y}_{mis} | \mathbf{y}_{obs}, \boldsymbol{\theta}, \Sigma \sim \mathrm{N}\left(\boldsymbol{\theta}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{y}_{obs} - \boldsymbol{\theta}_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right).$$

for each i is different.

We have to deal with the imputation one obs. at a time.

Gibbs sampling

$\theta \sim N(u_0, \Lambda_0)$

$\Sigma \sim IW(\nu_0, S_0)$

- Initialize $(\mathbf{y}_{mis}^{(0)}, \boldsymbol{\theta}^{(0)}, \Sigma^{(0)})$.
- For $t = 1, 2, \ldots$
  - Compute using the complete data $\mathbf{y} = (\mathbf{y}_{obs}, \mathbf{y}_{mis}^{(t-1)})$,

$$\left(\Lambda_n^{(t)}\right)^{-1} = n\left(\Sigma^{(t-1)}\right)^{-1} + \Lambda_0^{-1},$$

$$\boldsymbol{\mu}_n^{(t)} = \Lambda_n^{(t)}\left(n\left(\Sigma^{(t-1)}\right)^{-1}\bar{y} + \Lambda_0^{-1}\boldsymbol{\mu}_0\right)$$

  - Draw

$$\boldsymbol{\theta}^{(t)} \sim \mathrm{N}(\boldsymbol{\mu}_n^{(t)}, \Lambda_n^{(t)})$$

  - Draw

$$\Sigma^{(t)} \sim \mathrm{IW}\left(\nu_n, \boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\theta}^{(t)}}\right).$$

  - Draw for all $i = 1, 2, \ldots, n$.

$$\mathbf{y}_{i,mis}^{(t)} \sim \mathrm{N}\left(\boldsymbol{\theta}_2^{(t)} + \Sigma_{21}^{(t)}\left(\Sigma_{11}^{(t)}\right)^{-1}(\mathbf{y}_{i,obs} - \boldsymbol{\theta}_1^{(t)}), \Sigma_{22}^{(t)} - \Sigma_{21}^{(t)}\left(\Sigma_{11}^{(t)}\right)^{-1}\Sigma_{12}^{(t)}\right)$$

- Discard burn-ins.

Example: Air pollutant measurements

- Suppose in measuring two pollutants (e.g., PM2.5 and SO2), 16 times on a day, out of which, in 6 times, we measured only one of the pollutant.

  $(104, 100), (105, ??), (103, 101), (102, 104), (105, 108), (107, 108), (??, 103), (104, 104), (??, 106), (106, 107), (105, 105), (102, ??), (102, ??), (??, 106), (105, 105), (104, 105)$

- Also, we know that those missingness is either MCAR or MAR.
- Can you think of some plausible scenarios for MCAR, MAR, and MNAR?

Reading the data

```r
library(mvtnorm) # for drawing multivariate normal
library(MCMCpack) # for drawing inverse-Wishart
library(coda) # for MCMC diagnostics
options(digits=3) # tidy up print outs

y.original <- matrix(c(104,100,105,NA,103,101,102,104,105,108,107,108,NA,103,104,104,
              NA,106,106,107,105,105,102,NA,102,NA,NA,106,105,105,104,105),
             ncol=2,byrow=TRUE)
n <- nrow(y.original) # sample size
p <- ncol(y.original) # dimensionality

I <- !is.na(y.original) # missingness indicator, TRUE if present, 0 if missing
```

prior specification

```r
# prior for theta
mu.0 <- c(100,100)
Lambda.0 = matrix(c(100,15,15,25),ncol=2,byrow=TRUE)

# prior for Sigma
nu.0 <- p + 2  # a very weak prior
S0 <- matrix(c(4,0,0,4),ncol=2,byrow=TRUE)
```

Initialization

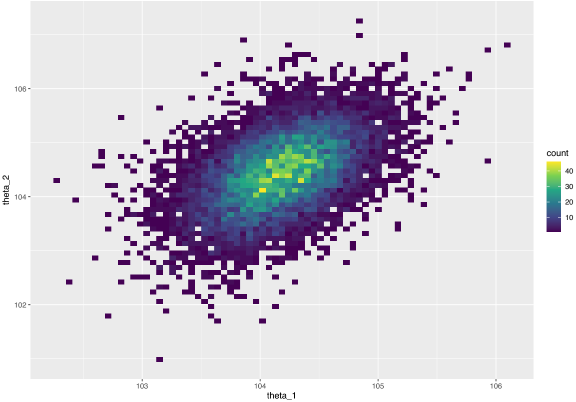
```r
niter <- 10000 # total number of iterations
nburnin <- 1000 # 1000 burn-in steps

ybar.original <- apply(y.original,2,mean,na.rm=TRUE) # the column means of the original data

y <- y.original ## y holds the imputed data (y.obs,y.mis)
# initialize y by filling in the NAs with the corresponding column means
for (i in 1:p) {
  y[I[,i]==0,i] <- ybar.original[i]
}

## Proceed as before like there are no missing data
ybar <- apply(y,2,mean)
nu.n <- nu.0 + n

THETA <- matrix(NA,nrow=niter,ncol=p) # matrix for storing the draws for theta
colnames(THETA) <- c("theta1","theta2")

THETA.init <- ybar # Initial values set to sample mean
THETA.curr <- THETA.init # the theta value at current iteration

SIGMA <- matrix(NA,nrow=niter,ncol=p*p) # matrix for storing the draws for Sigma
colnames(SIGMA) <- c("sigma11","sigma12","sigma21","sigma22")

SIGMA.init <- cov(y) # intial value set to sample covariance
SIGMA.curr <- SIGMA.init # the Sigma value at current iternation
```

Gibbs sampling

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

  Lambda.n <- solve(n*solve(SIGMA.curr)+solve(Lambda.0))
  mu.n <- Lambda.n %*% (n*solve(SIGMA.curr,ybar)+solve(Lambda.0,mu.0))

  ## Update theta
  THETA.curr <- rmvnorm(1,mean=mu.n,sigma=Lambda.n)

  ## Update Sigma
  S.theta <- (t(y)-c(THETA.curr))%*% t(t(y)-c(THETA.curr))
  SIGMA.curr <- riwish(v=nu.n,S=S0+S.theta)

  ## Impute the missing data
  for (i in 1:n) {
      var.obs = which(I[i,]) ## which variables are observed
      var.mis = which(!I[i,]) ## which variables are missing

      if (length(var.mis) > 0){ ## if there are missing values
        SIGMA.obs <- SIGMA.curr[var.obs,var.obs] # Sigma11
        SIGMA.mis <- SIGMA.curr[var.mis,var.mis] # Sigma22
        SIGMA.mis.obs <- SIGMA.curr[var.mis,var.obs] # Sigma21
        SIGMA.obs.mis <- t(SIGMA.mis.obs) # Sigma12
        y[i,var.mis] <- rnorm(1, mean=THETA.curr[var.mis]+
                          SIGMA.mis.obs%*%solve(SIGMA.obs,y[i,var.obs]-THETA.curr[var.obs]),
                          sd=sqrt(SIGMA.mis-SIGMA.mis.obs%*%solve(SIGMA.obs,SIGMA.obs.mis)))
      }
  }
  ybar <- apply(y,2,mean)

  ## Save the current iteration
  THETA[t,] <- THETA.curr
  SIGMA[t,] <- SIGMA.curr
}
```

Histogram of MCMC draws for $\boldsymbol{\theta}$

```r
ggplot(data.frame(THETA), aes(x=theta1, y=theta2) ) +
  labs(x=expression(theta_1),y=expression(theta_2)) +
  geom_bin2d(bins=70) +
  scale_fill_continuous(type = "viridis")
```



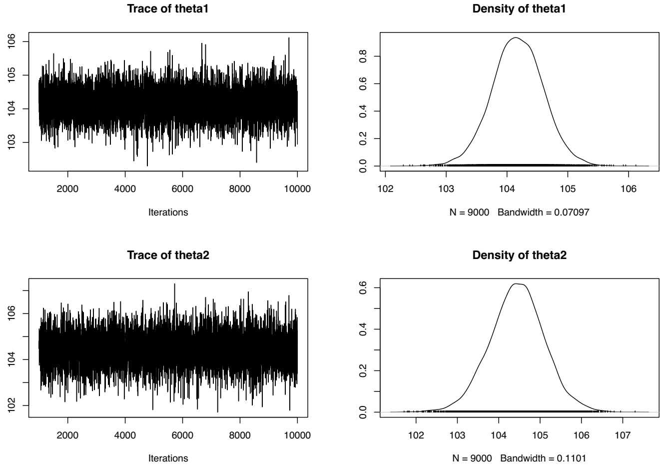
MCMC diagnostics

```r
THETA.mcmc <- mcmc(THETA[-(1:nburnin),],start=nburnin+1)
summary(THETA.mcmc)
```

```
##
## 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
## theta1  104 0.422  0.00445        0.00509
## theta2  104 0.662  0.00698        0.00829
##
## 2. Quantiles for each variable:
##
##         2.5% 25% 50% 75% 97.5%
## theta1  103 104 104 104   105
## theta2  103 104 104 105   106
```

Trace plots for $\boldsymbol{\theta}$

```r
plot(THETA.mcmc)
```



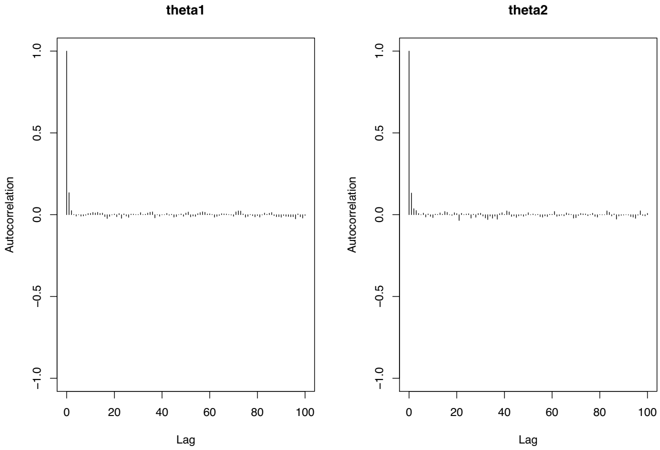
ESS for $\boldsymbol{\theta}$

```r
effectiveSize(THETA.mcmc)
```

```
## theta1 theta2
##   6864   6394
```

Autocorrelation plot for $\boldsymbol{\theta}$

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



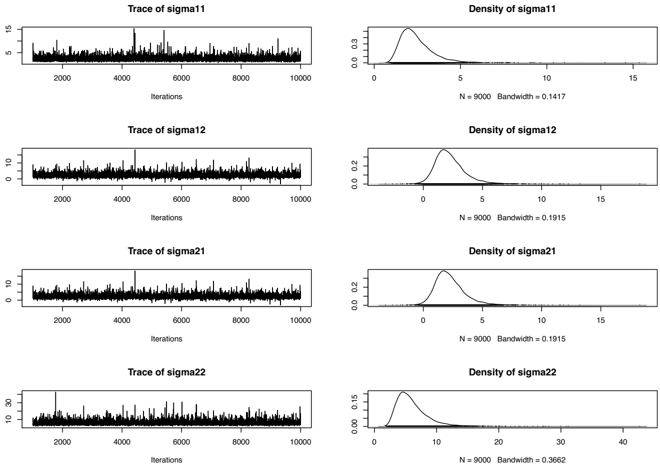
MCMC diagnostics

```r
SIGMA.mcmc <- mcmc(SIGMA[-(1:nburnin),],start=nburnin+1)
summary(SIGMA.mcmc)
```

```
##
## 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
## sigma11 2.45 0.982   0.0103         0.0119
## sigma12 2.26 1.315   0.0139         0.0178
## sigma21 2.26 1.315   0.0139         0.0178
## sigma22 6.08 2.690   0.0284         0.0370
##
## 2. Quantiles for each variable:
##
##          2.5%  25%  50%  75% 97.5%
## sigma11 1.186 1.78 2.25 2.89  4.80
## sigma12 0.286 1.41 2.07 2.91  5.37
## sigma21 0.286 1.41 2.07 2.91  5.37
## sigma22 2.852 4.30 5.50 7.16 12.87
```

Trace plots for Σ

```r
plot(SIGMA.mcmc)
```



ESS for Σ

```
effectiveSize(SIGMA.mcmc)
```

```
## sigma11 sigma12 sigma21 sigma22
##    6831    5431    5431    5295
```

Autocorrelation plot for Σ

```r
autocorr.plot(SIGMA.mcmc,lag.max=100)
```